كلية العلوم و التقنيات فاس
ⵜⴰⵖⵉⵡⴰⵏⵜ ⵏ ⵜⵎⴰⵙⵙⴰⵏⵉⵏ ⴷ ⵜⵙⵏⵉⵜⵉⵏ
Faculté des Sciences et Techniques de Fès

جامعة سيدي محمد بن عبد الله
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵙⵉⴷⵉ ⵎⵓⵃⵎⵎⴰⴷ ⴱⵏ ⵄⴰⴱⴷⵍⵍⴰⵀ
Université Sidi Mohamed Ben Abdellah

DEPARTEMENT DES MATHEMATIQUES

Master Mathématique et Application au Calcul Scientifique (MACS)

MEMOIRE DE FIN D'ETUDES

Pour l'obtention du Diplôme de Master Sciences et Techniques (MST)

Les systèmes d'aide à la décision médicale: Apprentissage automatique

Réalisé par: Aferhane Asma

Encadré par: Pr. Ammor Ouafae

Soutenu le 20 juin 2019 devant le jury composé de :

- **Pr. Ammor Ouafae** Faculté des Sciences et Techniques Fès.
- **Pr. Elhilali Alaoui Ahmed** Faculté des Sciences et Techniques Fès.
- **Pr. Ezzaki Fatima** Faculté des Sciences et Techniques Fès.
- **Pr. Rahmouni Hassani Aziza** Faculté des Sciences et Techniques Fès.

Année Universitaire 2018 / 2019

FACULTE DES SCIENCES ET TECHNIQUES FES – SAISS

B.P. 2202 – Route d'Imouzzer – FES

REMERCIEMENTS

Je remercie tout d'abord ALLAH le Tout Puissant de m'avoir donné la foi et de m'avoir permis d'en arriver là.

Mes remerciements vont également à mes parents particulièrement de tous les sacrifices qu'ils ont consentis pour me permettre de suivre mes études dans les meilleurs conditions possibles et n'avoir jamais cessé de nous encourager tout au long de mes années d'études.

Je voudrais dans un premier temps remercier, ma directrice de ce mémoire **Mme Ammor Ouafae**, pour sa patience, sa gentillesse, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter ma réflexion.

je désire aussi remercier aux membres du jury respectable : Pr. Ammor Ouafae, Pr. Elhilali Alaoui Ahmed, Pr. Rahmouni Hassani Aziza, Pr. Ezzaki Fatima pour avoir accepter d'évaluer ce travail et pour l'honneur qu'ils m'ont fait en portant leur attention sur ce mémoire.

Mon vifs remerciements s'adressent également à nos professeurs qui nous ont encouragés et nous ont aidés avec leurs remarques et observations durant la période de notre parcours universitaire.

Enfin, je voudrais exprimer ma reconnaissance envers les amis et les collègues qui m'ont apporté leur soutien moral et intellectuel tout au long de ma démarche.

Dédicace

Je dédie cet humble travail avec grand amour et sincérité à :
Mes chers parents sources de tendresse, d'amour et d'affection.
A ma soeur et mon frère, en témoignage de leurs amours, avec nos souhaits de bonheur, de santé et de réussite.
A tous les membres de ma famille. A tous mes amis.
A tous mes professeurs.
A tous ce qui ont participé à ce modeste travail.

TABLE DES MATIÈRES

1 Quelques notion sur les système d'aide a la décision médicale 5
1.1 Notion d'aide à la décision 6
1.2 Notion de la décision médicale 6
1.3 Les systèmes d'aide a la décision médicale 6
1.4 Nature et paramètre de la décision médicale 7
1.5 Typologie des SADM 8
1.6 Les systèmes d'expert et l'intelligence artificielle 8
1.7 L'apprentissage automatique et la classification 9
1.7.1 Maching learning et la classification 9
1.7.2 Les phases d'apprentissage automatique 10
1.7.3 Généralité sur Les méthodes de classification 11
1.8 Indicateurs de performance en classification 13
1.9 Exemple d'applications 15
1.10 la médecine et la prise de décision au futur 16

2 Méthode de classification pour la prise de décision 17
2.1 Arbre de décision 17
2.1.1 Construction d'arbre 18
2.1.2 Apprentissage 19
2.2 Réseau de neurone artificiel (RNA) 20
2.2.1 Structure de réseau 20
2.2.2 Apprentissage des poids 21
2.3 Régression logistique 22
2.3.1 L'estimation des coefficient 24
2.3.2 Tests dans le modèle 24
2.3.3 Interprétation des coefficients 25
2.3.4 Adéquation du modèle 25
2.4 Classification bayésienne (Naïve bayésienne) 26
2.4.1 Les étapes de la classification bayésienne 26
2.4.2 Apprentissage 27

2.5 K plus proche voisin (K-ppv) 28
2.5.1 Principe de K-PPV 28

3 Test et application **31**
3.1 Présentation de diabète 31
3.1.1 Définition 31
3.1.2 Cause de diabète 32
3.1.3 Classification de diabète 32
3.2 Présentation de la base de donnée utilisées 33
3.3 Simulation sous R 34
3.3.1 L'application de l'algorithme d'arbre de décision 35
3.3.2 L'application de l'algorithme de réseau de neurone 37
3.3.3 L'application de l'algorithme de régression logistique 38
3.3.4 L'application d'algorithme naïve bayésien 40
3.3.5 L'application de K plus proche voisin 41
3.4 Discussion 43
3.5 Conclusion 46

TABLE DES FIGURES

1.1 Illustration la tache de la classification. 10
1.2 Phase d'apprentissage. 10
1.3 Phase de test. 11
1.4 Tableau de contingence. 14
1.5 Courbe ROC. 15

2.1 Schéma d'un arbre de décision. 18
2.2 partition des classes 18
2.3 L'architecture d'un réseau de neurones multicouches. 20
2.4 Représentation matricielle du modèle d'un neurone artificiel 21
2.5 Représentation de la fonction logistique 23
2.6 Exemple de classification de K-NN. 29

3.1 Fonctionnement de l'insuline. 33
3.2 Échantillon de la base du diabète de type 2. 34
3.3 Arbre de décision pour la prédiction de les variables. 35
3.4 la taille d'arbre optimale. 36
3.5 Arbre de décision élagué. 36
3.6 Perception multicouche de la base de diabète. 37
3.7 Mise à jour des poids synaptique. 38
3.8 Coefficient estimée. 39
3.9 Les Odds ratio. 39
3.10 Graphe de corrélation entre les attributs de la base. 40
3.11 Matrice de corrélation. 40
3.12 Tableaux des probabilités conditionnée par l'attribut catégorique "diabetes" 41
3.13 Nombre optimale de k. 42
3.14 représentation de la courbe ROC de la régression logistique. 43
3.15 Comparaison graphique entre le l'algorithme régression logistique et l'arbre de décision. 44
3.16 représentation de la courbe ROC de trois méthodes. 45
3.17 représentation de la courbe ROC de quatre algorithmes. 45

3.18 représentation de la courbe ROC de nos algorithmes. 46

LISTE DES TABLEAUX

3.1 Description des attributs de la base. 34
3.2 Matrice de confusion de l'arbre de décision. 37
3.3 Matrice de confusion de réseau de neurone. 38
3.4 Matrice de confusion de la régression logistique. 39
3.5 Matrice de confusion de naive bayes. 41
3.6 Matrice de confusion de K-NN. 42
3.7 Comparaison entre les algorithmes en utilisant les taux d'évaluation. 43

ABREVIATIONS

IA : intelligence artificielle.

SADM : système d'aide à la décision médicale.

RNA : réseau de neurone artificiel.

k-ppv ou **K-nn** : K plus proche vosin.

ROC : Caractéristique de fonctionnement du récepteur.

INTRODUCTION

Chaque jour, nous entendons le terme intelligence artificielle, l'IA définie, plus généralement, comme la capacité d'une machine capable d'agir, par elle-même ou sous le contrôle de l'homme à reproduire des actions ou des fonctions qui sont habituellement celles des êtres humains. Aujourd'hui, nous la retrouvons dans nos machines informatiques, objets connectés, applications, réseaux sociaux, transports, et dans le secteur médical.

En 1958, le laboratoire d'intelligence artificielle du Massachusetts Institute of Technology (MIT) est créé, les scientifiques et les médecins sont captivés par le potentiel de l'IA en médecine, des ordinateurs intelligents capables de stocker et de traiter de vastes quantités de connaissances, aidant les cliniciens avec des activités de diagnostic. Une communauté d'informaticiens et de professionnels de la santé entreprennent alors de mettre au point des systèmes d'aide à la décision médicale (SADM) aident les médecins et autres professionnels de la santé pour obtenir des informations pertnentes au diagnostic et au thérapeutique des maladie, ils tentent ensuite de prévoir l'évolution de la maladie.

Ces systèmes décisionnelles définissent comme la construction de programmes d'IA qui effectuent un diagnostic et font des recommandations thérapeutiques, ils sont capables de faire ce que l'homme fait. Cependant, l'application de l'intelligence artificielle à la médecine offre une perspective essentielle à l'essor de ces nouvelles technologies, qu'il s'agisse de renforcer la relation entre patients et médecins, de poser des diagnostics plus rapides et plus précis ou encore d'améliorer la qualité des soins et aide les médecins de prendre la décision.

En d'autre termes, L'IA permet aux médecins de gagner du temps en laissant la machine analyser elle-même les données et fournir des estimations, dans le but est de réussir à prédire de nombreuses maladies, afin que les médecins puissent intervenir le plus tôt possible.

Le diagnostic médical est un processus pour la prise de décision, avec l'utilisation de l'informatique pour la leur réalisation devient de plus en plus fréquente. Les systèmes d'aide à la décision médicale fournissent une aide substantielle, car elles réduisent les erreurs dues à la fatigue et le temps nécessaire pour le diagnostic. Actuellement, la plupart des hôpitaux modernes sont bien équipés avec des dispositifs de collecte de données. Cependant, L'augmentation du volume de données entraîne des difficultés à extraire des informations utiles pour l'aide à la décision. Les méthodes traditionnelles d'analyse de

données peuvent être complètes par l'utilisation de méthodes dites intelligentes.

L'apprentissage est la capacité de s'améliorer avec l'expérience, de se rappeler les décisions antérieures et des résultats similaires afin de faire de meilleurs choix à l'avenir dans des situations similaires. L'apprentissage automatique (Machine learning en anglais) cherche à trouver le moyen de construire des programmes informatiques qui s'améliorent automatiquement avec l'expérience. On distingue deux type d'apprentissage automatique les plus largement adoptées sont l'apprentissage supervisé qui forme des algorithmes basés sur des données d'entrée et de sortie étiquetées par l'homme et l'apprentissage non supervisé qui ne fournit pas à l'algorithme des données étiquetées pour lui permettre de trouver une structure et de découvrir une logique dans données entrées.

On entend que Machine Learning une somme d'algorithmes capables d'améliorer les performances de la machine au fur et à mesure qu'elle obtient des datas. C'est un principe d'entrées-sorties : on entre des informations en entrée, et on attend des informations en sortie. Si elles sont correctes, on le dit à la machine (on la supervise) et si elles sont fausses on le lui dit aussi. Et au fur et à mesure, elle apprend à avoir de plus en plus de conclusions correctes en sortie.

A travers ce mémoire de Master, on s'intéresse par l'approche numérique des SADM en utilisant l'apprentissage automatique dans leur processus de prise de décision. Plusieurs travaux ont vu le jour et c'est dans ce contexte que nous proposons dans notre travail d'approche numérique des système d'aide à la décision médicale qui sont utilisé des données est des connaissances des antécédents pour guider des nouveaux cas via les algorithme d'apprentissage automatique supervisé Ainsi que les approches algorithmiques courantes de l'apprentissage automatique, y compris l'algorithme de régression logistique, l'apprentissage par arbre décisionnel et la classification bayésienne, et d'autre algorithmes supervisés.

Ensuite, on effectue aussi un apprentissage automatique supervisé sur la base de données "Pima Indiens Diabètes" pour la prédiction du diabète de type 2 qui est un dysfonctionnement du système de régulation de la glycémie. Beaucoup de travaux ont été menés afin d'effectuer la classification ou le diagnostic du diabète. Hung-Chun Lin et al ont obtenu une précision de 75% en utilisant un ensemble de cellules mémoires qui ont subit un apprentissage leur permettant d'effectuer une classification et cela grâce à la méthode K-plus proche voisin, dans la classification du diabète. Une autre méthode a utilisé un réseau de neurones pour effectuer la classification du diabète. L'apprentissage du réseau de neurones a été effectué pour chaque patient sur une période de 24h et le test a été effectué sur une autre période de 24h. Finalement, on fait une comparaison entre les 4 algorithmes à travers la courbes ROC de chaque méthode ou encore par le choix de taux de précision le plus élevé en utilisant le logiciel R. [1] La meilleur algorithme peut considérer comme résultat une classification des patients qui peuvent être susceptibles d'être atteint du diabète et qui a comme but l'aide à la décision du médecin en termes de temps d'examination et de taux de précision du résultat.

Le manuscrit est organisé en trois principaux chapitres en plus de l'introduction générale.

-Dans le premier chapitre nous avons présentés un aperçu générales sur les système d'aide a la décision médicale et quelque notion sur essentiel l'apprentissage automatique.

1. R est un langage de programmation open-source utilisé principalement pour le calcul statistique dont Les packages les plus populaires pour l'apprentissage automatique incluant R sont CARET (abréviation de Classification And REgression Training) pour créer des modèles prédictifs et e1071 qui comprend des fonctions pour les statistiques et la théorie des probabilités.

-Le deuxième chapitre résume les méthodes ou les algorithmes d'apprentissage automatique.

-Le troisième chapitre présente un aperçu général sur la maladie du diabète en citant les causes et leurs classes, il représente encore les caractéristique de la base de données utilisée pour évaluer et comparer les approches d'apprentissage. Ensuite, les résultats sont présentés, comparés et interprétés a l'aide des indicateurs de performance.

CHAPITRE

1

QUELQUES NOTION SUR LES SYSTÈME D'AIDE A LA DÉCISION MÉDICALE

Introduction

Au cours des années 70, avec l'apparition de l'intelligence artificielle dont l'objectif était la recherche de moyens susceptibles de doter les systèmes informatiques de capacités intellectuelles comparables à celles des humains. La complexité croissante des connaissances médicales telles que les données physiopathologiques ou les moyens diagnostiques et thérapeutiques, obligent le médecin à gérer toujours plus d'informations pour soigner un patient. Une réponse à cet état de fait peut être une plus grande spécialisation des médecins mais ils ne peuvent être réunis simultanément et cette spécialisation peut entraîner des pertes de temps.

Lorsqu'il soigne un patient, le médecin doit prendre toute une série de décisions. Il agit en suivant un raisonnement qui implique tout à la fois des notions de connaissance, d'incertitude, d'expérience et de risque.Le développement de systèmes d'aide à la décision nécessite de modéliser cette pratique. Pour cela, il importe de retracer la démarche du médecin face à un malade et de faire une analyse de la décision médicale. Les développeurs de ces systèmes ont souvent des ambitions de recherche en matière d'AI et de processus de raisonnement et appliquent de nouvelles méthodes à des connaissances médicales plutôt que d'adopter la démarche inverse qui consiste à résoudre un problème médical donné en appliquant une méthode adaptée.Dans ce contexte nous allons dans ce chapitre présenter quelques notion sur SADM et les systèmes d'expert, puis on va voir les indicateurs des performances en apprentissage automatique.

1.1 Notion d'aide à la décision

Ce qu'on l'entend par le mot "décision" en différentes discipline concerne des phénomènes très distincts. Pour la plupart des chercheurs, la décision est définie comme un choix entre plusieurs alternatives,où comme le dit Castles[1] "*Une décision est un choix conscient entre au moins deux actions possibles*". Pour d'autres, la décision concerne le processus de sélection de buts et d'alternatives.

En d'autre terme, elle peut être définit comme étant un acte par lequel un ou des décideurs opèrent un choix entre plusieurs options permettant d'apporter une solution satisfaisante à un problème donné. Cette notion de décision a évolué dans le temps au fur et à mesure que ce sont transformés et complexifiés les procédures de prise de décision.

La prise de décisions peut apparaitre dans n'importe quel contexte de la vie quotidienne, aussi bien au niveau professionnel que sentimental ou familier...etc. Toute prise de décision doit inclure une grande connaissance du problème. En l'analysant et en le comprenant, il est alors possible de trouver une solution. Bien entendu, face à des questions simples par exemple,Lorsqu'un jeune doit choisir quelles études poursuivre après le lycée, il doit prendre une décision raisonnée, étant donné que ce choix donnera lieu à d'importantes conséquences, la prise de décisions a lieu pratiquement toute seule avec un raisonnement un peu complexe.

Cela constitue une aide qui permettra au décideur d'élaborer une représentation pertinente de la situation. Tout en sachant que tout programme sur les connaissances et les données a pour but d'aider à la décision.

1.2 Notion de la décision médicale

Lorsqu'on s'intéresse aux travaux sur la décision en santé, on rencontre d'abord la décision médicale, cette dernière est considérée comme étant le centre de l'acte médical[2]. Le médecin est confronté à la décision au choix entre plusieurs possibilités avec Le processus de la décision médicale consiste à poser un diagnostic,proposer un traitement ou le différer ...etc. Ainsi, de très nombreuses applications d'aide à la décision ont été développées dans ce domaine. Ces applications sont destinées à soutenir le personnel de santé dans leurs prises de décisions. Cela implique l'utilisation de divers outils d'aide à la décision.

Dans le cadre de ce domaine, l'aide à la décision consiste en l'extraction de connaissances dans de grands volumes de données pour en extraire des informations pertinentes. Cependant, les mécanismes de la décision en médecine générale sont complexes dont La décision est le fruit de différentes composantes : composantes biomédicales (objectives), et autres composantes, plus subjectives (liées à la relation, la personnalité du médecin, celle du patient, l'environnement...). Ceci peut expliquer que des situations qui semblent similaires au point de vue biomédical puissent donner lieu à des décisions différentes.

1.3 Les systèmes d'aide a la décision médicale

Un systèmes d'aide à la décision médicale SADM est un outil qui permet d'améliorer la qualité des soins, plusieurs définitions concernant les SADM ont été proposées :

1. Professeur en sciences cognitives et responsable du programme de lecture du Centre d'excellence ARC sur la cognition et ses troubles
2. Définition du Grand Usuel Larousse, Dictionnaire encyclopédique, 1996

▷ Un système d'aide à la décision médicale est un ensemble organisé d'informations, conçu pour assister le praticien dans son raisonnement en vue d'identifier un diagnostic et de choisir la thérapeutique adéquate, en opérant un dialogue entre l'homme et la machine.

▷ Les systèmes d'aide à la décision médicale (SADM) sont des applications informatiques dont le but est de fournir aux cliniciens en temps et lieux utiles les informations décrivant la situation clinique d'un patient ainsi que les connaissances appropriées à cette situation, correctement filtrées et présentées afin d'améliorer la qualité des soins et la santé des patients.

Ces systèmes décisionnels devraient pouvoir assister les médecins en leur proposant une synthèse dans leur pratique quotidienne, pourvu que celle-ci réponde réellement à leurs attentes. Mais, ces systèmes restent peu utilisés en routine pour plusieurs raisons. D'une part, les développeurs de ces systèmes ont souvent des ambitions de recherche en matière d'intelligence artificielle et de processus de raisonnement et appliquent de nouvelles méthodes à des connaissances médicales plutôt que d'adopter la démarche inverse qui consiste à résoudre un problème médical donné en appliquant une méthode adaptée. D'autre part,les médecins rencontrent la difficulté d'exprimer leurs besoins de façon pertinente, c'est à dire, il existe des difficultés de communication entre les développeurs (langage spécialisé informatique) et les cliniciens (langage spécialisé médical).

1.4 Nature et paramètre de la décision médicale

Les décisions diagnostiques, pronostiques et thérapeutiques reposent sur trois groupes d'arguments : la sémiologie qui correspond à l'étude des signes et des symptômes, la nosologie qui définit les maladies et leurs critères d'identification, et enfin les caractères de l'environnement reposant sur les données épidémiologiques, psychologiques ou socio-professionnelles.

Les signes cliniques et les résultats des examens para-cliniques sont des éléments d'information permettant au médecin d'estimer l'état pathologique du patient et peuvent être rassemblés sous le terme général de tests qui sont l'expression de variables. Il existe trois types de variables :

▷ les variables qualitatives non ordonnées : sont des variables catégorielles tels que La première information est dichotomique car elle ne peut prendre que deux valeurs tandis que la seconde peut prendre au mois trois valeurs.

▷ les variables qualitatives ordonnées : ces variables sont ordinales. Il existe alors une hiérarchie dans les catégories comme pour l'évaluation la sévérité dune dyspnée qui peut être cotée de un à quatre.

▷les variables quantitatives dites mesurables :ces variables s'expriment par un nombre. Elles sont continues lorsqu'elles peuvent prendre toutes les valeurs comprises dans l'intervalle des variations telles la taille. Par contre, Elles sont discontinues ou discrètes lorsqu'elles ne peuvent prendre que certaines valeurs comme la fréquence cardiaque par minute dont La gestion des informations est souvent facilitée par la transformation des

réponses quantitatives en réponses catégorielles où chaque classe correspond à un intervalle de mesure.

1.5 Typologie des SADM

Les objectifs d'un système dépendent des utilisateurs, de leurs fonctions et de leurs connaissances. Il existe trois grandes catégories de systèmes d'aide :

▷ **Les systèmes d'assistance documentaire** représentent les systèmes d'aide indirecte à la prise de décision. L'objectif de ces systèmes d'assistance documentaire est de faciliter l'accès aux informations pertinentes en un temps record mais ces systèmes n'ont pas de méthode de raisonnement à proprement parler,Le système d'aide au diagnostic médical est actuellement toujours utilisé dans sa fonction documentaire.

▷ **Les systèmes d'alerte** assistent les utilisateurs en déclenchant automatiquement des alarmes signifiant des erreurs à ne pas commettre des informations concernant une prise de décision pour une situation donnée. Ils sont plus actifs et plus appliqués dans la décision médicale. L'assistance fournie n'est pas une aide au raisonnement globale du cas du patient, mais plutôt un aide-mémoire fournissant une information utile dans une situation facile à définir a priori .Ces systèmes, comme les précédents, ne raisonnent pas véritablement.

▷ **Les systèmes consultants** raisonnent sur des situations médicales définies et fournissent à l'utilisateur des conclusions argumentées selon les méthodes de raisonnement employées. les développeurs s'intéressent principalement à ce type de système où l'on note le plus de réalisations en matière de système d'aide à la décision.

1.6 Les systèmes d'expert et l'intelligence artificielle

Depuis l'invention des machines, leur capacité a effectuer les différentes taches n'a cessé à s'accroitre de façon exponentielle, alors on a besoin de l'intelligence artificielles **(AI)** pour poursuivre la création des machines. AI est un domaine de recherche qui promis à un grand avenir, selon John McCarthy [3]l'intelligence artificielle est *" la science et l'ingénierie de la fabrication de machines intelligentes, en particulier des programmes informatiques intelligents"*. Ses applications concernent toutes les activités humaines, permettent d'améliorer la qualité des soins et aide les médecins de prendre la décision. En effet, les AI sont au cœur de la médecine du futur. l'un de sous domaine de l'AI est l'apprentissage automatique.
D'un autre coté, On ne peut pas voir les systèmes d'aide à la décision médicale sans parler des systèmes experts. Les systèmes d'expert sont l'une des plus importantes applications de l'intelligence artificielle,ils sont des systèmes d'information qui traitent les données pour produire des résultats qui facilitent la prise de décision, Un système expert est donc un logiciel ou une partie d'un logiciel qu'assister l'homme dans différentes domaines. Cependant, un système expert se compose de trois partie principale :

3. né en 1927 à Boston, l'un des principaux pionniers de l'intelligence artificielle est le principal pionnier de l'intelligence artificielle

▷ **base de connaissances (Knowledge Base)** : La base des connaissances comporte les données, les connaissances et les relations. Ainsi, les principes et les règles de décision utilisés par l'expert pour résoudre un type particulier de problème et Il convient de noter que le succès du système expert dépend de l'ampleur de sa base de connaissances et de sa précision.

▷ **un moteur d'inférences (Inference Engine)** : Le moteur d'inférences est un programme qui utilise les données et les règles de la base de connaissance pour conclure les résultats et suggérer les recommandations nécessaires.

▷ **interfaces de l'utilisateur (User Interface)** : L'interface un moyen de l'aide à l'acquisition des connaissances fournies par l'expert.

Les systèmes d'expert les plus célèbres, **Mycin** pour l'identification d'infections bactériennes ou **Sphinx** pour la détection d'ictères , s'appuient sur l'ensemble des connaissances médicales dans un domaine donné et une formalisation des raisonnements des spécialistes qui relies ces connaissances entre elles pour aboutir à un diagnostic.

1.7 L'apprentissage automatique et la classification

En général, l'apprentissage est né grâce aux technologies dont laquelle les ordinateurs peuvent apprendre sans être programmées pour effectuer des taches spécifiques. Les chercheurs intéresses par l'intelligence artificielle souhaitaient vérifier si les ordinateurs pouvaient apprendre à partir de données. Ainsi, les IA auront la possibilité, grâce au *Machine Learning*, d'expérimenter et de tirer des conclusions de leurs expérimentations. Particulièrement, ils peuvent être utilisé pour prendre une décision au diagnostic médical.

1.7.1 Maching learning et la classification

L'apprentissage automatique aussi appelée le maching learning en anglais, est un des champs de l'Intéligence Artificielle qui consiste en l'automatisation de l'apprentissage d'un algorithme, notamment par l'analyse, la sélection et le traitement de données. Ainsi, nous intéressons par le premier stade du Machine Learning, qui est la classification de données. Il consiste à regrouper les données en classes et il est constitué de deux principaux types :

► L'apprentissage supervisé (Supervised learning) : le programme a une base de données avec des classes connues. Il aura pour objectif de ranger des valeurs qui lui sont attribuées dans les classes connues.

► L'apprentissage non supervisé (Unsupervised learning) :le programme ne reçoit que des valeurs et doit créer les classes dans lesquelles les attribuer. Il va donc "décider" lui même le nombre de classes à créer pour ensuite ranger les données dans chaque classe.

Dans notre mémoire, nous intéressons par la classification supervisé binaire dont l'objectif est de définir des règles permettant de classer des objets dans des classes à partir de variables qualitatives ou quantitatives caractérisant ces objets. généralement, La classification vise à prédire une classe à partir d'une base d'observations , pour cela on dispose d'une liste d'attributs ou des variable (qualitatif ou quantitatif) avec la classe connue.

En médecine, on utilise la classification au diagnostic des maladie. On représente un exemple de classification binaire dont le nombre de classes est exactement deux. On choisi 6 individus (patients), alors le problème est de déterminer s'il sont la grippe à partir de leurs symptômes(les attributs). Puis il est possible de déterminer si un nouveau individu (patient)a la grippe (appartient à la classe 1 ou 0). De plus, les algorithmes de

classification génèrent un modèle qui effectue un ensemble de tests sur les symptômes (les attributs) pour déterminer la classe.

	ATTRIBUTS				CLASSE
Nom	Fièvre	Douleur Musculaire	Toux	Céphalée	Grippe?
Patient1	O	O	O		O
Patient2	O		O		
Patient3				O	
Patient4	O	O	O	O	O
Patient5	O	O		O	O
Patient6		O			
PatientX	O	O			?

OBSERVATIONS (connues)

FIGURE 1.1 – Illustration la tache de la classification.

Les règles de classification ont un cas particulier des règles d'associations, il se compose d'une condition et d'une prédiction sous forme : **Si** condition **Alors** prédiction dont la prédiction est fixe. Dans la suite, on va voir comment évaluer si cette ces règles donne des résultats efficace ou non.

1.7.2 Les phases d'apprentissage automatique

Un superviseur est disponible pour étiqueté correctement les données d'apprentissage, soit A correspond à un ensemble de n paires d'entrées et d'étiquette de sortie et la fonction f vas être définie d'une façon explicite par l'algorithme d'apprentissage dont ce type d'apprentissage est constitué de deux phase :

▶ **Phase d'apprentissage ou d'entrainement** : dans cette phase l'algorithme d'apprentissage reçoit en entrer des exemple d'apprentissage étiqueté A et produit un modèle de prédiction le plus performant possible.

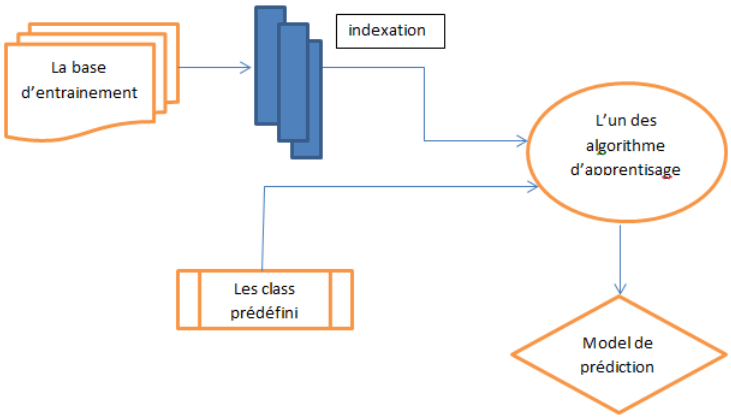

FIGURE 1.2 – Phase d'apprentissage.

► **La phase de test** :c'est la phase de prédiction[4] des nouveaux exemple en fonction du modèle obtenue dans la phase d'apprentissage qui vas associer variable d'entrée a une classe.

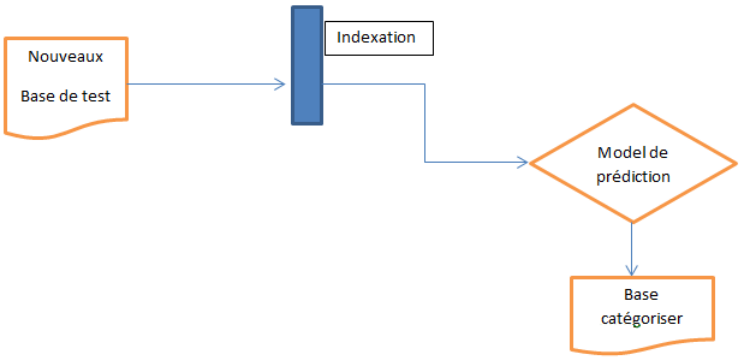

FIGURE 1.3 – Phase de test.

Ces deux phases peuvent être exécutées de manières consécutive ou itérative.

1.7.3 Généralité sur Les méthodes de classification

L'arbre de décision :

Pour certains domaines, on a besoin de produire des procédures de classification compréhensibles par l'utilisateur. Particulièrement, le cas pour l'aide au diagnostic médical où le médecin doit pouvoir interpréter les raisons du diagnostic. Les arbres de décision répondent à cette contrainte car ils représentent graphiquement un ensemble de règles et les algorithmes d'apprentissage par arbres de décision sont efficaces, disponibles dans la plupart des environnements de fouille de données.

Pour la construction,l'arbre de décision est basé sur deux plus importantes Les méthodes sont :

► La méthode CART développée par Breiman et al. en 1984. Elle construit des arbres de décision binaires.elle est fondée sur Le critère de Gini qui est utilisé pour associer un test à un noeud.

► La méthode ID3 développée par Quinlan en 1983 et leur amélioration en 1993 par une nouvelle version C4.5. On ne se restreint pas à des attributs binaires telle que le choix du test associé à un noeud se fait à l'aide de la fonction Gain basées sur la fonction entropie[5] (définit dans le chapitre suivant).

Les arbres de décision fournissent des méthodes effectives qui permet d'obtenir de bons résultats dans la pratique. Ils possèdent l'avantage d'être compréhensibles par tout utilisateur si la taille de l'arbre produit est raisonnable.

4. cette tache vise à prédire une ou plusieurs caractéristiques inconnues à partir d'un ou plusieurs caractéristique connues

5. L'entropie de Shanon due à Claude Shanon en 1948, est une fonction mathématique qui correspond à la quantité d'information contenue ou délivrée par une source d'information

Réseaux de neurone :

les principes des modèles connexionnistes ont été décrits récemment par Grumbach dont la premiére modélisation d'un neurone formels en 1940 par Mac Culloch et Pitts s'inspirant de leurs travaux sur les neurones biologiques. Elle est fondée sur le comportement de réseaux de neurones formel interconnectés dont chaque neurone formel pouvant être considéré comme une simplification extrême d'un neurone. Ensuite depuis le début de 1980, le développement des technologies d'investigations offrent de nouveaux horizons pour la compréhension des phénomènes neuronaux, avec, plusieurs études portant sur les possibilités d'extension des réseaux de neurones monocouche ont été réalisées et de nouveaux modèles sont apparus.

Un neurone formel est définit par cinq éléments suivants :

- La nature de ses entrées.
- La fonction d'entrée totale qui définit le pré-traitement effectué sur les entrées.
- La fonction d'activation du neurone qui définit son état interne en fonction de son entrée totale.
- La fonction de sortie du neurone.
- La nature de la sortie du neurone.

Les réseaux de neurones jouent un rôle important dans le domaine de l'aide à la décision médicale par les possibilités d'apprentissage qu'ils offrent. ils peuvent d'aider au diagnostic des cardiologie et des épilepsies.

régression logistique :

La régression logistique est largement utilisée en apprentissage automatique. Ainsi, elle est plus répandue dans de nombreux domaines. En particulier en médecine, elle permet par exemple d'associer un ensemble de symptômes avec une maladie. Elle est similaire à la régression linéaire mais elle convient aux modèles dans lesquelles les variables de sortie sont dichotomiques. Cette régression est basée sur une fonction qui vient de Joseph Berkson en 1944 et 1951 nommée logit dont l'estimation des coefficient de cette fonction se fait a l'aide de le principe du maximum de vraisemblance, si les coefficient ne sont pas bien estimer a l'aide "odd rations", il faut le corrigera et le réexaminera le bon ajustement des valeurs estimées, jusqu'à ce qu'aucune correction ne puisse atteindre un meilleur résultat (Howell, 1998).

Classification bayésienne(naïve bayésienne)

Parmi les différentes techniques probabilistes utilisées dans le domaine de diagnostic médicale, nous nous concentrons sur l'approche Naïve [6] Bayes. Cette dernière se comporte extrêmement bien et souvent mieux que d'autres modèles. On peut également l'utilise la méthode de Naive Bayes avec des entités continues, mais il est plus adapté aux variables binaire. Si toutes les entités en entrée sont catégoriques, Naive Bayes est recommandé. Cependant, dans le cas de caractéristiques numériques, cela suppose une autre hypothèse forte, à savoir que la variable numérique est normalement distribuée. Le choix d'un diagnostic peut se faire en évaluant les probabilités de la présence de la maladie en fonction des résultats du test et pour la prise de décision médicale.
la classification naïve bayésienne fut utilisée pour la prise de décision. Parmi ces origines les plus significatifs, on citera son apprentissage rapide qui ne nécessite pas un gros volume

6. la distribution normale de $X|Y$ qui suppose une corrélation nulle entre les composantes de X.

de données et son extrême rapidité d'exécution comparé à d'autres méthodes plus complexes. Finalement, malgré la forte hypothèse simplificatrice d'indépendance des variables, la classification naïve bayésienne obtient des résultats remarquables dans de nombreuses applications de la vie courante ce qui en fait un algorithme de choix parmi les outils du machine learning. Cette méthode est associé avec la table de probabilités conditionnelles de chaque variable sont appelées valeurs prédictives. Cette dernière est la probabilité de la présence de cette maladie si le test donne une réponse positive et la valeur prédictive négative est la probabilité de l'absence de la maladie si le test donne une réponse négative. Leur estimation s'effectue à partir du théorème de Bayes(théorème élémentaire de la théorie des probabilités).

$$P(C/X_1, ..., X_n) = \frac{p(C)P(X_1, ..., X_n/C)}{P(X_1, ..., X_n)}$$

où C est une variable de classe, conditionnée par plusieurs variables caractéristiques $\{X_1, ..., X_n\}$. Lorsque le nombre des variable est grand, ou lorsque ces variables peuvent prendre un grand nombre de valeurs, baser ce modèle sur des tableaux de probabilités devient impossible.

Parmi les premiers systèmes probabilistes, on propose celui de de Dombal dans la fin des années 60, il aidait au diagnostic des syndromes douloureux abdominaux avec huit maladies et 50 signes. Des évaluations ont montré qu'il était performant à Leeds.

k plus proche voisin :

L'un des algorithme de classification la plus simple et plus classique est K-Nearest Neighbors. On dispose d'un ensemble des points réparti en deux classes. Un nouveau point semblable aux précédents se présente, sa classe est inconnue. L'objectif est de lui attribuer une classe en utilisant le fait qu'on connaît la classe que les autres points appartient.

A partir de ce ensemble de points pour lesquels la classe d'appartenance est connue,le problème est de quel façon classer un nouveau point pour lequel cette classe est inconnue. Cette méthode consiste à attribuer à ce nouveau point la même classe que le plus proche des points appartenant au nuage initial. Elle est facile à implémenter mais peu utilisée.

1.8 Indicateurs de performance en classification

Matrice de confusion

Évaluation de la qualité d'une méthode sur un ensemble d'observations il s'agit de comparer les résultats donnés par la méthode à ce que l'on peut observer dans la population. Pour mesurer la qualité d'une règle ou bien méthode, on a besoin de la matrice de confusion sous forme de tableau de contingence.

Pour un test à réponse binaire visant à décider de la présence ou de l'absence d'une maladie M à partir d'une base de connaissance, quatre situations sont possibles :

►**Vrais positifs (VP)** : le nombre d'observation qui donne une réponse positive et M est effectivement présente, ils induisent une décision correcte.

►**Faux positifs (FP)** : le nombre d'observation qui donne une réponse positive et M est absente, ils induisent une décision erronée par excès.

►**Vrais négatifs (VN)** : le nombre d'observation qui donne une réponse négative et M est effectivement absente, ils induisent une décision correcte.

►**Faux négatifs (FN)** : le nombre d'observation qui donne une réponse négative et M est présente, ils induisent une décision erronée par défaut.

	M présente	M absente
réponse du test positive	Vrai positif décision correcte (sensibilité)	Faux positif décision erronée (erreur α)
réponse du test négative	Faux négatif décision erronée (erreur β)	Vrai négatif décision correcte (spécificité)

FIGURE 1.4 – Tableau de contingence.

-*La sensibilité* permet de représenter la fréquence des réponses positives au test. En d'autres termes, c'est la probabilité conditionnelle d'avoir un résultat du test positif si M est présente.Par conséquence, la sensibilité est égale à 1 s'il n'existe pas de faux négatif.

$$Sensibilité = \frac{VP}{VP + FN}$$

-La spécificité représente la fréquence des réponses négatives au test parmi les patients qui ne sont pas atteints de la maladie M. En outre, c'est la probabilité conditionnelle d'avoir un résultat du test négatif avec l'absence de M. Elle correspond à 1 s'il n'existe pas de faux positif.

$$Spécificité = \frac{VN}{VN + FP}$$

-Pour prédire les performance d'un classeur sur les données non observées,on utilise la méthode d'évaluation nommée *Accuracy* en anglais, elle est définit comme le ratio entre les exemples bien classés et le nombre total d'exemples :

$$Accuracy = \frac{VP + VN}{VP + FP + VN + FN}$$

- *Précision* est la proportion d'élèments bien classés pour une classe donnée :

$$Précision = \frac{VP}{VP + FP}.$$

-*L'erreur de classification* est l'inverse de l'accuracy. Elle est définie comme étant le ratio entre le nombre d'exemples mal classés et le nombre total d'exemples.

$$Erreur = \frac{FP + FN}{VP + FP + VN + FN}$$

la sensibilité et la spécificité sont en concurrence : lorsque l'on essaye d'augmenter la sensibilité l'autre est diminue.

Courbe ROC

Une courbe **ROC** (receiver operating characteristic) est un graphique représentant les performances d'un modèle de classification pour tous les seuils de classification. Cette courbe trace le taux de vrais positifs(sensibilité) en fonction du taux de faux positifs (spécificité). En d'autre part, ce courbe trace les valeurs TVP et TFP pour différents seuils de classification. Diminuer la valeur du seuil de classification permet de classer plus d'éléments comme positifs, ce qui augmente le nombre de faux positifs et de vrais positifs.

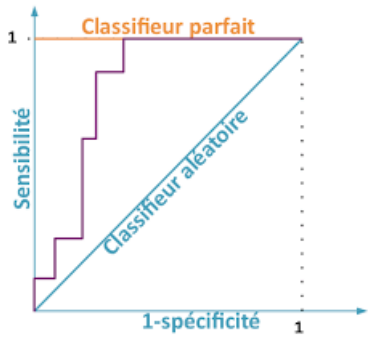

FIGURE 1.5 – Courbe ROC.

Pour calculer les points d'une courbe ROC, nous pourrions effectuer plusieurs évaluations d'un modèle de régression logistique en variant les seuils de classification, mais ce serait inefficace. Nous pouvons en revanche calculer efficacement l'aire sous cette courbe ou AUC(aire sous la courbe ROC)tel que la valeur mesure l'intégralité de l'aire à deux dimensions située sous l'ensemble de la courbe ROC de (0,0) à (1,1).

L'AUC fournit une mesure agrégée des performances pour tous les seuils de classification possibles. On peut interpréter l'AUC comme une mesure de la probabilité pour que le modèle classe un exemple positif aléatoire au-dessus d'un exemple négatif aléatoire.

AUC <0.5 : Aucune classification.
AUC $\in [0.5, 0.7[$: la classification est faible.
AUC $\in [0.7, 0.8[$: Classification acceptable.
AUC $\in [0.8, 0.9[$: Classification excellente.
AUC $\in [0.9, 1]$: Classification parfaite.

1.9 Exemple d'applications

Les revues récentes de la littérature montrent que les SADM utilisés en pratique de la médecine, ils peuvent être aider dans un ensemble des activités médicales, on peut citer Les maladies chroniques, les affections aigües et les urgences. Ainsi, on les utilise dans La plupart des spécialités médicales comme cardiologie, cancer, les affections psychiatriques, pédiatrie ou la gériatrie. En générale, les système d'aide a la décision médicale sont utilisée Les différentes catégories de médecins : la médecine générale ou bien spécialisée.

Les diagnostics sont une composante essentielle des soins de santé. Pour prendre en charge un patient efficacement, il est nécessairement d'identifier sa maladie avec précision. Ainsi le *Machine Learning* a été développer de nouvelles méthodes d'identification plus efficaces.

En 2016, IBM Watson Health et Quest Diagnostics s'associaient pour former Watson Genomics. Ce partenariat a permis de chercher de la façon intégrer les SADM et le séquençage génomique des tumeurs. De plus, Google DeepMind s'est associé avec plusieurs organisations médicales britanniques afin de développer des solutions de diagnostics basées sur le Machine Learning.

Un autre exemple, On peut évoquer la plateforme predict P1vital Predicting Response to Depression Treatment développée par l'université d'Oxford, qui permet de diagnostiquer et de soigner les maladies du cerveau grâce à l'analyse prédictive.

En termes de diagnostic, On peut citer quelques application :

▷ Your.MD, chatbot de prédiagnostic : cette application, développée en collaboration avec le NHS [7] au Royaume-Uni, utilise l'IA pour fournir des conseils médicaux sous forme de chatbot. [8]

▷ Cardiologs : est une solution web de lecture des électrocardiogrammes (ECG), en récupérant un signal d'ECG, il est possible, en le poussant sur la plateforme Cardiologs, d'obtenir un rapport personnalisé en temps réel.

1.10 la médecine et la prise de décision au futur

L'intelligence artificielle appliquée à la santé fournit des solutions d'aide à la décision pour les médecins oncologues, et contribue aussi à donner davantage d'autonomie aux patients atteints de diabète. D'une autre coté, les techniques d'analyse des données peuvent permettre de progresser dans la détection des pathologies, et éviter de soumettre les malades à des examens intrusifs, comme l'explique Thibault Pironneau. [9] Une récente étude montre qu'une IA pourrait dans le futur de pronostiquer la réponse d'un patient à son traitement comme l'institue Curie travaille avec une start-up française "Owkin" qui ambitionne d'utiliser de nombreuses données médicales pour trouver le meilleur traitement pour un patient et même découvrir de nouveaux médicaments.
Aujourd'hui, malgré l'existence des solutions technologiques, La transformation du parcours de soins devra être accompagnée par une responsabilisation du grand public. Ce processus de sensibilisation du patient représente une révolution sociétale. Ainsi, en remplaçant le modèle de soin curatif coûteux par un modèle préventif plus responsable.

7. National Health Service
8. est un outil de conversation automatisée avec l'usager.
9. directeur commercial en charge de la santé, la protection sociale et l'industrie pharmaceutique chez IBM France

CHAPITRE

2

MÉTHODE DE CLASSIFICATION POUR LA PRISE DE DÉCISION

Introduction

La classification est une méthode mathématique d'analyse des données utiliser pour faciliter l'étude d'une population d'effectif important comme malades, gènes,...etc,ils sont regroupés en un nombre restreint de classes de telle sorte que les individus d'une même classe soient les plus semblables possibles et que les classes soient les plus distinctes possibles. Son rôle est important pour résoudre plusieurs problèmes en statistiques, reconnaissances de formes, apprentissage automatique, réseaux de neurones et également dans différents domaine comme la médecine, la psychologie, la biologie,...etc. La plupart des algorithmes d'apprentissage tentent de trouver un modèle (une fonction mathématique) qui explique le lien entre des données d'entrée et les classes de sortie.
Nous allons présenter quelques méthode classiques de classification pour la prise de la décision médicale :

- ▷ *L'arbre de décision.*
- ▷ *Le réseau de neurone.*
- ▷ *Régression logistique.*
- ▷ *la classification bayésienne.*
- ▷ *K plus proches voisins.*

2.1 Arbre de décision

-Un arbre de décision est une représentation graphique d'une procédure de classification., on peut dire que c'est un classificateur interprétable et simple à utiliser tout en gardant des performances très acceptables,il est formellement une structure d'arbre comprenant des nœuds qui sont des tests sur les attributs dont Chaque nœud réalise un test

tel que le résultat indique la branche à suivre dans l'arbre, les feuilles(nœuds terminales) spécifient les catégories ou les classes tel que.

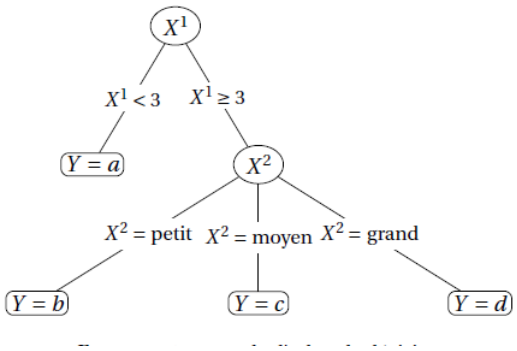

FIGURE 2.1 – Schéma d'un arbre de décision.

Telle que les attributs X^1 et X^2 ont chacun deux valeurs distinctes, si les exemples avec l'attribut X^1 est inférieure strictement à 3, ils correspondent à une seule classe $Y = a$. Dans le cas où l'attribut à X^1 supérieure ou égal à 3, les exemples correspondent à deux classes différentes, on a besoin alors de prendre l'attribut X^2 pour diviser les exemples dans leurs classes respectives.Cependant, on peut voir un arbre de décision comme une partition de l'espace des attribut avec une classe attribuée à chaque élément de cette partition, ces éléments de la partition étant les équivalents des feuilles de l'arbre.

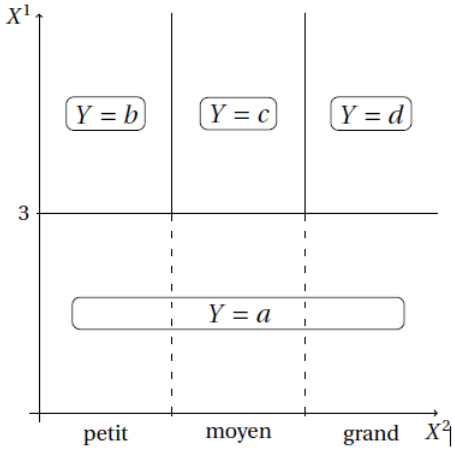

FIGURE 2.2 – partition des classes

2.1.1 Construction d'arbre

Le but de la construction d'un arbre de décision est de bien séparer les classes de manière à obtenir des feuilles le plus "pures" possible en terme de classe. Cette pureté, ou homogénéité, de classe se mesure à l'aide de fonctions d'impureté notées E.
On sélection des meilleur nœuds a l'aide de plusieurs critères, parmi les plus connues on peut citer l'entropie de Shanon pour calculer le gain d'information qui est la réduction de l'entropie à laquelle on peut s'attendre si on fait une partition sur la base de cet

attribut. Soit A une variable aléatoire discrète, prenant n valeurs $A_1, ..., A_n$ de probabilités d'obtention respectives $p_1, ..., p_n$.On définit l'entropie de A, généralement notée $E(A)$:

$$E(A) = -\sum_{i=1}^{n} p_i log p_i$$

où p_i est la proportion d'exemples de catégorie i dans A.
Le gain d'information d'un attribut X est donné par :

$$Gain(A, X) = E(A) - \sum_{j=1}^{n} \frac{|A_j|}{n} E(A_j).$$

Avec : $|A_j|$:taille de A_j.
L'entropie permet de donner la façon standard de quantifier l'information moyenne contenue par une observation x. Plus $p(x)$ est proche d'une loi uniforme, plus l'entropie est élevée.et Plus le gain est élevé, plus cet attribut nous fourni de l'information pour la classification des données. Pour la racine de l'arbre, on choisit donc l'attribut qui maximise ce gain.

2.1.2 Apprentissage

L'algorithme d'apprentissage doit fournir en sortie un arbre de décision. L'idée est de diviser récursivement et le plus efficacement possible les exemples de l'ensemble d'apprentissage par des tests définis (entropie) à l'aide des attributs jusqu'à ce que l'on obtienne des sous-ensembles d'exemples ne contenant (presque) que des exemples appartenant tous à une même classe en utilisant les trois opérateurs suivants :

. Décider si un nœud doit être étiqueté comme une feuille ou bien tous les exemples sont classifie.

. Sélectionner un test à associer à un nœud. Par exemple :l'entropie et le gain d'information.

. Affecter une classe à une feuille.

Algorithm 1 : Construction d'arbre

Entrée : Échantillon S
Répéter :
  Décider si le nœud courant est terminal.
  Si le nœud est terminal **alors**
    Affecter une classe
  Sinon
    Sélectionner un test et créer le sous-arbre
  FinSi
   Passer au nœud suivant non exploré s'il en existe.
  Jusqu'à obtenir un arbre de décision.
  Fin

2.2 Réseau de neurone artificiel (RNA)

Un réseau de neurones peut être vu comme un ensemble d'unités de traitement, appelés nœuds ou neurones, c'est un système composé de plusieurs nœuds fonctionnant en parallèle dont la fonction est déterminée par la structure du réseau et l'opération effectuée par les nœuds. Le principe de fonctionnement est le suivant : on dispose initialement d'une base de d'entrainement constituée de couples de données (entrées / sorties) et on souhaite utiliser cette base de données pour entrainer un algorithme à reproduire les associations constatées entre les entrées et les sorties de l'échantillon.le cas particulier le plus utiliser est souvent donné par le perception multicouches [1]

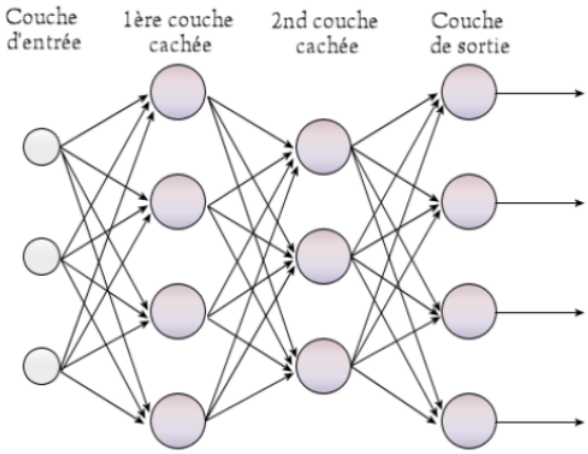

FIGURE 2.3 – L'architecture d'un réseau de neurones multicouches.

On remarque qu'un réseau de neurone est constitué d'une couche de neurones d'entrées, une ou plusieurs couche cachée intermédiaire et d'une couche de sortie où l'on lit la variable prédite dont les sommets sont reliés entre eux dans un et un seul sens, de plus aucun neurone n'est relié à un neurone situé sur la même couche. Le but est de cherche à trouver les poids reliant les neurones de façon à minimiser une fonction de coût dont les valeur de poids est à déterminer en fonction du problème via un algorithme d'apprentissage (propagation, rétro-propagation...). Les réseaux de neurones peuvent être utilisés pour effectuer une classification supervisée de la manière suivante : chaque noeud d'entrée correspond à un attribut de l'objet et on peut prendre un neurone de sortie par classe dont la valeur de sortie est la valeur de la fonction d'appartenance (probabilité que l'objet appartienne à cette classe).

2.2.1 Structure de réseau

Un réseau de neurone est en général composé d'une succession de couches 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 dont chaque

1. Le perceptron est un type de réseau neuronal formel,il a été inventé en 1957 par Frank Rosenblatt au Cornell Aeronautical Laboratory, inspiré par la théorie cognitive de Friedrich Hayek et celle de Donald Hebb, le mot multicouche montre l'existence de plusieurs couche dans le réseau.

synapse [2] est associé un poids synaptique. Généralement Un neurone [3] est constitué d'un intégrateur qui effectue la somme pondérée de ses entrées. Le résultat s de cette somme qui s'appelle le niveau d'activation du neurone est ensuite transformé par une fonction de transfert f comme fonction sigmoïde qui produit la sortie y du neurone. On considére qu'on a une seule cachée, soit n entrées du neurone correspondent au vecteur $X = [x_1, x_2, ..., x_n]^T$ et $w = [w_{11}, w_{21}, w_{n1}]^T$ représente le vecteur des poids et b le biais [4] du neurone. La sortie s est donnée par l'équation suivante :

$$s = \sum_{i=1}^{n} w_{i1} x_i + b = w^T + b$$

Lorsque le niveau d'activation atteint ou dépasse le seuil b, alors l'argument de f devient positif ou nul,sinon il est négatif, puis on ajoutons la fonction d'activation f pour obtenir la sortie du neurone :

$$y = f(s) = f(w^T + b).$$

on suppose que la fonction de transfert est un fonction sigmoïde qui définit par :

$$f(x) = \frac{1}{1+e^{-x}}$$

Ces équation nous amène à introduire un schéma de notre modèle plus compact dans le figure. Le résultat correspond au niveau d'activation qui est ensuite comparé au seuil b par soustraction. Finalement la sortie du neurone est calculée par la fonction d'activation f. La sortie d'un neurone est toujours un scalaire binaire ou non.

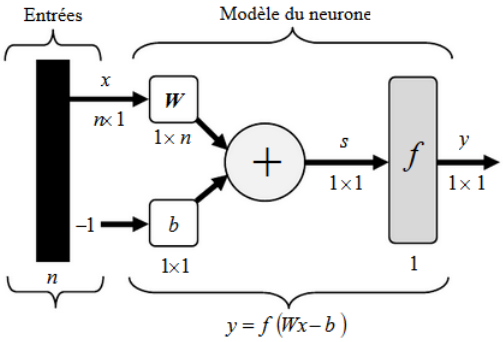

FIGURE 2.4 – Représentation matricielle du modèle d'un neurone artificiel

2.2.2 Apprentissage des poids

L'apprentissage repose sur l'idée de propager vers les couches internes l'erreur quadratique commise en sortie :

$$E = \frac{1}{2} \sum ||Y_{di} - Y_i||^2 = \frac{1}{2} \sum_{j=1}^{m} e_j^2$$

2. Les synapses du neurone quant à eux reçoivent les informations des autres neurones via l'axone et permettent donc aux neurones de communiquer entre eux

3. Un neurone est une cellule nerveuse constituant la base du système nerveux spécialisée dans le traitement des signaux électriques. C'est en 1881 que le terme de neurone (all. Neuron, n) fut intégré au vocabulaire médical par l'anatomiste allemand Heinrich Wilhelm Waldeyer.

4. Le biais est un neurone qui émet un signal d'intensité

pour modifier les poids synaptiques dont l'erreur entre la sortie réelle et la sortie désirée est calculée par somme quadratique des erreurs sur chaque neurone de sortie. En général, l'apprentissage consiste donc à minimiser l'erreur quadratique commise sur l'ensemble des exemples, par ajustement des poids en diminuant le gradient jusqu'au l'erreur commise est inférieure à un seuil choisi on peut dire alors que le réseau a convergé. Le calcul de l'erreur se fait par le principe de rétro-propagation du gradient telle que cette algorithme se résume aux étapes suivantes :
On suppose que le réseau est formé de :
n neurone en entrée, m neurones en sortie et possédant n couches. On note par n_k le nombre de neurones de la couche cachée numéro k.

Algorithm 2 :Rétro propagation du gradient

1-Initialisation

Initialisation la matrice des poids W et seuils du réseau.

2-Présentation d'un exemple en entrée :
On considère $X = (X_1, ..., X_n)$ le vecteur d'entrée et $Y_d = (Y_{d1}, ..., Y_{dn})$ le vecteur de la sortie désirée.
3-Calculer les sorties :
Calculer successivement les sortie des différentes couche à l'aide de la fonction d'activation(sigmoïde).
4- Mettre à jour des poids et des bais :

$$w_{ij} = w_{ij} - \alpha\delta_i y_j$$

$$b_j = b_j - \alpha\delta_i$$

avec α est le coefficient d'apprentissage et δ_i est l'erreur de l'unité i,nous écrivons :

$$\delta_i = (Y_i - Y_{di})\varphi'(e_i)$$

Avec e_i est est l'erreur associe à le neurone i. -Si i est un neurone de sortie :

$$e_i = Y_{di} - Y_i$$

-Sinon

$$e_i = \sum_k \delta_k w_{kj}$$

5-Critère d'arrêt :
Répéter les étapes 2,3 et 4 jusqu'à un nombre maximum d'itérations ou jusqu'à ce que l'erreur quadratique moyenne $E < \alpha$.

2.3 Régression logistique

L'un des principaux modèles d'analyse multivariée en médecine est la régression logistique,elle permet de mesurer l'association entre la survenue d'un évènement (variable expliquée qualitative) et les facteurs susceptibles de l'influencer (variables explicatives).On l'applique à la santé pour Identifier des facteurs liés à une maladie ou bien pour cherche des causes de décès ou de survie de patients. En général, la régression logistique est une

méthode de régression qui est le pendant de la régression linéaire quand Y ne doit prendre que deux valeurs possibles. Ainsi, utiliser pour étudier le lien entre une Variable a expliquer qualitative ou binaire Y et p variables explicatives quantitatives $\{X_j\}_{j=1..p}$.
En régression linéaire multiple dont le modèle est linéaire, la fonction d'hypothèse pourra s'écrire comme suit :

$$Y{=}f(X_1, X_2, .., X_p){+}\epsilon{=}\beta_0{+}\sum_{i=1}^{p} \beta_i X_i + \epsilon$$

avec :

- β_i : est un poids ou paramètre de la fonction hypothèse.
- β_0 :est une constante nommée le biais.

On peut observer qu'on peut réécrire β_0 en : $\beta_0 X_0$ avec $X_0 = 1$, on a définit précédemment X comme étant un vecteur de $X_1, X_2, ..., X_n$. Faisant la même chose pour les β_i, soit grand Θ le vecteur contenant $\beta_0, \beta_1, ..., \beta_p$. Cela nous permet de réécrire notre fonction $f(X)$ de façon plus compacte comme suit :

$$f(X) = \Theta X$$

Dans le cas dans Régression logistique, la fonction logistique f nommée aussi la fonction sigmoïde ne peut être une fonction linéaire car Y ne prend que deux valeurs.Nous écrivons :

$$Y = f(X) + \epsilon = \pi(X) + \epsilon = \frac{e^{\Theta X}}{1+e^{\Theta X}} + \epsilon = \mathbb{P}(Y = 1/X) + \epsilon$$

On appelle cette fonction hypothèse $\pi(X)$: la fonction score. L'idée est de trouver des coefficients $\beta_0, \beta_1, ..., \beta_p$ de sorte que :

- $\pi(X)$ proche de 1 quand la classe vaut 1.
- $\pi(X)$ proche de 0 quand la classe vaut 0.

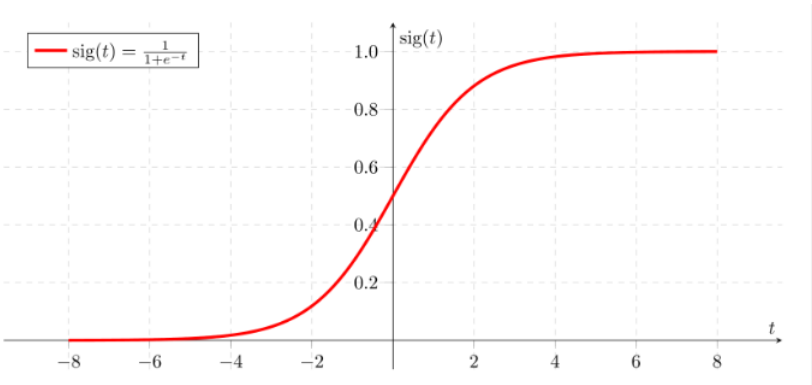

FIGURE 2.5 – Représentation de la fonction logistique

Par l'application de la transformation logit qui permet de revenir à un modèle linéaire, on obtient :

$$Logit(\pi(X)) = ln(\frac{\pi(X)}{1 - \pi(X)}) = \Theta X$$

2.3.1 L'estimation des coefficient

En régression linéaire multiple,on utilise Méthode des moindres carrés [5] mais en la régression logistique ne le permet pas d'obtenir une estimation des coefficients, pour cela, on utilise la méthode du maximum de vraisemblance "Maximum likelihood" qui permet d'estimer les paramètres.Pour simplifier, soit Y binaire et on suppose qu'il y a une seule variable explicative quantitative X,alors La vraisemblance d'un n-échantillon $y_1, y_2, ..., y_n$ est définie comme la probabilité d'observer cet échantillon,pour une observation i :

$$\pi(x_i) = \mathbb{P}(Y = 1/X = x_i)$$

alors la vraisemblance est définit par :

$$L = \mathbb{P}(Y_1 = y_1, Y_2 = y_2, .., Y_n = y_n)$$

Si les observations sont considérées indépendantes entre elles,et comme Y_i suit de loi de Bernoulli [6], on trouve :

$$L = \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i) \text{ avec } \mathbb{P}(Y_i = y_i) = \pi(x_i)^{y_i}[1 - \mathbb{P}(Y_i = y_i]^{1-y_i}.$$

Alors L s'écrit :

$$L = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$$

-L'objectif : trouver β_0 et β_1 qui maximisent la probabilité d'observer l'échantillon c'est à dire maximisation de la vraisemblance :

$$\max_{\beta_0,\beta_1}(L) = \max_{\beta_0,\beta_1}(log(L))$$

pour trouver β_0 et β_1 qui maximisent L, on a recours aux dérivées partielles :

$$\frac{\partial L}{\partial \beta_0} = 0 \text{ et } \frac{\partial L}{\partial \beta_1} = 0$$

2.3.2 Tests dans le modèle

Le tests de Wald et du rapport de vraisemblance s'étendent au test de plusieurs coefficients à la fois :

Test du rapport des vraisemblances :

Le test du rapport des vraisemblance maximales sur la signification globale des coefficients est basée sur statistique khi-deux χ^2 :

H_0 :$\beta_1 = \beta_2 = ... = \beta_p = 0$ (Vraisemblance L_0).
H_1 : $\exists j \in \{1...p\} \quad tel \quad que \quad \beta_j \neq 0$.

Le principe de ce test est des comparer la vraisemblance L_{X_i} (avec variable explicative) avec L_0.le test Sous l'hypothèse H_0 :

$$D = -2ln(\frac{L_0}{L_{X_j}}) \backsim \chi^2_{pd.l.l}.$$

· Si on ne rejette pas H_0 associée au L_0 alors STOP.
· Si on rejette H_0 alors test individuel de chaque coefficient.

5. La méthode des moindres carrés, indépendamment élaborée par Legendre et Gauss au début du $XIXe$ siècle, permet de comparer des données expérimentales, généralement entachées d'erreurs de mesure, à un modèle mathématique censé décrire ces données.

6. loi de Bernoulli du nom du mathématicien suisse Jacques Bernoulli, est une distribution discrète de probabilité, qui prend la valeur 1 avec la probabilité p et 0 avec la probabilité q = 1 − p.

Test de Wald :

Le test de Wald se fait sur un coefficient avec les hypothèses :

▪ $H_0 : \beta_j = 0$ ▪ $H_1 : \beta_j \neq 0$ Le test sur chaque coefficient β_j ,sous l'hypothèse H_0 :

$$K = \frac{\widehat{\beta_j}^2}{s^2_{\widehat{\beta_j}}} \backsim \chi^2_{1d.l.l}$$

Si on rejette H_0,le paramètre β_j est significatif,la variable j a une influence sur la probabilité d'apparition de l'évènement, sachant les autres variables du modèle.

2.3.3 Interprétation des coefficients

Dans le coté médicale, l'odds ratio c'est le rapport des cotes des probabilités d'avoir la maladie pour ceux qui ont un symptôme X d'une part et de ceux qui ne l'ont pas d'autre part,on peut dire aussi que un mesure d'association entre symptôme et la maladie. Pour simplifier, On considère un seul variable explicative X, soit :

$$Logit(\pi(X)) = \beta_0 + \beta_1 X$$

alors l'odds ratio entre les classe $X = 1$ et $X = 0$ est donné par :

$$OR = e^{\beta_1}$$

En effet : on a $$logit(\pi(1)) - logit(\pi(0)) = \beta_1$$
alors

$$logit(\frac{\pi(1)}{\pi(0)}) = ln(\underbrace{\frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}}}_{OR}) = \beta_1.$$

Interprétation de l'Odds-ratio :

·$OR = 1$: pas d'association.

·$OR > 1$: X est un facteur de risque de Y.

·$OR < 1$: X est un facteur protecteur de Y.

Donc l'exponentiel du coefficient peut être interpréter comme un odds-ratio. Le principe de cette interprétation est de présenter la relation linéaire entre X et $Logit(\pi(X))$ en trois étapes :

▷ Découper X en groupes.

▷ Pour chaque intervalle on calcule $\mathbb{P}(Y = 1/X = x_i)$.

▷ Représenter graphiquement $Logit(\pi(X))$ en fonction des groupes de X.

2.3.4 Adéquation du modèle

Le but de l'adéquation ou la validation de modèle est de déterminer la qualité d'ajustement du modèle aux données,Si l'ajustement est correct, les valeurs prédites seront proches des valeurs observées. La Comparaison des probabilités prédites par le modèle $\widehat{\pi}(X_j)$ à celles observées dans l'échantillon s'appelle la calibration du modèle. On peut déterminer la calibration par test de Hosmer-Lemeshow [7].

7. Le test est utilisé dans les modèles de prédiction du risque,il évalue si les taux d'événements observés correspondent aux taux d'événements attendus dans les sous-groupes de la population.

Test de Hosmer-Lemeshow :

-On calcule pour chaque observation la probabilité prédite par le modèle $\pi(X_j)$.
-On classe les observations par groupes de probabilités prédites.
-On compare dans chaque classe les effectifs ou l'écart observés et les effectifs théoriques, si dans chaque classe ces deux effectifs sont proches alors le modèle est calibrée,sinon s'il existe des classes dans lesquelles les effectifs sont trop différents, alors le modèle est mal calibrée,on utilisant une autre fois le test de χ^2 sous les hypothèse :

H_0 :distance faible (les probabilités prédites sont proches de celles observées).
H_1 :distance élevée (les probabilitès prédites sont différentes des observées).

Si on ne rejette pas H_0, on peut dire que le modèle est calibré
En pratique on ne rejette pas H_0 si $p > 0.2$

2.4 Classification bayésienne (Naïve bayésienne)

Un classification naïve bayésienne(naive bayes en anglais) est un modèle assez simple, robuste et rapide à utiliser pour les problématiques de classement avec des variables explicatives qualitatives. Ainsi, La raison pour laquelle l'algorithme Naive Bayésienne s'appelle Naive n'est pas parce qu'il est simple ou stupide mais c'est parce que l'algorithme repose sur une hypothèse très forte selon laquelle les données ont des attributs indépendantes entre eux, alors qu'en réalité, elles peuvent être dépendantes d'une manière ou d'une autre. En d'autres part, cela suppose que la présence d'une fonctionnalité dans une classe est complètement indépendante de la présence de toutes les autres fonctionnalités dont Cette méthode d'apprentissage basé sur l'application du théorème de Bayes avec l'hypothèse d'indépendance "naïve" entre toutes les caractéristiques.
Cet outil de classification doit dans un premier temps être entrainé sur une base de données d'apprentissage qui prédites la classe de sortie en fonction des entrées. Pendant la phase d'apprentissage, l'algorithme élabore ses règles de classification sur ce des donnée, pour les appliquer dans un second temps à la classification des données de prédiction. Le classificateur bayésien naïf implique que les classes de la phase d'apprentissage soit connu et fournit, d'où le caractère supervisé de l'outil.

2.4.1 Les étapes de la classification bayésienne

la classification est réalisé suivant deux étapes :

Identification des variables :

L'identification des variables s'agit de déterminer l'ensemble des variables X_i catégoriel ou numérique,qui caractérisent le système. Dans le coté médicale les variables X_i représentent les symptômes d'une maladie étudie, ensuite identifier les classes de chaque patient.

Loi de probabilité conjointe des variables :

Cette étape consiste à calculer pour chaque variable sous forme vecteur X_j, la probabilité conditionnelle $P(C_i/X_j)$ pour chacune classe de sortie C_i avec $i \in \{1, 2, ..., L\}$ à l'aide de la règle Bayes[8] ,puis affecter l'individu(patient) à la classe qui génère la probabilité maximale. La règle de Bayes est une façon de passer de $P(X_i/C_j)$ connu de l'ensemble de données d'apprentissage, à la recherche de $P(C_i/X_j)$,nous écrivons :

$$P(C_i/X_1, ..., X_n) = \frac{P(C_i)P(X_1, ..., X_n/C_j)}{P(X_1, ..., X_n)} \qquad \forall i \in \{1...k\}.$$

En langage courant, cela signifie :

$$postérieure = \frac{antérieur * vraisemblance}{évidence}.$$

En introduisant l'hypothèse naive : si chaque X_i est indépendant des autres caractéristiques alors :

$$P(X_1, ..., X_n/C_j) = P(X_1/C_j)P(X_2/C_j)...P(X_n/C_j) = \prod_{k=1}^{n} P(X_k/C_j)$$

Pour chacune des variables est associée une table de probabilités conditionnelles.

2.4.2 Apprentissage

On cherche à trouver une règle de décision qui maximise la probabilité a posteriori, définissant quelle décision prendre en fonction de l'entité observée à partir de trouver la classe d'une observation par une approche probabiliste "maximum a posteriori".

$$C_{MAP} = argmax_k P(C_k|X_1, ..., X_n)P(X_1, ..., X_n) = argmax_k P(C_k) \prod_{k=1}^{n} P(X_k/C_k)$$

avec l'estimation des probabilité $P(C_k)$ et $P(X_1, ..., X_n/C_k)$ trouver à partir de l'ensemble des observations par contre la probabilité $P(X_1, ..., X_n)$ est la même pour toutes les classes, on suppose que les variables sont discret alors estimation de $P(C_k)$ s'écrit :

$$\widehat{P}(C_k) = \frac{n_k}{n}$$

où n est le nombre total d'observations et n_k est le nombre d'observations de classe C_k. On suppose que a_i le $i^{éme}$ composantes de X_i, alors la probabilité $P(X_i/C_k)$ est estimée par la fréquence relative des éléments ayant la valeur v_i pour X_i et qui sont dans C_k :

$$\widehat{P}(X_i/C_k) = \prod P(a_i = v_i/C_k)$$

avec

$$P(a_i = v_i/C_k) = \frac{v_i}{n_k}.$$

où v_i est le nombre d'instances qui ont l'attribut a_i et qui appartiennent à la classe C_k.

Malgré les hypothèses d'indépendance relativement entre les observations, le classifieur bayésien naïf a plusieurs propriétés qui le rendent très pratique dans les cas réels. Comme, la dissociation des lois de probabilités conditionnelles de classe entre les différentes caractéristiques aboutit au fait que chaque loi de probabilité peut être estimée indépendamment en tant que loi de probabilité à une dimension.

8. Thomas Bayes est un mathématicien britannique et pasteur de l'Église presbytérienne, connu pour avoir formulé le théorème de Bayes.

2.5 K plus proche voisin (K-ppv)

La méthode des plus proches voisins noté parfois **k-PPV** ou **k-NN** pour (k-Nearest-Neighbor) a été utilisée dans la classification en particulier pour la prise de la décision médicale comme une technique non paramétrique. Elle consiste à déterminer pour chaque nouvel individu que l'on veut classer, la liste des plus proches voisins parmi les individus déjà classés. L'individu est affecté à la classe qui contient le plus d'individus parmi ces plus proches voisins.
La méthode **K-NN** suppose que les données se trouvent dans un espace de caractéristiques. Cela signifie que les points de données sont dans un espace métrique, pour cela il nécessite de choisir une distance le nombre de voisins à prendre en compte.

2.5.1 Principe de K-PPV

Pour trouver la classe d'un nouveau individu (patient), l'algorithme de K-PPV se base sur le principe suivant : il cherche les k plus proches voisins de ce nouveau individu, ensuite, il choisit parmi les candidats trouvés le résultat le plus proche et le plus fréquent. Pour affecter un nouvel individu à une classe, l'algorithme cherche les k plus proches voisins parmi les individus déjà classés. Ainsi, l'individu est affecté à la classe qui contient le plus d'individus parmi les candidats trouvés.
Cette méthode utilise principalement deux paramètres :

▷ Une fonction de similarité(la distance) pour comparer les individus dans l'espace de caractéristiques qu'elle permet de mesurer le degré de différence entre deux vecteurs.Il existe plusieurs types de distance parmi lesquels on trouve :

La distance euclidienne :

C'est une distance qui calcule la racine carrée de la somme des différences carrées entre les coordonnées de deux points :

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

où x et y sont des vecteurs.

Distance Manhattan :

la distance de Manhattan : calcule la somme des valeurs absolues des différences entre les coordonnées de deux points :

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

où x et y sont des vecteurs.

Il existe d'autres distances selon le cas d'utilisation de l'algorithme, mais la distance euclidienne reste la plus utilisée.

▷ le nombre k qui décide combien de voisins influencent la classification à l'aide de la méthode de validation croisé.

validation croisé

Validation croisé représente comme une technique de ré-échantillonnage, elle est une stratégie universelle pour tester la précision des prédictions de tout les modèles probabiliste. De plus, la validation croisée est aussi utilisé comme un outil empirique pour déterminer (à travers des tests successifs) des paramètres qui autrement seraient trop difficiles voir et impossible trouver comme paramètre K pour classifieur k-ppv :

▷ **Choix du partitionnement** : diviser l'échantillon générale en k partitions de même taille, avec la partition k_i est l'échantillon de validation et le reste représente l'échantillon d'entrainement.

▷ **Déroulement du test :** pour chaque partition on applique le modèle à l'échantillon de validation,puis on estime les performances du modèle.

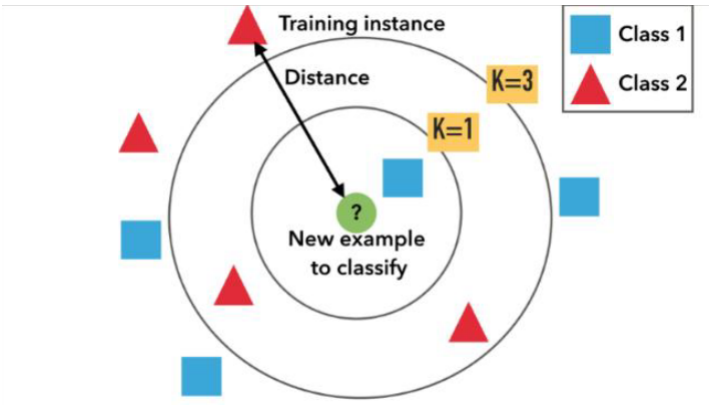

FIGURE 2.6 – Exemple de classification de K-NN.

Interprétation de l'exemple : L'échantillon de test (cercle vert) doit être classé soit dans la première classe de carrés bleus, soit dans la deuxième classe de triangles rouges. Si k = 3 (Cercle de trait plein), il est assigné à la deuxième classe parce qu'il y a 2 triangles et seulement 1 carré dans le cercle intérieur. Si k = 5 (cercle de lignes pointillées), il est assigné à la première classe (3 carrés contre 2 triangles à l'intérieur du cercle extérieur).

Algorithm 3 Apprentissage par K-PPV

Entrée :
-S ensemble d'entrainement.
-Y : étiquettes la classe de X. -K, le nombre des plus proches voisins de cet ensemble de données à retrouver.
-x l'ensemble de données à traiter.
Pour $i = 1...m$ **faire**
Calculer la distance $d(X_i, x)$
FinPour
Construire l'ensemble I contenant des indices pour k plus petite distance $d(X_i, x)$.
Retourner Étiquette majoritaire pour $\{Y_i \quad ou \quad i \in I\}$

CHAPITRE

3

TEST ET APPLICATION

Introduction

En apprentissage automatique (Machine learning), un problème de classification se présente sous la forme d'un ensemble de données contenant des exemples issus de l'observation d'un phénomène. Chaque exemple est constitué d'une description(Entrée) et d'une étiquète (sortie).

Un algorithme d'apprentissage automatique analyse ses données afin de construire un classificateur. C'est à ce classificateur que revient ensuite la tâche d'étiqueter de nouveaux exemples à partir de leur description. Afin d'étudier le problème de classification auquel nous nous intéresserons, dans un premier temps nous intéressons à l'application des méthodes d'apprentissage dans le domaine médical, notamment dans le diabète de type 2. Pour finir, nous nous effectuerons une comparaison entre les algorithme en utilisant les outils d'évaluation.

3.1 Présentation de diabète

3.1.1 Définition

L'Organisation Mondiale de la Santé définit le diabète comme un trouble du métabolisme d'étiologies multiples, caractérisé par une hyperglycémie chronique avec des troubles du métabolisme des glucides, lipides et de protéines résultant de défauts de sécrétion d'insuline, d'action de l'insuline, ou les deux.

La plupart de nourriture se décompose en glucose qui sera utilisé par nos cellules pour produire de l'énergie. Cependant, le glucose ne peut pas pénétrer les cellules sans la présence de l'insuline, une hormone produite par le pancréas. Après avoir mangé quelque chose, le pancréas sécrète automatiquement une quantité suffisante d'insuline pour transporter le glucose présent dans le sang aux cellules et pour diminuer le taux de sucre dans le sang. Une personne qui a du diabète souffre d'hyperglycémie c'est-à-dire que la quantité

de glucose dans le sang est trop élevée. car Le corps ne produit pas assez d'insuline ou les cellules ne réagissent pas correctement à l'insuline produite par le pancréas.

3.1.2 Cause de diabète

La prévalence de diabète été augmentée par cinq en moins de cinquante ans. Cette accroissement progressive est due à divers facteurs :

- le vieillissement global de la population,
- l'augmentation de l'espérance de vie du diabétique,
- l'augmentation de la fécondité des femmes diabétique,
- l'augmentation de l'obésité,
- l'incrémentation de la consommation des sucres raffinés.

Il existe d'autre facteurs qui peuvent servir comme déclencheur tels que :

- les régimes riches en graisse et protéine,
- la consommation réduite de fibre,
- une alimentation déficiente en hydrate de carbone complexe et vitamine E,
- le stress chronique,
- le tabagisme qui peut causer l'apparition de l'insulinores istance.

3.1.3 Classification de diabète

La classification du diabète se base sur son étiologie et caractéristique physiopathologique. On peut distingue trois type de diabète :

► **Diabète de type 1** : ce type de diabète survient lorsque le pancréas ne produit plus d'insuline en raison d'une réaction auto-immune [1], ceci est une conséquence de l'autodestruction progressive et sélective des cellules Béta des îlots de Langerhans, ce qui affecte l'utilisation des hydrates de carbone, protéine et graisse.
Puisque un pancréas sain sécrète une quantité d'insuline plus élevé de ce dont le corps à besoin, des mois ou des années peuvent passer avant que la maladie ne soit pas diagnostiquée. La vitesse à laquelle les cellules se détruisent dépendra de l'âge du malade. Une fois le traitement avec insuline établie, l'organisme passera par une période, allant jusqu'à un an, durant laquelle les sécrétions d'insuline se réinitialisent et les besoins exogènes diminuent, sans pour autant abandonner le traitement. Mais après une dizaine d'années, les cellules Béta seront entièrement détruites, donc toute l'insuline nécessaire devra être administrée par injection. A l'heure actuelle, Nous ne savons pas exactement pourquoi les gens développent un diabète, mais il y a de nombreuses recherches sont en cours pour trouver une explication.

► **Diabete de type 2** : Dans ce type de diabète l'altération métabolique n'est pas aussi intense que pour diabète de type 1 et l'évolution de la maladie est progressive. Ce type est caractérisé par la résistance ou la faible sensibilité du corps à l'insuline, c'est-à-dire que le taux d'insuline endogène peut se trouver dans les paramètres normaux, mais les tissus sont incapables de l'assimiler et par conséquent le taux de glucose dans le sang augmente.

La liaison insuline-récepteur active un deuxième messager qui induit la synthèse des protéines et l'activation et inhibition des enzymes intracellulaire. Les gens souffrant de diabète de type 2 ont des altérations dans les mécanismes post-récepteurs, ce qui oblige

1. le corps produit des anticorps contre ses propres cellules. Lorsque cela se produit, c'est que votre pancréas ne peut plus produire sa propre insuline

l'organisme à augmenter la sécrétion d'insuline pour compenser et ceci peut conduire à l'épuisement des cellules Béta. Pour des personnes avec une certaines prédisposition les cellules ne seront pas capable de maintenir un taux de glucose normale ce qui conduit à l'apparition du diabète. Même si les diabétiques de ce type ne nécessitent pas les injections d'insuline pour survivre, la plupart de diabétique de type 2 est la conséquence de leur obésité tel que la graisse abdominal est la plus dangereuse.

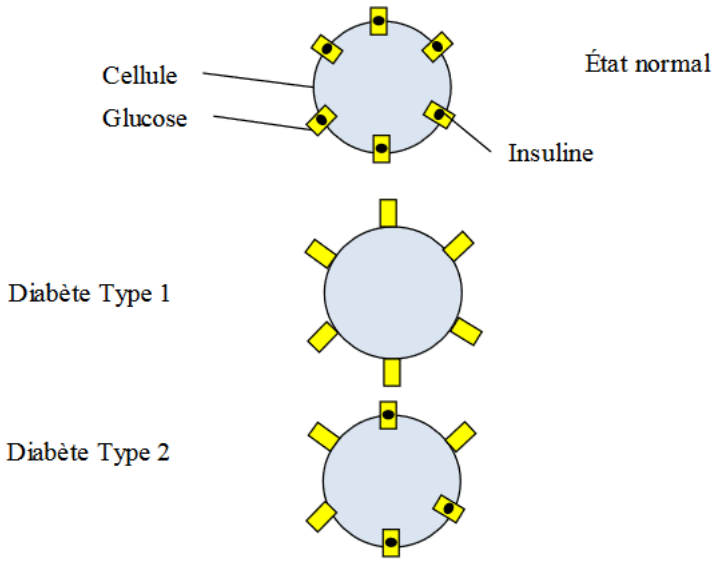

FIGURE 3.1 – Fonctionnement de l'insuline.

► **Diabète gestationnel** : Pendant la grossesse, surtout pendant le troisième trimestre, une incrémentation des besoins insuliniques se produit. Environ 4% des femmes subiront des altérations durant leur grossesse. Ce diabète d'intolérance disparait souvent à la fin de la grossesse, mais il est important de diagnostiquer ce type car les patientes peuvent expérimenter une augmentation du risque de mortalité fœtale si elles ne sont pas traitées correctement.

3.2 Présentation de la base de donnée utilisées

Les Pimas sont un groupe d'Amérindiens vivant en Arizona. Une prédisposition génétique a permis à ce groupe de survivre normalement pendant plusieurs années à un régime pauvre en glucides. Ces dernières années, en raison du passage soudain des cultures agricoles traditionnelles aux aliments transformés et du déclin de l'activité physique, ils ont développé la prévalence la plus élevée de diabète de type 2 et ont pour cette raison fait l'objet de nombreuses études. L'institut national du diabète et des maladies digestives et rénales de l'institut national de la santé étaient à l'origine de la base de données sur le diabète "Pima Indians Diabetes".

Cette base est une collection de rapports de diagnostic médicaux de 768 femmes (les Pimas) âgé de plus de 21 ans 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, le tableau suivant présente une description de ces attributs :

	description attribut	Moyenne	Déviation standard
1	Nombre de grossesses (Ngross)	3.8	3.4
2	Concentration de glucose plasmatique(mg/dl)	120.9	32.0
3	Pression artérielle diastolique(mm Hg)(PAD)	69.1	19.4
4	Épaisseur de la peau u niveau du triceps(mm)(Epai)	20.5	16.0
5	Taux d'insuline au bout de 2 heures(mU1/4ml)(INS)	79.8	115.2
6	Indice de masse corporelle(kg/m^2)(IMC)	32.0	7.9
7	Fonction pedigree du diabète(Ped)	0.50	0.3
8	Age	33.0	11.8

TABLE 3.1 – Description des attributs de la base.

Voilà un échantillon de cette base du diabète de type 2 qui comporte 19 patiente et la dernière case représente la classe de chaque patiente.

Data: newdata

	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age	diabetes
1	1	89	66	23	94	28.1	0.167	21	0
2	0	137	40	35	168	43.1	2.288	33	1
3	3	78	50	32	88	31.0	0.248	26	1
4	2	197	70	45	543	30.5	0.158	53	1
5	1	189	60	23	846	30.1	0.398	59	1
6	5	166	72	19	175	25.8	0.587	51	1
7	0	118	84	47	230	45.8	0.551	31	1
8	1	103	30	38	83	43.3	0.183	33	0
9	1	115	70	30	96	34.6	0.529	32	1
10	3	126	88	41	235	39.3	0.704	27	0
11	11	143	94	33	146	36.6	0.254	51	1
12	10	125	70	26	115	31.1	0.205	41	1
13	1	97	66	15	140	23.2	0.487	22	0
14	13	145	82	19	110	22.2	0.245	57	0
15	3	158	76	36	245	31.6	0.851	28	1
16	3	88	58	11	54	24.8	0.267	22	0
17	4	103	60	33	192	24.0	0.966	33	0
18	4	111	72	47	207	37.1	1.390	56	1
19	3	180	64	25	70	34.0	0.271	26	0

FIGURE 3.2 – Échantillon de la base du diabète de type 2.

Malheureusement, il existe quelques cas avec des données manquantes qui ont été remplacé par des zéros, ce qui donne des valeurs impossible tel qu'une pression artérielle égal à 0. Pour cela, on peut fire une élimination de ces cas, nous obtenons une base de 392 patientes dont 262 non diabétiques et 130 diabétiques.

3.3 Simulation sous R

Introduction

Cette section est consacré à nos différentes approches d'apprentissage automatique ; nous concernons de traiter notre base de données "Pima Indians Diabetes" avec chacun

des algorithmes d'apprentissage, puis nous réalisons une étape comparative. En particulier, nous analysons l'influence de leurs paramètres sur la prédiction ou bien les outils d'évaluation de algorithme.

Avant de concentrer sur les outils d'apprentissage, nous avons procédé aux points suivant :

1· Collection des données et les rendre propre c'est-à-dire renseigner les valeurs vides, supprimer et remplacer les valeurs aberrantes.

2· Découper les données en deux phase (données d'apprentissage et test).

3· 4. Apprentissage et évaluation, en appliquant la technique de prédiction.

4· Évaluation de performance de chaque algorithme.

5· Exportation du modèle (meilleur algorithme + prétraitement) sous format binaire.

3.3.1 L'application de l'algorithme d'arbre de décision

Aujourd'hui, il n'existe pas de fonction offrant une interface graphique pour réaliser des arbres décisionnels avec R. Heureusement, les commandes sont relativement simples, il suffit juste de cinq ou six lignes de code pour construire un arbre de décision et déterminer leur précision a l'aide de la fonction "rpart" et la commande "prp" avec l'échantillon d'apprentissage.

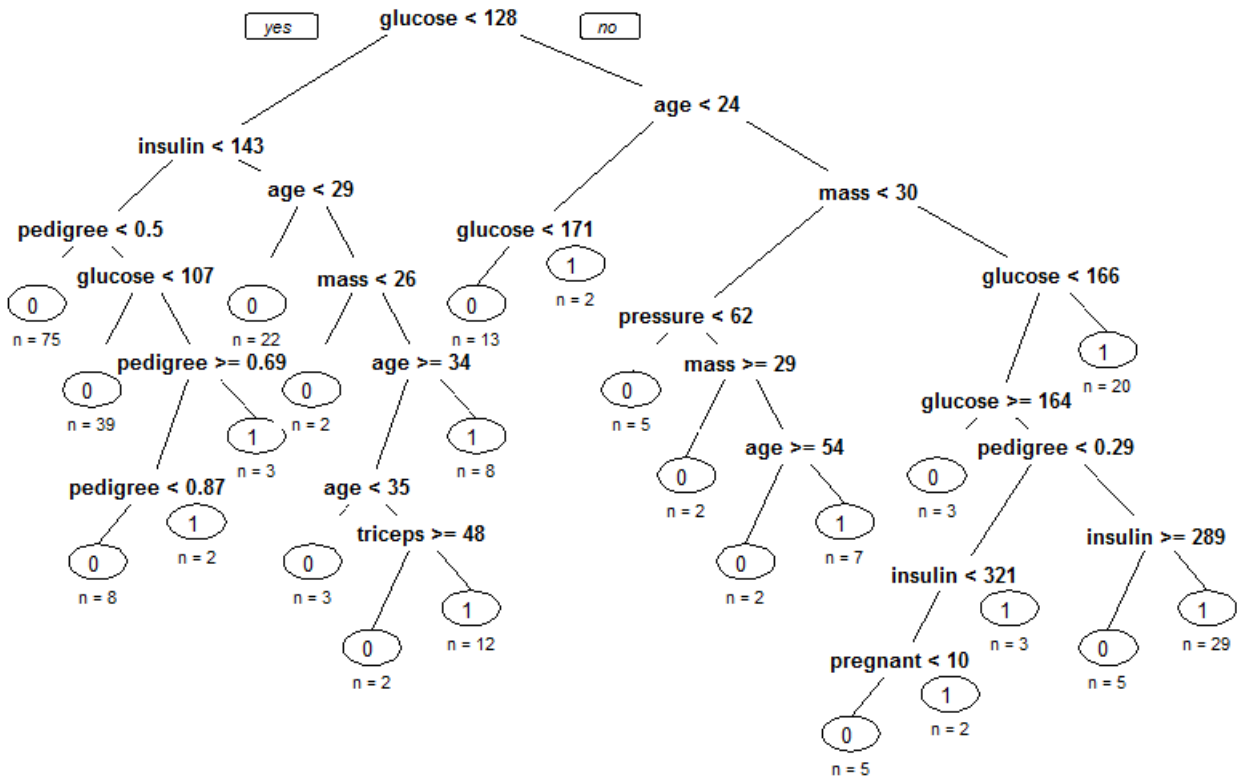

FIGURE 3.3 – Arbre de décision pour la prédiction de les variables.

Ici, pour chaque nœud terminal, logiciel R indique la classe prédite, le nombre des patients de la classe prédite à gauche et le nombre des patients de l'autre classes à droite.

Plus l'arbre est complexe (beaucoup de nœuds), plus il va bien apprendre l'échantillon d'apprentissage, mais aussi plus il va faire de l'overfitting c'est à dire adaptation uniquement à l'échantillon d'apprentissage, mais beaucoup d'erreurs sur un nouvel échantillon de test. La complexité doit donc être pénalisée, d'où un paramètre cp (complexity parameter).

	CP	nsplit	rel error	xerror	xstd
1	0.258427	0	1.00000	1.00000	0.087099
2	0.078652	1	0.74157	0.91011	0.084870
3	0.033708	3	0.58427	0.95506	0.086035
4	0.028090	6	0.48315	0.86517	0.083601
5	0.022472	11	0.33708	0.84270	0.082926
6	0.016854	15	0.24719	0.80899	0.081861
7	0.011236	17	0.21348	0.82022	0.082223
8	0.010000	24	0.13483	0.88764	0.084249

FIGURE 3.4 – la taille d'arbre optimale.

Sous R, on peut élaguer cet arbre de décision en utilisant la commande "printcp" pour déterminer la taille optimale et pour obtenir le même résultat mais sous la forme de tableau contingence avec l'échantillon de test :

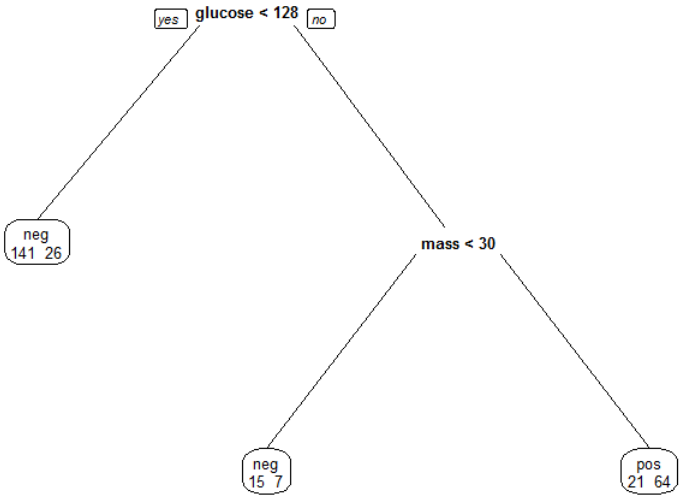

FIGURE 3.5 – Arbre de décision élagué.

Pour la matrice de confusion de l'algorithme arbre de décision, on observe que les valeurs de prédiction sont mentionnées par ligne et les valeurs réelles par colonne.

A partir de la matrice de confusion, nous allons constater la performance de cette méthode a l'aide des outils d'évaluation de qualité, on obtient :

$$Sensibilité = 0.779$$
$$Spécificité = 0.814$$
$$Arruracy = 0.779$$

L'algorithme Arbre de décision ayant taux de précisons de 77.9% avec 26 observations sont mal classée.

	0	1
0	66	11
1	15	26

TABLE 3.2 – Matrice de confusion de l'arbre de décision.

3.3.2 L'application de l'algorithme de réseau de neurone

Malgré l'utilisation du modèle de réseau de neurones n'ait pas été pertinente au début, mais elle a récemment gagné en notoriété grâce à la popularisation des outils de données volumineuses.

Tout d'abord, nous allons commencer l'algorithme d'apprentissage automatique réseau de neurone en utilisant la fonction nnet pour trouver la précision de cette méthode.

-On fixe le nombre de neurones à 5 à laquelle la méthode est converge.

```
# weights:  51
initial  value 179.661233
iter  10 value 156.695315
iter  20 value 155.543823
iter  30 value 155.208242
iter  40 value 154.719472
iter  50 value 154.717993
iter  60 value 154.717792
final  value 154.717759
converged
```

La convergence et les temps d'exécution sont rapides et ce sont des caractéristiques fortes de ce package. Ensuite, la création d'un réseau de neurone se fait à l'aide de la fonction "nnet".

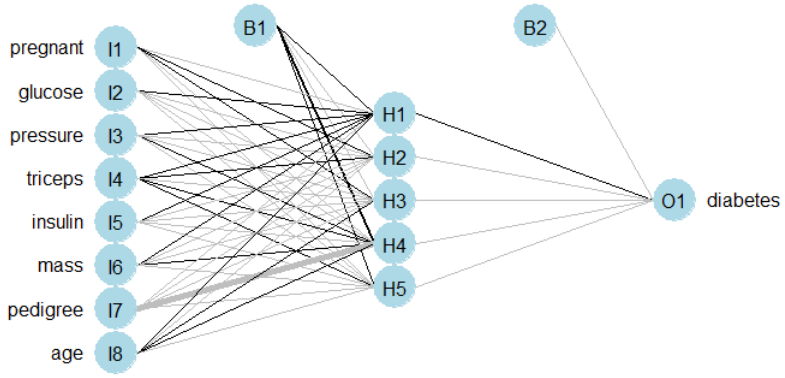

FIGURE 3.6 – Perception multicouche de la base de diabète.

Aprés de choisir le taux d'apprentissage $\alpha = 0.5$ et le biais éventuel $\theta = -\alpha$, le calcule des nouveaux poids synaptique se fait à l'aide de la fonction "predict".

```
b->h1  i1->h1  i2->h1  i3->h1  i4->h1  i5->h1  i6->h1  i7->h1  i8->h1
 0.37   -0.63    2.90    2.36    0.66    1.22    0.30   -0.11    0.90
b->h2  i1->h2  i2->h2  i3->h2  i4->h2  i5->h2  i6->h2  i7->h2  i8->h2
-0.64    0.02   -2.29   -0.54    0.20   -3.52   -0.95   -0.45   -0.73
b->h3  i1->h3  i2->h3  i3->h3  i4->h3  i5->h3  i6->h3  i7->h3  i8->h3
-0.43    0.57   -2.91   -0.89   -1.24   -4.22   -1.32   -0.02    0.05
b->h4  i1->h4  i2->h4  i3->h4  i4->h4  i5->h4  i6->h4  i7->h4  i8->h4
49.33  -31.82   -4.90   23.11   21.54  -10.38    4.61 -173.80   18.70
b->h5  i1->h5  i2->h5  i3->h5  i4->h5  i5->h5  i6->h5  i7->h5  i8->h5
 0.40   -0.05   -1.38   -0.10    0.21   -2.50   -0.96   -0.61   -0.76
 b->o   h1->o   h2->o   h3->o   h4->o   h5->o
-0.17    0.80   -1.65   -1.40   -1.77   -1.69
```

FIGURE 3.7 – Mise à jour des poids synaptique.

Maintenant, on va voir des performances prédictives c'est-à-dire nous effectuons la prédiction sur l'échantillon test, et nous confrontons les valeurs prédites et observées. A

	0	1
0	52	12
1	28	26

TABLE 3.3 – Matrice de confusion de réseau de neurone.

partir de ca tableau, nous allons constater la qualité de cette méthode a l'aide des outils d'évaluation de qualité suivant :

$$Sensibilité = 0.684$$
$$Spécificité = 0.65$$
$$Arruracy = 0.661$$

L'algorithme de réseau de neurone ayant taux de précisons de 66.1% avec 26 observations sont mal classée.

3.3.3 L'application de l'algorithme de régression logistique

la régression logistique consiste à tester l'ensemble de données diabétiques à l'aide de la vraisemblable et test de Hosmer-Lemeshow.

En effet, l'estimation du modèle de régression logistique à un modèle linéaire et le calcul de les coefficients se font à l'aide de la commande "glm". Sous les hypothèses avec $\alpha = 0.01$:

$$H_0 : \beta_0 \neq 0.$$
$$H_1 : \exists i \in \{1, ..., 8\} \quad tel \quad que \quad \beta_j \neq 0.$$

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.637623   1.624077  -7.166 7.74e-13 ***
pregnant      0.117047   0.069871   1.675  0.09390 .
glucose       0.042282   0.007769   5.442 5.26e-08 ***
pressure     -0.000523   0.014403  -0.036  0.97103
triceps       0.002458   0.020496   0.120  0.90455
insulin      -0.001566   0.001613  -0.971  0.33144
mass          0.109276   0.033468   3.265  0.00109 **
pedigree      1.209498   0.525573   2.301  0.02138 *
age           0.035719   0.022098   1.616  0.10601
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 3.8 – Coefficient estimée.

On remarque que les variable "glucose" et "mass" sont significatifs, par contre les autres variables sont non significatifs.

Pour l'interprétation des coefficients, on calcule "OR" pour chaque coefficient on trouve que tout les exponentielle de coefficient sont inférieur à 1 :

```
                OR      2.5 % 97.5 %         p
(Intercept) 8.8276e-06 2.9402e-07 0.0002 7.740e-13 ***
pregnant    1.1242e+00 9.8147e-01 1.2926  0.093896 .
glucose     1.0432e+00 1.0280e+00 1.0600 5.261e-08 ***
pressure    9.9948e-01 9.7174e-01 1.0286  0.971035
triceps     1.0025e+00 9.6260e-01 1.0435  0.904547
insulin     9.9843e-01 9.9526e-01 1.0016  0.331435
mass        1.1155e+00 1.0468e+00 1.1943  0.001094 **
pedigree    3.3518e+00 1.2240e+00 9.6266  0.021375 *
age         1.0364e+00 9.9350e-01 1.0842  0.106010
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

FIGURE 3.9 – Les Odds ratio.

Pour l'adéquation de modèle on calcule les probabilité prédite de variable "diabetes" pour chaque femme, puis on le compare avec les valeurs observée, on minimise l'erreur de entre les probabilité prédite et les valeurs observée avec un seuil $\alpha = 0.5$ à l'aide de la fonction "predict". Ensuite , on va construire le tableau de confusion pour calculer les méthodes d'évaluation de la qualité de cette algorithme d'apprentissage automatique : A

	0	1
0	73	18
1	7	20

TABLE 3.4 – Matrice de confusion de la régression logistique.

partir de ce tableau les indicateurs de performance sont :

$$Sensibilité = 0.526$$
$$Spécificité = 0.91$$
$$Arruracy = 0.788$$

Cette algorithme d'apprentissage automatique ayant taux de précisons de 78.8% avec 25 individus sont mal classée.

3.3.4 L'application d'algorithme naïve bayésien

Le classifieur naïf de Bayes fait une hypothèse simplificatrice pour permettre au calcul de s'échelonner. Dans cette base de diabètes nous supposons que les variables prédictives sont conditionnellement indépendantes les unes des autres compte tenu de la valeur de la réponse(classe). C'est une hypothèse extrêmement forte, nous pouvons voir rapidement que les données d'attrition enfreignent cette règle parce que nous avons plusieurs variables moyennement à fortement corrélées.

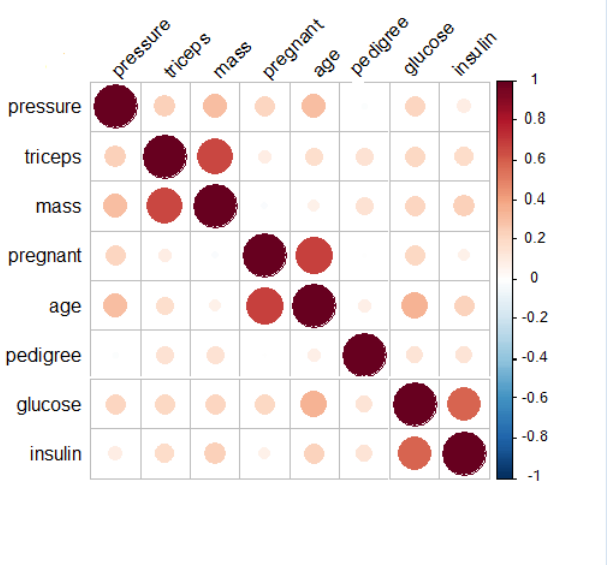

FIGURE 3.10 – Graphe de corrélation entre les attributs de la base.

Même si ces caractéristiques sont liées dans la réalité, un classifieur bayésien naïf déterminera qu'un patiente a le diabète ou non, en considérant indépendamment ses caractéristiques de mass, taux d'insuline, nombre de grossesses s... Il est bien évidemment possible d'introduire certaines relations de dépendance, avec l'utilisation par exemple d'un ratio "mass/insulin ", mais en considère l'indépendance comme même.

Cette corrélation peut représenter sous forme d'une matrice.

	pressure	triceps	mass	pregnant	age	pedigree	glucose
pressure	1.000	0.230	0.300	0.2100	0.300	-0.0160	0.21
triceps	0.230	1.000	0.660	0.0930	0.170	0.1600	0.20
mass	0.300	0.660	1.000	-0.0250	0.070	0.1600	0.21
pregnant	0.210	0.093	-0.025	1.0000	0.680	0.0076	0.20
age	0.300	0.170	0.070	0.6800	1.000	0.0850	0.34
pedigree	-0.016	0.160	0.160	0.0076	0.085	1.0000	0.14
glucose	0.210	0.200	0.210	0.2000	0.340	0.1400	1.00
insulin	0.099	0.180	0.230	0.0790	0.220	0.1400	0.58

FIGURE 3.11 – Matrice de corrélation.

Cependant, en faisant cette hypothèse, nous pouvons simplifier notre calcul de sorte que la probabilité a posteriori soit simplement le produit de la distribution de probabilité pour chaque variable conditionnée par l'attribut catégorique "diabetes".

Nous pouvons calculer le modèle naïf de Bayes dans le package "h2o" avec la commande "naiveBayes" et pour crée la probabilité conditionnelle pour chaque entité séparément. Nous avons aussi les probabilités a priori qui indique la distribution de nos données.

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1
0.6642336 0.3357664

Conditional probabilities:
   pregnant
Y        [,1]      [,2]
  0 0.1557854 0.1459002
  1 0.2851662 0.2402413

   glucose
Y        [,1]      [,2]
  0 0.3179998 0.1824807
  1 0.5695989 0.2176724

   pressure
Y        [,1]      [,2]
  0 0.5123307 0.1350414
  1 0.5782356 0.1587530

   triceps
Y        [,1]      [,2]
  0 0.4043508 0.2091470
  1 0.5306122 0.2003928

   insulin
Y        [,1]      [,2]
  0 0.1370647 0.1254769
  1 0.2384756 0.1655330

   mass
Y        [,1]      [,2]
  0 0.2754556 0.1359754
  1 0.3676536 0.1425734

   pedigree
Y        [,1]      [,2]
  0 0.4862253 0.3172108
  1 0.6092500 0.3552770

   age
Y        [,1]      [,2]
  0 0.1190476 0.1500066
  1 0.2675725 0.1823963
```

33% observations appartient à la catégorie où la classe "diabetes "est 0.

FIGURE 3.12 – Tableaux des probabilités conditionnée par l'attribut catégorique "diabetes"

Après avoir créé le modèle naïf Bayésien, nous pouvons utiliser la fonction de prédiction universelle pour créer une prédiction, ensuite pour aboutir la matrice de confusion.
Outils d'évaluation :

	0	1
0	62	12
1	18	26

TABLE 3.5 – Matrice de confusion de naive bayes.

$$Sensibilité = 0.684$$
$$Spécificité = 0.775$$
$$Arruracy = 0.746$$

Nous sommes en mesure de classer correctement 62 cas sur 80 "non" et 26 cas sur 38 "oui" correctement. Cela signifie que la capacité de l'algorithme Naive Bayes à prédire les cas "non" est d'environ 77 % mais ne tombe que dans 68% des cas oui ce qui donne une précision globale de 74,6%.

3.3.5 L'application de K plus proche voisin

Cette méthode est utilisée pour effectuer la classification du diabète telle que L'algorithme K-NN est basé sur la similarité des caractéristiques c'est-à-dire la mesure dans laquelle les caractéristiques hors échantillon ressemblent à notre ensemble d'apprentissage détermine la manière dont nous classons un point de données donné à l'aide d'une distance.

Tout d'abord, nous devons définir les variables catégorielles et choisir les étiquettes

appropriées pour les classer correctement avec la normalisation des données. Pour choisir le nombre approprié de k et pour obtenir le modèle k-NN la commande la plus largement utilisée est "kknn".

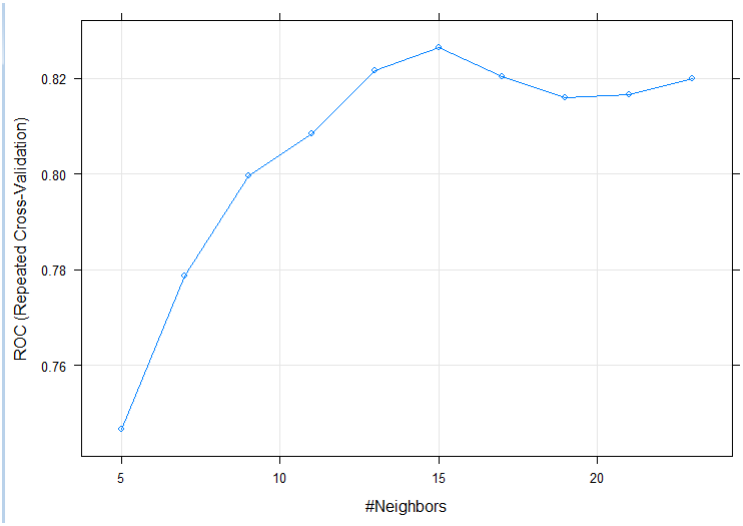

FIGURE 3.13 – Nombre optimale de k.

En k = 15, la précision maximale obtenue est de 75%. Après cela, il semble que K augmente, le classement augmente mais réduit le taux de réussite. En addition, le package "caret" permet de prédire le résultat. En utilisant la matrice de confusion, nous pouvons

	0	1
0	64	16
1	13	25

TABLE 3.6 – Matrice de confusion de K-NN.

obtenir des statistiques de nos résultats.

$$Sensibilité = 0.61$$
$$Spécificité = 0.83$$
$$Arruracy = 0.754$$

Le tableau suivant résume les outils d'évaluation des algorithme d'apprentissage automatique : Par l'utilisation de différents algorithmes d'apprentissage automatique, on remarque a partir de le tableau précèdent que la méthode de régression logistique possède un taux de précision plus élevée alors, on peut dire que cet algorithme permet de prédire efficacement le diabète de type 2.

	Accuracy %	Sensibilité %	Taux de Spécificité %
Régression logistique	**0.788**	0.526	**0.912**
Arbre de décision	0.779	**0.702**	0.814
K-plus proche voisin	0.754	0.609	0.83
Naïve bayésien	0.746	0.684	0.775
Réseau de neurone	0.661	0.684	0.65

TABLE 3.7 – Comparaison entre les algorithmes en utilisant les taux d'évaluation.

3.4 Discussion

Aujourd'hui, l'apprentissage gagne en popularité pour les scientifiques de données avec les applications d'apprentissage automatique sont hautement automatisées et se modifient d'elles-mêmes. Elles continuent de s'améliorer avec le temps, avec une intervention humaine minimale, à mesure qu'elles apprennent avec davantage de données. Afin de montrer les résultats obtenus pour les trois modèles, on illustre dans ce qui suit les résultats en termes de la comparaison entre les courbes de chaque résultat qui nous permettra de comparer trace le taux de vrais positifs en fonction du taux de faux négatifs. Par la suite, la meilleure courbe nous servira à déterminer l'algorithme le plus performant.

Cas 1 :

La courbe de ROC est un outil d'évaluation et de comparaison de de modèles efficace et robuste,avec l'intérêt d'établir visuellement la pertinence d'un modèle.

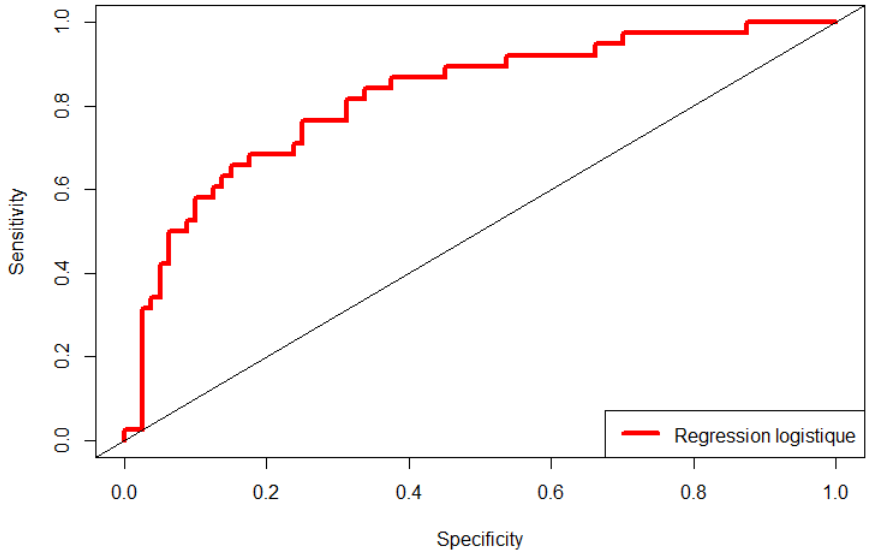

FIGURE 3.14 – représentation de la courbe ROC de la régression logistique.

Le graphique montre que notre modèle a une courbe ROC dont l'AUC est supérieure à 45 degrés tel que l'AUC mesure la qualité du classement des prédictions, plutôt que

leurs valeurs absolues.On peut dire que nous avons créé un modèle qui fait mieux que la prédiction aléatoire avec le taux de précision égale à 79 %.

Cas 2 :

La superposition des courbes permet d'évaluer rapidement et facilement le meilleur modèle.

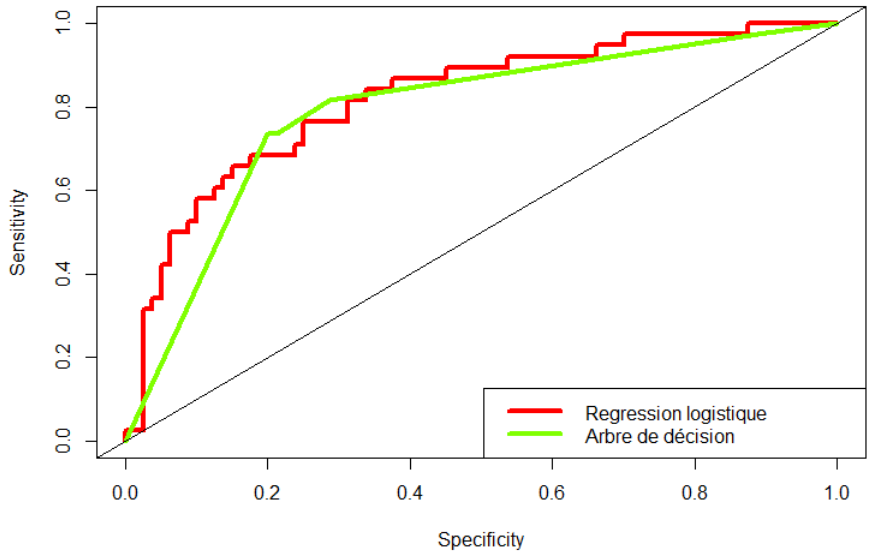

FIGURE 3.15 – Comparaison graphique entre le l'algorithme régression logistique et l'arbre de décision.

On peut dire que nous avons créé un modèle qui fait mieux que la prédiction aléatoire avec les taux de précision égale à 77%. Mais, on peut observer que les courbe correspondant au modèle 2 (arbre de décision) est au dessus de celle du modéle 1(régression logistique), donc nous remarquons qu'à nouveau la meilleure celui de la régression logistique qui est de 79 %.

Cas 3 :

Le calcul d'aire sous les courbes permet d'évaluer numériquement et plus précisément les modèles avec le taux de précision.

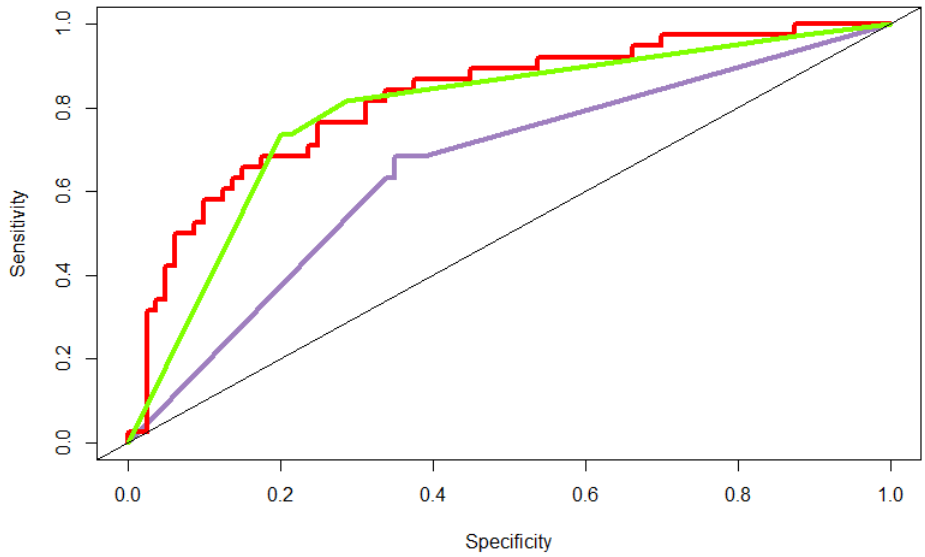

FIGURE 3.16 – représentation de la courbe ROC de trois méthodes.

Le graphe représente la spécificité en fonction de la sensibilité entres des deux précédant algorithmes et celui de réseau de neurone. Comme les précédentes comparaisons celui de régression logistique garde le meilleur taux de précisions et meilleur AUC c'est-à-dire les courbe ROC de deux dernière algorithme sont au dessus de la courbe de le premier modèle (régression logistique) .

Cas 4 :

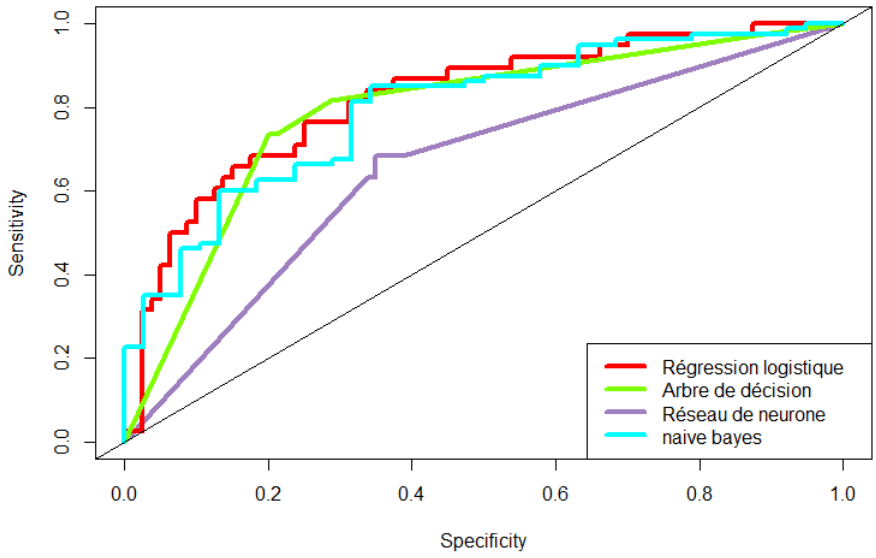

FIGURE 3.17 – représentation de la courbe ROC de quatre algorithmes.

Le graphe représente le taux de vrai négatifs en fonction du taux du taux vrai positifs entres des quatre précédant algorithmes et celui de Naïves Bayes. Contrairement a l'algorithme de réseau de neurone, l'algorithme Naïves Bayes reste le meilleur algorithme avec

un taux de spécificité plus élevée par rapport au réseau de neurone.

Cas 5 :

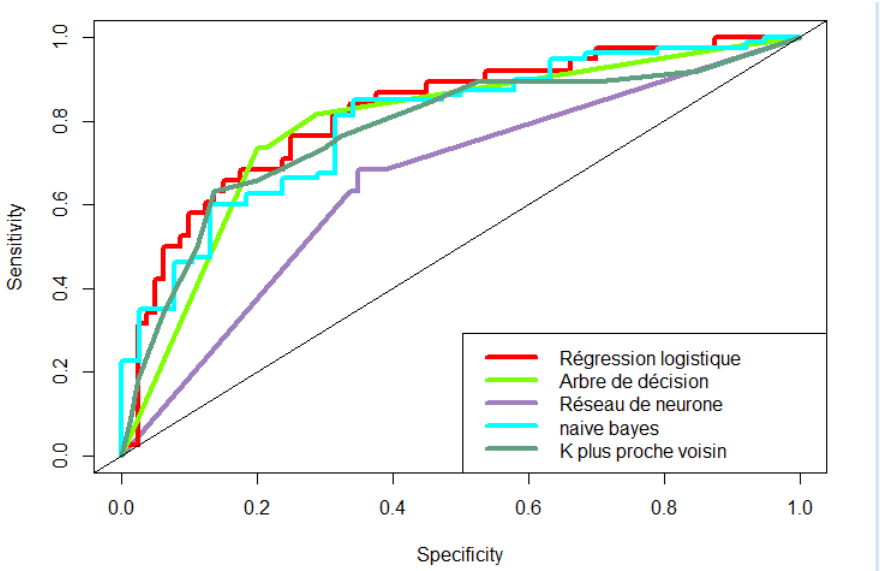

FIGURE 3.18 – représentation de la courbe ROC de nos algorithmes.

Le graphe représente le graphe ROC entre les cinq méthodes d'apprentissage automatique : la régression logistique ,l'arbre de décision, réseau de neurone, classification bayésienne et K-plus proche voisin tel que l'algorithme régression logistique ayant encore le meilleur AUC et le meilleur taus de précision contrairement à l'algorithme réseau de neurone avec un taux de précision de 75%.

3.5 Conclusion

Dans ce chapitre, nous avons vu les différents types de diabète. Même s'il existe des méthodes de prévention qui permettent de réduire le risque d'avoir le diabète, parfois il est impossible de l'éviter comme pour le diabète de type 1. Dans ces cas là, la seule solution est de pouvoir le diagnostiquer très tôt et faire tout son possible pour combattre les complications. Dans ce mémoire, nous présentons des approches différentes d'aide à la prise de décision en utilisant la classification de l'apprentissage dans la prédiction du diabète de type 2. Ensuite, nous avons proposé une approche de comparaison entre différents algorithmes de l'apprentissage automatique qui est une amélioration comparée aux études de l'état de l'art. Cette amélioration se résume dans l'évaluation des performances. Il s'est avéré que l'algorithme "Régression logistique" a le meilleur taux de précision et le meilleur "AUC". Les médecins, les patients et les proches du patient peuvent bénéficier des résultats de la prédiction. Dans les milieux cliniques à faibles ressources, il est nécessaire de prédire l'état du patient après l'admission pour répartir les ressources de manière appropriée.

Au future, l'objectif principal est d'appliquer ces différents algorithmes de classification sur des données concernant le risques de développer un diabète de type 2 récoltés au niveau

de deux infrastructures différentes dans le but d'aboutir à un résultat d'aide à la prédiction de ce type de diabète.

CONCLUSION GÉNÉRALE

Le travail présenté dans ce mémoire concerne les systèmes d'aide à la décision dans le secteur médicale dont le but de modéliser la qualité des données et prendre des décisions fondées sur les algorithmes d'apprentissage automatique.

L'apprentissage automatique est un domaine en constante innovation qui facilite l'utilisation des ordinateurs dans la construction de modèles à partir des données d'échantillonnage afin d'automatiser les processus de prise de décision en fonction des données saisies. Donc il est important de garder à l'esprit que les algorithmes, les méthodes et les approches continueront de changer.

Nous avons présenté dans ce travail une approche numérique des systèmes d'aides à la décision médicale basée sur les algorithmes K-plus proche voisin (K-ppv), réseau de neurone, régression logistique, classification bayésienne, arbre de décision au diagnostic médicale, ces algorithmes puissent d'apprendre en comparant sa sortie prédite avec les sorties observée pour trouver des erreurs et modifier le modèle en conséquence.

D'autre part, nous avons mené une étude critique sur la prédiction de la base du diabète type 2 tel que nous avons fait une comparaison entre les différentes méthodes dans laquelle nous avons repris l'essentiel des avantages et inconvénients des méthodes proposées. Pour finir, nous avons présenté une méthode pour la prédiction du diabète basée sur l'application des algorithmes d'apprentissage automatique supervisé. Nous avons testé les algorithmes d'apprentissage supervisé sur la base de données de diabètes de type 2. Par la suite, nous avons amélioré le meilleur algorithme d'apprentissage à savoir régression logistique en termes de taux de précision, temps d'exécution. Le taux de classification obtenu avec notre méthode est parmi les meilleurs résultats obtenus pour la classification du base diabète de type 2, par rapport aux autres algorithmes.

En termes de perspectives, la prédiction du diabète à l'aide des méthodes d'apprentissage peut être élargie en utilisant les méthodes de base de connaissance pour augmenter l'interopérabilité du diagnostic.

BIBLIOGRAPHIE

[1] https ://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[2] K. Polat, S. Sahan, H. Kodaz, and S. Günes,*A New Classification Method for Breast Cancer Diagnosis : Feature Selection Artificial Immune Recognition System (FS-AIRS) ICNC 2005, LNCS 3611.* Berlin Heidelberg : Springer-Verlag, 2005, pp. 830 – 838.

[3] Béchir Dola, *World Health Organization WHO, 'Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications'. Report of a WHO Consultation. Part 1 : Diagnosis and Classification.* Geneva, 1999.

[4] Mathieu FEUILLOY *Étude d'algorithmes d'apprentissage artificiel pour la prédiction de la syncope chez l'homme.*

[5] Projet de fin d'étude *L'aide de l'intelligence artificielle à la prise de décision.*

[6] Tom Mitchell *Machine Learning, McGraw Hill*, 1997.

[7] Yufei LUO,*Amélioration de la modélisation de sinistres graves à l'aide d'une approche d'apprentissage.*

[8] Frédéric Santos,*Arbres de décision.*CNRS, UMR 5199 PACEA.

[9] Ameni Bouaziz, thése de doctorat,*"Méthodes d'apprentissage interactif pour la classification des messages courts".*

[10] Rémi Lajugie,*Faire apprendre à des machines : les plus proches voisins*

[11] thèse de doctorat, Aicha Guelli, *la médecine a l'ère de l'intelligence artificielle.*

[12] Projet de fin d'étude,LABIAD ALI, *sélection des mots clés basée sur la classification et l'extraction des règles d'association.*