Evaluating Bias Detection and Mitigation in Student Failure and Crime Datasets Using Fairlearn and AIF360

Bernhard Kollmann, Clemens Thrakl

Scope of the work

- 2 Frameworks - AIF360 | Microsoft Fairlearn
- 2 Datasets - Student Failures | Crime Statistic
- 2 Iterations - Raw | Cleaned Data
- Sensitive Features Alone | Groups
- Mitigation Steps Individually | Iteratively

Research Quesition

- How effective are Fairlearn and AIF360 in detecting and mitigating bias in student failure and crime datasets, and what insights can be derived regarding the socio-economic implications of these biases?

  - **Socioeconomic Nature and Relevance:**

    How do the socio-economic backgrounds of students and crime rates influence bias, and what is the relevance of addressing bias in these datasets for societal equity?

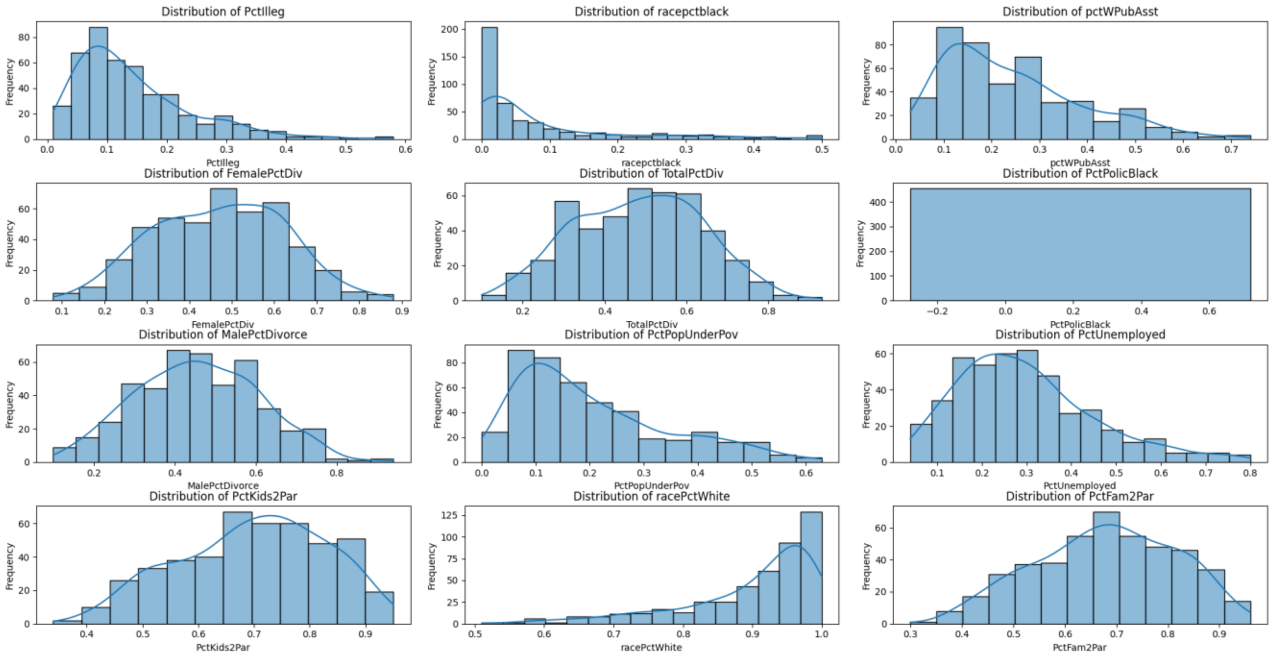
  - **Bias Detection and Mitigation:**

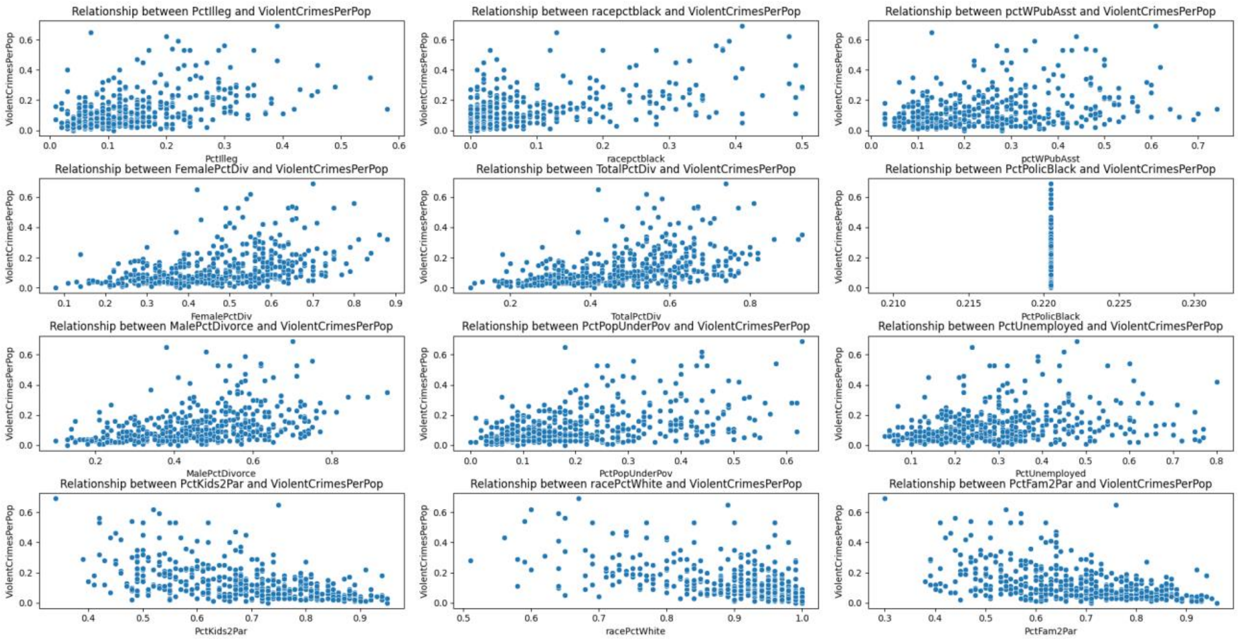
    Which key differences in bias detection and mitigation capabilities between Fairlearn and AIF360 when applied to student failure and crime datasets exist?

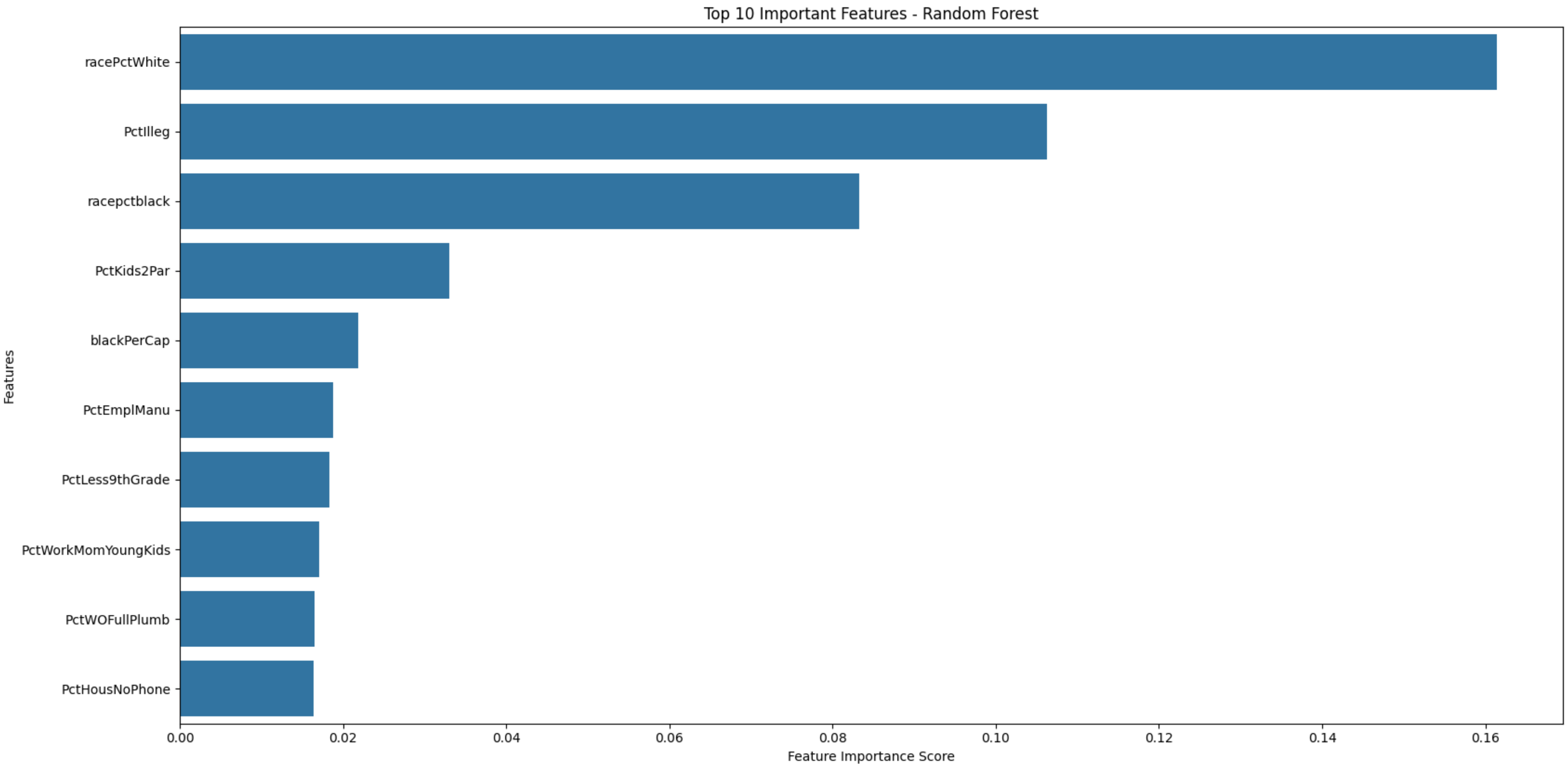
  - **Correlations Between Datasets:**

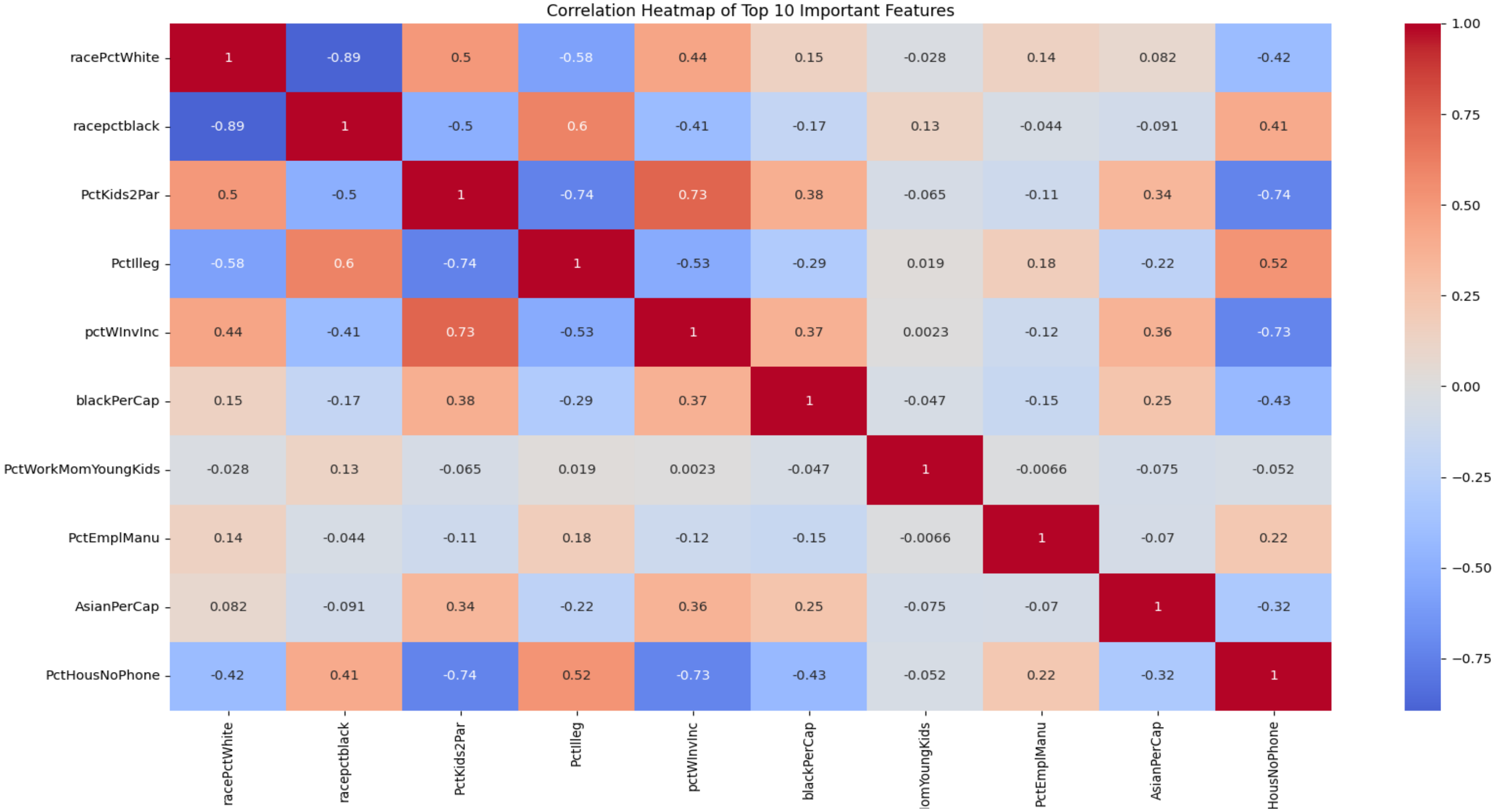
    What correlations can be drawn between biases in student failure rates and crime rates, and how do these correlations inform our understanding of socio-economic factors and systemic inequalities?

Socioeconomic Background & Relevancy Crime

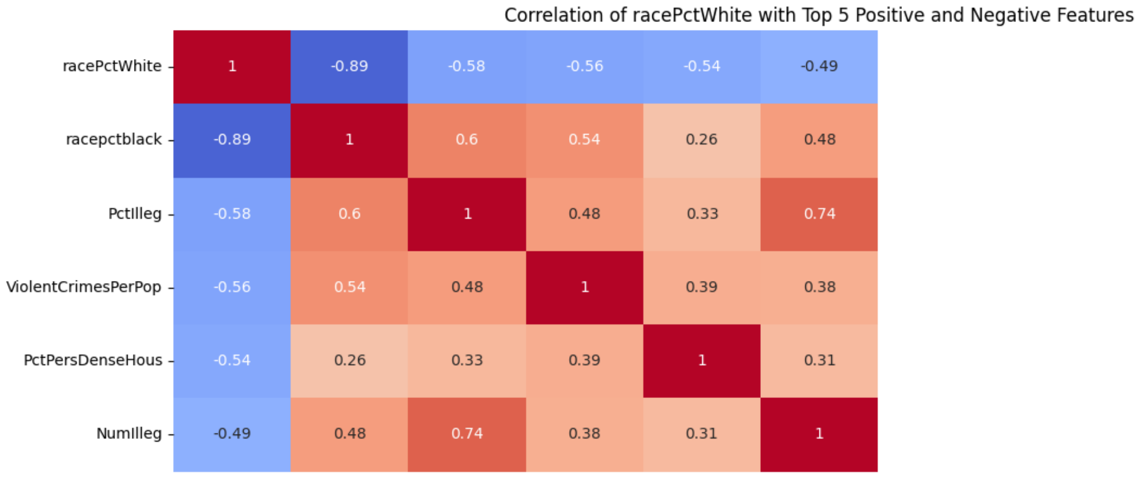




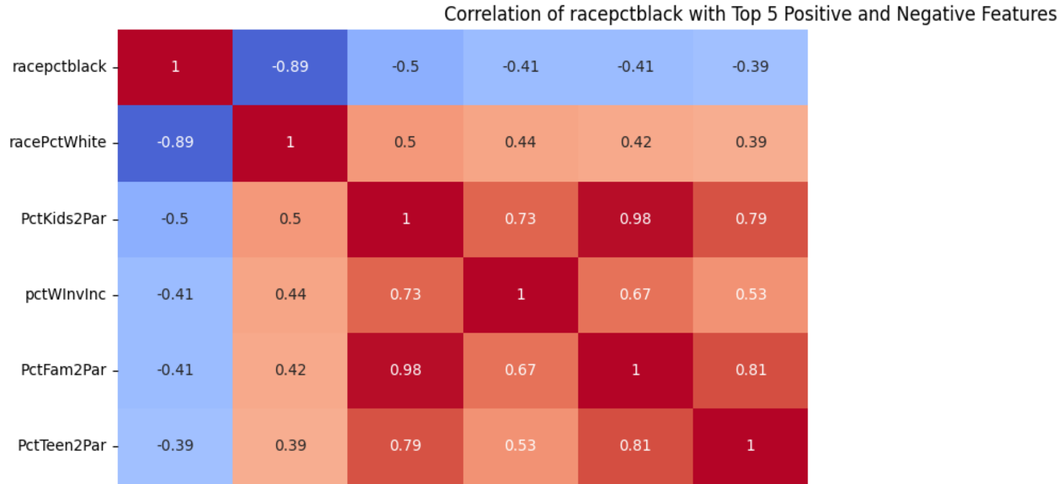

Top 10 Important Features - Random Forest
Features
racePctWhite
PctIlleg
racepctblack
PctKids2Par
blackPerCap
PctEmplManu
PctLess9thGrade
PctWorkMomYoungKids
PctWOFullPlumb
PctHousNoPhone
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Feature Importance Score



Correlation Heatmap of Top 10 Important Features

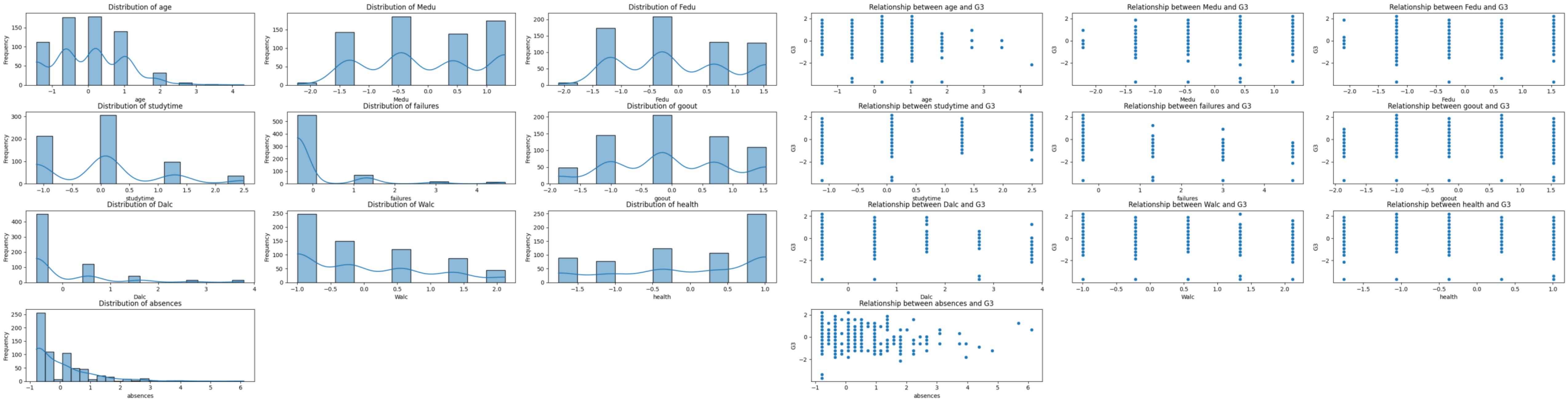


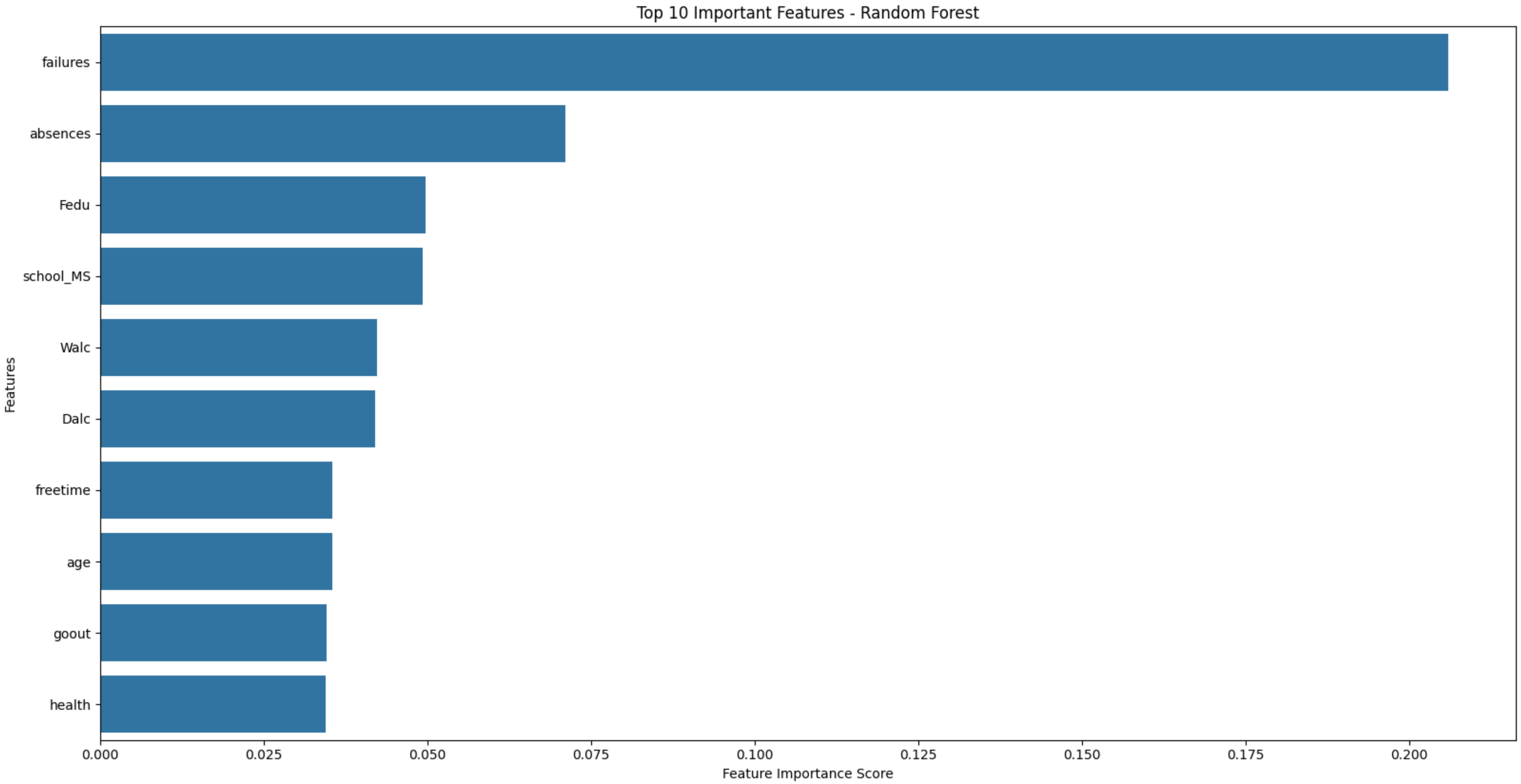
Correlation of racePctWhite with Top 5 Positive and Negative Features



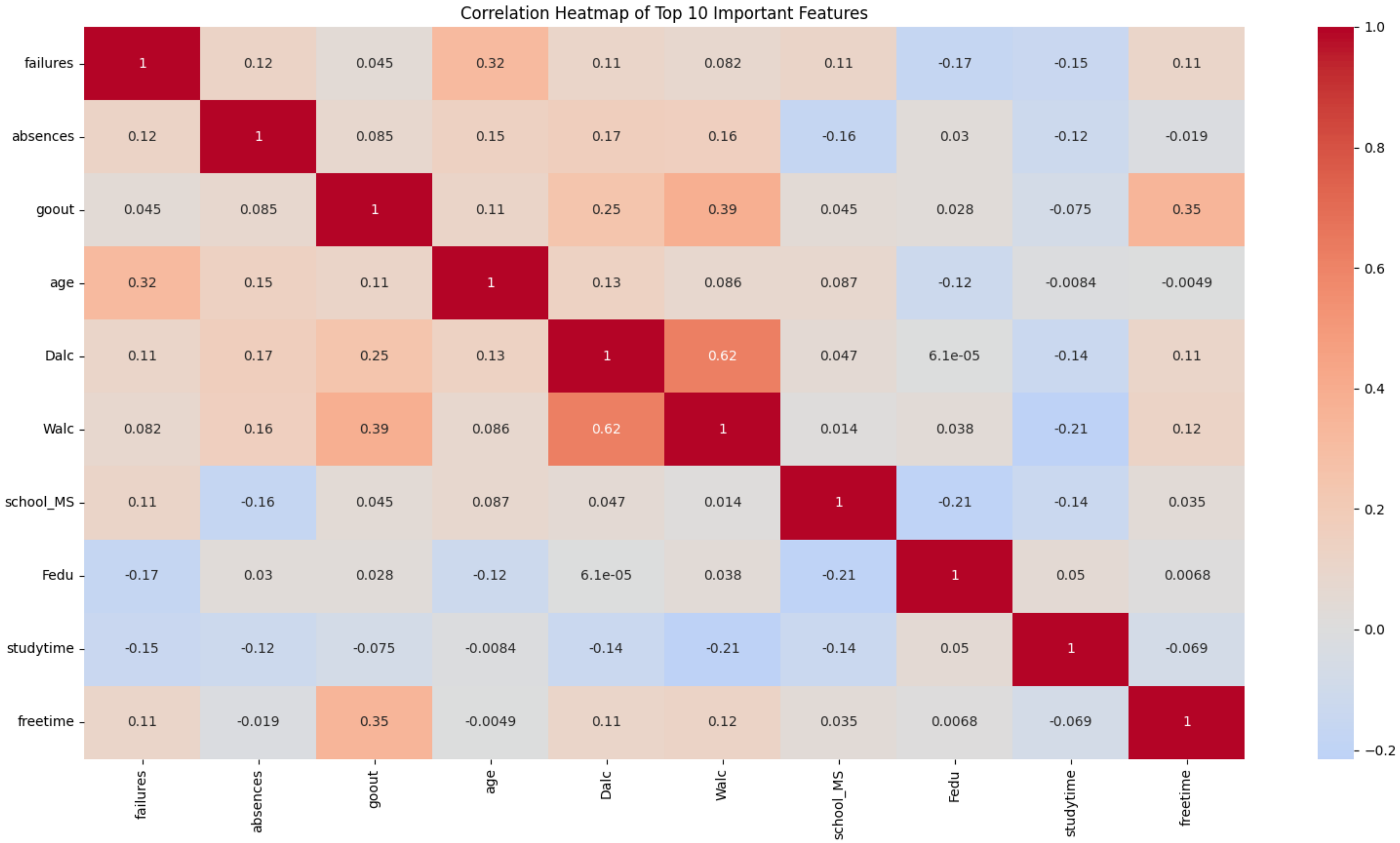
Correlation of racepctblack with Top 5 Positive and Negative Features

Socioeconomic Background & Relevancy Student

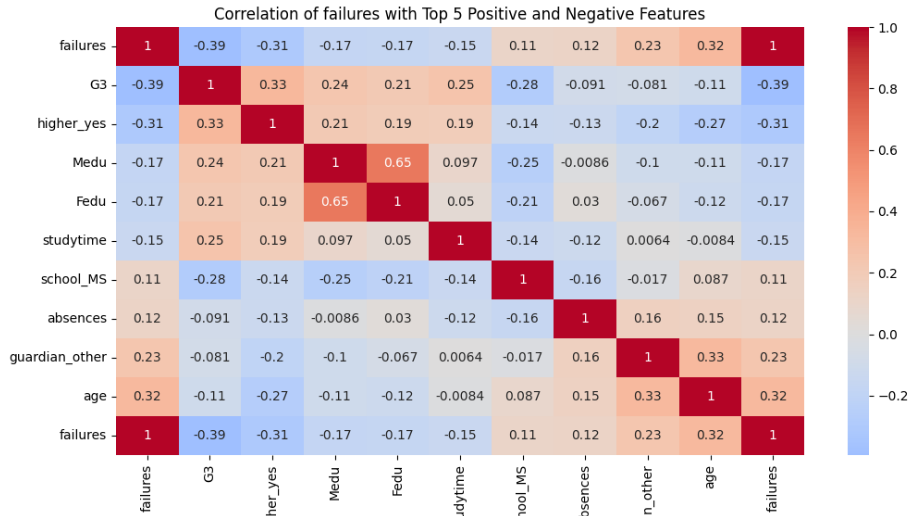



Top 10 Important Features - Random Forest
failures
absences
Fedu
school_MS
Walc
Dalc
freetime
age
goout
health
Features
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Feature Importance Score

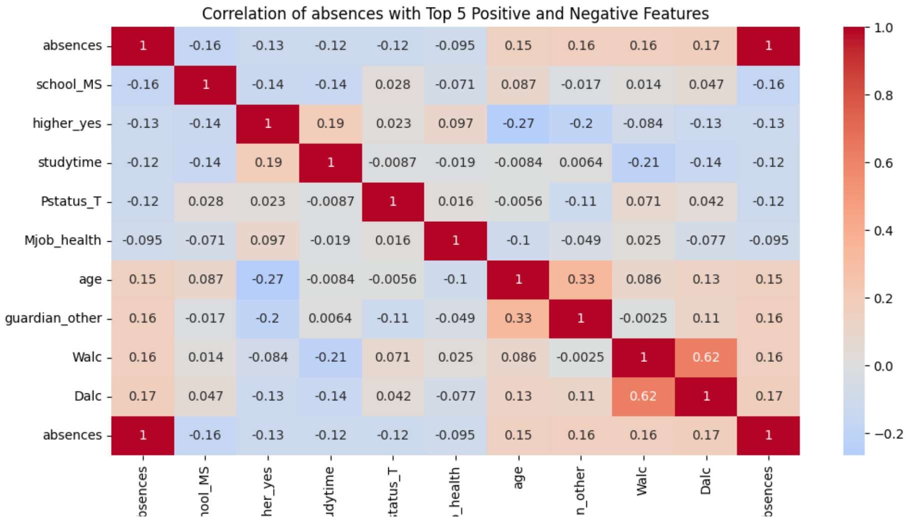
Correlation Heatmap of Top 10 Important Features



	failures	absences	goout	age	Dalc	Walc	school_MS	Fedu	studytime	freetime
failures	1	0.12	0.045	0.32	0.11	0.082	0.11	-0.17	-0.15	0.11
absences	0.12	1	0.085	0.15	0.17	0.16	-0.16	0.03	-0.12	-0.019
goout	0.045	0.085	1	0.11	0.25	0.39	0.045	0.028	-0.075	0.35
age	0.32	0.15	0.11	1	0.13	0.086	0.087	-0.12	-0.0084	-0.0049
Dalc	0.11	0.17	0.25	0.13	1	0.62	0.047	6.1e-05	-0.14	0.11
Walc	0.082	0.16	0.39	0.086	0.62	1	0.014	0.038	-0.21	0.12
school_MS	0.11	-0.16	0.045	0.087	0.047	0.014	1	-0.21	-0.14	0.035
Fedu	-0.17	0.03	0.028	-0.12	6.1e-05	0.038	-0.21	1	0.05	0.0068
studytime	-0.15	-0.12	-0.075	-0.0084	-0.14	-0.21	-0.14	0.05	1	-0.069
freetime	0.11	-0.019	0.35	-0.0049	0.11	0.12	0.035	0.0068	-0.069	1



Correlation of failures with Top 5 Positive and Negative Features



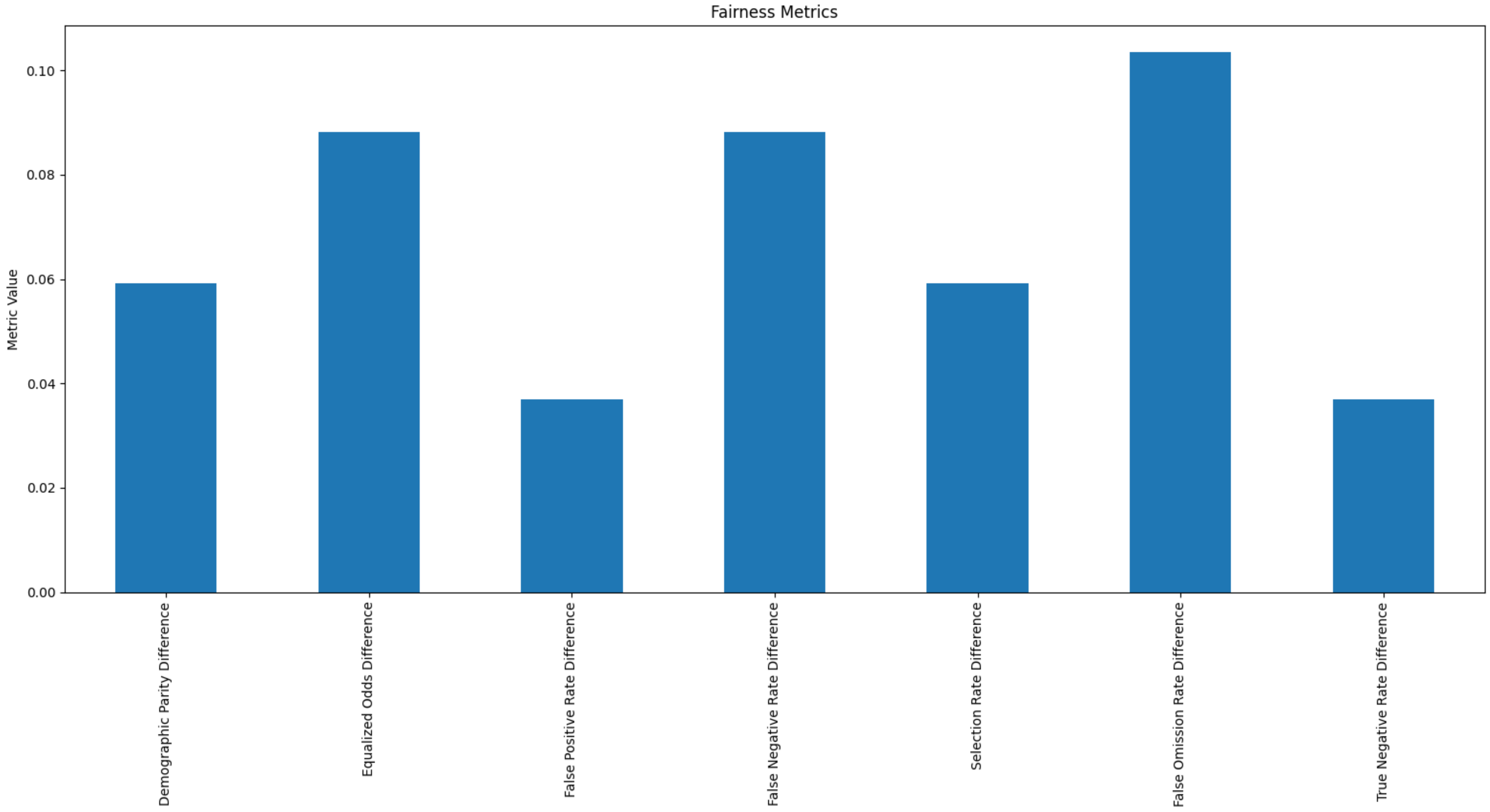
Correlation of absences with Top 5 Positive and Negative Features

Research Question 1

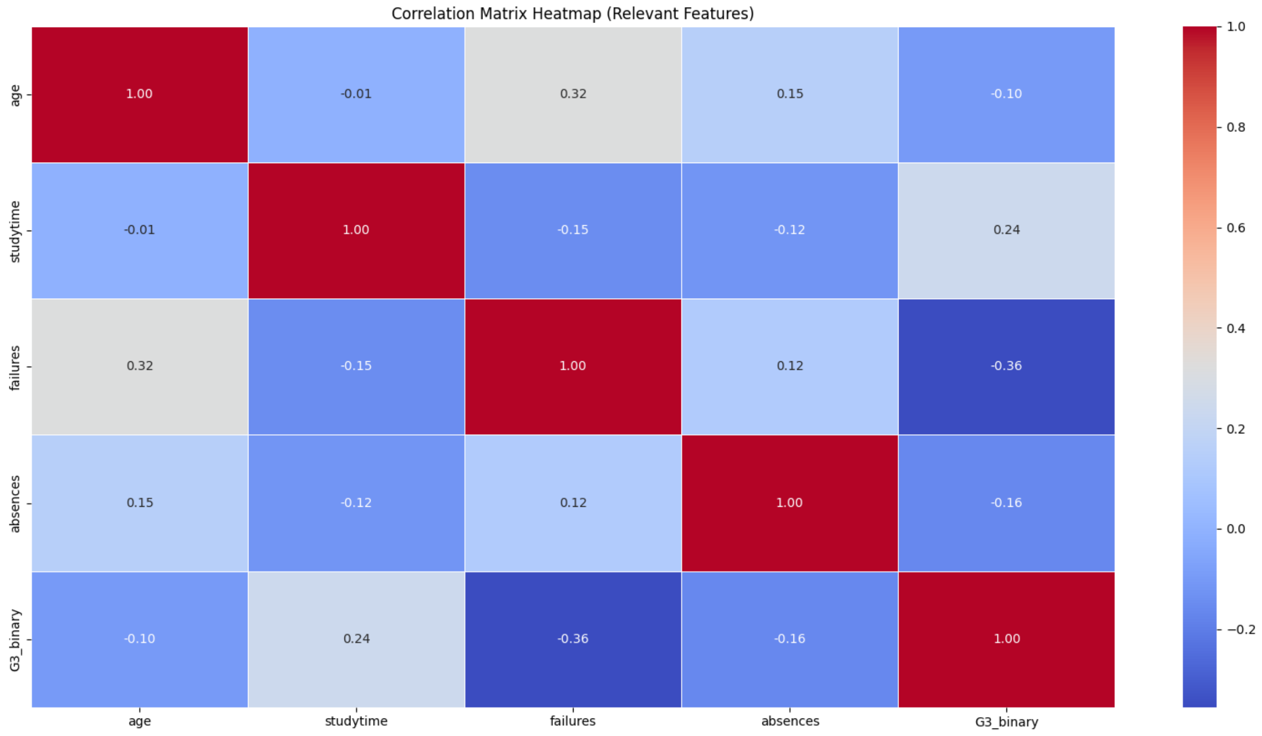
- Significant potential biases detected in crime and student performance datasets are influenced by socio-economic factors such as race, parental education, and income levels etc.
- Fairlearn and AIF360 perfect for identifying and mitigating these biases
- Crucial to ensure unbiased decision-making and supporting equitable policy interventions.

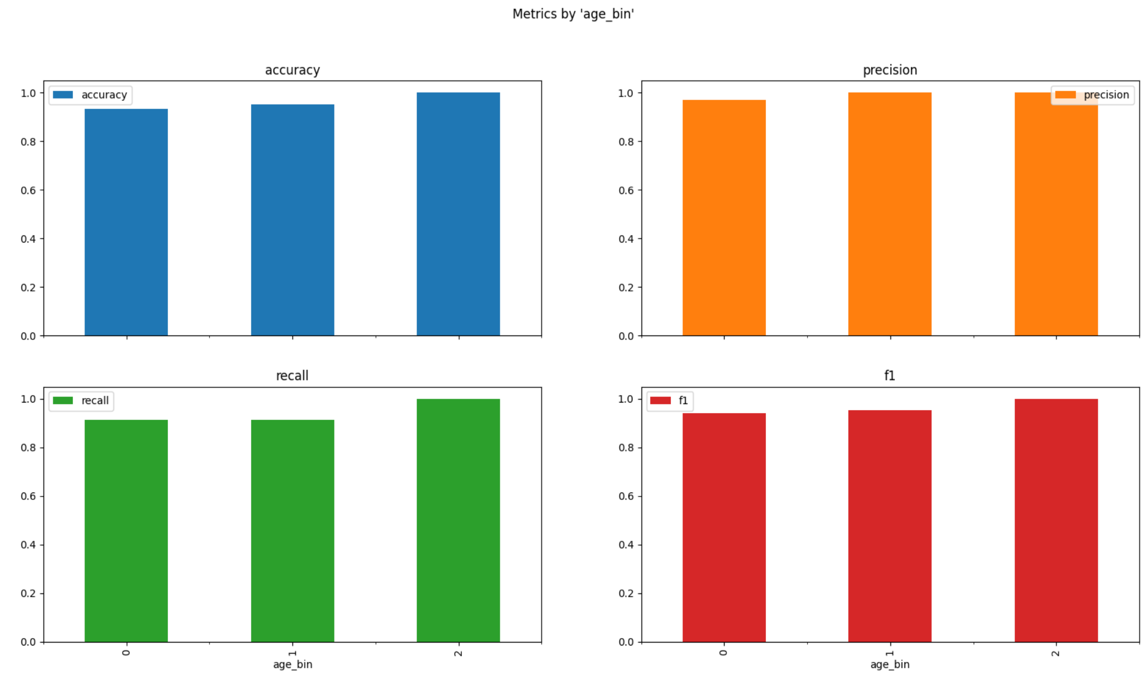
Bias Detection

Bias detection - Student

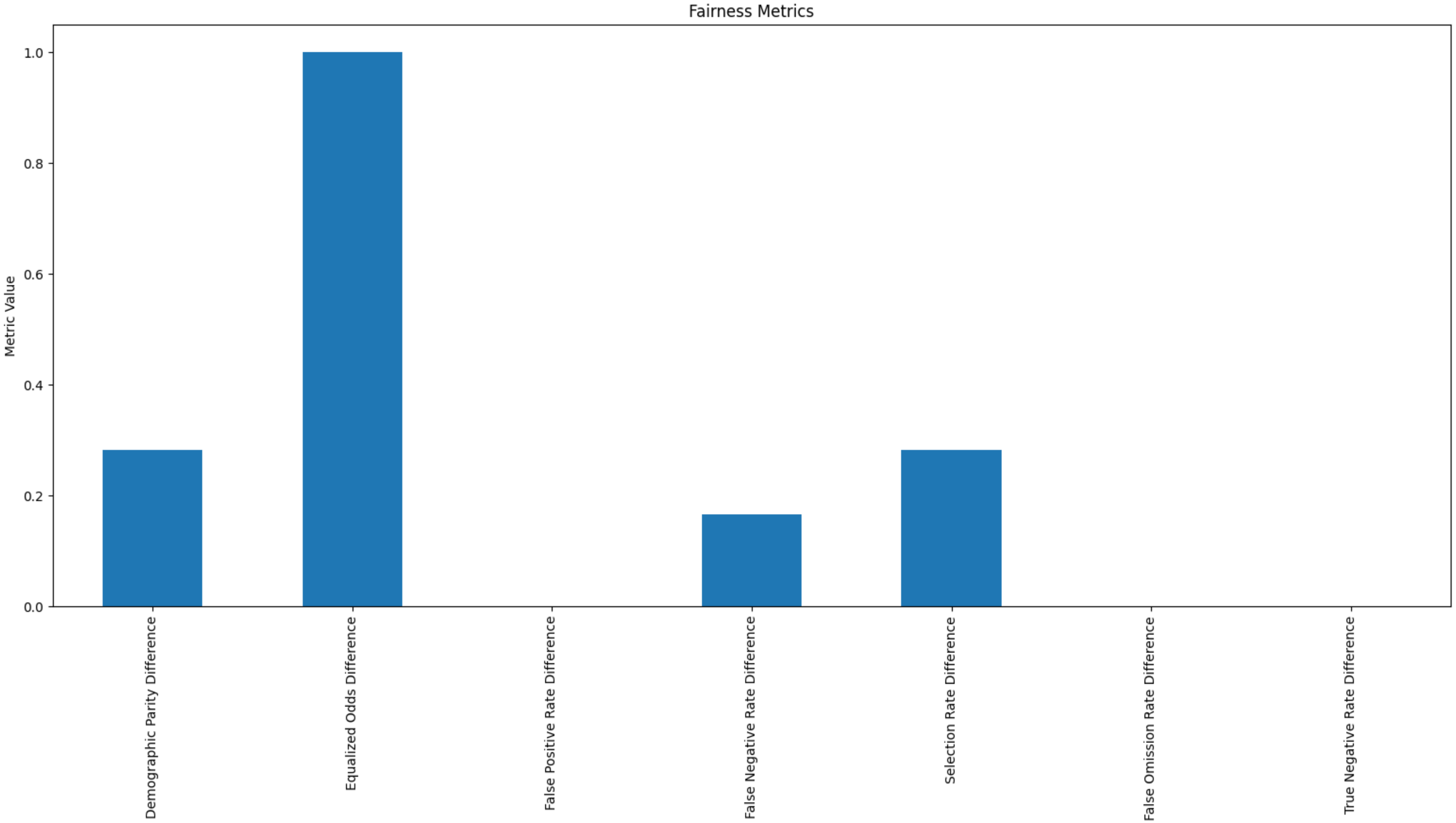


Bias detection - Student

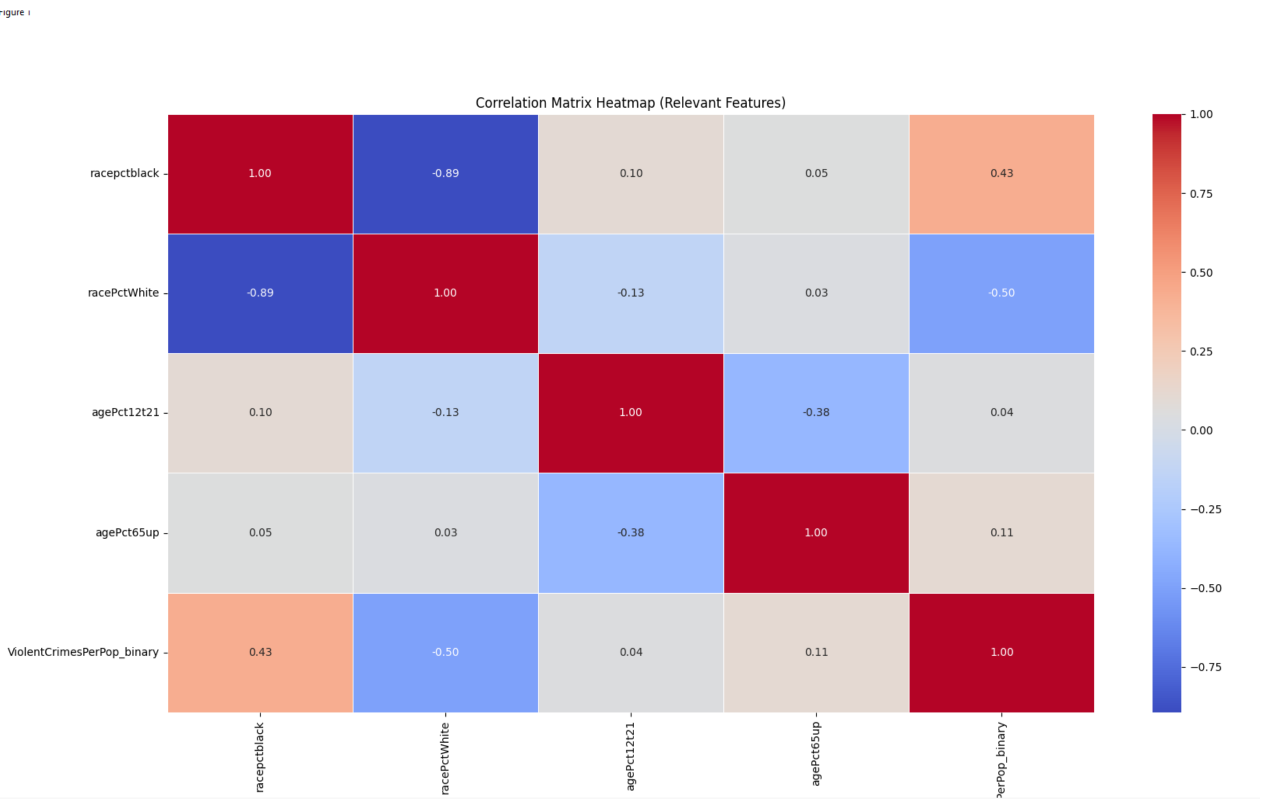


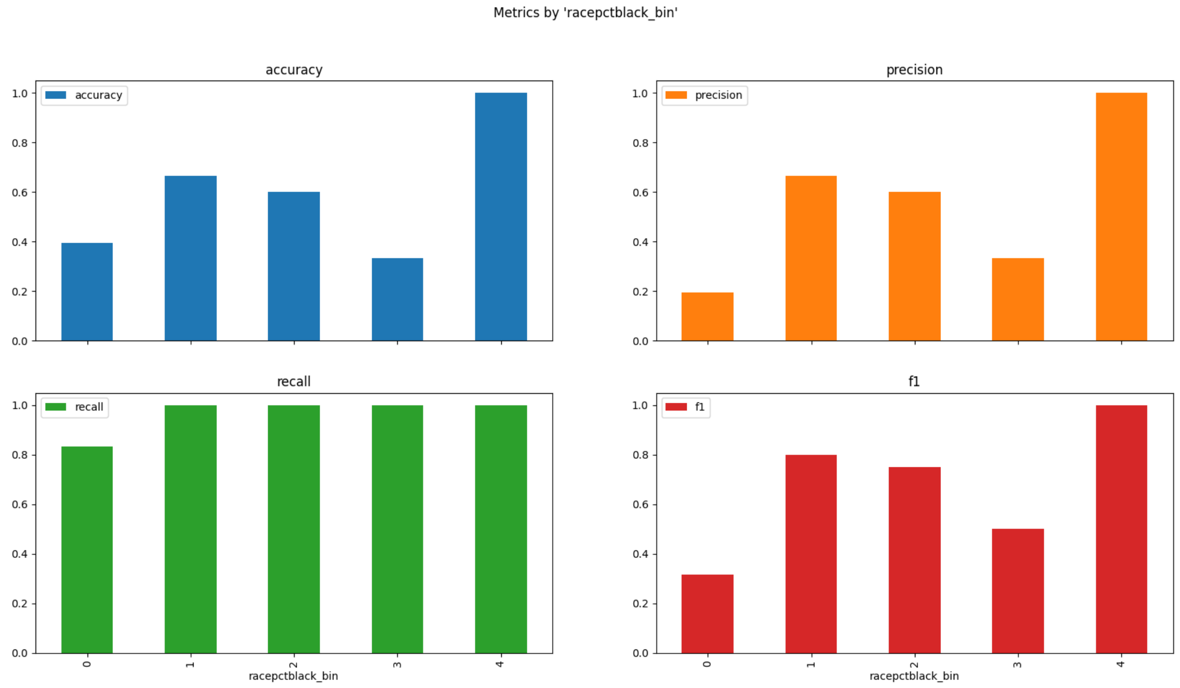


Bias detection - Crime



Bias detection - Crime

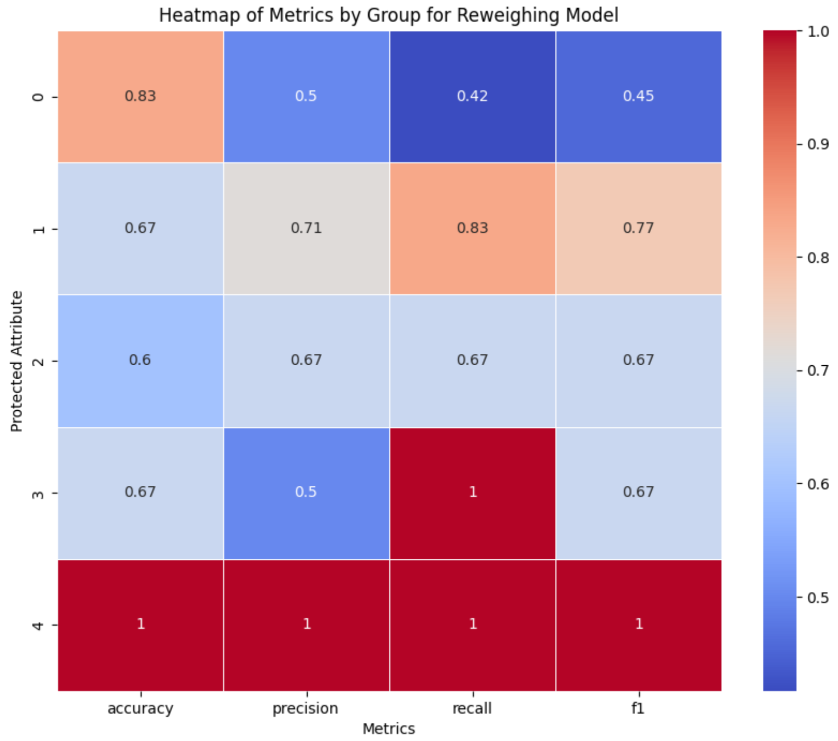




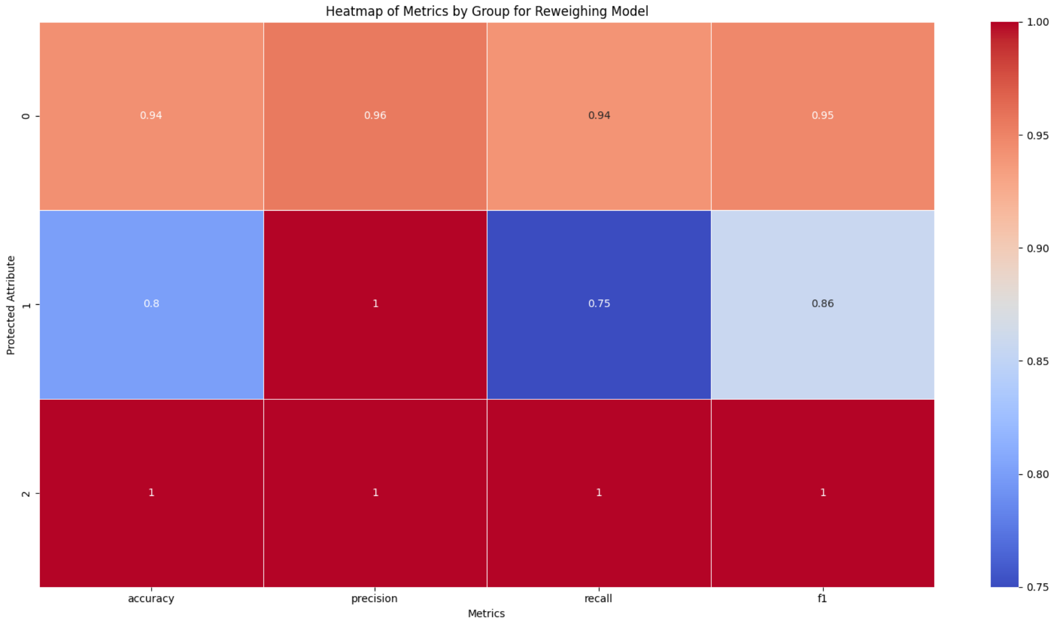
Bias Mitigation

Reweighing Fairlearn

Crime

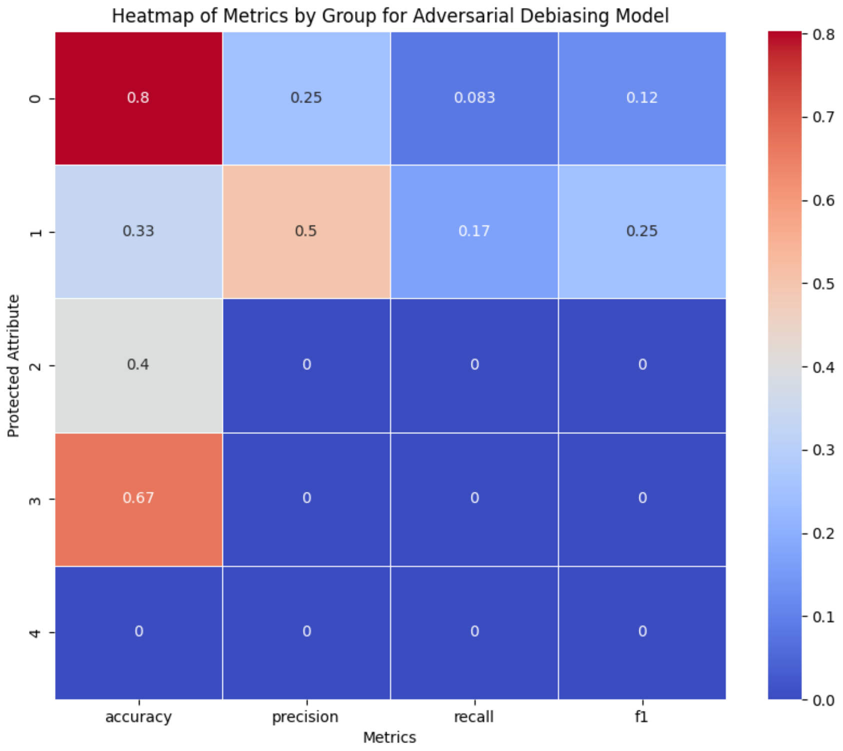


Student

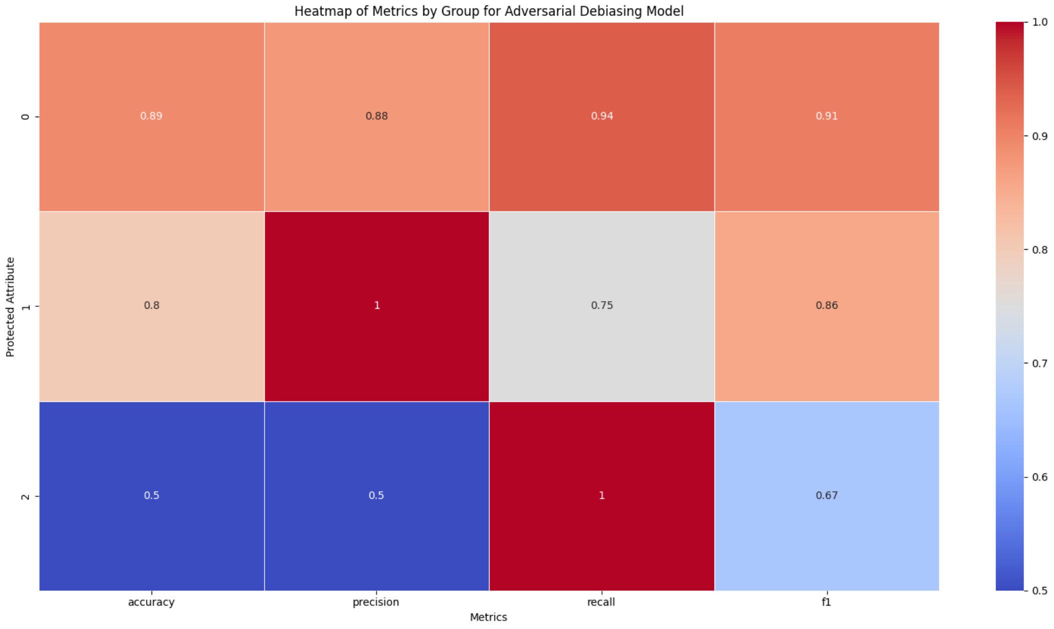


Adversarial Debiasing - Fairlearn

Crime

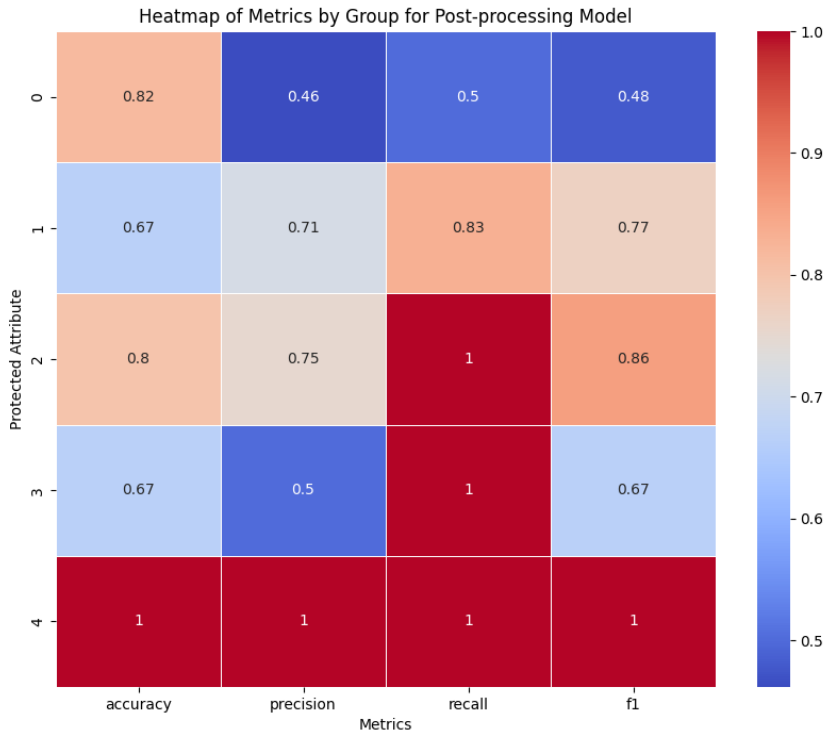


Student

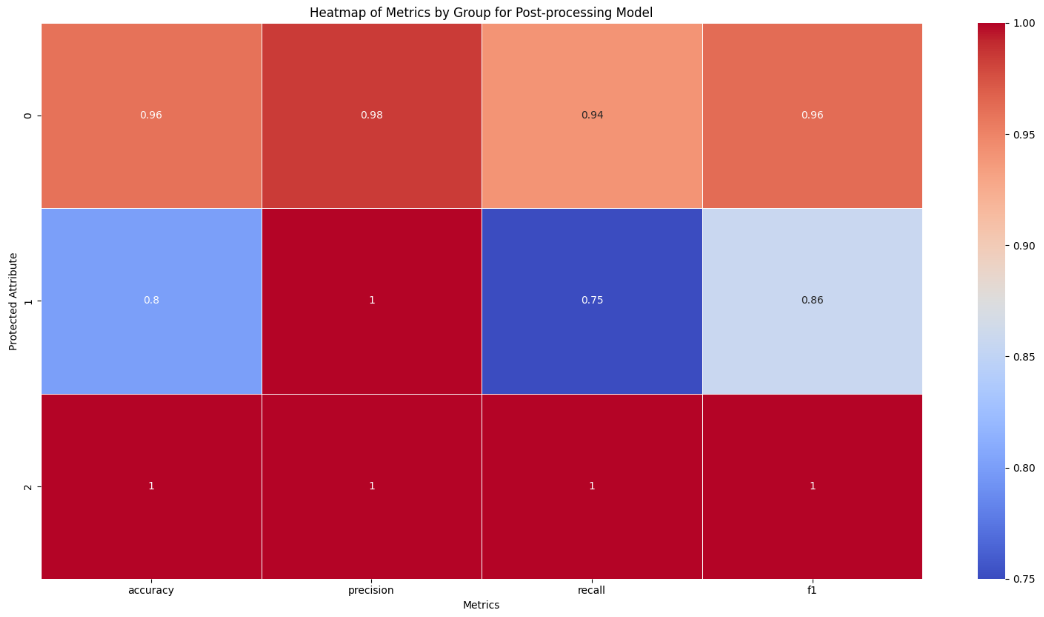


Post-Processing - Fairlearn

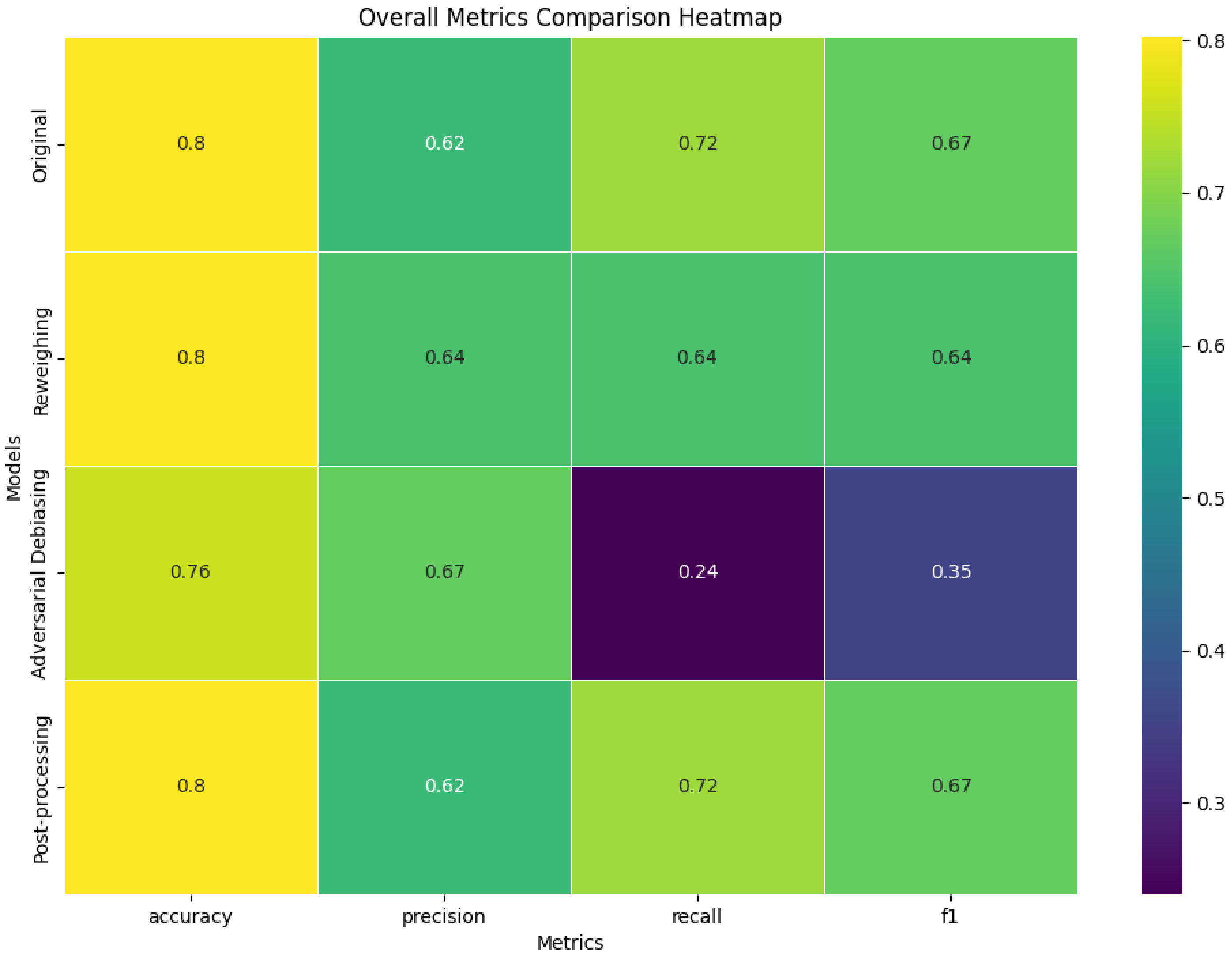
Crime



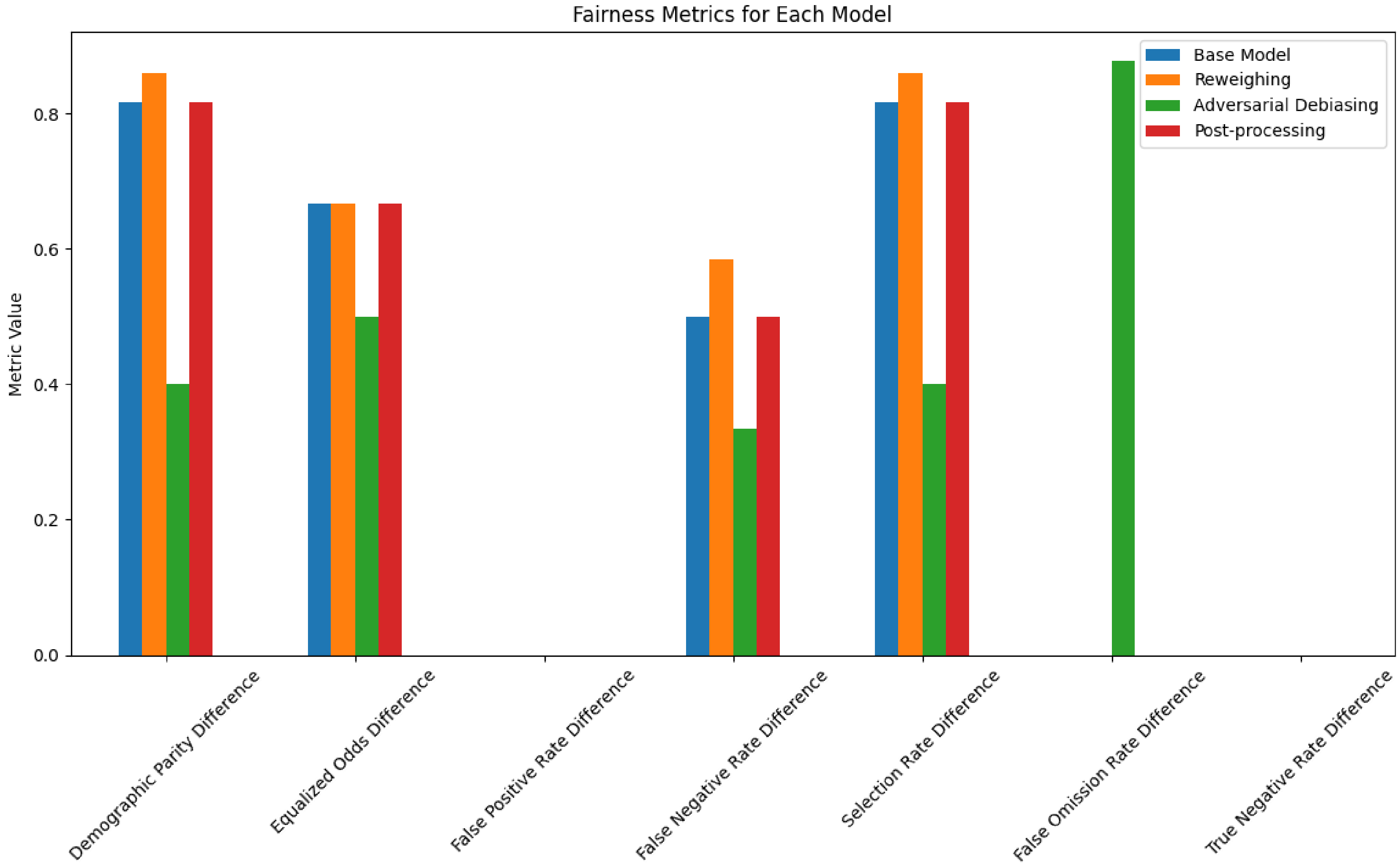
Student



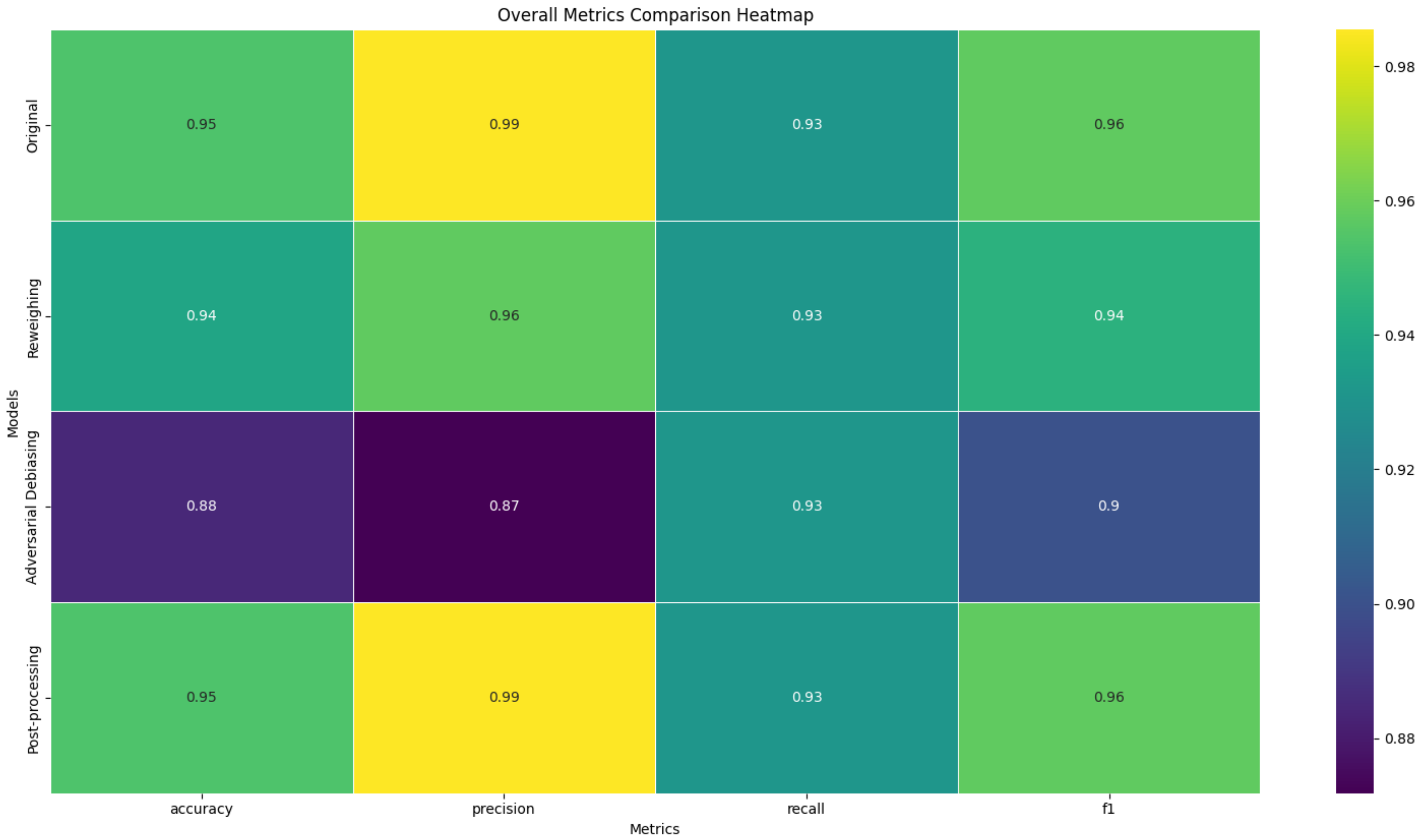
Fairlearn Combined Results - Crime



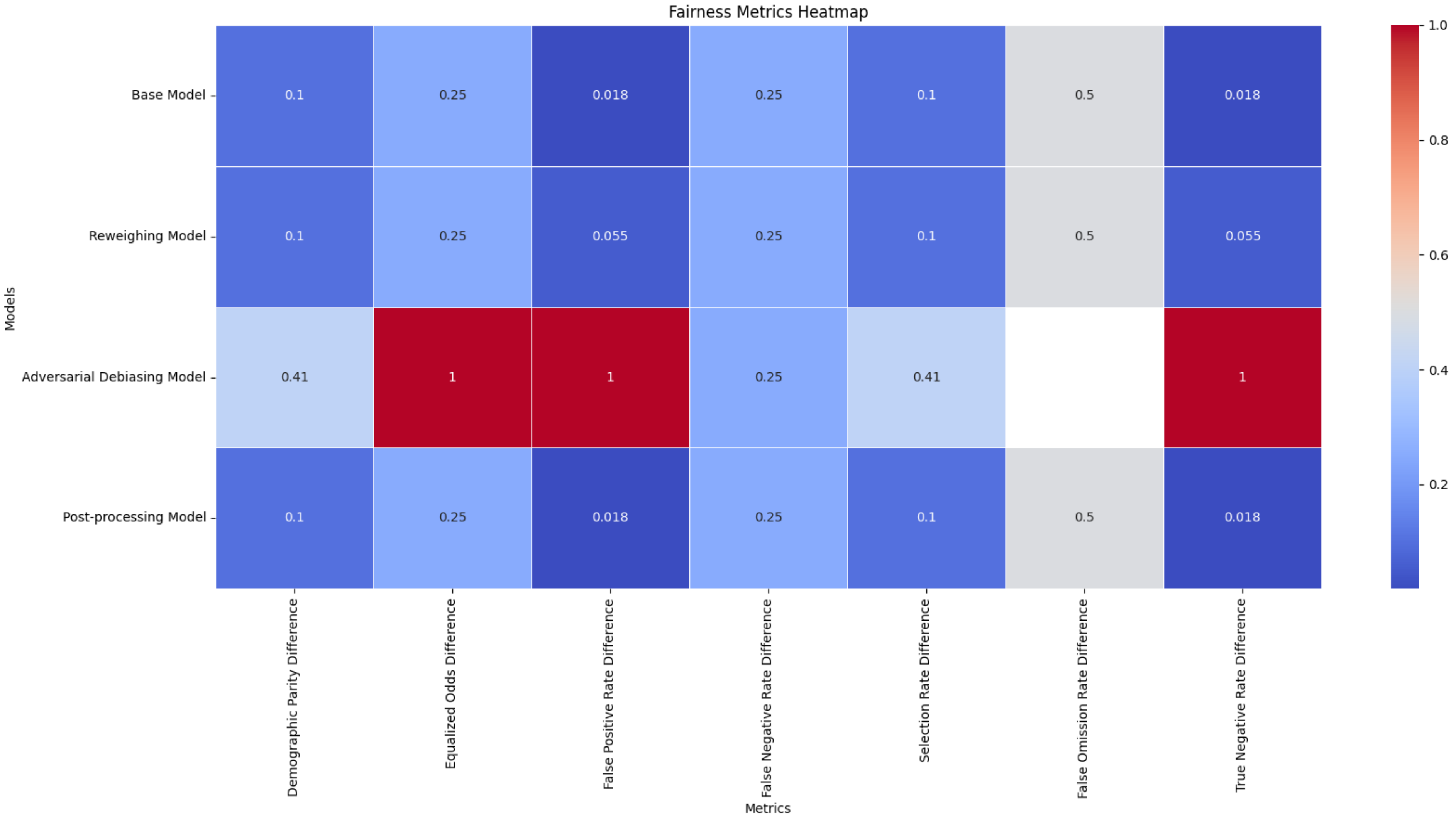
Fairlearn Combined Results - Crime



Fairlearn Combined Results - Student

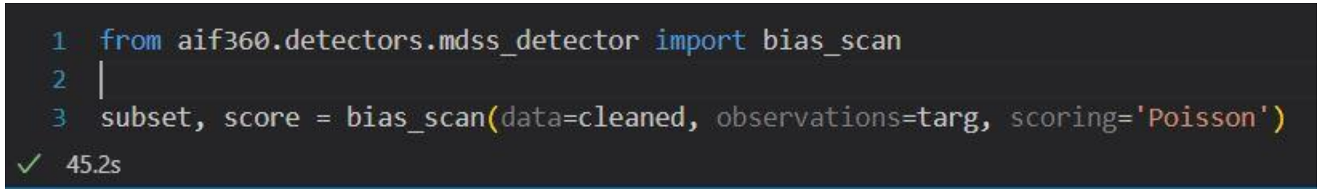


Fairlearn Combined Results - Student



	Demographic Parity Difference	Equalized Odds Difference	False Positive Rate Difference	False Negative Rate Difference	Selection Rate Difference	False Omission Rate Difference	True Negative Rate Difference
Base Model	0.1	0.25	0.018	0.25	0.1	0.5	0.018
Reweighing Model	0.1	0.25	0.055	0.25	0.1	0.5	0.055
Adversarial Debiasing Model	0.41	1	1	0.25	0.41		1
Post-processing Model	0.1	0.25	0.018	0.25	0.1	0.5	0.018

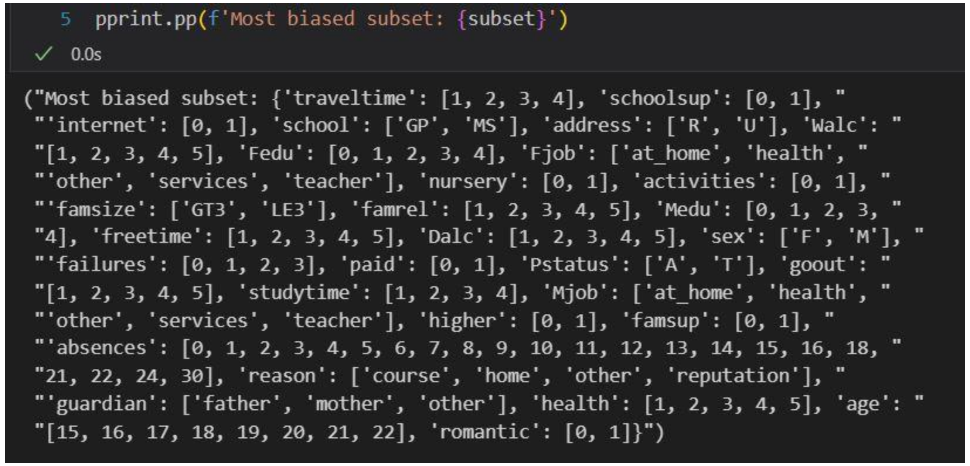
Bias Detection – Aif360



```
from aif360.detectors.mdss_detector import bias_scan

subset, score = bias_scan(data=cleaned, observations=targ, scoring='Poisson')
```

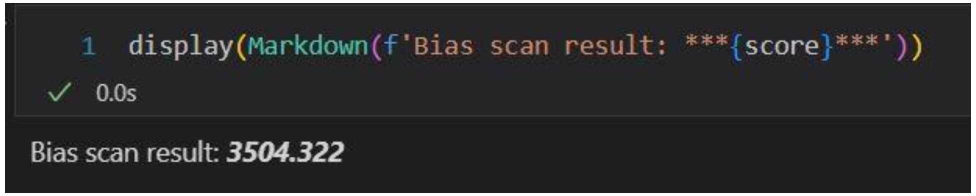
✓ 45.2s



```
pprint.pp(f'Most biased subset: {subset}')
```

✓ 0.0s

```
("Most biased subset: {'traveltime': [1, 2, 3, 4], 'schoolsup': [0, 1], "
 "'internet': [0, 1], 'school': ['GP', 'MS'], 'address': ['R', 'U'], 'Walc': "
 "[1, 2, 3, 4, 5], 'Fedu': [0, 1, 2, 3, 4], 'Fjob': ['at_home', 'health', "
 "'other', 'services', 'teacher'], 'nursery': [0, 1], 'activities': [0, 1], "
 "'famsize': ['GT3', 'LE3'], 'famrel': [1, 2, 3, 4, 5], 'Medu': [0, 1, 2, 3, "
 "4], 'freetime': [1, 2, 3, 4, 5], 'Dalc': [1, 2, 3, 4, 5], 'sex': ['F', 'M'], "
 "'failures': [0, 1, 2, 3], 'paid': [0, 1], 'Pstatus': ['A', 'T'], 'goout': "
 "[1, 2, 3, 4, 5], 'studytime': [1, 2, 3, 4], 'Mjob': ['at_home', 'health', "
 "'other', 'services', 'teacher'], 'higher': [0, 1], 'famsup': [0, 1], "
 "'absences': [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, "
 "21, 22, 24, 30], 'reason': ['course', 'home', 'other', 'reputation'], "
 "'guardian': ['father', 'mother', 'other'], 'health': [1, 2, 3, 4, 5], 'age': "
 "[15, 16, 17, 18, 19, 20, 21, 22], 'romantic': [0, 1]}")
```



```
display(Markdown(f'Bias scan result: ***{score}***'))
```

✓ 0.0s

Bias scan result: ***3504.322***

Aif360 Fairness Metrics

Smooth empirical differences in fairness = Probability of favourable and unfavourable outcomes between intersecting group

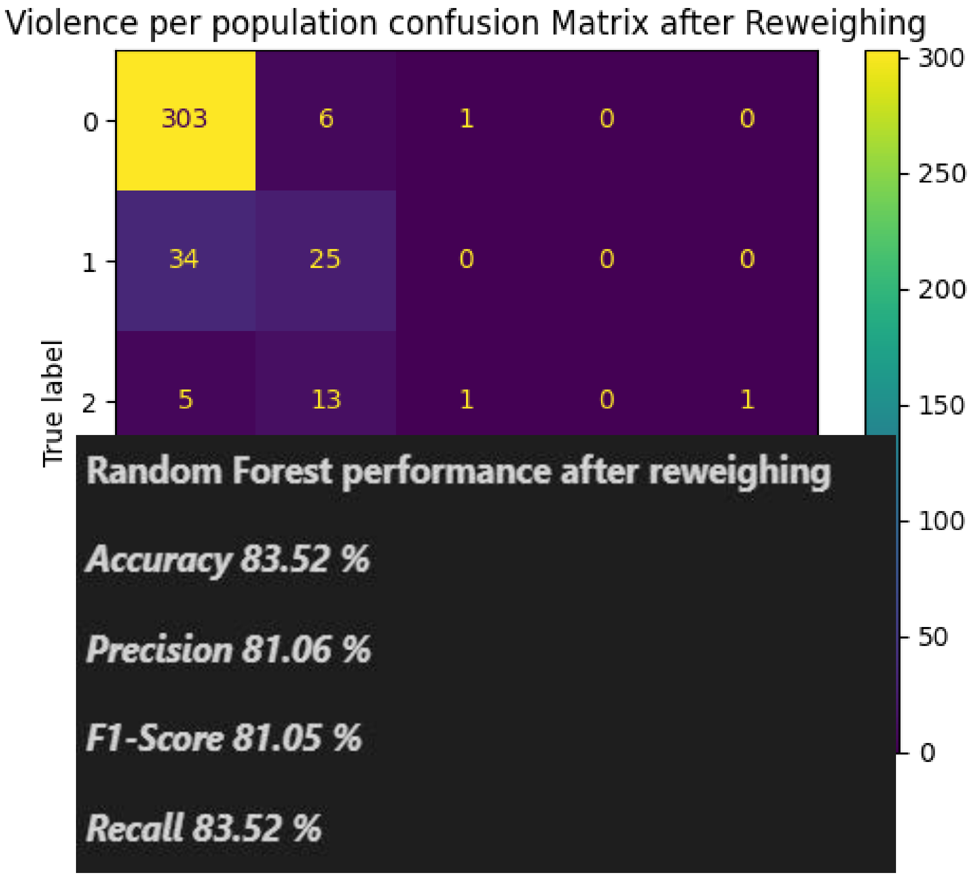
Disparate Impact = Rate of positive outcomes of unprivileged groups / Rate of positive outcomes of privileged groups

Statistical parity difference = Difference in the ratio of favourable outcomes between monitored groups and reference groups

1. Z. Zhang, D. Wang, B. Yang and J. Yin, "Weighted Multidimensional Scaling Localization Method With Bias Reduction Based on TOA," in IEEE Sensors Journal, vol. 23, no. 17, pp. 19803-19814, 1 Sept.1, 2023, doi: 10.1109/JSEN.2023.3296986. keywords: {Sensors;Location awareness;Estimation;Mathematical models;Noise measurement;Weight measurement;Time measurement;Bias reduction;Cramér–Rao lower bound (CRLB);multidimensional scaling (MDS);sensor-based localization;time-of-arrival (TOA) measurement}
2. IBM (2024). SED. Available at: https://dataplatform.cloud.ibm.com/docs/content/wsj/model/wos-smooth-empirical-diff.html?context=cpdaas
3. Truong, B. (2021). Mitigating Bias in AI with AIF360. [online] Medium. Available at: https://towardsdatascience.com/mitigating-bias-in-ai-with-aif360-b4305d1f88a9.
4. dataplatform.cloud.ibm.com. (n.d.). Statistical parity difference | IBM Cloud Pak for Data as a Service. [online] Available at: https://dataplatform.cloud.ibm.com/docs/content/wsj/model/wos-stat-parity-diff.html?context=cpdaas

Pre-Processing: Reweighing Aif360

Crime



Violence per population confusion Matrix after Reweighing

0	303	6	1	0	0
1	34	25	0	0	0
2	5	13	1	0	1

Random Forest performance after reweighing

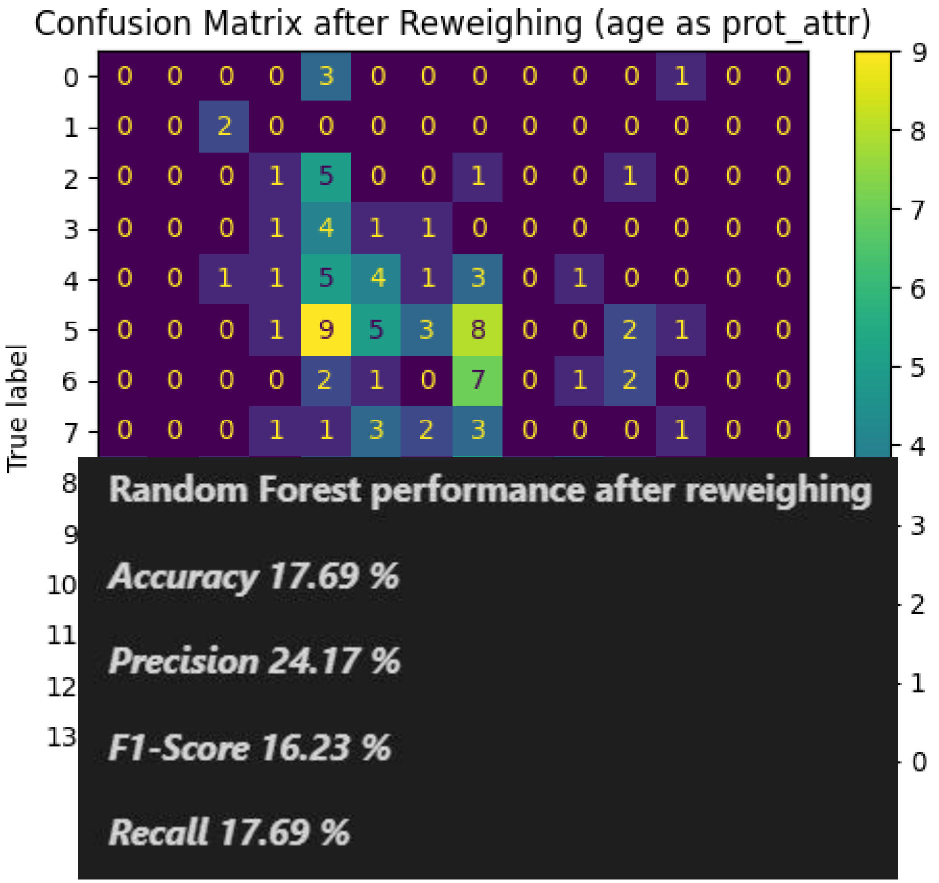
Accuracy 83.52 %

Precision 81.06 %

F1-Score 81.05 %

Recall 83.52 %

Student



Confusion Matrix after Reweighing (age as prot_attr)

0	0	0	0	0	3	0	0	0	0	0	0	1	0	0
1	0	0	2	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	5	0	0	1	0	0	1	0	0	0
3	0	0	0	1	4	1	1	0	0	0	0	0	0	0
4	0	0	1	1	5	4	1	3	0	1	0	0	0	0
5	0	0	0	1	9	5	3	8	0	0	2	1	0	0
6	0	0	0	0	2	1	0	7	0	1	2	0	0	0
7	0	0	0	1	1	3	2	3	0	0	0	1	0	0

Random Forest performance after reweighing

Accuracy 17.69 %

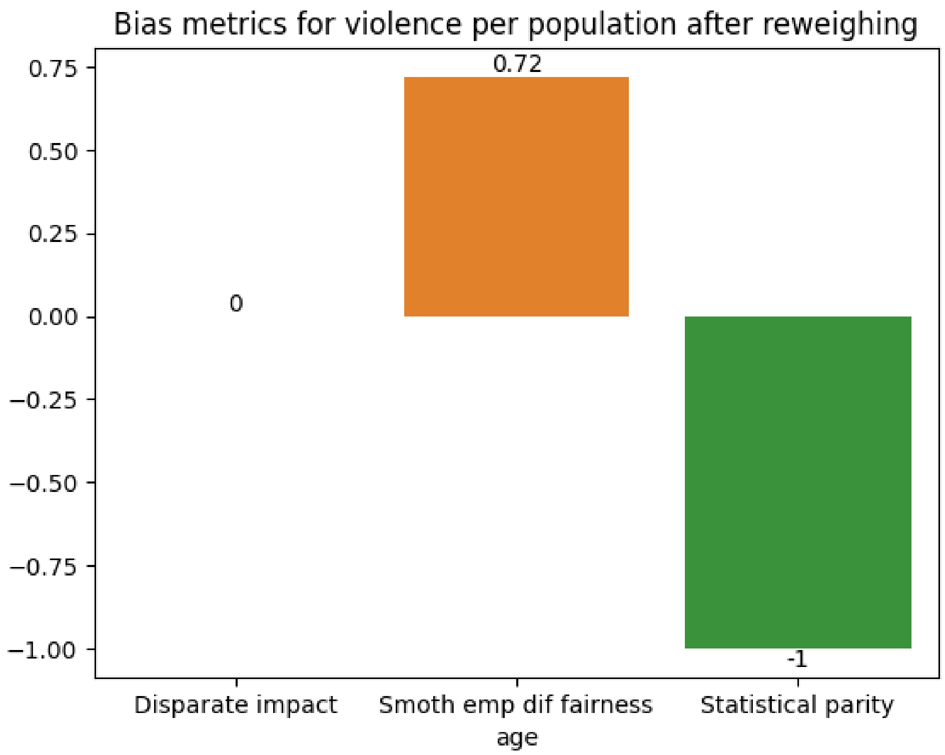
Precision 24.17 %

F1-Score 16.23 %

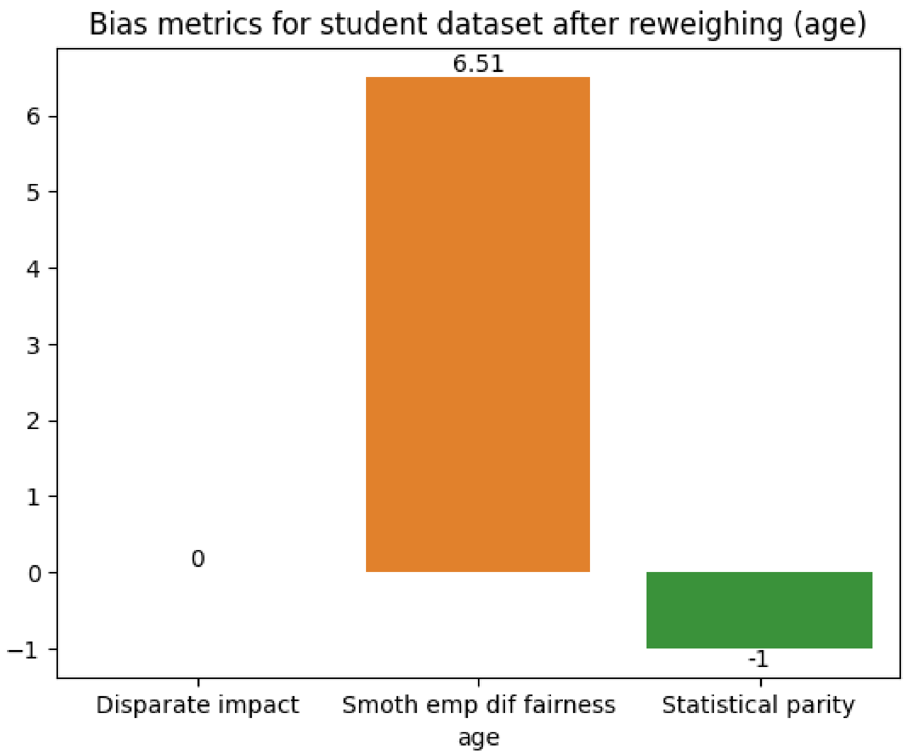
Recall 17.69 %

Reweighing Aif360

Crime

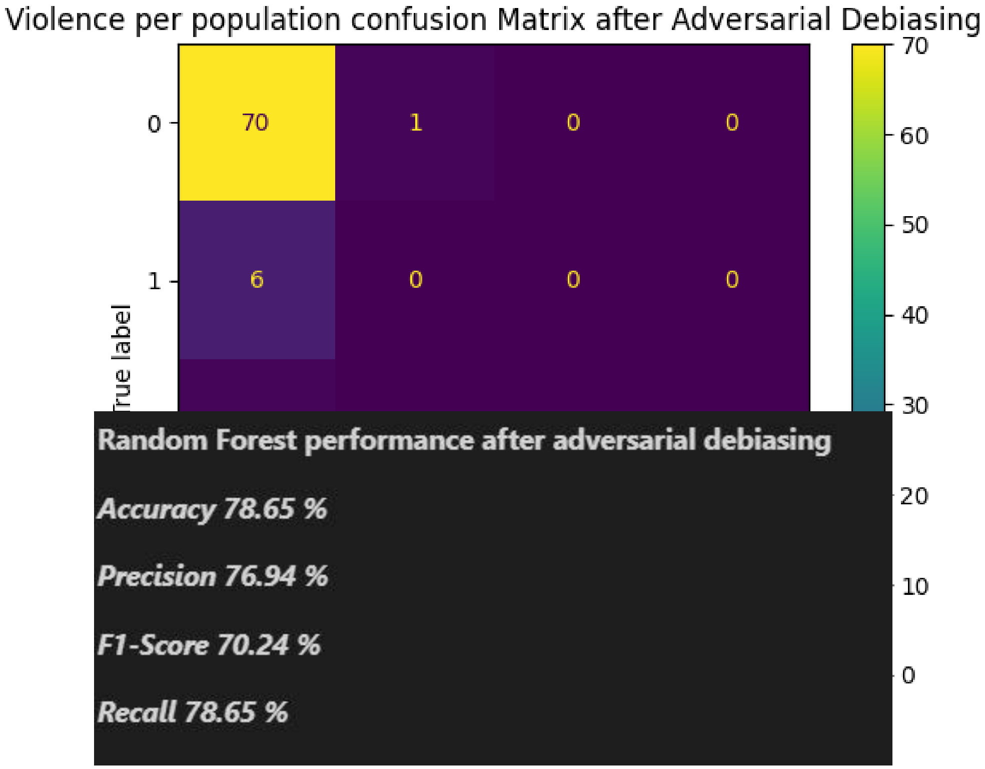


Student



Adversarial Debiasing - Aif360

Crime



Random Forest performance after adversarial debiasing

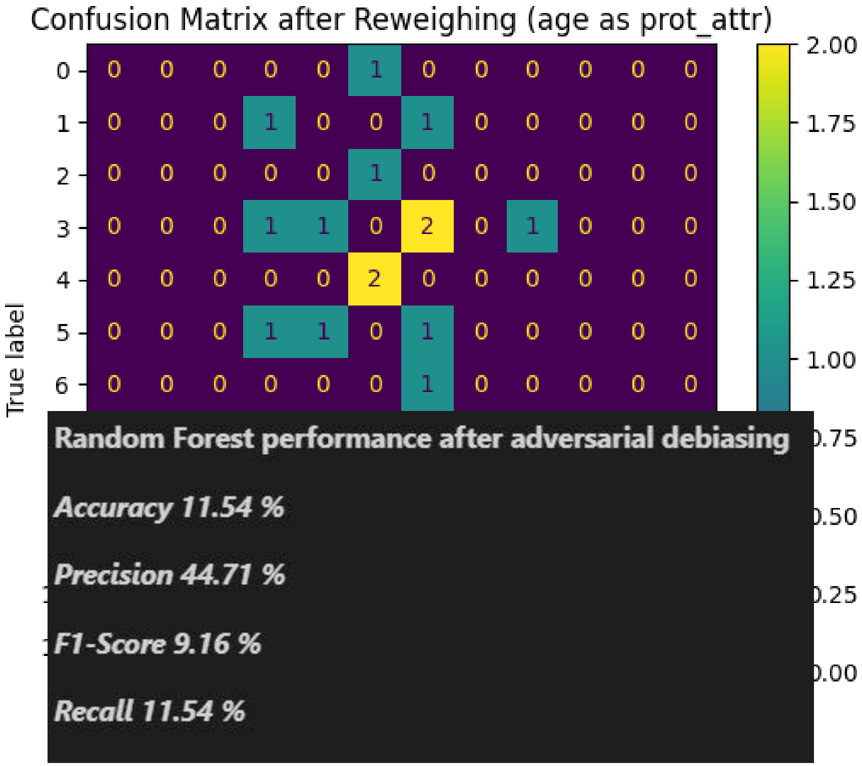
Accuracy 78.65 %

Precision 76.94 %

F1-Score 70.24 %

Recall 78.65 %

Student



Random Forest performance after adversarial debiasing

Accuracy 11.54 %

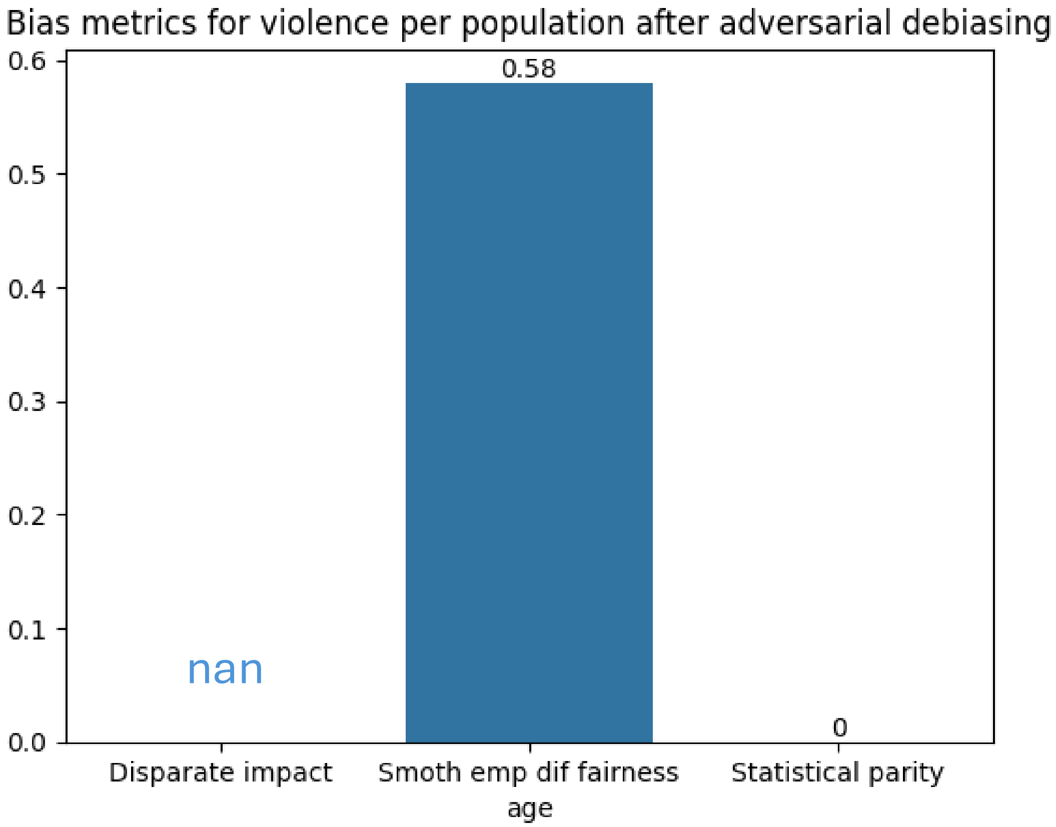
Precision 44.71 %

F1-Score 9.16 %

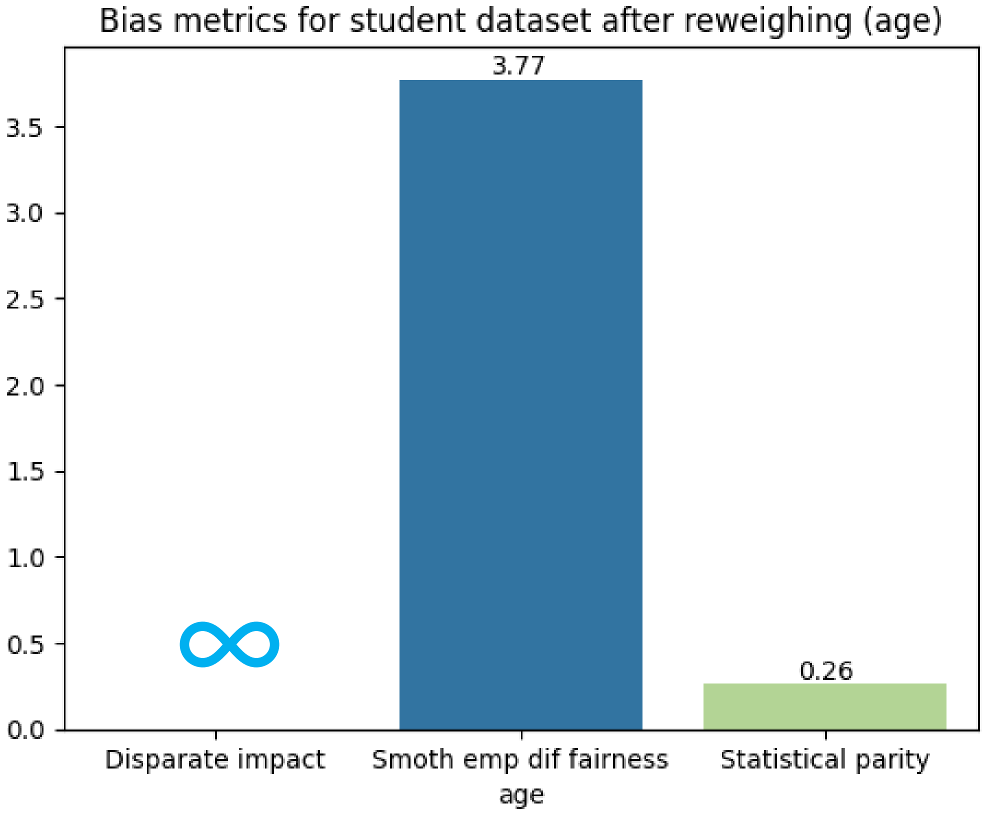
Recall 11.54 %

Adversarial Debiasing - Aif360

Crime

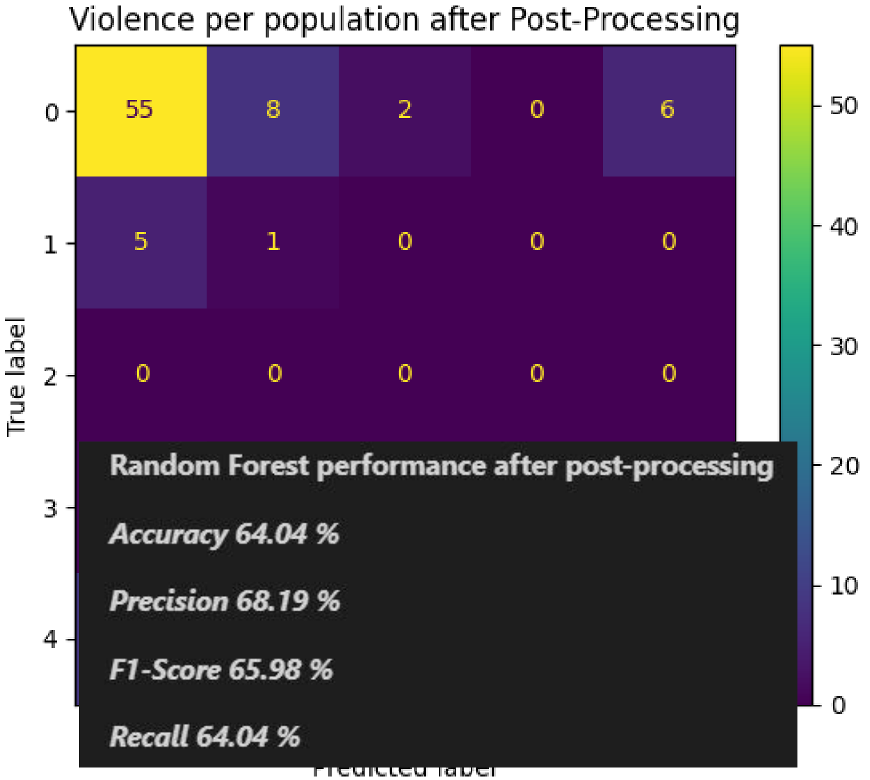


Student



Post-Processing Aif360

Crime



0	55	8	2	0	6
1	5	1	0	0	0
2	0	0	0	0	0

Random Forest performance after post-processing

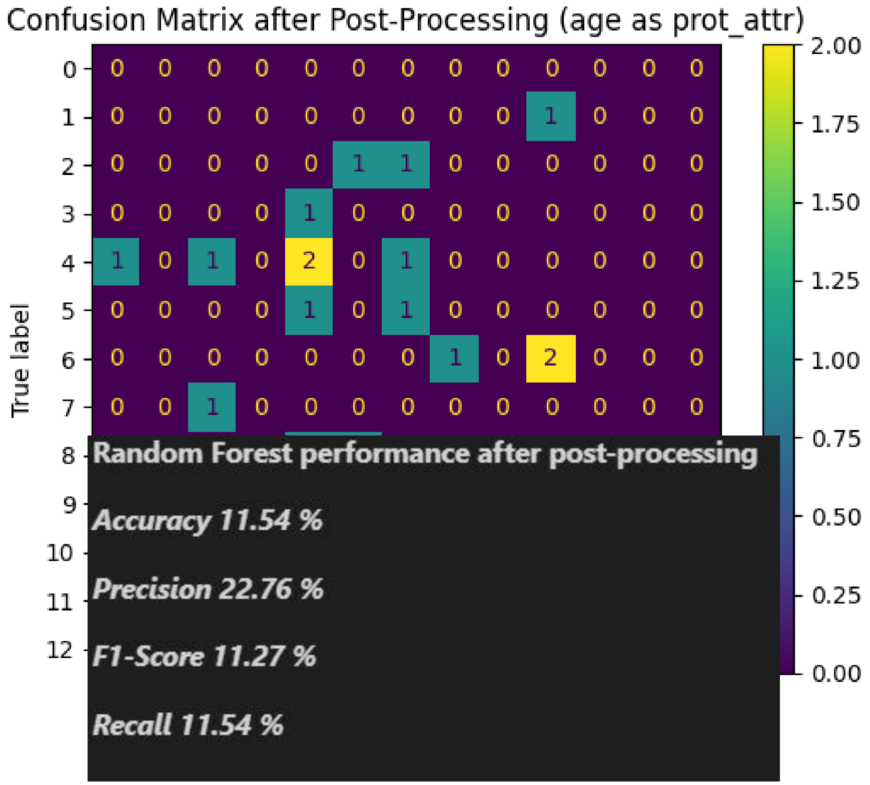
Accuracy 64.04 %

Precision 68.19 %

F1-Score 65.98 %

Recall 64.04 %

Student



0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	1	0	0	0
2	0	0	0	0	0	1	1	0	0	0	0	0	0
3	0	0	0	0	1	0	0	0	0	0	0	0	0
4	1	0	1	0	2	0	1	0	0	0	0	0	0
5	0	0	0	0	1	0	1	0	0	0	0	0	0
6	0	0	0	0	0	0	0	1	0	2	0	0	0
7	0	0	1	0	0	0	0	0	0	0	0	0	0

Random Forest performance after post-processing

Accuracy 11.54 %

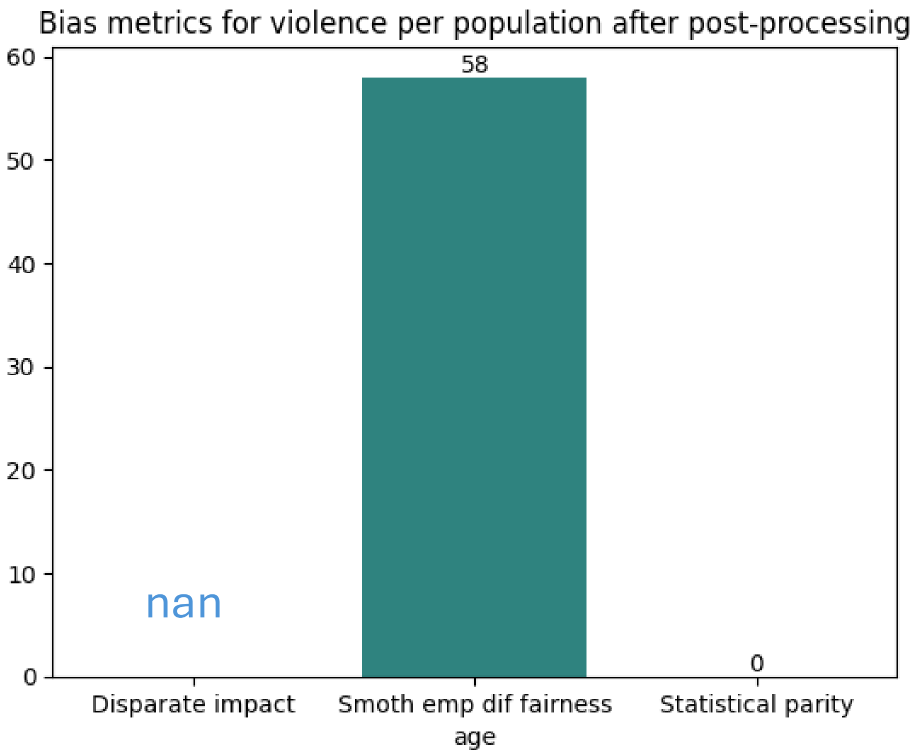
Precision 22.76 %

F1-Score 11.27 %

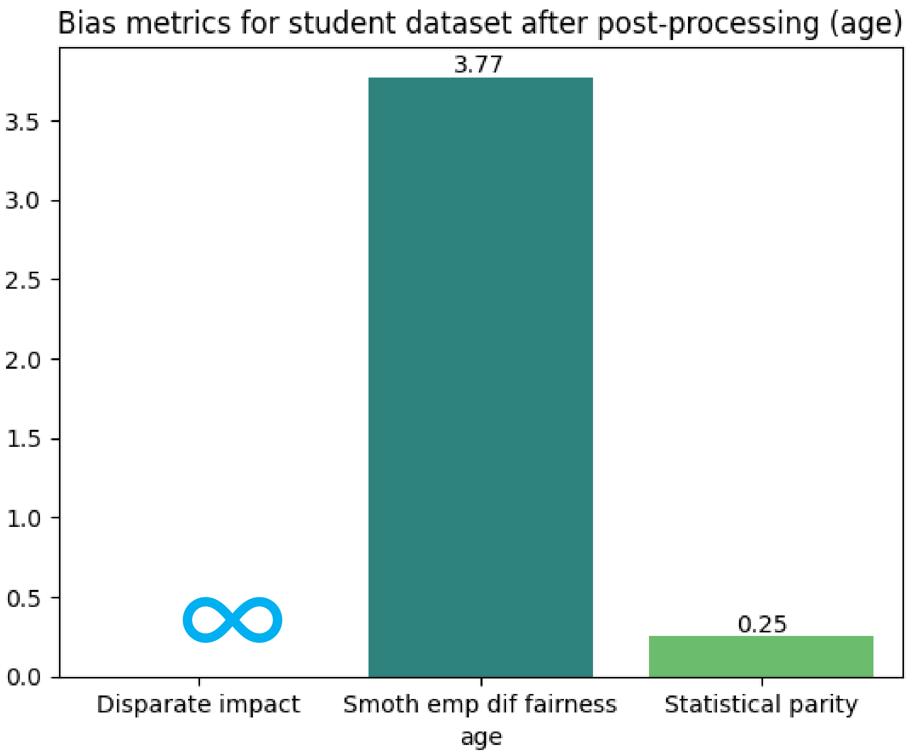
Recall 11.54 %

Post-Processing Aif360

Crime



Student



Research Question 2

Fairlearn

- Easier applicable
- Very well suited for bias detection
- Fast results
- Very well suited for gaining an overview
- Good results in combination with pre and post-processing
- In-Processing a bit difficult
  - If training data not adequately represents all sensitive groups, difficult to tweak correctly

Aif360

- Limited bias detection
- More challenging to set up and get to work propperly
- Once set up properly, more mitigation options and metrics
- (Api to sklearn)

Research Question 3

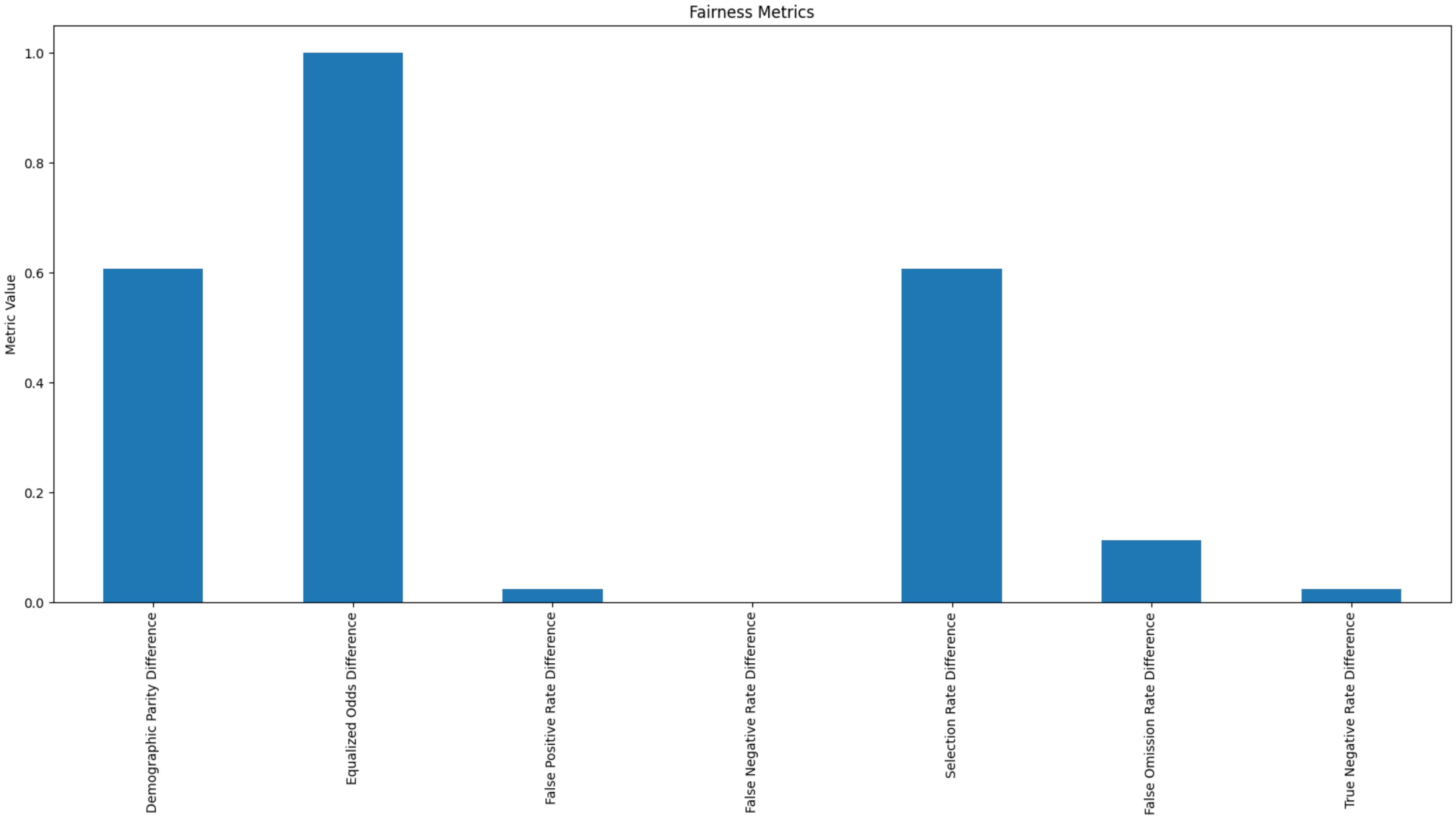
- Biases in the crime dataset are generally more extreme, particularly related to socio-economic factors
- Student dataset shows high biases primarily in absence and previous failures (personal behavior).
- Bias mitigation techniques reduce these biases in both dataset.
- But this often comes at the cost of decreased accuracy and precision.
- Highlight trade-offs involved in addressing systemic inequalities and socio-economic disparities, informing our understanding of the complex interplay between education and crime.

Thanks for your Attention

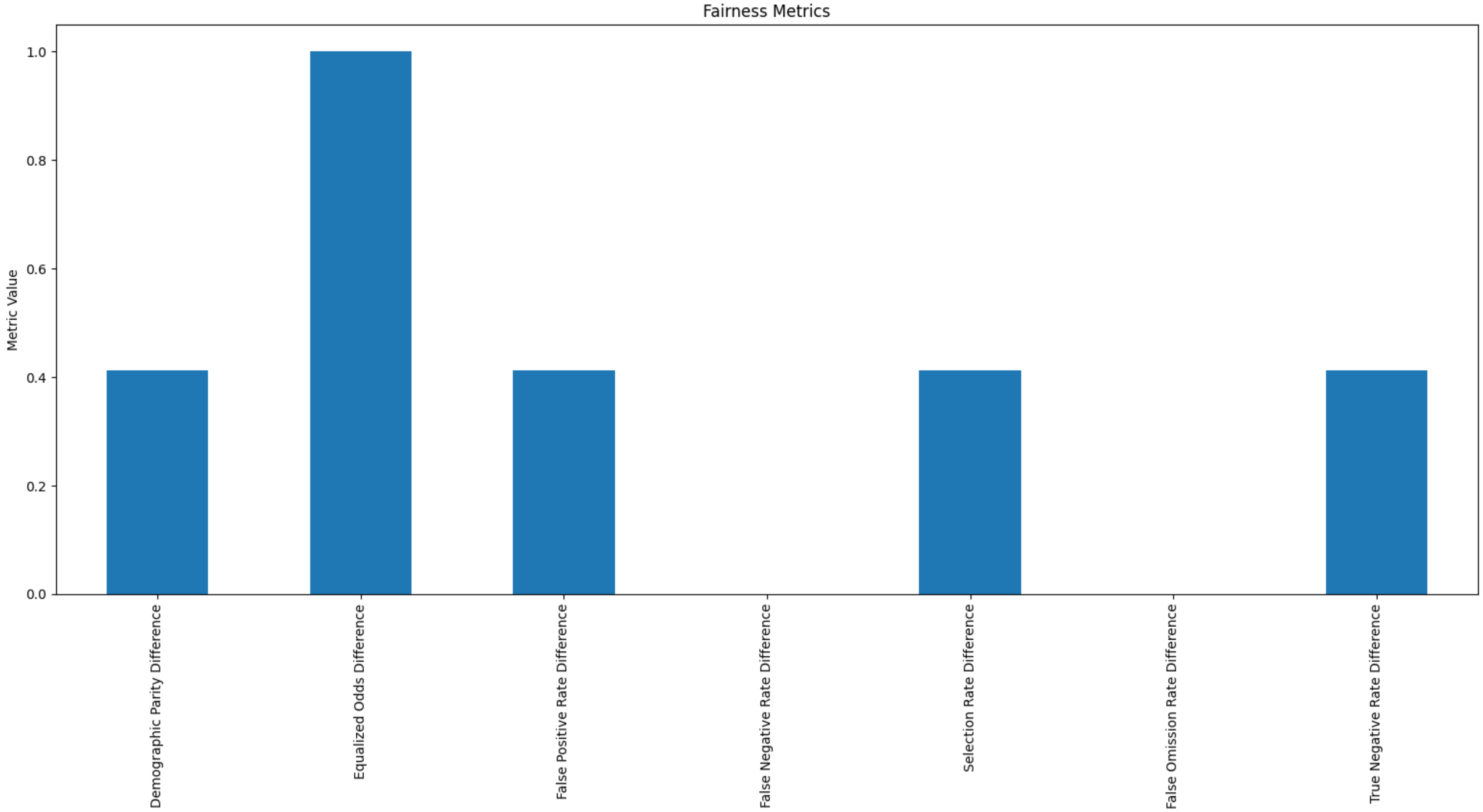
Q & A

Feel free to ask!

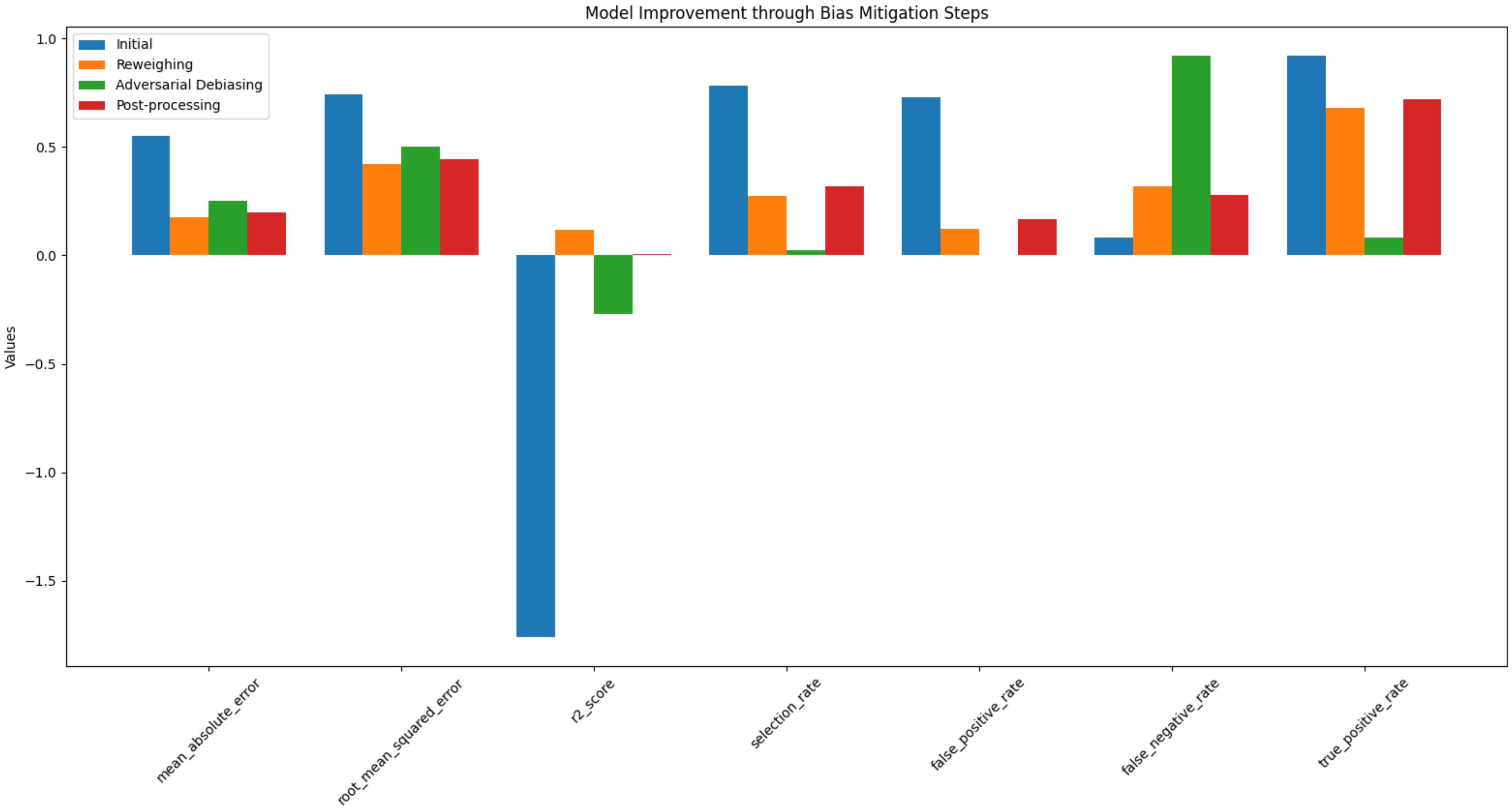
Bias detection results Student Failure



Bias detection results Crime PCTFam2Par

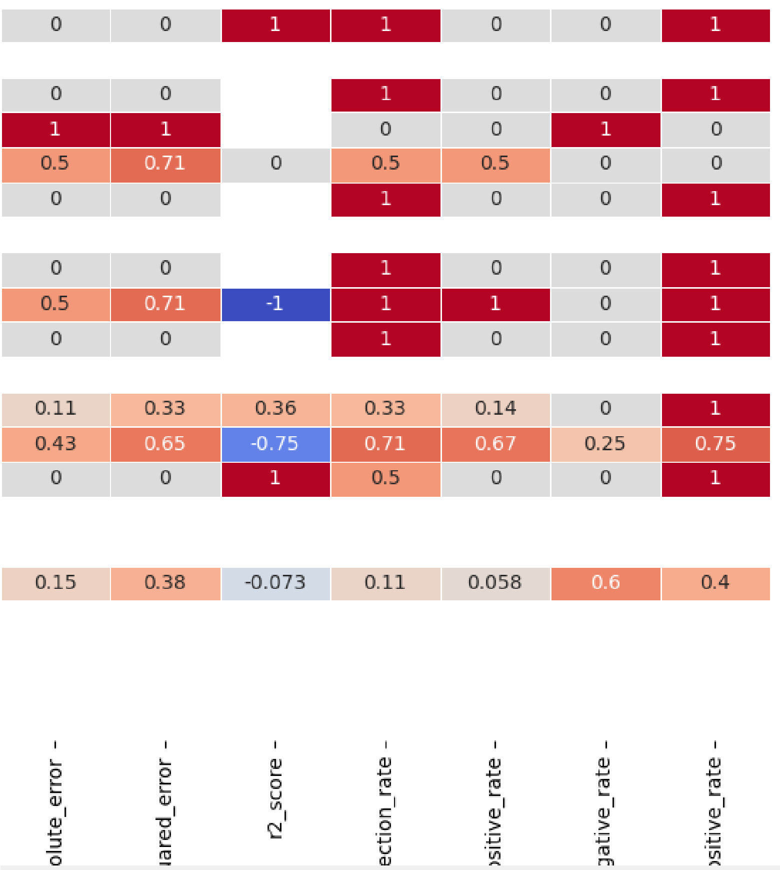


Bias Mitigation Crime Grouped features

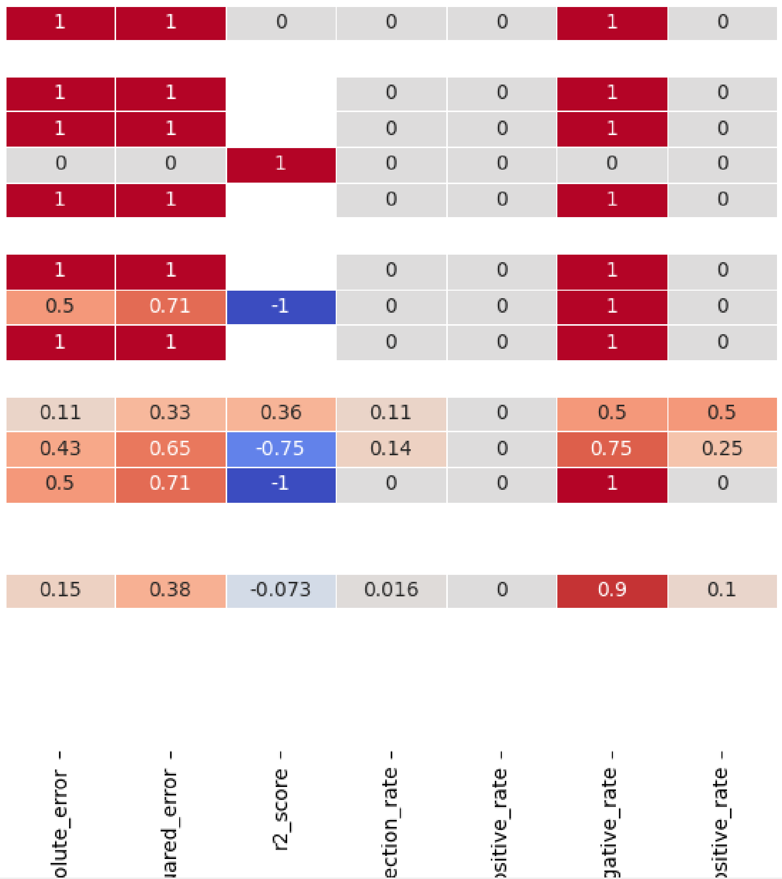


Bias Mitigation Crime Grouped features

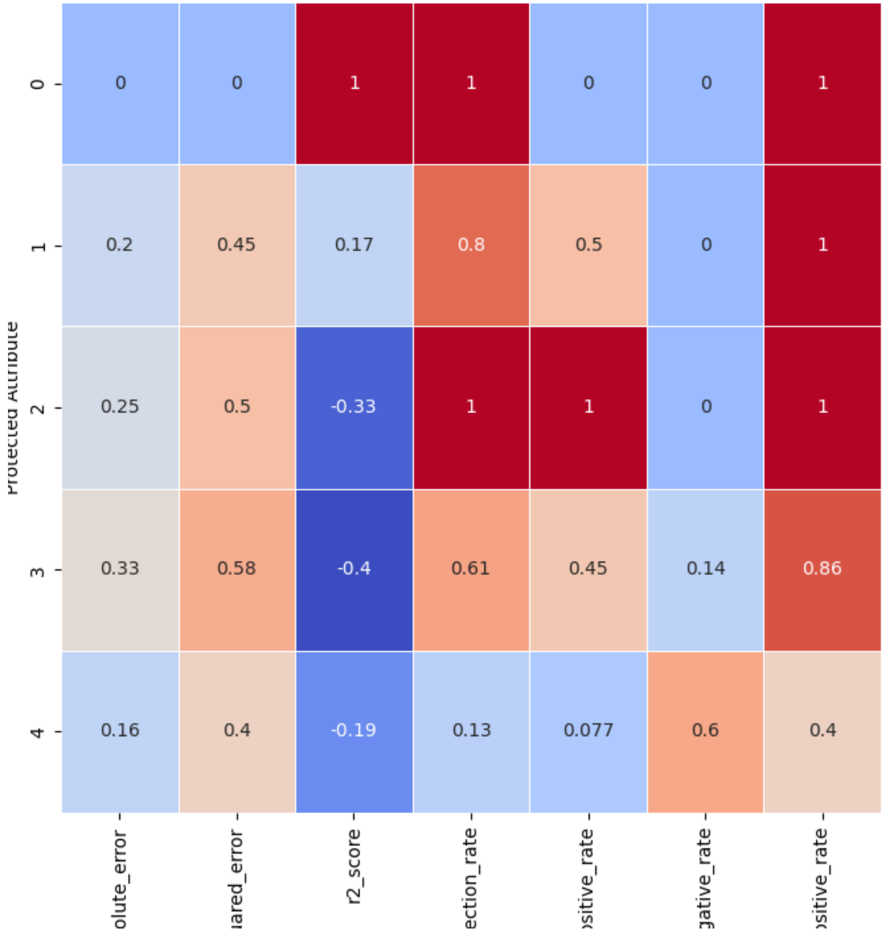
Heatmap of Metrics by Group for Reweighing Model



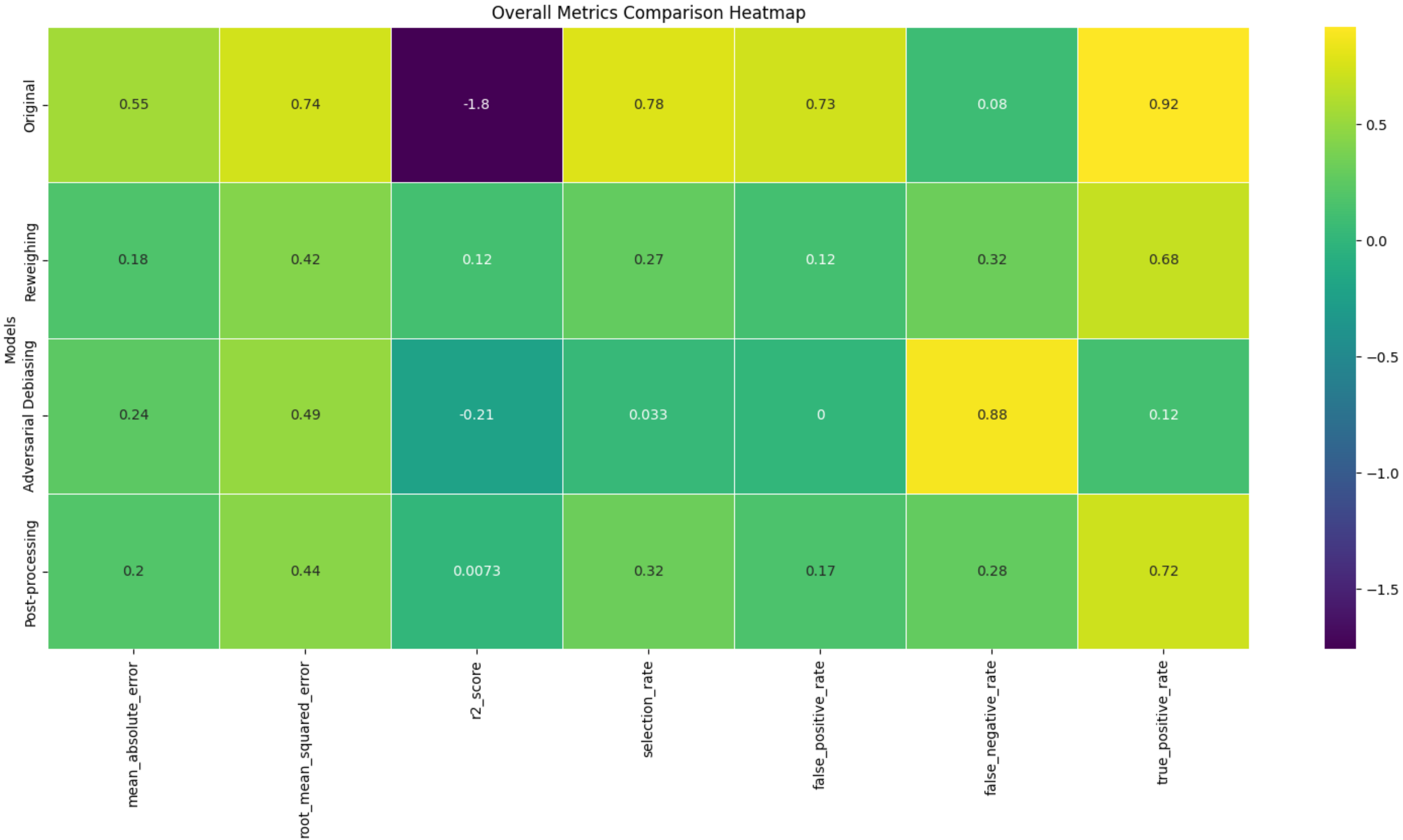
Heatmap of Metrics by Group for Adversarial Debiasing Model



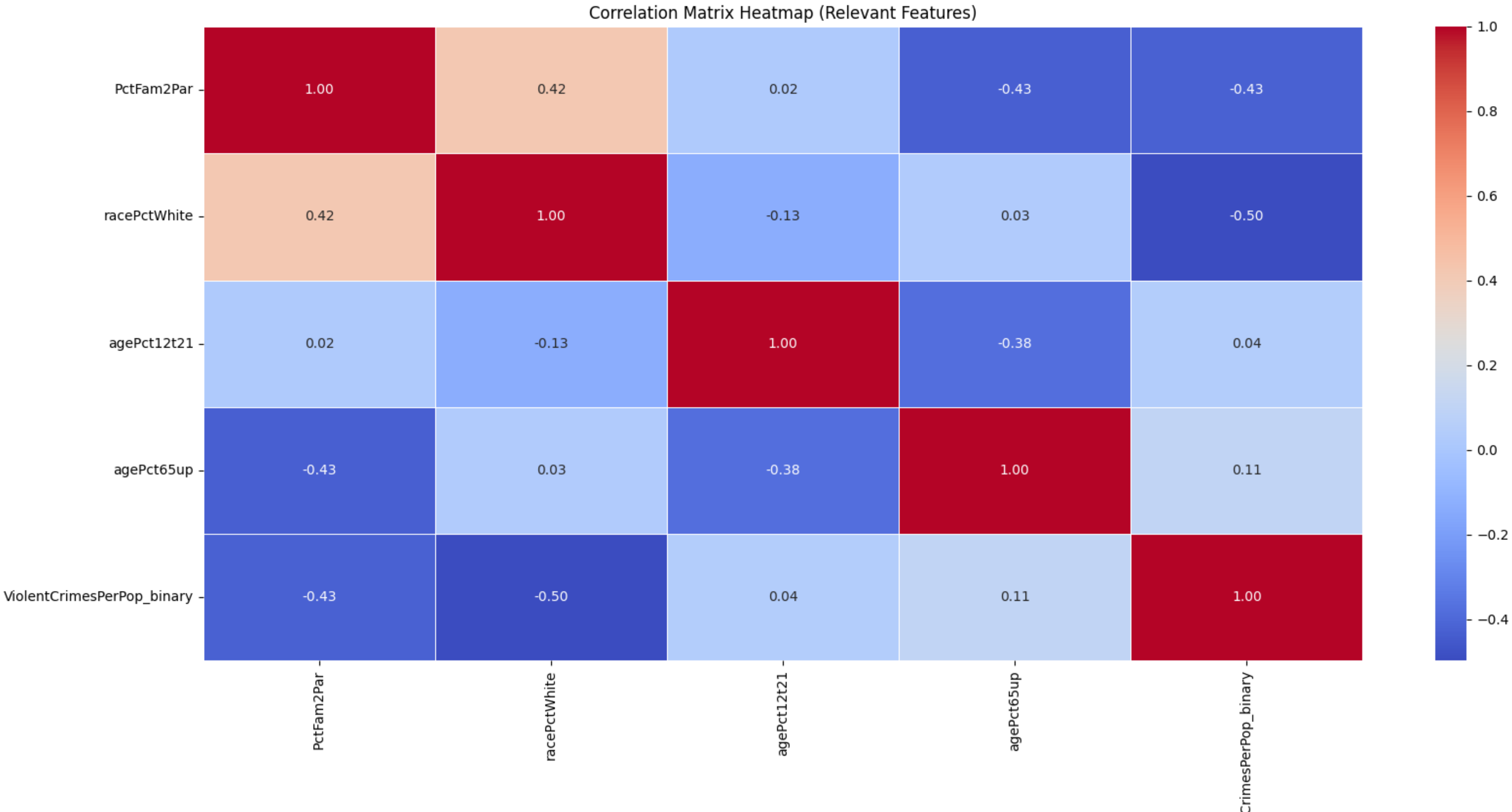
Heatmap of Metrics by Group for Post-processing Model



Bias Mitigation Crime Grouped features



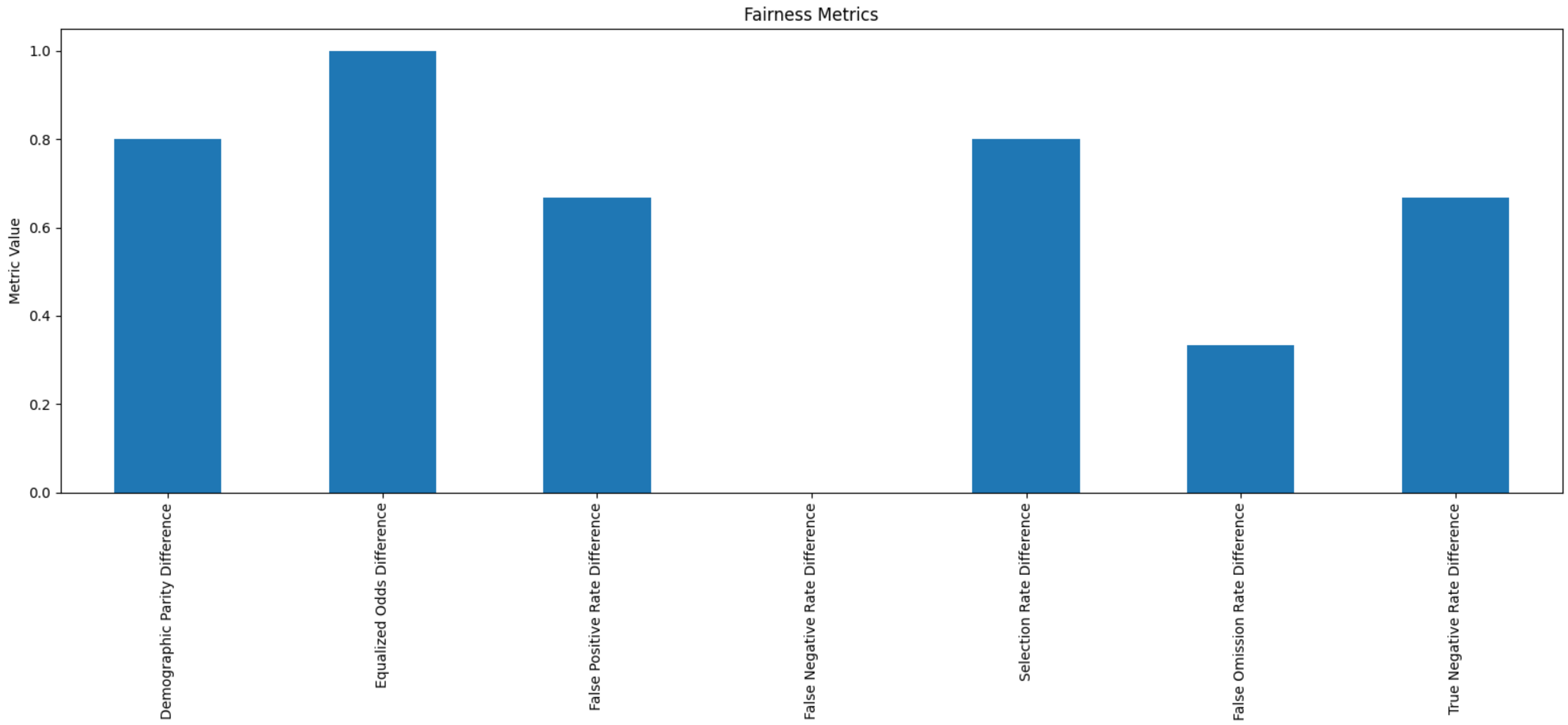
Bias Mitigation Crime indirect correlation



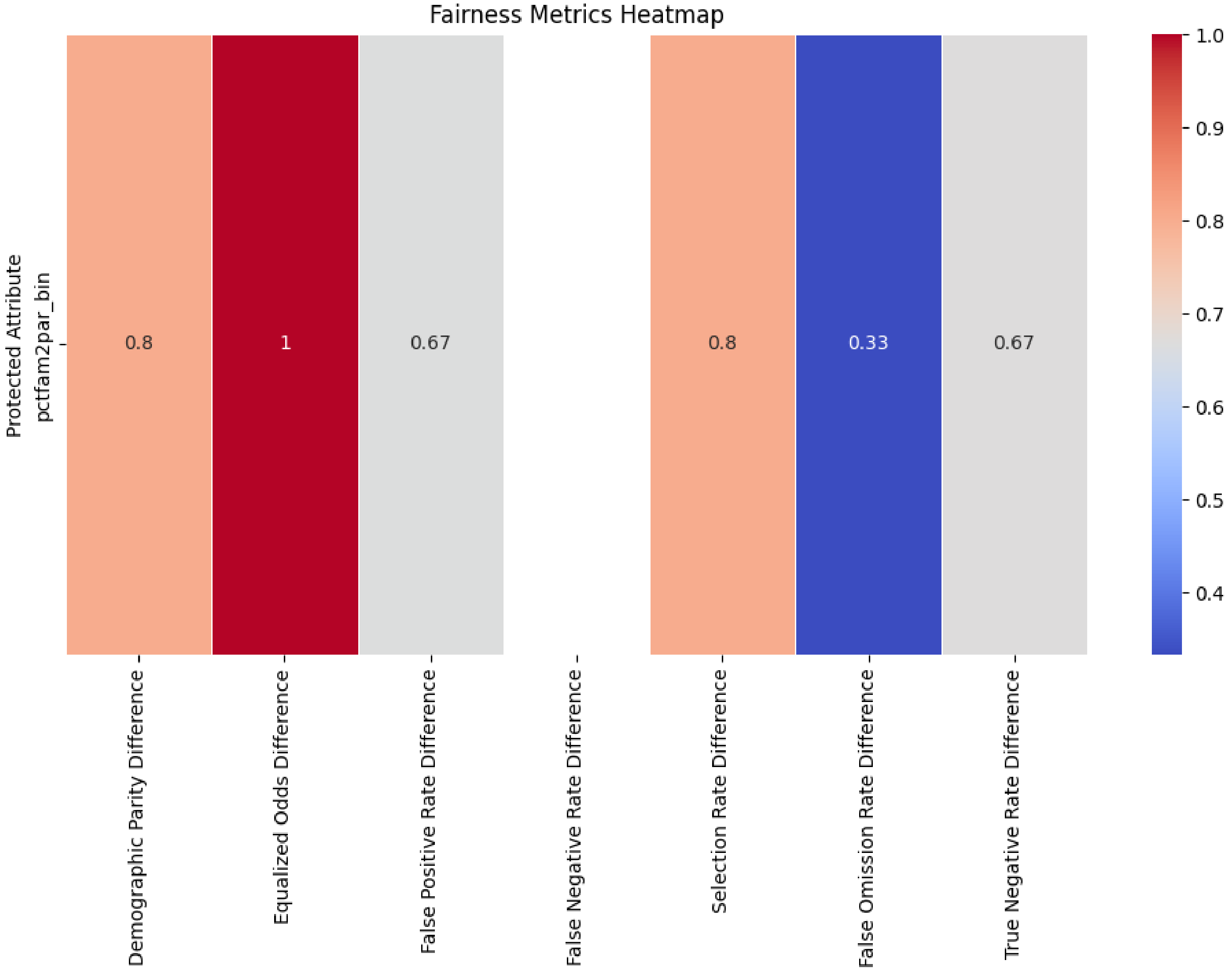
Correlation Matrix Heatmap (Relevant Features)

	PctFam2Par	racePctWhite	agePct12t21	agePct65up	CrimesPerPop_binary
PctFam2Par	1.00	0.42	0.02	-0.43	-0.43
racePctWhite	0.42	1.00	-0.13	0.03	-0.50
agePct12t21	0.02	-0.13	1.00	-0.38	0.04
agePct65up	-0.43	0.03	-0.38	1.00	0.11
ViolentCrimesPerPop_binary	-0.43	-0.50	0.04	0.11	1.00

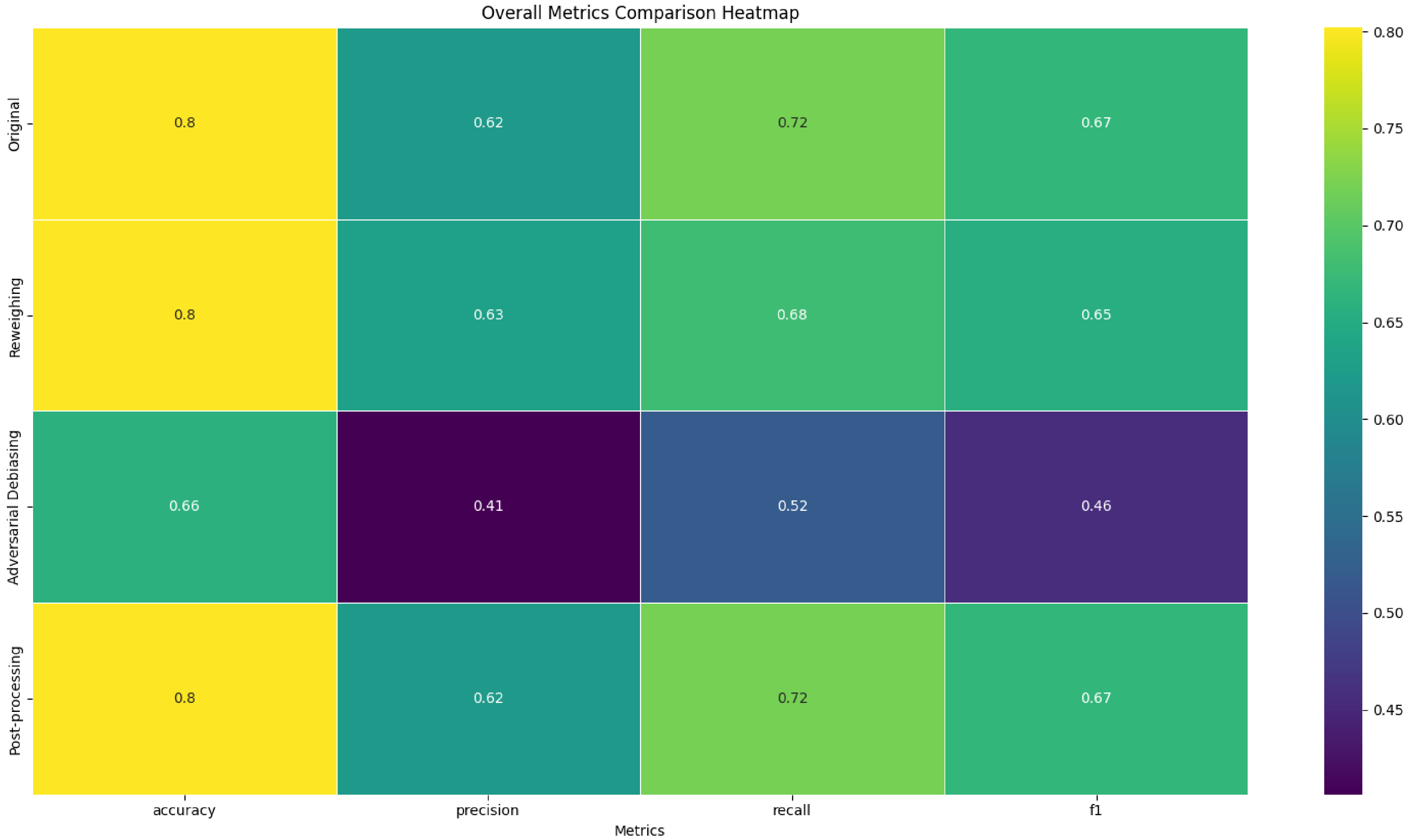
Bias Mitigation Crime indirect correlation



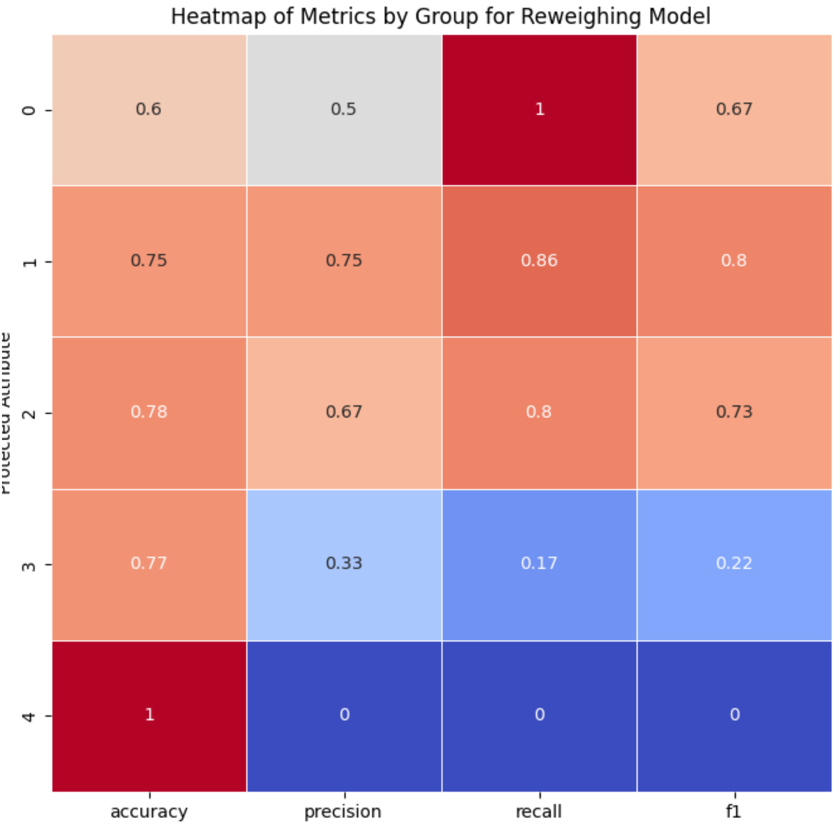
Bias Mitigation Crime indirect correlation



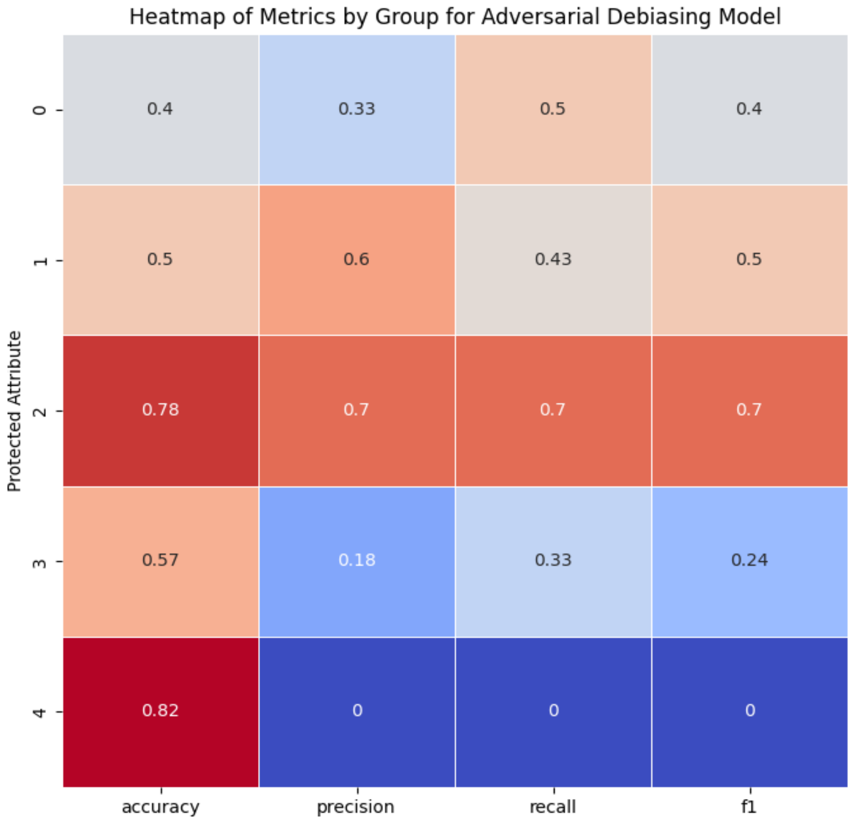
Bias Mitigation Crime indirect correlation



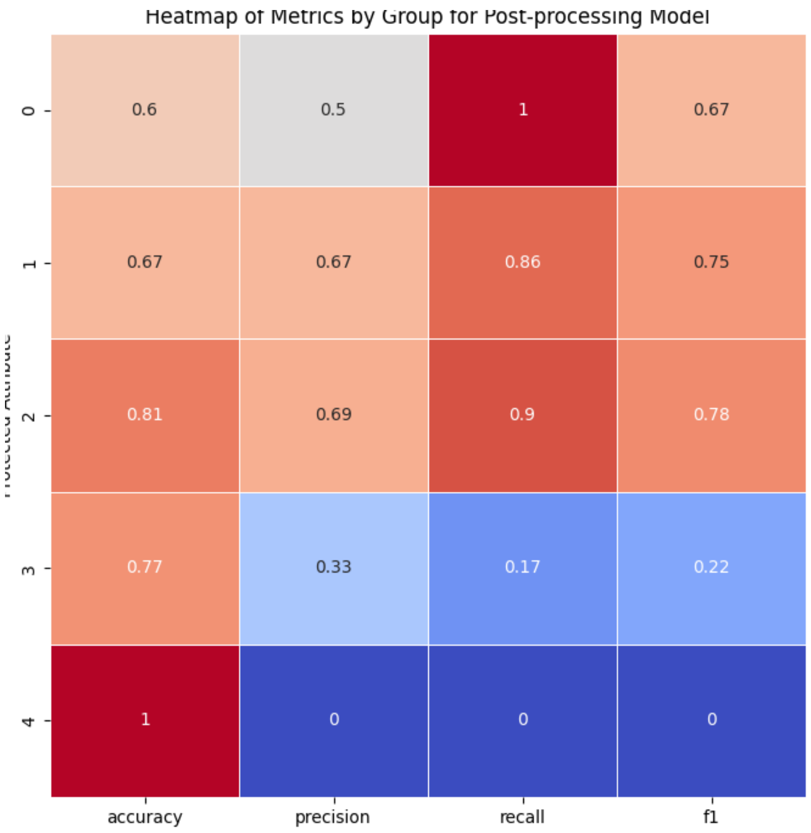
Bias Mitigation Crime indirect correlation


Heatmap of Metrics by Group for Reweighing Model

Protected Attribute	accuracy	precision	recall	f1
0	0.6	0.5	1	0.67
1	0.75	0.75	0.86	0.8
2	0.78	0.67	0.8	0.73
3	0.77	0.33	0.17	0.22
4	1	0	0	0

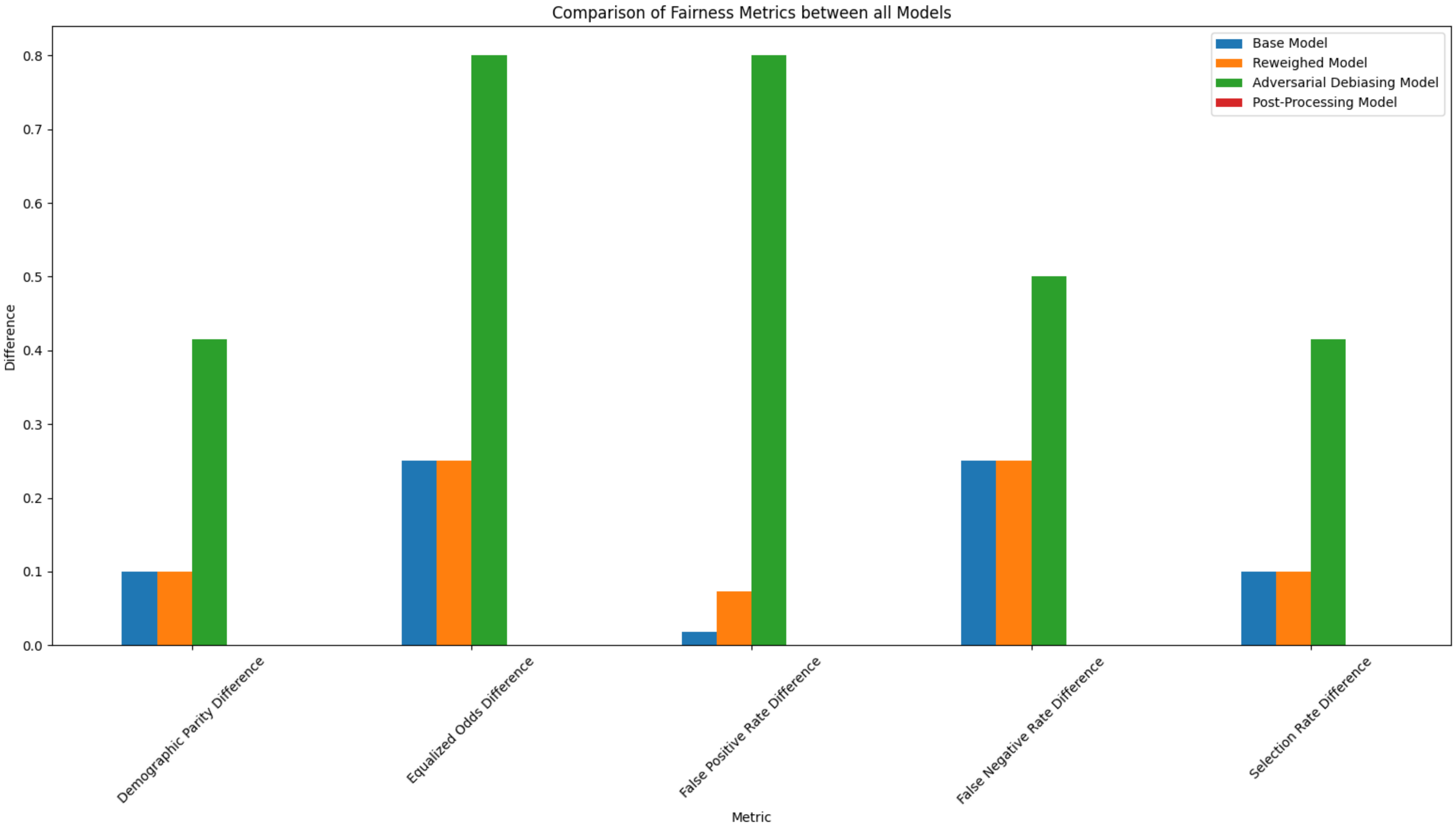

Heatmap of Metrics by Group for Adversarial Debiasing Model

Protected Attribute	accuracy	precision	recall	f1
0	0.4	0.33	0.5	0.4
1	0.5	0.6	0.43	0.5
2	0.78	0.7	0.7	0.7
3	0.57	0.18	0.33	0.24
4	0.82	0	0	0

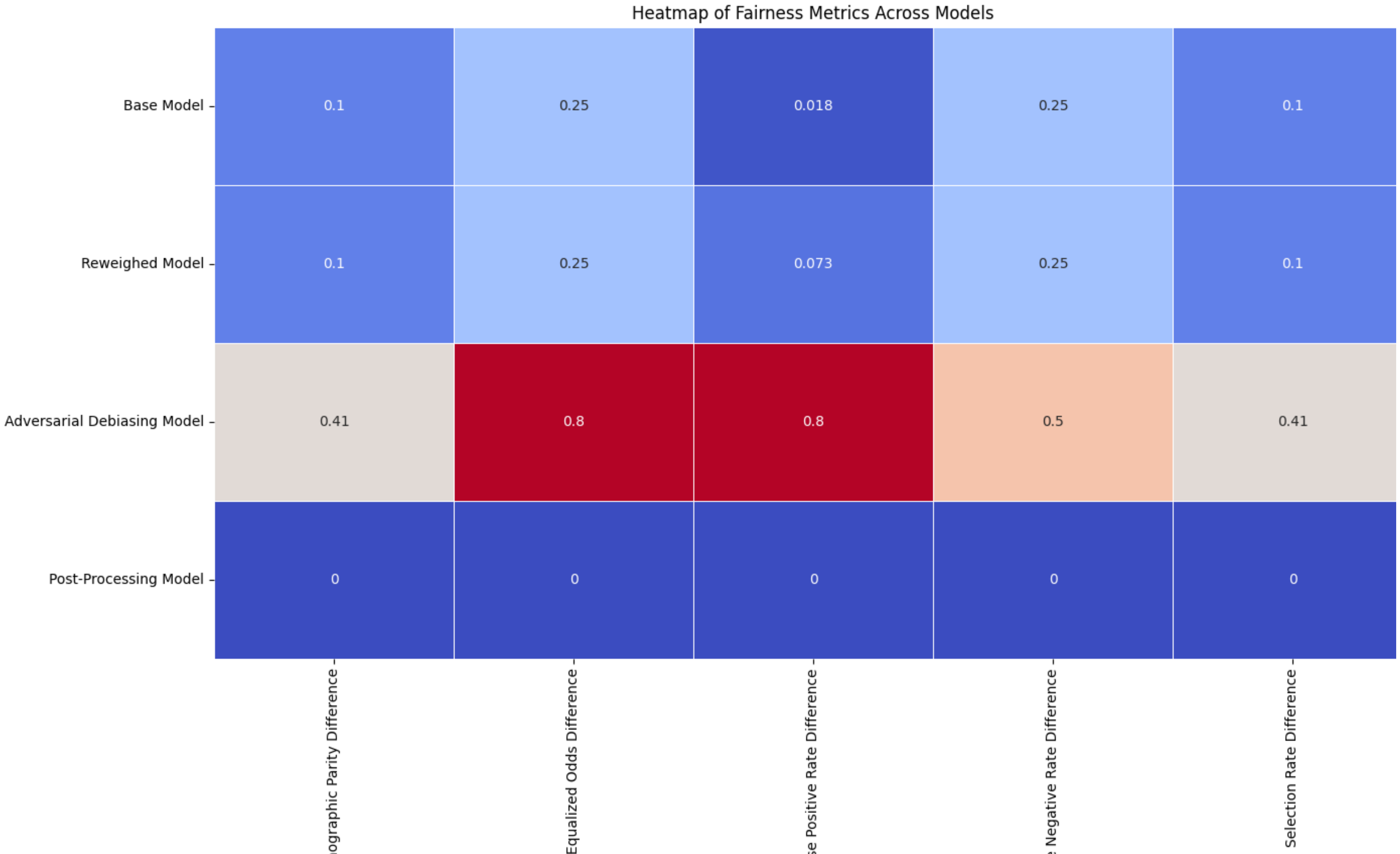

Heatmap of Metrics by Group for Post-processing Model

Protected Attribute	accuracy	precision	recall	f1
0	0.6	0.5	1	0.67
1	0.67	0.67	0.86	0.75
2	0.81	0.69	0.9	0.78
3	0.77	0.33	0.17	0.22
4	1	0	0	0

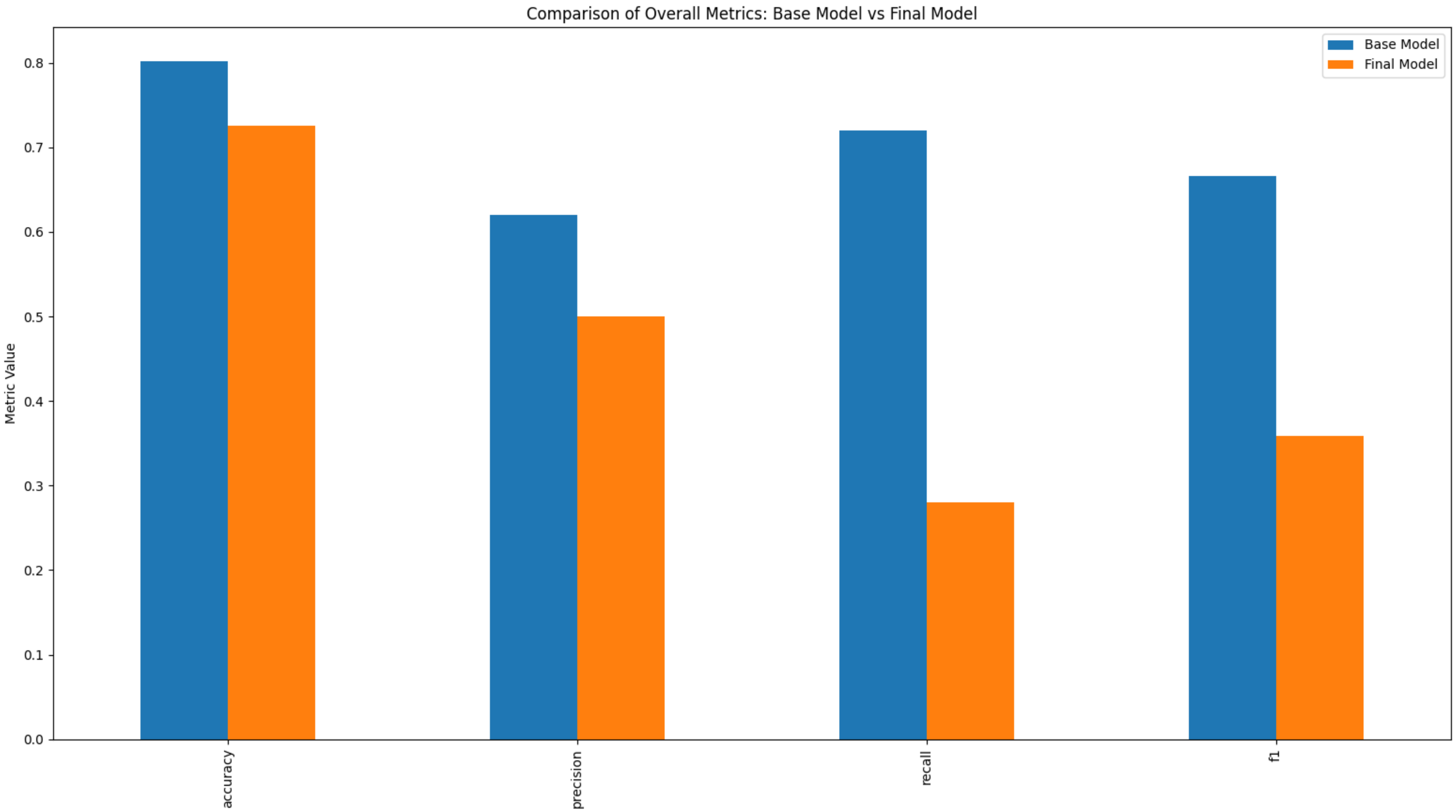
Bias Mitigation Student Grouped features



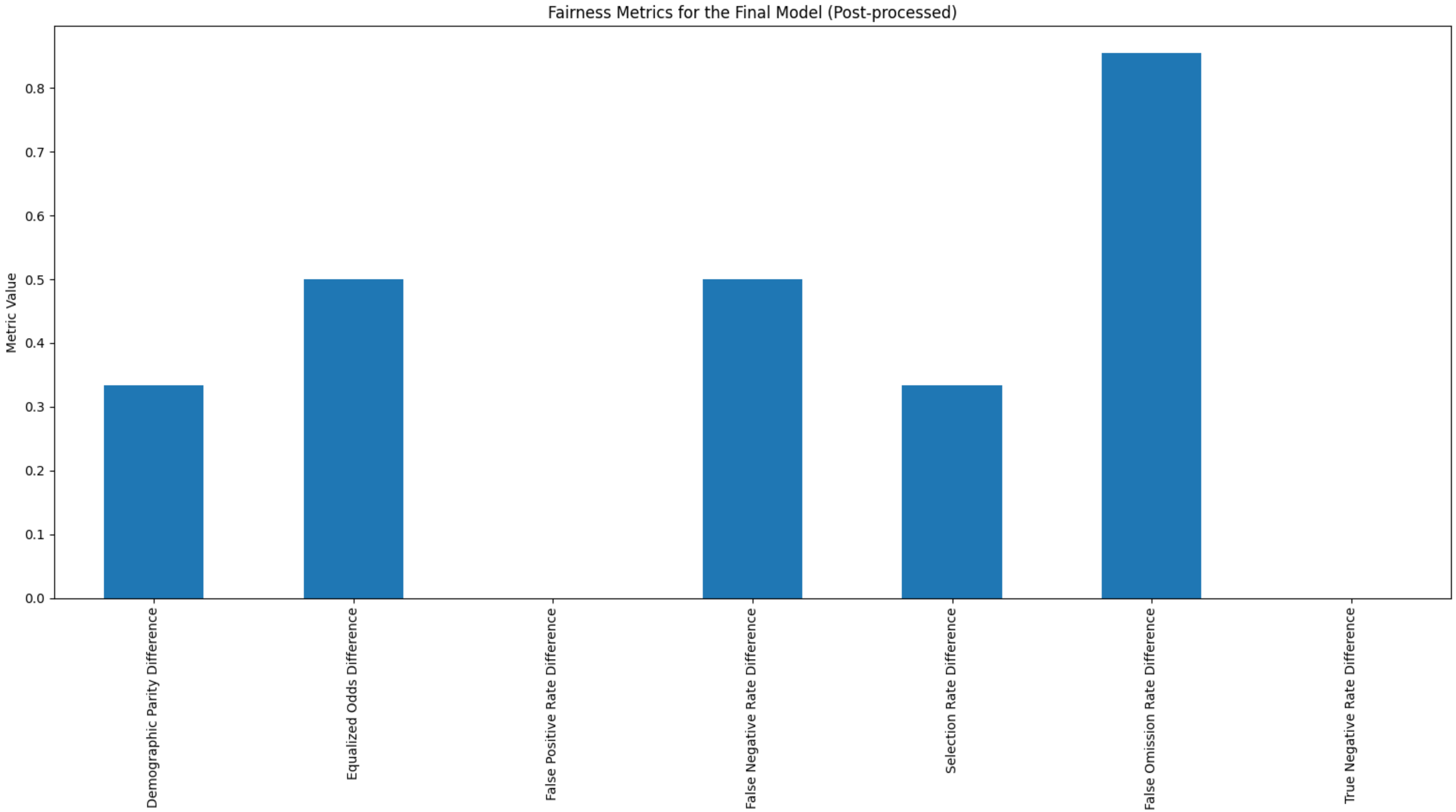
Bias Mitigation Student Grouped features



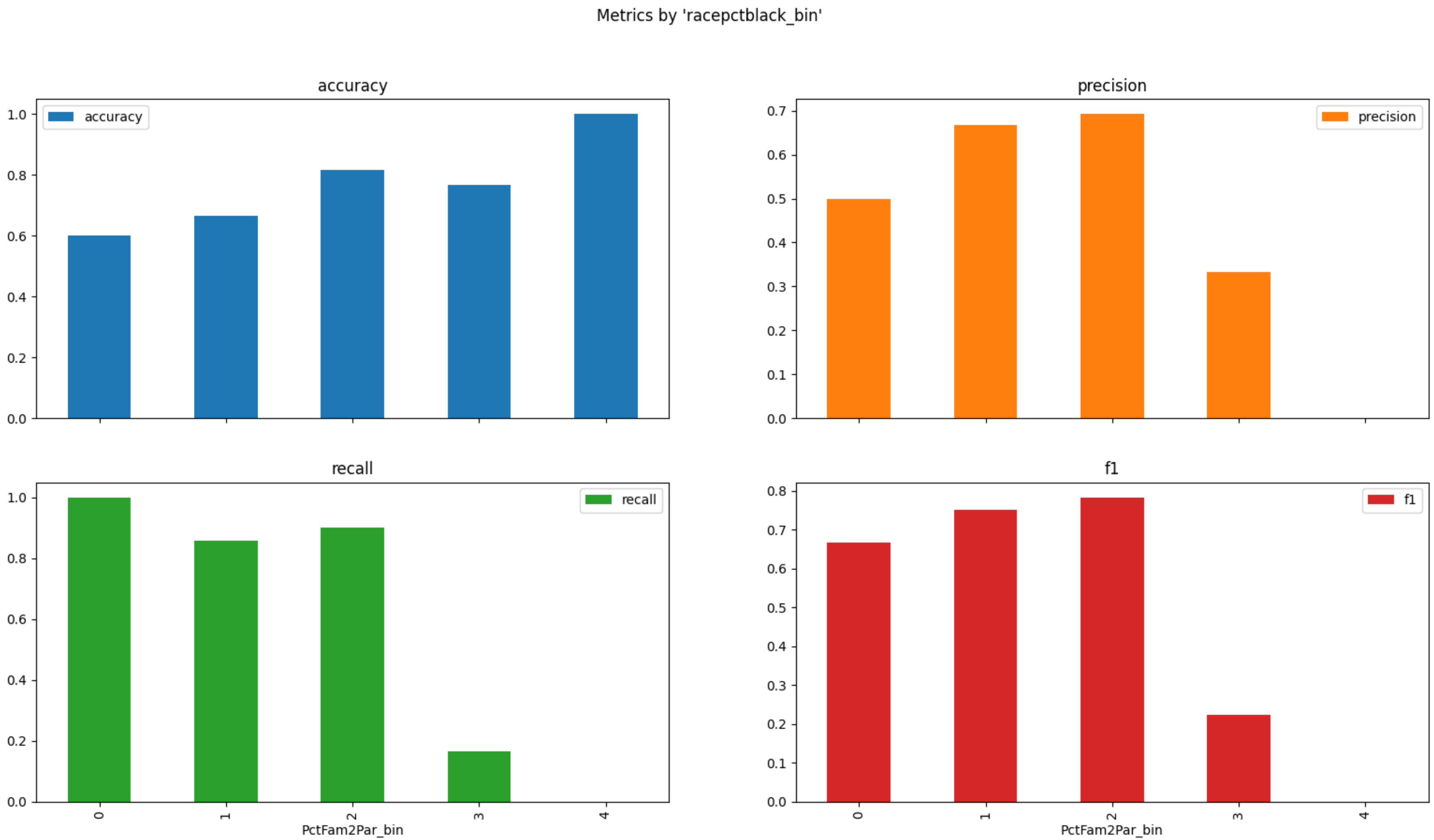
Bias Mitigation Crime Iterative Mitigation



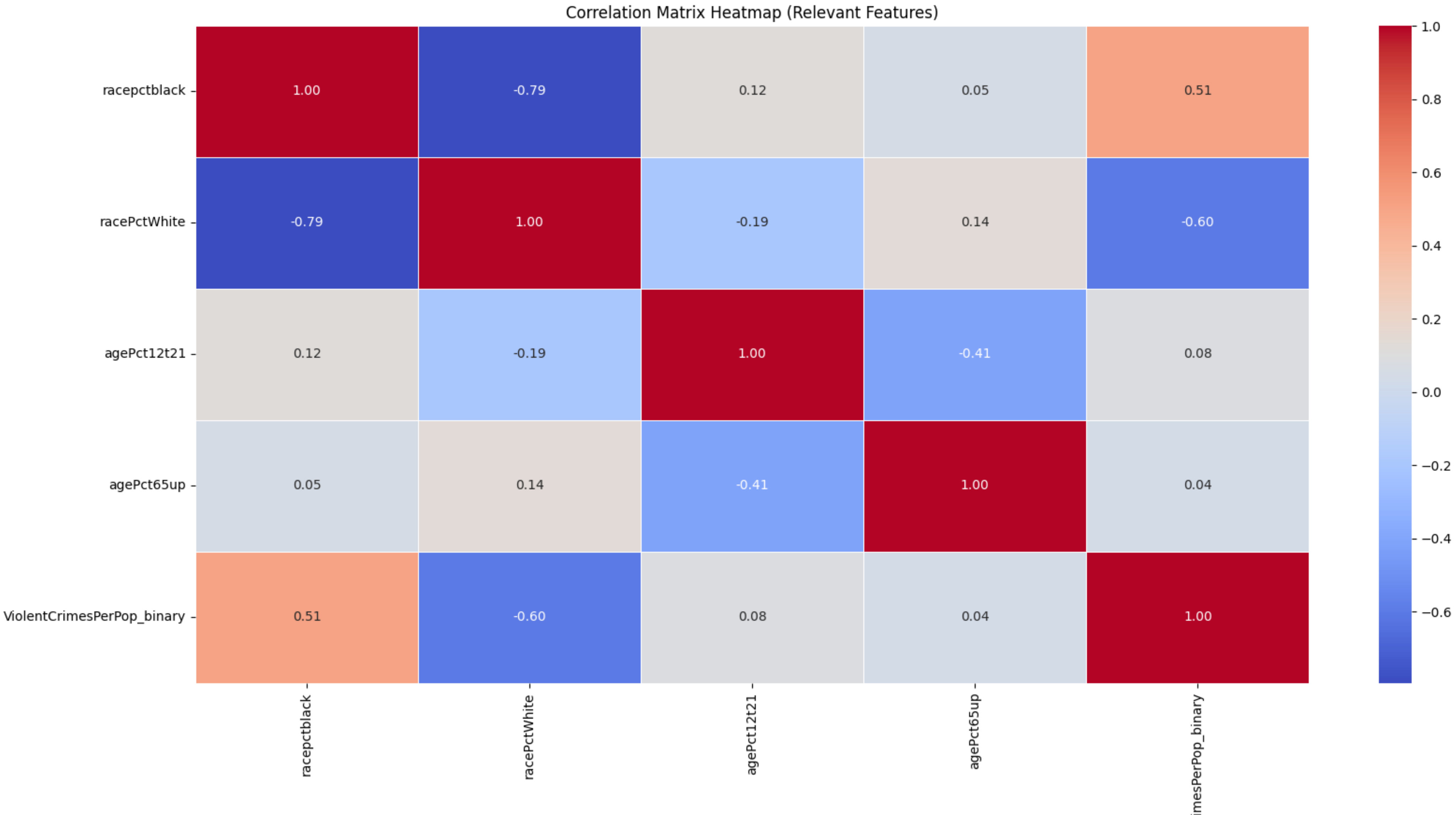
Bias Mitigation Crime Iterative Mitigation



Bias Mitigation Crime Raw Data



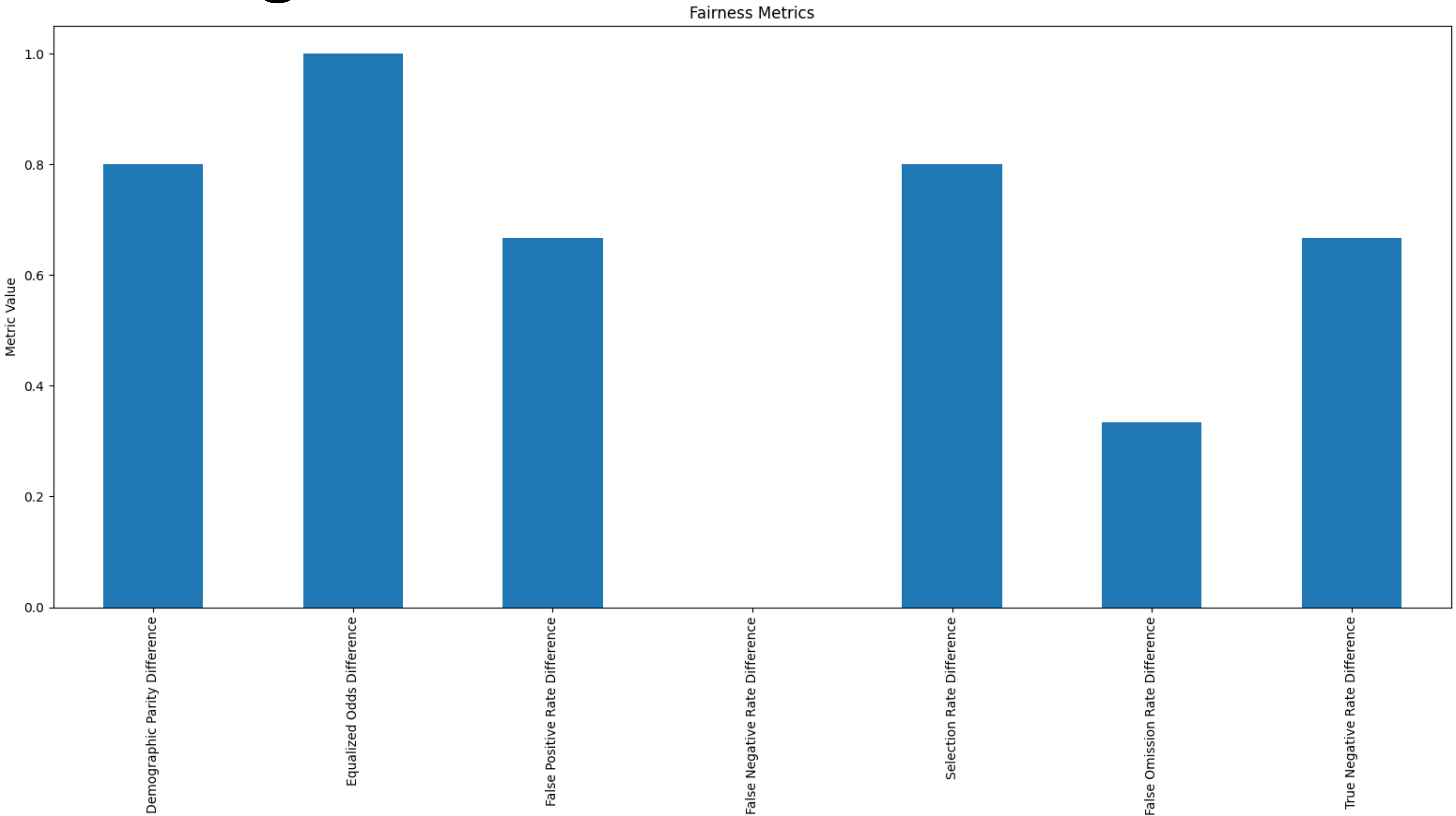
Bias Mitigation Crime Raw Data



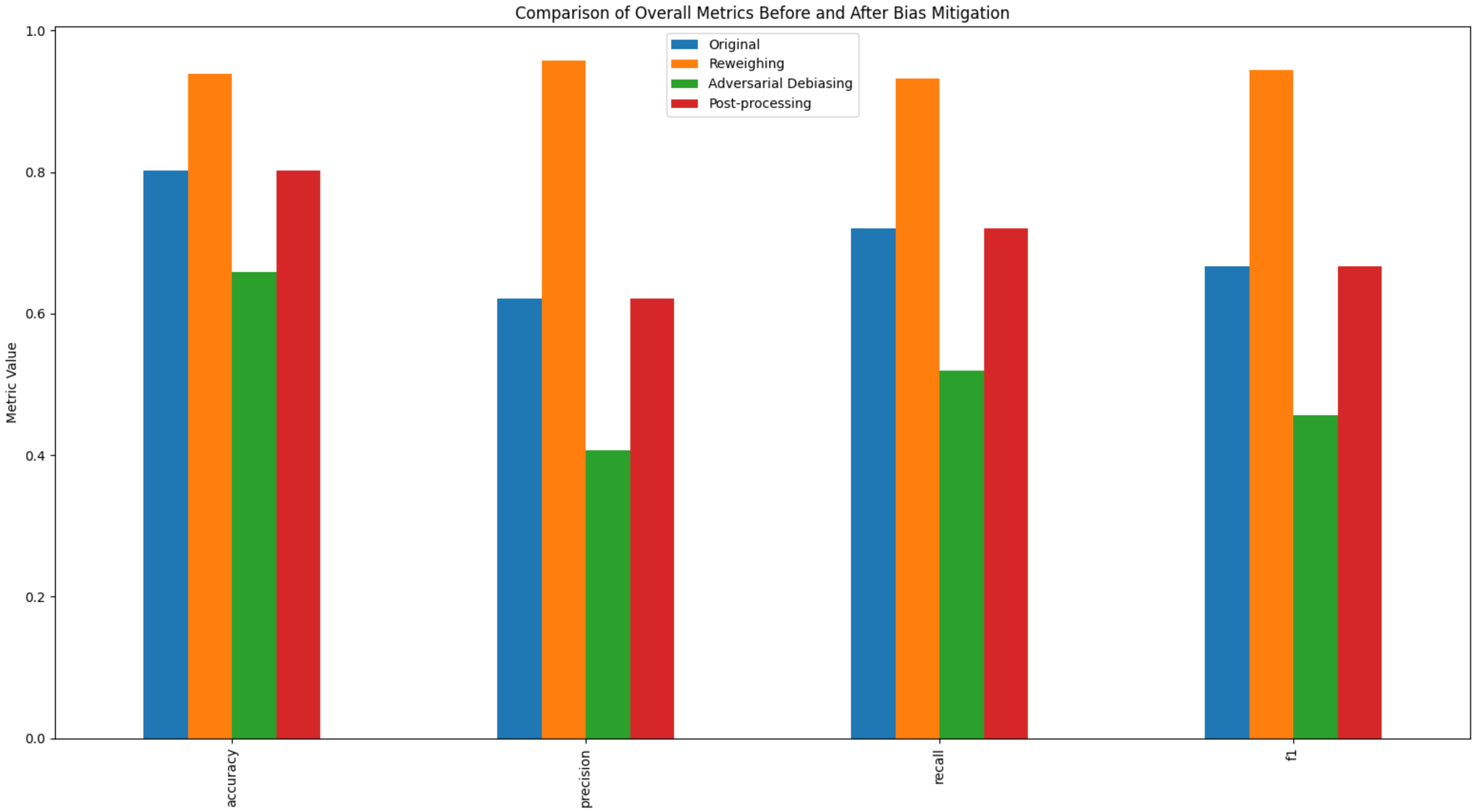
Correlation Matrix Heatmap (Relevant Features)

	racepctblack	racePctWhite	agePct12t21	agePct65up	imesPerPop_binary
racepctblack	1.00	-0.79	0.12	0.05	0.51
racePctWhite	-0.79	1.00	-0.19	0.14	-0.60
agePct12t21	0.12	-0.19	1.00	-0.41	0.08
agePct65up	0.05	0.14	-0.41	1.00	0.04
ViolentCrimesPerPop_binary	0.51	-0.60	0.08	0.04	1.00

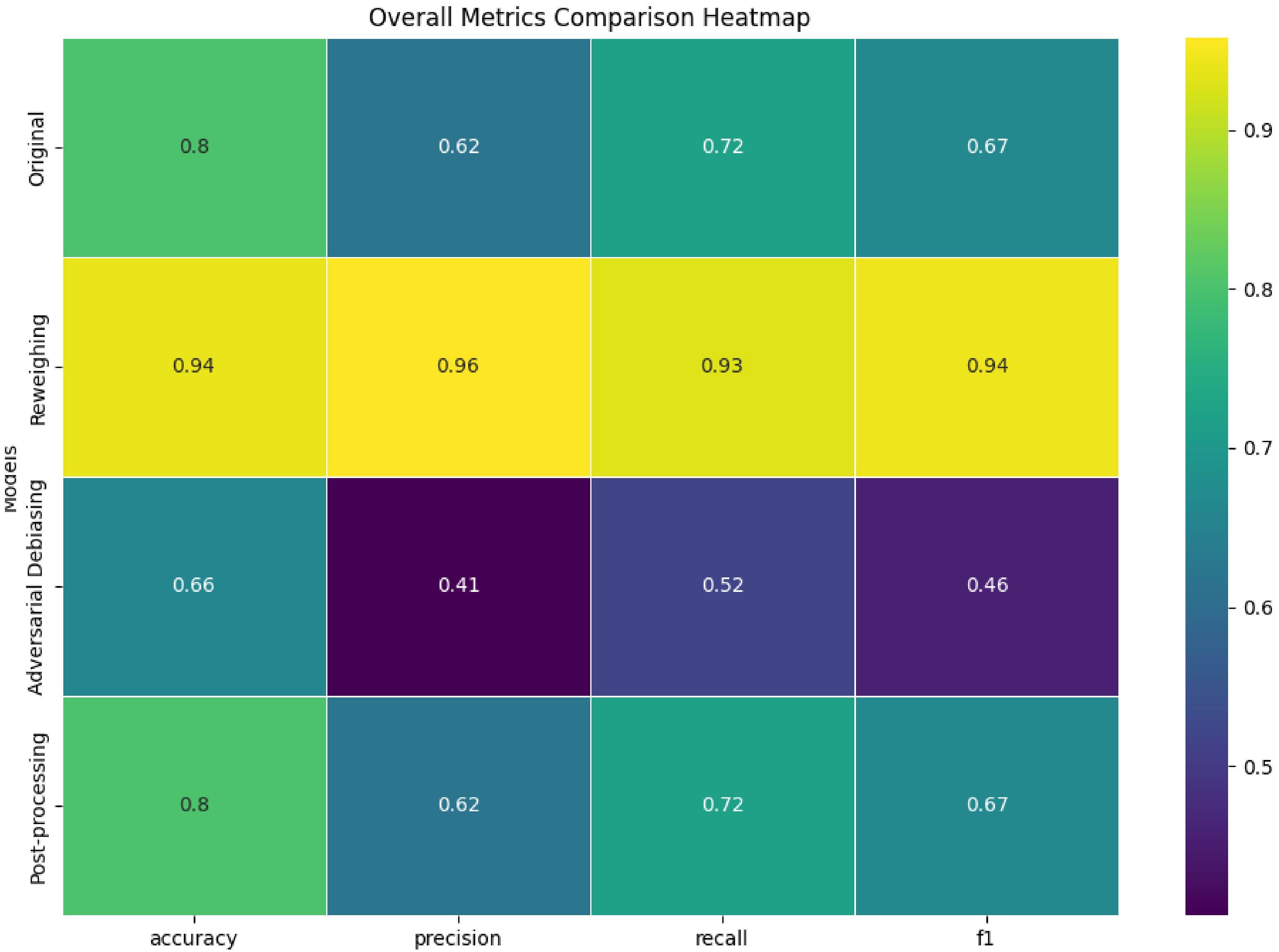
Bias Mitigation Crime Raw Data



Bias Mitigation Crime Raw Data

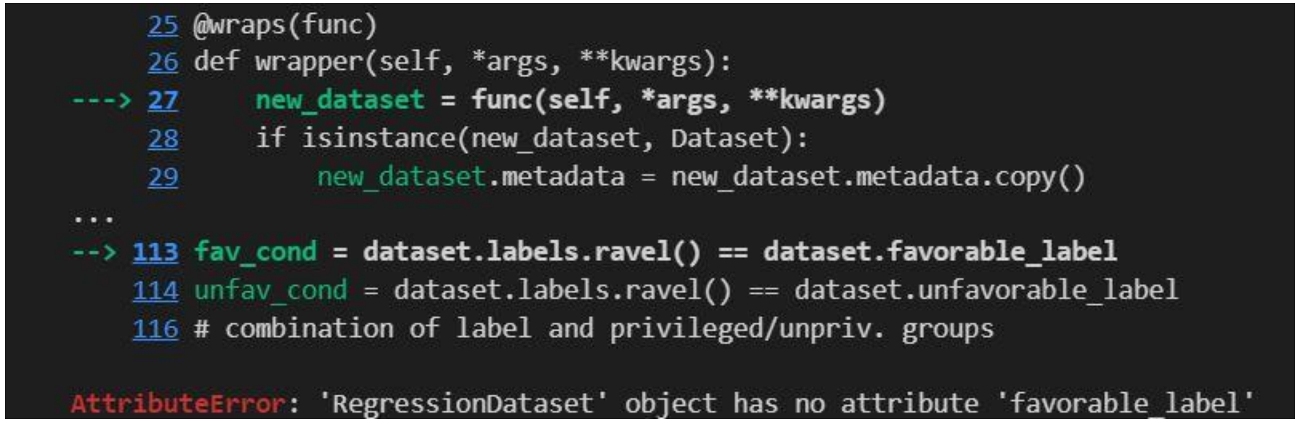


Bias Mitigation Crime Raw Data



Aif360 - Test of RegressionDataset

Attribute only available in a StructuredDataset



```
     25 @wraps(func)
     26 def wrapper(self, *args, **kwargs):
---> 27     new_dataset = func(self, *args, **kwargs)
     28     if isinstance(new_dataset, Dataset):
     29         new_dataset.metadata = new_dataset.metadata.copy()
...
--> 113 fav_cond = dataset.labels.ravel() == dataset.favorable_label
    114 unfav_cond = dataset.labels.ravel() == dataset.unfavorable_label
    116 # combination of label and privileged/unpriv. groups

AttributeError: 'RegressionDataset' object has no attribute 'favorable_label'
```

Bias Detection Metrics Explanation

- **Demographic Parity Difference**:
  - Measures the difference in selection rates (i.e., the proportion of positive predictions) across different demographic groups. A lower value indicates more equal treatment of different groups by the model.
- **Equalized Odds Difference**:
  - Compares the true positive rates and false positive rates between demographic groups. A lower value suggests that the model's error rates are more evenly distributed among different groups, indicating fairer performance.
- **False Positive Rate Difference**:
  - Indicates the difference in the rate of false positives (incorrectly predicting a positive outcome) between groups. A lower value means the model is equally likely to falsely identify a positive case across groups.
- **False Negative Rate Difference**:
  - Shows the difference in the rate of false negatives (incorrectly predicting a negative outcome) between groups. A lower value suggests that the model misses positive cases at similar rates across groups.
- **Selection Rate Difference**:
  - Measures the variation in the proportion of positive predictions made by the model across different demographic groups. A smaller value implies a more equitable distribution of positive predictions.
- **False Omission Rate Difference**:
  - Represents the difference in the rate at which actual positives are misclassified as negatives (false omissions) between groups. A lower value indicates that the model is equally reliable in predicting negatives across groups.
- **True Negative Rate Difference**:
  - Reflects the difference in the rate of true negatives (correctly predicting a negative outcome) between groups. A lower value means the model is similarly effective at identifying negatives across different demographic groups.

Privileged vs. Unprivileged

- **Privileged and unprivileged features** are used in bias mitigation to identify and address disparities in machine learning models.
- **Privileged features** refer to attributes that correspond to groups receiving favorable treatment or advantages (e.g., higher socioeconomic status, majority racial groups)
- **Unprivileged features** refer to attributes that correspond to disadvantaged or underrepresented groups (e.g., lower socioeconomic status, minority racial groups).
- These distinctions help in defining fairness constraints and ensuring that the model does not disproportionately benefit the privileged groups over the unprivileged ones, thereby promoting equity and reducing bias in predictions.

Data Cleaning

- **Imputation with Mean Values**
- **Dropping Non-Essential Columns**
- **IQR Method for Outlier Removal**
- **Standardization of Numerical Features**
- **Correlation Matrix Visualization**

Bias Mitigation - Worse Results

- **Reweighing:**
  - Reweighing adjusts the weights of different samples to reduce bias, but it can introduce noise or instability in the model's learning process, leading to worse performance.
  - It can also cause the model to overfit on certain samples with higher weights, reducing its ability to generalize and increasing bias in underrepresented groups.

Bias Mitigation - Worse Results

- **Adversarial Debiasing:**
  - Adversarial debiasing involves a complex training process with competing objectives, which can sometimes lead to suboptimal solutions if the adversary is not strong enough or if the training is not properly balanced.
  - The method may inadvertently amplify biases if the adversarial network overpowers the primary model, causing instability and poor bias mitigation performance.

Bias Mitigation - Worse Results

- **Post-Processing:**
  - Post-processing techniques adjust predictions after the model has been trained, which can sometimes lead to inconsistencies and unintended consequences if the adjustments are too aggressive or not well-calibrated.
  - These techniques can also fail to address underlying biases in the model's learned representations, resulting in persistent or even amplified biases in the final predictions.

Limitations

- One significant challenge was ensuring the quality and consistency of the datasets.
- Potential trade-off between reducing bias and maintaining model performance.
- Challenges in interpreting the results of certain bias mitigation techniques, particularly when they performed worse than the base model. - Solved with: Additional analyses and considering context-specific factors that might influence these outcomes
- Expand the groups of sensitive features even more and perform even more in depth tests

Further metrics

- **MDSS Score**
  - **Detect bias by "minimizing noise" in a multidimensional dataset, essentially detecting where most uneven distributions are located.**
- **SED**
  - **The Smoothed empirical differential (SED) metric compares the differential of smoothed probability between groups of features.**

1. Z. Zhang, D. Wang, B. Yang and J. Yin, "Weighted Multidimensional Scaling Localization Method With Bias Reduction Based on TOA," in IEEE Sensors Journal, vol. 23, no. 17, pp. 19803-19814, 1 Sept.1, 2023, doi: 10.1109/JSEN.2023.3296986. keywords: {Sensors;Location awareness;Estimation;Mathematical models;Noise measurement;Weight measurement;Time measurement;Bias reduction;Cramér–Rao lower bound (CRLB);multidimensional scaling (MDS);sensor-based localization;time-of-arrival (TOA) measurement}
2. IBM (2024). SED. Available at: https://dataplatform.cloud.ibm.com/docs/content/wsj/model/wos-smooth-empirical-diff.html?context=cpdaas