**REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE**

**MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE**

***Université de Mohamed El-Bachir El-Ibrahimi - Bordj Bou Arreridj***

***Institut d'électronique et télécommunication***

# *Mémoire*

*Présenté pour obtenir*

**LE DIPLOME DE MASTER**

FILIERE : Electronique

**Spécialité :** Système des télécommunications

Par

- **MERAH** Imene
- **LAKEHAL** Imane

*Intituléé*

## *Application de l'Intelligence Artificielle Pour la Prédiction et la Classification du Cancer du Pancréas*

***Soutenu le :*** *09/06/2026*

***Devant le Jury composé de :***

| *Nom & Prénom* | *Grade* | *Qualité* | *Etablissement* |
|---|---|---|---|
| *M. ASBAI Nassime* | *Prof* | *Président* | *Univ-BBA* |
| *Dr BEHIH Mohamed* | *MCB* | *Encadreur* | *Univ-BBA* |
| *Pr BOUTTOUT Farid* | *Prof* | *Co- Encadreur* | *Univ-BBA* |
| *M. ADOUI Ibtissem* | *MCA* | *Examinateur* | *Univ-BBA* |

***Année Universitaire 2025/2026***

# *Remerciements*

Nous remercions d'abord Dieu qui nous a donné la force et le courage d'accomplir cet humble travail.

Nous exprimons nos remerciements avec un grand plaisir et un grand respect à notre cher promoteur : **Dr. BEHIH Mohamed** pour sa disponibilité, sa gentillesse, son soutien et ses encouragements.

Nous remerciements vivement les membres du jury pour avoir accepté d'évaluer notre travail... Nous aimerions aussi remercier nos familles, nos amis et collègue, ainsi que tous ceux qui ont contribués de prêt ou de loin à la réalisation de ce travail.

Nous remercions tous nos professeurs du département d'électronique à l'université Mohamed El-Bachir El-Ibrahimi - Bordj Bou Arreridj.

# *Dédicace*

*À mon paradis, à la source de ma joie et de mon bonheur, à celle qui m'a entourée d'amour, de tendresse et de sacrifices, qui a toujours été à mes côtés pour me soutenir, me guider et m'encourager,*
*À ma cher Maman.*

*Ma source de vie, d'amour et d'affection, à mon support qui était toujours à mes côtés pour me soutenir et m'encourager,*
*À mon cher Papa.*

*À mon frère Toufik pour son amour et sa présence, sa confiance en moi tout au long de mon parcours.*

*À ma grande sœur Yasmine qui n'at pas cessé de me conseiller, encourager et de me soutenir tout au long de mes études.*

*À mes petites Youcef & Norsine .*

*À mon Tchiko.*

*À tous les membres de ma grande famille.*

*Sans oublier mon binôme pour leur soutien moral leur patience et leur compréhension tout au long de ce projet.*

*À mes chers amis Aya, Gigi pour notre amitié merci d'être à mes côtés.*

*À tous mes collègues de promotion de 2eme année master.*

*Je vous aime profondément.*

***Imene Merah***

*Je dédie cet ouvrage*
*A ma maman qui m'a soutenu et encouragé durant ces années d'études.*
*Qu'elle trouve ici le témoignage de ma profonde*

*reconnaissance A mon frère Marouane,*

*Ma sœur Chahra*

*Mes grands parents*

*Et Ceux qui ont partagé avec moi tous les moments de ma vie.*
*Ils m'ont chaleureusement supporté et encouragé tout au long de mon*

*parcours. A ma famille, mes proches et à ceux qui me donnent de l'amour*

*et de la vivacité.*

*A mes meilleurs amis Imene & Aya qui m'ont toujours encouragé, et à qui je souhaite plus de succès.*

*A tous ceux que j'aime.*

***Imane Lakehal***

## Résumé

Cette étude porte sur le développement et l'évaluation de modèles d'intelligence artificielle pour la prédiction du cancer du pancréas à partir de biomarqueurs et de données médicales. Plusieurs algorithmes d'apprentissage supervisé, notamment la régression logistique, la forêt aléatoire et la logique floue, ont été utilisés et comparés en termes de performance.

Des étapes de prétraitement des données (nettoyage, standardisation et division des données) ont été appliquées afin d'améliorer la qualité de l'apprentissage. Les modèles ont été évalués à l'aide de mesures telles que l'Accuracy, la précision, le rappel et le F1-score, ainsi que par des outils comme la matrice de confusion.

Les résultats montrent que la forêt aléatoire offre les meilleures performances, confirmant l'efficacité des techniques d'apprentissage automatique dans l'aide au diagnostic précoce du cancer du pancréas.

## ملخص

تتناول هذه الدراسة تطوير وتقييم نماذج الذكاء الاصطناعي للتنبؤ بسرطان البنكرياس اعتمادًا على المؤشرات الحيوية والبيانات الطبية. تم استخدام عدة خوارزميات تعلم مُراقَب، مثل الانحدار اللوجستي والغابة العشوائية والمنطق الضبابي، ومقارنتها من حيث الأداء

كما تم تطبيق خطوات معالجة مسبقة للبيانات مثل التنظيف، التقييس، وتقسيم البيانات بهدف تحسين جودة التعلم. وتم تقييم النماذج ، إضافة إلى مصفوفة F1 ، ودرجة(Recall) ، الاسترجاع(Precision) ، الدقة الإيجابية(Accuracy) باستخدام مقاييس مثل الدقة الالتباس.

أظهرت النتائج أن نموذج الغابة العشوائية حقق أفضل أداء، مما يؤكد فعالية تقنيات التعلم الآلي في دعم التشخيص المبكر لسرطان البنكرياس.

## Abstract

This study focuses on the development and evaluation of artificial intelligence models for predicting pancreatic cancer using medical data and biomarkers. Several supervised machines learning algorithms, including Logistic Regression, Random Forest, and Fuzzy Logic, were implemented and compared in terms of performance.

Data preprocessing steps such as cleaning, standardization, and data splitting were applied to improve the quality of learning. The models were evaluated using metrics such as Accuracy, Precision, Recall, and F1-score, as well as tools like the confusion matrix.

The results show that the Random Forest model achieved the best performance, confirming the effectiveness of machine learning techniques in supporting early diagnosis of pancreatic cancer.

# Table des matières

# Liste des tableaux

# Liste des figures

# Liste des abréviations

**IA :** Intelligence Artificielle

**ML :** Machine Learning (Apprentissage Automatique)

**DL :** Deep Learning (Apprentissage Profond)

**OMS :** Organisation Mondiale de la Santé

**RF :** Random Forest (Forêt Aléatoire)

**SVM :** Support Vector Machine

**KNN :** K-Nearest Neighbors (K plus proches voisins)

**CA19-9 :** Carbohydrate Antigen 19-9

**LYVE1:** Lymphatic Vessel Endothelial Hyaluronan Receptor 1

**REG1A:** Regenerating Family Member 1 Alpha

**REG1B:** Regenerating Family Member 1 Beta

**TFF1 :** Trefoil Factor 1

**IRM :** Imagerie par Résonance Magnétique

**CT SCAN:** Computed Tomography Scan (Tomodensitométrie)

**PDAC:** Pancreatic Ductal Adenocarcinoma

**NON-PDAC:** Non Pancreatic Ductal Adenocarcinoma

**TP:** True Positive

**TN:** True Negative

**FP:** False Positive

**FN:** False Negative

**XAI :** Explainable Artificial Intelligence

**NTIC :** Nouvelles Technologies de l'Information et de la Communication

**AIM :** Artificial Intelligence Model

**K-Means** : Algorithme de partitionnement en K-moyennes

**Softmax :** Fonction d'activation normalisée

**St-Alors :** Standard Algorithm for Logistic Regression (à préciser selon ton texte)

# Introduction générale

Le cancer pancréatique est l'un des cancers les plus redoutés dans le monde. Pour cause sa progression asymptomatique, la difficulté de son diagnostic, et le peu de marqueurs tumoraux spécifiques à ce type de cancer font de cette pathologie l'une des plus meurtrière. Le cancer pancréatique est divisé en deux types : le cancer exocrine qui dépend du pancréas exocrine, et le cancer endocrine au dépend du pancréas endocrine. La tumeur est le plus souvent localisée à la tête du pancréas (près de 70% des cas). Le type histologique le plus fréquent est l'adénocarcinome. [1]

A nos jours et avec l'inondation totale de l'intelligence artificielle aux différents aspects de nos vies quotidiennes, il est très utile d'unir les capacités énormes de cette essor technologique avec les efforts des praticiens pour arriver à un meilleur rondement de prédiction et de classification de différentes maladies et compris le cancer pancréatique.

Ce projet de fin d'études est consacré à l'exploitation et l'application des deux algorithmes de l'apprentissage automatique des forêts aléatoires, et régression logistique, et l'algorithme de l'intelligence artificielle de haut niveau de la logique floue pour la prédiction et la classification du cancer pancréatique.

Ce manuscrit est divisé en trois chapitres initialisés par cette introduction générale. Le premier chapitre est consacré à la définition de différentes notions de base qui dépendent du domaine de l'intelligence artificielle ainsi que la description de quelques l'application de cette technologie dans le domaine de la médicine, et du soin de santé.

Dans le deuxième chapitre, les algorithmes des forêts aléatoires, et régression logistique, et de la logique floue sont détaillés.

Les trois algorithmes détaillés dans le chapitre précédant sont appliqués sur une base de données contient des marqueurs biologiques liés au cancer pancréatique. Les données de cette base de données sont au début exploitées et analysées, puis traitées par les trois algorithmes, et les différentes sortes de prédiction et de classification sont ensuite représentés et comparés.

A la fin, une conclusion générale résumera l'ensemble des connaissances et résultats obtenus.

# Chapitre I. Intelligence artificielle appliquée au diagnostic médical

## I.1. Introduction

L'intelligence artificielle (AI) est un essor technologique qui s'articule sur nombreux domaines et qui cible des applications qui sont liées directement aux différents aspects de nos vies quotidiennes. L'IA transforme la médecine en facilitant le diagnostic, le traitement, et en soutenant la médecine prédictive et personnalisée. De cette manière, l'AI pu contribuer à une révolution réelle dans le domaine de la médicine et du soin de santé.

Ce chapitre est consacré définition des éléments de base de la technologie de l'intelligence artificielle qui nous aident de comprendre mieux ces principes, sous-domaines, domaines d'applications et plus précisément dans le domaine de la médicine de soin de santé.

Ce chapitre termine par une conclusion.

## I.2. Définitions

L'intelligence artificielle (IA) est une technologie qui permet aux ordinateurs et aux machines de simuler l'apprentissage, la compréhension, la résolution de problèmes, la prise de décision, la créativité et l'autonomie de l'être humain. L'intelligence artificielle est un englobe d'une vaste gamme de domaines intellectuels, elle peut être définie comme un style de programmation pour traiter des données varies et complexes selon des règles pour atteindre des objectifs bien déterminés. L'intelligence artificielle consiste à étudier les systèmes qui agissent de manière intelligente, perceptible par tout observateur. Cependant, cette définition ne capture pas entièrement la portée de l'intelligence artificielle. Souvent, les techniques d'IA sont utilisées pour résoudre des problèmes relativement simples ou des problèmes complexes à l'intérieur de systèmes plus vastes. Une définition plus complète de l'intelligence artificielle pourrait être formulée comme suit : l'utilisation de méthodes inspirées par le comportement intelligent des êtres humains et d'autres animaux pour résoudre des problèmes complexes.

Les applications et les appareils assistés par l'IA peuvent voir et identifier les objets. Ils sont capables de comprendre le langage humain et d'y répondre. Ils peuvent tirer des enseignements de nouvelles informations et de nouvelles expériences. Ils peuvent formuler des recommandations détaillées aux utilisateurs et aux experts. Ils peuvent agir indépendamment, remplaçant ainsi l'intelligence ou l'intervention humaine. [2]

## I.3. Caractéristiques de l'intelligence artificielle

L'intelligence artificielle est une technologie informatique simulant les capacités cognitives humaines. Les algorithmes de l'intelligence artificielle traitent grands volumes de données avec une autonomie dans la prise de décision et la capacité de résolution de problèmes complexes. Les caractéristiques communes à tous les types d'IA sont les suivantes :

- ✓ Perception de l'environnement en prenant en compte la complexité du monde réel.
- ✓ Collection et interprétation des données puis traitement d'informations.
- ✓ Prise de décision (y compris le raisonnement et l'apprentissage) et exécution de tâche en adéquation avec l'environnement, cela avec un certain degré d'autonomie.
- ✓ Réalisation d'objectifs spécifiques. [3] sont donnés par la figure I.1.

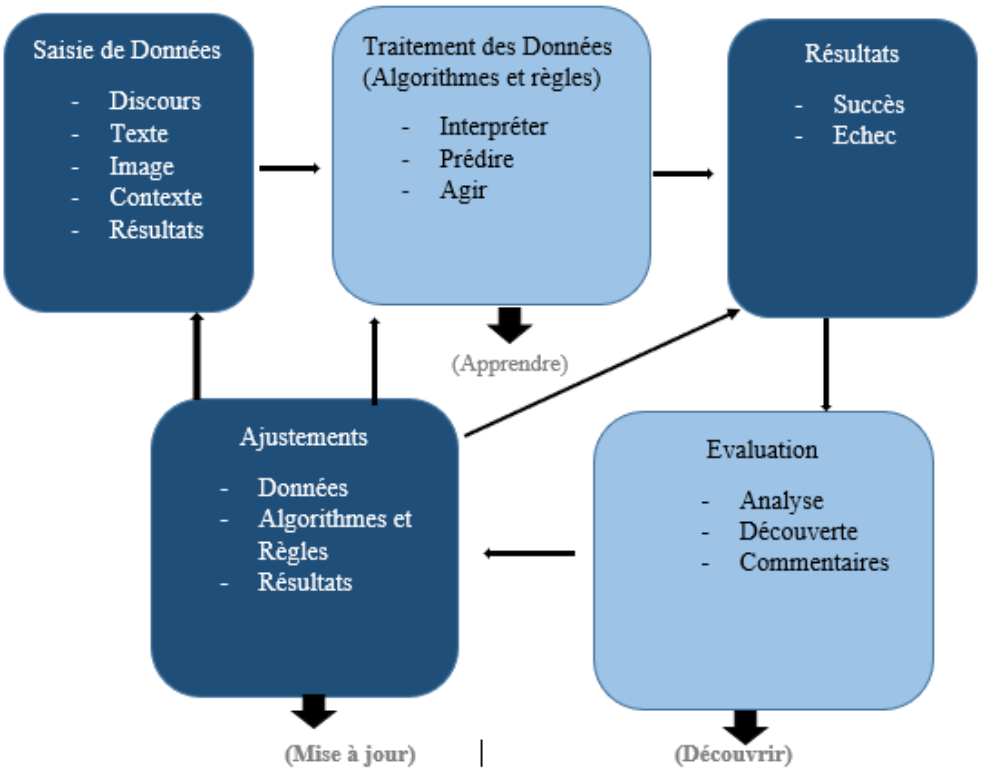


*Figure I.1: Schéma du mode de fonctionnement de l'IA.*

## I.4. Types d'intelligence artificielle

Les Intelligences Artificielles sont généralement classées selon deux grandes approches : leurs **capacités globales** et leur **degré d'évolution technologique**. En se basant sur deux approches, sept types d'intelligence artificielle peuvent être définis. Ces différents types sont donnés par la figure I.2.

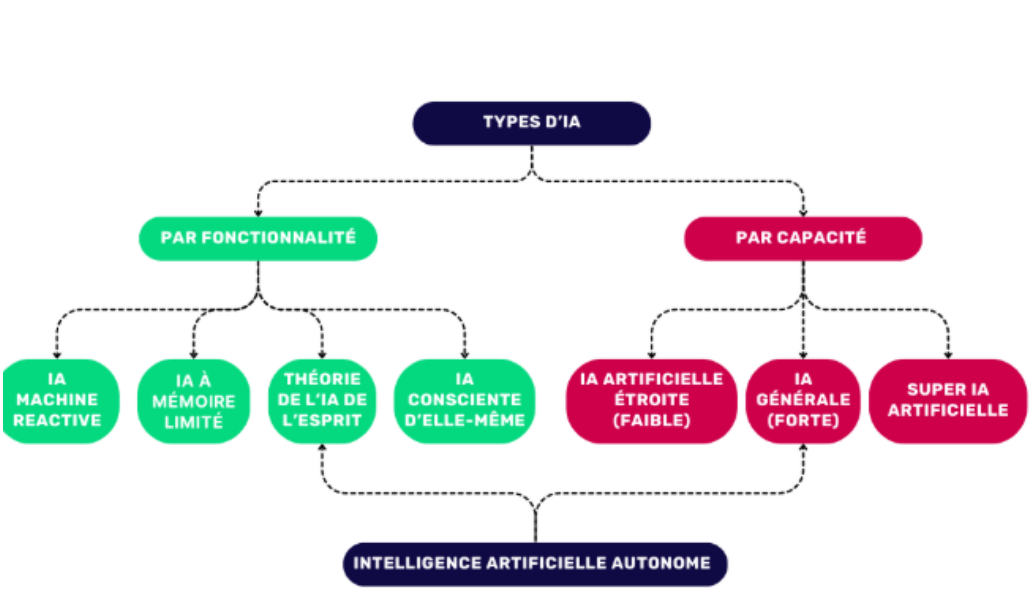


*Figure I.2: Les 7 types d'intelligence artificielle.*

**I.4.1. Intelligence artificielle étroite ou IA faible** : l'IA actuelle est limitée à des tâches spécifiques où elle surpasse parfois les humains, mais elle nécessite toujours notre intervention pour être formée. Toutes les autres formes d'IA sont théoriques.

**I.4.2. Intelligence artificielle générale (AGI) ou IA forte** : l'AGI est actuellement un concept théorique. Contrairement à l'IA étroite, elle réutiliserait ses connaissances pour de nouvelles tâches sans entraînement humain.

**I.4.3. Super IA artificielle** : la super IA artificielle serait capable de ressentir des émotions, de développer des besoins et des croyances, allant au-delà de simplement comprendre les sentiments humains.

**I.4.4. Intelligence artificielle réactive** : l'IA réactive se spécialise dans des tâches précises en exploitant les données pour produire des résultats intelligents. Par exemple, Depp Blue d'IBM a battu Garry Kasparov en analysant les configurations d'échecs, illustrant ainsi la nature ciblée et axée sur les données de l'IA réactive.

**I.4.5. IA à mémoire limitée** : l'IA à mémoire limitée se souvient des événements passés et surveille des situations spécifiques pour décider des actions futures, s'améliorant avec l'entraînement. Par exemple, un chat bot d'IA générative prédit le mot ou l'élément suivant dans un contexte donné.

**I.4.6. IA de l'esprit** : l'IA émotionnelle vise à comprendre et répondre aux pensées et émotions humaines, personnaliser les interactions en fonction des besoins émotionnels, et développer une théorie de l'esprit. Les chercheurs espèrent qu'elle pourra analyser diverses données, telles que les voix et les images, pour interpréter les sentiments humains.

**I.4.7. IA consciente d'elle-même** : elle percevrait ses états internes et développerait émotions, besoins et convictions. [4]

## I.5. Sous-domaines de l'intelligence artificielle

L'intelligence artificielle est la capacité d'un système à simuler l'intelligence humaine. Le pionnier américain de l'IA, Marvin Minsky, avait une définition plus précise : l'IA est "une science dont le but est de faire réaliser par une machine des tâches que l'homme accomplit en utilisant son intelligence". La technologie de l'intelligence artificielle regroupe trois sous-domaines principaux comme le montre la figure I.3. [5] Ces trois sous-domaines sont : l'apprentissage approfondi (*Deep Learning*), l'apprentissage automatique (*Machine Learning*), et intelligence artificielle de haut niveau.

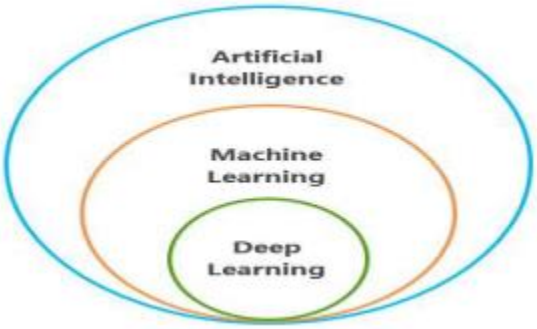


***Figure I.3: Relation entre les différents sous-domaines de l'intelligence.***

### I.5.1. Apprentissage profond

**I.5.1.1. Définition** : l'apprentissage profond est un ensemble de techniques d'apprentis - sage automatique qui a permis des avancées importantes en intelligence artificielle dans les dernières années. Dans l'apprentissage automatique, un programme analyse un ensemble de données afin de tirer des règles qui permettront de tirer des conclusions sur des nouvelles données. L'apprentissage profond est basé sur ce qui a été appelé, par analogie, des « Réseaux de neurones artificiels », composé de milliers d'unités (les neurones) qui effectuent chacune de petites opérations simples. Les résultats d'une première couche de « neurones » servent d'entrée aux calculs d'une deuxième couche et ainsi de suite. [6]

L'apprentissage en profondeur permet aux modèles informatiques avec plusieurs couches de traitement d'apprendre plusieurs degrés d'abstraction pour les représentations de données. Ces techniques ont considérablement amélioré l'état de l'art en matière de reconnaissance vocale, de reconnaissance visuelle d'objets, de détection d'objets et de divers autres domaines tels que le développement de médicaments et la génomique. [7] La figure I.4 montre le principe de fonctionnement de l'apprentissage profond.

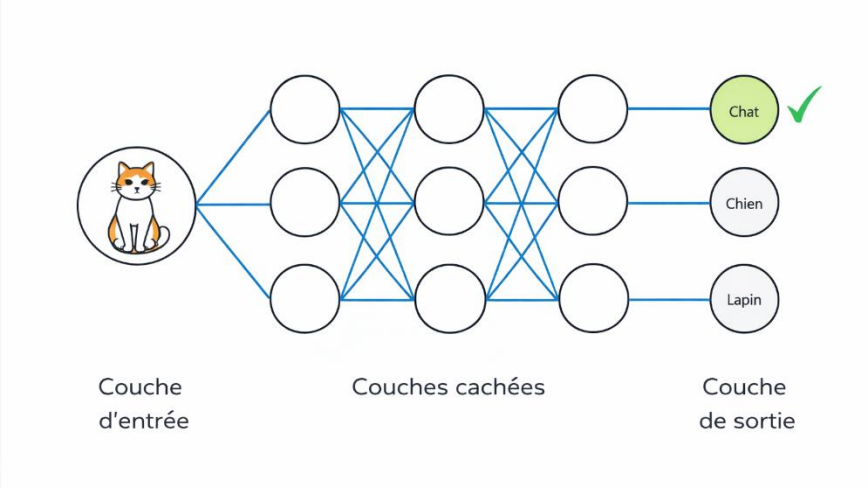


*Figure I.4: le principe de fonctionnement de l'apprentissage profond.*

L'application de l'apprentissage profond suit un processus universel en étapes itératives : définition du problème, collecte et préparation des données, choix ou conception d'une architecture de réseau de neurones, entraînement sur les données, évaluation des performances et déploiement du modèle dans une application finale. Le processus d'application de l'apprentissage profond : la méthode pour créer et déployer un modèle d'apprentissage profond se déroule selon les étapes clés suivantes.

- ✓ **Définition de l'objectif :** identifier clairement le problème à résoudre, par exemple : classification d'images, reconnaissance vocale, traduction, … etc.
- ✓ **Collection et labellisation des données :** l'apprentissage approfondi nécessite un immense volume de données brutes et annotées (étiquetées) correctement.
- ✓ **Prétraitement :** cette étape consiste à normaliser, redimensionner, et diviser les données en trois parties **: d'entraînement**, de **validation, et de test.**
- ✓ **Choix de l'architecture :** consiste à choisir une architecture de réseau de neurones adaptée aux données traitées. Les plus courantes sont : les réseaux de neurones convolutifs, et les réseaux de neurones récurrents ou les Transformers.
- ✓ **Entraînement du modèle :** alimenter le réseau avec les données d'entraînement. Durant cette phase, le modèle ajuste ses poids internes grâce à des algorithmes d'optimisation et une fonction de coût pour minimiser l'erreur de prédiction.
- ✓ **Évaluation et réglage :** tester le modèle avec l'ensemble de validation, et ajuster les hyperparamètres comme le taux d'apprentissage ou le nombre de couches pour éviter le surapprentissage ou le sous-apprentissage.
- ✓ **Déploiement et surveillance :** intégrer le modèle entraîné dans un environnement réel comme un site web et continuer à surveiller ses prédictions dans le temps.

**I.5.1.2. Domaine d'application de l'apprentissage profond** : les domaines d'application des techniques de l'apprentissage profond sont divers. En effet, ces techniques se développent dans le domaine de l'informatique appliquées aux NTIC, à la robotique, à la reconnaissance ou comparaison de formes, la sécurité, la santé, la pédagogie assistée par l'informatique, et plus généralement à l'intelligence artificielle. L'apprentissage profond permet à un ordinateur déformable d'analyser les émotions révélées par un visage photographié ou filmé, ou analyser les mouvements et position des doigts d'une main, ce qui peut être utile pour traduire le langage des signes, améliorer le positionnement automatique d'une caméra, etc. Elles sont utilisées pour certaines formes d'aide au diagnostic médical (ex : reconnaissance automatique d'un cancer en imagerie médicale), ou de prospective ou de prédiction (ex : prédiction des propriétés d'un sol filmé par un robot ou la prédiction des maladies). [8]

**I.5.1.3. Exemples d'application de l'apprentissage profond** : l'apprentissage profond est utilisé dans différents secteurs, de la conduite automatisée aux Dispositifs médicaux. Grâce au l'apprentissage profond, nous pouvons maintenant. [8] : faire une colorisation des images en noir et blanc, ajouter des sons à des films silencieux, faire de la traduction automatique, faire de la classification des objets en photographies, générer d'écriture automatique, génération de légende d'image, etc.

### I.5.2. Apprentissage automatique

**I.5.2.1. Définition** : l'apprentissage automatique, est un sous-type de l'intelligence artificielle qui permet aux ordinateurs d'apprendre sans avoir été explicitement programmés. Cette définition est celle donnée en 1959 par un de ses pionniers, Arthur Samuel. L'apprentissage automatique travaille à partir d'exemples. Il utilise des algorithmes pour analyser statistiquement des données et identifier des modèles. Ces modèles sont ensuite utilisés pour prédire des résultats (figure I.5), avoir une meilleure compréhension des processus qui génèrent ces données ou pour prendre des décisions. Il consiste à utiliser des données pour entraîner un modèle informatique afin d'accomplir une tâche spécifique. Ce sous-domaine de l'IA s'articule sur des algorithmes capables de prédire des tâches complexes. Il est utilisé dans des contextes variés, comme l'analyse de données expérimentales.

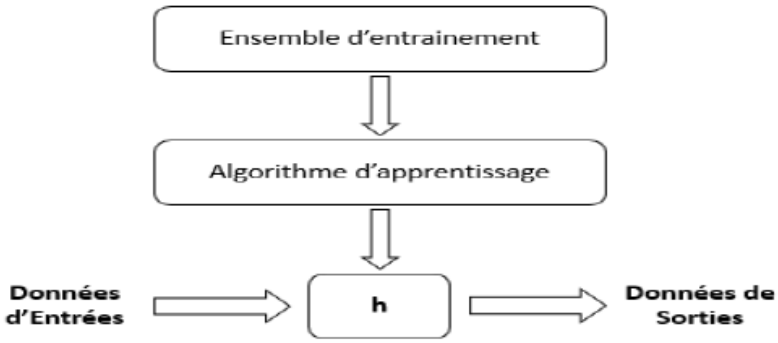


*Figure I.5: Fonctionnement de l'apprentissage automatique.*

L'application de l'apprentissage automatique passe par les phases suivantes :

- ✓ **Définition du problème :** clarifier le résultat attendu, par exemple : prédire des prix, classifier des E-mails en spam, détecter des anomalies.
- ✓ **Collection et préparation des données :** rassembler les données pertinentes et les préparer par la suppression des erreurs, traitement des valeurs manquantes, et le formatage pour garantir la qualité de l'apprentissage.
- ✓ **Choix de l'algorithme :** consiste à sélectionner la méthode d'apprentissage adaptée : **apprentissage supervisé, apprentissage non supervisé, ou apprentissage par renforcement.**
- ✓ **Entraînement du modèle :** fournir les données préparer à l'algorithme choisi pour qu'il ajuste ses paramètres internes et apprenne à reconnaître des motifs.
- ✓ **Évaluation :** tester le modèle avec un ensemble de données distinct pour mesurer ses performances en termes de la précision, et du taux d'erreur, par exemple.
- ✓ **Déploiement :** inclure le modèle final dans un environnement de production réel afin qu'il puisse traiter de nouvelles données et effectuer des prédictions.

Les algorithmes de l'apprentissage automatique les plus utilisés pour la classification sont : arbres de décision, forêts aléatoires, régression logistique, machines à vecteurs de support, K-plus proches voisins.

### I.5.2.2. Application de l'apprentissage automatique

- ✓ La classification : classer des fichiers en fonction de leur contenu, détecter des anomalies sur une chaine de production, détecter des spams…

- ✓ La régression : prédire une valeur numérique, utile pour connaître l'évolution de la météo ou du cours d'une action.
- ✓ Le regroupement : regrouper des clients par persona et habitudes d'achat.
- ✓ Moteurs de recherche, moteurs de recommandation…
- ✓ Agents conversationnels (chatbots)…

### I.5.2.3. Les algorithmes d'apprentissage automatique

Les SVMs (*Support Victor Machine* ou Machine à vecteurs de support), développés dans les années 1990, sont une famille d'algorithmes d'apprentissage automatique qui permettent de résoudre des problèmes de classification. Les SVMs ont pour but de séparer les données en classes à l'aide d'une frontière, de telle façon que la distance entre les différents groupes de données et la frontière qui les sépare soit maximale. Cette distance est aussi appelée « marge » et les SVMs sont ainsi qualifiés de « séparateurs à vaste marge », les « vecteurs de support » étant les données les plus proches de la frontière.

Pour un usage général, les arbres de décision sont utilisés pour représenter visuellement les décisions et montrer ou éclairer la prise de décision. Lorsqu'on travaille avec l'apprentissage automatique et l'exploration de données, les arbres de décision sont utilisés comme modèle prédictif. Ces modèles cartographient les observations de données et tirent des conclusions sur la valeur cible des données. L'objectif de l'apprentissage par arbre de décision est de créer un modèle qui prédira la valeur d'une cible en fonction de variables d'entrée. Dans le modèle prédictif, les attributs des données qui sont déterminés par l'observation sont représentés par les branches, tandis que les conclusions sur la valeur cible des données, sont représentées dans les feuilles.

Les forêts d'arbres décisionnels ou forêts aléatoires (*Random Forest classifier*). L'algorithme des forêts d'arbres décisionnels effectue un apprentissage sur de multiples arbres de décision entrainés sur des sous-ensembles de données légèrement différents.

La KNN, une abréviation de K-Nearest-Neighbors, est une méthode d'apprentissage supervisé. Elle est utilisée pour la classification. L'entrée consistera en les k exemples d'entraînement les plus proches dans un espace, la sortie est l'appartenance à une classe. L'algorithme assignera un nouvel objet à la classe la plus commune parmi ses k plus proches voisins. [9]

**I.5.2.4. Les types d'apprentissages** : les apprentissages automatiques peuvent se catégoriser selon le type d'apprentissage qu'ils emploient : l'apprentissage supervisé, l'apprentissage non supervisé, l'apprentissage semi-supervisé, l'apprentissage partiellement et l'apprentissage par renforcement. La figure I.6 montre les types de l'apprentissage automatique.

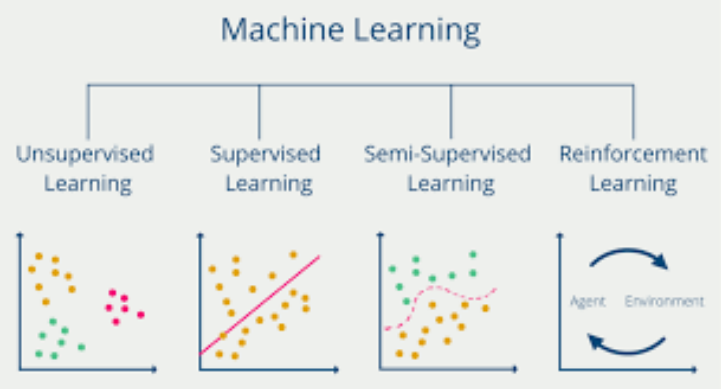


*Figure I.6: Différentes types de l'apprentissage automatique.*

- ✓ **L'apprentissage supervisé** : si les classes sont prédéterminées et les exemples connus, le système apprend à classer selon un modèle de classement ; on parle alors d'apprentissage supervisé (ou d'analyse discriminante). Un expert doit préalablement correctement étiqueter des exemples. L'apprenant peut alors trouver ou approximer la fonction qui permet d'affecter la bonne « étiquette » à ces exemples. Parfois il est préférable d'associer une donnée non pas à une classe unique, mais une probabilité d'appartenance à chacune des classes prédéterminées (on parle alors d'apprentissage supervisé probabiliste). L'analyse discriminante linéaire ou les SVM sont des exemples typiques. (Figure I.7)

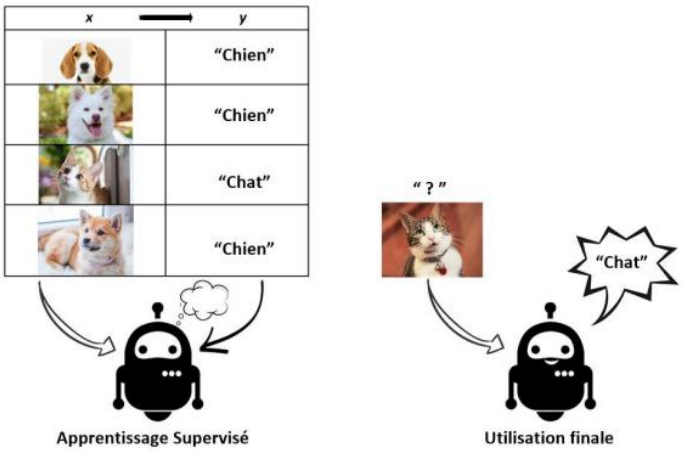


*Figure I.7: Fonctionnement de l'apprentissage supervisé.*

- ✓ **L'apprentissage non supervisé** : quand le système ou l'opérateur ne dispose que d'exemples, mais non d'étiquettes, et que le nombre de classes et leur nature n'ont pas été prédéterminés, on parle d'apprentissage non supervisé. Aucun expert n'est disponible ni requis. L'algorithme doit découvrir par lui-même la structure plus ou moins cachée des données. Le système doit ici dans l'espace de description cibler les données selon leurs attributs disponibles, pour les classer en groupe homogènes d'exemples. La similarité est généralement calculée selon la fonction de

distance entre paires d'exemples. C'est ensuite à l'opérateur d'associer ou déduire du sens pour chaque groupe. Divers outils mathématiques et logiciels peuvent l'aider. On parle aussi d'analyse des données en régression. Si l'approche est probabiliste (c'est à dire que chaque exemple au lieu d'être classé dans une seule classe est associé aux probabilités d'appartenir à chacune des classes), on parle alors de « soft clustering » (par opposition au « hard clustering »). (Figure I.8**)**

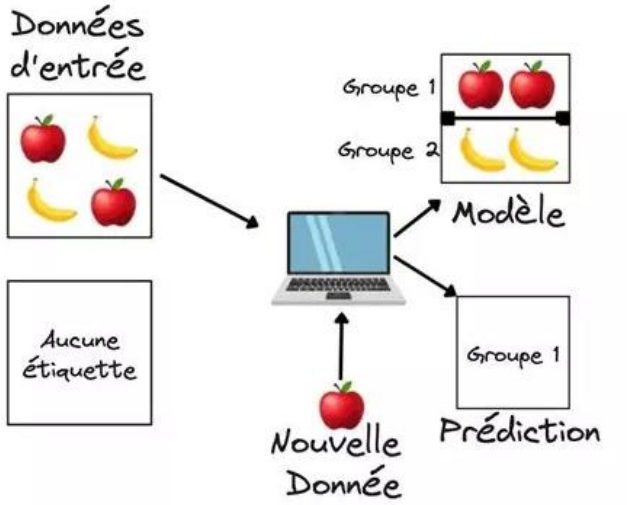


***Figure I.8: Fonctionnement de l'apprentissage non supervisé.***

- ✓ **L'apprentissage semi-supervisé** : effectué de manière probabiliste ou non, il vise à faire apparaître la distribution sous-jacente des « exemples » dans leur espace de description. Il est mis en œuvre quand des données (ou « étiquettes ») manquent... Le modèle doit utiliser des exemples non-étiquetés pouvant néanmoins renseigner. L'apprentissage partiellement supervisé (probabiliste ou non) quand l'étiquetage des données est partiel. C'est le cas quand un modèle énonce qu'une donnée n'appartient pas à une classe A, mais peut-être à une classe B ou C (A, B et C et 3 maladies par exemple évoquées dans le cadre d'un diagnostic différentiel).
- ✓ **L'apprentissage par renforcement** : l'algorithme apprend un comportement étant donné une observation. L'action de l'algorithme sur l'environnement produit une valeur de retour qui guide l'algorithme. [10] (Figure I.9**).**

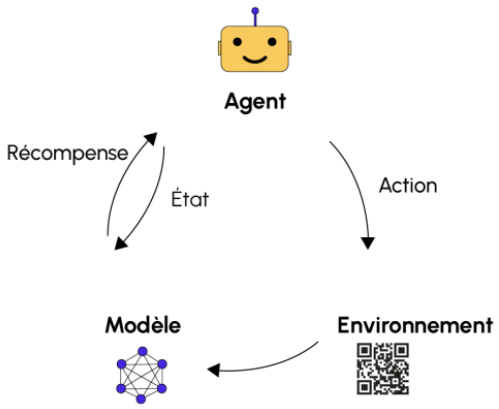


*Figure I.9: Exemple de l'apprentissage par renforcement.*

**I.5.3. Intelligence Artificielle** : l'IA est un sous-domaine en pleine expansion, combinant auto-apprentissage et adaptabilité pour traiter des données complexes et variées. Elle est utilisée dans des applications concrètes comme la recherche médicale et clinique, l'optimisation des réseaux mobiles, et la vision par ordinateur. Parmi les piliers de ce sous-domaine, on trouve : systèmes experts, logique floue, systèmes à base de règles, planification et ordonnancement, programmation par contraintes, représentation des connaissances, etc.

## I.6. Comparaison entre les différents domaines de l'intelligence artificielle :

La comparaison entre les différents domaines de l'intelligence artificielle est donnée par le tableau I.1.

| Aspect | Intelligence Artificielle | Apprentissage automatique | Apprentissage profond |
|---|---|---|---|
| **Définition** | Domaine global qui vise à imiter l'intelligence humaine | Sous-domaine de l'IA qui apprend à partir de données | Sous-domaine du ML utilisant des réseaux de neurones profonds |
| **Approche** | Basée sur règles, logique, et/ou apprentissage | Algorithmes qui s'améliorent avec l'expérience | Réseaux de neurones à plusieurs couches |
| **Complexité des données** | Peut traiter des données simples ou structurées | Données structurées ou semi-structurées | Données complexes (images, sons, textes) |
| **Exemples d'applications** | Chatbots, systèmes experts, robots | Détection de spam, recommandation de films | Reconnaissance faciale, traduction automatique |

| | | | |
|---|---|---|---|
| **Besoin en données** | Variable selon la méthode | Nécessite un volume modéré de données | Nécessite un très grand volume de données |
| **Puissance de calcul** | Moyenne | Moyenne à élevée | Très élevée (GPU, calcul parallèle) |
| **Performance** | Dépend des règles et modèles | Bonne pour tâches spécifiques | Excellente pour tâches complexes (vision, langage) |

*Tableau I.1: Comparaison entre l'AI, DL, et ML*

## I.7. Importance de l'intelligence artificielle dans les applications médicales

L'intelligence artificielle est un domaine multidisciplinaire de l'informatique qui vise à concevoir des systèmes capables de simuler, voire de dépasser, certaines facultés cognitives humaines telles que l'apprentissage, le raisonnement, la perception et la prise de décision. Elle repose sur un ensemble de technologies telles que le traitement automatique du langage naturel, la reconnaissance vocale, la vision par ordinateur, la robotique, ainsi que l'apprentissage automatique et l'apprentissage profond. [11] On distingue généralement trois niveaux d'IA : l'IA faible, spécialisée dans des tâches précises ; l'IA forte, capable de reproduire une intelligence comparable à celle de l'humain ; et l'IA super-intelligente, encore hypothétique, censée dépasser largement les capacités humaines.

L'impact de l'IA est aujourd'hui considérable dans de nombreux secteurs, notamment celui de la santé, où elle joue un rôle central dans l'amélioration de la qualité des soins, l'accélération des diagnostics et l'optimisation des traitements. [12] En analysant automatiquement de vastes ensembles de données médicales telles que les images radiologiques, les signaux physiologiques ou les dossiers médicaux électroniques l'IA permet une prise de décision plus rapide et plus précise, souvent comparable, voire supérieure, à celle des experts humains. Elle soutient également la médecine personnalisée, en identifiant des modèles complexes associés aux profils génétiques ou aux réponses thérapeutiques des patients. Par ailleurs, elle facilite la détection précoce des maladies chroniques, améliore la gestion hospitalière et assure un suivi en temps réel des patients, participant ainsi à une médecine plus prédictive, préventive et efficiente.

**I.7.1. Application de l'intelligence artificielle dans la santé** : l'intégration de l'IA dans le domaine médical répond à des enjeux majeurs tels que l'amélioration des résultats cliniques, la réduction des coûts de santé et l'optimisation des ressources humaines. Face à la complexité croissante des pathologies et à la pression exercée sur les systèmes de santé, l'IA se présente comme une solution

innovante pour renforcer la médecine préventive. Les technologies de dépistage précoce basées sur l'IA permettent par exemple d'identifier rapidement et à moindre coût les individus à risque, facilitant ainsi des interventions ciblées et efficaces. [13-15] Au cours de la dernière décennie, les progrès de l'apprentissage automatique (ML) et de l'apprentissage profond (DL) ont profondément transformé l'intelligence artificielle médicale (AIM). Ces techniques ont permis la mise au point de modèles prédictifs avancés capables de diagnostiquer un large éventail de maladies, d'évaluer la réponse aux traitements et de personnaliser les parcours de soins. Elles contribuent également à l'automatisation des flux de travail cliniques, au suivi de l'évolution des maladies, et à la prise de décisions thérapeutiques plus éclairées, entraînant ainsi une amélioration tangible de la qualité des soins. [16-18]

Les systèmes d'IA en santé exploitent la capacité des algorithmes à traiter et interpréter de grandes quantités de données en temps réel, apportant ainsi une aide précieuse à la décision clinique, au suivi des patients et à la personnalisation des soins. [13] Les domaines d'application couvrent plusieurs spécialités médicales telles que la radiologie, la génomique, l'analyse des dossiers médicaux électroniques, ainsi que l'épidémiologie et la santé publique. [19-20]

Cependant, malgré ces avancées, l'adoption de l'IA dans les systèmes de santé nécessite un encadrement rigoureux afin de garantir la protection des données sensibles des patients et de respecter les cadres éthiques et réglementaires en vigueur. Une utilisation responsable et transparente de ces technologies est indispensable pour assurer leur efficacité clinique, leur acceptabilité sociale et leur durabilité à long terme. [18]

### I.7.2. Diabète sucré

Le diabète sucré est une maladie chronique caractérisée par une élévation anormale du taux de glucose dans le sang due à un défaut de production ou d'utilisation de l'insuline, une hormone produite par le pancréas. Cette maladie peut entraîner diverses complications lorsqu'elle n'est pas correctement prise en charge.

On distingue principalement ces types de diabète comment montre la Figure I.10

- ✓ **Diabète de type 1** : causé par une absence de production d'insuline.
- ✓ **Diabète de type 2** : forme la plus fréquente, liée à une mauvaise utilisation de l'insuline et à des facteurs tels que l'obésité et la sédentarité.
- ✓ **Diabète gestationnel** : apparaît durant la grossesse.
- ✓ **Diabète néonatal** : forme rare apparaissant dans les premiers mois de vie. [21]

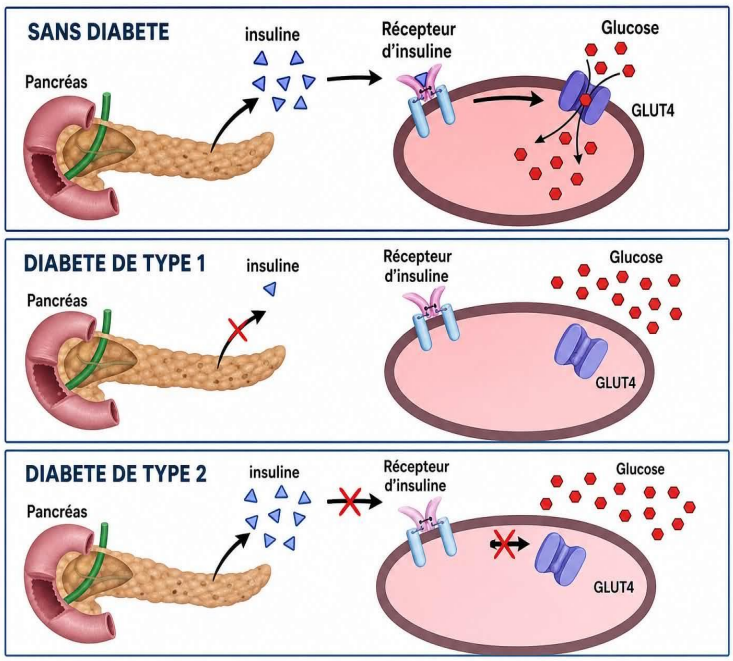


*Figure I.10: Schéma des mécanismes du diabète de type 1 et 2.*

Pour prédire si un patient est atteint de diabète ou non, cinq algorithmes différents ont été utilisés dans les modèles de prédiction en apprentissage automatique, à savoir : la régression logistique la machine à vecteurs de support *"Support Vector Machine "*, la méthode des k plus proches voisins *"K-Nearest Neighbours ",* l'arbre de décision *"Decision Tree "* et la forêt aléatoire, comme illustré dans la figure I.11.

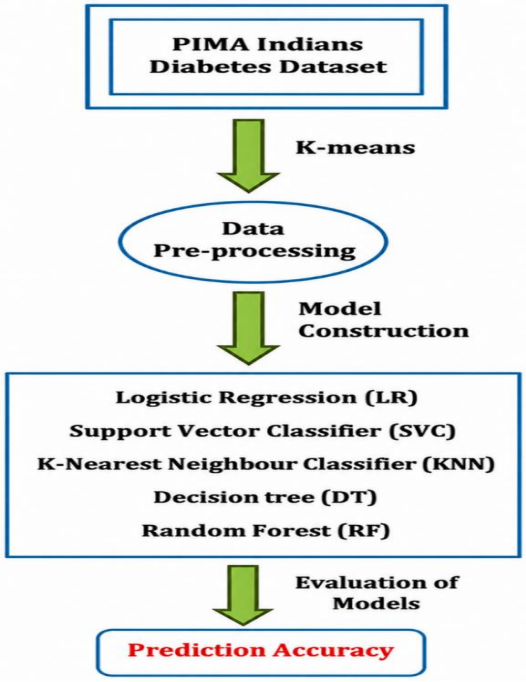


*Figure I.11: Schéma du processus de prédiction du diabète à l'aide du l'apprentissage automatique.*

### I.7.3. Cancer du pancréas

**I.7.3.1. Définition** : le pancréas est une glande située dans l'abdomen. Il contient deux types de cellules. D'une part, les **cellules exocrines** : grâce aux enzymes digestives qu'elles sécrètent, elles favorisent la digestion des aliments. D'autre part, les **cellules endocrines :** qui produisent majoritairement de l'insuline et participent ainsi à la régulation de la glycémie. En cas de **cancer du pancréas**, on a une prolifération anormale de ces cellules. Si la multiplication anarchique concerne les cellules exocrines des canaux phtèancréatiques, on a un cancer appelé **adénocarcinome canalaire du pancréas** (90 % des cas). Si cette prolifération cellulaire concerne les cellules endocrines, on a des **tumeurs neuroendocrines du pancréas**. [22]

Le cancer du pancréas est rare avant 45 ans ; son incidence augmente avec l'âge et la fréquence maximale se situe vers 75-80 ans. Les facteurs de risque sont peu établis en dehors des antécédents familiaux et du tabac. Le cancer du pancréas est souvent de diagnostic tardif : les symptômes n'apparaissent généralement que lorsque la tumeur a dépassé le pancréas ou est devenue métastatique. Son pronostic est effroyable ; c'est le 14ème cancer le plus répondu dans le monde avec plus de 216000 nouveaux cas enregistrés chaque année ; chez l'homme, il vient au quatrième rang des cancers mortels les plus courants, après les cancers du poumon, du côlon et du rectum, et de la prostate. Chez la femme, il représente la cinquième cause de décès, après les cancers du sein, colorectales, poumon et de l'utérus ou des ovaires étant plus fréquents. Il cause environ 213 000 décès chaque année. L'incidence et la mortalité des cancers du pancréas augmentent avec l'âge de façon linéaire à partir de l'âge de 45 ans (OMS, 2014).

En Algérie le cancer pancréatique fait particulièrement peur à cause de son pronostique très sombre, même si selon le registre des tumeurs d'Alger son incidence, est estimée seulement a à 2,45 /105 habitant. Le cancer du pancréas touche deux fois plus les hommes que les femmes, généralement après 50 ans. Toutes les tumeurs ne sont pas opérables, et les complications post opératoires restent importantes. Opérer un cancer du pancréas permet néanmoins d'augmenter les chances de survie à cinq ans (Registre des tumeurs d'Alger, 2012). [1] (Figure I.12)

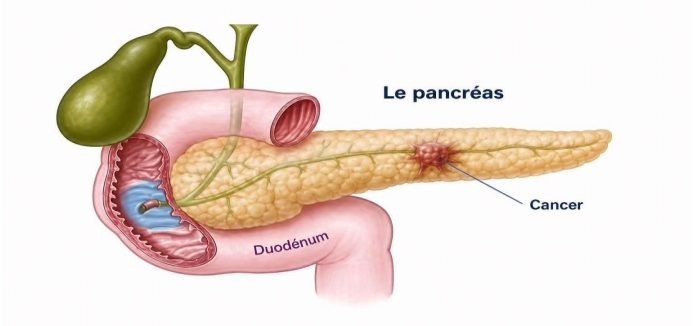


*Figure I.12: Cancer du pancréas.*

**I.7.3.2. Types du cancer du pancréas** : il existe plusieurs types de cancers du pancréas. Leur classification dépend du type de cellule à partir duquel la tumeur se développe, ce qui influence le pronostic et les options thérapeutiques.

L'adénocarcinome canalaire du pancréas est de loin le plus fréquent, représentant plus de 90 % des cas. Il prend naissance dans les cellules glandulaires qui tapissent les canaux pancréatiques. Ce type de cancer est souvent agressif et peut rapidement s'étendre aux tissus voisins, aux vaisseaux sanguins et aux organes adjacents.

Le carcinome neuroendocrine, ou tumeurs neuroendocrines du pancréas (environ 5 % des cas), se développe à partir des cellules endocrines du pancréas, responsables de la production d'hormones comme l'insuline et le glucagon. Ces tumeurs peuvent être fonctionnelles (sécrétant des hormones) ou non fonctionnelles, et leur évolution est souvent plus lente que celle des adénocarcinomes.

Il existe également d'autres formes beaucoup plus rares de cancer du pancréas, comme les tumeurs acineuses ou les cystadénocarcinomes, qui ne sont pas abordées ici en détail en raison de leur faible fréquence. [23]

### I.7.3.3. Diagnostic du cancer du pancréas

Lorsque le médecin suspecte la présence d'un cancer du pancréas, il peut faire pratiquer divers examens pour confirmer son diagnostic :

- ✓ Une échographie de l'abdomen.
- ✓ Un scanner de l'abdomen (tomodensitométrie) ou une IRM.
- ✓ Une écho endoscopie : une sonde fine est introduite par la bouche jusqu'au duodénum (la première partie de l'intestin grêle après l'estomac). Elle permet de visualiser le canal pancréatique et le canal cholédoque (par lequel la bile passe dans l'intestin). Le médecin peut profiter de cet examen pour injecter des produits dans le canal pancréatique afin d'obtenir de meilleures images par les autres modes d'examen.

Cependant, le diagnostic définitif de cancer du pancréas est confirmé par un examen au microscope d'un petit fragment de pancréas (la biopsie). Cette biopsie peut être pratiquée soit à travers la peau du ventre en s'aidant d'une échographie pour guider le geste, soit lors de l'écho endoscopie, soit lors d'une intervention chirurgicale de l'abdomen. [24]

La détection du cancer du pancréas repose sur une série de tests et d'examens car la maladie est souvent asymptomatique dans ses débuts. Les premiers signes peuvent inclure des douleurs abdominales ou dorsales, une jaunisse, une perte de poids inexpliquée et des troubles digestifs. Si ces symptômes sont

présents, le médecin peut réaliser des tests sanguins pour vérifier les niveaux de marqueurs tumoraux comme le CA 19-9**,** une protéine détectée à la surface des cellules cancéreuses.

L'imagerie médicale joue un rôle clé dans le diagnostic. Les techniques telles que la tomodensitométrie (CT scan), l'imagerie par résonance magnétique (IRM), et l'échographie endoscopique permettent de visualiser la tumeur et d'évaluer son étendue. Si une anomalie est détectée, une biopsie est généralement effectuée pour confirmer la présence de cellules cancéreuses.

Les personnes à risque élevé, comme celles ayant des antécédents familiaux ou des mutations génétiques, sont susceptibles de développer un cancer du pancréas et doivent effectuer des examens afin de détecter le cancer à un stade plus précoce.

Bien que le diagnostic précoce du cancer du pancréas soit difficile, ces méthodes permettent d'identifier la maladie, de planifier un traitement adapté et d'établir un pronostic pour le patient. [25]

### I.7.3.4. Mesures cliniques et de biomarqueurs :

✓ **Plasma CA19-9** : le CA 19-9 reste le principal marqueur tumoral sanguin utilisé. Une augmentation de son taux dans le plasma peut indiquer une récidive ou l'apparition du cancer. Il est utile pour prédire la résécabilité de la tumeur.

✓ **Creatinine** : la créatinine est un déchet métabolique produit par les muscles, éliminé par les reins. Son taux sanguin (créatininémie) est le meilleur indicateur de la fonction rénale : une augmentation signale généralement un défaut de filtration. Les taux normaux sont d'environ 60-105 µmol/L (hommes) et 45-90 µmol/L (femmes).

✓ **LYVE1** : est une glycoprotéine de membrane de type I qui agit comme récepteur principal de l'hyaluronane sur les cellules endothéliales lymphatiques (*Lymphatic Vessel Endothelial Hyaluronan Receptor 1*). Largement utilisé comme marqueur spécifique pour identifier les vaisseaux lymphatiques, il joue un rôle dans le transport de l'acide hyaluronique, la lymphangiogénèse et la métastase tumorale. Dans le cadre du cancer, son augmentation dans l'urine peut être liée à la formation de nouveaux vaisseaux lymphatiques (lymphangiogenèse) facilitant la propagation de la tumeur.

✓ **REG1B** : le gène REG1B (*Regenerating Family Member 1 Beta*) code pour une protéine sécrétée par le pancréas exocrine, impliquée dans la régénération cellulaire et la prévention de la précipitation du carbonate de calcium. Situé sur le chromosome 2p12, il est associé à la pancréatite et potentiellement à la néoangiogenèse cérébrale.

✓ **TFF1** : le Trefoil Factor 1 (TFF1), également connu sous le nom de pS2, est une protéine sécrétée appartenant à la famille des facteurs de trèfle. Elle joue un rôle essentiel dans la protection et la réparation de la muqueuse gastro-intestinale en stabilisant la couche de mucus et en favorisant la

cicatrisation de l'épithélium. La combinaison de TFF1 avec les autres marqueurs augmente la sensibilité du test pour détecter la présence de tumeurs précoces.

✓ **REG1A** : le REG1A (*Regenerating Family Member 1 Alpha*) est une protéine sécrétée codée par le gène éponyme situé sur le chromosome 2p12 chez l'humain. Les niveaux des REG1B et du REG1B augmente dans l'urine en cas d'inflammation chronique ou de cancer du pancréas.

## I.8. Conclusion

En conclusion, l'intelligence artificielle représente une avancée majeure dans le domaine de la santé. Elle contribue à améliorer la précision des diagnostics, à accélérer la prise de décision et à optimiser les traitements médicaux. Grâce à sa capacité d'analyse rapide et à son fonctionnement continu, l'IA réduit les erreurs humaines et facilite la gestion des données médicales. Cependant, malgré ses nombreux avantages, l'intelligence artificielle présente certaines limites, notamment le coût élevé des technologies, le risque de dépendance excessive aux machines et la diminution du contact humain dans les soins médicaux. Ainsi, l'IA ne doit pas être considérée comme un remplacement du personnel médical, mais plutôt comme un outil complémentaire destiné à soutenir et renforcer l'expertise humaine. L'avenir de la santé repose sur une collaboration équilibrée entre l'homme et la technologie.

# Chapitre II. Principes et Algorithmes

## II.1. Introduction

Dans ce chapitre, on présente les principales et les algorithmes d'intelligence artificielle utilisées dans les problèmes de classification, notamment dans le domaine de la santé. On commence par les différentes catégories d'apprentissage automatique (l'apprentissage supervisé, non supervisé et par renforcement). Ensuite, nous avons mentionné les algorithmes de classification couramment utilisés, tels que la régression logistique, les forêts aléatoires. Enfin, on présente les approches alternatives comme la logique floue et les systèmes experts, qui visent à simuler le raisonnement humain.

## II.2. Les applications de l'apprentissage automatique

**II.2.1. Apprentissage supervisé :** l'apprentissage supervisé, comme son nom l'indique, repose sur la supervision humaine. L'ensemble de données « étiquetées » sert à entraîner les machines à la technique d'apprentissage supervisé, puis ces dernières prédisent la sortie en se basant sur cet entraînement. Dans ce cas, les données étiquetées indiquent qu'il existe une correspondance préexistante entre les entrées et les résultats. Autrement dit, on fournit à l'ordinateur des exemples d'entrées et de sorties pendant l'entraînement, puis on lui demande d'effectuer des prédictions sur l'ensemble de données de test. [26] L'apprentissage automatique supervisé peut être classé en deux types de problèmes comme l'indique la figure II.1

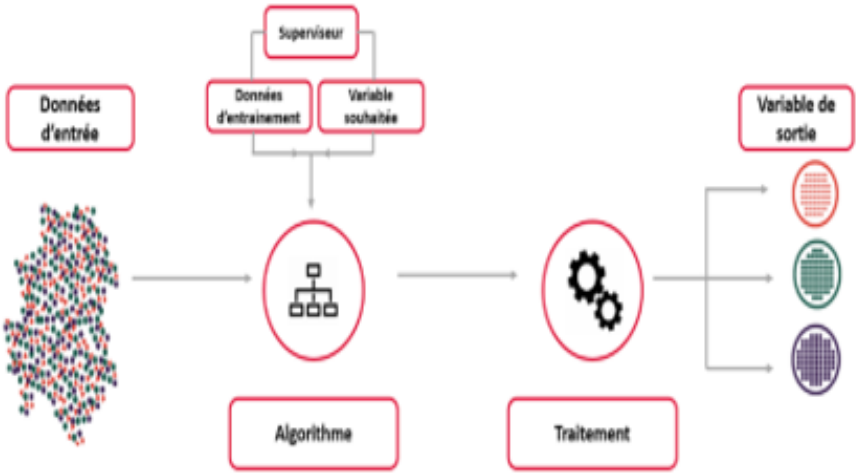


*Figure II.1: L'apprentissage supervisé.*

**II.2.1.1. Classification :** les algorithmes de classification servent à résoudre les problèmes de classification où la variable de sortie est catégorielle, comme oui ou non, homme ou femme, et rouge ou bleu. Ces algorithmes prédisent les catégories présentes dans l'ensemble de données. La détection de spams et le filtrage des courriels sont des exemples concrets d'application des algorithmes de classification. Les algorithmes de classification les plus populaires sont : les forêt aléatoires, l'arbre de

décision, régression logistique, et machine à vecteurs de support (SVM). [26]

**II.2.1.2. Régression** : il maintient l'apprentissage à partir de la fiche de données fournie et, par conséquent, fournit une sortie cohérente pour les nouvelles données qui sont introduites dans l'algorithme. Il établit un lien entre les variables dépendantes et indépendantes. Il anticipe des valeurs réelles telles que la température, l'argent, la taille et le prix. Il est utilisé pour prévoir la météo, le marché boursier et d'autres événements. [27]

Les algorithmes de régression les plus populaires sont : l'algorithme de régression linéaire simple, l'algorithme de régression multivariée, l'algorithme d'arbre de décision, et la régression Lasso. [26]

La figure II.2 montre la différence entre la classification et la régression.

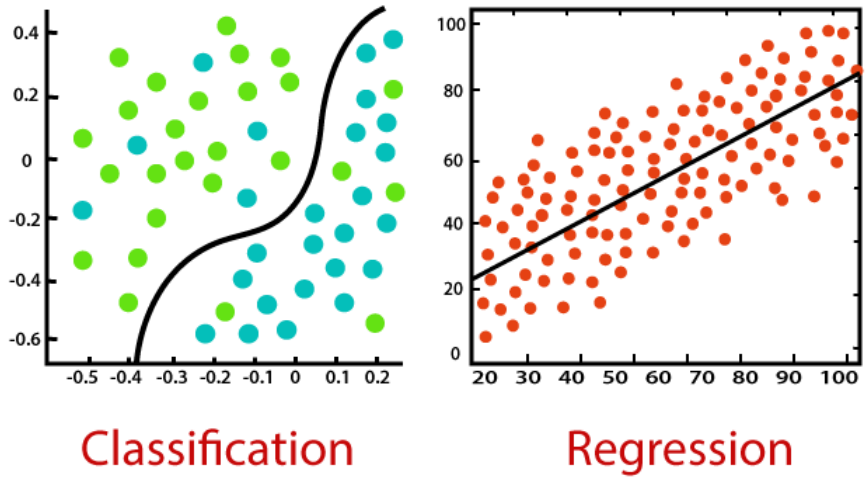


*Figure II.2: Différence entre la classification et la régression.*

**II.2.2. L'apprentissage non supervisé** : contrairement à l'apprentissage supervisé, l'apprentissage non supervisé ne repose pas sur la supervision humaine. L'apprentissage automatique non supervisé désigne un type d'apprentissage automatique où la machine est entraînée sur un ensemble de données non étiquetées, puis utilisée pour faire des prédictions sur la sortie sans aucune intervention humaine. En apprentissage non supervisé, les modèles sont entraînés sur des données qui n'ont été ni étiquetées ni catégorisées, puis ils prennent des décisions à partir de ces données sans aucune instruction extérieure. La fonction principale d'un algorithme d'apprentissage non supervisé est de classer un ensemble de données non structuré en catégories pertinentes en fonction de ses caractéristiques. L'objectif de ces systèmes automatisés est de découvrir des modèles jusqu'alors inconnus dans l'ensemble de données donné (Figure II.3). L'apprentissage non supervisé peut être classé en deux types : regroupement**,** association.

**II.2.2.1. Regroupement** : le regroupement est un outil précieux pour identifier des tendances dans les données. Cette méthode permet de regrouper les éléments de sorte que ceux qui ont le plus de points

communs restent ensemble, tandis que ceux qui en ont moins sont répartis dans d'autres groupes. Les algorithmes de regroupement sont utilisés dans de nombreux contextes, notamment pour la classification des clients selon leurs habitudes d'achat. Les algorithmes de regroupement les plus courants sont : l'algorithme de clustering K-means, l'algorithme de Mean-shift, l'algorithme DBSCAN, l'analyse en composantes principales (ACP), l'analyse en composantes indépendantes (ICA).

**II. 2.2.2. Association** : appliquée à un vaste ensemble de données, la technique d'apprentissage non supervisé des règles d'association révèle des corrélations surprenantes entre des variables jusque-là inconnues. Cet algorithme d'apprentissage vise principalement à identifier les dépendances entre les variables de données, puis à les organiser de manière à maximiser les résultats. L'analyse du panier d'achat, l'exploration des données d'utilisation du Web, la production en continu, etc., ne sont que quelques exemples des nombreuses applications de cette technique. Parmi les algorithmes d'apprentissage des règles d'association les plus populaires, on peut citer Apriori, Eclat et FP-growth.

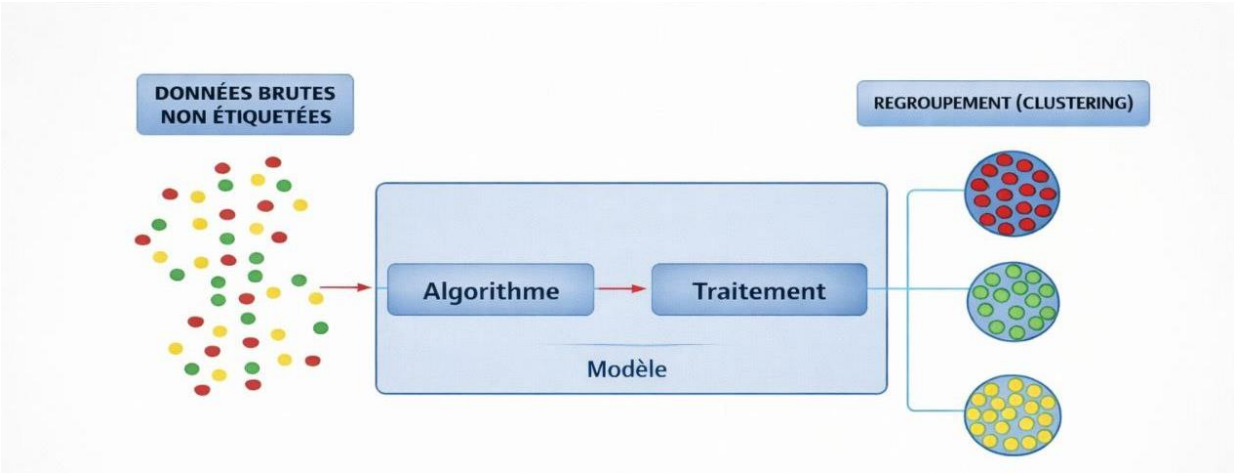


*Figure II.3: L'apprentissage non supervisé.*

**II.2.3. L'apprentissage par renforcement** : l'apprentissage par renforcement, qui utilise une technique basée sur le feedback, permet à un composant logiciel d'explorer automatiquement son environnement par une combinaison d'essais et d'erreurs, d'actions, d'apprentissage par l'expérience et d'amélioration des performances. L'objectif d'un agent d'apprentissage par renforcement est de maximiser les récompenses qu'il reçoit pour les actions souhaitables et l'évitement des actions indésirables. Ceci contraste avec l'apprentissage supervisé, qui repose sur des données étiquetées. [26] (Figure II.4)

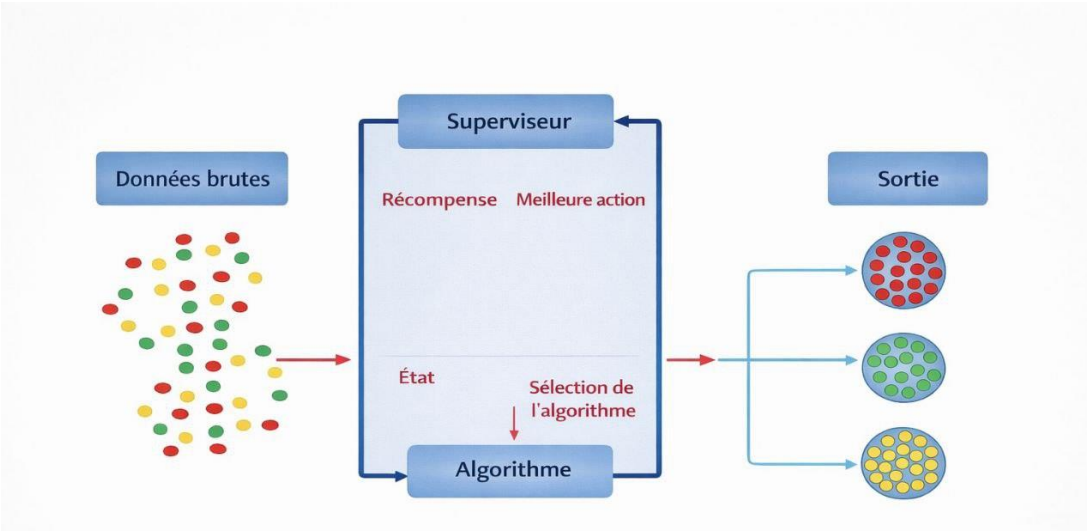


*Figure II.4: L'apprentissage par renforcement.*

# II.3. Algorithmes d'apprentissage automatique utilisés pour les problèmes de classification

## II.3.1. Régression logistique

**II.3.1.1. Définition** : l'une des algorithmes d'apprentissage automatique les plus connus, la régression logistique est un type d'apprentissage supervisé. La variable dépendante catégorique peut être prédite à partir d'un ensemble donné de facteurs indépendants avec cette méthode. L'objectif de la régression logistique est de prévoir la valeur d'une variable dépendante catégorique. Pour cette raison, la réponse doit être un nombre discret ou catégorique. Des valeurs probabilistes comprises entre 0 et 1 sont fournies au lieu des valeurs exactes de Oui et Non, 0 et 1, vrai et faux, etc. Si vous êtes familier avec la régression linéaire, vous trouverez de nombreuses similitudes entre les deux méthodes dans la régression logistique. Les problèmes de régression peuvent être résolus avec la régression linéaire, tandis que les problèmes de classification peuvent être traités avec la régression logistique. Au lieu d'une ligne droite, deux valeurs maximales sont prévues par une fonction logistique en forme de « S », qui est ajustée dans la régression logistique (0 ou 1). [26]

**II.3.1.2. Principe de fonctionnement de la régression logistique** : la régression logistique est un modèle de classification qui permet d'estimer la probabilité d'appartenance d'une observation à une classe donnée (généralement 0 ou 1).

**II.3.1.3. Représentation des données** : on considère un ensemble de variables explicatives (features) :

$$X = \begin{bmatrix} X11 & \cdots & X1m \\ \vdots & \ddots & \vdots \\ Xn1 & \cdots & Xnm \end{bmatrix} \qquad (II.1)$$

Et une variable cible binaire :

$y$ = [0 Si l'observation appartient à la classe 0,

1 si l'observation appartient à la classe 1] **(*II*. 2)**

Dans un premier temps, le modèle calcule une combinaison linéaire des variables d'entrée :

$$z = (\sum_{i=1}^{n} \omega_i \, x_i) + b \qquad \textbf{(}\mathit{II}\textbf{. 3)}$$

Où : $\boldsymbol{\omega_i}$: poids associés aux variables, $\boldsymbol{x_i}$ : variables explicatives, $\boldsymbol{b}$ : biais (intercept)

Cette expression peut s'écrire sous forme vectorielle :

$z = \omega . X + b$ **(*II*. 4)**

Pour transformer la valeur $z$ (qui peut être n'importe quel réel) en une **probabilité entre 0 et 1**, on applique la fonction sigmoïde :

$\sigma(z) = \frac{1}{1+e^{-z}}$ **(*II*. 5)**

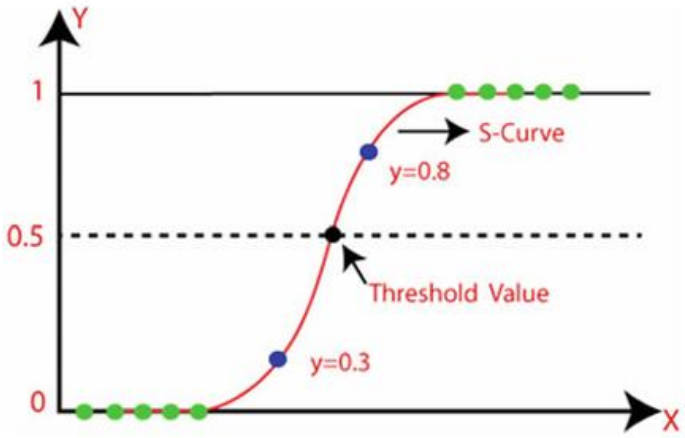


***Figure II.5: La fonction logistique.***

Comme le montre la figure 5 ci-dessus, Cette fonction possède les propriétés suivantes :

$\sigma(z) \rightarrow 1$ lorsque $z \rightarrow +\infty$

$\sigma(z) \rightarrow 0$ lorsque $z \rightarrow -\infty$

$0 \leq \sigma(z) \leq 1$

La sortie du modèle correspond à une probabilité :

$P(y = 1) = \sigma(z)$ **(II.6)**

$P(y = 0) = 1 - \sigma(z)$ **(II.7)**

On définit le rapport de chances comme :

$$\frac{p(x)}{1-p(x)} = e^{z} \quad \textbf{(II.8)}$$

En appliquant le logarithme naturel :

$$\log[\frac{p(x)}{1-p(x)}] = z \quad \textbf{(II.9)}$$

En remplaçant z :

$$\log\left[\frac{p(x)}{1-p(x)}\right] = \omega.X + b \quad \textbf{(II.10)}$$

$$\left[\frac{p(x)}{1-p(x)}\right] = e^{\omega.X+b} \quad \textbf{(II.11)}$$

$$p(x) = e^{\omega \cdot X^{+}b}.\,(1 - p(x)) \quad \textbf{(II.13)}$$

$$p(x) = e^{\omega \cdot X^{+}b} - e^{\omega \cdot X^{+}b}.\,p(x) \quad \textbf{(II.14)}$$

$$p(x) + e^{\omega \cdot X^{+}b}.\,p(x) = e^{\omega \cdot X^{+}b} \quad \textbf{(II.15)}$$

$$p(x)(1 + e^{\omega \cdot X^{+}b}) = e^{\omega \cdot X^{+}b} \quad \textbf{(II.16)}$$

$$p(x) = \frac{e^{\omega.X+b}}{1+e^{\omega.X+b}} \quad \textbf{(II.17)}$$

La formule finale : $p(x; b; \omega) = \frac{e^{\omega.X+b}}{1+e^{\omega.X+b}} = \frac{1}{1+e^{\omega.X+b}}$ **(II.18)**

**II.3.1.4. Types de régression logistique** : selon les catégories, la régression logistique peut être classée en trois types :

✓ **Binomiale :** La régression logistique binomiale ne comporte que deux valeurs possibles pour la variable dépendante, par exemple 0 ou 1, réussite ou échec, etc.

✓ **Multinomiale :** La régression logistique multinomiale comporte trois valeurs possibles ou plus pour la variable dépendante, sans ordre particulier, par exemple « chat », « chien » ou « mouton ».

✓ **Ordinale** : La régression logistique ordinale comporte trois valeurs possibles ou plus pour la variable dépendante, avec un ordre particulier, par exemple « faible », « moyen » ou « élevé ». [26]

**II.3.1.5. Hypothèses de la régression logistique** : les principales hypothèses sont les suivantes :

✓ **Indépendance des observations** : Chaque point de données est supposé indépendant des autres, ce qui signifie qu'il ne doit exister aucune corrélation ni dépendance entre les échantillons d'entrée.

✓ **Variables dépendantes binaires** : La variable dépendante est supposée binaire, c'est-à-dire qu'elle ne peut prendre que deux valeurs. Pour plus de deux catégories, la fonction SoftMax est utilisée.

✓ **Relation linéaire entre les variables indépendantes et le logarithme du rapport de cotes :** Le modèle suppose une relation linéaire entre les variables indépendantes et le logarithme du rapport

de côtes de la variable dépendante, ce qui signifie que les prédicteurs influencent le logarithme du rapport de côtes de manière linéaire.

✓ **Absence de valeurs aberrantes :** L'ensemble de données ne doit pas contenir de valeurs aberrantes extrêmes, car celles-ci peuvent fausser l'estimation des coefficients de régression logistique.

✓ **Taille d'échantillon suffisamment grande :** Un échantillon suffisamment grand est nécessaire pour produire des résultats fiables et stables. [28]

### II.3.2. Forêts aléatoires

**II.3.2.1. Définition** : RF est un algorithme d'apprentissage automatique qui utilise de nombreux arbres de décision pour améliorer la précision des prédictions. Chaque arbre analyse différentes parties aléatoires des données, et leurs résultats sont combinés par vote pour la classification ou par moyenne pour la régression, ce qui en fait une technique d'apprentissage d'ensemble. Cela permet d'améliorer la précision et de réduire les erreurs. [29]

#### II.3.2.2. Caractéristiques clés de la forêt aléatoire

✓ **Gestion des données manquantes** : Elle fonctionne même en présence de données manquantes, vous évitant ainsi de devoir les imputer manuellement.

✓ **Importance des variables :** Elle indique les variables (colonnes) les plus pertinentes pour les prédictions, facilitant ainsi la compréhension de vos données.

✓ **Performance avec les données volumineuses et complexes :** Elle gère efficacement les grands ensembles de données comportant de nombreuses variables, sans ralentissement ni perte de précision.

✓ **Utilisation pour diverses tâches** : Elle peut être utilisée pour la classification (prédiction de types ou d'étiquettes) et la régression (prédiction de nombres ou de quantités). [29]

#### II.3.2.3. Hypothèses de la forêt aléatoire

✓ **Chaque arbre prend ses propres décisions :** Chaque arbre de la forêt effectue ses propres prédictions sans dépendre des autres.

✓ **Utilisation aléatoire des données :** Chaque arbre est construit à partir d'échantillons et de caractéristiques aléatoires afin de réduire les erreurs.

✓ **Nombre suffisant de données :** Un volume de données suffisant garantit que les arbres sont différents et apprennent des modèles uniques et variés.

✓ **Des prédictions différentes améliorent la précision :** La combinaison des prédictions de différents arbres permet d'obtenir un résultat final plus précis. [29]

### II.3.2.4. Principe de fonctionnement de l'algorithme des forêts aléatoires

✓ **Création de nombreux arbres de décision :** l'algorithme crée de nombreux arbres de décision, chacun utilisant une partie aléatoire des données. Chaque arbre est donc légèrement différent.

✓ **Sélection aléatoire des caractéristiques :** Lors de la construction de chaque arbre, toutes les caractéristiques (colonnes) ne sont pas examinées simultanément. Quelques-unes sont sélectionnées aléatoirement pour déterminer comment diviser les données. Cela permet aux arbres de rester différents les uns des autres.

✓ **Prédiction de chaque arbre :** Chaque arbre fournit sa propre réponse ou prédiction en fonction de ce qu'il a appris de sa partie des données.

✓ **Combinaison des prédictions :** Pour la classification, une catégorie est choisie, la réponse finale étant celle sur laquelle la plupart des arbres s'accordent (vote majoritaire). Pour la régression, une valeur numérique est prédite, la réponse finale étant la moyenne des prédictions de tous les arbres. [29] (Figure II.6)

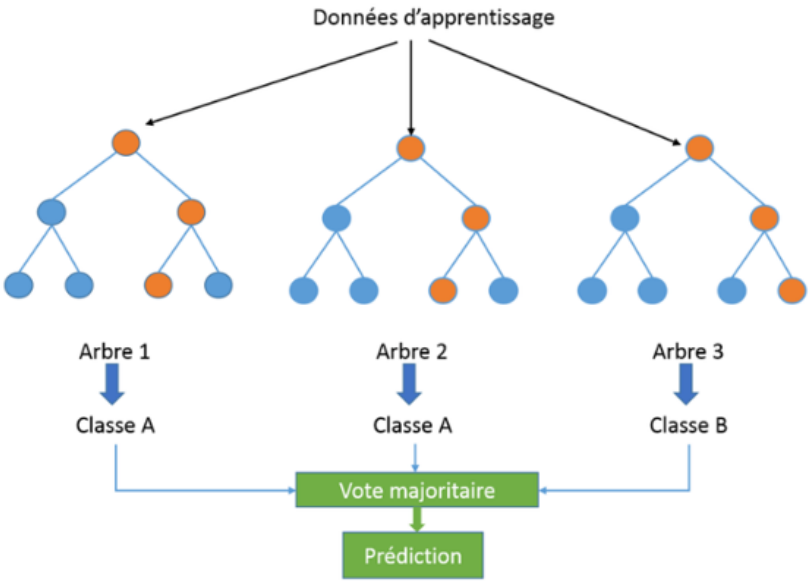


*Figure II.6: La forêt aléatoire.*

## II.4. Approches d'intelligence artificielle utilisées pour les problèmes de classification

### II.4.1. Logique floue

#### II.4.1.1. Architecture

✓ **Base de règles :** Elle contient l'ensemble des règles et les conditions SI-ALORS fournies par les experts pour gouverner le système de prise de décision, sur la base des informations linguistiques. Les développements récents de la théorie floue offrent plusieurs méthodes efficaces pour la conception et

le réglage des contrôleurs flous. La plupart de ces développements réduisent le nombre de règles floues.

✓ **Fuzzification :** Elle est utilisée pour convertir les entrées, c'est-à-dire les nombres précis, en ensembles flous. Les entrées précises sont essentiellement les entrées exactes mesurées par les capteurs et transmises au système de contrôle pour traitement, telles que la température, la pression, le nombre de tours par minute, etc.

✓ **Moteur d'inférence :** Il détermine le degré de correspondance de l'entrée floue actuelle par rapport à chaque règle et décide quelles règles doivent être activées en fonction du champ d'entrée. Ensuite, les règles activées sont combinées pour former les actions de contrôle.

✓ **Défuzzification :** Elle est utilisée pour convertir les ensembles flous obtenus par le moteur d'inférence en une valeur précise. Plusieurs méthodes de défuzzification sont disponibles et la plus appropriée est utilisée avec un système expert spécifique pour réduire l'erreur. [30]

**II.4.1.2. Les applications du floue logique** : il est utilisé dans le domaine aérospatial pour le contrôle de l'altitude des engins spatiaux et des satellites, dans le système automobile pour le contrôle de la vitesse et le contrôle du trafic, pour les systèmes de support à la prise de décision et l'évaluation personnelle dans les grandes entreprises. Il a des applications dans l'industrie chimique pour le contrôle du pH, le séchage et le processus de distillation chimique. La logique floue est utilisée dans le traitement du langage naturel et diverses applications intensives en intelligence artificielle. La logique floue est largement utilisée dans les systèmes de contrôle modernes tels que les systèmes experts. [45]

### II.4.2. Système expert

**II.4.2.1. Principe d'un système expert** : un système expert remplace un expert humain d'où sort un outil de travail et d'aide dans différents domaines tels que : le diagnostic (médical ou technique), la prévision (entreprise), la classification, le dépannage, etc. Ils présentent un intérêt réel pour les milieux professionnels, car ils permettent l'informatisation de certaines fonctions intellectuelles qualifiées, difficiles à modéliser sous forme d'algorithmes sûrs et définitifs. Ces fonctions concernent l'identification ou le diagnostic de situation, la prévision des évènements, la conception d'objets, la planification d'actions, etc. dans l'activité des entreprises. (Figure II.7)

**II.4.2.2. Objectif des systèmes experts** : le principal objectif des systèmes experts est de mettre en mémoire les connaissances théoriques de l'expert, de façon à ce qu'elles soient toujours disponibles et aussi de pouvoir retraduire les connaissances sous formes d'aide au diagnostic, c'est une sorte de questions / réponses avec l'utilisateur. [31]

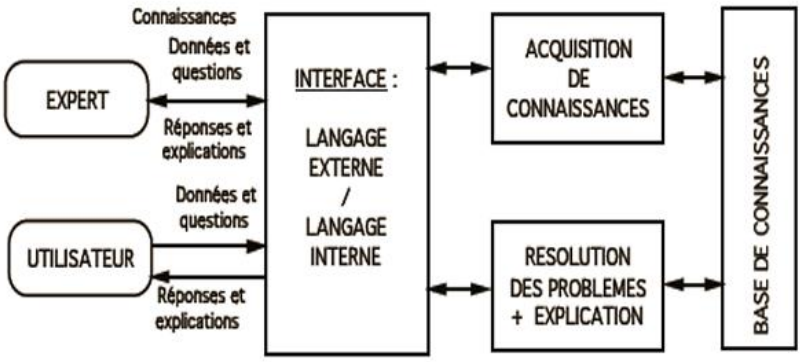


*Figure II.7: Présentation d'un système expert.*

**II.4.2.3. Principaux composants d'un système expert** : un système expert est principalement composé de (figure II.8) :

- ✓ Une base de connaissances
  - o Une base de faits
  - o Une base de règles
- ✓ Un moteur d'inférence

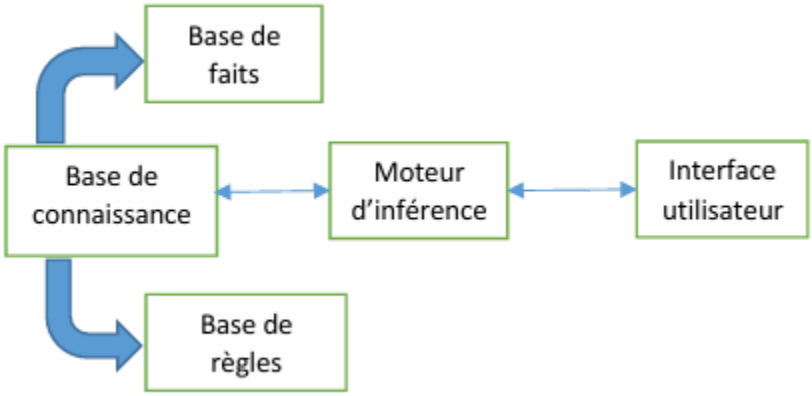


*Figure II.8: Les composants d'un système expert.*

**II. 4.2.4. Stratégie de raisonnement** : appelé également démonstrateur de théorème, le moteur d'inférence est chargé d'appeler et d'utiliser les informations contenues dans la base de connaissances, correctement, pour répondre à une question ou résoudre un problème. Il effectue des déductions grâce à deux principales stratégies de raisonnement qui sont : [31]

**a) Stratégie relative au chaînage :**

- ✓ **Chaînage avant :** Son principe est de déterminer une réponse à partir des faits constatés, contenus dans la base de faits, en utilisant les règles de déductions de la base de connaissance et en déduisant des conclusions à partir des prémisses vérifiées
- ✓ **Chaînage arrière :** Son principe consiste, à partir d'une question posée appelée but, de remonter jusqu'aux faits de la base de connaissance, grâce aux règles.
- ✓ **Chaînage mixte :** Comme son nom l'indique, cette approche utilise la combinaison des deux approches en alliant leurs avantages pour essayer de pallier leurs inconvénients.

**b) Stratégie relative au parcours :**

- ✓ Parcours en largeur d'abord.
- ✓ Parcours en profondeur d'abord.

### II.4.2.5. Fonctionnement général d'un système Expert

Le fonctionnement général d'un système expert se résume en ces points :

- ✓ Chercher à prouver le 1er diagnostic de la liste.
- ✓ Chercher la/les règles qui ont ce diagnostic en conclusion (chaînage arrière).
- ✓ Chercher à prouver les prémisses :
  - o Par des faits initiaux (s'ils existent),
  - o Par des déductions (enchaînées),
  - o Par des questions.
- ✓ Si le 1er diagnostic échoue, nous passons au deuxième et puis aux suivants
- ✓ Dès qu'un diagnostic est prouvé, nous affichons : <SUCCÈS et proposer la conclusion>
- ✓ Si aucun diagnostic ne peut être prouvé, nous affichons : <ECHEC>. [31]

## II.5. Conclusion

Ce chapitre met en évidence l'importance des algorithmes et leurs avantages variés. L'apprentissage automatique offre précision et automatisation, tandis que les systèmes flous et les systèmes experts assurent une meilleure transparence et explicabilité. Ainsi, dans le domaine du diagnostic médical, il est essentiel de comprendre leur fonctionnement, leur structure et les hypothèses sur lesquelles ils reposent.

# Chapitre III : prédiction et classification du cancer du pancréas

## III.1. Introduction

Ce chapitre est consacré à l'implémentation des codes sources pour la prédiction et la classification du cancer du pancréas en se basant sur les deux méthodes d'apprentissage automatique du forêt aléatoire "*Radom Forest*", et régression logistique "*Logistic Regression*", et la méthode de l'intelligence artificielle de haut de la logique floue *"fuzzy logic"*.

L'implémentation de ces différents codes source est basée sur une base de données agrée et utilisé par les praticiens ainsi que les chercheurs pour étudier ce type de cancer. Cette implémentation est achevée en utilisant le Python, et elle inclue cinq phases principales : l'importation des bibliothèques et modules nécessaires, l'exploitation de données, l'analyse de données, application de l'algorithme envisagé, et l'exploitation et la comparaison de différents résultats de prédication de classifications obtenus.

A la fin, une conclusion est donnée pour résumer les principaux travaux effectués dans ce chapitre.

## III.2. L'importation des bibliothèques et modules nécessaires

L'étude et l'application des algorithmes de l'intelligence artificielle faites pour la prédiction du cancer pancréatique sont basées sur l'utilisation d'une base donnée [32] de référence. Cette base de données est ensemble de données offre une collection de caractéristiques de patients et de marqueurs biologiques pertinents pour les études pancréatiques. Elle contient des données pour 590 patients où chaque patient est identifié par un identificateur "Sample_ID", un groupe d'étude "Patient_Cohort" et la source fournissant des données "Sample_origin". Pour chaque patient, les informations démographiques de l'âge "Age" et le sexe "Sex" sont fournies. Une évaluation clinique est incluse pour préciser l'état sanitaire du pancréas du patient "Diagnosis". La basse de données inclus les marqueurs cliniques et les mesures de biomarqueurs protéiques détectables dans l'urine indispensable pour le diagnostic pancréatique. Ces marqueurs cliniques sont le plasma CA19-9 "Plasma_CA19_9" et du taux du creatinine "Creatinine" tandis que les mesures de biomarqueurs sont la glycoprotéine de membrane de type I "LYVE1", les m**embre 1 alpha/bêta de la famille des protéines régénératrices des îlots** " REG1B" et "REG1A" et le **Trefoil Factor 1** "TFF1".

Pour l'état sanitaire du patient, l'évaluation "Diagnosis" peut indiquer : en bon santé "Healthy", le type de cancer du pancréas le plus courant et le plus agressif **"PDAC " (*Pancreatic Ductal AdenoCarcinoma*)**, ou les néoplasies pancréatiques "NON-**PDAC" (*non-pancreatic ductal adenocarcinoma*)** représentent les types de tumeurs restants, moins fréquents.

## III.3. Exploitation de données

**III.3.1. Âge** : l'âge est l'un des facteurs de risque les plus importants pour le cancer du pancréas. Pour la base de données utilisé et comme le montre la figure III.1, la majorité des patients se situent entre 50 et 70 ans, ce qui correspond à la tranche d'âge la plus exposée au cancer du pancréas. Les cas en dessous de 40 ans ou au-delà de 80 ans sont rares, confirmant que la maladie touche principalement les adultes et les personnes âgées.

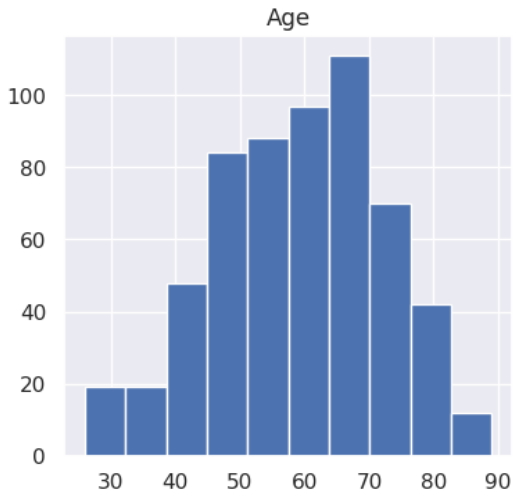


*Figure III.1 : Âge des patients de la base de données.*

**III.3.2. Plasma_C19_9** : le CA 19-9 est un marqueur tumoral biologique utilisé pour le suivi des patients atteints de cancer du pancréas. Il s'agit d'une molécule (antigène), portée à la surface de cellules cancéreuses et libérée dans le sang, qui peut être dosée dans un simple test sanguin. Un taux élevé de CA 19-9 peut indiquer une activité tumorale accrue, mais il ne s'agit pas d'un test de dépistage initial. Le CA 19-9 est particulièrement utilisé pour le suivi post-diagnostic, l'évaluation de l'efficacité des traitements, la détection de récidives et la surveillance de la progression tumorale.

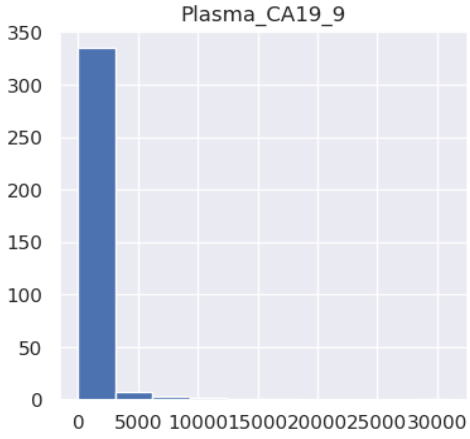


*Figure III.2 : Distribution du CA19-9.*

D'après la figure III.2, la distribution est asymétrique vers la droite : la majorité des patients ont des valeurs faibles, mais quelques-uns présentent des taux très élevés. Cela reflète le rôle de CA19-9 comme marqueur tumoral : bas chez les sujets sains, élevé en cas de tumeur maligne.

**III.3.3. Créatinine** : la créatinine n'est pas un marqueur tumoral direct, mais elle est **indispensable pour interpréter correctement les biomarqueurs urinaires** (LYVE1, REG1A, REG1B, TFF1). En cas d'insuffisance **rénale** (créatinine élevée), l'élimination des biomarqueurs peut être perturbée, faussant leur concentration dans l'urine.

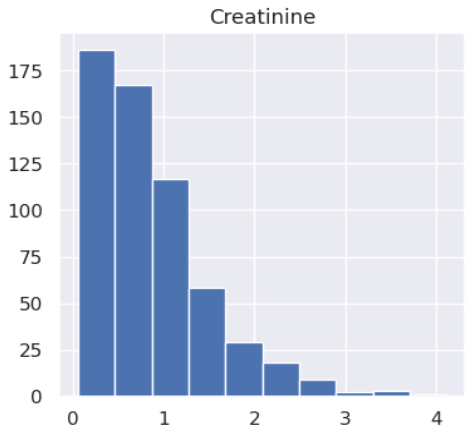


*Figure III.3 : Taux de créatinine.*

La plupart des valeurs se situent dans la plage normale (0,5–1,5 mg/dL). Les valeurs élevées sont rares et peuvent indiquer une altération de la fonction rénale. Cette variable agit comme un facteur de contrôle pour interpréter correctement les biomarqueurs urinaires.

**III.3.4. LYVE1** : LYVE1 **(***Lymphatic Vessel Endothelial Hyaluronan Receptor 1*) est une glycoprotéine membranaire utilisée comme indicateur de la formation de nouveaux vaisseaux lymphatiques (lymphangiogenèse), un processus clé dans la propagation du cancer du pancréas. Son augmentation dans l'urine est fortement corrélée aux cas de PDAC.

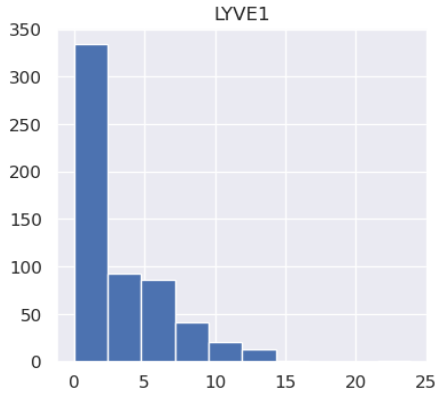


*Figure III.4 : Concentrations urinaires de LYVE1.*

La distribution est également asymétrique : la majorité des patients ont des niveaux faibles, mais certains présentent des valeurs très élevées. L'augmentation de **LYVE1** est souvent associée à la **lymphangiogenèse**, processus favorisant la propagation tumorale.

**III.3.5. REG1B** : REG1B (*Regenerating Family Member 1 Beta*) est une protéine sécrétée par le pancréas exocrine. Son rôle physiologique principal est lié à la régénération cellulaire et à la protection des tissus pancréatiques. En cas de **pancréatite chronique** ou de **cancer du pancréas (PDAC),** ses niveaux augmentent de manière significative ; Cette élévation est interprétée comme un **signal biologique d'activité tumorale ou inflammatoire.** REG1B est donc considéré comme un **biomarqueur urinaire** utile pour distinguer les patients sains des patients atteints de cancer.

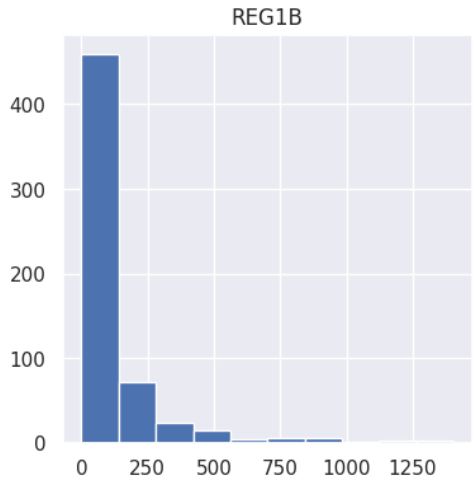


***Figure III.5 : Niveaux urinaires de REG1B.***

**III.3.6. TFF1** : TFF1 joue un rôle essentiel dans la protection et la réparation de la muqueuse gastro-intestinale. Elle stabilise la couche de mucus et favorise la cicatrisation de l'épithélium. En cancérologie, l'augmentation de TFF1 est observée dans les tumeurs précoces. L'histogramme dans la figure montre que la majorité des patients ont des valeurs faibles, mais certains présentent des pics élevés. Ces cas correspondent souvent à des patients atteints de tumeurs.

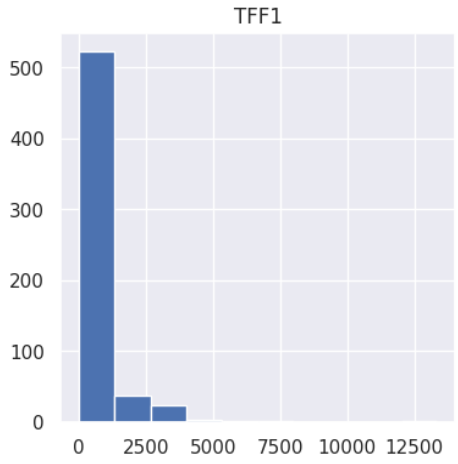


*Figure III.6 : Concentrations urinaires de TFF1.*

**III.3.7. REG1A** : c'est une protéine sécrétée par le pancréas, impliquée dans la régénération cellulaire et la protection des tissus pancréatiques. Elle participe aux mécanismes de réparation en cas de lésion ou d'inflammation. Ses niveaux augmentent dans l'urine en cas de pancréatite chronique ou de cancer du pancréas (PDAC). L'histogramme montre une distribution asymétrique, avec la majorité des patients ayant des valeurs faibles et quelques cas présentant des niveaux très élevés.

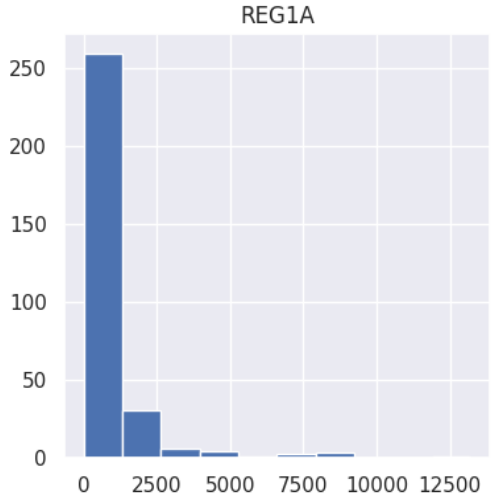


*Figure III.7 : Niveaux urinaires de REG1A.*

**III.3.8. Diagnostic** : on a trois cas :

- ✓ **Healthy** : en bon santé ;
- ✓ **PDAC :** le type de cancer du pancréas le plus courant et le plus agressif ;
- ✓ **NON-PDAC** : représentent les types de tumeurs restants, moins fréquents.

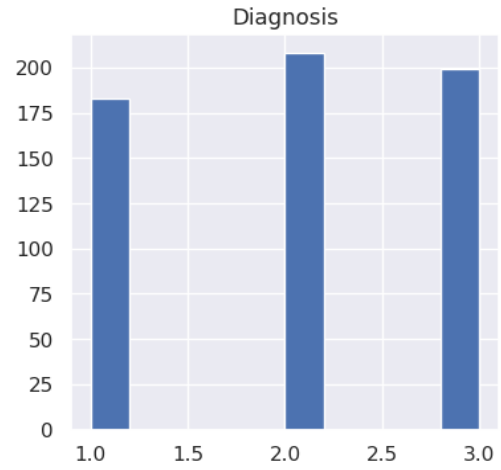


*Figure III.8 : Répartition des diagnostics (Healthy, PDAC, NON-PDAC).*

## III.4. Analyse des données

**III.4.1. Analyse comparative de la créatinine entre les groupes diagnostiques** : chez les personnes en bonne santé, la majorité des valeurs de la créatinine se situent approximativement entre 0,4 et 1, avec une distribution centrée sur des valeurs normales et sans présence de valeurs très élevées, ce qui reflète un fonctionnement rénal normal. En revanche, dans le groupe Non-PDAC, les valeurs de la créatinine montrent une légère augmentation par rapport au groupe sain ainsi qu'une plus grande diversité des valeurs, ce qui peut être expliqué par l'influence d'autres maladies susceptibles d'affecter les reins. Concernant le groupe PDAC, la courbe apparaît plus longue et plus étendue, avec certains patients présentant des taux de créatinine très élevés. Globalement, on observe une différence notable dans la distribution de la créatinine entre les groupes, caractérisée par une augmentation plus importante chez les patients atteints d'un adénocarcinome pancréatique, suggérant la présence de troubles métaboliques ou rénaux associés au cancer du pancréas.

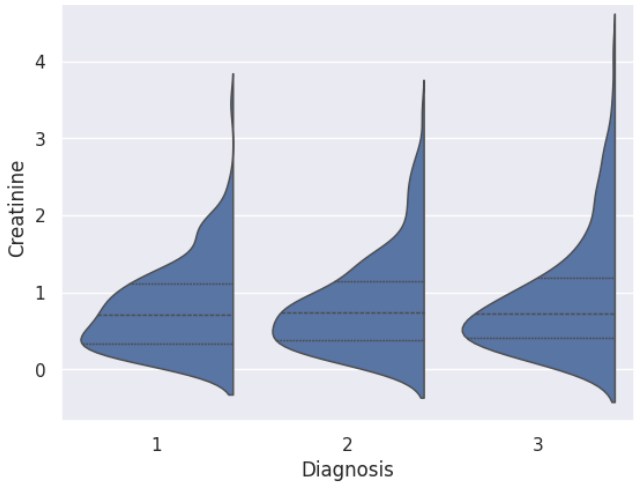


*Figure III.9 : Visualisation des distributions de la créatinine.*

**III.4.2. Analyse comparative de LYVE1 entre les groupes diagnostiques** : chez les personnes en bonne santé, les valeurs de LYVE1 sont très faibles et proches de zéro, avec une distribution étroite, ce qui indique un faible taux de ce biomarqueur dans des conditions normales. Dans le groupe Non-PDAC, une légère augmentation des valeurs est observée par rapport aux sujets sains, et certains patients présentent des valeurs moyennes, probablement en raison d'infections ou d'autres maladies pouvant influencer légèrement le taux de LYVE1. En revanche, chez les patients atteints de PDAC, l'augmentation est très marquée : la courbe est plus longue, atteint des valeurs très élevées et présente une forte densité dans les zones élevées. Ces résultats montrent que les niveaux de LYVE1 sont significativement plus élevés chez les patients atteints d'un adénocarcinome pancréatique, ce qui suggère son potentiel en tant que biomarqueur pour la détection du cancer du pancréas.

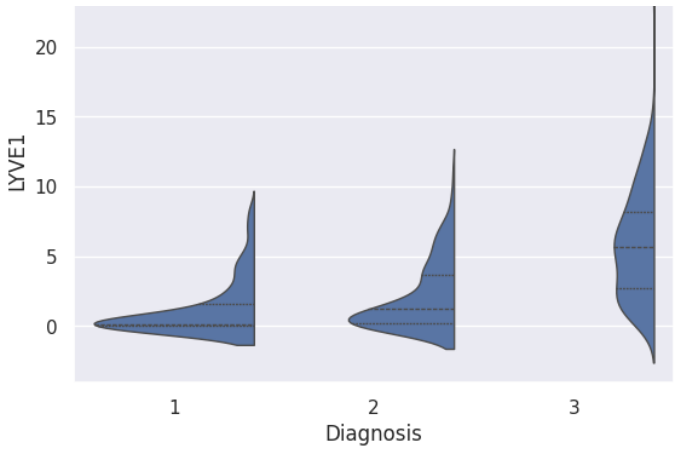


*Figure III.10 : Visualisation des distributions de LYVE1.*

**III.4.3. Analyse comparative de REG1B entre les groupes diagnostiques** : chez les personnes en bonne santé, la plupart des valeurs de REG1B sont très faibles et la distribution reste proche de zéro, ce qui correspond à une situation physiologique normale. Dans le groupe Non-PDAC, une élévation modérée des valeurs est observée, avec certains patients présentant des niveaux plus élevés, ce qui peut être lié à d'autres maladies pancréatiques influençant le taux de REG1B. En revanche, chez les patients atteints de PDAC, l'augmentation est très importante par rapport aux autres groupes : les valeurs atteignent des niveaux très élevés et montrent une large dispersion. Ces résultats indiquent que le gène REG1B est fortement associé au cancer du pancréas. De plus, les niveaux les plus élevés de REG1B ont été observés chez les patients atteints de PDAC, accompagnés d'une variabilité importante des valeurs, ce qui souligne son potentiel en tant que biomarqueur efficace pour la détection du cancer du pancréas.

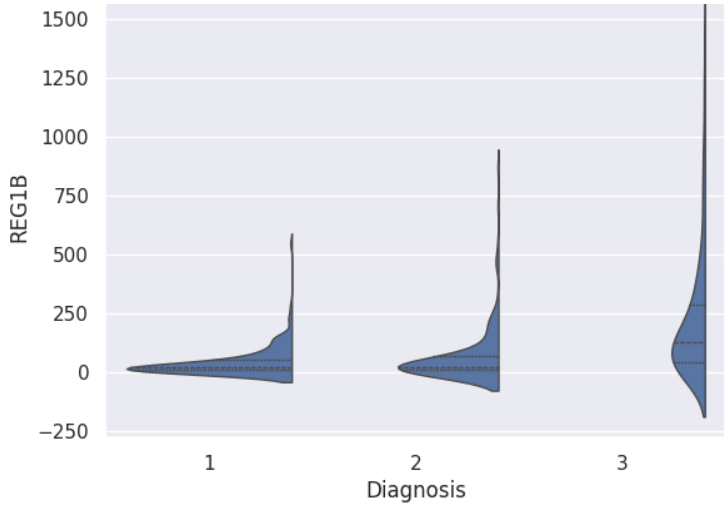


*Figure III.11 : Visualisation des distributions de REG1B.*

## III.5. Prétraitement et application des algorithmes envisagés

**III.5.1. Prétraitement** : avant d'entraîner un modèle de prédiction appliqué au cancer du pancréas, il est essentiel de préparer le jeu de données par une étape appelée **prétraitement**. Cette phase garantit que les informations cliniques et biologiques sont propres, cohérentes et correctement formatées afin que l'algorithme puisse apprendre de manière efficace. Dans le cadre des programmes de prédiction du cancer pancréatique présentés dans ce chapitre, le prétraitement comprend plusieurs tâches cruciales : l'identification et la gestion des valeurs manquantes, la normalisation des biomarqueurs sanguins et urinaires pour éviter les biais liés aux différences d'échelle, l'encodage des variables catégorielles telles que le sexe ou l'origine des échantillons, ainsi que la séparation des données en ensembles d'entraînement et de test.

**III.5.1.1. Identification des valeurs manquantes** : lors de l'exploration initiale du jeu de données, certaines colonnes présentaient des valeurs absentes (*NaN*), en particulier les biomarqueurs **Plasma_CA19-9** et **REG1A**. Leur présence pose un problème majeur, car les algorithmes d'apprentissage automatique ne peuvent pas traiter directement des valeurs manquantes.

L'identification s'effectue en parcourant chaque variable et en comptabilisant le nombre de valeurs absentes. Cette opération permet de dresser une cartographie précise des colonnes affectées et d'évaluer l'ampleur du problème.

**III.5.1.2. Interpolation des données** : après l'identification des valeurs absentes dans les colonnes **Plasma_CA19-9** et **REG1A**, une procédure d'**imputation par interpolation** a été appliquée. Cette méthode consiste à estimer les valeurs manquantes en se basant sur les observations voisines dans la série de données. Concrètement, lorsqu'une valeur était absente pour un patient, elle a été remplacée par une estimation calculée à partir des valeurs précédentes et suivantes, assurant ainsi une continuité

numérique. Par exemple, pour certains patients dont les mesures de **REG1A** étaient manquantes, l'interpolation a généré des valeurs intermédiaires cohérentes (≈ 253.60 ou ≈ 404.78), permettant de conserver l'intégralité des observations. De même, les valeurs manquantes de **Plasma_CA19-9** ont été remplacées par des estimations progressives (≈ 9.35 ou ≈ 17.50), garantissant une distribution continue de ce biomarqueur

**III.5.1.3. Encodage des variables catégorielles** : certaines variables sont de nature catégorielle comme le sexe du patient, la cohorte d'appartenance ou l'origine de l'échantillon. Ces variables ne peuvent pas être directement interprétées par les algorithmes d'apprentissage automatique, qui nécessitent des entrées numériques. L'étape d'encodage consiste donc à transformer ces catégories en valeurs numériques.

- ✓ **Sexe → M = 1, F = 0 ;**
- ✓ **Cohorte → Cohort1 = 1, Cohort2 = 0 ;**
- ✓ **Origine de l'échantillon → (BPTB = 0, LIV = 1, ESP = 2, UCL = 3)**

## III.6. Application de l'algorithme des forêts aléatoires et construction du modèle

**III.6.1. Application et construction :** la figure 12 présente les variables explicatives utilisées dans l'étude pour entraîner le modèle de régression logistique. Ces variables regroupent plusieurs informations médicales et biologiques des patients, telles que l'âge, le sexe, le taux de CA19-9, la créatinine ainsi que différents biomarqueurs associés au cancer du pancréas.

| | Patient_Cohort | Sample_Origin | Age | Sex | Plasma_CA19_9 | Creatinine | LYVE1 | REG1B | TFF1 | REG1A |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 33 | 0 | 11.700000 | 1.83222 | 0.893219 | 52.94884 | 654.282174 | 1262.000000 |
| 1 | 1 | 0 | 81 | 0 | 9.350000 | 0.97266 | 2.037585 | 94.46703 | 209.488250 | 228.407000 |
| 2 | 0 | 0 | 51 | 1 | 7.000000 | 0.78039 | 0.145589 | 102.36600 | 461.141000 | 253.603727 |
| 3 | 0 | 0 | 61 | 1 | 8.000000 | 0.70122 | 0.002805 | 60.57900 | 142.950000 | 278.800455 |
| 4 | 0 | 0 | 62 | 1 | 9.000000 | 0.21489 | 0.000860 | 65.54000 | 41.088000 | 303.997182 |
| 5 | 0 | 0 | 53 | 1 | 9.666667 | 0.84825 | 0.003393 | 62.12600 | 59.793000 | 329.193909 |
| 6 | 0 | 0 | 70 | 1 | 10.333333 | 0.62205 | 0.174381 | 152.27700 | 117.516000 | 354.390636 |
| 7 | 0 | 0 | 58 | 0 | 11.000000 | 0.89349 | 0.003574 | 3.73000 | 40.294000 | 379.587364 |
| 8 | 0 | 0 | 59 | 0 | 17.500000 | 0.48633 | 0.001945 | 7.02100 | 26.782000 | 404.784091 |
| 9 | 0 | 0 | 56 | 0 | 24.000000 | 0.61074 | 0.278778 | 83.92800 | 19.185000 | 429.980818 |
| 10 | 0 | 0 | 77 | 0 | 23.500000 | 0.29406 | 0.001176 | 6.21800 | 28.297000 | 455.177545 |
| 11 | 0 | 0 | 71 | 1 | 23.000000 | 1.05183 | 0.860337 | 243.08200 | 608.284000 | 480.374273 |
| 12 | 1 | 0 | 49 | 0 | 15.000000 | 0.85956 | 1.416314 | 151.83077 | 74.189903 | 505.571000 |
| 13 | 0 | 0 | 53 | 1 | 7.000000 | 1.91139 | 1.516773 | 150.89000 | 590.686000 | 487.775143 |
| 14 | 0 | 0 | 56 | 0 | 12.000000 | 0.91611 | 0.599645 | 93.81100 | 93.576000 | 469.979286 |

*Figure III.12 : Variables explicatives utilisées pour l'apprentissage du modèle.*

Les données affichées montrent une diversité importante entre les patients au niveau des valeurs biologiques et cliniques. Cette variation permet au modèle d'apprentissage automatique d'identifier les caractéristiques les plus pertinentes pour distinguer les différentes catégories diagnostiques. Les

variables explicatives jouent donc un rôle essentiel dans la qualité des prédictions du modèle. La Figure 13 représente la variable cible "Diagnosis" utilisée dans le processus de classification supervisée. Chaque valeur correspond à une catégorie diagnostique attribuée à un patient dans la base de données.

| | Diagnosis |
|---|---|
| 0 | 1 |
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 5 | 1 |

*Figure III.13 : Variable cible utilisée pour la classification des patients.*

Les valeurs numériques observées sont codées afin de faciliter leur traitement par le modèle de machine Learning. Cette variable constitue l'objectif principal de la prédiction, puisque le modèle doit apprendre à reconnaître correctement chaque catégorie diagnostique à partir des données médicales disponibles.

La figure 14 présente les données d'entraînement utilisées lors de l'apprentissage du modèle de régression logistique. Elle contient plusieurs observations de patients avec leurs caractéristiques médicales et biologiques.

| | Patient_Cohort | Sample_Origin | Age | Sex | Plasma_CA19_9 | Creatinine | LYVE1 | REG1B | TFF1 | REG1A |
|---|---|---|---|---|---|---|---|---|---|---|
| 573 | 1 | 2 | 85 | 1 | 787.533333 | 1.25541 | 9.522560 | 404.495350 | 3570.915000 | 342.510000 |
| 513 | 1 | 1 | 50 | 1 | 129.000000 | 0.42978 | 1.457756 | 62.254731 | 393.900700 | 205.046000 |
| 386 | 0 | 0 | 49 | 0 | 20.000000 | 0.71253 | 0.632433 | 22.602000 | 494.588000 | 121.653565 |
| 456 | 1 | 1 | 76 | 0 | 30.000000 | 0.22620 | 5.621266 | 406.730640 | 897.708948 | 1410.928000 |
| 249 | 0 | 0 | 70 | 0 | 19.000000 | 0.78039 | 6.018273 | 19.017000 | 566.488000 | 36.067857 |
| 170 | 0 | 0 | 65 | 1 | 10.333333 | 0.20358 | 0.000814 | 32.923000 | 70.350000 | 121.832273 |
| 55 | 0 | 0 | 49 | 1 | 2.116380 | 0.91611 | 0.013293 | 78.470300 | 13.293490 | 107.833091 |
| 90 | 0 | 0 | 71 | 0 | 1.178787 | 0.26013 | 0.000129 | 5.225890 | 0.129430 | 158.155000 |
| 560 | 1 | 0 | 63 | 0 | 433.000000 | 0.29406 | 3.199987 | 193.186000 | 321.529050 | 624.345000 |
| 589 | 1 | 0 | 74 | 1 | 1488.000000 | 1.50423 | 8.200958 | 411.938275 | 2021.321078 | 13200.000000 |
| 511 | 1 | 1 | 67 | 1 | 755.000000 | 0.37323 | 9.672798 | 369.256580 | 2719.353300 | 1646.490000 |
| 495 | 1 | 1 | 55 | 1 | 181.000000 | 0.50895 | 2.225879 | 24.173807 | 355.819100 | 82.651000 |
| 563 | 1 | 0 | 75 | 0 | 2202.000000 | 1.59471 | 9.667768 | 559.547520 | 1780.555200 | 1788.392000 |
| 475 | 0 | 0 | 71 | 1 | 434.000000 | 0.11310 | 0.773507 | 149.629000 | 98.272000 | 2512.000000 |
| 372 | 0 | 0 | 67 | 0 | 11.000000 | 0.80301 | 0.003212 | 3.751000 | 336.579000 | 43.215217 |

*Figure III.14 : Données d'entraînement utilisées pour l'apprentissage du modèle.*

Les données d'entraînement permettent au modèle d'apprendre les relations existantes entre les biomarqueurs et les diagnostics médicaux. Plus les données sont représentatives et variées, plus le modèle peut améliorer sa capacité de généralisation et produire des prédictions précises sur de

nouvelles données.

La figure 15 montre les catégories diagnostiques associées aux données d'entraînement. Les valeurs affichées correspondent aux classes utilisées pendant la phase d'apprentissage supervisé.

| | Diagnosis |
|---|---|
| 573 | 3 |
| 513 | 3 |
| 386 | 2 |
| 456 | 3 |
| 249 | 2 |
| 170 | 1 |
| 55 | 1 |
| 90 | 1 |

*Figure III.15 : Variable cible des données d'entraînement.*

La variable cible d'entraînement sert de référence au modèle afin qu'il puisse apprendre à associer chaque ensemble de caractéristiques médicales à son diagnostic correspondant. Cette étape est indispensable pour permettre au modèle de reconnaître correctement les différentes classes lors des prédictions futures.

La figure 16 présente les données de test utilisées pour évaluer les performances du modèle après l'apprentissage. Ces données n'ont pas été utilisées durant l'entraînement du modèle.

dtype: int64

| | Patient_Cohort | Sample_Origin | Age | Sex | Plasma_CA19_9 | Creatinine | LYVE1 | REG1B | TFF1 | REG1A |
|---|---|---|---|---|---|---|---|---|---|---|
| 225 | 0 | 0 | 41 | 1 | 5.600000 | 1.32327 | 1.897387 | 29.548000 | 46.814000 | 519.976818 |
| 14 | 0 | 0 | 56 | 0 | 12.000000 | 0.91611 | 0.599645 | 93.811000 | 93.576000 | 469.979286 |
| 85 | 0 | 0 | 64 | 1 | 1.286190 | 1.06314 | 0.023728 | 27.022365 | 23.728300 | 357.681091 |
| 418 | 1 | 1 | 64 | 0 | 1972.000000 | 0.83694 | 1.406857 | 13.618654 | 431.725959 | 27.737000 |
| 132 | 1 | 0 | 54 | 1 | 6.770000 | 0.95004 | 4.616530 | 54.969010 | 101.154050 | 173.162000 |
| 213 | 0 | 0 | 72 | 0 | 38.353333 | 1.51554 | 1.234861 | 42.507000 | 674.289000 | 476.904000 |
| 186 | 0 | 0 | 39 | 0 | 12.750000 | 0.89349 | 0.736352 | 21.550000 | 105.943000 | 1094.910882 |
| 340 | 1 | 3 | 80 | 1 | 94.371429 | 0.61074 | 6.698661 | 7.420854 | 61.373998 | 105.684000 |

*Figure III.16 : Données de test utilisées pour l'évaluation du modèle.*

Les données de test permettent de vérifier la capacité du modèle à généraliser ses prédictions sur de nouveaux patients. Cette étape est importante pour mesurer l'efficacité réelle du modèle et détecter d'éventuels problèmes de sur apprentissage.

La figure 17 représente les diagnostics réels associés aux données de test. Les valeurs correspondent aux catégories diagnostiques utilisées pour comparer les prédictions du modèle avec les résultats réels.

| | Diagnosis |
|---|---|
| 225 | 2 |
| 14 | 1 |
| 85 | 1 |
| 418 | 3 |
| 132 | 1 |
| 213 | 2 |
| 186 | 2 |
| 340 | 2 |
| 122 | 1 |
| 394 | 3 |
| 229 | 2 |
| 585 | 3 |
| 261 | 2 |
| 468 | 3 |
| 52 | 1 |

*Figure III.17 : Variable cible des données de test*

La comparaison entre les prédictions du modèle et les valeurs réelles de YTest permet d'évaluer la précision et les performances globales du modèle de classification. Cette étape joue un rôle essentiel dans la validation du système de détection du cancer du pancréas.

Les résultats ont montré que les variantes CA19-9, LYVE1, REG1B, TFF1 et REG1A ont une signification statistique significative dans la prédiction du cancer du pancréas.

```
two arrays of the FtrSlctn tuple displaying: the chi-squared statistics and their corresponding p-values for each feature:
(array([2.92167316e+01, 1.04346335e+02, 2.22383383e+02, 5.93786523e+00,
        4.53001322e+05, 3.37912687e+00, 6.84318750e+02, 3.38023914e+04,
        1.63691070e+05, 1.28373526e+05]), array([4.52550798e-007, 2.19523573e-023, 5.12934495e-049, 5.13581000e-002,
        0.00000000e+000, 1.84600097e-001, 2.52389568e-149, 0.00000000e+000,
        0.00000000e+000, 0.00000000e+000]))
```

| | 0 |
|---|---|
| Patient_Cohort | 4.525508e-07 |
| Sample_Origin | 2.195236e-23 |
| Age | 5.129345e-49 |
| Sex | 5.135810e-02 |
| Plasma_CA19_9 | 0.000000e+00 |
| Creatinine | 1.846001e-01 |
| LYVE1 | 2.523896e-149 |
| REG1B | 0.000000e+00 |
| TFF1 | 0.000000e+00 |
| REG1A | 0.000000e+00 |

dtype: float64

```
RandomForestClassifier
RandomForestClassifier(random_state=0)
```

*Figure III.18 : Résultats du test Chi-Square et des p-values.*

Le modèle Foret Aléatoire a été entrainé avec succès pour classer les patients en fonction de leur signes vitaux et de leurs données médicales. Les résultats ont montré que certains biomarqueurs, notamment CA19-9, LYVE1, REG1B, TFF1 et REG1A, sont fortement associés au diagnostic du cancer du pancréas, confirmant ainsi leur importance dans le processus de classification. De plus, les phases de

standardisation et de segmentation des données ont permis d'améliorer l'apprentissage, et le modèle forêt aléatoire a pu exploiter ces données pour une classification efficace des patients.

### III.6.2. Post-traitement

**III.6.2.1. Précision du modèle** : le score de précision mesure le pourcentage de prédictions correctes parmi toutes les prédictions effectuées.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{Nombre\ de\ prédictions\ correctes}}{\textbf{Nombre totale de préditions}} \quad \textbf{(III.1)}$$

Résultat : **84,3 %**, ce qui indique que le modèle classe correctement la majorité des patients

**III.6.2.2. Matrice de confusion** : la matrice de confusion obtenue permet d'évaluer la qualité des prédictions du modèle forêt aléatoire en comparant les classes réelles Healthy (0), NON-PDAC (1) et PDAC (2) aux classes prédites. Elle se compose de quatre groupes principaux :

- ✓ Les vrais positifs (TP) ;
- ✓ Les faux positifs (FP) ;
- ✓ Les faux négatifs (FN) ;
- ✓ Les vrais négatifs (TN).

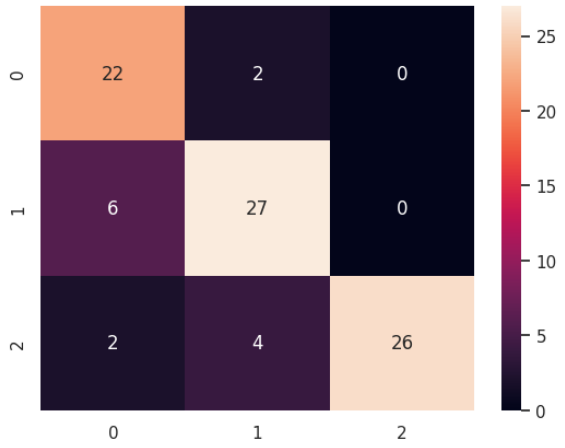


*Figure III.19 : Matrice de confusion du modèle forêt aléatoire.*

Les valeurs diagonales (22, 27, 26) représentent les vrais positifs (TP), c'est-à-dire les prédictions correctes. Les valeurs hors diagonale (2, 6, 2, 4) correspondent aux faux positifs (FP) et faux négatifs (FN).

**III.6.2.3. Rappel** : le rappel mesure la capacité du modèle à détecter correctement les cas positifs.

$$\boldsymbol{Rappel} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad \textbf{(III.2)}$$

**III.6.2.4. Précision** : la précision mesure la qualité des prédictions positives du modèle.

$$\textbf{précision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad \textbf{(III.3)}$$

**III.6.2.5. Score F1** : le F1-score combine la précision et le recall

$$F1 = 2 \times \frac{\text{Precision} + \text{Recall}}{\text{Precision} \times \text{Recall}} \quad \text{(III.4)}$$

**III.6.2.6. Spécificité** : mesurer la capacité du modèle à reconnaître correctement les vrais négatifs.

$$\text{Spécificité} = \frac{TN}{TN + FP} \quad \text{(III.5)}$$

On calcule la spécificité pour chaque classe.

| | |
|---|---|
| Spécificité classe 1 | **0.8688524590163934** |
| Spécificité classe **2** | **0.8888888888888888** |
| Spécificité classe **3** | **1.0** |

*Tableau III.1: Spécificité*

**III.6.2.7. Rapport de classification** : le rapport de classification présente trois indicateurs principaux pour chaque classe : la **précision,** le **rappel** et le **score F1**, ainsi que le nombre d'observations réelles (*support*), calculés pour chaque classe (Healthy, NON-PDAC, PDAC).

| **Classe** | **Précision** | **Rappel** | **Score F1** | **Support** |
|---|---|---|---|---|
| **Healthy** | 0,73 | 0,92 | 0,81 | 24 |
| **NON-PDAC** | 0,82 | 0,82 | 0,82 | 33 |
| **PDAC** | 1,00 | 0,81 | 0,90 | 32 |
| **Accuracy** | | | 0,84 | 89 |
| **Macro avg** | 0,85 | 0,85 | 0,84 | 89 |
| **Weighted avg** | 0,86 | 0,84 | 0,85 | 89 |

*Tableau III.2: Rapport de classification*

**III.6.2.8. Importance des variables** : la figure 20 illustre la contribution de chaque biomarqueur et variable clinique à la prédiction du diagnostic (Healthy, NON-PDAC, PDAC). Chaque barre du graphique représente un **poids relatif** attribué par l'algorithme la forêt aléatoire, indiquant l'influence de la variable correspondante dans la classification finale.

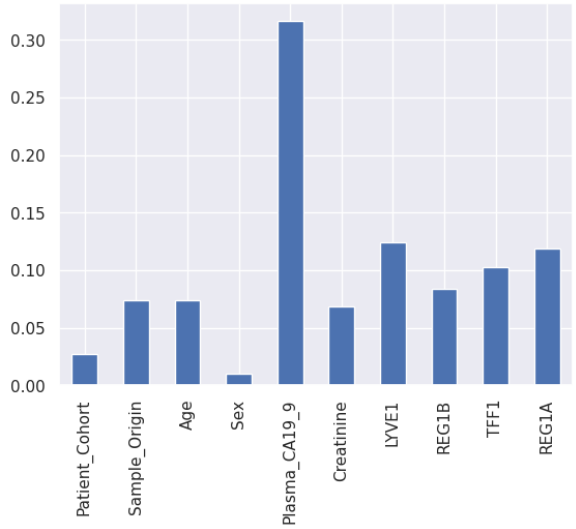


*Figure III.20 : Contribution des biomarqueurs et variables cliniques à la prédiction du diagnostic.*

## III.7. Application de l'algorithme de la regression logistique et construction du modèle

**III.7.1. Application de l'algorithme et construction du modèle** : la figure 21 présente un aperçu des variables médicales et biologiques utilisées dans l'étude avant l'apprentissage du modèle de régression logistique. On y trouve plusieurs caractéristiques importantes comme l'âge, le sexe, le taux de CA19-9, la créatinine ainsi que différents biomarqueurs liés au diagnostic du cancer du pancréas. Ces données représentent les variables explicatives utilisées pour entraîner le modèle.

| | Patient_Cohort | Sample_Origin | Age | Sex | Plasma_CA19_9 | Creatinine | LYVE1 | REG1B | TFF1 | REG1A |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 33 | 0 | 11.700000 | 1.83222 | 0.893219 | 52.94884 | 654.282174 | 1262.000000 |
| 1 | 1 | 0 | 81 | 0 | 9.350000 | 0.97266 | 2.037585 | 94.46703 | 209.488250 | 228.407000 |
| 2 | 0 | 0 | 51 | 1 | 7.000000 | 0.78039 | 0.145589 | 102.36600 | 461.141000 | 253.603727 |
| 3 | 0 | 0 | 61 | 1 | 8.000000 | 0.70122 | 0.002805 | 60.57900 | 142.950000 | 278.800455 |
| 4 | 0 | 0 | 62 | 1 | 9.000000 | 0.21489 | 0.000860 | 65.54000 | 41.088000 | 303.997182 |
| 5 | 0 | 0 | 53 | 1 | 9.666667 | 0.84825 | 0.003393 | 62.12600 | 59.793000 | 329.193909 |
| 6 | 0 | 0 | 70 | 1 | 10.333333 | 0.62205 | 0.174381 | 152.27700 | 117.516000 | 354.390636 |
| 7 | 0 | 0 | 58 | 0 | 11.000000 | 0.89349 | 0.003574 | 3.73000 | 40.294000 | 379.587364 |

*Figure III.21 : Présentation des variables médicales et biologiques des patients.*

Les valeurs affichées montrent une grande diversité entre les patients, ce qui permet au modèle d'apprendre les différences entre les différentes catégories diagnostiques. Certaines variables biologiques présentent des valeurs élevées chez certains patients, ce qui peut indiquer une relation avec la présence de la maladie. Cette étape est essentielle pour préparer les données avant la phase d'apprentissage automatique.

La figure 22 représente la variable cible « Diagnosis » utilisée dans le processus de classification. Chaque valeur correspond à une catégorie diagnostique attribuée à un patient dans la base de données.

| | Diagnosis |
|---|---|
| 0 | 1 |
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 5 | 1 |

*Figure III.22 : Variable cible "Diagnosis" utilisée pour la classification.*

On remarque que les données sont codées sous forme numérique afin de faciliter leur traitement par le modèle de Machine Learning. Le modèle utilisera ces classes pour apprendre à distinguer les patients sains, les patients non PDAC et les patients atteints de PDAC. Cette variable constitue l'élément principal à prédire.

Le tableau de la figure 23 contient une autre partie des données médicales utilisées pour l'entraînement du modèle. Il présente plusieurs observations de patients accompagnées de leurs caractéristiques cliniques et biologiques.

| | Patient_Cohort | Sample_Origin | Age | Sex | Plasma_CA19_9 | Creatinine | LYVE1 | REG1B | TFF1 | REG1A |
|---|---|---|---|---|---|---|---|---|---|---|
| 573 | 1 | 2 | 85 | 1 | 787.533333 | 1.25541 | 9.522560 | 404.495350 | 3570.915000 | 342.510000 |
| 513 | 1 | 1 | 50 | 1 | 129.000000 | 0.42978 | 1.457756 | 62.254731 | 393.900700 | 205.046000 |
| 386 | 0 | 0 | 49 | 0 | 20.000000 | 0.71253 | 0.632433 | 22.602000 | 494.588000 | 121.653565 |
| 456 | 1 | 1 | 76 | 0 | 30.000000 | 0.22620 | 5.621266 | 406.730640 | 897.708948 | 1410.928000 |
| 249 | 0 | 0 | 70 | 0 | 19.000000 | 0.78039 | 6.018273 | 19.017000 | 566.488000 | 36.067857 |
| 170 | 0 | 0 | 65 | 1 | 10.333333 | 0.20358 | 0.000814 | 32.923000 | 70.350000 | 121.832273 |
| 55 | 0 | 0 | 49 | 1 | 2.116380 | 0.91611 | 0.013293 | 78.470300 | 13.293490 | 107.833091 |
| 90 | 0 | 0 | 71 | 0 | 1.178787 | 0.26013 | 0.000129 | 5.225890 | 0.129430 | 158.155000 |
| 560 | 1 | 0 | 63 | 0 | 433.000000 | 0.29406 | 3.199987 | 193.186000 | 321.529050 | 624.345000 |
| 589 | 1 | 0 | 74 | 1 | 1488.000000 | 1.50423 | 8.200958 | 411.938275 | 2021.321078 | 13200.000000 |
| 511 | 1 | 1 | 67 | 1 | 755.000000 | 0.37323 | 9.672798 | 369.256580 | 2719.353300 | 1646.490000 |

*Figure III.23 : Données d'entraînement utilisées dans le modèle de régression logistique.*

Les données montrent des différences importantes entre les patients concernant les biomarqueurs et les valeurs biologiques. Ces variations permettent d'améliorer la capacité du modèle à identifier des schémas caractéristiques associés au cancer du pancréas. Plus les données sont variées, plus le modèle peut devenir performant et précis.

Cette image montre différentes valeurs de diagnostic associées aux patients après la préparation des données. Les classes affichées correspondent aux catégories utilisées dans la classification supervisée.

| | Diagnosis |
|---|---|
| 573 | 3 |
| 513 | 3 |
| 386 | 2 |
| 456 | 3 |
| 249 | 2 |
| 170 | 1 |

*Figure III.24 : Répartition des classes diagnostiques des patients.*

La présence de plusieurs catégories diagnostiques permet au modèle de réaliser une classification multi classes. Les différentes étiquettes servent de référence pendant l'apprentissage afin que le modèle puisse reconnaître correctement chaque type de diagnostic lors des prédictions futures.
Cette image affiche une partie supplémentaire des données d'entrée utilisées dans l'étude. Les colonnes représentent les caractéristiques médicales tandis que les lignes correspondent aux différents patients observés.

| | Patient_Cohort | Sample_Origin | Age | Sex | Plasma_CA19_9 | Creatinine | LYVE1 | REG1B | TFF1 | REG1A |
|---|---|---|---|---|---|---|---|---|---|---|
| 225 | 0 | 0 | 41 | 1 | 5.600000 | 1.32327 | 1.897387 | 29.548000 | 46.814000 | 519.976818 |
| 14 | 0 | 0 | 56 | 0 | 12.000000 | 0.91611 | 0.599645 | 93.811000 | 93.576000 | 469.979286 |
| 85 | 0 | 0 | 64 | 1 | 1.286190 | 1.06314 | 0.023728 | 27.022365 | 23.728300 | 357.681091 |
| 418 | 1 | 1 | 64 | 0 | 1972.000000 | 0.83694 | 1.406857 | 13.618654 | 431.725959 | 27.737000 |
| 132 | 1 | 0 | 54 | 1 | 6.770000 | 0.95004 | 4.616530 | 54.969010 | 101.154050 | 173.162000 |
| 213 | 0 | 0 | 72 | 0 | 38.353333 | 1.51554 | 1.234861 | 42.507000 | 674.289000 | 476.904000 |
| 186 | 0 | 0 | 39 | 0 | 12.750000 | 0.89349 | 0.736352 | 21.550000 | 105.943000 | 1094.910882 |
| 340 | 1 | 3 | 80 | 1 | 94.371429 | 0.61074 | 6.698661 | 7.420854 | 61.373998 | 105.684000 |
| 122 | 1 | 0 | 49 | 1 | 6.316667 | 1.45899 | 0.005836 | 28.332540 | 113.776998 | 84.123000 |
| 394 | 0 | 0 | 58 | 1 | 11.000000 | 0.44109 | 0.632433 | 188.253000 | 138.630000 | 321.551000 |

*Figure III.25 : Caractéristiques médicales utilisées pour l'apprentissage du modèle.*

Les informations biologiques affichées jouent un rôle important dans la détection des anomalies associées au cancer du pancréas. Certaines variables montrent des écarts significatifs entre les patients, ce qui aide le modèle à améliorer la séparation entre les différentes classes diagnostiques.

Cette image présente les résultats de la sélection des caractéristiques à l'aide du test statistique du chi carré (Chi-Squared Test).

| | Diagnosis |
|---|---|
| 225 | 2 |
| 14 | 1 |
| 85 | 1 |
| 418 | 3 |
| 132 | 1 |
| 213 | 2 |
| 186 | 2 |

*Figure III.26 : Résultats de la sélection des caractéristiques par le test du Chi carré.*

Les variables ayant des valeurs p très faibles sont considérées comme fortement significatives pour la classification. Cela signifie qu'elles possèdent une relation importante avec le diagnostic du cancer du pancréas. Cette étape permet de sélectionner les caractéristiques les plus pertinentes afin d'améliorer les performances du modèle et de réduire les variables inutiles.
Cette image présente une partie de la variable de diagnostic (Diagnosis) des patients utilisés dans le processus de classification. Les valeurs numériques affichées représentent les différentes catégories

diagnostiques utilisées dans le modèle de Machine Learning. Les données montrent la répartition des patients selon les différentes classes de diagnostic de l'étude.

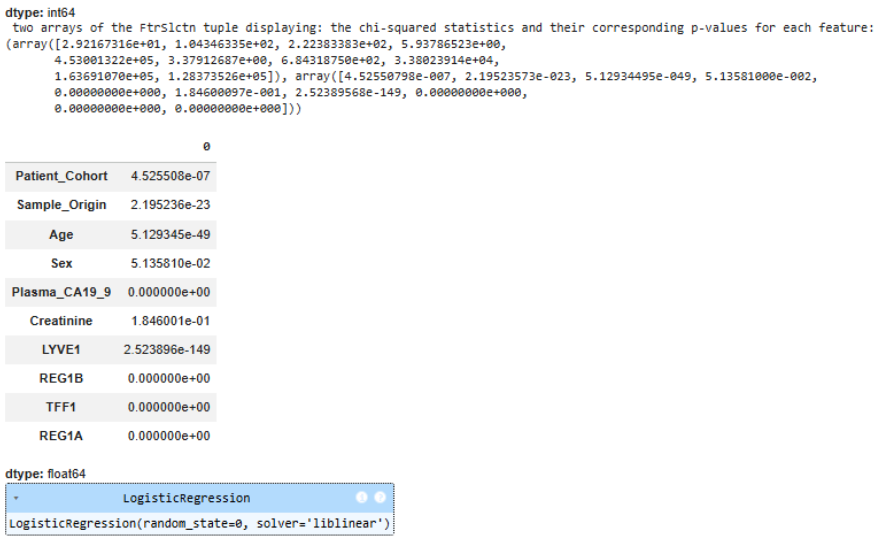


*Figure III.27 : Codage des catégories diagnostiques des patients.*

On remarque que la variable « Diagnosis » contient plusieurs classes codées par les valeurs 1, 2 et 3 afin de faciliter leur traitement par le modèle de régression logistique. Ces valeurs représentent les différents états diagnostiques des patients dans la base de données. Ce codage permet au modèle d'apprendre les différences entre les cas sains et les cas pathologiques, ce qui contribue à améliorer la précision des prédictions et la détection du cancer du pancréas à partir des données médicales et biologiques disponibles.

Les résultats obtenus à partir du modèle de régression logistique ont montré que plusieurs biomarqueurs, notamment CA19-9, LYVE1, REG1B, TFF1 et REG1A, présentent une forte relation avec le diagnostic du cancer du pancréas, ce qui confirme leur importance dans le processus de classification médicale. De plus, les étapes de préparation des données, telles que la séparation des variables explicatives et de la variable cible ainsi que la division des données en ensembles d'entraînement et de test, ont contribué à améliorer l'apprentissage du modèle. Ainsi, le modèle de régression logistique a pu exploiter efficacement les données médicales afin de réaliser une classification précise des patients et d'améliorer les performances de prédiction du diagnostic.

### III.7.2. Post-traitement

**III.7.2.1. Précision du modèle la régression logistique** : le résultat du précision (Logistic Regression) = 0.7865168539325843 signifie que le modèle de régression logistique a correctement prédit environ 79 % des cas dans l'ensemble de test.

**III.7.2.2. Matrice de confusion** : la figure 9 montre la répartition des prédictions du modèle pour les

trois classes : Healthy (0), NON-PDAC (1) et PDAC (2). Les valeurs diagonales (22, 24, 24) représentent les vrais positifs (TP), c'est-à-dire les cas correctement classés.

Les valeurs hors diagonale indiquent les faux positifs (FP) et faux négatifs (FN), correspondant aux erreurs de classification.

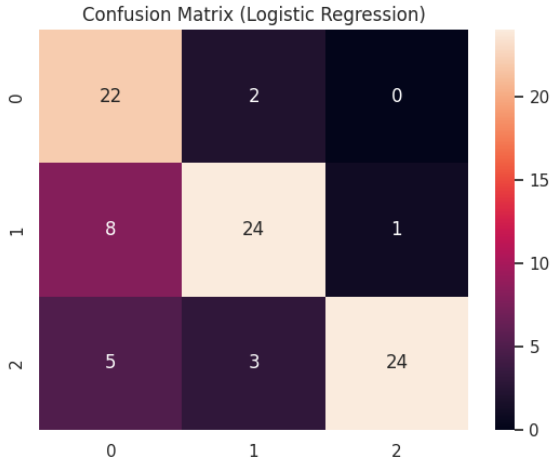


*Figure III.28 : Matrice du confusion "la régression logistique".*

**III.7.2.3. Évaluation statistique du modèle de régression logistique** : les performances du modèle de régression logistique montrent une précision globale satisfaisante mais perfectible. Le rappel (0,798) indique que près de 80 % des patients atteints de PDAC ont été correctement détectés, ce qui traduit une bonne sensibilité du modèle. La précision (0,805) révèle que plus de 80 % des prédictions positives sont exactes, limitant ainsi les fausses alertes. Le score F1 (0,787), qui combine rappel et précision, confirme un équilibre global entre détection et fiabilité. La spécificité varie selon les classes : elle est correcte pour la classe 1 (0,787), élevée pour la classe 2 (0,902) et excellente pour la classe 3 (0,979), ce qui montre que le modèle distingue bien les patients non atteints. Enfin, l'accuracy globale (0,787) traduit que près de 79 % des prédictions sont correctes, mais souligne que 21 % des cas restent mal classés.

| | |
|---|---|
| **Le rappel** | 0.797979797979798 |
| **La précision** | 0.8053858784893267 |
| **Le score F1** | 0.7873538411364661 |
| **La spécificité** | Classe1:0.7868852459016393<br>Classe2:0.9019607843137255<br>Classe3:0.9787234042553191 |
| **Accuracy** | 0.7865168539325843 |

*Tableau III .3 : Indicateur de performance (Rappel, Précision, F1, Spécificité, Accuracy).*

**III.7.2.4. Rapport de classification** : les résultats du modèle de régression logistique montrent des performances variables selon les classes. Pour la classe Healthy, la précision est relativement faible (0,63) mais le rappel est élevé (0,92), ce qui signifie que la majorité des patients sains sont correctement identifiés, au prix de nombreuses fausses alertes. La classe NON-PDAC présente une précision correcte (0,83) mais un rappel plus limité (0,73), traduisant une tendance du modèle à manquer certains cas. La classe PDAC affiche une précision très élevée (0,96), ce qui indique que presque toutes les prédictions positives sont correctes, mais le rappel reste modéré (0,75), révélant que certains patients atteints ne sont pas détectés.

Le score F1, qui équilibre précision et rappel, varie entre 0,75 et 0,84 selon les classes, confirmant une performance globalement satisfaisante.

L'accuracy globale atteint 0,79, ce qui signifie que près de 80 % des prédictions sont correctes. Les moyennes macro et pondérée (≈ 0,79–0,82) confirment un équilibre général

| Classe | Précision | Rappel | Score F1 | Support |
|---|---|---|---|---|
| **Healthy** | 0,63 | 0,92 | 0,75 | 24 |
| **NON-PDAC** | 0,83 | 0,73 | 0,77 | 33 |
| **PDAC** | 0,96 | 0,75 | 0,84 | 32 |
| **Accuracy** | | | 0,79 | 89 |
| **Macro avg** | 0,81 | 0,80 | 0,79 | 89 |
| **Weighted avg** | 0,82 | 0,79 | 0,79 | 89 |

***Tableau III .4 : Rapport de classification "la régression logistique".***

# III.8. Application de l'algorithme de la logique floue et construction du modèle

## III.8.1. Application de l'algorithme et construction du modèle

Le modèle basé sur le **logique floue** a été utilisé afin de regrouper les patients en trois catégories : *Healthy*, *NON-PDAC* et *PDAC*.

Contrairement aux méthodes de classification classiques, cet algorithme attribue à chaque observation un degré d'appartenance à chaque classe, permettant ainsi une représentation plus flexible des données.

Après transformation des degrés d'appartenance en classes finales, le modèle a obtenu une **précision globale de 51%**, ce qui indique une performance (modérée).

- ✓ Accuracy = 0. 515593220338983
- ✓ Rapport de classification

| Classe | Précision | Rappel | Score F1 | Support |
|---|---|---|---|---|
| **1** | 0,46 | 0,69 | 0,55 | 183 |
| **2** | 0,00 | 0,00 | 0,00 | 208 |
| **3** | 0,56 | 0,89 | 0,69 | 199 |
| **Accuracy** | | | 0,51 | 590 |
| **Macro avg** | 0,34 | 0,53 | 0,41 | 590 |
| **Weighted avg** | 0,33 | 0,51 | 0,40 | 590 |

***Tableau III .5 : Rapport de classification du modèle de logique floue.***

- ✓ Matrice de confusion : L'analyse de la matrice de confusion montre que les classes *Healthy* et *NON-PDAC* sont partiellement bien identifiées, avec un nombre acceptable de prédictions correctes. Cependant, la classe *PDAC* n'a pas été correctement détectée, avec un rappel égal à zéro. Cela signifie que les patients atteints de PDAC n'ont jamais été correctement classés par le modèle.

```
[[126   0   57]
 [125   0   83]
 [ 22   0  177]]
```
**…… (III.6)**

Ce résultat (figure 29) s'explique par le chevauchement important entre les différentes classes dans l'espace des caractéristiques, ce qui rend difficile leur séparation par une approche de clustering flou. En effet, certains patients PDAC présentent des caractéristiques similaires aux autres groupes, ce qui conduit le modèle à les attribuer à des classes différentes.

```
First 10 Predictions:
[1 3 1 1 1 1 3 1 1 1]
```

***Figure III.29 : Résultats des prédictions initiales du modèle de logique floue.***

En conclusion, bien que le logique floue permette une modélisation souple basée sur des degrés d'appartenance, ses performances restent limitées pour la détection de certaines classes complexes comme le PDAC dans cet ensemble de données.

### III.8.2. Post-traitement

#### III.8.2.1. La précision

Le résultat précision (FCM) = 0,497 montre que le modèle logique floue « *Fuzzy C-Means* » n'a correctement classé qu'environ 50 % des patients**.**

### III.8.2.2. Matrice de la confusion

La figure I.30 montre une forte concentration des prédictions dans les classes NON-PDAC et **PDAC**, tandis que la classe Healthy n'a été détectée dans aucun cas

Les valeurs principales sont :

- **Healthy** : 0 cas correctement identifiés, 113 confondus avec NON-PDAC et 70 avec PDAC.
- **NON-PDAC** : 119 cas correctement classés, mais 89 confondus avec PDAC.
- **PDAC** : 174 cas correctement détectés, 25 confondus avec NON-PDAC.

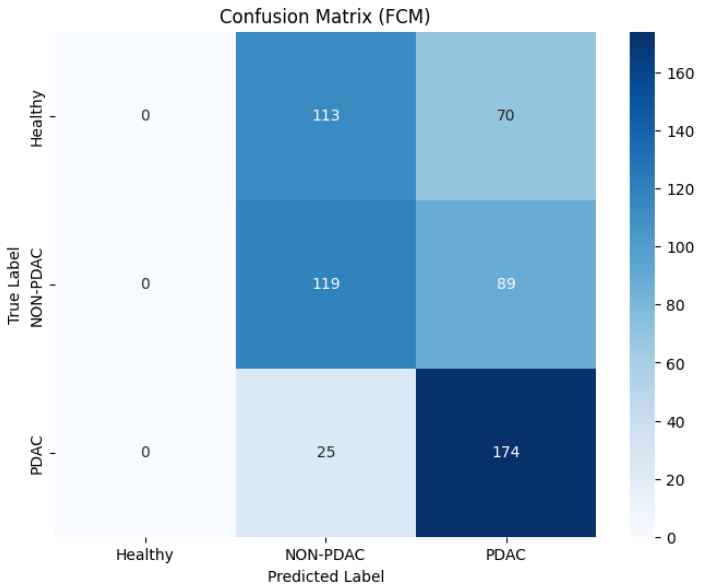


*Figure III.30 : Matrice de confusion du modèle floue logique.*

L'absence totale de détection des patients sains (aucun vrai négatif) montre que le modèle ne parvient pas à reconnaître la classe Healthy.

Les confusions fréquentes entre NON-PDAC et PDAC traduisent une mauvaise séparation des frontières entre ces deux catégories, probablement due à la nature floue du clustering.

Bien que la classe PDAC soit la mieux reconnue (174 cas corrects), le modèle tend à sur-classifier les observations dans cette catégorie, ce qui explique le grand nombre de faux positifs.

### III.8.2.3. Rapport de classification

| Classe | Précision | Rappel | Score F1 | Support |
|---|---|---|---|---|
| **Healthy** | 0,00 | 0,00 | 0,00 | 183 |
| **NON-PDAC** | 0,46 | 0,57 | 0,51 | 208 |
| **PDAC** | 0,52 | 0,87 | 0,65 | 199 |
| **Accuracy** | | | 0,50 | 590 |

| | | | | |
|---|---|---|---|---|
| **Macro avg** | 0,33 | 0,48 | 0,39 | 590 |
| **Weighted avg** | 0,34 | 0,50 | 0,40 | 590 |

*Tableau 6 : Rapport de classification du modèle floue logique.*

Le tableau montre que :

- ✓ **Classe Healthy** : Le modèle n'a aucune capacité à identifier les patients sains, ce qui confirme les observations de la matrice de confusion (aucun vrai positif).
- ✓ **Classe NON-PDAC** : Le modèle détecte un peu plus de la moitié des cas réels, mais avec de nombreuses erreurs de classification.
- ✓ **Classe PDAC** : Le modèle reconnaît bien les patients atteints de PDAC (rappel élevé), mais au prix d'un grand nombre de faux positifs, ce qui réduit la précision.
- ✓ Les moyennes globales (**macro avg = 0,39**, **weighted avg = 0,40**) et l'**accuracy = 0,50** confirment une faible performance globale : le modèle ne distingue correctement qu'un cas sur deux.

## III.9. Comparaison avec les modèles d'apprentissage automatique

| **Critère** | **Foret aléatoire** | **Régression logistique** | **Logique floue** |
|---|---|---|---|
| **Type** | Ensemble d'arbres de décision (bagging) | Classifieur linéaire supervisé | Clustering flou non supervisé |
| **Exactitude (Accuracy)** | Elevée (~84%) | Moyenne (~79%) | Faible (~50%) |
| **Vitesse d'exécution** | Moyenne | Très rapide | Rapide |
| **Interprétabilité** | Moyenne | Très élevée | Faible |
| **Contrôle du surapprentissage** | Bagging et aléa | Régularisation(L1/L2) | Aucun mécanisme |
| **Gestion de la non-linéarité** | Bonne | Faible | Très faible |
| **Importance des variables** | Basée sur les scores de Gini | Basée sur les coefficients | Non disponible |
| **Valeurs manquantes** | Gere partiellement les valeurs manquantes | Doivent être prétraitées | Doivent être prétraitées |
| **Mise a l'échelle requise** | Non | Oui | Oui |
| **Pertinence clinique/cas d'usage** | Meilleur modelé pour la prédiction du cancer du pancréas | Bon compromis entre simplicité et fiabilité | Inadapté au diagnostic |

*Tableau III .7 : Comparaison entre la logique floue / les systèmes experts et l'apprentissage automatique.*

## III.10. Conclusion

Dans ce chapitre, nous avons appliqué trois algorithmes : la régression logistique, les forêts aléatoires et la logique floue sur une base de données de 590 patients pour la classification du cancer du pancréas (Healthy, NON-PDAC, PDAC). Après une phase de prétraitement incluant l'interpolation des valeurs manquantes et la normalisation des données, les résultats obtenus montrent une nette supériorité de la forêt aléatoire avec une précision de 84,3 %, contre 78,7 % pour la régression logistique et seulement 51,3 % pour la logique floue. La forêt aléatoire se distingue par son excellent équilibre entre précision et rappel, notamment pour la classe PDAC (F1-score = 0,90), tandis que la logique floue s'avère inadaptée à ce type de données structurées en raison du fort chevauchement entre les classes. En perspective, l'amélioration des performances passerait par l'utilisation de bases de données plus larges et l'exploration d'algorithmes plus avancés.

# Conclusion générale

A travers ce travail, nous avons démontré l'efficacité des techniques d'intelligence artificielle dans la prédiction et la classification du cancer du pancréas. Les étapes de prétraitement (interpolation des valeurs manquantes, normalisation Min-Max) ont contribué à la robustesse des modèles, tandis que les algorithmes supervisés ont montré des performances solides. Parmi eux, la forêt aléatoire s'est distinguée par sa précision (84,3 %) et ses excellents scores F1 pour les trois classes diagnostiques (Healthy : 0,81, NON-PDAC : 0,82, PDAC : 0,90). La régression logistique a offert une transparence appréciable pour l'interprétation des biomarqueurs, mais avec une précision légèrement inférieure (78,7 %). En revanche, la logique floue, bien qu'intéressante pour son raisonnement explicable, a montré des limites importantes face à des données médicales complexes.

Cette comparaison met en évidence un compromis : les modèles d'apprentissage automatique offrent une haute précision, tandis que les systèmes flous privilégient l'explicabilité. Pour le diagnostic du cancer du pancréas, où la détection précoce est critique, la forêt aléatoire apparaît comme l'approche la plus fiable. Les perspectives futures incluent l'enrichissement de la base de données, l'exploration d'algorithmes plus puissants (XGBoost, deep learning), et l'intégration de techniques d'IA explicable (XAI) afin de concilier performance et transparence.

Ce travail confirme que l'intelligence artificielle, en particulier la forêt aléatoire, est un outil puissant pour la détection précoce du cancer du pancréas, non pas pour remplacer les médecins mais pour les soutenir dans la prise de décisions diagnostiques plus rapides et plus précises.

# Bibliographies

[1] « Innovations et thérapeutiques en oncologie », Cairn.info, 2024. [En ligne]. Disponible : https://stm.cairn.info/revue-innovations-et-therapeutiques-en-oncologie-2024-1-page-30. [Consulté le : 27 février 2026].

[2] « Intelligence artificielle », IBM, [En ligne]. Disponible : https://www.ibm.com/fr-fr/think/topics/artificial-intelligence. [Consulté le : 27 février 2026].

[3] Samoili S, Lopez CM, Delipetrev B, Martinez-Plumed F, Gomez GE, De PG. JRC Publications Repository. 2021 [cited 2024 May 14]. AI Watch. Defining Artificial Intelligence 2.0. Available from: https://publications.jrc.ec.europa.eu/repository/handle/JRC126426

[4] « Intelligence artificielle », Inventiv IT, [En ligne]. Disponible : https://inventiv-it.fr/intelligence-artificielle/. [Consulté le : 27 février 2026].

[5] C. Martin, « Tomographie d'impédance électrique en dermatologie », 2019.

[6] [Ghediri et al., 2021] Ghediri, Rebahi, KhawLa, Semri, Kenza, Belhouchette, & Imane. 2021. La Reconnaissance des émotions de base par Les réseaux de neurones : application de deep Learning.

[7] [Bird et al., 2009] Bird, Steven, Klein, Ewan, & Loper, Edward. 2009. Natural Language Processing with Python. Culemborg, Netherlands: Van Duuren Media.

[8] S.Sahli, "Prédiabète : Un système de détection et prédiction de diabète, Mémoire de Master, Univ.8 Mai 1945, Guelma, Algérie, 2022.

[9] J. Abbas et A. Hamdad, « Apprentissage automatique du dialecte algérien », Mémoire de fin d'études, Master en informatique, Université Mouloud Mammeri de Tizi-Ouzou, Algérie, 2020.

[10] «What is dermoscopy », Skin Repair, [En ligne]. Disponible : https://www.skinrepair.net.au/what-is-dermoscopy. [Consulté le : 27 février 2026].

[11] F. K. Dosilovic, M. Brcic et N. Hlupic, « Explainable artificial intelligence: A survey », in *Proc. MIPRO*, 2018, pp. 210–215.

[12] C. Krueger *et al.*, « Machine learning and artificial intelligence: Definitions, applications, and future directions », *Current Reviews in Musculoskeletal Medicine*, vol. 13, n° 1, pp. 69–76, 2020.

[13] T. Davenport et R. Kalakota, « The potential for artificial intelligence in healthcare », *Future Healthcare Journal*, vol. 6, n° 2, pp. 94–98, 2019.

[14] S. Reddy, J. Fox et M. P. Purohit, « Artificial intelligence-enabled healthcare delivery », *Journal of the Royal Society of Medicine*, vol. 112, n° 1, pp. 22–28, 2019.

[15] S. Castagno et M. Khalifa, « Perceptions of artificial intelligence among healthcare staff », *Frontiers in Artificial Intelligence*, vol. 3, 2020.

[16] M. C. Laï *et al.*, « Perceptions de l'intelligence artificielle en santé », *Journal of Translational Medicine*, vol. 18, n° 1, 2020.

[17] S. Secinaro *et al.*, « The role of artificial intelligence in healthcare », *BMC Medical Informatics and Decision Making*, vol. 21, n° 1, 2021.

[18] J. Amann *et al.*, « Explainability for artificial intelligence in healthcare », *BMC Medical Informatics and Decision Making*, vol. 20, n° 1, 2020.

[19] J. J. Xu, « Knowledge discovery and data mining », in *Computer Handbook*, 3e éd., CRC Press, 2014.

[20] M. M. Bukhari *et al.*, « Artificial neural network model for diabetes prediction », *Complexity*, 2021.

[21] M. Labed, « Classification pour le diagnostic du diabète utilisant des algorithmes d'apprentissage », Mémoire de master, Université Badji Mokhtar, Annaba, Algérie, 2021.

[22] « Cancer du pancréas: définition, causes et traitement », Elsan.

[23] Fondation contre le cancer, 'Types de cancers du pancréas et anatomie de l'organte.' [En ligne].Disponible : https://cancer.be/cancer/cancer-du-pancreas/types-de-cancers-du-panceas-et-anatomie-de-l-organe/[Consulté le : 29 avril 2026].

[24] « Dépistage et diagnostic du cancer du pancréas », VIDAL.

[25] « Cancer du pancréas et espérance de vie », Fondation ARC.

[26] J. Talukdar, T. P. Singh et B. Barman, *Artificial Intelligence in Healthcare Industry*. Cham, Suisse: Springer, 2023.

[27] D. Pradhan, P. K. Sahu, H. M. Tun et P. Chatterjee, *Artificial and Cognitive Computing for Sustainable Healthcare Systems in Smart Cities*. Londres, Royaume-Uni: ISTE, 2024.

[28] GeeksforGeeks, « Logistic Regression in Machine Learning », 2025. [En ligne]. Disponible : https://www.geeksforgeeks.org/understanding-logistic-regression/
[Consulté le : 3 mai 2026].

[29] GeeksforGeeks, « Random Forest Algorithm in Machine Learning », 2025. [En ligne]. Disponible :https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/
[Consulté le : 3 mai 2026].

[30] M. Cherfaoui et F. Chouiter, « Artificial intelligence application for diabetes prediction », mémoire de master, University of Mohamed El Bachir El Ibrahimi Bordj Bou Arréridj, Bordj Bou Arréridj, Algérie, 2025.

[31] M. Benbrahim, « Chapitre II : Système expert », support de cours, University of Batna 2, Batna, Algérie, 2020. [En ligne]. Disponible : (ajoute le lien PDF si tu l'as) [Consulté le : 3 mai 2026].

[32] Debernardi S., O'Brien H., Algahmdi A.S., Malats N., Stewart G.D., et al., *A Combination of Urinary Biomarker Panel and PancRISK Score for Earlier Detection of Pancreatic Cancer : A Case-Control Study*, PLOS Medicine, vol. 17, n° 12, e1003489, 2020.