**Explaining Model Behavior**

- **LACE - Individual** prediction
- **DivExplorer - Data subgroup**

Model agnostic → applicable for a generic classifier

**LACE - Explaining Individual Predictions**

Given a single prediction, we explain the reasons behind it.

- Quantitative insight → prediction difference
- Qualitative insight → local rules

It can reveal if model predictions are based on protected attributes.

**DivExplorer**

- **Automatic identification** of subgroups with divergent behavior
- **Local contribution** to divergence
- **Global contribution** to divergence

**Try it at divexplorer.org!**

**Try it at DIVEXPLORER.ORG!**

**ON THE NEED OF XAI**

Black box models → hide from users the reasons behind predictions

Why Explainable AI?

- **FAIRNESS**
- **TRUST**
- **MODEL VALIDATION**
- **DEBUG & ERROR ANALYSIS**

**Explainable AI and Fairness**

Eliana Pastor, Elena Baralis, Luca de Alfaro

POLITO  POLITO  UCSC

**DivExplorer - Explaining Subgroups**

Evaluation typically considers overall performance

However, the model behavior can differ across data subgroups

DivExplorer

- Automatic identification of subgroups with divergent behavior
- Local contribution to divergence
- Global contribution to divergence

**Try it at divexplorer.org!**

**Evaluation typically considers overall performance**

However, the model behavior can differ across data subgroups

**DivExplorer**

- Automatic identification of subgroups with divergent behavior
- Local contribution to divergence
- Global contribution to divergence

**Try it at divexplorer.org!**

**Why Explainable AI?**

**LACE - EXPLAINING INDIVIDUAL PREDICTIONS**

- Quantitative insight → prediction difference
- Qualitative insight → local rules

It can reveal if model predictions are based on protected attributes.