UNIVERSITÉ D'ARTOIS

# Avis de Soutenance

## Monsieur LOUENAS BOUNIA

## Informatique et applications

Soutiendra publiquement ses travaux de thèse intitulés

*Modèles formels pour l'IA explicable : des explications pour les arbres de décision*

dirigés par Monsieur Pierre MARQUIS et Monsieur Frédéric KORICHE

Soutenance prévue le ***vendredi 22 décembre 2023*** à 13h30
Lieu : Faculté des Sciences Jean Perrin, Salle des thèses. 13 Rue Jean Souvraz, 62300 Lens
Salle : des thèses

**Composition du jury proposé**

| | | |
|---|---|---|
| M. Pierre MARQUIS | Université d'Artois | Directeur de thèse |
| M. Frédéric KORICHE | Université d'Artois | Directeur de thèse |
| Mme Céline ROUVEIROL | Université Sorbonne Paris Nord - LIPN | Examinatrice |
| Mme Isabelle BLOCH | Sorbonne Université - LIP6 | Rapporteure |
| M. Martin COOPER | Université de Toulouse - IRIT | Rapporteur |

**Résumé :**

Le besoin et la motivation pour l'intelligence artificielle explicable (XAI) peuvent être résumés en deux objectifs principaux : justification et validation. Les explications permettent de justifier les résultats d'un modèle d'apprentissage automatique en fournissant un motif du raisonnement suivi par le modèle. Une fois que les explications sont extraites, il est essentiel de vérifier la validité de ces explications pour s'assurer qu'elles correspondent aux intentions et aux attentes de l'utilisateur du système d'IA considéré. Dans cette thèse, nous avons développé des méthodes de génération d'explications locales post-hoc. Nous avons commencé par étudier l'intelligibilité computationnelle des classifieurs booléens, caractérisée par leur capacité à répondre aux requêtes d'explicabilité en temps polynomial. Nous avons montré que l'intelligibilité computationnelle des arbres de décision est supérieure à celle de nombreux modèles d'apprentissage automatique, ce qui nous a conduit par la suite à nous focaliser sur les arbres de décision comme modèle d'apprentissage. Nous avons étudié la capacité des arbres de décision à extraire, minimiser et compter les explications abductives et contrastives. Nous avons montré que l'ensemble complet des raisons suffisantes pour une instance peut être de taille exponentielle, rendant ainsi difficile, voire impossible, la génération de l'ensemble complet. De plus, deux raisons suffisantes pour une même instance peuvent différer de manière significative. Pour tenter de synthétiser l'ensemble des raisons suffisantes pour une instance, nous avons introduit les concepts d'attribut nécessaire et d'attribut pertinent, ainsi que le concept de pouvoir explicatif d'un attribut. Pour réduire le nombre d'explications à fournir à l'utilisateur, nous avons aussi développé des modèles de préférence et des algorithmes pour générer des explications exploitant ces préférences. Cette démarche présente plusieurs avantages. En prenant en compte ces préférences, nous pouvons rechercher des explications abductives qui conviennent à l'utilisateur tout en réduisant considérablement leur nombre. Cependant, même quand on se restreint aux explications préférées, les explications abductives peuvent être de trop grande taille pour être interprétables par les humains en raison des limitations cognitives de ces derniers. Nous avons donc abordé le défi de réduire la taille des explications tout en maintenant une probabilité élevée de prédiction exacte. Ce problème est difficile en général, même pour les arbres de décision. Pour surmonter cette difficulté, nous avons exploré l'approximation des explications probabilistes en utilisant le concept de super-modularité. Nous avons développé deux algorithmes gloutons (GA et GD) pour la minimisation super-modulaire. L'efficacité de ces algorithmes dépend de la courbure de la fonction d'erreur non normalisée qui mesure la précision de l'explication. Empiriquement, nous avons montré l'efficacité de ces algorithmes.