Version longue

Depuis plusieurs années, l'Intelligence Artificielle (IA) fondée sur le machine learning ou apprentissage machine a produit des résultats impressionnants, notamment dans les domaines de la reconnaissance d'image et du traitement de la voix. Ce succès a été permis par l'augmentation de la puissance de calcul, mais aussi par la capacité de traitement de plus gros volumes de données et par des algorithmes plus efficients tels que les réseaux de neurones artificiels. Et aussi le développement et l'engouement rapides autour de l'IA qui ont suivi ont conduit à prioriser la performance des algorithmes au détriment de leur interprétabilité et ont conduit au problème de la « boite noire ». Les algorithmes de machine learning sont en effet comparés à des « boites noires » et le problème d'opacité est considéré comme un défi scientifique majeur.

Le manque d'interprétabilité des techniques de machine learning pose des problèmes juridique, opérationnels et éthiques. Tout d'abord sur le plan juridique car s'exprime la nécessité de satisfaire à des normes et des réglementations. Pour un certain nombre de secteurs l'urgence et le besoin d'interprétabilité sont plus pressants : la banque, la justice, la santé sont les plus demandeurs car pour ces domaines en particulier il s'applique des exigences légales et éthiques très strictes. Un second aspect est la prise en compte grandissante d'un apprentissage automatique centré vers l'utilisateur, un modèle compréhensible devrait permettre d'effectuer les tâches professionnels ou quotidiennes de façon plus pérenne et réaliste. L'aspect éthique est une autre préoccupation incontournable et qui prend tout son sens dans un contexte de collecte massive de données et d'appel à plus de transparence.

Un nouveau pan de recherche émerge et se concentre sur la question de l'impénétrabilité de l'intelligence artificielle : il s'agit du machine learning interprétable qui dessine une nouvelle dynamique dans laquelle l'interprétabilité pourrait devenir le nouveau critère d'évaluation des modèles.

Notre projet de E3 consiste à construire des explications de modèles de machine learning, que nous considérons comme des « boîtes noires », que nous déployons sous la forme d'une « boîte à outils ». Dans ce contexte plus précis, nous définissons l'interprétabilité comme la capacité à fournir pour un ensemble d'utilisateurs une explication des résultats obtenus par le machine learning qui puisse aider à répondre à la question « pourquoi a-t-on ces résultats ? ». Le projet s'est déroulé suivant quatre jalons : réalisation d'un état de l'art, réalisation d'un banc de test, expérimentation des techniques d'interprétabilité et réalisation de la boîte à outils.

La littérature fait état d'un nombre croissant de techniques et nous nous focalisons sur celles qui semblent être les plus pertinentes, à savoir LIME, SHAP, PDP, ICE, permutation features et shapley value. Cette sélection a pu être dressé après avoir réalisé un état de l'art d'une liste non exhaustive des techniques existantes - explicitant le fonctionnement mathématique et spécifiant les avantages et inconvénients. Lors de cette première

étape le travail a été réparti entre les membres du groupe et a été complétée par la réalisation de mises en pratique des techniques via des applications sur des jeux de données de faibles volumes.

Cette sélection a permis la réalisation d'un premier prototype de boîte à outil : nous avons implémenter des modèles de complexité croissante en commençant par une régression linéaire simple et en finissant avec XGboost et mesurer leurs performances. Le critère était donc l'inteprétabilité, nous avons appliqué les techniques énoncées précédemment en les mettant en parallèle afin de réaliser une étude comparée des résultats – et comparée vis-à-vis de la compréhension humaine des jeux de données- et afin de pouvoir penser notre boîte à outils. En pratique nous avons exploité un jeu de données venant du site Kaggle pour réaliser notre travail : jeu de données tabulaire regroupant des mesures physico chimiques réalisées sur le Vinho Verde, un vin portugais.

L'ensemble des techniques sélectionné est encapsulé dans la boîte à outil qui permet la génération automatique d'un ensemble de reports, graphiques et mesures pouvant être utilisées avec n'importe quel modèle de machine learning. La boite à outil génère donc les résultats des techniques.