الجـــــــــــــــــــــــمهـوريـــــــة الجـــــــــــــــزائـريـة الديمـــــــــقـراطيـة الـشـــــــــــعبيـة

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة التـــــــــــــــــــــــــــــعـليـم العـــــــــالي و البـــــــــــحث العـــــلـمـــي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جـــــــــــــــــامعة أبي بكـــر بـلـقـايد– تـلمســـــــــــــــــان –

Université Aboubakr Belkaïd– Tlemcen –

Faculté de TECHNOLOGIE

MEMOIRE

Présenté pour l'obtention du **diplôme** de **MASTER**

En : Télécommunications

Spécialité : Réseaux et Télécommunications

Par : Miraoui Amina&Moussaoui Djamila

Sujet

Classification du diabète à l'aide des méthodes du Deep Learning

Soutenu publiquement, le 08/06/2023, devant le jury composé de :

M.HADJILA Mourad	MCA	Université de Tlemcen	Président
M.BERKAT Abdellatif	MCA	Université de Tlemcen	Examinateur
M.MOUSSAOUI Djilali	MCA	Université de Tlemcen	Encadreur
M. FERHI WAFAA	DOCTORANTE	Université de Tlemcen	Co-encadrante

Année universitare :2022/2023

Remerciements

*À notre superviseur, **M. Moussaoui Djilali**,*

Nous tenons à exprimer notre profonde gratitude pour votre rôle de superviseur tout au long de notre parcours. Votre temps, vos conseils et vos orientations ont été d'une valeur inestimable pour nous. Votre soutien constant et votre expertise nous ont permis de progresser et de réaliser notre travail avec succès. Merci d'avoir partagé votre savoir et votre expérience avec nous.

*Nous aimerions également remercier chaleureusement **M. Hadjila Mourad** pour ses efforts remarquables. Sa contribution et sa collaboration ont été essentielles dans la réalisation de notre projet. Nous sommes reconnaissants d'avoir eu l'opportunité de travailler avec lui.*

*Par ailleurs, nous souhaitons exprimer notre sincère gratitude envers **Mme. Ferhi Wafaa**, notre superviseure de formation. Votre soutien constant, vos encouragements et votre disponibilité ont été d'une grande aide pour nous. Votre expertise et votre passion ont été une source de motivation tout au long de notre parcours. Nous vous remercions infiniment pour votre dévouement et votre investissement envers notre réussite.*

Enfin, nous aimerions exprimer notre sincère gratitude aux membres du jury pour leur attention en évaluant notre travail.

Dédicace

***À mon père bien-aimé**,*

Aujourd'hui, je termine cette étape importante de ma vie, et je me rends compte que je ne serais pas ici sans toi. Même si tu n'es plus parmi nous, tu as toujours été présent dans mon cœur et dans mon esprit.

***À ma chère mère**,*

Tu as été mon roc, ma force et ma lumière tout au long de ce voyage. Tu as été mon inspiration et ma source de soutien constant. Cette réussite est aussi la tienne, car tu as joué un rôle essentiel dans ma formation et mon épanouissement.

***À mes chères sœurs**,*

Je suis si reconnaissante d'avoir pu compter sur vous tout au long de mes études. Votre soutien, votre aide et vos encouragements m'ont donné la force de continuer lorsque les défis semblaient insurmontables.

***À ma famille bien-aimée**,*

Je tiens à vous exprimer toute ma gratitude et mon amour profond pour chacun d'entre vous. Vous avez été mes piliers de soutien tout au long de mon parcours d'études.

***À mes chers amis**,*

Les années d'études ont été remplies de moments inoubliables grâce à vous. Vous avez été mes compagnons de route, mes partenaires de projet et mes sources de joie et de divertissement. Votre amitié précieuse a rendu cette aventure encore plus belle et significative.

Amina

Dédicace

Je dédie ce mémoire

A la femme qui a souffert sans me laisser souffrir, et qui n'a jamais dit non

à mes exigences et qui n'a ménagé aucun effort pour me rendre heureuse

: Ma adorable mère est ***Houaria****.*

À ma sœur, mon amour et mon soutien, qui m'a accompagnée dans mes moments les plus difficiles: ***Yamina****.*

A toute ma famille et mes amis.

Djamila

ملخص

السكري هو مرض مزمن يتميز بارتفاع مستوى السكر في الدم، والذي يمكن أن يؤثر بشكل كبير على صحة ورفاهية الأفراد. في هذا العصر الرقمي، يلعب الذكاء الاصطناعي (AI) والتعلم العميق (Deep Learning) دورًا متزايد الأهمية في مجال تشخيص مرض السكري.

يوفر الذكاء الاصطناعي والتعلم العميق أدوات قوية لتحليل بيانات الصحة، مما يسمح بتحديد الأنماط والارتباطات والسمات الرئيسية المرتبطة بالسكري. من خلال استخدام خوارزميات متطورة وشبكات عصبية عميقة وطريقة التحقق المتقاطع، يمكن لهذه التقنيات استخلاص معلومات قيمة من كميات كبيرة من البيانات، مما يساهم في الكشف المبكر وتوقع المضاعفات وتحسين رعاية مرضى السكري.

في إطار هذه الدراسة، قدمنا نهجًا جديدًا قائمًا على الذكاء الاصطناعي لتوقع السكري. تم تقييم واختبار نموذجنا باستخدام مجموعة البيانات ."Diabetes Health Indicators Dataset" حققت النتائج المستخلصة إنجازات واعدة ومشجعة، حيث بلغت الدقة 99.99٪ في حالة البيانات الثنائية والمتوازنة. تفتح هذه النتائج آفاقًا تجارية لاستخدام نموذجنا في تقليل أخطاء تشخيص السكري.

الكلمات الرئيسية: السكري، الذكاء الاصطناعي، مجموعة البيانات، التعلم العميق، التحقق المتقاطع، الدقة.

Résumé

Le diabète est une maladie chronique caractérisée par un taux élevé de sucre dans le sang, qui peut avoir un impact significatif sur la santé et le bien-être des individus. Dans cette ère numérique, l'intelligence artificielle (IA) et l'apprentissage profond (deep learning) jouent un rôle de plus en plus important dans le domaine du diagnostic et d'administration du diabète.

L'intelligence artificielle et le deep learning offrent des outils puissants pour l'analyse des données de santé, permettant ainsi d'identifier des modèles, des corrélations et des caractéristiques clés liés au diabète. Grâce à l'utilisation d'algorithmes sophistiqués, de réseaux de neurones profonds et la méthode de validation croisée, ces techniques peuvent extraire des informations précieuses à partir de grandes quantités de données, ce qui contribue à la détection précoce, à la prédiction des complications et à l'amélioration de la prise en charge des patients diabétiques.

Dans le cadre de cette étude, nous avons proposé une nouvelle approche basée sur l'IA pour prédire le diabète. Notre modèle a été évalué et testé en utilisant l'ensemble de données "DiabetesHealthIndicatorsDataset". Les résultats obtenus sont prometteurs et encourageants, avec une accuracy de 99.99% et une fonction loss de 3.95% pour le cas binaire et les données équilibrées. Ces résultats ouvrent des perspectives commerciales pour l'utilisation de notre modèle afin de réduire les pertes de diagnostic du diabète.

Mots clés: diabète, intelligence artificielle,dataset,deeplearning, validation croisée, accuracy, fonction loss.

Abstract

Diabetes is a chronic disease characterized by high blood sugar levels, which can have a significant impact on the health and well-being of individuals. In this digital age, artificial intelligence (AI) and deep learning are playing an increasingly important role in the diagnosis and administration of diabetes.

Artificial intelligence and deep learning offer powerful tools for analyzing health data, enabling the identification of key patterns, correlations and characteristics related to diabetes. Through the use of sophisticated algorithms, deep neural networks, and the cross-validation method, these techniques can extract valuable information from large amounts of data, contributing to early detection, prediction of complications, and improved management of diabetes patients.

In this study, we proposed a novel AI-based approach to predict diabetes. Our model was evaluated and tested using the Diabetes Health Indicators Dataset. The results obtained are promising and encouraging, with an accuracy of 99.99% and a loss function of 3.95% for the binary case and the balanced data. These results open commercial perspectives for the use of our model to reduce loss in diabetes diagnosis.

Keywords: diabetes, artificial intelligence,dataset, deep learning, cross validation, accuracy, loss function.

Table des matières

Liste des tableaux

Table des figures

Liste des abréviations

DID	Diabète Insulino-Dépendant en Anglais Diabète de Type 1
DL	Deep Learning / Apprentissage en Profondeur
DNID	Diabète Non Insulino-Dépendant
DNN	Deep Neural Network en Anglais ou Réseau de Neurones Profonds
DQN	Deep Q-Network / Réseau de Neurones Q-Profond
GNMT	Google Neural Machine Translation system / Système de traduction neuronale de Google
GPS	General Problem Solver en Anglais ou Résolveur de problèmes général
IA	Intelligence Artificielle
IBM	International Business Machines Corporation
IDF	International Diabetes Federation
IMC	Indice de masse corporelle
K-NN	k-Nearest Neighbors / k-Plus Proches Voisins
LR	Logistic Regression / Régression Logistique
MIT	Massachusetts Institute of Technology en Anglais ou Institut de technologie du Massachusetts
ML	Machine Learning en Anglais ou Apprentissage automatique
NaN	Not a Number
NASA	National Aeronautics and Space Administration / Administration nationale de l'aéronautique et de l'espace
NN	Neural Network en Anglais ou Réseau de Neurones
OMS	Organisation Mondiale de la Santé
PMC	Perceptron Multicouche
RBF	Radial Basis Function Networks / Réseaux à fonctions de base radiale
RF	Random Forest / Forêt Aléatoire
SARSA	State-Action-Reward-State-Action / État-Action-Récompense-État-Action
SLP	Single-Layer Perceptron / perceptron à une seule couche
SVM	Support Vector Machine Machine à Vecteurs de Support
ART	
TD	Temporal Difference learning / Apprentissage par Différence Temporelle

Introduction Générale

Les maladies constituent un défi majeur pour la santé publique, affectant des millions de personnes à travers le monde. Parmi elles, le diabète est une maladie chronique qui se caractérise par un taux élevé de sucre dans le sang. Cependant, grâce aux progrès de l'intelligence artificielle, de nouvelles méthodes de détection et de diagnostic sont en train de révolutionner la manière dont nous abordons cette maladie.

Le diabète est une maladie complexe qui nécessite une gestion précise pour prévenir d'éventuelles complications. Afin d'éviter ces dernières, il est nécessaire d'utiliser des techniques de détection intelligentes telles que le deep learning, pour ouvrir de nouvelles possibilités de détection précoce et précise de cette maladie.

Le deep learning, branche de l'intelligence artificielle, utilise des réseaux de neurones artificiels pour analyser des données complexes. Dans le cas du diabète, les algorithmes de deep learning sont entraînés sur de grandes quantités de données médicales afin de détecter les schémas et caractéristiques liés à cette maladie. Ils permettent ainsi de classer les patients en fonction de leur probabilité de développer ou d'avoir le diabète.

Dans le premier chapitre intitulé "Intelligence Artificielle", On a abordé plusieurs aspects essentiels de cette discipline. Nous avons commencé par une introduction générale, suivie d'une définition claire de l'intelligence artificielle. Ensuite, nous avons retracé son évolution historique, mettant en lumière ses avancées majeures. Les applications de l'intelligence artificielle dans la vie quotidienne ont été discutées, ainsi que les avantages, inconvénients et limites associés. Nous avons aussi exploré le machine learning et le deep learning, deux approches fondamentales de l'intelligence artificielle. En conclusion, nous avons résumé les idées principales abordées, soulignant l'importance continue de l'intelligence artificielle et les défis à relever.

Dans le deuxième chapitre, consacré au diabète, nous avons exploré cette maladie complexe sous différents angles. Nous avons d'abord présenté les deux types de diabète, en mettant en évidence leurs caractéristiques distinctes. Ensuite, nous avons examiné les causes et les facteurs de risque associés au diabète, en mettant en relief sa prévalence mondiale croissante. Nous avons ensuite abordé le diagnostic et le traitement du diabète, en mettant en avant les approches médicales et les méthodes de gestion du mode de vie. Nous avons aussi abordé l'impact psychologique et social du diabète, ainsi que les avancées dans le

développement des traitements de cette maladie. En conclusion de ce chapitre, nous avons récapitulé les principaux éléments présentés.

Le troisième chapitre de notre travail se concentre sur la méthodologie que nous avons employée. Nous avons détaillé la sélection du dataset utilisé pour l'étude et précisé les critères de choix que nous avons pris en compte. Ensuite, nous avons décrit les étapes de prétraitement des données que nous avons suivies, notamment la normalisation et la suppression des données manquantes. Notre objectif était de garantir la qualité et la cohérence des données avant de les utiliser pour entraîner notre modèle. Par la suite, nous avons expliqué en détail la création de notre modèle basé sur le deep learning. Nous avons détaillé l'architecture du réseau neuronal que nous avons utilisée, en décrivant les différentes couches et les fonctionnalités clés du modèle. De plus, nous avons fourni des explications sur les hyperparamètres choisis et les techniques d'optimisation que nous avons appliquées pour améliorer les performances du modèle. Ce chapitre méthodologique joue un rôle crucial dans la démonstration de notre travail. Il fournit les informations nécessaires pour évaluer la validité de notre approche et interpréter les résultats de manière éclairée.

Chapitre 1 :

L’intelligence artificielle

1.1 Introduction

Lorsque nous parlons d'intelligence artificielle ou d'apprentissage automatique, nous parlons d'algorithmes qui interprètent les données pour reconnaître des modèles et développer des règles. En d'autres termes, l'algorithme apprend à partir d'exemples plutôt qu'à partir de règles définies par un agent extérieur. Dans les deux cas, la machine reste connectée aux instructions écrites dans l'algorithme. Par conséquent, les ne sont pas dotés d'intelligence au sens humain. Il s'agit plutôt d'un processus cognitif hautement complexe et ciblé qui est toujours optimisé. Mais comme les humains, les ordinateurs intégrés par la pratique et la répétition. Plus un algorithme est publié d'exemples, plus son modèle est précis[1].

L'intelligence artificielle (IA) est la simulation des processus d'intelligence humaine par des machines. Plus précisément, les systèmes informatiques. Ces processus comportent trois phases :

- D'abord, l'apprentissage, l'acquisition de l'information et les règles de son utilisation.
- Puis, le raisonnement, soit l'utilisation de règles pour tirer des conclusions définitives.
- Enfin, l'autocorrection.

L'IA peut être classée comme faible ou forte. L'IA faible ou Narrow AI est un système d'intelligence artificielle conçu et formé pour une tâche spécifique. En tant que tels, les assistants personnels virtuels comme Siri d'Apple sont des formes faibles d'IA. L'IA forte ou l'intelligence générale artificielle a des capacités cognitives humaines. Étant donné une tâche inconnue, un bon système d'IA peut trouver une solution sans intervention humaine.

Le coût du matériel, des logiciels et des ressources humaines de l'intelligence artificielle peut être élevé. De nombreux fournisseurs ont accès aux plates-formes AIaaS (intelligence artificielle en tant que service), y compris des composants d'IA dans leurs offres standard. AIaaS permet aux particuliers et aux entreprises d'expérimenter l'intelligence artificielle et de tester plusieurs plates-formes avant de s'engager. Certains des services de cloud computing d'intelligence artificielle les plus populaires incluent Amazon AI Services, IBM Watson Assistant, Microsoft Cognitive Services et Google AI Services [2].

1.2 Définition de l'intelligence artificielle

Le terme « intelligence artificielle » ou IA est couramment utilisé pour désigner les ordinateurs et les programmes informatiques dont les performances sont liées à l'intelligence humaine. Par exemple, la capacité d'interagir avec les gens, de traiter de grandes quantités de données ou de s'améliorer continuellement en apprenant progressivement. C'est donc un vaste

sujet en éternelle évolution ! Selon Larousse, l'intelligence artificielle peut être définie comme "un ensemble de théories et de méthodes appliquées pour créer des machines capables de mimer l'intelligence". Ordinateur ou programme doté d'une puissance de traitement généralement associée à l'intelligence humaine et susceptible d'être améliorée par la technologie:

- Capacité de raisonner.
- Capacité de traiter de grandes quantités de données.
- Faculté de discerner des patterns et des modèles indétectables par un humain.
- Aptitude à comprendre et analyser ces modèles.
- Capacités à interagir avec l'homme.
- Possibilité d'apprendre progressivement.
- Améliorer continuellement ses performances.

« L'intelligence artificielle » couvre ainsi un sujet vaste et en constante évolution. Et depuis 1950, année de création de l'IA, elle a fait des progrès incroyables [3].

1.3 Histoire de l'intelligence artificielle

Il est intéressant de remonter aux origines et à l'histoire de l'intelligence artificielle pour bien comprendre ses repères originels et ses perspectives [4].

1950 Alan M. Turing, mathématicien et théoricien précurseur de l'informatique, lance le concept d'intelligence artificielle.

1955-1956 Lancement du premier programme d"intelligence artificielle par Allen Newell, John C. Shaw et Herbert A. Simon, le Logic Theorist.

1957 Modélisation des jeux d'échec.

1958 John McCarthy invente le Perceptron, langage de programmation interactif (développement au MIT).

1958 Construction du premier réseau neuronal, le Perceptron, de Frank Rosenbaltt, machine dite connectionniste.

1959 Elaboration du premier GPS (genaral problem solver)-fin de la première période de l'intelligence artificielle.

1970 Néoconnectionnisme.

1989 Deep Thought, supercalculateur d'IBM, deux millions de coups par seconde.

1990-1997 Développement de Deep Blue rebaptisé Deeper Blue: conception d'un système de 256 processeurs fonctionnant en paralléle, chaque processeur peut calculer environ trois millions de coups par seconde.

2009 Le MIT a lancé un projet visant à repenser la rcherche en intelligence artficielle.

2011 Watson, le superordinateur d'IBM remporte deux des trois manches du jeu télévisé Jeopardy! La performance a consisté pour cette intelligence artificielle à réoindre à des questions de culture générale.

2013 Humain Brain Project.

Google ouvre un laboratoire de rcherches dans les locaux de la NASA.

2014 Deep Knowledge Ventures nomme à son conseil d'administration VITAL, un algorithme capable d'élaborer ces décision en analysant les bilans comptables des entreprises potentiellement intéressantes, les test cliniques, la propriété intellectuelle et les précédents investissements.

Eugene Goostman, programme informatique conçu en Russie, est parvenu, lors d'une compétition organisée par l'Université britaniue de Reading, tromper plusieurs personnes dans le cadre d'un test de Turing.

2015 Facebook Artifical Intelligence Research (FAIR).

Google rend sa technologie d'intelligence artificielle TensorFlow accessible à tous.

Développement d'une crainte que l'intelligence artificielle dépasse a terme les performances de l'intelligence humaine.

2016 Amelia d'IPSoft un agent virtuel.

AlphaGo bat trois fois consécutives le chanmpion du mande de jeu de go, Lee Se-Dol en cinq manches.

1.4 Applications de l'intelligence artificielle déployées dans la vie quotidienne

Dans la culture populaire, l'intelligence artificielle a longtemps été sens de robots autonomes qui soit détruisent les gens dans une version catastrophe, soit les sauvent d'une tâche ingrate dans une version humaniste [5].

1.4.1 Les e-mails

La messagerie s'appuie fortement sur l'intelligence artificielle pour rationaliser les performances et améliorer l'expérience utilisateur. La fonction Smart Reply fournit des messages courts pour répondre aux e-mails en un seul clic. La société a également développé une fonctionnalité appelée Smart Compose qui peut compléter les phrases des utilisateurs.

1.4.2 Les réseaux sociaux

L'intelligence artificielle a un impact énorme sur la façon dont l'information est présentée, en particulier sur les réseaux sociaux. Par exemple, Facebook trie puis filtre les publications de vos contacts et des pages que vous suivez, en mettant en évidence les publications qu'il juge les plus importantes et en masquant complètement le reste. Le site utilise une IA appelée DeepText pour analyser le contenu de vos publications. Il peut être utilisé non seulement pour organiser des flux d'actualités, mais aussi pour intervenir si un employé détecte des signes de suicide.

1.4.3 La traduction

Les services de traduction automatique, disponibles via des services tels que Google Translate ou directement intégrés à des sites tels que Facebook, ont parcouru un long chemin ces dernières années pour produire un texte parfaitement compréhensible. Ces avancées ont été rendues possibles en partie par des méthodes d'apprentissage en profondeur telles que les systèmes d'intelligence artificielle de Google appelés Neural Machine Translation (GNMT).

1.4.4 Les applications de navigation

L'intelligence artificielle a changé nos habitudes de conduite grâce aux applications de navigation comme Waze ou Google Maps. Ils déterminent l'itinéraire le plus court et estiment l'heure d'arrivée. Il peut même modifier automatiquement les trajectoires de circulation pour éviter les embouteillages, en tenant compte du trafic en temps réel.

1.4.5 Médecine, santé

L'intelligence artificielle enrichit les fonctions médicales, permettant aux utilisateurs de surveiller leur propre santé en temps réel, et notamment à l'aide de montres connectées, elle peut détecter des pathologies ou des anomalies liées à la santé de l'utilisateur. , chutes, mauvaise saturation en oxygène pendant l'entraînement, rythme cardiaque trop rapide ou trop lent, etc.

1.5 Avantages, inconvénients et Limites de l'intelligence artificielle

1.5.1 Les avantages

1.5.1.1 La réduction des erreurs

Il est appliqué dans divers domaines tels que l'exploration spatiale. Des robots intelligents reçoivent des informations et explorent l'univers. Parce qu'il s'agit d'une machine à coque métallique, elle est plus solide et a une plus grande capacité à résister aux environnements et atmosphères difficiles [6].

1.5.1.2 L'exploration difficile

L'intelligence artificielle et la robotique pourraient être utilisées dans l'exploitation minière et d'autres processus d'exploration de carburant. De plus, ces machines sophistiquées peuvent être utilisées pour explorer les fonds marins, dépassant ainsi les limites humaines. En programmant des robots, ils peuvent effectuer des tâches complexes qui prennent plus de temps et prennent plus de responsabilités. Il ne s'use pas facilement non plus [6].

1.5.1.3 L'application quotidienne

Les smartphones sont un exemple approprié et quotidien de l'utilisation de l'intelligence artificielle. Lorsque nous prenons une photo, un algorithme d'intelligence artificielle identifie et détecte le visage de la personne et l'affiche lorsque nous publions la photo sur les sites de médias sociaux. L'intelligence artificielle est largement utilisée par les institutions financières et les institutions financières pour organiser et gérer les données. L'intelligence artificielle est également utilisée pour détecter les fraudes [6].

1.5.1.4 Les travaux répétitifs

Des tâches répétitives de nature monotone peuvent être effectuées à l'aide de l'intelligence artificielle. Les machines peuvent penser plus vite que les humains et effectuer plusieurs tâches simultanément. L'intelligence artificielle peut être utilisée pour effectuer des tâches dangereuses. Contrairement aux humains, leurs paramètres peuvent être ajustés. Leur vitesse et leur temps ne sont que des paramètres basés sur des calculs [6].

1.5.1.5 Les applications médicales

Les médecins utilisent des machines à intelligence artificielle pour évaluer les risques pour les patients et la santé. Les professionnels de la santé sont souvent formés sur des simulateurs chirurgicaux artificiels. La robotique est souvent utilisée pour aider les personnes atteintes de maladie mentale à sortir de la dépression et à rester actives [6].

1.5.2 Les inconvénients

1.5.2.1 Un coût élevé

Créer de l'intelligence artificielle nécessite des coûts énormes car c'est une machine très complexe. La réparation et l'entretien sont également assez coûteux. Ils ont des logiciels qui doivent être mis à jour fréquemment pour répondre aux besoins de l'environnement changeant et à la nécessité pour les machines de devenir plus intelligentes chaque jour. En cas de panne catastrophique, le processus de récupération du code perdu et de réinstallation du système peut être long et coûteux [6].

1.5.2.2 Aucune initiative

Les machines n'ont pas d'émotions et pas de valeurs morales. Ils agissent comme ils sont programmés et ne peuvent pas décider ce qui est bien ou mal. Face à une situation inconnue, ils ne peuvent même pas prendre de décision. Dans ces circonstances, il ne fonctionnera pas correctement ou sera cassé [6].

1.5.2.3 Aucune amélioration avec l'expérience

Contrairement aux humains, l'IA ne peut pas être améliorée avec l'expérience. Il stocke beaucoup de données, mais la façon dont il y accède et les utilise est très différente de l'intelligence humaine. Dans le monde de l'intelligence artificielle, il n'y a rien de mieux que de travailler avec son cœur ou sa passion. S'inquiéter ou s'inquiéter ne fait pas partie du vocabulaire de l'intelligence artificielle [6].

1.5.2.4 Le chômage

Remplacer les gens par des machines pourrait entraîner un chômage important. Une personne qui n'a rien à faire peut utiliser la pensée créative de manière destructrice. L'utilisation généralisée de l'intelligence artificielle pourrait rendre les humains inutilement dépendants des machines. L'intelligence artificielle entre de mauvaises mains constitue une grave menace pour l'humanité dans son ensemble. Il y a aussi une peur constante que les machines remplacent ou remplacent les humains. Identifier et rechercher les risques de l'intelligence artificielle est une tâche très importante. Tout ce qui est créé dans ce monde et dans les sociétés individuelles est un produit durable de l'esprit. L'intelligence artificielle complète et améliore l'intelligence humaine [6].

1.5.3 Les limites

La gamme de l'IA et de ses applications sont illimitée et croît de façon exponentielle chaque jour. Cependant, les avancées technologiques actuelles peuvent poser certaines limites [7]:

- La reproduction émotionnelle est encore l'apanage des humains.
- Complexité du codage multitâche : les logiciels alimentés par l'IA excellent dans une seule tâche.
- Configuration logicielle initiale : Si l'algorithme est capable d'auto-apprentissage, l'homme doit impérativement contrôler la machine, par exemple en définissant des objectifs.

1.6 Machine Learning

1.6.1 Définition

L'apprentissage automatique, l'apprentissage artificiel ou l'apprentissage statistique est un domaine de recherche sur l'intelligence artificielle qui utilise des approches mathématiques et statistiques pour permettre aux ordinateurs «d'apprendre» à partir de données, c'est-à-dire d'améliorer les performances de résolution de problèmes sans programmation explicite pour chacun. . . Il s'agit de la conception, de l'analyse, de l'optimisation, du développement et de la mise en œuvre de ces méthodes au sens large [8].

1.6.2 Les différents types d'apprentissage automatique

1.6.2.1 L'apprentissage supervisé

La machine learning supervisé est une technologie élémentaire mais stricte. Les opérateurs présentent à l'ordinateur des exemples d'entrées et les sorties souhaitées, et l'ordinateur recherche des solutions pour obtenir ces sorties en fonction de ces entrées. Le but est que l'ordinateur apprenne la règle générale qui mappe les entrées et les sorties. Le machine learning supervisé peut être utilisé pour faire des prédictions sur des données indisponibles ou futures (on parle alors de "modélisation prédictive"). L'algorithme essaie de développer une fonction qui prédit avec précision la sortie à partir des variables d'entrée – par exemple, prédire la valeur d'un bien immobilier (sortie) à partir d'entrées telles que le nombre de pièces, l'année de construction, la surface du terrain, l'emplacement, etc. L'apprentissage automatique supervisé peut être divisé en deux types :

Classification : la variable de sortie est une catégorie.

Régression : la variable de sortie est une valeur spécifique.

Les principaux algorithmes d'apprentissage automatique supervisé sont les forêts aléatoires, les arbres de décision, les k-Nearest Neighbors (K-NN), la régression linéaire, Naive Bayes, la machine à vecteurs de support (SVM), la régression logistique et le Gradient Boosting [9].

1.6.2.2 L'apprentissage non-supervisé

Dans l'apprentissage automatique non supervisé, l'algorithme lui-même détermine la structure des données d'entrée (les étiquettes ne sont pas appliquées à l'algorithme). Cette approche peut être soit une fin en soi (vous permettant de découvrir des structures cachées dans les données), soit un moyen d'atteindre une fin. Cette approche est aussi appelée « feature learning ». Un exemple d'apprentissage automatique non supervisé est l'algorithme de reconnaissance faciale prédictif de Facebook qui identifie les personnes à partir de photos publiées par les utilisateurs. Il existe deux types d'apprentissage automatique 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 d'apprentissage automatique non supervisé sont les K-means, le clustering/hierarchical clustering et la réduction de la dimensionnalité [9].

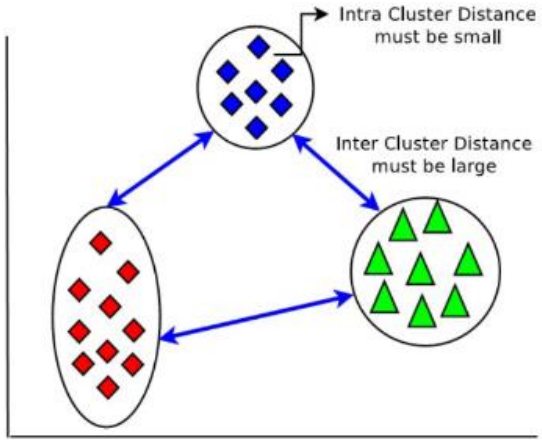

Figure 1.2- Apprentissage non supervisé **[9].**

1.6.2.3 L'apprentissage par renforcement

Dans l'apprentissage automatique par renforcement, un programme informatique interagit avec un environnement dynamique qui doit atteindre un objectif spécifique, comme conduire un véhicule ou affronter des adversaires dans un jeu. Les apprenants reçoivent des commentaires sous forme de « récompenses » et de « punitions » lorsqu'ils apprennent à naviguer dans l'espace du problème et à identifier les actions les plus efficaces dans un contexte donné. Il s'agissait d'un algorithme d'apprentissage par renforcement (Q-learning) qui était déjà devenu célèbre en 2013 pour apprendre à battre six jeux vidéo Atari sans l'intervention d'un programmeur. Il existe deux types d'apprentissage par renforcement:

Monte Carlo : le programme reçoit ses récompenses à la fin de l'état « terminal ».

Machine learning par différence temporelle (TD) : les récompenses sont évaluées et accordées à chaque étape.

Les principaux algorithmes d'apprentissage par renforcement sont Q-learning, Deep Q Network (DQN) et State-Action-Reward-State-Action (SARSA)... [9]

1.6.3 Les principaux algorithmes de Machine Learning

1.6.3.1 La régression linéaire

Cet algorithme est utilisé pour prédire une variable continue basée sur une autre variable. Il est souvent utilisé pour des problèmes de prévision tels que la prévision des ventes ou la prévision des prix de l'immobilier [10].

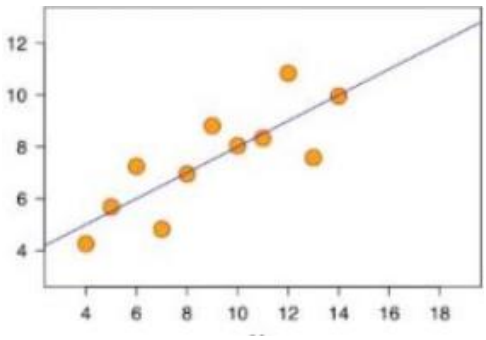

Figure 1.3- La régression linéaire

1.6.3.2 La régression logistique

Cet algorithme est utilisé pour prédire une variable binaire (0 ou 1) basée sur une autre variable. Il est souvent utilisé pour la classification binaire, comme la prédiction de la probabilité de défaut sur un prêt [10].

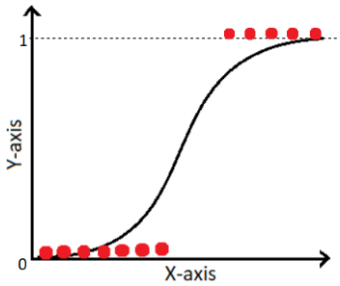

Figure 1.4- La régression logistique

1.6.3.3 Support Vector Machine (SVM)

Cet algorithme est utilisé pour classer les données. Il est souvent utilisé pour les problèmes de classification binaire et multi-classes [10].

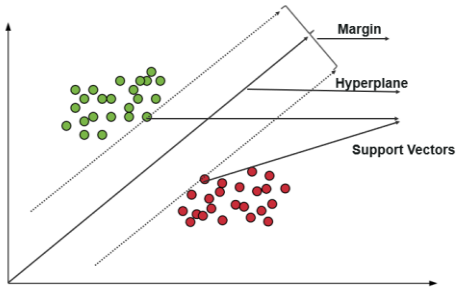

Figure 1.5- Support Vector Machine

1.6.3.4 L'arbre de décision

Cet algorithme est utilisé pour créer un modèle d'arbre de décision à partir de l'entrée. Il est souvent utilisé pour la classification et la prédiction [10].

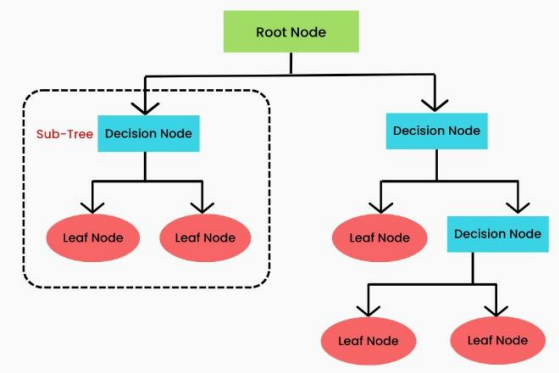

Figure 1.6- Les algorithmes d'arbre de décision

1.6.3.5 K-Means

Cet algorithme est utilisé pour partitionner les données en groupes. Il est souvent utilisé pour l'analyse de cluster et la segmentation du marché [10].

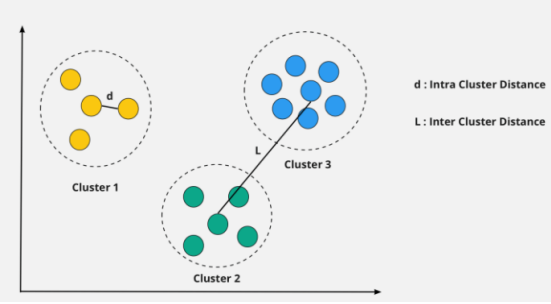

Figure 1.7- Les algorithmes k-moyennes

1.6.3.6 Le Gradient Boosting

Les algorithmes d'amplification de gradient produisent des modèles prédictifs qui combinent des modèles prédictifs faibles (généralement des arbres de décision) via un processus de construction qui améliore les performances globales du modèle [10].

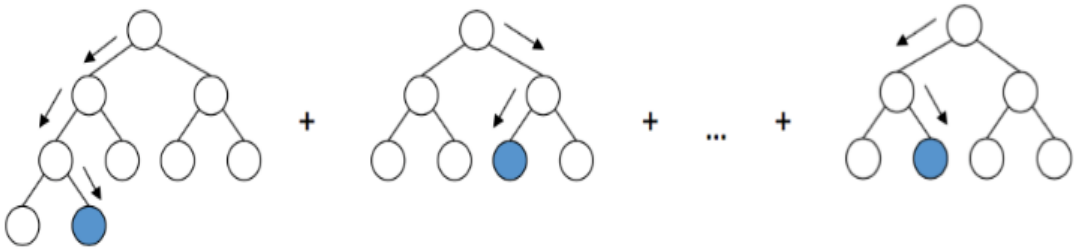

Figure 1.8- Le Gradient Boosting

1.7 Deep learning (DL)

1.7.1 Définition

Le deep learning ou apprentissage en profondeur est une technique d'intelligence artificielle qui utilise des réseaux de neurones artificiels pour apprendre à identifier des objets et à effectuer des tâches complexes [11] .

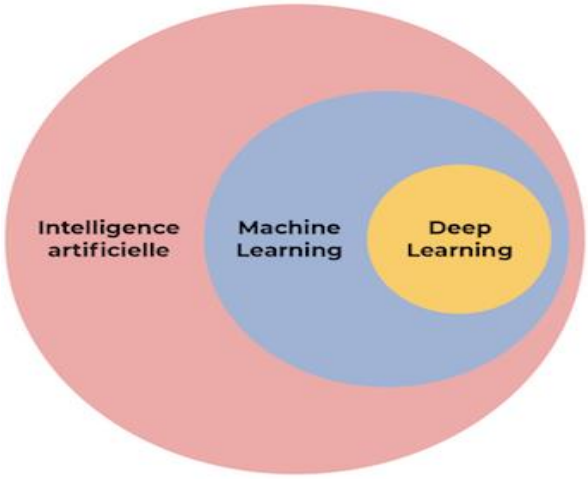

Figure 1.9- Définition de Deep Learning

1.7.2 La différence entre le machine learning et le deep learning

Tableau 1.1- La différence entre ML et DL [12]

	Tout l'apprentissage automatique	**Apprentissage en profondeur uniquement**
Nombre de points de données	Il peut être prédit en utilisant une petite quantité de données.	Nous devons utiliser une grande quantité de données d'entraînement pour la prédiction.
Dépendances matérielles	Vous pouvez travailler sur des machines faibles. Il ne nécessite pas beaucoup de puissance de calcul.	Dépend du véhicule parent. Essentiellement effectuer un grand nombre d'opérations de multiplication matricielle. Les GPU peuvent optimiser efficacement ces opérations.
Processus de caractérisation	Les utilisateurs doivent clairement identifier et créer des fonctionnalités.	Il apprend des fonctionnalités de haut niveau à partir des données et génère lui-même de nouvelles fonctionnalités.
Approche d'apprentissage	Divisez le processus d'apprentissage en étapes plus petites. Les résultats de chaque étape sont ensuite combinés en un seul résultat.	Vous passez par un processus d'apprentissage lorsque vous résolvez des problèmes du début à la fin.
Temps d'exécution	La formation prend relativement peu de temps, de quelques secondes à plusieurs heures.	Les algorithmes d'apprentissage en profondeur prennent généralement beaucoup de temps à s'entraîner car ils contiennent de nombreuses couches.
Sortir	Le résultat est généralement une valeur numérique telle qu'un score ou une classification.	La sortie peut être dans plusieurs formats tels que texte, partition ou son.

1.7.3 Réseaux de neurones artificiels

1.7.3.1 Définition

Un réseau de neurones artificiels, ou DNN en anglais, est un système informatique matériel et/ou logiciel dont le fonctionnement modélise le fonctionnement des neurones du cerveau humain [13].

1.7.3.2 Architecture

Les réseaux de neurones peuvent prendre de nombreuses formes en fonction de l'objet de données et de la complexité qu'ils traitent et de la manière dont les données sont traitées. Les architectures ont leurs propres forces et faiblesses, et nous pouvons les combiner pour optimiser nos résultats. Par conséquent, le choix de l'architecture est important et est largement déterminé par vos objectifs [14].

a. Les réseaux récurrents« FEED-BACK »

Les réseaux de ce type se caractérisent par la possibilité d'une diffusion récursive partielle ou complète de l'information. Les architectures les plus utilisées sont:

❖ **Les cartes de Kohonen auto-organisées**

Ce type de réseau utilise un apprentissage non supervisé pour ajuster des cartes discrètes et ordonnées basées sur le modèle d'entrée.

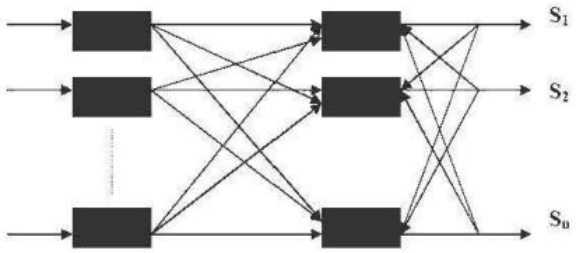

Figure 1.10- Les cartes de Kohonen auto-organisées

❖ **Les réseaux de Hopfield**

Les réseaux de Hopfield sont considérés comme des réseaux entièrement connectés et ne font pas la différence entre les neurones d'entrée et de sortie. Ce type de réseau agit comme une mémoire associative non linéaire capable de reconnaître des objets stockés dans l'espace de données.

❖ **Le perceptron monocouche « SLP »**

Avant de définir la structure collective d'un ensemble de neurones, il est important de définir un perceptron monocouche. Il s'agit d'un réseau très simple qui est généralement supervisé en raison de la composition des couches d'entrée et de sortie sans couches cachées. Suivez les règles de correction des erreurs ou les règles Hebb.

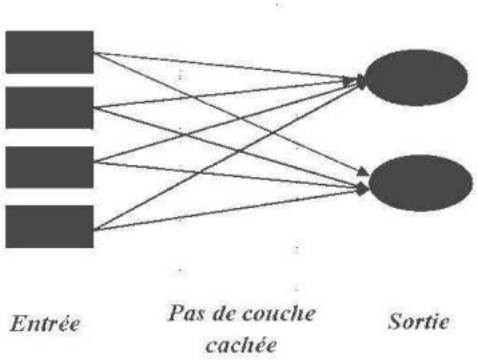

Figure 1.11- Le perceptron monocouche

❖ **Le perceptron multicouches« PMC »**

Les neurones de la couche d'entrée ont tendance à se connecter uniquement à la couche suivante, tandis que les neurones de la couche cachée ont tendance à se connecter à tous les neurones des couches précédente et suivante, et les connexions entre les neurones ne se chevauchent pas. Le choix du nombre de couches cachées dépend généralement de la complexité du problème à résoudre. En théorie, une couche cachée pourrait suffire à résoudre un problème particulier, mais plusieurs couches cachées pourraient potentiellement résoudre à la fois des problèmes plus simples et plus complexes.

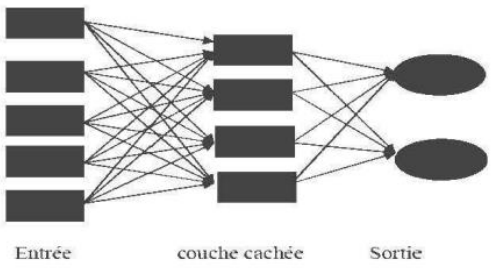

Figure 1.12- Le perceptron multicouches

❖ **Réseaux à fonction radiale« RBF »**

Les réseaux de fonctions radiales RBF sont très similaires aux réseaux PMC, mais se caractérisent par l'utilisation d'une fonction gaussienne comme fonction de base, de sorte que RBF est utilisé pour la même classe de problèmes que PMC : classification et prédiction. Parmi les types d'entraînement utilisés dans RBF figure un mode hybride avec des règles de correction d'erreurs.

1.7.3.3 Principe de fonctionnement

En règle générale, un réseau de neurones repose sur un grand nombre de processeurs opérant en parallèle et organisés en tiers. Le premier tiers reçoit les entrées d'informations brutes, un peu comme les nerfs optiques de l'être humain lorsqu'il traite des signaux visuels. Par la suite, chaque tiers reçoit les sorties d'informations du tiers précédent. On retrouve le même processus chez l'Homme, lorsque les neurones reçoivent des signaux en provenance des neurones proches du nerf optique. Le dernier tiers, quant à lui, produit les résultats du système [13].

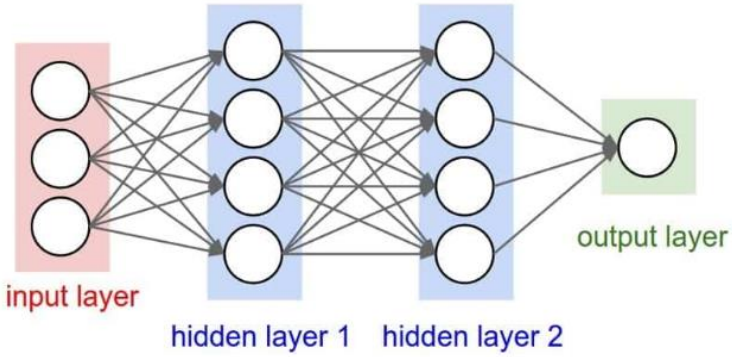

Figure 1.13- Principe de fonctionnement de réseaux neurones

1.7.3.4 Les types de réseaux de neurones artificiels

a. Les réseaux de neurones feed-forwarded

Dans une architecture feed-forward, les informations ne se propagent que vers l'avant et jamais vers l'arrière. Cette famille de réseaux de neurones comprend deux catégories de réseaux de neurones: les perceptrons simples et les perceptrons multicouches. Un perceptron simple est un réseau neuronal avec seulement deux couches de neurones : une couche d'entrée et une couche de sortie. Les deux couches sont directement connectées l'une à l'autre de sorte que le réseau n'a qu'une seule matrice de poids. Ce type d'algorithme n'est utile que pour classer linéairement un ensemble d'informations en deux catégories. Dans les perceptrons multicouches, il y a une ou plusieurs couches cachées entre les couches d'entrée et de sortie. C'est le meilleur algorithme pour gérer les fonctions non linéaires en combinant plusieurs matrices de poids. Pour traiter des données très complexes, vous pouvez créer des réseaux de neurones distincts, chacun traitant une information. Dans ce cas, nous parlons de réseaux de neurones convolutifs ou "réseaux de neurones convolutifs" [15].

b. Les réseaux de neurones à résonance

Dans un réseau résonnant, l'activation d'un neurone se reflète dans tous les autres neurones, les faisant osciller. Cette architecture peut prendre plusieurs formes avec un haut degré de complexité [15].

c. Les réseaux de neurones récurrents

Les réseaux de neurones récurrents effectuent un traitement récursif des informations. Les données peuvent être transmises dans les sens aller et retour. Cette architecture permet à l'algorithme de prendre en compte les informations contextuelles chaque fois qu'il traite les mêmes informations. Le réseau est donc autonome. Il existe des réseaux de neurones récurrents monocouches et multicouches. L'un des modèles à une couche les plus connus est le modèle de Hopfield [15].

d. Les réseaux de neurones autoorganisés

Les réseaux de neurones auto-organisés sont utilisés pour traiter les informations spatiales. Ils peuvent étudier un large éventail de distributions de données pour fournir des solutions aux problèmes de classification. Les cartes de Kohonen auto-organisées sont le modèle le plus connu de réseaux de neurones artificiels auto-organisés [15].

1.7.4 Cross validation

1.7.4.1 Définition

La validation croisée (CV) est une technique utilisée pour évaluer un modèle d'apprentissage et tester ses performances (ou sa précision). Cela implique de réserver un échantillon spécifique d'un jeu de données sur lequel le modèle n'est pas formé. Plus tard, le modèle est testé sur cet échantillon pour l'évaluer [16].

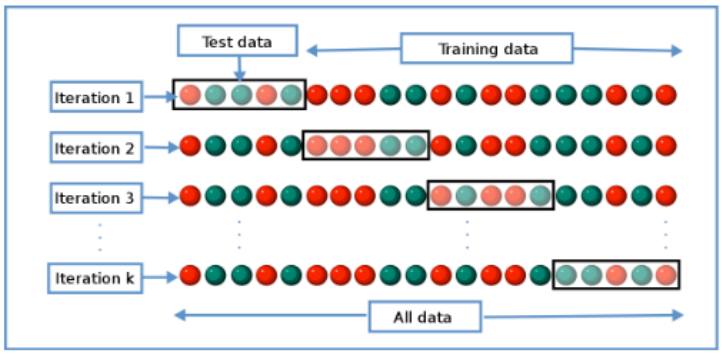

Figure 1.14- Cross-validation

1.7.4.2 Les différents types de cross-validation

a. Validation croisée k-fold

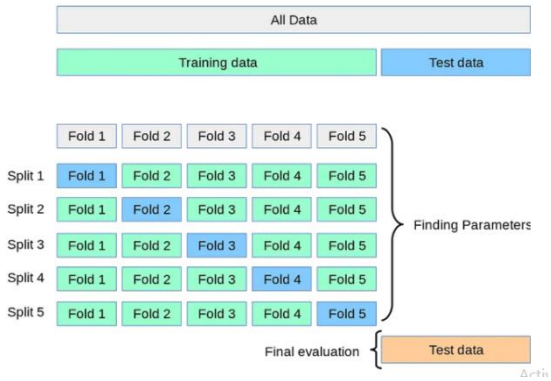

Figure 1.15- K-Fold Cross Validation

Principe: Les données sont divisées en k sous-ensembles de taille égale. Le modèle est entraîné k fois en utilisant k-1 sous-ensembles comme ensemble d'entraînement et un sous-ensemble comme ensemble de validation.

Avantages: Le modèle présente à la fois un faible biais, une faible complexité temporelle et utilise l'ensemble de données pour la formation et la validation [17].

b. Validation croisée stratifiée k-fold

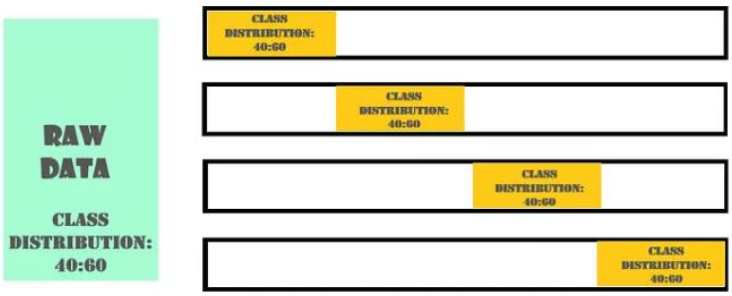

Figure 1.16- Validation croisée k-fold

Principe: Une variante de la validation croisée k-fold qui maintient la répartition des classes dans chaque fold. Utile pour les problèmes de classification où les classes sont déséquilibrées.

Avantages: Le modèle est efficace pour gérer les ensembles de données déséquilibrés [17].

c. Leave-one-out (LOO) cross-validation

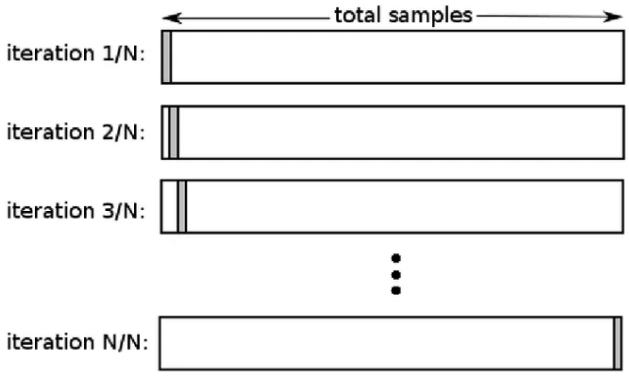

Figure 1.17- Leave-one-out

Principe: Chaque observation est utilisée comme ensemble de validation une seule fois, tandis que le reste des données est utilisé pour l'entraînement. Utile pour les ensembles de données de petite taille.

Avantages: Le modèle est caractérisé par sa simplicité, sa facilité de compréhension et de mise en œuvre [17].

d. Leave-p-out (LPO) cross-validation

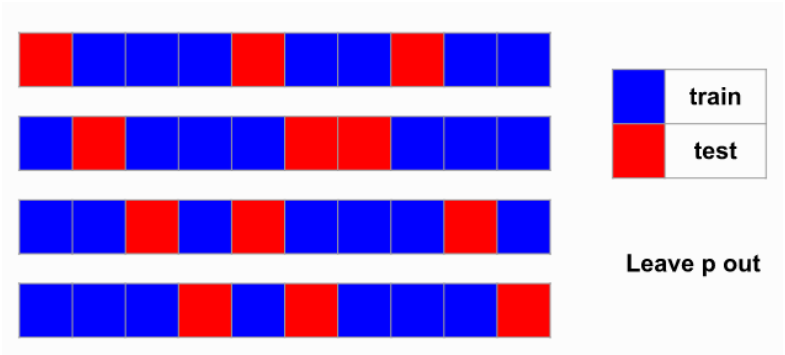

Figure 1.18- Leave-P-Out

Principe: Laisse p observations comme ensemble de validation et utilise le reste des données pour l'entraînement.

Avantages: Apporte une plus grande flexibilité dans le choix du nombre d'observations pour la validation [17].

1.7.5 Les domaines d'application du DL

1.7.5.1 Cyber sécurité

L'intelligence artificielle basée sur l'apprentissage profond est particulièrement adaptée aux cyberattaques car elle peut détecter les violations dans le fonctionnement des systèmes informatiques. C'est aussi un outil efficace pour la vidéosurveillance de sites sensibles comme les aéroports [18].

1.7.5.2 Les robots industriels

Les robots industriels optimisant de manière autonome les processus de production observent l'intelligence humaine. C'est ainsi que l'apprentissage en profondeur est utile dans de nombreux secteurs de l'industrie [18].

1.7.5.3 En médecine

En scannant les images avec beaucoup plus de précision qu'un œil humain entraîné, l'IA est une aubaine pour la médecine. Par exemple, l'intelligence artificielle permet de faire la distinction entre les tumeurs cancéreuses et non cancéreuses ou de détecter la maladie beaucoup plus tôt qu'auparavant [18].

1.7.5.4 Dans le domaine de l'agriculture

Dans la communauté de l'agriculture biologique, certains utilisent désormais des drones intelligents capables de scanner des hectares de cultures pour identifier les mauvaises herbes. C'est suffisant pour permettre aux agriculteurs de se concentrer sur les zones qui ont besoin d'être désherbées sans dépenser trop de temps et d'énergie [18].

1.7.5.5 Détection de fraude

Il existe de nombreuses façons d'utiliser l'apprentissage en profondeur pour la détection des fraudes. L'une consiste à former un modèle pour détecter les schémas frauduleux connus. Cela peut être fait en fournissant au modèle un ensemble de cas d'escroquerie connus. Vous pouvez ensuite utiliser le modèle pour signaler de nouveaux cas similaires à ceux de votre ensemble de données. C'est l'une des meilleures applications d'apprentissage en profondeur [18].

1.8 Conclusion

Dans ce chapitre nous avons parlé tout d'abord de l'intelligence artificielle : sa définition, son histoire ainsi que ces avantages ces inconvénients et ces limites.

La seconde partie de ce chapitre a été consacrée en Machine Learning sa définition les différents types d'apprentissage automatique et les principaux algorithmes de Machine Learning .En suite, Deep Learning nous avons également cité la définition, la différence entre le machine Learning et le deep Learning et les Réseaux de neurones artificiels ainsi que ces domaines d'application .Finalement, nous avons parlé sur la Dataset Comment créer un data set ,ou se trouve et son principe dans machine Learning .

Pour les chapitres suivants nous avons s'intéresser a la classification de diabète en utilisant le Deep Learning.

Chapitre 2 :

Le diabète

2.1 Introduction

Le diabète est une maladie qui touche de plus en plus de personnes. En Suisse, environ 0,3% de la population souffre de diabète de type 1, qui n'est pas très répandu. Or, le diabète de type 2 touche 15 à 20 % de la population. Il s'agit d'un phénomène inquiétant car il touche un nombre croissant d'enfants et d'adolescents, ainsi que des personnes d'âge moyen. Cependant, dans certains cas, une simple prévention peut réduire le développement de la maladie. Mais nous ne parlons pas de diabète, encore moins de sucre dans le sang. Le glucose est le niveau de sucre dans votre sang. Cette présence de sucre est appelée glucose. Lorsque la glycémie est supérieure à la normale, on parle d'hyperglycémie. En revanche, lorsque le taux de glucose est inférieur à la normale, on parle d'hypoglycémie.

Le diabète est une maladie à progression lente. Il existe deux types de diabète. Le diabète de type 1 est appelé « diabète juvénile » car il survient généralement chez les enfants et les adolescents. Ensuite, le diabète de type 2, appelé "diabète de l'adulte" par ce que il touche généralement les personnes de plus de 40 ans. Les personnes atteintes de diabète devraient recevoir un traitement adapté à leur type de diabète. L'objectif du traitement est de maintenir la glycémie de la personne dans la plage normale [19].

2.2 Définition de diabète

Le diabète est une maladie chronique caractérisée par une glycémie excessive ou une hyperglycémie. En raison de divers troubles fonctionnels, il existe deux principaux types de diabète : le diabète de type 1 et le diabète de type 2 [20].

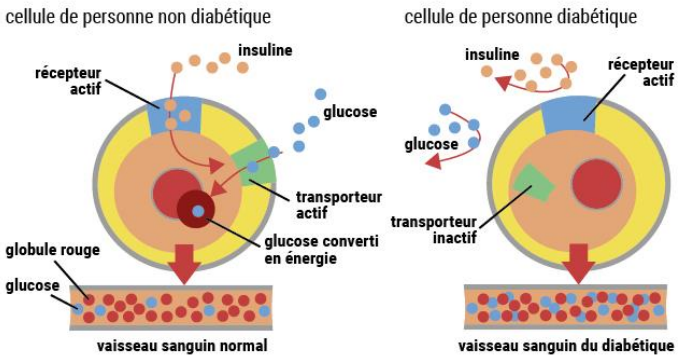

Figure 2.1- Qu'est ce que le diabète

2.3 Les deux types de diabète

Fondamentalement, il existe deux types de diabète. Le diabète de type 1 touche environ 6 % des personnes atteintes de diabète et le diabète de type 2qui en touche 92 %. Les autres types de diabète constituent les 2 % restants (MODY, LADA ou diabète secondaire à certaines maladies ou médicaments) [21].

2.3.1 Le diabète de type 1 (diabète insulino-dépendant ou DID)

Le diabète de type 1, anciennement appelé diabète insulino-dépendant (IDD), est généralement diagnostiqué chez les jeunes, tels que les enfants, les adolescents ou les jeunes adultes.

2.3.2 Le diabète de type 2

Le diabète de type 2 touche généralement les personnes de plus de 40 ans. Cependant, les premiers cas de la maladie chez les adolescents et les jeunes apparaissent en France. Le surpoids, l'obésité et le manque d'activité physique sont des causes évidentes du diabète de type 2 chez les personnes ayant une prédisposition génétique. L'évolution insidieuse et indolore du diabète de type 2 peut passer longtemps inaperçue. On estime qu'il s'écoule en moyenne 5 à 10 ans entre l'apparition de la première hyperglycémie et le diagnostic.

Le diabète de type 2, anciennement appelé non insulino-dépendant (DNID), a une évolution différente du diabète de type 1. Deux conditions peuvent provoquer une hyperglycémie:

- soit le pancréas fabrique toujours de l'insuline mais pas assez, par rapport à la glycémie: c'est l'insulinopénie.
- soit cette insuline agit mal

L'insuline ne peut plus réguler la glycémie, et cette résistance épuise progressivement le pancréas jusqu'à ce qu'il ne produise plus assez d'insuline. Ces deux mécanismes empêchent le glucose de pénétrer dans les cellules du corps et de rester dans la circulation sanguine. Les niveaux de sucre dans le sang ne sont pas contrôlés par l'insuline.

2.4 Les causes et facteurs de risque

2.4.1 Les causes

2.4.1.1 Les causes du diabète de type 1

La cause exacte du diabète de type 1 est inconnue. Le système immunitaire de l'organisme, qui combat normalement les bactéries et les virus nocifs, détruit les cellules productrices d'insuline (îlots) dans le pancréas. D'autres raisons possibles incluent [22]:

- La génétique.
- Exposition aux virus et autres facteurs environnementaux.

2.4.1.2 Les causes du diabète de type 2

Il n'y a pas de cause spécifique, mais les facteurs contributifs incluent [19]:

- Origine génétique: Le facteur familial prédomine absolument. Il existe souvent des antécédents familiaux du même type de diabète.
- Alimentation déséquilibrée, manque d'activité physique, excès de poids...

2.4.2 Les facteurs de risque

2.4.2.1 Les facteurs de risque du diabète de type 1

Les facteurs de risque du diabète de type 1 ne sont pas bien connus. Bien que la prédisposition génétique augmente, le risque de développer ce type de diabète est très modeste. Les experts ont également suggéré la possibilité d'une infection virale comme déclencheur. Par conséquent, sans connaissances suffisantes, il est actuellement difficile de prévenir le développement de ce type de diabète [20].

2.4.2.2 Les facteurs de risque du diabète de type 2

Voici les facteurs de risque les plus courants du diabète de type 2 [23]:

- Sexe: Les hommes sont plus susceptibles que les femmes.
- Âge: le risque augmente avec l'âge.
- En surpoids.
- Taille haute ou accumulation de graisse autour de l'abdomen.
- Niveau d'activité physique et habitudes alimentaires.
- Hypertension artérielle.
- Antécédents de glycémie anormalement élevée.
- Pour les femmes qui ont donné naissance à des bébés pesant plus de 4,1 kg (9 lb).
- Héritage.
- Origine ethnique : Aborigène, Africain, Asiatique, Latinx, etc…
- Niveau d'éducation.

2.5 Prévalence du diabète dans le monde

En 2014, le diabète touchait 422 millions de personnes dans le monde, mais en 1980, il ne touchait que 108 millions de personnes dans le monde et était le premier de l'Organisation mondiale de la santé (OMS) et de la Fédération internationale du diabète (FID). La quatrième prédiction était liée au risque de développer diabète. 1990. En 2025, 240 millions de personnes seront touchées... En 2019, plus de 463 millions de personnes dans le monde sont atteintes de diabète, dont 59 millions en Europe (Source : International Diabetes Federation Atlas 2019). Plus de 537 millions de personnes dans le monde (soit 1 sur 10) seront

diabétiques en 2021, dont 61 millions en Europe (Source : Atlas 2021 de la Fédération Internationale du Diabète) [24].

2.6 Le diagnostic et le traitement du diabète

2.6.1 Tests diagnostiques pour le diabète

Divers tests sont utilisés pour diagnostiquer (reconnaître) le diabète. Ce test est effectué en même temps qu'une prise de sang régulière. Il y a 4 épreuves. Ils sont expliqués ci-dessous. Tous ces tests sont utilisés pour vérifier le niveau de glucose (sucre) dans le sang [25].

- Test de glycémie aléatoire.
- Test de glycémie à jeun.
- Épreuve d'hyperglycémie provoquée par voie orale.
- Mesure de l'hémoglobine glyquée (A1c) .

2.6.1.1 Test de glycémie aléatoire

Ce test mesure le taux de sucre dans le sang. Il est basé sur un échantillon de sang.Pour ce test, l'heure à laquelle vous avez bu ou mangé pour la dernière fois n'a pas d'importance. Les résultats indiquent votre glycémie au moment de la prise de sang. Ces résultats peuvent être influencés par l'heure à laquelle vous avez mangé pour la dernière fois et par ce que vous avez mangé et bu pendant la journée.

2.6.1.2 Test de glycémie à jeun

Ce test mesure votre taux de sucre dans le sang. Il est basé sur un échantillon de sang.Vous devrez jeûner (ne pas manger ni boire) au moins huite heures pour ce test. Les résultats indiquent votre glycémie lorsque vous ne mangez ou ne buvez rien.

2.6.1.3 Épreuve d'hyperglycémie provoquée par voie orale

Ce test mesure votre taux de sucre dans le sang. Il est basé sur un échantillon de sang. Vous devrez boire une boisson sucrée pour ce test. Le résultat montre votre taux de sucre dans le sang après avoir mangé une certaine quantité de sucre. Ce test n'est pas effectué en même temps qu'une prise de sang régulière et doit être spécifiquement prescrit par votre médecin.

2.6.1.4 Mesure de l'hémoglobine glyquée (A1c)

Ce test évalue votre contrôle de la glycémie au cours des 3 derniers mois. Il est basé sur un échantillon de sang. Peu importe quand vous avez mangé ou bu pour la dernière fois pour ce test. Ce test compte le nombre de cellules sanguines contenant du sucre. Les résultats sont exprimés en pourcentages (%). Les lectures résultantes peuvent être converties en taux moyens de glycémie.

2.6.2 Options de traitement pour le diabète de type 1 et de type 2

Le diabète de type 1 et le diabète de type 2 sont les formes les plus courantes de diabète. Causées par diverses causes, ces maladies provoquent une hyperglycémie. Le seul traitement du diabète de type 1 consiste en des injections d'insuline à vie. Le traitement standard du diabète de type 2 est l'optimisation du mode de vie. Selon vos besoins, une perte de poids, une activité physique régulière et une alimentation équilibrée peuvent dans un premier temps suffire à contrôler votre glycémie. La deuxième intention est de prescrire des médicaments oraux et/ou injectables contre le diabète pour contrôler la glycémie [26].

2.6.3 Auto-surveillance de la glycémie et de l'hémoglobine A1c

La surveillance de la glycémie capillaire fournit des informations sur les tendances de la glycémie du patient dans un avenir proche, permettant une surveillance quotidienne. D'autre part, la mesure de l'hémoglobine A1c permet de connaître la quantité de glucose accumulée dans les globules rouges pendant 120 jours. Un test HbA1c tous les 3 mois peut vous donner des informations générales sur votre équilibre diabétique. L'HbA1c normale est un taux d'HbA1c de 7 % ou moins. Un taux d'HbA1c élevé est un taux d'HbA1c supérieur à 7 %, selon le profil et les objectifs du patient. Une HbA1c élevée entraîne souvent des complications, il est donc important de la surveiller de près pour effectuer les ajustements de traitement nécessaires si nécessaire [27].

2.6.4 Prise en charge du diabète gestationnel

La gestion du diabète gestationnel (DG) est très importante pour la santé de la mère et du fœtus. Voici quelques éléments clés du soutien de la DG [28]:

- Dépistage: toutes les femmes enceintes doivent être dépistées pour le DG.
- Régime alimentaire: une alimentation équilibrée est essentielle pour le contrôle de la glycémie chez les femmes atteintes de DG.
- Activité physique: l'exercice régulier peut aider à maintenir une glycémie normale chez les femmes atteintes de DG.
- Accouchement: les femmes atteintes de DG ont un risque accru de complications lors de l'accouchement, telles que des problèmes de glycémie chez le nouveau-né. Votre médecin peut recommander des mesures spéciales pendant l'accouchement pour assurer votre sécurité et celle de votre bébé.

2.7 La prévention du diabète et la gestion du mode de vie

2.7.1 Stratégies de prévention pour le diabète de type 2

Chaque jour, nous faisons des choix qui affectent notre santé. Prenez ces mesures importantes pour maintenir un mode de vie sain afin de prévenir ou de réduire votre risque de développer un diabète de type 2 ou un prédiabète [29].

- Maintenez un poids santé.
- Combinez une activité physique équilibrée avec une alimentation saine.
- Discutez avec votre médecin d'un poids santé.
- Apprenez à calculer votre indice de masse corporelle (IMC).
- Ayez une alimentation saine.
- Ayez une alimentation variée.
- 5 à 10 fruits et légumes par jour (Guide alimentaire canadien).
- Augmenter l'apport en fibres.
- Réduire l'apport en matières grasses et en sel.
- Limiter la consommation d'alcool.
- Choisir les bonnes portions.
- Faire de l'exercice régulièrement.
- Être actif pendant au moins 30 minutes par jour.
- Activités qui renforcent force et endurance Choix, souplesse.

2.7.2 Gestion de l'alimentation et de l'activité physique pour les personnes attentes le diabète

La gestion de l'alimentation et de l'activité physique est très importante pour les personnes atteintes de prédiabète, car elle peut aider à prévenir ou à retarder l'apparition du diabète de type 2. Voici quelques recommandations pour gérer votre alimentation et votre activité physique si vous souffrez de prédiabète [30].

Adopter une alimentation saine: les personnes atteintes de prédiabète doivent avoir une alimentation saine et équilibrée pour maintenir une glycémie normale. Ceux-ci peuvent inclure des aliments riches en fibres, en protéines, en graisses saines et en glucides complexes, et les aliments transformés riches en sucre ajouté doivent être évités.

Contrôler la taille des portions: manger la bonne quantité aide à maintenir une glycémie normale. Il est important de comprendre les portions recommandées et de mesurer vos aliments si nécessaire.

Éviter les boissons sucrées: les boissons sucrées, telles que l'eau gazeuse et les jus de fruits, peuvent rapidement augmenter la glycémie. Il est préférable de boire de l'eau, du thé ou du café non sucré.

Augmenter l'activité physique: l'exercice régulier peut aider à prévenir ou à retarder l'apparition du diabète de type 2.

Réduire le temps sédentaire: le temps passé assis peut augmenter votre risque de développer un diabète. 2. Il est recommandé de faire des pauses fréquentes, notamment de marcher ou de s'étirer pour se lever et bouger.

Perdre du poids si nécessaire: une perte de poids modérée peut aider à prévenir ou à retarder l'apparition du diabète de type 2 chez les personnes atteintes de prédiabète. Il est recommandé de viser une perte de 5 à 7 % du poids corporel initial.

2.7.3 Conseils pour maintenir une glycémie saine

Voici quelques conseils pour maintenir une glycémie saine [31]:

- Faites de l'exercice régulièrement.
- Mangez plus de fibres.
- Mangez moins de glucides.
- Buvez plus d'eau.
- Mangez des aliments à faible indice glycémique.
- Gérez le stress.

2.8 Impact psychologique et social du diabète

2.8.1 L'influence du diabète sur votre santé mentale

Le diabète est une maladie qui peut être un lourd fardeau mental et physique pour les patients et leurs familles. Et les diabétiques doivent constamment réfléchir à la manière de faire face à leur maladie. Sur le plan physique, le diabète peut entraîner d'autres problèmes de santé et complications [32].

Psychologiquement, les personnes atteintes de diabète peuvent ressentir une plus grande anxiété. En conséquence, des symptômes tels que:

- Vous vérifiez trop souvent votre glycémie.
- Faites toujours attention pour éviter de nouvelles complications.
- Vous vous inquiétez de l'impact du diabète sur votre vie personnelle et professionnelle.

Valérie S. Legendre est psychologue clinicienne et consultante principale en santé mentale pour la gestion de l'invalidité, l'assurance-vie et les soins de santé intégrés à la Sun Life dans l'Est du Canada. Selon elle, le diabète peut avoir un impact profond sur l'état émotionnel et mental d'une personne.

2.8.2 Diabète et vie sociale

Le diabète peut avoir un impact social important sur les personnes atteintes de la maladie, leurs familles et leurs communautés. Les conséquences sociales du diabète comprennent [33]:

Stigmatisation: Certaines personnes peuvent être blâmées pour leur maladie en raison de leur mode de vie ou de leurs habitudes alimentaires, ce qui peut entraîner une discrimination au travail ou dans d'autres domaines de la vie.

Impact économique: Les frais médicaux tels que les médicaments, les appareils de surveillance de la glycémie et les visites chez le médecin peuvent être élevés.

Impact sur la qualité de vie: Les complications du diabète, telles que les problèmes de vision, les problèmes rénaux et les problèmes de circulation, peuvent affecter la capacité des personnes à travailler, à se déplacer et à effectuer leurs activités quotidiennes.

Stress émotionnel: Les personnes atteintes de diabète peuvent se sentir anxieuses ou déprimées par leur maladie et les complications possibles.

Impact sur les relations sociales: Les personnes atteintes de diabète peuvent avoir des difficultés à suivre un régime ou à prendre des médicaments lorsqu'elles sont avec des amis et de la famille, ce qui peut affecter leur participation à des activités sociales.

2.9 Le développement des traitements de la maladie diabète

Il y a eu de nombreux développements dans le domaine du traitement et de la gestion du diabète au cours des dernières années. Voici quelques-unes des avancées les plus récentes:

Thérapies géniques: La thérapie génique est une technique prometteuse pour le traitement du diabète de type 1. Les chercheurs travaillent sur l'identification des gènes responsables de la production d'insuline et sur l'utilisation de la thérapie génique pour les remplacer ou les réparer [34].

Thérapies par cellules souches: Les chercheurs travaillent comment utiliser les cellules souches pour produire des cellules pancréatiques capables de produire de l'insuline pouvant être transplantée chez les personnes atteintes de diabète de type 1[35] .
Surveillance de la glycémie en continu: Les dispositifs de surveillance continue de la glycémie permettent aux diabétiques de surveiller leur glycémie en temps réel et d'adapter le traitement en conséquence [36].

Intelligences artificielles et apprentissages automatiques: Il utilise des algorithmes avancés pour analyser les données de surveillance de la glycémie et peut aider les diabétiques à prévoir et à prévenir les épisodes d'hypoglycémie et d'hyperglycémie [37].

Pancréas artificiel: Le système de pancréas artificiel combine un moniteur de glycémie en continu et une pompe à insuline pour ajuster automatiquement la quantité d'insuline administrée en fonction de la glycémie [38].

2.10 Conclusion

Dans ce chapitre nous avons présenté la maladie du diabète ,leur différents types, les causes et facteurs de risque ainsi que la prévalence du diabète dans le monde et le diagnostic et le traitement du diabète et nous avons aussi parlé de la prévention du diabète et la gestion du mode de vie , impact psychologique et social du diabète et a la fin le développement des traitements de la maladie diabète . Dans le prochain chapitre nous allons parler sur le fonctionnement et les techniques de deep learning dans le domaine médical et en particulier l'utilisation de deep learning pour la prédiction du diabète.

Chapitre 3 :

Le deep learning en détection et prédiction de diabète

3.1 Introduction :

Ces dernières années, la prévalence du diabète a augmenté rapidement, ce qui représente un défi important pour la santé mondiale. La détection précoce et le diagnostic précis du diabète sont essentiels pour une gestion efficace et la prévention des complications qui y sont associées. Avec les progrès technologiques et la disponibilité des datasets de santé à grande échelle, les techniques de deep learning sont devenues des outils puissants pour le diagnostic médical et les tâches de prédiction. Dans ce chapitre, nous nous concentrons sur le développement d'un modèle de deep learning pour la détection du diabète en utilisant le "Health Indicator Diabetes Dataset". Le Health Indicator Diabetes Dataset est une collection complète de dossiers médicaux de patients, de mesures cliniques et d'informations sur le mode de vie, spécialement conçue pour la détection du diabète.

3.2 Outils et environnement de développement

3.2.1 Le langage de programmation utilisé (Python)

Python est le langage de programmation le plus utilisé aujourd'hui. Lorsqu'il s'agit de résoudre des problèmes et des défis liés à la science des données, Python ne cesse de vous étonner. La plupart des data scientists utilisent quotidiennement la puissance de la programmation Python. Python est construit à l'aide d'excellentes bibliothèques de manipulation de données que les programmeurs utilisent quotidiennement pour résoudre des problèmes [39].

3.2.2 kaggle

Kaggle est une plateforme web interactive qui propose des compétitions d'apprentissage automatique dans le domaine de la science des données. La plate-forme fournit des ensembles de données, des blocs-notes et des didacticiels gratuits dont les scientifiques des données ont besoin pour mener à bien leurs projets d'apprentissage automatique [40] (voir tableau 3.1).

Tableau 1.1- Le cahier kaggle de base [41]

Type de bloc-note	**Noyaux**	**Mémoire**	**Nombre de bloc-note qui peut être exécuté à la fois**	**Durée par semaine**
CPU	4	16GB	10	Illimité
GPU	2	13GB	2	30heures
TPU	4	16GB	2	30heures

3.2.3 Les bibliothèques Python

Tableau 1.2- Bibliothèques de Python [42]

TensorFlow	une bibliothèque d'apprentissage en profondeur open source pour les tâches d'IA telles que la classification d'images, la reconnaissance vocale et la traduction automatique.
Keras	une bibliothèque d'apprentissage en profondeur qui facilite la création et l'entraînement de modèles de réseaux de neurones.
NumPy	une bibliothèque de calcul scientifique avec des fonctions pour les matrices, les polynômes, les fonctions mathématiques...
Pandas	une bibliothèque de manipulation de données pour Python, fournissant des structures de données pour manipuler facilement des tableaux de données.
Matplotlib	une bibliothèque de visualisation de données 2D pour créer des tracés d'aspect professionnel.
Scikit-learn	une bibliothèque d'apprentissage automatique avec des outils de classification, de régression, de clustering...

3.3 Dataset

3.3.1 Définition de Dataset

Figure 3.1- Diabetes Health Indicators Dataset

Diabetes Health Indicators Dataset est une collection de données cliniques et/ou démographiques relatives aux personnes atteintes de diabète. Ces ensembles de données peuvent inclure des paramètres de base tels que l'âge, le sexe, l'IMC, la pression artérielle et la glycémie, ainsi que des paramètres plus complexes tels que le traitement, les complications et les résultats pour la santé. Les ensembles de données des indicateurs de santé du diabète sont souvent utilisés pour l'analyse statistique et pour la construction de modèles prédictifs du diabète. Il peut également être utilisé pour étudier les tendances de santé publique, notamment pour mieux comprendre l'incidence et la prévalence du diabète dans différentes populations. Cet dataset se compose de 3 fichiers:

- Le premier fichier, intitulé "diabète_012_santé_indicateurs_BRFSS2015.csv", est une base de données contenant 253 680 réponses . La variable cible, Diabetes_012, comporte trois classes : 0 pour l'absence de diabète , 1 pour le prédiabète et 2 pour le diabète. Cependant, il y a un déséquilibre de classe dans cet ensemble de données. Il comprend aussi 21 variables de caractéristiques qui peuvent être utilisées pour analyser les facteurs liés au diabète.
- Le deuxième fichier, "diabète_binaire_5050split_santé_indicateurs_BRFSS2015.csv", est un ensemble de données épurées de 70 692 réponses . Il a une répartition égale 50-50 de répondants sans diabète et avec prédiabète ou diabète. La variable cible Diabetes_binary a 2 classes. 0 correspond à l'absence de diabète et 1 au prédiabète ou au diabète. Cet ensemble de données a 21 variables de caractéristiques et est équilibré.
- Le troisième fichier, "diabète_binaire_santé_indicateurs_BRFSS2015.csv", est un ensemble de données épurées de 253 680 réponses . La variable cible Diabetes_binary a 2 classes. 0 correspond à l'absence de diabète et 1 au prédiabète ou au diabète. Cet ensemble de données comporte 21 variables d'entité et n'est pas équilibré.

3.3.2 Informations sur dataset

Tableau 1.3- La base de données

Nom	Diabetes Health Indicators Dataset (en Anglais) ou Ensemble de données sur les indicateurs de santé du diabète.
Année	2015.
Type	Binaire(imbalanced)/Binaire (Balanced)/Multiclass.
Nombre d'observation	253 680/70 692/253 680.
Nmobre de feature (caractéristique)	21.

3.3.3 Description des variables d'ensemble de données

Tableau 1.4- Définition des variables

Abréviation	**Signification**
BP élevé(HighBP)	0 = pas de PA élevée 1 = PA élevée
HighChol(HighChol)	0 = pas d'hypercholestérolémie 1 = hypercholestérolémie
CholCheck(CholCheck)	0 = pas de contrôle du cholestérol dans 5 ans 1 = oui contrôle du cholestérol dans 5 ans
IMC(BMI)	Indice de masse corporelle
Fumeur(Smoker)	Avez-vous fumé au moins 100 cigarettes dans toute votre vie ? [Remarque : 5 paquets = 100 cigarettes] 0 = non 1 = oui
Accident vasculaire cérébral(Stroke)	vous avez eu un accident vasculaire cérébral. 0 = non 1 = oui
CoeurMaladieOuAttaque(HeartDiseaseorAttack)	maladie coronarienne (CHD) ou infarctus du myocarde (IM) 0 = non 1 = oui
PhysActivity(PhysActivity)	activité physique au cours des 30 derniers jours - travail exclu 0 = non 1 = oui
Des fruits(Fruits)	Consommer des fruits 1 fois ou plus par jour 0 = non 1 = oui
Légumes(Veggies)	Consommer des légumes 1 fois ou plus par jour 0 = non 1 = oui
HvyConsommation D'Alcool(HvyAlcoholConsump)	Gros buveurs (hommes adultes buvant plus de 14 verres par semaine et femmes adultes buvant plus de 7 verres par semaine) 0 = non 1 = oui
AnyHealthcare(AnyHealthcare)	Avoir tout type de couverture de soins de santé, y compris l'assurance maladie, les plans prépayés tels que HMO, etc. 0 = non 1 = oui
NoDocbcCost(NoDocbcCost)	Y a-t-il eu un moment au cours des 12 derniers mois où vous avez eu besoin de consulter un médecin, mais que vous n'avez pas pu en raison du coût ? 0 = non 1 = oui
GénSanté(GenHlth)	Diriez-vous qu'en général votre santé est : échelle de 1 à 5 1 = excellent 2 = très bon 3 = bon 4 = passable 5 = mauvais
Santé mentale(MentHlth)	En pensant maintenant à votre santé mentale, qui comprend le stress, la dépression et les problèmes émotionnels, pendant combien de jours au cours des 30 derniers jours votre santé mentale n'a-t-elle pas été bonne ? échelle 1-30 jours
PhysSanté(PhysHlth)	Maintenant, en pensant à votre santé physique, qui comprend les maladies physiques et les blessures, pendant combien de jours au cours des 30 derniers jours votre santé physique n'a-t-elle pas été bonne ? échelle 1-30 jours
Marche Diff(DiffWalk)	Avez-vous de sérieuses difficultés à marcher ou à monter des escaliers ? 0 = non 1 = oui
Sexe(Sex)	0 = femelle 1 = mâle
Âge(Age)	Catégorie d'âge à 13 niveaux (_AGEG5YR voir codebook) 1 = 18-24 9 = 60-64 13 = 80 ou plus
Éducation(Education)	Échelle de niveau d'éducation 1-6 1 = Jamais fréquenté l'école ou uniquement la maternelle 2 = De la 1re à la 8e année (élémentaire) 3 = De la 9e à la 11e année (certaines études secondaires) 4 = 12e année ou GED (Diplôme d'études secondaires) 5 = Collège 1 an à 3 ans (Certains collèges ou écoles techniques) 6 = Collège 4 ans ou plus (Diplôme collégial)
Revenu(Income)	Échelle de revenu échelle 1-8 1 = moins de 10 000 $ 5 = moins de 35 000 $ 8 = 75 000 $ ou plus

3.4 Implémentation et résultat

3.4.1 Implémentation proposée

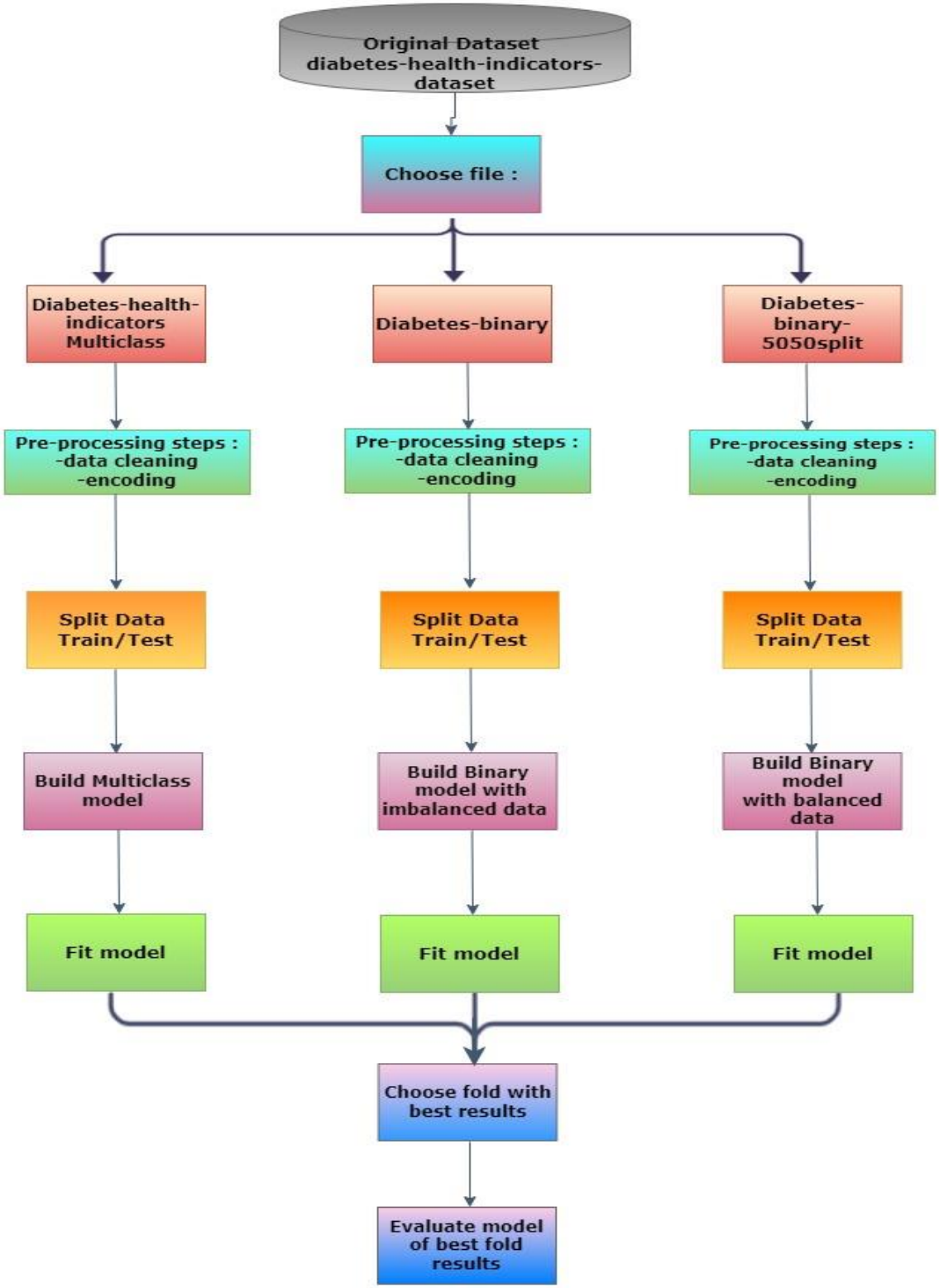

Figure 3.2- Architecture global des différentes approches effectuées

3.4.2 Prétraitement des données

Le prétraitement des données joue un rôle essentiel lors de l'utilisation du deep learning, car il permet de préparer les données de manière adéquate pour l'entraînement des modèles. L'importance du prétraitement des données réside dans le fait qu'il contribue à garantir la qualité, la fiabilité et la performance des résultats obtenus.

Le prétraitement des données implique la gestion des valeurs manquantes, des valeurs aberrantes et des erreurs dans le dataset. En éliminant ces anomalies, on obtient un jeu de données plus propre et plus fiable, ce qui permet d'éviter des erreurs ou des biais lors de l'entraînement du modèle.

Le prétraitement des données comprend souvent la normalisation ou la mise à l'échelle des données. Cela permet de ramener les valeurs des différentes variables à une même échelle, ce qui facilite l'apprentissage et la convergence des modèles. Sans une normalisation appropriée, certaines variables pourraient dominer l'apprentissage par rapport à d'autres, ce qui pourrait fausser les résultats.

Le prétraitement des données peut également inclure des techniques spécifiques en fonction du type de données utilisées. Par exemple, pour les données textuelles, la tokenisation, le stemming et la lemmatisation sont souvent utilisés pour représenter les mots de manière standardisée. Pour les données d'images, le redimensionnement, la normalisation des pixels et l'augmentation de données peuvent être nécessaires pour obtenir des résultats optimaux.

3.4.2.1 Nettoyage des données

Le nettoyage des données est une étape essentielle dans le processus de préparation des données. Il vise à garantir la qualité et l'intégrité des données en supprimant les valeurs manquantes, en traitant les valeurs aberrantes, en vérifiant la cohérence des données et en éliminant les doublons. Cette étape permet de s'assurer que les données sont prêtes à être utilisées dans des analyses ou des modèles d'apprentissage automatique, en minimisant les erreurs potentielles et en assurant des résultats fiables. Voici quelques fonctions couramment utilisées pour le nettoyage des données:

- **Drop:** la fonction "drop" supprime les lignes ou colonnes inutiles du jeu de données. Elle permet d'éliminer les données non pertinentes pour l'analyse.
- **Duplicate:** la fonction "duplicate" identifie et gère les doublons dans les données en trouvant les observations identiques ou similaires. Elle permet de prendre des mesures telles que la suppression ou la fusion des doublons pour éviter la redondance.
- **Fillna:** remplace les valeurs manquantes (NaN) dans un jeu de données en utilisant des valeurs spécifiées ou calculées (moyenne, médiane, mode), assurant ainsi la complétude des données pour l'analyse ultérieure.
- **Replace:** est une fonction qui permet de substituer des valeurs spécifiques par d'autres dans un jeu de données, ce qui permet de corriger des entrées incorrectes ou incohérentes et d'assurer la cohérence des données.
- **Dropna:** élimine les lignes ou les colonnes avec des valeurs manquantes, créant ainsi un jeu de données nettoyé et complet pour l'analyse ultérieure.

3.4.2.2 Normalisation des données

La normalisation des données est une étape essentielle qui vise à mettre toutes les variables d'un jeu de données sur une échelle commune, éliminant ainsi les biais liés aux unités de mesure et rendant les données comparables entre elles. Cela améliore les performances des modèles de deep learning en permettant une convergence plus rapide et une meilleure interprétation des résultats.

3.4.2.3 Encodage des variables catégorielles

L'encodage des variables catégorielles est une étape importante du prétraitement des données dans le deep learning. Il consiste à convertir les variables catégorielles en une forme numérique compréhensible par les algorithmes d'apprentissage automatique. Cela permet aux modèles d'apprendre à partir de ces variables et de les utiliser de manière appropriée lors de la prédiction ou de la classification.

3.4.2.4 Partitionnement du dataset

Le partitionnement du dataset consiste à diviser le dataset en ensembles distincts, tels que l'ensemble d'entraînement, l'ensemble de validation et l'ensemble de test, pour évaluer et optimiser les performances du modèle.

3.4.2.5 Gestion des déséquilibres de classe

La gestion des déséquilibres de classe est une approche visant à traiter les problèmes où les classes dans un dataset sont représentées de manière inégale, afin d'améliorer la performance des modèles sur les classes minoritaires.

3.4.3 Paramètres d'entraînement

3.4.3.1 Epoques (Epochs)

Les époques sont des itérations sur les données pendant la formation en apprentissage en profondeur. Augmenter les époques améliore les performances, mais trop d'époques peuvent causer un surapprentissage, donc il faut trouver un nombre optimal [43].

3.4.3.2 Fonction de perte(Loss)

Une fonction de perte évalue l'écart entre les prédictions et les valeurs réelles. On cherche à minimiser cette fonction pour améliorer le réseau de neurones. Les poids du réseau sont ajustés pour réduire l'écart entre les prédictions et les valeurs réelles [44].

Lors de la conception de notre modèle, nous avons choisi d'utiliser la fonction de perte [45] :

- **Binary_crossentropy**

La fonction de perte "binary_crossentropy": est utilisée dans les modèles de classification binaire. Elle mesure la différence entre les étiquettes réelles et les prédictions, en calculant l'entropie croisée. Cette fonction est idéale pour les problèmes où il n'y a que deux classes à prédire.

- **Categorical_crossentropy**

La fonction de perte "categorical_crossentropy" :est utilisée dans les modèles de classification multi-classes, où il y a plus de deux étiquettes de sortie. Dans ce cas, les étiquettes de sortie sont encodées sous forme de vecteurs binaires, où chaque classe a une valeur unique de 0 ou 1. Cette fonction mesure la divergence entre les distributions de probabilité des étiquettes réelles et les prédictions.

3.4.3.3 Fonction d'activation

Les fonctions d'activation sont essentielles en deep learning pour introduire une non-linéarité. Elles modélisent les relations complexes entre les données. Quelques fonctions couramment utilisées [46]:

- **Fonction d'activation Sigmoïde**

La fonction d'activation sigmoïde comprime les valeurs entre 0 et 1. Elle était utilisée dans les anciens réseaux de neurones, mais est moins populaire maintenant en raison de limitations, comme le problème de la disparition du gradient.

- **Fonction d'activation ReLU (Rectified Linear Unit)**

La fonction d'activation ReLU remplace les valeurs négatives par zéro et garde les valeurs positives. C'est une fonction populaire en raison de sa simplicité de calcul et de sa capacité à résoudre le problème de la disparition du gradient.

- **Fonction d'activation Softmax**

La fonction d'activation softmax est utilisée en sortie pour la classification multiclasse. Elle normalise les sorties en probabilités. Elle est définie par une exponentielle normalisée des valeurs d'entrée pour chaque classe.

3.4.3.4 Batch Normalization

La Batch Normalization (Batch Norm) est une méthode de normalisation qui est appliquée entre les couches d'un réseau neuronal, en utilisant des mini-lots au lieu de l'ensemble de données complet. Son objectif est d'accélérer l'apprentissage, de permettre l'utilisation de taux d'apprentissage plus élevés et d'améliorer la convergence du modèle [47].

3.4.3.5 Optimizer Adam

L'algorithme Adam est couramment utilisé pour l'entraînement de modèles de deep learning. Il étend la descente de gradient stochastique en exploitant les moyennes et les seconds moments des gradients pour ajuster de manière adaptative les taux d'apprentissage de chaque paramètre du modèle [48].

3.4.3.6 Optimizer SGD

La descente de gradient stochastique est un algorithme couramment utilisé en machine learning et deep learning. Il optimise les poids du modèle en se basfonction de perte calculé sur un sous-ensemble de données d'apprentissage aléatoire, plutôt que sur l'ensemble complet. Cela permet une optimisation efficace sur de grandes quantités de données [49].

3.4.4 Étude technique de la prédiction des diabètes

3.4.4.1 Classification "multi-classes"

Dans cette approche, nous avons utilisé le cas du dataset multiclass :

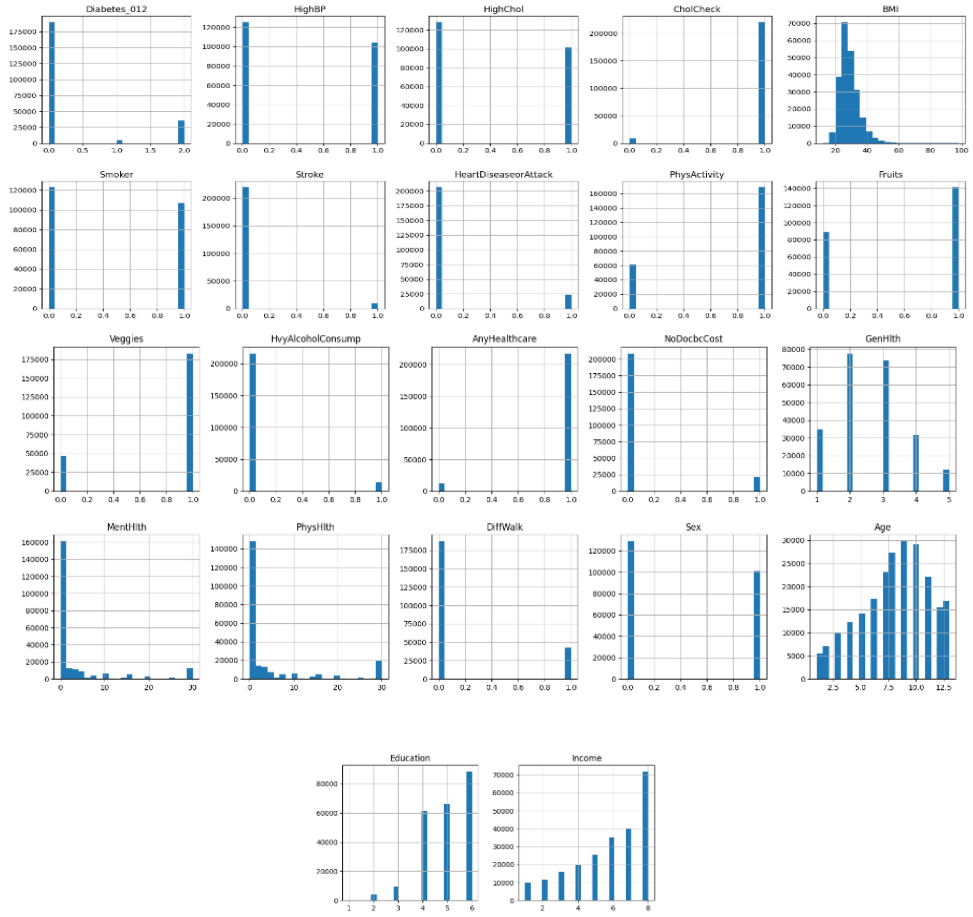

Figure 3.3- La représentation graphique des colonnes

Téléchargement des données

```
[1]: import pandas as pd
     data=pd.read_csv('/kaggle/input/diabetes-health-indicators-dataset/diabetes_012_health_indicators_BRFSS2015.csv')
```

Figure 3.4- Télécharger les données

Manipulation des données

```
[2]: data.head()
```

[2]:

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	...	1.0	0.0	5.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	...	0.0	1.0	3.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	5.0
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0

5 rows × 22 columns

Figure 3.5- Analyser les cinq premiers records de data set

```
[3]: data.shape
```

[3]: (253680, 22)

Figure 3.6- Déterminer le nombre de colonnes et de lignes présentes dans le data set

```
[4]: data.dtypes
```

[4]:

```
Diabetes_012            float64
HighBP                  float64
HighChol                float64
CholCheck               float64
BMI                     float64
Smoker                  float64
Stroke                  float64
HeartDiseaseorAttack    float64
PhysActivity            float64
Fruits                  float64
Veggies                 float64
HvyAlcoholConsump       float64
AnyHealthcare           float64
NoDocbcCost             float64
GenHlth                 float64
MentHlth                float64
PhysHlth                float64
DiffWalk                float64
Sex                     float64
Age                     float64
Education               float64
Income                  float64
dtype: object
```

Figure 3.7- Explorer les type des tous les colonnes de data

```
[5]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_012          253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
dtypes: float64(22)
memory usage: 42.6 MB
```

Figure 3.8- Explorer des informations sur le data

```
[6]: data.describe()
```

[6]:

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	
count	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.0
mean	0.296921	0.429001	0.424121	0.962670	28.382364	0.443169	0.040571	0.094186	0.756544	0.6
std	0.698160	0.494934	0.494210	0.189571	6.608694	0.496761	0.197294	0.292087	0.429169	0.4
min	0.000000	0.000000	0.000000	0.000000	12.000000	0.000000	0.000000	0.000000	0.000000	0.0
25%	0.000000	0.000000	0.000000	1.000000	24.000000	0.000000	0.000000	0.000000	1.000000	0.0
50%	0.000000	0.000000	0.000000	1.000000	27.000000	0.000000	0.000000	0.000000	1.000000	1.0
75%	0.000000	1.000000	1.000000	1.000000	31.000000	1.000000	0.000000	0.000000	1.000000	1.0
max	2.000000	1.000000	1.000000	1.000000	98.000000	1.000000	1.000000	1.000000	1.000000	1.0

8 rows × 22 columns

Figure 3.9- Un aperçu des statistiques numériques des valeurs présentes dans le data set

Nettoyage des données

```
[7]: data_clean=data.drop_duplicates(inplace=True)
```

```
[8]: data.shape
```

```
[8]: (229781, 22)
```

Figure 3.10- suppression des observation duplicate

Visualisation des données

[9]:
```
import matplotlib.pyplot as plt
vc = data['Diabetes_012'].value_counts(ascending=False)
classes = ['No Diabetes', 'Prediabetes', 'Diabetes']
values = vc.values
colors = ['#1f77b4', '#FF6347', '#FFA500']
plt.bar(classes, values, color=colors)
plt.title("Répartition des classes")
for i, value in enumerate(values):
    plt.text(i, value, str(value), ha='center', va='bottom')
plt.show()
```

[10]:
```
import matplotlib.pyplot as plt
vc = data['Diabetes_012'].value_counts(ascending=False)
colors = ['#1f77b4', '#FF6347', '#FFA500'] # bleu, rouge, orange, violet
plt.pie(x=vc.values,labels =['No Diabetes','Prediabetes', 'Diabetes'], explode=[0.0, 0.02,0.02], colors=colors, autopct='%1.1f%%')
plt.legend(title="Classes", loc="center left", bbox_to_anchor=(0.9, 0.8))#, 0.5, 1))
plt.show()
```

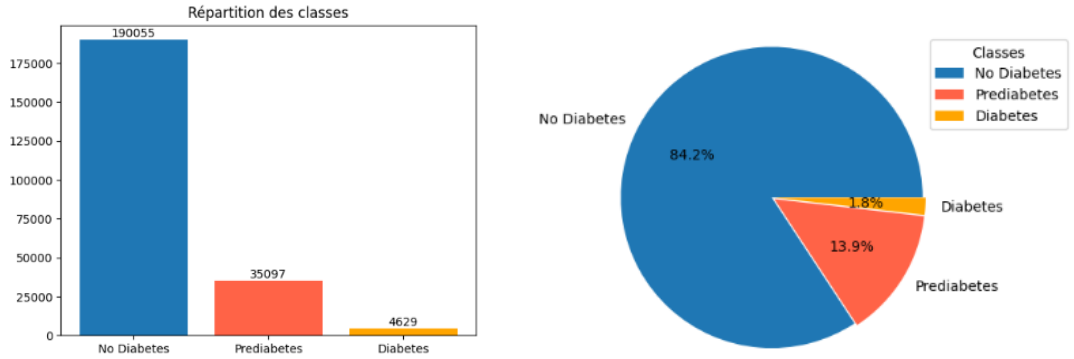

Figure 3.11- La distribution des données et des types de diabète dans dataset.

Diviser le data entre X et Y

[12]:
```
X=data.drop('Diabetes_012',axis=1)
y=data['Diabetes_012']
```

Figure 3.12- Apercu la division des données

A) Description de modèle

Nous avons utilisé Le modèle de réseau de neurones artificiels construit à l'aide de la bibliothèque Keras. Voici une description détaillée des différentes parties du modèle:

1. **Importation des bibliothèques:**

from keras.models import Sequential: Importe la classe Sequential de Keras pour créer le modèle.

from tensorflow import random: Importe le module random de TensorFlow pour la génération de nombres aléatoires.

import numpy as np: Importe la bibliothèque NumPy pour les opérations mathématiques.

2. **Initialisation des générateurs de nombres aléatoires:**

np.random.seed(42): Fixe la graine (seed) pour la génération de nombres aléatoires avec NumPy.

random.set_seed(42): Fixe la graine (seed) pour la génération de nombres aléatoires avec TensorFlow.

3. **Construction du modèle:**

model = Sequential(): Crée un modèle séquentiel vide.

4. **Importation des bibliothèques supplémentaires:**

keras.layers.Dense: Couche dense d'un réseau de neurones.

keras.regularizers.l2: Régularisation L2 utilisée dans la couche dense.

5. **Définition de la fonction build_model():** cette fonction crée et retourne le modèle Keras.

Un modèle est créé avec une couche d'entrée dense de 10 neurones, une fonction d'activation ReLU, et une régularisation L2 de 0.001. Il comprend également une couche cachée dense de 5 neurones avec une fonction d'activation ReLU et une régularisation L2 de 0.001. La couche de sortie est dense avec 3 neurones et utilise une fonction d'activation Softmax pour obtenir des probabilités de classe. Le modèle est compilé avec une fonction de perte categorical_crossentropy, un optimiseur Adam et une métrique de performance d'exactitude. Il est ensuite adapté aux données avec 5 époques et une taille de lot de 10.

Enfin, le modèle est encapsulé dans la classe KerasClassifier de scikit-learn, ce qui permet de l'utiliser dans des tâches de validation croisée. Nous avons utilisévalidation croisée K-Fold (K-Fold Cross Validation) avec 5 plis (folds) pour obtenir une estimation fiable des performances du modèle..

```
Model: "sequential_4"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 10)                220

 dense_1 (Dense)             (None, 5)                 55

 dense_2 (Dense)             (None, 3)                 18

=================================================================
Total params: 293
Trainable params: 293
Non-trainable params: 0
_________________________________________________________________
```

Figure 3.13- L'architecture de réseau de neurones profonds (DNN) utilisée.

- **Résultat obtenu**

Fold 1 :

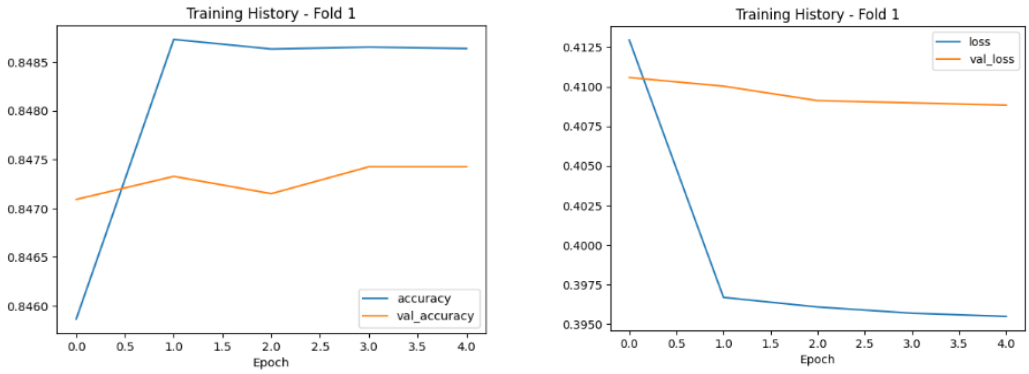

Figure 3.14- Un graphique illustrant l'accuracy et la perte pour la première validation croisée (fold 1)

Fold 2 :

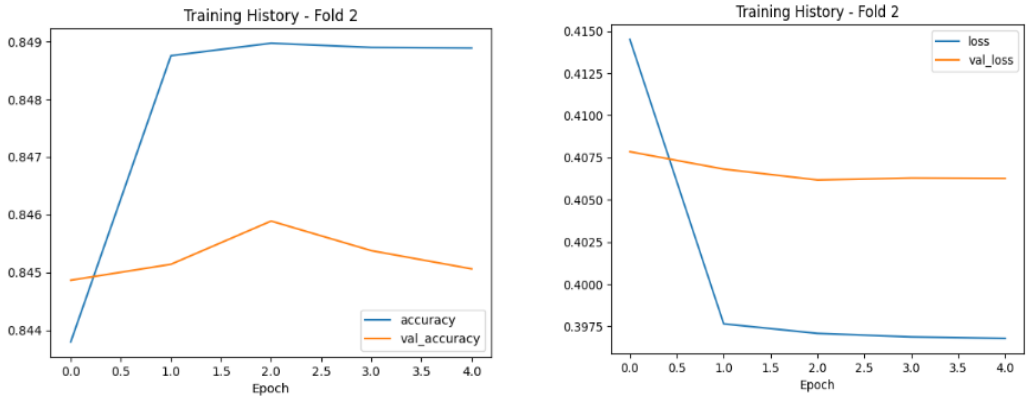

Figure 3.15- Un graphique présentant l'accuracy et la perte pour la deuxième validation croisée (fold 2)

Fold 3 :

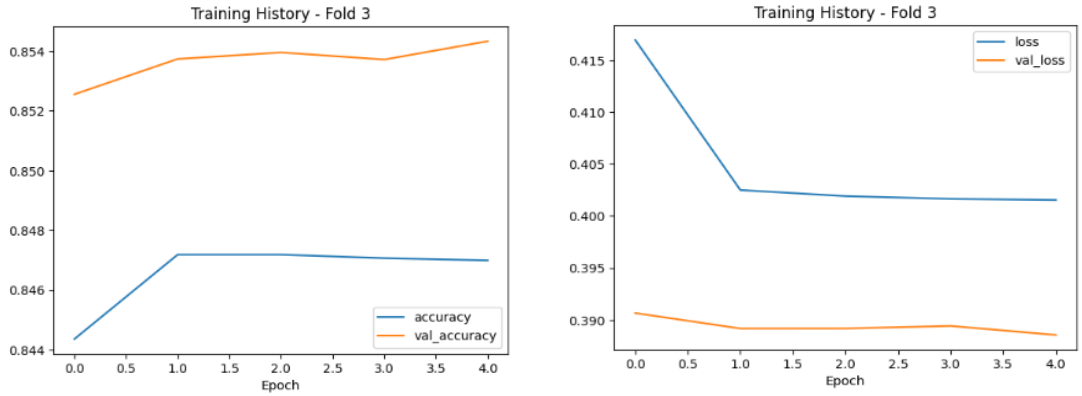

Figure3.16- Un graphique présentant l'accuracy et la perte pour la troisième validation croisée (fold 3)

Fold 4 :

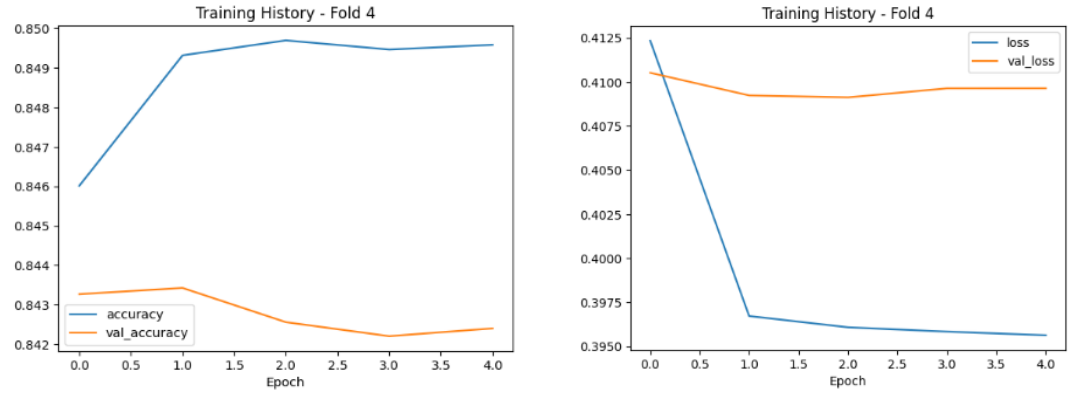

Figure 3.17- Un graphique montrant l'accuracy et la perte pour la quatrième validation croisée (fold 4)

Fold 5 :

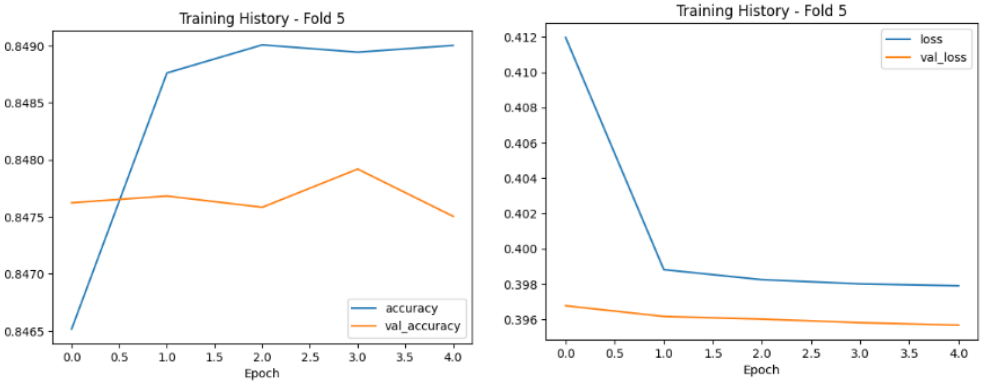

Figure 3.18- Un graphique montrant l'accuracy et la perte pour la cinquième validation croisée (fold 5)

Dans cette approche utilisant la validation croisée K-Fold avec 5 plis, le fold 3 s'est révélé être exceptionnel en termes de performance. Avec une précision (accuracy) de 85,43% et une perte (loss) de 38,81%, ce fold a démontré une capacité impressionnante à prédire avec précision les données de test.

3.4.4.2 Classification binaire

- **La première approche:** Dans cette approche, nous avons utilisé le cas du dataset binaire non-équilibré :

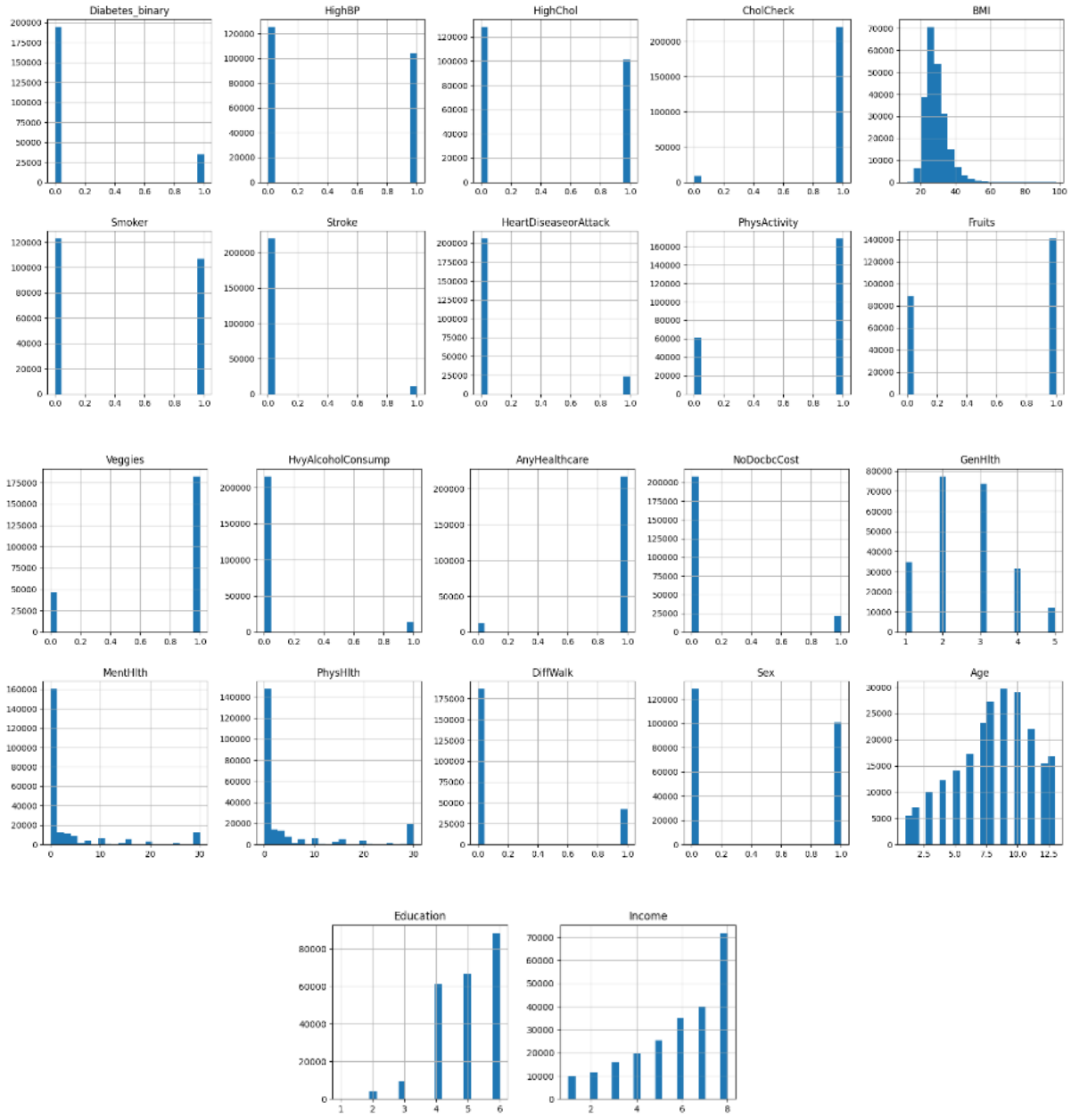

Figure 3.19- La représentation graphique des colonnes

Téléchargement des données

```
[1]: import pandas as pd
     data=pd.read_csv('/kaggle/input/diabetes-health-indicators-dataset/diabetes_binary_health_indicators_BRFSS2015.csv')
```

Figure 3.20- Télécharger les données

Manipulation des données

```
[2]: data.head()
```

[2]:

	Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	...	1.0	0.0	5.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	...	0.0	1.0	3.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	5.0
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0

5 rows × 22 columns

Figure 3.21- Analyser les cinq premiers records de data set

```
[3]: data.shape
```

```
[3]: (253680, 22)
```

Figure 3.22- Déterminer le nombre de colonnes et de lignes présentes dans le data set

```
[4]: data.dtypes
```

```
[4]: Diabetes_binary         float64
     HighBP                 float64
     HighChol               float64
     CholCheck              float64
     BMI                    float64
     Smoker                 float64
     Stroke                 float64
     HeartDiseaseorAttack   float64
     PhysActivity           float64
     Fruits                 float64
     Veggies                float64
     HvyAlcoholConsump      float64
     AnyHealthcare          float64
     NoDocbcCost            float64
     GenHlth                float64
     MentHlth               float64
     PhysHlth               float64
     DiffWalk               float64
     Sex                    float64
     Age                    float64
     Education              float64
     Income                 float64
     dtype: object
```

Figure 3.23- Explorer les type des tous les colonnes de data

```
[5]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_binary       253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
dtypes: float64(22)
memory usage: 42.6 MB
```

Figure 3.24- Explorer des informations sur le data

```
[6]: data.describe()
```

[6]:

	Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity
count	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000
mean	0.139333	0.429001	0.424121	0.962670	28.382364	0.443169	0.040571	0.094186	0.756544
std	0.346294	0.494934	0.494210	0.189571	6.608694	0.496761	0.197294	0.292087	0.429169
min	0.000000	0.000000	0.000000	0.000000	12.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	1.000000	24.000000	0.000000	0.000000	0.000000	1.000000
50%	0.000000	0.000000	0.000000	1.000000	27.000000	0.000000	0.000000	0.000000	1.000000
75%	0.000000	1.000000	1.000000	1.000000	31.000000	1.000000	0.000000	0.000000	1.000000
max	1.000000	1.000000	1.000000	1.000000	98.000000	1.000000	1.000000	1.000000	1.000000

8 rows × 22 columns

Figure 3.25- Un aperçu des statistiques numériques des valeurs présentes dans le data set

Nettoyage des données

```
[7]: data_clean=data.drop_duplicates(inplace=True)
```

```
[8]: data.shape
```

```
[8]: (229474, 22)
```

Figure 3.26- Suppression des observation duplicate

Visualisation des données

```
[9]:
import matplotlib.pyplot as plt
vc = data['Diabetes_binary'].value_counts(ascending=False)
classes = ['No Diabetes', 'Prediabetes ou Diabetes']
values = vc.values
colors = ['#1f77b4', '#FF6347']
plt.bar(classes, values, color=colors)
plt.title("Répartition des classes")
for i, value in enumerate(values):
    plt.text(i, value, str(value), ha='center', va='bottom')
plt.show()
```

```
[10]:
import matplotlib.pyplot as plt
vc = data['Diabetes_binary'].value_counts(ascending=False)
colors = ['#1f77b4', '#FF6347'] # bleu, rouge, orange, violet
plt.pie(x=vc.values, labels=['No Diabetes', 'Prediabetes ou Diabetes'], explode=[0.0, 0.02], colors=colors, autopct='%1.1f%%')
plt.legend(title="Classes", loc="center left", bbox_to_anchor=(0.9, 0.8))#, 0.5, 1))
plt.show()
```

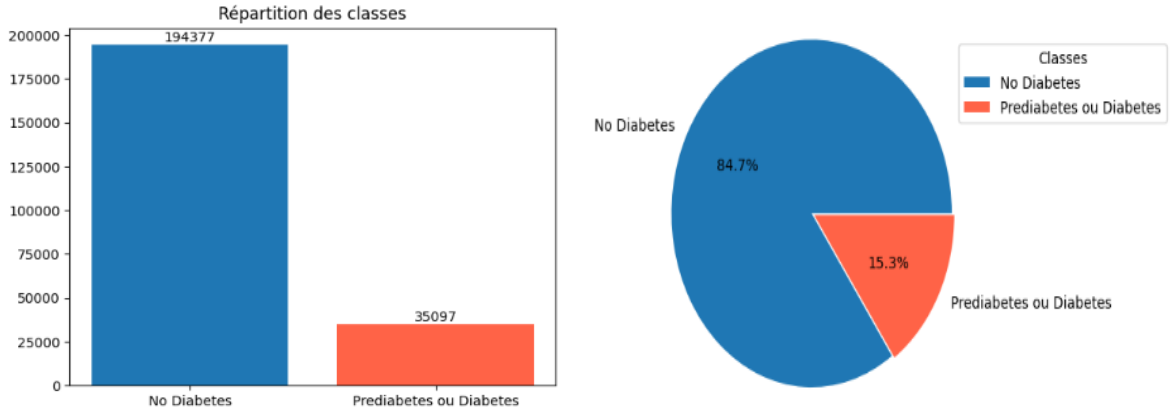

Figure 3.27- La distribution des données et des types de diabète dans dataset.

Diviser le data entre X et Y

```
[11]:
X=data.drop('Diabetes_binary',axis=1)
y=data['Diabetes_binary']
```

Figure 3.28- Apercu la division des données

A) Description de modèle

Nous avons utilisé un modèle construit avec un réseau de neurones artificiels en utilisant la bibliothèque Keras. Ce modèle est composé de trois couches denses, où chaque couche est connectée à la précédente. La première couche a 10 neurones avec une fonction d'activation rectifiée linéaire (ReLU) et une régularisation L2 avec un coefficient de pénalisation de 0.001. La deuxième couche a 5 neurones également avec une activation ReLU et une régularisation

L2. Enfin, la dernière couche a un seul neurone avec une fonction d'activation sigmoïde, qui est couramment utilisée pour la classification binaire.

Le modèle est compilé avec la fonction de perte binaire de la cross-entropy, optimisé par l'algorithme Adam, et évalué en utilisant la métrique d'exactitude (accuracy).Pour améliorer la stabilité du modèle et prévenir le surapprentissage (overfitting), une régularisation L2 est appliquée aux poids du modèle. Cela ajoute un terme de pénalité à la fonction de perte, encourageant les poids à rester petits et réduisant ainsi le risque de surapprentissage.

En utilisant l'ensemble de données préalablement mis à l'échelle, le modèle est entraîné en utilisant une validation croisée à k plis (k-fold cross-validation) avec un nombre de plis égal à 5. Cela signifie que l'ensemble de données est divisé en 5 parties égales, et le modèle est entraîné et évalué 5 fois, chaque fois en utilisant une partie différente comme ensemble de validation et les autres parties comme ensemble d'entraînement.

À chaque itération de la validation croisée, la fonction de perte (loss) et l'exactitude (accuracy) sont enregistrées pour évaluer les performances du modèle sur chaque pli. Ces mesures sont stockées dans les listes fold_loss et fold_accuracies respectivement, ce qui permet de suivre les performances du modèle sur l'ensemble des plis.

```
Model: "sequential_21"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_60 (Dense)            (None, 10)                220

 dense_61 (Dense)            (None, 5)                 55

 dense_62 (Dense)            (None, 1)                 6

=================================================================
Total params: 281
Trainable params: 281
Non-trainable params: 0
_________________________________________________________________
```

Figure 3.29- L'architecture de réseau de neurones profonds (DNN) utilisée.

Résultat obtenu

Fold 1 :

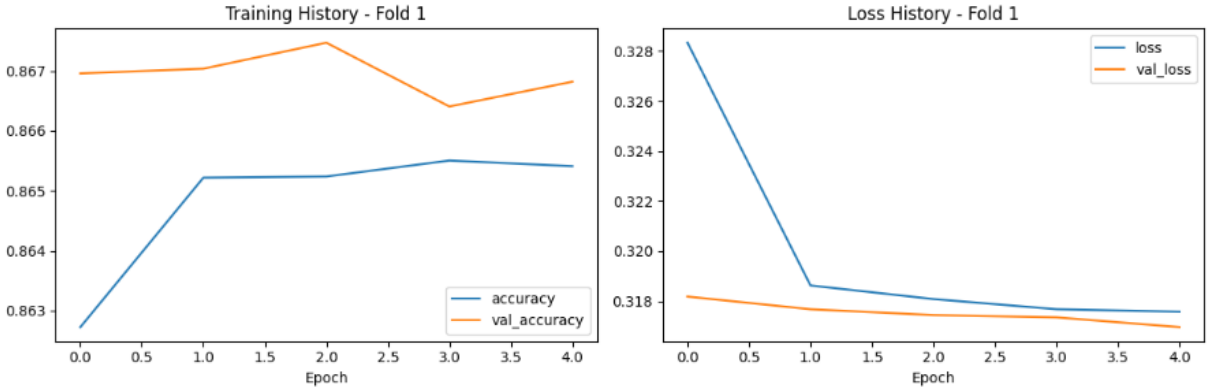

Figure 3.30- Un graphique illustrant l'accuracy et la perte pour la première validation croisée (fold 1)

Fold 2 :

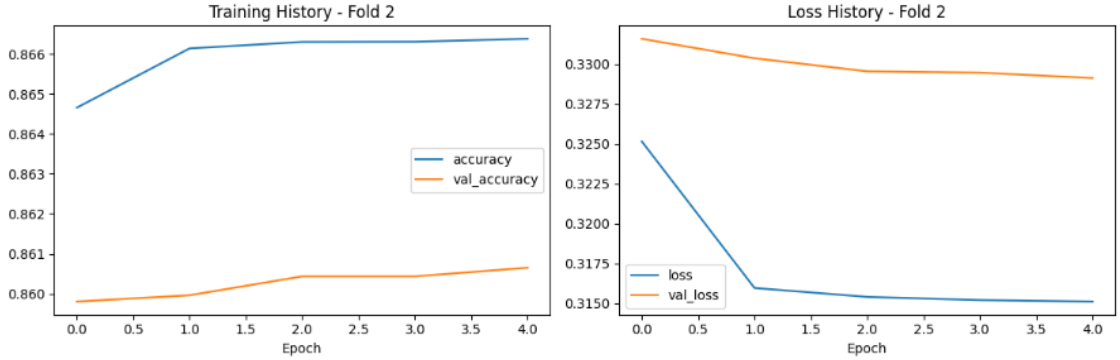

Figure 3.31- Un graphique présentant l'accuracy et la perte pour la deuxième validation croisée (fold 2)

Fold 3 :

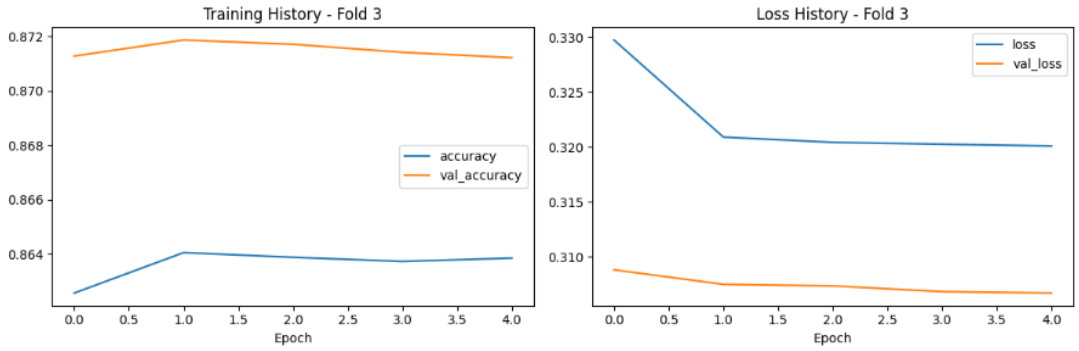

Figure 3.32- Un graphique présentant l'accuracy et la perte pour la troisième validation croisée (fold 3)

Fold 4 :

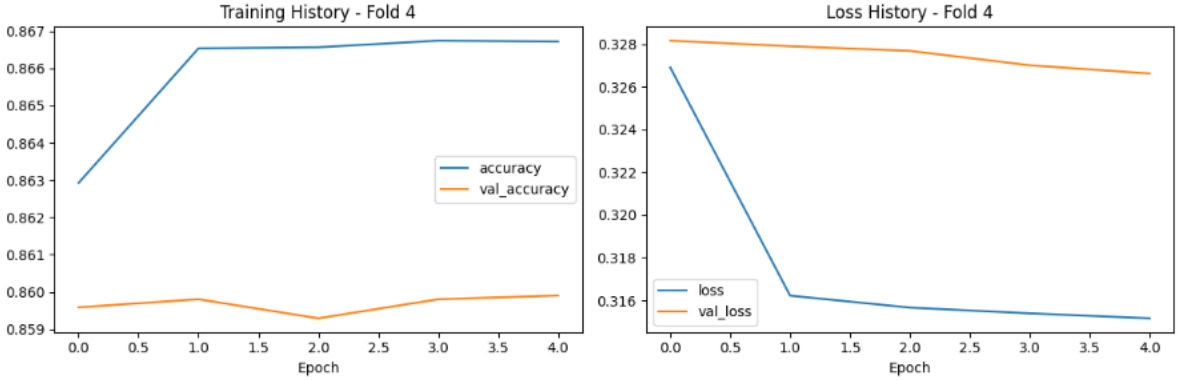

Figure 3.33- Un graphique montrant l'accuracy et la perte pour la quatrième validation croisée (fold 4)

Fold 5 :

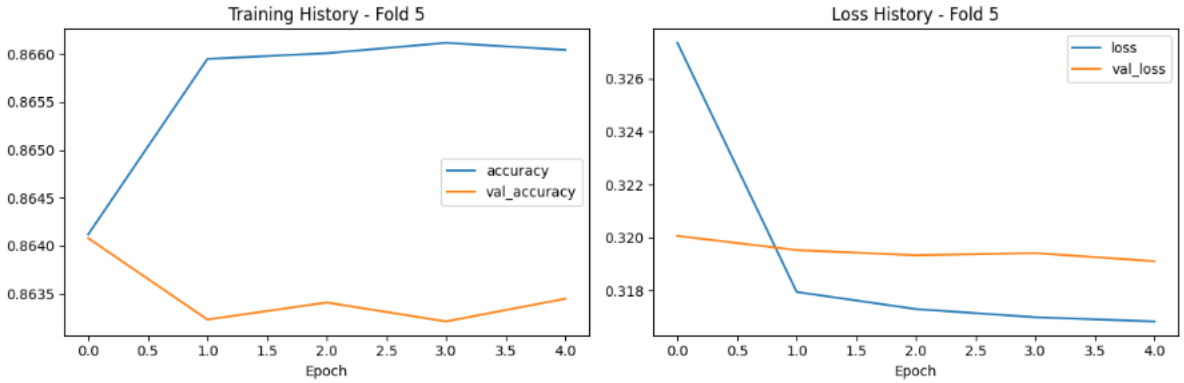

Figure 3.34- Un graphique montrant l'accuracy et la perte pour la cinquième validation croisée (fold 5)

L'approche utilisant la validation croisée K-Fold a révélé des résultats exceptionnels, où le fold 3 s'est démarqué comme étant le plus performant parmi tous les plis. Avec une précision(accuracy) de 87,14% et une perte de seulement 32,01%, le fold 3 a démontré une capacité remarquable à prédire avec précision les données de test.

- **La deuxième approche** Dans cette approche, nous avons utilisé le cas du dataset binaire équilibré:

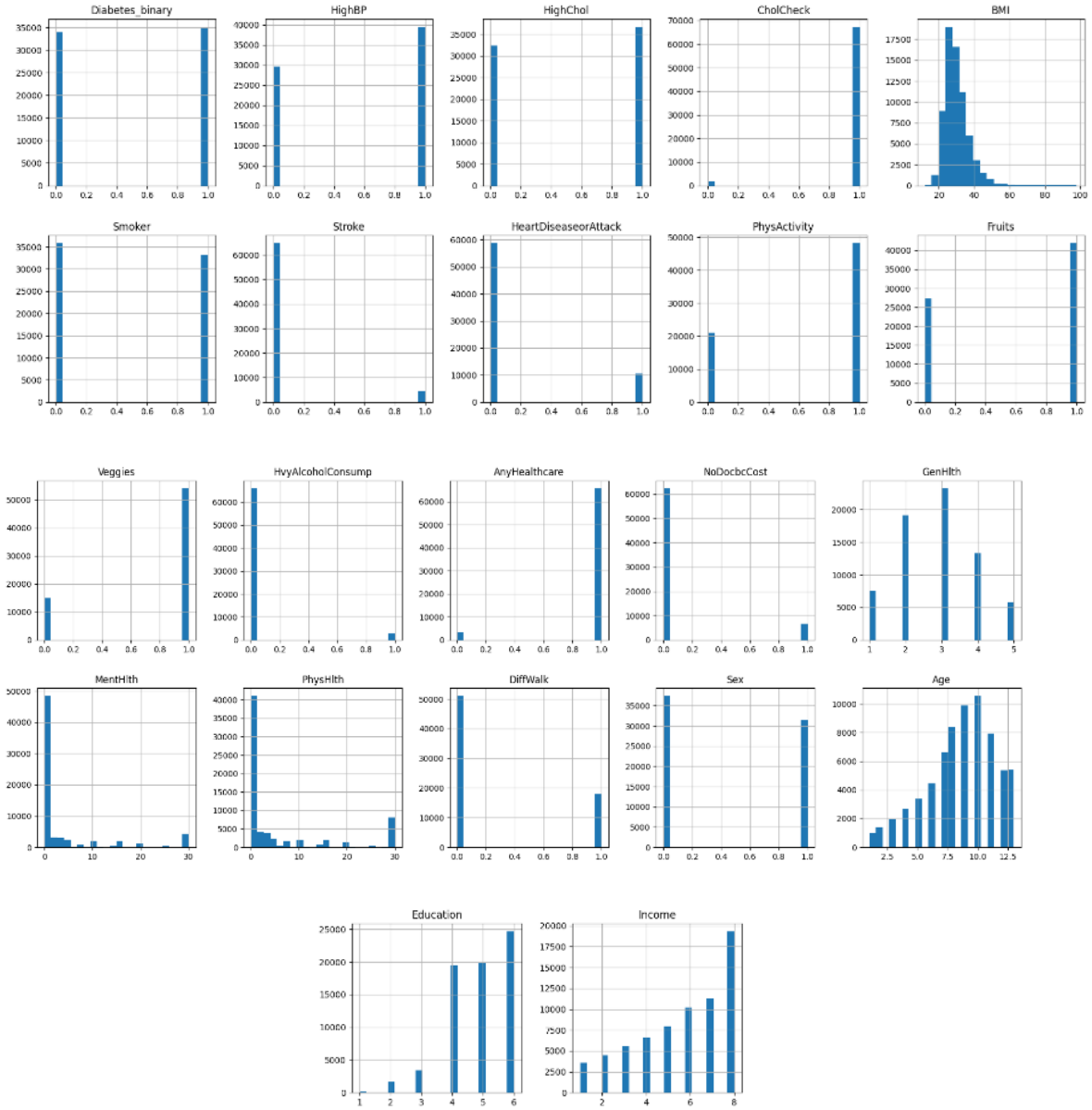

Figure 3.35- La représentation graphique des colonnes

Téléchargement des données

```
[1]: import pandas as pd
     data=pd.read_csv('/kaggle/input/diabetes-health-indicators-dataset/diabetes_binary_5050split_health_indicators_BRFSS2015.csv')
```

Figure 3.36-Télécharger les données

Manipulation des données

```
[2]: data.head()
```

[2]:

	Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth
0	0.0	1.0	0.0	1.0	26.0	0.0	0.0	0.0	1.0	0.0	...	1.0	0.0	3.0	5.0
1	0.0	1.0	1.0	1.0	26.0	1.0	1.0	0.0	0.0	1.0	...	1.0	0.0	3.0	0.0
2	0.0	0.0	0.0	1.0	26.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	1.0	0.0
3	0.0	1.0	1.0	1.0	28.0	1.0	0.0	0.0	1.0	1.0	...	1.0	0.0	3.0	0.0
4	0.0	0.0	0.0	1.0	29.0	1.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	0.0

5 rows × 22 columns

Figure 3.37- Analyser les cinq premiers records de data set

```
[3]: data.shape
```

```
[3]: (70692, 22)
```

Figure 3.38- Déterminer le nombre de colonnes et de lignes présentes dans le data set

```
[4]: data.dtypes
```

```
[4]: Diabetes_binary         float64
HighBP                  float64
HighChol                float64
CholCheck               float64
BMI                     float64
Smoker                  float64
Stroke                  float64
HeartDiseaseorAttack    float64
PhysActivity            float64
Fruits                  float64
Veggies                 float64
HvyAlcoholConsump       float64
AnyHealthcare           float64
NoDocbcCost             float64
GenHlth                 float64
MentHlth                float64
PhysHlth                float64
DiffWalk                float64
Sex                     float64
Age                     float64
Education               float64
Income                  float64
dtype: object
```

Figure 3.39- Explorer les type des tous les colonnes de data

```
[5]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70692 entries, 0 to 70691
Data columns (total 22 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Diabetes_binary       70692 non-null  float64
 1   HighBP                70692 non-null  float64
 2   HighChol              70692 non-null  float64
 3   CholCheck             70692 non-null  float64
 4   BMI                   70692 non-null  float64
 5   Smoker                70692 non-null  float64
 6   Stroke                70692 non-null  float64
 7   HeartDiseaseorAttack  70692 non-null  float64
 8   PhysActivity          70692 non-null  float64
 9   Fruits                70692 non-null  float64
 10  Veggies               70692 non-null  float64
 11  HvyAlcoholConsump     70692 non-null  float64
 12  AnyHealthcare         70692 non-null  float64
 13  NoDocbcCost           70692 non-null  float64
 14  GenHlth               70692 non-null  float64
 15  MentHlth              70692 non-null  float64
 16  PhysHlth              70692 non-null  float64
 17  DiffWalk              70692 non-null  float64
 18  Sex                   70692 non-null  float64
 19  Age                   70692 non-null  float64
 20  Education             70692 non-null  float64
 21  Income                70692 non-null  float64
dtypes: float64(22)
memory usage: 11.9 MB
```

Figure 3.40- Explorer des informations sur le data

[6]:
```
data.describe()
```

[6]:

	Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...
count	70692.000000	70692.000000	70692.000000	70692.000000	70692.000000	70692.000000	70692.000000	70692.000000	70692.000000	70692.000000	...
mean	0.500000	0.563458	0.525703	0.975259	29.856985	0.475273	0.062171	0.147810	0.703036	0.611795	...
std	0.500004	0.495960	0.499342	0.155336	7.113954	0.499392	0.241468	0.354914	0.456924	0.487345	...
min	0.000000	0.000000	0.000000	0.000000	12.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
25%	0.000000	0.000000	0.000000	1.000000	25.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
50%	0.500000	1.000000	1.000000	1.000000	29.000000	0.000000	0.000000	0.000000	1.000000	1.000000	...
75%	1.000000	1.000000	1.000000	1.000000	33.000000	1.000000	0.000000	0.000000	1.000000	1.000000	...
max	1.000000	1.000000	1.000000	1.000000	98.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...

8 rows × 22 columns

Figure 3.41-Un aperçu des statistiques numériques des valeurs présentes dans le data set

Nettoyage des données

[7]:
```
data_clean=data.drop_duplicates(inplace=True)
```

[8]:
```
data.shape
```

[8]: (69057, 22)

Figure 3.42- Supprimer les redoublant de la data

Visualisation des données

[9]:
```
import matplotlib.pyplot as plt
vc = data['Diabetes_binary'].value_counts(ascending=False)
classes = ['No Diabetes', 'Prediabetes ou Diabetes']
values = vc.values
colors = ['#1f77b4', '#FF6347']
plt.bar(classes, values, color=colors)
plt.title("Répartition des classes")
for i, value in enumerate(values):
    plt.text(i, value, str(value), ha='center', va='bottom')
plt.show()
```

[10]:
```
import matplotlib.pyplot as plt
vc = data['Diabetes_binary'].value_counts(ascending=False)
colors = ['#1f77b4', '#FF6347'] # bleu, rouge, orange, violet
plt.pie(x=vc.values, labels=['No Diabetes', 'Prediabetes ou Diabetes'], explode=[0.0, 0.02], colors=colors, autopct='%1.1f%%')
plt.legend(title="Classes", loc="center left", bbox_to_anchor=(0.9, 0.8))#, 0.5, 1))
plt.show()
```

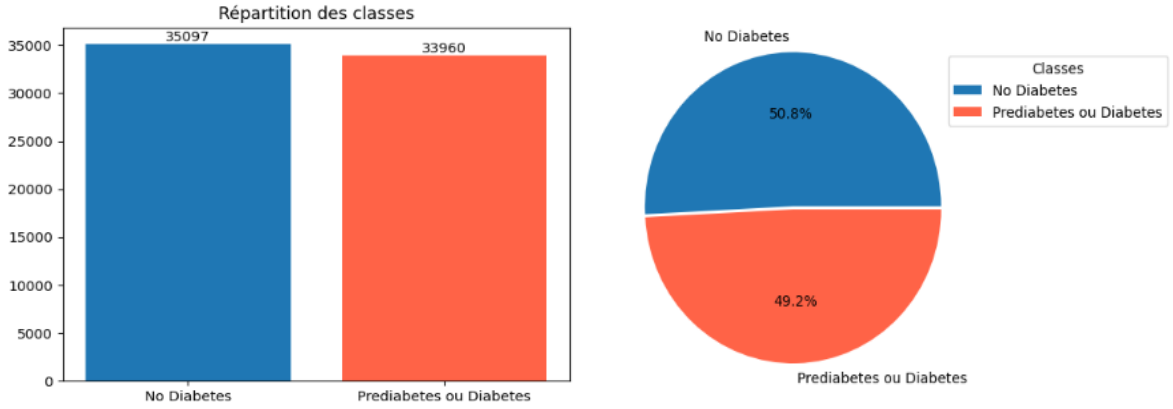

Figure 3.43- La distribution des flux de données et des types de diabète dans dataset.

Diviser le data entre X et Y

```
[11]: X=data.drop('Diabetes_binary',axis=1)
      y=data['Diabetes_binary']
```

Figure 3.44- Apercu la division des données

Description de modèle

Nous avons utilisé un modèle de réseau de neurones profonds (DNN) pour résoudre notre problème de classification binaire. Il a été construit en utilisant la bibliothèque Keras et comprend une architecture spécifique avec des couches d'entrée, des couches cachées et une couche de sortie.

Dans notre modèle, nous avons utilisé une couche d'entrée dense avec 10 neurones et une fonction d'activation ReLU. Cette couche est suivie d'une couche cachée dense avec 5 neurones et une fonction d'activation ReLU. Pour prévenir le surapprentissage, nous avons appliqué une régularisation L2 avec un paramètre de régularisation de 0.001 à ces deux couches.

Ensuite, nous avons ajouté une couche de sortie dense avec 1 neurone et une fonction d'activation sigmoïde. Cette couche nous permet d'obtenir une prédiction binaire pour notre problème de classification. Pour entraîner le modèle, nous l'avons compilé en utilisant la fonction de perte "binary_crossentropy" et l'optimiseur "adam". La fonction de perte nous aide à évaluer la qualité des prédictions du modèle, tandis que l'optimiseur ajuste les poids du modèle pour minimiser la perte. Nous avons également utilisé l'exactitude (accuracy) comme métrique de performance pour évaluer notre modèle. Cette métrique mesure la précision globale du modèle dans la prédiction des classes.

Enfin, pour obtenir une estimation fiable des performances de notre modèle, nous avons utilisé une validation croisée en divisant notre jeu de données en 5 plis (folds). À chaque itération de la validation croisée, le modèle a été entraîné sur un ensemble de plis d'entraînement et évalué sur le pli de validation restant.

```
Model: "sequential_17"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_45 (Dense)            (None, 10)                220

 dense_46 (Dense)            (None, 5)                 55

 dense_47 (Dense)            (None, 1)                 6

=================================================================
Total params: 281
Trainable params: 281
Non-trainable params: 0
_________________________________________________________________
```

Figure 3.45- L'architecture de réseau de neurones profonds (DNN) utilisée.

Résultat obtenu

Fold 1 :

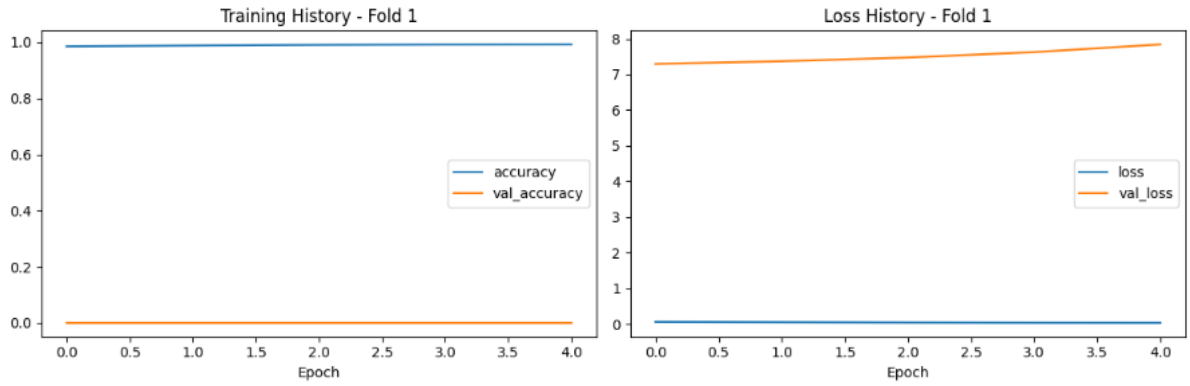

Figure 3.46- Un graphique illustrant l'accuracy et la perte pour la première validation croisée (fold 1)

Fold 2 :

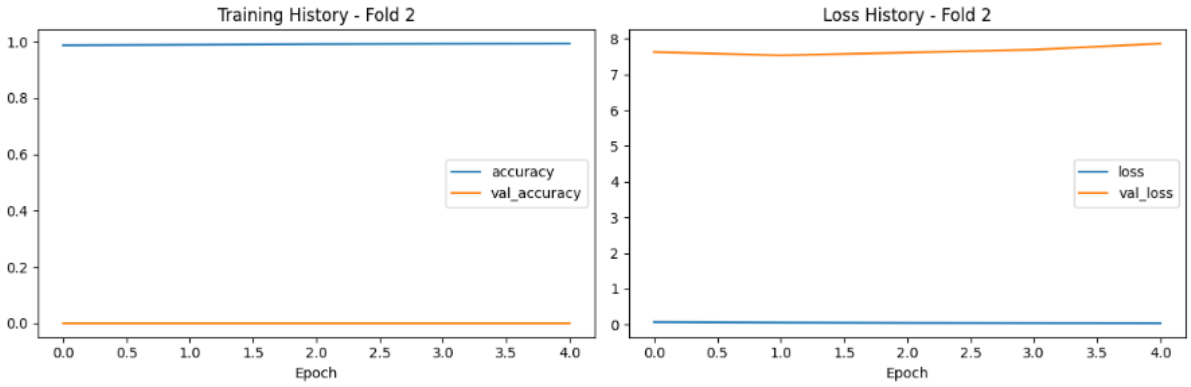

Figure 3.47- Un graphique présentant l'accuracy et la perte pour la deuxième validation croisée (fold 2)

Fold 3 :

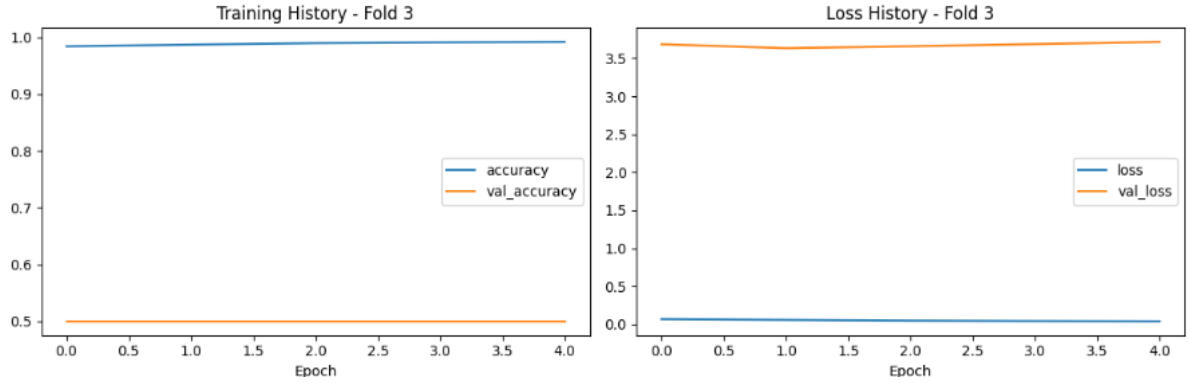

Figure 3.48- Un graphique présentant l'accuracy et la perte pour la troisième validation croisée (fold 3)

Fold 4 :

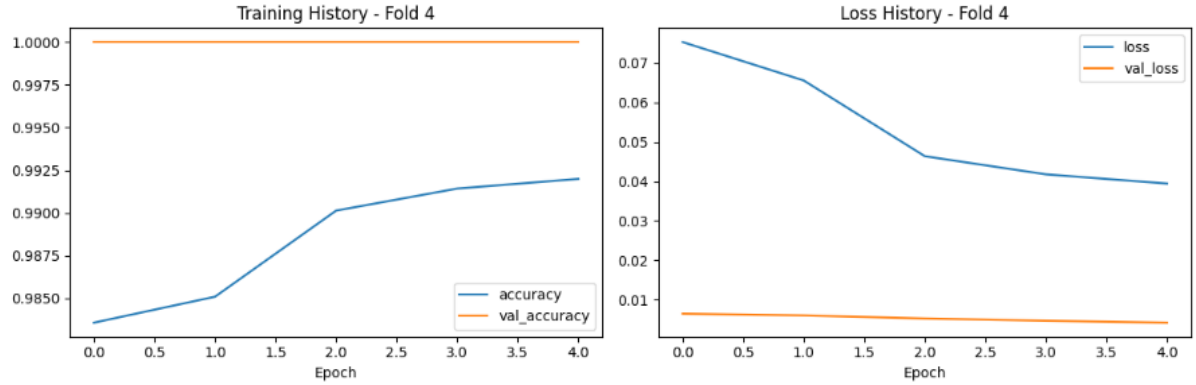

Figure 3.49- Un graphique montrant l'accuracy et la perte pour la quatrième validation croisée (fold 4)

Fold 5 :

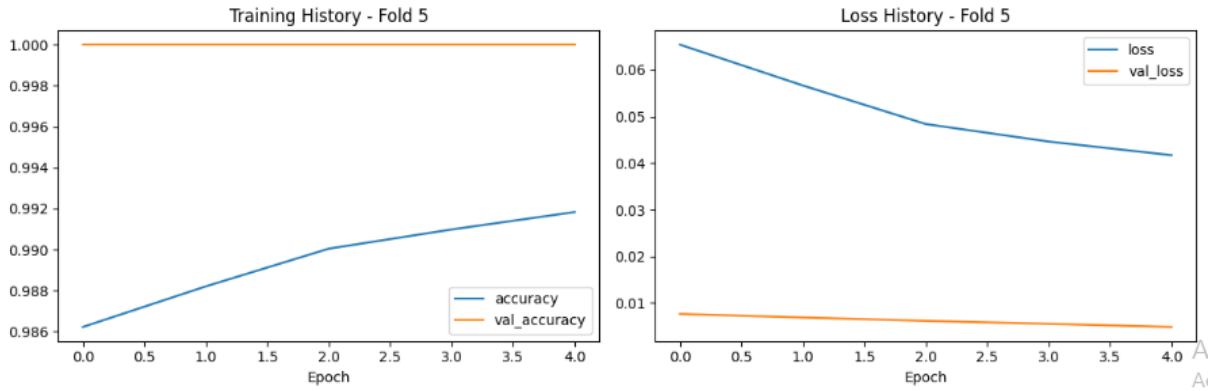

Figure 3.50- Un graphique montrant l'accuracy et la perte pour la cinquième validation croisée (fold 5)

Lors de l'utilisation de la validation croisée K-Fold, le fold 4s'est avéré être le plus performant parmi tous les plis. Cette approche a atteint une précision (accuracy) de 99.99% et une perte (loss) de seulement 3,95%. Ces résultats exceptionnels démontrent la capacité remarquable du fold 4 à prédire avec précision les données de test. Il est indéniable que le fold 4 représente le meilleur résultat de cette approche.

3.4.4.3 Comparaison des résultats

Le cas du dataset binaire équilibré présente le meilleur résultat avec une précision de 99.99%. La raison principale de cette supériorité réside dans le fait que les données sont équilibrées, c'est-à-dire qu'il y a une répartition égale entre les classes positives et négatives. Cette équilibration permet au modèle d'apprendre de manière plus équilibrée, évitant ainsi un biais envers l'une des classes. En revanche, les dataset multiclass et binaire non-équilibré peuvent présenter des déséquilibres dans les classes, ce qui peut influencer les performances du modèle. Ces déséquilibres peuvent entraîner des prédictions biaisées et des résultats moins fiables. Ainsi, en considérant la nature équilibrée des données, le cas du dataset binaire équilibré se démarque comme offrant les meilleurs résultats, par rapport aux autres fichiers dans cette comparaison.

3.4.5 Analyse comparative de différents datasets

Nous avons réalisé une comparaison approfondie de notre travail avec les études antérieures en utilisant l'ensemble de données BRFSS. Les modèles de prédiction développés pour les ensembles de données BRFSS (2014-2015-2017) ont démontré de bonnes performances, confirmant ainsi l'efficacité des approches . Cependant, notre méthode novatrice a surpassé ces modèles en termes de performance. En particulier, l'utilisation des réseaux de neurones profonds (DNN) a conduit à des résultats exceptionnels. Notre méthode a atteint une précision remarquable de 99,99%, ce qui représente les valeurs les plus élevées jamais enregistrées à ce jour . Ces performances supérieures démontrent clairement la

puissance et l'efficacité des DNN dans la prédiction des résultats sur l'ensemble de données BRFSS. Ces résultats sont prometteurs et soulignent l'importance de notre approche pour améliorer les prédictions dans divers domaines d'application. En somme, notre méthode représente une avancée significative dans le domaine des prédictions basées sur l'ensemble de données BRFSS (voir tableau 3.5).

Tableau 3.5- Comparaison de la méthode proposée avec les études existantes utilisant le dataset BRFSS [50].

Dataset	Méthode	Accuracy(%)
BRFSS-2014	NN	82.4
BRFSS-2017	RF	86.8
BRFSS-2015	KNN	98.36
BRFSS-2015	DNN	99.99

Nous avons également effectué une comparaison avec d'autres datasets , et les résultats ont confirmé nos attentes. Notre méthode s'est avérée être la plus précise et la plus efficace parmi toutes les alternatives évaluées. Pour mener cette comparaison, nous avons sélectionné plusieurs datasets pertinents dans le domaine d'étude.

En analysant les résultats obtenus, nous avons constaté que notre méthode surpassait les autres en termes de précision et d'efficacité(voir tableau 3.6).

Tableau 3.6- Comparaison de la méthode proposée avec les études existantes qui ont utilisé d'autres datasets[50].

Dataset	Méthode	Accuracy(%)
PIDD	RF	77.21
PIDD	LR,SVM	78.85, 77.71
Private	RF	80.84
PIDD	RF	88.31
PIDD	SVM	94.44
Private	LR	96.02
BRFSS	DNN	99.99

3.5 Conclusion

Le chapitre visait à explorer diverses approches pour prédire le diabète. Dans cette optique, nous avons abordé plusieurs aspects importants liés à cette problématique.L'introduction a mis en évidence l'importance de la prédiction du diabète et l'objectif était de trouver des solutions fiables pour anticiper et prévenir son développement, en raison de ses conséquences significatives sur la santé. Nous avons ensuite présenté les outils et l'environnement de

développement utilisés tout au long de notre étude. Cela incluait les langages de programmation et les bibliothèques qui nous ont permis de mettre en place nos modèles de prédiction.La base de données était essentielle dans notre étude, fournissant des informations précises sur les caractéristiques et les variables liées au diabète, garantissant ainsi des résultats fiables.Nous avons utilisé la validation croisée K-Fold pour évaluer nos modèles, et le fold 4 a été le plus performant avec une précision de 99.99% et une perte de 3,95%.

Notre étude a souligné l'importance de l'équilibrage des données dans la prédiction du diabète le cas du dataset binaire équilibré a présenté une répartition équilibrée des classes, conduisant à une précision exceptionnelle de 99.99% grâce à la validation croisée K-Fold. Ces résultats ouvrent de nouvelles perspectives pour des analyses futures dans ce domaine.

Conclusion Générale

En conclusion, ce mémoire d'étude intitulé "Classification du diabète à l'aide des algorithmes du Machine Learning" a examiné en détail l'application du deep learning dans la détection et la prédiction du diabète. Tout d'abord, nous avons posé les bases en présentant l'intelligence artificielle dans son ensemble, en définissant ses concepts et en retracant son évolution historique. Nous avons également exploré les différentes applications pratiques de l'intelligence artificielle dans la vie quotidienne, mettant en évidence ses avantages, ses inconvénients et ses limites.

Par la suite, nous avons approfondi notre compréhension du diabète, en commençant par une introduction détaillée définissant la maladie, ses différents types et les facteurs de risque qui y sont associés. Nous avons examiné la prévalence mondiale du diabète et mis en évidence l'importance d'un diagnostic précoce et d'un traitement adéquat pour minimiser les complications. Nous avons également abordé des aspects tels que la prévention du diabète et la gestion du mode de vie, en soulignant l'importance d'adopter des habitudes saines et d'adopter une approche holistique pour maintenir une bonne santé. Nous avons également examiné l'impact psychologique et social du diabète, mettant en évidence les défis auxquels sont confrontées les personnes atteintes de cette maladie chronique.

Dans la dernière partie de notre mémoire, nous nous sommes concentrés sur l'application du deep learning dans la détection et la prédiction du diabète. Nous avons présenté les outils et l'environnement de développement utilisés, ainsi que la base de données spécifique sur laquelle nous avons travaillé en utilisant la validation croisée K-Fold, nous avons obtenu des résultats prometteurs, en particulier avec le cas binaire qui a montré une répartition égale des classes et une précision exceptionnelle de 99.99%.

En résumé, ce mémoire d'étude a confirmé l'efficacité du deep learning dans la classification du diabète. Les avancées technologiques de l'intelligence artificielle ouvrent de nouvelles perspectives pour améliorer la détection précoce, la prise en charge et la prévention du diabète. Ce domaine émergent offre de vastes opportunités pour améliorer les soins de santé et la qualité de vie des personnes atteintes de diabète à l'avenir.

Bibliographie

[1] A. Turing, “Numérique (-numerique-296-) brève introduction au monde de l’intelligence artificielle”.

[2] M. Villani, “Qu ’ Est-Ce Que L ’ Intelligence Artificielle ?,” .

[3] I. Artificielle, “Intelligence artificielle : tout ce qu ’ il faut savoir Définition : Intelligence artificielle,”.

[4] M. Dadi, “Ms ELN Benmansour + Bouzouina + Oudinat PDF,” .

[5] “Chapitre 13 L’intelligence artificielle (IA).”

[6] J. Mccarthy, “L ’ Intelligence Artificielle , ses Avantages et PHILOSOPHIE DE L ’ IA :,” 2022.

[7] I. Artificielle, “Intelligence arti fi cielle : jusqu ’ où Quelles sont les limites de l ’ Intelligence Artificielle ? Intelligence Artificielle : quelles utilisations,” .

[8] D. Didaquest, “Apprentissage automatique Traduction Définition,” .

[9] M. Learning, D. Science, D. Science, and M. Learning, “Tout savoir sur le machine learning,”.

[10] F. C. Carri, A. Prendre, and M. Learning, “Les algorithmes de Machine Learning,”.

[11] psychomedia, “Apprentissage profond,” p. http://www.psychomedia.qc.ca/lexique/definition/ap, 2016, [Online]. Available: http://www.psychomedia.qc.ca/lexique/definition/apprentissage-profond

[12] A. M. Learning, “Apprentissage en profondeur et apprentissage automatique dans Azure Machine Learning Techniques d ’ apprentissage profond vs apprentissage automatique,”.

[13] L. Bastien, “Réseau de neurones artificiels : qu ’ est-ce que c ’ est et à quoi ça sert ?,”.

[14] “PAR El Mahdi BRAKNI RÉSEAUX DE NEURONES ARTIFICIELS APPLIQUÉS À LA MÉTHODE ÉLECTROMAGNÉTIQUE

TRANSITOIRE InfiniTEM," 2011.

[15] D. Learning, "e

n Deep Learning,".

[16] D. C. E. Poste, "validation croisée ? modèles d ' apprentissage,".

[17] S. Kumar, "Comprendre 8 types de validation crois é e,".

[18] L. D. Learning, "Deep Learning : avantages et inconvénients," pp.

[19] E. Suisse and B. Crottaz, "Diabete," .

[20] L. E. De and I. Comment, "Qu'est-ce que le diabète ? 03," pp.

[21] J. E. R. Inform and L. A. Newsletter, "Qu'est-ce que le diabète ?," .

[22] C. Dugas, "Diabète de type 1," *L'inclusion en éducation Phys.*, pp. doi: 10.2307/j.ctt1f1hd5q.22.

[23] "Ce qui me met à risque," .

[24] C. Manicot, "Les chiffres du SIDA.," *Rev. l"infirmiere*, vol. 41, no. 10, .

[25] D. Collegemc, U. H. N. Foundation, and S. Life, "En savoir plus sur le programme de diabète de Toronto Rehab myDiabetes Programme 12 semaines du programme THRiVE Welcome to Diabetes College," .

[26] Fédération Française des Diabétiques, "Les traitements du diabète," pp.

[27] A. Doutriaux, "Mesurer l ' HbA1c chez un patient diabétique," .

[28] C. Tran, M. Boulvain, and J. Philippe, "Prise en charge du diabète gestationnel: Nouvelles connaissances et perspectives futures," *Rev. Med. Suisse*, vol. 7, no. 298, , 2011.

[29] "Comment prévenir le diabète de type 2 Resources additionnelles," pp.

[30] T. Accepter and T. Refuser, "Alimentation , activité physique et diabète INTERVIEW DE MÉLANIE MERCIER ," vol. 2018, 2018.

[31] G. Devaux, "10 solutions faciles pour faire baisser la glycémie de façon naturelle," .

[32] P. Valerie, “L ’ influence du diabète sur votre santé mentale,” pp.

[33] F. Maladies, “Diabète et vie sociale condition de préparer son départ et de respecter quelques,” 2023.

[34] P. H. Jalini, “Une thérapie génique soigne (presque) le diabète de type 1 chez la souris,” pp.

[35] J. E. R. Inform and L. A. Newsletter, “THÉRAPIE CELLULAIRE : DES CELLULES SOUCHES DU CERVEAU GREFFÉES SUR UN,” pp.

[36] N. Contacter and A. I. S. Faq, “Surveillance de la glycémie : un guide complet,” pp.

[37] A. Martinez-millana *et al.*, “Artificial Intelligence for Diabetes Management and Decision Support : Literature Review,”doi: 10.2196/10775.

[38] “Pancréas artificiel Approches,”.

[39] La redaction, “Python : définition et utilisation de ce langage informatique,” 2020, [Online]. Available: https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1445304-python-definition-et-utilisation-de-ce-langage-informatique/

[40] A. Goldbloom, A. Goldbloom, and A. Alphabet, “Kaggle,”.

[41] K. Banachewicz, K. Banachewicz, L. Massaron, A. Goldbloom, L. Massaron, and A. Goldbloom, “The Kaggle Book Science,”.

[42] P. Libraries and D. Science, “Top 20 Python Libraries for Data Science for 2023,”.

[43] F. A. Q. S. U. R. L. Ia, “AI writing assistance Le Blog De L ’ IA,” .

[44] “https://www.editions-eni.fr/open/mediabook.aspx?idR=f6e7a7353a3574180124387fa03fdc1c&fbclid=IwAR0WVYgIRT5xz_YONJ0v8o7xARTxLksAN2Px-t5Mzwwx0Yo-LJg_j2DjeKc 1/1,” p. 7353, 2023.

[45] D. Learning and L. E. S. Bases, “Fonction de Loss – Laquelle choisir – Meilleur Tutoriel Classe vs label,”.

[46] L. E. S. Bases, “Fonction d ’ activation , comment ça marche ? – Une explication simple Qu ’ est-ce qu ’ une fonction d ’ activation ? Changer de point de vue,”.

[47] M. Riva, D. Learning, D. Learning, and B. Norm, “Normalisation par lots dans les réseaux de neurones convolutifs,”.

[48] R. Datafranca, “Optimisation Adam LES 101 MOTS DE L ’ INTELLIGENCE ARTIFICIELLE,” .

[49] S. G. Descent, “10 optimiseurs d ’ apprentissage automatique célèbres,”.

[50] Z. Ullah *et al.*, “Detecting High-Risk Factors and Early Diagnosis of Diabetes Using Machine Learning Methods,” *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/2557795.