

ProPublica recidivism Use Case (Compas risk assessment tool)

Transparency & Fairness in AI and Big Data Algorithms

LYOUSFI Youssef

Dimitris KOTZINOS
Vassilis CHRISTOPHIDES

M2 Data Science & Machine Learning

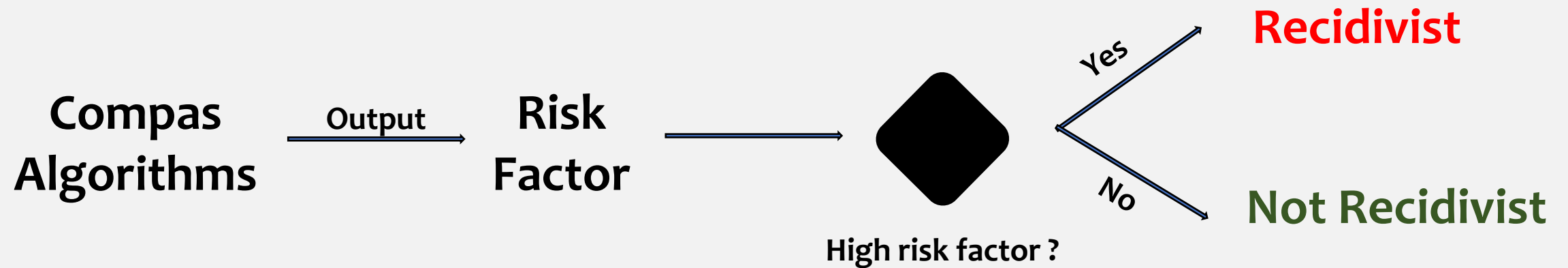
03/03/2023

Overview

- **Compas algorithm**
 - **Bias Detection & Mitigation**
 - **Reweighting preprocessing**
 - **Optimized preprocessing**
 - **Adversarial Inprocessing**
 - **Calibrated Odds**
- PostProcessing**

Compas

Correctional Offender Management Profiling for
Alternative Sanctions



The higher the risk factor, the more you become a recidivist

(predicted to reoffend, but don't)

False Positive

Black: 44.9%

White: 23.5%

(predicted not to reoffend, but did)

False Negative

Black: 28.1%

White: 47.7%

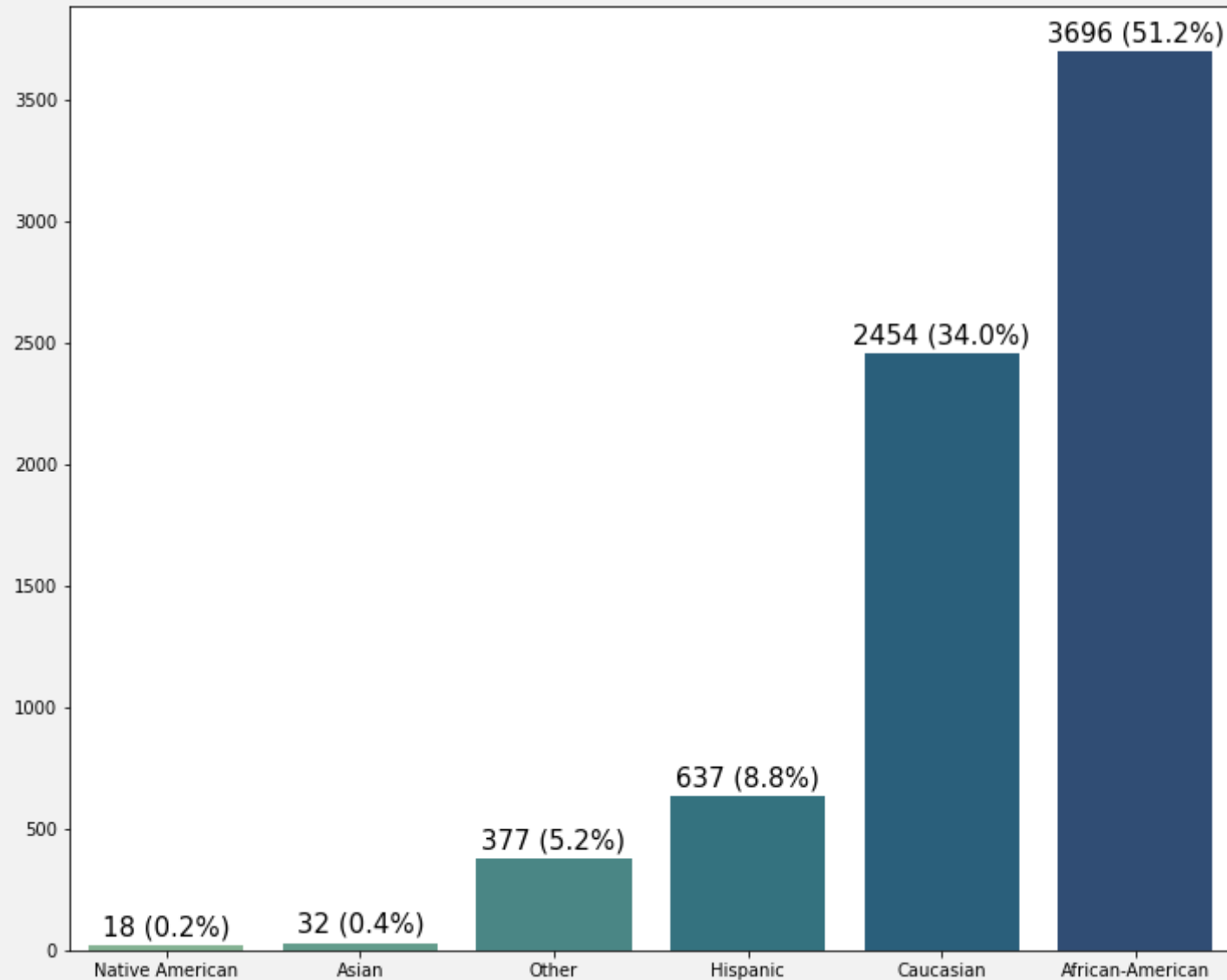
**If you are black, you are more likely to be assessed as high risk
than if you were white**

The accuracy, fairness, and limits of predicting recidivism | Science Advances



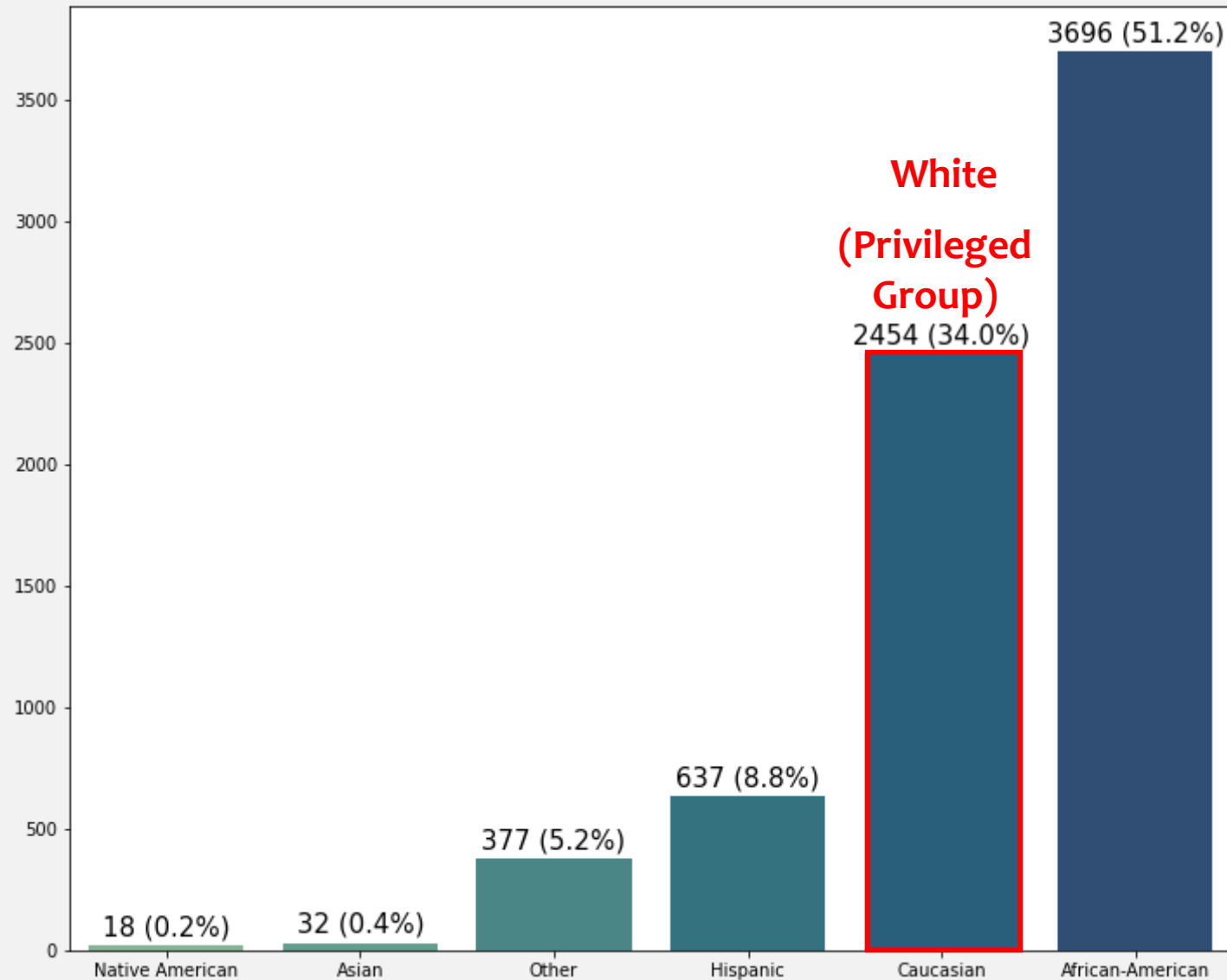
Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that

Machine Bias — ProPublica



Race attribute distribution

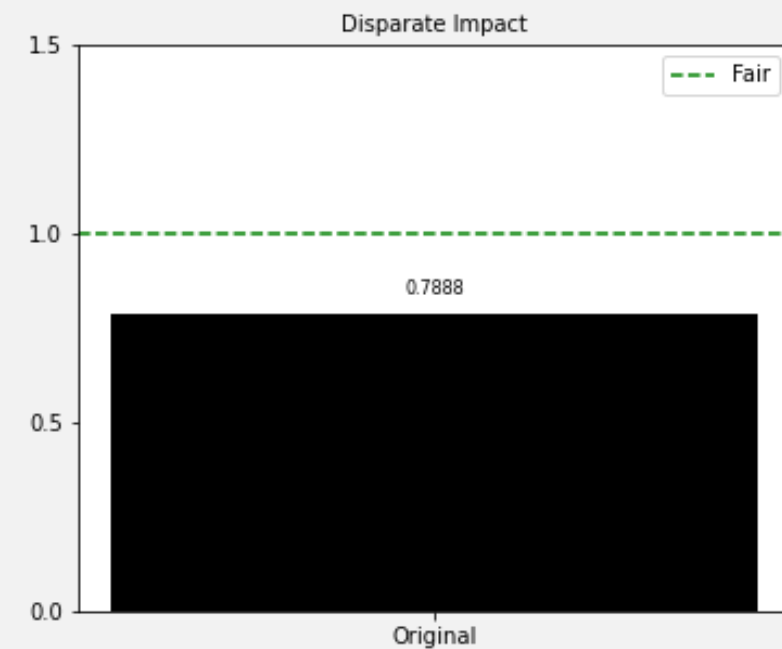
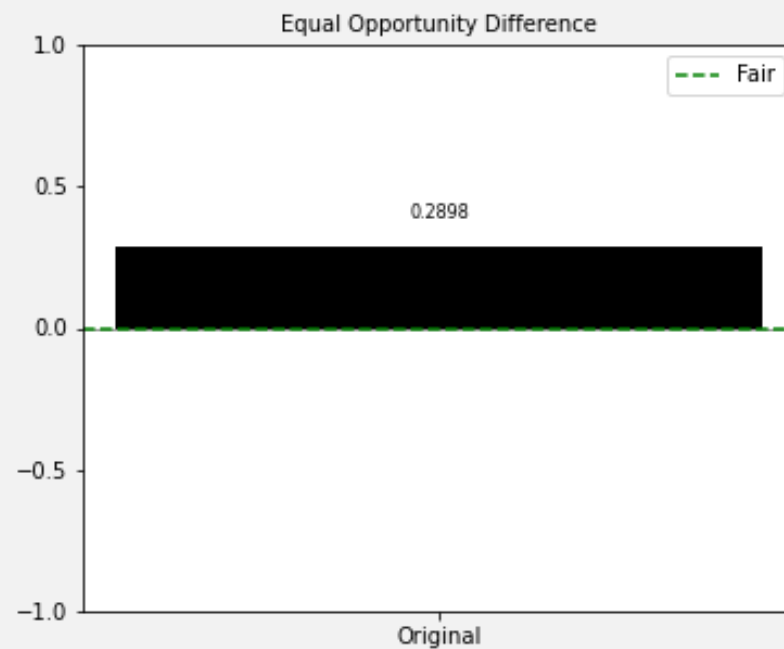
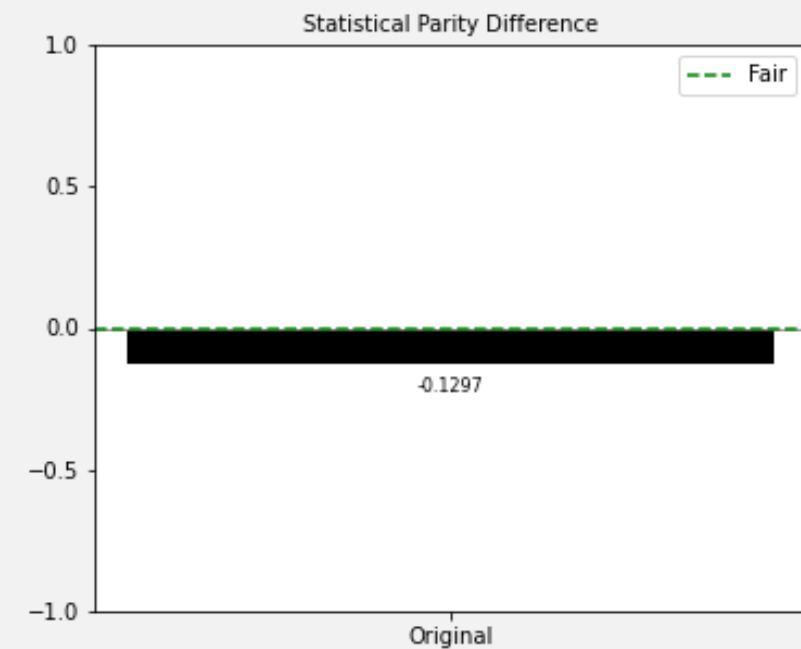
ProPublica recidivism dataset



Race attribute distribution

ProPublica recidivism dataset

Bias Detection & Mitigation



Fairness Metrics on Original Dataset

| Pre-Processing Algorithms Mitigate bias in training data | In-Processing Algorithms Mitigate bias in classifiers | Post-Processing Algorithms Mitigate bias in predictions |
|---------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------|
| Reweighting Modifies the weights of different training examples | Adversarial Debiasing Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions | Reject Option Classification Changes predictions from a classifier to make them more fair |
| Disparate Impact Remover Edits feature values to improve group fairness | Prejudice Remover Adds a discrimination-aware regularization term to the learning objective | Calibrated Equalized Odds Optimizes over calibrated classifier score outputs that lead to fair output labels |
| Optimized Preprocessing Modifies training data features and labels | Meta Fair Classifier Takes the fairness metric as part of the input and returns a classifier optimized for the metric | Equalized Odds Modifies the predicted label using an optimization scheme to make predictions more fair |
| Learning Fair Representations Learns fair representations by obfuscating information about protected attributes | | |

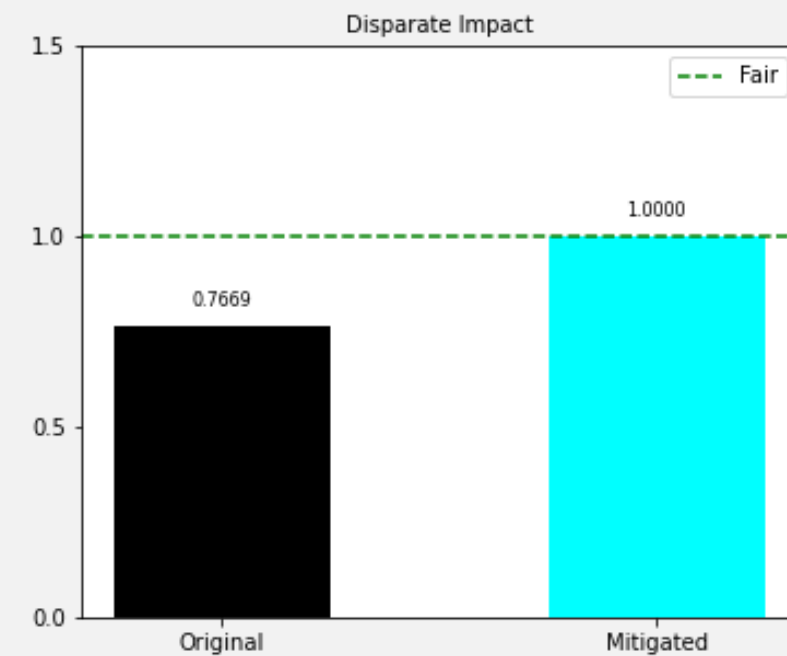
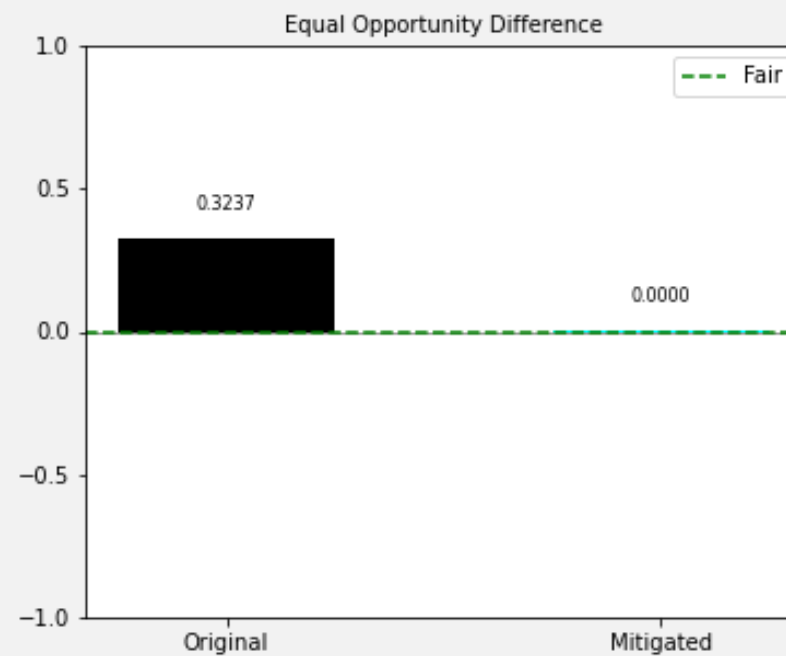
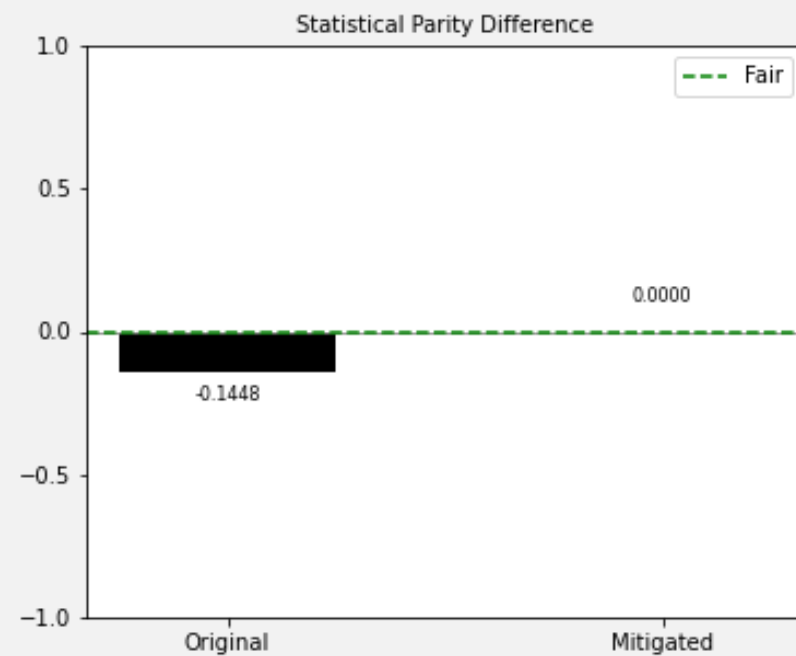
Bias Mitigation Algorithms

| Pre-Processing Algorithms Mitigate bias in training data | In-Processing Algorithms Mitigate bias in classifiers | Post-Processing Algorithms Mitigate bias in predictions |
|---------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------|
| Reweighting Modifies the weights of different training examples | Adversarial Debiasing Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions | Reject Option Classification Changes predictions from a classifier to make them more fair |
| Disparate Impact Remover Edits feature values to improve group fairness | Prejudice Remover Adds a discrimination-aware regularization term to the learning objective | Calibrated Equalized Odds Optimizes over calibrated classifier score outputs that lead to fair output labels |
| Optimized Preprocessing Modifies training data features and labels | Meta Fair Classifier Takes the fairness metric as part of the input and returns a classifier optimized for the metric | Equalized Odds Modifies the predicted label using an optimization scheme to make predictions more fair |
| Learning Fair Representations Learns fair representations by obfuscating information about protected attributes | | |

The Algorithms used

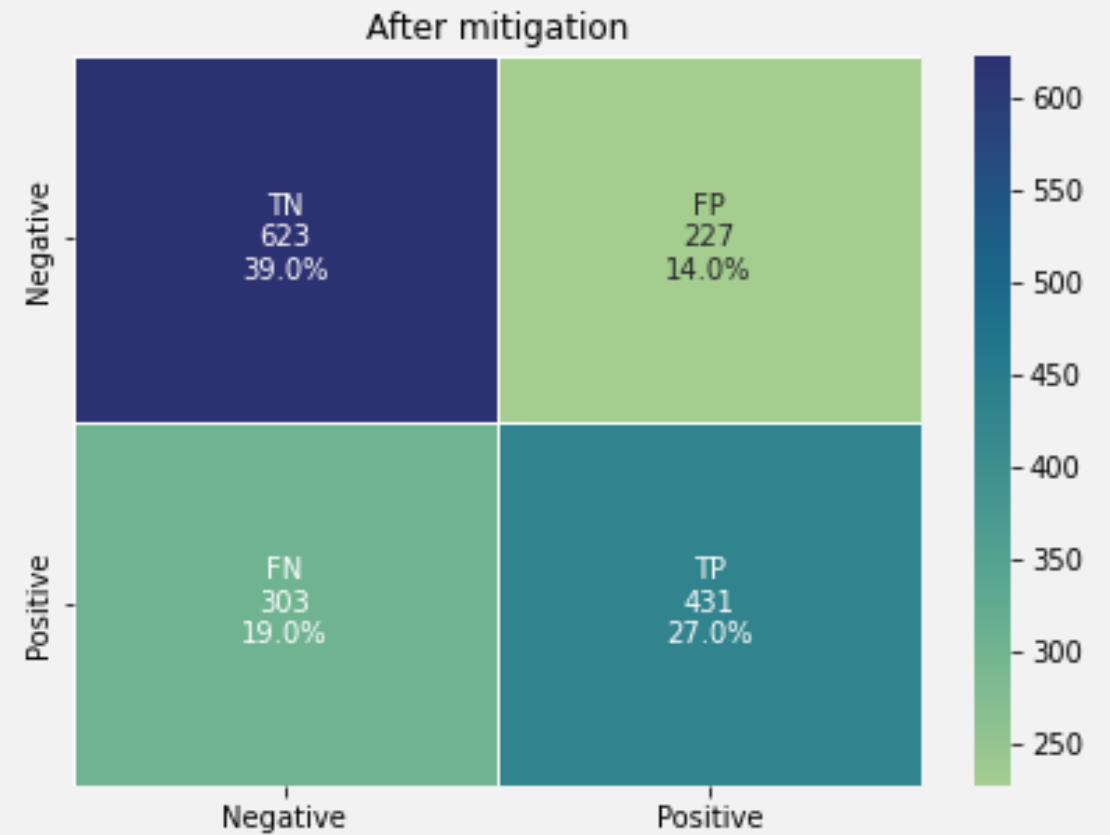
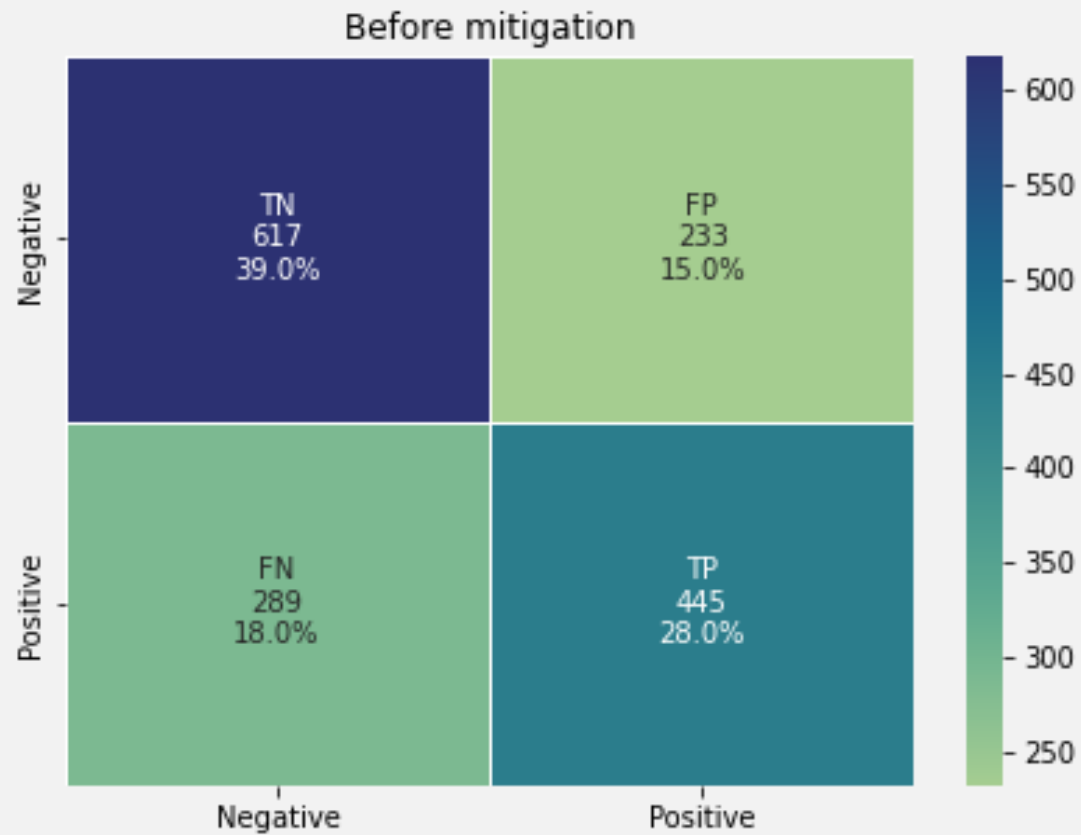
Bias Detection & Mitigation

Reweighting Preprocessing



Fairness Metrics after Reweighting technique

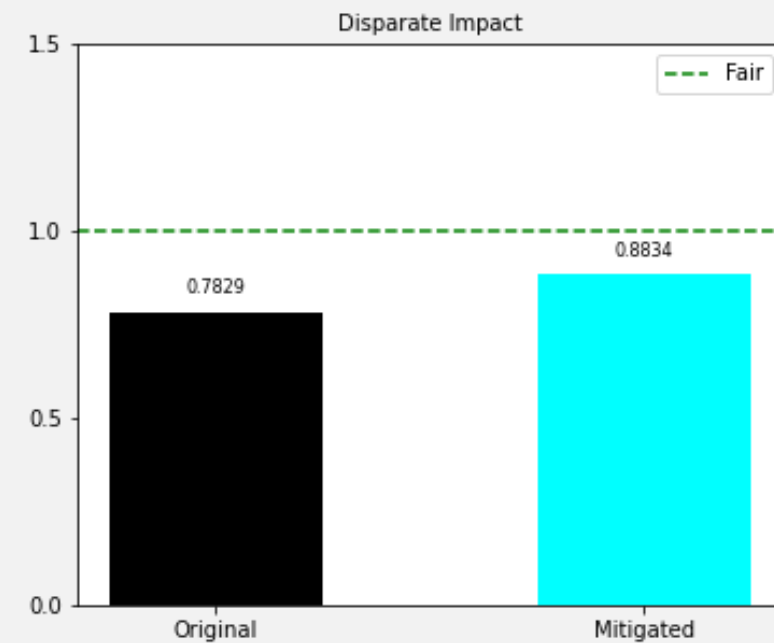
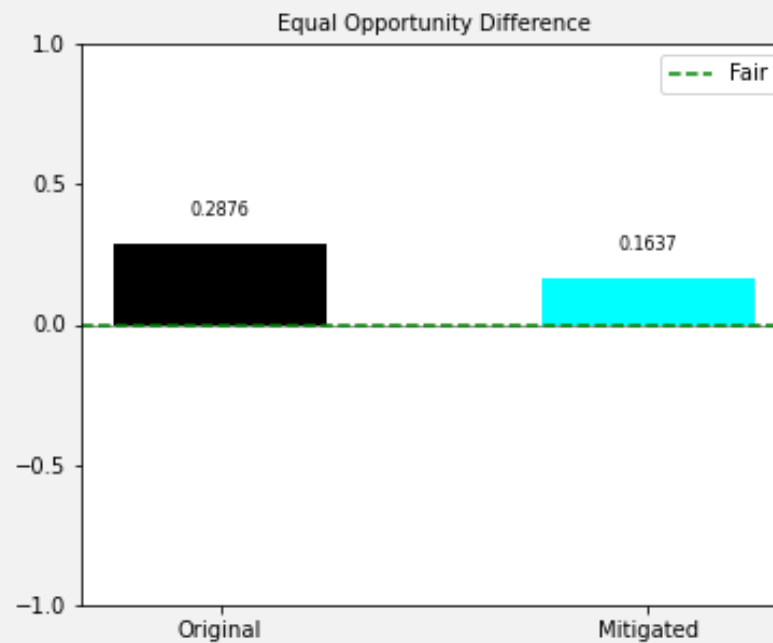
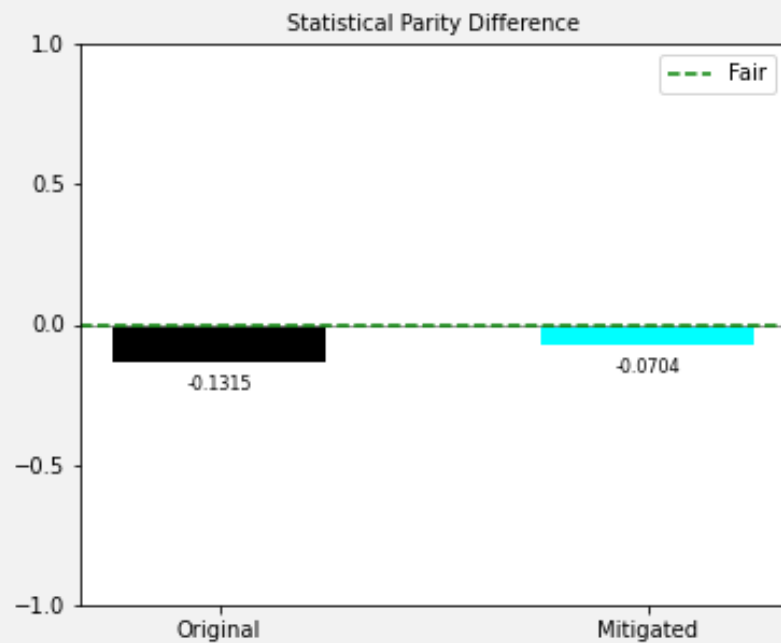
Reweighting processing



Confusion Matrix for Random Forest

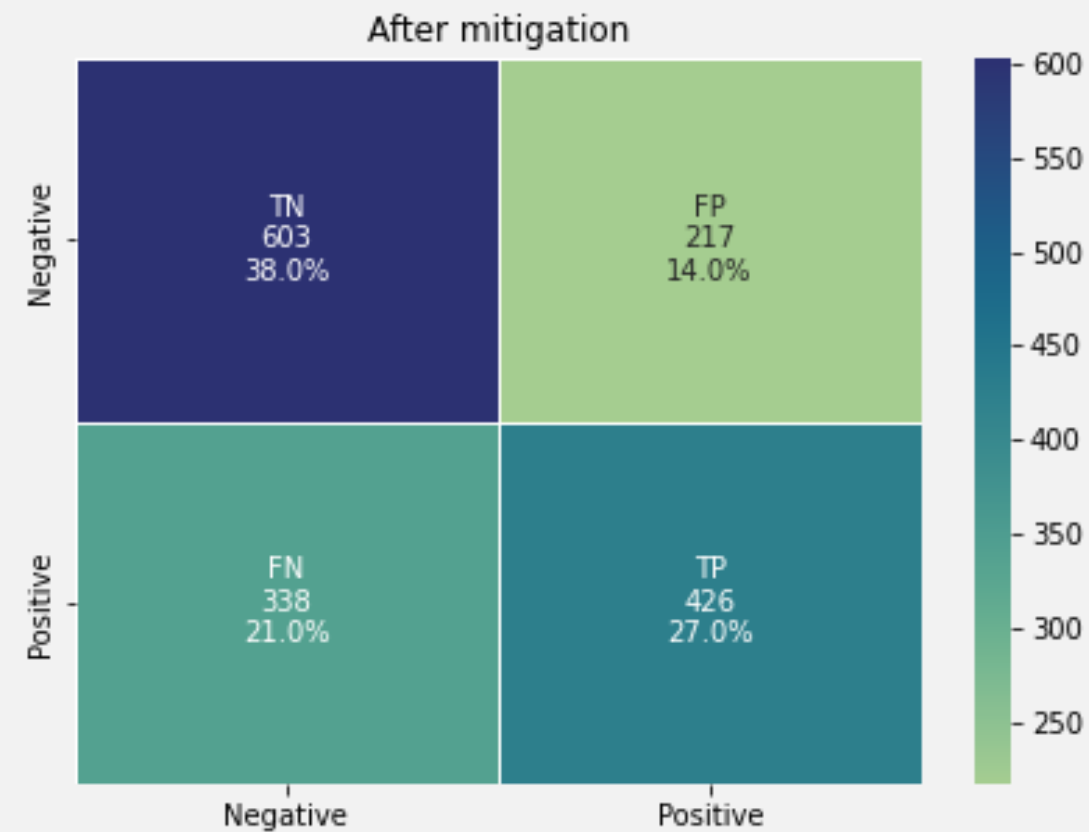
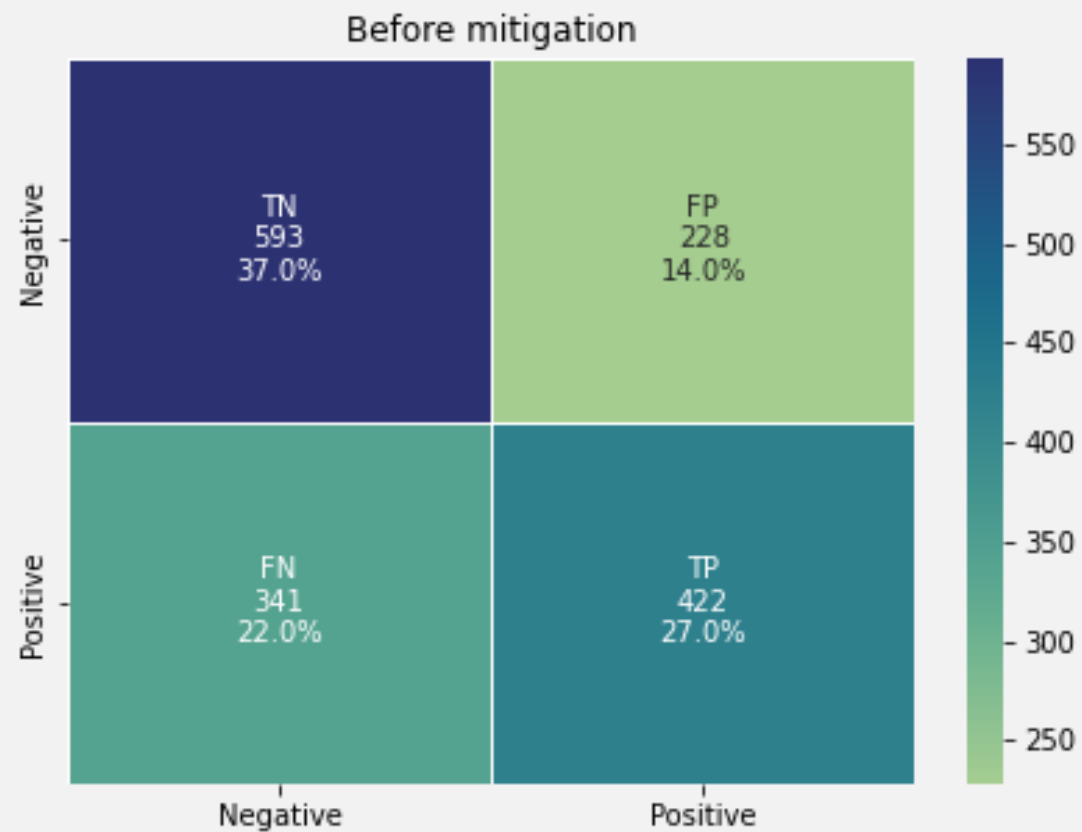
Reweighting processing

Optimized Preprocessing



Fairness Metrics after Optimized technique

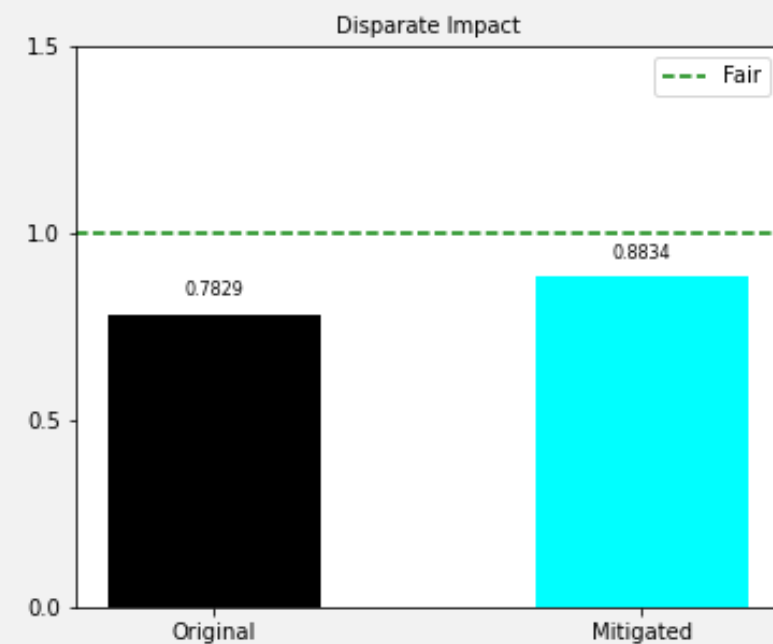
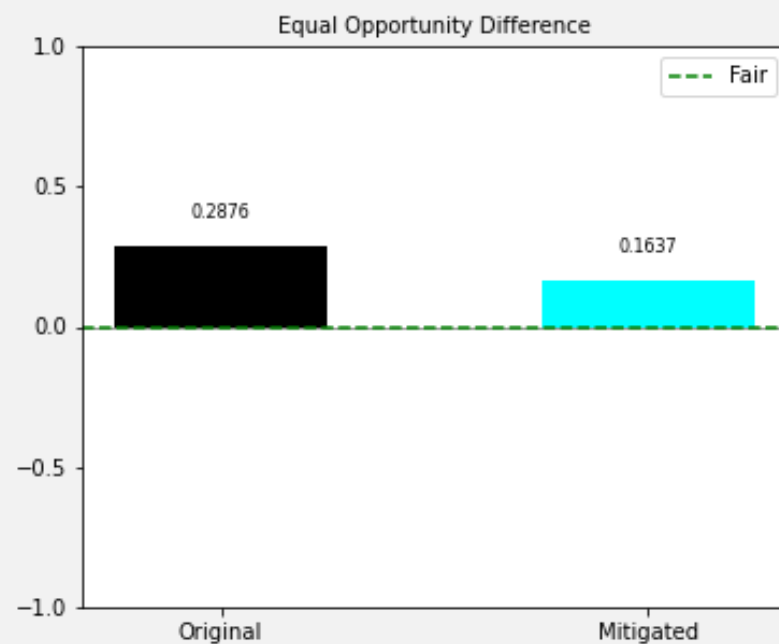
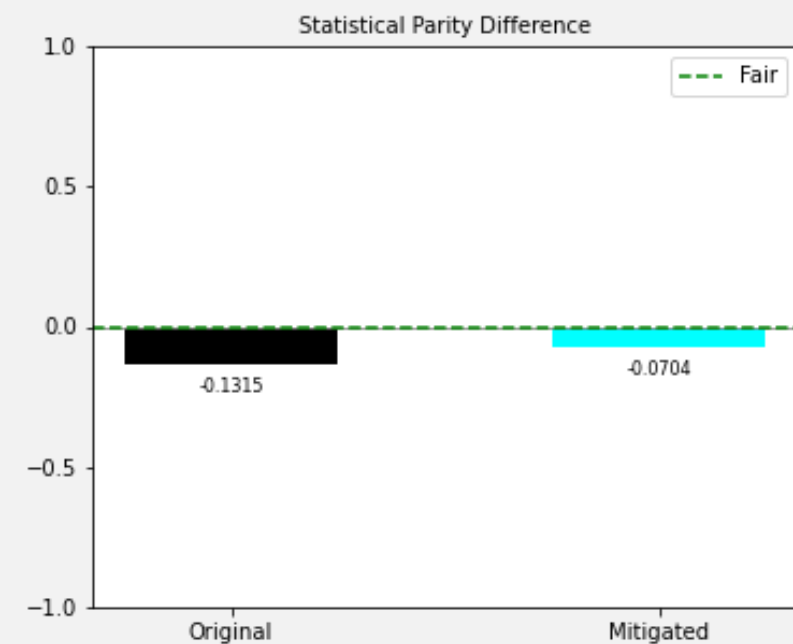
Optimized processing



Confusion Matrix for Random Forest

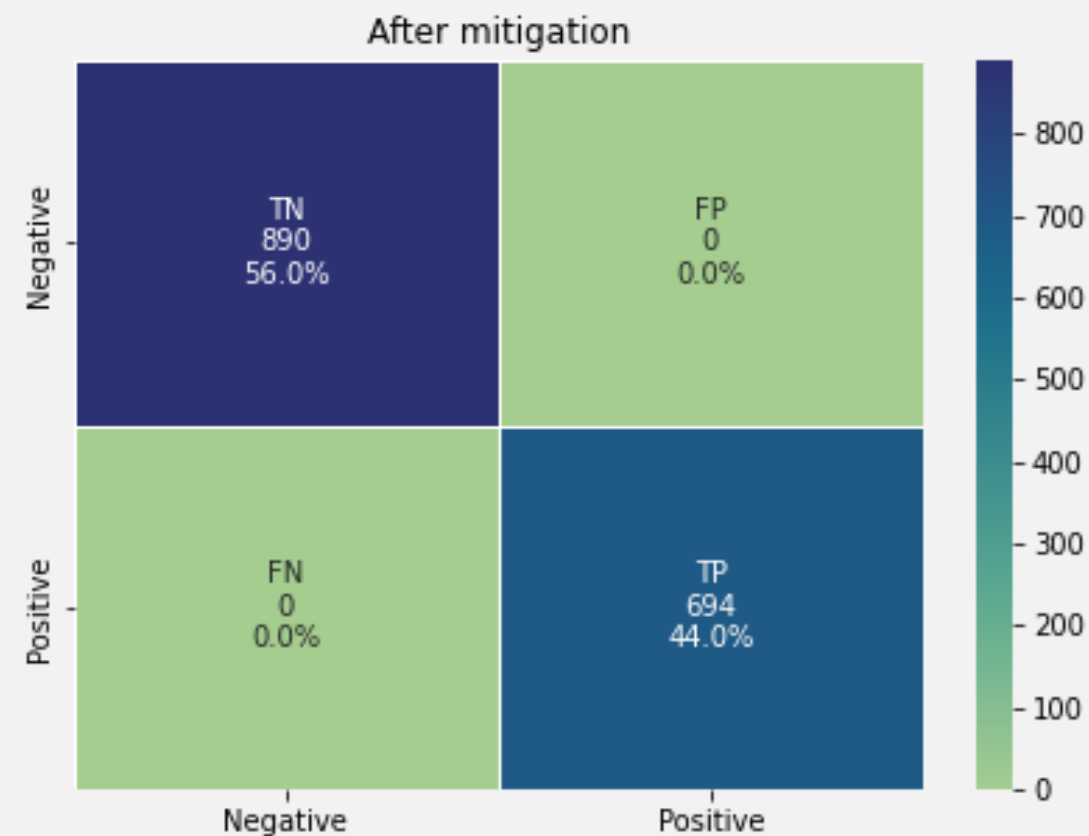
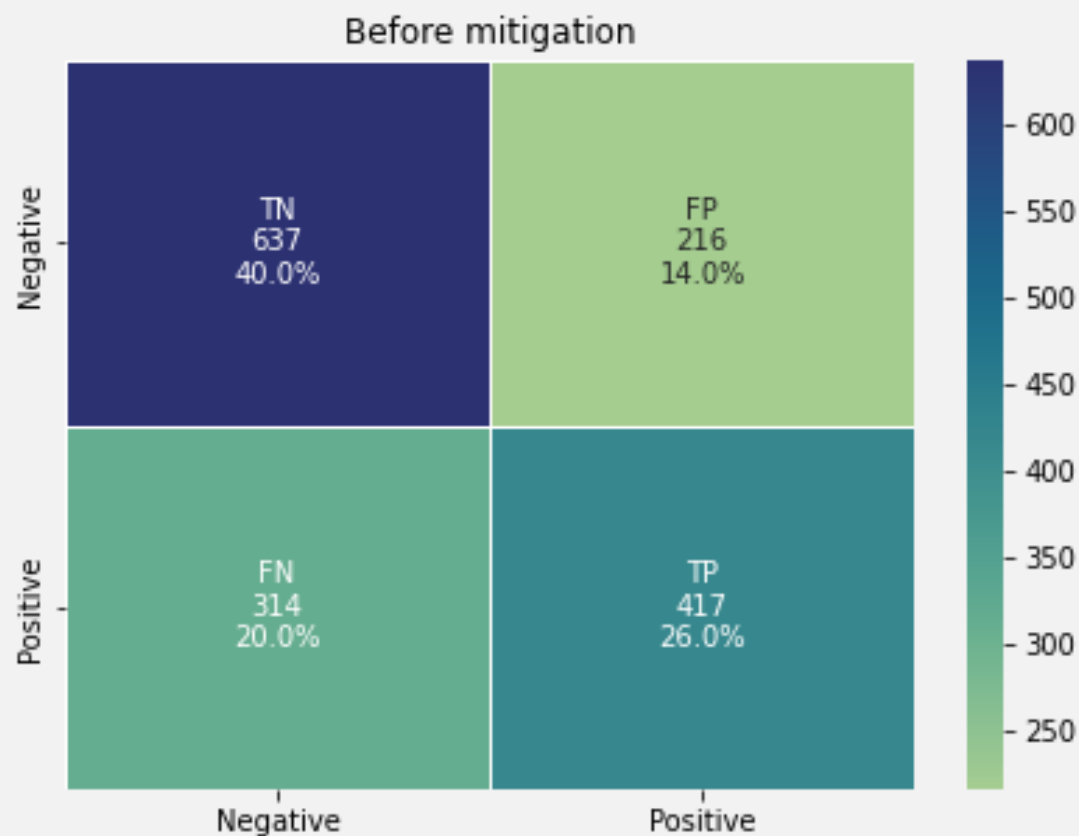
Optimized processing

Adversarial Inprocessing



Fairness Metrics after Adversarial technique

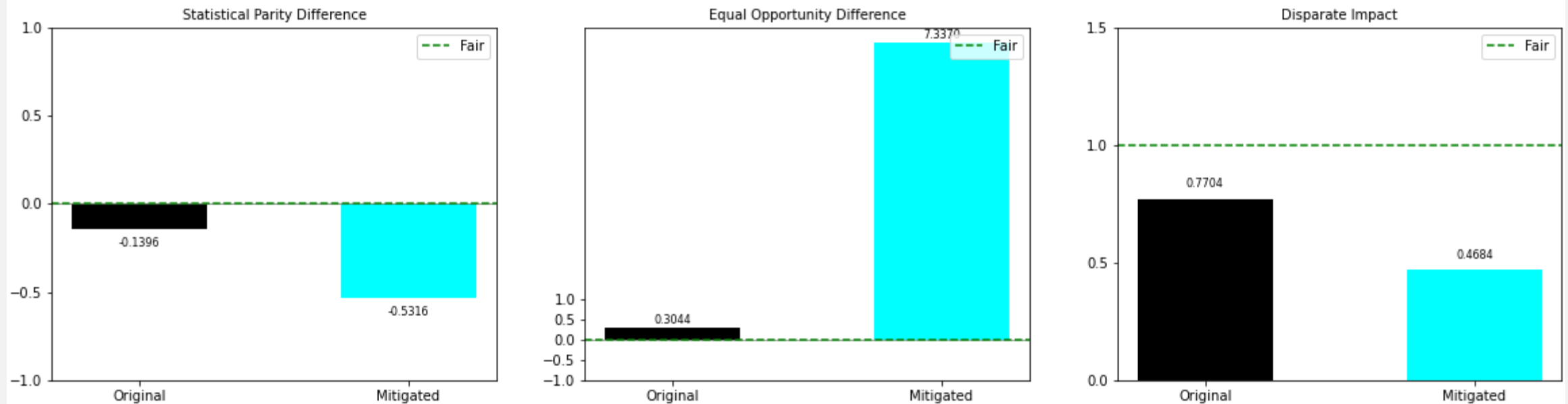
Adversarial processing



Confusion Matrix for Random Forest

Calibrated Odds post- processing

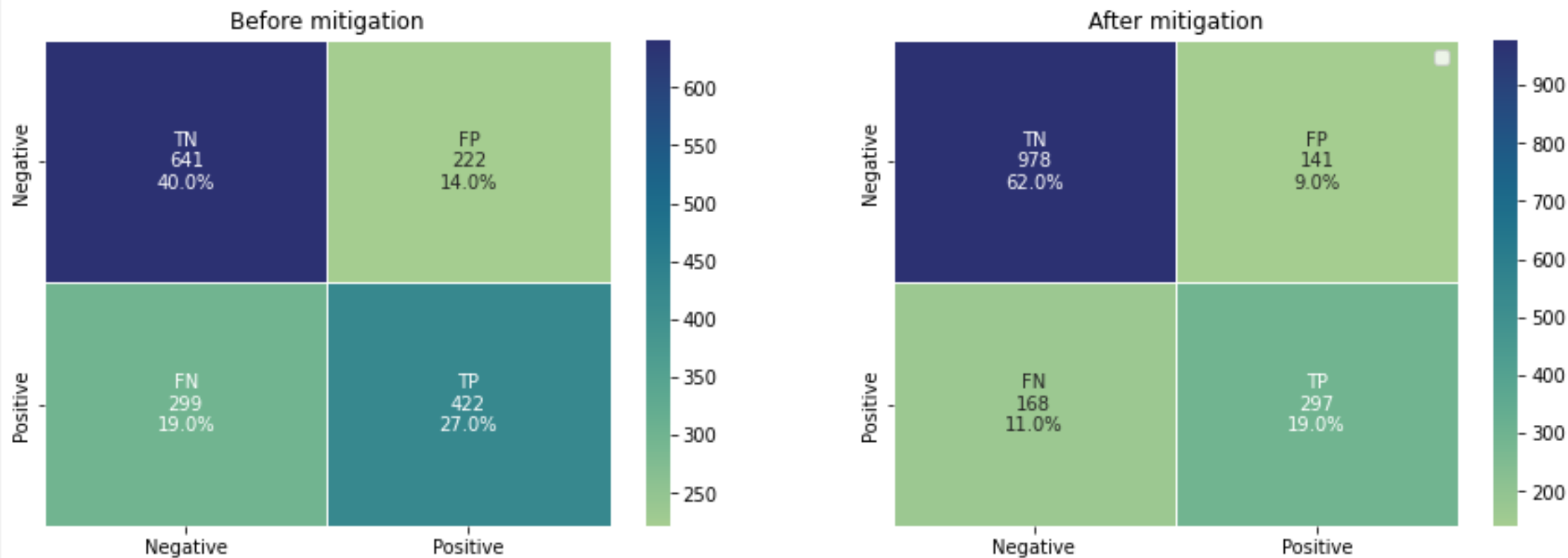
Fairness Metrics (Calibrated Odds)



Fairness Metrics after Calibrated Odds technique

Calibrated Odds processing

Confusion Matrix for Random Forest



Confusion Matrix for Random Forest

Calibrated Odds processing