République Algérienne Démocratique et Populaire

Ministère deL'enseignement Supérieur Et de la Recherche Scientifique

Université Yahia Fares Médéa

Faculté Des Sciences

Département de mathématique et Informatique



Mémoire de fin d'études

Pour l'obtention du diplôme de Master en Informatique

Option : Ingénierie des systèmes et technologie du web

Thème : Détection de virus l'hépatite C basé sur les modèles de Machine Learning

---

Présenté par :

- SARADOUNI Melissa

- RAIS Kaouther

Dirigé par :

Mr MAZARI Redha

Année Universitaire : 2022-2023

Remerciements

Nous remercions dieu tout puissant de nous avoir accordé la santé, le courage et la volonté pour accomplir ce modeste travail.

Nous tenons également à remercier notre encadreur Mr MAZARI REDHA pour l'aide et l'assistance qu'il nous à fournis afin de nous permettre de mener à bien et à terme ce travail.

Nous exprimons notre grand respect aux honorables nombres de jury qui ont accepte d'évaluer ce travail.

Nous tenons tout simplement à exprimer notre profonde gratitude a tous ceux qui nous ont soutenus de prés ou de loin durant tout notre cursus et espérant que ce mémoire servira de guide pour les promotions à venir.

Dédicaces

Je tiens tout d'abord à remercier ALLAH le tout puissant et miséricordieux, qui m'a donné La force et la patience d'accomplir ce Modeste travail.

À Mon père , j'espère que tu sois fier et que tu trouves ici le résultat de longues années de sacrifices et de privations pour m'aider à avancer dans la vie. Puisse Dieu faire en sorte que ce travail porte son fruit, Merci pour les valeurs nobles, l'éducation et le soutient permanent venu de toi.

À Ma mère mon exemple, qui a toujours été présente pour moi quelques soit les situations, qui m'a toujours soutenue, aidée, encouragée, et qui a fait de moi ce que je suis aujourd'hui. Sans elle je n'aurais jamais réussi, je sais que je ne pourrais jamais assez la remercier, alors j'espère qu'elle trouvera dans ce modeste travail un début de ma gratitude. Les mots ne sont pas assez fort pour exprimer la fierté que je ressens d'être sa fille.

À mes chères sœurs Linda, Radhia, Randa et Tina Pour leur soutien, et d'avoir accepté de m'écouter parler de ce mémoire pendant des mois.

À mes chers petits neveux Sid ali , Rmdhan Mohammed Amir ,Ahmed Ziri A ma chère nièce Soufia Iline .

À mes chères amies : Kaouther, Zahra , Warda , Youcef , Abdou , Oussama, Fateh , Djalel ....

À tous ceux qui me sont chères. A tous ceux qui m'aiment. A tous ceux que j'aime À toute personne qui a contribué à la réalisation de ce mémoire.

SARADOUNI Melissa

Dédicaces

Il n'y a que l'œuvre réalisé avec le soutien moral des personnes qui nous sont proches,

Je tiens à dédicacer mémoire à :

À ma grand-mère, puisse dieu lui accorde sa sainte miséricorde et l'accueillir en son vaste paradis.
À mes chers parents qui m'ont donné la vie et qui méritent le plus grand respect, je les souhaite le bonheur, la joie et la reconnaissance.

À mon grand-père, que Dieu le garde.

À mes très chers frères et sœurs : Nedjmeedine, Ziad, Rania et Chahinaze.
À mes chères tantes : Linda, Farida, Faiza, Dalal,, Amel, Nassira, Khadija et Hafidha.

À mes oncles : Mohammad et Houssem.
À mes cousines : Fatima Ezzahra Jouri, Maria, Sara et Hadil.

À mes amies qui ont été un vrai soutien durant toute cette période : Melissa, Hanane, Khadija, Ahlem et Manel.

À mes amis : Djalel, Riadh, Abdel slame, Fateh, Abd Elkadr, Islam, Oussama qui m'ont beaucoup aidé.

À toute ma grande famille Rais et Dahmani.

Que toute personne m'ait aidé, de près ou de loin, trouve ici l'expression de ma reconnaissance.

RAIS Kaouther

Résumé

L'hépatite C est une maladie du foie qui peut entraîner de graves complications si elle n'est pas détectée et traitée rapidement, donc il faut la détecter pour assurer une prise en charge efficace et réduire les risques liés à cette maladie. Cependant, les méthodes traditionnelles de dépistage de l'hépatite C peuvent présenter des limitations en termes de coûts élevés et de délais pour obtenir les résultats.

L'objectif principal de cette étude est d'améliorer ces méthodes de dépistage en termes de sensibilité, de coût et de rapidité des résultats.

Ainsi, notre étude se concentre sur la construction d'un modèle de Machine Learning (ML) qui vise à prédire de manière précise la présence de l'hépatite C.

Nous avons utilisé différents algorithmes de Machine Learning tels que Régression Logistique (LR), Forêt Aléatoire (RF), K plus proches voisins (KNN), Arbre de Décision (DT) et Machine à Vecteurs de Support (SVM).

Les résultats obtenus ont démontré que le SVM est la méthode la plus performante avec un taux de précision à 96% qui est une valeur très satisfaisante.

Ces résultats prometteurs suggèrent que cette étude peut servir de base pour de futures recherches dans le domaine de la détection de l'hépatite C.

En outre, nous avons développé une application web simple pour tester les prédictions du modèle construit.

Mots clés : l'hépatite C, Détectée, ML, LR, RF, KNN, DT, SVM, Précision.

Abstract

Hepatitis C is a liver disease that can lead to serious complications if not detected and treated quickly. Therefore, it is essential to detect it to ensure effective management and reduce the risks associated with this disease. However, traditional methods of hepatitis C screening may have limitations in terms of high costs and delays in obtaining results.

The main objective of this study is to improve these screening methods in terms of sensitivity, cost, and speed of results.

Thus, our study focuses on building a Machine Learning (ML) model that aims to accurately predict the presence of hepatitis C.

We used different Machine Learning algorithms such as Logistic Regression (LR), Random Forest (RF), K-nearest neighbors (KNN), Decision Tree (DT), and Support Vector Machine (SVM).

The results obtained demonstrated that SVM is the most effective method with an accuracy rate of 96%, which is very satisfying.

These promising results suggest that this study can serve as a basis for future research in the field of hepatitis C detection.

Additionally, we have developed a simple web application to test the predictions of the constructed model.

Key words: hepatitis C, Detected, ML, LR, RF, KNN, DT, SVM, Precision.

الملخص

التهاب الكبد هو مرض في الـكبد يمكن أن يؤدي إلى مضاعفات خطيرة إذا لم يتم اكتشافه وعلاجه بسرعة.
لذلك من الضروري اكتشافه لضمان إدارة فعالة وتقليل المخاطر المرتبطة بهذا المرض. ومع ذلك قد تواجه الطرق التقليدية لفحص التهاب الـكبد C قيودا فيما يتعلق بتكاليف مرتفعة وتأخير في الحصول على النتائج.
الهدف الرئيسي لهذه الدراسة هو تحسين هذه الطرق الفحص من حيث الحساسية والتكلفة وسرعة النتائج.
بالتالي ، تركز دراستنا على بناء نموذج تعلم آلي يهدف إلى التنبؤ بدقة وجود التهاب الـكبد C
لقد استخدمنا خوارزميات تعلم الآلة المختلفة مثل الانحدار اللوجستي، الغابة العشوائية، أقرب الجيران ، شجرة القرار، وآلة الدعم الناقل.
أظهرت النتائج المتحصلة أن آلة الدعم الناقلة هي الطريقة الأكثر فعالية بنسبة دقة تصل إلى 96٪، وهو أمر يبعث على الرضا جدًا.
تشير هذه النتائج الواعدة إلى أن هذه الدراسة يمكن أن تكون أساسا للبحوث المستقبلية في مجال كشف التهاب الـكبد.
.بالإضافة إلى ذلك ، قمنا بتطوير تطبيق ويب بسيط لاختبار توقعات النموذج المبني
.

الكلمات المفتاحية: التهاب الـكبد الوبائي، الـكشف، الطرق، تعلم الآلي، التحويل اللوجستي، الغابة العشوائية، أقرب الجيران، شجرة القرار، آلة الدعم الناقلة، الدقة.

Table des matières

Table des matières i

Table des figures iii

Liste des tableaux v

Liste des abréviations vi

Liste des abréviations vii

Introduction générale 1

Chapitre 1 : L'informatique médicale et l'hépatite c : 3
1.1 Introduction 3
1.2 Soins de santé en informatique : 3
1.2.1 Exemples d'outils informatiques pour le secteur de la santé : . . . 4
1.3 Généralité sur l'hépatite c : 4
1.3.1 Définition de l'hépatite c : 4
1.3.2 Découverte de l'hépatite c : 5
1.3.3 Évolution de l'hépatite C : 5
1.3.4 Modes de transmission de l'hépatite C : 6
1.3.5 Techniques de diagnostic du VHC : 7
1.3.6 Les conseils et les préventions contre le virus de l'hépatite c : . . 8
1.4 Conclusion : 9

Chapitre 2 : Machine Learning : 10
2.1 Introduction 10
2.2 L'Intelligence Artificielle (IA) : 10
2.3 Machine Learning : 11
2.3.1 Définition : 11
2.3.2 Les méthodes de Machine Learning : 11
2.3.3 Les principaux algorithmes de machine Learning : 13
2.4 Deep Learning : 17
2.5 L'apprentissage automatique pour la détection de l'hépatite : 17
2.6 Conclusion : 21

Chapitre 3 : Approche proposé : 22
3.1 Introduction 22
3.2 Méthodologie : 22
3.3 Préparation des données : 24
3.3.1 L'ensemble de données : 24

3.3.2 Equilibrage des données : 25
3.3.3 Nettoyage des données : 26
3.3.4 Répartition des données de train/test : 26
3.3.5 Transformation et normalisation des données : 26
3.3.6 Choisir les modèles de Machine Learning : 27
3.3.7 Évaluation des performances : 30
3.4 Conclusion : 32

Chapitre 4 : Résultats et Discussion : 33
4.1 Introduction 33
4.2 L'environnement de développement : 33
4.2.1 Python 33
4.3 Tests et Résultats : 35
4.3.1 Dataset : 35
4.3.2 Nettoyage de données 36
4.3.3 Transformation et normalisation des données : 36
4.4 Evaluation de chaque modèle : 37
4.5 La comparaison des modèles 42
4.6 Application : 44
4.7 Comparaison notre résultats avec les autres travaux connexes : 45
4.8 Conclusion : 47

Conclusion générale 48

Table des figures

1.1 représentation schématique de VHC. 4

2.1 représentation de l'intelligence artificielle. 10
2.2 Machine Learning. 11
2.3 les méthodes de machine Learning. 12
2.4 L'algorithme SVM. 14
2.5 L'algorithme forêt aléatoire. 14
2.6 L'algorithme KNN. 15
2.7 Exemple d'un arbre de décision. 15
2.8 Exemple d'une régression logistique. 16

3.1 Le flux général du processus. 23
3.2 Data Set 24
3.3 Pourcentage des catégories. 25
3.4 nouveaux Pourcentage des catégories. 26
3.5 code régression logistique 27
3.6 code de KNN 28
3.7 code arbre de décision 28
3.8 code de forêt aléatoire 29
3.9 code de SVM 29
3.10 Matrice de confusion pour une classification multi classe. 30

4.1 Résultats du Dataset. 35
4.2 Nombre des valeurs nulles avant le nettoyage. 36
4.3 Résultat de nettoyage de données. 36
4.4 Résultat de transformation de données en utilisant le Standard scaler 36
4.5 Résultats du modèle Régression logistique. 37
4.6 Matrice de confusion du modèle Régression logistique. 38
4.7 Résultats du modèle KNN. 38
4.8 Matrice de confusion du modèle KNN. 39
4.9 Résultats du modèle Arbre de décision. 39
4.10 Matrice de confusion du modèle Arbre de décision. 40
4.11 Résultats du modèle forêt aléatoire. 40
4.12 Matrice de confusion du modèle forêt aléatoire. 41
4.13 Résultats du modèle SVM. 41
4.14 Matrice de confusion du modèle SVM. 42

4.15 Comparaison des modèles basée sur leur précision (accuracy) 43
4.16 Application web de Test. 44
4.17 Résultats de Test. 44

Liste des tableaux

2.1 Comparaison des performances entre le classifier RF utilisant les para- mètres par défaut, l'ajustement des paramètres et la sélection des caractéristiques avec les nouveaux paramètres. 18
2.2 La comparaison de la méthode proposée avec d'autres méthodes. 19
2.3 Performance de différents modèles d'apprentissage automatique..... 20

4.1 Le résultat de la précision globale de cinq modèles proposés. 42
4.2 Comparaison des résultats obtenus entre le travail sélectionné et notre étude 45

Liste des abréviations

ADN	Acide désoxyribonucléique
ADNc	Acide désoxyribonucléique complémentaire
ALAT	Alanine aminotransférase
ALB	Albumine
ALP	Alkaline phosphatase
ALT	Alanine aminotransférase
ANN	Artificial neural networks
ARN	Acide ribonucléique
AST	Aspratate aminotransférase
BIL	Bilirubine
CHE	Cholinesterase
Chol	Cholestérol
CNN	Convolutional neural networks
CREA	Créatinine
DNN	Deep neural networks
DT	Decision trée
ELISA	Enzyme linked immunosorbent assay
EMR	Electronic Medical Record
FN	False negative
FP	False positive
GTT	Gamma glutamyltransférase
GNN	Generative adversarial networks
IA	Intelligence artificielle
IHSDS	Intelligent hepatits c stage diagnostic system
KNN	K-nearest neighbros
LR	Logistic regression
LSTM	Long short-term
memoryML	Machine learning
NN	Neural network
NS3	Protéine non structurale codée par hépatite c
NS4	Protéine non structurale codée par hépatite c
NS5	Protéine non structurale codée par hépatite c
PCR	Polymerase chain reaction
PROT	Protéine
RELP	Restriction enzyme lenght polymorphism

Liste des abréviations

RF	Random forest
RNN	Recrurrent neural network
SFS	Sequential forward selection
SHAP	Shapley additive explanations
SMOTE	Synthetic minority over sampling technique
SVM	Support vector machine
TMA	Transcription médite amplification
TN	True negative
TP	True postive
TROD	Test rapide d'orientation diagnostique
UCI	University of california,Irvine
VHA	Virus de l'hépatite A
VHB	Virus de l'hépatite B
VHC	Virus de l'hépatite C
VIH	Virus l'lmmunodéficiene humaine

Introduction générale

1. Contexte générale et problématique :

L'hépatite C est une maladie virale du foie qui représente un problème de santé majeur à l'échelle mondiale. Elle est causée par le virus de l'hépatite C (VHC) et peut entraîner des complications graves, telles que la cirrhose du foie et le cancer du foie.

La détection précoce de l'infection par le VHC est essentielle pour assurer une prise en charge rapide et efficace des patients. Traditionnellement, le diagnostic de l'hépatite C impliquait des tests sérologiques et moléculaires laborieux et coûteux. Cependant, l'avènement de l'apprentissage automatique et des modèles de Machine Learning (ML) offre de nouvelles opportunités pour améliorer le processus de diagnostic.

Les modèles de ML sont des outils informatiques capables d'apprendre à partir de données et de prendre des décisions ou de faire des prédictions sans être explicitement programmés. Ces modèles peuvent être formés à reconnaître des schémas complexes dans les données, ce qui en fait une approche prometteuse pour la détection de l'hépatite C.

2. Objectif :

L'objectif de ce mémoire est d'explorer l'utilisation des modèles de ML dans la détection du virus de l'hépatite C. Nous examinerons différentes méthodes et algorithmes de ML des classifications d'apprentissage supervise tels que KNN, DT, RF, LR et SVM.

Nous chercherons également à évaluer les performances des modèles de Machine Learningen termes de sensibilité, de spécificité et d'exactitude, comparativement aux méthodesde diagnostic conventionnelles.

Cela nous permettra de déterminer si les modèles de ML peuvent être utilisés comme outils efficaces de dépistage de l'hépatite C.

Introduction générale

3. Structure du travail :

Ce travail est organise en quatre principaux chapitres comme suit :

- Chapitre 1 : L'informatique médicale et l'hépatite C

Ce chapitre offre un aperçu général de la pandémie de l'hépatite C.

- Chapitre 2 : Machine Learning

Il s'agit d'aborder une introduction à l'apprentissage automatique et aux algorithmes d'apprentissage supervisé pouvant être utilisés pour détecter précocement l'hépatite C.

- Chapitre 3 : Approche proposé

Dans ce chapitre, nous détaillons l'approche proposée, la description de l'ensemble de données utilisé et les étapes de prétraitement des données.

- Chapitre 4 : Résultats et Discussion

Dans lequel nous exposerons les outils et les langages utilisés, nous discuterons les résultats obtenus, nous évaluerons la performance du modèle, puis nous comparerons les résultats expérimentaux de notre modèle avec ceux des études antérieures dans la dernière section.

Chapitre 1

Chapitre 1:L'informatique médicale et l'hépatite C :

1.1 Introduction

Ce chapitre se penche sur l'importance de l'informatique dans le domaine de la médecine, en particulier nous discuterons des avantages de l'utilisation de l'informatique pour améliorer la précision et l'efficacité des diagnostics et des traitements médicaux. Aussion définit la maladie de l'hépatite, puis son évolution et ses symptômes, les modes de transmission, les conseils de prévention les techniques de détection. Enfin on exposera les travaux de recherches antérieurs qui montrer l'efficacité de l'utilisation de l'informatiquepour la détection de l'hépatite.

1.2 Soins de santé en informatique :

L'informatique médicale est un domaine qui combine l'informatique avec la santé. Les applications de l'informatique médicale couvèrent un large éventail de domaines, notamment la gestion et le stockage des dossiers médicaux, générer des rapports et des statistiques sur la santé et bien plus encore surveiller l'état de santé du patient, et détection des anomalies ….

Les principaux avantages des systèmes informatiques sur le soin de santé : **[35]**

-Les systèmes informatiques peuvent aider les professionnels de la santé à réduire le temps nécessaire pour effectuer des diagnostics et des traitements ce qui permet de réduire les coûts et de traiter plus de patients.

-Les systèmes informatiques peuvent fournir des informations plus précises et fiables, ce qui peut aider les médecins à prendre des décisions plus éclairées et à fournir des soins plus efficaces et plus surs aux patients.

-Les systèmes d'analyse de données peuvent être utilisés pour identifier les risques potentiels et les facteurs de risque, ce qui peut aider les médecins à prendre des mesures préventives.

1.2.1 Exemples d'outils informatiques pour le secteur de la santé :

Il existe plusieurs outils informatiques qui peuvent être utilisés dans le secteur de la santé :

1. Gestion des dossiers médicaux électroniques (EMR) : Les dossiers médicaux électroniques (EMR) sont des techniques informatiques qui permettent aux médecins de stocker et de partager des informations sur les patients.
2. plateformes télémédecine : peuvent être utilisées pour développer l'accès aux soins de santé et fournir des consultations à distance.
3. Imagerie médicale : Les systèmes d'imagerie médicale peuvent être utilisés pour diagnostiquer et traiter plusieurs maladies
4. Systèmes de surveillance des patients : peuvent être utilisés pour surveiller les conditions médicales des patients et leur état de santé.
5.systèmes d'analyse et partage de données : peuvent être utilisés pour comprendre les évaluations et les conséquences de maladie .et facilite le partage des informations médicales entre les professionnels de la santé.
6. Systèmes d'assistance robotique : Les systèmes d'assistance robotique peuvent être utilisés pour aider les médecins à exécuter des procédures médicales complexes et à effectuerdes tâches difficiles. **[37]**

1.3 Généralité sur l'hépatite c :

1.3.1 Définition de l'hépatite c :

L'hépatite c (VHC) est une maladie infectieuse causée par le virus de l'hépatite c. Il s'agit d'une infection du foie qui peut entrainer des complications graves, notammentune cirrhose et un cancer du foie. Généralement ne produit aucun signe ou symptômeau cours de ses premiers stades. Lorsque les signes et symptômes se présentent, ils sont généralement bénins et grippaux et peuvent inclure : des douleurs abdominales, une fatigue persistante, des nausées et des vomissements.**[41]**

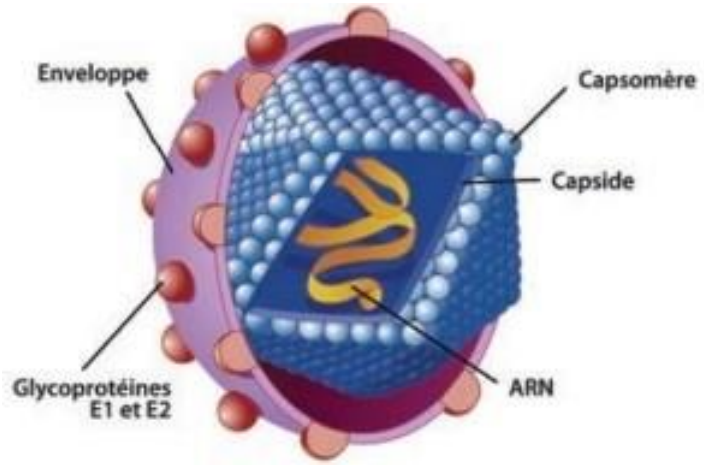


FIGURE 1.1 – représentation schématique de VHC **[26]**

1.3.2 Découverte de l'hépatite C :

En 1974, des tests sérologiques ont été mis au point pour détecter la présence du virus de l'hépatite A (VHA) ou de l'hépatite B (VHB) chez les patients enflamment d'hépatite. L'année suivante, Feinstone et son équipe ont testé le sang des patients infectés par l'hépatite après une transfusion pour détecter des anticorps contre différents virus. Ont constaté que certains patients ne produisaient pas des anticorps contre les virus des hépatites A ou B, ou contre le Virus Epstein Barr. Ils mirent donc en évidence l'existence d'un ou de plusieurs nouveaux agents inconnus causant des hépatites, baptisés (Hépatites Non A Non B) ou (HNANB). La principale percée a été la caractérisation du génome du virus en 1989 par Choo et son équipe à l'aide des techniques de biologie moléculaire. Ils ont utilisé du sérum de chimpanzés infectés par l'agent causal des (HNANB) pour produire une grande banque d'ADNc qui a été clonée dans un bactériophage "lambdagt11". Cette banque d'expression fut alors criblée par des sérums de patients humains atteints de (HNANB). Les auteurs isolèrent le clone, codant un antigène reconnu par les sérums de plusieurs patients. De plus, Choo et son équipe ont démontré que l'ADNc correspondant n'était pas dérivé d'un ARN encodé par le génome du primate, prouvant ainsi son origine virale. L'agent causal des (HNANB) ainsi identifié, prit alors le nom de VHC (Virus de l'hépatite C). À la cour de la même année, furent développés un test ELISA pour détecter la présence du VHC chez des patients et prouvent que ce virus était la principale cause des (HNANB) transfusionnelles à travers le monde.**[23]**

1.3.3 Évolution de l'hépatite C :

L'hépatite C est caractérisée par deux stades distincts un stade aiguë et un stade chronique :

Hépatite Aiguë :

L'hépatite aiguë apparait après 8 semaines d'incubation et plus souvent asymptomatique tandis que les formes symptomatiques durent en général 2 à 12 semaines et provoquant par des arthralgies.**[5]**

Hépatite Chronique :

L'hépatite chronique est une inflammation du foie qui subsiste pendant au moins 6 mois. C'est asymptomatique chez la plupart des gens. L'hépatite chronique est détectée durant d'examens de routine et se traduit par un taux d'ALAT souvent variables, une virémie plasmatique positive et la présence d'anticorps. Il existe plusieurs formes d'hépatites chroniques souvent associées aux degrés de lésions du foie. **[14]**
Il est schématisé en 3 formes :

- L'hépatite chronique avec transaminases normales.
- L'hépatite chronique minime.
- L'hépatite chronique modérée ou sévère.

1.3.4 Modes de transmission de l'hépatite C :

Le virus de l'hépatite C est transmis principalement par le sang, cette transmission peut se faire par :

La drogue :

L'usage de la drogue peut augmenter le risque de contracter l'hépatite C. Les principaux facteurs de risque sont l'utilisation de seringues et d'autres matériels médicaux contaminés, le partage d'objets personnels contaminés tels que (cuillère, filtre, eau, paille,la pipe à crack...) Et l'utilisation de drogues récréatives par voie intraveineuse, nasale ou orale. Les consommateurs des drogues récréatives doivent prendre des mesures afin de réduire le risque de transmission du virus.**[13]**

Transfusion de Produits sanguins :

Les produits sanguins sont les principales sources de transmission de l'hépatite C, car ils peuvent contenir des virus vivants. La transmission par transfusion de produits sanguins peut se produire lorsque des produits sanguins contaminés sont transfusés à une personne qui n'est pas infectée. Les risques de transmission de l'hépatite C par transfusion sont très faibles, car les produits sanguins sont testés et sélectionnés strictement. Les risques sont principalement associés à l'utilisation de produits sanguins non testés ou à l'utilisation de produits sanguins testés mais qui ne sont pas encore confirmés. Par conséquent, il est important de prendre des mesures de précaution pour réduire le risque de transmission de l'hépatite C par transfusion. Il est également important de s'assurer que les produits sanguins sont testés et sélectionnés avec soin. **[4]**

Transmission de la mère à l'enfant :

L'hépatite C se transmet de la mère à l'enfant dans des cas très rare, mais elle peut se produire. Cette transmission est estimée à environ 5%. Il existe différentes manières par lesquelles le virus peut être transmis, notamment par le sang, pendant la grossesse ou l'accouchement, les produits sanguins, le lait maternel et par contact direct avec la peau. Dans la plupart des cas, l'enfant ne présentera aucun symptôme, mais il est possible qu'il développe des complications plus tard dans la vie. **[40]**

Transmission par voie sexuelle :

La transmission par voie sexuelle de l'hépatite C constitue une forme de transmission interpersonnelle. Les personnes les plus à risque sont celles qui ont des relations sexuelles non protégées, celles qui ont des partenaires multiples et celles qui sont infectées par le VIH. Les personnes qui sont à risque doivent être testées pour l'hépatite C et se faire vacciner contre le virus de l'hépatite B. Les préservatifs peuvent réduire le risque de transmission par voie sexuelle, mais ils ne garantissent pas une protection complète.**[4]**

1.3.5 Techniques de diagnostic du VHC :

Pour diagnostiquer des infections par le VHC est basé sur deux types de tests : les tests indirects, qui recherchent les anticorps spécifiques dirigés contre le virus, et les tests directs basés sur les techniques de biologies moléculaires, qui recherchent les composants des particules virales.**[7]**

1.3.5.1 Les tests indirects ou sérologiques :

Les tests sérologiques sont des tests immuno enzymatiques. Ces tests utilisent des protéines recombinantes ou des peptides synthétiques codés par les régions structurales (capside, enveloppe) et les régions non structurales (NS3, NS4 et NS5) du génome viral. Ces tests permettent de détecter la présence d'anticorps spécifiques produits par le système immunitaire après l'exposition à un virus.**[7]**

Test ELISA : Les tests de troisième génération sont conçus pour détecter le virus de l'hépatite c et présentent une sensibilité et une spécificité excellente d'environ 98 et 99% respectivement. Bien que la sensibilité des tests de troisième génération soit satisfaisante chez les hémodialysés et les personnes coinfectées par le VIH, elle peut être faible chez les personnes immunodéprimées ou au tout début de l'infection par le VHC. **[7]**

Les TROD : Les tests rapides d'orientation diagnostique (TROD) sont des tests unitaires à lecture subjective, conçus pour fournir un résultat dans un délai court (généralement moins de 30 minutes). Ils peuvent être réalisés sur du sang total, de salive, du plasma ou des sérums selon la matrice revendiquée par le fabricant pour son produit. Le principe de ces tests est simple : ils consistent à capturer des antigènes ou des anticorps sur une surface solide, puis à ajouter des peptides spécifiques ou des immunoglobulines spécifiques lorsque ces analystes interagissent avec leurs immunoglobulines ou peptides, cela peut créer une marque visible à l'œil nu, qui permet de détecter le résultat.**[7]**

1.3.5.2 Les tests directs :

Les techniques de biologie moléculaire : Les techniques de biologie moléculaire appliquées au diagnostic des infections par le VHC peuvent être classées en trois catégories : les méthodes de détection de l'ARN du VHC, les méthodes de quantification de l'ARN et les méthodes de génotypage.**[7]**

Détection qualitative de l'ARN du VHC : Les techniques d'amplification de la cible, comme la réaction en chaine par polymérase (PCR) et la transcription médiate amplification (TMA), permettent de détecter l'ARN du VHC en quantités très faibles. Ces techniques d'amplification se basent sur la transcription inverse de l'ARN viral en ADN complémentaire, ce qui permet ensuite de produire un grand nombre de copies. Ces copies sont visibles grâce à l'électrophorèse et à la coloration spécifique de l'ADN, ou grâce à l'hybridation spécifique sur des microplaques. Parmi les techniques d'amplification de la cible. La PCR qualitative est actuellement la plus sensible pour détection de l'ARN du VHC. **[7]**

Quantification de l'ARN du VHC : La quantification de l'ARN du VHC (ou mesure de la charge virale) est une méthode pour évaluer le niveau de réplication virale dans l'organisme. Elle repose sur deux types de techniques d'amplification : les techniques d'amplification de la cible (PCR ou TMA) et les techniques d'amplification du signal. Dans la technique d'amplification de la cible, un standard interne en quantité connue est ajouté à l'échantillon et amplifié en compétition avec l'ARN viral, ce qui permet de quantifier les résultats de la cible grâce à la comparaison avec standard. De plus, des techniques en temps réel (real time PCR) sont en cours de développement pour améliorer la sensibilité, la spécificité et l'intervalle de quantification de cette méthode. Les techniques d'amplification du signal quantifient l'ARN viral, après hybridation à un support solide, grâce à un grand nombre de molécules "signal" (enzymes) qui catalysent la transformation d'un substrat en un composé détectable. La quantité de composé produite est comparée à un courbe étalon tracé avec des standards en quantités connues pour permettre la quantification. Les avantages de cette méthode Sont sa spécificité et sa reproductibilité, et elle quantifie de façon équivalente les différents génotypes du VHC.**[7]**

Les méthodes de génotypage : Le génotypage viral est une technique utilisée pour prédire le résultat du traitement et influencer le choix de la thérapie. Cette méthode repose sur une amplification initiale des fragments par PCR, suivie d'une analyse du polymorphisme de restriction (RELP) ou d'une technique standardisée d'hybridation inverse basée sur l'utilisation de sondes Oligo nucléotidiques spécifiques des différents génotypeset sous-types du VHC. **[7]**

1.3.6 Les conseils et les préventions contre le virus de l'hépatite C :

Il est important de prendre des mesures pour prévenir et réduire le risque de contracter l'hépatite C.

Voici quelques conseils et précautions :

- Ne partagez jamais d'aiguilles, de seringues, de tatouages ou de perçages et n'utilisez jamais d'objets coupants infectés.
- Utilisez un préservatif lors des relations sexuelles.
- Évitez les drogues injectables ou les pratiques qui les impliquent.
- Ne partagez pas votre brosse à dents, rasoirs ou autres objets.
- Si vous êtes une femme enceinte, faites-vous tester pour l'hépatite C pour protéger votre bébé.
- Si vous êtes un professionnel de la santé, portez des gants et des lunettes de protection lors de l'exposition à des fluides corporels.
- Faites-vous vacciner contre l'hépatite A et l'hépatite B si vous êtes à risque.
- Si vous avez des antécédents de comportements à risque ou si vous pensez que vous pourriez être infecté par le VHC, faites-vous tester. **[21]**

1.4 Conclusion :

Au cours de ce chapitre, nous avons décrié brièvement l'importance de l'informatique dans le monde médicale, ensuite nous avons expliqué en détail la maladie de l'hépatite C, sa découverte et évolution ses modes de transmission, ses techniques de diagnostic et quelques conseils de préventions contre le virus.

Dans le chapitre suivant nous passerons à présenter des travaux de recherches antérieurs sur les méthodes automatiques de détectionde VHC.

Chapitre 2

Chapitre 2 : Machine Learning

2.1 Introduction

Ce chapitre passe en revue certaines tâches d'apprentissage il s'agit d'aborder les définitions d'intelligence artificielle (IA), de machine Learning (ML) et leurs différentes méthodes et modèles en détails, plus le Deep Learning et discuterons sur l'importance d'utilisation de machine Learning pour améliorer le monde de sante.

2.2 L'Intelligence Artificielle (IA) :

L'intelligence artificielle (IA) est un domaine de recherche et de science vieux de 30 ans qui vise à développer des ordinateurs capables de simuler les capacités cognitives humaines, telles que l'apprentissage, la prise de décision et la pensée.
Les technologies d'IA peuvent être utilisées pour automatiser des tâches complexes, mieux comprendre les données et améliorer les performances commerciales.**[25]**
De plus, les principaux domaines fonctionnels de l'IA démontrent leur énorme potentiel et les innombrables réalisations qu'elle est susceptible d'entraîner. **[6]**
L'intelligence artificielle peut être divisée en deux domaines principaux : l'apprentissage automatique et l'apprentissage profond.

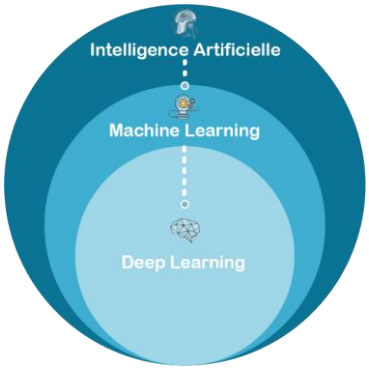


FIGURE 2.1 – représentation de l'intelligence artificielle.**[39]**

2.3 Machine Learning :

2.3.1 Définition :

Machine Learning (ML) est une branche de l'intelligence artificielle qui permet aux ordinateurs d'apprendre des données sans être explicitement programmés. C'est un processus par lequel des modèles informatiques sont formés à partir de données précédemment analysées, afin de prédire de nouvelles informations ou de prendre des décisions sans supervision humaine. La technologie a été développée en 1959 par Arthur Samuel, un informaticien. Depuis, cette technologie a fait d'importants progrès grâce à l'intérêt et aux investissements importants des chercheurs et des entreprises. L'apprentissage automatique est maintenant utilisé dans un large éventail d'applications, y compris la reconnaissance vocale, le traitement du langage naturel, l'auto-conduite et le marketing prédictif. Il est également utilisé pour améliorer les systèmes décisionnels et les processus opérationnels.**[45]** ML est un domaine de recherche à l'intersection de l'intelligence,mathématique et informatique. **[32]**

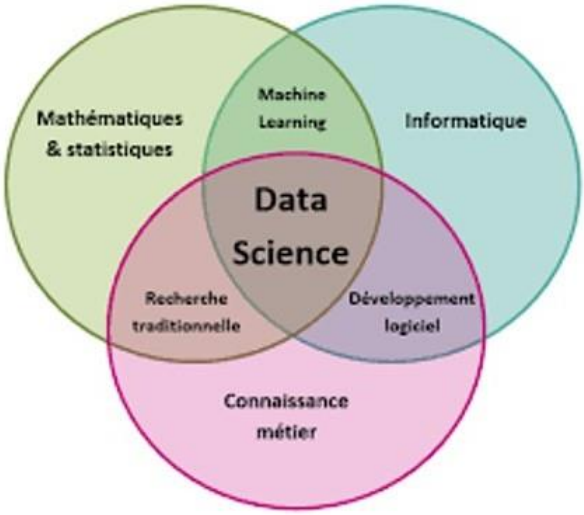


FIGURE 2.2 – Machine Learning.**[15]**

2.3.2 Les méthodes de Machine Learning :

Les méthodes d'apprentissage automatique sont un ensemble de techniques d'apprentissage automatique qui permettent aux ordinateurs d'acquérir et de diffuser des connaissances sur les données. Ils sont utilisés pour résoudre des problèmes complexes et exécuterdes tâches plus rapidement et avec précision. Ces méthodes sont utilisées dans de nombreux domaines, y compris la vision par ordinateur, la reconnaissance vocale, le traitement du langage naturel, la prédiction de séries chronologiques et la prédiction de données. Ils sont particulièrement utiles pour automatiser des processus complexes, améliorer la précision des prédictions et la compréhension du comportement humain. Les méthodes d'apprentissage automatique sont généralement classées selon le type d'apprentissage qu'elleseffectuent (Figure 2.3) :

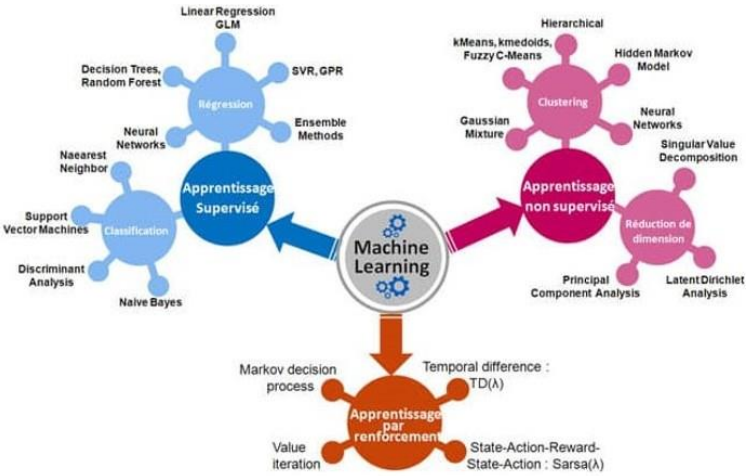


FIGURE 2.3 – les méthodes de machine Learning.**[20]**

2.3.2.1 Apprentissage supervisé :

L'apprentissage supervisé est une méthode d'apprentissage automatique où un algorithme apprend à prédire une variable de sortie (Y) à partir de variables d'entrée (x)en utilisant des données étiquetées. Il s'agit du paradigme le plus populaire en Machine Learning et en Deep Learning, et il est très précis. Des modèles d'apprentissage supervisé sont utilisés dans de nombreux domaines, notamment la reconnaissance vocale, la détection d'objets et la prévision des prix. L'apprentissage supervisé consiste à entraîner un modèle à partir de données étiquetées ou annotées.**[9]** Fonction H : X → Y de sorte que H(X) soit un bon prédicteur pour la valeur correspondante de Y. **[43]**
L'apprentissage supervisé est divisé en deux types :

2.3.2.1.1 Apprentissage supervisé par classification : C'est une façon de relier les données entre elles et de les diviser en groupes par ligne ou courbe. Il est divisé en deux types :

Classification binaire : Elle est une technique utilisée en machine Learning pour prédire si un exemple n'a que deux classes possibles, par exemple nous voulons savoir si la réclamation de client est à cause d'un problème technique ou un problème de respect. **[43]**

Classification multi classes : La classification en classes multiples consiste à répartir un ensemble de données entre plusieurs catégories. Ça veut dit qu'au lieu de Y = 0,1 nous allons étendre notre définition pour que Y = 0,1...N.**[43]**

2.3.2.1.2 Apprentissage supervisé par régression : L'apprentissage supervisé par régression est une technique d'apprentissage automatisée qui produit un modèle prédictif avec des données. Il peut prendre un exemple non nommé comme entrent et produireune cible. Il est généralement utilisé pour prédire des valeurs persistantes, telles que les prix de vente ou les salaires. Dépend des algorithmes d'apprentissage automatique tels que la régression linéaire et la régression logistique.**[9]**

2.3.2.2 Apprentissage non supervisé :

C'est un type d'auto-apprentissage qui permet au système d'apprendre des données sans correction ni supervision. Il est généralement utilisé pour détecter les structures et les modèles cachés dans les ensembles de données non nommés et donne la sortie désirée sans aucune interférence. Il est souvent basé sur des applications telles que l'analyse de données et d'images.**[16]**

2.3.2.3 Apprentissage supervisé par renforcement :

L'apprentissage supervisé par renforcement est une méthode de ML qui combine l'apprentissage supervisé et l'apprentissage par renforcement. Dans cette méthode, un algorithme d'apprentissage supervisé est utilisé pour générer un modèle prédictif qui est ensuite utilisé pour guider l'agent dans son environnement. L'agent apprend par l'expérience et s'améliore au fil du temps en prenant des actions qui lui permettront d'atteindre un objectif spécifique. Les données étiquetées fournies à l'algorithme d'apprentissage servent à générer ce modèle prédictif qui sera utilisé pour guider l'agent.**[17]**

2.3.3 Les principaux algorithmes de Machine Learning :

Les algorithmes d'apprentissage automatique sont des algorithmes qui aident les machines à apprendre et à prendre des décisions en fonction des données et de l'expérience. Ils sont utilisés dans de nombreuses applications, notamment la reconnaissance vocale, la traduction automatique et le traitement du langage naturel.
Voici quelques-unes des principaux algorithmes d'apprentissage automatique :

2.3.3.1 Réseaux de neurones :

Un réseau de neurones est un algorithme d'apprentissage automatique basé sur des réseaux de neurones artificiels conçus pour fonctionner de manière similaire à un cerveau biologique. Il se compose d'un contrat relié par des crochets et fonctionne en faisant passer des signaux entre les contrats. Il est utilisé pour résoudre des problèmes complexes en identifiant des modèles cachés dans de grandes quantités de données **[27]**. Ils sont couramment utilisés dans le traitement du langage naturel et la reconnaissance des images.

2.3.3.2 Support Victor Machine (SVM) :

SVM est une technique de classification supervisée qui peut devenir compliquée, mais reste intuitive au niveau le plus fondamental. L'objectif de SVM est d'identifier un hyperplan ou une frontière qui maximisera la marge entre les différentes classes de données, ce qui est similaire à l'objectif de la régression logistique, qui est de trouver la bonne ligne d'ajustement. Pour ce faire, nous devons trouver les points les plus proches (vecteurs de support) de la ligne des classes, et ensuite calculer la distance (marge) entre la ligne et les vecteurs de support. Enfin, notre objectif est de maximiser cette marge. L'hyperplan pour lequel la marge est maximale est l'hyperplan optimal (Figure 2.4).**[36]**

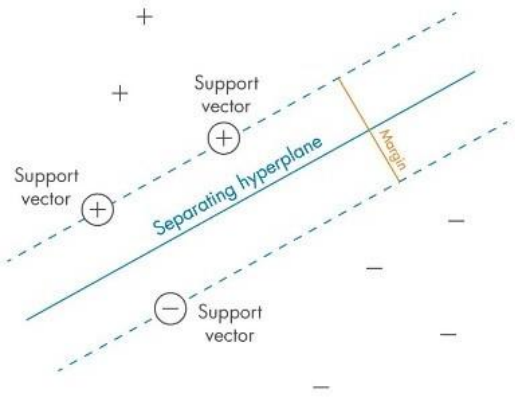


FIGURE 2.4 – L'algorithme SVM.**[30]**

2.3.3.3 forêt aléatoire :

C'est un algorithme d'apprentissage automatique combiné à des arbres de décision pour réduire les erreurs et améliorer la précision des prédictions. Il a été proposé par Leo Breiman en 2001 et consiste en un ensemble d'arbres de décision construits à partir de sous-échantillons aléatoires des données d'apprentissage. Chaque arbre de décision prédit le résultat, et les résultats finaux sont moyennés pour obtenir la prédiction finale. Les forêts aléatoires sont très robustes et peuvent être utilisées à la fois pour des tâches de classification et de régression.**[8]**

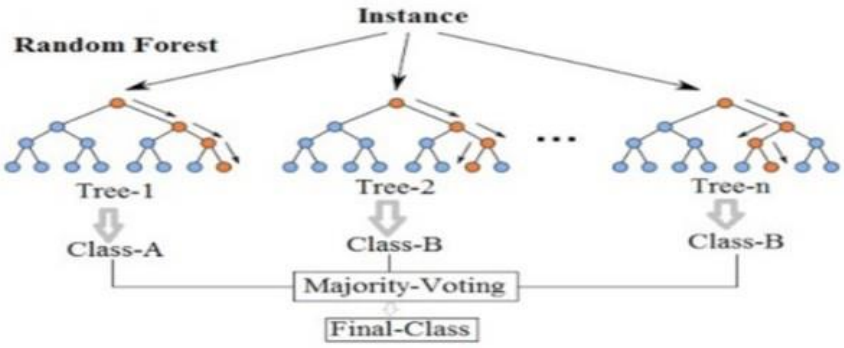


FIGURE 2.5 – L'algorithme forêt aléatoire.**[34]**

2.3.3.4 KNN (K Nearest Neighbors) :

KNN est une méthode d'apprentissage supervisé non paramétrique qui est utilisée pour la classification et la régression. Il fonctionne en trouvant les points de données de formation les plus proches d'un point de test donné et en utilisant la classe de ces points de formation pour classer le point de test. Cet algorithme est très utile pour l'exploration des données et les statistiques, car il peut apporter des prédictions sur la classe des points de données sans utiliser un modèle prédictif. **[46]**

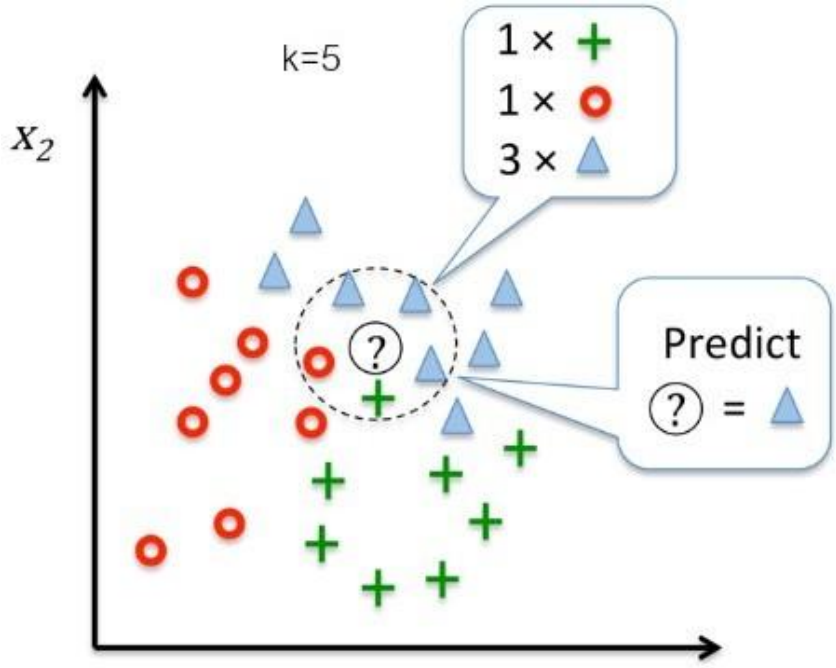


FIGURE 2.6 – L'algorithme KNN.**[44]**

2.3.3.5 Arbre de Décision :

L'arbre de décision est un algorithme de ML très répandu qui utiliseune structure en forme d'organigramme arborescent pour représenter de manière visuelle et explicite les décisions prises et les résultats possibles de ces décisions. En effet, l'arbrede décision présente graphiquement les différentes solutions envisageables selon certainesconditions, ce qui explique son nom. Il commence par une seule boîte, ou racine, qui se divise ensuite en un certain nombre de branches, tout comme un arbre. C'est un outil très efficace pour comprendre et interpréter les relations entre les variables et pour prendre des décisions en se basant sur des données. **[11]**

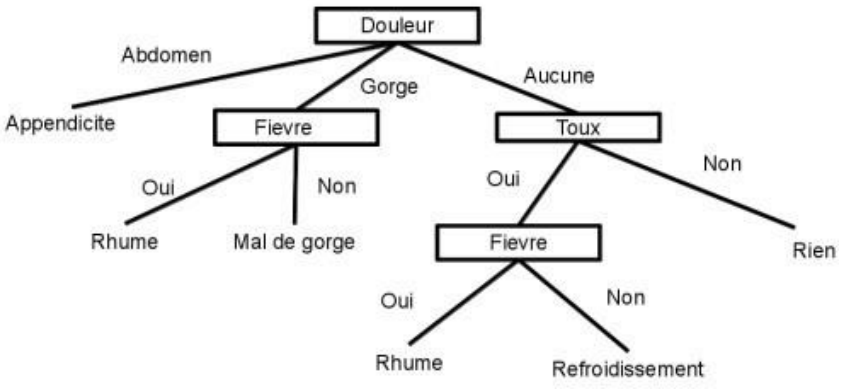


FIGURE 2.7 – Exemple d'un arbre de décision.**[11]**

2.3.3.6 Régression logistique :

La régression logistique est un algorithme de ML particulièrement efficace pour résoudre des problèmes de classification binaire, c'est-à-dire pour prédire si un événement aura lieu ou non. Il se base sur le calcul de la probabilité que l'événement se produise en utilisant les données disponibles. Il permet de classer les données en classes discrètes en étudiant la relation entre ces données et les étiquettes correspondantes. L'algorithme apprend une relation linéaire à partir d'un ensemble de données étiquetées, puis introduit une non-linéarité sous la forme de la fonction Sigmoïde pour améliorer les performances de la prédiction.**[33]**

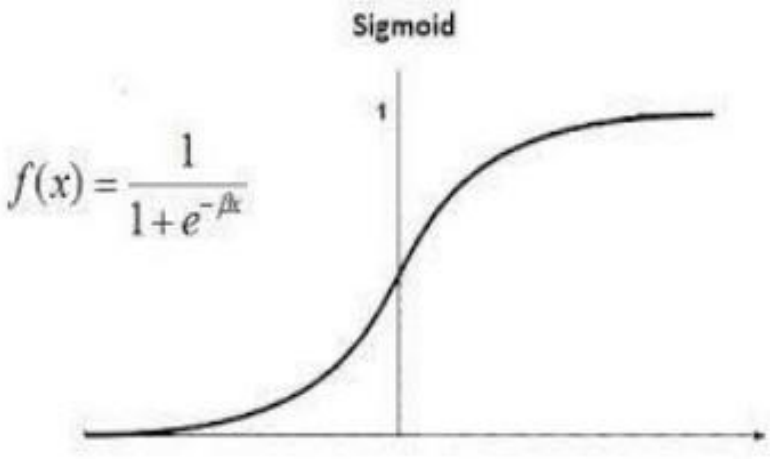


FIGURE 2.8 – Exemple d'une régression logistique.**[33]**

2.4 Deep Learning :

Deep Learning est un sous-ensemble spécifique de ML qui utilise les réseaux neuronaux profonds pour apprendre les données. Les réseaux neuronaux profonds peuvent êtreutilisés pour résoudre des problèmes de classification et de prédiction, tels que la reconnaissance vocale et la vision par ordinateur. L'apprentissage en profondeur est le moteurde l'évolution récente de l'intelligence artificielle. Le Deep Learning qui se concentre surla construction d'énormes neurones modèles de réseau qui peuvent prendre des décisionsprécises fondées sur des données simulant le comportement d'intelligence humaine.**[24]**Permet Les techniques qui se base le Deep Learning :

1. Réseaux de neurones profonds (Deep Neural Networks DNN)
2. Réseau de Neurones Convolutifs (Convolutional Neural Networks CNN).
3. Réseaux de Neurones Récurrents (Recurrent Neural Networks RNN).
4. Réseaux de neurones adverses (generative adversarial networks GNN).
5. Auto encoders.
6. Memoire à court et long Terme (Long short-term memory LSTM).

2.5 L'apprentissage automatique pour la détection de l'hépatite :

L'utilisation de l'apprentissage automatique dans les soins de santé ouvre de nouvelles perspectives pour le diagnostic et la gestion des maladies, telles que l'hépatite C. Grâce à des algorithmes spécifiques, il devient possible d'analyser les données des patients de manière rapide et précise afin d'identifier ceux qui sont atteints de l'hépatite C. De plus, cette approche permet de développer des traitements plus efficaces et de surveiller leur efficacité au fil du temps. Les algorithmes d'apprentissage automatique peuvent également aider à déterminer les facteurs de risque potentiels liés à la maladie, offrant ainsi une meilleure compréhension des causes et des stratégies de prévention. En outre, l'apprentissage automatique contribue à la création d'outils prédictifs qui ont le potentiel d'anticiper et de prévenir les complications associées à l'hépatite C.

Dans ce texte, nous explorons l'utilisation de l'apprentissage automatique dans le domaine médical, en mettant l'accent sur son application au diagnostic de l'hépatite C. Nous discutons des méthodes et des algorithmes utilisés dans ce contexte, ainsi que de leurs diverses applications pour la détection de la maladie.

En outre, nous examinons quelques études de recherche antérieures qui ont été menées dans ce domaine, offrant ainsi un aperçu des progrès réalisés et des résultats obtenus grâce à l'utilisation de l'apprentissage automatique dans le domaine de la santé, en particulier pour la détection de l'hépatite C :

(Muhammad Bilal Butt,Majed Alfayad, Shazia Saqib, M. A. Khan , Munir Ahmad , Muhammad Adnan Khan , and Nouh Sabri Elmitwally 2021), Cette étude présente un système intelligent de diagnostic du stade de l'hépatite C (IHSDS) utilisant l'apprentissage automatique pour prédire le stade de l'hépatite C chez les patients, en utilisant un réseau neuronal artificiel (ANN). L'ensemble de données utilisé dans l'étude est issu du référentiel d'apprentissage automatique UCI et contient 29 fonctionnalités, dont 19 ont été sélectionnées pour l'étude. 70% de l'ensemble de données a été utilisé pour l'entrainement et 30% à des fins de validation. La précision de l'IHSDS proposé a été comparée à celle des modèles précédemment présentés, et a atteint une précision de 98,89% pendant l'entrainement et une précision de 94,44% lors de la validation.**[10]**

(Heba Mamdouh Farghaly, Mahmoud Y. Shams, Tarek Abd El-Hafeez, 2023), Cet article présente un cadre de prédiction basé sur l'apprentissage automatique pour prédire le virus de l'hépatite C chez les travailleurs de la santé en Égypte. Les auteurs ont utilisé des données du National Liver Institute de l'Université de Menoufiya et comportent 859 patients avec 12 caractéristiques différentes. Ont effectué deux scénarios pour garantir la fiabilité de leur modèle : le premier sans sélection de fonctionnalités et le second après sélection de caractéristiques sur la base de la sélection avant séquentielle (SFS). Les algorithmes d'apprentissage, dont Naïve Bayes, la forêt aléatoire (RF), le voisin le plus proche K et la régression logistique ont été utilisés. Les résultats ont montré que le cadre proposé a obtenu des précisions plus élevées après la sélection SFS. Après le réglage des paramètres, le classificateur RF a obtenu une précision de classification de 94,88% avec seulement quatre fonctionnalités, et un temps d'apprentissage minimum de seulement 0,54 seconde. Table1 représenter la Comparaison des performances entre le classificateur RF en utilisant les paramètres par défaut, le réglage des paramètres et sélection de fonction avec le nouveau paramètre.**[29]**

Measure	Default parameters	Tuning parameters	Feature selection using new parameters
Accuracy	94.06%	94.29%	**94.88%**
No. of features	11	11	4

TABLE 2.1 – Comparaison des performances entre le classifieur RF utilisant les paramètres par défaut, l'ajustement des paramètres et la sélection des caractéristiques avec les nouveaux paramètres.**[29]**

(Taher M. Ghazal, Marrium Anam, Mohammad Kamrul Hasan , Muzammil Hussain, Muhammad Sajid Farooq , Hafiz Muhammad Ammar Ali , Munir Ahmad and Tariq Rahim Soomro , 2021) , objectif de cet article vise à fournir une approche efficace et efficiente pour prédire le stade de l'hépatite C. Pour cela, la technique proposée utilise un Algorithme d'apprentissage SVM gaussien fin, fournissant des résultats précis à 97,9 %. Table 2 montre une comparaison de la méthode proposée avec d'autres. **[18]**

Sr no.	Algorithm	Accuracy (%)
1	Decision tree	84.8
2	Fine-Gaussian SVM (proposed)	97.9

TABLE 2.2 – La comparaison de la méthode proposée avec d'autres méthodes.**[18]**

(TZUU-HSENG S. LI , (Member, IEEE), HUAN-JUNG CHIU , AND PING-HUAN KUO , (Associate Member, IEEE),2022), Cette étude vise à développer un modèle basé sur l'intelligence artificielle pour identifier les patients atteints du virus de l'hépatite C, détecter la maladie à un stade précoce et maintenir la performance des ensembles de données déséquilibrés. Les auteurs ont utilisé l'algorithme SMOTE pour traiter les données cliniques et ont combiné les algorithmes RF et LR pour créer deux modèles en cascade. L'algorithme ABC a été utilisé pour obtenir un modèle final nommé Cascade RF-LR (avec SMOTE), qui permet de détecter avec précision les probabilités multi classes de l'incidence du VHC et améliore l'efficacité des traitements pertinents. Les résultats ont été comparés à ceux d'autres algorithmes et ont montré que le modèle proposé est efficace pour détecter le virus de l'hépatite C à un stade précoce.**[28]**

(Ali Mohd Ali, Mohammad R. Hassan, Faisal Aburub, Mohammad Alauthman, Amjad Aldweesh, Ahmad Al-Qerem, Issam Jebreen and Ahmad Nabot, 2023) L'objectif de cette étude est d'évaluer la performance de plusieurs algorithmes d'apprentissage automatique dans le diagnostic des maladies hépatiques chroniques, en se concentrant sur l'hépatite C pour améliorer l'efficacité et la rentabilité du processus de diagnostic. Les auteurs ont utilisé un ensemble de données exhaustif de 1801 dossiers de patients provenant de l'Hôpital universitaire de Jordanie, chacun avec douze caractéristiques distinctes, et ont mené deux scénarios de recherche pour identifier les caractéristiques les plus pertinentes et améliorer la précision du modèle. La méthode Sequential Forward Selection (SFS) a été utilisée pour cela, ainsi que la technique de suréchantillonnage minoritaire synthétique (SMOTE) pour évaluer l'effet sur la précision des prédictions. Les modèles ont atteint une précision moyenne de 83%, sans effet significatif de SMOTE, et les explications additives de Shapley (SHAP) ont permis de résoudre le problème de l'interprétabilité des modèles. Cette étude met en avant l'importance de la rentabilité, de l'efficacité et de l'interprétabilité dans le diagnostic médical des maladies hépatiques chroniques, en se concentrant sur l'hépatite C.
Table 3 : montre une comparaison des performances de cinq modèles de machine learning différents (Logistic Régression (LR), K-plus proches voisins (KNN), arbre de décision (DT), forêt aléatoire (RF), et Neural Network (NN)) pour la classification binaire. Elle évalue l'utilisation de la technique de suréchantillonnage synthétique des minorités (SMOTE) en comparant les résultats avec et sans SMOTE. Les performances de chaque modèle ont été évaluées à l'aide de quatre mesures d'évaluation, à savoir la précision, le rappel, la précision et le F-mesure.**[1]**

		Without SMOTE				SMOTE			
	FS	Accuracy	Recall	Precision	F-measure	Accuracy	Recall	Precision	F-measure
LR	No	0.822	0.41	0.5	0.45	0.829	0.41	0.5	0.45
	Yes	0.829	0.41	0.5	0.45	0.823	0.41	0.5	0.45
KNN	No	0.831	0.92	0.51	0.46	0.831	0.92	0.51	0.46
	Yes	0.831	0.92	0.51	0.46	0.831	0.92	0.51	0.46
DT	No	0.820	0.41	0.49	0.45	0.820	0.41	0.49	0.45
	Yes	0.813	0.41	0.49	0.45	0.813	0.41	0.49	0.45
RF	No	0.822	0.57	0.51	0.43	0.824	0.51	0.58	0.76
	Yes	0.822	0.54	0.5	0.47	0.822	0.54	0.5	0.47
NN	No	0.829	0.41	0.5	0.45	0.822	0.41	0.5	0.45
	Yes	0.829	0.41	0.5	0.45	0.829	0.41	0.5	0.45

TABLE 2.3 – Performance de différents modèles d'apprentissage automatique.**[1]**

2.6 Conclusion :

Dans ce chapitre, nous avons introduit le concept d'intelligence artificielle et d'apprentissage automatique, puis nous avons expliqué en détail le terme « Machine Learning », ses méthodes, modèles, pertinence et avantages. Cette technologie peut être utilisée pour résoudre des problèmes complexes. Les algorithmes d'apprentissage automatique ont fourni des solutions précises et fiables à de nombreux problèmes. Des algorithmes ont également été utilisés pour cerner les tendances et les modèles à l'aide de données, ce qui a permis de prendre des décisions plus éclairées. Les progrès récents de l'apprentissage automatique permettent à la santé de tirer le meilleur parti de cette technologie, lui permettant de prendre des décisions plus rapidement et plus efficacement. Enfin, cette étude a montré que l'utilisation de l'apprentissage automatique est un outil puissant qui peut être utilisé pour améliorer la qualité de vie de l'être humain dans les différents domaines.

Chapitre 3

Chapitre 3 : Approche proposé.

3.1 Introduction

Dans le domaine du ML, l'évaluation des modèles est considérée comme la tâche la plus importante. Dans ce chapitre, nous commencerons par présenter une étude technique détaillée qui explique notre méthodologie des modèles sélectionnés. Ensuite, nous décrirons notre ensemble de données en détail, en fournissant une description de ses caractéristiques. Enfin, nous aborderons la phase de préparation des donnéeset mettrons en œuvre le code correspondant.

3.2 Méthodologie :

En se basant sur les études antérieures mentionnées dans le chapitre précédent, notre objectif est de développer un modèle capable de surpasser les modèles existants en termes d'exactitude et de précision pour la détection de virus de l'hépatite C.
Pour atteindre cet objectif, nous avons mis en place un processus en plusieurs étapes :
Tout d'abord, nous avons importé l'ensemble de données à partir du site Kaggle. Puis, nous avons nettoyé et normalisé les données pour les préparer à l'entraînement des modèles.
Ensuite, nous avons divisé l'ensemble de données en deux sous-ensembles : l'ensemble d'entraînement et l'ensemble de test.
Nous avons entraîné le modèle d'apprentissage automatique en utilisant des algorithmes tels que LR, DT, KNN, RF et le SVM pour détecter l'hépatite C à partir de dataset préparées.
Enfin, nous avons évalué les performances du modèle en utilisant des métriques d'évaluation telles que la précision, le rappale, le score F1, Spécificité et l'accuracy.
La figure (3.1) illustre les étapes de notre travail, de l'importation des données à l'obtention des résultats.

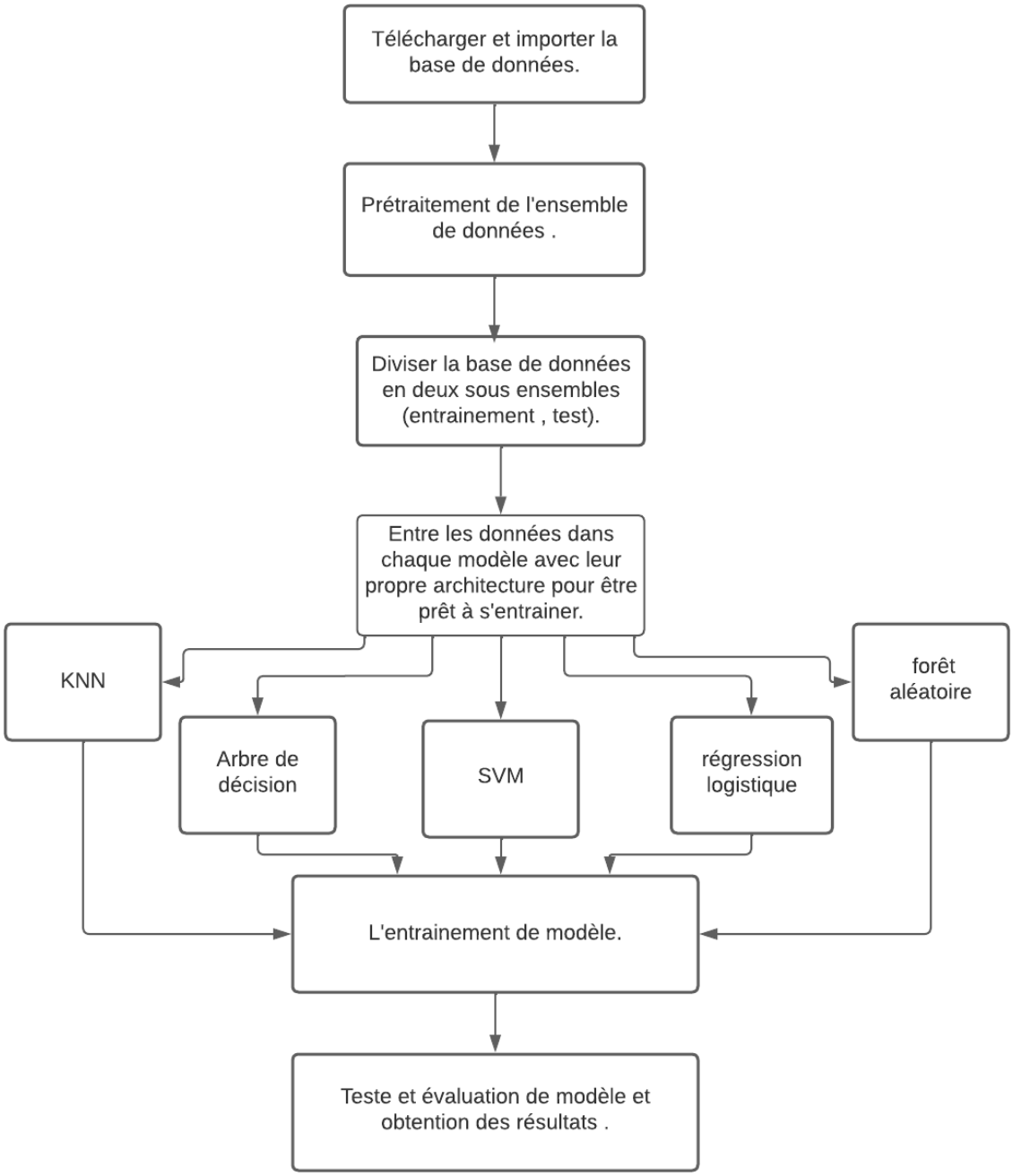


FIGURE 3.1 – Le flux générale du processus.

3.3 Préparation des données :

3.3.1 l'ensemble de données :

3.3.1.1 Définition l'ensemble de données utilisée :

L'ensemble de données est disponible sur le site Kaggle.com et comprend 615 lignes et 13 colonnes. Il contient 12 entrées qui décrivent les analyses effectuées, l'âge des patients et leur sexe. La sortie est présentée sous cinq catégories : donneur de sang, donneur de sang suspect, hépatite, fibrose et cirrhose.

Pour diagnostiquer l'hépatite, plusieurs tests sont effectués, notamment :
ALB : mesure la quantité d'albumine dans le sang, une faible quantité pouvant indiquer une maladie du foie ou des reins.
ALP : mesure le niveau de phosphatase alcaline dans le sang provenant du foie .
ALT : vérifie les dommages au foie.
AST : vérifie les dommages au foie et le succès du traitement des maladies du foie.
BIL : mesure la quantité de bilirubine, un liquide organique sécrété et recyclé par le foie.
CHE : vérifie le bon fonctionnement du système nerveux.
CHOL : mesure le taux de cholestérol, un corps gras indispensable au fonctionnement de l'organisme.
CREA : mesure la quantité de créatinine dans le sérum, le plasma et l'urine, principalement utilisée comme aide au diagnostic des maladies rénales.
GGT : mesure le niveau de GGT, une enzyme présente principalement dans le foie.
PROT : mesure la quantité de protéines dans le corps.
La figure (3.2) donner un aperçu de l'ensemble de données.

1	,"Category","Age","Sex","ALB","ALP","ALT","AST","BIL","CHE","CHOL","CREA","GGT","PROT"
2	1,"0=Blood Donor",32,"m",38.5,52.5,7.7,22.1,7.5,6.93,3.23,106,12.1,69
3	2,"0=Blood Donor",32,"m",38.5,70.3,18,24.7,3.9,11.17,4.8,74,15.6,76.5
4	3,"0=Blood Donor",32,"m",46.9,74.7,36.2,52.6,6.1,8.84,5.2,86,33.2,79.3
5	4,"0=Blood Donor",32,"m",43.2,52,30.6,22.6,18.9,7.33,4.74,80,33.8,75.7
6	5,"0=Blood Donor",32,"m",39.2,74.1,32.6,24.8,9.6,9.15,4.32,76,29.9,68.7
7	6,"0=Blood Donor",32,"m",41.6,43.3,18.5,19.7,12.3,9.92,6.05,111,91,74
8	7,"0=Blood Donor",32,"m",46.3,41.3,17.5,17.8,8.5,7.01,4.79,70,16.9,74.5
9	8,"0=Blood Donor",32,"m",42.2,41.9,35.8,31.1,16.1,5.82,4.6,109,21.5,67.1
10	9,"0=Blood Donor",32,"m",50.9,65.5,23.2,21.2,6.9,8.69,4.1,83,13.7,71.3
11	10,"0=Blood Donor",32,"m",42.4,86.3,20.3,20,35.2,5.46,4.45,81,15.9,69.9
12	11,"0=Blood Donor",32,"m",44.3,52.3,21.7,22.4,17.2,4.15,3.57,78,24.1,75.4
13	12,"0=Blood Donor",33,"m",46.4,68.2,10.3,20,5.7,7.36,4.3,79,18.7,68.6
14	13,"0=Blood Donor",33,"m",36.3,78.6,23.6,22,7,8.56,5.38,78,19.4,68.7
15	14,"0=Blood Donor",33,"m",39,51.7,15.9,24,6.8,6.46,3.38,65,7,70.4
16	15,"0=Blood Donor",33,"m",38.7,39.8,22.5,23,4.1,4.63,4.97,63,15.2,71.9
17	16,"0=Blood Donor",33,"m",41.8,65,33.1,38,6.6,8.83,4.43,71,24,72.7

FIGURE 3.2 – Data Set.

3.3.2 Equilibrage des données :

Le déséquilibre des données se produit lorsque les catégories dans un ensemble de données ne sont pas équitablement représentées en termes de leur taille ou de leur distribution. Les catégories majoritaires sont celles qui constituent une grande partie de l'ensemble de données, tandis que les catégories minoritaires représentent un pourcentage plus faible. Ce déséquilibre peut conduire à un biais dans le modèle de formation, car celui-ci tend à passer plus de temps à apprendre les catégories majoritaires plutôt que les catégories minoritaires.
Au départ, notre ensemble de données comprend cinq catégories de sortie : "Blood Donor", "Blood Donor Suspect", "Hepatitis", "Fibrosis" et "Cirrhosis". Les classe majoritaire sont "Blood Donor" et "Blood Donor Suspect ", tandis que les trois autres catégories sont considérées comme des classes minoritaires, comme illustré dans la Figure (3.3).

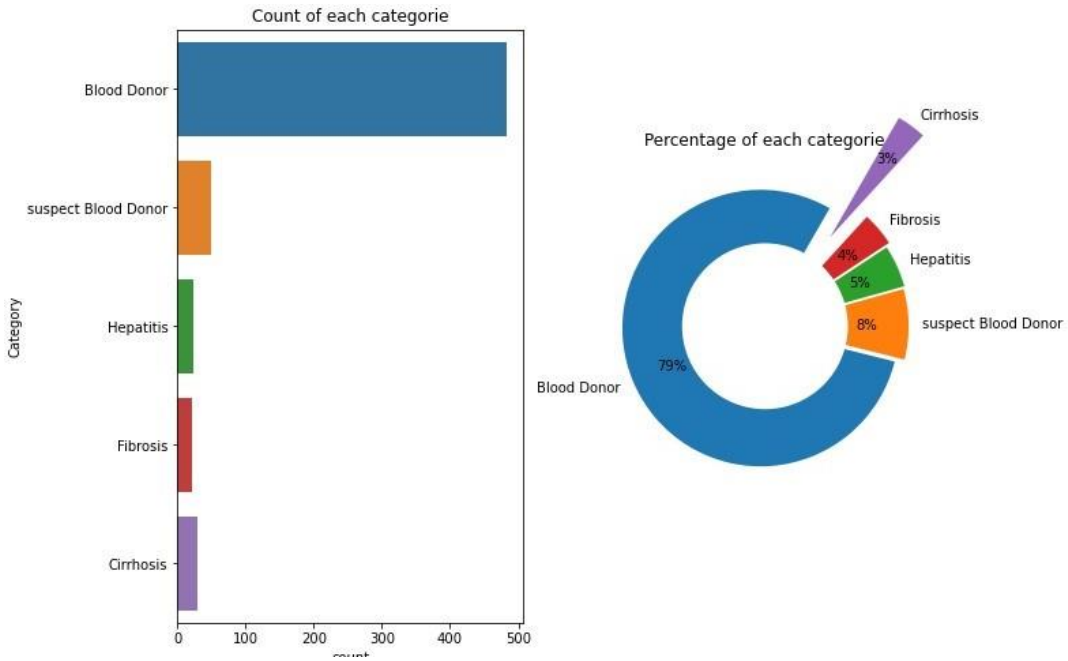


FIGURE 3.3 – Pourcentage des catégories.

Afin de corriger le déséquilibre de notre ensemble de données, nous avons entrepris une transformation consistant à ajouter des lignes supplémentaires aux classes minoritaires. Les valeurs ajoutées ont été choisies de manière à rester dans un champ spécifique pour chaque analyse. Parallèlement, nous avons supprimé les lignes de la classe majoritaire contenant des valeurs similaires qui n'amélioraient pas les résultats. Grâce à cette opération, nous avons obtenu un ensemble de données équilibré, comme illustré dans la Figure 3.4.

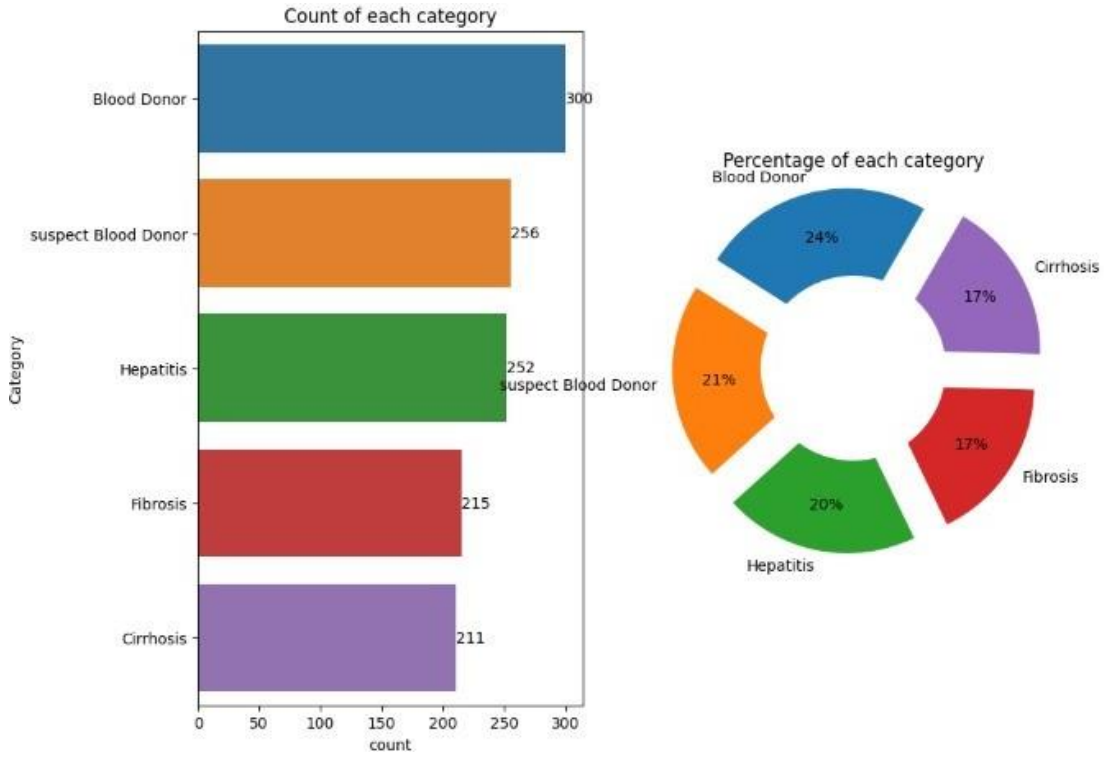


FIGURE 3.4 – Nouveaux Pourcentage des catégories.

3.3.3 Nettoyage des données :

L'objectif de cette étape est de détecter les valeurs nulles (NAN) et les données manquantes pour chaque ligne de notre base de données. Si des lignes contiennent des valeursnulles, nous pouvons remplacer par une moyenne (medain).

3.3.4 Répartition des données de train/test :

Nous avons utilisé la méthode "train_test_split" de la bibliothèque Scikit-learn pour diviser l'ensemble de données en deux ensembles distincts : un ensemble d'entraînement et un ensemble de test. Le paramètre "test_size" est utilisé pour indiquer la proportion des données à conserver pour le test, généralement autour de 30%, tandis que les 70% restants sont utilisés pour l'entraînement.

3.3.5 Transformation et normalisation des données :

3.3.5.1 standard scalaire :

La technique standard scalaire utilisée en prétraitement des données a pour le but de mettre à l'échelle les caractéristiques numériques d'un ensemble de données, afin qu'elles soient comparables entre elles. Elle est particulièrement utile pour réduire les

effets d'échelle indésirables lors de l'utilisation de certains algorithmes tels que les méthodes basées sur la distance, comme le k-plus proches voisins ou le SVM.**[12]**

La formule mathématique pour le standard scalaire est la suivante :

X scaled = X-μ /σ

où

X est la valeur originale, μ est la moyenne des valeurs de la variable et σ est l'écart type.

3.3.6 Choisir les modèles de Machine Learning :

Dans le chapitre précédent, nous avons exposé les différents modèles utilisés en apprentissage supervisé pour la classification. Toutefois, pour résoudre notre problématiquespécifique, nous devons sélectionner un modèle parmi plusieurs options, à savoir LR, KNN, DT ,SVM et RF.

3.3.6.1 Algorithme régression logistique :

Le code source de l'algorithme régression logistique est représenté dans l'image ci-dessous. Après avoir importé et divisé l'ensemble de données "dataset" à l'aide de la méthode "train_test_split", nous appliquons le classifieur de LR.

```
#LogisticRegression
lr = LogisticRegression(C=3)
lr =  OneVsOneClassifier(lr)
lr.fit(X_train, y_train)
def predicttLR(Z):
    return lr.predict(Z)
predictionsLR = predicttLR(X_test)
```

FIGURE 3.5 – code régression logistique.

Avec :

Le paramètre C est utilisé pour contrôler la régularisation dans la régression logistique.

3.3.6.2 Algorithme KNN :

L'image ci dessous montre le code du modèle K-NN : KNeighborsClassifier.

```
#KNN
knn = KNeighborsClassifier(n_neighbors = 8)
knn.fit(X_train, y_train)
def predicttKNN(Z):
    return knn.predict(Z)

predictionsKNN = predicttKNN(X_test)
```

FIGURE 3.6 – code de KNN

Le paramètre `n_neighbros` spécifie le nombre de voisins les plus proches à considérer lors de la classification.
Après nos expérimentations, nous avons constaté que définir le nombre de voisins `n_neighbros` à 8 donne de meilleurs résultats.

3.3.6.3 Algorithme d'arbre de décision :

L'image suivante représente le code source de l'algorithme DT :

```
#DecisionTree

tree = DecisionTreeClassifier(max_depth=10, min_samples_split=2,
                              random_state=42,
                              criterion='entropy',
                              splitter='random',
                              max_features='sqrt')


tree.fit(X_train, y_train)

def predicttDT(Z):
    return tree.predict(Z)

predictionsDT = predicttDT(X_test)
```

FIGURE 3.7 – code arbre de décision .

Les paramètres spécifiés dans le code ont les effets suivants :
- `max_depth` =10 : il restreint la profondeur maximale de l'arbre à 10, contrôlant ainsi sa complexité.
-`min_samples_split` =2 : il détermine le nombre minimum d'échantillons requis pou r diviser un nœud.
- `random_state`=42 : il initialise le générateur de nombres aléatoires afin de garantir la reproductibilité des résultats.
- criterion='entropy' : il utilise l'entropie comme critère pour décider de la division des nœuds de l'arbre.
- splitter='random' : il adopte une approche aléatoire pour sélectionner les divisions des nœuds.
-`max_features` ='sqrt' : il limite le nombre maximal de caractéristiques à prendre en

compte lors de la recherche de la meilleure division, en utilisant la racine carrée du nombre total de caractéristiques.

3.3.6.4 Algorithme de forêt aléatoire :

Nous utilisons la classe RandomForestClassifier pour construire notre forêt aléatoire.

```
#RandomForest
rfc = RandomForestClassifier(n_estimators=1)
rfc.fit(X_train, y_train)
def predicttRF(Z):
    return rfc.predict(Z)
predictionsRF = predicttRF(X_test)
```

FIGURE 3.8 – code de forêt aléatoire .

Le paramètre `n_estimators` spécifie le nombre d'arbres de décision à utiliser dans la forêt aléatoire.

3.3.6.5 SVM :

L'image ci-dessous illustre le code source de l'algorithme SVM ((Support Vector Machine) :

```
#SVM
svmm= svm.SVC(kernel='rbf', gamma = 0.9, C= 1)
svmm =  OneVsOneClassifier(svmm)
svmm.fit(X_train, y_train)
def predicttSVM(Z):
    return svmm.predict(Z)

predictionsSVM = predicttSVM(X_test)
```

FIGURE 3.9 – code de SVM .

Avec :
- kernel='rbf' : il utilise une fonction de base radiale (RBF) comme noyau pour construire l'hyperplan de séparation.
- gamma : il contrôle la flexibilité de la marge et influence la forme de la région de décision.
- C : il contrôle le compromis entre la maximisation de la marge et la minimisation de l'erreur d'entraînement.
- svmm = OneVsOneClassifier(svmm) : Cette ligne enveloppe le modèle SVM svmm dans un classifieur de type "One-vs-One". Cette approche est utilisée pour résoudre des problèmes de classification multiclasse en entraînant plusieurs classifieurs binaires, chacun distinguant entre deux classes différentes.

Après les tests que nous avons réalisés, nous avons trouvé que choisir un noyau RBF en

donnant les valeurs 1 et 0.9 à la constante C et à gamma respectivement, donne des meilleurs résultats.

3.3.7 Évaluation des performances :

Il est essentiel de vérifier la qualité du modèle utilisé, et pour ce faire, différentes mesures d'évaluation doivent être soigneusement sélectionnées, car le choix de mesure peut influencer la manière dont la performance est interprétée et évaluée.

3.3.7.1 Matrice de confusion :

La matrice de confusion est une représentation tabulaire qui permet de visualiser le nombre de prédictions correctes et incorrectes effectuées par un modèle de classification. Cette matrice croise la classification réelle et la classification prédite. **[19]**

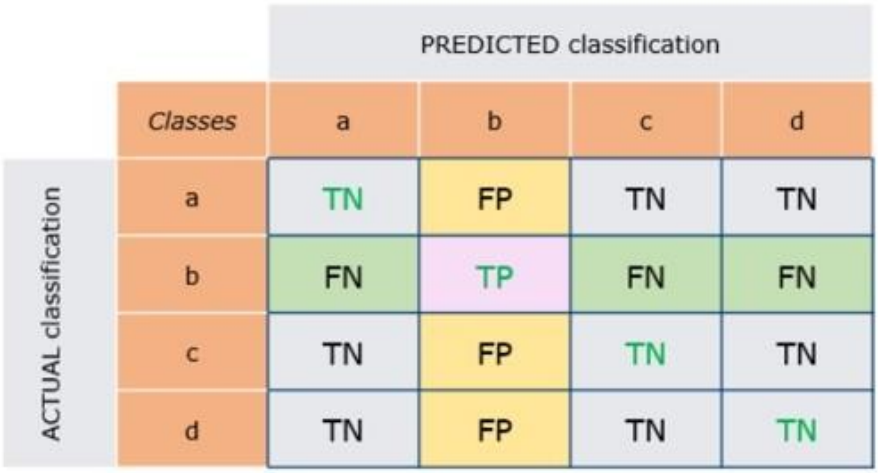


FIGURE 3.10 – matrice de confusion pour une classification multi classe.**[19]**

La matrice de confusion comprend quatre éléments clés :

- True Negative (TN) : les prédictions négatives correctes, où la classe réelle et la classe prédite sont toutes deux négatives.
- True Positive (TP) : les prédictions positives correctes, où la classe réelle et la classe prédite sont toutes deux positives.
- False Positive (FP) : les prédictions positives incorrectes, où la classe réelle est négative mais la classe prédite est positive, ce qui est une erreur de Type 1.
- False Negative (FN) : les prédictions négatives incorrectes, où la classe réelle est positive mais la classe prédite est négative, ce qui est une erreur de Type 2.

3.3.7.2 Mesures de classification :

Une fois que la matrice de confusion est générée, elle peut être utilisée pour effectuer des mesures plus détaillées, permettant d'évaluer de manière plus précise la qualité du modèle.**[38]** Parmi les mesures les plus courantes en classification, on peut citer :

Accuracy : La mesure de Accuracy représente le nombre de prédictions correctes effectuées par le modèle. Elle peut être calculée à partir des valeurs de la matrice de confusion à l'aide de la formule suivante :

$$Accuracy_k = (TP_k + TN_k)/(TP_k + FP_k + FN_k + TN_k)$$

Cette mesure est généralement utilisée lorsque le nombre de TP _k et de TN _k sont les plus importants et permet d'évaluer la performance globale du modèle pour toutes les classes. Le paramètre "k" désigne le numéro de la classe.**[38]**

Précision : Le concept de précision se réfère à la capacité d'un modèle à fournir des résultats exacts, et peut être mesuré en utilisant la formule suivante :**[38]**

$$Prcision_k = TP_k \,/\, (TP_k + FP_k)$$

Où :
TP représente le nombre de vrais positifs
et
FP représente le nombre de faux positifs.

Rappel (Recall) : Le rappel, également connu sous le nom de sensibilité, mesure la proportion de valeurs positives qui ont été correctement prédites par un modèle. Il peut être calculé en utilisant la formule suivante :**[38]**

$$Recall_k = TP_k/(TP_k + FN_k) = 1 - Typre2error$$

F1-Score : Le score F1 combine à la fois les mesures de rappel et de précision en une seule mesure globale de performance. Il est calculé en utilisant la formule suivante :**[38]**

$$F1Score_k = 2 \times (Recall_k * Prcision_k) \,/\, (Recall_k + Prcision_k)$$

où le rappel et la précision sont définis comme précédemment expliqué.

Spécificité (Specificiy) : La spécificité correspond au nombre de classes négatives prédites par le modèle, elle peut se calculer de la manière suivante :**[38]**

$$\text{Spécificité}(k) = \frac{TNk}{TNk + FPk} = 1 - Typre1error$$

3.4 Conclusion :

Dans ce chapitre, nous avons exposé notre méthodologie pour définir clairement les différentes étapes de notre travail. Nous avons commencé par décrire en détail notre dataset, en mettant en avant ses caractéristiques. Ensuite, nous avons sélectionné les algorithmes appropriés pour résoudre notre problème, à savoir : DT, RF, KNN, LR et SVM. Enfin Nous avons ensuite mis en œuvre ces algorithmes pour améliorer notre recherche.

Chapitre 4

Chapitre 4 : Résultats et Discussion

4.1 Introduction

Ce chapitre se concentre sur l'environnement de développement utilisé pour construire nos modèles et évalue leurs performances. Nous présentons également les résultats obtenus pour chaque modèle, suivis d'une brève discussion. De plus, nous comparons nos modèles entre eux ainsi qu'avec les travaux antérieurs pertinents.

4.2 L'environnement de développement :

4.2.1 Python :

Dans notre expérience, nous avons opté pour le langage de programmation Python. Ce dernier est reconnu pour sa facilité d'apprentissage et sa grande puissance. Il propose des structures de données de haut niveau, ainsi qu'une approche simple mais efficace de la programmation orientée objet. Grâce à sa syntaxe élégante et son typage dynamique, Python est souvent utilisé pour l'écriture de scripts et le développement rapide d'applications dans de nombreux domaines, sur la plupart des plates-formes.**[31]**
Nous avons pu mener notre expérience en utilisant les modules suivants :

4.2.1.1 Pandas :

La bibliothèque open-source Pandas est une solution logicielle conçue spécifiquement pour la manipulation et l'analyse de données dans le langage Python. Cette bibliothèque se distingue par sa flexibilité, ses performances et sa facilité d'utilisation. Avec Pandas, Python est désormais capable de charger, aligner, manipuler et fusionner des données de manière efficace. Les performances sont particulièrement remarquables lorsque le code source backend est écrit en C ou en Python. **[2]**

4.2.1.2 NumPY :

NumPy est un package essentiel pour le calcul scientifique en Python. En science informatique, l'algèbre linéaire est largement employée pour convertir des vecteurs en matrices, normaliser des données, effectuer des partitions de données et exécuter diverses opérations mathématiques. NumPy propose des structures de données de tableaux multidimensionnels performantes pour la manipulation de ces données, ainsi qu'un ensemble complet de fonctions mathématiques pour les opérations d'algèbre linéaire et d'autres opérations mathématiques courantes. **[31]**

4.2.1.3 Matplotlib :

C'est une bibliothèque utilisée pour la visualisation de données. Elle regroupe un ensemble de fonctions de style de commande qui permettent à Matplotlib de fonctionner de manière similaire à MATLAB. Par exemple, il est possible de créer une figure, de créer une zone de traçage dans une figure, etc. **[3]**

4.2.1.4 Seaborn :

Seaborn est une bibliothèque de visualisation de données Python basée sur Matplotlib. Elle permet de créer des graphiques statistiques à la fois informatifs et esthétiques, tels que des diagrammes en barres, des histogrammes, etc. Seaborn offre également des fonctionnalités avancées telles que la visualisation de relations linéaires et non linéaires entre des variables, la comparaison de distributions de données et la création de visualisations en facettes. En combinant les fonctionnalités de Matplotlib avec des graphiques plus sophistiqués et modernes, Seaborn est un outil puissant pour explorer et communiquerles tendances et les modèles des données. **[31]**

4.2.1.5 Sklearn :

Scikit-learn est une bibliothèque Python populaire utilisée pour déterminer la précision et le rappel dans les tâches d'apprentissage automatique. Elle offre également des outils pour évaluer et calculer des mesures de performance importantes telles que la matrice de confusion, la précision, la sensibilité, la spécificité et le score F1. **[31]**

4.2.1.6 Django :

Django est un Framework Web puissant développé en Python, qui suit le modèle architectural du contrôleur de vue de modèle. Son objectif principal est de simplifier la création de sites Web complexes qui reposent sur des bases de données.**[42]**

4.2.1.7 HTML :

(HyperText Markup Language) est un langage de balisage utilisé sur le Web pour définir la structure des pages. Il se compose de trois éléments principaux :

- Hypertexte : Il s'agit de texte, souvent accompagné d'éléments tels que des images, qui est organisé de manière à connecter des éléments pertinents entre eux.

- Markup : HTML agit comme un guide de style pour composer tout ce qui doit être affiché, que ce soit sur support papier ou électronique. Il permet de structurer le contenu en utilisant des balises et des attributs pour spécifier la présentation et le formatage.

- Language : HTML est un langage compris et interprété par les systèmes informatiques. Il fournit une syntaxe claire et définie pour décrire la structure et le contenu des pages Web, ce qui permet aux navigateurs Web de les afficher correctement.**[22]**

4.3 Tests et Résultats :

Pour commencer, nous exposons les résultats des étapes abordées dans le chapitre 03 :

4.3.1 Dataset :

Les résultats de l'augmentation du dataset pour les cinq catégories sont présentés dans la Figure 4.1.

My Dataset :

	Number	Category	Age	Sex	ALB	ALP	ALT	AST
0	1	0=Blood Donor	32	m	38.5	52.5	7.7	22.1
1	540	0s=suspect Blood Donor	53	m	19.3	116.0	49.2	146.6
2	548	1=Hepatitis	40	f	49.0	114.0	44.3	96.2
3	584	2=Fibrosis	52	m	41.0	45.2	42.0	34.0
4	615	3=Cirrhosis	59	m	44.0	NaN	100.0	80.0
...	...	...	...	..	...	...	...	...
1229	615	3=Cirrhosis	40	m	44.0	137.8	100.0	80.0
1230	582	2=Fibrosis	49	f	41.0	45.2	46.0	39.0
1231	583	2=Fibrosis	51	f	41.0	45.2	164.0	70.0
1232	584	2=Fibrosis	56	f	41.0	45.2	42.0	34.0
1233	585	2=Fibrosis	75	f	41.0	45.2	114.0	125.0

FIGURE 4.1 – Résultats du Dataset.

4.3.2 Nettoyage de données :

Dans le processus de nettoyage des données, nous avons identifié les valeurs manquantes dans dataset. Après avoir déterminé le nombre de valeurs nulles pour chaque variable, nous avons remplacé ces valeurs par la moyenne (ou la médiane). Le nombre des valeurs nulles pour chaque colonne est présenté dans la figure suivante (Figure 4.2) :

Avant le traitement :

```
Check if there is null data :
 Number       0
Category      0
Age           0
Sex           0
ALB           4
ALP         149
ALT           7
AST           0
BIL           0
CHE           0
CHOL         26
CREA          0
GGT           0
PROT          2
dtype: int64
```

FIGURE 4.2 – Nombre des valeurs nulles avant le nettoyage.

Après le traitement :

```
Number      0
Category    0
Age         0
Sex         0
ALB         0
ALP         0
ALT         0
AST         0
BIL         0
CHE         0
CHOL        0
CREA        0
GGT         0
PROT        0
dtype: int64
```

FIGURE 4.3 – Résultat de nettoyage de données.

4.3.3 Transformation et normalisation des données :

Nous avons utilisé la technique standard scaler pour transformer les données de notre dataset en une représentation numérique significative.

```
Instance 1
[ 0.47050192 -1.74583369  1.53557047  1.12208182 -1.3205397  -0.55737787
 -0.27626369  0.13231951  0.78819993  0.02362074 -0.062402   -0.29171373
 -0.2293503 ]

Instance 2
[ 0.75807942 -0.18185956  1.53557047 -0.55348027  0.02655475 -0.82896415
 -0.23831179  5.05654804 -1.89613903  0.22522766  0.0157792   0.61247462
  0.29523764]

Instance 3
[ 0.46507593  1.22571715 -0.65122378  1.12208182 -1.03681946 -0.69435182
 -0.64715727 -0.1474662   1.16356578  0.65084226 -0.20377075 -0.46866462
  1.28359173]

Instance 4
[ 0.39996404  1.22571715 -0.65122378 -1.90378619  1.06455561  0.24557289
  1.32979175 -0.34611405 -0.75742415 -0.10331695 -0.59574773  2.13957102
  0.39407305]

Instance 5
[-1.03792347  1.14751844 -0.65122378 -0.08038038  0.13496818 -0.31885426
 -0.85416764 -0.01876477 -0.06662482  0.15802535  0.21819355 -0.70596292
 -0.68551373]
```

FIGURE 4.4 – Résultat de transformation de données en utilisant le Standard scaler

4.4 Evaluation de chaque modèle :

Comme discuté précédemment, nous avons réalisé une détection de l'hépatite C en appliquant des algorithmes d'apprentissage automatique supervisé sur notre ensemble de données. Pour ce faire, nous avons crée un code Python qui nous a permis d'effectuer la classification et les tests nécessaires. Les résultats obtenus pour chaque modèle sont les suivants :

Régression logistique :

```
LR Report:
                    precision    recall  f1-score   support

       Blood Donor       1.00      1.00      1.00        96
         Cirrhosis       0.92      0.92      0.92        62
          Fibrosis       0.81      0.85      0.83        71
         Hepatitis       0.86      0.82      0.84        66
suspect Blood Donor       1.00      1.00      1.00        76

          accuracy                           0.92       371
         macro avg       0.92      0.92      0.92       371
      weighted avg       0.92      0.92      0.92       371

LR Accuracy = 0.9245283018867925 %
LR Accuracy : [1.         0.97304582 0.93261456 0.94339623 1.        ]
LR Precision:  [1.         0.91935484 0.81081081 0.85714286 1.        ]
LR Recall:  [1. 1. 1. 1. 1.]
LR Specifficty:  [1.         0.98381877 0.95333333 0.9704918  1.        ]
LR F1 Score :  [1.         0.95798319 0.89552239 0.92307692 1.        ]
```

FIGURE 4.5 – Résultats du modèle Régression logistique.

Le modèle de LR a obtenu une précision globale de 92,45%, ce qui témoigne de sa bonne performance. Les scores d'accuracy spécifiques à chaque classe confirment également la performance générale élevée du modèle, avec des scores élevés observés pour la plupart des classes.

Matrice de confusion :

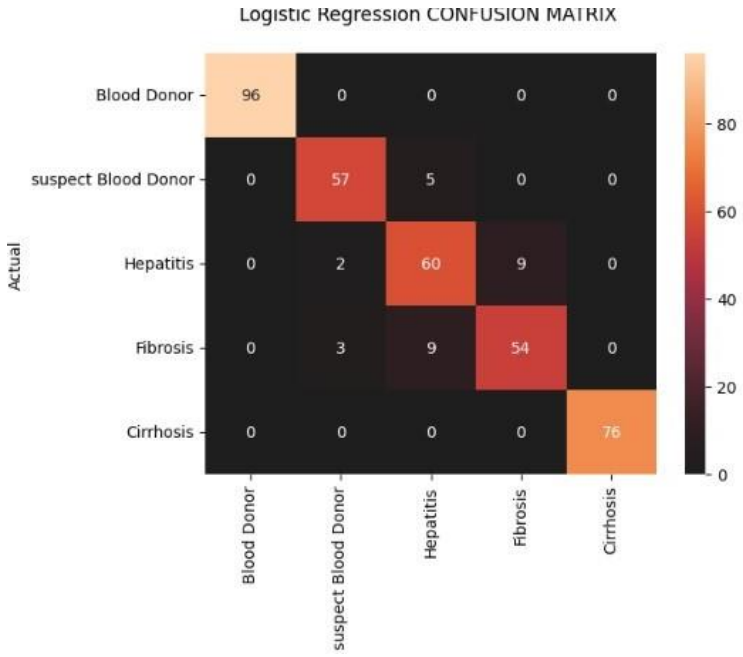


FIGURE 4.6 – Matrice de confusion du modèle Régression logistique.

KNN :

```
KNN Report:
                    precision    recall  f1-score   support

       Blood Donor       1.00      1.00      1.00        96
         Cirrhosis       0.92      0.87      0.89        62
          Fibrosis       0.73      0.93      0.81        71
         Hepatitis       0.98      0.79      0.87        66
suspect Blood Donor      1.00      0.95      0.97        76

          accuracy                           0.92       371
         macro avg       0.92      0.91      0.91       371
      weighted avg       0.93      0.92      0.92       371

KNN Accuracy = 0.9164420485175202 %
KNN Accuracy : [1.         0.96495957 0.91913747 0.95956873 0.98921833]
KNN Precision:  [1.         0.91525424 0.72527473 0.98113208 1.        ]
KNN Recall:  [1. 1. 1. 1. 1.]
KNN Specifficty:  [1.         0.98381877 0.91666667 0.99672131 1.        ]
KNN F1 Score :  [1.         0.95575221 0.84076433 0.99047619 1.        ]
```

FIGURE 4.7 – Résultats du modèle KNN.

Les résultats montrent que le modèle KNN a de bonnes performances avec une précision globale élevée de 92% et des scores élevés pour la plupart des métriques d'évaluation.Cela indique que le modèle est capable de bien généraliser et de prédire avec précision lesétiquettes de classe pour de nouvelles données.

Matrice de confusion :

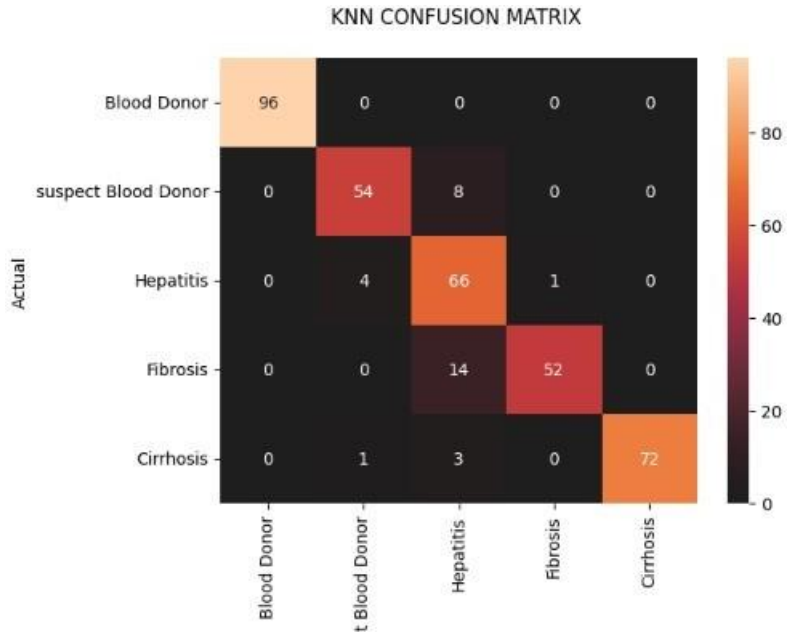


FIGURE 4.8 – Matrice de confusion du modèle KNN.

Arbre de décision :

```
Decision Tree Report:
                    precision    recall  f1-score   support

       Blood Donor       1.00      1.00      1.00        96
         Cirrhosis       0.95      1.00      0.98        62
          Fibrosis       0.98      0.82      0.89        71
         Hepatitis       0.82      0.98      0.90        66
suspect Blood Donor      1.00      0.95      0.97        76

          accuracy                           0.95       371
         macro avg       0.95      0.95      0.95       371
      weighted avg       0.96      0.95      0.95       371

Decision Tree Accuracy = 0.9514824797843666 %
Decision Tree Accuracy : [1.         0.99191375 0.96226415 0.95956873 0.98921833]
Decision Tree Precision:  [1.         0.95384615 0.98305085 0.82278481 1.        ]
Decision Tree Recall:  [1. 1. 1. 1. 1.]
Decision Tree Specifficty:  [1.         0.99029126 0.99666667 0.95409836 1.        ]
Decision Tree F1 Score :  [1.         0.97637795 0.99145299 0.90277778 1.        ]
```

FIGURE 4.9 – Résultats du modèle Arbre de décision.

L'accuracy (précision globale) pour l'algorithme de DT est de 0.95, soit environ 95.15%.

Cela indique que le modèle a correctement classé environ 95.15% des instances dans l'ensemble de données.

En analysant les autres mesures de performance, on peut constater que le modèle de l'arbre de décision obtient des scores élevés de précision, de rappel et de score F1 pour la plupart des classes, ce qui indique une bonne capacité à prédire les étiquettes de classe.

Matrice de confusion :

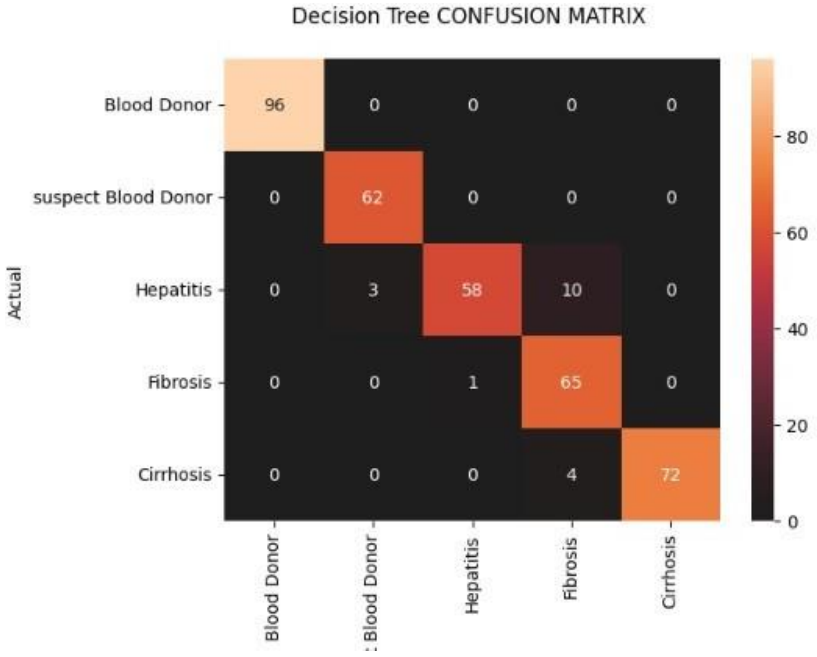


FIGURE 4.10 – Matrice de confusion du modèle Arbre de décision.

RF :

```
RF Report:
                    precision    recall  f1-score   support

       Blood Donor       0.94      0.94      0.94        96
         Cirrhosis       0.89      1.00      0.94        62
          Fibrosis       0.94      0.87      0.91        71
         Hepatitis       0.91      0.89      0.90        66
suspect Blood Donor      1.00      0.97      0.99        76

          accuracy                           0.94       371
         macro avg       0.93      0.94      0.93       371
      weighted avg       0.94      0.94      0.94       371

RF Accuracy = 0.9353099730458221 %
RF Accuracy : [0.99191375 0.97574124 0.98382749 0.98921833 0.97843666]
RF Precision:  [0.97938144 0.87323944 0.98507463 1.         0.95945946]
RF Recall:  [1. 1. 1. 1. 1.]
RF Specifficty:  [0.99272727 0.97087379 0.99666667 1.         0.98983051]
RF F1 Score :  [0.98958333 0.93233083 0.9924812  1.         0.97931034]
```

FIGURE 4.11 – Résultats du modèle forêt aléatoire.

Les résultats obtenus pour le modèle RF sont très prometteurs, avec une précision et une exactitude globale de 94%. Cela indique que le modèle est capable de fournir des prédictions précises et fiables pour les étiquettesde classe dans l'ensemble de données.

Matrice de confusion :

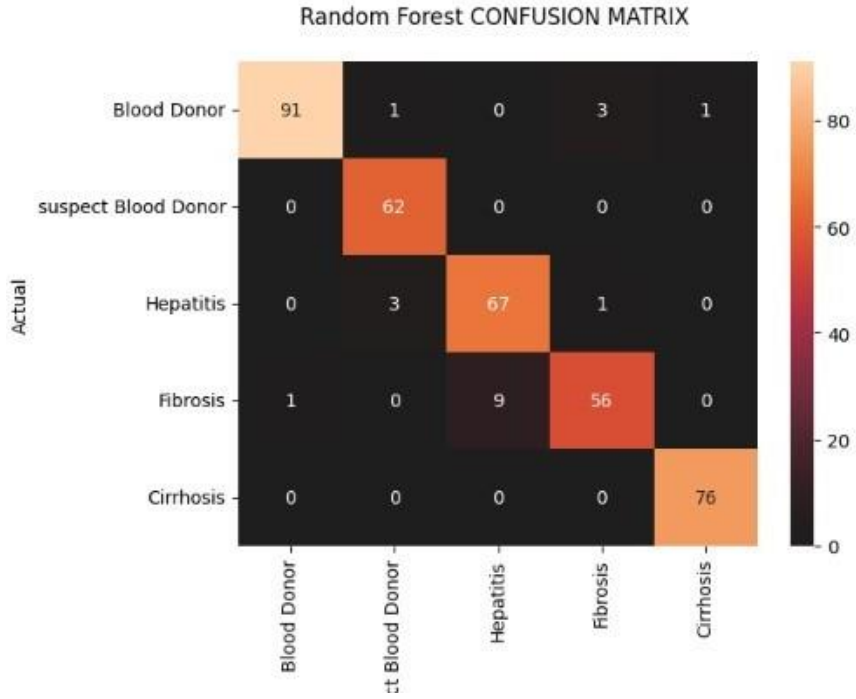


FIGURE 4.12 – Matrice de confusion du modèle forêt aléatoire.

SVM :

```
SVM Report:
                    precision    recall  f1-score   support

       Blood Donor       1.00      0.99      0.99        96
         Cirrhosis       0.84      1.00      0.91        62
          Fibrosis       0.97      0.99      0.98        71
         Hepatitis       0.98      0.91      0.94        66
suspect Blood Donor      1.00      0.91      0.95        76

          accuracy                           0.96       371
         macro avg       0.96      0.96      0.96       371
      weighted avg       0.96      0.96      0.96       371

SVM Accuracy = 0.9595687331536388 %
SVM Accuracy : [0.99730458 0.96765499 0.99191375 0.98113208 0.98113208]
SVM Precision:  [1.         0.83783784 0.97222222 0.98360656 1.        ]
SVM Recall:  [1. 1. 1. 1. 1.]
SVM Specifficty:  [1.         0.96116505 0.99333333 0.99672131 1.        ]
SVM F1 Score :  [1.         0.91176471 0.98591549 0.99173554 1.        ]
```

FIGURE 4.13 – Résultats du modèle SVM.

Les résultats obtenus pour le modèle SVM sont très prometteurs, avec une précision globale élevée de 95.96%. Ces résultats suggèrent que le modèle a la capacité de généraliser efficacement et de prédire avec précision les étiquettes de classe pour de nouvelles données. Les scores élevés dans les autres métriques d'évaluation confirment également la performance solide du modèle. En conclusion, le modèle SVM présente des performances encourageantes et peut être considéré comme une option fiable pour la résolution de ce problème spécifique.

Matrice de confusion :

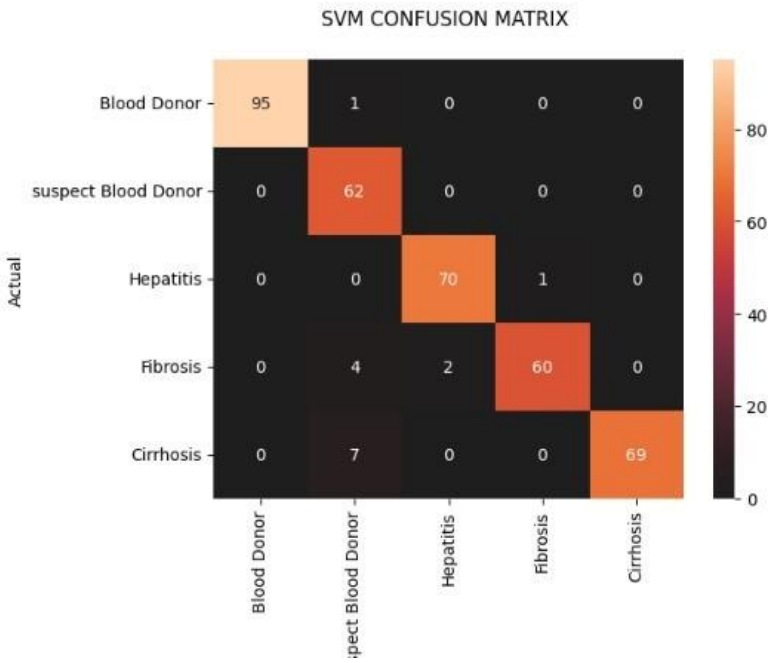


FIGURE 4.14 – Matrice de confusion du modèle SVM.

4.5 La comparaison des modèles :

Une fois les cinq modèles exécutés et les résultats obtenus, nous avons créé le tableau (4.1) pour présenter les performances en termes de précision globale (accuracy).

Les modèles	Régression logistique	KNN	Arbre de décision	Forêt aléatoire	SVM
L'Accuracy	92%	92%	95%	94%	96%

TABLE 4.1 – Le résultat de la précision globale de cinq modèles proposés.

Après avoir appliqué différents algorithmes d'apprentissage automatique et divisé les données, les résultats obtenus révèlent la précision globale de chaque algorithme. Il est remarquable que le SVM obtienne les meilleurs taux de précision, avec des pourcentages élevés. L'arbre de décision affiche un taux de précision de95%, tandis que forêt aléatoire atteint 94%. Le KNN et la régression logistique suivent avec un taux de précision de 92%. Ces résultats mettent en évidence l'efficacité du SVM dans la classification des données, tandis que DT, KNN, LR et RF présentent également de bonnes performances, bien que légèrement inférieures.

Le graphe suivant représente Les ' Accuracy ' des 5 modèles :

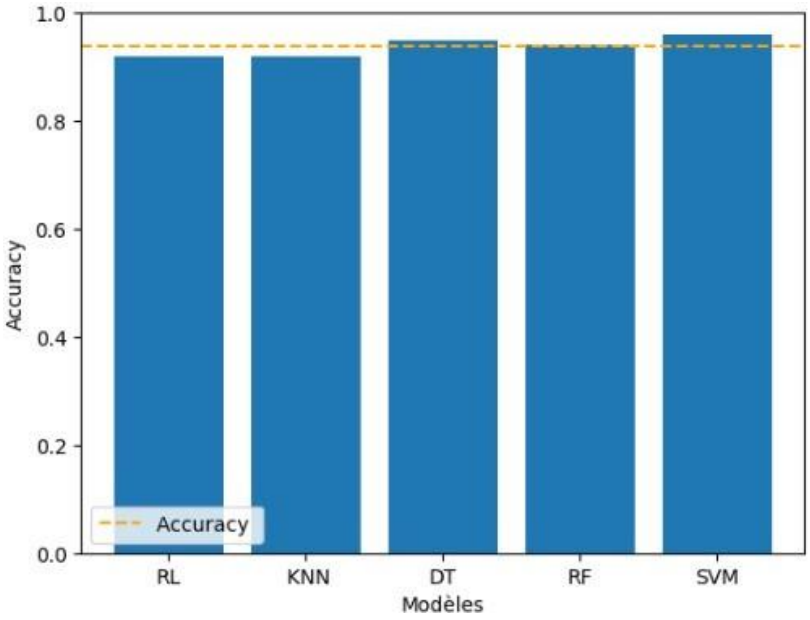


FIGURE 4.15 – Comparaison des modèles basée sur leur précision (accuracy).

4.6 Application :

Nous avons créé une application web simple pour tester notre modèle, qui permet de prédire les catégories de l'hépatite C pour chaque cas. La figure 4.16 présente l'interface de cette application.

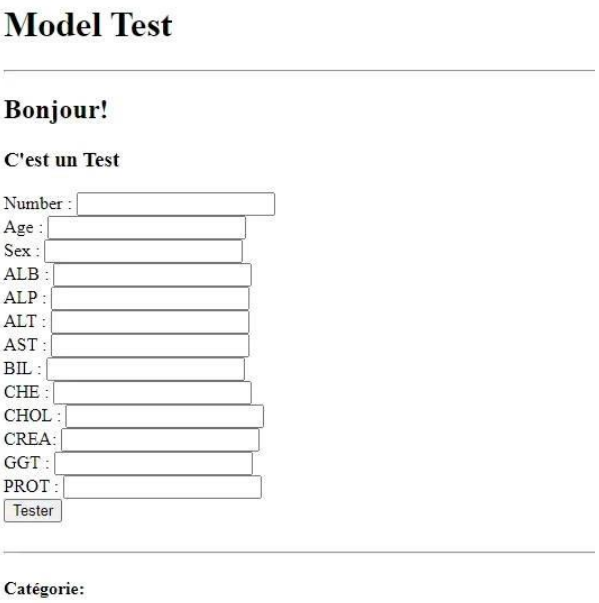


FIGURE 4.16 – Application web de Test.

Voici un exemple de résultats obtenus lors des tests :

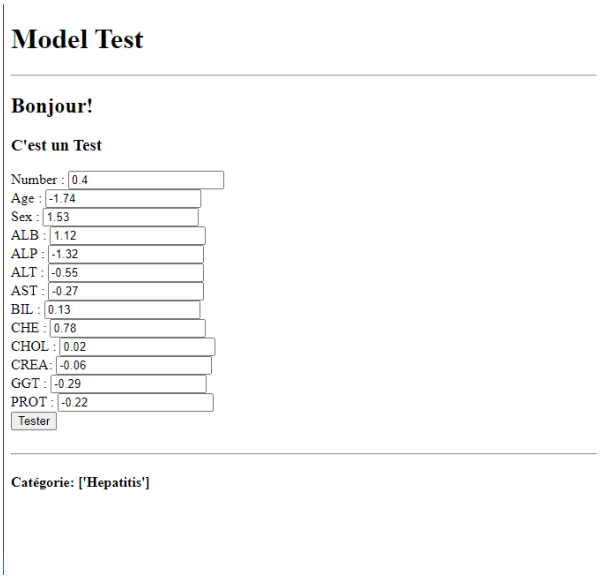


FIGURE 4.17 – Résultats de Test.

4.7 Comparaison notre résultats avec les autres travaux connexes :

Nous avons sélectionné un travail précédemment mentionné dans le chapitre 2 qui a obtenu de bons résultats, puis nous l'avons comparé à notre propre travail. Le tableau 4.2 présente la comparaison des résultats obtenus entre ce travail sélectionné et notre étude.

Choix des modèles pour l'article [1]	Résultats d'accuracy	Les modèles utilisés dans notre étude	Résultats d'accuracy
LR	83%	LR	92%
KNN	83%	KNN	92%
DT	82%	DT	95%
RF	82%	RF	94%
NN	83%	SVM	96%

TABLE 4.2 – Comparaison des résultats obtenus entre le travail sélectionné et notre étude.

Lors de la comparaison entre les résultats des modèles dans l'article et notre étude, des améliorations significatives de l'exactitude ont été observées pour la plupart des modèles. La régression logistique a connu une augmentation de 9 points de pourcentage, passant de 83Lors de la comparaison des résultats entre l'article et notre étude, nous avons observé des améliorations significatives de l'exactitude pour la plupart des modèles. Par exemple, la régression logistique a augmenté de 9 points de pourcentage, passant de 83% à 92% d'exactitude dans notre étude. De même, l'algorithme des k plus proches voisins a connu une amélioration de 9 points de pourcentage, atteignant un taux d'exactitude de 92%. Notre modèle d'arbre de décision a surpassé les résultats de l'article initial, passant de 82% à 95% d'exactitude. De plus, notre modèle de forêt aléatoire a également montré une amélioration en passant de 82% à 94% d'exactitude. Enfin, notre modèle SVM a dépassé le modèle de réseau neuronal mentionné dans l'article, avec un taux d'exactitude de 96% contre 83%. Ces résultats soulignent l'efficacité accrue de nos modèles dans la détection de l'hépatite C, ce qui suggère une amélioration potentielle de la précision du diagnostic de cette maladie. Toutefois, il est important de noter que ces comparaisons ont été effectuées sur des ensembles de données différents, ce qui nécessite une analyse

plus approfondie des mesures de performance et des méthodologies utilisées dans chaque étude pour une évaluation complète de l'efficacité de la détection de l'hépatite C.

4.8 Conclusion :

Dans ce dernier chapitre, nous avons exposé les outils et les langages que nous avons employés pour créer le modèle et l'application web. Nous avons également présenté les résultats de toutes les étapes abordées dans le troisième chapitre, en mettant l'accent sur les résultats des tests. Enfin, nous avons parlé sur un travail similaire au notre dans la dernière section

Conclusion générale

La détection de l'hépatite C revêt une importance capitale dans la lutte contre cette maladie du foie. Les méthodes traditionnelles de dépistage de l'hépatite C reposent généralement sur des tests sanguins pour détecter la présence du virus ou la présence d'anticorps spécifiques. Cependant, ces méthodes peuvent présenter certaines limitations, notamment en termes de sensibilité, de coût et de temps nécessaire pour obtenir les résultats. Par conséquent, une problématique majeure se pose sur la recherche de solutions plus performantes soulève des défis importants sur le plan technique et médical. Mais des avancées dans le domaine de la biotechnologie, de l'apprentissage automatique et de la recherche médicale pourraient apporter des réponses prometteuses.

Ce mémoire a pour objectif la construction d'un modèle de Machine Learning visant à améliorer la détection de l'hépatite C. Le but est de développer un modèle capable de fournir des prédictions aussi précises que possible concernant la détection de cette maladie.

Nos contributions sont regroupées en plusieurs étapes. Tout d'abord, nous avons préparé notre ensemble de données en ajoutant de nouvelles lignes, en supprimant les valeurs manquantes et en effectuant des opérations de nettoyage, de fractionnement et de transformation des données. Ensuite, nous avons sélectionné le modèle approprié parmi plusieurs options, notamment la régression logistique, le SVM, le KNN, DT et RF. Nous avons évalué les performances de ces modèles et nous avons constaté que le SVM était plus performant que les autres. Enfin, nous avons testéle fonctionnement du modèle en l'intégrant dans une application web que nous avons développée spécifiquement à cet effet. Cette dernière étape nous a permis de vérifier la qualité et la fiabilité de notre modèle dans un contexte d'utilisation réelle.

La nature de la science est caractérisée par son caractère inachevé et la perpétuelle génération d'idées nouvelles. C'est pourquoi nous restons engagés à développer constamment ce modèle afin qu'il atteigne son plein potentiel. Nous sommes enthousiastes à l'idée de réaliser une application Android dédiée à la prédiction de l'hépatite C qui pourrait intégrer deux modèles de ML pour améliorer les résultatsde détection et faciliter l'évaluation précoce de la maladie. Notre objectif est de créer une solution pratique et accessible qui puisse contribuer à la lutte contre l'hépatite C et fournir des informations précieuses aux professionnels de la santé et aux patients.

Bibliographie

[1] Ali Mohd ALI et al. “Explainable Machine Learning Approach for Hepatitis C Diagnosis Using SFS Feature Selection”. In : *Machines* 11.3 (2023), p. 391.

[2] ANALYTICS VIDHYA. *Pandas - A Hands-On Guide for Beginners*. https://www.analyticsvidhya.com/blog/2021/07/pandas-a-hands-on-guide-for-beginners/. Consulté le 11 juin 2023.

[3] HE-ARC. *Livres Python - Matplotlib*. https://he-arc.github.io/livre-python/matplotlib/index.html. Consulté le 4 mai 2023.

[4] Tarik ASSELAH, Marie-Pierre RIPAULT et Patrick MARCELLIN. “Hépatite chronique B : qui traiter et comment ?” In : *Gastroentérologie clinique et biologique* 29.4 (2005), p. 374-383.

[5] Kaddour Nesrine BABOURI RADJA. “Étude rétrospective d’une maladie épidémiologique : hépatite C dans la région de Guelma (2011-2013)”. In : (2013).

[6] Boris BARRAUD. *L'intelligence artificielle–Dans toutes ses dimensions*. Université Paris 1 Panthéon-Sorbonne-Université Jiao Tong de Shanghai-L …, 2019.

[7] Souhila BENNAMA et Cherifa BENABED. “Mémoire de fin d’études”. Français. Thème : L’hépatite C : Génétique et épidémiologie. Master en biologie. Mostaganem, Algérie : Faculté des Sciences de la Nature et de la Vie, Université Abdelhamid Ibn Badis-Mostaganem, juin 2020.

[8] Gérard BIAU et Erwan SCORNET. “A random forest guided tour”. In : *Test* 25 (2016), p. 197-227.

[9] Andriy BURKOV. *The hundred-page machine learning book*. T. 1. Andriy Burkov Quebec City, QC, Canada, 2019.

[10] Muhammad Bilal BUtt et al. “Diagnosing the stage of hepatitis C using machine learning”. In : *Journal of Healthcare Engineering* 2021 (2021).

[11] CEdRIc CNAM. *Cours Vertigo : Arbres de décision*. https://cedric.cnam.fr/vertigo/cours/ml2/coursArbresDecision.html. Consulté le 11 avril 2023.

[12] Thara D.K., PremaSudha B.G et Fan XIOnG. “Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques”. In : *Pattern Recognition Letters* 128 (2019), p. 544-550. Issn : 0167-8655. dOI : https://doi.org/10.1016/j.patrec.2019.10.029. URL : https://www.sciencedirect.com/science/article/pii/S0167865519303083.

[13] E. DELAROcqUE-AstAGnEAU et al. “Les modes de transmission du virus de l’hépatite C : approches méthodologiques : Modes of hepatitis C virus transmission : methodological approaches”. In : *Revue d’Épidémiologie et de Santé Publique* 54 (2006). Séminaire ANRS : Recherches en santé publique sur les hépatites virales, p. 5-14. Issn : 0398-7620. dOI : https://doi.org/10.1016/S0398-7620(06)76758-X. URL : https://www.sciencedirect.com/science/article/pii/S039876200676758X.

[14] S DEUffIc, T POynARd et A-J VALLERON. “Correlation between hepatitis C virus prevalence and hepatocellular carcinoma mortality in Europe”. In : *Journal of viral hepatitis* 6.5 (1999), p. 411-413.

[15] EURODECISION. *Data Science Algorithms*. https://www.eurodecision.com/algorithmes/data-science. Year Not Specified.

[16] GEEKSfORGEEKS. *Supervised vs Unsupervised Learning*. https://www.geeksforgeeks.org/supervised-unsupervised-learning/. Year Not Specified.

[17] Sylvain GELLY. “Une contribution à l’apprentissage par renforcement : application au Computer Go”. Thèse de doct. Paris 11, 2007.

[18] Taher M GhAZAL. “Hep-pred : hepatitis c staging prediction using fine gaussian svm”. In : *Computers, Materials & Continua* 69.1 (2021), p. 191-203.

[19] Margherita GRANDINI, Enrico BAGLI et Giorgio VISANI. “Metrics for multi-class classification : an overview”. In : *arXiv preprint arXiv:2008.05756* (2020).

[20] GROUPE HLI. *Machine Learning dans l'industrie*. https://www.groupe-hli.com/machine-learning-dans-industrie/. Year Not Specified.

[21] HEPATITES INFO SERVICE. *Prévention de l'hépatite C*. https://www.hepatites-info-service.org/prevention-de-lhepatite-c/. Consulté le 11 mars 2023. s.d.

[22] KOLADE CHRIS. *What is HTML ? Definition and Meaning*. https://www.freecodecamp.org/news/what-is-html-definition-and-meaning/. Consulté le 11 mai 2023.

[23] GQHGAR KUO et al. "An assay for circulating antibodies to a major etiologic virus of human non-A, non-B hepatitis". In : *Science* 244.4902 (1989), p. 362-364.

[24] Stanislas LAULY. "Exploration des réseaux de neurones à base d'autoencodeur dans le cadre de la Modélisation des données textuelles". Thèse de doct. thèse de doctorat, université de Sherbrooke, 2016.

[25] Jean-Louis LAURIERE. "Intelligence artificielle : résolution de problèmes par l'homme et la machine". In : (1987).

[26] Hélène LE GUILLOU-GUILLEMETTE et Véronique APAIRE-MARCHAIS. "Virus de l'hépatite C, aspects virologiques". In : *Actualités Pharmaceutiques* 58.582 (2019), p. 23-26. ISSN : 0515-3700. DOI : https://doi.org/10.1016/j.actpha.2018.11.005. URL : https://www.sciencedirect.com/science/article/pii/S0515370018304348.

[27] LEBIGDATA. *Réseau de neurones artificiels - Définition*. https://www.lebigdata.fr/reseau-de-neurones-artificiels-definition. Consulté le 15 mars 2023.

[28] Tzuu-Hseng S LI, Huan-Jung CHIU et Ping-Huan KUO. "Hepatitis C Virus Detection Model by Using Random Forest, Logistic-Regression and ABC Algorithm". In : *IEEE Access* 10 (2022), p. 91045-91058.

[29] Heba MAMDOUH FARGHALY, Mahmoud Y SHAMS et Tarek ABD EL-HAFEEZ. "Hepatitis C Virus prediction based on machine learning framework : a real-world case study in Egypt". In : *Knowledge and Information Systems* 65.6 (2023), p. 2595-2617.

[30] MATHWORKS. *Support Vector Machines*. https://fr.mathworks.com/discovery/support-vector-machine.html. Consulté le 20 mars 2023.

[31] F MENCZER, S FORTUNATO et CA DAVIS. “Python tutorial”. In : *A First Course in Network Science* (2020), p. 221-237.

[32] Andreas C MÜLLER et Sarah GUIDO. *Introduction to machine learning with Python : a guide for data scientists*. ” O’Reilly Media, Inc.”, 2016.

[33] Andrew NG. “Machine learning. coursera”. In : *Standford University* (2016).

[34] OPENCLASSROOMS. *Réduisez la corrélation entre les apprenants faibles à l’aide des forêts aléatoires*. https://openclassrooms.com/fr/courses/4470521-modelisez-vos-donnees-avec-les-methodes-ensemblistes/4664688-reduisez-la-correlation-entre-les-apprenants-faibles-a-l-aide-des-forets-aleatoires. Consulté le 18 mars 2023.

[35] PROHEALTHWARE. *Health IT Benefits for Patients and Health Care Providers*. https://prohealthware.com/fr/health-it-benefits-for-patients-and-health-care-providers/. Consulté le 11 juin 2023. s.d.

[36] Rushikesh PUPALE. “Support vector machines (svm)—an overview”. In : *A post at Towards data science available at https://towardsdatascience. com/https-medium-compupalerushikesh-svm-f4b42800e989* (2018).

[37] PURE STORAGE. *What is Healthcare IT ?* https://www.purestorage.com/fr/knowledge/what-is-healthcare-it.html. Consulté le 11 juin 2023. s.d.

[38] Fatima Zohra RAHMANI. “Une Approche basée TextING pour la classification des services web”. Thèse de doct. Université Ibn Khaldoun-Tiaret-, 2022.

[39] SEENOVATE. “Avis d’expert sur le Deep Learning”. In : *Seenovate Blog* (juin 2023).

[40] AK SERME et al. “Portage du virus de l’hépatite C chez les femmes enceintes et transmission mère-enfant à Ouagadougou, Burkina Faso”. In : *Santé publique : courte note* 2764 (2005), p. 108-109.

[41] Mohamed TAGAJDID et al. “Les nouvelles techniques de génotypage du virus de l’hépatite C”. In : *Journal de Biologie Médicale* 3 (juill. 2014).

[42] TECHOPEDIA. *Django*. https://www.techopedia.com/definition/28227/django. Consulté le 11 juin 2023.

[43] TIBCO. *What is Supervised Learning ?* https://www.tibco.com/fr/reference-center/what-is-supervised-learning. Year Not Specified.

[44] VitalFlux. *K-Nearest Neighbors Explained with Python Examples*. https://vitalflux.com/k-nearest-neighbors-explained-with-python-examples/. Consulté le jour mois année. année de publication.

[45] Wikimemoires. “Apprentissage automatique vs Intelligence artificielle”. In : *Wikimemoires* (déc. 2020).

[46] Shichao Zhang et al. “Learning k for knn classification”. In : *ACM Transactions on Intelligent Systems and Technology (TIST)* 8.3 (2017), p. 1-19.