STATISTICAL INFERENCE FOR INDIVIDUAL FAIRNESS

Presented by: Sathvick Sudarsan & Vineeth Ajith John

INTRODUCTION

OVERVIEW

This project delves into the critical issue of gender bias in machine learning, employing the UCI Adult Dataset to examine and mitigate discriminatory patterns in gender

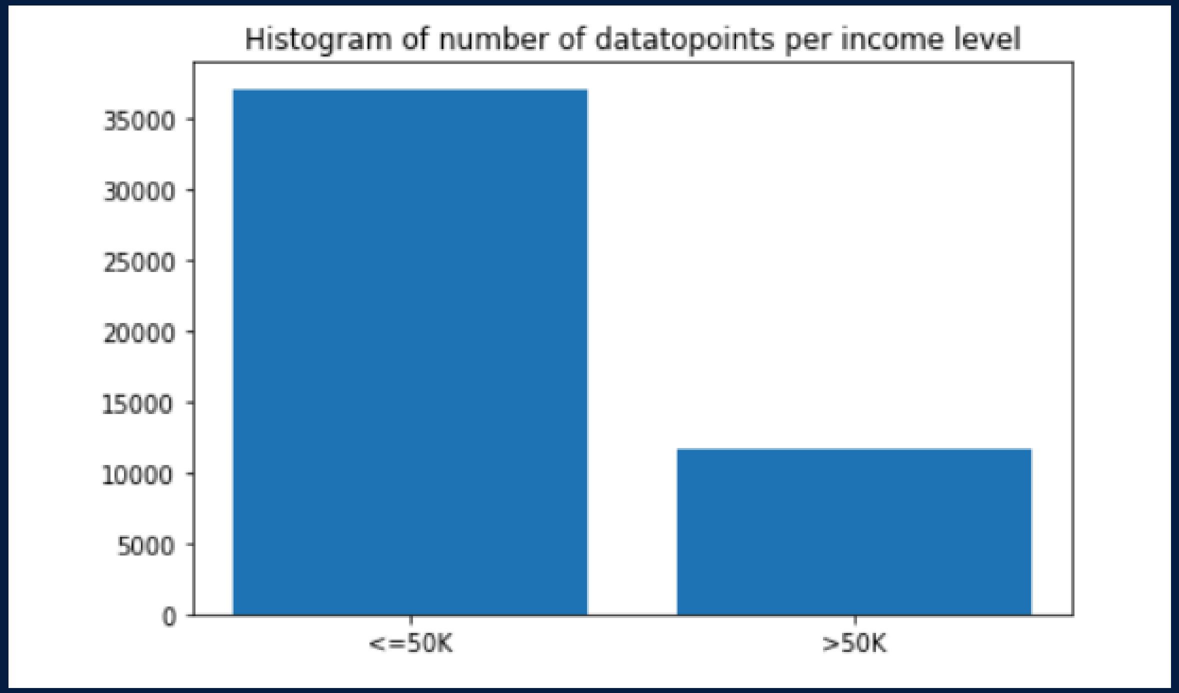
DATASET

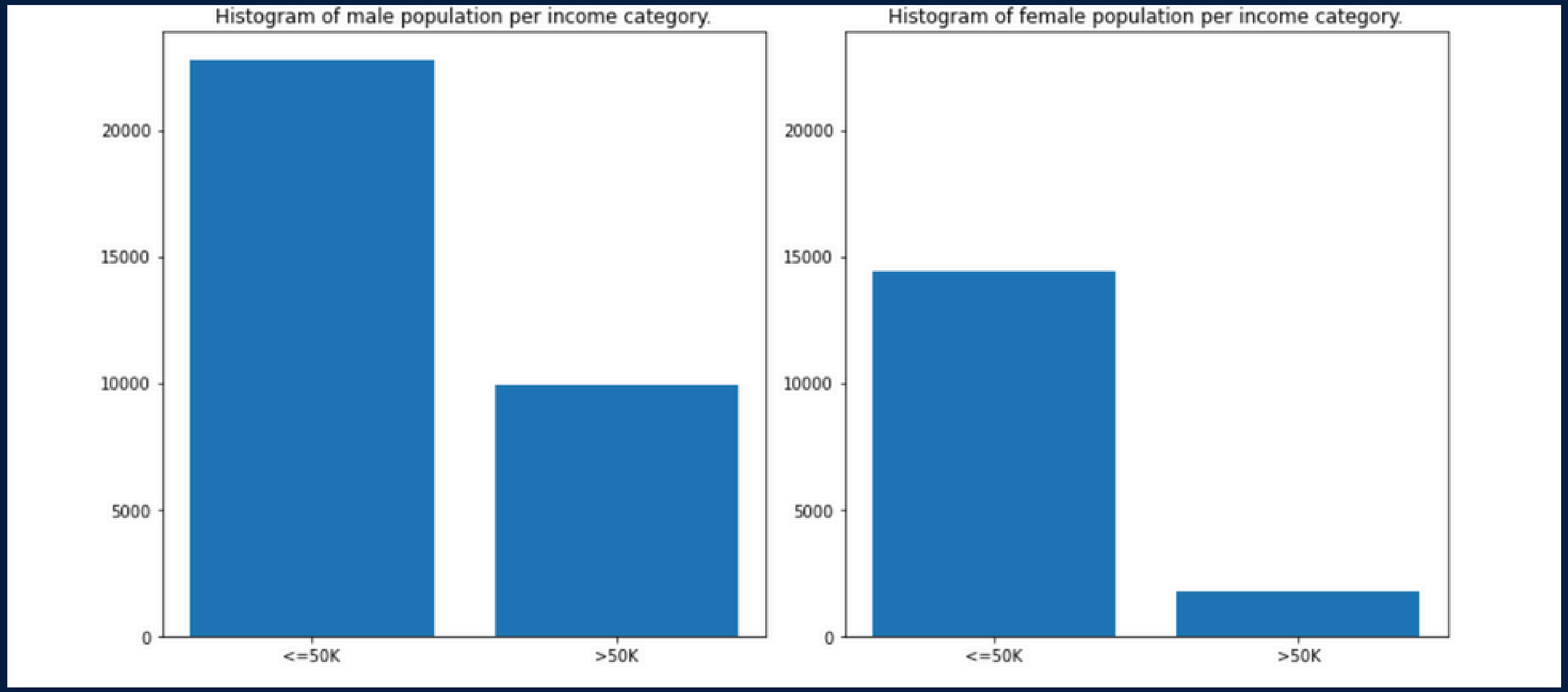
The UCI Adult dataset, commonly referred to as the "Census Income" dataset, is a widely recognized dataset in the field of machine learning, utilized for predicting whether an individual earns more than $50,000 per year based on census data. It contains 48,842 entries, each with 14 attributes that include demographic information such as age, work class, education, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours worked per week, and country of origin.

DATA PRE-PROCESSING

- Converted native country to binary one-hot for US vs non-US
- Converted sex and salary to binary one-hot
- Changed marital status to single or couple
- Converted relationships to one-hot
- Converted race to one-hot
- Transformed work-class feature
- Converted occupation to one-hot

DATA EXPLORATION





MLP CLASSIFIER

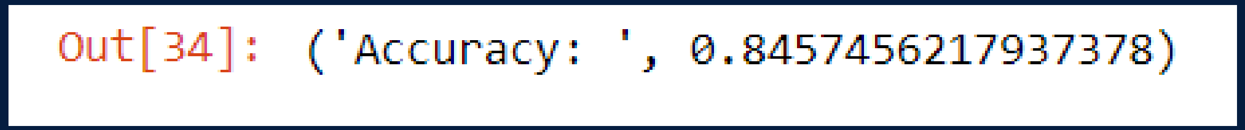
- we employ a Multi-layer Perceptron (MLP) classifier to predict salary categories based on the UCI Adult dataset.

- The MLP is a type of feedforward artificial neural network that maps sets of input data onto a set of appropriate outputs.

- The positive category refers to the high-income category: >50k a year and negative category refers to the low income category : <50k a year.

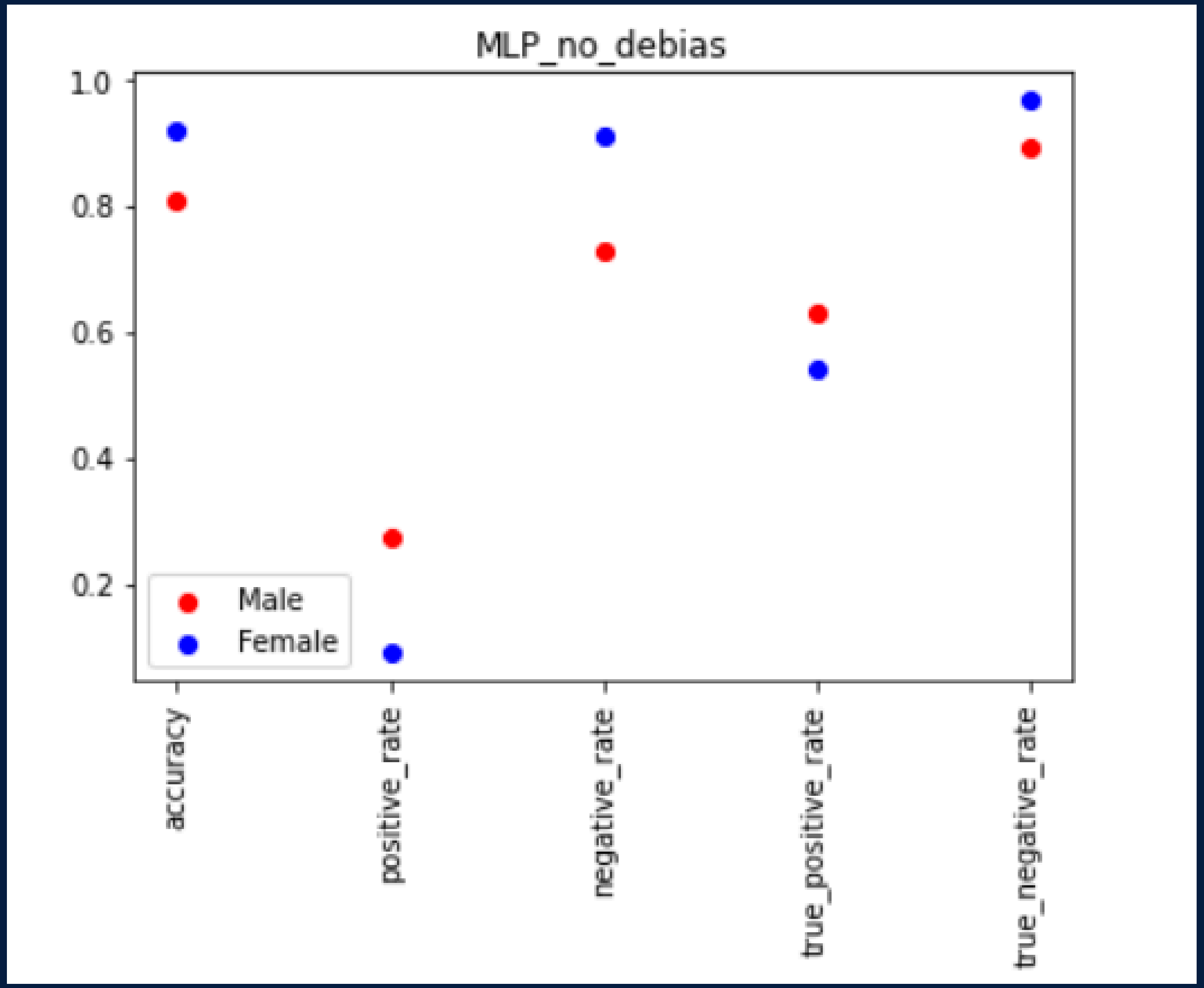
- The dataset is first shuffled and split into training and test sets, with 75% of the data used for training and the remaining 25% for testing.

- The MLP classifier is then trained on the training set with a maximum iteration limit set to ensure convergence.



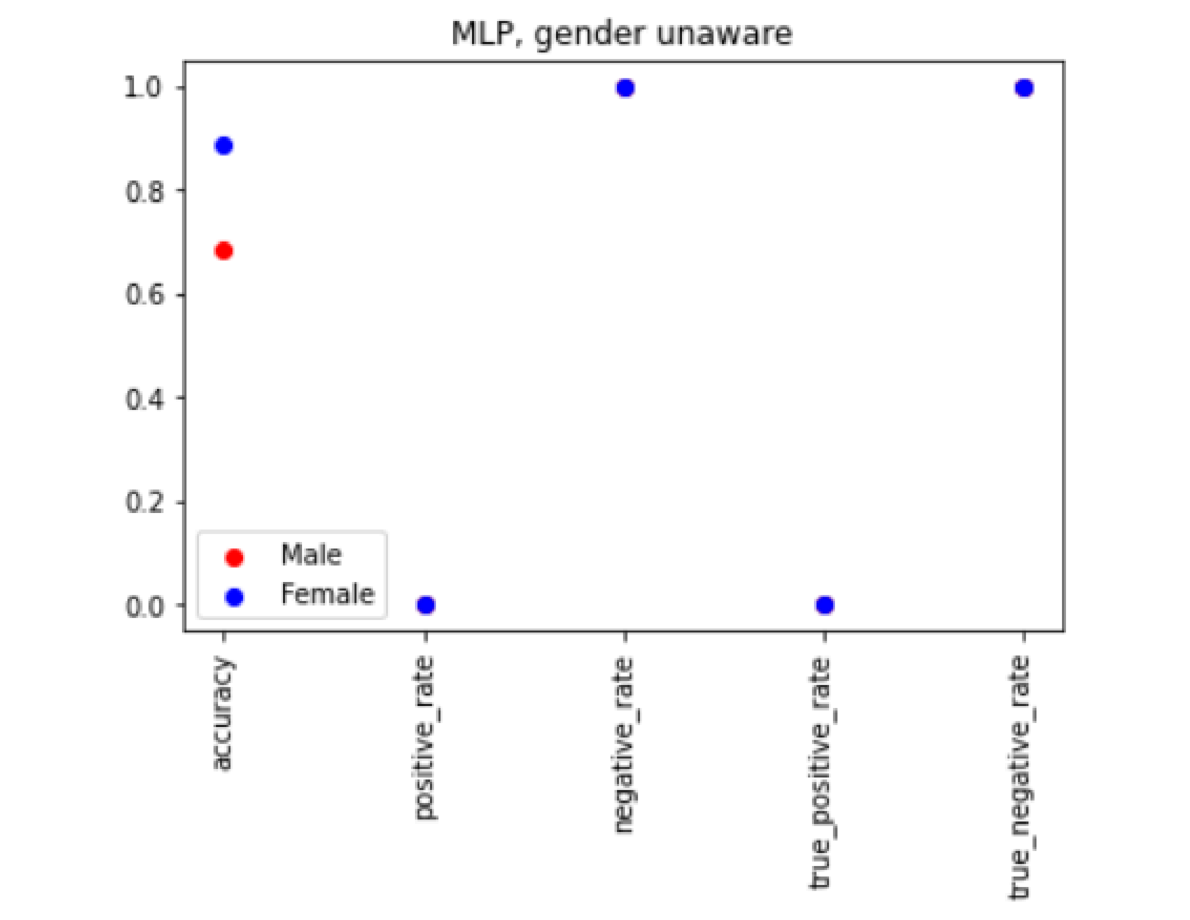
```
Out[34]:  ('Accuracy: ', 0.8457456217937378)
```

GENDER BIAS IN PREDICTION



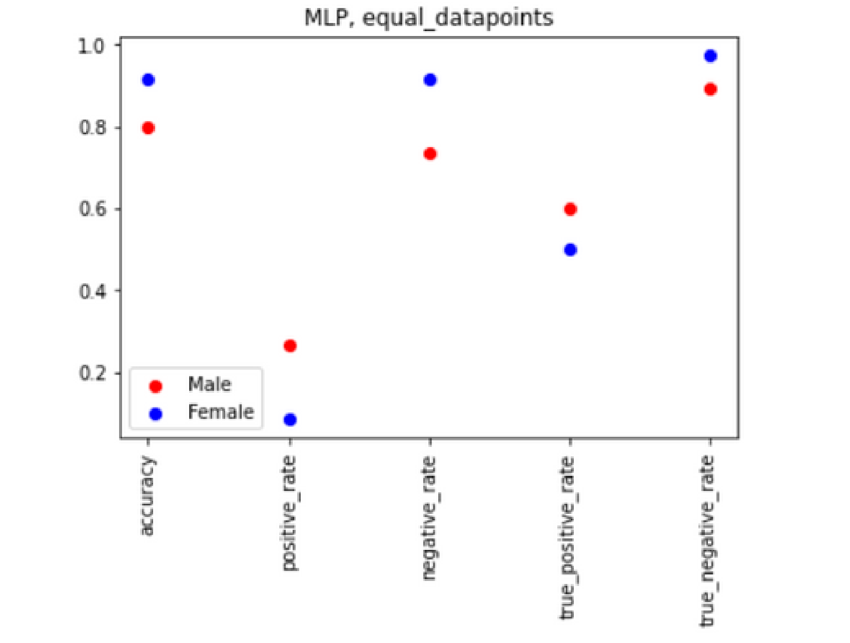
BIAS MITIGATION

Through Unawareness

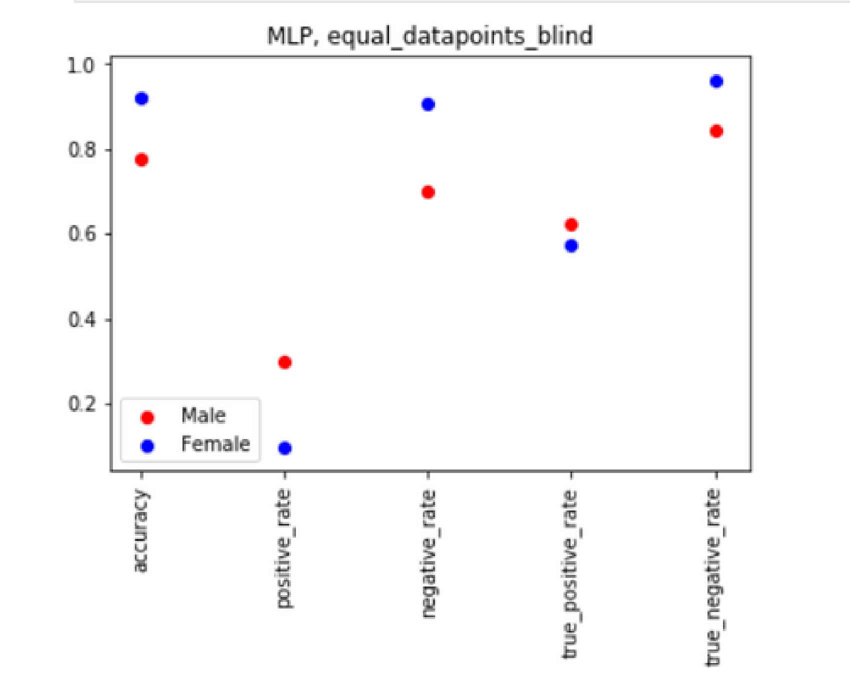


Through Dataset Balancing

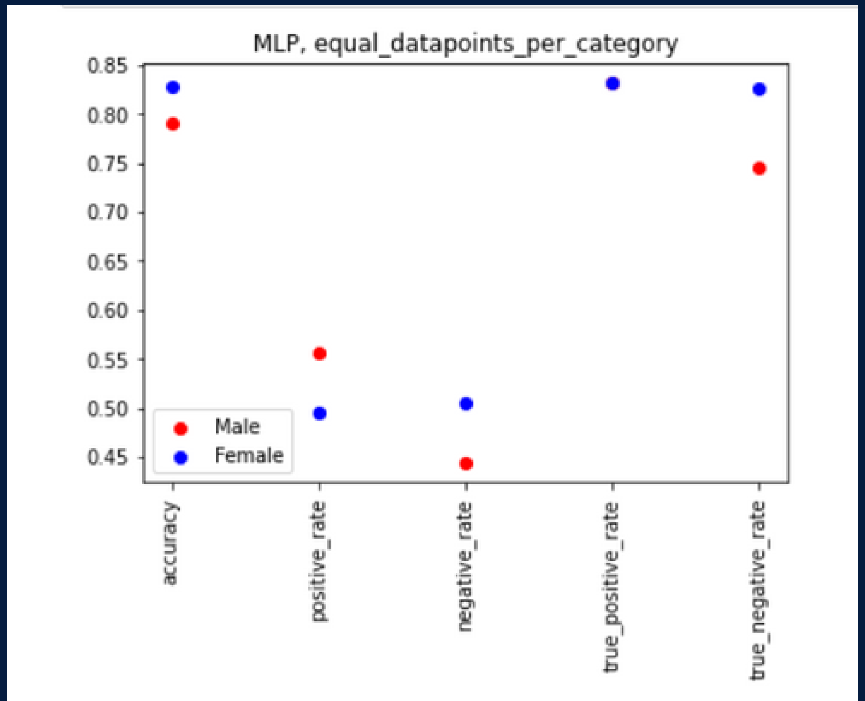
Equal number of datapoints per demographic



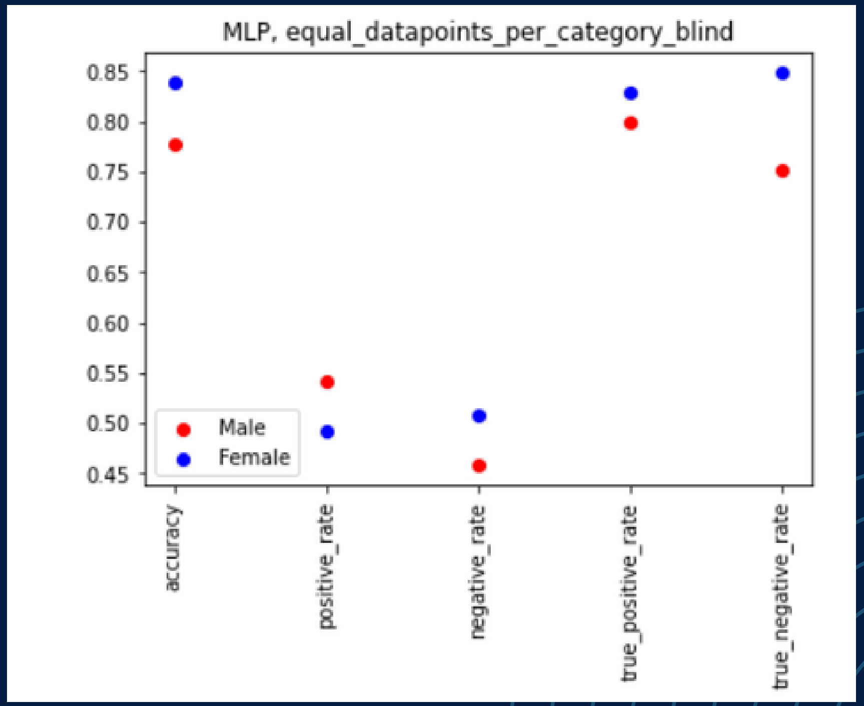
(unaware)



Equal number of datapoints per demographic in each category



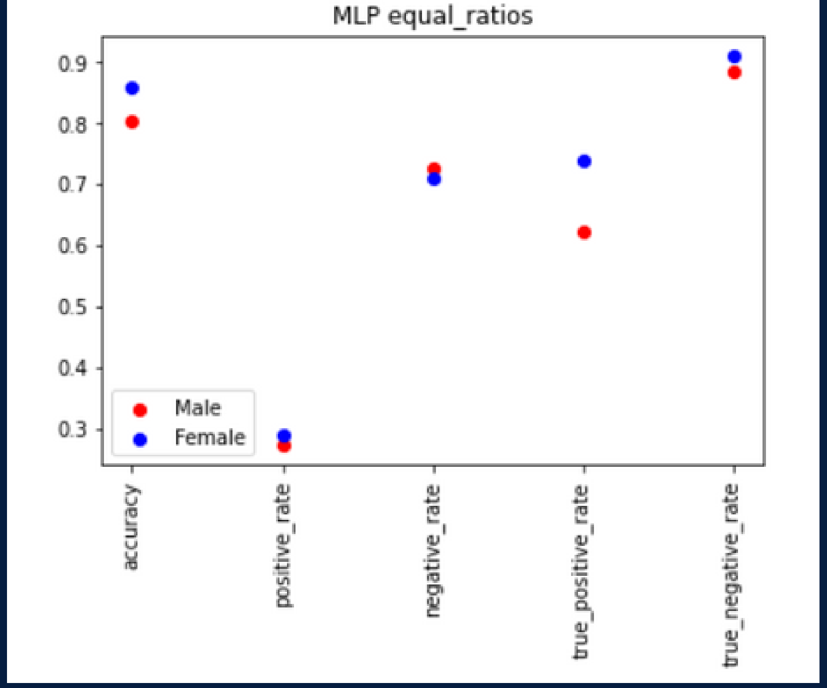
(unaware)



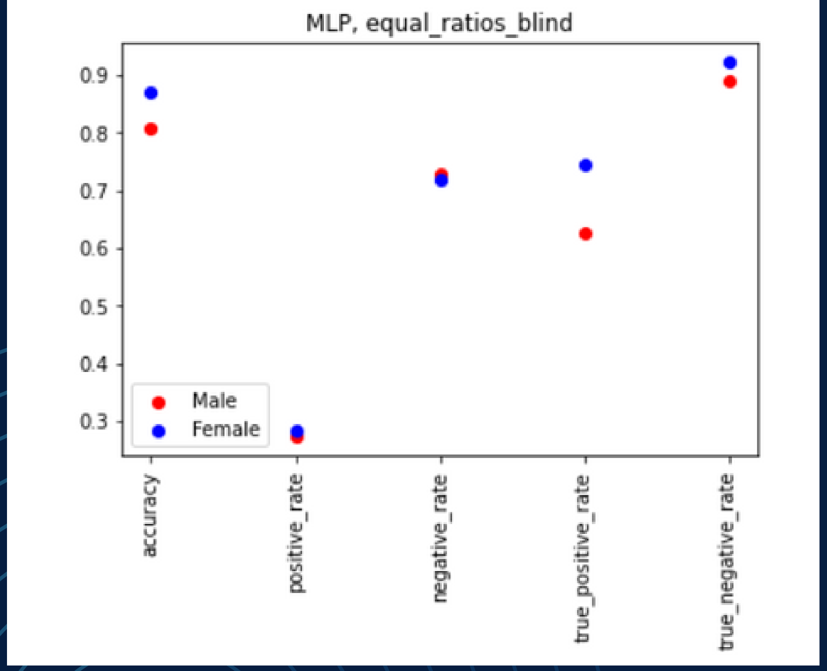
BIAS MITIGATION (contd.)

Through Dataset Balancing

Equal ratios instead of equal number of datapoints

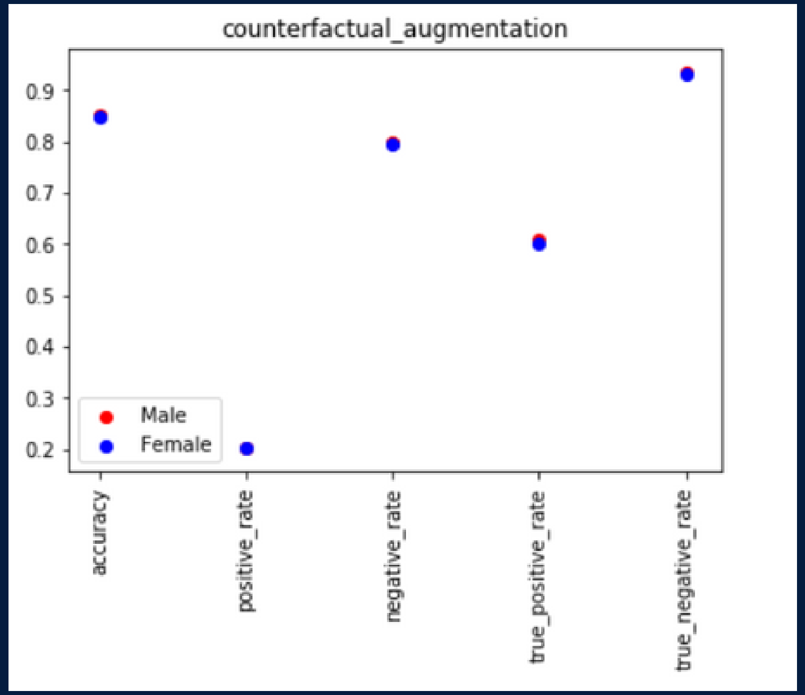


(unaware)

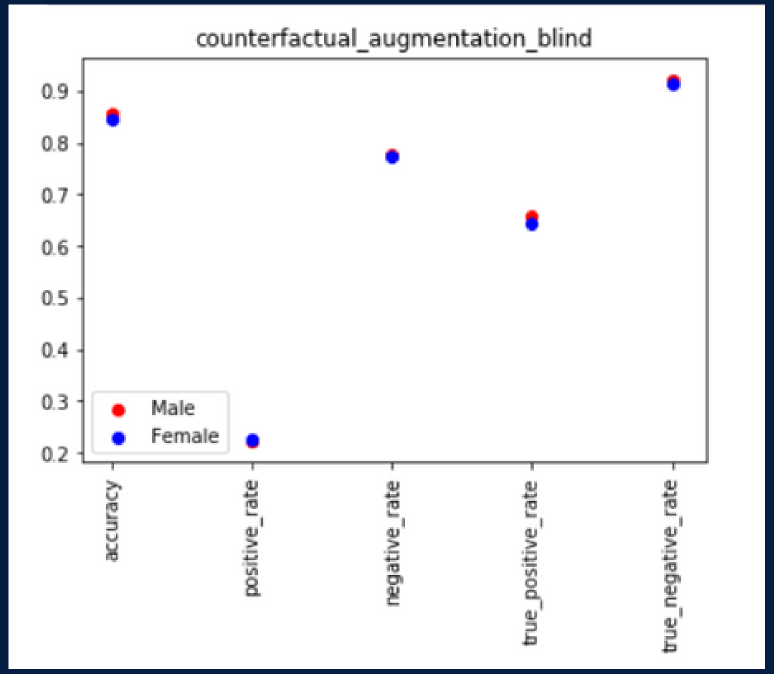


Through data augmentation

Counterfactual augmentation

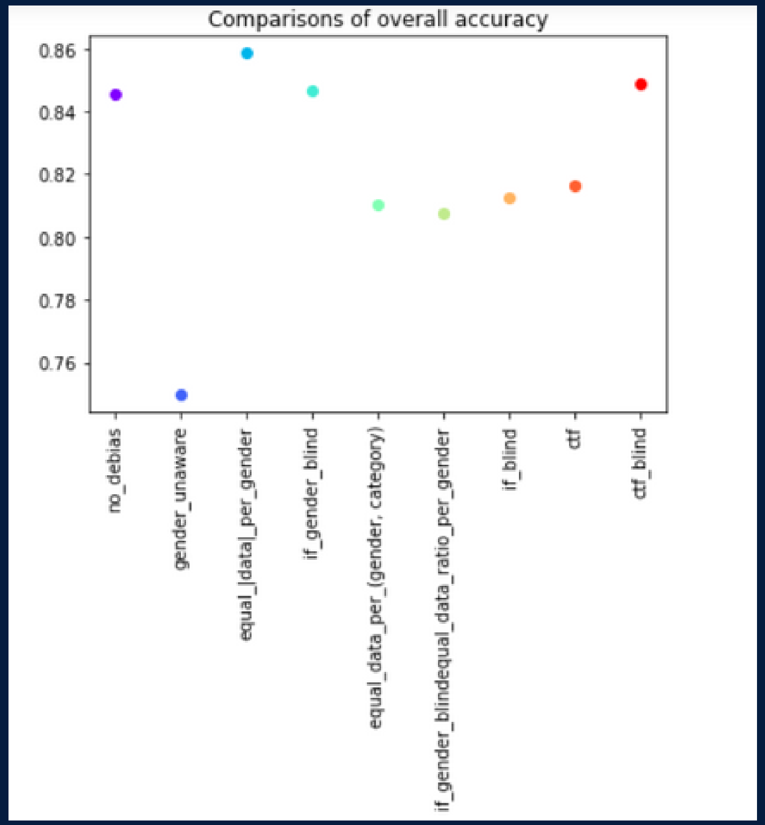


(unaware)

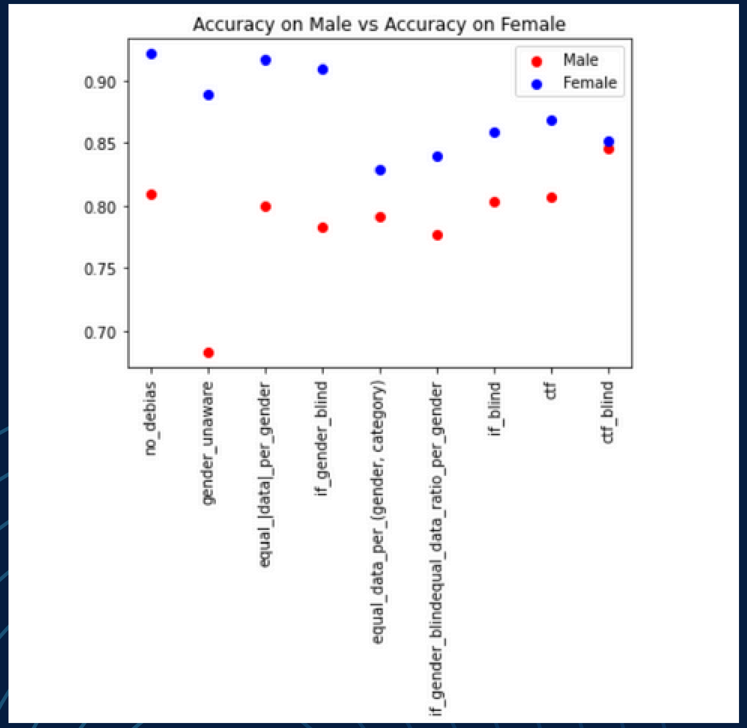


COMPARING BIAS MITIGATION APPROACHES

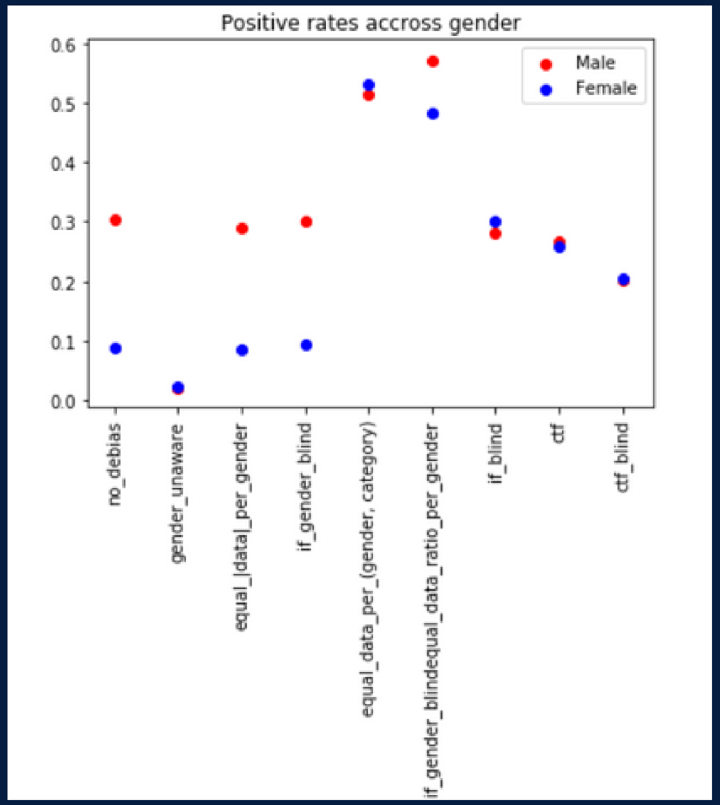
Comparing overall accuracies

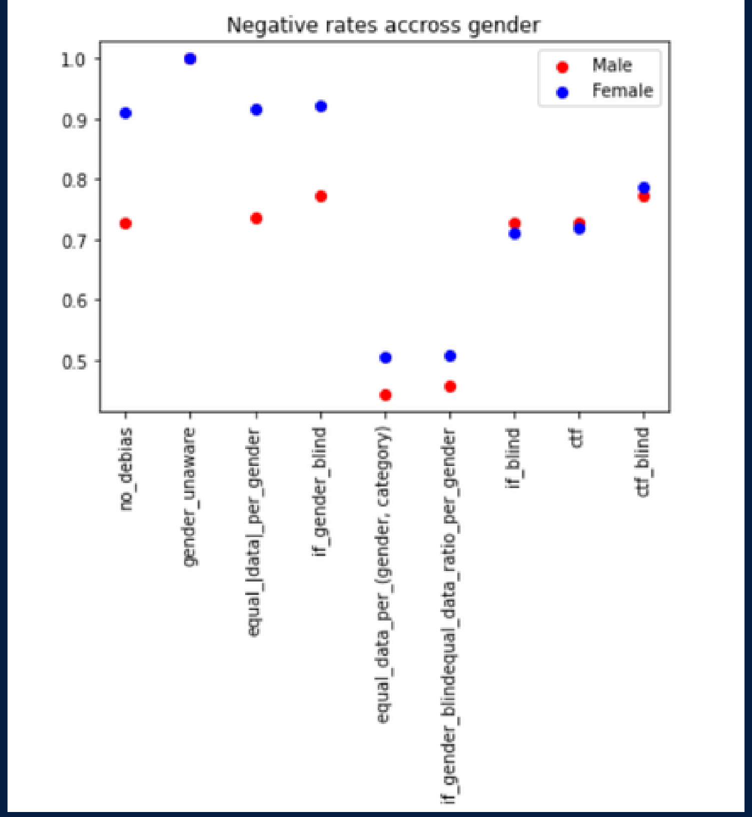


Comparing overall accuracy accross gender

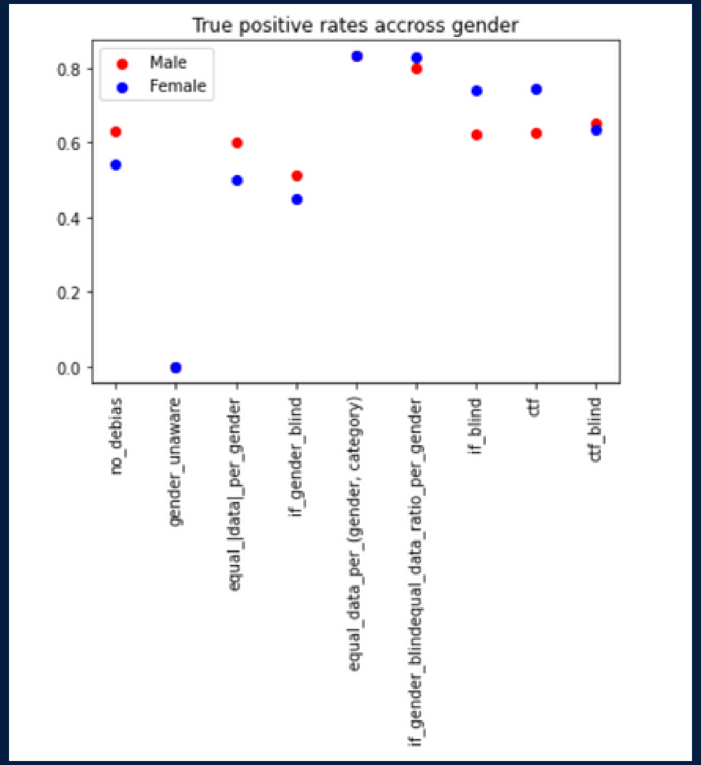


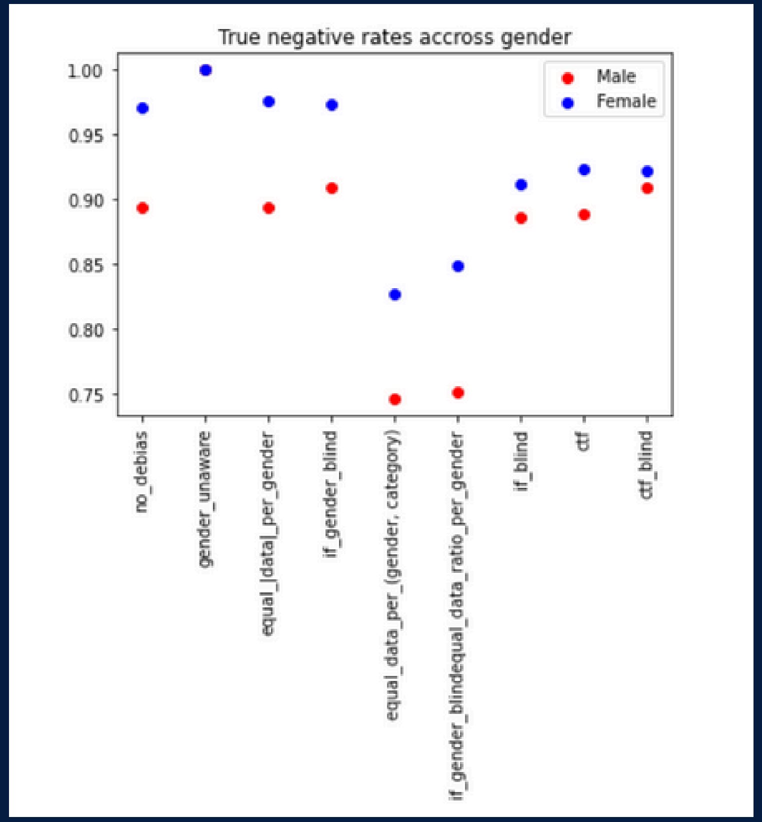
Positive and negative rates accross gender





True positive and True negative rates accross gender





THANK YOU