XIA — L'intelligence artificielle explicable

Rendu final des projets

JUILLARD Thibaut
thibaut.juillard@univ-lyon2.fr

GHIZLAN Moqim
moqim.ghizlan@univ-lyon2.fr

Université Lumière Lyon 2
M2 MALIA

06 février, 2025

Introduction

Dans un monde où les modèles d'apprentissage automatique sont de plus en plus intégrés dans des domaines critiques tels que la santé, la finance et la sécurité, la question de l'explicabilité des décisions prises par ces modèles devient essentielle. L'interprétation des modèles de Machine Learning, souvent considérés comme des boîtes noires, représente un défi majeur pour assurer la transparence, la confiance et la responsabilité dans les systèmes automatisés. Ce rapport présente une synthèse des travaux réalisés à travers plusieurs projets axés sur l'explicabilité des modèles, chacun explorant des méthodes et des approches variées pour comprendre et interpréter les décisions des algorithmes.

Les projets abordés couvrent un large éventail de techniques d'interprétation, allant des méthodes basées sur des contre-factuels pour les données tabulaires à des approches plus complexes telles que la propagation de la pertinence couche par couche (Layer-wise Relevance Propagation, LRP) appliquée à des réseaux de neurones traditionnels, des modèles de traitement du langage naturel comme BERT, et des réseaux de neurones graphiques (GNN). D'autres projets se sont concentrés sur des techniques de visualisation comme Grad-CAM, ou sur des méthodes d'explicabilité globale telles que SHAP, particulièrement utiles dans des contextes sensibles comme la prédiction des troubles bipolaires.

L'objectif de ce rapport est de détailler le travail accompli dans chacun de ces projets, en expliquant non seulement les méthodologies employées, mais aussi la réflexion sous-jacente aux choix des modèles et des techniques d'interprétation. À travers cette analyse, nous cherchons à mettre en lumière les défis rencontrés, les solutions apportées, et les enseignements tirés de l'application de ces méthodes à des cas d'usage concrets. En explorant la complémentarité de ces approches, nous visons à offrir une vision d'ensemble.

Contexte et Problématiques

L'expansion rapide des modèles d'apprentissage automatique dans divers domaines a mis en évidence un défi fondamental : comprendre comment et pourquoi ces modèles prennent certaines décisions. Bien que les algorithmes de Machine Learning, en particulier ceux reposant sur des architectures complexes comme les réseaux de neurones profonds, affichent des performances impressionnantes, ils manquent souvent de transparence. Ce phénomène, désigné sous le terme de boîte noire , soulève des questions éthiques, juridiques et pratiques, notamment dans des secteurs où la fiabilité des décisions est essentielle, tels que la santé, la finance ou la justice.

L'explicabilité des modèles vise à répondre à ce besoin de transparence en fournissant des explications claires sur les prédictions des algorithmes. Cela permet non seulement d'accroître la confiance des utilisateurs envers les systèmes automatisés, mais aussi de repérer d'éventuels biais ou erreurs dans les modèles. Cependant, trouver un équilibre entre performance et interprétabilité reste un défi majeur. En effet, les modèles les plus performants sont souvent les moins explicables, ce qui soulève des problématiques complexes quant à leur adoption dans des contextes sensibles.

Ce rapport s'inscrit dans ce cadre en explorant différentes méthodes d'interprétation des modèles d'apprentissage automatique. Chaque projet présenté illustre des approches variées pour rendre les modèles plus transparents, tout en mettant en lumière les défis spécifiques rencontrés, tels que la complexité des données, la diversité des algorithmes, et la nécessité d'adapter les techniques d'explicabilité aux cas d'usage particuliers. L'objectif

est de comprendre les forces et les limites de ces méthodes afin de favoriser un usage plus responsable et éclairé des technologies d'intelligence artificielle.

Méthodologie Générale

La méthodologie générale adoptée pour ces projets repose sur une approche rigoureuse qui combine des techniques d'analyse de données, de modélisation et d'interprétation des résultats. Le processus débute par une phase d'exploration des données, visant à comprendre la structure des ensembles de données utilisés, à identifier les variables clés et à repérer d'éventuelles anomalies. Cette étape comprend des analyses statistiques descriptives, des visualisations graphiques et des techniques de prétraitement des données, telles que la normalisation et la gestion des valeurs manquantes.

La sélection des modèles de Machine Learning se fait en fonction de la nature des données et des objectifs spécifiques de chaque projet. Divers algorithmes ont été employés, allant des modèles supervisés classiques comme les forêts aléatoires (Random Forest) et les machines à vecteurs de support (SVM), jusqu'à des architectures plus complexes telles que les réseaux de neurones profonds et les modèles de traitement du langage naturel (BERT). L'entraînement des modèles s'appuie sur des techniques de validation croisée pour garantir la robustesse des performances, en évaluant des métriques telles que la précision, le rappel, la F-mesure et l'aire sous la courbe ROC (AUC-ROC).

L'interprétation des modèles est une partie essentielle de la méthodologie. Plusieurs approches ont été mises en œuvre, y compris des méthodes basées sur l'analyse de l'importance des caractéristiques, la propagation de la pertinence couche par couche (Layer-wise Relevance Propagation), les explications par contre-factuels et des techniques de visualisation comme Grad-CAM. Chaque méthode a été sélectionnée en fonction de sa capacité à fournir des explications claires et pertinentes pour les modèles et les jeux de données concernés.

Enfin, une phase d'analyse critique des résultats permet d'évaluer non seulement la performance des modèles, mais aussi la qualité des explications fournies. Cette évaluation prend en compte la cohérence des interprétations, leur robustesse face à des variations des données d'entrée et leur utilité pour les utilisateurs finaux. L'objectif est d'établir des conclusions solides sur l'efficacité des différentes techniques d'explicabilité, tout en identifiant les points forts et les limites de chaque approche.

Analyse des Projets

Cette section présente une analyse détaillée des six projets réalisés, chacun illustrant des approches variées pour l'interprétation des modèles d'apprentissage automatique. Chaque projet a été conçu pour répondre à des problématiques spécifiques, en mettant en œuvre des méthodes adaptées à la nature des données et aux objectifs visés.

Le premier projet porte sur l'utilisation des SHAP (SHapley Additive exPlanations) Values. Cette méthode repose sur la théorie des jeux de Shapley pour attribuer à chaque caractéristique un score reflétant sa contribution à la prédiction finale d'un modèle. L'analyse des résultats a permis d'identifier les facteurs les plus influents, offrant ainsi des explications précieuses, notamment dans des contextes sensibles comme la santé ou la finance.

Le deuxième projet s'intéresse à l'interprétation des modèles tabulaires à l'aide de méthodes contre-factuelles. L'objectif est d'identifier les modifications minimales nécessaires sur les variables d'entrée pour changer la prédiction d'un modèle donné. Cette approche permet de mieux comprendre la sensibilité du modèle à certaines caractéristiques et d'offrir des explications intuitives sur les décisions prises.

Le troisième projet explore l'utilisation de Grad-CAM (Gradient-weighted Class Activation Mapping) pour l'analyse d'images. Cette technique génère des cartes de chaleur qui mettent en évidence les régions d'une image ayant le plus contribué à la prédiction du modèle. L'analyse a permis de valider la pertinence des explications en comparant les zones identifiées par le modèle avec des annotations humaines.

Le quatrième projet porte sur l'interprétation des modèles BERT (Bidirectional Encoder Representations from Transformers). À l'aide de la bibliothèque Captum, des techniques d'attribution de l'importance des tokens ont été utilisées pour identifier les parties des textes qui influencent le plus les prédictions. Cela permet d'améliorer la compréhension des modèles de traitement du langage naturel, souvent perçus comme des boîtes noires.

Le cinquième projet concerne l'application de la méthode Layer-wise Relevance Propagation (LRP) pour l'analyse d'images. LRP redistribue la prédiction finale du modèle à travers les différentes couches du réseau jusqu'aux entrées initiales, en attribuant des scores de pertinence à chaque pixel. L'analyse a été étendue aux variantes LRP- et LRP- afin d'évaluer leur robustesse et leur capacité à fournir des explications stables.

Enfin, le sixième projet traite de l'interprétation des réseaux de neurones graphiques (Graph Neural Networks, GNN) à l'aide de la méthode LRP. L'objectif est d'expliquer les prédictions des modèles sur des graphes, en mettant en évidence les nœuds et les arêtes les plus influents. Ce type d'analyse est particulièrement complexe en raison de la structure non euclidienne des données, nécessitant des adaptations spécifiques des méthodes d'explicabilité.

Chaque projet a permis d'approfondir la compréhension des forces et des limites des différentes techniques d'explicabilité, en mettant en évidence leur complémentarité selon les contextes d'application.

4.1 SHAPLEY Values

Le premier projet porte sur l'utilisation des SHAP (SHapley Additive exPlanations) Values, une méthode d'explicabilité des modèles de Machine Learning fondée sur la théorie des jeux de Shapley. Cette approche attribue à chaque caractéristique d'une observation un score représentant sa contribution à la prédiction du modèle. L'idée clé derrière SHAP est de mesurer l'impact marginal de chaque caractéristique en comparant les prédictions du modèle avec et sans cette caractéristique, sur l'ensemble des combinaisons possibles des autres variables.

Mathématiquement, la valeur de Shapley pour une caractéristique i est définie par la formule suivante :

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f(S \cup \{i\}) - f(S) \right]$$

où :

- N est l'ensemble de toutes les caractéristiques,

- S est un sous-ensemble de N ne contenant pas la caractéristique i,
- $f(S)$ représente la prédiction du modèle en considérant uniquement les caractéristiques de S,
- ϕ_i est la valeur de Shapley associée à la caractéristique i.

Cette formule calcule la contribution moyenne marginale de la caractéristique i à travers toutes les permutations possibles des autres caractéristiques. Cela garantit des propriétés d'équité telles que l'efficacité, la symétrie et l'additivité, ce qui fait de SHAP une méthode robuste pour l'explicabilité.

Dans le cadre du projet, nous avons appliqué la méthode SHAP à un modèle de classification supervisé. Après avoir entraîné le modèle, nous avons calculé les valeurs de Shapley pour chaque observation du jeu de données. Les résultats ont été visualisés à l'aide de graphiques en barres et de diagrammes de dispersion pour illustrer l'influence des caractéristiques sur les prédictions. L'analyse des résultats a montré que certaines variables avaient un impact significatif sur la prédiction, tandis que d'autres avaient une contribution négligeable.

Un point fort de SHAP est sa capacité à fournir des explications locales (pour une prédiction spécifique) et globales (pour l'ensemble du modèle). Cependant, le calcul exact des valeurs de Shapley est coûteux en termes de complexité computationnelle, surtout lorsque le nombre de caractéristiques est élevé. Pour pallier ce problème, des approximations comme Kernel SHAP ou Tree SHAP ont été utilisées, offrant un compromis entre précision et efficacité.

4.2 Counterfactuals pour Données Tabulaires

Le deuxième projet porte sur l'utilisation des méthodes contre-factuelles pour l'interprétation des modèles de Machine Learning appliqués aux données tabulaires. L'idée des contre-factuels repose sur la question suivante : Que faudrait-il changer dans les données d'entrée pour obtenir une prédiction différente ? . Cette approche permet de générer des exemples synthétiques, proches des observations réelles, mais dont la prédiction du modèle est inversée. Cela aide à comprendre la sensibilité du modèle par rapport à certaines variables et offre des explications intuitives et actionnables pour les utilisateurs.

Mathématiquement, la génération d'un contre-factuel peut être formulée comme un problème d'optimisation. Étant donné un point de données x avec une prédiction $f(x)$, l'objectif est de trouver un point x' tel que la prédiction $f(x')$ soit différente de $f(x)$ (par exemple, dans un problème de classification binaire, changer de classe), tout en minimisant la distance entre x et x'. Ce problème peut être formalisé par l'équation suivante :

$$\min_{x'} \; d(x, x') \quad \text{sous la contrainte} \quad f(x') \neq f(x)$$

où :

- $d(x, x')$ représente une mesure de distance (comme la distance euclidienne) entre l'observation initiale x et le contre-factuel x',
- f est le modèle de Machine Learning utilisé pour la prédiction,
- x' est la version modifiée de x que l'on souhaite générer.

Dans le cadre de ce projet, plusieurs techniques d'optimisation ont été explorées pour générer des exemples contre-factuels. L'une des approches courantes consiste à utiliser des méthodes basées sur des algorithmes de recherche par gradient, particulièrement efficaces pour les modèles différentiables. D'autres méthodes, comme l'utilisation de modèles génératifs ou d'algorithmes évolutionnaires, ont également été étudiées pour des modèles plus complexes.

L'analyse des résultats a montré que les contre-factuels offrent des explications précieuses, car ils permettent d'identifier les caractéristiques qui influencent directement la décision du modèle. Par exemple, dans un modèle de crédit, un contre-factuel peut indiquer qu'une légère augmentation du revenu ou une réduction du taux d'endettement suffirait à obtenir une décision de prêt positive. De telles explications sont non seulement informatives, mais aussi actionnables, car elles donnent des indications concrètes sur les changements nécessaires.

Cependant, la génération de contre-factuels soulève des défis importants. Il est essentiel que les contre-factuels soient à la fois plausibles et cohérents avec la distribution des données réelles. Pour cela, des contraintes supplémentaires peuvent être ajoutées au problème d'optimisation afin d'assurer la validité des exemples générés.

4.3 Grad-CAM pour l'Analyse d'Images

Le troisième projet se concentre sur l'utilisation de la méthode Grad-CAM (Gradient-weighted Class Activation Mapping) pour l'interprétation des modèles de deep learning appliqués à l'analyse d'images. Grad-CAM est une technique de visualisation qui permet d'identifier les régions d'une image ayant le plus contribué à la décision d'un modèle de classification. En générant des cartes de chaleur superposées aux images d'origine, cette méthode aide à comprendre quelles parties de l'image le modèle considère comme les plus importantes pour sa prédiction.

Sur le plan mathématique, Grad-CAM repose sur l'utilisation des gradients des scores de classe par rapport aux cartes d'activation d'une couche de convolution spécifique. Soit A^k la k-ième carte d'activation d'une couche convolutive et y^c le score de la classe c avant l'application de la fonction softmax. Le poids d'importance α_k^c pour la carte d'activation A^k est calculé en moyennant les gradients du score de la classe c par rapport à cette carte :

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

où :

- Z est le facteur de normalisation correspondant au nombre de pixels dans la carte d'activation,
- i et j parcourent les dimensions spatiales de la carte d'activation,
- $\frac{\partial y^c}{\partial A_{ij}^k}$ représente le gradient du score de la classe par rapport à l'activation A_{ij}^k.

La carte de chaleur $L_{\text{Grad-CAM}}^c$ pour la classe c est ensuite obtenue en combinant les cartes d'activation pondérées par leurs poids d'importance, puis en appliquant la fonction ReLU pour ne conserver que les contributions positives :

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right)$$

Dans le cadre du projet, nous avons appliqué Grad-CAM à des modèles de classification d'images basés sur des réseaux de neurones convolutifs (CNN). Après l'entraînement du modèle, des images de test ont été utilisées pour générer des cartes de chaleur illustrant les zones de l'image les plus influentes dans la prédiction. Ces cartes ont été superposées aux images d'origine pour une meilleure interprétation visuelle.

Les résultats ont montré que Grad-CAM permet de visualiser de manière intuitive les régions de l'image qui activent fortement les neurones responsables de la classification. Par exemple, dans un modèle de reconnaissance d'objets, la méthode a permis d'identifier que le modèle se concentrait principalement sur des caractéristiques pertinentes comme le contour des objets ou des motifs spécifiques. Cependant, des cas d'erreurs ont révélé que le modèle pouvait parfois se focaliser sur des détails non pertinents, mettant en évidence des faiblesses dans l'apprentissage.

L'un des avantages majeurs de Grad-CAM est sa capacité à être appliqué à une grande variété de modèles de vision par ordinateur sans nécessiter de modifications structurelles. Cependant, la qualité des explications dépend fortement du choix de la couche d'activation utilisée. Une mauvaise sélection peut entraîner des cartes de chaleur moins informatives.

4.4 Interprétation de BERT

Le quatrième projet est consacré à l'interprétation des modèles BERT (Bidirectional Encoder Representations from Transformers), une architecture de pointe dans le domaine du traitement du langage naturel (NLP). BERT repose sur des mécanismes d'attention qui permettent au modèle de capter des relations complexes entre les mots dans un texte. Cependant, cette complexité rend difficile la compréhension des décisions du modèle. L'objectif de ce projet était d'expliquer les prédictions de BERT en identifiant les tokens les plus influents dans le processus de classification.

L'interprétation de BERT s'appuie sur des techniques d'attribution de l'importance des caractéristiques, en particulier les méthodes basées sur les gradients. Une des approches utilisées est la méthode des gradients intégrés (Integrated Gradients), qui attribue à chaque entrée un score reflétant son influence sur la prédiction. Mathématiquement, pour une entrée x et un modèle F, le score d'importance pour une caractéristique i est défini par :

$$\text{IG}_i(x) = (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \cdot (x - x'))}{\partial x_i} d\alpha$$

où :

- x_i est la valeur de la caractéristique i de l'entrée réelle,
- x'_i est une valeur de référence (par exemple un vecteur de zéros),
- F est la fonction de prédiction du modèle,
- α est un coefficient d'intégration variant de 0 à 1.

Cette méthode permet de quantifier l'impact de chaque mot sur la prédiction finale du modèle en mesurant l'effet cumulatif des gradients le long d'un chemin reliant l'entrée de référence à l'entrée réelle.

Dans le cadre du projet, nous avons appliqué cette technique à des modèles BERT pré-entraînés sur des tâches de classification de texte. Les scores d'importance ont été visualisés en surlignant les mots les plus influents dans les phrases d'entrée. Cette approche a permis d'identifier les tokens clés qui orientent la décision du modèle. Par exemple, dans des tâches de classification de sentiments, des mots comme excellent ou terrible présentaient des scores d'importance élevés, confirmant la pertinence des explications.

Les résultats ont mis en évidence la capacité de BERT à capturer des dépendances contextuelles complexes, mais aussi des biais potentiels dans les prédictions. Certaines erreurs ont montré que le modèle pouvait accorder trop d'importance à des mots non significatifs, ce qui souligne la nécessité d'une interprétation rigoureuse pour identifier de tels biais.

Un des avantages de cette méthode est sa compatibilité avec les architectures de type Transformer sans nécessiter de modifications structurelles. Cependant, le calcul des gradients intégrés est coûteux en ressources, en particulier pour de longues séquences de texte.

4.5 Layer-wise Relevance Propagation pour l'Analyse d'Images

Le cinquième projet se concentre sur l'application de la méthode Layer-wise Relevance Propagation (LRP) pour l'analyse d'images. LRP est une technique d'interprétation des modèles de deep learning qui permet de comprendre comment un modèle de classification prend ses décisions en redistribuant la prédiction finale jusqu'aux pixels de l'image d'entrée. L'objectif de cette méthode est d'attribuer à chaque pixel une pertinence indiquant son influence sur la décision du modèle.

Mathématiquement, LRP repose sur la conservation de la pertinence à travers les couches du réseau de neurones. Soit un modèle de classification avec une sortie $f(x)$ pour une image d'entrée x. La pertinence R_j d'un neurone j dans une couche donnée est redistribuée aux neurones de la couche précédente i selon la règle suivante :

$$R_i = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j} + \epsilon} R_j$$

où :

- R_i et R_j sont les scores de pertinence des neurones i et j,
- z_{ij} représente la contribution de la connexion entre les neurones i et j (par exemple, le produit de l'activation a_i et du poids w_{ij}),
- ϵ est un petit terme ajouté pour éviter des divisions par zéro, utilisé dans la variante LRP-ϵ,
- la somme $\sum_{i'} z_{i'j}$ est calculée sur toutes les entrées i' connectées à j.

Cette redistribution est appliquée de manière itérative, couche par couche, jusqu'à atteindre la couche d'entrée, où chaque pixel de l'image reçoit un score de pertinence indiquant sa contribution à la prédiction finale.

Dans le cadre du projet, nous avons appliqué la méthode LRP à des modèles de classification d'images basés sur des réseaux de neurones convolutifs (CNN), notamment sur le dataset MNIST et des images plus complexes. En complément de la version standard de LRP, nous avons testé deux variantes : LRP-ϵ, qui améliore la robustesse en stabilisant les dénominateurs des fractions, et LRP-$\alpha\beta$, qui permet de contrôler la répartition des contributions positives et négatives avec les paramètres α et β tels que $\alpha - \beta = 1$.

Les résultats obtenus sous forme de cartes de chaleur ont permis d'identifier les régions les plus pertinentes des images pour la classification. Par exemple, dans la reconnaissance de chiffres manuscrits, les contours des chiffres étaient fortement mis en évidence, ce qui confirme la pertinence des explications fournies par LRP. L'analyse a également permis de détecter des cas où le modèle se concentrait sur des détails non significatifs, révélant ainsi des biais potentiels dans l'apprentissage.

L'un des avantages de LRP est sa capacité à fournir des explications précises à l'échelle des pixels, ce qui en fait un outil puissant pour l'interprétation des modèles de vision par ordinateur. Cependant, la qualité des explications dépend fortement des hyperparamètres choisis et de la structure du modèle. De plus, bien que LRP soit applicable à une grande variété de réseaux de neurones, son interprétation peut être plus complexe pour des architectures très profondes.

4.6 Interprétation des Réseaux de Neurones Graphiques

Le sixième projet est dédié à l'interprétation des réseaux de neurones graphiques (Graph Neural Networks, GNN) en utilisant la méthode Layer-wise Relevance Propagation (LRP). Les GNN sont des architectures puissantes conçues pour traiter des données structurées en graphes, telles que des réseaux sociaux, des molécules, ou des relations complexes entre entités. Cependant, leur complexité rend difficile la compréhension des décisions qu'ils produisent, d'où la nécessité d'approches explicatives robustes comme LRP.

L'adaptation de LRP aux GNN repose sur le même principe de redistribution de la pertinence à travers les couches du réseau, mais elle nécessite des ajustements spécifiques pour gérer la structure des graphes. Considérons un graphe $G = (V, E)$, où V est l'ensemble des nœuds et E l'ensemble des arêtes. La pertinence R_v d'un nœud v est redistribuée à ses nœuds voisins en fonction des poids des arêtes et des activations des nœuds, selon la règle suivante :

$$R_i = \sum_{j \in \mathcal{N}(i)} \frac{a_i w_{ij}}{\sum_{k \in \mathcal{N}(j)} a_k w_{kj} + \epsilon} R_j$$

où :

- R_i et R_j sont les scores de pertinence des nœuds i et j,
- a_i représente l'activation du nœud i,
- w_{ij} est le poids de l'arête entre les nœuds i et j,
- $\mathcal{N}(i)$ désigne l'ensemble des voisins du nœud i,
- ϵ est un terme de stabilisation pour éviter des divisions par zéro.

Cette approche permet de suivre la contribution des nœuds et des arêtes à la prédiction finale du modèle, en attribuant des scores de pertinence qui reflètent l'importance de chaque composant du graphe.

Dans le cadre du projet, nous avons appliqué LRP à des modèles GNN formés sur des graphes issus du dataset Cora, un ensemble de données largement utilisé pour la classification de nœuds dans des réseaux de citations scientifiques. L'analyse des graphes générés par le modèle de Barabási-Albert a permis d'observer comment les structures de réseau influencent les décisions de classification. En visualisant les scores de pertinence, nous avons pu identifier les nœuds et les arêtes ayant un impact significatif sur les prédictions.

Les résultats ont montré que LRP est capable de mettre en évidence des motifs structurels clés dans les graphes, tels que des hubs ou des communautés fortement connectées, qui jouent un rôle essentiel dans la classification. Par exemple, dans des graphes de réseaux sociaux, les nœuds avec un degré de centralité élevé apparaissaient souvent comme des points critiques dans les décisions du modèle.

Cependant, l'interprétation des GNN avec LRP présente des défis, notamment en raison de la complexité des interactions entre les nœuds et les arêtes. De plus, la distribution de la pertinence peut être sensible à la topologie du graphe, ce qui nécessite des ajustements des paramètres pour garantir des explications cohérentes.

Discussion

L'analyse des différents projets montre que chaque méthode d'interprétation offre des avantages spécifiques selon le type de données et de modèles. Les approches comme SHAP et les contre-factuels sont particulièrement adaptées aux données tabulaires, tandis que Grad-CAM et LRP permettent de visualiser efficacement les zones d'intérêt dans les images et les graphes. En combinant ces techniques, il est possible d'obtenir une compréhension plus complète des décisions des modèles de Machine Learning, renforçant ainsi leur transparence et leur fiabilité.

Conclusion

Ce rapport a exploré diverses méthodes d'interprétation des modèles de Machine Learning, démontrant leur utilité pour améliorer la transparence et la compréhension des décisions algorithmiques. Chaque approche, qu'il s'agisse de SHAP, des contre-factuels, de Grad-CAM ou de LRP, a permis d'apporter des éclairages spécifiques adaptés à différents types de données et de modèles. En combinant ces techniques, il est possible de renforcer la confiance des utilisateurs dans les systèmes d'intelligence artificielle et de promouvoir un usage plus éthique et responsable des modèles prédictifs.

Références

References

[1] Scott M. Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions.* Advances in Neural Information Processing Systems (NeurIPS), 2017.

[2] Sandra Wachter, Brent Mittelstadt, and Chris Russell. *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR.* Harvard Journal of Law & Technology, 2017.

[3] Ramprasaath R. Selvaraju et al. *Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.* Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[4] Lenaic Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. *Explaining Recurrent Neural Network Predictions in Sentiment Analysis.* EMNLP Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis, 2017.

[5] Thomas N. Kipf and Max Welling. *Semi-Supervised Classification with Graph Convolutional Networks.* arXiv preprint arXiv:1609.02907, 2016.

[6] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. *Axiomatic Attribution for Deep Networks.* Proceedings of the 34th International Conference on Machine Learning (ICML), 2017.

[7] ChatGPT. *Assistance à la correction des projets et à la rédaction du rapport final.*

[8] Sites et tutoriels. *Ressources en ligne utilisées pour l'assistance à la correction des projets.*

Rendu Final

L'ensemble des projets réalisés dans le cadre de ce rapport est disponible sur la page GitHub de Ghizlan Moqim. Vous pouvez les consulter en suivant le lien ci-dessous, où vous trouverez les codes sources, les notebooks, ainsi que des ressources complémentaires liées à l'interprétation des modèles de Machine Learning :

`https://github.com/moqim-ghizlan/XAI-Project-Interpretability`

Ce dépôt regroupe les différentes approches explorées, avec des explications détaillées et des exemples pratiques pour faciliter la compréhension des concepts abordés.