CS 6603: AI, Ethics, and Society

Georgia Tech College of Computing

Final Project

Abhijith Chakiat, Amisha Buch, Nirali Thakkar

abhijith.c@gatech.edu, abuch6@gatech.edu, nirali@gatech.edu

FINAL PROJECT REPORT

Google drive link for all the files -

https://drive.google.com/drive/folders/1loiBYn2VXVWYC-JW_qv40fn_xN-jzomh?usp=share_link

1 STEP 1

(i) Which dataset did you select?

We selected the 'Students Performance in Exams' dataset from Kaggle. Linked here:
https://www.kaggle.com/datasets/spscientist/students-performance-in-exams

(ii) Which regulated domain does your dataset belong to? - **Education**

(iii) How many observations are in the dataset? - **1000**

(iv) How many variables are in the dataset? **- 8**

(v) Which variables did you select as your dependent variables?

We derived 4 new outcome variables.

- math_result
- reading_result
- writing_result
- overall_result

All the 4 variables will have a value of either **Pass/Fail. (Favorable/Unfavourable)**

(vi) How many and which variables in the dataset are associated with a legally recognized protected class? Which legal precedence/law (as discussed in the lectures) does each protected class fall under?

- **gender (Equal Pay Act of 1963; Civil Rights Act of 1964, 1991)**
- **race/ethnicity (Civil Rights Act of 1964, 1991)**

2 STEP 2

(i) Table documenting the relationship between members and membership categories for each protected class variable (from Step 2.1)

Gender

1. Male
2. Female

Race

1. Group A (White)
2. Group B (Black or African American)
3. Group C (American Indian or Alaska Native)

4. Group D (Asian)
5. Group E (Native Hawaiian or other Pacific Islander)

Since the dataset is small, no subset relationships have been formed and the races from Group A to Group E have been considered individually.

Race/Gender	**Male**	**Female**	**Total**
Group A	53	36	89
Group B	86	104	190
Group C	139	180	319
Group D	133	129	262
Group E	71	69	140

Total	482	518	1000

(ii) Table documenting the relationship between values and discrete categories/numerical values associated with your dependent variables (from Step 2.2)

Gender

1. Male = 1
2. Female = 0 (Here, generally FALSE = 0, but for easier calculations we assign 0 as a category value)

Race

1. Group A = 1
2. Group B = 2
3. Group C = 3
4. Group D = 4
5. Group E = 5

Dependent variable –

Overall_result

1. Pass = 1
2. Fail = 0

Math_result

1. Pass = 1
2. Fail = 0

Reading_result

1. Pass = 1
2. Fail = 0

Writing_result

1. Pass = 1
2. Fail = 0

Dependent Variable	Discrete value	
	Pass	Fail
Overall_result	1	0
Math_result	1	0
Reading_result	1	0
Writing_result	1	0

(iii) Table providing the computed frequency values for the membership categories each protected class variable (from Step 2.3)

1. Dependent Variable - Overall_result

Sr. No.	Independent Variable	Dependent Variable – Overall_result
1.	Male	Frequency of Pass: 301 Frequency of Fail: 181
2.	Female	Frequency of Pass: 425 Frequency of Fail: 93
3.	Group A	Frequency of Pass: 51 Frequency of Fail: 38
4.	Group B	Frequency of Pass: 130 Frequency of Fail: 60
5.	Group C	Frequency of Pass: 230

		Frequency of Fail: 89
6.	Group D	Frequency of Pass: 204 Frequency of Fail: 58
7.	Group E	Frequency of Pass: 111 Frequency of Fail: 29

2. Dependent Variable - Math_result

Sr. No.	Independent Variable	Dependent Variable – Math_result
1.	Male	Frequency of Pass: 356 Frequency of Fail: 126

2.	Female	Frequency of Pass: 321 Frequency of Fail: 197
3.	Group A	Frequency of Pass: 47 Frequency of Fail: 42
4.	Group B	Frequency of Pass: 119 Frequency of Fail: 71
5.	Group C	Frequency of Pass: 206 Frequency of Fail: 113
6.	Group D	Frequency of Pass: 190 Frequency of Fail: 72

7.	Group E	Frequency of Pass: 115 Frequency of Fail: 25

3. Dependent Variable - Reading_result

Sr. No.	Independent Variable	Dependent Variable – Reading_result
1.	Male	Frequency of Pass: 319 Frequency of Fail: 163
2.	Female	Frequency of Pass: 427 Frequency of Fail: 91
3.	Group A	Frequency of Pass: 54

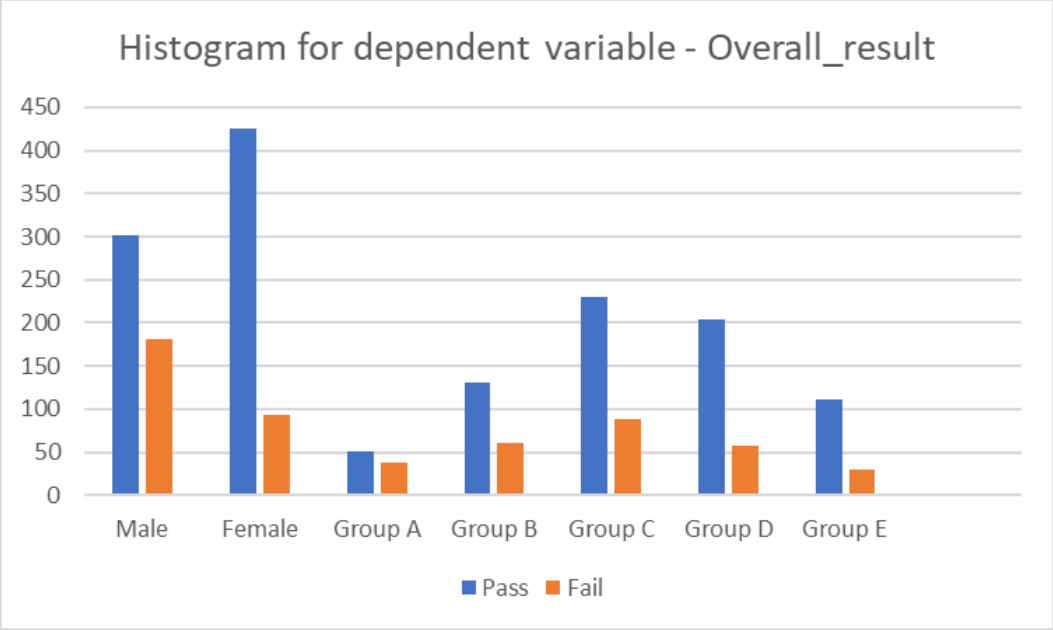
		Frequency of Fail: 35
4.	Group B	Frequency of Pass: 135 Frequency of Fail: 55
5.	Group C	Frequency of Pass: 244 Frequency of Fail: 75
6.	Group D	Frequency of Pass: 200 Frequency of Fail: 62
7.	Group E	Frequency of Pass: 113 Frequency of Fail: 27

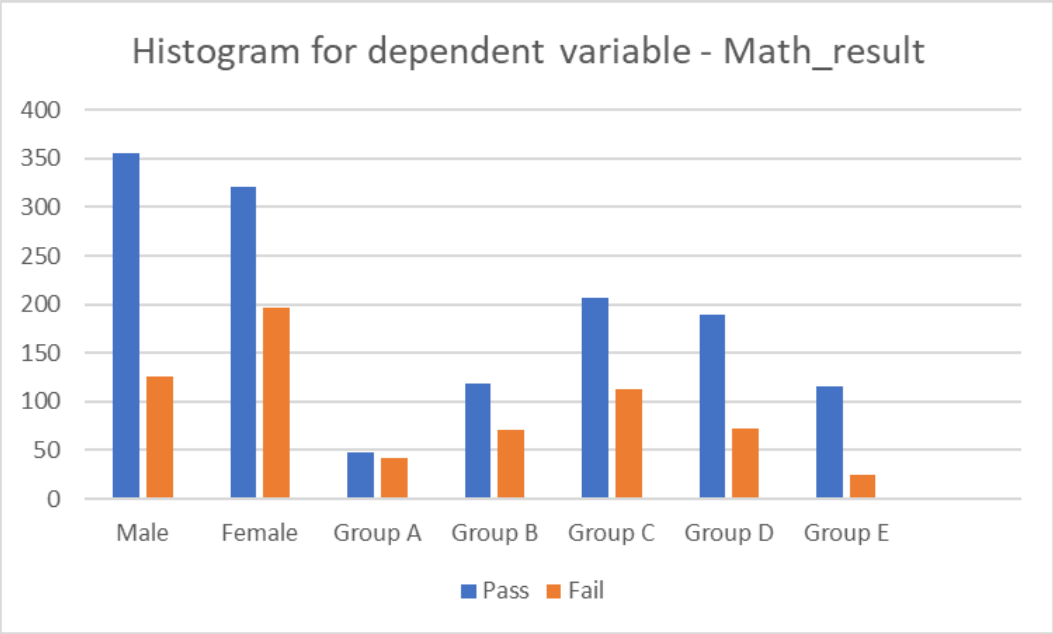
4. Dependent Variable - Writing_result

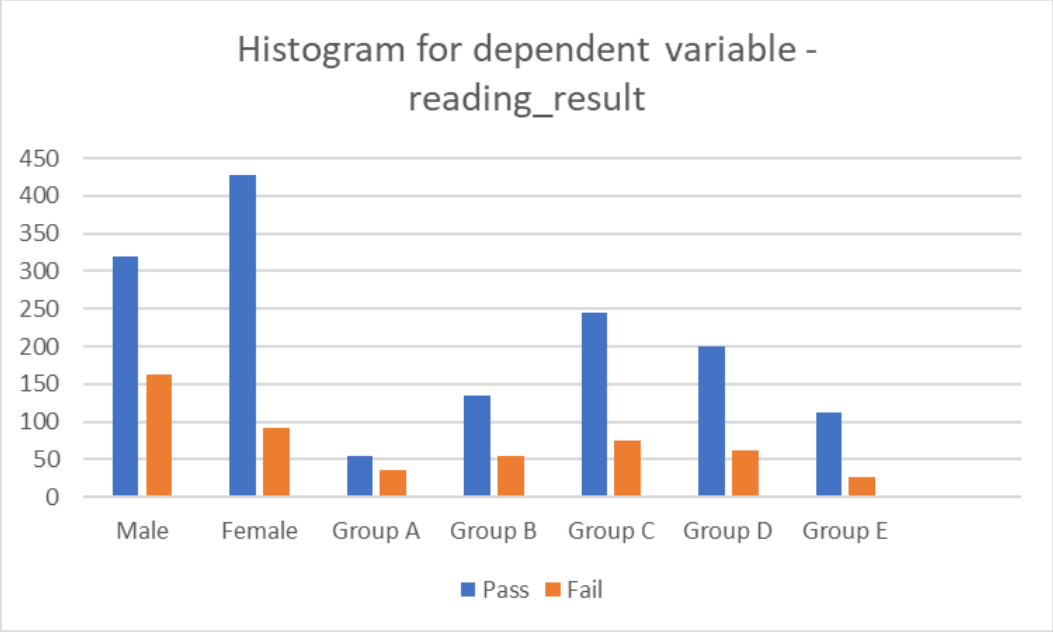
Sr. No.	Independent Variable	Dependent Variable – Writing_result
1.	Male	Frequency of Pass: 295 Frequency of Fail: 187
2.	Female	Frequency of Pass: 424 Frequency of Fail: 94
3.	Group A	Frequency of Pass: 52 Frequency of Fail: 37
4.	Group B	Frequency of Pass: 129 Frequency of Fail: 61
5.	Group C	Frequency of Pass: 225

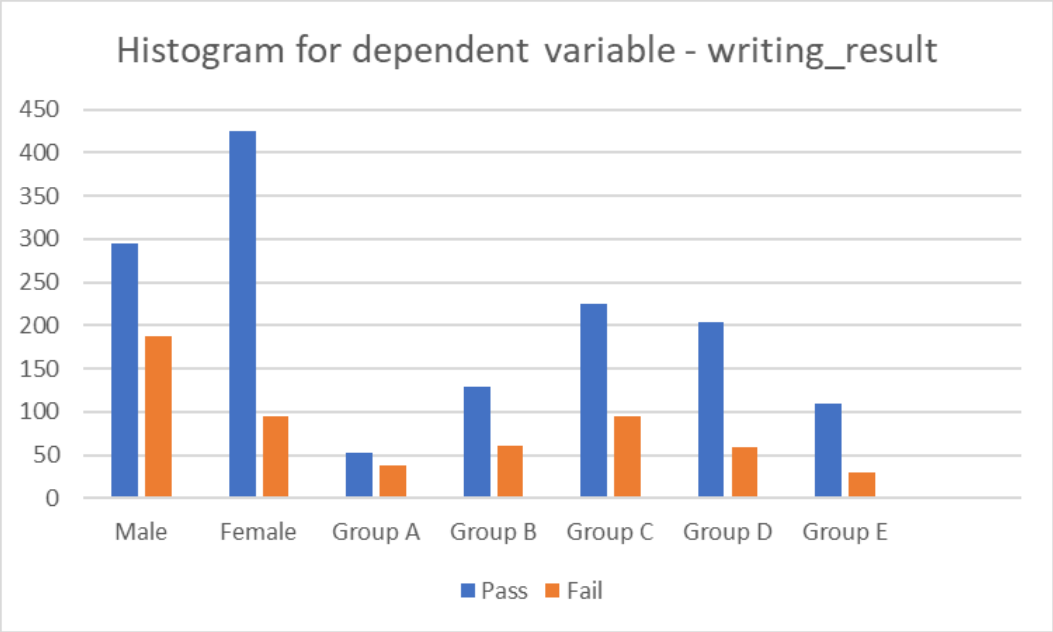
		Frequency of Fail: 94
6.	Group D	Frequency of Pass: 203 Frequency of Fail: 59
7.	Group E	Frequency of Pass: 110 Frequency of Fail: 30

(iv) Histograms derived from Step 2.4




Histogram for dependent variable - Math_result
400
350
300
250
200
150
100
50
0
Male
Female
Group A
Group B
Group C
Group D
Group E
Pass
Fail


Histogram for dependent variable - reading_result
450
400
350
300
250
200
150
100
50
0
Male
Female
Group A
Group B
Group C
Group D
Group E
Pass
Fail



3 STEP 3

(i) Provide the resulting code (can be as an additional .ipynb file if submitting a PDF)

(ii) Provide a table documenting the protected class variable selected, the privileged/unprivileged groups/values, the pre-processing bias mitigation function selected, and the fairness metrics/resulting values computed in Step 3.2 and Step 3.4

Pre-processing bias mitigation method - Disparate Impact Remover

Dependent Variables selected: Reading Result and Math Result

Bias Metrics - Disparate Impact Ratio and Statistical Parity Difference

For Protected Class Variable - Gender | **Dependent Variable** - Math Result

Gender - Priv/Un priv	**Dependent Value**	**Values**
Male - Privileged	Favorable	356
Female - Unprivileged	Favorable	321
Male - Privileged	Unfavorable	126
Female - Unprivileged	Unfavorable	197

DIR: 0.8390199991323586 | SPD: -11.88980919271377

For Protected Class Variable - Gender | **Dependent Variable** - Reading Result

Gender - Priv/Un priv	**Dependent Value**	**Values**
Male - Privileged	Favorable	319
Female - Unprivileged	Favorable	427
Male - Privileged	Unfavorable	163
Female - Unprivileged	Unfavorable	91

DIR: 1.2455307972549352 | SPD: 16.249859818324552

For Protected Class Variable - Race | **Dependent Variable** - Math Result

Race- Priv/Un priv	**Dependent Value**	**Values**
Privileged	Favorable	471
Unprivileged	Favorable	206
Privileged	Unfavorable	210
Unprivileged	Unfavorable	113

DIR: 0.9336900744763692 | SPD: -4.586193086876676

For Protected Class Variable - Race | **Dependent Variable** - Reading Result

Race- Priv/Un priv	**Dependent Value**	**Values**
Privileged	Favorable	502
Unprivileged	Favorable	244
Privileged	Unfavorable	179
Unprivileged	Unfavorable	75

DIR: 1.037630044086975 | SPD: 2.7739033967197315

4 STEP 4 OPTION A

(i) Provide the resulting code (can be as an additional .ipynb file if submitting a PDF)

(ii) Document 1) the privileged/unprivileged groups, 2) the dependent variable, 3) the quantitative results from applying the two fairness metrics on the classifier output associated with the original and transformed dataset, 4) a table documenting whether there was positive, negative, or no change in each of the fairness metrics after transforming the dataset, after training the classifier on the original dataset, and after training the classifier on the transformed dataset.

1. Privileged Group - Gender(Male)
2. Unpriviled Group - Gender(Female)
3. Dependent Variable - 'overall_result'
4. 4) & 5) (Combined Answer) Results from from applying two fairness metrics on the classifier output and the table representing Change
    a. DIR

	DIR Value	Change
Original Dataset	1.3138316294462473	N/A
Transformed Dataset	1.199524597483781	Negative
Original Testing Dataset(From Classifier)	1.2660256410256412	N/A
Transformed Testing Dataset(From Classifier)	1.5437198067632851	Positive

b. SPD

	DIR Value	Change
Original Dataset	19.59819926624907	N/A
Transformed Dataset	12.170172543616523	Negative
Original Testing Dataset(From Classifier)	17.510548523206765	N/A
Transformed Testing Dataset(From Classifier)	29.47878470403353	Positive

5 STEP 5

(i) List the members of your project team

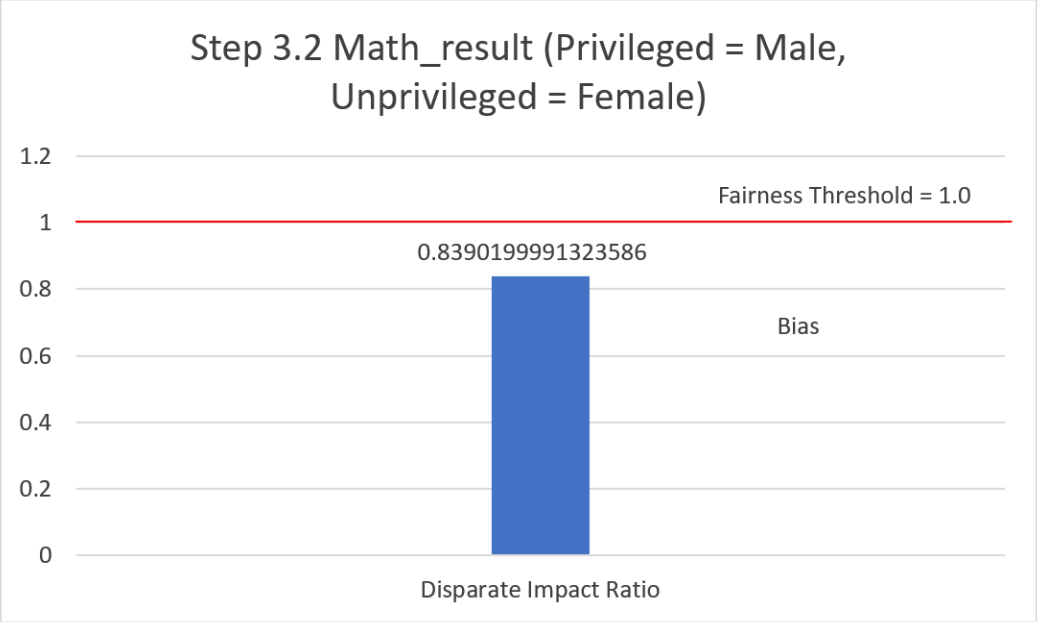
Abhijith Chakiat, Amisha Buch, Nirali Thakkar

(ii) Graph the results from applying the two fairness metrics on your privileged/unprivileged groups as derived from Step 3.2, 3.4, and 4.5

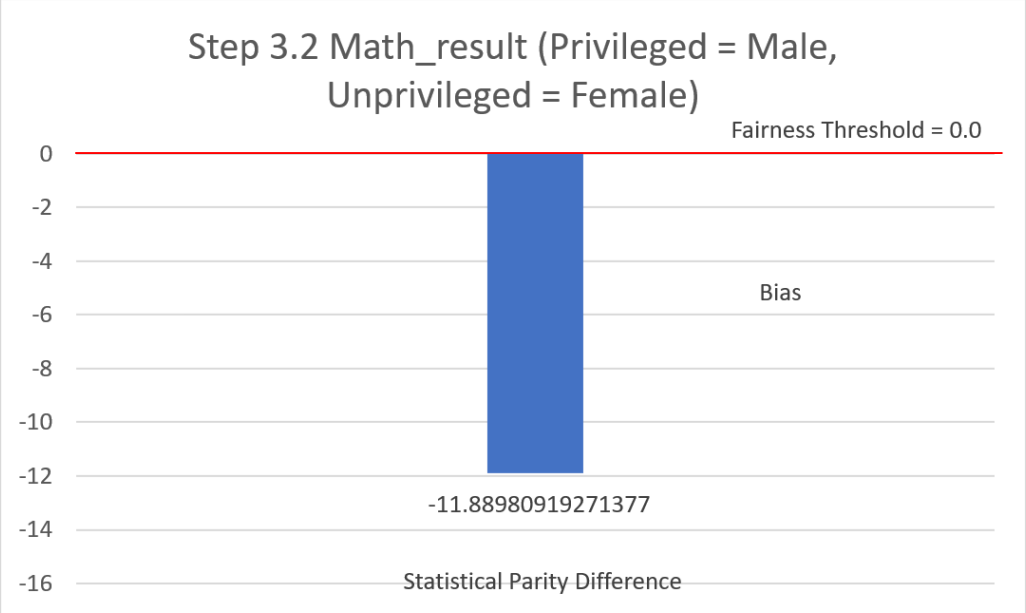
Step 3.2

Gender - Math_result

Disparate Impact Ratio

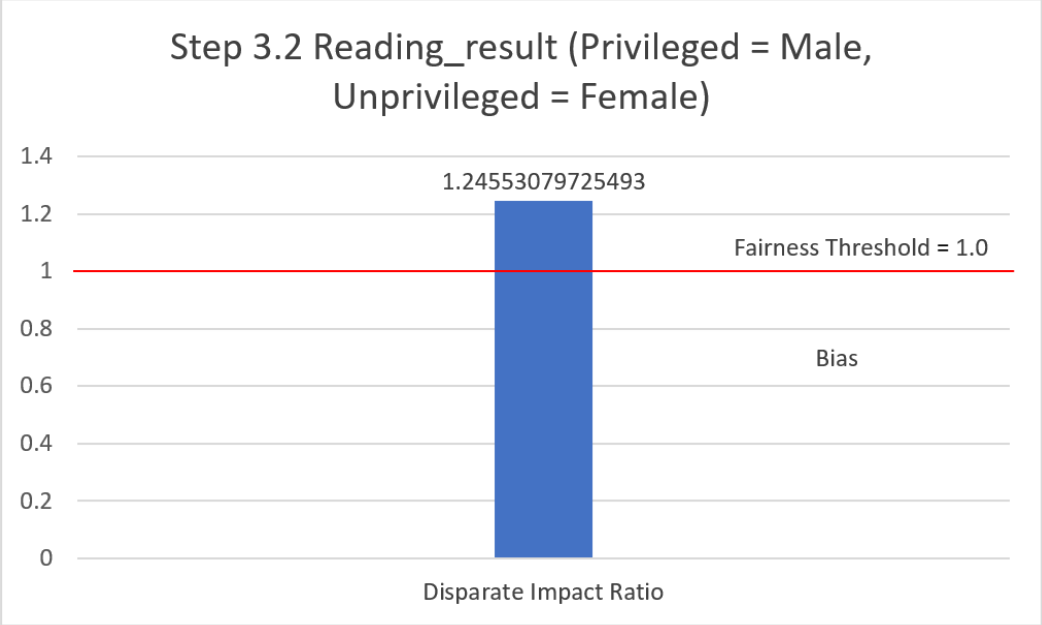


Statistical Parity Difference

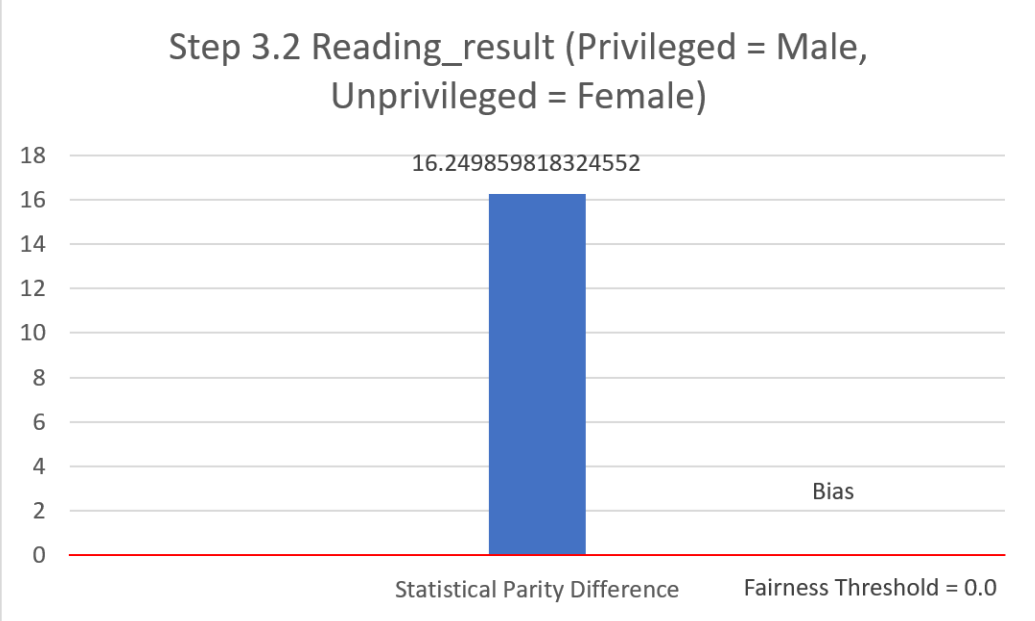


Gender - Reading_result

Disparate Impact Ratio

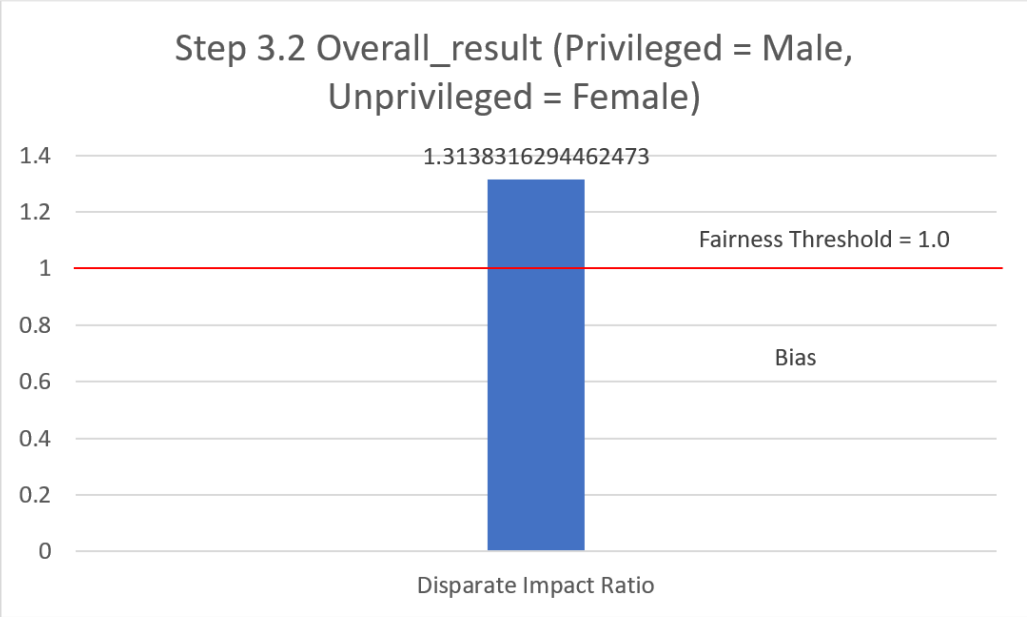


Statistical Parity Difference

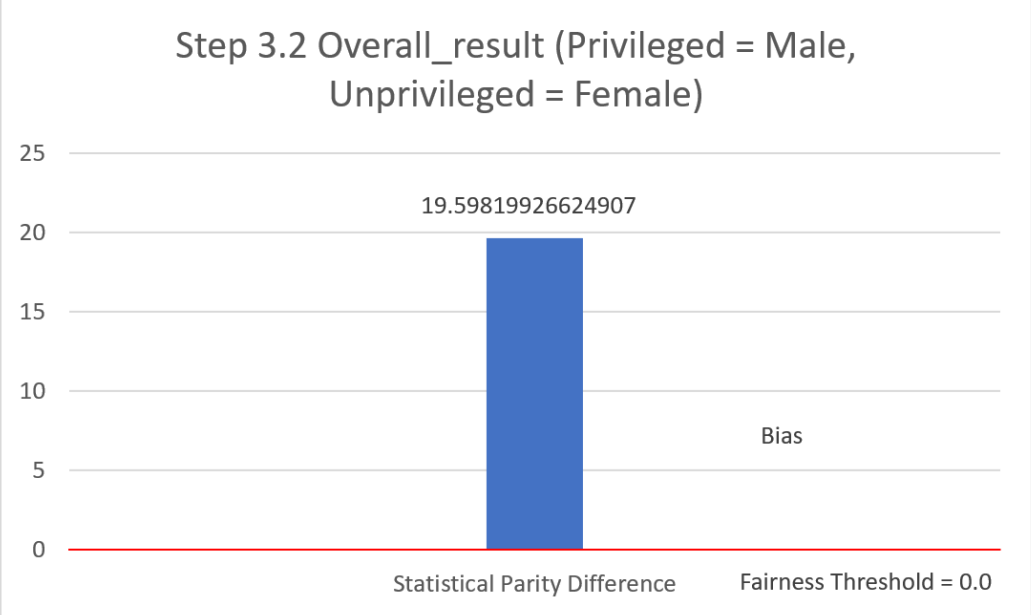


Gender - Overall_result

Disparate Impact Ratio

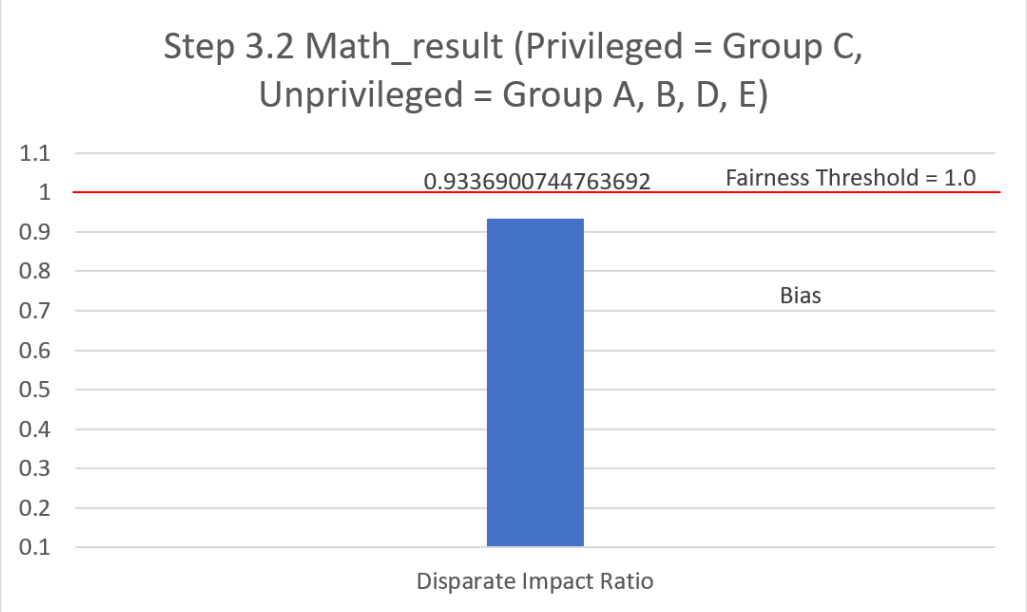


Statistical Parity Difference

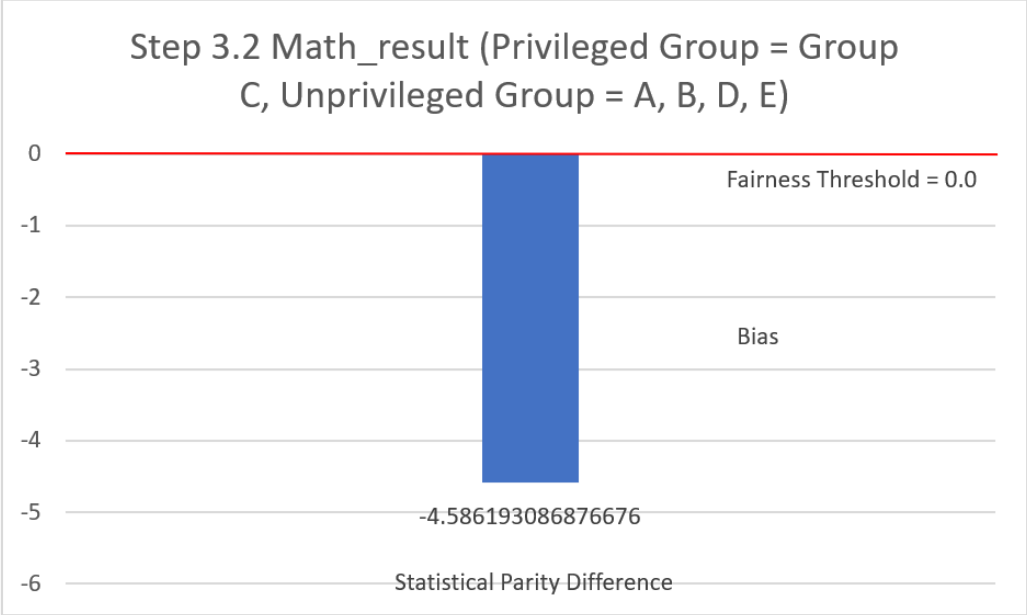


Race - Math_result

Disparate Impact Ratio

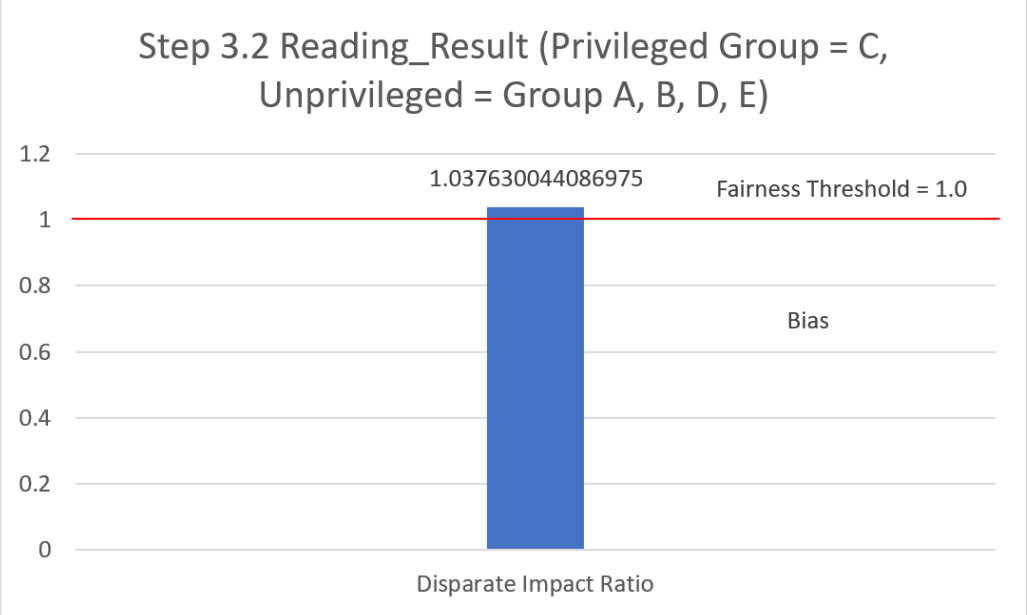


Statistical Parity Difference

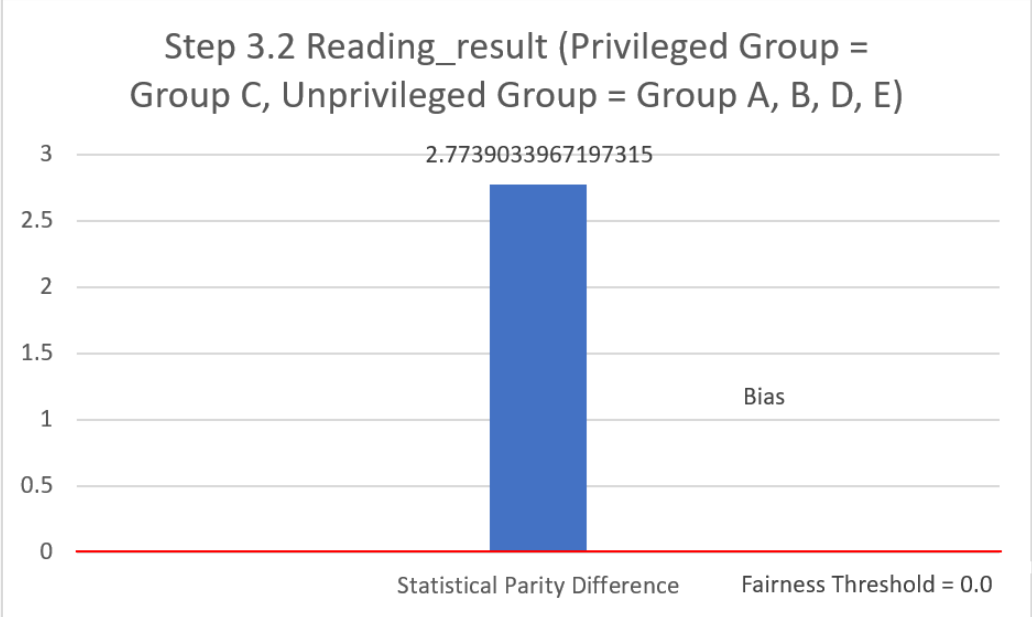


Race - Reading_result

Disparate Impact Ratio



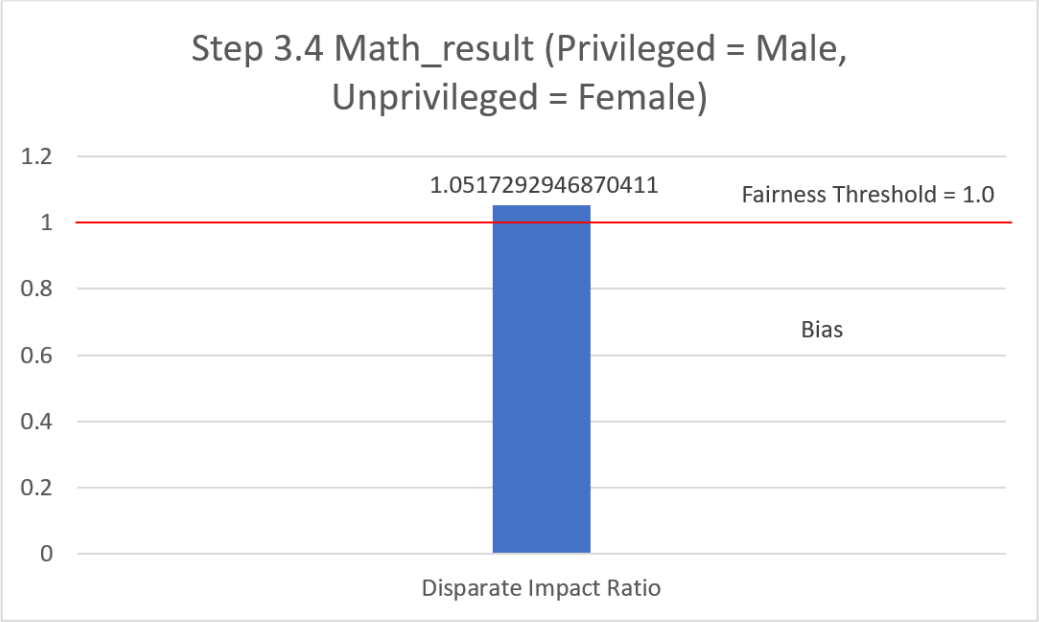
Statistical Parity Difference



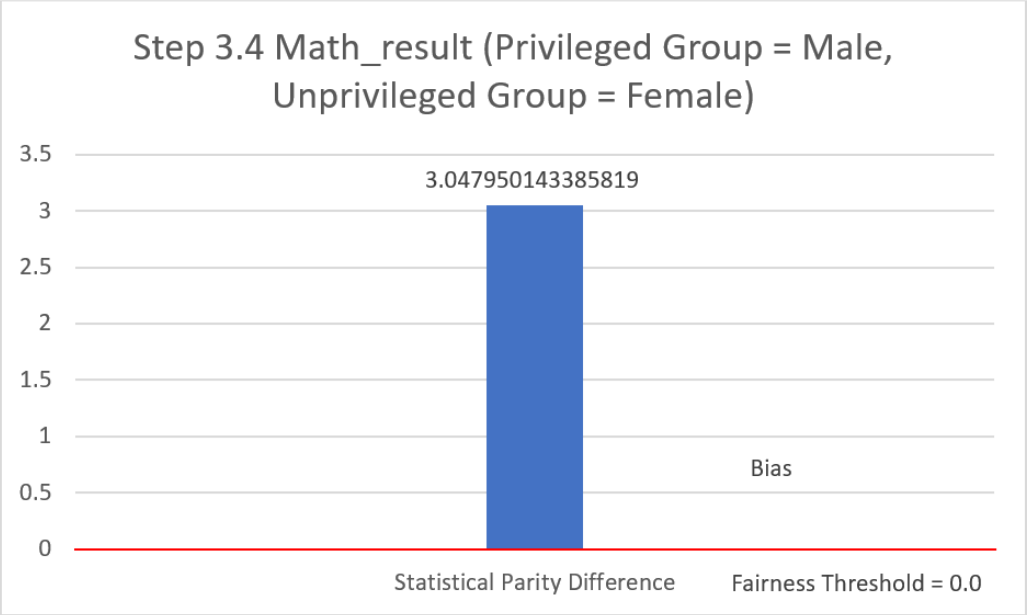
Step 3.4

Gender - Math_result

Disparate Impact Ratio

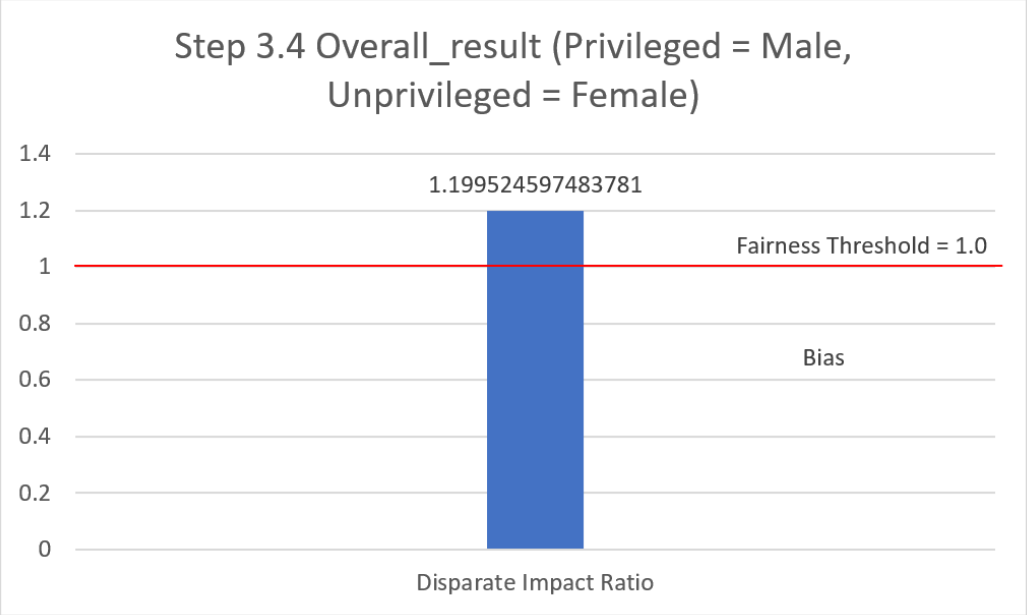


Statistical Parity Difference

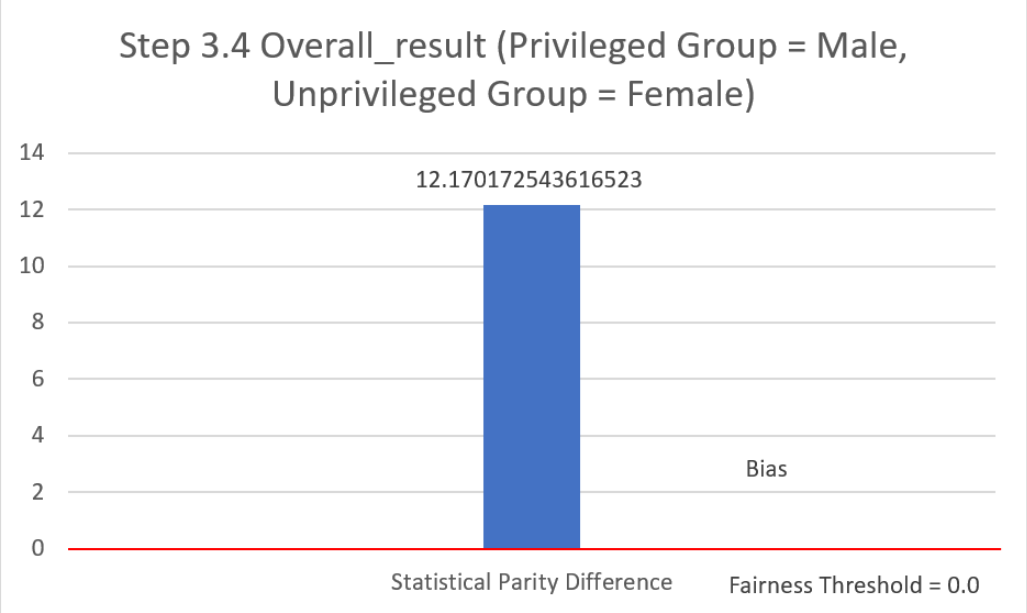


Gender - Overall_result

Disparate Impact Ratio



Statistical Parity Difference

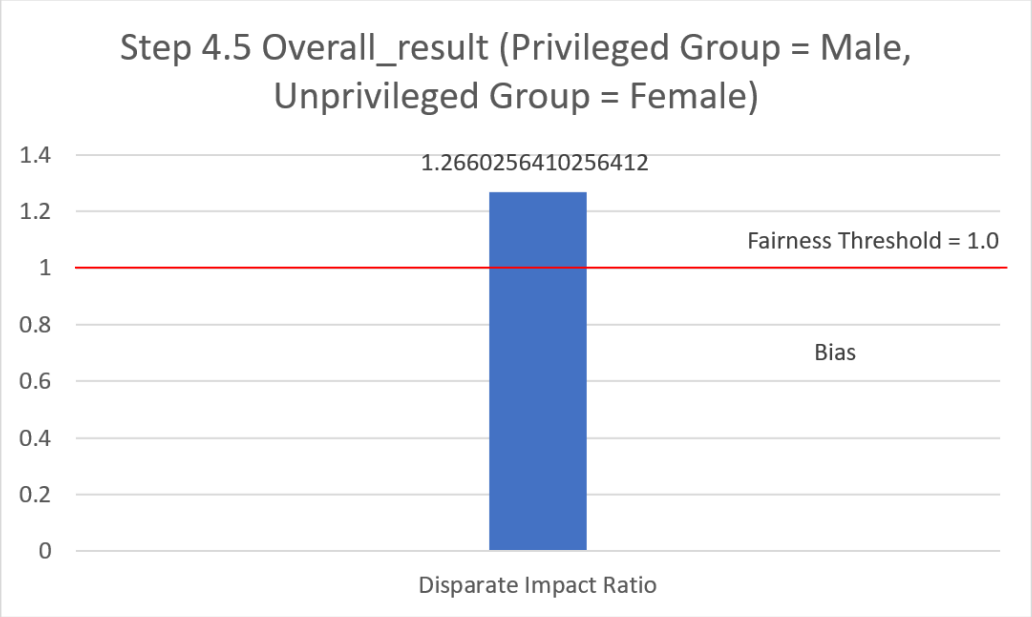


Step 4.5

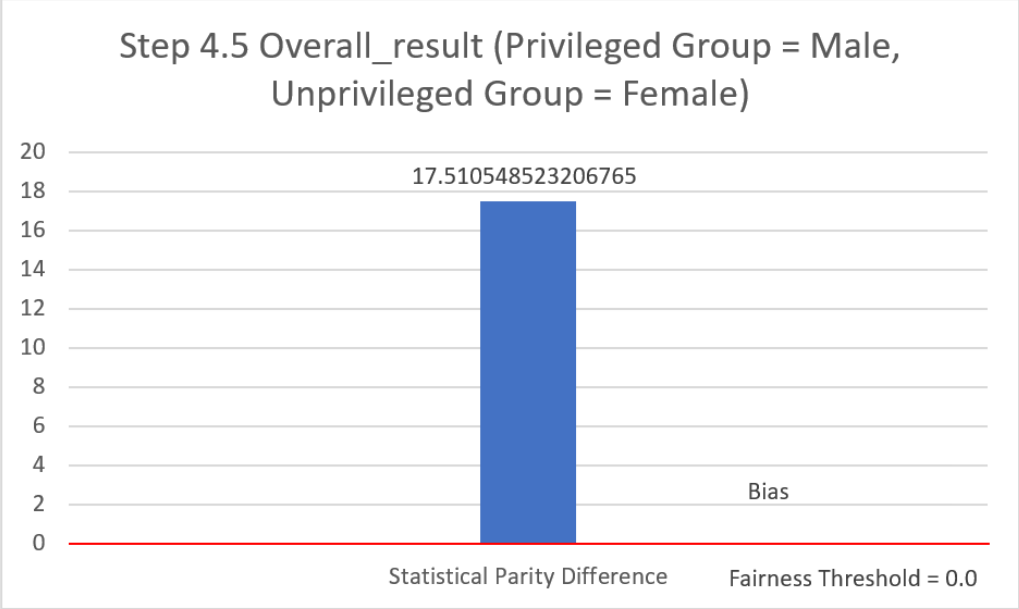
Gender - Overall_result

Classifier Output associated with the original testing dataset

Disparate Impact Ratio

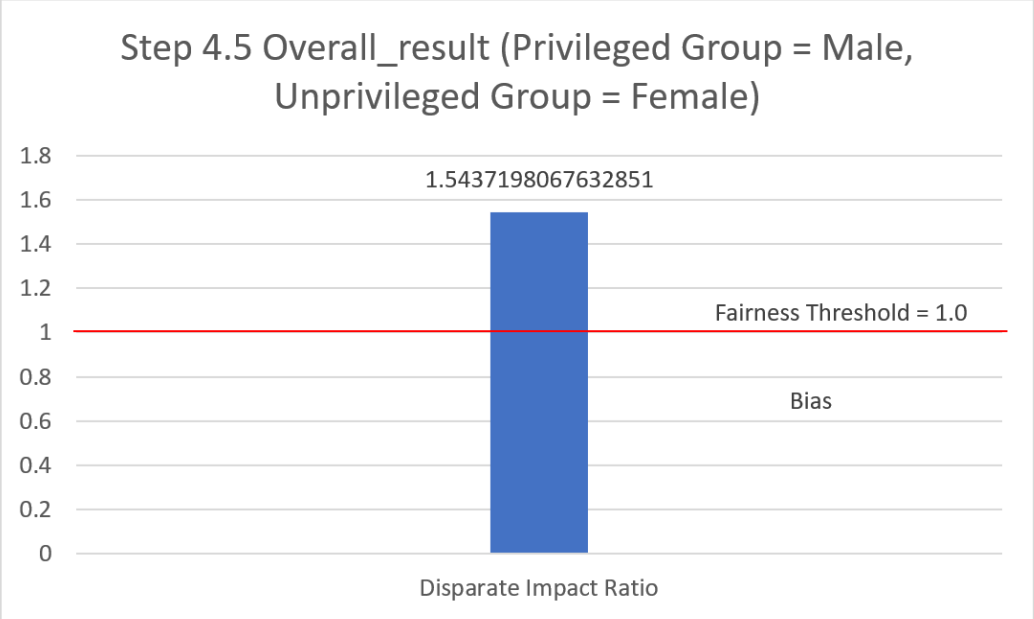


Statistical Parity Difference

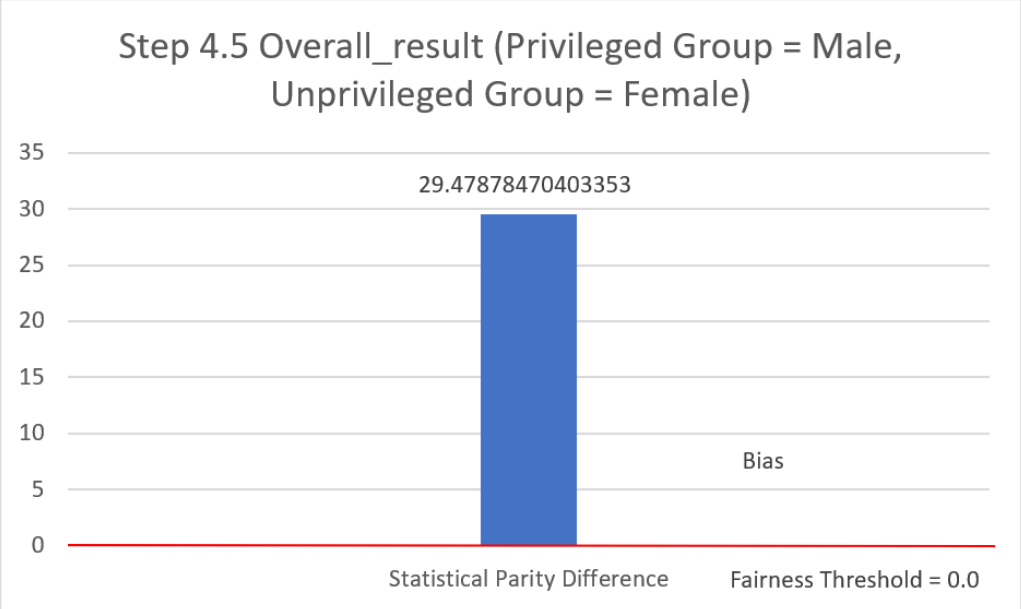


Classifer Output associated with the transformed testing dataset

Disparate Impact Ratio



Statistical Parity Difference



(iii) Explain which fairness metric (if any) is best and provide a justification for your answer

DIR comes out to be a better fairness metric in case of this dataset. This is because even with the original, transformed and classified datasets, the DIR value still remains closer to the threshold value than SPD. SPD shows a very high variance once classification is applied, which doesnot represent the data and the biases accurately.

(iv) Individual Answers- Did any of these approaches seem to work to mitigate bias (or increase fairness)? Explain your reasoning. Did any group receive a positive advantage? Was any group disadvantaged by these approaches? What issues would arise if you used these methods to mitigate bias?

Abhijith Chakiat's Answer

We did see a bias in the dataset that favored the privileged groups with respect to the dependent variables that were analyzed. We looked at two metrics to test this - Disparate Impact Ratio and Statistical Parity Difference. Both these metrics flagged the bias in the dataset that we reported. On application of the Disparate Impact Remover method, there was a change in the metrics that showed a decrease in the bias. The unprivileged group based on the dependent variable of gender received an advantage due to this, no disadvantage to the classes were seen. Since no class was at a disadvantage, in the context of this dataset and metrics, there are no issues that should arise with this method.

Amisha Buch's Answer

The Disparate Impact Remover bias mitigation method did seem to work in reducing the bias in the system. According to the definition, DIR is "a

pre-processing technique that edits values, which will be used as features, to increase fairness between the groups". This is precisely what it does. It balances out the DIR and SPD (Statistical Parity Difference) values, hence reducing bias. The unprivileged group (in terms of protected variable gender) received an advantage due to this. No group received a disadvantage because of this approach.

Though I would like to mention that the bias between the privileged and unprivileged groups was very less to start with. But DIR still helped in reducing the bias.

I would like to add that we also tried Logistic Regression to try and test if it reduces bias. But it clearly showed that it did not help in reducing bias or increasing fairness. It may be possible that this kind of classifier is not suited for this dataset. In order to know what works best, different classifiers can be applied and results can be compared to determine if classifying can actually reduce bias.

Nirali Thakkar's Answer

The two fairness metrics used to decrease the bias in the dataset were Disparate Impact Remover (DIR) and Statistical Parity Difference (SPD). Both these metrics were efficient in decreasing the bias. This is because DIR balances the important attributes and their correlated features to equalize the probability values between the privileged and unprivileged groups while SPD balances the differences in these values to equalize the probability values between the privileged and unprivileged groups. Both these algorithms were efficient in almost equalizing the probability values between the privileged and unprivileged groups.

The DIR metric proved to be more near to the threshold value 1 than the SPD value was near its threshold 0. Therefore, DIR was slightly more better in decreasing the bias. The bias was in favor of the privileged group; once the fairness metrics were applied, this bias decreased, thus making it advantageous for the unprivileged group. No group was particularly disadvantaged by these approaches.

No issue has arised due to the metrics applied; however, it is better to increase the threshold range to provide a better balance. That is, instead of keeping an ideal threshold of 1 (DIR) or 0 (SPD), a range can be kept such as 0.9 to 1.1 (for DIR) or -0.1 to + 0.1 (for SPD).