Classification probabiliste pour la prédiction et l'explication d'événements de santé défavorables et évitables en EHPAD

Clara Charon

Thèse présentée pour l'obtention du grade de

Docteur de Sorbonne Université

Spécialité

Informatique

École doctorale

Informatique, Télécommunication
et Électronique de Paris (ED130)

Classification probabiliste pour la prédiction et l'explication d'événements de santé défavorables et évitables en EHPAD

---

par Clara CHARON

Soutenue publiquement le : *9 juillet 2024*

Devant un jury composé de :

Sébastien Destercke, Directeur de recherche, CNRS, UTC	*Rapporteur*
Maturin Tabue-Teguo, Professeur, Université des Antilles	*Rapporteur*
Véronique Delcroix, Maître de Conférences, AMIH, UPHF	*Examinatrice*
Marie-Jeanne Lesot, Professeure, LIP6, Sorbonne Université	*Présidente du jury*
Patrice Perny, Professeur, LIP6, Sorbonne Université	*Directeur de thèse*
Pierre-Henri Wuillemin, Maître de Conférences, LIP6, Sorbonne Université	*Encadrant*
Joël Belmin, Professeur, APHP, Sorbonne Université	*Encadrant*

Merci à TOUMLILT, 2021 pour l'origine du template initial.

Remerciements

Tout d'abord, je tiens à remercier chaleureusement ceux qui ont créé ce sujet de stage, devenu thèse : Joël Belmin, Philippe Thalamy et Pierre-Henri Wuillemin. Merci pour tout ce qu'ils m'ont appris. Plus particulièrement, merci au Pr Belmin pour son importante expertise, mais aussi son remarquable investissement et encadrement. Merci à Philippe pour son intérêt sur le sujet, d'y avoir cru et de nous avoir fait confiance. Merci aussi à lui d'avoir créé autant de moments conviviaux chez Teranga ! Et bien sûr, un grand merci à Pierre-Henri pour son super encadrement, tout ce temps qu'il m'a accordé, sa bienveillance, sa patience, sa grande pédagogie et de m'avoir permis d'arriver au bout. Merci aussi pour tous ces chaleureux repas. Merci aussi à Patrice Perny d'avoir accepté de diriger cette thèse et pour tous ses bons conseils.

Je remercie sincèrement Sébastien Destercke et Maturin Tabue-Teguo d'avoir accepté de rapporter ma thèse. Un grand merci également à Véronique Delcroix et Marie-Jeanne Lesot d'avoir accepté d'évaluer ma thèse en tant qu'examinatrices. Je remercie aussi Xavier Tannier d'avoir fait partie de mon comité de suivi aux côtés de Marie-Jeanne Lesot.

J'ai eu la chance d'avoir deux super duos dans ce fameux bureau 401. Tout d'abord "mes chics types" Gaspard et Marvin, je ne me serais probablement pas lancée dans cette intimidante aventure sans leur exemple, imparfait et identifiable. Merci de m'avoir aussi bien accueillie à mon arrivée en stage, de m'avoir appris autant de choses utiles qu'inutiles, merci pour ces incessantes discussions, encouragements, débats et rires. Puis il y a eu Margot et Mahdi, ils vont bientôt finir eux aussi, et j'ai hâte de cette fameuse remise des diplômes tous ensemble l'année prochaine ! Merci pour toute cette entraide, ces démotivations et remotivations ensemble, cette bonne humeur et toute cette nourriture grignotée dans ce bureau ! Merci aussi à toutes les autres personnes que j'ai côtoyées dans ce bureau et qui ont égayé quelques journées de travail.

Pas en 401, mais presque comme si, il y avait Thibaut. Merci à lui d'accepter de traîner avec nous, de nous donner des bons conseils sur le travail ou sur le sport et d'être toujours partant pour une IPA ! Merci globalement à tous les membres de ce couloir avec qui j'ai pu partager des déjeuners aux arènes, des pauses-café dans

cette salle de "convivialité" ou des verres au Buisson Ardent. Et au-dessus, au 5ème, merci à Garance de m'avoir permis de ne pas passer cette première conférence à Nice seule, et pour tous les autres bons moments et incessantes questions que j'avais pour elle. Merci aussi à elle de m'avoir présenté aux doctorants de son équipe et merci à eux de m'avoir accepté quelques fois en exil ces derniers mois !

Cela fait presque 10 ans que j'ai commencé mes études à Jussieu. Cette université et moi-même avons eu le temps de bien changer entre temps, par exemple, de nom pour elle et moi de voie. Son ouverture sur la multidisciplinarité et ses gens passionnants m'ont permis d'en arriver à ce stade que je n'aurais jamais imaginé. Je suis reconnaissante de tout ce que j'aurais appris ici, ainsi que toutes les opportunités et belles rencontres que ce lieu m'a apportées.

Merci à Élodie de m'avoir managée et soutenue pendant ces trois années. Merci à l'équipe NETSmart pour tout ce que vous avez accompli. Merci aussi à tous mes anciens collègues de chez Teranga, ce travail de recherche était souvent isolé, mais cela ne m'a pas empêché de passer plein de bons moments au siège, en séminaire ou même sur TeamSpeak !

Merci à tous mes autres amis qui sont doctorants ou désormais docteurs, qu'importe la discipline et le labo, pouvoir s'allier dans l'universalité de la condition de doctorant fait toujours du bien. . . Et surtout, bon courage à ceux qui n'ont pas encore fini !

Merci aussi à tous mes amis, même ceux qui n'ont rien à voir avec la recherche, qui m'ont accompagné dans les moments de détente dont j'avais bien besoin, votre soutien et votre intérêt m'ont toujours beaucoup touché, et que je vous connaisse de primaire, de collège, de lycée, de Jussieu, du travail ou par d'autres biais détournés qui ont mené à des amitiés sincères, j'espère vous garder encore très longtemps auprès de moi. Je suis vraiment désolée de ne pas tous vous nommer individuellement, mais vous savez à quel point vous comptez pour moi !

Merci à toute ma famille évidemment, se retrouver tous ensemble le dimanche soir est un rituel essentiel pour moi. Vous avez pu suivre toutes mes avancées (et mes reculs), alors merci pour votre investissement ! Merci aussi à ma "jolie" famille pour tous ces bons moments passés ensemble et votre soutien.

Et enfin, merci infiniment à Zac. Merci de me regarder avec ces yeux qui me font sentir si exceptionnelle. Merci d'être mon plus grand supporter et de t'être autant investi dans cette thèse. J'ai hâte de toutes ces nouvelles aventures avec toi.

Résumé

L'EHPAD, établissement d'hébergement pour personnes âgées dépendantes, constitue une option à laquelle a recours une population nombreuse et croissante, lorsque pour diverses raisons, et notamment de santé, il n'est plus possible de vivre à domicile. Avec le développement des nouvelles technologies informatiques dans le domaine de la santé, un nombre croissant d'établissements de santé sont équipés de systèmes d'information regroupant les données administratives et médicales des patients ainsi que des informations sur les soins qui leur sont prodigués. Parmi ces systèmes, les dossiers médicaux électroniques (DME) émergent comme des outils essentiels, offrant un accès rapide et aisé aux informations des patients dans le but d'améliorer la qualité et la sécurité des soins. Dans ce travail, nous utilisons les données anonymisées des DME de NETSoins, un logiciel largement utilisé dans les EHPAD en France, afin de proposer et d'analyser des classifieurs capables de prédire plusieurs événements de santé défavorables chez les personnes âgées qui sont potentiellement modifiables par des interventions de santé appropriées. Notre démarche se concentre notamment sur l'utilisation de méthodes capables de fournir des explications, notamment les modèles graphiques probabilistes tels que les réseaux bayésiens. Après un prétraitement complexe pour adapter des données d'une base événementielle en données utilisables par un apprentissage statistique, tout en conservant leur cohérence médicale, nous avons développé une méthodologie d'apprentissage mise en œuvre dans trois expériences de classification probabiliste utilisant des réseaux bayésiens distincts, ciblant différents événements : le risque de survenue de la première escarre, le risque d'hospitalisation en urgence à l'entrée du résident en EHPAD, et le risque de fracture dans les premiers mois d'hébergement. Pour chaque cible, nous avons comparé les performances de notre classifieur de réseaux bayésiens selon divers critères avec d'autres méthodes de *machine learning* ainsi qu'avec les pratiques actuellement utilisées en EHPAD pour prédire ces risques. Nous avons aussi confronté les résultats des réseaux bayésiens à l'expertise clinique. Cette étude démontre la possibilité de prédire ces événements à partir des données déjà collectées en routine par les soignants, ouvrant ainsi la voie à de nouveaux outils de prédiction intégrables directement dans le logiciel déjà utilisé par ces professionnels.

Mots-clés : Réseaux Bayésiens, Classification Probabiliste, Apprentissage Machine, EHPAD, Personnes Âgées, Dossier Médicaux Électroniques

Abstract

Nursing homes, which provide housing for dependent elderly people, are an option used by a large and growing population when, for a variety of reasons, including health, it is no longer possible for them to live at home. With the development of new information technologies in the health sector, an increasing number of health care facilities are equipped with information systems that group together administrative and medical data of patients as well as information on the care they receive. Among these systems, electronic health records (EHRs) have emerged as essential tools, providing quick and easy access to patient information in order to improve the quality and safety of care. We use the anonymized data of the EHRs from NETSoins, a software widely used in nursing homes in France, to propose and analyze classifiers capable of predicting several adverse health events in the elderly that are potentially modifiable by appropriate health interventions. Our approach focuses in particular on the use of methods that can provide explanations, such as probabilistic graphical models, including Bayesian networks.

After a complex preprocessing step to adapt event-based data into data suitable for statistical learning while preserving their medical coherence, we have developed a learning method applied in three probabilistic classification experiments using Bayesian networks, targeting different events : the risk of occurrence of the first pressure ulcer, the risk of emergency hospitalization upon the resident's entry into the nursing home, and the risk of fracture in the first months of housing. For each target, we have compared the performance of our Bayesian network classifier according to various criteria with other machine learning methods as well as with the practices currently used in nursing homes to predict these risks. We have also compared the results of the Bayesian networks with clinical expertise.

This study demonstrates the possibility of predicting these events from the data already collected in routine by caregivers, thus paving the way for new predictive tools that can be integrated directly into the software already used by these professionals.

Keywords : Bayesian Network, Probabilistic Classification, Machine Learning, Nursing Homes, Elderly, Electronic Health Records.

Table des matières

Introduction

Les résidents des Établissements d'Hébergement pour Personnes Âgées Dépendantes (EHPAD) forment une population nombreuse dont la santé est complexe et leur prise en charge représente un enjeu majeur de santé publique, dans un contexte où la population vieillit et où le nombre de personnes âgées dépendantes augmente de façon importante. La prévention et la prédiction d'événements de santé défavorables et évitables sont donc des priorités pour améliorer la qualité des soins des résidents.

L'existence d'une base de données de grande taille en EHPAD nous offre une opportunité unique de tirer parti de cette expérience accumulée pour améliorer la prise en charge des résidents. Il s'agit des données des dossiers médicaux électroniques (DME) d'un logiciel de soins utilisé dans les EHPAD en France : NETSoins, édité par Teranga Software. Le logiciel NETSoins a atteint une certaine maturité avec 16 ans d'existence et actuellement plus de 220 000 dossiers résidents et plus de 100 000 utilisateurs, ce qui en fait le logiciel pour EHPAD le plus déployé en France.

Le projet est donc d'analyser ces données récoltées, et de voir dans quelle mesure celles-ci pourraient aider à faire des prédictions sur la santé des résidents et aller vers la création d'un véritable outil d'aide à la décision et à la prédiction pour les soignants. Cette version améliorée de NETSoins, appelée NETSmart, a pour but de simplifier la tâche des soignants en les aidant à prendre la bonne décision rapidement et en mettant en avant les éléments qu'ils auraient pu manquer.

Contrairement aux données hospitalières, les données en EHPAD sont collectées sur une période plus longue, ce qui permet de suivre l'évolution de l'état de santé des résidents dans le temps et d'identifier potentiellement des nouveaux facteurs de risque, spécifiques aux personnes âgées dans ce contexte.

Dans ce contexte, l'Intelligence Artificielle (IA) et plus particulièrement le *machine learning* offrent des perspectives prometteuses pour le développement de modèles de prédiction. Il est difficile de trouver une définition formelle du *machine learning*, mais il peut être considéré comme le sous-domaine de l'IA qui se concentre sur le développement d'algorithmes permettant aux ordinateurs de découvrir automatiquement des schémas dans les données et de s'améliorer avec l'expérience, sans recevoir d'instructions explicites [Sanchez-Martinez et al., 2022]. Plutôt que d'être spécifique à un domaine scientifique, le *machine learning* se situe à l'intersection des statistiques, des mathématiques et de l'informatique, avec des outils analytiques qui transcendent les frontières entre les trois disciplines [Beam et Kohane, 2018]. Les algorithmes de *machine learning* se distinguent par leur approche de l'apprentissage basée sur les données, contrairement aux modèles basés sur des règles qui s'appuient sur la connaissance du domaine.

L'utilisation de l'IA dans le domaine de la santé soulève des questions éthiques, réglementaires et techniques. En particulier, l'explicabilité des modèles est devenue un enjeu majeur pour les professionnels de santé, qui doivent être en mesure de comprendre et d'expliquer les prédictions du modèle pour avoir confiance en elles. C'est pourquoi nous nous intéresserons à l'*Explainable* AI (XAI), qui vise à développer des méthodes permettant de rendre les modèles de *machine learning* plus compréhensibles et plus transparents.

Pour garantir l'explicabilité et la pertinence, un modèle de *machine learning* doit forcément se construire en dialogue entre les personnes qui ont la connaissance du domaine d'application, que nous appellerons "expert", et celui qui le développe. Dans notre cadre, nous avons pu bénéficier de l'avis d'experts en gériatrie, la spécialité de médecine sur les soins aux personnes âgées, en particulier celles qui souffrent de maladies liées à l'âge et de problèmes de santé complexes.

Les réseaux bayésiens sont particulièrement adaptés à ce dialogue entre expert et développeur. Ce sont des modèles graphiques probabilistes utilisés pour représenter les dépendances conditionnelles entre des variables aléatoires et peuvent être utilisés comme modèle de *machine learning* explicable.

Dans cette thèse, nous proposons donc de développer et d'évaluer des approches de classification probabiliste et de réseaux bayésiens pour la prédiction et l'explication d'événements de santé défavorables et évitables en EHPAD, en utilisant des données issues du Dossier Médical Électronique (DME) et en prenant en compte les spécificités de la population gériatrique.

Notre objectif est de contribuer à l'amélioration de la prise en charge des résidents en EHPAD, en développant des outils de prédiction et d'explication fiables et transparents, qui permettront aux professionnels de santé de prendre des décisions éclairées et d'améliorer la qualité des soins.

Contexte de la thèse

Cette thèse a bénéficié du dispositif des conventions industrielles de formation par la recherche (CIFRE), géré par l'Association Nationale de la Recherche et de la Technologie (ANRT). Cette convention permet une collaboration entre une entreprise privée et un laboratoire de recherche public. Ici, Teranga Software s'est donc associé au laboratoire d'informatique de Sorbonne Université, le LIP6, et plus particulièrement avec Pierre-Henri Wuillemin de l'équipe DECISION, ainsi qu'au Pr

Joël Belmin, gériatre du Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé (LIMICS) de Sorbonne Université et Chef de service à l'hôpital universitaire Charles Foix (Ivry-sur-Seine).

Organisation du manuscrit et contributions

Ce manuscrit se compose de deux parties.

La première est dédiée à l'état de l'art et contient trois chapitres. Le chapitre 1 introduit les notions de base utiles à la compréhension de cette thèse et présente le modèle du réseau bayésien. Le chapitre 2 présente d'abord la classification probabiliste pour ainsi définir les réseaux bayésiens en tant que classifieur, puis présente d'autres méthodes de classification populaire, ainsi que les problématiques liées à la création des classifieurs, à leur évaluation, et enfin, à leur explicabilité. Enfin, le chapitre 3 contient un état de l'art plus concret, sur l'utilisation de modèle de *machine learning* dans le domaine de la santé particulièrement.

La deuxième partie contient quatre chapitres et se consacre à la présentation de nos contributions sur les réseaux bayésiens en tant qu'outil de classification pour la prédiction en EHPAD. Le chapitre 4 présente le contexte de la base de données que nous avons pu utiliser, ainsi que le prétraitement général appliqué. Le chapitre 5 expose nos résultats sur le premier événement de santé défavorable et évitable que nous avons voulu prédire : la survenue de la première escarre. Le chapitre 6 se consacre à la prédiction du risque d'hospitalisation en urgence à l'entrée du résident et le chapitre 7 à la prédiction du risque de fracture à partir des données des premiers mois dans l'établissement.

Le manuscrit se conclut par une analyse critique de nos résultats et des perspectives de développements futurs. Afin de faciliter la lecture de ce document et compte tenu de la diversité des sujets abordés, nous avons choisi d'inclure une bibliographie à la fin de chaque chapitre.

Certains des résultats présentés dans cette thèse ont été publiés dans :

- *One Month Prediction of Pressure Ulcers in Nursing Home Residents with Bayesian Network*, Charon, Wuillemin, Havreng-Théry et al., 2024
- *Improving Pressure Ulcers Prediction in Nursing Homes with ML Algorithm*, Charon, Wuillemin et Belmin, 2023
- *Learning Bayesian Networks for the Prediction of Unfavorable Health Events in Nursing Homes*, Charon, Wuillemin et Belmin, 2022

Deux autres articles sont également en finalisation de rédaction.

Références

BEAM, Andrew L. et Isaac S. KOHANE (2018). “Big Data and Machine Learning in Health Care”. In : *JAMA* 319.13, p. 1317-1318 (cf. p. 3).

CHARON, Clara, Pierre-Henri WUILLEMIN et Joël BELMIN (2023). “Improving Pressure Ulcers Prediction in Nursing Homes with ML Algorithm”. In : *Studies in Health Technology and Informatics* 302, p. 350-351 (cf. p. 5).

– (2022). “Learning Bayesian Networks for the Prediction of Unfavorable Health Events in Nursing Homes”. In : *Studies in Health Technology and Informatics* 294, p. 147-148 (cf. p. 5).

CHARON, Clara, Pierre-Henri WUILLEMIN, Charlotte HAVRENG-THÉRY et Joël BELMIN (2024). “One Month Prediction of Pressure Ulcers in Nursing Home Residents with Bayesian Networks”. In : *Journal of the American Medical Directors Association*, S1525–8610(24)00070-7 (cf. p. 5).

SANCHEZ-MARTINEZ, Sergio, Oscar CAMARA, Gemma PIELLA et al. (2022). “Machine Learning for Clinical Decision-Making : Challenges and Opportunities in Cardiovascular Imaging”. In : *Frontiers in Cardiovascular Medicine* 8. Publisher : Frontiers (cf. p. 3).

Partie I

État de l'art

1 Réseaux Bayésiens : modèle, interprétation et apprentissage

Le traitement de l'incertitude est un élément essentiel du travail de prédiction : il s'agit de prendre en compte les incertitudes issues d'une base bruitée, de questions médicales qui incluent naturellement une grande variabilité, voire une incertitude quant à la cible à prédire. Il existe de nombreuses façons de traiter cette incertitude, nous nous attacherons à utiliser l'une des mieux fondées : les probabilités [BILLINGSLEY, 1995]. Toutefois, gérer les incertitudes avec un modèle probabiliste pose un enjeu important de complexité. Les modèles probabilistes sont victimes de ce qu'on appelle la *malédiction de la dimensionnalité* : le nombre de paramètres évolue exponentiellement selon la dimension du modèle. Nous nous attaquerons ici à des problèmes impliquant un grand nombre de variables pour lesquels il n'est pas raisonnable de représenter ni de manipuler de façon exhaustive la distribution de probabilité associée. Les réseaux bayésiens fournissent une représentation compacte et donnent par ailleurs une grille de lecture plus aisée pour analyser les comportements induits par ces distributions complexes à grande dimension de probabilité. Il existe aussi dans ce domaine des algorithmes d'apprentissage automatique.

Ce chapitre présente succinctement les concepts fondamentaux de la théorie des probabilités. Puis il décrit, d'une manière plus précise, le modèle des réseaux bayésiens, ainsi que les outils d'interprétation et les algorithmes d'apprentissage statistique qui en découlent.

Bien évidemment, nous nous limitons dans cette partie aux concepts utiles pour la suite, et dans ce but, nous nous sommes inspirés et nous référons les lecteurs pour plus de détails aux ouvrages de référence [PEARL, 1988 et KOLLER et FRIEDMAN, 2009].

1.1 Rappel sur les probabilités

Le terme "probabilité" que nous employons régulièrement dans le langage commun fait référence au degré de confiance que nous avons dans la possibilité d'un événement de se produire. Dans un cadre plus formel, sans rentrer dans un débat plus que centenaire [FREEDMAN, 1997], le terme probabilité correspond soit à une fréquence d'apparition d'un événement (approche dite *fréquentiste*) soit en effet à un degré de croyance quant à l'apparition de l'événement (approche dite *bayésienne*).

1.1.1 Probabilité

Soit Ω un espace de résultats possible. Un événement se caractérise par l'ensemble des résultats qui peuvent être obtenus lorsqu'il se réalise. Par exemple, si nous considérons un lancer d'un dé à 6 faces, nous définirons $\Omega = \{1, 2, 3, 4, 5, 6\}$. L'événement "le lancer du dé est pair" se définit alors par $A = \{2, 4, 6\}$. Un événement est donc un sous-ensemble de Ω.

Parmi les sous-ensembles de Ω, on suppose qu'il existe un ensemble d'événements S auxquels nous sommes prêts à attribuer des probabilités (mesurables). En particulier Ω appartient à S car il s'agit de l'événement certain. De même, nous tenons pour acquis que le complémentaire d'un événement mesurable est mesurable, ainsi que l'union et l'intersection d'événements mesurables. À partir de ces concepts, nous pouvons maintenant définir la notion de distribution de probabilité :

Définition 1.1.1

Une distribution de probabilité $\mathbb{P}$ sur (Ω, S) est une représentation des événements de S en valeurs réelles qui satisfait les conditions suivantes :

- $\forall A \in S, \mathbb{P}(A) \geqslant 0$ *: une probabilité est positive.*
- $\Omega \in S$ *et* $\mathbb{P}(\Omega) = 1$ *: la probabilité de tous les résultats possibles, l'événement certain, est égale à 1.*
- *Si* $A, B \in S$ *et* $A \cap B = \varnothing$ *alors* $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ *: la probabilité que l'un de deux événements disjoints se produise est la somme des probabilités des deux événements.*

Cette définition est suffisante pour induire un grand nombre de propriétés, par exemple : $\mathbb{P}(\varnothing) = 0$ et $\forall A \in S$, $0 \geqslant \mathbb{P}(A) \geqslant 1$.

1.1.2 Variable aléatoire

Quand S, et surtout Ω, devient de très grande taille, il est nécessaire de se doter d'un outil permettant d'indexer et d'organiser cet ensemble combinatoire d'événements. Ce sont les variables aléatoires.

Définition 1.1.2 (*Variable aléatoire*)

Étant donné $\mathbb{P}$ sur (Ω, S), on appelle variable aléatoire X, une variable qui peut prendre un certain nombre de valeurs et qui est telle que :

1. *On note le domaine $dom(X)$, l'ensemble des valeurs possibles de X.*
2. *La valeur de la variable est fonction des résultats de Ω.*
3. *Tout résultat correspond à une unique valeur de la variable X.*
4. *$\forall x \in dom(X)$, la proposition $(X = x)$ correspond à l'ensemble des résultats de Ω qui donne la valeur x à X.*
5. *$\forall x \in dom(X), (X = x) \in S$.*

Exemple 1.1.1

Par exemple, si on définit P une variable aléatoire vérifiant :

- $dom(P) = \{0, 1\}$
- $(P = 0) = \{2, 4, 6\}$
- $(P = 1) = \{1, 3, 5\}$

alors la variable P représente la parité d'un lancer de dé. On peut en conclure que pour un dé non pipé, $\mathbb{P}(P = 0) = \mathbb{P}(\{2, 4, 6\}) = \frac{1}{2}$.

Si $dom(X)$ est dénombrable ou fini, on dit alors que X est une variable aléatoire discrète, sinon X est une variable aléatoire continue. Il est toutefois communément admis qu'on appelle "variable discrète", une variable discrète finie et que l'on précisera lorsqu'elle est discrète et seulement dénombrable.

Exemple 1.1.2

Pour illustrer différentes définitions, nous utiliserons un exemple sur l'asthme [GONZALES, 2018]. Supposons que vous vous rendiez sur votre lieu de travail et que vous vous interrogiez sur la pertinence d'utiliser votre vélo. Dans les circonstances où la pollution reste modérée, ce mode de déplacement peut s'avérer plaisant. Cependant, lorsque le niveau de pollution est élevé, cela peut déclencher des crises d'asthme, une situation que vous cherchez à éviter. Seulement, vous n'avez pas toujours accès directement à l'information sur le niveau de pollution. Et si on décrit de manière plus générale le problème, on peut remarquer que le niveau de pollution dépend de la densité de la circulation. La circulation, elle, peut dépendre de l'heure, si c'est "l'heure de pointe" par exemple, mais aussi de s'il y a eu un accident, ce qui est d'ailleurs souvent lié à la météo. Les variables disponibles et leurs modalités sont présentes dans la table 1.1.

Variable	Valeurs
Météo	{ensoleillé, nuageux, pluvieux, orageux, neigeux}
Accident	{oui, non}
Heure	$\{0, 1, 2, \ldots, 23\}$
Circulation	{faible, normale, dense, exceptionnelle}
Pollution	$\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$
Asthme	{crise, sans crise}

Tab. 1.1. : Les variables discrètes de l'exemple sur l'asthme

La discrétisation est un processus qui permet de transformer une variable continue X en une variable discrète (finie) X_d par la définition d'une partition (finie) de $dom(X)$ et le $dom(X_d)$ correspond alors à cette partition de dom(X). Dans la partie 2.5, nous parlerons plus en profondeur de ce processus. Dans notre exemple 1.1.2, l'heure, variable continue sur $[0, 24[$ est discrétisée en $[\![0, 23]\!]$.

1.1.3 Probabilité marginale, jointe et conditionnelle

Soit A et B deux variables aléatoires discrètes.

Définition 1.1.3 (*Probabilité marginale*)

La probabilité marginale $\mathbb{P}(A)$ *représente l'ensemble des probabilités de* $\mathbb{P}(A = a)$ *pour tout* $a \in dom(A)$. *La probabilité marginale* $\mathbb{P}(A)$ *est donc représentée par un vecteur ou tenseur de rang 1 de taille* $\mid dom(A) \mid$.

De manière compacte, s'il n'y a pas d'ambiguïté, on notera $\mathbb{P}(a)$ pour $\mathbb{P}(A = a)$.

Définition 1.1.4 (*Probabilité jointe*)

La probabilité jointe $\mathbb{P}(A, B)$ *représente l'ensemble des probabilités* $\mathbb{P}((A = a) \cap (B = b))$ *pour tout* $a \in dom(A)$ *et* $b \in dom(B)$.

La probabilité jointe $\mathbb{P}(A, B)$ *est donc représentée par une matrice ou tenseur de rang 2 de taille* $\mid dom(A) \mid \times \mid dom(B) \mid$.

Nous simplifierons $\mathbb{P}(a, b)$ pour $\mathbb{P}(A = a, B = b) = \mathbb{P}((A = a) \cap (B = b))$.

Cette notion de probabilité jointe peut être généralisée à plus de deux variables : $\mathbb{P}(X_1, \cdots, X_d)$ représentée par un tenseur de rang d et de taille $\prod_{i=1}^{d} \mid dom(X_i) \mid$

Définition 1.1.5 (*Marginalisation*)

La marginalisation est l'opération qui consiste à diminuer la dimension (le rang) d'une probabilité jointe en éliminant des variables, c'est-à-dire, en sommant sur les différentes valeurs de cette variable

$$\mathbb{P}(X_1, X_2) = \sum_{x_3 \in dom(X_3)} \mathbb{P}(X_1, X_2, x_3)$$

Nous simplifierons $\sum_{X_3} \mathbb{P}(X_1, X_2, X_3)$ pour $\sum_{x_3 \in dom(X_3)} \mathbb{P}(X_1, X_2, x_3)$

En particulier, on retrouve la probabilité de la marginale $\mathbb{P}(A)$ en marginalisant la probabilité jointe $\mathbb{P}(A, B)$:

$$\mathbb{P}(A) = \sum_{B} \mathbb{P}(A, B)$$

Définition 1.1.6 (*Probabilité conditionnelle*)

La probabilité conditionnelle $\mathbb{P}(A \mid B)$ est définie pour $\mathbb{P}(B) > 0$ et représente l'ensemble des probabilités $\mathbb{P}((A = a) \mid (B = b))$ pour tout $a \in dom(A)$ et $b \in dom(B)$.

La probabilité conditionnelle $\mathbb{P}(A \mid B)$ est donc représentée par une matrice ou tenseur de rang 2 de taille $\mid dom(A) \mid \times \mid dom(B) \mid$.

On définit alors le tenseur $\mathbb{P}(A|B)$ par une division tensorielle entre le tenseur $\mathbb{P}(A, B)$ et le tenseur $\mathbb{P}(B)$:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)}$$

Une conséquence immédiate de la définition de la probabilité conditionnelle est le théorème de Bayes :

Définition 1.1.7 (*Théorème de Bayes*)

Sous contrainte de positivité de $\mathbb{P}(B)$:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)}$$

Une autre conséquence de ces définitions est la *chain rule* :

Définition 1.1.8 (*Chain Rule*)

Soit $\mathbb{P}$ une probabilité jointe sur un ensemble de variables $\{X_1, \ldots, X_d\}$, $\mathbb{P}$ peut alors être factorisée comme suit :

$$\mathbb{P}(X_1, X_2, ..., X_d) = \mathbb{P}(X_1)\mathbb{P}(X_2|X_1) \cdot \mathbb{P}(X_3|X_1, X_2) \cdot ... \cdot \mathbb{P}(X_d|X_1, X_2, ..., X_{d-1})$$

Il est important de noter que l'ordre d'énumération des variables impacte fortement les probabilités conditionnelles qui composent cette factorisation. Autrement dit, si la loi jointe est supérieur à 0, la *chain rule* propose $d!$ factorisations de la loi jointe : une par ordre d'énumération.

1.1.4 Indépendance marginale et conditionnelle

D'après la section précédente, une probabilité jointe sur un grand nombre de variables est donc représentée par un tenseur dont le rang correspond au nombre de ses variables. C'est une représentation précise, mais exponentielle, de l'information probabiliste. Pourtant, il devrait être possible de discerner des relations qualitatives dans cette représentation quantitative.

Exemple 1.1.3

1. *"il est nécessaire de connaître X_1 pour prédire X_2"*
2. *"si je connais la valeur de A alors B n'apporte aucune information sur C"*
3. *"U n'apporte jamais d'information sur Z"*

Ces propriétés qui sont encodées dans les paramètres de la loi jointe sont une information importante pour l'interprétation et pour la compréhension avec les experts. Il est donc important de les extraire et les mettre en valeur.

Ces relations qualitatives prennent essentiellement la forme d'indépendance marginale et conditionnelle qu'il s'agit donc maintenant de définir formellement.

Définition 1.1.9 (*Indépendance marginale*)

Une probabilité $\mathbb{P}$ satisfait $(X \perp\!\!\!\perp Y)$ si et seulement si

$$\mathbb{P}(X, Y) = \mathbb{P}(X) \cdot \mathbb{P}(Y)$$

L'indépendance marginale permet la découverte de connaissances importantes dans la probabilité jointe, comme le montre l'item 1 de l'exemple 1.1.3. Toutefois, elle ne permet pas de discerner des comportements plus fins, mais tout aussi importantes dans la compréhension du modèle tel que celui de l'item 2. Il est alors nécessaire d'introduire la notion d'indépendance conditionnelle.

Définition 1.1.10 (*Indépendance conditionnelle*)

Une variable aléatoire (v.a.) X est indépendante conditionnellement à une v.a. Y sachant la v.a. Z dans $\mathbb{P}$ dénotée $X \perp\!\!\!\perp Y|Z$ si et seulement si :

$$\mathbb{P}(X, Y \mid Z) = \mathbb{P}(X|Z) \cdot \mathbb{P}(Y|Z)$$

De manière équivalente, $X \perp\!\!\!\perp Y \mid Z \iff \mathbb{P}(X|Y, Z) = \mathbb{P}(X|Z)$. Cette propriété est intéressante car parfois plus facilement interprétable pour les experts : si Z est connu, Y ne rajoute aucune nouvelle information sur X.

Exemple 1.1.4

Voici des exemples de probabilité marginale, jointe et conditionnelle provenant des données de l'exemple 1.1.2 :

asthme	
crise	sans_crise
0.3131	0.6869

$P(asthme)$

	asthme	
pollution	crise	sans_crise
1	0.0000	0.0000
2	0.0075	0.0280
3	0.0835	0.3399
4	0.0787	0.2000
5	0.0700	0.0775
6	0.0329	0.0300
7	0.0104	0.0060
8	0.0180	0.0038
9	0.0070	0.0012
10	0.0050	0.0005

$P(asthme, pollution)$

	asthme	
pollution	crise	sans_crise
1	0.1000	0.9000
2	0.2113	0.7887
3	0.1972	0.8028
4	0.2824	0.7176
5	0.4746	0.5254
6	0.5230	0.4770
7	0.6341	0.3659
8	0.8257	0.1743
9	0.8537	0.1463
10	0.9091	0.0909

$P(asthme|pollution)$

Fig. 1.1. : De gauche à droite, la probabilité marginale de la variable asthme, la probabilité jointe des variables asthme et pollution et la probabilité conditionnelle de la variable asthme sachant la variable pollution.

Notre objectif est donc de représenter la distribution jointe $\mathbb{P}$ sur un ensemble de variables aléatoires $\mathbf{X} = \{X_1, \ldots, X_k\}$. Cependant, même dans un cas très simple, la représentation explicite des distributions jointes entraîne une charge computationnelle importante.

1.2 Modèle du Réseau Bayésien

Extraire de la connaissance qualitative d'une probabilité jointe revient principalement à identifier le plus grand nombre d'indépendances marginales et conditionnelles possibles. De même que la représentation quantitative est exponentielle, la liste des indépendances identifiées peut être de très grande taille et donc difficile à maîtriser dans le cadre d'une interaction avec un expert. Il serait pertinent de trouver une représentation plus compacte de ces indépendances. Les réseaux bayésiens proposent d'utiliser un graphe entre variables à cet effet.

1.2.1 Notions principales de théorie des graphes

Cette section s'attache à présenter les quelques concepts de graphes suffisant pour la suite de ce manuscrit.

Définition 1.2.1

*Soit N un ensemble fini, un **graphe** $\mathcal{G}$ sur N est la représentation d'une relation binaire entre les éléments de N. Si la relation est symétrique, on dit que le graphe est **non orienté** et si la relation est anti-symétrique, on dit que le graphe est **orienté**. Un graphe $\mathcal{G}(N, A)$ est donc caractérisé par son ensemble N et un ensemble $A \subset N \times N$.*

*Dans ce contexte, on appelle **nœud** un élément de N et on appelle **arc** (respectivement **arête**) un élément de A si $\mathcal{G}$ est orienté (respectivement non orienté).*

Avec cette définition de $\mathcal{G}$, nous confondons arcs bi-orientés et arêtes, ce qui nous suffira pour la suite de notre présentation. Dans ce même cadre, nous pouvons introduire la notion de graphe mixte : un graphe possédant des arêtes et des arcs.

- Pour un arc $x \to y$, on nomme x le **parent** de y et y l'**enfant** de x.
- Pour une arête $x - y$, on dit que x et y sont **voisins**.

Définition 1.2.2

Dans un graphe orienté,

- *un* ***chemin*** *de longueur l est une séquence d'arcs $(x_i \to x_{i+1})_{i\in\{0,\cdots,l-1\}} \in A^l$,*
- *un* ***circuit*** *de longueur l est un chemin tel que $x_0 = x_l$.*

Dans un graphe quelconque,

- *une* ***chaîne*** *de longueur l est un ensemble de nœuds $(x_0, \cdots, x_l)$ tel que $\forall i < l, x_i - x_{i+1} \in A$ ou $x_i \to x_{i+1} \in A$ ou $x_i \leftarrow x_{i+1} \in A$,*
- *un* ***cycle*** *est une chaîne telle que $x_0 = x_l$.*

Un ***descendant*** *de x est un nœud y tel qu'il existe une chaine de x à y.*
Un ***ascendant*** *de y est un nœud x tel qu'il existe une chaine de x à y.*

On nomme **DAG** ("*directed acyclic graph*"), un graphe orienté sans circuit (et non pas sans cycle, le terme anglais étant un faux-ami).
Enfin, on appelle **squelette** le graphe non orienté ayant la même structure que le graphe d'origine.

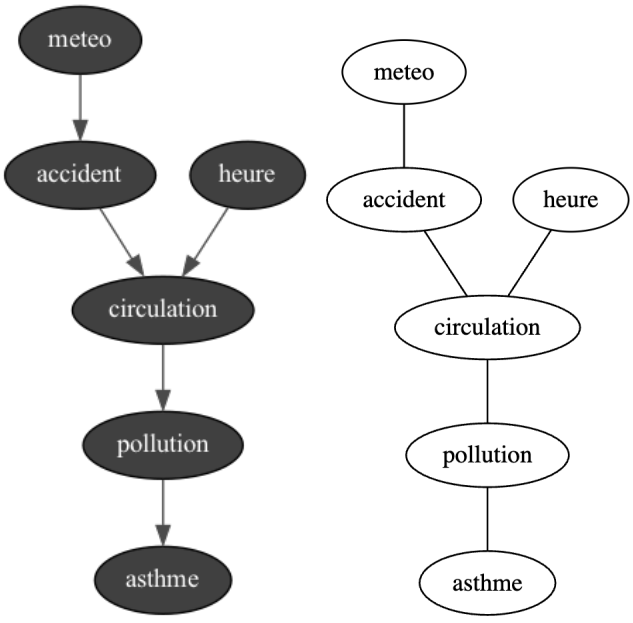


Fig. 1.2. : Le graphe orienté sur l'asthme à gauche et non orienté à droite de l'exemple 1.1.2.

Exemple 1.2.1

Dans le graphe orienté, $pollution \rightarrow asthme$ est un exemple d'arc où la variable pollution est le parent de la variable asthme. Dans le graphe non orienté, $pollution - asthme$ est un exemple d'arête où la variable pollution et la variable asthme sont voisins. Ici, le graphe orienté est un DAG et le graphe non orienté est son squelette.

1.2.2 Définition des réseaux bayésiens

Le rôle des graphes en modélisation probabiliste et statistique est de fournir des moyens pratiques pour exprimer des hypothèses structurelles sur le modèle, de faciliter la représentation efficace et l'interprétation des lois jointes et de permettre aussi d'accélérer le calcul d'inférence à partir d'observations [PEARL, 1988].

Définition 1.2.3 (*Réseau Bayésien*)

Un réseau bayésien $\mathcal{B} = (\mathbb{P}, \mathcal{G})$ utilise le DAG $\mathcal{G}(\boldsymbol{X}, \boldsymbol{A})$ pour la représentation compacte de $\mathbb{P}$, la distribution de probabilités jointe sur l'ensemble de variables $\boldsymbol{X} = \{X_1, \cdots, X_n\}$. Dans ce graphe, l'absence d'arc indique qu'une indépendance conditionnelle a été détectée :

$$\forall i \neq j, \text{ si } X_i \rightarrow X_j \notin \boldsymbol{A} \text{ alors } \exists Z, \varnothing \subseteq Z \subset \boldsymbol{X} \backslash \{X_i; X_j\} \text{ tel que } X_i \perp\!\!\!\perp X_j \mid Z$$

La structure de graphe permet donc de lire certaines indépendances dans la loi $\mathbb{P}$. Par exemple, la propriété de Markov locale (PML) indique que chaque variable est conditionnellement indépendante de ses non-descendants sachant ses parents. Ce qui conduit à la propriété fondamentale des réseaux bayésiens : la factorisation de la distribution jointe $\mathbb{P}$ des n variables du modèle.

Définition 1.2.4 (*Factorisation dans un réseau bayésien*)

Soit le réseau bayésien $\mathcal{B} = (\mathbb{P}, \mathcal{G})$,

$$\mathbb{P}(X_1, \cdots, X_n) = \prod_{i=1}^{n} \mathbb{P}\left(X_i | \textit{parents}(X_i)\right)$$

C'est une propriété fondamentale qui justifie la représentation compacte annoncée plus haut. Par exemple, dans notre exemple 1.1.2 sur l'asthme de 6 variables, chacune ayant 2 à 24 modalités, construire la distribution de probabilité jointe de toutes

ces variables, soit $\mathbb{P}(meteo, accident, heure, circulation, pollution, asthme)$, revient à une table de $5 \times 2 \times 24 \times 4 \times 10 \times 2$ soit 19 200 paramètres. Avec la factorisation, on peut représenter le modèle avec un ensemble de tableaux comprenant en tout $5 + 10 + 24 + 192 + 40 + 20$ soit 291 valeurs.

Cette factorisation n'est pas unique et n'est pas non plus parfaite. En effet, il est possible que certaines indépendances dans la loi qui permettraient d'être encore plus compacte, ne soient pas représentables dans le graphe. Par contre, la propriété de Markov globale indique que toutes les indépendances lisibles dans le graphe sont effectivement présentes dans la loi $\mathbb{P}$. Cela nous assure que cette décomposition, ni unique ni parfaite, est au moins exacte. Pour une définition plus précise de la propriété de Markov globale, nous référons le lecteur à la partie 1.3.

Il existe bien sûr un grand nombre d'extensions à ce modèle, notamment dans le cadre des variables continues. Il est vrai qu'expérimentalement, les bases de données de cas pratiques contiennent souvent des variables de type quantitatif qu'il est tentant de représenter par des variables aléatoires continues. Toutefois, ces modèles de réseaux bayésiens continus nécessitent généralement l'hypothèse de représentation paramétrique des distributions continues. Dans le cadre de cette thèse, il nous a semblé bien plus difficile de justifier du choix d'une hypothèse arbitraire du modèle plutôt que de se restreindre aux modèles discrets, en utilisant si nécessaire des techniques de discrétisation décrites dans la partie 2.5.

1.2.3 Inférence dans un réseau bayésien

Décrire la loi $\mathbb{P}$ par le réseau bayésien $\mathcal{B}$ de graphe $\mathcal{G}$ permet donc de représenter des distributions de très haute dimensionnalité, inaccessibles sans cet outil. Un des buts importants de pouvoir manipuler la loi $\mathbb{P}$ est le calcul d'impact probabiliste d'observation de certaines variables sur d'autres variables. Par exemple, comment se comporte la probabilité d'avoir une crise d'asthme si la circulation est faible. Ou encore, si le niveau de pollution est maximal, un accident est-il plus probable d'avoir eu lieu ? On nomme ces calculs d'impacts par le terme : inférence probabiliste.

Définition 1.2.5 (*Inférence Probabiliste*)

L'inférence consiste à calculer la distribution d'une variable d'intérêt (C) à partir d'observations sur certaines des autres variables ($\boldsymbol{A}$).

$$\mathbb{P}(C|A=\epsilon_a) = \mathbb{P}(C|\epsilon_a) = \frac{\mathbb{P}(\epsilon_a|C)\cdot\mathbb{P}(C)}{\mathbb{P}(\epsilon_a)} = \frac{\mathbb{P}(C,\epsilon_a)}{\mathbb{P}(\epsilon_a)} \propto \mathbb{P}(C,\epsilon_a)$$

Toutefois, dans le cadre des réseaux bayésiens, à cause de leur grande dimension impliquant une explosion combinatoire, cette équation fournit une procédure de calcul en un temps exponentiel [DAGUM et LUBY, 1993]. Il n'est pas nécessaire ici de se pencher précisément sur ces algorithmes, toutefois il est intéressant de noter que le graphe des réseaux bayésien servira aussi à obtenir des algorithmes de calculs exacts en des temps généralement beaucoup plus raisonnables.

Exemple 1.2.2

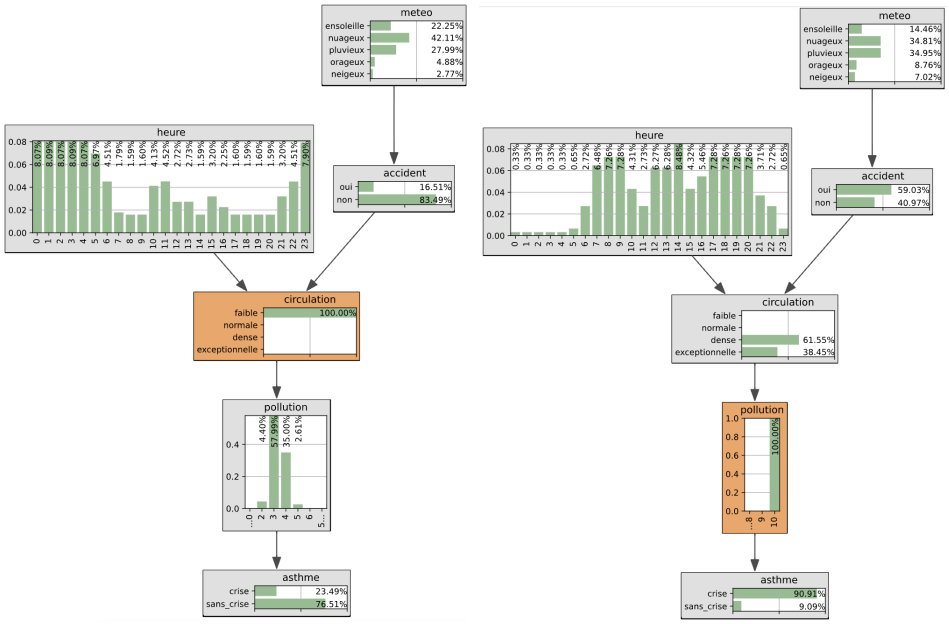


Fig. 1.3. : On peut lire $\mathbb{P}(asthme \mid circulation = faible)$ à gauche et $\mathbb{P}(accident \mid pollution = 9)$ à droite

Si on reprend les mêmes exemples, on observe qu'une crise d'asthme est peu probable (0.23) en cas de circulation faible, mais que si le niveau de pollution est maximal, la probabilité qu'un accident ait eu lieu est de 0.60.

1.3 Interprétation dans les réseaux bayésiens

Interpréter un modèle consiste à dégager et à exhiber des connaissances qualitatives qui permettent de comprendre son comportement en faisant abstraction de sa complexité. Dans le cadre des modèles probabilistes et notamment des réseaux bayésiens, ces connaissances prennent principalement la forme d'indépendances conditionnelles : quelles variables dépendent de quelles autres, et dans quels contextes. Il faut toutefois noter que la lecture de ces connaissances n'est pas directement une lecture du graphe, la preuve en est que le même réseau bayésien peut être représenté par différents DAG. Ainsi, le réseau bayésien à deux variables dépendantes A et B est aussi bien représenté par le graphe $A \rightarrow B$ que $B \rightarrow A$.

1.3.1 *d-séparation*

La lecture des connaissances s'effectue par l'intermédiaire d'une propriété spécifique aux réseaux bayésiens appelée la d-séparation. La d-séparation consiste à questionner la capacité d'une variable X à impacter une autre variable Y. Il s'agit d'identifier les chaines entre les deux variables qui sont aptes à transmettre de l'information, étant donné un contexte d'une certaine connaissance représentée par la valeur d'un ensemble de variables **Z**. Les chaines qui ne peuvent transmettre l'information sont nommées chaines bloquantes.

Définition 1.3.1 (*Chaîne bloquée*)

On dit qu'une chaîne $(X = U_0, U_1, \cdots, U_l, U_{l+1} = Y)$ *est communicante par* **Z** *si et seulement si :*

$$\forall i \in \{1, \cdots, l\}, \begin{cases} \text{si } U_{i-1} \rightarrow U_i \leftarrow U_{i+1} \text{ alors } U_i \text{ ou un de ses descendants } \in \mathbf{Z}, \\ \text{sinon } U_i \notin \mathbf{Z}. \end{cases}$$

Une chaîne qui n'est pas communicante par **Z** *est bloquée par* **Z**.

Exemple 1.3.1

Si l'on prend deux exemples classiques :

1. *La pointure P d'une personne et son niveau de lecture L sont dépendants, sauf si on connait l'âge A de la personne. La structure du modèle est donc $P \leftarrow A \rightarrow L$. Dans cette chaine, P et L sont bloqués par A.*
2. *Les résultats de deux lancers de dé (D_1, D_2) sont indépendants, sauf si on connait la somme S des deux dés. La structure du modèle est donc $D_1 \rightarrow S \leftarrow D_2$ où D_1 et D_2 sont communicants par S.*

Définition 1.3.2 (*d-séparation*)

Soit X et $Y \in \boldsymbol{X}$ et $\boldsymbol{Z} \subset \boldsymbol{X} \backslash \{X, Y\}$, on dit que X et Y sont d-séparées par $\boldsymbol{Z}$, si et seulement si toute chaîne de X à Y est bloquée par $\boldsymbol{Z}$.

La d-séparation est donc une notion purement graphique. C'est la propriété de Markov globale, précédemment citée en partie 1.2.2, qui la relie aux indépendances dans la loi $\mathbb{P}$.

Propriété 1.3.1 (*Propriété de Markov Globale*)

$\mathcal{B}(\mathbb{P}, \mathcal{G})$ est un réseau bayésien si et seulement si toute d-séparation dans $\mathcal{G}$ représente une indépendance conditionnelle dans $\mathbb{P}$.

On peut remarquer que d'après cette propriété, il peut exister des indépendances conditionnelles dans $\mathbb{P}$ qui ne sont pas représentés par une d-séparation. Il est donc pertinent de simplifier au maximum le graphe $\mathcal{G}$ afin de minimiser le nombre de ces indépendances mal représentées.

1.3.2 Couverture et frontière de Markov

Grâce à la d-séparation et la propriété de Markov globale, nous avons donc l'outil qui nous permet d'interpréter le graphe comme une source de connaissances qualitatives sur la loi $\mathbb{P}$ et sur lequel s'appuyer pour communiquer, confronter et valider le modèle avec l'expert. Par exemple, lorsqu'on s'intéresse localement à une variable, il peut être pertinent de déterminer les variables qui lui sont "proches", c'est-à-dire pour lesquels on peut s'attendre à un impact important.

Définition 1.3.3 (*Couverture de Markov*)

Une couverture de Markov d'une variable aléatoire $Y \in \boldsymbol{X}$ est un sous-ensemble $\boldsymbol{S}$ tel que :

$$Y \perp\!\!\!\perp \boldsymbol{X} \backslash \boldsymbol{S} \mid \boldsymbol{S}.$$

Autrement dit, **S** contient toutes les informations nécessaires pour l'inférence de Y : les variables de $\mathbf{X} \backslash \mathbf{S}$ sont superflues dans ce cas.

Définition 1.3.4 (*Frontière de Markov*)

Une frontière de Markov de Y est un élément minimal au sens de l'inclusion de l'ensemble des couvertures de Markov de Y.
*Plus particulièrement dans un réseau bayésien, la frontière de Markov d'un nœud précis contient ses parents, ses enfants et les autres parents de ses enfants [*PEARL*, 1988].*

Il est facile de généraliser et de proposer des niveaux de frontière de Markov : celle de niveau 1 est la minimale, celle de niveau 2 est la frontière de la frontière de Markov, jusqu'au niveau n qui inclue nécessairement toutes les variables du graphe. On structure alors le graphe en sous ensemble concentrique de moins en moins pertinents pour prédire Y. Cette information est très facile à décrire et à partager avec les experts.

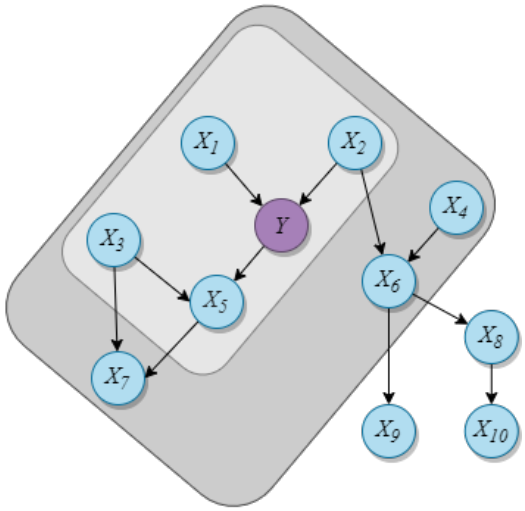


Fig. 1.4. : Un exemple de réseau bayésien. En gris clair la frontière de Markov de niveau 1 de Y et en gris foncé celle de niveau 2.

1.3.3 Causalité

L'idée même d'interprétabilité et d'explicabilité fait intervenir le concept de causalité. Pour fournir une explication d'un résultat, d'analyses ou de classifications, l'idéal serait de pouvoir en expliciter la cause. Toutefois, dans la cadre des sciences des données, la causalité a toujours été un sujet délicat [PEARSON, 1892]. Même s'il est connu depuis longtemps que "corrélation n'est pas causalité", les réseaux bayésiens, en ce qu'ils utilisent le concept intuitivement causal qu'est l'arc, rajoute encore à l'ambiguïté de ce concept. Dans un premier temps, il s'agira pour nous d'analyser la relation entre modèle causal et modèle graphique. Dans un second temps, nous définirons quelques concepts clés qui aideront à la compréhension des modèles appris dans nos contributions.

Comme nous l'avons déjà dit, il n'y a pas unicité du graphe représentant notre distribution $\mathbb{P}$. Plus précisément, on dira que deux graphes sont Markov-équivalents si on peut y lire (grâce à la d-séparation), les mêmes indépendances conditionnelles. On peut donc définir une classe d'équivalence de Markov comme un ensemble de réseaux bayésiens qui sont tous Markov-équivalents.

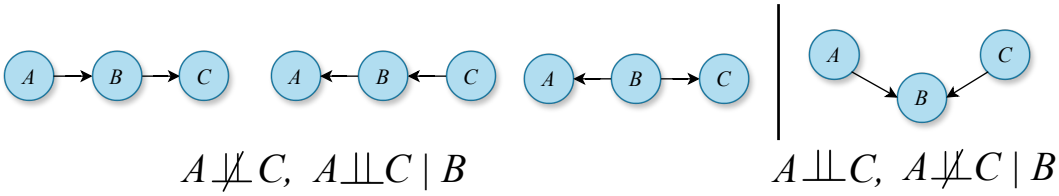


Fig. 1.5. : Deux classes d'équivalences de Markov.

Par exemple, dans cette figure 1.5, on présente deux classes d'équivalence différentes : une classe formée de 3 graphes différents et une classe n'en contenant qu'un. Ce dernier graphe représente des relations d'indépendances très caractéristiques, déjà rencontré dans la *d-séparation* : la v-structure, c'est-à-dire, un motif où deux nœuds non reliés entre eux possède un enfant commun. On peut d'ailleurs caractériser plus précisément les graphes qui font partie de la même classe d'équivalence : ils ont le même squelette et les mêmes v-structures [VERMA et PEARL, 1991]. Cette propriété nous permet donc de représenter un graphe caractéristique de la classe d'équivalence : le graphe essentiel.

Définition 1.3.5 (*Graphe essentiel*)

Le graphe essentiel est caractéristique d'une classe d'équivalence de Markov, c'est un graphe mixte, de même squelette que les graphes de la classe d'équivalence et qui ne possède un arc que lorsque cet arc est commun à tous les graphes de la classe d'équivalence.

Fig. 1.6. : Les graphes essentiels des exemples en figure 1.5.

Comme la classe d'équivalence de Markov est composé de réseaux bayésiens qui ont les mêmes indépendances probabilistes, une propriété surprenante est qu'ils sont indistinguables d'un niveau statistique, ce qui amène à deux conclusions : un algorithme d'apprentissage ne peut pas faire mieux qu'apprendre dans une classe d'équivalence et de ce fait, un apprentissage statistique ne peut pas prétendre obtenir un modèle causal.

Définition 1.3.6 (*Modèle causal*)

Un modèle causal est un DAG tel qu'un arc $X \to Y$ représente la relation causale : X est la cause de Y.

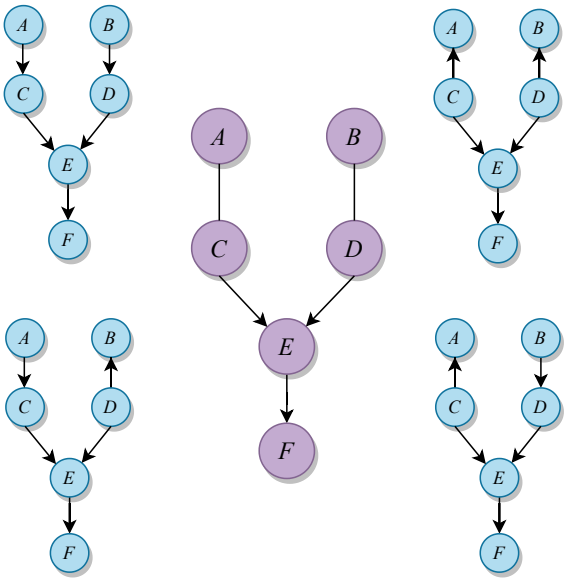


Fig. 1.7. : Une classe d'équivalence de Markov et son graphe essentiel au centre.

Il est communément admis qu'un des graphes de la classe d'équivalence de Markov est donc un modèle causal et que les autres ne le sont pas. Ainsi, même si le modèle causal n'est pas accessible par l'apprentissage, les arcs du graphe essentiel sont donc eux causaux.

Cette affirmation est un peu rapide, car il n'est pas possible de faire de modèle causal sans accepter l'idée de variables latentes qui introduisent des corrélations fallacieuses ("*spurious correlations*"), par exemple, entre deux conséquences d'une variable causale latente.

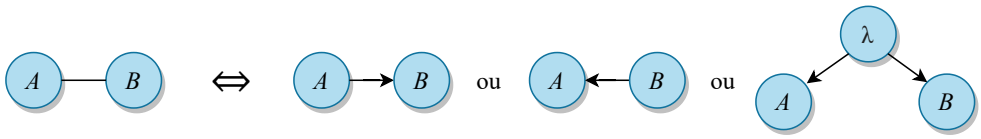


Fig. 1.8. : Différentes possibilités causales lorsque A et B sont corrélés. λ étant une variable non observée.

Même si le modèle causal n'est pas atteignable et que le modèle causal du graphe essentiel est une simplification, on peut toutefois remarquer qu'il existe des moyens de détecter des variables latentes (voir partie 1.4.3) et que la tâche de prédiction qui nous intéresse particulièrement ne nécessite pas nécessairement d'avoir un modèle causal. Par contre, l'induction causale (la recherche du modèle causal) est un sujet particulièrement intéressant, car c'est le moyen de passer d'un modèle de prédiction à un modèle de prescription, lui seul capable d'indiquer comment faire évoluer l'état de la variable d'intérêt.

1.4 Apprentissage

Un réseau bayésien est donc constitué d'une structure qui est un DAG et de paramètres qui sont les probabilités conditionnelles de chacun des nœuds sachant ses parents. Réussir à dériver ces connaissances à partir d'un échantillon iid suivant la loi jointe n'est pas trivial. Dans cette partie, nous présentons les méthodes classiques de cette tâche que l'on nomme communément apprentissage.

1.4.1 Apprentissage expert : modélisation

Dans de nombreux domaines, un processus automatique de construction du modèle n'est pas envisageable. Par exemple, isoler les variables pertinentes peut

être difficile, obtenir une base de données suffisante est délicat, voire impossible, etc. Plus généralement, il est fréquent d'avoir à sa disposition une base qui, soit en taille, soit en granularité, se révèle insuffisante pour l'apprentissage du modèle. Dans cette situation, il existe toutefois une possibilité pour que ces connaissances non dérivables d'une manière automatique puissent être acquises auprès d'experts dans le domaine d'application et de la littérature pertinente :

- **Choix de variables :** Elle est généralement basée sur des entretiens avec des experts, des descriptions du domaine et une analyse approfondie de l'objectif du réseau en cours de construction [LUCAS et al., 2004]. Cette étape est la plupart du temps nécessaire aussi dans un processus automatique. Même s'il est possible d'isoler les variables non pertinentes (voir partie 1.3.2), ou détecter l'existence de potentielles variables cachées (voir partie 1.4.3).

- **Estimation de la structure :** Les relations de dépendance et d'indépendance entre les variables doivent ensuite être analysées et exprimées dans une structure graphique. Pour ce faire, l'intuition de la causalité peut être utilisée comme principe directeur. Les relations obtenues sont alors exprimées en termes graphiques en prenant la direction de la causalité pour diriger les arcs entre les variables. Mais, il n'est pas toujours facile ou intuitif pour les experts d'exprimer les relations causalement. La connaissance des contraintes probabilistes qualitatives et des contraintes logiques entre les variables impliquées peut aider à la construction.

- **Estimation des paramètres :** Enfin, les distributions de probabilités conditionnelles locales pour chaque variable doivent être complétées. Ces probabilités peuvent difficilement être obtenues auprès d'experts du domaine. Il existe donc un grand nombre de méthodes qui permettent aux experts d'exprimer quantitativement leurs connaissances. Cela reste un processus extrêmement complexe et fastidieux. Toutefois, il faut noter que, la structure étant acquise, les probabilités nécessaires sont des probabilités locales dont l'estimation peut venir de diverses sources. Ainsi, connaitre les paramètres d'une loi jointe $\mathbb{P}(S, G)$ est bien plus complexe que d'estimer la loi *a priori* $\mathbb{P}(S)$ ainsi que le mécanisme probabiliste qui explique la valeur de G une fois la valeur de S connu : $\mathbb{P}(G \mid S)$. Par ailleurs, $\mathbb{P}(S)$ et $\mathbb{P}(G \mid S)$ n'ont pas à être obtenu par un même processus ou même par un même expert (voir exemple 1.4.1).

Exemple 1.4.1

Dans un service médical d'un hôpital, pour une maladie donnée, on cherche à comprendre l'impact du sexe S sur la guérison G d'une maladie. La cohorte contient 55 hommes et 30 femmes dont l'équipe médicale estime que 50 hommes vont guérir et seulement 15 femmes.

- *Les médecins se sont accordés pour dire que parmi les variables potentiellement explicatives retenues, seul le sexe impacte la guérison de la maladie.*
- *Les experts se sont naturellement fixés sur une structure causale $S \to G$.*
- *Il reste à estimer les paramètres $\mathbb{P}(S)$ et $\mathbb{P}(G \mid S)$. Les estimations de l'équipe médicale permettent de quantifier correctement $\mathbb{P}(G \mid S)$ puisqu'elle rend compte explicitement du mécanisme reliant le sexe de la personne et sa capacité à guérir. Par contre, il sera maladroit d'utiliser la proportion d'homme et de femmes dans cet hôpital pour estimer $\mathbb{P}(S)$. Une source extérieure comme l'INSEE pourrait permettre de fournir cette probabilité pour l'ensemble de la population française. En particulier, l'estimation de $\mathbb{P}(G)$ serait surestimée en utilisant cette probabilité locale de S puisqu'on dirait qu'en moyenne, on guérit beaucoup plus qu'en réalité.*

Dans les cas favorables, les données sont suffisantes pour pouvoir envisager des traitements statistiques. Des processus d'apprentissage automatique peuvent donc être envisagés pour la structure comme pour les paramètres.

1.4.2 Apprentissage des paramètres

Le problème de l'apprentissage de paramètres dans un réseau bayésien consiste à estimer les distributions de probabilités conditionnelles à structure fixée à partir d'une base de données.

La théorie classique des statistiques inférentielles propose comme meilleure estimation d'une probabilité, le calcul d'une fréquence dans la base de données. Par exemple, avec p_i la probabilité d'occurrence d'un certain événement i et N_i le nombre d'occurrences de i dans la base de taille N, on estime p_i par $\tilde{\mathbb{P}}_i = \frac{N_i}{N}$. Il suffit donc de compter N_i et N pour estimer p_i.

Les statistiques bayésiennes rajoutent la prise en compte d'un *a priori*. Pour peu que l'*a priori* soit bien choisi, cela se résume à adjoindre une base virtuelle de taille M représentant cette connaissance experte et à rajouter au comptage N_i des pseudo-comptages α_i dans cette base virtuelle. Dans le cas des probabilités discrètes,

l'*a priori* conjugué de la distribution multinomiale est la distribution de Dirichlet [SAPORTA, 2006].

Dans le cas des réseaux bayésiens, les probabilités à estimer sont plus complexes, car elles prennent la forme de probabilités conditionnelles : la probabilité qu'une variable X_i prenne la valeur k sachant que les parents pa_i de X_i prennent la valeur j. Les comptages sont donc triplement indicés, classiquement représentés par : $N_{i,j,k}$. De même pour l'a-priori : les pseudo-comptages dans la base virtuelle seront représentées par $\alpha_{i,j,k}$. On définit aussi $r_i = \mid dom(X_i) \mid$.

On peut donc résumer l'apprentissage des paramètres dans un réseau bayésien par le tableau suivant :

a priori	Nom	Formule
Sans	Maximum de vraisemblance	$\tilde{\mathbb{P}}_{\text{MLE}}(X_i = k \mid pa_i = j) = \frac{N_{i,j,k}}{N_{i,j}}$
Avec	Maximum a Posteriori	$\tilde{\mathbb{P}}_{\text{MAP}}(X_i = k \mid pa_i = j) = \frac{\alpha_{i,j,k}+N_{i,j,k}-1}{\alpha_{i,j}+N_{i,j}-r_i}$
Avec	Estimation a Posteriori	$\tilde{\mathbb{P}}_{\text{EAP}}(X_i = k \mid pa_i = j) = \frac{\alpha_{i,j,k}+N_{i,j,k}}{\alpha_{i,j}+N_{i,j}}$

Tab. 1.2. : Les types d'apprentissage de paramètres dans un cadre bayésien

Quelques remarques sur ce tableau :

- quand $N_{i,j,k} \to 0$, l'*a priori* est très important et l'apprentissage devient donc expert ;
- quand $N_{i,j,k} \to \infty$, $\alpha_{i,j,k}$ n'a plus d'utilité, les estimations sont consistantes et équivalentes ;
- il existe un cas particulier quand $\alpha_{i,j,k} = 1$ où on considère un *a priori* uniforme partout.

Des méthodes ont en effet été proposées dans le cas où peu de données rendaient l'estimation des paramètres fragiles sans *a priori* explicite. Par exemple, un ajustement des paramètres pour éviter les 0 :

Définition 1.4.1 (*Ajustement de Laplace (smoothing)*)

$$\tilde{\mathbb{P}} \approx \frac{N_{i,j,k} + 1}{N_{i,j} + r_i}$$

Cela revient au cas précédent avec $\alpha_{i,j,k} = 1$, il n'y a pas de connaissance experte, on a donc un *a priori* uniforme et une influence faible.

Un problème souvent rencontré est la présence dans la base de valeurs manquantes. Ces valeurs manquantes faussent les estimations des probabilités par comptage et sont donc un problème pour les algorithmes d'apprentissage. Une solution souvent proposée est l'utilisation de l'algorithme EM, pour Espérance - Maximisation [DEMPSTER et al., 1977] qui consiste à itérer sur des estimations des probabilités des valeurs manquantes, à partir d'une première estimation (voir algorithme 1.4.1). À chaque itération, cette estimation se précise. Cet algorithme a de bonnes propriétés, comme l'assurance d'une convergence, mais souffre de plusieurs défauts comme une lenteur de convergence, la convergence vers un optimum local et un manque de robustesse, car la solution trouvée dépend fortement de l'estimation initiale.

Algorithme 1.4.1 EM (Espérance - Maximisation)

Répéter jusqu'à convergence :

0. On choisit $\mathbb{P}^0(X_i \mid pa_i)$
1. Étape E : Estimer $N_{ijk}^{(t+1)}$ à partir des $\mathbb{P}^t(X_i \mid pa_i)$
2. Étape M : $\mathbb{P}^{t+1}(X_i = k \mid parents_i = j) = \frac{N_{ijk}^{(t+1)}}{N_{ij}^{(t+1)}}$

Ces contraintes sont très fortes et empêchent l'utilisation de cet algorithme dans de nombreux cas pratiques. Il existe alors d'autres méthodes plus simples, mais moins fondées mathématiquement, pour prendre en compte ces valeurs manquantes (voir partie 2.4).

1.4.3 Apprentissage de la structure

Nous nous plaçons à présent dans le cas où la structure du réseau bayésien n'est pas connue et nous intéressons aux méthodes permettant sa reconstruction à partir des données.

L'apprentissage de la structure d'un réseau bayésien est un problème NP-difficile [1] [CHICKERING, 2000], en premier lieu parce que le nombre de structures possible (NS) en fonction du nombre de variables n est super-exponentiel [ROBINSON, 1977] :

1. En théorie de la complexité, les problèmes NP-difficiles sont des problèmes pour lesquels aucune solution efficace (en temps polynomial) n'est connue [COOK, 1971]. Cela signifie qu'il n'existe pas d'algorithme qui puisse résoudre ces problèmes en un temps raisonnable à mesure que la taille des données augmente.

$$NS(n) = \begin{cases} 1 & , n \leqslant 1 \\ \sum_{i=1}^{n} (-1)^{i+1} \cdot C_i^n \cdot 2^{i \cdot (n-i)} \cdot NS(n-1) & , n > 1 \end{cases}$$

Ce qui implique qu'aucun algorithme d'exploration ne peut être efficace dans un tel espace d'état.

Il s'agit donc de proposer des algorithmes qui ne nécessiteront pas une phase d'exploration explicite. Pour ce faire, il existe deux méthodes principales : les algorithmes basés sur un score et les algorithmes basés sur l'indépendance (ou les contraintes).

Algorithmes à base de scores

Ces algorithmes se basent sur la proposition d'une fonction (le score) qui évalue la qualité de la structure d'un réseau bayésien en fonction de deux critères : (1) la vraisemblance de la structure par rapport à la base, et (2) un critère plus difficile à cerner par sa subjectivité, qui est censé représenter la préférence pour les modèles simples : le rasoir d'Occam. Ces deux critères sont antagonistes. En effet, le modèle le plus simple correspond à un ensemble de variables complètement indépendantes, ce qui est peu vraisemblable, et réciproquement, augmenter la vraisemblance tend à rendre tout dépendant de tout, ce qui donne le modèle le plus complexe possible. Cependant, l'existence d'un tel score permet de proposer des algorithmes de type glouton : étant donné une structure initiale et une notion de voisinage, l'algorithme explore l'espace des réseaux bayésiens de voisins en voisins, en maximisant l'augmentation du score à chaque itération. L'algorithme se termine évidemment lorsque tout le voisinage a un score moindre que la structure courante. Cette famille d'algorithmes se caractérise par le choix de la notion de voisinage et le choix de la notion du score. Le voisinage le plus communément admis correspond à l'application d'une opération élémentaire d'ajout, de retrait ou de modification d'un arc. Ce voisinage implique des contraintes supplémentaires sur le score pour être efficace : la décomposition locale.

Définition 1.4.2 (*Décomposition locale du score*)

Soit G une structure courante et G' une structure de son voisinage, alors il existe un arc $x \leftarrow y$ qui a été ajouté, retiré ou modifié et soit f un score sur les structures. f est décomposable localement si $f(G') - f(G)$ ne dépend que de x et y.

Cette définition nous assure un gain de score efficace dans tout le voisinage. Un grand nombre de scores peuvent être décrits et vérifient ces propriétés nécessaires d'un score. Nous allons uniquement en décrire deux :

1. le critère BIC (*Bayesian Information Criterion*)

$$\text{Score}_{\text{BIC}}(T, D) = log_2 L(\Theta^{\text{MV}}, T : D) - \frac{1}{2} \cdot Dim(T) \cdot log_2 N$$

   $\dim(T)$ est le nombre de paramètres du réseau bayésien. Ce critère existait bien avant les réseaux bayésiens dans le domaine de la sélection de modèles [SCHWARZ, 1978]. Sa formulation est très lisible, car on voit apparaitre sa vraisemblance et la $\dim(T)$ comme critère de complexité.

2. le score BDeu (*Bayesian Dirichlet score Equivalent*)

$$\text{Score}_{\text{BDe}}(T, D) = P(T) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{i,j})}{\Gamma(N_{i,j} + \alpha_{i,j})} \prod_{k=1}^{r_i} \frac{\Gamma(N_{i,j,k} + \alpha_{i,j,k})}{\Gamma(\alpha_{i,j,k})}$$

   Ce score bayésien se base sur l'expression de la vraisemblance du modèle, un *a priori* de Dirichlet pour les paramètres et un *a priori* sur la structure $P(T)$. Cet *a priori* permet d'exprimer une préférence pour certaines structures : on implémentera le rasoir d'Occam en exprimant que la probabilité *a priori* d'une structure sera d'autant plus grande que la structure sera simple. Mais, on pourra aussi intégrer d'autres préférences : par exemple, privilégier les structures connexes.

Comme tout algorithme glouton, les méthodes à base de score permettent d'obtenir un optimum local. Dès lors, il y a tout lieu d'utiliser l'ensemble des méta-heuristiques proposés dans la littérature afin d'optimiser le score de la structure finale : Random Restart [HOOS et STÜTZLE, 2005], recuit simulé [KIRKPATRICK et al., 1983], Tabu List [GLOVER et LAGUNA, 1998], algorithmes génétiques [GOLDBERG, 1989], etc.

Ces algorithmes ont longtemps été considérés comme les plus efficaces et les plus rapides, toutefois leur caractère pragmatique (l'ajout d'un arc ne correspond qu'à une augmentation d'un score) les rend peu séduisants dans le cadre d'une utilisation des réseaux bayésiens où l'interprétabilité serait un critère important. Depuis quelques années, avec l'avènement de l'XAI (*EXplainable Artificial Intelligence*), ce point de vue devient plus prégnant et contribue à renouveler l'intérêt pour la deuxième classe d'algorithmes d'apprentissage de structure.

Algorithmes à base de contraintes

Cette famille d'algorithmes part du principe que la structure d'un réseau bayésien encode une information sur les indépendances conditionnelles vérifiées par la loi jointe. Plus précisément, l'absence d'arc entre X et Y constitue la manifestation de l'existence d'une indépendance (potentiellement conditionnelle) entre les variables X et Y. Ces algorithmes se basent donc sur des tests d'indépendance conditionnelle dans la base de données, comme le test du χ^2, test du G2, etc.[SAPORTA, 2006], afin de construire peu à peu la structure vérifiant toutes ces contraintes d'indépendances révélées dans la base. Ces algorithmes doivent prendre en compte trois écueils principaux : le nombre exponentiel de tests d'indépendances, le taux d'erreur des tests statistiques d'indépendances (souvent fixé à 5%) et enfin le manque de robustesse de ces tests d'indépendances quand le nombre de données diminue. Le troisième point n'est pas à négliger, car, pour des tests d'indépendances conditionnelles, la taille de la partie de la base adressée par un test peut devenir très petite. Ces deux derniers écueils ne sont pas facilement évitables et les algorithmes de cette famille se distingueront donc principalement par la stratégie mise en œuvre pour minimiser le nombre de tests d'indépendances effectivement calculés.

Ainsi, l'algorithme PC, nommé d'après Peter Spirtes et Clark Glymour [SPIRTES et GLYMOUR, 1991], lui-même un raffinement de l'algorithme IC [VERMA et PEARL, 1991], propose une stratégie qui consiste, à partir d'un squelette complet, à trier les tests d'indépendances conditionnelles sur la taille de leur ensemble conditionnant. Les indépendances déjà trouvées se répercutent par la suppression d'arcs dans le graphe courant, ce qui réduit d'autant le nombre d'indépendances de plus haute complexité nécessaire par la suite. Dans un second temps, les indépendances calculées permettent de découvrir des v-structures et donc d'obtenir le graphe essentiel. Comme nous l'avons dit plus haut, le travail statistique d'apprentissage de structure se termine ici. Une troisième phase finalise l'orientation de manière arbitraire afin d'obtenir un réseau bayésien dans la même classe d'équivalence (voir algorithme 1.4.2).

Algorithme 1.4.2 PC (Peter-Clark)

1. Création du graphe non orienté reliant tous les nœuds.
2. Test et suppression de toutes les indépendances conditionnelles par des χ^2.
3. Recherche des v-structures.
4. Propagation des contraintes d'orientations.
5. Orientations des dernières arêtes en restant Markov-équivalent.

À ce jour, la stratégie de PC semble être la plus efficace pour organiser les tests d'indépendances. D'autres algorithmes se sont donc attaqués aux deux autres problèmes : le taux d'erreur et le manque de robustesse des tests d'indépendances. Par exemple, l'algorithme MIIC (*Multivariate Information-based Inductive Causation* de [AFFELDT et ISAMBERT, 2015 ; AFFELDT, VERNY et al., 2016] propose de remplacer ces critères d'indépendances statistiques par des concepts issus de la théorie de l'information : l'information mutuelle (à trois points), que l'on définira juste en dessous, permet de tester l'indépendance conditionnelle. L'intérêt de cette proposition est (1) d'obtenir un test plus robuste, (2) de fournir une information plus quantitative sur l'indépendance qui permet de trier l'ensemble des indépendances candidates et donc d'améliorer aussi la stratégie de minimisation de nombre de tests.

Pour définir l'information mutuelle à trois points, nous allons d'abord définir simplement l'entropie, la quantité d'information contenue dans une seule variable. Nous nous basons sur [KHINCHIN, 1957].

Définition 1.4.3 (*Entropie*)

Soit X, une variable discrète aléatoire, l'entropie $H(X)$ est définie par :

$$H(X) = -\sum_{x \in \boldsymbol{X}} \mathbb{P}(x) \log \mathbb{P}(x)$$

On peut étendre cette définition à l'entropie d'une paire de variables X et Y, qu'on appelle entropie jointe.

Définition 1.4.4 (*Entropie jointe*)

$$H(X,Y) = -\sum_{x \in \boldsymbol{X}} \sum_{y \in \boldsymbol{Y}} \mathbb{P}(x,y) \log \mathbb{P}(x,y)$$

Si on en revient à l'information mutuelle, elle mesure la quantité d'information partagée entre deux variables aléatoires.

Définition 1.4.5 (*Information mutuelle*)

$$I(X;Y) = \sum_{x \in \boldsymbol{X}} \sum_{y \in \boldsymbol{Y}} \mathbb{P}(x,y) \log \frac{\mathbb{P}(x,y)}{\mathbb{P}(x)\,\mathbb{P}(y)} = H(X) + H(Y) - H(X,Y)$$

Dans le cas de trois variables, l'information mutuelle à trois points, notée $I(X;Y;Z)$, quantifie la quantité d'information commune à toutes les paires possibles parmi les

variables X, Y et Z. À partir des définitions précédentes, on peut la définir par :

Définition 1.4.6 (*Information mutuelle à trois points*)

$$I(X;Y;Z) = H(X) + H(Y) + H(Z) - H(X,Y) - H(X,Z) - H(Y,Z) + H(X,Y,Z)$$

À partir du calcul de cette information mutuelle et si elle est négative, l'algorithme MIIC repère alors les indépendances. Il suit ensuite le modèle de l'algorithme PC. L'algorithme MIIC permet aussi de détecter les variables latentes [VERNY et al., 2017]. En effet, lorsqu'il existe un conflit d'orientation dans les arcs, cela peut être dû à l'existence d'une variable latente non observée.

Ce chapitre s'est appliqué à définir le modèle du réseau bayésien et introduire les concepts importants pour son interprétation et surtout son apprentissage. Nous nous intéresserons maintenant à la tâche de classification.

Références

AFFELDT, Séverine et Hervé ISAMBERT (2015). "Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information". In : *Proceedings of the UAI 2015 Conference on Advances in Causal Inference - Volume 1504*. ACI'15. Aachen, DEU : CEUR-WS.org, p. 1-29 (cf. p. 35).

AFFELDT, Séverine, Louis VERNY et Hervé ISAMBERT (2016). "3off2 : A network reconstruction algorithm based on 2-point and 3-point information statistics". In : *BMC Bioinformatics* 17.2, S12 (cf. p. 35).

BILLINGSLEY, Patrick (1995). *Probability and Measure*. Google-Books-ID : z39jQgAACAAJ. Wiley. 608 p. (cf. p. 10).

CHICKERING, David (2000). "Learning Bayesian Networks is NP-Complete". In : *Networks* 112 (cf. p. 31).

COOK, Stephen A (1971). "The Complexity of Theorem-Proving Procedures". In : (cf. p. 31).

DAGUM, Paul et Michael LUBY (1993). "Approximating probabilistic inference in Bayesian belief networks is NP-hard". In : *Artificial Intelligence* 60.1, p. 141-153 (cf. p. 21).

DEMPSTER, Arthur P., Nan M. LAIRD et Donald B. RUBIN (1977). "Maximum Likelihood from Incomplete Data Via the EM Algorithm". In : *Journal of the Royal Statistical Society : Series B (Methodological)* 39.1, p. 1-22 (cf. p. 31).

FREEDMAN, David (1997). "Some Issues in the Foundation of Statistics". In : *Topics in the Foundation of Statistics*. Sous la dir. de Bas C. van FRAASSEN. Dordrecht : Springer Netherlands, p. 19-39 (cf. p. 10).

GLOVER, Fred et Manuel LAGUNA (1998). “Tabu Search”. In : *Handbook of Combinatorial Optimization : Volume1–3*. Sous la dir. de Ding-Zhu DU et Panos M. PARDALOS. Boston, MA : Springer US, p. 2093-2229 (cf. p. 33).

GOLDBERG, David Edward (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Google-Books-ID : 2IIJAAAACAAJ. Addison-Wesley. 436 p. (cf. p. 33).

GONZALES, Christophe (2018). “Les réseaux bayésiens”. In : *Interstices*. Publisher : INRIA (cf. p. 12).

HOOS, Holger H. et Thomas STÜTZLE (2005). “3 - GENERALISED LOCAL SEARCH MACHINES”. In : *Stochastic Local Search*. Sous la dir. d’Holger H. HOOS et Thomas STÜTZLE. The Morgan Kaufmann Series in Artificial Intelligence. San Francisco : Morgan Kaufmann, p. 113-147 (cf. p. 33).

KHINCHIN, Alexandre (1957). *Mathematical Foundations Of Information Theory* (cf. p. 35).

KIRKPATRICK, Scott, Daniel GELATT et Manuela P. VECCHI (1983). “Optimization by Simulated Annealing”. In : *Science* 220.4598. Publisher : American Association for the Advancement of Science, p. 671-680 (cf. p. 33).

KOLLER, Daphne et Nir FRIEDMAN (2009). *Probabilistic graphical models : principles and techniques*. Adaptive computation and machine learning. Cambridge, MA : MIT Press. 1231 p. (cf. p. 10).

LUCAS, Peter J. F., Linda C. VAN DER GAAG et Ameen ABU-HANNA (2004). “Bayesian networks in biomedicine and health-care”. In : *Artificial Intelligence in Medicine*. Bayesian Networks in Biomedicine and Health-Care 30.3, p. 201-214 (cf. p. 28).

PEARL, Judea (1988). *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc. 552 p. (cf. p. 10, 19, 24).

PEARSON, Karl (1892). *The Grammar of Science*. Cambridge Library Collection - Physical Sciences. Cambridge : Cambridge University Press (cf. p. 25).

ROBINSON, Randall W. (1977). “Counting unlabeled acyclic digraphs”. In : *Combinatorial Mathematics V*. Sous la dir. de Charles H. C. LITTLE. Lecture Notes in Mathematics. Berlin, Heidelberg : Springer, p. 28-43 (cf. p. 31).

SAPORTA, Gilbert (2006). *Probabilités, analyse des données et statistique*. Google-Books-ID : rprNjztQYPAC. Editions TECHNIP. 664 p. (cf. p. 30, 34).

SCHWARZ, Gideon (1978). “Estimating the Dimension of a Model”. In : *The Annals of Statistics* 6.2. Publisher : Institute of Mathematical Statistics, p. 461-464 (cf. p. 33).

SPIRTES, Peter et Clark GLYMOUR (1991). “An Algorithm for Fast Recovery of Sparse Causal Graphs”. In : *Social Science Computer Review* 9.1. Publisher : SAGE Publications Inc, p. 62-72 (cf. p. 34).

VERMA, Thomas et Judea PEARL (1991). *Equivalence and Synthesis of Causal Models*. Google-Books-ID : ikuuHAAACAAJ. UCLA Computer Science Department. 9 p. (cf. p. 25, 34).

VERNY, Louis, Nadir SELLA, Séverine AFFELDT, Param Priya SINGH et Hervé ISAMBERT (2017). "Learning causal networks with latent variables from multivariate information in genomic data". In : *PLOS Computational Biology* 13.10, e1005662 (cf. p. 36).

Classification

2

Les méthodes de *machine learning* permettent d'apprendre des modèles et de prendre des décisions complexes à partir de jeux de données. La classification est une tâche d'apprentissage particulière dont l'objectif est de prédire les étiquettes de classe catégorielle de nouvelles instances, sur la base d'observations. Ainsi, on prédit la classe Y par une fonction des observations $\mathbf{X}$, soit :

$$\hat{y} = f(\mathbf{X}) \tag{2.1}$$

On distingue deux méthodes de classification : supervisée ou non supervisée. En classification non supervisée, le jeu de données ne contient pas d'exemple annoté. Le but est alors de détecter des similarités dans les données et de construire des groupes ou *clusters*. En classification supervisée, le jeu de données est composé d'exemples annotés par la classe (la cible). L'objectif principal est alors d'apprendre une correspondance entre les caractéristiques d'entrée et un ensemble prédéfini de classes. Les classes représentent différentes catégories ou groupes auxquels les instances peuvent appartenir. Si la cible n'a que deux valeurs possibles (généralement 0 et 1), on parle de classification binaire. En présence de multiples classes, on parle alors d'apprentissage multi-classe.

Dans ce chapitre, nous nous concentrons sur la classification supervisée. Nous décrirons d'abord la classification probabiliste pour ainsi introduire les réseaux bayésiens en tant que classifieur, puis nous parlerons de quelques autres méthodes de classification populaire. Nous aborderons ensuite les problématiques générales de la création des classifieurs et de leur évaluation, et les méthodes actuelles pour y remédier.

2.1 Classification probabiliste

Un classifieur probabiliste est un classifieur qui permet de prédire, sachant des observations d'une instance, une distribution de probabilité sur un ensemble de classes, plutôt qu'uniquement sélectionner la classe préférée.

Plus formellement, à partir de la base des variables $\mathbf{X} = (X_1, \ldots, X_n)$ et la classe Y, la classification probabiliste estime une distribution de probabilité. La fonction $f(\mathbf{X})$ qui précédemment (Équation 2.1) retournait la classe préférée, prend alors la forme de la recherche du maximum de cette distribution sur les valeurs de Y. Soit :

$$\hat{y} = \arg\max_{Y} \mathbb{P}(.) \tag{2.2}$$

On rappelle que la fonction de probabilité $\mathbb{P}$ est une estimation empirique à partir de la base de données. Ce principe général se spécialise en deux méthodes qui se distinguent par le type de $\mathbb{P}$.

2.1.1 Maximum de vraisemblance et maximum *a posteriori*

Pour mener à bien la tâche de classification qui consiste à trouver Y en fonction de $x_1, \ldots, x_n$, deux fonctions de probabilités sont pertinentes : la vraisemblance $\mathbb{P}(x_1, \ldots, x_n | Y)$ et la loi *a posteriori* $\mathbb{P}(Y | x_1, \ldots, x_n)$ [BISHOP, 2006].

De nombreuses méthodes consistent à proposer des modèles de vraisemblance afin de minimiser le nombre de paramètres à partir de la base (voir les prochaines sections). Par exemple, une mixture de gaussienne consistera à estimer la moyenne et la variance dans chaque classe comme sur la figure 2.1.

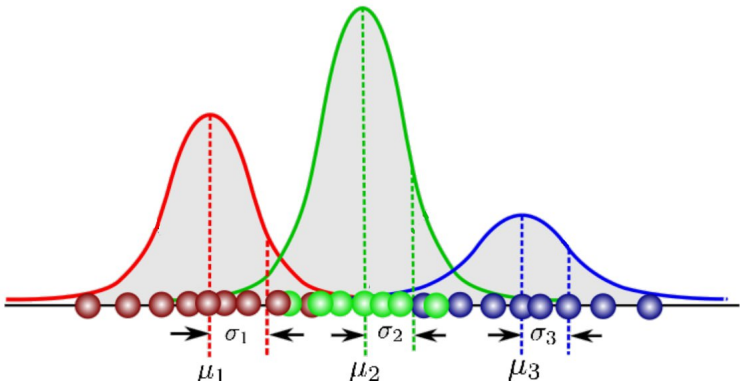


Fig. 2.1. : Exemple d'une mixture de gaussienne. Les points sur l'axe permettent de calculer les paramètres (μ_i, σ_i) et donc d'estimer les vraisemblances, par exemple : $\mathbb{P}(x \mid Rouge) \sim \mathcal{N}(\mu_1, \sigma_1^2)$

.

Par contre, la loi *a posteriori* peut être très difficile à obtenir, car sa connaissance demande de connaître la distribution jointe de toutes les variables, soit la vraisemblance et la distribution *a priori* de la classe.

$$\mathbb{P}(Y \mid x) \propto \mathbb{P}(x, Y) = \mathbb{P}(x \mid Y) \cdot \mathbb{P}(Y)$$

La classification probabiliste (Équation 2.2) peut alors se faire par maximum de vraisemblance (ML) :

$$\hat{Y}_{ML} = \arg\max_{Y} \mathbb{P}(x_1, \ldots, x_n | Y)$$

ou par maximum *a posteriori* (MAP) :

$$\hat{Y}_{MAP} = \arg\max_{Y} \mathbb{P}\left(Y|x_1, \ldots, x_n\right)$$

La connaissance de $\mathbb{P}(Y)$ est en effet exigeante, mais a contrario, si elle est accessible, il faut bien évidemment utiliser le MAP qui compare des probabilités d'une même distribution, contrairement au ML qui compare des probabilités de distributions différentes.

Dans le cas d'une classification binaire, Y ne peut prendre que deux valeurs (0 ou 1) et il est fréquent de représenter la fonction à estimer non pas par $\hat{Y}_{MAP}$ mais par la probabilité de la classe 1 :

$$f(\mathbf{x}) = \mathbb{P}\left(Y = 1|x_1, \ldots, x_n\right)$$

2.1.2 *Naive Bayes*

Comme dit précédemment, un classifieur probabiliste va proposer des hypothèses qui lui permettront de rendre calculables la vraisemblance ou les distributions *a posteriori*, c'est-à-dire de minimiser le nombre de paramètres nécessaires pour calculer cette probabilité. Généralement, c'est la vraisemblance qui est visée par cette simplification. Le classifieur *naive Bayes* est dans ce cadre le plus simple des classifieurs (non triviaux). Il consiste à faire l'hypothèse "naïve" que toutes les variables sont indépendantes entre elles sachant la classe : $\forall k \neq i$, $X_k \perp\!\!\!\perp X_i|Y$. Ce qui permet de réécrire la vraisemblance et l'*a posteriori* :

$$\mathbb{P}(X_1, \ldots, X_n \mid Y) = \prod_{i=1}^{n} \mathbb{P}(X_i \mid Y)$$

$$\mathbb{P}(Y \mid X_1, \ldots, X_n) \propto \mathbb{P}(Y, X_1, \ldots, X_n) = \mathbb{P}(Y) \cdot \prod_{i=1}^{n} \mathbb{P}(X_i \mid Y)$$

Ce qui permet de réécrire la tâche de classification par :

$$\hat{y}_{ML} = \arg\max_{y} \left(\prod_{i=1}^{n} \mathbb{P}(x_i|y) \right)$$

$$\hat{y}_{MAP} = \arg\max_{y} \left(\mathbb{P}(y) \cdot \prod_{i=1}^{n} \mathbb{P}(x_i|y) \right)$$

Le modèle *naive Bayes* demande un nombre linéaire de paramètres en fonction de la dimension plutôt qu'un nombre exponentiel. Toutefois, l'hypothèse d'indépendance est extrêmement forte et est rarement vérifiée en réalité. Cependant, dans le cadre d'une classification binaire, l'estimation obtenue est souvent suffisante pour prendre une décision pertinente [HAND et YU, 2001 ; ZHANG, 2004].

Par ailleurs, on peut remarquer que la factorisation *naive Bayes* correspond à celle d'un réseau bayésien dont la structure graphique est fixe et aisément identifiable : Y n'a pas de parent et chaque X_i n'a que Y comme parent. Estimer un *naive Bayes* revient donc à estimer un modèle graphique dont la structure fixe est représentée dans la figure 2.2.

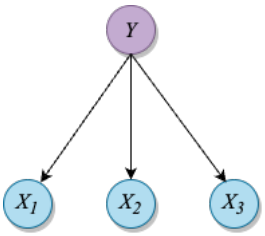


Fig. 2.2. : Représentation d'un *naive Bayes* comme un modèle graphique ($n = 3$).

Cette représentation des *naive Bayes* en tant que modèle graphique incite à proposer d'autres modèles graphiques un peu plus complexes, mais gardant au moins une partie de la structure *naive Bayes*.

2.1.3 *Tree Augmented Naive Bayes*

Le modèle graphique *naive Bayes* est caractérisé par un graphe où chaque X_i a un unique parent Y ; ce qui permet d'envisager un modèle légèrement plus complexe : autoriser que chaque X_i ait un second parent autre que Y. En respectant la contrainte d'acyclicité du graphe, les seconds parents décrivent donc un arbre (ou une forêt) parmi les X_i : on obtient le modèle *Tree Augmented Naive Bayes* (TAN) qu'on retrouve en figure 2.3.

La définition du MAP devient donc :

$$\hat{Y}_{MAP} = \arg\max_{y} \left(\mathbb{P}(Y) \cdot \prod_{i=1}^{n} \mathbb{P}(X_i \mid Y, parent(X_i)) \right)$$

où $parent(X_i) \in \mathbf{X}$ est le parent de X_i dans l'arbre s'il existe.

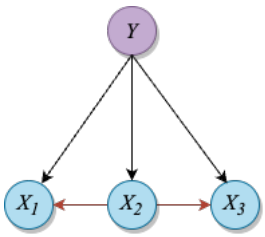


Fig. 2.3. : TAN : Y est parent de X_1, X_2, X_3 et $X_1 \leftarrow X_2 \rightarrow X_3$ forme un arbre.

Ce modèle est un peu plus riche que le *naive Bayes*, il reste toutefois d'une complexité contrôlée linéairement et possède un algorithme efficace pour identifier la partie variable de sa structure : l'arbre entre les attributs (voir algorithmes 2.1.1 et 2.1.2).

Algorithme 2.1.1 KRUSKAL, 1956

1. Trier les arêtes du graphe $\mathcal{G}$ par ordre décroissant de poids. Soit T l'ensemble des arêtes formant l'arbre de poids maximum. Initialiser $T \neq \varnothing$.
2. Ajouter la première arête à T.
3. Continuer à ajouter l'arête suivante à T seulement si elle ne crée pas de cycle dans T. Si aucune arête n'est plus disponible, sortir de l'algorithme et signaler que $\mathcal{G}$ est déconnecté.
4. Si T contient $n-1$ arêtes (où n est le nombre de nœuds dans $\mathcal{G}$), arrêter l'algorithme et afficher T. Sinon, retourner à l'étape 3.

Algorithme 2.1.2 Construit-TAN

1. $\forall i < j$, calculer $I_P(X_i; X_j \mid Y)$
2. Évaluer les arcs du graphe non-orienté complet des (X_i) par les $I(X_i; X_j \mid Y)$.
3. Calculer $\mathbb{E}(I_P(.;. \mid Z) = \frac{\sum_{i<j} I(X_i;X_j|Z)}{d(d-1)/2}$.
4. Supprimer les arcs de valeur $< \mathbb{E}(I(.;. \mid Z)$.
5. Appliquer Kruskal sur ce graphe pour obtenir une forêt couvrante de poids maximum.
6. Sur chaque partie connexe C de cette forêt, chercher $X_{root} = \arg\max_{X \in C} I(X; Y)$.
7. Utiliser ce X_{root} comme racine pour l'orientation de C
8. Rajouter Y dans le graphe, père de tous les X.
9. Apprendre les paramètres du BN.

2.1.4 Classifieur à partir de réseau bayésien

Les deux classifieurs précédemment évoqués sont basés sur des réseaux bayésiens dont la structure est en partie fixée, mais on peut aussi proposer un classifieur dont la structure serait un réseau bayésien totalement appris.

Ici, la structure du réseau est quelconque et la classe Y est traitée comme toute autre variable. La section 1.4.3 propose plusieurs méthodes permettant de trouver la structure de ce réseau bayésien. Le modèle obtenu est donc bien plus général et permet de spécifier tout type de relations probabilistes entre la classe et les attributs.

Afin de prédire la classe Y, il s'agit de calculer $\mathbb{P}(Y \mid \mathbf{X})$ qui prend dans ce cas la forme d'une inférence probabiliste (voir définition 1.2.5). Par ailleurs, la section 1.3.2 nous a montré que ce calcul revenait à calculer la probabilité de Y sachant la frontière de Markov de Y, ce qui revient à proposer un processus de sélection de variables strictement nécessaires à la classification, parmi l'ensemble des attributs $\mathbf{X}$. L'apprentissage des paramètres et les calculs de la probabilité de classification sont largement simplifiés.

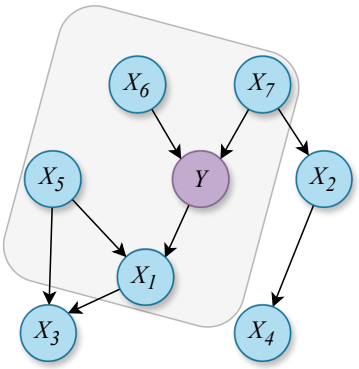


Fig. 2.4. : Exemple de réseau bayésien, sa structure nous permet de déduire que la classe Y ne dépend pas de X_2, X_3, X_4 et, car ils ne font pas partie de la frontière de Markov de Y.

Enfin, un avantage supplémentaire de l'apprentissage d'un tel réseau bayésien est la découverte de connaissances sur le domaine. En effet, la structure apprise expose les relations entre toutes les variables tant qualitativement (indépendances probabilistes) que quantitativement en explicitement les mécanismes probabilistes de génération de chacune des données. Ainsi, ces classifieurs fournissent un outil de validation peu fréquent, la confrontation directe du modèle et des experts [HECKERMAN et al., 1995].

2.2 Autres méthodes de classification

Dans le cadre de cette thèse, nous utiliserons principalement comme classifieurs les réseaux bayésiens. Pour justifier ce choix, il nous a paru important de pouvoir comparer cette méthode à d'autres classifieurs plus classiques. Il serait impossible de décrire ici toutes les méthodes de classification de l'état de l'art, nous en avons donc sélectionné quelques-unes qui sont populaires [JORDAN et MITCHELL, 2015], facilement implémentables avec des bibliothèques Python connues, comme scikit-learn [PEDREGOSA et al., 2011] et représentatives des différents types de fonctionnements. Ce sont ces méthodes que nous avons utilisées par la suite pour comparer les résultats à réseaux bayésiens de classification.

2.2.1 Régression Logistique

La régression logistique est une méthode de classification binaire qui se base sur une hypothèse de séparations linéaires des deux classes et sur des hypothèses sur la distribution de probabilité *a posteriori* ["Logistic regression" 2012].

Si les classes sont linéairement séparables, cela implique qu'il existe dans l'espace des attributs un hyperplan qui sépare les deux classes, comme illustré en figure 2.5.

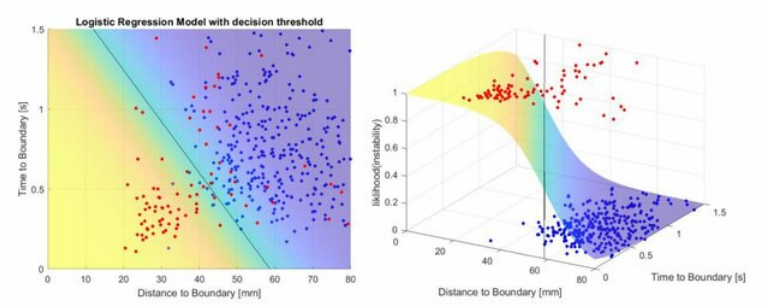


Fig. 2.5. : Visualisation du modèle de régression logistique sur un cas d'exemple [REIMEIR et al., 2021].

L'équation d'un hyperplan s'écrit :

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n = 0$$

Dans un cadre probabiliste de *maximum a posteriori*, pour une classe Y binaire, la séparation entre les deux classes vérifie :

$$\begin{aligned} \mathbb{P}(Y=0 \mid x) &= \mathbb{P}(Y=1 \mid x) \\ \iff \frac{\mathbb{P}(Y=0 \mid x)}{\mathbb{P}(Y=1 \mid x)} &= 1 \\ \iff \log \frac{\mathbb{P}(Y=0 \mid x)}{\mathbb{P}(Y=1 \mid x)} &= 0 \end{aligned}$$

Par analogie,

$$\log \frac{\mathbb{P}(Y=0 \mid x)}{\mathbb{P}(Y=1 \mid x)} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

En notant $p = \mathbb{P}(Y=1 \mid x)$ et $z = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$, on obtient :

$$\begin{aligned} \log \frac{p}{1-p} &= z \\ \iff \frac{1-p}{p} &= \exp^{-z} \\ \iff 1-p &= p \cdot \exp^{-z} \\ \iff 1 &= p(1+\exp^{-z}) \end{aligned}$$

On peut ainsi écrire la fonction de classification de la régression logistique :

$$f_{LR}(x) = P(Y=1 \mid x) = \frac{1}{1+e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}}$$

Les coefficients β sont optimisés pour maximiser la vraisemblance des observations sachant les variables d'entrée.

Cette méthode propose un nombre de paramètres très faible, car linéaire en fonction du nombre d'attributs. Par ailleurs, ces paramètres sont uniquement des poids associés à chaque attribut, ce qui permet de les ordonner par importance pour la classification, ce qui est une aide à l'explication. Mais il est à noter que cette modélisation repose sur des hypothèses arbitraires qu'il est difficile de vérifier.

2.2.2 Arbre de décision

Les arbres de décision présentent une structure hiérarchique en forme d'arbre. Ils partitionnent récursivement l'espace des variables en des régions dont une classe est de plus en plus majoritaire. Chaque feuille de l'arbre représente alors une région où l'on est raisonnablement certain de la classe [“Classification and regression trees (CART)” 2012].

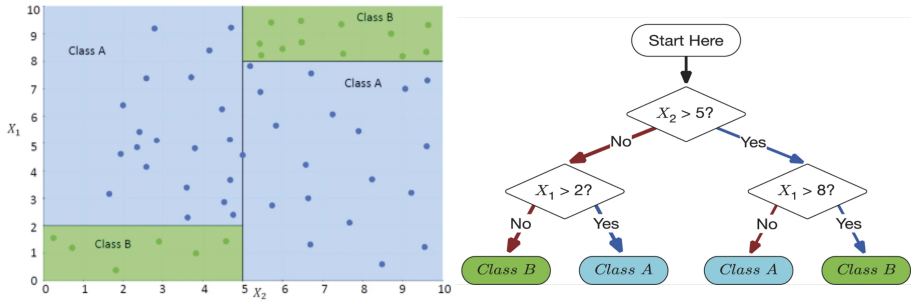


Fig. 2.6. : Exemple d'arbre de décision [VALDES et al., 2016]

Le processus de classification est alors direct. À chaque nœud interne de l'arbre, le choix d'une branche est pris en évaluant un test spécifique. Les feuilles de l'arbre représentent les résultats de la classification. Le processus de construction d'un arbre de décision implique la sélection des caractéristiques optimales à chaque nœud afin de maximiser le gain d'informations.

L'interprétabilité des arbres de décision est une caractéristique essentielle, car la logique régissant la classification est explicitement décrite dans un format basé sur des règles.

Ils sont donc simples à comprendre et à interpréter, facile à utiliser et versatiles. Mais ils ont des limites, les frontières de décision sont fréquemment alignées sur un des axes comme l'on peut apercevoir sur l'exemple de la figure 2.6. Mais surtout, les arbres de décisions manquent de robustesse : ils sont aussi très sensibles à des petites variations dans les données d'apprentissage [GÉRON, 2019].

2.2.3 Méthodes d'ensemble

Pour palier le manque de robustesse de certains classifieurs, une idée commune est d'agréger plusieurs modèles. On parle alors de méthodes d'ensemble. Le groupe (ou "ensemble") sera souvent plus performant que le meilleur modèle individuel ; en particulier si les modèles individuels produisent une grande variabilité de résultats.

L'instabilité d'un modèle devient alors une force. Il faut noter que cette robustesse acquise se fait au détriment de l'explicabilité, car une décision d'un modèle agrégé sera plus complexe à extraire que celle d'un modèle simple [HASTIE et al., 2009a].

Nous décrirons par la suite quelques méthodes ensemblistes.

Random Forest

Les arbres de décision en particulier sont un exemple type de classifieurs très sensibles aux modifications, même légères des données d'apprentissage, et en font un bon candidat pour une méthode d'ensemble : un *random forest* sera donc un ensemble d'arbres de décision, chacun appris sur une sous-partie aléatoire de la base d'apprentissage. Les prédictions sont obtenues par vote majoritaire sur l'ensemble des arbres appris [HASTIE et al., 2009c].

Expérimentalement, les *random forest* obtiennent de très bons résultats en termes de classification et de robustesse.

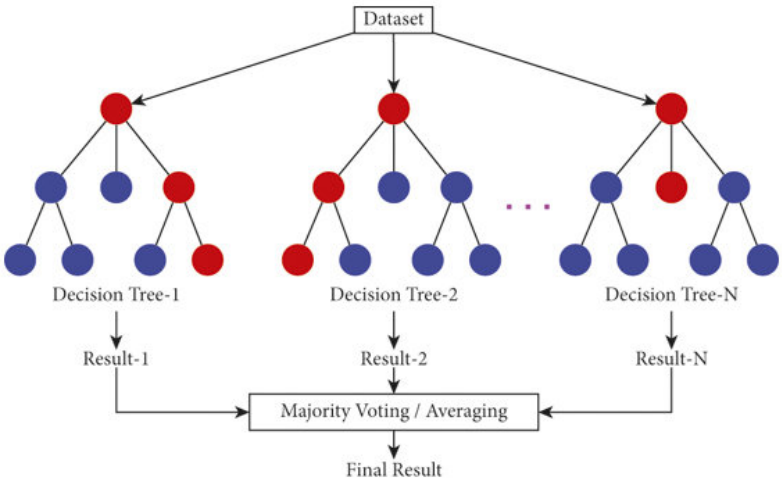


Fig. 2.7. : Illustration d'un Random Forest [KHAN et al., 2021]

Boosting

Le *boosting* désigne toute méthode d'ensemble capable de combiner plusieurs classifieurs faibles en un classifieur de plus en plus fort. L'idée générale de la plupart des méthodes de boosting est de former des classifieurs de manière séquentielle, chacun essayant de corriger son prédécesseur. Il existe de nombreuses méthodes

de *boosting*, et une des plus populaires s'appelle AdaBoost, pour *Adaptive Boosting* [Freund et Schapire, 1997].

La technique utilisée par AdaBoost est la suivante : pour qu'un nouveau classifieur corrige ses prédécesseurs, il fait plus attention aux données d'apprentissage qui viennent d'être mal prédites. Ainsi, on se concentre petit à petit sur les cas difficiles.

Plus précisément, quand on entraîne un classifieur AdaBoost, l'algorithme entraîne d'abord un classifieur simple (comme à nouveau un arbre de décision) et l'utilise pour faire des prédictions sur la base d'apprentissage. L'algorithme augmente ensuite le poids des instances mal classées. Puis, il entraîne un deuxième classifieur, en utilisant les poids mis à jour, et fait à nouveau des prédictions sur la base, actualise les poids et ainsi de suite.

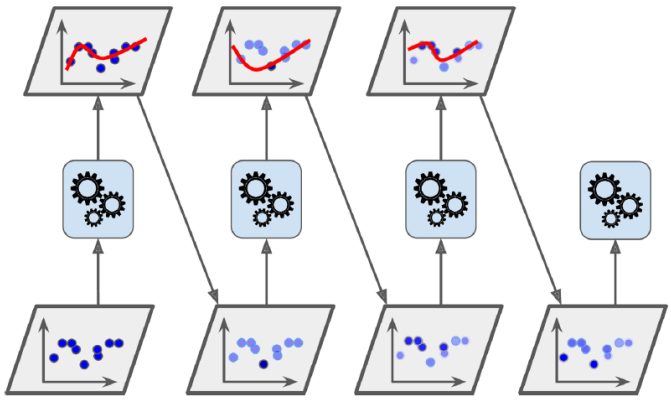


Fig. 2.8. : Illustration d'AdaBoost [Géron, 2019]

Une autre méthode de *boosting* populaire est le *Gradient Boosting* [Friedman, 2001]. L'algorithme utilise la descente de gradient pour minimiser les erreurs à chaque étape. Il ajuste les nouveaux modèles en fonction du gradient des erreurs par rapport aux prédictions du modèle combiné. Une implémentation optimisée du Gradient Boosting est disponible dans la bibliothèque XGBoost, pour "*Extreme Gradient Boosting*". Cette bibliothèque se veut extrêmement rapide, adaptable et efficace [Chen et Guestrin, 2016].

Les méthodes d'ensembles prouvent leur efficacité, car elles permettent d'améliorer les résultats et la robustesse, et même pour certaines réduire le biais des classifieurs. Il y a toutefois peu de résultats théoriques sur ces méthodes et les conclusions expérimentales sont parfois contradictoires [Hastie et al., 2009d].

2.2.4 Réseaux de neurones

Les premières méthodes de classification que nous avons décrites fournissent un modèle simple de calcul de la classe, soit en termes de nombre de paramètres, soit en termes de structure. Les méthodes d'ensembles quant à elles agrègent un ensemble de ce type de classifieur. Il existe une troisième catégorie de classifieurs que nous devons évoquer dans ce chapitre : les réseaux de neurones. Ils se caractérisent par une structure dense et un nombre de paramètres généralement importants. Même si leur capacité de prédiction est *a priori* reconnue, le manque d'explicabilité (boite noire) est dans notre cadre un inconvénient important.

Un réseau de neurones est un modèle inspiré des neurones biologiques. Un neurone formel modélise le neurone biologique comme une fonction qui se décompose en une première étape d'agrégation linéaire de ses entrées, suivi d'un opérateur non linéaire (la sigmoïde), censé représenter les phénomènes de saturations électrochimique du neurone [HASTIE et al., 2009b].

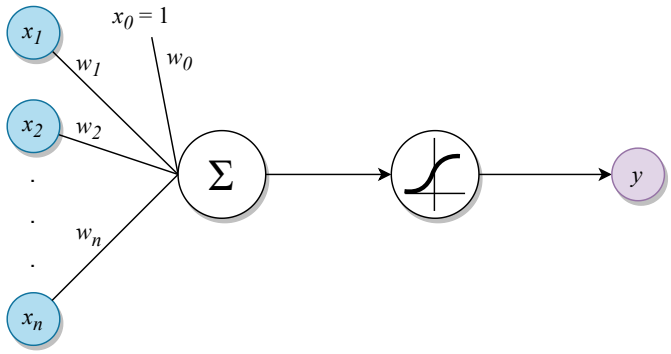


Fig. 2.9. : Modèle d'un neurone formel.

La fonction de classification de réseau de neurones est alors :

$$f_{RN}(x) = \sigma(\sum_{i=0}^{n} w_i x_i)$$

Tout comme la régression logistique, un neurone formel propose une classification binaire et linéairement séparable dont la frontière a pour équation :

$$\sum_{i=0}^{n} w_i x_i = 0$$

Toutefois, contrairement aux classificateurs de régression logistique, les perceptrons n'émettent pas de probabilité de classe ; ils effectuent plutôt des prédictions basées

sur un seuil strict.

À la façon d'un modèle d'ensemble hiérarchique, le perceptron multi-couche (MLP pour *multi-layer perceptron*) organise un grand nombre de neurones (donc de classifieurs binaires), soit en parallèle (à l'intérieur d'une couche), soit en série (entre les différentes couches). Un perceptron multi-couche est donc composé d'une couche d'entrée (traversante), d'une ou plusieurs couches appelées couches cachées, et d'une dernière couche appelée couche de sortie.

Les couches supérieures opèrent donc des classifications de résultats de classifications précédentes, ce qui fournit au MLP des outils de manipulation de concepts et d'abstraction de plus haut niveau. L'apprentissage d'un tel modèle est complexe, basé principalement sur la rétropropagation d'erreur à travers les différentes couches du réseau, converge lentement et nécessite une grande quantité de données supervisées.

Toutefois, les réseaux de neurones sont polyvalents, puissants et évolutifs, ce qui les rend idéaux pour s'attaquer à des tâches d'apprentissage automatique vastes et très complexes, telles que la classification de très grande base d'images.

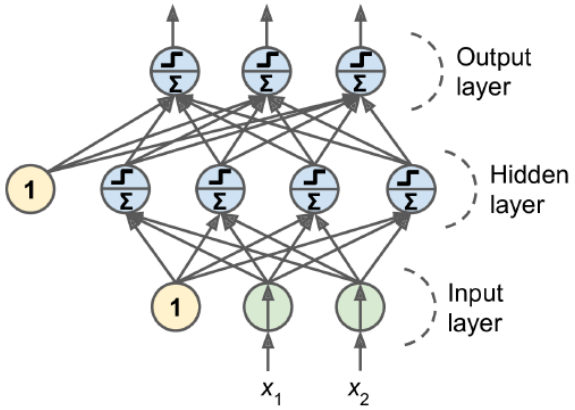


Fig. 2.10. : Illustration d'un perceptron multi-couche [GÉRON, 2019]

Enfin, un *deep neural network* est une extension des réseaux de neurones dans laquelle on spécialise certaines couches de neurones, on augmente le nombre de couches et on augmente considérablement la taille des bases de données. C'est le cadre du *deep learning*, ou apprentissage profond.

Ces explications pourraient être bien plus développées, et il existe des modèles que nous n'avons même pas évoqués, tel que les machines à vecteurs de support (SVM), mais nous justifierons plus tard ce choix.

2.3 Validation des modèles

Dans la section précédente, nous avons listé un ensemble de méthodes de classification qui ont toutes la capacité de mener à bien la tâche de prédiction. Toutefois, elles reposent chacune sur leur hypothèse propre, ont des cadres d'utilisations spécifiques et fonctionnent de manières très différentes en fonction des données d'apprentissage. Il devient alors essentiel de pouvoir évaluer et comparer les résultats de ces méthodes, autant pour les améliorer, que pour les discriminer ou les sélectionner.

La qualité d'un modèle dépend de la qualité de sa prédiction. Il est donc très important de valider le modèle avant son utilisation. L'option la plus commune est de séparer son jeu de donnée originel en deux jeux : un pour l'apprentissage et un pour la validation. On teste notre modèle sur le jeu de données qui n'a pas servi à l'apprentissage et cela permet d'éviter l'écueil du surapprentissage et donne une idée de son pouvoir prédictif sur des nouvelles données.

Le surapprentissage (ou *overfitting* en anglais) est un phénomène courant en classification, qui se produit lorsqu'un modèle est trop complexe et qu'il a été entraîné de manière trop précise sur les données d'entraînement, ce qui réduit sa capacité à généraliser à de nouvelles données.

2.3.1 Scores pour classifieur binaire

Les scores d'évaluation de modèles de classification permettent une quantification de la performance du modèle. Il existe de nombreux scores différents qui sont plus ou moins adaptés à chaque usage.

Matrices de confusion

La matrice de confusion est l'outil essentiel pour la construction de ces scores. Elle compte les prédictions correctes et incorrectes du modèle en séparant par classe. Sa taille dépend du nombre de classes du modèle.

Dans le cadre d'un classifieur binaire, la matrice de confusion se compose de :

- Vrais négatifs : le nombre de cas négatifs qui sont correctement classés en tant que négatif par le modèle.
- Faux négatifs : le nombre de cas positifs qui sont incorrectement classés en tant que négatif par le modèle.

- Faux positifs : le nombre de cas négatifs qui sont incorrectement classés en tant que positif par le modèle.
- Vrais positifs : le nombre de cas positifs qui sont correctement classés en tant que positif par le modèle.

	Classe réelle : 0	Classe réelle : 1
Classe Prédite : 0	Vrais Négatifs (VN)	Faux Négatifs (FN)
Classe Prédite : 1	Faux Positifs (FP)	Vrais Positifs (VP)

Tab. 2.1. : Matrice de confusion d'un classifieur binaire

Tous les scores décrits ensuite se basent sur cette matrice de confusion dans le cas binaire. En effet, dans le cas du multi-classe, la matrice de confusion est plus difficile à synthétiser.

Métriques d'évaluation

Le score le plus utilisé pour évaluer un classifieur binaire est l'*accuracy*. Pour éviter toute confusion, nous utiliserons son terme anglais, car bien que traduisible par "exactitude", il est peu employé et parfois confondu avec précision. Elle mesure le taux de "bonnes réponses" soit :

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN}$$

Un score proche de 1 indique que le modèle a une grande proportion de prédictions correctes. C'est un score simple qui peut être trompeur dans le cas de modèles avec des classes déséquilibrés. En effet, si la classe cible est peu représentée, le modèle pourra avoir une très bonne *accuracy* en ne prédisant quasiment jamais cette classe.

La précision ou valeur prédictive positive (VPP) est la proportion de vrais positifs parmi les cas prédits comme positifs par le modèle. Une précision faible indique donc que le modèle crée trop de faux positifs, et donc "surestime" cette classe.

$$\text{Précision} = \text{VPP} = \frac{VP}{VP + FP}$$

Le rappel ou sensibilité ou taux de vrais positifs (*TPR*) mesure la proportion de cas positifs qui sont correctement identifiés par le modèle. Un rappel faible indique

alors que le modèle "sous-estime" la classe positive.

$$\text{Rappel} = \text{Sensibilité} = TPR = \frac{VP}{VP + FN}$$

Il existe généralement un compromis entre précision et rappel qu'on essayera de synthétiser dans des scores plus sophistiqués par la suite.

La spécificité est la propension d'être bien classé négativement sur l'ensemble des cas négatifs.

$$\text{Spécificité} = \frac{VN}{VN + FP}$$

La valeur prédictive négative (VPN) indique la probabilité qu'un résultat négatif corresponde réellement à un cas négatif :

$$\text{VPN} = \frac{VN}{VN + FN}$$

Spécificité et VPN sont respectivement les "rappel" et "précision" pour la classe négative.

On peut de la même façon définir le taux de faux positifs (*FPR*) :

$$FPR = \frac{FP}{FP + VN} = 1 - \text{Spécificité}$$

Les scores, vus précédemment, basés sur les composantes de la matrice de confusion binaire, apportent chacun une information complémentaire sur le comportement du classifieur. Ainsi, améliorer un classifieur peut revenir à arbitrer une augmentation de la précision ou une augmentation du rappel. Ce compromis entre précision et rappel peut être explicité sous la forme d'une métrique appeler F-score.

Définition 2.3.1 (*F-score*)

Le F-score est une métrique qui combine précision et rappel pour une évaluation plus complète du modèle.

$$\textit{F-score} = 2 \times \frac{\textit{Précision} \times \textit{Rappel}}{\textit{Précision} + \textit{Rappel}}$$

Il s'agit de la moyenne harmonique des scores de précision et de rappel. Un F-score élevé indique donc que le modèle a à la fois une bonne précision et un bon rappel.

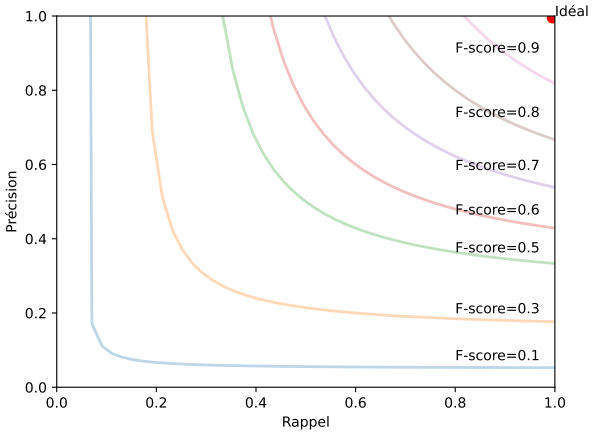


Fig. 2.11. : Courbe Précision-Rappel avec lignes iso F-scores.

Le F-score se généralise avec le F-beta score :

Définition 2.3.2 (*F-beta score*)

$$F_\beta score = (1+\beta^2) \times \frac{Précision \times Rappel}{\beta^2 \times Précision + Rappel}$$

Le F-beta score est une métrique qui permet de pondérer l'importance de la précision ou du rappel grâce au paramètre β. Le F-score est un cas particulier du F-beta score lorsque $\beta = 1$, c'est pourquoi on le nomme parfois F1-score. Pour $\beta < 1$, le score donne plus d'importance à la précision et pour $\beta > 1$, au rappel. Dans un contexte médical de prévention et diagnostic, il peut être intéressant de donner plus de poids au rappel pour minimiser le risque de non-diagnostic d'un patient malade, une discussion sur la valeur de β peut alors tout à fait être pertinente.

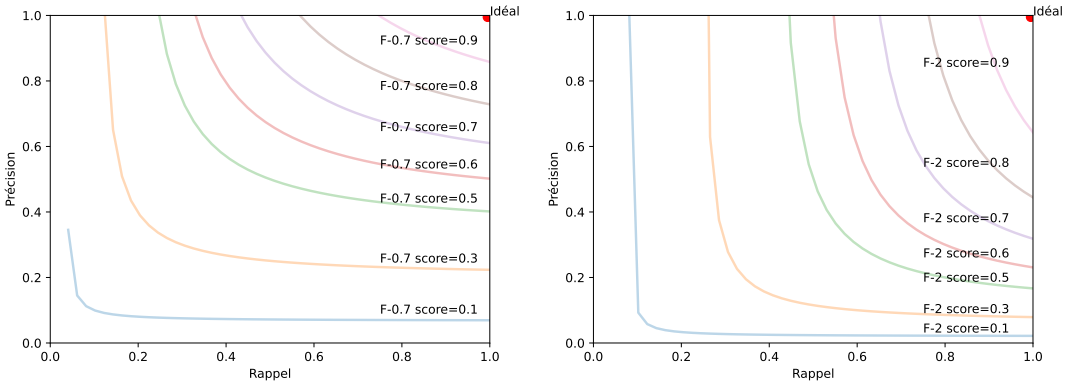


Fig. 2.12. : Courbe Précision-Rappel avec lignes iso F-beta scores avec $\beta = 0.7$ à gauche et $\beta = 2$ à droite

2.3.2 Courbes ROC et Précision-Rappel

Un classifieur probabiliste propose donc une valeur de probabilité d'appartenir à la classe positive pour chaque élément de la base de validation. Plutôt que d'utiliser le MAP, il prédit une classe pour chaque élément grâce à un seuil sur cette valeur de probabilité. C'est de la confrontation de cette valeur de cette prédiction et de la classe indiquée dans la base qu'émerge l'ensemble des critères ci-dessus. Pour des raisons de robustesse du modèle ou pour essayer de s'abstraire du choix arbitraire du seuil, des outils existent qui permettent d'évaluer un classifieur probabiliste uniquement à partir de la liste des probabilités prédites. Ce sont les courbes ROC (*Receiver Operating Characteristic*) [EGAN, 1975] et PR (Précision-Rappel) [DAVIS et GOADRICH, 2006].

Les deux courbes partent du même principe : si on trie l'ensemble des éléments de la base par probabilité prédite croissante dans une liste L, tout classifieur probabiliste consistera à séparer cette liste en deux parties (L_0, L_1) et à affecter la classe négative à la première partie L_0 et la classe positive à la seconde L_1. S'il y a N éléments dans la base, il y a donc N façon de séparer cette liste en deux, correspondant à N classifieurs différents dont on peut calculer pour chacun *TPR*, *FPR*, précision et rappel.

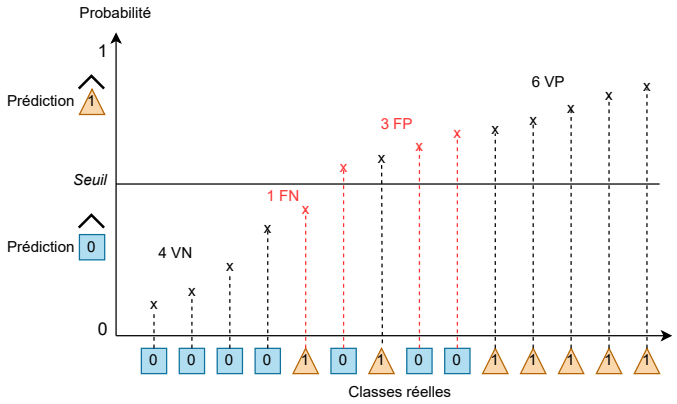


Fig. 2.13. : Tri des éléments de la base selon leur probabilité dans un cas exemple où la classe négative est représentée par un carré bleu et la classe positive un triangle orange.

Dans la figure 2.13, avec cette séparation arbitraire qu'on appellera seuil, on a alors :

$$TPR = \text{Rappel} = \frac{6}{7}\,;\; FPR = \frac{3}{7} \text{ et Précision} = \frac{6}{9}.$$

En affichant les N points (TPR, FPR) correspondant aux N seuils, on obtient

la courbe ROC ; en affichant les N points $(\text{Précision}, \text{Rappel})$ on obtient la courbe PR.

La courbe ROC se déploie dans un repère avec en abscisse *FPR* et en ordonnée *TPR*. Elle commence forcément en $(0, 0)$, qui correspond à une seconde partie vide de la liste triée. Elle se termine forcément en $(1, 1)$ qui correspond à la première partie vide de la liste triée. Un point (x, y) de la courbe tel que $x < y$, impliquerait qu'il y a plus de faux positifs que de vrais positifs. La courbe ROC est donc au-dessus de la diagonale principale. Enfin, un classifieur idéal aurait un FPR à 0 et un TPR à 1. Ces deux dernières remarques suggèrent que plus la courbe ROC passe près du point idéal, plus le classifieur est pertinent. Un critère plus formel pour expliciter cette propriété est que l'aire sous la courbe ROC soit maximale. On nomme ce critère AUC pour *area under the curve*. L'AUC appartient à $[0, 1]$ et représente la probabilité qu'un élément positif choisi aléatoirement aura une probabilité prédite supérieure qu'un élément négatif choisi aléatoirement de la même façon. Elle permet donc de comparer de manière robuste plusieurs classifieurs probabiliste.

La courbe PR se déploie dans un repère avec en abscisse le rappel et en ordonnée la précision. Lorsque la seconde partie de la liste triée est vide, le rappel est égal à 0 et la précision n'est pas définie, mais conventionnellement fixée à 1. Le classifieur idéal correspond au point $(1, 1)$. C'est pourquoi l'AUC est aussi un critère synthétique pour cette courbe.

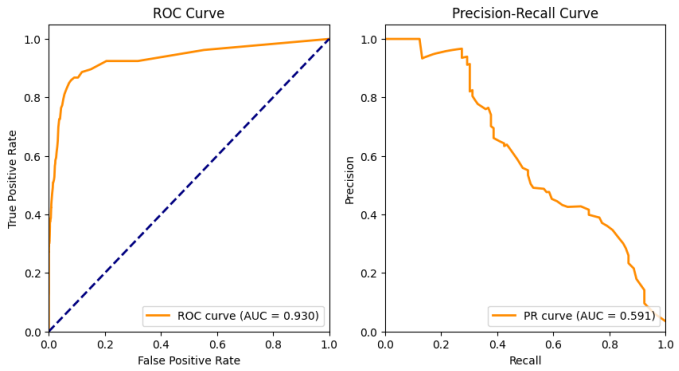


Fig. 2.14. : Exemple de courbes ROC et Précision-Rappel [CALLE, 2023].

2.3.3 Choix du seuil d'un classifieur probabiliste binaire

Même si les concepts de ROC et PR d'évaluer une estimation probabiliste sans expliciter la frontière de décision, le classifieur nécessite de prédire la classe en sélectionnant un seuil de probabilité qui sépare les éléments classés positivement et

négativement. On peut remarquer que cette séparation correspond à sélectionner un des couples (L_0, L_1) obtenu précédemment. Les courbes ROC et PR peuvent donc nous aider à sélectionner cette séparation.

En ce qui concerne la courbe ROC, il s'agirait par exemple de sélectionner le point de la courbe le plus "proche" du point idéal. En ce qui concerne la courbe PR, on choisira le point qui maximise le critère F-$beta$. Un avantage de la courbe PR par rapport à la courbe ROC est donc la prise en compte dans le choix du seuil du compromis précision/rappel dans le cadre de l'application. Par ailleurs, dans un contexte de classes déséquilibrées, il sera aussi plus avisé de choisir le seuil de la courbe Précision-Rappel, car cette courbe prend plus en compte la prévalence des classes.

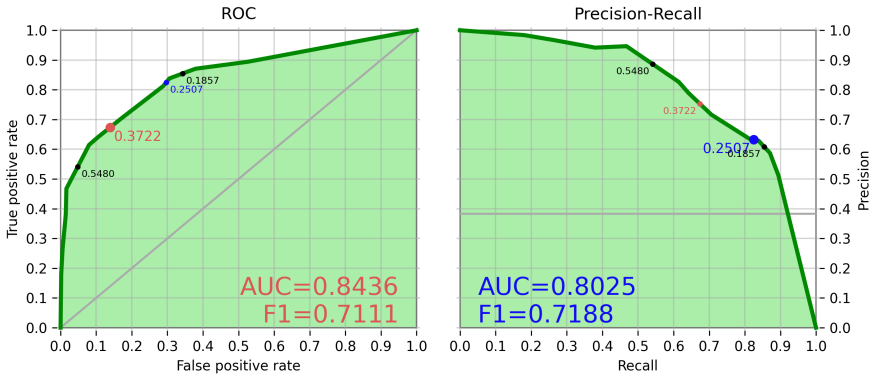


Fig. 2.15. : Exemple d'une courbe ROC avec son point optimal en rouge et d'une courbe précision-rappel avec son point optimal, qui maximise le F-score en bleu. Les seuils maximisant le F-2 score (0.1857) et le F-0.7 score sont aussi indiqués (0.5480) en noir. La localisation des mêmes seuils sur l'autre courbe est aussi représentée.

Le choix du seuil de probabilité qui permet de différencier les classes est un paramètre important dans les classifieurs probabilistes. En effet, le choix évident du seuil de probabilité égale à 0.5 correspond à fixer un nombre de faux positifs et de faux négatifs et donc à fixer tous les critères. Il est nécessaire de pouvoir changer ce seuil dans le cadre d'une décision sur le compromis précision-rappel. Comme on peut le voir sur la figure 2.13, le seuil agit comme une frontière de décision qui peut avoir beaucoup d'impact.

2.4 Gestion des données manquantes

L'imputation des valeurs manquantes est une étape cruciale dans le processus de prétraitement des données, en particulier lorsqu'il s'agit de données du monde "réel" où les informations peuvent être incomplètes. Les valeurs manquantes peuvent nuire aux performances des modèles d'apprentissage automatique, car de nombreux algorithmes ne peuvent pas fonctionner directement en leur présence. Par conséquent, des méthodes d'imputation sont appliquées pour estimer ou compléter les valeurs manquantes dans l'ensemble de données.

On distingue plusieurs types de données incomplètes [MACK et al., 2018] :

- **MCAR (*Missing Completely At Random*) :** ici, les données sont manquantes complétement aléatoirement, le fait que des valeurs soient manquantes est complètement indépendant des données observées et non observées. Par exemple, des relevés de laboratoires qui manquent à cause d'un problème technique exceptionnel. Il n'y a pas de différences systémiques entre les personnes avec ces variables présentes ou absentes.
- **MAR (*Missing At Random*) :** quand des données sont manquantes aléatoirement, le fait que les données soient manquantes est relié à des variables observées, et pas à des variables non observées. Par exemple, si dans une étude médicale, un questionnaire doit être rempli par le patient en ligne, le fait que la valeur soit manquante peut être lié à son âge. Ainsi, si la probabilité de complétion est liée à l'age, qui est une variable observée et non pas à la cible de l'étude, on considère les données comme MAR.
- **NMAR (*Not Missing At Random*) :** des données ne sont pas manquantes aléatoirement si le fait qu'elles manquent sont liées à une donnée non observée. Par exemple, des personnes en bonne santé générale sont moins susceptibles de se faire peser lors d'une consultation médicale ou en tout cas moins régulièrement qu'une personne en surpoids ou en sous-poids. Si une étude sur les poids est effectuée ici, les valeurs manquantes seront liées à un facteur qui n'est *a priori* pas mesuré dans l'étude.

Si les valeurs manquantes sont largement minoritaires, la suppression des lignes de la base dans lesquelles il y a des valeurs manquantes pourrait être envisagée, mais elle implique d'être sûre que ces variables sont manquantes complètement aléatoirement. Cette hypothèse est rarement facile à démontrer et on prend alors le risque d'introduire des biais dans la prédiction.

Dans un cas plus général, voici quelques méthodes courantes [ACUÑA et RODRIGUEZ, 2004] d'imputation des valeurs manquantes :

- **Imputation de la moyenne ou de la médiane :** Remplacer les valeurs manquantes par la moyenne, la médiane ou le mode des valeurs observées (non manquantes) dans la variable concernée. Cette méthode est simple et souvent utilisée lorsque les valeurs manquantes sont supposées être totalement aléatoires.
- ***Forward fill* :** Remplir les valeurs manquantes avec la valeur observée la plus récente (ou ultérieure). Cette méthode est souvent utilisée dans les données de séries chronologiques où les valeurs manquantes sont supposées suivre la dernière valeur observée.
- **Imputation par k plus proches voisins (KNN) :** Estimer les valeurs manquantes en prenant en compte les valeurs des k plus proches voisins dans l'ensemble de données. Cette méthode est utile lorsqu'il existe une structure locale dans les données. Cette méthode a d'abord été utilisée pour l'estimation des puces à ADN (*DNA microarrays*) avec la proposition des k plus proches voisins pondérés par la distance euclidienne [Troyanskaya et al., 2001].
- **Imputation à l'aide de modèles d'apprentissage automatique :** Former un modèle d'apprentissage automatique pour prédire les valeurs manquantes sur la base des autres caractéristiques de l'ensemble de données. Il peut s'agir de méthodes telles que les arbres de décision, les forêts aléatoires ou même des modèles plus complexes tels que le *gradient boosting* ou les réseaux neuronaux. Dans ce cadre, on peut aussi utiliser la méthode EM décrite précédemment dans l'algorithme 1.4.1.
- **Indicateurs d'absence :** Créer des variables indicatrices binaires pour indiquer si une valeur est manquante ou non. De cette façon, le modèle d'absence est préservé et les valeurs manquantes peuvent être traitées comme une catégorie distincte au cours de l'analyse.

Le choix de la méthode d'imputation dépend de la nature des données, de la raison de l'absence de données et des hypothèses sur le mécanisme des données manquantes. Il est souvent utile d'évaluer l'impact des différentes méthodes d'imputation sur les performances du modèle d'apprentissage automatique en aval. En outre, il convient de veiller à ne pas introduire de biais au cours du processus d'imputation.

[Acuña et Rodriguez, 2004] ont étudié l'effet de quatre méthodes de traitement des valeurs manquantes. Ces méthodes comprennent la suppression des cas manquants et trois méthodes d'imputation : l'imputation par la moyenne, l'imputation par la médiane et l'imputation par KNN. Leurs résultats montrent que l'imputation n'a pas d'effet significatif sur la précision de la classification, ce qui correspond aux

résultats relativement plus anciens de [DIXON, 1979]. Il faut cependant noter que la comparaison s'est arrêtée sur des méthodes relativement simples.

Une étude similaire a été faite par [BATISTA et MONARD, 2003]. Les résultats montrent ici que l'imputation par KNN permet d'obtenir une bonne précision, mais uniquement lorsque les variables ne sont pas fortement corrélées entre elles.

Il est à noter que c'est un domaine avec peu de bases théoriques, il est difficile de savoir ce qu'on optimise lorsqu'on impute les valeurs manquantes, ce qui permet de proposer beaucoup de méthodes différentes, généralement mal généralisable.

2.5 Méthodes de discrétisation

Dans le chapitre précédent, nous avons commencé à aborder la discrétisation en 1.1.2. En pratique, dans une base de données, une variable est représentée par l'ensemble des valeurs qu'elle prend dans la base et ne peut pas être réellement continue. Mais, si l'ensemble de ces valeurs est un ensemble de "grande taille" de nombres réels, alors on considère que cette variable est continue. Un ensemble de "grande taille" correspond à un ensemble dont la taille est du même ordre de grandeur que celle de la base. Dans le cas où l'ensemble de ces valeurs serait de petite taille, on considère que la variable est discrète.

Si la variable est continue, il est possible de la paramétrer en l'assimilant, par exemple, à une loi normale. Mais, contrairement à ce qui est communément admis, limiter l'ensemble des distributions continues aux distributions paramétriques est souvent extrêmement réducteur et biaise nécessairement les traitements ultérieurs. Cette proposition est d'autant plus vraie pour les classifieurs *model-based* comme les réseaux bayésiens et plus généralement la plupart des méthodes explicables. En ce sens, décider d'un modèle paramétrique particulier pour une distribution continue est un acte aussi arbitraire que de discrétiser en décidant l'ensemble des valeurs que prend effectivement la variable. Ainsi, comme dit précédemment en 1.2.2, dans le cadre de ce travail, nous nous intéresserons exclusivement à des modèles discrétisés.

La discrétisation est donc le processus de conversion de données continues en données discrètes. Elle consiste principalement à décider d'intervalles de valeurs pour les données continues et à transformer la variable en remplaçant dans la base chaque valeur de la variable par l'intervalle qui la contient. Le processus de discrétisation se caractérise donc par la méthode d'identification de ces intervalles.

Voici quelques méthodes courantes de discrétisation des données automatiques [GARCIA et al., 2013] :

- **Discrétisation à intervalle égale :** cette méthode consiste à diviser la plage de valeurs continues en intervalles de largeur égale. Il s'agit d'une méthode simple, mais qui est *a priori* et donc qui ne s'adapte pas aux données.
- **Discrétisation à quantile égal :** dans cette approche, les intervalles sont choisis de telle sorte que la proportion du nombre de cas dans l'intervalle soit la même. Les intervalles peuvent donc avoir des tailles très variées en fonction de leur densité. C'est une méthode simple qui minimise l'arbitraire, car il suffit de choisir le nombre de classes et qui prend en compte la distribution des données de la base d'apprentissage.
- **Discrétisation basée sur le *clustering* :** des techniques telles que le regroupement par *k-means* [MACQUEEN, 1967] peuvent être appliquées pour regrouper les valeurs similaires. Les *clusters* deviennent alors les valeurs discrètes. Cette méthode s'adapte plus étroitement à la distribution des données que la discrétisation à largeur ou à fréquence égales.
- **Discrétisation basée sur l'entropie :** l'entropie mesure l'incertitude ou le désordre dans un ensemble de valeurs. La discrétisation s'effectue en trouvant les points de séparation qui minimisent l'entropie. Par exemple, l'algorithme MDLP [FAYYAD et IRANI, 1993].
- **Discrétisation basée sur la fréquence :** pour le cas des variables catégorielles, dont l'ensemble de valeurs est de "grande taille", il existe des méthodes de discrétisation particulières comme celle-ci. Elle consiste à regrouper, dans une même classe, les valeurs catégorielles de même fréquence. Cette méthode consiste à discrétiser les données en fonction de la fréquence des valeurs. Par exemple, créer des catégories pour les valeurs courantes, rares et aberrantes. Cette méthode peut s'avérer utile par exemple lorsque les catégories ont perdu leur sens (codification, anonymisation, etc.). La variable discrétisée n'est alors que la représentation d'une information très partielle de la variable initiale.

Enfin, la connaissance du domaine peut être utilisée pour définir manuellement des intervalles basés sur les caractéristiques spécifiques des données. C'est ce qu'on appelle la discrétisation experte. Cette approche est certainement la plus pertinente, mais demande une connaissance experte qu'il est souvent difficile de collecter.

De même que pour les valeurs manquantes, le critère à optimiser lors de la discrétisation n'est pas facile à spécifier. Il en découle aussi un grand nombre de méthodes originales, comme les arbres de décision ou les méthodes basées sur

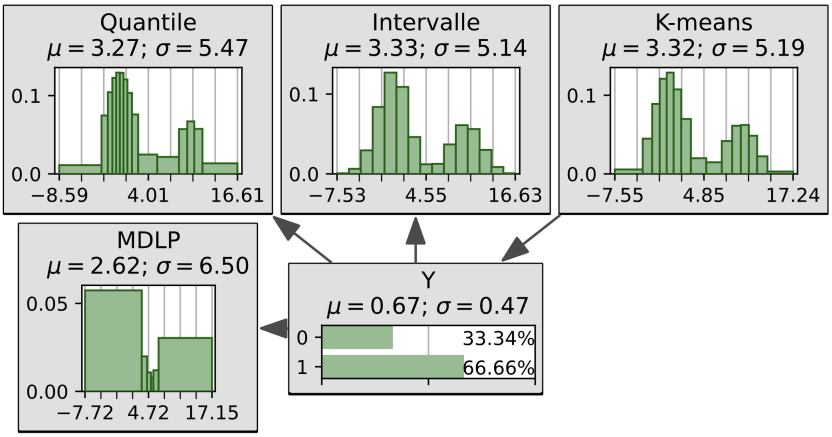


Fig. 2.16. : Comparaison de différentes méthodes de discrétisation sur une mixture de deux gaussiennes.

l'entropie qui sont souvent des approches intéressantes, mais dont il est difficile de cerner les cas d'utilisations ou même de tracer leurs utilisations en dehors des articles qui les proposent.

Le choix d'une méthode de discrétisation est principalement empirique, mais il est bien évidemment essentiel de prendre en compte la nature des données, les exigences de la tâche de modélisation et les caractéristiques des algorithmes utilisés. En pratique, il est souvent judicieux d'expérimenter différentes méthodes et d'évaluer leur impact sur les performances des tâches en aval [GARCIA et al., 2013].

2.6 Interprétation des modèles

Les récents progrès des classifieurs permettent de les utiliser pour des problèmes de plus en plus complexes, de plus en plus sensibles, et il n'est pas rare de voir des processus de décision complètement automatiques basés sur ces classifications. Que ce soit dans le processus de prise de décision ou de l'aide à la décision, outre les performances des classifieurs, la capacité du modèle à expliquer cette décision apparaît donc de plus en plus important. XAI est le nom de domaine qui s'intéresse à cette interprétabilité ou explicabilité des modèles de décision.

2.6.1 Boite noire, boite blanche

Une nouvelle catégorisation des modèles en découle : on distinguera les modèles dont les paramètres appris permettent un certain accès au processus de la décision,

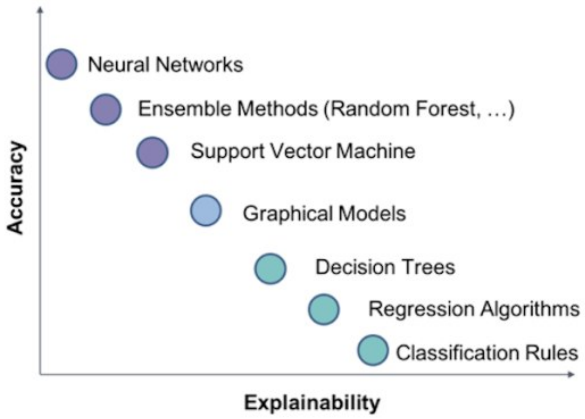


Fig. 2.17. : Classification des méthodes de classifications [DAM et al., 2018]

ce sont les modèles dit "boites blanches" et les modèles dont les paramètres en trop grand nombre ou trop diffus rendent opaque ce processus de décision (les "boites noires").

La figure 2.17 illustre, de manière simplifiée, une distribution des modèles en fonction de la pouvoir prédictif et de leur explicabilité. On remarque que les réseaux de neurones sont efficaces, mais très opaques, comme expliqué précédemment. Les régressions logistiques et les arbres de décision sont eux explicables, mais généralement moins exacts. Les modèles graphiques, dont les réseaux bayésiens, apparaissent comme un meilleur compromis entre les deux critères.

Toutefois, il est à noter que la représentation schématique de la figure 2.17 insiste sur une idée préconçue du compromis à trouver entre *accuracy* et explicabilité : même si les boites noires ont en effet généralement un meilleur pouvoir prédictif que les boites blanches, la disposition linéaire proposée nous semble très biaisée. Par exemple, d'après nos expériences, le pouvoir de prédiction des réseaux de neurones ne les placent pas si haut, au-dessus des Random Forest ou XGBoost. De même, pour les réseaux bayésiens, le fort gain en explicabilité tel que nous l'avons décrit dans le chapitre précédent en section 1.3 ne s'obtient pas contre une aussi grande perte de pouvoir prédictif.

Nous proposons donc une version amendée de ce schéma plus proche de nos expériences.

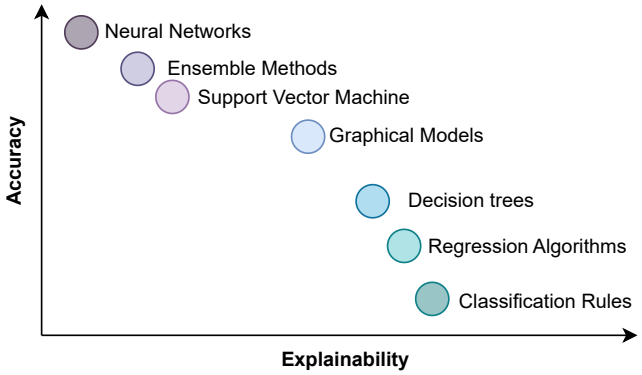


Fig. 2.18. : Classification des méthodes de classifications d'après nous.

2.6.2 Outils d'interprétations *Post-Hoc*

Il existe de nombreuses méthodes qui essayent de fournir des explications sur les modèles après sa conception et qui peuvent être appliquées à toutes les méthodes de classification.

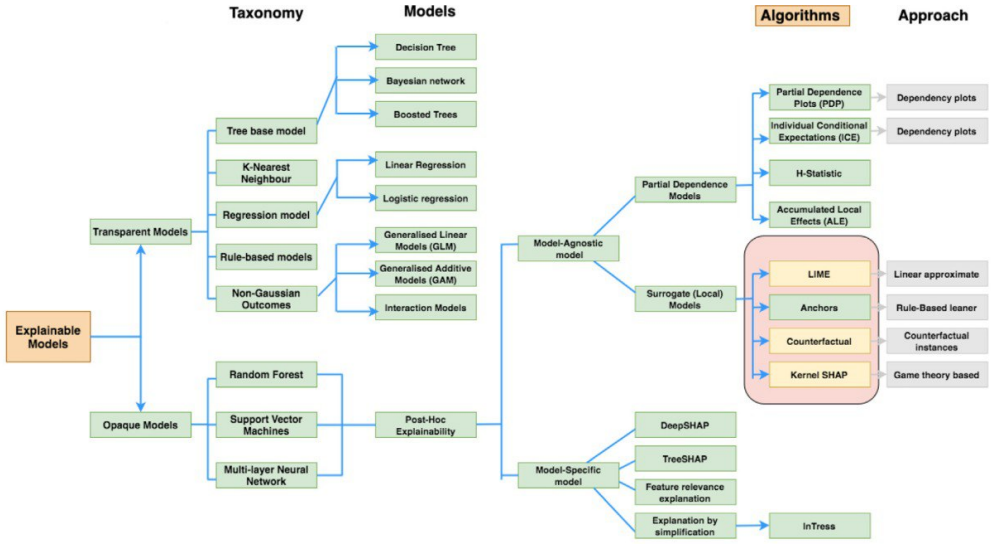


Fig. 2.19. : Taxinomie de l'*XAI* [RASHEED et al., 2022]

Une technique de *XAI* largement utilisée est dérivée de la théorie des jeux : les valeurs de Shapley [SHAPLEY, 1953]. C'est à l'origine une méthode pour répartir les valeurs de jeux entre un ensemble de joueurs dans un jeu de coalition. Dans le cadre du *machine learning*, en interprétant la fonction de classification comme

une fonction de valeur de jeu, les valeurs de Shapley permettent d'évaluer la contribution de chacune des variables à la valeur de prédiction pour chaque ligne de la base fournissant donc une interprétabilité locale expliquant l'importance de chaque variable dans la prise de décision pour ce cas précis. En calculant leur espérance sur la base complète, on obtient les contributions moyennes de chacune des variables, permettant de les ranger par ordre d'importance, ce qui fournit aussi une interprétabilité globale de l'importance de chaque variable. Il est toutefois à noter que ces calculs sont computationnellement intenses. L'une des méthodes pour diminuer le temps de calcul est de calculer l'espérance sur des petites bases extraites de la base de validation, ce qui pose des problèmes de robustesse.

L'explication fournie par les valeurs de Shapley constitue une excellente base pour comprendre le comportement d'un modèle prédictif. Ces valeurs offrent une explication indépendante du modèle, étayée par des fondements mathématiques solides. Cependant, il est important de souligner que dans un cadre causal, les valeurs de Shapley peuvent potentiellement conduire à des explications erronées, comme présenté dans [Hadj Ali et al., 2023], si des variables explicatives sont possiblement des conséquences de la classe.

Cette thèse s'appuie donc sur les réseaux bayésiens comme outil de classification. Afin de pouvoir justifier de ce choix, outre la classification probabiliste comme classifieur de réseaux bayésiens, ce chapitre présente aussi d'autres méthodes de classification, ainsi que les méthodes principales d'évaluation, et se termine sur des considérations sur l'interprétabilité des classifieurs. Cette présentation fournit donc une première justification théorique : les réseaux bayésiens offrent un bon compromis entre la qualité de la prédiction proposée et la qualité de l'explication de cette prédiction.

Références

Acuña, Edgar et Caroline Rodriguez (2004). "The Treatment of Missing Values and its Effect on Classifier Accuracy". In : *Classification, Clustering, and Data Mining Applications*. Sous la dir. de David Banks, Frederick R. McMorris, Phipps Arabie et Wolfgang Gaul. Studies in Classification, Data Analysis, and Knowledge Organisation. Berlin, Heidelberg : Springer, p. 639-647 (cf. p. 60, 61).

Batista, Gustavo E. A. P. A. et Maria Carolina Monard (2003). "An analysis of four missing data treatment methods for supervised learning". In : *Applied Artificial Intelligence* 17.5. Publisher : Taylor & Francis _eprint : https ://doi.org/10.1080/713827181, p. 519-533 (cf. p. 62).

BISHOP, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Springer New York, NY (cf. p. 41).

CALLE, Juan Esteban de la (2023). *How and Why I Switched from the ROC Curve to the Precision-Recall Curve to Analyze My Imbalanced...* Medium. URL : https://juandelacalle.medium.com/how-and-why-i-switched-from-the-roc-curve-to-the-precision-recall-curve-to-analyze-my-imbalanced-6171da91c6b8 (visité le 4 fév. 2024) (cf. p. 58).

CHEN, Tianqi et Carlos GUESTRIN (2016). "XGBoost : A Scalable Tree Boosting System". In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 785-794. arXiv : 1603.02754[cs] (cf. p. 50).

"Classification and regression trees (CART)" (2012). In : MURPHY, Kevin P. *Machine Learning : A Probabilistic Perspective*. Google-Books-ID : NZP6AQAAQBAJ. MIT Press, p. 544-550 (cf. p. 48).

DAM, Hoa Khanh, Truyen TRAN et Aditya GHOSE (2018). "Explainable Software Analytics". In : (cf. p. 65).

DAVIS, Jesse et Mark GOADRICH (2006). "The relationship between Precision-Recall and ROC curves". In : *Proceedings of the 23rd international conference on Machine learning*. ICML '06. New York, NY, USA : Association for Computing Machinery, p. 233-240 (cf. p. 57).

DIXON, John K. (1979). "Pattern Recognition with Partly Missing Data | IEEE Journals & Magazine | IEEE Xplore". In : *IEEE Transactions on Systems, Man, and Cybernetics (Volume : 9, Issue : 10* (cf. p. 62).

EGAN, James P. (1975). *Signal detection theory and ROC-analysis*. Academic Press series in cognition and perception. OCLC : 1499787. New York : Academic Press. 277 p. (cf. p. 57).

FAYYAD, Usama M. et Keki B. IRANI (1993). "Multi-interval discretization of continuous-valued attributes for classification learning". In : *Ijcai*. T. 93. Issue : 2. Citeseer, p. 1022-1029 (cf. p. 63).

FREUND, Yoav et Robert E SCHAPIRE (1997). "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". In : *Journal of Computer and System Sciences* 55.1, p. 119-139 (cf. p. 50).

FRIEDMAN, Jerome H. (2001). "Greedy function approximation : A gradient boosting machine." In : *The Annals of Statistics* 29.5. Publisher : Institute of Mathematical Statistics, p. 1189-1232 (cf. p. 50).

GARCIA, Salvador, J. LUENGO, José Antonio SÁEZ, Victoria LÓPEZ et F. HERRERA (2013). "A Survey of Discretization Techniques : Taxonomy and Empirical Analysis in Supervised Learning". In : *IEEE Transactions on Knowledge and Data Engineering* 25.4, p. 734-750 (cf. p. 63, 64).

GÉRON, Aurélien (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*. Google-Books-ID : HHetDwAAQBAJ. "O'Reilly Media, Inc." 851 p. (cf. p. 48, 50, 52).

HADJ ALI, Mahdi, Yann Le BIANNIC et Pierre-Henri WUILLEMIN (2023). "Interpreting Predictive Models through Causality : A Query-Driven Methodology". In : *The International FLAIRS Conference Proceedings* 36 (cf. p. 67).

HAND, David J. et Keming YU (2001). "Idiot's Bayes : Not So Stupid after All ?" In : *International Statistical Review / Revue Internationale de Statistique* 69.3. Publisher : [Wiley, International Statistical Institute (ISI)], p. 385-398 (cf. p. 43).

HASTIE, Trevor, Robert TIBSHIRANI et Jerome FRIEDMAN (2009a). "Ensemble Learning". In : *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Sous la dir. de Trevor HASTIE, Robert TIBSHIRANI et Jerome FRIEDMAN. Springer Series in Statistics. New York, NY : Springer, p. 605-624 (cf. p. 49).

– (2009b). "Neural Networks". In : *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Sous la dir. de Trevor HASTIE, Robert TIBSHIRANI et Jerome FRIEDMAN. Springer Series in Statistics. New York, NY : Springer, p. 389-416 (cf. p. 51).

– (2009c). "Random Forests". In : *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Sous la dir. de Trevor HASTIE, Robert TIBSHIRANI et Jerome FRIEDMAN. Springer Series in Statistics. New York, NY : Springer, p. 587-604 (cf. p. 49).

– (2009d). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY : Springer (cf. p. 50).

HECKERMAN, David, Dan GEIGER et David M. CHICKERING (1995). "Learning Bayesian Networks : The Combination of Knowledge and Statistical Data". In : *Machine Learning* 20.3, p. 197-243 (cf. p. 45).

JORDAN, Michael I. et Tom M. MITCHELL (2015). "Machine learning : Trends, perspectives, and prospects". In : *Science* 349.6245. Publisher : American Association for the Advancement of Science, p. 255-260 (cf. p. 46).

KHAN, Muhammad Yaseen, Abdul QAYOOM, Muhammad NIZAMI et al. (2021). "Automated Prediction of Good Dictionary EXamples (GDEX) : A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques". In : *Complexity* (cf. p. 49).

KRUSKAL, Joseph B (1956). "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem". In : *Proceedings of the American Mathematical Society* 7, p. 48-50 (cf. p. 44).

"Logistic regression" (2012). In : MURPHY, Kevin P. *Machine Learning : A Probabilistic Perspective*. Google-Books-ID : NZP6AQAAQBAJ. MIT Press, p. 245-279 (cf. p. 46).

MACK, Christina, Zhaohui SU et Daniel WESTREICH (2018). “Types of Missing Data”. In : *Managing Missing Data in Patient Registries : Addendum to Registries for Evaluating Patient Outcomes : A User’s Guide, Third Edition [Internet]*. Agency for Healthcare Research et Quality (US) (cf. p. 60).

MACQUEEN, J. (1967). “Some methods for classification and analysis of multivariate observations”. In : *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Statistics*. T. 5.1. University of California Press, p. 281-298 (cf. p. 63).

PEDREGOSA, Fabian, Gaël VAROQUAUX, Alexandre GRAMFORT et al. (2011). “Scikit-learn : Machine Learning in Python”. In : *Journal of Machine Learning Research* 12.85, p. 2825-2830 (cf. p. 46).

RASHEED, Khansa, Adnan QAYYUM, Mohammed GHALY et al. (2022). “Explainable, trustworthy, and ethical machine learning for healthcare : A survey”. In : *Computers in Biology and Medicine* 149, p. 106043 (cf. p. 66).

REIMEIR, Benjamin, Steven van ANDEL et Peter FEDEROLF (2021). *How does the postural control system determine whether or not it is "in trouble" ? The role of Distance-to-Boundary and Time-to-Boundary in detecting postural instability* (cf. p. 46).

SHAPLEY, Lloyd S. (1953). “17. A Value for n-Person Games”. In : *17. A Value for n-Person Games*. Princeton University Press, p. 307-318 (cf. p. 66).

TROYANSKAYA, Olga, Mike CANTOR, Gavin SHERLOCK et al. (2001). “Missing Value Estimation Methods for DNA Microarrays”. In : *Bioinformatics* 17, p. 520-525 (cf. p. 61).

VALDES, Gilmer, José Marcio LUNA, Eric EATON et al. (2016). “MediBoost : a Patient Stratification Tool for Interpretable Decision Making in the Era of Precision Medicine”. In : *Scientific Reports* 6.1. Number : 1 Publisher : Nature Publishing Group, p. 37854 (cf. p. 48).

ZHANG, Harry (2004). “The Optimality of Naive Bayes”. In : *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)* (cf. p. 43).

3 *Machine Learning* pour la médecine

Nous avons vu précédemment comment utiliser des données pour créer des modèles de classification, nous allons maintenant essayer de présenter plus concrètement leurs utilisations dans le domaine qui nous intéresse particulièrement : la santé. Après quelques réflexions sur le contexte général, ce chapitre s'applique à présenter les problématiques actuelles dans le domaine de l'apprentissage automatique appliqué à la médecine. Nous nous concentrons ensuite sur l'utilisation du *machine learning* dans le domaine de la gériatrie puis sur l'application des réseaux bayésiens dans le domaine médical.

3.1 Contexte général

Dès les débuts du développement de l'intelligence artificielle au XX^{ème} siècle, le domaine de la santé a été l'objet de travaux significatifs. Une des premières applications pratiques est le projet MYCIN [BUCHANAN et E. SHORTLIFFE, 1984], développé dans les années 1960-1970 et permettant d'identifier à l'aide d'un système expert les bactéries à l'origine de graves infections pour proposer des traitements adaptés. La structure modulaire du système s'avéra pratique et mènera même au développement de modèles graphiques tels que les réseaux bayésiens [HECKERMAN et E. H. SHORTLIFFE, 1992].

L'accroissement général du volume disponible de données de santé et le développement rapide de méthodes de *machine learning* ont rendu possibles de nombreuses récentes applications de l'intelligence artificielle dans le domaine de la santé.

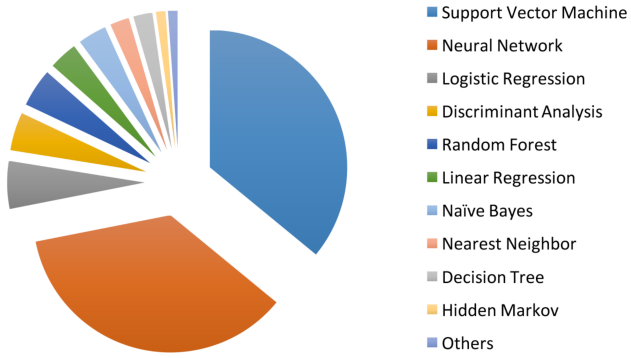


Fig. 3.1. : Les algorithmes de *machine learning* utilisés dans la littérature médicale. Les données sont générées par une recherche sur les algorithmes de *machine learning* dans la santé sur PubMed [F. JIANG et al., 2017].

Le *machine learning* peut être appliqué à divers types de données médicales, qu'elles soient structurées ou non structurées. Les techniques apparemment popu-

laires de méthodes d'apprentissage automatique pour les données structurées (en figure 3.1), sont les machines à vecteurs de support et les réseaux neuronaux, des méthodes donc dites "boites noires". Les réseaux bayésiens n'apparaissent explicitement ici que sous sa forme la plus simple : le *Naïve Bayes*. Pour les données non structurées, le traitement du langage naturel est très utilisé. Les domaines majeurs des maladies qui utilisent des outils d'IA comprennent le cancer, la neurologie et la cardiologie [F. JIANG et al., 2017].

Il existe différents domaines d'application en médecine que l'on résume généralement de cette façon :

- **Diagnostic médical :** diagnostic de diverses pathologies
- **Pronostic et traitement :** prédire l'apparition ou l'évolution des maladies et guider les choix thérapeutiques.
- **Imagerie médicale :** approfondissement des techniques de traitement d'images et de vision par ordinateur pour aider à l'analyse de ces images.
- **Bioinformatique et génomique :** l'analyse de données génomiques et la découverte de médicaments.

C'est principalement le deuxième point qui nous intéresse ici et plus particulièrement le pronostic. En effet, l'implémentation globalisée des dossiers électroniques médicaux (DME) dans les dernières décennies a augmenté drastiquement la disponibilité des données de patients [ADLER-MILSTEIN et al., 2017]. Dans ce contexte, un modèle pronostic est donc un modèle de prédiction multivariable utilisant les données du DME pour estimer la probabilité qu'un patient ait un résultat clinique particulier au cours d'une certaine période à venir.

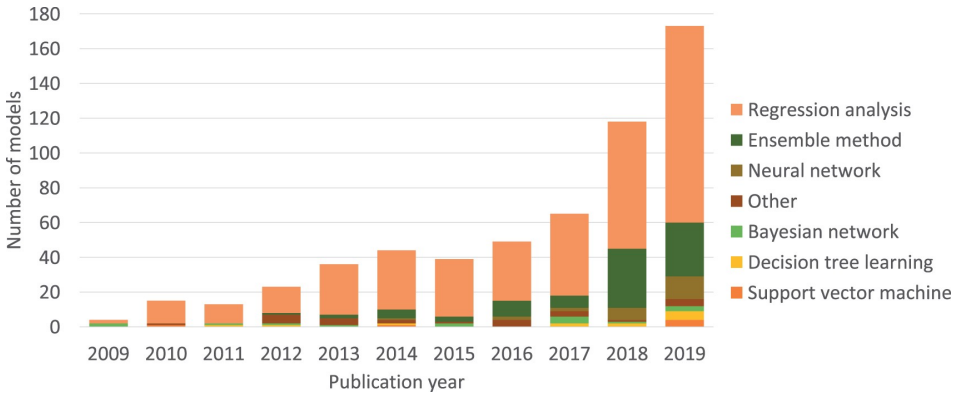


Fig. 3.2. : Tendance dans les méthodes de *machine learning* utilisées pour le développement de modèles de prédiction clinique [YANG et al., 2022].

Une revue de littérature de 2022, incluant tous les articles décrivant le développement d'un ou plusieurs modèles pronostics, nous montre en figure 3.2 qu'effective-

ment, ils sont de plus en plus populaires. On remarque que la méthode largement majoritaire est la simple régression, même si elle est proportionnellement de moins en moins utilisée. Les méthodes de plus en plus présentes sont les méthodes ensemblistes et les réseaux de neurones malgré leur faible explicabilité. Les réseaux bayésiens, eux, représentent un faible pourcentage qui ne semble pas évoluer.

3.2 Défis et limitations

Pourtant, malgré un intérêt global pour les applications en santé depuis de nombreuses années, les modèles de *machine learning* finalement réellement utilisés restent rares. Bien que de nombreuses craintes aient d'abord émergé sur le remplacement des docteurs par des ordinateurs, il est désormais bien admis : l'intelligence artificielle est là uniquement pour aider le soignant, et pas le détrôner. Mais, comme le montre la figure 3.3, la plupart des modèles ne dépassent pas le stade de la validation interne et ne sont donc finalement pas implémentés en usage clinique.

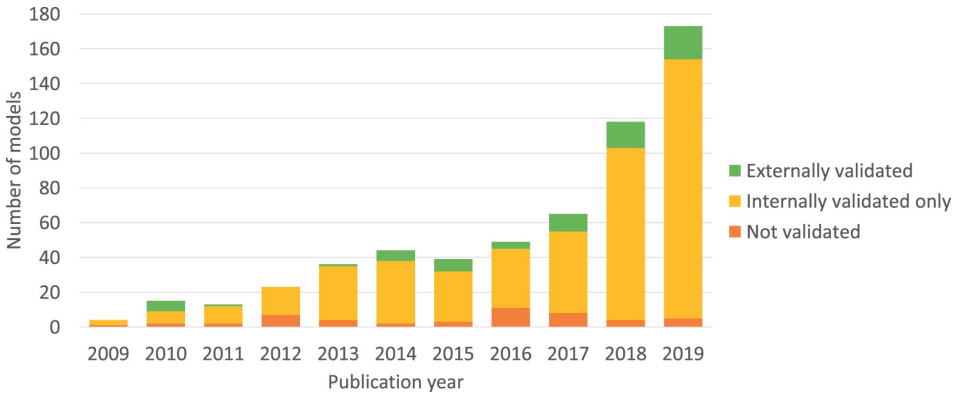


Fig. 3.3. : Tendance dans le type de validation utilisée pour le développement de modèles de prédiction clinique [YANG et al., 2022].

Pour justifier cette faiblesse, les défis dans la construction de modèles prédictifs pour la médecine souvent évoqués sont : les valeurs manquantes, l'équité ("fairness") et la définition claire de l'objectif. Nous verrons ici des pistes d'explications.

3.2.1 Données insuffisantes

"The biggest obstacle to using advanced data analysis isn't skill base or technology ; it's plain old access to the data." - Edd Wilder-James, Harvard Business Review (2016)

Il existe un problème majeur de disponibilité et de qualité des données médicales. Les enjeux d'anonymisation et de protection de la vie privée sont évidemment essentiels, mais de nombreux entrepôts de données de santé émergent actuellement en respectant ces problématiques. En France, notamment, il existe aujourd'hui 22 entrepôts de données de santé hospitaliers dont 17 qui proviennent de CHU et 5 d'autres établissements hospitaliers [DOUTRELIGNE et al., 2023].

Pour la qualité des données, le premier problème est les valeurs manquantes. Les données incomplètes peuvent être dues à de nombreuses causes, notamment à des limitations de coût en ce qui concerne l'équipement, un manque de temps ou de formation du personnel soignant censé remplir les données ou à une évaluation inadéquate des facteurs liés au patient. Les données médicales souffrent aussi beaucoup de la diversité des pratiques et des terminologies entre professionnels et entre pays. Différents pays, ou même des hôpitaux au sein d'un même pays, stockent les données en utilisant des normes de classification différentes, ce qui complique l'intégration et l'agrégation des ensembles de données sur les patients. C'est ainsi que même une base de données avec un très grand nombre de variables bien renseignées ne suffit pas si le nombre de patients est faible ou s'ils sont non représentatifs.

Un exemple médiatisé de données non représentatives qui a eu un impact sur la qualité du modèle est le cas du système Watson d'IBM pour l'oncologie, qui a été formé sur des données de patients avec un très grand nombre de variables pour faire des recommandations de traitement pour huit types de cancer différents. Watson a été formé sur de petits ensembles de données comprenant 106 cas de cancers de l'ovaire et 635 cas de cancers du poumon et s'est appuyé sur des données et des recommandations de traitement provenant d'un seul centre [SANDERS et al., 2019]. Lorsque le système a été utilisé ailleurs, les oncologues signalaient souvent des taux de concordance faibles entre leurs propres recommandations et celles de Watson. Bien que les recommandations que Watson ait assimilées des oncologues du Memorial Sloane Kettering Cancer Center puissent généralement profiter aux New-Yorkais aisés fréquentant cet hôpital, elles pourraient s'avérer inadéquates pour des patients présentant des complexités cliniques différentes [HARISH et al., 2021].

Il n'y a actuellement pas de preuve concluante que le *machine learning* est plus performant que la modélisation statistique pour la prédiction de risque médical. Mais la plupart des modèles de prédiction des risques utilisent des données d'entrée structurées avec un nombre relativement faible de variables. Il en ira différemment pour les modèles de prédiction des risques qui utiliseront des données d'entrée complexes (par exemple, des textes, des images, des données omiques). La qualité

des données des études de *machine learning* doit donc être améliorée pour prouver son intérêt [CHRISTODOULOU et al., 2019].

3.2.2 Éthique, explication et confidentialité

Il existe de nombreux exemples d'algorithmes de *machine learning* qui ont montré des biais discriminatoires dans leur prise de décision. Aux États-Unis, des recherches ont ainsi mis en évidence que les populations afro-américaines étaient plus pénalisées par les décisions de justice [ANGWIN et al., 2016]. Ces mêmes populations sont aussi plus discriminées sur des plateformes en ligne de locations d'appartement [EDELMAN et al., 2017]. Ou encore, des publicités ciblées en ligne sur des offres d'emploi dans les domaines des sciences, de la technologie, de l'ingénierie et des mathématiques seraient moins fréquemment proposées aux femmes qu'aux hommes [LAMBRECHT et TUCKER, 2018]. Cela provient généralement d'un jeu de données lacunaire et de l'utilisation d'un modèle *black box* [CARUANA, 2019].

L'augmentation de la complexité des modèles d'IA, tels que les réseaux neuronaux convolutionnels (CNN) et les architectures d'apprentissage profond, a suscité des inquiétudes concernant leur interprétabilité et leur explicabilité. À mesure que les systèmes d'IA deviennent essentiels aux processus de prise de décision critiques, il devient essentiel de comprendre et de faire confiance au raisonnement derrière leurs résultats. En effet, le manque d'interprétabilité dans la prise de décision basée sur l'IA soulève des préoccupations en matière de confiance et de responsabilité. Il s'agit alors de se concentrer sur des approches visant à renforcer la confiance, améliorer la sécurité des patients et fournir des informations exploitables aux professionnels de la santé [VEERAPPA et RINZIVILLO, 2023].

Mais, ces données sont évidemment sensibles et confidentielles. Il faut donc trouver un équilibre entre la transparence de l'algorithme et la protection des personnes dont les données ont servi à créer le modèle.

Pour maximiser l'inclusivité et minimiser les préjugés, il faut tenir compte des diverses préoccupations et risques en matière de protection de la vie privée : veiller à ce que les données utilisées dans les systèmes d'IA représentent diverses populations (ethnies, sexes, milieux socio-économiques et conditions de santé). Mais, à moins d'avoir accès à un immense jeu de données, cela nécessite de prendre en compte les préoccupations particulières des différents groupes et penser que les groupes minoritaires présentent généralement des risques plus élevés en matière de protection de la vie privée, ce qui entraîne des disparités et peut nécessiter des stratégies spécifiques de gestion des risques.

La collecte de données pour une IA éthique semble pourtant possible. D'après [HOLZINGER, 2021], les principes importants sont la confiance qui est une évaluation subjective, mais qui inclue la sécurité, la fiabilité, l'intégrité, la prévisibilité, la fiabilité et la robustesse, soit produire des résultats fiables même si les données d'entrée sont perturbées. On peut aller créer une IA digne de confiance ("Trustworthy IA") qui garantit la sécurité, la sûreté, la vie privée, la non-discrimination, l'équité, la responsabilité (traçabilité, réplicabilité), l'auditabilité et le bien-être environnemental, et surtout la robustesse et l'explicabilité.

3.2.3 Expliquer ou prédire ?

"Essentially, all models are wrong, but some are useful." - George E.P. Box (1978)

Tous les modèles sont imparfaits, mais certains sont pertinents. Et en général, ils ne sont pertinents que pour une seule chose précise. Il existe une distinction fondamentale entre les modèles causaux et les modèles de prédiction. Lorsque les modèles visent à nous aider à comprendre le monde réel, c'est généralement au prix d'une baisse de performance dans la prédiction. L'extrême pragmatisme caractéristique des modèles de prédiction signifie que leur relation avec le monde réel est obscurcie, mais ils peuvent encore être fonctionnels [HOLZINGER, 2021].

Ainsi, dans ce contexte, les modèles sont plus pertinents pour la prédiction que pour l'explication. Les modèles de prédiction sont conçus pour faire des prévisions précises, tandis que les modèles d'explication sont axés sur la compréhension des mécanismes sous-jacents du monde réel, même s'ils ne sont pas aussi performants en termes de prédiction.

Traditionnellement, les modèles de prédiction se concentrent et sont efficaces sur des résultats isolés. Mais chaque personne sera toujours soumise à des risques multiples, et ces risques ne sont souvent pas indépendants. Les modèles multi-états et les modèles de risques concurrents offrent des approches potentielles pour la prédiction de résultats multiples, ceci est particulièrement utile dans le contexte de la multimorbidité. Mais il en découle plusieurs défis méthodologiques (par exemple, l'explosion combinatoire) [MARTIN, MAMAS et al., 2018 ; MARTIN, SPERRIN et al., 2021].

D'après [HERNÁN et al., 2019], l'identification des patients présentant un mauvais diagnostic est très différente de l'identification du meilleur plan d'action pour la prévention ou le traitement d'une maladie. Pire encore, les algorithmes prédictifs,

lorsqu'ils sont mal utilisés pour l'inférence causale, peuvent conduire à un ajustement incorrect des facteurs de confusion et donc à des décisions erronées. Les modèles de prédiction nous informent que des décisions doivent être prises, mais ils ne peuvent pas nous aider à prendre ces décisions. En revanche, les analyses causales sont conçues pour nous aider à prendre des décisions, car elles s'attaquent aux questions "que se passerait-il si ...".

Un autre risque lié à une approche non causale est illustré par le paradoxe de la prédiction. Si on imagine un groupe de patients G qui présentent un risque élevé pour un événement indésirable si aucune mesure n'est prise. Le modèle de prédiction M est construit à partir de données historiques où aucune mesure n'a été prise. Les patients du groupe G sont correctement classés comme présentant un risque élevé par M. Le modèle de prédiction M est déployé. Les patients du groupe G reçoivent une intervention, évitant ainsi l'événement indésirable, ce que l'on recherche. Mais, si en utilisant de nouvelles données, le modèle M est mis à jour pour devenir M', les patients du groupe G sont maintenant incorrectement classés comme présentant un risque faible par M' (car ils n'ont pas déclaré l'événement indésirable). Ainsi, alors que la mise à jour d'un modèle de prédiction est importante, son déploiement ici est contre-productif.

D'après [Sperrin, Jenkins et al., 2019 ; Lin et al., 2021 ; Sperrin, Diaz-Ordaz et al., 2021], les résultats des modèles de prédiction sont habituellement interprétés comme des prédictions dans un contexte d'intervention, en supposant qu'aucune mesure ne soit prise. On considère aussi que toutes les personnes présentant un risque élevé de mauvais résultats bénéficieront de l'intervention proposée. Aucune de ces deux hypothèses n'est correcte, il n'y a aucune considération de causalité lorsque les modèles de prédiction sont développés avec des méthodes normales d'apprentissage supervisé. Il faut développer des méthodes statistiques pour permettre des prédictions dans le cadre d'interventions hypothétiques ("que se passerait-il si") pour passer ainsi de la prédiction des risques à la prédiction des bénéfices.

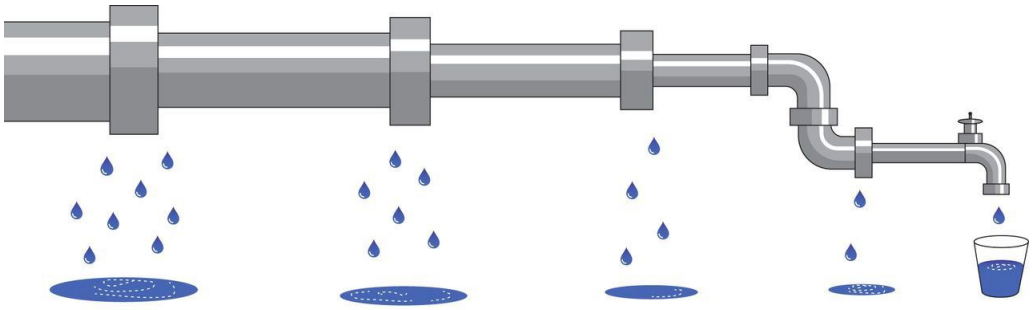
3.2.4 Bonnes pratiques

Même dans le cadre dans lequel un modèle de *machine learning* a réussi à être développé dans un cadre médical, le processus classique, ou "pipeline", a généralement quelques faiblesses. En effet, comme schématisés en figure 3.4, les modèles peuvent être :

- inadaptés à l'objectif,
- sans réelle validation,

- sans mise en utilisation pratique,
- et sans finalement d'adoption.

Pour éviter ces pipelines qui échouent à divers stades de développement et qui ne sont finalement pas utilisés, des bonnes pratiques peuvent être prises en compte.



Not fit for purpose	No validation	No implementation	Not adopted
Developed on wrong patient population	Lack of data or incentive to pursue validation studies	No impact on decision making or patient (health) outcomes	Prediction (perceived as) not useful
Expensive or non-available predictors	Incompletely reported prediction model	No software developed to implement and use the model	Predictions not trusted
Time intensive to use model	Poorly developed or overfitted model	Requirements for adherence to (medical device) regulations	Model not transparent enough, or no tools available to enhance its use in practice
Outcome measured unreliably	Proprietary model code	Cost(-effectiveness) of using proprietary model	Model (perceived as) outdated

Fig. 3.4. : Représentation des "fuites" à chaque étape de création d'un modèle prédictif médical [ROYEN et al., 2022].

Avant d'implémenter un modèle de prédiction dans la pratique clinique, il est important de s'assurer que sa performance de prédiction est généralisable et robuste en validant le modèle de manière externe dans un autre contexte, c'est ce qu'on appelle la validation externe.

Notamment, un modèle clinique a peu de chances d'être accepté s'il n'a pas été prouvé qu'il fonctionne sur des populations disparates. La validation externe peut utiliser [MOONS et al., 2012 ; S. H. PARK et HAN, 2018] :

- des données collectées dans le même hôpital, mais échantillonnées sur une période ultérieure ou antérieure, appelée validation temporelle ;
- des données provenant d'hôpitaux ou de pays différents, où les soins cliniques et les définitions peuvent être différents, appelée validation géographique ;
- des données provenant d'individus démographiquement différents de ceux à partir desquels le modèle a été développé, appelé validation du domaine.

Si nécessaire, le modèle peut être mis à jour en utilisant les connaissances acquises au cours du processus de validation.

La validation externe peut aussi se faire par d'autres chercheurs. Seulement, la définition du problème de prédiction n'est souvent pas clairement exposée, et le modèle final n'est souvent pas complètement décrit, avec peu ou pas d'amélioration au fil du temps. Il est donc important d'améliorer la diffusion des informations nécessaires pour permettre la validation externe par d'autres chercheurs afin d'accroître l'adoption clinique des modèles développés [YANG et al., 2022].

Pour résumer, les principes clés qu'il faut prendre en considération lors de l'utilisation de l'intelligence artificielle dans le domaine de la santé sont [SOLOMONIDES et al., 2022] :

- **l'explicabilité :** fournir une description en termes compréhensibles ;
- **l'interprétabilité :** le raisonnement plausible pour les décisions ;
- **la fiabilité :** la robustesse, sûreté, sécurité et résilience ;
- **"l'auditabilité" :** l'enregistrement de l'activité en vue d'une analyse ultérieure ;
- **la transparence :** le modèle est reconnaissable ou annoncé comme étant de l'IA ;
- **la responsabilité :** le modèle est sous surveillance active, les personnes en charge de la gestion sont définies ;
- **l'équité :** le modèle est sans parti pris et non discriminatoire.

À l'avenir, les soins de santé personnalisés grâce à des modèles de prédiction deviendront la norme plutôt que l'exception. De plus en plus, des informations phénotypiques approfondies provenant de la génétique, de la physiopathologie et de l'imagerie pourront aussi être utilisées pour faire des prédictions. Afin d'obtenir un bénéfice réel pour le patient, il est important de donner la priorité à la généralisation des modèles au delà de son contexte d'apprentissage plutôt que de la considérer comme une réflexion après coup.

3.3 *Machine learning* en gériatrie

Le domaine médical qui nous intéresse plus particulièrement est la gériatrie, soit la médecine des personnes âgées, la population présente en EHPAD. Des revues de littératures [WOODMAN et MANGONI, 2023] se sont intéressées spécifiquement aux applications d'algorithmes de *machine learning* dans la gériatrie et ont montré une

volonté d'éduquer les médecins spécialistes aux méthodes d'IA et aux utilités qu'elles pourraient avoir dans leur pratique.

Les opportunités d'IA en médecine sont particulièrement pertinentes pour la gestion des patients âgés, un groupe qui se caractérise par des schémas complexes de comorbidités et une variabilité interindividuelle significative dans leur fonctionnement et de la réponse au traitement. Les outils cliniques qui utilisent des algorithmes d'apprentissage automatique pour déterminer le choix optimal du traitement obtiennent lentement l'approbation nécessaire des organismes directeurs et sont mis en œuvre dans les soins de santé, avec des implications significatives pour pratiquement toutes les disciplines médicales. Au-delà de l'obtention de l'approbation réglementaire, un élément crucial de la mise en œuvre de ces outils est la confiance et le soutien des personnes qui les utilisent. Dans ce contexte, une meilleure compréhension par les cliniciens de l'intelligence artificielle et des algorithmes d'apprentissage automatique permet d'apprécier les avantages, les risques et les incertitudes possibles, et d'améliorer les chances d'une adoption réussie.

L'adoption récente et généralisée des dossiers médicaux électroniques (DME) et le large soutien du grand public au partage des dossiers médicaux dépersonnalisés [JONES et al., 2022] ont permis d'accéder à un volume considérable de données sur la santé pour la recherche et cela concerne aussi les personnes âgées.

Une revue de littératures d'environ 300 articles sur l'utilisation d'intelligence artificielle pour les soins infirmiers [SEIBERT et al., 2021] a démontré que les hôpitaux constituent le cadre de recherche le plus important, suivis par le soin à domicile. Les maisons de retraite, les soins ambulatoires de longue durée et les soins de santé externes étant moins souvent abordés. L'objectif de ces études revenant le plus souvent était le suivi de l'activité et de la santé (surveillance ou classification) dans 30 % des cas. La coordination des soins et la communication sont des sujets fréquents qui, entre autres, incluent des approches d'IA classant les informations dans la documentation infirmière, soutenant la prise de décision et fournissant des informations pour la coordination et la continuité des soins. La détection et la prévention des chutes, ainsi que la classification des risques de chute, sont aussi des objectifs fréquemment mentionnés pour les sujets liés à l'IA. Contrairement aux chutes, d'autres aspects liés à la mobilité ont suscité moins d'intérêt et n'ont été mentionnés que dans quelques études. Un autre objectif présentant un degré élevé de spécificité est la prédiction et la classification des risques d'escarres (3 % des études).

3.4 Utilisation des réseaux bayésiens dans un contexte médical

Les réseaux bayésiens existent dans le domaine de la médecine et des soins de santé depuis leur création et sont devenus de plus en plus populaires pour traiter les connaissances incertaines impliquées dans l'établissement de diagnostics de maladies, dans la sélection d'alternatives de traitement optimales et dans la prédiction des résultats des traitements dans différents domaines. Les réseaux bayésiens sont aussi de plus en plus développés dans des domaines des soins de santé qui ne sont pas directement liés à la gestion des maladies chez les patients individuels. Par exemple, les réseaux bayésiens sont utilisés en épidémiologie clinique pour la construction de modèles de maladies et en bioinformatique pour l'interprétation des données d'expression génique des microréseaux [P. J. F. Lucas et al., 2004].

L'intérêt que suscitent les réseaux bayésiens dans le domaine des soins de santé peut s'expliquer par leur capacité à modéliser des problèmes complexes avec des dépendances potentiellement causales où un degré significatif d'incertitude est présent, combiner différentes sources d'information telles que des données médicales et les connaissances d'experts, être présentées dans une structure graphique interprétable et modéliser des interventions et raisonner à la fois en termes de diagnostic et de pronostic [Kyrimi, McLachlan, Dube, Neves et al., 2021].

Plusieurs avantages ont été mentionnés dans la littérature. L'utilisation de systèmes basés sur les réseaux bayésiens pour optimiser les processus à tous les stades des soins peut être très bénéfique en ce qui concerne le coût des soins de santé [Dranca et al., 2018 ; Bukhanov et al., 2017 ; Vemulapalli et al., 2016]. Les systèmes basés sur les réseaux bayésiens, contrairement aux approches déterministes, permettent aussi d'avoir confiance dans les prédictions [Haddad et al., 2014 ; X. Jiang et al., 2014], ce qui se traduit par une plus grande flexibilité [Wang et al., 2015]. De plus, il est possible de travailler à partir de cas incomplets et d'interroger n'importe quel nœud du réseau bayésien, ce qui rend les systèmes basés sur le réseau bayésien nettement plus utiles dans la pratique clinique que les modèles construits à partir de variables de résultats spécifiques [Sesen et al., 2014 ; E. Park et al., 2018]. En outre, contrairement aux régressions logistiques et à d'autres techniques de modélisation, les systèmes basés sur les réseaux bayésiens ne sont pas limités aux relations linéaires. Au contraire, ils sont capables de modéliser des relations complexes entre les variables lorsque des conditions de causalité et d'indépendance conditionnelle sont impliquées, ce qui est très utile pour la prise de décision clinique [Wang et al., 2015 ; E. Park et al., 2018 ; Cai et al., 2017]. Enfin, les systèmes basés

sur les réseaux bayésiens ont la réputation d'obtenir des résultats justes [X. JIANG et al., 2014 ; CAILLET et al., 2015 ; SYAFIANDINI et WASITO, 2016].

Mais, la revue de littérature [KYRIMI, MCLACHLAN, DUBE et FENTON, 2020] a montré que la plupart des réseaux bayésiens médicaux publiés sont présentés sans explication sur la façon dont la structure du réseau a été développée et sans justification de la raison pour laquelle elle représente la structure correcte pour l'application médicale donnée. Les réseaux bayésiens dans la santé n'apparaissent pas utilisés à leur plein potentiel, notamment parce que les pipelines génériques de développement de réseaux bayésiens ne sont pas suffisamment répandus. Des limitations existent dans la manière dont les réseaux bayésiens dans la santé sont présentés dans la littérature, ce qui a un impact sur la compréhension, le consensus vers des méthodologies systématiques, la pratique et l'adoption. Bien que la littérature présente de nombreux outils d'aide à la décision médicale basés sur les réseaux bayésiens, aucun n'a *a priori* été adopté dans des soins cliniques de première ligne [KYRIMI, MCLACHLAN, DUBE et FENTON, 2020]. Par ailleurs, il est rare que les auteurs de réseaux bayésiens médicaux publiés fournissent des indications sur la manière dont leur outil peut être intégré facilement dans la routine clinique des soignants [KYRIMI, MCLACHLAN, DUBE, NEVES et al., 2021].

Certains problèmes globaux de l'adoption d'algorithmes de *machine learning* reviennent comme le manque de qualité et la disponibilité des données médicales [SESEN et al., 2014 ; JOCHEMS et al., 2016 ; BUKHANOV et al., 2017]. Quels que soient les efforts déployés pour pallier le manque de données lors de l'élaboration du réseau bayésien à l'aide de l'expertise médicale et/ou de la littérature, la mauvaise qualité et/ou le manque de données médicales restent un problème pour la performance et la facilité d'utilisation du modèle [OJEME et MBOGHO, 2016 ; LOGHMANPOUR et al., 2014].

Une barrière spécifique qui empêche l'adoption des réseaux bayésiens est sa capacité limitée à utiliser des données continues [KYRIMI, DUBE et al., 2021]. En effet, il faut généralement discrétiser les données, ce qui peut être considéré comme ne pas suivre un raisonnement clinique et résulter en une perte d'information [LUO et al., 2017 ; MERLI et al., 2016 ; BERCHIALLA et al., 2014]. Cependant, comme nous en avons discuté dans le chapitre précédent en 2.5, la paramétrisation souffre aussi du même défaut.

De nombreux réseaux bayésiens développés à ce jour pour des applications réelles en biomédecine et en santé ont été construits à la main avec des experts [P. J. LUCAS et al., 2000 ; HECKERMAN, HORVITZ et al., 1992 ; GAAG et al., 2002 ; ANDREASSEN et al., 1999]. Créer manuellement un réseau bayésien exige pourtant un accès direct

à des experts disponibles et demande forcément du temps et de l'implication. Or, dans de nombreux domaines de la biomédecine et des soins de santé, des données ont été collectées et gérées, parfois pendant de nombreuses années. Une telle collecte de données contient généralement des informations très précieuses sur les relations entre les variables discernées, même implicitement. Si un ensemble complet de données est disponible, un réseau bayésien peut bien sûr être appris à partir des données, c'est-à-dire qu'il peut être développé sans accès explicite aux connaissances d'experts humains (voir 1.4).

Ainsi, comme pour les autres méthodes de classification, il est important que les données utilisées aient été collectées en faisant attention aux biais, en quantité suffisante pour repérer les relations probabilistes entre les variables et que les valeurs manquantes aient été imputées de façon pertinente [P. J. F. LUCAS et al., 2004].

Il reste à déterminer si l'apprentissage d'un réseau bayésien de structure plus complexe est plus efficace que celui d'un classificateur bayésien simple. On pourrait s'attendre à ce que plus le graphe d'un réseau bayésien reflète fidèlement les dépendances et les indépendances encodées dans les données, meilleures sont ses performances. Des recherches ont toutefois montré que, lorsqu'ils sont utilisés pour des problèmes de classification, les réseaux bayésiens naïfs ont tendance à être plus performants que les réseaux plus sophistiqués [DOMINGOS et PAZZANI, 1997]. Cette constatation a conduit à suggérer que les structures de réseaux plus complexes ne sont pas rentables. En revanche, [FRIEDMAN et al., 1997] et [CHENG et GREINER, 2013] ont montré que les TAN qui, par rapport aux réseaux bayésiens naïfs, intègrent des dépendances supplémentaires entre leurs variables caractéristiques, sont souvent plus performants que ces réseaux bayésiens naïfs. On peut donc imaginer que le fait de permettre des relations encore plus complexes entre des variables de bonne qualité peut s'avérer encore plus performant. Par ailleurs, les modèles *naive Bayes* et TAN ayant une structure quasiment fixe, ils ne permettent pas d'apprendre la structure entre les variables et de gagner en explicabilité.

Ainsi, l'utilisation des réseaux bayésiens dans le milieu médical est souvent limitée par plusieurs facteurs :

- les réseaux bayésiens sont fréquemment créés à partir des connaissances médicales et non pas automatiquement à partir des données, ce qui peut s'expliquer par un manque de bases de données pertinentes et de bonne qualité ;
- dans les cas (rares) d'apprentissage automatique, les méthodes d'apprentissage sont souvent des structures simples, comme les *naive Bayes* et TAN, probablement faute d'implémentation facilement accessible et interprétable pour la

communauté médicale ;

- il y a peu de validation en conditions réelles avec intégration dans la routine des soignants, peut-être à cause d'un manque de dispositifs compatibles.

Dans cette première partie regroupant les trois chapitres de notre état de l'art, nous avons justifié notre choix des réseaux bayésiens comme modèle de classification privilégié pour ce contexte médical. Nous avons introduit tout d'abord les réseaux bayésiens, et notamment leurs spécificités et leurs méthodes d'apprentissage. Ensuite, nous avons abordé comment les utiliser pour faire de la classification probabiliste, tout en présentant d'autres méthodes de classification qui nous permettront de nous comparer. Enfin, nous avons dressé un état des lieux de l'utilisation de ces méthodes dans le domaine spécifique de la santé. Cette première partie fournit ainsi tous les concepts nécessaires à la compréhension de la deuxième partie de notre travail. De plus, elle expose les bonnes pratiques sur lesquelles nous nous sommes appuyés pour construire nos propres modèles de classification, que nous détaillerons par la suite.

Références

ADLER-MILSTEIN, Julia, A Jay HOLMGREN, Peter KRALOVEC et al. (2017). "Electronic health record adoption in US hospitals : the emergence of a digital "advanced use" divide". In : *Journal of the American Medical Informatics Association* 24.6, p. 1142-1148 (cf. p. 73).

ANDREASSEN, Steen, Christian RIEKEHR, Brian KRISTENSEN, Henrik C. SCHØNHEYDER et Leonard LEIBOVICI (1999). "Using probabilistic and decision–theoretic methods in treatment and prognosis modeling". In : *Artificial Intelligence in Medicine*. Prognostic Models in Medicine 15.2, p. 121-134 (cf. p. 83).

ANGWIN, Julia, Jeff LARSON, Surya MATTU et Lauren KIRCHNER (2016). *Machine Bias*. ProPublica. URL : `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing` (cf. p. 76).

BERCHIALLA, Paola, Ezio Nicola GANGEMI, Francesca FOLTRAN et al. (2014). "Predicting severity of pathological scarring due to burn injuries : a clinical decision making tool using Bayesian networks". In : *International Wound Journal* 11.3, p. 246-252 (cf. p. 83).

BUCHANAN, Bruce et Edward SHORTLIFFE (1984). *Rule-based Expert System – The MYCIN Experiments of the Stanford Heuristic Programming Project*. Journal Abbreviation : SERBIULA (sistema Librum 2.0) Publication Title : SERBIULA (sistema Librum 2.0) (cf. p. 72).

BUKHANOV, Nikita, Marina BALAKHONTCEVA, Sergey KOVALCHUK, Nadezhda ZVARTAU et Aleksandra KONRADI (2017). “Multiscale modeling of comorbidity relations in hypertensive outpatients”. In : *Procedia Computer Science*. CENTERIS 2017 - International Conference on ENTERprise Information Systems / ProjMAN 2017 - International Conference on Project MANagement / HCist 2017 - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2017 121, p. 446-450 (cf. p. 82, 83).

CAI, Zhi-qiang, Peng GUO, Shu-bin SI et al. (2017). “Analysis of prognostic factors for survival after surgery for gallbladder cancer based on a Bayesian network”. In : *Scientific Reports* 7.1. Number : 1 Publisher : Nature Publishing Group, p. 293 (cf. p. 82).

CAILLET, Pascal, Sarah KLEMM, Michel DUCHER, Alexandre AUSSEM et Anne-Marie SCHOTT (2015). “Hip Fracture in the Elderly : A Re-Analysis of the EPIDOS Study with Causal Bayesian Networks”. In : *PLOS ONE* 10.3. Publisher : Public Library of Science, e0120125 (cf. p. 83).

CARUANA, Richard (2019). “Friends Don’t Let Friends Deploy Black-Box Models : The Importance of Intelligibility in Machine Learning”. In : *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD ’19. New York, NY, USA : Association for Computing Machinery, p. 3174 (cf. p. 76).

CHENG, Jie et Russell GREINER (2013). *Comparing Bayesian Network Classifiers*. arXiv : 1301.6684[cs,stat] (cf. p. 84).

CHRISTODOULOU, Evangelia, Jie MA, Gary S. COLLINS et al. (2019). “A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models”. In : *Journal of Clinical Epidemiology* 110, p. 12-22 (cf. p. 76).

DOMINGOS, Pedro et Michael PAZZANI (1997). “On the Optimality of the Simple Bayesian Classifier under Zero-One Loss”. In : *Machine Learning*. Kluwer Academic Publishers 29, p. 103-130 (cf. p. 84).

DOUTRELIGNE, Matthieu, Adeline DEGREMONT, Pierre-Alain JACHIET, Antoine LAMER et Xavier TANNIER (2023). “Good practices for clinical data warehouse implementation : A case study in France”. In : *PLOS Digital Health* 2.7. Publisher : Public Library of Science, e0000298 (cf. p. 75).

DRANCA, Lacramioara, Lopez de ABETXUKO RUIZ DE MENDAROZKETA, Alfredo GOÑI et al. (2018). “Using Kinect to classify Parkinson’s disease stages related to severity of gait impairment”. In : *BMC Bioinformatics* 19.1, p. 471 (cf. p. 82).

EDELMAN, Benjamin, Michael LUCA et Dan SVIRSKY (2017). “Racial Discrimination in the Sharing Economy : Evidence from a Field Experiment”. In : *American Economic Journal : Applied Economics* 9.2, p. 1-22 (cf. p. 76).

FRIEDMAN, Nir, Dan GEIGER et Moises GOLDSZMIDT (1997). “Bayesian Network Classifiers”. In : *Machine Learning* 29.2, p. 131-163 (cf. p. 84).

GAAG, L. C. van der, S. RENOOIJ, C. L. M. WITTEMAN, B. M. P. ALEMAN et B. G. TAAL (2002). "Probabilities for a probabilistic network : a case study in oesophageal cancer". In : *Artificial Intelligence in Medicine* 25.2, p. 123-148 (cf. p. 83).

HADDAD, Tarek, Adam HIMES et Michael CAMPBELL (2014). "Fracture prediction of cardiac lead medical devices using Bayesian networks". In : *Reliability Engineering & System Safety* 123, p. 145-157 (cf. p. 82).

HARISH, Vinyas, Felipe MORGADO, Ariel D. STERN et Sunit DAS (2021). "Artificial Intelligence and Clinical Decision Making : The New Nature of Medical Uncertainty". In : *Academic Medicine : Journal of the Association of American Medical Colleges* 96.1, p. 31-36 (cf. p. 75).

HECKERMAN, David E., Eric J. HORVITZ et Bharat N. NATHWANI (1992). "Toward normative expert systems : Part I. The Pathfinder project". In : *Methods of Information in Medicine* 31.2, p. 90-105 (cf. p. 83).

HECKERMAN, David E. et Edward H. SHORTLIFFE (1992). "From certainty factors to belief networks". In : *Artificial Intelligence in Medicine* 4.1, p. 35-52 (cf. p. 72).

HERNÁN, Miguel A., John HSU et Brian HEALY (2019). "A Second Chance to Get Causal Inference Right : A Classification of Data Science Tasks". In : *CHANCE* 32.1, p. 42-49 (cf. p. 77).

HOLZINGER, Andreas (2021). "The Next Frontier : AI We Can Really Trust". In : *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Sous la dir. de Michael KAMP, Irena KOPRINSKA, Adrien BIBAL et al. Communications in Computer and Information Science. Cham : Springer International Publishing, p. 427-440 (cf. p. 77).

JIANG, Fei, Yong JIANG, Hui ZHI et al. (2017). "Artificial intelligence in healthcare : past, present and future". In : *Stroke and Vascular Neurology* 2.4, p. 230-243 (cf. p. 72, 73).

JIANG, Xia, Diyang XUE, Adam BRUFSKY, Seema KHAN et Richard NEAPOLITAN (2014). "A New Method for Predicting Patient Survivorship Using Efficient Bayesian Network Learning". In : *Cancer Informatics* 13. Publisher : SAGE Publications Ltd STM, CIN.S13053 (cf. p. 82, 83).

JOCHEMS, Arthur, Timo M. DEIST, Johan van SOEST et al. (2016). "Distributed learning : Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept". In : *Radiotherapy and Oncology* 121.3. Publisher : Elsevier, p. 459-467 (cf. p. 83).

JONES, Linda A., Jenny R. NELDER, Joseph M. FRYER et al. (2022). "Public opinion on sharing data from health services for clinical and research purposes without explicit consent : an anonymous online survey in the UK". In : *BMJ Open* 12.4. Publisher : British Medical Journal Publishing Group Section : Health policy, e057579 (cf. p. 81).

KYRIMI, Evangelia, Kudakwashe DUBE, Norman FENTON et al. (2021). "Bayesian networks in healthcare : What is preventing their adoption ?" In : *Artificial Intelligence in Medicine* 116, p. 102079 (cf. p. 83).

KYRIMI, Evangelia, Scott MCLACHLAN, Kudakwashe DUBE et Norman FENTON (2020). *Bayesian Networks in Healthcare : the chasm between research enthusiasm and clinical adoption*. Pages : 2020.06.04.20122911 (cf. p. 83).

KYRIMI, Evangelia, Scott MCLACHLAN, Kudakwashe DUBE, Mariana R. NEVES et al. (2021). "A comprehensive scoping review of Bayesian networks in healthcare : Past, present and future". In : *Artificial Intelligence in Medicine* 117, p. 102108 (cf. p. 82, 83).

LAMBRECHT, Anja et Catherine E. TUCKER (2018). *Algorithmic Bias ? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads*. Rochester, NY (cf. p. 76).

LIN, Lijing, Matthew SPERRIN, David A. JENKINS, Glen P. MARTIN et Niels PEEK (2021). "A scoping review of causal methods enabling predictions under hypothetical interventions". In : *Diagnostic and Prognostic Research* 5.1, p. 3 (cf. p. 78).

LOGHMANPOUR, Natasha A., Marek J. DRUZDZEL et James F. ANTAKI (2014). "Cardiac Health Risk Stratification System (CHRiSS) : A Bayesian-Based Decision Support System for Left Ventricular Assist Device (LVAD) Therapy". In : *PLOS ONE* 9.11. Publisher : Public Library of Science, e111264 (cf. p. 83).

LUCAS, Peter J., Nicolette C. de BRUIJN, Karin SCHURINK et Andy HOEPELMAN (2000). "A probabilistic and decision-theoretic approach to the management of infectious disease at the ICU". In : *Artificial Intelligence in Medicine* 19.3, p. 251-279 (cf. p. 83).

LUCAS, Peter J. F., Linda C. VAN DER GAAG et Ameen ABU-HANNA (2004). "Bayesian networks in biomedicine and health-care". In : *Artificial Intelligence in Medicine*. Bayesian Networks in Biomedicine and Health-Care 30.3, p. 201-214 (cf. p. 82, 84).

LUO, Yi, Issam El NAQA, Daniel L. MCSHAN et al. (2017). "Unraveling biophysical interactions of radiation pneumonitis in non-small-cell lung cancer via Bayesian network analysis". In : *Radiotherapy and Oncology* 123.1. Publisher : Elsevier, p. 85-92 (cf. p. 83).

MARTIN, Glen P., Mamas A. MAMAS, Niels PEEK, Iain BUCHAN et Matthew SPERRIN (2018). "A multiple-model generalisation of updating clinical prediction models". In : *Statistics in Medicine* 37.8, p. 1343-1358 (cf. p. 77).

MARTIN, Glen P., Matthew SPERRIN, Kym I. E. SNELL, Iain BUCHAN et Richard D. RILEY (2021). "Clinical prediction models to predict the risk of multiple binary outcomes : a comparison of approaches". In : *Statistics in Medicine* 40.2, p. 498-517 (cf. p. 77).

MERLI, Mauro, Marco MOSCATELLI, Giorgia MARIOTTI et al. (2016). "A minimally invasive technique for lateral maxillary sinus floor elevation : a Bayesian network study". In : *Clinical Oral Implants Research* 27.3, p. 273-281 (cf. p. 83).

MOONS, Karel G. M., Andre Pascal KENGNE, Diederick E. GROBBEE et al. (2012). "Risk prediction models : II. External validation, model updating, and impact assessment". In : *Heart* 98.9. Publisher : BMJ Publishing Group Ltd and British Cardiovascular Society Section : Review, p. 691-698 (cf. p. 79).

OJEME, Blessing et Audrey MBOGHO (2016). “Selecting Learning Algorithms for Simultaneous Identification of Depression and Comorbid Disorders”. In : *Procedia Computer Science*. Knowledge-Based and Intelligent Information & Engineering Systems : Proceedings of the 20th International Conference KES-2016 96, p. 1294-1303 (cf. p. 83).

PARK, Eunjeong, Hyuk-jae CHANG et Hyo Suk NAM (2018). “A Bayesian Network Model for Predicting Post-stroke Outcomes With Available Risk Factors”. In : *Frontiers in Neurology* 9 (cf. p. 82).

PARK, Seong Ho et Kyunghwa HAN (2018). “Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction”. In : *Radiology* 286.3. Publisher : Radiological Society of North America, p. 800-809 (cf. p. 79).

ROYEN, Florien S. van, Karel G. M. MOONS, Geert-Jan GEERSING et Maarten van SMEDEN (2022). “Developing, validating, updating and judging the impact of prognostic models for respiratory diseases”. In : *European Respiratory Journal* 60.3. Publisher : European Respiratory Society Section : ERJ Methods (cf. p. 79).

SANDERS, Samantha F., Mats TERWIESCH, William J. GORDON et Ariel Dora STERN (2019). “How Artificial Intelligence Is Changing Health Care Delivery”. In : *NEJM Catalyst* (cf. p. 75).

SEIBERT, Kathrin, Dominik DOMHOFF, Dominik BRUCH et al. (2021). “Application Scenarios for Artificial Intelligence in Nursing Care : Rapid Review”. In : *Journal of Medical Internet Research* 23.11, e26522 (cf. p. 81).

SESEN, M. Berkan, Michael D. PEAKE, Rene BANARES-ALCANTARA et al. (2014). “Lung Cancer Assistant : a hybrid clinical decision support application for lung cancer care”. In : *Journal of The Royal Society Interface* 11.98. Publisher : Royal Society, p. 20140534 (cf. p. 82, 83).

SOLOMONIDES, Anthony E, Eileen KOSKI, Shireen M ATABAKI et al. (2022). “Defining AMIA’s artificial intelligence principles”. In : *Journal of the American Medical Informatics Association* 29.4, p. 585-591 (cf. p. 80).

SPERRIN, Matthew, Karla DIAZ-ORDAZ et Romin PAJOUHESHNIA (2021). “Invited Commentary : Treatment Drop-in-Making the Case for Causal Prediction”. In : *American Journal of Epidemiology* 190.10, p. 2015-2018 (cf. p. 78).

SPERRIN, Matthew, David JENKINS, Glen P. MARTIN et Niels PEEK (2019). “Explicit causal reasoning is needed to prevent prognostic models being victims of their own success”. In : *Journal of the American Medical Informatics Association : JAMIA* 26.12, p. 1675-1676 (cf. p. 78).

SYAFIANDINI, Arida Ferti et Ito WASITO (2016). “Metastasis identification based on clinical parameters using Bayesian network”. In : *2016 4th International Conference on Information and Communication Technology (ICoICT)*. 2016 4th International Conference on Information and Communication Technology (ICoICT), p. 1-6 (cf. p. 83).

VEERAPPA, Manjunatha et Salvo RINZIVILLO (2023). *Explainable AI - Introduction to the Special Theme*. URL : https://ercim-news.ercim.eu/en134/special/explainable-ai-introduction-to-the-special-theme (cf. p. 76).

VEMULAPALLI, Vijetha, Jiaqi QU, Jeonifer M. GARREN et al. (2016). "Non-obvious correlations to disease management unraveled by Bayesian artificial intelligence analyses of CMS data". In : *Artificial Intelligence in Medicine* 74, p. 1-8 (cf. p. 82).

WANG, Zhao, Michael W. JENKINS, George C. LINDERMAN et al. (2015). "3-D Stent Detection in Intravascular OCT Using a Bayesian Network and Graph Search". In : *IEEE Transactions on Medical Imaging* 34.7. Conference Name : IEEE Transactions on Medical Imaging, p. 1549-1561 (cf. p. 82).

WOODMAN, Richard J. et Arduino A. MANGONI (2023). "A comprehensive review of machine learning algorithms and their application in geriatric medicine : present and future". In : *Aging Clinical and Experimental Research* 35.11, p. 2363-2397 (cf. p. 80).

YANG, Cynthia, Jan A. KORS, Solomon IOANNOU et al. (2022). "Trends in the conduct and reporting of clinical prediction model development and validation : a systematic review". In : *Journal of the American Medical Informatics Association : JAMIA* 29.5, p. 983-989 (cf. p. 73, 74, 80).

Partie II

Contributions des réseaux bayésiens à la prédiction d'événements de santé en EHPAD

La base de données

4

Nous allons maintenant aborder nos contributions sur les réseaux bayésiens en tant qu'outil de classification pour la prédiction en EHPAD. Tout d'abord, nous présenterons la base de données que nous avons eue la chance de pouvoir utiliser. Ensuite, nous consacrerons un chapitre aux trois événements de santé défavorables et évitables auxquels nous nous sommes intéressés en raison de leur prévalence en EHPAD : l'escarre, l'hospitalisation en urgence et la fracture.

Dans ce chapitre, nous exposerons plus spécifiquement le contexte d'application de nos contributions, ainsi que la réglementation qui nous a permis d'effectuer ces travaux. Enfin, nous détaillerons l'ensemble du prétraitement des données qui est commun à nos trois cibles médicales.

4.1 Contexte

4.1.1 Établissement d'Hébergement pour Personnes Âgées Dépendantes

L'EHPAD, établissement d'hébergement pour personnes âgées dépendantes, constitue une option à laquelle a recours une population nombreuse et croissante, lorsque pour diverses raisons, et notamment de santé, il ne leur est plus possible de vivre à leur domicile. La moyenne d'âge des résidents aujourd'hui est de 87 ans et la durée moyenne de séjour de 2.5 ans [Belmin et al., 2016]. Ces établissements hébergent des personnes qui ne peuvent plus être prises en charge à leur domicile et leur assurent une aide pour les gestes de la vie quotidienne.

Cette population est fortement touchée par les maladies chroniques et se trouve très exposée aux syndromes gériatriques et aux situations complexes parmi lesquels les maladies neurocognitives, la dénutrition, la dépression, les chutes, les fractures, la polymédication. Aussi, les soins médicaux représentent un aspect très important de la mission des EHPAD et ils occupent une part croissante dans l'activité des professionnels de ces établissements.

Dans les EHPAD, le patient est pris en charge par une équipe soignante pluridisciplinaire constituée d'infirmiers et d'aides-soignants qui réalisent les soins quotidiens (pansements, mesures de la glycémie, distribution des médicaments, etc.). Le médecin coordonnateur élabore le projet général de soins, et en coordonne la mise en œuvre au sein de l'EHPAD. Il arrive souvent que le médecin coordonnateur gère

plusieurs EHPAD en même temps, et il est peu présent physiquement au sein de l'établissement.

Avec le développement des nouvelles technologies informatiques dans le domaine de la santé, un nombre croissant d'établissements de santé sont équipés de systèmes d'information regroupant les données administratives et médicales des patients ainsi que des informations sur les soins qui leur sont prodigués. Ces systèmes, et notamment les dossiers patients informatisés, permettent un accès facile et rapide aux informations des patients, notamment dans un but d'amélioration de la qualité et de la sécurité des soins.

4.1.2 NETSoins

NETSoins est une solution logicielle complète sur l'accompagnement des résidents en EHPAD, éditée par Teranga Software depuis 2007 et utilisée par plus de 4 000 établissements en France. L'ensemble des fonctionnalités métiers pour l'accompagnement des résidents est disponible dans le dossier informatisé NETSoins. Par exemple :

- **les soins :** avec les plans de soins et les projets personnels du résident
- **le médical :** avec les observations des médecins, les relevés et les prescriptions
- **l'administratif :** avec les mouvements du résident (hospitalisations, rendez-vous, etc)
- **l'alimentation :** avec les habitudes et le suivi alimentaire
- **le projet personnalisé :** avec les activités et la vie sociale du résident
- **le paramédical :** avec les comptes-rendus des professionnels de santé et les évaluations
- **les interfaces :** avec une liaison possible avec les laboratoires, les pharmacies et le dossier médical partagé (DMP) ainsi que la télémédecine.

Le logiciel est utilisé et alimenté par le personnel de l'établissement : les aides soignantes, infirmiers, médecins, mais aussi les animateurs, les pharmaciens, le personnel administratif et paramédical, ainsi que d'autres intervenants.

NETSoins est accessible depuis un simple navigateur web. Il respecte le règlement général sur la protection des données (RGPD) et toutes les données sont centralisées et sécurisées dans des environnements d'HDS (hébergeurs de données de santé) certifié. Toutes les informations du logiciel NETSoins sont stockées dans une base

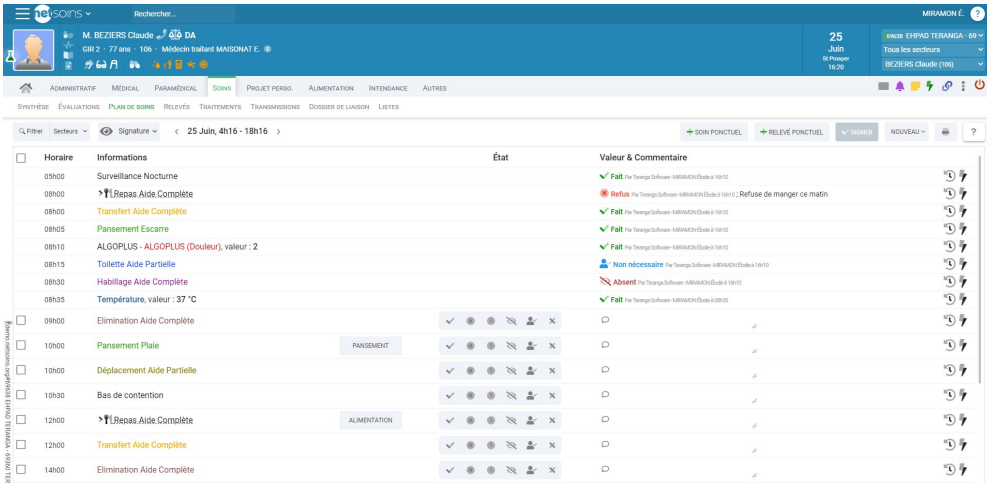


Fig. 4.1. : Capture d'écran du plan de soins d'un résident fictif du logiciel NETSoins [*Teranga Sofware - Prendre soin de ceux qui prennent soin des autres* 2023].

de données relationnelle en PostgreSQL. Elle contient plus de 300 tables avec en moyenne 13 colonnes par table. Les liens entre les tables sont multiples et complexes.

4.2 Réglementations et anonymisation

Travailler avec des données de santé individuelles nécessite évidemment de nombreuses précautions. La CNIL (commission nationale de l'informatique et des libertés), l'autorité compétente en la matière de protection des données personnelles en France, recommande d'intégrer dès la conception d'un système d'intelligence artificielle, les principes de protection des données personnelles (*privacy by design*) [CNIL, 2023].

4.2.1 Anonymisation et pseudonymisation

Il faut distinguer anonymisation et pseudonymisation. D'après la CNIL, l'anonymisation est le "traitement de données à caractère personnel qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la *personne* concernée par quelque moyen que ce soit et de manière irréversible". Pour cela, trois critères sont définis :

- **impossibilité d'individualisation :** il ne doit pas être possible d'isoler un individu dans le jeu de données ;

- **impossibilité de corrélation :** il ne doit pas être possible de relier entre eux des ensembles de données distincts concernant un même individu ;
- **impossibilité d'inférence :** il ne doit pas être possible de déduire, de façon quasi certaine, de nouvelles informations sur un individu.

La pseudonymisation, elle, est un traitement de données personnelles consistant à remplacer les données directement identifiantes (nom, prénom, etc.) d'un jeu de données par des données indirectement identifiantes (alias, numéro séquentiel, etc.), l'opération étant réversible [IA, 2023].

4.2.2 Mise en conformité

En pratique, les trois critères d'anonymisation sont difficiles à prouver, mais nous avons mis en place un certain nombre de mesures pour améliorer le respect de ces règles. Un long travail a été effectué avec des avocats spécialisés et les conditions générales de ventes du logiciel NETSoins ont dû être changées pour mentionner l'utilisation possible de données à des fins de recherche. Les établissements sont libres de s'y opposer dans leur entièreté, et à l'échelle individuelle aussi, chaque résident (ou proche aidant) peut refuser l'exploitation de ses données. Ces refus ou acceptations sont à présent tracés dans la base de données. Des protocoles stricts ont été conçus :

- Les données ne contiennent pas d'information non anonyme (nom, date de naissance, lieu de l'EHPAD, etc.) ni de champs libres de textes.
- Aucune date n'est extraite. Nous détaillerons ceci dans la partie 4.3.2.
- Le nom des variables est encodé différemment à chaque extraction.
- Aucune clé d'identification n'est nécessaire pour l'étude.

Ainsi, aucune information permettant l'identification d'une personne précise n'est récoltée. Par ailleurs, le traitement des données est effectué sur une machine virtuelle mise à disposition par un hébergeur certifié HDS.

Pour être sûr de n'exporter que des données anonymes et d'intérêts, chaque information présente dans la base a dû être passée en revue. Un tri d'abord des tables dans leur globalité a été effectué, pour écarter, par exemple, les données en lien avec la facturation ou la maintenance. Puis un tri interne aux tables avec l'élimination des noms, adresses, numéros, etc. a été fait. Les champs en textes libres ont été écartés dans la majorité des cas, car nous n'avons pour l'instant aucun moyen d'être sûr que le nom du patient ou autre élément identifiant n'est pas employé dans un commentaire (par exemple : "Mr X avait mal à la tête, je lui ai donné un

Doliprane"). Cela peut poser un problème, car si l'infirmière a bien noté l'information dans une transmission (ce sont les notes anecdotiques que les soignants écrivent sur des patients pour la traçabilité), il n'est pas certain que cette information soit aussi saisie dans le plan de soins du résident, qui lui peut être récupéré anonymement. Il est ainsi possible de passer à côté d'informations précieuses.

Toutes les données sont stockées dans des serveurs sécurisés, gérés par d'autres entreprises que Teranga. Chaque base de production réunit plusieurs établissements par groupe ou par date de souscription au logiciel. D'une base à l'autre, la structure est toujours la même, mais les habitudes des utilisateurs diffèrent et la qualité de saisie aussi. Chaque établissement peut paramétrer le logiciel, par exemple en imposant de remplir la date de naissance en créant un résident ou d'utiliser les appellations de la base de données Claude Bernard [BCB, 2019] pour déclarer une pathologie. Certains établissements ne le font pas, car ils préfèrent être libres dans leurs actions. Même si cela fournirait des données beaucoup plus homogènes, c'est aussi une stratégie de l'entreprise de laisser ce choix à ses clients. En effet, ils préfèrent éviter que les clients ne saisissent des informations fausses ou approximatives, car ils n'auraient pas trouvé assez rapidement ce qu'ils cherchaient.

4.3 Pipeline de prétraitement

Le but de ce projet a été d'utiliser les données du logiciel NETSoins, pour implémenter des classifieurs capables de prédire plusieurs événements de santé défavorables qui sont potentiellement modifiables par des interventions de santé appropriées. Dans le cadre d'un modèle de *machine learning* où il y a beaucoup de données à manipuler, il est courant d'utiliser une "pipeline". Il s'agit d'une séquence de différents traitements de données qui vise à transformer les données dans un format apte à l'apprentissage.

4.3.1 Inclusion

Les résidents inclus dans l'étude sont uniquement des résidents en EHPAD (pas d'établissement handicap, ni d'établissement test) qui ont été en hébergement permanent entre 2012 et 2022 inclus et qui ont au moins un relevé de poids saisi. En effet, nous avons remarqué que les dossiers résidents qui ne contiennent aucun poids étaient très peu ou mal remplis. Nous avons ainsi eu accès aux données de 168 666 résidents à ce stade. Les informations extraites sont les suivantes :

Type	Variables
Informations générales	sexe
	âge
	taille
	malentendant
	malvoyant
	fumeur
	consommation d'alcool
Relevés	poids
	tensions
	albuminémies
Évaluations	AGGIR
	PATHOS
	Braden
	MMSE
	MNA
	MNA simplifié
Mouvements administratifs	sortie de l'établissement définitive
	décès
	hospitalisations
Pathologies dans le dossier résident	incontinence
	dépression
	dénutrition
	diabète
	fracture
	escarre
	asthme
	BCPO
	ostéoporose
	hyperthyroïdie
Médicaments prescrits	pour la dépression
	pour le diabète
	psycholeptiques
	pour les infections
	pour l'épilepsie
	pour le Parkinson
	pour la démence
	pour le cholestérol
	pour l'ostéoporose
	pour la psychose
	œstrogène
	corticoïdes
	stéroïdes
	anti-inflammatoire non stéroïdien
	pour l'hyperthyroïdie
	pour l'hypotension
	pour la BCPO
	morphine
	perfusion
Événements ponctuels	participation à des activités
	chutes dans l'établissement
Plan de soins	Aide déplacement
	Pansements escarre

Tab. 4.1. : Variables extraites triées par type.

Nous détaillerons ensuite comment nous avons utilisé chaque donnée et leur objectif.

4.3.2 Temporalité

Nous sommes partis d'une base de données sous forme événementielle, où chaque information ajoutée sur un résident sera stockée sous forme d'événement ou d'occurrence. Des données supplémentaires peuvent aussi être incluses, telles que des détails sur l'événement, des mesures ou des résultats associés. À partir de cela, nous avons dû reconstruire la meilleure "photographie" d'un résident à un instant t malgré des données périodiques et intemporelles. Nous avons alors cherché à obtenir des données discrétisées en format tabulaire, tout en gardant un sens médical.

Chaque saisie dans NETSoins est enregistrée avec un *timestamp* (horodatage), donc avec une date et une heure. Puisque nous ne pouvions avoir accès à des dates précises pour des raisons d'anonymisations, ces *timestamps* ont été soustraits à la date d'admission du résident. Ainsi, pour se repérer dans le temps, seul le délai en seconde après l'admission est utilisé. Les résidents ont donc tous le même point de départ.

Exporter toutes les données saisies sur les variables sélectionnées n'était pas envisageable en termes de volume, mais aussi en termes de pertinence. Nous avons donc créé des marqueurs temporels, tels que l'admission, une semaine après, un mois après ou encore la dernière valeur saisie. Pour les variables binaires de type "présence/absence", la dénutrition par exemple, nous vérifions à chaque marqueur temporel si la pathologie a été ajoutée dans le dossier médical.

Pour les variables qui nécessitaient plus de réflexion sur leur prétraitement tels que les relevés, les différentes cibles de classifications et les événements ponctuels, nous les avons exportées sous forme de dictionnaires avec l'événement, sa valeur si elle a lieu et son délai tout au long du séjour.

Une temporalité a été définie pour chaque résident et chaque variable a été créée en fonction. Les marqueurs temporels, ainsi que les sélections des temporalités, sont différents selon les cibles et seront ainsi exposés en 5.3, 6.3 et 7.3.

4.3.3 Transformation des variables

Il faut noter que le logiciel est utilisé par différents types de personnel de l'établissement : les aides soignantes, infirmiers, médecins, mais aussi les animateurs, les pharmaciens, le personnel administratif et paramédical, etc. La saisie des données concernant un résident en EHPAD est donc effectuée par des personnes distinctes à des moments différents plus ou moins périodiques. L'exploitation des données temporelles est alors difficile et différentes transformations sont effectuées sur les variables selon le type.

Niveau de dépendance

Dans les EHPAD, la perte d'autonomie des personnes âgées est évaluée par la grille AGGIR (Autonomie Gérontologie Groupe Iso-Ressources). Cette évaluation vise à déterminer le degré de dépendance d'une personne âgée en fonction de différents critères.

La grille AGGIR est basée sur plusieurs dimensions de l'autonomie [LÉGIFRANCE, 2017] :

- **Cohérence :** Évaluation de la capacité à converser ou se comporter de façon sensée.
- **Orientation :** Évaluation de la capacité à se repérer dans le temps et l'espace.
- **Toilette :** Évaluation de la capacité de la personne à réaliser sa toilette quotidienne (lavage, habillage, coiffure, etc.).
- **Incontinence :** Évaluation de l'hygiène urinaire et fécale.
- **Transferts :** Mesure de la capacité de la personne à se lever, se coucher, s'asseoir, et à se déplacer.
- **Alimentation :** Évaluation de la capacité de la personne à se nourrir, à préparer ses repas, et à gérer son régime alimentaire.
- **Déplacement à l'intérieur et à l'extérieur :** Évaluation de la mobilité de la personne à l'intérieur de son domicile et à l'extérieur.
- **Communication à distance :** Évaluation de la capacité de la personne à utiliser un téléphone.

Chaque dimension est notée en fonction du degré de dépendance de la personne, et ces notes sont ensuite utilisées pour attribuer un score global correspondant à un

des six niveaux de GIR (groupes iso-ressources), allant de GIR 1 (perte d'autonomie la plus élevée) à GIR 6 (autonomie la plus élevée).

L'évaluation AGGIR est un outil essentiel pour déterminer les besoins en soins et en assistance des résidents en EHPAD, permettant ainsi d'adapter les services et les prises en charge en fonction du niveau de dépendance de chaque individu.

Cette évaluation est effectuée de façon sérieuse par les établissements, car elle détermine un certain nombre de financements, elle nous donne donc des informations potentiellement précieuses.

Nous avons donc utilisé comme variable le niveau global GIR à l'entrée et au point où nous regardons les données. Nous avons aussi créé une variable qui indique la différence entre ces deux niveaux, pour savoir si la dépendance du résident augmente, est stable ou diminue (ce qui est rare, mais possible). À partir de chacun des 24 items de l'évaluation qui sert à calculer ce niveau, nous avons de même créé une variable pour avoir des informations plus précises sur la mobilité et l'autonomie des résidents.

Pathologies

Pour les pathologies, la présence ou l'absence d'une pathologie déclarée dans le dossier médical est vérifiée au moment précis de la temporalité définie. Nous nous sommes rendu compte que certaines pathologies paraissent trop peu présentes pour une population d'EHPAD. Il est ainsi difficile de savoir si l'information n'a pas été saisie, car le résident n'a pas cette pathologie ou s'il s'agit d'une erreur. Nous avons donc cherché à trouver l'information de la présence d'une pathologie ailleurs que simplement déclarée dans le dossier médical. Pour cela, nous avons utilisé deux moyens : les médicaments prescrits et l'évaluation PATHOS.

Une liste de codes ATC de médicaments correspondant à des traitements pour des pathologies spécifiques a été créée avec des professionnels de santé. Nous avons alors extrait les prescriptions des résidents et ajouté les variables "médicaments associés à une pathologie".

Le modèle PATHOS, lui, est un outil utilisé par les professionnels de santé pour adapter la prise en charge médicale des personnes âgées, en fonction de leur état de santé et de la stratégie thérapeutique correspondante [CNSA, 2022]. Il décrit la situation clinique des personnes et mesure un certain nombre d'indicateurs, données utilisables au niveau d'un individu ou d'un ensemble d'individus.

Comme l'évaluation AGGIR, son utilisation est obligatoire en EHPAD et permet des financements à l'échelle de l'établissement. À l'échelle individuelle, elle constitue un élément qui permet la mise en place d'un plan d'aides et de soins personnalisés. C'est un bon instrument de description synthétique de l'état de santé d'une personne âgée. Le nombre de points qu'il donne d'un point de vue individuel n'est pas considéré comme informatif. Nous avons quand même choisi d'ajouter le score global en variable pour le confirmer.

Pour évaluer le PATHOS, nous avons analysé les pathologies du résident et leur gravité. Les pathologies sont classées par type :

- Affections cardio-vasculaires
- Affections neuro-psychiatriques et troubles psychologiques et/ou comportementaux
- Affections broncho-pulmonaires
- Pathologies infectieuses
- Affections dermatologiques
- Affections ostéo-articulaires
- Affections gastro-entérologiques
- Affections endocriniennes
- Affections uro-néphrologiques
- Autres domaines pathologiques

À partir de l'évaluation PATHOS, nous avons donc pu retrouver des pathologies. Nous avons aussi créé une variable qui devient positive dès qu'un état "grave" est déclaré pour n'importe quelle maladie. De même, une variable qui regroupe toutes les affections neurologiques et une variable qui regroupe toutes les affections cardiovasculaires ont été définies. Cela nous a aussi permis de différencier les personnes qui n'ont réellement pas une pathologie de celles où l'information n'est juste pas remplie.

Relevés périodiques

Pour les relevés tels que les poids et les tensions, nous avons calculé :

- le pourcentage d'augmentation ou de perte sur 1, 3 et 6 mois
- les évolutions du relevé divisé par la durée (1, 3 ou 6 mois)
- les moyennes des relevés sur 1, 3 et 6 mois.

Ces durées nous ont permis d'avoir des indications différentes selon le recul. Nous avons aussi testé différentes méthodes pour essayer de capturer différentes informations à partir de ses séries temporelles.

Des discrétisations expertes ont été effectuées pour les tensions et les pourcentages de différence de poids. Pour les tensions, nous avons observé les tensions systoliques et les avons divisées en :

- Supérieur à 140, ce qui correspond à de l'hypertension
- Entre 120 et 140, ce qui correspond à une tension normale
- Entre 120 et 100, ce qui correspond à une hypotension légère
- Entre 100 et 80, ce qui correspond à une hypotension modérée
- Inférieur à 80, ce qui correspond à une hypotension sévère

Pour les pourcentages de différence de poids sur 1, 3 et 6 mois :

- Catégorie 1 : inférieur à -15 %
- Catégorie 2 : entre -1 % et -10 %
- Catégorie 3 : entre -10 % et 0
- Catégorie 4 : entre 0 et 10 %
- Catégorie 5 : entre 10 % et 15 %
- Catégorie 6 : supérieur à 15 %.

Nous avons aussi calculé l'IMC à partir de la taille et du dernier poids.

Événements ponctuels

De la même façon, pour les événements de type chute, participation aux animations ou hospitalisations, nous avons regardé quand a eu lieu le dernier événement de ce type et la fréquence de ces événements sur 1, 3 et 6 mois.

Une discrétisation experte a aussi été effectuée :

- Catégorie 0 : l'événement n'a jamais eu lieu
- Catégorie 1 : dernier événement il y a plus de 6 mois
- Catégorie 2 : dernier événement entre 3 et 6 mois
- Catégorie 3 : dernier événement entre 1 et 3 mois
- Catégorie 4 : dernier événement dans le dernier mois

4.3.4 Critères d'exclusion

En exportant uniquement des résidents avec au moins un relevé de poids, nous avons obtenu 168 666 résidents. En analysant le pourcentage de valeurs manquantes sur les variables possibles sur la table 4.2b à gauche, on remarque que certaines sont quasiment vides.

On ne représente ici qu'uniquement les variables dont on peut quantifier l'absence, pour les autres variables telles que les chutes ou les pathologies, on ne peut pas être sûr que leur absence n'est pas en réalité une saisie manquante. Par ailleurs, pour les relevés et évaluations, nous avons analysé s'ils avaient été saisis au moins une fois durant le séjour du résident, peu importe la temporalité.

Variable	Valeurs manquantes
Poids	0 %
Âge	0 %
Sexe	0 %
GIR	9 %
Tensions	24 %
Taille	46 %
PATHOS	**49 %**
MMSE	70 %
Albuminémie	80 %
MNA simplifié	93 %
MNA	95 %
Braden	96 %
Norton	97 %
Taille	168 666

(a) à l'extraction

Variable	Valeurs manquantes
Poids	0 %
PATHOS	**0 %**
Âge	0 %
Sexe	0 %
GIR	1 %
Tensions	11 %
Taille	33 %
MMSE	59 %
Albuminémie	70 %
MNA simplifié	87 %
MNA	92 %
Braden	94 %
Norton	97 %
Taille	73 478

(b) après tri

Tab. 4.2. : Pourcentage de valeurs manquantes par variable à l'extraction (4.2a) et après suppression des résidents sans évaluation PATHOS (4.2b).

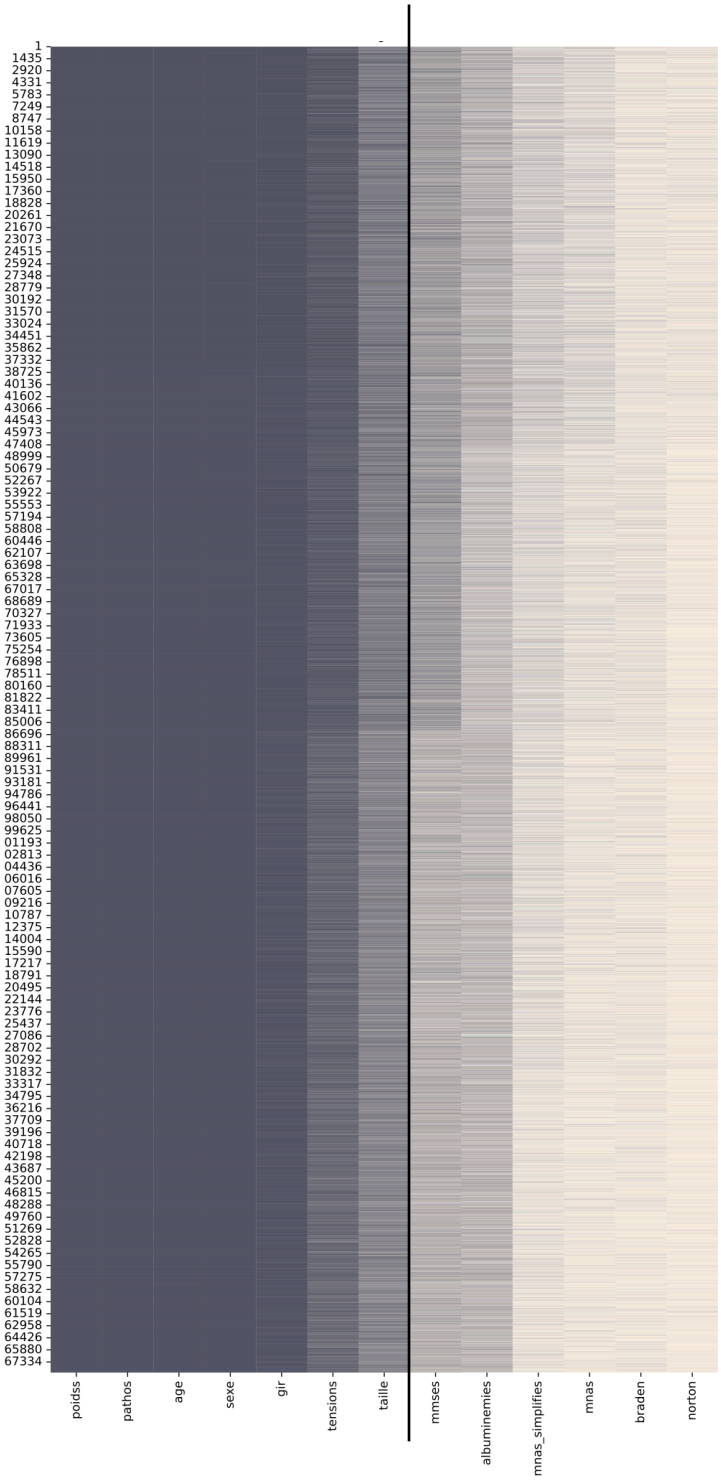


Fig. 4.2. : Représentation des données manquantes pour quelques variables. Chaque trait horizontal noir représente une donnée présente pour cette ligne dans cette colonne. La ligne noire verticale représente la frontière de suppression à 50% de variable manquante dans toute la colonne.

Nous avons fait le choix de garder uniquement les résidents qui ont eu au moins une évaluation PATHOS saisie dans leur dossier, ce qui nous a permis d'obtenir la table 4.2b et la figure 4.2. Ainsi, nous avons notamment des informations plus fiables sur les pathologies : lorsque l'évaluation est faite, le soignant a explicitement indiqué qu'une pathologie était absente. Nous obtenons à ce stade 73 478 résidents. Les pourcentages de valeurs manquantes baissent ainsi pour toutes les variables, mais pour les variables avec plus de 50% de valeurs manquantes, cela restait difficile de les exploiter et nous les avons donc supprimés. Il s'agit des relevés d'albuminémie, de l'évaluation MNA (*Mini Nutritional Assessment*) et sa version simplifiée, qui sont des critères utilisés notamment pour diagnostiquer la dénutrition ; de MMSE (*Mini Mental State Evaluation*), une évaluation des troubles cognitifs et des évaluations de Norton et Braden que nous aborderons plus en section 5.2.

De même, les valeurs aberrantes sont exclues de l'étude. Ainsi, les résidents avec des scores GIR qui ne sont pas entre 1 et 6, des âges inférieurs à 40 et supérieur à 110 ou des durées de séjour trop élevé sont supprimées.

Pour garantir des données de bonne qualité, une suppression des résidents avec trop de variables manquantes est effectuée après analyse et est différentes selon les cibles. Les résultats seront donc exposés dans chaque chapitre consacré.

4.4 Caractéristiques démographiques

Une analyse sur le jeu de données global a été effectuée pour vérifier que l'échantillon correspondait démographiquement aux données d'EHPAD françaises. La comparaison avec les données du rapport de la DREES de 2022 sur des données d'EHPAD française de 2019 [BALAVOINE, 2022] en table 4.3, permet d'observer une homogénéité en termes d'âge à l'admission, de sexe et de durée de séjour. Les résidents semblent par contre moins dépendants dans notre base de données, mais nous avons calculé ces statistiques à l'admission des résidents, contrairement au rapport qui les calcule à une date précise. En effet, pour des raisons d'anonymisation, nous n'avons pas accès à des dates, mais à des délais après entrée. Nous avons donc une base de données représentative de la population nationale, avec des suivis en établissement sur du long terme.

Variable	Jeu de données	Données nationales
Âge à l'admission, moyenne (écart-type)	84.4 (8.6)	85.9
Pourcentage de femmes	72.8 %	72.8 %
Niveau de dépendance :		
GIR1 (très sévère)	7.1 %	16.4 %
GIR2 (sévère)	32.0 %	38.1 %
GIR3 (modéré/sévère)	20.3 %	18.6 %
GIR4 (modéré)	29.5 %	20.2 %
GIR5 (léger)	6.7 %	4.4 %
GIR6 (pas de dépendance)	4.4 %	2.3 %
Durée de séjour en mois, médiane (Q1-Q3)	14.9 (3.0 - 40.4)	13.0 (2.0 - 42.0)

Tab. 4.3. : Caractéristiques démographiques de la base de données, comparaison avec les données du rapport de la DREES [BALAVOINE, 2022]

4.5 Pipeline global

La pipeline de la figure 4.3 représente l'ensemble des étapes, effectuées en Python, communes aux classifieurs de réseaux bayésiens que nous présenterons par la suite. Par ailleurs, les autres méthodes de classification testées sont implémentées de la même façon. Les principales étapes sont donc :

- **Extraction des données**
- **Préparation des données :** transformation des données pour passer d'une base de données événementielle à une base de données tabulaire exploitable pour de l'apprentissage statistique tout en gardant un sens médical
- **Apprentissage automatique :** création d'un classifieur construit à partir de réseaux bayésiens puis ajustement du classifieur avec nos données d'apprentissage
- **Évaluation du classifieur :** Évaluation sur la base test grâce à des scores, des résultats graphiques et des comparaisons avec d'autres méthodes.

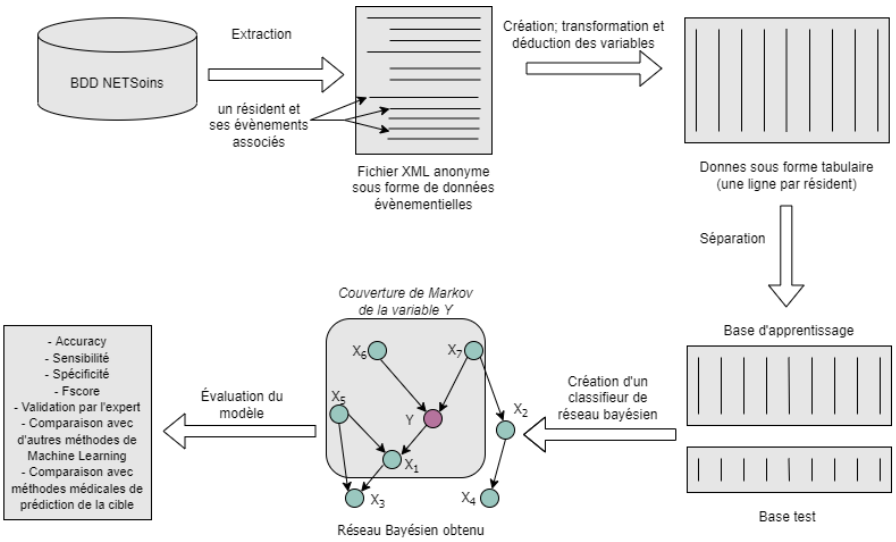


Fig. 4.3. : Pipeline commune à chaque cible

Nous nous sommes pour l'instant concentrés sur les deux premières étapes et avons démontré que nous bénéficions, après une mise en conformité et un certain nettoyage, d'une base de données de santé de grande taille, homogène par rapport à la population française en EHPAD et avec des variables utilisables et pertinentes. Nous verrons ensuite l'application de cette pipeline à différents événements de santé défavorables avec le prétraitement spécifique selon chaque cible, l'apprentissage et les résultats de chaque classifieurs.

Références

BALAVOINE, Angélique (2022). *Des résidents de plus en plus âgés et dépendants dans les établissements d'hébergement pour personnes âgées | Direction de la recherche, des études, de l'évaluation et des statistiques*. ÉTUDES ET RÉSULTATS N°1237 (cf. p. 107, 108).

BCB (2019). *Base Claude Bernard - La base de données sur les Médicaments et les produits de santé* (cf. p. 98).

BELMIN, Joël, Philippe CHASSAGNE, Patrick FRIOCOURT et al. (2016). *Gériatrie : pour le Praticien*. Paris : Elsevier Health Sciences. 1071 p. (cf. p. 94).

CNSA (2022). "Le modèle PATHOS. Guide d utilisation 2022". In : (cf. p. 102).

IA, HUB France (2023). *L'IA éthique en pratique* (cf. p. 97).

LÉGIFRANCE (2017). "Article Annexe 2-1". In : *Code de l'action sociale et des familles* (cf. p. 101).

5 Prédiction de la survenue de la première escarre en EHPAD

La survenue d'escarre est le premier événement de santé défavorable et évitable en EHPAD auquel nous nous sommes intéressés. Nous définirons d'abord plus précisément le sujet, nous présenterons ensuite les méthodes actuellement utilisées pour détecter leur risque, puis le prétraitement spécifique à la cible qui a dû être effectué sur la base de données. Nous exposerons ensuite les différents résultats sur trois temporalités différentes. Enfin, nous aborderons l'application logicielle de cet outil. Les résultats de ce chapitre sont publiés dans les articles [CHARON, WUILLEMIN et BELMIN, 2022], [CHARON, WUILLEMIN et BELMIN, 2023] et [CHARON, WUILLEMIN, HAVRENG-THÉRY et al., 2024].

5.1 Définition de l'escarre

Une escarre est une lésion cutanée d'origine ischémique (diminution de l'apport sanguin artériel à un organe) localisée au niveau de la peau ou des tissus mous sous-jacents, située en général sur une saillie osseuse. Elle est la conséquence d'un phénomène de pression ou de pression associée à un cisaillement.

Le cisaillement peut résulter du frottement de la peau lorsque le patient glisse sur un support. Le mécanisme menant aux escarres est bien défini, il faut qu'il y ait une force de pression supérieure à la pression capillaire des tissus ce qui entraîne l'hypoxie et la nécrose tissulaire [BELMIN et al., 2016].

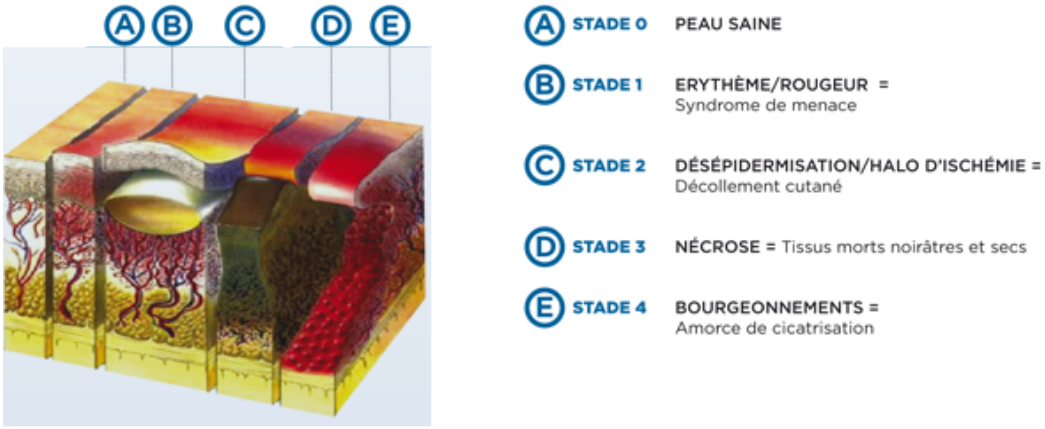


Fig. 5.1. : Schéma sur les stades des escarres [BELMIN et al., 2016]

C'est une pathologie éprouvante pour le patient, douloureuse, qui diminue significativement la qualité de vie. Elle est aussi difficile à traiter et le coût de cicatrisation est important et augmente exponentiellement avec le stade de la maladie [HAUTE

Autorité de Santé, 2001]. Soigner une escarre en EHPAD requiert donc un temps important pour un personnel infirmier trop peu nombreux. Les escarres sont une pathologie qui touche préférentiellement les personnes âgées. Selon une étude réalisée dans cinq pays européens, les deux tiers des escarres sont détectés chez les personnes de 70 ans et plus, avec un pic de fréquence entre 76 et 84 ans [Vanderwee et al., 2007]. Il est important de noter aussi que la prévalence des escarres était de 7.2% dans les EHPAD en France selon une enquête de 2018 [Barrois et al., 2018].

Les facteurs de risque d'escarre sont mentionnés dans la littérature scientifique depuis des dizaines d'années [Ahn et al., 2016]. Plus de 100 facteurs de risque d'escarre sont identifiés dans la littérature scientifique en 2003 [Lyder, 2003]. Les résultats de [Berlowitz et al., 2001] montrent que les facteurs de risques d'escarre pour les résidents en maison de retraite les plus élevés sont l'âge, le sexe masculin, l'ethnicité non blanche, une évaluation non habituelle, l'autonomie dans la mobilité au lit, l'autonomie dans les transferts, l'immobilisation au lit, l'incontinence urinaire, la détérioration de l'état cognitif, le diabète, des maladies vasculaires, une fracture de la hanche dans les 180 jours, l'indice de masse corporelle (IMC), une maladie en phase terminale et les antécédents d'escarre résolue.

Un modèle théorique du risque d'escarre contenant 4 éléments principaux a été créé par [Defloor, 1999] :

- **forces de compression :** lorsque des charges mécaniques exercent une pression sur les tissus d'un patient qui sont comprimés entre une surface, comme une chaise, et les saillies osseuses du corps humain, comme le sacrum ou les fesses. Ces charges mécaniques augmentent le risque d'escarre, surtout lorsqu'elles persistent pendant une longue période chez des patients immobiles et en position assise.
- **forces de cisaillement :** ce sont les forces de glissement qui se produisent lorsque le corps humain est déplacé le long d'une surface. Par exemple, lorsque l'infirmière déplace un patient dans son lit ou transfère un patient du lit à la chaise, actions qui impliquent le contact continu du patient avec une surface tout au long du mouvement et entraînent une séparation des couches de la peau.
- **tolérance tissulaire à la pression :** les facteurs qui changent la capacité du tissu à redistribuer la pression. Il s'agit de facteurs externes, tels que les charges mécaniques et l'humidité.
- **tolérance tissulaire à l'oxygène :** les facteurs qui influencent la distribution de l'oxygène dans le tissu ou le besoin en oxygène du tissu. Il s'agit ici de

facteurs internes tels qu'une pression artérielle basse ou une mauvaise oxygénation.

Bien qu'aucune étude n'ait précisément analysé la temporalité du développement des escarres, [Gefen, 2008] a montré que les escarres peuvent se développer chez les personnes à haut risque en moins d'une heure de pression soutenue sur des tissus corporels vulnérables.

Dans le modèle conceptuel de [Defloor, 1999], plusieurs facteurs affectent la tolérance tissulaire à la pression, notamment l'âge, le stress, la cognition, la sensibilité à la douleur et au confort (acuité), la déshydratation et la masse tissulaire, ainsi que les carences en protéines et en vitamine C. De même, les facteurs qui affectent la tolérance tissulaire à l'oxygène sont la température, les médicaments, la carence en protéines, le tabagisme, la pression artérielle et la présence de certaines maladies (c'est-à-dire celles qui affectent l'apport en oxygène, l'hyperémie réactive et l'occlusion vasculaire). Des études suggèrent que le développement des escarres peut résulter de modifications majeures du fonctionnement normal de ces mécanismes humains. Les escarres peuvent aussi se développer à partir des effets cumulatifs de changements mineurs dans plusieurs de ces mécanismes/facteurs, particulièrement en combinaison avec une pression externe soutenue (forces perpendiculaires à la peau) et/ou la présence d'humidité externe (incontinence fécale ou urinaire, transpiration excessive) et/ou des forces de friction et de cisaillement (glissements, frottements ou forces parallèles à la peau).

Plus récemment, les facteurs de risques pour les personnes âges en maisons de retraite relevés par [Ahn et al., 2016] ont été : l'anémie, la dénutrition, la déshydration, les infections, l'incontinence urinaire et fécale.

Malgré cette causalité multiple, les escarres sont pourtant hautement évitables. D'après [Haute Autorité de Santé, 2001], les mesures générales de prévention sont les suivantes :

- diminuer la pression en évitant les appuis prolongés par la mobilisation ;
- utiliser des supports (matelas, surmatelas, coussins de siège) ;
- observer régulièrement l'état cutané ;
- maintenir l'hygiène de la peau en évitant la macération ;
- assurer un équilibre nutritionnel ;
- favoriser la participation du patient et de son entourage à la prévention et identifier les facteurs de risque.

La prévention implique une prise en charge particulière et multidisciplinaire si bien qu'elle n'est pas applicable à tous les résidents en permanence. Il est ainsi nécessaire de pouvoir identifier les personnes à risque sur lesquelles les soins de prévention seront implémentés. L'objectif de cette étude est donc cette identification précoce du risque d'escarre.

5.2 Méthodes actuelles de détection de risque d'escarre

Les méthodes de détection de risque de survenue d'escarres sont des échelles d'évaluation qui ont été élaborées pour identifier de façon plus précise les malades à haut risque et mettre en place une prévention. Elles fournissent un score à partir de quelques réponses à des questions. La plus connue est l'échelle de Norton [NORTON et al., 1962], qui a été validée pour des personnes de plus de 65 ans. Elle est simple d'utilisation, mais ne prend pas en compte le statut nutritionnel.

Condition physique	Condition mentale	Activité	Mobilité	Incontinence
1 Mauvaise	1 Stuporeux	1 Couche	1 Immobile	1 Fécale et urinaire
2 Pauvre	2 Confus	2 Fauteuil	2 Très limitée	2 Urinaire
3 Moyenne	3 Apathique	3 Marche aidée	3 Peu limitée	3 Occasionnelle
4 Bonne	4 Alerte	4 Ambulant	4 Complète	4 Continent

Le score final est la somme des scores de chacun des 5 items. Plus le score est faible, plus le risque d'escarre est élevé.

<16 : risque élevé ≥16 : risque faible

Fig. 5.2. : Échelle de Norton pour l'évaluation du risque d'escarre [NORTON et al., 1962 ; ABDELLATIF, 2021]

On utilise aussi l'échelle de Braden [BERGSTROM et al., 1998) qui est simple et claire. Mais la simplicité de ces échelles est aussi un inconvénient, elles oublient des facteurs de risque important.

Sensibilité	Humidité	Activité
1 Complètement limitée	1 Constamment humide	1 Confiné au lit
2 Très limitée	2 Très humide	2 Confiné en chaise
3 Légèrement limitée	3 Parfois humide	3 Marche parfois
4 Pas de gêne	4 Rarement humide	4 Marche fréquemment
Mobilité	**Nutrition**	**Frictions et frottements**
1 Totalement immobile	1 Très pauvre	1 Problème permanant
2 Très limitée	2 Probablement inadéquate	2 Problème potentiel
3 Légèrement limitée	3 Correcte	3 Pas de problème
4 Pas de limitation	4 Excellente	
Score total :		

Le score final est la somme des scores de chacun des 6 items. Le risque est d'autant plus élevé que le score final est faible.

≥18 : risque faible | 13 - 17 : risque modéré | 8 - 12 : risque élevé | < 7 : risque élevé

Fig. 5.3. : Échelle de Braden pour l'évaluation du risque d'escarre [Bergstrom et al., 1998 ; Abdellatif, 2021]

La méta-analyse de [H.-L. Chen et al., 2016] n'a trouvé que 8 études évaluant l'échelle de Braden chez des résidents de maisons de retraite. Dans 6 d'entre elles, la performance de l'échelle a été étudiée de manière transversale sur différents types d'établissement de santé, ce qui a produit des résultats de validité contradictoires qui ne permettent pas d'évaluer la prédiction de l'escarre en maisons de retraite. Les 2 études prospectives étaient de petite taille ($n = 335$) et leurs sensibilités (0.73 et 0.79) et spécificités (0.74 et 0.76) pour prédire l'escarre avaient des intervalles de confiance assez larges [Braden et Bergstrom, 1994 ; Souza et al., 2010]. Dans les deux études, l'incidence de l'escarre était extrêmement élevée (21% en 90 jours dans [Souza et al., 2010] et 27% en 4 semaines dans [Braden et Bergstrom, 1994)] ce qui soulève la question de l'applicabilité de ces études anciennes aux maisons de retraites actuelles en France où l'incidence est bien plus faible.

Des recherches existent pour essayer d'améliorer la prédiction de cette pathologie, notamment en essayant de prendre en compte le contexte. Une revue systématique sur la prévention des escarres concluait sur la nécessité d'essayer d'inclure des variables sur les soins quotidiens prodigués pour améliorer la prédiction [Sullivan et Schoelles, 2013]. Les méthodes de *machine learning* ont déjà été utilisés pour prédire les escarres et ont donné des résultats prometteurs, en particulier dans le cadre des soins intensifs [Kaewprag et al., 2015 ; Alderden et al., 2018 ; Ladios-Martin et al., 2020].

Quelques études menées dans des maisons de retraites présentaient plusieurs limites, comme un petit ensemble de données, une faible sensibilité ou une fenêtre de prédiction trop courte pour mettre en place un traitement préventif efficace [Hu et al., 2020 ; Song et al., 2021]. En Corée, [Lee et al., 2021] ont étudié des méthodes

de *machine learning* pour prédire la prévalence des escarres dans 60 maisons de repos à partir des caractéristiques des établissements et des profils agrégés des résidents, mais ils n'ont pas exploré la prédiction individuelle d'escarre.

5.3 Prétraitement spécifique

Nous nous sommes concentrés uniquement sur la survenue de la première escarre d'un résident dans l'établissement, ce qui était plus facile à gérer d'un point de vue temporel. Cliniquement, cela a du sens, car une fois qu'une escarre a été déclarée, le risque de refaire une escarre reste élevé dans tous les cas. L'objectif était donc de prédire la survenue de la première escarre dans l'établissement.

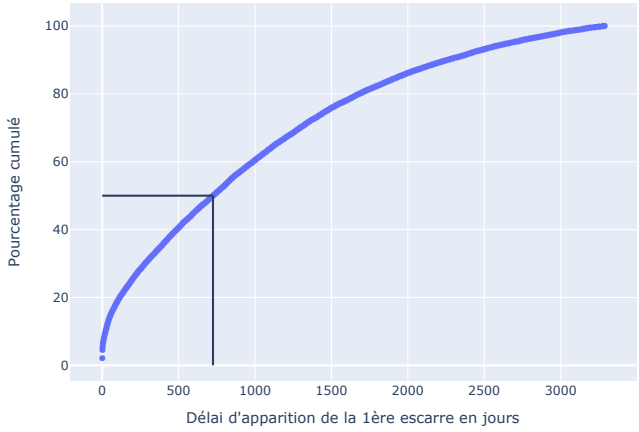


Fig. 5.4. : Distribution de l'apparition de la première escarre dans les EHPAD en pourcentage cumulé. La ligne noire indique que 50% des escarres ont lieu avant 725 jours.

Nous avons fixé 3 objectifs temporels de prédiction : 1 mois avant l'escarre, 2 et 3 mois avant. Ainsi, trois jeux de données différents sont créés en accord avec ces temporalités. Pour avoir un cas d'escarre, il faut qu'une escarre soit déclarée dans les pathologies du résident après l'admission du résident, ou qu'un pansement escarre soit présent dans le plan de soin ou dans les transmissions. Ainsi, si le résident est un cas d'escarre, nous regardons ses données 1 mois avant la survenue de sa première escarre pour le premier jeu de données, 2 mois avant pour le deuxième et 3 mois avant pour le troisième.

Les cas témoins, eux, sont les personnes qui n'ont pas eu d'escarre renseignée dans leur dossier. Pour choisir la temporalité où les données sont étudiées, le système

est plus complexe. Dans un premier temps, nous utilisions les dernières données disponibles pour les cas témoins, mais cela correspond souvent à la fin de vie du résident et engendrait donc des biais (état aggravé). Un système de dates index reflétant la distribution d'apparition de la première escarre a alors été créé à partir de la figure 5.4. Cette répartition est expliquée dans le graphe ci-dessous. De manière pratique, nous avons calculé dans l'ensemble des données de cas les déciles du délai de survenue d'une escarre depuis l'admission dans l'établissement, ce qui a permis de définir 10 groupes. Les résidents du groupe de contrôle ont été sélectionnés au hasard dans l'un des dix groupes, et leur date index a été déterminée sur la base du délai attribué.

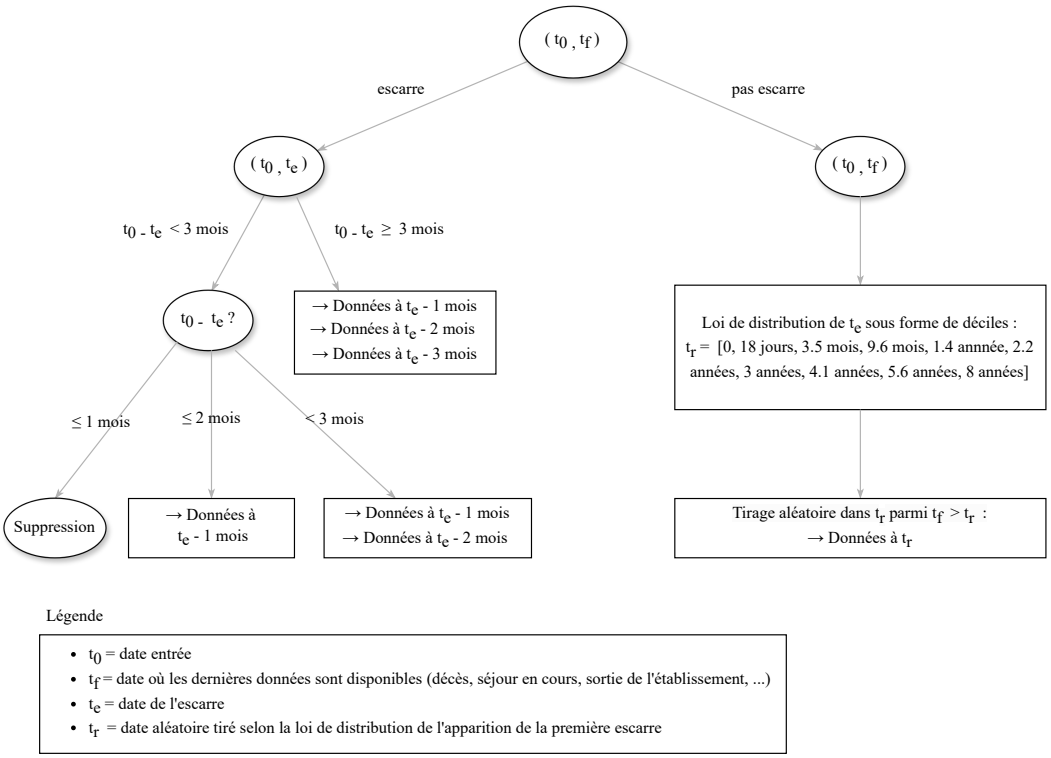


Fig. 5.5. : Arbre représentant la gestion de la temporalité des données

Les variables liées aux échelles de Braden et Norton avaient été supprimées en 4.3.4, car très peu remplies. Nous ne l'avons pas remis en question malgré les possibles informations sur des facteurs de risques qu'elles contiennent. En effet, en gardant en perspective l'implémentation des modèles en routine clinique, nous ne voulions pas qu'il puisse devenir important d'avoir effectué au préalable une évaluation de risque d'escarre pour que notre outil prédise précisément le risque d'escarre.

Nous avons finalement obtenu 148 variables. Un dernier tri a été effectué en

éliminant les lignes avec trop de valeurs manquantes. Nous avons utilisé seulement les résidents avec moins de 10 valeurs manquantes au total. Cela correspond uniquement à une suppression de 1.4 % de lignes, pour les trois bases de données, montrant que nous en avons finalement très peu. La méthode utilisée pour la complétion des dernières valeurs manquantes est celle des k plus proches voisins, présentée en 2.4, avec ici k $= 5$.

5.4 Implémentation du classifieur de réseau bayésien

Nous avons utilisé la bibliothèque pyAGrum, qui permet de construire des modèles et des algorithmes pour les modèles graphiques probabilistes en Python [DUCAMP et al., 2020]. Plus précisément, le module `skbn` permet de créer des classifieurs binaires et multi-classes qui utilisent un réseau bayésien pour prédire, tout en étant compatibles avec les fonctions de scikit-learn. Nous appellerons ces classifieurs "BNClassifier". Plusieurs méthodes d'apprentissage, présentées dans la section 1.4.3, sont disponibles : Chow-Liu, *naive Bayes*, TAN, MIIC, *Greedy Hill Climbing* et *Tabu list searching*. Il est de même possible d'ajouter un *a priori* de type *Smoothing*, BDeu ou Dirichlet (voir section 1.4.2), et un type de score parmi AIC (*Akaike information criterio*), BIC (*Bayesian Information criterion*), BD (*Bayesian-Dirichlet*), BDeu, K2, *log2 likelihood ratio test* (voir section 1.4.3). Des contraintes sur la structure peuvent aussi être imposées.

Puisque nous sommes dans un contexte de prévention où nous voulons éviter un maximum les faux négatifs, nous avons choisi d'optimiser nos modèles selon un score agrégeant la précision et le rappel, tout en donnant deux fois plus de poids au rappel : le $F_{\beta=2}$ score, tel que défini en 2.3.1 et que nous nommerons F-2 score.

Les différentes méthodes, combinées aux différents *a priori* ou scores ont donc été testés, et les paramètres qui optimisaient le F2-score était finalement la méthode d'apprentissage MIIC, sans *a priori* ni score particulier. Il n'est pas étonnant que les *a priori* ne soient pas nécessaires ici, car nous avons une base de données de grande taille. De plus, la méthode MIIC étant basée sur des contraintes, les scores ne sont pas compatibles. Nous n'avons pas, par contre, exploré la possibilité d'ajouter des contraintes, car nous voulions rester dans un contexte d'apprentissage du modèle à partir des données uniquement. Les courbes ROC et précision-rappel sont tracées à chaque fois à partir de l'échantillon d'apprentissage, pour permettre de choisir le seuil. De la même façon, nous avons privilégié un seuil permettant d'optimiser le F2-score.

La plupart des variables sont binaires ("présence" ou "absence"), mais les variables liées à l'évaluation AGGIR, elles, peuvent prendre trois valeurs ("indépendance", "dépendance partielle", "dépendance"). Il y a aussi des variables qui ont été discrétisées de façon experte, comme décrit précédemment (4.3.3). La variable ayant le plus de modalités (11) que nous voulions garder ainsi est la différence de GIR entre l'admission et actuellement. Cela correspond donc au seuil à partir duquel nous discrétisons les variables. Les 21 variables continues restantes ont été discrétisées en 18 catégories maximum, en utilisant la méthode des quantiles. Le nombre de catégories dans lequel discrétiser a été testé entre 5 et 20 et celui optimisant les résultats en termes de F2-score a été gardé. Puisque des temporalités étaient fixées pour les cas témoins, la variable "durée après entrée" ne pouvaient pas être discrétisées de la même façon pour ne pas prendre le risque que les cas témoins et les cas d'escarres se retrouvent dans des catégories différentes. Pour cette variable uniquement, nous avons donc discrétisé en déciles.

5.5 Prédiction de l'escarre 1 mois avant son apparition

Les données de 58 368 résidents ont été utilisées et parmi eux, 16 942 ont développé une escarre, correspondant à taux de cas positifs de 29%. Pour entraîner le classifieur, 75% des données ont été utilisées. Les 25% restant ont été utilisés pour la validation. La taille de ce jeu de données était suffisamment importante pour permettre une simple séparation aléatoire et non une validation croisée.

Nous avons confronté le F2-score, F1-score, précision, rappel, *accuracy* (voir section 2.3.1) avec d'autres modèles de *machine learning* de scikit-learn [Pedregosa et al., 2011] et XGBoost [T. Chen et Guestrin, 2016] que nous avons présentés en 2.1 et en 2.2. Nous avons aussi comparé avec le classifieur QDA (*Quadratic Discriminant Analysis*) de scikit-learn. Il s'agit d'une méthode probabiliste simple avec une frontière de décision quadratique [Hastie et al., 2009].

5.5.1 Résultats graphiques

L'interprétabilité du modèle nous a permis d'étudier le réseau bayésien obtenu, présent en entier en figure 5.6. Sa grande taille rend sa compréhension difficile.

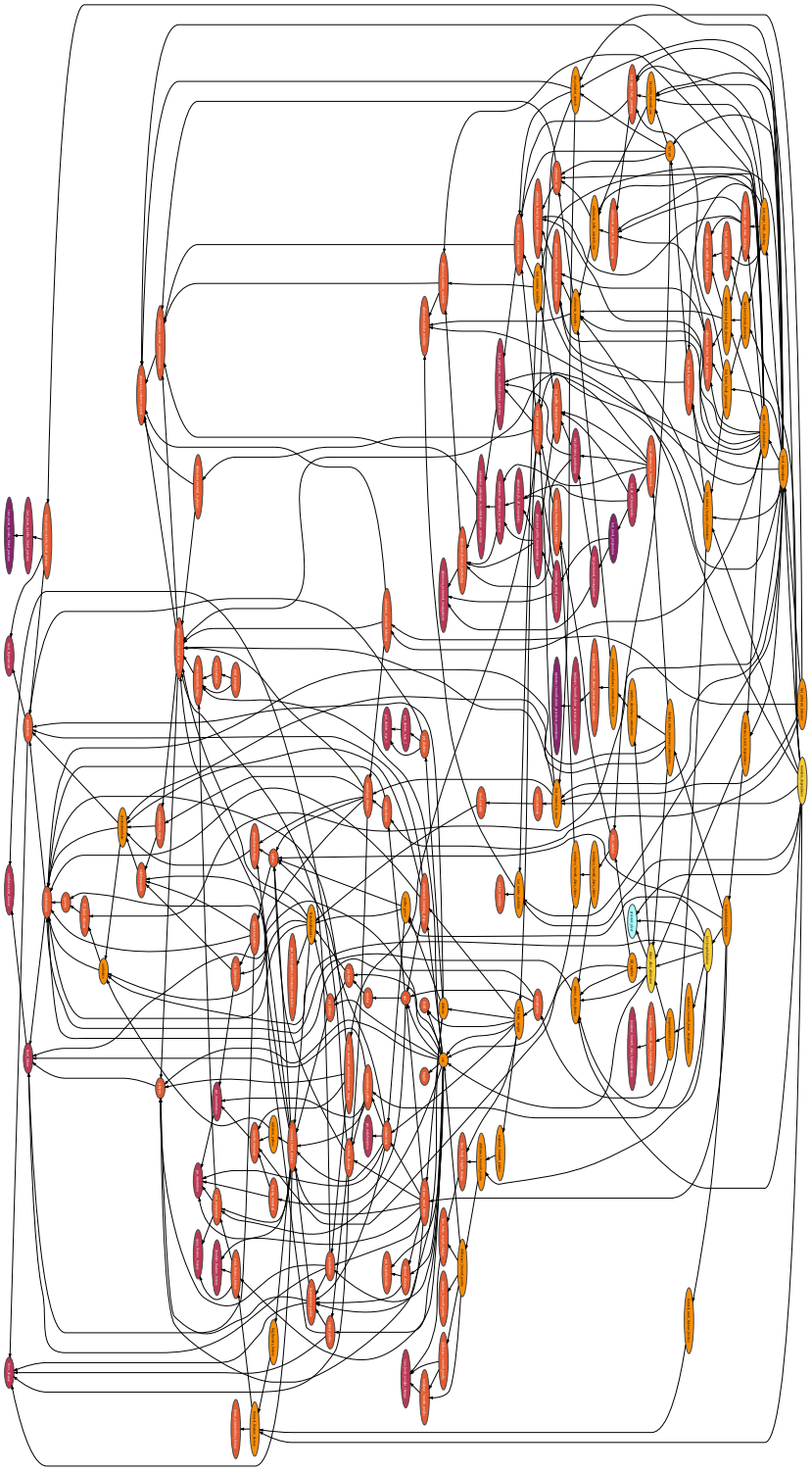


Fig. 5.6. : Réseau bayésien entier obtenu. La couleur des variables indique la distance à la cible en termes de couverture de Markov, la légende détaillée est disponible en annexe en table B.1.

Nous avons donc plus précisément étudié la frontière de Markov de la cible escarre (`pressure_ulcer`), définie en 1.3.2, et questionné des experts en gériatrie à son sujet. Les variables de la frontière de Markov (en figure 5.7) et leurs valeurs de Shapley (figure 5.8) ont été étudiées par des gériatres, qui les ont jugées cohérentes avec les connaissances cliniques et la littérature. Les valeurs de Shapley conditionnelles ont été obtenues grâce au module `explain` de pyAgrum et nous avons utilisé la librairie SHAP pour les illustrer.

En effet, les variables sélectionnées par le modèle comprenaient la dépendance aux transferts (`transfers_dependance`), venant de l'évaluation AGGIR, qui est fortement liée à une mobilité réduite, un facteur de risque classique pour l'escarre. De plus, elles contiennent une hospitalisation récente (`last_hospitalization`), qui est souvent associée à des facteurs favorisant l'escarre, comme une maladie aiguë, une aggravation de l'immobilité, une mauvaise alimentation et une inflammation systémique [DWYER et al., 2014]. Une troisième variable sélectionnée par le modèle était la durée actuelle du séjour en EHPAD (`delay_after_admission`). En effet, un séjour plus long en EHPAD est associé à une avancée en âge, au déclin de l'autonomie et à une durée plus longue des maladies chroniques, facteurs qui pourraient privilégier l'apparition de l'escarre.

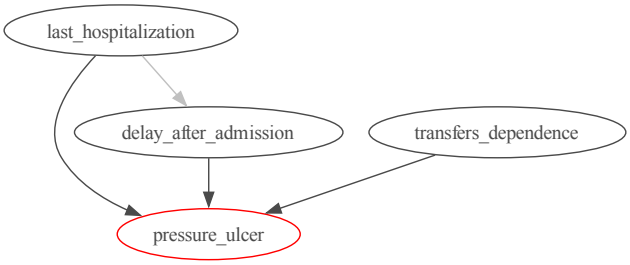


Fig. 5.7. : Frontière de Markov de la cible escarre (`pressure_ulcer`) du BNClassifier 1 mois avant. En gris, un arc présent dans le réseau bayésien inutilisé dans la construction de la frontière de Markov.

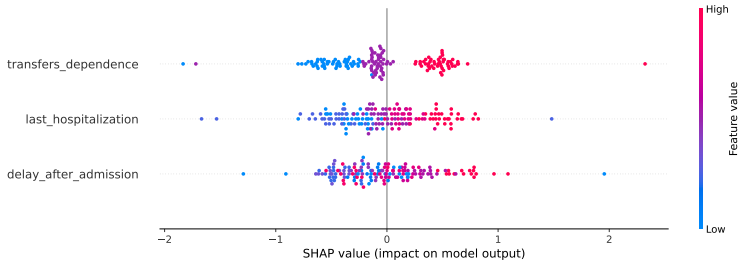


Fig. 5.8. : Valeurs de Shapley des variables de la frontière de Markov de niveau 1.

Il est important de souligner que, parmi les 148 variables disponibles, le réseau bayésien n'a identifié que 3 variables nécessaires pour prédire la survenue de l'escarre un mois avant. Nous nous sommes donc demandés de quelles autres variables plus précisément ces trois variables résumait l'information, et nous avons donc calculé la couverture de Markov de niveau 2 (figure 5.9) qui comprend 39 variables, triées dans la table 5.1, et liées à la dépendance, à l'hospitalisation, aux chutes et aux fractures du fémur, aux maladies infectieuses et au diabète, à la perte de poids et au manque de participation aux activités de loisirs, qui ont aussi été considérées comme pertinentes par les experts.

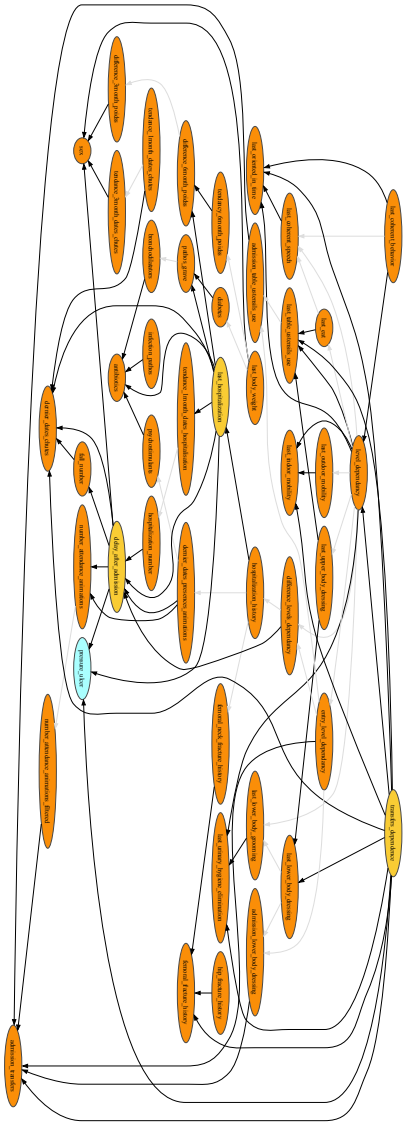


Fig. 5.9. : Représentation dans la couverture de Markov de niveau 2 de notre cible. La couleur des variables indique la distance à la cible en termes de couverture de Markov, la légende détaillée est disponible en annexe en table B.1.

Type	Variables
Démographique	sexe
Poids	6 mois tendance poids
	6 mois différence poids
	3 mois différence poids
	dernier poids
Variables liées au GIR	évolution niveau GIR
	niveau GIR admission
	habillage bas du corps admission
	utilisation couverts admission
	dépendance transferts admission
	cohérence discours
	cohérence comportement
	orientation temps
	toilette bas du corps
	habillage haut du corps
	habillage bas du corps
	utilisation couverts
	dépendance repas
	hygiène élimination urinaire
	mobilité intérieure
	mobilité extérieure
Activités	nombre participation activités
	délai dernière participation activités
Maladies	état grave PATHOS
	infection
	diabète
Hospitalisation	1 mois tendance hospitalisation
	nombre hospitalisation
	antécédent hospitalisation
Chutes et fractures	3 mois tendance chutes
	1 mois tendance chutes
	nombre chutes
	délai dernière chute
	antécédent fracture fémur
	antécédent fracture hanche
Médicaments	psychostimulants
	bronchodilatateurs

Tab. 5.1. : Variables présentes dans la couverture de Markov de niveau 2 de la cible, représentée en 5.9. Les variables sont triées par type.

5.5.2 Résultats numériques et comparaison avec d'autres méthodes

La courbe ROC et la courbe de précision-rappel offrent une évaluation complète des performances du classificateur à différents seuils de classification. Elles sont ici tracées dans la figure 5.10 à partir d'un échantillon d'apprentissage, pour nous permettre de visualiser le seuil de probabilité choisi. Le choix d'un seuil conventionnel, tel que le point optimal de la courbe ROC, peut permettre d'obtenir une précision légèrement supérieure, mais au détriment d'une réduction importante du rappel. Dans ce contexte, la sélection d'un seuil qui maximise le F2-score est apparue comme un choix plus approprié.

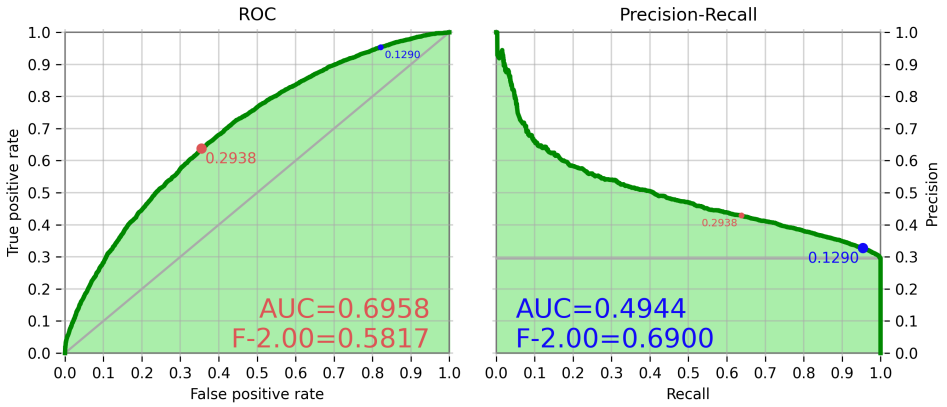


Fig. 5.10. : Courbes ROC et précision-rappel du BNClassifier pour la prédiction d'escarre 1 mois avant (le point rouge correspond au seuil minimisant la distance au point idéal de la courbe ROC, et le point bleu, celui maximisant le F2-score).

Pour vérifier comment se comporte BNClassifier sur d'autres données que celles de l'apprentissage ainsi qu'analyser sa robustesse, nous avons répété 25 fois les séparations aléatoires entre échantillons d'apprentissage et échantillons de validation. Chaque modèle a donc été appris sur un ensemble d'apprentissage et testé sur l'ensemble de validation complémentaire 25 fois différentes. Les résultats sont donc exprimés en moyenne. Les modèles sont classés par ordre décroissant des valeurs du score F2. Pour chaque colonne, la valeur en gras indique le meilleur score (maximum). Les intervalles de confiance sont disponibles en annexe et montrent que les résultats sont stables sur l'ensemble des échantillons de validation.

Si l'on considère le F2-score comme notre principale métrique d'évaluation, le BNClassifier a surpassé toutes les autres méthodes pour prédire les escarres. Toutefois, si l'on examine l'*accuracy* globale, notre méthode affiche des performances comparativement plus faibles. Il est important de noter que l'*accuracy* mesure la

capacité du classificateur à classer correctement les instances dans toutes les classes sans tenir compte des déséquilibres entre les classes. Dans notre ensemble de données, les cas d'escarres représentaient environ 29 % des échantillons, ce qui signifie qu'un classificateur qui ne prédit que les classes négatives obtiendrait une précision d'environ 71 %, similaire au résultat de *Random Forest*. Ces résultats soulignent que le BNClassifier a excellé dans la classification correcte des instances positives avec une précision élevée, mais au prix d'un plus grand nombre de faux positifs. On remarque par ailleurs que les scores exposés dans la courbe précision-rappel calculés sur un échantillon d'apprentissage, sont très proches de ceux obtenus en moyenne sur les échantillons de validation, montrant que le modèle n'effectue pas de surapprentissage.

	F2-score	F1-Score	Précision/ VPP	Rappel/ Sensibilité	Accuracy
BNClassifier	**0.67**	0.47	0.32	**0.94**	0.40
QDA	0.50	0.45	0.39	0.56	0.60
Naïve Bayes	0.47	**0.55**	0.38	0.63	0.60
Decision Tree	0.47	0.55	0.38	0.63	0.60
MLP	0.37	0.39	0.44	0.36	0.68
XgBoost	0.33	0.39	0.56	0.30	**0.73**
AdaBoost	0.24	0.30	0.56	0.21	0.72
Logistic Regression	0.21	0.27	0.56	0.18	0.72
Random Forest	0.14	0.20	**0.65**	0.12	0.72

Tab. 5.2. : F2-score et autres métriques évaluant les performances des différents modèles pour la prédiction à un mois des escarres chez les résidents en EHPAD.

Cependant, notre étude présente certaines limites qui reviendront au fil des applications. La qualité des données du dossier médical est imparfaite, avec certaines valeurs manquantes que nous ne pouvons pas repérer et éventuellement des erreurs et, pour les variables qui changent avec le temps, la fréquence d'acquisition varie fortement d'un résident à l'autre.

Et, plus particulièrement à cette application, le stade de l'escarre n'était pas disponible dans la plupart des dossiers. Il est probable que les escarres de stade 1 aient donc été largement sous-diagnostiquées ou sous-déclarées. De même, nous n'avons pas pu trouver dans le dossier des indications sur des soins préventifs aux escarres, ce qui aurait indiqué un risque repéré par les soignants, et potentiellement un faux négatif expliqué par la prise en charge des soignants qui ont réussi à éviter l'escarre.

5.6 Prédiction de l'escarre 2 mois avant son apparition

De la même façon, nous avons appliqué les mêmes méthodes pour créer et évaluer un classifieur qui prédit les escarres deux mois avant leur survenue. Ici, nous utilisons une base de 57 503 résidents avec 27.95% de cas positifs.

5.6.1 Résultats graphiques

La frontière de Markov que l'on peut voir en figure 5.11 est ici de plus grande taille, avec une seule variable en commun du modèle précédent : la dépendance dans les transferts. La dénutrition revient deux fois, sa présence (ou absence) à la fois dans le dossier médical et dans l'évaluation PATHOS. Il parait logique que les deux variables soient dépendantes entre elles, mais nous ne pouvions pas les agréger, car les informations n'avaient pas forcément la même temporalité.

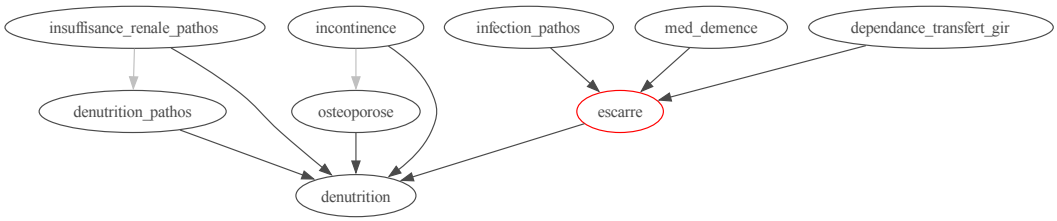


Fig. 5.11. : Frontière de Markov de la cible escarre du BNClassifier 2 mois avant. En gris, les arcs présents dans le réseau bayésien inutilisés dans la construction de la frontière de Markov.

L'analyse des valeurs de Shapley, en figure 5.12, nous indique que c'est bien une haute dépendance dans les transferts. C'est aussi la prise de médicament pour la démence et une indication de dénutrition dans le dossier médical qui sont aussi des variables explicatives du modèle. L'absence de dénutrition dans la dernière évaluation PATHOS impacte ici positivement le risque d'escarre. Une dénutrition présente dans le dossier médicale, mais pas dans l'évaluation PATHOS pourrait indiquer une dénutrition qui s'est déclarée récemment et que l'évaluation PATHOS n'est pas encore à jour. On retrouve donc des facteurs de risques connus des escarres. Par contre, on observe que la présence d'une infection dans l'évaluation PATHOS augmente le risque d'être classé négativement en risque d'escarre, mais que son

absence n'impacte pas la prédiction. Il est difficile d'interpréter ceci, mais on pourrait imaginer que des signes d'infection augmentent l'attention du personnel soignant sur un résident et sur son risque d'escarre, qui a donc été évité. En effet, nous savons qu'il y a des actions du personnel soignant auquel nous n'avons pas pu avoir accès et qui pourraient aider le modèle. Par ailleurs, il est plus difficile d'analyser l'impact de l'incontinence, l'insuffisance rénale et l'ostéoporose, mais ils sont aussi des facteurs de risques.

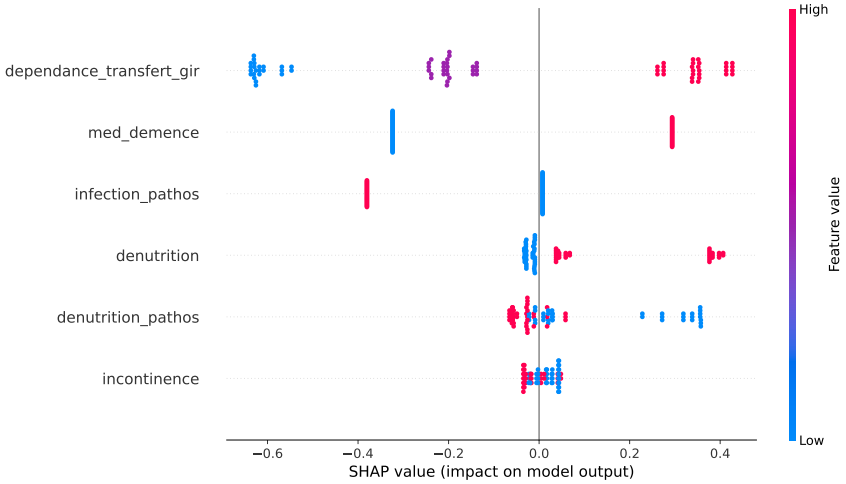


Fig. 5.12. : Valeurs de Shapley des variables de la frontière de Markov du BNClassifier 2 mois avant.

5.6.2 Résultats numériques et comparaison avec d'autres méthodes

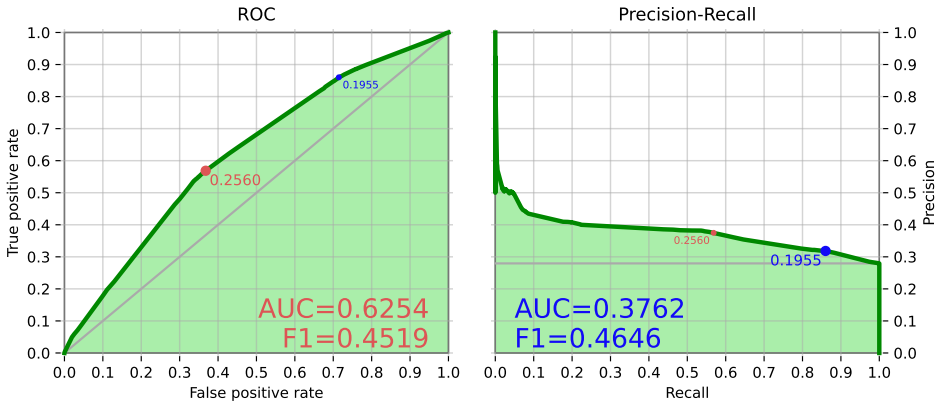


Fig. 5.13. : Courbes ROC et précision-rappel du BNClassifier pour la prédiction d'escarre 2 mois avant (le point rouge correspond au seuil minimisant la distance au point idéal de la courbe ROC, et le point bleu, celui maximisant le F2-score).

La courbe précision-rappel présente en figure 5.13 nous indique qu'ici, le seuil optimisant le F1-score permet déjà d'obtenir un rappel élevé. Le seuil optimisant le F2-score lui, se situe tout à droite, avec un rappel presque parfait, mais un *FPR* quasiment à 1. Ainsi, la plupart des cas sont classés positivement. Le choix d'utiliser le seuil du F1-score parait ici plus adapté.

	F2-score	F1-Score	Précision/ VPP	Rappel/ Sensibilité	Accuracy
BNClassifier	**0.64**	**0.46**	0.32	**0.85**	0.45
Naive Bayes	0.54	0.45	0.35	0.62	0.58
QDA	0.50	0.42	0.34	0.57	0.57
Decision Tree	0.35	0.34	0.33	0.35	0.62
Neural Net	0.33	0.34	0.37	0.32	0.66
XgBoost	0.21	0.26	0.47	0.18	0.71
Logistic Regression	0.10	0.15	0.50	0.09	**0.72**
AdaBoost	0.10	0.15	0.48	0.09	**0.72**
Random Forest	0.07	0.10	**0.51**	0.06	**0.72**

Tab. 5.3. : F2-score et autres métriques évaluant les performances des différents modèles pour la prédiction à deux mois des escarres chez les résidents en EHPAD.

Les résultats numériques présents en table 5.3 suivent les mêmes tendances que les précédents pour le modèle 1 mois avant. Le modèle est quasiment aussi précis, mais un peu moins sensible. BNClassifier reste quand même le meilleur en termes de F1-score, de F2-score et de Rappel. Les modèles avec une *accuracy* de 0.72, mais un rappel extrêmement bas sont en réalité des modèles que ne prédisent quasiment que des cas négatifs, donc sans escarre.

5.7 Prédiction de l'escarre 3 mois avant son apparition

Les mêmes méthodes ont été appliquées pour prédire le risque de façon encore plus lointaine, ici 3 mois avant sa survenue. La base de données utilisée se compose de 56 708 résidents avec ici 26.98% de cas positifs.

5.7.1 Résultats graphiques

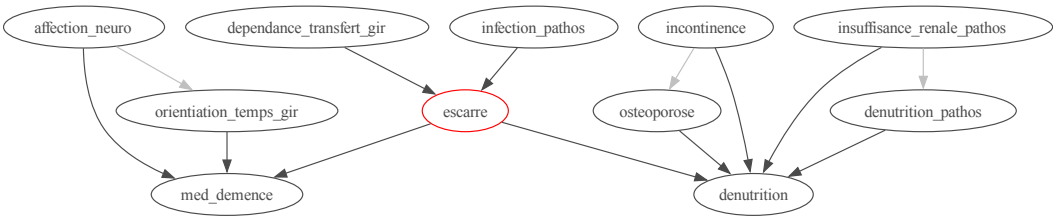


Fig. 5.14. : Frontière de Markov de la cible escarre du BNClassifier 3 mois avant. En gris, les arcs présents dans le réseau bayésien inutilisés dans la construction de la frontière de Markov.

Le nombre de variables sélectionnées par le réseau bayésien augmente encore, comme nous pouvons le voir en figure 5.14. Beaucoup de variables du modèle 2 mois avant sont reprises et l'analyse des valeurs de Shapley en figure 5.15 nous permet de tirer les mêmes conclusions. Deux nouvelles variables sont présentes dans la couverture de Markov : la présence ou l'absence d'affections neurologiques dans l'évaluation PATHOS, ainsi que la capacité d'orientation dans le temps indiqué dans l'évaluation AGGIR. Ces deux variables sont liées à une détérioration de l'état cognitif, facteur de risque que nous avons déjà évoqué pour les escarres.

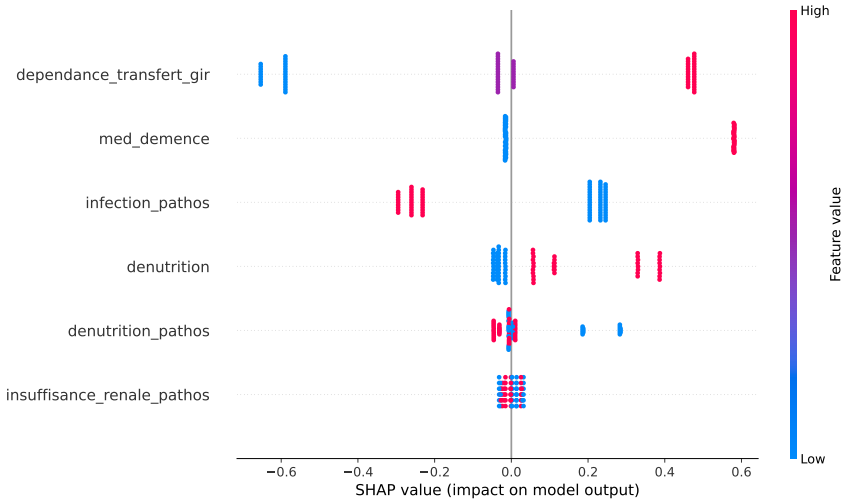


Fig. 5.15. : Valeurs de Shapley des variables de la frontière de Markov du BNClassifier 3 mois avant.

5.7.2 Résultats numériques et comparaison avec d'autres méthodes

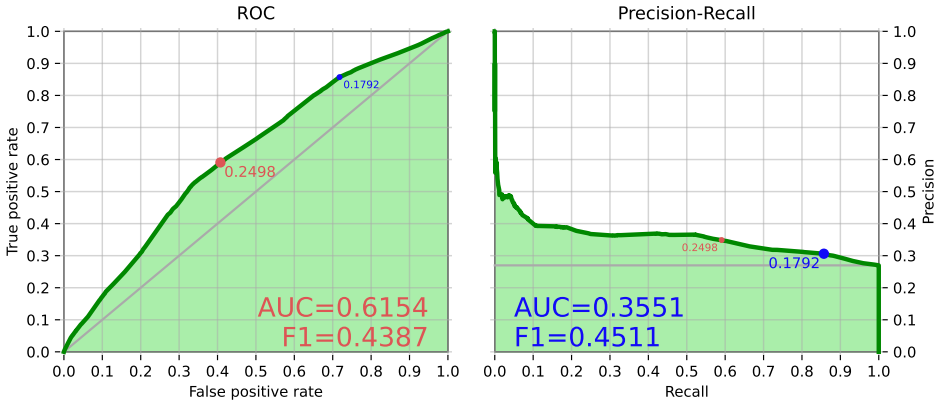


Fig. 5.16. : Courbes ROC et précision-rappel du BNClassifier pour la prédiction d'escarre 3 mois avant (le point rouge correspond au seuil minimisant la distance au point idéal de la courbe ROC, et le point bleu, celui maximisant le F1-score).

Sur la figure 5.16, nous pouvons faire la même analyse que pour le modèle 2 mois avant. Le seuil optimisant le F1-score est donc utilisé.

	F2-score	F1-Score	Précision/ VPP	Rappel/ Sensibilité	Accuracy
BNClassifier	**0.62**	**0.45**	0.31	**0.84**	0.44
Naive Bayes	0.53	0.44	0.34	0.61	0.58
QDA	0.49	0.41	0.33	0.56	0.56
Arbre de décision	0.33	0.33	0.32	0.34	0.62
MLP	0.32	0.33	0.36	0.31	0.66
XgBoost	0.18	0.23	0.45	0.15	0.72
AdaBoost	0.07	0.10	0.49	0.06	**0.73**
Regression Logistique	0.07	0.1	0.51	0.06	**0.73**
Random Forest	0.05	0.08	**0.52**	0.04	**0.73**

Tab. 5.4. : F2-score et autres métriques évaluant les performances des différents modèles pour la prédiction à trois mois des escarres chez les résidents en EHPAD.

La table 5.4 nous montre des résultats quasiment similaires au modèle de 2 mois, avec une très légère baisse. Ainsi, plus le modèle est proche, meilleure est la qualité de la prédiction. Les résultats des autres modèles suivent aussi les mêmes tendances.

5.8 Escarre développée à l'hôpital

Dans notre contexte, il est important de différencier les escarres développées en EHPAD des escarres développées au cours d'une hospitalisation. En effet, la prévention que nous pouvons mettre en place dans l'EHPAD pour l'éviter ne suivra pas nécessairement lors de l'hospitalisation. La littérature sur le sujet utilise le terme "HAPI" pour "*Hospital-Acquired Pressure Injuries*" et des méthodes de *machine learning* sont aussi utilisées dans ce contexte [DWEEKAT et al., 2023].

Nous avons donc effectué une analyse dans nos données en examinant le délai entre la première escarre et la dernière hospitalisation, s'il y en a eu une. Seule la date d'entrée en hospitalisation est indiquée, mais les escarres déclarées en dehors de l'EHPAD ne sont pas prises en compte.

Il y 49.86% des résidents ayant eu une escarre qui a eu une hospitalisation au préalable, mais les délais vont de quelques jours à plusieurs années.

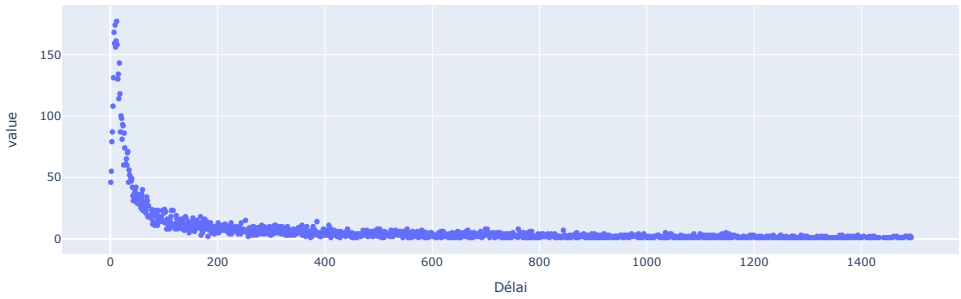


Fig. 5.17. : Répartition des délais entre hospitalisation et escarre en jours

Dans la figure 5.17, on aperçoit un pic à 9 jours, donnant l'impression que la plupart des escarres sont déclarées à ce moment-là. Si nous utilisons la somme cumulée pour analyser le pic à l'échelle du nombre d'hospitalisations globale, nous obtenons la figure 5.18.

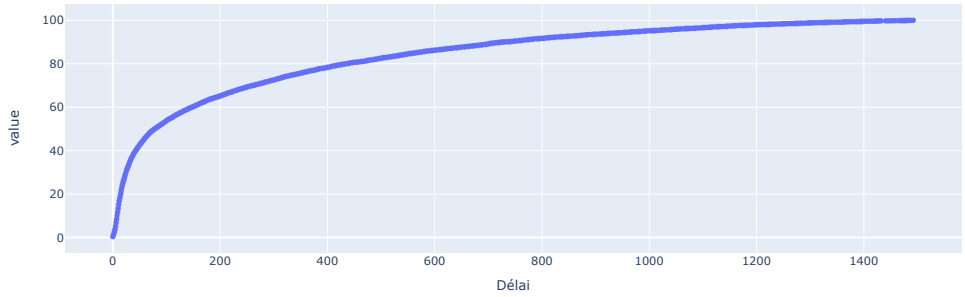


Fig. 5.18. : Répartition des délais entre hospitalisation et escarre en jours en cumulé du pourcentage de la base de cas d'escarre

La présence d'un pic d'incidence d'escarres autour de 9 jours après l'hospitalisation peut sembler préoccupante, mais il convient de le mettre en perspective avec l'évolution globale de l'incidence d'escarres en EHPAD. Si l'hospitalisation n'avait aucun impact sur l'incidence d'escarres, on s'attendrait à observer un nombre constant de déclarations d'escarres par jour sur une période raisonnable de la vie en EHPAD. C'est une hypothèse un peu forte, mais qui permet de poursuivre l'analyse.

Sur la figure 5.18, on observe une zone de linéarité aux alentours d'une cinquantaine de jours. En utilisant une analyse contrefactuelle, nous pouvons construire une droite (figure 5.19) représentant l'évolution du nombre d'escarres si l'hospitalisation n'avait aucun impact. Cette droite est obtenue en translatant la courbe bleue à partir de 40 jours. Si nous calculons la régression linéaire ajustée entre 40 et 80 jours, l'équation est $y = 32.43x + 2549$.

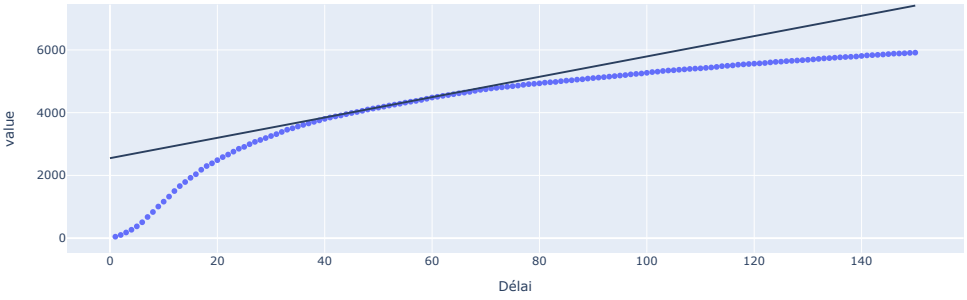


Fig. 5.19. : Répartition des délais entre hospitalisation et escarre en jours en cumulé du nombre de la base de cas d'escarre entre 1 et 140 jours. La droite représente une régression linéaire qui s'ajuste sur la courbe entre 40 et 80 jours avec un $R^2 = 0.99$ et dont l'équation est $y = 32.43x + 2549$.

Cette hypothèse nous permet donc de conclure que le nombre d'escarres attribuables à l'hospitalisation est de 2549 dans notre base de données, ce qui est relativement rare.

Par ailleurs, on peut observer que le décrochage de la courbe est important jusqu'à une vingtaine de jours. L'influence de l'hospitalisation dure ainsi jusqu'à 20 jours, sachant que nous utilisons dans le modèle uniquement des valeurs à partir de 30 jours avant l'escarre. Notre système ne devrait donc pas être impacté par ces cas.

De plus, si nous analysons la prédiction de nos BNClassifieur sur ces cas particuliers ayant eu une hospitalisation moins de 30 jours avant la survenue de l'escarre, on observe que ces cas ont été classés comme positif à 99.90% pour le modèle 1 mois avant, 78.892 % pour 2 mois avant et 77.63 % pour 3 mois avant. Ainsi, bien que l'algorithme n'ait pas accès à l'information sur cette hospitalisation, à un mois, il prédit que ces résidents sont à risque. À deux et à trois mois, ils le prédisent aussi dans la grande majorité des cas, mais à l'image des résultats généraux des modèles, ils sont un peu moins sensibles. L'hospitalisation est donc un facteur aggravant, mais pas déclenchant.

5.9 Évaluation de l'échelle de Braden et comparaison

Dans la base de données, les résidents dont le risque d'escarre avait été évalué par l'échelle de Braden ont été identifiés, cet outil étant largement plus utilisé que l'échelle de Norton. Cela a permis de comparer ses résultats avec notre classifieur de réseau bayésien calculé au même moment et dans le même échantillon, ainsi qu'avec le reste des résidents. La prédiction du risque par l'échelle de Braden avec un seuil à 12, correspondant au score maximum qui permet d'obtenir un risque élevé (voir figure 5.3), et le classifieur ont été comparés à la présence/absence d'escarre un mois plus tard.

Parmi les 37 231 dossiers résidents analysés, seulement 5.6 % avait un résultat d'échelle de Braden. Parmi eux, 33 % ont développé une escarre un mois plus tard. Cela contraste avec l'occurrence plus faible d'escarre (16 %) parmi les 35 121 résidents qui n'ont jamais eu d'échelle de Braden effectué. Cela montre que le simple fait d'effectuer une échelle de Braden est un meilleur prédicteur d'un risque élevé de développer une escarre que le résultat de l'échelle en elle-même. Il est probable que le personnel soignant réalise une évaluation préventive uniquement lorsque son impression clinique indique qu'il existe un risque. Les matrices de confusion des prédictions de l'échelle de Braden, de notre BNClassifier obtenu sur les résidents

qui ont été évalués par l'échelle de Braden (la population à haut risque) et du BNClassifier sur le reste des résidents (population à plus faible risque) sont présentes en Table 5.5.

État	**Échelle de Braden**		
1 mois après	Risque escarre	Pas de risque	Total
Escarre	107	588	695
Pas d'escarre	125	1 290	1 415
Total	232	1 878	2 110
	BNClassifier (HR)		
	Risque escarre	Pas de risque	Total
Escarre	693	2	695
Pas d'escarre	1 405	10	1 415
Total	2 098	12	2 110
	BNClassifier (FR)		
	Risque escarre	Pas de risque	Total
Escarre	4 867	789	5 655
Pas d'escarre	2576	26 889	29 465
Total	7 443	2 7678	35 121

Tab. 5.5. : Matrices de confusion de l'échelle de Braden, du BNClassifier dans l'échantillon à haut risque (HR) et à faible risque (FR) selon l'équipe médicale.

La sensibilité et spécificité pour les trois matrices de confusions sont présentes en Table 5.6. La performance de l'échelle de Braden est faible, avec une sensibilité à 0.15, indiquant qu'elle sous-estime largement le risque de survenue d'escarre. La sensibilité du BNClassifier dans la population FR est nettement meilleure que celle obtenue par l'échelle de Braden dans la population HR. Dans cette même population, la sensibilité à 0.99 et la spécificité à 0.01 du BNClassifier montrent qu'il capture essentiellement l'impression clinique : quasiment tous les cas identifiés par le personnel soignant le sont aussi par notre méthode. Ces résultats soulignent la pertinence des professionnels de santé, mais il faut noter que leurs facteurs déterminants ne sont pas connus et ne peuvent pas facilement être traduits en règles de décisions. Par ailleurs, dans la population FR, le BNClassifier identifie un grand nombre de vrais positifs qui semblent non identifiés par le personnel soignant, certainement par manque de temps ou de ressources.

	Échelle de Braden	BNClassifier (HR)	BNClassifier (FR)
Sensibilité	0.15	0.99	0.86
Spécificité	0.91	0.01	0.91

Tab. 5.6. : Sensibilité et Spécificité de l'échelle de Braden, du BNClassifier dans l'échantillon à haut risque (HR) et à faible risque (FR).

La principale limite de notre étude est liée au pourcentage plutôt faible de résidents évalués par une échelle de Braden, sélectionnés sur la base d'une impression clinique. De plus, nous ne pouvons exclure que chez les résidents avec un score de Braden positif, le personnel soignant ait commencé un protocole préventif pour éviter la survenue d'escarre, réduisant ainsi son incidence. Ainsi, une meilleure comparaison de l'efficacité de l'échelle de Braden et du BNClassifier pourrait être réalisée en évaluant le risque d'escarre par les deux méthodes sur des échantillons aléatoirement sélectionnés.

Pour évaluer le risque de survenue d'escarre, notre BNClassifier donne de meilleurs résultats que l'outil actuel de référence : l'échelle de Braden, qui a une sensibilité faible et sous-estime le risque d'escarre dans une grande proportion de résidents. Ainsi, 588 résidents avec une échelle de Braden qui ne donnait pas de risque important développaient une escarre dans le mois suivant.

Nous avons observé que l'impression clinique du personnel soignant menant à effectuer une échelle de Braden a une valeur prédictive intéressante, que le BNClassifier retrouve presque parfaitement de façon automatique, sans intervention du personnel soignant nécessaire. Notre méthode a aussi de bons résultats sur la population à plus faible risque, ce qui en fait un outil légitime d'aide à la décision pour la population à risque d'escarre. Le réseau bayésien capte donc la connaissance experte et son intérêt supplémentaire est qu'il peut l'appliquer sur toute la population automatiquement.

5.10 Application logicielle : NETSmart

Les résultats prometteurs dans la prédiction d'escarre en EHPAD ont rapidement voulu être implémentés par Teranga Software dans leur logiciel NETSmart, option spécialisée dans l'aide à la décision médicale de NETSoins.

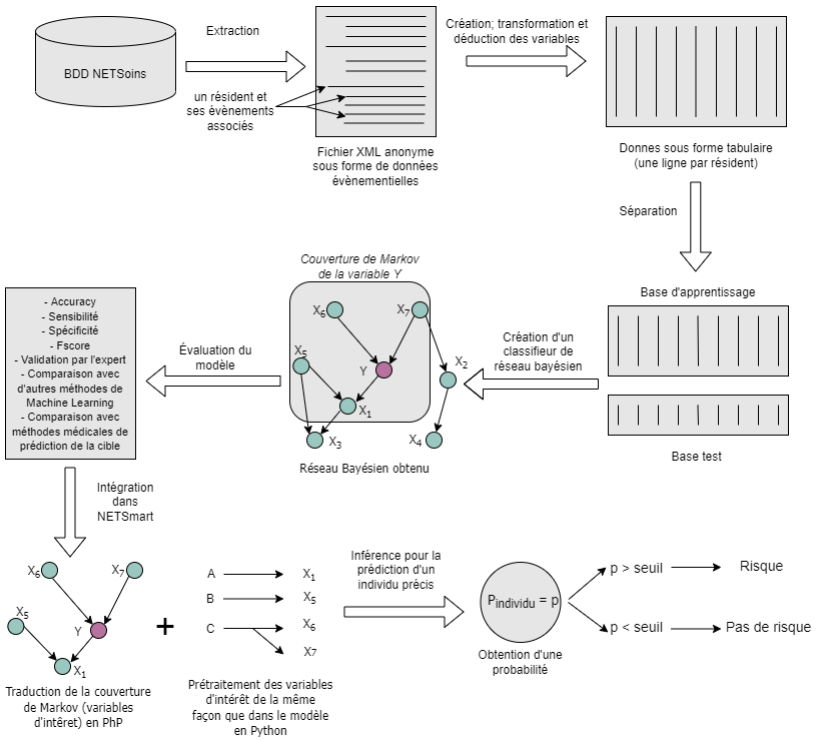


Fig. 5.20. : Pipeline allant jusqu'à l'implémentation dans NETSmart.

Plusieurs réseaux bayésiens ont donc été obtenus selon les modèles, nous avons donc exporté leur frontière de Markov.

Pour les intégrer, nous avons utilisé le module `metaGenBayes` qui permet de transcrire les calculs d'inférence probabiliste d'un réseau sous la forme de fonctions en PHP, le langage utilisé dans NETSoins. Ainsi, à partir de valeurs des observations, ces fonctions calculent les distributions des cibles. Ce module est capable de générer du code dans plusieurs langages de destination et est disponible sur le GitLab d'aGrUM/pyAgrum. Après une phase d'optimisation des équipes de Teranga, ce code a ainsi pu être intégré directement dans NETSmart. Avec ces fonctions, on obtient donc 3 probabilités associées aux 3 temporalités à partir d'un profil de résident. Si la probabilité dépasse le seuil défini, un risque d'escarre est annoncé dans le module NETSmart avec vue sur tous les résidents ou sur le tableau de bord d'un résident précis. La temporalité associée dépend du résident, s'il a un risque à 1 mois à afficher en priorité, sinon risque à 2 mois sinon risque à 3 mois.

NETSmart est un dispositif médical logiciel, de classe IIb selon le Règlement Européen 2017/745, et dans ce cadre, une investigation clinique doit avoir lieu avant sa commercialisation. C'est une solution de prédiction de l'escarre et d'aide à la décision dans les 4 thématiques suivantes :

- Escarre
- Dénutrition
- Risque iatrogénique
- Risque de chute

Sur la base des données des patients saisies dans NETSoins, NETSmart envoie des notifications aux utilisateurs, qui informent notamment de l'état de santé des patients dans les 4 thématiques précitées. Elles fournissent un niveau de risque actuel modéré à sévère dans ces 4 thématiques : les escarres, la dénutrition, le risque iatrogène et le risque de chutes, ainsi que la probabilité d'apparition d'escarres (i.e., probabilité de risque futur à 1, 2 et 3 mois). NETSmart propose des recommandations officielles (émanant principalement des guides de la Haute Autorité de Santé (HAS)) en lien avec le risque actuel ou futur décelé. Ces recommandations peuvent être des bonnes pratiques de suivi, des conseils de prise en charge médicale adaptée ou des conseils de traitements médicamenteux adaptés.

Deux investigations cliniques sont donc en cours. Une première investigation clinique avec recueil de données primaires a pour objectif d'évaluer l'utilisation de NETSmart par des professionnels de santé d'EHPADs français (médecins, infirmiers et aides-soignants) en démontrant la satisfaction des utilisateurs vis-à-vis de NETSmart et sa sécurité clinique. Une autre investigation clinique sur données secondaires est nécessaire, agissant comme validation externe de l'algorithme de la solution NETSmart pour la prédiction du risque d'escarre. Les données de validation externe indiqueront, pour chaque patient de la cohorte de validation, la probabilité de survenue de l'épisode d'escarre calculée par le modèle prédictif ainsi que la donnée réelle de survenue de l'escarre, et ce, pour chaque horizon temporaire (à 1, 2 et 3 mois après la date de prédiction). Les données individuelles pseudonymisées des participants seront extraites des EHPAD sélectionnés pour participer à l'étude et ayant accepté de participer et elles seront issues du logiciel de soins NETSoins. Les données des personnes seront extraites des établissements et traitées pour être disponibles au même format que celui sur lequel est utilisé l'algorithme.

Pour conclure, l'application sur les escarres a été la première cible à laquelle nous nous sommes intéressés et la plus concluante. Bien que les BNClassifier soient meilleurs lorsque la temporalité est plus proche, ils restent plus performants, selon nos critères préventifs, que les autres méthodes de classification testées, et que l'outil principalement utilisé actuellement pour prédire le risque d'escarre en EHPAD : l'échelle de Braden. De plus, ils repèrent et sélectionnent des variables qui sont des facteurs de risque d'escarre connu, sans qu'on ait apporté de connaissances

expertes aux modèles. La principale limite de notre analyse est de ne pas avoir accès à l'information de prévention d'escarre en cours sur le résident. Il est possible que des résidents que nous prédisons à risque aient finalement évité l'escarre, car les soignants ont aussi détecté ce risque et leur prise en charge préventive ont permis de l'éviter. Cependant, la possibilité fournie par Teranga Software de valider de façon externe les classifieurs de réseaux bayésiens et par la suite de les ajouter dans leurs logiciels est extrêmement innovante et permet d'espérer un outil de *machine learning* réellement utilisé en routine clinique, comme abordée dans le chapitre 3.

Références

ABDELLATIF, Abir (2021). "Amélioration des pratiques des soignants en EHPAD : développement et évaluation d'un assistant numérique intelligent". Thèse de doct. Sorbonne Université (cf. p. 115, 116).

AHN, Hyochol, Linda COWAN, Cynthia GARVAN, Debra LYON et Joyce STECHMILLER (2016). "Risk Factors for Pressure Ulcers Including Suspected Deep Tissue Injury in Nursing Home Facility Residents : Analysis of National Minimum Data Set 3.0". In : *Advances in Skin & Wound Care* 29.4, p. 178 (cf. p. 113, 114).

ALDERDEN, Jenny, Ginette Alyce PEPPER, Andrew WILSON et al. (2018). "Predicting Pressure Injury in Critical Care Patients : A Machine-Learning Model". In : *American Journal of Critical Care* 27.6, p. 461-468 (cf. p. 116).

BARROIS, Brigitte, Denis COLIN et François-André ALLAERT (2018). "Prevalence, characteristics and risk factors of pressure ulcers in public and private hospitals care units and nursing homes in France". In : *Hospital Practice (1995)* 46.1, p. 30-36 (cf. p. 113).

BELMIN, Joël, Philippe CHASSAGNE, Patrick FRIOCOURT et al. (2016). *Gériatrie : pour le Praticien*. Paris : Elsevier Health Sciences. 1071 p. (cf. p. 112).

BERGSTROM, Nancy, Barbara J. BRADEN, M. CHAMPAGNE, M. KEMP et E. RUBY (1998). "Predicting pressure ulcer risk : a multisite study of the predictive validity of study of the Braden Scale." In : *Nursing Research* 47.(5), p. 261-269 (cf. p. 115, 116).

BERLOWITZ, D. R., G. H. BRANDEIS, J. J. ANDERSON et al. (2001). "Evaluation of a risk-adjustment model for pressure ulcer development using the Minimum Data Set". In : *Journal of the American Geriatrics Society* 49.7, p. 872-876 (cf. p. 113).

BRADEN, Barbara J. et Nancy BERGSTROM (1994). "Predictive validity of the Braden Scale for pressure sore risk in a nursing home population". In : *Research in Nursing & Health* 17.6, p. 459-470 (cf. p. 116).

CHARON, Clara, Pierre-Henri WUILLEMIN et Joël BELMIN (2023). “Improving Pressure Ulcers Prediction in Nursing Homes with ML Algorithm”. In : *Studies in Health Technology and Informatics* 302, p. 350-351 (cf. p. 112).

– (2022). “Learning Bayesian Networks for the Prediction of Unfavorable Health Events in Nursing Homes”. In : *Studies in Health Technology and Informatics* 294, p. 147-148 (cf. p. 112).

CHARON, Clara, Pierre-Henri WUILLEMIN, Charlotte HAVRENG-THÉRY et Joël BELMIN (2024). “One Month Prediction of Pressure Ulcers in Nursing Home Residents with Bayesian Networks”. In : *Journal of the American Medical Directors Association*, S1525–8610(24)00070-7 (cf. p. 112).

CHEN, Hong-Lin, Wang-Qin SHEN et Peng LIU (2016). “A Meta-analysis to Evaluate the Predictive Validity of the Braden Scale for Pressure Ulcer Risk Assessment in Long-term Care”. In : *Ostomy/Wound Management* 62.9, p. 20-28 (cf. p. 116).

CHEN, Tianqi et Carlos GUESTRIN (2016). “XGBoost : A Scalable Tree Boosting System”. In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 785-794. arXiv : 1603.02754[cs] (cf. p. 120).

DEFLOOR, T. (1999). “The risk of pressure sores : a conceptual scheme”. In : *Journal of Clinical Nursing* 8.2, p. 206-216 (cf. p. 113, 114).

DUCAMP, Gaspard, Christophe GONZALES et Pierre-Henri WUILLEMIN (2020). “aGrUM/-pyAgrum : a toolbox to build models and algorithms for Probabilistic Graphical Models in Python”. In : *10th International Conference on Probabilistic Graphical Models*. T. 138. Proceedings of Machine Learning Research. Skørping, Denmark, p. 609-612 (cf. p. 119).

DWEEKAT, Odai Y., Sarah S. LAM et Lindsay MCGRATH (2023). “Machine Learning Techniques, Applications, and Potential Future Opportunities in Pressure Injuries (Bedsores) Management : A Systematic Review”. In : *International Journal of Environmental Research and Public Health* 20.1, p. 796 (cf. p. 132).

DWYER, Rosamond, Belinda GABBE, Johannes U. STOELWINDER et Judy LOWTHIAN (2014). “A systematic review of outcomes following emergency transfer to hospital for residents of aged care facilities”. In : *Age and Ageing* 43.6, p. 759-766 (cf. p. 122).

GEFEN, Amit (2008). “How much time does it take to get a pressure ulcer? Integrated evidence from human, animal, and in vitro studies”. In : *Ostomy/Wound Management* 54.10, p. 26-28, 30-35 (cf. p. 114).

HASTIE, Trevor, Robert TIBSHIRANI et Jerome FRIEDMAN (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY : Springer (cf. p. 120).

HAUTE AUTORITÉ DE SANTÉ, HAS (2001). *Prévention et traitement des escarres de l’adulte et du sujet âgé*. URL : https://www.has-sante.fr/jcms/c_271996/fr/prevention-et-traitement-des-escarres-de-l-adulte-et-du-sujet-age (cf. p. 112, 114).

HU, Ya-Han, Yi-Lien LEE, Ming-Feng KANG et Pei-Ju LEE (2020). “Constructing Inpatient Pressure Injury Prediction Models Using Machine Learning Techniques”. In : *CIN : Computers, Informatics, Nursing* 38.8, p. 415 (cf. p. 116).

KAEWPRAG, Pacharmon, Cheryl NEWTON, Brenda VERMILLION et al. (2015). “Predictive Modeling for Pressure Ulcers from Intensive Care Unit Electronic Health Records”. In : *AMIA Summits on Translational Science Proceedings* 2015, p. 82-86 (cf. p. 116).

LADIOS-MARTIN, Mireia, José FERNÁNDEZ-DE-MAYA, Francisco-Javier BALLESTA-LÓPEZ et al. (2020). “Predictive Modeling of Pressure Injury Risk in Patients Admitted to an Intensive Care Unit”. In : *American Journal of Critical Care* 29.4, e70-e80 (cf. p. 116).

LEE, Soo-Kyoung, Juh Hyun SHIN, Jinhyun AHN, Ji Yeon LEE et Dong Eun JANG (2021). “Identifying the Risk Factors Associated with Nursing Home Residents’ Pressure Ulcers Using Machine Learning Methods”. In : *International Journal of Environmental Research and Public Health* 18.6. Number : 6 Publisher : Multidisciplinary Digital Publishing Institute, p. 2954 (cf. p. 116).

LYDER, Courtney H. (2003). “Pressure ulcer prevention and management”. In : *JAMA* 289.2, p. 223-226 (cf. p. 113).

NORTON, Doreen, Rhoda MCLAREN et A.N. EXTON-SMITH (1962). *An investigation of geriatric nursing problems in hospital*. National Corporation for the Care of Old People. London (cf. p. 115).

PEDREGOSA, Fabian, Gaël VAROQUAUX, Alexandre GRAMFORT et al. (2011). “Scikit-learn : Machine Learning in Python”. In : *Journal of Machine Learning Research* 12.85, p. 2825-2830 (cf. p. 120).

SONG, Jie, Yuan GAO, Pengbin YIN et al. (2021). “The Random Forest Model Has the Best Accuracy Among the Four Pressure Ulcer Prediction Models Using Machine Learning Algorithms”. In : *Risk Management and Healthcare Policy* 14. Publisher : Dove Press, p. 1175-1187 (cf. p. 116).

SOUZA, Diba Maria Sebba Tosta de, Vera Lúcia Conceição de Gouveia SANTOS, Helena Keiko IRI et Miriam Yukiko SADASUE OGURI (2010). “Predictive validity of the Braden Scale for Pressure Ulcer Risk in elderly residents of long-term care facilities”. In : *Geriatric Nursing (New York, N.Y.)* 31.2, p. 95-104 (cf. p. 116).

SULLIVAN, Nancy et Karen M. SCHOELLES (2013). “Preventing In-Facility Pressure Ulcers as a Patient Safety Strategy : A Systematic Review”. In : *Annals of Internal Medicine* 158.5, p. 410 (cf. p. 116).

VANDERWEE, Katrien, Michael CLARK, Carol DEALEY, Lena GUNNINGBERG et Tom DEFLOOR (2007). “Pressure ulcer prevalence in Europe : a pilot study”. In : *Journal of Evaluation in Clinical Practice* 13.2, p. 227-235 (cf. p. 113).

6 Prédiction de l'hospitalisation en urgence à l'entrée du résident en EHPAD

Dans une seconde expérience, nous avons utilisé les mêmes méthodes pour essayer de prédire, grâce aux réseaux bayésiens, un deuxième événement de santé défavorable et évitable en EHPAD : l'hospitalisation en urgence des résidents. Nous définirons tout d'abord la cible et explorons les méthodes actuelles pour la prédire. Nous présenterons ensuite la base de données utilisée ici avec son prétraitement spécifique. Enfin, nous exposerons les résultats graphiques et numériques.

6.1 Contexte

Les passages aux urgences et les hospitalisations qui peuvent en découler sont des enjeux majeurs de santé publique pour les personnes âgées [Veyron et al., 2019 ; Kahn et al., 2016 ; Wolinsky et al., 2008]. En effet, une grande partie des personnes âgées se rendent aux urgences et subissent des hospitalisations non planifiées, et cette proportion augmente avec l'âge et la fragilité [Hwang et Morrison, 2007]. Les séjours aux urgences et les hospitalisations peuvent avoir un impact négatif sur la santé et l'autonomie des personnes âgées fragiles, avec des conséquences potentiellement à long terme [Belmin et al., 2022 ; Iwashyna et al., 2010 ; Chen et al., 2008].

Les situations d'urgence chez les personnes âgées découlent principalement de circonstances complexes, qui impliquent généralement un événement soudain, la crise, ainsi qu'un état de fragilité chronique résultant de l'accumulation de diverses affections fréquentes chez cette population, telles que les maladies et la fragilité sociale, psychologique ou socio-économique [Veyron et al., 2019 ; Paul Y. Takahashi et al., 2016].

Cependant, une grande proportion de ces visites aux urgences pourrait être évitée [Gasperini et al., 2017 ; Adams, 2013 ; Uscher-Pines et al., 2013]. Ainsi, la mise en place de stratégies pour identifier les patients à risque et leur permettre d'être traités en ambulatoire est une piste prometteuse ; et dans les cas où l'hospitalisation est nécessaire, une admission programmée dans un service adapté est une meilleure option qu'une hospitalisation après une visite aux urgences, tant que le patient ne se trouve pas dans une situation de danger vital [Hwang et Morrison, 2007].

6.2 État de l'art sur les outils de prédiction du risque

À notre connaissance, aucune échelle de risque n'est actuellement utilisée en routine clinique pour les hospitalisations en urgence spécifique aux personnes âgées en EHPAD. Des méthodes de *machine learning* ont cependant été utilisées pour développer des outils de prédictions de risque d'urgence chez les personnes âgées [Belmin et al., 2022 ; Shelton et al., 2000 ; Gasperini et al., 2017 ; Paul Y. Takahashi et al., 2016 ; Crane et al., 2010]. Ces études ont d'ailleurs identifié des facteurs de risques de passage en urgence chez les personnes âgées comme :

- les comorbidités et la prise de nombreux médicaments [Shelton et al., 2000 ; Gasperini et al., 2017 ; Crane et al., 2010] ;
- une hospitalisation antérieure [Shelton et al., 2000 ; Crane et al., 2010 ; Paul Y. Takahashi et al., 2016] ;
- un âge avancé [Gasperini et al., 2017 ; Crane et al., 2010] ;
- le fait d'être veuf ou séparé [Gasperini et al., 2017 ; Crane et al., 2010].

Il est à noter qu'aucune de ces études ne se concentre ou n'a été validée sur des résidents en EHPAD.

6.3 Spécificités de prétraitement

Les hospitalisations sont des événements fréquents au cours du séjour en EHPAD. Plus d'un résident sur deux de notre base de données en subira au moins une. Parmi ces hospitalisations, certaines seront programmées et certaines auront lieu suite à une urgence. Ce sont ces dernières qui nous intéressent ici.

Dans le logiciel NETSoins, les hospitalisations font partie des situations administratives qui sont saisies et qui peuvent être utiles pour la facturation. Plusieurs types de mouvements sont associés à l'hospitalisation en urgence : l'hospitalisation, le passage aux urgences et l'hospitalisation suite à un passage aux urgences. Les hospitalisations programmées ne sont pas reconnaissables ici, et sont mélangées dans les hospitalisations "générales" avec probablement quelques hospitalisations qui ont été en urgences, mais dont cela n'a pas été spécifié. Nous avons donc décidé d'utiliser ici comme cible uniquement les labels "urgences" et "hospitalisations suite urgence" que nous avons fusionnés sous le label "hospitalisation_urgence".

À l'origine, nous voulions pouvoir prédire ces urgences à court terme, soit dans les 14 jours. Seulement, ces événements peuvent arriver plusieurs fois à une même

personne à différentes temporalités. Cela posait donc des questions de gestion des données pour ces cas précis. Il faut déjà pouvoir identifier et séparer temporellement de façon précise les différentes urgences d'un même résident. Nous nous sommes ainsi questionnés sur la méthodologie permettant de les faire figurer dans une base de données tabulaire. Il est possible de dupliquer les résidents à différentes temporalités, mais cela peut augmenter considérablement le nombre de cas positifs de la base de données et biaiser en augmentant artificiellement le poids des résidents qui auraient eu beaucoup d'hospitalisations en urgence. Des méthodes de compensation de cas témoins ont été envisagés, mais aucune solution satisfaisante n'a encore été trouvée.

Nous avons donc voulu essayer de mettre en avant les résidents qui ont eu au moins une hospitalisation en urgence dans leur séjour en EHPAD, peu importe le nombre et de les mettre en opposition à ceux qui n'en ont jamais eu. Pour cela, prédire le risque d'un résident de faire une urgence pendant son séjour avec les données à entrée dans l'établissement semblait pertinent d'un point de vue clinique.

Nous avons donc utilisé le marqueur temporel à 18 jours après l'admission. Les variables calculant des évolutions temporelles sur trois mois et six mois ont donc été enlevées. Nous avons aussi observé qu'au bout de 18 jours en EHPAD, 84 % des résidents n'avait pas encore d'évaluation PATHOS effectuée, contrairement à l'évaluation AGGIR qui avait été faite chez 59 % des résidents. Cela n'est pas très étonnant, l'évaluation PATHOS est généralement faite pour tous les résidents en même temps de façon annuelle, et est donc difficilement exploitable à cette temporalité. Toutes les variables liées à cette évaluation ont donc été supprimées pour cette cible. Cependant, nous avons gardé uniquement les résidents avec un niveau GIR, mais nous n'avons pas effectué d'autre tri sur les lignes avec des valeurs manquantes.

Nous obtenons donc une base de 114 variables et de 64 547 lignes avec 17 120 cas positifs d'urgence, soit 26.52 %. Une comparaison des caractéristiques démographiques a été effectuée entre les deux populations en table 6.1. En ce qui concerne l'âge et le sexe, les deux populations étudiées sont similaires et présentent une certaine homogénéité. On remarque par contre que les cas témoins avec une dépendance très sévère sont bien plus importants que chez les personnes qui ont eu une urgence.

Variable	Cas urgence (n = 17 120)	Cas témoins (n = 47 427)
Âge, moyenne (écart-type)	86.36 (7.39)	85.82 (7.70)
Pourcentage de femmes	68.38 %	70.89 %
Niveau de dépendance :		
GIR1 (très sévère)	3.90 %	10.52 %
GIR2 (sévère)	34.48 %	35.71 %
GIR3 (modéré/sévère)	23.33 %	19.25 %
GIR4 (modéré)	30.86 %	25.36 %
GIR5 (léger)	4.99 %	5.04 %
GIR6 (pas de dépendance)	2.44 %	4.11 %

Tab. 6.1. : Caractéristiques démographiques des cas d'hospitalisation en urgence et des cas témoins de la base de données utilisée

6.4 Résultats graphiques

La frontière de Markov obtenue par BNClassifier sur le jeu de données d'apprentissage est disponible en figure 6.1. Elle est de grande taille, une vingtaine de variables y figurent. Nous avons donc effectué une analyse des valeurs de Shapley et les variables dont la moyenne de leurs valeurs de Shapley sont les plus importantes sont présentées en figure 6.2 par ordre d'importance, du haut vers le bas.

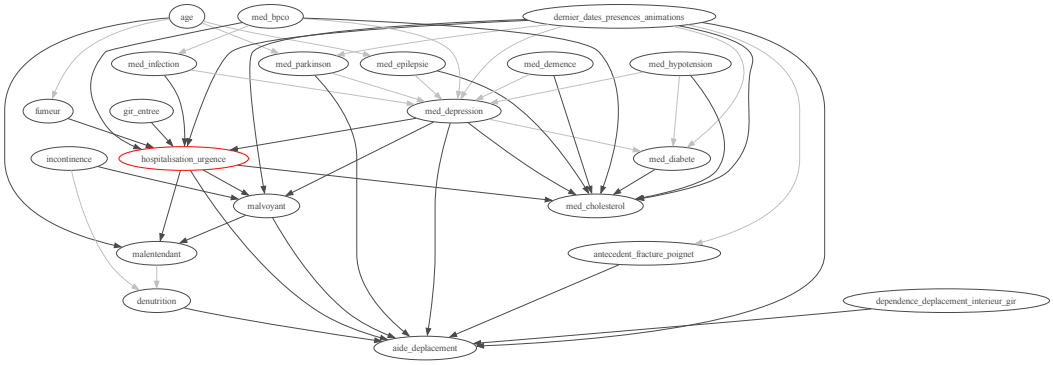


Fig. 6.1. : Frontière de Markov de la cible hospitalisation en urgence du BNClassifier. En gris, les arcs présents dans le réseau bayésien inutilisés dans la construction de la frontière de Markov.

Le niveau GIR global à l'entrée est donc la variable la plus importante, mais c'est un niveau GIR modéré (3-4) qui augmente ici le risque, comme ce qui a été observé dans les caractéristiques démographiques de la table 6.1.

On peut donc s'imaginer qu'il s'agit de la population qui n'est pas la plus surveillée, mais dont l'état de santé peut rapidement se dégrader. Le fait d'avoir des maladies chroniques est représenté ici par les médicaments contre la bronchopneumopathie chronique obstructive (BCPO), le cholestérol et la dépression. Le besoin d'assistance dans les déplacements, le fait d'être malvoyant, malentendant et fumeur sont aussi des facteurs aggravants qui sont signe de fragilité. De même, la prise de médicaments contre les infections est un facteur de risque présent dans le modèle.

On retrouve bien dans ces variables l'idée que l'hospitalisation en urgence est liée à une crise (une infection par exemple), ainsi qu'un état de fragilité chronique. Par ailleurs, on peut remarquer que ce sont les médicaments comme marqueur de pathologie qui sont pertinents lorsque l'évaluation PATHOS n'est pas disponible.

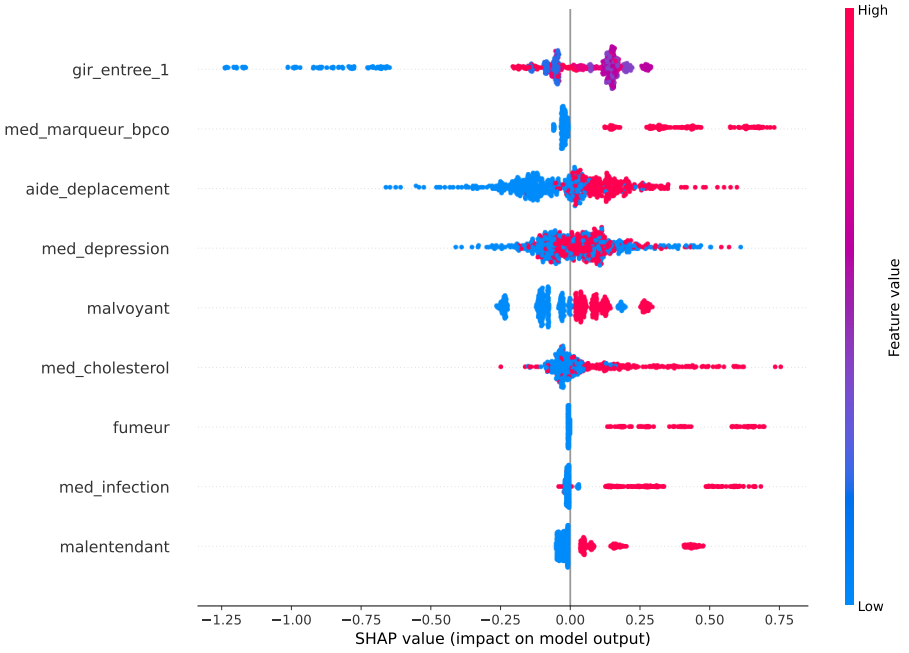


Fig. 6.2. : Valeurs de Shapley des variables de la frontière de Markov du BNClassifier pour la cible hospitalisation en urgence avec les valeurs de Shapley moyennes les plus importantes.

6.5 Résultats numériques

Au vu des résultats de la courbe ROC et précision-rappel en figure 6.3 sur la base d'apprentissage, nous avons cette fois utilisé le seuil qui maximise le F1-score. En effet, il permet d'obtenir un rappel de 0.80 et une précision de 0.40 et pour

augmenter légèrement le rappel, une baisse importante de précision aurait dû avoir lieu. L'AUC de la courbe ROC est ici de 0.71.

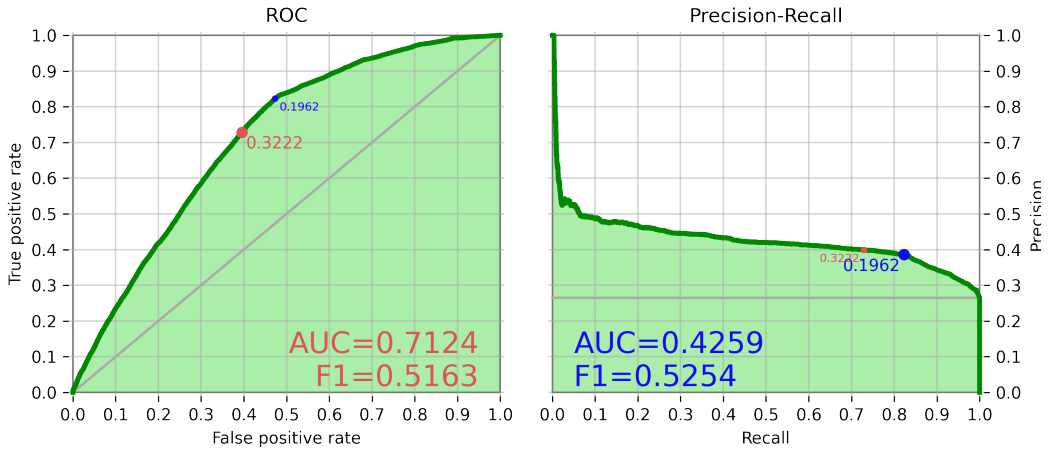


Fig. 6.3. : Courbe ROC précision-rappel du BNClassifier pour la prédiction d'hospitalisation en urgence (le point rouge correspond au seuil minimisant la distance au point idéal de la courbe ROC, et le point bleu, celui maximisant le F1-score).

Nous avons donc testé ce BNClassifier avec le seuil de 0.1962 sur des données de test et avons comparé les résultats avec les mêmes méthodes de *machine learning* précédemment utilisées. Nous avons répété 10 fois l'expérience et les moyennes des résultats sont présentées en table 6.2.

	F2-score	F1-Score	Précision/ VPP	Rappel/ Sensibilité	Accuracy
BNClassifier	**0.66**	**0.52**	0.38	0.80	0.60
QDA	0.63	0.45	0.30	**0.88**	0.42
Naive Bayes	0.50	0.46	0.40	0.53	0.66
Arbre de décision	0.38	0.37	0.36	0.38	0.65
MLP	0.29	0.33	0.40	0.28	0.70
XgBoost	0.22	0.27	**0.45**	0.20	0.72
AdaBoost	0.17	0.22	0.44	0.14	0.72
Régression Logistique	0.15	0.20	**0.45**	0.13	**0.73**

Tab. 6.2. : F2-score et autres métriques évaluant les performances des différents modèles pour la prédiction d'hospitalisation en urgence chez les résidents en EHPAD.

Le choix du seuil optimisant le F1-score ne permet pas au BNClassifier d'obtenir le meilleur rappel, détenus ici par QDA, mais il obtient une précision supérieure. Les scores synthétiques de ces deux métriques, le F1-score et le F2-score, montre que BNClassifier est le meilleur compromis parmi toutes les méthodes présentées.

Les stratégies actuelles de prédiction d'hospitalisation en urgence chez les personnes âgées sont rares, et aucune spécifique au contexte des EHPAD ne semble exister. À partir de notre base de données de grande taille, nous avons pu développer un classifieur à base de réseau bayésien qui permet de prédire le risque à l'entrée du résident en EHPAD. Les variables sélectionnées par le modèle sont pertinentes et les résultats numériques prometteurs. Encore une fois, nous avons un modèle plus sensible que spécifique, ce qui reste pertinent dans un contexte de prévention, et avec un meilleur F2-score que les autres méthodes de classification testées. Une publication sur ce sujet est en cours de finalisation.

Références

Adams, James G. (2013). "Emergency Department Overuse : Perceptions and Solutions". In : *JAMA* 309.11, p. 1173-1174 (cf. p. 144).

Belmin, Joël, Patrick Villani, Mathias Gay et al. (2022). "Real-world Implementation of an eHealth System Based on Artificial Intelligence Designed to Predict and Reduce Emergency Department Visits by Older Adults : Pragmatic Trial". In : *Journal of Medical Internet Research* 24.9, e40387 (cf. p. 144, 145).

Chen, Cheryl Chia-Hui, Charlotte Wang et Guan-Hua Huang (2008). "Functional trajectory 6 months posthospitalization : a cohort study of older hospitalized patients in Taiwan". In : *Nursing Research* 57.2, p. 93-100 (cf. p. 144).

Crane, Sarah J., Ericka E. Tung, Gregory J. Hanson et al. (2010). "Use of an electronic administrative database to identify older community dwelling adults at high-risk for hospitalization or emergency department visits : The elders risk assessment index". In : *BMC Health Services Research* 10.1, p. 338 (cf. p. 145).

Gasperini, Beatrice, Antonio Cherubini, Francesca Pierri et al. (2017). "Potentially preventable visits to the emergency department in older adults : Results from a national survey in Italy". In : *PLOS ONE* 12.12. Publisher : Public Library of Science, e0189925 (cf. p. 144, 145).

Hwang, Ula et R. Sean Morrison (2007). "The geriatric emergency department". In : *Journal of the American Geriatrics Society* 55.11, p. 1873-1876 (cf. p. 144).

Iwashyna, Theodore J, E Wesley Ely, Dylan M Smith et Kenneth M Langa (2010). "Long-term cognitive impairment and functional disability among survivors of severe sepsis". In : *JAMA* 304.16, p. 1787-1794 (cf. p. 144).

Kahn, Joseph H., Brendan G. Magauran, Jonathan S. Olshaker et Kalpana N. Shankar (2016). "Current Trends in Geriatric Emergency Medicine". In : *Emergency Medicine Clinics of North America*. Geriatric Emergencies 34.3, p. 435-452 (cf. p. 144).

Paul Y. Takahashi, M. D., M. S. Herbert C. Heien, M. P. H. Lindsey R. Sangaralingham, PhD Nilay D. Shah et ScD James M. Naessens (2016). “Enhanced Risk Prediction Model for Emergency Department Use and Hospitalizations in Patients in a Primary Care Medical Home”. In : July 2016 22. Publisher : MJH Life Sciences (cf. p. 144, 145).

Shelton, Paul, Mark A Sager et Cheryl Schraeder (2000). “The Community Assessment Risk Screen (CARS) : Identifying Elderly Persons at Risk for Hospitalization or Emergency Department Visit”. In : *THE AMERICAN JOURNAL OF MANAGED CARE* 6.8 (cf. p. 145).

Uscher-Pines, Lori, Jesse Pines, Arthur Kellermann, Emily Gillen et Ateev Mehrotra (2013). “Emergency department visits for nonurgent conditions : systematic literature review”. In : *The American Journal of Managed Care* 19.1, p. 47-59 (cf. p. 144).

Veyron, Jacques-Henri, Patrick Friocourt, Olivier Jeanjean et al. (2019). “Home care aides’ observations and machine learning algorithms for the prediction of visits to emergency departments by older community-dwelling individuals receiving home care assistance : A proof of concept study”. In : *PLOS ONE* 14.8. Publisher : Public Library of Science, e0220002 (cf. p. 144).

Wolinsky, Fredric D., Li Liu, Thomas R. Miller et al. (2008). “Emergency Department Utilization Patterns Among Older Adults”. In : *The Journals of Gerontology : Series A* 63.2, p. 204-209 (cf. p. 144).

7 Prédiction de fracture à partir des premiers mois dans l'établissement

Dans une troisième analyse, nous nous sommes intéressés à un dernier événement de santé défavorable et évitable : les fractures. Nous expliquerons tout d'abord plus précisément cette cible et quels outils existent actuellement pour prédire ce risque. Puis, nous exposerons le prétraitement spécifique appliqué ainsi que nos résultats. Enfin, ces résultats seront comparés avec une de ces méthodes validées de prédiction de risque.

7.1 Contexte médical

Une fracture est une rupture ou une cassure d'un os. Les fractures représentent une cause importante d'invalidité motrice chez les personnes âgées, et peuvent survenir en l'absence de traumatisme, ou après un traumatisme minime [Belmin, Chassagne et Friocourt, 2023].

L'ostéoporose est une maladie du tissu osseux qualitative et quantitative qui a pour conséquence une fragilité excessive du squelette et donc un risque associé de fracture, dont la redoutable fracture du col du fémur.

Les quatre principaux facteurs de risque de fractures sont : l'âge, la densitométrie osseuse montrant un seuil fracturaire significativement élevé, un antécédent de fracture ostéoporotique (telle une fracture du col du fémur) et la propension aux chutes [Belmin, Chassagne, Friocourt et al., 2016].

La densitométrie est un examen médical qui permet de mesurer la densité minérale osseuse (DMO), c'est-à-dire la quantité de calcium dans les os. C'est une méthode peu invasive qui mesure la quantité de tissu osseux, évalue le seuil fracturaire et peut estimer la probabilité dans le temps de survenue de fracture. Seulement, ces examens ne sont pas encore assez répandus, peu remboursés et non présents dans le dossier résident de NETSoins.

La fracture étant souvent une conséquence de l'ostéoporose, il peut donc être intéressant d'essayer de la prédire. Les facteurs favorisant l'ostéoporose sont [Belmin, Chassagne, Friocourt et al., 2016] :

- Âge avancé
- Ménopause
- Apports en calcium insuffisants
- Déficit en vitamine D
- Tabagisme

- Sédentarité
- Maigreur, dénutrition protéinoénergétique
- Facteurs génétiques
- Immobilisation prolongée
- Certains médicaments : corticostéroïdes, héparine au long cours, analogues de la LH-RH, phénobarbital, phénytoine, inhibiteurs de la pompe à protons, antiaromatases, anti-androgènes
- Maladies endocriniennes : hyperthyroidie, maladie de Cushing, hyperparathyroïdie primitive, insuffisance androgénique
- Polyarthrite rhumatoïde et autres rhumatismes inflammatoires

Certains facteurs valent la peine d'être aussi mentionnés, bien qu'ils n'accroissent pas le risque d'ostéoporose, mais celui de chute pouvant mener à une fracture :

- Alcoolisme
- Baisse de l'acuité visuelle
- Troubles neuromusculaires et/ou orthopédiques

Une prise en charge préventive pour éviter l'ostéoporose et ainsi la fracture est possible, mais est longue à mettre en place. L'activité physique, notamment, contribue à réduire la diminution liée à l'âge de la densité osseuse et a des effets positifs sur la fonction musculaire et la mobilité. Des programmes spécifiques d'exercices physiques adaptés aux personnes ostéoporotiques sont d'ailleurs disponibles.

7.2 Outils actuels de prédiction du risque

Les outils de prédiction de risque de fracture sont des instruments de plus en plus utilisés en clinique pour évaluer le risque de fracture chez les patients. Parmi les outils les plus couramment utilisés, on retrouve QFracture et FRAX.

QFracture est un outil de prédiction de risque de fracture développé au Royaume-Uni [Hippisley-Cox et Coupland, 2012]. Il prend en compte plusieurs facteurs de risque, tels que l'âge, le sexe, l'indice de masse corporelle, l'historique de tabagisme et de consommation d'alcool, les antécédents de fractures, les comorbidités et l'utilisation de certains médicaments (voir figure 7.1). Il permet d'estimer deux pourcentages : le risque de fracture de la hanche et celui de fracture de la colonne vertébrale, du poignet, de l'épaule et de la hanche sur une période de 1 à 10 ans.

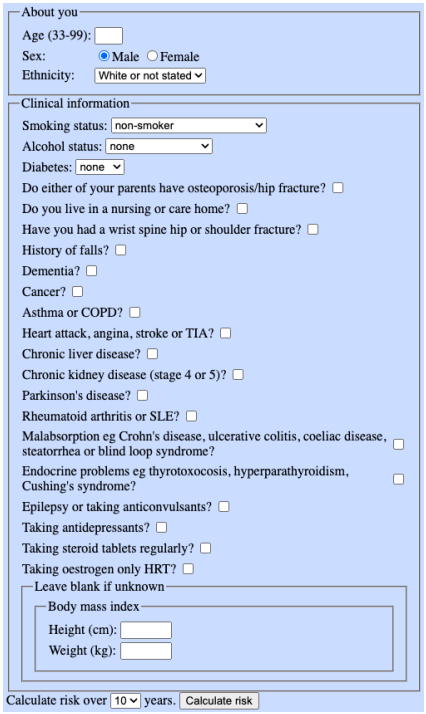


Fig. 7.1. : Vue de l'outil QFracture.

L'algorithme a été conçu pour estimer le risque absolu de fracture ostéoporotique et de fracture de la hanche dans le cadre des soins primaires. L'algorithme est basé sur des variables qui sont facilement disponibles dans les dossiers électroniques des patients ou que les patients eux-mêmes connaîtraient probablement sans avoir besoin de tests de laboratoire ou de mesures cliniques. Cette approche veut permettre aux algorithmes d'être facilement mis en œuvre dans la pratique clinique de routine ou utilisés par des patients individuels. Les modèles sont différents pour les hommes et les femmes.

Ce sont des régressions de Cox [Cox, 1972] qui sont utilisées sur l'ensemble de données pour estimer les coefficients et les rapports de risque associés à chaque facteur de risque. Le modèle de Cox permet d'estimer l'effet de différentes variables $X_1, \cdots, X_n$ sur le risque instantané de survenance de l'événement d'intérêt λ , en fonction du temps t écoulé depuis le début de l'observation. Il s'exprime sous la forme suivante :

$$\lambda\left(t, X_1, \cdots, X_n\right) = \lambda_0\left(t\right) \exp\left(\Sigma_{i=1}^{n} \beta_i X_i\right)$$

FRAX (*Fracture Risk Assessment Tool*) est un autre outil de prédiction de risque de fracture largement utilisé [Kanis, Johnell et al., 2001]. Il a été développé par l'Université de Sheffield et prend en compte des facteurs de risque similaires à ceux

de QFracture, tels que l'âge, le sexe, l'indice de masse corporelle, les antécédents de fractures, les comorbidités. FRAX permet d'estimer selon le pays deux probabilités distinctes : le risque de fracture de la hanche et de fractures ostéoporotiques majeures uniquement sur une période de 10 ans. Il est basé sur des modèles individualisés et des régressions qui intègrent les risques associés aux facteurs de risque cliniques ainsi que la densité minérale osseuse (DMO) au niveau du col du fémur si elle est disponible [KANIS, HARVEY et al., 2017].

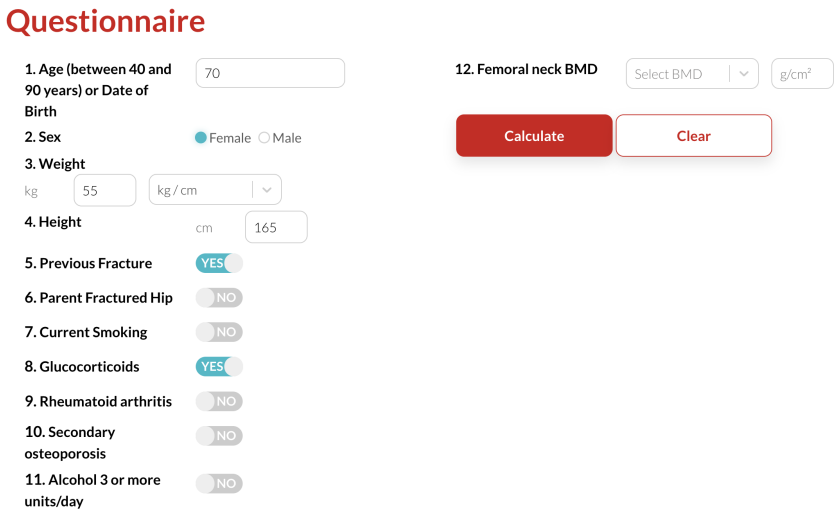


Fig. 7.2. : Vue de l'outil FRAX.

D'après une récente revue systématique [MARQUES et al., 2015] sur les outils de prédiction de fracture ostéoporotique, FRAX a le plus grand nombre de validations externes et d'études indépendantes. Bien que la méta-analyse précise que des limites méthodologiques ont été observées dans certaines études et que comparer ces outils uniquement par leur AUC doit être fait avec prudence, l'AUC global de FRAX est annoncé à 0.79 (95 % IC 0.73 à 0.85). Celle de QFracture atteint 0.89 (95 % IC 0.88 à 0.89). Toutes ces analyses se font avec une prédiction de risque sur 10 ans.

7.3 Spécificités de prétraitement

Nous avons initialement voulu nous concentrer sur le risque de fracture du col du fémur. Toutefois, nous nous sommes rendu compte que les fractures étaient mal saisies dans le logiciel NETSoins. En effet, en table 7.1 se trouvent le nombre de cas de fractures par type dans la base de données exportée. Nous avons effectué une recherche par mot-clé dans le dossier médical. Ainsi, les cas de fracture du

col du fémur se retrouvent aussi dans les cas de fracture du fémur. Au vu du peu de nombre de cas, nous avons cherché à retrouver l'information des fractures dans les comptes-rendus d'hospitalisation, mais ceux-ci sont généralement des documents PDF sur lequel il était impossible de faire des requêtes en restant certain de l'anonymisation.

Type	Nombre
Fracture col du fémur	1 376
Fracture fémur	2 212
Ostéosynthèse fémur	143
Fracture poignet	995
Fracture hanche	464
Fracture épaule	274
Fracture colonne vertébrale	227

Tab. 7.1. : Nombre de cas de fractures par type

Nous avons donc décidé de créer plutôt un algorithme qui prédirait le risque de fracture globale, de la même façon que FRAX et QFracture. Nous avons ainsi un nombre de cas positifs plus important, bien qu'il reste bien plus faible que les cibles précédentes.

Dans notre contexte, une prédiction sur 10 ans est difficilement exploitable, les personnes âgées en EHPAD étant déjà âgées et ne restant en moyenne que quelques années. Mais une prédiction de fracture doit être bien anticipée pour être évitable. Nous avons donc opté pour une prédiction du risque à l'entrée du résident pour tout son séjour en EHPAD. Afin de disposer d'un certain recul sur les données, nous avons utilisé le marqueur temporel de "3 mois et demi" pour définir la période de référence pour les cas témoins et les cas de fracture. Les cas de fracture sont les résidents qui ont souffert d'au moins une fracture avec un des intitulés de la table 7.1 durant leur séjour en EHPAD. Par ailleurs, les cas de fracture survenant avant le marqueur temporel sont supprimés.

Étant donné le faible nombre de cas de fractures, nous avons dû limiter la suppression des lignes avec des valeurs manquantes durant le prétraitement des données. Nous avons donc inclus tous les résidents ayant au moins 70 valeurs présentes sur les 140. Cela nous permet finalement d'obtenir 74 825 résidents dont 3 791 cas de fractures soit 5.06% de cas positifs. Par rapport au total de la table 7.1, nous avons 1 900 cas de fractures en moins. Cette différence peut s'expliquer par le fait que certaines fractures ont eu lieu avant le marqueur temporel, que plusieurs fractures

appartenaient au même résident ou que la moitié des variables du résident étaient manquantes. Cependant, une analyse au Danemark sur 2 601 personnes âgées en institution reportait 3.80% de cas de fractures de fémur entre 2018 et 2019 [SARWARI et al., 2024].

Variable	Cas fracture (n = 3 791)	Cas témoins (n = 71 034)
Âge, moyenne (écart-type)	85.04 (7.91)	84.41 (8.71)
Pourcentage de femmes	85.04 %	72.27 %
Niveau de dépendance :		
GIR1 (très sévère)	6.92 %	5.83 %
GIR2 (sévère)	33.73 %	34.35 %
GIR3 (modéré/sévère)	19.53 %	20.79 %
GIR4 (modéré)	29.45 %	28.72 %
GIR5 (léger)	6.08 %	6.45 %
GIR6 (pas de dépendance)	4.28 %	3.86 %

Tab. 7.2. : Caractéristiques démographiques des cas de fracture et des cas témoins de la base de données utilisée

En table 7.2, nous avons comparé les caractéristiques démographiques entre les cas positifs et les cas témoins. On remarque qu'au niveau de l'âge et de la dépendance, les groupes sont homogènes. Par contre, les femmes ont nettement plus de fractures, ce qui se vérifie par ailleurs [THÉLOT et al., 2017].

7.4 Résultats

Nous avons appliqué les mêmes méthodes que précédemment et utilisé un BN-Classifier appris avec la méthode MIIC et un seuil maximisant le F2-score.

7.4.1 Résultats graphiques

La frontière de Markov de niveau 1 en figure 7.3 est obtenue. L'analyse des valeurs de Shapley en figure 7.4 nous indique l'ordre d'importance des variables présentes dans la frontière de Markov et le sens de leur impact. Les variables avec une valeur de Shapley moyenne moins importante ne sont pas représentées.

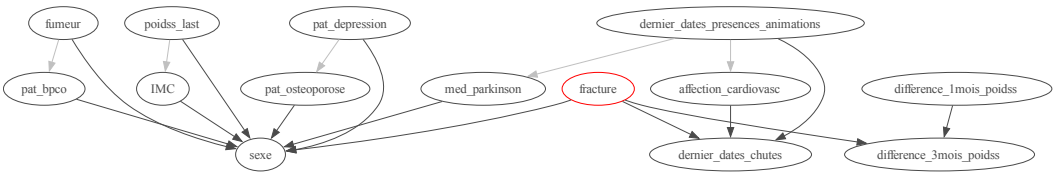


Fig. 7.3. : Frontière de Markov de la cible fracture du BNClassifier. En gris, les arcs présents dans le réseau bayésien inutilisés dans la construction de la frontière de Markov.

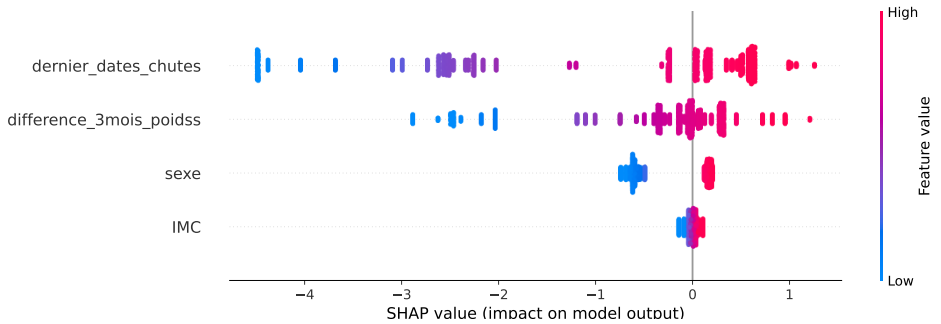


Fig. 7.4. : Valeurs de Shapley des variables de la frontière de Markov du BNClassifier pour la cible fracture.

Ainsi, le sexe est important pour prédire le risque, et c'est le fait d'avoir la valeur "1" (donc d'être une femme) qui l'augmente. Une ou plusieurs chutes récentes sont aussi des facteurs de risque attendu. Une grande différence de poids pourrait indiquer une dénutrition et ainsi faire appel à un autre facteur de risque, mais l'analyse sur l'IMC ne permet pas d'obtenir plus d'informations. Les variables fumeur et ostéoporose, bien qu'avec des valeurs de Shapley moyennes faibles, sont présentes dans la frontière de Markov et sont des facteurs de risques connus. Les variables sélectionnées par le modèle à 3 mois et demi de présence en EHPAD sont donc très pertinentes.

7.4.2 Résultats numériques

La figure 7.5 nous montre la courbe ROC et PR calculés sur l'échantillon d'apprentissage. L'AUC de la courbe ROC est à 0.79. Ici, sélectionner comme seuil le point optimal de la courbe ROC et non pas celui qui optimise le F2-score pourrait être intéressant. Mais on remarque que la précision est déjà très faible et le point optimisant le F2-score (en bleu sur la figure) nous permet de le garder au plus haut rappel possible pour cette précision.

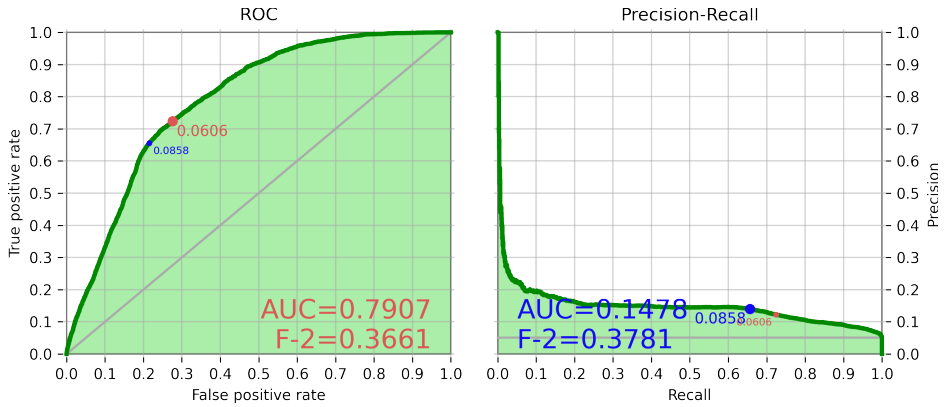


Fig. 7.5. : Courbes ROC et précision-rappel du BNClassifier pour la prédiction de fracture (le point rouge correspond au seuil minimisant la distance au point idéal de la courbe ROC, et le point bleu, celui maximisant le F2-score).

	F2-score	F1-Score	Précision/ VPP	Rappel/ Sensibilité	Accuracy
BNClassifier	**0.31**	**0.17**	0.10	0.62	0.70
Naive Bayes	0.26	0.13	0.07	0.80	0.48
QDA	0.24	0.12	0.06	**0.89**	0.33
Arbre de décision	0.16	0.15	0.14	0.17	0.90
MLP	0.12	0.14	0.20	0.11	0.93
XgBoost	0.10	0.14	0.58	0.08	**0.95**
Régression Logistique	0.01	0.02	**0.61**	0.01	**0.95**
AdaBoost	0.00	0.0	0.39	0.00	**0.95**

Tab. 7.3. : F2-score et autres métriques évaluant les performances des différents modèles pour la prédiction de fracture chez les résidents en EHPAD.

Nous avons répété 10 fois l'expérience de validation avec des extractions aléatoires de sous-base d'apprentissage testées sur leur base complémentaire. Les moyennes des scores de BNClassifier ainsi que les autres méthodes habituelles sur le jeu de données de test sont présentes en table 7.3. Le rappel de BNClassifier reste bon, mais la précision est faible. C'est tout de même la meilleure méthode de classification en termes de rappel, F1-score et de F2-score. Cependant, XgBoost et la régression logistique ont des précisions intéressantes, malgré un rappel faible. En analysant les matrices de confusion, on observe que les méthodes avec une *accuracy* à 0.95 % prédisent systématiquement moins de 1 % de cas positifs. BNClassifier obtient une *accuracy* intéressante de 0.70 %, tout en prédisant correctement la plupart des cas de fractures.

Une étude australienne [ESHETIE et al., 2024] a étudié le risque de fracture en

établissement de soins de longue durée et a créé un outil de prédiction de fracture sur une population avec 7.2% de cas de fracture. Le modèle obtient une AUC de 0.62 en moyenne sur l'échantillon test. Ici, l'AUC moyenne de BNClassifier sur les échantillons tests était de 0.72. Mais, l'interprétation de l'AUC dans ce contexte de classes très déséquilibrées reste délicate.

7.4.3 Séparation selon sexe

De même, nous avons essayé de créer des modèles différents selon les sexes, comme QFracture et FRAX. Les résultats sont présentés ici. Ils sont moins bons en termes de scores que le modèle général, nous ne les avons donc pas utilisés. Leurs frontières de Markov sont légèrement différentes selon les sexes, mais on ne retrouve pas *a priori* de facteurs de risques spécifiques au genre.

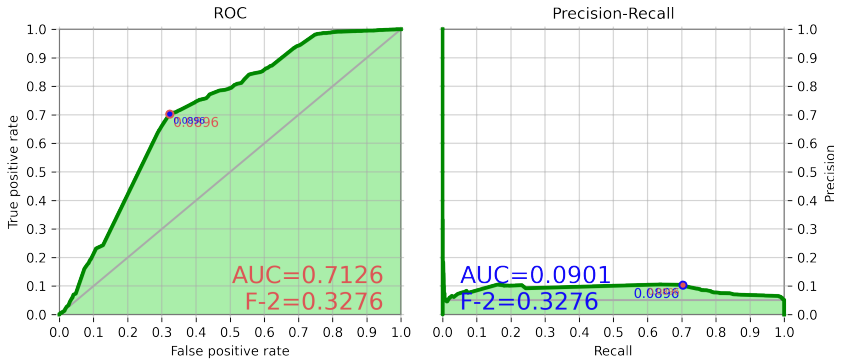


Fig. 7.6. : Courbes ROC et précision-rappel du BNClassifier pour la prédiction de fracture chez les femmes (le point rouge correspond au seuil minimisant la distance au point idéal de la courbe ROC, et le point bleu, celui maximisant le F2-score).

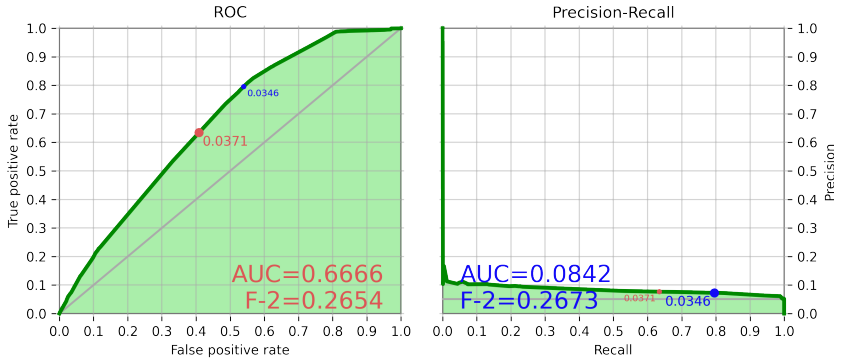


Fig. 7.7. : Courbes ROC et précision-rappel du BNClassifier pour la prédiction de fracture chez les hommes (le point rouge correspond au seuil minimisant la distance au point idéal de la courbe ROC, et le point bleu, celui maximisant le F2-score.

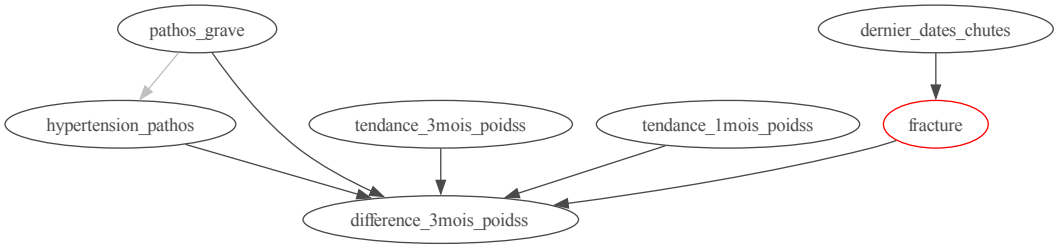


Fig. 7.8. : Frontière de Markov de la cible fracture du BNClassifier chez les femmes. En gris, les arcs présents dans le réseau bayésien inutilisés dans la construction de la frontière de Markov.

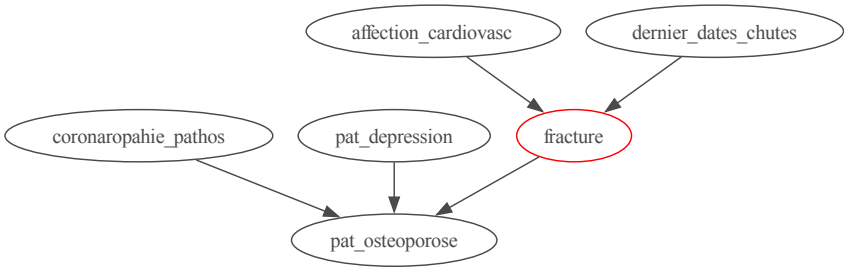


Fig. 7.9. : Frontière de Markov de la cible fracture du BNClassifier chez les hommes. En gris, les arcs présents dans le réseau bayésien inutilisés dans la construction de la frontière de Markov.

7.5 Comparaison avec QFracture

FRAX et QFracture sont des évaluations accessibles en ligne, mais elles ne sont pas intégrées dans le logiciel NETSoins. Le code du modèle QFracture est disponible en *open source*, nous avons donc pu implémenter ses fonctions en Python et l'appliquer sur nos données. Nous avons pu vérifier que notre implémentation donnait les mêmes résultats que l'outil en ligne. Les fonctions et les facteurs de risques sont différents pour les hommes et les femmes, nous les avons donc séparés pour notre analyse. Nous avons ici calculé les pourcentages de risque à 1 an.

Toutes les variables utilisées étaient disponibles dans nos données sauf l'ethnicité. Mais une catégorie spécifique lorsque la donnée était indisponible était possible, nous l'avons donc utilisé pour tous nos résidents. Pour la consommation d'alcool,

nous avons juste une variable binaire alors que dans le modèle de QFracture 6 catégories sont disponibles selon la quantité consommée par jour. Nous avons donc appliqué la quantité "1-2 unité par jour" pour les résidents avec une variable alcool à 1. De même pour la variable fumeur, nous avons utilisé la catégorie "fumeur léger, moins de 10 cigarettes par jour" pour les résidents avec une variable fumeur à 1.

QFracture renvoie un pourcentage de risque et n'indique pas d'interprétation avec un seuil précis pour obtenir un risque important, nous avons donc tracé les courbes ROC et précision-rappel pour les femmes (figure 7.10) et pour les hommes (figure 7.11).

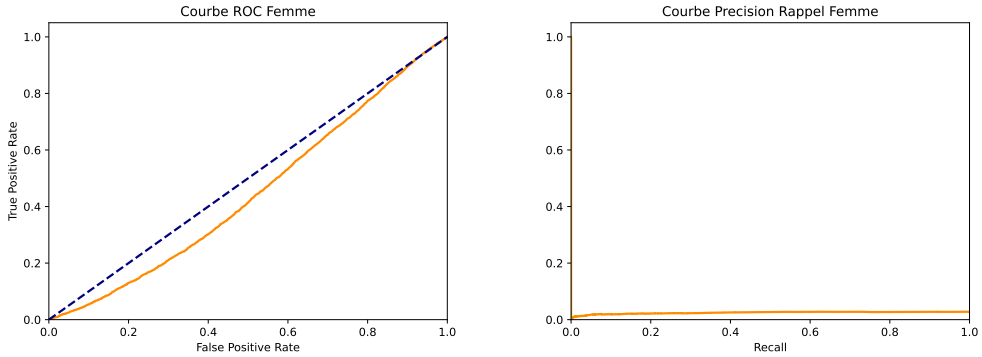


Fig. 7.10. : ROC à gauche et précision-rappel à droite pour les femmes.

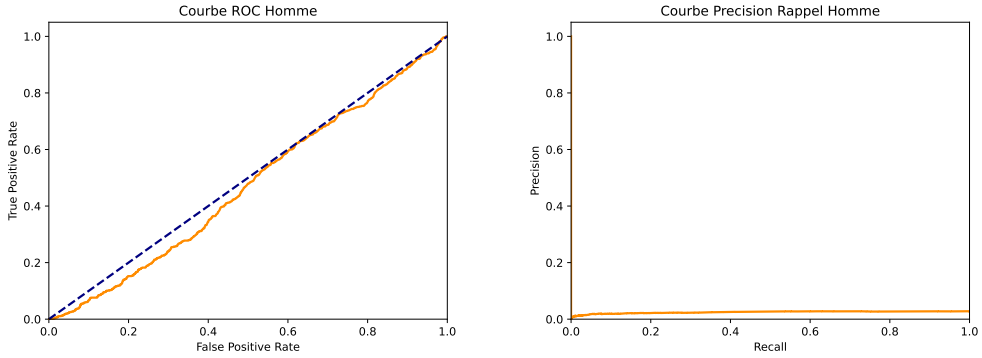


Fig. 7.11. : ROC à gauche et précision-rappel à droite pour les hommes

Il semble ne pas y avoir de différence entre le résultat de la prédiction de QFracture et le hasard, quel que soit le sexe. Nous avons aussi illustré le score obtenu par QFracture en fonction du sexe et du fait d'avoir développé une fracture en figure 7.12. Les médianes sont très faibles pour tous les cas. Et pour les personnes ayant eu une fracture, le pourcentage de risque ne dépasse jamais 40 %.

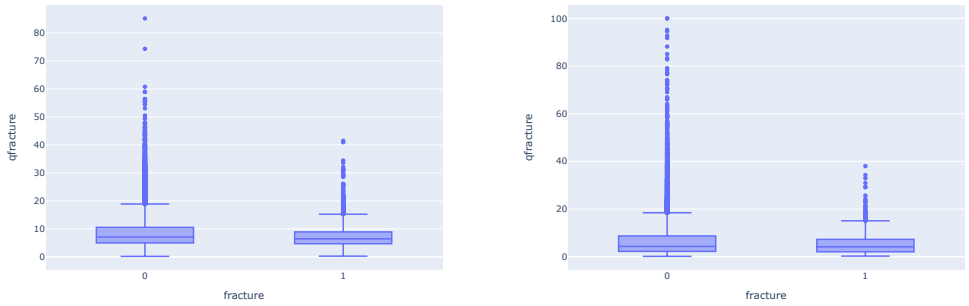


Fig. 7.12. : Boîte à moustaches de la répartition des scores de QFracture en fonction du fait d'avoir réellement eu une fracture pour les femmes à gauche et les hommes à droite.

Les pourcentages de risque prédits étaient en moyenne encore plus faibles avec une prédiction plus tardive. Bien que nous ayons dû faire des choix méthodologiques pour implémenter QFracture avec nos données, l'outil ne semble pas du tout adapté à la prédiction de risque de fracture en EHPAD. Nous n'avons pas pu implémenter FRAX et comparer, car aucune version n'est disponible en *open source*. Il faut aussi noter que FRAX ne permet de prédire le risque de fracture que sur 10 ans.

Pour conclure, nous obtenons un modèle de prédiction de risque de fracture qui pourrait certainement être amélioré en disposant d'une base de données avec plus de cas positifs, mais qui repère déjà des facteurs de risque intéressants et avec des meilleurs résultats que les autres méthodes de classification. L'outil de prédiction QFracture déjà existant ne semble pas du tout adapté à la population de personnes âgées et fragiles présentes en EHPAD. Une publication sur ce sujet est en cours de rédaction.

Références

BELMIN, Joël, Philippe CHASSAGNE et Patrick FRIOCOURT (2023). *Gériatrie : Pour Le Praticien*. Paris : Elsevier Health Sciences. 981 p. (cf. p. 154).

BELMIN, Joël, Philippe CHASSAGNE, Patrick FRIOCOURT et al. (2016). *Gériatrie : pour le Praticien*. Paris : Elsevier Health Sciences. 1071 p. (cf. p. 154).

COX, David R. (1972). "Regression Models and Life-Tables". In : *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2. Publisher : [Royal Statistical Society, Wiley], p. 187-220 (cf. p. 156).

ESHETIE, Tesfahun C., Gillian E. CAUGHEY, Craig WHITEHEAD et al. (2024). “The risk of fractures after entering long-term care facilities”. In : *Bone* 180, p. 116995 (cf. p. 161).

HIPPISLEY-COX, Julia et Carol COUPLAND (2012). “Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom : prospective open cohort study”. In : *BMJ* 344 (may22 1), e3427-e3427 (cf. p. 155).

KANIS, John A., Nicholas C. HARVEY, Helena JOHANSSON et al. (2017). “FRAX Update”. In : *Journal of Clinical Densitometry : The Official Journal of the International Society for Clinical Densitometry* 20.3, p. 360-367 (cf. p. 157).

KANIS, John A., Olof JOHNELL, Anders ODEN et al. (2001). “Ten Year Probabilities of Osteoporotic Fractures According to BMD and Diagnostic Thresholds”. In : *Osteoporosis International* 12.12, p. 989-995 (cf. p. 156).

MARQUES, Andréa, Ricardo J. O. FERREIRA, Eduardo SANTOS et al. (2015). “The accuracy of osteoporotic fracture risk prediction tools : a systematic review and meta-analysis”. In : *Annals of the Rheumatic Diseases* 74.11. Publisher : BMJ Publishing Group Ltd Section : Clinical and epidemiological research, p. 1958-1967 (cf. p. 157).

SARWARI, Zuhreh, Gitte Schultz KRISTENSEN, Sofie Ronja PETERSEN et Christian Backer MOGENSEN (2024). “Analysis of traumatic event emergency department visits among care home residents aged 65+ years in Southern Jutland, Denmark : implications for comprehensive care and subsequent hospital admissions - a register-based cohort study”. In : *BMC geriatrics* 24.1, p. 465 (cf. p. 159).

THÉLOT, Bertrand, Linda LASBEUR et Gaëlle PÉDRONO (2017). “La surveillance épidémiologique des chutes chez les personnes âgées”. In : *Bull Epidémiol Hebd*, p. 328-35 (cf. p. 159).

Conclusion & Perspectives

Tout au long de cette thèse, nous nous sommes intéressés à l'utilisation des réseaux bayésiens comme modèle de classification dans un contexte médical.

Pour ce faire, nous avons commencé par présenter le modèle des réseaux bayésiens, son interprétation et son apprentissage. Nous avons ensuite défini la tâche de classification en s'intéressant particulièrement à la classification probabiliste, dont font partie les réseaux bayésiens lorsqu'ils sont utilisés en tant que classifieurs. L'objectif de cette présentation était de montrer comment les réseaux bayésiens offrent un bon compromis entre la qualité de la prédiction proposée et la qualité de l'explication de cette prédiction. Nous avons ensuite introduit d'autres modèles de *machine learning* car il était important de pouvoir se comparer. De plus, nous avons abordé les problèmes de gestion des données manquantes, de discrétisation des données et de scores de validation.

Enfin, l'état des lieux des utilisations du *machine learning* dans le domaine de la santé a permis de nous guider durant tous les développements de cette thèse. En particulier, nous avons pris en considération les "fuites" schématisées dans le "pipeline" en figure 3.4 ; par exemple, l'objectif d'utilisation en routine clinique, l'intégration dans le logiciel pour une utilisation pratique et la validation externe conduite avec l'investigation clinique. Nous espérons ainsi avoir contribué à leurs adoptions par le personnel soignant.

Par ailleurs, l'utilisation des réseaux bayésiens dans le milieu médical nous paraissait subir trois problèmes principaux que nous exposions dans le chapitre 3 :

- les réseaux bayésiens sont fréquemment créés à partir des connaissances médicales et non pas automatiquement à partir des données, ce qui peut s'expliquer par un manque de bases de données pertinentes et de bonne qualité ;
- dans les cas (rares) d'apprentissage automatique, les méthodes d'apprentissage sont souvent des structures simples, comme les *naive Bayes* et TAN, probablement faute d'implémentation facilement accessible et interprétable pour la communauté médicale ;
- il y a peu de validation en conditions réelles avec intégration dans la routine des soignants, peut-être à cause d'un manque de dispositifs compatibles.

Dans ce travail, nous avons pu présenter des solutions à ces trois problèmes en développant des classifieurs de réseaux bayésiens avec pyAgrum, outil d'utilisation assez simple, mais qui peut apprendre des modèles bien plus complexes que les modèles TAN ou *naives Bayes*. Nous avons proposé des outils d'interprétation et de validation des réseaux obtenus avec les couvertures de Markov généralisées, les

valeurs de Shapley et les courbes ROC et précision-rappel. De plus, nous avions accès à une base de données de santé assez exceptionnelle (en taille, comme en qualité de suivi) qui nous a permis de développer une méthodologie d'apprentissage automatique de réseaux bayésiens où la connaissance experte sert uniquement de support à la construction et à la validation du modèle. En effet, après une mise en conformité longue et complexe, ainsi que des prétraitements assez sophistiqués comprenant entre autres la transformation d'une base événementielle en une base tabulaire et la création de variables médicalement pertinentes, nous avons pu créer un jeu de données homogène par rapport à la population en EHPAD française. Dans le cadre de la validation des modèles, il a été très intéressant d'observer que nous pouvions retrouver dans nos apprentissages automatiques des facteurs de risques connus des experts. Enfin, il est important de remarquer que nos classifieurs ne demandent pas de saisie supplémentaire de la part des soignants, et sont facilement intégrables dans un logiciel déjà déployé et utilisé par la plupart des EHPAD en France.

L'utilité d'un outil de prédiction des risques dépend de sa capacité à identifier des risques qui peuvent bénéficier d'interventions préventives spécifiques, mais aussi, de sa capacité à mettre en œuvre cette prévention à bon escient, c'est-à-dire, éviter les faux négatifs. Nous nous sommes donc concentrés sur des événements de santé défavorables, mais évitables : la survenue d'escarre, l'hospitalisation d'urgence et la fracture, et avons opté pour des métriques privilégiant le rappel (ou la sensibilité) tout en prenant en compte la précision : le F2-score.

L'application sur la première survenue d'escarre en EHPAD a été la première cible à laquelle nous nous sommes intéressées et la plus aboutie. En choisissant des temporalités de prédiction d'un, deux ou trois mois, une prise en charge préventive peut être effectuée et efficace. Nous avons pu remarquer que plus la temporalité était proche, plus juste était la prédiction et que BNClassifier obtenait les meilleurs résultats numériques en comparaison avec les autres méthodes de classification testées. Par ailleurs, en nous concentrant sur la sous-population de notre base ayant été testée avec l'outil principalement utilisé actuellement pour prédire le risque d'escarre en EHPAD : l'échelle de Braden, nous avons pu montrer d'une part son manque de pertinence et d'autre part, que sur cette sous-base, les résultats de BNClassifier étaient en accord avec l'impression clinique du personnel soignant menant à utiliser cet outil. Cette première étude a abouti à l'implémentation d'un outil de prédiction intégré dans le prototype de NETSmart actuellement en phase d'investigation clinique.

Dans une seconde expérience, nous avons étudié la prédiction de l'hospitalisation

en urgence à partir des informations obtenues à l'admission du résident en EHPAD. De manière similaire, nous avons privilégié un modèle plus sensible que spécifique, mais qui reste pertinent dans un contexte de prévention. Dans ce cadre, nous avons obtenu un meilleur F2-score que les autres méthodes de classification testées. Faute d'un modèle préexistant, nous n'avons pas pu nous comparer à des résultats antérieurs.

En ce qui concerne la prédiction de fractures à partir des premiers mois dans l'établissement, nous avons obtenu ici les résultats les moins probants numériquement parmi toutes les applications, ce qui pourrait certainement être amélioré par une base de données de plus grande taille : en effet, il s'agirait d'améliorer la variabilité des cas positifs malgré leur faible prévalence. Notre modèle a néanmoins permis d'identifier des facteurs de risque intéressants et a obtenu de meilleurs résultats que les autres méthodes de classification testées. Nous avons aussi constaté que l'outil de prédiction QFracture existant n'est pas du tout adapté à la population de personnes âgées et fragiles présentes en EHPAD. Il serait de même intéressant de se comparer à d'autres outils de prédiction du risque fracturaire comme FRAX, ce qui n'a pas pu être réalisé faute d'une implémentation *open source* de cet outil.

Dans toutes ces expériences, les évaluations PATHOS et GIR ont été des très bonnes alliées pour la prédiction, ce sont des évaluations obligatoires et importantes pour obtenir des financements pour les EHPAD, elles sont donc faites sérieusement et contiennent des informations précieuses. Il y a certainement d'autres variables pertinentes auxquelles nous n'avons pas pu avoir accès, mais l'objectif était de pouvoir proposer des classifieurs ne nécessitant pas de saisie supplémentaire et donc de composer avec ce manque d'information.

Dans le cadre de cette thèse, notre objectif était principalement de proposer une méthodologie permettant l'évaluation de classifieurs basés sur des réseaux bayésiens et sur la base de données NETSoins. Étant donné la base de données obtenue et le choix de la métrique F2-score, la comparaison entre les différents types de classifieurs s'est fait sans parti pris particulier. Nous nous attendions donc à des résultats numériques meilleurs pour des méthodes moins explicables, comme le présuppose le compromis classique entre explicabilité et qualité de la prédiction. Les résultats des trois expériences montrent que ce compromis ne s'applique pas systématiquement.

Dans la conception de ces outils de prédiction de risque et dans l'objectif de les intégrer chez notre partenaire industriel, nous avons remarqué que l'usage en gestion du risque est de favoriser des outils qui fournissent des niveaux de risque : faible, modéré, fort, par exemple. Cela a donné lieu à de nombreuses discussions avec

les médecins et les équipes de Teranga. La première idée serait bien évidemment d'apprendre des classifieurs non binaires. Toutefois, dans un cadre supervisé, cela nécessite l'estimation de ces différentes classes pour chaque patient, ce qui n'est évidemment pas disponible. Une seconde proposition serait de définir plusieurs seuils dans l'estimation de la probabilité de la classe cible. Néanmoins, il est difficile d'accepter que cette probabilité soit une estimation correcte du risque. En effet, nous ne pouvons pas affirmer que parmi les résidents ayant fait des escarres, celui dont la probabilité la plus élevée est celui avec le plus de risque ou encore celui avec l'escarre la plus grave. Cette différence s'amoindrirait dans un cadre parfait d'un modèle purement causal et exhaustif. Cette discussion intéressante a abouti dans le logiciel à la mise en place d'une notification binaire moins ergonomique en termes d'expérience utilisateur, mais plus scientifiquement fiable.

Pour faire progresser nos classifieurs, les perspectives semblent plus pertinentes dans l'amélioration du prétraitement des données que dans l'amélioration technique du modèle. Un grand nombre de variables ont déjà été synthétisées à partir des données brutes contenues dans les dossiers médicaux électroniques provenant d'EHPAD. Il est certain que des analyses ultérieures pourraient en dégager d'autres. Par exemple, avec des méthodes d'anonymisation plus poussées, il pourrait être possible d'analyser des données moins structurées comme les comptes-rendus d'hospitalisation ou des transmissions narratives. Par ailleurs, l'exploitation des données temporelles irrégulières est un problème très complexe qui est un sujet de recherche à lui seul.

La perspective d'utilisation des algorithmes dans NETSmart et l'investigation clinique qui permet une validation externe de l'algorithme sur les escarres sont des chances exceptionnelles. Il reste encore des questions réglementaires et pratiques sur comment et à quelles fréquences les algorithmes de prédiction pourront être mis à jour. Ces problématiques sont encore émergentes et leurs résolutions nécessitent naturellement beaucoup de temps. Actuellement, il est prévu que l'investigation clinique se termine fin 2024. Il est certain que les retours de cette investigation mettront en avant d'autres pistes d'améliorations.

Bibliographie

ABDELLATIF, Abir (2021). “Amélioration des pratiques des soignants en EHPAD : développement et évaluation d’un assistant numérique intelligent”. Thèse de doct. Sorbonne Université.

ACUÑA, Edgar et Caroline RODRIGUEZ (2004). “The Treatment of Missing Values and its Effect on Classifier Accuracy”. In : *Classification, Clustering, and Data Mining Applications*. Sous la dir. de David BANKS, Frederick R. MCMORRIS, Phipps ARABIE et Wolfgang GAUL. Studies in Classification, Data Analysis, and Knowledge Organisation. Berlin, Heidelberg : Springer, p. 639-647.

ADAMS, James G. (2013). “Emergency Department Overuse : Perceptions and Solutions”. In : *JAMA* 309.11, p. 1173-1174.

ADLER-MILSTEIN, Julia, A Jay HOLMGREN, Peter KRALOVEC et al. (2017). “Electronic health record adoption in US hospitals : the emergence of a digital “advanced use” divide”. In : *Journal of the American Medical Informatics Association* 24.6, p. 1142-1148.

AFFELDT, Séverine et Hervé ISAMBERT (2015). “Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information”. In : *Proceedings of the UAI 2015 Conference on Advances in Causal Inference - Volume 1504*. ACI’15. Aachen, DEU : CEUR-WS.org, p. 1-29.

AFFELDT, Séverine, Louis VERNY et Hervé ISAMBERT (2016). “3off2 : A network reconstruction algorithm based on 2-point and 3-point information statistics”. In : *BMC Bioinformatics* 17.2, S12.

AHN, Hyochol, Linda COWAN, Cynthia GARVAN, Debra LYON et Joyce STECHMILLER (2016). “Risk Factors for Pressure Ulcers Including Suspected Deep Tissue Injury in Nursing Home Facility Residents : Analysis of National Minimum Data Set 3.0”. In : *Advances in Skin & Wound Care* 29.4, p. 178.

ALDERDEN, Jenny, Ginette Alyce PEPPER, Andrew WILSON et al. (2018). “Predicting Pressure Injury in Critical Care Patients : A Machine-Learning Model”. In : *American Journal of Critical Care* 27.6, p. 461-468.

ANDREASSEN, Steen, Christian RIEKEHR, Brian KRISTENSEN, Henrik C. SCHØNHEYDER et Leonard LEIBOVICI (1999). “Using probabilistic and decision–theoretic methods in treatment and prognosis modeling”. In : *Artificial Intelligence in Medicine*. Prognostic Models in Medicine 15.2, p. 121-134.

ANGWIN, Julia, Jeff LARSON, Surya MATTU et Lauren KIRCHNER (2016). *Machine Bias*. ProPublica. URL : `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

BALAVOINE, Angélique (2022). *Des résidents de plus en plus âgés et dépendants dans les établissements d'hébergement pour personnes âgées | Direction de la recherche, des études, de l'évaluation et des statistiques*. ÉTUDES ET RÉSULTATS N°1237.

BARROIS, Brigitte, Denis COLIN et François-André ALLAERT (2018). "Prevalence, characteristics and risk factors of pressure ulcers in public and private hospitals care units and nursing homes in France". In : *Hospital Practice (1995)* 46.1, p. 30-36.

BATISTA, Gustavo E. A. P. A. et Maria Carolina MONARD (2003). "An analysis of four missing data treatment methods for supervised learning". In : *Applied Artificial Intelligence* 17.5. Publisher : Taylor & Francis _eprint : https ://doi.org/10.1080/713827181, p. 519-533.

BCB (2019). *Base Claude Bernard - La base de données sur les Médicaments et les produits de santé*.

BEAM, Andrew L. et Isaac S. KOHANE (2018). "Big Data and Machine Learning in Health Care". In : *JAMA* 319.13, p. 1317-1318.

BELMIN, Joël, Philippe CHASSAGNE et Patrick FRIOCOURT (2023). *Gériatrie : Pour Le Praticien*. Paris : Elsevier Health Sciences. 981 p.

BELMIN, Joël, Philippe CHASSAGNE, Patrick FRIOCOURT et al. (2016). *Gériatrie : pour le Praticien*. Paris : Elsevier Health Sciences. 1071 p.

BELMIN, Joël, Patrick VILLANI, Mathias GAY et al. (2022). "Real-world Implementation of an eHealth System Based on Artificial Intelligence Designed to Predict and Reduce Emergency Department Visits by Older Adults : Pragmatic Trial". In : *Journal of Medical Internet Research* 24.9, e40387.

BERCHIALLA, Paola, Ezio Nicola GANGEMI, Francesca FOLTRAN et al. (2014). "Predicting severity of pathological scarring due to burn injuries : a clinical decision making tool using Bayesian networks". In : *International Wound Journal* 11.3, p. 246-252.

BERGSTROM, Nancy, Barbara J. BRADEN, M. CHAMPAGNE, M. KEMP et E. RUBY (1998). "Predicting pressure ulcer risk : a multisite study of the predictive validity of study of the Braden Scale." In : *Nursing Research* 47.(5), p. 261-269.

BERLOWITZ, D. R., G. H. BRANDEIS, J. J. ANDERSON et al. (2001). "Evaluation of a risk-adjustment model for pressure ulcer development using the Minimum Data Set". In : *Journal of the American Geriatrics Society* 49.7, p. 872-876.

BILLINGSLEY, Patrick (1995). *Probability and Measure*. Google-Books-ID : z39jQgAACAAJ. Wiley. 608 p.

BISHOP, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Springer New York, NY.

BRADEN, Barbara J. et Nancy BERGSTROM (1994). "Predictive validity of the Braden Scale for pressure sore risk in a nursing home population". In : *Research in Nursing & Health* 17.6, p. 459-470.

BUCHANAN, Bruce et Edward SHORTLIFFE (1984). *Rule-based Expert System – The MYCIN Experiments of the Stanford Heuristic Programming Project*. Journal Abbreviation : SERBIULA (sistema Librum 2.0) Publication Title : SERBIULA (sistema Librum 2.0).

BUKHANOV, Nikita, Marina BALAKHONTCEVA, Sergey KOVALCHUK, Nadezhda ZVARTAU et Aleksandra KONRADI (2017). “Multiscale modeling of comorbidity relations in hypertensive outpatients”. In : *Procedia Computer Science*. CENTERIS 2017 - International Conference on ENTERprise Information Systems / ProjMAN 2017 - International Conference on Project MANagement / HCist 2017 - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2017 121, p. 446-450.

CAI, Zhi-qiang, Peng GUO, Shu-bin SI et al. (2017). “Analysis of prognostic factors for survival after surgery for gallbladder cancer based on a Bayesian network”. In : *Scientific Reports* 7.1. Number : 1 Publisher : Nature Publishing Group, p. 293.

CAILLET, Pascal, Sarah KLEMM, Michel DUCHER, Alexandre AUSSEM et Anne-Marie SCHOTT (2015). “Hip Fracture in the Elderly : A Re-Analysis of the EPIDOS Study with Causal Bayesian Networks”. In : *PLOS ONE* 10.3. Publisher : Public Library of Science, e0120125.

CALLE, Juan Esteban de la (2023). *How and Why I Switched from the ROC Curve to the Precision-Recall Curve to Analyze My Imbalanced...* Medium. URL : https://juandelacalle.medium.com/how-and-why-i-switched-from-the-roc-curve-to-the-precision-recall-curve-to-analyze-my-imbalanced-6171da91c6b8 (visité le 4 fév. 2024).

CARUANA, Richard (2019). “Friends Don’t Let Friends Deploy Black-Box Models : The Importance of Intelligibility in Machine Learning”. In : *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD ’19. New York, NY, USA : Association for Computing Machinery, p. 3174.

CHARON, Clara, Pierre-Henri WUILLEMIN et Joël BELMIN (2023). “Improving Pressure Ulcers Prediction in Nursing Homes with ML Algorithm”. In : *Studies in Health Technology and Informatics* 302, p. 350-351.

– (2022). “Learning Bayesian Networks for the Prediction of Unfavorable Health Events in Nursing Homes”. In : *Studies in Health Technology and Informatics* 294, p. 147-148.

CHARON, Clara, Pierre-Henri WUILLEMIN, Charlotte HAVRENG-THÉRY et Joël BELMIN (2024). “One Month Prediction of Pressure Ulcers in Nursing Home Residents with Bayesian Networks”. In : *Journal of the American Medical Directors Association*, S1525–8610(24)00070-7.

CHEN, Cheryl Chia-Hui, Charlotte WANG et Guan-Hua HUANG (2008). “Functional trajectory 6 months posthospitalization : a cohort study of older hospitalized patients in Taiwan”. In : *Nursing Research* 57.2, p. 93-100.

CHEN, Hong-Lin, Wang-Qin SHEN et Peng LIU (2016). “A Meta-analysis to Evaluate the Predictive Validity of the Braden Scale for Pressure Ulcer Risk Assessment in Long-term Care”. In : *Ostomy/Wound Management* 62.9, p. 20-28.

CHEN, Tianqi et Carlos GUESTRIN (2016). “XGBoost : A Scalable Tree Boosting System”. In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 785-794. arXiv : 1603.02754[cs].

CHENG, Jie et Russell GREINER (2013). *Comparing Bayesian Network Classifiers*. arXiv : 1301.6684[cs,stat].

CHICKERING, David (2000). “Learning Bayesian Networks is NP-Complete”. In : *Networks* 112.

CHRISTODOULOU, Evangelia, Jie MA, Gary S. COLLINS et al. (2019). “A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models”. In : *Journal of Clinical Epidemiology* 110, p. 12-22.

“Classification and regression trees (CART)” (2012). In : MURPHY, Kevin P. *Machine Learning : A Probabilistic Perspective*. Google-Books-ID : NZP6AQAAQBAJ. MIT Press, p. 544-550.

CNIL (2023). *Tenir compte de la protection des données dans la collecte et la gestion des données*. URL : `https://www.cnil.fr/fr/tenir-compte-de-la-protection-des-donnees-dans-la-collecte-et-la-gestion-des-donnees` (visité le 8 fév. 2024).

CNSA (2022). “Le modèle PATHOS. Guide d utilisation 2022”. In.

COOK, Stephen A (1971). “The Complexity of Theorem-Proving Procedures”. In.

COX, David R. (1972). “Regression Models and Life-Tables”. In : *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2. Publisher : [Royal Statistical Society, Wiley], p. 187-220.

CRANE, Sarah J., Ericka E. TUNG, Gregory J. HANSON et al. (2010). “Use of an electronic administrative database to identify older community dwelling adults at high-risk for hospitalization or emergency department visits : The elders risk assessment index”. In : *BMC Health Services Research* 10.1, p. 338.

DAGUM, Paul et Michael LUBY (1993). “Approximating probabilistic inference in Bayesian belief networks is NP-hard”. In : *Artificial Intelligence* 60.1, p. 141-153.

DAM, Hoa Khanh, Truyen TRAN et Aditya GHOSE (2018). “Explainable Software Analytics”. In.

DAVIS, Jesse et Mark GOADRICH (2006). “The relationship between Precision-Recall and ROC curves”. In : *Proceedings of the 23rd international conference on Machine learning*. ICML ’06. New York, NY, USA : Association for Computing Machinery, p. 233-240.

DEFLOOR, T. (1999). “The risk of pressure sores : a conceptual scheme”. In : *Journal of Clinical Nursing* 8.2, p. 206-216.

DEMPSTER, Arthur P., Nan M. LAIRD et Donald B. RUBIN (1977). “Maximum Likelihood from Incomplete Data Via the EM Algorithm”. In : *Journal of the Royal Statistical Society : Series B (Methodological)* 39.1, p. 1-22.

DIXON, John K. (1979). “Pattern Recognition with Partly Missing Data | IEEE Journals & Magazine | IEEE Xplore”. In : *IEEE Transactions on Systems, Man, and Cybernetics (Volume : 9, Issue : 10*.

DOMINGOS, Pedro et Michael PAZZANI (1997). “On the Optimality of the Simple Bayesian Classifier under Zero-One Loss”. In : *Machine Learning*. Kluwer Academic Publishers 29, p. 103-130.

DOUTRELIGNE, Matthieu, Adeline DEGREMONT, Pierre-Alain JACHIET, Antoine LAMER et Xavier TANNIER (2023). “Good practices for clinical data warehouse implementation : A case study in France”. In : *PLOS Digital Health* 2.7. Publisher : Public Library of Science, e0000298.

DRANCA, Lacramioara, Lopez de ABETXUKO RUIZ DE MENDAROZKETA, Alfredo GOÑI et al. (2018). “Using Kinect to classify Parkinson’s disease stages related to severity of gait impairment”. In : *BMC Bioinformatics* 19.1, p. 471.

DUCAMP, Gaspard, Christophe GONZALES et Pierre-Henri WUILLEMIN (2020). “aGrUM/-pyAgrum : a toolbox to build models and algorithms for Probabilistic Graphical Models in Python”. In : *10th International Conference on Probabilistic Graphical Models*. T. 138. Proceedings of Machine Learning Research. Skørping, Denmark, p. 609-612.

DWEEKAT, Odai Y., Sarah S. LAM et Lindsay MCGRATH (2023). “Machine Learning Techniques, Applications, and Potential Future Opportunities in Pressure Injuries (Bedsores) Management : A Systematic Review”. In : *International Journal of Environmental Research and Public Health* 20.1, p. 796.

DWYER, Rosamond, Belinda GABBE, Johannes U. STOELWINDER et Judy LOWTHIAN (2014). “A systematic review of outcomes following emergency transfer to hospital for residents of aged care facilities”. In : *Age and Ageing* 43.6, p. 759-766.

EDELMAN, Benjamin, Michael LUCA et Dan SVIRSKY (2017). “Racial Discrimination in the Sharing Economy : Evidence from a Field Experiment”. In : *American Economic Journal : Applied Economics* 9.2, p. 1-22.

EGAN, James P. (1975). *Signal detection theory and ROC-analysis*. Academic Press series in cognition and perception. OCLC : 1499787. New York : Academic Press. 277 p.

ESHETIE, Tesfahun C., Gillian E. CAUGHEY, Craig WHITEHEAD et al. (2024). “The risk of fractures after entering long-term care facilities”. In : *Bone* 180, p. 116995.

FAYYAD, Usama M. et Keki B. IRANI (1993). “Multi-interval discretization of continuous-valued attributes for classification learning”. In : *Ijcai*. T. 93. Issue : 2. Citeseer, p. 1022-1029.

FREEDMAN, David (1997). “Some Issues in the Foundation of Statistics”. In : *Topics in the Foundation of Statistics*. Sous la dir. de Bas C. van FRAASSEN. Dordrecht : Springer Netherlands, p. 19-39.

FREUND, Yoav et Robert E SCHAPIRE (1997). “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting”. In : *Journal of Computer and System Sciences* 55.1, p. 119-139.

FRIEDMAN, Jerome H. (2001). “Greedy function approximation : A gradient boosting machine.” In : *The Annals of Statistics* 29.5. Publisher : Institute of Mathematical Statistics, p. 1189-1232.

FRIEDMAN, Nir, Dan GEIGER et Moises GOLDSZMIDT (1997). “Bayesian Network Classifiers”. In : *Machine Learning* 29.2, p. 131-163.

GAAG, L. C. van der, S. RENOOIJ, C. L. M. WITTEMAN, B. M. P. ALEMAN et B. G. TAAL (2002). “Probabilities for a probabilistic network : a case study in oesophageal cancer”. In : *Artificial Intelligence in Medicine* 25.2, p. 123-148.

GARCIA, Salvador, J. LUENGO, José Antonio SÁEZ, Victoria LÓPEZ et F. HERRERA (2013). “A Survey of Discretization Techniques : Taxonomy and Empirical Analysis in Supervised Learning”. In : *IEEE Transactions on Knowledge and Data Engineering* 25.4, p. 734-750.

GASPERINI, Beatrice, Antonio CHERUBINI, Francesca PIERRI et al. (2017). “Potentially preventable visits to the emergency department in older adults : Results from a national survey in Italy”. In : *PLOS ONE* 12.12. Publisher : Public Library of Science, e0189925.

GEFEN, Amit (2008). “How much time does it take to get a pressure ulcer ? Integrated evidence from human, animal, and in vitro studies”. In : *Ostomy/Wound Management* 54.10, p. 26-28, 30-35.

GÉRON, Aurélien (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*. Google-Books-ID : HHetDwAAQBAJ. "O’Reilly Media, Inc." 851 p.

GLOVER, Fred et Manuel LAGUNA (1998). “Tabu Search”. In : *Handbook of Combinatorial Optimization : Volume1–3*. Sous la dir. de Ding-Zhu DU et Panos M. PARDALOS. Boston, MA : Springer US, p. 2093-2229.

GOLDBERG, David Edward (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Google-Books-ID : 2IIJAAAACAAJ. Addison-Wesley. 436 p.

GONZALES, Christophe (2018). “Les réseaux bayésiens”. In : *Interstices*. Publisher : INRIA.

HADDAD, Tarek, Adam HIMES et Michael CAMPBELL (2014). “Fracture prediction of cardiac lead medical devices using Bayesian networks”. In : *Reliability Engineering & System Safety* 123, p. 145-157.

HADJ ALI, Mahdi, Yann Le BIANNIC et Pierre-Henri WUILLEMIN (2023). “Interpreting Predictive Models through Causality : A Query-Driven Methodology”. In : *The International FLAIRS Conference Proceedings* 36.

HAND, David J. et Keming YU (2001). “Idiot’s Bayes : Not So Stupid after All ?” In : *International Statistical Review / Revue Internationale de Statistique* 69.3. Publisher : [Wiley, International Statistical Institute (ISI)], p. 385-398.

HARISH, Vinyas, Felipe MORGADO, Ariel D. STERN et Sunit DAS (2021). “Artificial Intelligence and Clinical Decision Making : The New Nature of Medical Uncertainty”. In : *Academic Medicine : Journal of the Association of American Medical Colleges* 96.1, p. 31-36.

HASTIE, Trevor, Robert TIBSHIRANI et Jerome FRIEDMAN (2009a). “Ensemble Learning”. In : *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Sous la dir. de Trevor HASTIE, Robert TIBSHIRANI et Jerome FRIEDMAN. Springer Series in Statistics. New York, NY : Springer, p. 605-624.

– (2009b). “Neural Networks”. In : *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Sous la dir. de Trevor HASTIE, Robert TIBSHIRANI et Jerome FRIEDMAN. Springer Series in Statistics. New York, NY : Springer, p. 389-416.

– (2009c). “Random Forests”. In : *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Sous la dir. de Trevor HASTIE, Robert TIBSHIRANI et Jerome FRIEDMAN. Springer Series in Statistics. New York, NY : Springer, p. 587-604.

– (2009d). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY : Springer.

HAUTE AUTORITÉ DE SANTÉ, HAS (2001). *Prévention et traitement des escarres de l'adulte et du sujet âgé*. URL : https://www.has-sante.fr/jcms/c_271996/fr/prevention-et-traitement-des-escarres-de-l-adulte-et-du-sujet-age.

HECKERMAN, David, Dan GEIGER et David M. CHICKERING (1995). “Learning Bayesian Networks : The Combination of Knowledge and Statistical Data”. In : *Machine Learning* 20.3, p. 197-243.

HECKERMAN, David E., Eric J. HORVITZ et Bharat N. NATHWANI (1992). “Toward normative expert systems : Part I. The Pathfinder project”. In : *Methods of Information in Medicine* 31.2, p. 90-105.

HECKERMAN, David E. et Edward H. SHORTLIFFE (1992). “From certainty factors to belief networks”. In : *Artificial Intelligence in Medicine* 4.1, p. 35-52.

HERNÁN, Miguel A., John HSU et Brian HEALY (2019). “A Second Chance to Get Causal Inference Right : A Classification of Data Science Tasks”. In : *CHANCE* 32.1, p. 42-49.

HIPPISLEY-COX, Julia et Carol COUPLAND (2012). “Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom : prospective open cohort study”. In : *BMJ* 344 (may22 1), e3427-e3427.

HOLZINGER, Andreas (2021). “The Next Frontier : AI We Can Really Trust”. In : *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Sous la dir. de Michael KAMP, Irena KOPRINSKA, Adrien BIBAL et al. Communications in Computer and Information Science. Cham : Springer International Publishing, p. 427-440.

HOOS, Holger H. et Thomas STÜTZLE (2005). “3 - GENERALISED LOCAL SEARCH MACHINES”. In : *Stochastic Local Search*. Sous la dir. d'Holger H. HOOS et Thomas STÜTZLE. The Morgan Kaufmann Series in Artificial Intelligence. San Francisco : Morgan Kaufmann, p. 113-147.

HU, Ya-Han, Yi-Lien LEE, Ming-Feng KANG et Pei-Ju LEE (2020). “Constructing Inpatient Pressure Injury Prediction Models Using Machine Learning Techniques”. In : *CIN : Computers, Informatics, Nursing* 38.8, p. 415.

HWANG, Ula et R. Sean MORRISON (2007). “The geriatric emergency department”. In : *Journal of the American Geriatrics Society* 55.11, p. 1873-1876.

IA, HUB France (2023). *L'IA éthique en pratique*.

IWASHYNA, Theodore J, E Wesley ELY, Dylan M SMITH et Kenneth M LANGA (2010). “Long-term cognitive impairment and functional disability among survivors of severe sepsis”. In : *JAMA* 304.16, p. 1787-1794.

JIANG, Fei, Yong JIANG, Hui ZHI et al. (2017). “Artificial intelligence in healthcare : past, present and future”. In : *Stroke and Vascular Neurology* 2.4, p. 230-243.

JIANG, Xia, Diyang XUE, Adam BRUFSKY, Seema KHAN et Richard NEAPOLITAN (2014). "A New Method for Predicting Patient Survivorship Using Efficient Bayesian Network Learning". In : *Cancer Informatics* 13. Publisher : SAGE Publications Ltd STM, CIN.S13053.

JOCHEMS, Arthur, Timo M. DEIST, Johan van SOEST et al. (2016). "Distributed learning : Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept". In : *Radiotherapy and Oncology* 121.3. Publisher : Elsevier, p. 459-467.

JONES, Linda A., Jenny R. NELDER, Joseph M. FRYER et al. (2022). "Public opinion on sharing data from health services for clinical and research purposes without explicit consent : an anonymous online survey in the UK". In : *BMJ Open* 12.4. Publisher : British Medical Journal Publishing Group Section : Health policy, e057579.

JORDAN, Michael I. et Tom M. MITCHELL (2015). "Machine learning : Trends, perspectives, and prospects". In : *Science* 349.6245. Publisher : American Association for the Advancement of Science, p. 255-260.

KAEWPRAG, Pacharmon, Cheryl NEWTON, Brenda VERMILLION et al. (2015). "Predictive Modeling for Pressure Ulcers from Intensive Care Unit Electronic Health Records". In : *AMIA Summits on Translational Science Proceedings* 2015, p. 82-86.

KAHN, Joseph H., Brendan G. MAGAURAN, Jonathan S. OLSHAKER et Kalpana N. SHANKAR (2016). "Current Trends in Geriatric Emergency Medicine". In : *Emergency Medicine Clinics of North America*. Geriatric Emergencies 34.3, p. 435-452.

KANIS, John A., Nicholas C. HARVEY, Helena JOHANSSON et al. (2017). "FRAX Update". In : *Journal of Clinical Densitometry : The Official Journal of the International Society for Clinical Densitometry* 20.3, p. 360-367.

KANIS, John A., Olof JOHNELL, Anders ODEN et al. (2001). "Ten Year Probabilities of Osteoporotic Fractures According to BMD and Diagnostic Thresholds". In : *Osteoporosis International* 12.12, p. 989-995.

KHAN, Muhammad Yaseen, Abdul QAYOOM, Muhammad NIZAMI et al. (2021). "Automated Prediction of Good Dictionary EXamples (GDEX) : A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques". In : *Complexity*.

KHINCHIN, Alexandre (1957). *Mathematical Foundations Of Information Theory*.

KIRKPATRICK, Scott, Daniel GELATT et Manuela P. VECCHI (1983). "Optimization by Simulated Annealing". In : *Science* 220.4598. Publisher : American Association for the Advancement of Science, p. 671-680.

KOLLER, Daphne et Nir FRIEDMAN (2009). *Probabilistic graphical models : principles and techniques*. Adaptive computation and machine learning. Cambridge, MA : MIT Press. 1231 p.

KRUSKAL, Joseph B (1956). "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem". In : *Proceedings of the American Mathematical Society* 7, p. 48-50.

KYRIMI, Evangelia, Kudakwashe DUBE, Norman FENTON et al. (2021). “Bayesian networks in healthcare : What is preventing their adoption ?” In : *Artificial Intelligence in Medicine* 116, p. 102079.

KYRIMI, Evangelia, Scott MCLACHLAN, Kudakwashe DUBE et Norman FENTON (2020). *Bayesian Networks in Healthcare : the chasm between research enthusiasm and clinical adoption*. Pages : 2020.06.04.20122911.

KYRIMI, Evangelia, Scott MCLACHLAN, Kudakwashe DUBE, Mariana R. NEVES et al. (2021). “A comprehensive scoping review of Bayesian networks in healthcare : Past, present and future”. In : *Artificial Intelligence in Medicine* 117, p. 102108.

LADIOS-MARTIN, Mireia, José FERNÁNDEZ-DE-MAYA, Francisco-Javier BALLESTA-LÓPEZ et al. (2020). “Predictive Modeling of Pressure Injury Risk in Patients Admitted to an Intensive Care Unit”. In : *American Journal of Critical Care* 29.4, e70-e80.

LAMBRECHT, Anja et Catherine E. TUCKER (2018). *Algorithmic Bias ? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads*. Rochester, NY.

LEE, Soo-Kyoung, Juh Hyun SHIN, Jinhyun AHN, Ji Yeon LEE et Dong Eun JANG (2021). “Identifying the Risk Factors Associated with Nursing Home Residents’ Pressure Ulcers Using Machine Learning Methods”. In : *International Journal of Environmental Research and Public Health* 18.6. Number : 6 Publisher : Multidisciplinary Digital Publishing Institute, p. 2954.

LÉGIFRANCE (2017). “Article Annexe 2-1”. In : *Code de l’action sociale et des familles*.

LIN, Lijing, Matthew SPERRIN, David A. JENKINS, Glen P. MARTIN et Niels PEEK (2021). “A scoping review of causal methods enabling predictions under hypothetical interventions”. In : *Diagnostic and Prognostic Research* 5.1, p. 3.

LOGHMANPOUR, Natasha A., Marek J. DRUZDZEL et James F. ANTAKI (2014). “Cardiac Health Risk Stratification System (CHRiSS) : A Bayesian-Based Decision Support System for Left Ventricular Assist Device (LVAD) Therapy”. In : *PLOS ONE* 9.11. Publisher : Public Library of Science, e111264.

“Logistic regression” (2012). In : MURPHY, Kevin P. *Machine Learning : A Probabilistic Perspective*. Google-Books-ID : NZP6AQAAQBAJ. MIT Press, p. 245-279.

LUCAS, Peter J., Nicolette C. de BRUIJN, Karin SCHURINK et Andy HOEPELMAN (2000). “A probabilistic and decision-theoretic approach to the management of infectious disease at the ICU”. In : *Artificial Intelligence in Medicine* 19.3, p. 251-279.

LUCAS, Peter J. F., Linda C. VAN DER GAAG et Ameen ABU-HANNA (2004). “Bayesian networks in biomedicine and health-care”. In : *Artificial Intelligence in Medicine*. Bayesian Networks in Biomedicine and Health-Care 30.3, p. 201-214.

LUO, Yi, Issam El NAQA, Daniel L. MCSHAN et al. (2017). “Unraveling biophysical interactions of radiation pneumonitis in non-small-cell lung cancer via Bayesian network analysis”. In : *Radiotherapy and Oncology* 123.1. Publisher : Elsevier, p. 85-92.

LYDER, Courtney H. (2003). “Pressure ulcer prevention and management”. In : *JAMA* 289.2, p. 223-226.

MACK, Christina, Zhaohui SU et Daniel WESTREICH (2018). “Types of Missing Data”. In : *Managing Missing Data in Patient Registries : Addendum to Registries for Evaluating Patient Outcomes : A User’s Guide, Third Edition [Internet]*. Agency for Healthcare Research et Quality (US).

MACQUEEN, J. (1967). “Some methods for classification and analysis of multivariate observations”. In : *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Statistics*. T. 5.1. University of California Press, p. 281-298.

MARQUES, Andréa, Ricardo J. O. FERREIRA, Eduardo SANTOS et al. (2015). “The accuracy of osteoporotic fracture risk prediction tools : a systematic review and meta-analysis”. In : *Annals of the Rheumatic Diseases* 74.11. Publisher : BMJ Publishing Group Ltd Section : Clinical and epidemiological research, p. 1958-1967.

MARTIN, Glen P., Mamas A. MAMAS, Niels PEEK, Iain BUCHAN et Matthew SPERRIN (2018). “A multiple-model generalisation of updating clinical prediction models”. In : *Statistics in Medicine* 37.8, p. 1343-1358.

MARTIN, Glen P., Matthew SPERRIN, Kym I. E. SNELL, Iain BUCHAN et Richard D. RILEY (2021). “Clinical prediction models to predict the risk of multiple binary outcomes : a comparison of approaches”. In : *Statistics in Medicine* 40.2, p. 498-517.

MERLI, Mauro, Marco MOSCATELLI, Giorgia MARIOTTI et al. (2016). “A minimally invasive technique for lateral maxillary sinus floor elevation : a Bayesian network study”. In : *Clinical Oral Implants Research* 27.3, p. 273-281.

MOONS, Karel G. M., Andre Pascal KENGNE, Diederick E. GROBBEE et al. (2012). “Risk prediction models : II. External validation, model updating, and impact assessment”. In : *Heart* 98.9. Publisher : BMJ Publishing Group Ltd and British Cardiovascular Society Section : Review, p. 691-698.

NORTON, Doreen, Rhoda MCLAREN et A.N. EXTON-SMITH (1962). *An investigation of geriatric nursing problems in hospital*. National Corporation for the Care of Old People. London.

OJEME, Blessing et Audrey MBOGHO (2016). “Selecting Learning Algorithms for Simultaneous Identification of Depression and Comorbid Disorders”. In : *Procedia Computer Science*. Knowledge-Based and Intelligent Information & Engineering Systems : Proceedings of the 20th International Conference KES-2016 96, p. 1294-1303.

PARK, Eunjeong, Hyuk-jae CHANG et Hyo Suk NAM (2018). “A Bayesian Network Model for Predicting Post-stroke Outcomes With Available Risk Factors”. In : *Frontiers in Neurology* 9.

PARK, Seong Ho et Kyunghwa HAN (2018). “Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction”. In : *Radiology* 286.3. Publisher : Radiological Society of North America, p. 800-809.

PAUL Y. TAKAHASHI, M. D., M. S. HERBERT C. HEIEN, M. P. H. LINDSEY R. SANGARALINGHAM, PhD NILAY D. SHAH et ScD JAMES M. NAESSENS (2016). “Enhanced Risk Prediction Model for Emergency Department Use and Hospitalizations in Patients in a Primary Care Medical Home”. In : July 2016 22. Publisher : MJH Life Sciences.

PEARL, Judea (1988). *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc. 552 p.

PEARSON, Karl (1892). *The Grammar of Science*. Cambridge Library Collection - Physical Sciences. Cambridge : Cambridge University Press.

PEDREGOSA, Fabian, Gaël VAROQUAUX, Alexandre GRAMFORT et al. (2011). "Scikit-learn : Machine Learning in Python". In : *Journal of Machine Learning Research* 12.85, p. 2825-2830.

RASHEED, Khansa, Adnan QAYYUM, Mohammed GHALY et al. (2022). "Explainable, trustworthy, and ethical machine learning for healthcare : A survey". In : *Computers in Biology and Medicine* 149, p. 106043.

REIMEIR, Benjamin, Steven van ANDEL et Peter FEDEROLF (2021). *How does the postural control system determine whether or not it is "in trouble" ? The role of Distance-to-Boundary and Time-to-Boundary in detecting postural instability*.

ROBINSON, Randall W. (1977). "Counting unlabeled acyclic digraphs". In : *Combinatorial Mathematics V*. Sous la dir. de Charles H. C. LITTLE. Lecture Notes in Mathematics. Berlin, Heidelberg : Springer, p. 28-43.

ROYEN, Florien S. van, Karel G. M. MOONS, Geert-Jan GEERSING et Maarten van SMEDEN (2022). "Developing, validating, updating and judging the impact of prognostic models for respiratory diseases". In : *European Respiratory Journal* 60.3. Publisher : European Respiratory Society Section : ERJ Methods.

SANCHEZ-MARTINEZ, Sergio, Oscar CAMARA, Gemma PIELLA et al. (2022). "Machine Learning for Clinical Decision-Making : Challenges and Opportunities in Cardiovascular Imaging". In : *Frontiers in Cardiovascular Medicine* 8. Publisher : Frontiers.

SANDERS, Samantha F., Mats TERWIESCH, William J. GORDON et Ariel Dora STERN (2019). "How Artificial Intelligence Is Changing Health Care Delivery". In : *NEJM Catalyst*.

SAPORTA, Gilbert (2006). *Probabilités, analyse des données et statistique*. Google-Books-ID : rprNjztQYPAC. Editions TECHNIP. 664 p.

SARWARI, Zuhreh, Gitte Schultz KRISTENSEN, Sofie Ronja PETERSEN et Christian Backer MOGENSEN (2024). "Analysis of traumatic event emergency department visits among care home residents aged 65+ years in Southern Jutland, Denmark : implications for comprehensive care and subsequent hospital admissions - a register-based cohort study". In : *BMC geriatrics* 24.1, p. 465.

SCHWARZ, Gideon (1978). "Estimating the Dimension of a Model". In : *The Annals of Statistics* 6.2. Publisher : Institute of Mathematical Statistics, p. 461-464.

SEIBERT, Kathrin, Dominik DOMHOFF, Dominik BRUCH et al. (2021). "Application Scenarios for Artificial Intelligence in Nursing Care : Rapid Review". In : *Journal of Medical Internet Research* 23.11, e26522.

SESEN, M. Berkan, Michael D. PEAKE, Rene BANARES-ALCANTARA et al. (2014). "Lung Cancer Assistant : a hybrid clinical decision support application for lung cancer care". In : *Journal of The Royal Society Interface* 11.98. Publisher : Royal Society, p. 20140534.

SHAPLEY, Lloyd S. (1953). “17. A Value for n-Person Games”. In : *17. A Value for n-Person Games*. Princeton University Press, p. 307-318.

SHELTON, Paul, Mark A SAGER et Cheryl SCHRAEDER (2000). “The Community Assessment Risk Screen (CARS) : Identifying Elderly Persons at Risk for Hospitalization or Emergency Department Visit”. In : *THE AMERICAN JOURNAL OF MANAGED CARE* 6.8.

SOLOMONIDES, Anthony E, Eileen KOSKI, Shireen M ATABAKI et al. (2022). “Defining AMIA’s artificial intelligence principles”. In : *Journal of the American Medical Informatics Association* 29.4, p. 585-591.

SONG, Jie, Yuan GAO, Pengbin YIN et al. (2021). “The Random Forest Model Has the Best Accuracy Among the Four Pressure Ulcer Prediction Models Using Machine Learning Algorithms”. In : *Risk Management and Healthcare Policy* 14. Publisher : Dove Press, p. 1175-1187.

SOUZA, Diba Maria Sebba Tosta de, Vera Lúcia Conceição de Gouveia SANTOS, Helena Keiko IRI et Miriam Yukiko SADASUE OGURI (2010). “Predictive validity of the Braden Scale for Pressure Ulcer Risk in elderly residents of long-term care facilities”. In : *Geriatric Nursing (New York, N.Y.)* 31.2, p. 95-104.

SPERRIN, Matthew, Karla DIAZ-ORDAZ et Romin PAJOUHESHNIA (2021). “Invited Commentary : Treatment Drop-in-Making the Case for Causal Prediction”. In : *American Journal of Epidemiology* 190.10, p. 2015-2018.

SPERRIN, Matthew, David JENKINS, Glen P. MARTIN et Niels PEEK (2019). “Explicit causal reasoning is needed to prevent prognostic models being victims of their own success”. In : *Journal of the American Medical Informatics Association : JAMIA* 26.12, p. 1675-1676.

SPIRTES, Peter et Clark GLYMOUR (1991). “An Algorithm for Fast Recovery of Sparse Causal Graphs”. In : *Social Science Computer Review* 9.1. Publisher : SAGE Publications Inc, p. 62-72.

SULLIVAN, Nancy et Karen M. SCHOELLES (2013). “Preventing In-Facility Pressure Ulcers as a Patient Safety Strategy : A Systematic Review”. In : *Annals of Internal Medicine* 158.5, p. 410.

SYAFIANDINI, Arida Ferti et Ito WASITO (2016). “Metastasis identification based on clinical parameters using Bayesian network”. In : *2016 4th International Conference on Information and Communication Technology (ICoICT)*. 2016 4th International Conference on Information and Communication Technology (ICoICT), p. 1-6.

Teranga Sofware - Prendre soin de ceux qui prennent soin des autres (2023). URL : `https://www.teranga-software.com/`.

THÉLOT, Bertrand, Linda LASBEUR et Gaëlle PÉDRONO (2017). “La surveillance épidémiologique des chutes chez les personnes âgées”. In : *Bull Epidémiol Hebd*, p. 328-35.

TOUMLILT, Ilyas (2021). “Colony : a Hybrid Consistency System for Highly-Available Collaborative Edge Computing”. Thèse de doct. Sorbonne Université.

TROYANSKAYA, Olga, Mike CANTOR, Gavin SHERLOCK et al. (2001). “Missing Value Estimation Methods for DNA Microarrays”. In : *Bioinformatics* 17, p. 520-525.

USCHER-PINES, Lori, Jesse PINES, Arthur KELLERMANN, Emily GILLEN et Ateev MEHROTRA (2013). "Emergency department visits for nonurgent conditions : systematic literature review". In : *The American Journal of Managed Care* 19.1, p. 47-59.

VALDES, Gilmer, José Marcio LUNA, Eric EATON et al. (2016). "MediBoost : a Patient Stratification Tool for Interpretable Decision Making in the Era of Precision Medicine". In : *Scientific Reports* 6.1. Number : 1 Publisher : Nature Publishing Group, p. 37854.

VANDERWEE, Katrien, Michael CLARK, Carol DEALEY, Lena GUNNINGBERG et Tom DEFLOOR (2007). "Pressure ulcer prevalence in Europe : a pilot study". In : *Journal of Evaluation in Clinical Practice* 13.2, p. 227-235.

VEERAPPA, Manjunatha et Salvo RINZIVILLO (2023). *Explainable AI - Introduction to the Special Theme*. URL : `https://ercim-news.ercim.eu/en134/special/explainable-ai-introduction-to-the-special-theme`.

VEMULAPALLI, Vijetha, Jiaqi QU, Jeonifer M. GARREN et al. (2016). "Non-obvious correlations to disease management unraveled by Bayesian artificial intelligence analyses of CMS data". In : *Artificial Intelligence in Medicine* 74, p. 1-8.

VERMA, Thomas et Judea PEARL (1991). *Equivalence and Synthesis of Causal Models*. Google-Books-ID : ikuuHAAACAAJ. UCLA Computer Science Department. 9 p.

VERNY, Louis, Nadir SELLA, Séverine AFFELDT, Param Priya SINGH et Hervé ISAMBERT (2017). "Learning causal networks with latent variables from multivariate information in genomic data". In : *PLOS Computational Biology* 13.10, e1005662.

VEYRON, Jacques-Henri, Patrick FRIOCOURT, Olivier JEANJEAN et al. (2019). "Home care aides' observations and machine learning algorithms for the prediction of visits to emergency departments by older community-dwelling individuals receiving home care assistance : A proof of concept study". In : *PLOS ONE* 14.8. Publisher : Public Library of Science, e0220002.

WANG, Zhao, Michael W. JENKINS, George C. LINDERMAN et al. (2015). "3-D Stent Detection in Intravascular OCT Using a Bayesian Network and Graph Search". In : *IEEE Transactions on Medical Imaging* 34.7. Conference Name : IEEE Transactions on Medical Imaging, p. 1549-1561.

WOLINSKY, Fredric D., Li LIU, Thomas R. MILLER et al. (2008). "Emergency Department Utilization Patterns Among Older Adults". In : *The Journals of Gerontology : Series A* 63.2, p. 204-209.

WOODMAN, Richard J. et Arduino A. MANGONI (2023). "A comprehensive review of machine learning algorithms and their application in geriatric medicine : present and future". In : *Aging Clinical and Experimental Research* 35.11, p. 2363-2397.

YANG, Cynthia, Jan A. KORS, Solomon IOANNOU et al. (2022). "Trends in the conduct and reporting of clinical prediction model development and validation : a systematic review". In : *Journal of the American Medical Informatics Association : JAMIA* 29.5, p. 983-989.

ZHANG, Harry (2004). "The Optimality of Naive Bayes". In : *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*.

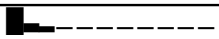
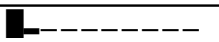
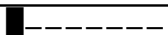
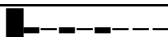
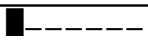
Table des figures

Liste des tableaux

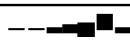
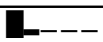
Annexe : Jeu de données

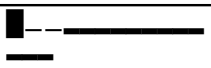
A

A.1 Description des variables du jeu de données pour la prédiction d'escarre 1 mois avant.

Feature name	Distribution	Modalities	Completion (%)
1mo_attendance_leisure_slope		[(0 1], (1, 2], (2, 3], (3, 4], (4, 5], (5, 6], (6, 8], (8, 11], (11, 15], (15, 23], (23, 93]]	100.0
1mo_body_weight_change		[<=-15%-, -15%,-10%, -10%-0, 0-10%, 10%-15%, >=15]	99.1
1mo_body_weight_difference (kilos)		[(-10, -2], (-2, -1], (-1, 0], (0, 1], (1, 6]]	99.1
1mo_mean_systolic_blood_pressure (mmHg)		[<80, 80-100, 100-120, 120-140, >140]	52.1
3mo_attendance_leisure_slope		[(0, 1], (1, 2], (2, 3], (3, 4], (4, 5], (5, 7], (7, 10], (10, 14], (14, 21], (21, 97]]	100.0
3mo_body_weight_change		[<=-15%-, -15%, -10%, -10%-0, 0-10%, 10%-15%, >=15]	99.1
3mo_body_weight_difference (kilos)		[(-10, -5], (-5, -3], (-3, -2], (-2, -1], (-1, 0], (0, 1], (1, 2], (2, 3], (3, 14]]	99.1
3mo_body_weight_slope		[(-33, -1], (-1, 0], (0, 1], (1, 38]]	99.1
3mo_falls_slope		[(0, 1], (1, 2], (2, 31]]	100.0
3mo_hospitalizations_slope		[0, 1, 2, 3, 4, 6, 7, 9]	100.0
3mo_mean_systolic_blood_pressure (mmHg)		[<80, 80-100, 100-120, 120-140, >140]	70.5
6mo_attendance_leisure_slope		[(0, 1], (1, 2], (2, 3], (3, 5], (5, 6], (6, 9], (9, 13], (13, 20], (20, 89]]	100.0
6mo_body_weight_change		[<=-15%-, -15%,-10%, -10%-0, 0-10%, 10%-15%, >=15]	99.1
6mo_body_weight_difference (kilos)		[(-11, -7], (-7, -5], (-5, -3], (-3, -2], (-2, -1], (-1, 0], (0, 1], (1, 2], (2, 3], (3, 4], (4, 13]]	99.1
6mo_bod_weight_slope		[(-18, -1], (-1, 0], (0, 22]]	99.1
6mo_fall_slope		[(0, 1], (1, 30]]	100.0
6mo_hospitalizations_slope		[0, 1, 2, 3, 4, 5, 6]	100.0
6mo_mean_systolic_blood_pressure (mmHg)		[<80, 80-100, 100-120, 120-140, >140]	77.7
adherence_to_treatment		[independent, partially dependent, dependent]	99.4
admission_adherence_to_treatment		[independent, partially dependent, dependent]	99.8
admission_coherent_behavior		[independent, partially dependent, dependent]	99.8
admission_coherent_speech		[independent, partially dependent, dependent]	99.8
admission_distance_purchasing		[independent, partially dependent, dependent]	99.8
admission_do_all_house works		[independent, partially dependent, dependent]	99.8
admission_eating_autonmy		[independent, partially dependent, dependent]	99.8
admission_fecal_hygiene_elimination		[independent, partially dependent, dependent]	99.8
admission_indoor_mobility		[independent, partially dependent, dependent]	99.8
admission_level_dependency		[independent, partially dependent, dependent]	99.8

admission_lower_body_dressing		[independent, partially dependent, dependent]	99.8
admission_lower_body_grooming		[independent, partially dependent, dependent]	99.8
admission_meal_preparation		[independent, partially dependent, dependent]	99.8
admission_middle_body_dressing		[independent, partially dependent, dependent]	99.8
admission_oriented_in_places		[independent, partially dependent, dependent]	99.8
admission_oriented_in_time		[independent, partially dependent, dependent]	99.8
admission_outdoor_mobility		[independent, partially dependent, dependent]	99.8
admission_participate_in_leisure		[independent, partially dependent, dependent]	99.8
admission_personnal_management		[independent, partially dependent, dependent]	99.8
admission_public_transportation_use		[independent, partially dependent, dependent]	99.8
admission_remote_communication		[independent, partially dependent, dependent]	99.8
admission_table_ustensils_use		[independent, partially dependent, dependent]	99.8
admission_transfers_dependence		[independent, partially dependent, dependent]	99.8
admission_upper_body_dressing		[independent, partially dependent, dependent]	99.8
admission_upper_body_grooming		[independent, partially dependent, dependent]	99.8
admission_urinary_hygiene_elimination		[independent, partially dependent, dependent]	99.8
age (years)		[(44, 72], (72, 78], (78, 81], (81, 83], (83, 85], (85, 86], (86, 87], (87, 88], (88, 89], (89, 90], (90, 91], (91, 92], (92, 93], (93, 94], (94, 95], (95, 96], (96, 98], (98, 113]]	100.0
aggravated disease		[absence, presence]	100.0
aid_for_mobility		[absence, presence]	100.0
antibiotics		[absence, presence]	100.0
antidementia_drugs		[absence, presence]	100.0
antidepressants		[absence, presence]	100.0
antidiabetics		[absence, presence]	100.0
antiepileptics		[absence, presence]	100.0
antiosteoporotic_drugs		[absence, presence]	100.0
antiparkinson_drugs		[absence, presence]	100.0
antipsychotics		[absence, presence]	100.0
arthritis		[absence, presence]	100.0
asthma		[absence, presence]	100.0
BMI (kg/m²)		[(6,0, 17], (17, 18], (18, 20], (20, 21], (21, 22], (22, 23], (23, 24], (24, 25], (25, 26], (26, 27], (27, 28], (28, 30], (30, 32], (32, 89]]	69.4
bpco		[absence, presence]	100.0

bronchodilatators		[absence, presence]	100.0
cancer		[absence, presence]	100.0
cardiovascular_disease		[absence, presence]	100.0
cholesterol_lowering_ drugs		[absence, presence]	100.0
chronic_renal_failure		[absence, presence]	100.0
coherent_behavior		[independent, partially dependent, dependent]	99.4
coherent_speech		[independent, partially dependent, dependent]	99.4
coronary_heart_disease		[absence, presence]	100.0
corticosteroids		[absence, presence]	100.0
delay_after_admission (month)		[(0, 7], (7,13], (13, 18], (18, 24], (24,31], (31, 41], (41,53], (53,68], (64, 84], (84,120]]	100.0
denutrition		[absence, presence]	100.0
depression		[absence, presence]	100.0
diabetes_type_1		[absence, presence]	100.0
diabetes_type_2		[absence, presence]	100.0
diabetes		[absence, presence]	100.0
difference_levels_ dependency		[-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5]	99.4
distance_purchasing		[independent, partially dependent, dependent]	99.4
do_all_houseworks		[independent, partially dependent, dependent]	99.4
eating_autonmy		[independent, partially dependent, dependent]	99.4
fall_number		[0, 1, 2, 3+]	100.0
fecal_hygiene_ elimination		[independent, partially dependent, dependent]	99.4
femoral_fracture_history		[absence, presence]	100.0
femoral_neck_fracture_ history		[absence, presence]	100.0
hearing_impaired		[absence, presence]	100.0
heart_failure		[absence, presence]	100.0
height (cm)		[(100, 148], (148, 150], (150, 153], (153, 154], (154, 155], (155, 156], (156, 157], (157, 158], (158, 160], (160, 161], (161, 162], (162, 163], (163, 165], (165, 167], (167, 169], (169, 173], (173, 200]]	69.9
hip_fracture_history		[absence, presence]	100.0
hospitalization_history		[absence, presence]	100.0
hospitalization_number		[0, 1, 2, 3, 4+]	100.0
hypertension		[absence, presence]	100.0
hyperthyroid		[absence, presence]	100.0
incontinence		[absence, presence]	100.0
indoor_mobility		[independent, partially dependent, dependent]	99.4
infection		[absence, presence]	100.0

IV_fluid_infusion		[absence, presence]	100.0
last_attendance_leisure		[never, +6mo, 3mo-6mo, 1mo-3mo, 1mo]	100.0
last_body_weight (kilos)		[(20, 41], (41, 45], (45, 47], (47, 50], (50, 52], (52, 54], (54, 56], (56, 58], (58, 60], (60, 63], (63, 65], (65, 67], (67, 70], (70, 73], (73, 76], (76, 81], (81, 89], (89, 200]]	99.1
last_fall		[never, +6mo, 3mo-6mo, 1mo-3mo, 1mo]	100.0
last_hospitalization		[never, +6mo, 3mo-6mo, 1mo-3mo, 1mo]	100.0
level_dependency		[independent, partially dependent, dependent]	99.4
liver_disease		[absence, presence]	100.0
lower_body_dressing		[independent, partially dependent, dependent]	99.4
lower_body_grooming		[independent, partially dependent, dependent]	99.4
malnutrition		[absence, presence]	100.0
meal_preparation		[independent, partially dependent, dependent]	99.4
med_hypotension		[absence, presence]	100.0
med_steroide		[absence, presence]	100.0
middle_body_dressing		[independent, partially dependent, dependent]	99.4
neurologic_disease		[absence, presence]	100.0
non_steroidal_ antiinflammatory_drugs		[absence, presence]	100.0
number_attendance_ leisure		[(0, 2], (2, 4], (4, 8], (8, 14], (14, 22], (22, 33], (33, 47], (47, 66], (66, 90], (90, 124], (124, 172], (172, 248], (248, 368], (368, 606], (606, 1000]]	100.0
opioids		[absence, presence]	100.0
oriented_in_places		[independent, partially dependent, dependent]	99.4
oriented_in_time		[independent, partially dependent, dependent]	99.4
osteoporosis		[absence, presence]	100.0
outdoor_mobility		[independent, partially dependent, dependent]	99.4
participate_in_cultural/ sports_activities		[independent, partially dependent, dependent]	99.4
personnal_management		[independent, partially dependent, dependent]	99.4
polyarthritis		[absence, presence]	100.0
pressure_ulcer		[absence, presence]	100.0
psychostimulants		[absence, presence]	100.0
psychotropics		[absence, presence]	100.0
public_transportation_ use		[independent, partially dependent, dependent]	99.4
remote_communication		[independent, partially dependent, dependent]	99.4
rhythm_disorder		[absence, presence]	100.0

sex		[absence, presence]	100.0
shoulder_fracture_ history		[absence, presence]	100.0
smoker		[absence, presence]	100.0
spine_fracture_history		[absence, presence]	100.0
table_ustensils_use		[independent, partially dependent, dependent]	99.4
thyroid_disease		[absence, presence]	100.0
transfers_dependence		[independent, partially dependent, dependent]	99.4
upper_body_dressing		[independent, partially dependent, dependent]	99.4
upper_body_grooming		[independent, partially dependent, dependent]	99.4
urinary_hygiene_ elimination		[independent, partially dependent, dependent]	99.4
visually_impaired		[absence, presence]	100.0
wrist_fracture_history		[absence, presence]	100.0

Annexe : Réseaux Bayésiens complets

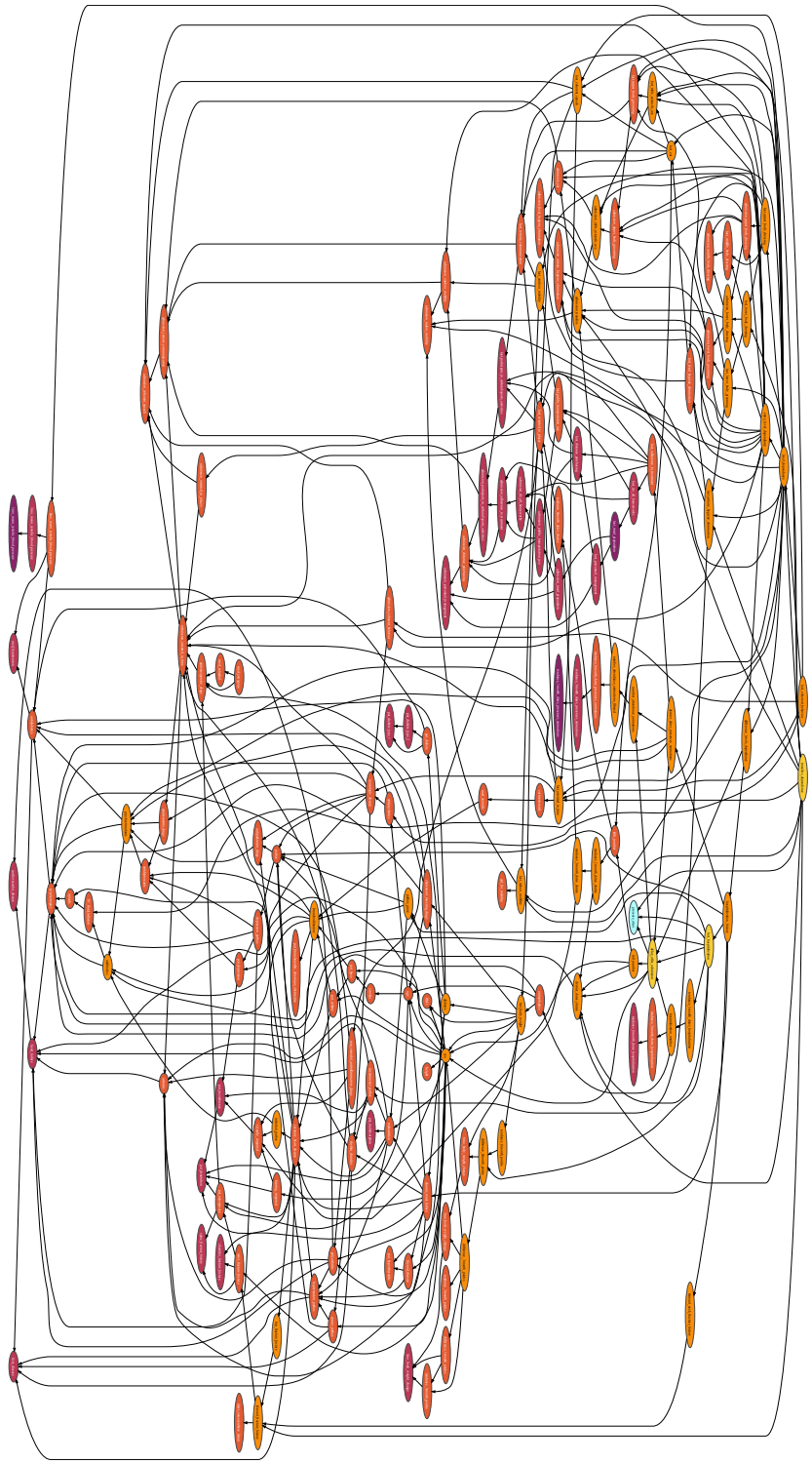
B

Réseau bayésien complet obtenu pour chaque cible. La couleur des variables indique la distance à la cible au sens des couvertures de Markov.

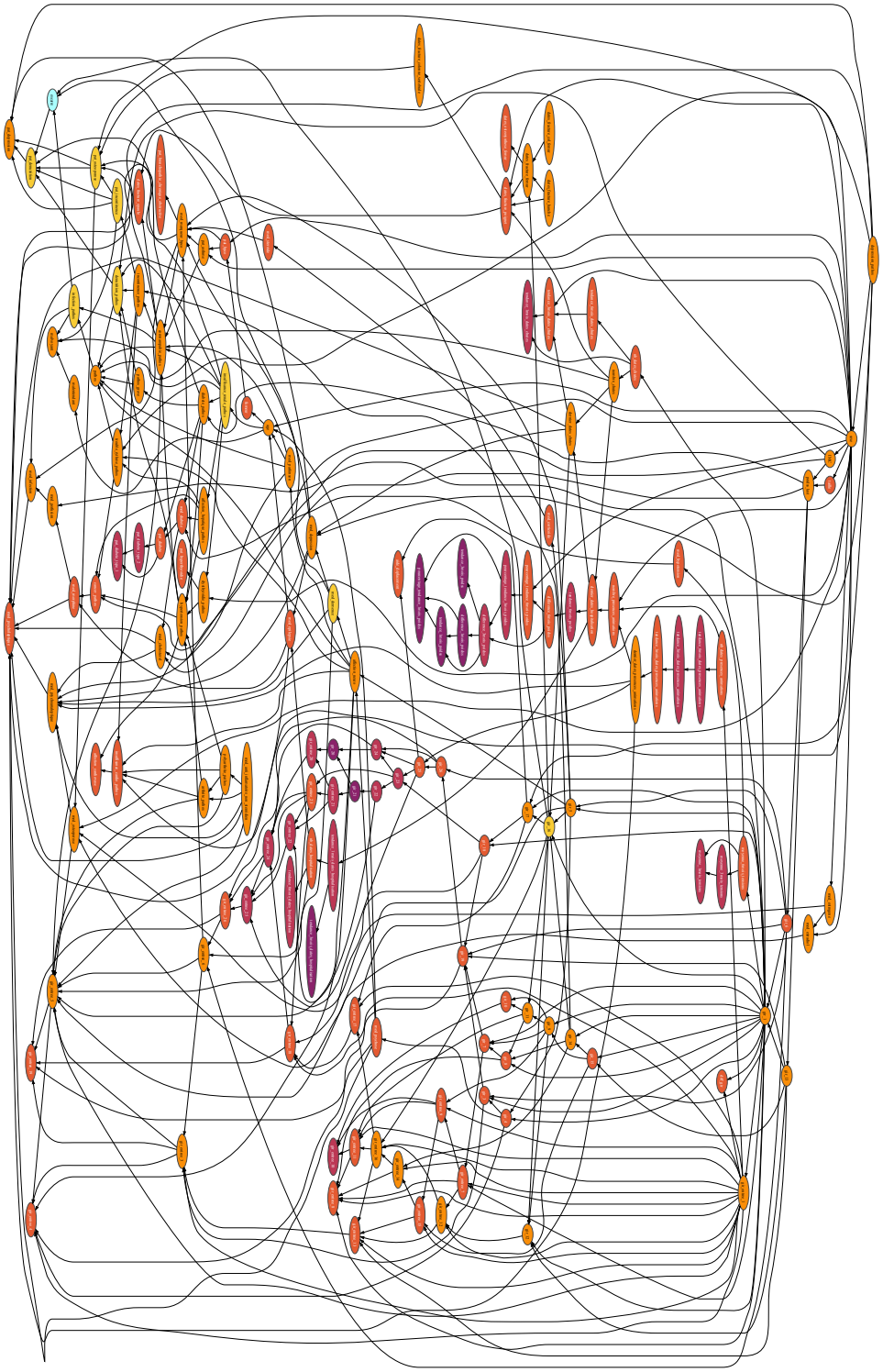
Couleur	Signification
	Cible
	Frontière de Markov
	Couverture de Markov de niveau 2
	Couverture de Markov de niveau 3
	Couverture de Markov de niveau 4
	Couverture de Markov de niveau 5

Tab. B.1. : Légende des couleurs des couvertures de Markov.

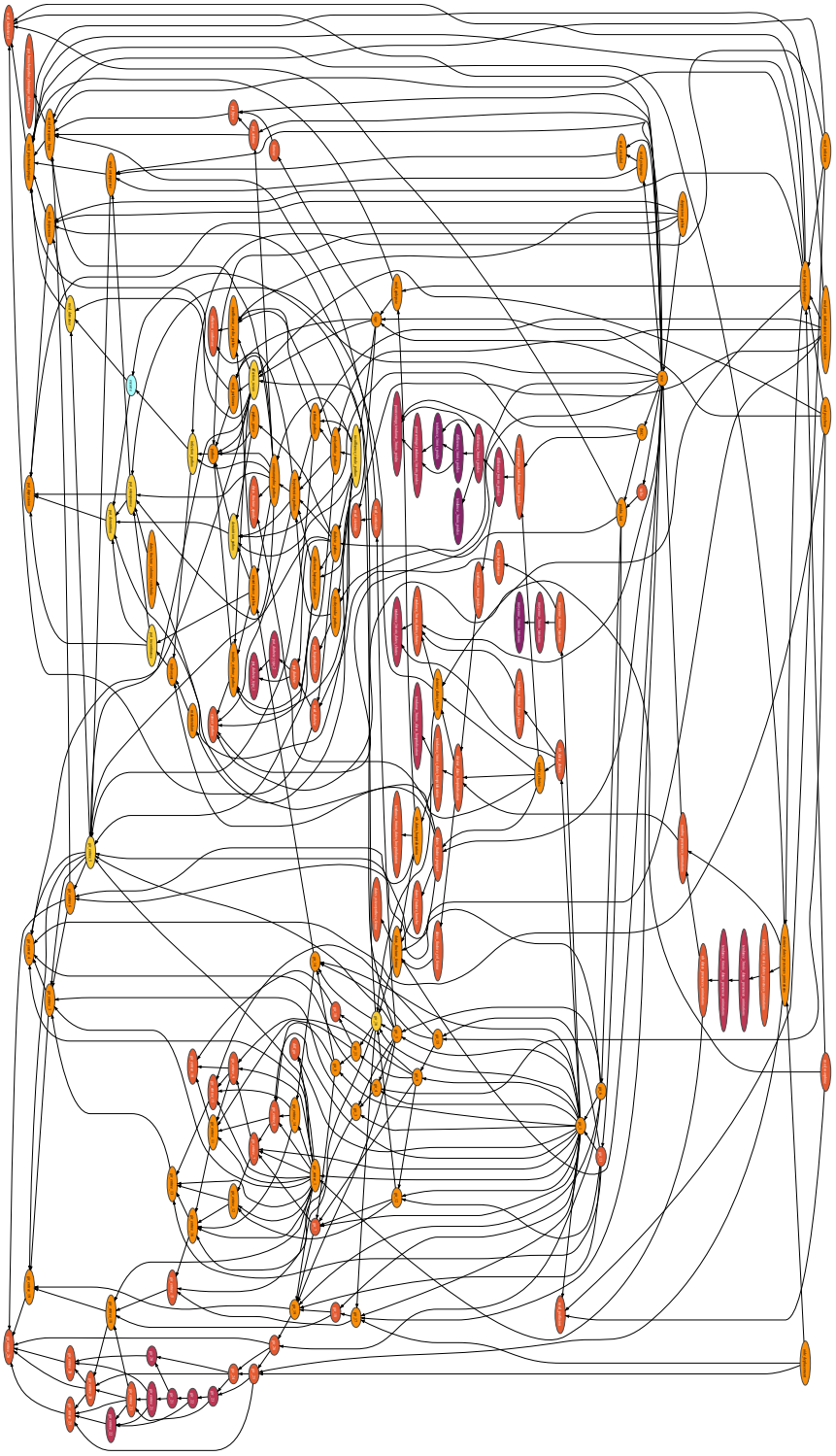
B.1 Cible : escarre 1 mois avant



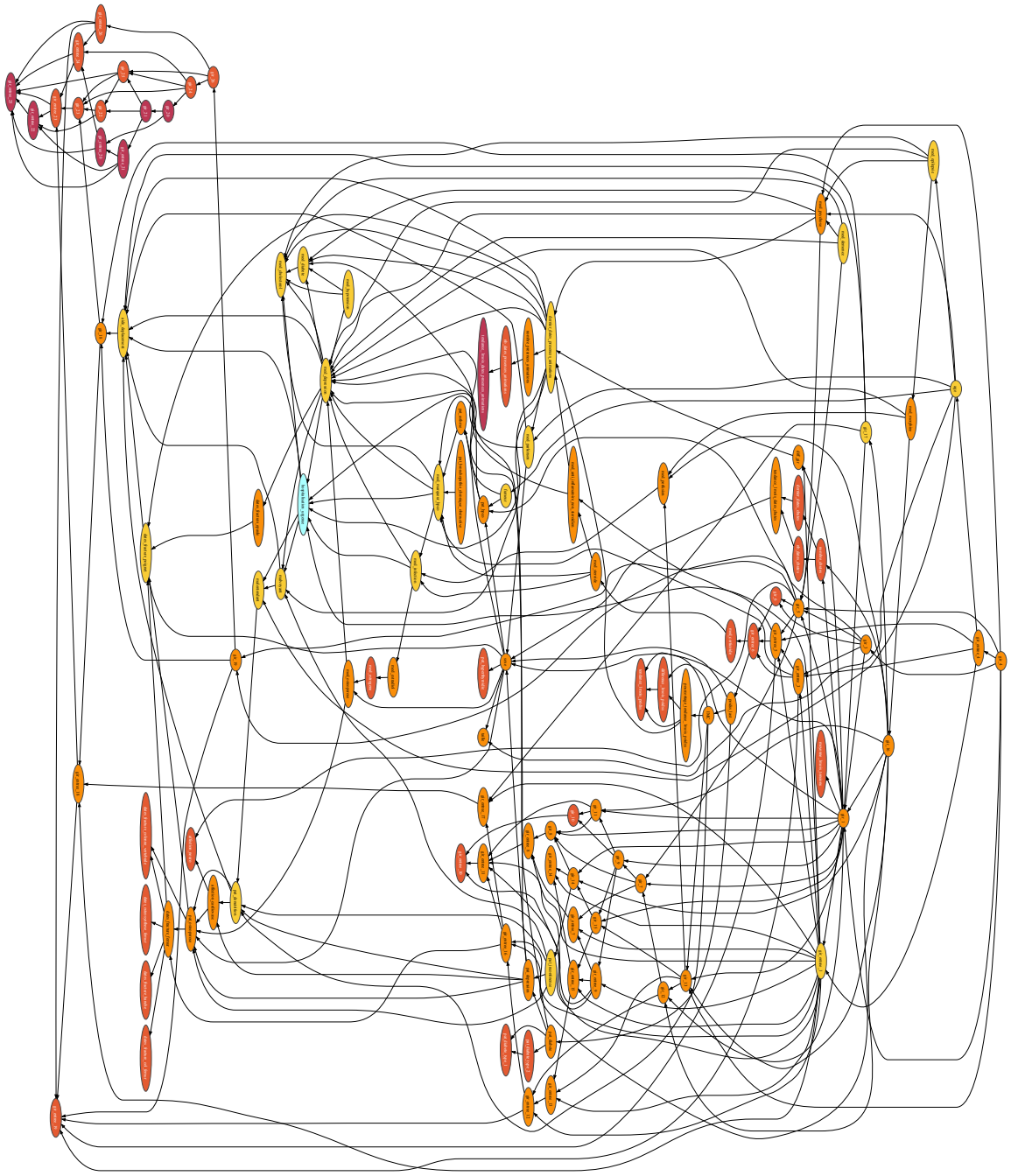
B.2 Cible : escarre 2 mois avant



B.3 Cible : escarre 3 mois avant



B.4 Cible : hospitalisation en urgence



B.5 Cible : fracture

C

Annexe : Tables des résultats étendus

C.1 Scores des classifieurs pour la prédiction d'escarre 1 mois avant

Modèle	F2-Score	F1-Score	Précision/ VPP	Rappel/ Sensibilité	Accuracy	Specificity	NPV
BNClassifier	0.67 (0.01)	0.47 (0.01)	0.32 (0.01)	0.94 (0.04)	0.40 (0.03)	0.18 (0.05)	0.88 (0.02)
QDA	0.50 (0.08)	0.45 (0.03)	0.39 (0.03)	0.56 (0.14)	0.60 (0.08)	0.62 (0.17)	0.78 (0.02)
Naive Bayes	0.47 (0.01)	0.55 (0.01)	0.38 (0.01)	0.63 (0.01)	0.60 (0.00)	0.58 (0.01)	0.79 (0.00)
Decision Tree	0.39 (0.01)	0.38 (0.01)	0.37 (0.01)	0.39 (0.01)	0.63 (0.00)	0.73 (0.00)	0.75 (0.00)
MLP	0.37 (0.04)	0.39 (0.03)	0.44 (0.01)	0.36 (0.05)	0.68 (0.01)	0.81 (0.04)	0.76 (0.01)
XgBoost	0.33 (0.01)	0.39 (0.01)	0.56 (0.01)	0.30 (0.01)	0.73 (0.00)	0.91 (0.00)	0.76 (0.00)
AdaBoost	0.24 (0.01)	0.30 (0.01)	0.56 (0.01)	0.21 (0.01)	0.72 (0.00)	0.93 (0.00)	0.74 (0.00)
Logistic Regression	0.21 (0.01)	0.27 (0.01)	0.56 (0.01)	0.18 (0.01)	0.72 (0.00)	0.94 (0.00)	0.74 (0.00)
Random Forest	0.14 (0.01)	0.20 (0.01)	0.65 (0.02)	0.12 (0.01)	0.72 (0.00)	0.97 (0.00)	0.73 (0.00)

Tab. C.1. : Métriques (moyenne (écart-type)) évaluant les performances des différents modèles pour la prédiction à un mois des escarres chez les résidents en EHPAD.

C.2 Scores des classifieurs pour la prédiction d'escarre 2 mois avant

Modèle	F2-score	F1-Score	Précision/ VPP	Rappel/ Sensibilité	Accuracy
Logistic Regression	0.10 [0.09, 0.11]	0.15 [0.14, 0.16]	0.50 [0.47, 0.54]	0.09 [0.08, 0.10]	0.72 [0.72 0.73]
Decision Tree	0.35 [0.33, 0.36]	0.34 [0.33, 0.35]	0.33 [0.32, 0.34]	0.35 [0.33, 0.37]	0.62 [0.61, 0.63]
Random Forest	0.07 [0.06, 0.08]	0.10 [0.09, 0.11]	0.51 [0.48, 0.55]	0.06 [0.05, 0.06]	0.72 [0.72, 0.73]
MLP	0.33 [0.24, 0.41]	0.34 [0.28, 0.40]	0.37 [0.36, 0.39]	0.32 [0.21, 0.42]	0.66 [0.63, 0.69]
AdaBoost	0.10 [0.09, 0.12]	0.15 [0.13, 0.17]	0.48 [0.45, 0.52]	0.09 [0.07, 0.10]	0.72 [0.71, 0.73]
Naive Bayes	0.54 [0.50, 0.58]	0.45 [0.44, 0.46]	0.35 [0.34, 0.37]	0.62 [0.54, 0.70]	0.58 [0.54, 0.62]]
QDA	0.50 [0.35, 0.65]	0.42 [0.36, 0.48]	0.34 [0.31, 0.38]	0.57 [0.31, 0.83]	0.57 [0.45, 0.69]
XgBoost	0.21 [0.20, 0.22]	0.26 [0.25, 0.28]	0.47 [0.44, 0.49]	0.18 [0.17, 0.2]	0.71 [0.71, 0.72]
BNClassifier	0.64 [0.62, 0.65]	0.46 [0.45, 0.47]	0.32 [0.31, 0.32]	0.85 [0.84, 0.87]	0.45 [0.43, 0.46]

Tab. C.2. : Métriques (moyenne [IC à 95%]) évaluant les performances des différents modèles pour la prédiction à deux mois des escarres chez les résidents en EHPAD.

C.3 Scores des classifieurs pour la prédiction d'escarre 3 mois avant

Modèle	F2-score	F1-Score	Précision/ VPP	Rappel/ Sensibilité	Accuracy
Logistic Regression	0.07 [0.06, 0.08]	0.10 [0.09, 0.12]	0.51 [0.47, 0.55]	0.06 [0.05, 0.07]	0.73 [0.73, 0.74]
Decision Tree	0.33 [0.31, 0.35]	0.33 [0.31, 0.34]	0.32 [0.30, 0.33]	0.34 [0.32, 0.36]	0.62 [0.62, 0.63]
Random Forest	0.05 [0.05, 0.06]	0.08 [0.07, 0.09]	0.52 [0.48, 0.55]	0.04 [0.04, 0.05]	0.73 [0.73, 0.74]
MLP	0.32 [0.27, 0.36]	0.33 [0.30, 0.36]	0.36 [0.34, 0.39]	0.31 [0.24, 0.37]	0.66 [0.64, 0.69]
AdaBoost	0.10 [0.09, 0.12]	0.15 [0.13, 0.17]	0.48 [0.45, 0.52]	0.09 [0.07, 0.1]	0.72 [0.71, 0.73]
Naive Bayes	0.53 [0.49, 0.57]	0.44 [0.42, 0.45]	0.34 [0.32, 0.36]	0.61 [0.53, 0.70]	0.58 [0.53, 0.62]
QDA	0.50 [0.35, 0.65]	0.42 [0.36, 0.48]	0.34 [0.31, 0.38]	0.57 [0.31, 0.83]	0.57 [0.45, 0.69]
XgBoost	0.21 [0.20, 0.22]	0.26 [0.25,0.28]	0.47 [0.44, 0.49]	0.18 [0.17, 0.20]	0.71 [0.71, 0.72]

Tab. C.3. : Métriques (moyenne [IC à 95%]) évaluant les performances des différents modèles pour la prédiction à trois mois des escarres chez les résidents en EHPAD.

C.4 Scores des classifieurs pour la prédiction d'hospitalisation en urgence

Modèle	F2-score	F1-Score	Précision/ VPP	Rappel/ Sensibilité	Accuracy
Logistic Regression	0.15 [0.15, 0.16]	0.2 [0.2, 0.21]	0.45 [0.43, 0.46]	0.13 [0.13, 0.14]	0.73 [0.72, 0.73]
Decision Tree	0.38 [0.36, 0.38]	0.37 [0.36, 0.38]	0.36 [0.35, 0.37]	0.38 [0.36, 0.39]	0.65 [0.65, 0.66]
Random Forest	0.09 [0.08, 0.1]	0.13 [0.12, 0.14]	0.44 [0.41, 0.48]	0.07 [0.07, 0.08]	0.73 [0.73, 0.74]
MLP	0.29 [0.24, 0.35]	0.33 [0.28, 0.37]	0.4 [0.39, 0.42]	0.28 [0.21, 0.34]	0.7 [0.69, 0.71]
AdaBoost	0.17 [0.16, 0.18]	0.22 [0.21, 0.23]	0.44 [0.42, 0.46]	0.14 [0.14, 0.15]	0.72 [0.72, 0.73]
Naive Bayes	0.50 [0.48, 0.51]	0.46 [0.44, 0.47]	0.4 [0.39, 0.41]	0.53 [0.51, 0.55]	0.66 [0.66, 0.67]
QDA	0.63 [0.56, 0.69]	0.45 [0.41, 0.48]	0.3 [0.24, 0.37]	0.88 [0.65, 1.11]	0.42 [0.20, 0.63]
XgBoost	0.22 [0.2, 0.24]	0.27 [0.25, 0.29]	0.45 [0.43, 0.46]	0.2 [0.18, 0.21]	0.72 [0.72, 0.73]
BNClassifier	0.66 [0.64, 0.67]	0.52 [0.51, 0.53]	0.38 [0.37, 0.39]	0.8 [0.77, 0.83]	0.6 [0.59, 0.62]

Tab. C.4. : Métriques (moyenne [IC à 95%]) évaluant les performances des différents modèles pour la prédiction d'hospitalisation en urgence chez les résidents en EHPAD.

C.5 Scores des classifieurs pour la prédiction de fracture

Modèle	F2-score	F1-Score	Précision/ VPP	Rappel/ Sensibilité	Accuracy
Logistic Regression	0.01 [0.0, 0.02]	0.02 [0.01, 0.03]	0.61 [0.34, 0.87]	0.01 [0.0, 0.01]	0.95 [0.95, 0.95]
Decision Tree	0.16 [0.14, 0.18]	0.15 [0.13, 0.17]	0.14 [0.12, 0.15]	0.17 [0.15, 0.19]	0.90 [0.90, 0.91]
Random Forest	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.02 [0.00, 0.06]	0.0 [0.0, 0.0]	0.95 [0.95, 0.95]
MLP	0.12 [0.07, 0.18]	0.14 [0.1, 0.18]	0.2 [0.15, 0.24]	0.11 [0.05, 0.18]	0.93 [0.92, 0.95]
AdaBoost	0.0 [0.0, 0.01]	0.0 [0.0, 0.01]	0.39 [0.0, 1.0]	0.0 [0.0, 0.0]	0.95 [0.95, 0.95]
Naive Bayes	0.26 [0.25, 0.28]	0.13 [0.12, 0.14]	0.07 [0.07, 0.08]	0.80 [0.73, 0.86]	0.48 [0.41, 0.54]
QDA	0.24 [0.22, 0.27]	0.12 [0.10, 0.13]	0.06 [0.05, 0.07]	0.89 [0.70, 1.0]	0.33 [0.14, 0.52]
XgBoost	0.10 [0.09, 0.11]	0.14 [0.12, 0.16]	0.58 [0.47, 0.68]	0.08 [0.07, 0.09]	0.95 [0.95, 0.95]
BNClassifier	0.31 [0.28, 0.33]	0.17 [0.16, 0.19]	0.1 [0.09, 0.11]	0.62 [0.55, 0.69]	0.7 [0.67, 0.73]

Tab. C.5. : Métriques (moyenne [IC à 95%]) évaluant les performances des différents modèles pour la prédiction de fracture chez les résidents en EHPAD.