الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي والبحث العلمي

UNIVERSITÉ BADJI MOKHTAR - ANNABA
BADJI MOKHTAR – ANNABA UNIVERSITY

جامعة باجي مختار – عنابـــة

Faculté : Science de l'Ingénieur

Département : Informatique

Domaine : Informatique.

Filière : Informatique

Spécialité : Système d'information décisionnel 'SID'

Mémoire

Présenté en vue de l'obtention du Diplôme de Master

Thème:

Classification pour diagnostic du diabète utilisant les algorithmes d'apprentissage.

Présenté par : *Labed Marwa*

Encadrant :Dr Nadjette Dendani *Grade MCB Université Badji Mokhtar Annaba*

Jury de Soutenance :

M[ed] Benabas Farouk	Professeur	Université Badji Mokhtar Annaba	Président
Nadjette Dendani	MCB	Université Badji Mokhtar Annaba	Encadrant
Guessoum Meriem	Professeur	Université Badji Mokhtar Annaba	Examinateur

Année Universitaire : 2020/2021

Remerciements

Avant tout, nous remercions en premier lieu Allah le tout puissant de nous avoir illuminé et ouvert les voies du savoir, et pour nous avoir accordé la volonté et le courage pour élaborer ce travail

Nos sincères gratitudes à notre encadrant Mme Dendani de nous encadrer et de nous accompagner pour la réalisation de ce mémoire.

Nous vous sommes très reconnaissantes de nous avoir accordé la chance de travailler sur ce thème très intéressant.

Enfin, à ceux qui comptent le plus à nos yeux, à nos chers parents

Dédicaces

Je dédie ce mémoire

A l'homme, mon précieux offre de dieu, qui doit ma vie, ma réussite et tout mon respect : mon cher père ***El Hani****.*

A la femme qui a souffert sans me laisser souffrir, qui n'a jamais dit non à mes exigences et qui n'a épargné aucun effort pour me rendre heureuse : mon adorable mère ***Nora****.*

A toute ma famille ainsi qu'à mes amis.

Table des matières

Tables des figures

Liste des tableaux

Liste des abréviations

RN :	Réseau de neurones.
SVM :	Machine à vecteurs de support (Support Vector Machine).
Kppv(Knn) :	K-plus proche voisins (K-nearest neighbors).
SMO :	Sequential Minimal Optimization.
IBk :	Instance Based Learner.
HGPO :	Test d'hyperglycémie provoquée.
HbA1C :	L'hémoglobine glyquée.
PIDD :	PIMA Indian Diabetes Dataset.
LMS :	Least Mean Square.
FPGA :	Field Programmable GateArrays.
PML (MLP) :	Perceptron multicouche (MultiLayer Percecptron).
FID :	Fédération Internationale du Diabète.
RBF :	Radial Basis Function.

Introduction

Depuis plusieurs années, grâce à l'émergence des nouvelles technologies de l'information et de la communication, l'information médicale est devenue de plus en plus disponible et accessible. Le domaine médical dispose aujourd'hui d'une très grande quantité de données permettant ainsi la recherche d'une information médicale quelconque. Cependant, l'exploitation de cette grande quantité de données rend la recherche et la classification des données médicales précises complexes et coûteuses en termes de temps. Cette difficulté a motivé le développement de nouveaux outils de classification des données adapté.

L'intelligence artificielle est un sous-domaine de l'informatique. Ces derniers temps, l'expression « intelligence artificielle » est fréquemment utilisé dans le public car il s'agit d'un domaine en constante évolution notamment grâce aux progrès des technologies informatiques et entre autres grâce aux capacités toujours plus grandes des machines pour effectuer les calculs.

Le Machine learning ou apprentissage automatique est le fait qu'une machine dotée d'une intelligence artificielle puisse être capable de s'autogérer mais surtout d'être autodidacte.

La classification est une méthode basique de l'apprentissage automatique, l'objectif déborde le cadre strictement exploratoire. C'est la recherche d'une typologie, ou segmentation, c'est-à-dire d'une partition, ou répartition des individus en classes, ou catégories. Ceci est fait en optimisant un critère visant à regrouper les individus dans des classes, chacune le plus homogène possible et, entre elles, les plus distinctes possible.

Plusieurs travaux sont effectués afin de développer des outils de classification, de prédiction et d'aide au diagnostic. Ces systèmes sont même considérés comme étant essentiels dans beaucoup de disciplines médicales car ils permettent à l'expert soit le médecin de réduire au maximum les erreurs et d'assurer un diagnostic plus exact en classant les patients suivant deux catégories : malade ou non malade.

Le diabète est une maladie qui se définit par **une hyperglycémie** (c'est-à-dire un taux de sucre élevé dans le sang en permanence). Le taux de sucre dans le sang ne doit pas dépasser 1,26 G/L **à jeun** Le diabète est **une maladie chronique qui ne se guérit pas**, mais qui se contrôle par un traitement adapté.

Problématique

- Comment aider les médecins dans le processus de prise de décision ?
- Comment réaliser un système de prédiction pour aider les médecins à poser un diagnostic ?
- Quelle méthode ou technique intelligente utilisé pour parvenir à de meilleurs résultats ?

Objectifs

Notre objectif dans ce travail est d'arriver à la meilleure classification possible, d'où l'étude comparative que nous proposons entre quatre algorithmes d'apprentissage (SVM, MLP, NB et Knn) sous PYTHON, afin de savoir lequel offrira les résultats les plus optimaux selon à plusieurs critères.

Les performances obtenues à partir du système sont tout à fait satisfaisantes et concluantes.

Le plan de ce projet de fin d'étude s'articule autour de trois chapitres :

- Chapitre 1 présente un aperçu général sur la maladie du diabète en citant les causes, symptômes, traitements, complications et le diagnostic.
- Chapitre 2 décrit les algorithmes et techniques de classification utilisé dans ce mémoire soient réseau de neurones, machine à vecteurs de support, naive bayeset et k-plus proche voisin .
- Chapitre 3 constitue le cœur de ce mémoire, où on expose la conception du travail, et l'implémentation de notre outil ainsi que les performances obtenues.

Nous clôturons ce mémoire avec une conclusion générale et des perspectives.

Chapitre1 : Diabète sucré

1. Introduction :

De nombreuses maladies constituent un problème de santé publique et d'éthique tant par la morbidité que par la mortalité. Entre autres, on site HTA, dyslipidémie, les insuffisances organiques et le diabète. Cette dernière touche plus de 463 millions de la population mondiale et est responsable de 1.5 millions de décès par an [web3]. Connaitre le diabète et ses facteurs favorisant, ainsi que sa physiopathologie et traitement est devenue une obligation aussi bien pour les sujets atteints que par les indemnes afin de permettre un suivi, une prévention et une lutte efficaces contre cette pathologie.

2. Définition :

Le diabète est une maladie chronique incurable qui se caractérise par l'impossibilité du corps à contrôler et utiliser convenablement la quantité de glucose présente dans le sang comme source d'énergie. Cette maladie est causée par un manque ou un défaut d'utilisation d'une hormone appelé insuline. Cette dernière qui est produite par les cellules bêta du pancréas, permet l'entrée du sucre dans les cellules du corps pour y être stockés en cas de besoin.

3. Type du diabète sucré :

1. Diabète de type 1 :

Le diabète de type 1 ou appelé diabète insulinodépendant apparait le plus souvent pendant l'enfance, à l'adolescence ou au début de l'âge adulte, rarement chez les personnes âgées. Il est caractérisé par l'absence totale de production d'insuline [web4].

2. Diabète de type 2 :

Connu sous le nom de diabète non-insulinodépendant, le diabète de type 2 touche 90% de la population diabétique. Il se manifeste principalement chez les personnes âgées de plus de 40ans.

Une prédisposition génétique, obésité et le manque d'activité physique contribuent à l'apparition de ce type de diabète, ajoutant à cela une insulinorésistance résultant d'une mauvaise utilisation de l'insuline et/ou une insulinopénie c'est-à-dire une carence en insuline [web5].

3. Diabète gestationnel :

C'est un état d'intolérance transitoire au glucose, il apparait au cours de la grossesse, généralement au cours du 3éme trimestre chez une femme sans diabète sucré comme antérieur. Il peut néanmoins être révélateur d'un diabète antérieur et disparaître après l'accouchement.

4. Diabète néonatal :

C'est un diabète **insulinodépendant** qui apparait durant les premiers mois de vie. Il peut être transitoire (l'insuline peut être arrêtée, généralement avant l'âge de 6 mois, mais une récidive du diabète est possible à la puberté ou à l'âge adulte) ou permanent (le traitement ne peut pas être arrêté) **[web6].**

4. Symptômes, causes et traitements du diabète :

TYPES	SYMPTOMES	CAUSES	TRAITEMENT ET PREVENTION
TYPE I	- Une polydipsie (soif importante). - Augmentation du volume des urines (polyurie). - Fort amaigrissement (malgré un appétit augmenté). - Vision trouble. - Asthénie. - Haleine à l'odeur fruitée de pomme verte.	- Les causes de ce type de diabète sont encore méconnues en ce jour. Néanmoins, certains chercheurs pensent que la mauvaise alimentation, la sédentarité et la prédisposition génétique peuvent être des facteurs de risque.	- Exige une insulinothérapie, des solutions adaptées à chaque patient sont proposées : • Multi injections. • Pompe à insuline. - Une bonne nutrition est importante.
TYPE II	- Une soif intense et une faim exagérée (polyphagie). - Somnolence et fatigue constante. - Besoin fréquent d'uriner (particulièrement le soir). - Vision brouillée. - Infections des organes vitaux. - Cicatrisation lente.	- La génétique et l'hérédité. - Le surpoids et l'obésité. - L'hypertension artérielle élevée. - Survenue d'un diabète sucré durant une grossesse. - Taux élevé de cholestérol. - Intolérance au glucose.	- L'adoption d'une meilleure alimentation. - Pratique régulière d'une activité physique. - Prise des médicaments tels que des antidiabétiques oraux. - Hygiène corporelle.
GESTATIONNEL	- Fatigue importante. - Soif accrue. - Des urines plus abondantes. - Envie plus fréquente d'uriner. - Maux de tête.	- La résistance des cellules à l'action de l'insuline causée durant la grossesse par les hormones du placenta. - Une importante prise de poids entre deux grossesses.	- Modifications de l'alimentation maternelle. - Contrôle du poids des femmes. - Une bonne hygiène de vie.
NEONATAL	- Déshydratation. - Hyperglycémie souvent très élevée (5 g/l en moyenne). - Insulinémie (concentration d'insuline dans le sang) basse.	- Une anomalie au niveau du chromosome 6. - Une mutation du gène Kir6.2.	- Réhydratation. - Insuline : injections sous cutanées ou pompe. - Alimentation adaptée au nouveau-né.

Tableau 1 : Symptômes, causes et traitement du diabète.

5. Les complications du diabète :

Quel qu'en soit le type, le diabète peut entraîner des complications qui affectent plusieurs parties de l'organisme et accroître le risque général de décès prématuré. Au nombre des complications possibles figurent l'infarctus du myocarde, l'accident vasculaire cérébral, l'insuffisance rénale, l'amputation des jambes, la perte de vision et des lésions nerveuses. Pendant la grossesse, un diabète mal maîtrisé accroît le risque de mortalité intra-utérine et d'autres complications [1].

La majorité des complications, liées au diabète, peuvent être évitées, diminuées ou retardées si le diabète est dépisté et traité précocement et correctement.

Qu'il s'agisse du type1, du type2, du diabète de grossesse ou du diabète néonatal, une consultation chez le médecin s'impose. Généralement, la simple mesure de la glycémie à jeun, par prise de sang, suffit pour dépister un diabète.

6. Diagnostic du diabète :

Le diagnostic du diabète est établi grâce à une prise de sang qui dose le taux de sucre dans le sang. Pour les périodes de mesure nous avons :

- Taux de glycémie à jeun :

Le diagnostic est posé lorsque cette glycémie à jeun soit la prise de sang qui a eu lieu sans apport calorique pendant huit heures au moins est égale ou supérieure à 1.26 g/l (ou 7mmol/L) et est constatée à deux reprises.

- Le test d'hyperglycémie provoquée (HGPO) :

Ce test consiste à mesurer les taux de variations de la glycémie soit le taux de sucre dans le sang après avoir ingéré 75g d'hydrate de carbone (glucides) sous forme de boisson. Le sucre sanguin est mesuré durant les 8 heures de jeûne puis toutes les 2 heures après l'ingestion de la solution. Si le taux de la glycémie est supérieur à 11.1mmol/L, on diagnostique un diabète, ou bien un pré-diabète si le taux de glucose se trouve entre 7.8mmol/L et 11.0mmol/L.

Il permet de détecter un diabète de type 2 ou un diabète gestationnel chez la femme enceinte.

- L'hémoglobine glyquée (HbA1c) :

Le dosage de l'hémoglobine glyquée ou l'HbA1c s'effectue par prise de sang dans un laboratoire d'analyse médicale. Il est préconisé à intervalles réguliers, tous les 2 à 3 mois environ. Il n'est pas nécessaire d'être à jeun pour la prise de sang.

7. Conclusion :

Dans ce chapitre, nous avons présenté la maladie du diabète sucré, qui est le domaine d'application de ce projet ainsi que tous les concepts qui la caractérise que ce soit les symptômes, causes, traitements, diagnostique et complications. Dans le prochain chapitre, nous décrirons les méthodes de classification retenues pour notre étude.

Chapitre 2: Description détaillée des techniques de classification

1. Introduction

Les méthodes de classification ont pour but d'identifier les classes auxquelles appartiennent des objets à partir de certains paramètres descriptifs. Elles s'appliquent à un grand nombre d'activités humaines et conviennent en particulier au problème de la prise de décision automatisée.

Les méthodes utilisées pour la classification sont nombreuses, citons : les réseaux de neurones, machine à vecteurs de support, la méthode de naïve bayésienne…etc

Nous présentons dans la suite de ce chapitre, une étude détaillée des quatre techniques SVM, kplus proche voisins, NB et réseaux de neurones. Ces méthodes ont montrés leurs efficacités dans des domaines d'applications très variés tels que le traitement d'image, la bioinformatique, la finance, la catégorisation de textes et le diagnostic médical [2].

2. Les réseaux de neurones (RN) :

2.1 Définition :

Un réseau neuronal est un modèle mathématique de calcul dont la conception est inspirée de la structure des aspects fonctionnels des réseaux de neurones biologiques. Les principaux réseaux se distinguent par l'organisation du graphe(en couches, complets…), c'est-à-dire leur architecture, son niveau de complexité(le nombre de neurones), par le type de neurones (leurs fonctions de transition ou d'activation) et enfin l'objectif visé : apprentissage supervisé ou non, optimisation, systèmes dynamiques…etc **[3].**

Les neurones sont organisés en couches successives dont chacune prend ses entrées sur les sorties de la précédente. Chaque couche (i) est composée de N_i neurones, prenant leurs entrées sur les N_{i-1} neurones de la couche précédente. À chaque synapse est associé un poids synaptique, de sorte que les N_{i-1}sont multipliées par ce poids, puis additionnées par les neurones de niveau i, ce qui est équivalent à multiplier le vecteur d'entrée par une matrice de transformation **[web8].**

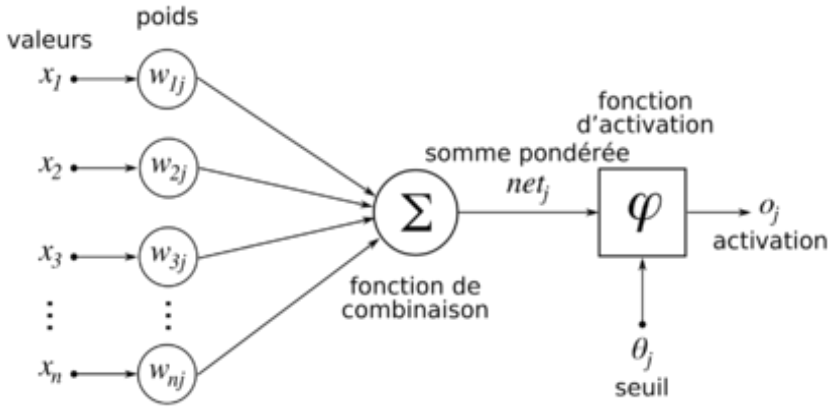

Figure 1.1 : Structure d'un réseau de neurone artificiel.

2.2 Perceptron multicouche (PML) :

Le perceptron multicouche est un modèle non linéaire, le plus utilisé dans les réseaux de neurones. Il peut contenir deux ou plusieurs couches (Figure1.2). La couche d'entrée est la première couche dont le nombre de neurones égal au nombre de caractéristiques spécifiques sélectionnées. La couche de sortie est la dernière couche qui détermine la sortie, le nombre de neurones dans cette couche dépend du nombre de classes désirées. Les couches cachées représentent des résultats d'analyse intermédiaires afin de renforcer la capacité d'apprentissage du réseau.

Le réseau de neurone multicouche est basé sur l'algorithme de la retro propagation pour rétro propagé l'erreur entre l'état des neurones de sortie et la réponse désirée sur les poids de connexions des couches antérieures.

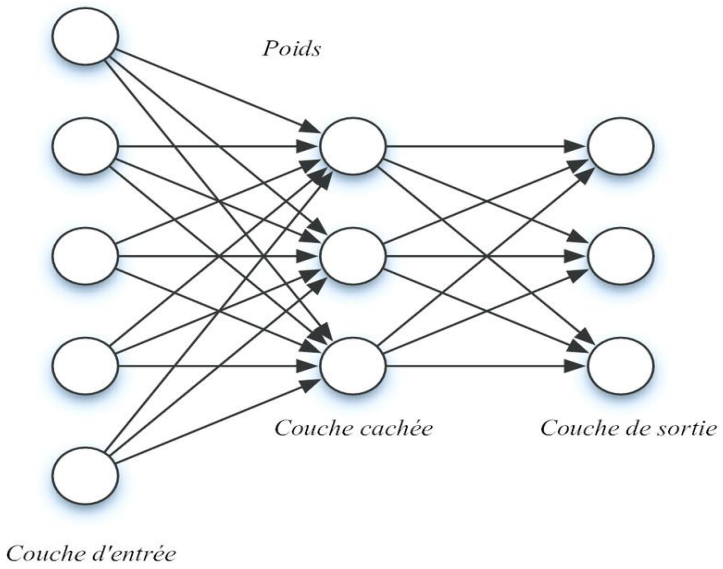

Figure 1.2 : Un réseau perceptron multicouche.

2.3 L'algorithme d'apprentissage :

Initialisation des poids du réseau.

Présentation du vecteur d'apprentissage à l'entrée du réseau (couche d'entrée).

Calcul du vecteur de sortie S : le système propage les activités neuronales à travers le réseau. Chaque neurone calcule la somme pondérée de ses entrées et transmet le résultat par une fonction de type sigmoïde pour produire sa valeur de sortie. Le vecteur de sortie S est le résultat du calcul de la dernière couche (couche de sortie).

Calcul de l'erreur. On compare alors les valeurs de sortie actuelles S avec les valeurs désirées de R appelé vecteur de référence.

Rétro propagation de l'erreur : l'algorithme de rétro propagation du gradient permet récursivement de rétro propager l'erreur de la couche de sortie vers les couches cachées jusqu'à la première couche du réseau.

Modification des poids : les poids de chaque neurone sont modifiés, soit à chaque présentation d'un vecteur d'apprentissage (méthode du gradient stochastique), soit après cumul de l'erreur pour un certain nombre de vecteurs d'apprentissage (méthode du gradient standard) [4].

2.4 Avantages et inconvénients des RN :

Avantages

- Possibilité de faire le parallélisme (les éléments de chaque couche peuvent fonctionner en parallèle).
- Apprentissage automatique des poids.
- Résistance aux pannes c'est-à-dire que si un neurone ne fonctionne plus, le réseau ne se perturbe pas.
- De nouvelles variables peuvent être présentées en entrée pour améliorer les prévisions.

Inconvénients

- Difficile de choisir la structure (type, nombre de nœuds, connexions…) la mieux adaptée au problème.
- Paramètres difficiles à interpréter (boite noire).
- Difficulté de paramétrage surtout pour le nombre de neurones dans la couche caché.

3. Machine à vecteurs de support (SVM) :

3.1 Définition :

Une machine à vecteurs de support, est un algorithme d'apprentissage automatique supervisé qui permet de résoudre des problèmes tant de classification que de régression ou de détection d'anomalie. Ils sont connus pour leurs solides garanties théoriques, leur grande flexibilité ainsi que leur simplicité d'utilisation même sans grande connaissance de data mining.

Les SVMs reposent sur l'idée de trouver un hyperplan soit une frontière dans le but de séparer les données en classes, de telle façon que la distance entre les différents groupes de données et la frontière qui les sépare soit maximale. Cette distance est aussi appelée « marge » et les SVMs sont ainsi qualifiés de « séparateurs à vaste marge », les « vecteurs de support » étant les données les plus proches de la frontière.

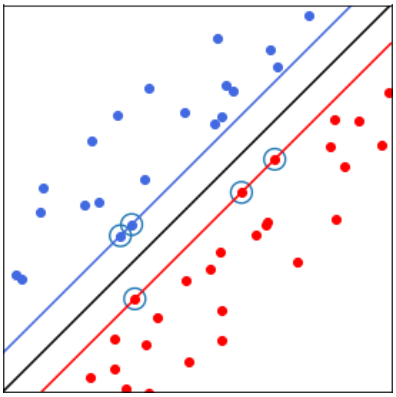

Figure 2.1 : Exemple de classe linéairement séparable.

Cette notion de frontière suppose que les données soient linéairement séparables, ce qui est rarement le cas. Pour y pallier, les SVMs reposent souvent sur l'utilisation de « noyaux ». Ces fonctions mathématiques permettent de séparer les données en les projetant dans un espace vectoriel (Figure2.2). La technique de maximisation de marge permet, quant à elle, de garantir une meilleure robustesse face au bruit et donc un modèle plus généralisable.

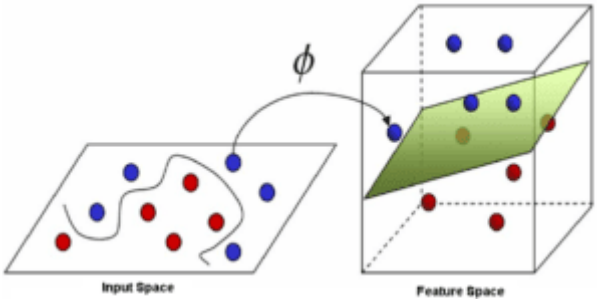

Figure 2.2 : Exemple d'un problème non linéairement séparable.

3.2 Classifieur linéaire :

Un classifieur est dit linéaire lorsqu'il est possible d'exprimer sa fonction de décision par une fonction linéaire comme suit : $h(x) = wTx + b = \sum wixi + b\ n\ i=1$

Où $w \in Rn$ est le vecteur de poids et $b \in R0$ le biais, alors que x est la variable du problème. X est l'espace d'entrée et qui correspond à Rn, où n est le nombre de composantes de vecteurs contenant les données.

Pour décider à quelle catégorie un exemple estimé x' appartient, il suffit de prendre le signe de la fonction de décision : y = sign ((h (x')). La fonction sign () est appelée classifieur [2].

A toute fonction de décision linéaire on peut y associer une frontière de décision : □(w,b) = $\{x \in Rn \mid wTx + b = 0\}$. Tout comme la fonction de décision linéaire, cette frontière de décision est définie à un terme multiplicatif prés dans le sens ou la frontière définie par le couple (w,b) est la même engendrée par (kw,kb) ∀ k ∈ R. Cela est lié à la définition de l'hyperplan affine associé à la fonction caractéristique. Pour garantir l'unicité de la solution on peut soit considérer l'hyperplan standard (tel que ||w|| = 1) soit l'hyperplan canonique par rapport à un point x (tel que $wTx + b = 1$).

3.3 Marge maximale de l'hyperplan :

Dans les SVMs, la frontière de séparation est choisie comme celle qui maximise la marge. La marge géométrique représente la distance euclidienne prise perpendiculairement entre l'hyperplan et l'exemple xi. En prenant un point quelconque xj se trouvant sur l'hyperplan, la marge géométrique peut s'exprimer par : $w\ ||w||$. (xi-xj)

L'hyperplan à marge maximale est le modèle le plus utilisé dans les machines à vecteurs supports. L'estimation des paramètres (w*, b*) de l'hyperplan qui maximise la marge se fait en résolvant le problème d'optimisation suivant :

$(w^*,b^*) = \arg(w,b)\{mini(yi(wxi + b)), ||w|| = 1\}$

Dire que les deux classes de l'échantillon d'apprentissage S sont linéairement séparable est équivalent à dire qu'il existe des paramètres (w*,b*) tels que l'on a pour tout i=1,2,…,n :

$w^*xi + b* > 0\ siyi = 1$

$w^*xi + b* < 0\ siyi = -1$

Ce qui est équivalent à : $(w*xi + b*) > 0$; ∀ $i = 1,2, \dots, n$

La définition consiste à dire qu'il doit exister un hyperplan laissant d'un côté toutes les données positives et de l'autre, toutes les données négatives. Dès lors, on peut définir deux plans se trouvant

de part et d'autre de l'hyperplan et parallèle à celui-ci, sur lesquels reposent les exemples les plus proches. La figure 2.3 illustre cette situation [2].

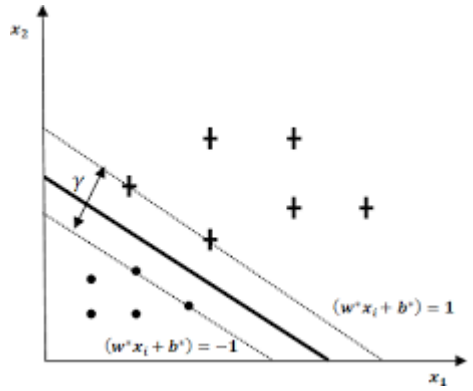

Figure 2.3 : Hyperplan marge maximale.

3.4 Le cas des multi-classes :

La plupart des problèmes ne se contentent pas de deux classes de données. Il existe plusieurs méthodes pour faire la classification multi-classes. Les plus utilisées sont :

La première méthode, est une méthode dite un-contre-un. Au lieu d'apprendre N fonctions de décisions, ici chaque classe est discriminée d'une autre.

La deuxième méthode est appelé un-contre-tous. C'est une approche étendant la notion de marge aux multi-classes. Cette formulation intéressante permet de poser u problème d'optimisation unique. Le problème fait intervenir N fonctions de décision [2].

3.5 Avantages et inconvénients des SVMs :

Avantages

- Les SVMs possèdent des fondements mathématiques solides.
- Décision rapide. La classification d'un nouvel exemple consiste à voir le signe de la fonction de décision f(x).
- Sa grande précision de prédiction.
- Ils peuvent être plus efficaces car ils utilisent un sous-ensemble de points d'entraînement.

Inconvénients

- Classification binaire d'où la nécessité d'utiliser l'approche un-contre-un.
- Grande quantité d'exemples en entrées implique un calcul matriciel important.
- Temps de calcul élevé lors d'une régularisation des paramètres de la fonction noyau.

- Moins efficace sur les jeux de données contenant de bruits.

4 K-plus proche voisin (Kppv/Knn) :

4.1 Définition :

La méthode k-plus proche voisins (k-nearest neighbors en anglais) est une méthode non paramétrique très intuitive où une nouvelle observation est classée dans la classe d'appartenance de l'observation de l'échantillon d'apprentissage qui lui est la plus proche, au regard des co-variables utilisées. La détermination de leur similarité est basée sur des mesures de distance. L'algorithme des k-plus proches voisins est un algorithme intuitif, aisément paramétrable pour traiter un problème de classification avec un nombre quelconque d'étiquettes [5].

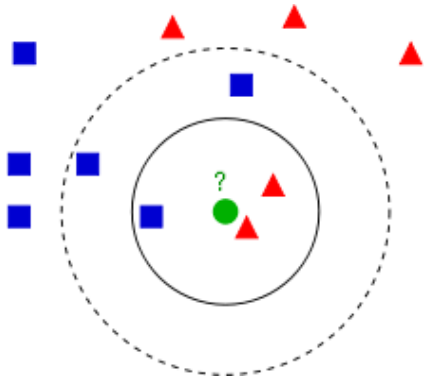

Figure 3 : Méthode des k-plus proches voisins.

Dans l'exemple de la figure 3, l'échantillon de test (cercle vert) pourrait être classé soit dans la première classe de carré bleu ou la seconde classe de triangles rouges. Si k = 3 (cercle en ligne pleine) il est affecté à la seconde classe car il y a deux triangles et seulement un carré dans le cercle considéré. Si k = 5 (cercle en ligne pointillée) il est affecté à la première classe (3 carrés face à deux triangles dans le cercle externe) [web9].

La méthode des k-plus proche voisins a l'avantage d'être très simple à mettre en œuvre et d'utiliser directement l'ensemble d'apprentissage. Elle ne fait aucune hypothèse a priori sur les données. La qualité de la discrimination par cette méthode dépend du choix du nombre k de voisins considérés.

4.2 Choix de k :

La valeur de *k* est un des paramètres à déterminer lors de l'utilisation de ce type de méthode. La valeur que l'on choisit pour k va être plus critique, plus déterminante en rapport avec la performance du classificateur (Figure 3). On peut se permettre de considérer un plus grand nombre de voisins, sachant que plus ils diffèrent du document à classer, moins ils ont d'impact sur la prise de décision. Cependant, il demeure nécessaire de limiter le nombre de voisins pour s'en tenir à un temps de calcul raisonnable.

L'emploi de *k* voisins, au lieu d'un seul, assure une plus grande robustesse à la prédiction. Classiquement, dans le cas où la variable à prédire comporte deux étiquettes, ce paramètre *k* est une valeur impaire afin d'avoir une majorité plus facilement décidable et éviter les votes égalitaires **[6].**

4.3 L'algorithme des k-plus proches voisins :

L'algorithme des k-plus proches voisins est un des algorithmes de classification les plus simples d'apprentissage automatique supervisé. Le seul outil dont on a besoin est une distance entre les éléments que l'on veut classifier. Si on représente ces éléments par des vecteurs de coordonnées, il y a en général pas mal de choix possibles pour ces distances, partant de la simple distance usuelle (euclidienne) en allant jusqu'à des mesures plus sophistiquées pour tenir compte si nécessaire de paramètres non numériques comme la couleur, la nationalité, etc.

ALGORITHME DES K-PLUS PROCHES VOISINS :

Input : Données d'apprentissage ; $X^{train} = (x_1^{train},\ldots,x_n^{train})$; classes des données d'apprentissage $Z^{train} = (z_1^{train},\ldots,z_n^{train})$; $X^{test} = (x_1^{test},\ldots,x_m^{test})$

Algorithme Knn ;

For $i \leftarrow$ 1 **to** m **do**

  For $j \leftarrow$ 1 **to** n **do**

    Calculer la distance euclidienne entre x_i^{test} et x_j^{train} en utilisant l'équation

    $d_j \leftarrow d(x_i^{test}, x_j^{train})$

  End

  Calculer la classe z_i^{train} de l'iéme exemple qui vaut la classe de son pvv :

  Trouver l'indice du pvv de x_i^{test} :

  $Ind\_pvv_i \leftarrow \text{argmin}_{j=1}^{n} d_j$

  Trouver la classe du pvv de x_i^{test} (qui est $x_{ind\_pvv_i}^{train}$) :

  $z_i^{test} = z_{ind\_pvv_i}^{train}$

End

Result : classes des données de test $Z^{test} = (z_1^{test},\ldots,z_n^{test})$

4.3 Avantages et inconvénients des Kppv :

Avantages

- Simplicité, apprentissage rapide.
- Bonnes performances en général.
- Méthode facile à comprendre.
- Adapté aux domaines où chaque classe est représentée par plusieurs prototypes et où les frontière sont irrégulières (ex. Reconnaissance de chiffre manuscrits ou d'images satellites) **[8].**

Inconvénients

- Paramétrage difficile (choix de la taille du voisinage).
- Lenteur en classement (passage en revue de tous les individus).
- Méthode gourmande en place mémoire.
- Sensibilité à la dimensionnalité (et aux variables non pertinentes et corrélés) **[9].**

5 Naïve bayésienne

5.1 Définition :

La classification naïve bayésienne est un type de classification bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses. Elle met en œuvre un classifieur bayésien naïf, ou classifieur naïf de Bayes, appartenant à la famille des classifieurs linéaires.

Un terme plus approprié pour le modèle probabiliste sous-jacent pourrait être « modèle à caractéristiques statistiquement indépendantes ».

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$$
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Théorème de Bayes

P(A) : désigne la probabilité à priori de A.

P(A|B) : désigne la probabilité a posteriori de A sachant B (ou encore de A sous condition B).

P(A∩B) : désigne la probabilité que A et B aient tous les deux lieu.

5.2 Estimation de la valeur et des paramètres :

Tous les paramètres du modèle (probabilités *a priori* des classes et lois de probabilités associées aux différentes caractéristiques) peuvent faire l'objet d'une approximation par rapport aux fréquences relatives des classes et caractéristiques dans l'ensemble des données d'entraînement. Il s'agit d'une estimation du maximum de vraisemblance des probabilités.

5.3 Avantages et inconvénients :

Avantages :

- Simplicité, rapidité de calcul, capacité à traiter de très grandes bases.
- Modèle linéaire même niveau de performances.
- Incrémentalité (table des probas conditionnelles à maintenir).
- Robustesse (performant même si hypothèse non-respectée).

Inconvénients :

- Nombre de règles égal au nombre de combinaisons de descripteurs.
- Pas de modèle explicite.
- Très utilisé en recherche, peu en marketing.

6 Les travaux similaires :

Ubeyli a obtenu un taux de 98.14% en appliquant le classifieur ANFIS pour la reconnaissance du diabète et la technique de validation croisée de K-Fold sur la base de données PIDD **[10].**

Les chercheurs *Pradhan* et *Sahu* ont proposé un perceptron multicouche et l'algorithme génétique pour classer les sujets diabétiques ou non. Cette classification a été appliquée sur la base de données du diabète PIDD, ils ont obtenu une précision de 72.2% **[11].**

En 2011, *Bayu Adhi Tama & al*. ont réalisé une série d'expériences pour prédire le diabète sur un ensemble de données privées. Le SVM a surpassé les autres classifieurs ainsi que les méthodes basées sur un ensemble avec une précision moyenne de 96,49% en utilisant hold out et le test de validation croisée (10-fold cross validation) **[12].**

En 2012, *Asha Gowda Karegowda, M.A. Jayaram, A.S. Manjunath* dans leurs article, ont proposé une catégorisation des patients diabétiques utilisant K moyenne en cascade et K-plus proche

voisin. Ils ont classé les patients diabétiques grâce aux résultats obtenus de l'utilisation de Kppv et K moyenne. Le système proposé a obtenue une précision de 82% **[13].**

Kumari & al. ont appliqué SVM avec le noyau RBF sur la base de données PIDD. Ils ont réalisé une précision de 78%, 80% de sensibilité et 76.5% de spécificité **[14].**

Akanksha a exposé un système qui emploi une interface floue en cascade avec un réseau de neurone Feed-Forward afin d'obtenir une décision optimale concernant l'état physiologique et pathologique futur d'un patient. Le classifieur neuro-flou a été implémenté sur un FPGA (implantation *Hardware*) pour une détection à temps réel de l'état critique d'un sujet diabétique. La précision du système a été confirmée en prédisant l'état diabétique d'un patient 30 jours avant l'état critique **[15].**

Le travail présenté par *Belarouci*, propose une méthode de pondération basée sur l'algorithme des moindres carrés moyens LMS dans le but d'affecter des poids forts aux différents échantillons de la classe minoritaire et des poids faibles aux différents échantillons de classes majoritaires afin de traiter le problème d'apprentissage des données déséquilibrées. Après cette phase, Ils se sont servis de plusieurs approches de classifications comme le perceptron multicouche qui a réalisé une bonne reconnaissance des données. La performance du classifieur a été évaluée en utilisant la base de données PIDD à un taux de classification de 99.24% **[16].**

Zhilbert Tafa & al. en 2017, ils ont proposé une implémentation conjointe du SVM et Naïve Bayes appliqué sur la base de données des diabétiques de Kosovo. Le rapport de répartition pour l'ensemble d'apprentissage et de test étant de 50 :50%, le taux de classification a été élevé à 95,25% et 94 ,52% pour SVM et Naïve Bayes respectivement **[17].**

Pradhan et al. ont conçu un classifieur pour la détection du diabète à l'aide du réseau neuronal et l'algorithme flou du kppv. Les résultats obtenus avec l'utilisation de WEKA a montré que l'ensemble de formation a fourni une précision de 100% par rapport à 10CV qui a donné 73,047% en utilisant 768 instances de PIDD **[18].**

7 Conclusion :

Ce chapitre était destiné à présenter de manière simple et complète les différentes méthodologies de la classification sélectionnées pour une étude comparative. Dans le chapitre suivant, nous exposerons la partie conception réalisation de notre système.

Chapitre 3 : Conception et Réalisation

1. Introduction

Dans les chapitres précédents nous avons exposé les méthodes d'apprentissage et les différents outils de classification data mining. Nous présentons dans cette dernière partie notre étude comparative, nous avons réalisé une étude comparative entre les résultats des quatre outils de classification des données médicales. Plusieurs méthodes sont utilisées dans ce domaine nous nous sommes intéressés à quatre méthodes qui sont: Knn, SVM , BN et RN. Nous avons choisi ces quatre méthodes car elles sont très utilisées dans la littérature.

La figure 4.1 donne une vue générale de notre travail.

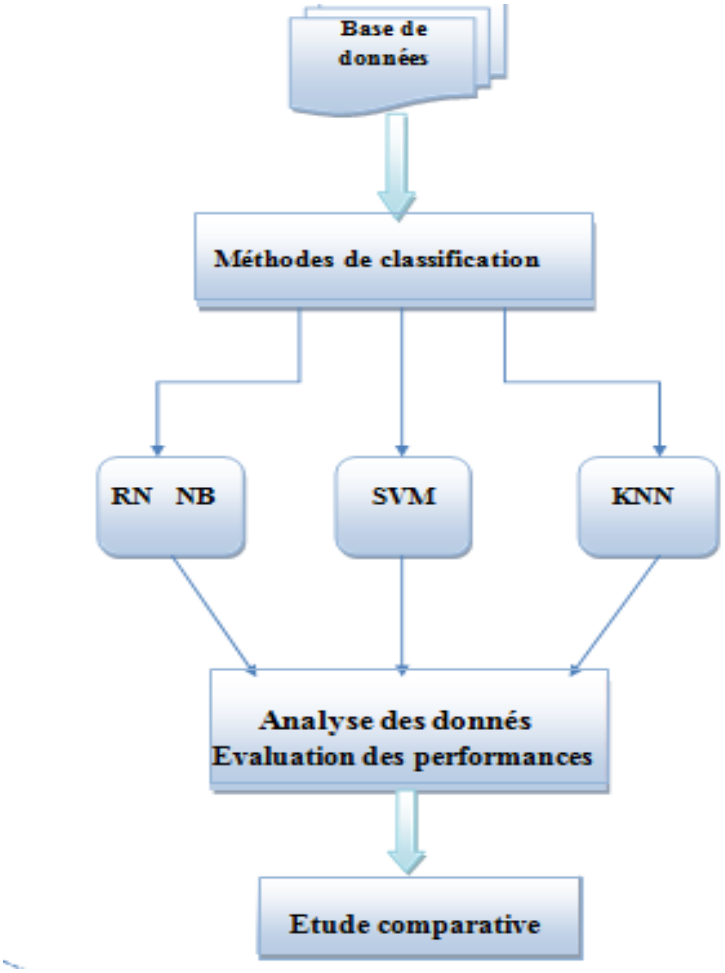

Figure 4.1 : Architecture générale du système.

2. Base de données

Pour mener à bien notre étude comparative, nous avons choisi une base de donnée médicale réelle extraite du dépôt d'UCI [27] qui est: Pima. Le choix de cette base est justifié par les critères suivants :

1. La taille de la base.
2. Nombre des attributs.
3. Nombre de classe.

Description de la base de données Pima :

La base Pima indien Diabète est constitué de 768 cas dont 268 sont diabétique et 500 non diabétique. Chaque cas est formé de 9 attributs, dont 8 représentent des facteurs de risque et le 9eme représente la classe du patient. Les huit descripteurs cliniques sont :

1. Npreg : nombre de grossesses,
2. Glu : concentration du glucose plasmatique,
3. BP : tension artérielle diastolique
4. SKIN : épaisseur de pli de peau du triceps,
5. Insuline : dose d'insuline,

06. BMI : index de masse corporelle,

7. PED : fonction de pedigree de diabète (l'hérédité)
8. Age : âge

Et 2 classes (1 ET 0)

- Si classe =1 implique Présence de la maladie
- 0Si classe =0 implique Absence de la maladie

3. Critères d'évaluation et Mésure de performance

Le critère d'évaluation est un facteur clé à la fois dans l'évaluation de la performance de classification et guidance de la modélisation de classificateur. Pour comparer de façon synthétique les performances des différentes méthodes et de différents outils retenues pour notre étude nous avons calculé : le taux de classification(TC), la sensibilité(SE) et la spécificité(SP), leurs définitions respectives sont les suivantes :

TC =100*[(VP + VN)/ (VP + VN + FP + FN)]

Se =100*[(VP/VP + FN)]

Sp =100*[(VN/VN + FP)]

- Vrai positive (VP) : les cas positives classé positives

- Vrai négative (VN) : les cas négatif classé négative
- Faux positive (FP) : les cas positive classé négative
- Faux négative (FN) : les négatives classées positive
- Taux de classification : le pourcentage des exemples classé correctement
- Sensibilité (Se) : le pourcentage des exemples positive classé correctement
- Spécificité (Sp) : le pourcentage des instances négative classé correctement

Les mesures de performance utilisées sont :

- Accuracy
- Precision
- Recall
- F1 score
- Auc_Roc curve

3.1 Accuracy

Mesure la capacité du modèle à correctement classifier toutes les classes.

Formule de calcul : Accuracy = $\frac{tp+tn}{N}$

3.2 Precision

Mesure la capacité du modèle à ne pas classifier comme positive une instance négative.

Formule de calcul : Precision = $\frac{tp}{tp+fp}$

3.3 Recall

Mesure la capacité du modèle à retrouver toutes les instances positives.

Formule de calcul : Recall = $\frac{tp}{tp+fn}$

3.4 F1 score

Représente la moyenne pondérée de Precision et Recall. Ces deux dernières apportent une contribution relative équitable au F1 score.

Formule de calcul : F1 score = $2 * \frac{precision*recall}{precision+recall}$

3.5 Auc_Roc curve

La courbe AUC - ROC est une mesure de performance pour les problèmes de classification à divers réglages de seuil. ROC est une courbe de probabilité et AUC représente le degré ou la mesure de séparabilité. Il indique dans quelle mesure le modèle est capable de faire la distinction entre les classes. Plus l'AUC est élevée, mieux le modèle prédit 0 classes comme 0 et 1 classes comme 1. Par analogie, plus l'AUC est élevée, plus la distinction est meilleure entre les patients atteints de la maladie et les patients sans maladie.

La courbe ROC est tracée avec TPR par rapport au FPR où TPR est sur l'axe des y et FPR est sur l'axe des x.

4 Résultat et Interprétation :

Cette partie relate les résultats de l'étude comparative entre les quatre outils de classification des données en utilisant les quatre algorithmes (RN, BN, SVM, Knn) appliqués au base de donnée (Pima).

L'objectif de chaque algorithme est de parvenir à mieux classer les futures observations en minimisant l'erreur de classification. La question que nous nous sommes posé est : Quel est le meilleur outil? Ou quel outil donne les meilleurs résultats ?

4.1 Méthode de classification :

4.1.1 Méthode RN :

L'objectif est de prédire la classe des futures observations en mesurant le taux d'erreur entre les sorties désirées et les sorties observées par l'algorithme. Au cours de l'apprentissage, chaque instance est présentée à l'entrée du réseau et les poids de chacun des neurones sont réajustés de manière à minimiser l'erreur. L'apprentissage s'arrête quand l'erreur diminue jusqu'à atteindre un niveau 0 ou constant.

Paramètre	Valeur
Interface graphique	Désactivée
Taux d'apprentissage	0.3
Momentum	0.2
Nombre d'époques	500
Couches caches	a = (nombre attributs + nombre classes)/2 = (23+6)/2 = 14

Tableau 3 : Paramètres utilisés pour RN.

4.1.2 Méthode RN :

Le principe est de maximiser la marge entre les classes en utilisant un minimum de vecteurs de support. Dans sa phase d'apprentissage, l'algorithme utilise les données livrées en exemple pour découvrir un ensemble de fonctions linéaires qui permettent de séparer le mieux les différentes classes. L'algorithme parvient à séparer les classes avec une fonction simple en transportant les données dans un espace de dimension plus élevé. Suite à l'apprentissage, les frontières de chaque classe sont définies. Il suffit alors de présenter une nouvelle instance pour directement connaître sa classe prédite [19].

Paramètre	Valeur
Complexité (C)	1.0
Tolérance	0.001
Nombre de partitions	-1
Epsilon	1.0E-12
Noyau	Polynomial

Tableau 4 : Paramètres utilisés pour SVM.

4.1.3 Méthode KNN :

Cet algorithme détermine la classe d'une instance selon la classe majoritaire chez les K plus proches voisins. Pendant la phase d'apprentissage, les données livrées en exemple sont mémorisées de façon à optimiser la recherche des voisins. Par la suite, lorsqu'une instance doit être classifiée, on trouve parmi les K voisins les plus proches selon la distance Euclidienne la classe qui est la plus populaire. L'instance est alors attribuée à cette classe.

Paramètre	Valeur
Nombre de voisins	1
Erreur quadratique moyenne	Désactivée
Algorithme de recherché	Linéaire avec distance Euclidienne

Tableau 5 : Paramètres utilisés pour Knn.

4.1.4 Méthode NB :

La méthode de classification naïve bayésienne est un algorithme d'apprentissage supervisé (supervised machine learning) qui permet de classifier un ensemble d'observations selon des règles déterminées par l'algorithme lui-même.

Tolérance	0.001
Nombre de partitions	-1
Epsilon	1.0E-12
Noyau	Polynomial

Tableau 6 : Paramètres utilisés pour NB.

4.2 Résultat et interprétation :

Résultat 1 :

	Diabetic	No_diabetic
Diabetic	*VP = 110*	*FN = 0*
No_diabetic	*FP = 0*	*VN = 45*

Tableau 7 : Matrice de confusion pour MLP.

	Diabetic	No_diabetic
Diabetic	*VP = 108*	*FN = 2*
No_diabetic	*FP = 0*	*VN = 45*

Tableau 8: Matrice de confusion pour SVM.

	Diabetic	No_diabetic
Diabetic	*VP = 110*	*FN = 0*
No_diabetic	*FP = 0*	*VN = 45*

Tableau 9: Matrice de confusion pour Knn.

	Diabetic	No_diabetic
Diabetic	*VP = 108*	*FN = 2*
No_diabetic	*FP = 0*	*VN = 45*

Tableau 10: Matrice de confusion pour NB.

Résultat 1 :

Les tableaux et courbes suivantes intègrent les résultats de mesures de performances en pourcentages de chaque méthode et hybridation pour dataset en utilisant les modèles de classification. Les diagrammes en barres suivant chaque tableau sont à titre illustratif et offrent une meilleure visualisation des résultats.

	Accuracies	**Precisions**	**Recalls**	**f1_scores**
MLP	70.31	62.5	54.79	58.39
SVM	77.6	86.71	49.32	62.61
NB	76.56	71.88	63.01	67.15
KNN	73.53	69.49	56.16	62.12

Tableau 11: Matrice de confusion pour PIDD data set.

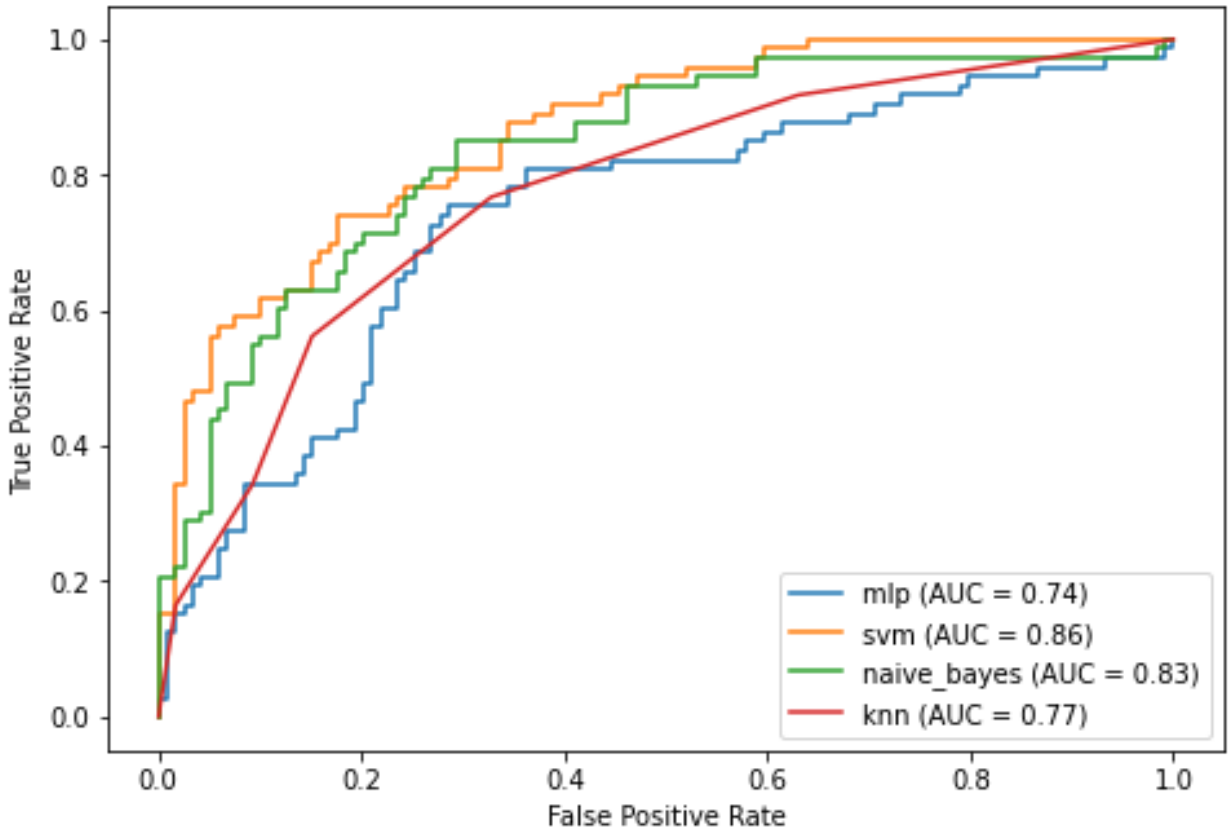

Figure 6 : Courbes ROC des modèles en utilisant le dataset PIMA

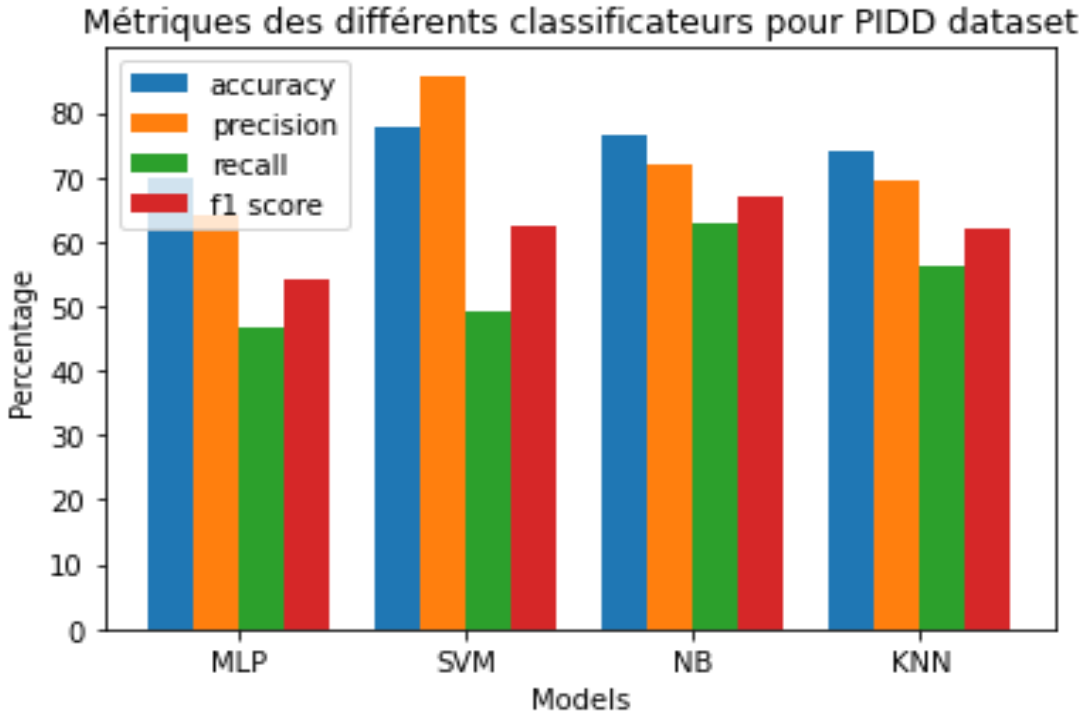

Figure 7 : Comparaison des résultats avec les différents métriques.

Discussion 1:

- L'ensemble des données utilisé pour prédire le diabète est présenté dans la figure 7. Ici, le paramètre diabétique a été considéré comme dépendant variable et le reste des paramètres ont été pris comme variables indépendantes. Le paramètre diabétique ne prend que de valeurs binaires où 0 représente non diabétique et 1 représente diabétique.
- Les quatre techniques de classification (Réseau de neurones, Machine à Vecteurs Support, Naive bayes et k-plus proche voisin) ont ensuite été appliqués à l'ensemble d'apprentissage pour prédire les résultats de l'ensemble de test qui ont abouti à la matrice de confusion comme indiqué dans le tableau 11.
- Une autre observation du tableau est que la précision de toutes les méthodes est plus sur notre ensemble de données puisque la première a plus de champs pertinents pour évaluer le risque de diabète.
- Le modèle Machine à Vecteurs Support dispose du meilleur score en termes d'accuracy et precision. Celui de Naive Bayes tient ceux de recall et f1 score.
- Machine à Vecteurs Support dispose de bonne performance en classification binaire (où il y a seulement 2 classes d'instances comme dans notre cas).
- Le modèle Réseau de neurones semble ne pas être adapté à la classification binaire dans notre cas avec des résultats moins satisfaisant que les autres.
- On peut dire que le modèle Naive Bayes est celui qui dispose des meilleures performances en générale en tenant en compte le fait que ses scores pour les quatres métriques sont plus ou moins proches les uns des autres avec le meilleur recall possible, c'est-à-dire une meilleure capacité à retrouver toutes les instances positives (de classe malade qui est la classe d'intérêt).

- À partir du Tableau11 et la figure qui montre les performances graphiques de tous les algorithmes de classification sur la base d'instances classées. nous pouvons conclure que l'algorithme de classification Naive Bayes surpasse comparativement les autres algorithmes. L'algorithme Naïve Bayes est considéré comme la meilleure méthode d'apprentissage automatique supervisé de cette expérience car il donne une plus grande Accurancy par rapport aux autres algorithmes de classification avec une précision de 76,30 %.

Résultat 2:

Pour augmenter les performances et arrivée a la meilleure classification possible on propose notre base de données de diabète (IBN NAFIS). Le paramètre diabétique ne prend pas que des valeurs binaires, il prend plus que deux classes (multi classes) où les classes représentent les types de diabètes.

Les quatre techniques de classification (Réseau de neurones, Machine à Vecteurs Support, Naive bayes et k-plus proche voisin) ont ensuite été appliqués à l'ensemble d'apprentissage pour prédire les résultats de l'ensemble de test qui ont abouti à la matrice de confusion comme indiqué dans ce tableau 11.

	Accuracies	**Precisions**	**Recalls**	**f1_scores**
MLP	82.76	80.94	82.76	80.99
SVM	55.17	30.44	55.17	39.23
NB	48.28	59.84	48.28	49.2
KNN	27.59	7.61	27.59	11.93

Tableau 12: Matrice de confusion pour notre data set.

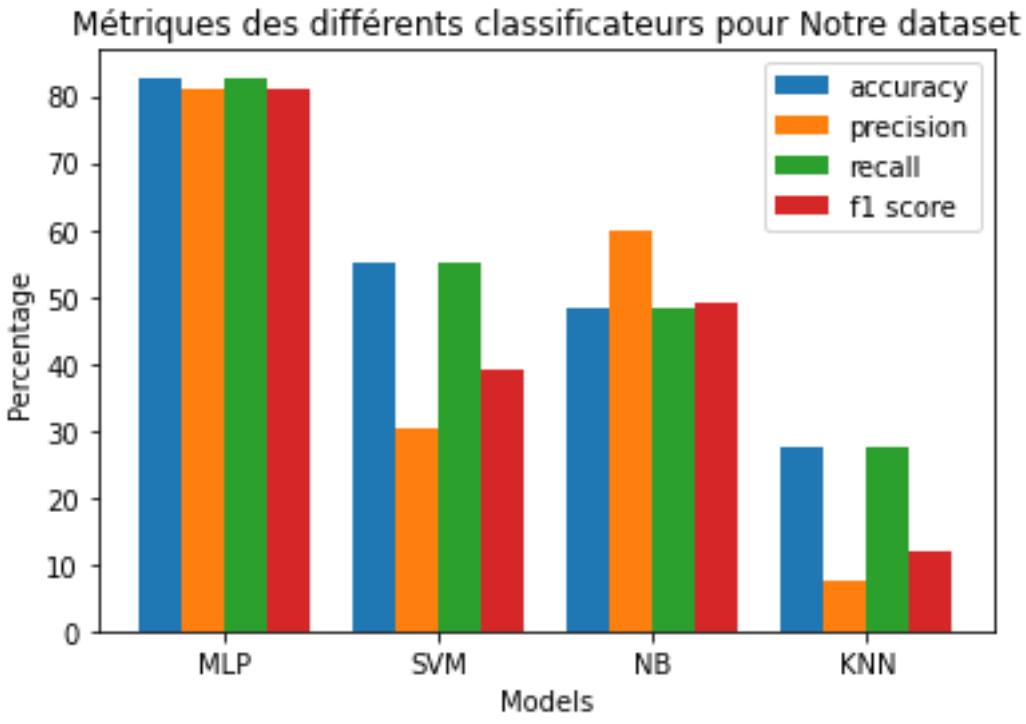

Figure 8 : Comparaison des résultats avec les différents métriques.

Discussion 2:

Dans le cas de ce dataset, les rôles sont inversés et le modèle mlp s'avère être plus efficace étant donné qu'on se retrouve face à un cas de classification multi classes (plus de 2).

5 Conclusion :

A travers ce chapitre, nous avons exposé les étapes de réalisation de notre système qui propose une étude comparative entre les quatre algorithmes de classification dans le but de fournir la meilleure classification possible comme diagnostic du patient.

Notre objectif qui est d'évaluer et d'étudier les algorithmes sélectionnés a bien été atteint.

La conclusion qu'on peut tirer de cette étude est qu'il est difficile d'affirmer avec certitude qu'un tel algorithme est meilleur par rapport à un autre, le choix dépend du problème où il va être appliqué et de plusieurs autres facteurs.

Conclusion

Diagnostiquer une maladie chez un patient sain ne produit pas les mêmes conséquences que de prédire la bonne santé chez un individu malade. Dans le premier cas, le patient sera soigné à tort, ou peut être demandera-t-on des analyses supplémentaires superflues; dans le second cas, il ne sera pas soigné, au risque de voir son état se détériorer de manière irrémédiable, et pour cela on a étudié la classification des données médicales afin d'éviter cette difficulté.

Ce travail nous a amené au développement d'une étude comparative entre les quatres outils de classification des données en utilisant les quartes algorithmes RN, NB, SVM et Knn appliqués aux deux bases de données sélectionnées (Pima, Ibn Nafis) du domaine médical.

L'objectif est de donner la possibilité aux professionnels d'étudier et d'appliquer les différentes techniques et outils de classifications. Donc nous avons tout d'abord étudié les approches et les notions fondamentales de la classification des données.

En premier lieu, nous avons présenté l'intelligence artificielle ainsi que les différentes techniques intervenant dans la classification. Puis, nous avons distingué les différentes approches adoptées pour l'apprentissage automatique et les méthodes d'apprentissage supervisé et non supervisé. Cette analyse nous a permis de constater l'importance de chaque méthode de classification pour le bon fonctionnement des bases de données dans le domaine médical.

Enfin, nous présentons les limites actuelles de notre étude comparative entre les résultats des trois outils de classification des données médicales et les difficultés rencontrées car on a conclue qu'il est difficile d'affirmer qu'un tel outil est meilleur par rapport à un autre.

Globalement, cette étude a permis d'exposer concrètement la problématique de classification de données dans le domaine médical, pouvant en cela contribuer a posteriori à trouver les solutions adéquates aux questions auxquelles est confronté ce domaine de recherche.

Références

[1] Dr. Margaret Chan. **'Rapport Mondial sur le Diabète'**. *Organisation mondiale de la Santé (OMS)*. Page 6, 2016.

[2] Hamza Cherif Ikram, **'Classification des tracés CardioTocoGraphiques (CTG) d'un fœtus à l'aide de classifieurs multiples'**. Thèse soutenue à l'Université Abou Bakr Belkaid de Tlemcen, 2011.

[3] Wikistat, **'Réseaux de neurones'**, 2016.

[4] Soumia Benikhlef, Bendimerad El Batoul, Nesma Settouti, **'Extraction des caractéristiques pour la classification de la maladie de Parkinson'**, 2013.

[5] Eve Mathieu-Dupas, **'Algorithme des k plus proches voisins pondères et application en diagnostic'**, Marseille, France. pp : 02-04, 2010.

[6] Barigou Fatiha, **'Contribution à la catégorisation de textes et à l'extraction d'information'**. Thèse soutenue à l'Université d'Oran, 2013.

[7] Reguieg Samia, Zeghaba Achouak, **'Etude comparative de quelques outils de classification dur des données médicales'**. Thèse soutenue à l'Université Abou Bakr Belkaid de Tlemcen, 2017.

[8] Ricco Rakotomalala, **'Autres méthodes supervisées extrapolées du schéma bayesien** Quelques approches pour rendre calculable P (Y/X)**'**. Page 8.

[9] Wassim Lahbibi, **'Algorithme KNN (k nearest neighbours, ou k plus proches voisins)'**. Ecole Supérieur de Commerce, 2013.

[10] Ubeyli ED, **'Automatic diagnosis of diabetes using adaptive neuro-fuzzy inference systems'**. Expert System. 27(4):259-266. Volume 27, N°4, pp: 259-266, 2010.

[11] Manaswini Pradhan, Dr. Ranjit Kumar Sahu, **'Predict the onset of diabetes disease using Artificial Neural Network (ANN)'**. *International Journal of Computer Science & Emerging Technologies* (E-ISSN: 2044-6004). Volume 2, N°2, pp: 303-311, 2011.

[12] B. A. Tama, F.Rodiyatul, H. Hermansyah, **'An Early Detection Method of Type-2 Diabetes Mellitus in Public Hospital'**, *TELEKOMNIKA* (*Telecommunication Computing Electronics and Control*), Volume 9, N°2, pp 287-294, 2011.

[13] Asha Gowda Karegowda, M.A. Jayaram, A.S. Manjunath, **'Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients'** IJEAT (*International Journal of Engineering and Advanced Technology*) Volume 1, N°3, pp 147-151, 2012.

[14] V. A. Kumari, R. Chitra, **'Classification of diabetes disease using support vector machine'**, *Internationnal Journal of Engineering Reaserch and Applications*, Volume 3, No2, pp 1797-1801, 2013.

[15] Akanksha Nilosey, **'FPGA Based Diabetic Patient Health Monitoring Using Fuzzy Neural Network'**. *International Journal of Science and Research (IJSR)* ISSN (Online): 2319-7064. Volume 5, N°10, 2016.

[16] Sara Belarouci, **'Traitement et classification intelligente des données médicales non équilibrées'**. Thèse de doctorat, 2016.

[17] M. Jahangir, M. Ahmed, h. Afzal, K. Khurshid, **'An expert system for diabetes prediction using auto tuned Multi-layer Perceptron'**, *IEEE Inteligent Systems Conference*, 2017.

[18] M. Pradhan, K. Kohale, P. Naikade, A. Pachore, & E. Palwe, **'Design of classifier for detection of diabetes using neural network and fuzzy k-nearest neighbor algorithm'**. *International Journal of Computational Engineering Research*, Volume2, N°5, pp: 1384-1387.

[19] Jonathan Bouchard, **'Optimisation des méthodes de classification'**. Projet de fin d'étude en génie de la production automatisée, Université du Québec, Ecole de technologie supérieur, 2008.

[20] Rayane Allouani, **'Conception et réalisation d'un système expert basé sur une ontologie pour le diagnostic du diabète'**. Mémoire de licence soutenue à l'Université de Badji Mokhtar, 2018.

Webographies

[web1] https://www.federationdesdiabetiques.org/information/definition-diabete/chiffres-monde#:~:text=5%20millions%20de%20personnes%20sont,l'apanage%20des%20pays%20d%C3%A9velopp%C3%A9s.

[web2] https://www.prnewswire.com/news-releases/federation-internationale-du-diabete-les-dernieres-conclusions-indiquent-que-463-millions-de-personnes-vivent-aujourd-hui-avec-le-diabete-a-travers-le-monde-et-une-hausse-des-chiffres-en-general-852954523.html

[web3] https://www.diabete.fr/comprendre/diabete/le-diabete-dans-le-monde

[web4] https://www.diabete.qc.ca/fr/comprendre-le-diabete/tout-sur-le-diabete/types-de-diabete/le-diabete-de-type-1/

[web5] https://www.diabetevaud.ch/comprendre-le-diabete/diabete-de-type-2/

[web6] https://www.ajd-diabete.fr/le-diabete/les-autres-types-de-diabete/le-diabete-neonatal/

[web7] https://www.passeportsante.net/fr/Maux/analyses-medicales/Fiche.aspx?doc=dosage-hemoglobine-glyquee

[web8] https://fr.wikipedia.org/wiki/R%C3%A9seau_de_neurones_artificiels

[web9] https://fr.wikipedia.org/wiki/M%C3%A9thode_des_k_plus_proches_voisins#Algorithme

[web10] https://web.maths.unsw.edu.au/~lafaye/CCM/merise/concintro.htm#:~:text=MERISE%20est%20une%20m%C3%A9thode%20de,plusieurs%20mod%C3%A8les%20conceptuels%20et%20physiques.

[web11] https://www.datanovia.com/en/fr/blog/interpretation-du-coefficient-kappa/

Résumé

Ce mémoire présente la conception et réalisation d'une étude comparative de plusieurs méthodes de classification et d'apprentissage dans le but de mettre en avant les différences entre ces dernières. Les méthodes sélectionnées sont : SVM, RN, NB, KNN ainsi que leurs possibles hybridations. Ces dernières sont appliquées de façon à présenter séparément les mesures de performances.

Les résultats ont montré que la performance de chaque méthode ou hybridation dépend des caractéristiques des datasets et modèles de classification utilisés.

Mots clefs: SVM, RN, NB, KNN, méthodes d'apprentissage.

Abstract

his thesis presents the design and realization of a comparative study of several classification and learning methods in order to highlight the differences between them. The methods selected are: SVM, RN, NB, KNN as well as their possible hybridizations. These are applied so that the performance measures are presented separately.

The results showed that the performance of each method or hybridization depends on the characteristics of the datasets and classification models used.

Keywords: SVM, RN, NB, KNN, learning methods.