

Des explications garanties pour l'IA

Jeudi 22 octobre 2026

IA^2 2026 – Institut d'Automne en Intelligence Artificielle
Le Touquet-Paris-Plage, France
CRIL, Université d'Artois & CNRS

Gilles AUDEMARD
Steve BELLART
Louenas BOUNIA
Mathis KONARSKI
Jean-Marie LAGNIEZ
Pierre MARQUIS
Mehdi SABIRI
Nicolas SZCZEPANSKI

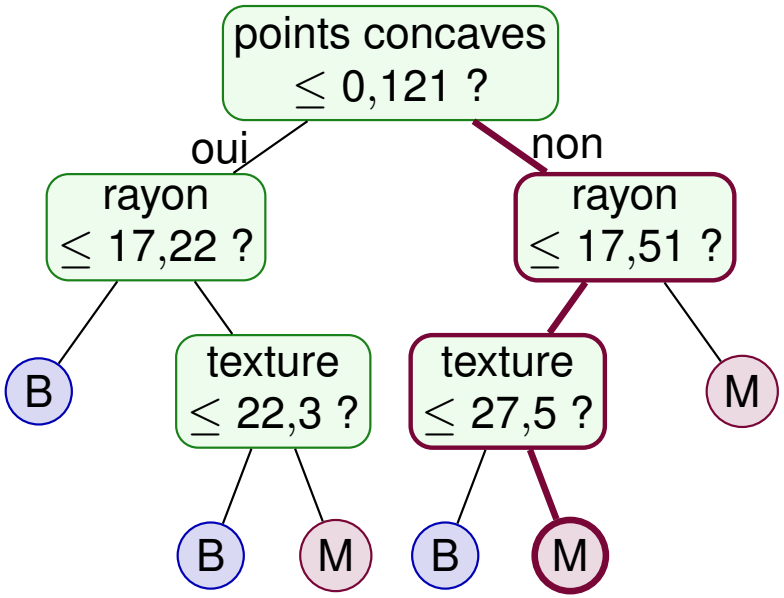








Abductive ou contrastive ? Un exemple réel



B : bénigne, M : maligne (à gauche : oui)

Une patiente : rayon 15,2 ; points concaves 0,198 ; texture 30,15
$\rightarrow$ **tumeur maligne**

Pourquoi ? (abductive)

Points concaves $>$ 0,121 et texture $>$ 27,5 **suffisent** : le rayon ne compte pas

Pourquoi pas bénigne ? (contrastive)

Avec une texture $\leq$ 27,5, et le reste inchangé, elle serait **bénigne**

Arbre appris sur Breast Cancer Wisconsin (569 tumeurs, 94 % de bonnes prédictions en test) : rayon, points concaves et texture des noyaux cellulaires. Explications calculées avec PyXAI.

Pourquoi Alice a-t-elle obtenu son prêt ?

Alice : revenu de 2 200 €, salariée, mariée $\rightarrow$ **prêt accordé**

Méthode	**Explication**
PyXAI (formelle)	**revenu** $>$ **2 000 €**
SHAP, LIME (agnostiques)	revenu 0,2 ; emploi 0,7 ; mariage 0,7

- **PyXAI** : tout client qui gagne plus de 2 000 € obtient le prêt, c'est **garanti**
- **SHAP, LIME** : « salariée et mariée » semble décisif, pourtant Bob (1 900 €, salarié, marié) est refusé

Partout où l'IA décide, une explication doit être **fidèle au modèle**

SHAP et LIME peuvent se tromper

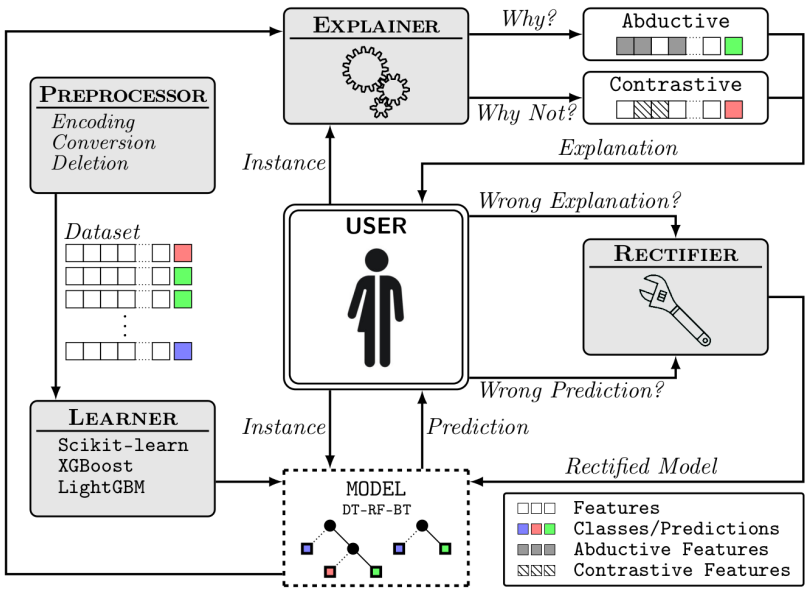
Une explication est **incorrecte** s'il existe un **contre-exemple** : un cas qui la respecte, mais que le modèle classe autrement.

Jeu de données	LIME	Anchor	SHAP
adult	61,3 %	80,5 %	70,7 %
lending	24,0 %	3,0 %	17,0 %
rcdv	94,1 %	99,4 %	85,9 %
compas	71,9 %	84,4 %	60,4 %
german	85,3 %	99,7 %	63,0 %

Part d'explications incorrectes

PyXAI : des explications **correctes par construction** (garantie formelle)

PyXAI en bref



- **Modèles à base d'arbres** : arbres de décision, forêts aléatoires, arbres boostés
- scikit-learn, XGBoost, LightGBM
- **Pourquoi ?** et **pourquoi pas ?** : explications abductives et contrastives
- **Rectification** du modèle avec les connaissances du domaine
- Interface graphique, plus de 10 notebooks

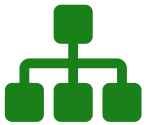
`pip install pyxai` (version 2.0.1, licence MIT)

Publications autour de PyXAI

Année	Lieu	Titre
2026	IJCAI	Computing Coverage-Based Prime Implicant Explanations for Tree-Based Models
2026	KR	A Rectification-Based Approach for Distilling Boosted Trees into Decision Trees
2025	EGC	Une interface XAI pour des modèles d'apprentissage automatique à base d'arbres
2024	ECAI	Designing an XAI Interface for Tree-Based ML Models
2024	IJCAI	On the Computation of Example-Based Abductive Explanations for Random Forests
2024	IJCAI	Deriving Provably Correct Explanations for Decision Trees: The Impact of Domain Theories
2024	IJCAI (démo)	**PyXAI: An XAI Library for Tree-Based Models**
2023	AISTATS	Computing Abductive Explanations for Boosted Trees
2023	ECAI	On Contrastive Explanations for Tree-Based Classifiers
2023	ECAI	Rectifying Binary Classifiers
2023	IJCAI	Computing Abductive Explanations for Boosted Regression Trees
2023	EGC	PyXAI : calculer en Python des explications pour des modèles d'apprentissage supervisé
2022	DKE	On the Explanatory Power of Boolean Decision Trees
2022	IJCAI	On Preferred Abductive Explanations for Decision Trees and Random Forests
2022	AAAI	Trading Complexity for Sparsity in Random Forest Explanations
2021	KR	On the Computational Intelligibility of Boolean Classifiers
2020	KR	On Tractable XAI Queries based on Compiled Representations

Auteurs : G. Audemard, S. Bellart, L. Bounia, S. Coste-Marquis, F. Koriche, J.-M. Lagniez, P. Marquis, M. Sabiri et N. Szczepanski (CRIL).

Perspectives



Réseaux de neurones

Des explications garanties au-delà des arbres



CNN

Expliquer des modèles qui travaillent sur des images



Valeurs manquantes

Expliquer quand des données sont absentes (NaN), fréquent en santé

Garder la **garantie de correction** sur des modèles et des données plus réalistes