An Auditor's Guide for Discrimination Detection

Federica Picogna – f.picogna@nyenrode.nl

1. Development of Artificial Intelligence: Exploring Benefits and Risks

2. Regulating Artificial Intelligence: An Overview of the AI Act

3. Limitations of the AI Act: Current Gaps and Challenges

4. Addressing the Challenge of Defining Fairness

5. Solving the Fairness Measures Selection Challenge

6. Identifying New Challenges in the AI Field

The Growing Impact of Artificial Intelligence: Benefits and Risks Tied with Its Use

Artificial Intelligence is affecting our decisions and our lifestyle every day

→ Benefits

Increased decision-making speed

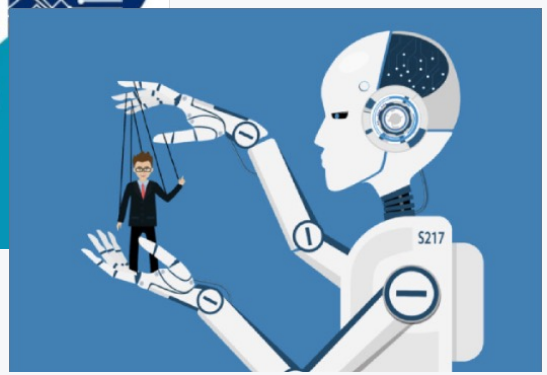
Automated repetitive tasks

Boosted productivity

→ Downsides

Privileged group vs Unprivileged group

Possible discrimination



Understanding the Benefits and Risks of AI Use with a Real-Life Case Study: The COMPAS Case

How can we decide whether to impose a long or short sentence? How can we decide whether to allow early release or not?

Automatic classification of offenders according to their risk of committing another crime within two years of release

HOWEVER

AI leads to systematic and incorrect lengthy incarceration and denial of early release

Balancing the Benefits and Risks of AI Use through Regulations like the AI Act

The Artificial Intelligence Act

→ Benefits of AI use

→ Respect for the rights recognized for all EU citizens

AI Classification Based on Risk

Who can verify the existence of this delicate balance and classify AI? **The auditor**

What Are the Auditor's Challenges in AI Assessment? Fairness Definition and Fairness Measure Selection

1. What is meant by "Fair"?

2. Which fairness measure should they choose?

The Ambiguity of the AI Act in Defining the Auditor's Task: What Does “Fairness” Mean?

> ***Diversity, non-discrimination and fairness means that AI systems are developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law. Social and environmental well-being means that AI***

College Admissions:

- Merit-Based Criteria
- Inherent Disadvantages for Certain Communities

What Is the Solution for the Fairness Definition Problem? Non-Generic, Context-Specific, and Measure-Based Definitions

We define Fairness according to the fairness measure used to conduct the AI auditing process

Resolving the issue of choosing the appropriate fairness measure

EQUALS

Resolving the issue of defining Fairness

What Are the Auditor's Challenges in AI Assessment? Fairness Definition and Fairness Measure Selection

1. What is meant by "Fair"?

2. Which fairness measure should they choose?

Understanding Fairness Measures through a Real-Case Example: Description of the DUO Case

- 615 students
- Different information are included, such as home visit results
- Migration Background: Sensitive Attribute
  - 474 Privileged Students
  - 141 Unprivileged Students

No-Abuse of the independent-living allowance	Abuse of the independent-living allowance
465 Privileged Students	9 Privileged Students
131 Unprivileged Students	10 Unprivileged Students
596 Students in Total	19 Students in Total

The Majority of Fairness Measures Stem from the Confusion Matrix: Understanding it through the DUO Case Dataset

		True Value		
		n	p	
Predicted Value	$\hat{n}$	455 **True Negative (TN)**	14 **False Negative (FN)**	469 **Predicted Negative**
	$\hat{p}$	141 **False Positive (FP)**	5 **True Positive (TP)**	146 **Predicted Positive**
		596 **Actual Negative**	19 **Actual Positive**	615 **Total**

Summary of the Terminology Explained

- Negative = Label provided by AI to indicate a student who is not abusing the housing grants.
- Positive = Label provided by AI to indicate a student who is abusing the housing grants.
- True Negative (TN) = The number of students that AI correctly predicts are not abusing the housing grants.
- True Positive (TP) = The number of students that AI correctly predicts are abusing the housing grants.
- False Negative (FN) = The number of students that AI incorrectly predicts are not abusing the housing grants.
- False Positive (FP) = The number of students that AI incorrectly predicts are abusing the housing grants.

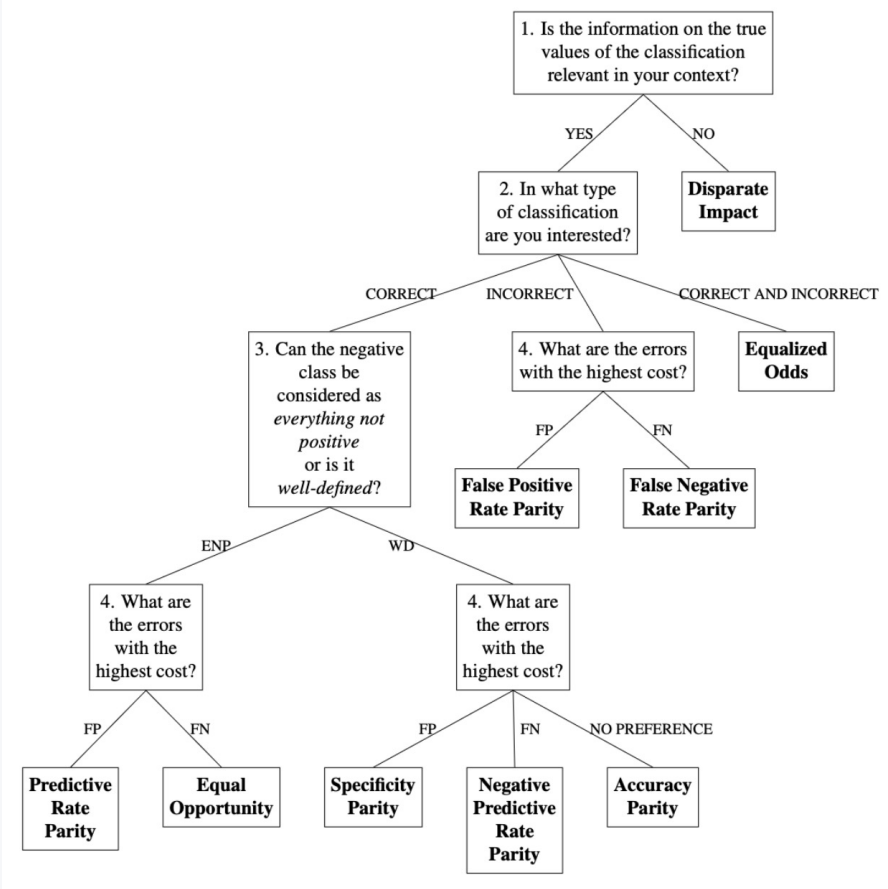
So many Fairness Measures Can Be Obtained by Combining the Quantities of the Confusion Matrix: Which One Should the Auditor Use?

"Column" based measures	TPR or Sensitivity	Equal Opportunity	$\frac{TP}{TP + FN}$
	FPR	False Positive Rate Parity	$\frac{FP}{TN + FP}$
	TNR or Specificity	Specificity Parity	$\frac{TN}{TN + FP}$
	FNR	False Negative Rate Parity	$\frac{FN}{TP + FN}$

"Row" based measures	Precision	Predictive Rate Parity	$\frac{TP}{TP + FP}$
	Negative Predictive Value	Negative Predictive Rate Parity	$\frac{TN}{TN + FN}$
	Positive Rate	Disparate Impact	$\frac{\text{Predicted Positive}}{\text{Population Size}}$

"Combined" measure	Accuracy	Accuracy Parity	$\frac{TP + TN}{TP + FN + TN + FP}$

The Decision-Making Workflow the Auditor Can Use as a Guide in Selecting the Most Appropriate Fairness Measure



Choosing a Fairness Measure Means Choosing the Associated Fairness Definition

Group Fairness Measures	Fairness Definitions
Disparate Impact	Is the percentage of people whose label according to AI is Positive (=change in the status quo)the same in the privileged group as in the unprivileged group? If the answer is yes then the AI is considered fair
Equalized Odds	AI is fair if its use results in the same number of people from different groups experiencing correct and incorrect changes in the status quo.
Equal Opportunity and Predictive Rate Parity	AI is fair if its use results in the same number of people from different groups experiencing correct changes in the status quo.
Specificity Parity and Negative Predictive Rate Parity	AI is fair if its use results in the same number of people from different groups correctly experiencing no changes in the status quo.
False Positive Rate Parity	AI is fair if its use results in the same number of people from different groups experiencing incorrect changes in the status quo.
False Negative Rate Parity	AI is fair if its use results in the same number of people from different groups incorrectly experiencing no changes in the status quo.
Accuracy Parity	AI is fair if its use results in the same number of people from different groups correctly experiencing changes and and correctly experiencing no changes in the status quo.

Analyzing Discrimination in AI Predictions: A Practical Approach with the DUO Case Dataset

Privileged Students

		True Value		
		n	p	
Predicted Value	$\hat{n}$	359	5	364
	$\hat{p}$	106	4	110
		465	9	474

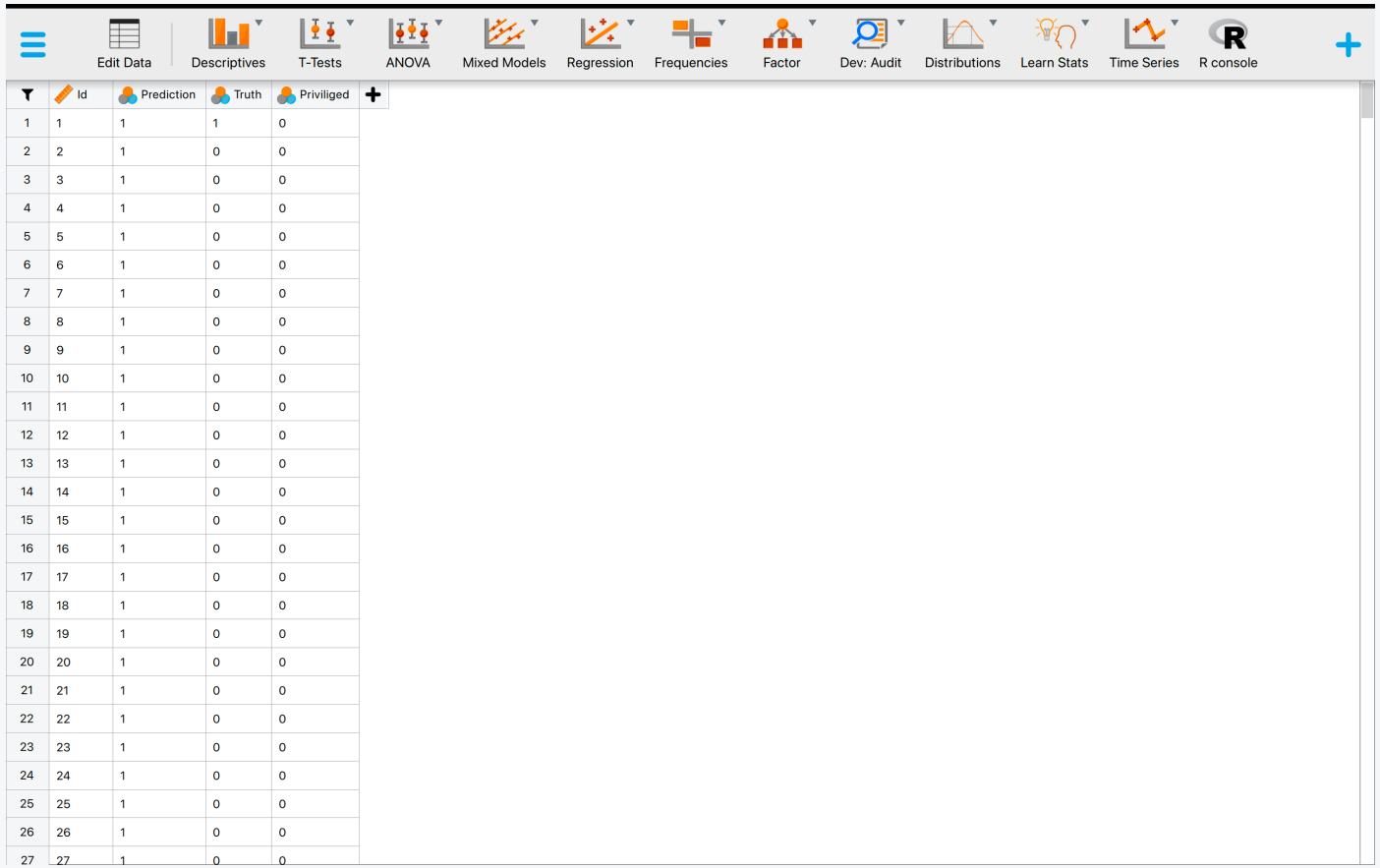
Unprivileged Students

		True Value		
		n	p	
Predicted Value	$\hat{n}$	96	9	105
	$\hat{p}$	35	1	36
		131	10	141

How different are the AI's predictions for Privileged versus Unprivileged students?

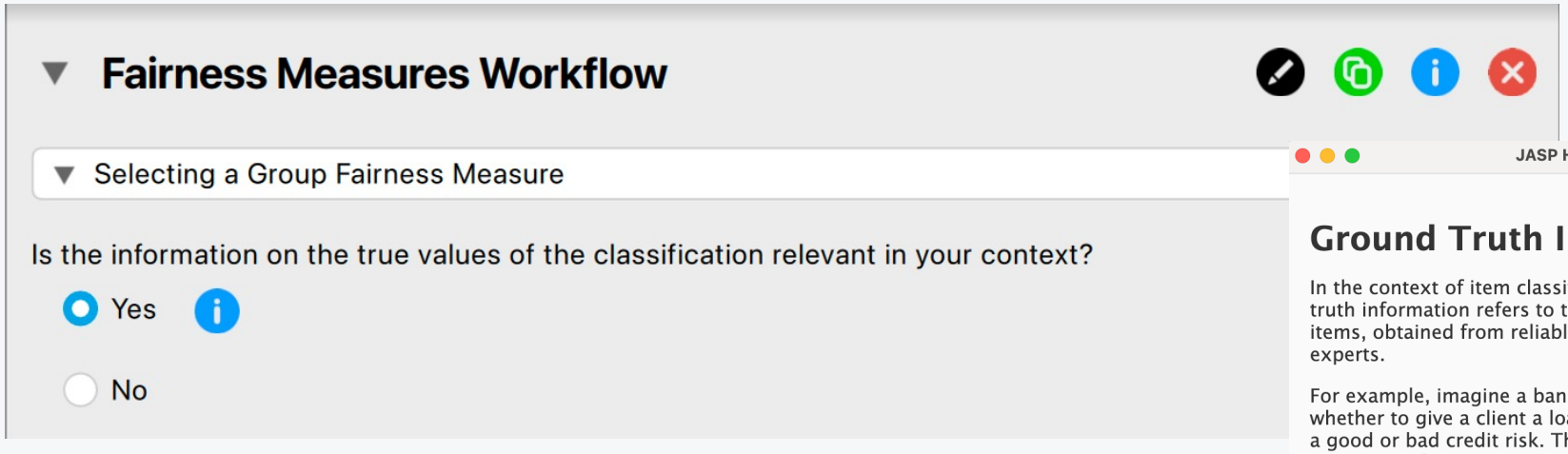
- A confusion matrix is obtained for each group: how AI correctly and incorrectly classifies the students
- Auditor's task: answering decision-making workflow questions to obtain the best fairness measure to use and use it to draw conclusions on potential discrimination in AI outcomes.

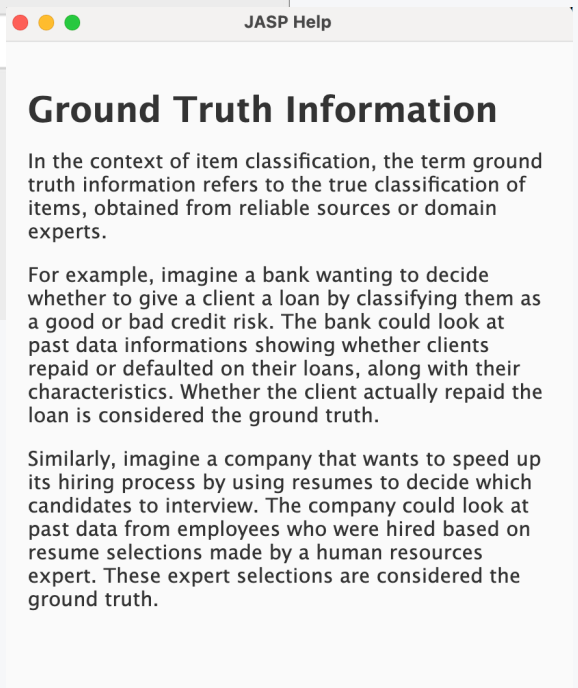
JASP Demo: The DUO Case Dataset



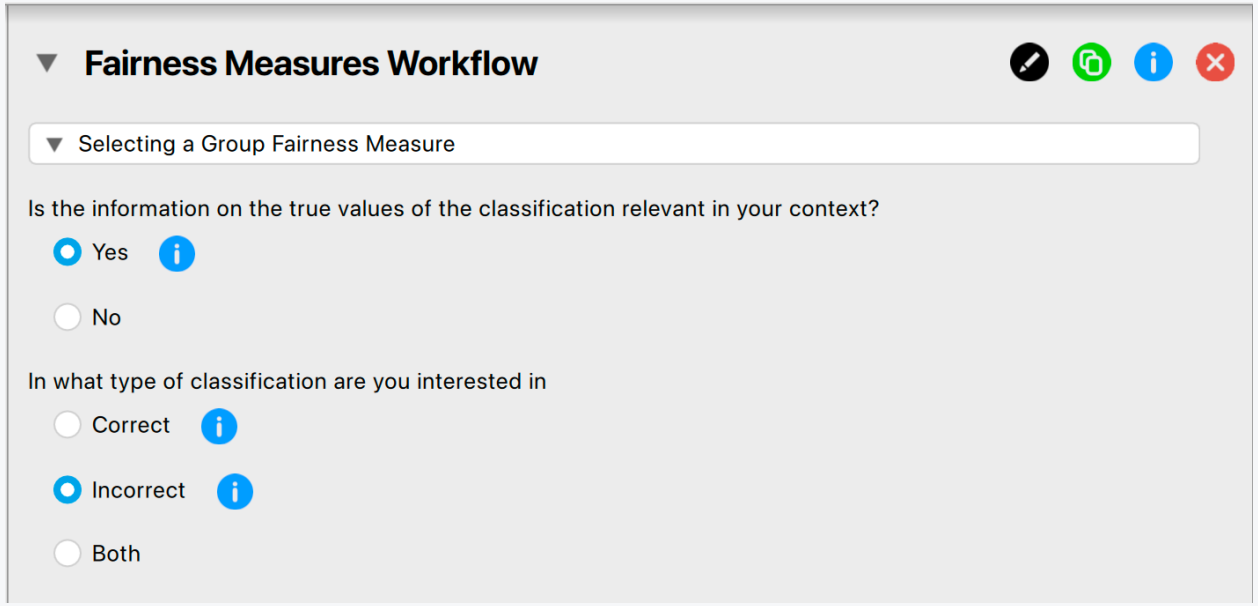
	Id	Prediction	Truth	Priviliged
1	1	1	1	0
2	2	1	0	0
3	3	1	0	0
4	4	1	0	0
5	5	1	0	0
6	6	1	0	0
7	7	1	0	0
8	8	1	0	0
9	9	1	0	0
10	10	1	0	0
11	11	1	0	0
12	12	1	0	0
13	13	1	0	0
14	14	1	0	0
15	15	1	0	0
16	16	1	0	0
17	17	1	0	0
18	18	1	0	0
19	19	1	0	0
20	20	1	0	0
21	21	1	0	0
22	22	1	0	0
23	23	1	0	0
24	24	1	0	0
25	25	1	0	0
26	26	1	0	0
27	27	1	0	0

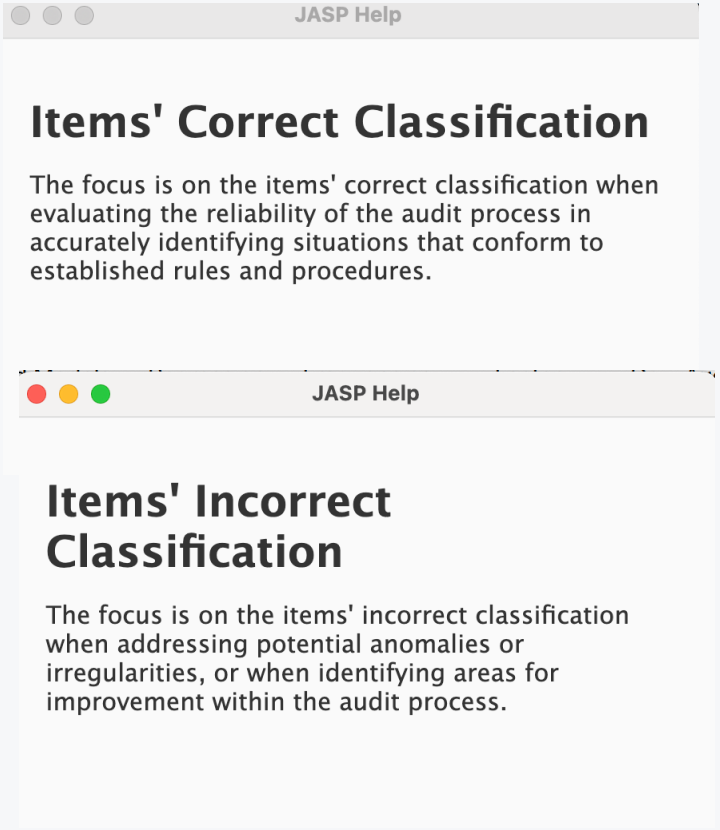
JASP Demo: The First Question in the Decision-Making Workflow



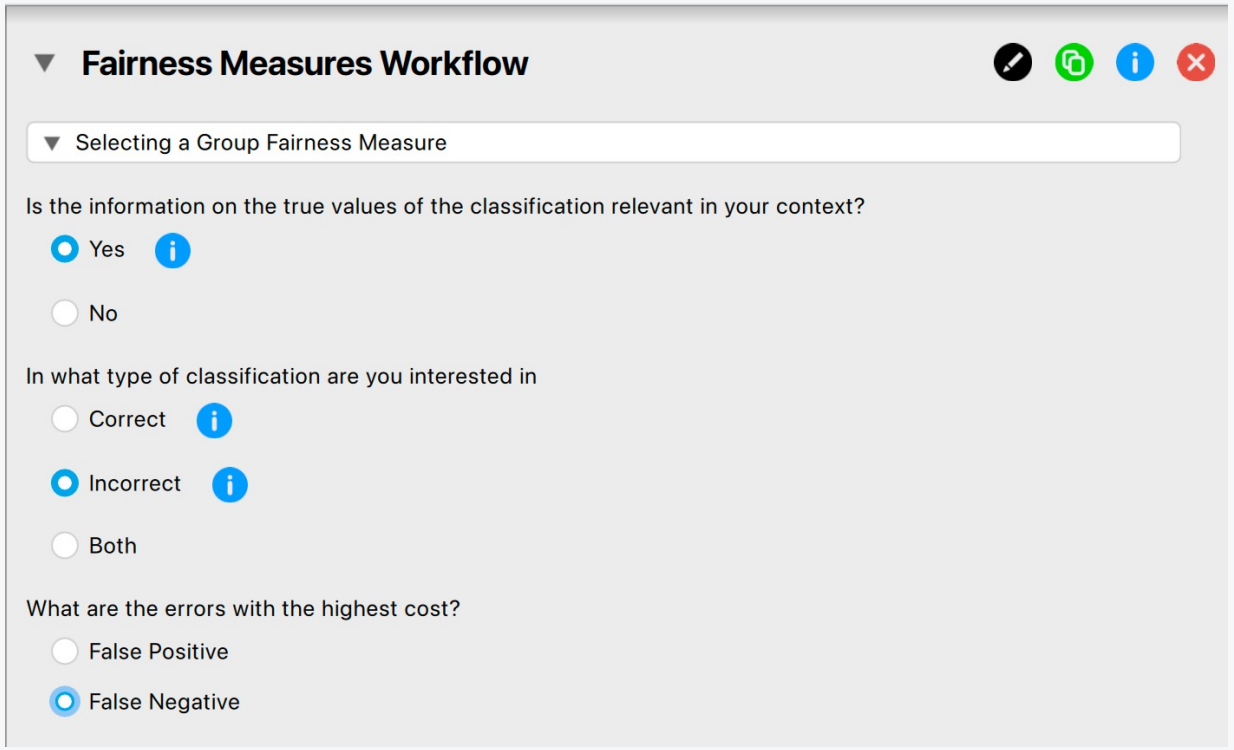


JASP Demo: The Second Question in the Decision-Making Workflow





JASP Demo: The Fourth Question in the Decision-Making Workflow



JASP Demo: The Outcome

Results

Fairness Measures Workflow

The selected Fairness Metric is False Negative Rate Parity

Details regarding the fairness measure

Fairness Definition:
AI is considered fair if it provides the same amount of incorrect negative predictions for both privileged and unprivileged groups. In other words, AI fairness is achieved when the same number of items from these two groups incorrectly experience no changes from the status quo. This change in status quo can refer to favorable outcomes, such as being selected for a job interview or receiving reimbursement for medical expenses. However, it can also represent negative outcomes, such as being deemed high risk for reoffending within two years of release or defaulting on a bank loan.
The term items refers to what is being classified; these items can be people, like job applicants, or objects, such as bank accounts.
The term negative predictions refers to one of the two possible predictions the AI can make: positive or negative. This is because we are working within the framework of binary classification, where there are only two possible classes for items to be categorized: positive or negative.

Fairness Measure Formula:
The False Negative Rate Parity is based on the False Negative Rate. Therefore, the False Negative Rate, whose formula is FN/(TP+FN), is applied to both the privileged group and the unprivileged group.
FN indicates the number of False Negatives, meaning the number of items with a true positive classification that the AI classifies as negative, and TP indicates the number of True Positives, meaning the number of items with a true positive classification that the AI also classifies as positive.

Do We Have Discrimination in the DUO Case Dataset? Application of False Positive Rate Parity as a Fairness Measure

False Negative Rate for Privileged Students :

$$\frac{\text{FN}}{\text{TP} + \text{FN}} = \frac{5}{4 + 5} = 0.55$$

The number of students that AI incorrectly predicts are NOT abusing the housing grants.

The total number of students abusing the housing grants (equal to the sum of the students whom AI correctly predicts are abusing the grants and the students whom AI incorrectly predicts are not abusing them).

False Negative Rate for Unprivileged Student:

$$\frac{\text{FN}}{\text{TP} + \text{FN}} = \frac{9}{1 + 9} = 0.90$$

Addressing Additional AI Challenges for Future Development: The Need for a Tolerance Threshold for Tolerable Unfairness

1. When can the difference be interpreted as evidence of discrimination in the AI outcome?

2. How can the auditor compare different metrics?

Tolerable Difference and Thresholds
Audit Risk

Disparate Impact: The Only Fairness Measure with a Threshold That However Lacks a Statistical Interpretation

Positive Rate for Privileged Students :

$$\frac{\text{Predicted Positive}}{\text{Population Size}} = \frac{110}{474} = 0.232$$

Positive Rate for Unprivileged Students:

$$\frac{\text{Predicted Positive}}{\text{Population Size}} = \frac{36}{141} = 0.255$$

Disparate Impact:

$$\frac{\text{PRPS}}{\text{PRUS}} = \frac{0.232}{0.255} = 0.91 > 0.80$$

For each group, the total number of students that AI correctly and incorrectly predicts are abusing the housing grants divided by the total amount of students

MEANING: How many students does AI think are abusing the housing grants among all those for whom information is available? And how does this perception vary between the two groups? Does AI believe that Privileged Students are behaving better?

The Importance of Choosing the Right Fairness Measure

False Positive Rate for Privileged Students :

$$\frac{\text{FP}}{\text{TN} + \text{FP}} = \frac{106}{106 + 359} = 0.228$$

False Positive Rate for Unprivileged Student:

$$\frac{\text{FP}}{\text{TN} + \text{FP}} = \frac{35}{35 + 96} = 0.267$$

The number of students that AI incorrectly predicts are abusing the housing grants.

The total number of students NOT abusing the housing grants.

So the overall conclusion depends on which stakeholder the auditor deems the most important, as the chosen fairness measure is influenced by the responses to the workflow questions, which will vary for each stakeholder.

WHAT WE DID SO FAR:

- Development of a decision-making workflow for the auditor to select the appropriate fairness measure for evaluating discrimination and therefore the appropriate fairness definition

WHAT WE WILL WORK ON:

- Translation of the concept of Audit Risk into the AI field

- Development of a Tolerance Difference in the AI field

Thank you for the attention!