UNIVERSITÄT DES SAARLANDES

Designing Fair Machine Learning Model

Deborah D. Kanubala & Angel Gabriel

20.06.2025

Group Activity

- Use this dataset and fit a simple classification model of your choice:
- https://github.com/Ghana-Data-Science-Summit-IndabaX-Ghana/IndabaX25/blob/main/Designing%20Fair%20Machine%20Learning%20Algorithms%20(Advanced%20Level)/default%20of%20credit%20card%20clients.xls
- Evaluate the performance of your model with any metric of your choice. However, provide justification for your choice of metric
- You are welcome to work in groups

Tutorial Outline

UNIVERSITÄT DES SAARLANDES

MOTIVATION FOR FAIRNESS

MEASURING UNFAIRNESS

MITIGATION OF UNFAIRNESS

Part I: Motivation

Motivation

MAY 5, 2020

Black drivers get pulled over by police less at night when their race is obscured by 'veil of darkness,' Stanford study finds

After analyzing 95 million traffic stop records, filed by officers with 21 state patrol agencies and 35 municipal police forces from 2011 to 2018, researchers concluded that "police stops and search decisions suffer from persistent racial bias."

Image from: [Joshua Loftus]

Google apologizes after its Vision AI produced racist results

by Nicolas Kayser-Bril

A Google service that automatically labels images produced starkly different results depending on skin tone on a given image. The company fixed the issue, but the problem is likely much broader.

STORY 7 APRIL 2020 #DISCRIMINATION

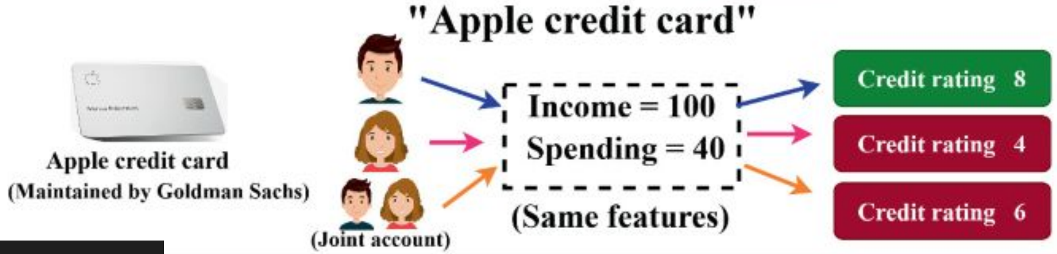
Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

AI 'EMOTION RECOGNITION' CAN'T BE TRUSTED

The belief that facial expressions reliably correspond to emotions is unfounded, says a new review of the field

Jul 25, 2019, 11:56am EDT



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Protected Attributes

Legally Recognized Protected Classes

Race (Civil Rights Act of 1964); **Color** (Civil Rights Act of 1964); **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964); **Religion** (Civil Rights Act of 1964);**National origin** (Civil Rights Act of 1964); **Citizenship** (Immigration Reform and Control Act); **Age** (Age Discrimination in Employment Act of 1967);**Pregnancy** (Pregnancy Discrimination Act); **Familial status** (Civil Rights Act of 1968); **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

[Boracas & Hardt 2017]

Sources of Bias

- **Sample Bias:** Occurs when one population is overrepresented or underrepresented in a training dataset.

- **Label Bias:** Occurs when annotation process introduce bias during creation of training data.

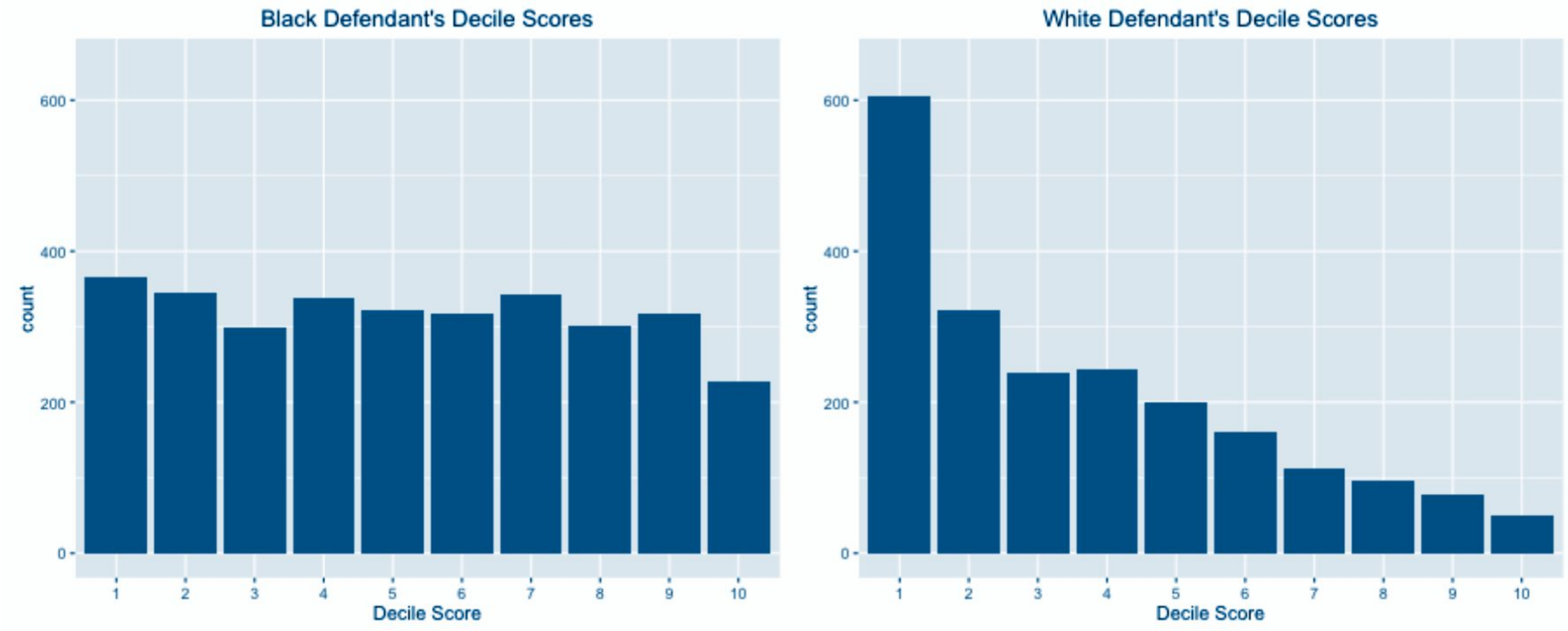
- **Outcome proxy Bias:** Occurs when the ML task is not specified appropriately. (Arrest - > Police arrest, Cost of health system -> quality of health.)

- **Human Biases in Historical data:** Historical data contains human biases and stereotypes.

Why Fairness is Hard using COMPASS as case Study

Compass Study Case

- 2016 ProPublica article analyzed COMPAS scores for >7000 people arrested in Broward county, Florida



Question: How many of these people ended up committing new crimes within 2 years?

Error Metrics

	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	True Negative (TN)	False Positive (FP)
Outcome: Recidivated	False Negative (FN)	True Positive (TP)

$$\text{Error Rate} = \frac{FP+FN}{TN+FP+FN+TP}$$ *How often is the prediction wrong?*

Defendants care about this

$$\text{False Positive Rate} = \frac{FP}{FP+TN}$$ *How often were non-offenders predicted to reoffend?*

Judges care about this

$$\text{False Negative Rate} = \frac{FN}{FN+TP}$$ *How often were offenders predicted not to reoffend?*

Error Estimation

Black Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	990 (TN)	805 (FP)
Outcome: Recidivated	532 (FN)	1369 (TP)

Error Rate ≈ 36.2%

White Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	1139 (TN)	349 (FP)
Outcome: Recidivated	461 (FN)	505 (TP)

Error Rate ≈ 33.0%

Similar error rates between white and black defendants

Error Estimation

Black Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	990 (TN)	805 (FP)
Outcome: Recidivated	532 (FN)	1369 (TP)

White Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	1139 (TN)	349 (FP)
Outcome: Recidivated	461 (FN)	505 (TP)

Error Rate ≈ 36.2%

Error Rate ≈ 33.0%

False Positive Rate ≈ 44.9%

False Positive Rate ≈ 23.5%

Black defendants have 1.9x higher False Positive Rate!

Why Fairness is Hard

UNIVERSITÄT DES SAARLANDES

Black Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	990 (TN)	805 (FP)
Outcome: Recidivated	532 (FN)	1369 (TP)

White Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	1139 (TN)	349 (FP)
Outcome: Recidivated	461 (FN)	505 (TP)

Error Rate ≈ 36.2%

Error Rate ≈ 33.0%

False Positive Rate ≈ 44.9%

False Negative Rate ≈ 28.0%

False Positive Rate ≈ 23.5%

False Negative Rate ≈ 47.7%

White defendants have 1.7x higher False Negative Rate

Why Fairness is Hard

UNIVERSITÄT DES SAARLANDES

Black Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	990 (TN)	805 (FP)
Outcome: Recidivated	532 (FN)	1369 (TP)

White Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	1139 (TN)	349 (FP)
Outcome: Recidivated	461 (FN)	505 (TP)

Fairness through unawareness

Surprising fact: COMPAS gives very different outcomes for white vs black defendants, but it does not use race as an input to the algorithm!

Why Fairness is hard

⚖️ Tension	💡 What it looks like in practice	🚀 Take-away
Which metric to satisfy?	Demographic parity, equalized odds, calibration, predictive parity … you can't have them all at once (*impossibility theorems*). Choosing one redistributes harm/benefit.	Metric choice is a **normative** decision, not only technical.
"Fairness through unawareness" doesn't work	Dropping the sensitive feature leaves proxies (zip code ⇒ race, career gaps ⇒ gender). The model re-learns group labels and bias becomes harder to spot.	Blindness ≠ fairness; **redlining could happen.**
Data imbalance & representation	Minority classes often have sparse, noisy labels; majority groups dominate the loss signal. Over-sampling or re-weighting can help, but may inflate variance.	Garbage in → garbage out; **fix upstream data** as well as models.

Types of Fairness

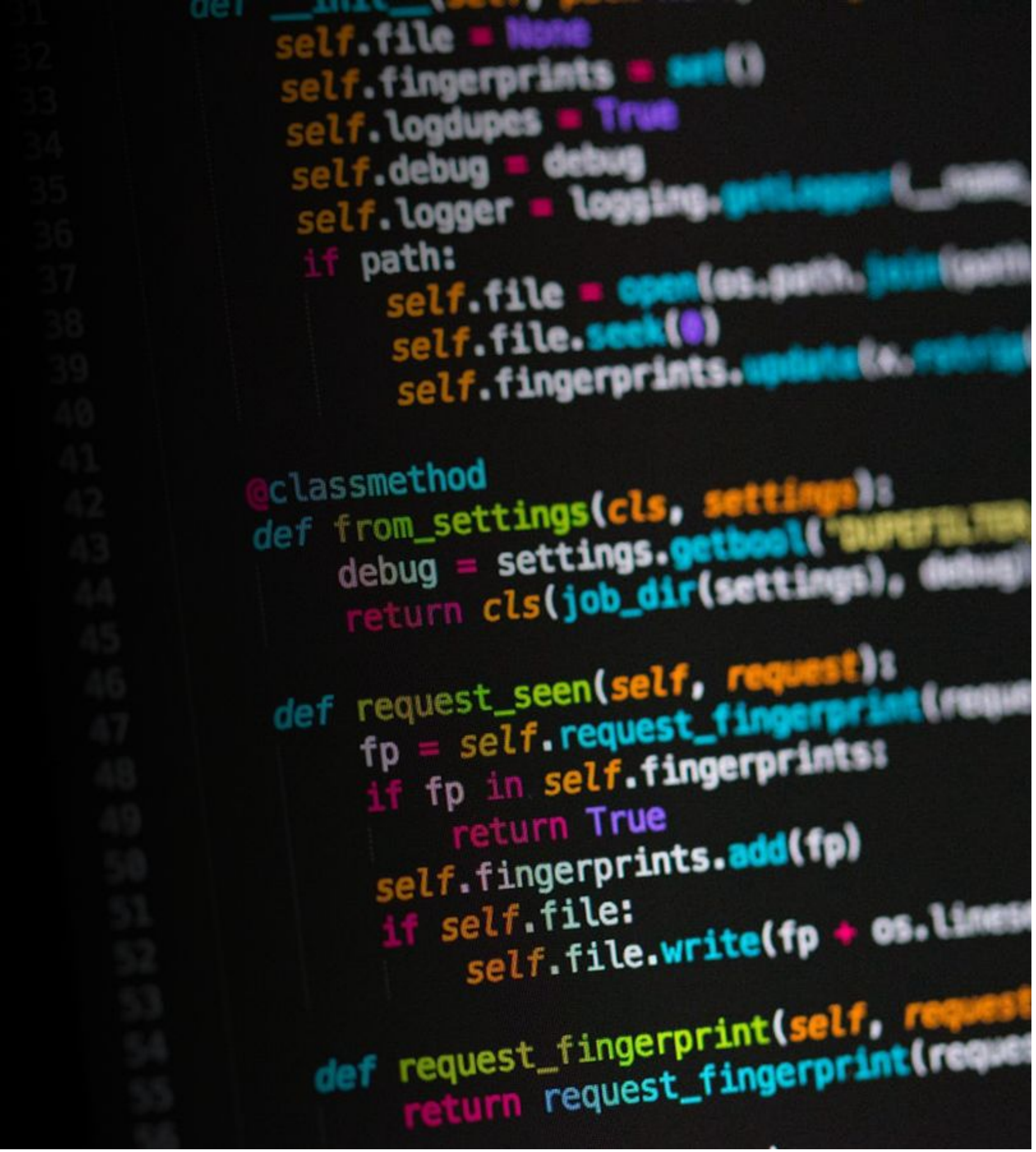
- Individual Fairness: Similar individuals should be treated similarly.

- Group Fairness: Outcomes of a decision making system should not differ systematically between two demographic groups.

- Counterfactual Fairness: Outcomes of an algorithm would not changed if in a counterfactual world if the individual had a different demographic characteristics.

Hands on

Coding Time

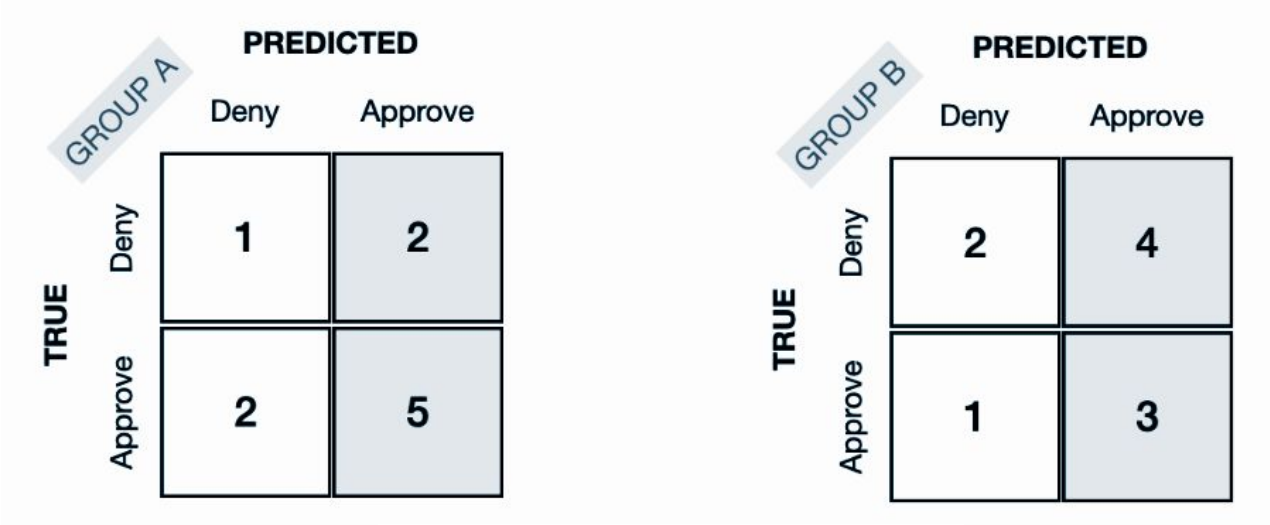


Part II: Measuring Unfairness

Demographic Parity

Demographic parity is a fairness metric whose goal is to ensure a machine learning model's predictions are independent of membership in a sensitive group.

$$\mathbb{P}\{\widehat{Y}=1 \mid A=a\}=\mathbb{P}\{\widehat{Y}=1 \mid A=b\}.$$



Demographic parity

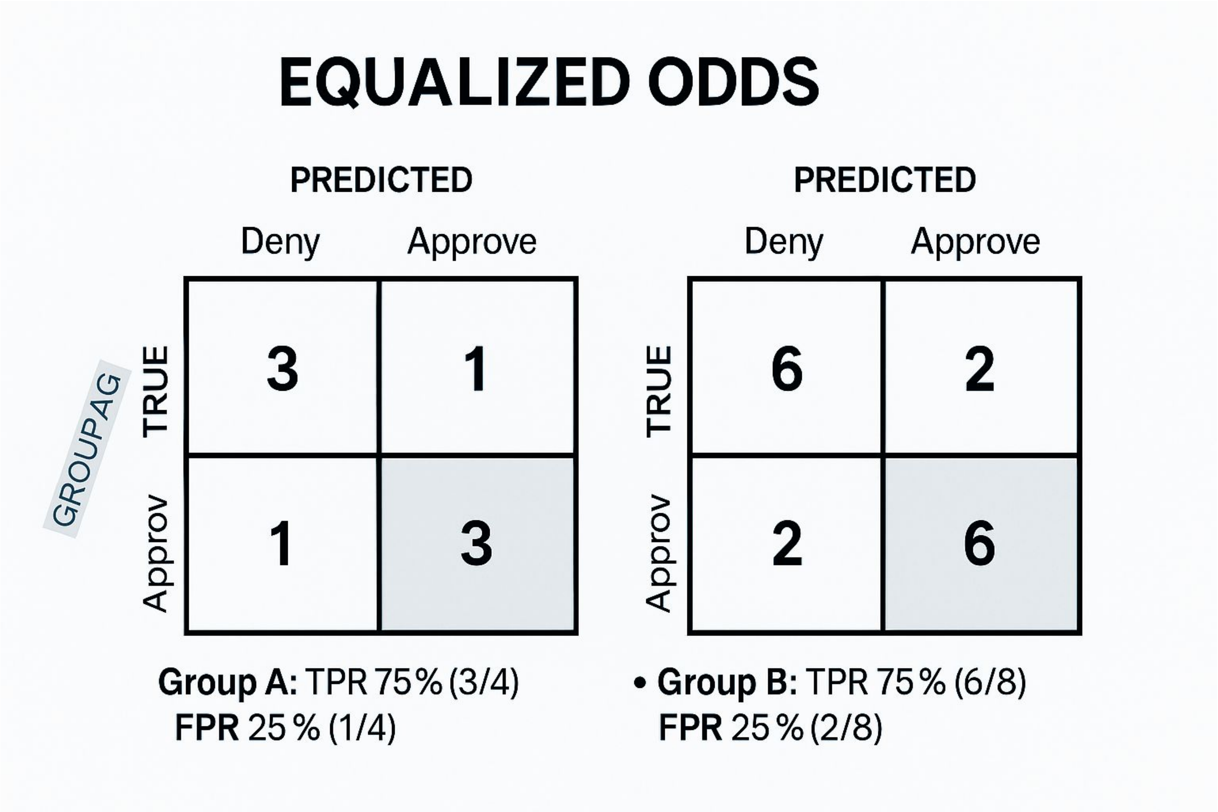
20 applicants (50% from **Group A**)
14 approvals (50% from **Group A**)

Equalized Odds

The goal of the equalized odds fairness metric is to ensure a machine learning model performs equally well for different groups.

It requires that the machine learning model's predictions are not only independent of sensitive group membership, but that groups have the same false positive rates and true positive rates.

TPR = TP/(TP + FN)
FPR = FP/(FP+TN)

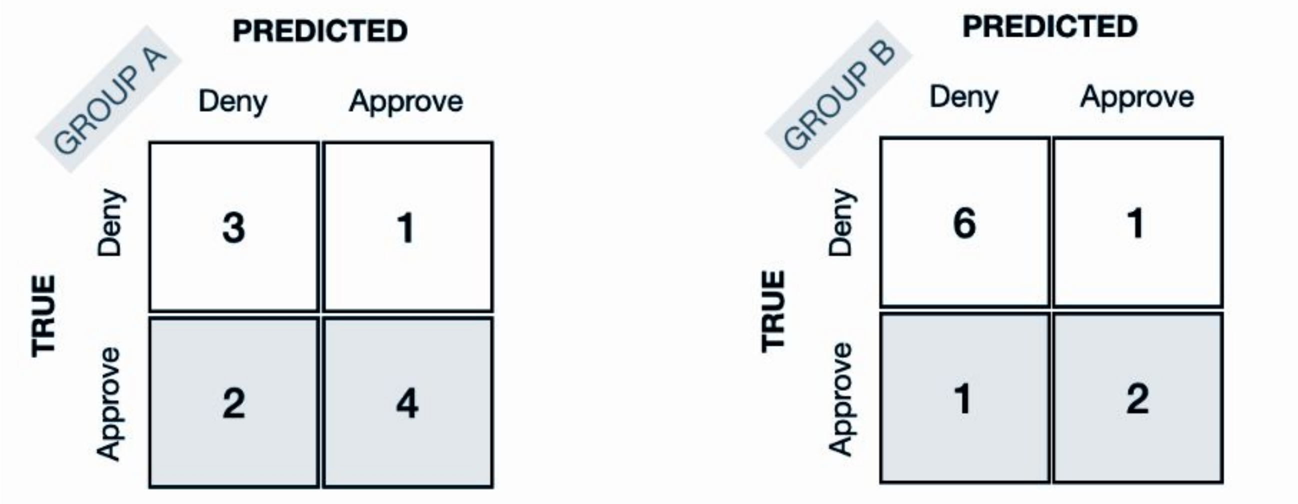


Equality of Opportunity

Equal opportunity is a relaxed version of equalized odds that only considers conditional expectations with respect to positive labels, i.e.,

$$\mathbb{P}\{\widehat{Y} = 1 \mid Y = 1, A = a\} = \mathbb{P}\{\widehat{Y} = 1 \mid Y = 1, A = b\}$$

TPR = TP /(TP + FN)

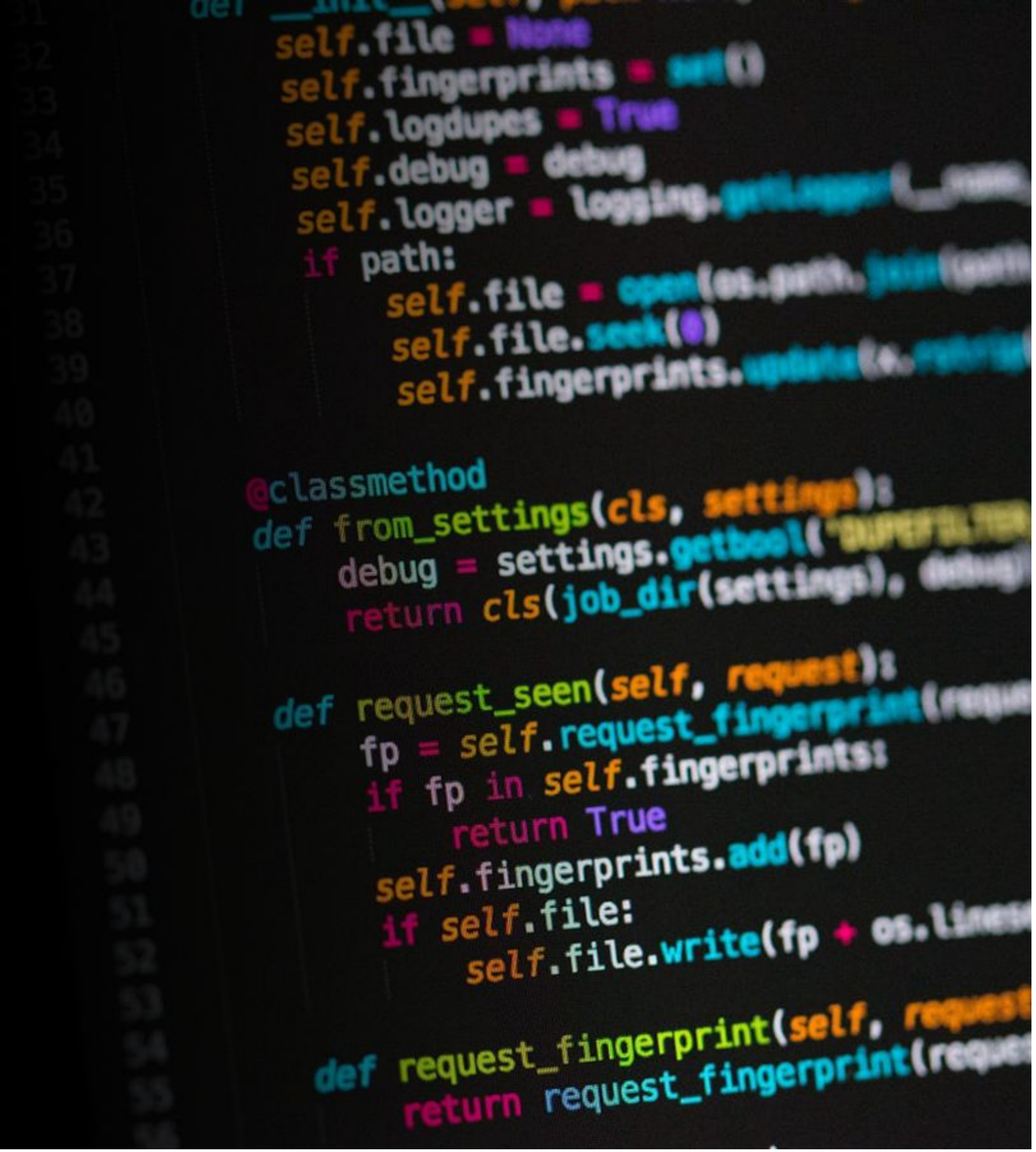


Equal opportunity

Group A: 66% true positive rate: 4/(4+2)
Group B: 66% true positive rate: 2/(1+2)

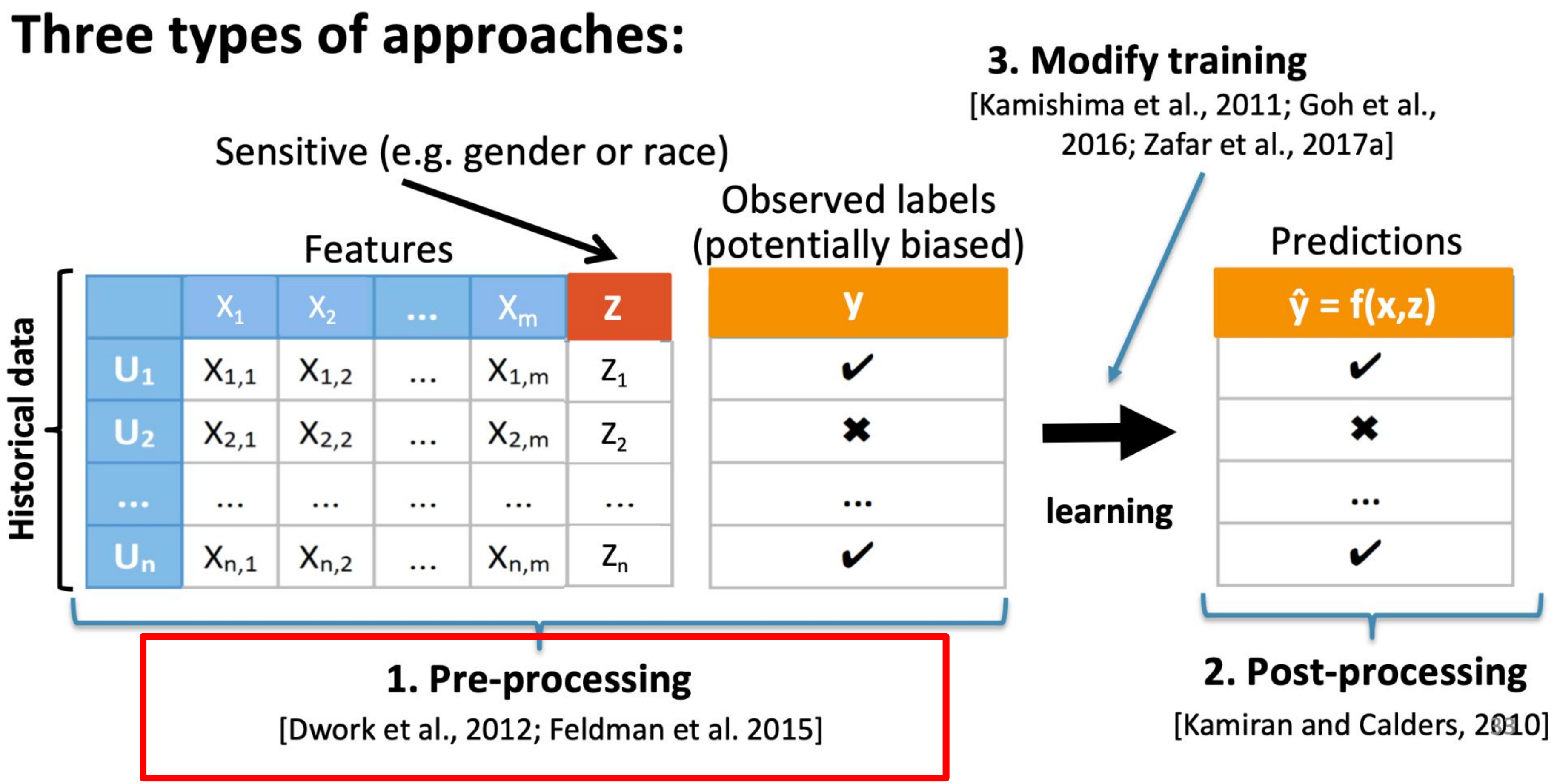
Hands on

Coding Time



Part III: Mitigation Unfairness

Fairness Mitigation Approaches



Pre Processing Approach

- Reweighting generates weights for the training samples in each (group, label) combination differently, ensuring fairness before classification

- Higher weights are assigned to instances that are underrepresented and lower weights are assigned to instances that are overrepresented.

Additional methods:

- **Fair Representations**
  Finds a latent representation that encodes the data well.

- **Disparate Impact Remover**
  Edits feature values to increase group fairness while preserving rank ordering within groups

In Processing Approach

Example using Demographic Parity as a metric

MLE loss binary cross entropy. Can be squared, hinge etc

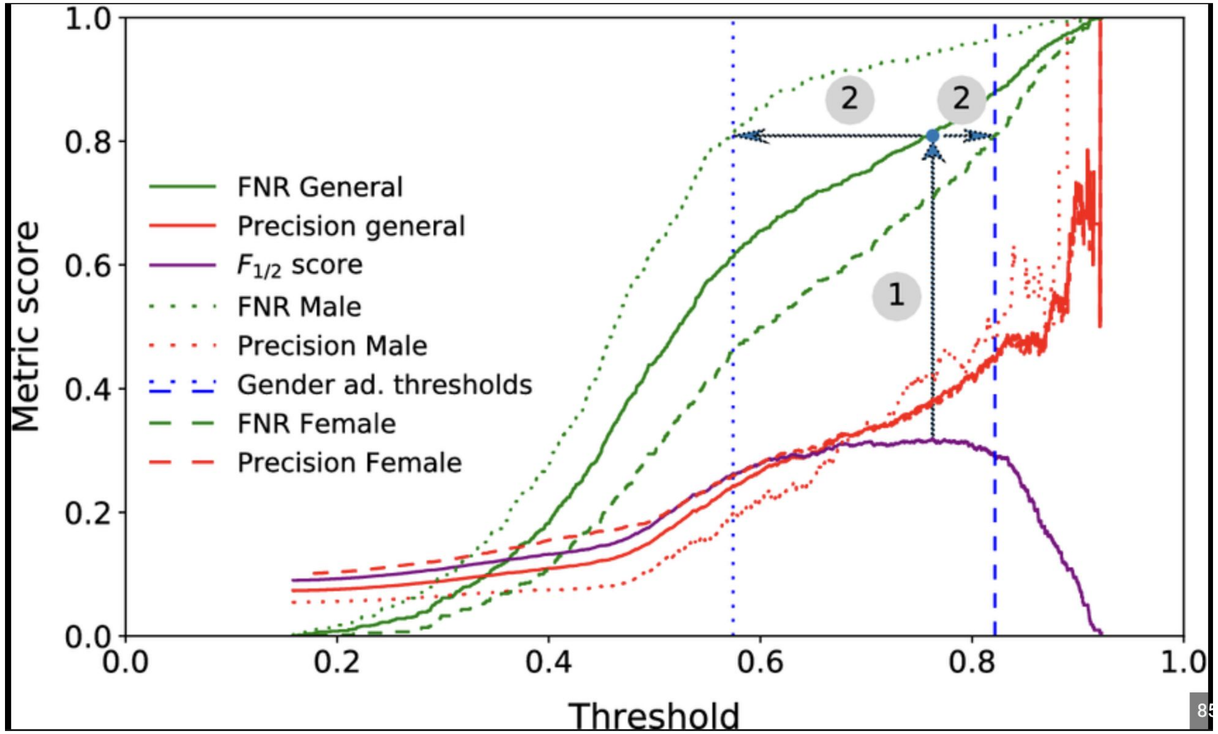
$$DP - Loss = BCE + \lambda * \underbrace{(\mathbb{E}[\widehat{Y} = 1 \mid Z = 1] - \mathbb{E}[\widehat{Y} = 1 \mid Z = 0])^2}_{\text{Fairness metric}}$$

Controls the strength of fairness

Fairness metric. For example, here it demographic parity bit can be replace with equality of opportunity etc

- <u>Adversarial Debiasing</u>
  Learns classifier to maximize prediction accuracy and simultaneously reduces adversary's ability to determine protected attribute from predictions
- <u>Prejudice Remover</u>
  Adds a discrimination aware regularization term to the learning objective
- <u>Meta Fair Classifier</u>
  Takes fairness metric as part of the input and returns a classifier optimized for the metrics

Post Processing Approach



Threshold Optimizer

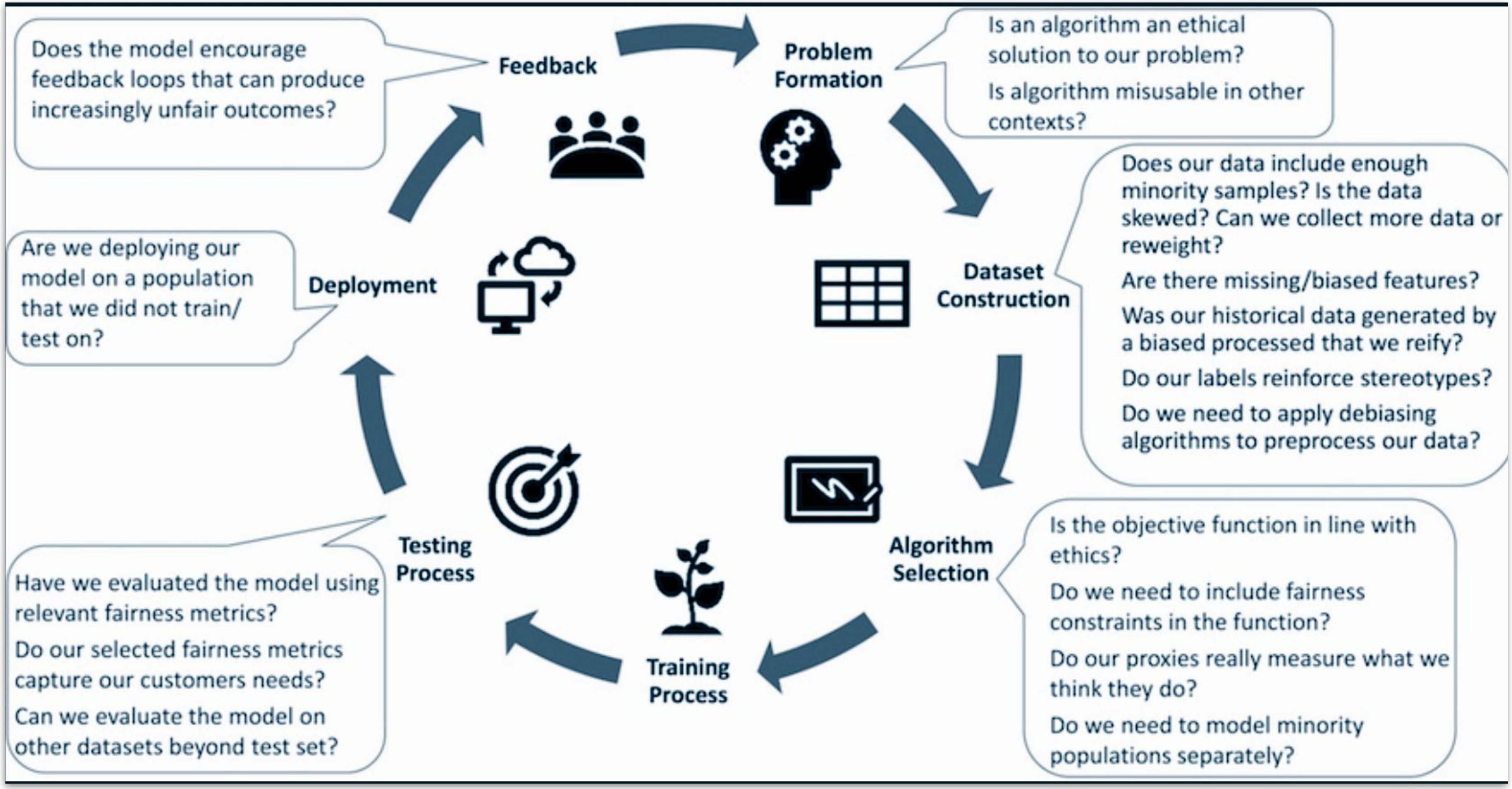
Equalized Odds:
Modifies predicted label using an optimization scheme to make predictions more fair

Interactive tool

https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Building Fair Models

UNIVERSITÄT DES SAARLANDES



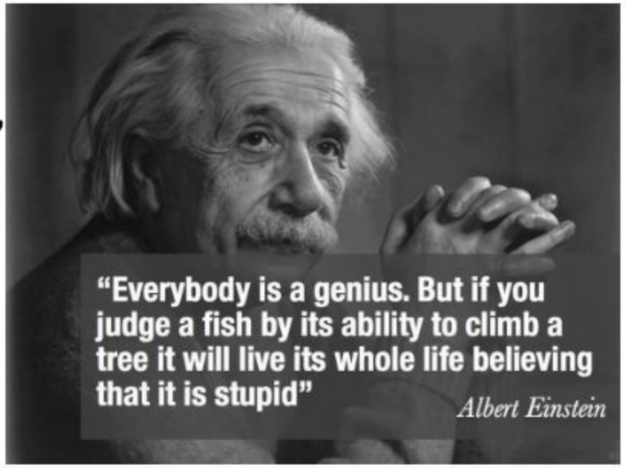
Fairness-aware Machine Learning, Bennett et al., WSDM Tutorial (2019).

Take Home Message

UNIVERSITÄT DES SAARLANDES

- For **Ethical ML**, first **bear in mind the assumptions –** wrong assumption come often at a high social cost.
- Assumptions start with the **data collection process –** which features, data vs population distribution, feedback loops, etc.
- More realistic assumptions may require of **probabilistic approaches** (e.g., stochastic decisions).
- **Essential to have a holistic view of the algorithm** – starting from the data collection process before training, all the way to the deployment in the real-world.

Not because something is technically possible, it is the right thing to do!



Slides developed using the following materials

- Slides: Justin Johnson & David EECS 442 WI 2021 lecture: https://web.eecs.umich.edu/~justincj/slides/eecs442/WI2021/442_WI2021_lecture18.pdf
- AI Fairness Learn about four different types of fairness. Assess a toy model trained to judge credit card applications: https://www.kaggle.com/code/alexisbcook/ai-fairness

Additional materials for interested participants

UNIVERSITÄT DES SAARLANDES

- Fairness Tutorial Notebook: https://colab.research.google.com/drive/1HN-sLQXQ3hClQbv3OyGUX9uEujQLBbwj#scrollTo=Ik_FsjBWhJfX
- What is Fair about Individual Fairness: https://philsci-archive.pitt.edu/18889/1/Fleisher%20-%20Individual%20Fairness.pdf
- Equality of Opportunity in Supervised Learning: https://arxiv.org/pdf/1610.02413.pdf
- AI Fairness How to measure and Reduce Unwanted Bias in ML: https://krvarshney.github.io/pubs/MahoneyVH2020.pdf
- AIF360 Library: https://aif360.res.ibm.com/
- A Survey on Bias and Fairness in Machine Learning: https://arxiv.org/pdf/1908.09635.pdf
- A clarification of the nuances in the fairness metrics landscape: https://www.nature.com/articles/s41598-022-07939-1
- Fairness in ML Survey paper: https://dl.acm.org/doi/pdf/10.1145/3616865

Feedback Form

THANK YOU FOR YOUR ATTENTION

If you have any questions or interested in learning or doing research in this area, please contact: **dkanubala@aimsammi.org** **or** **adaambiikgabz45@gmail.com**