الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي

جامعة سعيدة د .مولاي الطاهر
كلية التكنولوجيا
قسم: الإعلام الآلي

Mémoire de Master

Spécialité : Modélisation Informatique des Connaissances et du Raisonnement

Thème

Machine learning dans l'épidémiologie (application au maladies transmissibles et non transmissibles)

Présenté par :

Mehanguef Ghizlane

Hassad Ahlem

Dirigé par :

Pr. Hamou Mohamed Reda

Dr. Zerrouki Kadda

Année universitaire 2022-2023

Table des matières

Liste des figures

Liste des tableaux:

Liste des abréviations:

IA Intelligence Artificielle

ML Machine Learning

MNT Maladies non transmissibles

MT Maladies transmissibles

CART Classification and Regression Tree

MDO Maladies à déclaration obligatoire

UCU University and College Union

MCV Maladies cardiovasculaires

KNN K-Nearest Neighbors

KPPV k plus proche voisins

AVC accident vasculaire

DL deep learning (apprentissage profond)

Dédicace

On dédie ce travail à nos chers parents et à toutes les personnes qui nous ont aidé, soutenu et encouragé durant notre parcours universitaire, ce projet vous est dédié.

Remerciement

Nous tenons à exprimer toute notre reconnaissance à notre directeur de recherche, Pr.HAMOU Mohamed Redha ainsi que notre co-encadreur ZERROUKI Kadda pour leur patience, leur disponibilité et leur judicieux conseils.

Nous adressons notre sincère gratitude à tous les professeurs, intervenants et toutes les personnes qui par leurs paroles, leurs écrits, leurs conseils et leurs critiques ont guidé nos réflexions.

Nous souhaitons également remercier les membres du jury pour l'intérêt qu'ils ont porté à notre projet de fin d'études, en acceptant de l'examiner et en nous faisant part de leurs propositions constructives pour son amélioration.

Nous remercions aussi nos très chers parents pour leur soutien constant et leurs encouragements, ainsi que nos amis qui ont toujours été là pour nous. Leur soutien inconditionnel et leurs encouragements ont été d'une grande aide.

À tous ces intervenants, nous présentons nos remerciements, nos respect et notre gratitude.

ملخص

تقدم هذه المذكرة استخدام التعلم الآلي للتنبؤ بالتهاب الكبد C وأمراض القلب والأوعية الدموية باستخدام خوارزميات التعلم الآلي الخاضعة للإشراف. الهدف هو تطوير نموذج تنبؤ دقيقة وموثوق لتحديد الأفراد المعرضين لخطر الإصابة بهذه الأمراض. تم معالجة البيانات وتطبيعها ثم إجراء تحليل البيانات الاستكشافية. تم استخدام خوارزميات التعلم الآلي لبناء نماذج التنبؤ. تقييم أداء كل خوارزمية باستخدام مقاييس مثل الدقة وقياس F ... وأظهرت النتائج أن خوارزميات Random Forest و Decision Tree كانت أفضل أداء من حيث التصنيف.

الكلمات المفتاحية:الذكاء الاصطناعي،التعلم الآلي ،التهاب الكبد c ، أمراض القلب والأوعية الدموية،التنبؤ .

Abstract

This paper presents the use of machine learning to predict hepatitis C and cardiovascular diseases using supervised machine learning algorithms. The objective is to develop accurate and reliable prediction models to identify individuals at high risk of developing these diseases. The data was preprocessed and normalized, followed by exploratory data analysis. Machine learning algorithms were then used to construct the prediction models. The performance of each algorithm was evaluated using measures such as accuracy, precision, and F-measure… The results demonstrated that Random Forest and Decision Tree algorithms outperformed others in terms of classification.

Keywords: Artificial intelligence; machine learning;hepatitis c; cardiovascular disease; prediction.

Résumé

Ce mémoire présente l'utilisation du machine learning pour prédire l'hépatite C et les maladies cardiovasculaires à l'aide des algorithmes de Machine Learning supervisée. L'objectif est de développer des modèles de prédiction précis et fiables pour identifier les individus à risque élevé de développer ces maladies. Les données ont été prétraitées, normalisées ensuite une analyse exploratoire des données a été effectuée. Les algorithmes de machine learning ont ensuite été utilisés pour construire les modèles de prédiction. Les performances de chaque algorithme ont été évaluées en utilisant des mesures telles que l'exactitude, la précision, la F-mesure... Les résultats ont démontré que les algorithmes Random Forest et Decision Tree étaient les plus performants en termes de classification.

Mot clés: Intelligence artificiel; machine learning; prédictions; maladies; hépatite C; cardiovasculaire

Introduction général

Contexte:

L'apprentissage automatique (machine learning) est devenu un outil puissant dans le domaine des soins de santé pour la modélisation prédictive et la prise de décision. L'une des applications cruciales de l'apprentissage automatique est la prédiction de maladies telles que l'hépatite C et les maladies cardiovasculaires.
Les maladies cardiovasculaires et l'hépatite C sont deux problèmes de santé majeurs qui ont des conséquences significatives sur la mortalité dans le monde. Les maladies cardiovasculaires, telles que les maladies cardiaques et les accidents vasculaires cérébraux, sont responsables d'un nombre élevé de décès depuis plusieurs décennies. Selon les statistiques, un pourcentage élevé de la population est touché par ces maladies, et un nombre important d'hommes et de femmes en meurent chaque année.
“L'hépatite C”, est une maladie virale qui affecte le foie, est également une préoccupation majeure pour la santé publique. Les statistiques montrent que cette maladie a causé un nombre considérable de décès dans le monde. Une détection tardive de “l'hépatite C” peut entraîner une réduction significative de l'espérance de vie, et le traitement de cette maladie peut être complexe et exigeant.
Les domaines de l'intelligence artificielle permettent de prédire des résultats en utilisant des processus d'apprentissage à partir de données d'entrée.
Dans le cas des maladies cardiovasculaires et de l'hépatite C, les méthodes de ML utilisent des données d'entrée pour y appliquer des algorithmes supervisés, non supervisés ou semi supervisés. Ces algorithmes permettent de traiter des ensembles de données complexes et d'identifier des schémas ou des relations entre les données qui peuvent être utilisées pour le diagnostic et la prédiction.
Il est important de noter que les progrès technologiques et l'utilisation de l'IA et du ML ont contribué aux recherches sur les maladies cardiovasculaires et l'hépatite C, en offrant de nouvelles opportunités pour améliorer le diagnostic, le traitement et la prévention de ces maladies.

Problématique :

La détection précoce et l'intervention opportune peuvent grandement améliorer les résultats des patients et réduire la charge pesant sur les systèmes de santé. Les méthodes traditionnelles de prédiction des maladies reposent souvent sur l'analyse manuelle et la prise de décision subjective, ce qui peut être chronophage et sujet aux erreurs. Les algorithmes d'apprentissage automatique offrent une solution

prometteuse en exploitant de vastes ensembles de données et une reconnaissance automatisée de motifs pour générer des prédictions précises.

Cependant, l'efficacité des différents algorithmes d'apprentissage automatique dans la prédiction de l'hépatite C et des maladies cardiovasculaires reste un sujet d'investigation. Le choix de l'algorithme (ou des algorithmes) le plus adapté pour cette application spécifique est essentiel pour garantir la fiabilité et l'exactitude des modèles de prédiction. De plus, comprendre les mesures de performance et comparer les résultats de différents algorithmes est crucial pour sélectionner l'approche optimale.

Par conséquent, le problème clé abordé dans ce mémoire est le suivant : Comment utiliser efficacement les algorithmes d'apprentissage automatique pour développer des modèles de prédiction précis et fiables afin d'identifier les individus à haut risque de développer l'hépatite C et les maladies cardiovasculaires ? En explorant et en évaluant différents algorithmes et mesures de performance, cette étude vise à fournir des connaissances précieuses pour le développement de modèles de prédiction efficaces pour ces maladies.

Objectif:

l’objectif de notre mémoire est de créer un modèle basé sur le machine learning pour la prédiction des maladies transmissibles et non transmissibles, plus précisément de” l’Hépatite C” et les ” MCV” (Maladies cardiovasculaire) en ce basant sur des données médicales de différent patient à travers le monde.

Chapitre I : Etat de l’art et travaux connexes

1. Etat de l’art et travaux connexe :

L'avancement de l'intelligence artificielle (IA) et de l'informatique a révolutionné l'aide à la preuve scientifique en médecine, en permettant une analyse plus approfondie des données médicales, l'interprétation des images médicales et la gestion efficace des vastes quantités de données. Cette évolution a ouvert de nouvelles perspectives pour la découverte de connaissances, la personnalisation des traitements et l'amélioration de la médecine fondée sur des preuves.[42]

Il y a d’abord eu l'approche symbolique, qui utilise des règles logiques pour raisonner, a donné lieu au développement de systèmes experts dans le domaine médical. Ces systèmes exploitent les connaissances médicales spécifiques à un domaine donné et formalisent les raisonnements des experts pour parvenir à des diagnostics. [42]

De manière simultanée, l'approche numérique a également connu des améliorations significatives grâce à l'utilisation de l'intelligence artificielle (IA) pour l'analyse de grandes quantités de données. Les algorithmes d'apprentissage profond, également connus sous le nom de deep learning, sont inspirés du fonctionnement du cerveau. Ils simulent un réseau de neurones organisés en différentes couches qui interagissent entre elles. L'algorithme apprend la tâche qui lui est assignée par des essais et des erreurs successifs.[43]

Pour donner un exemple concret, les applications utilisées pour analyser des photographies de peau à la recherche de mélanome ont commencé par l'intégration initiale de 50 000 images, qu'elles soient pathologiques ou non. La performance de l'analyse s'améliore au fil du temps à mesure que le système d'IA effectue de nouvelles analyses et apprend des résultats obtenus.[43]

Au début, les travaux se sont principalement concentrés sur des données reproductibles et directement utilisables, telles que des résultats biologiques comme la présence ou l'absence de cellules de mélanome. Cela était possible grâce à la disponibilité de bases de données contenant un grand nombre de photographies numérisées ainsi que des listes informatisées de résultats d'analyses biologiques ou anatomo-pathologiques. Ces données ont fourni une base solide

pour le développement de modèles d'apprentissage automatique capables d'analyser et d'interpréter ces informations de manière efficace et précise.[44]

En obstétrique, les recherches se sont intensifiées ces dernières années, en particulier dans les domaines du dépistage de la trisomie 21 par le calcul du risque combiné et de l'aide au diagnostic ou à la détection des anomalies cardiaques fœtales. Cette spécificité de l'étude de la grossesse et du fœtus in utero nous amène à passer d'une médecine prédictive, à une confrontation avec la réalité.[44]

Les avancées technologiques ont contribué à l'amélioration de ces systèmes de prédiction, rendant certaines prévisions de plus en plus fiables dans certains domaines. Cependant, les développements dans d'autres domaines sont plus délicats en raison de paramètres mal définis, nombreux et parfois difficiles à quantifier, ainsi que des données peu utilisables en l'état.[45]

Il est important de noter que la fiabilité et l'applicabilité des prédictions varient en fonction des paramètres spécifiques étudiés. Malgré ces défis, la recherche se poursuit pour affiner les modèles et les algorithmes, en intégrant de nouvelles données et en développant des méthodes d'analyse plus avancées.[45]

2. Machine learning dans la détection et prédiction des maladies:

L'application de l'apprentissage automatique aux données multimodales de la médecine de précision permet une analyse approfondie des ensembles de données massifs. Grâce à cette approche, une meilleure compréhension de la santé et des maladies humaines peut être obtenue. L'accent est mis sur l'utilisation de l'apprentissage automatique pour traiter les "mégadonnées" de la médecine, en tenant compte de la génétique, de la génomique et d'autres domaines connexes. voici quelques exemples de l’utilisation du machine learning pour la prédiction des maladies:[41]

- **Prédiction de COVID-19** : Le COVID-19 s’est révélé être une maladie virale infectieuse et mortelle, et sa propagation rapide et massive est devenue l’un des plus grands défis du monde.

Les chercheurs ont fourni un examen complet du rôle de l’apprentissage profond et l’apprentissage automatique dans la recherche de techniques de prédiction pour le COVID-19.

Un modèle mathématique a été formulé pour analyser et détecter sa menace potentielle. Le modèle proposé est un algorithme de détection intelligent basé sur le cloud utilisant une machine à vecteurs de support (CSDC-SVM) avec des tests de validation croisés.
Les résultats expérimentaux ont atteint une précision de 98,4%.[46]

- **Le projet SCRUM-Japan Genesis** : vise à établir un algorithme, appelé séquençage virtuel (VSQ), en utilisant la technologie d'apprentissage profond (DL) et les diagnostics pathologiques pour la prédiction des anomalies du génome du cancer. [47]

- **Prédiction du développement de la maladie d'Alzheimer** : pour des patients atteints d'une déficience cognitive légère. [48]

- **Les maladies cardio-vasculaires (MCV)** désignent, pour la plupart, des affections comprenant des veines limitées ou obstruées qui peuvent provoquer une crise cardiaque, une angine de poitrine ou un accident vasculaire cérébral. Le classificateur d'apprentissage automatique prédit l'affection en fonction de l'état de l'effet secondaire subi par le patient. [49]

Chapitre II : épidémiologie

1.L'épidémiologie:

1.1. Introduction :

L'Épidémiologie est une science aux contours incertains.

Le terme “EPIDÉMIOLOGIE” est lui-même ambiguë et son sens a varié au cours des âges.

Son champ d’intérêt s’accroît d’année en année, et sa méthodologie est encore en pleine évolution.

De nombreuses définitions proposées témoignent de l’évolution de l’épidémiologie.

Dans le sens littéral : l’épidémiologie est la science des phénomènes qui concerne l’ensemble d’une population vivant sur un territoire.

Épi : Sur

Demos : Peuple ou Population

Logos : Étude ou Discipline scientifique

1.2.Définition :

L'Épidémiologie est classiquement définie comme l’étude de la distribution des maladies dans les populations humaines, ainsi que les influences qui déterminent cette distribution. Son champ s’est étendu pour couvrir aussi l’étiologie de l’ensemble des problèmes de santé ainsi que leur contrôle.[26]

« L’étude de la distribution et des déterminants des états ou phénomènes liés à la santé dans une population déterminée et l’application de cette étude à la prévention et à la maîtrise des problèmes de santé »[27]

1.3.Objectif:

L’épidémiologie est une science qui participe aux actions de santé publique. Elle est une des disciplines qui permet d’étayer la prise de décision. Elle apporte aux responsables de la politique de santé, des mesures, des prévisions et des évaluations.

L’épidémiologie se décompose en activités :

de surveillance ;

d’investigation ;

de recherche ;

d’évaluation.

1.4. Buts des pratiques épidémiologiques[28]

1- L’épidémiologie peut aider à la compréhension des états de santé et des maladies.

2- L’épidémiologie permet la mesure de l’état de santé d’une population.

3- La méthode épidémiologique peut aider à identifier les agents pathogènes et trouver la source de ces agents.

4- L’épidémiologie peut aider à comprendre comment la maladie est transmise.

5- L’épidémiologie peut découvrir qui risque de devenir malade et permet la mesure des risques individuels et collectifs.

6- L’épidémiologie peut dévoiler l’exposition spécifique qui a causé directement la maladie (Facteur de risque).

7- L’épidémiologie permet la prévention de la survenue des maladies et des phénomènes pathologiques.

8- L'épidémiologie permet l’évaluation des méthodes d’intervention.

“La finalité est d’améliorer la santé des populations grâce à une meilleure compréhension et connaissance des maladies”

2. **Maladies** :

2.1.Introduction:

1,8 millions d'enfants meurent dans le monde par des maladies évitables par la vaccination Sur 57 millions de décès dans le monde, nous estimons qu'environ 40% sont dues aux maladies transmissibles.

Les deux tiers surviennent en Afrique.[29]

2.2. Problèmes des Maladies chroniques:

Les maladies non transmissibles chroniques posent un grand problème de santé publique.

Plusieurs facteurs intimement liés au rythme de vie des populations humaines interviennent dans la survenue de ces maladies.

L'allongement de la durée de vie des populations (Espérance de vie), augmente les risques de survenue des maladies non transmissibles et/ou chroniques.[30]

2.3.Les Maladies non transmissibles:

•les maladies non transmissibles (MNT), ne se transmettent pas d'une personne à l'autre.

•Elles peuvent être des maladies chroniques à progression lente et peuvent mener à une mort plus rapide comme certains types d'AVC.

Elles se composent essentiellement de maladies cardiovasculaires, les cancers, les maladies respiratoires chroniques (comme la broncho-pneumopathie chronique obstructive OU l'asthme) et le diabète.[27]

2.4.Les Maladies transmissibles:

sont des maladies infectieuses ayant la capacité de se transmettre à plusieurs individus et entre individus.

Une maladie infectieuse est un déséquilibre résultant de l'interaction entre :

- Un agent infectieux, agent « causal »
- Son hôte, l'homme
- Les facteurs environnementaux[29]

Chapitre III : Machine Learning

1.Qu’est ce que le Machine Learning ?

Le machine learning est une branche de l'intelligence artificielle qui permet aux ordinateurs d'apprendre et de prendre des décisions ou faire des prédictions à partir de données, sans être explicitement programmés. En utilisant des échantillons de données, des modèles d'apprentissage automatique sont créés en entraînant ces modèles sur les données disponibles. Ces modèles peuvent ensuite être utilisés pour analyser de nouvelles données et effectuer des prédictions ou prendre des décisions basées sur les schémas et les tendances identifiés lors de l'apprentissage.
L'objectif principal du machine learning est de permettre aux ordinateurs de généraliser à partir des exemples passés et de les appliquer à de nouvelles situations pour prédire des événements futurs.

Dans le domaine de la technologie, le machine learning est une compétence hautement recherchée. Des entreprises leaders telles que Google, Facebook et Amazon sont activement à la recherche d'experts en machine Learning. De plus, des entreprises comme Microsoft, Uber et Airbnb utilisent des algorithmes de machine learning pour améliorer divers aspects de leurs activités, tels que l'expérience client, la personnalisation des recommandations et l'optimisation des processus.

En résumé, le machine learning est un domaine clé de l'intelligence artificielle qui permet aux ordinateurs d'apprendre à partir de données et de faire des prédictions. Son utilisation est répandue dans divers secteurs technologiques et offre des opportunités d'innovation et d'amélioration des performances.[1]

2.L’histoire du Machine Learning :

L'histoire du Machine Learning remonte à plusieurs décennies et a connu une évolution remarquable. Voici un aperçu de son parcours :

· **Les débuts** : Dans les années 1950 et 1960, des chercheurs pionniers tels qu'Arthur Samuel ont commencé à explorer la possibilité de créer des programmes capables d'apprendre à partir de l'expérience. Des travaux tels que le programme de jeu de dames d'Arthur Samuel ont ouvert la voie à l'utilisation de l'apprentissage automatique pour améliorer les performances des machines.
· **L'ère des réseaux de neurones** : Dans les années 1980 et 1990, les réseaux de neurones artificiels ont suscité un intérêt croissant. Des chercheurs tels que Geoffrey Hinton ont contribué à développer des modèles de réseaux neuronaux

profonds et ont démontré leur efficacité pour résoudre des problèmes complexes tels que la reconnaissance vocale et la vision par ordinateur.

· **L'explosion du Big Data** : Au cours des dernières décennies, l'explosion des données numériques a joué un rôle clé dans l'évolution du Machine Learning. La disponibilité de vastes ensembles de données a permis de développer des modèles plus puissants et précis, alimentant ainsi les progrès de l'apprentissage automatique.

· **Avancées algorithmiques** : Le développement de nouveaux algorithmes et techniques a joué un rôle crucial dans l'évolution du Machine Learning. Des méthodes telles que les machines à vecteurs de support, les arbres de décision, les forêts aléatoires et les réseaux de neurones profonds ont permis de résoudre une gamme variée de problèmes.

· **Applications pratiques** : Le Machine Learning a connu une adoption généralisée dans divers domaines grâce à des applications pratiques. Des domaines tels que la reconnaissance d'images, la recommandation de produits, la prédiction de fraudes et la traduction automatique ont bénéficié des avancées du Machine Learning, améliorant ainsi notre quotidien.

· **Intelligence artificielle moderne** : Le Machine Learning est devenu une composante clé de l'intelligence artificielle moderne. Les algorithmes de Machine Learning permettent de créer des modèles d'IA capables de comprendre, de raisonner et de prendre des décisions, conduisant ainsi à des systèmes d'IA avancés et sophistiqués.[2]

3.Types de Machine Learning :

Différentes catégories d'apprentissage automatique (machine learning) sont utilisées pour résoudre une variété de problèmes. Voici quelques principaux types d'apprentissage automatique :

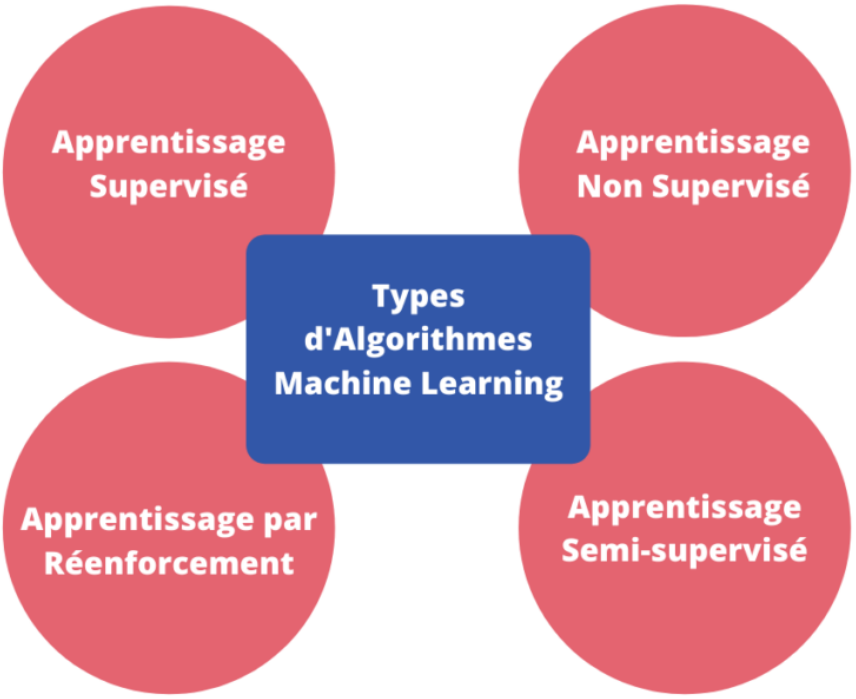


Figure 1 : Types de Machine Learning[31]

3.1.L’apprentissage supervisé :

L'apprentissage supervisé est considéré comme l'une des formes les plus simples de machine learning, car il repose sur des données d'entrée avec des étiquettes ou des annotations spécifiques pour chaque sortie. Les algorithmes utilisés sont formés sur ces données afin de développer un modèle capable de comprendre les relations entre les entrées et les sorties, et de produire des résultats précis pour de nouvelles données. Pendant le processus d'apprentissage, le modèle ajuste ses paramètres afin de mieux représenter les motifs et les corrélations présents dans les données, permettant ainsi de faire des prédictions correctes.[3]
Il existe deux catégories principales d'algorithmes : la régression et la classification.

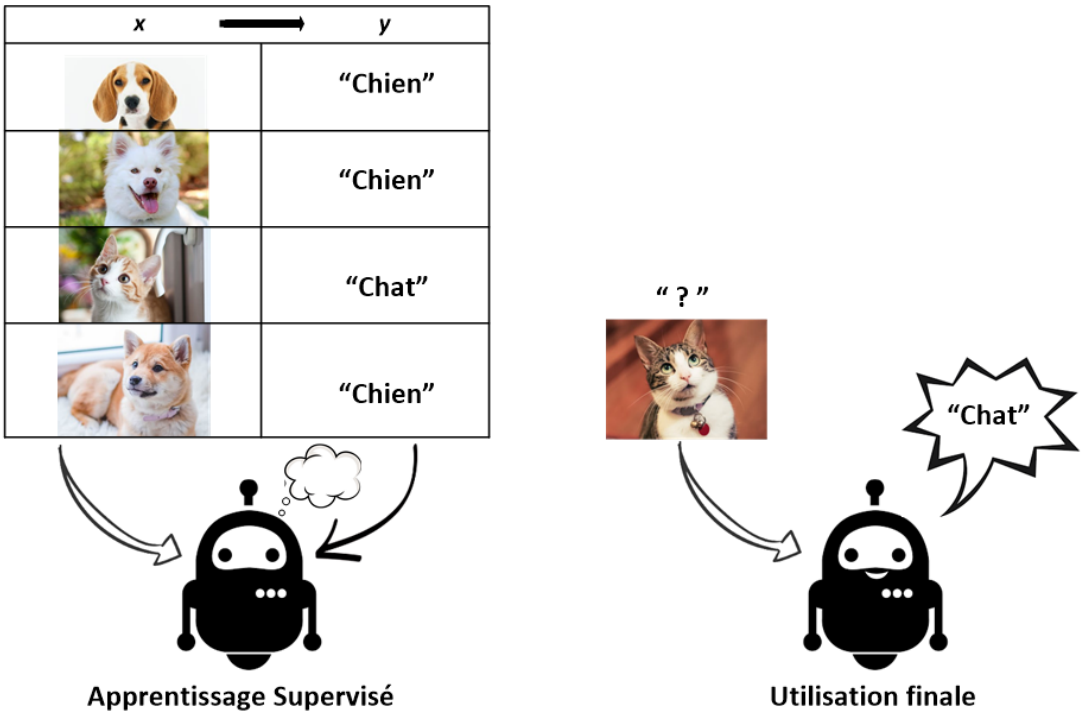


Figure 2 :Apprentissage supervisé [32]

Les algorithmes de régression, qui cherchent à prédire une valeur continue, une quantité.

Les algorithmes de classification, qui cherchent à prédire une classe/catégorie.

Les algorithmes de régression: sont utilisés pour estimer ou prédire des valeurs continues, c'est-à-dire des quantités numériques. Ces algorithmes cherchent à établir une relation mathématique entre les variables d'entrée et de sortie afin de pouvoir prédire une valeur numérique pour de nouvelles données. Ils sont largement utilisés dans des domaines tels que l'économie, la finance, la

météorologie et d'autres disciplines où la prédiction de valeurs continues est essentielle.[33]

Les algorithmes de classification: sont conçus pour prédire la classe ou la catégorie à laquelle une nouvelle donnée appartient. Ils analysent les caractéristiques ou les attributs des données d'entrée et les associent à des catégories prédéfinies. L'objectif est de développer un modèle capable de généraliser cette association afin de classifier de nouvelles données. Les algorithmes de classification sont largement utilisés dans des domaines tels que la reconnaissance d'images, la détection de spam, la prédiction de maladies et d'autres applications où la catégorisation précise des données est cruciale.

3.1.1.Algorithmes de ML supervisé:

3.1.1.1. **Naïve Bayes**:

La classification naïve bayésienne repose sur l'hypothèse d'indépendance conditionnelle entre les caractéristiques. Elle utilise le théorème de Bayes pour calculer la probabilité d'un événement en se basant sur des connaissances préalables des conditions liées. Ce théorème, découvert par le statisticien Thomas Bayes au 18e siècle, a été publié par Richard Price après sa mort. La formule du théorème de Bayes est la suivante :

$P(A|B) = P(B|A)P(A) / P(B)$

Dans un problème de classification, l'objectif est de trouver la classe la plus probable (étiquette A) étant donné les caractéristiques observées (B). En utilisant le

théorème bayésien et en supposant l'indépendance entre les caractéristiques et le nombre d'événement l'équation devient :

$P(y|x1, \dots, xn) = P(x1,\dots,xn|y)P(y) / P(x1,\dots,xn)$

Ici, *y* représente l'événement que nous cherchions à classer et *n* est le nombre de caractéristiques.
En appliquant cette approche, le modèle naïf bayésien prédit la classe qui présente la probabilité la plus élevée étant donné les caractéristiques observées.[4]

3.1.1.2. **Arbres de décisions:**

Les algorithmes J48 et Random Forest sont fréquemment utilisés pour construire des arbres de décision. Ils évaluent l'importance de chaque attribut en termes de sa capacité à séparer les données, en déterminant les seuils qui permettent de diviser les instances en différentes catégories cibles. Les arbres de décision résultants sont facilement interprétables, car ils fournissent une représentation explicite des règles de décision.

Dans le domaine de l'exploration de données, le logiciel Weka propose également un outil de visualisation des arbres de décision. La visualisation des données présente des avantages distincts par rapport à l'analyse textuelle. Elle peut gérer efficacement des données hétérogènes et bruitées, tout en offrant une approche intuitive qui ne nécessite pas de connaissances approfondies des paramètres, des mathématiques ou des statistiques complexes des algorithmes.

Ces algorithmes sont largement utilisés dans de nombreuses études portant sur la prédiction de défauts ou d'erreurs, où ils permettent d'analyser les caractéristiques des données pour identifier les facteurs prédictifs.[5,6,7]

3.1.1.3 **knn:**

L'algorithme de construction du KNN (K-Nearest Neighbors) se déroule généralement comme suit :

- Sélectionnez le nombre K de voisins.
- Pour chaque exemple dans l'ensemble de données :
- Calculez la distance entre l'exemple de requête et l'exemple actuel en utilisant les données.
- Ajoutez la distance et l'indice de l'exemple à une collection ordonnée.
- Triez cette collection de distances et d'indices dans l'ordre croissant des distances
- Sélectionnez les K premiers éléments de la collection.
- Attribuez l'exemple de requête à la classe où le nombre de voisins k est maximal (classe la plus fréquente).

Cet algorithme permet de prédire la classe d'un exemple de requête en se basant sur les classes des exemples les plus proches (les k voisins les plus proches).
La classe attribuée est celle qui est la plus représentée parmi les k voisins les plus proches

3.1.1.4 **La régression linéaire:**
La régression linéaire est une méthode statistique utilisée pour modéliser la relation entre une variable dépendante (la variable que l'on souhaite prédire) et une ou plusieurs variables indépendantes (les variables explicatives). L'objectif de la régression linéaire est de trouver une relation linéaire entre ces variables qui permet de prédire avec précision la variable dépendante.
La formulation de base de la régression linéaire est donnée par l'équation :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta p Xp + \varepsilon$$

où Y représente la variable dépendante,

X_1, X_2, ..., Xp représentent les variables indépendantes,
β_0, β_1, β_2, ..., βp sont les coefficients de régression à estimer,
et ε est le terme d'erreur qui capture l'écart entre la valeur prédite et la valeur réelle de la variable dépendante.

L'estimation des coefficients de régression se fait généralement en minimisant la somme des carrés des écarts entre les valeurs prédites et les valeurs réelles, à l'aide de la méthode des moindres carrés.
Une fois les coefficients estimés, on peut utiliser le modèle pour prédire la valeur de la variable dépendante en fonction des valeurs des variables indépendantes.

La régression linéaire peut être utilisée dans une variété de domaines, tels que l'économie, la finance, la sociologie, la biologie, etc. Elle permet de comprendre les relations entre les variables et de faire des prédictions sur la base de ces relations. De plus, la régression linéaire peut être étendue pour inclure des termes non linéaires, des interactions entre les variables, ou encore pour gérer des données catégorielles.

Bien que d'autres méthodes d'apprentissage statistique plus avancées existent, la régression linéaire reste populaire et largement utilisée en raison de sa simplicité, de son interprétabilité et de sa performance satisfaisante dans de nombreux cas. Elle peut également servir de point de départ pour explorer des méthodes plus complexes et avancées.

3.1.1.5. Régression logistique:
La régression logistique est devenue un outil important dans l'apprentissage du sujet automatique. Cette approche permet l'utilisation d'algorithmes dans des applications d'apprentissage Classez automatiquement les données entrantes par rapport aux données historiques.
Plus les données d'entrée sont pertinentes, meilleure est la capacité de l'algorithme à prédire la classification au sein de l'ensemble de données. La régression logistique ou modèle logit est un modèle Régression binomiale. Comme avec tous les modèles de régression binomiaux, ce sont meilleur modèle Un modèle mathématique simple utilisé pour un grand nombre d'observations réelles.
Autrement dit, relier une variable au vecteur de variables aléatoires (x1,…,xk) Un binôme aléatoire est généralement noté y. La régression logistique est un cas Particularités des modèles linéaires généralisés [21]

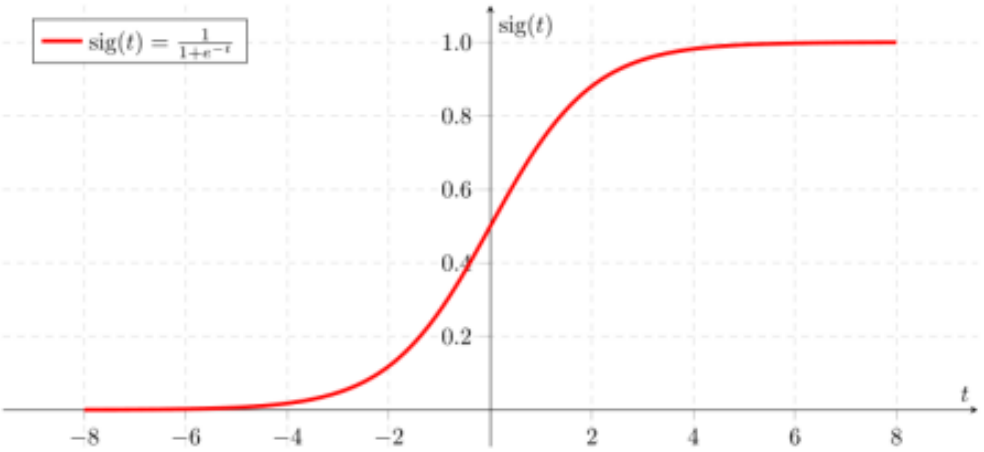


Figure 3 : Graphe et expression de la fonction sigmoïde [34]

3.1.1.6. forêt aléatoire (Random Forest):

Est un autre algorithme très couramment utilisé. L'algorithme construit plusieurs arbres de classification et de régression (CART, Classification and Regression Tree), chaque arbre est associé à différents scénarios et différentes variables initiales. L'algorithme est aléatoire, les données ne le sont pas. Ce type

d'algorithme est utilisé dans la classification prédictive et la modélisation de régression.

exemple : vous avez 1000 observations sur une population avec 10 variables. Pour construire le modèle CART à utiliser, l'algorithme Random Forest tire au hasard un échantillon de 100 observations et 5 variables aléatoires. L'algorithme répète ce processus plusieurs fois avant de faire une prédiction finale pour chaque

observation. La prédiction finale n'est qu'une fonction correspondant à la somme de différentes prédictions.

3.2. L’apprentissage non supervisé :

Ce type d'algorithme se focalise sur l'exploration de données non étiquetées en identifiant des motifs, des groupes ou des relations. Il est utilisé pour la segmentation de marché, l'analyse de sentiments et la détection d'anomalies.[35]

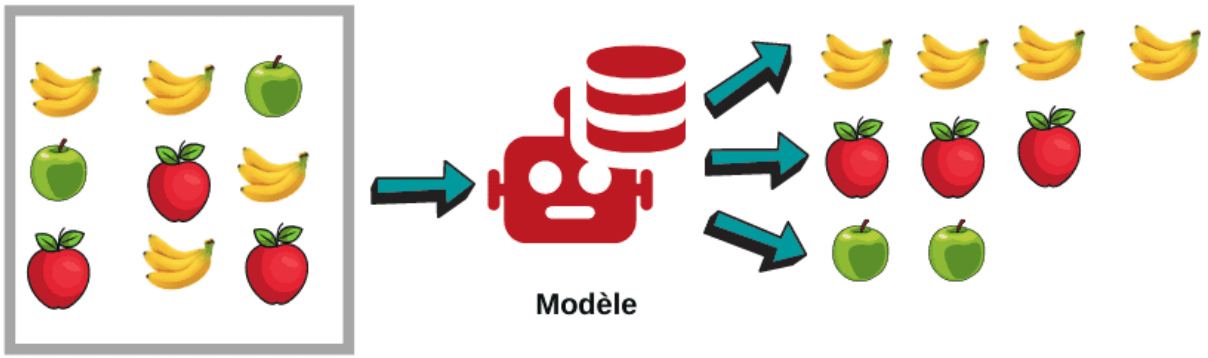


Figure 4 :Apprentissage non supervisé [35]

3.2.1.clustering:

Le clustering est le processus de séparation ou de division d'un ensemble de données en plusieurs ensembles de données.
Plusieurs groupes pour que les ensembles de données appartiennent au même groupe ensembles de données plus similaires que les autres groupes.
Tout simplement,L'objectif est de séparer les groupes ayant des caractéristiques similaires et de les affecter à grappe. [36]

3.2.1.1.K-Means:

Est un algorithme d'apprentissage automatique non supervisé pour résoudre les problèmes de clustering. Ils divisent et classent un ensemble de points données non étiquetées dans des groupes appelés “ clusters” .

Chaque itération de l'algorithme alloue à chaque Indique des groupes avec des caractéristiques similaires. Les points de données peuvent effectuer un suivi dans le temps pour détecter les changements qui se produisent dans le cluster.

L'algorithme K-Means peut confirmer les hypothèses sur les types de groupes présents dans des ensembles de données spécifiques ou utilisés pour découvrir des clusters inconnus. [35]

3.2.2 Avantages et inconvénients des algorithmes de machine learning:

Apprentissage	Type	Algo	Avantages	Inconvénients
Supervisé	**Classification**	**KNN**	Facile à implémenter. Efficace. [8] L'algorithme est polyvalent [9]	Calculer chaque fois la similarité entre les k. [10] Grande capacité de stockage. Utilise de nombreuses données de références pour classifier les nouvelles entrées. [8]
		SVM	Leur capacité à manipuler de grandes quantités de données Le faible nombre d'hyper paramètres. Elles sont bien fondées théoriquement. [11]	Complexes pour la classification des corpus. Demande un temps énorme pendant les phases de test. [12]
		Arbre de décision	Faciles à comprendre. Ils permettent de sélectionner l'option la plus appropriée parmi plusieurs. Il est facile de les associer à d'autres outils de prise de décision. [13]	Instables. [13] Certains concepts sont difficiles à exprimer à l'aide d'arbres de décision (comme XOR). [14]

		Naïve Bayes	La facilité et la simplicité de leur implémentation. Leur rapidité. Les méthodes Naïve Bayes donnent de bons résultats. [15]	Faire le même travail de classification. [16] [17]
	Régression	**Linéaire**	Simplicité d'interprétation. Facilité de calcul [17]	Elle ne traite pas les valeurs manquantes de variables continues sensible aux valeurs hors norme de variables continues [18]
Non Supervisé	**Clustering**	**K means**	Simple Flexible Efficace Complexité temporelle. [19]	Ensemble non optimal de clusters Manque de cohérence Limitation des calculs Spécifiez les valeurs k [19]

Tableau 1 : Avantages et inconvénients des algorithmes de machine learning

3.3.L'apprentissage semi supervisé :

Comme nous pouvons le deviner, l'apprentissage semi-supervisé implique l'apprentissage d'étiquettes à partir d'un ensemble de données partiellement étiqueté. L'un des avantages clés de cette approche est d'éviter d'étiqueter l'ensemble complet des exemples d'apprentissage. Cela s'avère particulièrement utile lorsque l'acquisition de données est facile, mais que l'étiquetage de ces données demande un effort considérable de la part des humains. [37]

3.4.L’apprentissage par renforcement :

Implique l'apprentissage par essais et erreurs. Un agent d'apprentissage interagit avec un environnement, prend des actions et reçoit des récompenses ou des punitions en fonction de ses performances. L'objectif de l'agent est de maximiser les récompenses sur le long terme en apprenant quelles actions sont les plus appropriées dans différentes situations. [37]

3.4.1.Quelques exemples de renforcement learning: [37]

- L'apprentissage par renforcement est utilisé dans de nombreux domaines tel que:
- La robotisation des usines et entrepôts de marchandises pour permettre aux automates d'apprendre par eux-mêmes à poser un nouveau modèle de pièce sans programmation préalable.
- La calibration et le contrôle qualité des systèmes industriels, qu'ils soient centrés sur la fabrication, la supply chain ou la production d'énergie.
- La finance pour optimiser le trading automatisé ou la gestion des risques de marché.
- La synthèse de texte pour estimer la qualité globale d'un résumé en s'extrayant d'une logique de mot à mot.
- Les jeux et les moteurs de recommandation pour développer des stratégies en environnement incertain.
- La voiture autonome pour améliorer la capacité du véhicule à réagir à tel ou tel événement de circulation.

4.Les éléments fondamentaux du ML:

Le Machine Learning repose sur deux piliers fondamentaux :

Premièrement, les données constituent les exemples sur lesquels l'algorithme va apprendre. Ce sont les données d'entraînement qui servent de base pour la création du modèle. Sans des données pertinentes et de qualité, aucun algorithme

d'apprentissage ne pourra produire un modèle efficace. Comme le dit l'Adage, "garbage in, garbage out", ce qui signifie qu'un algorithme d'apprentissage ne peut générer que des prédictions de mauvaise qualité si les données fournies sont de mauvaise qualité. [38]

Deuxièmement, l'algorithme d'apprentissage est la procédure utilisée pour traiter ces données et générer le modèle. L'entraînement consiste à exécuter cet algorithme sur un jeu de données spécifique. Il est essentiel de choisir un algorithme adapté aux données et au problème à résoudre.

Même avec des données de haute qualité, si l'algorithme d'apprentissage utilisé n'est pas approprié, le modèle résultant ne sera pas de bonne qualité.

Ainsi, les données et l'algorithme d'apprentissage sont tous les deux d'une importance considérable. Les données doivent être pertinentes et de qualité pour obtenir des résultats précis, tandis que l'algorithme d'apprentissage doit être adapté aux données afin de générer un modèle de haute qualité. [1]

5.Conclusion:

En résumé, le Machine Learning est une discipline puissante qui permet aux ordinateurs d'apprendre à partir de données et d'effectuer des tâches complexes de manière automatique. Grâce à ses nombreuses applications, il offre un potentiel énorme pour résoudre des problèmes et améliorer notre compréhension du monde qui nous entoure.

Chapitre IV: Implémentation

1.Introduction:

Dans cette phase d'implémentation, nous avons mis en pratique les concepts théoriques et les méthodes discutées dans les chapitres précédents de notre recherche. Notre objectif était de concrétiser ces idées en utilisant une approche basée sur l'apprentissage automatique (Machine Learning) à l'aide de bibliothèques Python et d'un environnement spécifique.

Nous avons utilisé des outils tels que pandas, scikit-learn (sklearn) et d'autres bibliothèques Python pour manipuler et prétraiter nos données.

Pour mettre en œuvre nos modèles d'apprentissage automatique, nous avons utilisé des bibliothèques spécialisées telles que scikit-learn, qui offrent une large gamme d'algorithmes d'apprentissage automatique pré-implémentés. Cela nous a permis d'entraîner et d'évaluer différents modèles en utilisant des techniques telles que la régression, la classification, etc.

2.Outils et Librairies utilisés:

Pandas: est une bibliothèque Python couramment utilisée pour la manipulation et l'analyse de données. Son principal avantage réside dans sa fonctionnalité de nettoyage des données, qui résout efficacement le problème de la préparation des données dans les projets d'apprentissage automatique. En effet, de nombreux ensembles de données disponibles contiennent des valeurs manquantes ou nulles, ce qui peut avoir un impact considérable sur la performance de notre modèle. Grâce à Pandas, nous pouvons facilement traiter ces valeurs problématiques et les remplacer par des données valides, ce qui facilite grandement le processus de préparation des données.

Scikit-learn: est une bibliothèque extrêmement populaire pour l'apprentissage automatique. Elle propose une multitude d'algorithmes et de fonctionnalités pour diverses tâches telles

que la sélection de modèles, le prétraitement des données, la validation croisée, l'évaluation des performances des modèles, et bien plus encore. Elle offre une mise en œuvre efficace des algorithmes d'apprentissage automatique, ce qui facilite la construction de pipelines de traitement des données. Grâce à scikit-learn, les utilisateurs peuvent bénéficier d'une large gamme d'outils pour développer des modèles d'apprentissage automatique de manière rapide et efficace.

NumPy: est une bibliothèque Python qui étend les fonctionnalités du langage en permettant la manipulation de tableaux multidimensionnels de manière efficace.

Matplotlib: est une bibliothèque complète en Python qui offre une large gamme de fonctionnalités pour la création de visualisations, que ce soit des graphiques statiques, des animations ou des visualisations interactives.

3.Environnement d'implémentation:

Dans le cadre de notre étude, nous avons opté pour l'utilisation de l'environnement de développement Google Colab.

3.1.Google Colab:

Google Colab ou Colaboratory est un service cloud, offert par Google (gratuit), basé sur Jupyter Notebook et destiné à la formation et à la recherche dans l'apprentissage automatique. Cette plateforme permet d'entraîner des modèles de Machine Learning directement dans le cloud. Sans donc avoir besoin d'installer quoi que ce soit sur notre ordinateur à l'exception d'un navigateur.[23]

3.2. Les avantages de l'utilisation de Google Colab pour notre implémentation :[39]

1. Accès gratuit aux ressources de calcul puissantes et à une mémoire suffisante.
2. Environnement de développement basé sur le cloud, offrant une accessibilité depuis n'importe où et à tout moment.
3. Interface conviviale et fonctionnalités intégrées facilitant l'utilisation de Google Colab.
4. Collaboration aisée avec d'autres utilisateurs grâce au partage de notebooks.
5. Intégration native avec d'autres services Google tels que Google Drive, permettant le stockage des données et les sauvegardes automatiques.
6. Communauté d'utilisateurs étendue, facilitant la résolution de problèmes et l'apprentissage des meilleures pratiques.
7. Expérimentation interactive de modèles d'apprentissage automatique grâce à l'utilisation de GPU et TPU disponibles sur Colab.

4.Définir l'ensemble des données utilisées

4.1. Collecte de données:

La période de stage que nous avons effectuée durant la période allant du 01/02/2023 au 02/03/2023 au niveau du service de prévention de la santé publique, avait pour but la collecte de données sur les maladies transmissibles et non transmissibles au niveau de la wilayas de Saida (commune de Saïda et ouled khaled).

Il importe de préciser que nos recherches nous ont permis de dégager des listes faisant ressortir les maladies à déclaration obligatoire (MDO) recensées durant la période de 2019 jusqu'à 2022, néanmoins ces listes comportent dans leur majorité des maladies transmissibles. ci dessous les données après numérisation :

index	Date	Commune	Age	Sex	Maladie	Meningite Germe	Meningite Evolution	TEP	Type de maladie
475	2021-11-28 00:00:00	saida	1	M	Mg LC	NaN	NaN	NaN	Mg
476	2021-11-28 00:00:00	saida	46	F	VIH/SIDA	NaN	NaN	NaN	VIH/SIDA
477	2021-11-08 00:00:00	saida	23	F	TBC EP NPRV	NaN	NaN	gang	TBC
478	2021-11-09 00:00:00	ouled khaled	42	F	TBC EP PRV	NaN	NaN	Mastite	TBC
479	2021-11-10 00:00:00	saida	38	F	TBC EP NPRV	NaN	NaN	Multiviscérale	TBC
480	2021-11-11 00:00:00	ouled khaled	28	M	TBC EP PRV	NaN	NaN	gang	TBC
481	2021-11-12 00:00:00	ouled khaled	15	M	TBC EP PRV	NaN	NaN	gang	TBC
482	2021-11-13 00:00:00	saida	37	F	TBC EP PRV	NaN	NaN	gang	TBC
483	2021-11-14 00:00:00	saida	9	F	Mg P	NaN	NaN	NaN	Mg
484	2021-11-15 00:00:00	saida	37	M	TBC EP PRV	NaN	NaN	Urinaire	TBC
485	2021-11-16 00:00:00	saida	28	F	TBC P M+	NaN	NaN	NaN	TBC
486	2021-11-17 00:00:00	ouled khaled	32	F	TBC EP NPRV	NaN	NaN	Péritonéale	TBC
487	2021-11-18 00:00:00	ouled khaled	23	F	TBC P M+	NaN	NaN	NaN	TBC
488	2021-11-19 00:00:00	saida	73	F	TBC EP NPRV	NaN	NaN	gang	TBC
489	2021-11-20 00:00:00	ouled khaled	40	F	TBC EP NPRV	NaN	NaN	gang	TBC
490	2021-11-21 00:00:00	saida	44	F	TBC EP PRV	NaN	NaN	gang	TBC
491	2021-11-22 00:00:00	saida	50	M	TBC EP NPRV	NaN	NaN	Multiviscérale	TBC
492	2021-11-23 00:00:00	ouled khaled	24	F	TBC P M+	NaN	NaN	NaN	TBC
493	2021-11-24 00:00:00	ouled khaled	10	M	TBC EP PRV	NaN	NaN	gang	TBC
494	2021-11-28 00:00:00	saida	1	M	Mg LC	NaN	NaN	NaN	Mg
495	2021-11-28 00:00:00	saida	46	F	VIH/SIDA	NaN	NaN	NaN	VIH/SIDA
496	2021-12-12 00:00:00	saida	3	M	Mg P	NaN	NaN	NaN	Mg

Figure 5: Aperçu du dataset des MDO de la wilayas de saida.

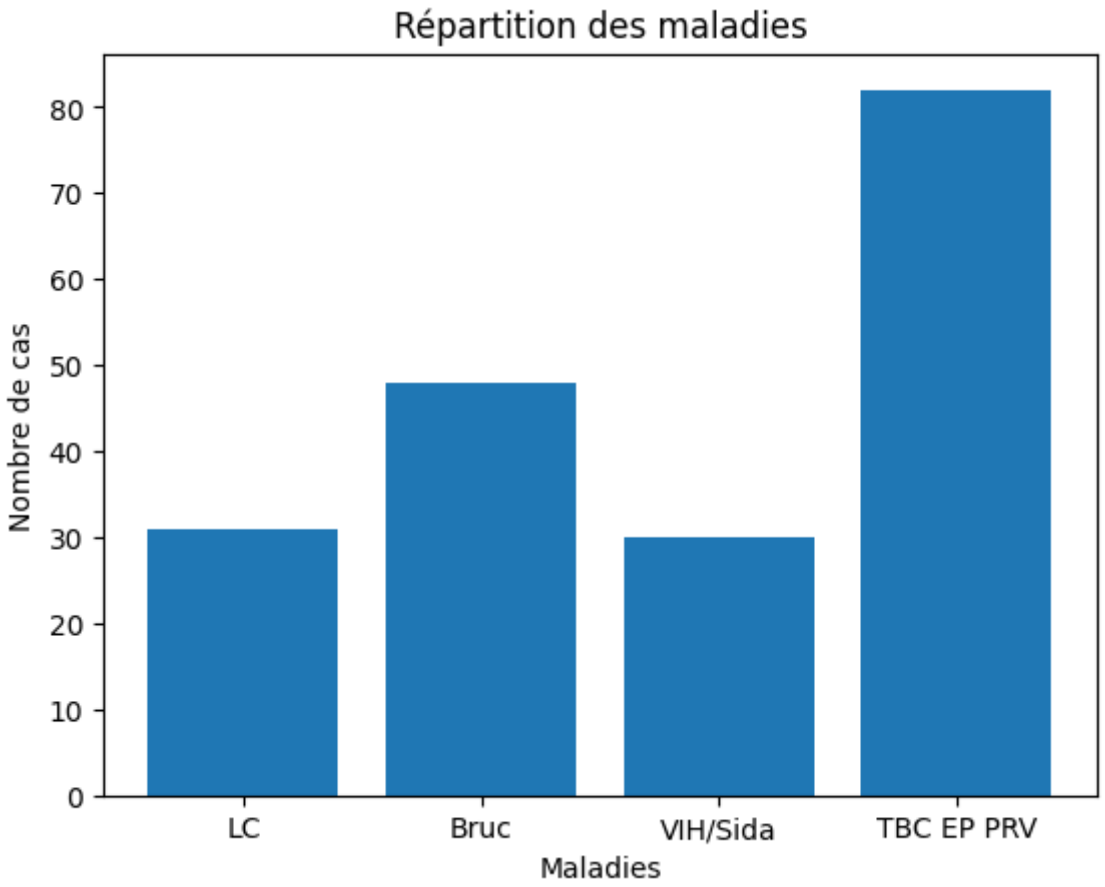


Figure 6: graphes montrant la répartition des maladies sur le nombre de cas

Après l'étape du prétraitement, dans laquelle nous avons effectué les étapes suivantes:

1- Lecture du fichier Excel : on utilise la bibliothèque pandas pour lire le fichier Excel et stocker les données dans un DataFrame appelé 'df'.

2- Remplissage des valeurs manquantes : utilisant la méthode fillna() pour remplir les valeurs manquantes dans les colonnes 'Méningite Germe', 'TEP', 'Meningite Evolution' et 'Germe' avec la valeur '0'.

3- Suppression de la colonne 'Date' : avec la méthode drop() pour supprimer la colonne 'Date' du DataFrame 'df'.

4- Encodage des variables catégorielles : on utilise la classe LabelEncoder de la bibliothèque scikit-learn pour encoder les variables catégorielles en valeurs numériques. Les colonnes 'Commune', 'Sex' et 'Maladie' sont encodées à l'aide de cet encodeur.

5- Séparation des variables indépendantes et de la variable cible :on divise les données en deux ensembles distincts : les variables indépendantes (X) et la variable cible (y). Les caractéristiques (X) sont toutes les colonnes du DataFrame 'df' sauf la colonne 'Maladie', qui est la variable cible.

6- Encodage des variables catégorielles dans les caractéristiques : Les mêmes étapes d'encodage des variables catégorielles sont répétées pour les colonnes 'Commune' et 'Sex' dans l'ensemble de caractéristiques (X).

7- Séparation des ensembles d'entraînement et de test : on utilise la fonction train_test_split de scikit-learn pour diviser les ensembles de caractéristiques (X) et de variable cible (y) en ensembles d'entraînement et de test. L'ensemble de test est défini à 20% de l'ensemble de données total, et le paramètre random_state est utilisé pour garantir la reproductibilité des résultats.

On a ensuite appliqué les méthodes de machine learning suivante sur ces données et voici leur résultats:

	Algorithm	Precision	Recall	Accuracy	F1 Score
0	Random Forest	0.422470	0.404959	0.404959	0.409950
1	Decision Tree	0.380666	0.355372	0.355372	0.355115
2	KNN	0.419181	0.396694	0.396694	0.398469
3	Logistic Regression	0.402360	0.479339	0.479339	0.419513

Figure 7: Algorithmes utilisés et leur résultat (dataset MDO SAIDA)

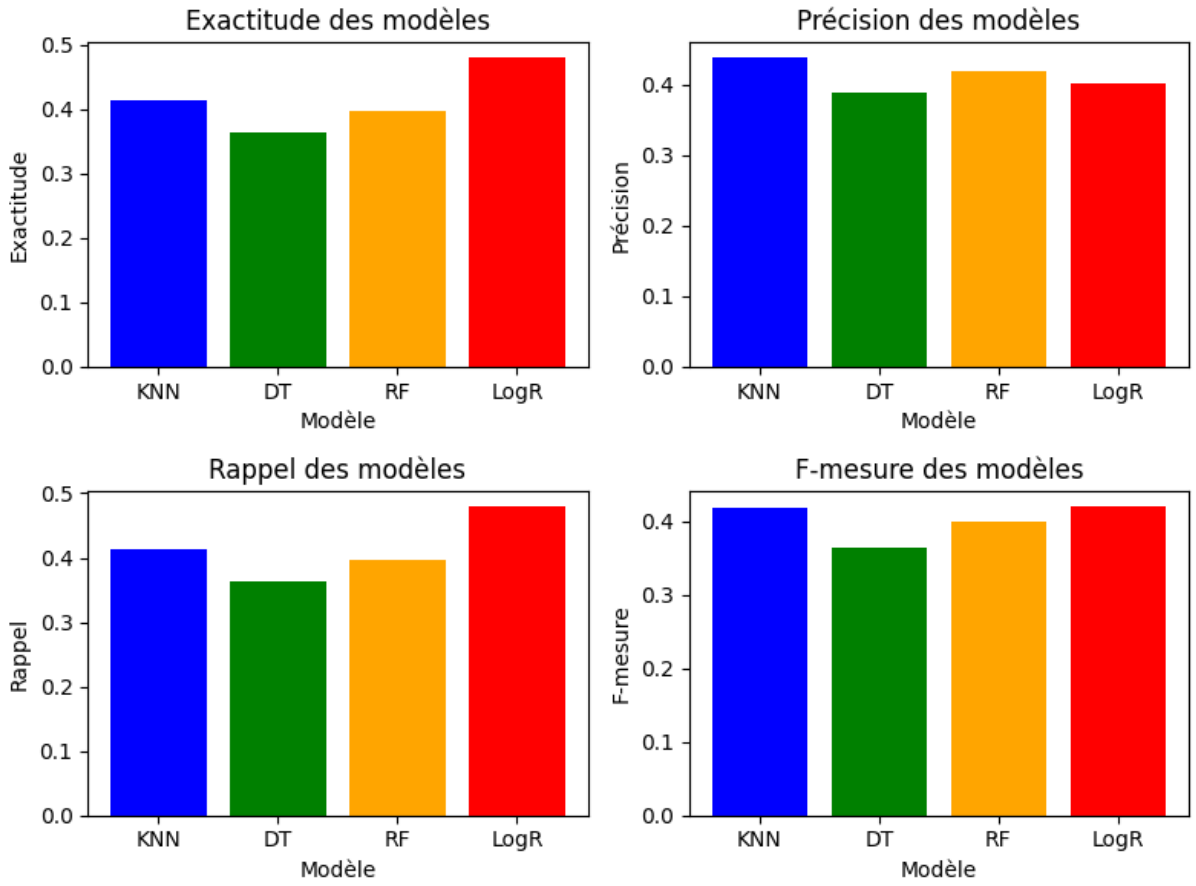


Figure 8: graphe à barres montrant les performances des méthodes de ML utiliser

Les méthodes citées supra nous ont conduit à des résultats très faibles due au manque de caractéristique nécessaire à une prédiction fiable.

Force est de constater que les données recueillies ne sont pas d'une grande utilité, on s'est orienté vers des données disponibles en ligne.

4.2.maladies transmissibles(dataset hépatite c):

	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
0	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	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	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	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	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...
610	3=Cirrhosis	62	f	32.0	416.6	5.9	110.3	50.0	5.57	6.30	55.7	650.9	68.5
611	3=Cirrhosis	64	f	24.0	102.8	2.9	44.4	20.0	1.54	3.02	63.0	35.9	71.3
612	3=Cirrhosis	64	f	29.0	87.3	3.5	99.0	48.0	1.66	3.63	66.7	64.2	82.0
613	3=Cirrhosis	46	f	33.0	NaN	39.0	62.0	20.0	3.56	4.20	52.0	50.0	71.0
614	3=Cirrhosis	59	f	36.0	NaN	100.0	80.0	12.0	9.07	5.30	67.0	34.0	68.0

615 rows × 13 columns

Figure 9: Aperçu Dataset d'hépatite C.

Ce dataset a été obtenu à partir du référentiel d'apprentissage automatique de l'UCI. l'ensemble de données contient les valeurs de laboratoire des donneurs de sang et des patients atteints d'hépatite C et de valeurs démographiques comme l'âge …

Caractéristiques de l'ensemble de données :	Multivarié	**Nombre d'instances :**	615	**Zone:**	Vie
Caractéristiques des attributs :	Entier, Réel	**Nombre d'attributs :**	14	**Date du don**	2020-06-10
Tâches associées :	Classification, regroupement	**Valeurs manquantes ?**	Oui	**Nombre de visites Web :**	93236

Tableau 2: Caractéristiques du dataset Hépatite C

l'attribut cible pour la classification est la catégorie (donneur de sang par rapport (y compris sa progression)).[25]

4.2.1.Informations sur les attributs :

Tous les attributs sauf Catégorie et Sexe sont numériques. Les données de laboratoire sont les attributs de 1-13.

1. Category : Une variable catégorique indiquant la catégorie ou le groupe auquel chaque enregistrement appartient. (valeurs : '0=Donneur de sang', '0s=Donneur de sang suspect', '1=Hépatite', '2=Fibrose', '3=Cirrhose ')
2. Age : L'âge du patient ou de l'individu associé à l'enregistrement.
3. ALB : La valeur associée à l'albumine.
4. ALP : La valeur associée à l'alcaline phosphatase.
5. ALT : La valeur associée à l'alanine aminotransférase.
6. AST : La valeur associée à l'aspartate aminotransférase.
7. BIL : La valeur associée à la bilirubine.
8. CHE : La valeur associée à la cholinestérase.
9. CHOL : La valeur associée au cholestérol.
10. CREA : La valeur associée à la créatine.
11. GGT : La valeur associée à la gamma-glutamyl transférase.
12. PROT : La valeur associée à la protéine totale.
13. Sexe : Une variable binaire indiquant le sexe de l'individu (1 pour masculin, 0 pour féminin).

4.3.maladies non transmissibles(maladies cardiovasculaires):

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

Figure 10:Aperçu du Dataset des MCV

Ce dataset, datant de 1988, comprend quatre bases de données distinctes : Cleveland, Hongrie, Suisse et Long Beach V. Il comprend initialement 76 attributs, y compris l'attribut cible, mais les études publiées sur ce dataset se concentrent sur l'utilisation d'un sous-ensemble de 14 attributs spécifiques. L'attribut cible fait référence à la présence d'une maladie cardiaque chez le patient, où la valeur 0 indique l'absence de maladie et la valeur 1 indique la présence de maladie.

Ce dataset regroupe des données provenant de différentes sources et contient des informations sur des patients, en mettant l'accent sur la prédiction de la présence ou de l'absence de maladie cardiaque. Les 14 attributs sélectionnés sont utilisés dans les études pour tenter de prédire cette condition médicale.[24]

4.3.1.Informations sur les attributs :

1. âge
2. sexe
3. type de douleur thoracique (4 valeurs)
4. tension artérielle au repos
5. cholesterol en mg/dl
6. glycémie à jeun > 120 mg/dl

7. résultats électrocardiographiques au repos (valeurs 0,1,2)
8. fréquence cardiaque maximale atteinte
9. angine d'effort
10. oldpeak = dépression ST induite par l'exercice par rapport au repos
11. la pente du segment ST d'effort maxima
12. nombre de vaisseaux principaux (0-3) colorés par fluoroscopie

tal : 0 = normal ; 1 = défaut corrigé ; 2 = défaut réversible

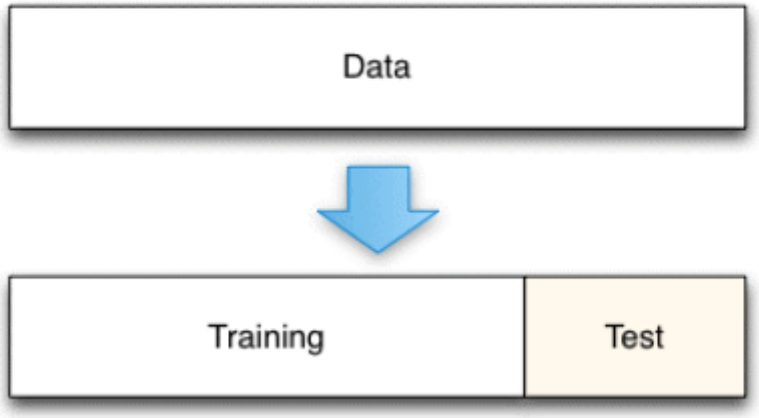
Les noms et numéros de sécurité sociale des patients ont été récemment supprimés de la base de données, remplacés par des valeurs fictives.

5.Étapes de prétraitement des données:

5.1.dataset hépatite(maladie transmissible):

1. Extraction du premier caractère de la colonne "Category" pour réduire les classes de prédictions à une seule lettre.
2. Suppression des lignes contenant des valeurs manquantes (NaN).
3. Réinitialisation des index des lignes pour éviter les incohérences.
4. Encodage binaire des colonnes catégorielles à l'aide de la fonction pd.get_dummies(). Cela permet de convertir les variables catégorielles en variables binaires (0 ou 1) pour les utiliser dans les modèles d'apprentissage automatique.
5. Création d'une instance du scalateur MinMaxScaler pour normaliser les colonnes sélectionnées. Le MinMaxScaler transforme les valeurs de chaque colonne en les mettant à l'échelle dans un intervalle spécifié, généralement entre 0 et 1.
6. Application de la normalisation Min-Max sur les colonnes sélectionnées du DataFrame encodé.

7. Séparation des ensembles de données en ensembles d'entraînement et de test à l'aide de la fonction train_test_split().



Cela divise les données en deux ensembles, un pour l'entraînement du modèle et l'autre pour évaluer les performances du modèle.

	Category	Age	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT	Sex_m
0	0	0.166667	0.350669	0.101653	0.020962	0.036694	0.032181	0.367578	0.218447	0.091495	0.011757	0.580336	1
1	0	0.166667	0.350669	0.145571	0.052713	0.044990	0.014890	0.650434	0.408981	0.061619	0.017172	0.760192	1
2	0	0.166667	0.475483	0.156427	0.108816	0.134014	0.025456	0.494997	0.457524	0.072822	0.044400	0.827338	1
3	0	0.166667	0.420505	0.100419	0.091554	0.038290	0.086936	0.394263	0.401699	0.067221	0.045328	0.741007	1
4	0	0.166667	0.361070	0.154947	0.097719	0.045310	0.042267	0.515677	0.350728	0.063486	0.039295	0.573141	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
584	3	0.648148	0.283804	0.086603	0.043465	0.444799	0.034582	0.322882	0.309466	0.044814	0.069926	0.858513	0
585	3	0.666667	0.358098	0.098692	0.057645	0.878111	0.188280	0.290193	0.373786	0.119597	0.149443	0.616307	0
586	3	0.722222	0.254086	1.000000	0.015413	0.318124	0.236311	0.276851	0.591019	0.044534	1.000000	0.568345	0
587	3	0.759259	0.135215	0.225759	0.006165	0.107849	0.092219	0.008005	0.192961	0.051349	0.048577	0.635492	0
588	3	0.759259	0.209510	0.187515	0.008015	0.282068	0.226705	0.016011	0.266990	0.054803	0.092358	0.892086	0

589 rows × 13 columns

Figure 11: Aperçu du dataset d'hépatite C aprés près-traiter et normaliser

5.2.Dataset MCV:

1. Extraction du premier caractère de la colonne "target" :

Convertit la colonne "target" en chaîne de caractères, puis extrait le premier caractère de chaque valeur et le remplace dans la colonne "target".

2. Suppression des lignes contenant des valeurs manquantes (NaN)

Supprime toutes les lignes qui contiennent des valeurs manquantes (NaN) dans le dataframe "data".

3. Réinitialisation des index des lignes

Réinitialise les index des lignes du dataframe "data" de manière séquentielle, en supprimant les anciens index.

4. Encodage binaire des colonnes catégorielles
5. Effectue un encodage binaire des colonnes catégorielles du dataframe "data" à l'aide de la fonction get_dummies() de pandas. Cela crée de nouvelles colonnes binaires pour chaque valeur unique dans les colonnes catégorielles.
6. Création de l'instance du MinMaxScaler :

Créer une instance de la classe MinMaxScaler, qui sera utilisée pour normaliser les données.

7. Colonnes sélectionnées pour la normalisation :

Définit une liste de noms de colonnes à normaliser.

8. Normalisation des colonnes sélectionnées :

Applique la normalisation Min-Max (mise à l'échelle) aux colonnes sélectionnées du dataframe "data_encoded" en utilisant la méthode fit_transform() de l'instance du MinMaxScaler. Les valeurs normalisées sont ensuite assignées aux mêmes colonnes dans le dataframe "data_encoded".

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target_0	target_1
0	0.479167	1	0	0.292453	0.196347	0	1	0.740458	0	0.161290	2	2	3	1	0
1	0.500000	1	0	0.433962	0.175799	1	0	0.641221	1	0.500000	0	0	3	1	0
2	0.854167	1	0	0.481132	0.109589	0	1	0.412214	1	0.419355	0	0	3	1	0
3	0.666667	1	0	0.509434	0.175799	0	1	0.687023	0	0.000000	2	1	3	1	0
4	0.687500	0	0	0.415094	0.383562	1	1	0.267176	0	0.306452	1	3	2	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1020	0.625000	1	1	0.433962	0.216895	0	1	0.709924	1	0.000000	2	0	2	0	1
1021	0.645833	1	0	0.292453	0.301370	0	0	0.534351	1	0.451613	1	1	3	1	0
1022	0.375000	1	0	0.150943	0.340183	0	0	0.358779	1	0.161290	1	1	2	1	0
1023	0.437500	0	0	0.150943	0.292237	0	0	0.671756	0	0.000000	2	0	2	0	1
1024	0.520833	1	0	0.245283	0.141553	0	1	0.320611	0	0.225806	1	1	3	1	0

Figure 12: Aperçu du dataset des MCV après près-traiter et normaliser

6.Application des modèles :

6.1.Hépatite C (transmissible):

Méthode	Type	Hyperparamètre
Knn	**Classification**	**K= 5**
Forêt aléatoire	**Classification**	**n_estimators=100, random_state=42**
Arbres de décision	**Classification**	**random_state=42**
Régression Logistique	**Classification**	**default paramètres**
Naïve Bayes	**Classification**	**default paramètres**
SVM	**Classification**	**default paramètres**

Tableau 3: Méthodes de machine learning Appliquées sur le dataset (hépatite C)

6.2.MCV (non transmissible):

Méthode	Type	Hyperparamètre
Kppv	**Classification**	**K= 5**
Forêt aléatoire	**Classification**	**n_estimators=100, random_state=42**
Arbres de décision	**Classification**	**random_state=42**
Régression Logistique	**Classification**	**default paramètres**
Naïve Bayes	**Classification**	**default paramètres**
svm	**Classification**	**default paramètres**

Tableau 4: Méthodes de machine learning Appliquées sur le dataset (MCV)

7.Résultat:

7.1.Hépatite C:

Modèle	Exactitude	Précision	Rappel	F-mesure	moyenne
KPPV	0.915	0.871	0.915	0.891	0.895
Régression Logistique	0.898	0.871	0.898	0.875	0.881
Arbre de décision	0.915	0.891	0.915	0.894	0.903
orêt aléatoire	0.932	0.927	0.932	0.926	0.929
Naive bayes	0.872881	0.854116	0.872881	0.861147	0.865
SVM	0.898305	0.855869	0.898305	0.875815	0.882

Tableau 5: Résultat des performances des modèles du ML sur le dataset (hépatite C)

Dans ce tableau nous avons les résultats des performances des modèles de ML appliquées sur le dataset.

en calculant la moyenne des résultats des performances nous déduisant que les algorithmes Random forest et Decision Tree ont les meilleurs performances pour la prédiction de l'Hépatite C avec les résultats respectives suivant: 92% et 90%

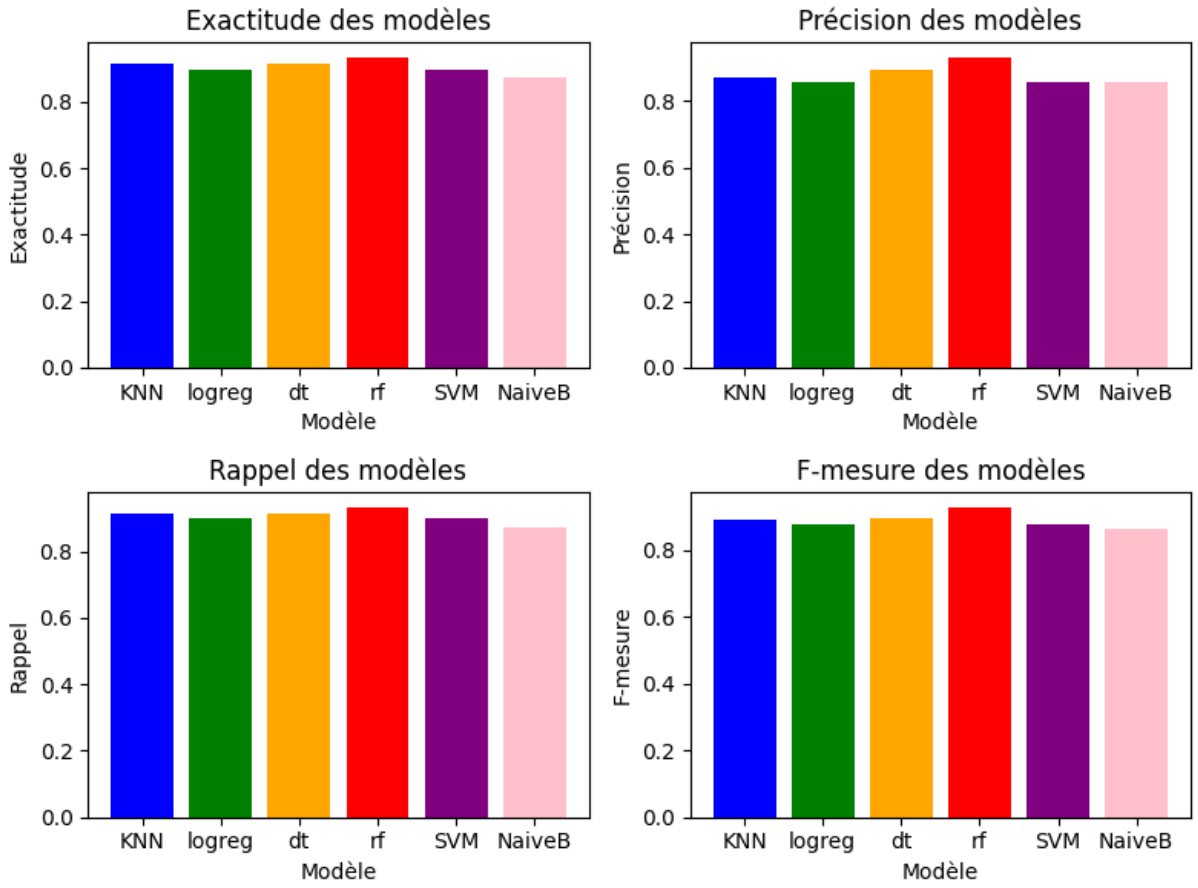


Figure 13: graphe en barres horizontales montrant les Résultats des performances des modèles de ML utilisées sur le dataset d'Hépatite C

7.2.MCV:

Modèle	Exactitude	Précision	Rappel	F-mesure	moyenne
KPPV	0.858	0.858	0.858	0.858	0.858
Arbre de décision	0.985	0.985	0.985	0.985	0.985
Forêt aléatoire	0.985	0.985	0.985	0.985	0.985
Régression Logistique	0.932	0.927	0.932	0.926	0.929
Naive bayes	0.868	0.870	0.868	0.868	0.868
SVM	0.808	0.860	0.868	0.846	0.845

Tableau 6: Résultat des performances des modèles du ML sur le dataset (MCV)

Dans ce tableau nous avons les résultats des performances des modèles de ML appliquées sur le dataset, en calculant la moyenne des résultats des performances nous déduisant que les algorithmes Random forest et Decision Tree ont les meilleures performances pour la prédiction de l'Hépatite C avec 98%.

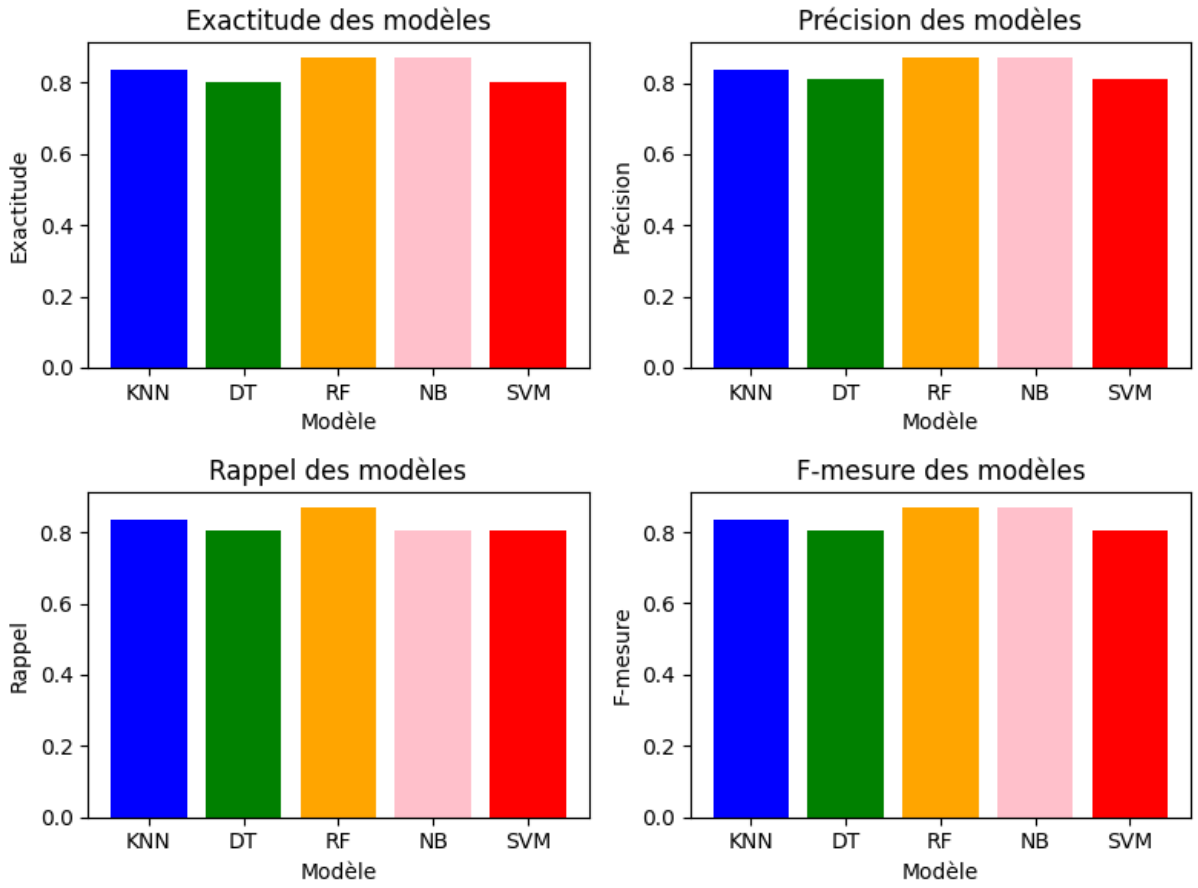


Figure 14: graphe en barres horizontales montrant les scores des performances des méthodes du ML utilisé sur le dataset (MCV)

8.Explicabilité du modèle :

L'un des défis d'un projet de machine leaning est d'expliquer la prédiction du modèle. Un modèle peut considérer certaines caractéristiques plus importantes que d'autres pour sa prédiction. Un autre modèle pourrait peser d'autres caractéristiques comme plus importantes.Permutation importance et SHAP sont deux méthodes que l'on peut utiliser pour comprendre quelles caractéristiques ont été sélectionnées pour avoir le plus d'impact sur la prédiction de notre modèle.

8.1. Permutation Importance:

La permutation importance, comme mentionnée précédemment, évalue l'importance de chaque caractéristique en mesurant la diminution de la performance du modèle lorsque les valeurs de cette caractéristique sont mélangées aléatoirement. Une caractéristique qui a un impact important sur les prédictions du modèle entraînera une plus grande baisse de performance lorsqu'elle est mélangée, indiquant ainsi son importance.

8.2. SHAP:

Quant à SHAP, il s'agit d'une méthode basée sur la théorie des jeux qui attribue une valeur d'importance à chaque caractéristique en calculant la contribution marginale de cette caractéristique à la prédiction finale. SHAP mesure l'importance d'une caractéristique en considérant toutes les combinaisons possibles de caractéristiques et en évaluant comment la présence ou l'absence d'une caractéristique affecte la prédiction.

Les deux méthodes fournissent des mesures d'importance des caractéristiques, mais elles diffèrent dans leur approche et leur calcul. La permutation importance évalue l'importance relative des caractéristiques en les mélangeant aléatoirement, tandis que SHAP attribue des valeurs d'importance en utilisant des concepts de la théorie des jeux.

L'utilisation de ces deux méthodes peut aider à obtenir une compréhension plus complète des caractéristiques qui influencent le plus les prédictions du modèle. Elles permettent de hiérarchiser les caractéristiques en fonction de leur impact, de détecter les caractéristiques clés et d'expliquer les prédictions du modèle de manière plus détaillée et interprétable.

8.2.1. Résultat SHAP pour les MCV:

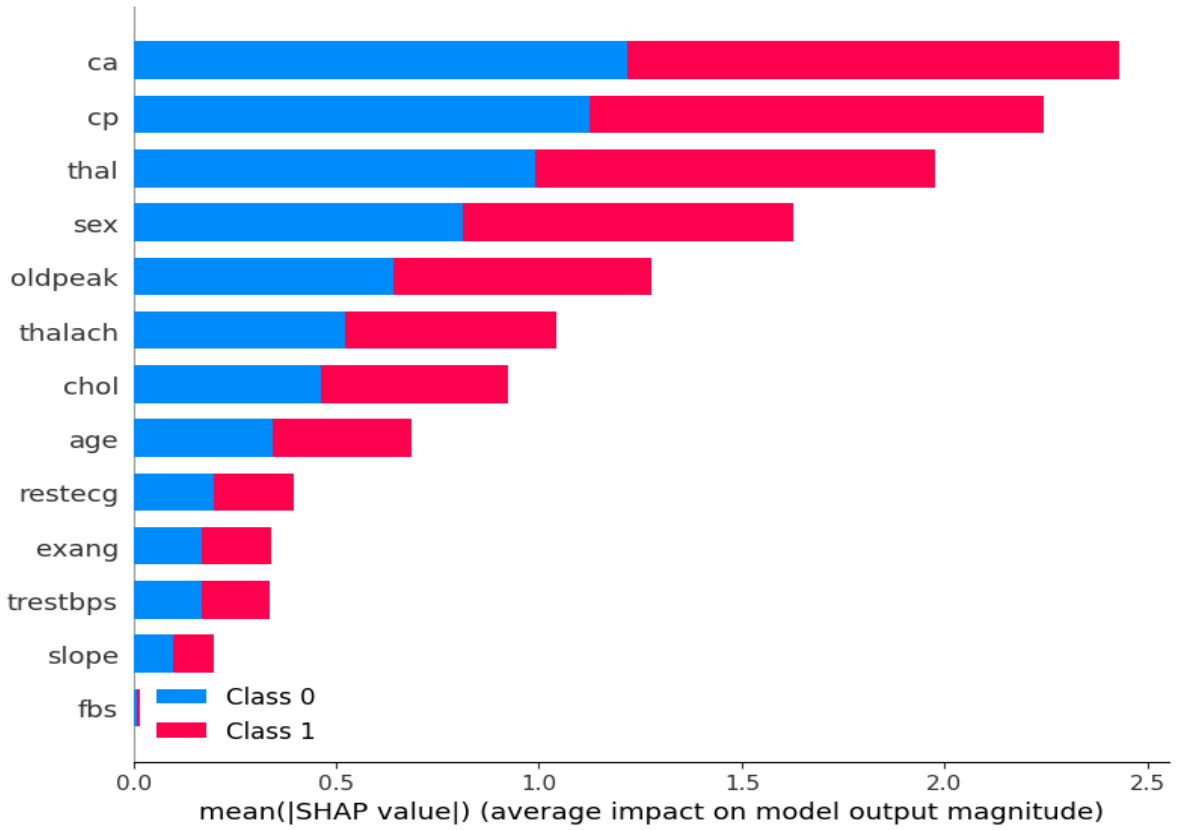


Figure 15: graphe en barres verticales montrant les résultats SHAP pour les MCV

8.2.1. Résultat SHAP pour les HEPATITE C:

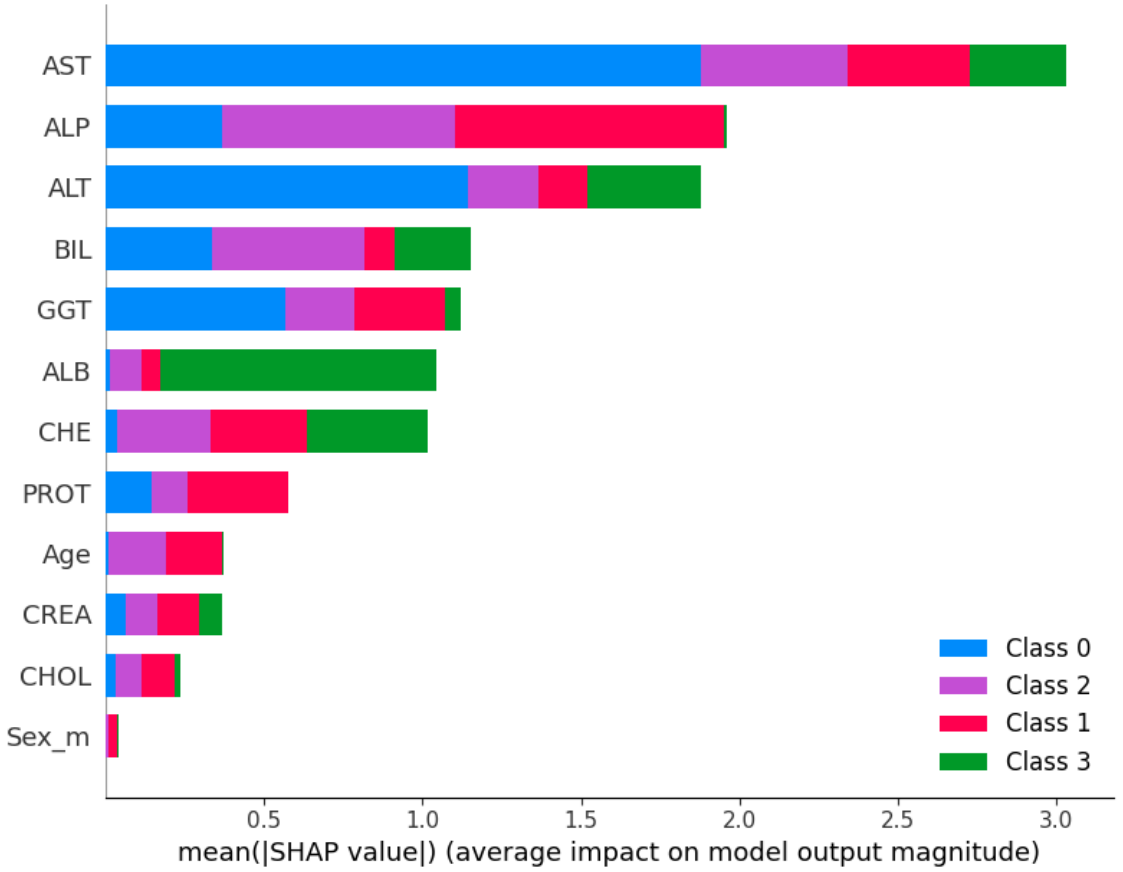


Figure 16: graphe en barres verticales montrant les résultats SHAP pour hépatite C

Conclusion:

Dans ce chapitre, nous avons présenté les différentes étapes de pré traitement des données tels que l'exploration et la visualisation des données ainsi que le nettoyage des valeurs aberrantes et la normalisation. L'application des méthodes d'évaluation nous a permis de sélectionner le modèle Random forest comme le meilleur modèle pour la prédiction des MCV et le Random forest et decision tree comme meilleurs modèles pour la prédiction d'Hépatite C. A la fin, nous avons extrait la caractéristique dominante pour les les modèles utilisés.

Conclusion générale:

Les maladies cardiovasculaires sont un groupe de conditions médicales qui affectent le cœur et les vaisseaux sanguins. Elles sont l'une des principales causes de décès dans le monde, représentant un fardeau considérable pour la santé publique.

L'hépatite C est une maladie infectieuse causée par le virus de l'hépatite C (VHC). Elle affecte principalement le foie et peut entraîner une inflammation chronique, une cirrhose et, dans certains cas, un cancer du foie. L'hépatite C est considérée comme un problème de santé mondial, touchant des millions de personnes dans le monde.

Le machine learning peut apporter une valeur ajoutée en aidant à la prédiction, au diagnostic précoce et à la gestion des maladies.

Dans ce mémoire nous avons utilisé les méthodes de machine learning pour faire une prédiction des maladies citées plus haut au vu de leur gravité et l'importance de la détection précoce qui peut sauver la vie d'un patient et dans lesquelles nous avons constatées que le Random Forest et Decision tree sont les meilleurs méthodes avec les scores les plus élevés.

Toutefois le manque de données a été un ralentissant dans ce projet, car il est difficile de trouver des données exploitables.

En conclusion, les maladies épidémiologiques représentent un défi majeur pour la santé publique, mais le machine learning offre des opportunités pour améliorer la prévention, le dépistage et la gestion de ces affections. En combinant les connaissances médicales avec les capacités avancées d'analyse des données, le machine learning peut contribuer à réduire le fardeau des épidémies, améliorer les résultats pour les patients et promouvoir une meilleure santé au niveau individuel et de la population.

Bibliographie

1. Bayliss, L., & Jones, L. D. (2019). The role of artificial intelligence and machine learning in predicting orthopedic outcomes. The bone & joint journal, 101(12), 1476-1478.
2. Ma, Y. (2020). Machine Learning Based Applications for Data Visualization, Modeling, Control, and Optimization for Chemical and Biological Systems. Louisiana State University and Agricultural & Mechanical College.
3. Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT), 48(3), 128-138.
4. Moeyersoms, J., de Fortuny, E. J., Dejaeger, K., Baesens, B., & Martens, D. (2015). Comprehensible software fault and effort prediction: A data mining approach. Journal of Systems and Software, 100, 80-90.
5. Gyimóthy, T., Ferenc, R., & Siket, I. (2005). Empirical validation of object-oriented metrics on open source software for fault prediction. IEEE Transactions on Software engineering, 31(10), 897-910.
6. Kaur, A., & Kaur, K. (2014, September). Performance analysis of ensemble learning for predicting defects in open source software. In 2014 international conference on advances in computing, communications and informatics (ICACCI) (pp. 219-225). IEEE.
7. Mezili, H. (2021). Vers une amélioration de la détection d'intrusion par les méthodes de sélection des fonctionnalités à l'aide des arbres de décision (Doctoral dissertation, Université Ibn Khaldoun-Tiaret-).
8. Chen, L. P. (2019). Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar: Foundations of machine learning: The MIT Press, Cambridge, MA, 2018, 504 pp., CDN $96.53 (hardback), ISBN 9780262039406.
9. Quang, C. T. (2005). Classification automatique des textes vietnamiens Hanoi. Institut de la Francophonie pour l' informatique.
10. Zhao, X., & Kuh, A. (2002, November). Adaptive kernel least square support vector machines applied to recover DS-CDMA signals. In Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, 2002. (Vol. 1, pp. 943-947). IEEE.
11. Hasan, M., & Boris, F. (2006). Svm: Machines à vecteurs de support ou séparateurs à vastes marges. Rapport technique, Versailles St Quentin, France. Cité, 64.
12. Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning. Springer series in statistics. New York, NY, USA.

13. Mezili, H. (2021). Vers une amélioration de la détection d'intrusion par les méthodes de sélection des fonctionnalités à l'aide des arbres de décision (Doctoral dissertation, Université Ibn Khaldoun-Tiaret-).
14. Mezili, H. (2021). Vers une amélioration de la détection d'intrusion par les méthodes de sélection des fonctionnalités à l'aide des arbres de décision (Doctoral dissertation, Université Ibn Khaldoun-Tiaret-).
15. LAHRACHE, F. Classification des textes prophétiques (Doctoral dissertation, Faculté des mathématiques et de l'informatique-Université Mohamed Boudiaf-M'SILA).

16. Zeitouni, K. (2006). Analyse et extraction de connaissances des bases de données spatio-temporelles (Doctoral dissertation, Université de Versailles-Saint Quentin en Yvelines).
17. Alouaoui, H., Turki, S. Y., & Faiz, S. (2017). Knowledge extraction from geographical databases for land use data production. In Handbook of Research on Geographic Information Systems Applications and Advancements (pp. 321-343). IGI Global.
18. Stéphane, T. (2012). Data mining et statistique décisionnelle: l'intelligence des données. Editions Technip.
19. Hilali, H. (2009). Application de la classification textuelle pour l'extraction des règles d'association maximales (Doctoral dissertation, Université du Québec à Trois-Rivières).
20. Mezili, H. (2021). Vers une amélioration de la détection d'intrusion par les méthodes de sélection des fonctionnalités à l'aide des arbres de décision (Doctoral dissertation, Université Ibn Khaldoun-Tiaret-).
21. Im, V., & Briex, M. (2020). Médecine prédictive, deep learning, algorithmes et accouchement. Spirale, (1), 204-209.

22. Bahado-Singh, R. O., Vishweswaraiah, S., Aydas, B., Yilmaz, A., Saiyed, N. M., Mishra, N. K., ... & Radhakrishna, U. (2022). Precision cardiovascular medicine: artificial intelligence and epigenetics for the pathogenesis and prediction of coarctation in neonates. The Journal of Maternal-Fetal & Neonatal Medicine, 35(3), 457-464.

23. Henri Michel.(2021). Google Colab : Le guide Ultime. https://ledatascientist.com/google-colab-le-guide-ultime/

24. SVETLANA ULIANOVA. (2019).Cardiovascular Disease dataset. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

25. HCV data. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5D612

26. Giroux, É. (2012). De l'épidémiologie de santé publique à l'épidémiologie clinique. Quelques réflexions sur la relation entre épidémiologie et clinique (1920-1980). Bulletin d'histoire et d'épistémologie des sciences de la vie, 19(1), 21-43..

27. Brault, N. (2017). Le concept de biais en épidémiologie (Doctoral dissertation, Université Sorbonne Paris Cité).

28. Gérin, M., & Gosselin, P. La référence bibliographique de ce document se lit comme suit: Bouyer J, Cordier S, Levallois P (2003) Épidémiologie. In: Environnement et santé publique-Fondements et.

29. Porta, M. (Ed.). (2008). A dictionary of epidemiology. Oxford university press.

30. Porta, M. (Ed.). (2014). A dictionary of epidemiology. Oxford university press.

31. Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.".

32. Vasilev, I., Slater, D., Spacagna, G., Roelants, P., & Zocca, V. (2019). Python Deep Learning: Exploring deep learning techniques and neural network architectures with Pytorch, Keras, and TensorFlow. Packt Publishing Ltd.

33. Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning. Springer series in statistics. New York, NY, USA.

34. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: Springer.

35. HAMMOUDI, M., & KECHRA, R. A. (2022). Systèmes d'Identification Automatique des Véhicules (Doctoral dissertation, Université Ibn Khaldoun-Tiaret-).

36. AMIA, O., & LARBAOUI, A. (2020). Reconnaissance Automatique des plaques d'immatriculation (Doctoral dissertation)

37. Swain, D., Pattnaik, P. K., & Gupta, P. K. (2021). Machine Learning and Information Processing. Springer Singapore.

38. Hasnaoui, A. (2019). Identification d'indicateurs stratégiques dans les documents (Doctoral dissertation, Université du Québec à Trois-Rivières).

39. Foster, G., & Stefik, M. (1986, December). Cognoter: theory and practice of a colab-orative tool. In Proceedings of the 1986 ACM conference on Computer-supported cooperative work (pp. 7-15).

40. Korzeniewski, S. J., Sutton, E., Escudero, C., & Roberts, J. M. (2022). The global pregnancy collaboration (CoLab) symposium on short-and long-term outcomes in offspring whose mothers had preeclampsia: a scoping review of clinical evidence. Frontiers in Medicine, 9, 984291.

41. Jafarpour Khameneh, N. (2014). Machine Learning for Disease Outbreak Detection Using Probabilistic Models [Thèse de doctorat, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/1659/

42. J. Léonard, « Médecine et statistiques au xixe siècle », Publications des séminaires de mathématiques et informatique de Rennes, fascicule 2, « Séminaires de mathématiques. Science, histoire et société contemporaine », 1983, p. 1-14.

43. D.L. Sackett et coll., « Evidence-based medicine: What it is and what it isn't », BMJ, vol. 312, no 7023, janvier 1996, p. 71-72.

44. R.O. Bahado-Singh, S. Vishweswaraiah, B. Aydas, A. Yilmaz et coll., "Precision cardiovascular medicine: artificial intelligence and epigenetics for the pathogenesis and prediction of coarctation in neonates", J Matern Fetal Neonatal Med., 2020 Feb 4, p. 1-8.

45. J. Balayla, G. Shrem, "Use of artificial intelligence (ai) in the interpretation of intrapartum fetal heart rate (fhr) tracings: a systematic review and meta-analysis", Arch Gynecol Obstet., 2019 Jul., 300(1), p. 7-14.

46. Boulekcher, R. & Kabour, O. & Chebout, M. (2021). (Une Approche Basée Machine Learning Pour La Prédiction Du Covid-19 En Algérie). [Mémoire de Master, Université Larbi Ben M'hidi - Om-el-bouaghi].

47. Camille Kergal. Méthode d'apprentissage profond pour l'analyse génomique des cancers canins comme modèle des cancers humains. Médecine humaine et pathologie. Université de Rennes, 2022

48. Muhammed Niyas K.P., Thiyagarajan P., Feature selection using efficient fusion of Fisher Score and greedy searching for Alzheimer's classification, Journal of King Saud University - Computer and Information Sciences,Volume 34, Issue 8, Part A,2022,Pages 4993-5006,

49. S.K.M.G.M.M. Animesh Hazra, «Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques A Review,» Advances in Computational Sciences and Technology, pp. 2137-2159, 2017