Arg-XAI: a Tool for Explaining Machine Learning Results

15 maggio 2023

A.D. 1308
unipg
UNIVERSITÀ DEGLI STUDI
DI PERUGIA

Overview

- Some motivations
- Tool demo
- Validation
- Future work

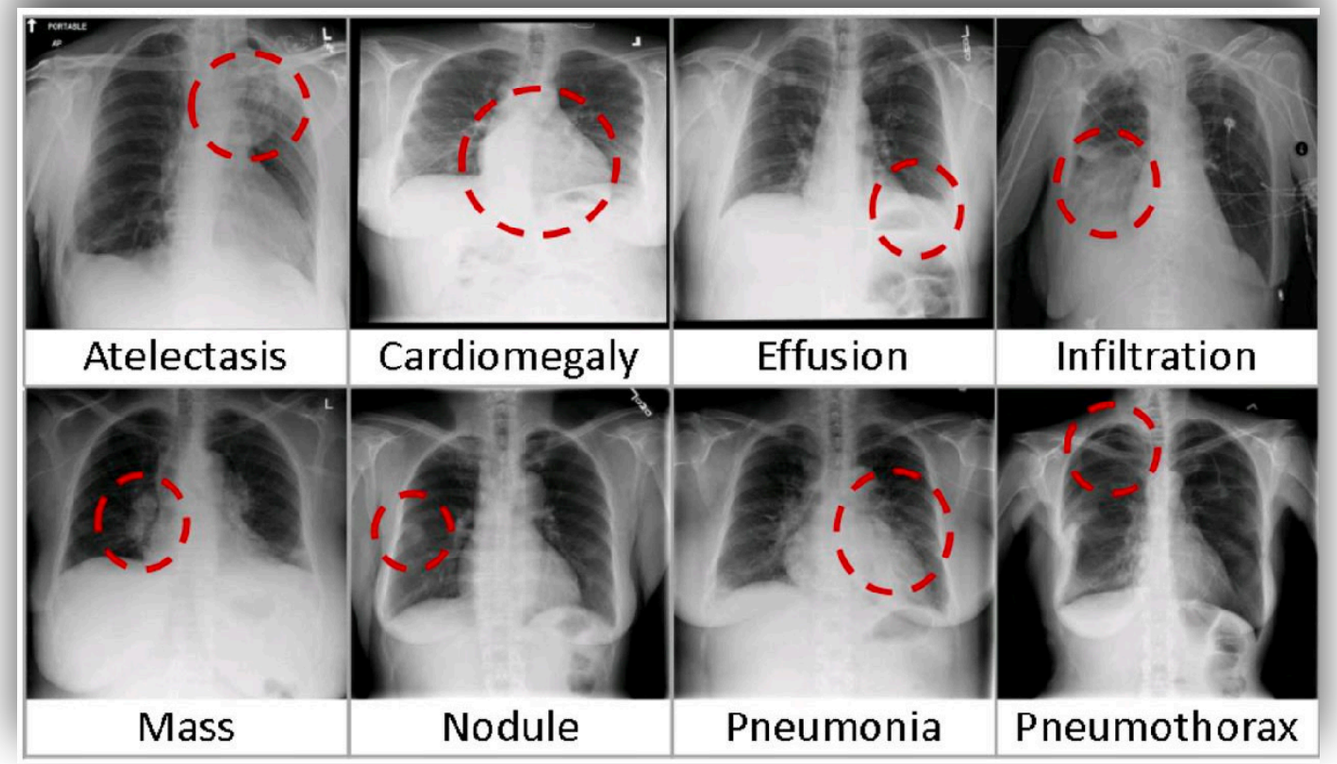
Titanic: a Class Prediction Example

1. Get a dataset
2. Train the model
3. ???
4. Profit

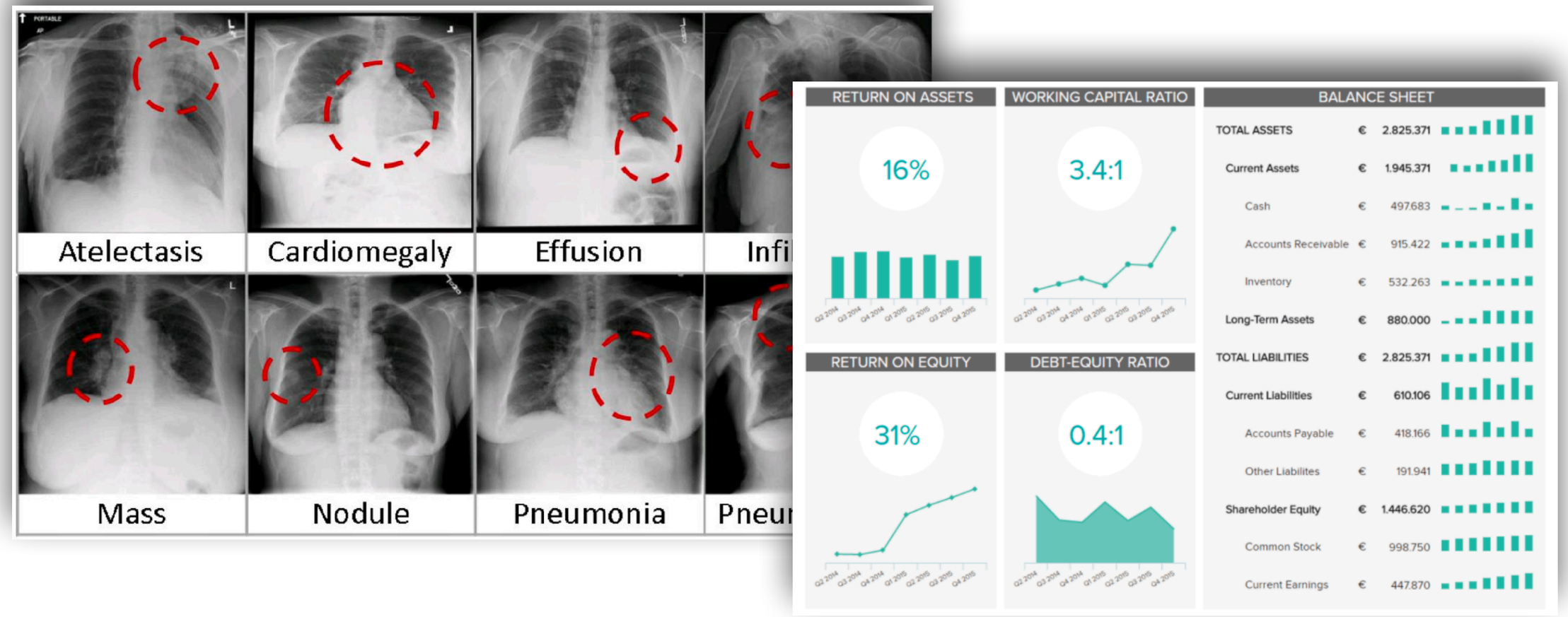
Feature	Values	Type	Description
Pclass	$1, 2, 3$	categorical	Ticket class
sex	$0, 1$	categorical	passenger gender
SibSp	$0 - 8$	categorical	# of siblings/spouses
Parch	$0 - 6$	categorical	# of parents/children
Embarked:	C, Q, S	categorical	port of embarkation
Survived:	$0, 1$	categorical	passenger survival
Age	$0.17 - 76$	numerical	passenger age
Fare	$0 - 512$	numerical	passenger fare

- Anna survived
- … but why?

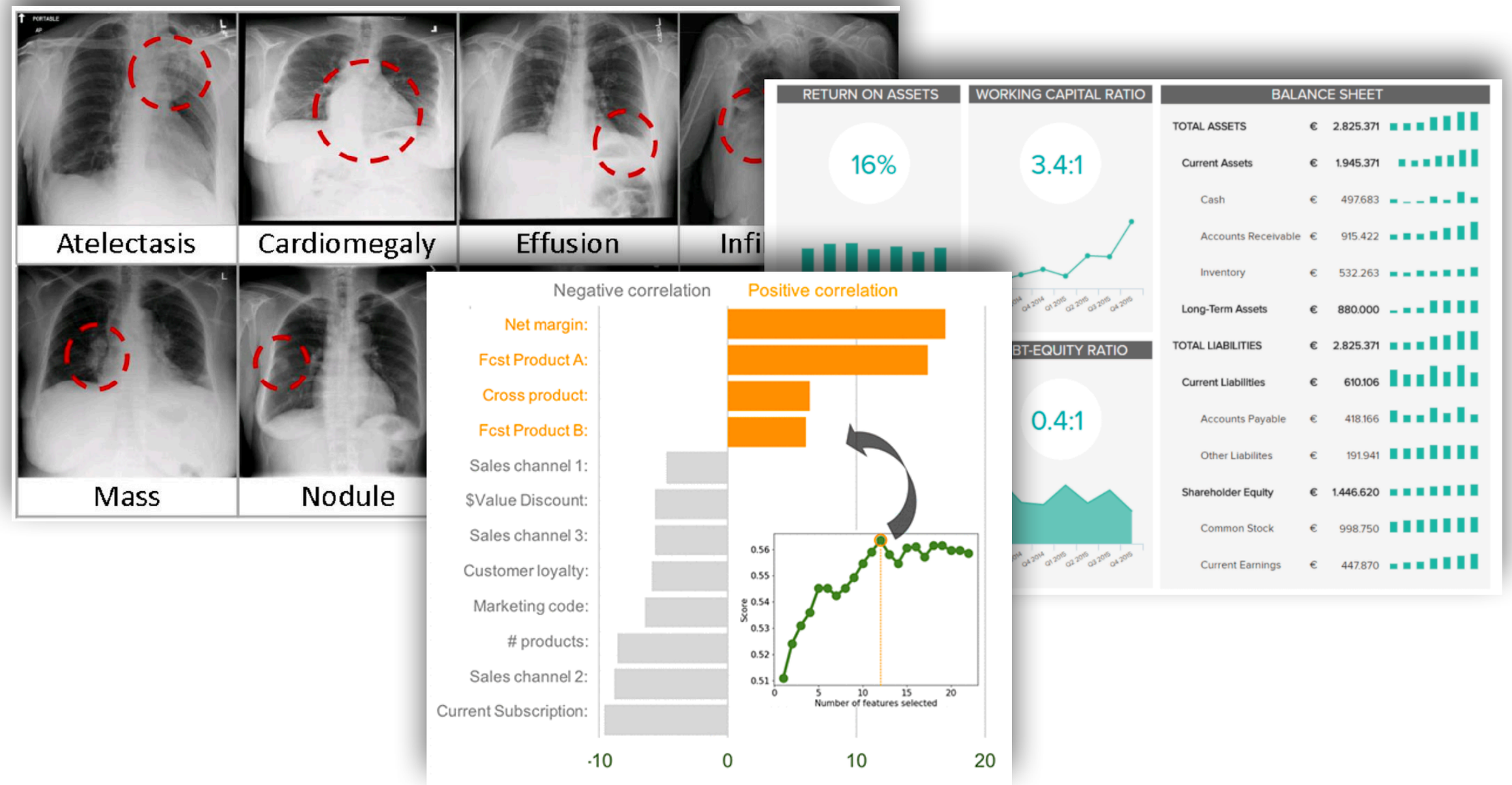
Explanation in Critical Fields



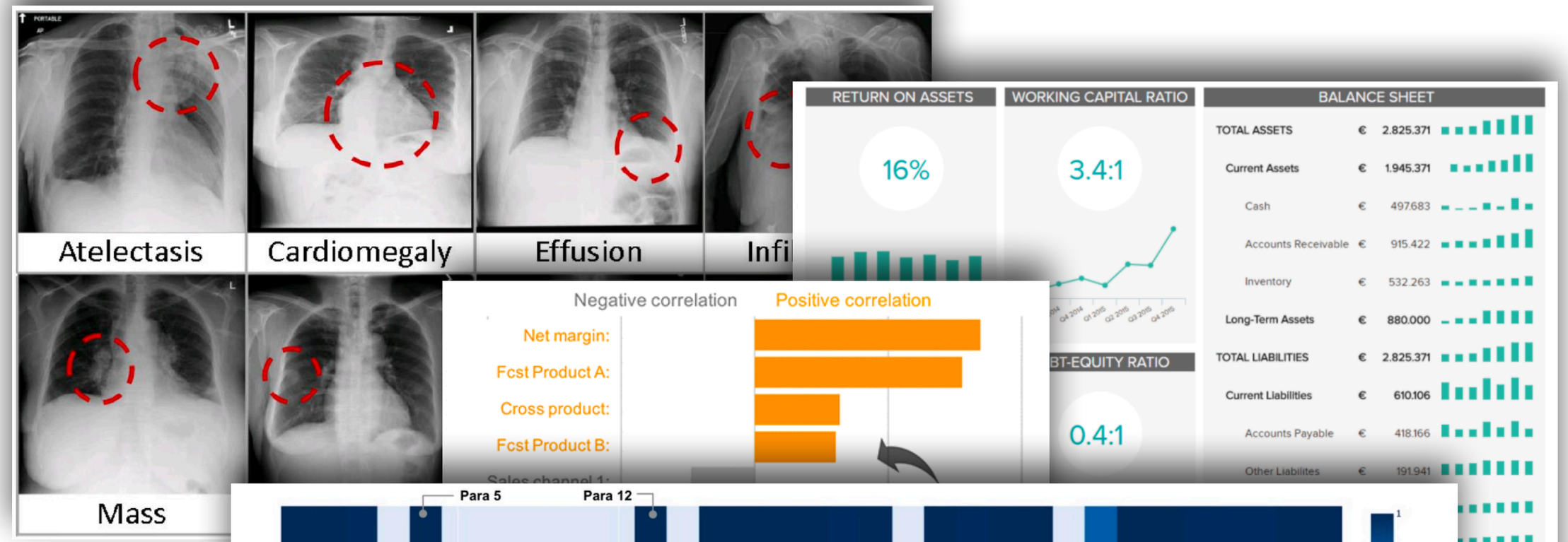
Explanation in Critical Fields

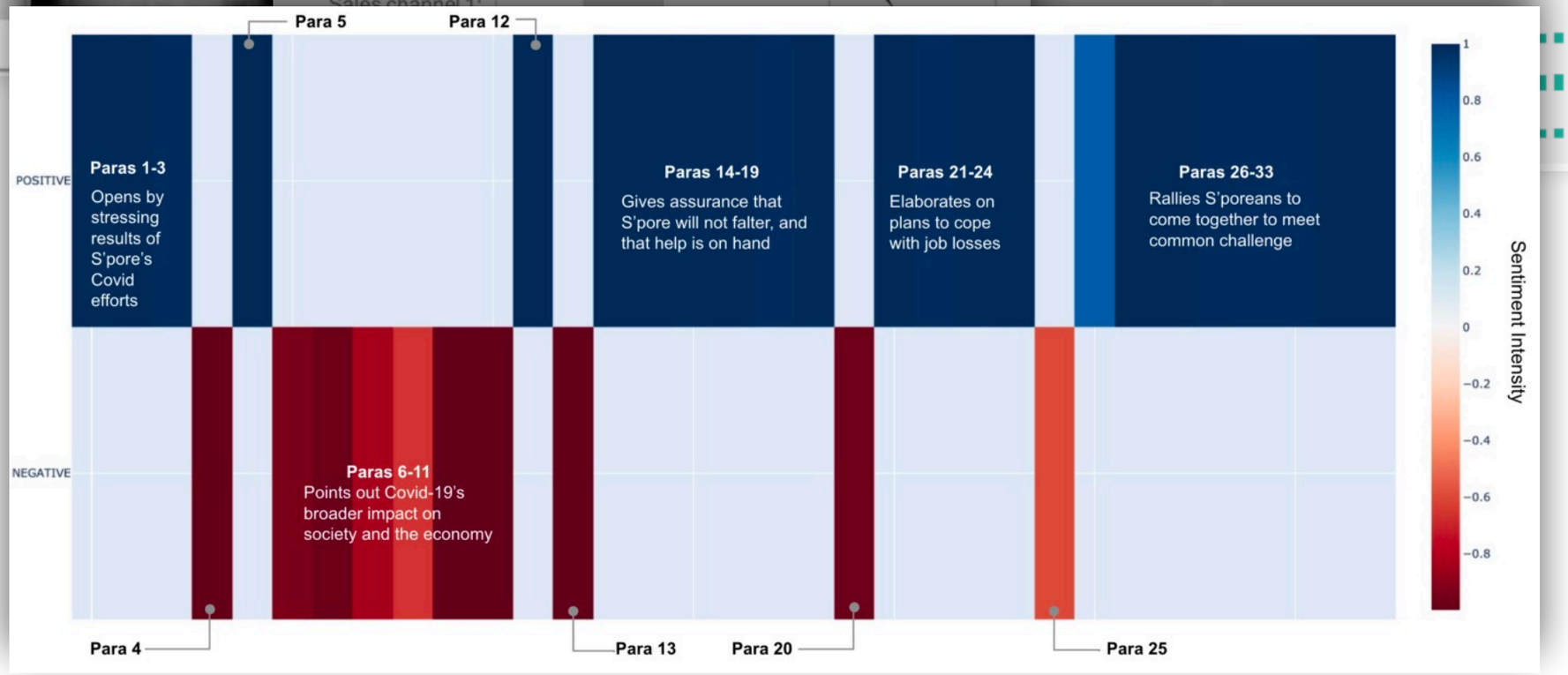


Explanation in Critical Fields



Explanation in Critical Fields





In a nutshell…

- We want to explain the results of a classifier
- Basic idea of the tool:
  - Transform the dataset into a Bipolar Argumentation Framework
  - Use argumentation semantics to select acceptable arguments
  - Produce an explanation tree

In a nutshell…

- We want to explain the results of a classifier
- Basic idea of the tool:
  - Transform the dataset into a Bipolar Argumentation Framework
  - Use argumentation semantics to select acceptable arguments
  - Produce an explanation tree

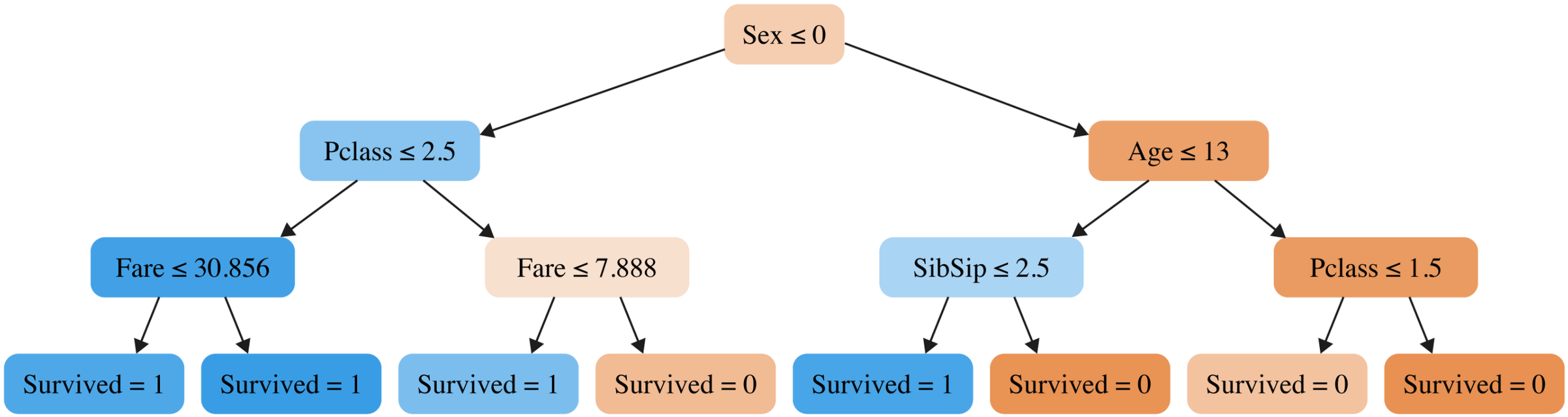
Demo

arg-xai.dmi.unipg.it

Validation A: Decision Tree

- Semi-stable extension:

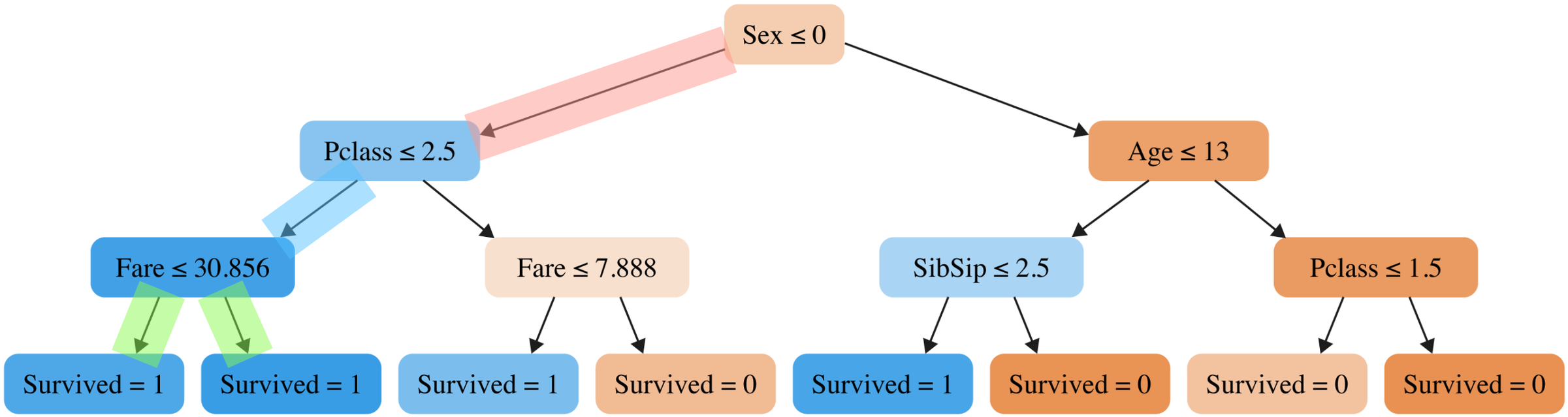
Sex=0, Pclass=1, Fare≥10.4812, Survived=1,
Age<0.96, Embarked=C, Parch=1, SibSp=1



Validation A: Decision Tree

- Semi-stable extension:

Sex=0, **Pclass=1**, **Fare≥10.4812**, **Survived=1**,
Age<0.96, Embarked=C, Parch=1, SibSp=1



Validation B: Rule-Based Classifier

- Semi-stable extension:

  thall=2, caa=0, slp=2, exng=0, sex=0, cp=2, chol<245.5, oldpeak<1.7, thalach≥147.5, trtbps<107, age<54.5, restecg=1, **Heart Attack=1**, fbs=1

- We use RIPPER to derive a set of rules for **Heart Attack=1**

  thall=2 ∧ caa=0 ∧ slp=2, exng=0 ∧ caa=0 ∧ sex=0, exng=0 ∧ thall=2 ∧ cp=2, caa=0 ∧ thall=2 ∧ sex=1, trtbps=130.0-138.0 ∧ chol=187.0-207.0

Validation B: Rule-Based Classifier

- Semi-stable extension:

thall=2, caa=0, slp=2, exng=0, sex=0, cp=2, chol<245.5, oldpeak<1.7, thalach≥147.5, trtbps<107, age<54.5, restecg=1, **Heart Attack=1**, fbs=1

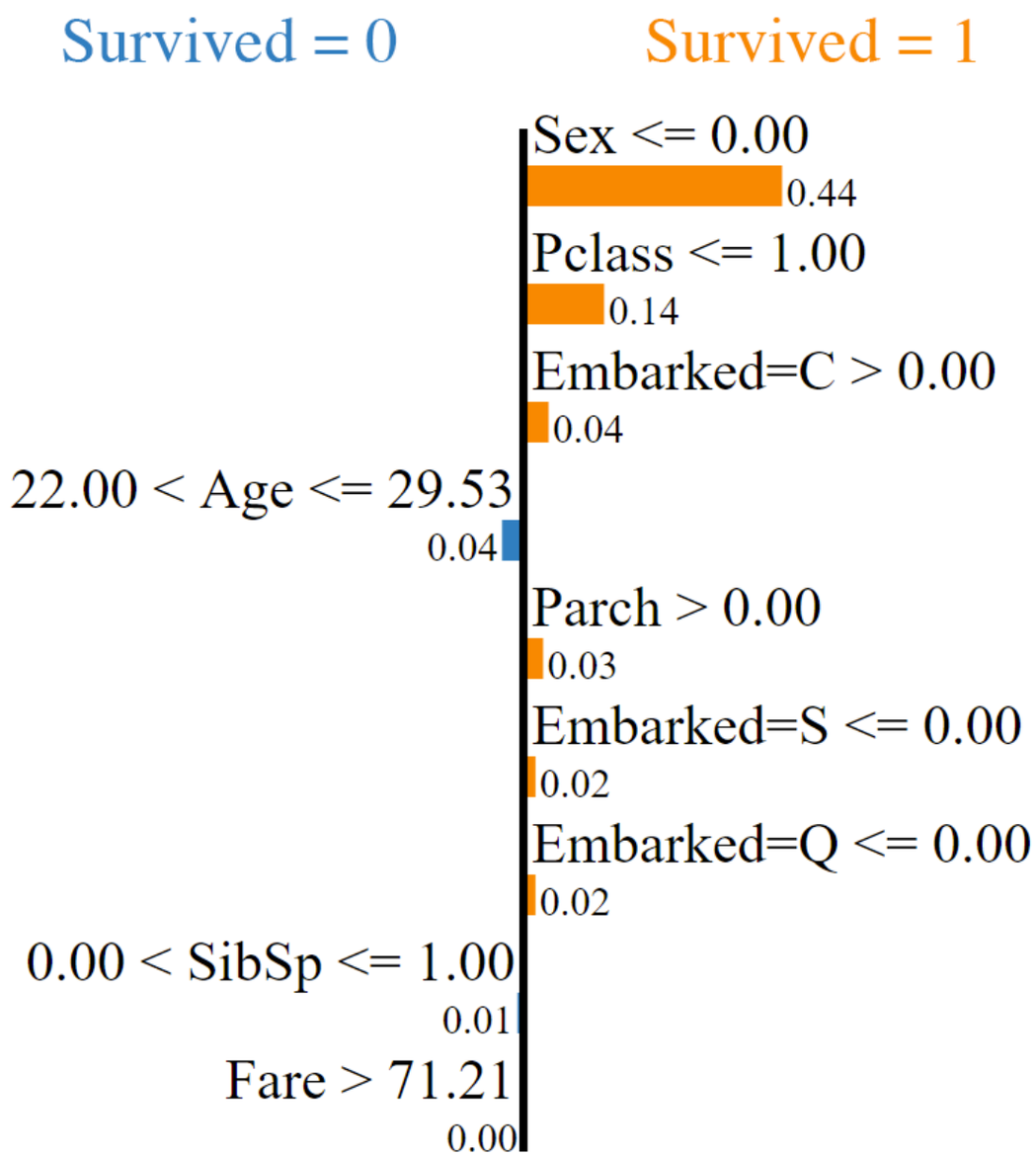
- We use RIPPER to derive a set of rules for **Heart Attack=1**

thall=2 ∧ caa=0 ∧ slp=2, exng=0 ∧ caa=0 ∧ sex=0, exng=0 ∧ thall=2 ∧ cp=2, caa=0 ∧ thall=2 ∧ sex=1, trtbps=130.0-138.0 ∧ chol=187.0-207.0

Validation C: LIME

- Semi-stable extension:

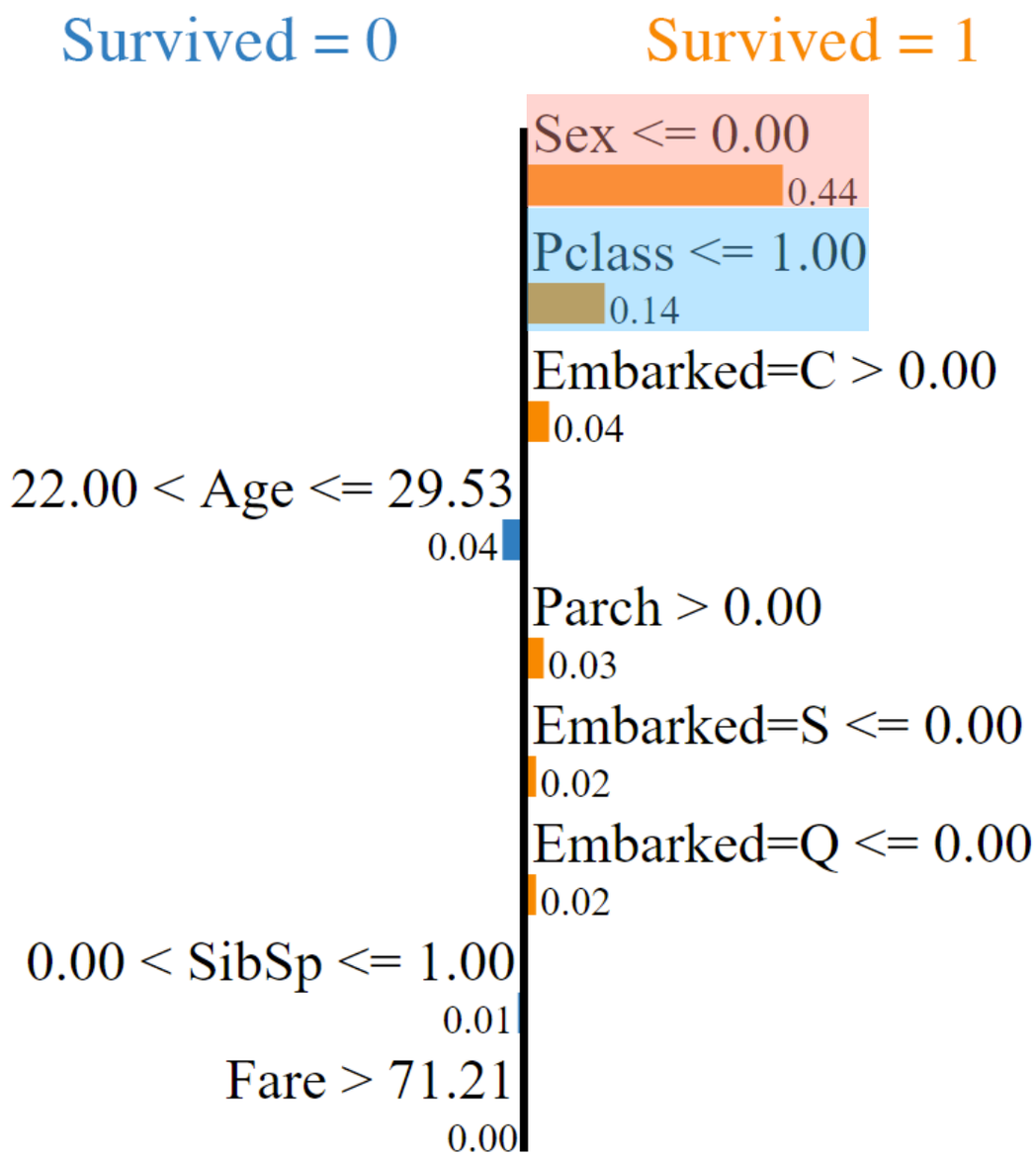
Sex=0,
Pclass=1,
Fare≥10.4812,
Survived=1,
Age<0.96,
Embarked=C,
Parch=1,
SibSp=1



Validation C: LIME

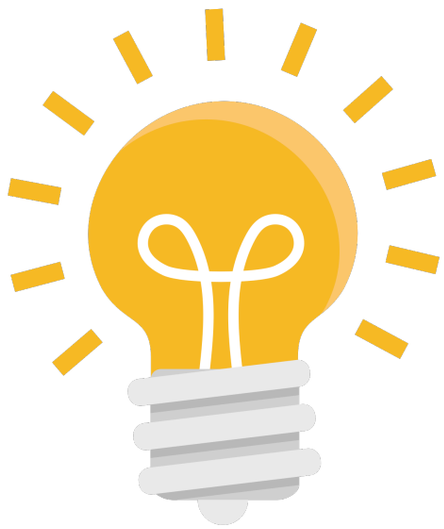
- Semi-stable extension:

Sex=0,
Pclass=1,
Fare≥10.4812,
Survived=1,
Age<0.96,
Embarked=C,
Parch=1,
SibSp=1



Future Work

- Three research lines:
  - Derive causality between features (break symmetric relations)
  - Techniques for detecting spurious correlation
  - Devise efficient algorithms for probabilistic semantics



THANK YOU FOR YOUR ATTENTION

Arg-XAI: a Tool for Explaining Machine Learning Results

A.D. 1308
unipg
UNIVERSITÀ DEGLI STUDI
DI PERUGIA