République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique
Université Mohamed el-Bachir el-Ibrahimi Bordj Bou Arréridj
Faculté de Mathématique et Informatique

MEMOIRE

Présente en vue de l'obtention du diplôme
Master en informatique

Spécialité : Technologie de l'information et de la communication.

THEME :

Un système de prédiction et de prévision du diabète de type 2

Présenté Par :

- Benbelouaer ghada.

Soutenue le :

Devant le jury composé de :

Président	Mr :
Examinateur	Mr :
Examinateur	Mr :
Encadreur	Dr: Saad Saoud Manel.

Année Universitaire 2020-2021

Remerciements

Je remercie tout d'abord le bon Dieu d'e m'avoir donné la santé et la volonté d'entamer et de terminer ce mémoire.

Je tiens à remercier également mon encadreur Dr. saad saoud manet d'avoir accepté de ma guider tout au long de ce travail.

Merci à l'ensemble des enseignants de l'université El Bachir El Ibrahimi de Bordj Bou Arreridj

Dédicaces

Je dédie ce travail à mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études.

A mes chères sœurs, Lilya, Sirine, Yakine pour leurs encouragements permanents, et leur soutien moral.

A ma sœur du cœur « Anissa » pour leur encouragement, et soutien.

A toute mes amies, pour l'amitié et les moments agréables que nous avons passés ensemble.

A toute ma famille pour leur soutien tout au long de mon parcours universitaire.

الملخص :

مـرض السـكري هـو مـرض مـزمن يمكـن أن يكـون سـببه عـدم قـدرة الجسـم علـى صـنع أو اسـتخدام الأنسـولين الـذي ينتجـه. مـع مـرور الوقـت، يمكـن للسـكري أن يضـر القلـب والأوعيـة الدمويـة والعينـين والكليتـين والأعصـاب ... الـخ الوقايـة مـن هـذا المـرض هـو موضـوع سـاخن للبحـث. وقـد اسـتخدمت العديـد من التقنيات لتطوير انظمة موثوقة للتنبؤ و لتشخيص مرض السكري.

الهـدف مـن هـذا العمـل هـو تحقيـق نظـام للتنبـؤ بالسـكري مـن النـوع 2 والوقايـة منـه باسـتخدام الشـبكات العصبية الاصطناعية (التنبؤ متعدد الطبقات).

الكلمات المفتاحية

السكري ، التنبؤ ، التعلم الآلي ، الشبكات العصبية الاصطناعية.

Résumé :

Le diabète est une maladie chronique qui peut être causée par l'incapacité du corps de produire ou d'utiliser l'insuline qu'il produit. Au fil du temps, le diabète peut endommager le cœur, les vaisseaux sanguins, les yeux, les reins et les nerfs ... etc. La prévention de cette maladie est un sujet de recherche brûlant. De nombreuses techniques ont été utilisées pour mettre au point un système de prévision fiable permettant de diagnostiquer le diabète.

L'objectif de ce travail est de réaliser un système de prédiction et de prévention du diabète type 2 en utilisant les réseaux de neurones artificiels (la prédiction multicouche).

Mots clés :

Diabète, prédiction, apprentissage automatique, réseaux de neurones artificiels.

Abstract:

Diabetes is a chronic disease that can be caused by the body's inability to make or to use the insulin produced. Over time, diabetes can damage the heart, blood vessels, eyes, kidneys and nerves ... etc. The prevention of this disease is a hot topic of research. Many techniques have been used to develop a reliable predictor system for diagnosing diabetes.

The objective of this work is to realize a system for predicting and preventing diabetes type 2 using artificial neural networks (multilayer prediction).

Key words:

Diabetes, prediction, machine learning, artificial neural networks.

Table de matière:

Chapitre03: Réalisation.

Liste des tableaux:

Liste des figures:

Liste d'abréviations

DID : Diabète insulinodépendant.

NDID : non insuline dépendant

DT1 : diabète type 1.

DT2 : diabète type 2.

CEED : center européen d'étude de diabète.

HbA1c: hémoglobine glyquée

RNA : réseau de neurone Artificiel.

AR : Auto régressifs.

MGC : Mesure de glucose continue.

CMG : continuos Glucose Monitoring.

PMC : perception multicouches.

INTRODUCTION GENERALE

Introduction générale:

Le diabète est une maladie chronique qui provoque une augmentation de la glycémie. Cette maladie a commencé à se propager rapidement, ce qui nous incités à penser à un système qui nous aide à réduire le risque d'infection avec une prédiction précoce et une connaissance des facteurs les plus importants qui la contrôlent.

L'apprentissage automatique est un champ d'étude de l'intelligence artificielle qui peut être la meilleure solution de développer des systèmes de prévision fiables afin de diminuer la propagation de cette maladie.

A travers de ce mémoire de Master, nous intéressons à l'utilisation des réseaux de neurones artificiels (RNA) pour développer un système de prédiction et de prévision du diabète type 2, dans l'objectif de réduire les risques de complications de cette maladie chronique sur la santé du patient.

Ce mémoire est organisé en trois principaux chapitres:

1. Le premier chapitre présente un aperçu général sur la maladie du diabète, ses différents types, les symptômes, le diagnostic ainsi que la prédiction du diabète et les techniques de prédiction.

2. Le deuxième chapitre donne un aperçu sur les réseaux de neurones artificiels, les algorithmes d'apprentissage, les problèmes résoudre par RNA.

3. Le dernier chapitre présente d'abord une étude technique dans laquelle nous définissons l'ensemble des données, l'environnement utilisé pour construire le système. Puis une définition détaillée de la base et l'ensemble de données utilisé. Ensuite, les résultats du traitement et l'algorithme appliqué. En fini le chapitre par une conclusion qui résume les traitements et les idées appliqué dans ce système.

Chapitre 01

Généralités sur le diabète

1. Introduction :

Le diabète est une maladie qui empêche le corps d'utiliser correctement l'énergie fournie par les aliments ingérés. Par ailleurs, la maladie survient lorsque le pancréas ne s´écrêté plus d'insuline ou lorsque le corps devient résistant à la quantité d'insuline produite. Il existe principalement deux types de diabète : Le type 1 se caractérise par une production insuffisante d'insuline dans l'organisme pour lequel la survie du patient n´nécessite des injections d'insuline. Ces symptômes sont notamment les suivants : ´émission d'urine, soif excessives, faim constante, perte de poids, altération de la vision et la fatigue. Le type 2 appelé diabète non insulinodépendant ou diabète de l'adulte, il résulte de l'utilisation inefficace de l'insuline par l'organisme. Les symptômes peuvent être similaires à ceux du diabète de type 1, mais ils sont souvent moins marqués ou absents. En outre, il existe un autre type de diabète appelé diabète gestationnel qui se d´développe pendant la grossesse il est associe à un risque à long terme de diabète de type 2. Le surpoids, le manque d'exercice, les antécédents familiaux et le stress est augmenté le risque possible de diabète et le mauvais contrôle de dosage de sucre (glucose) dans le sang peut entrainer des complications très grave (c´excité, cataracte, thrombose, n´néphropathie….).

Qu'est-ce que le diabète ?

2. Définition:

Le diabète est un trouble de l'assimilation, de l'utilisation et du stockage des sucres apportés par l'alimentation. Cela se traduit par un taux de glucose dans le sang (encore appelé glycémie) élevé : on parle d'hyperglycémie.

Les aliments sont composés de lipides (graisses), protéines (protéines animales ou végétales) et glucides (sucres, féculents). Ce sont eux qui fournissent l'essentiel de l'énergie dont a besoin le corps pour fonctionner, passent dans l'intestin, puis rejoignent la circulation sanguine. Quand on mange, le taux de sucre dans le sang augmente, les glucides sont alors transformés essentiellement en glucose. Le pancréas détecte l'augmentation de la glycémie. Les cellules bêta du pancréas, regroupées en amas appelés îlots de Langerhans, sécrètent de l'insuline. L'insuline fonctionne comme une clé, elle permet au glucose de pénétrer dans les cellules de l'organisme : dans les muscles, dans les tissus adipeux et dans le foie où il va pouvoir être transformé et stocké le glucose diminue alors dans le sang. Une autre hormone, le glucagon, permet de libérer le glucose stocké dans le foie, en dehors des repas, lors d'une baisse énergétique ou d'une baisse

de glycémie, C'est l'équilibre de ces hormones qui permet de maintenir la glycémie stable dans le corps. En cas de diabète, ce système de régulation ne fonctionne pas. [1]

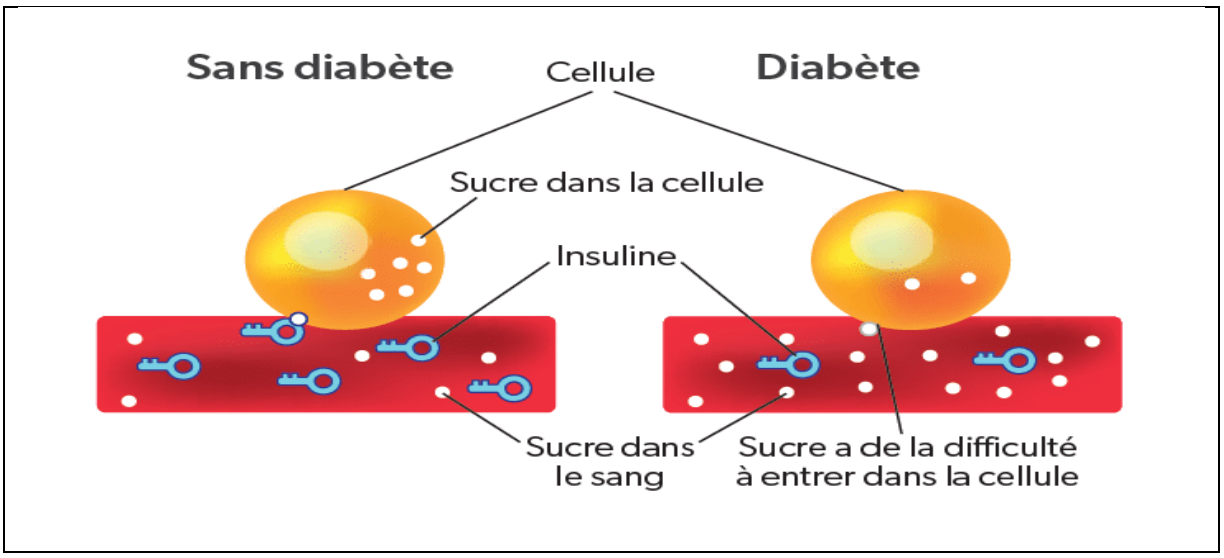


Figure 1 : une photo de la cellule normale /diabétique .[2]

4 63millions de personnes sont atteintes du diabète dans le monde d'après la Fédération internationale du diabète. Celle-ci qualifie le phénomène de véritable pandémie, car la progression est considérable. Ainsi, l'OMS prévoit 622 millions de diabétiques d'ici 2040. [3]

Figure 2: les nombres de personnes sont atteints du diabète dans le monde. [4]

3. Examen et diagnostic : comment savoir si on a du diabète ?

Un dosage de la glycémie est pratiqué en laboratoire d'analyses médicales. Un diabète est avéré lorsque la glycémie à jeun est égale ou supérieure à 1.26 g/l à deux reprises ou égale ou supérieure à 2 g/l à n'importe quel moment de la journée. [1]

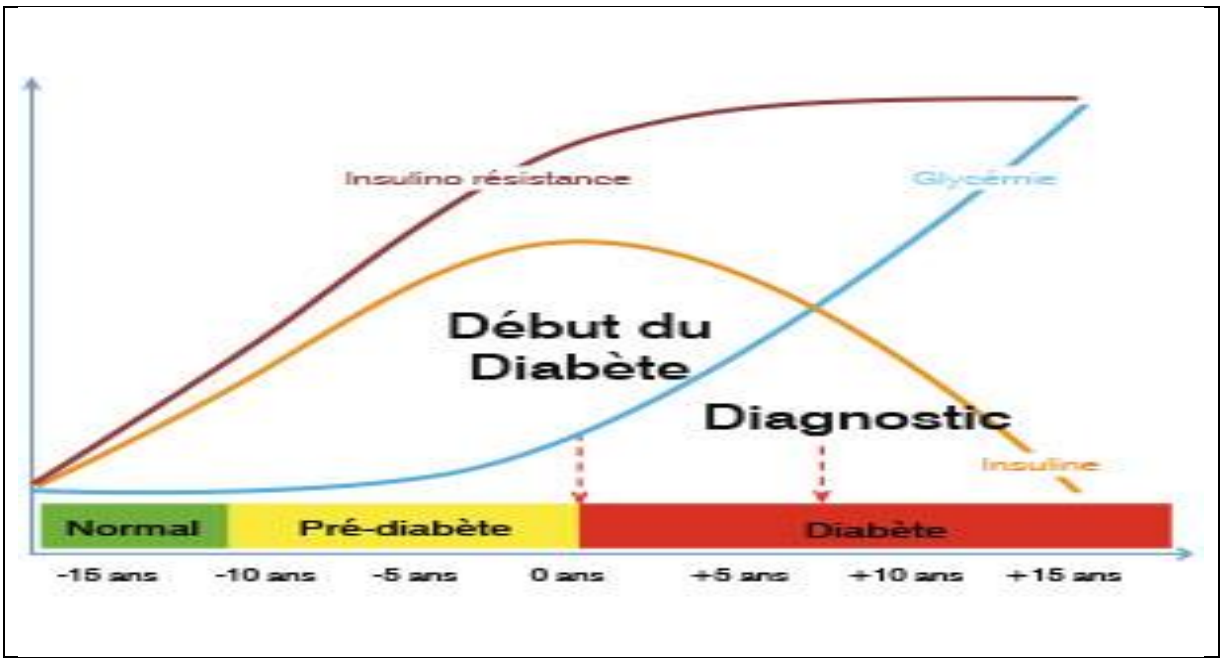


Figure 3:comment savoir si on a du diabète. [5]

4. Les types de diabetes:

On distingue principalement deux types de diabète : **le diabète de type 1** qui touche environ 6% des diabétiques et **le diabète de type 2** qui en touche 92 %. Les autres types de diabète concernent les 2 % restants (Diabète gestationnel, MODY, LADA ou diabète secondaire à certaines maladies ou prises de médicaments). [1]

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

Le diabète de type 1, appelé autrefois diabète insulinodépendant (DID), est habituellement découvert chez les personnes jeunes : enfants, adolescents ou jeunes adultes. [1]

4.1.1 Les symptômes du diabète de type 1 :

Les symptômes sont généralement une soif intense, des urines abondantes, un amaigrissement rapide. Ce diabète résulte de la disparition de la cellule bêta du pancréas entraînant une carence totale en insuline. [4]

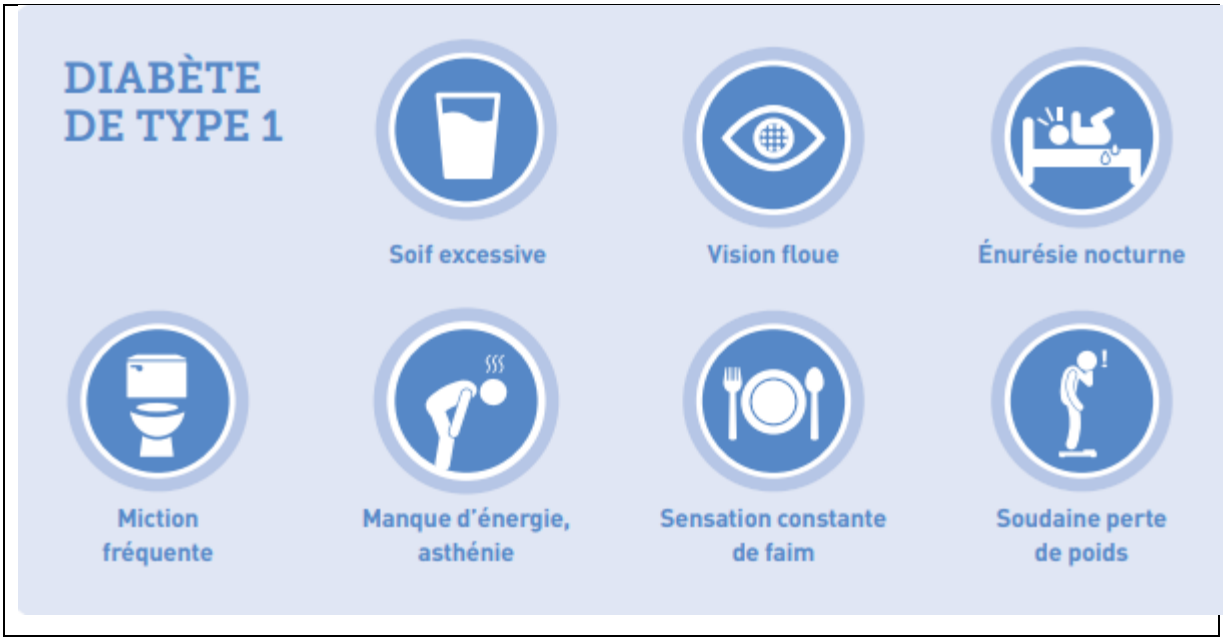


Figure 4:Les symptômes du diabète de type 1. [4]

L'organisme ne reconnaît plus ces cellules bêta et les détruit (les cellules bêta sont détruites par des anticorps et des cellules de l'immunité, les lymphocytes, fabriquées par l'organisme): **on dit que le diabète de type 1 est une maladie auto-immune**. Le glucose ne pouvant entrer dans les cellules retourne dans le sang. Le taux de glucose dans le sang s'élève alors.

4.2 Diabète de type 2 :

Précédemment appelé diabète non insulinodépendant ou diabète de la maturité ou de l'adulte, est une maladie chronique, silencieuse et indolore, qui se caractérise par un taux de sucre (glucose) trop élevé dans le sang (hyperglycémie). Cette anomalie est causée par un défaut de la sécrétion ou de l'utilisation de l'insuline qu'est la conséquence d'une perte de fonctionnalité des ˆılots pancréatiques. Cette perte de

fonctionnalité est la conséquence de l'interaction de facteurs génétiques, volontiers héréditaires et de facteurs environnementaux liés au mode de vie. Contrairement au diabète de type 1, le diabète de type 2 est le plus souvent asymptomatique. De ce fait, la maladie peut ˆêtre diagnostiquée plusieurs années après son apparition, une fois les complications déjà présentes. [6]

Figure 5: Résistance à l'insuline. [7]

4.2.1 Les symptômes cliniques du diabète de type 2:

Sont les mêmes que celles du type 1, auxquelles s'ajoutent les risques cardio-vasculaires, mais aussi une incidence sur le développement de certains cancers, troubles du comportement ou maladies mentales. [8]

4.2.2 Le traitement du diabète de type 2 :

Le traitement repose prioritairement sur des alimentations équilibrée et pratique d'une activité physique régulière .si ces deux éléments sont insuffisants, il faudra ajouter un traitement par anti-diabétique oral. Le traitement à l'insuline peut s'avérer nécessaire, si les glycémies restent néanmoins élevées. [8]

Figure 6: Le traitement hygiéno-diététiques. [7]

4.3 Diabète gestationnel:

Le diabète gestationnel est un diabète qui survient chez une femme enceinte, du fait des modifications métaboliques provoquées par la grossesse (mais pas toutes les femmes enceintes). Il est appelé aussi diabète de grossesse. [8]

Contrairement aux diabètes de type DT1 et DT2 qui sont des pathologies évolutives et à surveiller à vie, le diabète gestationnel disparait le plus souvent après la naissance du bébé. Lorsqu'une femme souffre de diabète gestationnel au cours de sa grossesse, elle est plus susceptible d'en souffrir à nouveau lors de sa prochaine grossesse et elle est exposée à un risque plus ´élevé de d´développer un diabète de type 2 par la suite. Plus une femme est enceinte à un âge avancé, plus le risque de d´développer un diabète gestationnel au cours de sa grossesse est ´élevé.

4.3.1 Le traitement du diabète gestationnel :

Selon le Centre européen d'´étude du Diabète (Ceed) le traitement par le recours `à l'insuline est n´nécessaire dans 50% des cas et dans quelques cas plus rares un traitement par anti-diabétique oral peut ˆêtre mis en place. Dans tous les cas, des mesures hygiène diététiques doivent rapidement ˆêtre mises en place, avec la particularité qu'elles doivent prendre en compte `à la fois le diabète de la mère et les besoins nutritionnels du fœtus. [8]

4.3.2 Les conséquences du diabète gestationnel :

- maternelles : toxémie gravidique.

- La macrosomie fœtale : risques pour l'enfant d'ˆêtre trop gros et d'entrainer des complications `a l'accouchement.
- n´néonatales : risques d'hypoglycémie et d'hypocalcémie.

5. Les complications du diabète:

Le but du traitement dans les deux types de diabète est de normaliser la glycémie : les hyperglycémies répétées et prolongées entraînent à long terme une altération des nerfs et des vaisseaux sanguins présents dans tout le corps. Ce sont les complications du diabète qui peuvent se traduire par une cécité, des atteintes des pieds pouvant conduire à des amputations, des infarctus et des accidents vasculaires cérébraux, des troubles de l'érection ou une insuffisance rénale.

6. La prédiction du diabète:

La prédiction est l'action d'annoncer à l'avance un événement par calcul, par raisonnement, par induction; *par métonymie*, ce qui est ainsi annoncé. [9]

- Vérification régulièrement le niveau de diabète avec auto-teste : un lecteur de glycémie pour contrôler plusieurs fois par jour sur une goutte de sang à des moments précis.
- Faite le test AC1 : ce test vous indique comment la glycémie s'écoulera au fil du temps.
- Le risque de développer un diabète augmente si vous souffrez d'hypertension.
- Test de sucre à jeun : teste en laboratoire d'analyses médicales pour mesurer sa glycémie à jeun et tous les 3 mois, son hémoglobine glyquée (HbA1c).
- Surpoids et obésité : les personnes surpoids sont plus à risque de développer un diabète.
- Diabète héréditaire : Le poids de l'hérédité diffère selon qu'il s'agit du diabète de type 1 ou du diabète de type 2.
- Vieillesse.
- Taux de cholestérol élevé.

7. Les techniques de prédiction :

- Les techniques basées sur les données : dépendent principalement des données d'entrées-sorties d'expérience et ne nécessitent aucune connaissance sur la physiologie du diabète.
- Application axées sur la connaissance dans l'exploration de données (système de gestion concerne la modélisation prédictive du métabolisme du glucose).
- Les modèles compartimentaux, les réseaux de neurones à convolution de séries chronologique et les réseaux de neurones récurrents (RNN).
- Le capteur CGM : Un système de mesure du glucose en continu est composé d'un lecteur et d'un capteur qui doit être placé sur la peau et remplacé chaque semaine.[10]
- L'analyse des séries chronologiques : Une série chronologique est la réalisation d'un processus aléatoire indicé par le temps, noté {Xt}. Pour chaque t, Xt est une variable aléatoire dont on a une réalisation, xt.[11]
- Méthodes d'apprentissage automatique.
- Les processus gaussiens.
- Les modèles de prédiction autorégressifs (AR) basés sur données CGM.
- Prédiction à l'aide d'un réseau de neurone artificiel.

8. Conclusion :

Les deux principaux types de diabète sont des maladies différentes mais caractérisées par un excès de sucre dans le sang et doivent être prises au sérieux et traitées efficacement.il n'y a pas de « petits diabètes » ou de diabètes plus graves que d'autres.

Malgré la recherche médicale qui avance tous les jours, le diabète reste une maladie qui se soigne très bien mais qui ne se guérit pas. Il faut donc, toute sa vie, se surveiller, garder de bonnes habitudes alimentaires, pratiquer une activité physique et prendre régulièrement son traitement. Un diabétique peut donc être un malade en bonne santé ! **OUI à la qualité de vie !**.

Chapitre 02 :

Réseaux de neurones artificiels

1. Introduction :

Un réseau de neurones artificiels (RNA), (ou Artificiel Neural Network en anglais), est un système informatique matériel et / ou logiciel **dont le fonctionnement est calqué sur celui des neurones du cerveau humain.**

Il s'agit là d'une variété de technologie Deep Learning (apprentissage profond), qui fait elle-même partie de la sous-catégorie d'intelligence artificielle du Machine Learning (apprentissage automatique). Dans ce chapitre, nous définirons les réseaux de neurones, les problèmes résoudre par les RNA, quelle que domaines d'application, ses principaux types et l'algorithme choisie pour atteindre le but dans notre travaille, afin de réduire les risques de complication de cette maladie sur la santé d'un patient. [12]

2. Historique :

Le concept des réseaux de neurones artificiels **fut inventé en 1943 par deux chercheurs de l'Université de Chicago** : le neurophysicien « Warren McCullough », et le mathématicien « Walter Pitts ».Dans un article publié dans le journal Brain Theory, les deux chercheurs présentent leur théorie selon laquelle l'activation de neurones est l'unité de base de l'activité cérébrale. [12]

En 1949: D. Hebb

Présente dans son ouvrage « The Organization of Behavior » une règle d'apprentissage pour les réseaux de neurones artificiels. De nombreux modèles de réseaux aujourd'hui s'inspirent encore de la règle de Heb.

En 1958: F. Rosenblatt

Développe le modèle du perceptron et démontre son théorème de convergence. Dans la même période, Le modèle de L'Adaline (ADAptive LINar Element) a été présenté par B. Widrow et Hoff. Ce modèle sera par la suite le modèle de base des réseaux multi-couches.

En 1969: M. Minsky & S. Papert

Démontrent les limitations du modèle du perceptron.

Plus récemment, dans les années quatre-vingt, de nouveaux modèles mathématiques (les réseaux à couches, les réseaux auto-adaptatifs, les mémoires associatives, etc.) ont permis de dépasser les limites du perceptron.

Aujourd'hui, la discipline des RNA concerne un public de plus en plus large d'étudiants, de chercheurs, d'ingénieurs et d'industriels. Des revues spécialisées et un flux très important d'articles ne cessent de marquer leur importance.

3. Neurones biologiques :

Un neurone biologique est une cellule vivante particulière possédant des extensions par lesquelles il peut distribuer des signaux à d'autres cellules ou en recevoir

Les neurones sont reliés entre eux par des liaisons appelées axones. Ces axones vont eux-mêmes jouer un rôle important dans le comportement logique de l'ensemble. Ces axones conduisent les signaux électriques de la sortie d'un neurone vers l'entrée (synapse) d'un autre neurone. [13]

Les neurones font une sommation des signaux reçus en entrée et en fonction du résultat obtenu vont fournir un courant en sortie.

La structure d'un neurone se compose de trois parties principales (figure1):

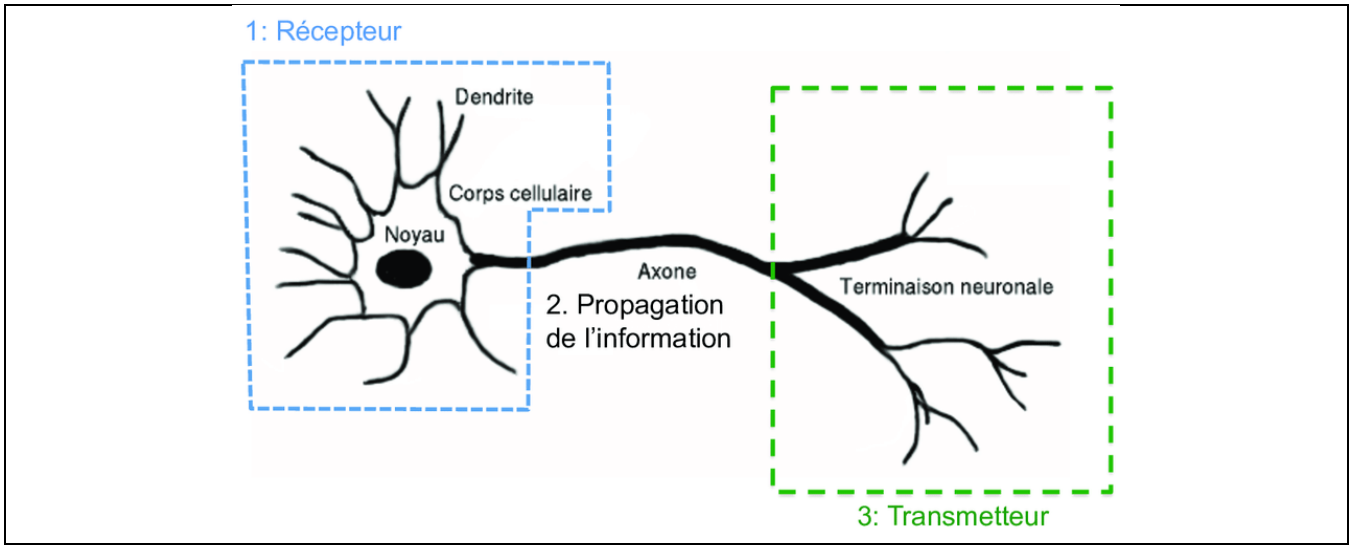


Figure 7: Neurones biologiques. [13]

- **Le corps cellulaire :** composé du centre de contrôle traitant les informations reçues par les dendrites.
- **Les dendrites :** sont les principaux fils conducteurs par lesquels transitent l'information venue de l'extérieur.

- **L'axone :** est fil conducteur qui conduit le signal de sortie du corps cellulaire vers d'autres neurones.

4. Neurones formels (artificiels):

Un neurone formel est une représentation mathématique et informatique d'un neurone biologique. Le neurone formel possède généralement plusieurs entrées et une sortie qui correspondent respectivement aux dendrites et au cône d'émergence du neurone biologique (point de départ de l'axone). Les actions excitatrices et inhibitrices des synapses sont représentées, la plupart du temps, par des coefficients numériques (les poids synaptiques) associés aux entrées. Les valeurs numériques de ces coefficients sont ajustées dans une phase d'apprentissage. Dans sa version la plus simple, un neurone formel calcule la somme pondérée des entrées reçues, puis applique à cette valeur une fonction d'activation, généralement non linéaire. La valeur finale obtenue est la sortie du neurone. [14]

Le neurone formel est l'unité élémentaire des réseaux de neurones artificiels dans lesquels il est associé à ses semblables pour calculer des fonctions arbitrairement complexes, utilisées pour diverses applications en intelligence artificielle.

Mathématiquement, le neurone formel est une fonction à plusieurs variables et à valeurs réelles. [14]

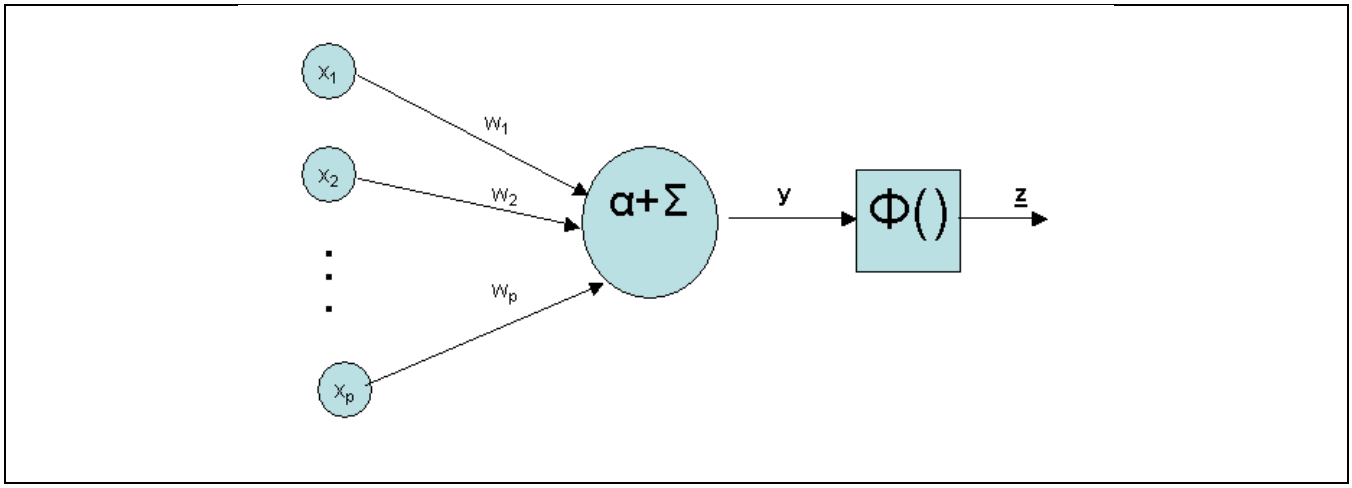


Figure 8 : Neurones formels (artificiels). [15]

4.1 Modélisation d'un neurone formel :

La modélisation consiste à mettre en œuvre un système de réseaux neuronaux sous un aspect non pas biologique mais artificiel, cela suppose que d'après le principe biologique on aura une correspondance pour chaque élément composant le neurone biologique, donc une modélisation pour chacun d'entre eux. Un neurone se compose généralement d'une entrée formée des variables sur lesquelles opèrent ce neurone et une sortie représentant la valeur de la fonction réalisée (fonction d'activation). La sortie du neurone est une fonction non linéaire d'une combinaison des entrées xi (signaux d'entrées) pondérées par les paramètres wi (poids synaptiques). Graphiquement le neurone est présenté sous la forme indiquée sur la figure (Figure 2).

Mathématiquement le neurone est une fonction algébrique non linéaire, paramétrée, à valeurs bornées. La sortie du neurone est donnée par l'expression suivante :

$$Y_K = \varphi\left[\sum_{j=1}^{p} W_{KJ}\,.X_J - \theta_K\right]$$

Où : x 1, x 2, x 3,…... x p : entrées,

w_{k1},w_{k2}, w_{k3} ……w_{kp}: Poids synaptiques du neurone.

θ_k: Le Seuil.

φ ( ): La fonction d'activation.

y_k: est la sortie d'activation.

5. Le lien entre les neurones biologiques et neurones artificiels:

Le neurone artificiel était inspiré par le neurone biologique.

Résoudre des problèmes de la même manière que le cerveau humain.

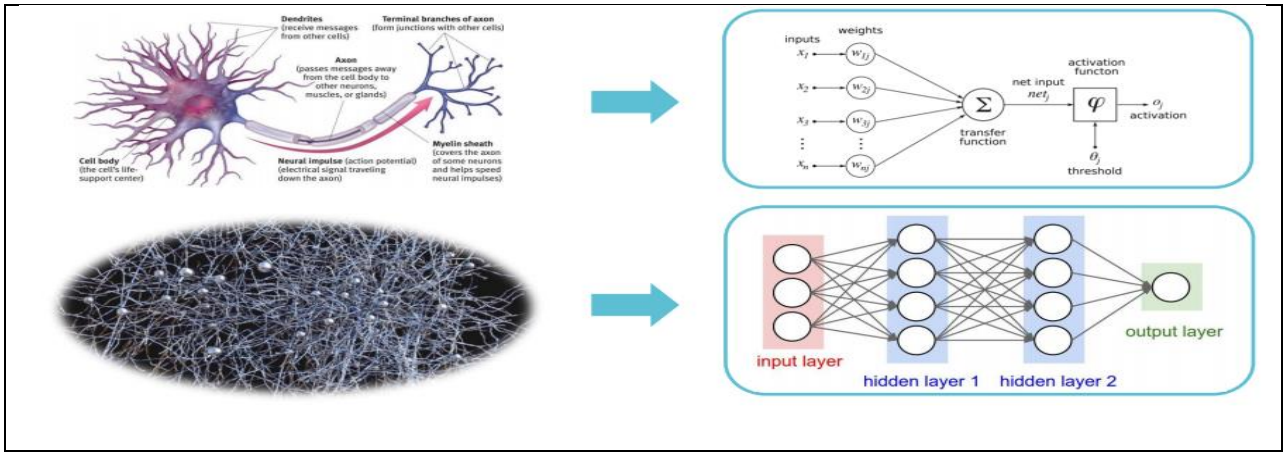


Figure 9:Le lien entre les neurones biologiques et neurones artificiels.[16]

6. Les réseaux de neurones artificiels:

6.1 Définitions :

Les réseaux de neurones artificiels sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit. Toute structure hiérarchique de réseaux est évidemment un réseau.

Les réseaux de neurones artificiels, sont des modèles inspirés du fonctionnement du cerveau animal, et dont le but est de voir surgir des propriétés analogues au système biologique.

Ils en reprennent quelques grands principes :

- Le parallélisme : les neurones sont des entités réalisant une fonction très simple, mais ils sont très fortement interconnectés entre eux, ce qui rend le traitement du signal massivement parallèle.

- Les poids synaptiques : les connexions entre les neurones ont des poids variables, qui déterminent la force de l'interaction entre chaque paire de neurones.

- L'apprentissage : ces coefficients synaptiques sont modifiables lors de l'apprentissage, dans le but de faire réaliser au réseau la fonction désirée. [12]

7. Problèmes résolus par les réseaux de neurones artificiels :

Les RNA sont considérés comme une nouvelle approche de traitement de l'information par apprentissage de cette information et la rendent disponible à l'utilisation afin de résoudre un tel problème.

7.1 La classification :

La classification est le processus de classement des entrées en groupes. Par exemple, une compagnie d'assurance peut vouloir classer les demandes d'assurance dans les différentes catégories de risque, ou une organisation en ligne peut vouloir de son système de messagerie de classer le courrier entrant dans des groupes de messages spam et non-spam.

7.2 La reconnaissance des formes (pattern recognition) :

La reconnaissance des formes est l'une des utilisations les plus courantes des réseaux neuraux. Pattern recognition est une forme de classification et est tout simplement la capacité de reconnaître un motif. Le modèle doit être reconnu même s'il n'est pas clair.

Exemple : la reconnaissance des visages.

7.3 L'optimisation :

Une autre application des réseaux de neurones est l'optimisation qui peut être appliquée à de nombreux problèmes pour lesquels une solution est recherchée. Le réseau de neurones peut ne pas toujours trouver la solution optimale mais il cherche à trouver une solution acceptable.

7.4 La prévision :

La prévision est une autre application des réseaux neuraux artificiels. Etant donné une série temporelle de données d'entrée, un réseau de neurones peut prédire les valeurs futures. La précision de la prévision dépend de nombreux facteurs, tels que la quantité et la pertinence des données d'entrée. Par exemple, les réseaux de neurones sont généralement appliqués à des problèmes de prédiction de l'évolution des marchés financiers.

8. Les domaines application des RNA :

- **Traitement d'image :** compression d'images, reconnaissance de caractères et de signatures, reconnaissance de formes et de motifs, chiffrement, classification, ……
- **Traitement du signal :** traitement de la parole, identification de sources, filtrage, classification, …….

- **Traitement automatique des langues :** segmentation en mots, représentation sémantique des mots (plongements lexicaux), étiquetage morphosyntaxique, traduction automatique, …….
- **Contrôle :** diagnostic de pannes, commande de processus, contrôle qualité, robotique, ……
- **Optimisation :** allocation de ressources, planification, régulation de trafic, gestion, finance, ..
- **Simulation :** simulation boîte noire, prévisions météorologiques.
- Classification d'espèces animales étant donnée une analyse ADN.
- Modélisation de l'apprentissage et perfectionnement des méthodes de l'enseignement.
- Approximation d'une fonction inconnue ou modélisation d'une fonction connue mais complexe à calculer avec précision.
- **En gestion et finance:** Banque, Cartes de crédit, Finance, Assurance, Marketing.
- **Autres domaines :** Archéologies**,** Défense, Environnement, Sécurité, Production, Médecine, Energies, Pharmacie, Psychologie, Recherche scientifique, Télécommunication, transport … **[17]**

9. Architecture des réseaux de neurones :

L'adaptation d'un réseau à un problème donné passe par un choix de la topologie et des poids de liaison entre les neurones.

La topologie des réseaux de neurones peut être très variée. On peut concevoir plusieurs types de réseaux seulement en modifiant les règles de connexion.

9.1 Réseau à une Couche:

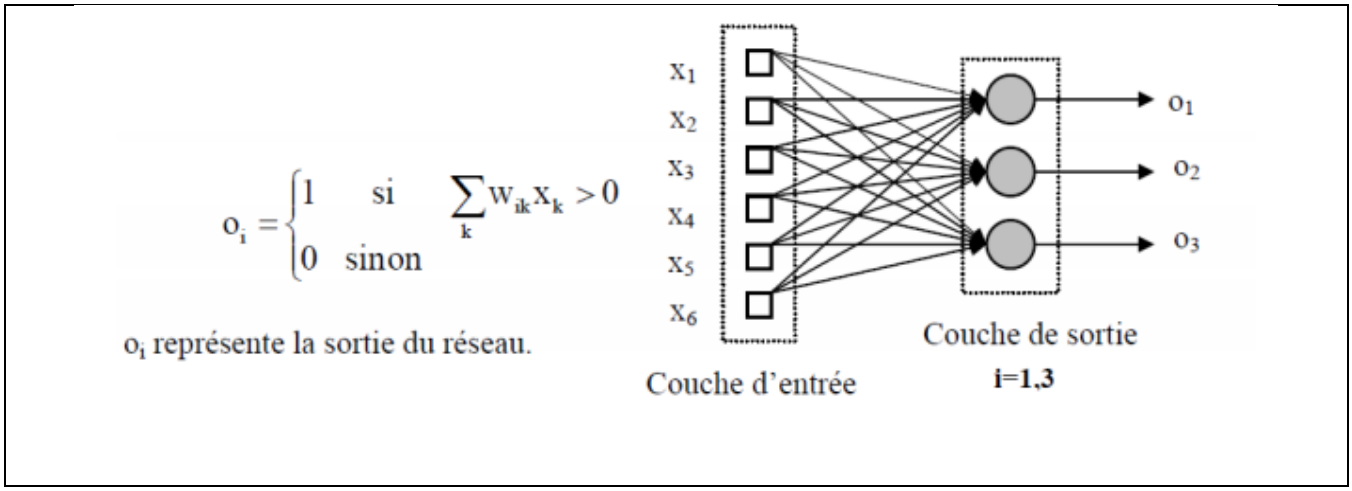


Figure 10: exemple d'un réseau à une couche.

9.2 Réseaux récursifs :

L'idée des connexions récurrentes est que le réseau est capable de « se rappeler » des valeurs des états précédents par l'intermédiaire de leurs poids synaptiques.

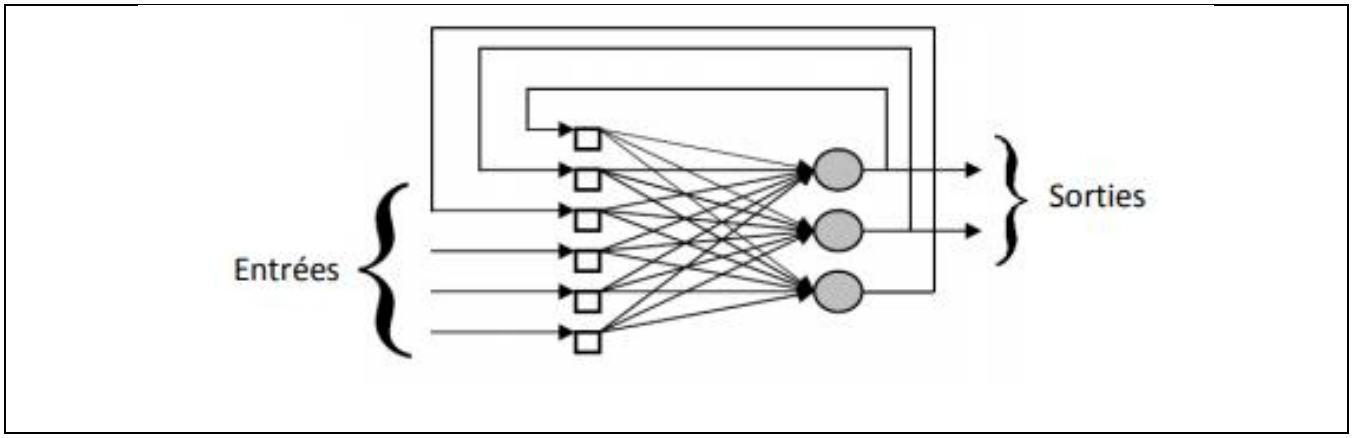


Figure 11: Exemple d'un réseau récursif

9.3 Réseaux multicouches :

On appelle :

- Couche d'entrée : l'ensemble des neurones d'entrée,

- Couche de sortie : l'ensemble des neurones de sortie.

- Couches cachées : l'ensemble des couches intermédiaires, elles n'ont aucun contact avec l'extérieur.

- (ne) indique les neurones de la couche d'entrée et (ns) indique les neurones de la couche de sortie.

La (figure.6) représente un exemple d'un perceptron multicouche (PMC), notant les couches cachées par la lettre 'c' et les couches de sorties par la lettre 'o'.

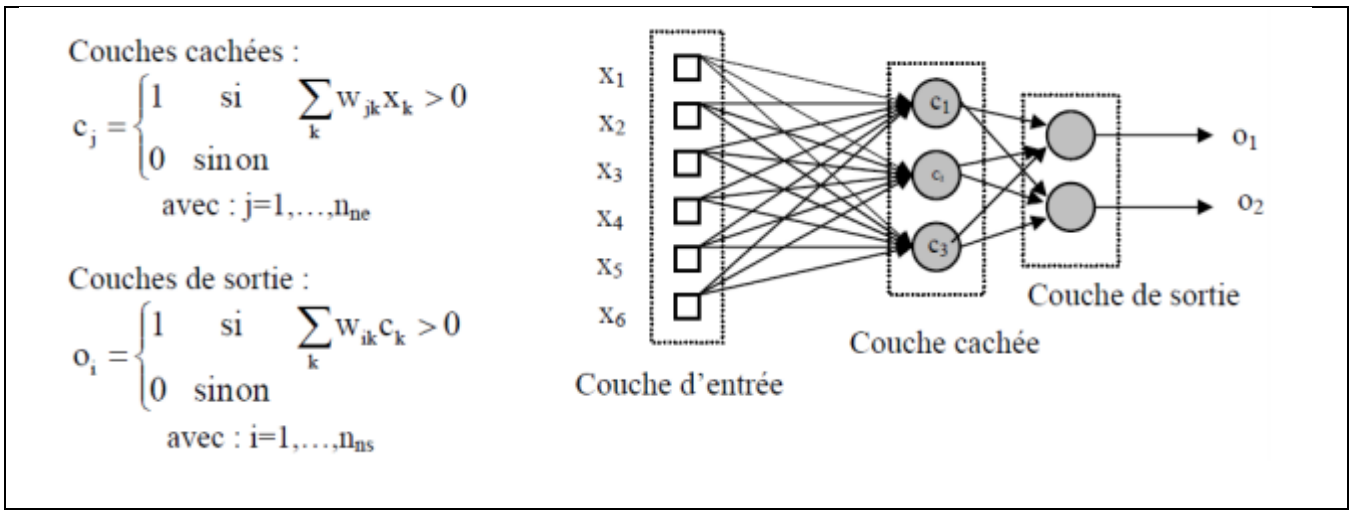


Figure 12: Exemple d'un réseau multicouche.

10. Les types de réseau de neurones :

Il existe de nombreux types de réseaux neuronaux, on peut les diviser en deux grandes catégories selon la nature de leur algorithme d'apprentissage. En effet, les premiers sont dits ***supervisés*** car, lors de l'apprentissage, ils doivent disposer d'un professeur capable de leur indiquer ce qui devrait être produit en sortie pour chaqu' une des informations fournies en entrées. Les seconds sont dits ***non supervisés*** car ils arrivent à s'auto-organiser.[18]

Nous allons décrire ci-dessous différents types de réseaux pour leurs objectifs différents :

10.1 Le perceptron (réseau à couches):

C'est un réseau à couche qui permet de positionner en entrées des éléments devants être appris (éléments linéairement séparables afin de ne pas avoir de confusion). Ceux-ci tracent un chemin à travers

le réseau. Une fois l'apprentissage effectué, en repassant les mêmes éléments en entrée ils réutilisent le même chemin et activent les mêmes neurones de sorties. Plusieurs neurones de sorties sont activées pour chaque comparaison, les résultats ne sont pas identiques à chaque comparaison, des approximations sont effectuées.

Le perceptron peut être utilisé pour l'apprentissage et la reconnaissance d'image. [18]

10.2 Kohonen:

Ce réseau de neurones peut être considéré comme dynamique, des neurones peuvent être détruits et crées, le réseau n'a pas de taille fixe. Généralement ce réseau est appelée carte de kohonen, en effet ce réseau est représenté à plat comme une grille rectangulaire à 1, 2, 3 ou 4 dimensions. Les applications sont multiples : sélection de données représentatives dans une grande base de cas, compression d'images, diagnostic de pannes, optimisation combinatoire (dont le fameux "voyageur de commerce", modélisation de la cartographie des aires visuelles...

Un exemple d'utilisation: Hop Field : Représente un réseau sans structure de couches, ni de sens de propagation, composé de N cellules. Ce réseau ce rapproche le plus du fonctionnement du cerveau humain. [18]

Aussi les types de réseau de neurones diffèrent par plusieurs paramètres :

- la topologie des connexions entre les neurones ;
- la fonction d'agrégation utilisée (somme pondérée, distance pseudo-euclidienne...) ;
- la fonction de seuillage utilisée (sigmoïde, échelon, fonction linéaire, fonction de Gauss, ...) ;
- l'algorithme d'apprentissage (rétro propagation du gradient, cascade corrélation) ;
- d'autres paramètres, spécifiques à certains types de réseaux de neurones, tels que la méthode de relaxation pour les réseaux de neurones (e.g : réseaux de Hopfield) qui ne sont pas à propagation simple (e.g : Perceptron Multicouche). [19]

De nombreux autres paramètres sont susceptibles d'être mis en œuvre dans le cadre de l'apprentissage de ces réseaux de neurones par exemple :

- la méthode de dégradation des pondérations (weight decay), permettant d'éviter les effets de bord et de neutraliser le sur-apprentissage. [19]

10.3 Le perceptron multicouche (PMC):

Le perceptron multicouche *(*multilayer perceptron MLP) est un type de réseau neuronal artificiel organisé en plusieurs couches au sein desquelles une information circule de la couche d'entrée vers la couche de sortie uniquement ; il s'agit donc d'un réseau à propagation directe (feedforward). Chaque couche est constituée d'un nombre variable de neurones, les neurones de la dernière couche (dite « de sortie ») étant les sorties du système global. [20]

Le perceptron a été inventé en 1957 par Frank Rosenblatt au *Cornell* Aeronautical Laboratory. Dans cette première version le perceptron était alors mono-couche et n'avait qu'une seule sortie à laquelle toutes les entrées étaient connectées. [20]

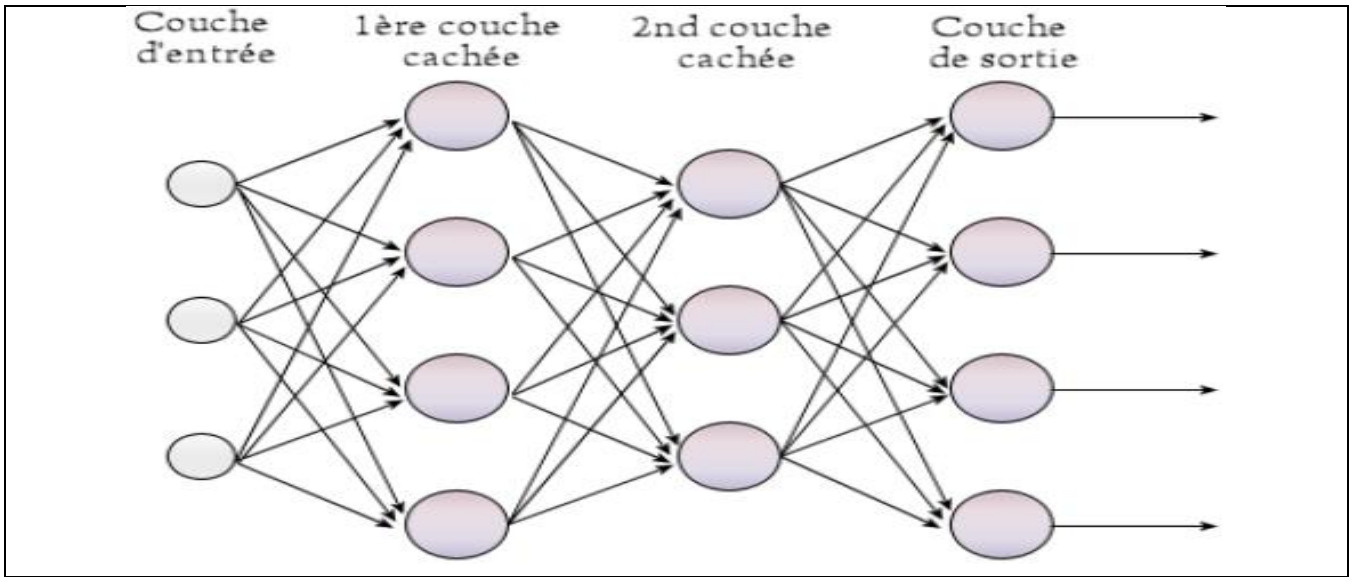


Figure 13:Un perceptron multicouche(PMC).

10.4 Perceptron multicouche à rétropropagation:

Les premiers réseaux de neurones n'étaient pas capables de résoudre des problèmes non linéaires ; cette limitation fut supprimée au travers de la rétropropagation du gradient de l'erreur dans les systèmes

multicouches, proposé par Paul Werbos **)** en 1974 et mis au point douze années plus tard, en 1986 par David Rumelhart.

Dans le perceptron multicouche à rétropropagation, les neurones d'une couche sont reliés à la totalité des neurones des couches adjacentes. Ces liaisons sont soumises à un coefficient altérant l'effet de l'information sur le neurone de destination. Ainsi, le poids de chacune de ces liaisons est l'élément clef du fonctionnement du réseau : la mise en place d'un Perceptron multicouche pour résoudre un problème passe donc par la détermination des meilleurs poids applicables à chacune des connexions inter-neuronales. Ici, cette détermination s'effectue au travers d'un algorithme de rétropropagation. [20]

10.5. Apprentissage :

10.5.1 Calcul de l'erreur:

En connaissant la valeur attendue e_i à la sortie d'un perceptron pour des entrées données, on peut calculer l'écart avec la prédiction grâce à une fonction objectif C^2, le plus souvent l'erreur quadratique moyenne $(abrégée\ MSE)^3$, telle que :

$$MSE(e_i, y_i) = \frac{1}{n}\sum_{i=0}^{n}(y_i - e_i)^2$$

Cette fonction n'est pas linéaire, et sa dérivée est plus grande si la prédiction est éloignée de la valeur attendue, permettant ainsi un apprentissage plus rapide· Au contraire, l'erreur moyenne absolue (MAE) a une dérivée constante, et donc un taux d'apprentissage qui ne varie pas.

$$MAE(e_i, y_i) = \frac{1}{n}\sum_{i=0}^{n}(|y_i - e_i|)^2$$

En minimisant ces fonctions objectives, les prédictions gagnent en précision. [20]

10.5.2. Calcul du gradient:

Durant la phase d'apprentissage, après avoir calculé les erreurs du réseau de neurones, il est nécessaire de les corriger afin d'améliorer ses performances. Pour minimiser ces erreurs – et donc la fonction objectif –, l'algorithme du gradient est le plus souvent utilisé. Le gradient ∇ est calculé afin de connaître la variation de la fonction objectif par rapport aux paramètresθ. Il permet ensuite de modifier ces paramètres

proportionnellement à leur impact sur la précision de la prédiction, dans le but d'atteindre après plusieurs itérations le minimum global de la fonction objectif.

La modification des paramètres θ à un instant t se fait tel que :

$$\theta_{t-1} = \theta_t - \alpha \nabla C$$

avec α un scalaire, le taux d'apprentissage, et ∇C le gradient de la fonction objectif. L'algorithme du gradient permet donc de trouver les paramètres θ du réseau tel que la somme des erreurs faites par les prédictions sur des données d'entrainement X soit la plus faible possible, c'est-à-dire que :

$$C_X(\theta) = \min_{\theta} C_X(\theta).$$

Le gradient se calcule avec la dérivée partielle de la fonction objectif par rapport à chacun des paramètres. Lorsqu'il y a plusieurs paramètres à optimiser, il est exprimé comme un vecteur, parfois noté $\vec{\nabla}$, puis ensuite ajouté au vecteur θ des paramètres, après avoir été multiplié par le taux d'apprentissage. Le gradient indique la direction vers le maximum de la fonction objectif, et son opposé mène donc vers le minimum. Son expression est donc :

$$\nabla C_x(\theta) = (\frac{\partial C}{\partial \theta_1}, \frac{\partial C}{\partial \theta_2}, \frac{\partial C}{\partial \theta_3} \ldots\ldots)^T$$

Soit ∇_I le gradient sur un perceptron de la couche k, alors l'ensemble des gradients de cette couche peuvent être stockés et manipulés dans une matrice jacobienne J_k , c'est-à-dire une matrice contenant les dérivées partielles de la fonction objectif vectorielle sur toute la couche avec :

$$\boldsymbol{K_j} = \begin{pmatrix} \boldsymbol{\nabla_1} \\ \boldsymbol{\nabla_2} \\ \boldsymbol{\nabla_3} \\ . \\ . \\ . \\ \boldsymbol{\nabla_m} \end{pmatrix} = \begin{pmatrix} \frac{\boldsymbol{\partial C}}{\boldsymbol{\partial \theta_{1,1}}} & . & . & . & \frac{\boldsymbol{\partial C}}{\boldsymbol{\partial \theta_{1?n}}} \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ \frac{\boldsymbol{\partial C}}{\boldsymbol{\partial \theta_{m,1}}} & . & . & . & \frac{\boldsymbol{\partial C}}{\boldsymbol{\partial \theta_{m,n}}} \end{pmatrix}$$

$\nabla \mathbf{C} = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\theta}_{ij}}$

En utilisant le théorème de dérivation des fonctions composées, la variation de la fonction objectif par rapport à l'un des poids est :

$$\nabla C = \frac{\partial \theta}{\partial \omega} = \frac{\partial o}{\partial \omega}\frac{\partial y}{\partial o}\frac{\partial C}{\partial y}$$

$$\frac{\partial C}{\partial y_i} = \frac{\partial}{\partial y_i}\left(\frac{1}{n}\sum_{i=0}^{n}(y_i - e_i)^2\right) = 2(y_i - e_i)$$

$\frac{\partial y}{\partial o} = y(1-y)$ si la fonction sigmoïde sert d'activation, ou $\frac{\partial y}{\partial o} = 1 - y^2$ pour la tangente hyperbolique ;

$$\frac{\partial o}{\partial \omega} = y_i - 1$$

L'apprentissage s'arrête lorsque les paramètres convergent vers des valeurs, et que la dérivée de la fonction objectif vaut 0. [20]

10.6 Algorithme de rétropropagation :

1. Présentation d'un motif d'entraînement au réseau.
2. Comparaison de la sortie du réseau avec la sortie ciblée.
3. Calcul de l'erreur en sortie de chacun des neurones du réseau.
4. Calcul, pour chacun des neurones, de la valeur de sortie qui aurait été correcte.
5. Définition de l'augmentation ou de la diminution nécessaire pour obtenir cette valeur (erreur locale).
6. Ajustement du poids de chaque connexion vers l'erreur locale la plus faible.
7. Attribution d'un blâme à tous les neurones précédents.
8. Recommencer à partir de l'étape 4, sur les neurones précédents en utilisant le blâme comme erreur. [20]

11.Conclusion :

Dans ce chapitre, nous avons présenté les réseaux de neurones artificiels qui peuvent aidez-nous à détecter le diabète à un stade précoce, ce qui peut aider à réduire le risque de complication de cette maladie sur la santé du patient. Dans l'étude qui suit , l'objectif principale est d'appliquer le type de réseau neurone choisie «le perception multicouches (MLP) » aux données ext

Chapitre 03

Réalisation

1. Introduction:

Dans ce dernier chapitre, nous présentons d'abord une étude technique dans laquelle nous définissons l'environnement logiciel utilisé pour construire notre application, puis nous définirons notre datas et avec une description de ses caractéristiques et les étapes de prétraitement des données (explorer, nettoyer, sélection de modèle ...) pour corriger les valeurs aberrantes et choisir le meilleur modèle à suivre. A la fin, c'est la partie application ou nous fournissons des interfaces graphiques importantes développées pour clarifier les performances des activités du système et nous terminerons par une conclusion.

2. Outils et Libraries utilisés:

Anaconda :

Anaconda est une distribution python pour les applications de data science et d'apprentissage automatique. C'est un logiciel gratuit et open source qui contient plusieurs packages. Le principal avantage de l'utilisation d'anaconda est que, anaconda est comme un point central pour les bibliothèques qui auraient besoin pour le traitement de données, l'analyse prédictive et les calculs scientifiques. [21]

Jupyter notebook :

Jupyter Notebook est un environnement de programmation qui prend en charge plusieurs langages de programmation, dont Python. Jupyter Notebook nous permet de créer des documents contenant du code, des équations, des visualisations et du texte. Ses utilisations comprennent : le nettoyage et la transformation des données, la simulation numérique, la modélisation statistique, la visualisation des données, l'apprentissage automatique et bien plus encore. [22]

Python :

Python est un langage de programmation multi-paradigme et le langage de programmation dominant dans la data science avec de nombreuses implémentations ce qui le rend encore plus intéressant. Concernant le domaine de l'apprentissage automatique Python se distingue tout particulièrement en offrant une pléthore de librairies de très grande qualité, couvrant tous les types d'apprentissages disponibles qui combine la facilité d'utilisation et d'apprentissage avec la puissance des librairies qu'elles possèdent. Parmi ces bibliothèques, nous avons utilisé : [23]

Matplotlib :

Matplotlib est une bibliothèque du langage de programmation Python destinée à tracer et visualiser des données sous formes de graphiques.[24]

Seaborn :

Seaborn est une bibliothèque de visualisation de données Python basée sur matplotlib. Il fournit une interface de haut niveau pour dessiner des graphiques statistiques attrayants et informatifs.

Pandas :

Pandas est une autre bibliothèque Python utilisée pour la manipulation et l'analyse des données, le point fort de cette bibliothèque est qu'elle possède une fonctionnalité importante appelée nettoyage des données qui résout le problème du temps passé à nettoyer les données dans un projet d'apprentissage automatique car de nombreux ensembles de données disponibles contiennent des champs vides ou nuls, ce qui peut avoir un impact n´négatif ´énorme sur notre modèle. [25]

NumPy:

NumPy est une extension du langage de programmation Python, destinée à manipuler des tableaux multidimensionnels.

Scikit-learn :

Scikit-learn est la bibliothèque Python la plus importante pour ce qui concerne l'apprentissage automatique telle qu'il contient de nombreux algorithmes (forêts aléatoires, des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support). [26]

3. Définition l'ensemble de données utilisé :

C'est un ensemble de données sur le diabète, extrait de « Pima Indian database », il se compose de plusieurs variables prédictives médicales < Inputs > et une variable cible < Outcome >.

Les variables sont les suivants :

1. Glucose : Concentration plasmatique de glucose à 2 heures dans un test oral de tolérance au glucose.

2. Pregnancies : Nombre de fois enceinte.

3. BloodPressure : Pression artérielle diastolique (mm Hg).

4. SkinThickness : Epaisseur de pli cutané du triceps (mm).

5. Insuline : Insuline sérique 2 heures (mu U/ ml).

6. BMI : (ou IMC) Indice de masse corporelle (poids en kg / (taille en m)2).

7. DiabetesPedigreeFunction : Fonction généalogique du diabète.

8. Age : l'âge en années.

9. Outcome : variable de classe (0 ou 1) où 0 indique que le patient ne souffre pas de diabète et 1 indique que le patient est diabétique.

3.1 Description des variables:

Variable	Description	Analyse de données
Glucose	Une valeur de 2 heure entre (140 et 200 mg)/dl (7.8 et 11.1 mmol/L) est appelé tolérance au glucose altéré signifie que il y a un risque accru de développe le diabète au fil de temps. Un taux de glucose de 200 mg/dL(11.1 mmol/L) ou plus utilisé pour diagnostiquer le diabète.	Minimum = 0 Maximum = 199
Pregnancies	Nombre de fois enceinte	Minimum = 0 Maximum = 17

BloodPressure	Si un TA diastolique > 90 signifie une pression artérielle élevé (probabilité élevé de diabète) Un TA diastolique < 60 signifie une pression artérielle base (mois probabilité de diabète)	Minimum = 0 Maximum = 122
SkinThikness	Valeur estimé pour la graisse corporelle. épissure normal du plicutané chez les femmes est de 23 mm. Une épissure plus élevée conduit à l'obésité et les chances de diabète augmente.	Minimum = 0 Maximum = 99
Insulin	Insuline sérique 2 heures (mu U/ ml) et niveau d'insuline normal 16-166 mUI/L, les valeurs au-dessus de cette plage peuvent être alarmante.	Minimum = 0 Maximum = 846
BMI	(poids en kg / taille en m2) IMC de 18.5 à 20 c'est normal IMC entre 25 et 30 situer dans une plage surpoids Et de 30 ou plus situer dans la fourchette d'obésité.	Minimum = 0 Maximum = 67.1
DiabetePredigme Function	Fournit des informations sur les antécédentes chez les parents et la relation génétique avec les patients. Une fonction de pedigree plus élevée signifie que le patient plus	Minimum = 0.08 Maximum = 2.42

	susceptible de souffrir un diabète	
Age	Age d'une personne en années	Minimum = 21 Maximum = 81
Outcome	Indique si une personne est diabétique ou non	0 (non diabétique) :500 1 (diabétique) : 268

Tableau 1:Description des variables d'ensemble de données.

4. Les étapes à suivre dans ce travail :

- Chargement des données.
- Vérification et l'affichage des informations de la base.
- Normalisation.
- Tracer un graphique de densité pour chaque variable.
- Division des données.
- Visualisation et vérification des valeurs aberrantes.
- Standardisation.
- Analyse de corrélation.
- Importance de la fonctionnalité.
- Former MLP (train MLP).
- Tester la MLP.

4.1 Chargement des données :

La base utilisée est « Pima Indians Diabetes DataBase ».

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns

Figure 14: Aperçu de l'ensemble de données.

Pour visualiser l' ensemble de données on a utilisé la bibliothèque « pandas » qui génère un rapport de profil à partir d'un ensemble de données et qui aide à obtenir et connaitre des informations globales et approfondies sur l'ensemble de données et les variables qui contiennent.

4.2 Vérification et l'affichage des informations de la base :

Les entités sont affectées à data_X et les étiquettes correspondantes à data_Y. les informations Pandas montrent les types de données de colonne (fonctionnalité) et le nombre de valeurs non nulles.

```
data_X info:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 8 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
dtypes: float64(2), int64(6)
memory usage: 48.1 KB

data_Y info:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 1 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   Outcome  768 non-null    int64
dtypes: int64(1)
memory usage: 6.1 KB

0    500
1    268
Name: Outcome, dtype: int64
```

Figure 15: les informations de la base de données.

4.3 Normalisation des données :

La normalisation des données est une méthode de prétraitement des données qui permet de réduire la complexité des modèles, est pour effectuer cette opération on a utilisé le code suivant.(figure 13)

La normalisation des données fait référence au décalage des valeurs de vos données afin qu'elles se situent entre 0 et 1. La normalisation des données, dans ce contexte, est utilisée comme technique de mise à l'échelle pour établir la moyenne et l'écart type à 0 et 1, respectivement. [27]

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	0.352941	0.743719	0.590164	0.353535	0.000000	0.500745	0.234415	0.483333
1	0.058824	0.427136	0.540984	0.292929	0.000000	0.396423	0.116567	0.166667
2	0.470588	0.919598	0.524590	0.000000	0.000000	0.347243	0.253629	0.183333
3	0.058824	0.447236	0.540984	0.232323	0.111111	0.418778	0.038002	0.000000
4	0.000000	0.688442	0.327869	0.353535	0.198582	0.642325	0.943638	0.200000

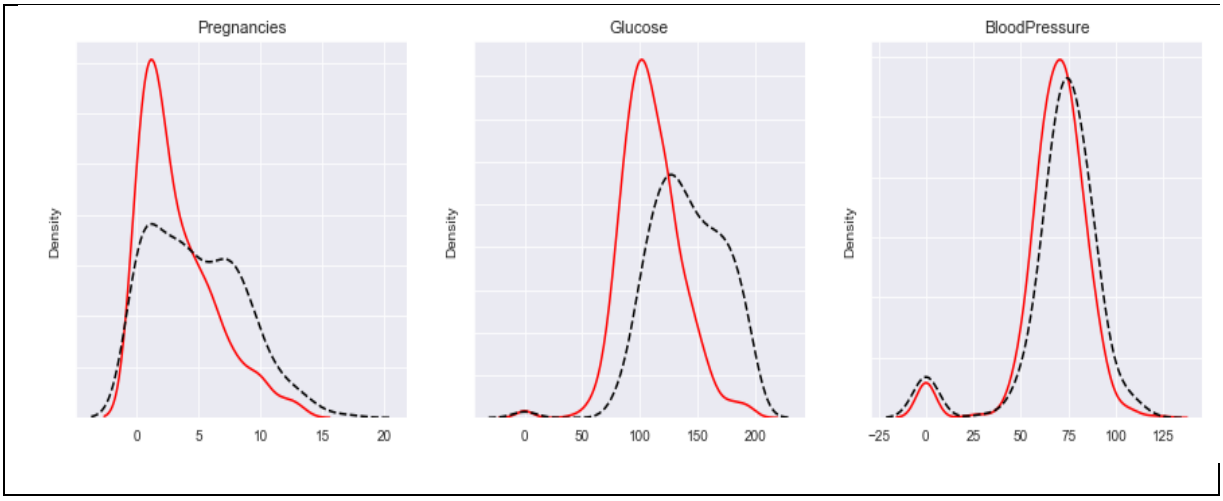
Figure 16: la normalisation de la base de données.

```
#normalisation
feature_cols = ['Pregnancies','Glucose','BloodPressure','SkinThickness','Insulin','BMI','DiabetesPedigreeFunction','Age']
X = data [feature_cols]
Y = data ['Outcome']
X_norm = X.apply(lambda x:(x-x.min())/(x.max()-x.min()))
X_norm.head()
```

Figure 17: le code de la normalisation.

4.4 Tracer un graphique de densité pour chaque variable :

Visualisation.ipynb : consiste en une étude de l'ensemble de données à l'aide des bibliothèques matplotlib et seaborn.



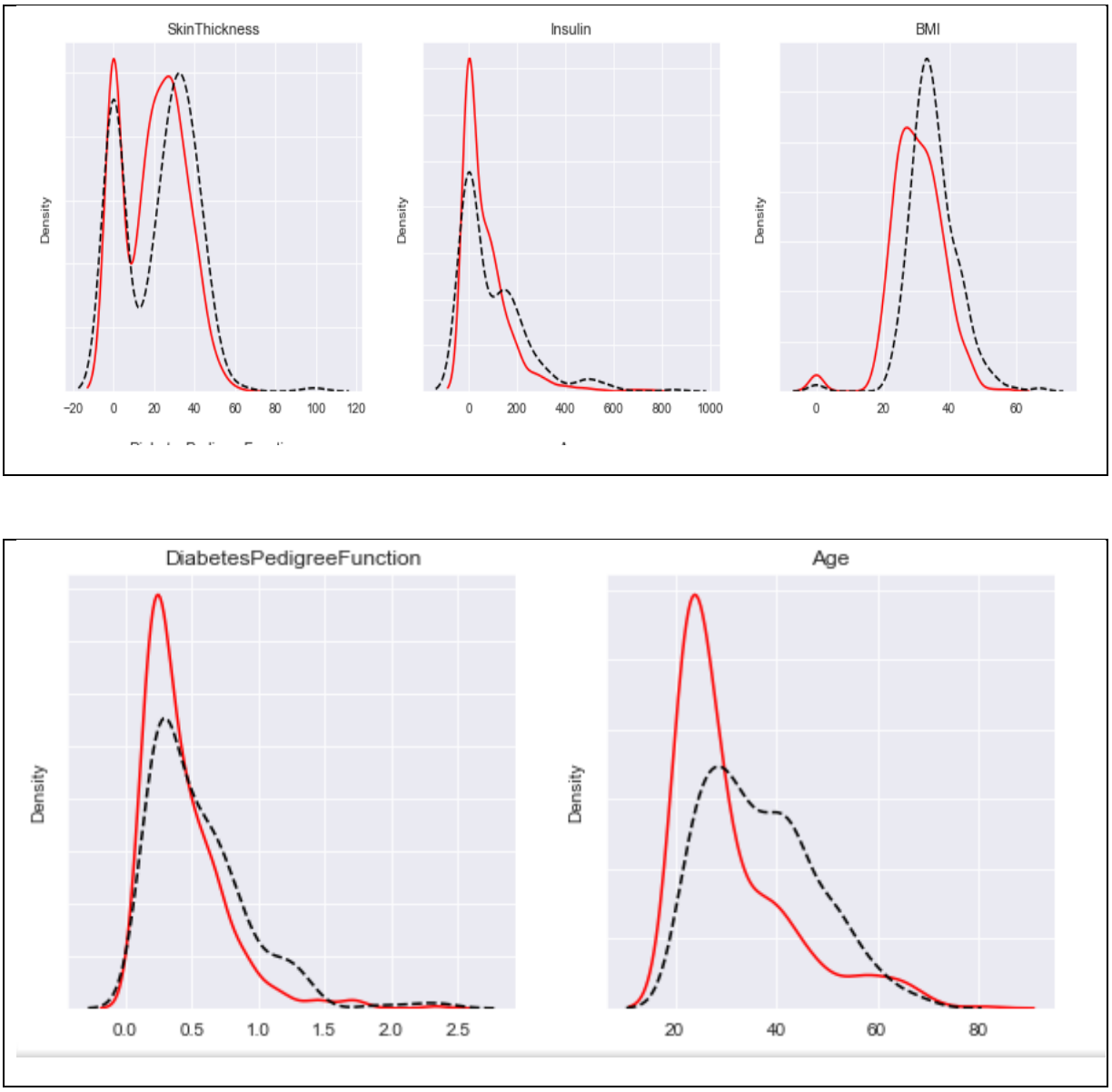


Figure 18: graphiques de densités pour chaque variable.

Il nous permet de comprend l'étude des données avant le prétraitement et l'ingénierie des fonctionnalités elle-même.

4.5 Vérification des valeurs null :

```
data.isnull().any()
```

```
Pregnancies                 False
Glucose                     False
BloodPressure               False
SkinThickness               False
Insulin                     False
BMI                         False
DiabetesPedigreeFunction    False
Age                         False
Outcome                     False
dtype: bool
```

Figure 19:Vérification des valeurs null.

4.6 Description de la base :

Pour les données numériques, l'index du résultat comprend le nombre, la moyenne, la norme, le min, le max ainsi que les centiles inférieur, 50 et supérieur. Par défaut, le centile inférieur est de 25 et le centile supérieur est de 75. Le 50 centile est le même que la médiane. [28]

```
data.describe()
```

Figure 20:code utilisé pour la description de la base.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure 21:Description de la base.

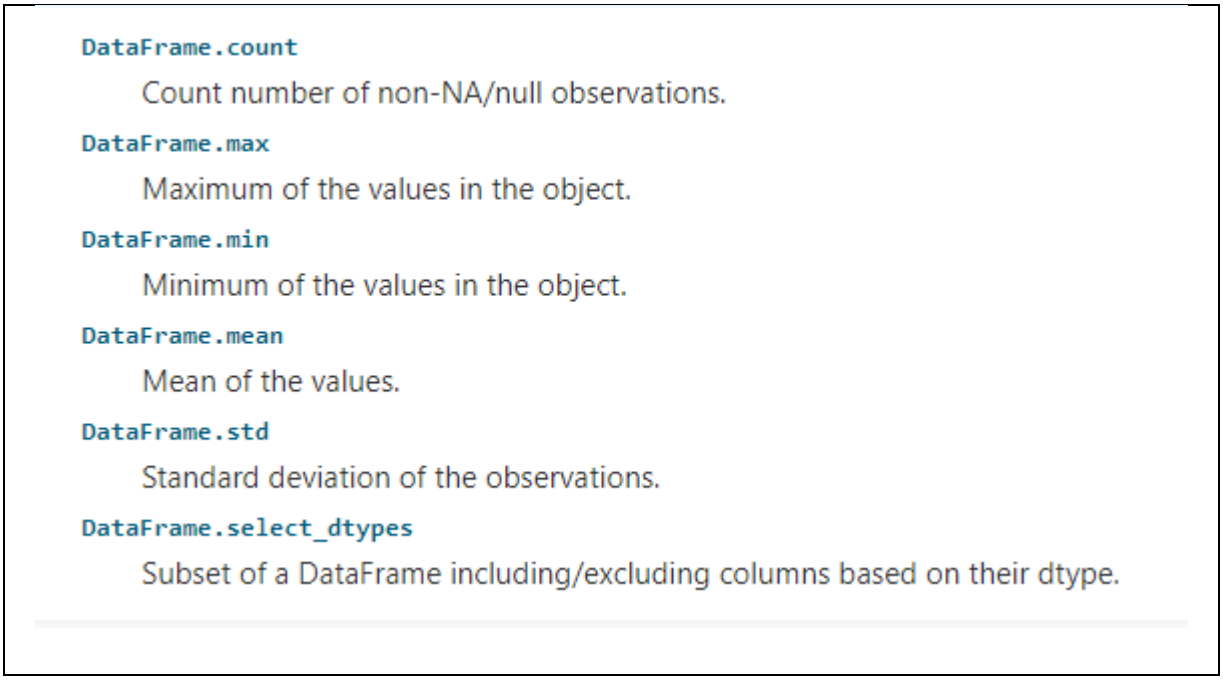


Figure 22: description de DataFrame. [28]

4.7 Division des données :

L'ensemble de données est divisé en ensembles d'apprentissage (70%) et de test (30%). Nous utilisons le paramètre stratify de la fonction train_test_split pour obtenir la même distribution de classe sur les ensembles de train et de test.

4.8 Visualisation et vérification des valeurs aberrantes :

Les valeurs aberrantes dégradent les performances d'apprentissage. Une analyse des valeurs aberrantes est effectuée pour chaque caractéristique une par une. Nous utilisons l'analyse des quartiles pour la détection des valeurs aberrantes. Pour chaque caractéristique, il y a deux tracés ci-dessous. La distribution des caractéristiques est sur la gauche. La boîte à moustaches de la même entité est à droite. Les deux sont analysés ensemble pour avoir une idée des valeurs aberrantes. À partir de ce moment, la moustache inférieure de la boîte à moustaches est désignée par LW et la moustache supérieure est désignée par UW.

Afin de dessiner une boîte à moustaches, les données d'entités sont divisées en quatre. Trois points de coupe sont nécessaires. Ces points sont le quartile inférieur (ou premier quartile), la

médiane (ou deuxième quartile) et le quartile supérieur (ou troisième quartile). Le premier quartile est la médiane des données inférieure au deuxième quartile. Le troisième quartile est la médiane des données supérieure au deuxième quartile. L'intervalle interquartile (IQR) est trouvé en soustrayant le quartile inférieur du quartile supérieur. Les valeurs aberrantes sont déterminées à l'aide des quartiles inférieurs et supérieurs et de l'IQR.

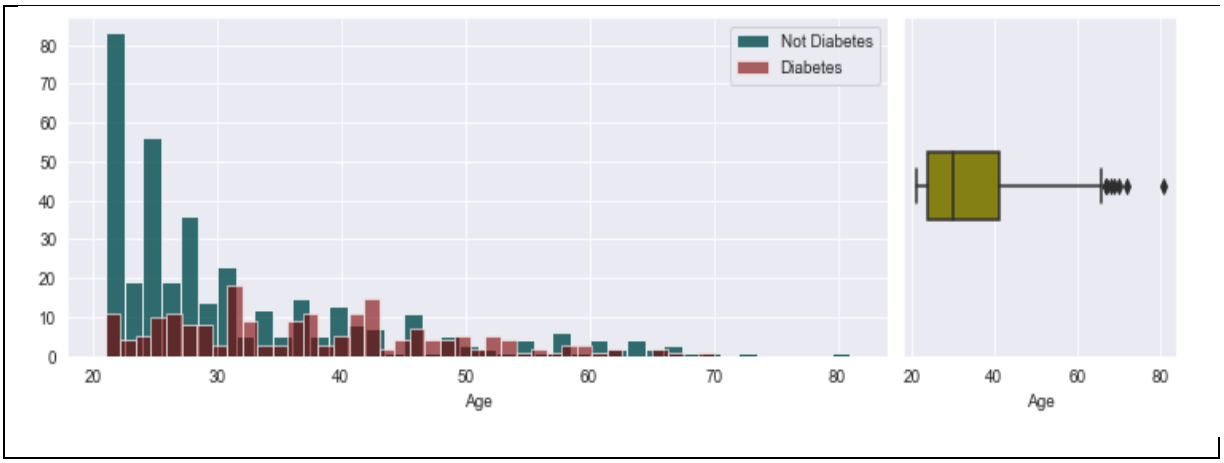


Figure 23:visualisation et vérfication de variable age.

Il y a quelques mesures au-dessus de UW en raison d'événements rares. Nous les remplaçons par le 95e quantile.

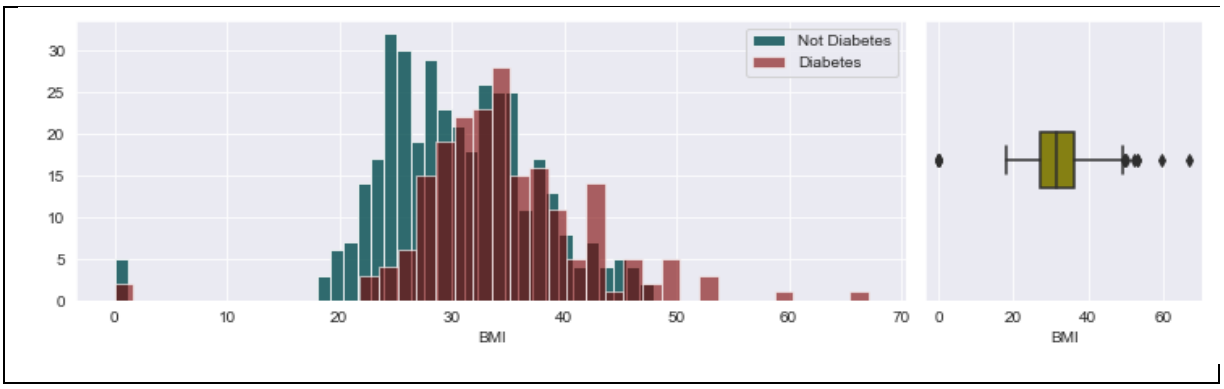


Figure 24:visualisation et vérification de variable BMI.

Il y a des valeurs 0 pour BMI. Nous les remplaçons par des médianes. De plus, nous remplaçons les valeurs supérieures à UW par q95th.

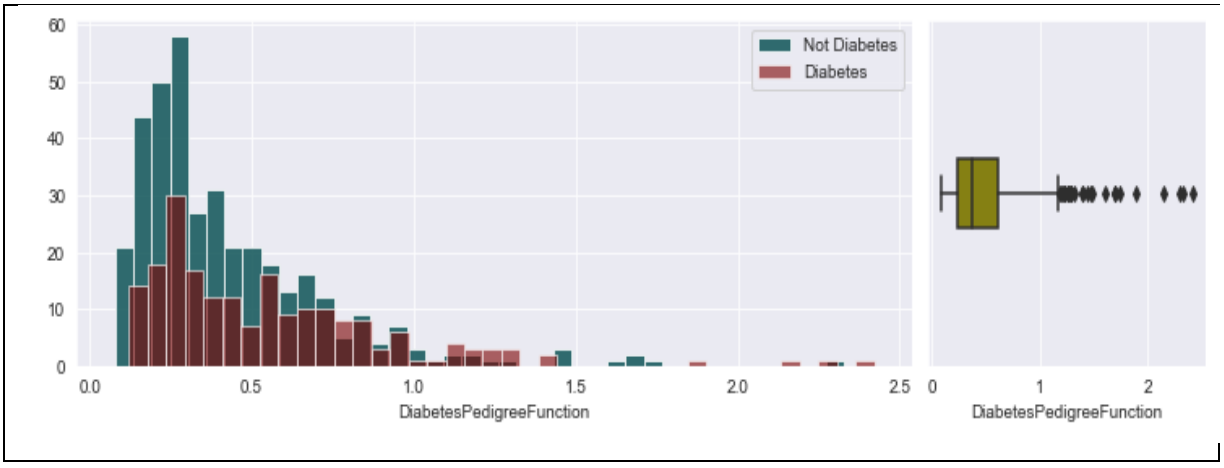


Figure 25:visualisation et vérification de variable DiabetesPedigreeFunction.

Nous remplaçons les valeurs supérieures à UW par le 95e quantile.

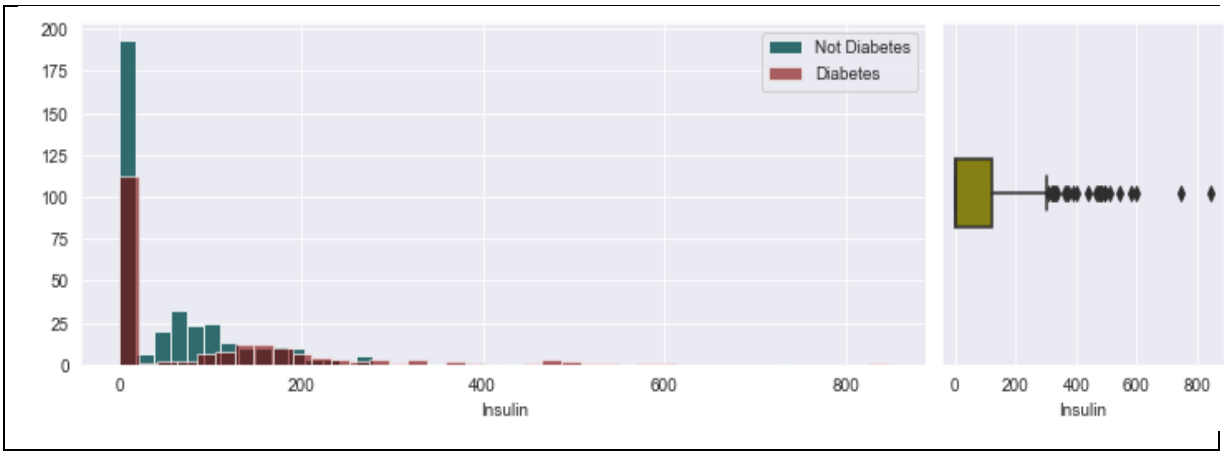


Figure 26:visualisation et vérification de variable insulin.

Il y a des valeurs 0 pour insulin, ce qui est peu probable. Nous les remplaçons donc par le 60e quantile car la médiane est de 0. De plus, nous remplaçons les valeurs supérieures à UW par le 95e quantile.

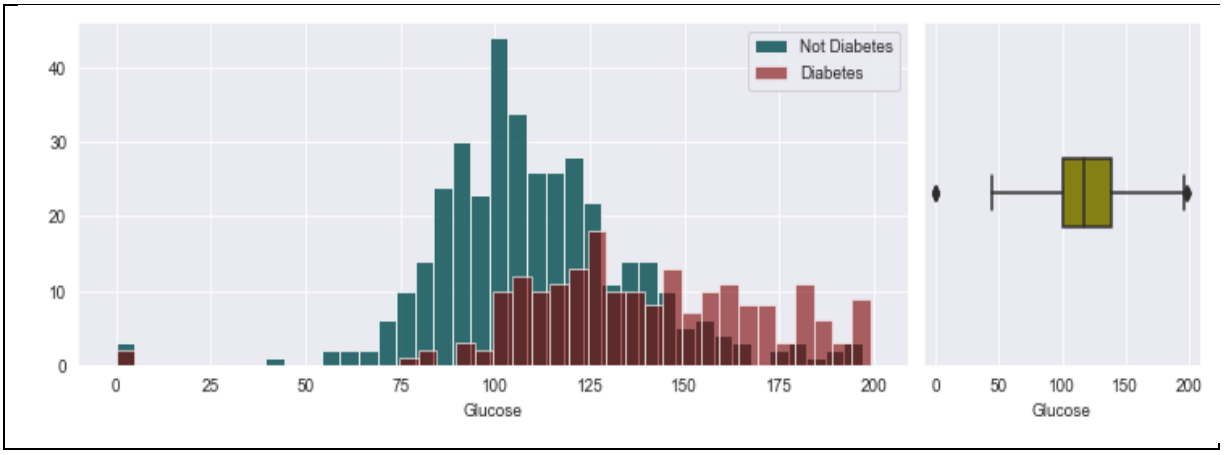


Figure 27:visualisation et vérification de variable Glucose.

Il y a des valeurs 0 pour le glucose. Nous pouvons considérer les valeurs 0 comme espace réservé pour les données manquantes. Nous les remplaçons donc par la médiane.

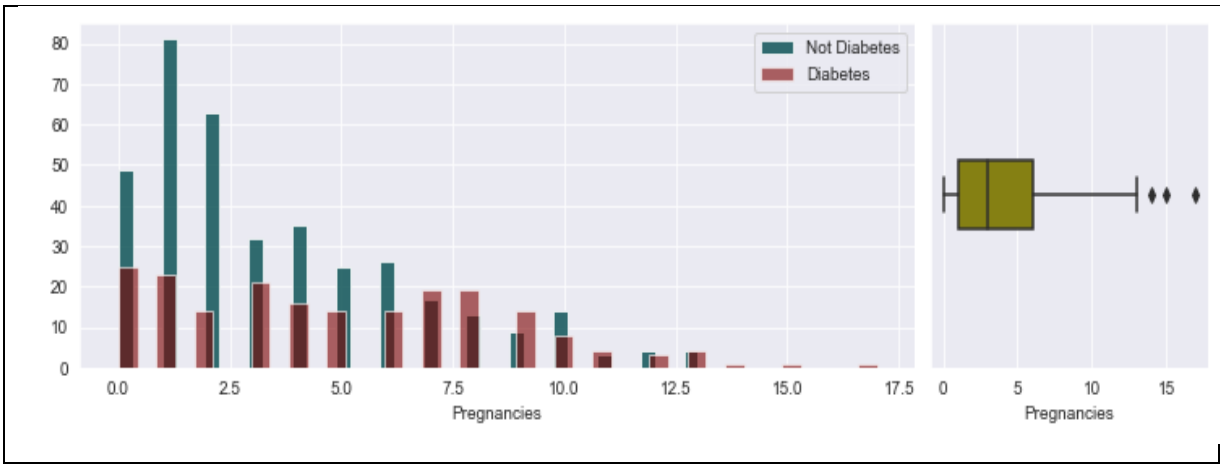


Figure 28: visualisation et vérification de variable pregnancies.

Pour la fonction Grossesse, il y a quelques mesures au-dessus de la moustache supérieure. Ce sont des événements rares. Nous les remplaçons par le 95e quantile.

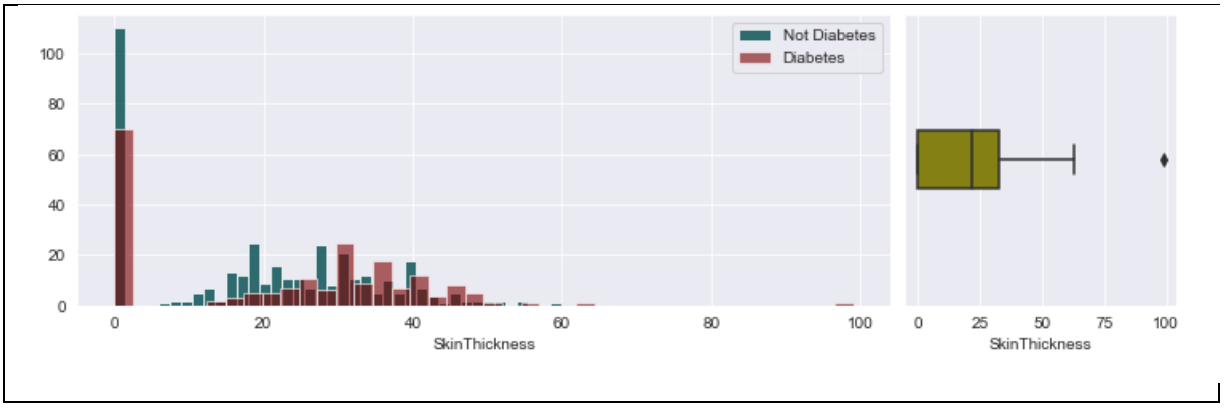


Figure 29: visualisation et vérification de variable SkinThickness.

Il y a des valeurs 0 pour SkinThickness, ce qui est peu probable. Nous les remplaçons donc par la médiane. De plus, nous remplaçons les valeurs supérieures à UW par le 95e quantile.

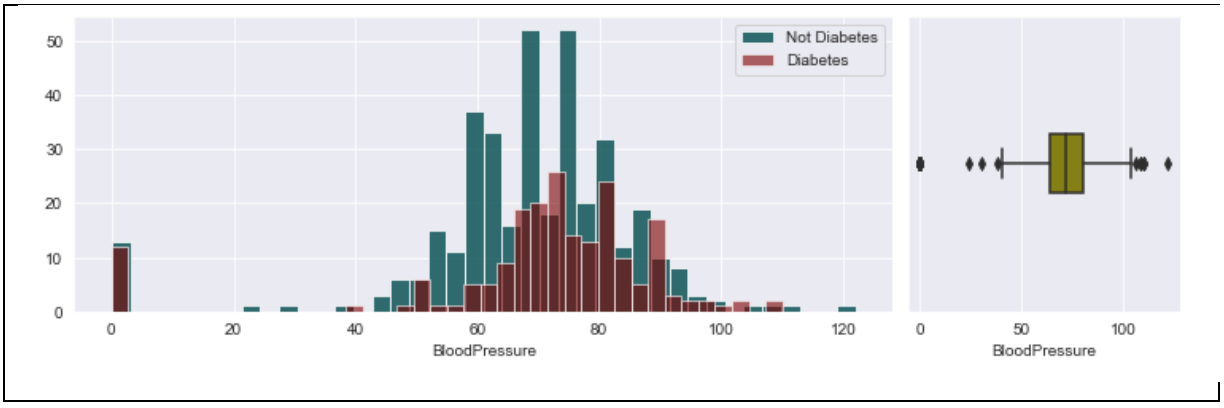


Figure 30: visualisation des variables BloodPressure.

Il y a des valeurs 0 pour Blood Pressure, ce qui est peu probable. Nous les remplaçons donc par la médiane. De plus, nous remplaçons les valeurs inférieures à LW (sauf les zéros) par le 5e quantile et les valeurs supérieures à UW par le 95e quantile.

4.9 Standardisation :

Pour augmenter les performances d'apprentissage, les caractéristiques d'entrée sont standardisées. La moyenne et l'écart type de la caractéristique sont calculés. Ensuite, la moyenne est soustraite de chaque échantillon de la caractéristique et le résultat est divisé par l'écart type. L'objectif est de transformer la fonctionnalité pour avoir une moyenne de 0 et un écart type de

1. StandardScaler de scikit-learn est utilisé. Un StandardScaler est adapté à la fonctionnalité dans train_X, puis ce scaler transforme la même fonctionnalité dans train_X et test_X.

4.10 Analyse de corrélation :

Des corrélations linéaires entre les caractéristiques ainsi qu'entre les caractéristiques et la sortie sont calculées. La fonction corr de Pandas est utilisée pour calculer la matrice de corrélation et la carte thermique Seaborn est utilisée pour le traçage.

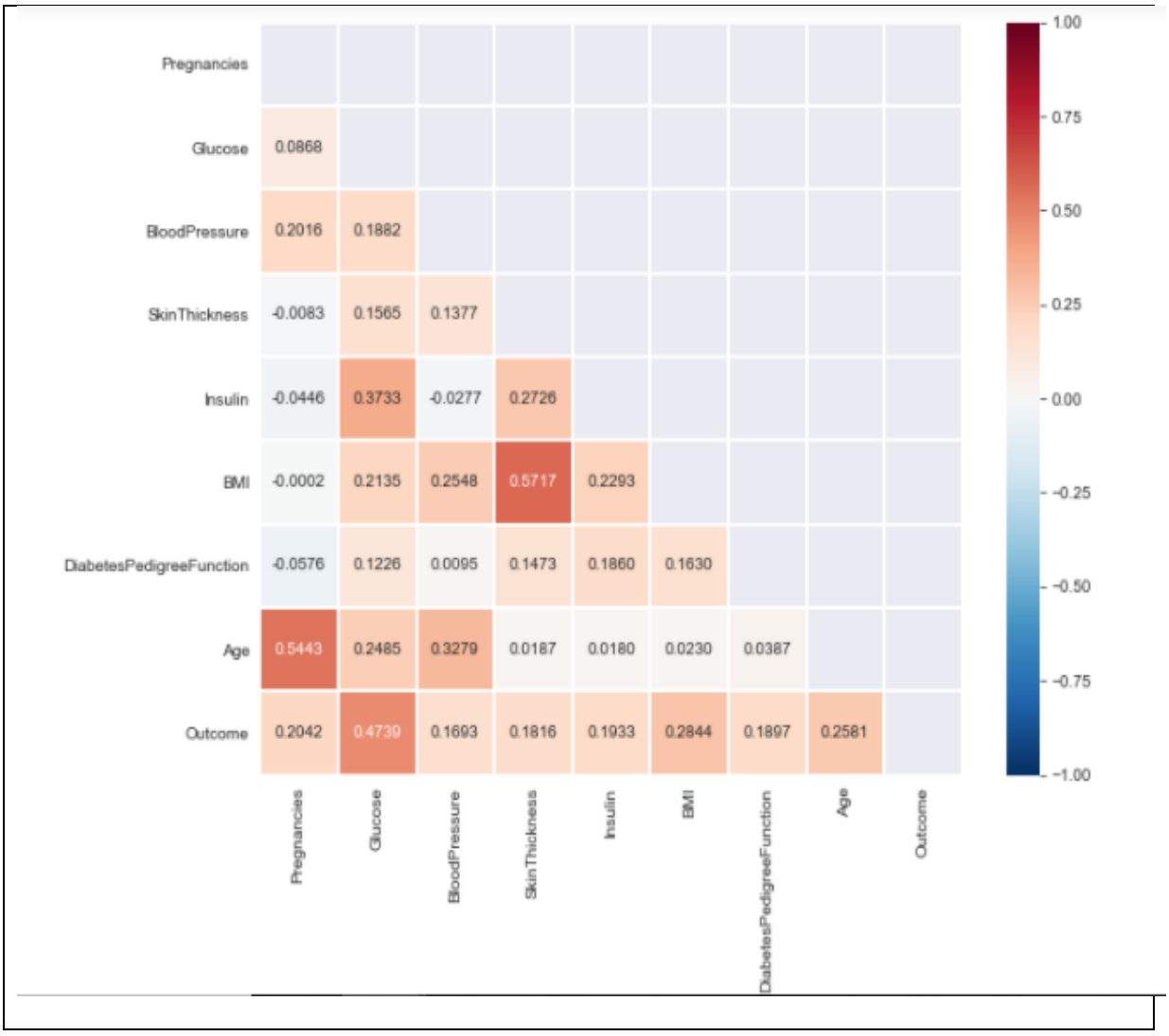


Figure 31: matrice de corrélation.

La matrice de corrélation montre qu'il existe une forte corrélation entre Outcome – Glucose et Age – pregnancies. Le résultat à la corrélation linéaire la plus élevé avec SkinThickness - BMI

4.11 Future importance:

Les méthodes d'analyse de l'importance des caractéristiques peuvent être divisées en 3 grandes catégories :

- **Méthode de filtrage** : calcul d'une métrique comme le coefficient de corrélation entre chaque caractéristique et sortie séparément comme nous l'avons fait ci-dessus. Dans cette méthode, toutes les caractéristiques sont évaluées indépendamment.

- **Méthodes embarquées** (Embedded methods): des méthodes telles que la régression logistique ou la régression linéaire apprennent les coefficients qui multiplient chaque caractéristique. L'amplitude des coefficients est associée à l'importance des caractéristiques. De plus, les méthodes basées sur des arbres, telles que les forêts aléatoires ou l'amélioration des arbres à gradients, apprennent l'importance des fonctionnalités pendant le processus de formation. Dans les méthodes intégrées, toutes les fonctionnalités sont évaluées conjointement.

```
array([[ 3.70346358e-01,  1.08294280e+00,  0.00000000e+00,
         8.09842689e-04, -3.12335579e-02,  5.71375139e-01,
         3.29023521e-01,  2.62399865e-01]])
```

Figure 32:méthode embarquée avec L1 pénalité.

```
array([[ 0.37471198,  1.08735792, -0.00450346,  0.01483766, -0.04650527,
         0.57332743,  0.33707886,  0.26982317]])
```

Figure 33:méthode embarquée avec L2 pénalité.

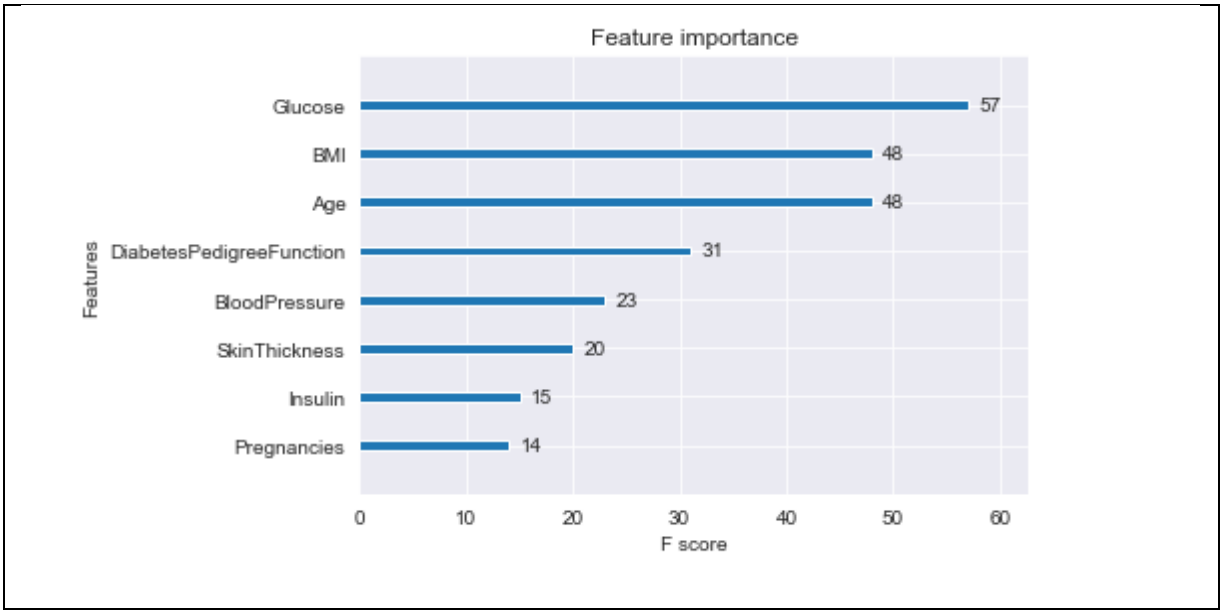


Figure 34: méthode embarquée -xgboost-

Les facteurs les plus important selon à la méthode embarquée 'xgboost' c'est le glucose, BMI et l'âge

- **<u>Méthodes wrapper</u>** : Fondamentalement, vous disposez d'un estimateur et vous entraînez cet estimateur avec les sous-ensembles de caractéristiques. Le sous-ensemble donnant le meilleur score est sélectionné et les autres caractéristiques sont éliminées.

```
[ True  True False  True  True  True  True  True]
```

Figure 35: Méthode wrapper.

4.12 MLP (train / test) :

	precision	recall	f1-score	support
Not Diabetes	0.8496	0.9040	0.8760	125
Diabetes	0.7966	0.7015	0.7460	67
accuracy			0.8333	192
macro avg	0.8231	0.8027	0.8110	192
weighted avg	0.8311	0.8333	0.8306	192

Figure 36:résulta de MLP test.

5. Conclusion :

Dans ce chapitre, nous avons présenté les différentes étapes de traitement des données, la vérification et l'affichage des informations de la base ainsi que les déférents algorithmes appliquée pour un meilleur résultat. L'application de le modèle MLP choisi pour faire cette traitement, à la fin on a développé un système qui nous permet de prédire si une personne diabétique ou pas à partir de ces information médicales.

Conclusion générale

6. Conclusion générale :

Ce mémoire présente un système de perdition du diabète type 2 en utilisant les réseaux de neurones artificiels.

Dans notre projet, nous avons étudié le diabète en détail, nous avons abordé ses types, ses causes, ses symptômes, ses causes, en plus de la façon de le prédire.

Pour obtenir les meilleurs résultats et atteindre l'objectif, ce projet a été implémenté en utilisant anaconda et jupyter notebook en plus des bibliothèques telles que seaborn, pandas.

Le but de ce système était d'obtenir le meilleur taux de prédiction pour assurer son efficacité, et c'est ce que nous avons atteint en utilisant la perception multicouche.

7. Références:

[1]:Fédération des diabétiques, [en ligne] sur le site : https://www.federationdesdiabetiques.org/information/diabete.

[2] : Équipe des professionnelles de la santé de Diabète Québec.

[3] : Atlas IDF 2017.

[4] : L'ATLAS DU DIABÈTE DE LA FID 9ème Édition 2019.

[5] : pharmacie principale sur le site : https://m.pharmacie-principale.ch/themes-sante/diabete

[6] : CEED : Centre européen d'étude du Diabète. Le dépistage au cœur des actions de prévention. [En ligne].Disponible sur : http : //ceed-diabete.org/blog/diabete-ledepistage-au-coeur-des-actions-de-prevention/

[7] : Xpnworld.La sensibilité à l'insuline c'est quoi ?[en ligne].Disponible sur : https : //xpnworld.com/sensibilie-insuline/

[8] : CEED : Centre européen d'étude du Diabète. Diabètes et complications. [en ligne].Disponible sur : http ://ceed-diabete.org/fr/le-diabete/diabete-etcomplications/

[9] : PRÉDICTION : Définition de PRÉDICTION [en ligne] sur : https://www.cnrtl.fr/definition/pr%C3%A9diction

[10] : Diabète info.fr [en ligne] sur le site : https://diabete-infos.fr/mesure-du-glucose-en-continu/

[11] : Analyse de Séries Chronologiques J.J. Daudin, C. Duby, S. Robin & P. Trécourt (INA-PG, Mathématiques) Mai 1996 disponible sur : http://www2.agroparistech.fr/IMG/pdf/Polychro.pdf

[12] : LE BIG DATA sur le site : https://www.lebigdata.fr/reseau-de-neurones-artificiels-definition.

[13] : DataScientest sur le site : https://datascientest.com/deep-learning-reseau-de-neurones-biologiques-ou-artificiels

[14] : https://datascientest.com/deep-learning-reseau-de-neurones-biologiques-ou-artificiels

[15] : https://www-ljk.imag.fr/membres/Adriana.Climescu/rncours1.htm

[16] : https://www.nextinpact.com/article/27283/105231-ia-nano-neurones-pour-repenser-larchitecture-interne-lelectronique

[17] :https://fr.m.wikiversity.org/wiki/R%C3%A9seaux_de_neurones/Applications_des_r%C3%A9seaux_de_neurones?fbclid=IwAR3Fo3IkRQ9Ls8lw6o-4ixsZ4sp8zo1DFJ0bffc9Cr3ZPLISJBiIr1q6UEw

[18] : http://www-igm.univ-mlv.fr/~dr/XPOSE2001/seguin/final/ReseauNeuro.html

[19] : https://www.techno-science.net/glossaire-definition/Reseau-de-neurones-artificiels-page-5.html

[20] : https://fr.wikipedia.org/wiki/Perceptron_multicouche.

[21] : https://fr.wikipedia.org/wiki/Anaconda_(distribution_Python)

[22]: https://jupyter.org/

[23]: https://fr.wikipedia.org/wiki/Python_(langage)

[24]: https://fr.wikipedia.org/wiki/Matplotlib

[25]: https://fr.wikipedia.org/wiki/Pandas

[26]: https://fr.wikipedia.org/wiki/Scikit-learn

[27]: https://www.sisense.com/glossary/data-standardization/

[28]:https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.describe.html