# Evaluation of group fairness measures in student performance prediction problems

Tai Le Quy[1], Thi Huyen Nguyen[1], Gunnar Friege[1], Eirini Ntoutsi[2]

[1] Leibniz University Hannover (LUH)

[2] Bundeswehr University Munich (UniBw-M)

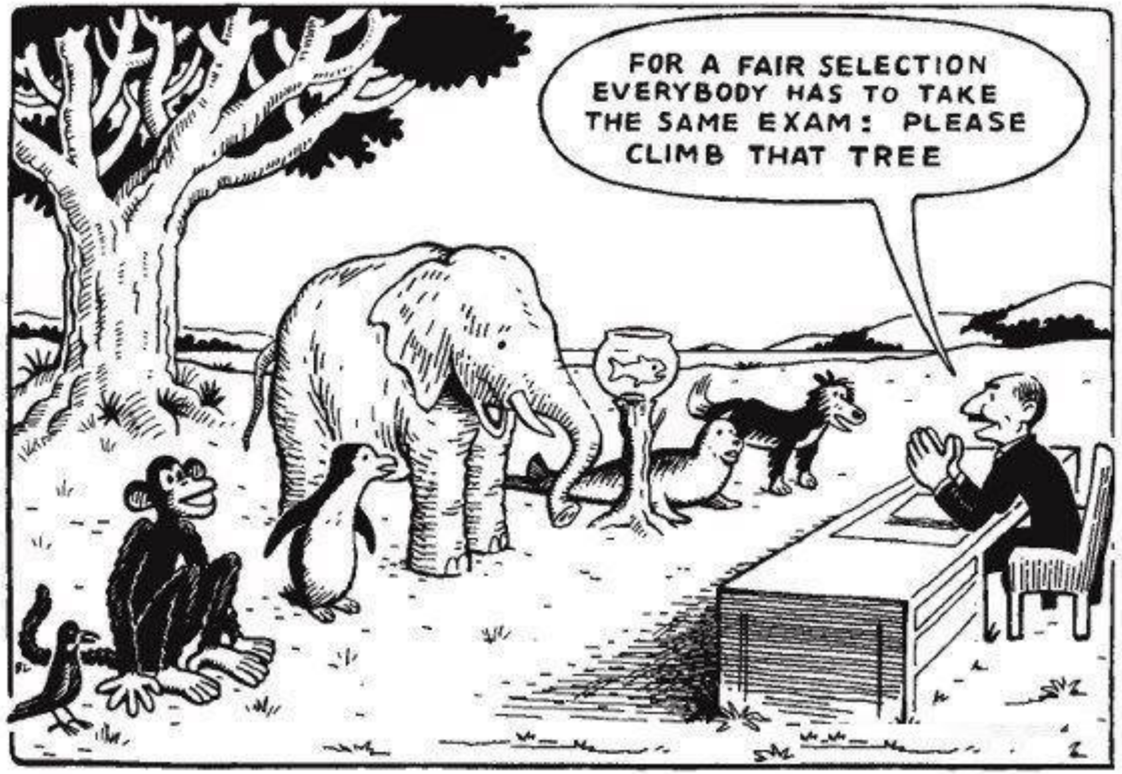
Grenoble, France, 23.09.2022

# Outline

- Introduction
- Problem definition
- Fairness measures
- Evaluation
- Conclusion and outlook

# Introduction (1/3)

- Fairness is a fundamental concept of education
  - All students must have an equal opportunity in study or,
  - be treated fairly regardless of their household income, assets, gender, or race, etc.




Source: https://educationrightsblog.wordpress.com/2016/06/11/comparison-to-canada/

# Introduction (2/3)

- Student performance prediction is a common task of the EDM community, that:
  - Supports in selecting courses and designing appropriate future study plans for students.
  - Helps teachers and managers to monitor students
  - Reduces the official warning signs as well as expelling students
- ML-based decisions can be biased to protected attributes such as gender or race due to historical discrimination embedded in the data

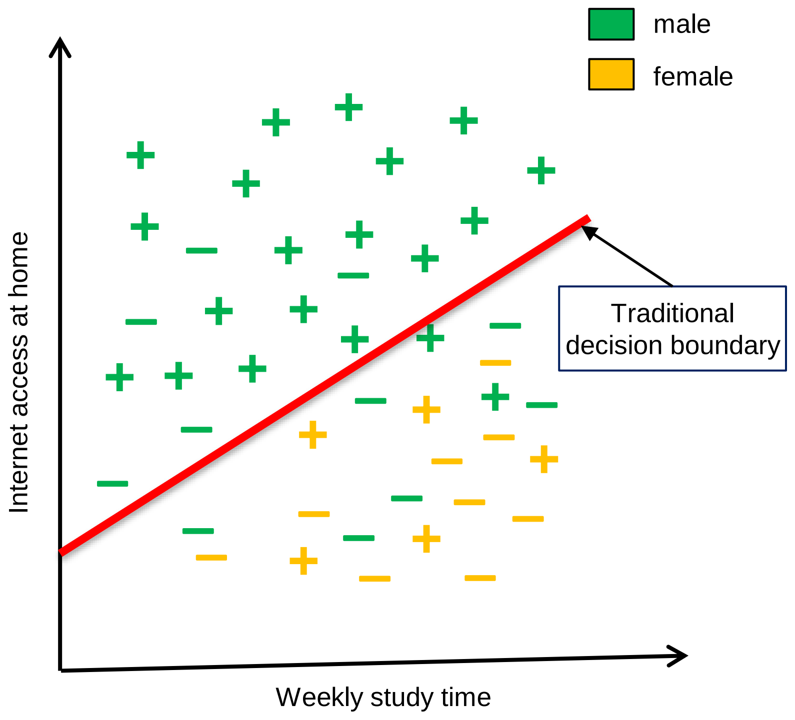

Source: https://machinextreme.com/publication/

# Introduction (3/3)

- A large variety of fairness measures have been introduced in ML area.
- Choosing proper measures can be cumbersome due to the dependence of fairness on context
- There is no metric that fits all circumstances !!!

# Problem definition

- Student performance prediction problem is considered as a binary classification task:
  - *D*: a binary classification dataset
  - Class attribute $Y$ = {+, −}, e.g., $Y$ = {pass, fail}
  - *S*: binary protected attribute, $S \in \{s, \bar{s}\}$, e.g., Gender ∈ {female, male}
    - *s*: the discriminated group (protected group), e.g., female
    - $\bar{s}$ : the non-discriminated group (non-protected group), e.g., male
  - Predicted outcome $\hat{Y}$ = {+, −}



# Fairness measures (1/5)

- The most prevalent group fairness notions used in ML

| Measures | Proposed by | Published year | #Citations |
|---|---|---|---|
| Statistical parity | Dwork et al. | 2012 | 2,367 |
| Equal opportunity | Hardt et al. | 2016 | 2,575 |
| Equalized odds | Hardt et al. | 2016 | 2,575 |
| Predictive parity | Chouldechova et al. | 2017 | 1,430 |
| Predictive equality | Corbett-Davies et al. | 2017 | 878 |
| Treatment equality | Berk et al. | 2018 | 626 |
| Absolute Between-ROC Area | Gardner et al. | 2019 | 84 |

- Example:
  - A dataset with 100 instances

| | | Predicted class | |
|---|---|---|---|
| | | **Positive +** | **Negative -** |
| **Actual class** | **Positive +** | True Positive (TP)<br>$TP_{prot} + TP_{non-prot}$<br>**70** (32:38) | False Negative (FN)<br>$FN_{prot} + FN_{non-prot}$<br>**10** (4:6) |
| | **Negative -** | False Positive (FP)<br>$FP_{prot} + FP_{non-prot}$<br>**9** (4:5) | True Negative (TN)<br>$TN_{prot} + TN_{non-prot}$<br>**11** (6:5) |

# Fairness measures (2/5)

- Statistical parity (SP ∈ [−1, 1])
  - Measures the difference (bias) in the predicted outcome ($\hat{Y}$) between any two groups

$$SP = P(\hat{Y} = +|S = \bar{s}) - P(\hat{Y} = +|S = s)$$

  - E.g.,

$$SP = \frac{38+6}{54} - \frac{32+4}{46} \approx 0.0322$$

- Equal opportunity (EO ∈ [0, 1])
  - The classifier should give similar results for students of both genders with actual “pass” class

$$EO = |P(\hat{Y} = -|Y = +, S = \bar{s}) - P(\hat{Y} = -|Y = +, S = s)|$$

  - E.g.,

$$EO = \left| \frac{38}{38+6} - \frac{32}{32+4} \right| \approx 0.0253$$

| Actual class | Predicted class: Positive + | Predicted class: Negative - |
|---|---|---|
| Positive + | True Positive (TP)<br>$TP_{prot} + TP_{non-prot}$<br>**70** (32:38) | False Negative (FN)<br>$FN_{prot} + FN_{non-prot}$<br>**10** (4:6) |
| Negative - | False Positive (FP)<br>$FP_{prot} + FP_{non-prot}$<br>**9** (4:5) | True Negative (TN)<br>$TN_{prot} + TN_{non-prot}$<br>**11** (6:5) |

# Fairness measures (3/5)

- Equalized odds (EOd ∈ [0, 2])
  - Predicted true positive and false positive probabilities should be the same between male and female student groups

$$EOd = \sum_{y \in \{+,-\}} |P(\hat{Y} = +|S = s, Y = y) - P(\hat{Y} = +|S = \bar{s}, Y = y)|$$

  - E.g., $EOd = |\frac{32}{32+4} - \frac{38}{38+6}| + |\frac{4}{4+6} - \frac{5}{5+5}| \approx 0.1253$
- Predictive parity (PP∈ [0, 1])
  - The probability of a student predicted to “pass” actually having “pass” class should be the same, for both male and female students

$$PP = |P(Y = +|\hat{Y} = +, S = s) - P(Y = +|\hat{Y} = +, S = \bar{s})|$$

  - E.g., $PP = \frac{32}{32+4} - \frac{38}{38+5} \approx 0.0052$

| | | Predicted class | |
|---|---|---|---|
| | | Positive + | Negative - |
| Actual class | Positive + | True Positive (TP) $TP_{prot} + TP_{non-prot}$ **70** (32:38) | False Negative (FN) $FN_{prot} + FN_{non-prot}$ **10** (4:6) |
| | Negative - | False Positive (FP) $FP_{prot} + FP_{non-prot}$ **9** (4:5) | True Negative (TN) $TN_{prot} + TN_{non-prot}$ **11** (6:5) |

# Fairness measures (4/5)

- Predictive equality (PE ∈ [0, 1])
  - The probability of students with an actual “fail” class being incorrectly assigned to the “pass” class should be the same for both male and female students

$$P(\hat{Y}=+|Y=-,S=s)=P(\hat{Y}=+|Y=-,S=\overline{s})$$

  - E.g., $PE=|\frac{4}{6+4}-\frac{5}{5+5}|=0.1$

- Treatment equality (TE)
  - The ratios of false negatives and false positives are the same for both male and female students

$$\frac{FN_{prot.}}{FP_{prot.}}=\frac{FN_{non-prot.}}{FP_{non-prot.}}$$

  - E.g., TE = −0.2

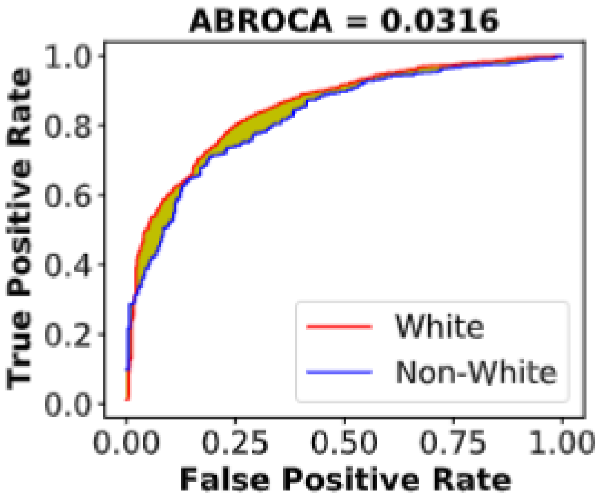
| | | Predicted class | |
|---|---|---|---|
| | | Positive + | Negative - |
| Actual class | Positive + | True Positive (TP) $TP_{prot}+TP_{non-prot}$ **70** (32:38) | False Negative (FN) $FN_{prot}+FN_{non-prot}$ **10** (4:6) |
| | Negative - | False Positive (FP) $FP_{prot}+FP_{non-prot}$ **9** (4:5) | True Negative (TN) $TN_{prot}+TN_{non-prot}$ **11** (6:5) |

# Fairness measures (5/5)

- Absolute Between-ROC Area (ABROCA ∈ [0, 1])
  - Measures the divergence between the protected ($ROC_s$) and non-protected group ($ROC_{\bar{s}}$) curves across all possible thresholds $t \in [0,1]$ of FPR and TPR

$$\int_0^1 \mid ROC_s(t) - ROC_{\bar{s}}(t) \mid \, dt$$

  - E.g.,



# Evaluation

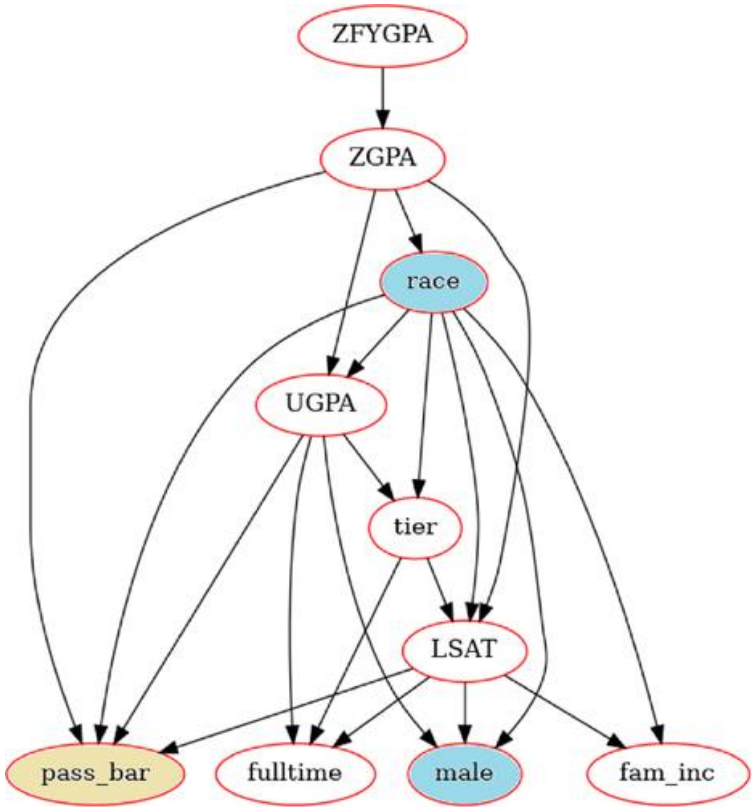
- Datasets

| Datasets | #Instances | #Instances (cleaned) | #Attributes | Protected attribute | Class label | IR (+:-) |
|---|---|---|---|---|---|---|
| Law school | 20,798 | 20,798 | 12 | Race | Pass the bar exam | 8.07:1 |
| PISA | 5,233 | 3,404 | 24 | Gender | Reading score | 1.35:1 |
| Studden academics | 131 | 131 | 22 | Gender | ESP | 3.70:1 |
| Student performance | 649 | 649 | 33 | Gender | Final grade | 5.49:1 |
| xAPI-Edu-Data | 480 | 480 | 17 | Gender | Grade level | 2.78:1 |

- Binarize class labels:
  - PISA dataset: *reading score* {<500, ≥500} ~ {low, high}
  - Student academics: *ESP* (end semester percentage) {pass, good-and-higher}
  - Student performance dataset: *final grade* {<10, ≥10} ~ {fail, pass}
  - xAPI-Edu-Data: *grade level* {Low,Medium−High}
- 70% of data for training and 30% for testing (single split)

# Datasets (1/2)

- Bayesian network[1]
  - If there is any direct/indirect edge from any protected attribute to the class attribute, we may infer that the dataset is biased w.r.t. the specific protected attribute
- Law school dataset:
  - The bar exam's result is conditionally dependent on the law school admission test (*LSAT*) score, undergraduate grade point average (*UGPA*) and *Race*

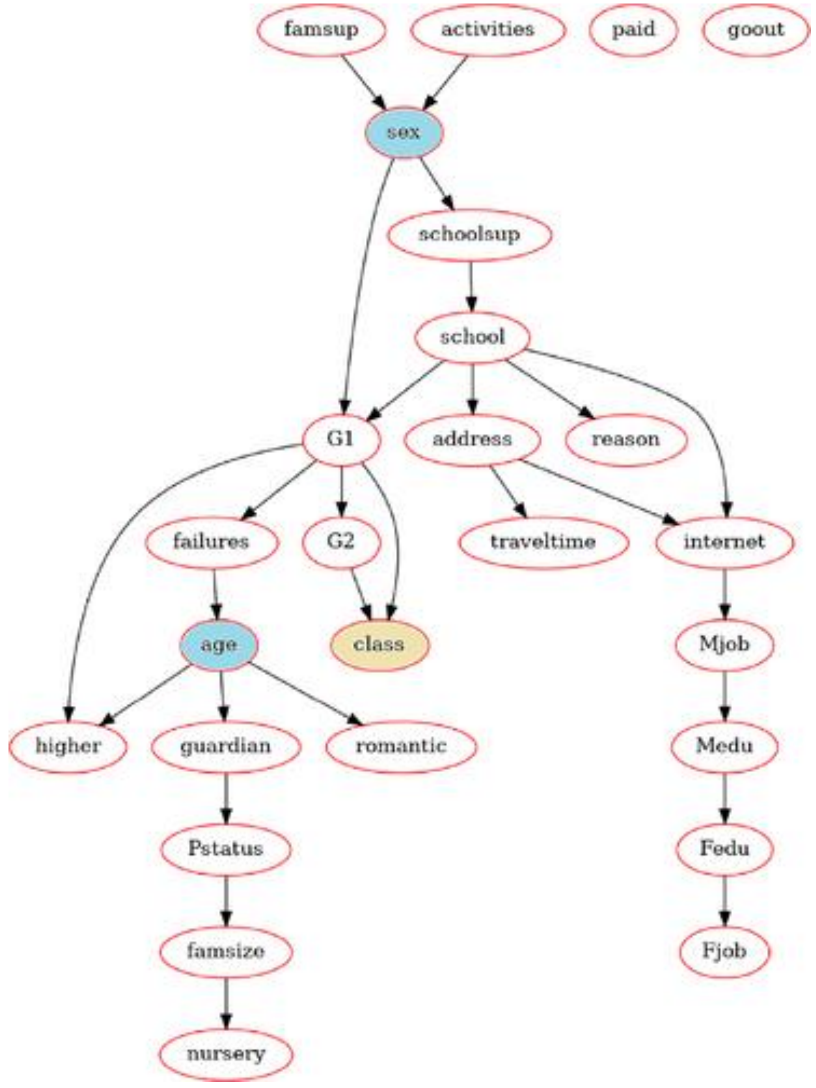



**Law school**: Bayesian network (class label: *pass_bar*, protected attributes: *male, race*)

[1] Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., & Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1452. https://doi.org/10.1002/widm.1452

# Datasets (2/2)

- Student performance dataset:
  - The class label attribute is conditionally dependent on the grade G2




**Student performance-Portuguese subject**: Bayesian network (class label: class, protected attributes: age, sex)

# Evaluation setups

- Predictive models
  - Traditional models
    - Decision Tree
    - Naive Bayes
    - Multi-layer Perceptron
    - Support Vector Machines
  - Fairness-aware models
    - Agarwal’s: reduces the fair classification to a sequence of cost-sensitive classification problems with the lowest (empirical) error subject to the desired constraints
    - AdaFair: updates the weights of the instances in each boosting round

# Experimental results (1/4)

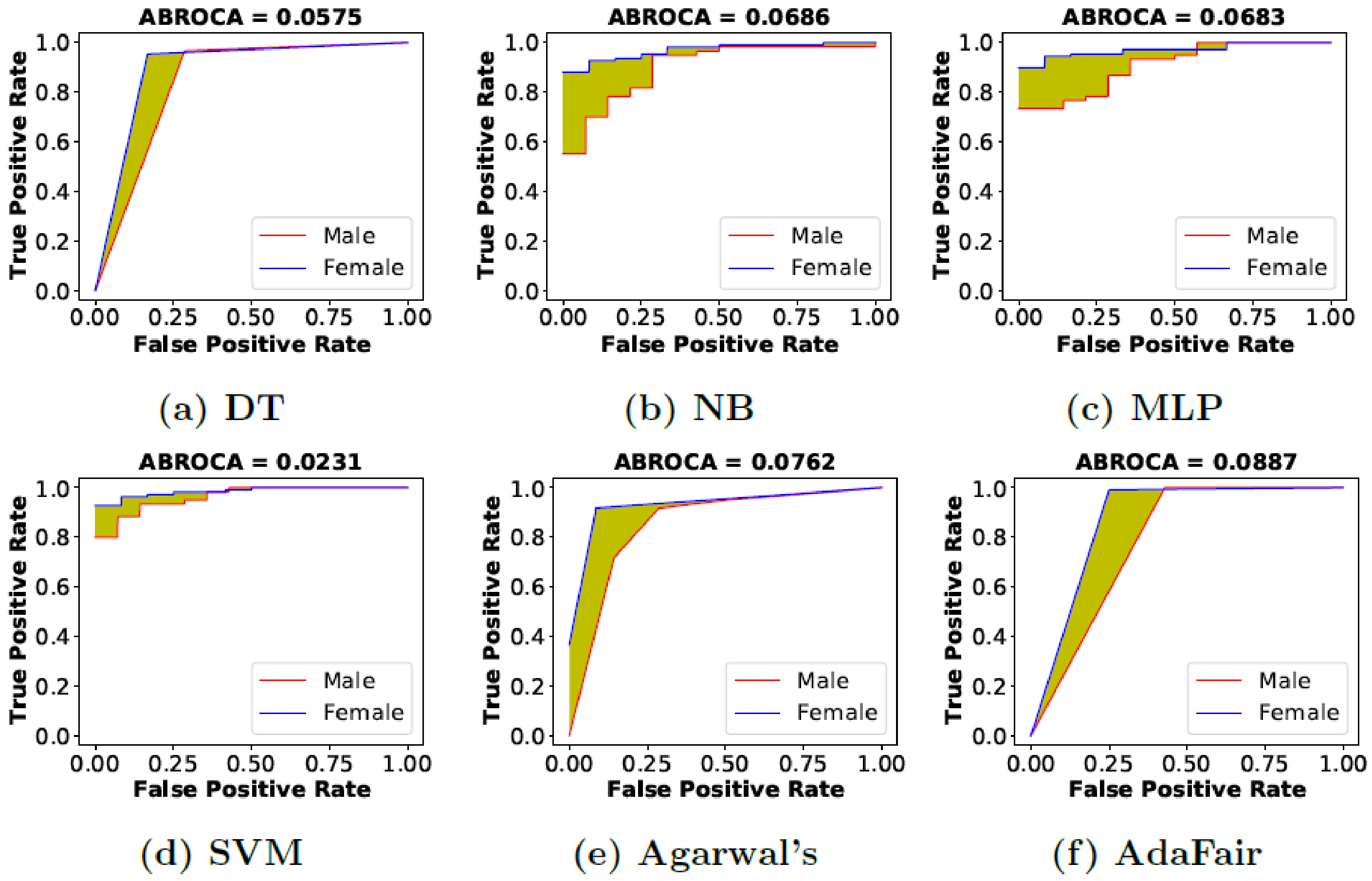
- ML model's accuracy variation over each value of the protected attribute

| Measures | DT | NB | MLP | SVM | Agarwal's | AdaFair |
|---|---|---|---|---|---|---|
| Accuracy | 0.9333 | 0.8974 | 0.9077 | 0.9231 | 0.8923 | **0.9487** |
| Balanced accuracy | **0.8639** | 0.8595 | 0.7840 | 0.7441 | 0.8565 | 0.8240 |
| Statistical parity | -0.0382 | -0.0509 | -0.0630 | **0.0151** | -0.0209 | -0.0255 |
| Equal opportunity | 0.0125 | 0.0174 | 0.03 | 0.0183 | 0.0176 | **0.0092** |
| Equalized odds | 0.1316 | 0.2198 | **0.1252** | 0.3279 | 0.2200 | 0.1877 |
| Predictive parity | **0.0456** | 0.0591 | 0.0601 | 0.0944 | 0.0577 | 0.0639 |
| Predictive equality | 0.1190 | 0.2024 | **0.0952** | 0.3095 | 0.2024 | 0.1786 |
| Treatment equality | 2.0 | 7.5 | **0.3333** | 0.5 | 9.75 | **0.3333** |
| ABROCA | 0.0575 | 0.0686 | 0.0683 | **0.0231** | 0.0762 | 0.0887 |

**Student performance dataset**: performance of predictive models

# Experimental results (2/4)

- ABOCA is the measure with the lowest variability across predictive methods and datasets



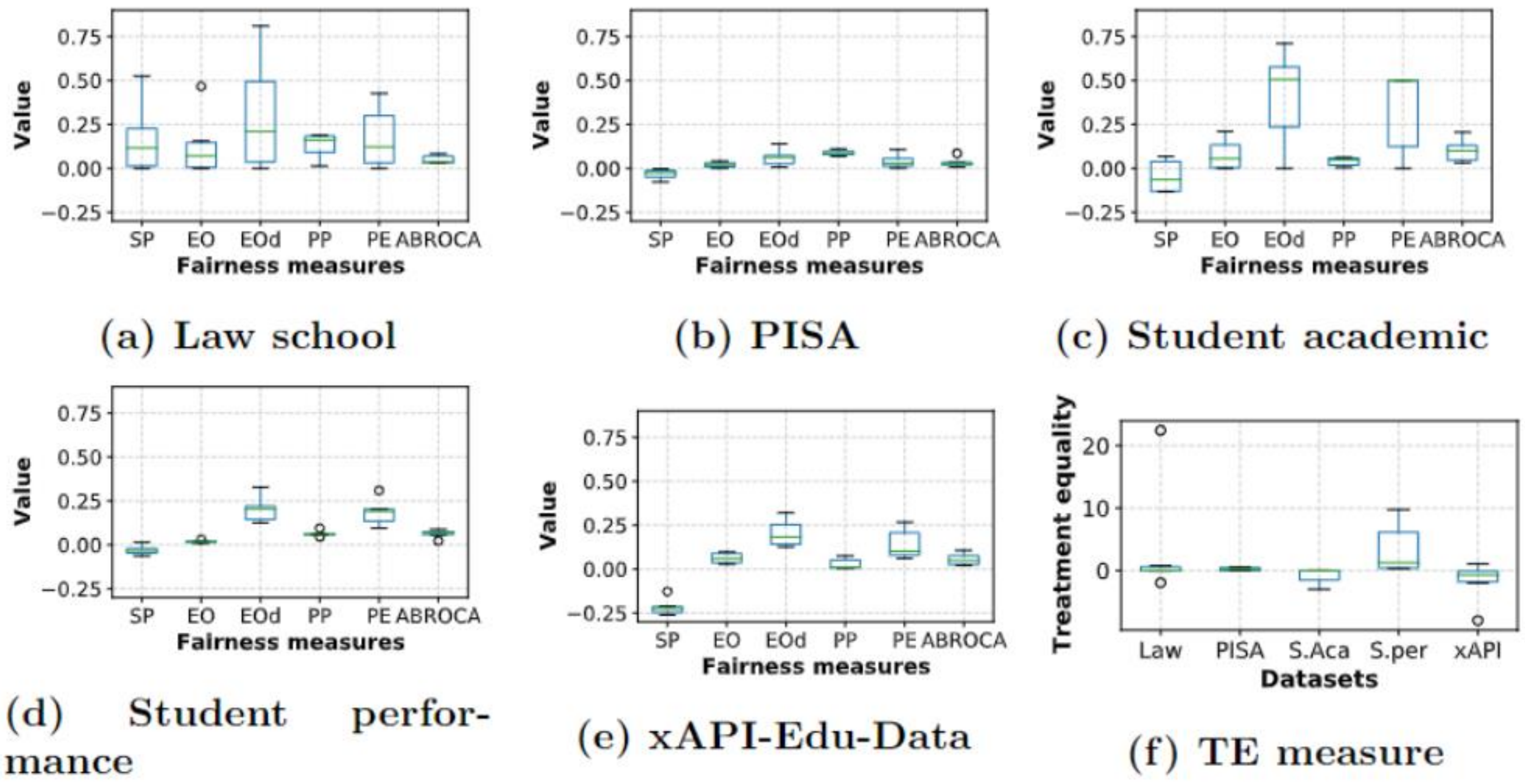

(a) DT (b) NB (c) MLP

(d) SVM (e) Agarwal's (f) AdaFair

**Student performance dataset**: ABROCA slice plots

# Experimental results (3/4)

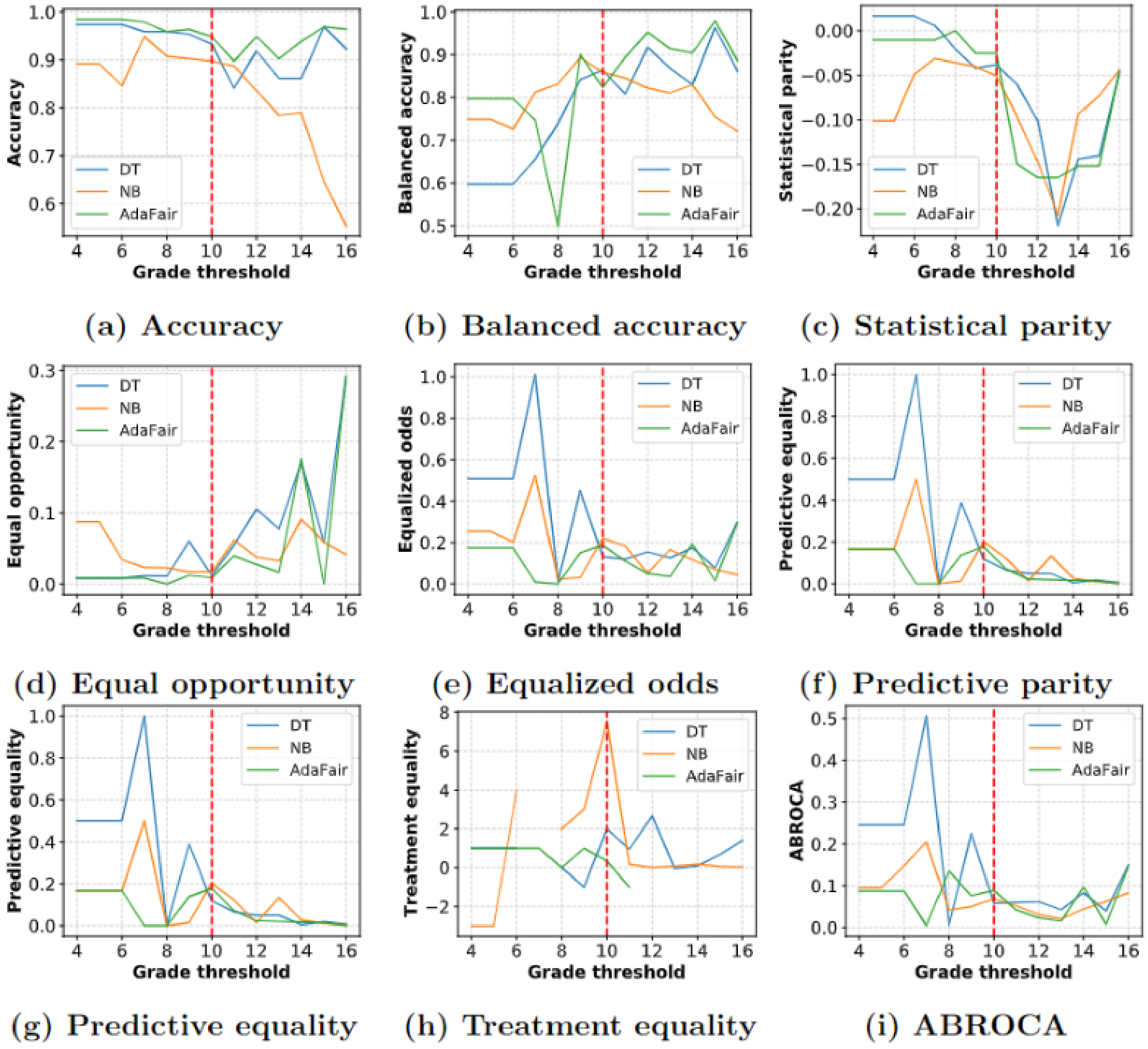
- Equal opportunity and predictive parity also have a slight variation across methods and datasets.
- Equalized odds can represent two measures equal opportunity and predictive equality
- Treatment equality has a very wide range of values (the value may not be bounded)




Variation of fairness measures

# Experimental results (4/4)

- Effect of varying grade threshold on fairness
  - All fairness measures are affected by the grade threshold
  - When the grade threshold is gradually increased, the predictive models tend to be fairer (equalized odds, predictive equality, and ABROCA)




Accuracy and fairness interventions with varying grade threshold on Student performance dataset (Decision Tree)

# Conclusion and outlooks

- We evaluate 7 popular group fairness measures for student performance prediction problems.
- The experimental results reflect variations and correlations of fairness measures across datasets and predictive models.
- The choice of fairness measures is important, and it should be based on the fact that all genders and races, etc., have an equal opportunity.
- Choosing the threshold is an important factor contributing to ensuring fairness in the output of the ML models.
- We plan to extend our evaluation of fairness w.r.t. multiple protected attributes and individual fairness notions.

ECML PKDD 2022

# Thank you for your attention!

# Question?

LernMINT

The work is supported by LernMINT - The Lower Saxony doctoral program (Ministry for Science and Culture, Lower Saxony, funding period 2019-2024)
https://lernmint.org/

tai@l3s.de
tailequy.github.io
tailequy


https://aiml-research.github.io/