Stratégies coalitionnelles pour une explication efficace des prédictions individuelles.

Elodie Escriva *,**, Gabriel Ferrettini **,
Julien Aligon **, Jean-Baptiste Excoffier *,
Chantal Soulé-Dupuy **

*Kaduceo, 31 Allée Jules Guesde, 31400 Toulouse, France
firstName.lastName@kaduceo.com
**Université de Toulouse-Capitole, IRIT, (CNRS/UMR 5505)
2 rue du Doyen-Gabriel-Marty 31042 Toulouse cedex 9, France
firstName.lastName@irit.fr

Résumé. Ce papier est un résumé des travaux publiés dans le journal Information Systems Frontiers (Ferrettini et al., 2021). Face aux nombreuses applications de l'apprentissage machine (ML) dans de nombreux domaines, la nécessité de comprendre le fonctionnement des modèles en boite noire est devenu croissante, particulièrement chez les non-experts. Plusieurs méthodes fournissant des explications sur les prédictions des modèles existent, avec des temps de calculs longs ou des hypothèses restrictives sur les interactions entre attributs. Ce papier détaille des méthodes basées sur la détection de groupes d'attributs pertinents – appelés coalitions – influençant la prédiction. Nos résultats montrent que les méthodes coalitionnelles sont plus performantes que celles existantes, comme SHAP. Le temps d'exécution est réduit en préservant la précision des explications. Ces méthodes permettent une augmentation des cas d'utilisation afin d'accroître la confiance entre les modèles ML, les utilisateurs et toute personne affectée par une décision impliquant ces modèles.

1 Introduction

Dans de nombreux domaines, l'absence d'explications des prédictions et le manque de transparence des modèles "boite noire" est un frein à l'utilisation de l'apprentissage automatique. La majorité des algorithmes de la littérature ne fournit pas d'informations claires sur la production de prédictions. Cependant, plusieurs techniques agnostiques ont été décrites pour comprendre les prédictions individuelles des modèles (Strumbelj et Kononenko, 2010; Ribeiro et al., 2016; Casalicchio et al., 2018). Les méthodes agnostiques sont applicables à tous types de modèles, indépendamment de leur structure interne, contrairement aux méthodes spécifiques à un seul type de modèles. Pour l'explication d'une prédiction individuelle, les valeurs de Shapley sont une base intéressante pour calculer les influences de chaque attribut sur cette prédiction. Cette technique se base sur la théorie des jeux collaboratifs en considérant les attributs comme des joueurs