

PyXAI : calculer en Python des explications pour des modèles d'apprentissage supervisé

Gilles Audemard *, Steve Bellart *, Louenas Bounia *
Jean-Marie Lagniez *, Pierre Marquis **, Nicolas Szczepanski ¹ *

*Univ. Artois, CNRS, CRIL, F-62300 Lens

**Institut universitaire de France
nom@cril.fr

Résumé. L'intelligence artificielle explicable est un sous-domaine de l'IA, qui connaît un essor important depuis quelques années. Le but poursuivi est de développer des méthodes et outils pour expliquer les résultats produits par des algorithmes d'IA. Consacrée à cette tâche, PyXAI est une librairie Python permettant de calculer des explications pour des prédictions réalisées à partir de plusieurs modèles d'apprentissage supervisé bien connus à base d'arbres : les arbres de décision (*decision trees*), les forêts aléatoires (*random forests*) et les arbres optimisés (*boosted trees*). PyXAI prend en charge deux bibliothèques d'apprentissage automatique : Scikit-Learn et XGBoost. Plusieurs types d'explication peuvent être calculés : *abductive* (pourquoi cette prédiction ?) et *contrastive* (pourquoi pas une autre prédiction ?). Divers types d'explications abductives sont proposés permettant de réaliser différents compromis taille / temps de calcul.

1 Introduction

L'essor de l'intelligence artificielle et de l'apprentissage automatique (ML pour *machine learning*) à travers ses nombreuses applications (diagnostic médical, reconnaissance vocale, conduite autonome, ...) a conduit au développement rapide de l'intelligence artificielle explicable (XAI pour *eXplainable Artificial Intelligence*). En effet, l'apprentissage automatique produit souvent des modèles de type « boîtes noires » et se pose ainsi un problème de confiance envers les décisions prises. Pour pallier ce problème, il importe de développer de nouvelles approches pour garantir confiance et transparence envers les systèmes d'IA exploitant de tels modèles. Cela passe, en particulier, par la possibilité de produire des explications (aussi appelées raisons) pour les décisions obtenues. Dans cette mouvance, l'Union Européenne a introduit un droit à l'explication dans le droit général de la protection des données (RGPD).

Diverses approches de calcul d'explications, indépendantes du modèle d'apprentissage considéré, ont été proposées. Cependant, il a été montré dans Ignatiev (2020) que les approches les plus populaires basées sur les *explications agnostiques* au modèle, telles que LIME (Ribeiro et al., 2016), Anchors (Ribeiro et al., 2018) et SHAP (Lundberg et Lee, 2017), fournissent un

1. Ce travail a été réalisé dans le cadre de la chaire ANR d'enseignement et de recherche EXPEKCTATION (ANR-19-CHIA-0005-01).