Vers une aide intelligible à l'analyse prédictive

Julien Aligon

MÉMOIRE

En vue de l'obtention de l'

HABILITATION A DIRIGER DES RECHERCHES DE L'UNIVERSITE DE TOULOUSE

Délivré par : *l'Université Toulouse Capitole (UT Capitole)*

Présentée et soutenue le *25/06/2024* **par :**

Julien ALIGON

Vers une aide intelligible à l'analyse prédictive.

JURY

Nicolas Labroche	Maître de conférences HDR	Univ. de Tours
Anne Laurent	Professeure des Universités	Univ. de Montpellier
Marie-Jeanne Lesot	Professeure des Universités	Sorbonne Université
Paul Monsarrat	Professeur des Universités - Praticien hospitalier	Univ. Toulouse Paul Sabatier
Chantal Soulé-Dupuy	Professeure des Universités	Univ. Toulouse Capitole
Olivier Teste	Professeur des Universités	Univ. Toulouse Jean Jaurès
Cédric Wemmert	Professeur des Universités	Univ. de Strasbourg

---

École doctorale et spécialité :
MITT : Domaine STIC : Intelligence Artificielle

Unité de Recherche :
Institut de Recherche en Informatique de Toulouse, UMR 5055 (CNRS)

Professeur référent :
Chantal Soulé-Dupuy

Rapporteurs :
Anne Laurent, *Marie-Jeanne Lesot* et *Cédric Wemmert*

Résumé

L'analyse de données est un terme très générique et touche une variété d'analyses possibles, incluant, par exemple, l'analyse décisionnelle (BI) ou l'analyse statistique, dont fait partie l'analyse prédictive. Dans le cadre de l'analyse prédictive, la construction et l'usage de modèles prédictifs, réalisés à l'aide de techniques d'intelligence artificielle, ne sont pas sans poser de multiples problèmes, notamment liés à la compréhension des résultats produits par le modèle, connus sous le nom d'effet boîte noire. En effet, l'apprentissage supervisé ne permet pas, par lui-même, de comprendre aisément les raisons influençant les prédictions faites pour des instances d'un jeu de données.

Le besoin d'une aide à l'analyse de données est alors un enjeu important et particulièrement criant dans le cadre d'une analyse prédictive réalisée via des modèles d'apprentissage. Notamment, la prise en compte de l'utilisateur dans cette aide à l'analyse est primordiale. L'aspect « human-in-the-loop » est en effet trop souvent négligé dans un environnement où l'aide à la construction de modèles prédictifs fait souvent la part belle au tout automatique, en particulier par les approches d'apprentissage machine automatique (« AutoML »). Ces approches accentuent alors l'effet boîte noire où l'utilisateur ne maîtrise plus la construction du modèle et sa compréhension.

L'ambition soutenue dans cette Habilitation à Diriger des Recherches est alors de questionner la problématique de savoir comment offrir à un utilisateur les moyens d'une analyse prédictive intelligible ? Pour cela, nos travaux traitent trois thématiques générales de recherche, dans le contexte de l'apprentissage supervisé : (1) l'aide à l'analyse prédictive basée sur l'apprentissage machine ("Machine Learning", ML), (2) les systèmes de recommandation de modèles prédictifs, et (3) l'explicabilité des modèles prédictifs ("eXplainable Artificial Intelligence", XAI).

En ce sens, les contributions réalisées durant ces années portent principalement sur :

1. Conception de systèmes de recommandation de modèles ML et de modèles XAI.
2. Définition d'une nouvelle méthode XAI locale attributive coalitionnelle.
3. Définition d'une approche méthodologique de l'exploration et l'analyse de données guidées par l'XAI

Ces travaux se sont en particulier déroulés dans un contexte multidisciplinaire, avec le laboratoire RESTORE (biologie/santé) spécialisé dans les gérosciences. Cette collaboration a mené à plusieurs co-encadrements de thèses traitant de l'analyse du vieillissement, et de travaux menant à une meilleure compréhension des phénomènes sous-jacents.

Remerciements

L'exercice d'une habilitation à diriger des recherches est toujours une étape importante dans la vie d'un chercheur.

Mes premiers remerciements vont, naturellement, à :

— Anne Laurent, Professeure des universités - Université de Montpellier,

— Marie-Jeanne Lesot, Professeure des universités - Sorbonne Université,

— Cédric Wemmert, Professeur des universités - Université de Strasbourg,

pour avoir accepté de rapporter ce mémoire d'HDR et de participer à ce jury.

C'est aussi l'occasion pour moi de revenir sur ces années passées à l'université Toulouse Capitole, depuis ma nomination en tant que maître de conférences en 2016. En particulier, je tiens à remercier toutes les personnes qui, de près ou de loin, m'ont permis de soutenir mes activités d'enseignant-chercheur. Et ils sont nombreux...

Je l'écris sans équivoque : j'ai eu, et j'ai encore aujourd'hui, le plaisir de trouver à l'université Toulouse Capitole des conditions de travail facilitantes.

Indéniablement, les services administratifs de l'université y contribuent pour une bonne part. Je remercie tout particulièrement le secrétariat informatique (Lydie Ballabriga, Michèle Cuesta et Florence Thery) pour leurs précieux conseils et aides dans l'organisation quotidienne de la vie de la faculté informatique. Le soutien de la scolarité informatique à nos activités d'enseignements et d'organisation des formations est tout aussi inestimable. Un immense merci à la remarquable Mélanie Buzet, cheffe de service, ainsi qu'à Jérôme Barathieu (qui a su m'accompagner avec bienveillance dans mes débuts de responsable de formation...) mais aussi à Christophe Erta, Evelyne Fabien et Mélanie Hochet.

Je n'oublie évidemment pas tous les collègues de la faculté informatique, représentée par notre doyen Laurent Perrussel. En particulier, je voudrais remercier chaleureusement Jean-Marc Thevenin qui, dès mon arrivée à l'université, a eu la gentillesse de m'impliquer à ses côtés dans le montage de nouveaux enseignements ainsi que de me proposer de co-gérer avec lui la formation de Master 2 MIAGE ISIAD. Un clin d'œil tout aussi amical à Harold Parpex, vacataire de la faculté, qui se trouve, par les hasards de la vie, à avoir été l'un de mes derniers étudiants lorsque j'étais encore en thèse à l'antenne universitaire de Blois (et qui a su bien me le rendre, lors de cette fameuse journée de découverte au

ski !). Un grand merci également à tous les collègues pour la confiance qui m'a été faite de coordonner la mention de master MIAGE de la faculté. C'est à la fois un défi et un véritable plaisir que de pouvoir agir collectivement dans l'évolution de nos offres de formation afin d'offrir le meilleur à nos étudiants ! Plus généralement, je souligne vivement tous les moments de convivialité, de bonne humeur et de rire que nous avons partagés ensemble, et bien d'autres certainement encore à venir ! C'est aussi cela de bonnes conditions de travail...

Une faculté va aussi de pair avec ses équipes de recherche. Je dois bien avouer que l'équipe SIG de l'IRIT, dont je suis membre, sait très bien intégrer les nouveaux venus ! L'humanité et l'esprit de solidarité qui animent cette équipe sont effectivement, pour moi, une profonde source de satisfaction. Un grand merci à tous les membres de cette équipe et tout particulièrement à Olivier Teste, chef de l'équipe et qui me fait le plaisir de participer à ce jury d'HDR, et Max Chevalier pour leurs précieux conseils tout au long de ces années ! Un merci très vif aux membres SIG/UT-Capitole, qui m'ont permis de m'intégrer dans d'excellentes conditions, en particulier Franck Ravat, Chantal Soulé-Dupuy, Ronan Tournier et Nathalie Vallès-Parlangeau (qui est parti rejoindre d'autres contrées depuis...). Je n'oublie pas non plus Jiefu Song et Moncef Garouani, nouveaux arrivants sur ces dernières années, à qui je souhaite de très belles carrières parmi nous !

Je vais me permettre de remercier plus longuement Chantal Soulé-Dupuy qui a très gentiment accepté d'être la garante de cette HDR. Merci beaucoup pour sa relecture attentive de ce mémoire, ses avis et ses conseils ! Plus globalement, c'est un réel plaisir que de co-encadrer avec elle, depuis 2017, de multiples thèses ayant permis de faire évoluer nos thématiques de recherche vers l'apprentissage automatique. Je suis aussi très heureux que nous ayons porté ensemble la thématique de l'explicabilité dans l'équipe ! Enfin, un merci très vif pour la confiance qu'elle m'a toujours accordée, son accompagnement et les opportunités offertes, en particulier avec nos collègues en santé. Je salue aussi ses qualités humaines indéniables et sa grande capacité d'écoute !

Je voudrais aussi remercier plus particulièrement les membres du laboratoire pluridisciplinaire RESTORE dont j'ai la chance d'être partenaire. Un très grand merci à Philippe Valet, directeur de cette unité, de porter ce formidable projet rassemblant des équipes en biologie/santé avec d'autres domaines, dont l'informatique, afin d'avancer dans ce grand défi qui nous concerne tous : le vieillissement en bonne santé ! Un merci tout aussi appuyé à l'équipe 4 m'ayant permis d'y rencontrer des personnes visionnaires dans la complémentarité des recherches en biologie/santé et informatique. Un immense merci à Louis Casteilla et Valérie Planat, chefs de cette équipe, et à Paul Monsarrat avec qui j'ai le bonheur d'avancer dans une recherche passionnante et fructueuse ! Merci aussi à Isabelle Ader, Jean-Philippe Pradère et Cédric Dray pour nos collaborations tout aussi enrichissantes !

Il est, cependant, toujours navrant de constater, encore en 2024, les réticences d'une

partie de la communauté informatique à considérer la recherche interdisciplinaire comme un atout, valorisée au même titre que toute autre recherche. Je ne nie pas les difficultés à construire une collaboration interdisciplinaire. C'est un raisonnement sur le temps long, qui se compte en décennie, qui suppose des efforts soutenus de chacun, et qui ne permet naturellement pas une valorisation immédiate des travaux de recherche. Apprendre des autres communautés, échanger sur nos pratiques, partager nos expériences : c'est pourtant bien tout cela qui devrait animer l'esprit de tout chercheur ! Aujourd'hui, la recherche en biologie/santé ne peut avancer sans l'implication de multiples domaines comme l'informatique, les mathématiques, la physique, etc. La recherche en informatique, quant à elle, ne peut avancer sans se donner du sens et tâcher à se faire comprendre et accepter par le plus grand nombre.

Cette HDR ne serait évidemment rien sans les jeunes docteurs et doctorants que j'ai eus ou j'ai encore le plaisir d'encadrer. Je tiens ainsi à remercier et féliciter pour leurs thèses : Franck Boizard, Gabriel Ferrettini, Elodie Escriva ainsi que les doctorants actuels : Robin Cugny, Emmanuel Doumard et Haomiao Wang à qui je souhaite une très bonne fin de thèse !

Ces dernières années ont aussi été l'occasion de très belles collaborations avec Nicolas Labroche, de l'université de Tours. Merci à lui de m'avoir proposé de co-gérer, depuis 2020, diverses actions liées à la thématique de l'explicabilité tant au niveau national qu'international. Merci aussi pour tous ces traits d'humour assurant une bonne humeur à chaque instant ! Je souhaite de tout cœur que nous puissions avancer ensemble encore longtemps.

Il y a parfois des personnes, dans la vie, qui éclairent, qui passionnent, qui motivent. Cette HDR est aussi un peu une conséquence du bonheur de ces trois années de thèse passées sous la direction de Patrick Marcel. Merci infiniment à lui : je mesure encore aujourd'hui la chance d'avoir appris les fondamentaux de ce métier à ses côtés. Je me rappelle encore de ses premiers cours de bases de données dispensés dans une salle du bâtiment L de la faculté des sciences de Grandmont... c'était il y a déjà 17 ans !

Je remercie tous mes amis pour leurs appuis tout au long de ces années. Je pense, en particulier, à mes amis nantais dont je garde toujours un exceptionnel souvenir de mes années de post-doctorat !

Enfin, merci à ma famille pour son soutien sans faille. Merci à mes parents, Michèle et Jean-Yves, qui m'ont toujours encouragé, en particulier depuis mes débuts en thèse. Merci à mon frère Stéphane, ma soeur Caroline et mon beau-frère Florent pour tous nos moments ensemble et leur gentillesse. Merci aussi à mon neveu Aleksander qui me remplit de joie à chacune de nos rencontres ! Même si Toulouse m'éloigne forcément un peu de vous tous, il n'en reste pas moins que je pense toujours beaucoup à vous. Je suis heureux d'avoir une famille comme la nôtre, restant forte et unie malgré les épreuves de la vie.

Table des matières

Table des figures

Chapitre 1

Introduction

1.1 Contexte et ambition de nos travaux

L'analyse de données est un terme très générique et touche une variété d'analyses possibles, incluant, par exemple, l'analyse décisionnelle (BI) ou l'analyse statistique, dont fait partie l'analyse prédictive. Les différents types d'analyse de données dépendent ainsi de la manière dont sont *modélisées* les données [Tukey, 1962].

Dans le cadre de l'analyse prédictive, la construction et l'usage de modèles prédictifs, réalisés à l'aide de techniques d'intelligence artificielle, ne sont pas sans poser de multiples problèmes. Comme noté dans [Attaran and Attaran, 2019], les difficultés à obtenir des quantités de données les plus à jour et le manque d'experts en intelligence artificielle dans les organisations (en particulier les petites structures) ne peut que freiner l'adoption de ce type d'analyse. De plus, même lorsqu'un modèle prédictif est mise en place, le plus souvent par l'utilisation de techniques d'apprentissage supervisé [Jordan and Mitchell, 2015, Hastie et al., 2009b], les difficultés de compréhension des résultats produits par le modèle, connues sous le nom d'*effet boîte noire*, peut mener à ce que l'utilisateur perde confiance en ce type de technologie. En effet, l'apprentissage supervisé ne permet pas, par lui-même, de comprendre aisément les raisons influençant les prédictions faites pour des instances d'un jeu de données.

Comme nous le constatons, le besoin d'une aide à l'analyse de données est un enjeu important et particulièrement criant dans le cadre d'une analyse prédictive réalisée via des modèles d'apprentissage supervisé.

Pour répondre à cet enjeu, les systèmes d'aide à la décision (decision support system ou DSS) sont généralement des approches proposées, notamment par l'usage de systèmes de recommandation. De nombreuses propositions ont été faites dans les domaines de l'analyse décisionnelle [Aligon et al., 2015, Drushku et al., 2019, Drushku et al., 2017] et plus récemment dans un objectif d'analyse prédictive [Rudnichenko et al., 2020, Yahyaoui et al., 2019, Safdar et al., 2018], montrant ainsi l'intérêt apporté par les DSS.

Lors d'une aide à l'analyse prédictive, la prise en compte de l'utilisateur dans les choix proposés par un DSS est également primordiale. L'aspect *human-in-the-loop*

[Mosqueira-Rey et al., 2023] est en effet trop souvent négligé dans un environnement où, par exemple, l'aide à la construction de modèles prédictifs fait souvent la part belle au tout automatique, en particulier par les approches AutoML [Feurer et al., 2015a, He et al., 2021]. Ces approches accentuent alors l'effet boîte noire où l'utilisateur ne maîtrise plus la compréhension de construction de son modèle.

Au vu des éléments présentés précédemment, l'ambition portée par ce mémoire est alors de questionner la problématique suivante : *comment offrir à un utilisateur les moyens d'une analyse prédictive intelligible ?*

Pour cela, nous proposons de nous attaquer à trois thématiques générales de recherche, dans le contexte de l'apprentissage supervisé :

— L'aide à l'analyse prédictive
— Les systèmes de recommandation de modèles prédictifs
— L'explicabilité des modèles prédictifs

Nous détaillons ces trois thématiques via quatre verrous présentés dans l'orientation de nos travaux de la section suivante.

1.2 Orientation de nos travaux

L'orientation de nos travaux est illustrée en Figure 1.1 et se décompose en quatre verrous.

L'aide à l'analyse de données, telle que présentée Section 1.1, implique alors un premier verrou (point 1 de la Figure 1.1) : **proposer un système recommandant un ou plusieurs modèles prédictifs, le/les plus adapté(s) aux préférences de l'utilisateur** et aux données qu'il souhaite analyser. Ainsi, nous proposons une méthode, basée sur le méta-apprentissage et l'AutoML, permettant d'identifier les données déjà analysées dans le passé et les plus similaires au jeu de données à analyser. Par la suite les modèles prédictifs déjà exécutés sur ces données similaires sont filtrés de sorte à correspondre aux préférences utilisateur (sous la forme d'indicateurs de performance à maximiser) à l'aide d'un front de Pareto. En général, ces systèmes de recommandation fournissent des modèles prédictifs très précis. Cependant, l'interaction avec ces modèles prédictifs se limite simplement à... les exécuter ! En effet, il est souvent peu facile, voire impossible, de chercher à les valider ou les personnaliser. Il s'agit là d'un inconvénient majeur par lequel un utilisateur peut perdre confiance, en raison d'un manque d'explications sur les résultats du système de recommandation.

Une analyse prédictive intelligible nécessite également de s'intéresser à un deuxième verrou (point 2 de la Figure 1.1) consistant à **proposer le modèle d'explication le plus adapté**. En effet, la littérature offre désormais tout un panel de méthodes pour expliquer un modèle prédictif. Nous proposons alors un système recommandant le modèle d'explication adapté à un contexte de données et de préférences utilisateurs (nommé AutoXAI, basé en partie sur le principe de l'AutoML).

Le troisième verrou (point 3 de la Figure 1.1) consiste à **étudier les limites des méthodes d'explication**, en particulier les méthodes dites "post-hoc locales attributives", actuellement très populaires pour expliquer ce qu'a pu influencer le classement d'une instance vis-à-vis d'un modèle prédictif et du jeu de données considéré. À l'aide de ces explications, il est alors possible de mieux cerner les variables impliquées dans les décisions du modèle. Nous proposons ainsi à la fois une nouvelle méthode de calcul d'explications attributives, palliant la non-prise en compte des interactions entre variables par les méthodes de la littérature, ainsi qu'un comparatif montrant les avantages et inconvénients de l'utilisation des méthodes attributives.

Le quatrième verrou (point 4 de la Figure 1.1) consiste à **voir les explications comme un nouvel espace de données analysable et utilisable**. Différentes façons complémentaires permettent d'étudier cette question, à savoir :

- La **proposition d'un système, orienté *human-in-the-loop*, de sélection de modèles prédictifs à l'aide d'explications**. Nous avons ainsi proposé un cadre permettant à un utilisateur d'agir directement dans la vérification et la personnalisation de son modèle prédictif, à l'aide d'explications fournies pour chaque instance de son jeu de données.
- La **recherche d'une sélection de variables la plus adaptée à un profil d'explication** souhaité. Notre ambition est de compléter le paradigme du domaine de la sélection d'attributs, principalement basé sur la recherche du sous-ensemble de variables le plus petit possible assurant une précision du modèle prédictif la plus élevée possible, en y ajoutant un principe d'explicabilité. En effet, nous faisons l'hypothèse qu'une sélection de variables, bien que permettant d'obtenir un modèle prédictif très performant, peut n'avoir qu'un sens très relatif du côté métier. Il s'agit alors d'identifier à quel point une sélection de variables peut impacter les explications finales.
- La **sélection d'instances ayant un pouvoir explicatif significatif**. À partir des explications locales, il est important de pouvoir en identifier des groupes ayant un comportement similaire et offrir à l'utilisateur les grandes tendances d'explications. Nous avons ainsi proposé une approche par clustering assurant que les clusters d'explications obtenus sont bien caractéristiques de relations particulières entre variables, vis-à-vis d'un prédictif donné et ayant un sens côté métier.
- L'**analyse de données aidée par les explications**. Au travers d'un cas médical, nous montrons comment les explications locales et les clusters associés permettent de compléter une analyse statistique et prédictive classique, en mettant en avant les nouvelles relations découvertes entre variables.

1.3 Plan du mémoire

Les travaux de recherches exposés dans ce mémoire s'articulent en six chapitres :

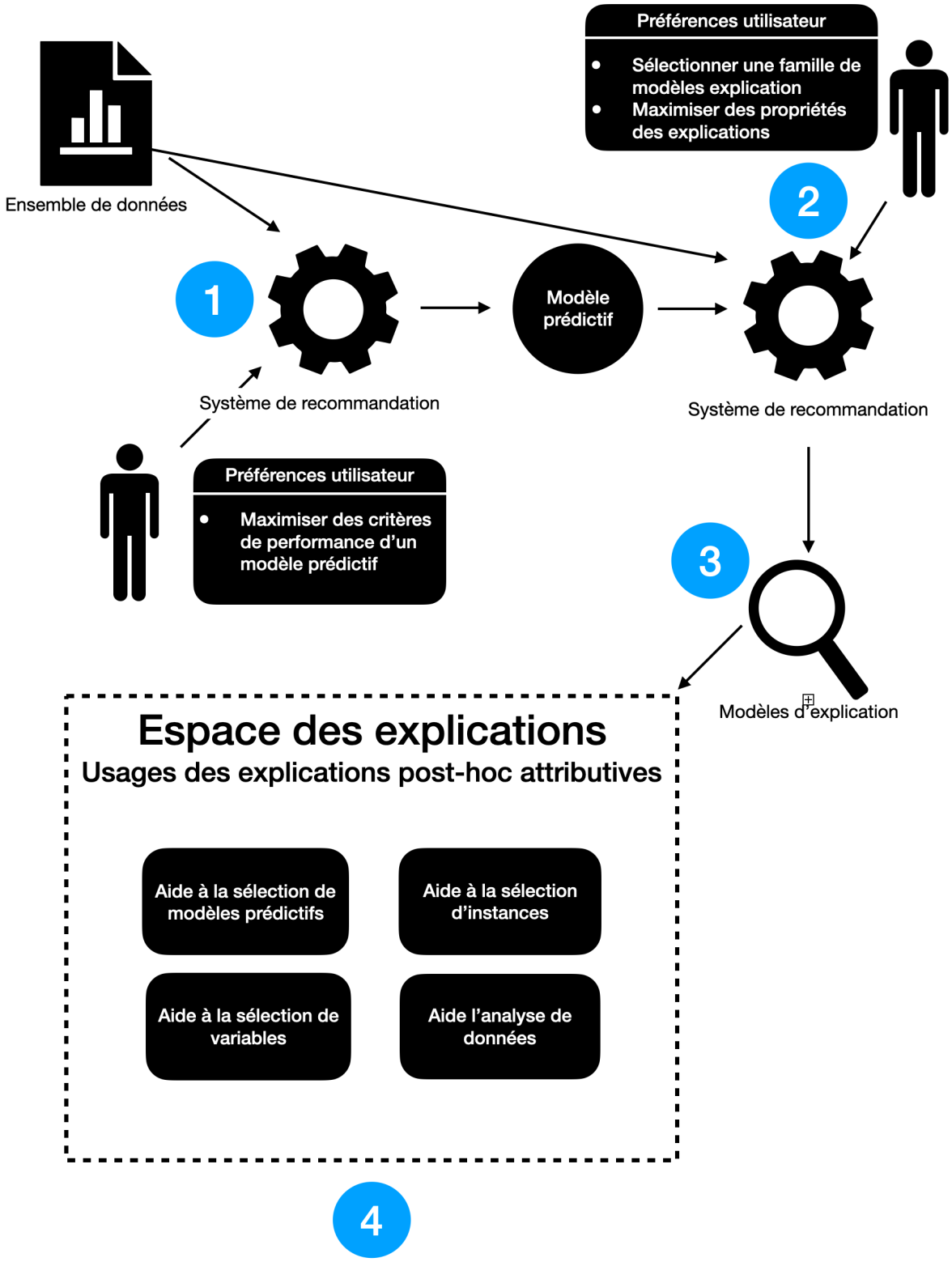


FIGURE 1.1 – Organisation générale des travaux de recherche menés dans ce mémoire.

— Le Chapitre 2 présente les domaines abordés tout au long de ce mémoire, à savoir l'apprentissage automatique, la recommandation de modèles prédictifs et l'explicabilité de modèles prédictifs. En particulier, nous détaillons, pour chacun de ces domaines, leur rapport avec l'utilisateur et, plus généralement, le système d'information qu'ils représentent.
— Le Chapitre 3 présente notre cadre de recommandation de modèles prédictifs. Notre proposition est notamment basée sur les concepts provenant de la recommandation contextuelle et de l'AutoML, en prenant en compte les préférences de l'utilisateur sous la forme de critères de performances souhaités dans le modèle à recommander.
— Le Chapitre 4 présente un cadre pour la recommandation de modèles XAI. Ce cadre recommande les modèles d'explication à l'aide de préférences utilisateurs matérialisées par les propriétés et métriques souhaitées sur ces modèles. Ce cadre s'inspire aussi de l'AutoML pour l'hyperparamétrage automatique des méthodes XAI.
— Le Chapitre 5 présente quelques limites à l'usage des explications post-hoc locales attributives. Nous y exposons le manque de considération des interactions entre variables dans le calcul des explications par des techniques populaires comme *SHAP* ou *LIME*. Afin de pallier ce problème, nous développons notre solution basée sur des calculs de coalitions issues de variables corrélées. Nous comparons ensuite les différentes méthodes locales afin de montrer les avantages et inconvénients de chacune.
— Le Chapitre 6 présente les explications locales attributives comme un nouvel espace de données utilisable dans quatre cas d'application : la sélection de modèles, la sélection de variables, la sélection d'instances et, plus globalement, l'aide à l'analyse de données.
— Le Chapitre 7 présente la valorisation des travaux de recherche exposés dans ce mémoire. Nous revenons sur les encadrements de thèse et mémoires de Master 2, les publications ainsi que les projets et collaborations ayant contribué à l'ensemble des résultats présentés.
— Le Chapitre 8 conclut et introduit les perspectives de recherches envisagées.

Chapitre 2

L'analyse prédictive et l'utilisateur

2.1 Introduction

Nous proposons, dans ce chapitre, un état de la littérature concernant l'analyse prédictive et les difficultés rencontrées par l'utilisateur quant à son utilisation, motivant ainsi notre proposition d'une aide à ce type d'analyse.

L'orientation prise dans ce mémoire est de ne se concentrer que sur les tâches d'apprentissage supervisé, notamment pour des raisons liées à la grande popularité de son usage, dans de multiples domaines comme la biologie/santé, la finance etc.

Ainsi, nous étudierons les avantages et limites de l'utilisation de l'apprentissage supervisé et de la manière dont les travaux de la littérature y répondent, en particulier via la proposition de systèmes de recommandation et de techniques d'explicabilité.

Tout au long de ce chapitre, nous discuterons également de la relation de l'apprentissage supervisé avec ses usages dans un Système d'Information afin d'en souligner à la fois le nombre limité de contributions dans ce domaine de recherche et de motiver la communauté à répondre aux limites bien réelles de l'apprentissage supervisé, surtout au vu de son usage toujours plus grandissant et populaire.

2.2 L'analyse de données et l'apprentissage automatique

L'analyse de données peut être assimilée à un processus de découverte de connaissances dans les bases de données (Knowledge Discovery Databases, ou KDD), incluant des tâches de prétraitement de données (comme le nettoyage et la transformation des données) ainsi que de modélisation [Runkler, 2020, Fayyad et al., 1996a, Fayyad et al., 1996b]. Une illustration de ce processus est proposée en Figure 2.1.

Il existe de nombreuses manières de modéliser les données afin de répondre à des finalités propres, qu'elles soient décisionnelles (BI) ou statistiques.

Alors que l'analyse décisionnelle, en particulier celle basée sur l'analyse OLAP (OnLine Analytical Processing), est aujourd'hui très bien définie et structurée dans la littérature

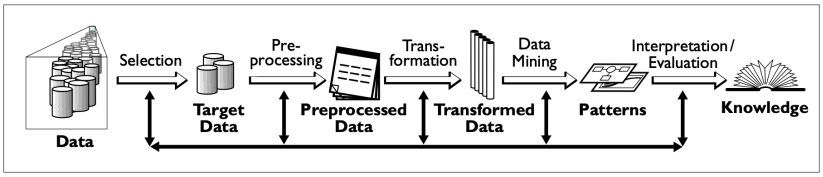


FIGURE 2.1 – Processus de KDD pour l'analyse de données, repris de [Fayyad et al., 1996b]

[Kimball, 1996, Golfarelli et al., 1998, Francia et al., 2022, Forresi et al., 2021], l'analyse statistique regroupe des méthodes variées et parfois très complexes, rendant plus difficile la mise en place d'un processus d'analyse unique, simple et clair pour l'utilisateur.

L'analyse statistique regroupe généralement les méthodes liées à l'analyse descriptive et l'analyse inférentielle. L'analyse descriptive [Ross, 2017] cherche à produire un résumé statistique des données (mesures quantitatives) alors que l'analyse inférentielle [Casella and Berger, 2021] s'attache à apprendre sur les données et les modéliser à l'aide de probabilités. La complexité de l'analyse inférentielle tient au fait qu'elle peut faire appel, entre autres, à des techniques d'intelligence artificielle (en particulier l'apprentissage automatique) menant à l'analyse prédictive [Siegel, 2013].

Nous proposons dans les sections suivantes de détailler la notion d'apprentissage automatique afin de montrer toute la difficulté de son usage et motivant le besoin d'aide à son utilisation dans le cadre d'une analyse prédictive.

2.3 L'apprentissage automatique : concepts clés

L'intelligence artificielle (IA) est aujourd'hui, et depuis plusieurs décennies maintenant, un enjeu majeur à la fois d'un point de vue technique, d'innovation/recherche et de société. *Une entreprise sur dix utilise maintenant dix applications d'IA ou plus ; les chatbots, l'optimisation des processus et l'analyse de la fraude sont à la tête des principaux cas d'utilisation*", d'après Forbes en 2020 [1].

L'apprentissage automatique (machine learning) est l'une des branches majeures de l'intelligence artificielle et est aujourd'hui appliqué dans une très grande variété de secteurs, comme la finance, le droit, la médecine, etc. Une étude Statista de 2023 [2] précise d'ailleurs les parts de marché entre ces secteurs, proposé en Figure 2.2, où l'industrie, la finance, la santé et le transport prédominent.

Toujours d'après Statista, une étude menée en 2019 [3] a montré qu'environ 80 milliards de dollars cumulés ont été investis dans l'intelligence artificielle, dont plus de la moitié (environ 42 milliards) a été consacrée au développement de l'apprentissage automatique seul.

1. https://www.forbes.com/sites/louiscolumbus/2020/07/09/10-ways-ai-is-improving-new-product-development/?sh=6da60e595d3c
2. https://www.statista.com/outlook/tmo/artificial-intelligence/machine-learning/worldwide
3. https://www.statista.com/chart/17966/worldwide-artificial-intelligence-funding/

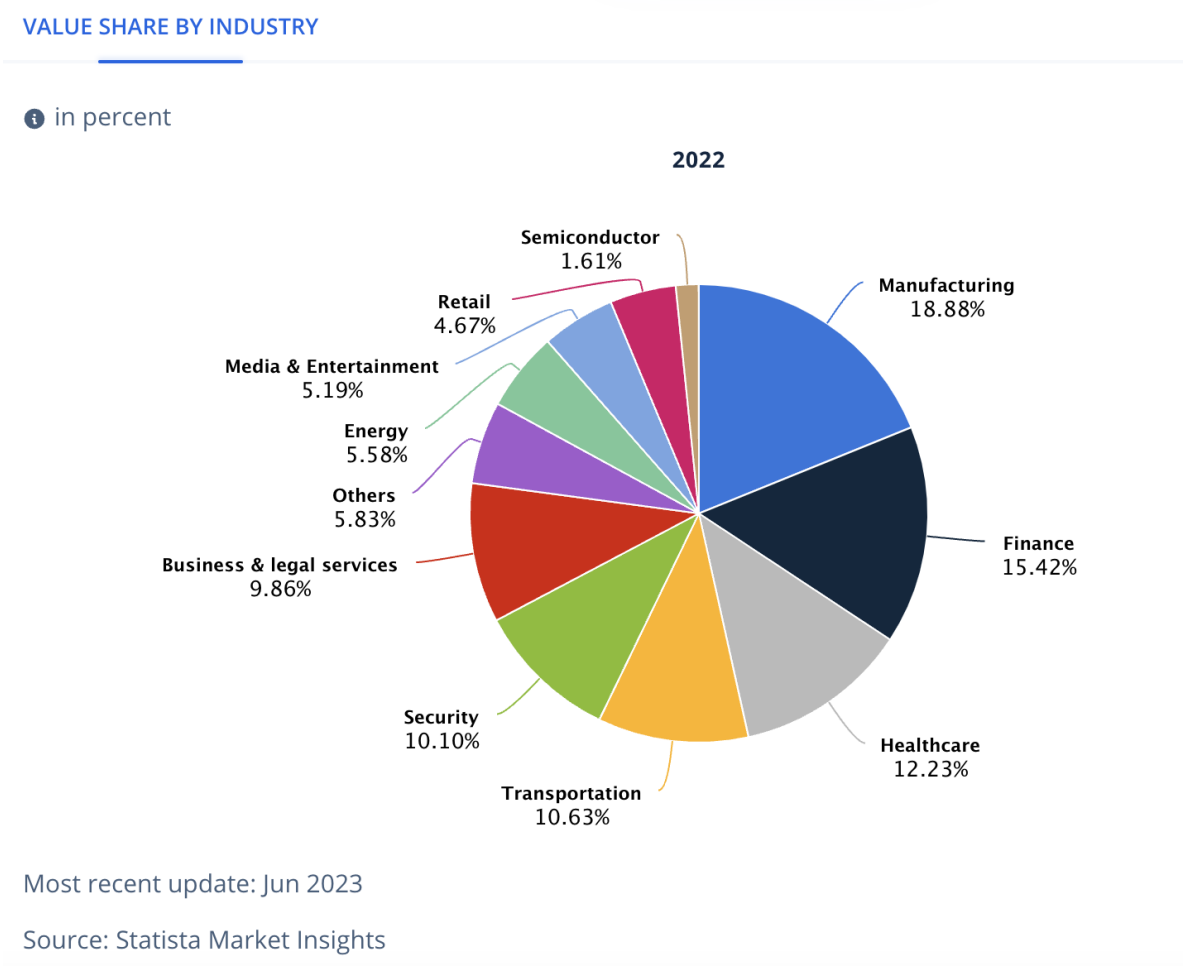


FIGURE 2.2 – Répartition des parts de marché de l'apprentissage automatique, par secteur d'activité

La réduction des coûts d'entreprise (38%), la génération d'informations et de connaissances sur les clients (37%), et l'amélioration de l'expérience client sont les trois cas d'utilisation les plus populaires en apprentissage automatique en entreprise, d'après l'étude de 2020 de Forbes [4].

Plus formellement, l'apprentissage automatique s'attache à tirer bénéfice d'expériences passées afin d'en construire, automatiquement, une modélisation la plus adaptée et précise possible [Jordan and Mitchell, 2015]. Il est traditionnellement décomposé en trois catégories :

— l'*apprentissage par renforcement*
— l'*apprentissage non-supervisé*
— l'*apprentissage supervisé*

L'apprentissage par renforcement [Watkins and Dayan, 1992, Kaelbling et al., 1996, Sutton and Barto, 2018] est basé sur la notion d'agent permettant de proposer, à partir d'expériences passées et face à un environnement donné, l'ensemble des actions maximisant l'obtention d'une récompense donnée.

L'apprentissage non supervisé [Hastie et al., 2009b, Ghahramani, 2003, Celebi and Aydin, 2016] consiste à construire un modèle automatiquement à l'aide des seules caractéristiques des données d'entrée. Dans le cadre du partitionnement de données, il s'agit de regrouper, dans une même classe, les données partageant les mêmes caractéristiques, à l'aide d'une mesure de similarité. Des techniques très populaires comme les k-means ou les méthodes hiérarchiques permettent leur mise en application.

L'apprentissage supervisé [Caruana and Niculescu-Mizil, 2006, Hastie et al., 2009a, Cunningham et al., 2008] considère, au contraire de l'apprentissage non supervisé, un jeu de données d'apprentissage et labélisées (où l'on indique, pour chaque donnée, une vérité terrain). Une fois le modèle produit, à l'aide de techniques comme SVM, Random Forest, XGBoost, Naive bayes etc., il s'agit de prédire la classe la plus probable (pour la prédiction d'une variable qualitative) ou une régression (pour la prédiction d'une variable quantitative) d'une nouvelle donnée n'ayant pas servi à l'apprentissage du modèle.

Ce dernier domaine étant au coeur des recherches développées dans ce mémoire, les sections suivantes reviennent en détail sur sa démocratisation, son usage et ses limites.

2.3.1 Apprentissage supervisé : motivation et définitions

Motivation

Le choix fait d'orienter les recherches proposées dans ce mémoire sur l'apprentissage supervisé tient du constat que ce domaine est encore (et toujours) largement utilisé parmi les autres domaines de l'apprentissage automatique [Jordan and Mitchell, 2015, Hastie et al., 2009b, Jurado et al., 2022, Hamida et al., 2021, Sharifani and Amini, 2023]. Ce que confirme aussi une étude de Gartner de 2020 [5] où la majeure partie de la valeur économique actuelle obtenue

4. https://www.forbes.com/sites/louiscolumbus/2020/01/19/roundup-of-machine-learning-forecasts-and-market-estimates-2020/?sh=6c0cbea55c02
5. https://www.gartner.com/smarterwithgartner/understand-3-key-types-of-machine-learning

grâce à l'apprentissage automatique est basée sur des cas d'utilisation de l'apprentissage supervisé. L'apprentissage non-supervisé est surtout efficace sur des problèmes plus spécifiques et l'apprentissage par renforcement est surtout limité par sa démocratisation, mais semble cependant très prometteur pour les années à venir.

Plus spécifiquement, le papier [Cohen, 2021] établit aussi que la grande majorité des algorithmes d'apprentissage automatique utilisés dans le domaine de la santé est surtout dominée par les approches d'apprentissage supervisé (en particulier les techniques SVM et de deep-learning, issu d'une étude de 2019 mené ici [6]).

Le succès de l'apprentissage supervisé doit sans doute beaucoup à son accès facilité par des bibliothèques du type *Scikit-learn* [7], à tout le moins pour des modèles de prédictions simples. L'avantage principal est aussi dû aux classes à prédire, connues à l'avance, permettant ainsi de construire des modèles de prédiction fidèles à une certaine réalité attendue [Pugliese et al., 2021].La dimension du "tout automatique" y est sans doute pour beaucoup également : à l'inverse de l'apprentissage non-supervisé, il n'y a pas besoin d'une intervention humaine pour choisir ce qui discriminera les données. Même si l'inconvénient de l'apprentissage supervisé est de devoir considérer un jeu de données d'apprentissage toujours plus important (d'autant plus accentué avec l'avènement du deep-learning), ce problème semble désormais limité aujourd'hui, au vu des masses de données gigantesques disponibles, et cela pour de nombreux secteurs. Nous reviendrons sur ces critiques et limites de l'apprentissage supervisé dans la section 2.3.3.

Définitions

L'*apprentissage supervisé* consiste à produire un modèle prédictif à partir de données d'apprentissage formées d'un ensemble de variables d'instances et d'une variable cible à prédire.

Plus formellement, nous adoptons les définitions classiques suivantes.

Définition 2.1 (Jeu de données). Considérons X, Y un jeu de données, avec $X = \{x_i\}_{i=1}^{n} | x_i \in \mathbb{R}^d$ les instances et $Y = \{y_i\}_{i=1}^{n} | y_i \in \mathbb{R}$ les labels correspondants, n est le nombre d'instances et d est le nombre de variables dans le jeu de données.

Définition 2.2 (Modèle prédictif). Un modèle prédictif est entraîné sur un jeu de données X, Y en déduisant des relations statistiques entre X et Y. Ce modèle peut alors être utilisé comme une fonction de prédiction notée $f : X \to \hat{Y}$ avec $\hat{Y} = \{\hat{y}_i\}_{i=1}^{n} | \hat{y}_i \in \mathbb{R}$, les prédictions produites.

2.3.2 L'apprentissage supervisé et l'usager

L'apprentissage supervisé et le système d'information

Les recherches menées et discutées dans ce mémoire sont majoritairement centrées sur l'utilisateur. L'apprentissage supervisé, tourné vers l'utilisateur, implique de s'intéresser

6. https://medium.com/sciforce/top-ai-algorithms-for-healthcare-aa5007ffa330
7. https://scikit-learn.org/

tout d'abord à sa relation avec un système d'information. De manière surprenante, la recherche dans le domaine du système d'information s'intéresse assez peu aux usages de l'apprentissage automatique. Comme l'a montré un travail particulièrement intéressant de [Abdel-Karim et al., 2021], assez peu de publications traitent de ce sujet dans les grands journaux reconnus en système d'information (Information Systems, Decision Support System, etc.). Même si le taux de publications a été multiplié par un peu plus de 4 entre 2009 et 2019, l'apprentissage automatique en reste le parent pauvre. Les raisons évoquées par [Abdel-Karim et al., 2021] sont multiples. Une première raison est due à l'aspect technique de l'apprentissage automatique pour lequel bon nombre de chercheurs en système d'information ne sont pas formés, marquant ainsi un manque d'expertise et de compréhension pour ce domaine. Le paramétrage des algorithmes d'apprentissage n'est pas non plus une chose aisée : il n'existe pas de processus ni de règles claires et génériques, applicables en toutes circonstances. Un autre aspect limitant, et indéniable, de l'apprentissage automatique tient à l'effet boîte noire des modèles. En effet, cette limite aura tendance à entraver la confiance de l'utilisateur dans le système et le conduire à rejeter le système (notamment pour des contextes plus sensibles, comme la santé) [Rai, 2020].

Il y a pourtant des atouts à motiver une recherche liant le système d'information et l'apprentissage automatique. [Abdel-Karim et al., 2021] a relevé que plus de la moitié des papiers traitant de ces sujets étaient composés d'équipes pluridisciplinaires et impliquaient une plus forte participation des entreprises ! De plus, les implications sociétales et économiques sont très fortes : mieux adapter et comprendre les risques liés à l'utilisation de l'apprentissage automatique c'est la démocratiser et permettre aux organisations (notamment de petites tailles) de rester compétitives. Une première solution proposée par [Abdel-Karim et al., 2021] est de voir l'application de l'apprentissage automatique dans un système d'information comme un processus de découverte de connaissances (KDD), de manière similaire à ce qui a été évoqué en Section 2.2 sur l'analyse de données. Un modèle d'apprentissage automatique est alors surtout issu d'un long processus mêlant une séquence d'étapes de prétraitement des données, de sélection de variables et d'ingénierie des fonctionnalités. Le choix d'un modèle de prédiction dépend ainsi des propriétés et structures des données à analyser.

L'apprentissage supervisé et le *human-in-the-loop*

Depuis quelques années maintenant, un nombre croissant de recherches [Wu et al., 2022] s'est investi dans les interactions possibles entre l'utilisateur et un système afin de l'aider à améliorer la construction d'un modèle prédictif [Vartak et al., 2015, Vartak et al., 2016]. Comme indiqué dans la section précédente, l'une des difficultés à l'élaboration d'un modèle d'apprentissage supervisé est son paramétrage délicat, et implique souvent des itérations du type essai/erreur [Xin et al., 2018]. D'après [Mosqueira-Rey et al., 2023], le principe du *human-in-the-loop* se base sur trois approches possibles, généralement assurées aussi par un processus itératif :

— l'*Active learning* [Settles, 2009] : l'utilisateur est considéré comme un Oracle et se limite à annoter des données que lui renvoie le système. Cette approche est notamment très

utile pour de l'apprentissage non-supervisé.
— l'*Interactive machine learning* [Amershi et al., 2014] : l'interaction est plus complète entre le système et l'utilisateur. L'utilisateur peut être amené à valider, nettoyer ou corriger des résultats proposés par le système. Il peut aussi directement proposer au système tout élément qu'il juge important à prendre en compte.
— le*Machine teaching* [Ramos et al., 2020] : l'utilisateur transfère sa connaissance et son expertise au système, à l'aide d'exemples fournis ou de correction d'erreurs détectées.

Ce principe de *human-in-the-loop* est, en particulier, très présent dans les milieux du traitement automatique du langage [De Angeli et al., 2021] ainsi que l'analyse d'images. Même si cela peut paraître étonnant, ce n'est sans doute pas complètement dû au hasard ! Comme le souligne [Mosqueira-Rey et al., 2023], la principale limite à l'application du *human-in-the-loop* pour l'apprentissage supervisé, est son effet boîte noire (comme souligné aussi dans la section précédente) et son manque de production d'explications. Or, les domaines du traitement automatique du langage et de l'analyse d'images sont, par essence, bien mieux compréhensibles pour un utilisateur à qui l'ont pourrait demander d'analyser ou valider, par exemple, des morceaux de phrases ou d'images.

Comme nous l'avons vu, les interactions de type *human-in-the-loop* permettent de mieux s'approprier un modèle d'apprentissage et certainement d'améliorer la confiance de l'utilisateur, surtout quand celui-ci n'est pas expert en informatique ou en apprentissage automatique. Les méthodes proposant l'élaboration de modèles prédictifs de manière totalement automatique, de type Auto-ML, peuvent d'ailleurs être une réponse à une meilleure interactivité avec l'utilisateur [Mosqueira-Rey et al., 2023]. Mais à la seule condition que ces méthodes soient capables d'expliquer le plus clairement possible les résultats produits.

Dans ce mémoire, nous nous focaliserons sur une approche d'*Interactive machine learning* car celui-ci assure une plus grande interaction entre l'utilisateur et le système au contraire des deux autres approches, tournées quasiment exclusivement sur l'avis de l'utilisateur.

2.3.3 Les limites de l'apprentissage supervisé

Depuis ses débuts, l'apprentissage supervisé souffre d'un certain nombre de limites rendant ce domaine difficile à prendre en main, surtout pour un non spécialiste. Nous pouvons lister quatre limites majeures, à savoir :
— l'*effet Rashomon* ;
— l'*effet de dilution* ;
— l'*hyperparamétrage* des algorithmes ;
— l'effet *boîte noire* des modèles prédictifs.

L'*effet Rashomon* fait référence au film japonais du même nom, datant de 1950, où plusieurs témoins d'un même meurtre fournissent différentes analyses de la scène. Le domaine de l'épistémologie [Anderson, 2016] a repris ce principe permettant de décrire des phénomènes, souvent complexes, à partir de différentes façons de penser ces mêmes phénomènes. Dans le cadre de l'apprentissage supervisé, la multiplicité du nombre d'algorithmes possibles provoque alors cet effet [Wang and Tao, 2008, Semenova et al., 2022]. En d'autres termes, les modèles

prédictifs produits sont avant tout subjectifs et ne sont que des points de vue possibles sur leur manière de généraliser des données d'apprentissage. Dès lors, comment choisir le meilleur modèle ? Les mesures classiques des modèles prédictifs, telle la précision, sont un premier indice de la qualité d'un modèle, mais ne peuvent être une fin en soi. Comme l'a montré [Alanazi et al., 2017], différents modèles produits pour un même jeu de données peuvent être évalués de manière très différente dans la littérature, montrant tout simplement la complexité à produire un modèle unique, bien paramétré et facilement adapté aux données.

L'*effet de dilution* [Wang and Tao, 2008] concerne le problème de haute dimensionnalité dans les jeux de données. Pour un même ensemble d'apprentissage, plus le nombre de variables est important, plus l'ensemble d'apprentissage sera dilué dans cette haute dimensionnalité. Cela signifie qu'il est possible de générer beaucoup de modèles prédictifs, mais dont seulement une toute petite partie d'entre eux s'avéreront capables de généraliser correctement l'ensemble d'apprentissage. Cela rend donc encore plus difficile la recherche d'un bon modèle prédictif. Ce phénomène montre également toute l'importance que représente l'application d'un pré-processing adéquat, en particulier la sélection de variables.

L' *hyperparamétrage* des algorithmes d'apprentissage implique de devoir renseigner un certain nombre de paramètres pour la construction d'un modèle [Probst et al., 2019]. La recherche adéquate de ces valeurs de paramètres est souvent très fastidieuse, spécifique au type de modèle de prédiction et des données à analyser. L'utilisateur a le choix entre garder les valeurs par défaut proposées par l'algorithme ou bien de les spécifier par son expérience passée (et parfois recommandées par la littérature) ou encore par essai/erreur. Pour pallier ce problème, l'automatisation de la recherche de l'hyperparamétrage est aussi possible à l'aide de techniques comme grid ou random search [Bergstra and Bengio, 2012] ou par optimisation Bayésienne [Snoek et al., 2012]. Cette automatisation présente aussi ses défauts, où la recherche optimale des hyperparamètres n'est pas toujours acquise et peut mener à des temps de calcul considérables.

La plupart des modèles prédictifs a pour conséquence de générer un *effet boîte noire* pouvant être problématique pour l'utilisateur [Goodman and Flaxman, 2017]. Cet effet empêche l'utilisateur de comprendre les raisons pour lesquelles un modèle prédictif propose telle ou telle décision. Ce problème est aussi d'autant plus accentué lorsqu'il touche des utilisateurs non spécialistes de l'apprentissage automatique, en particulier ceux touchant à des secteurs plus sensibles comme la médecine, la défense ou le juridique [Barredo Arrieta et al., 2020]. Ce phénomène va aussi grandissant, à mesure que les algorithmes d'apprentissage automatique se perfectionnent, souvent d'ailleurs dans le seul objectif d'améliorer la précision des résultats obtenus, sans se soucier alors de les rendre plus transparents aux utilisateurs. Les propositions du "tout automatique" comme les approches autoML [Feurer et al., 2015a, He et al., 2021] amplifient, en outre, cet *effet boîte noire*. Comme indiqué dans [Barredo Arrieta et al., 2020, Došilović et al., 2018], il y a bien un compromis à trouver entre la précision d'un modèle prédictif et sa capacité à offrir un résultat le plus transparent possible.

Ce passage en revue des limites de l'apprentissage supervisé montre qu'il est difficile, pour un utilisateur seul, de pouvoir choisir le modèle le plus approprié, et correctement paramétré,

à son cas d'usage. Voilà pourquoi un certain nombre de travaux se sont concentrés sur la proposition de systèmes de recommandation spécifiques à l'apprentissage supervisé. La section suivante revient sur ces propositions.

2.4 La recommandation de modèles prédictifs

Les limites de l'apprentissage supervisé, évoquées dans la section précédente, montrent bien qu'il est difficile, surtout pour un non-expert, de choisir le modèle le plus approprié avec les paramétrages adéquats pour un jeu de données à analyser. Ainsi, l'idée de pouvoir recommander, automatiquement, le modèle prédictif le plus pertinent semble être la solution idéale.

En particulier, l'apprentissage automatique automatisé (Automated machine learning, AutoML [Feurer et al., 2015a, He et al., 2021]) est une voie très prometteuse depuis quelques années. Le papier [Vanschoren, 2019] en fait d'ailleurs un très bon panorama en listant les principales catégories possibles. Elles ont toutes en commun de se baser sur une analyse des workflows d'apprentissage déjà réalisés par le passé par d'autres utilisateurs. C'est ce que l'on appelle le *Méta-apprentissage (Meta-learning)*.

En complément des définitions de jeux de données et de modèles précisées en Section 2.3.1, nous définissons la notion de workflow et de tâches pour l'apprentissage supervisé.

Définition 2.3 (Workflow). Un Workflow consiste en une série d'algorithmes et d'actions d'analyse de données, avec leurs hyperparamètres, menant à la création d'un modèle supervisé, suivant un processus de KDD, comme discuté en Section 2.2. À titre d'exemple, un workflow pourrait consister en l'élimination des valeurs manquantes du jeu de données étudié, puis en une sélection de variables, suivie de l'apprentissage d'un arbre de décision sur le jeu de données résultant.

Définition 2.4 (Tâche). Une Tâche d'apprentissage supervisé décrit les objectifs d'un utilisateur, pour un jeu de données à analyser. Elle comprend les mesures de performance (propriétés) les plus importantes pour l'utilisateur. Il s'agit, par exemple, de la précision du modèle, son nombre de faux positifs ou négatifs, ou des mesures plus complexes comme l'aire sous la courbe ROC [Caruana and Niculescu-Mizil, 2006].

2.4.1 Le meta-apprentissage et la recommandation de modèles prédictifs

Selon [Vanschoren, 2019], nous considérons trois types de méta-apprentissage :
- Méta-apprentissage à partir de l'évaluation de modèles ;
- méta-apprentissage à partir de modèles précédents ;
- méta-apprentissage à partir des propriétés des tâches.

Méta-apprentissage à partir de l'évaluation de modèles

A partir d'un ensemble de workflows passés et d'évaluations faites de ces workflows pour des tâches d'apprentissage prédictif, l'idée est de recommander le workflow le plus optimal pour une nouvelle tâche prédictive. Plusieurs méthodes de la littérature ont été proposées pour répondre à cette idée. Parmi celles-ci, on peut citer le *transfert de workflow* : si un ou plusieurs workflows performent très bien sur des tâches passées, il est fort probable que ces workflows fonctionnent correctement sur une nouvelle tâche similaire. Le problème de similarité entre tâches peut être résolu à l'aide de techniques comme le *relative landmarks* [Fürnkranz and Petrak, 2001, Leite et al., 2012] ou les *modèles de substitution* [Wistuba et al., 2015, Feurer et al., 2018], notamment. Concernant la technique *relative landmarks*, il s'agit d'employer des modèles simples et rapides (de type Naive Bayes, arbre de décision, etc.) jouant le rôle de landmarks et de mesurer l'écart de performance vis-à-vis de tous les worflows passés. Le landmarker qui performe le mieux parmi tous les workflows sera considéré comme applicable pour la nouvelle tâche. Cette technique permet notamment de gagner du temps sur l'exécution des modèles prédictifs. La technique de *modèles de substitution* génère, quant à elle, des modèles de substitution (par exemple à l'aide d'un processus gaussien) pour chaque modèle des workflows passés. Le même principe que la technique précédente est ensuite appliquée : le modèle de substitution performant le mieux parmi les workflows passés est sélectionné pour être appliqué à la nouvelle tâche.

Méta-apprentissage à partir de modèles précédents

Il s'agit principalement des techniques liées à l'apprentissage par transfert (transfert-learning) [Thrun and Pratt, 1998]. Le but est de réutiliser un modèle précédent (mais relativement proche de la nouvelle tâche à réaliser) et l'utiliser comme point de départ pour répondre à une variable cible similaire [Evgeniou et al., 2005, Sharif Razavian et al., 2014]. Par exemple, si un modèle prédictif est capable de bien repérer un chat dans une image, on peut supposer qu'il servira d'une bonne base pour repérer un félin, plus généralement.

Méta-apprentissage à partir des propriétés des tâches

Le principe de cette méthode est d'utiliser une autre source d'information disponible dans les workflows : les métadonnées. Ces métadonnées sont en général des métriques liées aux ensembles de données analysés (ou à analyser). Il existe un nombre très varié et important de ces métadonnées dédiées aux ensembles de données : on appelle d'ailleurs ces métadonnées des meta-feature. Une liste possible est disponible dans [Vanschoren, 2019] ou encore dans [Raynaut, 2018]. Il peut s'agir de meta-feature portant simplement sur le nombre d'instances, de variables, mais aussi plus statistiques comme des scores de corrélations ou bien provenant de la théorie de l'information (comme l'entropie) ou de la qualité des données (cohérence des données, par exemple). Il peut s'agir aussi d'informations portant sur des scores de performance des modèles landmarker exécutés sur le jeu de données visé.

L'idée est ensuite d'utiliser ces meta-feature au sein d'une mesure de similarité permettant d'estimer la proximité entre les jeux de données des tâches passées et de la tâche nouvelle. Pour le jeu de données de la tâche passée le plus similaire, on identifie ensuite le workflow le plus performant pour être recommandé à la tâche nouvelle. Le workflow recommandé est souvent accompagné d'une optimisation des hyperparamètres, visant à améliorer la performance du modèle final [Feurer et al., 2019].

Ce type de méta-apprentissage à partir des propriétés des tâches nous semble la voie la plus prometteuse, qui est aussi à la base de la proposition du populaire Auto-sklearn [Feurer et al., 2019]. En effet, contrairement à l'approche par évaluation des modèles, cette approche permet de conserver les modèles originaux tout en les paramétrant automatiquement au nouveau jeu de données à analyser. La solution basée sur le transfert learning, quant à elle, requiert une expertise en apprentissage automatique trop importante pour être utilisable facilement.

Dans les sections suivantes, nous détaillons la recommandation de workflows par méta-apprentissage des propriétés des tâches.

2.4.2 La recommandation de workflow par meta-apprentissage

Les étapes générales pour la recommandation de workflow par méta-apprentissage des propriétés des tâches sont présentées Figure 2.3. Ce schéma est celui adopté par [Feurer et al., 2019] et [Raynaut, 2018, Raynaut et al., 2017a] et détaillé comme suit :

1. A partir des jeux de données analysées par de précédents utilisateurs, le système identifie le sous-ensemble similaire au jeu de données à analyser (à l'aide des meta-features, comme indiqué dans la section précédente).
2. Parmi le sous-ensemble trouvé, le système classe les workflows utilisés pour ces ensembles de données, en fonction des indicateurs de performance souhaités par l'utilisateur.
3. Le workflow le plus performant sur un jeu de données similaire à celui de l'utilisateur est alors recommandé à ce dernier.

La difficulté de ce type d'approche, et plus généralement pour tout ce qui concerne le meta-apprentissage, est de disposer de meta-base contenant à la fois des ensembles de données (ou au moins ses meta-features) ainsi que des workflows préalablement conçus et exécutés sur ces données (afin d'en récupérer toute sorte d'indicateurs de performance). Une solution possible peut venir de plateformes du type de celles proposées par OpenML [8] [Vanschoren et al., 2013]. Le but de cette plateforme est de collecter et stocker un large nombre de workflows d'apprentissage automatique, ainsi que les ensembles de données correspondant, afin d'encourager la collaboration entre les utilisateurs. L'avantage de cette plateforme est que l'utilisateur peut spécifier le type de tâche qu'il souhaite réaliser sur un jeu de données avec les paramètres lui permettant, en particulier, de spécifier les critères d'évaluation qu'il souhaite. Cette plateforme

8. https://www.openml.org/

a permis de faciliter l'évaluation de travaux de recherches comme dans [Feurer et al., 2015a] ou [Doshi-Velez and Kim, 2017]. Pour toutes ces raisons, c'est également cette plateforme sur laquelle nous nous baserons dans ce mémoire pour l'évaluation de la plupart de nos contributions.

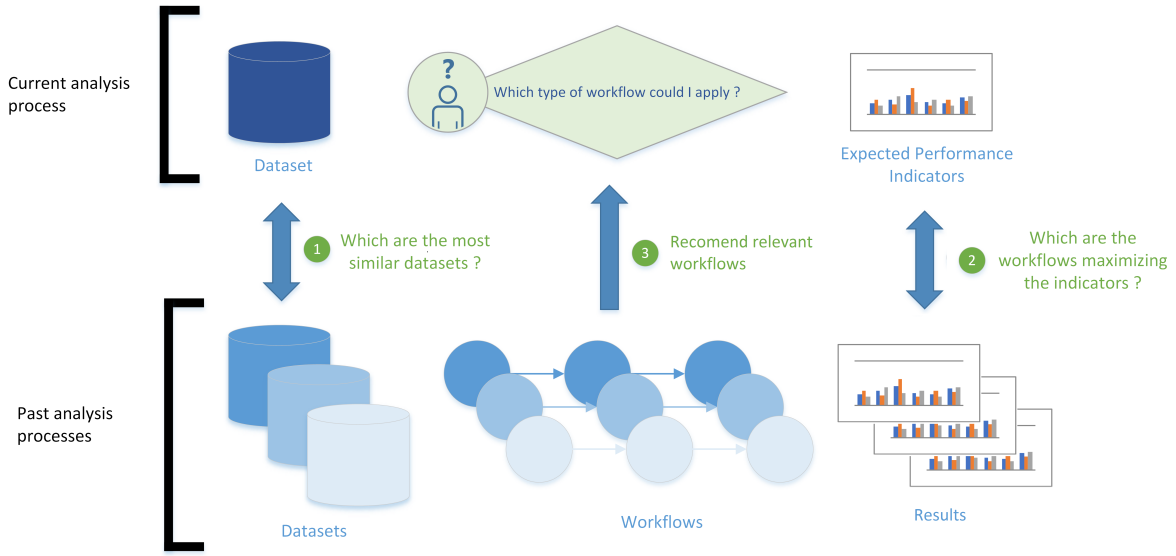


FIGURE 2.3 – Etapes principales pour la recommandation de workflows selon [Feurer et al., 2019] et [Raynaut, 2018, Raynaut et al., 2017a]

Déterminer la dissimilarité entre deux ensembles de données

La recommandation de workflow basé sur la comparaison de meta-features entre jeux de données suppose de pouvoir en définir une mesure de similarité. Les propositions de [Feurer et al., 2019] et [Raynaut, 2018, Raynaut et al., 2017a] abordent, en particulier, cette problématique. Le travail de [Raynaut, 2018] ayant montré que leur proposition de mesure de dissimilarité, par meta-feature, entre jeux de données, surpassait les autres propositions de la littérature, nous proposons de la détailler ci-après.

[Raynaut, 2018] propose notamment une mesure de dissmilarité assurant que chaque meta-feature est bien pris en compte dans le calcul. En effet, si plusieurs meta-feature sont calculés pour une même variable (moyenne, écart-type des valeurs par exemple), la littérature propose en général d'en faire une simple moyenne... Cette technique aura alors tendance à faire perdre une quantité d'information potentiellement importante. Ainsi, l'approche clé de [Raynaut, 2018] est de comparer les variables par paire, entre deux jeux de données. La comparaison entre paires de variables mène alors à mesurer une dissimilarité entre l'ensemble des meta-feature de l'attribut visé. Cette astuce permet notamment de s'abstraire d'un ordre entre variables (qui n'a pas lieu d'être lorsque l'on raisonne par meta-feature) Si le nombre

de variables entre deux jeux de données est différent, les meta-feature du ou des variables en sus sont comparés avec des valeurs vides. La proposition repose aussi sur deux mesures de dissimilarité : une dissimilarité entre meta-features des variables et une dissimilarité entre meta-features des jeux de données.

Le choix d'une dissimilarité et non d'une distance réside dans le fait que même si deux jeux de données sont physiquement différents (par l'ordre différent de leurs variables), il faut pouvoir les considérer comme parfaitement similaires si leur dissimilarité est nulle. Ainsi, et toujours d'après [Raynaut, 2018], la mesure de dissimilarité doit répondre aux propriétés nécessaires suivantes :

Définition 2.5. Supposons A un ensemble et d une fonction : $A^2 \to \mathbb{R}$.
d est une **fonction de dissimilarité** sur A si et seulement si, $\forall x, x' \in A$:
- $d(x, x') \geq 0$ (Positivité)
- $x = x' \to d(x, x') = 0$ (Indiscernabilité des identiques)
- $d(x, x') = d(x', x)$ (Symétrie)

Le détail des mesures de dissimilarités employées peut être retouvé dans [Raynaut, 2018].

2.4.3 Les limites à la recommandation de modèles prédictifs

Le manque de prise en compte du contexte utilisateur

Comme l'a déjà très bien montré l'étude de [Adomavicius and Tuzhilin, 2010], le prise en compte du contexte utilisateur dans un système de recommandation ne peut qu'améliorer la pertinence de ses résultats. Les systèmes de recommandation tenant compte du contexte devraient être préférés quand le contexte utilisateur est d'autant plus important et complexe. Un certain nombre d'approches ont été proposées ces dernières années, dans des applications très différentes. Comme par exemple, la détection d'émotions [Ishanka et al., 2017] ou encore la suggestion de nouvelles collaborations entre entreprise et universitaires à l'aide de contextes liés aux chercheurs [Wang et al., 2017]. [Adomavicius and Tuzhilin, 2010] structure le domaine de la recommandation basée sur le contexte utilisateur en trois principaux types de méthodes :

- par pré-filtrage : le contexte de l'utilisateur est considéré comme des données additionnelles aux données d'origine servant au système de recommandation. Cela peut permettre, par exemple, de donner plus de poids aux données d'origine partageant un même contexte.
- par post-filtrage : le contexte de l'utilisateur n'est considéré qu'en filtrage des résultats déjà produits par le système de recommandation. Cela peut servir, par exemple, à re-classer une liste d'items recommandés selon un contexte d'utilisateur donné.
- par filtrage basé sur le modèle : le contexte de l'utilisateur est considéré comme une nouvelle dimension des données à analyser par le système de recommandation. Le contexte n'influence ainsi pas directement les données d'origine. Cette approche permet, par exemple, de générer plusieurs classements en fonction des différentes préférences à prendre dans le contexte de l'utilisateur. La recommandation finale peut être

ainsi générée à l'aide d'un front de Pareto entre les différents classements proposés, afin d'identifier l'item le mieux classé entre tous les classements.

Dans le cadre de la recommandation de workflow par meta-apprentissage, il s'agit souvent de ne maximiser qu'un seul indicateur de performance (traditionnellement, un score de précision). Cependant, il nous semble utile de pouvoir considérer un vrai contexte utilisateur dans ce domaine où celui-ci peut être intéressé par la recommandation d'un modèle maximisant à la fois un score de précision, de rappel ou encore un Kappa de Cohen [Cohen, 1968]. Ainsi, la méthode de filtrage basé sur le modèle nous semble tout indiquée pour répondre à ce problème, comme nous le détaillerons dans le Chapitre 3.

Le manque de confiance dans les recommandations

La recommandation de modèle d'apprentissage supervisé recherche, en général, à proposer le modèle le plus précis possible. Egalement, ce type de système limite souvent les interactions possibles par l'utilisateur, par exemple pour vérifier la validité du modèle proposé ou bien lui permettre de le personnaliser davantage.

Ces inconvénients peuvent mener à ce que l'utilisateur perde confiance dans les modèles proposés et finisse, tout simplement, pas ne plus les utiliser. Le problème principal porte sur le manque de transparence et donc d'explications fournies à l'utilisateur. Ce problème est d'autant plus accentué qu'il touche à la fois sur le besoin de comprendre le processus ayant permis d'obtenir les recommandations (et leur classement) ainsi que sur l'explication des modèles de prédiction eux-mêmes.

L'importance de la transparence est reconnue depuis très longtemps, en particulier dans les systèmes experts [Buchanan and Shortliffe, 1984], et plus récemment dans les systèmes de recommandation [Tintarev and Masthoff, 2015, Chanson et al., 2021, Zhong and Negre, 2022]. En particulier, [Tintarev and Masthoff, 2015] considère qu'un système de recommandation doit prendre en compte les notions suivantes :

— Transparence (i.e. comment le système fonctionne comme dans [Cramer et al., 2008]),
— Confiance (i.e. la confiance de l'utilisateur dans les recommandations comme dans [Chen and Pu, 2005]),
— Acceptation (i.e. l'acceptation de l'utilisateur dans les recommendations fournies comme dans [Cramer et al., 2008]),
— Efficacité (i.e. produire de meilleures décisions comme dans [Shani et al., 2013]),

Les travaux de [Chanson et al., 2021, Zhong and Negre, 2022] ont pour objectif de mieux comprendre ce qui a pu influencer les recommandations via le système, notamment en étudiant l'influence des données d'entrée (et plus particulièrement leurs variables).

Le cadre juridique et politique pousse aussi à toujours plus de transparence et d'explications liées aux algorithmes d'intelligence artificielle que ce soit au niveau français (les obligations liées au RGPD [9] ainsi que le rapport Villani [10]) ou européen (notamment via le futur AI

9. https://www.cnil.fr/fr/rgpd-de-quoi-parle-t-on
10. https://www.enseignementsup-recherche.gouv.fr/fr/rapport-de-cedric-villani-donner-un-sens-l-intelligence-artificielle-ia-49194

act [11]).

Pour toutes ces raisons, il nous paraît indispensable d'associer à un système de recommandation de workflow par meta-apprentissage un système permettant d'expliquer à l'utilisateur le modèle prédictif proposé. Au vu de la complexité que représente un modèle prédictif en soi, il nous semble plus judicieux d'investir d'abord sur les moyens de fournir un système permettant d'expliquer au mieux les prédictions faites par le modèle, afin de maximiser les chances de son adoption par l'utilisateur. Il n'en reste pas moins que l'explication du processus de recommandation de workflow par meta-apprentissage reste un enjeu stratégique et sera bien évidemment évoquée dans les perspectives à la fin de ce mémoire, disponible en Chapitre 8.

2.5 L'explicabilité pour les modèles prédictifs

2.5.1 L'explicabilité : concepts clés

L'explicabilité en intelligence artificielle (XAI), et plus particulièrement dans le domaine de l'apprentissage automatique, est à la fois une notion ancienne et une préoccupation plutôt récente [Miller, 2019]. Ancienne car l'explicabilité a connu un bond important il y a de cela trente ans, à l'époque des contributions de l'IA symbolique avec les systèmes experts [Chandrasekaran et al., 1989, Swartout and Moore, 1993, Buchanan and Shortliffe, 1984]. D'ailleurs, l'explicabilité était aussi une préoccupation de l'époque pour les chercheurs en Système d'Information, comme le rappelle très bien [Meske et al., 2022]. Récente car l'avènement de l'apprentissage automatique a accentué le problème du manque de confiance dans les applications de l'IA, dû à des problèmes éthiques par exemple [Angwin et al., 2016].

C'est ce que confirme aussi une étude de McKinsey [12], en affirmant que les organisations qui établissent la confiance numérique des consommateurs par des pratiques telles que l'explication de l'IA sont plus susceptibles de voir leur chiffre d'affaires annuel croître à des taux de 10% ou plus. Cette étude considère aussi que l'explicabilité pour l'IA ne pourra être que profitable aux organisations, et cela pour au moins cinq raisons : (1) l'augmentation de la productivité, (2) l'instauration de la confiance et d'une meilleure adoption de l'IA, (3) l'émergence de nouvelles pratiques, génératrices de valeur, (4) l'assurance que l'IA apporte une valeur ajoutée à l'entreprise et (5) Atténuer les risques réglementaires. Elle en conclut aussi :

> *"Les gens utilisent ce qu'ils comprennent et ce en quoi ils ont confiance. C'est particulièrement vrai pour l'IA. Les entreprises montrant facilement comment leurs idées et recommandations en matière d'IA y parviennent en sortiront gagnantes, non seulement auprès des utilisateurs de l'IA de leur organisation, mais aussi auprès des régulateurs et des consommateurs, et en termes de résultats financiers."*

11. https://artificialintelligenceact.eu
12. https://www.mckinsey.com/capabilities/quantumblack/our-insights/why-businesses-need-explainable-ai-and-how-to-deliver-it

L'explicabilité de l'IA dans le Système d'Information

L'explicabilité est une notion, de fait, très pluridisciplinaire, mêlant à la fois des compétences liées à l'intelligence artificielle, les sciences sociales, l'interaction humain-machine ([Bove et al., 2023]) et l'interaction humain-agent, comme souligné par [Miller, 2019]. Ce caractère pluridisciplinaire ne peut que renforcer l'importance de l'étudier aussi dans le domaine du Système d'Information.

Comme discuté dans [Meske et al., 2022], et finalement en concordance avec l'étude de [Abdel-Karim et al., 2021] sur les faibles contributions de la recherche en Système d'Information appliquée à l'apprentissage automatique (voir Section 2.3.2), les contributions dans le domaine de l'XAI sont encore balbutiantes [Eiras-Franco et al., 2019, Giboney et al., 2015, Martens and Provost, 2014], même si une plus forte augmentation est constatée depuis 2021 [Brasse et al., 2023]. Pourtant l'enjeu est là encore très important et l'explicabilité ne peut pas se limiter qu'à une considération technique : elle offre aussi de multiples perspectives concernant la place de l'XAI (et son processus) dans une organisation, l'influence de l'explicabilité auprès des utilisateurs, la proposition de métriques adéquates pour l'évaluation des explications, les problèmes éthiques et moraux, etc. Comme souligné par [Brasse et al., 2023], repris de [Chromik and Butz, 2021], dans l'action humaine quotidienne, *l'explication est un processus social et itératif entre un explicateur (i.e celui qui explique) et un destinataire de l'explication.*

C'est d'ailleurs dans cet objectif itératif que s'inscrit une partie des travaux présentés dans ce mémoire mêlant explicabilité et recommandation de modèle prédictif, comme détaillé dans le Chapitre 6.

L'explicabilité en apprentissage automatique

Comme indiqué dans les sections précédentes, l'effet boîte noire lié aux modèles d'apprentissage automatique est un problème majeur. De nombreuses méthodes sont aujourd'hui disponibles permettant d'expliquer un modèle d'apprentissage automatique. Selon la taxonomie de [Molnar, 2020], nous pouvons retrouver la classification suivante pour ces méthodes :

- les modèles intrinsèques ou post hoc : les explications sont déduites soient d'un modèle dit intrinsèquement explicable (comme peut l'être un arbre de décision par exemple), soient issues des réponses du modèle prédictif, après entraînement.
- les modèles spécifiques ou agnostiques au modèle : les modèles spécifiques ne sont applicables que pour un type d'algorithme d'apprentissage particulier. Par exemple, les modèles intrinsèquement explicables sont toujours spécifiques aux modèles prédictifs. Il peut aussi s'agir de certaines implémentations de type post-hoc comme avec TreeSHAP (voir Section 2.5.2). Les méthodes agnostiques au modèle sont applicables pour tout modèle prédictif et sont toutes, de fait, post-hoc.
- les modèles locaux or globaux : les modèles locaux produisent des explications pour des instances ou des groupes d'instances. Les modèles globaux peuvent simplement agréger les résultats de modèles locaux ou bien permettre de comprendre le comportement

interne du modèle, en jouant sur l'influence de certains éléments internes (par exemple étudier les coupures dans un arbre de décision).

Les types de résultats produits par les méthodes d'explication sont aussi très larges : certaines vont produire des instances (nouvelles ou non), d'autres des vecteurs de poids ou encore des modèles de substitutions au modèle de prédiction initial.

Dans les sections suivantes, nous ferons un focus sur les méthodes locales post-hoc agnostiques au modèle, principalement utilisées dans ce mémoire. Ces dernières étant sans doute plus facile d'accès à tout utilisateur (expert ou non) que les méthodes intrinsèques (souvent liées à des modèles trop simplistes), ou celles cherchant à mieux comprendre les phénomènes internes d'un modèle.

Dans la suite de ce mémoire, nous faisons aussi le choix de considérer, indifféremment, les notions d'explication et d'interprétation et de les réunir autour du seul terme d'*explicabilité*. Nous soutenons cette proposition par le fait que tout ce qui a trait à produire une justification pour mieux comprendre une instance, au travers d'un modèle d'apprentissage supervisé et au regard du jeu de données de départ, est une explication. Les moyens de mettre en valeur ces dernières permettront alors de mieux guider l'utilisateur à la recherche de possibles relations, voire d'identifier des causalités, dans le jeu de données. Nous insistons d'ailleurs sur le fait qu'une explication n'implique pas nécessairement une causalité, comme pour toute exploitation de méthode statistique. L'important étant que l'explicabilité soit au service de l'utilisateur et l'aide à imaginer toute hypothèse crédible le menant à conduire des analyses ultérieures et complémentaires permettant de la confirmer ou non.

2.5.2 Les explications locales post-hoc, agnostiques au modèle

Comme indiqué précédemment, ce mémoire a pour objectif de se concentrer principalement sur l'usage d'explications locales post-hoc. Nous proposons ainsi un panorama des méthodes existantes de la littérature, en revenant sur les avantages et inconvénients de chacune. Nous nous basons notamment sur la structure et le travail de synthèse proposée par [Molnar, 2020].

Espérance conditionnelle individuelle (ICE)

Cette méthode [Goldstein et al., 2015] est très simple et basique : elle permet de représenter comment évolue, pour une instance, le changement d'une prédiction en fonction des valeurs d'une variable donnée. On représente cette évolution sous la forme d'un diagramme d'*Espérance conditionnelle individuelle*, où une ligne correspond à une instance. Le diagramme moyennant l'ensemble des lignes est appelé diagramme de dépendance partielle (faisant partie des méthodes d'explication globale).

Cette méthode, bien que très simple à comprendre, reste avant tout très basique et ne permet pas, par exemple, de pouvoir analyser plusieurs variables à la fois, rendant une vue globale de l'instance à expliquer difficile.

Explications contrefactuelles

Une explication contrefactuelle, pour une prédiction donnée, est la plus petite modification à apporter aux valeurs d'une ou de plusieurs variables afin de faire changer la classe/régression prédite par le modèle. L'idée est, plus généralement, d'identifier ce qui dans les données d'entrée, influe sur la sortie du modèle pour une prédiction donnée. Le but est donc de chercher le ou les plus petites modifications venant à changer la prédiction. Pour cela, plusieurs techniques ont été proposées, notamment [Wachter et al., 2017] et [Dandl et al., 2020]

Même si cette méthode a l'avantage de proposer des explications claires, puisqu'adoptant le même schéma que l'instance à expliquer, son désavantage est surtout lié au fait qu'il est possible de proposer plusieurs explications contrefactuelles pour expliquer une même prédiction. Nous nous ramenons alors au même problème indiqué dans la Section 2.3.3 avec l'effet Rashomon. Dès lors, quelle(s) contrefactuelle(s) choisir ?

LIME

LIME est une méthode d'explication locale décrite dans [Ribeiro et al., 2016]. *LIME* utilise des modèles explicables afin d'approximer localement un modèle boîte noire complexe et, pour chaque instance, explique l'influence de chaque variable sur la prédiction. Pour chaque instance à expliquer, *LIME* génère de nouvelles données dans un voisinage proche et calcule les prédictions de ces nouvelles instances à l'aide du modèle boîte noire. Un modèle de régression linéaire, modèle intrinsèquement interprétable, est entraîné sur le nouveau jeu de données. Ce modèle local (vu comme un modèle de substitution) est alors utilisé pour expliquer la prédiction de l'instance visée, sous la forme d'un vecteur de poids associant à chaque variable le score d'influence sur la prédiction. Une limitation bien connue de *LIME* est son hypothèse restrictive portant sur la linéarité locale (cela rend donc toute analyse, généralisée à plusieurs points, plus délicate) ainsi que sur l'indépendance des variables [Slack et al., 2020, Garreau and von Luxburg, 2020]. Définir une localité autour de l'instance à expliquer a possiblement tout d'une gageure, puisqu'il dépend de la bonne adéquation du modèle de substitution : cela peut avoir un impact significatif sur la précision des explications produites.

Le code complet de *LIME* est disponible sur GitHub [13].

Anchor

La méthode *Anchor* prend pour base le principe issu de *LIME* (les deux méthodes sont proposées par les mêmes auteurs [Ribeiro et al., 2018]) mais en cherchant à générer, cette fois-ci, des règles de décisions. L'idée est de trouver la règle de décision qui permette de délimiter une explication (*scoped rule*) cohérente entre l'instance visée et son voisinage. Comme avec *LIME*, *Anchor* génère des perturbations autour de l'instance à expliquer afin de produire une règle de type "IF..THEN.." (au contraire de LIME qui génère un modèle local de substitution).

13. https://github.com/marcotcr/lime

Les règles possèdent donc une partie conditionnelle, faite d'un sous-ensemble des variables et de valeurs associées, et une partie conséquent incluant tout simplement la classe prédite. Des scores de précision et de couverture sont aussi associés à la règle produite : - la précision indique, parmi le voisinage de l'instance expliquée, combien d'entre elles suivent bien la règle (et ne change pas de prédiction, donc) - la couverture indique le nombre d'instances total du jeu de données qui suivent la règle.

Même si cette méthode reste rapide à exécuter et claire dans les explications produites, *Anchor* souffre des mêmes problèmes que *LIME* et notamment sur la génération des perturbations autour de l'instance à expliquer, particulièrement difficile à configurer.

Shapley values

Pour expliquer des prédictions individuellement, une méthode basée sur les valeurs de Shapley est décrite dans [Štrumbelj and Kononenko, 2008, Strumbelj and Kononenko, 2010, Štrumbelj and Kononenko, 2014]. Les valeurs de Shapley permettent de pondérer équitablement les groupes de variables selon leur importance relative pour un gain donné [Shapley et al., 1953], au contraire d'une méthode comme *LIME* qui ne garantit pas que le gain soit réparti équitablement entre les variables. Dans le domaine de l'apprentissage automatique, le gain peut être lié à la prédiction faite par le modèle. L'influence de chaque variable est calculée selon l'impact de la prédiction pour chaque coalition de variables. Toutes les coalitions sont évaluées avec et sans la présence d'une variable donnée et les variations constatées, sur la prédiction à expliquer, sont utilisées pour le calcul de l'influence de cette variable. Cette méthode est bien évidemment très coûteuse en temps de calcul, avec une complexité exponentielle liée au nombre de variables dans le jeu de données. Une méthode plus récente, comme *SHAP*, est basée sur ce principe des valeurs de Shapley, avec pour but de résoudre (en partie) ce problème de complexité.

SHAP

SHAP (SHapley Additive exPlanations) [Lundberg and Lee, 2017] combine les méthodes *LIME* [Ribeiro et al., 2016] and Shapley values [Štrumbelj and Kononenko, 2014], ainsi qu'un certain nombre d'autres méthodes de la littérature [Lipovetsky and Conklin, 2001, Bach et al., 2015, Datta et al., 2016, Shrikumar et al., 2017], en un cadre unique pour produire des explications locales. La principale idée est de créer des perturbations pour simuler l'absence d'une variable et d'utiliser un modèle local linéaire afin d'approximer le changement dans la prédiction, comme pour *LIME*. Cela évite d'avoir à ré-entrainer un modèle complexe sans la variable visée. Ces explications locales peuvent être agrégées afin d'expliquer le comportement global du modèle. Les explications locales et globales sont ainsi cohérentes entre elles dans le sens où elles sont fondées sur une même base. La méthode *SHAP* est disponible en plusieurs versions : une version agnostique, *KernelSHAP* ainsi que des versions spécifiques à un modèle, comme *TreeSHAP*, *LinearSHAP* et *DeepSHAP* pour des modèles basés arbre, linéaire et profond, respectivement. *SHAP* est certainement la méthode d'explication locale

	Compréhension des explications	Analyse entre variables	Temps de calcul	Considération de l'indépendance entre variables	Précision des explications attributives
ICE	Très facile	Impossible	Faible	N/A	N/A
Contrefactuelles	Modérée	Difficile	Très Elevé	N/A	N/A
LIME	Facile	Possible	Faible	Oui	Faible
Anchor	Facile	Possible	Faible	Oui	Faible
Shapley values	Facile	Possible	Très Elevé	Oui	Importante
SHAP	Facile	Possible	Elevé	Oui	Modérée

TABLE 2.1 – Résumé des avantages et inconvénients des méthodes XAI locales post-hoc agnostiques

la plus utilisée de nos jours dans le domaine de l'apprentissage automatique. Cependant, la méthode n'est pas exempte de tout défaut et souffre toujours d'un manque de précision [Slack et al., 2020, Kumar et al., 2020], principalement due à des hypothèses restrictives (linéarité locale et indépendance entre variables) même si cela est moins prégnant qu'avec la méthode *LIME*. De plus, son temps de calcul reste potentiellement conséquent, en dehors de *TreeSHAP* pour les modèles basés arbre [Van den Broeck et al., 2021].

Le code complet de *SHAP* est disponible sur GitHub [14].

Quelle explication considérer ?

Au vu de toutes les méthodes, locales et agnostiques au modèle exposées dans cette section nous faisons le choix dans ce mémoire de nous focaliser particulièrement sur les trois méthodes attributives de la littérature, à savoir *LIME*, *Valeurs de Shapley* et *SHAP*. Une première justification porte sur leur très grande popularité (sans doute plus encore pour *SHAP*) [Linardatos et al., 2021], applicable et utilisée dans de multiples domaines en particulier en finance [Futagami et al., 2021] et en médecine [Bibault et al., 2021]. Comme montré dans la Table 2.1, résumant les avantages et inconvénient des méthodes XAI évoquées, nous justifions l'usage des méthodes atributives par leur grande simplicité d'utilisation et d'analyse qui se résume à produire de simples vecteurs de poids par instance. Il est vrai que ces méthodes ne sont pas exemptes de tout défaut, notamment quant au problème de temps de calcul et de robustesse (que nous détaillerons dans la section suivante), mais nous considérons que c'est un moindre mal face aux avantages évoqués ci-dessus.

Nous faisons aussi une hypothèse claire sur l'enjeu du choix de ces méthodes d'explication : celui de considérer les explications attributives comme une nouvelle source de données à exploiter et analyser. Nous reviendrons sur ce point dans le Chapitre 6.

14. https://github.com/slundberg/shap

2.5.3 Evaluation des explications

En apprentissage automatique, l'évaluation des explications produites est encore aujourd'hui un enjeu majeur. Cela amène à comprendre et évaluer ce qu'est une bonne explication. Cette question, en apparence simple, est pourtant éminemment complexe à résoudre. [Miller, 2019] indique que l'évaluation des méthodes d'explication reste très subjective et aucun consensus n'existe encore pour la proposition de métriques pertinentes.

Dans le domaine des explications locales, [Robnik-Šikonja and Bohanec, 2018] définit des propriétés comme la précision, la fidélité, la représentativité, la compréhensibilité ou encore la cohérence. Les explications qui respectent ces propriétés peuvent être vues comme de bonnes explications, car considérées comme fidèles au modèle ou aux données, dignes de confiance et faciles à comprendre pour les utilisateurs finaux. Toutefois, ces propriétés sont essentiellement subjectives, car elles ne sont pas définies mathématiquement et la manière de les mesurer n'est pas évidente. La définition d'une bonne explication peut également varier en fonction de l'utilisateur final, du domaine d'application, des objectifs de l'utilisation des explications, ce qui rend difficile l'évaluation objective de la qualité des explications, comme le souligne [Miller, 2019].

Dans le cadre des explications locales attributives, un point plus particulier porte sur la notion de monotonicité et complexité des explications [Nguyen and Martínez, 2020]. La monotonicité est particulièrement intéressante, car elle évalue la relation entre les valeurs d'une explication et ses attentes. La complexité effective est liée à la concision et évalue le nombre minimum de variables nécessaires à l'explication. La robustesse est une autre mesure fréquemment mentionnée dans la littérature, définie comme la capacité de l'explication à être similaire lorsque les données d'entrée sont similaires. Malheureusement, il existe plusieurs formulations mathématiques de cette métrique, en fonction de la manière dont les auteurs déterminent ce que signifie la similarité et comment la calculer [Alvarez-Melis and Jaakkola, 2018].

Ces multiples implémentations produisent des mesures qui permettent d'évaluer chaque méthode d'explication de manière spécifique. Cependant, comparer des mesures qui n'ont pas été calculées de la même manière ou qui sont mathématiquement incohérentes entre elles, rend encore plus difficile la rechercher d'une qualité universelle des explications. L'une des difficultés réside donc dans la capacité à produire des mesures applicables à toutes les méthodes d'explication.

La problématique de l'évaluation des explications, de ses multiples propriétés désirables possibles et de leurs multiples implémentations, montre finalement bien les mêmes limites que celles exposées pour l'apprentissage automatique (voir Section 2.3.3) : comment choisir la bonne méthode d'explication ? Comment bien hyperparamétrer ces méthodes ? Comment bien les évaluer ? Ces questions restent aujourd'hui récurrentes et non résolues. Une des voies alors possibles serait, à l'image de ce qui a été proposé pour la recommandation de modèle prédictif via des techniques comme l'AutoML, de recommander automatiquement la méthode d'explication la plus adaptée aux données à analyser et aux préférences utilisateurs. C'est ce que nous détaillerons dans le Chapitre 4 et pour lequel nous préciserons certains manquements à la littérature, notamment les problèmes de lisibilité des explications, comme évoqués dans

le Chapitre 5.

2.6 Conclusion

Dans ce chapitre, nous avons considérés les trois domaines de recherche abordés dans ce mémoire, à savoir : l'apprentissage automatique, la recommandation de modèle prédictif et l'explicabilité pour l'apprentissage automatique.

L'apprentissage automatique, et notamment l'apprentissage supervisé, objet de notre étude, montre clairement ses limites et ses difficultés à une adoption facilitée pour un utilisateur. Les quatre limites évoquées portant sur l'*effet Rashomon*, l'*effet de dilution*, l'*hyperparamétrage* des algorithmes et l'effet *boîte noire* doivent pouvoir être traitées en offrant les moyens techniques à leur résolution, sous peine d'une moindre confiance de l'utilisateur.

A cette fin, la recommandation de modèles prédictifs peut être un premier pas pour combler ces manques. La recommandation par méta-apprentissage est une voie prometteuse car permettant de sélectionner et d'hyperparamétrer automatiquement le modèle prédictif le plus adapté à un jeu de données à analyser, en fonction de jeux de données passés similaires (et de leurs workflows correspondants). Nous avons pu y noter, en particulier, un manque de prise en compte du contexte utilisateur, notamment par ses préférences sur les performances du modèle. La considération de ce contexte devrait ainsi permettre d'améliorer davantage les recommandations proposées.

L'effet *boîte noire* trouve une réponse via les méthodes d'explicabilité proposées dans la littérature. Nous avons pu nous rendre compte à la fois du nombre important des propositions faites, y compris dans le domaine des méthodes d'explication locale post-hoc, objets également de ce mémoire. Cependant, les mêmes types de limites sont également constatés pour l'apprentissage automatique. Ces limites étant accentuées par le manque de consensus sur la manière d'évaluer les explications, même si quelques propositions s'y attaquent, notamment en préconisant des propriétés désirables à la génération de bonnes explications. Nous avons également pu constater que les organisations peinaient, en général, à s'approprier l'usage de l'apprentissage automatique et les méthodes de XAI associées. Cela est notamment dû au fait qu'il n'existe pas de processus ni de règles claires et génériques, applicables en toutes circonstances. Il y a sans doute aussi un a priori sur une certaine technicité de l'apprentissage automatique, alors que l'usager et son environnement dans un Système d'Information sont essentiels à prendre en compte pour une meilleure diffusion de son usage.

Dans l'objectif de fournir à un utilisateur les moyens d'une analyse prédictive intelligible, nous proposons les solutions suivantes afin de répondre aux limites évoquées :

— Le Chapitre 3 s'intéresse à la recommandation de modèles prédictifs, par une méthode de méta-apprentissage, en considérant les multiples préférences utilisateur afin de maximiser les performances souhaitées pour le modèle prédictif.

— Le Chapitre 4 offre un cadre pour une recommandation de modèles d'explication, en fonction d'un jeu de données et de préférences utilisateur. Le cadre reprend, pour partie, l'idée de l'AutoML et sa capacité à hyperparamétrer automatiquement un al-

gorithme d'apprentissage. Cette méthode se base, en particulier sur les propriétés et métriques de la littérature dans le domaine de l'explicabilité. Elle peut être considérée, par ailleurs, comme suffisamment générique pour être applicable à n'importe quelle méthode d'explication post-hoc, qu'elle soit globale ou locale.

— Le Chapitre 5 détaille les approches d'explications locales attributives en proposant une nouvelle solution d'explication palliant le manque de prise en compte des dépendances entre variables ainsi qu'une étude quantitative comparant l'ensemble de ces approches. Nous montrerons que les seules propriétés et métriques de la littérature ne suffisent pas à montrer les différences entre méthodes attributives et que chaque méthode a ses propres avantages et inconvénients dans l'identification de relation entre les données, via un modèle prédictif.

— Le Chapitre 6 considère les explications produites par les méthodes locales attributives comme un nouvel espace de données. Ce nouveau paradigme nous permet d'imaginer des pistes très variées :

  ➢ une démarche itérative (human-in-the-loop), complémentant la proposition faite dans le Chapitre 3, afin d'affiner et personnaliser un modèle prédictif à l'aide d'explications locales.

  ➢ une aide à la sélection de variables considérant les explications comme une nouvelle dimension au choix de variables à considérer. Nous montrerons d'ailleurs une sélection de variables, quelle qu'elle soit, a une influence sur les profils d'explication, pour une même méthode de XAI.

  ➢ une aide à la sélection d'instances en démontrant que l'usage d'explications locales permet de mieux identifier les instances importantes d'un jeu de données, notamment à l'aide de techniques de clustering.

  ➢ une aide à l'analyse de données, plus généralement, en décrivant l'intérêt de l'usage combiné d'explications et de règles de décisions (expliquant les explications !) pour améliorer les conclusions possibles à retirer d'un jeu de données. Nous montrerons également l'intérêt d'une démarche hiérarchique de l'analyse des explications afin, là encore, de mieux éliciter les relations dans les données.

Chapitre 3

Un cadre pour l'aide à la sélection automatique de modèles prédictifs

3.1 Introduction

Dans ce chapitre, nous nous intéressons à la recommandation de modèles prédictifs, par l'intermédiaire des workflows d'analyse. Plus précisément, il s'agit de proposer une solution considérant les multiples préférences de l'utilisateur (les performances à maximiser dans l'analyse prédictive de son jeu de données) pour la recommandation de workflows, via une approche AutoML [Feurer et al., 2015a, He et al., 2021]. Cela implique de s'intéresser aux verrous suivants :

— comment concilier les préférences utilisateur entre elles ?
— comment sélectionner les workflows les plus pertinents au sens des préférences utilisateurs ?
— comment rendre applicable l'approche AutoML, et notamment la recherche de jeux de données similaires dans des bases d'analyses passées ?

Les travaux présentés dans ce chapitre font référence aux publications suivantes [Raynaut et al., 2017a, Ferrettini et al., 2020c] et à une partie du travail de thèse de Gabriel Ferrettini [Ferrettini, 2021].

3.2 Les préférences de l'utilisateur dans la recommandation de modèles

Comme discuté dans le Chapitre 3, il nous semble important de prendre en compte le contexte utilisateur dans la recommandation de workflows d'apprentissage, par une approche AutoML. Il s'agit de considérer un ensemble de préférences, souhaitées par l'utilisateur, prenant la forme de critères de performance que le modèle prédictif recommandé doit maximiser. D'autre part, l'AutoML étant basé sur la recherche du jeu de données le plus similaire au

jeu de données à analyser, il est important de considérer aussi ce critère dans le contexte de l'utilisateur. Dès lors, comment concilier recherche du jeu de données le plus similaire et maximisation des performances souhaitées par l'utilisateur ? C'est ce que nous détaillons dans les sous-sections suivantes.

3.2.1 Modélisation des préférences utilisateur

Dans l'apprentissage automatique, les critères de performance sont donc potentiellement multiples. Par exemple, dans le cas d'une tâche de prédiction sensible (comme le diagnostic précoce d'une maladie dangereuse), l'utilisateur a tout intérêt à être vigilant sur le taux de faux négatif. Il serait donc pertinent de sélectionner les workflows ayant montré un bon *rappel* sur des jeux de données similaires passés. Notre approche consiste alors à considérer des critères capables de caractériser différents aspects de la performance des workflows, modélisant les préférences des utilisateurs. Même en ne considérant que les problèmes de classification supervisée, de nombreux critères ont été proposés pour caractériser les différents aspects de la performance, comme le *Kappa* de Cohen [Cohen, 1968], qui est une mesure d'accord par rapport au hasard, ou le *Score d'information* plus complexe de [Kononenko and Bratko, 1991], qui mesure la quantité d'informations *non triviales* produites par le modèle.

Nous pouvons donc représenter la modélisation des préférences de l'utilisateur comme un ensemble de critères de performance qui l'intéressent, chacun étant associé à un poids qualifiant son importance relative. Par exemple, l'utilisateur qui souhaite éviter les faux négatifs considère le *rappel* comme son critère le plus important, mais cela ne signifie pas qu'il ne se soucie pas de la *précision* ! Un poids plus élevé associé au *rappel* représentera alors cette préférence.

3.2.2 Un problème multicritères

En considérant un utilisateur, en possession d'un jeu de données et ayant défini ses préférences, l'objectif de notre système est de recommander des workflows à partir d'analyses passées via une approche AutoML. Cela implique l'accès à une base d'expériences d'analyse de données passées, dans laquelle les utilisateurs ont déjà préalablement effectué d'autres analyses. Une telle analyse passée consiste alors en un jeu de données, sur lequel a été appliqué un workflow, produisant un résultat (en l'occurrence lié à une tâche de prédiction). Comme nous l'avons vu précédemment, notre objectif est donc de trouver des analyses passées *pertinentes*, pour un utilisateur, en fonction de deux critères :

1. l'analyse passée doit avoir été produite sur un jeu de données similaire à celui de l'utilisateur actuel.
2. ses résultats, évalués selon les critères de performance donnés par l'utilisateur, doivent être satisfaisants de son point de vue.

Nous nous retrouvons donc face à un problème d'optimisation multicritères, où à la fois la similarité avec un jeu de données et ses performances passées sont à prendre en compte.

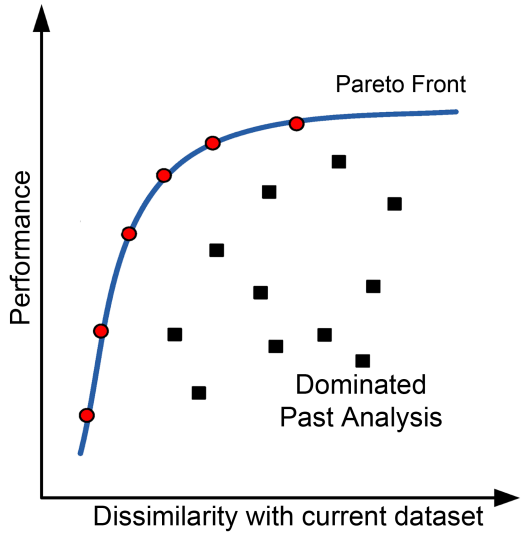


FIGURE 3.1 – Front de Pareto des meilleures analyses passées selon nos deux critères.

En effet, le jeu de données le plus similaire ne garantit pas forcément de maximiser tous les critères de performance souhaités alors que peut-être un autre jeu de données un peu moins similaire le garantirait. Pour résoudre cela, deux approches sont considérées :

1. **Recommandation moyenne** : La première approche consiste à simplement agréger par une moyenne les deux critères, avec la même importance. Cette approche, formalisée dans l'algorithme 1, produit une recommandation unique qui est la plus susceptible de convenir à l'utilisateur sans nécessiter de calcul supplémentaire. Elle perd toutefois la plupart des informations contenues dans les deux critères initiaux.

2. **Recommandation par Pareto** : La deuxième approche consiste à explorer l'ensemble des solutions optimales de Pareto. Nous considérerons alors le front de Pareto complet des analyses passées, non dominées, comme un ensemble de recommandations. L'examen des meilleurs candidats selon nos deux critères (comme le montre la Figure 3.1) augmente les chances de trouver celui qui convient le mieux à l'utilisateur, mais nécessite une étape supplémentaire pour faire la distinction entre les candidats. En effet, la fourniture de l'ensemble des recommandations serait probablement utile aux utilisateurs experts, mais risquerait fort de submerger un non-expert. Des tests complémentaires sont donc indispensables pour sélectionner le workflow le plus adapté parmi les workflows candidats. Cette approche est formalisée dans l'algorithme 2.

Algorithm 1 Recommendation by criteria average.

Require: $\mathcal{B} = (D_i, W_i, R_i)_{i \in [1..n]}$ The base of past analysis, each consisting in a workflow W_i applied on a dataset D_i yielding a result R_i
$\mathcal{D}$ The current user's dataset
$\mathcal{P}$ The current user's preferences

for all analysis $(D_i, W_i, R_i) \in \mathcal{B}$ **do**
Compute and normalize $d(\mathcal{D}, D_i)$
Compute and normalize $\mathcal{Q}_{\mathcal{P}}(R_i)$
end for
Find i maximizing $d(\mathcal{D}, D_i) + \mathcal{Q}_{\mathcal{P}}(R_i)$
return W_i

Algorithm 2 Recommendation from Pareto front.

Require: $\mathcal{B} = (D_i, W_i, R_i)_{i \in [1..n]}$ The base of past analysis
$\mathcal{D}$ The current user's dataset
$\mathcal{P}$ The current user's preferences

for all analysis $(D_i, W_i, R_i) \in \mathcal{B}$ **do**
Compute $d(\mathcal{D}, D_i)$
Compute $\mathcal{Q}_{\mathcal{P}}(R_i)$
end for
Compute the Pareto front $F \in \mathcal{B}$ of non-dominated analysis (analyses where neither $d(\mathcal{D}, D_i)$ nor $\mathcal{Q}_{\mathcal{P}}(R_i)$ can be improved in value without degrading the other)

for all non-dominated analysis $(D_j, W_j, R_j) \in F$ **do**
Realize experiment $(\mathcal{D}, W_j, \mathcal{R}_j)$
Compute $\mathcal{Q}_{\mathcal{P}}(\mathcal{R}_j)$
end for
Find j maximizing $\mathcal{Q}_{\mathcal{P}}(\mathcal{R}_j)$ **return** W_j

3.3 Evaluation

3.3.1 Recommandation de workflow par dissimilarité

Le système de recommandation proposé reprenant le concept de l'AutoML (voir Chapitre 2.4), il est nécessaire de disposer d'une mesure de dissimilarité comme discuté également dans le Chapitre 2.4.2. Nous nous proposons alors d'utiliser la mesure de dissimilarité proposé par [Raynaut, 2018]. En effet, celle-ci a démontré qu'elle surpassait d'autres propositions de

la littérature. Cette mesure est constituée de deux dissimilarités, sur deux niveaux : l'une calculant des dissimilarités entre meta-features de variables, l'autre entre meta-features de jeux de données. Nous détaillons ci-après les moyens pour collecter les meta-feature sur ces deux niveaux.

Meta-feature d'un jeu de données

Afin de disposer d'une large sélection de meta-features issus de diverses catégories, nous avons choisi d'utiliser ceux de la base OpenML [Vanschoren et al., 2013], qui comporte plus d'une centaine de meta-features issus de différentes approches statistiques, de la théorie de l'information et du landmarking (liste complète disponible sur : http://www.openmlorg/).

Meta-feature d'une variable

Les variables des jeux de données peuvent être caractérisées à l'aide d'un ensemble de mesures, consistant principalement en des versions non agrégées des meta-features pour jeux de données, décrites précédemment. Pour construire notre ensemble de meta-features, nous utilisons les 72 mesures proposées dans [Raynaut et al., 2017b], caractérisant les variables, une à une. Pour rappel, les variables de deux jeux de données sont comparées entre paires les plus similaires : pour deux jeux de données A et B, chaque variable de A est associée à une variable de B de telle sorte que la dissimilarité totale de chaque paire soit aussi faible que possible.

3.3.2 Protocole d'évaluation du système de recommandation

Cette section décrit les expériences menées pour évaluer notre système de recommandation. L'objectif principal est de montrer comment la prise en compte du contexte utilisateur peut permettre de recommander des workflows très similaires à ce qu'un expert aurait fait. Nos évaluations reposent sur un schéma de validation croisée ([Stone, 1974]) sur un nombre fixe d'analyses passées, divisant itérativement cette base en ensembles d'entraînement et de test. Notre système de recommandation utilise ensuite les analyses des ensembles d'entraînement pour recommander des workflows sur l'ensemble de test. Ceux-ci peuvent ensuite être comparés aux workflows conçus par les utilisateurs dans le cadre de ces analyses.

Pour effectuer cette comparaison, nous avons besoin de définir une dissimilarité empirique entre les workflows selon l'intuition suivante : *Deux workflows sont similaires s'ils présentent des performances similaires sur des jeux de données similaires.* Supposons que nous souhaitions évaluer la dissimilarité entre les workflows W_A et W_B. Nous disposons de deux ensembles d'analyses $\mathcal{B}_A = (D_i, W_A, R_i)_{i\in[1..n]}$ et $\mathcal{B}_B = (D_j, W_B, R_j)_{j\in[1..m]}$, utilisant respectivement W_A et W_B. En considérant une dissimilarité d entre les jeux de données D_i, D_j et une dissimilarité d' entre les résultats de performances R_i, R_j, nous pouvons construire les séquences $d(D_i, D_j)$ et $d'(R_i, R_j)$. La *corrélation* entre ces séquences exprime alors notre intuition de dissimilarité des workflows : des workflows similaires ont tendance à donner des résultats plus semblables sur des jeux de données similaires que sur des jeux de données très différents. Sans aucune

hypothèse sur la distribution de ces dissimilarités, nous pouvons calculer cette corrélation en utilisant la corrélation de Spearman. Nous pouvons alors définir notre dissimilarité de manière à ce qu'elle soit nulle sur des worfklows très similaires (ρ de Spearman de 1) et qu'elle diverge à l'infini sur des workflows se comportant de manière exactement opposée (ρ de Spearman de -1). Cette dissimilarité entre les workflows peut alors être exprimée comme suit :

$$\Delta(W_A, W_B) = log_2 \frac{2}{1 + \rho(\ d(D_i, D_j),\ d'(R_i, R_j)\ )}$$

3.3.3 Base d'analyses passées

Nous utiliserons comme base d'analyses passées la plateforme OpenML [Vanschoren et al., 2013], mettant à disposition des millions d'analyses faites par apprentissage automatique. Cependant, parmi toutes ces analyses, un certain nombre d'entre elles ont été générée artificiellement (pour des besoins de recherche par exemple) et constitue une bonne partie de la base OpenML. Comme l'objectif est d'évaluer la capacité de notre système de recommandation à proposer des workflows pertinents, il nous parait donc nécessaire de nous focaliser sur les analyses produites réellement par des utilisateurs. OpenML n'indiquant pas explicitement la différence entre une analyse artificielle ou manuelle, nous décidons de sélectionner les analyses selon les conditions suivantes :

- seulement 5 analyses ont été effectuées, au plus, pour un même jeu de données ;
- le jeu de données doit être disponible publiquement (afin de calculer les dissimilarités) ;
- les critères de performance préférés par l'utilisateur doivent être disponibles ;
- OpenML doit contenir au moins 50 autres analyses utilisant le même workflow afin de pouvoir calculer les dissimilarités entre workflow ;
- Chaque analyse doit porter sur un jeu de données différent afin d'éviter que le nouveau jeu de données analysé par l'utilisateur ne se retrouve pas dans les analyses passées.

Ainsi, 324 analyses ont été sélectionnées selon ces conditions. Pour aller plus loin, nous avons décidé de subdiviser ces analyses en plusieurs sous-ensembles, afin de mesurer notre système de recommandation selon différentes caractéristiques (résumés dans le tableau 3.1) :

- Jeux de données : nous souhaitons étudier l'impact de la diversité des jeux de données sur les recommandations. L'idée est de construire deux sous-ensembles, l'un avec des jeux de données très similaires entre eux, et un autre très dissimilaires entre eux. Pour cela, on choisit un jeu de données considéré comme "central". A l'aide de la dissimilarité entre jeux de données, on construit donc les deux sous-ensembles en fonction de leur proximité avec ce jeu de données "central". Ce jeu de données "central" est choisi de sorte à maximiser la différence de dissimilarité moyenne des jeux de données, entre les deux sous-ensembles.
- Workflow : nous étudions comment la diversité des workflows peut affecter les recommandations. Le processus est très similaire au précédent : nous remplaçons simplement la dissimilarité entre jeux de données par celle entre workflows. Il en résulte un sous-ensemble d'expériences avec des worflows très similaires et un autre avec des workflows plus différents.

— Résultat de performance : cette fois-ci nous utilisons la dissimilarité entre résultats de performances, à l'aide d'une distance de Manhattan normalisée sur les critères de performance disponibles ; ceci afin d'obtenir nos deux sous-ensembles.

Ainsi, comme indiqué dans le tableau 3.1, nous obtenons l'*Ensemble le plus proche* et l'*Ensemble le plus éloigné* composés d'analyses liées soit à des jeux de données, des workflows ou des résultats de performances les plus similaires et dissimilaires possibles, respectivement. Enfin, l'*Ensemble complet* réunit les 324 analyses disponibles.

TABLE 3.1 – Dissimilarités moyennes, entre les différents sous-ensembles, selon différentes caractéristiques

	Jeu de données	Workflow	Résultat de performance
Ensemble le plus proche	0.122	0.429	0.242
Ensemble complet	0.153	0.841	0.266
Ensemble le plus éloigné	0.170	1.252	0.300

3.3.4 Référentiel de comparaison

Nous comparons notre système à deux processus de recommandation simple, nous servant ainsi de référence :

1. **Random** : Recommande de manière aléatoire un workflow issu de l'une des analyses passées. Cette stratégie nous permettra d'évaluer grossièrement les performances d'un tel système, réalisables sans utiliser aucune des informations disponibles, notamment les meta-feature. En raison du caractère manifestement aléatoire de cette approche, la moyenne des résultats a été calculée sur dix répétitions de l'ensemble de l'évaluation.
2. **BestPerformance** : Recherche dans la base des analyses passées, celle donnant les meilleures performances, en fonction des préférences de l'utilisateur, et recommande le workflow associé. Par exemple, un utilisateur qui ne demande qu'une grande précision se voit proposer le workflow qui a obtenu la plus grande précision dans la base. L'objectif de cette base est d'évaluer ce système en n'utilisant que les informations liées aux scores de performances.

3.3.5 Résultats

Les résultats sont disponibles dans les Figures 3.2 et 3.3. Ces résultats donnent les scores de précision pour des seuils de dissimilarité compris entre 0.0 et 1.6 (la mesure de dissimilarité utilisée étant définie entre 0.0 et $+\infty$.). Plus particulièrement, nous remarquons qu'un seuil supérieur à 1 permet de considérer des workflows similaires avec des scores de précision convergeant vers 1.0 pour chacune des méthodes. Ce phénomène est parfaitement attendu : notre évaluation étant basée sur le principe de la validation croisée, toutes les recommandations seront donc considérées comme similaires aux workflows de tests pour ce seuil.

La Figure 3.2 décrit les résultats obtenus par validation croisée, pour tous les jeux de données. On y constate que le système de *Recommandation par Pareto* donne les meilleurs résultats pour toute valeur de seuil, même dans le cas d'un seuil faible. Par exemple, si l'on considère un seuil de 0.6, le score de précision est d'environ 0.75, alors que toutes les autres approches sont inférieures à 0.52. En particulier, la recommandation de Pareto surpasse continuellement la recommandation basée sur la stratégie $BestPerformance$, sélectionnant le workflow uniquement sur le critère de performance. Ce résultat indique que la prise en compte du contexte utilisateur pour filtrer les workflows pertinents est efficace. Cependant, lorsque l'on analyse la recommandation basée sur la stratégie de moyenne des critères (au lieu du front de Pareto), la performance apparaît beaucoup plus faible. Ce résultat incite donc à la prudence lors de l'utilisation d'informations contextuelles : elles sont utiles si chaque information du contexte est prise en compte séparément dans la sélection du workflow. Dans le cas contraire, la dilution de ces informations par une agrégation montre un intérêt plus que limité (la stratégie par *Random* pouvant faire mieux dans la plupart des cas.).

La Figure 3.3 affine les résultats précédents en divisant la base des expériences passées en sous-ensembles, en fonction de la diversité des jeux de données, des workflows et des résultats de performance (comme indiqué dans 3.3.3). En fonction de la diversité des jeux de données, nous pouvons constater que *ParetoRecommendation* surpasse toujours les autres méthodes, que les sous-ensembles de jeux de données soient très similaires ou très différents entre eux. Si l'on compare ses résultats à ceux de la Figure 3.2, la *Recommandation par Pareto* ne semble alors pas très sensible à la diversité des jeux de données. Ce résultat indique que notre système peut recommander des workflows pertinents dans un contexte où des jeux de données passés ne seraient pas totalement similaires au jeu de données à analyser.

En séparant par diversité de workflows, nous remarquons que les résultats sur l'ensemble proche atteignent un maximum très rapidement. Ce résultat est attendu puisque la probabilité de tirer un workflow similaire est plus grande lorsque la plupart sont similaires les uns aux autres (c'est pourquoi l'approche *Random* est plus performante que les autres avec des seuils plus bas que dans les autres cas). Si l'on considère l'ensemble le plus éloigné, la *Recommandation par Pareto* surpasse à nouveau les autres méthodes, mais avec un score de précision généralement plus faible (avec un seuil de 0.6, le score de précision est d'environ 0.59, alors que nous avons atteint 0.75 pour tous les jeux de données). Ce résultat est également attendu, car il est plus difficile de fournir des workflows similaires lorsque la plupart sont très différents, mais l'observation d'une perte de performance limitée (par rapport aux autres approches) est encourageante.

En analysant par diversité des résultats de performance, nous pouvons remarquer que les scores de précision liés à l'ensemble le plus proche semblent moins bons qu'avec l'ensemble le plus éloigné. Par exemple, pour un seuil de 0.6, la *Recommandation par Pareto* obtient un score de 0.64 en utilisant l'ensemble le plus proche, et de 0.82 en utilisant l'ensemble le plus éloigné. Cela peut s'expliquer par le fait que la diversité des résultats semble être une propriété importante pour la recommandation, car elle augmente les chances de trouver des expériences passées correspondant aux préférences de l'utilisateur. Nous devons cependant

noter que des résultats de performance similaires ne signifient pas nécessairement que des workflows similaires ont été utilisés. Par exemple, un score de rappel peut tout à fait être maximisé par deux algorithmes radicalement différents.

Au vu de tous ces résultats, nous pouvons remarquer que la *Recommandation par Pareto* semble toujours surpasser les autres méthodes, garantissant une robustesse relative sur la plupart des diversités possibles liées à une analyse de données prédictive. Cela augure bien de ses résultats sur des bases d'analyses passées potentiellement beaucoup plus larges, même lorsque les conditions d'exploitation limitent sa diversité.

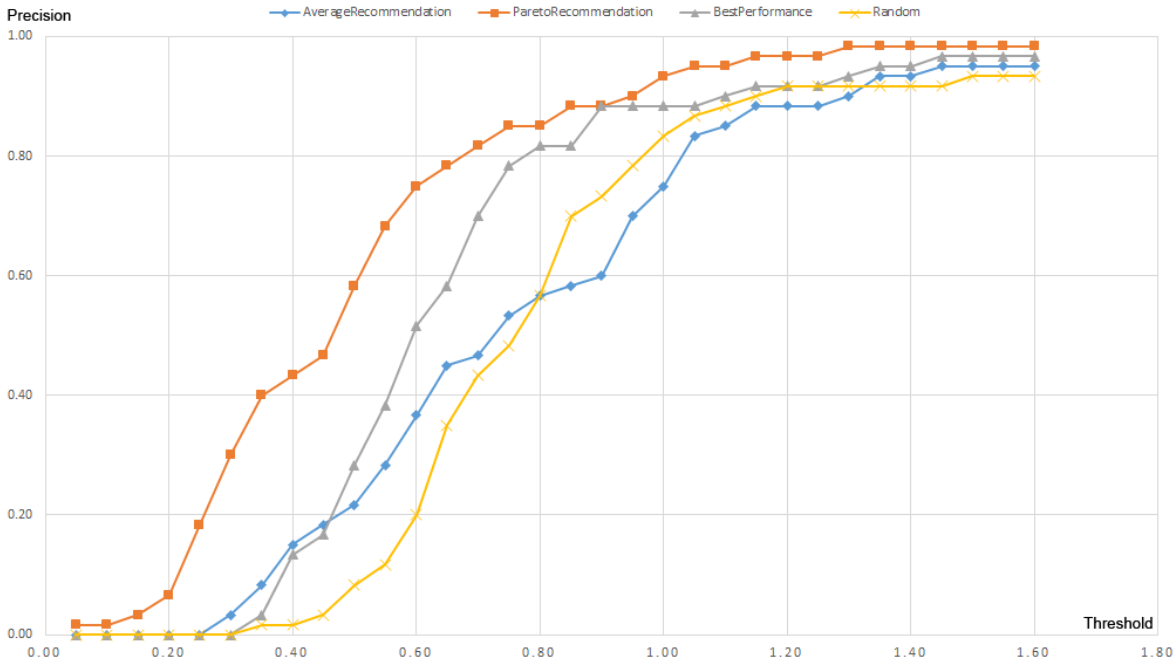


FIGURE 3.2 – Scores de précision par seuil de similarité, sur l'ensemble complet.

3.4 Conclusion

3.4.1 Bilan

Les travaux présentés dans ce chapitre ont mené à trois propositions principales :

— la prise en compte du contexte utilisateur incluant le jeu de données à analyser ainsi que ses préférences sur les performances à atteindre pour un modèle prédictif,

— la proposition de deux versions du système de recommandation prenant en compte la problématique multi-critère liée au contexte utilisateur,

— l'intégration du système de recommandation dans une approche AutoML.

La prise en compte du contexte utilisateur nous a amenés à nous pencher à la fois sur la notion de dissimilarité entre jeux de données et la notion de préférences multiples. La mesure de dissimilarité entre jeux de données a été choisie de sorte à couvrir la plupart des caractéristiques possible (méta-données) disponibles au niveau du jeu de données en lui-même ainsi qu'au niveau de ses variables. La modélisation des préférences utilisateur nous a menés à les représenter comme un ensemble de critères de performance où chaque critère est pondéré selon l'importance souhaitée par l'utilisateur.

Les deux versions du système de recommandation répondent à la problématique évoquée au début de ce chapitre concernant le besoin de concilier la recherche de jeux de données similaires et la recherche des meilleures performances. La première version considère une moyenne de ces deux critères et la seconde version prend en compte l'ensemble du front de Pareto pour ces critères.

Les deux systèmes sont pensés dans une intégration AutoML où la proposition d'un modèle prédictif se fait en explorant des bases d'analyses passées. Pour cela, nous avons considéré la base OpenML, car elle répondant à la fois aux besoins d'identifier des meta-données sur les analyses passées et disposaient d'une quantité suffisante d'analyses passées pour pouvoir tester et mettre en pratique nos propositions.

Les expériences ont montré que notre système basé sur le front de Pareto surpasse, en termes de précision, toutes les autres approches, même dans le cas d'un seuil bas (où la dissimilarité est la plus exigeante). Notre approche est également efficace pour gérer des jeux de données passés assez peu similaires au jeu de données à analyser.

3.4.2 Perspectives

Les perspectives possibles aux travaux évoqués dans ce chapitre touchent à des domaines de recherche variés.

Tout d'abord, nous pouvons évoquer le **besoin de prendre en compte le niveau d'expertise de l'utilisateur** dans les recommandations, à intégrer alors à son contexte. Comme souligné par [Knijnenburg et al., 2011], les besoins liés aux systèmes de recommandation diffèrent selon le niveau d'expertise de l'utilisateur. Un utilisateur novice sera ainsi satisfait d'un résultat par défaut, sans intégrer de préférence particulière, au contraire d'un utilisateur plus expert. Cela pose aussi la question de la manière d'interagir avec l'utilisateur afin de récolter ses préférences. Dans notre proposition, c'est à l'utilisateur lui-même de fixer ses pondérations entre les différents types de performances souhaitées pour un modèle prédictif. Cela peut être limitant, même pour un expert. Des solutions basées sur des priorités entre performances (plutôt que des scores), ou encore des suggestions automatiques de performances faciliteraient l'usage du système par l'utilisateur.

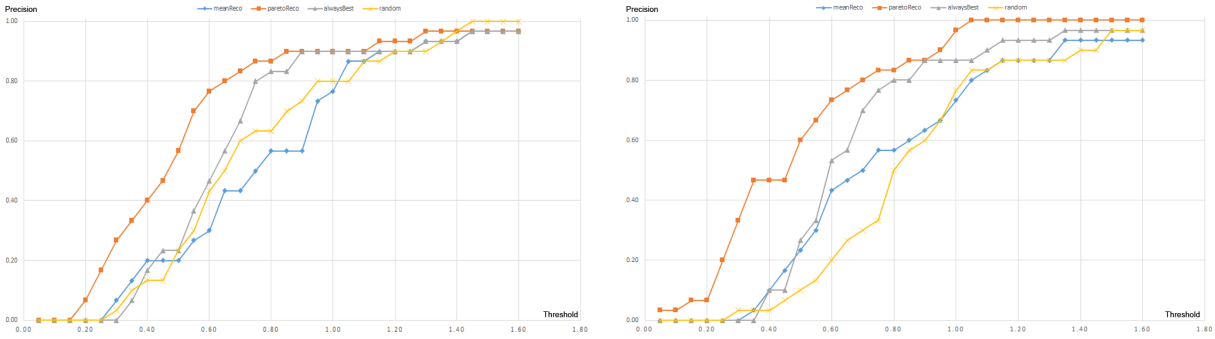
Une autre perspective concerne le fait que les workflows sélectionnés par notre système de recommandation sont repris tels quels depuis les analyses passées. Une **optimisation limitée des hyperparamètres** [**Feurer et al., 2015b**] **ou un réglage "intelligent par défaut"** [**Mantovani et al., 2015**] pourraient contribuer à accroître les performances générales avec un coût supplémentaire limité. De plus, en considérant des informations cette fois-ci sémantiques sur les workflows (c'est-à-dire toute information permettant de décrire chaque étape du workflow), il serait possible d'effectuer des recherches heuristiques plus avancées sur l'obtention de workflows pertinents. Cela permettrait aussi, sans doute, de générer des workflows mieux personnalisés au contexte de l'utilisateur ainsi qu'à son niveau d'expertise comme évoqué précédemment. Nous pourrions aussi envisager de combiner des sous-séquences de différents workflows permettant d'en générer un nouveau (par exemple des étapes de pré-traitement de l'un et un algorithme prédictif d'un autre), d'une manière similaire aux algorithmes génétiques. Un tel processus nécessiterait toutefois des connaissances sémantiques supplémentaires, par exemple sous la forme d'une ontologie complète d'analyse de données [Keet et al., 2015], pour laquelle aucune mise en oeuvre effectivement n'est encore disponible...

Comme évoqué dans le Chapitre 2.4.3, recommander des modèles, au travers des workflows proposés, est une chose, mais les faire comprendre à l'utilisateur est une condition nécessaire à une meilleure adoption de leurs usages. Le manque de confiance dans les recommandations de modèles prédictifs peut ainsi être pallié selon deux grandes voies disctinces : **en expliquant à l'aide du fonctionnement interne du système de recommandation** (comme proposé dans [Chanson et al., 2021, Zhong and Negre, 2022]), soit en **expliquant directement le résultat recommandé**, c'est-à-dire le modèle prédictif en lui-même. En ce qui nous concerne, cette dernière solution est sans doute préférable à considérer, dans un premier temps, au vu de la complexité que représente un modèle prédictif. Hors modèle intrinsèquement interprétable (voir Chapitre 2.5), un modèle prédictif est, de fait, difficile à appréhender pour un utilisateur. Lui offrir les moyens de le comprendre est donc la condition pour qu'il puisse se l'approprier. Cela est d'ailleurs tout l'enjeu des travaux présentés dans les chapitres suivants, en particulier via l'explication des prédictions fournies par le modèle.

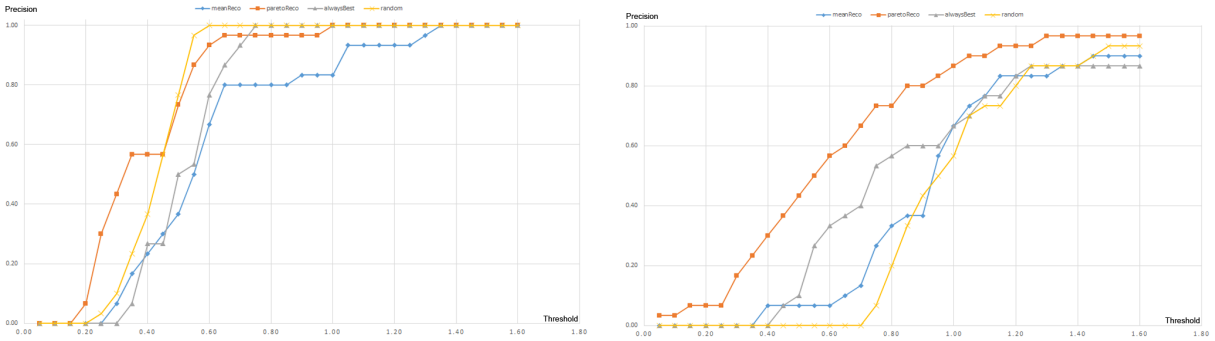
Les explications proposées vis-à-vis d'un modèle prédictif seraient très certainement utiles dans un cadre plus large d'interaction avec un système de recommandation. En effet, **ces explications pourraient être à la base d'un cycle itératif entre le système de recommandation et l'utilisateur** où ce dernier, à l'aide des explications, permettrait de produire des retours (feedback) au système afin d'en améliorer les résultats proposés. Cela pourrait être, par exemple, d'ajouter ou supprimer une variable donnée au vu d'une incohérence que l'utilisateur aurait pu détecter dans les explications liées au modèle prédictif. C'est d'ailleurs l'objet d'une des sections du Chapitre 6, pour une aide à la sélection de modèles prédictifs.

Ensemble le plus proche Ensemble le plus éloigné

Sous-ensembles de jeux de données



Sous-ensembles de workflow



Sous-ensemble de résultats de performance

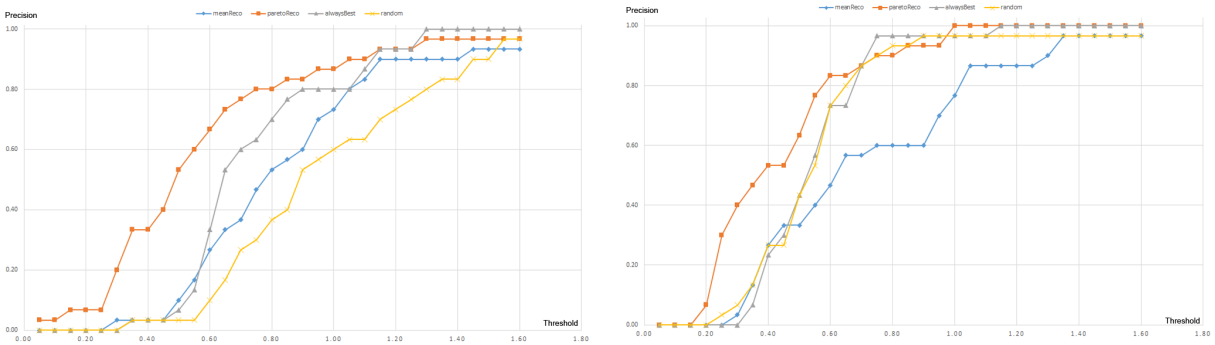


FIGURE 3.3 – Scores de précision par seuil de similarité, sur les différents sous-ensembles.

Chapitre 4

Un cadre pour la sélection automatique de modèles d'explications

4.1 Introduction

Dans ce chapitre, nous proposons de nous intéresser à la recommandation de modèles d'explication (voir état de l'art Chapitre 2.5). Comme discuté dans le Chapitre 2.5, le manque de compréhension dans les modèles d'apprentissage automatique produits retarde leurs adoptions dans des domaines à forts enjeux tels que le domaine médical [Markus et al., 2021], la sécurité numérique [Brown et al., 2018], le domaine judiciaire [Tan et al., 2018], ou la conduite autonome [Omeiza et al., 2021], par exemple.

Au vu du nombre de types d'explications désormais à disposition, il nous semble pertinent de pouvoir proposer un système de recommandation proposant le modèle d'explication le plus adapté à un contexte utilisateur (AutoXAI). L'idée est de reprendre la même philosophie de recommandation tenant compte du contexte, comme proposé dans le chapitre précédent, afin d'identifier les modèles XAI maximisant les propriétés que les utilisateurs souhaitent obtenir dans les explications (exactitude ou compacité des explications, par exemple). Afin d'optimiser les modèles XAI recommandés, nous proposons de nous inspirer des techniques AutoML pour la recherche automatique des hyperparamètres les plus appropriés, en particulier l'optimisation bayésienne [Zhou et al., 2021, Nauta et al., 2022].

Nous illustrons les recommandations faites par AutoXAI à l'aide de deux cas d'utilisation portant sur des contraintes et des besoins utilisateurs différents ainsi que des jeux de données et des modèles différents. Les études faites sur ces cas d'utilisation nous permettent de mettre en évidence les interactions entre les hyperparamètres et les propriétés des modèles XAI, ainsi que les interactions entre les propriétés elles-mêmes.

En outre, les travaux présentés dans ce chapitre font référence à la publication suivante [Cugny et al., 2022] et au travail de thèse de Robin Cugny.

4.2 La prise en compte du contexte utilisateur

Actuellement, si un utilisateur souhaite sélectionner une solution XAI, il peut s'appuyer sur des bibliothèques XAI, quelques benchmarks et solutions AutoML disponibles. Les bibliothèques XAI disponibles telles que DeepExplain [Ancona et al., 2018], AI Explainability 360 [Arya et al., 2019], et Alibi [Klaise et al., 2021] rassemblent les solutions XAI les plus actuelles. Cependant, elles n'intègrent pas l'évaluation automatique des explications et ne recommandent pas de solutions XAI en fonction des besoins et des contraintes des utilisateurs.

Les comparatifs et tentatives de benchmarks disponibles [Yeh et al., 2019, Hooker et al., 2019, Alvarez-Melis and Jaakkola, 2018] comparent, quant à elles, l'efficacité des solutions XAI en utilisant des métriques d'évaluation XAI. Mais ils sont souvent associés à un modèles prédictif ou un jeu de données spécifique, qui ne sont pas forcément ceux souhaités par l'utilisateur. De plus, les hyperparamètres des solutions XAI ne sont pas optimisés pour maximiser les diverses propriétés dont l'utilisateur a besoin. Ce dernier point est problématique, car certaines propriétés XAI telles que l'exactitude et la compacité ne sont pas toujours indépendantes.

Enfin, renforçant l'intérêt de notre proposition, [Vermeire et al., 2021] souligne que les utilisateurs devraient être guidés dans le choix des solutions XAI et propose d'ailleurs une méthodologie sur cette question, tandis que [Palacio et al., 2021] propose un cadre théorique pour faciliter la comparaison entre les solutions XAI.

4.2.1 Aider l'utilisateur face aux multiples solutions XAI

Au cours de la dernière décennie, l'XAI a proposé une grande variété de solutions pour faciliter la compréhension des modèles d'apprentissage automatique [Adadi and Berrada, 2018, Carvalho et al., 2019, Molnar, 2020, Barredo Arrieta et al., 2020, Doshi-Velez and Kim, 2017, Lipton, 2018, Gilpin et al., 2018].

Compte tenu du nombre croissant de propositions d'XAI [Barredo Arrieta et al., 2020], l'évaluation de la qualité des explications est devenue nécessaire pour choisir une solution d'XAI appropriée ainsi que ses hyperparamètres. Il convient de noter que l'évaluation des explications peut être effectuée soit subjectivement par des humains, soit objectivement à l'aide de mesures [Nguyen and Martínez, 2020, Zhou et al., 2021, Nauta et al., 2022]. Cependant, les utilisateurs qui souhaitent inclure une solution XAI sont confrontés aux problèmes suivants :

- Ils doivent vérifier quelles solutions XAI sont compatibles avec le type de données et le modèle d'apprentissage.
- Les solutions XAI doivent expliquer spécifiquement ce que les utilisateurs souhaitent comprendre, et ce, dans une restitution appropriée.
- Ils doivent évaluer l'efficacité des explications produites par les solutions XAI sélectionnées.
- Ce contexte exige que les explications répondent à des critères de qualité spécifiques (appelés propriétés des explications), ce qui impose l'utilisation de métriques d'évalua-

tion appropriées.
— Ils doivent trouver les meilleurs hyperparamètres pour chacune des solutions XAI sélectionnées afin de conserver la meilleure d'entre elles.

Ainsi, la proposition discutée dans ce chapitre a pour ambition de fournir un système automatique de recommandation de solutions XAI, en prenant en compte le contexte de l'utilisateur, constitué de : son jeu de données à analyser, le modèle prédictif utilisé sur ce jeu de données et les besoins et contraintes liés aux explications souhaitées.

Proposer une solution XAI adaptée nécessite de définir les éléments des contextes des utilisateurs et de les utiliser pour filtrer les solutions XAI compatibles. Cette tâche est difficile, car il y a autant de formalisations que d'auteurs dans le domaine du XAI et très peu de travaux ont tenté d'unifier les éléments du XAI avec une formalisation [Lundberg and Lee, 2017, Palacio et al., 2021, Nauta et al., 2022]. Comme il faut pouvoir évaluer les solutions XAI, nous devons trouver automatiquement les métriques d'évaluation XAI compatibles avec les bonnes solutions XAI et qui sont significatives pour le contexte utilisateur. De plus, il est nécessaire de trouver un moyen d'évaluer de multiples propriétés sur les explications, tout en tenant compte des préférences des utilisateurs. En outre, le classement des solutions XAI nécessite de trouver les meilleurs hyperparamètres en utilisant les métriques d'évaluation XAI. Comme cela est coûteux en termes de calcul, nous comptons nous inspirer des stratégies d'évaluation de modèles dans AutoML [He et al., 2021] qui permettent d'économiser du temps.

4.3 Les propriétés et métriques des modèles d'explication

Avant de rentrer dans les détails formels des propriétés et métriques pour les modèles d'explication, nous proposons un exemple illustrant l'intérêt de notre approche.

4.3.1 Exemple illustratif

Prenons tout d'abord l'exemple d'une data-scientist dans un laboratoire médical, Alice, et de Bob, un collègue médecin. Bob utilise un modèle boîte noire d'apprentissage automatique comme outil d'aide à la décision et demande s'il est possible d'obtenir une explication pour les prédictions du modèle afin de vérifier certains cas rares. Alice a accès au modèle, ainsi qu'aux données qui ont été utilisées pour le construire, et souhaite maintenant mettre en œuvre une solution XAI adaptée pour produire des explications. Ici, les besoins de Bob, le médecin, sont les suivants : les explications doivent se concentrer sur les prédictions (puisqu'il est demandé pourquoi elles sont obtenues) et la solution XAI doit expliquer le modèle entraîné sans le modifier. De plus, Bob veut savoir comment les données collectées pour un patient (les variables) influencent le résultat du modèle.

En ce qui concerne les contraintes du contexte, les décisions à fort enjeu imposent l'utilisation d'un modèle précis et des explications les plus fidèles possibles (propriété d'exactitude). Néanmoins, les explications doivent être les plus cohérentes entre elles, avec le moins de perturbations possible (dû à des mesures de prises de sang bruitées par exemple) : des explications

stables sont donc obligatoires (propriété de continuité). Enfin, puisque Bob sera le principal utilisateur de ces explications, des explications concises doivent être proposées et mettre de côté les variables sans importance (propriété de compacité).

4.3.2 Définitions

Préalablement, nous considérons les définitions classiques des jeux de données et modèles prédictifs pour l'apprentissage supervisé proposés dans le Chapitre 2.3.1.

Définition 4.1 (Explanandum et explanan). Nous notons $\mathcal{E} = \{\mathcal{E}_i\}_{i=1}^{k}$ L'ensemble de tous les explanandum possibles, où l'explanandum $\mathcal{E}_i$ est un descripteur pour les fonctions d'explication spécifiant *ce qui est expliqué*. Nous notons aussi $\mathcal{E}' = \{\mathcal{E}'_j\}_{j=1}^{k'}$ l'ensemble de tous les explanan possibles, où l'explanan $\mathcal{E}'_j$ est un descripteur pour les fonctions d'explication spécifiant *comment c'est expliqué*.

Dans notre exemple illustratif, $\mathcal{E}_i$ =*Pourquoi cette prédiction ?* et $\mathcal{E}'_j$ =*résumé par variable.*

Définition 4.2 (Propriétés XAI). Les propriétés XAI sont des critères de qualité descriptifs pour les explications. Nous notons P_r, l'ensemble des propriétés que les explications vérifient ou non.

Dans notre exemple illustratif, les propriétés d'intérêt sont l'exactitude, la continuité et la compacité. Ainsi, dans AutoXAI, le data-scientist peut spécifier ses besoins avec $(\mathcal{E}, \mathcal{E}')$ et ses contraintes avec P_r.

Définition 4.3 (Solution XAI). Une solution XAI agit comme une fonction qui produit une ou plusieurs explications. Nous notons $E = \{e_t\}_{t=1}^{l}$, l'ensemble des explications avec $l \in \mathbb{N}$ le nombre d'explications. Nous notons $f_e^{(h)} : P(X, Y, F, \hat{Y}) \to E$ la fonction d'explication avec $P(X, Y, F, \hat{Y})$ une partition de $\{X, Y, F, \hat{Y}\}$ et h les hyperparamètres de la solution XAI. $f_e^{(h)} \in F_e$ avec F_e l'ensemble des fonctions d'explication. Les hyperparamètres font référence aux paramètres statiques qui déterminent les comportements de la solution XAI. Pour des modèles intrinsèquement interprétables $f = f_e^{(h)}$.

Dans notre exemple illustratif, les solutions XAI comme LIME [Ribeiro et al., 2016] et Kernel SHAP [Lundberg and Lee, 2017] sont considérées comme $f_e^{(h)} : (X, F) \to E$. Les deux solutions produisent une explication par influence de variables $e_t \in \mathbb{R}^d$ pour chaque patient de sorte que $E = \{e_t\}_{t=1}^{n}$.

Définition 4.4 (Métriques d'évaluation XAI). Une métrique d'évaluation XAI évalue une propriété et est souvent adaptée à un type d'explication spécifique. Nous notons l'ensemble des métriques d'évaluation XAI $M = \{m_q\}_{q=1}^{c}$, où $m_q : P(X, F, F_e, Y) \to \mathbb{R}$, avec $P(X, F, F_e, Y)$ une partition de $\{X, F, F_e, Y\}$, de sorte que m_q évalue $p_q \in P_r$.

Dans notre exemple illustratif, la robustesse $m_q : (X, F_e) \to \mathbb{R}$ est une métrique d'évaluation vérifiant la propriété de continuité p_q.

4.4 AutoXAI

Nous décrivons d'abord le processus d'AutoXAI globalement, par étape et pour chaque composant, comme le montre la Figure 4.1a, puis nous détaillons l'optimiseur d'hyperparamètres dans la Figure 4.1b et enfin chaque composant en utilisant les définitions de la section 4.3.2.

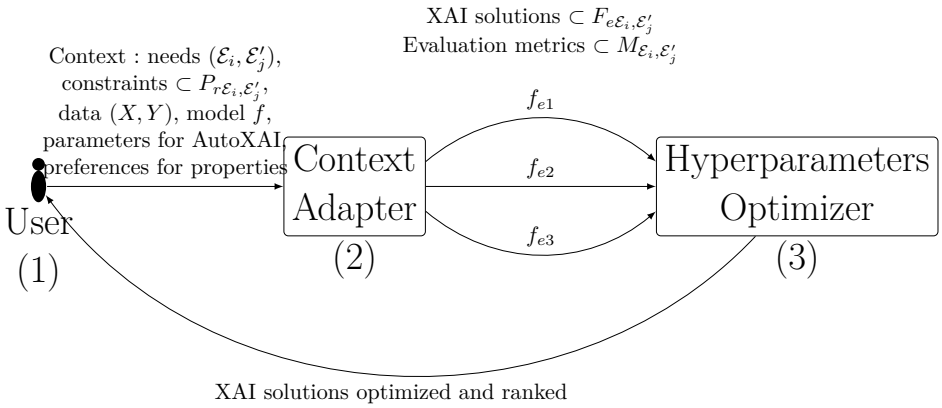
Voici les opérations telles qu'elles sont effectuées, en commençant par la Figure 4.1a :

1. L'utilisateur donne les éléments du contexte, les paramètres pour l'AutoXAI et ses préférences concernant les propriétés XAI.
2. Le composant *Context Adapter* sélectionne un sous-ensemble de solutions XAI correspondant aux besoins et un sous-ensemble de métriques d'évaluation pour s'assurer que les contraintes sur les propriétés sont respectées.
3. Pour chaque solution XAI, l'optimiseur d'hyperparamètres recherche les hyperparamètres qui réduiront la fonction de perte sur la base des scores agrégés des métriques d'évaluation. Pour ce faire, il effectue les opérations suivantes en boucle, voir la Figure 4.1b.
   (a) l'estimateur d'hyperparamètres propose de nouveaux hyperparamètres en fonction de l'algorithme d'optimisation choisi.
   (b) L'*Explainer* utilise la solution XAI et les nouveaux hyperparamètres proposés pour produire des explications.
   (c) L'évaluateur applique les métriques d'évaluation aux explications et agrège les scores ainsi obtenus.

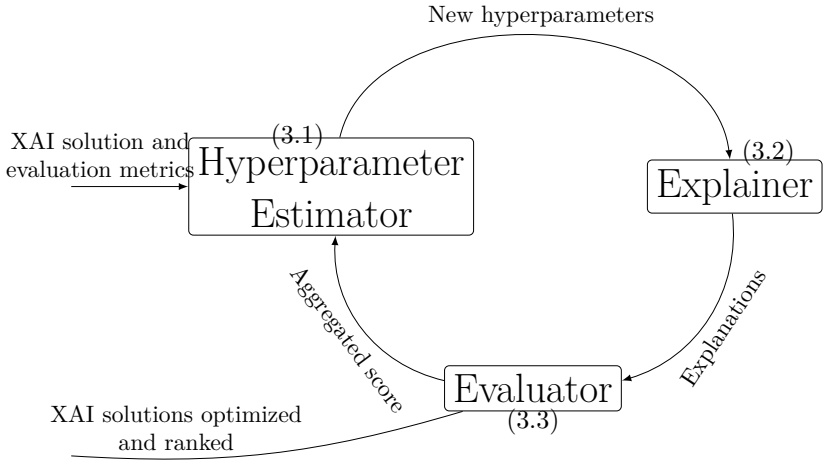
4.4.1 Context adapter

Come détaillé dans la section 4.3.2, les solutions XAI peuvent être regroupées selon $(\mathcal{E}_i, \mathcal{E}'_j)$, leur explanandum et leur explanan. Ce regroupement détermine aussi $P_{r\mathcal{E}_i,\mathcal{E}'_j}$, les propriétés décrivant les solutions XAI et par conséquent $M_{\mathcal{E}_i,\mathcal{E}'_j}$, les métriques d'évaluation XAI qui peuvent être appliquées. Pour obtenir $(\mathcal{E}_i, \mathcal{E}'_j)$, nous demandons à l'utilisateur ce qu'il veut à l'aide de réponses pré-écrites en langage naturel. Pour ce faire, AutoXAI utilise une banque de questions de [Liao et al., 2020] afin de proposer des explanandum, laissant ainsi l'utilisateur la possibilité de choisir à quelle question la solution XAI devra répondre. Cette banque de questions a été spécifiquement conçue afin de soutenir la conception d'applications XAI centrées sur l'utilisateur. Cette banque inclut des questions sur les données, leur traitement et les résultats pour l'apprentissage automatique. Concernant les explanan, AutoXAI utilise la liste des types d'explication disponibles dans [Carvalho et al., 2019]. Ce dernier liste les principaux types d'explications existantes : influence de variables, modèles internes, points de données, modèle de substitution intrinsèquement interprétable, ensembles de règles, explications en langage naturel et réponses aux questions.

Avec ces connaissances, un système de recommandation tenant compte du contexte est alors possible en sélectionnant $F_{e\mathcal{E}_i,\mathcal{E}'_j}$ dans F_e, $P_{r\mathcal{E}_i,\mathcal{E}'_j}$ dans P_r et $M_{\mathcal{E}_i,\mathcal{E}'_j}$ dans M. Ceci est



(a) Architecture globale d'AutoXAI



(b) Vue détaillée de l'optimisation des hyperparamètres

FIGURE 4.1 – Architecture d'AutoXAI.
Les figures sont lues en suivant le numéro des étapes. Dans la Figure 4.1a, pour chaque solution XAI, l'étape (3) optimise ses hyperparamètres par rapport aux scores agrégés des mesures d'évaluation en entrant dans la boucle de la Figure 4.1b.

possible en étiquetant chacune de ces propositions avec un tuple de $(\mathcal{E}, \mathcal{E}')$ en utilisant des tables de correspondance [Liao et al., 2020, Nauta et al., 2022]. $P_{r\mathcal{E}_i,\mathcal{E}'_j}$ sert à la modélisation contextuelle du système de recommandation. En effet, l'utilisateur peut choisir les poids (degré d'importance) pour chacune des propriétés dans $P_{r\mathcal{E}_i,\mathcal{E}'_j}$. Ces poids sont utilisés dans l'évaluation des solutions XAI afin de guider l'optimisation et par conséquent le classement des solutions XAI produites. 1 est la valeur par défaut ; augmenter un poids w_q rend sa propriété p_q plus importante et inversement, 0 signifie que la propriété est ignorée. En plus de ces éléments, AutoXAI peut récupérer, si nécessaire, le modèle et le jeu de données sur lequel il a été entraîné.

4.4.2 Estimateur d'hyperparamètres

L'objectif de ce composant est de proposer de nouveaux hyperparamètres pour obtenir le meilleur score agrégé. Les algorithmes d'optimisation sont itératifs et certains, comme l'optimisation bayésienne [Snoek et al., 2012], associent les hyperparamètres à un score pour construire un modèle probabiliste. Ce modèle probabiliste estime les hyperparamètres qui devraient donner le meilleur score. Ici, f_e, avec des hyperparamètres précédemment estimés, est évaluée avec les scores agrégés de $M_{\mathcal{E}_i,\mathcal{E}'_j}$ (détaillé davantage dans la Section 4.4.4) et il en résulte une nouvelle entrée pour mettre à jour le modèle probabiliste. Ce composant estime les hyperparamètres en fonction des résultats des scores précédents, le cas échéant. Sinon, les valeurs des hyperparamètres sont fixées en fonction de l'initialisation de l'algorithme choisi, par exemple au hasard.

4.4.3 L'explainer

L'objectif de ce composant est de produire des explications en utilisant la solution XAI et les hyperparamètres définis. Comme les implémentations des solutions XAI varient selon leur paradigme de programmation et dans la structure de données qu'elles utilisent et renvoient, il est nécessaire de mettre en place un wrapper qui standardise l'entrée et la sortie pour chaque solution XAI d'un type donné. Ce composant sert de base pour inclure toutes les solutions XAI mises en œuvre et est conçu pour être complété par de nouvelles solutions XAI à venir.

4.4.4 L'évaluateur

Ce composant vise à calculer les scores des métriques d'évaluation XAI correspondants aux propriétés XAI demandées par l'utilisateur. Comme pour le composant précédent, il agit également comme un wrapper qui normalise l'entrée et la sortie des métriques d'évaluation XAI et sert de base pour inclure n'importe quelle métrique d'évaluation XAI. Il agrège également les scores des propriétés afin de fournir un objectif d'optimisation unique pour la recherche des hyperparamètres adéquats. Pour trouver les meilleurs hyperparamètres h de la solution XAI f_e, nous définissons l'objectif d'optimisation comme suit :

$$\max_{h \in H} A(f_e, h) \tag{4.1}$$

Avec A la fonction d'agrégation et H l'ensemble de tous les hyperparamètres possibles. A doit rassembler les multiples scores renvoyés par les métriques d'évaluation XAI évaluant les propriétés choisies et pondérées en fonction des préférences de l'utilisateur. Pour ce faire, nous optons pour une scalarisation linéaire [Miettinen, 2012] :

$$A(f_e, h) = \frac{1}{c'} \sum_{q=1}^{c'} w_q \times sc_q(m_q(f_e^{(h)})) \tag{4.2}$$

La métrique d'évaluation XAI pour p_q est notée $m_q(f_e^{(h)})$ pour être plus concise, bien qu'elle puisse utiliser n'importe quelle partition de $\{X, F, F_e, Y\}$ telle que définie dans 4.4. c' est le nombre de propriétés choisies et donc de métriques d'évaluation XAI. Les poids w_q représentent le degré d'importance défini par l'utilisateur et $sc_q(.)$ est une fonction de mise à l'échelle basée sur les résultats précédents pour une propriété p_q. Cette fonction de mise à l'échelle permet de ne pas favoriser une métrique d'évaluation XAI par rapport à une autre. Pour la première itération de AutoXAI, il n'y a pas de résultats antérieurs et la mise à l'échelle ne peut pas être appliquée. Par conséquent, nous initialisons $sc_q(.)$ en utilisant les scores d'évaluation des solutions XAI avec les hyperparamètres par défaut. Ce démarrage à froid permet également de vérifier si les solutions XAI peuvent être performantes avec les hyperparamètres par défaut.

Bien que ce cadre produise déjà un classement des solutions XAI, le temps de calcul peut être un problème et doit être pris en compte, c'est ce que nous allons voir dans la prochaine section.

4.5 Stratégies d'évaluation pour minimiser le temps de calcul

Certaines solutions XAI et métriques d'évaluation XAI n'ont pas été conçues pour être utilisées plusieurs fois de suite et présentent une grande complexité algorithmique. Pour réduire le coût en temps de ces algorithmes, sans changer leur architecture, nous proposons d'adapter les stratégies heuristiques existantes dans le domaine de l'AutoML. AutoXAI adapte trois stratégies provenant de l'AutoML pour réduire le temps de calcul : l'échantillonnage, l'arrêt précoce et l'échange d'information. Ces stratégies sont détaillées ci-après.

4.5.1 L'échantillonnage

La réduction du nombre d'explications produites permet de réduire le nombre d'opérations liées à l'élaboration et à l'évaluation des explications. Il est également possible d'utiliser un sous-ensemble du jeu de données pour créer des explications, ce qui peut réduire le nombre d'opérations pour la production des explications. Ces approximations peuvent être suffisamment précises en fonction de la diversité de l'ensemble de données, de la complexité du modèle et de la sensibilité de la solution XAI. Ce problème est résolu en spécifiant le pourcentage d'explications à traiter.

4.5.2 L'arrêt précoce

L'arrêt précoce peut être effectué lors du calcul de la métrique d'évaluation XAI ou de l'optimisation des hyperparamètres. Dans les deux cas, cette stratégie peut faire gagner beaucoup de temps, mais elle doit être mise en place correctement pour éviter une approximation trop grossière. Pour le calcul des métriques d'évaluation, cette option n'est possible que si la métrique d'évaluation XAI est appliquée de manière séquentielle aux explications. En outre, cette stratégie est plus efficace si les explications sont également calculées de manière séquentielle ; en effet, moins d'explications sont calculées de cette manière. La condition d'arrêt est que la métrique d'évaluation XAI n'évolue plus que par un petit pourcentage ; en dessous d'un seuil donné et pendant plusieurs itérations. Nous considérons alors qu'elle converge. Concernant l'optimisation des hyperparamètres, le raisonnement est le suivant : si le meilleur score ne change pas pendant plusieurs itérations, alors il a été trouvé. Ici, le choix d'un seuil est important pour éviter de passer à côté d'une meilleure solution.

4.5.3 L'échange d'information

Un autre moyen d'économiser du temps de calcul consiste à réutiliser les résultats intermédiaires et à partager les informations entre les évaluations. Dans l'AutoML, la stratégie de partage des poids d'un ancien modèle accélère l'apprentissage d'un nouveau modèle [He et al., 2021]. Dans AutoXAI, certains calculs intermédiaires peuvent être réutilisés de la même manière. Cette stratégie est particulièrement efficace pour certains calculs intermédiaires coûteux comme un processus gaussien [Alvarez-Melis and Jaakkola, 2018].

4.6 Prototype et évaluation

Dans les expériences suivantes, une implémentation du cadre AutoXAI, décrit dans la section précédente, est appliquée à deux cas d'utilisation ainsi qu'à l'exemple illustratif. Le code permettant de reproduire les résultats de ces cas d'utilisation est disponible à l'adresse suivante : `https://github.com/RobinCugny/AutoXAI`

4.6.1 Estimation du diabète

En reprenant l'exemple illustratif décrit en début de chapitre, nous utilisons les jeux de données du Diabète [Efron et al., 2004] et Indiens Pima [Smith et al., 1988]. Le jeu de données du Diabète comporte 10 variables et est conçu pour une tâche de régression visant à prédire l'évolution de la maladie. Le jeu de données sur les Indiens Pima comporte 8 variables et est conçu pour une classification binaire visant à prédire si les patients sont atteints de diabète, ou non. Le modèle boite noire utilisé est l'implémentation Scikit-learn d'un Perceptron Multi-couches [Pedregosa et al., 2011]. Concernant la tâche de régression, nous utilsons

MLPRegressor [1] et pour la classification nous utilisons MLPClassifier [2].

Les solutions XAI mises en œuvre dans cet exemmple sont LIME [Ribeiro et al., 2016] et Kernel SHAP [Lundberg and Lee, 2017], renvoyant tous deux des scores d'importance, pour chaque variable, vis-à-vis d'un modèle prédictif, et pour chaque instance.

Les métriques d'évaluation XAI mises en œuvre et leurs propriétés correspondantes sont les suivantes :

— **Robustesse** [Alvarez-Melis and Jaakkola, 2018] *Continuité*
— **Fidélité** [Yeh et al., 2019] *Exactitude*
— **Nombre de variables** [Rosenfeld, 2021] *Compacité*

[Alvarez-Melis and Jaakkola, 2018] propose la métrique de **robustesse** qui évalue la *Continuité* en adaptant la continuité Lipschitzienne. L'objectif est de mesurer les changements dans les explications alors que le jeu de données d'entrée subit de petites perturbations. En effet, les explications ne doivent pas changer radicalement si l'instance ne change pas non plus. [Yeh et al., 2019] propose d'évaluer l'*Exactitude* avec une métrique de **Fidélité**. Elle consiste à perturber l'entrée, sur la base des scores d'influences des variables dans les explications, et à mesurer pour chaque nouvelle entrée le changement dans la sortie de la fonction prédictive f. La *compacité* est évaluée avec le **nombre de variables** qui correspond à la taille du vecteur d'explications. Voir le tableau 4.1 pour les formules.

1. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html
2. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

TABLE 4.1 – Métriques d'évaluation XAI

Propriété	Métrique	Formule
Continuité	**Robustesse**	$\hat{L}(x_i) = \max_{x_j \in B_\epsilon(x_i)} \frac{\|f_e^{(h)}(x_i) - f_e^{(h)}(x_j)\|_2}{\|x_i - x_j\|_2}$ Avec $B_\epsilon(x_i)$ une boule de rayon ϵ^* centrée sur x_i le point de donnée étudié. La méthode d'optimisation recherche le point dans l'espace de données avec le ratio le plus élevé dans le voisinage. Cette valeur est conservée comme mesure du manque de robustesse $\hat{L}(x_i)$.
Exactitude	**Fidélité**	$INFD(e_i, f, x_i) = \mathbb{E}_{I \sim \mu I}[(I^T e_i - (f(x_i) - f(x_i - I)))^2]$ avec I, les perturbations autour de x_i tel que $I = x_i - x_i'$, Nous choisissons ici une mise en œuvre bruitée avec $x_i' = x_i + \epsilon$ et ϵ un bruit uniforme.
Compacité	**Nombre de variables**	$NoF(e_i) = \text{Card}(e_i)$ avec $e_i \in E$, le i-ème vecteur d'explications, par influence de variables.
Complétude	**Non-représentativité**	$NR(E) = \frac{1}{n} \sum_{i=1}^{n} \min_{e_j \in E} d(x_i, e_j)$ avec $n = \text{Card}(X)$, d une fonction de distance et E l'ensemble des explications composé de points de données représentaifs e_j, aussi appelées prototypes.
Compacité (*Redondance*)	**Diversité**	$Div(E) = \sum_{\{e_i, e_j\} \in P_2(E)} \frac{d(e_i, e_j)}{C_2^l}$ avec $P_2(E)$, l'ensemble des combinaisons de deux prototypes, d une fonction de distance et C_2^l le nombre de combinaisons avec $l = \text{Card}(E)$ les prototypes.
Compacité (*Size*)	**Number of prototypes**	$NoP(E) = \text{Card}(E)$ avec E, l'ensemble des prototypes.

*L'implémentation est une zone ayant la taille de l'écart-type, il s'agit d'une variation proposée par les auteurs originaux.

Pour l'agrégation dans ce scénario, Alice et Bob fixent les poids à 1, 2 et 0,5 pour la **Robustesse**, **Fidélité** et le **Nombre de variables** respectivement. Alice fixe le nombre d'itérations à 25.

La stratégie d'optimisation des hyperparamètres est une optimisation bayésienne. Nous utilisons l'implémentation par processus gaussien [Nogueira, 14]. Afin d'optimiser le temps de calcul, nous utilisons *l'arrêt précoce* pour le calcul des mesures d'évaluation XAI et de recherche d'hyperparamètres, ainsi que l'échange d'informations pour les calculs de robustesse et de fidélité. Pour la robustesse, les informations partagées entre les itérations de AutoXAI sont les points de données donnant le score maximal de robustesse (voir le tableau 4.1), que nous appelons maxima en abrégé. Pour la fidélité, les informations partagées entre les itérations sont les points de perturbation générés et les prédictions du modèle pour ceux-ci.

Un extrait du classement produit par AutoXAI pour le jeu de données sur le Diabète figure dans le tableau 4.2 et celui pour le jeu de données sur les Indiens Pima dans le tableau 4.3. Les solutions XAI sont classées par ordre décroissant en fonction du score agrégé produit par l'équation 4.2. Pour illustrer la diversité des solutions XAI, nous présentons trois combinaisons d'hyperparamètres avec LIME et trois avec SHAP. Le choix du nombre de variables pour l'explication est un processus subjectif qui nécessite de visualiser les explications, nous avons opté pour un nombre de variables de 1, 3 et 5. Ainsi, l'utilisateur peut vérifier si des explications courtes sont suffisantes pour comprendre la prédiction ou si davantage de variables seraient utiles. Dans les colonnes *Hyperparamètres*, les deux premiers hyperparamètres pour LIME comme pour SHAP sont : premièrement, le nombre de variables dans l'explication, et deuxièmement, le nombre de perturbations utilisées pour construire le modèle linéaire. Le dernier hyperparamètre pour SHAP est la régularisation *l1* à utiliser pour la sélection des variables.

TABLE 4.2 – Extrait du classement produit par AutoXAI sur le jeu de données Diabète.

Score agrégé	Robustesse	Fidélité	Nombre de Variables	Solution XAI	Hyperparamètres
1.023	0.727	0.833	1.351	LIME	1 ;3656
1.019	0.703	0.991	0.745	LIME	3 ;8782
0.963	0.682	1.068	0.139	LIME	5 ;5392
-0.287	0.310	-0.924	1.351	SHAP	1 ;1304 ;auto
-0.633	-0.319	-0.975	0.745	SHAP	3 ;1571 ;aic
-0.639	0.014	-1.000	0.139	SHAP	5 ;1148 ;aic

En ce qui concerne le tableau 4.2, LIME est systématiquement mieux classé que SHAP avec ces métriques d'évaluation XAI. Plusieurs facteurs pourraient expliquer pourquoi SHAP n'obtient pas de meilleurs résultats : avec les hyperparamètres par défaut, SHAP présente une robustesse moyenne inférieure à celle de LIME sur certains jeux de données de l'UCI d'après [Dua and Graff, 2017] (Glass, Wine et Leukemia) et surtout un écart-type plus important [Alvarez-Melis and Jaakkola, 2018]. Pour le jeu de données Diabète, comme nous pouvons le

TABLE 4.3 – Extrait du classement produit par AutoXAI sur le jeu de données des Indiens Pima.

Score agrégé	Robustesse	Fidélité	Nombre de Variables	Solution XAI	Hyperparamètres
1.412	0.744	1.435	1.243	LIME	1 ;5347
1.282	0.575	1.325	1.243	SHAP	1 ;509 ;bic
0.361	0.633	0.117	0.430	LIME	3 ;8329
0.176	0.339	-0.014	0.430	SHAP	3 ;713 ;auto
0.070	0.262	0.070	-0.383	SHAP	5 ;537 ;bic
-0.185	0.599	-0.481	-0.383	LIME	5 ;7023

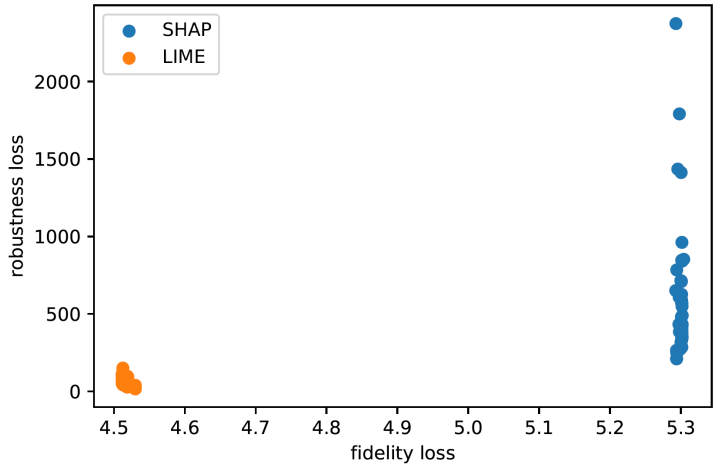


FIGURE 4.2 – Perte de robustesse et perte de fidélité pour le jeu de données sur le diabète

voir dans la Figure 4.2, nous observons également un écart-type important et une robustesse moyenne plus élevée pour SHAP avec différents hyperparamètres. Nous observons en particulier qu'il a constamment une perte de fidélité plus élevée avec une distribution resserrée sur ce score. Selon [Yeh et al., 2019], cela signifie que SHAP capture moins bien la façon dont la prédiction du modèle change en réponse aux perturbations. Comme Kernel SHAP s'appuie sur la stratégie LIME pour construire un modèle linéaire local, cela signifie que la perte pourrait provenir de l'inclusion des valeurs de Shapley. Une hypothèse serait que le modèle boîte noire ne raisonne pas de la même manière, vis-à-vis des relations entre les variables, que Kernel SHAP en les détectant avec son approximation linéaire.

En ce qui concerne le jeu de données des Indiens Pima, il semble toutefois que SHAP soit légèrement plus performant. Dans la Figure 4.3, nous pouvons voir que SHAP est moins robuste que LIME la plupart du temps, mais qu'il a une fidélité équivalente. Ici, SHAP semble réussir à capturer les changements de la fonction prédictive, et pourrait donc trouver plus d'interactions entre variables en commun avec le modèle prédictif.

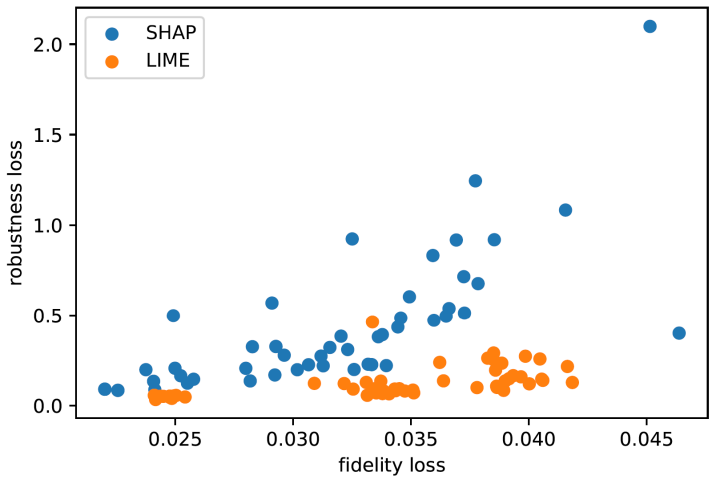


FIGURE 4.3 – Perte de robustesse et perte de fidélité pour le jeu de données des Indiens Pima

La compacité a un impact sur les autres propriétés [Nauta et al., 2022]. En effet, Bob, le médecin, devrait analyser les explications pour confirmer le nombre de variables nécessaires à la compréhension de la prédiction. Il convient donc de comparer les solutions XAI à l'aide du score agrégé en tenant compte du nombre de variables. Pour cela, prenons une instance particulière du jeu de données sur le Diabète. Le modèle fait une prédiction et Bob demande *Pourquoi cette prédiction ?* et veut connaître les variables qui contribuent à cette prédiction. Les explications produites par les solutions XAI recommandées par AutoXAI sont présentées dans la Figure 4.4. En bas à droite se trouve LIME avec les hyperparamètres par défaut. Nous pouvons voir que certaines variables ont peu d'influence sur la prédiction et sont inutiles pour répondre à la question de Bob. Avec ces explications de tailles différentes, Bob peut voir ce qui est important et ce qui est négligeable pour lui. Il peut ainsi choisir la taille de l'explication qu'il souhaite, en gardant à l'esprit les scores des autres propriétés et le score agrégé.

4.6.2 Détection de SPAM

Pour le deuxième cas d'utilisation, nous considérons Charli, responsable de la sécurité informatique dans un laboratoire. Le personnel se plaint de recevoir des spams dans le service de chat du laboratoire. Charli gère ce service et n'a que peu de connaissances en apprentissage automatique. Charli souhaite utiliser un algorithme prédictif pour détecter automatiquement les spams. Par conséquent, Charli trouve un *GloVe embedding* prêt à l'emploi [Pennington et al., 2014] et met en œuvre un LSTM [Hochreiter and Schmidhuber, 1997]. Charli entraîne le modèle sur un jeu de données de spams, obtient une bonne précision et décide de l'essayer une semaine sur le service de chat. À la fin de la semaine, un collègue de Charli lui montre un message qui n'a pas pu être envoyé parce qu'il est suspecté d'être un spam. Charli se demande alors "Où le modèle échoue-t-il ?", et plus précisément "À quoi ressemblent les faux positifs et les faux négatifs ?". Pour répondre à ces questions, Charli souhaite mettre

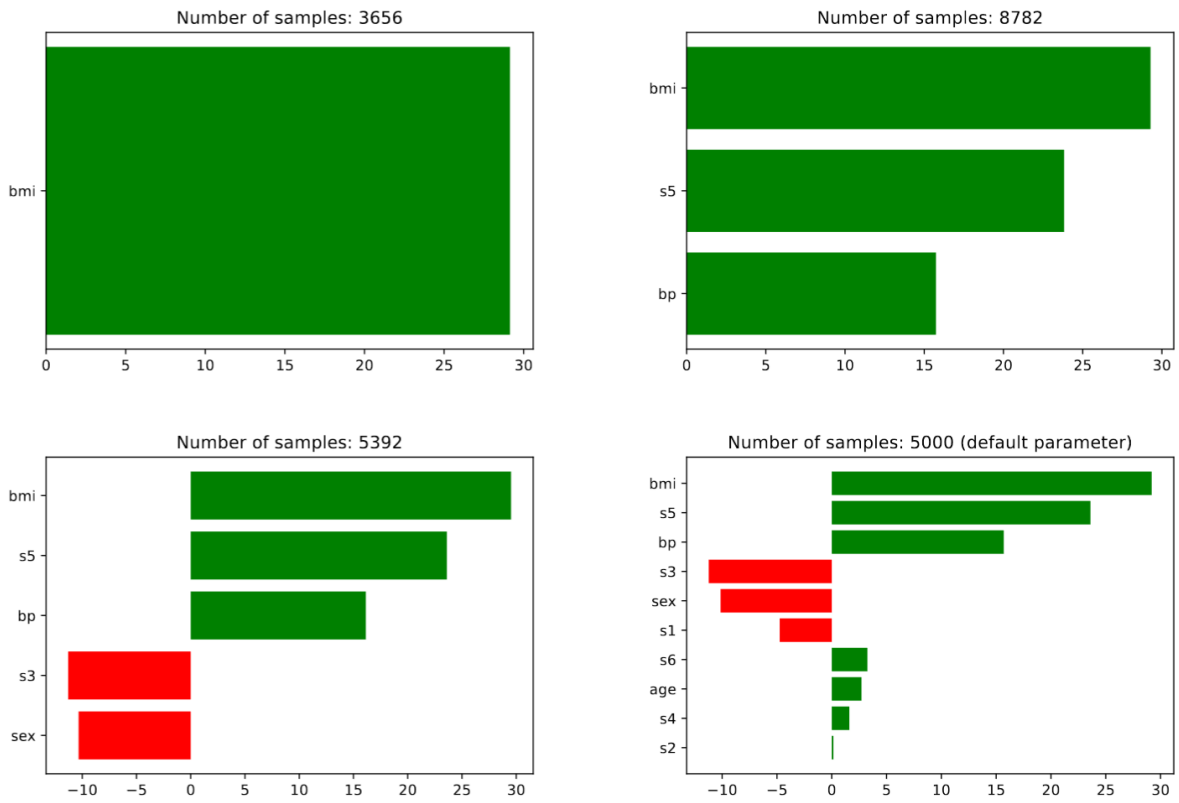


FIGURE 4.4 – Différentes tailles d'explications produites par LIME pour une instance tirée du jeu de données sur le diabète

en œuvre une solution XAI. En fin de compte, Charli veut utiliser ces connaissances pour pouvoir expliquer aux utilisateurs les décisions prises à partir de leurs données et éventuellement réduire le taux d'erreur.

Charli a besoin d'exemples de messages contenant des prédictions spécifiques pour savoir à quoi ils ressemblent. Nous définissons donc $\mathcal{E}_{i'}$ =*Quel type de données conduit à cette prédiction ?* et $\mathcal{E}'_{j'}$ =*Points de données en tant qu'explications* (également appelés prototypes).

Pour les contraintes du contexte, Charli ne veut manquer aucun type de message conduisant à une erreur. Idéalement, Charli souhaite que chaque point de données ait un prototype similaire (complétude). Cependant, Charli veut aussi éviter d'avoir trop de prototypes et éviter les redondances (compacité).

Le jeu de données utilisé est dérivé de celui de l'UCI SMS Spam [Dua and Graff, 2017]. Il possède 8714 variables pour 5572 instances et a été construit en utilisant le TFIDF [Jones, 1972]. Charli sépare ce jeu de données en quatre sous-ensembles en fonction des résultats du modèle : vrais positifs, vrais négatifs, faux positifs et faux négatifs. Ainsi, le modèle prédictif (GloVe et LSTM) n'est plus nécessaire, en tant que tel, pour la suite de l'expé-

TABLE 4.4 – Extrait du classement produit par AutoXAI sur le jeu de données SMS Spam

Aggregated score	Scaled Representativeness	Scaled Diversity	Scaled NoP	XAI solution	Hyperparameters
0.483	0.904	0.248	-0.303	k-medoids	heuristic ;300 ;pam ;cosine ;16
0.466	0.463	0.412	0.030	MMD-critic	1.000 ;13
0.384	0.224	0.201	0.251	Protodash	gaussian ;11.90 ;11
0.367	-0.660	0.589	0.917	MMD-critic	1.000 ;5
0.331	-0.444	0.048	0.917	k-medoids	build ;224 ;alternate ;cosine ;5
0.255	-0.580	0.092	0.917	Protodash	gaussian ;21.43 ;5

rience. Les solutions XAI mises en œuvre ici sont MMD-critic [Kim et al., 2016], Protodash [Gurumoorthy et al., 2019] et k-medoids [Kaufman and Rousseeuw, 1990].

MMD-critic propose des prototypes comme explications. Pour représenter fidèlement la distribution des données, il minimise l'écart entre les distributions des prototypes et celles des données. Il propose également des critiques, c'est-à-dire des points qui ne sont pas bien représentés par les prototypes. Protodash généralise [Kim et al., 2016], il s'agit d'une sélection rapide de prototypes qui associe également des poids (non négatifs) aux prototypes, ce qui indique leur importance. Enfin, bien qu'il n'ait pas été proposé pour l'XAI, nous utilisons k-medoids car il s'agit d'une référence donnant des résultats souvent comparables. Il trouve des médoïdes (prototypes dans notre cas) tels que la distance entre un prototype et les autres points de son groupe de données est minimale.

Les métriques d'évaluation XAI mises en œuvre et leurs propriétés correspondantes sont les suivantes :

— **Représentativité** *Complétude*
— **Diversité** [Nguyen and Martínez, 2020] *Compacité* (*redondance*)
— **Nombre de prototypes** *Compacité* (*taille*)

Nous proposons une nouvelle métrique de **représentativité** pour évaluer s'il existe un prototype proche pour chaque échantillon de données en moyenne. Contrairement à [Nguyen and Martínez, 2020], la proposition est indépendante du modèle, car les solutions XAI qui n'utilisent pas de modèle prédictif ne doivent pas être évaluées en fonction de ce modèle. Pour évaluer la *compacité*, nous utilisons le **nombre de prototypes** pour la taille de l'explication et la **diversité**. Pour la **diversité**, nous adaptons la proposition de [Nguyen and Martínez, 2020] pour mesurer la distance moyenne entre les prototypes. Comme la **diversité** et le **nombre de prototypes** sont fondamentalement différents, nous considérons qu'ils correspondent à deux sous-propriétés différentes (*redondance* et *taille* respectivement) et laissons l'utilisateur attribuer un poids à chacune d'elles. Le tableau 4.1 donne le détail des formules.

Pour l'agrégation, dans ce scénario, Charli fixe les pondérations à 2, 1, 2 pour la **représentativité**, la **diversité** et le **nombre de prototypes** respectivement. Charli, fixe le nombre d'itérations à 25. La stratégie d'optimisation des hyperparamètres concerne l'échange d'information à l'aide d'un processus gaussien.

Un extrait du classement produit par AutoXAI pour le jeu de données SMS Spam figure dans le tableau 4.4. Comme précédemment, les solutions XAI sont triées par ordre décroissant

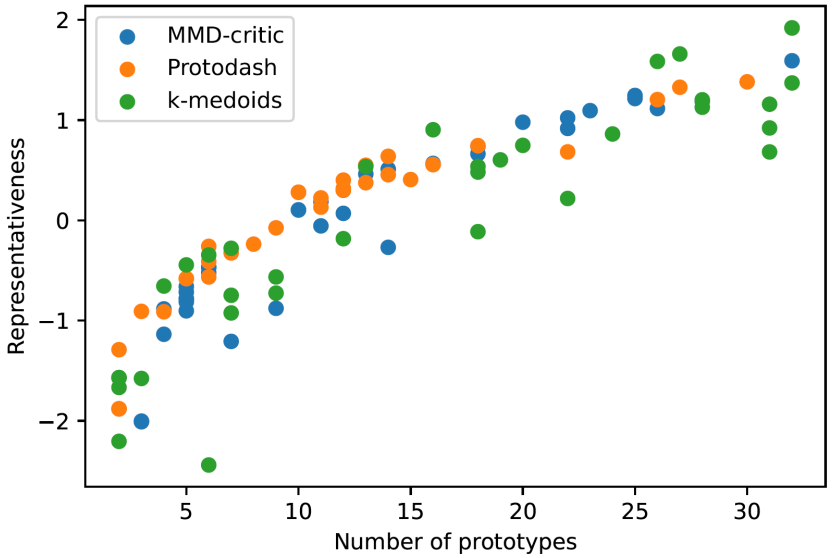


FIGURE 4.5 – Influence du nombre de prototypes sur la représentativité

en fonction de leur score agrégé. Pour chaque solution XAI, nous présentons deux résultats, le meilleur score global et le meilleur score pour un nombre de prototypes inférieur ou égal à 5. Dans les colonnes *Hyperparamètres*, k-medoids dispose des hyperparamètres suivants : la méthode d'initialisation, le nombre maximum d'itérations, l'algorithme à utiliser, la métrique et le nombre de medoïds à générer [3]. MMD-critic dispose des hyperparamètres suivants : la valeur gamma et le nombre de prototypes à trouver. Protodash dispose des hyperparamètres suivants : le noyau à utiliser, la valeur de sigma et le nombre de prototypes à trouver.

En ce qui concerne le tableau 4.4, nous observons que K-medoïds a le meilleur score agrégé avec une représentativité élevée. Nous observons également que le score de représentativité est systématiquement plus faible lorsque le nombre de prototypes est réduit. Cette tendance est illustrée dans la Figure 4.5. On constate que la représentativité est plus importante avec un plus grand nombre de prototypes. Ceci est évidemment attendu, car plus il y a de prototypes, plus il est probable qu'un point de données soit proche de l'un d'entre eux. Il en résulte un choix à faire entre compacité et exhaustivité qui incite à choisir des pondérations appropriées pour les propriétés. Dans le tableau 4.4, nous observons que les deux algorithmes pour les K-medoïds (*Pam* et *Alternate*) sont performants, tandis que Protodash semble avoir de meilleurs résultats avec le noyau gaussien. Cela est confirmé par la Figure 4.6 où Protodash obtient de meilleurs résultats en termes de diversité avec le noyau gaussien, tandis que les K-medoïds obtiennent des résultats équivalents avec les algorithmes *Pam* et *Alternate*. MMD-critic obtient régulièrement les meilleurs résultats en termes de diversité et K-medoïds présente une plus grande variance des résultats en termes de diversité et de représentativité.

3. https://scikit-learn-extra.readthedocs.io/en/stable/generated/sklearn_extra.cluster.KMedoids.html

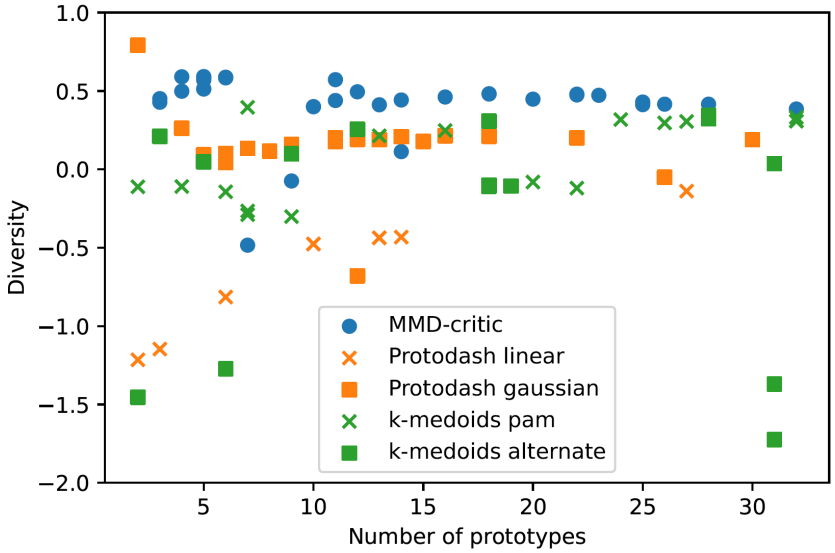


FIGURE 4.6 – Influence du nombre de prototypes sur la diversité

En utilisant la solution XAI la plus performante sur le jeu de données SMS Spam (K-medoïds en l'occurrence), Charli obtient les prototypes suivants pour les faux positifs :

— *Hey pple...$700 or $900 for 5 nights...Excellent location wif breakfast hamper!!!*
— *Unlimited texts. Limited minutes.*

Ainsi que les prototypes suivants pour les faux-négatifs :

— *FROM 88066 LOST £12 HELP*
— *Money i have won wining number 946 wot do i do next*

Grâce à ces phrases représentatives, Charli peut expliquer directement aux utilisateurs les types de messages que le modèle risque de mal classer et peut travailler sur les données et le modèle prédictif en analysant les prédictions et les prototypes, pour ces messages.

4.6.3 Evaluation des stratégies pour minimiser le temps de calcul

Les résultats préliminaires pour les stratégies d'évaluation minimisant le temps de calcul (voir section 4.5) sont obtenus sur le premier cas d'utilisation avec le jeu de données du Diabète et le modèle *MLPRegressor*. La solution XAI utilisée est LIME avec les hyperparamètres par défaut et la robustesse et la fidélité pour les mesures d'évaluation XAI. AutoXAI est exécuté sur un ordinateur portable équipé d'un processeur octa-core de 2,40 GHz. Le tableau 4.5 montre la moyenne avec son écart-type du temps de calcul et du score d'évaluation. L'arrêt précoce permet d'économiser 96.42% du temps pour la robustesse et 96.13% pour la fidélité. La stratégie par échange d'information permet d'économiser 93% du temps pour la robustesse et 21.26% du temps pour la fidélité. Concernant la robustesse, les maxima utilisés pour calculer les scores sont obtenus avec LIME par d'autres hyperparamètres, d'où la différence de scores

TABLE 4.5 – Temps de calcul et scores d'évaluation pour LIME sans et avec les stratégies d'évaluation pour minimiser le temps de calcul

	Sans stratégie		Arrêt précoce		Echange d'information		Les 2 stratégies ensembles	
	Temps	Score	Temps	Score	Temps	Score	Temps	Score
Robustesse	488.04 ± 14.79	-70.67 ± 0.64	17.45 ± 1.37	-67.41 ± 3.26	34.15 ± 1.70	-69.22 ± 2.24	1.22 ± 0.13	-67.62 ± 3.12
Fidélité	11.84± 0.27	-5.29 ± 0.06	0.46 ± 0.05	-5.00 ± 0.78	9.32 ± 0.37	-5.40 ± 0.12	0.35 ± 0.03	-5.48 ± 0.6

Les temps de calcul sont exprimés en secondes. Pour les scores d'évaluation, plus ils sont élevés, mieux c'est.

avec les résultats de la section précédente. Concernant la fidélité, les points de perturbation générés et leurs prédictions correspondantes sont obtenus avec un autre *seed*, ce qui explique la faible différence de score par rapport à la baseline sans stratégie. L'utilisation des deux stratégies permet d'économiser 99.75% du temps pour la robustesse et 97% pour la fidélité.

4.7 Conclusion

4.7.1 Bilan

Dans ce chapitre, nous avons proposé un système de recommandation de modèles XAI, nommé AutoXAI. AutoXAI automatise la tâche, classiquement fastidieuse, de sélection d'une solution XAI et de ses hyperparamètres. Il produit un classement des solutions en tenant compte des préférences de l'utilisateur. Comme nous l'avons vu, l'originalité est donc de combiner :

— la prise en compte du contexte de l'utilisateur, incluant : le jeu de données, le modèle prédictif et des préférences liées aux types d'explications souhaitées ainsi que leurs évaluations
— l'optimisation automatique des hyperparamètres des algorithmes de XAI, inspiré de l'AutoML

La prise en compte du contexte l'utilisateur dans un cadre XAI nécessite de récolter ses préférences. A l'aide de questions posées à l'utilisateur, résumées autour des notions de Explanandum (ce qui est expliqué) et Explanan (comment c'est expliqué), il est possible d'identifier le type d'explication possible. Cependant, et pour un même type d'explication, il est nécessaire de pouvoir évaluer les solutions de XAI correspondantes. Nous nous sommes donc concentrés sur les propriétés et métriques évaluant ces solutions afin de les classer. La notion de "bonne" explication n'est pas triviale et nous avons pu nous rendre compte de la difficulté à obtenir, par la littérature, une formalisation complète des propriétés et métriques faisant consensus.

L'optimisation automatique des hyperparamètres permet de renseigner automatiquement les hyperparamètres les plus adéquats pour les algorithmes XAI, afin que le modèle XAI produit réponde aux propriétés et métriques voulues par l'utilisateur. L'optimisation

est réalisée à l'aide d'un processus itératif, contrôlé par les scores agrégés des différentes métriques considérées (que l'utilisateur peut pondérer afin de prioriser des métriques par rapport à d'autres). Les stratégies d'optimisation du temps de calcul, inspiré aussi de l'AutoML et adapté au contexte XAI, ont montré leur claire efficacité.

4.7.2 Perspectives

L'approche originale proposée dans ce chapitre mène à une multitude de problématiques et travaux possibles.

Tout d'abord, et de manière similaire à notre proposition de recommandation de modèles prédictifs, il peut sembler délicat pour un utilisateur, qui plus est non expert, de devoir saisir des poids pour chaque propriété XAI souhaitée. Nous pourrions alors plutôt imaginer demander à l'utilisateur de saisir des préférences entre ces propriétés ou bien lui en recommander automatiquement par défaut. Les questions liées aux notions d'Explanandum et Explanan sont essentielles à l'utilisation d'AutoXAI, et nécessitent forcément une certaine connaissance des techniques d'explications existantes. Là encore, cela peut sembler délicat pour une utilisation par un non expert. Nous pourrions alors imaginer mieux automatiser cette phase en cherchant à **proposer une restitution XAI possible, en fonction du type de jeu de données et du modèle prédictif utilisés**. Par exemple, dans le cas de données textuelles (comme illustré dans notre deuxième cas d'usage dans les évaluations), utiliser une explication basée sur l'exemple semble naturellement plus pertinente qu'une solution basée sur l'influence de variables.

AutoXAI a été évalué sur deux cas d'usage possibles et des mesures d'évaluation d'explications objectives. En revanche il n'a pas été testé dans des situations réelles ni avec de vrais utilisateurs. Dans ce but, notre approche pose cependant deux problèmes majeurs : il faut pouvoir **évaluer subjectivement des explications et le système recommandant des explications**. Ces deux problèmes restent jusqu'à maintenant totalement ouverts. Dans le domaine du XAI, l'évaluation utilisateur reste très confidentielle, parce que réputée très difficile, et très peu de recherches se sont attaquées à ce problème même s'il est reconnu comme un enjeu important [Miller, 2019].

Le choix du modèle XAI recommandé peut également avoir un fort impact sur la compréhension du modèle prédictif ou de ses résultats. L'effet Rashomon en XAI (comme discuté dans le Chapitre 2.5) n'est pas sans poser des **questions éthiques sur les explications générées**, possiblement biaisées. Dans le cas d'une explication basée sur les exemples, la réduction du nombre de prototypes peut, par exemple, conduire à ne pas prendre en compte des sous-ensembles de données moins bien représentés. Dans le cas d'une explication attributive, un petit nombre de variables sélectionné peut aussi cacher un biais dans un modèle.

Un point particulièrement sensible, et d'ailleurs très lié à la notion d'explication biaisée,

porte sur la **robustesse des explications**. Une explication robuste admet généralement que pour deux instances similaires, leurs explications seront également similaires (l'inverse n'étant pas forcément vrai). Même si ce fait semble naturel, la littérature a déjà démontré, depuis quelques années, les problèmes de robustesse liés à l'usage de techniques comme SHAP et LIME. Ces techniques, particulièrement populaires pour de nombreux usages, peuvent s'avérer problématiques si le problème de robustesse est mal compris par les utilisateurs. Cela pourrait d'ailleurs mener à une moindre confiance de leur part quant à leur utilisation. Il semble donc important de pouvoir, a minima, avertir l'utilisateur quand un problème de robustesse est détecté et que deux instances proches possèdent des explications très différentes. Il peut aussi s'agir d'identifier le sous-ensemble de données possiblement problématique dans le jeu de données de départ.

Au delà de la propriété de robustesse, il semble aussi important de pouvoir mieux **analyser les interactions entre les données, les modèles prédictifs et les modèles XAI** mis en oeuvre. L'étude des propriétés des explications et de leurs métriques est insuffisante dans la littérature. Par exemple, des métriques peuvent-elles se contredirent pour une même propriété ? Et si oui, comment les concilier ? Comme nous avons pu le voir pour le cas d'usage du SPAM, certaines métriques (comme la représentativité) sont encore manquantes dans la littérature ou mal définies. Cela peut clairement freiner l'usage, à l'heure actuelle, du type de système de recommandation que nous proposons.

Plus généralement, le **manque de benchmark en explicabilité** est criant, mais peu étonnant au vu des problèmes particulièrement difficiles soulevés précédemment. Cela implique de travailler sur un ou plusieurs jeux de données acceptés par la communauté, d'identifier et se mettre d'accord sur toutes les propriétés et métriques indispensables pour juger de la bonne pertinence d'un modèle XAI, d'identifier les modèles prédictifs appropriés à l'utilisation de ces modèles XAI. Tout cela en étant le plus objectif possible et en évitant tout biais provenant d'un modèle prédictif et/ou XAI...

D'ailleurs, dans ce sens, il est intéressant de pouvoir **étudier les limites des modèles XAI**, en particulier attributifs (car les plus populaires). En particulier, nous pouvons citer le manque de considération des interactions entre variables pour des techniques comme SHAP (comme nous avons d'ailleurs pu le constater dans le cas d'usage du Diabète). Il semble aussi important de faire remonter les bonnes pratiques d'utilisation d'une méthode XAI par rapport à une autre. En effet, même si notre système de recommandation AutoXAI est capable de proposer un modèle XAI bien adapté, dans les limites des métriques que l'on peut lui fournir, il n'est pas capable de recommander un modèle XAI selon ses avantages et inconvénients connus. Par exemple, est-ce qu'un modèle XAI est mieux adapté à un jeu de données de plus ou moins grande taille, ou pour détecter des corrélations particulières, etc. C'est tout l'objet du chapitre suivant.

Chapitre 5

Quelques limites et préconisations à l'usage des modèles d'explication

5.1 Introduction

Dans le cas des méthodes post-hoc locales attributives, objet d'étude de ce chapitre, $SHAP$ et $LIME$ font partie des méthodes les plus populaires en explicabilité. Ces méthodes sont largement utilisées aujourd'hui dans des domaines très différents comme la finance, les assurances, la santé, le biomédical, etc. Elles ne sont pourtant pas exemptes de tout défaut, comme rappelé dans les chapitres 2.5 et 4.

En particulier, le manque de considération dans les interactions entre variables peut sembler une limite importante alors que la plupart des jeux de données à analyser présentent ce problème, même après un prétraitement adéquat. C'est ce que nous souhaitons investiguer dans la première partie de ce chapitre, en proposant des méthodes d'explication locales attributives et coalitionnelles prenant en compte les interactions entre variables.

Dans le chapitre précédent, nous avons abordé la recommandation de modèles d'explication et souligné, entre autres, le manque de propositions de la littérature pour évaluer la qualité des explications fournies. Nous pourrions ajouter aussi le besoin de quantifier le bon usage d'une méthode d'explication selon le contexte de l'utilisateur : par exemple, en fonction de la dimensionnalité d'un jeu de données ou du type de modèle prédictif.

En considérant les nouvelles méthodes coalitionnelles de ce chapitre ainsi que les méthodes SHAP et LIME, il est intéressant de pouvoir les comparer entre elles, afin d'en souligner leurs avantages et inconvénients, en fonction des jeux de données et des modèles prédictifs utilisés.

Pour cela, il est important de proposer des métriques d'intérêt objectives entre les méthodes attributives, en se basant surtout sur l'interprétation des résultats produits (et moins sur les fonctionnements internes de chacune des méthodes d'explication). En effet, nous considérons qu'il est important de pouvoir évaluer ce à quoi l'utilisateur est confronté directement : les scores d'influence des variables (que les scores soient liés au niveau d'une instance ou bien agrégés pour tout un jeu de données). Nous proposons six métriques d'intérêt pour comparer

ces méthodes incluant : le temps de calcul, les différences de scores d'influence, la répartition des scores entre variables, la robustesse entre explications, la lisibilité et la "clusterabilité".

A l'aide de ces métriques et des analyses comparatives, nous serons en mesure de proposer, en fin de chapitre, une feuille de route pour le bon usage des méthodes locales attributives.

Ce chapitre fait référence aux travaux publiés dans [Ferrettini et al., 2020a, Ferrettini et al., 2020b, Ferrettini et al., 2022, Doumard et al., 2022, Doumard et al., 2023] et regroupe une partie des travaux de thèse de Gabriel Ferrettini, Elodie Escriva et Emmanuel Doumard.

5.2 Une solution d'explication basée sur les interactions entre variables

Nous proposons dans cette section de nous intéresser aux approches d'explication locale attributive, basées sur les calculs de coalitions. Nous commençons par détailler la méthode *complète*, générant toutes les coalitions possibles entre variables (principe des valeurs de Shapley) et nous poursuivons ensuite par des propositions d'approximation, et notamment celles basées sur des calculs de groupes de variables corrélées.

5.2.1 Méthode complète (valeurs de Shapley)

Cette méthode est tout simplement la proposition du calcul complet des coalitions par valeur de Shpley [Štrumbelj and Kononenko, 2014], présenté dans le Chapitre 2.5.2.

Nous définissons ainsi l'influence *complète* d'une variable $a_i \in A$ sur la classification d'une instance x : étant donné un jeu de données décrit selon les variables de A, l'influence *complete* de la variable a_i sur la classification d'une instance x par le classifieur f sur la classe C dépend de l'influence de tous les sous-groupes possibles $A' \subseteq A$ qui ne contiennent pas a_i. Ainsi, l'influence *complète* de a_i est :

$$\mathcal{I}^C_{a_i}(x) = \sum_{A' \subseteq A \setminus a_i} p(A', A) * (inf^C_{f,(A' \cup a_i)}(x) - inf^C_{f,A'}(x)) \tag{5.1}$$

Avec $p(A', A)$ une fonction de pénalisation tenant compte de la taille du sous-ensemble A'. En effet, si une variable influence beaucoup le résultat d'un classifieur, qui est basé sur un grand groupe de variables, alors elle peut être considérée comme très influente par rapport aux autres. Dans le cas d'un petit nombre de variables, son influence serait moindre. Les valeurs de Shapley définissent la pénalisation comme suit :

$$p(A', A) = \frac{|A'|! * (|A| - |A'| - 1)!}{|A|!} \tag{5.2}$$

Cette *influence complète* d'une variable base son calcul sur son importance parmi toutes les configurations de variables possibles. Cependant, le calcul d'une influence complète pour une seule instance est extrêmement coûteux, avec une complexité de $\bigcirc(2^n * l(n, x))$, avec n le

nombre de variables, x le nombre d'instances dans le jeu de données, et $l(n, x)$ la complexité de l'apprentissage du modèle à expliquer. Il n'est donc pas pratique d'utiliser *l'influence complète* dans la plupart des tâches d'analyse de données. Par conséquent, il devient nécessaire de rechercher un moyen plus efficace d'expliquer les prédictions. Bien que *l'influence complète* soit trop lourde en termes de calcul, elle peut être considérée comme une excellente base de référence [Štrumbelj and Kononenko, 2014]. Nous pouvons donc évaluer d'autres méthodes d'explication en étudiant leurs différences avec cette méthode.

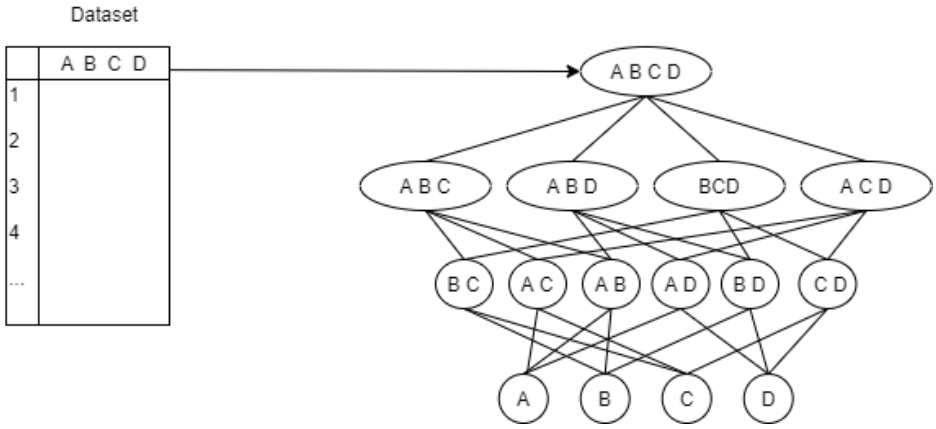


FIGURE 5.1 – Représentation des groupes calculés par la méthode *complète* pour un jeu de données comportant 4 variables. Chaque combinaison possible de variables est calculée pour garantir une valeur d'influence aussi proche que possible de la réalité.

Exemple 5.1. Comme le montre la Figure 5.1, l'influence d'une variable dépend de son influence seule, mais aussi de chaque groupe possible de variables qui la contiennent. Ainsi, pour un jeu de données comportant 4 variables A B C D, l'influence de la variable A est composée de l'influence de $\{A\}$ seule, ainsi que de l'influence des groupes $\{A, B\}$, $\{A, C\}$, $\{A, D\}$, $\{A, B, C\}$, $\{A, B, D\}$, $\{A, C, D\}$ et $\{A, B, C, D\}$.

5.2.2 Méthode K-complète

Nous proposons une approximation de la méthode *complète*. Cette approximation, qui consiste à rechercher un sous-ensemble parmi tous les sous-groupes de la méthode *complète*, pourrait être plus pratique en termes de complexité. Cette solution devrait produire une explication, a priori, plus précise que la simple considération d'indépendance entre variables individuellement (*influence linéaire*). Nous considérons alors la méthode *complète de profondeur-k* défini comme la méthode *complète*, mais en ignorant les groupes de variables A' de taille supérieure à k :

$$\mathcal{I}_{a_i}^{C_k}(x) = \sum_{A' \subseteq A \setminus a_i,\ |A'|<k} p_k(A', A) * (inf_{f,(A' \cup a_i)}^{C}(x) - inf_{f,A'}^{C}(x)) \tag{5.3}$$

$$p_k(A', A) = \frac{|A'|! * (|A| - |A'| - 1)!}{k * (|A| - 1)!} \tag{5.4}$$

En particulier, nous pouvons noter qu'une influence *linéaire* est en fait identique à l'influence par la méthode *complète de profondeur*-1. L'intuition derrière cette approche est d'éliminer les plus grands groupes, qui ont un impact moindre sur la valeur de Shapley alors que ce sont les plus coûteux à calculer.

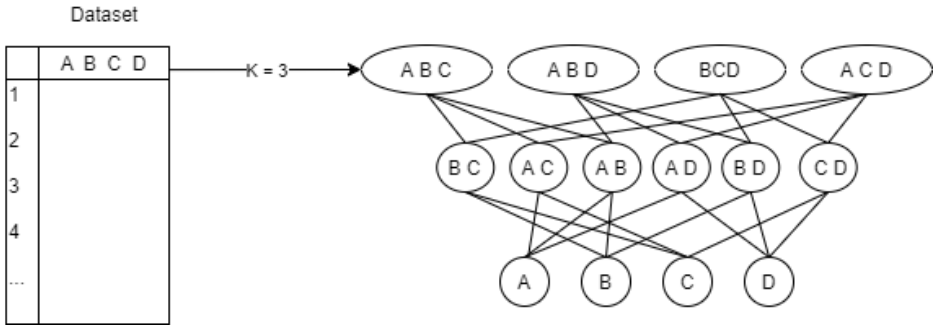


FIGURE 5.2 – Représentation des groupes calculés par la méthode *k-complète* pour un jeu de données à 4 variables. La taille des groupes est limitée par le paramètre k : ici, la taille maximale des groupes est de 3.

Exemple 5.2. Comme illustré dans la Figure 5.2, pour le même jeu de données de 4 variables et un paramètre $k = 3$, l'influence totale de la variable A dépend seulement de l'influence de A seule et des groupes de variables contenant A et d'une taille maximum de 3 : $\{A, B\}$, $\{A, C\}$, $\{A, D\}$, $\{A, B, C\}$, $\{A, B, D\}$ and $\{A, C, D\}$.

5.2.3 Limitation de ces méthodes

Les deux méthodes décrites précédemment, ainsi que la méthode SHAP exposée dans le Chapitre 2.5.2 ont plusieurs limitations.

La méthode *Complète* a une complexité exponentielle par rapport au nombre de variables, ce qui la rend inutilisable dans la plupart des cas pratique. Elle peut être approximée par les méthodes *SHAP -KernelSHAP* et ses variantes - mais au prix d'hypothèses très restrictives telles que la linéarité locale, qui ne tient pas pleinement compte des dépendances entre variables, ce qui peut biaiser les résultats de l'explication. En outre, le calcul peut prendre beaucoup de temps dans une configuration d'interdépendance élevée des variables, ce qui est souvent le cas dans la pratique. La méthode *k-complète* est une autre façon d'approximer la méthode *complète* qui permet de considérer une complexité donnée à l'aide du paramètre k.

L'inconvénient de toutes ces méthodes est que les groupes générés peuvent inclure des sous-groupes inutiles ou redondants, ce qui augmente considérablement le temps de calcul sans gain significatif en termes de précision comparé à la méthode *complète*.

5.2.4 Méthodes Coalitionnelles

Les limites des méthodes d'approximation de la méthode *complète* évoquées précédemment montrent le besoin de prendre en compte les interactions potentielles entre les variables. La combinaison de variables non liées doit être évitée au maximum pour minimiser la complexité, et donc le temps de calcul, tout en conservant une grande précision par rapport à la méthode *complète*. À cette fin, et parce qu'il existe une multitude de possibilités pour prendre en compte ces interactions, nous proposons plusieurs méthodes telles que celles basées sur le *Coefficient de corrélation de Spearman*, le *Facteur d'inflation de la variance (VIF)* ainsi que sur une *Analyse en composantes principales (ACP)* ou encore sur une solution basée sur les interactions entre les variables et le modèle prédictif [Henelius et al., 2014].

Nous développons également des méthodes *Inverse* - basées sur Spearman ou VIF - qui ne rassemblent que les variables non corrélées, puisque les groupes formés uniquement de variables fortement corrélées contiennent principalement des informations redondantes. Pour chaque algorithme, un paramètre contrôle la taille des sous-groupes générés. Une valeur plus élevée de ce paramètre génère des groupes plus grands, tandis qu'une valeur plus faible produit des groupes plus petits, donc moins complexes. Les explications par influence pour chaque variable d'un jeu de données sont ensuite calculées à l'aide d'une influence *coalitionelle*, prenant comme paramètre la liste des groupes générés par la méthode de regroupement.

Pour des raisons de synthèse, nous ne détaillons que la méthode basée sur Spearman, car donnant les meilleurs résultats (meilleur ratio entre le temps de calcul et la qualité des explications produites) comparés aux autres méthodes coalitionnelles. Nous ne présentons pas non plus les évaluations entre ces méthodes par souci de clarté dans ce chapitre. Le détail de toutes les méthodes et leurs évaluations sont cependant disponibles dans l'article [Ferrettini et al., 2022].

Coalitions basées sur la corrélation de Spearman

Cette méthode basée sur le coefficient de corrélation de Spearman prend en compte les corrélations non linéaires et le calcul de la corrélation entre les variables doit se faire par paires. Ainsi, la méthode consiste à générer la matrice de toutes les corrélations de chaque paire et à décider ensuite quelles variables font partie d'un groupe. Pour cette méthode, nous pouvons soit privilégier le calcul des variables fortement corrélées, soit au contraire privilégier les groupes de variables non corrélées. Ces deux approches sont appelées respectivement *coalition de Spearman* et *coalition de Spearman inversée*.

Pour un jeu de données $D = (A, X)$, avec $A = \{a_1, ..., a_n\}$ la matrice de corrélations C est obtenue en calculant le coefficient de corrélation de Spearman de chaque paire de variables : $C(1, 2) = corr(a_1, a_2)$. Ainsi C est symétrique et les valeurs de la diagonale sont à 1. Pour chaque ligne i de la matrice C, nous considérons comme groupées avec a_i les variables fortement corrélées (ou peu) avec a_i, pour la *Coalition de Spearman* (ou la *coalition de Spearman inversée*).

L'algorithme 3 détaille la méthode de *coalition de Spearman*. La méthode de *coalition de*

Algorithm 3 Extraction des coalitions basées sur Spearman.

Require: a threshold t, the set of variables of the dataset A, and a function $spearman(A)$ calculating the matrix of all the absolute Spearman correlation coefficient of all the subsets of a set of variables. a max and min functions which returns the maximum and minimum of a matrix line.

Ensure: σ a coalition of variables

```
σ ← {}
corrmat ← spearman(A)                    ▷ calculating the correlation matrix
for all a ∈ A do
    g ← {}
    for all a ∈ A do
        if corrmat(a, a′) > max(corrmat(a)) ∗ (1 − t) and max(corrmat(a)) > 0.1 then
            ▷ If the most correlated variable has a coefficient less than 0.1, we consider a as
a singleton
            add a′ to g
        end if
    end for
    add g to σ
end for
return σ
```

Spearman inversée peut être obtenue en remplaçant la condition d'ajout d'une variable à un groupe par $corrmat(a, a') < min(corrmat(a)) + max(corrmat(a)) * t$ et $min(corrmat(a)) < 0.5$. Ceci permet d'ajouter au groupe les variables les moins corrélées jusqu'à un seuil : si la variable la moins corrélée à a a sa corrélation de Spearman supérieure à 0.5, nous considérons la variable a comme un singleton.

Exemple 5.3. Étant donné notre jeu de données précédent de 4 variables, nous calculons la matrice des coefficients de corrélation de Spearman comme indiqué dans la Figure 5.3. Dans cette matrice, nous itérons sur chaque ligne de la matrice afin de créer des groupes basés sur les variables les plus corrélées. Dans la première ligne, nous voyons que la variable la plus corrélée à A est B, et que les deux autres variables sont très faiblement corrélées à A. Nous avons donc un premier groupe : $\{A, B\}$. La deuxième ligne nous indique que A et C sont tous deux fortement corrélés à B. Nous avons donc un deuxième groupe : $\{A, B, C\}$. De même, la troisième ligne indique que B et D sont corrélées à C, et nous ajoutons donc un troisième groupe : $\{B, C, D\}$. Enfin, en examinant la dernière ligne, nous apprenons que seul C est fortement corrélé à D et notre dernier groupe est donc $\{C, D\}$. Comme les deux groupes de cardinalité 2 sont contenus dans les deux groupes de cardinalité 3, nous avons nos coalitions finales : $\{\{A, B, C\}, \{B, C, D\}\}$. Avec cette coalition, l'influence complète de la variable A est composée de $\{A\}$, $\{A, B\}$, $\{A, C\}$ et $\{A, B, C\}$. B est composé de $\{B\}$, $\{B, D\}$, $\{A, B\}$, $\{B, C\}$, $\{A, B, C\}$ et $\{B, C, D\}$. Enfin, l'influence complète de D est composée de $\{D\}$, $\{C, D\}$,

$\{B, D\}$ et $\{B, C, D\}$.

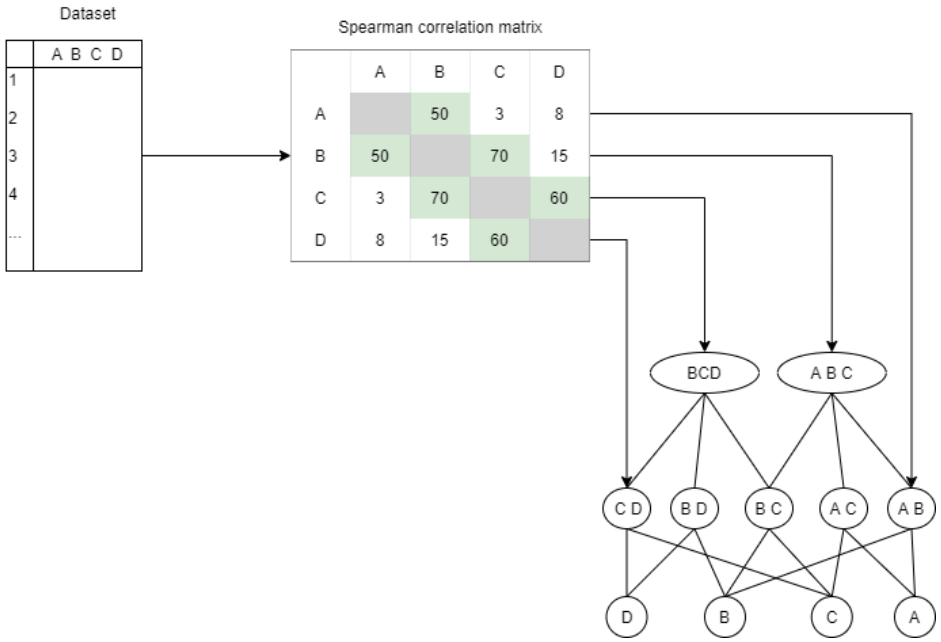


FIGURE 5.3 – Représentation des groupes calculés par la méthode de coalition basée sur la méthode de Spearman pour un jeu de données à 4 variables. La matrice de corrélation de Spearman est calculée. Pour chaque ligne, les variables les plus corrélées avec la variable courante de la ligne sont considérées comme faisant partie d'un groupe.

L'implémentation complète de notre proposition est disponible ici [1].

Autres approches coalitionnelles possibles

Les autres approches coalitionnelles proposées sont basées sur :

— le Facteur d'Inflation de la Variance (VIF) : afin d'être plus exhaustifs dans les calculs de corrélations, nous avons fait le choix de proposer une méthode supportant la multicolinéarité, pour une variable cible donnée. L'idée est de calculer chaque VIF pour chaque variable du jeu de données considéré. Ensuite un nouveau VIF est calculé en supprimant l'une des variables. Si la différence de scores est significative, alors la variable est déterminante pour le calcul et est considérée dans le groupe formant les coalitions finales.

— l'Analyse en Composante Principale (ACP) : il s'agit de bénéficier des calculs de réduction des dimensions du jeu de données. Pour ce faire, les différentes variables sont combinées linéairement, avec pour résultat un nouvel ensemble de variables (chaque

1. https://github.com/kaduceo/coalitional_explanation_methods

nouvelle variable étant une combinaison linéaire des précédents). Notre raisonnement, pour cette approche, est de considérer l'ensemble des variables combinées (résumées par la nouvelle variable de l'ACP) comme un groupe d'influences.

— Le modèle : les groupes de variables sont créés en utilisant le modèle prédictif lui-même pour détecter les interactions. Dans cette approche, aucune corrélation n'est détectée, mais seulement une interaction au sens de l'utilisation des variables par le modèle. Pour ce faire, on randomise certaines valeurs du jeu de données et on étudie l'évolution des prédictions du modèle. Cela nous permet de déterminer les groupes de variables changeant le moins les prédictions sur le jeu de données.

5.3 Métriques d'intérêt pour la comparaison des méthodes locales attributives

Afin d'évaluer l'intérêt des méthodes d'explication et les comparer sur un grand nombre de jeux de données, nous proposons six métriques différentes qui ne prennent en compte que les valeurs d'influence produites par la méthode. Dans toutes les définitions suivantes, supposons que X est un jeu de données avec n instances, d le nombre de variables et f une méthode d'explication qui peut être appliquée à chaque instance du jeu de données en fonction d'un modèle d'apprentissage automatique.

Définition 5.1. *Temps moyen de calcul.* La première mesure est le temps de calcul moyen par instance, c'est-à-dire le temps nécessaire à une méthode d'explication donnée pour calculer les influences locales d'un jeu de données, divisé par le nombre d'instances du jeu de données.

Définition 5.2. *Erreur comparée à la méthode complète.* La seconde mesure est une quantification de l'écart moyen entre l'influence donnée par une méthode et la méthode *Complète*, considérée comme une base de référence puisque calculant toutes les coalitions possibles de variables (voir la section 5.2.1).

Soit $f_k(x)$ l'influence d'une variable k produite par une méthode d'explication f pour une instance donnée x, et un modèle d'apprentissage automatique donné, et $f_k^C(x)$ l'influence donnée par la méthode *Complète* pour le même modèle, la même variable et la même instance. Nous définissons l'erreur moyenne de la méthode d'explication comme suit :

$$err(f, X) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{p} \sum_{k=1}^{d} \left| f_k(X_i) - f_k^C(X_i) \right| \tag{5.5}$$

Définition 5.3. *Aire sous la courbe d'importance des variables cumulées (AUC).* La troisième métrique s'inspire du principe de complexité effective défini dans [Nguyen and Martínez, 2020]. Elle évalue la concision d'une explication en fonction de la distribution de l'importance des variables. L'importance des variables (valeur absolue des moyennes de l'influence attribuée aux instances pour une variable donnée) est classée par

ordre décroissant, puis la somme cumulative est calculée. Par exemple, dans un jeu de données comportant deux variables, si une méthode accorde 80% d'importance à la variable la plus importante (et donc 20% à la seconde), elle aura un vecteur de proportion d'importance cumulée de $[0, 0.8, 1]$. Nous pouvons alors définir l'aire sous la courbe (AUC) normalisée comme suit :

Soit $C \in [0; 1]^{d+1}$ le vecteur de proportion d'importance cumulée donnée par une méthode d'explication sur un jeu de données, avec C_i la proportion d'importance totale prise par les i variables les plus importantes. Nous définissons l'aire sous la courbe d'importance cumulative des caractéristiques comme suit :

$$AUC(X) = \frac{1}{d} \sum_{i=0}^{d-1} \frac{C_i + C_{i+1}}{2} \tag{5.6}$$

Cette métrique indique si une méthode d'explication favorise les scores d'influence de grande importance à seulement quelques variables ou, au contraire, une répartition plus homogène entre un plus grand nombre de variables. Comme cette somme cumulative est triée par valeur décroissante, cette valeur est comprise entre 0.5 et 1. Une valeur de 0.5 signifie que la méthode d'explication accorde la même importance à toutes les variables, tandis qu'une valeur de 1 signifie que la méthode d'explication n'accorde des influences non nulles qu'à une seule variable, expliquant ainsi les prédictions du modèle par une seule variable.

Définition 5.4. *Robustesse (estimation locale de Lipschitz).* La quatrième mesure concerne la robustesse des méthodes d'explication. Une méthode est robuste si des instances similaires conduisent à des explications similaires. Formalisée dans [Alvarez-Melis and Jaakkola, 2018], nous utilisons la version discrète de l'estimation locale de Lipschitz.

Soit $\mathcal{N}_\epsilon = \{x_j \in X | \|x_i - x_j\| \leq \epsilon\}$ le voisinage ϵ d'une instance x_i.

$$\tilde{L}_X(x_i) = \max_{x_j \in \mathcal{N}_\epsilon(x_i) \leq \epsilon} \frac{\|f(x_i) - f(x_j)\|_2}{\|x_i - x_j\|} \tag{5.7}$$

Une valeur élevée de $\tilde{L}_X(x_i)$ signifie que la méthode d'explication n'est pas robuste pour l'instance x_i sur le jeu de données X, et une valeur faible signifie que l'explication est robuste pour l'instance x_i sur le jeu de données X. Nous faisons la moyenne de cette valeur sur toutes les instances d'un jeu de données afin d'obtenir la valeur de la métrique d'une méthode pour un jeu de données entier.

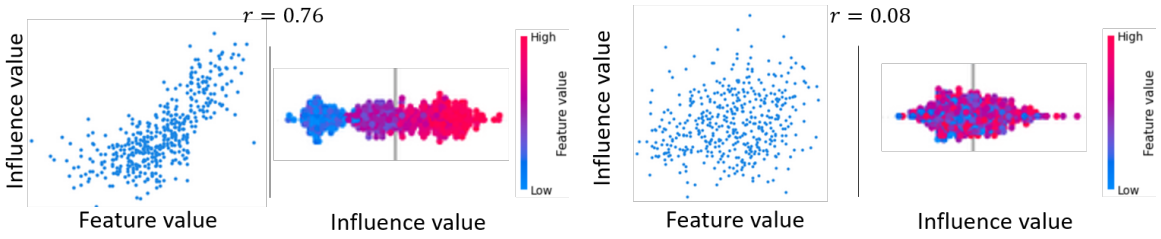
Définition 5.5. *Lisibilité.* La cinquième mesure est une mesure de la lisibilité de l'explication globale. Elle s'inspire de la métrique de monotonicité définie dans [Nguyen and Martínez, 2020], et nous l'adaptons afin d'analyser la corrélation entre les valeurs des données et les influences pour une variable. Même si les explications sont calculées pour chaque instance, nous voulons que ces explications aient un sens lorsque nous les comparons les unes aux autres, globalement. Pour évaluer cela, nous examinons la relation entre

la valeur d'une variable et la valeur de l'explication de cette même variable, pour toutes les instances, en utilisant la corrélation de Spearman r.

Soit $X_i \in \mathbb{R}^n$ la variable i d'un jeu de données, $f(X_i) \in \mathbb{R}^n$ l'explication de chaque instance pour la variable i en question et $r(X, Y)$ le coefficient de corrélation de Spearman de deux vecteurs de même taille. Nous définissons la lisibilité d'une méthode d'explication sur un jeu de données X comme suit :

$$\mathcal{R}(X) = \frac{1}{d} \sum_{i=1}^{d} |r(X_i, f(X_i))| \tag{5.8}$$

Dans la Figure 5.4, nous montrons un exemple visuel de ce que nous considérons comme lisible ou illisible, selon notre définition.



(a) (b)

FIGURE 5.4 – (a) Exemple d'explication lisible. Chaque point correspond à une instance. A droite (représentation compacte), la couleur représente la valeur de la variable. (b) Exemple d'explication illisible.

Définition 5.6. *Clusterabilité.* La sixième et dernière métrique mesure l'interaction des variables par paire, en fonction des explications. Pour ce faire, pour chaque paire de variables au sein d'une explication globale, nous utilisons une méthode de clustering pour créer une partition de l'explication entre toutes les instances pour la paire de variables, puis nous évaluons la qualité du clustering ainsi créé. Nous calculons la moyenne de cette valeur pour toutes les paires de caractéristiques et appelons cette mesure la "clusterabilité" (bidimensionnelle) :

Soit $f(X_i) \in \mathbb{R}^n$ l'explication de chaque instance d'une variable i, K une fonction de clustering, et S une fonction d'évaluation du clustering. Nous définissons la *clusterabilité* comme suit :

$$Cl(X) = \frac{2}{d * (d-1)} \sum_{\substack{i,j \in [1,\dots,d] \\ i \neq j}} S(K(f(X_i), f(X_j))) \tag{5.9}$$

Un score de clusterabilité élevé signifie que la méthode d'explication établit des relations entre les paires de variables, par leur contribution conjointe aux prédictions. Pour nos expérimentations, nous utilisons l'algorithme K-Means comme méthode de clustering et le score de Silhouette comme mesure de la qualité du clustering.

5.4 Comparaison des méthodes

Nous proposons dans cette section de comparer les quatre méthodes locales attributives incluant les deux méthodes populaires de la littérature, à savoir $SHAP$ et $LIME$, ainsi que les deux méthodes coalitionnelles discutées dans ce chapitre : la méthode *complète* (valeurs de Shapley) et la méthode basée sur *Spearman*.

5.4.1 Protocole

Toutes les expériences sont réalisées sur un processeur Intel Xeon Gold 6230 avec 125 Go de RAM en utilisant Python 3.9.7. Toutes les exécutions sont effectuées sur un seul coeur de CPU pour des raisons d'optimisation et de facilité de reproductibilité. Pour comparer les méthodes d'explication, nous les appliquons à un large éventail de 304 jeux de données disponibles sur OpenML. En raison des contraintes de complexité de calcul des méthodes d'explication, nous n'avons pris en compte que les jeux de données comportant au maximum 13 variables et au maximum 10 000 instances. Nous n'avons également pris en compte que les tâches de classification afin d'utiliser des modèles prédictifs et des mesures comparables.

Étant donné qu'une méthode d'explication nécessite l'application d'un modèle, nous choisissons quatre types de modèles prédictifs largement utilisés pour la classification : Régression logistique (LR), Machines à vecteurs de support (SVM), Forêts aléatoires (RF) et Machines à gradient boosté (GBM). Pour les trois premiers, nous utilisons l'implémentation de la bibliothèque Python *scikit-learn* version 1.0.1. Pour les GBM, nous utilisons la bibliothèque Python *XGBoost* version 1.5. Nous utilisons des valeurs par défaut pour les hyperparamètres des modèles. Pour les méthodes d'explication, nous utilisons les bibliothèques Python *shap* 0.40, *lime* 0.2.0.1 ainsi que la méthode coalitionnelle, basée sur Spearman, détaillée en Section 5.2.4.

Précisions sur les expérimentations

Les expérimentations sont présentées en deux parties. La Section 5.4.2 compare les quatre méthodes attributives. La méthode *complète* sert de référence pour la métrique d'intérêt sur l'écart moyen entre influences. La méthode *Spearman* est utilisée avec un seuil de 25% de toutes les coalitions de variables (voir [Ferrettini et al., 2022]).

En ce qui concerne *SHAP*, nous utilisons la méthode agnostique *KernelSHAP* sur tous les jeux de données. Comme cette méthode est très lente à exécuter si nous considérons tout le jeu des données comme échantillon possible pour les permutations, nous choisissons de suivre la recommandation de *SHAP*. La documentation de *KernelSHAP* propose [2], pour accélérer le temps de calcul, d'effectuer un clustering *K-Means* sur le jeu des données d'entrée, puis en prenant les centroïdes comme échantillons possibles. Nous choisissons $K = 10$ clusters pour chaque jeu de données.

2. https://shap-lrjball.readthedocs.io/en/latest/generated/shap.KernelExplainer.html

En outre, pour les deux modèles prédictifs à base d'arbres XGBoost et RF, nous utilisons l'explainer spécifique au modèle *TreeSHAP* au moyen de deux implémentations. La première détermine les valeurs de *SHAP* avec des échantillons d'arrière-plan, de manière similaire à *KernelSHAP* mais optimisée pour les méthodes basées sur les arbres. Nous utilisons le jeu des données complet comme échantillon pour cette méthode. La deuxième méthode approxime les valeurs de *SHAP* en tenant compte des structures des arbres et ne nécessite pas d'échantillonnage, c'est pourquoi nous l'appelons *TreeSHAPapprox*.

Enfin, nous considérons *LIME*, qui nécessite la création d'un certain nombre d'échantillons perturbés pour expliquer chaque instance. Nous avons choisi de fixer ce nombre à 100 échantillons pour tous les jeux de données.

Avec une méthodologie similaire, la Section 5.4.3 identifie l'impact du modèle prédictif sur des méthodes d'explication spécifiques.

L'ensemble des expérimentations réalisées dans ce chapitre sont disponibles sur Github [3].

5.4.2 Comparaison des méthodes attributives

Temps de calcul

Nous montrons dans la Figure 5.5 l'évolution du temps de calcul moyen de chaque méthode pour chaque modèle prédictif, sur les jeux de données partageant le même nombre de variables.

LIME, dont la complexité est linéaire en fonction du nombre de variables, est très coûteuse par rapport à d'autres méthodes en basse dimension (peu de variables), mais moins coûteuse que les méthodes basées sur les coalitions et *KernelSHAP*, en haute dimension. *LIME* semble également présenter une très faible variabilité de temps de calcul entre les jeux de données, ce qui se traduit par des barres d'erreur plus petites sur le graphique.

Les méthodes coalitionnelles présentent une complexité exponentielle avec le nombre de variables, avec un temps d'exécution élevé en haute dimension, mais un temps d'exécution similaire à celui des autres méthodes en basse dimension. Le temps d'exécution de la méthode *Spearman* semble naturellement corrélé au temps d'exécution de la méthode *Complete*, prenant une fraction du temps (environ 25%) de la méthode *Complete*.

KernelSHAP, malgré son optimisation liée à l'échantillonnage, a un temps d'exécution élevé en haute dimension, comparable aux méthodes basées sur les coalitions, pour SVM et la régression logistique. Pour les modèles prédictifs basés sur les arbres, *KernelSHAP* est plus lent en basse dimension, mais plus rapide en haute dimension que les méthodes basées sur la coalition. Enfin, les explainers basés sur les arbres semblent avoir un temps d'exécution constant par instance, quel que soit le nombre de variables, et la version de *TreeSHAP* par approximation a le temps d'exécution le plus faible.

3. https://github.com/EmmanuelDoumard/local_explanation_comparative_study

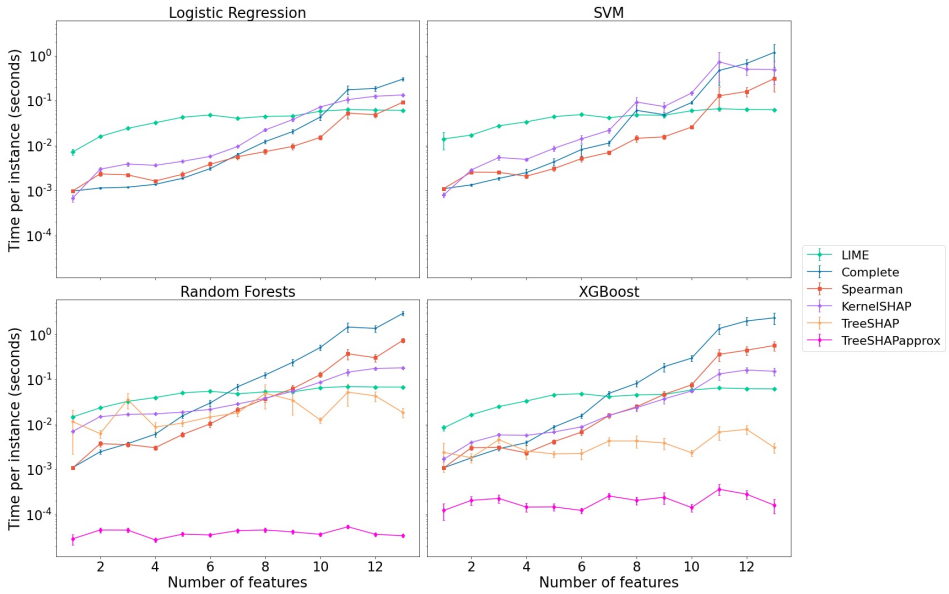


FIGURE 5.5 – Temps d'exécution moyen de chaque méthode par instance et par nombre de variables, pour chaque modèle

Erreur

En ce qui concerne l'erreur, la Figure 5.6 montre la différence absolue moyenne d'influence entre chaque méthode et la méthode *complète* (base de référence). Tout d'abord, nous pouvons constater que, dans l'ensemble, plus un jeu de données comporte de variables, plus les influences sont proches de la méthode *Complete*.

Cela est probablement dû au fait qu'en général, plus il y a de variables, moins l'amplitude de l'influence de chaque caractéristique individuelle sur la prédiction est importante. Nous constatons également que, quel que soit le modèle prédictif, les méthodes sont positionnées de la même manière. En basse dimension (moins de 6 variables), *KernelSHAP* est la plus proche de la méthode *Complète*, suivie par *Spearman*, tandis que *LIME* est la plus éloignée. Dans les dimensions supérieures, *Spearman* devient plus précis que *KernelSHAP*. *TreeSHAP* (par approximation ou non) est plus précis que *KernelSHAP*, mais toujours moins précis que *Spearman* dans les dimensions élevées. Notez que la version par approximation de *TreeSHAP* n'apparaît pas sur le graphique pour XGBoost car sa mise en œuvre oblige ses valeurs *SHAP* à être exprimées en log-odds plutôt qu'en probabilités, ce qui rend impossible toute comparaison avec d'autres méthodes.

AUC

La Figure 5.7a montre la moyenne de la proportion d'importance cumulée des variables les plus importantes pour les 37 jeux de données comportant 10 variables. De cette manière,

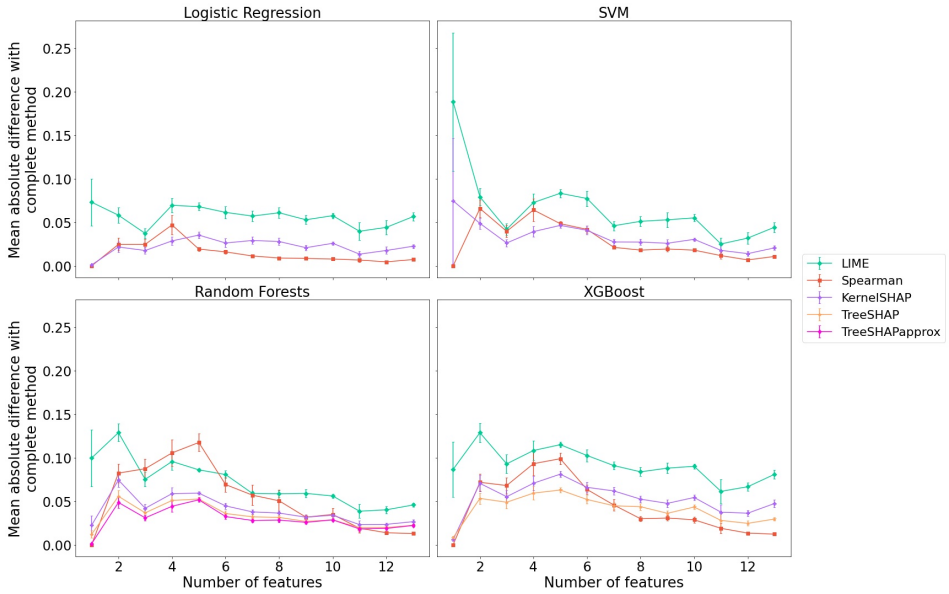
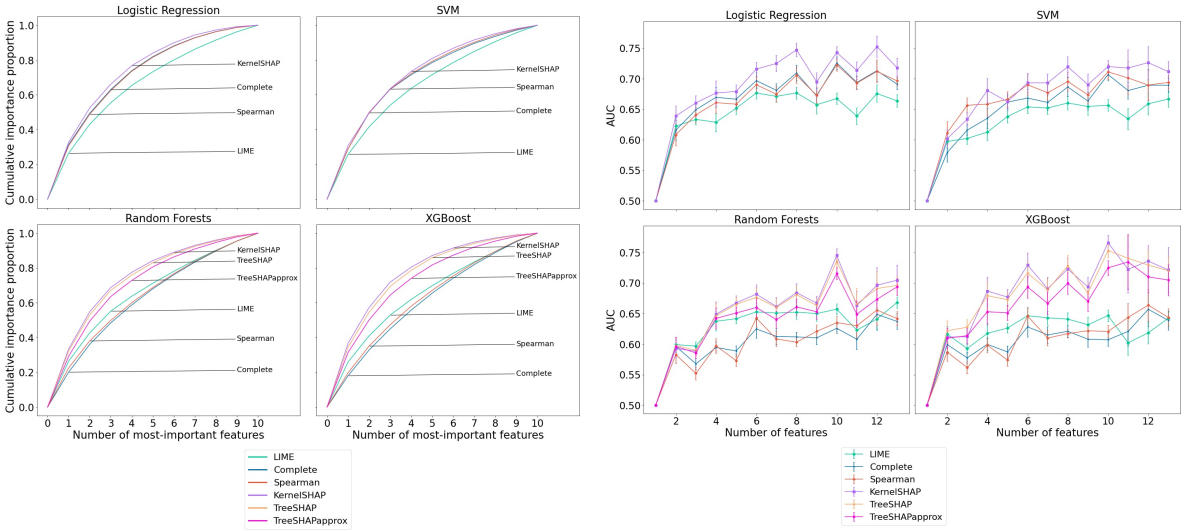


FIGURE 5.6 – Différence absolue moyenne de chaque méthode par rapport à la méthode *complète*, moyenne calculée en fonction du nombre de variables, pour chaque modèle.

pour chaque modèle prédictif et pour chaque méthode, nous obtenons une courbe à partir de laquelle nous calculons notre troisième métrique : l'AUC de la courbe.

Nous voyons sur la figure que certaines méthodes présentent des courbes plus raides que d'autres. Par exemple, avec la régression logistique et SVM, *LIME* accorde une proportion moindre à l'importance totale des premières caractéristiques les plus importantes, par rapport aux méthodes basées sur les coalitions et à *SHAP*. Pour les modèles par arbre, nous constatons que *SHAP*, quelle que soit la méthode, accorde beaucoup plus d'importance aux premières variables les plus importantes que les autres méthodes.

Conformément à la méthode de calcul de l'AUC illustrée dans la Figure 5.7a, nous représentons les valeurs moyennes de l'AUC pour les jeux de données de 2 à 13 variables pour chaque modèle prédictif et méthode d'explication dans la Figure 5.7b. Pour tous les modèles, nous pouvons constater que les méthodes basées sur *SHAP* ont tendance à produire des influences avec un AUC plus élevé que les autres méthodes. Cela signifie que les méthodes *SHAP* ont tendance à attribuer la majeure partie de l'importance des influences à un nombre réduit de variables les plus importantes, tandis que les autres méthodes ont tendance à répartir l'importance des variables de manière plus uniforme sur l'ensemble de toutes les variables. Les deux méthodes coalitionnelles semblent générer des AUC similaires. Enfin, *LIME* tend à produire des influences avec des AUC plus faibles que les autres méthodes pour les modèles SVM et de régression logistique, alors qu'elle produit des AUC plus proches des méthodes coalitionnelles pour les modèles basés arbre.



(a) (b)

FIGURE 5.7 – (a) Proportion d'importance cumulée des variables les plus importantes par méthode, pour chaque modèle. Seules les influences calculées sur des jeux de données comportant 10 variables sont indiquées. (b) AUC de chaque méthode, moyennée en fonction du nombre de variables, pour chaque modèle.

Robustesse

En ce qui concerne la robustesse, nous montrons dans la Figure 5.8 les estimations locales de Lipschitz (robustesse), pour chaque modèle prédictif, regroupées par méthode. Nous avons utilisé la formule 5.7 avec $\epsilon = 0,3$. Dans l'ensemble, la méthode d'explication n'a pas tellement d'impact sur la robustesse, sauf pour LIME avec les modèles de régression logistique et SVM, pour lesquels la méthode est beaucoup moins robuste. Nous pouvons également constater que la méthode *Spearman* est légèrement moins robuste que les méthodes *complète* et *SHAP*.

Lisibilité

La Figure 5.9, de la même manière, représente la lisibilité de chaque modèle, regroupée par méthode. La méthode d'explication n'a pas non plus beaucoup d'impact sur la lisibilité. Les méthodes *complète* et *Spearman* ont une lisibilité légèrement inférieure aux autres. Cela signifie que le lien entre une variable et ses explications tend à être moins évident avec ces méthodes qu'avec les autres. Cela est peut-être dû à la nature coalitionnelle de ces méthodes : en se concentrant sur les coalitions, ces méthodes sont souvent en mesure de saisir des interactions complexes entre plusieurs variables, ce qui signifie que la contribution marginale d'une variable est trop complexe pour être expliquée uniquement par la valeur de cette même variable.

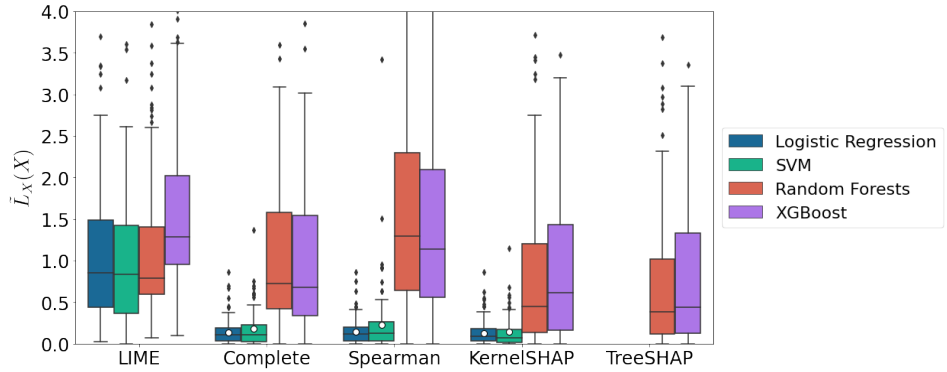


FIGURE 5.8 – Estimation locale de Lipschitz pour chaque modèle, regroupée par méthode. Chaque boîte représente les résultats agrégés pour tous les jeux de données. Le point blanc représente la valeur moyenne. En raison de valeurs aberrantes, nous avons repositionné le graphique à $\tilde{L}_X(X) = 4$

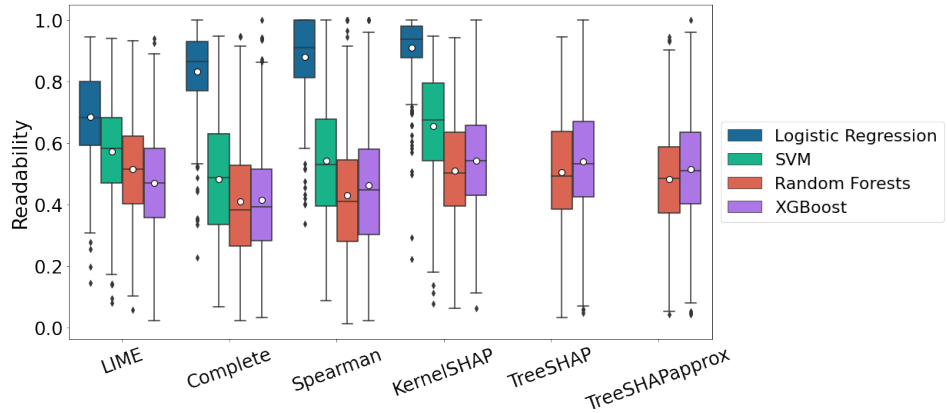


FIGURE 5.9 – Lisibilité de chaque modèle, regroupée par méthode. Chaque boîte représente les résultats agrégés pour tous les jeux de données. Le point blanc représente la valeur moyenne.

Clusterabilité

Enfin, nous montrons dans la Figure 5.10 la clusterabilité bidimensionnelle des méthodes appliquées à chaque modèle prédictif. Nous pouvons voir que *LIME* a une clusterabilité significativement plus faible que les autres méthodes, qui ont elles-mêmes une clusterabilité similaire entre elles. Cela signifie que *LIME* a tendance à capturer moins d'interactions entre les paires de variables par groupes d'instances. Cela peut être dû à la discrétisation imposée par *LIME* sur chaque variable indépendamment des autres.

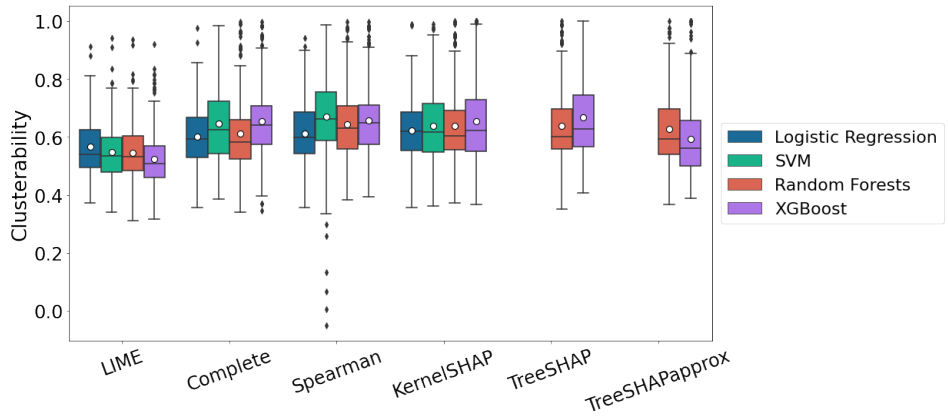


FIGURE 5.10 – Capacité de clusterabilité pour chaque modèle, groupé par méthode. Chaque boîte représente les résultats agrégés pour tous les jeux de données. Le point blanc représente la valeur moyenne.

5.4.3 Impact des modèles prédictifs sur les explications

Temps de calcul

Nous montrons dans la Figure 5.11 le temps de calcul par instance nécessaire pour calculer les explications de chaque modèle prédictif, pour chaque méthode d'explication.

Nous constatons que le temps d'exécution de *LIME* ne varie quasiment pas entre les modèles : le temps de calcul par instance est le même, quel que soit le modèle. Pour les autres méthodes, le classement des performances de calcul de la méthode d'explication en fonction du modèle est à peu près le même, du plus lent au plus rapide : Random Forests, XGBoost, SVM et Régression logistique. Le SVM présente une variabilité globalement plus élevée, avec des courbes plus raides et des barres d'erreur plus élevées. Le SVM présente même des résultats aberrants lorsqu'il est appliqué à *KernelSHAP* dans des dimensions plus élevées. Dans l'ensemble, nous n'observons pas de comportement spécifique du temps de calcul de la méthode par rapport au modèle utilisé, sauf pour *TreeSHAPapprox* où les Random Forests sont plus rapides à calculer. Cela peut être lié au fait que *TreeSHAPapprox* ne prend en compte que les structures arborescentes, plus simples avec Random Forests qu'avec XGBoost.

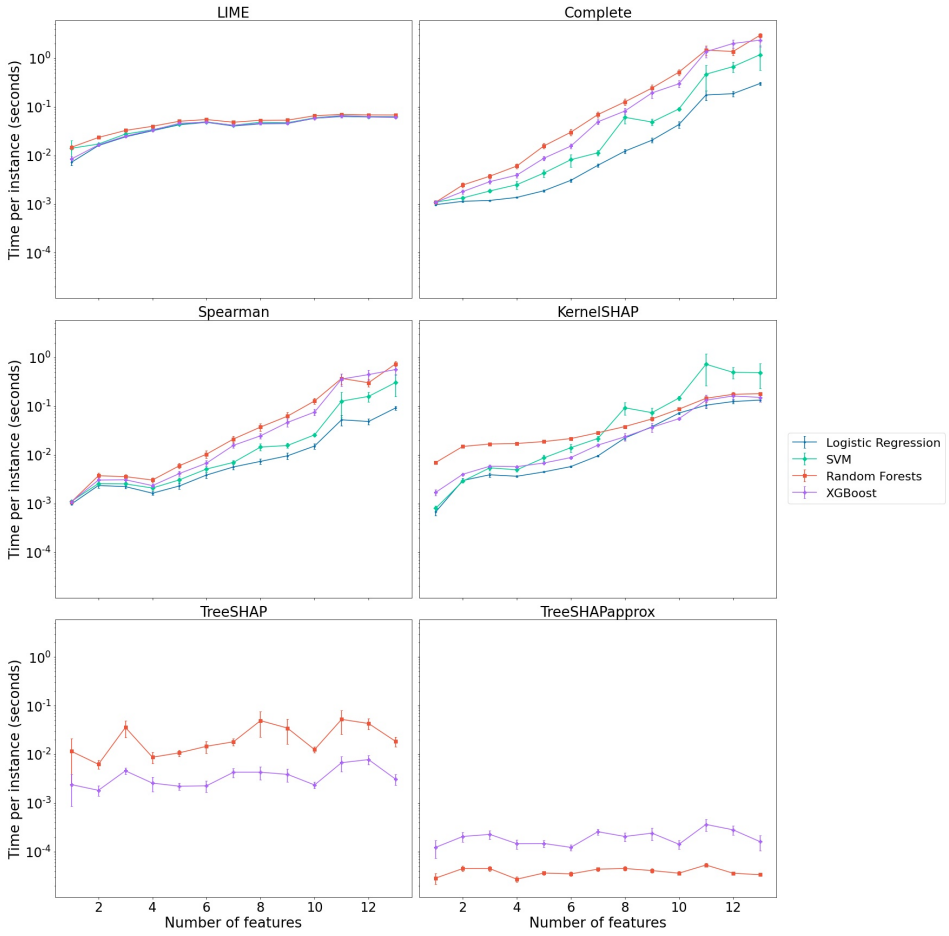


FIGURE 5.11 – Temps d'exécution de chaque modèle par instance, en moyenne par nombre de variables, pour chaque méthode d'explication

En général, plus un modèle prédictif est rapide à obtenir et à prédire des valeurs et plus il est simple, plus les explications sont rapides à calculer, quelle que soit la méthode.

Erreur

Nous présentons dans la Figure 5.12 l'erreur de chaque méthode d'explication pour chaque modèle. La figure ne présente pas les résultats pour *TreeSHAPapprox* car le seul modèle pertinent pour cette méthode est Random Forests et il n'y a donc pas d'autre modèle avec lequel comparer ses résultats.

Pour les trois méthodes agnostiques (*LIME*, *KernelSHAP* et *Spearman*), les modèles de régression logistique et SVM proposent les explications les plus précises. Nous pouvons constater que les explications basées sur la régression logistique sont généralement plus précises que celles des SVM, en particulier pour les faibles dimensions. Les explications de XGBoost sont moins précises que celles de Random Forest, sauf pour la méthode *Spearman* (des résultats similaires sont observés). Dans l'ensemble, il semble que plus le modèle est simple, plus il est précis vis-à-vis de la méthode *complète*.

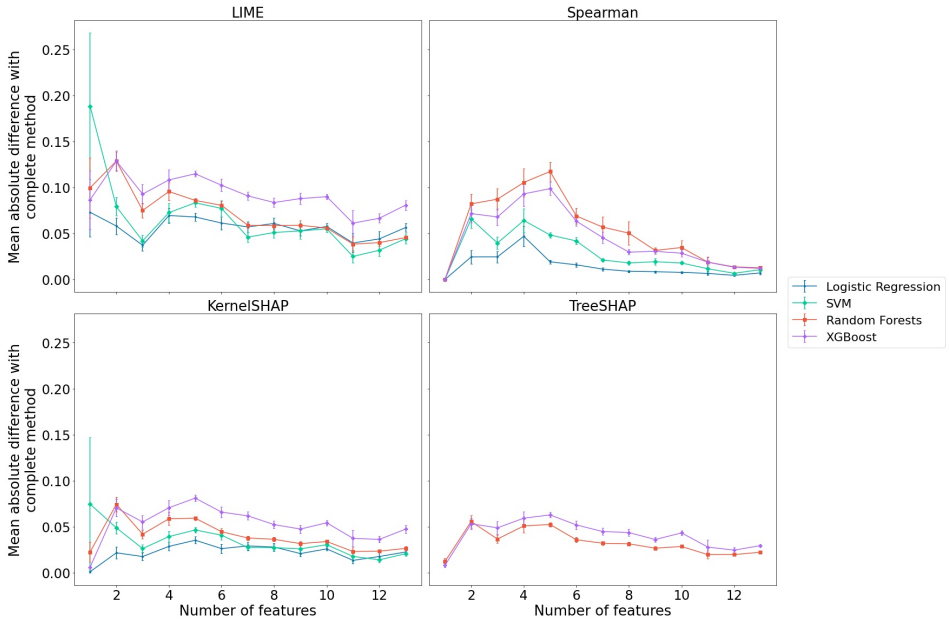


FIGURE 5.12 – Différence absolue moyenne de chaque méthode d'explication par rapport à la méthode *complète*, moyenne calculée en fonction du nombre de variables, pour chaque modèle.

AUC

En ce qui concerne l'AUC, nous présentons tous les résultats dans la Figure 5.13. Nous observons que pour *LIME* et *KernelSHAP*, il n'y a pas de différence significative entre les AUC

des différents modèles. Cependant, pour les méthodes coalitionnelles, nous pouvons observer une séparation claire entre les modèles basés ou non sur les arbres : ces dernières ont une AUC plus élevée que les autres. Cela signifie que, lors de l'utilisation de méthodes coalitionnelles, il faut être conscient que les différents modèles peuvent produire des distributions d'importances différentes sur les variables. En ce qui concerne les méthodes basées sur les arbres, nous pouvons constater que XGBoost génère des explications avec des AUC légèrement plus élevées que Random Forests, en moyenne.

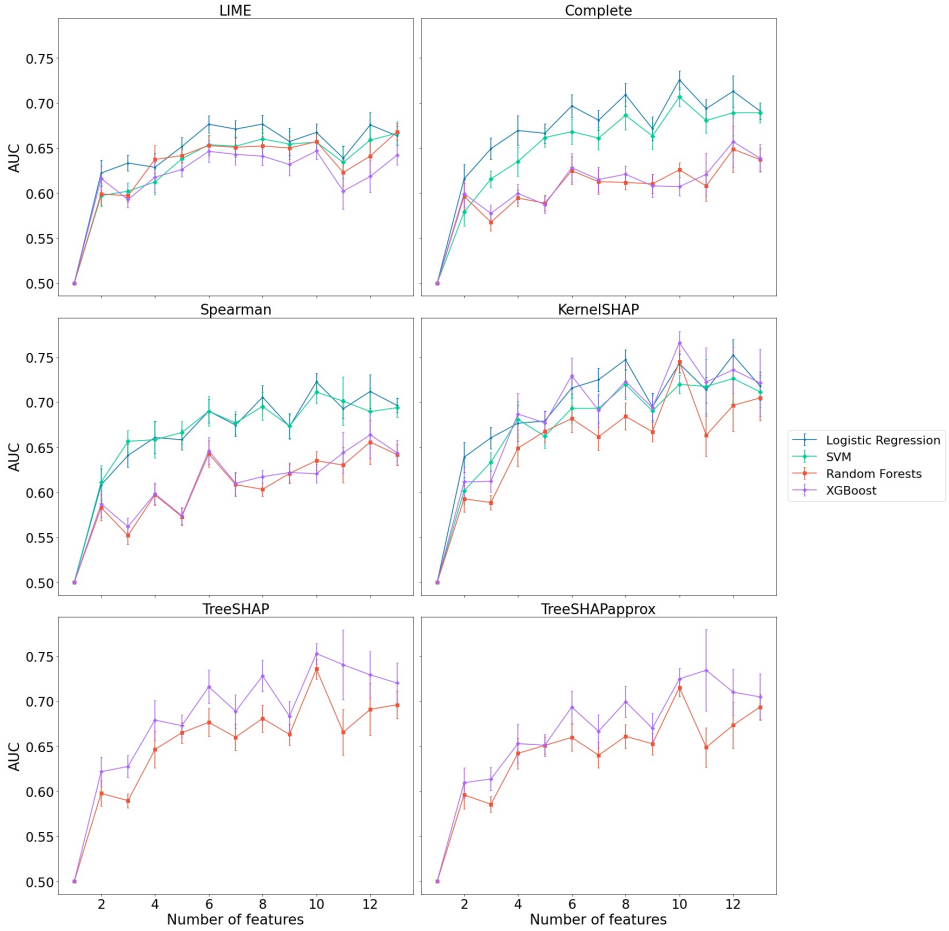


FIGURE 5.13 – AUC de chaque modèle, moyenné par le nombre de variables, pour chaque méthode d'explication

Robustesse

Concernant la robustesse (ainsi que pour la lisibilité et la clusterabilité), nous utilisons les mêmes graphiques que la section précédente pour analyser l'impact du modèle prédictif sur

les explications.

La Figure 5.8 confirme que, à l'exception de LIME, les modèles de régression logistique et de SVM produisent des explications beaucoup plus robustes que les modèles Random Forests et XGBoost. Ceci est probablement lié à la complexité des modèles. D'une part, un modèle plus complexe est généralement plus difficile à expliquer, même pour les méthodes d'explication agnostiques, et d'autre part, un modèle plus complexe conduit à des fonctions hautement non linéaires, ce qui signifie que des instances proches les unes des autres peuvent avoir des prédictions différentes et, par conséquent, des explications différentes.

Lisibilité

En ce qui concerne la lisibilité, la Figure 5.9 montre que les explications fournies par le modèle de régression logistique sont beaucoup plus lisibles que celles fournies par les autres modèles. Les explications fournies par le modèle SVM se situent entre le modèle de régression logistique et les modèles basés arbres, pour la plupart des méthodes d'explication. Cela est probablement dû au fait que les modèles plus simples ont tendance à établir des relations entre les variables (individuellement) et le résultat, sans nécessairement tenir compte des interactions entre variables, produisant ainsi des explications qui peuvent être plus facilement lues variable par variable.

Clusterabilité

Enfin, nous examinons en Figure 5.10 la capacité de clusterabilité des explications appliquée aux modèles prédictifs. Nous pouvons constater que tous les modèles ont une capacité de clusterabilité similaire entre eux. Cela peut indiquer que le modèle n'est pas important pour déterminer des sous-populations particulières d'explications par paires de variables, ou que cela dépend davantage du jeu de données considéré que du modèle.

5.5 Feuille de route pour l'usage des méthodes locales attributives

Le Tableau 5.1 résume les avantages et les inconvénients de chaque méthode d'explication évaluée. Dans l'ensemble, nous soulignons le fait que les méthodes coalitionneles devraient être plus efficaces pour produire des explications locales précises, tandis que *SHAP* devrait être plus efficace pour produire des explications globales cohérentes et faciles à interpréter. Cela est également confirmé par le fait que *SHAP* a tendance à accorder plus d'importance à quelques variables que les autres méthodes, produisant ainsi des explications globales plus concises, mais masquant potentiellement d'autres contributions et interdépendances de variables. Les explications basées sur *Spearman* sont dans l'ensemble légèrement moins robustes que les autres méthodes, et les méthodes coalitionneles sont légèrement moins lisibles dans l'ensemble. *LIME* présente plusieurs inconvénients, l'un des plus remarquables étant sa tendance à ne pas tenir compte des interactions entre les variables ainsi que des influences complexes.

Nom de la méthode		Avantages		Inconvénients	
Basée sur les coalitions	Complète	Considère les dépendances entre variables	Valeurs de Shapley exactes	Lente en haute dimension Explications globales difficiles à lire	
	Spearman		Paramètre α pour contrôler le niveau d'approximation		Moins robuste pour les modèles basés arbre
LIME		Rapide en haute dimension Nombreux paramètres pour trouver un compromis entre robustesse et localité		Lente en basse dimension Qualité faible des explications Tendance à ne pas détecter les influences non linéaires et non monotones Manque de robustesse avec des modèles simples Peut manquer des relations entre paires de variables	
SHAP	KernelSHAP	Facile à interpréter les explications globales	Paramètres nombreux	Les approximations peuvent être imprécises	Lente en haute dimension
	TreeSHAP		Très rapide en basse et haute dimensions		Spécifique aux modèles basés arbre
	TreeSHAPapprox				

TABLE 5.1 – Tableau récapitulatif des avantages et des inconvénients de chaque méthode d'explication

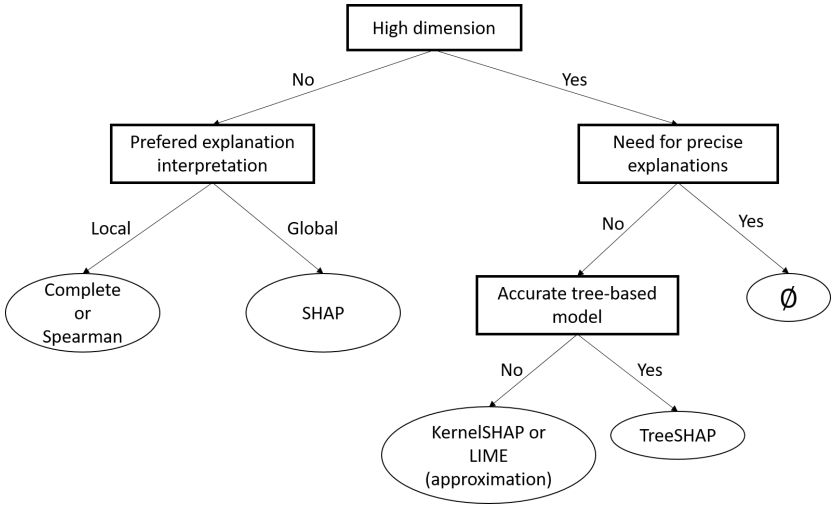


FIGURE 5.14 – Feuille de route pour un usage approprié des méthodes d'explication

En ce qui concerne les hyperparamètres des méthodes testées, chaque méthode propose un nombre et des types de paramètres différents. Le paramètre α de *Spearman* permet aux utilisateurs de contrôler facilement le compromis entre le temps de calcul et le degré d'approximation des explications. Les paramètres de *LIME*, nombreux et complexes, permettent un réglage fin de la méthode, mais nécessitent une connaissance approfondie de son comportement en tenant compte aussi du type de modèle prédictif et du jeu de données considérés. Les paramètres de *KernelSHAP* sont similaires à ceux de *LIME*, mais ils semblent induire des changements moins importants dans les explications obtenues, ce qui permet de ne pas trop dégrader la qualité des explications tout en optimisant les calculs.

A l'aide des résultats obtenus lors des évaluations, nous proposons une feuille de route simplifiée, sous la forme d'un arbre de décision présenté dans la Figure 5.14, avec l'intention d'aider les utilisateurs à identifier la méthode d'explication la plus appropriée en fonction de leurs jeux de données et de leurs objectifs.

Sur cette figure, une dimension élevée représente le nombre de variables présentes dans le jeu de données étudié. En effet, il n'y a pas de seuil "strict" pour définir le passage d'une dimension faible à une dimension élevée, mais nous pouvons raisonnablement considérer que ce seuil se situe quelque part entre 11 et 15 variables, en fonction de la complexité du jeu de données et du temps de calcul tolérable ainsi que du matériel disponible. Un "modèle basé arbre précis" représente la capacité d'entraîner un modèle basé arbre satisfaisant (défini par les objectifs de l'utilisateur) sur le jeu de données. Le modèle peut ensuite être expliqué grâce à l'optimisation réalisée dans *TreeSHAP*. Si le modèle souhaité n'est pas basé arbre, nous conseillons à l'utilisateur d'examiner les paramètres de *KernelSHAP* et de *LIME* afin de réduire le temps de calcul de leur phase d'échantillonnage, jusqu'à ce que les explications soient calculées en un temps raisonnable. Cependant, nous mettons en garde contre la perte potentielle de précision et de robustesse induite par de telles approximations.

Enfin, nous montrons que *SHAP* et *LIME* peuvent faire des approximations importantes dans certains cas, et que les méthodes coalitionnelles ne peuvent pas être exécutées en un temps raisonnable, en haute dimension. Cela laisse de futurs travaux de perspectives possibles pour des explications précises en haute dimension, cette problématique n'étant pas encore abordée dans la littérature, à notre connaissance.

5.6 Conclusion

5.6.1 Bilan

Dans ce chapitre, nous nous sommes intéressés aux limites de certaines méthodes d'explication post-hoc locales attributives. Nous y avons souligné le manque de considération de l'interaction entre variables dans les approches traditionnelles comme $SHAP$ ou $LIME$ et soulevé le besoin de mieux cerner les avantages et inconvénients de chaque méthode afin d'en définir les bonnes pratiques.Ainsi, nos contributions consistent en trois points :

- La proposition de plusieurs méthodes d'explication dites coalitionnelles, basés sur des calculs de corrélation entre variables.
- La proposition de six métriques d'intérêt pour les méthodes attributives.
- la proposition d'une feuille de route pour l'usage de ces méthodes attributives.

Concernant **les méthodes coalitionnelles**, nous sommes d'abord revenus sur le principe de base du calcul des coalitions par la méthode *complète* (valeur de Shapley) en y soulignant un problème de complexité exponentielle. Pour y remédier en partie, nous avons ensuite proposé les méthodes *k-complète* ainsi que les méthodes coalitionnelles basées sur des sous-groupes de variables corrélées. Le principal but de toutes ces méthodes est d'approcher le plus possible les explications produites par la méthode *complète*, tout en minimisant le nombre de sous-groupes d'interactions entre variables à calculer. Les résultats ont montré que l'approche coalitionnelle basée sur la corrélation de Spearman donnait les meilleurs résultats en termes de gain de temps tout en étant relativement proche des influences produites par la

méthode *complète*.

La multitude des méthodes attributives, composées de $SHAP$, $LIME$ ainsi que des méthodes coalitionnelles, implique le besoin d'identifier leurs particularités et notamment d'évaluer l'intérêt pour l'utilisateur des différentes stragégies d'explication. Pour cela, nous avons proposé **six métriques d'intérêt** spécifiquement conçues pour étudier les vecteurs d'influences produits par les méthodes attributives. En particulier, nous avons proposé des métriques permettant d'analyser la capacité des méthodes d'explication à mettre en évidence les corrélations et clusters entre valeurs des données et influences afin d'en déduire un score de lisibilité et de clusterabilité des méthodes.

A l'aide des métriques proposées, nous avons finalement pu comparer et analyser les différences entre les méthodes attributives pour en produire une **feuille de route de leur bonne utilisation**. Cette feuille de route est issue des conclusions issues de nos évaluations et prend la forme d'un arbre de décision dont les critères de séparation prennent en compte la dimensionnalité des jeux de données, le type de modèle prédictif et les préférences en précision des explications. Cette feuille de route, et les évaluations préalablement conduites soulèvent ainsi que le lien étroit qui existe entre les caractéristiques des jeux de données, le type de modèle prédictif utilisé et les explications produites par les méthodes attributives. Nous soulignons donc qu'il n'existe donc pas de méthode d'explication applicable dans tous les cas de figure.

5.6.2 Perspectives

Comme nous avons pu le constater dans notre proposition de feuille de route, il n'existe pas, à l'heure actuelle, de proposition de solution d'explication locale attributive efficace et précise pour des jeux de données de grande dimension. Une solution pour pallier, en partie, ce problème de grande dimension pourrait être d'effectuer préalablement une **étude globale de l'importance des variables** à l'aide de mesures telles que *l'importance des variables*, indépendant du modèle prédictif, ou *l'indice de Gini* pour les modèles basés arbre. Ces informations sont ensuite utilisées pour calculer les influences uniquement pour les variables les plus importantes lors de la génération d'explications individuelles. Cette solution serait certainement intéressante pour les approches coalitionnelles proposées dans ce chapitre.

A vu de nos résultats concernant les **méthodes coalitionnelles**, leurs logiques pourraient d'ailleurs être intégrées à **au sein de techniques comme SHAP** afin de ne concentrer les calculs d'influences que sur les sous-groupes de variables corrélées.

Les métriques proposées dans ce chapitre peuvent être considérées comme un premier apport à la constitution d'un benchmark, comme discuté dans les perspectives du chapitre précédent. Il faudrait sans doute pouvoir **étendre ces métriques d'intérêt à toute méthode d'explication post-hoc**, en particulier les explications contrefactuelles. Comment s'assurer qu'une contrefactuelle reste lisible ? Que l'étude d'un sous-ensemble de

contrefactuelles reste informative pour l'utilisateur ?

Ces métriques peuvent être aussi vues comme une première réponse à la question de savoir si **les explications générées sont susceptibles de provoquer de l'intérêt pour l'utilisateur et notamment de lui permettre d'imaginer de possibles causalités** à valider (ou non) ultérieurement. Dans le cas d'une analyse de données, et notamment exploratoire, c'est certainement l'une des raisons de pouvoir utiliser des explications (comme nous le verrons dans le chapitre suivant). Cela pose évidemment la question d'une définition d'une causalité possible et plausible. Dans ce chapitre, nous l'avons surtout vu sous le prisme de la corrélation entre valeurs des données et explications et dans leur capacité à identifier des clusters.

Chapitre 6

Les explications comme un nouvel espace de données

6.1 Introduction

Comme nous avons pu le voir tout au long des différents chapitres précédents, les méthodes post-hoc locales attributives génèrent des vecteurs d'influences, pour toute instance d'un jeu de données, vis-à-vis d'un modèle prédictif. En tout état de cause, cette multitude d'explications peut se voir comme la base d'un nouvel espace de données à valoriser et analyser.

Ainsi, l'ambition de ce chapitre est d'étudier l'impact et l'usage des explications dans la réalisation de tâches liées à l'analyse prédictive. Nous montrons que les explications peuvent aider à des tâches aussi variées que (1) la sélection de modèles, (2) la sélection de variables, (3) la sélection d'instances, ou encore (4) l'analyse de données.

Concernant la sélection de modèles, nous montrons, à l'aide d'un cadre orienté "human-in-the-loop", que les explications locales peuvent être une aide à la fois dans le choix de modèles prédictifs recommandés (voir Chapitre 3) ainsi que dans la compréhension et le raffinement de celui-ci.

Pour une tâche de sélection automatique de variables, nous mettons en évidence que l'explicabilité peut être vue comme une dimension supplémentaire à prendre en compte pour obtenir le sous-ensemble de variables le plus adéquat. En particulier, nous montrons que le choix de ce sous-ensemble, via un modèle prédictif, peut impacter significativement les explications produites.

Nous mettons également en évidence l'utilité des explications pour la sélection d'instances représentatives, au sens du modèle prédictif. Le nombre d'explications générées étant aussi important que la taille du jeu de données à analyser, la possibilité d'identifier les instances découlant des explications les plus représentatives permettrait alors une meilleure compréhension du modèle prédictif.

Enfin, nous montrons que les explications peuvent être une aide complémentaire à une analyse de données. Au travers d'un cas d'usage médical, nous détaillons l'intérêt de capita-

liser sur l'analyse des explications afin de détecter des relations non linéaires entre variables (difficilement identifiables par l'analyse seule des données brutes) ainsi que de caractériser plus simplement des sous-groupes de patients.

Ce chapitre fait référence aux travaux publiés dans [Ferrettini et al., 2020c, Wang et al., 2023, Excoffier et al., 2022, Escriva et al., 2023a, Escriva et al., 2023b], incluant une partie des travaux de thèse de Gabriel Ferrettini, Haomiao Wang, Elodie Escriva et Emmanuel Doumard.

6.2 Aide à la sélection de modèles

Nous présentons tout d'abord l'intérêt des explications pour une aide à la sélection et raffinement de modèles.

Notre cadre d'aide à la sélection de modèles est divisé en trois composantes successives, orienté dans une stratégie "human-in-the-loop", illustrée en Figure 6.1, que nous développons dans les sections suivantes :

— Etapes (1) et (2) : A partir d'un ensemble de workflows à disposition (obtenus par exemple par le système de recommandation de modèles vu dans le Chapitre 3), nous montrons comment un utilisateur, expert de ses données, peut être guidé, à l'aide d'explications, dans le processus complexe de la sélection de modèles.

— Etapes (3) et (4) : Au cours de cette sélection, l'utilisateur est également guidé dans le processus de *feature engineering* pour son jeu de données, si nécessaire.

— Etapes (5) et (6) : Enfin, une fois le modèle créé, nous offrons la possibilité à l'utilisateur d'exploiter ce modèle, par le biais de nouvelles instances à expliquer, illustrant ainsi les nouvelles perspectives qu'offrent l'XAI dans ce cadre

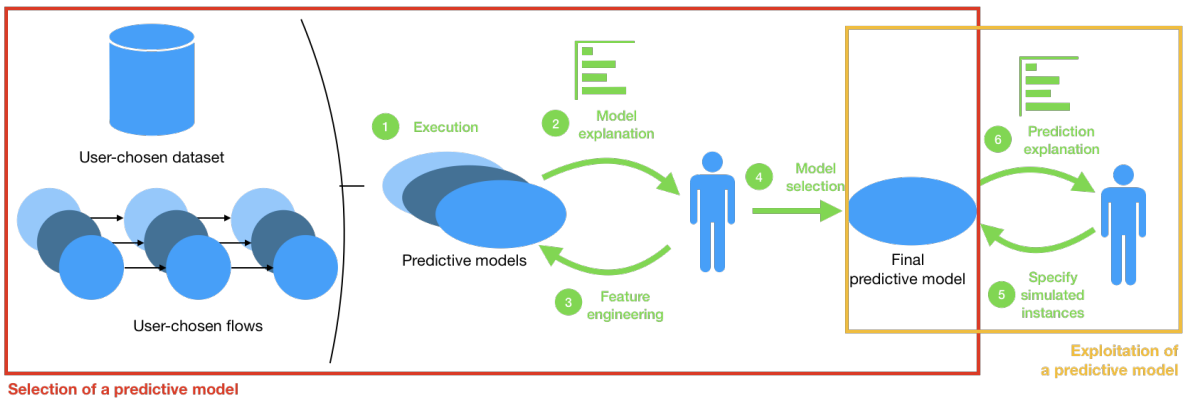


FIGURE 6.1 – Cadre pour l'aide à la sélection et raffinement d'un modèle prédictif et son usage.

6.2.1 Sélection d'un modèle à l'aide de l'explication de prédiction

Nous détaillons, dans cette section, les étapes (1) et (2) de la Figure 6.1 pour la sélection de modèles.

1. *Exécution* - Parmi une sélection de workflows, l'utilisateur peut accéder à une description de chacun d'entre eux et de son fonctionnement interne s'il le souhaite. Les workflows sont ensuite exécutés et produisent un ensemble de modèles prédictifs.
2. *Explication du modèle* - À l'aide de ces modèles, le système peut générer la classification de toute instance du jeu de données de départ et fournir son explication, pour chaque modèle. Ces explications prennent la forme de vecteurs d'influences de variables. Grâce à ces explications, l'utilisateur peut alors détecter l'influence des variables sur toutes les prédictions souhaitées.

6.2.2 Raffinement d'un modèle par feature engineering et explicabilité

Le raffinement du modèle prédictif sélectionné est détaillé dans les étapes 3 et 4 de la Figure 6.1.

3. *Feature engineering* - Grâce aux explications sur les prédictions, un utilisateur peut accéder au raisonnement qui sous-tend chaque modèle, ce qui lui permet de détecter d'éventuelles failles dans les modèles proposés. Par exemple, l'explication des prédictions peut permettre au personnel d'un hôpital réalisant une étude médicale décrite dans [Ribeiro et al., 2016] de se rendre compte que certaines variables n'auraient pas dû être incluses dans leur ensemble de données. En outre, sur la base de son propre domaine d'expertise, un utilisateur peut évaluer l'importance de chaque variable par rapport à l'importance que leur accordent les modèles. Il peut ainsi sélectionner les variables indésirables et les supprimer du jeu de données et réentraîner le modèle prédictif.
4. *Sélection de modèle* - Une fois que les variables finales souhaitées ont été déterminées, l'utilisateur exploite sa connaissance du domaine pour évaluer le raisonnement lié à chaque modèle. Cette évaluation est basée à la fois sur une évaluation globale, à l'aide par exemple d'un kappa de Cohen ou l'aire sous la courbe ROC, et sur les explications locales relatives à la prédiction. L'utilisateur sélectionne ensuite le modèle final souhaité en choisissant le modèle le plus performant, mais aussi celui qui utilise de la manière la plus pertinente les variables de l'ensemble de données.

6.2.3 Exploitation d'un modèle à l'aide de l'explicabilité

Après avoir obtenu le modèle prédictif affiné à l'aide des explications locales fournies, notre cadre propose une dernière phase d'exploitation du modèle. Les étapes (5) et (6) de la Figure 6.1 permettent de générer des prédictions pour de nouvelles instances (non présentes dans le jeu de donnée initial), en y associant les explications locales correspondantes.

5. *Explication de prédictions* - Le modèle prédictif étant prêt, l'utilisateur peut créer de nouvelles instances et les introduire dans le modèle afin d'en obtenir des prédictions. Ces prédictions sont produites avec leurs explications, qui indiquent à l'utilisateur comment les variables ont influencé le modèle pour générer les prédictions.
6. *Instances simulées* - L'utilisateur peut ensuite utiliser ces fonctionnalités pour explorer les possibilités de nouvelles prédictions. Pour ce faire, il peut par exemple randomiser les données d'une instance afin d'analyser les explications correspondantes ou bien tester des hypothèses en modifiant précisément les valeurs des instances.

Dans ce contexte de bac à sable, l'utilisateur peut tester différents cas, réels ou hypothétiques. Les prédictions pour ces nouvelles instances donnent à l'utilisateur un nouvel aperçu du comportement du modèle en fonction de ces nouvelles données. Par exemple, un médecin pourrait tester l'évolution des risques de développer un diabète si le patient modifie son régime alimentaire ou son activité physique. De plus, en fixant les valeurs d'une variable et en randomisant les autres pour générer un nouvel ensemble d'instances, l'utilisateur peut voir les effets de cette variable particulière sur une plus grande population. Cette fonctionnalité vise à offrir un large éventail de possibilités à l'utilisateur tout en l'aidant à interpréter les résultats obtenus à l'aide d'une explication de la prédiction.

6.2.4 Illustration du cadre

Afin de montrer le potentiel de notre cadre d'aide à la sélection de modèles, nous proposons une maquette possible illustrant son utilisation. Pour cela, nous utilisons le jeu de données des Indiens Pimas [Smith et al., 1988] pour des tâches d'apprentissage supervisé. Dans notre cas d'utilisation, un biologiste souhaite étudier ce jeu de données et l'utiliser dans notre système de recommandation (Chapitre 3) pour fournir des workflows d'analyses possibles. Tout d'abord, l'utilisateur saisit le jeu de données en tant qu'entrée du système de recommandation et lui demande d'effectuer une recommandation.

Appropriation du modèle par l'utilisateur

L'utilisateur se voit donc présenter un ensemble des meilleures recommandations du front de Pareto. Une description de chaque workflow est mise à la disposition de l'utilisateur (voir Figure 6.2) afin qu'il puisse effectuer une première sélection parmi les différentes options. Bien que ces descriptions soient nécessairement techniques, elles sont essentielles pour permettre à l'utilisateur de comprendre le fonctionnement de chaque workflow. À titre d'exemple, nous pouvons voir dans la Figure 6.2 qu'un workflow n'est pas seulement la production d'un modèle prédictif, mais aussi des opérations successives de transformation appliquées à l'ensemble de données.

Ces workflows sont ensuite exécutés et présentés à l'utilisateur par le biais d'un ensemble d'instances sélectionnées. Ces instances sont sélectionnées de manière à favoriser une grande diversité dans les explications de prédiction. L'algorithme exact utilisé ici est celui présenté dans [Ribeiro et al., 2016]. L'utilisateur peut ainsi explorer chaque modèle prédictif à travers

cet ensemble d'instances, en visualisant un ensemble varié de points clés illustrant les modèles. Il en déduit ensuite le fonctionnement de l'ensemble du modèle, avec un minimum d'informations. Les instances et les explications associées sont représentées dans la Figure 6.3. À gauche, l'utilisateur peut sélectionner l'instance qu'il souhaite étudier et décider de supprimer éventuellement des variables du jeu de données. Sur la droite est présentée l'explication de la prédiction de l'instance sélectionnée pour chacun des modèles (comme celui entouré en vert). Par exemple, les modèles Random Forest et Bagging J48 (un arbre de décision optimisé) basent principalement leurs prédictions de cette instance sur la *tension artérielle* et l'*âge*, tandis que le modèle Naive Bayes est principalement influencé par la variable *masse*. Ainsi, en présentant ces résultats et la manière dont ils ont été obtenus, l'utilisateur comprend mieux ce qui a influencé le modèle pour la prédiction de telle ou telle instance, sans se fier aveuglément au modèle seul. Dans notre cas d'utilisation, nous pouvons voir que le biologiste sélectionne l'instance 49.

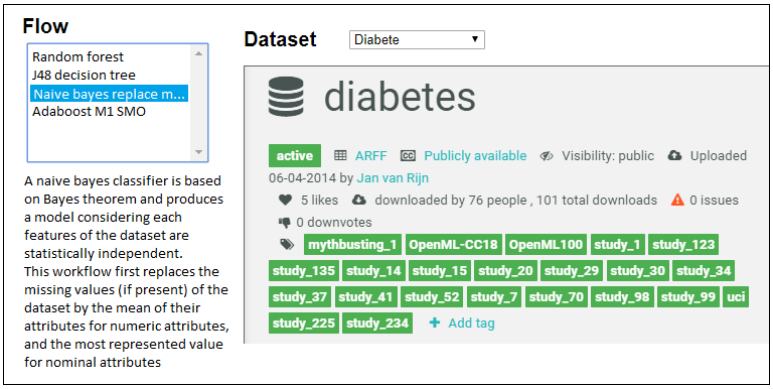


FIGURE 6.2 – Recommandation de workflow

6.2.5 Confiance de l'utilisateur dans les résultats produits

Grâce à cette méthode d'explication, l'utilisateur peut choisir entre les modèles sans devoir s'appuyer uniquement sur des mesures globales de performance (en précision par exemple). Il peut utiliser son propre jugement plutôt que de ne devoir faire qu'avec la seule proposition de workflow obtenue par un processus entièrement automatisé (AutoML). En outre, cette solution permet d'évaluer les défauts éventuels des modèles, ce qui n'est pas toujours possible avec les seules mesures conventionnelles. Par exemple, la précision globale d'un modèle ou le score de Kappa n'avertissent pas l'utilisateur d'une variable inappropriée qui devrait être supprimée de l'ensemble de données.

Dans notre exemple, l'utilisateur peut décider que l'*âge* d'un patient n'est pas très important pour déterminer s'il est susceptible de souffrir de diabète. En même temps, si notre utilisateur considère la *masse* d'un patient comme un indicateur valable, cela indique que le modèle Naive Bayes est plus intéressant dans son cas (en supposant que les instances qu'il

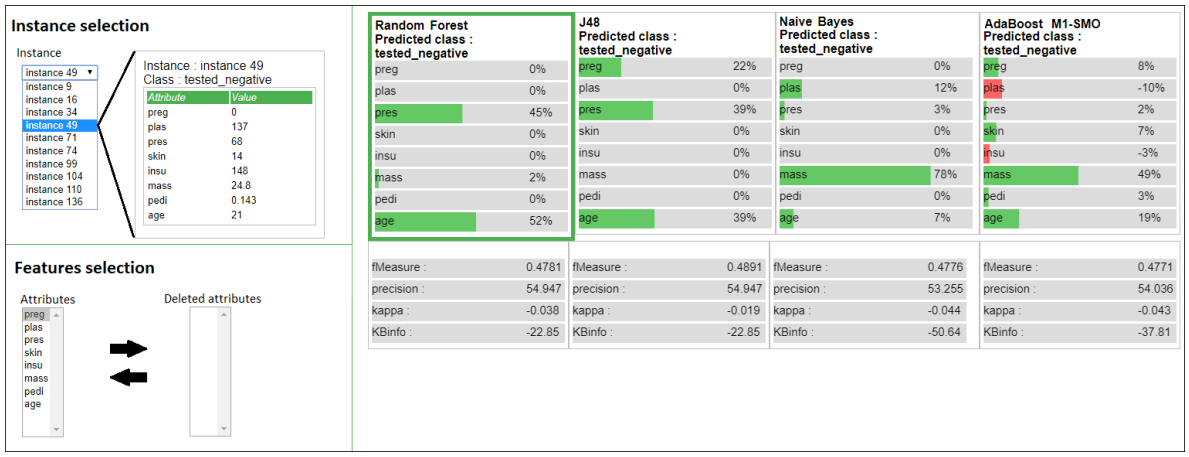


FIGURE 6.3 – Visualisation des résultats de prédictions avec les explications associées

a examinées soient cohérentes avec cette explication). Cette compréhension d'un modèle, de ses forces et de ses faiblesses peut donner à l'utilisateur une plus grande confiance dans ce qu'il accomplit au cours de son processus d'analyse des données. En repérant les éventuels problèmes du modèle prédictif, il est également en mesure de savoir dans quelles circonstances le modèle reste fiable.

6.2.6 Personnalisation d'un modèle

Une fois que l'utilisateur a étudié ses modèles, il peut évaluer les workflows répondant le mieux à ses besoins. En particulier, l'utilisateur peut identifier les variables principalement considérées par les workflows et décider lesquelles sont importantes pour son étude. Dans notre exemple, le biologiste pourrait vouloir étudier l'impact d'indicateurs de diabète moins évidents et décider de supprimer les caractéristiques de l'insuline et du plasma de son jeu de données (comme le montre la Figure 6.4). En effet, certaines variables pourraient avoir une influence importante sur un sous-ensemble d'instances, bien que cela ne soit pas le cas dans le jeu de données entier. Par exemple, nous pouvons voir sur la Figure 6.4 que le workflow J48 a considérablement modifié son comportement par rapport à la Figure 6.3, tandis que le modèle Adaboost a simplement ajusté l'importance de chaque variable.

6.3 Aide à la sélection de variables

Comme entraperçu dans le cadre de la sélection de modèles, au travers de l'étape de *feature engineering* notamment, la sélection de variables reste un enjeu très important, notamment pour répondre au fameux problème du *Fléau de la dimension (curse of dimensionality)*. Il s'agit, traditionnellement, d'identifier le sous-ensemble de variables le plus petit possible, maximisant la précision du futur modèle prédictif construit.

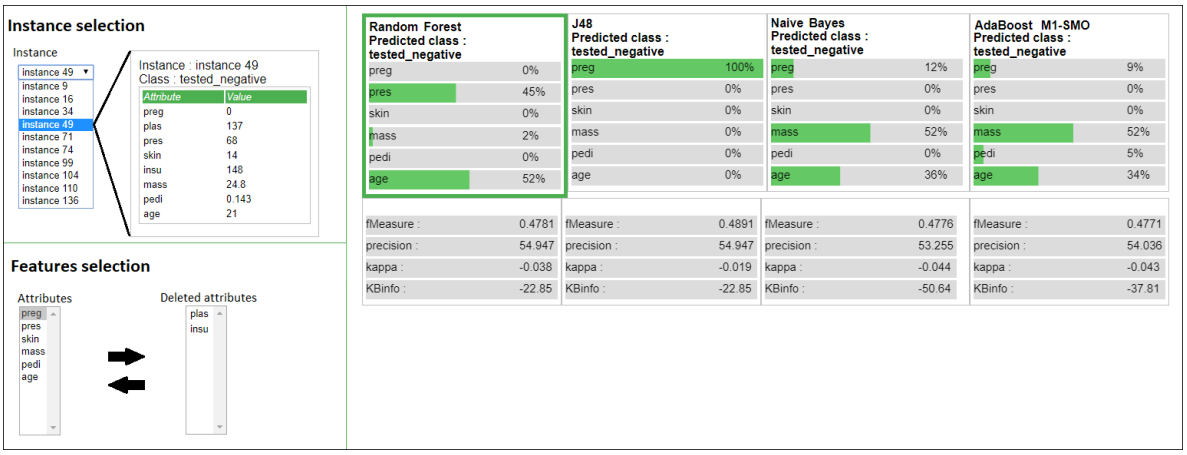


FIGURE 6.4 – Nouvelles explications de prédiction une fois que les variables plasma et insuline ont été supprimées.

Cependant, selon le théorème du "no free lunch" (NFL) [Wolpert and Macready, 1997], il n'existe pas de méthode universelle pour la sélection des variables. Les méthodes de sélection de variables peuvent être classées en trois catégories principales en fonction de leur dépendance à l'égard d'un modèle d'apprentissage automatique : *filter, wrapper* et *embedded*.

Les méthodes *filter* sont indépendantes du modèle d'apprentissage et se concentrent sur la métrique d'évaluation calculée uniquement à partir des données (variable et cible). Les méthodes *wrapper* sont "enveloppées" dans le modèle d'apprentissage ; elles utilisent directement le modèle final dans l'évaluation du sous-ensemble. La méthode *wrapper* doit alors être considérée comme un problème d'optimisation. Les méthodes *embedded* signifient que les algorithmes d'apprentissage intègrent intrinsèquement le processus de sélection de variables. Par exemple, étant donné que les modèles basés arbres ou basés règles intègrent des étapes de séparation pendant l'apprentissage, une partie des variables peut par conséquent être éliminée.

La définition de critères de qualité est considérée comme un point essentiel dans la comparaison des méthodes de sélection de variables. Les mesures d'accuracy sont traditionnellement utilisées dans ce but afin d'évaluer si le modèle prédictif obtenu est d'au moins aussi bonne précision (ou avec une perte minime) que si l'étape de sélection de variables n'avait pas été appliquée. Cependant, les chercheurs ont démontré que la précision n'est pas suffisante pour déterminer la pertinence d'un modèle prédictif [McNee et al., 2006, Hossin and Sulaiman, 2015]. Par conséquent, substituer la précision, ou la complémenter, pour l'évaluation des méthodes de sélection de variables reste encore un problème ouvert.

Nous proposons alors dans cette section de considérer l'explicabilité comme un moyen de pouvoir évaluer, en partie, la pertinence d'une sélection de variables. Mais cela suppose de pouvoir mesurer l'impact que peut avoir chaque méthode de sélection de variables sur les profils d'explications générés par le modèle prédictif.

6.3.1 La sélection de variables et l'explicabilité

Avant leur utilisation dans le domaine de l'explicabilité, la théorie des jeux et les valeurs de Shapley ont été utilisées comme mesures d'évaluation pour la sélection de variables [Cohen et al., 2005, Cohen et al., 2007, Sun et al., 2012]. Les valeurs de Shapley ont également été intégrées dans d'autres techniques de sélection de variables telles que Borutashap[Keany, 2020, Keany, 2022]. En effet, étant donné que les méthodes locales attributives affectent des valeurs d'influence à chaque variable, certains travaux utilisent ces valeurs sous la forme d'une explication globale comme méthode de sélection de variables [Man and Chan, 2021, Liu et al., 2022]. Une autre approche, nommée *SCI-XAI* [Moreno-Sanchez, 2021], a intégré ensemble les concepts de sélection de variables et d'XAI dans un seul workflow, mais ce travail n'a pris en compte que les modèles ensemblistes à base d'arbres et a utilisé un modèle intrinsèquement explicable pour quantifier l'impact des méthodes de sélection de variables. Il est donc difficile de généraliser à tous les cas réels en raison des limites du modèle et de l'explainer.

6.3.2 Cadre expérimental

Les étapes suivantes présentées en Figure 6.5 ont été mises en oeuvre pour étudier l'impact des méthodes de sélection de variables sur les explications obtenues : sélection des jeux de données, sélection de variable, l'apprentissage et l'explication du modèle ainsi que le calcul des métriques appropriées. Les résultats des expériences sont accessibles sur Github [1].

Sélection des jeux de données

Nos expérimentations se basent sur les jeux de données de la base OpenML. Nous considérons le sous-ensemble des jeux de données respectant les conditions suivantes : 1) tâches de classification binaire, 2) variables à expliquer uniquement continues, 3) pas de données manquantes, 4) 10 à 150 variables, 5) moins de 12 000 instances. 144 jeux de données répondent à ces conditions après suppression de jeux de données doublons ou liés à de trop nombreuses études de séries temporelles (sur les taux de changes).

Sélection de variables

Un large éventail de méthodes de sélection de variables a été considéré afin de couvrir à la fois les différentes familles (*filter, wrapper* et *embedded*) et les différentes stratégies de calcul au sein d'une même famille :

— Filter : nous considérons les méthodes basées sur la similarité (*fisher*, *reliefF* [Kira and Rendell, 1992] et *spec* [Zhao and Liu, 2007]), les méthodes basées sur les statistiques (*f* et *chi2*), les méthodes basées sur le sparse learning (*rfs* [Nie et al., 2010]) et les méthodes basées sur la théorie de l'information (*mrmr* [Peng et al., 2005], *cmim* [Fleuret, 2004] et *jmi* [Yang and Moody, 1999])

1. https://github.com/haomiaow/XAI_feature_selection

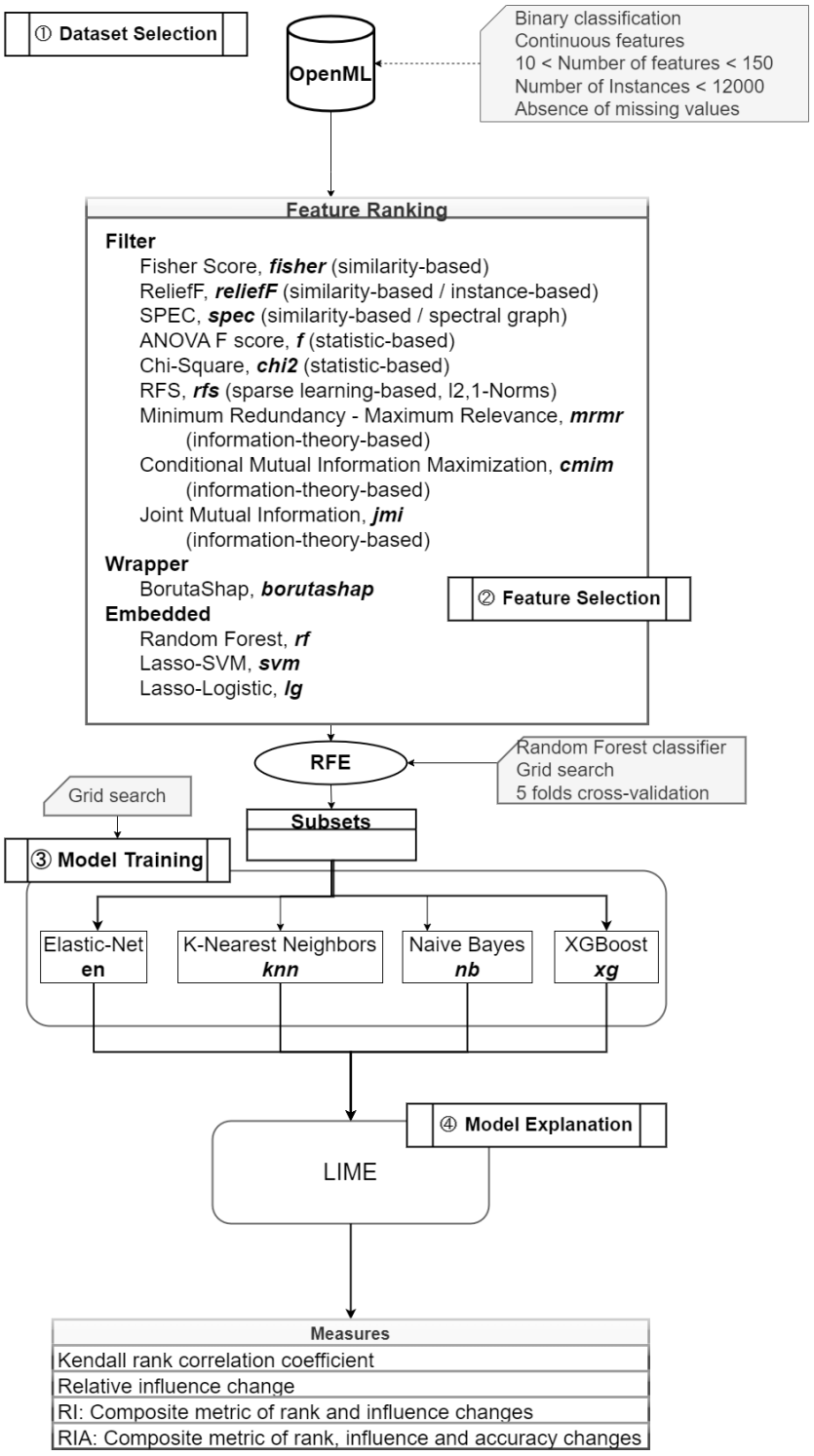


FIGURE 6.5 – Schéma du cadre expérimental

— Wrapper : nous considérons les méthodes basées sur l'importance des variables (Random Forest, *rf*) ou le coefficient (Linear Support Vector - *svm*, Logistic Regression - *lg*)
— Embedded : nous considérons la méthode *borutaShap*, combinant la technique Boruta avec les valeurs de SHAP [Keany, 2020].

Notre méthodologie comprend aussi une étape d'élimination récursive des variables avec validation croisée 5 fois (RFECV) [Guyon et al., 2004] afin de déterminer un seuil approprié de manière reproductible. Pour cela, nous utilisons un RandomForest dont les hyperparamètres ont été réglés (*max_depth, min_sample*) au préalable par un grid-search. Cette étape est essentielle, car elle permet de déterminer le nombre minimal de variables à sélectionner entre chaque méthode de sélection de variables. Il peut arriver, en effet, qu'à précision égale, différents sous-ensembles de variables sélectionnés soient possibles. Par exemple un sous-ensemble de 30 variables et un autre de 20 variables impliquant, par la suite, un modèle prédictif avec une précision de 80% dans les deux cas ; Dans ce cas ci, 20 variables seront préférées puisque le nombre de variables est plus petit.

Apprentissage du modèle

Quatre algorithmes de classification ont été choisis pour refléter une certaine diversité de stratégies algorithmiques d'apprentissage : Elastic-Net (*i.e.*, *en*, modèle linéaire avec pénalisation)[Zou and Hastie, 2005], K-Nearest Neighbors (*i. e.*, *knn*, modèle basé sur une distance), Naive Bayes (*i.e.*, *nb*, modèle probabiliste), XGBoost (*i.e.*, *xg*, modèle ensembliste basé sur les arbres) [Chen and Guestrin, 2016]. Les hyperparamètres de *en* (*l1_ratio*), *knn* (*n_neighbors*) et *xg* (*max_depth, min_child_weight, gamma, eta*) ont été réglés par un grid-search avec validation croisée 5 fois. Chaque classifieur a été entraîné séparément avec chaque sous-ensemble de variables généré par chaque méthode de sélection de variables (comme décrit précédemment). Toutes les instances ont été utilisées pour l'entraînement afin d'éviter tout biais dû à l'échantillonnage. Un score d'accuracy a également été calculé comme mesure de performance.

Explication du modèle

Malgré ses défauts, comme nous avons pu le constater dans les Chapitres 4 et 5, nous avons choisi d'utiliser LIME pour expliquer nos modèles. En tant que méthode attributive, elle affecte une valeur d'influence à chaque variable de chaque instance, qui représente sa contribution à la prédiction. En outre, l'un des avantages de LIME, par rapport à d'autres méthodes attributives telles que KernelSHAP et les méthodes coalitionnelles, est une complexité de calcul moindre lorsque le nombre de variables augmente (quel que soit le modèle prédictif), ce qui est essentiel pour la faisabilité de la présente étude.

6.3.3 Métriques utilisées

Afin de comparer nos différentes méthodes de sélection de variables sous le prisme des explications fournies par LIME via les modèles prédictifs, il est nécessaire de définir plusieurs métriques permettant de mesurer les différences de profils d'explications. Nous proposons, à ce titre, quatre métriques.

Corrélation de rang de Kendall

En tant que méthode locale, LIME calcule une explication globale avec classement de l'importance des variables à l'aide de l'équation 6.1, où $M_{i,j}$ désigne l'explication de la $j^{ème}$ variable pour une instance i d'un jeu de données avec n instances et f variables. La fonction $argsort(v_f)$ renvoie l'ordre décroissant d'un vecteur de f éléments (i.e., vecteur d'importance de la variable). :

$$ranking = argsort(\sum_{i=1}^{n}|M_{i,j}|) \tag{6.1}$$

La corrélation de rang de Kendall [Kendall, 1938], *i.e.*, aussi appelé τ de Kendall, est une statistique non paramétrique qui mesure la similarité entre deux classements. Cette méthode compare la position de chaque paire d'éléments dans les deux classements pour déterminer si la paire est concordante ou discordante. L'équation 6.2 définit le τ en utilisant le nombre de paires concordantes et discordantes :

$$\tau = \frac{\#\ of\ concordant\ pairs - \#\ of\ discordant\ pairs}{\frac{1}{2} \times n \times (n-1)} \tag{6.2}$$

où n est le nombre d'éléments dans le classement.

Un τ de 1, 0 ou -1 signifie que les deux classements sont identiques, indépendants, ou inversés, respectivement. Comme les mêmes variables sont nécessaires pour la comparaison, seule l'intersection de deux sous-ensembles a été prise en compte.

Changement d'influence relatif

Cette mesure est complémentaire du τ de Kendall afin d'observer les différences dans le classement des variables ayant de faibles changements entre les influences. La contribution d'une variable éliminée par la sélection de variables est ici considérée comme nulle. Afin de rendre la métrique d'influence indépendante du nombre de variables, une normalisation a été effectuée par rapport à l'influence totale.

$$inf'_f = \frac{inf_f}{\sum_{j=1}^{f}|inf_j|} \tag{6.3}$$

Dans l'Equation 6.3, pour une instance décrite par n variables, inf'_f représente la valeur normalisée de la valeur d'influence inf_i de la $f^{ème}$ variable.

$$diff = \sum_{j=1}^{f} \overline{|M_{i,j}^{FS\prime} - M_{i,j}^{O\prime}|}\Big|_{i=1}^{n} \tag{6.4}$$

L'Equation 6.4 montre le *Changement d'influence relatif* entre deux matrices d'explication d'un jeu de données comportant n variables et m instances, $M^{O\prime}$ désigne la matrice d'explication originale normalisée, $M^{FS\prime}$ représente l'explication normalisée obtenue après l'application d'une méthode de sélection de variables.

Métrique composite des changements de rang et d'influence : la métrique RI

L'intuition derrière cette métrique est de limiter la pénalité du changement de classement entre différentes variables si leurs influences sont proches, et inversement. En fait, la métrique combine le changement d'influence relative et le changement de classement. C représente les matrices M' d'explication normalisées, réorganisées par ordre décroissant d'importance des variables ; l représente la taille du sous-ensemble de variables sélectionnées ; la fonction $PR(M, f)$ renvoie le rang centile d'une variable f dans le classement, calculé à partir de la matrice d'explication M à l'aide de la formule 6.1. La racine de quatre sert à ajuster les deux pénalités à la même échelle.

$$\begin{aligned} RI = & \overline{(|PR(M^{FS}, j) - PR(M^{O}, j)| + \epsilon)} \\ & \overline{\times((\overline{|C_{i,j}^{FS} - C_{i,j}^{O}|}\Big|_{i=1}^{n})^{\frac{1}{4}} + \epsilon) - \epsilon^2}\Big|_{j=1}^{l} \end{aligned} \tag{6.5}$$

Métrique composite des changements de rang, d'influence et de précision : la métrique RIA

Sur la base de la métrique précédente, la RIA pénalise un modèle dont la précision est fortement dégradée par rapport au modèle original. La fonction $Acc(M)$ renvoie la précision du modèle associée à une explication donnée M.

$$\begin{aligned} RIA = & \overline{(|PR(M^{FS}, j) - PR(M^{O}, j)| + \epsilon)} \\ & \overline{\times((\overline{|C_{i,j}^{FS} - C_{i,j}^{O}|}\Big|_{i=1}^{n})^{\frac{1}{4}} + \epsilon)} \\ & \overline{\times(Acc(M^{O}) - Acc(M^{FS})) + \epsilon) - \epsilon^3}\Big|_{j=1}^{l} \end{aligned} \tag{6.6}$$

6.3.4 Résultats

Pour des raisons de concision, nous ne présentons qu'une petite partie des résultats produits (l'ensemble des résultats sont consultables dans [Wang et al., 2023]). Nous nous focaliserons surtout sur une analyse statistique globale ainsi que sur une analyse plus locale sur

un jeu de données, afin de montrer les différences concrètes d'explications entre méthodes de sélection de variables.

Analyse statistique

La Figure 6.6-A représente le τ *de Kendall* qui indique les classements des influences des variables, générés pour chaque méthode de sélection de variables, en utilisant l'explication sans sélection (*all*) comme référence. Les résultats sont globalement similaires pour une même technique de sélection de variables en fonction des différents modèles prédictifs, avec une tendance à un meilleur τ pour le modèle *xg*. Le τ le plus élevé est obtenu avec le modèle *xg* avec la méthode *borutashap* FS (.58), mais dans d'autres modèles, les τ les plus élevés sont trouvés avec *reliefF*. Le τ le plus faible est observé pour le modèle *knn* (*spec* FS model, .21). Le sélection de variables avec *spec* présente les coefficients les plus faibles quel que soit le modèle prédictif, avec un τ compris entre 0,21 et 0,29.

La Figure 6.6-B décrit le *changement d'influence relatif* entre les explications générées par chaque méthode de sélection de variables et les explications originales (*all*). Les changements les plus significatifs sont obtenus avec les méthodes *rfs* et *rf*, tandis que les changements les moins significatifs sont obtenus pour *reliefF* de manière cohérente avec tous les modèles prédictifs.

La valeur la plus faible de *RI* a été observée pour *reliefF*, avec des métriques et des écarts types en moyenne, réduits quel que soit le modèle ML, et inversement pour *spec* (Figure 6.6-C). Pondérée par la variation de la précision (Figure 6.6-D), la valeur positive de la RIA pour le modèle *knn* pourrait être liée à une amélioration de la précision pour toutes les méthodes de sélection de variables. En revanche, pour le modèle *xg*, la précision s'est légèrement dégradée, bien que les différences entre les méthodes de sélection de variables soient très mineures. La *RIA* d'une méthode de sélection de variables dépend largement du choix du modèle prédictif.

Jeux de données *Indian Liver Patient*

Le jeu de données *Indian Liver Patient* (OpenML ID 41945, [Ramana et al., 2011]) contient 583 instances, dont 416 patients présentant des lésions hépatiques et 167 patients sains. Le jeu de données a été utilisé pour évaluer les algorithmes de prédiction des maladies du foie et 10 variables sont présentes.

Nous présentons en Figure 6.7, les profils d'explicabilité pour le modèle *xg*, où nous pouvons remarquer que tous les modèles de sélection de variables appliqués ont une précision similaire. Cependant, leurs explications diffèrent considérablement ; *spec* et *mrmr*, bien que liés à des modèles de précision identiques, fournissent des explications radicalement différentes. L'explication de *mrmr* n'a même pas de sens véritable (ce problème est dû à un taux de rétention extrêmement faible et à un fort déséquilibre dans la classe cible) Le meilleur sous-ensemble a été généré par *borutashap*, qui a fourni les explications les plus similaires à l'ensemble de variables complet, bien que ce sous-ensemble contienne plus de variables que les autres méthodes de sélection de variables.

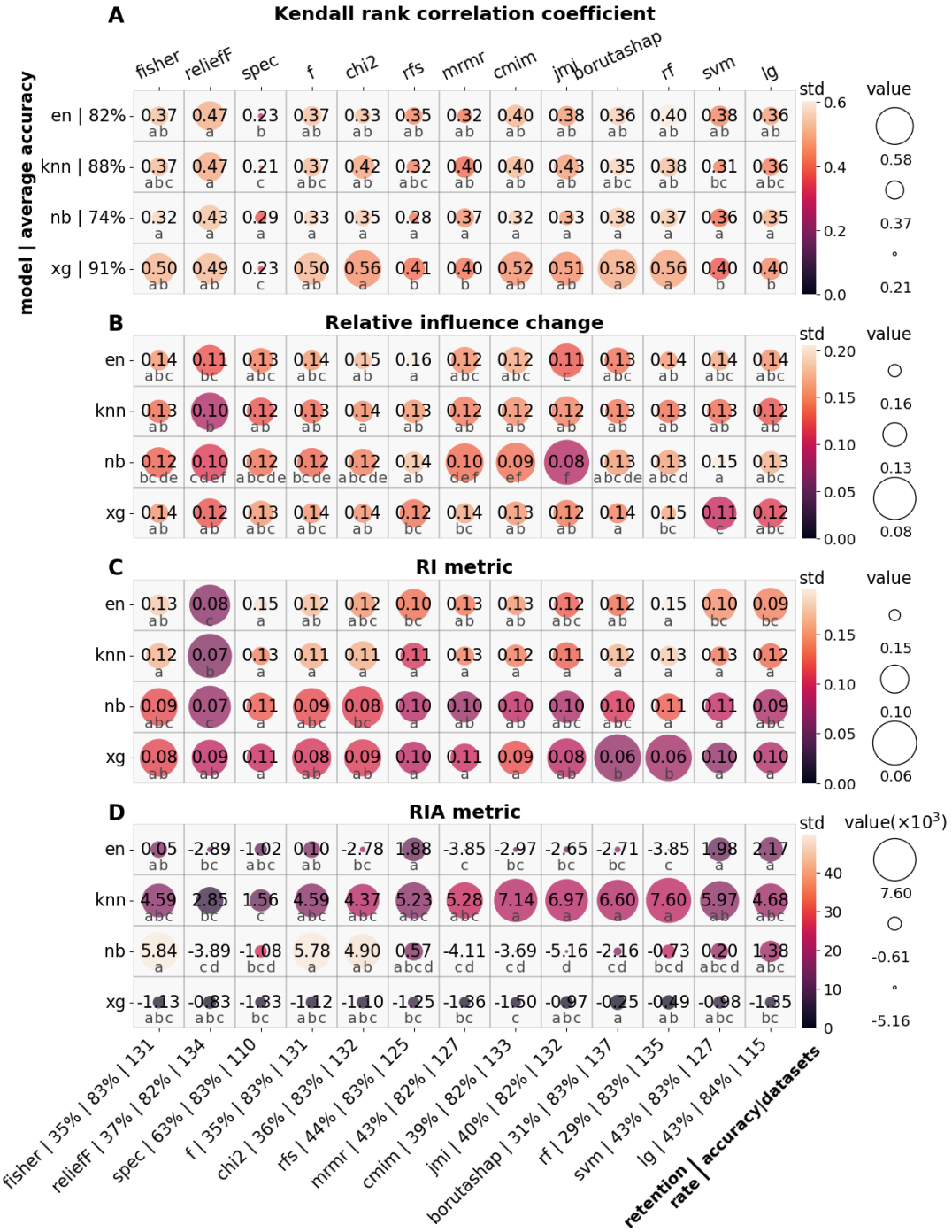


FIGURE 6.6 – Corrélation de rang de Kendall (A), changement d'influence relatif (B), métrique RI (C) et RIA (D). La taille et la valeur des cercles représentent les différences entre les explications générées par la méthode de sélection de variables sur l'axe x et l'ensemble des variables originales, pour le classifieur sur l'axe y. La couleur du cercle indique l'écart-type. Pour un modèle donné (en ligne), les lettres identiques indiquent qu'il n'y a pas de différence significative de métrique entre les méthodes de sélection de variables après ajustement sur la précision du modèle, l'identifiant de l'ensemble de données, le nombre de variables et d'instances et le taux de rétention. L'axe des abscisses indique les méthodes de sélection de variables et le taux de rétention, l'accuracy moyenne et le nombre de jeux de données concernés ; l'axe des ordonnées indique les classifieurs et l'accuracy moyenne.

Comme nous pouvons donc le constater avec cet exemple, le choix d'une sélection de variables peut aussi se voir à l'aide des profils d'explications, en complément de l'accuracy du modèle. Nous montrons aussi qu'une sélection de variables maximisant seulement l'accuracy peut mener à des prises de décision possiblement inappropriées, en rendant les explications incompréhensibles. Ainsi, vaut-il mieux une sélection de variables avec un profil explicable très clair, quitte à dégrader (légèrement) l'accuracy ? Dans tous les cas, nous défendons l'idée de considérer l'explication comme un critère à part entière dans le choix d'une sélection de variables, au même titre que l'accuracy.

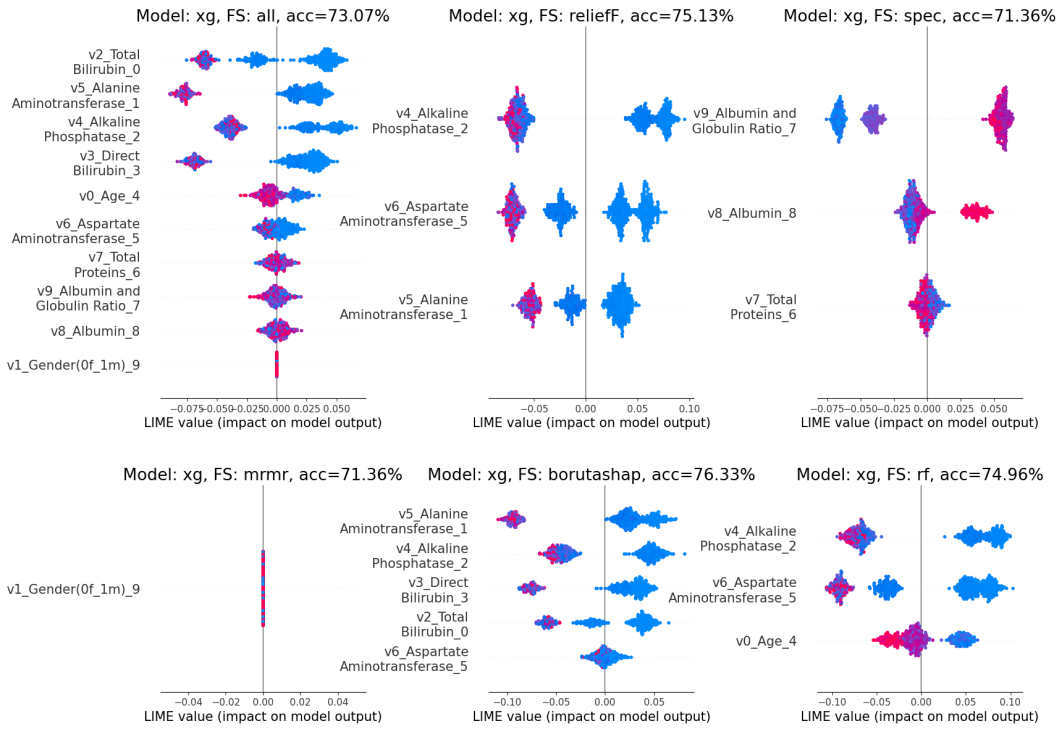


FIGURE 6.7 – Graphiques récapitulatifs pour le jeu de données *Indian Liver Patient* avec différentes méthodes de sélection de variables pour le modèle *xg*.

La sélection de variables dans un espace à trois dimensions

Afin de répondre aux conclusions faites sur nos expérimentations, nous proposons de considérer l'accuracy, le taux de rétention et l'explication (métrique RI) comme les trois dimensions permettant de choisir une méthode de sélection de variables appropriée. Dans la Figure 6.8, nous positionnons les différentes méthodes de sélections de variables dans cet espace en trois dimensions pour le modèle prédictif *en*.

Comme nous pouvons le voir, ces trois dimensions sont difficiles à concilier : la méthode de sélection de variables optimale est différente pour chaque dimension. *reliefF* et *spec* ont

des comportements respectifs complètement différents, avec une grande précision/un taux de rétention élevé/une grande variation des explications pour l'un et une plus faible précision/un faible taux de rétention/une faible variation des explications pour l'autre. *lg* et *rf* ont respectivement la meilleure précision et le meilleur taux de rétention. Un compromis doit donc être trouvé entre les trois dimensions, à prendre en compte dans un futur système de recommandation de sélection de variables à adapter aux priorités de l'utilisateur.

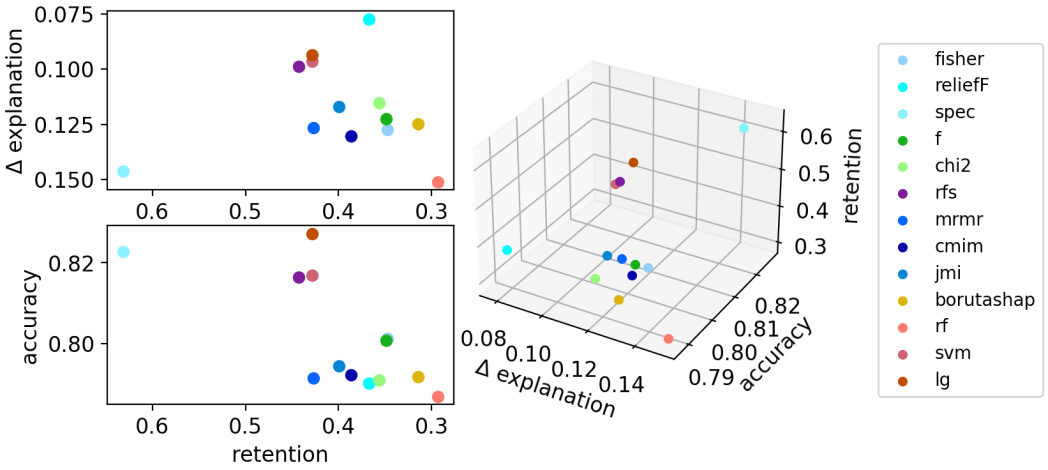


FIGURE 6.8 – Positionnement des méthodes de sélection de variables pour le modèle *en* dans un espace tridimensionnel dont les axes représentent respectivement l'explication (RI), l'accuracy et le taux de rétention.

6.4 Aide à la sélection d'instances

Comme souligné précédemment dans la proposition de cadre pour la sélection de modèles, la sélection d'instances peut être un moyen intéressant pour mieux appréhender et comprendre un modèle prédictif.

Les explications locales attributives associées à des instances que l'utilisateur souhaite visualiser pourraient être suffisantes pour mieux comprendre le jeu de données, associé au modelé prédictif. Ce qui fait d'ailleurs la popularité de ces méthodes d'explication est leur capacité à lier l'impact de chaque variable à la prédiction faite pour chaque instance et permettre ainsi de détecter des différences plus fines entre toutes les instances. Cependant le fait de ne fournir que des influences locales semble insuffisant pour améliorer l'efficacité de la prise de décision. En effet, [Weerts et al., 2019, Zhang et al., 2020] montrent que l'affichage seul des influences avec une prédiction individuelle n'a pas amélioré de manière significative l'utilité et la compréhension pour l'utilisateur par rapport à la prédiction. En outre, le fait de connaître

toutes les explications locales d'un jeu de données ne garantit pas une compréhension complète des données puisqu'il y a autant d'explications que d'instances dans les données brutes d'origine.

En revanche, les influences ont un avantage important à considérer : elles apportent de nouvelles informations grâce au modelé prédictif qui lui même prend en compte les phénomènes complexes et les interactions entre les données. L'analyse des influences peut ainsi permettre d'identifier les grandes tendances des explications, c'est-à-dire les relations particulières entre les variables. En outre, il peut être intéressant de fournir une vue globale des explications afin de déterminer si les instances sont des cas typiques ou atypiques des données. Dans cette optique, une approche de clustering, basée sur les influences, est sans doute un bon candidat pour détecter des sous-groupes d'influences plus homogènes et comprendre le comportement de la modélisation et du jeu de données sous-jacent. Ainsi, dans cette section, nous souhaitons proposer un cadre pour l'analyse des influences par le biais d'une approche de clustering, permettant d'identifier ensuite des instances à étudier par l'utilisateur.

6.4.1 Cadre du clustering basé sur les influences

La Figure 6.9 montre le processus, étape par étape, de clustering des instances en fonction de leurs influences :

1. Un modèle prédictif est formé à partir des données brutes et prédit les classes de toutes les instances du jeu de données brutes.
2. Une méthode d'explication attributive explique le modèle formé. Les utilisateurs peuvent choisir les données utilisées en entrée de la méthode. Les influences sont calculées pour expliquer pourquoi le modèle prédictif a fait telle ou telle prédiction.
3. Un algorithme de clustering utilise les influences pour créer des groupes homogènes d'instances afin de détecter leurs variables importantes sur la base du modelé prédictif. Les utilisateurs peuvent définir le nombre de clusters qu'ils souhaitent calculer.

Dans ce cadre, divers éléments peuvent être modifiés en fonction des préférences de l'utilisateur. N'importe quel modèle de classification peut être utilisé à l'étape 1, car ils sont tous conçus pour calculer des prédictions, et l'étape 3 permet d'utiliser n'importe quelle méthode de clustering.

À l'étape 2, le cadre est conçu pour accepter les méthodes d'explication locale attributive. Ces influences sont représentées sous forme de vecteurs, où chaque instance a une valeur associée à chaque variable. Nous utilisons directement ces données d'influence comme données d'entrée pour l'étape de clustering. Pour les tâches supervisées, les méthodes d'explication attributive génèrent généralement un jeu de données pour chaque classe avec des dimensions identiques à celles des données brutes. Par exemple, si les données brutes consistent en n instances et m variables et que la tâche supervisée est un problème multiclasses avec c classes, l'ensemble de données généré (également appelé ensemble de données d'influence) a une dimension de $n \times m \times c$. Pour obtenir un ensemble de données d'influence ayant la même dimension que les données brutes ($n \times m$), on ne peut sélectionner qu'une seule classe et ses

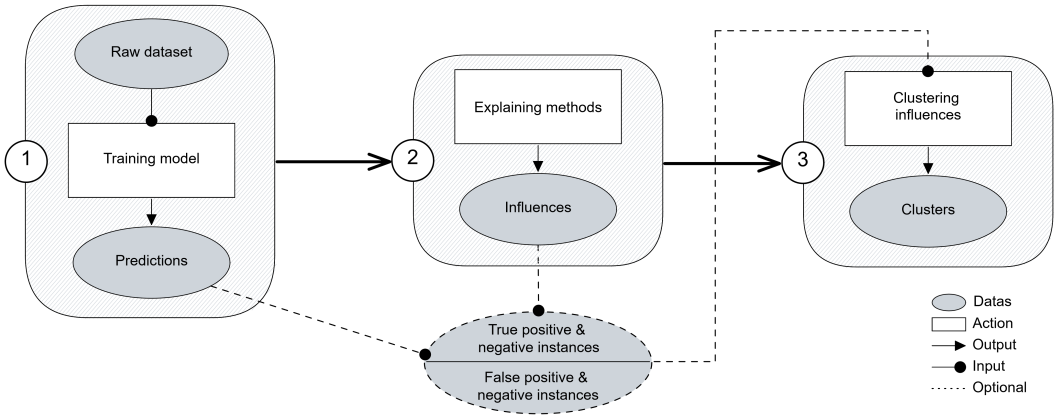


FIGURE 6.9 – Schéma du cadre du clustering basé sur les influences

influences associées. Par exemple, en ce qui concerne la classification binaire, la classe positive est souvent choisie comme classe d'intérêt pour les influences. Enfin, les méthodes d'explication attributive permettent de sélectionner des données d'entrée différentes de celles utilisées pour la formation du modèle d'apprentissage automatique. Ces données sont utilisées par les méthodes attributives pour expliquer le modèle, en créant des perturbations dans SHAP ou LIME, par exemple.

Une étape supplémentaire et facultative consiste à sélectionner un sous-ensemble particulier des données pour le clustering. En effet, il est possible d'étudier séparément les instances correctement et incorrectement classées par le modèle via le clustering d'instances. Compte tenu des prédictions du modèle par rapport aux labels des données, les influences sont séparées en deux groupes distincts avant d'être regroupées. Deux ensembles différents de clusters sont alors proposés aux utilisateurs. Cette étape peut présenter plusieurs avantages. Étant donné que les influences représentent les décisions du modèle, la séparation des instances peut apporter de nouvelles connaissances. L'étude des instances bien classées peut aider à identifier des patterns caractéristiques en éliminant le bruit et les valeurs aberrantes des instances mal classées. Cela peut donner une idée plus précise des modèles généraux, par exemple pour vérifier qu'il n'y a pas de biais dans le jeu de données. En ce qui concerne les instances mal classées, elles peuvent avoir plusieurs représentations. Elles peuvent être des valeurs aberrantes dans les données et ne pas correspondre aux comportements généraux sans biais ni erreur. Toutefois, les instances mal classées peuvent également constituer un sous-groupe particulier des données à étudier. C'est le cas, par exemple dans le domaine de la santé, où des enfants atteints de cancers sont généralement associés aux personnes âgées. En raison de l'âge, le modèle peut mal comprendre ce sous-groupe, car il y a peu d'enfants atteints de cancers non pédiatriques, où les variables d'entrée peuvent être insuffisantes pour identifier ce sous-groupe. Cependant, il est nécessaire d'étudier ce sous-groupe pour comprendre s'il existe un comportement spécifique et, en fin de compte, pour comprendre le jeu des données. La séparation des instances peut donc permettre d'explorer de nouveaux modèles invisibles si toutes les données étaient

conservées. Cela peut être encore plus important pour les influences en raison de leur lien direct avec le modèle. En effet, lorsque la prédiction du modèle est incorrecte, les influences reflètent cette erreur et sont directement affectées par la prédiction erronée du modèle.

L'implémentation complète est disponible sur ce lien GitHub [2].

6.4.2 Protocole expérimental

104 jeux de données ont été utilisés pour les expérimentations, répondant aux besoins suivants : classification binaire, plus de 100 instances, plus de quatre variables et au plus neuf variables en raison du coût de calcul vis à vis des influences (en particulier l'utilisation de Kernel SHAP). Nous entraînons un modèle Random Forest (RF) avec une validation croisée par grid-search pour optimiser les hyperparamètres. Ce modèle a été choisi pour tester les méthodes d'explication spécifiques aux arbres tout en conservant un nombre limité d'hyperparamètres pour éviter le surapprentissage (par rapport aux modèles par arbres boostés). Pour évaluer les performances des modèles, chaque jeu de données est divisé en ensembles d'entraînement et de test en fonction des ratios 75% et 25%.

Certains modèles que nous avons générés ont également une très faible précision, le minimum étant de 0.42. Nous avons choisi de séparer les modèles sur la base d'un seuil fixé à 0.8 afin d'évaluer le comportement de notre cadre sur les modèles ayant une précision élevée et faible. Ainsi, les modèles à haute précision ont une précision équilibrée médiane de 0.92, tandis que les modèles à faible précision ont une médiane de 0.6.

Nous étudions également le nombre d'instances bien classées et mal classées par les modèles prédictifs. Dans toutes les expériences, nous appelons *vraies instances* les instances bien classées, en référence aux termes Vrai positif et Vrai négatif. Les *fausses instances* sont alors liées aux instances classées en faux positif et faux négatif, c'est-à-dire aux instances mal classées. Nous utilisons trois séparations différentes des données : toutes les instances ensemble, uniquement les vraies instances et uniquement les fausses instances.

À des fins d'exhaustivité, nous choisissons les trois méthodes d'explication attributive pour calculer les influences : KernelSHAP/TreeSHAP, LIME et Spearman coalitional. Comme expliqué dans le Chapitre 5, chaque méthode XAI fournie des influences avec des avantages et des inconvénients différents. Nous souhaitons donc étudier la pertinence de l'utilisation de la classification par influence locale par rapport à la classification brute de manière globale.

Une fois les influences calculées, les instances sont regroupées par l'approche basée sur l'influence avec les K-médoids comme méthode de clustering. Cette méthode présente l'avantage de toujours sélectionner des instances réelles comme centroïdes, contrairement à d'autres méthodes de clustering telles que *k-means*. La distance euclidienne est celle considérée dans ce clustering. Nous utilisons 10 pourcentages différents pour choisir le nombre de clusters : 1%, 2%, 3%, 4%, 5%, 10%, 20%, 30%, 40% and 50%. Le nombre de clusters est ainsi défini en fonction du pourcentage selon $n_{cluster} = p * n_{instance}$ avec p le pourcentage de sélection entre 0 et 1. Nous fixons un nombre minimum de deux pour éviter un trop petit nombre de clusters.

2. https://github.com/kaduceo/XAI-based-instance-selection

La taille des jeux de données étant très variable, nous préférons sélectionner un pourcentage plutôt qu'un nombre fixe d'instances afin de tenir compte de la diversité des jeux de données.

Pour évaluer la qualité des clusters, nous utilisons l'*Entropie* et la *Pureté*. L'entropie mesure la distribution des classes de prédiction dans un cluster, c'est-à-dire la capacité de l'algorithme à différencier les données qui n'ont pas la même classe "réelle". Une entropie parfaite signifie que toutes les instances de la même classe se trouvent dans les mêmes clusters. En outre, la pureté mesure la taille relative de la classe majoritaire dans un cluster afin d'évaluer sa prédominance sur les autres classes. Une pureté parfaite signifie que chaque cluster ne comporte qu'une seule classe. Ces deux mesures donnent des valeurs comprises entre 0 et 1. Un clustering parfait aura généralement une entropie égale à 0 et une pureté égale à 1. Ces mesures sont définies comme suit [Conrad et al., 2005] :

$$Entropy = \sum_{k=1}^{K} \frac{n_k}{n} \Big(-\frac{1}{\log q} \sum_{i=1}^{q} \frac{n_k^i}{n_k} \log \frac{n_k^i}{n_k} \Big) \qquad Purity = \sum_{k=1}^{K} \frac{1}{n} \max_i (n_k^i)$$

où C_k est un cluster particulier de taille n_k, q est le nombre de classes dans le jeu de données, K le nombre de clusters et n_k^i est le nombre d'instances de la $i$ème classe affectée au $k$ième cluster.

6.4.3 Résultats

Comparaison des clusters basés sur les données brutes et les influences

Lorsque l'on compare les clusters de données brutes aux clusters d'influence, pour toutes les instances, la Figure 6.10 montre que les clusters de données brutes ont une pureté plus faible et une entropie plus grande que les autres clusters, quels que soient les pourcentages, les méthodes XAI ou les performances du modèle. Les différences d'entropies sont encore plus marquées lorsque le modèle a une précision supérieure à 80%. Les clusters issus de données brutes sont de moins bonne qualité que les clusters issus d'influences, ce qui indique que le clustering d'instances sur la base de leurs influences donne de meilleurs résultats qu'un clustering basé sur les seules données brutes. En outre, comme prévu, lorsque les modèles sont moins précis, les clusters ont une pureté plus faible et une entropie plus grande, quels que soient les pourcentages de données ou de cluster. En effet, lorsque les performances du modèle sont médiocres alors que le modèle est correctement entraîné, cela peut indiquer que les données sont moins généralisables ou de moindre qualité. Cette hypothèse semble se refléter dans la qualité des clusters créés.

Si l'on ne tient compte que des *vraies instances* (les instances bien prédites par le modèle), la Figure 6.11a montre des résultats similaires à ceux de la Figure 6.10 : les clusters basés sur les influences sont de meilleure qualité que celles basées sur les données brutes (pour toutes les méthodes XAI, les pourcentages de sélection et la précision du modèle). La pureté et l'entropie sont presque parfaites, même avec de faibles pourcentages de sélection. Les clusters sont également de meilleure qualité avec les seules *vraies instances* qu'avec toutes les instances du jeu de données.

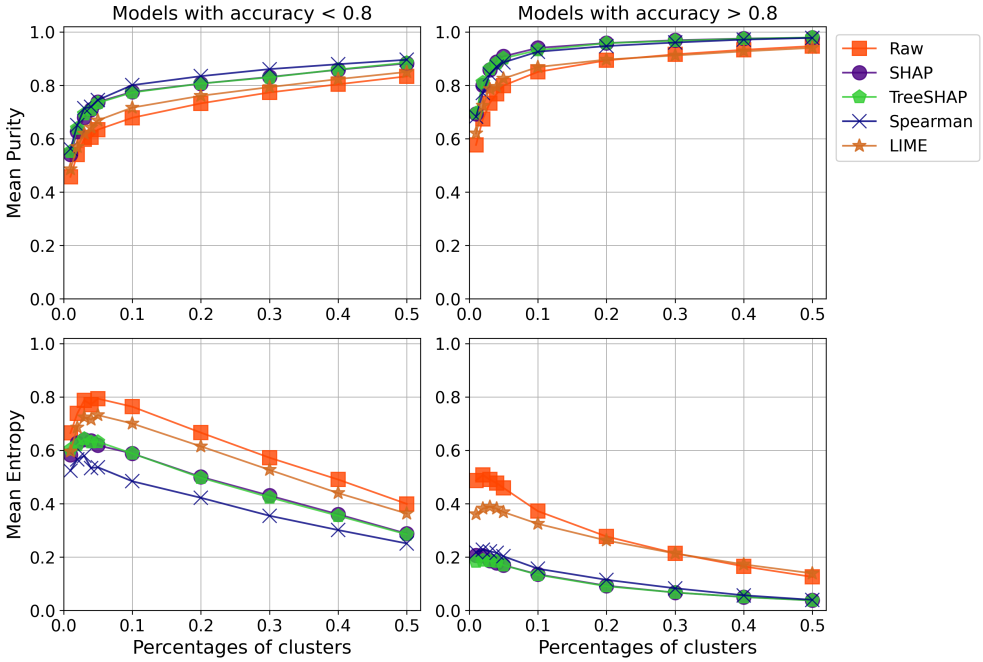


FIGURE 6.10 – Comparaison du clustering pour les méthodes XAI entraînées sur toutes les instances.

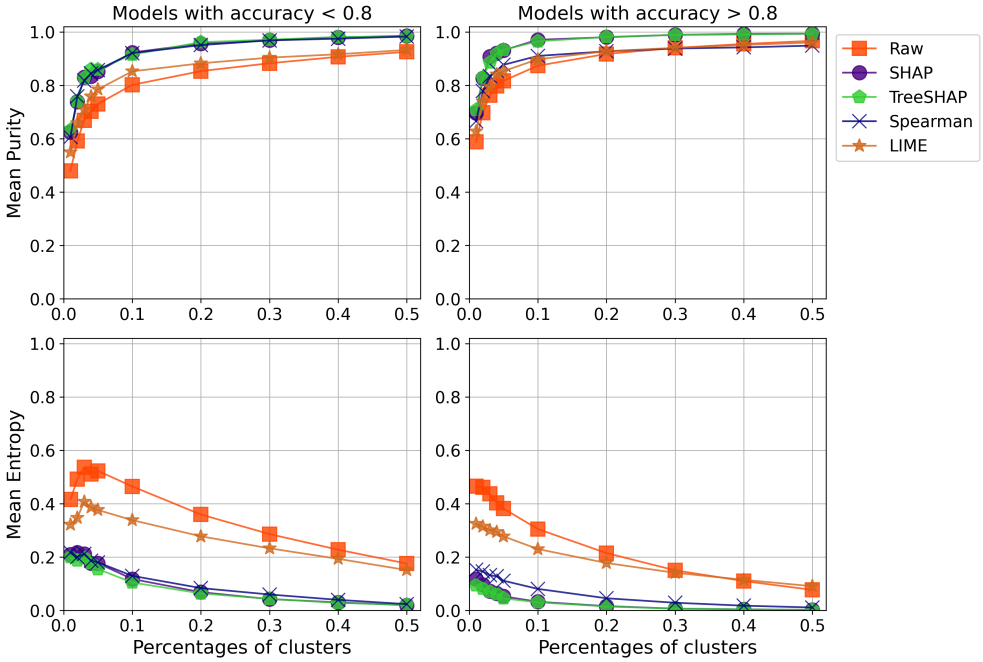
La Figure 6.11b ne prend en compte que les *fausses instances* (instances mal classées par le modèle). Globalement, la qualité des clusters est dégradée, en particulier la pureté. La pureté vérifiant la proportion de la classe majoritaire dans chaque cluster, le clustering des instances mal classées par le modèle diminue logiquement la pureté du cluster. Aucune méthode XAI ne semble avoir de bons résultats sur les petits pourcentages, même si elles ont toutes de meilleurs résultats que le clustering brut. Avec les *fausses instances*, nous analysons les cas où le modèle ne parvient pas à généraliser ou à décrire correctement les données. Comme les influences représentent la décision du modèle, les influences des instances mal classées peuvent être de moins bonne qualité que les vraies instances. Elles peuvent toutefois être représentatives de la raison pour laquelle le modèle ne généralise pas et ne comprend pas ces données. Ces clusters peuvent donc indiquer où se situent les problèmes dans les données ou dans le modèle.

Comparaison de l'impact d'utilisation des différents sous-groupes de données

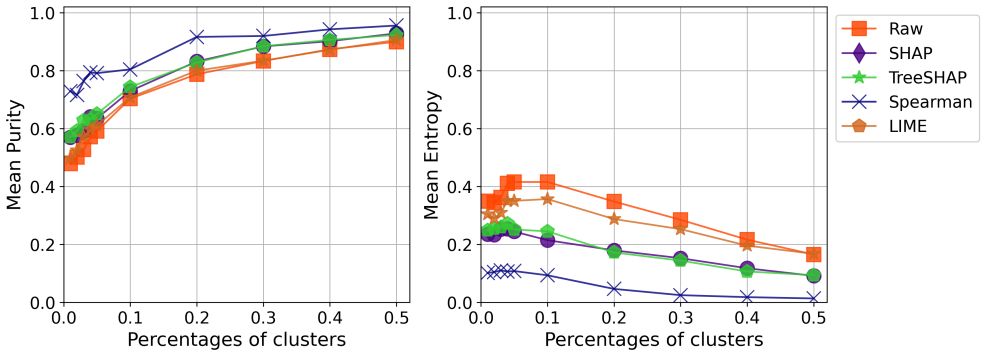
Nous montrons ici dans quelles circonstances des instances bien classées ou mal classées peuvent être utilisées pour produire des clusters de bonne qualité (ou non), notamment dans le pire des cas (précision dégradée sur un ensemble d'instances mal classées). Les Figures 6.12 et 6.13 montrent la qualité des clusters pour les trois modalités de données, avec les influences respectives de KernelSHAP et de la coalition basée sur Spearman.

La Figure 6.12 montre qu'il y a peu de différence dans la qualité des clusters entre les sous-groupes *toutes les instances* et *vraies instances* pour les modèles à haute précision. La pureté est élevée et presque égale pour les deux modalités, et les sous-groupes *toutes les instances* ont une entropie légèrement plus élevée. Les influences des *vraies instances* produisent des clusters presque parfaits, même avec des pourcentages de cluster faibles, et sont peu affectées par la précision du modèle. Étant donné que les modèles à haute précision ont moins de *fausses instances*, leurs influences peuvent ne produire que des bruits pour le clustering. En les supprimant, on obtient des résultats globaux légèrement meilleurs, car les clusters ont une meilleure entropie. Pour les modèles peu précis, les différences entre les sous-groupes sont plus marquées, probablement parce que la proportion de *fausses instances* est plus importante. Les sous-groupes *toutes les instances* et *vraies instances* présentent une différence de 0.4 en termes d'entropie et de 0.1 en termes de pureté pour la quasi-totalité des pourcentages. Les sous-groupes *fausses instances* ont également une pureté similaire et une meilleure entropie que les sous-groupes *toutes les instances*. La séparation des instances vraies et fausses, afin de les étudier séparément, produit des groupes plus homogènes et plus cohérents que le maintien de toutes les instances ensemble, en particulier pour les modèles à faible précision. Avec ces modèles, le nombre de fausses instances est plus élevé, et elles représentent souvent des comportements qui n'ont pas été détectés par le modèle.

Pour la méthode coalitionnelle basée sur Spearman, la Figure 6.13 révèle un comportement général similaire à celui de KernelSHAP en ce qui concerne la qualité des clusters en fonction des sous-groupes, en particulier pour les modèles à haute précision et pour les sous-groupes *vraies instances*. Toutefois, pour les modèles à faible précision et contrairement à



(a) "True" instances.



(b) "False" instances.

FIGURE 6.11 – Comparaison du clustering pour les méthodes XAI formées à partir (a) uniquement les "vraies" instances et (b) uniquement les "fausses" instances pour les modèles dont l'accuracy est inférieure à 0,8

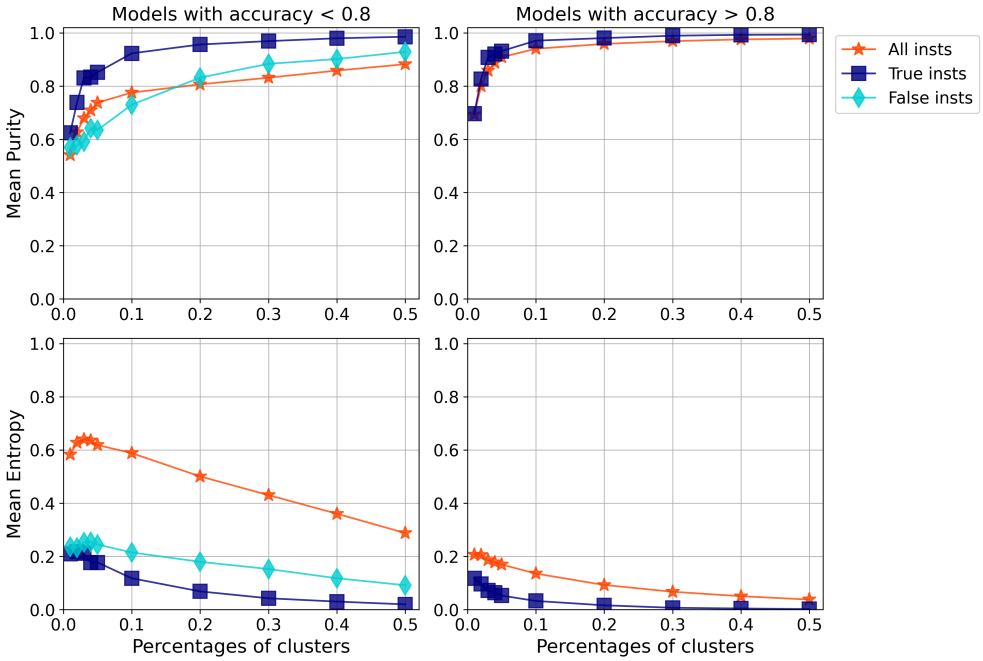


FIGURE 6.12 – Comparaison du clustering à partir des influences obtenues par KernelSHAP.

KernelSHAP, il existe des différences lorsque l'on utilise uniquement des instances erronées. Les sous-groupes *false instances* ont une pureté légèrement plus élevée et une entropie plus faible, en particulier pour les faibles pourcentages. L'utilisation différente des données d'entrée par les deux méthodes peut expliquer ce comportement. KernelSHAP utilise les données d'entrée pour produire des perturbations pour le modèle, en créant de nouvelles instances et en étudiant une zone plus large de l'espace de données que les seules données d'entrée (ici, les fausses instances). En revanche, la coalition de Spearman ne produit aucune perturbation et utilise les données d'entrée telles quelles pour expliquer le modèle. L'espace de données est alors plus petit et donc moins exhaustif. L'utilisation de fausses instances uniquement peut conduire à des influences plus précises pour ce sous-groupe, par rapport à l'utilisation de toutes les instances ou avec perturbations, d'où la différence entre les deux sous-groupes pour la méthode coalitionnelle avec Spearman et la différence avec KernelSHAP. En outre, pour les modèles à faible précision, les clusters issus des sous-groupes *vraies instances* et *fausses instances* sont meilleurs que les clusters issus de toutes les instances.

6.4.4 Discussions

Le clustering sur les influences XAI a donné de meilleurs résultats que le clustering sur les données brutes, indépendamment du pourcentage de clustering, de la méthode XAI ou de la performance du modèle. Les influences semblent contenir des informations permettant

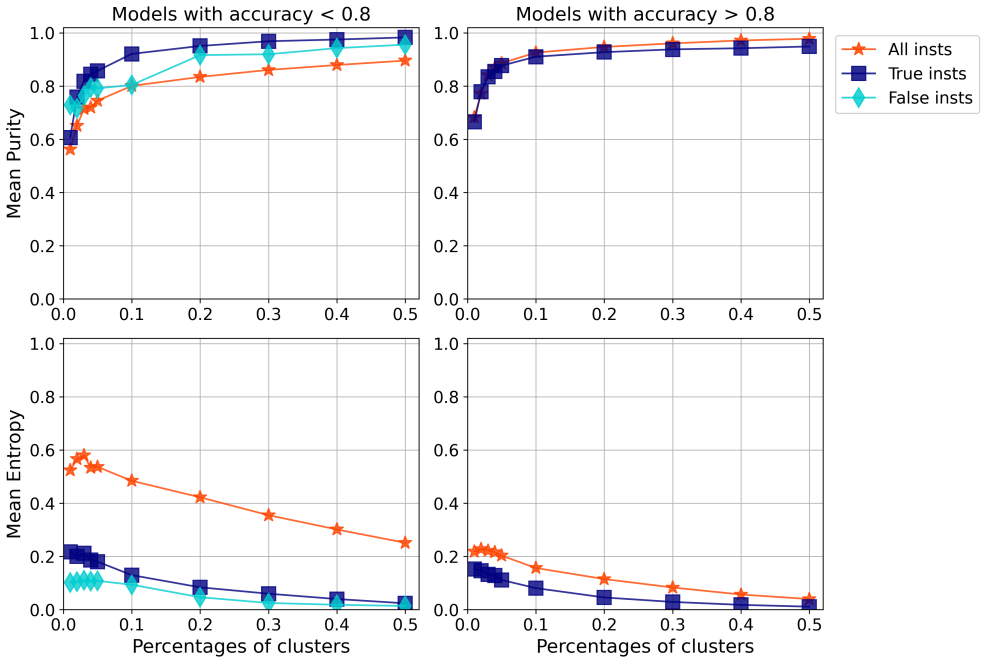


FIGURE 6.13 – Comparaison du clustering à partir des influences obtenues par les coalitions de Spearman.

un meilleur clustering, probablement en mettant en évidence les variables les plus significatives pour chaque instance ou en éliminant les bruits des données brutes. Cette conclusion semble cohérente avec les résultats de [Cooper et al., 2021] et renforcée par notre cadre général de clustering d'explications, évalué à l'aide de plusieurs méthodes XAI attributives sur une centaine de jeux de données.

La séparation des instances correctement et incorrectement classées par le modèle semble également donner de meilleurs résultats que le maintien de toutes les instances ensemble. Étant donné que les informations des deux sous-groupes sont différentes, chacun d'eux semble créer du bruit dans les informations de l'autre sous-groupe. En effet, les instances mal classées sont souvent des valeurs aberrantes ou des instances critiques du jeu de données. Leur comportement est différent du comportement général des données, alors que les instances correctement classées suivent le comportement détecté par le modèle. Cependant, comme certaines erreurs de classification peuvent résulter d'un biais dans un sous-groupe de données ou du comportement atypique de ce sous-groupe par rapport au jeu de données global, il est très intéressant de les étudier en priorité.

Une limite à la séparation de ces sous-groupes concerne la diminution de leur pertinence lorsque la précision du modèle augmente. En effet, le nombre de fausses instances diminue logiquement avec l'augmentation de la précision. La création d'un modèle XAI et de clusters avec un faible nombre d'instances n'a pas de sens et ne peut que conduire à une mauvaise compréhension des données. Cependant, à mesure que la précision augmente, les fausses instances deviennent principalement des valeurs aberrantes du jeu de données ou des instances biaisées plutôt que des sous-groupes avec leurs comportements à analyser. Leur petit nombre peut être analysé manuellement sans méthode de regroupement particulière.

6.5 Aide à l'analyse de données

Au vu des résultats obtenus dans la section précédente, montrant qu'un clustering basé sur les influences peut être bénéfique comparé aux seules données brutes, nous proposons dans cette section un usage plus global de ces types de clusters pour l'analyse de données.

Notre objectif est d'appliquer une approche bottom-up d'analyse exploratoire des données (sur un jeu de données médicales), à la fois sur les explications et les données brutes ; ceci afin de mettre en évidence et de comparer les connaissances récupérées dans les deux espaces de données. Nous montrons que les explications peuvent permettre un examen plus approfondi du jeu de données. Cette étude peut également montrer l'utilité de considérer les explications non seulement comme un résultat mais aussi comme un moyen.

6.5.1 Méthodologie

Jeu de données

Pour que les résultats soient reproductibles, nous utilisons un jeu de données en libre accès : *Acute Inflammation dataset*[3]. Le jeu de données porte sur l'inflammation aiguë et a été créé pour développer un système expert pour les maladies urinaires. Il se compose de 120 patients, décrits par six variables : Température *(35°C-42°C)*, Présence de nausées *(oui-non)*, Douleur lombaire *(oui-non)*, Poussée urinaire (besoin continu d'uriner, *oui-non*), Douleur mictionnelle *(oui-non)* et Brûlure de l'urètre, démangeaison, gonflement de l'orifice de l'urètre (abrégé en Brûlure de l'urètre, *oui-non*). Chaque patient peut souffrir de deux maladies différentes du système urinaire : l'inflammation aiguë de la vessie (AIUB) et la néphrite aiguë d'origine rénale. Les patients peuvent souffrir des deux maladies simultanément, de sorte que ce jeu de données constitue un problème multiclasses. Nous nous concentrons uniquement sur la maladie AIUB afin de se limiter à problème de classification binaire. Le personnel médical a défini l'AIUB comme *"une apparition soudaine de douleurs dans la région de l'abdomen et de la miction sous la forme de poussées urinaires constantes, de douleurs à la miction et parfois d'une absence de rétention d'urine. La température du corps augmente, le plus souvent sans dépasser 38°C. L'urine excrétée est trouble et parfois sanglante.* [Czerniak and Zarzycki, 2003]. Ceci forme ainsi la base de l'expertise de l'utilisateur pour l'analyse de données proposée dans cette section.

Méthode

La méthode proposée vise à analyser et à explorer des jeux de données par le biais d'un modèle prédictif et des influences. Basée sur un jeu de données d'intérêt constitué des dossiers médicaux des patients et du diagnostic de leur maladie, cette méthode permet de comprendre les interactions entre les variables des patients et la maladie. Elle est divisée en trois parties, inspirées par [Excoffier et al., 2022] :

1. La première consiste en une *modélisation prédictive*, afin d'évaluer le risque de maladie AIUB pour chaque patient sur la base de la compréhension de la relation statistique complexe du jeu de données. Un modèle XGBoost, par ensemble d'arbres boostés, est utilisé pour son efficacité. Nous utilisons une procédure de validation croisée imbriquée pour fournir une modélisation non biaisée (optimisation des hyperparamètres avec une validation croisée interne 5 fois) et pour évaluer les performances et calculer les explications locales (par le biais d'une validation croisée externe 5 fois).
2. La deuxième étape est l'*explication de la modélisation* pour fournir des explications individuelles de la prédiction pour chaque patient, correspondant aux facteurs de risque et aux facteurs de protection. TreeSHAP est utilisée pour calculer les explications par influence.

3. https://archive.ics.uci.edu/ml/datasets/Acute+Inflammations

3. La dernière étape consiste à *identifier des sous-groupes de patients similaires* pour découvrir des modèles locaux dans les données et expliquer les variables de ces sous-groupes. L'algorithme K-Medoids est utilisé pour la tâche de clustering afin d'assurer une robustesse contre les valeurs aberrantes, tandis que le nombre optimal de groupes a été choisi avec le score de Silhouette [Rousseeuw, 1987]. L'algorithme K-Medoids est utilisé pour les explications de l'influence de l'étape (2), avec l'avantage de prendre en compte les interactions non linéaires découvertes par le modèle tout en ayant toutes les variables à la même unité. Les règles de décision pour tous les groupes sont calculées avec l'algorithme Skope-Rules [Gardin et al., 2019]. Les règles sont calculées de manière à garantir une précision et un rappel parfaits de toutes les règles : toutes les instances du cluster respectent la règle, et toutes les instances respectant la règle appartiennent au cluster.

6.5.2 Résultats

6.5.3 Analyse des données brutes

Tests statistiques et de population

Le Tableau 6.1 présente les principales caractéristiques du jeu de données en utilisant uniquement les données brutes, avec les résultats des tests statistiques effectués sur les patients AIUB et non AIUB : Tests de Student pour les variables quantitatives et test du Khi-deux pour les variables qualitatives. Trois variables sont définies comme statistiquement significatives pour détecter l'AIUB : la douleur lombaire, la poussée urinaire et la douleur à la miction. Les patients souffrant de douleurs lombaires semblent présenter moins d'AIUB, tandis que les douleurs liées à la poussée urinaire et à la miction sont en corrélation avec un diagnostic d'AIUB.

Analyse par clustering et règles de décision

Pour créer des groupes homogènes de patients, l'application d'un clustering est la méthode la plus classique. Le nombre optimal de clusters était de 11, sur la base des scores de silhouette du Tableau 6.2. Le Tableau 6.3 montre les règles définies par Skope-Rules pour décrire chaque groupe. Les règles ont une médiane de 2,5 variables par règle. Toutes les règles ont une précision et un rappel parfaits avec un maximum de trois variables, ce qui est un nombre suffisamment faible pour faciliter l'interprétation de chaque règle. Les variables les plus utilisées sont la brûlure de l'urètre et la température, avec six occurrences distinctes, toutes deux précédemment définies comme non significativement discriminants pour le diagnostic de l'AIUB dans le Tableau 6.1. Un seul groupe, le groupe 2, n'utilise que des variables significativement discriminantes. De plus, le fait d'avoir onze clusters rend difficile la compréhension des règles et de ces groupes.

Nous proposons alors, en complément de cette analyse, d'étudier les explications produites par le modèle prédictif.

TABLE 6.1 – **Caractéristiques de la population** La moyenne et l'écart-type sont présentés pour les variables quantitatives, et les nombres et proportions pour les variables qualitatives binaires. Les p-values ont été ajustées à l'aide de la correction de Bonferroni pour contrôler le taux d'erreur.

		Total	**Non-AIUB**	**AIUB**	**p-value**
	Nb patients	120	61 (50.8)	59 (49.2)	
Quanti.	Temperature	38.72 (±1.8)	39.15 (±1.9)	38.29 (±1.7)	0.0552
Quali.	Nausea	29 (24.2)	10 (16.4)	19 (32.2)	0.4224
	Lumbar pain	70 (58.3)	51 (83.6)	19 (32.2)	<0.01
	Urine pushing	80 (66.7)	21 (34.4)	59 (100.0)	<0.01
	Micturition pain	59 (49.2)	10 (16.4)	49 (83.1)	<0.01
	Urethra Burning	50 (41.7)	21 (34.4)	29 (49.2)	0.8814

TABLE 6.2 – Score de Silhouette pour plusieurs nombres de clusters pour les données brutes.

K	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Raw	0.56	0.44	0.37	0.42	0.46	0.51	0.54	0.54	0.56	**0.57**	0.56	0.56	0.56	0.56

6.5.4 Analyse par explicabilité

Explications post-hoc locales

Un modèle XGBoost est produit et expliqué par la méthode SHAP. Le modèle avait une précision de 98,33%, une sensibilité de 96.72%, une spécificité de 100% et un score ROC AUC de 99.06%. La figure 6.14 montre les influences absolues moyennes du SHAP et les distributions des influences en fonction de la valeur de la variable. Les trois variables les plus importantes sont la douleur à la miction, la poussée urinaire et la température. La douleur à la miction et la poussée urinaire augmentent le risque de souffrir d'AIUB. Au contraire, une température plus élevée diminue la probabilité de souffrir d'AIUB. En particulier, la poussée urinaire semble également avoir moins d'impact sur la prédiction que l'absence de poussée urinaire. En revanche, les nausées et les brûlures de l'urètre ont peu ou pas d'impact sur les prédictions. Pour les nausées, SHAP décrit que leur présence augmente le risque d'AIUB pour certains patients et qu'un sous-groupe de patients est identifié.

La Figure 6.15 montre la distribution des influences uniquement pour les patients souffrant de nausées. Si l'on examine ces patients en détail, on constate qu'ils souffrent tous de douleurs lombaires, de douleurs mictionnelles et d'une température supérieure à 40°C (qui est plus élevée que la moyenne dans le jeu des données). Il semble qu'il existe un sous-groupe de patients présentant une relation étroite entre ces quatre variables. De plus, pour ce sous-groupe de patients, il existe une forte corrélation entre la variable Poussée urinaire et la présence d'AIUB : lorsqu'un patient a une poussée urinaire, il a une AIUB ; lorsqu'il n'a

TABLE 6.3 – Règles de décision pour les clusters basés sur les données brutes, avec le nombre de patients par cluster et le pourcentage moyen du risque AIUB.

N° groupe	Rules	Nb	Mean %
1	Nausea = 1 & Urine pushing = 0	10	45.6
2	Lumbar pain = 0 & Urine pushing = 0	10	10.7
3	Nausea = 1 & Urethra burning = 1	9	72.2
4	Temperature < 39.85 & Micturition pain = 0 & Urethra burning = 1	10	13.0
5	Lumbar pain = 0 & Urethra burning = 1	20	97.1
6	Temperature < 38.95 & Temperature > 36.65 & Urine pushing = 0	13	11.0
7	Temperature < 38.95 & Lumbar pain = 0 & Micturition pain = 0	10	59.9
8	Nausea = 1 & Urine pushing = 1 & Urethra burning = 0	10	73.6
9	Temperature > 39.85 & Nausea = 0 & Urethra burning = 1	11	11.2
10	Lumbar pain = 0 & Micturition pain = 1 & Urethra burning = 0	10	97.1
11	Temperature < 36.65 & Urethra burning = 0	7	11.2

pas de poussée urinaire, il n'y a pas d'AIUB. Il est probablement préférable d'étudier ce sous-groupe, car la variable nausée peut créer un biais en raison de sa forte association avec d'autres variables du jeu de données.

Analyse par clustering et règles de décision

Comme un sous-groupe a déjà été découvert, le clustering peut sans doute aider à en trouver d'autres aussi intéressants. Pour le clustering des influences par SHAP, le nombre optimal de clusters est fixé à 7, sur la base du score de silhouette du Tableau 6.4. Le Tableau 6.5 présente les règles définies par Skope-Rules pour les clusters basés sur les influences. Ces règles ont une médiane de deux variables par règle et se concentrent principalement sur les variables statistiquement pertinentes. Une seule règle comporte trois variables, et la variable la plus utilisée est la Poussée urinaire, avec cinq occurrences. Comme indiqué précédemment pour les "sous-groupes nausées", cette variable est la plus importante pour les patients souffrant de nausées (groupes 4 et 6) ainsi que pour les patients souffrant de douleurs lombaires (groupes 3 et 5). La variable Poussée Urinaire n'apparait pas dans les règles à l'exception des seuls groupes 2 et 7, les deux groupes les plus importants, où le risque d'AIUB est respectivement très faible et très élevé. Ces groupes peuvent être intéressants à étudier d'un point de vue médical afin de comprendre les caractéristiques des patients et pourquoi la variable Poussée Urinaire n'est pas la plus pertinente pour les distinguer des autres groupes. De même, bien que la douleur mictionnelle soit la variable la plus influente pour SHAP, elle n'est pas très

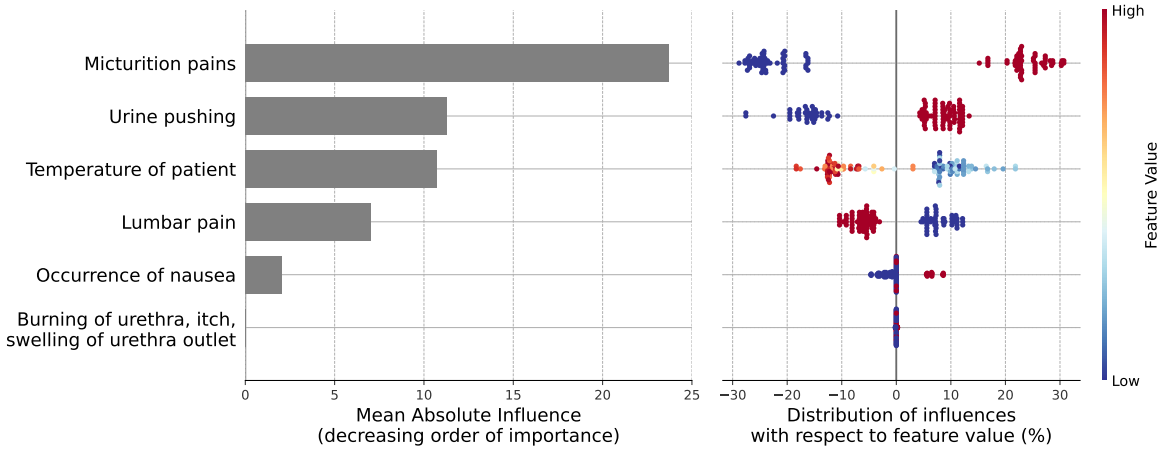


FIGURE 6.14 – Influences absolues moyennes de SHAP et distribution des influences pour le modèle entraîné.

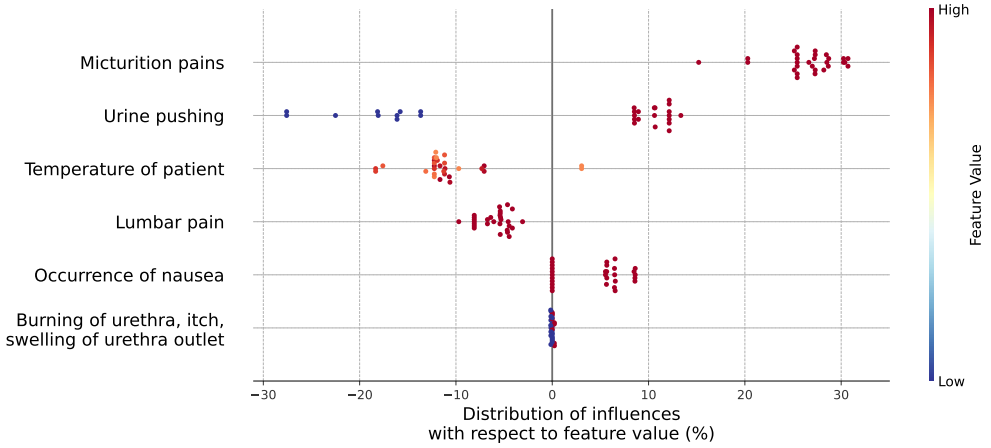


FIGURE 6.15 – Distribution des influences de SHAP pour les patients souffrant de nausées.

TABLE 6.4 – Score de silhouette pour plusieurs nombres de clusters pour les données XAI.

K	2	3	4	5	6	7	8	9	10	11	12	13	14	15
XAI	0.59	0.59	0.52	0.62	0.69	**0.76**	0.74	0.69	0.63	0.61	0.61	0.61	0.62	0.67

TABLE 6.5 – Règles de décision pour les clusters basés sur les influences, avec le nombre de patients par cluster et le pourcentage moyen de risque AIUB.

N° groupe	**Rules**	**Nb**	**Mean %**
1	Temperature <= 38.89 & Urine pushing = 0	20	11.0
2	Micturition pain = 0 & Urethra burning = 1	21	12.0
3	Lumbar pain = 0 & Urine pushing = 0	10	10.7
4	Nausea = 1 & Urine pushing = 0	10	45.6
5	Lumbar pain = 0 & Urine pushing = 1 & Micturition pain = 0	10	59.9
6	Nausea = 1 & Urine pushing = 1	19	73.0
7	Lumbar pain = 0 & Micturition pain = 1	30	97.2

présente dans les règles, principalement parce que cette variable semble remplacée par la variable Nausée dans les clusters, étant donné qu'il existe un lien étroit entre la nausée et la douleur mictionnelle.

6.5.5 Discussions

Dans cette étude, les méthodes d'analyse des données brutes et par explications ont permis de détecter des patterns dans les données, des sous-groupes de patients et des informations sur la relation entre la maladie AIUB et les symptômes des patients. Outre les informations connues dans la littérature et trouvées dans l'analyse des données brutes, l'analyse des données basée sur l'explication a permis d'identifier les facteurs de risque et de protection de manière plus concise. Les règles sont principalement basées sur des variables statistiquement significatives, repérant de nouvelles interactions entre les variables et la variable cible, par rapport à une simple analyse des données brutes. Le nombre réduit de clusters et de variables dans chaque règle simplifie également la compréhension des sous-groupes de patients et la relation de chaque variable avec le risque d'AIUB. Avec les données brutes, plusieurs groupes ont des pourcentages moyens similaires de risque d'AIUB et des patients presque identiques. Les différences entre ces groupes sont souvent basées sur des variables qui ne sont pas importantes pour la détection de l'AIUB. Ce comportement peut être bénéfique pour étudier le jeu de données en profondeur, moins pour découvrir les variables qui ont un véritable impact sur le diagnostic de la maladie et pour capturer des connaissances concises. La concision fournie par les influences facilite également l'affectation d'un nouveau patient à un sous-groupe de patients afin d'étudier leur maladie et leurs facteurs de risque. Cet avantage provient de la capacité de la modélisation par apprentissage automatique à saisir des relations plus complexes

que les méthodes statistiques traditionnelles. Enfin, les données d'explication ont permis de découvrir des sous-groupes de patients pertinents, notamment ceux souffrant de nausées. Ce sous-groupe présente des relations fortes entre plusieurs variables, et la présence d'AIUB est basée uniquement sur la variable Poussée urinaire, ce qui rend son étude intéressante pour comprendre les mécanismes de la maladie chez certains patients. La découverte de ce type de sous-groupe peut aider à étudier les biais du jeu de données, en particulier autour de la variable Nausée.

6.6 Conclusion

6.6.1 Bilan

Dans ce chapitre nous avons pu montrer l'intérêt de voir les explications post-hoc locales attributives comme un nouvel espace de données à analyser et exploiter. Nous l'avons décliné sur quatre tâches liées à de l'analyse prédictive, incluant la sélection de modèles, la sélection de variables, la sélection d'instances et l'analyse de données plus généralement.

Nous avons ainsi, tout au long de ces cas d'usage, apporté les contributions suivantes :

— un cadre itératif, à base d'explications, entre l'utilisateur et la recommandation de modèles prédictifs afin de sélectionner un modèle prédictif adéquat et personnalisé.
— l'ajout de la notion d'explication pour la sélection de variables, associé à des métriques mesurant différents profils d'explication.
— un cadre pour la sélection d'instances, basée sur les explications, afin de produire des clusters d'intérêt.
— une méthode d'analyse mêlant l'analyse traditionnelle par données brutes et l'analyse de clusters d'explications associées à des règles de décision.

Concernant la **sélection de modèles**, nous proposons un cadre "human-in-the loop" visant à aider dans chaque étape l'utilisateur dans sa recherche du modèle prédictif répondant à son besoin, à partir d'explications de prédiction. Nous avons montré qu'à l'aide de ces explications, l'utilisateur peut repérer d'éventuelles erreurs dans la formation de son modèle, choisir entre différents workflows possibles et des variables pertinentes à considérer dans son jeu de données. De plus, une fois le modèle formé, l'utilisateur dispose d'un environnement de type bac à sable pour expérimenter et expliquer de nouvelles données.

L'intégration de l'explication comme troisième dimension à prendre en compte lors d'une **sélection de variables** (en plus de l'accuracy et du taux de rétention) a permis de montrer la variabilité des profils d'explication en fonction de la méthode de sélection de variables utilisée. Pour cela, nous avons proposé plusieurs métriques d'intérêt permettant de mesurer les différences d'explications en prenant en compte, en particulier, les rangs et les taux d'influence entre les variables. Aucune méthode de sélection de variables n'est efficace dans tous les cas et nous préconisons de choisir une méthode de sélection de variables en

fonction des préférences de l'utilisateur sur les trois dimensions. Cependant, la précision pourrait être utilisée comme condition préalable au calcul des explications, car l'équilibre entre les trois dimensions n'a de sens que lorsque la précision se situe dans une fourchette acceptable.

Au sujet de la **sélection d'instances**, nous avons présenté un cadre permettant de produire des clusters d'explications plus homogènes et globalement de meilleure qualité que seulement basé sur les données brutes. Ces clusters peuvent être vus comme une base pour une aide à de futures analyses, en fournissant des informations plus précises sur les relations entre données et modèle prédictif. Les centroïdes peuvent notamment servir à proposer, à l'utilisateur, les instances du jeu de données, ayant de fait un pouvoir explicatif particulier. L'intégration de l'analyse de ces centroïdes dans notre cadre de sélection de modèles pourrait d'ailleurs renforcer cet effet "human-in-the-loop" mis en avant. D'autre part, nous avons mis en évidence l'intérêt de dissocier les explications liées aux instances bien ou mal classées par le modèle, en représentant simultanément les sous-groupes de données les plus importants ainsi que des comportements liés à des valeurs aberrantes.

Notre dernière contribution, consacrée à l'**aide à l'analyse de données**, exploitant les clusters d'explications obtenus par la sélection d'instances, est décrite au travers d'un cas d'étude médicale. L'analyse des données brutes et des explications permettent de détecter des patterns complémentaires dans les données, les sous-groupes d'individus ainsi que des informations liées aux la relation entre variables. L'analyse des données basée sur les explications a permis d'identifier des facteurs de risque et de protection de manière plus concise. Les règles de décision sont principalement basées sur des variables statistiquement significatives, ajoutant des interactions entre les variables non détectées par une analyse des données brutes. Avec les données brutes, les clusters générés étaient surtout basés sur des variables peu importantes pour la détection de la maladie visée (AIUB). Ce comportement peut être utile pour étudier un jeu de données en profondeur, moins pour découvrir les variables qui ont un véritable impact sur le diagnostic de la maladie. La concision offerte par les influences facilite également l'affectation possible d'un futur patient à un sous-groupe de patients afin d'étudier leur maladie et leurs facteurs de risque. Cet avantage provient de la capacité du modèle prédictif à saisir des relations plus complexes que les méthodes statistiques traditionnelles.

6.6.2 Perspectives

Les perspectives de recherche liées à l'usage des explications comme aide à la réalisation de tâches prédictives sont naturellement nombreuses.

L'aspect bac à sable développé dans le cadre de la sélection de modèles mériterait d'être davantage développé. En ajoutant des fonctionnalités semblables à celles d'un **outil de type "What-if"** [Wexler, 2018], combinées à de l'interaction avec des explications de prédictions, cela permettrait à l'utilisateur d'expérimenter facilement des hypothèses plus complexes en comprenant mieux les interactions entre les variables et le modèle. Nos démarches proposées

pour l'analyse de données et la sélection d'instances, basées sur les explications, sont certainement des marqueurs de la faisabilité de ce type d'approche. Plus spécifiquement, le domaine médical pourrait aussi se satisfaire de ce type d'outil, notamment pour la construction de cohortes virtuelles. Il s'agirait, par exemple, de constituer des groupes de patients virtuels répondant à certains critères définis par un chercheur en santé. Si ses hypothèses sont vérifiées, notamment à l'aide de l'explicabilité, il pourrait alors chercher à la valider réellement et, sinon, identifier les raisons pour lesquelles le groupe de patients virtuel ne répondait pas à son besoin.

L'explicabilité comme dimension supplémentaire à la sélection de variables mène à considérer quelques axes d'améliorations possibles. Il faudrait tout d'abord pouvoir tester l'**impact de la sélection de variables à l'aide d'autres méthodes d'explications**. Au niveau local, les méthodes d'explications de type TreeSHAP, adaptées aux modèles prédictifs basés arbres, semblent être les seules à pouvoir être utilisées sur des jeux de données de relatives grandes dimensions. Il pourrait aussi être intéressant d'évaluer l'impact d'une sélection de variables sur les explications contrefactuelles produites (voir Chapitre 2.5). Dans l'idée de proposer un système de recommandation basée sur les trois dimensions que sont l'accuracy, le taux de rétention et donc les explications (comme évoqué en Section 6.3 ; à l'aide d'un front de Pareto par exemple), la **prise en compte de l'expertise utilisateur** est certainement aussi une voie prometteuse. Comme nous l'avons vu, chaque méthode de sélection de variables impacte différemment les explications produites. De plus, comme nous l'avons aussi constaté dans le Chapitre 5, chaque méthode d'explication attributive a ses propres capacités à détecter de manière plus au moins fine les corrélations entre variables et produit des explications plus ou moins lisibles (dont nous avons proposé une première définition). A l'aide de métriques appropriées, il s'agirait donc de proposer l'ensemble de variables le plus à même d'être compris par l'utilisateur en fonction de son niveau d'expertise dans les données. Il s'agit, finalement, de concilier une sélection de variables à la fois dirigée par les données seules (mesurées par l'accuracy et le taux de rétention) mais aussi dirigée par les explications.

Les clusters, générés par notre cadre de sélection d'instances, peuvent permettre de produire des instances représentatives des explications locales possibles (liées aux modèles prédictifs). Nous pourrions alors **comparer les instances identifiées à celles produites par les méthodes d'explication basées sur les exemples** (incluant les contrefactuelles ainsi que les prototypes/critiques et les exemples contradictoires notamment). Nos expérimentations étaient aussi limitées aux seules évaluations objectives et ne considéraient pas d'évaluation utilisateur. L'évaluation utilisateur dans le domaine de l'explicabilité n'est pas sans poser de difficulté à l'heure actuelle, comme évoquée dans le Chapitre 2.5.3, mais nous pourrions imaginer une campagne d'évaluation limitée aux instances sélectionnées et mesurer à quel point celles-ci peuvent servir à mieux comprendre un jeu de données et le modèle prédictif associé (ainsi que les éventuels soucis détectés).

D'autre part, les clusters basés sur les influences peuvent sans doute aider à comprendre *pourquoi* un modèle est faux et pas seulement *où* le modèle est faux, par exemple en **facilitant la détection de bais dans les données, ou pour améliorer la qualité des données**. La détection de biais dans les données est un enjeu important et l'explicabilité pourrait alors aider à raffiner un modèle en éliminant les instances ou groupes d'instances problématiques. La même stratégie pourrait être reprise avec une sélection de variables dont les explications assurent une analyse non biaisée. Au vu du nombre possiblement conséquent de clusters d'explication générés (si le jeu de données est important), une solution possible serait d'**offrir à l'utilisateur les moyens de les explorer**. Une approche facilitant cette exploration peut être basée sur une solution hiérarchique des clusters produits. Ainsi, à un niveau plus ou moins agrégé de la hiérarchie, il serait possible d'identifier des instances pertinentes en fonction d'un niveau d'explication plus au moins grossier.

Finalement, les solutions variées que nous proposons dans ce chapitre (sélection de modèles, sélection de variables, sélection d'instances) ont aussi pour ambition d'être proposées dans un **cadre unique d'aide à l'analyse de données**. Ce cadre unique, résolument orienté "human-in-the-loop", permettrait ainsi de considérer les feedback de l'utilisateur afin de mieux personnaliser chaque étape de son analyse de données : le choix du modèle, les variables à considérer et les instances à explorer/analyser en priorité.

Chapitre 7

Valorisation des travaux

Ce chapitre présente une synthèse de mes activités d'encadrements et de diffusion de la recherche, depuis que je suis Maitre de conférences en 2016. Les projets de recherche sont également détaillés.

7.1 Encadrements et publications

Depuis 2016, j'ai assuré les co-encadrements des activités de recherche d'étudiants en Master (M1 et M2) ainsi que d'étudiants en doctorat et en post-doctorat, dont voici les détails :

— 7 thèses dont 3 soutenues en 2019, 2020 et 2023.
— 1 post-doctorat.
— 2 stages de recherche en M1 et 2 stages de recherche en M2.

	Nombre	%Encadrement
Encadrement d'étudiants en Master 1	2	
Encadrement d'étudiants en Master 2	2	
Encadrement d'étudiants en Post-doctorat	1	50%
Encadrement d'étudiants en Doctorat		
Ayant soutenus	3	10%, 50%, 50%
En cours	4	33%, 33%, 33%, 33%

7.1.1 Encadrements de thèses de doctorat

J'ai co-encadré trois étudiants dont les thèses ont été soutenues :

Elodie ESCRIVA : *"Interprétation de modèles et de résultats de prédiction : cas du secteur santé et de l'aide à l'appropriation par les médecins."*

— Thèse de Doctorat de l'Université Toulouse Capitole, Laboratoire IRIT.
— Projet CIFRE avec la société Kaduceo.

— Débuté en Février 2021, soutenue en Mars 2024.
— Encadrement à 50% - Directeur de Recherche : Chantal Soulé-Dupuy.

Cette thèse vise à expliciter un modèle prédictif en identifiant les localités dans les données les plus intéressantes, à destination de professionnels de santé.

Gabriel FERRETTINI : *"Système adaptatif pour l'aide à la conception de processus d'analyse"*

— Thèse de Doctorat de l'Université Toulouse Capitole, Laboratoire IRIT.
— Bourse Ministère.
— Débuté en Octobre 2017, soutenue en Mars 2021.
— Encadrement à 50% - Directeur de Recherche : Chantal Soulé-Dupuy.
— Devenir : qualifié pour les postes de Maitre de conférences. Actuellement en Post-doctorat.

Cette thèse a eu pour objectif de proposer un système itératif de recommandation de modèles prédicrifs, basé sur le contexte utilisateur et reprenant une démarche AutoML. Les itérations sont assurées par des explications locales aux prédictions permettant à l'utilisateur de réaliser des tâches de feature engineering, par exemple.

Franck BOIZARD : *"Application de la biologie des systèmes pour l'identification de marqueurs moléculaires des maladies rénales dans les fluides biologiques"*

— Thèse de Doctorat de l'Université Paul Sabatier - Toulouse 3, Laboratoire INSERM/IRIT.
— Projet région Occitanie.
— Débuté en Octobre 2016, soutenue en Octobre 2019.
— Encadrement à 10% - Directeur de Recherche : Joost-Peter Schanstra (INSERM).
— Devenir : Ingénieur d'étude en statistiques.

Cette thèse a permis de produire un graphe mettant en relation les interactions en protéines de la littérature. Cela a permis de raisonner et détecter, notamment via la détection de communautés dans les graphes, les interactions problématiques à l'orgine de maladies rénales.

Je participe actuellement à un projet région et PIA3 (collaborations IRIT/RESTORE) et dont je suis co-porteur avec Chantal Soulé-Dupuy et Paul Monsarrat :

Haomiao WANG : *"PREDINSIGHT 4D - Système d'aide à l'analyse de jeux de données multimodaux pour la détection de biomarqueurs visuels et biologiques du vieillissement."*

— Thèse de Doctorat de l'Université Paul Sabatier - Toulouse 3, Laboratoire RESTORE.
— Projet région Occitanie.
— Débuté en Octobre 2021 pour une soutenance prévue en Octobre 2024.
— Encadrement à 33% - Directeur de Recherche : Paul Monsarrat (RESTORE).

Cette thèse se consacre notamment à l'étude de la sélection de variables et son impact sur les explications des prédictions, dans le cas de données de vieillissement.

Emmanuel DOUMARD *"How to explore physiological aging ? A new framework for an in-depth explainability of machine learning models."*

— Thèse de Doctorat de l'Université Paul Sabatier - Toulouse 3, Laboratoire IRIT/Laboratoire RESTORE
— Projet PIA3 EUR CARE.
— Débuté en Octobre 2021 pour une soutenance prévue en Octobre 2024.
— Encadrement à 33% - Directeur de Recherche : Chantal Soulé-Dupuy.

Cette thèse a pour objet la proposition d'un cadre d'exploration d'explications de prédictions, vues comme une nouvelle source d'analyse et validées sur des données de vieillissement.

Je suis également impliqué dans un projet CIFRE :

Robin CUGNY : "Expliquer et valider par l'exemple un modèle : Application à un Sosie Virtuel Projectif"

— Thèse de Doctorat de l'Université Paul Sabatier - Toulouse 3, Laboratoire IRIT.
— Projet CIFRE avec la société SolutionData Group.
— Débuté en Septembre 2021 pour une soutenance prévue en Septembre 2024.
— Encadrement à 33% - Directeur de Recherche : Max Chevalier.

Cette thèse a pour objectif de valider un modèle prédictif au travers de modèles d'explications fournies à l'utilisateur.

Je suis aussi impliqué dans un co-encadrement en co-tuelle avec l'Académie des sciences d'Arménie. Malheureusement, suite à la guerre en Arménie de 2020, la thèse est actuellement suspendue.

Elisa GYUlGYULYAN : "Quality Measures for User-centric System Results in an OLAP Context"

— Thèse de Doctorat de l'Université Paul Sabatier - Toulouse 3, Laboratoire IRIT.
— Projet de co-tutelle avec l'Académie des Sciences d'Arménie.
— Débuté en Octobre 2017, actuellement suspendue.
— Encadrement à 33% - Directeur de Recherche : Franck Ravat.

Cette thèse a pour objectif de mesurer et avertir l'utilisateur quant aux problèmes de qualité des données dans un contexte de Big Data.

7.1.2 Encadrement de post-doctorat

Philippe ROUSSILLE : "Vers un environnement logiciel d'aide à la simulation pour la prédiction des pathologies associées à l'âge"

— De Février 2018 à Février 2019
— Encadrement à 50% avec Chantal Soulé-Dupuy

Ce post-doctorat avait pour objectif de valoriser le travail de thèse de William Raynaut, dans l'équipe, en construisant l'environnement logiciel pour la recommandation de modèles prédictifs à l'aide d'une dissimilarité entre méta-attributs.

7.1.3 Encadrements de stages de recherche de Master 2 et Master 1

Les encadrements de stages de recherche ont permis soit d'appuyer des travaux de thèse en cours (Zheng LIU, Tom LEFRERE, Manon MARTIN), soit de confirmer la faisabilité de perspectives de recherche (Yacine MOKHTARI). Les détails de ces stages sont disponibles ci-dessous.

Yacine MOKHTARI (Master 2) : "L'explicabilité au service du clustering interactif"

— Débuté en Mars 2023, soutenue en Septembre 2023

— Encadrement à 50% avec Olivier Teste

Ce stage avait pour objectif de montrer l'intérêt des explications locales sur les solutions existantes de clustering actif (en particulier COBRA).

Zheng LIU (Master 2) : "Optimisation de la recherche de jeux de données similaires par meta-analyse"

— Débuté en Avril 2018, soutenue en Septembre 2018.

— Encadrement à 100%.

Ce stage avait pour objectif d'optimiser les comparaisons entre jeux de données à l'aide d'une mesure de dissimilarité (basée sur des méta-attributs), proposée lors de la thèse de William Raynaut, précédent doctorant dans l'équipe.

Tom LEFRERE et Manon MARTIN (Master 1) : "Experimentations pour la sélection d'instances à l'aide d'explications locales"

— Débuté en Mai 2023, fin en Aout 2023 (3 mois chacun)

— Encadrement à 50% avec Nicolas Labroche (Université de Tours)

Ce stage avait pour objectif de compléter les expérimentations menées lors de la thèse de Elodie ESCRIVA, en montrant l'intérêt de l'usage des explications pour la sélection d'instances, à l'aide d'autres algorithmes de clustering.

7.1.4 Publication des travaux

Type	**Depuis 2016**	**Elements emblématiques**
Revues internationales avec comité de sélection	8	Dont IS, ISF, JBHI
Conférences internationales avec comité de sélection et actes	13	Dont CIKM, DOLAP, ADBIS
Co-édition d'actes de conférences nationales avec comité de sélection	2	EDA
Conférences nationales avec comité de sélection et actes	5	Dont EGC, INFORSID
Brevet déposé	1	

Parmis les revues internationales avec comité de sélection, trois publications ont été acceptées dans des journaux en biologie/santé (avec un fort usage de l'apprentissage automatique et

de l'explicabilité), en particulier à Aging Cell (Q1, IF= 11), Scientific reports (Q1, IF=4.99) et Journal of personalized medicine (Q2, IF=3.4).

Voici le détail des publications, depuis 2016 :

— Revues internationales avec comité de rédaction : [Wang et al., 2023], [Doumard et al., 2023], [Bernard et al., 2023], [Ferrettini et al., 2022], [Monsarrat et al., 2022], [Boizard et al., 2021], [Drushku et al., 2019], [Astsatryan et al., 2018].

— Conférences internationales avec comité de sélection et actes : [Escriva et al., 2023a], [Escriva et al., 2023b], [Cugny et al., 2022], [Doumard et al., 2022], [Excoffier et al., 2022], [Zhao et al., 2021], [Ferrettini et al., 2020b], [Ferrettini et al., 2020a], [Ferrettini et al., 2020c], [Gyulgyulyan et al., 2019], [Gyulgyulyan et al., 2018], [Drushku et al., 2017], [Raynaut et al., 2017a].

— Conférences nationales avec comité de sélection et actes : [Cugny et al., 2023], [Escriva et al., 2022], [Drushku et al., 2020], [Ferrettini et al., 2019], [Aligon et al., 2016]

7.2 Projets et collaborations

7.2.1 Projets

Je participe actuellement à un projet région et PIA3 (collaborations IRIT/RESTORE) et dont je suis co-porteur avec Chantal Soulé-Dupuy et Paul Monsarrat :

— Oct. 2021 – Oct. 2024 : Projet région Occitanie "PREDINSIGHT 4D - Système d'aide à l'analyse de jeux de données multimodaux pour la détection de biomarqueurs visuels et biologiques du vieillissement.", pour le financement d'une thèse. Cette thèse se consacre notamment à l'étude de la sélection de variables et son impact sur les explications des prédictions, dans le cas de données de vieillissement.

— Oct. 2021 – Oct. 2024 : Projet PIA3 EUR CARE, "How to explore physiological aging ? A new framework for an in-depth explainability of machine learning models." pour le financement d'une thèse. Cette thèse a pour objet la proposition d'un cadre d'exploration d'explications de prédictions, vues comme une nouvelle source d'analyse et validées sur des données de vieillissement.

Je suis également impliqué dans deux projets CIFRE :

— Fev. 2021 – Fev. 2024 : Thèse CIFRE "Interprétation de modèles et de résultats de prédiction : cas du secteur santé et de l'aide à l'appropriation par les médecins", avec un montant de 18K€ de la société Kaduceo. J'en suis le co-porteur avec Chantal Soulé-Dupuy (définition du sujet, négociation financière avec la société). Cette thèse vise à expliciter un modèle prédictif en identifiant les localités dans les données les plus intéressantes, à destination de professionnels de santé.

— Oct. 2021 – Fev. 2024 : Thèse CIFRE "Expliquer et valider par l'exemple un modèle : Application à un Sosie Virtuel Projectif" avec un montant de 30K€, de la société SolutionData Group. Max Chevalier en est le porteur et j'en ai co-rédigé le sujet de

thèse. Cette thèse a pour objectif de valider un modèle prédictif au travers de modèles d'explications fournies à l'utilisateur.

7.2.2 Collaborations

Depuis mon arrivée à l'université Toulouse Capitole, je me suis toujours appliqué à développer des collaborations aussi bien au niveau local (collaboration entre laboratoires toulousains), au niveau national (par mon investissement scientifique dans le GDR Madics) et au niveau international (collaborations avec l'Arménie et l'Imperial College of London).

Au niveau local :
Au niveau local, ma contribution la plus significative est le développement de collaborations pluridisciplinaires avec le laboratoire RESTORE. Deux équipes de l'IRIT sont membres partenaires de ce laboratoire : l'équipe REVA, par sa spécialité dans les modèles agents, ainsi que l'équipe SIG pour ses spécialités en gestion et analyse de données. J'ai, en particulier, pris part avec Chantal Soulé-Dupuy, à définir les axes importants de recherche liés à l'équipe SIG et que RESTORE a souhaité développer depuis sa création, à savoir :

- L'aide à la gestion de données : les 4 équipes de RESTORE possèdent une multitude de données hétérogènes et volumineuses. L'objectif de ce laboratoire étant de rendre mieux interopérables ces données entre les équipes, l'importance d'une bonne gestion des données, facilement accessible, est donc un enjeu crucial. Il a ainsi été proposé l'élaboration d'un lac de données permettant, à terme, de proposer de meilleures pratiques dans l'usage des données et de leurs analyses.
- L'aide à l'analyse de données : Le machine learning est un domaine prenant de plus en plus d'importance dans les recherches en biologie/santé. Pour cette raison, il a été proposé d'orienter les recherches sur la recommandation de modèles prédictifs, ainsi que de solutions d'explicabilité permettant de mieux comprendre les modèles utilisés ainsi que les résultats produits. Je suis, depuis, le correspondant scientifique, auprès de l'équipe SIG, pour ce laboratoire. Cette responsabilité implique, actuellement, les tâches suivantes :
  - Organisation de séminaires en informatiques proposés à RESTORE (en gestion et analyse de données) par des membres de l'équipe SIG
  - Participation aux journées scientifiques de RESTORE en juin 2022, avec l'animation d'une session sur le machine learning et l'explicabilité
  - Participation à un consortium pour une réponse à un appel à projet I-Demo, mêlant entreprises dans le secteur de la santé, le laboratoire RESTORE et l'équipe SIG de l'IRIT (projet non encore soumis)

Depuis la mise en place de cette collaboration en 2021, j'ai notamment travaillé avec Paul Monsarrat et Philippe Kemoun, tous deux PU-PH à RESTORE. Cette collaboration a permis :

- Le co-encadrement, en cours, de 2 thèses (Haomiao Wang et Emmanuel Doumard) dans le domaine de l'explicabilité en machine learning

— La réalisation de 5 articles de recherche [Monsarrat et al., 2022, Doumard et al., 2022, Doumard et al., 2023, Wang et al., 2023, Bernard et al., 2023] (dont trois revues Q1 à Information Systems, Journal of Biomedical and Health Informatics et Aging Cells)
— Un dépôt de brevet, assuré par INSERM Transfert. Ce brevet a pour but de prédire l'âge biologique personnalisé à l'aide d'une démarche d'analyse originale basée sur l'explicabilité post-hoc additive (via la méthode SHAP).

Au niveau national :
J'ai développé les collaborations suivantes :

— Avec Christophe Denis (Sorbonne Université, LIP6), pour l'organisation de l'atelier EXPLAIN'AI en 2021 à Blois
— Avec Nicolas Labroche (Université de Tours, LIFAT), pour l'organisation de plusieurs ateliers (EXPLAIN'AI, EXEC-MAN) et le co-encadrement prochain de 2 stagiaires de M1 afin de développer nos recherches en explicabilité
— Avec Cyril de Runz (Université de Tours, LIFAT) pour l'organisation de la journée interassociations EGC/Inforsid

Au niveau international :
J'ai développé deux collaborations ces dernières années :

— Avec Hrachya Astsatryan, de l'académie des sciences d'Arménie, pour une co-tutelle de thèse sur la qualité des données dans un contexte de Big data [RI6][CI8][CI9].
— Avec Matthieu Komorowski, de l'Imperial College of London (Faculté de médecine, Département de chirurgie et de cancérologie), pour un travail commun sur l'explicabilité dans les méthodes par apprentissage par renforcement, avec une application dans un traitement optimal de la septicémie. Emmanuel Doumard, doctorant que je co-encadre, y a effectué un stage de 4 mois, entre mai et septembre 2023 dans ce laboratoire.

7.3 Animation scientifique

7.3.1 Comités de programme et de lecture de revues et conférences

Je participe tous les ans à des comités de programmes en conférence et ateliers internationaux et nationaux ainsi qu'à des comités de lecture de revues internationales dans les domaines de l'explicabilité, l'aide à la décision, les bases de données et les systèmes d'information. En moyenne, sur ces 4 dernières années, je participe tous les ans à :

— 1 comité de lecture de revues internationales (Information Systems, Data and Knowledge Engineering),
— 2 comités de programme de conférences ou ateliers internationaux (DOLAP, ICEIS),
— 3 comités de programmes de conférences ou ateliers nationaux (EXLAIN'AI, EDA et INFORSID)

7.3.2 Comités de suivi et jurys de thèse

J'ai participé au suivi, dans le cadre de comités de thèses, de deux doctorants en 2022 et 2023. Je suis également invité prochainement à deux jurys de thèse, en tant qu'examinateur.

Année	Années de thèse	Etudiant	Domaine	Directeurs de thèse
2022 et 2023	2ème et 3ème année	Adulam Jeyasothy	Explications contrefactuelles	Marie-Jeanne Lesot et Christophe Marsalat (Sorbonne Université)
2022	2ème année	Benjamin Beltzung	Application de l'intelligence artificielle à l'étude de l'apprentissage et de l'évolution du dessin chez les Hominidés.	Cédric Sueur et Marie Pelé (Université de Strasbourg)

7.3.3 Participation à un réseau de recherche

Ces 3 dernières années, je suis impliqué dans le GDR Madics au travers de :

— 2021 : Co-responsable d'un atelier FENDER avec Nicolas Labroche (Université de Tours - LIFAT) et Michael Baker (Telecom Paris Tech - I3)

— 2022-2023 : Co-responsable d'une action HELP avec les mêmes collègues.

Cet atelier puis cette action ont pour but de fédérer la communauté liée au domaine de l'explicabilité, avec un point de vue original : étudier l'explicabilité des pipelines de Machine Learning à la fois du point de vue de l'utilisateur et des données. Cette implication a permis la constitution d'un réseau de participants d'une trentaine de personnes, impliquant des industriels et des académiques de divers laboratoires (doctorants compris). Entre 2021 et 2022, des réunions mensuelles entre les participants ont eu lieu par visio-conférence, afin d'échanger sur nos recherches en explicabilité. Nous proposons aussi régulièrement des interventions aux symposiums organisés par Madics, annuellement (08/07/2021 à Rennes (avec distanciel), 12/07/2022 à Lyon, 25/05/2023 à Troyes)

7.3.4 Organisation d'ateliers et conférences

Depuis ces 4 dernières années, j'ai été amené à co-organiser 3 ateliers (2 hébergés à la conférence nationale EGC, et 1 à l'international, hébergé à ADBIS) et 2 éditions de la conférence nationale EDA, dont voici les détails ci-dessous.

Via le réseau développé dans l'atelier FENDER puis l'action HELP du GDR Madics, j'ai co-organisé les ateliers suivants :

— L'atelier EXPLAIN'AI, hébergé à la conférence EGC, en 2022 à Blois puis en 2023 à Lyon. Cet atelier propose un moment d'échange sur les avancées et sur les projets scientifiques menés autour du triptyque "Données, utilisateurs, Explications". Au vu du succès d'audience lors de ces éditions (environ 60 participants), l'association EGC nous

a proposé de pérenniser cet atelier sous la forme d'un groupe de travail soutenu par l'association EGC. Nous avons alors proposé de co-labelliser l'action HELP du GDR Madics, afin de ne pas créer de structures qui feraient doublon... Notre action du GDR Madics est devenue ainsi la première à être également soutenue par l'association EGC.

— L'atelier international AIDMA (EXEC-MAN) hébergé à la conférence ADBIS, à Barcelone, qui a eu lieu en septembre 2023. Cet atelier a pour but de rassembler un public interdisciplinaire afin de discuter des défis que posent l'explication et la transparence des systèmes de décision intelligents automatisés basés sur l'IA appliquée à la médecine et aux données de santé.

En 2021, j'ai été président de l'organisation de la conférence nationale EDA (Business Intelligence & Big Data ; organisé en distanciel) et président du comité de programme, en 2022, de cette même conférence à Clermont-Ferrand. En 2021 également, j'ai été mandaté par les associations Inforsid et EGC pour proposer une journée commune fédérant les deux communautés. Cette journée a eu lieu en septembre 2022 à Toulouse et a réuni une quarantaine de participants autour de la problématique du "human in the loop".

Chapitre 8

Conclusion générale et perspectives

Nous proposons, dans ce chapitre, de faire le bilan des travaux proposés dans ce mémoire ainsi qu'une synthèse des perspectives de travail présentées dans chaque chapitre de contribution. Enfin, nous évoquerons les nouveaux axes de recherches possibles à développer.

8.1 Bilan

En réponse à la problématique *comment offrir à un utilisateur, les moyens d'une analyse prédictive intelligible ?*, les contributions exposées dans ce mémoire ont permis de s'attaquer aux verrous suivants :

— la recommandation de modèles prédictifs, basée sur le contexte utilisateur ;
— la recommandation de modèles XAI, en fonction de préférences utilisateur ;
— l'étude des limites de l'usage des explications
— l'usage des explications, vues comme un nouvel espace de données analysable.

La recommandation de modèles prédictifs

Nous avons montré l'intérêt de prendre en compte le contexte utilisateur dans la recommandation de modèles prédictifs. Plus précisément, nous avons considéré les multiples préférences souhaitées sur les performances à atteindre pour un modèle prédictif. Le principe de notre système est basé sur l'AutoML afin de rechercher les jeux de données les plus similaires au jeu de données courant (à l'aide d'une mesure de dissimilarité de la littérature). Les analyses passées correspondantes sont alors sélectionnées de sorte à maximiser les performances du modèle, souhaitées par l'utilisateur. Deux versions du système de recommandation sont proposées, afin de considérer différemment la recherche d'un jeu de données similaire et la recherche de performances maximales. Ainsi une version par la moyenne de ces deux critères et une autre basée sur un front de Pareto sont réalisées. Cette dernière obtient les meilleurs résultats selon nos expérimentations, même en cas de difficultés à identifier des jeux de données similaires.

La recommandation de modèles XAI

Au vu de la variété des propositions des modèles XAI, il nous est apparu important de pouvoir offrir un cadre permettant de recommander le modèle XAI le plus adapté au contexte de l'utilisateur, incluant son jeu de données, le modèle prédictif ainsi que ses souhaits sur le type d'explication et ses propriétés à maximiser. En particulier, nous avons proposé une stratégie permettant de sélectionner les modèles XAI correspondants au type d'explication recherché et d'optimiser automatiquement leurs hyperparamètres de sorte à maximiser les propriétés souhaitées (à l'image de ce qui est fait, en partie, dans le domaine de l'AutoML). Notre proposition a été validée au travers de deux cas d'usage, considérant des explications basées sur les exemples et des explications attributives, montrant ainsi l'adaptabilité de notre approche à des types d'explication différents.

Les limites de l'usage des explications

Les explications post-hoc locales attributives, largement utilisées de nos jours et dans de nombreux domaines d'applications, ne sont pas sans défauts. Ces problèmes nous ont alors motivés à proposer une nouvelle méthode d'explication, basée sur les principes des valeurs de Shapley, afin de répondre au manque de considération des interactions entre variables. Nous avons ainsi présenté une solution calculant des coalitions de variables corrélées, à l'aide de différentes méthodes de calculs de corrélations. La méthode basée sur la corrélation de Spearman a montré qu'elle obtenait les meilleurs résultats en termes de temps de calcul et de taux de précision obtenu, comparés aux autres méthodes de calculs.

Nous avons aussi été amenés à mieux étudier l'applicabilité des méthodes d'explication attributives, en proposant une feuille de route sur leurs usages, en fonction de la taille d'un jeu de données à analyser et le type de modèle prédictif utilisé. Nous avons, par exemple, montré qu'il était préférable d'utiliser notre proposition, basée sur les corrélations de Spearman, pour expliquer localement de petits jeux de données et TreeSHAP lorsqu'il s'agit d'un jeu de données plus volumineux et qu'un modèle prédictif, basé arbre, est utilisé. Il est à noter qu'il n'existe, à ce jour, aucune proposition permettant de calculer des explications pour de grands jeux de données, de manière précise et agnostique à tout modèle prédictif.

Les explications, vues comme un nouvel espace de données

Via les explications attributives, nous avons mis en évidence comment celles-ci pouvaient être bénéfiques à différentes tâches liées à l'analyse prédictive. Dans le domaine de la sélection de modèles, les explications peuvent servir à un cadre itératif plus global permettant de mieux comprendre un modelé prédictif afin de mieux le personnaliser (via des tâches de préprocessing par exemple) ou de le valider. Les explications ont aussi montré leur avantage dans le cadre d'un usage de type "bac à sable" d'un modèle prédictif, afin de simuler des données et étudier ce qui a pu influencer leurs prédictions.

Nous avons aussi présenté l'impact que peuvent avoir les profils d'explications dans le choix d'une sélection de variables. Cela nous a permis de soutenir l'idée de considérer l'explicabilité

comme une dimension à part entière dans le choix d'une méthode de sélection de variables, aux côtés des classiques accuracy et taux de rétention, pour un modèle prédictif donné.

Afin de sélectionner des instances ayant un pouvoir explicatif représentatif du modèle, nous avons proposé un cadre permettant d'obtenir des clusters d'explications produisant des informations plus précises sur les relations entre les données et le modèle prédictif. Les évaluations ont montré que les clusters basés sur les explications étaient plus homogènes et de meilleure qualité que ceux basés sur les seules données brutes. Plus particulièrement, les évaluations ont aussi mis en évidence que notre cadre reste robuste, même en cas de modèles prédictifs dont la précision est dégradée.

En exploitant ces clusters d'explication, nous avons montré, au travers d'un cas d'étude médicale, la complémentarité que peut avoir une analyse des explications avec une analyse de données plus large. En ajoutant un système de génération de règles de décision pour expliquer les clusters d'explications, nous avons pu faire remonter des interactions qu'il était impossible de déduire avec des statistiques plus classiques. La concision des explications proposées renforce aussi l'intérêt à une adoption plus large dans une aide à l'analyse de données.

8.2 Synthèse des perspectives de recherche

Tout au long des chapitres de contributions, nous avons pu faire état d'un certain nombre de perspectives de recherche possibles. Ces perspectives peuvent se résumer aux quatre principaux axes de recherche suivants :

— L'interaction avec les explications ;
— L'approche *human-in-the-loop* et le feedback utilisateur ;
— La caractérisation et la prévention des limites en XAI ;
— La proposition d'un benchmark pour l'XAI.

L'interaction avec les explications

Comme étudié dans le Chapitre 6, concernant l'aide à la sélection d'instances et l'aide à l'analyse de données plus généralement, les explications regroupées en clusters permettent de mieux mettre en avant les interactions entre les données. Le nombre de clusters générés, dépendant sans doute de la quantité de données d'origine, peut cependant devenir important. Un moyen d'y remédier serait de proposer une solution structurant ces clusters selon différents niveaux de hiérarchies (en faisant alors appel à un algorithme de clustering hiérarchique). Associé à des opérateurs de navigation et à une représentation condensée des clusters (à l'aide de règles de décision entre les divisions de clusters, par exemple), il serait alors possible d'offrir un moyen d'explorer et d'interagir plus facilement avec une masse d'explications. En cela, le domaine de la navigation OLAP peut sans doute être une source d'inspiration intéressante pour ce travail [Aligon et al., 2014].

Dans un autre registre, comme indiqué dans le Chapitre 5, les méthodes d'explications de type SHAP présentent une limite importante au sujet de l'indépendance entre les variables (ce

qui est pourtant rarement le cas dans l'analyse de jeux de données réels). Notre proposition, basée sur les calculs de coalitions de variables corrélées, a montré tout l'intérêt de mieux considérer les dépendances entre variables. Comparée à d'autres méthodes locales attributives, notre approche coalitionnelle semble en effet mieux détecter certaines interactions. Cependant, notre méthode souffre du problème de réentraînement du modèle prédictif lors d'absence de variables dans les coalitions. Il pourrait donc être pertinent de proposer une solution hybride mixant notre méthode avec la méthode SHAP pour ne se focaliser que sur les coalitions de variables corrélées tout en simulant l'absence de variables par des perturbations sur un échantillon d'instances, évitant alors le réentraînement du modèle prédictif.

L'approche *human-in-the-loop* et le feedback utilisateur

Nous avons pu montrer, dans ce mémoire, tout l'intérêt de considérer une approche "human-in-the-loop" pour de l'aide à la sélection de modèles prédictifs ou XAI (voir Chapitres 3, 4 et 6). Afin de compléter cette approche, la prise en compte de l'expertise utilisateur pourrait permettre de recommander des modèles prédictifs (au travers d'une sous séquence de workflows, par exemple) ou XAI s'approchant au plus près de ses besoins. Cette expertise, dans le cas de la sélection de modèles, pourrait se définir en fonction du niveau de compréhension des explications qui lui sont fournies. En effet, comme nous avons pu le voir dans le cas particulier de la sélection de variables (qui n'est qu'un élément parmi d'autres d'un workflow), les explications produites peuvent diverger sensiblement en fonction de la méthode de sélection de variables considérée. Ainsi, la recommandation d'un modèle prédictif ou XAI pourrait être contrainte par des profils d'explication que l'utilisateur serait en mesure de mieux comprendre. Par conséquent, peu importe alors la complexité du modèle proposé, tant qu'il reste compréhensible par l'utilisateur (sans une dégradation trop importante de l'accuracy).

Néanmoins, un modèle prédictif et les explications fournies ne sont pas exempts d'erreurs. Il semble donc important de compléter l'approche "human-in-the-loop" en intégrant les feedback de l'utilisateur concernant les erreurs relevées. Cette démarche itérative devrait alors permettre de rendre le modèle plus cohérent vis-à-vis des besoins utilisateur. L'usage d'une sélection d'instances, la plus concise possible et ayant un pouvoir explicatif important, pourrait permettre de recueillir ces feed-back, par exemple en pointant les instances ayant des prédictions ou explications problématiques.

L'usage de modèles plus cohérent et compréhensibles permettrait de mieux généraliser les approches du type "what-if" [Wexler et al., 2019, Wexler et al., 2020, Singh et al., 2021], permettant une démarche plus exploratoire des prédictions, afin de tester des hypothèses métiers, et d'y faire émerger d'éventuels effets de causalités dans les variables analysées pour des groupes d'instances bien définies.

La caractérisation et la prévention des limites en XAI

Comme nous avons pu le voir dans les Chapitres 2 et 5, le domaine de l'XAI, et en particulier les approches locales post-hoc attributives, ne sont pas exempt de défauts [Slack et al., 2020, Dimanov et al., 2020]. Un des grands problèmes actuels est lié à la *ro-*

bustesse de ce type d'explications pour lesquelles des instances similaires n'ont pas toujours d'explications similaires : cela peut amener l'utilisateur à s'interroger sur la fiabilité des explications qui lui sont fournies.

Dans l'objectif de résoudre ces incohérences, de récents travaux ont proposé de préférer l'usage des explications logiques, qu'elles soient abductives (pourquoi le classifieur a prédit l'instance pour cette classe ?) [Ignatiev et al., 2019, Marques-Silva and Ignatiev, 2022] ou contrastives/contrefactuelles (pourquoi le classifieur n'a pas prédit l'instance pour cette classe ?) [Ignatiev et al., 2020, Bloch and Lesot, 2022], assurant par essence leur fiabilité. Ces explications ont cependant deux défauts majeurs : leurs temps de calcul sont exponentiels et elles sont soumises aussi à l'effet *Rashomon*. Il existe en effet une multitude d'explications possibles pour une même instance.

Comme souligné par Christoph Molnar dans un récent article [1], les problèmes identifiés dans les méthodes d'explications attributives ne doivent pas faire oublier leurs avantages et bénéfices. En particulier, ces méthodes permettent une représentation des explications concise, agrégeable et conviviale (notamment dans les visualisations proposées) justifiant leur large utilisation dans de nombreux domaines, y compris sensibles comme le médical. De plus, des analyses via l'utilisation de diagrammes de dépendance partielle (PDP) permettent de rechercher des corrélations (et donc possibles causalités) entre valeurs de variables et explications, très simplement. C'est notamment sur ces bases que nous avons soutenu, tout au long de ce mémoire, l'usage de ces méthodes. Par un pragmatisme certain, il nous semble alors délicat de mettre de côté des méthodes aussi populaires et qui ont, de facto, fait leurs preuves ces dernières années.

Néanmoins, les incohérences soulevées pourraient être source de perspectives particulièrement intéressantes. Par exemple, la détection automatique d'instances ou de groupes d'instances, dont le manque de robustesse serait jugé trop fort, permettrait de limiter, voire interdire, le calcul des explications les concernant. Plus largement, les incohérences logiques identifiées permettraient de mieux guider, prévenir et avertir l'utilisateur sur certains cas d'usage des explications attributives.

Une autre limite importante à l'usage des méthodes attributives concerne leur "bonne applicabilité" en fonction de caractéristiques de jeux de données et d'un modèle prédictif. En effet, il n'existe pas, à ce jour, d'études fines sur les influences que peuvent avoir les jeux de données, modèles prédictifs et explications attributives entre eux. Par exemple, une explication reste-t-elle fiable même dans le cas d'un modèle prédictif dont l'*accuracy* serait fortement dégradé ? Y a-t-il des caractéristiques de jeux de données pouvant impacter fortement la compréhension des explications ?

Cette dernière question peut mener à s'intéresser aux problèmes de qualité et biais dans les données et leurs impacts dans les explications. Une première étude pourrait s'intéresser aux clusters d'explication proposés dans le Chapitre 6 afin d'identifier des sous-groupes d'explications qui n'obéiraient pas ou divergeraient trop (dans le sens d'influences de variables peu ordinaires) par rapport à un comportement général pourtant attendu. Ces explications

1. https://mindfulmodeler.substack.com/p/should-we-stop-interpreting-ml-models

remontées à l'utilisateur pourraient alors permettre de mieux valider le bon usage d'un modèle prédictif.

La proposition d'un benchmark pour l'XAI

Nous avons pu constater dans les Chapitres 4 et 5 le manque de références de la littérature afin de pouvoir comparer les méthodes XAI existantes. La proposition d'un benchmark, associant génération synthétique de jeux de données et ensemble de propriétés et métriques, est une perspective importante afin d'être en mesure de choisir, objectivement, la méthode XAI la plus adaptée au besoin de l'utilisateur. Quelques travaux [Arras et al., 2022, Agarwal et al., 2022, Liu et al., 2021] se sont récemment penchés sur ce type de proposition. En particulier, les travaux proposés dans [Agarwal et al., 2022] semblent un excellent début dans la constitution de ce type de benchmark. Les auteurs proposent OpenXAI [2] afin d'évaluer les méthodes post-hoc attributives. Ce cadre inclut un générateur de données synthétiques, quelques jeux de données réels, des modèles prédictifs préentraînés, ainsi que la plupart des métriques de la littérature permettant d'évaluer les méthodes attributives.

Néanmoins, cette proposition se limite aux seules approches attributives sans inclure d'autres types de méthodes XAI. De plus, même si ce benchmark reprend les propriétés importantes de la littérature pour l'évaluation des explications, rien n'est dit sur les possibles contradictions que peuvent avoir les différentes propositions de métriques pour une même propriété (comme souligné dans le Chapitre 4). C'est un point d'investigation important avant toute évaluation objective des méthodes XAI. Comme nous avons également pu le tester dans le Chapitre 5, la notion de lisibilité est sans doute une propriété manquante de la littérature afin de s'assurer que les explications fournies soient véritablement informatives pour l'utilisateur.

Enfin, l'évaluation subjective des explications est tout autant d'un intérêt majeur. Les campagnes d'évaluation, auprès d'utilisateurs réels, des méthodes XAI ou de toute proposition y faisant appel sont encore aujourd'hui trop limitées. Ces campagnes pourraient d'ailleurs faire partie intégrante d'ensembles de vérités terrain utilisables dans un benchmark pour l'XAI.

8.3 Nouveaux champs d'études à développer

Au-delà des perspectives évoquées dans la section précédente, les travaux proposés dans ce mémoire peuvent aussi mener à l'exploration de champs d'études plus larges.

L'ensemble des champs d'études que nous allons détailler peuvent se rapporter à une même problématique principale : assurer une analyse prédictive intelligible au sein d'un Système d'Information. En effet, comme noté dans le Chapitre 2, le domaine du Système d'Information s'est assez peu approprié les problématiques liées à l'apprentissage automatique et l'explicabilité. Cela provient notamment de la complexité à industrialiser facilement ces tâches.

Quatre grands objectifs sont alors envisagés, comme indiqué dans la Figure 8.1.

2. https://github.com/AI4LIFE-GROUP/OpenXAI

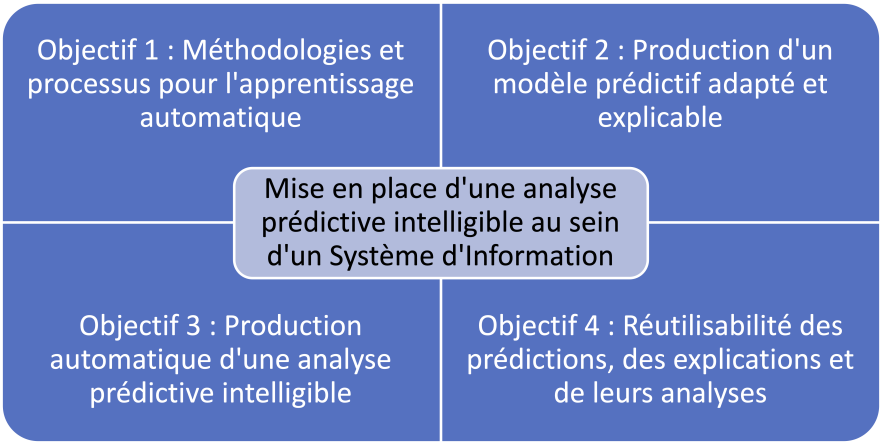


FIGURE 8.1 – Problématique principale : l'analyse prédictive intelligible dans un Système d'Information

Dans l'idée d'une meilleure démocratisation, notamment au sein de petites ou moyennes organisations, la recherche doit ainsi cibler les méthodologie et processus assurant le déploiement et l'usage de modèles prédictifs fiables, explicables et répondant à des obligations juridiques nouvelles. La recherche doit aussi pouvoir s'intéresser à la production automatique d'analyses prédictives intelligibles : cela ne peut que faciliter le travail des experts de domaines ou de tout expert en sciences des données. Enfin, l'ensemble de ces actions et usages réalisés au sein d'un Système d'Information peuvent se voir comme des sources de données, à part entière. La gestion de données a alors tout son rôle à jouer en permettant leur stockage, leur interrogation et leur réutilisabilité.

Les quatre objectifs sont illustrés dans la Figure 8.2 et détaillés ci-après.

8.3.1 Objectif 1 : Méthodologie et processus pour l'apprentissage automatique

La définition de méthodologies, processus et modèles, notamment conceptuels, ne pourra que faciliter le déploiement et l'usage de l'apprentissage automatique au sein d'un Système d'Information. Dans ce but, [Maass and Storey, 2021] pose d'ailleurs les bases d'une première proposition de modèle conceptuel applicable à l'apprentissage automatique. Comme indiqué par les auteurs, le principal problème lié à l'apprentissage automatique au sein d'un Système d'Information vient du fait que les connaissances métiers ne sont pas nécessairement un prérequis à l'élaboration de modèles prédictifs, au contraire de domaines comme les bases de données ou la business intelligence. Cela peut, certes, faciliter les démarches exploratoires, mais rend plus difficiles la conception et le développement de solutions d'apprentissage automatique.

La problématique d'une analyse intelligible pose aussi la question de sa place et sa prise

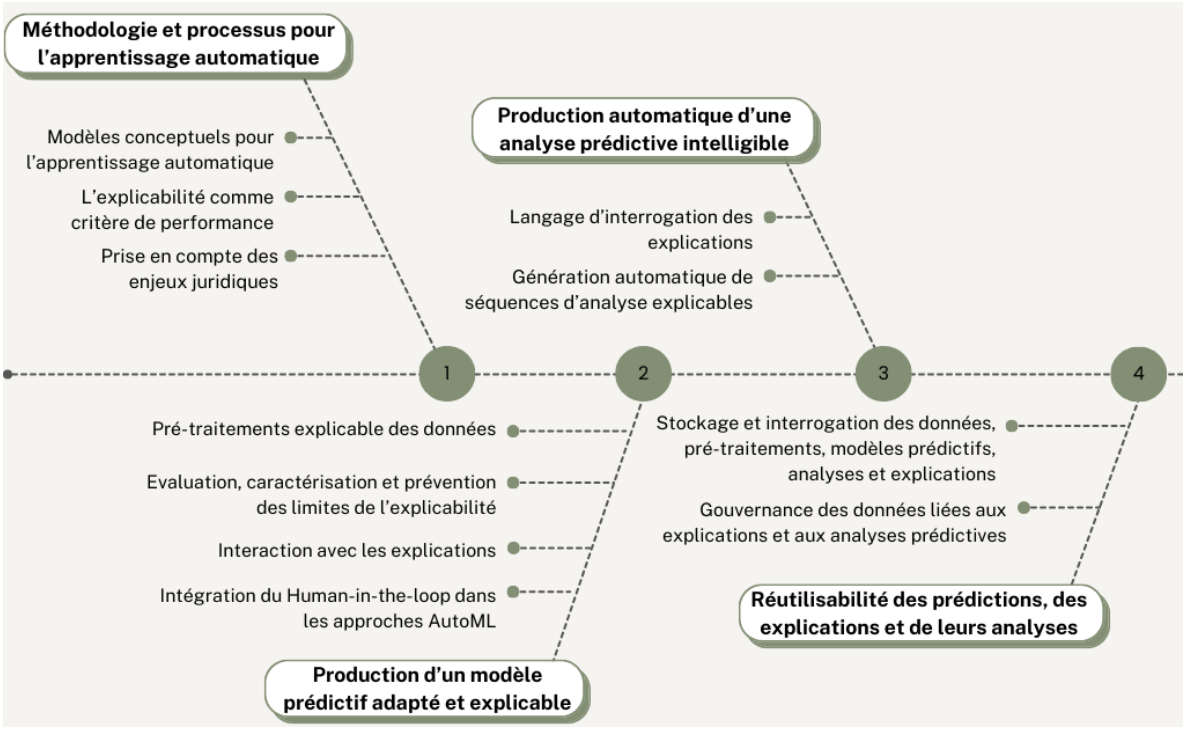


FIGURE 8.2 – Détails des objectifs pour la mise en place d'une analyse prédictive intelligible au sein de Système d'Information

en compte dans un Système d'Information : comment assurer que tous les processus mis en place puissent répondre correctement à un objectif d'analyse prédictive et que chaque étape du processus soit compréhensible et explicable pour l'utilisateur, et partageable entre tous ?

La législation européenne prévue sur l'intelligence artificielle (AI act [3]) exigera certainement plus de transparence dans les algorithmes d'apprentissage et modèles prédictifs.

Dans cet objectif, une perspective possible serait de considérer les contraintes juridiques non pas comme une validation a posteriori à la création d'un modèle prédictif, mais de les considérer dès leur conception initiale (position également défendue dans [Maass and Storey, 2021]). Ainsi un algorithme d'apprentissage automatique aurait à intégrer les contraintes juridiques d'abord, quitte à devoir simplifier le modèle produit ou bien interdire automatiquement à certaines variables ou sous-ensembles de données à servir à l'apprentissage. L'explicabilité pourrait là encore être une source intéressante afin de faire le lien entre les données, le modèle prédictif et les contraintes juridiques.

8.3.2 Objectif 2 : Production d'un modèle prédictif adapté et explicable

L'investissement dans l'aide à la construction de modèles prédictifs (de type AutoML) peut être vu comme une approche intéressante. Cependant, et comme nous l'avons constaté tout au long de ce mémoire, l'usage seul de l'AutoML et son manque de transparence génèrent possiblement un manque de confiance limitant son adoption. Ainsi l'aide à l'analyse prédictive doit pouvoir s'accompagner de moyens la rendant intelligible. Cela impose de s'intéresser au besoin d'expliquer chaque étape du pipeline d'analyse : les données brutes, les prétraitements effectués, les modèles prédictifs choisis ainsi que les modèles d'explication compatibles. Plus encore, les relations entre ces étapes sont tout aussi importantes à comprendre : l'impact d'une gestion de la qualité des données brutes a des conséquences sur une sélection de variables qui elle-même contraint l'élaboration d'un modèle prédictif.

Ajouté aux perspectives discutées dans la Section 8.2, considérant les limites de l'explicabilité, les moyens d'interagir avec elle et considérant une approche human-in-the-loop, il serait alors possible de produire des modèles fiables, vérifiés et explicables.

Quelques applications directes sont d'ailleurs déjà envisagées pour la justification de modèles [4] ou bien l'amélioration et le débogage de modèles [5].

8.3.3 Objectif 3 : Production automatique d'une analyse prédictive intelligible

La narration de données (data storytelling) est un domaine récent et de plus en plus étudié en recherche [Marcel et al., 2023, El Outa et al., 2022, Chanson et al., 2022]. L'idée est, comme son nom l'indique, de pouvoir raconter une histoire à partir d'un jeu de données

3. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
4. https://mindfulmodeler.substack.com/p/use-interpretability-to-justify-your
5. https://mindfulmodeler.substack.com/p/how-i-use-ml-interpretability-for

et cela automatiquement. Elle a pour objectif de produire automatiquement une séquence logique (composée de visualisations, mots clés, textes, etc.) à partir d'éléments d'un jeu de données, considérés comme ayant un intérêt majeur pour l'analyse. On y trouve un lien très fort avec le journalisme de données (data journalism), dont l'objectif est similaire et permet de produire des articles sur une thématique précise et mettant en valeur un ou plusieurs jeux de données. La principale difficulté de ce domaine est de pouvoir détecter automatiquement les tendances véritablement intéressantes à analyser dans un jeu de données. Dans le cadre d'une analyse prédictive, la narration de données est complexifiée par le besoin de prendre en compte un modèle prédictif, en plus du jeu de données initial. Cependant, les techniques de XAI devraient être un atout particulièrement intéressant afin de repérer dans les données des tendances cachées (notamment les possibles causalités entre variables concernant des sous-ensembles particuliers de données). Comme indiqué dans la Section 8.2, la proposition d'un langage d'interrogation des explications, à l'aide d'opérateurs, permettrait d'explorer un jeu de données via un modèle prédictif afin de mieux cibler les clusters d'instances souhaités.

8.3.4 Objectif 4 : Réutilisabilité des prédictions, des explications et de leurs analyses

Un point d'élargissement de nos travaux de recherche peut aussi se situer dans le domaine de la gestion de données. Dans le cadre de l'industrialisation des processus de Big Data Analytics, les architectures de lac de données [Ravat and Zhao, 2019a, Ravat and Zhao, 2019b, Derakhshannia et al., 2020, Dolhopolov et al., 2023] permettent de stocker, interroger et analyser/croiser de multiples sources de données, processus de traitement et d'analyses conduites. Cependant, ces architectures ne peuvent fonctionner qu'avec l'aide d'un système de gouvernance de métadonnées efficace. En complément des activités de recherche exprimées dans ce mémoire, nous avons proposé un système de gouvernance de métadonnées couvrant tous les aspects d'une analyse prédictive [Zhao et al., 2021]. Cette gouvernance de métadonnées permet alors de répondre à des questions du type : "quelles sont les analyses effectuées sur des jeux de données passées proches d'un jeu de données courant" ou encore "quel est le modèle prédictif le plus performant pour un jeu de données à analyser".

L'aspect XAI n'est cependant pas pris en compte dans ce système de métadonnées. En considérant, par exemple, les explications attributives comme des données en tant que telles, des métadonnées orientées XAI pourraient certainement servir à répondre à des questions complémentaires du type : "Quels sont les jeux de données, via une analyse prédictive, s'expliquant de la même manière ?" ou encore "Quelles sont les différences d'influences constatées entre différents modèles prédictifs pour un même jeu de données ?"

Bibliographie

[Abdel-Karim et al., 2021] Abdel-Karim, B. M., Pfeuffer, N., and Hinz, O. (2021). Machine learning in information systems-a bibliographic review and open research issues. *Electronic Markets*, 31 :643–670.

[Adadi and Berrada, 2018] Adadi, A. and Berrada, M. (2018). Peeking inside the black-box : a survey on explainable artificial intelligence (xai). *IEEE access*, 6 :52138–52160.

[Adomavicius and Tuzhilin, 2010] Adomavicius, G. and Tuzhilin, A. (2010). Context-aware recommender systems. In *Recommender systems handbook*, pages 217–253. Springer.

[Agarwal et al., 2022] Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., and Lakkaraju, H. (2022). Openxai : Towards a transparent evaluation of model explanations. *Advances in Neural Information Processing Systems*, 35 :15784–15799.

[Alanazi et al., 2017] Alanazi, H. O., Abdullah, A. H., and Qureshi, K. N. (2017). A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. *J. Med. Syst.*, 41(4) :1–10.

[Aligon et al., 2014] Aligon, J., Boulil, K., Marcel, P., and Peralta, V. (2014). A holistic approach to olap sessions composition : The falseto experience. In *Proceedings of the 17th International Workshop on Data Warehousing and OLAP*, pages 37–46.

[Aligon et al., 2015] Aligon, J., Gallinucci, E., Golfarelli, M., Marcel, P., and Rizzi, S. (2015). A collaborative filtering approach for recommending olap sessions. *Decision Support Systems*, 69 :20–30.

[Aligon et al., 2016] Aligon, J., Guillet, F., Blanchard, J., Picarougne, F., and Duke, E. (2016). Défi egc 2016 : Analyse par motifs fréquents et topic modeling. In *EGC*, pages 395–406.

[Alvarez-Melis and Jaakkola, 2018] Alvarez-Melis, D. and Jaakkola, T. S. (2018). On the Robustness of Interpretability Methods. *Workshop on Human Interpretability for Machine Learning (WHI) - International Conference on Machine Learning (ICML)*. arXiv : 1806.08049.

[Amershi et al., 2014] Amershi, S., Cakmak, M., Knox, W. B., and Kulesza, T. (2014). Power to the people : The role of humans in interactive machine learning. *Ai Magazine*, 35(4) :105–120.

[Ancona et al., 2018] Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2018). Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

[Anderson, 2016] Anderson, R. (2016). The rashomon effect and communication. *Canadian Journal of Communication*, 41(2) :249–270.

[Angwin et al., 2016] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. propublica, may 23.

[Arras et al., 2022] Arras, L., Osman, A., and Samek, W. (2022). Clevr-xai : A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81 :14–40.

[Arya et al., 2019] Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., and Zhang, Y. (2019). One explanation does not fit all : A toolkit and taxonomy of ai explainability techniques.

[Astsatryan et al., 2018] Astsatryan, H. V., Grogoryan, H., Gyulgyulyan, E., Hakobyan, A., Kocharyan, A., Narsisian, W., Sahakyan, V., Shoukourian, Y., Abrahamyan, R., Petrosyan, Z., et al. (2018). Weather data visualization and analytical platform. *Scalable Computing : Practice and Experience*, 19(2) :79–86.

[Attaran and Attaran, 2019] Attaran, M. and Attaran, S. (2019). Opportunities and challenges of implementing predictive analytics for competitive advantage. *Applying business intelligence initiatives in healthcare and organizational settings*, pages 64–90.

[Bach et al., 2015] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7). Publisher : Public Library of Science.

[Barredo Arrieta et al., 2020] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (xai) : Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58 :82–115.

[Bergstra and Bengio, 2012] Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).

[Bernard et al., 2023] Bernard, D., Doumard, E., Ader, I., Kemoun, P., Pagès, J.-C., Galinier, A., Cussat-Blanc, S., Furger, F., Ferrucci, L., Aligon, J., et al. (2023). Explainable machine learning framework to predict personalized physiological aging. *Aging Cell*, page e13872.

[Bibault et al., 2021] Bibault, J.-E., Chang, D. T., and Xing, L. (2021). Development and validation of a model to predict survival in colorectal cancer using a gradient-boosted machine. *Gut*, 70(5) :884–889.

[Bloch and Lesot, 2022] Bloch, I. and Lesot, M.-J. (2022). Towards a formulation of fuzzy contrastive explanations. In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. IEEE.

[Boizard et al., 2021] Boizard, F., Buffin-Meyer, B., Aligon, J., Teste, O., Schanstra, J. P., and Klein, J. (2021). Prynt : A tool for prioritization of disease candidates from proteomics data using a combination of shortest-path and random walk algorithms. *Scientific Reports*, 11(1) :5764.

[Bove et al., 2023] Bove, C., Lesot, M.-J., Tijus, C. A., and Detyniecki, M. (2023). Investigating the intelligibility of plural counterfactual examples for non-expert users : an explanation user interface proposition and user study. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 188–203.

[Brasse et al., 2023] Brasse, J., Broder, H. R., Förster, M., Klier, M., and Sigler, I. (2023). Explainable artificial intelligence in information systems : A review of the status quo and future research directions. *Electronic Markets*, 33(1) :26.

[Brown et al., 2018] Brown, A., Tuor, A., Hutchinson, B., and Nichols, N. (2018). Recurrent neural network attention mechanisms for interpretable system log anomaly detection. In *Proceedings of the First Workshop on Machine Learning for Computing Systems*, MLCS'18, New York, NY, USA. Association for Computing Machinery.

[Buchanan and Shortliffe, 1984] Buchanan, B. G. and Shortliffe, E. H. (1984). *Rule based expert systems : the mycin experiments of the stanford heuristic programming project (the Addison-Wesley series in artificial intelligence)*. Addison-Wesley Longman Publishing Co., Inc.

[Caruana and Niculescu-Mizil, 2006] Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168.

[Carvalho et al., 2019] Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability : A survey on methods and metrics. *Electronics*, 8(8).

[Casella and Berger, 2021] Casella, G. and Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

[Celebi and Aydin, 2016] Celebi, M. E. and Aydin, K. (2016). *Unsupervised learning algorithms*, volume 9. Springer.

[Chandrasekaran et al., 1989] Chandrasekaran, B., Tanner, M. C., and Josephson, J. R. (1989). Explaining control strategies in problem solving. *IEEE Intelligent Systems*, 4(01) :9–15.

[Chanson et al., 2022] Chanson, A., Labroche, N., Marcel, P., Rizzi, S., and t'Kindt, V. (2022). Automatic generation of comparison notebooks for interactive data exploration. In *EDBT*, pages 2–274.

[Chanson et al., 2021] Chanson, A., Labroche, N., and Verdeaux, W. (2021). Towards local post-hoc recommender systems explanations. In *Proceedings of the 23rd International*

Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP).

[Chen and Pu, 2005] Chen, L. and Pu, P. (2005). Trust building in recommender agents. In *Proceedings of the Workshop on Web Personalization, Recommender Systems and Intelligent User Interfaces at the 2nd International Conference on E-Business and Telecommunication Networks*, pages 135–145.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

[Chromik and Butz, 2021] Chromik, M. and Butz, A. (2021). Human-xai interaction : a review and design principles for explanation user interfaces. In *Human-Computer Interaction–INTERACT 2021 : 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18*, pages 619–640. Springer.

[Cohen, 1968] Cohen, J. (1968). Weighted kappa : nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4) :213.

[Cohen, 2021] Cohen, S. (2021). Chapter 1 - the evolution of machine learning : past, present, and future. In Cohen, S., editor, *Artificial Intelligence and Deep Learning in Pathology*, pages 1–12. Elsevier.

[Cohen et al., 2007] Cohen, S., Dror, G., and Ruppin, E. (2007). Feature selection via coalitional game theory. *Neural Computation*, 19(7) :1939–1961.

[Cohen et al., 2005] Cohen, S., Ruppin, E., and Dror, G. (2005). Feature selection based on the shapley value. *other words*, 1 :98Eqr.

[Conrad et al., 2005] Conrad, J. G., Al-Kofahi, K., Zhao, Y., and Karypis, G. (2005). Effective document clustering for large heterogeneous law firm collections. In *AIL Proceedings*.

[Cooper et al., 2021] Cooper, A., Doyle, O., and Bourke, A. (2021). Supervised clustering for subgroup discovery : An application to covid-19 symptomatology. In *ECML-PKDD Proceedings*.

[Cramer et al., 2008] Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., Aroyo, L., and Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-adapted interaction*, 18 :455–496.

[Cugny et al., 2022] Cugny, R., Aligon, J., Chevalier, M., Roman Jimenez, G., and Teste, O. (2022). Autoxai : A framework to automatically select the most adapted xai solution. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 315–324.

[Cugny et al., 2023] Cugny, R., Aligon, J., Chevalier, M., Roman-Jimenez, G., and Teste, O. (2023). Autoxai : Un cadre pour sélectionner automatiquement la solution d'xai la plus adaptée. In *EGC*, pages 491–498.

[Cunningham et al., 2008] Cunningham, P., Cord, M., and Delany, S. J. (2008). Supervised learning. In *Machine learning techniques for multimedia : case studies on organization and retrieval*, pages 21–49. Springer.

[Czerniak and Zarzycki, 2003] Czerniak, J. and Zarzycki, H. (2003). Application of rough sets in the presumptive diagnosis of urinary system diseases. In *AI and Security in Computing Systems*.

[Dandl et al., 2020] Dandl, S., Molnar, C., Binder, M., and Bischl, B. (2020). Multi-objective counterfactual explanations. In *International Conference on Parallel Problem Solving from Nature*, pages 448–469. Springer.

[Datta et al., 2016] Datta, A., Sen, S., and Zick, Y. (2016). Algorithmic transparency via quantitative input influence : Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617.

[De Angeli et al., 2021] De Angeli, K., Gao, S., Alawad, M., Yoon, H.-J., Schaefferkoetter, N., Wu, X.-C., Durbin, E. B., Doherty, J., Stroup, A., Coyle, L., et al. (2021). Deep active learning for classifying cancer pathology reports. *BMC bioinformatics*, 22(1) :1–25.

[Derakhshannia et al., 2020] Derakhshannia, M., Gervet, C., Hajj-Hassan, H., Laurent, A., and Martin, A. (2020). Data lake governance : Towards a systemic and natural ecosystem analogy. *Future internet*, 12(8) :126.

[Dimanov et al., 2020] Dimanov, B., Bhatt, U., Jamnik, M., and Weller, A. (2020). You shouldn't trust me : Learning models which conceal unfairness from multiple explanation methods.

[Dolhopolov et al., 2023] Dolhopolov, A., Castelltort, A., and Laurent, A. (2023). Trick or treat : Centralized data lake vs decentralized data mesh. In *International Conference on Management of Digital*, pages 303–316. Springer.

[Doshi-Velez and Kim, 2017] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv :1702.08608*.

[Doumard et al., 2022] Doumard, E., Aligon, J., Escriva, E., Excoffier, J.-B., Monsarrat, P., and Soulé-Dupuy, C. (2022). A comparative study of additive local explanation methods based on feature influences. In *24th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data ((DOLAP 2022)*, volume 3130, pages 31–40. CEUR-WS. org.

[Doumard et al., 2023] Doumard, E., Aligon, J., Escriva, E., Excoffier, J.-B., Monsarrat, P., and Soulé-Dupuy, C. (2023). A quantitative approach for the comparison of additive local explanation methods. *Information Systems*, 114 :102162.

[Došilović et al., 2018] Došilović, F. K., Brčić, M., and Hlupić, N. (2018). Explainable artificial intelligence : A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215.

[Drushku et al., 2019] Drushku, K., Aligon, J., Labroche, N., Marcel, P., and Peralta, V. (2019). Interest-based recommendations for business intelligence users. *Information Systems*, 86 :79–93.

[Drushku et al., 2020] Drushku, K., Aligon, J., Labroche, N., Marcel, P., and Peralta, V. (2020). Recommandations basées sur les centres d'intérêts utilisateurs en business intelligence. In *INFORSID 2020*.

[Drushku et al., 2017] Drushku, K., Aligon, J., Labroche, N., Marcel, P., Peralta, V., and Dumant, B. (2017). User interests clustering in business intelligence interactions. In *Advanced Information Systems Engineering : 29th International Conference, CAiSE 2017, Essen, Germany, June 12-16, 2017, Proceedings 29*, pages 144–158. Springer.

[Dua and Graff, 2017] Dua, D. and Graff, C. (2017). UCI machine learning repository.

[Efron et al., 2004] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, 32(2) :407–499.

[Eiras-Franco et al., 2019] Eiras-Franco, C., Guijarro-Berdinas, B., Alonso-Betanzos, A., and Bahamonde, A. (2019). A scalable decision-tree-based method to explain interactions in dyadic data. *Decision Support Systems*, 127 :113141.

[El Outa et al., 2022] El Outa, F., Marcel, P., Peralta, V., da Silva, R., Chagnoux, M., and Vassiliadis, P. (2022). Data narrative crafting via a comprehensive and well-founded process. In *European Conference on Advances in Databases and Information Systems*, pages 347–360. Springer.

[Escriva et al., 2023a] Escriva, E., Aligon, J., Excoffier, J.-B., Monsarrat, P., and Soulé-Dupuy, C. (2023a). How to make the most of local explanations : effective clustering based on influences. In *European Conference on Advances in Databases and Information Systems*, pages 146–160. Springer.

[Escriva et al., 2023b] Escriva, E., Doumard, E., Excoffier, J.-B., Aligon, J., Monsarrat, P., and Soulé-Dupuy, C. (2023b). Data exploration based on local attribution explanation : A medical use. In *New Trends in Database and Information Systems : ADBIS 2023 Short Papers, Doctoral Consortium and Workshops : AIDMA, DOING, K-Gals, MADEISD, PeRS, Barcelona, Spain, September 4–7, 2023, Proceedings*, page 315. Springer Nature.

[Escriva et al., 2022] Escriva, E., Ferrettini, G., Aligon, J., Excoffier, J.-B., and Soulé-Dupuy, C. (2022). Stratégies coalitionnelles pour une explication efficace des prédictions individuelles. In *Conférence francophone sur l'Extraction et la Gestion des Connaissances (EGC 2022)*, pages 395–402. RNTI : Revue des Nouvelles Technologies de l'Information.

[Evgeniou et al., 2005] Evgeniou, T., Micchelli, C. A., Pontil, M., and Shawe-Taylor, J. (2005). Learning multiple tasks with kernel methods. *Journal of machine learning research*, 6(4).

[Excoffier et al., 2022] Excoffier, J.-B., Escriva, E., Aligon, J., and Ortala, M. (2022). Local explanation-based method for healthcare risk stratification. In *Medical Informatics Europe 2022*, volume 294.

[Fayyad et al., 1996a] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996a). From data mining to knowledge discovery in databases. *AI magazine*, 17(3) :37–37.

[Fayyad et al., 1996b] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996b). The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11) :27–34.

[Ferrettini, 2021] Ferrettini, G. (2021). *Adaptive system for analysis process design assistance.* PhD thesis, Université Toulouse Capitole.

[Ferrettini et al., 2019] Ferrettini, G., Aligon, J., and Soulé-Dupuy, C. (2019). Un cadre d'aide à l'exploitation des résultats de prédictions, à destination d'experts de domaine.

[Ferrettini et al., 2020a] Ferrettini, G., Aligon, J., and Soulé-Dupuy, C. (2020a). Explaining single predictions : A faster method. In *SOFSEM 2020 : Theory and Practice of Computer Science : 46th International Conference on Current Trends in Theory and Practice of Informatics, SOFSEM 2020, Limassol, Cyprus, January 20–24, 2020, Proceedings 46*, pages 313–324. Springer.

[Ferrettini et al., 2020b] Ferrettini, G., Aligon, J., and Soulé-Dupuy, C. (2020b). Improving on coalitional prediction explanation. In *Advances in Databases and Information Systems : 24th European Conference, ADBIS 2020, Lyon, France, August 25–27, 2020, Proceedings 24*, pages 122–135. Springer.

[Ferrettini et al., 2020c] Ferrettini, G., Aligon, J., Soulé-Dupuy, C., and Raynaut, W. (2020c). A framework for user assistance on predictive models. In *Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020)*, volume 2621, page 32. ceur-ws. org.

[Ferrettini et al., 2022] Ferrettini, G., Escriva, E., Aligon, J., Excoffier, J.-B., and Soulé-Dupuy, C. (2022). Coalitional strategies for efficient individual prediction explanation. *Information Systems Frontiers*, 24(1) :49–75.

[Feurer et al., 2015a] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., and Hutter, F. (2015a). Efficient and robust automated machine learning. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

[Feurer et al., 2019] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., and Hutter, F. (2019). Auto-sklearn : efficient and robust automated machine learning, part of the springer series on challenges in machine learning book series (sscml). *Berlin, Germany : Spinger. doi*, 10 :978–3.

[Feurer et al., 2018] Feurer, M., Letham, B., and Bakshy, E. (2018). Scalable meta-learning for bayesian optimization using ranking-weighted gaussian process ensembles. In *AutoML Workshop at ICML*, volume 7.

[Feurer et al., 2015b] Feurer, M., Springenberg, J., and Hutter, F. (2015b). Initializing bayesian hyperparameter optimization via meta-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

[Fleuret, 2004] Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine learning research*, 5(9).

[Forresi et al., 2021] Forresi, C., Gallinucci, E., Golfarelli, M., and Hamadou, H. B. (2021). A dataspace-based framework for olap analyses in a high-variety multistore. *The VLDB Journal*, 30(6) :1017–1040.

[Francia et al., 2022] Francia, M., Gallinucci, E., and Golfarelli, M. (2022). Cool : A framework for conversational olap. *Information Systems*, 104 :101752.

[Fürnkranz and Petrak, 2001] Fürnkranz, J. and Petrak, J. (2001). An evaluation of landmarking variants. In *Working Notes of the ECML/PKDD 2000 Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning*, pages 57–68. Citeseer.

[Futagami et al., 2021] Futagami, K., Fukazawa, Y., Kapoor, N., and Kito, T. (2021). Pairwise acquisition prediction with shap value interpretation. *The Journal of Finance and Data Science*, 7 :22–44.

[Gardin et al., 2019] Gardin, F., Gautiern, R., Goix, N., Ndiaye, B., and Schertzer, J.-M. (2019). Skope-rules.

[Garreau and von Luxburg, 2020] Garreau, D. and von Luxburg, U. (2020). Explaining the explainer : A first theoretical analysis of lime. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 1287–1296. PMLR.

[Ghahramani, 2003] Ghahramani, Z. (2003). Unsupervised learning. In *Summer school on machine learning*, pages 72–112. Springer.

[Giboney et al., 2015] Giboney, J. S., Brown, S. A., Lowry, P. B., and Nunamaker Jr, J. F. (2015). User acceptance of knowledge-based system recommendations : Explanations, arguments, and fit. *Decision Support Systems*, 72 :1–10.

[Gilpin et al., 2018] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations : An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

[Goldstein et al., 2015] Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015). Peeking inside the black box : Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1) :44–65.

[Golfarelli et al., 1998] Golfarelli, M., Maio, D., and Rizzi, S. (1998). The dimensional fact model : A conceptual model for data warehouses. *International Journal of Cooperative Information Systems*, 7(02n03) :215–247.

[Goodman and Flaxman, 2017] Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3) :50–57.

[Gurumoorthy et al., 2019] Gurumoorthy, K. S., Dhurandhar, A., Cecchi, G. A., and Aggarwal, C. C. (2019). Efficient data representation by selecting prototypes with importance weights. In Wang, J., Shim, K., and Wu, X., editors, *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*, pages 260–269. IEEE.

[Guyon et al., 2004] Guyon, I., Weston, J., Barnhill, S. D., and Vapnik, V. N. (2004). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46 :389–422.

[Gyulgyulyan et al., 2019] Gyulgyulyan, E., Aligon, J., Ravat, F., and Astsatryan, H. (2019). Data quality alerting model for big data analytics. In *New Trends in Databases and Information Systems : ADBIS 2019 Short Papers, Workshops BBIGAP, QAUCA, SemBDM, SIMPDA, M2P, MADEISD, and Doctoral Consortium, Bled, Slovenia, September 8–11, 2019, Proceedings 23*, pages 489–500. Springer.

[Gyulgyulyan et al., 2018] Gyulgyulyan, E., Ravat, F., Astsatryan, H., and Aligon, J. (2018). Data quality impact in business inteligence. In *2018 Ivannikov Memorial Workshop (IVMEM)*, pages 47–51. IEEE.

[Hamida et al., 2021] Hamida, A. B., Devanne, M., Weber, J., Truntzer, C., Derangère, V., Ghiringhelli, F., Forestier, G., and Wemmert, C. (2021). Deep learning for colon cancer histopathological images analysis. *Computers in Biology and Medicine*, 136 :104730.

[Hastie et al., 2009a] Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., and Friedman, J. (2009a). Overview of supervised learning. *The elements of statistical learning : Data mining, inference, and prediction*, pages 9–41.

[Hastie et al., 2009b] Hastie, T., Tibshirani, R., and Friedman, J. H. (2009b). *The elements of statistical learning : data mining, inference, and prediction*, volume 2. Springer.

[He et al., 2021] He, X., Zhao, K., and Chu, X. (2021). Automl : A survey of the state-of-the-art. *Knowledge-Based Systems*, 212 :106622.

[Henelius et al., 2014] Henelius, A., Puolamaki, K., Boström, H., Asker, L., and Papapetrou, P. (2014). A peek into the black box : exploring classifiers by randomization. *Data mining and knowledge discovery*, 28(5-6) :1503–1529. QC 20180119.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8) :1735–1780.

[Hooker et al., 2019] Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. (2019). *A Benchmark for Interpretability Methods in Deep Neural Networks*. Curran Associates Inc., Red Hook, NY, USA.

[Hossin and Sulaiman, 2015] Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2) :1.

[Ignatiev et al., 2020] Ignatiev, A., Narodytska, N., Asher, N., and Marques-Silva, J. (2020). From contrastive to abductive explanations and back again. In *International Conference of the Italian Association for Artificial Intelligence*, pages 335–355. Springer.

[Ignatiev et al., 2019] Ignatiev, A., Narodytska, N., and Marques-Silva, J. (2019). Abduction-based explanations for machine learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1511–1519.

[Ishanka et al., 2017] Ishanka, U., Yukawa, T., et al. (2017). The prefiltering techniques in emotion based place recommendation derived by user reviews. *Applied Computational Intelligence and Soft Computing*, 2017.

[Jones, 1972] Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

[Jordan and Mitchell, 2015] Jordan, M. I. and Mitchell, T. M. (2015). Machine learning : Trends, perspectives, and prospects. *Science*, 349(6245) :255–260.

[Jurado et al., 2022] Jurado, X., Reiminger, N., Benmoussa, M., Vazquez, J., and Wemmert, C. (2022). Deep learning methods evaluation to predict air quality based on computational fluid dynamics. *Expert Systems with Applications*, 203 :117294.

[Kaelbling et al., 1996] Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning : A survey. *Journal of artificial intelligence research*, 4 :237–285.

[Kaufman and Rousseeuw, 1990] Kaufman, L. and Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). *Finding groups in data : an introduction to cluster analysis*, 344 :68–125.

[Keany, 2020] Keany, E. (2020). Is this the best feature selection algorithm "borutashap"? https://medium.com/analytics-vidhya/is-this-the-best-feature-selection-algorithm-borutashap-8bc238aa1677. [Online ; accessed 2022-09-08].

[Keany, 2022] Keany, E. (2022). Borutashap package. https://github.com/Ekeany/Boruta-Shap/, Last accessed on 2022-10-22.

[Keet et al., 2015] Keet, C. M., Ławrynowicz, A., d'Amato, C., Kalousis, A., Nguyen, P., Palma, R., Stevens, R., and Hilario, M. (2015). The data mining optimization ontology. *Journal of web semantics*, 32 :43–53.

[Kendall, 1938] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2) :81–93.

[Kim et al., 2016] Kim, B., Khanna, R., and Koyejo, O. O. (2016). Examples are not enough, learn to criticize ! criticism for interpretability. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

[Kimball, 1996] Kimball, R. (1996). *The data warehouse toolkit : practical techniques for building dimensional data warehouses*. John Wiley & Sons, Inc.

[Kira and Rendell, 1992] Kira, K. and Rendell, L. A. (1992). A practical approach to feature selection. In *Machine learning proceedings 1992*, pages 249–256. Elsevier.

[Klaise et al., 2021] Klaise, J., Looveren, A. V., Vacanti, G., and Coca, A. (2021). Alibi explain : Algorithms for explaining machine learning models. *Journal of Machine Learning Research*, 22(181) :1–7.

[Knijnenburg et al., 2011] Knijnenburg, B. P., Reijmer, N. J., and Willemsen, M. C. (2011). Each to his own : how different users call for different interaction methods in recommender

systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 141–148.

[Kononenko and Bratko, 1991] Kononenko, I. and Bratko, I. (1991). Information-Based Evaluation Criterion for Classifier's Performance. *Machine Learning*, 6(1) :67–80.

[Kumar et al., 2020] Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. (2020). Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pages 5491–5500. PMLR.

[Leite et al., 2012] Leite, R., Brazdil, P., and Vanschoren, J. (2012). Selecting classification algorithms with active testing. In *Machine Learning and Data Mining in Pattern Recognition : 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings 8*, pages 117–131. Springer.

[Liao et al., 2020] Liao, Q. V., Gruen, D., and Miller, S. (2020). *Questioning the AI : Informing Design Practices for Explainable AI User Experiences*, page 1–15. Association for Computing Machinery, New York, NY, USA.

[Linardatos et al., 2021] Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable ai : A review of machine learning interpretability methods. *Entropy*, 23(1) :18.

[Lipovetsky and Conklin, 2001] Lipovetsky, S. and Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4) :319–330.

[Lipton, 2018] Lipton, Z. C. (2018). The mythos of model interpretability : In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3) :31–57.

[Liu et al., 2021] Liu, Y., Khandagale, S., White, C., and Neiswanger, W. (2021). Synthetic benchmarks for scientific research in explainable machine learning. *arXiv preprint arXiv :2106.12543*.

[Liu et al., 2022] Liu, Y., Liu, Z., Luo, X., and Zhao, H. (2022). Diagnosis of parkinson's disease based on shap value feature selection. *Biocybernetics and Biomedical Engineering*, 42(3) :856–869.

[Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

[Maass and Storey, 2021] Maass, W. and Storey, V. C. (2021). Pairing conceptual modeling with machine learning. *Data & Knowledge Engineering*, 134 :101909.

[Man and Chan, 2021] Man, X. and Chan, E. P. (2021). The best way to select features ? comparing mda, lime, and shap. *The Journal of Financial Data Science*, 3(1) :127–139.

[Mantovani et al., 2015] Mantovani, R. G., Rossi, A. L., Vanschoren, J., and de Carvalho, A. C. (2015). Meta-learning recommendation of default hyper-parameter values for svms in classification tasks. In *MetaSel@ PKDD/ECML*, pages 80–92.

[Marcel et al., 2023] Marcel, P., Peralta, V., and Amer-Yahia, S. (2023). Data narration for the people : Challenges and opportunities.

[Markus et al., 2021] Markus, A. F., Kors, J. A., and Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care : A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113 :103655.

[Marques-Silva and Ignatiev, 2022] Marques-Silva, J. and Ignatiev, A. (2022). Delivering trustworthy ai through formal xai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12342–12350.

[Martens and Provost, 2014] Martens, D. and Provost, F. (2014). Explaining data-driven document classifications. *MIS quarterly*, 38(1) :73–100.

[McNee et al., 2006] McNee, S. M., Riedl, J., and Konstan, J. A. (2006). Being accurate is not enough : How accuracy metrics have hurt recommender systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, page 1097–1101, New York, NY, USA. Association for Computing Machinery.

[Meske et al., 2022] Meske, C., Bunde, E., Schneider, J., and Gersch, M. (2022). Explainable artificial intelligence : objectives, stakeholders, and future research opportunities. *Information Systems Management*, 39(1) :53–63.

[Miettinen, 2012] Miettinen, K. (2012). *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media.

[Miller, 2019] Miller, T. (2019). Explanation in artificial intelligence : Insights from the social sciences. *Artificial intelligence*, 267 :1–38.

[Molnar, 2020] Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

[Monsarrat et al., 2022] Monsarrat, P., Bernard, D., Marty, M., Cecchin-Albertoni, C., Doumard, E., Gez, L., Aligon, J., Vergnes, J.-N., Casteilla, L., and Kemoun, P. (2022). Systemic periodontal risk score using an innovative machine learning strategy : an observational study. *Journal of Personalized Medicine*, 12(2) :217.

[Moreno-Sanchez, 2021] Moreno-Sanchez, P. A. (2021). An automated feature selection and classification pipeline to improve explainability of clinical prediction models. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 527–534. IEEE.

[Mosqueira-Rey et al., 2023] Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., and Fernández-Leal, Á. (2023). Human-in-the-loop machine learning : A state of the art. *Artificial Intelligence Review*, 56(4) :3005–3054.

[Nauta et al., 2022] Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. (2022). From anecdotal evidence to quantitative evaluation methods : A systematic review on evaluating explainable ai. *arXiv preprint arXiv :2201.08164*.

[Nguyen and Martínez, 2020] Nguyen, A.-p. and Martínez, M. R. (2020). On quantitative aspects of model interpretability. *arXiv :2007.07584 [cs, stat]*. arXiv : 2007.07584.

[Nie et al., 2010] Nie, F., Huang, H., Cai, X., and Ding, C. (2010). Efficient and robust feature selection via joint $\ell 2$, 1-norms minimization. *Advances in neural information processing systems*, 23.

[Nogueira, 14] Nogueira, F. (2014–). Bayesian Optimization : Open source constrained global optimization tool for Python.

[Omeiza et al., 2021] Omeiza, D., Webb, H., Jirotka, M., and Kunze, L. (2021). Explanations in autonomous driving : A survey. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–21.

[Palacio et al., 2021] Palacio, S., Lucieri, A., Munir, M., Ahmed, S., Hees, J., and Dengel, A. (2021). Xai handbook : Towards a unified framework for explainable ai. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3766–3775.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830.

[Peng et al., 2005] Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8) :1226–1238.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

[Probst et al., 2019] Probst, P., Boulesteix, A.-L., and Bischl, B. (2019). Tunability : Importance of hyperparameters of machine learning algorithms. *The Journal of Machine Learning Research*, 20(1) :1934–1965.

[Pugliese et al., 2021] Pugliese, R., Regondi, S., and Marini, R. (2021). Machine learning-based approach : Global trends, research directions, and regulatory standpoints. *Data Science and Management*, 4 :19–29.

[Rai, 2020] Rai, A. (2020). Explainable ai : From black box to glass box. *Journal of the Academy of Marketing Science*, 48 :137–141.

[Ramana et al., 2011] Ramana, B. V., Babu, M. S. P., Venkateswarlu, N., et al. (2011). A critical study of selected classification algorithms for liver disease diagnosis. *International Journal of Database Management Systems*, 3(2) :101–114.

[Ramos et al., 2020] Ramos, G., Meek, C., Simard, P., Suh, J., and Ghorashi, S. (2020). Interactive machine teaching : a human-centered approach to building machine-learned models. *Human–Computer Interaction*, 35(5-6) :413–451.

[Ravat and Zhao, 2019a] Ravat, F. and Zhao, Y. (2019a). Data lakes : Trends and perspectives. In *Database and Expert Systems Applications : 30th International Conference, DEXA 2019, Linz, Austria, August 26–29, 2019, Proceedings, Part I 30*, pages 304–313. Springer.

[Ravat and Zhao, 2019b] Ravat, F. and Zhao, Y. (2019b). Metadata management for data lakes. In *New Trends in Databases and Information Systems : ADBIS 2019 Short Papers, Workshops BBIGAP, QAUCA, SemBDM, SIMPDA, M2P, MADEISD, and Doctoral Consortium, Bled, Slovenia, September 8–11, 2019, Proceedings 23*, pages 37–44. Springer.

[Raynaut, 2018] Raynaut, W. (2018). *Perspectives de Méta-Analyse pour un Environnement d'aide à la Simulation et Prédiction*. Thèse de doctorat, Université de Toulouse, Université Toulouse III-Paul Sabatier, Toulouse, France.

[Raynaut et al., 2017a] Raynaut, W., Aligon, J., Roussille, P., Soulé-Dupuy, C., and Valles-Parlangeau, N. (2017a). Towards a meta-analysis-based user assistant for analysis processes.

[Raynaut et al., 2017b] Raynaut, W., Soulé-Dupuy, C., and Vallès-Parlangeau, N. (2017b). Dissimilarités entre jeux de données. *Ingénierie des Systèmes d Inf.*, 22(3) :35–63.

[Ribeiro et al., 2016] Ribeiro, M., Singh, S., and Guestrin, C. (2016). "why should i trust you ?" : Explaining the predictions of any classifier. pages 97–101.

[Ribeiro et al., 2018] Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors : High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

[Robnik-Šikonja and Bohanec, 2018] Robnik-Šikonja, M. and Bohanec, M. (2018). *Perturbation-Based Explanations of Prediction Models*, pages 159–175. Springer International Publishing, Cham.

[Rosenfeld, 2021] Rosenfeld, A. (2021). Better metrics for evaluating explainable artificial intelligence. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, page 45–50, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

[Ross, 2017] Ross, S. M. (2017). *Introductory statistics*. Academic Press.

[Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20 :53–65.

[Rudnichenko et al., 2020] Rudnichenko, N., Vychuzhanin, V., Petrov, I., and Shibaev, D. (2020). Decision support system for the machine learning methods selection in big data mining. In *CMIS*, pages 872–885.

[Runkler, 2020] Runkler, T. A. (2020). *Data analytics*. Springer.

[Safdar et al., 2018] Safdar, S., Zafar, S., Zafar, N., and Khan, N. F. (2018). Machine learning based decision support systems (dss) for heart disease diagnosis : a review. *Artificial Intelligence Review*, 50 :597–623.

[Semenova et al., 2022] Semenova, L., Rudin, C., and Parr, R. (2022). On the existence of simpler machine learning models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1827–1858, New York, NY, USA. Association for Computing Machinery.

[Settles, 2009] Settles, B. (2009). Active learning literature survey.

[Shani et al., 2013] Shani, G., Rokach, L., Shapira, B., Hadash, S., and Tangi, M. (2013). Investigating confidence displays for top-n recommendations. *Journal of the American Society for Information Science and Technology*, 64(12) :2548–2563.

[Shapley et al., 1953] Shapley, L. S. et al. (1953). A value for n-person games.

[Sharif Razavian et al., 2014] Sharif Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). Cnn features off-the-shelf : an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813.

[Sharifani and Amini, 2023] Sharifani, K. and Amini, M. (2023). Machine learning and deep learning : A review of methods and applications. *World Information Technology and Engineering Journal*, 10(07) :3897–3904.

[Shrikumar et al., 2017] Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR.

[Siegel, 2013] Siegel, E. (2013). *Predictive analytics : The power to predict who will click, buy, lie, or die*. John Wiley & Sons.

[Singh et al., 2021] Singh, R., Abbas, A., Beqiri, S., Korot, E., Struyven, R., Keane, P., et al. (2021). Exploring the what-if-tool as a solution for machine learning explainability in clinical practice. *Investigative Ophthalmology & Visual Science*, 62(8) :79–79.

[Slack et al., 2020] Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling lime and shap : Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

[Smith et al., 1988] Smith, J. W., Everhart, J. E., Dickson, W., Knowler, W. C., and Johannes, R. S. (1988). Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, page 261. American Medical Informatics Association.

[Snoek et al., 2012] Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25.

[Stone, 1974] Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2) :111–147.

[Štrumbelj and Kononenko, 2008] Štrumbelj, E. and Kononenko, I. (2008). Towards a model independent method for explaining classification for individual instances. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 273–282. Springer.

[Strumbelj and Kononenko, 2010] Strumbelj, E. and Kononenko, I. (2010). An Efficient Explanation of Individual Classifications Using Game Theory. *J. Mach. Learn. Res.*, 11 :1–18. Publisher : JMLR.org.

[Sun et al., 2012] Sun, X., Liu, Y., Li, J., Zhu, J., Chen, H., and Liu, X. (2012). Feature evaluation and selection with cooperative game theory. *Pattern recognition*, 45(8) :2992–3002.

[Sutton and Barto, 2018] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning : An introduction*. MIT press.

[Swartout and Moore, 1993] Swartout, W. R. and Moore, J. D. (1993). Explanation in second generation expert systems. In *Second generation expert systems*, pages 543–585. Springer.

[Tan et al., 2018] Tan, S., Caruana, R., Hooker, G., and Lou, Y. (2018). Distill-and-compare : Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 303–310, New York, NY, USA. Association for Computing Machinery.

[Thrun and Pratt, 1998] Thrun, S. and Pratt, L. (1998). Learning to learn : Introduction and overview. In *Learning to learn*, pages 3–17. Springer.

[Tintarev and Masthoff, 2015] Tintarev, N. and Masthoff, J. (2015). Explaining recommendations : Design and evaluation. In *Recommender systems handbook*, pages 353–382. Springer.

[Tukey, 1962] Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics*, 33(1) :1–67.

[Van den Broeck et al., 2021] Van den Broeck, G., Lykov, A., Schleich, M., and Suciu, D. (2021). On the tractability of shap explanations. In *Proceedings of AAAI*.

[Vanschoren, 2019] Vanschoren, J. (2019). Meta-learning. *Automated machine learning : methods, systems, challenges*, pages 35–61.

[Vanschoren et al., 2013] Vanschoren, J., van Rijn, J. N., Bischl, B., and Torgo, L. (2013). OpenML : Networked Science in Machine Learning. *SIGKDD Explorations*, 15(2) :49–60. Place : New York, NY, USA Publisher : ACM.

[Vartak et al., 2015] Vartak, M., Ortiz, P., Siegel, K., Subramanyam, H., Madden, S., and Zaharia, M. (2015). Supporting fast iteration in model building. In *NIPS Workshop LearningSys*, pages 1–6.

[Vartak et al., 2016] Vartak, M., Subramanyam, H., Lee, W.-E., Viswanathan, S., Husnoo, S., Madden, S., and Zaharia, M. (2016). Modeldb : a system for machine learning model management. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, pages 1–3.

[Vermeire et al., 2021] Vermeire, T., Laugel, T., Renard, X., Martens, D., and Detyniecki, M. (2021). How to choose an explainability method ? towards a methodical implementation of xai in practice. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 521–533, Cham. Springer International Publishing.

[Wachter et al., 2017] Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box : Automated decisions and the gdpr. *Harv. JL & Tech.*, 31 :841.

[Wang et al., 2023] Wang, H., Doumard, E., Soulé-Dupuy, C., Kémoun, P., Aligon, J., and Monsarrat, P. (2023). Explanations as a new metric for feature selection : a systematic approach. *IEEE Journal of Biomedical and Health Informatics*.

[Wang and Tao, 2008] Wang, J. and Tao, Q. (2008). Machine learning : The state of the art. *IEEE Intelligent Systems*, 23(6) :49–55.

[Wang et al., 2017] Wang, Q., Ma, J., Liao, X., and Du, W. (2017). A context-aware researcher recommendation system for university-industry collaboration on r&d projects. *Decision Support Systems*, 103 :46–57.

[Watkins and Dayan, 1992] Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8 :279–292.

[Weerts et al., 2019] Weerts, H. J., van Ipenburg, W., and Pechenizkiy, M. (2019). A human-grounded evaluation of shap for alert processing. *arXiv preprint arXiv :1907.03324*.

[Wexler, 2018] Wexler, J. (2018). The what-if tool : code-free probing of machine learning models. *Google AI blog*.

[Wexler et al., 2019] Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., and Wilson, J. (2019). The what-if tool : Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1) :56–65.

[Wexler et al., 2020] Wexler, J., Pushkarna, M., Robinson, S., Bolukbasi, T., and Zaldivar, A. (2020). Probing ml models for fairness with the what-if tool and shap : hands-on tutorial. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 705–705.

[Wistuba et al., 2015] Wistuba, M., Schilling, N., and Schmidt-Thieme, L. (2015). Learning hyperparameter optimization initializations. In *2015 IEEE international conference on data science and advanced analytics (DSAA)*, pages 1–10. IEEE.

[Wolpert and Macready, 1997] Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1) :67–82.

[Wu et al., 2022] Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135 :364–381.

[Xin et al., 2018] Xin, D., Ma, L., Liu, J., Macke, S., Song, S., and Parameswaran, A. (2018). Accelerating human-in-the-loop machine learning : Challenges and opportunities. In *Proceedings of the second workshop on data management for end-to-end machine learning*, pages 1–4.

[Yahyaoui et al., 2019] Yahyaoui, A., Jamil, A., Rasheed, J., and Yesiltepe, M. (2019). A decision support system for diabetes prediction using machine learning and deep learning techniques. In *2019 1st International informatics and software engineering conference (UBMYK)*, pages 1–4. IEEE.

[Yang and Moody, 1999] Yang, H. and Moody, J. (1999). Feature selection based on joint mutual information. In *Proceedings of international ICSC symposium on advances in intelligent data analysis*, volume 1999, pages 22–25. Citeseer.

[Yeh et al., 2019] Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. (2019). On the (in)fidelity and sensitivity of explanations. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Zhang et al., 2020] Zhang, Y., Liao, Q. V., and Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *FAccT Proceedings*.

[Zhao et al., 2021] Zhao, Y., Ravat, F., Aligon, J., Soulé-Dupuy, C., Ferrettini, G., and Megdiche, I. (2021). Analysis-oriented metadata for data lakes. In *Proceedings of the 25th International Database Engineering & Applications Symposium*, pages 194–203.

[Zhao and Liu, 2007] Zhao, Z. and Liu, H. (2007). Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1151–1157.

[Zhong and Negre, 2022] Zhong, J. and Negre, E. (2022). A 3 r : Argumentative explanations for recommendations. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–9. IEEE.

[Zhou et al., 2021] Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. (2021). Evaluating the quality of machine learning explanations : A survey on methods and metrics. *Electronics*, 10(5).

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society : series B (statistical methodology)*, 67(2) :301–320.

[Štrumbelj and Kononenko, 2014] Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3) :647–665.