الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي والبحث العلمي

جامعة سعيدة د. مولاي الطاهر

كلية الرياضيات و الإعلام الآلي و الاتصالات السلكية و اللاسلكية

قسم: الإعلام الآلي

Mémoire de Master en informatique

Spécialité :

Modélisation Informatique des Connaissances et du Raisonnement

T h è m e

Les Techniques de Machine Learning pour la Détection et la Prédiction des Maladies Infectieuses

▪ **Présenté par :**
Mosbahi Aya Ferdous

Nedjadi Imene

▪ **Dirigé par :**
Dr. Zerrouki Kadda

Année universitaire 2024-2025

Remerciements

Nous souhaitons à exprimer notre profonde gratitude au Dr Zerrouki Kadda pour son encadrement, pour ses précieux conseils et sa disponibilité tout au long de ce travail.

Nous remercions également les membres du jury pour avoir accepté d'évaluer et de juger ce travail.

Enfin, nous adressons nos sincères remerciements à toutes les personnes qui, de près ou de loin, ont participé à l'aboutissement de ce travail.

Dédicace

Je dédie ce travail :

À moi-même, pour la persévérance et les efforts fournis tout au long de ce parcours malgré les défis.

À ma famille, pour son soutien moral et ses encouragements tout au long de mon parcours universitaire.

Ce travail vous est dédié avec toute ma reconnaissance et ma gratitude.

Mosbahi Aya Ferdous.

Dédicace

À moi-même… pour avoir enduré et lutté, pour être restée forte malgré tout ce que j'ai traversé.

À ma mère… source de tendresse et de force, la première à m'aimer sans condition.

À mon père… mon premier soutien, un pilier inébranlable, la lumière de mon chemin.

À ma sœur… complice de mes jours et de mes rires, qui a été une amie avant d'être une sœur.

À ma chère amie… qui a été à mes côtés lorsque j'avais besoin d'un appui, merci d'avoir été la lumière dans mes moments sombres.

Nedjadi Imene.

Résumé

Le Machine Learning s'est révélé hautement puissant pour lutter contre les maladies infectieuses en permettant une détection précoce, une prédiction précise des épidémies, leurs risques et un diagnostic assisté. Grâce à ses différentes techniques, il analyse des données complexes (symptômes, imagerie médicale, facteurs environnementaux) pour identifier des patterns invisibles à l'œil humain. Cependant, la qualité et la complexité des données et les questions éthiques liées à leur utilisation restent des défis majeurs. En dépit de ces limites, le Machine Learning offre des perspectives prometteuses pour améliorer la santé publique et sauver des vies grâce à des décisions médicales plus rapides et plus précises. Dans le cadre de cette étude, nous avons mené une recherche scientifique et examiné différentes techniques de Machine Learning pour la détection et la prédiction des maladies infectieuses à partir de trois ensembles de données sélectionnés, chacun correspondant à une maladie différente : SIDA/VIH, hépatite C et COVID-19. Pour chaque pathologie, nous avons adapté une approche de modélisation spécifique ce qui a donné des résultats prometteurs et pertinents. Ces résultats ouvrent la voie à plusieurs perspectives, notamment l'utilisation d'ensembles de données plus vastes, l'exploration des approches de Deep Learning, ainsi que l'intégration de ces modèles dans les systèmes d'aide à la décision médicale pour un diagnostic plus précis.

Mots-clés : Maladies Infectieuses, Détection Précoce, Stratification du risque, Machine Learning, Classification.

Abstract

Machine learning has proven to be highly powerful in the fight against infectious diseases by enabling early detection, accurate prediction of epidemics and their risks, and assisted diagnosis. Through its various techniques, it analyzes complex data (symptoms, medical imaging, environmental factors) to identify patterns invisible to the human eye. However, the quality and complexity of the data and the ethical issues related to its use remain major challenges. Despite these limitations, machine learning offers promising prospects for improving public health and saving lives through faster and more accurate medical decisions. In this study, we conducted scientific research and examined different machine learning techniques for the detection and prediction of infectious diseases from three selected datasets, each corresponding to a different disease: AIDS/HIV, hepatitis C, and COVID-19. For each pathology, we adapted a specific modeling approach, which yielded promising and relevant results. These results open the way to several perspectives, including the use of larger data sets, the exploration of Deep Learning approaches and the integration of these models into medical decision support systems for more precise diagnosis.

Keywords: Infectious Diseases, Early Detection, Risk Stratification, Machine Learning, Classification.

ملخص

أثبت التعلم الآلي فعاليته الكبيرة في مكافحة الأمراض المعدية، إذ يُمكّن من الكشف المبكر، والتنبؤ الدقيق بالأوبئة ومخاطرها، والمساعدة في التشخيص. ومن خلال تقنياته المتنوعة، يُحلل البيانات المعقدة (الأعراض، والتصوير الطبي، والعوامل البيئية) لتحديد الأنماط غير المرئية للعين البشرية. ومع ذلك، لا تزال جودة البيانات وتعقيدها، بالإضافة إلى القضايا الأخلاقية المتعلقة باستخدامها، تُشكل تحديات رئيسية. ورغم هذه القيود، يُقدم التعلم الآلي آفاقًا واعدة لتحسين الصحة العامة وإنقاذ الأرواح من خلال اتخاذ قرارات طبية أسرع وأكثر دقة. في هذه الدراسة، أجرينا بحث علمي ودراسة تقنيات التعلم الآلي المختلفة للكشف عن الأمراض المعدية والتنبؤ بها من ثلاث مجموعات بيانات مختارة، كل منها يُمثل مرضًا مختلفًا: الإيدز/فيروس نقص المناعة البشرية، والتهاب الكبد الوبائي سي، وكوفيد-19. لكل مرض، اعتمدنا نهج نمذجة محددًا، مما أسفر عن نتائج واعدة وذات صلة. تفتح هذه النتائج الباب أمام العديد من وجهات النظر، بما في ذلك استخدام مجموعات بيانات أكبر، واستكشاف أساليب التعلم العميق، ودمج هذه النماذج في أنظمة دعم القرار الطبي لتحقيق تشخيص أكثر دقة.

الكلمات المفتاحية: الأمراض المعدية، الكشف المبكر، تصنيف المخاطر، التعلم الآلي، التصنيف.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

Abréviation	Expression Complète
SIS	Susceptible-Infectious-Susceptible.
SIR	Susceptible-Infectious-Recovered.
SEIR	Susceptible-Exposed-Infectious-Recovered.
SIRS	Susceptible -Infectious-Recovered-Susceptible.
VIH	Virus de l'Immunodéficience Humaine.
SIDA	Syndrome d'Immunodéficience Acquise.
TP	True positive.
TN	True Negative.
FP	False Positive.
FN	False Negative.
TPR	True Positive Rate.
FPR	False positive Rate.
AUC-ROC	Area Under the Receiver-Operating Characteristic Curve.
LR	Logistic Regression.
RF	Random Forest.
DT	Decision Tree.
C4.5	C4.5 Decision Tree Algorithm.
C5.0	C5.0 Decision Tree Algorithm.
KNN	K-Nearest Neighbors.
NB	Naïve Bayes.
SVM	Support Vector Machines.
MLP	Multi-Layer Perceptron.
ANN	Artificial Neural Network.
MSO-MLP	Multi Swarm Optimization avec MLP.
PSO	Particle Swarm Optimization.
PSO-ANN	Particle Swarm Optimization avec ANN.

GBC	Gradient Boosting Classifier.
XGBoost	Extreme Gradient Boosting.
ET	Extra Trees Classifier.
AdaBoost	Adaptive Boosting.
LightGBM	Light Gradient Boosting Machine.
CatBoost	Categorical Boosting.
1-DCNN	One-Dimensional Convolutional Neural Network.
GRU	Gated Recurrent Unit.
LSTM	Long Short-Term Memory.
MARS	Multivariate Adaptive Regression Splines.
BGLM	Bayesian Generalized Lineaer Model.
HPM	Hybrid Prediction Model.
IRF	Improved Random Forest (RF + Bootsrap).
ELM	Extreme Learning Machine.
Acc.	Exactitude (Accuracy).
Prec.	Précision.
Rec.	Rappel (Recall).
RUS	Random Over Sampling.

Introduction générale

Contexte

Au niveau mondial, les maladies infectieuses représentent une menace sérieuse pour la santé publique. Chaque année, les maladies saisonnières comme la grippe saisonnière et le rhume ainsi que les pandémies comme le COVID-19 engendrent des millions de pertes humaines et ont une influence négative sur le développement socio-économique. De ce fait, une meilleure prise en charge est cruciale pour faire face à ces maladies infectieuses et il est indispensable que les systèmes de santé assurent une détection précoce et surtout efficace des maladies.

L'application du Machine Learning dans le domaine médical a marqué une avancée significative, en facilitant la détection précoce et la prédiction des maladies infectieuses, il apporte la possibilité d'analyse automatisée, rapide et plus précise des grandes quantités de données médicales ce qui permet aux professionnels de santé de mieux analyser et comprendre les dynamiques épidémiologiques ainsi qu'à améliorer leur aptitude aux crises sanitaires. Dans ce cadre, le Machine Learning est une technologie stratégique pour une gestion plus efficace des maladies infectieuses.

Problématique

Les maladies infectieuses représentent une menace très grave pour la santé mondiale, et ont un impact négatif dans le domaine sanitaire, économique et social. En dépit des progrès médicaux, la détection précoce et la prévention de ces maladies restent un défi, en raison de plusieurs facteurs notamment :

- Complexité des données : l'analyse des données médicales est complexe à cause de leur volume massif, leur hétérogénéité (textes, séries temporelles, images, etc…), leur incomplétude ainsi que leur déséquilibre.
- Propagation rapide des épidémies : les maladies infectieuses se propagent d'une manière très rapide, cela nécessite des outils prédictifs performants et efficaces pour faire face aux crises sanitaires.
- Limite des méthodes traditionnelles : les méthodes de diagnostic traditionnelles sont parfois lentes et coûteuses, ce qui les rend moins efficaces.
- Exigence de personnalisation : chaque épidémie évolue dans un contexte spécifique, ainsi chaque patient présente des caractéristiques uniques, ce qui nécessite des modèles

prédictifs personnalisés et l'adaptation des interventions en fonction des risques individuels.

Dans ce contexte, Comment contribue le Machine Learning dans l'amélioration de la détection et la prédiction des maladies infectieuses et l'estimation du risque de mortalité chez les patients infectés ? Quelles techniques les plus adaptées pour traiter les données médicales complexes, incomplètes et déséquilibrées ? Dans un contexte médical, quelles sont les défis techniques, pratiques et éthiques associées à l'application de ces modèles ? Enfin, comment intégrer ces technologies de manière efficace et surtout responsable dans les systèmes de santé existants.

Objectif

L'objectif de ce travail est d'explorer et d'analyser les techniques de Machine Learning pour la détection et la prédiction des maladies infectieuses et l'estimation de risque de mortalité chez les patients infectés. L'étude a pour but de montrer comment les algorithmes de Machine Learning peuvent améliorer la précision diagnostique et renforcer les décisions médicales face aux épidémies. Ce document vise également à identifier les avantages offerts par ces technologies ainsi que les perspectives pour son intégration dans les systèmes de santé.

Chapitre 01 : Les maladies infectieuses

1.1 Introduction

L'épidémiologie est la science qui étudie la manière dont les maladies se propagent dans les populations, ainsi que leurs causes et les moyens de les contrôler. Parmi ces maladies, les maladies infectieuses sont causées par des agents pathogènes comme des virus, des bactéries ou des parasites, et peuvent se transmettre d'une personne à une autre de plusieurs manières. Comprendre comment ces maladies se transmettent et comment les épidémies évoluent est essentiel pour mieux les prévenir et les combattre.

Dans ce chapitre, nous allons explorer les différents mécanismes de transmission des maladies infectieuses, ainsi que la dynamique des épidémies à travers des modèles mathématiques simples comme SIS et SIR, qui sont utiles pour comprendre et construire des outils prédictifs modernes, notamment en Machine Learning.

1.2 Mécanismes de transmission des maladies infectieuses

Les maladies infectieuses sont des affections causées par des agents pathogènes tels que les bactéries, les virus, les champignons ou les parasites. Ces agents pathogènes peuvent se transmettre d'un individu à un autre ou entre individus ou via l'environnement selon différents modes de transmission, notamment :

1.2.1 Transmission directe

Elle se produit par le contact physique ou contact direct entre individus (par exemple, les maladies sexuellement transmissibles). Ces agents pathogènes peuvent aussi se transmettre pendant la toux ou l'éternuement par dispersion de gouttelettes. Une autre transmission directe est la transmission mère-enfant ; pendant la grossesse, les maladies infectieuses peuvent être transmises de la mère au fœtus.

Parmi les maladies infectieuses qui peuvent se transmettre directement on note : VIH/SIDA, la syphilis, COVID-19, hépatite B…etc.

1.2.2 Transmission indirecte

Elle se produit par l'intermédiaire de plusieurs facteurs notamment : Les organismes vivants (Appelé vecteurs), qui transportent l'agent pathogène d'un hôte à un autre via une piqûre comme les moustiques qui sont des vecteurs pour le paludisme et la dengue par exemple. Il y a aussi la transmission fécale-orale, qui est transmises par la consommation d'aliments ou d'eau

contaminés par des matières fécales. Une autre transmission indirecte est la transmission par aérosols, qui se distingue par la suspension d'agent pathogène dans l'air et son déplacement sur de longues distances comme la tuberculose. Quant à la transmission par fomites, elle désigne la propagation par des objets ou surfaces contaminées qui ont été en contact avec un certain agent pathogène.

1.3 La dynamique des épidémies et les modèles de propagation

L'épidémiologie des maladies infectieuses cherche à identifier les agents pathogènes, à comprendre leur mode de propagation [1], analyse comment une certaine maladie transmissible apparaît, atteint un pic, puis persiste au sein d'une population ainsi qu'à modéliser la dynamique des épidémies afin de guider la prévention et la prédiction de l'évolution des maladies infectieuses. La dynamique des épidémies dépend de plusieurs facteurs, notamment les propriétés biologiques du pathogène, la durée de l'infection, les comportements sociaux et les conditions environnementales.

1.3.1 Les modèles de propagation

La modélisation mathématique de l'épidémiologie des maladies infectieuses constitue un véritable outil de santé publique. Elle offre la possibilité d'évaluer, de manière rapide et efficace, les stratégies de lutte proposées [1].

Plusieurs modèles mathématiques compartimentaux permettent de représenter les interactions entre individus selon l'état par rapport à la maladie (susceptible à se faire infecter, infecté mais non contagieux, infecté contagieux, immunisé, décédé, etc.), et servent à décrire et à prédire l'évolution d'une épidémie.

Les compartiments les plus couramment utilisés sont [1] :

- **S (Susceptibles)** : individus sains pouvant être infectés.
- **E (Exposés)** : individus en incubation, exposés à la maladie mais non encore contagieux.
- **I (Infectés)** : individus infectieux, capables de transmettre la maladie.
- **R (Résistants ou Retirés)** : Résistants. Contient les individus retirés de la dynamique de la maladie.

1.3.1.1 Modèle SIS (Susceptible-Infecté-Susceptible)

Le modèle SIS est le modèle le plus simple, qui n'attribue pas l'immunité, il considère que les individus guéris reviennent dans le groupe des susceptibles sans acquérir d'immunité permanente. Ce modèle divise la population en deux compartiments : les Susceptibles S et les Infectés I, comme le montre la figure 1 :

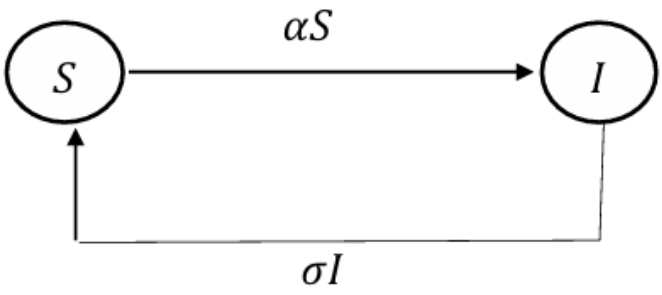

Figure 1. Schéma du modèle SIS [1].

1.3.1.2 Modèle SIR (Susceptible-Infecté-Retiré)

C'est un modèle mathématique simple utilisé en épidémiologie pour représenter la propagation des maladies infectieuses. Il divise la population en trois compartiments : Les susceptibles S, les infectés I et les guéris R, comme la figure montre :

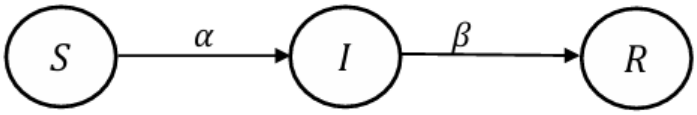

Figure 2. Schéma du modèle SIR [1].

Les transitions sont modélisées par des équations différentielles :

- Les Susceptibles deviennent Infectés au taux β (taux de transmission).
- Les Infectés deviennent Retirés au taux γ (taux de guérison ou de retrait).

Ainsi, l'évolution de l'épidémie dépend du nombre de reproduction de base $R_0 = \beta / \gamma$:

- L'épidémie peut se propager Si $R_0 > 1$.

- L'épidémie finit par disparaître Sinon.

Il existe aussi d'autres extensions du modèle SIR, notamment : SIRS qui introduit la possibilité que les individus guéris perdent leur immunité après un certain temps et redeviennent susceptibles à l'infection. La deuxième extension du modèle SIR est SEIR où l'on ajoute un compartiment E (Exposés) pour modéliser la période d'incubation, qui représente les personnes infectées mais pas encore infectieuses.

En divisant la population en compartiments : susceptibles, infectés, Retirés et exposés, des modèles mathématiques tels que SIR, SIS, SIRS et SEIR nous permettent de comprendre la propagation des maladies infectieuses. Ils peuvent prédire comment une épidémie se développe et simuler les effets de diverses interventions médicales (comme la vaccination et le confinement). Ces modèles servent aussi pour créer des modèles prédictifs plus avancés, qui utilisent des données réelles, ce qui permet d'aider les autorités sanitaires à mieux décider quoi faire et à anticiper l'évolution d'une épidémie.

1.4 Données et sources épidémiologiques

De nombreuses sources de données sont utilisées dans l'analyse et la modélisation des maladies infectieuses. Les informations cliniques (symptômes, résultats d'examens, dossiers médicaux) permettent de suivre l'évolution des cas et de caractériser les patients. Une représentation visuelle des effets pathologiques est fournie par les données d'imagerie médicale (radiogrammes, scanners), ce qui est utile en cas d'infections respiratoires comme le COVID-19 et la pneumonie. De plus, des données en temps réel sur le comportement de la population, la perception des risques et la diffusion de l'information sont fournies par les réseaux sociaux comme Twitter, Facebook et d'autres plateformes numériques.

1.4.1 Défis liés aux données

- Collecte incomplète : dans certaines régions, les données médicales sont rares.
- Qualité et bruit : certaines données, notamment issues des réseaux sociaux, peuvent contenir des erreurs ou être biaisées.
- Diversité des formats : les données cliniques, images médicales et données numériques ne sont pas structurées de la même manière, ce qui rend leur intégration complexe.

- Enjeux éthiques et confidentialité : l'utilisation de données personnelles, surtout en ligne, soulève des questions de confidentialité.

1.5 Conclusion

Bien comprendre comment une maladie se transmet et la dynamique d'une épidémie aide à mieux contrôler. Les modèles comme SIS ou SIR permettent de voir comment les gens passent d'un état sain à malade, puis ont guéri. Ces modèles simples sont très utiles pour mieux prévoir les risques. Ils servent aussi de base aux méthodes plus avancées, comme les modèles prédictifs utilisant des données réelles et le Machine Learning, pour mieux protéger la santé publique.

Chapitre 02 : Application du Machine Learning aux maladies infectieuses

2.1 Introduction

Le Machine Learning est un sous domaine de l'Intelligence Artificielle qui est souvent utilisé dans le domaine médical dans la détection précoce et la prédiction des maladies.

Dans ce chapitre, nous allons présenter les concepts clés de Machine Learning. Nous commencerons par présenter les différents types de Machine Learning notamment l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement. Nous verrons aussi les algorithmes importants et couramment utilisés de Machine Learning et les différentes métriques d'évaluation des performances des modèles ainsi que des études précédentes qui montrent l'application de ces techniques dans la détection et la prédiction des maladies infectieuses, ce qui permet de bien situer et comprendre le cadre dans lequel s'inscrit notre travail.

2.2 Machine Learning

2.2.1 Définition

Machine Learning est l'étude scientifique des algorithmes et des modèles statistiques que les systèmes informatiques utilisent pour effectuer une tâche spécifique sans être explicitement programmés [2].

Le machine Learning est une méthode d'analyse de données traitant de la construction et de l'évaluation des algorithmes. C'est la science qui donne aux machines à calculer la capacité d'agir sans être explicitement programmées. Il est défini par la capacité de choisir des caractéristiques efficaces pour la reconnaissance, la classification et la prédiction de modèles en fonction de modèles dérivés de données existantes [3].

Selon Arthur Samuel le Machine Learning est défini comme : "le domaine d'étude qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmé " [2] .

Une autre définition plus technique de Tom Mitchell 1997 : "lorsqu'on donne une tâche T et une mesure de rendement P, on dit qu'un programme informatique apprend d'une expérience E si les résultats obtenus sur T, mesurés par P, s'améliorent avec l'expérience E" [3].

2.2.2 Les différentes approches d'apprentissage en Machine Learning

Les différents algorithmes de Machine Learning sont classés généralement en trois catégories.

2.2.2.1 Apprentissage supervisé

Ce type d'apprentissage est nommé supervisé par ce que les données d'entraînement sont étiquetées par un expert. En d'autres termes, on connait déjà la classe de chaque instance dans les données. L'idée est de trouver la relation entre les entrées et les sorties et de trouver une fonction de correspondance en se basant sur les données étiquetées afin de prédire correctement la sortie des nouvelles données non étiquetées. Parmi les algorithmes d'apprentissage supervisé, on trouve : Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), Neural Networks (NN), etc.

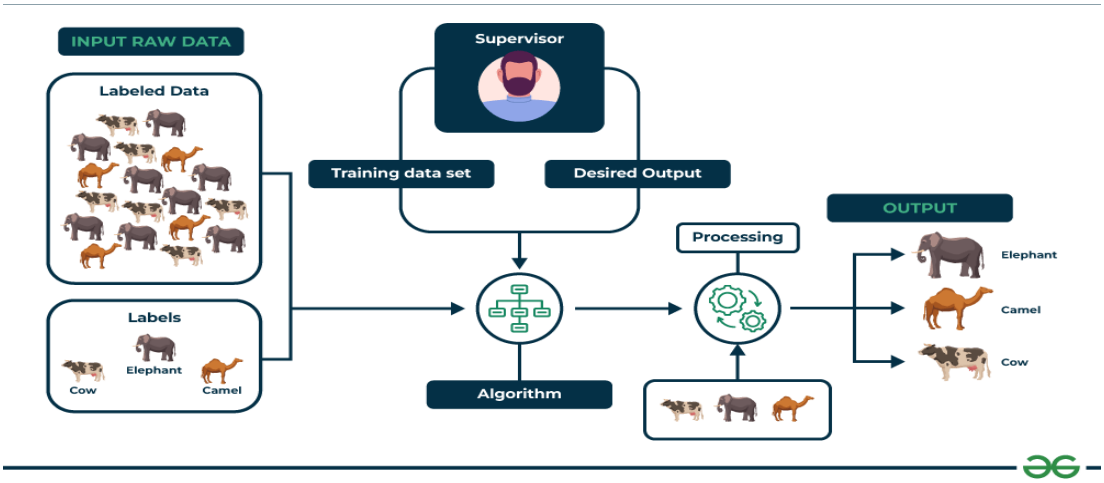

Figure 3. Apprentissage supervisé [4].

L'apprentissage supervisé est utilisé généralement pour deux tâches : La classification ou la régression.

- **La classification** : On parle de la classification quand la variable cible est catégorielle. C'est une tâche consistant à choisir une classe (valeur) parmi toutes celles possibles, Exemple : Un algorithme classifiant une tumeur comme « bénigne » ou « maligne » [5].
- **La régression** : On parle de la régression quand la variable cible est une variable continue. Exemple : Un algorithme prédisant le prix d'un appartement en fonction de ses caractéristiques.

2.2.2.2 Apprentissage non supervisé

Dans ce type d'apprentissage, les données ne sont pas étiquetées. L'objectif du système est d'identifier des caractéristiques communes aux données d'entraînement [5]. Dans cette

approche, la machine analyse les informations non organisées et les regroupe en fonction de similitudes, de modèles ou de différences. Contrairement à l'apprentissage supervisé, il n'y a pas de superviseur ou d'entraînement impliqué. La machine doit découvrir elle-même des structures cachées dans les données [4].

Les problèmes qui peuvent être résolus par ce type d'apprentissage automatique sont les problèmes de regroupement (clustering), les règles d'association d'apprentissage et la réduction de dimensionnalité. Parmi les algorithmes d'apprentissage non supervisé, on trouve : l'algorithme Apriori et l'algorithme K-means (k-moyennes).

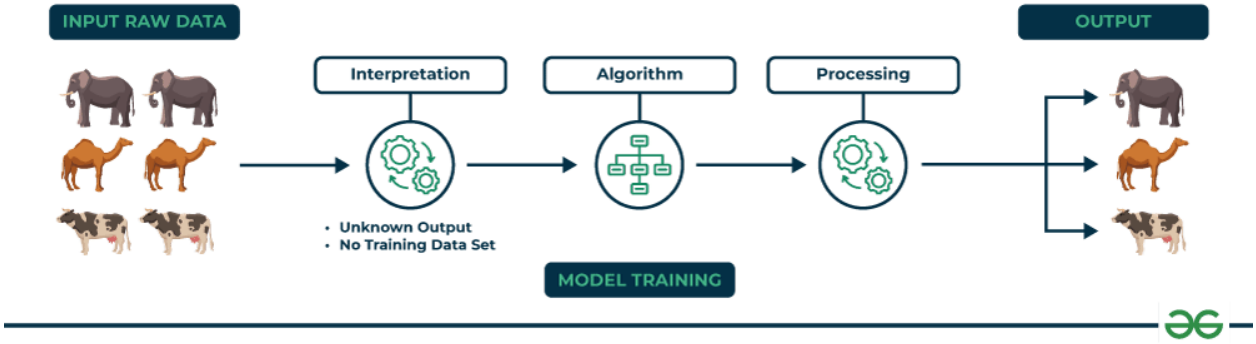

Figure 4. Apprentissage non supervisé [4].

2.2.2.3 Apprentissage par renforcement

C'est un domaine de Machine Learning concerné par la façon dont les agents devraient prendre des mesures dans un environnement pour maximiser une certaine notion de récompense cumulative. La différence entre l'apprentissage par renforcement et l'apprentissage supervisé est le signal de l'enseignant. Le signal de renforcement fourni par l'environnement dans l'apprentissage du renforcement est utilisé pour évaluer l'action (signal scalaire) plutôt que de dire au système d'apprentissage comment effectuer les actions correctes [6].

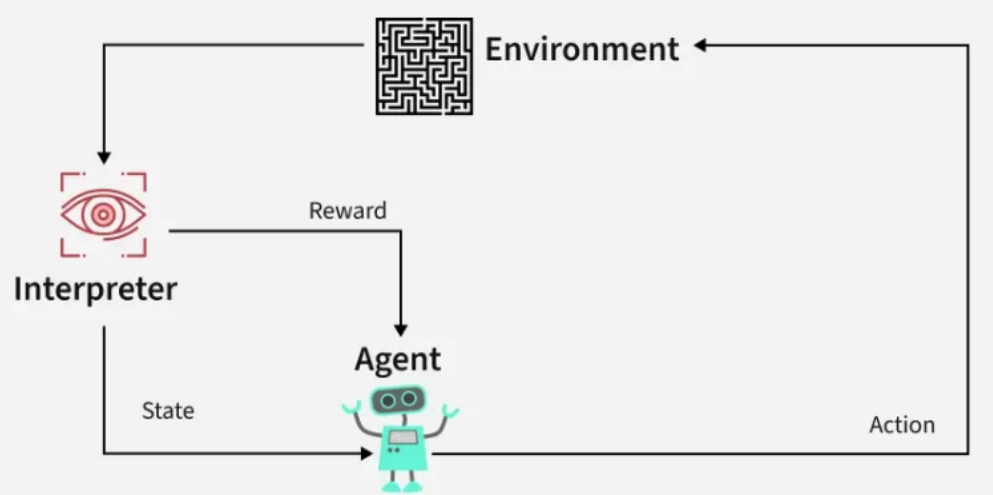

Figure 5. Apprentissage par renforcement [7].

2.2.3 Algorithmes de Machine Learning

2.2.3.1 Régression logistique

La régression logistique est un algorithme de Machine Learning supervisé qui est utilisé spécifiquement pour les tâches de classification binaire. Le modèle prédit une probabilité (allant de 0 à 1) qui détermine si une instance appartient à une classe spécifique avec l'utilisation de la fonction sigmoïde. Et applique un seuil (Threshold) pour prendre une décision de classification finale. Et selon le seuil choisi (par défaut 0.5), la décision de classification finale est prise (si la probabilité estimée > seuil, alors l'instance est classée dans une classe spécifique, sinon elle sera classée dans l'autre classe). Cet algorithme est utilisé dans diverses applications, par exemple : Classer les e-mails comme spam ou non spam, ou dans le domaine médical pour prédire si un patient s'est rétabli ou non, etc.

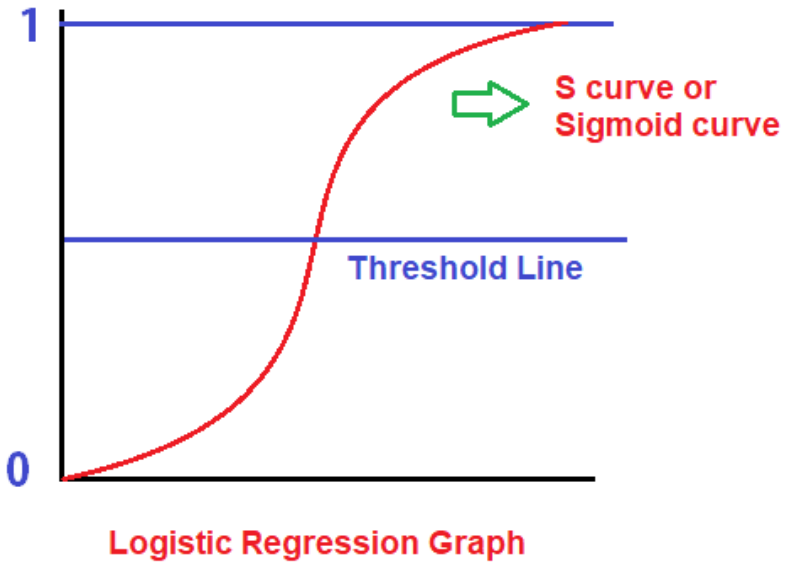

Figure 6. Courbe sigmoïde en régression logistique [8].

2.2.3.2 Arbres de décision

C'est un algorithme de Machine Learning supervisé qui est utilisé pour les tâches de classification ou de régression. Cette structure est constituée d'un nœud racine qui contient tout l'ensemble de données, des nœuds internes, chacun séparant les données selon une caractéristique, des branches qui représentent une valeur que le nœud peut prendre et des feuilles qui représentent la classe finale attribuée par le modèle. Le processus des arbres de décision commence d'abord par la sélection du nœud racine en utilisant un critère comme l'entropie ou l'indice de Gini. Ensuite, chaque nœud interne sépare les données en plusieurs groupes en prenant une décision selon différentes branches et crée de nouveaux sous-groupes. Ce processus se répète jusqu'à atteindre une feuille qui représente la classe finale attribuée.

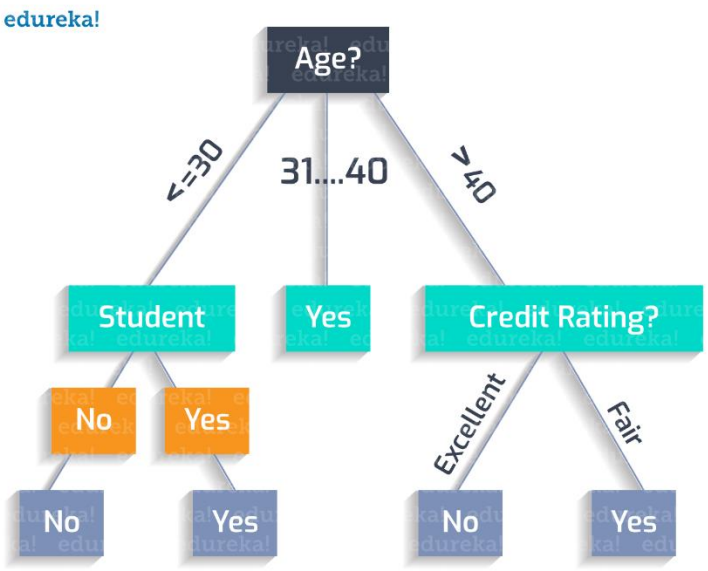

Figure 7. Architecture de l'Arbre de décision [9].

2.2.3.3 Forêts aléatoires

La forêt Aléatoire (ou le Random forest) est l'un des algorithmes de Machine Learning supervisé les plus couramment utilisés pour les tâches de classification ou de régression. Cette forêt aléatoire est constituée d'une combinaison de plusieurs arbres de décision. L'idée est d'entraîner chaque arbre dans la forêt sur des sous-échantillons aléatoires des données. Ces sous-échantillons sont obtenus grâce à la méthode du bootstrap. Afin d'obtenir la prédiction finale, chaque arbre vote pour une classe, ensuite leurs prédictions sont combinées : dans la classification, la classe majoritaire est choisie, tandis qu'en régression, la moyenne des prédictions des arbres est calculée. Les forêts aléatoires ont plusieurs applications dans plusieurs domaines notamment : en domaine de la médecine dans le diagnostic des maladies, dans la détection de fraude, etc.

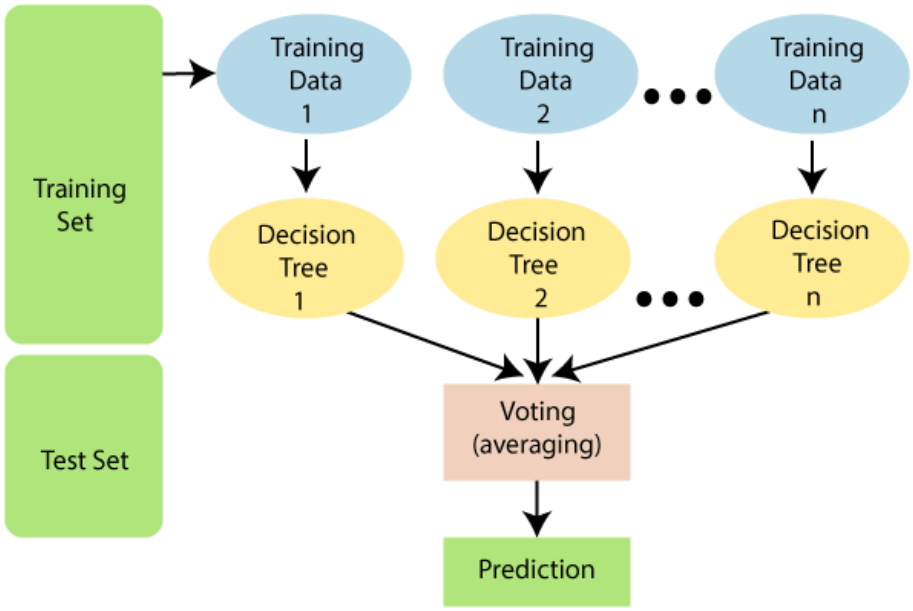

Figure 8. Illustration de processus de classification par la Foret aléatoire [10].

2.2.3.4 SVM

C'est un algorithme de Machine Learning supervisé qui est utilisé pour les tâches de classification ou de régression. En cas des données qui sont linéairement séparables, cet algorithme sépare deux classes de données avec une frontière optimale appelée hyperplan, qui est utilisée pour séparer les points de données. Inversement, pour des données non linéaires, l'algorithme SVM utilise une fonction noyau (ou kernel en anglais) afin de les transformer comme : noyau linéaire, polynomial et gaussien.

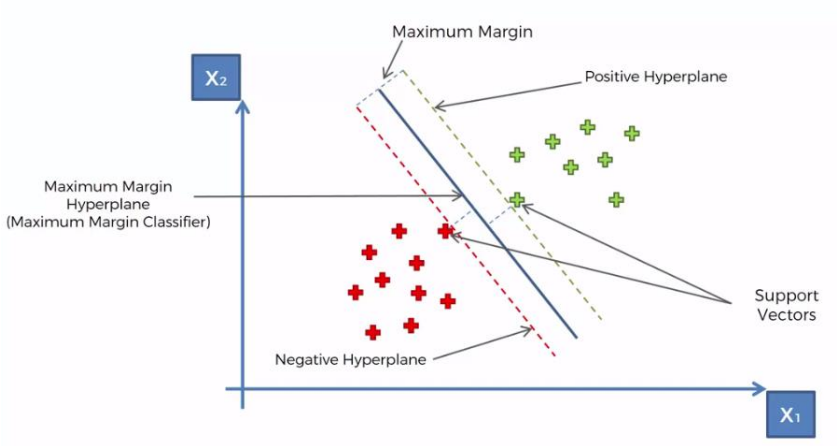

Figure 9. Marges et vecteurs de support dans un modèle SVM [11].

2.2.3.5 KNN

KNN (K Nearest Neighbors) est un algorithme de Machine Learning supervisé qui est utilisé pour les tâches de classification ou de régression. Cet algorithme trouve les K voisins (points de données) les plus proches de la nouvelle entrée en calculant une mesure de distance (similitude) entre cette entrée et tous les points d'entrainement. En cas de classification, la prédiction finale est faite en se basant sur la classe majoritaire des k voisins, alors que dans le cas de la régression, la prédiction finale est faite en se basant sur la valeur moyenne des k voisins. Parmi les mesures de distance on note : la distance Euclidienne, la distance de Manhattan et la distance de Minkowski.

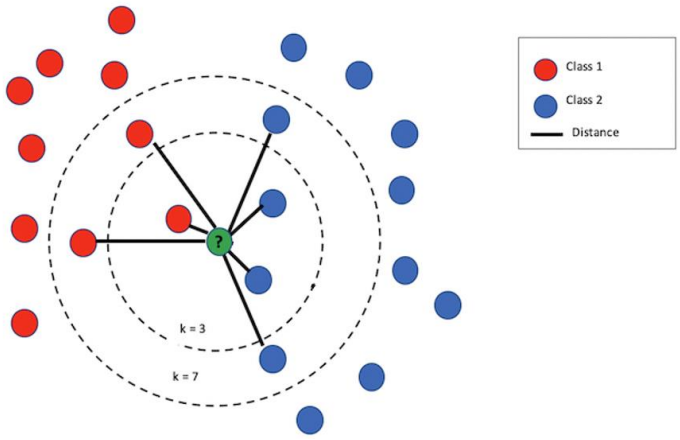

Figure 10. Illustration de processus de classification par KNN [12].

2.2.3.6 Réseau de neurones

Un réseau de neurones est un modèle d'apprentissage qui est inspiré des fonctions du cerveau humain. Il est constitué de plusieurs neurones artificiels groupés en couches (Couche d'entrée, couches cachées et couche de sortie). Le réseau reçoit les données dans la couche d'entrée qui se compose des neurones chacun correspond à une caractéristique des données d'entrée [13]. Un réseau de neurones peut avoir un ou plusieurs couches cachées, ces couches effectuent le calculs [13]. Son fonctionnement est come ceci : Après que les données entrent dans la couche d'entrée et passent à travers les couches cachées en calculant une somme pondérée et ajouter un biais à la somme, ce résultat est passé à une fonction d'activation (ReLU, sigmoide, etc…) jusqu'à arriver à la couche de sortie qui donne la prédiction et ici on parle de la propagation vers l'avant.

Après la propagation vers l'avant, le réseau de neurones compare les prédictions avec les sorties réelles et évalue les performances via une fonction de perte qui cherche à la minimiser et il mit à jour les poids et le biais avec un algorithme d'optimisation comme la descente du gradient après d'avoir calculé les gradients de la fonction de perte, et ici on parle de la rétropropagation.

Ce processus est répété plusieurs fois (epochs) jusqu'à atteindre des prédictions précises.

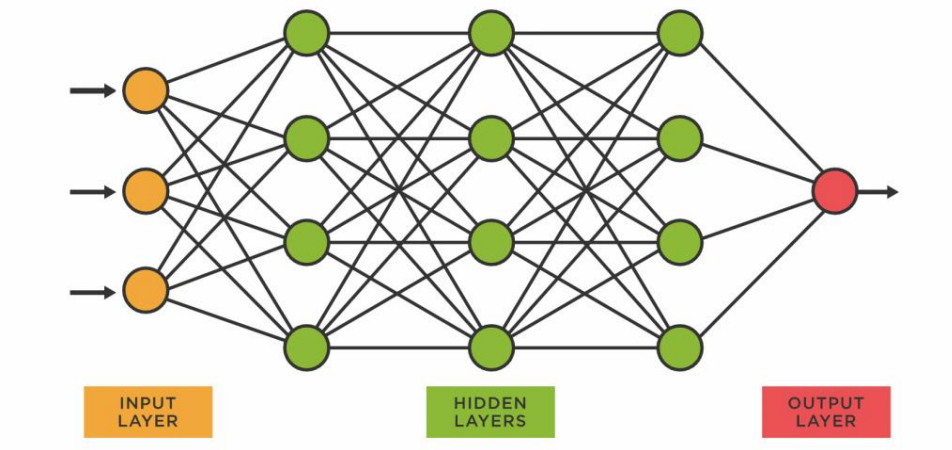

Figure 11. Architecture d'un réseau de neurones [14].

2.2.4 Les métriques d'évaluation

Les métriques d'évaluation sont des métriques utilisées pour mesurer la performance et la qualité d'un modèle de Machine Learning. Il existe plusieurs métriques d'évaluation notamment :

2.2.4.1 La matrice de confusion

La matrice de confusion est une évaluation des performances des classificateurs binaires ou multi-classes. Elle est sous forme d'une matrice N x N (N représente le nombre des classes cibles) qui résume les prédictions du modèle ou chaque cellule représente le nombre de fois où le modèle a prédit correctement ou bien incorrectement une classe.

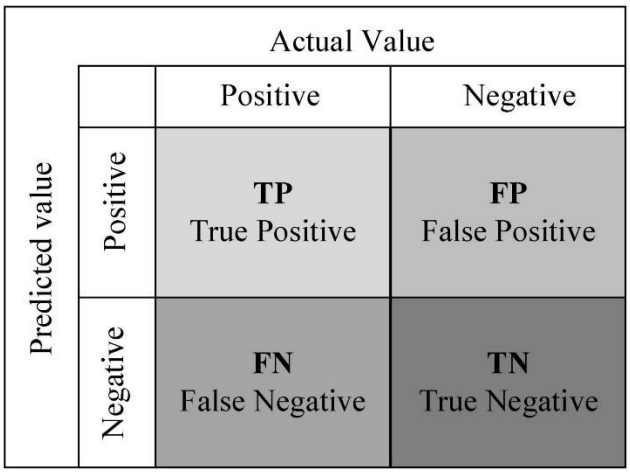

Figure 12. Matrice de confusion [15].

- **True Positive** (TP) : la prédiction faite par le modèle et la valeur réelle sont positives.
- **True Negative** (TN) : la prédiction faite par le modèle et la valeur réelle sont négatives.
- **False Positive** (FP) : ce sont les fausses alarmes, la prédiction faite par le modèle est positive alors que la valeur réelle est négative.
- **False Negative** (FN) : la prédiction faite par le modèle est négative alors que la valeur réelle est positive.

La matrice de confusion permet d'identifier quel type d'erreur commis par le modèle, elle indique la confusion entre les classes. Pour une interprétation efficace, elle doit être accompagnée avec d'autres métriques d'évaluation**.**

2.2.4.2 Exactitude (Accuracy)

L'exactitude est le rapport entre le nombre de classes correctement prédites (TP et TN) et le nombre total de prédictions. Elle mesure le pourcentage de prédiction correctes.

$$Exactitude = \frac{TP+TN}{TP+FP+TN+FN} \quad (2.1)$$

En cas des données déséquilibrées, la mesure d'exactitude n'est pas fiable parce que le modèle privilégie la classe majoritaire et ignore la classe minoritaire.

2.2.4.3 Précision

La précision indique que parmi tous prédictions positives fait par le modèle, combien sont réellement correctes (positives).

$$Précision = \frac{TP}{TP+FP} \tag{2.2}$$

2.2.4.4 Rappel (Recall)

Le rappel indique que parmi tous les cas positifs, combien ont été correctement détectés par le modèle, il est important lorsqu'on ne veut pas rater un vrai cas positif, mais cela peut augmenter les fausses alarmes (FP).

$$Rappel = \frac{TP}{TP+FN} \tag{2.3}$$

2.2.4.5 F1-Score

C'est la moyenne harmonique de la précision et le rappel, elle pénalise le déséquilibre entre la précision et le rappel, et elle est utile en cas des données déséquilibrées.

$$F1\_Score = 2 \times \frac{Précision \times Rappel}{Précision+Rappel} \tag{2.4}$$

2.2.4.6 AUC-ROC (Area Under Curve – Receiver Operating Characteristic)

La courbe ROC trace le TPR (rappel) en fonction du FPR). L'AUC est l'aire (surface) sous cette courbe. L'AUC-ROC mesure la capacité du modèle à distinguer entre les classes positives et négatives.

2.2.4.7 Courbe PR (Précision-Rappel)

La courbe PR trace la précision en fonction de rappel pour chaque seuil de décision comme la courbe ROC, L'aire sous cette courbe (appelé AUC-PR) est la performance globale pour la classe positive, donc elle est adaptée pour le cas des données déséquilibrées.

Dans le domaine médical, le choix des métriques d'évaluation dépend de l'équilibre des données pour ne pas ignorer les classes minoritaires. Les données médicales sont souvent déséquilibrées et dans ces situations, les métriques comme la précision, rappel, f1-score et courbe PR sont utilisés parce qu'ils donnent une mesure fiable de la performance du modèle.

2.3 Machine Learning dans la détection et la prédiction des maladies infectieuses

L'application du Machine Learning en domaine de la médecine spécialement en détection et prédiction des maladies infectieuses offre des perspectives prometteuses et permet d'une exploration approfondie des mégadonnées médicales. Les algorithmes de Machine Learning jouent un rôle important dans la détection précoce et la prédiction de ces maladies. Son intégration ouvre la voie vers une médecine plus préventive et fondée sur les données.

Plusieurs études dans la littérature ont été consacrés à la détection et la prédiction des maladies infectieuses en utilisant les techniques du Machine Learning en explorant différent approches algorithmiques et méthodologies, voici quelques travaux qui ont contribué à ce domaine :

Tableau 1. Revue des contributions liée à l'application du Machine Learning aux maladies infectieuses.

Référence	Maladie(s)	Dataset utilisé	Technique(s) de ML utilisée(s)	Mesures de performance	Meilleur résultats
[16]	Hépatite Dengue	Obtenu depuis le dépôt de l'Université de Californie à Irvine (155 instances et 20 attributs). Données collectées via des questionnaires et auprès des centres médicaux à Chennai (340 instances et 28 attributs).	LR, RF, DT, MLP, C4.5. DT, ANN, MSO-MLP, PSO-ANN, PSO.	TPR, FPR, Précision, Rappel, ROC, Exactitude.	RF (TPR=0.9245, Acc.=90.32%) MSO-MLP (TPR= 0.865, Acc.=85.18%)

[17]	Tuberculose	The tuberculosis gene expression dataset (composé de 48 803 gènes pour 498 instances)	weighted voting ensemble technique combinant SVM et NB	Exactitude, Spécificité, Sensitivité	Acc.= 95%
[18]	Dengue	Jeu de données clinique personnalisé (400 instances et 12 attributs)	KNN, SVM, RF, DT, NB, LR, ANN	Exactitude	NB & DT (Acc.=100%)
[19]	COVID-19	Données cliniques de l'hôpital Sírio-Libanês (Brésil) (1925 instances et 231 attributs)	LR, DT, RF, SVM, KNN, NB, XGBoost, ExtraTrees, AdaBoost, LightGBM, CatBoost, 1-DCNN.	Précision, Rappel, AUC, Exactitude, F1-score	RF (Rec.=83%)
[20]	AIDS/HIV	Les données ont été collectées depuis Data World.	SVM, RF, NB, GRU, LSTM	Exactitude, précision, rappel, et F1-score	LSTM & GRU (Acc. = 97.65% et 96.00%, Prec. = 77.35% et 84.00%, Rec. = 87.93% et 82.98%, F1-score=82.03% et 83.20%, respectivement.
[21]	Hépatite C	HCV dataset de UCI (1756 instances et 29 attributs.)	SVM, MARS, RF, DT, BGLM, HPM	Précision, Rappel, Exactitude,	HPM (Acc. = 96,82 %)

			model (IRF, SVM)	F-measure	
[22]	Dengue	Données du compétition DengAI	KNN, DT, RF, SVM, Gaussian Neighbor Boundaries	Moyenne d'exactitudes.	RF (Acc. mean= 8.72)
[23]	Hépatite C	HCVdata de UCI (615 instances)	LR, SVM, KNN, DT, RF, AdaBoost	Précision, Rappel, Exactitude, F1-score, AUC	AdaBoost (Acc. =97.8% AUC=0.994)
[24]	Paludisme	historical meteorological and clinical datasets	C5.0, DT, ANN, KNN, SVM, RL, XGBoost, RF	Exactitude, Rappel, Spécificité, AUC	XGBoost & DT (Spec. = 93.3%)
[25]	Tuberculose	Clinical data collected at the Lung Poly of RSUD Prof. Dr W Z Johannes Kupang	MLP, ELM	Exactitude	MLP (Acc. = 95%)
[26]	COVID-19	Kaggle – "Diagnosis of COVID-19 and its clinical spectrum"	MLP, RF, XGBoost	Exactitude, rappel, ROC_AUC	MLP (Acc. = 98.17 %)
[27]	AIDS/HIV	Le dataset utilisé provient de Kaggle (2139 instances et 23 attributs)	RF, ET, XGBoost, LightGBM, DT, GradientBoosting, SVM, Adaboost, LR	Exactitude, précision, rappel, F1-score, AUC-ROC	ET (AUC = 0.99 F1-Score = 0.94)

2.4 Conclusion

Dans ce chapitre, Nous avons présenté les concepts essentiels de Machine Learning notamment ses différents types d'apprentissage et ses algorithmes et les métriques d'évaluation. Ensuite, nous avons montré son rôle dans la détection précoce et la prédiction des maladies infectieuses et présenté plusieurs études antérieures dans ce domaine.

Chapitre 3 : Expérimentations et résultats

3.1 Introduction

De nos jours, les systèmes de santé basés sur le Machine Learning deviennent des outils essentiels pour la détection précoce et la prédiction des maladies infectieuses. Plusieurs étapes préliminaires sont indispensables pour construire un modèle de classification pour la détection de ces maladies, notamment le prétraitement des données pour qu'elles soient exploitables par les algorithmes de Machine Learning et la sélection des attributs les plus pertinents pour une prédiction finale fiable. Le déséquilibre et le choix optimal des hyperparamètres restent des défis lors de la construction du modèle et impactent fortement ses performances.

Dans ce chapitre, nous présentons le processus expérimental mis en œuvre dans notre projet. Nous commençons par décrire l'environnement logiciel utilisé. Après, nous présentons les trois datasets utilisés, leurs descriptions, le prétraitement effectué sur chaque ensemble de données, les différentes expérimentations menées sur chaque dataset, ainsi que la discussion des résultats obtenus.

3.2 Environnement logiciel

Notre approche a été développée en Python en s'appuyant sur l'environnement Google Colab.

3.2.1 Python

Python est l'un des langages de programmation les plus simples et les plus utiles et est largement utilisé dans l'industrie du logiciel. Il offre une grande flexibilité et une facilité d'utilisation. Python est largement utilisé en création de sites web, le développement de logiciels, la Science des données et l'apprentissage automatique [28].

Plusieurs bibliothèques du langage Python ont été utilisé dans notre travail, notamment :

- **Pandas** *:* C'est une bibliothèque Python qui est utilisée pour la manipulation et l'analyse de données. Elle est essentielle pour la préparation et l'analyse des données, grâce à ses structures de données telle que les DataFrames qui facilitent la manipulation et le traitement des données structurées.

- **NumPy :** C'est une bibliothèque qui permet de manipuler tableaux et matrices multi-dimensionnels efficacement, en offrant des outils permettant d'effectuer des calculs mathématiques et statistiques sur ces tableaux.
- **Matplotlib :** C'est une bibliothèque Python qui permet de créer des visualisations statiques, animées ou interactives en Python. Elle est couramment utilisée pour représenter graphiquement les jeux de données et les résultats de performance des modèles.
- **Seaborn :** C'est une bibliothèque Python basée sur Matplotlib. Seaborn offre une interface de haut niveau pour dessiner des graphiques statistiques.
- **Scikit-learn :** C'est la bibliothèque Python à utiliser pour l'apprentissage automatique. Elle offre la plupart des algorithmes d'apprentissage supervisé (SVM, KNN…) et non supervisé (Comme K-means) et peut être utilisé aussi pour l'exploration et l'analyse de données. Ce qui en fait au final un outil idéal pour les débutants en Machine Learning.

3.2.2 Google Colab

Google Colab (abréviation de Colaboratory) est une plateforme en ligne gratuite fournie par Google, permettant d'écrire et d'exécuter du code Python dans un environnement Jupyter Notebook en offrant un accès facile à des ressources GPU/TPU. Cela permet de réaliser des projets de data science et d'apprentissage automatique directement en ligne, sans nécessiter d'une installation locale.

3.3 Datasets utilisés

3.3.1 Dataset AIDS/HIV

3.3.1.1 Description du dataset

Ce dataset contient des Informations personnelles, des antécédents médicaux, Historique de traitement et des Résultats de laboratoire sur des patients infectés et non infectés avec le SIDA [29]. Il représente une collection de données médicales de 2139 patients et ne contient pas de valeurs manquantes. La variable cible ''infected'' est binaire, où 0 désigne les individus non infectés et 1 les individus infectés par le VIH. Parmi eux, 521 individus ont été infectés avec le SIDA, tandis que 1618 autres n'en ont pas. La Figure 13 présente la répartition des individus selon le sexe (0 = femme, 1 = homme) et leur cas d'infection avec une majorité d'hommes parmi les personnes non infectées, mais aussi une présence importante d'hommes parmi les

personnes infectées. La Figure 14 illustre La distribution globale de l'âge dans l'ensemble de la population étudiée, On observe une concentration maximale des individus entre 25 et 35 ans, avec un pic légèrement au-dessus de 100. La Figure 15 montre la distribution de la variable cible ''infected'', montrant un déséquilibre entre les classes : 75.64 % des individus ne sont pas infectés, contre 24.36 % infectés.

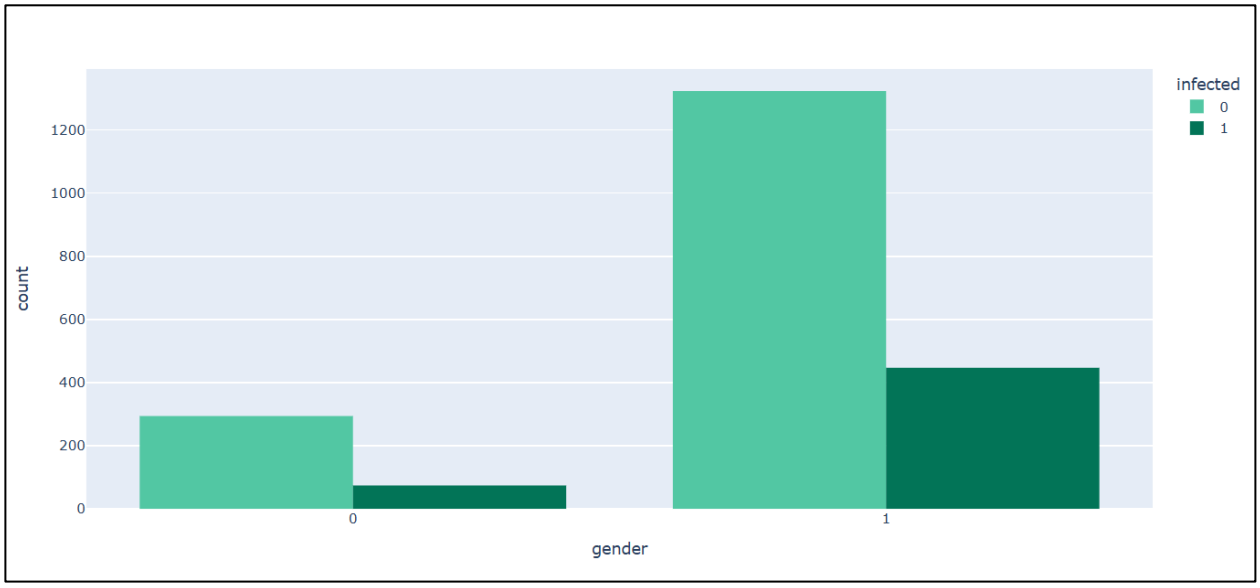

Figure 13. La répartition des individus selon le sexe (0 = femme, 1 = homme) et leur cas d'infection.

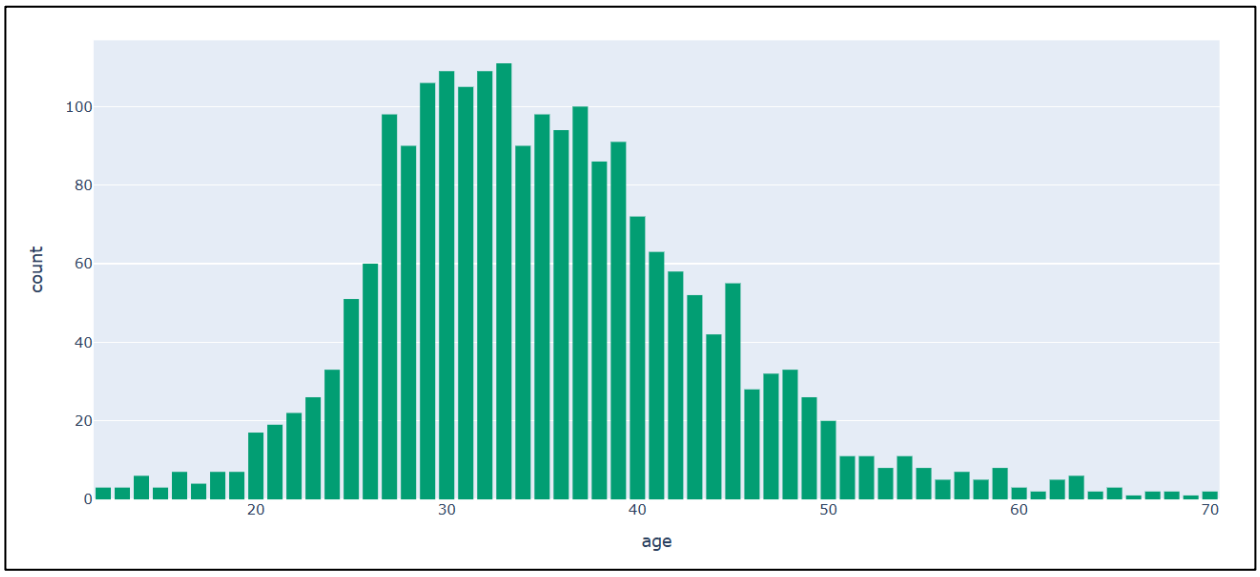

Figure 14. La distribution globale de l'âge dans l'ensemble de la population étudiée.

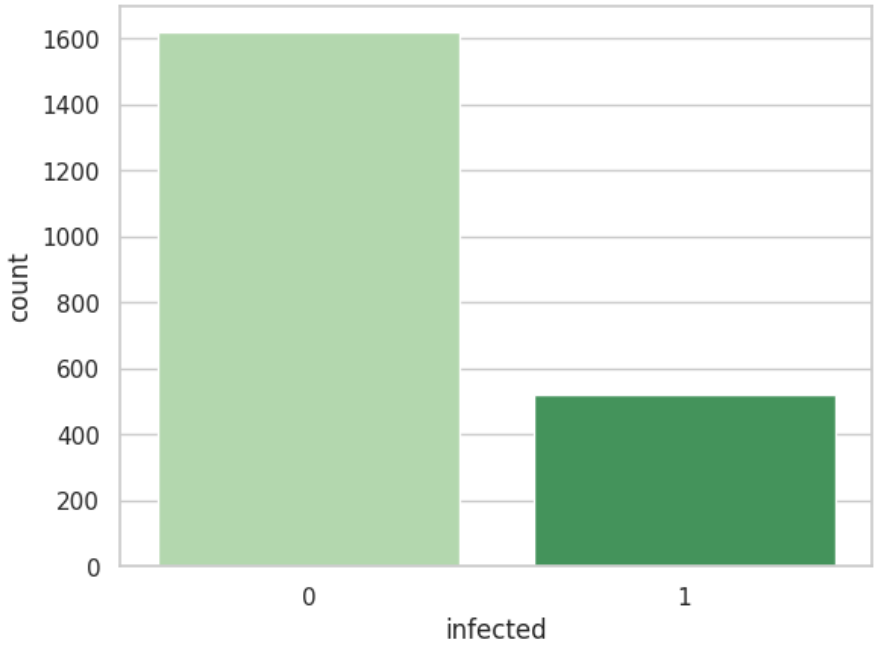

Figure 15. La distribution de la variable cible ‘’infected”.

Table 1. Description des variables du dataset AIDS/HIV [29].

Variable	**Description**
Age	Âge au début de l’étude (en années).
Wtkg	Poids au début de l’étude (en kg).
homo	Activité homosexuelle (0 = non, 1 = oui).
Race	Race (0 = blanc, 1 = non-blanc).
gender	Sexe (0 = femme, 1 = homme).
Hemo	Hémophilie (0 = non, 1 = oui).
drugs	Usage antérieur de drogues injectables (0 = non, 1 = oui).

Trt	Indicateur de traitement (0 = ZDV seul, 1 = ZDV + ddI, 2 = ZDV + Zal, 3 = ddI seul).
oprior	Traitement antirétroviral autre que ZDV avant jour 175 (0 = non, 1 = oui).
z30	ZDV dans les 30 jours précédant le jour 175 (0 = non, 1 = oui).
preanti	Jours de traitement antirétroviral avant le jour 175.
str2	Historique de traitement antirétroviral (0 = naïf, 1 = expérimenté).
Strat	Stratification du traitement (1 = naïf, 2 = 1-52 semaines, 3 = >52 semaines).
Treat	Indicateur de traitement (0 = ZDV seul, 1 = autre).
offtrt	Arrêt du traitement avant 96 ± 5 semaines (0 = non, 1 = oui).
cd40	CD4 au début de l'étude.
cd420	CD4 à 20 ± 5 semaines.
cd80	CD8 au début de l'étude.
cd820	CD8 à 20 ± 5 semaines.
Symptom	Présence de symptômes (0 = asymptomatique, 1 = symptomatique).
karnof	Score de Karnofsky (sur une échelle de 0 à 100).
Time	Durée avant échec ou censure.
infected	Est infecté par le SIDA (0 = non, 1 = oui).

3.3.1.2 Prétraitement des données

Le prétraitement des données est une étape importante pour rendre les données exploitables par les algorithmes de Machine Learning. Cette étape inclut la gestion des valeurs manquantes, le codage des variables catégorielles et la mise à l'échelle des données numériques.

- Les dix premières lignes de dataset AIDS/HIV ont été affichées pour avoir une vue brève sur le dataset.

	time	trt	age	wtkg	hemo	homo	drugs	karnof	oprior	z30	...	str2	strat	symptom	treat	offtrt	cd40	cd420	cd80	cd820	infected
0	948	2	48	89.8128	0	0	0	100	0	0	...	0	1	0	1	0	422	477	566	324	0
1	1002	3	61	49.4424	0	0	0	90	0	1	...	1	3	0	1	0	162	218	392	564	1
2	961	3	45	88.4520	0	1	1	90	0	1	...	1	3	0	1	1	326	274	2063	1893	0
3	1166	3	47	85.2768	0	1	0	100	0	1	...	1	3	0	1	0	287	394	1590	966	0
4	1090	0	43	66.6792	0	1	0	100	0	1	...	1	3	0	0	0	504	353	870	782	0
5	1181	1	46	88.9056	0	1	1	100	0	1	...	1	3	0	1	0	235	339	860	1060	0
6	794	0	31	73.0296	0	1	0	100	0	1	...	1	3	0	0	0	244	225	708	699	1
7	957	0	41	66.2256	0	1	1	100	0	1	...	1	3	0	0	0	401	366	889	720	0
8	198	3	40	82.5552	0	1	0	90	0	1	...	1	3	1	1	1	214	107	652	131	1
9	188	0	35	78.0192	0	1	0	100	0	1	...	1	3	0	0	1	221	132	221	759	1

Figure 16. Dix premières lignes du dataset AIDS/HIV.

- Ensuite, on a affiché les statistiques pour chaque colonne du dataset pour obtenir un aperçu statistique des variables numériques. En consultant les statistiques de chaque colonne, nous pouvons déduire que : les valeurs existantes sont dans le rang réalistique.
- Nous avons aussi affiché la matrice de corrélation sous forme de carte thermique (heatmap), qui est un outil qui permet de visualiser la relation entre les variables dans un dataset.

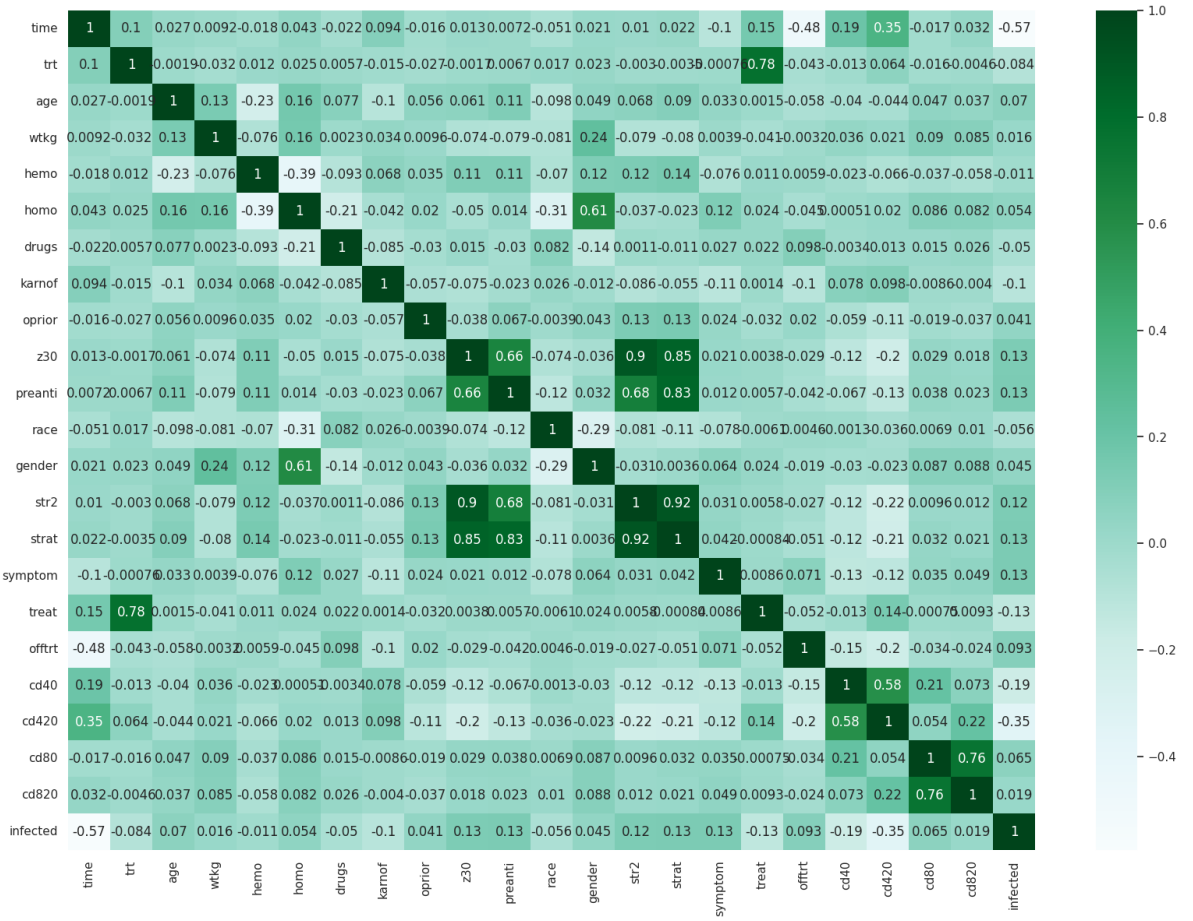

Figure 17. Matrice de corrélation de dataset AIDS/HIV.

Les principales observations à partir du heatmap présenté dans la figure 17 sont : la variable cd420 (r = -0.35) montre une corrélation négative modérée, indiquant qu'une baisse dans cette valeur peut être liée à une infection avec le SIDA, ce qui est cohérent avec le rôle des cellules CD4 dans la réponse immunitaire. La variable time est la variable la plus fortement corrélée négativement avec la classe infected (r = -0.57). Cela signifie que plus cette durée est longue, moins il y a de chances que la personne soit infectée, ce qui peut refléter l'efficacité du traitement ou une meilleure résistance de l'organisme. Par contre, plusieurs variables présentent une corrélation légèrement positive avec la variable cible infected qui sont : str2, treat, symptom, z30, preanti, strat et ont toutes des corrélations autour de r ≈ +0.13. Les autres variables, comme : age, gender, race, wtkg, hemo et drugs, présentent des corrélations très faibles avec la variable cible ($|r| < 0.1$), ce qui montre qu'elles ont peu d'influence directe sur la classe infected.

- Division des données : Les données prétraitées sont divisées en deux parties : une partie d'entraînement sur laquelle le modèle est entrainé et qui représente 80% de l'ensemble de données, et une deuxième partie sur laquelle le modèle est testé et représente un pourcentage de 20% des données.
- Standardisation des données : Cette étape consiste à transformer les variables numériques pour qu'elles aient une moyenne de 0 et un écart-type de 1, ce qui peut aider à améliorer les performances de certains algorithmes de Machine Learning qui sont sensibles à la variance des données.

3.3.1.3 Classification et évaluation des résultats

Dans la première expérimentation, nous avons utilisé des modèles classiques simples notamment : SVM, LR, RF, AdaBoost, KNN, DT avec le paramètre class_weight= balanced. Le tableau 2 et la figure 18 présentent les résultats des performances de ces modèles et une visualisation des résultats obtenus selon chaque métrique d'évaluation :

Tableau 2. Les performances des modèles classiques pour le dataset AIDS/HIV.

Modèle	Exactitude(%)	Précision (%)	Rappel (%)	F1-score (%)	ROC-AUC (%)
SVM	89	86	81	84	90
LR	84	78	81	79	88
RF	89	87	83	85	92
AdaBoost	87	82	82	82	92
KNN	81	76	70	72	80
DT	82	76	77	77	77

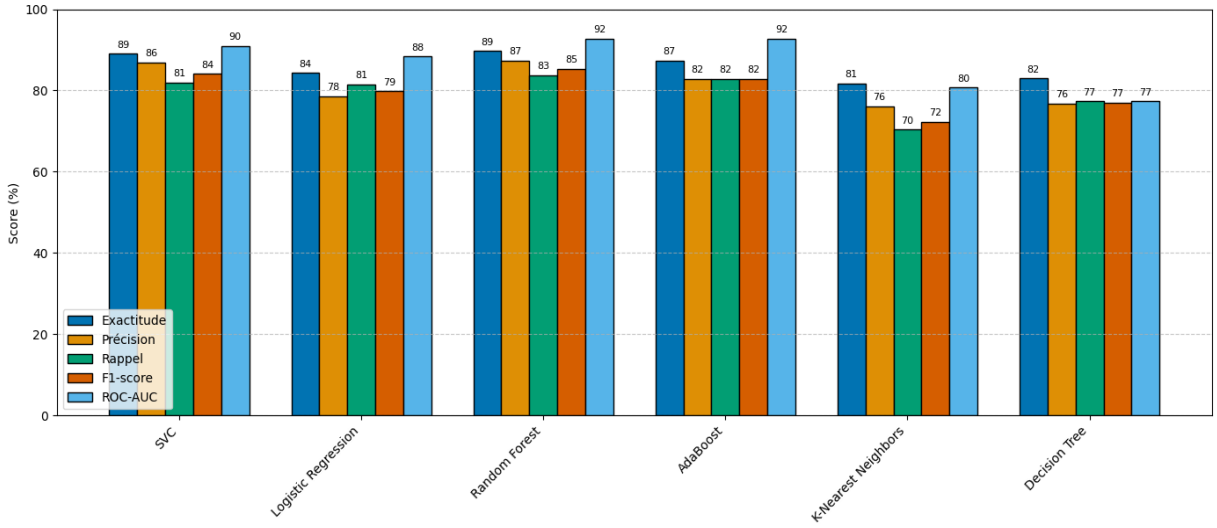

Figure 18. Visualisation des performances des modèles classiques pour le dataset AIDS/HIV.

Face à ces résultats relativement faibles et pour améliorer les résultats, nous avons exploré dans une deuxième expérimentation des modèles plus puissants notamment : GBC, lightGBM, CatBoost et LR, avec une optimisation des hyperparamètres de chaque modèle avec Optuna qui est une bibliothèque d'optimisation automatique basée sur l'optimisation bayésienne, et une sélection des attributs pour identifier les variables les plus pertinentes avec la méthode SFS (Sequential Forward Selection), qui est une méthode de sélection des attributs de type enveloppement (wrapper). Elle a retenu 19 des 21 attributs initiaux ce qui montre que la plupart des attributs étaient pertinents. Ces modèles avec les hyperparamètres optimaux sont intégrés comme des estimateurs de base dans un classifieur d'ensemble Stacking avec RF comme meta-learner. Le tableau 3 montre les hyperparamètres optimaux obtenus avec optuna pour chaque estimateur de base. Le tableau 4 présente les résultats de performances du Stacking et la figure 19 montre une visualisation de ces résultats obtenus :

Tableau 3. Les hyperparamètres optimaux obtenus par optuna pour chaque modèle.

Modèle	Hyperparamètres optimales
GBC	n_estimators= 143, max_depth= 10, learning_rate=0.18011892661817977, min_samples_split= 3, min_samples_leaf= 3, subsample=0.5992364960517184.
LightGBM	n_estimators= 151, max_depth=14, num_leaves= 40, learning_rate=0.018483372741887622, min_child_samples= 94, subsample=0.6411519350602914, colsample_bytree=0.6527861626053972, scale_pos_weight= 8.306211216650125.
CatBoost	Iterations= 298, Depth=5, learning_rate=0.012242513412851859, l2_leaf_reg=3.923271928869185, border_count=128, scale_pos_weight= 9.288621032584121.
LR	C=0.09142669797128761, Solver= 'lbfgs'.

Tableau 4. Les performances du modèle du Stacking avec SFS pour le dataset AIDS/HIV.

Modèle	Exactitude(%)	Précision(%)	Rappel(%)	F1-score(%)	ROC-AUC(%)
Stacking	91	90	85	88	94

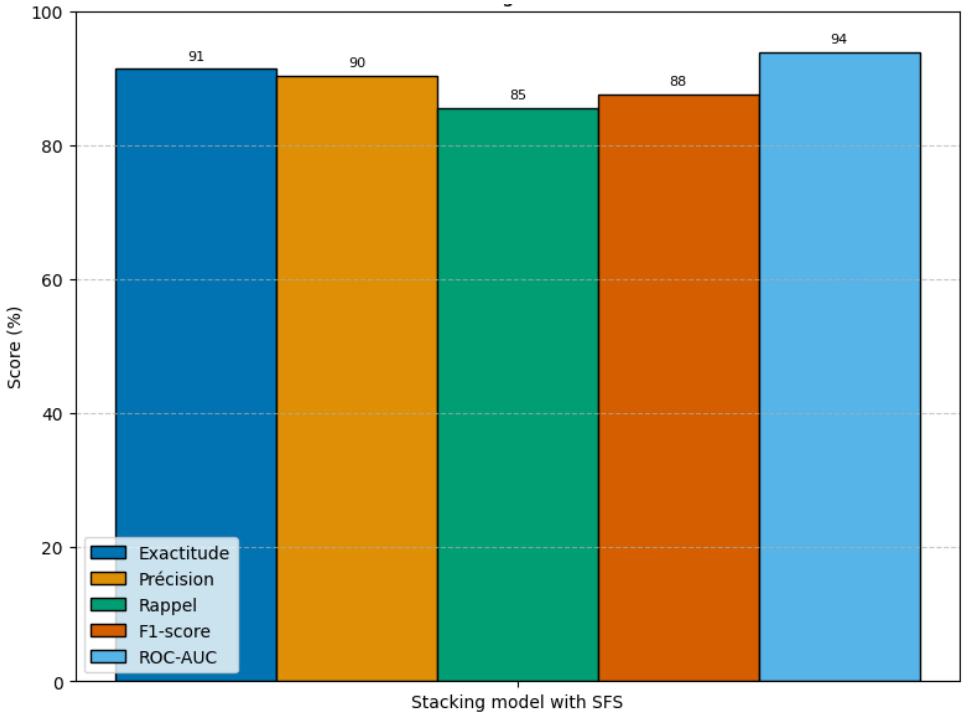

Figure 19. Visualisation des performances du modèle de Stacking avec SFS pour le dataset AIDS/HIV.

Ces résultats montrent que la méthode d'ensemble Stacking de ces modèles a légèrement amélioré les performances par rapport aux modèles classiques simples, ce qui confirme l'intérêt de cette approche.

Bien que l'étude [27] qui a utilisé le même dataset a obtenu des résultats élevés en termes de toutes les métriques d'évaluation (Acc. =98 %, Prec. =93 %, Rec.=96 %, F1-score=95 %, Roc-auc=99 %) en utilisant les méthodes de sur-échantillonnage : SMOTE, ADASYN et Random Over Sampling, il est important de noter que l'application de sur-échantillonnage surtout en domaine de la médecine ne préserve pas l'intégrité clinique des données car il génère des données synthétiques et donc les résultats seront gonflés et peu représentatifs du monde réel.

Néanmoins, notre méthode est plus fiable et mieux adaptée pour les applications réelles médicales et préserve l'intégrité clinique des données et a atteint des résultats solides et prometteurs.

3.3.2 Dataset Hépatite C

3.3.2.1 Description du dataset

Ce dataset contient des informations personnelles ainsi que des résultats de laboratoire sur des personnes infectées et non infectées avec l'hépatite C [30]. Il représente une collection de

données médicales de 615 patients et 14 variables. La variable cible pour la classification est Category (Blood donors, suspect blood donors et Hepatitis C, y compris sa progression : Hepatitis C, Fibrosis, Cirrhosis). La Figure montre la distribution de la variable cible ‘’Category’’, montrant un grand déséquilibre entre les classes.

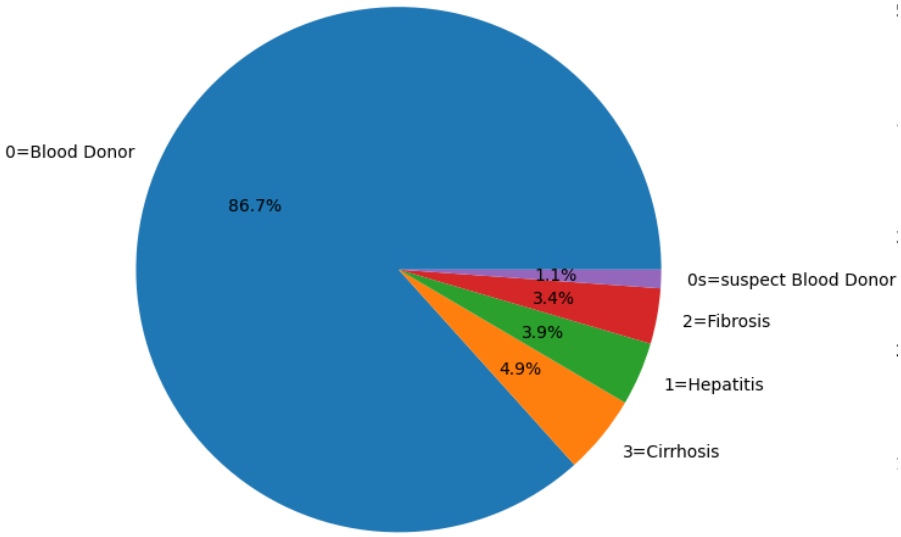

Figure 20. La distribution de la variable cible ‘’Category’’.

Tableau 5. Description des variables du dataset Hépatite C [30].

Variable	Description	Valeurs manquantes ?
Unnamed: 0	L'identifiant du patient.	Non
Age	L'âge de l'individu associé à l'observation.	Non
Sex	Variable binaire indiquant le sexe de l'individu	Non
ALB	La valeur associée à l'albumine.	Oui (1)
ALP	La valeur associée à l'alcaline phosphatase.	Oui (18)
AST	La valeur associée à l'aspartate aminotransférase.	Non
BIL	La valeur associée à la bilirubine.	Non
CHE	La valeur associée au cholinestérase.	Non
CHOL	La valeur associée au cholestérol.	Oui (10)
CREA	La valeur associée à la créatine.	Non
GGT	La valeur associée à la gamma-glutamyl transférase.	Non
PROT	La valeur associée à la protéine totale.	Oui (1)
ALT	La valeur associée à l'alanine aminotransférase.	Oui (1)
Category	Variable cible qui représente le diagnostic de l'individu (valeurs : '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis').	Non

3.3.2.2 Prétraitement des données

Les étapes effectuées de prétraitement pour ce jeu de données sont comme suit :

- Les dix premières lignes de dataset Hépatite C ont été affichées pour avoir une vue brève sur le dataset.

	Unnamed: 0	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
0	1	0=Blood Donor	32	m	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106.0	12.1	69.0
1	2	0=Blood Donor	32	m	38.5	70.3	18.0	24.7	3.9	11.17	4.80	74.0	15.6	76.5
2	3	0=Blood Donor	32	m	46.9	74.7	36.2	52.6	6.1	8.84	5.20	86.0	33.2	79.3
3	4	0=Blood Donor	32	m	43.2	52.0	30.6	22.6	18.9	7.33	4.74	80.0	33.8	75.7
4	5	0=Blood Donor	32	m	39.2	74.1	32.6	24.8	9.6	9.15	4.32	76.0	29.9	68.7
5	6	0=Blood Donor	32	m	41.6	43.3	18.5	19.7	12.3	9.92	6.05	111.0	91.0	74.0
6	7	0=Blood Donor	32	m	46.3	41.3	17.5	17.8	8.5	7.01	4.79	70.0	16.9	74.5
7	8	0=Blood Donor	32	m	42.2	41.9	35.8	31.1	16.1	5.82	4.60	109.0	21.5	67.1
8	9	0=Blood Donor	32	m	50.9	65.5	23.2	21.2	6.9	8.69	4.10	83.0	13.7	71.3
9	10	0=Blood Donor	32	m	42.4	86.3	20.3	20.0	35.2	5.46	4.45	81.0	15.9	69.9

Figure 21. Les dix premières lignes du dataset Hépatite C.

- Suppression de la colonne d'index « Unnamed : 0 » inutile.
- Suppression des lignes où la catégorie est "0s=suspect Blood Donor" : Cette classe ne contient que 7 échantillons sur un total de 615, ce qui représente 1,1 % du dataset et elle peut introduire un déséquilibre extrême.
- Encodage des variables ''Category'' et ''Sex'' : La transformation des valeurs textuelles en étiquettes numériques avec un mappage manuel :
    - 0 : Blood donor
    - 1 : Hepatitis
    - 2 : Fibrosis
    - 3 : Cirrhosis
    - m → 0
    - f → 1
- Imputation des valeurs manquantes : Les valeurs manquantes dans les colonnes numériques sont remplacées par la médiane de chaque colonne.

3.3.2.3 Classification et évaluation des résultats

Dans un premier temps, nous avons appliquées une approche classique de classification. Nous avons utilisé plusieurs algorithmes de classification avec le paramètre class_weight= balanced à cause du déséquilibre élevé des classes notamment :

- Régression logistique
- Random Forest
- SVM
- Decision Tree
- XGBoost

Après avoir entrainé ces modèles et évalué leur performance en utilisant les métriques d'évaluation : Exactitude, précision, rappel et f1-score, nous avons obtenu les résultats suivants présenté dans le tableau 6 :

Tableau 6. Les performances des modèles classiques pour le dataset Hépatite C.

Modèle	**Exactitude (%)**	**Précision (%)**	**Rappel (%)**	**F1-score (%)**
Régression logistique	94	74	84	78
Random Forest	93	82	57	62
SVM	93	70	68	67
Decision Tree	93	53	52	53
XGBoost	95	75	68	71

La figure 22 montre une visualisation des résultats obtenus par ces différents modèles selon chaque métrique d'évaluation :

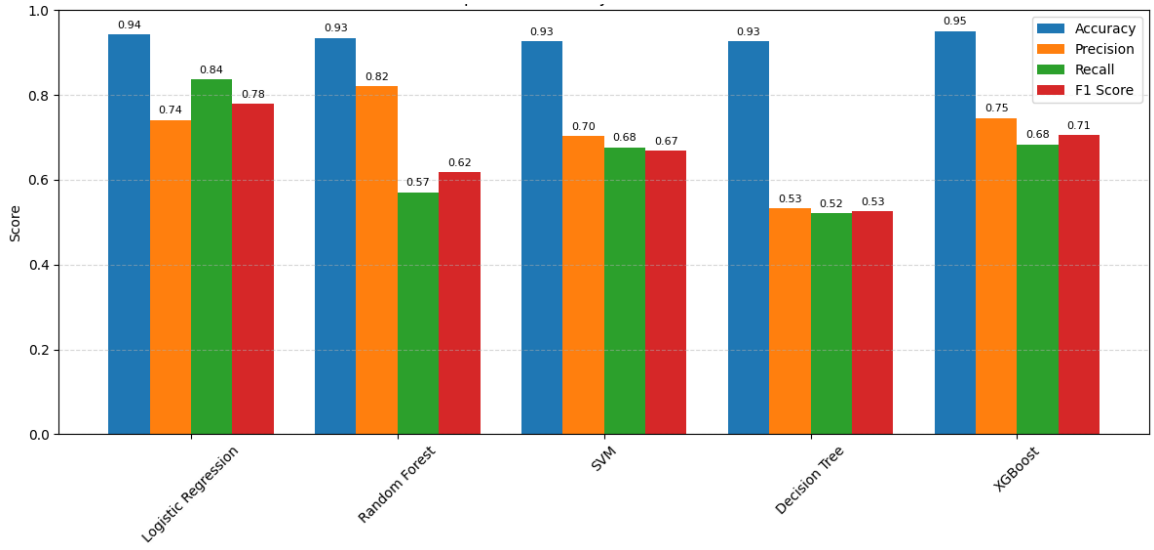

Figure 22. Visualisation des performances des modèles classiques pour le dataset Hépatite C.

Selon ces résultats, le classifieur XGBoost a eu la meilleure performance parmi tous les classifieurs, atteignant une exactitude de 95% et un f1-score de 71%, tandis que le classifieur DT présente la performance la plus faible avec une exactitude de 93% et f1-score de 53%.

Globalement, ces modèles classiques ont atteint des bonnes performances en termes d'exactitude mais on observe des performances faibles en termes de précision, rappel et f1-score qui peuvent s'exprimer par le fort déséquilibre des classes malgré l'utilisation de paramètre class_weight=balanced.

Les résultats obtenus avec ces modèles classiques montrent quelques limites notamment la confusion entre les classes minoritaires mal présentées à cause du déséquilibre élevé dans ce dataset. Pour surmonter ce problème, nous avons proposé une méthode multistage (2 stages) qui est inspiré du raisonnement d'un expert de santé. Cette approche proposée vise à décomposer le problème en 2 étapes : La premiére étape (ou stage) vise à prédire si un individu est dans un état sain ou malsain (Healthy ou unhealthy), après les cas malsains sont passé au 2éme stage pour prédire la progression (ou le stade) de la maladie de l'hépatite C chez ces individus malsains si elle est dans un stade précoce ou avancé (Early stage ou Advanced stage). La figure 23 montre le schéma de cette méthode proposée :

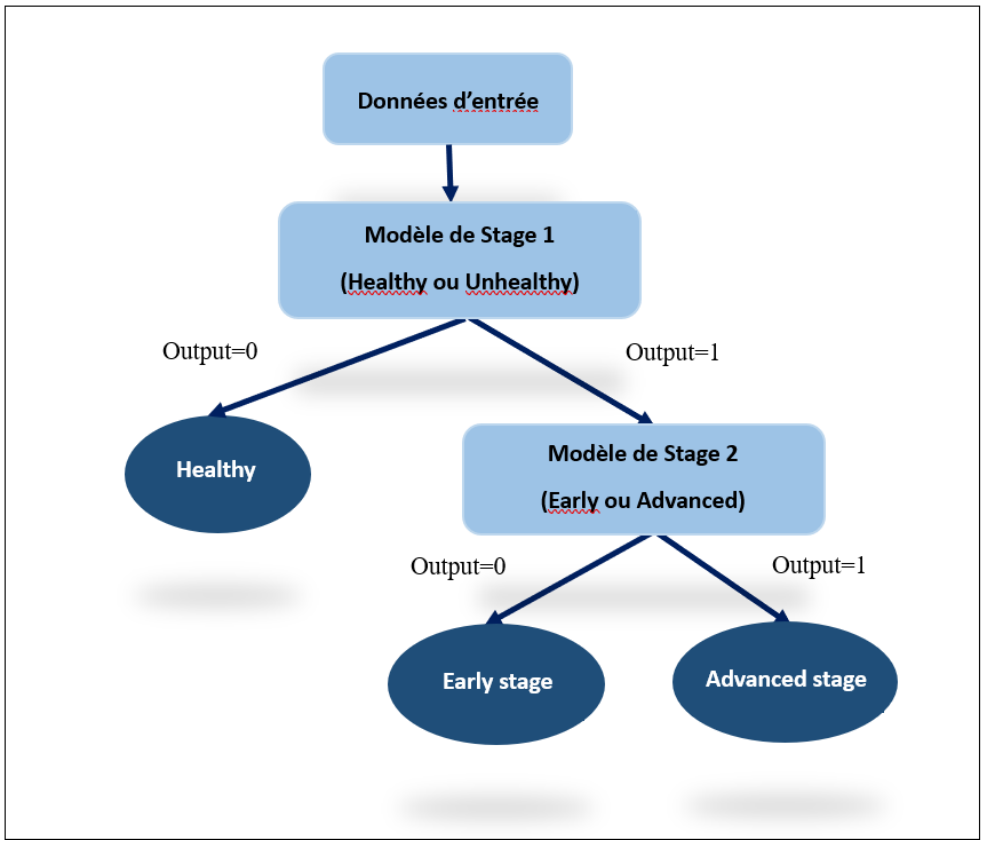

Figure 23. Le schéma de la méthode Multistage proposée.

Nous avons ajouté 3 variables cibles pour cette approche : Stage1_Target, Stage2_Target et 3classes_Target en se basant sur la variable cible originale Category :

```
df['Stage1_Target'] = df['Category'].apply(lambda x: 0 if x == 0 else 1)
df['Stage2_Target'] = df['Category'].apply(lambda x: np.nan if x == 0 else (0 if x == 1 else 1))
df['3classes_Target']=df['Category'].apply(lambda x: np.nan if x==0 else (1 if x==1 else (2 if x==2 else 3)))
```

Figure 24. Code python appliqué pour ajouter les nouvelles variables cibles.

Après nous avons divisé l'ensemble de données en deux parties (entrainement et test) avec la méthode train test split selon un ratio de 80/20 en utilisant le paramètre stratify=df [Category] pour préserver la proportion des classes dans les ensembles d'entrainement et de test.

- Pour le premier stage : Les mêmes modèles classiques mentionnés ci-dessus ont été entrainé et évalué sur l'ensemble d'entraînement en utilisant la technique de validation croisé StratifiedKFold (k=5). Après, le meilleur modèle qui a obtenu le meilleur f1-score est XGBoost et a été choisi et sauvegardé avec joblib et enfin été

évalué sur l'ensemble de test pour une évaluation finale. Le tableau 7 et la figure 25 présentent les performances ainsi que la matrice de confusion du meilleur modèle sélectionné pour le premier stage :

Tableau 7. Performance du modèle sélectionné pour le stage 1.

Modèle	Exactitude (%)	Précision (%)	Rappel (%)	F1-score (%)
XGBoost	98	99	93	96

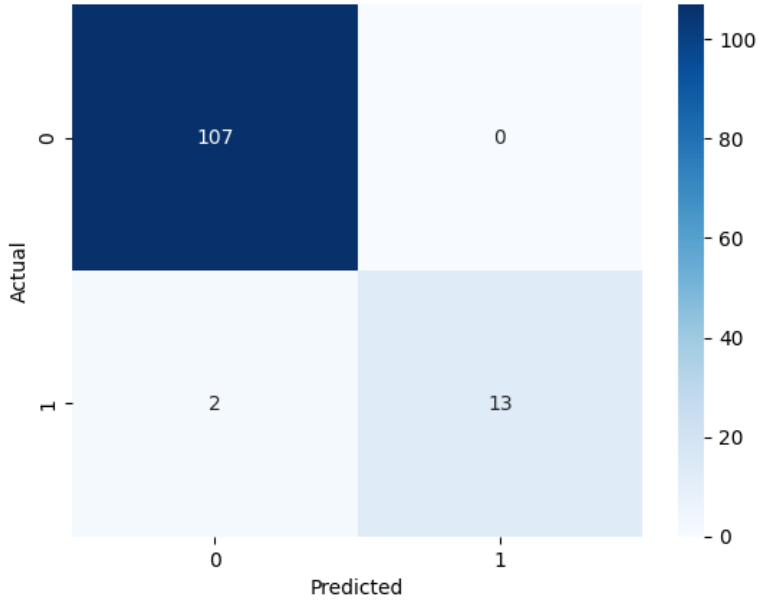

Figure 25. Matrice de confusion du modèle séléctionné pour le stage 1.

Les observations prédites comme malsains (Unhealthy) et qui présentent les vrais positifs (TP) sont passés au 2éme stage pour prédire le stade de l'hépatite C.

Dans un premier temps, nous avons expérimenté la prédiction des trois stades de cette maladie y compris : hepatitis, fibrosis et cirrhosis en utilisant la variable cible ajouté '3classes_Target', mais les résultats obtenus étaient insuffisants à cause de la confusion entre ces classes et aussi à cause de petit nombre d'échantillons par classe. Le tableau 8 et la figure 26 présentent les performances de cette expérimentation :

Tableau 8. Performance du modèle sélectionné pour le stage 2 à 3 classes.

Modèle	Exactitude (%)	Précision (%)	Rappel (%)	F1-score (%)
SVM	80	82	78	80

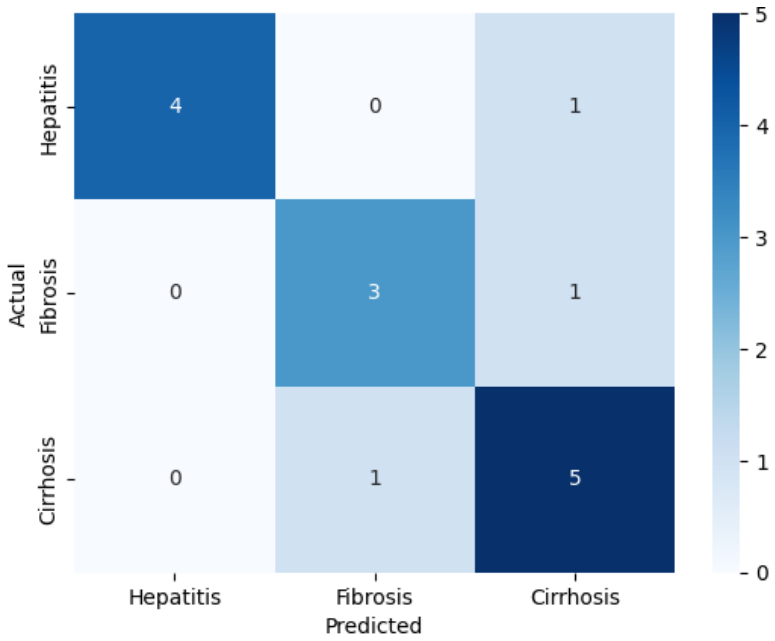

Figure 26. Matrice de confusion du modèle sélectionné pour le stage 2 à 3 classes.

La limite des résultats obtenus ci-dessus, nous a conduit à reformuler le 2éme stage en classification binaire ce qui a permis d'améliorer la stabilité du modèle et les résultats des performances :

- Early stage : Hepatitis.
- Advanced stage : en groupant Fibrosis et Cirrhosis.

Le tableau 9 montre les performances de cette expérimentation, tandis que les figures 27 et 28 présentent la matrice de confusion ainsi qu'une comparaison des performances entre les approches à trois classes (hepatitis, fibrosis et cirrhosis) et à deux classes (early et advanced) :

Tableau 9. Performance du modèle séléctionné pour le stage 2 à 2 classes

Modèle	Accuracy (%)	Précision (%)	Rappel (%)	F1-score (%)
SVM	93	95	90	92

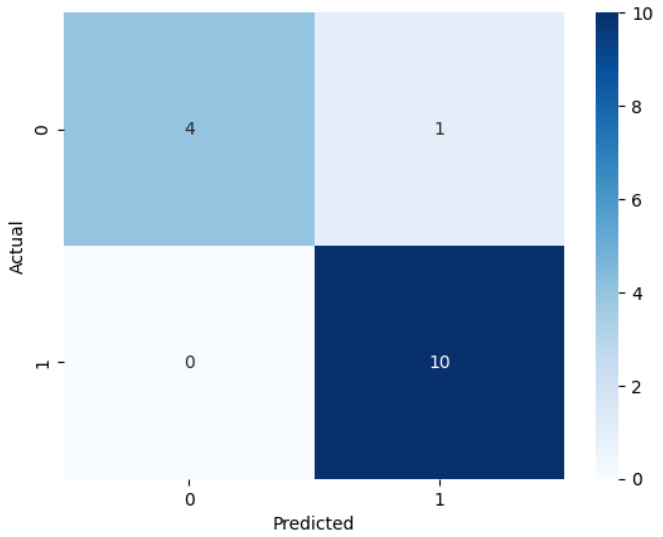

Figure 27. Matrice de confusion du modèle séléctionné pour le stage 2 à 2 classes.

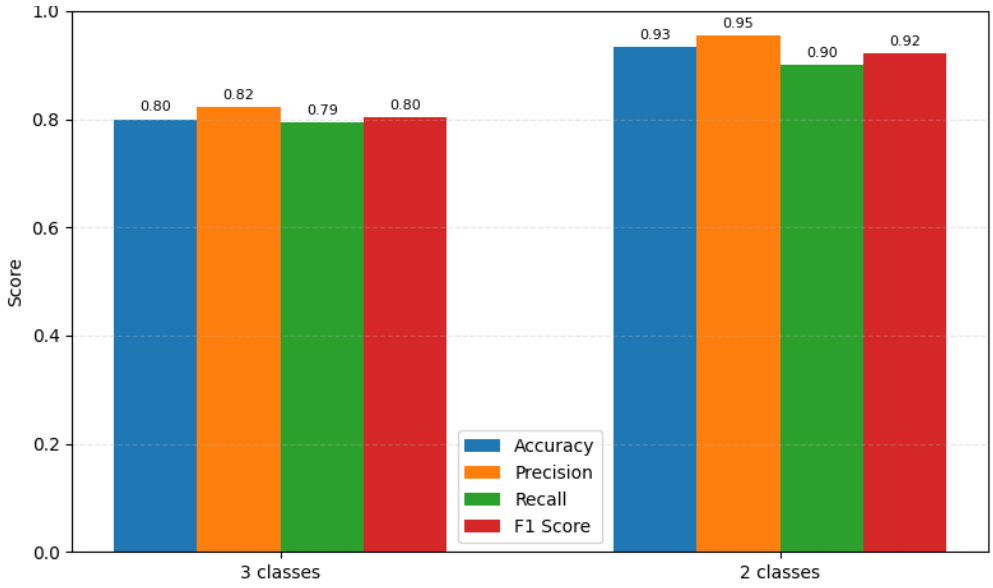

Figure 28. Comparaison entre les deux expérimentations du stage 2 (3 classes et 2 classes).

Pour le 2éme stage, les modèles ont été entrainé sur les cas malsains seulement, et le meilleur modèle a été choisi et sauvegardé par joblib de la même manière que le premier stage, mais en ajoutant une optimisation de seuil de classification (Threshold) pour maximiser le plus les performances du modèle surtout le f1-score avec des variations de seuil entre 0.1 et 0.9. Le meilleur seuil a été estimé par 0.41 avec un f1-score de 92%. Cette optimisation a amélioré les performances de modèle de stage 2 de 85% à 92% en termes de f1-score.

Finalement, nous avons relié les deux stages pour une évaluation finale de cette méthode proposée : chaque donnée d'entrée est d'abord évaluée par le modèle de stage 1 qui était sauvegardé avec joblib, si la prédiction est healthy elle est classifié comme healthy, sinon elle est passée au modèle de stage 2 pour finalement être classé en early ou advanced stage.

L'évaluation finale de notre méthode proposée est présentée dans le tableau 10, tandis qu'une comparaison entre les modèles classiques de classification et la méthode multistage proposé en termes des métriques d'évaluation : exactitude, précision, rappel et f1-score est présenté dans la figure 29 :

Tableau 10. Performance finale de la méthode Multistage proposée

Méthode	**Exactitude (%)**	**Précision (%)**	**Rappel (%)**	**F1-score (%)**
Multistage	98	94	83	86

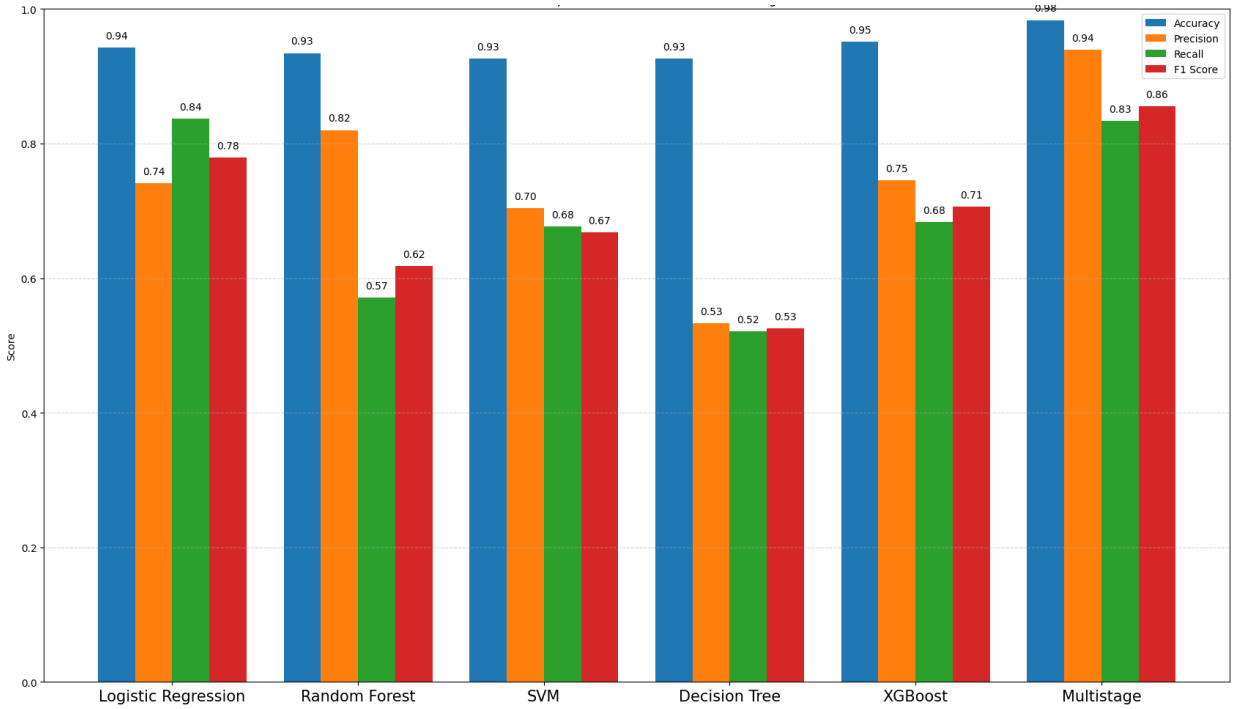

Figure 29. Comparaison des performances des modèles classiques avec la méthode multistage proposée.

En comparant les résultats de performances de la méthode classique par notre méthode proposée, notre méthode multistage à montrer une amélioration des performances surtout en termes de f1-score et elle offre une meilleure capacité à classifier les cas sains et malsains ainsi à distinguer entre les stades critiques des cas malsains.

Bien que l'étude de [23] a obtenu des résultats supérieurs en termes de f1-score (91%) pour la détection binaire (Cas sains vs Hépatite C), il est important de noter que la tâche qu'elle aborde est moins complexe. Néanmoins, notre approche multistage proposée, évalue les stades de gravité de la maladie de l'hépatite C (Stade précoce ou stade avancé). Malgré que nos résultats soient légèrement inférieures, ils demeurent pertinents et prometteurs d'un point de vue clinique.

3.3.3 Dataset COVID-19

3.3.3.1 Description du dataset

Ce dataset contient 1 048 576 données médicales anonymisés et collectés pendant la pandémie de COVID-19 y compris des informations démographiques (Age et sexe), des préconditions médicales (Maladies chroniques, grossesse, etc.) [31]. L'objectif de ce dataset est la prédiction du risque de mortalité du patient à partir de ses symptômes, préconditions et son

état clinique. Dans les caractéristiques booléennes, 1=Oui et 2=Non, ainsi les valeurs 97, 98 et 99 sont des données manquantes [31].

Tableau 11. Description des variables du dataset COVID-19 [31].

Variable	Description
Sex	Female=1 et Male=2.
Age	Age du patient.
classification	Résultats du test Covid. Valeurs 1 à 3 : patient positif à différents degrés de Covid. Valeur ≥ 4 : patient négatif ou test non concluant.
patient type	Type de prise en charge (1=renvoyé à domicile, 2=Hospitalisé)
pneumonia	Présence ou absence d'une pneumonie.
pregnancy	Enceinte ou pas.
diabetes	Diabétique ou pas.
copd	Si le patient souffre de COPD (maladie pulmonaire obstructive chronique) ou non.
asthma	Si le patient a de l'asthme ou non.
inmsupr	Si le patient est immunodéprimé.
hypertension	Si le patient a de l'hypertension ou non.
cardiovascular	Si le patient a une maladie cardiovasculaire ou non.
renal chronic	Si le patient a une maladie rénale chronique ou non.
other disease	Si le patient a une autre maladie.
obesity	Si le patient est obèse ou non.
tobacco	Si le patient consomme du tabac ou non.
usmr	Niveau de l'unité médicale.
medical unit	Type de l'établissement médical qui assure le soin.
intubed	Si le patient a été intubé ou non.
Icu	Si le patient a été admis en soins intensifs
date died	Date de décès (9999-99-99 = patient est vivant sinon il est décédé).

3.3.3.2 Prétraitement des données

Les étapes effectuées de prétraitement pour ce jeu de données sont comme suit :

- Les dix premières lignes de dataset COVID-19 ont été affichées pour avoir une vue brève sur le dataset.

	USMER	MEDICAL_UNIT	SEX	PATIENT_TYPE	DATE_DIED	INTUBED	PNEUMONIA	AGE	PREGNANT	DIABETES	...	ASTHMA	INMSUPR	HIPERTENSION	OTHER_DISEASE	CARDIOVASCULAR	OBESITY	RENAL_CHRONIC	TOBACCO	CLASIFFICATION_FINAL	ICU
0	2	1	1	1	03/05/2020	97	1	65	2	2	...	2	2	1	2	2	2	2	2	3	97
1	2	1	2	1	03/06/2020	97	1	72	97	2	...	2	2	1	2	2	1	1	2	5	97
2	2	1	2	2	09/06/2020	1	2	55	97	1	...	2	2	2	2	2	2	2	2	3	2
3	2	1	1	1	12/06/2020	97	2	53	2	2	...	2	2	2	2	2	2	2	2	7	97
4	2	1	2	1	21/06/2020	97	2	68	97	1	...	2	2	1	2	2	2	2	2	3	97
5	2	1	1	2	9999-99-99	2	1	40	2	2	...	2	2	2	2	2	2	2	2	3	2
6	2	1	1	1	9999-99-99	97	2	64	2	2	...	2	2	2	2	2	2	2	2	3	97
7	2	1	1	1	9999-99-99	97	1	64	2	1	...	2	1	1	2	2	2	1	2	3	97
8	2	1	1	2	9999-99-99	2	2	37	2	1	...	2	2	1	2	2	1	2	2	3	2
9	2	1	1	2	9999-99-99	2	2	25	2	2	...	2	2	2	2	2	2	2	2	3	2

Figure 30. Les dix premières lignes du dataset COVID-19.

- Toutes les valeurs manquantes ont été supprimées car leur proportion était importante.
- Les variables originales ont été encodées avec les valeurs 1 et 2 (Avec 1= oui, 2=non). Un re-mappage a été effectué pour une interprétation facile dans les différents modèles de Machine Learning avec : (0=non, 1=oui), Et pour la variable 'SEX' (0 =female, 1=male).
- Nous avons créé une nouvelle variable binaire qui serait la variable cible à partir de la variable DATE_DIED où :
  - 0 signifie que le patient n'est pas mort.
  - 1 signifie que le patient est mort.
- Nous avons supprimé la variable DATE_DIED.
- L'ensemble de données a été divisé en deux sous-ensembles : ensemble d'entrainement et un ensemble de test avec la méthode Train_Test_Split avec un pourcentage de 80% pour entrainement et 20% pour le test, en utilisant le paramètre stratify=y.
- Enfin un scaling de la variable AGE a été effectué.

La figure 31 présente la distribution de la variable cible Died qui montre un déséquilibre élevé :

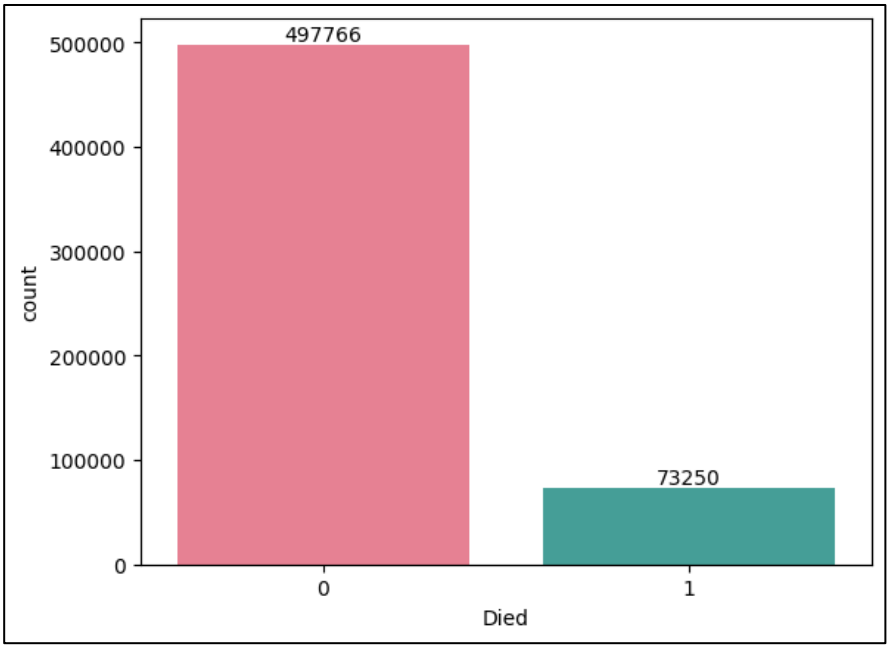

Figure 31. La distribution de la variable cible 'Died'.

3.3.3.3 Classification et évaluation des résultats

Lors de la pandémie de COVID-19, les experts de santé ont rencontré le problème de manque et de la gestion de distribution des ressources médicales comme les lits de réanimation et les respirateurs. Donc, il est essentiel de disposer d'un outil d'aide à la décision pour les experts de la santé, qui permet d'identifier le niveau de risque des patients pour une meilleur gestion de ces ressources médicales limitées et de prioriser les interventions.

L'objectif de notre approche est de fournir un outil d'aide à la décision pour les experts de santé dans les temps difficiles où il y a une surcharge dans les établissements de santé. Nous avons proposé un modèle de prédiction du risque de mortalité chez les patients pour une meilleure gestion des allocations des ressources médicales.

Dans la première expérimentation, nous avons utilisé des modèles classiques et simples, notamment LR, RF, AdaBoost avec le paramètre class_weight= balanced. Le tableau 12 et la figure 32 présentent les résultats des performances de ces modèles et une comparaison visuelle des résultats obtenus selon chaque métrique d'évaluation :

Tableau 12. Les performances des modèles simples pour le dataset COVID-19.

Modèle	Exactitude(%)	Précision(%)	Rappel(%)	F1-score(%)	ROC- AUC(%)
LR	87	73	88	77	94
RF	90	77	82	79	93
AdaBoost	91	81	78	80	95

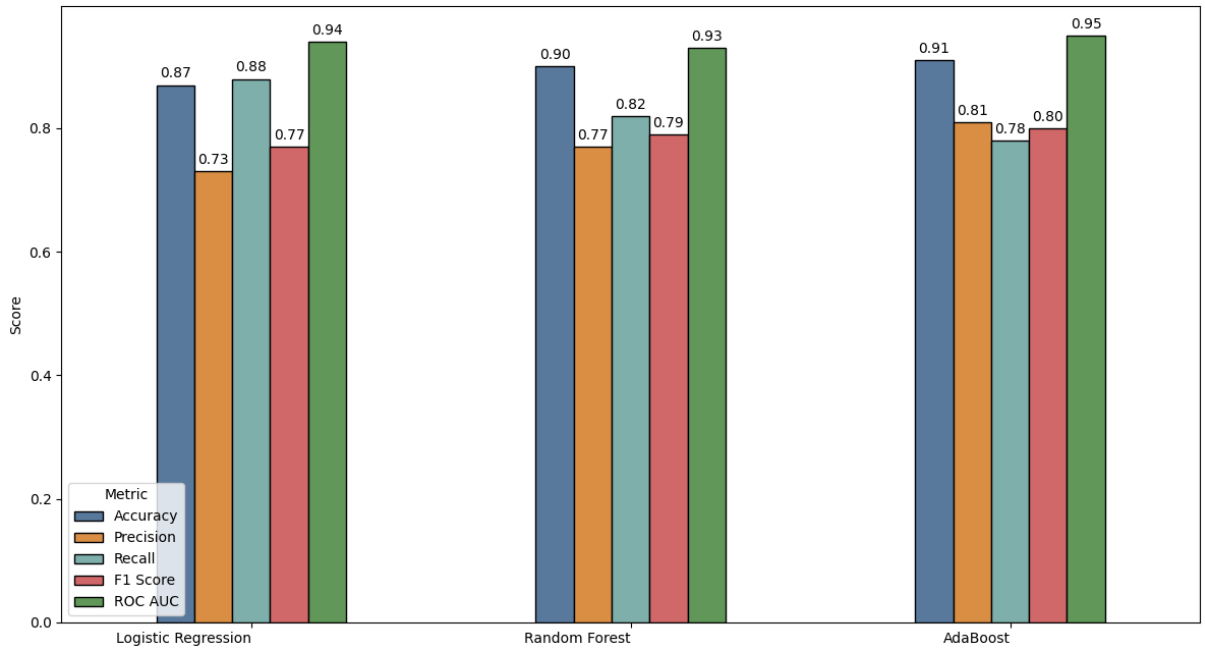

Figure 32. Visualisation des performances des modèles simples pour le dataset COVID-19.

Les résultats obtenus sont relativement faibles en termes de F1-score, donc nous avons exploré dans la deuxième expérimentation un classifieur d'ensemble Voting (avec la méthode de voting soft) avec des estimateurs plus puissants, notamment GBC, LightGBM, CatBoost et LR. LR a été choisi car il a donné un bon résultat en termes de Rappel (88%) ce qui est important dans notre contexte où les données sont déséquilibrées. Nous avons également appliqué le réglage de seuil de décision (Threshold) pour maximiser le F1-score.

Le tableau 13 présente les résultats de performances du modèle de Voting tandis que la figure 33 montre une visualisation de ces résultats obtenus :

Tableau 13. Les résultats de performances de modèle Voting.

Modèle	Exactitude(%)	Précision(%)	Rappel(%)	F1-score(%)	ROC- AUC (%)
Voting	92	82	84	83	96

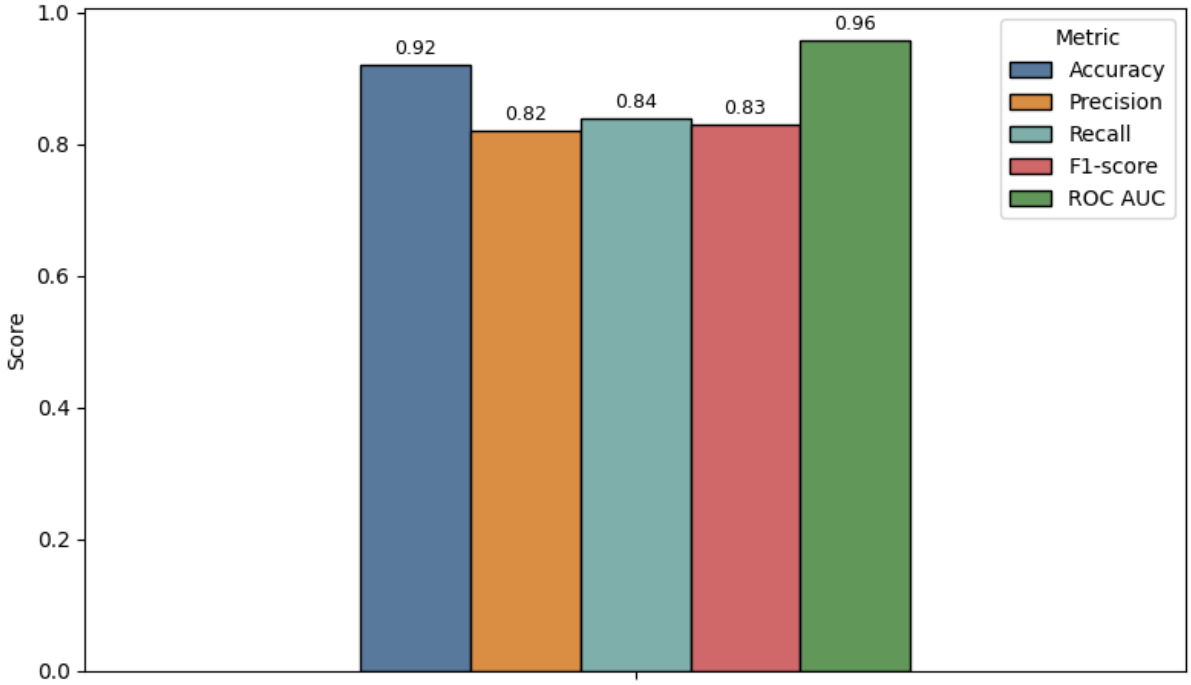

Figure 33. Visualisation des performances de modèle Voting.

Ces résultats ont été obtenus avec un seuil =0.42 et sont également faibles en termes de Rappel et f1-score à cause du déséquilibre des classes, le modèle est biaisé vers la classe majoritaire. Pour cela nous avons appliqué un sous-échantillonnage des données avec la méthode RUS (Random Under Sampling) pour que le nombre des observations de la classe majoritaire devienne égal à celui de la classe minoritaire, puis nous avons entrainé le même modèle de Voting sur ces données équilibrées. Les figures 34 et 35 montrent la distribution de la variable cible et une comparaison des résultats du modèle de Voting avant et après avoir appliqué RUS respectivement, tandis que le tableau 14 présente les résultats de performances de ce modèle :

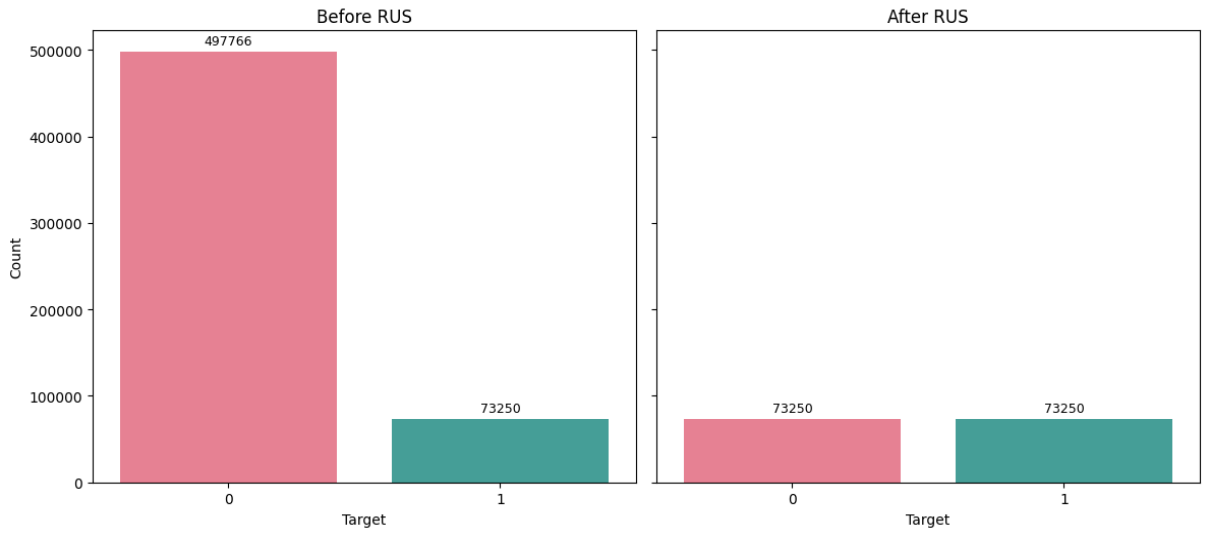

Figure 34. Distribution de la variable cible 'Died' avant et aprés l'application de RUS.

Tableau 14. Les résultats de performance de modèle Voting aprés l'application de RUS.

Modèle	Exactitude(%)	Précision(%)	Rappel(%)	F1-score(%)	ROC- AUC(%)
Voting avec RUS	89	90	89	89	96

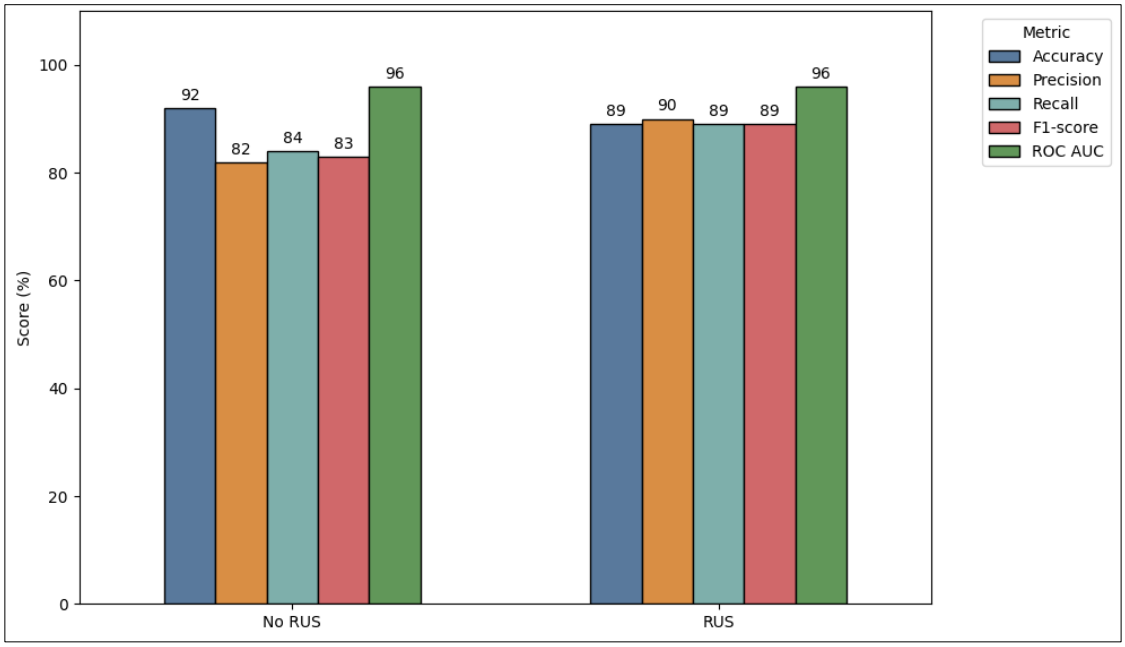

Figure 35. Comparaison des performances du modèle Voting avant et après l'application de RUS.

Après l'application de RUS et avec le seuil 0.53, nous observons une baisse de l'exactitude de 92% à 89%, ce qui est attendu car elle a favorisé la classe majoritaire avant l'application de RUS. La précision est améliorée avec un taux de 8 % ce qui indique qu'il y a moins de faux positifs (personnes qui n'ont pas décédé mais prédites par décédé). Également, on observe une amélioration en termes de rappel avec un taux de 5 % ce qui indique que notre modèle détecte mieux les cas positifs.

Les résultats obtenus indiquent que notre approche peut constituer un outil d'aide à la décision médicale pour estimer le risque de mortalité des patients pour une meilleure gestion de l'allocation des ressources médicales où il y a une surcharge dans les établissements de santé.

Pour illustrer comment notre modèle pourrait être intégré dans les systèmes cliniques, nous avons réalisé un prototype de visualisation des prédictions. Le risque de mortalité du patient prédit par notre modèle serait affiché avec le seuil optimal de décision du modèle. Et enfin, la décision finale demeure entre les mains des experts de la santé. Les figures 36 et 37 présentent une illustration de notre proposition de visualisation du résultat prédit par notre modèle en cas de survie et en cas de la mort, respectivement :

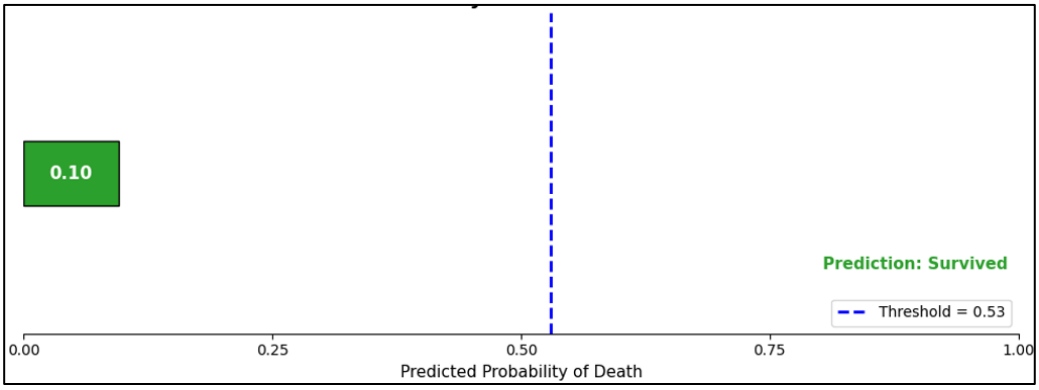

Figure 36. Visualisation du résultat prédit par notre modèle en cas de survie.

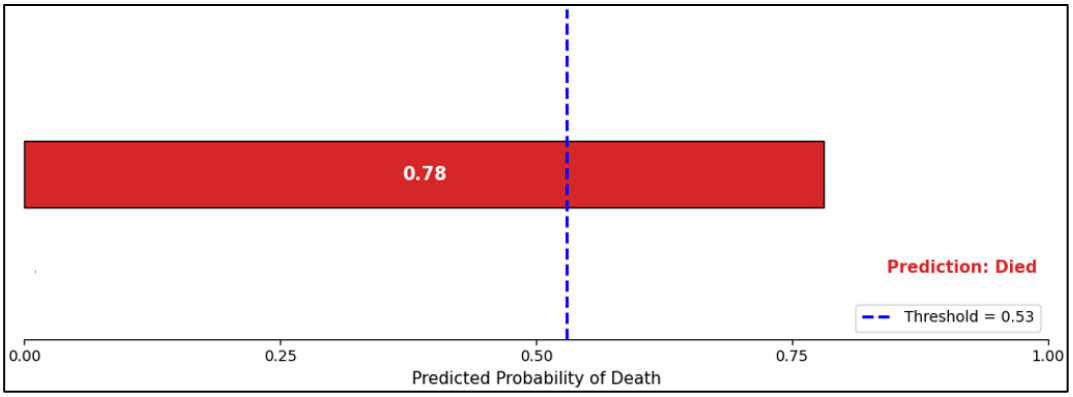

Figure 37. Visualisation du résultat prédit par notre modèle en cas de la mort.

3.4 Conclusion

Dans ce chapitre, nous avons présenté le processus expérimental, en commençant par décrire l'environnement logiciel utilisé. Ensuite nous avons décrit les ensembles de données utilisés, le prétraitement effectué et présenté les différentes expérimentations menées sur chaque ensemble de données ainsi que l'évaluation des résultats obtenus.

Conclusion générale

Conclusion générale

Les maladies infectieuses causent des millions de pertes humaines chaque année en représentant une menace très sérieuse pour la santé publique. Dans ce contexte, il devient indispensable d'une meilleure prise en charge des cas critiques et assurer une détection précoce de ces maladies. L'objectif de ce travail est d'explorer les techniques de Machine Learning pour la détection et la prédiction des maladies infectieuses ainsi que l'estimation de risque de mortalité chez les personnes infectées.

Dans ce travail, nous avons analysé trois ensembles de données relatifs à différentes maladies et obtenu des résultats prometteurs, ce qui a démontré la capacité des modèles de Machine Learning à la détection de ces maladies à un stade précoce ainsi d'estimer le risque de mortalité chez les patients pour sauver des vies. Parmi les différentes approches testées, les méthodes d'ensemble comme le Stacking et le Voting et l'ajustement des seuils de décision sont révélés efficaces. Cependant, le déséquilibre des données a été un ralentissant dans ce projet.

Ce travail ouvre des perspectives vers l'utilisation des données plus vastes, l'exploration des approches de Deep Learning et le déploiement de ces modèles dans des systèmes d'aide à la décision médicale.

Bibliographie

[1] H. Berrezoug, "La modélisation mathématique de la dynamique du choléra," 2019.

[2] B. Mahesh and others, "Machine learning algorithms-a review," *International Journal of Science and Reseach (IJSR).[Internet],* vol. 9, no. 1, pp. 381-386, 2020.

[3] S. E. ZADI and B. HADEF , "A learning-based approach for Tuberculosis detection," 2022.

[4] GeeksforGeeks, "Supervised and Unsupervised Learning," 27 February 2025. [Online]. Available: https://www.geeksforgeeks.org/supervised-unsupervised-learning/. [Accessed 28 March 2025].

[5] L. DE MATTEIS, S. Janny, S. Nathan and W. Shu-Quartier, "Introduction à l'apprentissage automatique," *Culture Sciences de l'Ingénieur,* p. 7, 2022.

[6] C. E. Bouanani and M. R. Bahoussi, "Une approche basée sur le machine learning pour la sécurité informatique : Application à la détection d'intrusion," 2022.

[7] GeeksforGeeks, "Reinforcement Learning," 24 February 2025. [Online]. Available: https://www.geeksforgeeks.org/what-is-reinforcement-learning/. [Accessed 29 May 2025].

[8] P. Das, "Logistics Regression in python," 16 June 2019. [Online]. Available: https://www.codespeedy.com/logistics-regression-in-python/. [Accessed 29 March 2025].

[9] Upasana, "Decision Tree: How to create a perfect decision tree?," 25 November 2020. [Online]. Available: https://www.edureka.co/blog/decision-trees/.

[10] A. Kalbande, "Random Forest algorithm in machine learning," 22 April 2022. [Online]. Available: https://www.fireblazeaischool.in/blogs/random-forest-algorithm/. [Accessed 29 March 2025].

[11] A. Manglick, "Support Vector Machine (SVM)," 2 July 2017. [Online]. Available: https://arun-aiml.blogspot.com/2017/07/support-vector-machine-svm.html. [Accessed 15 May 2025].

[12] J. Obukwelu, "What's the KNN? - nerd for tech - medium," 16 December 2021. [Online]. Available: https://medium.com/nerd-for-tech/whats-the-knn-74e84458bd24. [Accessed 15 May 2025].

[13] GeeksforGeeks, "What is a Neural Network?," 3 April 2025. [Online]. Available: https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/. [Accessed 15 May 2025].

[14] P. Ingle, "Top Neural network Architectures for Machine learning Researchers," 12 January 2025. [Online]. Available: https://www.marktechpost.com/2022/09/23/top-neural-network-architectures-for-machine-learning-researchers/. [Accessed 15 May 2025].

[15] Y. Su and Y. Shen, "A Deep Learning-Based Sentiment Classification model for real online consumption.," *Frontiers in Psychology,* vol. 13, 2022.

[16] N. Kumar and K. Sikamani, "Prediction of chronic and infectious diseases using machine learning classifiers-A systematic approach," *Int J Intell Eng Syst,* vol. 13, no. 4, pp. 11-20, 2020.

[17] V. C. Osamor and A. F. Okezie, "Enhancing the weighted voting ensemble algorithm for tuberculosis predictive diagnosis," *Scientific Reports,* vol. 11, no. 1, p. 14806, 2021.

[18] M. S. Islam, S. A. Khushbu, A. S. A. Rabby and T. Bhuiyan, "A study on dengue fever in bangladesh: Predicting the probability of dengue infection with external behavior with machine learning," in *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, 2021, pp. 1717-1721.

[19] V. V. Khanna, K. Chadaga, N. Sampathila, S. Prabhu and R. Chadaga, "A machine learning and explainable artificial intelligence triage-prediction system for COVID-19," *Decision Analytics Journal,* vol. 7, p. 100246, 2023.

[20] M. Alehegn, "Application of machine learning and deep learning for the prediction of HIV/AIDS," *HIV \& AIDS Review. International Journal of HIV-Related Problems,* vol. 21, no. 1, pp. 17-23, 2022.

[21] U. K. Lilhore, P. Manoharan, J. K. Sandhu, S. Simaiya, S. Dalal, A. M. Baqasah, M. Alsafyani, R. Alroobaea, I. Keshta and K. Raahemifar, "Hybrid model for precise hepatitis-C classification using improved random forest and SVM method," *Scientific Reports,* vol. 13, no. 1, p. 12473, 2023.

[22] G. Gupta, S. Khan, V. Guleria, A. Almjally, B. I. Alabduallah, T. Siddiqui, B. M. Albahlal, S. A. Alajlan and M. Al-Subaie, "DDPM: A dengue disease prediction and diagnosis model using sentiment analysis and machine learning algorithms," *Diagnostics,* vol. 13, no. 6, p. 1093, 2023.

[23] Y. Wang and B. a. Z. Q. Yin, "Application of Machine Learning Algorithms in Predicting Hepatitis C," in *Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science*, 2023, pp. 359-365.

[24] O. Khan, J. O. Ajadi and M. P. Hossain, "Predicting malaria outbreak in The Gambia using machine learning techniques," *Plos one,* vol. 19, no. 5, p. e0299386, 2024.

[25] A. W. Jannah and B. Al Kindhi, "Optimization of Early Detection of Tuberculosis: Use of Multilayer Perceptron and Extreme Learning Machine with Clinical Data," *Jurnal Indonesia Sosial Teknologi,* vol. 5, no. 5, 2024.

[26] S. Melchane, Y. Elmir and F. Kacimi, "Infectious diseases prediction based on machine learning: The impact of data reduction using feature extraction techniques," *Procedia Computer Science,* vol. 239, pp. 675-683, 2024.

[27] A. Mizwar, B. Hartato, A. Ridwan and F. Asharudin, "Machine Learning-Based Approach for HIV/AIDS Prediction: Feature Selection and Data Balancing Strategy," *Journal of Applied Informatics and Computing,* vol. 9, pp. 338-347, 02 2025.

[28] GeeksforGeeks, " What is Python? Its Uses and Applications.," 4 April 2025. [Online]. Available: https://www.geeksforgeeks.org/what-is-python/#uses-and-applications-of-python.

[29] "AIDS virus infection prediction 💉," 28 April 2024. [Online]. Available: https://www.kaggle.com/datasets/aadarshvelu/aids-virus-infection-prediction/data. [Accessed 21 April 2025].

[30] R. Lichtinghagen, F. Klawonn and G. Hoffmann. [Online]. Available: https://archive.ics.uci.edu/dataset/571/hcv+data. [Accessed 13 April 2025].

[31] "COVID-19 Dataset," 2022. [Online]. Available: https://www.kaggle.com/datasets/meirnizri/covid19-dataset/data. [Accessed 12 May 2025].

ملخص

أثبت التعلم الآلي فعاليته الكبيرة في مكافحة الأمراض المعدية، إذ يُمكّن من الكشف المبكر، والتنبؤ الدقيق بالأوبئة ومخاطرها، والمساعدة في التشخيص. ومن خلال تقنياته المتنوعة، يُحلل البيانات المعقدة (الأعراض، والتصوير الطبي، والعوامل البيئية) لتحديد الأنماط غير المرئية للعين البشرية. ومع ذلك، لا تزال جودة البيانات وتعقيدها، بالإضافة إلى القضايا الأخلاقية المتعلقة باستخدامها، تُشكل تحديات رئيسية. ورغم هذه القيود، يُقدم التعلم الآلي آفاقًا واعدة لتحسين الصحة العامة وإنقاذ الأرواح من خلال اتخاذ قرارات طبية أسرع وأكثر دقة. في هذه الدراسة، أجرينا بحث علمي ودراسة تقنيات التعلم الآلي المختلفة للكشف عن الأمراض المعدية والتنبؤ بها من ثلاث مجموعات بيانات مختارة، كل منها يُمثل مرضًا مختلفًا: الإيدز/فيروس نقص المناعة البشرية، والتهاب الكبد الوبائي سي، وكوفيد-19. لكل مرض، اعتمدنا نهج نمذجة محددًا، مما أسفر عن نتائج واعدة وذات صلة. تفتح هذه النتائج الباب أمام العديد من وجهات النظر، بما في ذلك استخدام مجموعات بيانات أكبر، واستكشاف أساليب التعلم العميق، ودمج هذه النماذج في أنظمة دعم القرار الطبي لتحقيق تشخيص أكثر دقة.

الكلمات المفتاحية: الأمراض المعدية، الكشف المبكر، تصنيف المخاطر، التعلم الآلي، التصنيف.

Abstract

Machine learning has proven to be highly powerful in the fight against infectious diseases by enabling early detection, accurate prediction of epidemics and their risks, and assisted diagnosis. Through its various techniques, it analyzes complex data (symptoms, medical imaging, environmental factors) to identify patterns invisible to the human eye. However, the quality and complexity of the data and the ethical issues related to its use remain major challenges. Despite these limitations, machine learning offers promising prospects for improving public health and saving lives through faster and more accurate medical decisions. In this study, we conducted scientific research and examined different machine learning techniques for the detection and prediction of infectious diseases from three selected datasets, each corresponding to a different disease: AIDS/HIV, hepatitis C, and COVID-19. For each pathology, we adapted a specific modeling approach, which yielded promising and relevant results. These results open the way to several perspectives, including the use of larger data sets, the exploration of Deep Learning approaches and the integration of these models into medical decision support systems for more precise diagnosis.

Keywords: Infectious Diseases, Early Detection, Risk Stratification, Machine Learning, Classification.

Résumé

Le Machine Learning s'est révélé hautement puissant pour lutter contre les maladies infectieuses en permettant une détection précoce, une prédiction précise des épidémies et leurs risques et un diagnostic assisté. Grace à ses différentes techniques, il analyse des données complexes (symptômes, imagerie médicale, facteurs environnementaux) pour identifier des patterns invisibles à l'œil humain. Cependant, la qualité et la complexité des données et les questions éthiques liées à leur utilisation restent des défis majeurs. En dépit de ces limites, le Machine Learning offre des perspectives prometteuses pour améliorer la santé publique et sauver des vies grâce à des décisions médicales plus rapides et plus précises. Dans le cadre de cette étude, nous avons mené une recherche scientifique et d'examiner différentes techniques de Machine Learning pour la détection et la prédiction des maladies infectieuses à partir de trois ensembles de données sélectionnés chacun correspond à une maladie différente : SIDA/VIH, hépatite C et le COVID-19. Pour chaque pathologie, nous avons adapté une approche de modélisation spécifique ce qui a donné des résultats prometteurs et pertinents. Ces résultats ouvrent la voie à plusieurs perspectives, notamment l'utilisation des ensembles de données plus vastes, l'exploration des approches de Deep Learning ainsi l'intégration de ces modèles dans les systèmes d'aide à la décision médicale pour un diagnostic plus précis.

Mots-clés : Maladies Infectieuses, Détection Précoce, Stratification du risque, Machine Learning, Classification.