Exploration et application des méthodes d'explicabilité pour l'amélioration des modèles de Machine Learning

par

Mathilde Boucly et Michel Daher Mansour

CentraleSupélec Exed – MS IAC

22/06/2025

Contents

[**Résumé** 3](#_Toc200061781)

[**1.** **Introduction** 4](#_Toc200061782)

[**2.** **Revue sur l’explicabilité** 5](#_Toc200061783)

[**3.** **Méthodologie** 6](#_Toc200061784)

[**4.** **Expérimentations** 7](#_Toc200061785)

[**A.** **Base de donnée : Sécurité Forensique** 7](#_Toc200061786)

[**1.** **Modèles Globales** 7](#_Toc200061787)

[**2.** **Modèles Locales** 7](#_Toc200061788)

[**B.** **Base de donnée :** 7](#_Toc200061789)

[**1.** **Modèles Global** 7](#_Toc200061790)

[**2.** **Modèles Local** 7](#_Toc200061791)

[**5.** **Résultats et Discussions** 8](#_Toc200061792)

[**6.** **Conclusion & Perspectives** 9](#_Toc200061793)

[Référence 10](#_Toc200061794)

# **Résumé**

# **Introduction**

Intro général :

Explicabilité – sujet du projet et intro des deux datasets – méthodologie – section des rapports

# **Revue sur l’explicabilité**

L’explicabilité des modèles d’intelligence artificielle (XAI) constitue aujourd’hui un levier stratégique dans l’implémentation de solutions basées sur l’IA en environnement industriel. Elle répond à des exigences croissantes en matière de **transparence**, de **traçabilité** des décisions, de **conformité réglementaire** (notamment avec le RGPD et l’AI Act européen), mais aussi de **compréhension métier**. L’essor des modèles complexes ou non interprétables « boîtes noires » (Random Forest, XGBoost, réseaux de neurones et du coup le Deep Learning) a entraîné le développement de nombreuses techniques XAI visant à rendre intelligibles leurs prédictions. Ces méthodes peuvent être classées selon deux axes principaux : **l’échelle d’explication** (globale ou locale) et leur **agnosticisme au modèle**.

Parmi les méthodes globales, les premières approches comme les Partial Dependence Plots (PDP), introduites par Friedman en 2001, permettaient de visualiser l’effet moyen d’une variable sur la prédiction. Les Accumulated Local Effects (ALE), proposées par Apley et Zhu en 2016, sont venues corriger les limites des PDP en tenant compte des corrélations entre variables. La méthode LOFO (Leave-One-Feature-Out), utilisée dès 2008 avec les forêts aléatoires par Breiman, quantifie l’importance d’une variable en mesurant la dégradation de performance lorsque celle-ci est exclue du modèle. Ces outils sont adaptés à une vision synthétique du modèle, utile pour des audits, des analyses de robustesse ou la présentation de résultats à des responsables techniques.

Les **méthodes locales**, comme **LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive exPlanations), ICE (Individual Conditional Expectation)** et **Anchors** visentà expliquer des prédictions individuelles. ICE, publiée en 2015 par Goldstein et al., permet de visualiser l’effet d’une variable sur une prédiction pour une observation donnée. LIME, proposée par Ribeiro et al. en 2016, approxime localement le modèle complexe par un modèle interprétable (linéaire ou à base de règles). Tandis que SHAP, développée par Lundberg & Lee en 2017, s’appuie sur la théorie des jeux pour attribuer une contribution équitable à chaque variable. Enfin, Anchors , introduit par Ribeiro en 2018, génère des règles simples de type *if-then* expliquant pourquoi une prédiction tient sous certaines conditions. Ces approches sont particulièrement utiles pour des cas d’usage métiers tels que la **justification de décisions automatisées** (ex. refus de crédit), la **compréhension d’erreurs,** ou encore la **détection de biais.** Ces approches locales sont particulièrement utiles dans des contextes métiers où il est essentiel de justifier une décision automatisée (ex : acceptation de prêt, rejet de candidature), analyser un comportement inattendu du modèle, ou identifier un biais.

Un atout clé dans le monde industriel est que la majorité de ces méthodes sont **modèle-agnostiques**,c’est-à-dire qu’elles fonctionnent indépendamment de l’algorithme utilisé. Cela permet de les intégrer dans des pipelines IA hétérogènes, sans avoir à adapter les explications à chaque nouveau modèle. L’implémentation concrète passe souvent par des outils intégrés (ex. SHAP dans XGBoost, LIME en Python) ou via des dashboards explicatifs accessibles aux métiers.

Pour terminer, une stratégie XAI efficace doit combiner des approches globales pour la gouvernance du modèle, et des approches locales pour l’aide à la décision au cas par cas. Elle doit aussi s’adapter aux différents profils d’utilisateurs, de l’ingénieur data au professionnel métier, en fournissant des explications à la fois **techniquement valides, visuellement claires**, et **contextuellement pertinentes**.

# **Méthodologie**

Ce qu’on va faire, RF ou gradientboosting sur deux dataset,

Expliquer les deux datasets

# **Expérimentations**

1. **Base de donnée : Sécurité Forensique**
2. **Modèles Globales**
   1. PDP
   2. Shap (summary plot)
   3. ALE
   4. LOFO
3. **Modèles Locales**
   1. Lime
   2. ICE
   3. Anchors
   4. Shap (force plot)
4. **Base de donnée :**
5. **Modèles Global**
   1. PDP
   2. Shap (summary plot)
   3. ALE
   4. LOFO
6. **Modèles Local**
   1. Lime
   2. ICE
   3. Anchors
   4. Shap (force plot)

# **Résultats et Discussions**

# **Conclusion & Perspectives**

# Référence