الجمهورية الجزَائِرية الديمقرَاطية الشعبية
RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
وزَارة التعليم العَالي و البحث العلمي
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE
جَامعة عمَار ـتليجي بِالأَغوَاط
UNIVERSITE AMAR TELIDJI LAGHOUAT

FACULTE DES SCIENCES
DEPARTEMENT D'INFORMATIQUE

Mémoire de Master

Domaine : Mathématiques et Informatique

Filière : Informatique

Option : Système d'information décisionnel

Réalisé Par :

Bouaziz Meriem **Sebaa Sarah Djihane**

Thème :

La prédiction du diabète en utilisant les techniques d'intelligence artificielle

---

Soutenue publiquement le 02-06-2022 devant le membre de jury composé de :

Mr. Laradj Chellama	M.A.(A)	Président
Mr. Guellouma Younes	M.C.(A)	Examinateur
Mr. Bouakkaz Mustapha	M.C.(A)	Encadreur

Année universitaire 2021/2022

REMERCIEMENT

Nous remercions Dieu le tout puissant de nous avoir donné la volonté, La patience et la santé pour accomplir ce travail.

Nous tenons à remercier notre encadreur Dr. Bouakkaz Mustapha pour ses conseils et ses recommandations.

Nos vifs remerciements aussi pour les membres du jury qui ont accepté d'examiner et d'évaluer notre travail.

Nos dédicaces sont également adressés au Professeur Lagraa Nasreddine pour ses précieux conseils et orientations qui nous ont aidé à réaliser ce travail.

Nous exprimons aussi notre gratitude pour Dr. Bousbaa Fatima pour tous ses encouragements et son aide dans la réalisation de ce travail.

Sans oublier de remercier tous les enseignants du département d'informatique et à tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail.

DÉDICACES

Je commence à dédier ce travail aux deux personnes les plus chers à mon cœur mes parent, merci d'avoir fait de moi la femme que je suis aujourd'hui.

Je tiens à dédier ce travail aussi pour mes chères sœurs d'avoir toujours été là pour moi et qui ont toujours été ma fierté et ma source d'inspiration, sans oublier mes deux neveux chérie Mohamed et Lokmane et ma future niéce.

Je tiens surtout à dédier ce travail à mon binôme et mon amie Meriem de m'avoir accompagné durant toutes ces cinq ans d'université et à toute sa famille.

A toute ma famille et mes amis qui méritent d'y trouver leur nom.

Sebaa Sarah Djihane

DÉDICACES

Je dédie ce travail à toute ma famille, Particulièrement à mes chers parents, qui m'avez toujours soutenus et encouragés durant ces années d'études, je vous remercie énormément pour tous ce que vous m'avez appris dans cette vie.

A mes deux chers frères houssem eddine et amine, merci d'être à mes côtés quand j'avais besoin de vos avis, vous êtes ma source de motivation et d'inspiration et à mon petit bout de sucre, ma nièce Iline.

Mes dédicaces sont également adressées à mon binôme et ma chère amie Sarah, c'était toujours un plaisir de travailler avec vous, Merci pour tous les souvenirs que nous avons passé pendant ce cycle universitaire, je vous souhaite un avenir plein de succès et que vous réalisez vos souhaits.

Bouaziz Meriem

RÉSUMÉ

Le diabète est une maladie chronique qui menace la vie de milliers de personnes dans le monde. En raison des graves complications qui peuvent être engendrées par cette maladie, un diagnostic précoce est alors nécessaire. A travers ce travail de master nous nous intéressons à l'utilisation des techniques de machine learning (ML) et deep learning (DL) pour la prédiction du diabète de type 2 sur la base de données PIma Indian Diabetes afin de réduire les risques de l'atteindre. Pour cela nous avons commencé par une étude comparative de deux algorithmes de machine learning qui sont : support vector machine (SVM) et Random Forest (RF) que nous avons par la suite choisi le modèle RF avec 81% d'accuracy pour l'étape de Sélection d'attribut avec l'outil Weka. Après avoir sélectionné les attributs les plus influents nous les avons utilisés pour construire le modèle de deep learning avec qui nous avons eu un score d'accuracy de 83%.

Mots clés : ML, DL, Base de données PIma Indian Diabetes, SVM, RF.

ABSTRACT

Diabetes is a chronic disease that threatens the lives of thousands of people around the world. Because of the serious complications that can be caused by this disease, an early diagnosis is necessary. Through this master's work we are interested in the use of machine learning (ML) and deep learning (DL) techniques for the prediction of diabetes type 2 on the PIma Indian Diabetes database in order to reduce the risks of reaching it. For this we endeavored a comparative study of two machine learning algorithms that are support vector machine (SVM) and random forest (RF) , then we chose the RF model with 81% accuracy for the Attribute Selection step with the Weka tool. After selecting the most influential attributes we used them to build the deep learning model with which we had an accuracy score of 83%.

Keywords : ML, DL, PIma Indian Diabetes data base, SVM, RF.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS

— OMS : Organisation Mondiale de la Santé.
— FIGO : la Fédération Internationale de Gynécologie et d'Obstétrique.
— DT1 : Diabète de type 1.
— DT2 : Diabète de type 2.
— IA : Intelligence Artificielle.
— ML : Machine Learning.
— DL : Deep Learning.
— API : Application Programming Interface.
— SVM : Support Vector Machine
— RF : Random Forest

INTRODUCTION GÉNÉRALE

Aujourd'hui, le développement du secteur médical ne se dément pas, ce qui est favorable à la santé publique et qui permet de réduire significativement le taux de mortalité dans le monde. Malgré ces avancées médicales, les maladies non transmissibles causent de réels problèmes et menacent la vie de milliers de personnes dans le monde entier, nous mentionnons le diabète.

Le diabète est une maladie dans laquelle les patients souffrent de problèmes de glycémie en raison d'une production insuffisante d'insuline ou par l'incapacité de l'organisme à utiliser efficacement l'insuline produite, Le manque de cette hormone dû à un dysfonctionnement du pancréas peut entraîner plusieurs complications comme : l'insuffisance rénale et rétinienne, destruction pathologique de cellules bêta pancréatiques, dysfonctionnement cardiovasculaire, cerveau etc.

Cependant, il n'y a pas de remède à long terme pour le diabète, mais il peut être contrôlé et prévenu si une prédiction précoce est possible avec précision.

Les chercheurs ont appliqué divers processus pour prédire le diabète tels que : les algorithmes d'apprentissage automatique (SVM, arbre de décision, KNN ...), les approches ensemblistes ainsi que les techniques des réseaux de neurones. L'objectif de notre étude applicative est d'utiliser la discipline d'intelligence artificielle (IA) et ses différentes techniques (l'apprentissage automatique et l'apprentissage profond) pour élaborer un modèle de prédiction du diabète de type 2 afin de diminuer les complications engendrées par cette maladie. Dans la partie d'apprentissage automatique, nous avons implémenté deux algorithmes (SVM, Random Forest) ensuite, une étude comparative était établie sur les résultats obtenus, Dans l'apprentissage profond nous avons élaboré un modèle optimisé basé sur la sélection des attributs. Ce travail est organisé en quatre chapitres comme suit :

— Le 1er chapitre est dédié aux concepts théoriques de la maladie, il donne un aperçu sur le diabète, ses différentes classes, les symptômes ainsi que le diagnostic et le traitement de la maladie et enfin quelques précautions pour éviter le diabète.

— Le 2éme chapitre représente l'intelligence artificielle, ses types ainsi que ses branches, ce

chapitre explique aussi la notion d'apprentissage automatique et d'apprentissage profond et les techniques que nous avons utilisées.

— Le 3éme chapitre est consacré aux travaux récents des chercheurs sur le diabète en utilisant les techniques de l'intelligence artificielle.

— Le dernier chapitre présente en détail notre contribution, en commençant tout d'abord par définir l'environnement dans lequel nous avons travaillé (outils de programmation, librairies, base d'apprentissage utilisée). Ensuite, nous avons présenté en détail le processus de l'implémentation de notre travail en expliquant chaque étape avec une discussion des résultats obtenus.

Le travail est clôturé par une conclusion générale qui résume notre contribution et indique les perspectives et les travaux futurs.

CHAPITRE 1

PRÉSENTATION DU DIABÉTE

1.1 Introduction

Le diabète est une maladie chronique non contagieuse qui menace la vie des millions de personnes dans le monde. Il se définit lorsque le taux de glycémie d'une personne est élevé à cause du pancréas qui ne produit pas suffisamment d'insuline ou lorsque le corps ne consomme pas efficacement l'insuline qu'il produit. L'Organisation Mondiale de la Santé (OMS) l'identifie comme étant une épidémie car il représente la quatrième cause de décès [1].

Comme le montre la Figure 1.1 (des statistiques sur la mortalité qui ont été faites pour la région Moyen-Orient et Afrique du Nord en 2019).

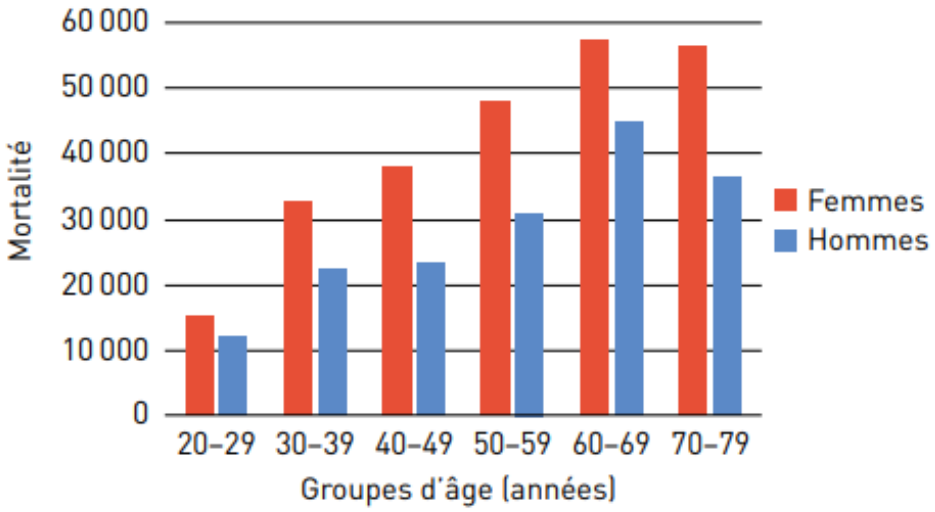

FIGURE 1.1 – Mortalité due au diabète par âge et par sexe, Région MoyenOrient et Afrique du Nord, 2019 [2]

1.2 Définition

Le diabète sucré ou simplement le diabète, est une maladie chronique qui se définie par l'élévation de la concentration du glucose dans le sang dû à l'insuffisance d'insuline ou à sa mal consommation par l'organisme.

L'insuline est une hormone régulatrice du glucose dans le sang produite par le pancréas, elle permet de transporter le glucose circulé dans le sang aux cellules du corps pour enfin produire de l'énergie et diminuer le taux de sucre dans le sang comme il est montré dans la figure 1.2. Une personne diabétique souffre d'une quantité de glucose dans le sang trop élevé. Si cette hyperglycémie n'est pas bien contrôlée et traitée elle peut engendrer plusieurs complications aux différents organes du corps, certains d'entre eux peuvent même être mortel comme les maladies cardiovasculaire, néphropathie, neuropathie. néamoins, une bonne prise en charge permet d'éviter ou tout au moins retardée ces complications [2].

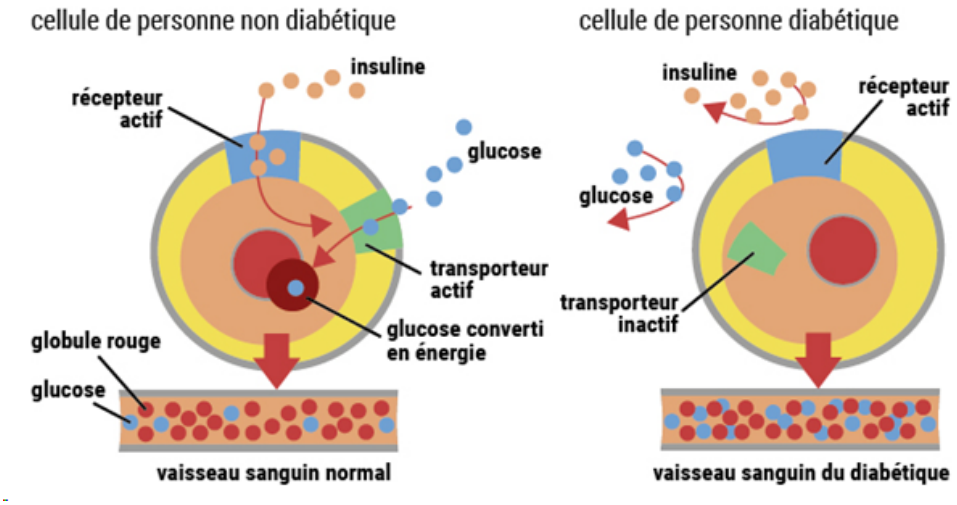

FIGURE 1.2 – Cellule d'une personne diabétique et non diabétique [3].

1.3 Classification du diabéte

1.3.1 Diabète de type 1

Le diabète est de type 1 lorsque l'organisme ne produit pas ou très peu d'insuline. il se produit à cause d'une attaque auto-immune c'est-à-dire les cellules béta qui produit de l'insuline dans le pancréas sont détruites par le système immunitaire. Les causes de ce processus destructeur sont encore inconnues, les personnes atteintes par ce type de diabète sont généralement des enfants ou des jeunes et il est impossible de le prévenir [1].

1.3.2 Diabète de type 2

Le diabète de type 2 est le plus courant, il touche environ 90 % de la population diabétiques totale, précédemment appelé diabète non insulino-dépendant ou diabète de la maturité, cependant ce type de diabète touche de plus en plus les jeunes à cause de la malnutrition et le manque d'activité physique.

Il se caractérise par un taux de sucre trop élevé dans le sang (hyperglycémie), cette anomalie est causée par une mauvaise utilisation de l'insuline par l'organisme. Ces symptômes sont similaires au diabète de type 1 mais beaucoup moins marqués et dans certains cas il peut être complètement asymptomatique, le patient peut être diagnostiqué plusieurs années après son apparition. la figure 1.3 montre la différence entre la consommation d'insuline par une personne normal et par les personnes atteintes du diabète de type 1 et 2 [1].

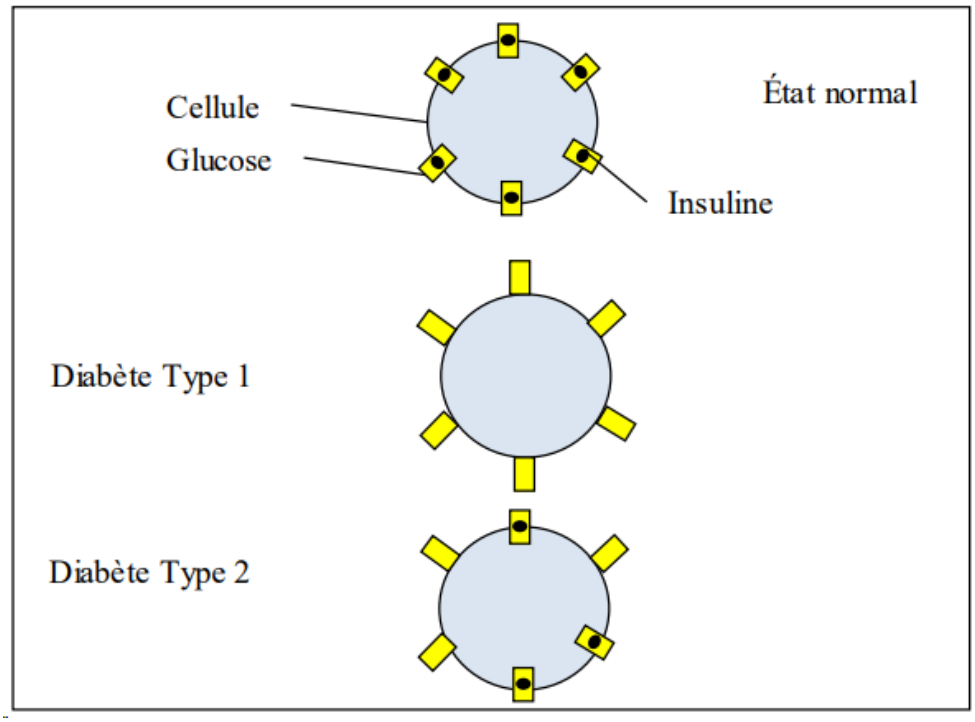

FIGURE 1.3 – Fonctionnement de l'insuline [4].

1.3.3 Diabète gestationnel

Selon l'OMS et la Fédération Internationale de Gynécologie et d'Obstétrique (FIGO), l'hyperglycémie durant la grossesse peut être classée en diabète gestationnel c'est à dire une élévation du taux du glucose dans le sang par rapport au valeurs normal mais inférieur à celle conduisant au diabète. Il touche une femme enceinte sur 10 et il s'applique aux femmes enceintes qui ont déjà développé un diabète ou une hyperglycémie, l'OMS estime que (75 à 90 %) des cas

d'hyperglycémie durant la grossesse sont dus au diabète gestationnel [2].

1.4 Les symptômes du diabète

Les symptômes peuvent se différer d'un type de diabète à un autre mais les cas les plus fréquents sont les suivants :

- La sensation de soif.
- La fatigue.
- Une vision floue.
- Soudaine perte de poids.
- Augmentation de la faim.
- Une lente cicatrisation des blessures.
- La polyurie.
- Engourdissement dans les mains et les pieds [5].

1.5 Le Diagnostic du diabéte

Il existe plusieurs tests qui permettent d'identifier le diabète, nous citons :

- **Test sanguin :**

  Cet examen permet de présenter le taux de sucre en mesurant la glycémie ou le taux de sucre sanguin, ce test est révélateur de la bonne santé de divers organes (foie, reins et pancréas). La personne est diabétique si la glycémie est supérieure ou égale à 1,26 g/l à jeun ou supérieure ou égale à 2 g/l après le repas (glycémie postprandiale) [6].

- **La glycosurie, rechercher l'albumine et le sucre dans les urines :**

  La glycosurie est un examen très simple et rapide qui consiste à chercher la présence de glucose et l'albumine (protéine) dans les urines, ce test est fait à l'aide des bandelettes qui sont trempées directement dans les urines détectant la présence de sucre, il est moins fiable puisqu'il faut avoir 1.80 g/l pour que la bandelette soit positive [6].

- **Un test pour détecter le diabète gestationnel :**

  appelé L'HGPO (hyperglycémie provoquée par voie orale), Il s'agit d'un test sanguin recommandé pour détecter un diabète gestationnel chez la femme enceinte entre 24 et 28 semaines de grossesse. Selon l'OMS, La personne est atteinte par le diabéte gestationnel si une seule mesure est supérieure ou égale : de 0,92 g/l à jeun, 1,80 g/l une heure après absorption du glucose et 1,53 g/l deux heures plus tard [6].

- **Les glycémies capillaires pour mesurer quotidiennement sa glycémie :**

  Ce test est connu chez les personnes diabétiques, il consiste à analyser une gouttelette de sang au bout de doigt par un appareil (glucomètre), ≪Un résultat massivement élevé

n'est pas trompeur. Mais cela ne suffit pas et cet examen doit être vérifié par une prise de sang», précise le Pr Altman [6].

— **L'hémoglobine glyquée :**

La mesure du taux sanguin de ce que l'on appelle "l'hémoglobine glyquée" donne une idée de la glycémie moyenne au cours des trois derniers mois, ce test est prescrit aux diabétiques pour surveiller l'évolution du diabète, la meilleure mesure se situe en dessous de 7 % [6].

1.6 Les causes du diabète

les causes du diabète sont nombreuses et se diffèrent d'un type à un autre.

1.6.1 Les causes du diabète de type 1

Le rôle des facteurs génétiques et environnementaux que nous venons mentionner ne sont qu'au stade d'hypothèses et d'observations. Les causes précises du DT1 sont encore nébuleuses. 1- Facteurs génétiques : Ce type de diabète pourrait trouver sa source dans des causes de nature héréditaires, c'est-à-dire lorsque les parents portent le gène (diabétiques). 2- Facteurs externes (environnementaux) :

— Les infections virales ou bactériennes.
— La nature de l'alimentation.
— Le stress psychologique.
— Les maladies qui touchent le pancréas (inflammation, kyste, cancer, etc.) [7].

1.6.2 Les causes du diabète de type 2

La cause précise de ce type de diabète n'a pas encore été scientifiquement établie. Mais un ensemble de facteurs de risque ont été identifiés, notamment :

— Le surpoids et l'obésité.
— La sédentarité et le manque d'activité physique.
— L'âge, le risque est plus élevé après 45 ans.
— Changement hormonal dans la période de la puberté.
— Antécédents familiaux.
— L'utilisation de certains médicaments [7].

1.6.3 Les causes du diabète gestationnel

La grossesse augmente les besoins de la mère en insuline d'un facteur de 2 à 3, un diabète gestationnel peut être développé si :

— Avoir des antécédents familiaux qui ont déjà eu un diabète
— La femme a eu un diabète gestationnel lors d'une grossesse précédente.
— Naissance d'un bébé de poids élevé(4kg).
— Avoir déjà pris un traitement de longue durée avec des corticostéroïdes (cortisone).
— âge supérieur à 35 ans.
— Etre en surpoid ou en obésité.[8]

1.7 Les complications du diabète

Le diabète peut engendrer des complications à court terme ainsi qu' à long terme. Elles peuvent toucher toutes les parties du corps, qui peuvent être très graves et affecter la vie des gens si elles ne sont pas prises en charge d'une manière efficace. La figure suivante montre les complications engendrées par le diabète.

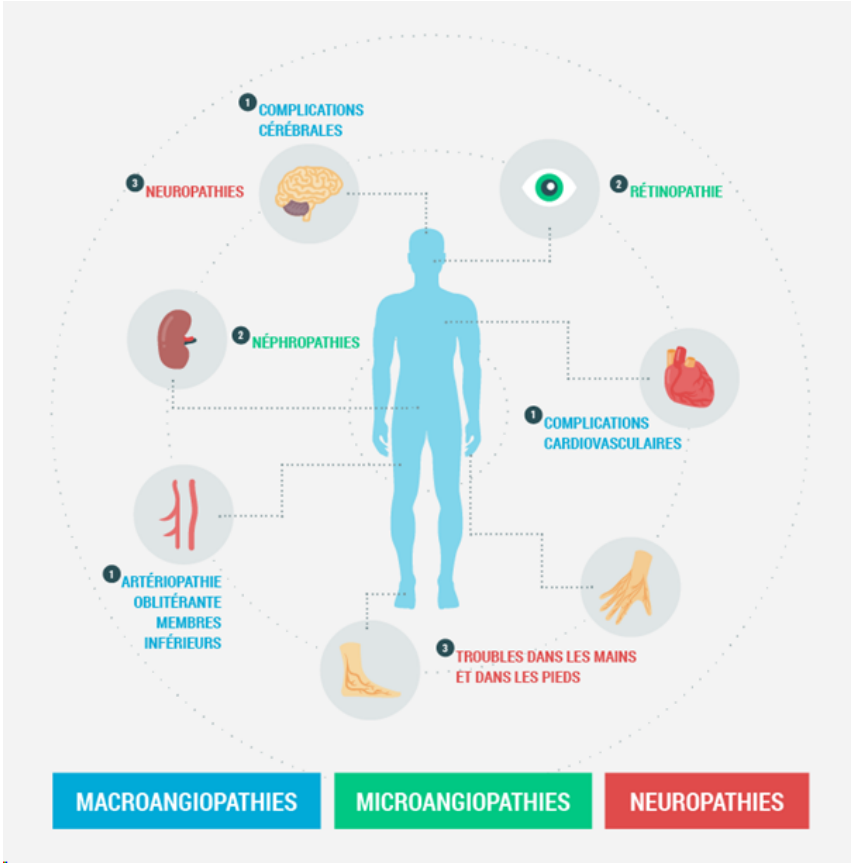

FIGURE 1.4 – Diabètes et complications [9].

1- Complications de la grossesse

Les études ont montré que les femmes qui ont eu le diabète durant la grossesse ont un grand risque d'avoir le diabète de type 2 dans les 15 ans suivant la grossesse de référence. Ces complications ne s'arrêtent pas qu'à la maman mais également à son nouveau-né qui a un risque important à l'obésité et cela peut entraîner à une résistance à l'insuline et par la suite l'intolérance au glucose et au diabète de type 2 [2].

2- Complications cardiovasculaires

Le diabète peut entraîner des problèmes cardiovasculaire 2 à 4 fois plus fréquemment chez les personnes diabétiques que chez les personnes non diabétiques, les études ont montré que ce risque est plus élevé chez les hommes que chez les femmes. Pour limiter le taux de risque d'avoir ces maladies cardiovasculaires il faut réduire la tension artérielle et la glycémie, et en prenant des médicaments hypolipémiants [2].

3- Maladies oculaires liées au diabète :

Les maladies oculaires qui peuvent être provoquées par le diabète comprennent essentiellement la rétinopathie diabétique (RD), l'œdème maculaire diabétique (OMD), la cataracte et le glaucome, mais également la diplopie et l'incapacité de focalisation. Il est possible de réduire l'impact de ces maladies qui peuvent même conduire à la cécité comme la RD en faisant un diagnostic régulièrement pour traiter plus rapidement et avec une meilleure gestion de la glycémie [2].

4- Néphropathies liées au diabète

Les personnes diabétiques ont un grand risque d'avoir une insuffisance rénale (néphropathie) qui est fortement associée aux maladies cardiovasculaires. Le diabète est la première cause de dialyse en France et représente 80 % des insuffisances rénale à l'échelle mondiale [2], celle-ci est dû au mal fonctionnement de la glycémie qui va causer l'hypertension et les reins seront obliger à travailler beaucoup plus pour éliminer l'excès de sucre à travers les urines ce qui va donc causer une insuffisance rénale ou une néphropathie à travers le temps. C'est pourquoi le contrôle de la glycémie et de la tension artérielle est essentiel pour réduire le risque [10].

5- Complications chez les enfants

Les enfants et adolescents (0 à 19 ans) peuvent développer le diabète de type 1 mais également le diabète de type 2 à cause de la malnutrition, manque d'activité physique et les facteurs génétiques. Ces enfants diabétiques ont un risque d'avoir les mêmes complications micro et macro-cardiovasculaire que chez les adultes si le diabète n'est pas bien pris en charge, c'est pour cela que les enfants et adolescents diabétique ont besoins de soutien et beaucoup plus d'attention de la part des parents car ils ont tendance d'avoir du mal à respecter leur traitement [2].

1.8 Prédiabète

Contrairement au diabète de type 1, Le diabète de type 2 ne vient pas brutalement, il est généralement précédé par une phase de prédiabète. Le prédiabète signifie que le taux de glycémie est plus élevé qu'une personne normale (Glycémie à jeun entre 1.10 et 1.25 g/L) mais pas au point de développer un diabète (Glycémie à jeun $\succ$ 1.25 g/L) [11]. Les facteurs de risque du prédiabète et du diabète de type 2 sont similaire, ils comprennent :

- Être en surpoids.
- Le sexe : les hommes sont plus vulnérables.
- L'âge : avoir plus de 45 ans.
- Avoir des antécédents qui ont eu le diabète.
- Mauvaise alimentation et activité physique.
- Pour les femmes : avoir un diabète gestationnel ou donner naissance à un bébé de plus de 4,1 kg [12].

Les personnes atteintes de prédiabète ne vont pas tous avoir le diabète de type 2 mais beaucoup en seront, une bonne alimentation et une perte de poids peuvent aider à prévenir la maladie et retarder ces complications [12].

1.9 Traitement et prévention

Comme précédemment mentionné, le diabète est une maladie chronique et donc la guérison totale est impossible, cependant les traitements sont utilisés pour éviter ou tout au moins retarder l'apparition des complications engendrées par cette maladie qui peuvent être mortelle dans certains cas.

Le traitement du diabète de type 2, prédiabète et diabète gestationnel est essentiellement d'adopter un mode de vie sain (alimentation équilibrée, activités physiques etc.), ou par des médicaments antidiabétiques dans le cas où la maladie n'est pas équilibrée par des mesures saines [13] [14] [15].

Le diabète de type 1 est traité par l'insulinothérapie, elle représente le traitement de référence de ce type de diabète, cette thérapie repose sur des injections sous-cutanées d'insuline plusieurs fois par jour en utilisant des analogues produits par des bactéries génétiquement modifiées, nous distinguons :

- Les analogues « rapides » : cet analogue aide à baisser le taux du glucose immédiatement après la prise d'un repas.
- Les analogues d'action ultra lente (insulines basales) : assure la présence d'insuline dans 24h. Si la maladie n'est pas équilibrée par l'insulinothérapie, la greffe d'îlots de Langerhans

dans le foie peut être un traitement [16].

La prévention est la meilleure façon de réduire l'apparition et les complications causées par la maladie, malheureusement pour un diabétique de type 1 les médecins et les chercheurs n'arrivent pas à comprendre sa cause principale ni à le prévenir, à la différence de diabète de type 2. La prévention du diabète de type 2 est étroitement liée à des règles d'hygiène de vie, Cette prévention peut être réalisée en trois niveaux :

1- La prévention primaire : Dans la pratique, c'est toute activité qui a lieu avant l'apparition de la maladie dans le but d'éviter son apparition, on mentionne :

— L'adoption d'un style de vie sain (alimentation équilibrée, activités sportives, l'abandon du tabac etc.).

— L'identification des personnes qui ont plus la possibilité de développer le diabète par un questionnaire.

— Test de la glycémie, qui permet d'empêcher le développement d'un diabète de type 2 chez les prédiabétiques [7].

2- La prévention secondaire : Elle permet principalement d'éviter les complications en cas d'apparition de la maladie par une détection précoce et un bon contrôle de la glycémie [7].

3- La prévention tertiaire : Elle vise à éviter l'invalidité fonctionnelle et sociale et la réhabilitation des patients handicapés et de réduire la progression des complications chroniques de la maladie [7].

1.10 Problématique

Après avoir eu une vue détaillée sur le diabète, nous pouvons déduire que cette maladie chronique est une menace pour la santé publique car :

— Les médecins n'arrivent pas à trouver un remède.

— La prévalence rapide.

— Les complications qu'il provoque.

— Le taux de mortalité est trop élevé dans le monde entier.

1.11 Solution proposée

L'objectif de ce travail est de proposer une solution basée sur les techniques de l'intelligence artificielle (apprentissage automatique et apprentissage profond) qui vont permettre de réduire l'atteinte du diabète de type 2 et éviter ses complications par une prédiction préalable.

1.12 Conclusion

Dans ce chapitre nous avons présenté une vue détaillée sur le diabète et ses différents types et nous avons surtout présenté les complications qui peuvent être engendrées par cette maladie afin de présenter la problématique et la solution proposé de notre étude. Dans le chapitre suivant nous allons présenter l'intelligence artificielle et ses différentes techniques qui vont être utilisées pour répondre à notre objectif.

CHAPITRE 2

INTELLIGENCE ARTIFICIELLE

2.1 Introduction

Dans le chapitre précédent nous avons parlé sur la maladie chronique, le diabète, et tout ce qui est en rapport avec cette maladie, les causes, le diagnostic, les complications... Et nous avons aussi entamé le but d'utiliser l'intelligence artificielle pour faire une prédiction qui va servir à sensibiliser les gens qui ont un risque d'avoir le diabète de type 2 à prendre un mode de vie sain pour enfin éviter la maladie et ses complications.

Dans ce chapitre nous allons présenter plus en détails l'intelligence artificielle, ses différents domaines et ses techniques plus spécifiquement ceux que nous allons utiliser dans la présente étude.

2.2 L'intelligence artificielle

2.2.1 Définition

La définition de l'intelligence artificielle diffère d'une méthode à une autre. Certaines d'entre eux sont centrées humain et d'autre sont centrées méthode ou problème, on retrouve dans ce tableau quelque définitions de l'IA à travers l'histoire, en haut on retrouve des définitions basées sur la pensée et en bas des définitions basées sur l'action tandis que ceux de droite sont basées sur l'intelligence humaine et ceux en gauche sont basées sur une mesure de performance idéale appelée "la rationalité" [17].

Penser humainement	**Penser rationnellement**
"La tentative nouvelle et passionnante d'amener les ordinateurs à penser...[d'en faire] des machines dotées d'un esprit au sens le plus littéral." (Haugeland, 1985) "[L'automatisation d'activités que nous associons à la pensée humaine,des activités telles que la prise de décision, la résolution de problèmes d'apprentissage... " (Bellman, 1978)	"L'étude des facultés mentales grâce à des modèles informatiques." (Charniak and McDermott, 1985) "L'étude des moyens informatiques qui rendent possible la perception, le raisonnement et l'action." (Winston, 1992)
Agir humainement	**Agir rationnellement**
"L'art de créer des machines capables de prendre en charge des fonctions exigeant de l'intelligence quand elles sont réalisées par des gens." (Kurzweil, 1990) "L'étude des moyens à mettre en oeuvre pour faire en sorte que des ordinateurs accomplissent des choses pour lesquelles il est préférable de recourir à des personnes pour le moment" (Rich and Knight, 1991)	"L'intelligence artificielle (computational intelligence) est l'étude de la conception d'agents intelligents.." (Poole et al., 1998) "L'IA... étudie le comportement intelligent dans des artefacts." (Nilsson, 1998)

TABLE 2.1 – Définitions de l'intelligence artificielle en quatre catégories [17].

Cependant des décennies avant ces définitions, un test qui s'appel le test de Turing a été introduit par Alain Turing en 1950 dans son ouvrage " Computing Machinery and Intelligence ". Ce test consiste à laisser un interrogateur humain faire la distinction entre une réponse textuelle humaine et une autre informatique pour mesurer l'IA et sa capacité de penser comme l'humain. Bien que ce test soit minutieux depuis le développement de l'IA, il reste une partie importante de l'histoire de l'IA et de la philosophie [18].

2.2.2 Les types de l'intelligence artificielle

Il existe deux principaux types de l'intelligence artificielle : IA faible et IA forte.

— **Intelligence artificielle faible (Weak AI) :**

L'intelligence artificielle faible appelée aussi intelligence artificielle étroite (Artificial Narrow Intelligence) est simplement toutes les applications de l'IA qui existe aujourd'hui, ce sont des programmes formés pour exécuter des tâches précises en utilisant des algorithmes avancés. Nous pouvons prendre comme exemple l'application Siri d'Apple, Alexa d'amazone et les véhicules autonomes [18].

— **Intelligence artificielle forte (Strong AI) :**

L'IA forte est une forme théorique de l'intelligence artificielle, elle est composée de deux catégories : l'Intelligence Artificielle Générale (IAG) et la Super Intelligence Artificielle (SIA). L'intelligence artificielle général (IAG) est lorsque la machine aura une intelligence égale à celle des humains et qu'elle aura la capacité d'apprendre, d'être autonome et indépendante dans ses choix. Tandis que la Super Intelligence Artificielle (SIA) est définie comme étant une intelligence extraordinaire qui surpassera l'intelligence et les capacités du cerveau humain [18]. La figure 2.1 présente une estimation sur quand est ce que nous allons atteindre l'Intelligence Artificielle Général et la Super Intelligence Artificielle.

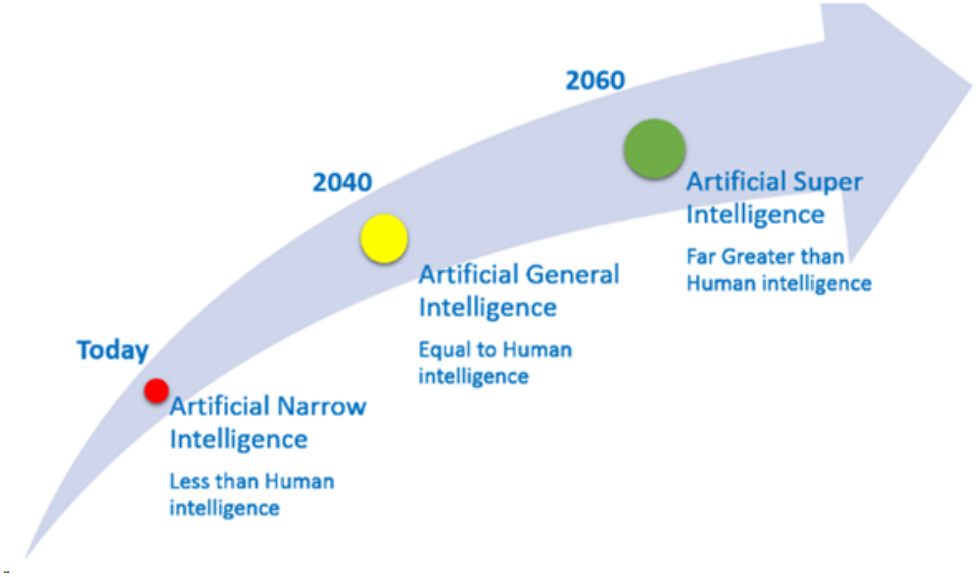

FIGURE 2.1 – Future de l'IA [19].

2.2.3 Les branches de l'intelligence artificielle

Il existe différentes branches de l'intelligence artificielle, certaines d'entre elles sont présentées dans la figure suivante :

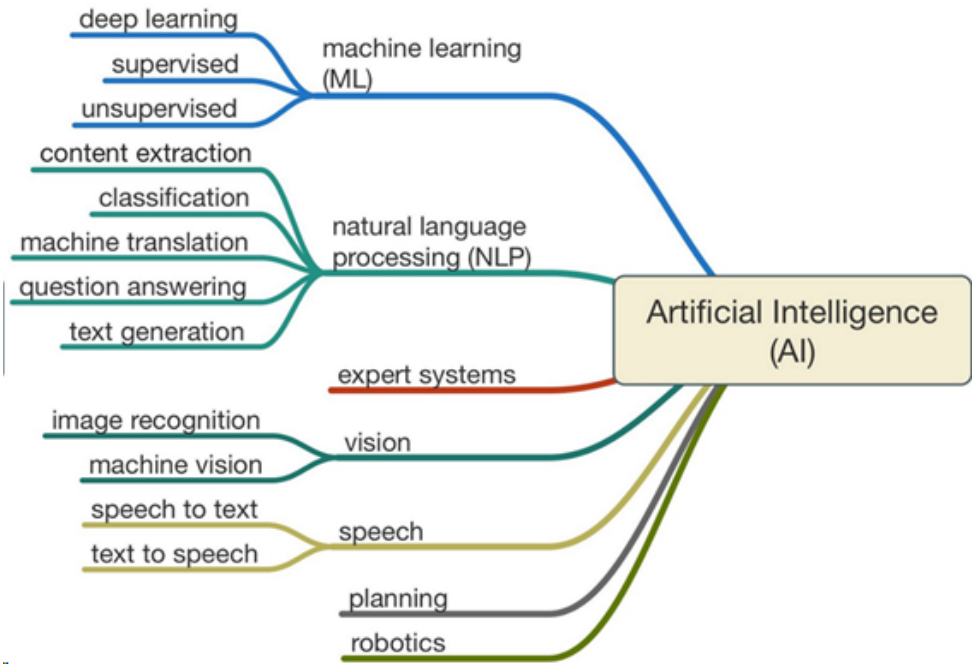

FIGURE 2.2 – Les branches de l'IA [20].

Toutes ces branches sont de l'intelligence artificielle faible car l'intelligence artificielle forte est toujours théorique, voici la description de quelqu'une (voir plus en détail l'apprentissage automatique et l'apprentissage profond dans la suite du chapitre).

— **Natural language processing (NLP) :**

Le traitement de langage naturel est l'un des branches de l'IA les plus connu, il est dédié a tout ce qui est linguistique. il permet aux ordinateurs de lire les textes, d'entendre la parole, de l'interpréter, d'analyser la parole et les émotions et de déterminer quelles parties sont importantes. Tout ça d'une manière plus efficace que les humains et sans épuisement [20].

— **Robotics :**

Les robots sont des machines programmées par les humains pour faire des tâches précises. ils ont des capteurs qui servent à collecter les informations de l'environnement extérieur tel que la température, les mouvements, les paroles afin de les interpréter et faire des actions par la suite. Ils sont aussi capables d'apprendre de leurs erreurs et de s'adapter au nouvel environnement [17].

— **Système Expert :**

Les systèmes experts visent à simuler le processus décisionnel d'un expert humain. Ces systèmes sont conçus pour être très réactifs, fiables et compréhensibles lorsqu'ils aident à la prise de décision humaine. Ce type d'intelligence artificielle est aujourd'hui utilisé pour des applications d'aide à la décision, gestion des entreprises et ainsi dans le domaine médical pour aider à faire des diagnostics [21].

2.3 L'apprentissage automatique

2.3.1 Définition

L'apprentissage automatique ou Machine Learning en anglais, est un sous-domaine de l'IA qui se fond sur des approches mathématiques et statistiques pour développer des algorithmes permettant aux ordinateurs d'accomplir des tâches sans être explicitement programmés pour chacune tout en se basant sur des situations rencontrées (données) [22].

L'apprentissage automatique implique généralement deux phases :

— La phase d'apprentissage ou phase d'entraînement, est effectuée avant que le modèle ne soit réellement utilisé. Elle consiste à estimer le modèle à partir des données disponibles.

— Phase de test : La deuxième phase correspond à la mise en production, le modèle est déterminé, puis de nouvelles données peuvent être testées pour obtenir les résultats correspondant à la tâche souhaitée.

2.3.2 les données

Comme précédemment indiqué dans la définition, les algorithmes d'apprentissage automatique sont basés sur les données, autrement dit samples, observations ou exemples. Deux types de données peuvent être utilisés :

— les données étiquetées (labellisées, annotées) ou chaque observation X est associée à une étiquette Y.

— les données non-étiquetées ou aucune étiquette Y n'est associée aux observations.

La construction d'un jeu de données (dataset) non étiquetées est généralement plus facile que celle d'un jeu de données étiquetées, car dans cette dernière, une intervention humaine est nécessaire pour définir les étiquettes à chaque observation, contrairement dans un dataset non étiqueté, il suffit simplement de collecter, après un prétraitement automatique minimal, les données [23].

2.3.3 les types d'apprentissage automatique

L'apprentissage automatique est un domaine vaste qui peut résoudre plusieurs types de problèmes, nous distinguons quatre formes d'apprentissage :

— Apprentissage supervisé.

— Apprentissage non supervisé.

— Apprentissage semi-supervisé.

— Apprentissage par renforcement.

1- L'apprentissage supervisé :

L'apprentissage supervisé est une branche qui traite les problèmes, dont les données sont étiquetées où chaque observation n est associée à une étiquette y, cette dernière est produite par une fonction inconnue f(x). L'objectif est de trouver une fonction H, aussi appelé modèle, qui approche à la fonction f afin de prédire l'étiquette Y de la nouvelle observation X, à partir de la connaissance fournie par les N observations étiquetées de la base d'apprentissage.

L'apprentissage est donc une recherche dans un espace des hypothèses possibles, d'une hypothèse qui prédit correctement la valeur de la nouvelle étiquette.

L'exactitude d'une hypothèse est mesurée par un ensemble de données de test contenant des exemples distincts de ceux de la base d'apprentissage [24]. Le workflow d'apprentissage supervisé est représenté par la figure suivante :

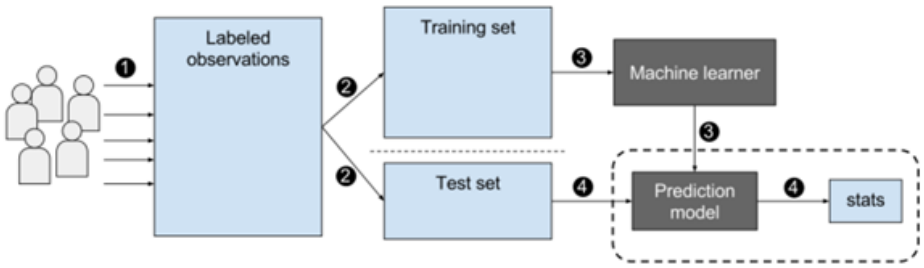

FIGURE 2.3 – Workflow d'apprentissage supervisé [25].

Dans la figure 2.3, la méthode train test split est utilisée pour l'évaluation de la performance des algorithmes d'apprentissage automatique utilisés. Plusieurs techniques existent, Nous mentionnons la technique de validation croisée.

Il existe deux types de modèles d'apprentissages supervisés :

— **Classification :**

Un modèle de classification permet de prédire une valeur qualitative, c'est-à-dire lorsque l'ensemble des valeurs de sortie Y prennent leurs valeurs dans un ensemble fini dont les éléments correspondent à des catégories (ou classes) à identifier. Lorsque l'étiquette Y

prend que deux valeurs, on parle de la classification binaire.

Exemple : identifier si un patient est diabétique ou non.

Dans le cas où Y est discrète et finie on parle de la classification multi-classe

Exemple : identifier le type de diabète chez un patient [24].

La figure ci-dessous permet de présenter les types du modèle de classification

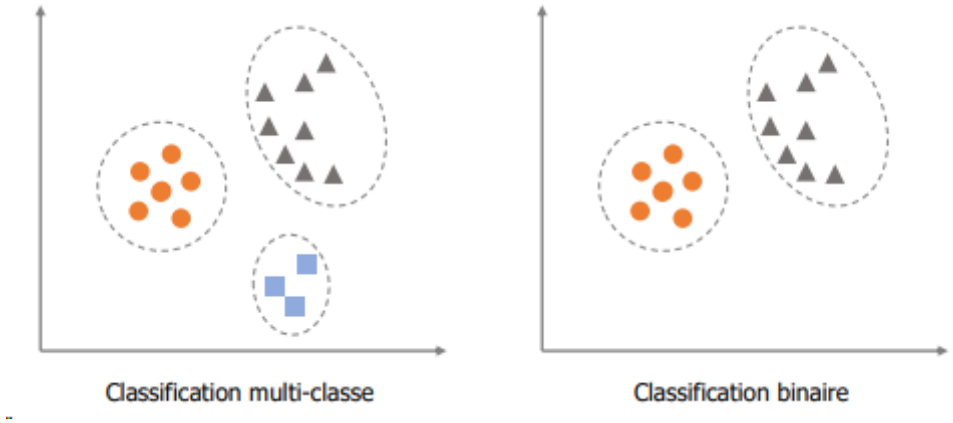

FIGURE 2.4 – Représentation des observations [26].

— **Régression :**

Une technique de modélisation qui permet de prédire une valeur quantitative, c'est à dire lorsque l'ensemble des valeurs de sortie Y prennent des valeurs réelles [24]. Exemple : Prédire le taux du glucose chez un patient selon son poids, l'âge etc. Les types de régression sont les suivants :

- Régression linéaire.
- Régression polynomiale.
- Régression logistique.
- Régression de Crête.
- Régression de Lasso [27].

2- L'apprentissage non supervisé :

Contrairement à l'apprentissage supervisé, ce type d'apprentissage traite des problèmes, dont les données ne sont pas étiquetées, dans l'objectif d'extraire des caractéristiques communes entre les N observations [24].

Les techniques d'apprentissage non supervisé peuvent être utilisées pour résoudre les problèmes suivants :

— **Partitionnement :**

Partitionnement ou clustering, consiste à regrouper les observations en des clusters (groupes) homogènes selon un critère de similarité, cette dernière est calculée par une

distance D [23].

Les méthodes de partitionnement utilisées visent à maximiser la similarité intra-groupe et de minimiser similarité inter-groupes [23].

Voici un exemple sur le clustering comme indiqué dans la figure suivante.

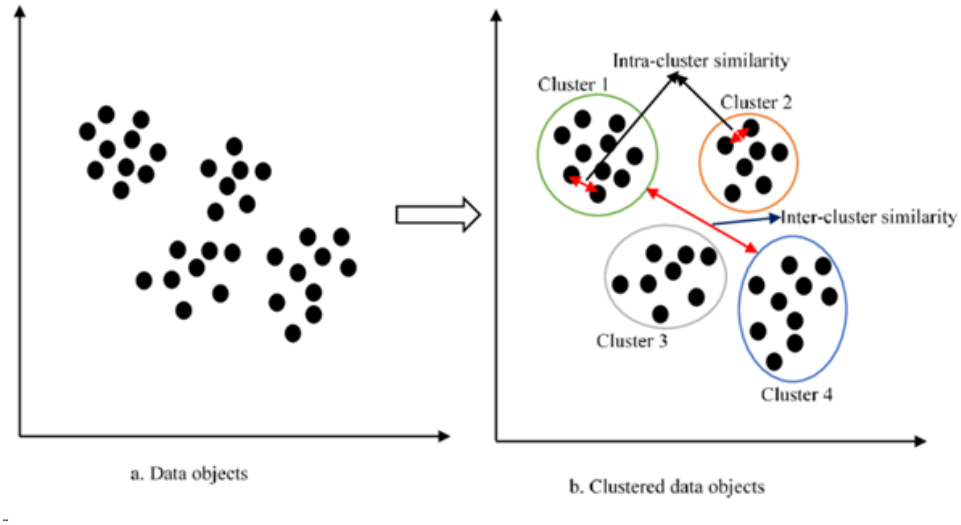

FIGURE 2.5 – Exemple de clustering avec illustrations intra et inter-clustering [28].

— **Réduction de dimension :**

Cette technique est utilisée lorsque le nombre de variables p utilisées pour représenter les données est très grand. Son principe est de représenter les données dans un espace de dimension plus petite que celle de l'espace dans lequel elles sont représentées à l'origine, dans le but de réduire le temps d'exécution et l'espace mémoire ainsi d'améliorer les performances d'un algorithme d'apprentissage supervisé [24].

— **Estimation de densité :**

Les fonctions de densité sont utilisées pour représenter la distribution des données dans leur espace vectoriel, cela permet de bien comprendre les données et de construire des modèles décisionnels entre les classes.

Supposant que le jeu de données est un échantillon aléatoire X, l'objectif est de trouver la fonction de densité de probabilité f qui a généré X.

La figure ci-dessous représente un exemple où les observations (présentées par les points bleus sur le plan horizontal) sont issues de R^2 et la densité qui a généré ces points est la surface courbe.

Le sommet (en rouge) sur cette surface montre la région où les observations sont les plus denses [29].

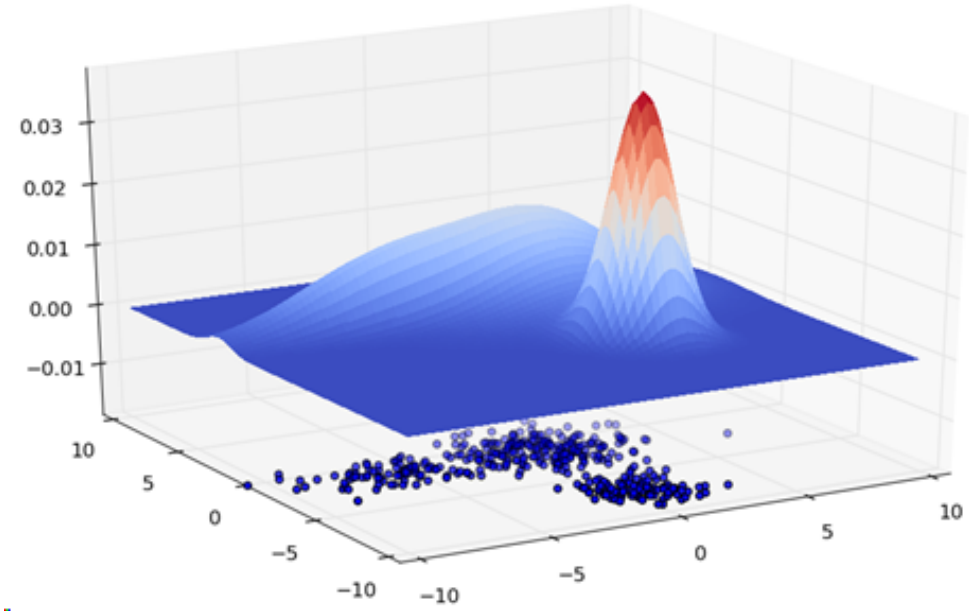

FIGURE 2.6 – Exemple de données de R^2 et de densité estimée [29].

3- L'apprentissage semi supervisé :

L'apprentissage semi-supervisé consiste à apprendre des étiquettes à partir d'un jeu de données partiellement étiqueté (contient des données étiquetées et non étiquetées), l'avantage de cette technique est au lieu d'étiqueter manuellement les données non étiquetées, nous donnons des étiquettes approximatives sur la base des données étiquetées [30]. Les étapes suivantes expliquent le principe d'apprentissage semi-supervisé :

— Entraîner un modèle à partir des données étiquetées.

— Utiliser le modèle pour prédire les étiquettes des données non-étiquetées.

— Réentraîner le modèle avec les données étiquetées et pseudo-étiquetées [31].

4- L'apprentissage par renforcement :

L'apprentissage par renforcement est le type d'apprentissage automatique où l'agent (machine) a la possibilité d'interagir avec son environnement, elle apprend par la méthode d'essais et d'erreurs. L'agent effectue des actions, l'environnement évalue ces actions et répond par une récompense ou une punition, sur la base de cette évaluation, l'agent déterminera si cette action est considérée comme le bon choix ou non, afin d'établir une stratégie qui permet d'obtenir la meilleure récompense possible. Ce type d'apprentissage est souvent utilisé dans les jeux (échecs, go) et la robotique [24].

Le mécanisme d'apprentissage par renforcement est montré dans la figure ci-dessous.

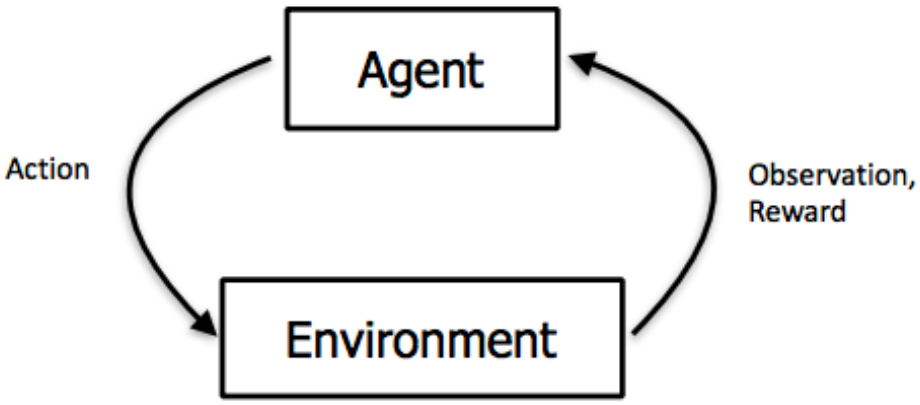

FIGURE 2.7 – Mécanisme d'apprentissage par renforcement, interactions entre l'environnement et l'agent [32].

2.4 Algorithmes d'apprentissage automatique

Voici quelques algoithmes utilisées dans la classification et la régession.

2.4.1 Forêts aléatoires (RF)

- Définition :

Forêts aléatoires ou Random forest en anglais, est une technique d'apprentissage automatique supervisée qui peut résoudre les problèmes de classification ainsi que les problèmes de régression. Elle est basée sur le concept d'apprentissage d'ensemble, qui est un processus de combinaison de plusieurs classificateurs pour résoudre un problème complexe et améliorer les performances du modèle. Le principe est de créer plusieurs arbres de décision (c'est la raison pour laquelle il est appelé une forêt) sur divers sous-ensembles de données et par la suite, L'algorithme établit la sortie finale selon des votes majoritaires sur les prédictions des arbres de décision, l'augmentation du nombre d'arbres augmente la précision du résultat [33].

Le schéma ci-dessous explique le fonctionnement de l'algorithme Random Forest :

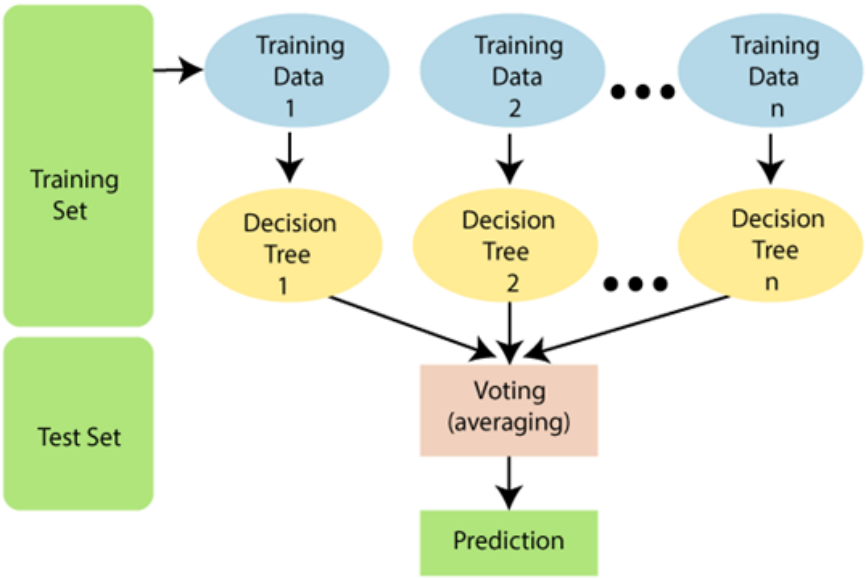

FIGURE 2.8 – Principe de fonctionnement de l'algorithme Random Forest [34].

- Principe de fonctionnement :

Comme indiqué précédemment dans la définition, la technique des forêts aléatoire se base sur le concept d'ensemble, elle utilise la méthode de Bagging.

Bagging ou Bootstrap Aggregation, est l'une des méthodes d'apprentissage d'ensemble qui sert à choisir à partir de la base d'apprentissage, un ensemble de données d'une façon aléatoire pour construire des nouvelles bases d'apprentissage. Cette étape est appelée Bootstrap, Chaque arbre de décision est créé à partir de ces nouvelles bases d'apprentissage.

À la réception d'une nouvelle donnée, tous les arbres de décision seront exécutés ou chaque arbre (modèle) fournit un résultat de prédiction. Le résultat final est basé sur le vote à la majorité après avoir combiné les résultats de tous les modèles. Cette étape est connue sous le nom d'agrégation [35].

la figure 2.9 explique le mode de fonctionnement de la technique "bagging".

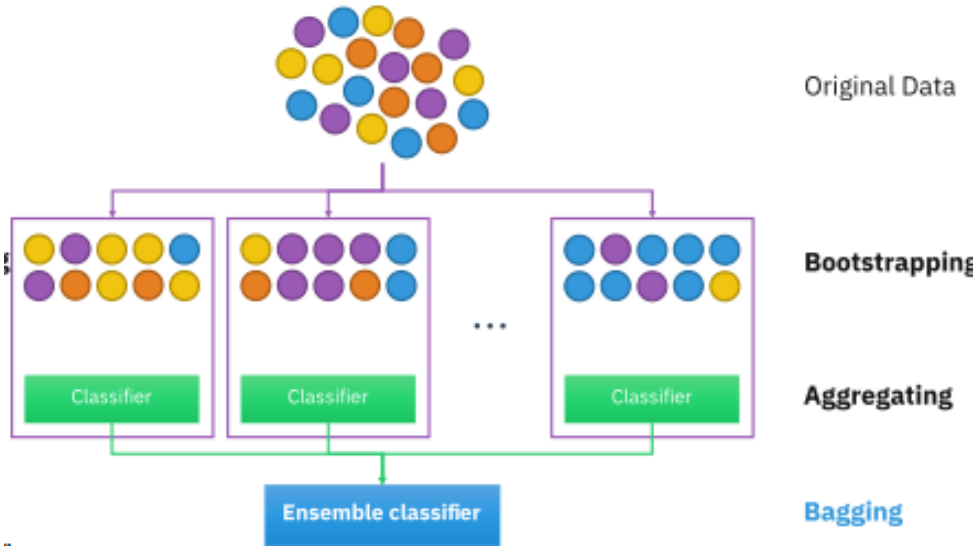

FIGURE 2.9 – Principe de la méthode bagging [35].

- Exemple d'application :

Dans cet exemple, le processus d'entraînement et de classification est élaboré par un modèle de forêt aléatoire dont la base d'apprentissage est composée d'un ensemble d'attribut qui caractérisent le diabète.

A) cette phase sert à créer les arbres de décision sur un échantillon aléatoire des données de la base d'apprentissage origine, qui contient des exemples positifs pour le cas d'une personne diabétiques (étiquettes vertes) et négatifs dans le cas contraire (étiquettes rouges).

B) Chaque arbre sera exécuté de la façon suivante : pour chaque nouvelle instance X, l'algorithme commence au nœud racine d'un arbre de décision et parcourt l'arbre en testant les valeurs des variables dans chacun des nœuds visités, en fonction de chacun, il sélectionne la prochaine branche à suivre. Ce processus est répété jusqu'à l'atteinte d'un nœud feuille, ce qui affecte une classe à cette instance (positive ou négative).

À la fin du processus, chaque arbre vote pour l'étiquette de classe préférée, et le mode des sorties est choisi comme prédiction finale à la base des votes majoritaires parmi tous les arbres individuels.

Le processus est représenté dans la figure suivante :

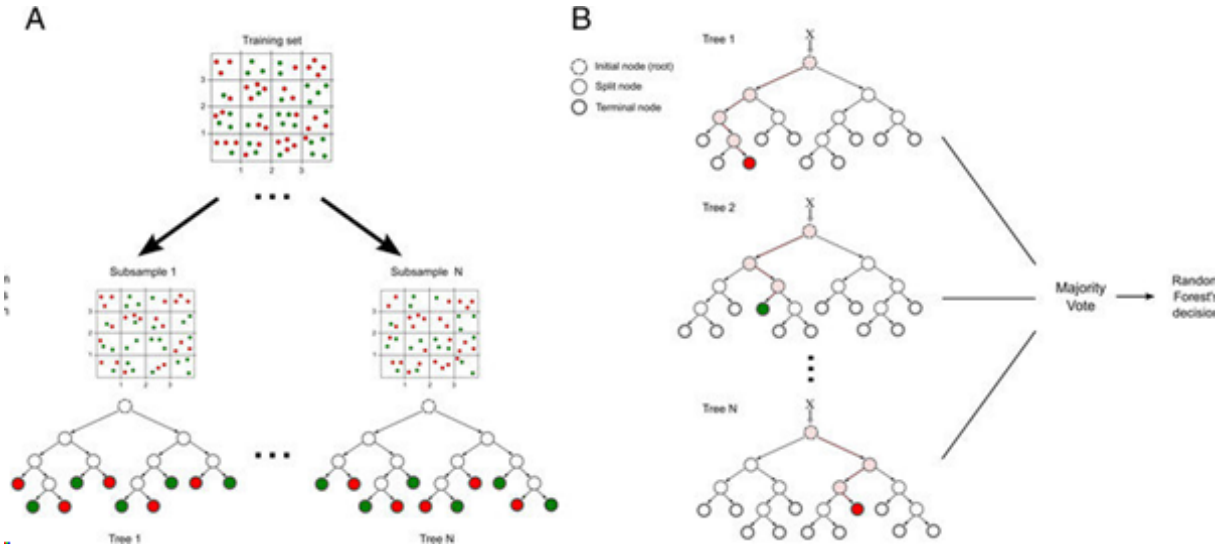

FIGURE 2.10 – Processus d'entraînement et de classification en utilisant le classificateur forêt aléatoire [36].

- Avantages :

— Il fonctionne efficacement avec les grandes bases de données.

— Il améliore la précision du modèle et évite le problème de surajustement(overfitting).

— Il dispose d'une méthode efficace pour gérer les données manquantes [33].

- Inconvénient :

Le problème majeur de cette méthode est, d'une part, la consommation du temps. En général, ces types d'algorithmes sont rapides lors de la phase d'entraînement, mais assez lent dans la phase de test à cause du nombre d'arbre de décision utilisé [37], d'autre part la consommation de ressources [33].

2.4.2 Machine à vecteur de support (SVM)

- Définition :

Support Vector Machine ou SVM est un algorithme d'apprentissage automatique supervisé qui peut être utilisé pour les problèmes de classification ainsi que pour les problèmes de régression. Le principe de cet algorithme est de tracer des points dans un espace de n dimensions (n est le nombre d'attribut ou features) ou chaque point représente une donnée suivant ses coordonnées. Ensuite le but est de trouver la meilleure ligne autrement appelé Hyper-Plane qui peut séparer les données en classes distinctes afin de faire une meilleure classification comme il est présenté dans la figure 2.11 [38].

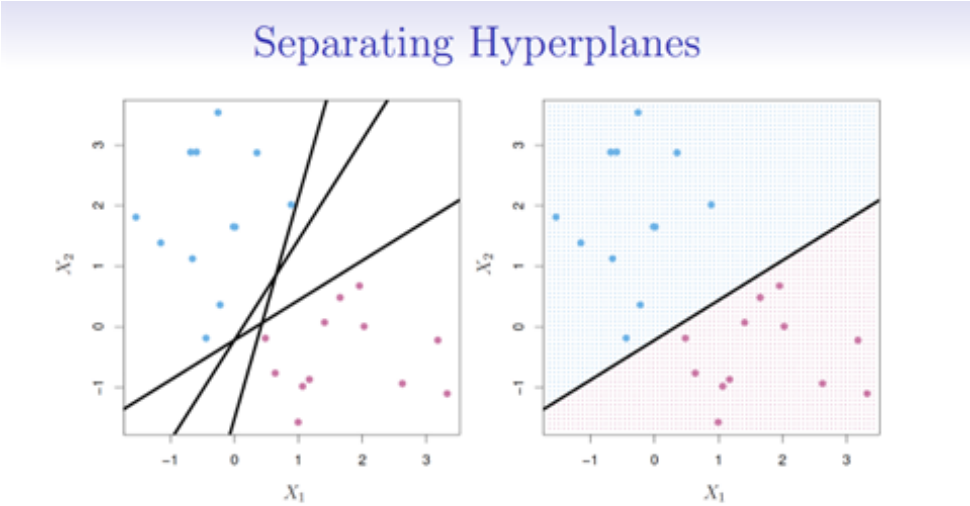

FIGURE 2.11 – Séparation des deux classes avec le meilleur HyperPlane [39].

- Exemple d'application :

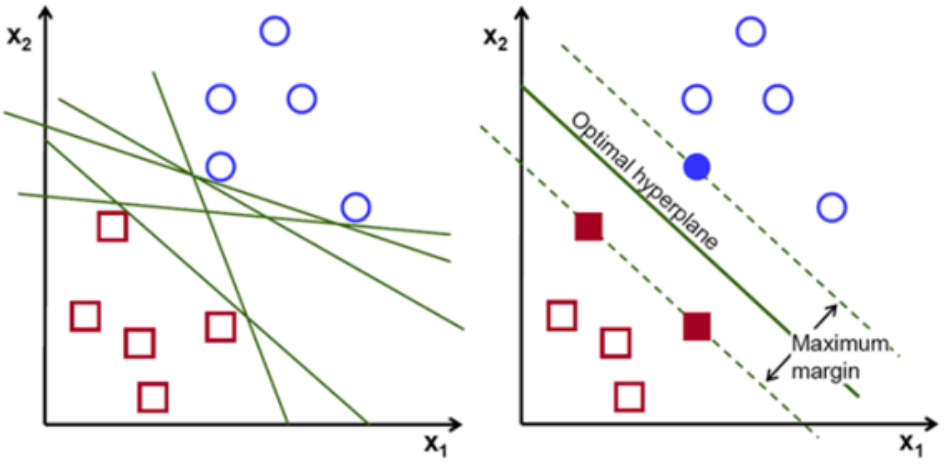

FIGURE 2.12 – Un exemple sur le fonctionnement de l'algorithme SVM [40].

Dans cet exemple d'application nous pouvons constater que l'espace de dimension contient deux classes : la classe des points de données carré et la classe des points de données cercle. L'objectif de l'algorithme SVM est de trouver un Hyperplan optimal, c'est-à-dire la recherche d'un Hyperplan qui non seulement sépare parfaitement les deux classes mais qu'il soit aussi éloigné au maximum de tout point de donnée. En effet, pour obtenir un Hyperplan optimal, il faut garder une marge maximale pour qu'il puisse placer les nouveaux points de données dans les classes qui leur conviennent le plus.

- Hyperplan et les vecteurs de support dans l'algorithme SVM :

1- Hyperplan :

Les Hyperplans sont des limites de décision qui aident à classifier les différents points de données. Plusieurs sont possibles pour séparer les points de données en classes mais l'objectif est de trouver le meilleur entre eux. Pour cela il faut choisir l'hyperplan qui maximise la marge entre les points de données les plus proches aux hyperplans, plus la distance et la marge sont grandes plus nous pouvons faire une classification plus exacte et une prédiction plus précise [40].

2- Vecteurs de support :

Les vecteurs de support sont les points de données les plus proches de l'Hyperplan comme nous pouvons le constater dans la figure 2.13, qui aident à maximiser la marge. C'est des points de données qui influent sur la position et l'orientation de l'Hyperplan et nous aident à construire notre modèle SVM [40].

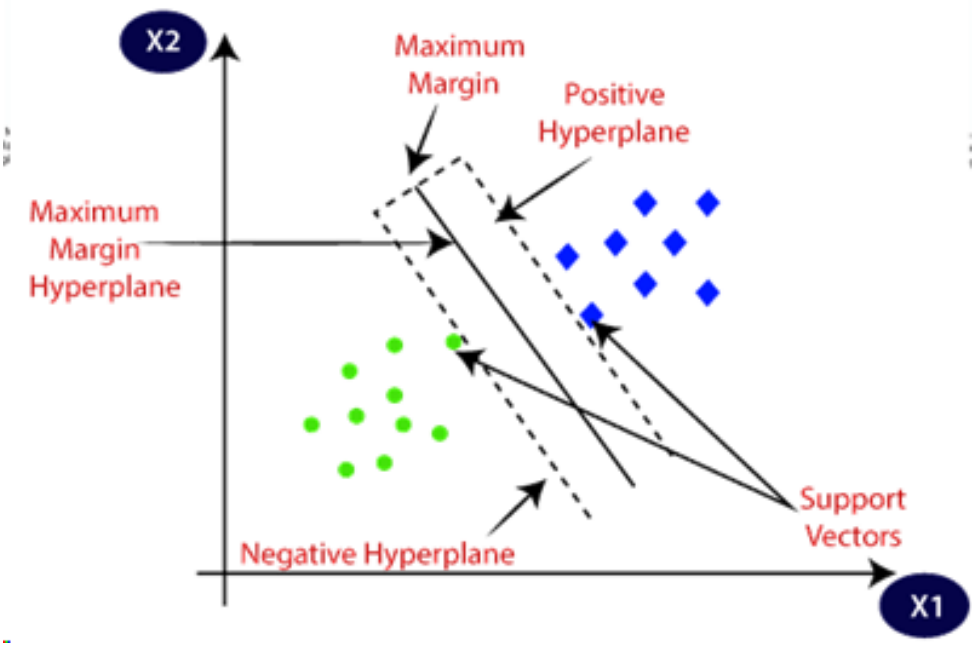

FIGURE 2.13 – L'Hyperplan optimale et les vecteurs de support [41].

- Les types de Support Vector Machine :

Il existe deux principaux types de l'algorithme SVM : linéaire et non linéaire, ce qui différencie l'un sur l'autre est essentiellement les points de données.

1. SVM Linéaire :

Ce type de SVM est utilisé seulement pour les points de données qui sont linéairement séparable et que nous pouvons facilement trouver un Hyperplan Linéaire pour séparer les deux classes de données. Ce type de SVM est souvent appelé Classificateur SVM ou SVC [39].

2. SVM Non linéaire :

Quand les points de données ne sont pas linéairement séparables (Dans la plupart du temps) et que les points de données ne peuvent pas être séparés par un Hyperplan, le non linéaire SVM est utilisé, en faisant appel à des techniques plus avancées comme le Kernel [42].

- Principe de fonctionnement :

L'algorithme Support Vector Machine (SVM) tente à tracer un Hyperplan sur les données d'une base d'apprentissage afin de les classer en deux groupes. Cet Hyperplan peut être une simple ligne si nous sommes en 2 dimensions, un plan en 3 dimensions et Hyperplan en (4D+) [39].

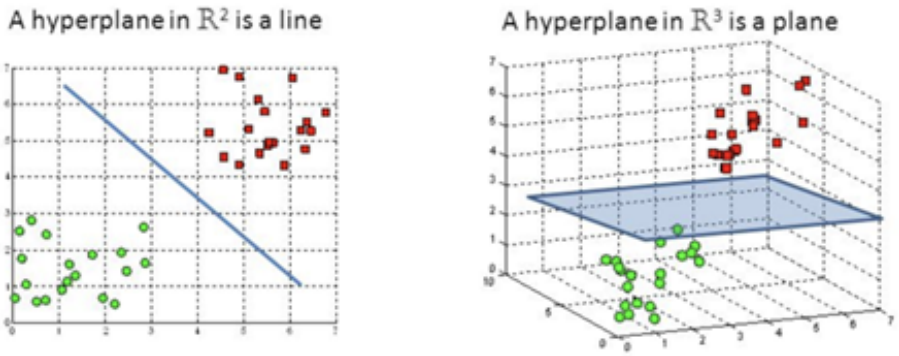

FIGURE 2.14 – Les hyperplan en 2D et en 3D [40].

- Nous pouvons trouver une infinité d'Hyperplans possibles pour classer les points de données, mais l'Hyperplan optimale reste celui qui maximise la marge entre les vecteurs de support. Il est appelé le "Maximum Margin Classifier" nous le considère comme optimal car plus la marge est grande plus il est robuste et plus il donne une classification précise surtout en ajoutant de nouvelles données[39].

- Dans certains cas les points de données ne sont pas parfaitement linéairement séparable, là le Maximum Margin Classifier ne marchera pas et nous devons faire appelle à une nouvelle notion qui s'appelle Soft Margin, cette marge permet de garder l'objectif de maximiser la marge mais en autorisant quelques mauvaises classifications. Ces mauvaises classifications peuvent être contrôlées à l'aide d'un paramètre qui s'appelle Tuning Parameter (dénoté en C) plus la valeur de C est grande plus le programme va autoriser les mauvaises classifications [39].

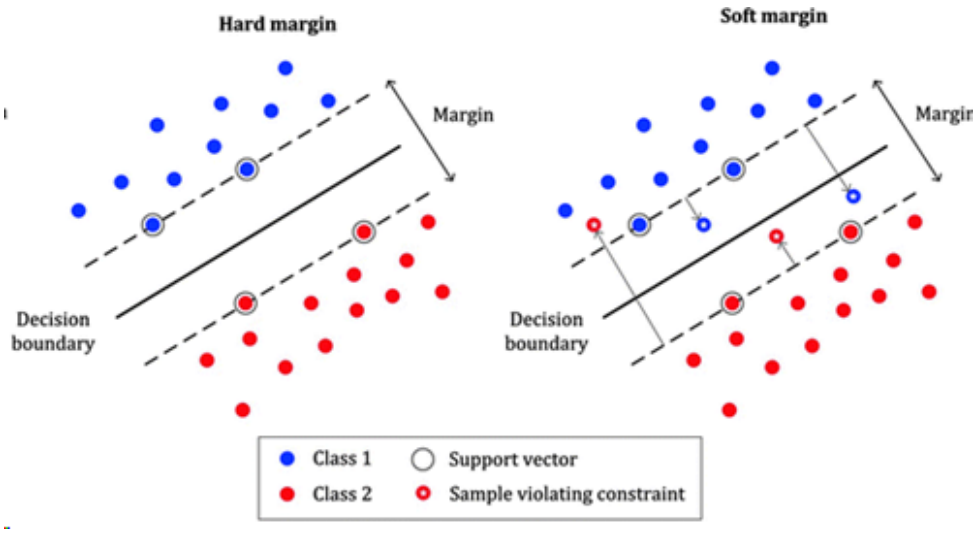

FIGURE 2.15 – Soft and Hard Margin [43].

- Dans d'autres cas, les points de données ne sont pas linéairement séparables comme le montre la figure suivante :

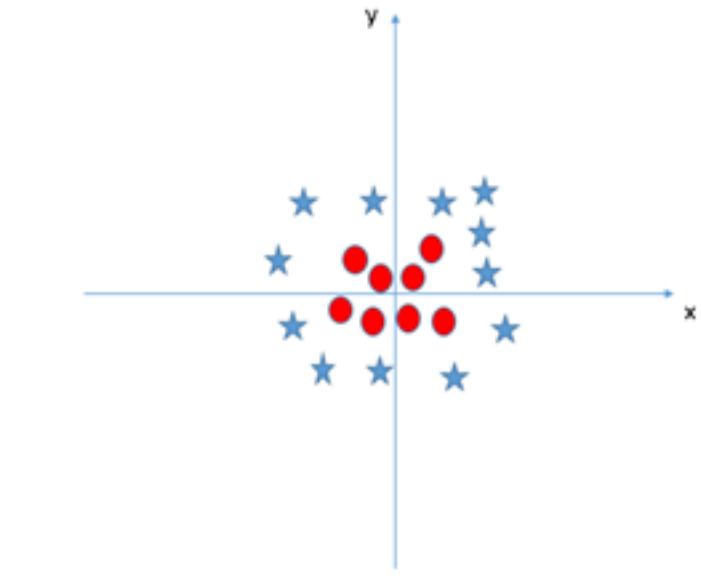

FIGURE 2.16 – Points de données non linéairement séparables [38].

Pour résoudre ce problème, une méthode a été développée en 1992 par Vapnik qui s'appelle The Kernel Trick [39], cette fonction consiste à transformer un espace de basses dimensions en un espace à des dimensions plus élevées, en d'autres termes il convertit un problème non linéairement séparable en un problème linéairement séparable [38]. Plusieurs types de Kernel existent certain d'entre eux sont Polynomial Kernels, Radial Basis Kernels et Linear Kernels. Tous ces Kernels

ont pour but de transformer les données pour permettre de construire un HyperPlane optimale [39].

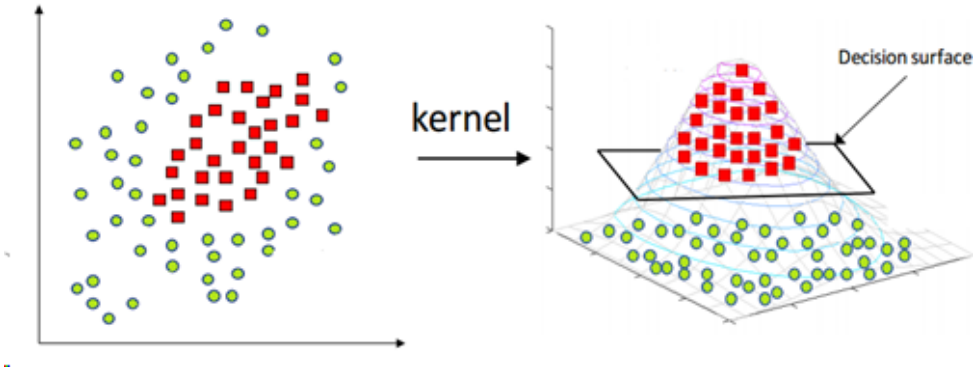

FIGURE 2.17 – Transformation des données en un espace linéairement séparable [42].

2.5 Réseaux de neurones et apprentissage profond

2.5.1 Réseaux de neurones

- Définition :

Un réseau de neurones est un ensemble d'éléments ou de nœuds interconnectés entre eux par des liens, sa capacité de traitement est dû à un ensemble de paramétrage de poids et de puissance qui est obtenue par un processus d'adaptation et d'apprentissage à partir d'un ensemble d'entraînements [44].

Les réseaux de neurones sont principalement inspirés de la cellule nerveuse, pour bien comprendre les réseaux de neurones il faut avoir une brève compréhension sur la neurobiologie.

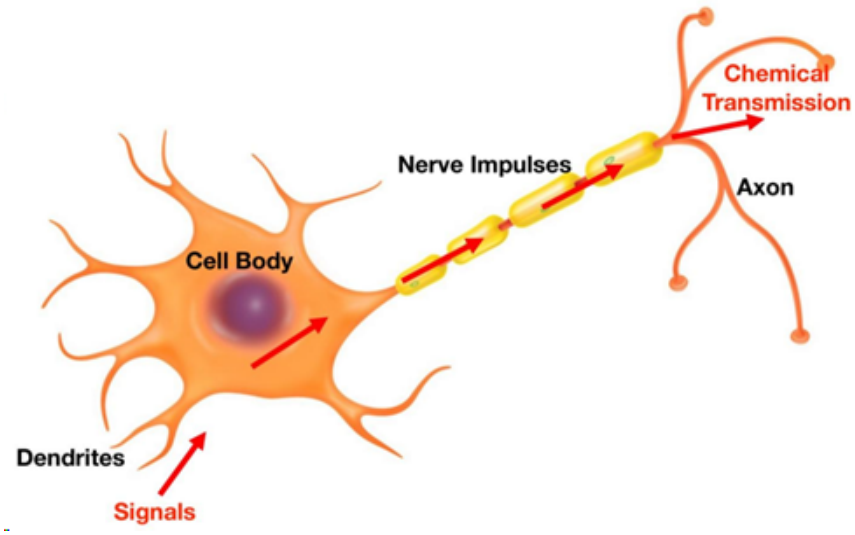

FIGURE 2.18 – Le modèle d'un neurone biologique [45].

Le cerveau humain est constitué d'environ 100 milliards de neurones qui communiquent entre eux par des signaux électriques, la communication entre les neurones est faite à l'aide des jonctions électrochimiques appelées synapses qui se situent sur les dendrites qui sont les branches de la cellule comme il est présenté dans la figure 2.18 [44].

Chaque neurone reçoit constamment des milliers de signaux par d'autres neurones via les dendrites qui vont éventuellement atteindre le corps cellulaire. Ensuite ils vont être additionnés ensemble d'une certaine manière puis décide s'il va générer une impulsion électrique ou pas qui va être transmise via les axons aux autres cellules [44].

- Structure d'un réseau de neurones artificielle :

Celle-ci est la structure la plus simple d'un réseau de neurones artificielle qui est inspiré par les neurones biologiques, ce qu'on appelle un "Perceptron".

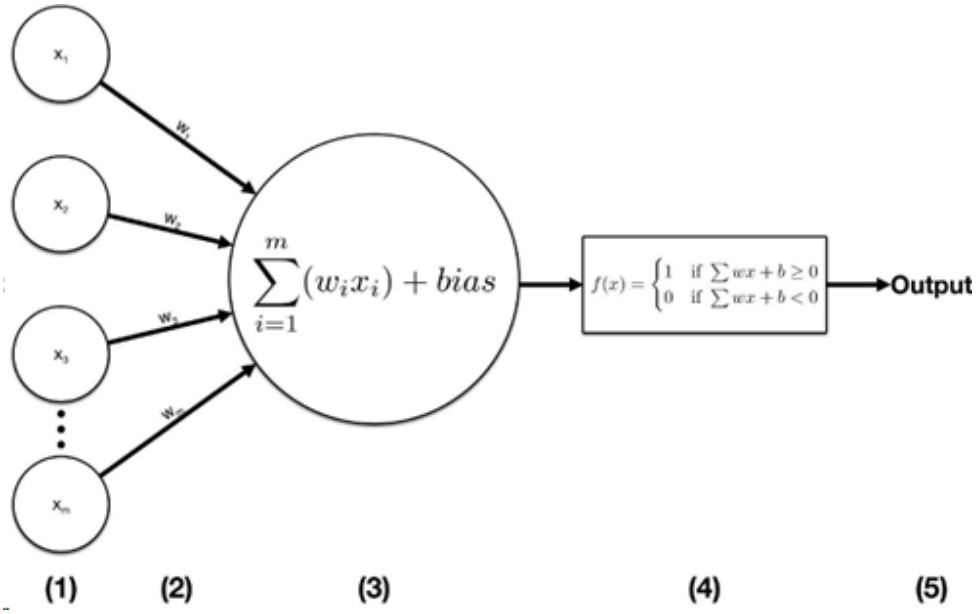

FIGURE 2.19 – modèle mathématique d'un réseau de neurones [45].

1. Les réseaux de neurones sont composés d'un ensemble de nœuds autrement appelés ensemble d'unités, chaque unité contient une donnée.

2. Chaque nœud est relié par un lien, et chaque lien a une valeur numérique appelée poids qui détermine la puissance de la connexion.

3. Une fonction d'addition (Summation) sera exécutée où chaque donnée sera multipliée par son poids individuellement, ensuite on fait l'addition de toutes ces données en ajoutant une autre valeur appelée Bias qui sert à décaler la fonction d'activation.

4. Le résultat de la précédente opération (Summation) représente l'entrée de la fonction d'activation, dans cet exemple la fonction d'activation donne 1 si l'entré est supérieur à 0 et donne un 0 sinon.

5. Le résultat de la fonction d'activation est soit 0 soit 1.

- L'apprentissage :

Apprentissage ou méthode de paramétrage des poids, est un processus par lequel les poids d'un réseau de neurone s'adaptent, les principaux modes d'apprentissage sont les suivant :

1- Supervisé :

Dans ce type d'apprentissage, les entrées et les sorties sont connues à l'avance c'est-à-dire que la base d'apprentissage est étiquetée. Le réseau traite ensuite les entrées et compare les résultats obtenus aux sorties souhaitées. Les poids sont ensuite ajustés grâce aux erreurs propagées à travers le système. Ce processus est répété tant que les poids sont continuellement améliorés. L'ensemble de données qui permet l'apprentissage est appelé l'ensemble d'apprentissage [44] [46].

2- Non supervisé :

Contrairement à l'apprentissage supervisé, dans ce type d'apprentissage les sorties ne sont pas fournies au préalable, le réseau est fourni seulement avec des entrées alors c'est au système de choisir quelles fonctionnalités à utiliser pour regrouper les données d'entrée. Cette méthode est appelée l'auto-organisation ou l'adaptation [44] [46].

- Fonctions d'activation :

Plusieurs fonctions d'activation ont été testées mais peu d'entre elles ont été trouvées pratiques. Ici nous allons vous présenter quatres fonctions d'activation les plus utilisé dans la figure suivante :

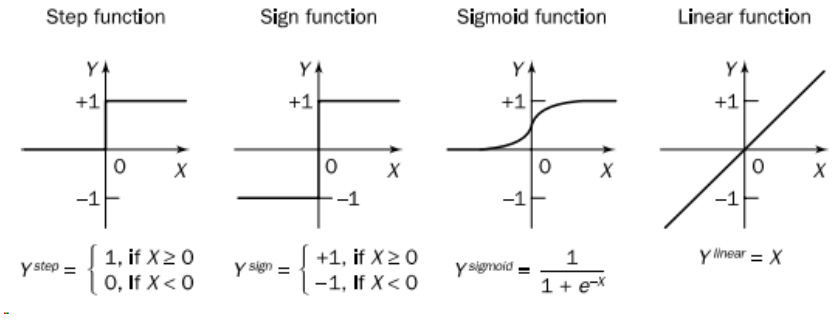

FIGURE 2.20 – Les fonctions d'activation utilisées par les neurones [47].

1- Fonctions Step et Sign :

Ces deux fonctions sont également appelées fonctions à limite dure, elles sont souvent utilisées dans les problèmes de classification, reconnaissance des motifs ainsi que dans les neurones décisionnels [47].

2- Fonction Sigmoid :

C'est l'une des fonctions d'activation non linéaire les plus utilisées, elle transforme n'im-

porte quelle donnée de moins l'infinie à plus l'infinie en une valeur raisonnable entre 0 et 1 [47].

3- Fonction Linéaire :

Cette fonction fournit un résultat égal à la sortie de l'opération d'addition de l'ensemble des données multiplié par leurs poids, ce type de fonction est utilisé dans l'approximation linéaire [47].

- Les différents types d'architecture des réseaux de neurones :

L'architecture des réseaux de neurones consiste à décrire comment les neurones sont connectés entre eux, voici quelques exemples d'architectures :

1- Propagation vers l'avant une seule couche (One layer feed forward) :

Ce type d'architecture est aussi appelé un réseau de neurones non bouclé (statique), il est représenté par un graphe où nous avons seulement deux couches, à savoir la couche d'entrée et la couche de sortie, la couche d'entrée ne compte pas car aucun calcul est effectué sur cette couche. Le signal se propage à partir de la couche d'entrée jusqu'à la couche de sortie d'une façon linéaire sans retour en arrière [48].

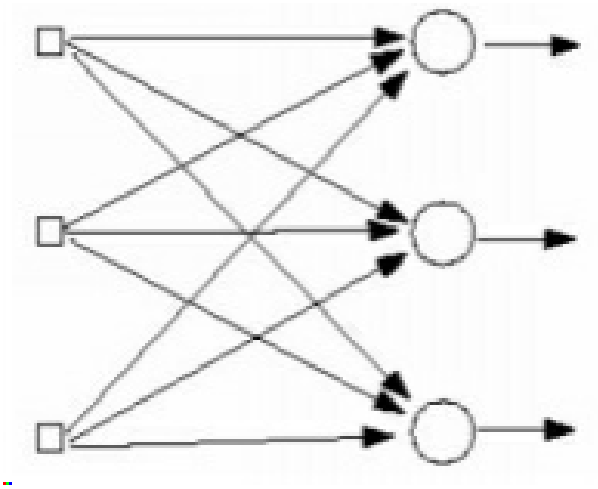

FIGURE 2.21 – Exemple d'un réseau de neurones non bouclé [48].

2- Propagation vers l'avant multicouche (Multi layer feed forward) :

Cette architecture est définie par une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie, où chacune d'entre elle à sa propre fonction. La couche d'entrée sert à récupérer les signaux du monde extérieur et à les redistribuer à tous les neurones de la première couche cachée. Tandis que les couches cachées servent à extraire les caractéristiques cachées et à les transférer à la couche cachée suivante jusqu'à y arriver à la couche de sortie pour déterminer le modèle de sortie, il s'agit donc d'un réseau de neurones de type Feed Forward[47]. Cette architecture n'est pas similaire à celle que nous avons vu auparavant puisque le nombre de couches n'est pas le même et c'est pour cette raison qu'il est plus puissant puisqu'il augmente

l'interaction entre les neurones [48].

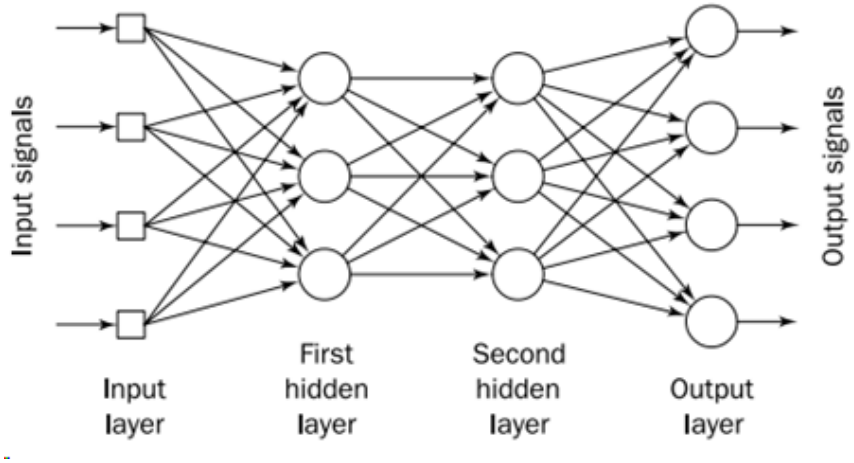

FIGURE 2.22 – Exemple d'un réseau de neurones non bouclé multicouche [47].

3- Réseau récurrent (Feed back) :

Les réseaux de neurones récurrents est un type de réseau de neurones où les sorties de la phase précédente sont sauvegardées dans une mémoire, ensuite ils sont transmis comme entrée dans la phase actuelle. Dans les réseaux de neurones typique, les sorties précédentes ne sont pas prises en considération pour générer la sortie actuelle tandis qu'il existe dans la plupart des applications actuelle une dépendance temporelle et que la sortie précédente va servir comme entrée qui va aider le programme à générer la sortie actuelle d'une façon plus précise, parmis ces applications nous pouvons citer : la reconnaissance vocale, traitement de langage naturelle, générateur de texte, traduction etc [49][50].

- L'algorithme de la descente du gradient :

Dans pratiquement tous les problèmes des réseaux de neurones, l'apprentissage se concentre sur la recherche de la bonne valeur des poids et des biais afin de minimiser une certaine fonction appelée "Cost Function" ou fonction de coût en français, cette fonction permet de mesurer l'erreur entre les résultats prédits et les résultats réels [50].

L'algorithme la descente du gradient est un algorithme d'optimisation itératif utilisé pour trouver une valeur appelée maximum ou minimum locale d'une fonction, cette fonction doit avoir deux principales exigences : il faut qu'elle soit dérivable et convexe. Cet algorithme est principalement utilisé dans l'apprentissage automatique et l'apprentissage profond mais nous pouvons le trouver également dans d'autres domaines comme : la robotique, les jeux d'ordinateur et le génie mécanique [51].

Les étapes d'application de l'algorithme la descente du gradient dans l'apprentissage profond sont les suivantes où **W** est un vecteur de poids :

1. Première itération : choisir des poids aléatoires pour tous les neurones **W0**.

2. Itération n+1 : les poids de cette itération **W**(n + 1) vont être mis à jour par rapport à l'itération précédente en utilisant cette formule :

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n - \gamma \nabla J(\boldsymbol{w}_n)$$

Avec ∇J(W) qui mesure la dérivée partielle de la fonction de coût par rapport à tous les composants du vecteur de poids W qui est définie comme suit :

$$\nabla J(\boldsymbol{w}) = \begin{pmatrix} \partial J(\boldsymbol{w}) / \partial w_1 \\ \vdots \\ \partial J(\boldsymbol{w}) / \partial w_{n_x} \end{pmatrix}$$

Et avec la valeur Y qui s'appelle le pas d'apprentissage, qui est une valeur paramétrable qui permet de déterminer à quelle point le modèle est adapté au problème, plus cette valeur est petite plus le modèle va effectuer des petits changements de poids et plus il aura besoin d'entraînements. Si la valeur de pas d'apprentissage est grande plus il va effectuer des changements de poids rapide et va nécessiter moins d'entraînements.

3. Le critère d'arrêt est défini soit par un seuil qu'il ne faut pas dépasser (le problème avec cette méthode c'est qu'elle est coûteuse en calcul) soit en définissant un certain nombre d'itérations fixé que nous pouvons par la suite l'augmenter si nous n'obtenons pas le résultat désiré [50].

2.5.2 Apprentissage profond

- Définition :

L'apprentissage profond ou deep learning en anglais, est un sous-domaine de l'apprentissage automatique, qui a permis des avancées importantes en intelligence artificielle dans les dernières années.

Un modèle d'apprentissage profond est basé sur les réseaux de neurones artificiels, composé de un ou de plusieurs couches de neurones qui effectuent chacun des opérations simples. Les résultats des neurones de la première couche sont utilisés comme entrée pour les calculs de la deuxième couche, et ainsi de suite [52].

Le développement du deep learning fut motivé pour développer les limites des algorithmes d'apprentissage automatique dans quelques tâches de l'IA après l'apparition de Big Data [53].

La figure ci-dessous montre l'architecture d'un modèle d'apprentissage profond :

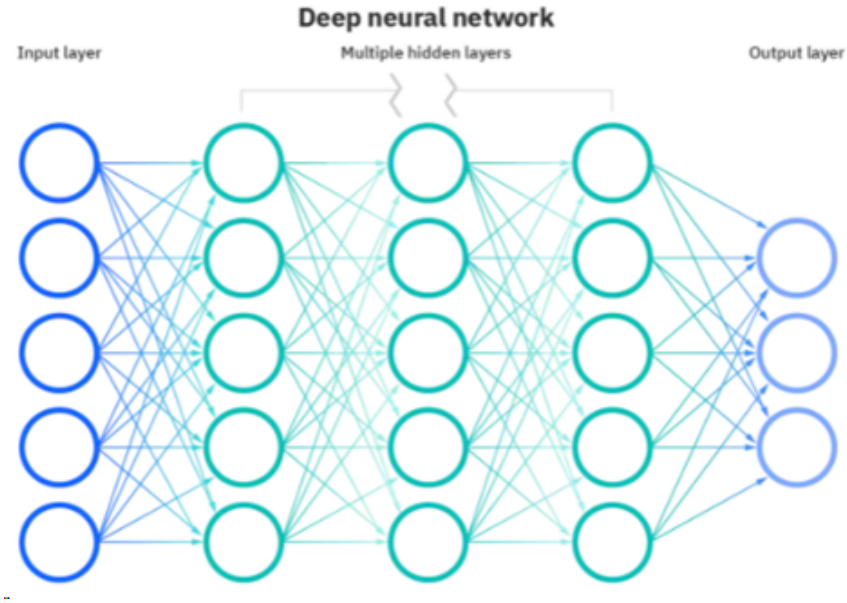

FIGURE 2.23 – Modèle de l'apprentissage profond [54].

- Principe de fonctionnement :

Comme précédemment indiqué dans la définition, l'apprentissage profond se base sur les réseaux de neurones artificiels inspirés du cerveau humain. Un modèle d'apprentissage profond est constitué de couches de dizaines ou centaines de neurones, les couches sont regroupées en trois types :

1- La couche d'entrée : son rôle est de recevoir les données d'entrée et de les transmettre à la première couche masquée.

2- Les couches cachées : dans cette couche, chaque nœud effectue des calculs mathématiques sur les entrées (fonction d'addition et d'activation) et transmet le résultat obtenu comme entrée à la couche suivante.

L'un des défis de la construction d'un modèle d'apprentissage profond est de déterminer le nombre de couches cachées et le nombre de neurones dans chaque couche.

3-La couche de sortie : une fois que l'ensemble de données d'entrée a traversé toutes les couches, la couche de sortie renvoie le résultat final.

Un modèle n'est toujours pas entraîné, ses sorties seront erronées. Pour entraîner un modèle d'apprentissage profond nous avons besoin :

- D'une grande puissance de calcul.

- D'une grande base d'apprentissage étiquetée, l'étiquetage nous aide à faire une comparaison entre le résultat (sortie) obtenu par le modèle et les sorties de l'ensemble de données (sortie réelle).

Une fonction du coût est utilisée pour mesurer la différence entre la valeur réelle et la valeur prévue, dans le cas où la valeur de la fonction est nulle c'est alors que les sorties de notre modèle sont les mêmes que les sorties de l'ensemble de données, dans le cas contraire la réduction de la valeur de la fonction du coût est nécessaire.

Pour réduire la fonction du coût, la technique de descente de gradient est utilisée pour ajuster les poids [55].

2.6 Domaine d'application de l'intelligence artificielle

L'intelligence artificielle est utilisée dans une variété de domaines tels que :

1- La santé :

Grâce aux techniques d'apprentissage profond, le secteur médical a eu un grand développement dans les méthodes de diagnostic et aussi dans la détection des maladies [56]. Voici un exemple d'application de l'intelligence artificielle en ophtalmologie : Google a développé avec des ophtalmologistes un modèle pour détecter la rétinopathie diabétique qui est une des complications causées par le diabète et qui peut mener à la cécité si les personnes atteintes du diabète ne font pas le dépistage précoce. Cette solution peut aider les médecins à faire le dépistage de la rétinopathie dans les pays qui n'ont pas suffisamment de spécialistes de vue comme l'inde et la thaïlande [57].

2- La finance :

L'intelligence artificielle s'est, depuis longtemps, invitée dans le secteur bancaire, grâce aux résultats du traitement automatisé des données. Les interventions d'IA permettent de regrouper les données d'une façon rapide, y compris celles qui peuvent être fournies par le client et de prédire des scénarios amélioreront considérablement les prévisions [56].

Prenons l'exemple de l'entreprise avec l'une des plus gros employeurs dans les états unis environ 240 000 employés est la banque JP Morgan Chase et Co. JP Morgan qui a développé un programme qui s'appelle COIN basé sur le machine learning, l'intelligence artificielle et la reconnaissance d'image pour raccourcir le temps d'analyse des documents qui prenait manuellement 360 000 heures chaque années par ses employeurs pour des tâches qui sont plutôt banales. JP Morgan assume que cette évolution ne cause pas le chômage mais plutôt de libérer les gens à faire des choses plus importantes [58].

3- La sécurité :

L'intelligence artificielle est en pleine croissance dans le domaine de la sécurité, nous pouvons constater sa présence dans notre vie quotidienne, par exemple les logiciels de reconnaissance faciale, l'empreinte digitale, la reconnaissance vocale etc. l'IA est également utilisée pour la prédiction et la protection contre les cyberattaques en utilisant les techniques d'apprentissage

automatique [56].

4- Le commerce :

Les algorithmes de l'IA comme le machine learning permettent de faire de la recommandation des produits ou la génération du contenu personnalisé suivant les méthodes de classement ou de clustering, et en utilisant les données de comportement passé de l'utilisateurs comme les cookies. Ceci est utilisé pour améliorer les stratégies de vente [59].

2.7 Conclusion

Dans ce chapitre nous avons introduit les notions de l'intelligence artificielle, ses types et ses différentes branches, nous avons également présenté les techniques d'intelligence artificielle qui seront utilisées dans notre étude qui sont : les deux algorithmes d'apprentissage automatique ainsi que l'apprentissage profond. Le chapitre est terminé par les différentes applications de l'intelligence artificielle sur plusieurs secteurs.

Dans le chapitre qui suit, nous allons présenter quelques travaux de recherche sur l'application de l'apprentissage automatique ainsi que l'apprentissage profond sur la prédiction du diabète de type 2.

CHAPITRE 3

ÉTAT DE L'ART

3.1 Introduction

Dans le chapitre précédent nous avons présenté l'intelligence artificielle, ses différents types ainsi que les branches de cette technologie, ensuite nous avons parlé sur l'apprentissage automatique et les réseaux de neurones.

Dans la partie consacrée à l'apprentissage automatique, nous avons parlé sur les différents types de problèmes résolus par l'apprentissage automatique ainsi que les deux algorithmes les plus utilisés dans la classification.

Dans la partie consacrée à la représentation de réseaux de neurones, nous avons parlé sur la structure d'un réseau de neurones artificielle, type d'apprentissage, les fonctions d'activations et l'architecture de réseaux de neurones ensuite nous avons présenté le concept d'apprentissage profond et pour conclure nous avons indiqué les différents domaines d'application de l'intelligence artificielle.

Beaucoup d'efforts de recherche ont été consacrés au développement des systèmes efficaces d'aide à la décision dédiés aux patients diabétiques pour les aider à la gestion de cette maladie chronique. Dans ce chapitre, nous allons citer quelques travaux de recherche d'application de l'intelligence artificielle dans la prédiction du diabète de type 2.

3.2 la prédiction du diabète en utilisant l'apprentissage automatique

Les chercheurs ont appliqué différentes techniques d'apprentissage automatique pour la prédiction du diabète, afin d'améliorer la précision des systèmes de soins de santé, Voici un ensemble d'études qui ont utilisé l'apprentissage automatique dans la diabétologie :

3.2.1 Etude 01 : Random Forest, KNN, Naïve Bayes, et J48 [60]

- L'approche proposée :

Pour l'élaboration de cette étude, les deux bases d'apprentissage suivantes ont été utilisées : PIDD (Pima Indian Diabetes Dataset) qui implique 768 enregistrements et 8 caractéristiques avec une classe cible (indique si la personne est diabétique ou non) et la base d'apprentissage 130-US qui contient un ensemble de données hospitalières sur le diabète aux États-Unis et qui se compose de 93743 instances et 48 attributs.

L'approche proposée a pour but de classifier et de prédire la maladie du diabète, le principe est de combiner un ensemble de techniques d'apprentissage automatique pour avoir de meilleurs résultats, les techniques combinées sont les suivantes : les forêts aléatoires, KNN, Naïve Bayes et J48(arbre de décision). La figure ci-dessous représente le workflow de l'approche proposée :

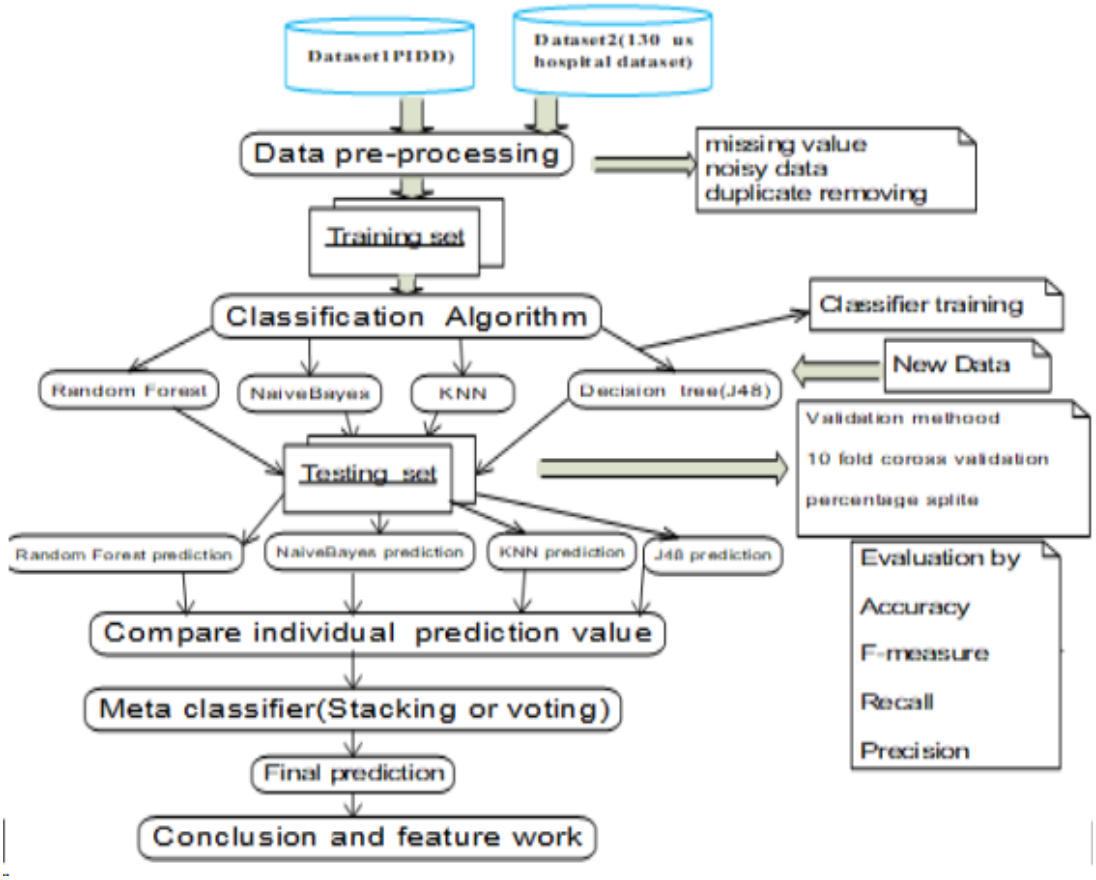

FIGURE 3.1 – workflow de l'approche proposée de l'étude 01 [60].

- Implémentation et résultats obtenus :

Dans cette étude applicative, l'ensemble de données est divisé en deux parties, les données d'entraînement (training data) et les données de test (testing data), ces deux parties constituant respectivement 90 % et 10 % de la base d'apprentissage.

Accuracy de l'approche d'ensemble proposée est de 93,62 % pour PIDD et 88,56 % pour l'ensemble de données d'hôpitaux 130-US. Ils ont conclu que les techniques d'ensemble offrent une meilleure précision que la méthode unique (une seule technique). Voici les résultats obtenus pour les deux base d'apprentissage :

Classification algorithm	Correctly classified	Incorrectly classified
J48	87.89	12.11
KNN	82.94	17.06
Naive Bayes	88.41	11.59
Random forest	89.84	10.16
Proposed ensemble using stacking	93.62	6.38

TABLE 3.1 – Les résultats de la précision des algorithmes de classification d'étude 01 en utilisant la base d'apprentissage PIDD [60]

Classification algorithm	accuracy before fs	incorrectly classified	accuracy after fs	incorrectly classified
J48	56.96	43.04	84.53	15.47
KNN	46.04	53.96	74.59	25.41
Naive Bayes	56.18	43.82	82.26	17.74
Random forest	48.68	51.32	82.80	17.20
Proposed ensemble using stacking	58.04	41.96	88.56	11.44

TABLE 3.2 – Les résultats de la précision des algorithmes de classification d'étude 01 en utilisant la base d'apprentissage 130-US [60].

3.2.2 Etude 02 : SVM [61]

- Principe de l'approche :

Cette étude se repose sur l'une des techniques de l'apprentissage automatique qui est SVM, le but de l'approche proposée est de prédire le diagnostic du diabète tout en se basant sur les attributs suivants : l'âge, l'indice de masse corporelle et la concentration du glucose dans le sang. L'application était sur 500 patients colombiens et sur un ensemble de données de patients d'une ethnie différente afin de savoir si le patient était diabétique, non diabétique ou en prédiabète.

- Implémentation et résultats obtenus :

Cette étude était établie conformément aux normes de l'OMS, dont 80 % de l'ensemble de données a été utilisé pour former un classificateur SVM non linéaire et les 20 % restants pour la phase du test. Le noyau utilisé à la fois pour l'entraînement et la prédiction était basé sur le radial et la méthode "10-fold cross-validation " a été employée pour valider le modèle de calcul. La performance du classificateur SVM a été mesuré selon les critères suivants :

- Précision.
- Sensibilité.
- Spécificité.
- Valeurs de prédiction positives et négatives.
- Matrice de confusion.

Le résultat du classificateur est obtenu avec une accuracy de 99,2 % sur des patients colombiens et avec une accuracy de 65,6 % sur l'ensemble de données des patients d'une ethnie différente.

3.2.3 Etude 03 : KNN [62]

- Principe de l'approche :

Dans cette étude, le diabète sucré était diagnostiqué à l'aide de l'algorithme du voisin le plus proche (KNN), dans lequel plusieurs paramètres ont été testés tels que le nombre de voisin (k) ainsi que les mesures de distances ou de similarités. Par la suite, une étude comparative était établie sur les résultats obtenus par les cinq différents algorithmes de classifications sur la base de données de 'Pima indian diabetes database' selon les critères suivants : Précision, rappel et f-measure.

- Implémentation et résultats obtenus :

La méthode du test utilisée dans cette étude est la technique d'échantillonnage Train et Test qui sert à réserver une partie de données pour l'apprentissage et une autre pour la phase test.

Avant de créer un classificateur basé sur l'algorithme KNN, Des différentes valeurs du "K" ont été testées (de 1 à 13) dont le meilleur score est enregistré au niveau de k=11. Ils ont aussi constaté que K=1 est le meilleur choix dans la phase du test contrairement à la phase d'entraînement ou la meilleure valeur de k est 11 d'où ils ont parlé sur le surapprentissage. Dans la partie suivante l'algorithme KNN est testé, en fixant k=11, sur les différentes distances suivantes : Euclidien, Manhattan, Minkowski, Hamming et Chebyshev. L'utilisation des métriques euclidien et Minkowski donne de meilleure performance.

Finalement, une série d'expériences étaient réalisées sur la base d'apprentissage Pima dans le but de faire une comparaison entre les techniques de classification suivantes : Knn, Svm,

Random Forest et D.trees. Les démarches de cette étude peuvent se résument dans la figure suivante :

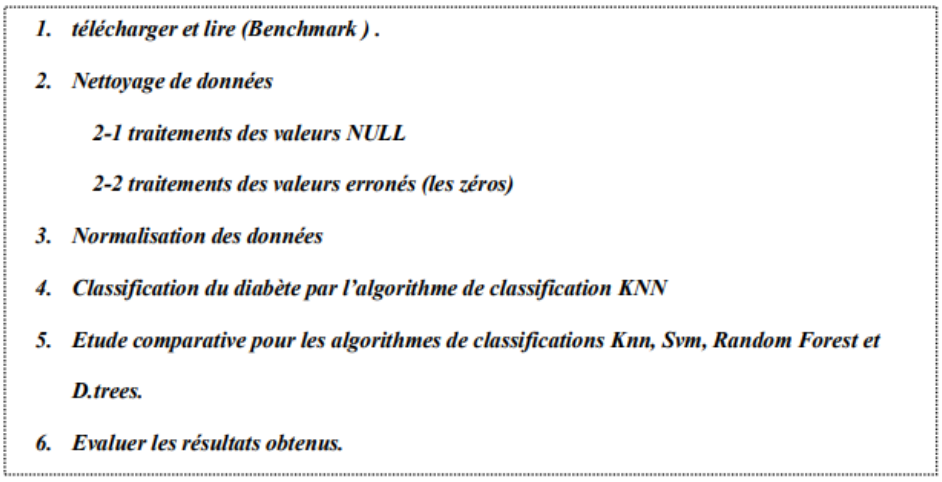

FIGURE 3.2 – démarche de l'approche proposée de l'étude 03 [62].

Le tableau 3.3 expose les résultats des expérimentations de ces 5 algorithmes de classification.

algorithme	précision	rappel	f1-score
KNN	0.68	0.61	0.64
SVM	0.67	0.04	0.08
Random forest classifier	0.63	0.56	0.60
GaussianNB	0.6	0.61	0.60
Tree	0.63	0.55	0.58

TABLE 3.3 – Les résultats des attributs d'évaluations pour les différents modèles [62].

Les résultats de performances ont montré clairement l'avance de l'algorithme KNN contre tous les autres algorithmes choisis dans cette étude.

3.3 la prédiction du diabète en utilisant l'apprentissage profond

3.3.1 Etude 01 : Réseau de neurones profond [63]

- Principe de l'approche :

Dans ce travail, les chercheurs ont proposé un système efficace de prise de décision médicale pour la prédiction du diabète basé sur les réseaux de neurones profonds (DNN). Ensuite, une étude comparative du DNN et de plusieurs méthodes d'apprentissage automatique a été établie.

Pour réaliser cette étude, ils ont utilisé un ensemble de données sur le diabète tiré de l'Hôpital de Francfort. Ce jeu de données contient 2000 enregistrements chacun avec 9 attributs. La figure 3.3 montre le workflow de l'approche proposée :

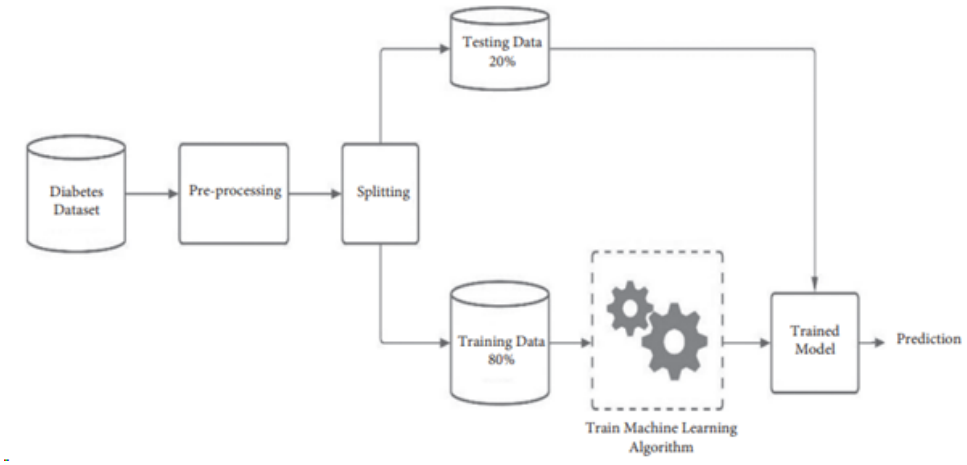

FIGURE 3.3 – workflow de l'approche proposée de l'étude 01 [63].

- Implémentation et résultat obtenu :

Le système proposé de prédiction de la maladie du diabète se compose d'un ensemble d'étapes liées les unes aux autres pour obtenir les résultats désirés. La première étape consiste à diviser l'ensemble de données en deux sous-ensembles, les données d'entraînement (80 %) et de test (20 %). La deuxième étape sert à appliquer deux catégories différentes (méthodes d'apprentissage automatique et d'apprentissage profond) afin de réaliser la phase d'entraînement, Par la suite une étude comparative ètait faite entre le modèle d'apprentissage profond proposé et les techniques d'apprentissage automatique.

Le prétraitement des données est fait à l'aide de la technique de la normalisation des données pour que les données puissent servir d'entrées aux algorithmes d'apprentissage automatique et à l'apprentissage profond.

Voici les méthodes de prédiction utilisées :

- Régression logistique.
- SVM.
- Extreme Gradient Boosting (XGBoost).
- Les arbres de décision.
- Les forêts aléatoires.
- DNN (modèle proposé).

Les métriques d'évaluation utilisées sont :

- Accuracy.
- Sensibilité.
- Spécificité.
- Précision.

- F1-score.

Après nombreuses expériences du changement dans le nombre de couches, le nombre de neurones dans chaque couche et les différents types de couches les chercheurs ont opté pour un modèle d'apprentissage profond avec les paramétres suivants Epochs=500, Batch-size =200, et Random-state=0.

Ce travail offre un modèle d'apprentissage profond avec une accuracy de 99,75 % et un score F1 de 99,66 % qui est supérieur aux accuracy des méthodes d'apprentissage automatique utilisées.

3.3.2 Etude 02 : Une approche d'apprentissage en profondeur [64]

- Principe de l'approche :

Les chercheurs proposent une stratégie de diagnostic du diabète à l'aide d'un réseau de neurones profonds par l'entraînement des attributs en mode de validation croisée quintuple et décuple.

- Implémentation et résultats obtenus :

Pour prédire le diabète sucré à l'aide de l'apprentissage profond, ils ont suivi ces étapes :

- Collecte de données : comme indiqué précédemment, la base d'apprentissage utilisée pour l'élaboration de ce travail est "Pima Indian diabetes".

- Préparation de données : Ils ont divisé les données en 2 façons en utilisant la validation croisée k-fold.

1- Five-fold cross-validation.

2- Ten-fold cross-validation.

- Mettre en œuvre un réseau de neurones profonds : ils ont choisi 4 couches cachées du réseau de neurones dont le nombre de neurones dans ceux-ci est respectivement 12,16,16,14 avec une couche d'entrée de 8 et la couche de sortie de 1.

- Critère d'évaluation : Pour visualiser les performances du modèle proposé, les critères suivantes sont utilisés : La précision, la sensibilité, la spécificité, score F1 et coefficient de corrélation de Matthews(MCC).

Le tableau ci-dessous montre les résultats obtenus :

Métriques d'évaluation	Méthode five-fold	Méthode ten-fold
accuracy	98.04	97.27
sensibilité	98.80	97.80
spécificité	96.64	96.27
f1 score	0.99	0.98
MCC	0.96	0.94

TABLE 3.4 – Mesures d'évaluation du système de prédiction du diabète [64].

Les résultats sur le jeu de données PID démontrent que l'approche d'apprentissage en profondeur conçoit un système de bon augure pour la prédiction du diabète avec une accuracy de prédiction de 98,35%, score F1 de 98% et MCC de 97% par la méthode Five-fold cross-validation. De plus, une précision de 97,11%, sensibilité de 96,25% et spécificité de 98,80% sont obtenues par la méthode Ten-fold cross-validation. L'expérimental résultats montrent que le système proposé fournit des résultats prometteurs en utilisant la méthode Five-fold cross-validation.

3.4 Discussion et critique

Le résultat obtenu par l'approche ensembliste proposée par l'étude de la réference [60] ne peut pas étre nier, cependant cette approche consomme beaucoup de temps et de ressource en pratique pour donner le résultat de la prédiction. De plus, Un manque d'information sur la sélection d'attribut qui a été faite sur la base d'apprentissage 130-US.

Pour l'étude de la réference [61], La taille de la base d'apprentissage (500 enregistrements) est insufisante pour que le modèle sois réellement utilisé.

Dans notre étude, nous avons choisi d'utiliser les deux algorithmes les plus performants dans la classification selon les études existantes, ainsi que l'utilisation de la sélection d'attribut afin de créer un modèle d'apprentissage profond prenant en considération les critiques mentionnées précédemment.

3.5 conclusion

Dans ce chapitre nous avons cité quelques techniques récentes développées pour le diagnostic du diabète sucré. Dans le chapitre qui suit, nous allons présenter en détail notre approche pour prédire le diabète de type 2 à l'aide des techniques d'intelligence artificielle (apprentissage automatique et apprentissage profond).

CHAPITRE 4

IMPLÉMENTATION DES ALGORITHMES ET PRÉSENTATION DES RÉSULTATS

4.1 Introduction

Dans ce dernier chapitre nous allons présenter notre contribution à la prédiction du diabète de type 2 en utilisant les techniques de l'intelligence artificielle (apprentissage automatique et apprentissage profond), en commençant d'abord par une présentation des logiciels et des librairies utilisés ensuite nous allons décrire la base de données que nous avons choisie pour enfin présenter l'implémentation des deux algorithmes de l'apprentissage automatique (Random Forest et SVM) ainsi que l'implémentation de notre modèle d'apprentissage profond qui sera basé sur les attributs les plus influents de la base de données que nous avons sélectionné.

4.2 Logiciels et librairies utilisés dans l'implémentation

4.2.1 Python

Python est un langage de programmation interprété, orienté objet et interactif. c'est l'un des langages les plus utilisés dans l'industrie avec des millions d'utilisateurs car il est très puissant et facile à apprendre avec une syntaxe simple et claire. Il prend en charge plusieurs paradigmes de programmation pas seulement l'orienté objet mais aussi la programmation procédurale et fonctionnelle, Python est également portable : il fonctionne sur différentes variantes d'Unix , de macOS ainsi que Windows. Ce langage est très utilisé dans les domaines de data science surtout en domaine d'apprentissage profond, il offre une variété de librairies qui rend la programmation et l'implémentation des algorithmes plus facile.

4.2.2 Tensorflow

Tensorflow est une plateforme Open-Source basée sur l'approche du passage en production qui propose plein d'outils, de bibliothèques et de ressources dédiés au Machine Learning. Son architecture simple et flexible permet au débutant ainsi qu'aux experts de développer facilement et d'avancer dans le domaine de Machine Learning que ce soit sur des serveurs , des appareils de périphériques ou sur le web.

4.2.3 Keras

Keras est une API de réseaux de neurones artificiels la plus utilisée selon le classement de Kaggle, elle offre une meilleure expérience d'utilisation grâce à ses simples API qui minimise les actions d'utilisateurs très fréquente et grâce à la convivialité et l'extensibilité. Il fournit aussi des messages d'erreurs simples à comprendre et à exploiter avec une documentation vaste et des guides de développement.

4.2.4 Weka

Weka est un logiciel Open-Source écrit en java conçu pour le prétraitement des données et l'implémentation de plusieurs algorithmes de Machine Learning ainsi que la visualisation des données comme : la Classification, le clustering, l'association, la sélection d'attributs. Après avoir choisi une de ces options, Weka offre une vue statistique des résultats grâce aux outils de visualisation pour mieux inspecter les données. Son utilisation facilite la tâche de développement des techniques d'apprentissage automatique et leur application dans les problèmes de Data mining.

4.3 Définition de l'ensemble de données utilisé et description des variables

4.3.1 Description de la base de données utilisée

Dans ce mémoire, nous avons choisie la base de données Pima Indian Diabetes, disponible sur le site officiel de Kaggle [65]. Qui a été prise originalement par Vincent Sigillito membre de l'institut national du diabète et maladies digestives et rénale, chef de groupe RMI Laboratoire de physique appliquée université Johns Hopkins, qui ont collecté ces données dans la région des indiens d'arizona Pima (voir figure 4.1).

Cette base de données est constituée de 768 cas des femmes indiennes Pima vivantes près de Phoenix Arizona qui se situe au Etats-Unis, ces femmes sont âgées de plus de 21 ans ou 500 d'entre elles sont non diabétiques (65.1 %) et 268 sont diabétiques (34.9%) qui est représenté

par le 9eme attribut de la base “Outcome” et les 8 autres attributs sont un ensemble de données médicales qui caractérisent le diabète que vous pouvez les voir dans le tableau 4.1.

FIGURE 4.1 – Zone de présence des indiens Pima [4].

4.3.2 Description des variables

La base de données contient 8 attributs de valeur numérique et 1 attribut de valeur booléenne “Outcome” qui est la variable prédictive indiquant si une personne est diabétique ou non, cette base de données contient aussi quelques valeurs manquantes qui sont prise en charge dans la phase de prétraitement de données. Le tableau suivant donne une description des attributs :

Nom d'attribut	Description	plage
Pregnancies	Nombre de fois enceinte	0 - 17
Glucose	Concentration du glucose plasmatique en 2 heurs (mg/dl)	0 - 199
Blood Pressure	Pression artérielle diastolique (mm Hg)	0 - 122
Skin Thickness	épaisseur de la peau au niveau du triceps (mm)	0 - 99
Insulin	Taux d'insuline au bout de 2 heurs	0 - 846
BMI	Body Mass Index (poid (kg)/ (hauteur(m))^2 , indique si une personne est obèse ou non ou une valeur plus de 25 est considérée comme en surpoids	0 - 67.1
Diabetes Pedigree Function	Fournit des informations sur les antécédents de diabète 2 chez le patient	0.078 - 2.42
Age	L'âge en année	21 - 81
Outcome	attribut qui indique si une personne est diabètique ou non (1 : diabètique , 0 : non diabètique)	268 sont diabètique 500 sont non diabètique

TABLE 4.1 – Description des attributs.

- Normes de la valeur du glucose :

Valeurs du glucose(g/l)	Interpretation
≤ 0.60	Très faible
0.61-0.80	Faible
0.81-1.40	Normal
1.41-1.80	Diabète précoce
≥ 1.81	Diabète

TABLE 4.2 – Les normes de glucose.

- Normes de la valeur du BMI :

valeurs du BMI	Interpretation
≤ 19	Maigre
19-24	Normal
25-30	En surpoid
31-40	Obèse
≥ 41	Très obèse

TABLE 4.3 – Les normes de BMI.

- Normes de la valeur du Blood pressure :

valeurs du Blood pressure	Interpretation
≤ 60	très faible
61-75	Faible
75-90	Normal
91-100	Élevé
≥ 100	Hypertension

TABLE 4.4 – Les normes de BloodPressure.

4.4 Implémentation et mise en œuvre

Pour l'élaboration de notre étude, nous avons choisi d'utiliser les techniques de l'apprentissage automatique et l'apprentissage profond afin de prédire le diabète de type 2 en utilisant la base d'apprentissage Pima Indian Diabetes Dataset.

Comme première étape, nous avons opté pour l'évaluation de deux algorithmes de machine learning selon trois critères (accuracy, précision et rappel) ainsi qu'une autre méthode d'évaluation de performance pour assurer le choix du meilleur modèle de prédiction.

Par la suite, nous avons fait une étude de sélection des attributs les plus importants de la base d'apprentissage dans l'objectif de créer un modèle d'apprentissage profond optimisé. La figure suivante montre les étapes suivies lors de l'élaboration de notre étude :

1. Téléchargement de la base de données
2. Prétraitement de données
3. Implémentation des modèles de machine Learning
   a. Implémentation du modèle SVM
   b. Implémentation du modèle Random Forest
   c. Sélection du meilleur modèle selon des critères d'évaluation
4. Évaluation de l'importance des variables en utilisant Weka et le meilleur modèle de machine Learning.
5. Sélection des attributs les plus influent.
6. Implémentation du Deep Learning en utilisant les attributs les plus influents.
7. Évaluation des résultats obtenu.

FIGURE 4.2 – Processus de notre approche proposée.

4.4.1 Prétraitement des données

La phase de prétraitement de données est cruciale lors de la création de n'importe quel modèle de machine learning et de data mining, elle consiste à nettoyer la base de données des données fausses, inexacte et à traiter les valeurs manquantes, elle consiste également à la réduction des données et des variables (que nous allons voir prochainement dans le chapitre) pour accroître la performance et la validité du modèle.

Pour cette base de données, il existe certaines valeurs manquantes des différents attributs qui sont remplacées par la valeur 0. Ces valeurs manquantes peuvent affecter l'efficacité lors de la création de nos modèles d'apprentissage automatique et d'apprentissage profond. Le tableau suivant montre le nombre des valeurs manquantes pour chaque attribut :

Nom d'attribut	**Valeurs manquantes**
Glucose	5
Blood Pressure	35
Skin Thickness	227
Insulin	374
BMI	11
Diabete Pedigree Function	0
Age	0

TABLE 4.5 – Le nombre de valeurs manquantes pour chaque attribut.

Nous pouvons constater d'aprés le tableau 4.5 que nous n'avons pas pris en considération l'attribut "Outcome" car c'est un attribut prédictif et que la valeur 0 dans cet attribut signifie qu'une personne n'est pas diabètique, ainsi que dans l'attribut "Pregnancies" n'était aussi pas pris en considération puisqu'il est logique qu'une femme n'a jamais tombé enceinte.

Pour traiter ces valeurs manquantes nous avons choisi de les remplacer par la moyenne de l'attribut correspondant. Le tableau ci-dessous montre la valeur moyenne de chaque attribut.

Nom d'attribut	**Valeurs moyennes**
Pregnancies	4
Glucose	120.895
Blood Pressure	69.105
Skin Thickness	20.536
Insulin	79.799
BMI	31.993
Diabete Pedigree Function	0.472
Age	33

TABLE 4.6 – La valeur moyenne de chaque attribut.

4.4.2 Implémentation des algorithmes de l'apprentissage automatique

Dans notre étude, nous avons choisi d'implémenter deux algorithmes d'apprentissage automatique les plus couramment utilisés en classification qui sont : Random Forest (RF) et Support Vector Machine (SVM) pour comparer les deux modèles et choisir le meilleur entre eux pour la phase suivante.

Les critères d'évaluation :

Les critères utilisés pour l'évaluation des deux modèles sont :

a. Accuracy : c'est une métrique utilisée dans les problèmes de classification, qui sert à calculer le pourcentage des prédictions correcte, il est calculé en divisant le nombre de prédictions correctes sur l'ensemble de toutes les prédictions

Accuracy =(Nombre des prédiction correcte)/(Nombre totale des Prédictions) il peut être calculé dans la classification binaire aussi par :

Accuracy = $\left(\frac{(VP+VN)}{(VP+FP+VN+FN)}\right)$

Avec VP, VN, FP, FN représente respectivement :

- Vrai positif : un diabètique classé diabètique
- Vrai négatif : un non diabètique classé non diabètique
- Faux positif : un non diabètique classé diabètique
- Faux négatif : un diabètique classé non diabètique

b. Precision : c'est une métrique qui mesure le nombre de prédictions pertinentes (vrai positif) parmi tous les exemples qui ont été prédits comme appartenant à la classe positive .

Précision = $\left(\frac{(VP)}{(VP+FP)}\right)$

c. Rappel : cette métrique mesure le nombre de prédictions positives par rapport à tous les exemples qui appartiennent réellement à la classe positive

Rappel = $\left(\frac{(VP)}{(VP+FN)}\right)$

B. Train-Test Split Evaluation :

Ces critères d'évaluation sont ensuite utilisés dans la technique Train-Test Split, c'est une technique d'évaluation qui sert à diviser la base de donnée en deux, Training data et Testing data, c'est-à-dire un morceau de la base de donnée sera utilisé pour former le modèle et un autre morceau sera utilisé pour tester l'efficacité du modèle.

L'objectif est de tester différentes configurations pour voir quel pourcentage de Training Data et Testing Data est meilleur selon les critères d'évaluation. Voici les trois configurations choisies pour comparer entre les deux modèles :

- Train 80%, Test 20% .
- Train 67%, Test 33% .
- Train 50%, Test 50% .

C. Implémentation de l'algorithme SVM :

Ce premier modèle d'apprentissage automatique a été construit en utilisant la bibliothèque Sklearn via la fonction " svm.SVC()" ou le mot svc signifie Support Vector Classifier car il y'a une autre fonction " svm.SVR()" pour les probèlmes de régression.

Nous avons utilisé la méthode " train-test-split" toujours de la même bibliothèque Sklearn pour effectuer les 3 configurations suivant ce code :

from sklearn.model-selection import train-test-split x-train, x-test, y-train, y-test = train-test-split(x,y, test-size = 0.2, stratify=y, random-state=2)

Avec les valeur de x-train, x-test, y-train, y-test signifie respectivement :

- x-train : ensemble des valeurs explicatives pour la collection d'entraînement
- x-test : ensemble des valeurs explicatives pour la collection de test
- y-train : ensemble des valeurs prédictives pour la collection d'entraînement
- y-test : ensemble des valeurs prédictives pour la collection de test

L'ensemble des valeurs prédictives de notre base de données sont ceux de la variable "Outcome" tandis que les valeurs explicatives sont ceux des 8 attributs restants. En ce qui concerne le paramétrage de ce modèle, l'utilisation du Kernel a été très bénéfique pour l'amélioration des performances, plus spécifiquement le kernel linéaire qui a été le plus adapté à notre base d'apprentissage que les autres types. La figure suivante montre les résultats d'accuracy selon les 3 configurations avec et sans utilisation du kernel.

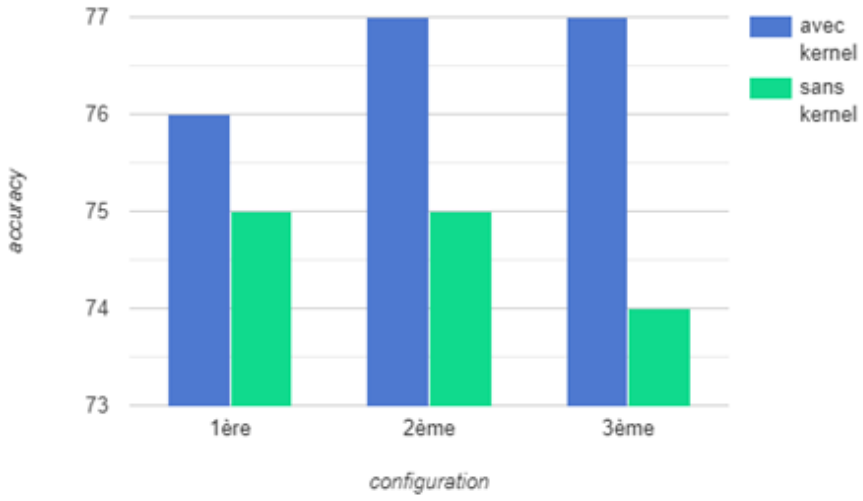

FIGURE 4.3 – Résultats d'accuracy score avec et sans kernel.

Dans le tableau qui suit, il est affiché les résultats des 3 métriques d'évaluation avec l'utilisation du kernel selon les 3 configurations.

Configuration	Accuracy	Precision	Recall
Train 80 %, Test 20%	76	76	73
Train 67 %, Test 33%	77	77	71
Train 50 %, Test 50%	77	78	71

TABLE 4.7 – Résultat du modèle SVM avec utilisation du Kernel.

Pour calculer l'accuracy et les autres métriques d'évaluation nous avons importé les métriques "accuracy-score" et "classification-report" de la bibliothèque Sklearn.

Nous pouvons constater que la configuration 50% pour l'ensemble d'entraînement et 50% pour l'ensemble de test est celle qui a donné le meilleur résultat pour ce modèle.

D. Implémentation de l'algorithme Random Forest :

Pour l'élaboration de ce modèle en appliquant la technique des forêts aléatoires, nous avons fait un appel à la fonction "RandomForestClassifier()" en utilisant toujours la bibliothéque Sklearn et la méthode "train-test-split" avec les paramètres suivant : random-state=7, le test-size est changé selon la configuration choisies. Signification de quelques paramètres d'entrées :

- n-estimators : prend une valeur entière qui désigne le nombre d'arbres dans la forêt, par défaut=100.

- criterion : La fonction pour mesurer la qualité d'une division. Les critères pris en charge sont "gini" pour l'impureté Gini et "entropie" pour le gain d'information. Ce paramètre est spécifique aux arbres de décision, par défaut="gini".

- max-depth : variable entière qui détermine la profondeur maximale de l'arbre de décision. par défaut="none"

- min-samples-split : exprime le nombre minimum d'échantillons requis pour diviser un nœud interne, par défaut=2.

- min-samples-leaf : c'est le nombre minimum d'échantillons requis pour être à un nœud feuille, par défaut = 1.

- max-features : par défaut="auto", désigne le nombre d'attributs à prendre en compte lors de la recherche de la meilleure répartition :

1- Si "auto", alors max-features=sqrt(n-features).

2- Si "sqrt", alors max-features=sqrt(n-features) (identique à "auto").

3- Si "log2", alors max-features=log2(n-features).

4- Si aucun, alors max-features=n-features.

- bootstrap : Si la valeur de cette variable booléenne est vrai alors les échantillons de bootstrap sont utilisés lors de la construction d'arbres. Sinon l'ensemble entier de données est utilisé pour

construire chaque arbre, par défaut="true".

- oob-score : S'il faut utiliser des échantillons hors sac (out of bag) pour estimer le score de généralisation. Uniquement disponible si bootstrap=True, par défaut="false" [66].

Dans notre implémentation, le changement était fait sur le paramètre d'entrée "n-estimators" afin d'augmenter l'accuracy du modèle dont la valeur choisie de "n-estimators" est 200.

Le tableau 4.8 montre les résultats obtenus des trois métriques avec les configurations choisies :

Configuration	**Accuracy**	**Précision**	**Rappel**
Train 80 %, Test 20%	81	82	80
Train 67 %, Test 33%	75	74	74
Train 50 %, Test 50%	78	77	76

TABLE 4.8 – Résultat du modèle random forest.

Nous pouvons constater que la première configuration de 80% de données pour l'ensemble d'entraînements et 20% pour l'ensemble de test est celle qui a donné le meilleur résultat dans toutes les métriques pour ce modèle.

E. Discussion des résultats obtenus :

Nous pouvons conclure d'après les résultats obtenus par les deux modèles d'apprentissage automatique SVM et Random Forest (présentés dans les tableaux 4.7 et 4.8) que l'algorithme Random Forest a donné de meilleurs résultats pour la classification binaire surtout pour la première configuration qui a donné 81% pour l'accuracy, 82% pour la précision et 80% pour le rappel est meilleur que tous les résultats du modèle SVM dans toutes les configurations. La figure suivante montre les différents scores d'accuracy pour les deux modèles Random Forest et SVM selon les trois configurations.

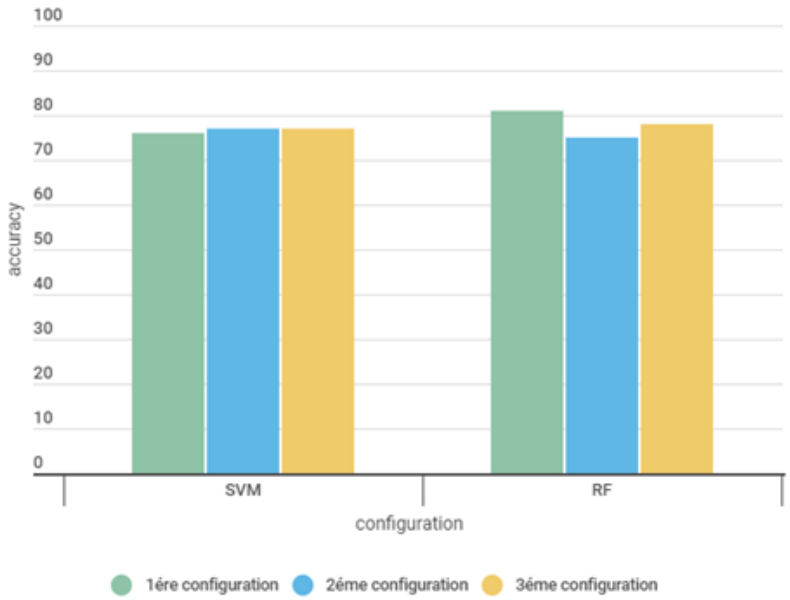

FIGURE 4.4 – Accuracy score pour les deux modèles SVM et Random Forest.

Nous choisissons alors le modèle Random Forest pour la phase de sélection des attributs les plus influents.

4.4.3 Sélection d'attributs

La sélection d'attributs (feature selection en anglais) est une étape qui nous permet de choisir correctement les attributs influents de la base de données dans l'objectif de réduire le nombre d'entrées pour notre modèle de deep learning et de ne prendre que les attributs qui ont une grande influence comme entrée.

Cette technique est utilisée pour réduire le nombre d'entrées et d'éliminer les attributs redondants, non pertinents et bruyants qui peuvent affecter négativement la phase d'apprentissage dans les modèles de machine learning et de deep learning afin de tirer le meilleur substitut d'attributs de la base d'apprentissage. Parmi les avantages de l'utilisation de la sélection d'attributs nous avons :

- dimensionnalité réduite : moins de dimensions signifie que le modèle est plus optimisé.

- Amélioration au niveau d'accuracy : quand il y'a moins d'attributs non pertinents signifie que l'accuracy du modèle va s'améliorer

- Réduction des coûts de calcul : moins d'entrées signifie que le modèle va s'entraîner plus rapidement [67]

plusieurs techniques sont utilisées pour la sélection d'attributs, pour notre étude nous avons choisi d'utiliser le logiciel Weka qui permet de faire une sélection d'attributs automatique

et de le comparer avec le résultat de " feature importance " qui a été pris après la construction du modèle de Random Forest.

La sélection d'attributs en utilisant Random Forest :

La sélection d'attribut dans notre modèle de Random Forest a été exécutée à l'aide de la bibliothèque Sklearn en utilisant l'instruction rfc.feature_importances_. Ou rfc est le nom de notre classifieur et feature_importances_ est la fonction utilisée. Cette fonction calcule la diminution de l'impureté du nœud pondéré par la probabilité d'atteindre ce nœud. Ou la probabilité d'atteindre le nœud peut être calculée par le nombre de samples qui atteignent le nœud divisé par le nombre total des samples. Plus la valeur est grande, plus l'attribut a une grande importance. Le résultat obtenu est le suivant :

[0.07684946, 0.25643635, 0.08952599, 0.08437176, 0.08552636, 0.14911634, 0.11751284, 0.1406609]

Pour mieux visualiser les résultats nous avons utilisé l'instruction suivante à l'aide de la bibliothèque Pandas :

(pd.Series(rfc.feature-importances-,index=x.columns).plot(kind='barh'))

Résultat :

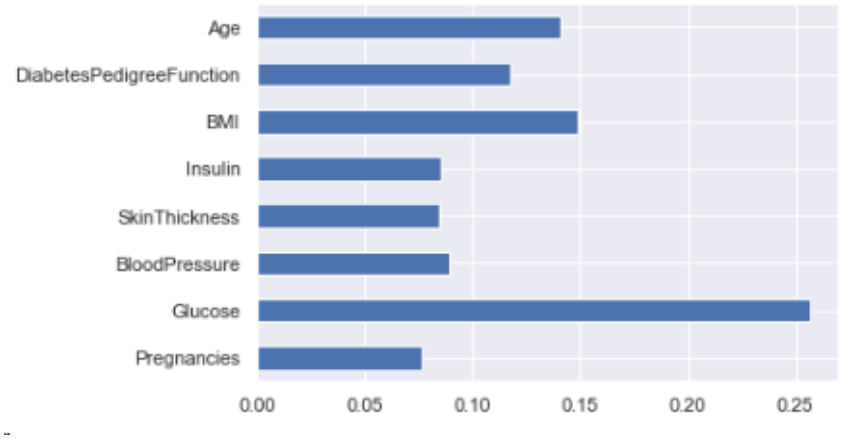

FIGURE 4.5 – La sélection d'attributs selon le modèle Random Forest

- La sélection d'attributs en utilisant l'outil Weka :

Pour effectuer la selection d'attribut "Feature Selection" en utilisant l'outil Weka, il faut passer par l'onglet "Select Attribut" et spécifier deux paramètres :

- Attribut Evaluator : c'est une technique qui permet d'évaluer les attributs de la base de données dans le contexte de la variable de sortie. Dans notre étude, nous avons choisi la fonction CfsSubsetEval comme "attribut evaluator" :

CfsSubsetEval : c'est une fonction qui permet d'extraire un ensemble d'attribut dont ils sont fortement corrélés avec la classe en ayant une faible intercorrélation sont préférés.

- Search Method : c'est une technique qui permet de naviguer plusieurs combinaisons possibles d'attributs afin d'arriver à une courte liste d'entités choisies, la fonction de recherche choisie pour notre étude est :

BestFirst : elle cherche l'espace d'un sous ensemble d'attribut en utilisant greedy hillclimbing augmenté d'une fonction de retour en arrière.

Le résultat obtenu est :

```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 46
        Merit of best subset found:    0.517

Attribute Subset Evaluator (supervised, Class (numeric): 9 Outcome):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,2,6,7,8 : 5
                     Pregnancies
                     Glucose
                     BMI
                     DiabetesPedigreeFunction
                     Age
```

FIGURE 4.6 – Les attributs sélectionnés par Weka

Les résultats suivants montrent qu'il y a une substitution de cinq attributs (Pregnancies, Glucose, BMI, DiabetesPedigreeFunction, Age) comme ceux qui ont une forte corrélation entre classe et moins d'intercorrélation avec un mérite de 0.51.

- Interprétation des résultats :

Les résultats obtenus après l'exécution de la fonction Feature Importances du modèle de Random Forest nous montrent la distribution des attributs selon leur degré d'influence lors de la construction de notre classificateur Random Forest que nous pouvons les classer comme suit du plus influent jusqu'au moins influent :

[Glucose, BMI, Age, DiabetesPedigreeFunction, BloodPressure, Insulin, SkinThickness, Pregnancies].

Tandis que dans l'outil Weka les résultats ont trouver une substitution de cinq variables qui sont les plus corrélées entre classe dont nous pouvons constater qu'elle contient parmis eux les quatres variables les plus influents selon les résultats de la fonction "feature-importances-" de Random Forest (Glucose, BMI, Age, DiabetesPedigreeFunction).

Dans notre étude, nous avons choisi de créer un modèle de Deep Learning plus optimisé avec un nombre d'entrée plus réduit pour les raisons déjà expliquées avec Feature Selection. Les attributs que nous avons sélectionnés sont : (Glucose, BMI, Age, DiabetesPedigreeFunction) qui sont les attributs les plus influent de Random Forest et qui se trouve aussi dans la substitution des attributs les plus corrélés de Weka en négligeant l'attribut pregnancies pour les deux raisons suivantes :

1. C'est l'attribut le moins influent selon Random Forest

2. C'est un attribut qui pourra être utilisé qu'avec les femmes dans notre étude nous avons voulu créer un modèle de Deep Learning pour la prédiction du diabète de type 2 optimisé qui sera utilisable pour les deux sexes.

4.4.4 Implémentation du modèle d'apprentissage profond

L'implémentation de ce modèle a été faite avec les quatres attributs que nous avons sélectionné à partir de la phase précédente (Glucose, BMI, DiabetesPedigreeFunction, Age) à l'aide de la bibliothèque Keras avec une configuration de 80 % pour l'ensemble d'entraînement et 20 % pour l'ensemble de test en utilisant la bibliothèque Sklearn.

La construction et l'amélioration de l'accuracy de notre modèle a été faite en jouant sur plusieurs paramètres, nous pouvons citer :

- Topologie du réseau : parfois un nombre de couches réduit avec beaucoups de nœuds par couche est meilleur et parfois plusieur couches avec moins de nœuds par couche est favorable c'est lié au problème.

- Le type de fonction d'activation : chaque fonction d'activation est meilleur pour un type de problème, il est possible de combiner plusieurs fonctions dans l'architecture du modèle.

- Le choix de la fonction de perte : pareil pour la fonction de perte, chaque fonction est

préférable pour chaque type de problème

- La taille du batch : signifie le nombre de samples à traiter avant de mettre à jour les poids.

- Le nombre des epochs : signifie le nombre de fois ou le modèle va s'entraîner sur toute la base d'apprentissage batch par batch.

L'architecture que nous avons utilisée dans notre approche est :

- La couche d'entrée : 4 noeuds d'entrées avec la fonction d'activation "Sigmoid".
- 1ère couche cachée : 12 noeuds avec la fonction d'activation "Sigmoid".
- 2ème couche cachée :8 noeuds avec la fonction d'activation "Sigmoid".
- 3ème couche cachée :4 noeuds avec la fonction d'activation "Sigmoid".
- La couche de sortie : 1 noeud de sortie avec la fonction d'activation "Sigmoid".

La fonction Sigmoid a été utilisée dans toutes les couches car c'est celle qui a donnée un meilleur résultat par rapport à d'autres fonctions que nous avons testées, elle est aussi la plus recommandée pour les problèmes de classification binaire ainsi que la fonction de perte "binary cross-entropy" que nous avons utilisée.

Par rapport à la taille de batch et le nombre d'epochs utilisé, nous avons choisi la taille 16 pour le batch avec le nombre d'epochs 1000, nous avons aussi utilisé l'optimiseur Adam pour le Backpropagation. Les résultats d'exécution de notre modèle Deep Learning sont :

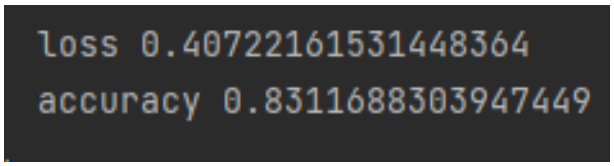

FIGURE 4.7 – Résultats d'exécution de notre modèle d'apprentissage profond.

L'enregistrement du modèle finalisé a été effectué à l'aide de la fonction " load-model" de la bibliothèque Keras sous format "h5", qui nous a permis de sauvegarder à chaque fois le modèle avec qui nous avons trouvé un bon résultat jusqu'à ce que nous avons trouver celui de 83 % d'accuracy pour ensuite avoir la possibilité de le réutiliser, faire des prédictions avec et d'afficher à nouveau la valeur de loss et d'accuracy.

- Discussion des résultats obtenus :

Les résultats obtenus de notre modèle d'apprentissage profond nous a permis de constater qu'il est possible d'obtenir un modèle avec un bon résultat d'accuracy compétitif aux modèles d'apprentissage automatique avec moins d'attributs, c'est-à-dire avec la bonne combinaison d'attributs et avec la bonne configuration d'apprentissage profond nous pouvons avoir des bons résultats.

Nous pouvons aussi constater que la phase de sélection d'attributs est très bénéfique pour améliorer les performances d'un modèle surtout pour les bases de données avec un nombre

d'attribut énorme cette phase peut apporter des bénéfices très remarquables.

En ce qui concerne l'utilisation de notre modèle d'apprentissage profond pour la prédiction du diabète de type 2 pour les femmes ainsi que pour les hommes est possible, car dans les deux attributs suivants : Glucose, BMI les valeurs de plage des hommes diabétiques sont inclus dans notre base de données. Nous pouvons constater dans le tableau suivant que les valeurs de plage des deux attributs Glucose et BMI dans notre base de données (après avoir effectué le nettoyage de données) couvrent bien toute les normes du Glucose et de BMI (voir tableau 4.2 et 4.3)

Attribut	**Valeurs de plage dans notre base de données**
Glucose(g/l)	44 - 199
BMI(kg/m^2)	18.2 - 67.1

TABLE 4.9 – Les valeurs de plage des deux attributs dans notre base de données.

Et par rapport l'attribut "DiabetesPedigreeFunction" c'est une fonction qui fournit des informations sur les antécédents diabétiques chez le patient ou plus la valeur de cet attribut est grande plus le patient est susceptible d'avoir le diabète. La méthode pour calculer cette valeur n'est pas fourni mais il est clair que c'est une valeur indépendante du sexe du patient tout comme l'âge donc nous pouvons déduire que notre modèle peut être utilisable pour les deux sexes.

4.5 Conclusion

Dans ce chapitre nous avons présenté notre contribution par rapport à la prédiction du diabète de type 2 de la base donnée Indian Pima Diabetes en utilisant les techniques d'intelligence artificielle. La première est l'utilisation des algorithmes d'apprentissage automatique SVM et Random Forest avec une étude comparative entre eux qui à amener à conclure que le modèle de Random Forest est meilleur selon toutes les métriques d'évaluation.

La seconde été en utilisant l'apprentissage profond avec les quatres attributs les plus influents selon Random Forest et Weka qui nous a permis d'avoir un modèle avec 83% d'accuracy et qui peut être utilisable pour les deux sexes.

CONCLUSION GÉNÉRALE ET PERSPECTIVES

Beaucoup de chercheurs ont mené des travaux afin d'effectuer la prédiction du diabète sucré. Dans ce mémoire de master nous avons choisi d'utiliser les techniques de l'apprentissage automatique et l'apprentissage profond sur la base de données Pima Indian Diabetes, en ce qui concerne l'apprentissage automatique nous avons implémenté deux algorithmes SVM et Random Forest, nous avons conclu suivant les métriques d'évaluation (accuracy, précision, rappel) et suivant trois configurations différentes de fragmentation de données que le modèle Random Forest été meilleur que celui de SVM avec un pourcentage de 81 % d'accuracy, 82 % de précision et 80 % de rappel par rapport à 77% d'accuracy et 78% de précision et 71% de rappel.

Nous avons ensuite utilisé la méthode de sélection d'attributs les plus influents de la base de données pour l'élaboration de notre modèle d'apprentissage profond afin d'optimiser notre modèle. Pour cela nous avons utilisé le modèle Random Forest pour la sélection des attributs les plus importants et le logiciel Weka. Après plusieurs essais d'architectures d'apprentissage profond et après avoir changé plusieurs paramètres nous avons obtenu un modèle avec un pourcentage 83 % d'accuracy qui peut être utilisable pour les deux sexes après avoir éliminé quatre attributs où un parmis eux est "Pregnancies" qui est dédié uniquement aux femmes.

Les principales perspectives qui apparaissent à l'issue de notre travail sont : l'utilisation de la technique de sélection d'attributs sur d'autres base d'apprentissage qui contiennent un nombre important d'attributs pour voir son impact, tester différentes combinaisons d'attributs pour l'apprentissage profond tout en essayant d'améliorer l'accuracy ainsi nous souhaitons le développement d'une application web ou mobile qui permet de mesurer le pourcentage d'atteinte du diabète dans le futur et qui sera accessible pour tout le monde.

BIBLIOGRAPHIE

[1] Organisation mondiale de la santé. le diabète. https://www.who.int/fr/news-room/fact-sheets/detail/diabetes. consulté le 17 octobre 2021.

[2] International Diabetes Federation (IDF). *L'Atlas Du Diabète 9ème édition*. 2019.

[3] Fédération française de cardiologie. réduire le risque cardio-vasculaire le diabete. https://www.fedecardio.org/je-minforme/le-diabete/. consulté le 15 octobre 2021.

[4] Mlle. SAIDI Meryem. Traitement de données médicales par un système immunitaire artificiel, reconnaissance automatique du diabète. *Mémoire de Magister en Informatique, Université ABOUBAKR BELKAID–TLEMCEN*, 2011.

[5] Vidal, diabète de type 2. https://www.vidal.fr/maladies/metabolisme-diabete/diabete-type-2.html. consulté le 20 novembre 2021.

[6] Le figaro. diabète : quel sont les examens pour le diagnostiquer. https://sante.lefigaro.fr/article/diabete-quels-sont-les-examens-pour-le-diagnostiquer-/. consulté le 19 octobre 2021.

[7] Sahnine nabil et Yahiaoui Yacine. Analyse des moyens à mettre en œuvre pour lutter contre le diabète. *Cas CHU l'hopital belloua Tizi- Ouzou, mémoire , master en science ´economique , université mouloud mammeri Tizi-ouzou*, 2017.

[8] Vidal. diabète gestationnel. https://www.vidal.fr/maladies/metabolisme-diabete/diabete-gestationnel.html. consulté le 24 octobre 2021.

[9] Ceed : Centre européen d'étude du diabète. diabetes et complications. http://ceed-diabete.org/fr/le-diabete/diabete-et-complications/. consulté le 24 octobre 2021.

[10] Diabète Québec. comprendre le diabète. https://www.diabete.qc.ca/fr/comprendre-le-diabete/tout-sur-lediabete/complications/la-nephropathie/. consulté le 17 novembre 2021.

[11] Fédération française des diabètiques. comprendre le diabète. https://www.federationdesdiabetiques.org/information/rechercheinnovations-diabete/actualites/le-prediabete. consulté le 17 novembre 2021.

[12] Diabète Québec. le prédiabète. https://www.diabete.qc.ca/fr/comprendre-le-diabete/tout-sur-le-diabete/types-dediabete/le-prediabete/. consulté le 19 novembre 2021.

[13] VIDAL. La prise en charge du diabète de type 2. https://www.vidal.fr/maladies/metabolisme-diabete/diabete-type-2/traitement.html. consulté le 15 novembre 2021.

[14] L'assurance maladie. Le traitement du diabète gestationnel. https://www.ameli.fr/assure/sante/themes/diabete-gestationnel/traitement-suivi-femme-enceinte-bebe. consulté le 26 novembre 2021.

[15] Passeport santé. Prédiabète : Ce qu'il faut savoir avant qu'il soit trop tard. https://www.passeportsante.net/fr/Actualites/Dossiers/DossierComplexe.aspx?doc=prediabete. consulté le 26 novembre 2021.

[16] Inserm : institut national de la santé et de la recherche médicale. diabète de type 1. https://www.inserm.fr/dossier/diabete-type-1/. consulté le 2 novembre 2021.

[17] Stuart Russell et al. *Artificial intelligence a modern approach, third edition*. 2010.

[18] IBM Cloud Education. Artificial intelligence. https://www.ibm.com/cloud/learn/what-is-artificial-intelligence. consulté le 21 novembre 2021.

[19] R. Saracco. Computers keep getting better . . . than us. *IEEE Future Directions*, 2018.

[20] Z Mohammed. Artificial intelligence definition, ethics and standards. *The British University In Egypt*, 2019.

[21] Journal Du Net. Système expert : définition, fonctionnement et exemples. https://www.journaldunet.fr/webtech/guide-de-l-intelligence-artificielle/1501897-systeme-expert-definitionfonctionnement-et-exemples/. consulté le 25 novembre 2021.

[22] Data Analytics. Qu'est ce que le machine learning. https://ia-data-analytics.fr/machine-learning/. consulté le 5 mars 2022.

[23] Frédéric Sur. *introduction 'a l'apprentissage automatique*. 2021.

[24] Chloé-Agathe Azencott. *Introduction au Machine Learning*. 2.

[25] Pensée Artificielle. Machine learning pour d´ebutant : Introduction au machine learning. https://penseeartificielle.fr/introduction-au-machine-learning/. consulté le 12 mars 2022.

[26] Morgane LAUR. Anticipation des changements de notes des obligations du portefeuille d'un assureur par méthode de machine learning. *master Actuariat et l'admission à l'Institut des Actuaires, Université Paris-Dauphine.*

[27] Analytics & Insights. Les types de regression. https://analyticsinsights.io/top-5-des-types-de-regression/. consulté le 22 mars 2022.

[28] Amit & Agbaje Moyinoluwa & José-Garcıa Adan & Olaide Oyelade Agushaka Ovre Ezugwu, Absalom & Shukla. Automatic clustering algorithms : a systematic review and bibliometric analysis of relevant literature. neural computing and applications. 2021.

[29] Cnam. estimation de densité. https://cedric.cnam.fr/vertigo/Cours/ml/coursEstimationDensite.html. consulté le 25 mars 2022.

[30] Analytics Vidhya. Introduction to pseudo-labelling : A semi-supervised learning technique. https://www.analyticsvidhya.com/blog/2017/09/pseudo-labelling-semi-supervised-learning-technique/. consulté le 1 avril 2022.

[31] Analytics Vidhya. Pseudo-labeling a simple semi-supervised learning method. https://www.analyticsvidhya.com/blog/2017/09/pseudo-labelling-semi-supervised-learning-technique/. consulté le 3 avril 2022.

[32] La revue IA. Apprentissage par renforcement. https://larevueia.fr/apprentissage-par-renforcement/. consulté le 22 mars 2022.

[33] Section. Introduction to random forest in machine learning. https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/. consulté le 29 mars 2022.

[34] JavaTPoint. Random forest algorithm. https://www.javatpoint.com/machine-learning-random-forest-algorithm. consulté le 27 mars 2022.

[35] Analytics Vidhya. Understanding random forest. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/. consulté le 1 avril 2022.

[36] Mariana & Corbellini Luis Machado, Gustavo & Recamonde-Mendoza. What variables are important in predicting bovine viral diarrhea virus ? a random forest approach. veterinary research. 2015.

[37] Sid Ahmed Amel et Rabhi Karima. La prédiction du diabète en utilisant les algorithmes de machine learning. *mémoire de master en informatique , Université de bouira*, 2020.

[38] Analytics Vidhya. Understanding support vector machine(svm) algorithm from examples (along with code). https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/. consulté le 26 février 2022.

[39] Towards Data Science. Support vector machines a brief overview. https://towardsdatascience.com/support-vectormachines-a-brief-overview-37e018ae310f. consulté le 27 février 2022.

[40] Towards Data Science. Support vector machine - introduction to machine learning algorithms. https://towardsdatascience.com/support-vector-machine-introduction-to-machinelearning-algorithms-934a444fca47. consulté le 27 février 2022.

[41] java T point. Support vector machine algorithm. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm. consulté le 1 mars 2022.

[42] Analytics Vidhya. Support vector machine (svm) a complete guide for beginners. https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/. consulté le 10 mars 2022.

[43] Analytics Vidhya. The a-z guide to support vector machine. https://www.analyticsvidhya.com/blog/2021/06/support-vector-machine-better-understanding/. consulté le 5 mars 2022.

[44] k. Gurney. *an introduction to neural networks*. 1997.

[45] Towards Data Science. Introducing deep learning and neural networks — deep learning for rookies (1). https://towardsdatascience.com/introducing-deep-learning-and-neural-networksdeep-learning-for-rookies-1-bd68f9cf588
consulté le 1 avril 2022.

[46] l.v. Fausett. *fundamentals of neural networks : architectures, algorithms and applications*. 1993.

[47] Michael Negnevitsky. *Artificial Intelligence , a guide to intelligent systems*. 2005.

[48] Pat Nakamoto. *Neural Networks & Deep Learning : Deep Learning explained to your granny – A visual introduction for beginners who want to maketheirownDeep Learning Neural Network (Machine Learning)*. 2017.

[49] Medium. Understanding recurrent neural network. https://medium.com/analyticsvidhya/understanding-rnns-652b7d77500e. consulté le 5 avril 2022.

[50] Umberto Michelucci. *Applied Deep Learning - A Case Based Approach To Understanding Deep Neural Networks*. 2018.

[51] Robert Kwiatkowski. Gradient descent algorithm — a deep dive. https://towardsdatascience.com/gradient-descent-algorithma-deep-dive-cf04e8115f21. consulté le 8 avril 2022.

[52] Boughaba Mohammed et Boukhris Brahim. L'apprentissage profond (deep learning) pour la classification et la recherche d'images par le contenu. *mémoire , master en informatique , université kasdi merbah ouargla*, 2017.

[53] Moualek Djelloul Youcef. Deep learning pour la classification des images. *mémoire de master en informatique, Université Abou Bakr Belkaid Tlemcen*, 2017.

[54] IBM. Neural networks. https://www.ibm.com/cloud/learn/neural-networks. consulté le 1 avril 2022.

[55] freecodecamp. Want to know how deep learning works ? here's a quick guide for everyone. https://www.freecodecamp.org/news/want-to-know-how-deep-learning-works-heres-a-quick-guide-for-everyone-1aedeca88076/. consulté le 5 avril 2021.

[56] hight tech. Les différents domaines d'application de l'intelligence artificielle. https://www.cmsinfo.org/high-tech/domaines-application-intelligence-artificielle/. consulté le 6 avril 2022.

[57] Google Health. Using artificial intelligence in ophthalmology. https://health.google/for-clinicians/ophthalmology/. consulté le 18 avril 2022.

[58] D. GALEON. Une ia a réalisé 360 000 heures de travail financier en quelques secondes. 2017.

[59] IBM Cloud Education. Artificial intelligence. https://www.ibm.com/cloud/learn/what-is-artificial-intelligence. consulté le 15 avril 2022.

[60] Rahul & Joshi Preeti & Mulay-Preeti Alehegn, Minyechil & Joshi. . diabetes analysis and prediction using random forest, knn, na¨ıve bayes, and j48 : An ensemble approach. 2020.

[61] Danelys Cabrerac Omar Bonerge Pinedad Amelec Viloriaa, Yaneth Herazo-Beltranb. Diabetes diagnostic prediction using vector support machines. 2020.

[62] Zouache hanen et Bendib ichrak. Classification du diabète avec l'algorithme knn. *mémoire , master en informatique , Universit´e de Bordj Bou Arreridj*, 2021.

[63] Youcef Brik Bilal Attallah Samir Brahim Belhaouari Tawfik Beghriche, Mohamed Djerioui. An efficient prediction system for diabetes disease based on deep neural network. *mémoire , master en informatique , Universit´e de Bordj Bou Arreridj*, 2021.

[64] Md. Milon Islam Safial Islam Ayon. Diabetes prediction : A deep learning approach. 2019.

[65] Kaggle, pima indians diabetes dataset. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. consulté le 15 avril.

[66] Scikitlearn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. consulté le 25 avril.

[67] T. R. N and R. Gupta. . feature selection techniques and its importance in machine learning : A survey. *IEEE International Students' Conference on Electrical*, 2020.