# Analyse Comparative des Méthodes d'Interprétabilité en IA

Études de Cas sur Méthodes Tabulaires, Visuelles, NLP et réseaux de neurones graphiques

Jules Odje

01 février 2025

**Résumé**

Dans ce rapport, nous avons analysé des méthodes d'Explainable Artificial Intelligence (XAI). Nous présentons des techniques qui peuvent être utilisées pour différents types de données : SHAP Values, Counterfactuals (DiCE), Grad-CAM pour les modèles standards, Layer-wise Relevance Propagation (LRP) sur le dataset Iris et des graphiques moléculaires, des méthodes d'analyse des tâches de NLP pour BERT. Chaque méthode a été appliquée dans des exemples pratiques et nous sommes parvenus à des conclusions concernant des méthodes XAI qui composent la littérature.

## Table des matières

# 1 Introduction

## 1.1 Contexte et Problématique

Avec la montée en puissance des applications d'intelligence artificielle. Ce rapport présente des techniques d'explication pour des modèles encore plus compliqués, en incluant des approches pour les données tabulaires, les images, le langage naturel et pour les réseaux de neurones graphiques.

## 1.2 Objectifs

— Comparer et analyser les méthodes XAI sur différents types de données.
— Évaluer leurs forces et faiblesses selon des critères d'interprétabilité.
— Explorer des applications dans des contextes variés, notamment en NLP et chimie moléculaire.

## 1.3 Méthodes Étudiées

Les méthodes suivantes sont comparées dans ce rapport :

— **SHAP Values** : Analyse tabulaire.
— **Counterfactuals (DiCE)** : Explications basées sur des scénarios alternatifs.
— **Grad-CAM** : Visualisation dans des modèles d'images.
— **Analyse de BERT** : Interprétabilité pour des tâches de NLP.
— **Layer-wise Relevance Propagation (LRP)** : Explications pour des relations complexes.
— **Graph Neural Networks (GNN)** : Analyse d'interprétabilité sur des données moléculaires.

# 2 Fondements Théoriques

# 3 Études de Cas

## 3.1 Cas d'Étude 1 : SHAP sur Approbation de Prêts

### 3.1.1 Description des Données

Le dataset comprend des variables telles que l'âge, le salaire, le ratio dette, et le score de crédit.

### 3.1.2 Résultats

— Variables influentes : Salaire et score de crédit.
— SHAP permet une visualisation intuitive des contributions des variables.

## 3.2 Cas d'Étude 2 : Counterfactuals sur Admissions Universitaires

### 3.2.1 Description des Données

Variables clés : GRE, GPA, Expérience de recherche, Classement universitaire.

### 3.2.2 Résultats

— Exemple contrefactuel : "Augmenter le GRE de 5 points et obtenir une recommandation forte."
— DiCE garantit que les recommandations soient réalistes et actionnables.

## 3.3 Cas d'Étude 3 : Grad-CAM sur Classification d'Images

### 3.3.1 Description des Données

Modèle : ResNet50 appliqué à des images de chats.

### 3.3.2 Résultats

— Visualisation des zones importantes : visage, oreilles, corps.
— Grad-CAM offre une interprétation visuelle intuitive pour des modèles d'images.

## 3.4 Cas d'Étude 4 : Analyse de BERT en Question-Answering

### 3.4.1 Description des Données

— Question : "What causes COVID-19 ?"
— Contexte : "COVID-19 is caused by the SARS-CoV-2 virus."

### 3.4.2 Résultats

— Méthode **Layer Integrated Gradients** : Identifie les tokens clés comme "causes" et "SARS-CoV-2."
— Méthode **Layer Conductance** : Met en évidence les contributions des couches du modèle.

## 3.5 Cas d'Étude 5 : LRP sur Dataset Iris

### 3.5.1 Description des Données

Dataset Iris : 150 échantillons, 4 caractéristiques (longueur et largeur des sépales et pétales).

### 3.5.2 Résultats

— Contributions dominantes : Largeur du sépale (42%) et longueur du pétale (33%).
— Validation des propriétés de conservativité et de positivité.

## 3.6 Cas d'Étude 6 : Graph Neural Network (GNN) sur Données Moléculaires

### 3.6.1 Description des Données

— Type : Structures moléculaires (SMILES).
— Molécules : Aspirine, Caféine, Testostérone, Salbutamol.

### 3.6.2 Résultats

— Analyse des relevances pour chaque molécule :
  — **Aspirine** : Zones influentes identifiées autour de la liaison ester.
  — **Caféine** : Maximum de relevance autour des cycles imidazole.
  — **Testostérone** : Contributions équilibrées sur toute la molécule.
  — **Salbutamol** : Fortes variations autour des groupes hydroxyles.
— LRP sur GNN montre une interprétabilité prometteuse pour des structures complexes.

# 4 Analyse Comparative

## 4.1 Comparaison des Méthodes

| **Critère** | **SHAP** | **DiCE** | **Grad-CAM** |
|---|---|---|---|
| Type de données | Tabulaire | Tabulaire | Images |
| Interprétabilité | Globale | Instance | Visuelle |
| Granularité | Feature | Instance | Pixel |

TABLE 1 – Comparaison des méthodes XAI standards

## 4.2 Synthèse sur GNN et Données Moléculaires

— **Pertinence** : Les GNN permettent une compréhension détaillée des interactions chimiques.
— **Limites** : Sensibilité aux hyperparamètres et complexité des graphes.

# 5 Conclusion

Il est important de choisir la méthode selon le type de données et l'objectif. Par moment, il est bien de combiner plusieurs méthodes pour une explication robuste.