



REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE  
MINISTRE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

**UNIVERSITE IBN KHALDOUN - TIARET**

# MEMOIRE

Présenté à :

**FACULTÉ MATHÉMATIQUES ET INFORMATIQUE  
DÉPARTEMENT D'INFORMATIQUE**

Pour l'obtention du diplôme de :

**MASTER**

Spécialité : Génie Informatique

Par :

**Fettah Lilia**

**Mehloul Hind**

Sur le thème

---

## **Application de Deep Learning pour un système de santé Intelligent dans un environnement cloud**

---

Soutenu publiquement le 10 / 07 / 2023 à Tiaret devant le jury composé de :

Mr Djaafri Laoun	M.C.A	Université Ibn-khaldoun	Président
Mr Messaoud Hameurlaine	M.C.A	Université-Tissemsilt	Encadreur
Mr GAFOUR Yacine	M.C.A	Université Ibn-khaldoun	Examineur
Mr Benghani Abd-Malek	M.C.A	Université Ibn-khaldoun	CO-Encadreur

2022-2023

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

# *Remerciement*

*En premier lieu, nous tenons à exprimer notre profonde gratitude à Dieu, merci à Allah, pour nous avoir accordé la santé, la persévérance et les opportunités nécessaires pour mener à bien ce projet. Sa bénédiction et sa guidance ont été inestimables tout au long de notre parcours.*

*À nos parents, nous sommes infiniment reconnaissants pour Leurs encouragements ont été une source de motivation pour nous, et nous leur sommes extrêmement reconnaissants d'avoir cru en nous et de nous avoir soutenus à chaque étape de notre vie .*

*Nous remercions Mr Messaoud HAMEURLAINE, notre encadreur et MR Benghani Abdelmalik le CO-Encadreur, pour son expertise, sa patience et ses précieux conseils. Sa guidance a été essentielle dans la réalisation de notre projet.*

*Nous adressons également nos remerciements aux membres du jury, Mr Djaafri Laoun et Mr GAFOUR Yacine, pour avoir consacré leur temps et leur expertise à l'évaluation de notre projet. et nous sommes honorés d'avoir bénéficié de leurs connaissances et de leur évaluation.*

*Enfin, un immense merci à tous les professeurs universitaires qui ont partagé leur savoir et leur passion avec nous tout au long de nos études.*

# *Dédicace*

*Je dédie ce modeste travail*

*À mes chers parents, Votre amour, votre soutien inconditionnel et vos encouragements ont été essentiels tout au long de ce parcours. Votre confiance en moi m'a donné la force de persévérer et de repousser mes limites.*

*À Ma sœur Fella, Ton amour, tes conseils et ton encouragement m'ont donné la force de persévérer et de réaliser mes rêves. Je suis profondément reconnaissante d'avoir une sœur aussi merveilleuse que toi.*

*À mes chers frères Karim et Soufien, Vous êtes les piliers de ma vie, Je suis fière d'avoir des frères aussi exceptionnels que vous. Notre lien fraternel est un cadeau précieux que je chérirai toujours*

*À ma chérie HANAN, Je tiens à te dédier ces mots remplis d'affection et de reconnaissance. Bien que nous ayons des liens familiaux, notre relation va bien au-delà. Tu es plus qu'une cousine, tu es une amie proche et une complice précieuse dans ma vie.*

*À mes neveux bien-aimés, Rahef et Iyad, Je serai toujours là pour vous, pour vous soutenir et vous encourager à réaliser vos rêves. Je vous aime infiniment.*

*À mes amis les plus proches, Hind Mehloul et Manina Noukase.*

*Sans oublier tous les professeurs qui ont contribué à ma formation de l'enseignement primaire, moyen et secondaire jusqu'à l'enseignement supérieur.*

*Lilia Fettah*

# *Dédicace*

*À mes chers parents, Votre amour, votre soutien inconditionnel et vos encouragements ont été essentiels tout au long de ce parcours. Votre confiance en moi m'a donné la force de persévérer et de repousser mes limites .*

*À ma chère sœur Douaa et mon cher frère Islam,  
Mes précieux complices de vie. Vous êtes bien plus que de simples frères et sœurs, vous êtes mes amis les plus proches et mes alliés les plus fidèles. Je suis fière de vous avoir à mes côtés et je chéris notre lien indéfectible.*

*À mon cher mari Mohammed, Cette dédicace est spécialement pour toi . Tu as été mon soutien inconditionnel mon compagnon de vie et ma source de bonheur. Ta présence m'a donné la force et la motivation nécessaires pour atteindre mes objectifs.*

*À toute ma famille Mehoul et Mahious. A tous ceux qui, de près ou de loin, ont participé à mon éducation, et m'ont aidé Dans les moments difficiles à surmonter mes problèmes .*

*À ma belle famille Bouaricha, merci de m'avoir accueillie à bras ouverts et d'avoir fait de moi l'un des vôtres. Je suis honorée de faire partie de votre famille et je suis reconnaissante de tout l'amour et le bonheur que vous apportez dans ma vie.*

*À ma chère amie Lilia FETTAH, Cette dédicace est spécialement pour toi, mon amie précieuse. Tu es une source constante de soutien, d'écoute et de réconfort. tu as toujours été là pour moi, prête à partager les rires, les larmes et les aventures. Je te chéris et je suis honoré(e) de t'avoir à mes côtés.*

*Hind Mehoul*

# Résumé

Ce mémoire se concentre sur l'identification et la prédiction du diabète en utilisant des techniques d'apprentissage automatique et profond. L'objectif est de construire un modèle capable de prédire si une personne est diabétique ou non en se basant sur des algorithmes tels que Decision Trees (Arbre de décision), Deep Neural Network (DNN), Support Vector Machine (SVM) , Les réseaux neuronaux de long terme à court terme (LSTM).

Les résultats obtenus à partir de l'ensemble de données de l'hôpital de Frankfort en Allemagne ont démontré que les réseaux neuronaux LSTM étaient l'algorithme le plus performant en termes de précision, aussi bien pour les méthodes d'évaluation que pour les paramètres de performance.

**Mots Clés :** Prédiction du diabète, Apprentissage automatique, Apprentissage profond, DNN, SVM, LSTM , Arbres de décision.

## Abstract

This dissertation focuses on the identification and prediction of diabetes using machine learning and deep learning techniques. The goal is to build a model capable of predicting whether a person is diabetic or not based on algorithms such as Decision Trees, Deep Neural Network (DNN), Support Vector Machine (SVM), Short-term memory long neural networks (LSTM).

Results obtained from the Frankfurt Hospital data set in Germany demonstrated that LSTM neural networks are the most accurate algorithm for both assessment methods and performance parameters.

**Key words:** Diabetes Prediction, Machine Learning, Deep Learning, DNN, SVM, LSTM, Decision Trees.

## ملخص

تركز هذه الأطروحة على تحديد مرض السكري والتنبؤ به باستخدام تقنيات التعلم العميق والآلي. الهدف هو بناء نموذج قادر على التنبؤ بما إذا كان الشخص مصابًا بمرض السكري أم لا بناءً على خوارزميات مثل، أشجار القرار، آلة ناقلات الدعم، الشبكة العصبية العميقة و الشبكات العصبية طويلة المدى للذاكرة.

أظهرت النتائج التي تم الحصول عليها من مجموعة بيانات مستشفى فرانكفورت في ألمانيا أن الشبكات العصبية طويلة المدى للذاكرة هي الخوارزمية الأكثر دقة لكل من طرق التقييم و معايير الأداء.

**الكلمات المفتاحية :** التنبؤ بمرض السكري، التعلم الآلي، التعلم العميق، أشجار القرار، آلة ناقلات الدعم، الشبكة العصبية العميقة، الشبكات العصبية طويلة المدى للذاكرة.

## TABLE DES MATIERES

INTRODUCTION GÉNÉRALE.....	1
<b>Chapitre I: Les systèmes de santé intelligents et le diabète</b> .....	3
1.Les systèmes de santé intelligents et le diabète .....	3
1.1 INTRODUCTION.....	3
1.2 INTELLIGENCE ARTIFICIELLE EN MEDICINE.....	3
1.2.1 Systèmes d'Aide à la Décision (SADs).....	4
1.2.2 Les systèmes d'aide au diagnostic médical (SADMs) .....	5
2. LE DIABETE.....	7
2.1. DEFINITION DE DIABETE .....	7
2.2. DIABÈTES ET COMPLICATIONS.....	8
2.3. SYMPTOMES DE DIABETE .....	9
2.4. DIAGNOSTIC DE DIABETE .....	11
2.4.1. Qui est concerné par le dépistage diabétique ?.....	12
2.4.2. Comment savoir si l'on est diabétique ?.....	12
2.4.3. Décoder et comprendre les résultats de la glycémie.....	13
2.4.4. A quelle fréquence dois-je contrôler ma glycémie ?.....	14
2.5. CLASSIFICATION DE DIABETE .....	15
2.5.1 Diabète de type 1 .....	15
2.5.2 Diabète de type 2 .....	16
2.5.3. Diabète gestationnel .....	17
2.6. PREVENTION DE DIABETE TYPE 2.....	19
2.7 LE RÔLE DES SYSTÈMES DE SANTÉ INTELLIGENTS DANS LA GESTION EFFICACE DU DIABÈTE .....	20
2.8 CONCLUSION .....	211
<b>Chapitre II : L'apprentissage automatique &amp; Approfondi</b> .....	23
2.1 INTRODUCTION.....	23
2.2 APPRENTISSAGE AUTOMATIQUE.....	23
2.3 LES TYPES D'APPRENTISSAGE AUTOMATIQUE .....	24
2.3.1 Apprentissage Supervisé.....	24

2.3.2 Apprentissage Non Supervisé.....	25
2.3.3 Apprentissage par renforcement.....	26
2.4. LES ALGORITHMES DE L'APPRENTISSAGE AUTOMATIQUE:.....	28
2.4.1 K nearest neighbors (KNN).....	28
2.4.2 Decision Trees (Arbre de décision).....	30
2.4.3 Random Forest (forêts aléatoires).....	31
2.4.4 Support Vector Machine (SVM).....	32
3.1 L'APPRENTISSAGE APPROFONDI.....	34
3.2 LES TYPES DE MODÈLES D'APPRENTISSAGE EN PROFONDEUR.....	35
3.2.1 Les réseaux neuronaux convolutifs (CNN).....	35
3.2.2 Les réseaux neuronaux récurrents (RNN).....	40
3.2.3 Les réseaux de neurones artificiels (ANN).....	41
3.2.4 Deep Neural Network (DNN) :.....	42
3.2.5 Les réseaux antagonistes génératifs (GAN).....	43
3.2.6 Les réseaux neuronaux à long terme et court terme(LSTM).....	44
3.2.7 Modèle VGG16.....	46
3.2.8 LES TRAVAUX DE RECHERCHE SUR L'APPLICATION DES ALGORITHMES DE MACHINE LEARNING ET DU DEEP LEARNING POUR LA PRÉDICTION DU DIABÈTE ...	47
3.3 CONCLUSION.....	50
<b>Chapitre III :Contribution et résultats.....</b>	<b>53</b>
3. INTRODUCTION.....	53
3.1 ENVIRONNEMENT CLOUD.....	53
3.1.1 Définition du cloud computing.....	53
3.2 OUTILS ET LIBRAIRIES UTILISÉS.....	54
3.2.1 Google Colab.....	54
3.2.2 Jupyter notebook.....	55
3.2.3 TensorFlow.....	55
3.2.4 Keras.....	55
3.2.5 Python.....	55
3.3 BIBLIOTHEQUES DE PYTHON.....	55
3.4 DESCRIPTION DU DATASET.....	56
3.4.1 Définition l'ensemble de données utilisé.....	56

3.4.2 Les étapes de pré-traitement de données .....	59
3.4.3 Exploration et visualisation de données .....	60
3.4.4 Nettoyage des données .....	61
3.4.5 Corrélation .....	63
3.4.6 Proportion des variables .....	64
3.5 Entraînement et Test des données .....	65
3.5.1 Sélection de modèle.....	65
3.5.3 Définir les modèles et passer les données pour trainer.....	66
3.5.4 Evaluations des modèles.....	70
3.5.5 Comparaison entre les différents modèles.....	74
3.6 CONCLUSION .....	755
CONCLUSION GÉNÉRALE ET PERSPECTIVES.....	77

# Liste des Figures

<b>Chapitre I: Les systèmes de santé intelligents et le diabète</b> .....	3
<b>Figure 1.1 :</b> Composants d'un SADM.....	6
<b>Figure 1.2:</b> Qu'est ce que le diabète.....	8
<b>Figure 1.3:</b> Complications du diabète .....	9
<b>Figure 1.4 :</b> Glucomètre .....	11
<b>Figure 1.5 :</b> Lecteur de la glycémie.....	14
<b>Figure 1.6:</b> Le diabète de type 2.....	17
<b>Figure 1.7:</b> Traitement du diabète type 2 .....	18
<b>Chapitre II : L'apprentissage automatique &amp; Approfondi</b> .....	22
<b>Figure 2. 1:</b> Les classes d'apprentissage automatique .....	24
<b>Figure 2. 2:</b> Exemple d'apprentissage supervisé classification et régression .....	25
<b>Figure 2. 3:</b> Apprentissage non-supervisé.....	26
<b>Figure 2. 4:</b> Principe de l'apprentissage par renforcement .....	27
<b>Figure 2. 5:</b> Les algorithmes d'apprentissage automatique.....	28
<b>Figure 2. 6:</b> Exemple simple sur KNN.....	29
<b>Figure 2. 7:</b> Arbre de décision répondre à la question si un personne diabétique ou non ?.....	31
<b>Figure 2. 8:</b> Structure de l'algorithme random fores.....	32
<b>Figure 2. 9:</b> Un simple exemple sur le fonctionnement de l'algorithme SVM .....	33
<b>Figure 2. 10:</b> Séparation parfait de deux classes avec un hyperplan.....	34
<b>Figure 2. 11:</b> Exemple de réseau de neurones profond .....	35
<b>Figure 2. 12:</b> Architecture d'un réseau de neurones convolutif.....	36
<b>Figure 2. 13:</b> Opération de convolution .....	37
<b>Figure 2. 14:</b> MaxPooling .....	38
<b>Figure 2. 15:</b> Les Trois fonctions d'activation.....	39
<b>Figure 2. 16:</b> Fully connected layer.....	40
<b>Figure 2. 17:</b> l'architecture de réseau récurrent .....	41
<b>Figure 2. 18:</b> Réseau de neurones biologique vs réseau de neurones artificielles .....	41
<b>Figure 2. 19:</b> Schéma d'un élément de traitement unique (PE) .....	42
<b>Figure 2. 20:</b> Représentation graphique d'un perceptron multicouche .....	42
<b>Figure 2. 21:</b> Artificial neuron .....	43
<b>Figure 2. 22:</b> réseaux antagonistes génératifs (GAN) .....	44
<b>Figure 2. 23:</b> L'architecture de LSTM.....	45
<b>Figure 2. 24:</b> Schéma fonctionnel du réseau VGG-16 .....	46
<b>Chapitre III :Contribution et résultats</b> .....	49
<b>Figure 3. 1:</b> Architecture du cloud computing .....	54
<b>Figure 3. 2:</b> Google Colab.....	54
<b>Figure 3. 3:</b> Données d'entrainement .....	59
<b>Figure 3. 4:</b> Informations de la base de données .....	60
<b>Figure 3. 5:</b> Le graphe histogramme des colonnes.....	61
<b>Figure 3. 6:</b> Après le remplacement des valeurs zéro .....	62

<b>Figure 3. 7:</b> Matrice de corrélation.....	63
<b>Figure 3. 8:</b> Proportion des variables pour les personnes diabétiques et non diabétique.....	64
<b>Figure 3. 9:</b> Aperçu la division des données .....	65
<b>Figure 3. 10:</b> Définir les (04) modèles .....	66
<b>Figure 3. 11:</b> Définition du modèle DNN .....	67
<b>Figure 3. 12:</b> Création de modèle LSTM .....	68
<b>Figure 3. 13:</b> Compilation du modèle .....	69
<b>Figure 3. 14:</b> Entraîner Le modèle .....	70
<b>Figure 3. 15:</b> évaluation de modèle Decision Trees .....	71
<b>Figure 3. 16:</b> évaluation de modèle DNN .....	72
<b>Figure 3. 17:</b> Plote de modèle accuracy et loss .....	72
<b>Figure 3. 18:</b> évaluation de modèle SVM .....	73
<b>Figure 3. 19:</b> évaluation de modèle LSTM .....	73
<b>Figure 3. 20:</b> Plote de modèle accuracy et loss .....	74
<b>Figure 3. 21:</b> Les résultats des attributs d'evaluations pour les différents modèles .....	74
<b>Figure 3. 22:</b> Représentation graphique de Accuracy,F1 score,precision et recall .....	74
<b>Figure 3. 23:</b> Plot de Comparaison entre les différents modèle .....	75

### Liste des tableaux

<b>Tableau 2. 1:</b> Résumé des travaux techniques de Deep Learning.....	47
<b>Tableaux 3. 1:</b> Description des variables d'ensemble de données.....	58

## Liste des abréviations

<b>DL:</b>	Deep Learning.
<b>ML:</b>	Machines Learning.
<b>IA:</b>	L'intelligence artificielle .
<b>SAD:</b>	Système d'Aide à la Décision .
<b>SADM:</b>	Systèmes d'Aide à la Décision Médicale.
<b>OMS:</b>	Organisation mondiale de la santé.
<b>KNN:</b>	K nearest neighbors.
<b>DT:</b>	Decision Trees.
<b>RF:</b>	Random Forest.
<b>SVM :</b>	Support Vector Machine.
<b>CNN :</b>	Convolution Neural Network.
<b>RELU :</b>	Rectified Linear Unit.
<b>RNN :</b>	Réseaux neuronaux récurrents.
<b>ANN :</b>	Artificial Neural Networks.
<b>DNN :</b>	Deep Neural Network.
<b>GAN :</b>	Generative Adversarial Network.
<b>LSTM :</b>	Long Short-Term Memory.
<b>MLP :</b>	Multi-Layer Perceptron's.

# Introduction générale

## **INTRODUCTION GÉNÉRALE**

Les systèmes de santé intelligents, soutenus par l'intelligence artificielle (IA), ont le potentiel de transformer la manière dont nous abordons les soins de santé. Grâce aux avancées technologiques, l'IA peut être utilisée pour améliorer la gestion des maladies chroniques telles que le diabète.

Le diabète est une maladie chronique qui touche des millions de personnes dans le monde entier. Il est devenu un problème de santé majeur, nécessitant une gestion adéquate pour prévenir les complications potentiellement graves associées à cette maladie. L'intégration de l'intelligence artificielle et plus spécifiquement du deep learning dans le domaine médical offre de nouvelles perspectives pour améliorer la détection précoce, le diagnostic précis et la prédiction du diabète.

Ce mémoire se concentre sur l'application du deep learning et de machine learning pour le développement d'un système de détection et de prédiction du diabète. L'objectif principal de cette recherche est d'explorer comment les avancées récentes de l'intelligence artificielle peuvent être exploitées pour améliorer la prise en charge des patients diabétiques, en fournissant des outils d'analyse et de prédiction plus précis et efficaces. Elle vise à :

- Aider les patients et les professionnels de santé pour prédire la maladie du diabète (en utilisant les algorithmes d'apprentissage profond et automatique), grâce aux symptômes qui apparaissent sur le patient .

- Travailler sur un dataset des maladies de diabète, Cela permet à la machine intelligente de donner les meilleurs résultats en s'entraînant sur des données plus importantes.

- Evaluer plusieurs modèles d'apprentissage et de choisir le meilleur modèle pour faire la prédiction de diabète.

Notre problématique face à la prévalence croissante du diabète et aux difficultés croissantes de sa prise en charge, il est essentiel de développer des outils et des méthodes efficaces pour prévenir la survenue et les complications de cette maladie. Les professionnels de la santé recommandent aux individus de suivre certaines consignes pour réduire les facteurs de risque associés au diabète. Cependant, avec l'augmentation du nombre de patients diabétiques, il devient de plus en plus difficile de contrôler la maladie de manière efficace.

Cependant, malgré les avancées réalisées dans ce domaine, certaines questions demeurent : comment les techniques de deep learning peuvent-elles être appliquées de manière optimale pour la détection et la prédiction du diabète ? Quels sont les facteurs de risque les plus significatifs qui doivent être pris en compte ? Comment ces systèmes peuvent-ils contribuer à améliorer la prise en charge des patients diabétiques et à réduire les complications associées à la maladie ?

La problématique de ce mémoire consiste donc à étudier l'application du deep learning et de machine learning dans le développement d'un système de détection et de prédiction du diabète, en se concentrant sur l'identification des facteurs de risque les plus importants. L'objectif est de déterminer comment ces techniques peuvent être utilisées de manière optimale pour améliorer la prise en charge du diabète, en fournissant des outils de prévention et de gestion plus précis et efficaces.

La méthode utilisée dans ce travail est l'application des différents algorithmes du deep learning (Deep Neural Network (DNN) et Les réseaux neuronaux de longue mémoire à court terme (LSTM)) et en utilisant les 2 classifieurs de Machine learning (Decision Trees, Support Vector Machine (SVM)). Les données utilisées dans notre étude ont été extraites de l'hôpital de Frankfurt en Allemagne et ont été obtenues à partir de la plateforme Kaggle. Notre objectif était de déterminer le meilleur algorithme qui donnera comme résultat une classification des patients en termes de taux de précision et de la sensibilité du modèle.

Ce travail est organisé en trois principaux chapitres comme suit :

1. Dans le premier chapitre nous avons exploré les systèmes d'aide à la décision clinique et examiné les différentes formes de diabète, les facteurs de risque, les symptômes, les méthodes de diagnostic, les traitements et les mesures de prévention. Nous avons mis en évidence l'utilisation de l'intelligence artificielle pour améliorer la gestion du diabète et les résultats cliniques des patients.

2. Le deuxième chapitre nous présente les algorithmes d'apprentissage automatique et les techniques de deep learning pour détecter l'apparition précoce du diabète. Il examine les travaux de recherche existants sur l'application de ces techniques pour la prédiction du diabète, en mettant l'accent sur le machine learning et le deep learning.

3. Le dernier chapitre nous avons décrit notre contribution spécifique à ce domaine. On présente les outils, les bibliothèques et les langages de programmation utilisés, ainsi que les étapes de prétraitement des données. On évalue plusieurs modèles, tels que les arbres de décision, les réseaux de neurones profonds (DNN), les machines à vecteurs de support (SVM) et Les réseaux neuronaux de longue mémoire à court terme (LSTM), pour la prédiction du diabète, en comparant leur précision et leur performance.

A la fin, ce travail se termine par une conclusion générale et des perspectives.

# Chapitre I

Les systèmes de santé intelligents  
et le diabète

## **1.1 INTRODUCTION**

L'intelligence artificielle en médecine est l'utilisation de modèles d'apprentissage automatique pour examiner les données médicales et générer des idées pour améliorer la santé. Grâce aux récents progrès de l'informatique et des sciences de l'information, l'intelligence artificielle est devenue une partie intégrante de la médecine moderne.

Les systèmes d'aide à la décision clinique (SAD) sont des outils informatiques qui aident les professionnels de santé dans leur prise de décision clinique en fournissant des informations pertinentes et des conseils personnalisés. Ces systèmes utilisent souvent des algorithmes pour analyser les données de santé et fournir des recommandations en matière de diagnostic, de traitement et de gestion des patients. Les SAD sont largement utilisés dans le domaine de la médecine pour améliorer la qualité des soins et réduire les erreurs cliniques.

Le diabète est un problème de santé mondial, avec une prévalence croissante dans tous les pays, en particulier dans les pays à revenu faible et intermédiaire. Selon l'Organisation mondiale de la santé, le nombre de personnes atteintes de diabète dans le monde est passé de 108 millions en 1980 à 422 millions en 2014.

Dans ce chapitre de recherche, on présente les différents systèmes d'aide à la décision clinique. Nous allons examiner les différentes formes de diabète, les facteurs de risque, les symptômes, les méthodes de diagnostic, les traitements et les préventions disponibles pour le diabète. En fin de compte, on montre comment l'utilisation de systèmes d'intelligence artificielle pour améliorer la gestion du diabète peut contribuer à améliorer les résultats cliniques pour les patients, ainsi qu'à réduire les coûts et la charge sur le système de santé dans son ensemble.

## **1.2 INTELLIGENCE ARTIFICIELLE EN MEDICINE**

En médecine, la recherche en intelligence artificielle (IA) se concentre de plus en plus sur l'application des techniques d'apprentissage automatique.

L'IA est devenue un outil précieux dans le domaine médical en raison de sa capacité à améliorer la précision et l'efficacité des diagnostics. Les algorithmes d'IA peuvent analyser de grandes quantités de données sur les patients et identifier des modèles et des tendances qui seraient difficiles à repérer pour les médecins. Cette analyse approfondie peut aider les médecins à effectuer des diagnostics plus précis et plus rapides, ce qui peut finalement conduire à de meilleurs résultats pour

les patients. De plus, l'IA peut automatiser les tâches fastidieuses et chronophages des médecins, telles que la saisie de données et l'analyse de routine, ce qui permet aux médecins de se concentrer sur des tâches plus complexes et importantes. Les algorithmes d'IA peuvent également examiner une grande quantité de littérature médicale et identifier des articles et des résultats pertinents pour un patient ou un état médical particulier, ce qui permet aux médecins de rester informés des derniers traitements et approches et de fournir les meilleurs soins possibles à leurs patients. Le monde médical s'intéresse principalement à l'IA dans quatre de ses domaines d'application : surveillance, prédiction, diagnostic et médecine personnalisée [1].

### **1.2.1 Systèmes d'Aide à la Décision (SADs)**

Un Système d'Aide à la Décision (SAD) est l'ensemble des outils informatiques (matériels et logiciels) qui permettent l'analyse des données opérationnelles issues du système d'information des organisations. Ces SADs sont présents dans de nombreux domaines et ont pour objectif d'aider le décideur dans sa tâche en lui fournissant tous les éléments pertinents pour la prise de décision. Cette dernière est devenue récemment un thème actif et important dans l'extraction et la gestion des connaissances. Cependant, les modèles de décision traditionnels sont confrontés à d'énormes problèmes tels que l'hétérogénéité des données, leur analyse sémantique, leur représentation multidimensionnelle, le problème de personnalisation et celui de la visualisation. Pour résoudre ces problèmes, il serait intéressant de proposer une approche traitant tous les types de données, permettant leurs analyses sémantiques et leurs représentations multidimensionnelles. Ces données répondant aux besoins du décideur, sont transformées en une vision orientée décideur puis présentées au moyen d'une interface graphique significative basée sur une technique métaphorique 3D [2] .

Les systèmes d'aide à la décision (SAD) sont des outils informatiques conçus pour aider les professionnels de la santé dans la prise de décisions complexes liées au diagnostic, à la thérapie et à la gestion de la santé des patients. Les SAD peuvent être utilisés dans diverses situations, notamment pour:

- **Le diagnostic:** peuvent aider les médecins à établir un diagnostic plus précis et rapide en analysant les symptômes et les antécédents médicaux du patient.
- **La prescription:** les SAD peuvent aider à recommander le traitement optimal pour un patient en utilisant des algorithmes basés sur des données cliniques, des protocoles de traitement et des preuves scientifiques.

- **La gestion de la santé:** surveille les maladies chroniques en fournissant des recommandations sur la fréquence et les types de tests et de suivis nécessaires.

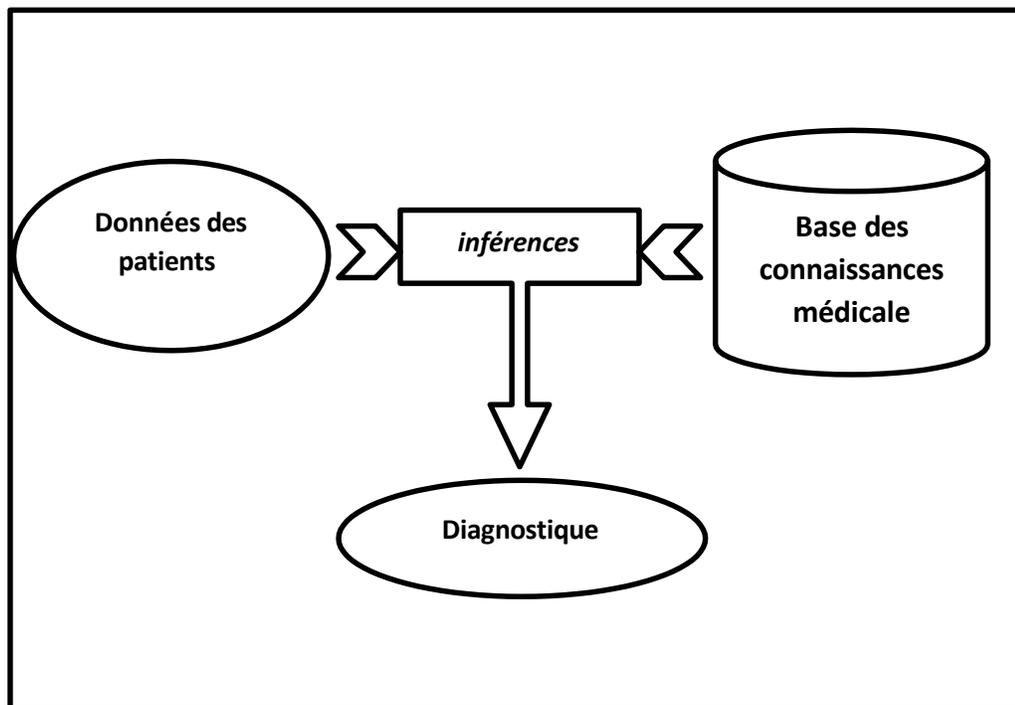
L'importance des SADs dans le domaine médical réside dans leur capacité à améliorer la qualité des soins, à réduire les erreurs de diagnostic et de traitement, à améliorer l'efficacité des soins de santé, à réduire les coûts et à améliorer la sécurité des patients. Les SAD peuvent également aider à réduire le fardeau des professionnels de la santé en leur fournissant des informations précieuses pour leur permettre de prendre des décisions éclairées en un temps record.

Cependant, il est important de noter que les SADs ne doivent jamais remplacer les professionnels de la santé, mais plutôt les aider dans leur travail en fournissant des informations et des recommandations basées sur des preuves scientifiques solides. Les SADs sont un complément précieux à l'expertise des professionnels de la santé, mais ils ne doivent pas être utilisés comme une solution unique pour le diagnostic et le traitement médical [3] .

### **1.2.2 Les systèmes d'aide au diagnostic médical (SADMs)**

Dans le domaine médical, les Systèmes d'Aide à la Décision Médicale (SADM) sont "des applications informatiques dont le but est de fournir aux cliniciens en temps et lieux utiles les informations décrivant la situation clinique d'un patient ainsi que les connaissances appropriées à cette situation, correctement filtrées et présentées afin d'améliorer la qualité des soins et la santé des patients [4] .

Les Systèmes d'Aide à la Décision Médicale (SADM) sont typiquement constitués de deux parties: une base de connaissances médicales et un moteur d'inférence. La base de connaissances est élaborée à partir des symptômes des maladies, des relations entre ces symptômes et des relations entre les différentes maladies. Le moteur d'inférence, quant à lui, se sert de ces connaissances médicales pour analyser les données des patients et prendre des décisions éclairées. La figure 1.1 illustre les différents composants logiques d'un SADM [5].



*Figure 1.1 : Composants d'un SADM*

Les Systèmes d'Aide au Diagnostic Médical (SADMs) sont des outils informatiques conçus pour aider les professionnels de la santé dans le processus de diagnostic en fournissant des suggestions et des recommandations de diagnostic basées sur les symptômes et les antécédents médicaux du patient. Les SADM peuvent être utilisés dans diverses situations, notamment pour:

- **L'analyse des symptômes:** les SADM peuvent aider à analyser les symptômes présentés par le patient et fournir une liste de maladies possibles qui pourraient causer ces symptômes.
- **L'interprétation des tests:** les SADM interprètent les résultats des tests diagnostiques tels que les radiographies, les scanners, les échographies, etc.
- **L'identification des maladies rares:** les SADM aident les médecins à identifier les maladies rares qui peuvent être difficiles à diagnostiquer en raison de leur rareté.

L'importance des SADM dans le domaine médical réside dans leur capacité à améliorer la précision du diagnostic, à réduire les erreurs de diagnostic, à réduire les coûts et à améliorer la sécurité des patients. Les SADM peuvent également aider les médecins à gagner du temps en leur fournissant des suggestions de diagnostic plus rapidement que s'ils devaient tout faire manuellement.

## **2. LE DIABETE**

### **2.1. DEFINITION DE DIABETE**

#### **Compréhension commune de la définition du diabète:**

- Le diabète est une maladie chronique caractérisée par un taux élevé de glucose dans le sang. Le glucose provient de la nourriture que nous consommons. L'insuline est une hormone produite par le pancréas qui aide à réguler la quantité de glucose dans le sang en permettant au glucose de pénétrer dans les cellules pour être utilisé comme source d'énergie. En cas de diabète, le corps ne produit pas suffisamment d'insuline ou ne peut pas utiliser efficacement l'insuline qu'il produit, ce qui entraîne une accumulation de glucose dans le sang.

#### **Définitions de sources médicales:**

- Selon l'Organisation mondiale de la santé (OMS), le diabète est défini comme "un état chronique caractérisé par une concentration élevée de glucose dans le sang (hyperglycémie) résultant soit d'une insuffisance d'insuline sécrétée par le pancréas, soit de l'incapacité des cellules à utiliser l'insuline sécrétée de manière efficace, soit des deux à la fois".[6]
- L'American Diabetes Association (ADA) définit le diabète comme "un groupe de maladies caractérisées par une hyperglycémie due à des défauts de la sécrétion d'insuline, de l'action de l'insuline, ou les deux".[7]
- Le National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) définit le diabète comme "un groupe de maladies métaboliques caractérisées par une hyperglycémie résultant de l'insuffisance de la production d'insuline, de la résistance à l'insuline ou des deux".[8]

#### **Définitions de sources non-médicales:**

- Le diabète est défini comme "une maladie chronique qui affecte la façon dont le corps traite les aliments et les transforme en énergie" .[9]
- Le diabète est également décrit comme "une maladie qui survient lorsque le pancréas ne produit pas suffisamment d'insuline ou lorsque le corps ne peut pas utiliser efficacement l'insuline qu'il produit".[10]

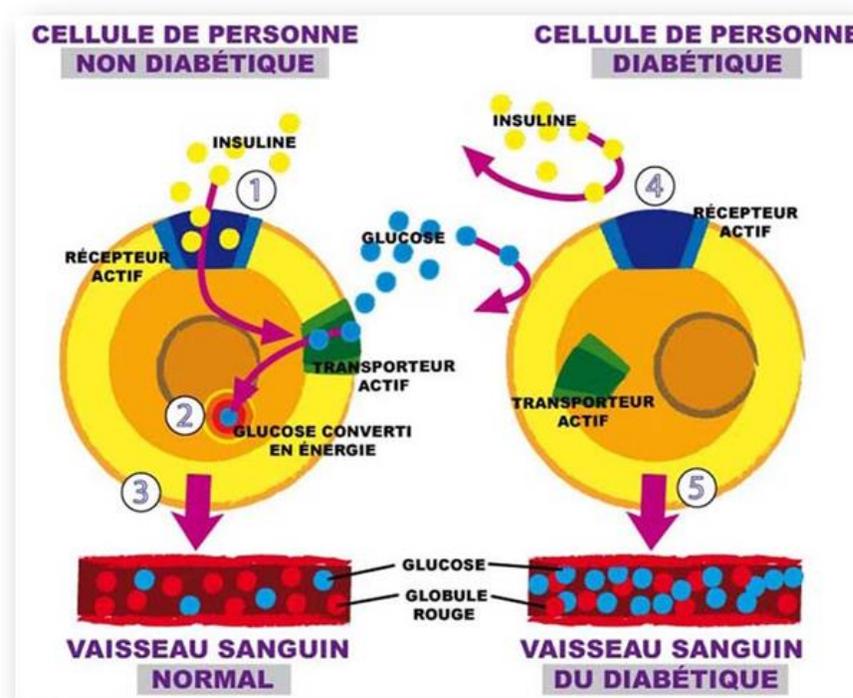


Figure 1.2: Qu'est ce que le diabète

## 2.2. DIABÈTES ET COMPLICATIONS

Le diabète est une maladie qui peut avoir des conséquences graves sur la santé, notamment si elle n'est pas traitée ou contrôlée de manière adéquate. Les complications liées au diabète peuvent toucher plusieurs organes et systèmes du corps. Voici une liste des principales complications du diabète :

- Les maladies cardiovasculaires: Les personnes atteintes de diabète sont deux à quatre fois plus susceptibles de développer une maladie cardiovasculaire que les personnes sans diabète. Les maladies cardiovasculaires liées au diabète comprennent la maladie coronarienne, l'insuffisance cardiaque, l'hypertension artérielle et les accidents vasculaires cérébraux.
- Les troubles de la vue: Le diabète peut endommager les petits vaisseaux sanguins de la rétine, provoquant une rétinopathie diabétique, qui peut entraîner une perte de vision ou même la cécité.
- Les maladies rénales: Les dommages aux petits vaisseaux sanguins des reins peuvent entraîner une insuffisance rénale et nécessiter une dialyse ou une transplantation rénale.

- Les troubles nerveux: Le diabète peut endommager les nerfs dans tout le corps, provoquant des douleurs, des picotements, des engourdissements et des faiblesses musculaires.
- Les problèmes de pieds: Les lésions nerveuses et les problèmes de circulation sanguine associés au diabète peuvent causer des ulcères et des infections du pied, qui peuvent finalement entraîner une amputation. [12]

Selon l'Organisation mondiale de la santé, le diabète est responsable de plus de 4 millions de décès chaque année dans le monde [11]. En France, environ 3,3 millions de personnes vivent avec le diabète, ce qui représente environ 5 % de la population totale. Selon une étude de l'Assurance maladie, le coût total du diabète en France est estimé à plus de 10 milliards d'euros par an, ce qui en fait l'une des maladies les plus coûteuses en termes de traitement médical et de perte de productivité. [13]



*Figure 1.3: Complications du diabète*

### 2.3. SYMPTOMES DE DIABETE

Les symptômes du diabète varient selon le type de diabète et la gravité de la maladie. Les symptômes courants du diabète de type 1 et du diabète de type 2 comprennent:

1. **Polydipsie:** une soif excessive. Les personnes atteintes de diabète peuvent avoir besoin de boire beaucoup plus d'eau que d'habitude.
2. **Polyurie:** augmentation de la miction, ou besoin fréquent d'uriner. Les personnes atteintes de diabète peuvent avoir besoin d'uriner plusieurs fois par jour, même la nuit.
3. **Polyphagie:** une faim excessive. Les personnes atteintes de diabète peuvent avoir besoin de manger plus souvent et en plus grande quantité que d'habitude.
4. **Fatigue:** une fatigue générale peut survenir en raison d'une incapacité à utiliser le glucose comme source d'énergie.
5. **Vision floue:** le diabète peut affecter la vision et causer des problèmes de vue, tels que la vision floue.
6. **Perte de poids involontaire:** les personnes atteintes de diabète de type 1 peuvent perdre du poids même si elles mangent plus que d'habitude. Cela est dû à une incapacité du corps à utiliser le glucose comme source d'énergie et à la dégradation des graisses et des protéines pour compenser.
7. **Engourdissement ou picotement dans les mains et les pieds:** le diabète peut causer des dommages aux nerfs, ce qui peut entraîner une perte de sensation ou une sensation de picotement dans les mains et les pieds.
8. **Infections fréquentes:** les personnes atteintes de diabète sont plus susceptibles de développer des infections, en particulier des infections des voies urinaires et des infections fongiques. [14] [15]

Il est important de noter que certaines personnes atteintes de diabète peuvent ne pas présenter de symptômes du tout, en particulier dans les premiers stades de la maladie. C'est pourquoi il est important de subir des tests de dépistage réguliers si vous présentez des facteurs de risque de diabète. [16]

## 2.4. DIAGNOSTIC DE DIABETE

Le diagnostic du diabète implique généralement des tests sanguins pour mesurer la glycémie. Les tests de glycémie mesurent la quantité de glucose dans le sang à un moment précis. Les tests de glycémie couramment utilisés pour diagnostiquer le diabète comprennent :

1. **Test de glycémie à jeun** : Ce test mesure la glycémie après une période de jeûne d'au moins huit heures. Si la glycémie est supérieure à 126 mg/dL (7,0 mmol/L), le test doit être répété pour confirmer le diagnostic.
2. **Test de tolérance au glucose** : Ce test mesure la glycémie avant et après la consommation d'une boisson sucrée contenant 75 grammes de glucose. Si la glycémie est supérieure à 200 mg/dL (11,1 mmol/L) deux heures après la consommation de la boisson, le test confirme un diagnostic de diabète.
3. **Test de glycémie aléatoire** : Ce test mesure la glycémie à un moment quelconque de la journée, sans exigence de jeûne. Si la glycémie est supérieure à 200 mg/dL (11,1 mmol/L) et que d'autres symptômes du diabète sont présents, le test confirme un diagnostic de diabète.

Il y a aussi des tests supplémentaires peuvent être nécessaires pour confirmer un diagnostic de diabète, notamment un test d'hémoglobine glyquée (HbA1c) qui mesure la glycémie moyenne sur une période de 2 à 3 mois. [17] [18]



*Figure 1.4 :Glucomètre*

### 2.4.1. Qui est concerné par le dépistage diabétique ?

Le dépistage du diabète est recommandé pour toutes les personnes ayant des facteurs de risque de développer la maladie. Les facteurs de risque de diabète comprennent :

- **Antécédents familiaux de diabète** : Si un membre de la famille proche (parent, frère ou sœur) a été diagnostiqué avec le diabète, il est recommandé de subir un dépistage régulier.
- **Surpoids ou obésité** : Les personnes en surpoids ou obèses sont plus à risque de développer le diabète de type 2.
- **Sédentarité** : Le manque d'exercice physique peut augmenter le risque de diabète.
- **Âge** : Le risque de développer le diabète augmente avec l'âge. Les personnes de plus de 45 ans sont plus susceptibles d'avoir le diabète de type 2.
- **Origine ethnique** : Les personnes d'origine africaine, hispanique, amérindienne, asiatique ou des îles du Pacifique sont plus à risque de développer le diabète de type 2.
- **Hypertension artérielle** : Les personnes ayant une pression artérielle élevée sont plus à risque de développer le diabète.
- **Niveaux de cholestérol anormaux** : Les personnes ayant des niveaux de cholestérol anormaux sont plus susceptibles de développer le diabète.

Le dépistage du diabète est particulièrement important pour les personnes ayant plusieurs facteurs de risque, ainsi que pour celles présentant des symptômes de la maladie tels que la soif excessive, la miction fréquente, la fatigue, la perte de poids involontaire et des infections récurrentes. [19] [20]

### 2.4.2. Comment savoir si l'on est diabétique ?

Le seul moyen de diagnostiquer le diabète est de passer un test de glycémie. Plusieurs types de tests peuvent être utilisés, notamment le test de glycémie à jeun, le test de tolérance au glucose et le test d'hémoglobine glyquée (HbA1c). Le test de glycémie à jeun mesure le taux de sucre dans le sang après une période de jeûne et un résultat de 126 mg/dL (7,0 mmol/L) ou plus sur deux tests différents indique un diagnostic de diabète. Le test de tolérance au glucose mesure la glycémie deux heures après

avoir bu une solution sucrée et un résultat de 200 mg/dL (11,1 mmol/L) ou plus indique un diagnostic de diabète. Le test d'HbA1c mesure la quantité d'hémoglobine glyquée dans le sang, qui est une indication du taux de sucre dans le sang au cours des deux à trois derniers mois et un résultat d'HbA1c de 6,5% ou plus indique un diagnostic de diabète. [21]

Si on présente des symptômes de diabète tels que la soif excessive, la miction fréquente, la fatigue, la perte de poids involontaire et des infections récurrentes, il est important de consulter un professionnel de la santé pour un dépistage de la maladie. Il est également important de se faire dépister régulièrement si on a des facteurs de risque de diabète, tels que des antécédents familiaux de diabète, un surpoids ou une obésité, une sédentarité, un âge avancé, une origine ethnique à risque, une hypertension artérielle ou des niveaux de cholestérol anormaux. Il est crucial de détecter le diabète tôt pour prévenir les complications à long terme de la maladie, notamment les maladies cardiovasculaires, les maladies rénales et la cécité. [22]

### **2.4.3. Décoder et comprendre les résultats de la glycémie**

Le décodage et la compréhension des résultats de la glycémie sont essentiels pour les personnes atteintes de diabète et pour celles qui souhaitent surveiller leur santé en général. Les résultats d'un test glycémique peuvent être exprimés en milligrammes de glucose par décilitre (mg/dl) ou en millimoles de glucose par litre (mmol/l). Le choix de l'unité de mesure dépend du pays où l'on se trouve. [23]

Pour interpréter les résultats d'un test glycémique, il est important de comprendre les plages de glycémie normales et anormales. Les niveaux normaux de glycémie varient généralement entre 70 et 99 mg/dl (3,9 à 5,5 mmol/l) pour une personne à jeun. Cependant, les plages de glycémie normales peuvent varier en fonction de l'âge, du sexe, de l'état de santé général, de l'activité physique et d'autres facteurs. [24] [25]

Si les résultats du test glycémique sont supérieurs à la plage normale, cela peut indiquer une hyperglycémie, qui est un signe de diabète ou d'un pré-diabète. Si les résultats du test sont inférieurs à la plage normale, cela peut indiquer une hypoglycémie, qui peut être causée par une carence en glucose ou par un traitement excessif du diabète. [26]

Il est important de noter que les résultats des tests de glycémie peuvent varier en fonction de différents facteurs tels que l'alimentation, l'exercice physique et l'état de santé général. Par conséquent, il est crucial de suivre les instructions de votre professionnel de la santé pour garantir des résultats

précis et de discuter de tout résultat anormal avec votre médecin pour évaluer les options de traitement appropriées. [27]

#### 2.4.4. A quelle fréquence dois-je contrôler ma glycémie ?

La fréquence à laquelle on doit contrôler la glycémie dépend du type de diabète, de la gravité de la maladie, de traitement et de l'état de santé général. En général, les personnes atteintes de diabète de type 1 doivent contrôler leur glycémie plusieurs fois par jour, car elles ont besoin d'insuline pour maintenir leur taux de sucre dans le sang dans une fourchette normale. Les personnes atteintes de diabète de type 2 peuvent également avoir besoin de contrôler leur glycémie régulièrement, en particulier si elles prennent des médicaments pour abaisser leur taux de sucre dans le sang.

Il est important de travailler avec un professionnel de la santé pour déterminer cette fréquence. En général, il est recommandé de contrôler la glycémie avant les repas et avant le coucher, ainsi que deux heures après les repas pour voir comment le corps réagit à la nourriture. Si le malade est une femme enceinte ou s'il a des complications liées au diabète, il est recommandé de contrôler votre glycémie plus fréquemment.

Il est également important de noter les résultats des tests de glycémie et de les partager avec un professionnel de la santé lors de rendez-vous de suivi. Cela lui permettra de surveiller le taux de sucre dans le sang au fil du temps et de modifier le traitement si nécessaire. En fin de compte, le contrôle régulier de la glycémie peut aider à mieux gérer le diabète et à prévenir les complications à long terme de la maladie.



*Figure 1.5 :Lecteur de la glycémie*

## 2.5. CLASSIFICATION DE DIABETE

Il existe plusieurs types de diabète, chacun ayant des causes et des caractéristiques différentes. Les principaux types de diabète sont le diabète de type 1, le diabète de type 2 et le diabète gestationnel.

### 2.5.1 Diabète de type 1

Le diabète de type 1, également connu sous le nom de diabète insulino-dépendant (DID) ou diabète juvénile, est une maladie auto-immune chronique caractérisée par une destruction des cellules bêta du pancréas qui produisent l'insuline. L'insuline est une hormone essentielle qui régule la glycémie en aidant les cellules à absorber le glucose de la circulation sanguine. En l'absence d'insuline, le glucose s'accumule dans le sang, provoquant une hyperglycémie qui peut entraîner des complications graves et potentiellement mortelles. [28]

Le diabète de type 1 représente environ 5 à 10% de tous les cas de diabète, avec une incidence de plus en plus élevée chez les enfants et les jeunes adultes. Selon les estimations de l'Organisation mondiale de la santé (OMS), environ 1,6 million de décès sont directement attribuables au diabète chaque année, et le diabète de type 1 représente environ 10% de tous les cas de diabète dans le monde. [30] [31]

Les causes exactes du diabète de type 1 ne sont pas encore connues, mais il est généralement considéré comme une maladie multifactorielle, impliquant à la fois des facteurs génétiques et environnementaux. On pense que certaines infections virales et bactériennes peuvent déclencher une réponse auto-immune qui attaque les cellules bêta du pancréas, entraînant une diminution progressive de la production d'insuline.

Les symptômes du diabète de type 1 peuvent apparaître rapidement et sont souvent plus graves que ceux du diabète de type 2. Ils comprennent une soif excessive, une miction fréquente, une perte de poids, une fatigue, une vision floue, une faim excessive et des infections fréquentes de la peau, de la vessie et des gencives. Si ces symptômes sont présents, il est important de consulter immédiatement un professionnel de la santé pour un diagnostic et un traitement précoces. [32]

Le traitement du diabète de type 1 consiste en une supplémentation en insuline pour réguler la glycémie et prévenir les complications à long terme, telles que les maladies cardiovasculaires, les

neuropathies, les troubles oculaires et rénaux. Les personnes atteintes de diabète de type 1 doivent surveiller régulièrement leur glycémie, suivre un régime alimentaire équilibré, faire de l'exercice régulièrement et éviter les facteurs de risque tels que le tabagisme. [29]

La recherche dans le domaine du diabète de type 1 est en constante évolution, avec des efforts pour développer des traitements plus efficaces, des outils de surveillance de la glycémie améliorés, des thérapies géniques et des traitements qui pourraient inverser ou prévenir la progression de la maladie. Cependant, il est important de souligner que la prévention reste la clé pour réduire le fardeau du diabète de type 1, en particulier chez les enfants et les jeunes adultes.

### **2.5.2 Diabète de type 2**

Le diabète de type 2 est un trouble métabolique chronique caractérisé par des taux élevés de glucose dans le sang résultant de l'incapacité du corps à utiliser efficacement l'insuline. Il représente environ 90 % de tous les cas de diabète dans le monde, ce qui en fait la forme la plus courante de la maladie. Contrairement au diabète de type 1, le diabète de type 2 se développe généralement à l'âge adulte et est souvent associé à des facteurs de risque modifiables. [33]

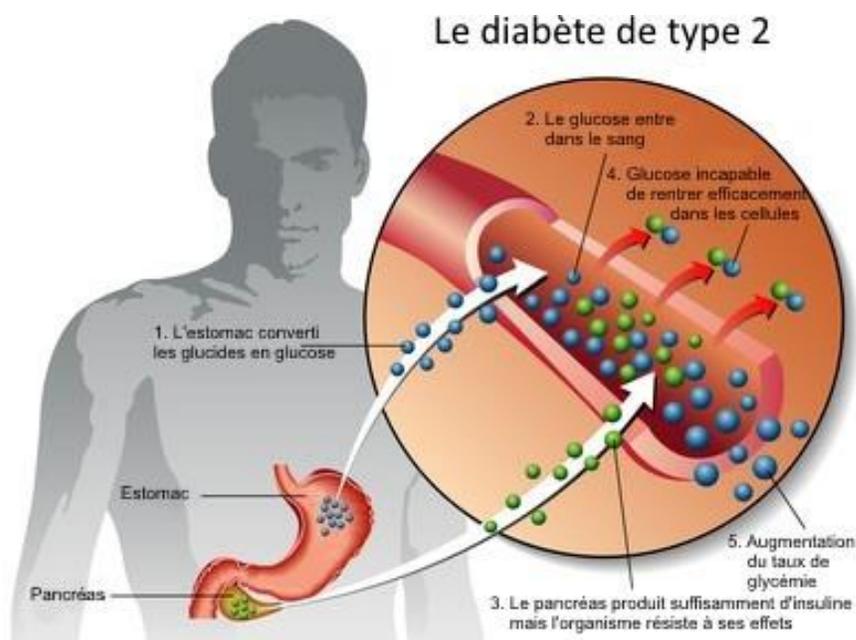
Le diabète de type 2 est fortement associé à l'obésité et au surpoids, ainsi qu'à un manque d'activité physique régulière et à une mauvaise alimentation. Selon l'Organisation mondiale de la santé (OMS), environ 1,9 milliard d'adultes dans le monde sont en surpoids, dont plus de 650 millions sont obèses. Ces chiffres augmentent régulièrement, entraînant une augmentation parallèle du nombre de personnes atteintes de diabète de type 2. [34]

Les statistiques montrent que le nombre de personnes atteintes de diabète de type 2 est en constante augmentation. En 2020, il y avait environ 463 millions de personnes atteintes de diabète dans le monde, et ce chiffre devrait atteindre 578 millions d'ici 2030 et 700 millions d'ici 2045, selon l'International Diabetes Federation (IDF). Le diabète de type 2 représente environ 90 % de tous les cas de diabète, avec une prévalence plus élevée dans les pays à revenu élevé. [36]

Les symptômes du diabète de type 2 peuvent être similaires à ceux du diabète de type 1, notamment une soif excessive, une miction fréquente et une fatigue. Cependant, le diabète de type 2 peut également être asymptomatique pendant une période prolongée, ce qui peut retarder le diagnostic

et le traitement. Par conséquent, il est essentiel de connaître les facteurs de risque de la maladie, notamment l'obésité, l'hypertension artérielle et les antécédents familiaux de diabète, afin de se faire dépister régulièrement. [35]

La prévention du diabète de type 2 repose sur une combinaison de changements de mode de vie sains, notamment une alimentation équilibrée et riche en fibres, une activité physique régulière, une gestion du poids appropriée et l'arrêt du tabac. Les personnes atteintes de pré-diabète peuvent également bénéficier d'un programme de modification du mode de vie pour prévenir ou retarder le développement du diabète de type 2.



*Figure 1.6:Le diabète de type 2*

### 2.5.3. Diabète gestationnel

Le diabète gestationnel est une affection qui se développe chez les femmes enceintes qui n'ont pas été diagnostiquées auparavant comme diabétiques. Cette maladie survient généralement vers la fin de la grossesse et peut avoir des conséquences graves pour la mère et le bébé si elle n'est pas traitée.

- **Causes:** Le diabète gestationnel est causé par une résistance à l'insuline, qui est provoquée par les hormones produites par le placenta pendant la grossesse. Ces hormones peuvent bloquer l'action de l'insuline dans le corps de la mère, ce qui entraîne une augmentation du taux de sucre dans le sang.

- **Facteurs de risque:** Les femmes qui ont des antécédents familiaux de diabète, qui ont un indice de masse corporelle élevé, qui ont déjà eu un bébé pesant plus de 4 kg à la naissance, ou qui ont eu un diabète gestationnel lors d'une grossesse précédente ont un risque plus élevé de développer un diabète gestationnel.
- **Symptômes:** Les symptômes du diabète gestationnel sont souvent légers et peuvent passer inaperçus. Cependant, certains signes peuvent inclure une soif excessive, une miction fréquente, une fatigue, une vision floue et une sensation de faim.
- **Complications:** Si le diabète gestationnel n'est pas traité, il peut entraîner des complications pour la mère et le bébé. Pour la mère, cela peut inclure une hypertension artérielle, un risque accru de césarienne, une pré-éclampsie, une infection urinaire et une polyhydramnios (excès de liquide amniotique). Pour le bébé, cela peut inclure une macrosomie (bébé de poids élevé à la naissance), un risque accru de blessures à la naissance, une hypoglycémie et un risque accru de développer un diabète plus tard dans la vie.
- **Diagnostic:** Le diagnostic du diabète gestationnel se fait généralement par un test de dépistage de glucose entre la 24<sup>ème</sup> et la 28<sup>ème</sup> semaine de grossesse. Si les résultats du test de dépistage sont supérieurs à une certaine plage de glycémie, un test de tolérance au glucose plus poussé sera effectué pour confirmer le diagnostic.
- **Traitement:** Le traitement du diabète gestationnel comprend généralement un régime alimentaire sain, de l'exercice physique régulier et une surveillance étroite de la glycémie. Dans certains cas, des médicaments peuvent également être prescrits pour aider à contrôler la glycémie. Le traitement précoce et efficace du diabète gestationnel peut aider à prévenir les complications pour la mère et le bébé. [43] [45] [47]



*Figure 1.7: Traitement du diabète type 2*

Selon l'Organisation mondiale de la santé, le diabète gestationnel affecte environ 1 sur 7 naissances dans le monde. En France, il est estimé que jusqu'à 10% des femmes enceintes développent un diabète gestationnel. Les femmes qui ont eu un diabète gestationnel ont un risque accru de développer un diabète de type 2 plus tard dans leur vie. De plus, les enfants nés de mères atteintes de diabète gestationnel ont également un risque plus élevé de développer un diabète de type 2 à l'avenir. [44] [46]

## **2.6. PREVENTION DE DIABETE TYPE 2**

La prévention du diabète est un sujet crucial pour réduire le fardeau de cette maladie chronique. Il existe plusieurs mesures préventives qui peuvent aider à prévenir le diabète de type 2, qui représente environ 90% de tous les cas de diabète dans le monde. Les mesures préventives comprennent des modifications du mode de vie et des stratégies de prévention pharmacologiques pour les personnes à haut risque.

La première étape importante dans la prévention du diabète de type 2 est d'adopter un mode de vie sain, qui comprend une alimentation saine et équilibrée et une activité physique régulière. Les personnes devraient essayer de manger des aliments riches en nutriments, tels que des fruits et légumes, des céréales complètes, des protéines maigres et des graisses saines, tout en limitant les aliments transformés, les aliments riches en sucre et en graisses saturées. L'exercice physique régulier est également important pour maintenir un poids santé, réduire l'insulino-résistance et améliorer la sensibilité à l'insuline. [48]

En plus des modifications du mode de vie, il existe également des stratégies pharmacologiques pour prévenir le diabète de type 2 chez les personnes à haut risque. Le médicament le plus couramment utilisé est la metformine, qui aide à réduire le taux de glucose dans le sang et améliore la sensibilité à l'insuline. D'autres médicaments, tels que les inhibiteurs de SGLT2 et les agonistes du récepteur GLP-1, ont également été approuvés pour la prévention du diabète de type 2 chez les personnes à haut risque. [48] [50]

En outre, plusieurs programmes de prévention du diabète ont été développés dans le monde entier pour aider à réduire le risque de diabète de type 2 chez les personnes à haut risque. Ces programmes sont généralement axés sur des interventions de style de vie, telles que des programmes

de perte de poids, des conseils diététiques et de l'exercice physique, ainsi que des programmes éducatifs pour sensibiliser les personnes aux risques de diabète. [50]

Selon l'Organisation mondiale de la santé, environ 70% des décès liés au diabète surviennent dans les pays à revenu faible ou intermédiaire. Le diabète est également en constante augmentation dans le monde entier, passant de 108 millions de cas en 1980 à 422 millions de cas en 2014. Cela met en évidence l'importance d'une prévention efficace du diabète pour réduire la charge de cette maladie chronique. [49]

## **2.7 LE RÔLE DES SYSTÈMES DE SANTÉ INTELLIGENTS DANS LA GESTION EFFICACE DU DIABÈTE**

Les systèmes de santé intelligents peuvent jouer un rôle crucial dans la prise en charge du diabète. Le diabète est une maladie chronique qui nécessite une surveillance et une gestion constantes de la glycémie pour prévenir les complications à long terme. Les systèmes de santé intelligents peuvent aider les patients atteints de diabète à surveiller leur glycémie de manière plus efficace et à prendre des décisions de traitement éclairées.

Il existe plusieurs types de systèmes de santé intelligents qui peuvent être utilisés pour la gestion du diabète. Les glucomètres connectés sont des appareils qui mesurent la glycémie et envoient des données à une application sur un téléphone ou un ordinateur. Les patients et les professionnels de la santé peuvent accéder à ces données pour surveiller la glycémie et ajuster le traitement en conséquence.

Les systèmes de gestion de l'insuline intelligents sont également de plus en plus utilisés pour la gestion du diabète. Ces systèmes utilisent un capteur de glucose en continu pour surveiller la glycémie et ajuster automatiquement la dose d'insuline administrée en fonction des résultats de la glycémie.

En outre, les systèmes de santé intelligents peuvent également aider les patients atteints de diabète à surveiller leur alimentation, leur activité physique et leur poids, qui sont tous des facteurs importants dans la gestion du diabète. Les applications de santé intelligente peuvent fournir des conseils nutritionnels personnalisés et aider les patients à suivre leurs progrès en matière d'activité physique et de perte de poids.

En résumé, les systèmes de santé intelligents peuvent aider les patients atteints de diabète à surveiller leur glycémie, à ajuster leur traitement en conséquence et à gérer leur mode de vie pour prévenir les complications à long terme. Ils peuvent également aider les professionnels de la santé à

suivre les progrès des patients et à ajuster les plans de traitement en fonction des résultats de la glycémie.[51][52]

## **2.8 CONCLUSION**

En conclusion, nous avons exploré les différents types de SAD et leur fonctionnement .Nous avons discuté des généralités du diabète, notamment de sa classification, de ses symptômes et des complications potentielles associées à cette maladie. Nous avons également souligné l'importance de la surveillance régulière du diabète et de la prévention des complications pour une gestion efficace de cette maladie chronique.

Cependant, il est important de noter que les SAD ne remplacent pas les professionnels de la santé et ne doivent pas être utilisés de manière isolée pour la prise de décision médicale. Leur utilisation peut améliorer les résultats de santé des patients atteints du diabète, en aidant les professionnels de la santé à prendre des décisions éclairées et à fournir des soins de qualité supérieure.

# Chapitre II

L'apprentissage automatique  
& Approfondi

## **2.1 INTRODUCTION**

L'apprentissage automatique est un domaine de recherche qui s'applique à étudier et améliorer les systèmes informatiques qui peuvent apprendre et s'adapter à partir des données et de l'expérience. Dans le domaine de la santé, l'apprentissage automatique est utilisé pour développer des systèmes qui peuvent prédire et détecter les maladies, diagnostiquer les patients et fournir des outils pour aider les professionnels de la santé à prendre des décisions plus éclairées. Les technologies d'apprentissage automatique peuvent aider à analyser les données médicales, prédire quel traitement sera le plus efficace pour un patient donné et même diagnostiquer des maladies.

Le deep learning est un sous-domaine de l'apprentissage automatique qui se concentre sur l'utilisation de réseaux de neurones profonds pour analyser des grandes quantités de données médicales et identifier les modèles et les tendances qui peuvent indiquer la présence d'une maladie. Les algorithmes de deep learning peuvent être formés à partir de données médicales telles que les résultats de tests de laboratoire, les images d'imagerie médicale, les antécédents médicaux et les signes cliniques.

Dans ce chapitre, nous avons présenté les algorithmes d'apprentissage automatique et les techniques de deep learning qui peuvent nous aider à détecter l'apparition précoce du diabète. Et on va parler sur les travaux de recherche sur l'application d'algorithmes de machine learning et du deep learning pour prédire le diabète .

## **2.2 APPRENTISSAGE AUTOMATIQUE**

L'apprentissage automatique en anglais «Machine learning » est une forme d'intelligence artificielle qui utilise des algorithmes informatiques pour identifier des modèles dans de grands ensembles de données et peut continuellement s'améliorer avec des données supplémentaires. Les algorithmes de modèle résultants peuvent faire des prédictions sur les résultats futurs. [53]

L'apprentissage automatique implique la collecte de données, leur traitement et leur analyse pour construire un modèle mathématique qui peut être utilisé pour effectuer des prévisions ou des classifications d'autres données.

## 2.3 LES TYPES D'APPRENTISSAGE AUTOMATIQUE

Des techniques d'apprentissage automatique sont nécessaires pour améliorer la précision des modèles prédictifs. Selon la nature du problème métier à résoudre, il existe différentes approches, qui varient selon le type et la quantité de données. Ce travail comprend environ Au moins trois techniques d'apprentissage automatique différentes l'apprentissage supervisé, l'apprentissage non supervisé, l'apprentissage par renforcement.

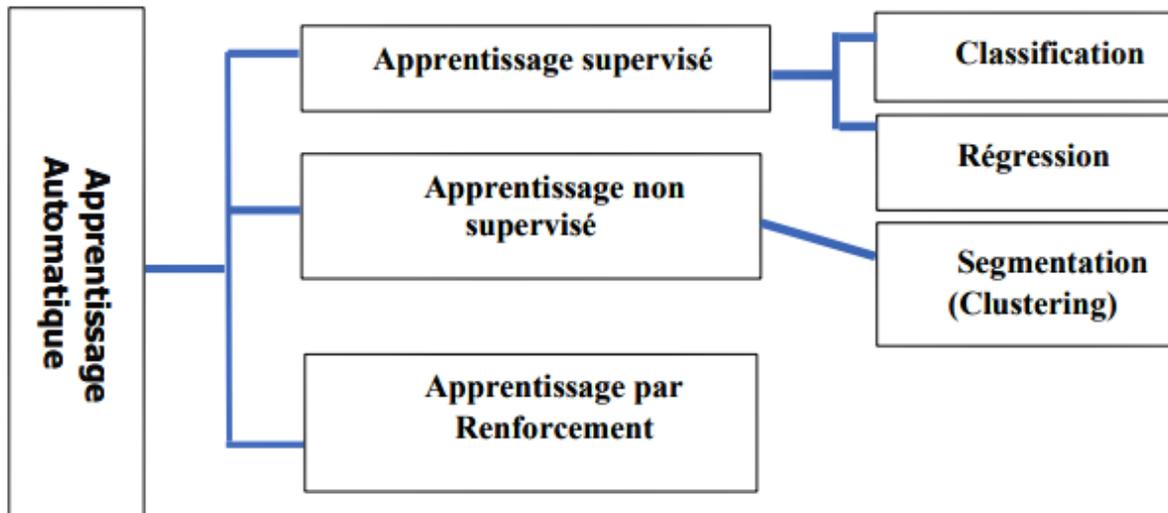


Figure 2. 1: Les classes d'apprentissage automatique

### 2.3.1 Apprentissage Supervisé

L'apprentissage supervisé, également connu sous le nom d'apprentissage automatique supervisé, est défini comme l'utilisation d'ensembles de données étiquetés pour former des algorithmes afin de classer les données ou de prédire avec précision les résultats.

L'algorithme est formé à l'aide d'une base de données d'apprentissage contenant des exemples de cas réels pour le traitement et la validation. L'objectif est de trouver des corrélations entre les données d'entrée (variables explicatives) et les données de sortie (variables à prédire), puis d'en déduire les connaissances extraites sur les entrées avec des sorties inconnues.

Chaque exemple, appelé aussi instance, est un couple d'entrée-sortie  $(x_n, y_n)$   $1 \leq n \leq N$ , avec  $x_n \in X$  et  $y_n \in Y$  et où :

**X** est l'ensemble d'attributs (discrets ou continues),

**Y** est l'ensemble des valeurs de sortie (la variable cible ou dépendante),

$Y$  peut être discrète ou continue[54]

Dans l'apprentissage supervisé, on a deux types d'algorithmes :

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

**Par exemple :**

\_Prédire le prix d'un appartement ( $y$ ) selon sa surface habitable ( $x$ )

\_ Prédire la quantité d'essence consommée ( $y$ ) selon la distance parcourue ( $x$ )

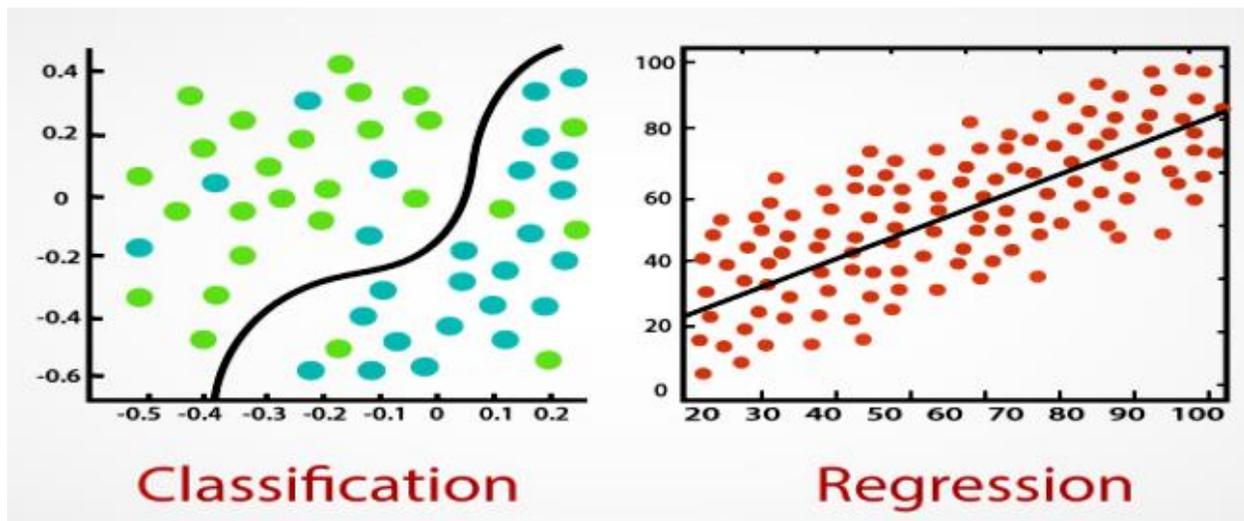
- Les algorithmes de **classification**, qui cherchent à prédire une classe/catégorie.

**Par exemple :**

\_Prédire si un email est un spam (*classe*  $y = 1$ ) ou non (*classe*  $y = 0$ ) selon le nombre de liens présent dans l'email ( $x$ )

\_ Prédire si une tumeur est maligne ( $y = 1$ ) ou bénigne ( $y = 0$ ) selon la taille de la tumeur ( $x_1$ ) et l'âge du patient ( $x_2$ )

Dans le cas d'un problème de classification, on représente souvent les classes par des symboles, plutôt que par leur valeur numérique (0, 1, ...)



*Figure 2.2: Exemple d'apprentissage supervisé classification et régression*

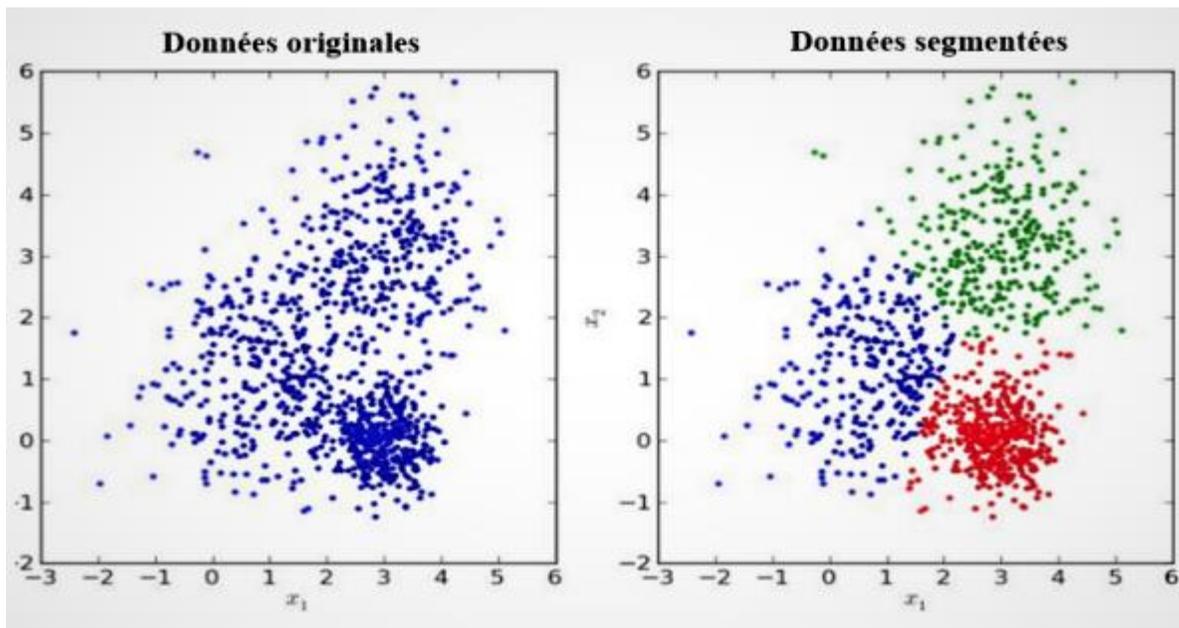
### 2.3.2 Apprentissage Non Supervisé

L'apprentissage non supervisé consiste à trouver des structures, des modèles et des motifs dans des données non étiquetées ou non catégorisées. Contrairement à l'apprentissage supervisé, il n'y a pas de réponse correcte ou de valeur de sortie connue à l'avance. Les algorithmes d'apprentissage non supervisé sont souvent utilisés pour explorer des données et découvrir des relations intéressantes entre

les variables ou pour effectuer une réduction de dimensionnalité en projetant les données dans un espace de dimension inférieure.[55]

Il existe plusieurs types d'apprentissage non supervisé :

- **Clustering** : l'objectif consiste à trouver des regroupements dans les données.
- **Association** : l'objectif consiste à identifier les règles qui permettront de définir de grands groupes de données. Les principaux algorithmes du machine learning non supervisé sont les suivants : K-Means, clustering/regroupement hiérarchique et réduction de la dimensionnalité.

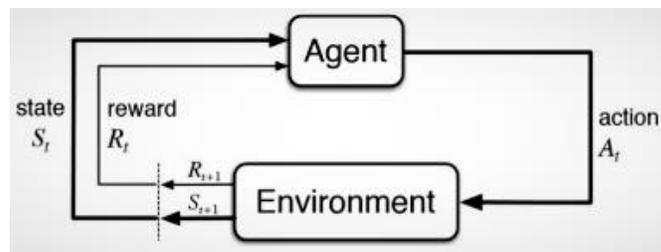


*Figure 2. 3: Apprentissage non-supervisé*

Dans le cas supervisé, la finalité est de déterminer une nouvelle sortie  $Y$  à partir d'une nouvelle entrée  $X$ , connaissant un ensemble d'observations  $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ . Lorsque les  $Y_i$  prennent des valeurs discrètes, il s'agit d'un problème de classification – en classification binaire, par exemple, on cherche à attribuer à  $X$  une étiquette 0 ou 1 –, tandis que des  $Y_i$  à valeurs réelles nous placent dans le cadre de la régression. En apprentissage non supervisé, en revanche, il n'y a pas de sortie, et il s'agit alors de construire un modèle permettant de représenter au mieux les observations  $X_1, \dots, X_n$ , de manière à la fois précise et compacte [56].

### 2.3.3 Apprentissage par renforcement

C'est un apprentissage non supervisé qui fonctionne avec un système de récompense. Au cours d'un épisode, un agent va effectuer des actions en fonction des informations en provenance d'un environnement (l'environnement correspond à l'ensemble de l'univers à l'exception de l'agent). L'agent va donc recevoir des récompenses en fonction de « l'efficacité » des actions entreprises. Personne ne dit à l'agent quelle action il doit effectuer à un instant  $t$ . Grâce à ce système de récompense, l'agent devra découvrir par lui-même les actions à entreprendre en fonction de la situation. Nous avons donc, avec l'apprentissage par renforcement, un système basé sur la « recherche par essai-erreur ». Une action qui se révélera à plus ou moins long terme peu judicieuse sera sanctionnée par une récompense négative (par exemple -1). A contrario, si l'action effectuée s'avère positive (toujours à plus ou moins long terme), l'agent recevra une récompense positive (par exemple +1). À chaque pas de temps  $t$ , l'agent reçoit une représentation de l'état de l'environnement notée  $S_t$ . L'ensemble du processus est résumé par la Figure 2.4 [64].



*Figure 2. 4: Principe de l'apprentissage par renforcement*

L'apprentissage par renforcement consiste à laisser un algorithme apprendre de ses erreurs pour atteindre un objectif. L'algorithme essaiera de nombreuses approches différentes pour tenter d'atteindre son but. En fonction de ses performances, il sera récompensé ou pénalisé pour l'inciter à poursuivre dans une voie ou à changer d'approche .

## 2.4. LES ALGORITHMES DE L'APPRENTISSAGE AUTOMATIQUE:

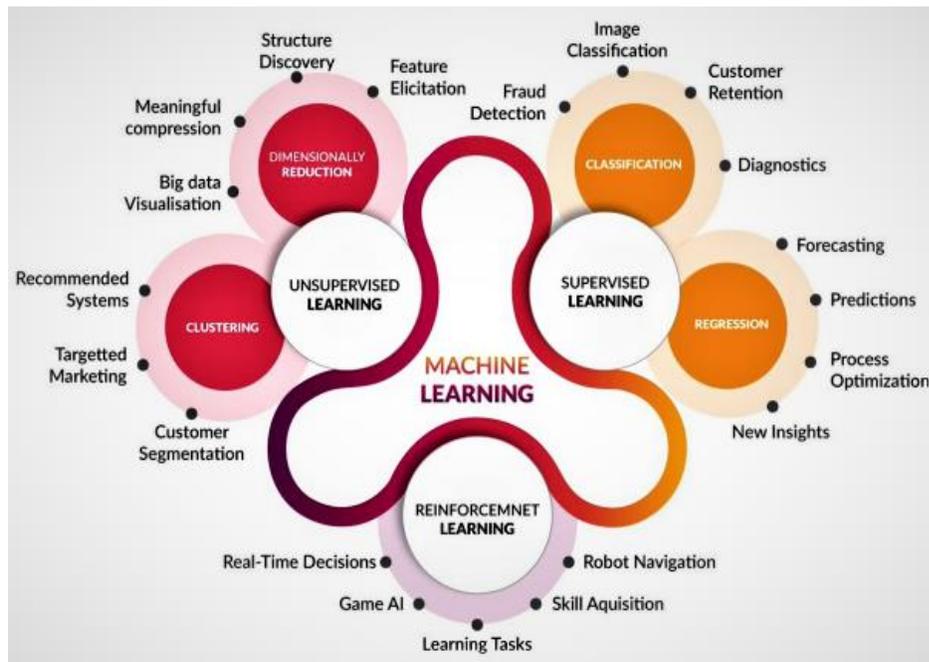
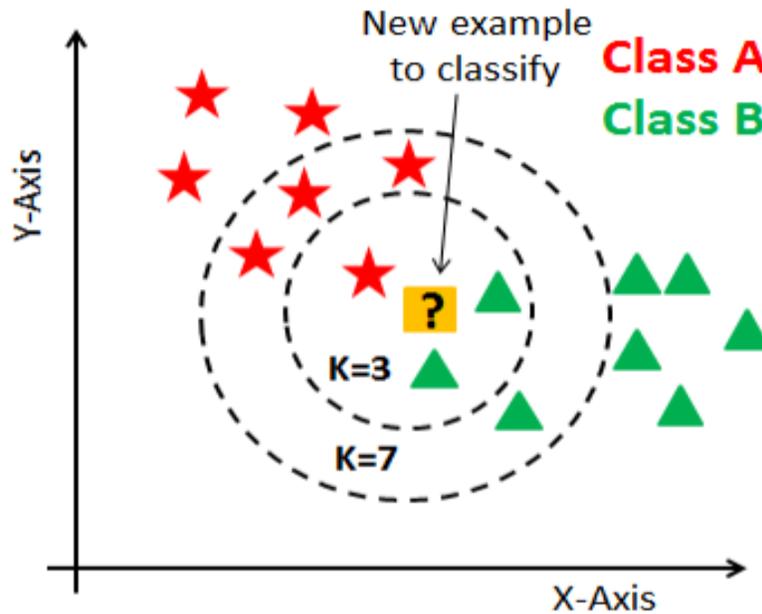


Figure 2. 5: Les algorithmes d'apprentissage automatique

### 2.4.1 K nearest neighbors (KNN)

La classification du voisin le plus proche (KNN) est une technique de classification non paramétrique inventée en 1951 par Evelyn Fix et Joseph Hodges. KNN convient à la classification ainsi qu'à l'analyse de régression. Le résultat de la classification KNN est l'appartenance à une classe. Des mécanismes de vote sont utilisés pour classer l'élément. Des techniques de distance euclidienne sont utilisées pour déterminer la distance entre deux échantillons de données. La valeur projetée dans l'analyse de régression est la moyenne des valeurs du KNN.[65]

K nearest neighbors (KNN) est un algorithme d'apprentissage automatique supervisé utilisé pour la classification et la régression. Il fonctionne en trouvant les K échantillons les plus proches d'un nouvel échantillon dans un ensemble de données et en utilisant ces échantillons pour prédire une étiquette ou une valeur pour le nouvel échantillon.



*Figure 2. 6: Exemple simple sur KNN*

Dans cet exemple, nous avons une donnée non classée, toutes les autres données sont catégorielles (étoile et triangle), et chaque donnée a sa catégorie (classe A et classe B).

— Si  $k=3$ , les données les plus proches des nouvelles données sont celles avec à l'intérieur du premier cercle, la classe la plus dominante est le triangle (classe B), puisqu'il y a 2 triangles et seulement 1 étoile, les données non classées classeront un triangle (classe B).

— Si  $k=7$  les données les plus proches des nouvelles données sont celles du deuxième cercle, la classe la plus dominante est l'étoile (classe A), puisque nous avons 4 étoiles et 3 triangles, les données non classées seront classées comme une étoile (catégorie A).

Le calcul de la proximité entre les échantillons peut être effectué en utilisant différentes mesures de distance, telles que la distance Euclidienne, la distance de Manhattan ou la distance de Minkowski. Une fois que les  $K$  échantillons les plus proches ont été identifiés, l'algorithme de KNN utilise la majorité de leurs étiquettes pour prédire l'étiquette du nouvel échantillon dans le cas d'un problème de classification, ou la moyenne de leurs valeurs dans le cas d'un problème de régression.

L'algorithme KNN est simple à comprendre et à mettre en œuvre, mais peut être sensible à la présence de valeurs manquantes ou de dimensions supplémentaires dans les données. Il est également

important de normaliser les variables d'entrée pour s'assurer que chaque variable contribue de manière égale à la distance entre les échantillons.

Il existe plusieurs types de distance parmi lesquels on trouve :

- Distance euclidienne :  $x, y$  sont des vecteur
- La distance de Minkowsky :
- $x, y$  sont des vecteurs
- $p$  : parametre

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

$$d(x, y) = (\sum_{i=1}^n |X_i - Y_i|^p)^{1/p}$$

- La distance de Manhattan :
- $x, y$  sont des vecteurs

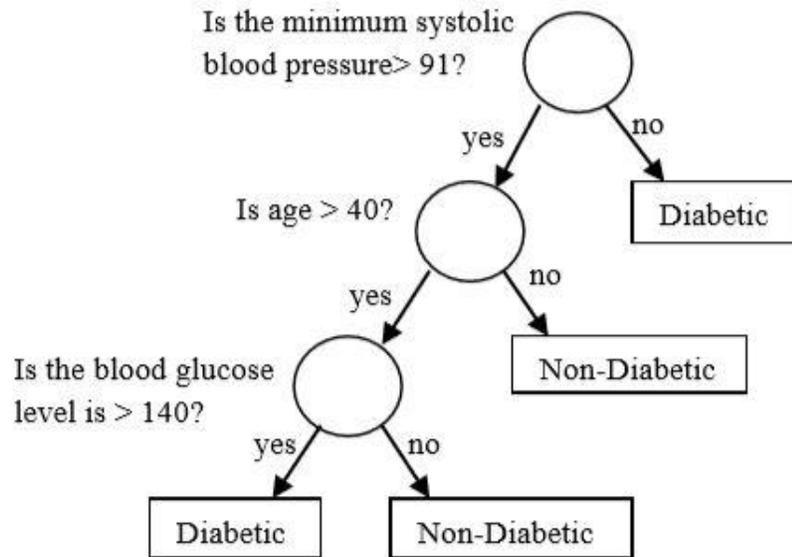
$$d(x, y) = \sum_{i=1}^k |X_i - Y_j|$$

### 2.4.2 Decision Trees (Arbre de décision)

Un arbre de décision est une technique basée sur un arbre dans laquelle tout chemin à partir de la racine est décrite par une séparation des données jusqu'à ce qu'un résultat booléen au niveau du nœud de feuille soit atteint.

Les arbres de décision sont l'un des puissants metthreshold valeur dans chaque test . Les règles conceptuelles sont beaucoup plus faciles à construire que les poids numériques dans le réseau neuronal de connexions entre nœuds . Principalement à des fins de regroupement, decision trees est utilisé.

De plus, decision trees est un modèle de classification habituellement utilisé dans l'exploration de données. Les nœuds et les branches sont composés de chaque arbre. Chaque nœud représente les caractéristiques d'une catégorie à classer et chaque sous-ensemble définit une valeur qui peut être prise par le nœud . En raison de leur analyse simple et de leur précision sur plusieurs formes de données, les arbres de décision ont trouvé de nombreux champs de mise en œuvre . Fig. 2.7 montre un exemple de DT pour prédire si une personne est diabétique ou non [59] .



**Figure 2. 7: Arbre de décision répondre à la question si un personne diabétique ou non ?**

Cette modèle contient trois attributs qui sont <<minimum systolic blood pressure>>, <<age>> et <<glucose >> avec deux classes <<diabetic >> et <<non-diabetic >>.

\_Si minimum systolic blood pressure > 91 = no , alors : personne = diabetic ;

\_Si minimum systolic blood pressure > 91 = yes AND age > 40 = no , alors : personne = non-diabetic ;

\_Si minimum systolic blood pressure > 91 = yes AND age > 40 = yes AND glucose =no ,alors : personne = non-diabetic ;

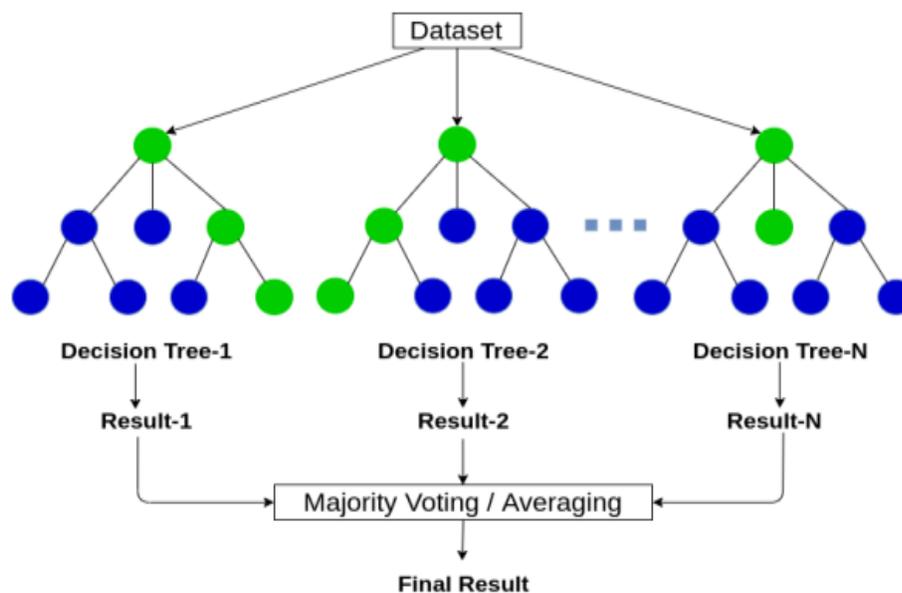
\_Si minimum systolic blood pressure > 91 = yes AND age > 40 = yes AND glucose =yes , alors : personne = diabetic ;

### 2.4.3 Random Forest (forêts aléatoires)

La forêt aléatoire est l'une des méthodes dites d'ensemble pour la classification, car un comité d'apprenants (des arbres dans ce cas) est généré et chacun vote pour l'étiquette prédite d'une instance donnée. Les arbres sont construits en utilisant la méthodologie des arbres de classification et de régression (CART). Dans la construction de l'ensemble d'arbres, RF utilise deux types de hasard : premièrement, chaque arbre est cultivé en utilisant une version bootstrap des données de formation. Un deuxième niveau de randomité est ajouté lors de la croissance de l'arbre en sélectionnant un

échantillon aléatoire de prédicteurs à chaque noeud pour choisir la meilleure division. Le nombre de prédicteurs sélectionnés à chaque noeud et le nombre d'arbres dans l'ensemble sont les deux principaux paramètres de l'algorithme RF. Les développeurs RF ont signalé que la méthode ne nécessite pas beaucoup de réglage des paramètres et les valeurs par défaut produisent souvent de bons résultats pour de nombreux problèmes. Une fois la forêt construite, l'attribution d'une nouvelle instance à une classe est accomplie en combinant les arbres, en utilisant un vote majoritaire.[61]

Random Forest est un classificateur composé d'un ensemble de classificateurs arborescents avec des vecteurs aléatoires indépendants distribués de façon identique et chaque arbre lançant un vote unitaire à l'entrée  $x$  pour la classe la plus populaire. Un vecteur aléatoire indépendant des vecteurs aléatoires précédents de la même distribution est généré et un arbre est généré à l'aide du test de formation. une borne supérieure est extraite pour Random Forests pour obtenir l'erreur de généralisation en termes de deux paramètres Exactitude et interdépendance des classificateurs individuels. La figure 2.8 montre l'organigramme de Random Forest.[62]



*Figure 2. 8: Structure de l'algorithme random forest*

#### 2.4.4 Support Vector Machine (SVM)

L'apprentissage SVM est l'une des nombreuses méthodes machine-learning. Par rapport aux autres méthodes ML, SVM est très efficace pour reconnaître les modèles subtils dans les ensembles

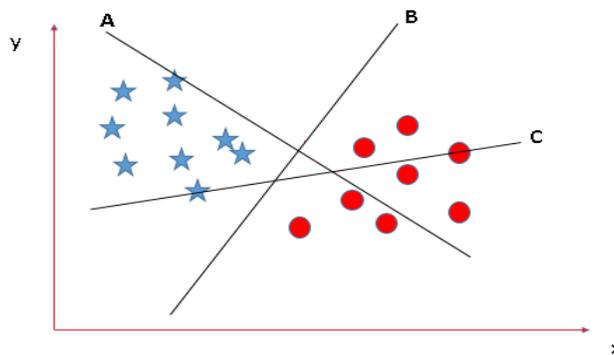
de données complexes . SVM peut être utilisé pour reconnaître l'écriture, reconnaître les cartes de crédit frauduleuses, identifier un haut-parleur, ainsi que détecter le visage . Le cancer est une maladie génétique où les caractéristiques génomiques ou les caractéristiques fonctionnelles peuvent représenter les sous-types de cancer, le pronostic des résultats, la prédiction des avantages médicamenteux, la tumorigenèse ou un processus biologique spécifique à la tumeur. Par conséquent, l'intelligence artificielle de SVM peut nous aider à reconnaître ces modèles dans une variété d'applications.[63]

Un Support Vector Machine (SVM) est un classificateur discriminant formellement défini par un hyperplan séparateur. En d'autres termes, compte tenu des données de formation étiquetées (apprentissage supervisé), l'algorithme produit un hyperplan optimal qui catégorise les nouveaux exemples.

Le problème le plus important qui se pose lors de l'utilisation de SVM est de savoir comment choisir le bon hyperplan. Considérez le scénario suivant [64]:

### Scénario 1 :

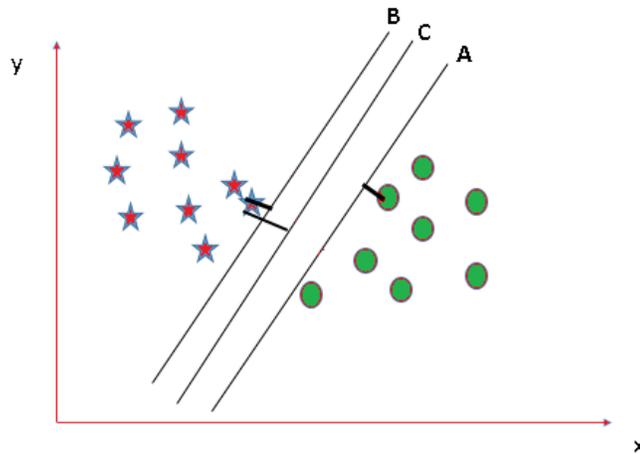
Dans ce cas, il y a trois hyperplans, appelés A, B, C. Le problème est maintenant de déterminer le bon hyperplan qui distingue le mieux l'étoile du cercle.



*Figure 2. 9: Un simple exemple sur le fonctionnement de l'algorithme SVM*

La règle générale à connaître est que jusqu'à ce que vous trouviez le bon hyperplan, lors de la classification des étoiles et des cercles, vous devez choisir l'hyperplan qui sépare le mieux les deux classes. Dans ce cas, B classe mieux les étoiles et les cercles, c'est donc un hyperplan droit [65].

**Scénario 2 :** Considérons maintenant un autre cas où les trois plans séparent bien les classes. La question est maintenant de savoir comment déterminer le bon plan dans cette situation



*Figure 2. 10: Séparation parfaite de deux classes avec un hyperplan*

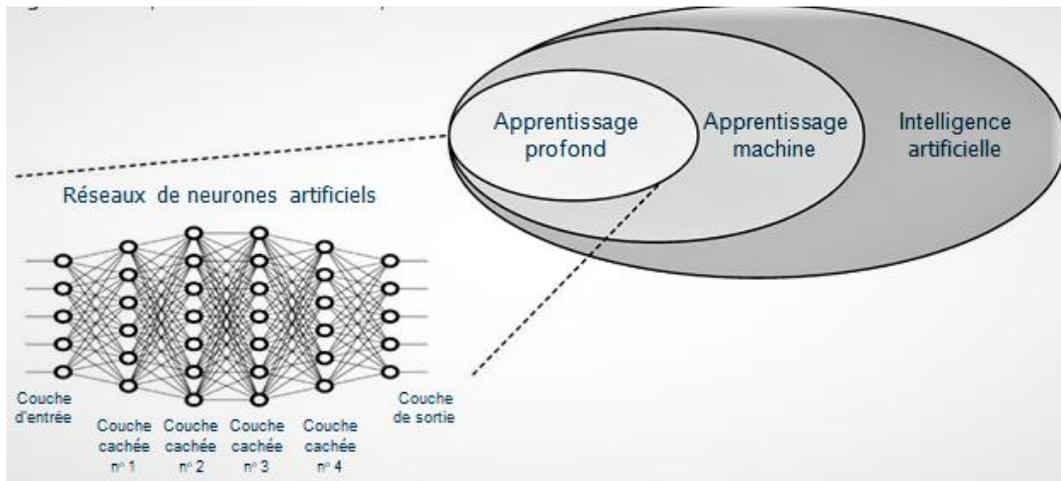
Dans ce cas, calculez la marge, qui est la distance entre le point de données le plus proche et l'hyperplan. Le plan à la distance maximale sera considéré comme l'hyperplan correct pour un meilleur ordonnancement des classes. Ici, C a la plus grande marge, il sera donc considéré comme l'hyperplan droit.[66]

### 3.1 L'APPRENTISSAGE APPROFONDI

L'apprentissage en profondeur est une forme d'intelligence artificielle dérivée de l'apprentissage automatique, il est donc nécessaire de comprendre son principe avant d'essayer de comprendre ce qu'est l'apprentissage en profondeur. Le concept d'apprentissage automatique remonte au milieu du 20ème siècle. Dans les années 1950, le mathématicien britannique Alan Turing a créé une machine capable d'apprendre, une "learning machine". Diverses techniques d'apprentissage automatique ont été développées en créant des algorithmes capables d'apprendre et de s'améliorer de manière autonome et continue .

Le deep learning est un sous-domaine de l'apprentissage automatique qui se concentre sur l'utilisation de réseaux de neurones profonds pour effectuer des tâches complexes telles que la reconnaissance d'images, la reconnaissance de la parole et la traduction automatique. Les réseaux de neurones profonds la figure 2.11 sont des algorithmes de traitement de données complexes qui peuvent

être formés à partir de grandes quantités de données pour effectuer des tâches en utilisant une approche hiérarchique. Cette approche hiérarchique permet aux réseaux de neurones profonds de capturer des caractéristiques complexes des données et de les utiliser pour effectuer des tâches de manière plus efficace que les algorithmes d'apprentissage automatique plus simples.



*Figure 2. 11: Exemple de réseau de neurones profond*

## 3.2 LES TYPES DE MODÈLES D'APPRENTISSAGE EN PROFONDEUR

### 3.2.1 Les réseaux neuronaux convolutifs (CNN)

CNN est un type de réseau de neurones spécialisés en traitement de données ayant une topologie pareille à une grille. Qui se sont classer très efficaces dans des divers domaines comme la reconnaissance et la classification d'images et vidéos. Il est utilisé pour identifier les isages, les objets, panneaux de circulation et auto-conduite des voitures.

Les réseaux neuronaux convolutifs (CNN) utilise pour l'analyse d'images. Ils fonctionnent en appliquant une série de filtres (appelés couches de convolution) à l'image d'entrée pour extraire des caractéristiques importantes, telles que des bords ou des textures. Ces caractéristiques sont ensuite agrégées dans des couches de pooling pour réduire la taille de l'image. Enfin, les caractéristiques sont envoyées à une ou plusieurs couches de neurones entièrement connectés pour produire des sorties, telles que la classification de l'image. Les poids de chaque couche sont appris lors de l'entraînement du réseau en utilisant un ensemble de données étiquetées.

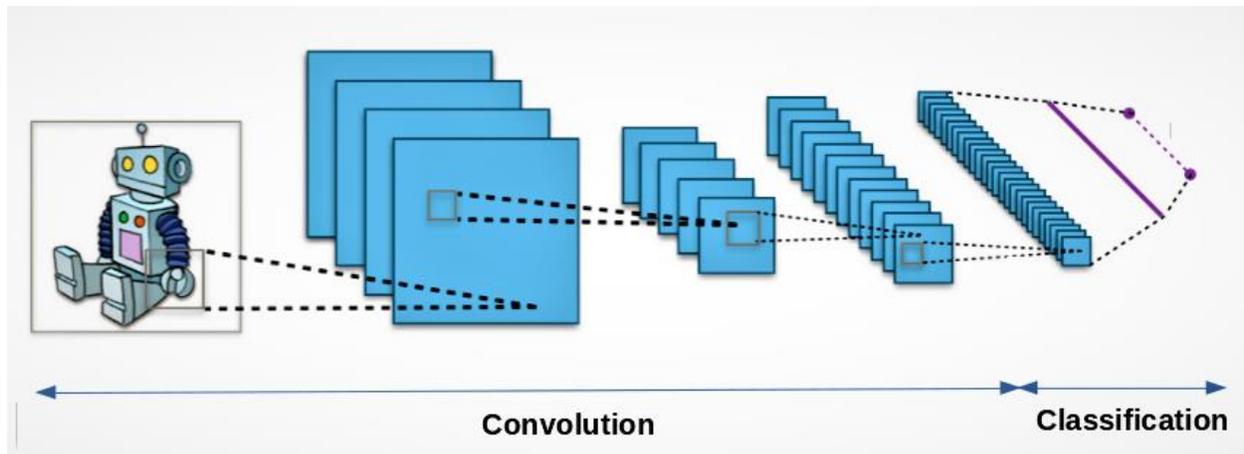


Figure 2. 12: Architecture d'un réseau de neurones convolutif

Il existe quatre types de couches pour un réseau de neurones convolutif : la couche de convolution, la couche de pooling, la couche de correction ReLU et la couche fully-connected.

#### • la couche de Convolution

La convolution est l'une des opérations les plus importantes dans le domaine de traitement de signal et elle est utilisée dans plusieurs applications en traitement naturel des langues, vision par ordinateur et traitement d'images. L'opération de convolution peut également être appliquée à des fonctions multidimensionnelles. Comme indique la figure 2.13 la convolution peut être appliquée aux images pour effectuer diverses transformations ; ici, les images sont traitées comme des fonctions bidimensionnelles.

On note  $I$  l'image en entrée,  $K$  le filtre 2D dont les dimensions  $m \times n$  et  $F$  la carte de caractéristiques qui représente le résultat de la convolution de l'image avec le filtre. On peut exprimer mathématiquement cette opération comme suit[71]:

$$F(i, j) = (I * K)(i, j) = \sum_m \sum_n I(i - m, j - n) K(m, n)$$

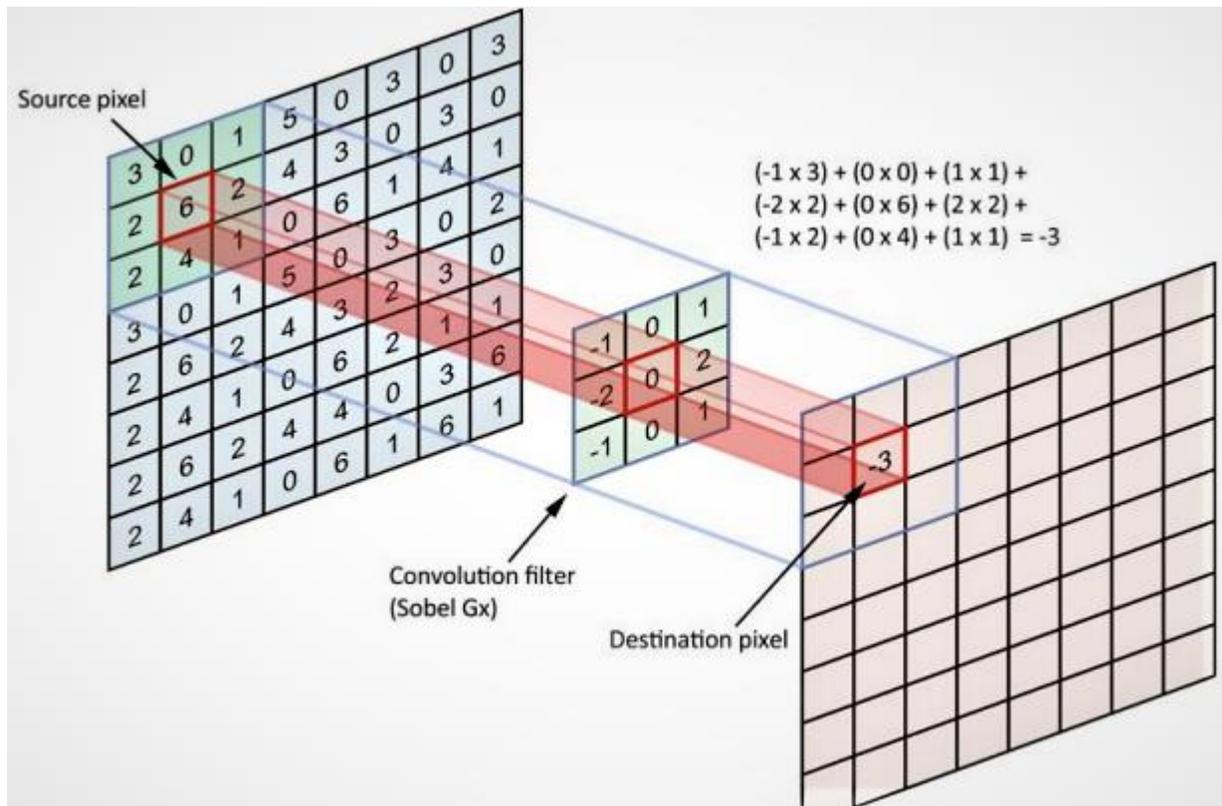
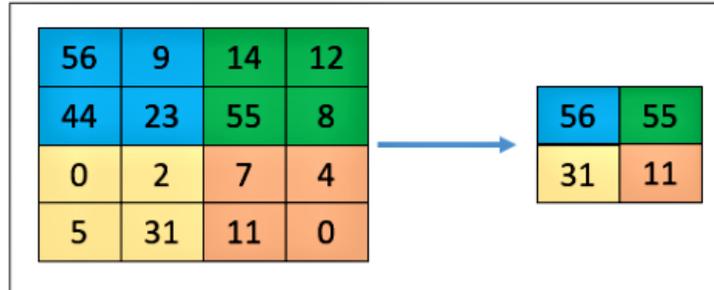


Figure 2. 13: Opération de convolution

### • La couche de Pooling

La mise en commun des couches vise à réduire progressivement la dimensionnalité de la représentation, et ainsi réduire davantage le nombre de paramètres et la complexité de calcul du modèle.

La couche de Pooling opère sur chaque carte d'activation de l'entrée, et ajuste sa dimensionnalité en utilisant la fonction « MAX ». Dans la plupart des CNN, ceux-ci se présentent sous forme de **max-pooling layers** avec des noyaux d'une dimensionnalité de 2x2 appliqués avec une enjambée de 2 le long des dimensions spatiales de l'entrée. Cela réduit la carte d'activation à 25% de la taille originale - tout en maintenant le volume de profondeur à sa taille standard.[72]



*Figure 2. 14: MaxPooling*

- **la couche de correction ReLU**

On fait passer les cartes de convolution à travers une couche d'activation non linéaire telle que Rectified Linear Unit (ReLU), qui consiste à remplacer les nombres négatifs des images filtrées par des zéros.

L'efficacité du traitement est améliorée en insérant une couche entre les couches de traitement qui exécutera une fonction mathématique (fonction d'activation) sur le signal de sortie.

Dans ce cadre on trouve ReLU (Rectified Linear Units) spécifie une fonction réelle non linéaire définie par  $\text{ReLU}(x) = \max(0, x)$ .

Par conséquent, la couche de correction ReLU remplace toutes les valeurs négatives reçues en entrée par des zéros. Il agit comme une fonction d'activation, Habituellement, la correction Relu est préférable, mais il existe d'autres formes.[73]

- La correction par tangente hyperbolique  $f(x) = \tanh(x)$ ,
- La correction par la tangente hyperbolique saturante:  $f(x) = |\tanh(x)|$ ,
- La correction par la fonction sigmoïde  $f(x) = \frac{1}{1 + e^{-x}}$ .

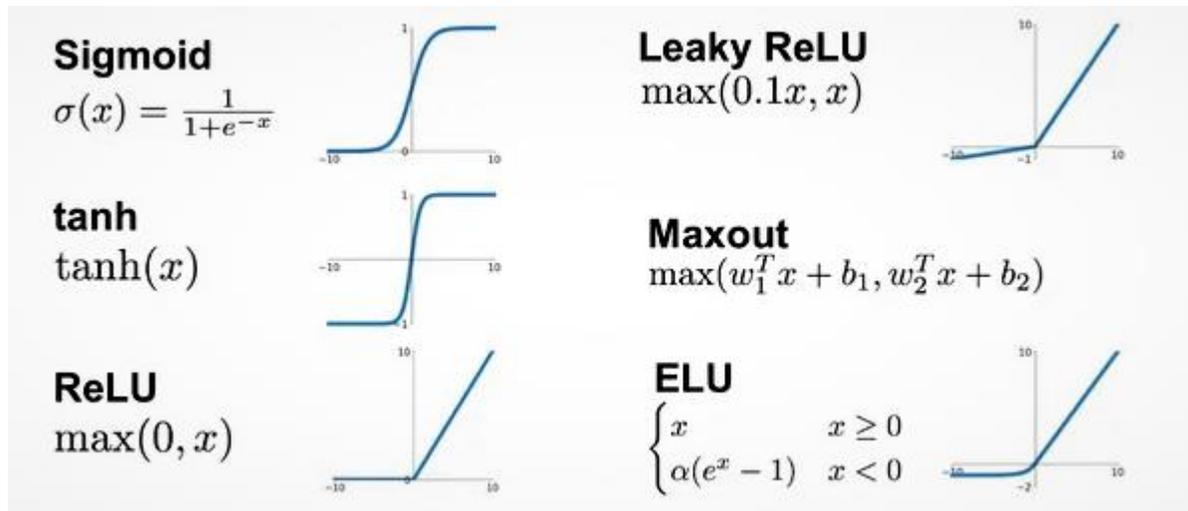


Figure 2.15: Les Trois fonctions d'activation

➤ **Rectified Linear Unit (ReLU)** Utilisée en conjonction avec une transformation linéaire transforme l'entrée en  $f(x) = \max(0; wx + b)$

La fonction d'activation sous-jacente est  $f(x) = \max(0; x)$  Récemment, le ReLU est plus couramment utilisé comme unité cachée. Les résultats montrent que les ReLU conduisent à des gradients importants et cohérents, ce qui facilite l'apprentissage par gradient Figure 3.5. Bien qu'un ReLU ressemble à une unité linéaire, il a une fonction dérivée et permet ainsi de calculer le gradient des pertes. Ces derniers temps, le ReLU a été le choix le plus populaire pour l'activation réseau cachée. Dans la plupart des cas, un ReLU peut être un choix par défaut qui conduirait à des résultats souhaitables dans un délai opportun.[88]

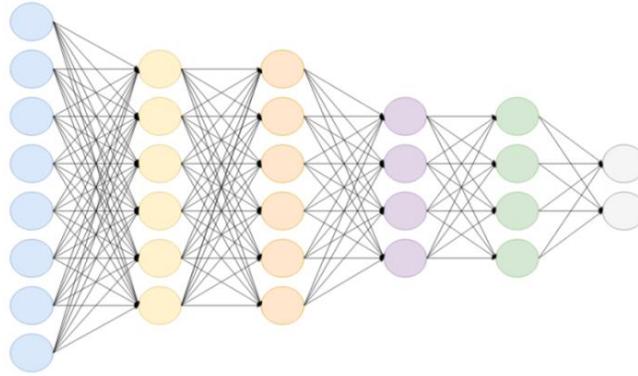
➤ **La Fonction Sigmoid** L'activation sigmoïde transforme l'entrée  $x$  comme suit:  $f(x) = \frac{1}{1+e^{-x}}$

Les unités sigmoïdes peuvent être utilisées dans la couche de sortie en conjonction avec l'entropie croisée binaire pour les problèmes de classification binaire. La sortie de cette unité peut modéliser une distribution Bernoulli sur la sortie  $y$  conditionnée sur  $x$ . [89]

- **Fully-connected Layer**

Les couches entièrement connectées de CNN sont entièrement connectées les unes aux autres comme leur nom l'indique. Ceci se trouve dans la dernière étape de CNN qui est complètement connecté avec les couches précédentes. Il dépeint le vecteur de caractéristique pour l'entrée. A partir de la figure 2.16 nous pouvons voir la structure d'une couche FC.

Les neurones d'une couche entièrement connectée ont toutes les connexions d'activation de la couche précédente. Multiplication matricielle avec un id offset de base utilisé pour activer les connexions. Lorsque le réseau est formé, les vecteurs de fonction régulent la perte. Ce vecteur est également utilisé pour la classification, la régression et l'alimentation d'autres couches.



*Figure 2. 16: Fully connected layer*

### 3.2.2 Les réseaux neuronaux récurrents (RNN)

Les réseaux de neurones récurrents (RNN) sont une forme de réseau de neurones qui peuvent prendre en compte des séquences d'entrées. Ils ont des boucles récurrentes qui leur permettent de conserver une mémoire à court terme, ce qui est particulièrement utile pour des tâches telles que la prédiction de séquences ou la traduction de langues. Les RNN traitent chaque élément de la séquence en utilisant une fonction d'activation non linéaire pour produire une sortie, qui est ensuite transmise à l'élément suivant de la séquence en même temps que l'état caché de la mémoire à court terme.

Les réseaux de neurones récurrents, recurrent neural network (RNN), ont été introduits pour que le modèle puisse traiter plus efficacement les séquences de données où l'ordre des entrées a de l'importance. Les RNN traitent la séquence d'entrée un élément à la fois et utilisent le retour d'information des éléments précédents dans le calcul des éléments suivants de la séquence. Ils conviennent donc parfaitement au traitement des données de nature séquentielle, comme l'ECG. Le modèle de mémoire à long terme (LSTM) est une version améliorée du RNN conventionnel [74].

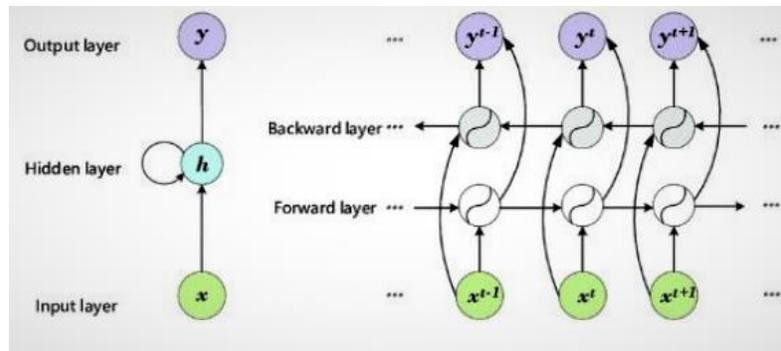


Figure 2. 17: l'architecture de réseau récurrent

### 3.2.3 Les réseaux de neurones artificiels (ANN)

ANN joue un rôle de premier plan dans deux principaux types de réseaux utilisés : les réseaux de transmission (supervisés) et les cartes auto-organisées (non supervisées). Ces réseaux sont constitués par une série de neurones connectés. Un seul neurone prend des variétés d'entrées numériques et fournit un ensemble pondéré de sorties transformées. La fonction de transformation fréquemment utilisée implique des fonctions sigmoïdes et tanh.[75]

Un ANN avec exactement trois couches de neurones est connu sous le nom de réseau de neurones à trois couches. La couche d'entrée est responsable de la réception des données initiales ou des caractéristiques d'entrée, la couche cachée effectue le traitement et la transformation des données, tandis que la couche de sortie produit les résultats ou les prédictions finales.

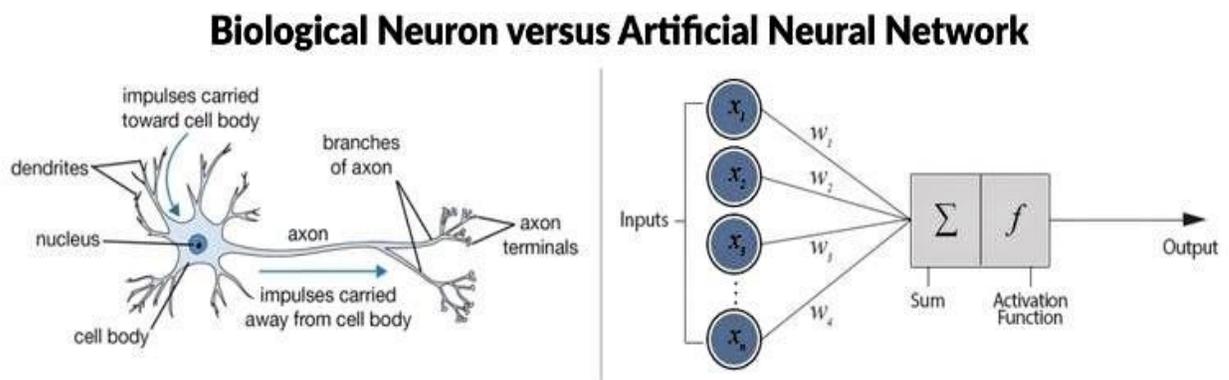
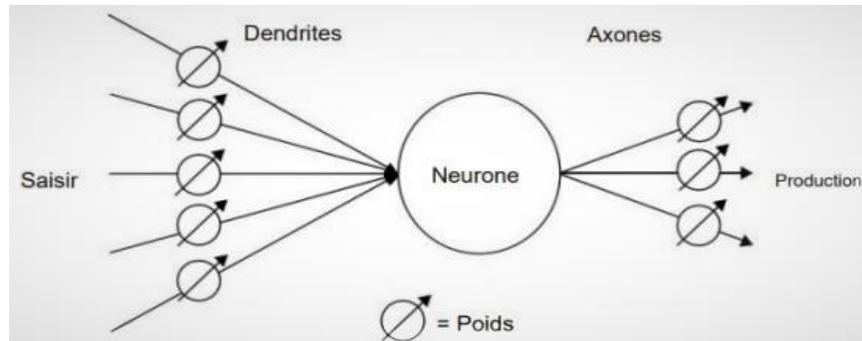


Figure 2. 18: Réseau de neurones biologique vs réseau de neurones artificielles

Les éléments de base de l'ANN sont les nœuds, également appelés éléments de traitement (PE), et les connexions. Chaque nœud a sa propre entrée, à partir de laquelle il reçoit des

communications d'autres nœuds et/ou de l'environnement et sa propre sortie, à partir de laquelle il communique avec d'autres nœuds ou avec l'environnement. Enfin, chaque nœud a une fonction  $f$  par laquelle il transforme sa propre entrée globale en sortie Figure 2.19

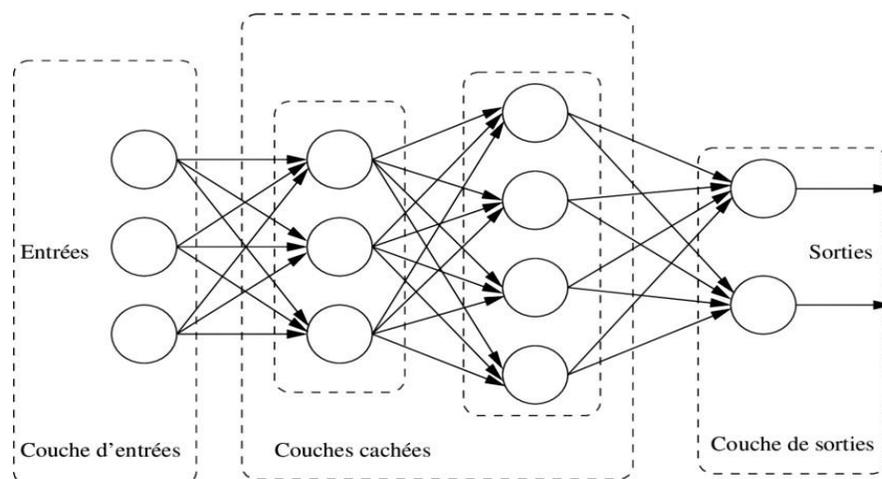


*Figure 2. 19: Schéma d'un élément de traitement unique (PE)*

### 3.2.4 Deep Neural Network (DNN) :

Deep Neural Network (DNN) est un type de réseau de neurones artificiels qui est conçu pour apprendre des modèles complexes à partir de données en utilisant un grand nombre de couches de neurones. Le DNN est un réseau de neurones artificiels qui peut être utilisé pour des tâches de classification, de régression, de reconnaissance de motifs, de traitement du langage naturel et d'autres tâches liées à l'apprentissage automatique.

Réseau de neurones profond DNN est une extension d'un perceptron multicouche MLP avec plusieurs couches de neurones artificiels.



*Figure 2. 20: Représentation graphique d'un perceptron multicouche*

Le DNN se compose de plusieurs couches de traitement qui sont capables d'extraire les caractéristiques hiérarchiques des données d'entrée. Le fonctionnement du DNN est émulé du fonctionnement du cerveau humain. Chaque couche du DNN est constituée de multiples unités de traitement appelées neurones. Un neurone effectue la somme pondérée des entrées  $[x_1, x_2, x_n]$  et alimente la somme résultante à une fonction d'activation qui génère la sortie désirée. Chaque neurone se compose d'un ensemble de poids  $[w_1, w_2, w_n]$  et d'un biais qui est optimisé pendant le processus d'entraînement. Le fonctionnement du neurone artificiel est représenté à la Fig. 2.21. [89]

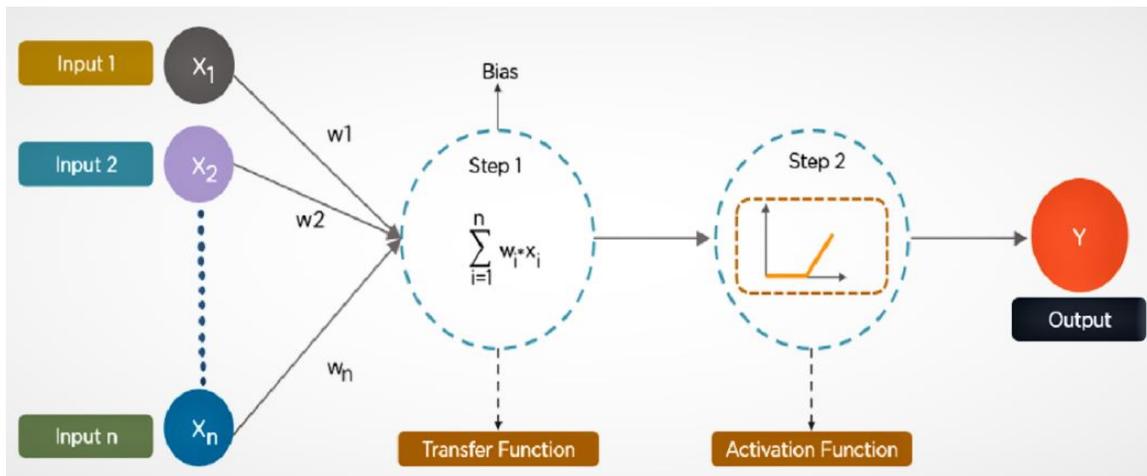


Figure 2. 21: Artificial neuron

### 3.2.5 Les réseaux antagonistes génératifs (GAN)

GAN est un modèle semi-supervisé ou non supervisé, et ils sont des modèles génératifs qui génèrent de nouvelles données échantillons comme l'entrée. Adversarial GAN terme implique qu'il existe une concurrence entre le générateur et le discrimination. Il est plus largement mis en œuvre dans les applications CV et NLP. Produit de contraste et imagerie moléculaire [76].

Les réseaux de neurones génératifs antagonistes (GAN) sont une forme de réseau de neurones qui permettent de générer de nouvelles données en s'appuyant sur un ensemble de données existant. Les GAN se composent de deux réseaux de neurones distincts : un générateur et un discriminateur. Le générateur prend en entrée un vecteur de bruit aléatoire et utilise une série de couches cachées pour produire une sortie qui ressemble à l'ensemble de données existant. Le discriminateur prend en entrée soit une instance de données générée par le générateur, soit une instance de données de l'ensemble de données existant, et essaie de déterminer si l'instance est vraie ou fausse.

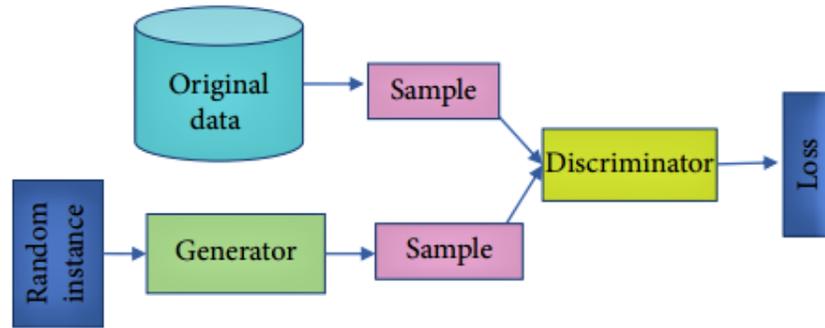


Figure 2. 22: réseaux antagonistes génératifs (GAN)

### 3.2.6 Les réseaux neuronaux de longue mémoire à court terme (LSTM)

Les réseaux neuronaux de longue mémoire à court terme (LSTM) sont une variante des réseaux neuronaux récurrents (RNN). Ces réseaux sont conçus pour être capables de traiter des séquences de données à long terme, comme des mots ou des phrases. Ils sont constitués de cellules de longue mémoire à court terme (LSTM) qui permettent aux réseaux de retenir plus longtemps des informations pertinentes. Les réseaux LSTM sont couramment utilisés pour la génération de texte, l'analyse de sentiment et la classification de documents [84].

Un réseau LSTM est constitué d'une séquence de cellules LSTM. Chaque cellule LSTM est dotée de trois portes qui contrôlent le flux d'informations à travers le réseau. La porte d'entrée régule l'arrivée des informations, la porte de mémorisation et d'oubli s'assure que les informations non pertinentes sont oubliées, et la porte de sortie détermine quelles informations sont transmises à l'étape suivante. Cette architecture permet au réseau de prendre des décisions en se basant sur des expériences passées et d'identifier les dépendances à long terme dans les données [102].

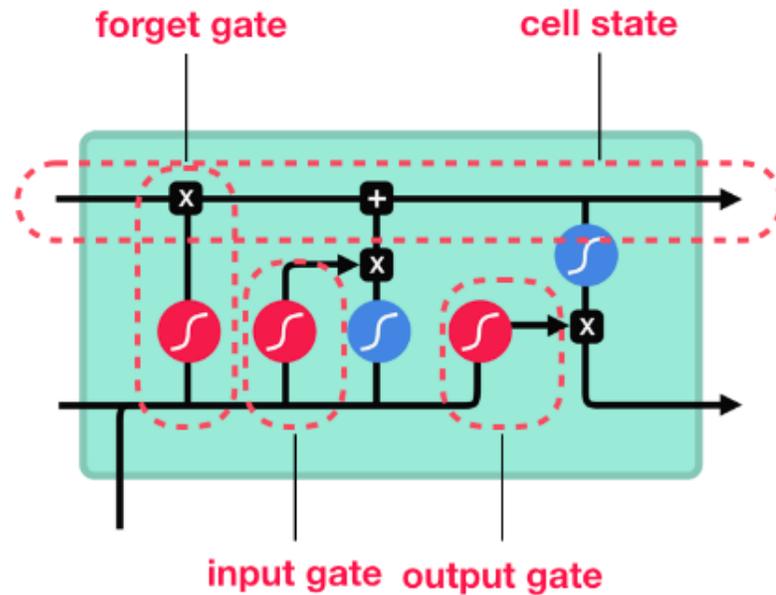


Figure 2. 23: L'architecture de LSTM

L'architecture LSTM se compose de plusieurs éléments importants:

- **Porte d'entrée** La porte d'entrée contrôle le flux des nouvelles données qui entrent dans la cellule de mémoire. Elle utilise une fonction d'activation sigmoïde pour déterminer quelles valeurs doivent être conservées et lesquelles doivent être rejetées.
- **Porte d'oubli ( Forget Gate )** La porte d'oubli régule le flux d'informations provenant de l'étape précédente qui doivent être oubliées. Elle utilise également une fonction d'activation sigmoïde pour décider quelles informations doivent être oubliées.
- **Porte de sortie** La porte de sortie contrôle la sortie de la cellule de mémoire. Elle utilise à la fois une fonction d'activation sigmoïde et une fonction de tangente hyperbolique pour déterminer quelles informations doivent être émises.
- **Cellule de mémoire** La cellule de mémoire est le composant principal de l'architecture LSTM. Elle stocke les informations au fil du temps et peut choisir sélectivement d'oublier certaines informations ou d'ajouter de nouvelles informations à son état interne.

À chaque pas de temps, le modèle LSTM reçoit un vecteur d'entrée ainsi qu'un vecteur d'état caché provenant de l'étape précédente. Le vecteur d'entrée est traité par la porte d'entrée et la porte d'oubli, et les valeurs résultantes sont utilisées pour mettre à jour la cellule de mémoire [102].

### 3.2.7 Modèle VGG16

Le Modèle VGG-16 a reçu une formation sur la base de données ImageNet .En raison de la formation approfondie que le réseau VGG-16 a subi, il donne d'excellentes précisions même lorsque les ensembles de données d'image sont petits. Le réseau VGG-16 se compose de 16 couches de convolution et a un petit champ réceptif de  $3 \times 3$ . Il a une couche Max pooling de taille  $2 \times 2$  et a un total de 5 couches de ce type. Il y a 3 couches entièrement connectées après la dernière couche Max pooling.

Ceci est suivi par trois couches entièrement connectées. Il utilise le classificateur softmax comme couche finale. L'activation ReLu est appliquée à toutes les couches cachées. Un schéma de l'architecture VGG-16 est présenté à la Fig 2.24 [77] .

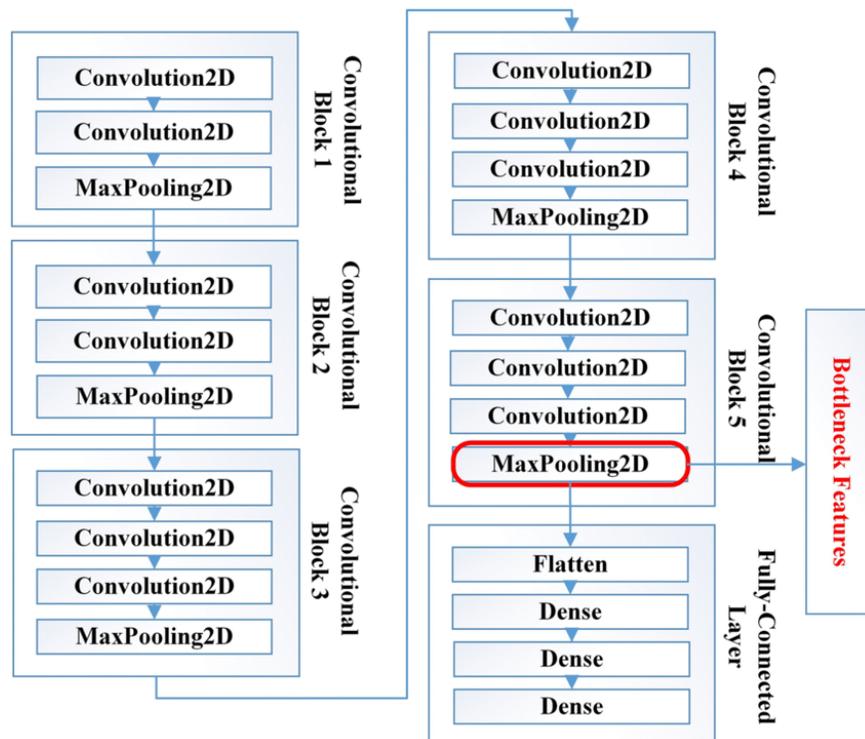


Figure 2. 24: Schéma fonctionnel du réseau VGG-16

### 3.2.8 LES TRAVAUX DE RECHERCHE SUR L'APPLICATION DES ALGORITHMES DE MACHINE LEARNING ET DU DEEP LEARNING POUR LA PRÉDICTION DU DIABÈTE

L'objectif principal de notre travail est de proposer un outil développé pour une prédiction plus précise et une détection précoce du diabète. Il existe une grande quantité de données disponibles sur Internet ou des sources externes, nous présenterons leurs méthodes et les résultats obtenus.

Les chercheurs ont commencé à réaliser les capacités des techniques deep Learning dans le traitement de grands ensembles de données. Par conséquent, la prédiction du diabète a également été effectuée en utilisant des techniques DL. Sept études ont été publiées au cours des six dernières années, comme le montre dans la figure 2.25 [78].

Years	technique	Result	Dataset
2016	Unsupervised Deep Learning Neural Network (Deep Patient) Convolutional Neural Network (CNN)	Area Under the ROC Curve (AUC-ROC) 0.91 Accuracy: 77.5%	Electronic Health Records
2017	Multilayer Perceptron Neural Network (MLP) Logistic Regression	Accuracy: 72.5% Accuracy: 65.2%	Continuous Glucose Monitoring (CGM) signals
2017	Recurrent Deep Neural Network (RNN) Deep Neural Network Long Short-Term Memory (LSTM)	Accuracy: Type 1 Diabetes = 78% Type 2 Diabetes = 81% Precision: 59.6%	Pima Indians Diabetes Dataset
2017	Markov Chain Neurs Network Plain Recurrent Deep Neural Network (RNN)	Precision: 34.1% Precision: 58.0	large regional Australian hospital dataset
2018	Modified Convolution Neural Network (CNN) Convolutional Neural Network (CNN)	Receiver Operating Characteristic (ROC) 0.96 Accuracy: 90.9%	Breath Dataset Private
2018	Convolutional Neural Network (CNN) combined with Long Short-Term Memory (LSTM)	Accuracy: 95.1%	Electrocardiograms (ECG) Private
2018	Deep Learning Architecture (MLP/General regression neural network (GRNN)/Radial Basis Function (RBF)	Accuracy 88.41%	Pima indians Diabetes Databse

**Tableau 2. 1: Résumé des travaux techniques de Deep Learning**

De plus, Ashiquzzaman et al. [79] ont utilisé un réseau neural profond (DNN). L'architecture du DNN composé de Perceptron multicouches (MLP), Réseau neuronal de régression générale (GRNN), et Radial Basis Function (RBF). L'évaluation de l'approche était fondée sur l'ensemble de données de Pima Indian. Les auteurs n'ont pas pré-traité l'ensemble de données intentionnellement

car DNN peut filtrer les données et acquérir les biais. L'ensemble de données est divisé en 192 échantillons pour l'ensemble d'essais et le reste pour la formation. Le taux de précision rapporté par les auteurs était de 88,41%.

Une autre étude de Swapna et al. [80] a utilisé deux techniques d'AD pour améliorer la précision de la prédiction du diabète. Un ensemble de données privées appelé électrocardiogrammes a été utilisé pour évaluer le rendement de CNN et de CNN-LSTM. Il comprenait 142 000 échantillons et huit attributs. La validation croisée en cinq volets a été utilisée pour diviser l'ensemble de données en ensembles de formation et d'essai. Les auteurs n'ont pas prétraité les données ni appliqué la méthode de sélection des caractéristiques en raison de l'auto-apprentissage du DNN. Les taux de précision générés pour les modèles étaient de 90,9 % et 95,1 %, respectivement.

Mohebbi et al. [81] ont utilisé la régression logistique comme base pour le réseau neuronal à plusieurs couches et le réseau neuronal conventionnel (CNN). L'objectif était de détecter les patients diabétiques à partir d'un ensemble de données de surveillance continue du glucose (CGM). L'ensemble de données est composé de neuf patients et chaque patient a reçu 10800 jours de données CGM, ce qui donne un total de 97200 jours CGM simulés. Les attributs utilisés dans cette étude n'ont pas été abordés. L'ensemble de données a été divisé en ensembles de formation, de validation et d'essai fondés sur la technique de validation croisée sans patient. En fait, les auteurs ont sélectionné six patients pour la formation et la validation, et trois patients pour les tests. Le CNN a atteint la plus haute précision de 77,5%.

Pham et al. [82] ont appliqué trois techniques d'AD différentes à un ensemble de données recueillies manuellement dans un hôpital régional australien. L'ensemble de données est composé de 12 000 échantillons (patients) contenant 55,5 % d'hommes. Certaines techniques de prétraitement (non mentionnées dans leur article) ont été appliquées pour nettoyer et réduire les échantillons à 7191 patients. Pour la validation, l'ensemble de données a été divisé en 2/3 pour l'ensemble de formation, 1/6 pour la validation et 1/6 pour l'essai. Les méthodes étaient la mémoire à court terme longue (LTSM), Markov et Plain RNN. La valeur de précision a été utilisée pour comparer les performances des techniques. La meilleure valeur de précision de 59,6 % a été obtenue en utilisant le LTSM.

De plus, Ramesh et al. [83] ont utilisé le Réseau neural récurrent (RNN) pour prédire les deux types de diabète. Les auteurs ont utilisé l'ensemble de données de Pima Indian avec 768 échantillons

et huit attributs. Les attributs sont classés selon leur plus haute importance, comme indiqué dans leur étude « Glucose, IMC, âge, grossesses, fonction généalogique du diabète, pression artérielle, épaisseur de la peau et insuline ». Pour valider l'étude, l'ensemble de données a été divisé en 80 % pour la formation et 20 % pour l'essai. L'exactitude de la prédiction du diabète de type 1 était de 78 %, alors qu'elle était de 81 % pour le diabète de type 2.

Les algorithmes de ML sont très bien connus dans le domaine médical pour prédire les maladies. De nombreux chercheurs ont utilisé des techniques de ML pour prédire le diabète afin d'obtenir les meilleurs résultats et les plus précis.[84]

Yuvaraj et Sripreethaa [85] ont présenté une application de prédiction du diabète utilisant trois algorithmes ML différents, dont Random Forest, Decision Tree et Naïve Bayes. L'ensemble de données sur le diabète chez les Indiens pimas (IDP) a été utilisé après le prétraitement. Les auteurs n'ont pas mentionné comment les données ont été pré-traitées, mais ils ont discuté de la méthode de gain d'information utilisée pour la sélection des fonctionnalités pour extraire les fonctionnalités pertinentes. Ils n'ont utilisé que huit attributs principaux parmi 13. De plus, ils ont divisé l'ensemble de données en 70 % pour la formation et 30 % pour les tests. Les résultats ont montré que l'algorithme de forêt aléatoire avait le taux de précision le plus élevé de 94 %.

De plus, Tafa et al. [86] ont proposé un nouveau modèle intégré amélioré de SVM et de Naïve Bayes pour prédire le diabète. Le modèle a été évalué à l'aide d'un ensemble de données recueillies à trois endroits différents au Kosovo. L'ensemble de données contient huit attributs et 402 patients chez lesquels 80 patients avaient le diabète de type 2. Les auteurs n'ont pas mentionné si les données étaient prétraitées ou non. Pour le test de validation, ils ont divisé l'ensemble de données en 50 % pour chacun des ensembles de formation et d'essai. Les algorithmes combinés proposés ont amélioré la précision de la prédiction pour atteindre 97,6 %. Cette valeur a été comparée à la performance de SVM et Naïve Bayes atteignant 95,52% et 94,52%, respectivement.

En outre, Deepti et Dilip [87] ont utilisé les classificateurs Decision Tree, SVM et Naive Bayes pour détecter le diabète. Le but était d'identifier le classificateur avec la plus grande précision. L'ensemble de données de Pima Indian a été utilisé pour cette étude. La partition de l'ensemble de données se fait au moyen d'une validation croisée à 10 plis. Les auteurs n'ont pas discuté du prétraitement des données. Le rendement a été évalué à l'aide des mesures de l'exactitude, de la

précision, du rappel et de la mesure F. La précision la plus élevée a été obtenue par le Bayes naïf, qui a atteint 76,30%.

Rakshit et al. [88] ont utilisé un réseau neuronal à deux classes pour prédire le diabète à l'aide de l'ensemble de données Pima Indian. Les auteurs ont prétraité l'ensemble de données en normalisant toutes les valeurs des attributs de l'échantillon en utilisant la moyenne et l'écart-type de chaque attribut afin d'obtenir une stabilité numérique. En outre, ils ont extrait les caractéristiques pertinentes en utilisant la corrélation. Cependant, les auteurs n'ont pas mentionné ces caractéristiques discriminatoires. L'ensemble de données a été divisé en un ensemble de formation contenant 314 échantillons et un ensemble d'essais comprenant 78 échantillons. Le résultat de ce modèle a atteint la plus haute précision de 83,3% par rapport aux autres précisions obtenues des études précédentes.

Dans leur quête pour améliorer le mécanisme de prédiction du diabète, les chercheurs ont construit des modèles combinés afin d'augmenter la précision. Les modèles peuvent être une combinaison de classificateurs d'apprentissage automatique ou d'un classificateur avec optimiseur d'intelligence artificielle. Ces modèles ont révélé une grande précision.

### **3.3 CONCLUSION**

Dans ce chapitre, nous avons exploré les différents types d'algorithmes de machine learning ainsi que des modèles d'apprentissage en profondeur et de leur application.

Nous avons constaté que l'utilisation de ces techniques peut nous aider à détecter et à prédire le diabète, ce qui peut aider à la prise en charge précoce et efficace de la maladie. Nous avons également examiné des recherches qui utilisaient des modèles d'apprentissage automatique et profond pour améliorer la précision de la prédiction du diabète de type 1 et de type 2.

# Chapitre III

## Contribution et résultats

### 3. INTRODUCTION

Dans ce dernier chapitre, nous allons tout d'abord présenter les différents outils, librairies, et les langages de programmation utilisés. Ensuite, nous procéderons à la définition de notre ensemble de données en décrivant ses caractéristiques ainsi que les différentes étapes de prétraitement des données telles que l'exploration, le nettoyage et la sélection du modèle. Ces étapes nous permettront de corriger les valeurs aberrantes et de choisir le modèle le plus approprié à suivre.

A la fin, nous allons montrer une évaluation de tout les modèles d'apprentissage utiliser (Decision Trees ,DNN,SVM et LSTM) et faire une comparaison entre les modèles afin de choisir le meilleur modèle qui a une grande précision pour prédire le diabète.

#### 3.1 ENVIRONNEMENT CLOUD

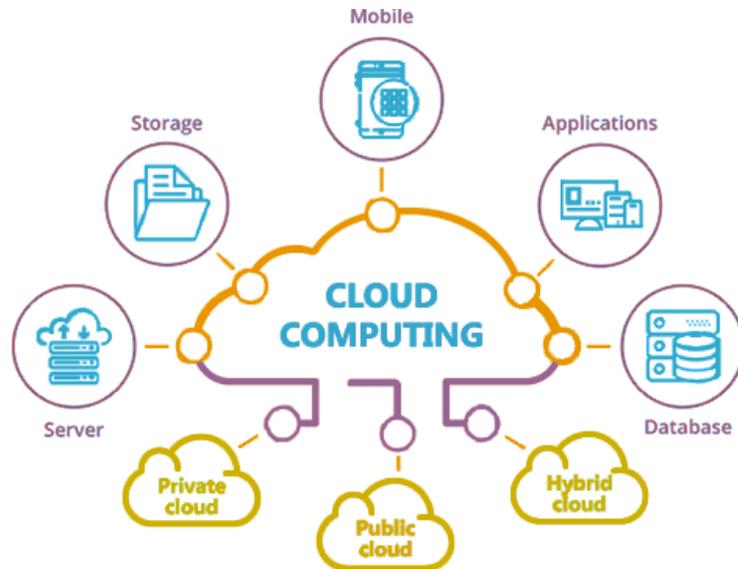
Environnement cloud est un modèle informatique qui permet d'accéder à des ressources informatiques à distance via Internet. Au lieu de disposer d'un serveur physique sur place, les utilisateurs peuvent utiliser des services cloud pour stocker des données, exécuter des applications et accéder à des ressources de calcul à la demande.

##### 3.1.1 Définition du cloud computing

Cette technologie permet la délocalisation des données et des applications sur des infrastructures dématérialisées accessibles depuis internet. De plus, elle consiste à exploiter la puissance de calculs et de stockages des données sur des serveurs informatiques distants par l'intermédiaire d'internet. Le cloud computing peut être divisé en trois catégories principales : le cloud public, le cloud privé et le cloud hybride.

De ce fait, les utilisateurs bénéficient de la baisse des coûts, grâce aux économies d'échelle de la ressource informatique et de l'augmentation de la fiabilité. Par conséquent, les entreprises ne seront plus obligées d'investir dans les serveurs, ni d'employer du personnel qualifié. Puisque, toute la partie informatique est sous-traitée à des prestataires de services.

D'ailleurs, le Cloud a créé de nouvelles opportunités pour les fournisseurs d'externalisation des services informatiques, dans la mesure où ces derniers ne seront appropriés qu'en fonction des besoins des clients.[90]

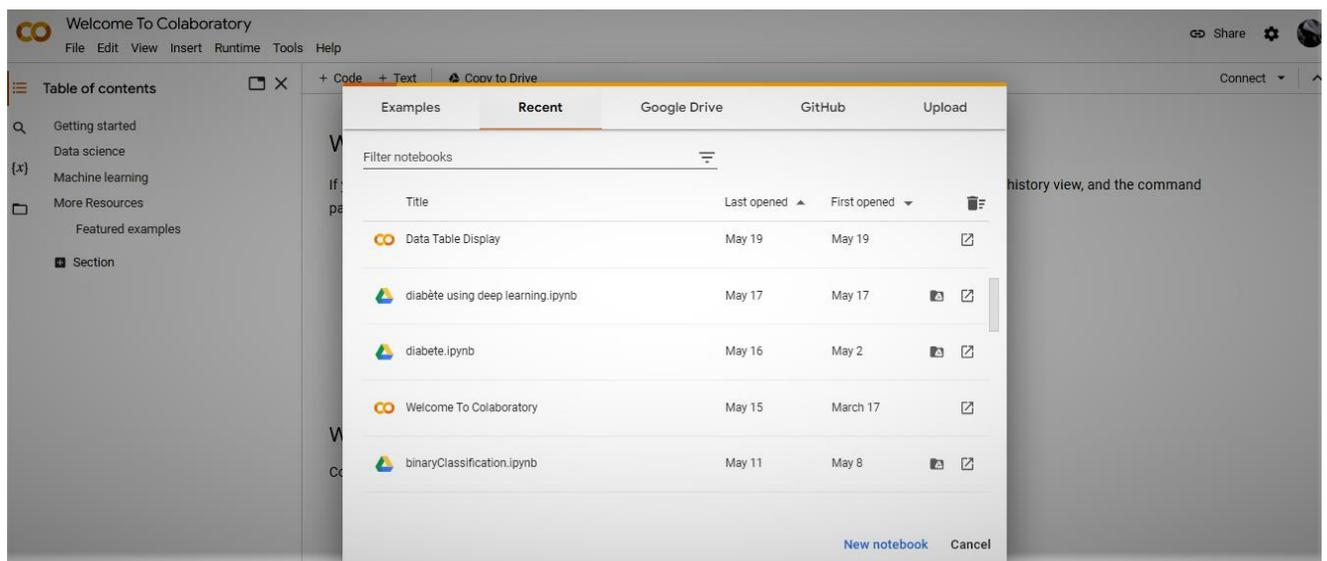


*Figure 3. 1: Architecture du cloud computing*

## 3.2 OUTILS ET LIBRAIRIES UTILISÉS

### 3.2.1 Google Colab

Colaboratoire, ou « Colab », est un produit de Google Research. Colab permet à quiconque d'écrire et d'exécuter du code python arbitraire via le navigateur, et est particulièrement bien adapté à l'apprentissage automatique, l'analyse des données et l'éducation. Plus Techniquement, Colab est un service de notebook Jupyter hébergé qui ne nécessite aucune configuration, tout en offrant un accès gratuit aux ressources informatiques, y compris les GPU [91] .



*Figure 3. 2:Google Colab*

### 3.2.2 Jupyter notebook

Jupyter Notebook est une application web open-source qui vous permet de créer et de partager des documents contenant du code en direct, des équations, des visualisations et du texte narratif. Utilisations : nettoyage et transformation des données, simulation numérique, modélisation statistique, visualisation des données, apprentissage automatique et bien plus encore [93] .

### 3.2.3 TensorFlow

TensorFlow est une plateforme open source de bout en bout pour l'apprentissage automatique. Il dispose d'un écosystème complet et flexible d'outils, de bibliothèques et de ressources communautaires qui permet aux chercheurs de pousser l'état de l'art en ML et aux développeurs de facilement construire et déployer des applications alimentées par ML [94].

### 3.2.4 Keras

Keras est une API d'apprentissage profond écrite en Python, fonctionnant sur la plateforme d'apprentissage automatique TensorFlow. Il a été développé en mettant l'accent sur l'expérimentation rapide. Il est essentiel de pouvoir passer d'une idée à un résultat le plus rapidement possible pour effectuer de bonnes recherches [95] .

### 3.2.5 Python

Python est un langage de programmation puissant et facile à apprendre. Il dispose de structures de données de haut niveau efficaces et d'une approche simple mais efficace de la programmation orientée objet.

La syntaxe élégante et la frappe dynamique de Python, ainsi que sa nature interprétée, en font un langage idéal pour le script et le développement rapide d'applications dans de nombreux domaines sur la plupart des plates-formes.

## 3.3 BIBLIOTHEQUES DE PYTHON

- **Pandas**

Pandas est un outil open-source très populaire dans le domaine de l'analyse de données et de la manipulation de données en Python. Pandas fournit des structures de données et des fonctions qui facilitent le traitement, l'exploration et la manipulation de jeux de données.

- **Numpy**

La bibliothèque NumPy (Numerical Python) est une bibliothèque fondamentale en Python pour la manipulation de tableaux multidimensionnels et la réalisation de calculs mathématiques et numériques efficaces.

- **Matplotlib**

Matplotlib est une bibliothèque de visualisation de données en Python qui permet de créer une grande variété de graphiques, tels que des graphiques linéaires, des graphiques à barres, des graphiques circulaires, des diagrammes de dispersion, des histogrammes, des cartes, des diagrammes en boîte, etc. Elle offre également des fonctionnalités avancées pour la personnalisation des graphiques, la gestion des axes, des légendes, des titres, des couleurs, des annotations, etc.[96]

Scikit-learn est une bibliothèque d'apprentissage automatique en Python qui propose une implémentation efficace d'un large éventail d'algorithmes d'apprentissage automatique, tels que la régression linéaire, la classification, le regroupement, l'analyse des composantes principales (PCA), les machines à vecteurs de support (SVM), les arbres de décision, les forêts aléatoires, les réseaux de neurones et bien d'autres. Elle fournit également des outils pour le prétraitement des données, la sélection de fonctionnalités, la validation croisée, l'optimisation des hyperparamètres et l'évaluation des modèles.[97]

### 3.4 DESCRIPTION DU DATASET

#### 3.4.1 Définition l'ensemble de données utilisé

C'est un ensemble de données sur le diabète provient de l'hôpital de Frankfort en Allemagne et a été téléchargé depuis Kaggle [98]. Il est au format CSV, ce qui facilite son traitement avec Python. . La taille du Dataset est 62,06 kB et contient des informations sur 2000 patients, qu'ils soient diabétiques ou non.

Le Dataset est composé de neuf (9) colonnes :

- **Glucose** : Il représente la concentration plasmatique de glucose mesurée deux heures après un test oral de tolérance au glucose.
- **Pregnancies** : Il indique le nombre de grossesses qu'a connues la patiente.
- **BloodPressure** : Il donne la pression artérielle diastolique mesurée en millimètres de mercure (mm Hg).
- **SkinThickness** : Il fournit l'épaisseur du pli cutané du triceps mesurée en millimètres (mm).

- **Insulin** : Il indique le niveau d'insuline sérique mesuré deux heures après une injection, exprimé en unités internationales par millilitre (mu U/ml).
- **BMI** : C'est l'Indice de Masse Corporelle (IMC) calculé en divisant le poids en kilogrammes par le carré de la taille en mètres (kg/m<sup>2</sup>).
- **DiabetesPedigreeFunction** : Il représente la fonction généalogique liée au diabète.
- **Âge** : C'est l'âge du patient en années.
- **Outcome** : Il s'agit de la variable de classe qui prend les valeurs 0 ou 1. 0 signifie que le patient ne souffre pas de diabète, tandis que 1 indique que le patient est diabétique.

Variable	Description	Analyse de données
Glucose	Une valeur de 2 heure entre (140 et 200 mg)/dl (7.8 et 11.1 mmol/L) est appelé tolérance au glucose altère signifie que il y a un risque accru de développer le diabète au fil de temps. Un taux de glucose de 200 mg/dL (11.1 mmol/L) ou plus utilisé pour diagnostiquer le diabète.	Minimum : 0 Maximum : 199
Pregnancies	Nombre de fois enceinte	Minimum : 0 Maximum : 17
BloodPressure	Si un TA diastolique supérieur à 90 signifie une pression artérielle élevée (probabilité élevée de diabète) Un TA diastolique inférieur à 60 signifie une pression artérielle basse (moins probabilité de diabète)	Minimum : 0 Maximum : 122
SkinThickness	Valeur estimée pour la graisse corporelle. épaisseur normale du pli cutané chez les femmes est de 23 mm. Une épaisseur plus élevée conduit à l'obésité et les chances de diabète augmente	Minimum : 0 Maximum : 110
Insulin	poids en kg / taille en m <sup>2</sup> ) IMC de 18.5 à 20 c'est normal IMC entre 25 et 30 situer dans une plage surpoids Et de 30 ou plus situer dans la fourchette d'obésité	Minimum : 0 Maximum : 80.6
DiabetePredigme Function	Fournit des informations sur les antécédentes chez les parents et la relation génétique avec les patients. Une fonction de pedigree plus élevée signifie que le patient plus susceptible de souffrir un diabète	Minimum : 0.078 Maximum : 2.42
Age	Age d'une personne en années	Minimum : 21 Maximum : 81
Outcome	Indique si une personne est diabétique ou non	0(non diabétique) :1316 1(diabétique) : 684

*Tableaux 3. 1: Description des variables d'ensemble de données*

### 3.4.2 Les étapes de pré-traitement de données

La création d'un modèle de Deep Learning est un processus complexe qui se déroule en plusieurs étapes. Chaque étape présente ses propres défis techniques et conceptuels. Le prétraitement des données pour le deep learning implique la visualisation des données afin d'obtenir des informations et d'effectuer des analyses sur les caractéristiques d'un ensemble de données. Il englobe également le nettoyage des données, qui consiste à supprimer ou à corriger les enregistrements contenant des valeurs corrompues ou non valides, dans le but d'améliorer la qualité des données. Enfin, la sélection du modèle joue un rôle crucial, car il doit être capable de fournir des prédictions plus précises que les autres modèles candidats.[99]

```
[ ] df = pd.read_csv('/content/diabetes (3).csv')
df.head()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	62	35	0	33.6	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.630	31	1
3	0	135	68	42	250	42.3	0.365	24	1
4	1	139	62	41	480	40.7	0.536	21	0

```
[ ] df.tail()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1995	2	75	64	24	55	29.7	0.370	33	0
1996	8	179	72	42	130	32.7	0.719	36	1
1997	6	85	78	0	0	31.2	0.382	42	0
1998	0	129	110	46	130	67.1	0.319	26	1
1999	2	81	72	15	76	30.1	0.547	25	0

*Figure 3. 3: Données d'entraînement*

### 3.4.3 Exploration et visualisation de données

La visualisation des données implique l'exploration visuelle et interactive de jeux de données de différentes tailles. Elle permet de mettre en évidence des informations qui étaient auparavant difficiles à percevoir. La visualisation facilite la communication universelle des informations et favorise le partage d'idées avec autrui. En ce qui concerne l'ensemble de données, il présente des caractéristiques similaires On remarque que :

- Le nombre des observations (2000 patients) dont 1316 diabétiques et 684 non diabétiques.
- les technique utilisées pour prédire le diabète.
- Les valeurs manquantes.
- La taille de dataset...

#### ➤ Visualisation de données

```
[ ] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  ---                -
0   Pregnancies           2000 non-null   int64
1   Glucose               2000 non-null   int64
2   BloodPressure         2000 non-null   int64
3   SkinThickness         2000 non-null   int64
4   Insulin               2000 non-null   int64
5   BMI                   2000 non-null   float64
6   DiabetesPedigreeFunction 2000 non-null   float64
7   Age                   2000 non-null   int64
8   Outcome               2000 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 140.8 KB

df.shape
(2000, 9)
```

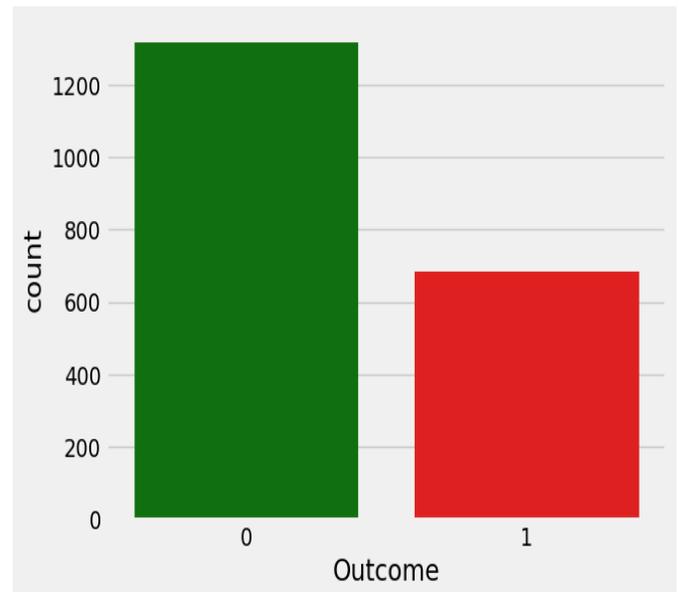
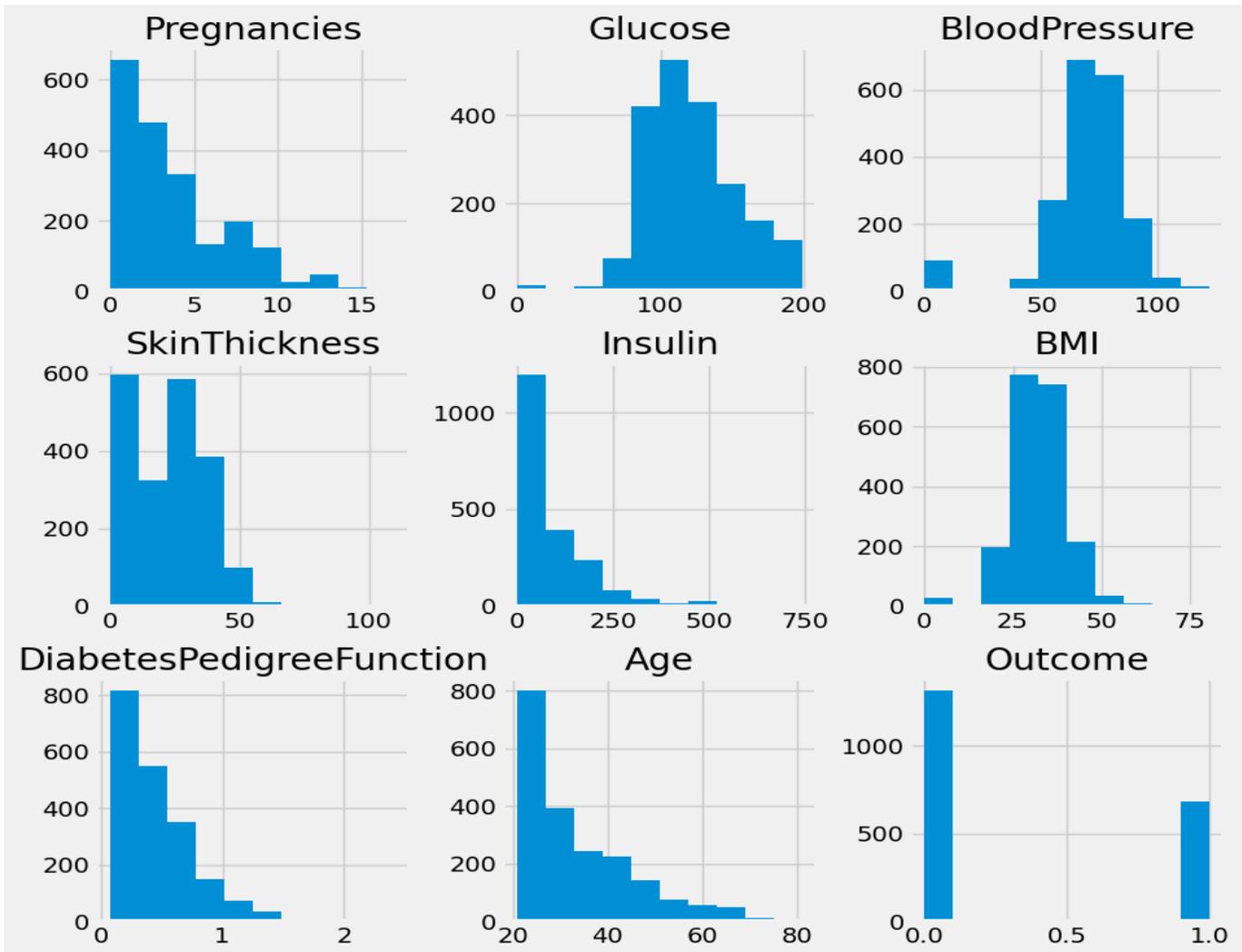


Figure 3. 4: Informations de la base de données

➤ **Distribution des variables**



*Figure 3.5: Le graphe histogramme des colonnes*

### 3.4.4 Nettoyage des données

Dans notre ensemble de données, les variables Glucose, BloodPressure, SkinThickness, Insulin et BMI contiennent des valeurs nulles (0). Par conséquent, toutes les valeurs nulles ont été remplacées. Cette étape a été effectuée afin de réduire les risques associés à ces valeurs manquantes dans l'ensemble de données.

Nous avons remplacé les cellules vides par une valeur de 0 avec NaN (pas un nombre). Seules les colonnes "Grossesses" et "Outcome" ont été laissées intactes, car une valeur de 0 peut être valide dans ces cas. Cela nous permet de traiter les données plus efficacement.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DPF	Age	Outcome
0	2	138.0	62.0	35.0	NaN	33.6	0.127	47	1
1	0	84.0	82.0	31.0	125.0	38.2	0.233	23	0
2	0	145.0	NaN	NaN	NaN	44.2	0.630	31	1
3	0	135.0	68.0	42.0	250.0	42.3	0.365	24	1
4	1	139.0	62.0	41.0	480.0	40.7	0.536	21	0
5	0	173.0	78.0	32.0	265.0	46.5	1.159	58	0
6	4	99.0	72.0	17.0	NaN	25.6	0.294	28	0
7	8	194.0	80.0	NaN	NaN	26.1	0.551	67	0
8	2	83.0	65.0	28.0	66.0	36.8	0.629	24	0
9	2	89.0	90.0	30.0	NaN	33.5	0.292	42	0

*Figure 3. 6: Après le remplacement des valeurs zéro*

Lors de l'analyse visuelle des données, nous avons identifié la présence de valeurs aberrantes dans certaines colonnes :

- **Blood Pressure** : Il est impossible qu'une personne ait une pression artérielle diastolique de 0 mm Hg, et nous avons observé 90 occurrences de cette valeur dans la colonne.
- **BMI** : Le poids d'une personne ne peut jamais être égal à 0 ou proche de zéro, car cela mettrait sa vie en danger. Nous avons constaté la présence de 28 occurrences avec une valeur de 0 dans cette colonne.
- **Glucose** : Le taux de glucose ne devrait jamais être égal à 0, cependant, après analyse des données, nous avons identifié 13 occurrences avec cette valeur dans la colonne.
- **Insulin** : Bien qu'il soit rare qu'une personne ait une valeur de 0 pour l'insuline, nous avons observé 956 occurrences de cette valeur dans cette colonne.
- **Pregnancies** : Il est normal d'avoir une valeur de zéro pour cette colonne, il n'est donc pas nécessaire de la nettoyer.
- **Skin Thickness** : Pour une personne en bonne santé, l'épaisseur du pli cutané ne peut pas être inférieure à 10 mm, mais nous avons trouvé 573 occurrences avec une valeur de 0 dans cette colonne. Ces valeurs aberrantes doivent être traitées afin de garantir l'intégrité et la qualité des données utilisées dans notre analyse.

### 3.4.5 Corrélation

Un ensemble de données de qualité se caractérise par des caractéristiques fortement corrélées à la variable cible et fortement non corrélées entre elles. L'objectif est de trouver des attributs qui présentent peu de corrélation entre eux.

Le coefficient de corrélation est une mesure qui quantifie la force de la relation entre deux variables. Il existe différents types de coefficients de corrélation, mais le plus couramment utilisé est le coefficient de corrélation de Pearson, noté "r". Ce coefficient est défini comme le rapport entre la covariance des variables X et Y et le produit de leurs écarts-types :  $r = \text{Cov}(X, Y) / (\sigma_X * \sigma_Y)$ .

Le coefficient de corrélation de Pearson varie de -1 à 1. Une valeur de -1 indique une corrélation négative parfaite, 1 indique une corrélation positive parfaite, et 0 indique l'absence de corrélation linéaire entre les variables.[100]

L'analyse de corrélation est importante dans le domaine de l'analyse des données, car elle permet de comprendre les relations entre les variables et d'identifier les caractéristiques les plus pertinentes pour la prédiction ou l'analyse de la variable cible.

La Figure 3.7 met en évidence l'importance des caractéristiques telles que l'âge, la glycémie (Glucose) et l'indice de masse corporelle (BMI) pour le diagnostic de la maladie du diabète. Ces caractéristiques sont fortement indépendantes les unes des autres.

En examinant la visualisation des données, on constate qu'il n'y a aucun point de données manquantes ou nulles dans l'ensemble des données.

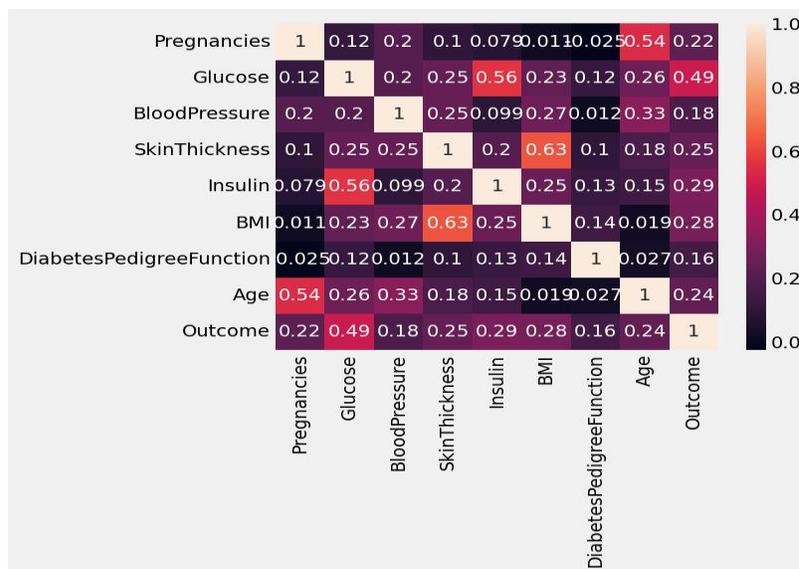


Figure 3. 7: Matrice de corrélation

### 3.4.6 Proportion des variables

La Figure 3.8 met en évidence la répartition des variables en fonction du statut de diabète (diabétique ou non diabétique). On remarque des variations significatives entre les deux groupes. Après avoir effectué le prétraitement des données et éliminé les valeurs nulles, les distributions des variables ont été ajustées pour mieux correspondre aux valeurs normales attendues.

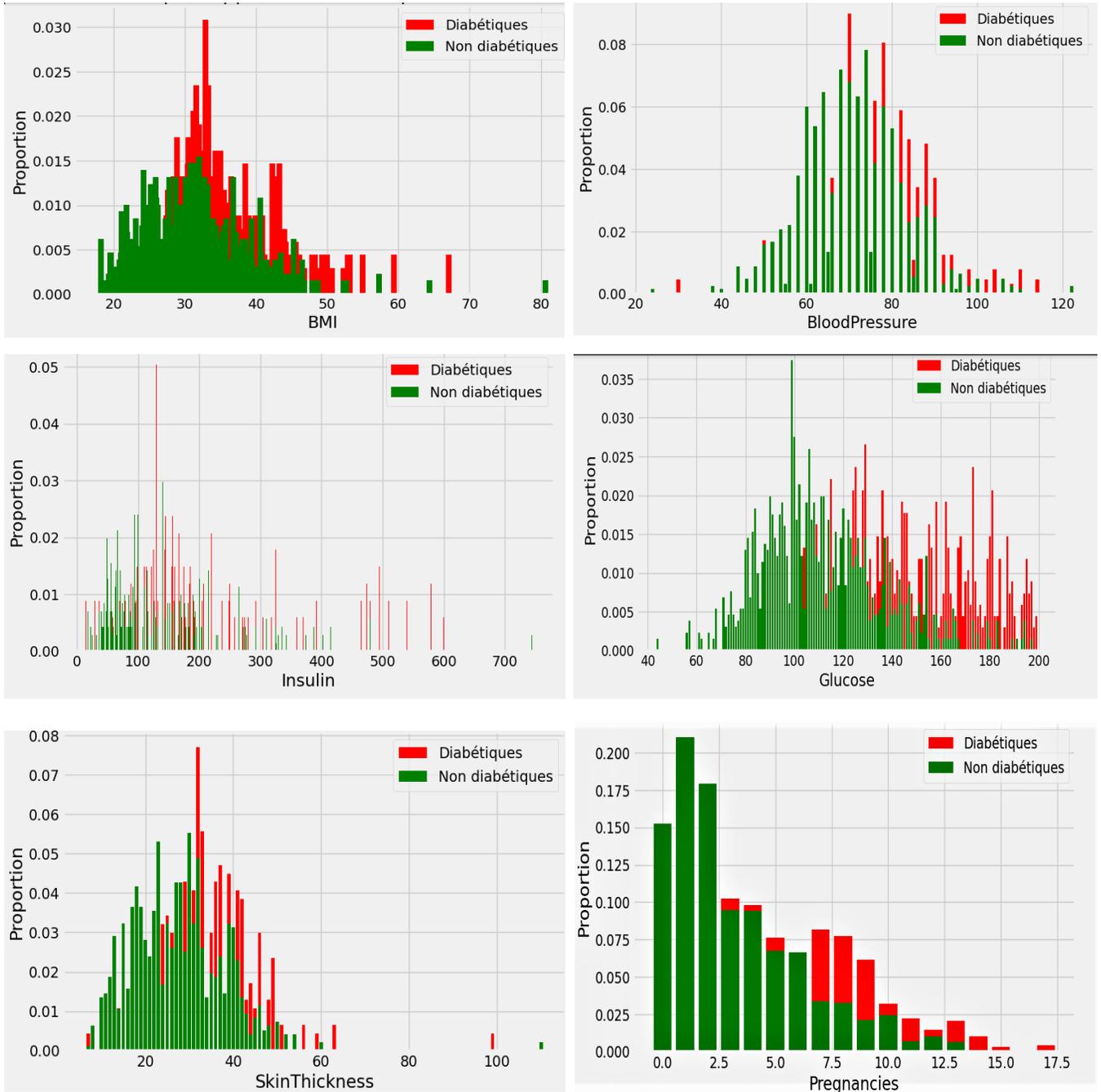


Figure 3. 8: Proportion des variables pour les personnes diabétiques et non diabétique

### 3.5 Entraînement et Test des données

Afin d'obtenir de bons résultats de prédiction, il est crucial de former et de tester un modèle. Cette approche implique la division de l'ensemble de données en deux parties distinctes : l'ensemble d'entraînement, sur lequel le modèle apprend, et l'ensemble de test, sur lequel le modèle est évalué et ses performances sont mesurées. L'objectif est de sélectionner les classifieurs qui fonctionnent de manière optimale sur les deux ensembles de données, ce qui se traduit par une précision améliorée.

#### 3.5.1 Sélection de modèle

Après la division des données, la prochaine étape consiste à choisir un modèle pour prédire le diabète. Cette phase implique l'exploration de divers algorithmes d'apprentissage automatique et d'apprentissage profond. Quatre modèles ont été utilisés pour la prédiction précoce du diabète, à savoir:

1. Decision Trees (Arbre de décision)
2. Deep Neural Network (DNN)
3. Support Vector Machine (SVM)
4. Les réseaux neuronaux à long terme et court terme (LSTM)

#### ➤ Train/Test Split

Cette approche implique la division de l'ensemble de données en deux parties distinctes : une partie d'entraînement sur laquelle le modèle effectue son apprentissage, et une partie de test utilisée pour évaluer les performances du modèle.

#### 3.5.2 Diviser le data entre deux partie (train ,test)

```
[ ] x=df.drop('Outcome',axis=1)
    y=df['Outcome']

[ ] from sklearn.model_selection import train_test_split
    x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2, random_state=7)

[ ] x_train.shape,y_train.shape

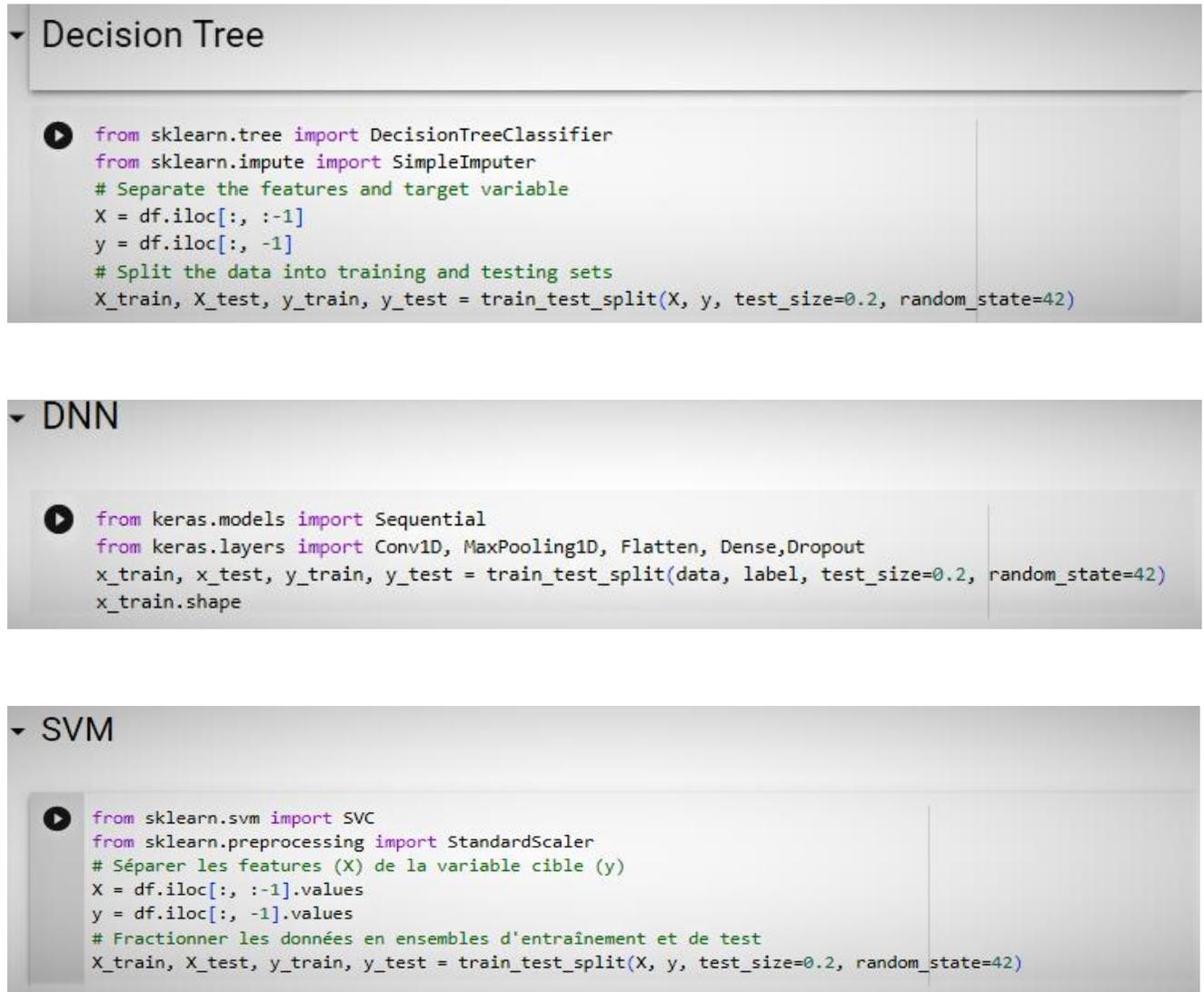
((1600, 8), (1600,))

[ ] x_test.shape,y_test.shape

((400, 8), (400,))
```

*Figure 3. 9: Aperçu la division des données*

### 3.5.3 Définir les modèles et passer les données pour trainer



*Figure 3. 10: Définir les (04) modèles*

➤ **Construction du modèle DNN:** Une fois que les données du dataset ont été préparées et divisées en ensembles de test et d'apprentissage, nous sommes prêts à construire notre modèle de réseau de neurones. Pour cela, nous avons utilisé la bibliothèque Keras.

```
# Define the model architecture
model2 = Sequential()
model2.add(Dense(500, input_dim=8, activation='sigmoid'))
model2.add(Dense(100, activation='sigmoid'))
model2.add(Dense(2, activation='softmax'))
```

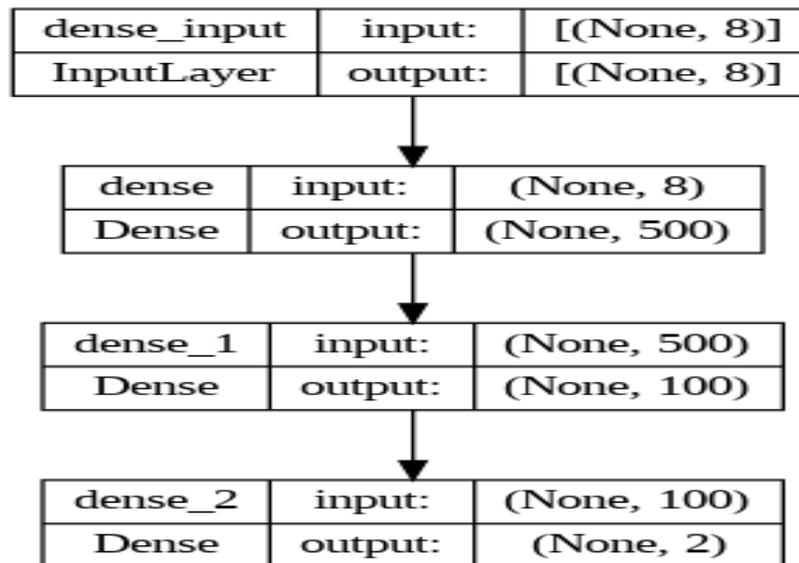


Figure 3. 11: Définition du modèle DNN

- **model2 = Sequential():** Cette ligne crée un modèle séquentiel, qui est une pile linéaire de couches. Cela nous permet de construire facilement le modèle en ajoutant des couches les unes après les autres.
- **model2.add(Dense(500, input\_dim=8, activation='sigmoid')):** ajoute une couche dense au modèle. La couche **Dense** représente une couche entièrement connectée dans le réseau neuronal. Dans ce cas, nous spécifions que la couche a 500 unités/neurones. Le paramètre **input\_dim** indique la forme d'entrée de la couche, qui est de 8 dans ce cas. Le paramètre **activation='sigmoid'** spécifie la fonction d'activation à appliquer aux sorties de la couche.
- **model2.add(Dense(100, activation='sigmoid')):** ajoute une autre couche dense au modèle avec 100 unités/neurones. Elle ne nécessite plus le paramètre **input\_dim** car elle prend automatiquement la forme de sortie de la couche précédente comme forme d'entrée.
- **model2.add(Dense(2, activation='softmax')):** ajoute la dernière couche dense au modèle avec 2 unités/neurones. Cette couche représente la couche de sortie du modèle et le paramètre

**activation='softmax'** spécifie la fonction d'activation à appliquer. Dans ce cas, l'activation softmax est utilisée pour obtenir des distributions de probabilité sur les classes.

- **model2.compile(loss='mean\_squared\_error', optimizer='adam', metrics=['accuracy']):** Cette ligne compile le modèle. Le paramètre **loss** spécifie la fonction de perte à minimiser lors de l'entraînement, qui est l'erreur quadratique moyenne dans ce cas. Le paramètre **optimizer** détermine l'algorithme d'optimisation à utiliser, et **'adam'** est un choix populaire. Le paramètre **metrics** spécifie les métriques d'évaluation à calculer pendant l'entraînement, dans ce cas, l'exactitude (accuracy).
- **model2.fit(x\_train, y\_train, epochs=1000, batch\_size=70, validation\_data=(x\_test, y\_test)):** entraîne le modèle. La fonction **fit** est utilisée pour entraîner le modèle sur les données d'entraînement (**x\_train** et **y\_train**). Le paramètre **epochs** spécifie le nombre de fois où les données d'entraînement doivent être itérées. Le paramètre **batch\_size** détermine le nombre d'échantillons par mise à jour du gradient. Le paramètre **validation\_data** est utilisé pour évaluer les performances du modèle sur les données de validation (**x\_test** et **y\_test**) après chaque époque.

### ➤ Construction du modèle LSTM:

```
# Créer le modèle LSTM
model = Sequential()
model.add(LSTM(64, input_shape=(1, X_train.shape[2])))
model.add(Dense(1, activation='sigmoid'))
```

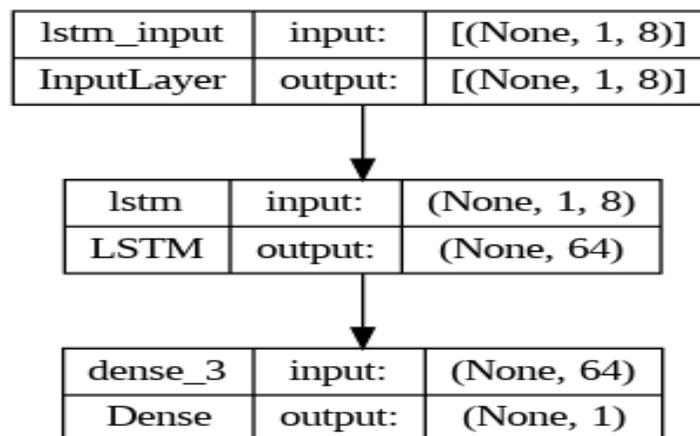


Figure 3. 12: Création de modèle LSTM

### ➤ La compilation du modèle :

La fonction **model2.compile** est utilisée pour compiler le modèle de réseau de neurones avec les paramètres spécifiés. Voici les détails de chaque paramètre:

- **optimizer='adam'**: Le paramètre **optimizer** spécifie l'algorithme d'optimisation à utiliser pour mettre à jour les poids du modèle pendant l'entraînement. Dans ce cas, nous utilisons l'optimiseur Adam, qui est un algorithme d'optimisation populaire pour les réseaux de neurones.
- **loss='binary\_crossentropy'**: Le paramètre **loss** spécifie la fonction de perte à minimiser lors de l'entraînement du modèle. Dans le cas d'un problème de classification binaire, la fonction de perte couramment utilisée est la `binary_crossentropy` (entropie croisée binaire). Cette fonction mesure la distance entre les prédictions du modèle et les valeurs réelles dans un problème de classification binaire.
- **metrics=['binary\_accuracy']**: Le paramètre **metrics** spécifie les métriques d'évaluation à calculer pendant l'entraînement et l'évaluation du modèle. Dans ce cas, nous utilisons la métrique `'binary_accuracy'` (exactitude binaire) pour évaluer la performance du modèle en termes de classification binaire. Cette métrique calcule le taux d'exactitude du modèle en prédisant correctement les étiquettes binaires.

```
# Compile the model
model2.compile(
    optimizer='adam',
    loss='binary_crossentropy',
    metrics=['binary_accuracy'],
)
```

*Figure 3. 13: Compilation du modèle*

➤ **Ajustement du modèle** : Pour adapter notre modèle aux données chargées, nous utilisons la fonction `fit()` sur le modèle, ce qui nous permet de le former en fonction de ces données.

**history=model2.fit(x\_train, y\_train, epochs=1000, batch\_size=70, validation\_split=0.4)**: Cette ligne entraîne le modèle `model2` sur les données d'entraînement `x_train` avec les étiquettes correspondantes `y_train`. Voici ce que font les différents paramètres :

**epochs=1000** : Spécifie le nombre d'époques d'entraînement, c'est-à-dire le nombre de fois que le modèle parcourra l'ensemble de données complet. Dans ce cas, le modèle sera entraîné sur 1000 époques.

**batch\_size=70** : Définit la taille des mini-lots utilisés pour la mise à jour des poids du modèle. Ici, chaque mini-lot contiendra 70 exemples d'entraînement. Les poids du modèle seront mis à jour après chaque mini-lot.

**validation\_split=0.4** : Spécifie la fraction des données d'entraînement à utiliser comme ensemble de validation. Dans ce cas, 40% des données d'entraînement seront utilisées pour la validation pendant

l'entraînement. Cela permet d'évaluer les performances du modèle sur un ensemble de données distinct et de détecter le surapprentissage.

L'historique de l'entraînement est stocké dans la variable `history`. Cela permet d'accéder aux métriques et aux valeurs de perte calculées pour chaque époque et pour les ensembles d'entraînement et de validation.

```
# Train the model
history = model2.fit(x_train, y_train, epochs=1000, batch_size=70, validation_split=0.2)
y_pred_prob = model2.predict(x_test)
y_pred = (y_pred_prob > 0.5).astype(int)
```

*Figure 3. 14: Entraîner Le modèle*

### 3.5.4 Evaluations des modèles

Pour évaluer la performance d'un modèle de classification et fournissent des indications sur sa capacité à prédire correctement les classes. Il est important de considérer ces mesures : Accuracy, La précision , recall score, F1 score dans le contexte spécifique du problème et de l'ensemble de données pour obtenir une évaluation précise de la performance du modèle.

**1. Accuracy (Exactitude) :** L'accuracy est une mesure courante pour évaluer la performance globale d'un modèle de classification. Elle représente le taux de prédictions correctes par rapport au nombre total d'échantillons. La formule mathématique de l'accuracy est [101]:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Où TP (True Positive) représente le nombre de vrais positifs, TN (True Negative) représente le nombre de vrais négatifs, FP (False Positive) représente le nombre de faux positifs et FN (False Negative) représente le nombre de faux négatifs. L'accuracy donne une indication générale de la performance du modèle, mais elle peut être biaisée si les classes sont déséquilibrées.

**2. Precision (Précision) :** La précision mesure la proportion d'échantillons positifs prédits correctement par rapport à tous les échantillons prédits comme positifs. La formule mathématique de la précision est **Precision** =  $TP / (TP + FP)$

Une précision élevée indique que le modèle a tendance à prédire correctement les échantillons positifs, minimisant ainsi les faux positifs.

**3. (Rappel) :** Le recall mesure la proportion d'échantillons positifs prédits correctement par rapport à tous les échantillons réellement positifs. La formule mathématique du recall est :

$$\text{Recall} = TP / (TP + FN)$$

Un recall élevé indique que le modèle a tendance à identifier correctement la majorité des échantillons positifs, minimisant ainsi les faux négatifs.

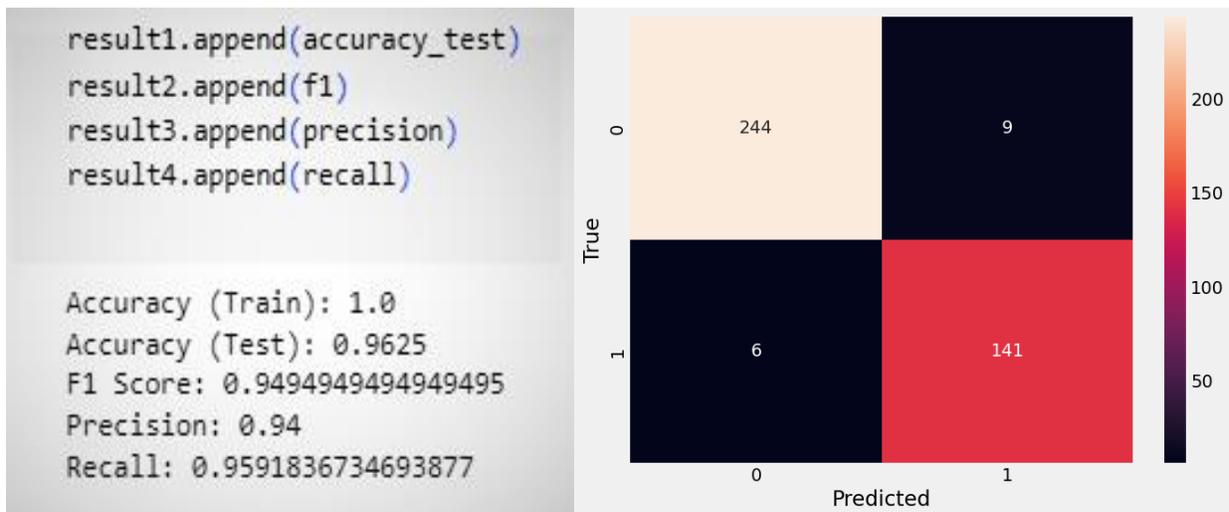
**4. F1 score :** Le F1 score est une mesure qui combine à la fois la précision et le recall en une seule valeur. Il fournit une mesure globale de la performance du modèle en tenant compte à la fois des vrais positifs et des faux positifs. Le F1 score est calculé à partir de la formule mathématique suivante :

$$\mathbf{F1\ score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

Le F1 score donne une idée globale de l'équilibre entre la précision et le recall du modèle. Il est utile lorsque les classes sont déséquilibrées ou lorsque les erreurs de faux positifs et de faux négatifs doivent être prises en compte simultanément.

➤ **Decision Trees (Arbre de décision)**

**La précision= 94%**



*Figure 3.15: évaluation de modèle Decision Trees*

➤ Deep Neural Network (DNN)

La précision= 92%

```
Epoch 995/1000
19/19 [=====] - 0s 12ms/step - loss: 0.036
Epoch 996/1000
19/19 [=====] - 0s 10ms/step - loss: 0.038
Epoch 997/1000
19/19 [=====] - 0s 8ms/step - loss: 0.0354
Epoch 998/1000
19/19 [=====] - 0s 8ms/step - loss: 0.0349
Epoch 999/1000
19/19 [=====] - 0s 8ms/step - loss: 0.0425
Epoch 1000/1000
19/19 [=====] - 0s 7ms/step - loss: 0.0312
13/13 [=====] - 0s 3ms/step
Accuracy = 0.9693750143051147
F1 Score = 0.9222087432878067
Precision = 0.9222277268260727
Recall = 0.9225
```

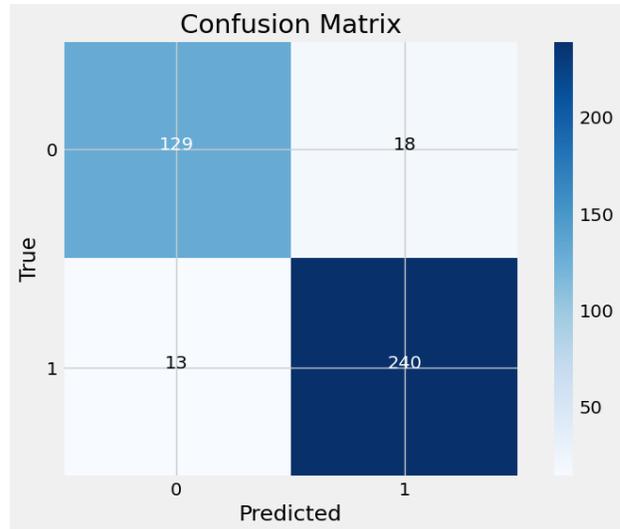


Figure 3. 16: évaluation de modèle DNN

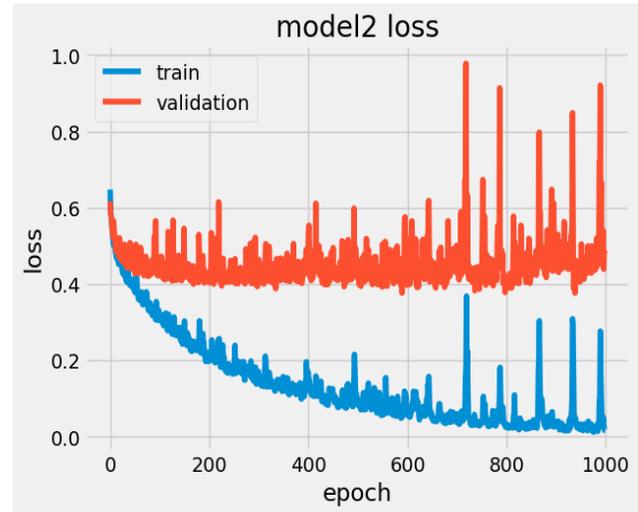
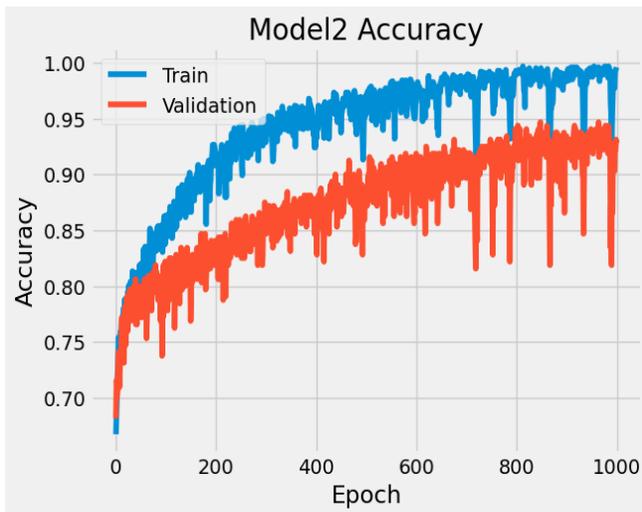


Figure 3. 17: Plote de modèle accuracy et loss

➤ Support Vector Machine (SVM)

La précision= 75%

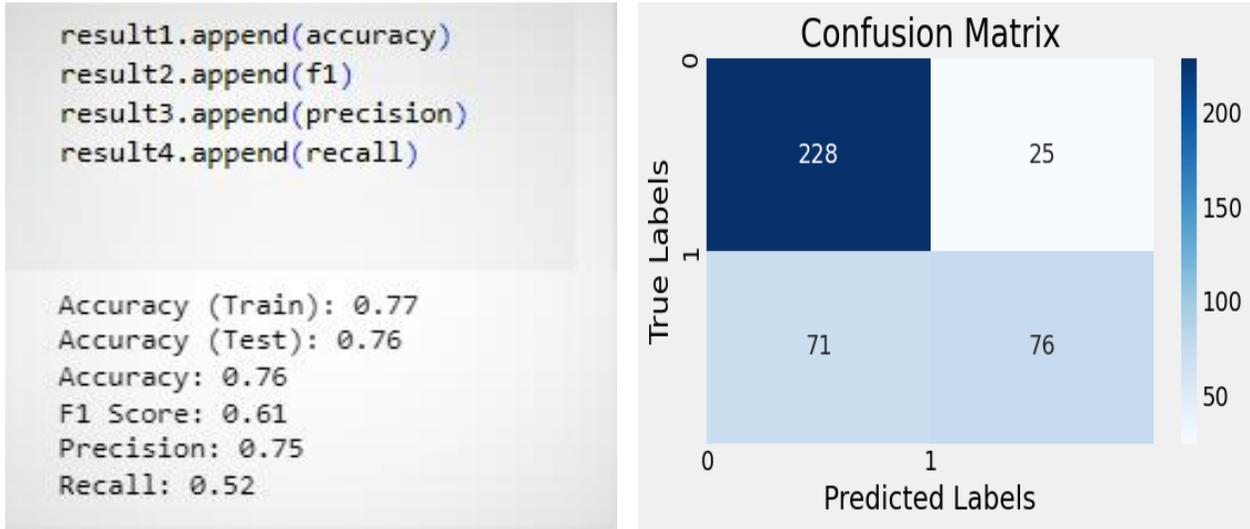


Figure 3. 18: évaluation de modèle SVM

➤ Les réseaux neuronaux à long terme et court terme (LSTM)

La précision= 99%

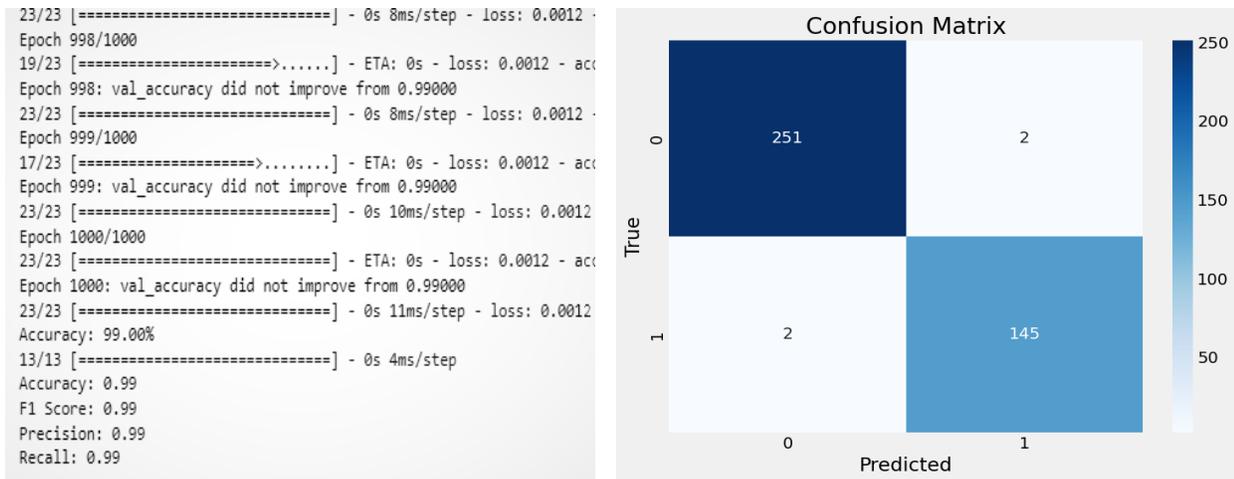


Figure 3. 19: évaluation de modèle LSTM

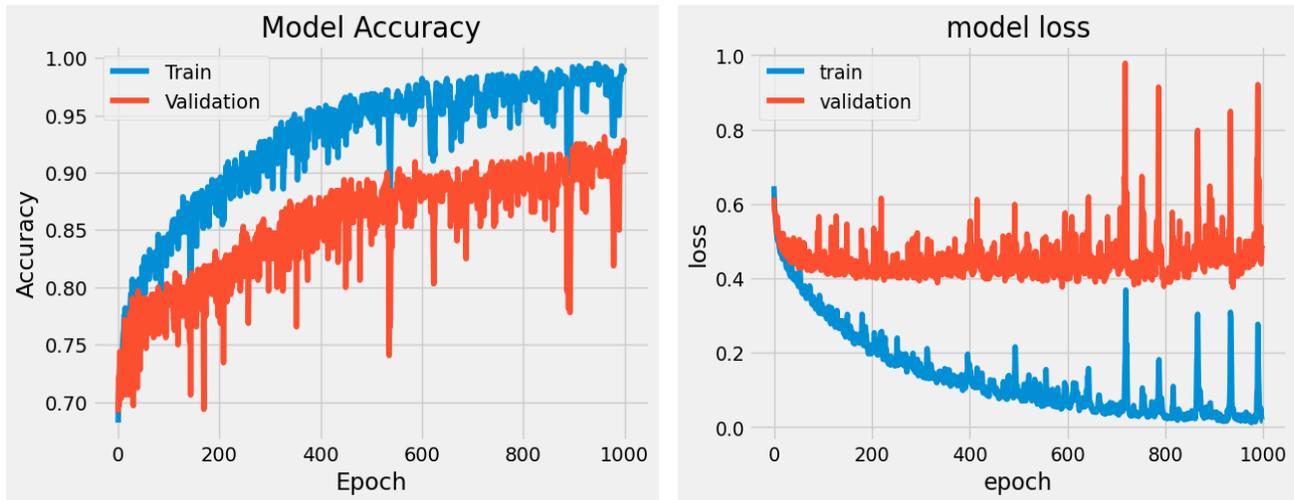


Figure 3. 20: Plote de modèle accuracy et loss

### 3.5.5 Comparaison entre les différents modèles

Algorithms	Accuracies	F1 Score	precision	recall
LSTM	0.990000	0.986395	0.986395	0.986395
DecisionTreeClassifier	0.962500	0.949495	0.940000	0.959184
Sequential	0.969375	0.922209	0.922228	0.922500
SVC	0.760000	0.612903	0.752475	0.517007

Figure 3. 21: Les résultats des attributs d'évaluations pour les différents modèles

D'après le tableau ci-dessus le modèle LSTM obtenu la meilleure précision qui égal à 99% c'est-à-dire que sur toutes les patients diabétiques 99% d'entre eux sont correctement classé à l'aide de mesure de diagnostiques médicales. Nous sélectionnons le LSTM comme le modèle le plus optimale et qui fonctionne mieux pour notre ensemble de données en raison de sa grande précision et score de rappel.

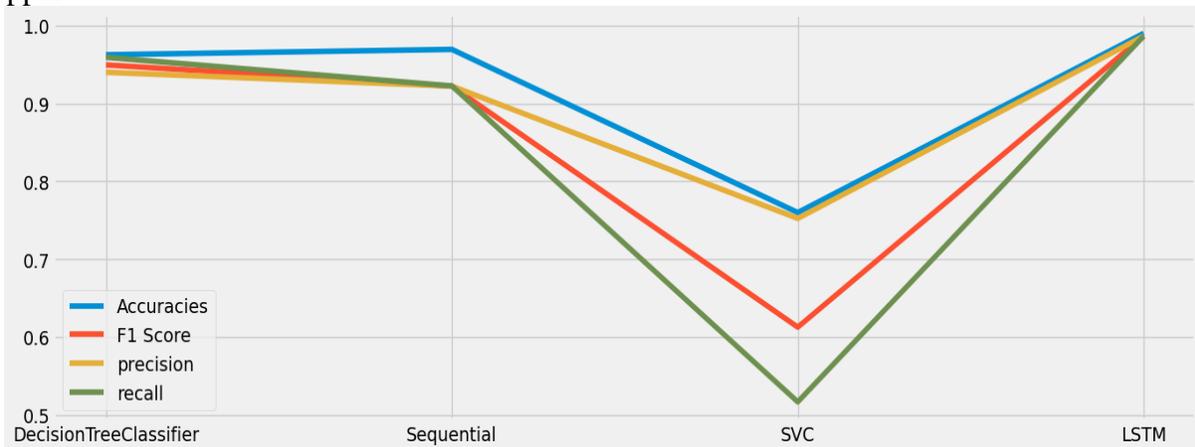


Figure 3. 22: Représentation graphique de Accuracy, F1 score, precision et recall

En analysant ces résultats, on peut conclure que le modèle LSTM est le plus performant avec une précision élevée, un F1-score équilibré et un rappel parfait. Le modèle DNN et le modèle Decision Tree Classifier montrent également de bonnes performances, tandis que le modèle SVM a des résultats relativement moins satisfaisants.

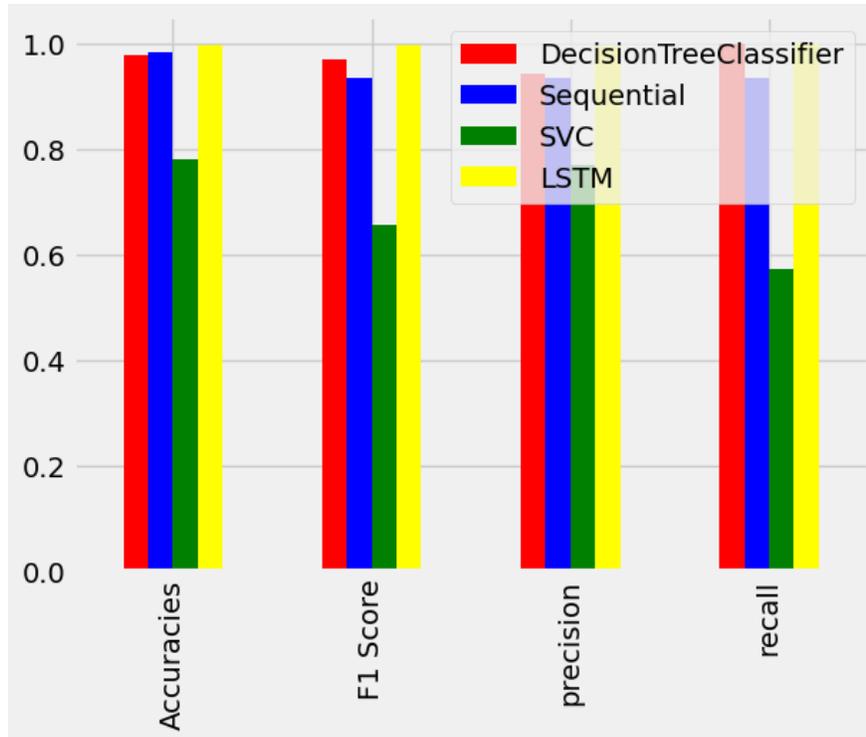


Figure 3. 23: Plot de Comparaison entre les différents modèles

### 3.6 CONCLUSION

Dans ce chapitre, nous avons abordé plusieurs étapes essentielles du prétraitement des données, telles que l'exploration et la visualisation des données, ainsi que le nettoyage des valeurs aberrantes. Notre objectif principal était de développer un modèle capable de prédire avec précision si une personne est diabétique ou non. Pour atteindre cet objectif, nous avons exploré et expérimenté différents algorithmes tels que les arbres de décision, les réseaux neuronaux profonds (DNN), les machines à vecteurs de support (SVM) ainsi que les réseaux neuronaux de longue mémoire à court terme (LSTM).

Grâce à l'application rigoureuse des méthodes d'évaluation, nous avons comparé les performances de ces différents algorithmes. Nos résultats ont démontré que le modèle LSTM se démarque comme le meilleur choix, offrant un taux de précision élevé. Cette conclusion solide renforce notre confiance dans l'utilisation du LSTM pour la prédiction du diabète .

# Conclusion générale et perspectives

### CONCLUSION GÉNÉRALE ET PERSPECTIVES

Le diabète est une maladie chronique qui a un impact profond sur la vie quotidienne et la santé des individus. En tant que pandémie mondiale, sa prévalence ne cesse d'augmenter, touchant des millions de personnes dans le monde entier. En Algérie, la prévalence estimée est d'environ 14,4%, ce qui souligne l'ampleur du problème et l'importance de trouver des solutions efficaces.

L'utilisation de systèmes d'intelligence artificielle et du deep learning dans la gestion du diabète offre des perspectives prometteuses. Ces technologies permettent une analyse plus approfondie des données médicales, ce qui peut conduire à une détection plus précoce du diabète, une prise de décision plus éclairée et une meilleure gestion des patients. Les résultats obtenus dans notre étude démontrent l'efficacité de ces techniques dans la prédiction du diabète.

Dans cette mémoire, nous avons abordé le défi de l'identification et de la prédiction du diabète en utilisant des techniques d'apprentissage automatique et approfondi. Notre objectif était de construire un modèle capable de prédire si une personne est diabétique ou non en se basant sur des algorithmes de machine learning et deep learning. Pour ce faire, nous avons comparé quatre algorithmes populaires : Decision Trees (Arbre de décision), Deep Neural Network (DNN), Support Vector Machine (SVM), Les réseaux neuronaux de longue mémoire à court terme (LSTM). Les résultats expérimentaux obtenus à partir de l'ensemble de données de l'hôpital de Frankfort en Allemagne ont montré que LSTM était le meilleur algorithme en termes de précision, à la fois pour les méthodes d'évaluation et pour les attributs d'évaluation.

Cependant, il convient de souligner certaines limites et perspectives pour de futures recherches. Tout d'abord, l'accès à des ensembles de données de grande qualité et de grande taille est essentiel pour améliorer la précision des modèles de prédiction. De plus, la collaboration entre les professionnels de la santé et les experts en intelligence artificielle est nécessaire pour développer des solutions adaptées aux besoins réels des patients diabétiques.

En ce qui concerne les perspectives d'avenir, nous envisageons d'explorer l'application du deep learning dans la prédiction du diabète. Cette approche utilise des réseaux de neurones profonds pour analyser de grandes quantités de données médicales et identifier des modèles et des tendances prédictifs. En exploitant les avantages du deep learning, nous pourrions obtenir des résultats encore plus précis et personnalisés pour la prédiction du diabète.

## CONCLUSION GÉNÉRALE

---

La réalisation de ce travail nous a permis d'acquérir une expertise approfondie dans les techniques de Machine Learning et de Deep Learning. Cette étude a été une opportunité précieuse pour développer nos compétences et nous a fourni des connaissances approfondies qui serviront de base solide pour nos futurs défis.

# Bibliographie

## **Bibliographie**

---

- [1] Briganti, G. (2023). Intelligence artificielle: une introduction pour les cliniciens. *Revue des Maladies Respiratoires*.
- [2] Ayadi, A., Sassi, S., & Tissaoui, A. Vers un Méta-Modèle de Système d'Aide à la Décision: application au domaine médical.
- [3] Systèmes d'aide à la décision - Pro Healthcare Services <https://prohealthware.com/fr/decision-support-systems/> (consulté le 07Avril 2023)
- [4] Ayadi, A., Sassi, S., & Tissaoui, A. Vers un Méta-Modèle de Système d'Aide à la Décision: application au domaine médical
- [5] Quel est le rôle des systèmes informatiques d'aide à la décision médicale ?Julien Claudet·15 septembre 2020 . (consulté le 08Avril 2023) <https://trustmyscience.com/role-systemes-informatiques-d-aide-decision-medicale/>
- [6] mondiale de la Santé, O. (2021). Améliorer le sort de toutes les personnes atteintes de diabète cent ans après la découverte de l'insuline: rapport du Sommet mondial sur le diabète.
- [7] American Diabetes Association. (2022). Introduction: standards of medical care in diabetes—2022. *Diabetescare*, 45(Supplement\_1),S1-S2..
- [8] National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). (2021). Diabetes Overview. <https://www.niddk.nih.gov/health-information/diabetes/overview> (consulté le 15 Mars 2023)
- [9]Diabetes Canada. (s.d.). What is Diabetes? <https://www.diabetes.ca/about-diabetes/what-is-diabetes> (consulté le 15 Mars 2023)
- [10] Médart, J. (2009). *Manuel pratique de nutrition: l'alimentation préventive et curative*. De Boeck Supérieur.

## Bibliographie

---

- [11] Wendland, E. M., Torloni, M. R., Falavigna, M., Trujillo, J., Dode, M. A., Campos, M. A., ... & Schmidt, M. I. (2012). Gestational diabetes and pregnancy outcomes-a systematic review of the World Health Organization (WHO) and the International Association of Diabetes in Pregnancy Study Groups (IADPSG) diagnostic criteria. *BMC pregnancy and childbirth*, 12, 1-13.
- [12] Médart, J. (2009). *Manuel pratique de nutrition: l'alimentation préventive et curative*. De Boeck Supérieur.
- [13] Charbonnel, B., Simon, D., Dallongeville, J., Bureau, I., Gourmelen, J., & Detournay, B. (2017). Coût du diabète de type 2 en France: une analyse des données de l'EGB. *Médecine des Maladies Métaboliques*, 11, IIS24-IIS27..
- [14] Cheng, A. Y., Feig, D. S., Ho, J., Siemens, R., Bajaj, H., Gilbert, J., ... & Diabetes Canada Clinical Practice Guidelines Expert Working Group. (2021). Blood glucose monitoring in adults and children with diabetes: update 2021. *Canadian Journal of Diabetes*, 45(7), 580-587.
- [15] CHABANE Saliha, BOUDJENNAD Lynda, (2017/2018) L'estime de soi chez les étudiants diabétiques, Etude de quatre (04) inscrits à l'université Abderrahmane mira de Bejaia, au niveau du EUP centre médical de l'université de Bejaia, Université Abderrahmane mira Bejaia, Faculté des sciences humaines et sociales, Département des sciences sociales.
- [16] CHABANE Saliha, BOUDJENNAD Lynda,(2017/2018) L'estime de soi chez les étudiants diabétiques, Etude de quatre (04) inscrits à l'université Abderrahmane mira de Bejaia, au niveau du EUP centre médical de l'université de Bejaia, Université Abderrahmane mira Bejaia, Faculté des sciences humaines et sociales, Département des sciences sociales .
- [17] American Diabetes Association. (2021). Standards of Medical Care in Diabetes-2021. *Diabetes Care*, 44(Supplement 1), S15-S33. <https://doi.org/10.2337/dc21-S002>
- [18] International Diabetes Federation. (2019). IDF Clinical Practice Recommendations for managing Type 2 Diabetes in Primary Care. International Diabetes Federation. <https://www.idf.org/component/attachments/attachments.html?id=1678&task=download> (consulté le 17 Mars 2023)
- [19] American Diabetes Association. (2021). Standards of Medical Care in Diabetes-2021. *Diabetes Care*, 44(Supplement 1), S15-S33. <https://doi.org/10.2337/dc21-S002>

## **Bibliographie**

---

[20] International Diabetes Federation. (2019). IDF Clinical Practice Recommendations for managing Type 2 Diabetes in Primary Care. International Diabetes Federation.

<https://www.idf.org/component/attachments/attachments.html?id=1678&task=download> (consulté le 20 Mars 2023)

[21] American Diabetes Association. (2021). Standards of Medical Care in Diabetes-2021. Diabetes Care, 44(Supplement 1), S15-S33. <https://doi.org/10.2337/dc21-S002>

[22] International Diabetes Federation. (2019). IDF Clinical Practice Recommendations for managing Type 2 Diabetes in Primary Care. International Diabetes Federation.

<https://www.idf.org/component/attachments/attachments.html?id=1678&task=download> (consulté le 20 Mars 2023)

[23] American Diabetes Association. (2021). Diagnosis. Retrieved from

<https://www.diabetes.org/a1c/diagnosis>. (consulté le 21 Mars 2023)

[24] World Health Organization. (2020). Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia. Retrieved from <https://www.who.int/publications/i/item/9789241548009>. (consulté le 23 Mars 2023)

[25] Mayo Clinic. (2021). Blood sugar testing: Why, when and how. Retrieved from <https://www.mayoclinic.org/diseases-conditions/diabetes/in-depth/blood-sugar/art-20046628>. (consulté le 23 Mars 2023)

[26] National Institute of Diabetes and Digestive and Kidney Diseases. (2021). Continuous Glucose Monitoring. Retrieved from <https://www.niddk.nih.gov/health-information/diabetes/overview/tests-diagnosis/continuous-glucose-monitoring>. (consulté le 24 Mars 2023)

[27] Centers for Disease Control and Prevention. (2021). Diagnosis of Diabetes. Retrieved from <https://www.cdc.gov/diabetes/basics/diagnosis.html>. (consulté le 23 Mars 2023)

[28] American Diabetes Association. Type 1 Diabetes. <https://www.diabetes.org/diabetes/type-1> (consulté le 23 Mars 2023)

[29] Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2020. <https://www.cdc.gov/diabetes/library/features/diabetes-stat-report.html> (consulté le 21 Mars 2023)

## **Bibliographie**

---

[30] International Diabetes Federation. Type 1 Diabetes. <https://www.idf.org/our-activities/care-prevention/type-1-diabetes.html> . (consulté le 24 Mars 2023)

[31] Juvenile Diabetes Research Foundation. Type 1 Diabetes Facts. <https://www.jdrf.org/t1d-resources/about/facts/> (consulté le 24 Mars 2023)

[32] National Institute of Diabetes and Digestive and Kidney Diseases. Type 1 Diabetes. <https://www.niddk.nih.gov/health-information/diabetes/overview/type-1-diabetes>(consulté le 24 Mars 2023)

[33] American Diabetes Association. (2021). Type 2 Diabetes. Retrieved from <https://www.diabetes.org/diabetes/type-2>(consulté le 20 Mars 2023)

[34] World Health Organization. (2020). Diabetes. Retrieved from [https://www.who.int/health-topics/diabetes#tab=tab\\_1](https://www.who.int/health-topics/diabetes#tab=tab_1)(consulté le 23 Mars 2023)

[35] Centers for Disease Control and Prevention. (2021). National Diabetes Statistics Report, 2020. Retrieved from <https://www.cdc.gov/diabetes/library/features/diabetes-stat-report.html>(consulté le 25 Mars 2023)

[36] International Diabetes Federation. (2019). IDF Diabetes Atlas. Retrieved from <https://www.diabetesatlas.org/en/resources/>(consulté le 25 Mars 2023)

[37] American Heart Association. (2021). Type 2 Diabetes. Retrieved from <https://www.heart.org/en/health-topics/diabetes/type-2-diabetes>(consulté le 22 Mars 2023)

[38] National Institute of Diabetes and Digestive and Kidney Diseases. (2020). Type 2 Diabetes. Retrieved from <https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/type-2-diabetes>(consulté le 25 Mars 2023)

[39] Mayo Clinic. (2021). Type 2 Diabetes. Retrieved from <https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193>(consulté le 24 Mars 2023)

[40] Kaur, H., & Bishnoi, R. (2021). Type 2 Diabetes Mellitus. StatPearls [Internet]. Retrieved from <https://www.ncbi.nlm.nih.gov/books/NBK554640/> (consulté le 25 Mars 2023)

[41] American College of Cardiology. (2021). Type 2 Diabetes. Retrieved from

## Bibliographie

---

<https://www.acc.org/latest-in-cardiology/articles/2020/01/09/15/13/type-2-diabetes>(consulté le 26 Mars 2023)

[42]International Diabetes Federation. (2019). IDF Diabetes Atlas, 9th Edition. Retrieved from <https://www.diabetesatlas.org/en/sections/demographic-and-geographic-outline.html>(consulté le 26 Mars 2023)

[43]American Diabetes Association.(2021).Gestational Diabetes. <https://www.diabetes.org/diabetes/gestationaldiabetes>(consulté le 25 Mars 2023)

[44] Ministère des Solidarités et de la Santé. (2021). Diabète gestationnel. <https://solidarites-sante.gouv.fr/soins-et-maladies/maladies/maladies-chroniques/diabete/diabete-gestationnel>(consulté le 25 Mars 2023)

[45] National Institute of Diabetes and Digestive and Kidney Diseases. (2020). Gestational Diabetes. <https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/gestational>(consulté le 26 Mars 2023)

[46] Organisation Mondiale de la Santé. (2013). Diagnostic Criteria and Classification of Hyperglycaemia First Detected in Pregnancy. [https://www.who.int/diabetes/publications/Hyperglycaemia\\_In\\_Pregnancy/en/](https://www.who.int/diabetes/publications/Hyperglycaemia_In_Pregnancy/en/)(consulté le 26 Mars 2023)

[47]O'Sullivan, J. B., & Mahan, C. M. (1964). Criteria for the Oral Glucose Tolerance Test in Pregnancy. *Diabetes*, 13(4), 278-285. <https://doi.org/10.2337/diab.13.4.278>

[48] American Diabetes Association. Standards of Medical Care in Diabetes—2021 Abridged for Primary Care Providers. *Clinical Diabetes*. 2021;39(1):9-37.

[49] Wendland, E. M., Torloni, M. R., Falavigna, M., Trujillo, J., Dode, M. A., Campos, M. A., ... & Schmidt, M. I. (2012). Gestational diabetes and pregnancy outcomes-a systematic review of the World Health Organization (WHO) and the International Association of Diabetes in Pregnancy Study Groups (IADPSG) diagnostic criteria. *BMC pregnancy and childbirth*, 12, 1-13.

[50] Knowler WC, Barrett-Connor E, Fowler SE, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med*. 2002;346(6):393-403.

## Bibliographie

---

- [51] American Diabetes Association. (2010). Standards of medical care in diabetes—2010. *Diabetes care*, 33(Suppl 1), S11.
- [52] Garg, S. K., Weinzimer, S. A., Tamborlane, W. V., Buckingham, B. A., Bode, B. W., Bailey, T. S., ... & Kaufman, F. R. (2017). Glucose outcomes with the in-home use of a hybrid closed-loop insulin delivery system in adolescents and adults with type 1 diabetes. *Diabetes technology & therapeutics*, 19(3), 155-163.
- [53] Waljee, A. K., & Higgins, P. D. (2010). Machine learning in medicine: a primer for physicians. *Official journal of the American College of Gastroenterology| ACG*, 105(6), 1224-1226
- [54] Mifdal, R. (2019). Application des techniques d'apprentissage automatique pour la prédiction de la tendance des titres financiers (Doctoral dissertation, École de technologie supérieure).
- [55] Hinton, G., & Sejnowski, T. J. (Eds.). (1999). *Unsupervised learning: foundations of neural computation*. MIT press.
- [56] Fischer, A. (2014). Deux méthodes d'apprentissage non supervisé: synthèse sur la méthode des centres mobiles et présentation des courbes principales. *Journal de la Société Française de Statistique*, 155(2), 2-35.
- [57] Roche, D. Deep learning et apprentissage par renforcement pour la conception d'une Intelligence Artificielle pour le jeu Yokai No Mori. *Ce travail est publié sous licence Creative Common*.
- [58] Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022, March). Machine-learning-based disease diagnosis: A comprehensive review. In *Healthcare* (Vol. 10, No. 3, p. 541). MDPI.
- [59] Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01), 20-28.
- [60] beginners-guide-to-decision-trees-why-arethey-crucial-for-data. <https://analyticsindiamag.com> (consulté le 04 Avril 2023).
- [61] Casanova, R., Saldana, S., Chew, E. Y., Danis, R. P., Greven, C. M., & Ambrosius, W. T. (2014). Application of random forests methods to diabetic retinopathy classification analyses. *PLOS one*, 9(6), e98587.
- [62] Abdulkareem, N. M., & Abdulazeez, A. M. (2021). Machine learning classification based on Radom Forest Algorithm: A review. *International Journal of Science and Business*, 5(2), 128-142.

## Bibliographie

---

- [63] Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer genomics & proteomics*, 15(1), 41-51.
- [64] Classifying data using Support Vector Machines (SVMs) in R 26 Oct, 2021. <https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-r/?ref=rp>. (consulté le 04Avril 2023).
- [65] Learn How to Use Support Vector Machines (SVM) for Data. <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>. (consulté le 04Avril 2023)
- [66] Support Vector Machine Algorithm Works With Examples .<https://mindmajix.com/support-vector-machine-algorithm> . (consulté le 05Avril 2023).
- [67] arget. Collecte de données. <https://www.lemagit.fr/definition/Collecte-de-donnees>, (consulté le 05 Avril 2022).
- [68] Abidin, N. Z., Ismail, A. R., & Emran, N. A. (2018). Performance analysis of machine learning algorithms for missing value imputation. *International Journal of Advanced Computer Science and Applications*, 9(6).
- [69] Support Vector Machine Algorithm (Français) ,NOVEMBRE 11,2020 <https://ottima-power.com/fr/support-vector-machine-algorithm-5/> (consulté le 07Avril 2023)
- [70] Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*.
- [71] Ben Lazreg, M. (2020). *Recherche de l'information dans les réseaux de neurones convolutifs pré-entraînés* (Doctoral dissertation, École de technologie supérieure).
- [72] O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*. <https://doi.org/10.48550/arXiv.1511.08458> .
- [73] Les réseaux de neurones convolutifs Benlahmar October 12 2018. <https://www.datasciencetoday.net/index.php/en-us/deep-learning/173-les-reseaux-de-neurones-convolutifs>.( consulté le 08 Avril 2022).

## Bibliographie

---

- [74] Grégoire, J. M. Intelligence artificielle et ECC: revue critique Jean-Marie Grégoire, Cédric Gilon, Hugues Bersini, Stéphane Carlier.
- [75] Nayariseri, A., Khandelwal, R., Tanwar, P., Madhavi, M., Sharma, D., Thakur, G., ... & Singh, S. K. (2021). Artificial intelligence, big data and machine learning approaches in precision medicine & drug discovery. *Current drug targets*, 22(6), 631-655.
- [76] Saba Raouf, S., & Durai, M. A. (2022). A Comprehensive Review on Smart Health Care: Applications, Paradigms, and Challenges with Case Studies. *Contrast Media & Molecular Imaging*, 2022.
- [77] Theckedath, D., & Sedamkar, R. R. (2020). Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks. *SN Computer Science*, 1, 1-7.
- [78] Larabi-Marie-Sainte, S., Aburahmah, L., Almohaini, R., & Saba, T. (2019). Current techniques for diabetes prediction: review and case study. *Applied Sciences*, 9(21), 4604.
- [79] Ashiquzzaman, A.; Kawsar Tushar, A.; Rashedul Islam, M.D.; Shon, D.; Kichang, L.M.; Jeong-Ho, P.; Dong-Sun, L.; Jongmyon, K. Reduction of overfitting in diabetes prediction using deep learning neural network. In *IT Convergence and Security; Lecture Notes in Electrical Engineering*; Springer: Singapore, 2017; Volume 449
- [80] Swapna, G.; Soman, K.P.; Vinayakumar, R. Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. *Procedia Comput. Sci.* 2018, 132, 1253–1262
- [81] Mohebbi, A.; Aradóttir, T.B.; Johansen, A.R.; Bengtsson, H.; Fraccaro, M.; Mørup, M. A deep learning approach to adherence detection for type 2 diabetics. In Proceedings of the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Korea, 11–15 July 2017; pp. 2896–2899.
- [82] Pham, T.; Tran, T.; Phung, D.; Venkatesh, S. Predicting healthcare trajectories from medical records: A deep learning approach. *J. Biomed. Inform.* 2017, 69, 218–229.
- [83] Balaji, H.; Iyengar, N.; Caytiles, R.D. Optimal Predictive analytics of Pima Diabetics using Deep Learning. *Int. J. Database Theory Appl.* 2017, 10, 47–62
- [84] Deo, R.C. Machine Learning in Medicine. *Circulation* 2015, 132, 1920–1930.
- [85] Yuvaraj, N.; SriPreethaa, K.R. Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. *Clust. Comput.* 2017, 22, 1–9.

## Bibliographie

---

- [86] Tafa, Z.; Pervetica, N.; Karahoda, B. An intelligent system for diabetes prediction. In Proceedings of the 2015 4th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 14–18 June 2015; pp. 378–382.
- [87] Sisodia, D.; Sisodia, D.S. Prediction of Diabetes using Classification Algorithms. *Procedia Comput. Sci.* 2018, *132*, 1578–1585.
- [88] Somnath, R.; Suvojit, M.; Sanket, B.; Riyanka, K.; Priti, G.; Sayantan, M.; Subhas, B. Prediction of Diabetes Type-II Using a Two-Class Neural Network. In Proceedings of the 2017 International Conference on Computational Intelligence, Communications, and Business Analytics, Kolkata, India, 24–25 March 2017; pp. 65–71.
- [89] Saleem, T. J., & Chishti, M. A. (2021). Deep learning for the internet of things: Potential benefits and use-cases. *Digital Communications and Networks*, *7*(4), 526-542.
- [90] SABOUK, L. N., & SIDMOU, M. L. (2022). L'émergence du Cloud Computing au service de la transparence des collaborations inter-organisationnelles et de la confiance numérique: revue de littérature. *Revue Internationale des Sciences de Gestion*, *5*(2).
- [91] Colaboratory— Google. <https://research.google.com/colaboratory/faq.html#resource-limits> (consulté le 20 Mai 2023) .
- [92] AnacondaNavigator—Anacondadocumentation.<https://docs.anaconda.com/anaconda/navigator> (consulté le 20 Mai 2023)
- [93] “Project Jupyter.” <https://www.jupyter.org> (consulté le 20 Mai 2023)
- [94] TensorFlow. <https://www.tensorflow.org/> (consulté le 20 Mai 2023)
- [95] K. Team, “Keras documentation: About Keras.” <https://keras.io/about/> (consulté le 20 Mai 2023)
- [96] documentation officielle de Matplotlib : <https://matplotlib.org/stable/contents.html>(consulté le 21 Mai 2023)
- [97] Guide utilisateur de Scikit-learn : [https://scikit-learn.org/stable/user\\_guide.html](https://scikit-learn.org/stable/user_guide.html) (consulté le 21 Mai 2023)
- [98] <https://www.kaggle.com/johndasilva/diabetes> (consulté le 22 Mai 2023)

## **Bibliographie**

---

[99] Rabhi Karima Sidahmed Amel. La prédiction du diabète en utilisant les algorithmes de machine learning. (2) :85.

[100]Corrélation Introductionaux statistiques [https://www.jmp.com/fr\\_fr/statistics-knowledge-portal/what-is-correlation.html](https://www.jmp.com/fr_fr/statistics-knowledge-portal/what-is-correlation.html) (consulté le 22 Mai 2023)

[101] Gupta, P., Chouhan, A. V., Wajeed, M. A., Tiwari, S., Bist, A. S., & Puri, S. C. (2023). Prediction of health monitoring with deep learning using edge computing. *Measurement: Sensors*, 25, 100604.

[102]LSTM Mémoire à court terme - Long Short-Term Memory 19.03.2023 <https://konfuzio.com/fr/lstm/> (consulté le 15 Mai 2023)