HomeWork 3 Report

Fairness and Bias Analysis in Machine Learning

NAME: Kavya Gengedulla Yeranarasappa
USC ID: 3206274199

1) Finding Racial Biases using ANOVA test

After loading and preprocessing of data, mean and median values for all traits among the given races were calculated and presented in the notebook, the analysis of ANOVA test demonstrates significant differences in trait perceptions across 3 racial groups white, black, east asian. Traits like "Egotistic," "Intelligent," and "Responsible" exhibit substantial variations among groups, with p-values well below 0.05 (e.g., p < 1.44e-07 for "Egotistic"). Similarly, traits like "Kind," "Trustworthy," and "Caring" also display significant differences, although with slightly higher p-values (e.g., p < 0.001 for "Caring"). These findings highlight the presence of racial biases in trait evaluations.

Below is the list of p-values after ANOVA test for all traits:

```
Significant biases :
Trait: Egotistic, p-value: 1.4412211501758323e-07
Trait: Intelligent, p-value: 1.6711926732092035e-15
Trait: Kind, p-value: 0.00012878104123732718
Trait: Responsible, p-value: 8.005057161350754e-11
Trait: Trustworthy, p-value: 3.672594877204227e-06
Trait: Aggressive, p-value: 2.7078370672104353e-11
Trait: Caring, p-value: 0.00041893692225684995
Trait: Emotional, p-value: 0.02319487807768113
Trait: Friendly, p-value: 4.9774464052824096e-05
Trait: Sociable, p-value: 0.03449616164486741
```

2) Finding Gender Biases using t-test

Later, mean and median values were calculated for all the given traits for different genders (female and male) and The t-test analysis uncovers significant gender-based differences in trait perceptions across all evaluated traits. For

instance, traits like "Confident" (p = 0.00025), "Egotistic" (p = 5.72e-103), and "Intelligent" (p = 6.62e-10) exhibit substantial variations. These p-values suggest strong evidence of gender-based biases in the given trait perceptions.

Below is the list of p-values after T-test for all traits:

```
Significant biases for gender:
Trait: Confident, p-value: 0.0002490671279198179
Trait: Egotistic, p-value: 5.724535735483598e-103
Trait: Intelligent, p-value: 6.617541566954897e-10
Trait: Kind, p-value: 3.6772754632964984e-84
Trait: Responsible, p-value: 5.314276336798082e-34
Trait: Trustworthy, p-value: 5.706506511591532e-73
Trait: Aggressive, p-value: 6.29311355462135e-101
Trait: Caring, p-value: 6.745136923994714e-105
Trait: Emotional, p-value: 1.029377924430242e-164
Trait: Friendly, p-value: 3.7108186188491455e-64
Trait: Sociable, p-value: 3.1946748840509533e-62
```

3) Top 5 most significant biases from both race and gender

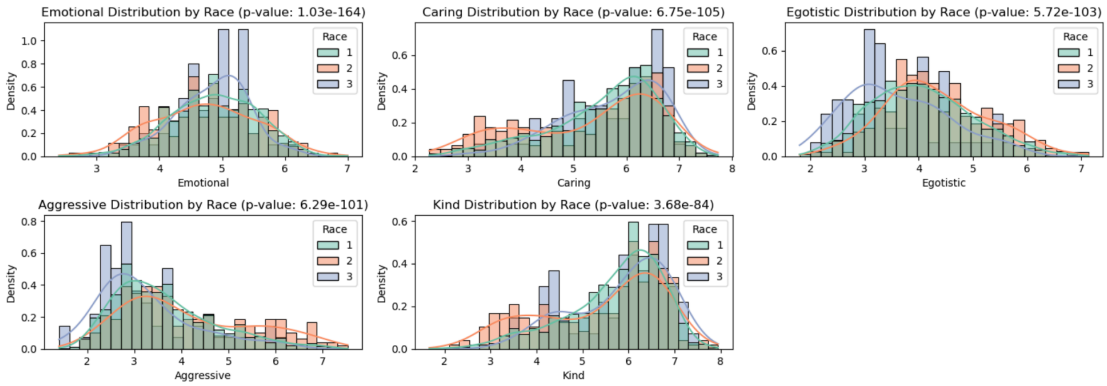
The top five most significant biases across race and gender include traits such as "Emotional," "Caring," "Egotistic," "Aggressive," and "Kind." These traits stand out with exceptionally low p-values, indicating strong biases.

```
Top 5 Most Significant Biases (Combined Race and Gender):
Trait: Emotional, p-value: 1.029377924430242e-164
Trait: Caring, p-value: 6.745136923994714e-105
Trait: Egotistic, p-value: 5.724535735483598e-103
Trait: Aggressive, p-value: 6.29311355462135e-101
Trait: Kind, p-value: 3.6772754632964984e-84
```

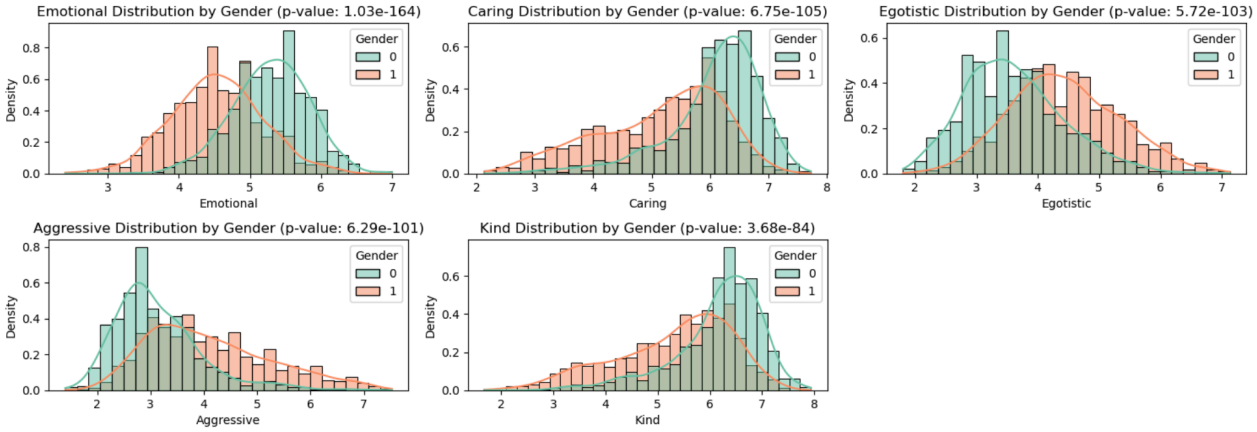
4) Histograms for the most significant biases

Histograms were calculated for both gender and race groups using matplotlib.The histograms below show how people from different genders and races perceive certain traits.

i) Histogram for race:



ii) Histogram for gender:



5) Finding possible adversities using the four-fifth rule

Using the "four-fifths rule.", a fairness problem in the hiring process based on both race and gender was found. It compares how many people from each group got hired. If a group's rate is less than four-fifths of the highest rate, it's flagged for potential discrimination.For **race, group 2 - Black** and for **gender, group 0 - Female** had a selection rate of **0.12**, which is below the four-fifths threshold causing fairness problems.

```
Adverse impact detected for the following race group(s):
- Group 2: Selection rate = 0.12

Adverse impact detected for the following gender group(s):
- Group 0: Selection rate = 0.12
```