Project 4

1. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

2. Fairness-aware Classifier with Prejudice Remover Regularizer

Group 4
Rishabh Ganesh, Qu Fei An, Rhoan Lee, Yerin Cho

Algorithm 1

Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

Algorithm 1: Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

Dataset and experimental setup

- List of all offenders screened through the COMPAS (Alternative Sanctions for Correctional Offender Management Profiling)
- offenders' demographic characteristics (gender, race, age)
- criminal history (charges arrested, number of previous offenses)
- risk scores assigned by COMPAS
- whether these individuals reoffended within two years of screening

Algorithm 1: Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

Analysis

Race	Yes	No	Total
Black	1,661(52%)	1,514(48%)	3,175(100%)
White	8,22(39%)	1,281(61%)	2,103(100%)
Total	2,483(47%)	2,795(53%)	5,278(100%)

Recidivism rates in ProPublica COMPAS data for both races

Feature	Description
Age Category	< 25, between 25 and 45, > 45
Gender	Male or Female
Race	White or Black
Priors Count	0–37
Charge Degree	Misconduct or Felony
2-year-rec. (target feature)	Whether (+ve) or not (-ve) the defendant recidivated within two years

Description of features used from ProPublica COMPAS data

Algorithm 1: Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

Algorithms

1. Fair Classifier with Disparate Mistreatment Constraints:
    - This classifier integrates fairness constraints directly into the training process. It aims to equalize misclassification rates across groups by modifying the decision boundary. The constraints are formulated as convex-concave programming problems, allowing for efficient optimization despite the non-convex nature of fairness constraints.

2. Baseline and Comparative Algorithms:
    - Baseline methods that don't incorporate fairness constraints, as well as other contemporary methods that aim to address fairness, such as those adjusting decision thresholds post-training to balance misclassification rates among groups.

Algorithm 1: Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

		Predicted Label		
		$\hat{y} = 1$	$\hat{y} = -1$	
True Label	$y = 1$	True positive	False negative	$P(\hat{y} \neq y \mid y = 1)$ False Negative Rate
	$y = -1$	False positive	True negative	$P(\hat{y} \neq y \mid y = -1)$ False Positive Rate
		$P(\hat{y} \neq y \mid \hat{y} = 1)$ False Discovery Rate	$P(\hat{y} \neq y \mid \hat{y} = -1)$ False Omission Rate	$P(\hat{y} \neq y)$ Overall Misclass. Rate

What this algorithm does

The method addresses fairness in binary classification tasks, where classifiers predict whether an instance belongs to one of two classes (-1 or 1).

1. Binary Classification and Sensitive Features:

The goal is to learn a function that maps user feature vectors to class labels, taking into account a sensitive feature, like race or gender, which should not influence the prediction outcomes.

2. Disparate Treatment:

This is avoided if the prediction made by a classifier does not change after observing the sensitive feature.

Formally, the probability $P(\hat{y}|x, z) = P(\hat{y}|x)$, meaning that the sensitive attribute z does not affect the predicted outcome $\hat{y}$.

3. Disparate Impact:

A classifier does not have a disparate impact if it assigns positive class labels to different groups (defined by the sensitive feature) at the same rate. Mathematically, $P(\hat{y} = 1 \mid z = 0) = P(\hat{y} = 1 \mid z = 1)$.

Algorithm 1: Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

Evaluation Methodology

1. Metrics:
    - The main metrics used for evaluation are **accuracy** and **fairness**, specifically examining **false positive** and **false negative rates** across different groups. The paper strives to achieve a balance where modifying the decision boundaries to reduce bias does not unduly sacrifice the overall accuracy of the classifiers.

2. Multiple Fairness Notions:
    - The evaluation also considers different notions of fairness simultaneously, such as ensuring that both the false positive and false negative rates are balanced across groups, which is challenging due to inherent trade-offs in classifier design.

Algorithm 1: Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

Main Results

Racial Group	Accuracy	False Positive Rate (FPR)	False Negative Rate (FNR)	Disparate Impact (DFPR)	Disparate Impact (DFNR)
Blacks	0.668	0.35	0.31	0.18	-0.30
Whites	0.668	0.17	0.61		

→ The bias inherent in the model's predictions
→ showing how different groups are unfairly treated either by higher false positives or false negatives.

Algorithm 1: Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

Main Results

1. Effectiveness of Fairness Constraints:
    - The results show that integrating fairness constraints into classifiers effectively reduces disparate mistreatment with a minimal decrease in overall accuracy. For instance, classifiers trained with these constraints could achieve fairness in terms of balanced misclassification rates without significant losses in predictive performance.

2. **Trade-offs** Between **Fairness** and **Accuracy**:
    - While the fair classifiers manage to reduce bias, there is a notable, though often small, trade-off in terms of accuracy. The extent of this trade-off depends on the stringency of the fairness constraints applied during training.

Algorithm 1: Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

	Feature	Importance
4	is_recid	0.770981
3	decile_score	0.047353
1	priors_count	0.045542
0	age	0.043250
5	jail_time_days	0.034544
9	score_text_High	0.021395
2	days_b_screening_arrest	0.014142
6	c_charge_degree_F	0.007136
11	sex_Male	0.004780
10	score_text_Medium	0.004327
8	age_cat_Greater than 45	0.003281
7	age_cat_25 - 45	0.003270

Feature Importance

- #1: ‘is_recid’
- Other features → significantly less important

Potential issues within the justice system itself

- Similar Accuracy when predicting, but we can observe an issue within the treatment of the two groups of people.

Group	Validation Accuracy	Actual Recidivism Cases	Total Cases	Actual Recidivism Rate	Predicted Recidivism Cases	Predicted Recidivism Rate
Total	**0.9574**	**471**	**1056**	**0.4460**	**508**	**0.4811**
White	**0.9680**	**176**	**437**	**0.4027**	**190**	**0.4348**
Black	**0.9499**	**295**	**619**	**0.4766**	**318**	**0.5137**

Algorithm 2

Fairness-aware Classifier with Prejudice Remover Regularizer

Algorithm 2: Fairness-aware Classifier with Prejudice Remover Regularizer

Dataset and Experimental Setup

- The evaluation was performed using a real dataset previously utilized in studies by Calders and Verwer.
- Includes attributes like age, work class, education, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, and native country
- These features are used to predict the likelihood of an individual earning more than $50,000 per year.
- This dataset helps in understanding **how effectively each model handles biases** related to gender in income predictions.
- Sensitive and Target Features: The primary sensitive feature of interest in the dataset is gender. The target variable is whether an individual's income exceeds $50,000 a year.

Algorithm 2: Fairness-aware Classifier with Prejudice Remover Regularizer

Algorithms Discussed

Prejudice Remover Regularizer:

This algorithm introduces a regularization component in the training process of probabilistic discriminative models like **logistic regression**. It is specifically designed to minimize the dependency of predictions on sensitive features, thereby reducing both direct and indirect biases. The algorithm penalizes the model for using sensitive features excessively, aiming to make the outcomes more equitable.

→ We used **sklearn logistic regression** & **Fairness aware classifier with prejudice remover regularizer**

Algorithm 2: Fairness-aware Classifier with Prejudice Remover Regularizer

sklearn Logistic Regression

- We make a Logistic Regression model that predicts `two_year_recid` without any prejudice remover.
- Notice that without any prejudice removal, the model predicts recidivism more accurately for Caucasians compared to African Americans.

Accuracy for privileged group (race == 1)	0.9759576202118989
Accuracy for unprivileged group (race == 0)	0.963474025974026

Without the `is_recid` feature, the accuracies from the Logistic Regression model for the privileged and unprivileged groups both decrease significantly, while the difference between the two accuracies remains almost the same.

Accuracy for privileged group (race == 1)	0.7224938875305623
Accuracy for unprivileged group (race == 0)	0.7126623376623377

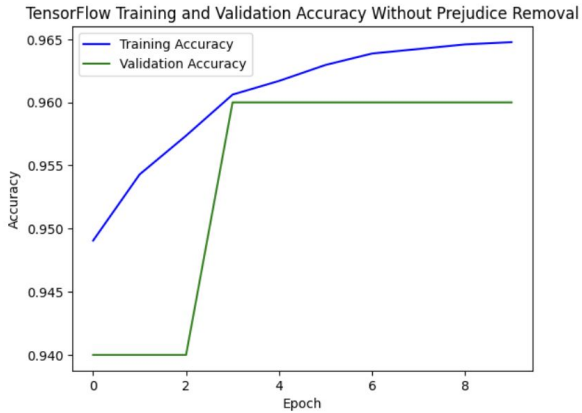
Store the class Y, the non-sensitive features X, and the sensitive feature S separately.

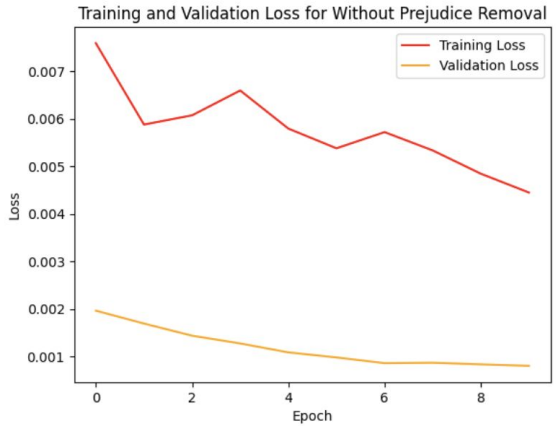
Algorithm 2: Fairness-aware Classifier with Prejudice Remover Regularizer

Fairness Aware Classifier with Prejudice Remover Regularizer

- We implement the prejudice removal regularizer as a loss function for our logistic regression classifier.
- To make the loss function compatible with our model, we normalize the training and validation data.

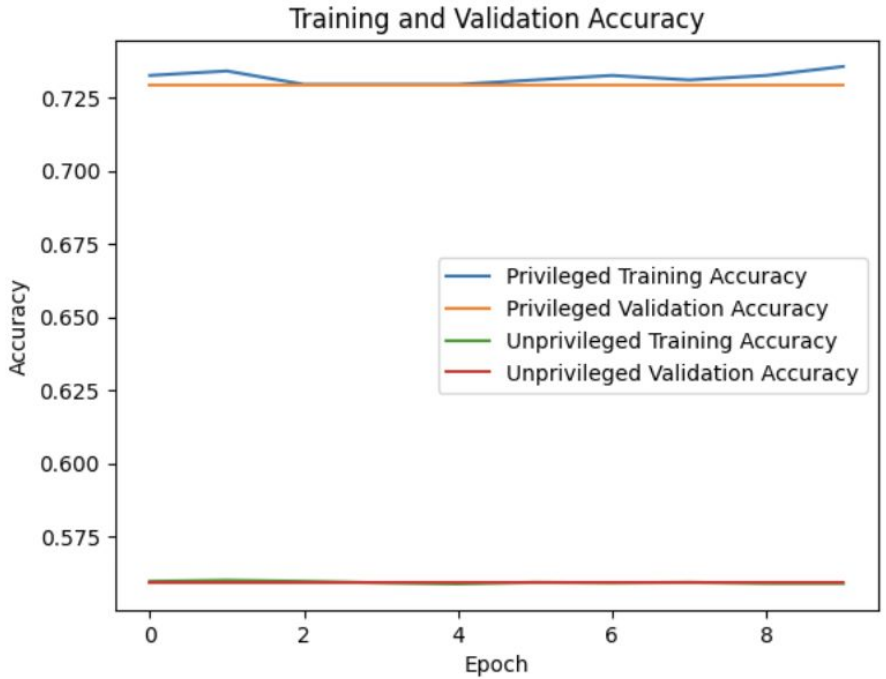
- Here, we visualize the performance on our data before using the loss function that incorporates the prejudice index calculation.
- Notice how the accuracy steadily improves and the loss decreases throughout training. We interpret this as a sign that the model is overfitting to the training data.





Algorithm 2: Fairness-aware Classifier with Prejudice Remover Regularizer

Evaluation



Algorithm 2: Fairness-aware Classifier with Prejudice Remover Regularizer

Main Results

- Performance of Prejudice Remover Regularizer:
  The Prejudice Remover Regularizer was effective in reducing bias without causing significant loss in prediction accuracy. It showed better performance in balancing accuracy and fairness compared to logistic regression and naïve Bayes models, especially when sensitive features were included.

Algorithm 2: Fairness-aware Classifier with Prejudice Remover Regularizer

Conclusion

- From evaluating the model, there is significant parity between the privileged and unprivileged groups based on the accuracy.
- Our model is overfitting to label the unprivileged group as "good" or in this case, "negative" for two year recidivism.
- As a result, the model incorrectly predicts 30% of the unprivileged group while maintaining its original accuracy for the privileged group, without using `is_recid` as a feature.
- We believe the model is "too aware" of fairness due to the overfitting, causing it to misclassify a significant portion of the unprivileged group as negative for `two_year_recid`.
- Using `is_recid` as the sole feature for predicting `two_year_recid` provided the most accurate classification with the most equal parity for us, despite being algorithmically unaware of fairness.