Fairness is not just about **data:** two models trained on the same data may be **fair** or **may not**.

We propose a **new way** to **analyze** and **measure** algorithmic fairness that goes **beyond predictive performance**.

IFAFAM: Individual Feature Algorithmic Fairness Analysis Method applied on predicting students' performance

Mélina Verger[1], Vanda Luengo[1], Francois Bouchet[1], Sébastien Lallé[1]
[1] *MOCAH team, Sorbonne University, CNRS, LIP6*
✉ <melina.verger@lip6.fr>

Motivation

- Increasing need for algorithmic fairness analysis and quantification
- Lack of **sensitive features discovery** – from what is learned from the models
- Proliferation of *performance-oriented* fairness metrics and *performance-oriented* analysis approaches

IFAFAM (contribution)

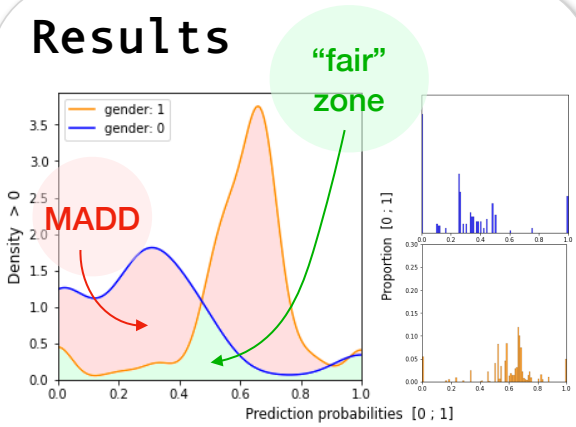
1. **General necessary steps**
   * prediction probabilities (PP) retrieval
   * sensitive-feature subset reduction
2. **Qualitative analysis**
   * smoothing of PP distributions
   * visual identification of **unequal treatment** and **stereotypical judgement**
3. **Quantitative analysis**
   * a new way to measure algorithmic unfairness: the **mean absolute density distance** (MADD)

MADD (MAE-like but on density, model-independent)

$$\frac{1}{N}\sum_{i=1}^{N}|d_i - d'_i|$$

N: total number of density observations
d_i: density value related to the prediction probability p_i of one group
d'_i: same as d_i but for the other group

Results



	MADD (OULAD dataset)		
	gender	level of poverty	disability
LR	**1.85***	1.79*	1.61*
NB	0.61	**0.98**	0.69
DT	**1.6**	1.5	1.26

Experimental validation :
* on several **educational** datasets and several binary classifier models (cross-tables) that **predict students' success to courses**
* comparison with existing fairness metrics (ABROCA, DI, TPR, ...)

Discussion

* binary sensitive features
* supervised-learning oriented
* for **any tabular data** without preprocessing

Take a photo to **visit my website**