Utilisation de modèles de Machine Learning combinés avec des méthodes d'interprétation innovantes pour identifier des facteurs pronostiques

Romane Péan, Adrien Darbier, Rim Ghorbal, David Pau, Cyril Esnault, Mélina Gilberg, Julien Dupin, Alexandre Civet

Roche, France

INTRODUCTION. Le cancer du sein est le cancer le plus fréquemment diagnostiqué chez les femmes dans le monde avec une prévalence à 5 ans de 228 252 cas en France [1]. L'objectif de cette analyse exploratoire est d'identifier les facteurs pronostiques de la réponse pathologique complète (pCR) après le traitement néoadjuvant en utilisant des méthodes de Machine Learning. Cette recherche complète une étude observationnelle nationale rétrospective incluant des patientes atteintes d'un cancer du sein précoce HER2+ recevant une thérapie ciblée (Trastuzumab) en situation néoadjuvante et adjuvante.

MÉTHODE. Une approche en 2 étapes a été utilisée pour identifier les facteurs pronostiques de la pCR. La première consiste à utiliser des modèles de Machine Learning destinés à prédire le statut pCR des patientes à partir de leurs caractéristiques cliniques et des caractéristiques des centres de prise en charge. Ces algorithmes (SVM, KNN, DecisionTree, RandomForest, XGBoost, AdaBoost) sont entraînés et optimisés par GridSearch avec une validation croisée sur un jeu de données d'apprentissage (75% des patientes), puis testés sur un jeu de données de validation (25%). Les performances en apprentissage et en validation des modèles de Machine Learning ont été calculées afin de sélectionner le modèle le plus robuste (peu de sur-apprentissage). La deuxième étape consiste à identifier les variables influençant la prédiction de la pCR en utilisant trois méthodes d'interprétation agnostiques. La méthode SHAP (SHapley Additive exPlanations) permet d'obtenir une interprétation globale en identifiant les variables les plus importantes et l'ampleur de leur impact sur le modèle. LIME (Local Interpretable Model-agnostic Explanations) fournit des explications locales en présentant la contribution de chaque variable à la prédiction de la pCR pour une patiente donnée. La méthode PDP (Partial Dependence Plot) permet de visualiser l'impact global de chaque variable sur la prédiction de la pCR.

RÉSULTATS. Adaboost offre le meilleur compromis entre une bonne AUC en test (0.62), une bonne spécificité en test (0.74) et une spécificité robuste (**Δ** = 0.09).

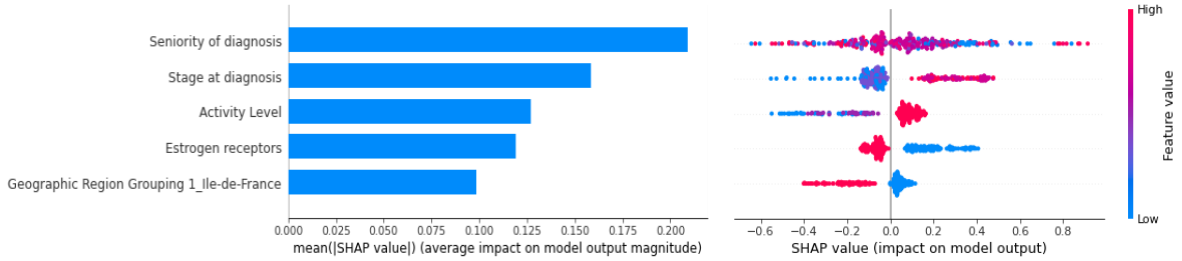

Figure : Résultats SHAP pour le modèle AdaBoost

Les résultats du SHAP (voir Figure) pour ce modèle ont montré que le stade au diagnostic et l'ancienneté du diagnostic sont les variables ayant le plus d'impact sur le statut pCR. L'impact de la variable Ancienneté du diagnostic semble suivre une distribution dissymétrique difficilement interprétable. La variable Stade au diagnostic semble avoir un impact linéaire et positif. Les résultats individuels des approches LIME et PDP tendent à renforcer les résultats globaux de la méthode SHAP.

DISCUSSION. Les modèles de Machine Learning utilisés permettent une prise en compte de toutes les variables sans a priori ce qui est un atout pour une analyse plus approfondie des données médicales. De plus, ils sont susceptibles de mieux identifier le modèle de données sous-jacent, et donc de permettre une meilleure identification et mesure des facteurs pronostiques qu'avec des modèles classiques. Toutefois, leur interprétation et le feature engineering des données nécessitent une vigilance clinique et méthodologique. Par conséquent, l'utilisation de méthodes innovantes d'interprétation permet d'obtenir des interprétations locales et globales du fonctionnement de ces modèles ainsi que l'identification des variables les plus importantes et de leurs effets.

[1] Global Cancer Observatory n.d. http://gco.iarc.fr;