# IMPROVE ALGORITHMIC FAIRNESS OF RECIDIVISM PREDICTION

Akshit Nanda (MS18216)

Department of Mathematical Sciences, IISER Mohali

## Introduction

No model is perfect, i.e., they make errors/ incorrect predictions. We call the model biased if these errors cause a systematic disadvantage toward a particular group. Fairness, unwanted bias, and discrimination have always concerned humans. The increasing growth in the use of artificial intelligence in varied sectors directly or indirectly affects people. Therefore, the models must be bias-free and give accurate and fair predictions.

**Algorithmic Fairness** is the idea that algorithmic systems should behave or treat people without unjust or prejudicial treatment on the grounds of sensitive characteristics.

## Problem Statement

In 2017, Propublica found that COMPAS, a recidivism prediction algorithm used by judges in the United States, failed differently for African-American defendants than for white Americans.

| | White American | African-American |
|---|---|---|
| False Positive | 23.5% | 44.9 % |
| False Negative | 47.7 % | 28 % |

Table 1. Disparity in Predictions

The above table shows how miserably the algorithm failed to achieve fairness by almost twice the rate of positive recidivate for black defendants.

**To find the best method to increase the fairness in recidivism prediction and minimize the trade-off between accuracy and fairness.**

## Data Cleaning and Profiling

Recidivism data, a real-world dataset, is used for all the experiments. The US judiciary uses this data as a pre-trial risk assessment to decide whether a defendant is detained or released. Since using the dataset greatly impacts an individual's life, it is very important to ensure the assessment made are accurate and fair.

In the **data cleaning** stage, the unwanted features were first removed; second, the rows with at least one missing value were removed; and third, the two sensitive features, sex and race were combined. Ultimately, the cleaned dataset had 7214 instances, nine features, and one label.

In the **data profiling** stage, it was observed that the base rate for the African-American males was much lower than that of all other groups, indicating clear biases towards African-American defendants.

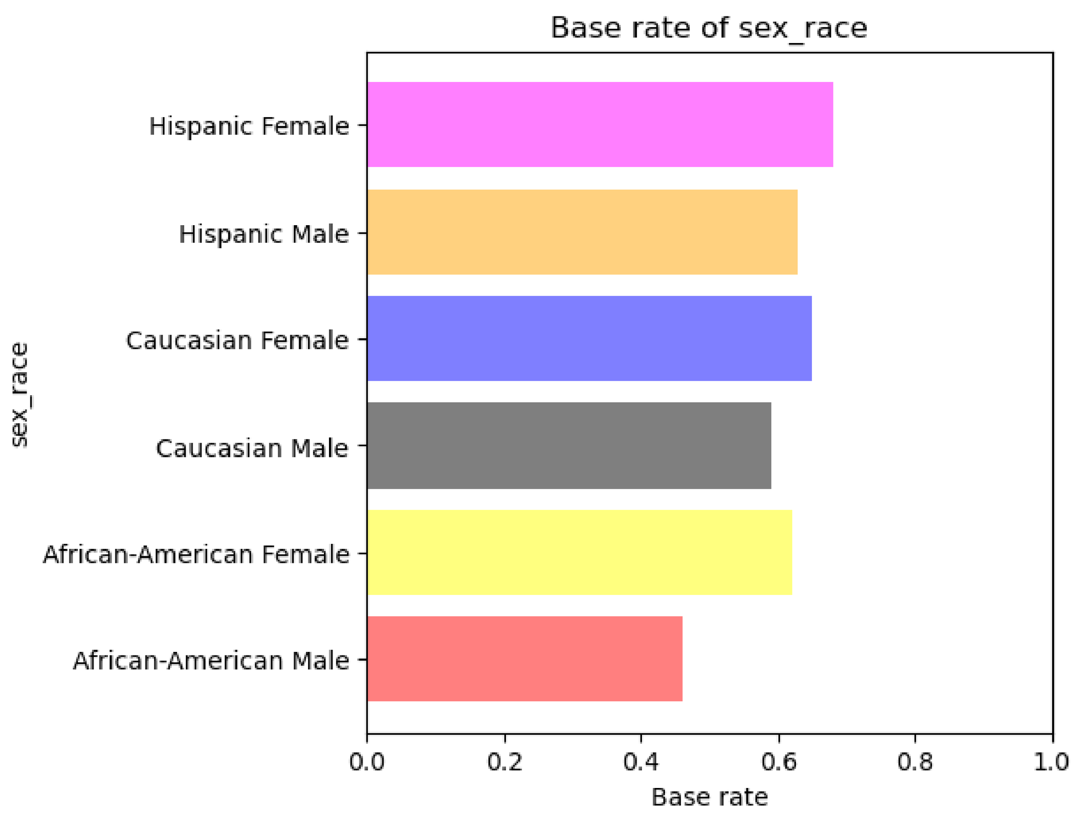


Figure 1. Base rate

## Definitions

1. **True positive**: The number of cases correctly classified as positive by the model.
2. **True negative**: The number of cases correctly classified as negative by the model.
3. **False positive**: The number of cases incorrectly classified as positive by the model.
4. **False negative**: The number of cases incorrectly classified as negative by the model.
5. **Accuracy**: For accuracy, the Balanced accuracy score is used to evaluate the performance of the classification algorithm. It is mathematically formulated as follows:

$$BAC = \frac{1}{2}\left(\frac{TP}{TP+FN}\frac{TN}{TN+FP}\right)$$

6. **Fairness**: For fairness, the Disparate impact ratio is used to evaluate the biases of the classification algorithm. It is mathematically calculated as follows:

$$DI = \frac{P(\hat{Y}=1|G=0)}{P(\hat{Y}=1|G=1)}$$

## Fairness-enhancing Techniques

**Pre-processing**
The data is modified before feeding into the machine learning algorithm in this technique. For example, CND, DIR

**In-processing**
The machine learning algorithm is modified to mitigate biases while training the model. For example, PR

**Post-processing**
In these, we take a subset of samples and change their predicted labels to meet the fairness requirement. For example, IGD, TO

**CND**
This aims to improve fairness by changing the labels of the instances that are nearest to the decision boundary since these are most prone to be falsely classified.

**DIR**
This aims to improve DI by changing the values of the features to have more equality between the underprivileged and privileged groups.

**Prejudice Remover(PR)**
This aims to remove the prejudice in the data, i.e., the dependence of the sensitive feature on the predicted labels. This is done by adding a regularization term.

**IGD**
This method aims to achieve fairness by choosing samples from a particular group with high individual bias and predicting its outcome by flipping the group membership.

**Threshold Optimizer(TO)**
This method aims to find the optimal thresholds for each group to predict the positive outcome.

## Choice of Algorithm

We start with a baseline algorithm (a model without any interventions) to evaluate the performance of different fairness-enhancing methods. For this, we measure the stability of the four common binary classification machine learning algorithms, Logistic regression, Decision tree, Gaussian naive bayes, and Support vector machine.

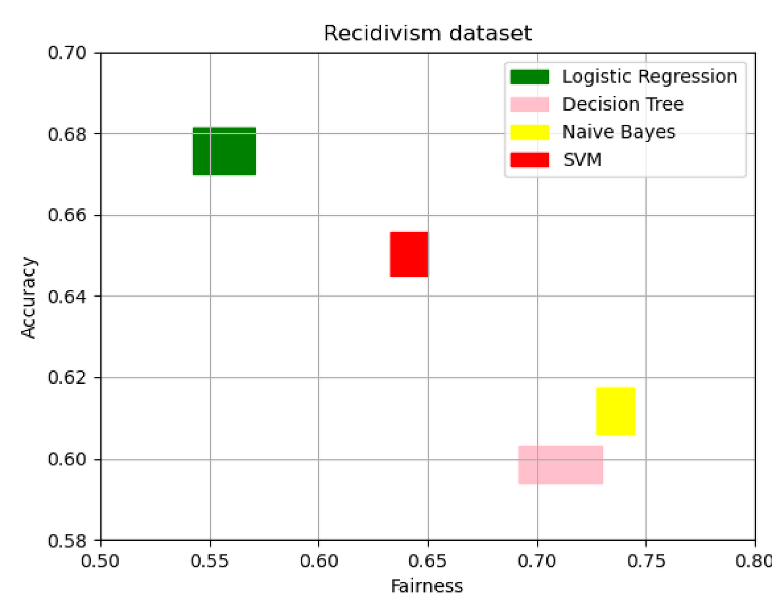


Figure 2. Stability of different algorithms

**Stability**
The stability of an algorithm is defined as the algorithm's performance, each tested on ten random train/test splits. A rectangle is drawn centered on the mean, and a width and height equal to the standard deviation along that measure are plotted.

## Baseline Algorithm

Logistic regression was chosen as it has the highest accuracy and low standard deviation along both accuracy and fairness measure.

Logistic regression uses a sigmoid function to get a probabilistic score between 0 and 1 and then uses a threshold to convert the score into binary. The model is represented mathematically by:

$$f_{w,b}(x) = \frac{1}{1+e^{-(wx+b)}}$$

The parameters w and b is optimized by maximizing the log-likelihood function:

$$log(L_{w,b}) = ln(L_{w,b}(x)) = \sum_{i=1,2,...,N} y_i ln(f_{w,b}(x)) + (1-y_i)ln(1-f_{w,b}(x))$$

## Result and Observations

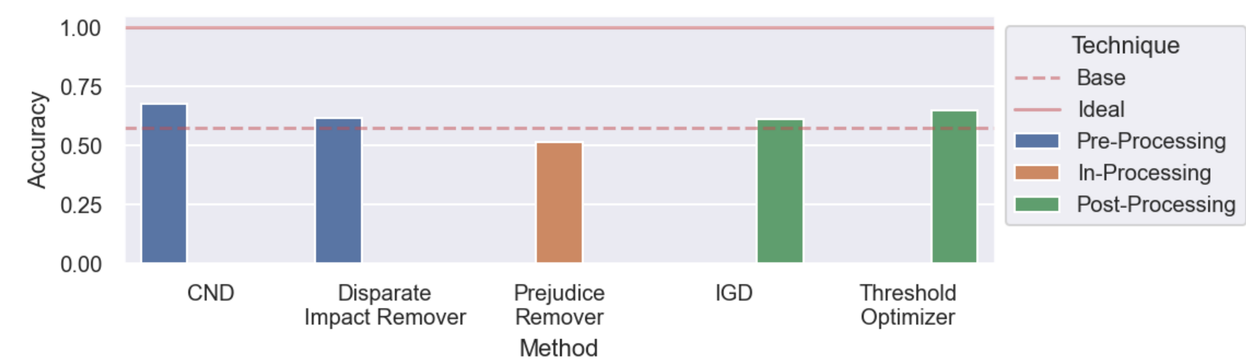


Figure 3. Accuracy of different fairness-enhancing methods

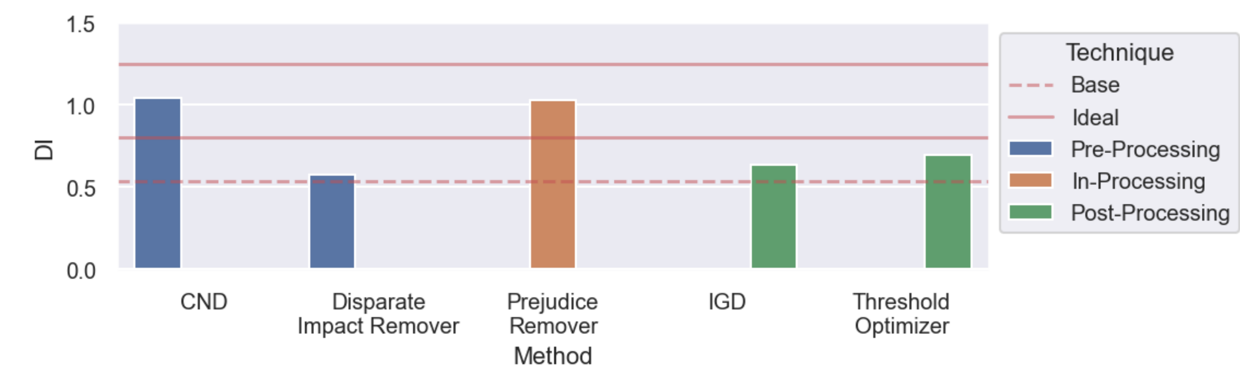


Figure 4. Fairness of different fairness-enhancing methods

In the above two figures, we observe that CND and Prejudice Remover perform the best in enhancing the fairness of the baseline algorithm and have higher accuracy. A clear trade-off can also be seen, indicating it's impossible to achieve high accuracy and fairness simultaneously.

## References

[Fel15] Michael Feldman, *Computational fairness: Preventing machine-learned discrimination*, Ph.D. thesis, 2015.

[FFM+15] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian, *Certifying and removing disparate impact*, proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 259–268.

[FSV+19] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth, *A comparative study of fairness-enhancing interventions in machine learning*, Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 329–338.

[HAÅ+20] Knut T Hufthammer, Tor H Aasheim, Sølve Ånneland, Håvard Brynjulfsen, and Marija Slavkovik, *Bias mitigation with aif360: A comparative study*, Norsk IKT-konferanse for forskning og utdanning, no. 1, 2020.

[HPS16] Moritz Hardt, Eric Price, and Nati Srebro, *Equality of opportunity in supervised learning*, Advances in neural information processing systems **29** (2016).

[JLA23] Lauren Kirchner Jeff Larson, Surya Mattu and Julia Angwin, *Compas recidivism risk score data and analysis*, 2023.

[LRB+19] Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri, *Bias mitigation post-processing for individual and group fairness*, Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp), IEEE, 2019, pp. 2847–2851.

[Wou22] Fenna Woudstra, *Algorithmic fairness: Which algorithm suits my purpose?*, 2022.

## Acknowledgements

I want to thank my supervisor Dr. Sarab Anand, Dr. Vaibhav Vaish, my family members, and my friends for their constant support.