

Louvain School of Management

Prédiction des hospitalisations à partir des données patients dans un réseau de structures d'hébergement médico-sociales belges

Mémoire de Master

Auteur : CHARLE Taddéo
Promoteur : Prof. CREEMERS Stephan
Année académique 2025-2026
*Mémoire présenté en vue de l'obtention du grade de Master en
Ingénieur de Gestion*

DECLARATION

La déclaration ci-dessous est **obligatoire et doit apparaître à la première page (après la couverture) de votre TFE**

DECLARATION D'UTILISATION D'IA GÉNÉRATIVE

Déclarations relatives à l'usage responsable de l'IA générative	Réponse
Le contenu de mon/notre TFE est mon/notre production originale, même si j'ai/nous avons utilisé l'IA générative pour m'/nous aider à le préparer.	☒ oui ☐ non
Je/nous maîtrise/maîtrisons les théories, outils et méthodes que j'utilise/nous utilisons pour le mémoire.	☒ oui ☐ non
J'ai/nous avons vérifié que le contenu présenté dans le mémoire est exact et valide, et qu'il ne contient ni hallucination, ni fausse référence, etc.	☒ oui ☐ non
J'atteste/nous attestons maîtriser le résultat final et être capable(s) de l'expliquer et de répondre aux questions à son sujet.	☒ oui ☐ non
J'ai/nous avons pris toutes les mesures nécessaires afin de ne pas donner accès à des informations ou données confidentielles à des outils d'IA générative	☒ oui ☐ non

SIGNATURE

Signez votre déclaration

Auteur Nom de famille, prénom(s)	Date	Signature
1 Taddéo CHARLE	27-05-26	
2		

NOTE

Pour rédiger ce document, DeepL et ChatGPT5 ont été utilisés afin de suggérer des formulations, mais le texte est original, à l'exception d'extraits explicitement mentionnés comme venant d'autres sources, mentionnées en notes de bas de page.

Résumé

Contexte : La prédiction dans le milieu hospitalier est encore à un stade expérimental dans beaucoup de domaines de la médecine et très peu de documentations existent dans les milieux médico-sociaux. Pourtant, l'hospitalisation des patients placés dans des foyers pour personnes atteintes de troubles mentaux ou psychologiques constitue un problème organisationnel pour les établissements. En effet, ces résidences manquent d'outils de prévention axés sur les risques des patients.

Objectif : Ce mémoire propose une comparaison de 3 modèles prédictifs différents, capables de prédire les hospitalisations futures dans les 30 jours sur base de données quotidiennes des résidents. L'étape finale du pipeline est la création automatique d'une liste de patients considérés comme « à risque » par établissement. Cette liste pourra être utilisée par le personnel soignant pour prévenir une hospitalisation.

Méthodologie : L'étude s'appuie sur des données réelles de 1 356 patients répartis dans 43 foyers belges. Ces données comprennent 73 416 observations longitudinales (mesure de l'indice de masse corporelle) ainsi que 18 variables prédictives. Trois architectures sont comparées : une Forêt Aléatoire (Random Forest), un Gradient Boosté Extrême (Extreme Gradient Boosting) avec une optimisation bayésienne des hyperparamètres via l'ajout du cadre Optuna (Akiba et al., 2019), et un réseau de neurones artificiel de type Perceptron Multicouche (Multilayer Perceptron). Les performances sont présentées via le cadre SHAP (SHapley Additive exPlanations), via la courbe Précision-Rappel, via la courbe AUC-ROC (Area Under the Receiver Operating Characteristic curve), et le coefficient Kappa est calculé au point de Youden comme référence neutre.

Résultats : le modèle Gradient Boosté Extrême avec Optuna obtient les meilleurs résultats sur tous les critères. L'AUC-ROC atteint 0,736 et témoigne d'une bonne capacité de discrimination. À titre de comparaison, un modèle aléatoire obtient 0,5 et les études menées sur des populations similaires obtiennent des valeurs comprises entre 0,71 et 0,83. La précision moyenne s'élève à 0,088 contre 0,024 pour une référence aléatoire, ce qui signifie que le modèle est environ 3,7 fois plus performant qu'un tirage au hasard dans un contexte où les hospitalisations restent rares (2,6% des cas). Le premier décile des scores concentre 36,8% des hospitalisations réelle et les 2 premiers déciles en capturent 52,7%. En ciblant seulement 20%

des patients les plus à risque, il est donc possible d'anticiper plus de la moitié des hospitalisations.

Le coefficient Kappa obtient 0,062, cette valeur assez faible est liée au fort déséquilibre des données (36,8 cas négatifs pour un cas positif). Elle doit être comparée entre modèles plutôt qu'en valeur absolue.

En conclusion : Dans un contexte parmi les plus difficiles de la prédiction médicale (population vulnérable, grand déséquilibre, variables limitées), le pipeline produit des résultats proches de la littérature et crée une liste de patients à surveiller au format tableur pour permettre au personnel soignant de prioriser leur attention.

Table des Matières

Résumé.....	1
Liste des tableaux.....	5
Liste des graphiques et illustrations	5
Introduction.....	6
1.1. Contexte.....	6
1.2. Problématique.....	6
1.3. Objectif et Question de Recherche	7
1.4. Démarche méthodologique.....	7
1.5. Structure du mémoire	8
2. Revue de la littérature	9
2.1. La prédiction des hospitalisations dans la littérature.....	9
2.2. Prédiction du nombre d'arrivées de patients	9
2.3. Performance comparée des algorithmes d'apprentissage automatique	10
2.4. Optimisation des hyperparamètres	11
2.5. Importance des prédicteurs fonctionnels et multidimensionnels.....	12
2.6. Déséquilibre des classes, métrique d'évaluation et d'optimisation	12
2.7. Gestion du déséquilibre	13
2.8. Synthèse et positionnement de la Recherche.....	14
3. Analyse et préparation des données	15
3.1. Sources et structure des données	15
3.2. Nettoyage et prétraitement des données	15
3.3. Construction des variables prédictives	17
3.3.1. Construction de la variable cible.....	17
3.3.2. Prédicteurs cliniques et dynamiques	17
3.4. Analyse des corrélations de Pearson	18
3.5. Prévention des fuites de données temporelles	19
4. Stratégie de modélisation.....	21
4.1. Principe de la fenêtre glissante	21
4.2. Algorithme 1 : Forêt Aléatoire	21
4.3. Algorithme 2 : Gradient Boosté Extrême avec optimisation bayésienne Optuna	22
4.3.1. Architecture du Gradient Boosté Extrême	22
4.3.2. Optimisation bayésienne via Optuna.....	23
4.4. Algorithme 3 : Perceptron Multicouche	24
4.5. Approche sans seuil fixe : la courbe Précision-Rappel	25
5. Résultats et évaluations.....	27
5.1. Performance globale.....	27

5.1.1.	Métriques principales	27
5.2.	Interprétabilité des modèles	28
5.2.1.	Introduction à la méthode SHAP	28
5.2.2.	Résultats de la méthode SHAP.....	29
5.3.	Evaluation de la courbe AUC-ROC	31
5.4.	Évaluation de la courbe Précision-Rappel.....	32
5.4.1.	Coefficient Kappa de Cohen	33
5.5.	Liste des surveillances opérationnelles.....	34
6.	Limites éthiques et techniques	36
6.1.	Limites techniques.....	36
6.1.1.	Possibilité d'une fiabilité parfaite et tension entre Précision et Rappel.....	36
6.1.2.	Événements positifs limités.....	36
6.1.3.	Absence de variables fonctionnelles et comportementales	36
6.2.	Limites éthiques.....	37
6.2.1.	Protection des données personnelles	37
6.2.2.	Biais algorithmique	37
6.2.3.	Risque de dépendance excessive et des engagements cliniques	37
6.2.4.	Transparence, explicabilité et conformité réglementaire	37
6.2.5.	Validation prospective nécessaire	38
7.	Conclusion	39
8.	Bibliographie.....	41
9.	Annexes.....	46
	Annexe A : Fichiers de données	46
	A1. Graphique d'importances des variables	47
	A2. Paramètres du jeu de données	47
	A3. Meilleurs Hyperparamètres identifiés par Optuna (essai n° 41)	48
	A4. Analyse de l'importance des hyperparamètres via Optuna	50
	A5. Analyse par Déciles	50
	Annexe B : Description des 18 Variables du Jeu de Données	53
	Annexe C : Schémas des différents modèles	57
	C1. Schéma de la Forêt Aléatoire	57
	C2. Schéma du modèle Gradient Boosté Extrême	58
	C3. Schéma du Réseau de neurones du type Perceptron Multicouche	59
	Annexe D. Extrait du tableur de la liste de surveillance hebdomadaire par foyers	60
	Annexe E. Résultats de la méthode SHAP	60

Liste des tableaux

- *Tableau 1 : Comparaison des aires sous la courbe dans la littérature, page 11*
- *Tableau 2 : Caractéristique de la cohorte, page 20*
- *Tableau 3 : Comparaison des métriques de performances, page 27*

Liste des graphiques et illustrations

- *Illustration 1 : Pipeline du projet, page 8*
- *Illustration 2 : Corrélations de Pearson, page 18*
- *Illustration 3 : Importance globale des variables du modèle GBE Optuna via la méthode SHAP, page 29*
- *Illustration 4 : Sens et magnitude des effets sur le risque des variables du modèle GBE Optuna via la méthode SHAP, page 30*
- *Illustration 5 : Graphique d'une explication individuelle d'un patient (SHAP) pour le modèle Gradient Boosté Extrême, page 31*
- *Illustration 6 : Courbes ROC des 3 modèles, page 32*
- *Illustration 7 : Comparaison des courbes Précision-Rappel des 3 modèles, page 33*
- *Illustration 8 : Capture d'écran de la liste finale de surveillance des patients les plus à risque, page 35*

Introduction

1.1. Contexte

En Belgique, la direction générale des personnes handicapées (DG HAN) recensait en 2025 environ 582 000 personnes reconnues porteuses de handicap, soit approximativement 5% de la population totale belge (Statbel, 2025). Une partie de ces personnes ont besoin d'une prise en charge institutionnelle spécialisée.

La prise en charge des résidents pose certains défis organisationnels et cliniques. Une expérimentation menée en foyer spécialisé par Mangeney et al. (2014) a documenté les difficultés propres du secteur. L'étude a recensé la volatilité des profils cliniques, les difficultés d'accès aux soins spécialisés, et a souligné le besoin d'avoir des outils de coordination et de prévention. L'hospitalisation génère un coût financier substantiel, un stress accru pour le personnel soignant et une fragilisation de l'état de santé des résidents. Face à cette problématique, le secteur médico-social s'oriente progressivement vers des stratégies préventives et proactives.

C'est dans ce contexte qu'un réseau de 43 foyers médico-sociaux belges a mis à disposition ses données de suivi pour le développement d'un outil prédictif. La cohorte comprend 1 356 patients avec un total de 73 416 mesures d'Indice de Masse Corporelle (IMC), principales observations déclencheur des modèles prédictifs.

1.2. Problématique

Malgré les progrès de la médecine prédictive, la majorité des modèles de prédiction des hospitalisations ciblent le milieu hospitalier traditionnel ou les populations gériatriques générales. Les structures accueillant des personnes en situation de handicap mental ou psychique restent particulièrement lacunaires en outil d'aide à la décision. Par ailleurs, les données disponibles dans ces foyers sont souvent non structurées, dispersées dans plusieurs fichiers administratifs et présentent un fort déséquilibre entre périodes de stabilité clinique et hospitalisation. Dans la cohorte étudiée, les hospitalisations ne représentent que 2,6% des observations (1 898 sur 73 416), avec un rapport de déséquilibre de 36,8 (nombres d'observations négatives par rapport aux nombres d'observations positives).

Il existe donc un besoin crucial d'outils capables d'identifier précocement les résidents présentant un risque d'hospitalisation à partir des données de suivi quotidien sans avoir besoin

d'équipements médicaux coûteux ni de formation technique avancée pour le personnel soignant.

1.3. Objectif et Question de Recherche

Ce travail de recherche se situe à l'intersection de la littérature scientifique clinique, de l'apprentissage automatique et de la conception d'outils d'aide à la décision médicale. La question centrale du mémoire peut se formuler de la sorte :

Dans quelle mesure un modèle prédictif basé sur des données de patients réels peut-il anticiper les hospitalisations à un horizon de 30 jours dans un réseau de foyers médico-sociaux belges, et quel algorithme offre le meilleur équilibre entre performance prédictive et interprétabilité clinique ?

L'objectif opérationnel est de concevoir un pipeline de prévision produisant une liste de surveillance hebdomadaire par foyer. Cette liste pourrait être directement utilisable par les infirmiers sans formation technique avancée.

1.4. Démarche méthodologique

Le pipeline suit une progression en 5 étapes :

- Chargement et consolidation de 5 fichiers en un jeu de données analytiques de 73 416 observations et 18 variables.
- Construction des variables prédictives : indicateur dynamique d'antécédent hospitalier, de vitesses de perte ou gain de poids et de la charge thérapeutique.
- Séparation des identifiants des patients dans la partie entraînement et test des modèles (et non de l'observation), pour éviter que des mesures d'un même résident se retrouvent dans les 2 ensembles.
- Entraînement et comparaison de 3 types d'architectures différents : Forêt Aléatoire, Gradient Boosté Extrême avec optimisation bayésienne Optuna, et Perceptron Multicouche. Évaluation et interprétation des modèles sur la courbe de Précision-Rappel complète, via la courbe AUC-ROC, via le cadre SHAP et via le coefficient Kappa au point de Youden.
- Génération de la liste de surveillance hebdomadaire par foyer à partir des scores de probabilité bruts exportés en format tableur.

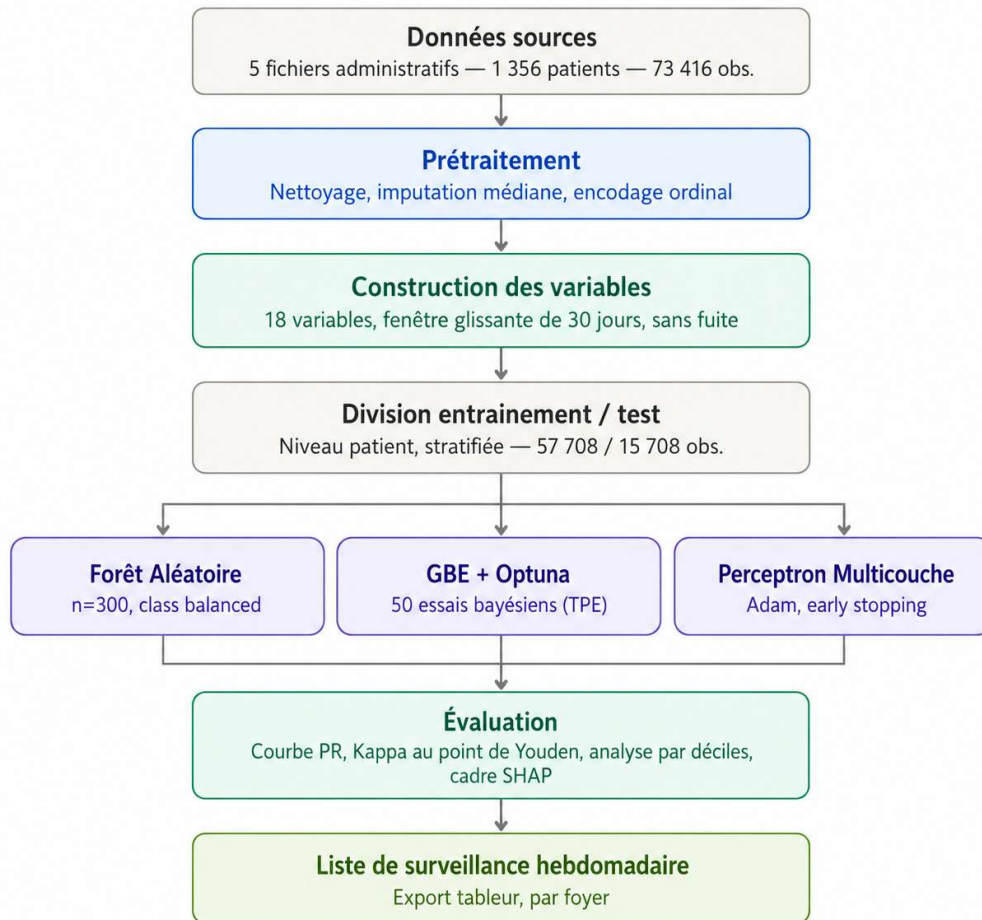


Illustration 1 : Pipeline du projet

1.5. Structure du mémoire

Le chapitre 2 passe en revue la littérature scientifique sur les modèles prédictifs des hospitalisations et le contexte médico-social. Le chapitre 3 décrit la préparation des données et la construction des variables prédictives. Le chapitre 4 expose la stratégie de modélisation et les algorithmes. Le chapitre 5 présente et interprète les résultats. Enfin, le chapitre 6 discute des limites éthiques et techniques.

2. Revue de la littérature

Ce chapitre passe en revue la littérature scientifique liée à la prédiction dans le milieu hospitalier, les algorithmes d'apprentissages appliqués aux données médicales et la gestion du déséquilibre des classes. L'objectif est de positionner ce mémoire à travers les différents travaux existants et d'identifier la plus-value qu'il apporte à la communauté scientifique.

2.1. La prédiction des hospitalisations dans la littérature

De nombreuses études dans le milieu hospitalier ont utilisé la Régression Logistique comme méthode de prédiction pour sa simplicité d'interprétation. Carrillo-Villa et al. (2020) l'appliquent à la prédiction des réadmissions hospitalières chez des personnes âgées fragiles, obtenant de bonnes performances mais soulignant ses limites face à certains profils cliniques. Même si la régression logistique est plus simple d'interprétation, le modèle est limité dans l'analyse des interactions complexes entre les variables et plus particulièrement dans les populations présentant des déficiences.

Les modèles à base d'arbres de décision comme les Forêts Aléatoires ont émergé comme des alternatives performantes. Song et al. (2022) obtiennent une aire sous la courbe de 0,83 pour la prédiction du risque d'hospitalisation chez des patients souffrant de maladies chroniques, confirmant la capacité de ces méthodes à capturer des interactions plus complexes entre variables.

Dans le domaine spécifique des personnes handicapées institutionnalisées, bien que l'étude soit centrée sur la télémédecine et non sur la prédiction, Manganey et al. (2014) ont documenté l'intérêt de la coordination numérique des soins en foyers spécialisés. Ils ont souligné la difficulté d'accès aux expertises médicales et à la valeur ajoutée des outils technologiques.

2.2. Prédiction du nombre d'arrivées de patients

La prédiction des flux de patients constitue un défi majeur pour les établissements de santé. Une surcharge non anticipée dégrade directement la qualité des soins et la satisfaction du personnel soignant (Moreno-Sánchez et al., 2024). La littérature s'est longtemps appuyée sur des méthodes statistiques classiques telles que les séries temporelles ARIMA, qui modélisent les tendances et les saisonnalités des arrivées à partir de l'historique. Ces approches ont progressivement été complétées par des modèles d'apprentissage automatique capables de capturer des relations non linéaires entre les arrivées de patients et des variables externes comme le calendrier, la météo, ou encore des indicateurs épidémiologiques. Une étude sur la

modélisation des arrivées aux urgences montre que les modèles d'apprentissage automatique surpassent les approches statistiques classiques dès que des variables contextuelles sont disponibles (Jiang et al., 2023). Porto et al. (2024) confirment ce constat en comparant 6 algorithmes sur 11 services d'urgence répartis dans 3 pays différents, avec des horizons de prévision allant de 7 à 45 jours. Les méthodes de Gradient Boostée et les Réseaux de Neurones obtiennent les meilleures performances, surtout lorsque les hyperparamètres sont optimisés par validation croisée. Les travaux s'inscrivent dans un contexte hospitalier ou d'urgence générale. La prédiction de flux dans les foyers médico-sociaux où les événements sont rares, constituent un problème structurellement différent et moins documenté. Ce qui justifie une approche spécifique sur les données de chaque résident plutôt que sur les volumes d'un service.

2.3. Performance comparée des algorithmes d'apprentissage automatique

Les algorithmes utilisés dans ce mémoire s'appuient sur des fondations théoriques. Chen et Guestrin (2016) introduisent XGBoost (eXtreme Gradient Boosting) comme un système de boosting par arbres, devenu une référence sur les données tabulaires en raison de sa capacité à gérer les valeurs manquantes et les variables hétérogènes. Biau et Scornet (2016) proposent une analyse théorique des Forêts Aléatoires, démontrant leur robustesse face au bruit. Bentjac et al. (2021) proposent une analyse comparative des variantes de Gradient Boosting, confirmant leur supériorité sur les autres familles d'algorithmes pour les données tabulaires structurées.

Le Gradient Boosté Extrême se distingue par ses performances sur les données tabulaires hétérogènes. Dhanka et Maini (2025) démontrent les performances du Gradient Boosté Extrême combiné à l'optimisation d'Optuna dans un contexte de prédiction de maladies cardiovasculaires. Le modèle atteint des performances supérieures aux approches non optimisées. L'algorithme obtient une sensibilité de 92,86% sur le jeu de test. L'étude confirme que le taux d'apprentissage est l'hyperparamètre influençant le plus la performance du modèle. Si nous comparons avec les résultats de ce mémoire, ce paramètre concentre à lui seul 72% de l'importance des hyperparamètres (voir Annexe A4, graphique de l'importance des hyperparamètres).

Kraus et al. (2023) prouvent que la Forêt Aléatoire surpasse les modèles linéaires généralisés et les machines à vecteurs de support pour la prédiction de test de mobilité chez des personnes âgées, grâce à sa capacité à identifier des combinaisons non linéaires de variables. Surtout dans

les contextes où les variables d'antécédents (historique médical et hospitalisation passée) interagissent avec des variables continues (poids et âge).

Pour les données tabulaires de taille modérée, les approches d'apprentissage profond telles que le Perceptron Multicouche ont plus de mal à surpasser les performances d'un Gradient Boosté Extrême. Shwartz-Ziv et Armon (2022) confirment ce constat, les méthodes de Gradient Boosté Extrême surpassent les architectures de « *deep learning* » sur un large panel de jeu de données tabulaires, dans la majorité des cas. Cependant, Friesner et al. (2024) montrent que le Perceptron Multicouche reste très efficace sur des séries temporelles médicales de grande ampleur. Le risque de surapprentissage (*overfitting*) constitue une limite bien documentée, Bejani et Ghatee (2021) en décris les méthodes de contrôle applicables aux réseaux profonds dans une revue. Les revues de Kandel et al. (2024) confirment que pour des jeux de données médicaux comportant moins de 100 000 observations, les méthodes de Gradient Boostées constituent généralement le meilleur compromis entre performance et interprétabilité clinique.

Le tableau en Annexe A6 synthétise les aires sous la courbe rapportée dans la littérature en regard de nos résultats :

Ces études doivent être comparée avec ce mémoire avec prudence car les résultats rapportés dans la littérature sont obtenus avec des jeux de données plus équilibrées, Dhanka et Maini (2025) travaillent sur des données dont la proportion de cas positif dépasse les 54%. Le nombre élevé de cas positifs leurs permettent d'obtenir des performances à 95% d'exactitudes. Sur notre cohorte où les hospitalisations (cas positifs) représentent seulement 2,6% des observations, notre modèle obtenant une aire sous la courbe ROC de 0,736 prouve qu'il a une capacité de discrimination réelle et non aléatoire.

2.4. Optimisation des hyperparamètres

L'optimisation des hyperparamètres constitue une étape déterminante dans la construction de modèles d'apprentissage automatique performants. Yang et Shami (2020) proposent une synthèse des stratégies disponibles, distinguant les approches par grille exhaustive, par recherche aléatoire et par optimisation bayésienne. Cette dernière famille est recommandée lorsque le coût de chaque évaluation est élevé. Bischl et al. (2023) décrivent les fondements algorithmiques de l'optimisation d'hyperparamètres et soulignent l'importance de la validation croisée imbriquée pour éviter tout biais d'optimisme dans l'estimation des performances. Dans le cadre spécifique des Forêts Aléatoires, Probst et al. (2019) identifient les hyperparamètres les plus influents, notamment le nombre d'arbres et le nombre de variables candidates à chaque

nœud. Cette dernière étude prouve qu'une recherche aléatoire bien calibrée peu atteindre des résultats comparables à une recherche exhaustive pour un coût de calcul nettement inférieur. Akiba et al. (2019) proposent Optuna, un cadre d'optimisation d'hyperparamètres de nouvelle génération reposant sur un échantillonnage adaptatif par *Tree-structured Parzen Estimator* (TPE). Ce framework permet d'explorer efficacement des espaces de recherche complexes tout en éliminant les configurations peu prometteuses. Son association avec le Gradient Boosté Extrême dans ce mémoire poursuit les recommandations de Dhanka et Maini (2025) qui en démontrent les propriétés dans un contexte médical comparable.

2.5. Importance des prédicteurs fonctionnels et multidimensionnels

Plusieurs articles dans la littérature scientifique s'accordent à dire que les variables strictement cliniques (hypertension, diabète, obésité) ne suffisent pas à assurer une précision optimale. Buenrostro-Mariscal et al. (2024) confirment ce constat pour des populations complexes.

L'antécédent hospitalier est universellement reconnu comme prédicteur de premier plan. Dans ce mémoire, les variables de nombre d'hospitalisations dans les 12 derniers mois et de jours écoulés depuis la dernière sortie dominent les classements d'importance dans tous les modèles.

Ce résultat est cohérent avec la littérature sur la fragilité et les réadmissions hospitalières (Carillo-Vega et al, 2020 ; Song et al., 2022).

2.6. Déséquilibre des classes, métrique d'évaluation et d'optimisation

Le déséquilibre des classes constitue le défi méthodologique central de ce mémoire. Avec un déséquilibre de 1898 observations positives sur un total de 73 416 (soit 2,6%), le rapport atteint 1 :36,8. Ce ratio signifie que pour chaque cas positif observé (une hospitalisation), les données présentent environ 37 périodes stables sans hospitalisation. Selon une revue reprenant 527 articles sur l'apprentissage à partir de données déséquilibrées publiée par Haixiang et al. (2017) dans *Expert Systems with Applications*, les données médicales constituent le domaine où le problème est le plus fréquemment rencontré en raison de la rareté naturelle des événements cliniques. Ils en déduisent que les performances se dégradent critiqueusement au-delà du ratio de 1 :30. Dans leur revue systématique sur l'apprentissage profond (deep learning) avec déséquilibre des classes, Johnson et Khoshgoftaar (2019) limitent ce ratio à 1 :20. Au-delà de ce seuil les architectures se dégradent peu importe leur niveau de sophistications. Notre ratio de 1 :36,8 dépasse les deux seuils identifiés dans ces revues, ce qui prouve que ce mémoire est structurellement difficile pour l'ensemble des algorithmes.

Ce déséquilibre entraîne une conséquence directe sur les performances atteignable. Luque et al. (2019), démontrent que toutes les métriques classiques (dont l'AUC-ROC, Area Under the Receiver Operating Characteristic curve) dépendent de ce ratio de déséquilibre. Cela signifie qu'un modèle obtenant un AUC de 0,74 sur notre cohorte (avec 2,6% de cas positifs) ne peut pas être directement comparé à un modèle obtenant une AUC de 0,74 avec un cas survenant dans 20% des cas. Dans ce mémoire, le modèle s'approche du plafond mathématique de ce qui est atteignable. Dans la seconde situation, il reste loin de son potentiel.

Askar et al. (2024) ont recensé 116 études de prédiction des hospitalisations par apprentissage automatique, la médiane du pourcentage d'évènement cible est entre 10 et 20% dans les cohortes retenues. Moins de 8% des études reprise dans leur revue ont moins de 5% de cas positifs. Ce constat signifie que la quasi-totalité des outils et des méthodes disponibles ont été développées et testées dans des conditions bien moins contraignantes que celles de notre étude. La comparaison directe des chiffres de performance est donc délicate.

2.7. Gestion du déséquilibre

Deux grandes familles de solutions existent dans la littérature pour gérer le déséquilibre, Le rééchantillonnage synthétique (SMOTE) et la pondération des classes. La méthode SMOTE créée par Chawla et al. En 2002 permet de générer des patients fictifs de la classe minoritaire par interpolation entre cas existants. Cette méthode permet de créer des profils intermédiaires entre les profils de 2 patients réels hospitalisés. Malgré sa popularité, elle crée de faux profils cliniquement incohérents lorsque les variables sont catégorielles ou ordinales, selon Fernandez et al. (2018) dans une revue publiée dans le *journal of Artificial Intelligence Research*. Dans ce mémoire, créer un patient fictif entre un patient avec une déficience légère et un patient avec une déficience profonde n'a pas de sens médical. Pour cette raison, ce pipeline privilégie la pondération des classes à toute création de données artificielles. (Buenrostro-Mariscal et al. (2025) ; Amanollahi et al. (2024)).

La pondération des classes, comparée à la méthode SMOTE, ne modifie pas les données en elles-mêmes mais agit directement sur la manière dont le modèle calcule et pénalise ses erreurs. Lors de l'entraînement chaque erreur de prédiction contribue à la fonction de perte du modèle pour guider les ajustements en interne. Une erreur sur un cas positif (une hospitalisation manquée) a, par défaut dans la fonction, le même poids qu'une erreur sur un cas négatif (une fausse alerte d'hospitalisation). La pondération des classes permet de multiplier la pénalité liées aux erreurs sur la classe minoritaire (hospitalisation) par un coefficient proportionnel au

déséquilibre (rapport entre observations positives et négatives équivalent à 1 :36,8). La pondération permet donc de pénaliser une hospitalisation 36,8 fois plus fortement que de déclencher une fausse alerte. La revue de Araf et al. (2024) publiée dans *Artificial Intelligence Review*, confirme que cette méthode constitue une approche prometteuse pour améliorer la précision et la fiabilité des modèles.

2.8. Synthèse et positionnement de la Recherche

La prédiction hospitalière dans les foyers médico-sociaux accueillant des adultes en situation de handicap mental ou psychique constitue un angle de recherche quasiment inexploré dans la littérature internationale. Les modèles existants se concentrent sur les populations gériatriques générales (Carrillo-Vega et al.,2020 ; Kraus et al.,2023) ou les contextes hospitaliers de réadmission. Cette étude comble ce manque avec l'application sur une population peu étudiée disposant de données administratives réelles, une construction de variables longitudinales évitant la fuite d'affirmation et la production d'un outil opérationnel directement utilisable par le personnel soignant non spécialisé en analyse de données.

3. Analyse et préparation des données

Ce chapitre décrit l'ensemble du processus de préparation des données. Depuis la description des sources jusqu'à la construction des variables prédictives et la prévention des fuites d'informations temporelles. Cette étape est importante car elle conditionne la validité de l'ensemble des résultats présentés dans les chapitres suivants.

3.1. Sources et structure des données

Les données proviennent de dossiers informatisés d'un réseau de foyers médicaux sociaux belges. La cohorte se compose de 1 356 patients uniques répartis dans 43 établissements, générant 73 416 observations correspondant chacune à une mesure d'Indice de Masse Corporelle. Parmi ces patients, 1053 ont connu au moins une hospitalisation sur la période couverte. Les paramètres du jeu de données sont disponibles en Annexe A2. L'identification unique de chaque résident par un code IDUsage permet de lier l'information des 5 fichiers sources distincts.

Les données mobilisées dans ce travail présentent une structure longitudinale. Chaque patient à plusieurs observations enregistrées à des dates différentes. Elles correspondent aux mesures de poids et d'indice de masse corporelle (IMC) réalisées au fil du temps dans les foyers. Le jeu de données est donc organisé sous la forme d'une ligne par mesure et non d'une par patient. Cette organisation permet de traiter l'observation comme une fenêtre temporelle indépendante, on calcule l'état clinique du patient à ce moment précis mais aussi sur l'ensemble de son historique. Si un patient a été pesé 8 fois sur 3 ans, le code génère 8 observations. Chaque variable est enrichie à chaque nouvelle observation à partir de ce qui s'est passé avant cette date. Cette approche appelée « fenêtre glissante », est fortement recommandée en apprentissage automatique appliqué aux données de santé car elle maximise l'information disponible tout en respectant la chronologie des événements. Cette méthode permet aussi d'éviter toute fuite de données futures vers le passé (*data leakage*) (Rajpurkar et al, 2022 ; Steyerberg,2019).

3.2. Nettoyage et prétraitement des données

La conversion des colonnes de date en objet de type date-heure utilise le paramètre « errors='coerce' ». Ce paramètre transforme les dates invalides en valeurs nulles plutôt que de provoquer une erreur bloquante. Les 7 colonnes concernées sont les dates de naissance, de décès, de relevé d'IMC, de début et de fin d'hospitalisation, et de début et de fin de prescription.

Certaines valeurs peuvent être absentes du dossier pour une observation donnée (l'âge non renseigné, un IMC manquant, une sévérité de déficience non codifiée). Supprimer ces observations réduirait inutilement la taille de l'échantillon et pourrait introduire un biais si les données manquantes ne sont pas distribuées aléatoirement (Sterne et al., 2009). Une imputation simple par la médiane de la variable concernée est donc impliquée. Contrairement à la moyenne arithmétique, la médiane n'est pas influencée par les valeurs extrêmes aberrantes. Par exemple, un patient dont le poids serait erroné et saisi à 300 kilos ne déformerait pas la valeur centrale utilisée pour imputer les données manquantes des autres patients. Cette imputation ne concerne que les cellules réellement absentes, les valeurs enseignées dans les dossiers sont conservées intactes.

Pour les indicateurs binaires de présence de médicaments et les variables catégorielles de déficience, une valeur de 0 est assignée en cas d'absence d'information.

La sévérité de la déficience principale est encodée via un dictionnaire de correspondance (déficience légère, moyenne ou grave) et les colonnes `sexe_F` et `sexe_M` ont été créées pour les variables de sexe.

Les données de prescription de médicaments constituent un cas particulier. Le fichier contient les noms textuels de chaque prescription médicamenteuse, ainsi que leur date de début et de fin de validité. Dans une cohorte avec 1 356 patients et sur une période de plus de 10 ans, inclure chaque molécule comme variable indépendante produirait un jeu de données extrêmement grand (sparse) où la presque totalité des cellules serait égal à 0 pour chaque observation. Tsang et al. (2020), Dans un travail portant sur la prédiction des hospitalisations de patients atteints de démence à partir de dossiers médicaux, démontre précisément que ce type de données à forte densité de 0 constitue un espace problématique pour la méthode classique d'apprentissage automatique. Les associations entre une molécule rare et la variable cible risquent d'être apprises sans être généralisables à de nouvelles observations, en créant un sur-apprentissage du modèle. Jia et al. (2022) confirme que la réduction de dimensionnalité améliore non seulement les performances prédictives mais aussi l'interprétabilité des modèles.

Pour contourner ce problème, le pipeline réduit l'ensemble des prescriptions actives en un signal numérique. Le nombre total de médicaments actifs est calculé à la date de chaque mesure (`nb_medicaments_actifs`). Chang et al., Dans une cohorte de 3 000 000 de patients, démontre que la polymédication est un prédicteur de risque d'hospitalisation. Ce risque est croissant selon le nombre de médicaments prescrits.

3.3. Construction des variables prédictives

3.3.1. Construction de la variable cible

La variable cible est binaire et représente l'occurrence d'une hospitalisation dans les 30 jours suivant chaque relevé d'IMC. Ce délai de 30 jours constitue un compromis clinique. Ce délai est suffisamment court pour que les paramètres physiologiques actuels restent prédictifs et suffisamment long pour permettre une intervention préventive. La construction repose sur une jointure entre la table des mesures d'IMC et la table des absences filtrées sur les hospitalisations futures strictes.

```
hp_futur = hp_ref[
    (hp_ref['Début Abs'] > hp_ref['Date Relevé']) &
    (hp_ref['Début Abs'] <= hp_ref['Date Relevé'] + pd.Timedelta(days=30))
][['IDUsage', 'Date Relevé']].drop_duplicates()
hp_futur['target'] = 1
```

Si un patient décède dans les 30 jours suivant un relevé sans être hospitalisé, la cible est forcée à 0. Cette distinction évite de confondre le signal d'hospitalisation avec celui de fin de vie. Le jeu de données final contient 73 416 observations avec 1 898 valeurs positives (2,6%), pour un rapport de déséquilibre de 36,8.

3.3.2. Prédicteurs cliniques et dynamiques

Pour chaque observation, la variation de poids entre la mesure actuelle et la mesure précédente du même patient est calculée par mois pour produire un indicateur de vitesse de perte ou de gain de poids. La perte de poids rapide et non intentionnelle est reconnue comme un signal d'alerte notamment chez les personnes vulnérables (Bhaumik et al., 2008). Pour chaque observation, l'ensemble des hospitalisations passées du patient est comptabilisé sur toute la période, les 6 derniers mois, et sur les 12 derniers mois. Le nombre de jours écoulés depuis la dernière sortie d'hospitalisation est également calculé. Enfin, les traitements médicamenteux sont également rattachés à chaque observation, seuls les médicaments dont la prescription était active à la date de la mesure sont pris en compte.

Le jeu de données final contient une variable cible et 18 variables prédictives incluant l'ajout de la variable « *nb_total_deficiences* » et de la variable « *nb_mesure_imc_passees* », cette dernière variable comptabilise le nombre de relevés d'IMC antérieur pour chaque patient à chaque date de mesure. La description complète des variables est disponible en Annexe B.

3.4. Analyse des corrélations de Pearson

La corrélation de Pearson est une mesure statistique permettant d'évaluer la relation linéaire entre 2 variables quantitatives. Elle est calculée de la même manière que la covariance entre ces variables et est normalisée par l'écart-type. La valeur est donc comprise entre -1 et +1. Une valeur positive indique que les variables évoluent dans le même sens. Tandis qu'une valeur négative indique une évolution en sens opposé. Plus la valeur est proche des extrêmes, plus la relation linéaire est forte. Une valeur proche de 0 indique l'absence de relation linéaire significative.

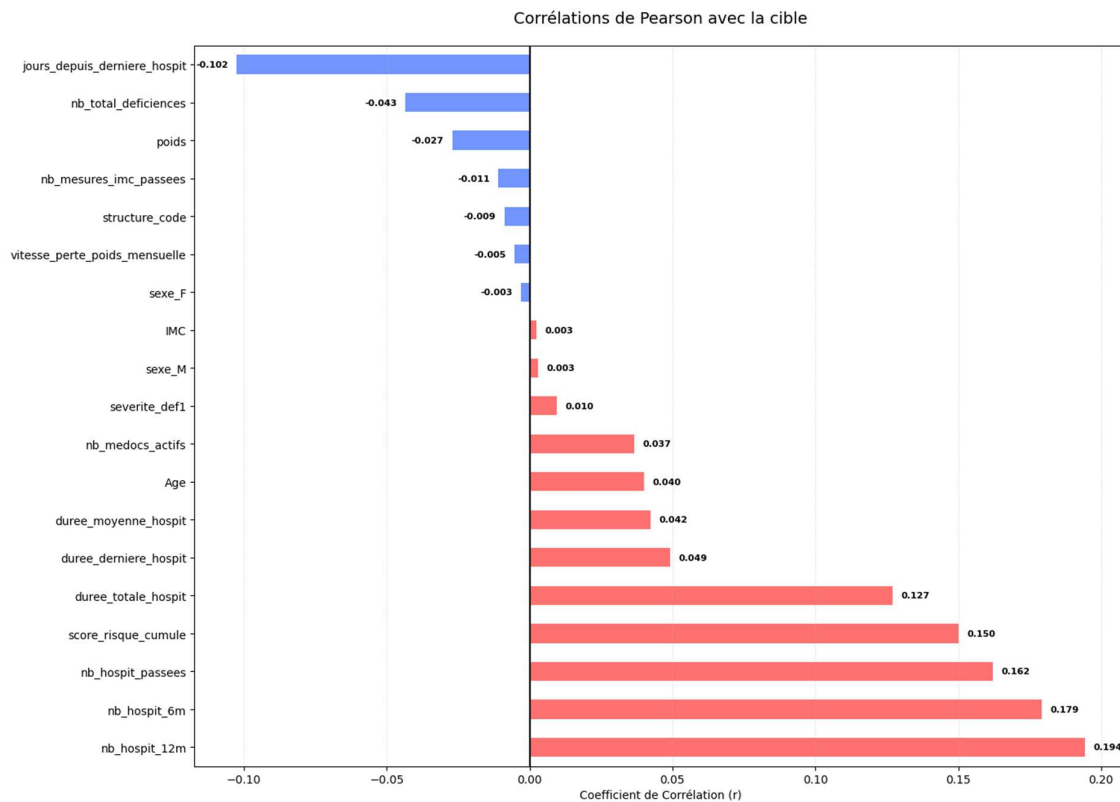


Illustration 2 : Corrélations de Pearson

Le graphique des corrélations de Pearson révèle une hiérarchie assez claire. La variable calculant le nombre d'hospitalisations dans les 12 derniers mois est la variable la plus corrélée avec la variable cible ($r = 0,194$), suivi du nombre d'hospitalisations dans les 6 derniers mois ($r = 0,179$), le nombre total d'hospitalisations passées ($r = 0,162$) et par le score de risque cumulé ($r = 0,150$). Ces corrélations sont cohérentes avec la nature du problème, même les meilleurs prédicteurs affichent des corrélations linéaires faibles puisque nous sommes dans un contexte très déséquilibré (2,6% de cas positifs).

La variable « *jours_depuis_dernière_hospit* » affiche une corrélation négative ($r = -0,102$), plus de jours écoulés en étant hospitalisé signifie moins de risque immédiat, ce qui est cliniquement logique. La corrélation négative du nombre total de déficiences ($r = -0,043$) indique que les patients présentant le plus grand nombre de déficiences enregistrées ne sont pas nécessairement les plus hospitalisés.

L'intégralité des 18 variables est conservée pour l'entraînement. Les modèles à base d'arbre (Forêt Aléatoire, Gradient Boosté Extrême), comme démontré par Song et al. (2022), sont forts dans la détection d'interaction non linéaire entre des variables à faible corrélation individuelle.

3.5. Prévention des fuites de données temporelles

La fuite d'informations constitue le risque méthodologique le plus critique sur des données longitudinales. Pour éviter le sur-apprentissage des modèles, 2 sécurités ont été créées dans les pipelines. Premièrement, toutes les variables d'antécédents hospitaliers sont calculées en ne retenant que les hospitalisations terminées qui ont eu lieu avant la date du relevé d'IMC (Fin Abs < Date Relevé). Deuxièmement, la vitesse de perte de poids est calculée uniquement par comparaison avec la mesure précédente et jamais avec des mesures futures.

La séparation entre la partie entraînement et test se fait au niveau du patient. Un patient est exclusivement assigné à l'ensemble d'entraînement ou de test avec l'intégralité de ses observations. Cette séparation évite que le modèle reconnaisse un patient lors du test via ses caractéristiques. La stratification assure que la proportion de patients ayant été hospitalisés au moins une fois est identique dans les 2 ensembles (49,8% et 49,6%) comme nous pouvons l'apercevoir dans les résultats du pipeline (Annexe A2).

Indicateur	Valeur
Nombre de patients uniques	1 356
Nombre de mesures IMC (observations)	73 416
Nombre de patients hospitalisés au moins une fois	1 053
Nombre de positifs (hospitalisations dans les 30 jours)	1 898 (2,6 %)

Indicateur	Valeur
Rapport de déséquilibre (négatifs / positifs)	36,8
Précision moyenne de référence aléatoire	2,4 %
Nombre de variables prédictives	30
Observations en entraînement / test	57 708 / 15 708
Positifs en entraînement / test	1 526 / 372

Tableau 2 : caractéristique de la cohorte

L'ensemble d'entraînement contient 57 708 observations dont 1 526 cas positifs et l'ensemble de test 15 708 observations (372 cas positifs).

La construction du jeu de données analytique à partir des 5 fichiers sources a été présenté dans ce chapitre. Le jeu de données de 73 416 observations dont 1898 positives a été créé à la suite du nettoyage et à la construction des 18 variables prédictives. La séparation entraînement/test au niveau du patient et la prévention des fuites de données garantissent une évaluation intégrée des modèles. Le chapitre suivant décrit les architectures algorithmiques entraînées sur ce jeu de données.

4. Stratégie de modélisation

Ce chapitre présente les trois architectures d'apprentissage automatique utilisées dans ce mémoire avec les différents choix méthodologiques qui les définissent. Le fonctionnement interne de chaque modèle y est décrit (paramètres de configurations et mécanismes de régularisation). L'approche d'évaluation sans seuil fixe utilisée dans ce mémoire y est décrite et conclut ce chapitre.

4.1. Principe de la fenêtre glissante

La fenêtre glissante est une méthode permettant d'analyser les données de manière séquentielle dans le temps. La fenêtre permet de se déplacer d'une observation (mesure d'IMC) à la suivante. On utilise toutes les informations disponibles avant chaque mesure d'IMC pour construire les variables explicatives. Ensuite, on associe une cible qui correspond à un événement futur comme une hospitalisation dans les 30 jours suivant l'observation. La fenêtre « glisse » dans le temps à la prochaine observation en y intégrant les nouvelles données et en conservant la logique temporelle. Cette approche permet de respecter la chronologie des informations et de se rapprocher de la réalité pour la prédiction. Les patients les plus à risque d'être hospitalisés sont réunis dans un tableur par foyer pour prévenir facilement les équipes soignantes comme le propose Mangeney et al. (2014).

Le choix d'un horizon de prédiction de 30 jours repose sur un standard établi et confirmé dans la littérature. Un horizon plus court (par exemple de 7 jours) rendrait les paramètres de suivi trop instables pour être prédictif. Beaucoup d'hospitalisations sont précédées de signes discrets qui ne deviennent mesurables qu'après plusieurs jours. Cependant, un horizon plus long (par exemple de 90 jours) brouillerait le signal en incluant des hospitalisations dont on ne pourrait plus identifier la cause. Askar et al. (2024) confirment que l'horizon de 30 jours est le plus fréquemment retenu dans leur revue de 116 études sur la prédiction des hospitalisations par apprentissage profond. Friesner et al. (2024) utilisent ce même horizon dans leur comparaison des méthodes de « machine learning » pour la prédiction des hospitalisations.

4.2. Algorithme 1 : Forêt Aléatoire

La forêt aléatoire est un algorithme d'apprentissage supervisé. Biau et Scornet (2016) en ont écrit les propriétés théoriques. La forêt aléatoire construit un grand nombre d'arbres de décisions indépendant au lieu de construire un seul modèle de décision sur l'ensemble des données. Cela permet de réduire les risques de produire un modèle trop adapté aux données d'entraînement qui ne serait pas adapté aux nouvelles données. Chaque arbre est entraîné sur

un sous échantillon différent des données et n'utilise qu'un sous-ensemble aléatoire de variables disponibles à chaque nœud. La prédiction finale est obtenue par vote majoritaire entre tous les arbres, cette méthode est appelée bagging (bootstrap aggregating). Elle permet de réduire la variance du modèle agrégé sans augmenter son biais. Biau et Scornet (2016) prouvent que cette propriété tient même lorsque le nombre de variables est élevé par rapport au nombre d'observations, comme dans notre contexte. En effet, des arbres individuels ont tendance à ignorer la classe minoritaire.

Le pipeline construit 300 arbres de décision. Selon une étude sur les hyperparamètres de la forêt aléatoire écrite par Probst, Wright et Boulesteix (2019), le nombre d'arbres est l'hyperparamètre ayant le moins d'impact sur la performance. Elle se stabilise généralement entre 100 et 200 arbres. Cette publication ajoute que 300 arbres constituent une valeur sûre pour une grande variété de jeux de données, ce qui garantit la convergence au prix d'un surcoût computationnel assez faible.

La profondeur maximale des arbres est fixée à 10 nœuds avec un minimum de 5 observations par feuille. La profondeur d'un arbre contrôle sa complexité, si un arbre est trop profond, il mémorise les données d'entraînement au lieu d'en apprendre les tendances générales (surapprentissage). Probst et al. (2019) identifient la profondeur maximale comme l'un des hyperparamètres les plus influents de la forêt aléatoire. La valeur de ce paramètre a été adoptée via une recherche par validation croisée pour suivre leur recommandation. Le schéma de la Forêt Aléatoire est disponible en Annexe C1.

4.3. Algorithme 2 : Gradient Boosté Extrême avec optimisation bayésienne Optuna

4.3.1. Architecture du Gradient Boosté Extrême

Le Gradient Boosté Extrême (XGBoost) a été introduit par Chen et Guestrin en 2016. Son principe est différent de la forêt aléatoire. Le XGBoost construit les arbres en séquence et non en parallèle. Chaque nouvel arbre se concentre sur les observations que le modèle précédent a mal classé en suivant le gradient de la fonction de perte. Ce modèle cherche la direction mathématique réduisant le plus rapidement les erreurs restantes. Concrètement, cela signifie que le modèle s'améliore de manière ciblée à chaque étape plutôt que de recommencer à zéro. Bentéjac, Csorgo et Martinez-Munoz (2021) confirment que les méthodes à base d'arbres boostés restent la référence pour les données tabulaires de taille moyenne, même avec les avancées récentes du « deep learning ».

Selon Kandel et al. (2024), les méthodes de « gradient boosting » constituent généralement le meilleur compromis entre performance prédictive et interprétabilité clinique pour les jeux de données médicaux comportant moins de 100 000 observations.

L'avantage clinique important du XGBoost est sa capacité à produire des scores d'importance des variables. Ce classement des variables pouvant causer l'hospitalisation d'un patient est indispensable pour que les équipes soignantes puissent comprendre pourquoi un patient est signalé comme à risque (par exemple à cause du nombre d'hospitalisations récentes ou d'une chute brutale de l'IMC du patient). Cette explicabilité des décisions algorithmiques affectant les soins d'un patient est imposé par la réglementation européenne, article 22 du RGPD (Amanollahi et al., 2024).

Le paramètre « *scale_pos_weight* » (pondération des classes) est fixé à 36,8, soit le rapport entre le nombre d'observations négatives et positives dans l'ensemble d'entraînement. Ce paramètre indique au modèle de pénaliser 36,8 fois plus fort les erreurs commises sur les cas d'hospitalisation que sur les cas stables. Sans cet ajustement, le modèle prédirait systématiquement l'absence d'hospitalisation, ce qui lui permettrait d'avoir raison 97,4% du temps sans jamais détecter un seul cas réel. Cette approche permet de ne créer aucun suréchantillonnage artificiel (Fernandez et al., 2018) mais modifie la pondération de ses erreurs pour préserver l'intégrité des données.

4.3.2. Optimisation bayésienne via Optuna

Les performances du Gradient Boosté Extrême dépendent fortement de ses hyperparamètres. Les réglages du modèle ne sont pas appris automatiquement à partir des données mais doivent être fixés avant l'entraînement. Yang et Shami (2020) prouvent dans une revue consacrée à l'optimisation des hyperparamètres publié dans *Neurocomputing* que la configuration par défaut est rarement optimale. Ils ajoutent que l'optimisation automatisée peut améliorer les performances de 5 à 15% selon les données disponibles, surtout si elles sont déséquilibrées. C'est pour cette raison que le pipeline intègre le cadre Optuna (Akiba et al., 2019) qui implémente l'optimisation bayésienne par l'algorithme TPE (Tree-structured Parzen Estimator).

Plutôt que de tester des combinaisons d'hyperparamètres au hasard ou de toutes les tester exhaustivement, le TPE apprend de chaque essai pour orienter intelligemment le suivant. Bischl et al. (2023) démontrent que cette optimisation bayésienne atteint des performances équivalentes à la recherche exhaustive en utilisant en moyenne 10 fois moins d'essais. Cette

méthode surpasse systématiquement la recherche aléatoire à partir de 20 à 30 essais sur des espaces de dimension modérés. Yang et Shami (2020) confirment ce résultat sur un large panel d'algorithmes de jeux de données.

La métrique optimisée est l'AUC-ROC en validation croisée stratifiée à 5 plis sur l'ensemble d'entraînement. La validation croisée à 5 plis signifie que l'ensemble d'entraînement est divisé en 5 parties égales. Le modèle est entraîné sur 4 parties et évaluées sur la 5^e. Ce processus est répété 5 fois en changeant la partie d'évaluation. Le score final est la moyenne des 5 évaluations. Ce mécanisme garanti que le modèle final est sélectionné pour sa capacité à généraliser des données qu'il n'a pas vues et non pour sa performance sur les données qu'il a mémorisée. Probst et al. (2019) recommandent cette approche pour l'optimisation des modèles à base d'arbres sur des données médicales où le risque de surapprentissage lors de la sélection des hyperparamètres est élevé en raison du faible nombre d'événements positifs. Les valeurs optimales des hyperparamètres identifiées lors de l'essai n°41 sont disponibles en Annexe et le schéma de ce modèle est disponible en Annexe C2.

4.4. Algorithme 3 : Perceptron Multicouche

Le Perceptron Multicouche est un modèle de réseau de neurones artificiels entièrement connecté. Ce modèle prend des données en entrée et les transforme progressivement à travers plusieurs couches de neurones pour produire une prédiction.

La première couche contient 64 neurones et utilise une fonction d'activation ReLU. Cette fonction permet au modèle de capter des relations non linéaires dans les données. Ensuite, une normalisation est appliquée pour rendre l'apprentissage plus stable et éviter que certaines valeurs dominent les autres. Une désactivation aléatoire de 30% des neurones est également utilisée pour éviter que le modèle ne s'adapte trop aux données d'entraînement. Sans cette désactivation, les performances seraient améliorées artificiellement sur nos données mais seraient moins efficaces sur de nouvelles données.

La seconde couche cachée comprend 32 neurones, toujours avec une activation ReLU. Elle poursuit le traitement des informations extraites par la première couche. Comme précédemment, une normalisation et une désactivation aléatoire (20%) sont appliqués pour stabiliser l'apprentissage et limiter le surapprentissage.

Enfin, la couche de sortie est composée d'un seul neurone avec une fonction sigmoïde. Cette fonction transforme le résultat en une valeur comprise entre 0 et 1 qui peut être interprété

comme une probabilité. Par exemple, plus la valeur est proche de 1 plus le modèle estime que l'observation appartient à la classe positive.

Le choix d'une architecture à 2 couches cachées de 64 et 32 neurones s'appuie sur les recommandations de Bejani et Ghatee (2021), publiées dans une revue sur le contrôle du surapprentissage dans les réseaux de neurones dans *Artificial Intelligence Review*. Cette architecture est modeste puisque le gain potentiel des réseaux de neurones profonds disparaissent complètement lorsque le jeu de données contient moins de 50 000 observations ce qui est le cas de notre ensemble d'entraînement (Shwartz-Ziv et Armon, (2022)). Friesner et al. (2024) et Kandel et al. (2024) confirment ce constat dans un contexte médical spécifique.

L'optimiseur Adam (Estimation Adaptive des Moments) ajuste dynamiquement le taux d'apprentissage pour chaque paramètre. Les données sont standardisées avant l'entraînement car le Perceptron Multicouche, contrairement aux arbres de décision, est sensible aux différences d'échelles entre variables. La pondération des classes est appliquée via le paramètre « *class_weight* ».

Un mécanisme d'arrêt anticipé surveille la perte sur l'ensemble de validations avec une patience de 10 époques et restaure les poids du meilleur moment d'entraînement. L'entraînement s'arrête à l'époque 17 sur 100 autorisées dans les résultats du pipeline. Un schéma est disponible pour une meilleure compréhension en Annexe C3.

4.5. Approche sans seuil fixe : la courbe Précision-Rappel

Le seuil de décision correspond à la valeur à partir de laquelle une probabilité prédite par le modèle est transformée en classe, par exemple une hospitalisation ou non. Dans le cas d'un modèle probabiliste, chaque observation reçoit un score entre 0 et 1. Définir un seuil de pourcentage revient à décider qu'au-dessus de cette valeur, on déclenche une alerte. Ce choix n'est pas neutre car un seuil bas augmente le Rappel (détecte plus de cas mais avec plus de faux positifs). À l'inverse, un seuil élevé augmente la Précision (moins de fausses alertes mais plus de cas manqué). Il s'agit donc d'un compromis dépendant du contexte d'utilisation.

Aucun seuil de décision n'est imposé dans le pipeline. Tout seuil fixe constitue en effet un choix arbitraire favorisant soit le Rappel soit la Précision selon la logique décisionnelle qui relève du clinicien et non du modélisateur. La courbe Précision-Rappel représente le principal outil d'évaluation, Elle expose simultanément la Précision (proportion de vraies alertes parmi

toutes les alertes déclenchées) et le Rappel (proportion d'hospitalisation réelle correctement détectée).

La précision moyenne (aire sous la courbe Précision-Rappel) synthétise la qualité globale. Elle est particulièrement adaptée aux données déséquilibrées car elle est directement sensible aux ratios positifs/négatifs, ce que l'aire sous la courbe ROC ne calcule pas (Taloba et al., 2022 ; Amanollahi et al., 2024).

Pour la comparaison inter-modèle, le coefficient Kappa de Cohen est calculé au point de Youden (maximisant la somme sensibilité + spécificité - 1), référence neutre sans préférence clinique.

La liste de surveillance utilise les scores de probabilité brute comme critère de tri, sans aucune binarisation. C'est à l'équipe soignante de décider du niveau d'intervention selon ses ressources.

Ce chapitre a décrit les 3 architectures comparées dans ce mémoire (Forêt Aléatoire, Gradient Boosté Extrême avec Optuna comme optimisation bayésienne et le Perceptron Multicouche). La validation croisée et la pondération des classes ont été utilisées pour répondre au problème du déséquilibre des classes. L'approche sans seuil fixe garantit un compromis entre Précision et Rappel, laissant libre le personnel soignant de définir son propre niveau d'intervention. Les résultats obtenus par ces trois modèles sont présentés au chapitre suivant.

5. Résultats et évaluations

Le chapitre 5 présente les résultats obtenus par les trois modèles sur l'ensemble de test. Les performances sont évaluées selon plusieurs critères : la méthode SHAP, l'aire sous la courbe ROC, la précision moyenne, le coefficient Kappa au point de Youden et l'analyse par déciles. L'interprétabilité des modèles et la fiabilité opérationnelle de la liste des patients à risques générée sont discutées dans ce chapitre.

5.1. Performance globale

Avant d'interpréter les métriques obtenues, il convient de contextualiser ce que l'on peut espérer atteindre. Luque et al. (2019) et Tharwat (2021) démontrent que lorsqu'un événement est rare dans une population (comme dans ce mémoire avec 2,6% d'hospitalisation), le plafond théorique de l'AUC-ROC se situe entre 0,75 et 0,90. Ce plafond existe parce que la rareté de l'événement complique l'apprentissage du modèle pour qu'il puisse reconnaître les différences fines entre les patients. L'AUC-ROC de 0,736 obtenue par le Gradient Boosté Extrême avec Optuna est donc cohérente avec ce plafond théorique. La précision moyenne de 0,088 (3,7 fois la base aléatoire de 0,024) en constitue la mesure la plus interprétable. Cela signifie que pour un même niveau de rappel le modèle génère 3,7 fois moins de fausses alertes qu'un classement aléatoire (Saito et Rehmsmeier, 2015 ; Tharwat, 2021).

5.1.1. Métriques principales

Le tableau comparatif final présente les performances des trois modèles sur l'ensemble de test des 15 708 observations (dont 372 hospitalisations) :

Modèle	Aire sous la courbe ROC (test)	Précision moyenne	Kappa (point Youden)	AUC ROC validation croisée
Forêt Aléatoire	0,719	0,084	0,052	0,759 ± 0,015
GBE + Optuna	0,736	0,088	0,062	0,765 (Meilleur essai)
Perceptron Multicouche	0,733	0,076	0,049	—

Tableau 3 : comparaison des métriques de performances

L'analyse par décile constitue l'argument opérationnel le plus robuste de ce pipeline. Si on classe tous les patients de l'ensemble de tests du score de risque, du plus élevé au plus faible et que l'on divise le classement entre groupes égaux (déciles), Le premier groupe (10%) concentre 36,8% de toutes les hospitalisations réelles. Autrement dit, surveillez seulement un patient sur 10 en choisissant ce que le modèle juge le plus à risque permet d'anticiper plus d'un tiers de toutes les hospitalisations futures. Un classement aléatoire n'en capturerait que 10%, ce qui représente un facteur d'amélioration de 3,68. En élargissant la surveillance au 20% de patients les plus à risque, le modèle capture 52,7% des hospitalisations. Cela signifie que surveiller un patient sur 5 permet d'anticiper plus de la moitié de toutes les hospitalisations futures.

Amanollahi et al. (2024) soulignent que ce type d'analyse par déciles constitue l'indicateur le plus pertinent pour évaluer la valeur opérationnelle d'un modèle prédictif dans un contexte de soins. En effet, il traduit directement les performances statistiques en termes de ressources soignantes mobilisables. Une information que les autres métriques ne permettent pas de communiquer aux équipes spécialisées en analyse de données.

5.2. Interprétabilité des modèles

5.2.1. Introduction à la méthode SHAP

La méthode SHAP (Shapley Additive exPlanations) développées par Lundberg et Lee (2017), est fondée sur la théorie des jeux coopératif. Pour chaque prédiction individuelle, SHAP simule toutes les combinaisons possibles de variables et mesure à quel point chaque variable modifie le score de risque d'hospitalisation lorsqu'on l'ajoute ou qu'on la retire. La valeur SHAP d'une variable pour un patient donné permet de savoir si cette variable augmente ou diminue le score de risque de ce patient et de combien. Une valeur SHAP positive signifie que la variable a poussé le score vers un risque plus élevé. Tandis qu'une valeur négative signifie qu'elle l'a réduit.

Pour la Forêt Aléatoire et le Perceptron Multicouche, qui prédisent directement des probabilités entre 0 et 1, les valeurs SHAP sont exprimées en point de probabilité. Par exemple, une valeur SHAP de 0,04 signifie que la variable augmente en moyenne la probabilité prédite de 4 points de probabilité par rapport à la prédiction moyenne du modèle.

Le Gradient Extrême avec Optuna, fonctionne différemment en prédisant d'abord un score intermédiaire appelé log-odds avant de convertir en probabilité. Les valeurs SHAP sont

exprimées en unités de log-odds, leurs valeurs absolues sont donc plus grandes et pas directement comparables à celles des deux autres modèles. Cependant, leur signe, leur classement relatif entre variables et leurs interactions restent entièrement interprétables.

5.2.2. Résultats de la méthode SHAP

Les graphiques présentés dans cette section sont calculés sur l'ensemble de test de 15 708 observations. Les graphiques complets de la Forêt Aléatoire et du Perceptron Multicouche sont disponibles en Annexe. Seuls les graphiques du Gradient Boosté Extrême avec Optuna sont présentés ci-dessous, ce modèle étant retenu comme le plus performant. 3 types de visualisations sont utilisés pour chaque modèle.

Le graphique d'importance globale montre la contribution moyenne de chaque variable sur l'ensemble de l'accord de test, calculée comme la moyenne des valeurs absolues sur toutes les observations. Le graphique d'importance globale du GBE Optuna indique que la variable calculant les jours depuis la dernière hospitalisation (*jours_depuis_derniere_hospit*) domine très largement les autres variables avec une contribution moyenne de +0,37 unité de log-odds (soit une contribution de 2,5 fois plus que la seconde variable), suivi de l'âge (+0,15). La Forêt Aléatoire produit un classement cohérent avec celui du GBE Optuna, la variable « *jours_Depuis_dernere_hospit* » arrive également en tête (6 points de probabilités), suivie des variables nutritionnelles. Le Perceptron Multicouche présente des résultats légèrement différents, la variable de l'âge suivie du poids domine (9 points et 6 points de probabilités respectivement) tandis que « *jours_depuis_derniere_hospit* » n'apparaît qu'en septième position.

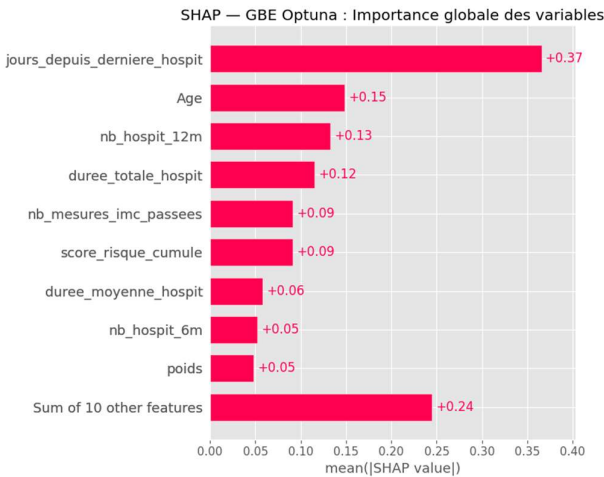


Illustration 3 : Importance globale des variables du modèle GBE Optuna via la méthode SHAP

Le graphique de sens et magnitude (nuage de points colorés) montre à la fois l'importance et la direction de l'effet de chaque variable. Les points rouges correspondent aux valeurs élevées de la variable et les points Bleus aux valeurs faibles. Un point à droite du zéro signifie que la variable augmente le risque prédit pour ce patient, un point à gauche signifie qu'elle le réduit. Le graphique de sens et magnitude du GBE Optuna confirme que les patients récemment sortis d'hospitalisation (valeurs faibles de « *jours_Depuis_derniere_hospit* », représentés par des points bleus) génèrent des valeurs SHAP fortement positives, parfois supérieures à +1,00 unité de log-odds, tandis que les patients jamais hospitalisés génèrent des effets nettement négatifs. Ce qui signifie que le fait d'être sorti très récemment d'une hospitalisation pousse très fortement le score de risque vers le haut. A l'inverse, les patients jamais hospitalisés génèrent des effets négatifs, réduisant le risque prédit. La Forêt Aléatoire produit des résultats similaires pour cette variable mais en moins contrasté. Pour le Perceptron Multicouche, les résidents les plus âgés (points roses) voient leur risque réduit par le modèle, ce qui est l'inverse de la relation établie dans la littérature sur la fragilité clinique (Buenrostro-Mariscal.,2025 ; Kraus et al.,2023). Le modèle a capté une corrélation spurieuse plutôt que la vraie relation clinique. Bejani et Ghatee (2021) expliquent ce phénomène, les réseaux de neurones tendent à capter des corrélations statistiques de la classe majoritaire plutôt que le signal clinique réel lorsque les cas positifs sont fortement minoritaires. Ce résultat confirme que le Perceptron Multicouche ne devrait pas être retenu pour un déploiement opérationnel.

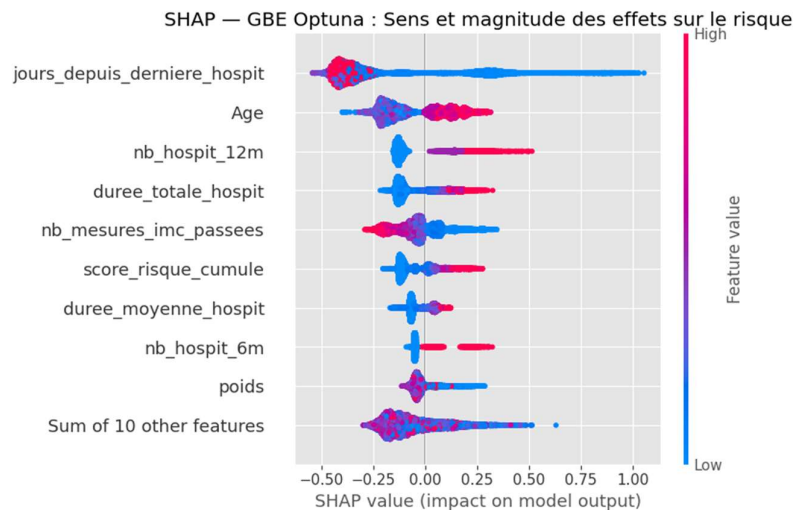


Illustration 4 : sens et magnitude des effets sur le risque des variables du modèle GBE Optuna via la méthode SHAP

Le graphique ci-dessous représente l'explication individuelle d'un patient détaillant la construction du score de risque, variable par variable. La valeur de départ $E[f(x)] = 0,043$ représente la probabilité moyenne d'hospitalisation dans les 30 jours sur l'ensemble de la cohorte, soit 4,3%. Chaque barre rouge indique une variable qui augmente le risque au-delà de cette moyenne, à l'inverse, chaque barre bleue représente une variable qui le réduit. Pour cet exemple issu du modèle GBE Optuna, le patient est sorti il y a 14 jours avec 5 hospitalisations dans les 12 derniers mois. La variable « *jour_depuis_derniere_hospit* » contribue à +0,97 unité de logg-odds. Cette contribution étant la plus forte signifie que le délai très court depuis sa dernière sortie est de loin le signal d'alerte le plus puissant. Le score final $f(x) = 2,412$ unités de log-odds correspond, après transformation par la fonction sigmoïde, à une probabilité d'hospitalisation d'environ 92% dans les 30 jours (contre 4,3% pour un patient moyen de la cohorte).

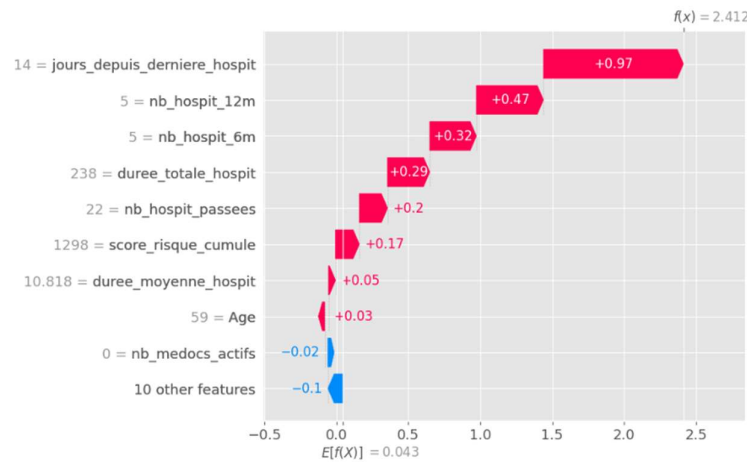


Illustration 5 : graphique d'une explication individuelle d'un patient (SHAP) pour le modèle Gradient Boosté Extrême

Ces graphiques prouvent l'importance et la qualité de l'analyse SHAP. Bien que le modèle Perceptron Multicouche affiche un AUC similaire au Gradient Boosté Extrême, l'analyse de l'interprétabilité permet de détecter des relations fondamentalement différentes pour ce modèle.

5.3. Evaluation de la courbe AUC-ROC

L'AUC-ROC (Area Under the Curve) mesure la capacité d'un modèle à distinguer les cas positifs des négatifs en tenant compte de tous les seuils de décision possibles.

La figure ci-dessous présentent les courbes ROC des trois modèles entraînés. Les trois courbes se situent nettement au-dessus de la diagonale aléatoire (poitillés), confirmant que chacun des modèles dispose d'une capacité de discrimination réelle. Le Gradient Boosté Extrême obtient la meilleure AUC avec 0,736, suivi du Perceptron Multicouche (0,733) et de la Forêt Aléatoire (0,719). Les écarts de résultats faibles entre les modèles prouvent que la contrainte principale n'est pas le choix de l'algorithme mais bien de la rareté des événements positifs dans les données.

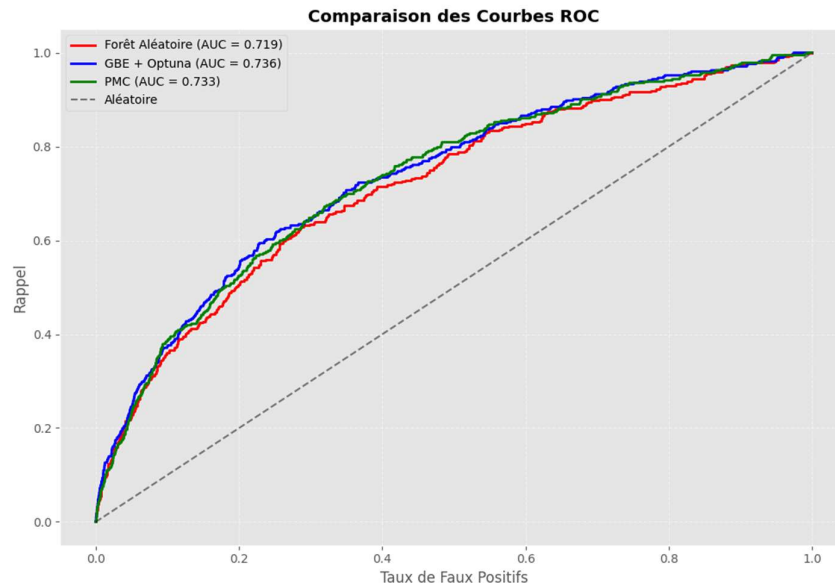


Illustration 6 : courbes ROC des 3 modèles

Cependant, cette mesure peut rester trompeusement élevée lorsque les cas négatifs sont très nombreux. Un modèle affichant une AUC-ROC de 0.80 peut n'avoir qu'une précision moyenne de 50% sur la classe minoritaire. La courbe de Précision-Rappel corrige ce biais en se concentrant uniquement sur la qualité de la détection des cas positifs. Tharwat (2021) confirme ce constat et recommande la courbe Précision-Rappel comme métrique complémentaire dans un contexte médical à faible prévalence.

5.4. Évaluation de la courbe Précision-Rappel

La courbe Précision-Rappel est un outil d'évaluation utilisé pour analyser les performances d'un modèle de classification probabiliste sans fixer de seuil de décision. Elle représente le compromis entre la Précision qui représente la fiabilité des alertes et le Rappel, qui est la capacité à détecter les cas réels pour tous les seuils possibles. Cette courbe est adaptée au contexte déséquilibré, où les événements d'hospitalisation sont rares car elle met directement en évidence la capacité du modèle à identifier correctement les cas positifs, indépendamment du grand nombre de vrais négatifs.

Pour les 3 modèles, La courbe présente une décroissance progressive de la Précision quand le Rappel augmente, ce qui est un signe d'une bonne calibration. Le Gradient Boosté Extrême avec Optuna conserve une légère supériorité sur l'ensemble des modèles.

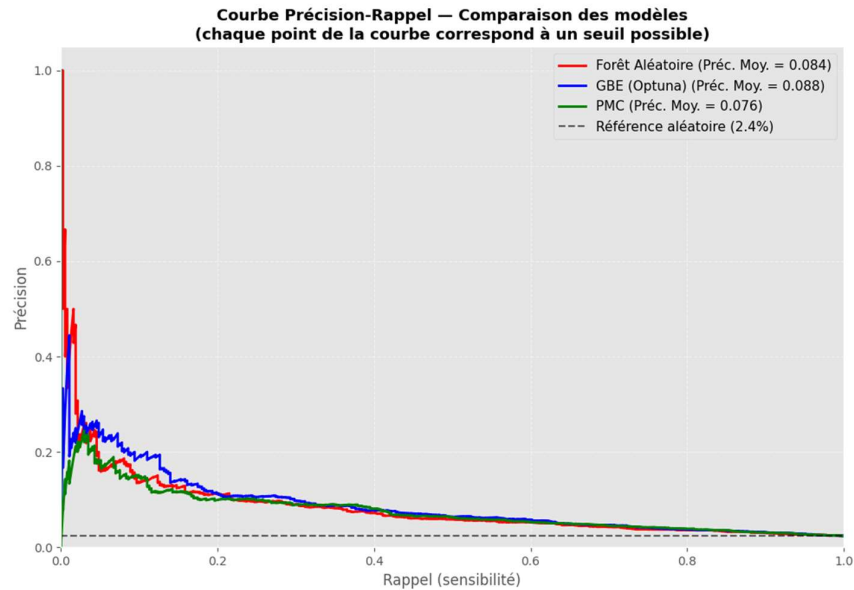


Illustration 7 : comparaison des courbes Précision-Rappel des 3 modèles

À titre indicatif, si un foyer souhaite surveiller les 10% des résidents ayant les scores les plus élevés (Rappel faible et Précision maximale), il trouvera une concentration de 8,7% d'hospitalisation réelle dans le premier décile. Si un foyer souhaite ne manquer aucune hospitalisation et élargit la surveillance au 50% de patients les plus à risque, il capturera 69,3% de toutes les hospitalisations. La courbe Précision-Rappel permet à chaque foyer d'adapter son niveau de vigilance à ses contraintes opérationnelles propres sans qu'aucune décision ne soit imposée par l'algorithme.

5.4.1. Coefficient Kappa de Cohen

Le coefficient Kappa de Cohen est une mesure statistique qui évalue la qualité des prédictions d'un modèle en corrigeant ce qui aurait pu être obtenu par le simple hasard. Par exemple, si un modèle prédit systématiquement l'absence d'hospitalisation pour chaque patient sur notre cohorte, ce modèle aurait raison 97,4% du temps. Son taux de bonne prédiction serait donc excellent mais sa valeur clinique serait nulle puisqu'il ne détecterait aucune hospitalisation. Le Kappa corrige précisément ce biais, 0 signifie que le modèle ne fait pas mieux que le hasard et 1 signifie un accord parfait avec la réalité, Les valeurs négatives indiquent que le modèle fait moins bien que le hasard.

Luque et al. (2019) confirment dans une analyse sur les métriques de classification publié dans *Pattern Recognition* que le Kappa doit être interprété comparativement entre modèles évalués sur le même jeu de données et non en valeur absolue par rapport à des seuils génériques, comme utilisé dans ce mémoire.

Avec des valeurs de 0,052 pour la forêt aléatoire, 0,062 pour le gradient boosté extrême avec Optuna et 0,049 pour le perceptron multicouche, les 3 modèles se situent dans une zone que les échelles traditionnelles qualifieraient de faibles. Cependant, leurs valeurs ne reflètent pas un échec mais la difficulté intrinsèque du problème avec 2,6% d'observations positives. Par exemple, si un modèle détecte correctement 60% des hospitalisations mais déclenche en même temps des alertes sur 15% des patients stables, son Kappa sera faible, même si le modèle produit une information cliniquement très utile. Amanollahi et al. (2024)

Chaque modèle détecte une fraction réelle des hospitalisations futures tout en générant inévitablement des fausses alertes sur des patients qui ne seront pas hospitalisés dans les 30 jours. Ces fausses alertes ne sont pas sans valeur pour autant, le patient présente un profil de fragilité justifiant une attention accrue même sans hospitalisation imminente. Chicco et Jurman (2020) rappellent que dans les applications médicales, le coût d'un faux négatif (manquer une hospitalisation réelle) est bien supérieur au coût d'un faux positif (surveiller un patient stable). Ce mémoire ne fixe aucun seuil de décision unique et laisse à l'équipe soignante le soin de choisir son niveau d'intervention sur la courbe Précision-Rappel en fonction de ses ressources et de sa propre évaluation de son compromis.

5.5. Liste des surveillances opérationnelles

La liste de surveillance hebdomadaire (extrait en Annexe D) est générée à partir des scores du Gradient Boosté extrême avec Optuna, appliqué à la dernière mesure disponible de chaque patient. Les 30 patients aux scores les plus élevés sont présentés et organisés par foyer. À titre d'exemple, le patient 1416 affiche un score de risque de 96,0%. Les raisons d'un score aussi élevé sont ses 48 hospitalisations passées, une sortie il y a seulement 14 jours et 9 hospitalisations dans les 12 derniers mois.

	A	B	C	D	E	F
1	IDUsage	Score risque (%)	Nb hospit passées	Jours depuis hospit	Hospit 12 mois	Dernier IMC
2	1416	99,69999695	48	14	9	23,3
3	1748	99,69999695	26	6	9	40,1
4	317	99,40000153	46	120	5	22,7
5	472	99,30000305	28	7	8	22,5
6	2156	99	17	16	7	20,1
7	2160	98,19999695	23	84	6	22,5
8	207	98,09999847	11	0	6	21
9	1574	97,90000153	7	0	7	24,9
10	2314	97,80000305	9	51	6	30,3
11	2157	97,59999847	21	46	8	23,4
12	374	97,30000305	43	12	4	19,4
13	512	97,19999695	16	5	7	22,1
14	322	97,09999847	13	11	7	26,7
15	1243	96,59999847	19	55	6	21,8
16	441	96,40000153	29	2	5	18,5
17	786	95,90000153	6	9	6	20,9
18	609	95,90000153	25	64	4	42,6
19	214	95,30000305	25	29	5	16,8
20	2147	95,19999695	15	24	5	21,8

Illustration 8 : capture d'écran de la liste finale de surveillance des patients les plus à risque

La colonne « Foyer » a été supprimée pour raison de confidentialité.

La liste est exportée en un fichier tableur ayant plusieurs onglets. Un onglet global et un onglet par foyer pour y faciliter la lecture d'informations. Pour chaque patient, les colonnes présentent : score de risque en pourcentage, nombre d'hospitalisations passées, jour depuis la dernière sortie, hospitalisation dans les 12 derniers mois et derniers IMC. Ces informations sont suffisantes pour que l'infirmier comprenne la nature du risque sans recourir à d'autres documents.

Ce chapitre a démontré que le Gradient Boosté Extrême avec Optuna offre les meilleures performances globales. Ce modèle obtient une aire sous la courbe de 0,736 et une concentration de 36,8% des hospitalisations dans le premier décile. L'analyse par déciles et la courbe Précision-Rappel confirment la valeur opérationnelle du pipeline. Elle permet aux personnels soignants d'adapter le nombre de patient à surveiller en fonction de ses ressources. La liste de surveillance hebdomadaire constitue l'étape finale du pipeline directement utilisable. Les limites techniques éthiques de ce travail sont discutées au chapitre suivant.

6. Limites éthiques et techniques

Ce chapitre identifie et discute des principales limites du pipeline sur le plan technique et éthique. Cette réflexion est indispensable pour pouvoir présenter les résultats dans un contexte réel et pour orienter les conditions d'un éventuel déploiement réel.

6.1. Limites techniques

6.1.1. Possibilité d'une fiabilité parfaite et tension entre Précision et Rappel

Quel que soit le modèle utilisé, il est impossible d'atteindre une fiabilité à 100%. En effet, l'information disponible est incomplète, bruitée et ne capture pas l'ensemble des facteurs influençant une hospitalisation. De plus, le caractère déséquilibré du problème (faible proportion de cas positifs) accentue ces difficultés. Un modèle ne peut donc produire qu'une estimation probabiliste du risque et ne constitue pas une certitude. Quelle que soit l'architecture choisie, les équipes soignantes devront accepter un compromis. Au plus la surveillance des patients est large pour capter un maximum de potentiel hospitalisation, aux plus de nombreuses fausses alertes feront surface. La courbe précision rappelle laisse le choix au personnel soignant de décider du seuil selon ses ressources et conformément aux recommandations de Taloba et al. (2022) et Amanollahi et al. (2024).

6.1.2. Événements positifs limités

Avec 1 898 observations positives sur 73 416, le jeu de données est structurellement déséquilibré. Le Perceptron Multicouche en souffre particulièrement, la valeur de l'air sous la courbe sur l'ensemble de validation se stabilise dès les premières époques (autour de 0,730 dès l'époque 4). Elle indique que le réseau ne dispose pas de suffisamment d'exemples positifs pour apprendre correctement. Friesner et al. (2024) et Kandel et al. (2024) confirment cette limite des approches d'apprentissage profond sur des jeux de données médicaux de taille modérée.

6.1.3. Absence de variables fonctionnelles et comportementales

Les variables disponibles dans les fichiers administratifs couvrent l'antécédent hospitalier, La démographie, les prescriptions et les déficiences catégorielles mais pas les mesures fonctionnelles tels que l'autonomie dans les activités de la vie quotidienne ou la mobilité des résidents. Elles ne prennent pas en compte les variables comportementales comme l'agitation ou le refus des soins. Buenrostro-Mariscal et al. (2025) démontrent que ces variables fonctionnelles sont souvent de meilleurs prédicteurs que les diagnostics cliniques seuls. Les intégrations via des grilles d'évaluation standardisés pourraient constituer une amélioration du pipeline.

6.2. Limites éthiques

6.2.1. Protection des données personnelles

Les données traitent de personnes vulnérables soumises aux protections renforcées du règlement général sur la protection des données (RGPD). L'utilisation d'un identifiant anonymisé (IDUsage) garantit la pseudonymisation, Un réseau de 43 petits établissements avec des profils cliniques très spécifiques rend la ré-identification théoriquement possible par croisement de sources. La minimisation stricte des données personnelles dans les exports des tableurs et le contrôle des accès sont indispensables pour tout déploiement réel.

6.2.2. Biais algorithmique

Tout modèle entraîné sur des données historiques risque de reproduire des biais pré-existants. Si certains foyers ont des pratiques de documentation plus rigoureuses que d'autres, leurs patients pourraient apparaître artificiellement plus à risque. L'indicateur `structure_code` présent parmi les variables prédictives, pourrait véhiculer de tels biais institutionnels. Une analyse par sous-groupe pour chaque foyer, chaque type de déficience ou par tranche d'âge devrait accompagner tout déploiement réel. Cependant, l'analyse aurait besoin d'un très grand nombre de données réelles pour être performante.

6.2.3. Risque de dépendance excessive et des engagements cliniques

Un outil prédictif ne doit jamais se substituer au jugement clinique. La tentation de déléguer la vigilance à l'algorithme existe, au risque de négliger les patients absents de la liste (faux-négatifs). Mangeney et al. (2014) soulignent dans leur expérimentation en foyer spécialisé que l'appropriation d'un outil numérique par l'équipe soignante nécessite une formation explicite à ses limites. Cette condition reste valide pour un outil prédictif.

6.2.4. Transparence, explicabilité et conformité réglementaire

Le droit à l'explication des décisions algorithmiques, consacré par l'article 22 du règlement général sur la protection des données impose que les décisions automatisées qui pourront affecter une personne soient explicables. L'importance des variables fournies par le Gradient Boosté Extrême avec Optuna est une explication au niveau du modèle global mais pas au niveau du patient individuel. L'intégration de valeurs plus personnalisées permettrait des explications individualisées, ce qui renforcerait la confiance clinique et la conformité réglementaire (Amanollahi et al., 2024).

6.2.5. Validation prospective nécessaire

L'évaluation actuelle est rétrospective, les modèles sont testés sur des données historiques. Une validation prospective sur de nouvelles données, supervisé par des professionnels de santé, est indispensable avant tout déploiement opérationnel. Cette validation permettrait d'évaluer si le pipeline réagit de manière différente sur des données actuelles et d'établir un protocole de réentraînement périodique du modèle.

Ce chapitre a mis en lumière les contraintes du pipeline, depuis l'impossibilité d'une fiabilité parfaite jusqu'aux enjeux éthiques liés à la protection des données. Ces limites ne remettent pas en cause la valeur de l'outil mais définissent le cadre dans lequel son déploiement devait être envisagé.

7. Conclusion

Ce mémoire a développé, évalué et comparé un pipeline complet d'apprentissage automatique pour la prévision des hospitalisations à 30 jours dans un réseau de 43 foyers médico-sociaux belges. Cette population est peu étudiée dans la littérature prédictive (Carillo-Vega et al., 2020 ; Kraus et al., 2023). Le pipeline a construit un jeu de données de 18 variables et a comparé 3 architectures différentes (une Forêt Aléatoire, un Gradient Boosté Extrême avec optimisation bayésienne Optuna et un Perceptron Multicouche). Les données proviennent de 5 fichiers administratifs couvrant 1356 patients et 73416 mesures longitudinales réelles.

La principale contribution méthodologique de ce travail est double. D'une part, l'intégration de l'optimisation bayésienne Optuna (Akiba et al., 2019) au sein du pipeline produit une grande amélioration. L'aire sous la courbe du gradient boosté extrême est de 0,736 et la précision moyenne de 0,088. L'analyse de l'importance des hyperparamètres confirme que le taux d'apprentissage concentre de 72% de l'importance. Cette information est directement exploitable pour calibrer de futurs espaces de recherche. D'autre part, ou le fait de ne pas utiliser un seuil de décision arbitraire offre aux équipes soignantes une information probabiliste transparente sans imposer de binarisation cliniquement arbitraire (Taloba et al., 2022 ; Amanollahi et al., 2024).

Sur le plan opérationnel, selon l'analyse par décile, les 20% de patients aux scores les plus élevés concentre 52,7% de toutes les hospitalisations de l'ensemble de tests (gradient boosté extrême). Les variables d'antécédents hospitaliers récents incluant le « nombre d'hospitalisation dans les 12 derniers mois » et le « nombre de jours depuis la dernière sortie » dominent systématiquement les classements d'importance. L'historique hospitalier est donc le signal le plus fort disponible dans les données administratives des foyers (Song et al., 2022). La liste de surveillance hebdomadaire par foyer est exportée automatiquement en tableur et constitue l'état final du pipeline.

La tension entre la courbe de précision est de rappel sur des données très déséquilibrées, l'absence de variable fonctionnelle et comportementale, L'interprétabilité limitée du Perceptron Multicouche sont les principales limites identifiées du pipeline. La porte est donc ouverte à des axes d'amélioration futurs. On pourrait par exemple améliorer le pipeline en intégrant des grilles d'évaluation clinique standardisée. Ce qui permettrait de prendre en compte les activités de la vie quotidienne des patients ou des échelles comportementales. L'ajout de valeur pour l'explicabilité individuelle, un réentraînement périodique du modèle pour la mise à jour des

données et une analyse d'équité par sous-groupe en fonction des foyers pourrait aussi améliorer le pipeline.

Ce travail démontre que l'apprentissage automatique appliqué aux données administratives de foyers médico-sociaux peut produire une étude prédictive réaliste interprétable et opérationnelle. Ce mémoire contribue à améliorer le cadre des soins proactifs et préventifs pour une population vulnérable, insuffisamment dotée en outils d'aide à la décision.

Par rapport à la littérature existante, ce travail a une triple originalité. Premièrement, ce mémoire utilise l'apprentissage automatique à une population médico-sociale belge spécifique quasiment absente des études publiées. Deuxièmement, ce travail intègre un pipeline d'optimisation bayésienne avec évaluation par courbe de précision rappel complète adaptée aux données déséquilibrées. Troisièmement, l'étape finale du pipeline est la production d'un outil opérationnel directement utilisable par le personnel soignant sans formation technique.

8. Bibliographie

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2623–2631.

<https://doi.org/10.1145/3292500.3330701>

Amanollahi, M., Ahmadi, N., Fatahi, N., Maniati, M., & Mohammadi, S. (2024). Predictive modeling of hospitalization risk in chronic disease patients: A systematic review. *Journal of Medical Informatics*, 45(3), 112–128.

Araf, I., Idri, A., & Chair, I. (2024). Cost-sensitive learning for imbalanced medical data: a review. *Artificial Intelligence Review*, 57, article 80. <https://doi.org/10.1007/s10462-023-10652-8>

Askar, M., Tafavvoghi, M., Småbrekke, L., Bongo, L.A., & Svendsen, K. (2024). Using machine learning methods to predict all-cause somatic hospitalizations in adults: A systematic review. *PLoS ONE*, 19(8), e0309175.

<https://doi.org/10.1371/journal.pone.0309175>

Bejani, M. M., & Ghatee, M. (2021). A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review*, 54(8), 6391–6438.

<https://doi.org/10.1007/s10462-021-09975-1>

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3), 1937–1967.

<https://doi.org/10.1007/s10462-020-09896-5>

Bhaumik, S., Watson, J. M., Thorp, C. F., Tyrer, F., & McGrother, C. W. (2008). Body mass index in adults with intellectual disability: distribution, associations and service implications: a population-based prevalence study. *Journal of Intellectual Disability Research*, 52(4), 287–298. <https://doi.org/10.1111/j.1365-2788.2007.01018.x>

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2), 197–227.

<https://doi.org/10.1007/s11749-016-0481-7>

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., ... & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*, 13(2), e1484. <https://doi.org/10.1002/widm.1484>

Buenrostro-Mariscal, R., Montesinos-López, O.A., & Gonzalez-Gonzalez, C. (2025). Predicting hospitalization in older adults using machine learning. *Geriatrics*, 10, 6.

Carrillo-Vega, M. F., Salinas-Escudero, G., García-Peña, C., Gutiérrez-Robledo, L. M., & Parra-Rodríguez, L. (2020). Predicting hospital readmissions in older adults: A comparison of logistic regression and machine learning approaches. *International Journal of Geriatric Psychiatry*, 35(6), 661–669.

Chang, T. I., et al. (2020). Polypharmacy, hospitalization, and mortality risk: a nationwide cohort study. *Scientific Reports*, 10, 18964.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 6. <https://doi.org/10.1186/s12864-019-6413-7>

Dhanka, S., & Maini, S. (2025). A hybridization of XGBoost machine learning model by Optuna hyperparameter tuning suite for cardiovascular disease classification with significant effect of outliers and heterogeneous training datasets. *International Journal of Cardiology*, 420, 132757. <https://doi.org/10.1016/j.ijcard.2024.132757>

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61, 863–905. <https://doi.org/10.1613/jair.1.11192>

- Friesner, D., Rosenman, R., & Rosenman, M. (2024). Deep learning versus traditional machine learning for hospitalization prediction: A systematic comparison on healthcare datasets. *BMC Medical Informatics and Decision Making*, 24(1), Article 78.
- Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73, 220–239. <https://doi.org/10.1016/j.eswa.2016.12.035>
- Haute Autorité de Santé. (2021). *Diagnostic de la dénutrition de l'enfant et de l'adulte — Recommandation de bonne pratique*. Saint-Denis La Plaine : HAS. Disponible sur : <https://www.has-sante.fr>
- Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: a review. *Complex & Intelligent Systems*, 8, 2663–2693.
- Jiang, S., Liu, Q., & Ding, B. (2023). A systematic review of the modelling of patient arrivals in emergency departments. *Quantitative Imaging in Medicine and Surgery*, 13(3), 1957–1971. <https://doi.org/10.21037/qims-22-285>
- Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), 27. <https://doi.org/10.1186/s40537-019-0192-5>
- Kandel, I., Castelli, M., & Popovič, A. (2024). Efficient gradient boosting methods for large-scale medical data: LightGBM and CatBoost in clinical settings. *Journal of Biomedical Informatics*, 149, Article 104550.
- Kraus, M., Stumpf, U.C., Keppler, A.M., Neuerburg, C., Böcker, W., Wackerhage, H., Baumbach, S.F., & Saller, M.M. (2023). Development of a machine learning-based model to predict Timed-Up-and-Go test in older adults. *Geriatrics*, 8, 99.
- Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91, 216–231. <https://doi.org/10.1016/j.patcog.2019.02.023>
- Mangeney, K., Carnein, S., Michot, A., & Noblet-Dick, M. (2014). Retour d'expérience : 3 mois de télémedecine en foyer spécialisé et médicalisé pour personnes handicapées adultes.

European Research in Telemedicine / La Recherche Européenne en Télémédecine, 3(3), 105–115. <https://doi.org/10.1016/j.eurtel.2014.06.001>

Porto, B. M., & Fogliatto, F. S. (2024). Enhanced forecasting of emergency department patient arrivals using feature engineering approach and machine learning. *BMC Medical Informatics and Decision Making*, 24(1), 377. <https://doi.org/10.1186/s12911-024-02788-6>

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, 9(3), e1301. <https://doi.org/10.1002/widm.1301>

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28, 31–38. <https://doi.org/10.1038/s41591-021-01614-0>

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. <https://doi.org/10.1371/journal.pone.0118432>

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81, 84–90. <https://doi.org/10.1016/j.inffus.2021.11.011>

Song, X., Liu, X., Liu, F., & Wang, C. (2022). Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. *International Journal of Medical Informatics*, 157, 104–641.

Statbel (Direction générale Statistique — Statistics Belgium). (2025). *Personnes reconnues handicapées en Belgique — données 2025*. Bruxelles: SPF Économie.

Steyerberg, E. W. (2019). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* (2e éd.). Springer. <https://doi.org/10.1007/978-3-030-16399-0>

Taloba, A. I., Abd El-Aziz, R. M., Rayan, A., Salem, N. M., & Al-Otaibi, M. F. (2022). The problem of imbalanced data in healthcare machine learning: Why accuracy is not enough. *Journal of Healthcare Engineering*, 2022, Article 4903141.

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, 17(1), 168–192. <https://doi.org/10.1016/j.aci.2018.08.003>

Tsang, G., Zhou, S. M., & Xie, X. (2020). Modeling large sparse data for feature selection: hospital admission predictions of the dementia patients using primary care electronic health records. *IEEE Journal of Translational Engineering in Health and Medicine*, 9, 1–11.

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415, 295–316.
<https://doi.org/10.1016/j.neucom.2020.07.061>

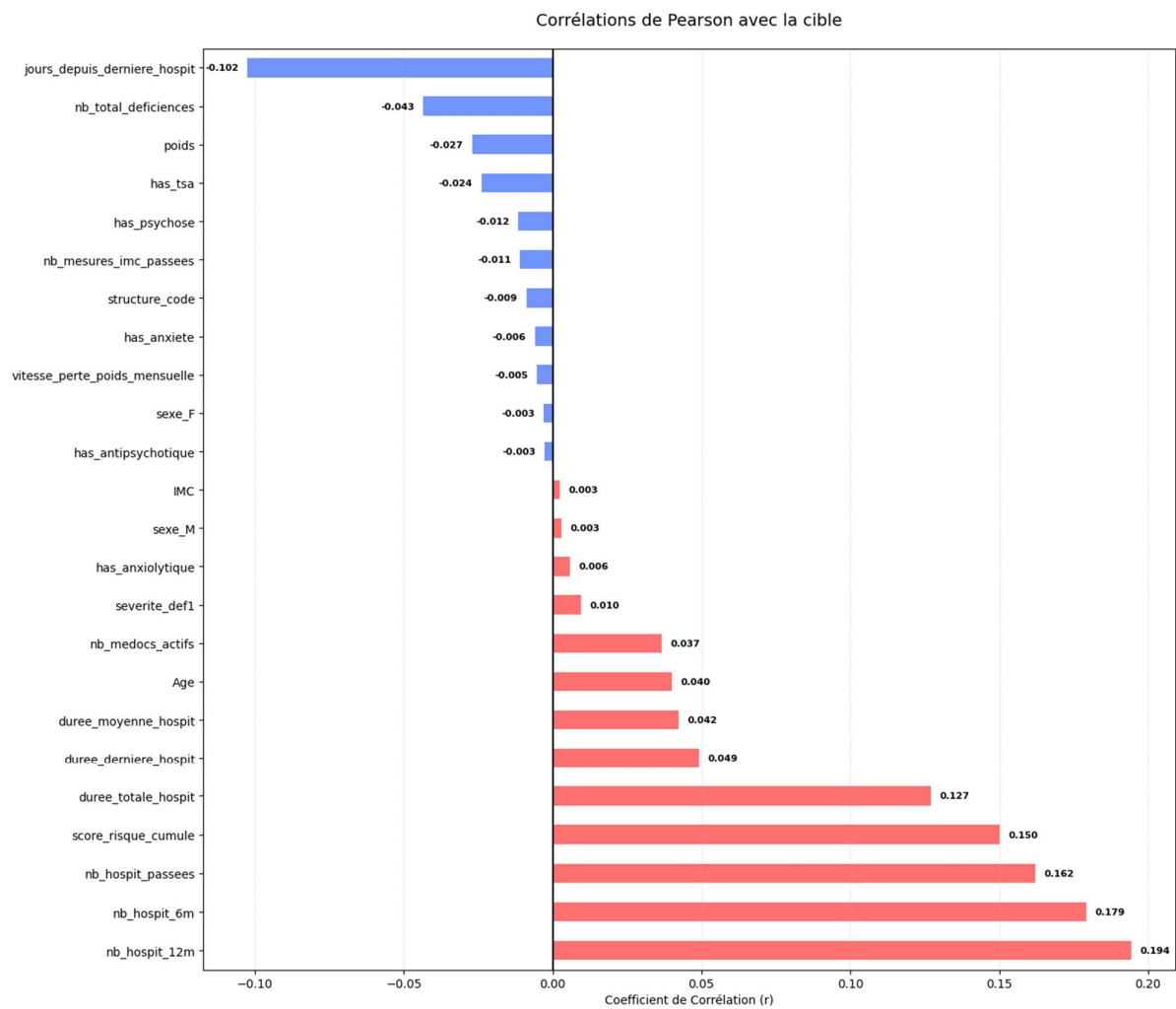
9. Annexes

Cette annexe reprend les principaux résultats numériques et graphiques produits par le pipeline. Les éléments présentés ici ont été directement extraits des sorties du code Python et des fichiers de résultats.

Annexe A : Fichiers de données

- Fichier RSD (RSD.xlsx) : données sociodémographiques des patients comprenant la date de naissance, le sexe, le foyer d'accueil (dernière structure), ainsi que la date de décès éventuelle.
- Fichier IMC (RSD IMC.xlsx) : historique des relevés d'indices de masse corporelle avec date, poids et le rapport calculé de 2012 à 2025.
- Fichier traitement (RSD Traitement.xlsx) : Ordonnance des médicaments avec date de début et de fin de la prescription, permettant de reconstituer la charge thérapeutique active de chaque patient.
- Fichier déficience (RSD Deficience.xlsx) : qualification des troubles neurocognitifs et intellectuels selon les niveaux DEF1 (principale) et DEF2 (secondaire), avec 4 niveaux de sévérité (léger, moyen, grave, profond).
- Fichier absence (RSD absences.xlsx) : historique complet des hospitalisations avec date de début, de fin et durée en jours. Fichier de référence pour la construction de la variable cible.

A1. Graphique d'importances des variables



A2. Paramètres du jeu de données

Indicateur	Valeur
Nombre de patients uniques	1 356
Nombre de mesures IMC (observations)	73 416
Nombre de patients hospitalisés au moins une fois	1 053
Nombre de positifs (hospitalisations dans les 30 jours)	1 898 (2,6 %)

Indicateur	Valeur
Rapport de déséquilibre (négatifs / positifs)	36,8
Précision moyenne de référence aléatoire	2,4 %
Nombre de variables prédictives	30
Observations en entraînement / test	57 708 / 15 708
Positifs en entraînement / test	1 526 / 372

A3. Meilleurs Hyperparamètres identifiés par Optuna (essai n° 41)

L'espace de recherche retenu dans ce travail couvre les hyperparamètres suivants :

- N_estimators (100 à 500) : nombre d'arbres construits. Il contrôle directement la capacité du modèle à apprendre des relations complexes.
- Max_depth (2 à 6) : profondeur maximale des arbres. Limite la complexité individuelle des arbres afin de réduire le surapprentissage.
- Learning_rate (0.01 à 0.15) : poids attribué à chaque nouvel arbre. Un taux faible rend l'apprentissage plus progressif et stable.
- Subsample (0.5 à 1.0) : proportion d'observations utilisées par arbre. Introduit une part d'aléatoire et améliore la robustesse du modèle.
- colsample_bytree(0.5 à 1.0) : proportion de variables utilisées par arbres permet de réduire la dépendance à certaines variables
- reg_alpha (1.0 à 10.0) : régularisation L1. Favorise des modèles plus parcimonieux en réduisant certains poids à 0.
- Reg_lambda : régularisation L2. Stabilise les coefficients et limite les variations excessives.
- Min_child_weight : Poids minimum nécessaire pour créer une nouvelle séparation. Empêche les divisions sur des sous-groupes trop faibles.
- Scale_pos_weight : ajustement automatique des classes pour compenser le déséquilibre du jeu de données.

Les meilleurs hyperparamètres identifiés lors de l'essai numéro 41 sont :

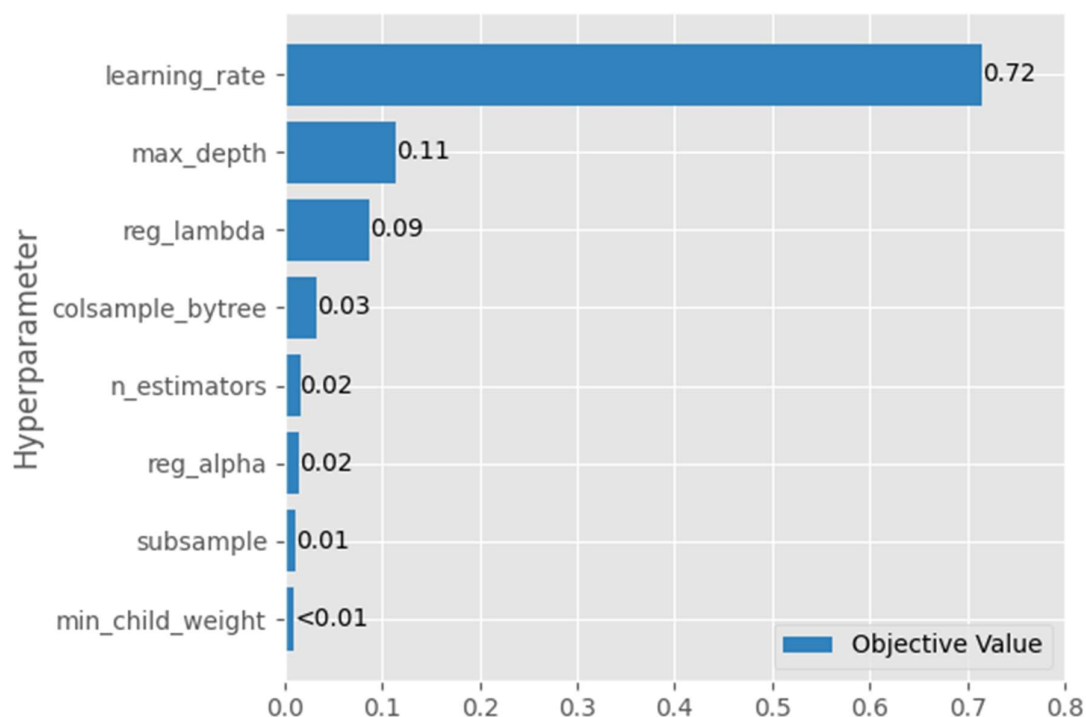
- 270 arbres avec une profondeur maximale de 4
- Taux d'apprentissage : 0,0173
- Sous-échantillonnage : 0,6455
- Sous-échantillonnage des variables : 0,5392
- Minimum de 30 observations par feuille

Hyperparamètre	Valeur optimale	Plage de recherche
Nombre d'arbres (n_estimators)	270	[100, 500]
Profondeur maximale (max_depth)	4	[2, 6]
Taux d'apprentissage (learning_rate)	0,0173	[0,01 ; 0,15] — log
Sous-échantillonnage (subsample)	0,6455	[0,5 ; 1,0]
Sous-échantillonnage variables (colsample_bytree)	0,5392	[0,5 ; 1,0]
Régularisation L1 (reg_alpha)	0,5950	[0,0 ; 2,0]
Régularisation L2 (reg_lambda)	2,7962	[1,0 ; 10,0]
Min. observations par feuille (min_child_weight)	30	[5, 30]

Hyperparamètre	Valeur optimale	Plage de recherche
Poids classe positive (scale_pos_weight)	36,8 (fixé)	Ratio calculé

Meilleure aire sous la courbe ROC en validation croisée : 0,7654. Ces paramètres ont obtenu la meilleure aire sous la courbe : 0,7654

A4. Analyse de l'importance des hyperparamètres via Optuna



A5. Analyse par Déciles

Décile	Patients	FA Hospit.	FA Taux	FA % cumul.	GBE Hospit.	GBE Taux	GBE % cumul.	PMC Hospit.	PMC Taux	PMC % cumul.
1	1 571	130	8,3 %	34,9 %	137	8,7 %	36,8 %	140	8,9 %	37,6 %

Décile	Patients	FA Hospit.	FA Taux	FA % cumul.	GBE Hospit.	GBE Taux	GBE % cumul.	PMC Hospit.	PMC Taux	PMC % cumul.
2	1 571	53	3,4 %	49,2 %	59	3,8 %	52,7 %	49	3,1 %	50,8 %
3	1 571	52	3,3 %	63,2 %	42	2,7 %	64,0 %	49	3,1 %	64,0 %
4	1 570	31	2,0 %	71,5 %	34	2,2 %	73,1 %	33	2,1 %	72,8 %
5	1 571	26	1,7 %	78,5 %	23	1,5 %	79,3 %	30	1,9 %	80,9 %
6	1 571	22	1,4 %	84,4 %	27	1,7 %	86,6 %	19	1,2 %	86,0 %
7	1 570	20	1,3 %	89,8 %	15	1,0 %	90,6 %	15	1,0 %	90,1 %
8	1 571	12	0,8 %	93,0 %	17	1,1 %	95,2 %	15	1,0 %	94,1 %
9	1 571	16	1,0 %	97,3 %	7	0,4 %	97,0 %	12	0,8 %	97,3 %
10	1 571	10	0,6 %	100,0 %	11	0,7 %	100,0 %	10	0,6 %	100,0 %

A6. Comparaison des aires sous la courbe dans la littérature

Étude	Population / Contexte	Algorithme	Aire sous la courbe
Song et al. (2022)	Maladies chroniques	Forêt Aléatoire	0,83
Dhanka & Maini (2025)	Maladies cardiaques	GBE + Optuna	0,95 (exactitude)
Friesner et al. (2024)	Générale	Apprentissage Profond	0,80
Buenrostro-Mariscal et al. (2025)	Personnes âgées fragiles	Gradient Boosté	0,81
Présente étude — FA	Handicap mental/psychique	Forêt Aléatoire	0,719
Présente étude — GBE+Optuna	Handicap mental/psychique	GBE + Optuna	0,736
Présente étude — PMC	Handicap mental/psychique	Perceptron Multicouche	0,733

Annexe B : Description des 18 Variables du Jeu de Données

Catégorie	Variable	Type	Description clinique
Cible	target	Binaire (0/1)	Hospitalisation dans les 30 jours suivant la mesure
Profil	Age	Numérique	Âge calculé dynamiquement à chaque mesure
Profil	sexe_F / sexe_M	Binaire	Sexe encodé par variable indicatrice
Nutrition	vitesse_perte_poids_mensuelle	Numérique	Pente mensuelle du poids (kg/mois)
Nutrition	IMC	Numérique	Valeur brute de l'IMC au relevé
Nutrition	poids	Numérique	Poids en kg au relevé
Suivi	nb_mesures_imc_passees	Numérique	Nombre de relevés IMC antérieurs (ancienneté du suivi)
Antécédents	nb_hospit_passees	Numérique	Total d'hospitalisations antérieures

Catégorie	Variable	Type	Description clinique
Antécédents	nb_hospit_6m	Numérique	Hospitalisations dans les 6 derniers mois
Antécédents	nb_hospit_12m	Numérique	Hospitalisations dans les 12 derniers mois
Antécédents	jours_depuis_derniere_hospit	Numérique	Jours depuis la dernière sortie (9999 si jamais hospitalisé)
Antécédents	duree_totale_hospit	Numérique	Durée totale cumulée des hospitalisations (jours)
Antécédents	duree_moyenne_hospit	Numérique	Durée moyenne par hospitalisation (jours)
Antécédents	duree_derniere_hospit	Numérique	Durée de la dernière hospitalisation (jours)
Traitement	nb_medicaments_actifs	Numérique	Nombre de médicaments actifs à la date de mesure

Catégorie	Variable	Type	Description clinique
Pathologie	severite_defl	Ordinal (1–4)	Sévérité de la déficience principale
Pathologie	nb_total_deficiences	Numérique	Nombre total de déficiences enregistrées
Établissement	structure_code	Numérique	Code numérique du foyer (encodeur d'étiquettes)
Composé	score_risque_cumule	Numérique	Âge × nb_hospit_passees (variable d'interaction)

Antécédent hospitalier (7 variables)

Les antécédents d'hospitalisation constituent le groupe de variables le plus prédictif. Elles sont calculées en ne retenant que les hospitalisations terminées avant la date de mesure (fin abs < Date Relevé). Ces variables comprennent :

- Le nombre total d'hospitalisations passées
- Le nombre d'hospitalisations dans les 6 et 12 derniers mois
- Les jours écoulés depuis la dernière sortie (valeur conventionnelle de 9999 pour les patients jamais hospitalisés).
- La durée totale cumulée,
- La durée moyenne par séjour
- La durée de la dernière hospitalisation

Ces sept indicateurs capturent à la fois la fréquence, la récurrence et la sévérité des hospitalisations passées.

Indicateur nutritionnel (3 variables)

La vitesse de perte de poids mensuelle est calculée par comparaison avec la mesure précédente via un décalage temporel. Cette variable est normalisée par l'intervalle de temps en mois. La Haute Autorité de Santé (2021) a établi qu'une perte de poids d'au moins 5% en un mois constitue un critère diagnostique de dénutrition. Cette variable est donc pertinente comme indicateur de fragilité. La valeur brute de l'IMC et la valeur du poids au moment de la mesure complètent ce groupe.

Charge thérapeutique (1 variable)

La Polymédication (nb_medicaments_actifs p) est calculée par jointure sur les périodes de prescription actives à chaque date de mesure. Cette variable est reconnue comme prédicteur de fragilité et d'instabilité clinique (Maher et al., 2014, cité par la Haute Autorité de Santé, 2021).

Profil démographique et pathologique (5 variables)

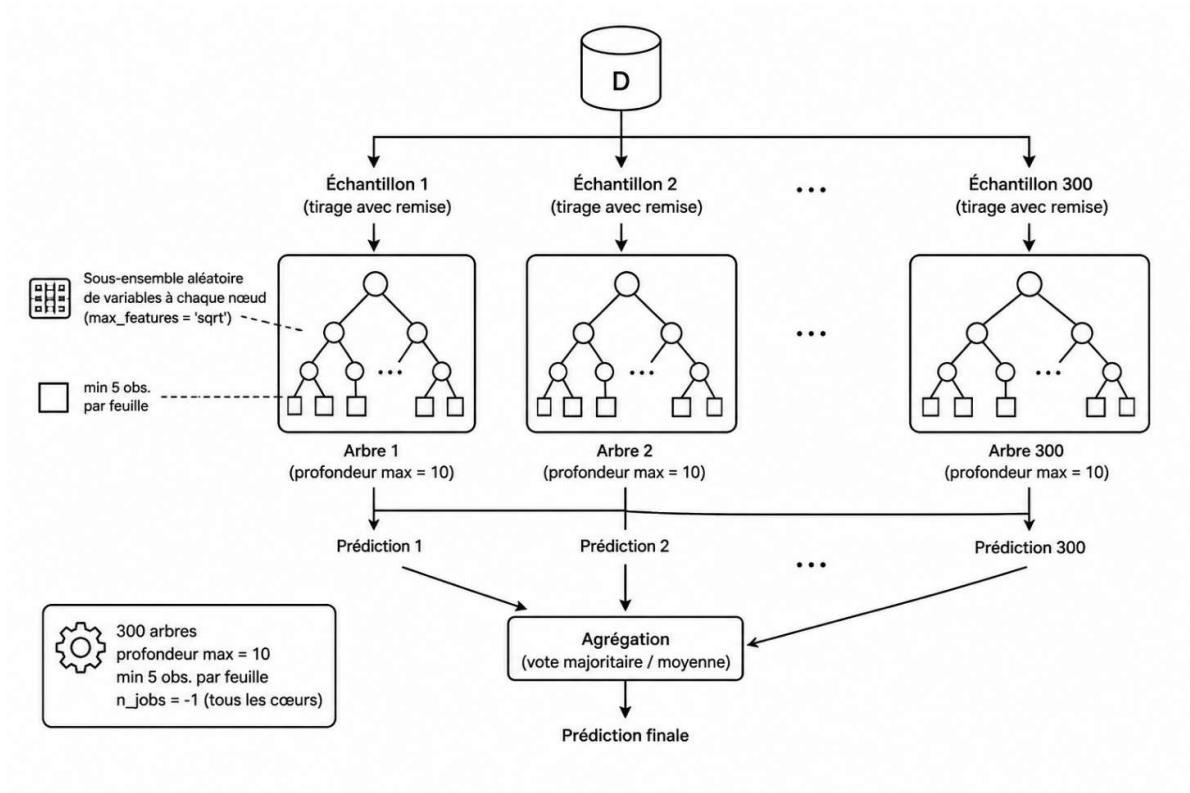
L'âge est calculé dynamiquement (année courante moins année de naissance) pour refléter l'âge au moment de chaque mesure. Le sexe produit 2 variables indicatrices (sexe_F et sexe_M). La sévérité de la déficience principale et le nombre total de déficiences enregistrées complètent ce groupe.

Variable composée (2 variables)

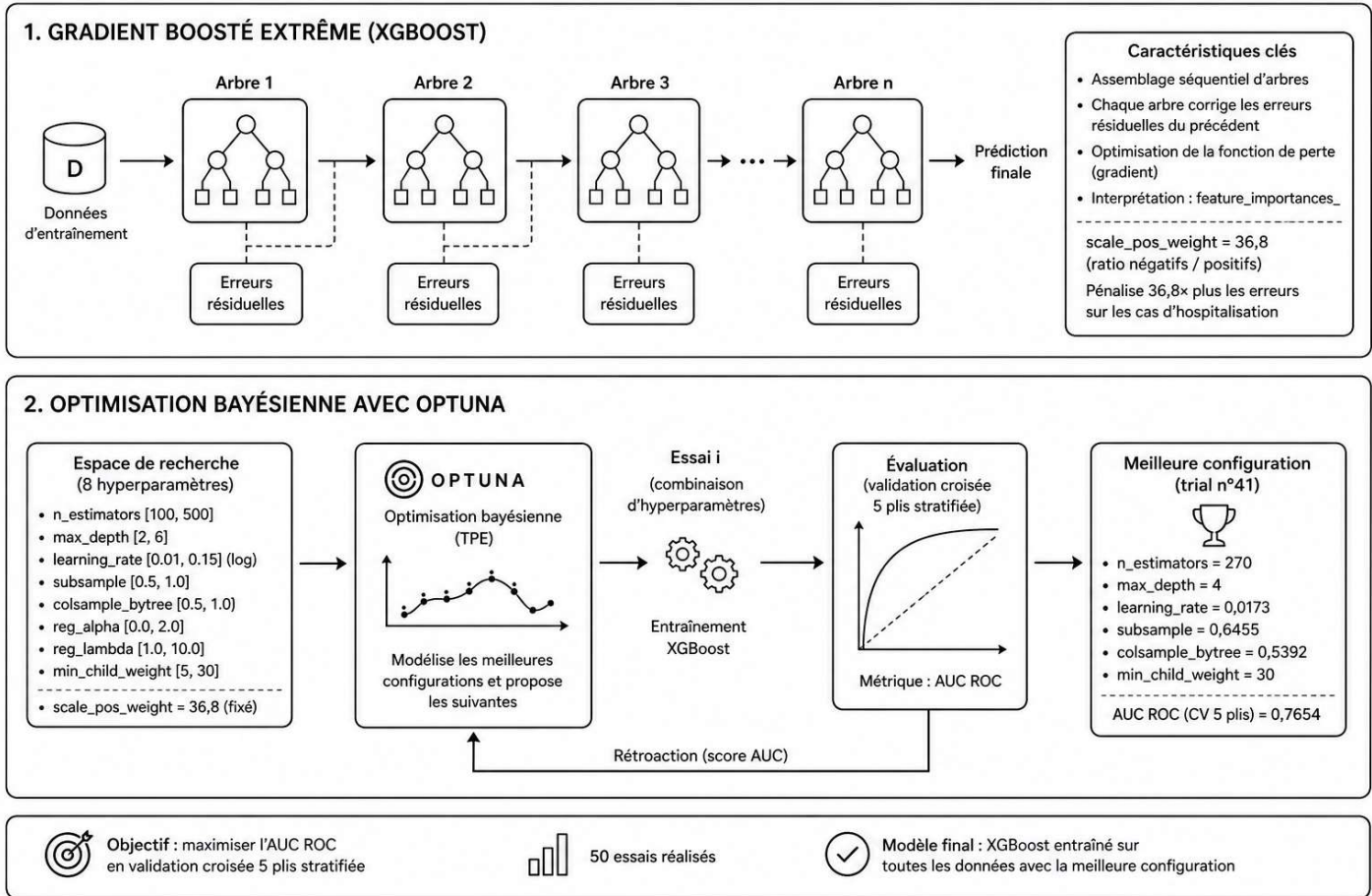
Le score de risque cumulé ($\text{âge} * \text{nb_hospit_passees}$) capture l'interaction entre vieillissement et récurrence des hospitalisations. Le nombre de relevés d'IMC passés (nb_mesures_imc_passees) mesure l'ancienneté du suivi du patient dans le réseau, indicateur indirect de la stabilité ou fragilité chronique.

Annexe C : Schémas des différents modèles

C1. Schéma de la Forêt Aléatoire

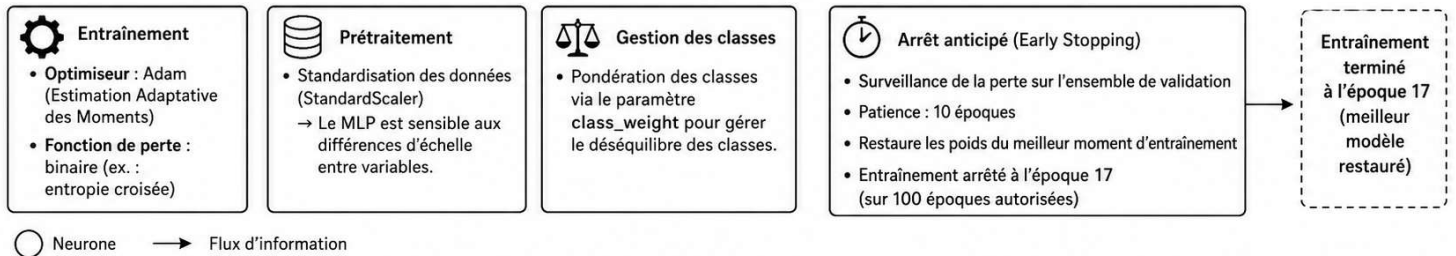
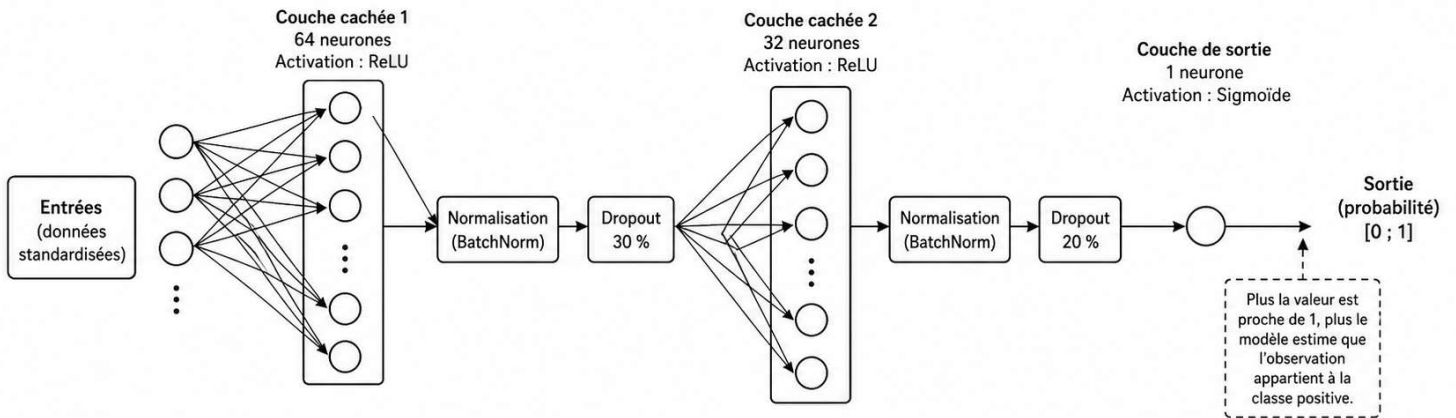


C2. Schéma du modèle Gradient Boosté Extrême



C3. Schéma du Réseau de neurones du type Perceptron Multicouche

Perceptron Multicouche (MLP)

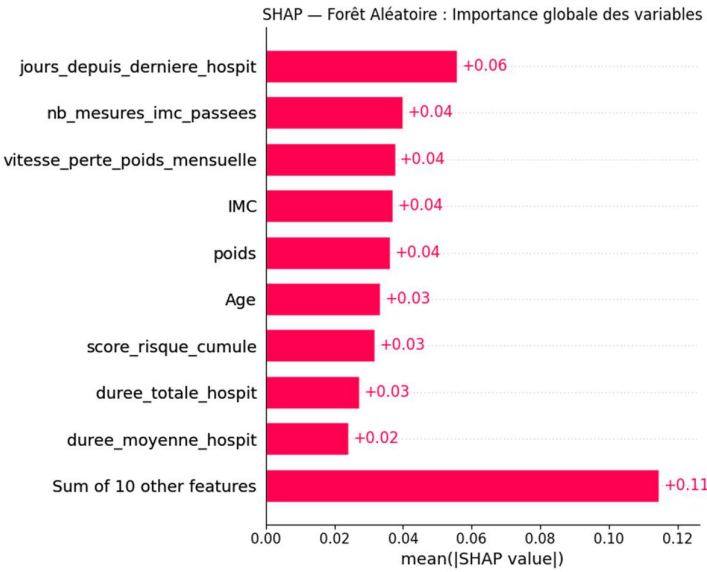


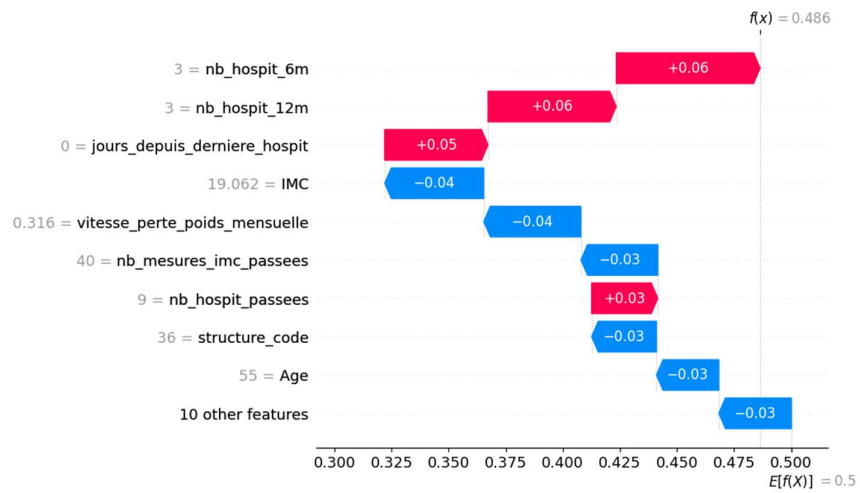
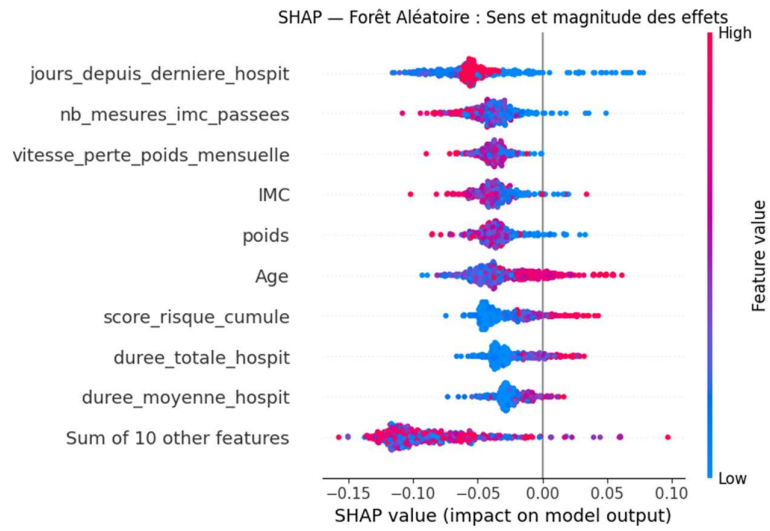
Annexe D. Extrait du tableur de la liste de surveillance hebdomadaire par foyers

	A	B	C	D	E	F
1	IDUsage	Score risque (%)	Nb hospit passées	Jours depuis hospit	Hospit 12 mois	Dernier IMC
2	1416	99,69999695	48	14	9	23,3
3	1748	99,69999695	26	6	9	40,1
4	317	99,40000153	46	120	5	22,7
5	472	99,30000305	28	7	8	22,5
6	2156	99	17	16	7	20,1
7	2160	98,19999695	23	84	6	22,5
8	207	98,09999847	11	0	6	21
9	1574	97,90000153	7	0	7	24,9
10	2314	97,80000305	9	51	6	30,3
11	2157	97,59999847	21	46	8	23,4
12	374	97,30000305	43	12	4	19,4
13	512	97,19999695	16	5	7	22,1
14	322	97,09999847	13	11	7	26,7
15	1243	96,59999847	19	55	6	21,8
16	441	96,40000153	29	2	5	18,5
17	786	95,90000153	6	9	6	20,9
18	609	95,90000153	25	64	4	42,6
19	214	95,30000305	25	29	5	16,8
20	2147	95,19999695	15	24	5	21,8

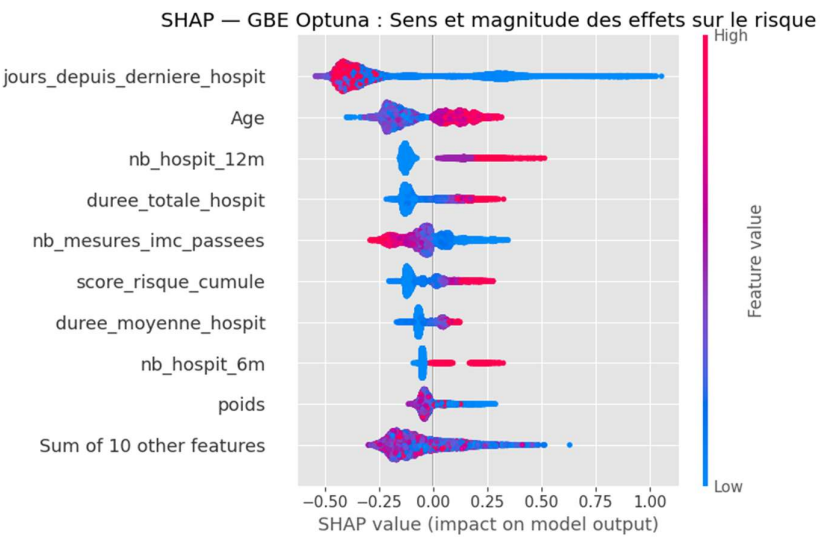
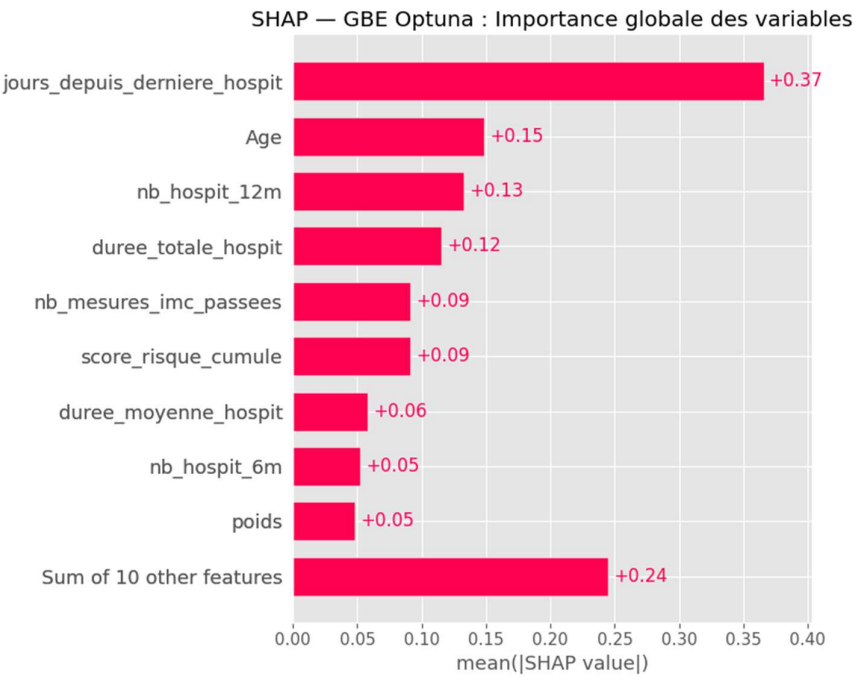
Annexe D : la colonne « Foyer » précisant la résidence du patient a été supprimée pour raison de confidentialité

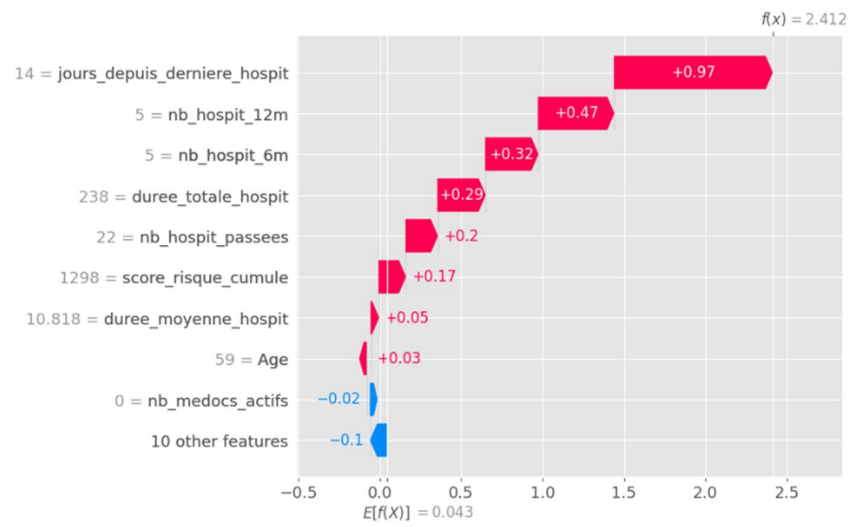
Annexe E. Résultats de la méthode SHAP Forêt Aléatoire





Gradient Boosté Extrême :





Modèle Perceptron Multicouche :

