

## Question 2:

**Objective:** To perform classification for Specs and Non Specs Faces and Mitigate the bias using 2 methods, adding more data and Multitasking approach.

**Task:** Perform classification using 5 layers of Neural Network with architecture [ 128 -- 128 -- 128 -- 64 -- 1 ) ]

### Performance Metrics:

For Detection:

1. ROC	4. Precision-Recall
2. Confusion Matrix	5. Accuracy
3. Area Under Curve(AUC)	

### Inferences from the observed results:

#### Bias Mitigation

Pre-Processing Algorithms:

Pre-processing algorithms are used to mitigate bias prevalent in the training data. The idea is to apply one of the following techniques for preprocessing the training data set and then apply classification algorithms for learning an appropriate classifier.

Reweighting:

Reweighting is a data preprocessing technique that recommends generating weights for the training examples in each (group, label) combination differently to ensure fairness before classification. The idea is to apply appropriate weights to different tuples in the training dataset to make the training dataset discrimination-free with respect to the sensitive attributes. Instead of reweighting, one could also apply techniques (non-discrimination constraints) such as suppression (remove sensitive attributes) or massaging the dataset — modify the labels (change the labels appropriately to remove discrimination from the training data). However, the reweighting technique is more effective than the other two mentioned earlier.

Optimized preprocessing:

The idea is to learn a probabilistic transformation that edits the features and labels in the data with group fairness, individual distortion, and data fidelity constraints and objectives.

Learning fair representations:

The idea is to find a latent representation that encodes the data well while obfuscating information about protected attributes.

Disparate impact remover:

Feature values are appropriately edited to increase group fairness while preserving rank-ordering within groups.

### **In-Processing Algorithms**

Adversarial Debiasing:

A classifier model is learned to maximize prediction accuracy and simultaneously reduce an adversary's ability to determine the protected attribute from the predictions. This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can exploit.

Prejudice remover:

The idea is to add a discrimination-aware regularization term to the learning objective.

### **Post-Processing Algorithms**

Equalized odds postprocessing:

The algorithm solves a linear program to find probabilities with which to change output labels to optimize equalized odds.

Calibrated equalized odds postprocessing:

The algorithm optimizes over calibrated classifier score outputs to find probabilities with which to change output labels with an equalized odds objective.

Reject option classification:

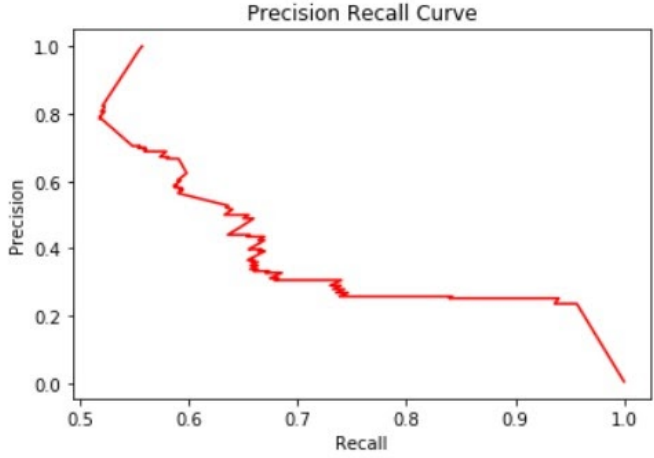
The idea is to give favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty.

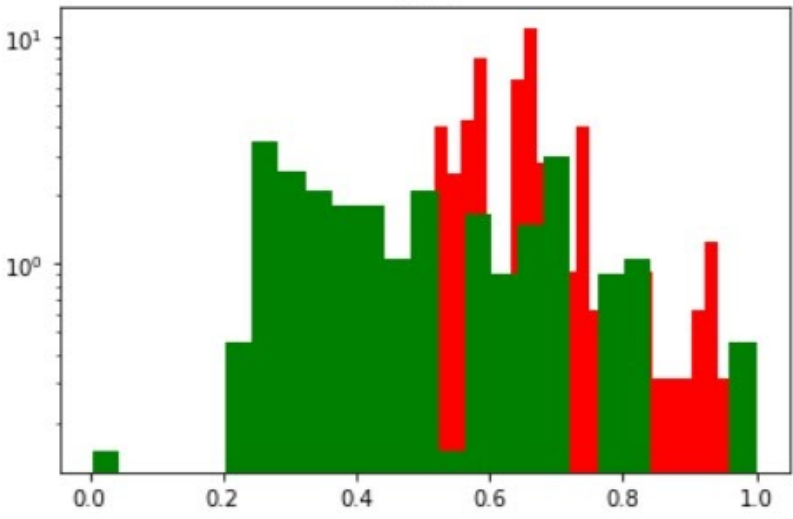
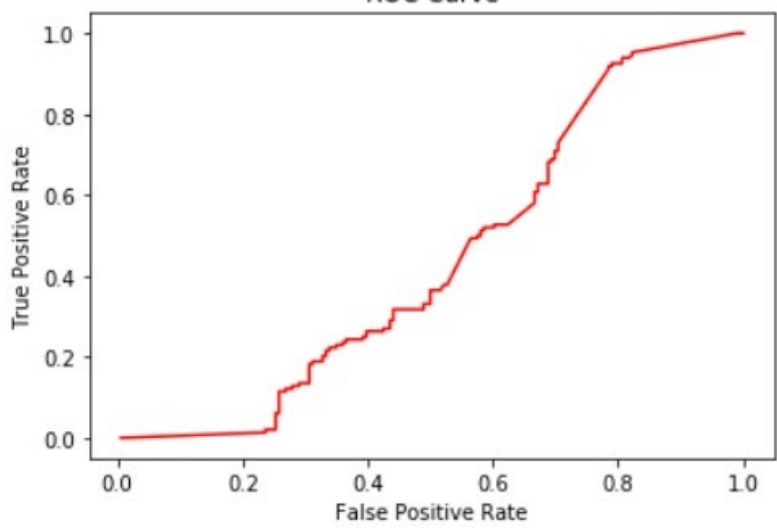
## QUESTION 2:

**Accuracy:**

41.916 %

**Evaluation Metrics to detect Bias:**

Metrics	Result
Precision Recall	 <p>A Precision-Recall curve plot titled "Precision Recall Curve". The x-axis is labeled "Recall" and ranges from 0.5 to 1.0 with major ticks every 0.1. The y-axis is labeled "Precision" and ranges from 0.0 to 1.0 with major ticks every 0.2. A red line represents the curve, starting at a precision of 1.0 for a recall of approximately 0.52. It then drops to a precision of about 0.8 at recall 0.55, and continues to decrease with some fluctuations, reaching a precision of approximately 0.3 at recall 0.75. From recall 0.75 to 0.95, the precision remains relatively flat around 0.25. Finally, it drops sharply to a precision of 0.0 at recall 1.0.</p>
Confusion Matrix	<pre>[[ 36 112]  [ 82 104]]</pre>

<p>Area Under Curve(AUC)</p>	<p>AUC</p> 
<p>Receiver Operating Characteristics</p>	<p>ROC Curve</p> 

**Bias Mitigation:**

Before adding New Data	After adding New Data
$\begin{bmatrix} 36 & 112 \\ 82 & 104 \end{bmatrix}$	$\begin{bmatrix} 106 & 42 \\ 102 & 84 \end{bmatrix}$
Accuracy: 41.916 %	Accuracy: 56.886 %

**Conclusion:**

I have learned the basic idea behind a bias system, ways of detection the bias, the metrics to be used when to measure the bias and the ways to mitigate the bias.