République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université de Mohamed El Bachir El Ibrahimi de Bordj Bou Arréridj
Faculté des Mathématiques et d'Informatique
Département d'informatique

MEMOIRE

Présenté en vue de l'obtention du diplôme

Master en informatique

Spécialité : Ingénierie de l'informatique décisionnelle

THEME

Application d'aide à la prévention et la prédiction de la maladie du foie

Présenté par :

KHRAMSSIA CHAIMA

KHENICHAT ZAHRA

Soutenu publiquement le : /06/2023

Devant le jury composé de :

Président

Examinateur

Encadreur : Dr.Boutouhami Sara

2022/2023

Dédicace

Je dédie humblement ce travail à mes chers parents, Abdelhakim et Samira. Aucun hommage ne saurait suffire pour exprimer l'amour, l'estime et le respect que je vous porte. Vous représentez pour moi la bonté incarnée et la source de tendresse. Ma mère, tu es la lumière de notre vie, celle qui a tout fait pour notre réussite et notre bonheur.

Je remercie également mes sœurs pour leur aide et leur courage, ainsi que tous ceux qui nous ont soutenues dans l'accomplissement de ce travail.

Je suis reconnaissant envers mes frères, Amine, Taki et Hamza, ainsi qu'envers mes sœurs Imane, Houda et mon fiancé Fadel. Je n'oublie pas non plus ma deuxième famille.

À mes meilleures amies, votre présence dans ma vie apporte un bonheur quotidien, je vous aime.

Enfin, je remercie les personnes que j'ai rencontrées à l'Université et qui sont devenues de véritables amis. Un merci particulier à Zahra, ma collègue, je suis ravie d'avoir partagé cette aventure avec toi.

Chaima.

Dédicace

Je dédie ce travail :

A mon cher père Abdelkader

A ma chère mère Aida

Qui n'ont jamais cessé de formuler des prières à mon égard, de me soutenir

Et de m'épauler pour que je puisse atteindre mes objectifs.

A mes frères Yousef, Djamel et Mohammed

A mes sœurs Djamila et Chaima

Pour leurs soutiens moraux tout au long de mes études,

Et qui m'ont toujours encouragé, et à qui je souhaite plus de succès.

À tous mes amies

A toute personne qui occupe une place dans mon cœur.

Zahra.

Remerciement

Nous commençons par exprimer notre gratitude à Dieu le tout puissant, qui nous a donné la Volonté, le courage et la patience nécessaires pour mener à bien ce travail.

Nous souhaitons adresser nos sincères remerciements à toutes les personnes qui nous ont aidés et ont contribué à la réalisation de ce mémoire. En particulier, nous tenons à remercier

Chaleureusement notre encadrante, Mme ***Boutouhami Sara****, pour sa direction et son accompagnement précieux tout au long de ce travail.*

Nous exprimons également notre gratitude envers les membres du jury pour l'intérêt qu'ils ont porté à notre projet de fin d'études et d'avoir accepté d'examiner notre travail.

Nous souhaitons exprimer notre profonde gratitude envers notre famille et tous nos amis pour leur soutien et leurs encouragements tout au long de ce travail. Leur présence et leur soutien moral ont été des sources d'inspiration et de motivation essentielles.

Résumé

Les maladies du foie suscitent une attention considérable dans la recherche médicale en raison de leur impact sur la santé humaine. Elles restreignent la capacité du foie à accomplir les activités quotidiennes et altèrent la qualité de vie. Les maladies du foie peuvent entraîner des incapacités et de graves complications, en étant parmi les principales causes de décès. La prévention, la prédiction et la gestion de ces maladies constituent une approche de soins de santé visant à aider les personnes touchées à maintenir leur autonomie et à préserver leur santé grâce à la détection précoce, la prévention et la gestion efficace. Une préoccupation courante des individus est de savoir quels changements ils doivent apporter pour se trouver dans une autre catégorie (type de maladie). Cette transition entre les catégories peut être perçue comme une possibilité d'amélioration de l'état de santé d'un patient ou, malheureusement, comme une détérioration de son état de santé.

Dans cette étude, nous utilisons des algorithmes d'apprentissage automatique supervisé, notamment KPP et AD, pour prédire la cirrhose du foie. Pour chaque technique, nous proposons un algorithme permettant de calculer de nouvelles valeurs de paramètres pour la transition d'une catégorie à une autre.

Mots clés : KPP, AD, maladies foie, fouille de données, migration entre classes.

Abstract

Liver diseases are receiving considerable attention in medical research due to their impact on human health. They restrict the ability of the liver to perform daily activities and impair quality of life. Liver disease can lead to disability and serious complications, being among the leading causes of death. The prevention, prediction and management of these diseases is a health care approach to help those affected maintain their independence and preserve their health through early detection, prevention and effective management. A common concern of individuals is what changes they need to make to find themselves in another category (type of illness). This transition between categories can be seen as an opportunity for improvement in a patient's health status or, unfortunately, as a deterioration in their health status.

In this study, we use supervised machine learning algorithms, including KNN and AD, to predict liver cirrhosis. For each technique, we propose an algorithm to calculate new parameter values for the transition from one category to another.

Keywords: KNN, AD, liver disease, search for data, migration between classes.

ملخص

تسبب أمراض الكبد اهتمامًا كبيرًا بالبحث الطبي بسبب تأثيرها على صحة الإنسان. أنها تقيد قدرة الكبد على أداء الأنشطة اليومية وتغيير نوعية الحياة. يمكن أن تؤدي أمراض الكبد إلى إعاقات ومضاعفات خطيرة، كونها من بين الأسباب الرئيسية للوفاة. إن الوقاية من هذه الأمراض والتنبؤ بها وإدارتها هي نهج الرعاية الصحية لمساعدة الأشخاص المتأثرين للحفاظ على صحتهم من خلال الكشف المبكر والوقاية والإدارة الفعالة من القلق المشترك للأفراد و معرفة التغييرات التي يجب عليهم إحضارها لإيجاد أنفسهم في فئة أخرى (نوع المرض). يمكن اعتبار هذا الانتقال بين الفئات بمثابة إمكانية لتحسين حالة صحة المريض أو لسوء الحظ تدهورًا في حالته الصحية.

للتنبؤ بتليف الكبد. لكل ADو KPP في هذه الدراسة، نستخدم خوارزميات التعلم التلقائي الخاضعة للإشراف، وخاصة تقنية، نقدم خوارزمية لحساب قيم المعلمات الجديدة للانتقال من فئة إلى أخرى.

الكلمات المفتاحية:KPP ،AD، مرض الكبد، ابحث عن البيانات، والترحيل بين الفصول.

Table des matières

Liste des abréviations

DM : Data Mining

ML : Machine Learning

LR: Logistic Regression

DT: Decision Tree

RL: Regression Logistique

KPP : K plus proches voisins

KNN: K- Nearest Neighbour

AD : Arbres de décision

SVM: Support Vector Machine

TP: Ture Positive

FP: False Negative

TN: True Negative

RF : Forêts aléatoires

RN : Réseaux de neurones

RB : Réseau bayésien

Liste des figures

Liste des tableaux

Introduction générale

1. Contexte et problématique

Une maladie chronique est une affection de longue durée qui évolue progressivement et peut avoir un impact significatif sur la vie quotidienne d'une personne. Elle peut limiter sa capacité à réaliser des activités quotidiennes et affecter sa qualité de vie. Les maladies chroniques peuvent entraîner des incapacités et des complications graves. Bien que toute personne puisse être touchée par une maladie chronique, il est possible de réduire les risques en identifiant les facteurs de risque et d'aggravation associés.

La cirrhose du foie est une maladie chronique qui se développe progressivement et peut prendre des années avant de présenter des symptômes. La cirrhose est caractérisée par une cicatrisation du tissu hépatique, qui altère la fonction normale du foie. Les causes courantes de la cirrhose sont l'hépatite, l'alcoolisme chronique et la stéatose hépatique non alcoolique. Il est important de diagnostiquer et de traiter la cirrhose du foie le plus tôt possible pour prévenir ou ralentir la progression de la maladie.

Depuis de nombreuses années, les institutions de santé ont choisi de collecter et de stocker des données médicales de patients sur de longues périodes. Ces données fournissent un grand nombre de facteurs et d'indices sur les patients ainsi que leur environnement. L'objectif de ces institutions est d'utiliser des techniques d'apprentissage pour exploiter ces collections et trouver des relations ou des corrélations entre les symptômes, les maladies et les traitements. Cela peut aider à identifier les facteurs de risque de maladie, à faciliter le diagnostic, à prévenir les maladies, à choisir et à surveiller l'efficacité des traitements, et à effectuer une surveillance épidémiologique.

La prévention ou la gestion des facteurs de risque peut permettre de contrôler ou de réduire les effets de nombreuses maladies chroniques. En identifiant les facteurs qui peuvent influencer l'évolution de la maladie et en comprenant leur rôle dans la maladie, il est possible d'être plus vigilant et de prendre des mesures appropriées. L'utilisation de techniques d'apprentissage

automatique s'avère efficace pour aider à prendre des décisions et des prévisions à partir de la grande quantité de données produites par l'industrie des soins de santé.

2. Objectifs et contribution

L'objectif des soins de santé pour les maladies chroniques est de prévenir, prédire et gérer ces maladies afin d'aider les personnes touchées à maintenir leur indépendance et leur bien-être grâce à une détection précoce, une prévention et une gestion efficaces.

La problématique que nous abordons dans notre travail est dans un premier temps l'application de techniques d'apprentissage automatique pour la prédiction de la maladie de foie « la cirrhose ». Les techniques que nous allons utiliser sont : le Plus Proche voisin (KPP) et les Arbres de Décision (AD). Nous avons utilisé les données médicales « Mayo Clinic Primary Biliary Cirrhosis Data ». C'est une base de données sur les maladies de foie que nous avons utilisé pour la construction de deux modèles d'apprentissages capables de prédire (classer) le stade de la maladie de foie (1,2 3 ou bien 4).

La présence de différents stades de maladies nous a incités à considérer la possibilité de passage d'une catégorie à une autre. En d'autres termes, en utilisant les informations disponibles sur une personne en particulier, notre système est capable de prédire le stade de la maladie dont elle souffre, comme le stade 3 par exemple. Les gens sont souvent préoccupés par les changements qu'ils doivent apporter pour éviter de passer à stade différent de la maladie. Ce passage (migration) peut représenter une amélioration ou, malheureusement, une détérioration de l'état de santé du patient. L'objectif principal est de ralentir l'évolution vers une forme invalidante de la maladie. L'idée est donc d'intervenir le plus tôt possible et de manière continue pour prévenir la progression et la gravité de la maladie.

Nous avons développé des algorithmes permettant la migration entre classes pour les deux méthodes de classification KPP et les arbres de décision. Pour KPP, nous avons conçu un algorithme qui calcule de nouveaux paramètres en cherchant parmi les plus proches voisins ceux qui ont réussi la migration vers la classe souhaitée tout en prenant en compte les paramètres que le patient souhaite maintenir ou accepter leur modification. Pour les arbres de décision, nous

avons proposé de stocker l'arbre complet dans une table et d'effectuer des requêtes pour répondre aux exigences de migration vers la classe souhaitée en respectant les contraintes imposées par le patient, telles que le maintien des valeurs initiales ou l'acceptation de leur modification.

Nous avons choisi de développer une application, ce qui est très pratique pour les patients qui souhaitent surveiller leur état de santé et voir comment les changements, même minimes, des paramètres affectent leur santé.

3. Plan du mémoire

Le mémoire se structure de la manière suivante :

Le chapitre 1 expose brièvement le processus de fouille de données (Data Mining DM), ainsi que les méthodes et techniques utilisées dans ce domaine. Nous nous concentrons particulièrement sur les techniques d'apprentissage supervisé qui nous intéressent, à savoir le Plus Proche Voisin (KPP) et les Arbres de Décision (AD).

Le chapitre 2 est divisé en trois parties. La première partie présente la base de données maladie de foie que nous avons utilisée. La deuxième partie traite de la construction du modèle de classification pour chacune des techniques précédemment évoquées et de leur application pour prédire le stade de la maladie de foie chez de nouveaux patients. Dans la troisième partie, nous abordons le principe de la migration entre classes et nous décrivons les algorithmes que nous avons proposés pour chaque technique. Le chapitre se conclut par une discussion des résultats obtenus.

Dans le chapitre 3, nous présentons les différents outils et langages de programmation que nous avons utilisés pour développer notre application, ainsi que ses différentes interfaces.

Enfin, nous clôturons ce mémoire avec une conclusion générale qui résume les contributions de notre travail, ainsi que quelques perspectives pour de futurs développements.

Chapitre 1 : l’apprentissage supervisé

1.1. Introduction

Dans ce chapitre, nous abordons le processus d'extraction de connaissances à partir des données, également connu sous le nom de fouille de données, ainsi que les tâches qui en découlent, en mettant l'accent sur la classification supervisée.

La classification supervisée est l'une des tâches principales de la fouille de données pour extraire des connaissances à partir des données. Elle implique la construction d'un modèle de classification à partir d'un ensemble d'exemples étiquetés avec leur classe, suivie de la prédiction de la classe de nouveaux exemples à l'aide de ce modèle. Cette tâche est importante car elle permet de prendre des décisions basées sur les caractéristiques des données, en les classant dans des catégories prédéfinies (Arlot, 2009). Pour cela, plusieurs techniques d'apprentissage supervisé sont utilisées, telles que le Plus Proche Voisin (KPP) et les Arbres de Décision (AD), qui sont présentées en détail dans ce chapitre.

1.2. Définition de la fouille de données

La fouille de données est une partie intégrante du processus d'extraction de connaissances à partir des données, également connu sous le nom de KDD (Knowledge Discovery from Data) ou datamining. En utilisant des techniques statistiques, mathématiques et de reconnaissance de formes, la fouille de données permet de découvrir des corrélations, des modèles et des tendances générales dans les données. L'analyse de problèmes pour comprendre leurs principes et développer des modèles mathématiques adéquats est une pratique courante en science et en ingénierie moderne (Pang-Ning Tan, 2006).

Avec la croissance rapide des quantités de données stockées et des vitesses de traitement, la fouille de données est devenue une discipline clé pour aider les entreprises et les organisations à prendre des décisions éclairées et à découvrir des informations précieuses à partir de leurs données. Les techniques de fouille de données comprennent l'utilisation de méthodes statistiques,

d'apprentissage automatique et de l'intelligence artificielle pour découvrir des modèles, des tendances et des relations dans les données. Le processus de fouille de données implique souvent la collecte, la préparation, la modélisation et l'interprétation des données. La fouille de données est devenue un domaine de recherche et de pratique important dans de nombreux domaines, notamment la finance, la santé, la sécurité, la vente au détail et les médias sociaux (Java T point).

La fouille de données a acquis une importance économique majeure en raison de sa capacité à optimiser la gestion des ressources humaines et matérielles dans divers domaines tels que l'industrie du crédit, l'optimisation des réservations (avions, hôtels), l'organisation des rayonnages en supermarché, la planification de campagnes publicitaires et de promotions, le diagnostic médical, l'analyse du génome, la classification d'objets astronomiques, le commerce électronique, l'analyse de pratiques commerciales et leurs impacts sur les ventes, la recherche sur le Web, la fouille de textes, et la fouille de séquences pour analyser l'évolution temporelle des données. Ces exemples concrets montrent que la fouille de données est une discipline essentielle pour obtenir des connaissances utiles à partir des vastes quantités de données disponibles dans divers domaines.

1.3. Classification des méthodes de fouille de données

Les méthodes de la fouille de données peuvent être classées selon plusieurs paramètres.

Il existe différentes méthodes de fouille de données, qui peuvent être classées selon leur objectif, leur type de traitement ou leur approche (Plantevit, 2019).

- D'un point de vue traitement, on peut classer les méthodes en deux grandes catégories : les méthodes supervisées et les méthodes non supervisées. Les méthodes supervisées sont utilisées lorsque les données sont étiquetées, c'est-à-dire que chaque exemple est associé à une classe ou à une valeur de sortie. Ces méthodes consistent alors à construire un modèle à partir de ces exemples étiquetés, pour prédire la classe ou la valeur de sortie d'exemples non étiquetés. Les méthodes non supervisées sont utilisées lorsque les données ne sont pas étiquetées, et visent à trouver des structures

ou des régularités dans les données, sans a priori sur les classes ou les valeurs de sortie.

- D'un point de vue objectif, on peut distinguer les méthodes de classification, qui permettent de regrouper des données selon des critères prédéfinis, les méthodes de segmentation, qui cherchent à identifier des sous-groupes homogènes dans les données, et les méthodes de prédiction, qui visent à établir des modèles pour anticiper les valeurs futures d'une variable en fonction des autres variables.
- D'un point de vue approche, on peut distinguer les méthodes paramétriques et les méthodes non paramétriques. Les méthodes paramétriques supposent que les données suivent une distribution connue ou modélisable, et cherchent à estimer les paramètres de cette distribution à partir des données. Les méthodes non paramétriques ne supposent pas de distribution particulière, et cherchent plutôt à identifier des structures ou des régularités dans les données.

1.4. La classification supervisée

La classification supervisée est une tâche de fouille de données qui utilise la connaissance à priori sur l'appartenance d'un exemple à une classe. Elle permet d'apprendre à l'aide d'un ensemble d'entrainement, une procédure de classification qui permet de prédire l'appartenance d'un nouvel exemple à une classe (Lev.Kiw, 2018)**.** C'est une méthode de fouille de données qui consiste à construire un modèle de classification à partir d'un ensemble de données d'entraînement étiquetées par leurs classes. Le modèle ainsi construit est ensuite utilisé pour prédire les classes des nouvelles données non étiquetées.

C'est une technique largement utilisée en médecine pour aider à diagnostiquer et prédire les résultats cliniques. Par exemple, elle peut être utilisée pour identifier les caractéristiques qui sont associées à une maladie particulière, ou pour prédire si un patient est à risque de développer une maladie spécifique en utilisant des données médicales telles que l'historique familial, les résultats de tests sanguins et d'imagerie médicale. La classification supervisée est un outil précieux pour aider les professionnels de la santé à diagnostiquer, prédire et traiter les maladies de manière plus efficace.

Dans la classification supervisée, le nombre ainsi que l'identité des classes sont connus à l'avance. La classification a donc pour objectif d'identifier les classes auxquelles appartiennent des objets à partir de traits descriptifs.

Le fonctionnement de la classification supervisée se décompose en deux points :

1. Le premier est la phase d'apprentissage, tout ce qui est appris par l'algorithme est représenté sous la forme des règles de classification que l'on appelle le modèle d'apprentissage.
2. Le second point est la phase de la classification proprement dite, dans laquelle les données tests vont être utilisées pour estimer la précision des règles de classification générées pendant la première phase. Si la précision du modèle est considérée comme acceptable, les règles de classification peuvent être appliquées à des nouvelles données (Mr Mint) .

La construction d'un modèle prédictif se fait généralement en trois phases :

- Une phase d'entraînement : On utilise l'échantillon d'entraînement pour créer le modèle.
- Une phase de validation : On utilise l'échantillon de validation pour évaluer la performance du modèle sur des données qui n'ont pas servi à l'entraînement, de façon à éviter le sur-apprentissage. La performance du modèle peut se baser sur différents indicateurs.
- Une phase de test : On utilise l'échantillon de test pour évaluer la performance finale du modèle. Elle est utile lorsque l'on souhaite une évaluation rigoureuse de la performance finale du modèle.

Généralement on sépare aléatoirement l'échantillon en trois (un échantillon pour chaque phase).

- L'échantillon d'entraînement comprend généralement entre 50% et 80% des données.

- L’échantillon de validation comprend entre 20% et 40% des données et l’échantillon de test utilise entre 5% et 10% des données (il est fréquent, en pratique, que cette étape soit omise).

Il existe de nombreuses méthodes de classification supervisée :

- K plus proches voisins « KPP »
- Analyse (factorielle) discriminante
- Régression logistique « RL »
- Arbres de décision « AD »
- Forêts aléatoires « RF »
- Réseaux de neurones « RN »
- Support Victor machines « SVM »
- Réseau bayésien « RB »

1.4.1. La méthode de classification « K plus proches voisins K-PP»

La classification du plus proche voisin (ou K-plus proches voisins, en anglais "K-nearest neighbors") est une technique d'apprentissage automatique utilisée pour classer des objets ou des données en fonction de leurs caractéristiques. Cette méthode repose sur le principe que des objets similaires ont tendance à appartenir à la même catégorie. Le fonctionnement de l'algorithme est simple : pour chaque nouvel objet à classer, l'algorithme recherche les K objets les plus similaires dans l'ensemble de données d'entraînement, où K est un nombre entier spécifié à l'avance. Ensuite, la catégorie la plus fréquente parmi les K voisins les plus proches est attribuée à l'objet en question.

La similarité entre les objets peut être calculée à l'aide de différentes mesures, telles que la distance euclidienne, la distance de Manhattan ou la distance de Minkowski. Le choix de la mesure dépend de la nature des données à classer (Arlot, 2009).

Pour pouvoir effectuer une prédiction, K-NN se base sur la base de données pour produire un résultat. Principe de K-NN : dis-moi qui sont tes voisins, je te dirais qui tu es ! K est le

nombre de voisin à vérifier. L'algorithme K-NN se baser sur la base de données en entier. Pour une observation, qui ne fait pas parti de la base de données, qu'on souhaite prédire, l'algorithme va chercher les K instances du jeu de données les plus proches (selon une mesure de similarité) de notre observation. Ensuite pour ces voisins, l'algorithme se basera sur leurs variables de sortie (output variable) pour calculer la valeur de la variable de l'observation qu'on souhaite prédire.

La classification du plus proche voisin est une méthode simple et efficace pour la classification d'objets. Elle ne nécessite pas de connaissances a priori sur la distribution des données et peut être utilisée avec des ensembles de données de différentes tailles.

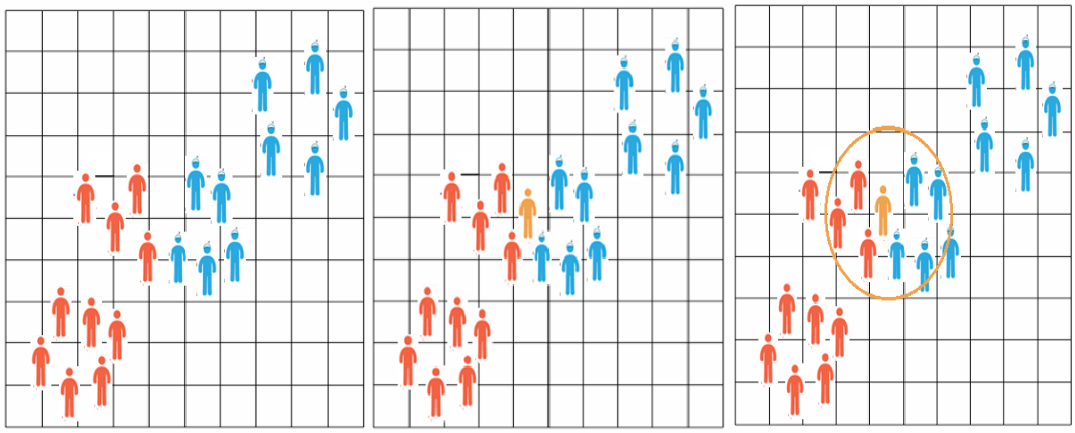

Figure 1: Classification par KPP.

La figure précédente explique mieux le principe du k plus proche voisin. Dans cet exemple les instances sont classées en deux classes soit les hommes bleus ou bien les rouges.

Si on souhaite classer une nouvelle instance l'homme en orange.

L'idée est voir ces voisins (il ressemble à ses voisins). Nous allons voir l'ensemble des voisins S'il parle avec plus des rouges que des bleus, on pourra supposer avec un certain degré de certitude qu'il est rouge. Au contraire, s'il converse avec plusieurs bleus et peu des rouges, on peut assumer qu'il est bleu. La distance qui les sépare est égale à k, la classe du nouvel élément est la classe majoritaire de ses voisins.

Dans cet exemple, on choisit K = 7 voisins proches du sujet mystérieux. Si on trace un cercle autour de lui, on peut voir que de ces 7 individus, il y a trois hommes rouges et quatre bleus.

Donc, on peut assumer que la probabilité selon laquelle l'invité mystérieux est rouge est de 3/7 tandis que la probabilité qu'il soit bleu est de 4/7.

On peut schématiser le fonctionnement de K-NN en l'écrivant en pseudocode suivant :

Début Algorithme

Données en entrée :

- Un ensemble de données **D**.
- Une fonction de définition distance **d**.
- Un nombre entier **K.**

Pour une nouvelle observation **X** dont on veut prédire sa variable de sortie **y** Faire :

1. Calculer toutes les distances de cette observation **X** avec les autres observations du jeu de données **D**
2. Retenir les **K** observations du jeu de données **D** les proches de **X** en utilisant la fonction de calcul de distance **d**
3. Prendre les valeurs de **y** des **K** observations retenues :
   - Si on effectue une régression, calculer la moyenne (ou la médiane) de **y** retenues
   - Si on effectue une classification, calculer le mode de **y** retenues
4. Retourner la valeur calculée dans l'étape 3 comme étant la valeur qui a été prédite par KNN pour l'observation **X.**

Fin Algorithme

Figure 2: Algorithme Méthode KNN.

1.4.1.1. Les mesures de similarités

Comme on vient de le voir l'algorithme K-NN a besoin d'une fonction de calcul de distance entre deux observations. Il existe plusieurs fonctions de calcul de distance : la distance euclidienne, la distance de Manhattan, la distance de Minkowski, celle de Jaccard, la distance Hamming…etc. La fonction de distance est choisie en fonction des types de données manipulées. Ainsi pour les données quantitatives (exemple : poids, salaires, taille, montant de panier électronique etc…) et du même type, la distance euclidienne est un bon candidat. Quant à la distance de Manhattan, elle est une bonne mesure à utiliser quand les données (input variables) ne sont pas du même type (exemple : Age, sexe, longueur, poids etc…) (Benzaki, 2018).

Il est inutile de coder, soi-même ces distances, généralement, les librairies de Machine Learning comme Scikit Learn, effectue ces calculs en interne. Il suffit juste d'indiquer la mesure de distance qu'on souhaite utiliser.

Pour les curieux, voici les définitions mathématiques des distances qu'on vient d'évoquer.

- **La distance euclidienne :**

Distance qui calcule la racine carrée de la somme des différences carrées entre les coordonnées de deux points :

$$De(x, y) = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}$$

- **Distance Manhattan :**

La distance de Manhattan : calcule la somme des valeurs absolues des différences entre les coordonnées de deux points :

$$Dm(x, y) = \sum_{i=1}^{k} |x_i - y_i|$$

- **Distance Minkowski :**

La distance entre deux points donnés est la différence maximale entre leurs coordonnées sur une dimension :

$$d(p,q) = (\sum_{i=1}^{n} |p_i - q_i|^c)^{\frac{1}{c}}$$

- **Distance Hamming :**

La distance entre deux points donnés est la différence maximale entre leurs coordonnées sur une dimension.

$$Dh(x, y) = \sum_{i=1}^{k} |x_i - y_i|$$

- Avec : x = y=> D = 0 / x ≠ y => D = 1

Notez bien qu'il existe d'autres distances selon le cas d'utilisation de l'algorithme, mais la distance euclidienne reste la plus utilisée.

1.4.1.2. Comment choisir la valeur K *?*

Le choix de la valeur K à utiliser pour effectuer une prédiction avec K-NN, varie en fonction du jeu de données. En règle générale, moins on utilisera de voisins (un nombre K petit) plus on sera sujette au problème du sous apprentissage (underfitting). Par ailleurs, plus on utilise de voisins (un nombre K grand) plus, sera fiable dans notre prédiction. Toutefois, si on utilise K nombre de voisins avec K=N et N étant le nombre d'observations, on risque d'avoir le problème de sur-apprentissage (overfitting) et par conséquent un modèle qui se généralise mal sur des observations qu'il n'a pas encore vu (Benzaki, 2018).

Pour sélectionner la valeur de K qui convient aux données, il faut exécuter plusieurs fois l'algorithme KPP avec différentes valeurs de K. Puis choisir le K qui réduit le nombre d'erreurs rencontrées tout en maintenant la capacité de l'algorithme à effectuer des prédictions avec précision lorsqu'il reçoit des données nouvelles.

1 .4.1.3. Avantages de K-NN

- Algorithme simple pour expliquer et comprendre / interpréter.
- Haute précision (relativement).
- L'algorithme est polyvalent. Il peut être utilisé pour la classification, la régression et la recherche d'informations.

1.4.1.4. Inconvénients de K-NN

- Stocke toutes (ou presque toutes) les données d'entraînement.
- L'étape de prédiction peut être lente (avec un grand N).
- Sensible aux fonctionnalités non pertinentes et à l'échelle des données.
- L'algorithme ralentit considérablement à mesure que le nombre d'observations et/ou de variables dépendantes/indépendantes augmente. En effet, l'algorithme parcourt l'ensemble des observations pour calculer chaque distance.
- Pas efficace pour des bases des données larges.
- L'estimation de ce modèle devient de mauvaise qualité quand le nombre de variables explicatives est grand.

1.4.2. La méthode de classification « Arbre de décision »

Les arbres de décision représentent une méthode très efficace d'apprentissage supervisé. C'est une technique d'apprentissage automatique utilisée pour classifier des données en fonction de leurs caractéristiques. Cette méthode est basée sur la construction d'un arbre de décision à partir de l'ensemble de données d'entraînement. La classification par arbre de décision présente plusieurs avantages, notamment sa facilité d'interprétation et d'explication. En effet, l'arbre de décision peut être représenté graphiquement, ce qui permet aux utilisateurs de comprendre facilement comment les décisions de classification sont prises.

Il s'agit de partitionner un ensemble de données en des groupes les plus homogènes possible du point de vue de la variable à prédire. On prend en entrée un ensemble de données

classées, et on fournit en sortie un arbre qui ressemble beaucoup à un diagramme d'orientation où chaque nœud final (feuille) représente une décision (une classe) et chaque nœud non-final (interne) représente un test. Chaque feuille représente la décision d'appartenance à une classe des données vérifiant tous les tests du chemin menant de la racine à cette feuille (Songul, 2016).

Un arbre de décision est un schéma représentant les résultats possibles d'une série de choix interconnectés. Il permet à une personne ou une organisation d'évaluer différentes actions possibles en fonction de leur coût, leur probabilité et leurs bénéfices. Il peut être utilisé pour alimenter une discussion informelle ou pour générer un algorithme qui détermine le meilleur choix de façon mathématique.

Un arbre de décision commence généralement par un nœud d'où découlent plusieurs résultats possibles. Chacun de ces résultats mène à d'autres nœuds, d'où émanent d'autres possibilités. Le schéma ainsi obtenu rappelle la forme d'un arbre. Dans ces structures d'arbre, les feuilles représentent les valeurs de la variable-cible et les embranchements correspondent à des combinaisons de variables d'entrée qui mènent à ces valeurs.

1.4.2.1. Principe de la construction

Au départ, les points de la base d'apprentissage sont tous placés dans le noeud racine. Une des variables de description des points est la classe du point, cette variable est dite « variable cible ». La variable cible peut être catégorielle (problème de classement) ou valeur réelle (problème de régression). Chaque noeud est coupé (opération split) donnant naissance à plusieurs noeuds descendants. Un élément de la base d'apprentissage situé dans un noeud se retrouvera dans un seul de ses descendants. L'arbre est construit par partition récursive de chaque noeud en fonction de la valeur de l'attribut testé à chaque itération (*top-down induction*).Le processus s'arrête quand les éléments d'un nœud ont la même valeur pour la variable cible.

L'idée de construction de l'arbre de décision est simple : il faut commencer par diviser récursivement et le plus efficacement possible les exemples de l'ensemble d'apprentissage par des tests définis à l'aide des attributs jusqu'à ce que l'on obtienne des sous-ensembles d'exemples ne contenant (presque) que des exemples appartenant à une même classe. Cette idée

débouche sur des méthodes de construction *Top-Down*, c'est-à-dire construisant l'arbre de la racine vers les feuilles récursives.

Dans toutes les méthodes, on trouve les trois opérateurs suivants :

- Décider si un nœud est terminal, c'est-à-dire décider si un nœud doit être étiqueté comme une feuille ou porter un test.
- Si un nœud n'est pas terminal, sélectionner un test à lui associer.
- Si un nœud est terminal, lui affecter une classe.

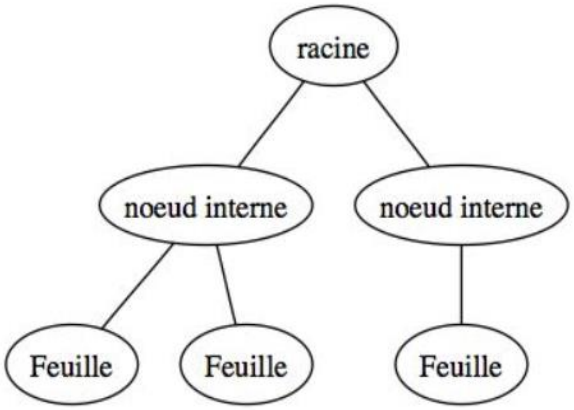

Figure 3: Modèle arbre de décision.

Les mesures de sélection d'attributs :

- **Le Gain**

La formule du gain est utilisée pour mesurer l'utilité d'une caractéristique dans la construction d'un arbre de décision. Elle permet de sélectionner la caractéristique la plus discriminante pour diviser l'ensemble de données en sous-ensembles de classes homogènes. Le gain est calculé en utilisant une mesure d'impureté, telle que l'entropie ou l'indice de Gini, qui permet de quantifier l'hétérogénéité des classes dans l'ensemble de données. Plus l'impureté est élevée, plus les exemples sont mélangés et plus il est difficile de séparer les différentes classes.

La formule du gain peut être exprimée comme suit :

$$Gain(S, A) = Impurté(S) - \sum \frac{|S_v|}{|S|} * Impurté(S_v)$$

La formule du gain mesure la réduction de l'impureté obtenue par la division de l'ensemble de données *S* en sous-ensembles S_v, en utilisant la caractéristique *A*. Plus le gain est élevé, plus la caractéristique *A* est discriminante pour la classification.

- **Le critère de Gini**

Le critère de Gini est utilisé pour sélectionner la caractéristique la plus discriminante pour diviser l'ensemble de données en sous-ensembles plus homogènes. L'objectif est de maximiser la réduction de l'impureté obtenue par la division, mesurée par le gain d'information. Les caractéristiques qui conduisent à une plus grande réduction de l'impureté, selon le critère de Gini, sont considérées comme les plus discriminantes pour la classification.

$$Gini(S) = 1 - \sum (\frac{|C|}{|S|})^2$$

Le critère de Gini est calculé comme la somme des carrés des probabilités d'appartenance à chaque classe, pondérées par le nombre d'exemples dans chaque classe, et soustrait de 1. Cette mesure prend des valeurs entre 0 et 1, où 0 correspond à un ensemble de données parfaitement homogène (tous les exemples appartiennent à la même classe) et 1 correspond à un ensemble de données parfaitement hétérogène (chaque classe contient le même nombre d'exemples).

- **L'entropie**

L'entropie est une mesure d'incertitude utilisée dans la construction d'arbres de décision pour quantifier l'hétérogénéité des classes dans un ensemble de données. Elle est basée sur la probabilité d'occurrence de chaque classe dans l'ensemble. L'entropie de Shannon, la mesure d'entropie la plus couramment utilisée, peut être calculée comme suit :

$$Entropie(S) = -\sum (\frac{|C|}{|S|}) * \log_2 (\frac{|C|}{|S|})$$

Où :

- *Gain (S, A)* est le gain de l'ensemble de données *S* après la division en utilisant la caractéristique *A*.
- *Impureté(S)* est la mesure d'impureté de l'ensemble de données *S*.
- S_v est le sous-ensemble de *S* correspondant à la valeur *v* de la caractéristique *A*.
- $|S_v|$ est le nombre d'exemples dans le sous-ensemble S_v.
- $|S|$ est le nombre total d'exemples dans *S*.
- *Gini(S)* est le critère de Gini pour l'ensemble de données *S*.
- *C*'est une classe dans *S*.
- $|C|$ est le nombre d'exemples dans *S* qui appartiennent à la classe *C*.
- *Entropie(S)* est l'entropie de l'ensemble de données S.

La figure suivante illustre le fonctionnement de l'algorithme de l'arbre de décision :

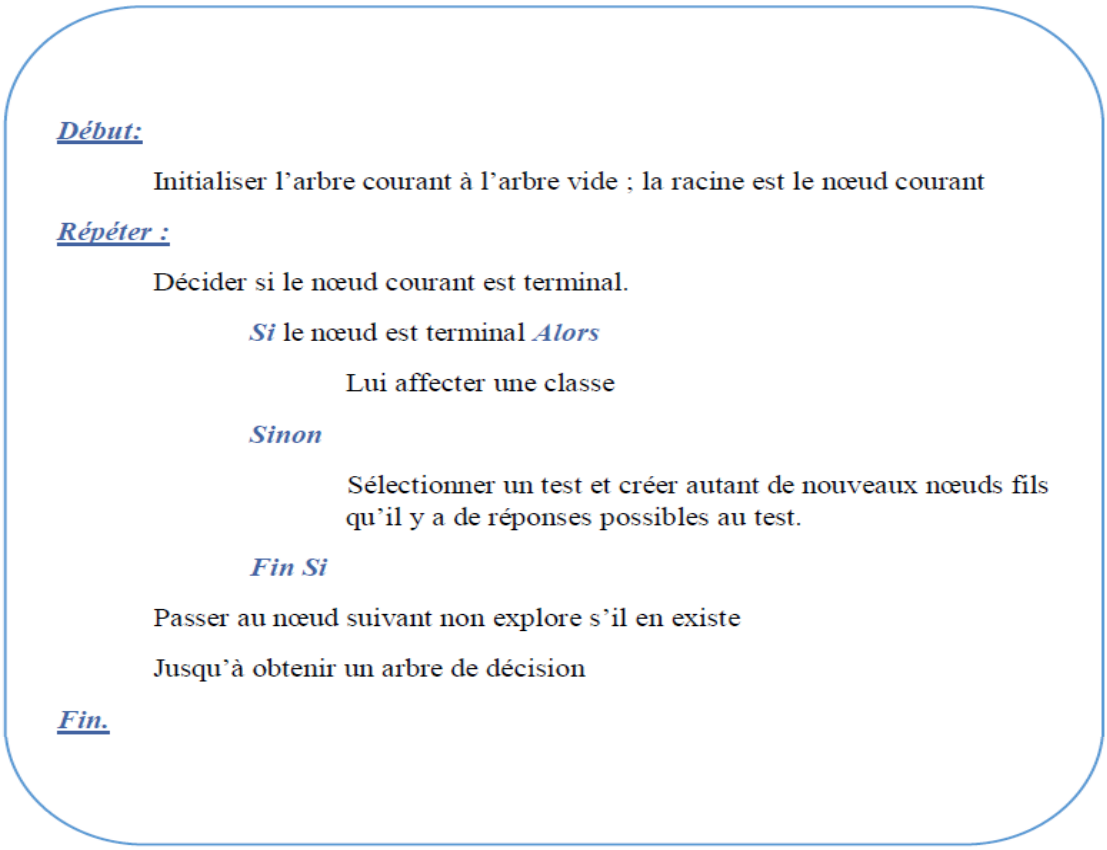

Figure 4: Le fonctionnement de l'algorithme de l'arbre de décision.

Un arbre de décision est considéré comme optimal lorsqu'il représente la plus grande quantité de données possible avec un nombre minimal de niveaux ou de questions. Les algorithmes conçus pour créer des arbres de décision optimisés incluent notamment CART, ASSISTANT, CLS et ID3/4/5. Il est également possible de créer un arbre de décision en générant des règles d'associations, en plaçant la variable cible sur la droite.

Chaque méthode doit déterminer quelle est la meilleure façon de répartir les données à chaque niveau. Les méthodes courantes pour ce faire comprennent l'indice d'impureté de Gini, le gain d'information et la réduction de la variance.

1.4.2.2. Choix de la bonne taille de l'arbre

Il n'est pas toujours souhaitable en pratique de construire un arbre dont les feuilles correspondent à des sous-ensembles parfaitement homogènes du point de vue de la variable-cible. Plus le modèle est complexe (plus l'arbre est grand, plus il a de branches, plus il a de feuilles), plus l'on court le risque de voir ce modèle incapable d'être extrapolé à de nouvelles données, c'est-à-dire de rendre compte de la réalité que l'on cherche à appréhender.

1.4.2.3. Avantages de AD

La popularité des arbres de décision se justifie par les raisons suivantes :

- Ils sont faciles à comprendre.
- Ils peuvent être utiles avec ou sans données concrètes, et les données - quelles qu'elles soient - nécessitent une préparation minimale.
- De nouvelles options peuvent être ajoutées aux arbres existants.
- Ils permettent de sélectionner l'option la plus appropriée parmi plusieurs.
- Le coût d'utilisation de l'arbre pour prédire des données diminue à chaque point de donnée supplémentaire.
- Ils fonctionnent aussi bien pour les données de catégorie que numériques.
- La modélisation des problèmes est possible avec plusieurs données de sortie.
- Ils utilisent un modèle de boîte blanche, ce qui rend les résultats faciles à expliquer.
- La fiabilité d'un arbre peut être testée et quantifiée.

1.4.2.4. Les inconvénients de AD

- Lors de la gestion de données de catégorie comportant plusieurs niveaux, le gain d'information est biaisé en faveur des attributs disposant du plus de niveaux.
- Les calculs peuvent devenir compliqués lorsqu'une certaine incertitude est de mise et que de nombreux résultats sont liés entre eux.
- Les conjonctions entre les nœuds sont limitées à l'opérateur « ET », alors que les graphiques décisionnels permettent de connecter des nœuds avec l'opérateur « OÙ ».

1.5. Conclusion

La collecte de données massives dans différents domaines est en constante augmentation et leur analyse est de plus en plus complexe. Dans ce chapitre, nous avons abordé la problématique de la fouille de données et les différentes techniques utilisées pour extraire des connaissances à partir de ces données. Nous avons ensuite étudié deux techniques d'apprentissage supervisé, les K plus proches voisins et les arbres de décision, que nous allons utiliser dans notre application de prédiction pour le diagnostic médical. Nous avons présenté le fonctionnement de chaque technique ainsi que leurs avantages et inconvénients. Dans le prochain chapitre, nous allons détailler l'application de ces techniques sur des données issues d'une base médicale portant sur la maladie de foie et plus principalement sur le passage entre classe (la migration).

Chapitre 2 : Architecteur de modélisation

2.1. Introduction

L'apprentissage automatique est largement utilisé dans l'industrie des soins de santé pour aider à prendre des décisions et des prévisions en se basant sur les grandes quantités de données disponibles. Les maladies du foie sont une cause majeure de morbidité et de mortalité dans le monde. Dans notre travail, nous avons abordé la problématique de prédire si une personne souffre d'une maladie du foie (classée en 4 types) en utilisant deux techniques d'apprentissage automatique : le plus proche voisin et les arbres de décision. Nous avons utilisé les données médicales « Mayo Clinic Primary Biliary Cirrhosis Data ». C'est une base de données sur les maladies de foie que nous avons utilisé pour la construction de deux modèles d'apprentissages capables de prédire (classer) le stade de la maladie de foie (1,2 3 ou bien 4).

La diversité des types de maladies du foie, nous a menés à la réflexion sur la possibilité de migration d'une classe à une autre. Pour chaque technique, nous avons proposé un algorithme permettant de calculer de nouvelles valeurs des paramètres pour cette migration d'une classe à une autre.

Le schéma suivant illustre de façon exhaustive notre travail :

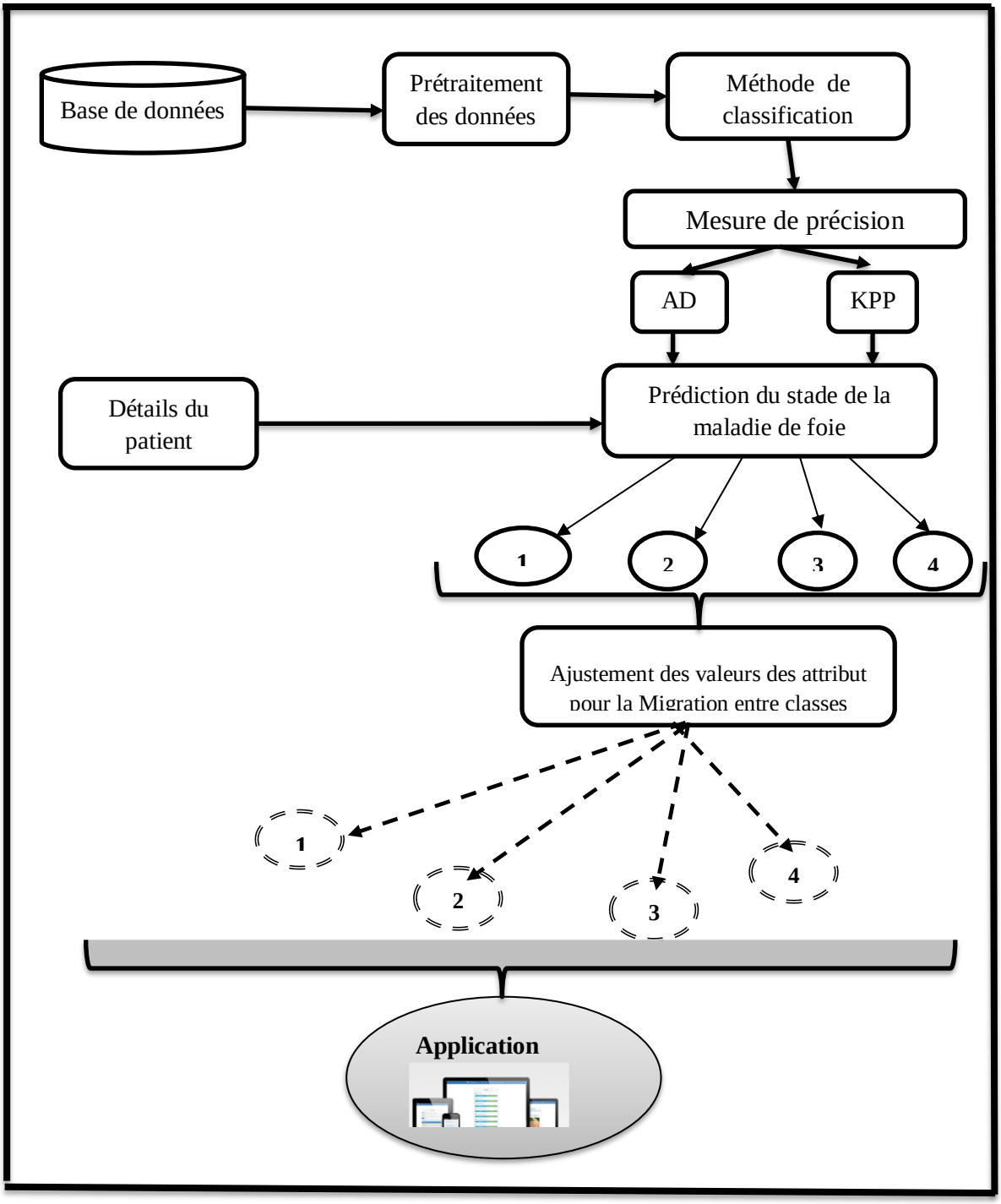

Figure 5: Schéma global de notre architecture.

2.2. La base de données médicale de la maladie de foie

Pour évaluer de manière adéquate notre système proposé, nous avons utilisé une base de données provenant de sources médicales authentiques. Notre choix s'est porté sur cette base de données en raison de la nature non binaire des résultats de classification, qui comporte plutôt

quatre classes distinctes. Cette décision nous offre une plus grande diversité d'options et de perspectives pour notre travail futur.

La base de données "Mayo Clinic Primary Biliary Cirrhosis Data" est une ressource importante que nous avons utilisée dans notre étude. Elle contient des données médicales spécifiques sur la cirrhose biliaire primitive, une forme spécifique de cirrhose du foie. Elle contient des informations cliniques et des résultats de tests biologiques provenant de patients atteints de cirrhose biliaire primitive. En utilisant cette base de données, nous avons pu exploiter les caractéristiques cliniques des patients, telles que les symptômes présents, les résultats des tests sanguins et d'autres facteurs pertinents, pour construire nos modèles d'apprentissage automatique. La base de données "Mayo Clinic Primary Biliary Cirrhosis Data" est une ressource précieuse pour la recherche médicale et la compréhension de la cirrhose biliaire primitive.

Pendant une période de dix ans, un total de 424 patients atteints de cirrhose biliaire primitive (CBP) a été référés à la Mayo Clinic. Parmi ces patients, 312 ont été inclus dans un essai randomisé contrôlé par placebo portant sur le médicament D-pénicillamine. Ces 312 patients constituaient le groupe de participants à l'essai et les données les concernant étaient largement complètes. Les 112 patients restants n'ont pas participé à l'essai clinique, mais ils ont consenti à ce que leurs mesures de base soient enregistrées et à être suivis pour la survie. Toutefois, peu de temps après le diagnostic, six de ces patients ont été perdus de vue, ce qui signifie que les données dont nous disposons ici concernent 106 cas supplémentaires, en plus des 312 participants randomisés.

Il est important de noter que les 312 participants à l'essai randomisé et les 106 cas supplémentaires représentent l'ensemble de données utilisé dans cette étude, soit un total de 418 cas pour les analyses et les évaluations.

1. **N_Days:** le nombre de jours entre l'inscription et le décès, la transplantation ou le moment de l'analyse de l'étude en juillet 1986. Cette variable mesure la durée entre l'enregistrement initial du patient et l'événement le plus précoce survenu parmi le décès, la transplantation ou le moment où les données ont été analysées dans le cadre de l'étude.

2. **Statut :** le statut du patient, indiqué par C (censuré), CL (censuré en raison d'une transplantation hépatique) ou D (décès). Cette variable indique si le patient a été censuré dans l'étude (par exemple, s'il est toujours en vie à la fin de l'étude), s'il a subi une transplantation hépatique ou s'il est décédé. L'étude a examiné les facteurs d'évaluation nutritionnelle chez 74 patients subissant une première transplantation hépatique. Les patients étaient répartis en quatre catégories en fonction du type de maladie hépatique dont ils souffraient. Les résultats de l'évaluation nutritionnelle ont révélé la présence de malnutrition avant l'opération dans tous les groupes, mais chaque groupe présentait des caractéristiques distinctes. Malgré une fonte musculaire et graisseuse extrême, le groupe atteint de cirrhose biliaire primitive semblait avoir la meilleure fonction de synthèse hépatique. Le groupe atteint d'hépatite aiguë était le plus malnutri parmi les différents groupes de maladies, selon tous les critères. Il est recommandé d'encourager un soutien nutritionnel agressif, comprenant un apport adéquat en nutriments et une supplémentation en vitamines et en oligo-éléments, pour tous les patients envisageant une transplantation hépatique.
3. **Drug :** l'étude visait à évaluer l'efficacité de la D-pénicillamine, un médicament utilisé pour le traitement de certaines maladies hépatiques, par rapport à un placebo. Les patients ont été assignés de manière aléatoire à l'un des deux groupes de traitement et ont reçu soit de la D-pénicillamine, soit un placebo pendant la durée de l'étude. Les chercheurs ont ensuite comparé les résultats et les taux de survie entre les deux groupes pour évaluer l'efficacité du médicament.
4. **Age :** l'âge des patients a été enregistré dans le cadre de l'étude. Il s'agit d'une variable démographique importante qui peut être associée à différents résultats, y compris la mortalité et la réponse au traitement.

5. **Sex:** gender of the patient: le sexe des patients a été enregistré dans le cadre de l'étude. Il s'agit d'une variable démographique qui peut également influencer les résultats et la réponse au traitement. Les études révèlent des variations entre les sexes dans les effets des médicaments en termes d'efficacité et de toxicité. Les femmes présentent une sensibilité accrue aux lésions hépatiques aiguës causées par des substances étrangères par rapport aux

hommes. Ces différences sont généralement attribuées à des facteurs physiologiques, à la pharmacocinétique et à la pharmacodynamique, mais aucune de ces explications n'est suffisante pour expliquer pleinement les différentes réponses aux substances étrangères.

6. **Ascites :** l'ascite est une accumulation de liquide dans l'abdomen souvent associée à une cirrhose du foie. Dans cette étude, la présence ou l'absence d'ascite a été enregistrée pour chaque patient, car cela peut être un indicateur de la gravité de la maladie hépatique.
7. **Hepatomegaly:** l'hépatomégalie désigne une augmentation de la taille du foie. Sa présence ou son absence a été enregistrée chez les patients de l'étude, car cela peut être un signe de certaines affections hépatiques.
8. **Spiders :** Les naevus araignée sont de petites dilatations des vaisseaux sanguins présentes à la surface de la peau, souvent associées à des maladies hépatiques. Dans cette étude, la présence ou l'absence de naevus araignée a été enregistrée chez les patients, car cela peut être un signe d'une affection hépatique sous-jacente.
9. **Edema:** l'œdème est l'accumulation de liquide dans les tissus, souvent observée dans les jambes et l'abdomen chez les patients atteints de maladies hépatiques. Dans cette étude, la présence ou l'absence d'œdème a été enregistrée chez les patients.
10. **Bilirubin:** la bilirubine est un pigment jaune produit lors de la dégradation des globules rouges, et des niveaux élevés peuvent indiquer des problèmes hépatiques ou biliaires. Des niveaux de bilirubine plus élevés que d'habitude peuvent indiquer différents types de problèmes de foie ou de voies biliaires. Parfois, des taux de bilirubine plus élevés peuvent être causés par un taux accru de destruction des globules rouges. Les niveaux de bilirubine dans le sang ont été mesurés chez les patients de l'étude.
11. **Cholesterol:** le cholestérol est une substance lipidique présente dans le sang, et des niveaux élevés peuvent être associés à des risques. Le cholestérol et d'autres graisses sont transportés dans votre circulation sanguine sous forme de particules sphériques appelées lipoprotéines. Les deux lipoprotéines les plus connues sont les lipoprotéines de basse densité (LDL) et les lipoprotéines de haute densité (HDL).
12. **Albumin:** l'albumine, produite par le foie, joue un rôle essentiel dans le maintien du liquide dans la circulation sanguine et le transport de diverses substances dans le corps. Un test sanguin d'albumine peut être prescrit pour évaluer le fonctionnement du foie et des reins.

13. **Coppre:** le dosage du cuivre excrété dans l'urine sur une période de 24 heures est utilisé pour diagnostiquer la maladie de Wilson et surveiller son traitement. Des taux élevés de cuivre urinaire sont souvent caractéristiques de cette maladie.
14. **Alk_Phos (alkaline phosphatase):** la phosphatase alcaline (ALP) est une enzyme présente dans tout le corps, principalement produite par le foie et les os. Des niveaux élevés d'ALP dans le sang peuvent indiquer des problèmes hépatiques ou osseux, mais un test d'ALP seul ne permet pas de poser un diagnostic précis.
15. **SGOT :** la SGOT (aspartate aminotransférase) est une enzyme hépatique produite par les cellules du foie. Son niveau dans le sang peut augmenter lorsque les cellules hépatiques sont endommagées, ce qui peut indiquer des problèmes hépatiques.
16. **Triglycérides** : le foie produit des triglycérides qui sont transportés vers les tissus périphériques par les particules de VLDL. Les triglycérides sont des lipides présents dans ces particules et sont impliqués dans le métabolisme énergétique.
17. **Platelets :** la numération plaquettaire peut être utilisée pour détecter une diminution des plaquettes sanguines (thrombocytopénie), qui peut être un signe de cirrhose du foie ou d'autres conditions. Cependant, chez les patients alcooliques, l'alcool lui-même peut entraîner une thrombocytopénie, ce qui rend l'interprétation plus complexe.
18. **Prothrombin:** la prothrombine est une protéine produite par le foie qui contribue à la coagulation sanguine. Le temps de prothrombine (TP) mesure le temps nécessaire pour que le sang forme un caillot. Un TP normal indique que les protéines de coagulation sanguine sont présentes en quantité adéquate. Ce test peut être utilisé pour évaluer la capacité du sang à coaguler correctement.
19. **La variable "Stage"** dans le contexte de la cirrhose du foie fait référence au stade histologique de la maladie. La cirrhose du foie peut être classée en différents stades en fonction de la gravité des lésions et des cicatrices hépatiques. Le système de classification le plus couramment utilisé est la classification de Child-Pugh ou le score du Model for End-Stage Liver Disease (MELD). Le stade histologique de la cirrhose du foie est généralement échelonné de 1 à 4, chaque stade représentant une augmentation progressive de la sévérité de la fibrose et de la cirrhose hépatique. Ces stades sont déterminés par une biopsie hépatique, où un petit échantillon de tissu hépatique est examiné au microscope pour évaluer l'étendue

des cicatrices et de l'inflammation. Les critères spécifiques et les systèmes de notation utilisés pour la classification peuvent varier, mais en général, le stade 1 indique une fibrose minimale ou une cirrhose précoce, tandis que le stade 4 indique une cirrhose avancée avec une fibrose importante et une fonction hépatique altérée.

2.3. Application des méthodes de classification sur la maladie de foie.

Les problèmes d'apprentissage sont énoncés sous forme de données. Ces séries caractérisant une série d'instances du phénomène à apprendre, que l'on nomme patients.

Chaque patient **P** est constitué d'une description **D** et d'une sortie **S**

P = (D, S)

Où :

- D ∈ X = {'Status', 'Drug', 'Age', 'Sex', 'Ascites', 'Hepatomegaly', 'Spiders', 'Edema', 'Bilirubin', 'Cholestérol', 'Albumin', 'Copper', 'Alk_Phos', 'SGOT', 'Tryglicerides', 'Platelets', 'Prothrombin'}
- S ∈ Y = {1, 2, 3,4}

Voici un échantillon de la base de d'apprentissage que nous avons utilisé.

Tableau 1: Echantillon de la base de données.

Status	Drug	Age	Sex	Ascites	Hepatomegaly	Spiders	Edema	Bilirubin	Cholesterol	Albumin	Copper	Alk_Phos	SGOT	Tryglicerides	Platelets	Prothrombin
-1	0	59	1	1	1	1	1	14,5	261	2,6	156	1718	137,95	172	190	12,2
0	0	56	1	0	1	1	0	1,1	302	4,14	54	7394,8	113,52	88	221	10,6
-1	0	70	0	0	0	0	-1	1,4	176	3,48	210	516	96,1	55	151	12
-1	0	55	1	0	1	1	-1	1,8	244	2,54	64	6121,8	60,63	92	183	10,3
1	1	38	1	0	1	1	0	3,4	279	3,53	143	671	113,15	72	136	10,9
-1	1	66	1	0	1	0	0	0,8	248	3,98	50	944	93	63	251	11
0	1	56	1	0	1	0	0	1	322	4,09	52	824	60,45	213	204	9,7
-1	1	53	1	0	0	0	0	0,3	280	4	52	4651,2	28,38	189	373	11
-1	0	43	1	0	0	1	0	3,2	562	3,08	79	2276	144,15	88	251	11
-1	1	71	1	1	0	1	1	12,6	200	2,74	140	918	147,25	143	302	11,5

Soit le nouvel patient P ayant les caractéristiques présentées dans le tableau suivant, que l'on souhaite classifier.

Tableau 2: Les données d'un nouvel patient.

Status	Drug	Age	Sex	Ascites	Hepatomegaly	Spiders	Edema	Bilirubin	Cholesterol	Albumin	Copper	Alk_Phos	SGOT	Tryglicerides	Platelets	Prothrombin	Stage
-1	1	66	1	0	1	0	0	0,8	248	3,98	50	944	93	63	251	11	?

2.3.1. Application de la méthode KPP

2.3.1.1. Classification du nouvel patient P par le KPP

Les étapes pour classifier ce patient P par le KPP sont :

1. Calculer toutes les distances de ce nouveau patient P avec les autres observations (patients) de la base d'apprentissage **D.** Pour calculer la distance on utilisant **La distance euclidienne :** De (x, y) = $\sqrt{\sum_{j=1}^{n}(x_j - y_j)^2}$

   Avec :

   - X_j : La valeur de lattribut de la base d'apprentissage
   - y_j: La valeur de l'attribut D'un patient p

2. Retenir les K observations de la base d'apprentissage D les proches de **P en** utilisation la fonction de calcul de distance **d.**
3. Retourner la valeur calculée **S** comme étant la valeur qui a été prédite par KNN pour l'observation **P.**

On Applique ces étapes à ce patient **P :**

Normalisation toutes des données : Après avoir effectué le processus normalisation, nous obtenons les valeurs suivantes :

Tableau 3: Normalisation de l'échantillon1 de la base médicale.

Status	Drug	Age	Sex	Ascites	epatomega	Spiders	Edema	Bilirubin	Cholestero	Albumin	Copper	Alk_Phos	SGOT	rygliceride	Platelets	rothrombi
-1,15654	-0,76842	0,801912	0,347571	4,458139	0,763504	1,842481	2,755915	2,645406	-0,46424	-2,14895	0,810942	-0,04049	0,307466	0,938197	-0,68441	1,472707
0,53769	-0,76842	0,514429	0,347571	-0,22431	0,763504	1,842481	0,163026	-0,4783	-0,24459	1,502031	-0,49933	3,094972	-0,17191	-0,55559	-0,36204	-0,12692
-1,15654	-0,76842	1,856018	-2,87711	-0,22431	-1,30975	-0,54275	-2,42986	-0,40837	-0,91961	-0,06268	1,504618	-0,70439	-0,51374	-1,14244	-1,08998	1,272754
-1,15654	-0,76842	0,418601	0,347571	-0,22431	0,763504	1,842481	-2,42986	-0,31512	-0,55532	-2,2912	-0,37088	2,391858	-1,20976	-0,48446	-0,75721	-0,42685
2,231922	1,301364	-1,21047	0,347571	-0,22431	0,763504	1,842481	0,163026	0,057859	-0,36781	0,055862	0,643946	-0,61878	-0,17918	-0,84012	-1,24597	0,173013
0,53769	1,301364	0,514429	0,347571	-0,22431	0,763504	-0,54275	0,163026	-0,50161	-0,13745	1,383492	-0,52503	-0,53427	-1,21329	1,667307	-0,53883	-1,0267
-1,15654	1,301364	0,226945	0,347571	-0,22431	-1,30975	-0,54275	0,163026	-0,66479	-0,36245	1,170123	-0,52503	1,579603	-1,84259	1,240511	1,21863	0,27299
-1,15654	-0,76842	-0,73133	0,347571	-0,22431	-1,30975	1,842481	0,163026	0,011237	1,148311	-1,01098	-0,17819	0,267711	0,429127	-0,55559	-0,05007	0,27299
-1,15654	1,301364	0,801912	0,347571	-0,22431	-1,30975	1,842481	0,163026	0,104482	-0,59818	0,032154	0,0145	-0,66296	-0,78748	-0,43111	-1,92191	2,872377
-1,15654	-0,76842	1,376879	0,347571	-0,22431	-1,30975	-0,54275	0,163026	-0,54823	-0,62496	0,861923	1,029321	3,986983	0,10653	-0,41333	0,407497	0,27299
-1,15654	-0,76842	0,322773	0,347571	-0,22431	0,763504	1,842481	2,755915	1,922758	-0,9089	-1,6748	6,360347	-0,4586	3,105655	1,436126	0,282707	1,67266
0,53769	-0,76842	-0,06054	0,347571	-0,22431	0,763504	-0,54275	-2,42986	-0,57154	-0,60353	0,126985	-0,69202	0,049542	-0,57457	0,06682	-0,48683	0,27299
-1,15654	1,301364	0,89774	0,347571	-0,22431	0,763504	-0,54275	0,163026	0,45415	0,141135	0,008447	0,605409	0,07053	0,003315	0,280218	0,688274	2,272518
0,53769	1,301364	1,281051	-2,87711	-0,22431	0,763504	1,842481	0,163026	-0,59486	-0,51246	0,767093	-0,66633	-0,52378	-1,12204	-0,64451	0,833862	0,672895
-1,15654	-0,76842	0,514429	0,347571	-0,22431	-1,30975	1,842481	0,163026	0,057859	-0,41067	0,292939	4,767462	-0,22938	-0,0271	-1,14244	-0,8612	0,872848
-1,15654	1,301364	0,514429	0,347571	4,458139	0,763504	1,842481	2,755915	3,321431	0,253639	-1,34289	5,974972	2,360375	2,055647	1,276077	-0,43483	0,972825
-1,15654	-0,76842	-0,53968	-2,87711	-0,22431	0,763504	-0,54275	0,163026	-0,24519	0,580435	1,170123	0,399875	2,169379	1,954394	1,969622	-1,93231	-0,82675
0,53769	1,301364	-0,53968	0,347571	-0,22431	-1,30975	-0,54275	0,163026	-0,57154	-0,26602	1,4072	-0,67918	-0,6243	-0,30084	-0,94682	0,709072	0,572919
-1,15654	1,301364	0,131118	0,347571	-0,22431	0,763504	1,842481	0,163026	0,477461	4,180555	0,411477	-0,51218	0,793528	0,854938	0,831497	1,717789	-0,82675
-1,15654	1,301364	0,322773	0,347571	4,458139	0,763504	1,842481	-2,42986	4,300503	-0,92497	-0,46571	1,645922	1,052792	-0,39973	0,867064	-1,82832	1,272754

Les valeurs normalisées de patient **p :**

Tableau 4: Normalisation des caractéristiques du patient p.

Status	Drug	Age	Sex	Ascites	epatomega	Spiders	Edema	Bilirubin	Cholestero	Albumin	Copper	Alk_Phos	SGOT	rygliceride	Platelets	rothrombin
2,231922	-0,76842	0,801912	0,347571	4,458139	0,763504	1,842481	2,755915	2,645406	-0,46424	-2,14895	0,810942	-0,04049	0,307466	0,938197	-0,68441	1,472707

Calculer toutes les distances de ce patient **P** avec les autres observations de la base d′apprentissage **D :** pour calculer la distance on utilisant La distance euclidienne

- Après calcul, nous obtenons les valeurs suivantes :

Tableau 5: Les distances des instances de l'échantillon par rapport au nouvel patient.

Status	Drug	Age	Sex	Ascites	epatomega	Spiders	Edema	Bilirubin	Cholestero	Albumin	Copper	Alk_Phos	SGOT	ryglicerid	Platelets	rothrombi	Stage	Distances
-1	0	59	1	1	1	1	1	14,5	261	2,6	156	1718	137,95	172	190	12,2	4	61,86352
0	0	56	1	0	1	1	0	1,1	302	4,14	54	7394,8	113,52	88	221	10,6	3	77,60525
-1	0	70	0	0	0	0	-1	1,4	176	3,48	210	516	96,1	55	151	12	4	79,21068
-1	0	55	1	0	1	1	-1	1,8	244	2,54	64	6121,8	60,63	92	183	10,3	4	82,47321
1	1	38	1	0	1	1	0	3,4	279	3,53	143	671	113,15	72	136	10,9	3	91,39999
0	1	56	1	0	1	0	0	1	322	4,09	52	824	60,45	213	204	9,7	3	99,17691
-1	1	53	1	0	0	0	0	0,3	280	4	52	4651,2	28,38	189	373	11	3	101,3139
-1	0	43	1	0	0	1	0	3,2	562	3,08	79	2276	144,15	88	251	11	2	108,6771
-1	1	59	1	0	0	1	0	3,6	236	3,52	94	591	82,15	95	71	13,6	4	114,3681
-1	0	65	1	0	0	0	0	0,8	231	3,87	173	9009,8	127,71	96	295	11	3	122,603
-1	0	54	1	0	1	1	1	11,4	178	2,8	588	961	280,55	200	283	12,4	4	130,6081
0	0	50	1	0	1	0	-1	0,7	235	3,56	39	1881	93	123	209	11	3	133,9049
-1	1	60	1	0	1	0	0	5,1	374	3,51	140	1919	122,45	135	322	13	4	136,7795
0	1	64	0	0	1	1	0	0,6	252	3,83	41	843	65,1	83	336	11,4	4	143,6815
-1	0	56	1	0	0	1	0	3,4	271	3,63	464	1376	120,9	55	173	11,6	4	144,039
-1	1	56	1	1	1	1	1	17,4	395	2,94	558	6064,8	227,04	191	214	11,7	4	145,5384
-1	0	45	0	0	1	0	0	2,1	456	4	124	5719	221,88	230	70	9,9	2	149,5926

- Pour **k=4, nous** avons trois patients classés comme suit :
  - Trois classés dans la **4**.
  - Un classé dans la classe **3**.

La classe majoritaire est la classe 4, ⟹ *la prédiction de P par KNN est=[4]*

2.3.2. Application de la méthode des Arbres de Décision

2.3.2.1. Classification du nouvel patient P par AD

2.3.2.1.1. Principe de la construction de l'arbre de décision

La construction d'un arbre de décision repose sur une idée simple : diviser de manière récursive et efficace les exemples de l'ensemble d'apprentissage en utilisant des tests basés sur les attributs, jusqu'à ce que l'on obtienne des sous-ensembles d'exemples contenant principalement des exemples appartenant à une même classe. Cette approche suit une méthode de construction appelée Top-Down, qui construit l'arbre de la racine vers les feuilles de manière récursive.

Le processus commence généralement par la sélection d'un attribut, puis le choix d'un certain nombre de critères pour diviser le nœud correspondant. Pour chaque critère choisi, un nouveau nœud est créé pour les données qui satisfont ce critère. L'algorithme se poursuit de manière récursive jusqu'à ce que des nœuds soient créés pour les données de chaque classe.

La figure de l'arbre de décision n'est pas lisible nous avons retranscrit sous forme textuelle.

Tableau 6: La forme textuelle de l'arbre de décision.

```
--- Hepatomegaly <= 0.50
|  |--- SGOT <= 63.55
|  |  |--- SGOT <= 56.38
|  |  |  |--- Prothrombin <= 10.80
|  |  |  |  |--- Alk_Phos <= 373.00
|  |  |  |  |  |--- class: 3.0
|  |  |  |  |--- Alk_Phos >  373.00
|  |  |  |  |  |--- class: 2.0
|  |  |  |--- Prothrombin >  10.80
|  |  |  |  |--- Age <= 50.50
|  |  |  |  |  |--- class: 1.0
|  |  |  |  |--- Age >  50.50
|  |  |  |  |  |--- class: 3.0
|  |  |--- SGOT >  56.38
|  |  |  |--- Platelets <= 266.00
|  |  |  |  |--- class: 1.0
|  |  |  |--- Platelets >  266.00
|  |  |  |  |--- class: 2.0
|  |--- SGOT >  63.55
|  |  |--- Platelets <= 310.00
|  |  |  |--- Age <= 58.50
|  |  |  |  |--- Alk_Phos <= 1813.50
|  |  |  |  |  |--- Platelets <= 174.00
|  |  |  |  |  |  |--- class: 2.0
|  |  |  |  |  |--- Platelets >  174.00
|  |  |  |  |  |  |--- class: 3.0
|  |  |  |  |--- Alk_Phos >  1813.50
|  |  |  |  |  |--- Prothrombin <= 10.20
|  |  |  |  |  |  |--- class: 4.0
|  |  |  |  |  |--- Prothrombin >  10.20
|  |  |  |  |  |  |--- class: 2.0
|  |  |  |--- Age >  58.50
|  |  |  |  |--- Alk_Phos <= 833.00
|  |  |  |  |  |--- class: 4.0
|  |  |  |  |--- Alk_Phos >  833.00
|  |  |  |  |  |--- SGOT <= 128.18
|  |  |  |  |  |  |--- class: 3.0
|  |  |  |  |  |--- SGOT >  128.18
|  |  |  |  |  |  |--- class: 2.0
|  |  |--- Platelets >  310.00
|  |  |  |--- Platelets <= 319.00
|  |  |  |  |--- Status <= 0.50
|  |  |  |  |  |--- class: 1.0
|  |  |  |  |--- Status >  0.50
|  |  |  |  |  |--- class: 2.0
|  |  |  |--- Platelets >  319.00
|  |  |  |  |--- Bilirubin <= 2.65
|  |  |  |  |  |--- Prothrombin <= 10.45
|  |  |  |  |  |  |--- class: 3.0
|  |  |  |  |  |--- Prothrombin >  10.45
|  |  |  |  |  |  |--- class: 2.0
|  |  |  |  |--- Bilirubin >  2.65
|  |  |  |  |  |--- SGOT <= 189.22
|  |  |  |  |  |  |--- class: 3.0
|  |  |  |  |  |--- SGOT >  189.22
|  |  |  |  |  |  |--- class: 1.0
```

```
|--- Hepatomegaly >  0.50
|   |--- Prothrombin <= 10.95
|   |   |--- Platelets <= 132.00
|   |   |   |--- Cholesterol <= 382.75
|   |   |   |   |--- class: 4.0
|   |   |   |--- Cholesterol >  382.75
|   |   |   |   |--- class: 2.0
|   |   |--- Platelets >  132.00
|   |   |   |--- Age <= 60.50
|   |   |   |   |--- Alk_Phos <= 1508.00
|   |   |   |   |   |--- Platelets <= 247.50
|   |   |   |   |   |   |--- class: 3.0
|   |   |   |   |   |--- Platelets >  247.50
|   |   |   |   |   |   |--- class: 3.0
|   |   |   |   |--- Alk_Phos >  1508.00
|   |   |   |   |   |--- Tryglicerides <= 136.50
|   |   |   |   |   |   |--- class: 4.0
|   |   |   |   |   |--- Tryglicerides >  136.50
|   |   |   |   |   |   |--- class: 3.0
|   |   |   |--- Age >  60.50
|   |   |   |   |--- Albumin <= 3.04
```

```
|   |--- Prothrombin >  10.95
|   |   |--- Cholesterol <= 351.50
|   |   |   |--- Albumin <= 3.92
|   |   |   |   |--- Platelets <= 452.50
|   |   |   |   |   |--- Alk_Phos <= 10049.10
|   |   |   |   |   |   |--- class: 4.0
|   |   |   |   |   |--- Alk_Phos >  10049.10
|   |   |   |   |   |   |--- class: 3.0
|   |   |   |   |--- Platelets >  452.50
|   |   |   |   |   |--- class: 2.0
|   |   |   |--- Albumin >  3.92
|   |   |   |   |--- Age <= 69.00
|   |   |   |   |   |--- class: 2.0
|   |   |   |   |--- Age >  69.00
|   |   |   |   |   |--- class: 4.0
|   |   |--- Cholesterol >  351.50
|   |   |   |--- Age <= 49.50
|   |   |   |   |--- Edema <= -0.50
|   |   |   |   |   |--- Alk_Phos <= 2685.00
|   |   |   |   |   |   |--- class: 4.0
|   |   |   |   |   |--- Alk_Phos >  2685.00
```

```
|   |   |   |   |   |--- Tryglicerides <= 136.50
|   |   |   |   |   |   |--- class: 4.0
|   |   |   |   |   |--- Tryglicerides >  136.50
|   |   |   |   |   |   |--- class: 3.0
|   |   |   |--- Age >  60.50
|   |   |   |   |--- Albumin <= 3.04
|   |   |   |   |   |--- Albumin <= 2.88
|   |   |   |   |   |   |--- class: 4.0
|   |   |   |   |   |--- Albumin >  2.88
|   |   |   |   |   |   |--- class: 3.0
|   |   |   |   |--- Albumin >  3.04
|   |   |   |   |   |--- Prothrombin <= 9.85
|   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |   |--- Prothrombin >  9.85
|   |   |   |   |   |   |--- class: 4.0
```

```
|   |   |--- Cholesterol >  351.50
|   |   |   |--- Age <= 49.50
|   |   |   |   |--- Edema <= -0.50
|   |   |   |   |   |--- Alk_Phos <= 2685.00
|   |   |   |   |   |   |--- class: 4.0
|   |   |   |   |   |--- Alk_Phos >  2685.00
|   |   |   |   |   |   |--- class: 3.0
|   |   |   |   |--- Edema >  -0.50
|   |   |   |   |   |--- class: 3.0
|   |   |   |--- Age >  49.50
|   |   |   |   |--- Bilirubin <= 3.60
|   |   |   |   |   |--- Age <= 57.50
|   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |   |--- Age >  57.50
|   |   |   |   |   |   |--- class: 3.0
|   |   |   |   |--- Bilirubin >  3.60
|   |   |   |   |   |--- Bilirubin <= 25.00
|   |   |   |   |   |   |--- class: 4.0
|   |   |   |   |   |--- Bilirubin >  25.00
|   |   |   |   |   |   |--- class: 2.0
```

2.3.2.2. Prédiction par Arbre de décision du nouvel patient P

Le processus de décision est équivalent à une « descente » dans l'arbre (de la racine vers une des feuilles) : à chaque étape un attribut est testé et un sous-arbre est choisi, la parcours s'arrête dans une feuille (une décision est prise).

Soit le nouvel patient P que l'on souhaite classifier.

Tableau 7: Les données d'un nouvel patient.

Status	Drug	Age	Sex	Ascites	Hepatomegaly	Spiders	Edema	Bilirubin	Cholesterol	Albumin	Copper	Alk_Phos	SGOT	Trygliceride	Platelets	Prothrombin	Stage
-1	1	66	1	0	1	0	0	0,8	248	3,98	50	944	93	63	251	11	?

Tableau 8: La classification du nouvel patient par AD.

```
|--- Hepatomegaly >  0.50
|   |--- Prothrombin <= 10.95
|   |   |--- Platelets <= 132.00
|   |   |   |--- Cholesterol <= 382.75
|   |   |   |   |--- class: 4.0
|   |   |   |--- Cholesterol >  382.75
|   |   |   |   |--- class: 2.0
|   |   |--- Platelets >  132.00
|   |   |   |--- Age <= 60.50
|   |   |   |   |--- Alk_Phos <= 1508.00
|   |   |   |   |   |--- Platelets <= 247.50
|   |   |   |   |   |   |--- class: 3.0
|   |   |   |   |   |--- Platelets >  247.50
|   |   |   |   |   |   |--- class: 3.0
|   |   |   |   |--- Alk_Phos >  1508.00
|   |   |   |   |   |--- Trygliceride <= 136.50
|   |   |   |   |   |   |--- class: 4.0
|   |   |   |   |   |--- Trygliceride >  136.50
|   |   |   |   |   |   |--- class: 3.0
|   |   |   |--- Age >  60.50
|   |   |   |   |--- Albumin <= 3.04
|   |   |   |   |   |--- Albumin <= 2.88
|   |   |   |   |   |   |--- class: 4.0
|   |   |   |   |   |--- Albumin >  2.88
|   |   |   |   |   |   |--- class: 3.0
|   |   |   |   |--- Albumin >  3.04
|   |   |   |   |   |--- Prothrombin <= 9.85
|   |   |   |   |   |   |--- Prothrombin >  10.95
|   |   |   |   |   |   |   |--- Cholesterol <= 351.50
|   |   |   |   |   |   |   |   |--- Albumin <= 3.92
|   |   |   |   |   |   |   |   |   |--- Platelets <= 452.50
|   |   |   |   |   |   |   |   |   |   |--- Alk_Phos <= 10049.10
|   |   |   |   |   |   |   |   |   |   |   |--- class: 4.0
|   |   |   |   |   |   |   |   |   |   |--- Alk_Phos >  10049.10
|   |   |   |   |   |   |   |   |   |   |   |--- class: 3.0
|   |   |   |   |   |   |   |   |   |--- Platelets >  452.50
|   |   |   |   |   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |   |   |   |   |--- Albumin >  3.92
|   |   |   |   |   |   |   |   |   |--- Age <= 69.00
|   |   |   |   |   |   |   |   |   |   |--- class: 2.0

|   |   |   |   |--- Age >  69.00
|   |   |   |   |   |   |   |   |   |   |--- class: 4.0
|   |   |   |   |   |   |   |--- Cholesterol >  351.50
|   |   |   |   |   |   |   |   |--- Age <= 49.50
|   |   |   |   |   |   |   |   |   |--- Edema <= -0.50
|   |   |   |   |   |   |   |   |   |   |--- Alk_Phos <= 2685.00
|   |   |   |   |   |   |   |   |   |   |   |--- class: 4.0
|   |   |   |   |   |   |   |   |   |   |--- Alk_Phos >  2685.00
|   |   |   |   |   |   |   |   |   |   |   |--- class: 3.0
|   |   |   |   |   |   |   |   |   |--- Edema >  -0.50
|   |   |   |   |   |   |   |   |   |   |--- class: 3.0
|   |   |   |   |   |   |   |   |--- Age >  49.50
|   |   |   |   |   |   |   |   |   |--- Bilirubin <= 3.60
```

| | | | | | | |--- class: 2.0
| | | | | | |--- Prothrombin > 9.85
| | | | | | | |--- class: 4.0 | | | | | | |--- Age <= 57.50
| | | | | | | |--- class: 2.0
| | | | | | |--- Age > 57.50
| | | | | | | |--- class: 3.0
| | | | | |--- Bilirubin > 3.60
| | | | | |--- Bilirubin <= 25.00
| | | | | | |--- class: 4.0
| | | | | |--- Bilirubin > 25.00

| | | | | | |--- class: 2.0 |

Le patient P est dans la classe 2

2.4. Calcul des valeurs pour la Migration entre classes

Une fois la maladie du foie détectée, cela déclenche généralement d'autres préoccupations chez les individus, comme le désir de prendre le contrôle de la maladie pour améliorer leur état de santé ou éviter toute détérioration possible. L'objectif principal demeure le ralentissement de l'évolution vers une forme invalidante de la maladie du foie. Ainsi, l'idée est d'intervenir le plus tôt possible et de manière soutenue afin de prévenir une aggravation et de réduire la gravité de la maladie.

Ceci se traduit simplement par des ajustements de certaines valeurs de paramètres spécifiques. La maladie du foie peut être classée en cinq catégories, ce qui nous pousse à réfléchir à la possibilité d'une transition d'une classe à une autre.

Pour un patient donné P, en fonction de ses informations (valeurs des paramètres restants) notre système permet la prédiction du stade de sa maladie par exemple ***classe_i***.

P(**Status**=Status_valeur1, **Drug**=Drug_valeur1, **Age**=Age_valeur1, **Sex**=Sex_valeur1, **Ascites** = Ascites_valeur1, **Hepatomegaly** =Hepatomegaly_valeur1, **Spiders** =Spiders_valeur1, **Edema** =Edema_valeur1, **Bilirubin** =Bilirubin_valeur1, **Cholesterol** = Cholesterol Cholesterol_valeur1, **Albumin** =Albumin_valeur1, **Copper** =Copper_valeur1, Alk_**Phos**

=Alk_Phos_valeur1, **SGOT** =SGOT_valeur1, **Tryglicerides** = Tryglicerides_valeur1, **Platelets** =Platelets_valeur1, **Prothrombin** = Prothrombin_valeur1)

Ceci peut être schématisé par le schéma explicatif suivant :

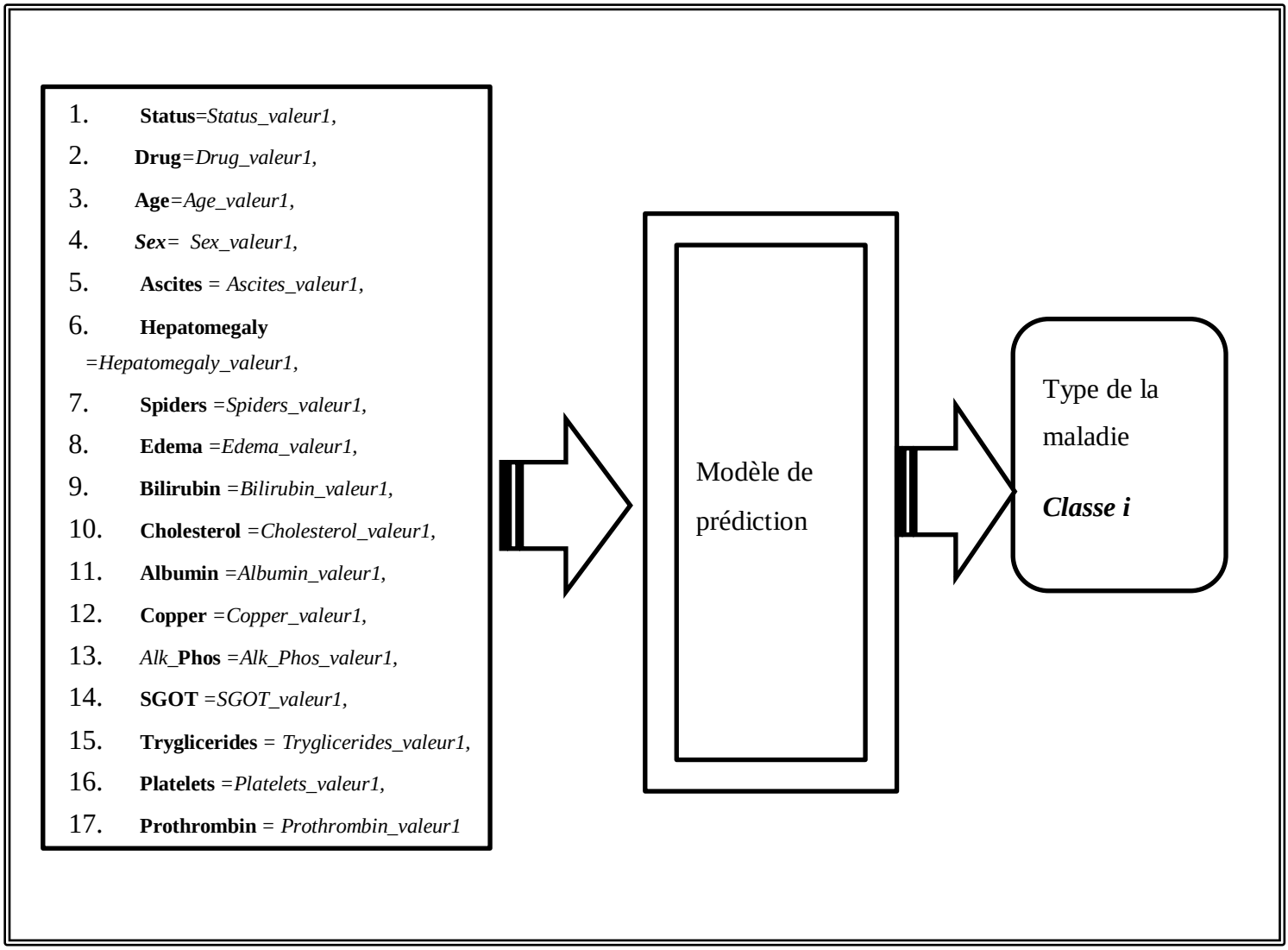

Figure 6: Schéma explicatif de la prédiction.

La migration entre la classe revient à trouver de nouvelles valeurs de certains paramètres ou bien de tous les paramètres. Ces nouvelles valeurs une fois réinjectées dans le système de prédiction permettent de prédire la classe souhaitée par le patient ***classe-j***.

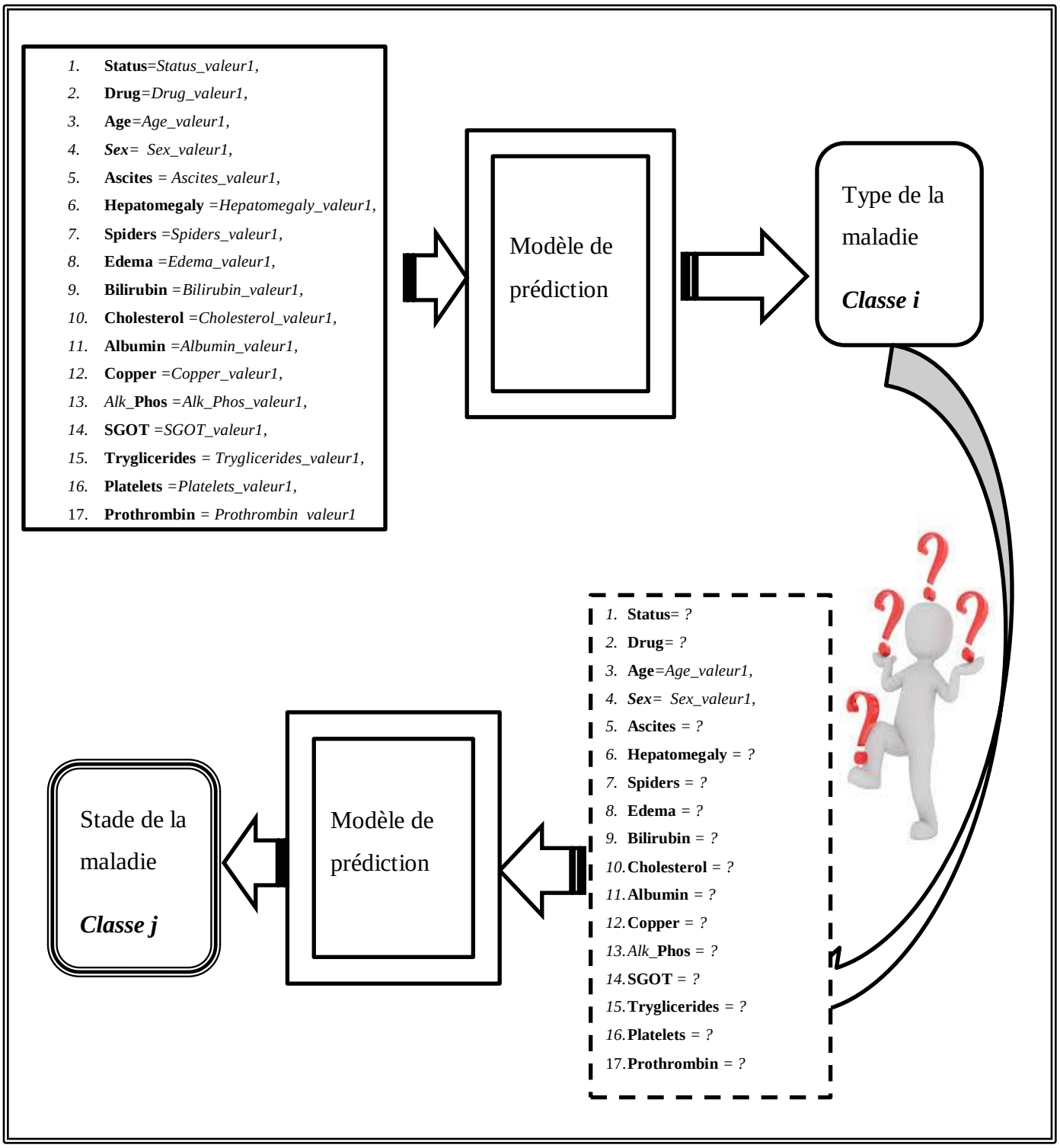

Figure 7: Schéma explicatif de la migration entre classe

2.4.1. Migration entre classes en utilisant le KNN

Pour faciliter la transition entre les classes, nous avons développé un algorithme basé sur la technique KPP (K Plus Proches Voisins) qui permet de calculer de nouveaux paramètres. Cet algorithme utilise une recherche parmi les plus proches voisins pour trouver ceux qui satisfont le passage souhaité vers une classe spécifique, en prenant en compte les paramètres que le patient souhaite maintenir ou accepter de modifier (Ghebouli farida, 2020).

Le processus consiste à identifier, parmi les plus proches voisins, ceux qui répondent à deux types de contraintes : la classe souhaitée (contrainte_ stricte) et la liste des paramètres inchangés (contraintes de valeurs). Dans ce cas, seuls les voisins ayant les mêmes valeurs que le patient pour ces paramètres spécifiques sont sélectionnés.

Ainsi, cet algorithme permet de déterminer les nouveaux paramètres en prenant en considération les contraintes de classe et les valeurs des paramètres inchangés, en se basant sur les plus proches voisins qui satisfont ces critères.

Partie 1

- ***Entrée*** : paramètres du patient P
- ***Prédiction*** de la classe du patient P
- ***Sortie*** :
  - ***Classe-i =*** La classe du patient P
  - ***Voisins*** : Tous les voisins utilisés pour la prédiction du patient P ***ordonné*** selon la distance.

Partie 2

- ***Entrée*** : ***Voisins***
- Prédiction de la classe de tous les voisins
- ***Sortie*** : ***Voisins*** : Tous les voisins utilisés pour la prédiction du patient P ***ordonné*** selon la distance en plus de la colonne *Classe_ prédite* par le système pour chaque voisin.

Partie 3

- ***Entrée :***
  - ***Classe-j=*** la classe souhaitée
  - ***Voisins***
  - ***Liste des paramètres inchangés L***
- ***Etape1 :***
  - *Sélection des voisins dont la classe prédite par KNN est la classe_j*
  - ***Voisins*** *= select * from* ***Voisins*** *where classe_predite=classe_j*
- ***Etape 2:***
  - *Pour chaque paramètre* ***li*** *dans la liste* ***L*** *faire une sélection dans les* ***Voisins*** de ceux ayant la même valeur pour le paramètre ***li*** du patient ***P***
  - ***Voisins*** *= select * from* ***Voisins*** *where li=****li_val_patient_P***
- ***Sortie :***
  - ***Les voisins satisfaisant les contraintes***
  - ***Les valeurs du premier voisins satisfaisant l'enseble des contraintes***

Figure 8: Méta-algorithme de la migration entre classe pour KNN.

2.4.1.1. Exemple de Migration par KNN

Soit le nouvel patient P ayant les caractéristiques présentées dans le tableau suivant :

Tableau 9: Les données d'un patient qui souhaite faire des migrations.

Status	Drug	Age	Sex	Ascites	epatomega	Spiders	Edema	Bilirubin	Cholestero	Albumin	Copper	Alk_Phos	SGOT	rygliceride	Platelets	rothrombi	Classe prédite
0	1	45	1	0	0	0	0	0,7	298	4,1	40	661	106,95	66	324	11,3	2

- **<u>Migration vers la classe1</u>**
  - **Attribut inchangés** (Age, Sex, Edema).
    - **Résultat :** Nous n'avons pas trouvé de résultats.
- **<u>Migration vers la classe3</u>**
  - **Attribut inchangés** (Age, sex, Edema).
    - **Résultat :** Nous avons trouvé trois combinaisons.

Tableau 10: Résultat 1 de migration vers la classe3 (KNN).

Status	Drug	Age	Sex	Ascites	epatomega	Spiders	Edema	Bilirubin	Cholestero	Albumin	Copper	Alk_Phos	SGOT	rygliceride	Platelets	rothrombi	Classe_Predite
0	0	45	1	0	0	0	0	0,6	266	3,97	25	1164	102,3	102	201	10,1	3
-1	0	45	1	0	0	1	0	3,9	350	3,22	121	1268	272,8	231	270	9,6	3
0	1	45	1	0	0	0	0	0,9	400	3,6	31	1689	164,3	166	327	10,4	3

- **Attribut inchangés** (Age, sex, Status, Drug).
  - **Résultat :** Nous avons trouvé une combinaison.

Tableau 11: Résultat 2 de migration vers la classe3 (KNN).

Status	Drug	Age	Sex	Ascites	epatomega	Spiders	Edema	Bilirubin	Cholestero	Albumin	Copper	Alk_Phos	SGOT	rygliceride	Platelets	rothrombi	Classe_Predite
0	1	45	1	0	0	0	0	0,9	400	3,6	31	1689	164,3	166	327	10,4	3

- **<u>Migration vers la classe4</u>**
  - **Attribut inchangés** (Age, sex, Edema).
    - **Résultat :** Nous avons trouvé 3 combinaisons.

Tableau 12: Résultat 1 de migration vers la classe4 (KNN).

Status	Drug	Age	Sex	Ascites	epatomega	Spiders	Edema	Bilirubin	Cholestero	Albumin	Copper	Alk_Phos	SGOT	rygliceride	Platelets	rothrombi	Classe_Predite
0	0	45	1	0	1	1	0	1,4	248	3,58	63	554	75,95	106	79	10,3	4
0	1	45	1	0	0	0	0	1	393	3,57	50	1307	74	103	295	10,5	4
0	1	45	1	0	0	0	0	3,6	374	3,5	143	1428	188	44	151	10,1	4

- **Attribut inchangés** (Age, sex, Status, Drug).
  - **Résultat** : Nous avons trouvé deux combinaisons.

Tableau 13: Résultat 2 de migration vers la classe4 (KNN).

Status	Drug	Age	Sex	Ascites	epatomega	Spiders	Edema	Bilirubin	Cholestero	Albumin	Copper	Alk_Phos	SGOT	rygliceride	Platelets	rothrombi	Classe_Predite
0	1	45	1	0	0	0	0	1	393	3,57	50	1307	74	103	295	10,5	4
0	1	45	1	0	0	0	0	3,6	374	3,5	143	1428	188	44	151	10,1	4

2.4.2 . Migration entre classes en utilisant l'AD

Nous avons proposé une approche pour la technique des arbres de décision, qui consiste à sauvegarder l'arbre de décision complet (tous les chemins) dans une table, puis à effectuer des requêtes afin de satisfaire la migration vers la classe souhaitée, tout en respectant les contraintes spécifiées par le patient, c'est-à-dire maintenir les valeurs initiales ou accepter des modifications.

Le processus implique donc la recherche des chemins qui satisfont les contraintes imposées par le patient. Il est important de noter qu'un chemin ne représente pas une seule combinaison de valeurs, mais plutôt une sous-population entière.

Ainsi, en utilisant cette approche, nous pouvons effectuer des requêtes dans la table de l'arbre de décision afin de trouver les chemins qui répondent aux contraintes spécifiques établies par le patient. Cela permet de prendre en compte les différentes combinaisons de valeurs associées à chaque chemin et de déterminer les options qui satisferont les préférences du patient tout en respectant les contraintes de migration. Nous avons sauvegardé l'arbre de décision dans le tableau suivant :

Tableau 14: Table des chemins de l'arbre de décision (part 1).

comp_Albumin	Albumin	Albumin_max	comp_Age	Age	Age_max	comp_Bilirubin	Bilirubin	Bilirubin_max	comp_Edema	Edema	comp_Alk_Phos	Alk_Phos	comp_Tryglicerides	Tryglicerides	Classe_Predite
QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	InferieurOuEgale	373	QuelqueSoit	-1	3
QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	Superieur	373	QuelqueSoit	-1	2
QuelqueSoit	-1	-1	InferieurOuEgale	50.5	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	QuelqueSoit	-1	QuelqueSoit	-1	1
QuelqueSoit	-1	-1	Superieur	50.5	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	QuelqueSoit	-1	QuelqueSoit	-1	3
QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	QuelqueSoit	-1	QuelqueSoit	-1	1
QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	QuelqueSoit	-1	QuelqueSoit	-1	2
QuelqueSoit	-1	-1	InferieurOuEgale	58.5	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	InferieurOuEgale	1813.5	QuelqueSoit	-1	2
QuelqueSoit	-1	-1	InferieurOuEgale	58.5	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	InferieurOuEgale	1813.5	QuelqueSoit	-1	3
QuelqueSoit	-1	-1	InferieurOuEgale	58.5	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	Superieur	1813.5	QuelqueSoit	-1	4
QuelqueSoit	-1	-1	InferieurOuEgale	58.5	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	Superieur	1813.5	QuelqueSoit	-1	2
QuelqueSoit	-1	-1	Superieur	58.5	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	InferieurOuEgale	833	QuelqueSoit	-1	4
QuelqueSoit	-1	-1	Superieur	58.5	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	Superieur	833	QuelqueSoit	-1	3
QuelqueSoit	-1	-1	Superieur	58.5	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	Superieur	833	QuelqueSoit	-1	2
QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	QuelqueSoit	-1	QuelqueSoit	-1	1
QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	QuelqueSoit	-1	QuelqueSoit	-1	2
QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	InferieurOuEgale	2.65	-1	QuelqueSoit	-1	QuelqueSoit	-1	QuelqueSoit	-1	3
QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	InferieurOuEgale	2.65	-1	QuelqueSoit	-1	QuelqueSoit	-1	QuelqueSoit	-1	2
QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	Superieur	2.65	-1	QuelqueSoit	-1	QuelqueSoit	-1	QuelqueSoit	-1	3
QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	Superieur	2.65	-1	QuelqueSoit	-1	QuelqueSoit	-1	QuelqueSoit	-1	1
QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	QuelqueSoit	-1	QuelqueSoit	-1	4
QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	QuelqueSoit	-1	QuelqueSoit	-1	2
QuelqueSoit	-1	-1	InferieurOuEgale	60.5	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	InferieurOuEgale	1508	QuelqueSoit	-1	3
QuelqueSoit	-1	-1	InferieurOuEgale	60.5	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	InferieurOuEgale	1508	QuelqueSoit	-1	3
QuelqueSoit	-1	-1	InferieurOuEgale	60.5	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	Superieur	1508	InferieurOuEgale	136.50	4
QuelqueSoit	-1	-1	InferieurOuEgale	60.5	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	Superieur	1508	Superieur	136.50	3
InferieurOuEgale	2.88	-1	Superieur	60.5	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	QuelqueSoit	-1	QuelqueSoit	-1	4
Entre	2.88	3.04	Superieur	60.5	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	QuelqueSoit	-1	QuelqueSoit	-1	3
Superieur	3.04	-1	Superieur	60.5	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	QuelqueSoit	-1	QuelqueSoit	-1	2

Tableau 15: Table des chemins de l'arbre de décision (part 2).

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
comp_Hepatomegaly	Hepatomegaly	comp_SGOT	SGOT	SGOT_max	comp_Prothrombin	Prothrombin	Prothrombin_max	comp_Platelets	Platelets	Platelets_max	comp_Cholesterol	Cholesterol	comp_Albumin	Albumin
InferieurOuEgale	0.5	InferieurOuEgale	56.38	-1	InferieurOuEgale	10.80	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	InferieurOuEgale	56.38	-1	InferieurOuEgale	10.80	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	InferieurOuEgale	56.38	-1	Superieur	10.80	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	InferieurOuEgale	56.38	-1	Superieur	10.80	-1	QuelqueSoit	-1	-1	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	Entre	56.38	63.55	QuelqueSoit	-1	-1	InferieurOuEgale	266	-1	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	Entre	56.38	63.55	QuelqueSoit	-1	-1	Superieur	266	-1	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	Superieur	63.55	-1	QuelqueSoit	-1	-1	InferieurOuEgale	174	-1	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	Superieur	63.55	-1	QuelqueSoit	-1	-1	Entre	174	310	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	Superieur	63.55	-1	InferieurOuEgale	10.20	-1	InferieurOuEgale	310	-1	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	Superieur	63.55	-1	Superieur	10.20	-1	InferieurOuEgale	310	-1	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	Superieur	63.55	-1	QuelqueSoit	-1	-1	InferieurOuEgale	310	-1	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	Entre	63.55	128.18	QuelqueSoit	-1	-1	InferieurOuEgale	310	-1	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	Superieur	128.18	-1	QuelqueSoit	-1	-1	InferieurOuEgale	310	-1	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	Superieur	63.55	-1	QuelqueSoit	-1	-1	Entre	310	319	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	Superieur	63.55	-1	QuelqueSoit	-1	-1	Entre	310	319	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	Superieur	63.55	-1	InferieurOuEgale	10.45	-1	Superieur	319	-1	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	Superieur	63.55	-1	Superieur	10.45	-1	Superieur	319	-1	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	Entre	63.55	189.22	QuelqueSoit	-1	-1	Superieur	319	-1	QuelqueSoit	-1	QuelqueSoit	-1
InferieurOuEgale	0.5	Superieur	189.22	-1	QuelqueSoit	-1	-1	Superieur	319	-1	QuelqueSoit	-1	QuelqueSoit	-1
Superieur	0.5	QuelqueSoit	-1	-1	InferieurOuEgale	10.95	-1	Superieur	319	-1	InferieurOuEgale	382.75	QuelqueSoit	-1
Superieur	0.5	QuelqueSoit	-1	-1	InferieurOuEgale	10.95	-1	Superieur	319	-1	Superieur	382.75	QuelqueSoit	-1
Superieur	0.5	QuelqueSoit	-1	-1	InferieurOuEgale	10.95	-1	Entre	132	247.5	QuelqueSoit	-1	QuelqueSoit	-1
Superieur	0.5	QuelqueSoit	-1	-1	InferieurOuEgale	10.95	-1	Superieur	247.5	-1	QuelqueSoit	-1	QuelqueSoit	-1
Superieur	0.5	QuelqueSoit	-1	-1	InferieurOuEgale	10.95	-1	Superieur	132	-1	QuelqueSoit	-1	QuelqueSoit	-1
Superieur	0.5	QuelqueSoit	-1	-1	InferieurOuEgale	10.95	-1	Superieur	132	-1	QuelqueSoit	-1	QuelqueSoit	-1
Superieur	0.5	QuelqueSoit	-1	-1	InferieurOuEgale	10.95	-1	Superieur	132	-1	QuelqueSoit	-1	InferieurOuEgale	2.88
Superieur	0.5	QuelqueSoit	-1	-1	InferieurOuEgale	10.95	-1	Superieur	132	-1	QuelqueSoit	-1	Entre	2.88

2.4.2.1. Explication d'une ligne du tableau

Nous avons défini pour chaque attribut trois colonnes, les deux premières colonnes pour sauvegarder deux opérateurs de comparaison et l'autre colonne pour sauvegarder la valeur de cet attribut.

Nous avons utilisé trois valeurs possibles dans les colonnes **comp1 et comp2** :

- Ind : dans le cas où l'attribut en question ne figure pas dans le chemin.
- <= : inferieur ou égale à la valeur qui est dans la colonne de l'attribut en question.
- > : supérieur à la valeur qui est dans la colonne de l'attribut en question.

Pour chaque chemin de l'arbre nous avons une ligne dans notre tableau :

Partie 1

- ***Entrée*** : paramètres du patient P
- ***Prédiction*** de la classe du patient P
- ***Sortie*** :
  - ***Classe-i =*** La classe du patient P

Partie 2

- ***Entrée :***
  - ***Classe-j=*** la classe souhaitée
  - ***Table des chemins de l'AD***
  - ***Liste des paramètres inchangés L***
- ***Etape1 :***
  - *Sélection des chemins dont la classe prédite par AD est la classe_j*
  - ***Chemins*** *= select * from* ***Table-des-chemins-AD*** *where classe_predite=classe_j*
- ***Etape 2:***
  - *Pour chaque paramètre* ***li*** *dans la liste* ***L*** *faire une sélection dans les* ***chemins*** de ceux ayant la même valeur pour le paramètre ***li*** du patient ***P***
  - ***Chemins*** *= select * from* ***Table-des-chemins-AD*** *where li=****li_val_patient_P***
- ***Sortie :***
  - ***Les chemins satisfaisant les contraintes***

Figure 9: Méta-algorithme de la migration entre classe pour AD.

2.4.2.2. Exemple de Migration par AD

Soit le nouvel patient P ayant les caractéristiques présentées dans le tableau suivant :

Tableau 16: Les données d'un patient qui souhaite faire des migrations.

Status	Drug	Age	Sex	Ascites	epatomega	Spiders	Edema	Bilirubin	Cholestero	Albumin	Copper	Alk_Phos	SGOT	ryglicerid	Platelets	Prothrombin	Classe prédite
1	0	59	1	1	1	1	1	14,5	261	2,6	156	1718	137,95	172	190	12,2	4

La classe prédite est : 4

- **Migration vers la classe1**
  - **Attribut inchangés** (Prothrombin, Platelets, Age, Albumin).
    - **Résultat** : Nous avons trouvé un seul chemin.

Tableau 17: Résultat 1 de migration vers la classe1 (AD).

_Hepatom	patomega	omp_SGO	SGOT	SGOT_max	p_Prothro	rothrombi	hrombin_	mp_Platel	Platelets	atelets_ma	p_Cholest	Cholestero	mp_Albun	Albumin	bumin_ma	comp_Age	Age	Age_max	mp_Biliru	Bilirubin	lir
InferieurO	0,5	Entre	56,38	63,55	QuelqueSc	-1	-1	InferieurO	266	-1	QuelqueSc	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	QuelqueSc	-1	

mp_Platel	Platelets	atelets_ma	p_Cholest	Cholestero	mp_Albun	Albumin	bumin_ma	comp_Age	Age	Age_max	mp_Biliru	Bilirubin	lirubin_ma	mp_Edem	Edema	mp_Alk_Pl	Alk_Phos	p_Tryglice	rygliceride	asse_Predite
InferieurO	266	-1	QuelqueSc	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	QuelqueSc	-1	QuelqueSc	-1	QuelqueSc	-1	1

- **Attribut inchangés** (Age, Prothrombin).
  - **Résultat** : Nous avons trouvé 3 chemins.

Tableau 18: Résultat 2 migration vers la classe1 (AD)

_Hepatom	patomega	omp_SGO	SGOT	SGOT_max	p_Prothro	rothrombi	hrombin_	mp_Platel	Platelets	atelets_ma	p_Cholest	Cholestero	mp_Albun	Albumin	bumin_ma	comp_Age	Age	Age_max	mp_Biliru	Bilirubin	l
InferieurO	0,5	Entre	56,38	63,55	QuelqueSc	-1	-1	InferieurO	266	-1	QuelqueSc	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	QuelqueSc	-1	
InferieurO	0,5	Superieur	63,55	-1	QuelqueSc	-1	-1	Entre	310	319	QuelqueSc	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	QuelqueSc	-1	
InferieurO	0,5	Superieur	189,22	-1	QuelqueSc	-1	-1	Superieur	319	-1	QuelqueSc	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	Superieur	2,65	

Platelets	atelets_ma	p_Cholest	Cholestero	mp_Albun	Albumin	bumin_ma	comp_Age	Age	Age_max	mp_Biliru	Bilirubin	lirubin_ma	mp_Edem	Edema	mp_Alk_Pl	Alk_Phos	p_Tryglice	rygliceride	Classe_Predite
266	-1	QuelqueSc	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	QuelqueSc	-1	QuelqueSc	-1	QuelqueSc	-1	1
310	319	QuelqueSc	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	QuelqueSc	-1	QuelqueSc	-1	QuelqueSc	-1	1
319	-1	QuelqueSc	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	Superieur	2,65	-1	QuelqueSc	-1	QuelqueSc	-1	QuelqueSc	-1	1

- **Migration vers la classe2**
  - **Attribut inchangés** (Age, Prothrombin).
    - **Résultat :** Nous avons trouvé 7 chemins.

Tableau 19: Résultat 1 migration vers la classe2 (AD).

_Hepatom	patomega	omp_SGO	SGOT	SGOT_max	p_Prothro	rothrombi	hrombin_	mp_Platel	Platelets	atelets_ma	p_Cholest	Cholestero	mp_Albun	Albumin	bumin_ma	comp_Age	Age	Age_max	mp_Biliru	Bilirubin	l
InferieurO	0,5	Entre	56,38	63,55	QuelqueSc	-1	-1	Superieur	266	-1	QuelqueSc	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	QuelqueSc	-1	
InferieurO	0,5	Superieur	128,18	-1	QuelqueSc	-1	-1	InferieurO	310	-1	QuelqueSc	-1	QuelqueSc	-1	-1	Superieur	58,5	-1	QuelqueSc	-1	
InferieurO	0,5	Superieur	63,55	-1	QuelqueSc	-1	-1	Entre	310	319	QuelqueSc	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	QuelqueSc	-1	
InferieurO	0,5	Superieur	63,55	-1	Superieur	10,45	-1	Superieur	319	-1	QuelqueSc	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	InferieurO	2,65	
Superieur	0,5	QuelqueSc	-1	-1	Superieur	10,95	-1	Superieur	452	-1	InferieurO	351,5	InferieurO	3,92	-1	QuelqueSc	-1	-1	QuelqueSc	-1	
Superieur	0,5	QuelqueSc	-1	-1	Superieur	10,95	-1	QuelqueSc	-1	-1	InferieurO	351,5	Superieur	3,92	-1	InferieurO	69	-1	QuelqueSc	-1	
Superieur	0,5	QuelqueSc	-1	-1	Superieur	10,95	-1	QuelqueSc	-1	-1	Superieur	351,5	QuelqueSc	-1	-1	Superieur	49,5	-1	Superieur	25	

Platelets	atelets_map_Cholest	Cholestero	mp_Albun	Albumin	bumin_ma	comp_Age	Age	Age_max	mp_Bilirut	Bilirubin	lirubin_m	mp_Edem	Edema	mp_Alk_Pl	Alk_Phos	p_Tryglice	rygliceride	asse_Predite
266	-1 QuelqueSc	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	QuelqueSc	-1	QuelqueSc	-1	QuelqueSc	-1	2
310	-1 QuelqueSc	-1	QuelqueSc	-1	-1	Superieur	58,5	-1	QuelqueSc	-1	-1	QuelqueSc	-1	Superieur	833	QuelqueSc	-1	2
310	319 QuelqueSc	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	QuelqueSc	-1	QuelqueSc	-1	QuelqueSc	-1	2
319	-1 QuelqueSc	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	InferieurO	2,65	-1	QuelqueSc	-1	QuelqueSc	-1	QuelqueSc	-1	2
452	-1 InferieurO	351,5	InferieurO	3,92	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	QuelqueSc	-1	QuelqueSc	-1	QuelqueSc	-1	2
-1	-1 InferieurO	351,5	Superieur	3,92	-1	InferieurO	69	-1	QuelqueSc	-1	-1	QuelqueSc	-1	QuelqueSc	-1	QuelqueSc	-1	2
-1	-1 Superieur	351,5	QuelqueSc	-1	-1	Superieur	49,5	-1	Superieur	25	-1	QuelqueSc	-1	QuelqueSc	-1	QuelqueSc	-1	2

- **<u>Migration vers la classe3</u>**
  - **Attribut inchangés** (Age, Prothrombin).
    - **Résultat :** Nous avons trouvé 5 chemins.

Tableau 20:Résultat 1 migration vers la classe3 (AD).

1	_Hepatom	patomeg	omp_SGO	SGOT	SGOT_ma	p_Prothro	rothrombi	hrombin_	mp_Platel	Platelets	atelets_ma	p_Cholest	Cholestero	mp_Albun	Albumin	bumin_ma	comp_Age	Age	Age_max	mp_Bilirut	Bilirubin
2	InferieurO	0,5	InferieurO	56,38	-1	Superieur	10,8	-1	QuelqueSc	-1	-1	QuelqueSc	-1	QuelqueSc	-1	-1	Superieur	50,5	-1	QuelqueSc	-1
3	InferieurO	0,5	Entre	63,55	128,18	QuelqueSc	-1	-1	InferieurO	310	-1	QuelqueSc	-1	QuelqueSc	-1	-1	Superieur	58,5	-1	QuelqueSc	-1
4	InferieurO	0,5	Entre	63,55	189,22	QuelqueSc	-1	-1	Superieur	319	-1	QuelqueSc	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	Superieur	2,65
5	Superieur	0,5	QuelqueSc	-1	-1	Superieur	10,95	-1	InferieurO	452	-1	InferieurO	351,5	InferieurO	3,92	-1	QuelqueSc	-1	-1	QuelqueSc	-1
6	Superieur	0,5	QuelqueSc	-1	-1	Superieur	10,95	-1	QuelqueSc	-1	-1	Superieur	351,5	QuelqueSc	-1	-1	Superieur	57,5	-1	InferieurO	3,6

1	mp_Platel	Platelets	atelets_ma	p_Cholest	Cholestero	mp_Albun	Albumin	bumin_ma	comp_Age	Age	Age_max	mp_Bilirut	Bilirubin	lirubin_m	mp_Edem	Edema	mp_Alk_Pl	Alk_Phos	p_Tryglice	rygliceride	asse_Predite
2	QuelqueSc	-1	-1	QuelqueSc	-1	QuelqueSc	-1	-1	Superieur	50,5	-1	QuelqueSc	-1	-1	QuelqueSc	-1	QuelqueSc	-1	QuelqueSc	-1	3
3	InferieurO	310	-1	QuelqueSc	-1	QuelqueSc	-1	-1	Superieur	58,5	-1	QuelqueSc	-1	-1	QuelqueSc	-1	Superieur	833	QuelqueSc	-1	3
4	Superieur	319	-1	QuelqueSc	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	Superieur	2,65	-1	QuelqueSc	-1	QuelqueSc	-1	QuelqueSc	-1	3
5	InferieurO	452	-1	InferieurO	351,5	InferieurO	3,92	-1	QuelqueSc	-1	-1	QuelqueSc	-1	-1	QuelqueSc	-1	Superieur	10049,1	QuelqueSc	-1	3
6	QuelqueSc	-1	-1	Superieur	351,5	QuelqueSc	-1	-1	Superieur	57,5	-1	InferieurO	3,6	-1	QuelqueSc	-1	QuelqueSc	-1	QuelqueSc	-1	3

2.5. Conclusion

Dans notre étude, nous avons examiné deux techniques d'apprentissage supervisé : les K plus proches voisins et les arbres de décision. Notre objectif était de prédire précocement les maladies du foie afin de réduire les risques pour la santé des patients.

L'objectif principal de cette prédiction est d'intervenir le plus tôt possible et de manière continue pour prévenir l'aggravation et la sévérité de la maladie. Il est essentiel que les patients prennent conscience de leur état médical réel afin de comprendre pleinement les enjeux liés à leur traitement actuel et futur. L'objectif ultime est de ralentir la progression vers une forme invalidante de la maladie.

En outre, nous avons pris en considération la migration entre les différentes classes de maladies. Cette migration peut être perçue comme une opportunité d'amélioration de l'état de santé du patient ou, malheureusement, comme une détérioration de son état avec l'apparition de complications. Pour chaque technique, nous avons développé un algorithme permettant de calculer de nouvelles valeurs de paramètres afin de faciliter la transition d'une classe à une autre.

Le chapitre suivant sera consacré à la présentation de la mise en œuvre et de l'implémentation de ces techniques dans notre application.

Chapitre 3 : Implémentation et Bilan

3.1. Introduction

Dans ce chapitre, nous mettons en avant notre application et décrivons l'implémentation des différentes techniques de classification que nous avons utilisées. La structure du chapitre est la suivante :

- Dans la première partie, nous procédons à l'évaluation des résultats obtenus par les algorithmes utilisés et les comparons entre eux. Nous analysons les performances de chaque technique de classification afin de déterminer leur efficacité respective dans la prédiction des maladies de foie.
- Dans la deuxième partie, nous présentons les outils et langages que nous avons employés pour la réalisation de notre projet. Nous mettons en avant les ressources technologiques qui ont été essentielles à la mise en place de notre application et à la manipulation des données médicales nécessaires à notre étude.
- Enfin, nous dévoilons notre application pour la classification et la migration entre les différentes classes de maladies de foie. Notre application porte le nom de "PMF Prédiction et Migration de la maladie de foie". Nous décrivons ses fonctionnalités et ses objectifs, mettant en avant son rôle dans la prédiction précise des maladies de foie et dans la gestion de la transition entre ces différentes classes.

A travers ce chapitre, nous souhaitons fournir une vision claire de notre application et de l'ensemble des méthodes utilisées, mettant ainsi en évidence les résultats obtenus et l'utilité de notre outil dans le domaine de la prédiction et de la gestion des maladies de foie.

3.2. Evaluation des résultats

Dans cette section, nous présenterons les résultats obtenus à l'aide de nos différentes techniques utilisées. Dans le domaine de l'apprentissage automatique, diverses métriques sont utilisées pour évaluer la précision prédictive d'un modèle. L'une de ces mesures est la validation

croisée, qui permet d'évaluer les performances d'un modèle en simulant son utilisation dans le monde réel.

La validation croisée est une méthode utilisée pour évaluer la performance d'un modèle d'apprentissage automatique en utilisant les données disponibles de manière plus efficace. Plutôt que de simplement diviser les données en un seul ensemble d'apprentissage et de test, la validation croisée consiste à diviser les données en plusieurs sous-ensembles appelés "plis" ou "folds".

Le processus de validation croisée se déroule de la manière suivante : le modèle est entraîné sur une partie des données appelée ensemble d'apprentissage, puis évalué sur les données restantes, appelées ensemble de validation. Ce processus est répété plusieurs fois, chaque fois en utilisant un pli différent comme ensemble de validation, tandis que les autres plis sont utilisés comme ensemble d'apprentissage. Les performances du modèle sur chaque pli sont ensuite agrégées pour obtenir une estimation globale de sa performance. Elle permet d'évaluer la capacité du modèle à généraliser à de nouvelles données en simulant différentes situations d'apprentissage et de test. Cela permet d'obtenir une mesure plus robuste de la performance du modèle, car elle est évaluée sur plusieurs jeux de données différents. En utilisant la validation croisée, il est possible de détecter les problèmes de sur apprentissage (overfitting) ou de sous-apprentissage (underfitting) du modèle.

La validation croisée est une technique qui permet d'estimer la performance d'un modèle d'apprentissage automatique en utilisant plusieurs sous-ensembles de données pour l'entraînement et l'évaluation, offrant ainsi une évaluation plus fiable et plus représentative de la capacité du modèle à généraliser à de nouvelles données. Nous avons deux annexes Annexe1 pour le choix du paramètre K pour l’algorithme KPP et Annex2 pour le choix de la profondeur de l’arbre pour la classification par arbre de décision.

Pour mesurer cette performance, il existe des indices ou critères qui permettent de quantifier l'écart entre les prédictions du modèle et les valeurs réelles. Ces critères servent à évaluer la précision, la sensibilité, la spécificité ou d'autres aspects de la performance prédictive du modèle. Dans cette section, nous examinerons ces différents indices et critères afin d'évaluer la

performance de nos techniques d'apprentissage automatique dans la prédiction des maladies de foie.

3.2.1. Critères et mesures d'évaluation

- **Matrice de confusion**

Dans les problématiques de classification, la plupart des indices de performance sont calculés à partir d'une matrice de confusion. Cette matrice affiche le nombre de succès et d'échecs de prédiction pour chaque catégorie de la variable à prédire. La matrice de confusion est une table qui montre chaque classe dans les données d'évaluation, ainsi que le nombre ou le pourcentage de prédictions correctes et incorrectes.

Dans le cas d'une tâche de classification supervisée binaire, où la modalité de la variable à prédire correspond à la classe « positive » et l'autre à la classe « négative », on nomme les coefficients de la matrice de confusion de la manière suivante :

- **VN** : Nombre de vrais négatifs (True Négatif TN)
- **FN** : Nombre de faux négatifs (False Négatif FN)
- **FP** : Nombre de faux positifs (False Positif FP)
- **VP** : Nombre de vrais positifs *(*True Positif TP)

Tableau 21: Matrice de Confusion

		Y prédit par le modèle	
		Y=1	Y=0
Y réel(Y')	Y'=1	Nombre de 1 prédits correctement Vrai Positifs (VP) True Positif (TP)	Nombre de 1 prédits en 0 Faux Négatif (FN) False Négatif (FN)
	Y'=0	Nombre de 0 prédits en 1 Faux Positifs (FP) False Positif (FP)	Nombre de 0 prédits correctement Vrai Négatif (VN) True Négatif (TN)

- **Accuracy (**Exactitude, justesse**)** (la proportion de prédictions correctes) : il s'agit d'une description d'erreurs systématiques, d'une mesure du biais statistique ; faible précision provoque une différence entre un résultat et une valeur "vraie".

$$\textbf{Accuracy} = \frac{TP+TN}{\text{TP+TN+FP+FN}} * \textbf{100}\%$$

.

- **Précision :** Proportion de solutions trouvées qui sont pertinentes. A quel point les prédictions positives sont précises.

$$\textbf{Précision}=\frac{TP}{\text{TP+FP}} * \textbf{100}\%$$

- **Rappel (sensitivity, reccal) :** Proportion des solutions pertinentes qui sont trouvées. Mesure la capacité du système à donner toutes les solutions pertinentes. Couverture des observations vraiment positives.

$$\textbf{Rappel (sensitivity, recall)} = \frac{TP}{\text{TP+FN}} * 100\%$$

- **F-mesure (F-score) :** La F-mesure correspond à un compromis de la précision et du rappel donnant la performance du modèle. Moyenne harmonique de la précision et du rappel. Mesure la capacité du modèle à donner toutes les solutions pertinentes et à refuser les autres.

$$\textbf{F1 score=2.}\frac{\textbf{Rappel} * \textbf{Précision}}{\textbf{Rappel + Précision}} * 100\%$$

Nous pouvons résumer ces indicateurs principaux dans le tableau suivant :

Tableau 22: Les différents critères d'évaluation d'un modèle de classification

Indicateur	Formule	Interprétation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Perforance globale du modéle
Précision	$\frac{TP}{TP+FP}$	À quel point les prédictions positives sont précises
Rappel"Sensibilité"	$\frac{TP}{TP+FN}$	Couverture des observations vraiment positivs
Spécificité	$\frac{TN}{TN+FP}$	Couverture des observations vraiment négatives
F-mesure	$\frac{2TP}{2TP+FP+FN}$	Indicateur hybride utilisé pour les classes non-balancées

Matrice de confusion pour la prédiction de maladie de foie

Notre variable de prédiction n'est pas binaire mais prend ses valeurs dans l'ensemble {1, 2, 3,4}.

Tableau 23: La matrice de confusion de notre modèle.

		Classe prédite par le modèle			
		Y=1	Y=2	Y=3	Y=4
Classe réel	Y'=1	**m_{11}**	m12	m13	m14
	Y'=2	m21	**m_{22}**	m 23	m24
	Y'=3	m31	m32	**m_{33}**	m34
	Y'=4	m41	m42	m43	**m_{44}**

m_{ij} : représente le nombre de patients de **classe i** prédit **classe j** par le modèle.

m_{cc} : Le nombre de patients de la **classe c** prédits correctement par le modèle (classe i) (représente les true positif de la **classe c**).

Accuracy

Correspond à la proportion d'observations bien classées.

$$\boldsymbol{Accuracy = \frac{\sum_i m_{ii}}{\sum_{i,j} m_{ij}}}$$

Taux d'erreur global

Le taux d'erreur global, correspond à la proportion d'observations mal classées, qui dépend du ratio entre la trace de la matrice de confusion (c'est-à-dire la somme des coefficients diagonaux, donc le nombre de bonnes prédictions), et la somme de tous les coefficients (autrement dit le nombre total de prédictions) : $\boldsymbol{E = 1 - \frac{\sum_{i=1}^{4} m_{ii}}{\sum_{i,j} m_{ij}}}$

Précision par rapport à une classe

La précision d'un classifieur par rapport à une certaine classe (autrement dit, par rapport à une certaine modalité de la variable à prédire), se mesure comme la proportion d'individus, parmi tous ceux pour lesquels le classifieur a prédit cette classe, qui appartiennent réellement à celle-ci.

$$\boldsymbol{Précision_{classe\ c}(P_c) = \frac{m_{cc}}{\sum_i m_{ic}}}$$

Rappel par rapport à une classe

Le rappel d'un classifieur par rapport à une certaine classe se mesure, quant à lui, comme la proportion d'individus, parmis tous ceux qui appartiennent réellement à cette classe, pour lesquels le classifieur a prédit cette classe c.

$$\boldsymbol{Rappel_{classe\ c}(R_c) = \frac{m_{cc}}{\sum_i m_{ci}}}$$

F-mesure par rapport à une classe

On peut résumer les mesures de précision de rappel par rapport à une classe c en un seul indicateur, en calculant la moyenne harmonique :

$$\boldsymbol{F_{classe\ c} = \frac{P_c \times R_c}{P_c + R_c}}$$

3.2.2. Le classifier KPP

3.2.2.1. Choix de la valeur de K

Quelques règles sur le choix de k : Le paramètre k doit être déterminé par l'utilisateur : k∈N. En classification binaire, il est utile de choisir k impair pour éviter les votes égalitaires. Le meilleur choix de k dépend du jeu de donnée. En général, les grandes valeurs de k réduisent l'effet du bruit sur la classification et donc le risque de sur-apprentissage, mais rendent les frontières entre classes moins distinctes. Il convient donc de faire un choix de compromis entre la variabilité associée à une faible valeur de k contre un 'oversmoothing' ou surlissage (i.e

gommage des détails) pour une forte valeur de k. Un bon k peut être sélectionné par diverses techniques heuristiques, par exemple, de validation-croisée. Nous choisirons la valeur de k qui minime l'erreur de classification (1mathieu-Dupas, 2010).

Nous avons appliqué l'algorithme KNN avec plusieurs valeurs du paramètre K et nous avons conclu que la valeur de **K=6** est le meilleur choix à faire pour notre base de données Mayo Clinik originale et modifiée. L'Annex1 contient toutes les valeurs de précision et de rappel pour une variation de k entre 1 et 10 en utilisation la technique de validation croisée. La classe 1 étant minoritaires nous avons opté pour le choix ou le taux de prédiction pour cette classe n'est pas nul, car la plupart des cas la prédiction pour cette classe est nul.

3.2.2.2. Matrice de confusion du classifier KPP

Tableau 24: Rapport de classification pour l'algorithme « KPP ».

Paramètres de précision rappel et accuracy du model KNN

	precision	recall	f1-score	support
1.0	0.09	0.25	0.13	4
2.0	0.21	0.21	0.21	19
3.0	0.34	0.38	0.36	32
4.0	0.47	0.31	0.38	29
accuracy			0.31	84

KNN Training Score: 0.6107784431137725

KNN Testing Score: 0.30952380952380953

Tableau 25: Matrice de confusion de « KPP »

		Classe prédite par le modèle			
		Classe 1	Classe 2	Classe 3	Classe 4
Classe réel	classe 1	**1**	1	2	0
	classe 2	2	**3**	10	4
	classe 3	0	7	**20**	5
	classe 4	1	4	10	**14**

Discussion :

Les résultats de la précision du classifieur KPP sont : 9% pour la classe 1 sachant que nous ne disposons que de 17 personnes de classe1 dans base de données, et de 21% pour la classe 2, 34% pour la classe 3 et de 47% pour la classe 4. La précision globale du classification KPP et 30%.

Malgré que ces résultats ne semblent pas satisfaisant pour un classifieur. Le problème se pose au niveau de l'ensemble des données. Cet avis est partagé par les discussions sur le site [REF : https://www.kaggle.com/code/consistentshourov/liver-cirrhosis-prediction-with-4-different-m-l]

3.2.3. Le classifier AD

3.2.3.1 Choix de la bonne taille de l'arbre

En pratique, il n'est pas toujours préférable de construire un arbre de décision où chaque feuille correspond à un sous-ensemble parfaitement homogène en termes de variable cible. Plus le modèle est complexe (avec un arbre plus grand, plus de branches et plus de feuilles), plus il y a un risque que le modèle ne puisse pas être généralisé à de nouvelles données, c'est-à-dire qu'il ne puisse pas refléter la réalité que l'on cherche à comprendre.

Pour notre base nous avons appliqué l'algorithme AD en variant la profondeur de l'arbre et nous avons trouvé que la profondeur 6 est la meilleure. Voir Annex2 pour plus de détails.

3.2.3.2. Matrice de confusion du classifier AD

Tableau 26: Rapport de classification pour l'algorithme arbre de décision.

Paramètres de précision rappel et accuracy du model Arbre

	precision	recall	f1-score	support
1.0	0.25	0.25	0.25	4
2.0	0.20	0.16	0.18	19
3.0	0.48	0.62	0.54	32
4.0	0.61	0.48	0.54	29
accuracy			0.45	84
macro avg	0.38	0.38	0.38	84
weighted avg	0.45	0.45	0.44	84

AD Training Score: 0.7724550898203593

AD Testing Score: 0.4523809523809524

Tableau 27: Matrice de confusion arbre de décision.

		Classe prédite par le modèle			
		Classe 1	Classe 2	Classe 3	Classe 4
Classe réel	classe 1	**1**	1	2	0
	classe 2	3	**4**	11	1
	classe 3	3	8	**12**	9
	classe 4	4	6	10	**9**

Les résultats de la précision du classifier AD sont : 25% pour la classe 1, et de 20% pour la classe 2, 48%pour la classe 3 et de 61% pour la classe 4. La précision globale de la classification AD et 45%.

Le classifier arbre de décision présente des résultats un peu meilleurs que ceux du KPP, malgré que ces résultats ne semblent pas satisfaisant pour un classifier.

3.2.4. La migration entre classes

Il est crucial de noter que l'avis des professionnels de la santé aurait été précieux pour mieux expliquer le concept de migration entre classes dans cette étape. Certains paramètres ne peuvent pas être modifiés, mais dans notre application, nous permettons ces changements. L'application dans un domaine moins sensible, tel que le domaine commercial, permet une meilleure appréciation. La raison de choisir cette base de données est l'idée d'avoir plusieurs classes, et l'absence de bases de données avec plusieurs classes dans les références disponibles a motivé ce choix.

3.2.4.1. La migration entre classes par KPP

Nous avons décidé d'appliquer notre algorithme de migration sur l'ensemble des données d'apprentissage et calculer le pourcentage de passage entre toutes les classes.

Nous avons décidé de ne maintenir que les deux paramètres **age** et **sex** :

Tableau 28: Pourcentage des possibilités de migration en utilisant KPP (1).

			Migration vers la classe			
			Classe 1	Classe 2	Classe 3	Classe 4
Classe rprédite						
	classe 1		-	89%	94%	67%
	classe 2		43%	-	89%	79%
	classe 3		37%	73%	-	72%
	classe 4		30%	64%	80%	-

Nous avons décidé de ne maintenir les cinq paramètres **age, sex, status et drug** :

Tableau 29: Pourcentage des possibilités de migration en utilisant KPP (2).

			Migration vers la classe			
			Classe 1	Classe 2	Classe 3	Classe 4
Classe rprédite						
	classe 1		-	33%	72%	22%
	classe 2		11%	-	59%	29%
	classe 3		10%	30%	-	32%
	classe 4		10%	28%	52%	-

3.2.4.2. La migration entre classes par AD

Nous avons décidé d'appliquer notre algorithme de migration sur l'ensemble des données d'apprentissage et calculer le pourcentage de passage entre toutes les classes.

Nous avons décidé de ne maintenir que les deux paramètres **age** et **sex**.

Tableau 30: Pourcentage des possibilités de migration en utilisant AD (1).

			Migration vers la classe			
			Classe 1	Classe 2	Classe 3	Classe 4
Classe rprédite						
	classe 1		-	100%	100%	100%
	classe 2		100%	-	100%	100%
	classe 3		100%	100%	-	100%
	classe 4		100%	100%	100%	-

Nous avons décidé de ne maintenir les cinq paramètres **age, sex, drug, status, Sgot *:***

Tableau 31: Pourcentage des possibilités de migration en utilisant AD (2).

			Migration vers la classe			
			Classe 1	Classe 2	Classe 3	Classe 4
Classe rprédite						
	classe 1		-	100%	100%	100%
	classe 2		89%	-	98%	98%
	classe 3		100%	100%	-	100%
	classe 4		95%	95%	95%	-

Discussion :

On constate que la technique des arbres de décisions donne des pourcentages plus élevés que la technique du plu proche voisin. Ceci peut être expliqué par le fais que les arbres de décision donnent des prédictions basées sur des intervalles de valeurs alors que la technique du KPP est basée sur l'égalité des paramètres.

Le choix de la méthode à utiliser pour la migration entre classes va dépendre du domaine d'application et à quel point le changement d'un paramètre est critique par rapport à la décision finale.

3.3. Outils et langage utilisés

3.3.1. Les outils

PyCharm : est un environnement de développement intégré utilisé pour programmer en Python. Il permet l'analyse de code et contient un débogueur Graphique. Il permet également la gestion des tests unitaires, l'intégration de logiciel de gestion de versions, et supporte le développement web avec Django.

Développé par l'entreprise tchèque JetBrains, c'est un logiciel multiplateforme qui fonctionne sous Windows, Mac OS X et Linux. Il est décliné en édition professionnelle, diffusé sous licence propriétaire, et en édition communautaire diffusé sous licence Apache (P.Makina, 17).

3.4.1. Les langages

Python : Est un langage de programmation puissant et facile Apprendre. Il dispose de structures de données de haut niveau et permet une approche simple mais efficace de la programmation orientée objet. Parce que sa syntaxe est élégante, que son typage est

dynamique et qu'il est facile à interpréter, Python est un langage idéal pour l'écriture de scripts et le développement rapide d'applications dans de nombreux domaines et sur la plupart des plateformes (P.Makina, 17).

L'interpréteur Python et sa vaste bibliothèque standard sont disponibles librement, sous forme de sources ou de binaires, pour toutes les plateformes majeures depuis le Site Internet https://www.python.org/ et peuvent être librement redistribués. Ce même site distribue et pointe vers des modules, des programmes et des outils tiers. Enfin, il constitue une source de documentation (W3). Avec Python, je me suis appuyé sur un Framework et les bibliothèques Python.

Flask : Ce Framework permet de développer des serveurs Web backend basés sur Python. Ce Framework est considéré comme le plus populaire et le meilleur pour les débutants.

```
from flask import Flask, render_template,
```

Scikit-Learn : Est une bibliothèque qui fournit une gamme d'algorithmes D'apprentissage supervisés et non supervisés via une interface cohérente en Python. La vision de la bibliothèque est un niveau de robustesse et de support requis pour une utilisation dans les Systèmes de production. Cela signifie qu'il faut se concentrer sur des préoccupations telles que la simplicité d'utilisation, la qualité du code, la collaboration, la documentation et les performances (Gupta, 17).

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import confusion_matrix
from sklearn import metrics
```

```
from sklearn.externals import joblib
```

Pandas : Est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques. Pandas est un logiciel libre sous licence.

Ce ne sont pas toutes les bibliothèques utilisées dans notre projet. Nous avons utilisé D'autres bibliothèques comme : Numpy, Matplotlib, scipy, Seaborn**.**

```
import pandas as pd
```

Numpy : est une extension du langage de programmation Python, destinée à manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableaux.


```
import numpy as np
```

3.5. Présentation de l'application

Figure 10: Page Principale.

Cette figure (la première fenêtre) de notre application. Elle contient deux boutons : le premier (**prédiction**) donne à l'utilisateur la possibilité de prédire le stade de sa maladie de foie après l'étape de remplissage des paramètres en entrées alors que le deuxième (**Fermer**) permet de fermer la fenêtre.

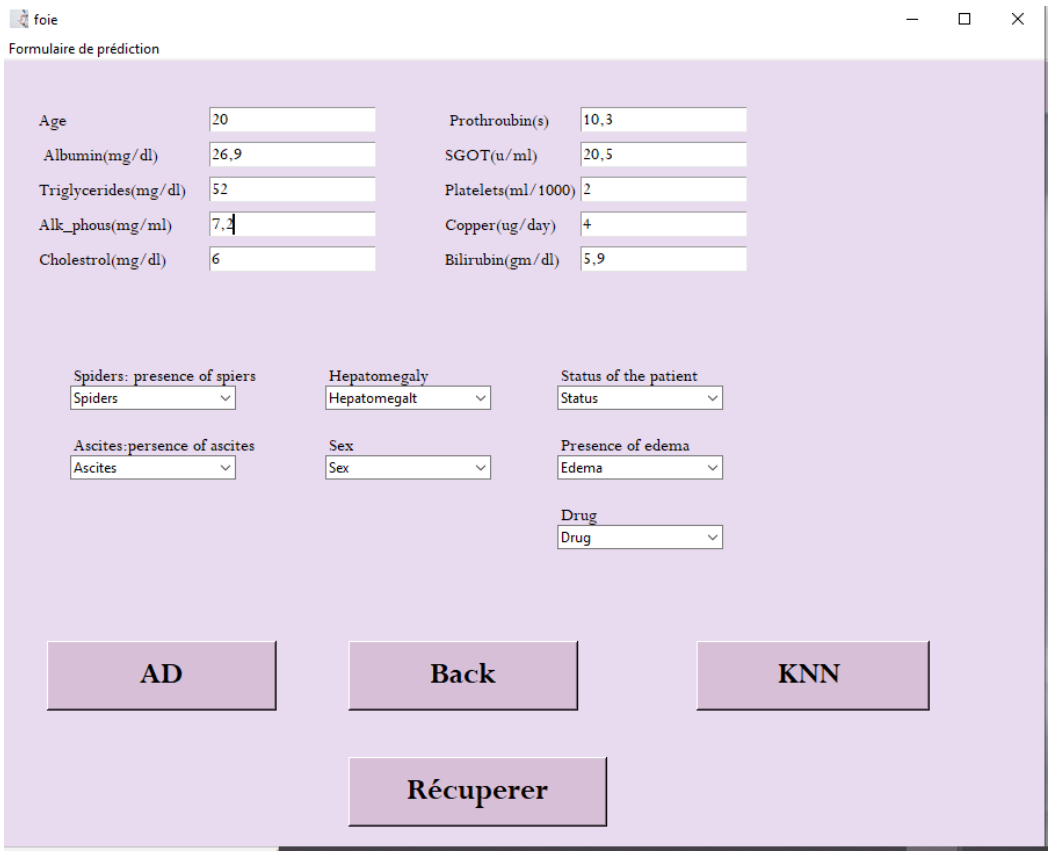

Figure 11: Formulaire de prédiction du stade de la maladie de foie.

Le formulaire Au-dessus donne la main à l'utilisateur pour remplir les informations le concernant et prédire à quelle classe est-il associé (stade de sa malade de foie {1.2.3.4}). Une fois ceci est fait le patient à la possibilité de choisir l'une des méthodes utilisées pour la prédiction : **(AD ou bien KNN)** et le résultat s'affiche sur l'écran de la fenêtre suivante.

Les boutons **«AD, KNN »** permettent de choisir laquelle des méthodes précédentes utilisée pour classer le nouveau patient et Le bouton **« récupérer »** pour récupérer les informations qui sont remplies par l'utilisateur.

Les résultats des prévisions sont affichés comme suit :

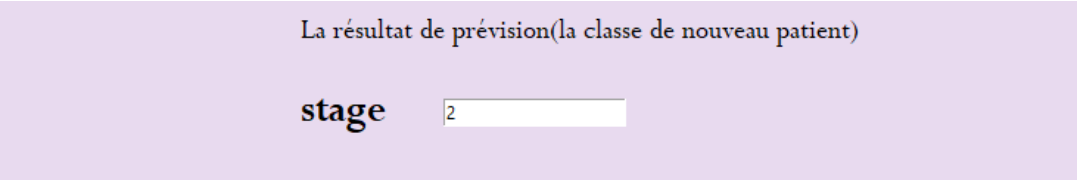

Figure 12: Résultats de prédiction du stade de la maladie de foie.

3.5. Conclusion

Ce dernier chapitre présente une étude liée à la mise en œuvre des différents algorithmes utilisés en plus des algorithmes proposés pour la migration entre classes. Cette étude nous a permis de faire une validation et une évaluation des performances de chacune des méthodes présentées.

Conclusion générale

Les maladies chroniques, telles que les maladies du foie, le diabète et le cancer, nécessitent une attention particulière en termes de prévention et de gestion des facteurs de risque pour réduire leurs effets néfastes. Comprendre les facteurs qui influent sur l'évolution de la maladie et leur rôle dans son développement est essentiel pour prendre les mesures appropriées. Cependant, la détermination manuelle des chances qu'un patient développe une maladie chronique en fonction de ses facteurs de risque est souvent complexe, en raison des interactions complexes entre ces facteurs. L'apprentissage automatique joue un rôle crucial dans ce contexte. Les techniques de prédiction basées sur l'apprentissage automatique se sont avérées efficaces pour faciliter la prise de décision et les prévisions à partir de la masse de données générées par l'industrie des soins de santé. Ces outils d'aide à la décision utilisant l'apprentissage automatique visent à améliorer la qualité des soins prodigués aux patients et à prédire l'apparition de maladies de manière plus précise. Grâce à ces avancées, il devient possible d'exploiter pleinement les données médicales pour obtenir des informations utiles et prendre des décisions éclairées pour la santé des patients.

Notre étude s'est focalisée sur l'application de techniques d'apprentissage automatique pour la prédiction de la présence et du type de maladies du foie. Nous avons exploité les méthodes des K plus proches voisins (KPP) et des arbres de décision (AD), en utilisant comme base de données médicales les informations fournies par le "Mayo Clinic Primary Biliary Cirrhosis Data". Cette base de données se concentre spécifiquement sur les maladies du foie et a été utilisée pour construire deux modèles d'apprentissage capables de classer les différents stades de la maladie (stades 1, 2, 3 ou 4). Les stades de la maladie de foie peuvent varier en fonction de leur gravité et de leur progression. Ces stades peuvent représenter différents degrés de fibrose, d'inflammation ou de dysfonctionnement hépatique. Il est essentiel de comprendre le stade spécifique de la maladie de foie d'un patient afin de déterminer le traitement approprié et d'évaluer le pronostic. Les techniques d'apprentissage automatique peuvent jouer un rôle important dans la prédiction et la classification des différents stades de la maladie de foie en utilisant des données médicales pertinentes. Notre étude s'est concentrée sur la construction de

modèles d'apprentissage automatique capables de prédire et de classifier les stades de la maladie de foie, tels que les stades 1, 2, 3 ou 4, en utilisant la base de données médicales "Mayo Clinic Primary Biliary Cirrhosis Data".

Les résultats de notre étude ont démontré que ces techniques d'apprentissage automatique sont non seulement efficaces pour prédire les maladies de foie, comme cela a été établi précédemment, mais elles peuvent également être utilisées avec succès pour améliorer la prise en charge des patients atteints de maladies du foie. En identifiant avec précision le stade de la maladie du foie, les médecins et les professionnels de la santé peuvent mettre en place des stratégies de traitement et de gestion adaptées à chaque patient, afin de ralentir la progression de la maladie et d'améliorer les résultats cliniques.

Ces résultats ouvrent des perspectives prometteuses pour l'application de l'apprentissage automatique dans le domaine de la médecine hépatique, en permettant une prise de décision plus éclairée et une meilleure personnalisation des soins pour les patients atteints de maladies du foie.

La diversité des stades de maladies de foie nous a conduits à réfléchir sur la possibilité d'une transition (migration) d'une classe à une autre. En d'autres termes, notre système permet de prédire, à partir des données réelles d'une personne, le type de maladie de foie dont elle souffre, tel que le type 3 par exemple. Les gens sont souvent préoccupés par les changements qu'ils doivent apporter à leur vie pour éviter de passer à une autre classe de maladie. Cette transition d'une classe à une autre peut entraîner une amélioration ou une dégradation de l'état de santé du patient. L'objectif principal est toujours de ralentir la progression de la maladie vers une forme invalidante. L'idée est donc d'intervenir le plus tôt possible et de manière continue pour prévenir l'aggravation et la gravité de la maladie.

Nous avons conçu un algorithme pour chacune des techniques, permettant de calculer les nouvelles valeurs des paramètres lors d'une migration d'une classe à une autre. De plus, nous avons développé une application web pour faciliter l'accès à cette méthode, à la fois pour les médecins et les patients souhaitant suivre leur état de santé. Cette application leur permet de visualiser l'impact des changements de paramètres, même minimes, sur leur état de santé et ainsi intervenir précocement pour prévenir toute aggravation.

Perspectives

Ce projet nous a permis d'explorer le domaine de l'apprentissage automatique et son application dans le domaine médical pour la prédiction des maladies de foies. Nous sommes très enthousiastes quant aux perspectives futures de notre travail, qui sont nombreuses :

- Explorer d'autres techniques de classification et évaluer leur performance pour ne retenir que celles qui fournissent les résultats les plus précis.
- Intégrer davantage de paramètres dans le modèle. En ajoutant des informations supplémentaires, nous pourrions améliorer la précision de nos prédictions pour des recherches futures.
- Collaborer avec des experts en médecine pour valider empiriquement nos résultats. En travaillant avec des médecins, nous pourrions confirmer la pertinence clinique de nos prédictions et les améliorer en fonction de leurs retours d'expérience.
- La mise en place d'une proposition de médicaments basée sur les résultats obtenus pour les patients.
- L'ajout d'une option de chat en direct pour permettre aux patients de discuter avec un médecin disponible concernant leur traitement.
- Utiliser notre projet comme un outil de formation pour les infirmières et les médecins nouvellement introduits dans le domaine de la maladie de foie.
- Permettre aux patients de choisir leurs médicaments pour une vie plus saine. S'il est mis en œuvre à grande échelle, il peut être utilisé dans des établissements médicaux tels que des hôpitaux et des cliniques, réduisant ainsi les temps d'attente pour les patients présentant des symptômes liés aux maladies des foies.
- À l'avenir, nous prévoyons également d'ajouter des comptes pour chaque utilisateur, afin de surveiller l'historique de leurs antécédents foies pour voir si leur état s'est amélioré ou détérioré.

Bibliographie

1mathieu-Dupas, E. (2010). *Algorithme des k plus proches voisins pondérés et application en diagnostic.* Plate-forme bioinformatique & biostatistique.

Arlot, S. (2009). *Classification supervisée : des algorithmes et leur calibration automatique.* Ecolde centrale de Paris.

Benzaki, Y. (2018). *Introduction à l'algorithme K Nearst Neighbors (K-NN)*. Récupéré sur https://mrmint.fr/introduction-k-nearest-neighbors

Brains, J. (2020). *py*. Récupéré sur https://www.jetbrains.com/lp/pycharm-pro/

Ghebouli farida, L. A. (2020). *Application pour la classification et la migration entre classes de la maladie cardiaque.mémoire de master:université* de bba.

Gupta. (2017). *Gupta, P.to wards datascience.(2017).Decision Trees in Machine Learning.[enligne].Disponiblesu: https ://towardsdatascience.com/decision-trees-in-machinelearning-641b9c4e8052.*

Ile.Alt. (2019). *IlemonaS.Atawodi.(2019).A Machine Learning Approach to Network Intrusion Detection System Using K NearestNeighbor and Random Forest. de master : universit´e de Southern nearest-neighbor-classification-scikit-learn.*

Ismaili, & Atawodi. (2019). *A Machine Learning Approach to Network Intrusion Detection System Using K NearestNeighbor and Random Forest. de master : universit´e de.*

Java T point. (s.d.). *K-Nearest Neighbor(KNN) Algorithm for Machine Learning*. Récupéré sur https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

Lev.Kiw. (2018). *Apprentissage et Machine Learning*. Récupéré sur Lev kiwi: https ://levkiwi.ch/apprentissage-et-machine-learning/

Mr Mint. (2013). *Mr. Mint Machine Learning made easy*. Récupéré sur Mr Mint: https://mrmint.fr/apprentissage-supervise-machine-learning

P.Makina. (2017). *Ga¨el, P.Makina Corpus.(2017).Initiation au Machine Learning avec Python.[enligne].Disponible sur : https ://makina-corpus.com/blog/metier/2017/initiation-aumachine-learning-avec-python-pratique.*

Pang-Ning Tan, M. S. (2006). *Introduction to Data Mining.* Pearson Addison Wesley.

Plantevit, M. (2019). *Fouille de Données : Processus ECD, fouille d emotids et clusturing.* Université Claude Bernard Lyon 1.

Songul, C. (2016). Comparison of Performance of Decision Tree Algorithms and Random Forest: An Application on OECD Countries Health Expenditures. *International Journal of Computer Applications, 1*(138), 37-41.

St.Kum. (2005). *P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining:*.

Usama Fayyad, G. P.-S. (1996). From Data mining to knowledge discovery in databases. *AI Magazine Volume*, 37-55.