République Algérienne Démocratique
et Populaire Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique Université de Bordj Bou Arreridj
Faculté des mathématiques et d'informatiques
Département Informatique

MÉMOIRE

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

Spécialité : Ingénierie de l'Informatique Décisionnelle.

Thème :

Classification du diabète avec l'algorithme KNN

Encadré par :

Belazzoug .M

Réalisé par :

- **Zouache hanen**
- **Bendib ichrak**

Promotion : 2020/ 2021

Remerciements

Nous remercions **le DIEU** de nous avoir donné la patience, la santé et le courage pour réaliser ce travail.

A travers ce modeste travail, nous tenons à remercier vivement notre encadreur **Belazzoug Mouhoub** pour ses conseils, ses encouragements, sa gentillesse, son suivi permanent qui nous ont permis de réaliser ce travail dans les meilleures conditions.

Nous remercions sincèrement les membres du jury d'avoir accepté d'examiner et d'évaluer notre travail.

Nous exprimons également notre gratitude à tous les professeurs et les enseignants de département informatique.

Sans oublier bien sûr de remercier profondément tous ceux qui ont contribué de près ou de loin à la réalisation du présent travail.

Merci à tous

Résumé

Le diabète est l'un des principaux problèmes de santé mondiaux. Selon le rapport de l'OMS 2011, environ 346 millions de personnes dans le monde souffrent de diabète sucré. Le diabète sucré est une maladie métabolique dans laquelle une mauvaise gestion de la glycémie entraîne le risque de nombreuses maladies comme la crise cardiaque, la maladie rénale et l'insuffisance rénale. Dans ce mémoire, nous diagnostiquons le diabète sucré à l'aide de l'algorithme du voisin le plus proche, dans lequel plusieurs paramètres ont été testés tels que le nombre de voisin (k) ainsi que les mesures de distances ou de similarités. Dans l'autre côté, une étude comparative est accompagnée de différents algorithmes de classification d'apprentissage supervisé sur les données de 'Pima indian diabetes database. Les résultats de performances ont montré clairement l'avance de l'algorithme KNN contre tous les autres algorithmes choisis dans cette étude.

Abstract

Diabetes is one of the world's major health problems. According to the WHO 2011 report, around 346 million people worldwide have diabetes mellitus. Diabetes mellitus is a metabolic disease in which poor blood sugar management puts you at risk for many diseases like heart attack, kidney disease, and kidney failure. In this Master dissertation, we diagnose diabetes mellitus using the nearest neighbour algorithm, in which several parameters were tested such as neighbour number (k) as well as distance or similarity measurements. On the other hand, a comparative study is accompanied by different supervised learning classification algorithms on data from the 'Pima indian diabetes' database. The performance results obviously showed the advance of the KNN algorithm against all other algorithms chosen in this study.

ملخص

مرض السكري هو أحد المشاكل الصحية العالمية الكبرى. حسب تقرير منظمة الصحة العالمية لعام 2011، يعاني حوالي 346 مليون شخص في جميع أنحاء العالم من مرض السكري. داء ا لسكري هو عبارة عن اضطراب في هرمون الأنسولين الذي ينتجه البنكرياس لمساعدة الجسم في استخدام السكر والدهون وتخزين بعضها مما يؤدي الى الإصابة بالعديد من الأمراض مثل النوبات القلبية وأمراض الكلى والفشل الكلوي .يمكن أن تكون تقنيات التعلم الآلي (MLTs) حلا للتشخيص المبكر والتنبؤ بمرض السكري. يعد التعلم الآلي (machine Learning)جانبًا من جوانب الذكاء الاصطناعي بحيث يتم استخدامه للتنبؤ بالمرض في المراحل المبكرة. في هذه المذكرة سوف نقوم بتشخيص داء السكري باستخدام خوارزمية أقرب الجيرانKNN ، حيث تم اختبار العديد من المتغيرات مثل عدد الجار (k)وقياس المسافات. و في

Résumé

الجانب الآخر ، ستكون الدراسة للمقارنة بين خوارزميات التصنيف (Support Vector ،Random Forest,KNN Machine، Naïves Bayes) على قاعدة بيانات **Pima indian diabète** ، واستنتاج افضل خوارزمية الي تفي بهذا الغرض اين أظهرت نتائج الدراسة بوضوح ان خوارزمية KNN هي الاحسن مقابل جميع الخوارزميات الأخرى المختارة في هذه الدراسة.

Table de matière

Liste de figures

Introduction générale

• Contexte générale et problématique

L'intelligence artificielle est une science qui vise à simuler et reproduire des comportements intelligents voire complexes chez l'être humain par des machines ou des ordinateurs en général. Elle repose sur plusieurs disciplines, à savoir le l'informatique, analyse de données, les ontologies, la fouille de données…etc. L'apprentissage automatique est une discipline prometteuse parmi celles de l'intelligence artificielle dans laquelle on cherche de trouver un moyen de créer des programmes informatiques qui s'améliorent automatiquement avec l'expérience et cela sans réécriture du code source du programme. L'application de l'IA en médecine offre à la machine une option pour prédire et également analyser les données médicales. La prédiction automatique de la maladie permettra ensuite aux médecins d'intervenir sur les patients, ce qui évitera les complications de la maladie et même de sauver la vie des patients.

• Position du problème

La maladie du diabète est un véritable problème de santé publique dans le monde. Près de 366 millions de diabétiques dans le monde en 2011 et les prévisions pour 2030 sont d'environ 552 millions selon la FID (Fédération Internationale du Diabète). Les systèmes de classification sont d'une grande aide car ils réduisent les erreurs dues à la fatigue et au temps nécessaire au diagnostic. L'utilisation d'un processus automatique de classification devient de plus en plus fréquente pour la fiabilité ainsi que l'efficacité du diagnostic de diabète. Dans cette optique, nous allons présenter un système de classification automatique de diabète.

• Contribution

A travers ce présent mémoire de Master, nous nous concentrerons sur l'utilisation d'algorithmes d'apprentissage automatique et notamment l'algorithme KNN pour la prédiction du diabète afin de réduire tout sort de risque de complications de cette maladie. Pour enrichir et donner plus de crédibilité à notre travail nous appliquerons d'autres algorithmes de classification d'apprentissage supervisé tels que : les Arbres de décision, les forêts aléatoires, la méthode de machine à vecteurs de support, et l'algorithme Naïves Bayes sur la base de données 'Pima Indian Diabète Database '. Les résultats de performance sont exprimés quantitativement en termes de Précision, Rappel et F-mesure.

- **organisation du manuscrit**

Ce travail est organisé en trois principaux chapitres comme suit :

Le 1er chapitre présente un aperçu général sur la maladie du diabète, leur différent types, les symptômes ainsi que le diagnostic et le traitement de la maladie et à la fin quelques préventions pour éviter le diabète.

Le 2ème chapitre présente les notions générales de l'apprentissage automatique. Les domaines d'application et les algorithmes d'apprentissage sont également exprimés dans cette partie. Ensuite, nous introduisons un état de l'art sur l'application des algorithmes de classification de diabète.

Le 3ème chapitre présente d'abord une étude technique dans laquelle nous définissons l'environnement logiciel utilisé pour la partie expérimentale de notre projet. Pour la partie réalisation, la base de données utilisée '**Pima indian diabetes database'** a était choisie pour tester les performances des classificateurs utilisés dans cette étude, tel que KNN, SVM et DT.

A la fin, nous terminerons ce travail par une conclusion générale et quelques perspectives.

I. Le diabète

Chapitre 1 : Le diabète

1.1 Introduction

Le diabète est considéré comme l'une des causes très fréquentes de morbidité et de mortalité dans le monde.

Le diabète est un problème de santé publique mondiale. En France, le nombre de diabétiques est environ 3,5 %. En Europe, le nombre de diabétiques est évalué à 30 millions, et en Etats-Unis, il y a 15 millions de diabétiques. Dans le monde entier, on dénombre 100 millions de diabétiques [1]

En fait, Le diabète se divise en deux formes différentes :

Le diabète insulinodépendant (type 1), ou le diabète juvénile, qui survient le plus souvent avant l'âge de 20 ans et représente 10 à 15 % des diabètes et le diabète non insulinodépendant (type 2), qui survient le plus souvent après l'âge de 50 ans et représente 85 à 90 % des diabètes. [1]

Le diabète non insulinodépendant est un véritable problème de santé publique. Sa prévalence augmente parallèlement avec l'augmentation de l'âge, de la concentration de la population dans les villes. De la sédentarisation et aussi l'augmentation de l'obésité dans les populations des pays industrialisés. Le diabète type 2 touche aussi les pays sous-développés où il atteint parfois une prévalence de 20 à 30 %.[1]

Le coût de la prise en charge de diabète est très élevé à cause du taux élevé de de la prise en charge de ses complications dégénératives (par exemple : le dialyse, l'amputation du pied ou jambe.)

Dans ce chapitre, on va donner une représentation générale sur le diabète, son diagnostique, ses types et ses majores complications.

1.2 Définitions

«Le diabète sucré est un trouble du métabolisme hydrocarboné lié soit à un déficit d'insuline, soit à une résistance anormale à cette hormone, d'où une accumulation de glucose dans les tissus». *Garnier Delamare* [2].

Le diabète est une maladie chronique caractérisée par une élévation chronique et permanente de la glycémie sanguine.

1.3 Rappel physiopathologique

De diabète s'accompagne d'une glycémie à jeun élevé supérieure ou égale à 1,26g/l (7mmol/L) ou 2g/l (11mmol/L) après absorption de 75g de glucose. Il existe deux variétés de diabètes, le type 1

caractérisé par un défaut de sécrétion d'insuline par le pancréas. C'est une forme rare qui touche les enfants et les jeunes adulte. Il est déclenché le plus souvent par une agression auto immune détruisant les îlots de Langerhans.il exige toujours un traitement insulinique.

Le type 2 est dû à une résistance des tissus à l'action de l'insuline. Il est plus fréquent à la maturité. Il existe un substrat génétique affectant l'action de l'insuline au niveau de ses récepteurs. Il existe des autres facteurs d'environnement aggravants ce trouble métabolique par exemple : le vieillissement tissulaire, de mauvaises habitudes alimentaires, la sédentarité, l'obésité viscérale. . Son traitement repose sur une bonne hygiène de vie entrainant une perte pondérale et des médicaments agissant à divers niveaux du métabolisme du glucose. [3]

Le diabète est la source de multiples complications qui font toute la gravité et le coût très élevé de la prise en charge de cette pathologie.

1.4 Données épidémiologiques

1.4.1 Nombre de diabétiques

87 à 91 % des diabétiques sont des adultes de type 2, non insulinodépendants 463millions de sujets âgés de 19 à 79 ans sont atteints de diabète dans le monde, le plus grand nombre dans les trois régions les plus peuplées, le Pacifique occidental incluant la Chine, l'Asie du Sud Est avec l'Inde et l'Europe.79,4 % vivent dans des pays à faible ou moyen revenus. Les zones urbaines sont plus affectées (279 millions) que les zones rurales (146 millions).Un tiers de cette population diabétique (135,6 millions) a plus de 65ans. En 2019, 231,9 millions de personnes, soit la moitié de la population diabétique, méconnaissaient leur maladie principalement dans les pays les plus pauvres en raison du peu de moyens affectés au dépistage.9 à 13 % des diabétiques sont des enfants et des adolescents.1.110000 sujets de moins de 19 ans présentent très majoritairement un diabète de type 1insulinodépendant.600.900 ont moins de 14 ans vivant majoritairement en Europe (162600) et en Amérique du Nord (121400). L'incidence du diabète de type 1 augmente régulièrement (3 %/ an) chez des enfants de plus en plus jeunes, avec des taux d'incidence pour 100.000 habitants particulièrement élevés en Finlande(62,3), Suède(43,2) et au Koweït(41,7).Cette situation évoque le rôle de facteurs d'environnement dans cette progression mais aucun n'a pu, à ce jour, être confirmé.

1.4.2 Prévalence du diabète, du pré-diabète et de l'hyperglycémie gravidique

Prévalence du diabète : 9,3 %de la population adulte mondiale est atteinte de diabète (9 % pour les femmes-9,6% pour les hommes avec une progression en fonction de l'âge (18,8% après 65 ans)et de fortes disparités régionales. Les taux de prévalence les plus élevés concernent l'Amérique

du nord (13,3%), la région MENA(12,8%) et le Pacifique occidental(9,6%), mais plusieurs territoires et régions insulaires dépassent largement ces chiffres. Chez les plus de 65 ans, l'Amérique du Nord présente la prévalence la plus élevée (27,7 %), suivie par la région MENA (24,2 %), l'Amérique du Sud (22,7%) et l'Europe (20,1 %), l'Afrique offrant la plus faible (8,4%).

Prévalence du pré-diabète : 373,9millions de personnes âgées de 20 à 79 ans présentent un pré-diabète (7,5%) avec une égale répartition hommes –femmes. Les pays les plus affectés sont les plus peuplés: Chine (54,5 millions), USA (37,4 millions), Indonésie (29,1 millions), Inde (25,2 millions). La région Amérique du Nord-Caraïbes présente la plus forte prévalence (12,3 %) contre l'Europe qui a la plus faible (4,4 %). La prévalence s'accroit avec l'âge (12-15% après 65 ans). Parmi les sujets affectés la moitié a moins de 50 ans et un tiers se situe dans la tranche d'âge 20-39 ans, les plus susceptibles d'évoluer vers un diabète caractérisé.

Prévalence de l'hyperglycémie gravidique : 20,4 millions d'enfants (15,8 %)recensés en 2019 sont nés d'une mère affectée par une élévation anormale de la glycémie pendant sa grossesse, diabète gestationnel (83,6 %) ou diabète authentique (16,4%). 88 % des cas concernent les pays les plus pauvres. L'Asie du Sud Est enregistré la plus forte prévalence (27 %). Cette dernière augmente très fortement avec l'âge atteignant 37 % chez les femmes de 45 à 49 ans. L'hyperglycémie gravidique précède parfois de plusieurs années la survenue d'un diabète authentique tant chez la mère que chez les enfants exposés in utero à ce désordre métabolique. [3]

1.5 Facteurs de risque

1.5.1 Facteurs de risque pour le diabète de type 1

Il y a plusieurs facteurs de risque de **DID** :

- les antécédents familiaux. Le facteur héréditaire est incriminé dans le développement de ce type de diabète

- les conditions environnementales. Par exemple : les infections virales,

- l'auto-immunité, c.-à-d. la présence des auto-anticorps qui attaquent les cellules de pancréas.

- la géographie .Par exemple, les gens dans certains pays, tels que la Finlande et la Suède, ont une présence beaucoup plus forte de DID. [4]

1.5.2 Facteurs de risque pour le diabète de type 2

Il y a plusieurs facteurs de risque pour le DNID :

- obésité est le facteur le plus fréquent . le tissu graisseux augmente la résistance des cellules à l'insuline.
- la sédentarité : peu d'exercices phasiques, longues durées de sédentarité .
- l'obésité d'enfance augmente significativement le risque de diabète de type 2.
- les antécédents familiaux.
- Les habitudes alimentaires.
- certaines populations et races se caractérisent par un développement naturel de diabète type2. Par exemple, les Asiatique-Américains, les hispaniques et les personnes de race noire sont jusqu'à 4 fois plus de périodes en danger d'avoir le diabète. [4]

1.5.3 Facteur de risque de diabète gestationnel

Quelque facteurs sont associés à son apparition tels que:

• ethnie non-caucasienne

• obésité

• âge > 30 ans

• anamnèse familiale de diabète de type 2 positive

• femme ayant déjà accouché d'un nouveau-né de plus de 4kg.

Ces patientes nécessitent un traitement d'insuline ainsi qu'une surveillance étroite de leur glycémie durant la grossesse et en post-partum. Le nouveau-né sera aussi encadré étroitement par une équipe multidisciplinaire. [4]

D'autres conditions peuvent augmenter le risque de diabète. Par exemple, personnes qui ont précédemment eu une crise cardiaque ou une rappe. Supplémentaire, le schizophrène, bipolaires et des personnes avec la dépression sont beaucoup pour obtenir le diabète, ainsi que les femmes avec le syndrome ovarien poly kystiques. [5]

1.6 Diagnostique de diabète

1.6.1 Objectifs

Bien définir les critères biologiques du diabète a deux objectifs:

A. à court terme : atteindre l'euglycimie pour lutter contre les signes cliniques associés à l'hyperglycémie (perte de poids, syndrome polyuro-polydipsique).

B. à long terme : déceler les personnes qui ont un risque de complications dégénératives. La prise en charge des complications comme celles de la rétinopathie, la néphropathie et neuropathie diabétique.

1.6.2 Critères de diagnostique

Le diagnostique de diabète est purement biologique ,repose sur la présence d'au moins 2 taux d'hyperglycémie dans 2 reprises déférentes.

Le diagnostique est confirmé dans l'une des 3 situations suivantes :

* la glycémie à jeun est supérieure à 1,26 g/l (7 mmol/l).

* la glycémie, à 2h, sous HGPO est supérieure à 2 g/l.

* une glycémie, dans n'importe quel moment dans le jour, est supérieure à 2 g/l (11,1 mmol/l) avec la présence à des signes cliniques évocateurs. [6]

1.7 Classification de diabète

Il existe trois types de diabète : le diabète de type 1, le diabète de type 2 et le diabète gestationnel.

1.7.1 Le diabète de type 1

Le diabète de type 1 est une maladie auto-immune, le pancréas produit une quantité insuffisante d'insuline pour réguler la glycémie.

Ce type de diabète se développe le plus souvent pendant l'enfance ou l'adolescence, mais peut aussi survenir chez l'adulte. [7]

Les cellules du pancréas, qui fabriquent l'insuline, sont détruites par des anticorps cibles

fabriqués par le corps humain. Le pancréas, ayant perdu ses cellules, ne peut plus produire d'insuline.

Après chaque alimentation, la glycémie s'augmente, le pancréas sécrète de l'insuline. Grâce à cette hormone, la glycémie diminuée car le glucose présent dans le sang entre dans les cellules, pour être stocké ou être utilisé pour donner l'énergie. S'il y a un manque d'insuline, la glycémie reste toujours élevée.

1.7.2 Le diabète de type 2

Cette forme de diabète est la dominante. Son diagnostic est plus tardif que dans DID.

Il est appelé « diabète gras » à cause de l'obésité des malades et ses surpoids. on a 2 types d'anomalies pour le DNID :

***L'insulino-résistance** : ou il ya des moindres effets de l'insuline sur ses tissus cibles tels que le foie et le muscle, ce qui entraîne une résistance de l'organisme à l'action de l'insuline.il y a donc une mauvaise utilisation d'insuline par l'organisme.

***L'insulinopénie** : ou il y a une insuffisance de production d'insuline par le pancréas.

1.7.3 Diabète gestationnel

Est un trouble de la tolérance au glucose qui entraine une hyperglycémie , qui est découverte pour la première fois pendant la grossesse et apres le 28 ieme SA.
Après l'accouchement, la glycémie peut redevenir normale ou le diabète peut persister.
Il peut également récidiver à chaque grossesse ou même en dehors de toute grossesse. C'est pour cette raison que la glycémie doit être contrôlée 3 mois après l'accouchement, puis annuellement et à chaque nouvelle grossesse. Le diabète gestationnel concerne moins d'1 femme sur 10. [7]

1.7.4 Autres diabètes

Les autres formes sont plus rares. On distingue par exemple:

• **les MODY (Maturity on set Diabetes in the Young),** en général non insulino-dépendants sont fortement déterminés par une composante génétique

• **les diabètes secondaires à d'autres maladies** telles que des maladies pancréatiques, endocriniennes ou hépatiques. L'hémochromatose ou certaines mutations de l'ADN mitochondrial.

• **le diabète lipoatrophie** (disparition du tissu adipeux, hyperlipidémie, stéatose hépatique, insulino-résistance majeure).

• **le diabète induit par des traitements médicamenteux** (ex: corticoïdes, diurétiques, neuroleptiques, certains immunosuppresseurs...) [8]

1.8 Les Complications de diabète

1.8.1 Complication métaboliques

1.8.1.1 « Coma » céto-acidosique

La définition du « coma » céto-acidosique est la suivante :

- Acétonurie ? 2 + ;
- glycosurie > 2 + ;
- Glycémie ? 2,5 g/L ;
- PH veineux < 7,25 ;
- Bicarbonate < 15 mEq/L.

Il s'agit d'un coma vrai, au sens nosologique du terme, rare : inférieur à 10 %.
L'incidence est de 2 à 4 % par an et par patient.

- **Étiologie**

Il peut s'agir :

✓ D'un déficit absolu en insuline, inaugural dans le diabète de type 1 (10 % des cas) ou d'un arrêt, volontaire ou non, de l'insulinothérapie.

✓ D'un déficit relatif en insuline : association d'un diabète non obligatoirement insulinodépendant et d'un facteur surajouté (infarctus, infection, corticothérapie).
L'étiologie est inconnue dans 25 % des cas

- **Diagnostic et évolution**

Phase de cétose

Un syndrome cardinal aggravé est observé, associé à des troubles digestifs (nausées, vomissements, douleurs abdominales).

Phase de céto-acidose

Elle est caractérisée par une dyspnée de Kussmaul associée à des troubles de la conscience (état stuporeux) et à une déshydratation mixte à prédominance extracellulaire. Un diagnostic rapide peut être établi par bandelettes et pH veineux et artériel (GDS).

L'ionogramme en urgence est réalisé pour le dosage de la kaliémie. En cas d'absence d'urine, les nouvelles bandelettes pour le dosage des corps cétoniques sanguins peuvent être utiles.

- **Critères de gravité**

Les critères de gravité imposant l'hospitalisation en réanimation sont les suivants :

- Sujet âgé ;
- pH < 7 ;
- kaliémie < 4 ou > 6 mmol/L ;
- coma profond ;
- instabilité tensionnelle ;
- non-reprise de diurèse après 3 heures ;
- vomissements incoercibles.

- **Diagnostic différentiel** Le diagnostic différentiel s'établit selon:
  - ✓ l'urgence abdominale (augmentation physiologique des enzymes) ;
  - ✓ le coma hyperosmolaire (calcul de la natrémie corrigée).

- **Évolution**

  On note l'évolution suivante :
- régression sous traitement en 24 à 48 h ;
- complication iatrogène : œdème cérébral, surcharge hydrosodée.

- **Traitement**

  a. Traitement préventif

  Le traitement préventif consiste à établir des règles éducatives en cas de cétose (maintien des injections même si inappétence, supplément en insuline rapide, acétonurie systématique si glycémie > 2,5 g/L).

  b. Premiers gestes

  Ils concernent le scope et la surveillance sang-urine.

Les gestes non systématiques concernent la sonde gastrique (sauf si vomissement), la sonde urinaire (sauf si absence de diurèse après 3 heures), le bilan infectieux et les enzymes (sauf orientation), et le cathéter central (sauf si désordre majeur).

C. traitement curatif

Le traitement curatif requiert :

- L'insuline rapide ou ultrarapide à la seringue électrique IV en débit constant, tant que dure la cétose (10 à 15 unités/heure) ; la recharge volumique par sérum salé isotonique, 4 à 7 L au mieux au perfuseur électrique ;

- les apports potassiques importants, si possible à la seringue électrique, à ajuster à la kaliémie répétée ;

- les apports glucosés intraveineux à la demande (G 10 %) pour maintenir la glycémie à 2,5 g/L ;

- le traitement du facteur déclenchant éventuel.

1.8.1.2 Coma hyperosmolaire

Il s'agit de la décompensation classique du sujet âgé diabétique de type 2, ou inaugurale du diabète, lorsque la polyurie a été compensée par des boissons sucrées, ou insuffisamment compensée (rôle de l'inaccessibilité aux boissons).Ce coma induit 20 à 40 % de mortalité chez le sujet âgé.

Les signes cliniques sont la déshydratation intense avec des troubles de la vigilance qui sont parfois révélateurs d'un diabète de type 2 méconnu.

- **Diagnostic biologique**

Le diagnostic biologique s'établit selon les critères suivants :

- glycémie > 6 g/L ;
- osmolarité > 350 mmol/kg : calculée selon la formule : (Na+ + 13) x 2 + G, où la concentration en sodium Na+ et la glycémie G sont en mmol/L ;
- natrémie corrigée > 155 mmol/L ; calculée selon la formule : Nap + [(G – 1) x 1,6], où Nap représente le sodium plasmatique, et la glycémie G est en g/L ;
- absence de cétose et d'acidose.

- **Étiologie**

Les facteurs de risque sont :

- l'âge > 80 ans ;
- l'infection aiguë
- les diurétiques ;
- la mauvaise accessibilité aux boissons : maisons de retraite, état de démence, etc.
- la corticothérapie.

- **Traitement**

Le traitement concerne :

- une mise en conditions : voie veineuse, éventuellement centrale, prévention des complications de décubitus ;

- une réhydratation prudente et lente, selon le terrain, avec 6 à 10 litres de sérum salé isotonique dans les premières 24 heures : la 1reh : 1 litre, 1 à 4 h : 2 à 3 litres, 4 à 24 h : 4 à 6 litres ;

- l'insulinothérapie intraveineuse continue à la seringue électrique : 2 à 3 unités/h en maintenant la glycémie > 2,5 g/L, selon les glycémies capillaires horaires ;
- la surveillance clinique (conscience, pouls, PA, température, diurèse), et biologique (ionogramme sanguin et créatininémie) ;
- l'héparinothérapie préventive ;

1.8.1.3 Hypoglycémies

L'hypoglycémie est inévitable chez tout diabétique de type 1 « bien équilibré » : 3 à 5 hypoglycémies modérées en moyenne par semaine ;

- nécessité de combattre les fausses croyances : l'hypoglycémie n'est pas mortelle et ne laisse pas de séquelles cérébrales (sauf cas extrêmes et hypoglycémie très profonde et prolongée) ; elle ne participe pas aux complications du diabète, ne déclenche pas d'accident vasculaire ou cardiaque, elle ne provoque pas de rebond d'hyperglycémies et ne fait pas prendre de poids ;
- nécessité de connaître les vrais dangers de l'hypoglycémie : peur n° 1 du diabétique +++ l'incitant à se maintenir en hyperglycémie, déstabilisation du diabète, prudence chez le sujet âgé, danger réel en cas d'alcoolisme concomitant, danger dans certaines situations ou sports à risque ;

- la non-perception et/ou la perception tardive des signes d'hypoglycémie accroît le risque d'hypoglycémie sévère ; les facteurs favorisants sont : les hypoglycémies mineures répétées plus ou moins ignorées (nocturne), la neuropathie végétative (longue durée du diabète) ;
- les causes les plus fréquentes sont : les repas sautés, insuffisants ou retardés, l'effort physique non pris en compte dans les doses d'insuline, l'erreur d'injection d'insuline [9]

1.8.2 Complications dégénérative

Ce sont des complications vasculaires chroniques du diabétique. Elles concernent l'intégralité des vaisseaux de l'organisme. On distingue la macroangiopathie (Atteinte des gros vaisseaux) non spécifique, plus fréquente chez le DST2 et la microangiopathie (Atteinte des artérioles et les capillaires) spécifique du diabète sucré, plus fréquente chez le DST1et qui se manifeste principalement le rein, la rétine et les nerfs. Actuellement la mortalité par infection est très faible dans la population diabétique grâce au progrès de l'antibiothérapie, et les complications aiguës du diabète sucré sont moins fréquentes et moins graves grâce aux programmes d'éducation des diabétiques. Toute la gravité de la maladie diabétique réside donc dans les complications dégénératives en termes de mortalité et de morbidité. L'importance de leur dépistage et de leur prévention se trouve ainsi justifiée et prouvée par de grandes études telles que l'étude de l'UKPDS pour le diabète sucré type 2 et celle de la DCCT pour le diabète sucré type1

UKPDS= United Kingdom Prospective Diabetes Study

DCCT= Diabetes Control Complications Trial

1.8.2.1 La macro-angiopathie

Le diabète sucré est un facteur de risque d'athérome :

- 30 à 40 % de la population diabétique présentent une **HTA**.

- la coronaropathie est plus fréquente et plus précoce.

- le risque de cardiopathie et d'AVC est multiplié par 2 à 4 et 1,5 à 2 respectivement : 75% des diabétiques meurent d'accidents cardiovasculaires, dont 50 % de cardiopathie ischémique. - le risque d'artériopathie oblitérante des membres inférieurs est multiplié par 6 à 10 et le risque de gangrène et d'amputation est multiplié par 20 par rapport à la population générale (50 % des amputations par artérite des membres inférieurs sont diabétiques)

1.8.2.2 La micro-angiopathie

Quelques données épidémiologiques rendent compte de la gravité du diabète sucré :

Espérance de vie : diminuée de 10 à 15 ans chez la femme et de 6 à 9 ans chez l'homme, à cause des problèmes cardiovasculaires essentiellement.

Rétinopathie diabétique (RD): 1ere cause de nouveaux cas de cécité non traumatique de l'adulte de moins de 50 ans. Le risque de cécité est multiplié par 10 par rapport à la population générale. Elle est déjà présente au moment du diagnostic du DST2 dans 20 % des cas. Après 15 ans d'évolution la prévalence de la RD est de 50 à 80 % pour le DST2 et de 80 à 90 % pour le DST1, 12 % des diabétiques type 1 et 7 % des types 2 sont aveugles. L'HTA est un grand facteur de risque de la RD.

Néphropathie diabétique (ND) : elle la seule cause croissante de l'insuffisance terminale (IRC) et tend à être la 1ère cause de néphropathie mortelle, la survie en dialyse est de 2 ans dans moins de 60 % des cas. Sa prévalence est de 30 à 35 % après 15 ans d'évolution. A partir de 5 ans d'évolution, son incidence augmente jusqu'à 10 ans pour diminuer et 50 % des patients décèdent par IRC ou insuffisance cardiaque dans les 20 ans qui suivent l'apparition de la protéinurie.3 facteurs exposent au risque de ND : l'HTA, la durée d'évolution, et l'équilibre glycémique.

1.8.2.3 Autres complications :

- Neuropathie **:** présente dans 20 à 30 % des cas, invalidante dans 2 % des cas, participe à la sur morbidité et la surmortalité.
- Complications périnatales lors d'une grossesse diabétique : prématurité, mort in utero, malformations, hypotrophie, macrosomie ....
- Coût économique élevé **:** Le diabète est une maladie grave qui coûte très chère à la santé, à la sécurité sociale et la famille :
  - Hospitalisations fréquentes et longues
  - Arrêts de travail souvent de longue durée
- Chirurgie.
- Handicap physique (amputation, cécité...etc.)
- Mobilisation du personnel important …….. [10]

1.9 Histoire de diabète

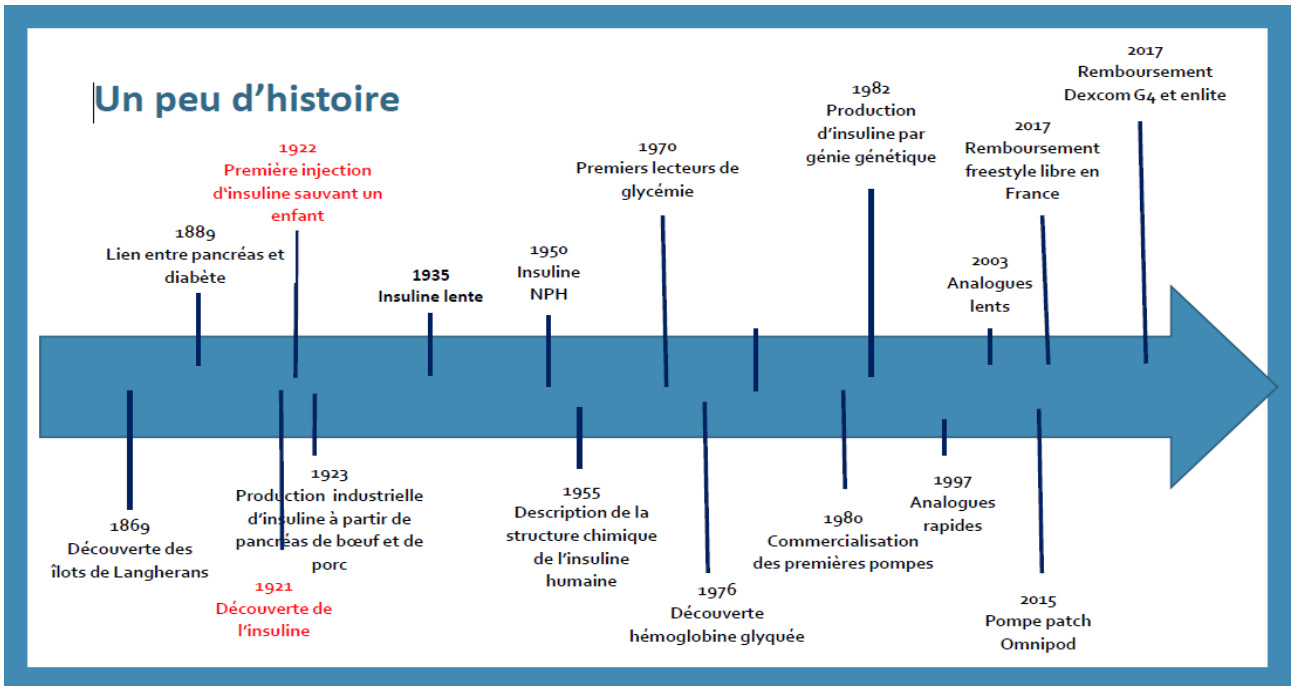

Figure 2-1 histoire de diabète [11]

C'est 4000 ans avant J.C. en Chine que le diabète a été mentionné pour la première fois. On parlait alors d'urine sucrée ou d'urine de miel.

En 1500 avant J.C. un papyrus égyptien, appelé papyrus Ebers, décrit des symptômes similaires à ceux du diabète.

Au fil des siècles et des avancées de la recherche médicale, des expériences ont été réalisées pour comprendre et trouver des traitements au diabète. Voici les étapes clés de la découverte de cette maladie et des progrès réalisés par la recherche. [12]

Quelques dates...

- **1869 :** découverte des îlots de Langerhans par l'étudiant allemand Paul Langerhans.
- **1889 :** lien établi entre le pancréas et le diabète par les Allemands Oskar Minkowski et Josef Von Mering.
- **1921 :** découverte de l'insuline par Frederick Grant Banting et Charles Best.
- **1922 :** première injection d'insuline (extraite du pancréas de porc) sauvant un enfant de 14 ans.

- **1923 :** prix Nobel décerné à Frederick Grant Banting pour cette grande avancée.
  Début de la production industrielle et commercialisation d'insuline par des laboratoires à partir du pancréas de bœuf et de porc.

- **1955 :** Frederick Sanger décrit la structure chimique de l'insuline humaine.

- **1978-1982 :** grâce aux progrès des technologies, l'insuline est produite par génie génétique.

En parallèle les premières pompes à insuline sont commercialisées.

Depuis, la recherche continue ses avancées technologiques dont l'objectif est toujours d'améliorer le quotidien du patient diabétique et de guérir la maladie.

1.10 Conclusion

Dans ce chapitre nous avons présenté la maladie du diabète, leur différents types, le Symptômes ainsi que le diagnostic et le traitement de la maladie, les facteurs de risque, et à la fin nous avons cité quelques complications et histoire de la maladie diabète.

Dans le prochain chapitre, nous présenterons des approches différentes d'aide au diagnostic préventif en utilisant les algorithmes de machine Learning dans la prédiction du diabète.

II. L’apprentissage automatique

Chapitre 2 : apprentissage automatique

2.1 Introduction

L'apprentissage automatique permet aux ordinateurs sans être explicitement programmés d'effectuer des tâches complexes, notamment l'apprentissage à partir des données. Dans ce chapitre, nous nous penchons sur l'apprentissage automatique pour lequel on va introduire ses principaux types ainsi que les algorithmes utilisés dans l'apprentissage automatique. Ensuite nous enchaînons par un état de l'art sur quelques algorithmes de classification appliqués dans la prédiction de diabète. Cette application apport un grand avantage à partir de laquelle on peut réduire les risques de complications de cette maladie sur la sante d'un patient.

2.2 Définition de l'apprentissage automatique

L'apprentissage automatique (en anglais : machine Learning) est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d' apprendre à partir de données, c'est-à-dire résoudre des tâches sans être explicitement programmés pour chacune. L'objectif est de rendre la machine capable de traiter une grande quantité d'information et effectuer des tâches extrêmement complexes afin d'obtenir des résultats en temps réel qu'il est difficile à obtenir avec des algorithmes classiques. [13]

L'apprentissage automatique comporte généralement deux phases : **La première** : Cette phase dite « d'apprentissage » ou « d'entraînement » est généralement réalisée préalablement à l'utilisation pratique du modèle. Consiste à estimer un modèle à partir de données, appelées observations, qui sont disponibles et en nombre fini. **La seconde** phase correspond à la mise en production : le modèle étant déterminé, de nouvelles données peuvent alors être soumises afin d'obtenir le résultat correspondant à la tâche souhaitée. En pratique, certains systèmes peuvent poursuivre leur apprentissage une fois en production, pour peu qu'ils aient un moyen d'obtenir un retour sur la qualité des résultats produits. [13]

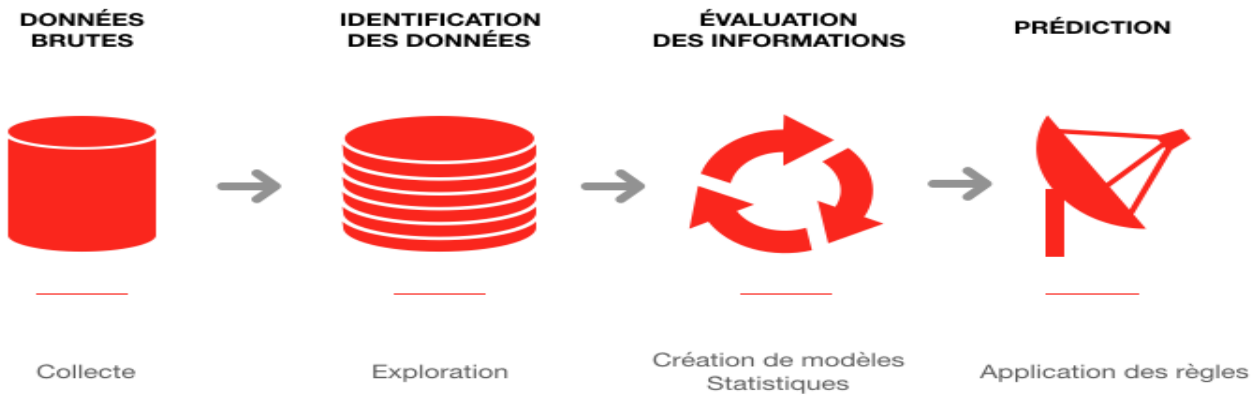

Figure 2-1 Le processus typique de l'apprentissage automatique

2.3 Les domaines d'applications de l'apprentissage automatique

L'apprentissage automatique s'applique à un grand nombre d'activités humaines et convient en particulier au problème de la prise de décision automatisée. Il s'agira, par exemple:

- Etablir un diagnostic médical à partir de la description clinique d'un patient
- Donner une réponse à la demande de prêt bancaire de la part d'un client sur la base de sa situation personnelle
- Déclencher un processus d'alerte en fonction de signaux reçus par des capteurs
- La reconnaissance des formes
- La reconnaissance de la parole et du texte écrit
- Contrôler un processus et de diagnostiquer des pannes.

2.4 Méthodologie d'un projet de machine Learning

Un projet de Machine Learning s'avère très différente de ceux des autres projets informatiques classiques. Le processus de Machine Learning est un processus qui comporte plusieurs étapes, chaque étape présente ses propres défis, techniques et conceptuels. La réussite d'un projet de machine Learning rejoint donc le respect des étapes ci-dessous:

- **Définition des objectifs**

Pour réussir un projet et notamment en Machine Learning il faut bien déterminer ces objectifs. Cela revient au type de projet et aussi il faut avoir une bonne lecture et de l'expérience sur le domaine en application. Dans cette optique, il faut déterminer de quelle typologie de problème nous

devons résoudre. Alors, nous devons savoir si nous avons des données d'expérimentation avec résultat ou non, afin de déterminer si nous abordons un problème de type supervisé ou non-supervisé. Ensuite il faut savoir quelle est la typologie du problème à résoudre : Régression, Classification ou Regroupement.

- **Définition d'ensemble de données utilisé et description des variables**

Une fois que nous avons décidons de notre projet, le moment est venu pour la première étape du projet **la collecte de données**. Cette étape est très importante car c'est la qualité et la quantité des données que vous collectez qui détermineront la qualité de votre modèle à venir. Dans certains cas, vous pourrez être amené à produire des données "artificielles" à partir des vraies données collectées.

- **Le nettoyage et la normalisation des données**

Le nettoyage des données est considéré comme l'une des étapes cruciales du flux de travail, car elle peut faire ou défaire le modèle. Il existe plusieurs facteurs à prendre en compte dans le processus de nettoyage des données. Observations en double ou non pertinentes. Mauvais étiquetage des données, même catégorie se produisant plusieurs fois. Points de données manquants ou nuls. Des valeurs aberrantes inattendues.

- **Choisir un modèle**

L'étape suivante du flux de travail consiste à choisir un modèle. Les chercheurs et les data scientiste ont créé de nombreux modèles ces dernières années. Certaines sont très bien adaptées aux images, d'autres aux données séquentielles, d'autres encore aux données textuelles, … et doit être aussi pris en compte le type de problème : un problème de classification, de régression, de recommandation,
de gaming.

- **La Séparation des données train /test**

  Dans cette étape Il faut faire attention à deux ou trois détails. A savoir, si le problème est un de problème de classification, est ce que les données sont **temporelles**. Il faut aussi tenir en compte si les données sont **groupées.**

- **évaluations des modèles**

Évaluer les performances d'un modèle de classification est un enjeu de grande importance car ces performances peuvent être utilisées pour l'apprentissage en tant que tel ou pour optimiser les valeurs des hyper-paramètres du classificateur ou bien pour faire la comparaison entre plusieurs classificateurs pour choisir le meilleur pour une telle base de données. On a présenté 4 indicateurs, adaptés pour évaluer la performance d'un modèle de classification et qui sont calculés à partir de la matrice de confusion. Ils sont assez simples à comprendre et sont très complémentaires.

- **Précision** : La précision est le rapport entre les observations positives correctement prédites et le total des observations positives prédites

Précision = TP/TP+FP……2.1

- **Rappel** (sensibilité) : Le rappel est le rapport entre les observations positives correctement prédites et toutes les observations de la classe réelle

Rappel = TP/TP+FN ……. …2.2

- **Score F1** : Le **score** F1 est la moyenne pondérée de la précision et du rappel. Par conséquent, ce score prend en compte à la fois les faux positifs et les faux négatifs.

Score F1 = 2*(Rappel * Précision) / (Rappel + Précision) ……. 2.3

2.5 Les types d'apprentissage automatique

L'apprentissage automatique procédé deux principaux types d'apprentissage : l'apprentissage supervisé et l'apprentissage non supervisé.

2.5.1 Apprentissage Supervisé

L'apprentissage Supervisé ou la méthode statistique d'apprentissage de classes consistant à apprendre une fonction de prédiction de classe de la nouvelle d'éléments à partir d'exemples étiquetés, il s'appelle aussi (un modèle) [14].

Il existe deux types de modèles d'apprentissages supervisés : le modèle de classification et le modèle de régression. Dans le modèle de classification permet de prédire une valeur qualitative. Par contre le modèle de régression permet de prédire une valeur quantitative. Cela signifie que l'ensemble des valeurs de sortie Y qu'on essaie d'estimer avec la fonction f est un ensemble de

réels, exemple prix des voitures. La figure 2-2 montre le diagramme de processus d'apprentissage supervisé.

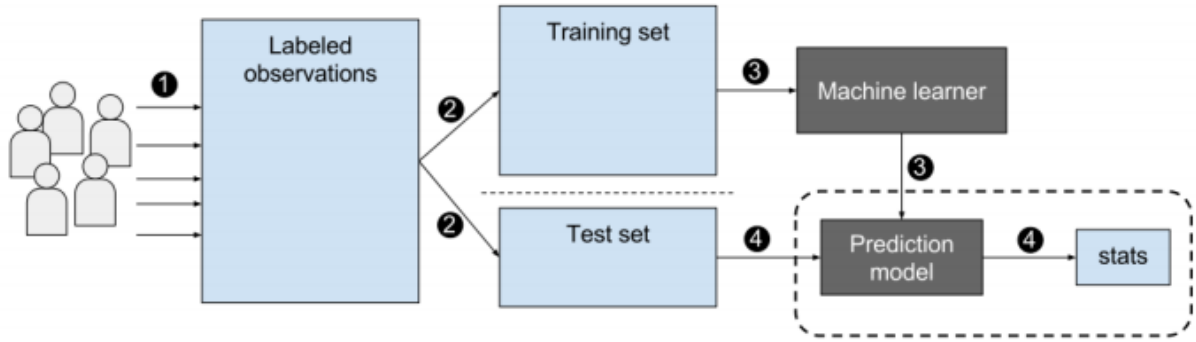

Figure 2-2 Diagramme de processus d'apprentissage supervisé [14]

Le diagramme au-dessus comporte trois partie principales : la base de données, la phase Train et la Phase Test. La base de données représente l'ensemble d'apprentissage ou la partie de données d'apprentissage qui sont étiquetés au préalable. La phase Train consiste à la création du modèle ou la fonction de prédiction de et à la fin la phase de test qui sert à tester la qualité du modèle généré dans la phase Train en lui appliquant sur ensemble de données réservés cette phase de test.

2.5.2 Apprentissage non supervisé

L'apprentissage non supervisé ne contient pas la variable de sortie correspondantes comme est le cas dans l'apprentissage supervisé. Alors son objectif est de modéliser la structure ou la distribution sous-jacente dans les données afin d'extraire automatiquement les catégories à associer aux données qu'on lui soumet [15].

Cette discipline est connue dans ce type d'apprentissage sont par le regroupement (clustering). Une définition courante est que le Regroupement consiste à regrouper un ensemble d'´éléments hétérogènes sous forme de sous-groupes homogène qui sont cachés auparavant. Un problème très courant dans cette discipline est le problème de grade dimensionnalité. Une solution évidente face à ce problème est de réduire la dimensionnalité. Cette dernière consiste à prendre des données dans un espace de grande dimension, et à les remplacer par des données dans un espace de plus petite dimension sans perdre la variance [16].

Tableau 2-1 comparaison entre apprentissage supervise et non supervise [15]

	Apprentissage supervisé	**Apprentissage non supervise**
Données entrée	Utilisé les données connues et étiquetés comme entrées	Données inconnues en entrée
Complexité informatique	Très complexe	Moins de complexités informatiques
Temps réel	Utilise analyse hors ligne	Utilise analyse en temps réel des donnes
Sous-domaines	Classification et régression	Exploitation de règles Clustering et d'association
Précision	Produit des résultats précis	Génère des résultats modérés
Nombre de classes	Nombre de classes connues	Le nombre de classes n'est pas connu

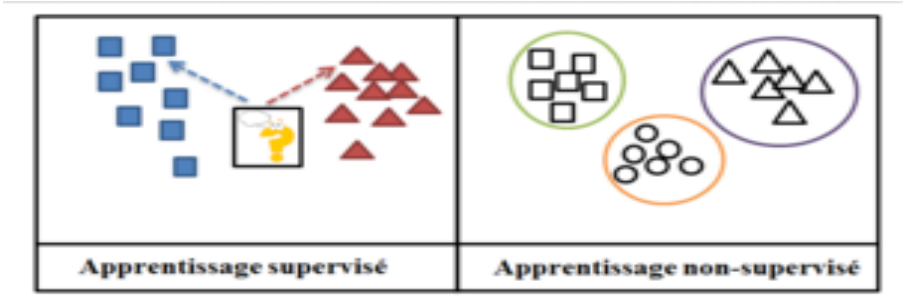

Figure 2-3 Différence entre deux types d'apprentissage [17]

2.5.3 L'apprentissage semi-supervisé

L'apprentissage semi-supervisé est une solution alternative à l'apprentissage supervisé quand on à faire avec des données non complètement étiquetées. Donc, il vise à faire apparaitre la distribution sous-jacente des « exemples » dans leur espace de description. À savoir, le cas de diagnostic : il peut constituer une aide pour le choix des moyens les moins onéreux de tests de diagnostics. Dans la section suivante on présente, quelques algorithmes d'apprentissage supervisé

et non supervisé les plus connus dans la littérature et notamment appliqués dans le diagnostic de diabète.

2.6 Les algorithmes de classification

2.6.1 K nearest neighbors (KNN)

K plus proche voisins ou K- Nearest Neighbors (KNN) en Anglais est l'un des méthodes d'apprentissage supervisé le plus simple, utilisé pour résoudre des problèmes de classification et de la régression. Son fonctionnement est de classer les nouveaux points de données en fonction de la similarité aux points de données voisins [18].

KNN est un algorithme qui ne fait aucunes hypothèses sur la structure des données et de la distribution, ce qui signifie qu'il s'agit d'un algorithme non paramétrique. Il est également appel algorithme de l'apprenant paresseux, car il n'apprend pas immédiatement de l'ensemble d'apprentissage, mais stocke l'ensemble de données et, au moment de la classification, il exécute une action sur l'ensemble de données. KNN fonctionne par classification ou prédiction sur la base d'un nombre fixe (K) de points de données les plus proches de point d'entrée. Cela signifie que pour une valeur choisie de K, un point d'entrée serait classée ou devrait appartenir à la même classe que la classe la plus proche des nombre des points K voisins [19]. Voici une illustration simplifie est présentée dans la Figure **2.4 en-dessous.**

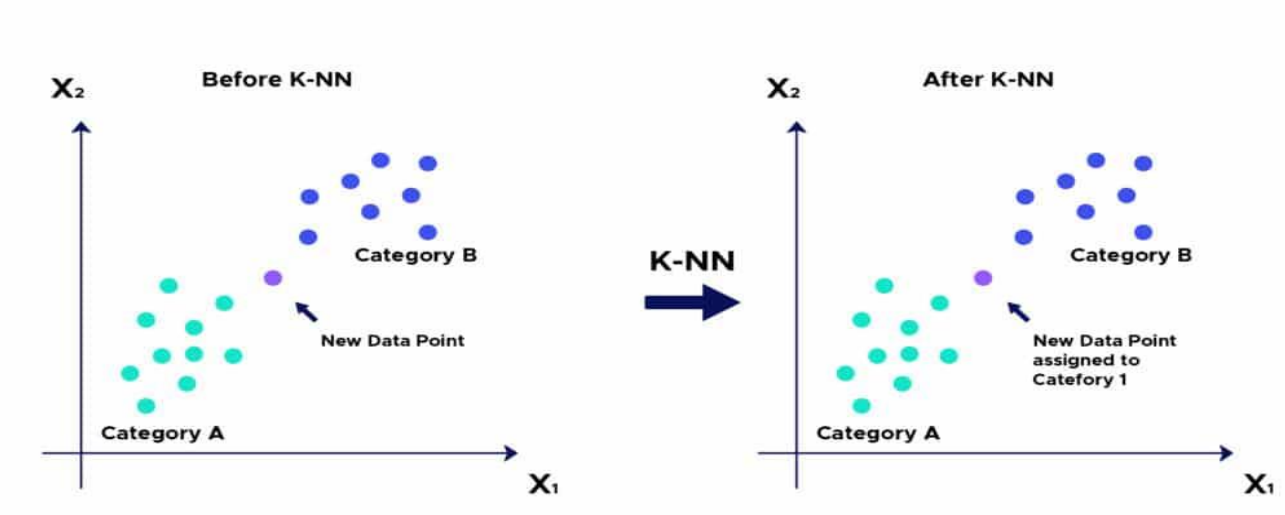

Figure 2-4 le fonctionnement de KNN.

1. **Pseudo code de KNN**

Un pseudo code de l'algorithme KNN est donnés dans l'algorithme 1 en-dessous.

Algorithme 1 : KNN

Début

1. Lire les données DATA: choisir D = distance et, k = nombre de voisins
2. **Pour** chaque donnée X en test **faire**
    - 2-1 Calculer la distance avec tous les données DATA en appliquant la distance D.
    - 2-2 Retenir les premiers K lignes de DATA les proches de X en utilisation D.
    - 2-3 Prendre les valeurs de y des K observations retenues
        - 2-3.1 Si le cas d'une régression, calculer la moyenne (ou la médiane) des y.
        - 2-3.2 Si il s'agit d'une classification, choisir la classe majoritaire des y
    - 2-4 affecter la valeur y calculée dans l'étape 2-3 à l'observation en test X.

Fin

2. Calcul de similarité dans l'algorithme KNN

L'algorithme K-NN a besoin d'une fonction de calcul de distance entre deux données. Pour en faire Il existe plusieurs fonctions de distance à savoir : la distance euclidienne, Manhattan, Minkowski, la similarité de Jaccard et la distance de Hamming…etc. le choix de la distance utilisée dans l'algorithme KNN dépond fortement de types des données en cours. D'après les expérimentations, la distance euclidienne semble plus adéquate lorsqu'il s'agit des données de même type ainsi que pour les données quantitatives. De l'autre côté, la distance de Manhattan est une bonne mesure et qui peut-être appliquer sur des données de diffèrent type. En-dessous on va présenter les définitions de quelques distances les plus utilisées.

La distance euclidienne:

$$\mathbf{D}\,(x, y) = \sqrt{\sum_{j=1}^{n} (xi - yj)^{\,2}} \ldots\ldots\ldots\ldots \mathbf{2.4}$$

Distance Manhattan :

$$\mathbf{D(x,y) = \sum_{i=1}^{k} /xi - yi/ \ldots\ldots\ldots\ldots\ldots\ 2.5}$$

Distance Minkowski

$$\boldsymbol{D(x,y) = \sqrt[p]{\sum_{i=1}^{k} /xi - yi/^{p}} \quad \ldots\ldots\ldots\ 2.6}$$

Il existe d'autres distances selon le cas d'utilisation de l'algorithme, mais la **distance euclidienne** reste la plus utilisée. [20]

3. Le choix du paramètre K : nombre de voisins

Le choix de la valeur K qui exprime le nombre de voisins dépond généralement du jeu de données. Dans la plupart des cas, plus k est petit, c.-à-d. Moins on utilise de voisins dans la prédiction plus on risque de tomber dans le cas de sous apprentissage. Par ailleurs, plus on utilise de voisins important l'algorithme KNN performe de mieux en mieux. Cependant, si K nombre de voisins égalent à N, et sachant que N est le nombre d'observations, on risque d'avoir le phénomène de sur-apprentissage

Exemple

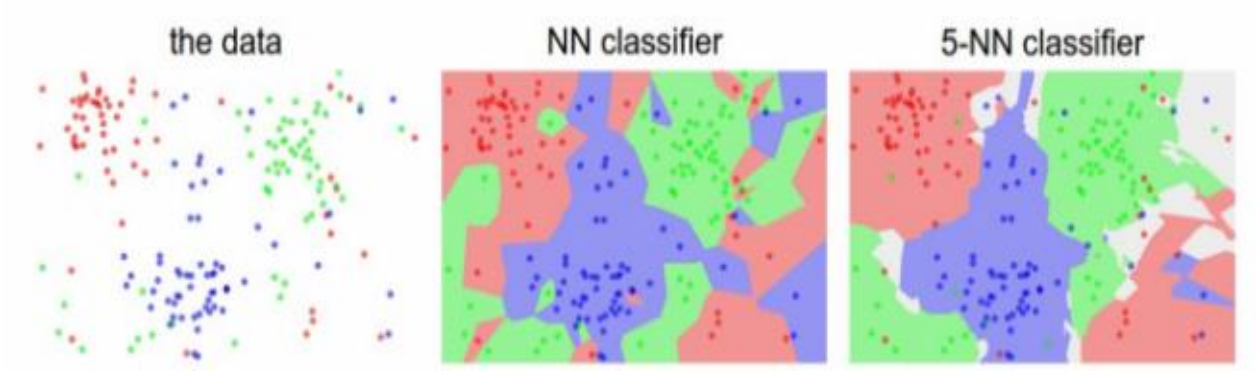

Figure 2-5 Classification des données avec KNN dans un plan 2d [20]

L'image ci-dessus, dans la partie la plus à gauche représente des points dans un plan 2D avec trois types d'étiquetages possibles (rouge, vert, bleu). Pour le classificateur 5-NN, les limites entre chaque région sont assez lisses et régulières. Quant au N-NN Classifier, on remarque que les limites sont "chaotiques" et irrégulières. Cette dernière provient du fait que l'algorithme tente de

faire rentrer tous les points bleus dans les régions bleues, les rouges avec les rouges etc… c'est un cas sur-apprentissage. Il est claire dans ce cas discuté que le classificateur 5-NN donne de meilleur résultats que le KNN.

4. Limitations de K-NN

K-NN est un algorithme très simple dans son comportement. Il ne nécessite pas une phase d'entrainement (lazy algorithm). De ce fait il prédit directement les données en teste à partir un ensemble d'entrainement. En revanche, il doit **stocker en mémoire l'ensemble des observations** pour pouvoir effectuer sa prédiction ainsi que la taille de l'échantillon **d'entrainement,** le choix de la méthode de calcul de la distance. Cependant, le nombre de voisins peut ne pas être évident. Il faut essayer plusieurs combinaisons et faire du tuning de l'algorithme (réglage de hyper-paramètres) pour avoir un résultat satisfaisant. [20]

2.6.2 Les arbres de décision

Les arbres de décisions l'est un algorithme d'apprentissage supervisé le plus utilisé et le plus connu. Il est adapté à la solution des problèmes de classifications ou également de régressions. Un arbre de décision est une structure arborescente semblable à un organigramme où un nœud interne représente une caractéristique (ou un attribut), la branche représente une règle de décision et chaque nœud feuille représente le résultat, cette structure aide pour prendre la décision. Un chemin de la racine vers une feuille présente une règle de décision. Le plus grand avantage de cet algorithme réside dans l'explication et l'interprétation de ces résultats. Les arbres de décisions appartiennent au type non-paramétrique et qui signifie qu'il n'y a pas d'hypothèse sous-jacente sur la distribution des données [21].

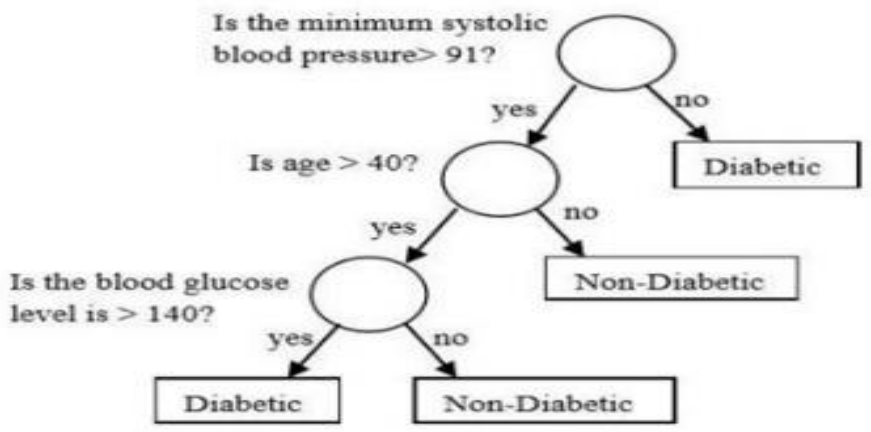

Figure 2-6 Arbre de décision répondre à la question si un personne diabétique ou non ? [22]

Comme nous l'avons déjà mentionné au-dessus, les arbres de décision sont bien adaptés aux problèmes de catégorisation où les attributs sont vérifiés pour déterminer une catégorie finale à cause de sa construction naturelle qui représente un arbre. De ce fait, pour prédire un nouveau cas il suffit juste de faire passer les valeurs de ses attributs dans l'arbre de la racine vers la feuille. La règle de décision alors aura une forme de ce type Si . . . alors . . . sinon. . . . A titre d'exemple et si on prenait l'exemple présenté dans la figure au-dessus on aura la règle de prédiction suivante :

Si minimun systolic blood pressure > 91 = no , alors : personne = diabetic .

Le choix d'attribut dans la construction des nœuds

Le principal problème qui se pose lorsque la construction d'un arbre de décision si comment choisi ou sélectionné le meilleur attribut pour le nœud racine et qui sépare mieux l'ensemble de données. Il existe deux mesures principales et populaires sont :

1. Indice de Gini

   Insérer définition et formule…..

2. Gain d'information

   Insérer définition et formule…..

Algorithme de construction d'un arbre de décision :

1. lire les données
2. sélectionner le meilleur attribut (nœud racine) en appliquant IG ou I.Gini
3. diviser pour chaque branche s'étendant à partir de nœud, répéter récursivement (3).
4. Arrête la division si arrive à :
   - un nœud pur,
   - très peu de points,
   - On atteint une certaine profondeur.

Malgré les avantages des arbres de décision à savoir ils sont faciles à expliquer et comprendre, Fonctionne avec des données catégorielles et numériques et ils sont également peu couteux en termes de calcul. Les arbres de décisions souffrent de quelque problème comme ils prennent beaucoup de temps pour former le modèle. Ils deviennent plus complexes à mesure qu'il s'approfondit. En plus, lorsqu'on a un petit changement dans les données peut entrainer un changement global de la structure de l'arbre de décision.

2.6.3 Support Vector Machine (SVM)

Machine à vecteurs de support ou SVM (Support Vector Machine en anglais) est l'un des algorithmes d'apprentissage supervisé les plus populaires, peut être le meilleur en termes de performance selon les travaux de recherche. Il est utilisé pour les problèmes de classification et également de régression. Le principe de do fonctionnement réside à la création de la meilleure ligne ou limite de décision qui peut séparer l'espace en deux classes tout en maximisant la distance entre la ligne de séparation et les deux espaces des deux classes, autrement dit maximiser la marge. Cette ligne de est appelée hyperplan. L'algorithme se base sur les points vecteurs extrêmes pour créer son hyperplan. Ces points extrêmes sont appelés les vecteurs de support. La figure suivante illustre deux classes (classe des points bleus et classe des points roses) différant qui sont classés avec un hyperplan.

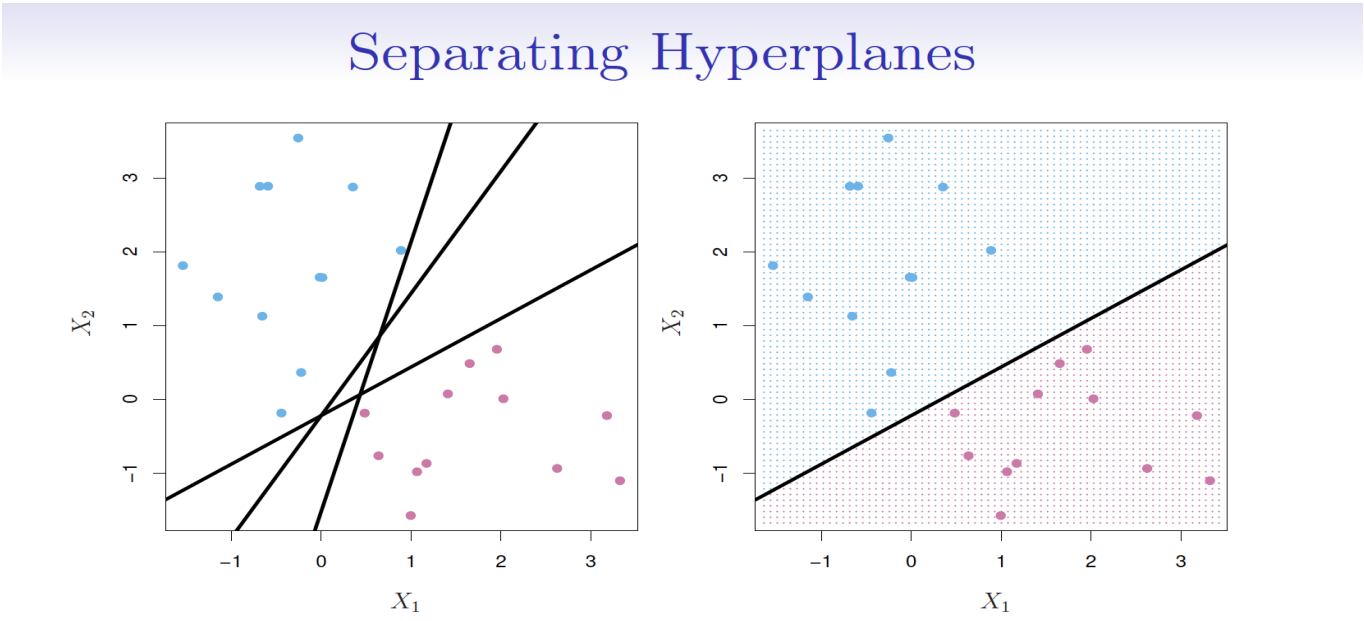

Figure 2-7 Séparation parfait de deux classes avec un hyperplan [23]

Exemple

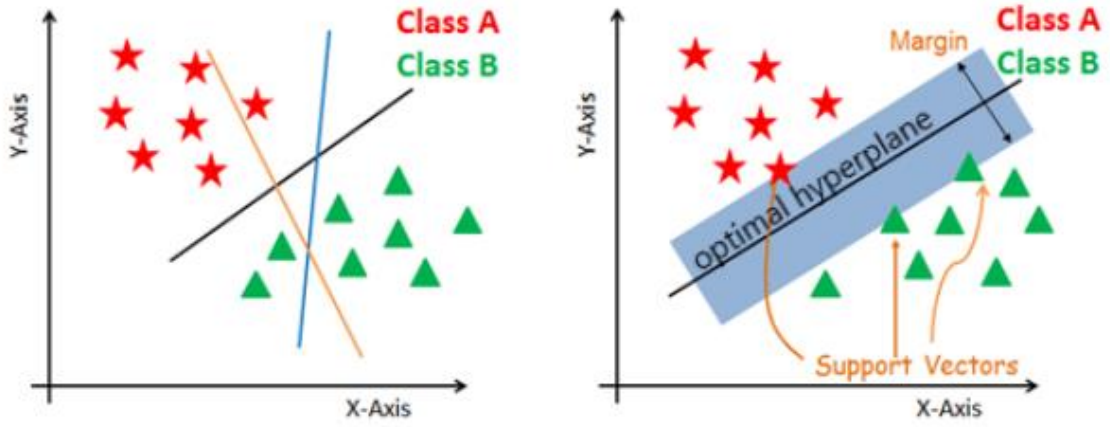

Figure 2-8 Un simple exemple sur le fonctionnement de l'algorithme SVM [24]

Malgré son simple principe qui est basé sur la maximisation de la marge dans la création de son séparateur, Il a la capacité de gérer de grands espaces fonctionnels. Fonctionne bien avec même des données non structurées et semi-structurées comme du texte, des images et des arbres. Il s'adapte relativement bien aux données de grande dimension. Dans l'autre côté, Il est sensible au bruit. Difficile de comprendre et d'interpréter le modèle final, les poids variables et l'impact individuel. L'extension de la classification `a plus de deux classes est problématique

2.6.4 Naïve bayésienne

Naïve Bayes (Mccallum et Nigam, 1998) se base sur la règle de Bayes pour prédire qu'un nouveau cas appartient à une catégorie donnée. Dans cet algorithme la catégorie la plus probable sera choisie. Plus particulièrement, l'algorithme Naïve Bayes utilise la probabilité conditionnelle d'ensemble des attributs v et d'une catégorie c pour calculer la probabilité d'un attribut qui appartient à un certain nombre d'observations Oj. A noter que, l'hypothèse bayésienne suppose que tous les attributs d'une Observation O soient indépendants les unes des autres selon le contexte d'une catégorie c. le calcul de la probabilité qu'une observation Oj appartient à une catégorie ci est présentées dans Equation **2.7** :

$$\mathbf{P\,(A|B)} = \frac{\mathbf{P(B|A)P(A)}}{\mathbf{P(B)}} \quad \text{.....2.7}$$

Dont,

P(A|B) : la probabilité conditionnelle que l'évènement A se produise, étant donné que B s'est produit. Ceci est ´également connu comme la probabilité postérieure.

P(B|A) : la probabilité conditionnelle que l'évènement B se produise, étant donné qu'A s'est produit.

P(A) et P(B) : probabilité de A et B sans égard l'un à l'autre. [24]

2.7 Etat de l'art sur la prédiction du diabète en appliquant les classificateurs

Les chercheurs ont appliqué différentes techniques de ML (machine Learning) pour la prédiction du diabète afin d'améliorer la précision des systèmes de soins de santé. En 2011, AlJarullah a appliqué les arbres de décision pour la prédiction du diabète sur **'pima idian diabète'**. Les techniques étudiées et évaluées pour cette recherche sont la sélection d'attributs, gestion des valeurs manquantes et discrétisation numérique. La précision obtenue grâce à ce modèle était de 78,17 % [25]. Ensuite, Xue-HuiMeng a réalisé une analyse comparative des trois modèles de prédiction du diabète. Parmi ces trois, l'algorithme C5.0 a surpassé les autres en donnant la meilleure précision [26]. Asha Gowda Karegowda, M.A. Jayaram, A.S. Manjunath ont utilisé un algorithme en cascade de K moyenne et K du plus proche voisin pour la catégorisation des patients diabétiques. Ils ont classé les patients diabétiques en proposant des résultats en utilisant KNN et K moyenne. La précision atteinte par le système proposé est de 82 % [27]. En 2013, Christobel Y.A et al ont proposé un nouvel algorithme de classification par classe K-Nearest Neighbor (CKNN) pour la classification des données sur le diabète. Ils ont utilisé un ensemble de données sur le diabète pour tester l'algorithme CKNN et le comparé avec le simple KNN par les mesures de performances telles que la précision, la sensibilité et la spécificité. Le modèle CKNN proposé donne une meilleure précision de classification de 78,16% par rapport au KNN simple [28]. Kumari V. Anuja ont proposé une machine à vecteurs de support (SVM) avec une fonction de noyau à base radiale pour la classification des données sur le diabète. Le modèle proposé atteint une précision de 78 % qui peut être utilisée avec succès pour diagnostiquer la maladie du diabète [29]. En 2013, Subham Khanna et al a proposé une application classificatrice basée sur le poids pour la prédiction du diabète Binning. Les mesures de performance utilisées sont la précision, la sensibilité, la spécificité et les valeurs kappa donnant valeurs de 83,2 %, 70,9 %, 89,7 % et 1,003, respectivement [30]. Ensuite, Parashar A et AL ont proposé une analyse discriminante linéaire et une machine à vecteurs de support pour le diagnostic de l'ensemble de données sur le diabète **Pima indian diabete**, où LDA réduit les sous-ensembles de caractéristiques et SVM est chargé de classer les

données. Ils ont également comparé SVM avec un réseau de neurones à action directe (FFNN).. ils ont trouvé que SVM + LDA proposé donne une meilleure précision de classification de 77,60% avec deux fonctionnalités [31] . Lin Li, a diagnostiqué un diabète en utilisant le poids approche de vote ajustée en formant le modèle proposé sur PIDD. L'auteur a utilisé un ensemble de SVM, ANN et naïve bayes pour prédire le diabète. Pendant la phase de prétraitement, le les enregistrements avec des valeurs biologiquement impossibles sont supprimés. La méthode Wrapper est utilisée pour la sélection de caractéristiques avec cinq caractéristiques pour la classification au lieu de neuf. Réglage du poids approche est utilisée pour combiner les résultats de classificateurs. Sur la base de cette approche, ils ont atteint une précision de 77,0%, spécificité 86,8% et sensibilité 58,3% [32]. Longfei Han et al ont appliqué sur un ensemble de données de diabète collectées pour China Health des algorithmes SVM et RF. L'ensemble d'entraînement est d'abord entraîné sur SVM en ajustant les paramètres pour obtenir la plus grande précision, suivi de l'extraction des règles à l'aide de RF en ajustant la règle paramètres d'induction pour obtenir les meilleures règles. Ces règles sont alors utilisées pour prédire la classe de chaque enregistrement à partir des données de test. Les traitements utilisées sont l'exclusion des données vacantes, le bruit annulation des données et sélection des fonctionnalités. Les valeurs de précision, rappel et valeur f calculés après validation croisée 10 fois étaient respectivement de 81,8 %, 75,6 % et 0,786 [33]. Dans l'autre côté, Farahmandian M. et AL ont appliqué un ensemble de données sur le diabète à divers algorithmes de classification tels que SVM, KNN, Naïve bayes, ID3, CART et C5.0 pour classer les données sur le diabète. Ils ont comparé la précision de classification de ces modèles. SVM donne la meilleure précision de classification puisque 81,77 % se comparent aux autres [34].

2.8 Conclusion

Dans ce chapitre, nous avons présenté les fondements théoriques de l'apprentissage automatique, le processus général de machine Learning, les types d'apprentissage que ce soit supervisé ou non-supervisé. Des algorithmes d'apprentissage automatiques notamment les classificateurs ont été clairement montrés avec leur définition et concept. Ensuite, nous avons introduit un état de l'art sur l'application des algorithmes de classification sur le diabète. La suite de ce mémoire est consacré à la partie objectif principal dont on applique les algorithmes de classification notamment l'algorithme KNN sur la base données **Pima Indians Diabetes Database** .

III. Implémentation et résultats

Chapitre 3 : implémentation et résultats

3.1 Introduction

Dans ce dernier chapitre, nous présentons la partie expérimentale de notre projet dans laquelle nous définissons l'environnement logiciel et matériel utilisés. Nous introduisons la base de données ou le banc d'essai qui est '**Pima indian diabetes database'**. Une description détaillée est affichée concernant ses caractéristiques à savoir le nombre d'observations et les variables descriptifs avec leurs types ainsi que les abréviations avec leurs significations. Ensuite, nous allons décrire les différentes étapes de prétraitement appliqués sur cette base de données comme le cas du traitement des valeurs NULL. Concernant le classificateur, nous avons choisi l'algorithme H-Plus-Proche-voisins KNN dans lequel plusieurs paramètres ont été testés tels que le nombre de voisin (k) et les mesures de distances ou de similarités. Dans l'autre côté, une étude comparative est accompagnée dans ce chapitre dans laquelle nous montrons l'avantage de notre choix de l'algorithme KNN contre les autres classificateurs à savoir les machines à support vecteur (SVM), les arbres de décision ainsi que l'algorithme des forets aléatoire. Les résultats des expérimentations sont communiqués à la fois qualitativement et quantitativement en termes de Précision, Rappel et F-mesure. À la fin nous terminerons ce chapitre par une conclusion.

3.2 Outils et environnement de développement

3.2.1 Kaggle

Kaggle est une plateforme web organisant des compétitions en science des données. Kaggle propose une plateforme pour coder et tester les modèles directement en ligne. C'est une fonctionnalité très intéressante puisqu'elle nous permet d'utiliser la puissance d'un GPU sans forcément avoir le hardware qui correspond. [35]

3.2.2 Python

Python C'est un langage de programmation multi-paradigme et le langage de programmation dominant dans la data science avec de nombreuses implémentations ce qui le rend encore plus intéressant .concernant le domaine de l'apprentissage automatique Python se distingue tout particulièrement en offrant une pléthore de librairies de très grande qualité, couvrant tous les types d'apprentissages disponibles qui combine la facilite d'utilisation et d'apprentissage avec la puissance des librairies qu'elles possèdent. Parmi ces bibliothèques, nous avons utilisé : [36]

- **Matplotlib** : est une bibliothèque complète pour créer des visualisations statiques, animées et interactives en Python.

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

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

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

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

3.3 Base de données utilisée

Dans ce qui suit nous allons présenter la base de données utilisée dans cette étude en introduisant une définition de cette dernière et la description ses variables prédictives.

- **Définition de l'ensemble de données**

Pima indian diabetes database est un ensemble de données provient à l'origine de l'Institut national du diabète et des maladies digestives et rénales. L'objectif de l'ensemble de données est de prédire par diagnostic si un patient souffre ou non de diabète sur la base de certaines mesures diagnostiques incluses dans l'ensemble de données. Plusieurs contraintes ont été placées sur la sélection de ces instances à partir d'une base de données plus importante. En particulier, tous les patients ici sont des femmes d'au moins 21 ans d'origine indienne Pima .Cette base données se composent de plusieurs variables prédictives médicales et d'une variable cible. Les variables prédictives incluent le nombre de grossesses que la patiente a eues, son IMC, son niveau d'insuline,

son âge, …etc. la figue **Figure 3-1** en-dessous montre un aperçu sur les premiers enregistrements de l'ensemble de données ***Pima indian diabetes database.***

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 3-1 Aperçu de l'ensemble de données

Tableau 3-1 Description des variables d'ensemble de données

Numéro	**Abréviation**	**signification**
1	Pregnancies	Nombre de fois enceintes
2	Glucose	Concentration de glucose plasmatique à 2 heures dans un test de tolérance au glucose par voie orale
3	BloodPressure	BloodPressure Pression artérielle diastolique (mm Hg)
4	SkinThickness	Épaisseur du pli cutané du triceps (mm)
5	Insulin	Insuline sérique 2 heures (mu U / ml)
6	BMI	Indice de masse corporelle (poids en kg /
Numéro	**Abréviation**	**Signification**
7	DiabetesPedigreeFunction	Fonction pedigree du diabète
8	Age	âge en années

9	Outcome	Variable de classe (0 ou 1) 268 sur 768 sont 1, les autres sont 0

L'ensemble de données contient 768 lignes et 9 colonnes. La variable 'Résultat' est la colonne que nous allons prédire qui signifié si le patient est diabétique ou non. 1 signifie que la personne est diabétique et 0 qui veut dire 'non diabétique'. Pour cette base utilisée, sur les 768 cas on trouve 500 sont étiquetées par 0 (non diabétique) et 268 par des 1 (qui veut dire diabétique).

3.4 Démarche expérimentale suivie

Le processus de la catégorisation passe généralement par quatre étapes : lecture des données, prétraitement, classification et ensuite l'évaluation des performances résultats obtenus. Nous avons opté à appliquer la stratégie suivante et qui est présentée sous forme d'un script dans la **Figure 3-2** en-dessous.

1. ***télécharger et lire (Benchmark) .***
2. ***Nettoyage de données***
   - ***2-1 traitements des valeurs NULL***
   - ***2-2 traitements des valeurs erronés (les zéros)***
3. ***Normalisation des données***
4. ***Classification du diabète par l'algorithme de classification KNN***
5. ***Etude comparative pour les algorithmes de classifications Knn, Svm, Random Forest et D.trees.***
6. ***Evaluer les résultats obtenus.***

Figure 3-2 processus suivi pour classification de Diabètes

3.5 Mesures utilisées pour l'évaluation

Pour évaluer les algorithmes de classification, nous avons utilisé les mesures : Précision, *Rappel* et *F-mesure* qui sont communément utilisées. Pour définir ces mesures, nous avons besoin de définir les valeurs suivantes par rapport à une catégorie C:

- ***TP : le nombre de vrais positifs (True positives).*** C'est le nombre d'instances correctement classés dans la catégorie *C*.
- ***FP : le nombre de faux positifs (False positives).*** C'est le nombre d'instances incorrectement classés dans la catégorie *C*.
- ***FN : le nombre de faux négatifs (False negatives).*** C'est le nombre d'instances incorrectement classés en dehors de la catégorie *C*.

Ainsi les mesures de précision et de rappel sont définies comme suit :

Précision : la *précision (P)* est le rapport du nombre de documents correctement attribués à la catégorie *C* au nombre total de documents classés comme appartenant à la catégorie *C*.

$$\mathbf{P} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FP}} \dots\dots\dots\dots\dots\dots\dots 3.1$$

Rappel : le *Rappel (R)* présente le rapport du nombre de documents correctement attribués à la catégorie *C* au nombre total de documents appartenant réellement à la catégorie *C*.

$$\mathbf{R} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}} \dots\dots\dots\dots\dots\dots\dots 3.2$$

De plus, une troisième mesure commune est appelée *F-mesure* (FM) est définie comme suit :

F-mesure : la *F-mesure (F)* désigne la moyenne harmonique entre la précision et le rappel.

$$\boldsymbol{F} = \mathbf{2}.\frac{\boldsymbol{P}.\boldsymbol{R}}{\boldsymbol{P} + \boldsymbol{R}} \dots\dots\dots\dots\dots\dots\dots 3.3$$

3.6 Expérimentation

Nous avons réalisé une série d'expérimentations sur la base de données *Pima indian diabetes database* . En effet, Nous avons testé et comparé les résultats des 5 algorithmes de classification à titre d'exemple KNN et SVM. Le tableau 3.4 montres les résultats de performances de ces 5 algorithmes et les métriques d'évaluation sont exprimées en termes de taux de Précision, de Rappel et de F-Measure. Le vecteur des caractéristiques résultant de la phase de prétraitements contient 9 variables descriptives et qui sera pris sans réduction dans la phase suivante de classification. Dans l'autre côté et concernant la méthode de test, nous avons utilisé la technique d'échantillonnage Train et Test, c.-à-d. réserver une partie de données pour l'apprentissage et une autre pour la phase test.

3.6.1 Fractionnement de l'ensemble de données « séparation tain/test »

Nous utilisons la méthode « train test split » importé de la bibliothèque ***sklearn*** pour effectuer le fractionnement train/test. Test size=0.3 à l'intérieur de la fonction indique le pourcentage des données qui doivent être conservées pour le test, autour 25% pour le test et le reste de 75% pour l'entrainement ce qui signifie 524 observations partie d'entrainement et 262 observations partie test.

3.6.2 Les algorithmes de classification choisis

Dns notre expérimentation, nous allons tester et comparer les résultats des cinq (5) algorithmes de classifications, Les classificateurs appliqués sur la base de données **Pima indian diabetes database** et qui sont utilisés pour la prédiction de diabète sont :

1. KNN (K-Nearest Neighbors)
2. Les arbres de décision (Décision tree)
3. SVM (Support Vector Machin)
4. Random Forest (foret aleatoire)
5. Naïve bayésienne (Gaussian Naive Bayes)

On a mis l'accent dans notre étude sur l'algorithme KNN qui est très promoteur en termes de résultats de performances.

3.6.3 Le nettoyage et la normalisation des données

1. Traitement des valeurs manquantes et zéros

Avant d'appliquer la classification automatique sur les données, nous devons d'abord supprimerons les zéros dans les colonnes où zéro n'a pas de sens, à savoir les colonnes BloodPressure et SkinThickness. Aussi pour les valeurs nulles (valeurs manquantes) sont en effet des données manquantes qui auraient dû être étiquetées NULL. Dans notre cas, nous avons utilisé la fonction médiane qui est la moyenne des valeurs non nulles dans chaque colonne pour remplacer les valeurs nulles ainsi que les zéros.

2. Normalisation des données

Notre base de données **Pima indian diabetes database** utilisée contient malheureusement des variables avec des ordres de grandeurs différents ce qui produit des performances médiocres. Pour pallier à ce problème nous avons appliqué la technique de normalisation '' Feature Scaling ''.

Plusieurs formulations existent dans la littérature, nous avons choisi la formule (**3.4**) de normalisation suivante.

$$z = \frac{xi - \mu}{\sigma} \quad \text{……..3.4}$$

- x_i la valeur qu'on veut normaliser (variable en entrée).
- μ la moyenne (*mean*) *des* observations pour cette variable.
- α est l'écart-type *(std*) *des* observations pour cette variable.

3.6.4 Classification de diabètes par l'algorithme KNN

Dans notre travail proposé, nous avons appliqué l'algorithme K- plus proche voisin sur les données de l'échantillon d'apprentissage et de test. Nous avons obtenu des résultats pour différentes valeurs de K qui est le nombre de voisins les plus proches. Les mesures de performances précision, Rappel er F-mesure ont été calculés pour chaque valeur de K dans l'ensemble {1, 3, 5, 7, 9, 11, 13}. Les résultats obtenus des expérimentions effectuées sont communiqués dans **le tableau 3.2** en-dessous.

Tableau 3-2 précision, rappel, f1_score en variant le paramètre K

La valeur de K	précision	rappel	F1-score
K=1	0.62	0.58	0.60
K=3	0.65	0.58	0.62
K=5	0.64	0.58	0.61
K=7	0.65	0.58	0.62
K=9	0.65	0.58	0.62
K=11	**0.68**	**0.61**	**0.64**
K=13	0.68	0.54	0.6

D'après les résultats montrés dans le tableau **3.2** on constate que les résultats de performance varient entre 0.54 et 0.68 pour toutes les mesures précision, rappel et F-measure. On remarque aussi, que plus le paramètre k n'augmente, plus le taux de mesures précision, rappel et F-measure augmentent également. Le meilleur score est enregistré au niveau de k=11 dans lequel on a obtenu

0.68, 0.61 et 0.64 pour les mesures précision, rappel et F-measure consécutivement. Cela est plus explicite dans la figure **3-4** ci-dessous.

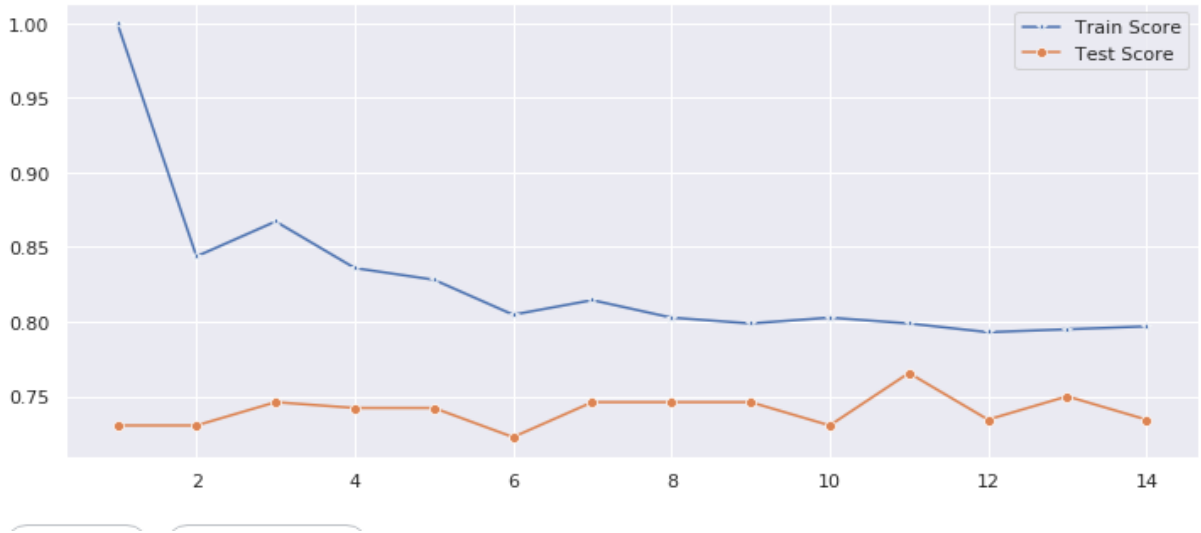

Figure 3-3 Evolution de F1score selon k dans la phase Train et Test.

Dans la figure **3-3**, on constate clairement que le meilleur résultat est capturé à k = 11 et cela dans la phase test, Par contre dans la phase d'entrainement (train) le meilleur score est de 1 pour une valeur de k égale à 1 avec un score moins de 0.75 dans la phase d'apprentissage. Cette divergence de performance entre l'apprentissage et la phase test est connu sous le nom de sur apprentissage. De ce fait, k=1 est le mauvais choix pour l'algorithme KNN. On remarque aussi écart considérable entre les résultats de phase d'entrainement et ceux de test l'lorsque K est entre [1..7]. Cette écart diminue de moins en moins l'lorsque k est supérieur ou égale à 11.

3.6.5 Combinaison de KNN avec Les Métriques

Concernant les métriques il existe plusieurs fonctions dans lesquelles on trouve les distances: Euclidien, Manhattan, Minkowski, Hamming et Chebyshev. Dans cette partie on va tester notre algorithme KNN sur le DataSet **Pima indian diabetes database** en se basant à chaque fois sur l'une des distances mentionnées au-dessus. On pense en tous les cas dans ces cinq distances que chacune d'entre elles peut conduire à une performance différente de celles des autres distances. Dans le **tableau 3-3** en-dessous montre le comportement de notre algorithme KNN sur les cinq métriques différentes tout en fixant k sur la **valeur 11**. Dans chaque expérimentation on relève les mesures de : Précision, Rappel et F1_score.

Tableau 3-3 Les mesures de performances avec les différentes métriques

	Précision	**Rappel**	**F1-score**
Euclidien	0.68	0.61	0.64
Manhattan	0.67	0.54	0.60
Minkowski	0.68	0.61	0.64
hamming	0.47	0.09	0.15
chebyshev	0.65	0.51	0.60

D'après le tableau ci-dessus le modèle KNN obtenu la meilleure précision qui égal à 0.68, le meilleure score de rappel égal à 0.61 et le meilleur F1_score égal 0.64 dans l'utilisation des métriques euclidien et Minkowski.

Nous sélectionnons les distances euclidien et Minkowski comme des métriques les plus optimale et qui fonctionne mieux pour notre ensemble de données en raison de sa grande précision et score de rappel.

3.6.6 Comparaison de KNN avec quelques algorithmes

On a réalisé une série d'expériences sur la base de données Pima indian diabetes database sur laquelle nous avons appliqué en premier lieu l'algorithme de classification KNN. On a ensuite comparé les performances de cet algorithme contre quatre algorithmes de classification les plus connus dans la littérature, notamment : Naïve bayes, arbre de décision, SVM, RandomForestClassifier. .

Le résultat dans le tableau 3.4 expose les résultats des expérimentations de ces 5 algorithmes de classification y compris KNN. Les résultats de performances sont exprimés en termes de Précision, Rappel et F1_score. Rappelons que tous les algorithmes ont été testés dans le même environnement de développement Kaggle.

Tableau 3-4 Les résultats des attributs d'évaluations pour les différents modèles

Algorithme	**Précision**	**rappel**	**F1-score**

KNN	**0.68**	**0.61**	**0.64**
SVM	0.67	0.04	0.08
RandomForestClassifier	0.63	0.56	0.60
GaussianNB	0.6	0.61	0.60
Tree	0.63	0.55	0.58

D'après le tableau ci-dessus le modèle KNN paramétré avec k=11 et la distance Euclidienne a obtenu le meilleur résultat en termes de précision qui égal à 0.68, le meilleure score en Rappel qui égal à 0.61 ainsi que en F1_score il a enregistré un score de 0.64. C'est-`a-dire que sur toutes les patients diabétiques 68% d'entre eux sont correctement classé à l'aide de mesure de diagnostiques médicales en appliquant l'algorithme KNN. De l'autre côté, l'algorithme SVM a enregistré 0.04 en termes de Rappel. Cette valeur est strictement inacceptable lors de la prédiction automatique. Pour les trois autres algorithmes Random Forest, Naive Bayes et les arbres de décisions, on trouve que ces trois algorithmes sont compétitifs et ils sont obtenus presque les mêmes résultats de performances : autour de 0.6 en terme de F1-score. En conclusion KNN est le meilleur algorithme sur les cinq algorithmes testés dans cette expérimentation.

3.7 Conclusion

Ce chapitre nous a permis de conduire la partie expérimentale de notre projet. En effet, on a présenté la stratégie suivie pour le processus de classification, les outils de développement utilisé ainsi que la base de données **Pima indian diabetes database** utilisé dans ce travail. Dans cette étude et lors de la phase de prétraitement, nous avons appliqué quelques fonctions sur les données à savoir, la gestion des valeurs manquantes, le traitement des valeurs erronés, et le cas des zéros et ainsi que l'étape de normalisation des données. En plus et dans les expérimentations, nous avons choisi 5 algorithmes de classification pour la partie prédiction de diabètes dans laquelle nous avons comparés leur comportements. Rappelons, que ces algorithmes de classification sont : les SVMs, D.Tree et N.Bayes Random Faorest et l'algorithme KNN. Les résultats de performances ont montré clairement l'avance de l'algorithme KNN contre tous les autres algorithmes choisis dans cette étude.

IV. Conclusion générale et perspective

Conclusion générale et perspective

Le diabète est l'un des problèmes de santé majeurs dans le monde. Selon le rapport de l'OMS 2011, environ 346 millions de personnes dans le monde souffrent de diabète sucré. Un diagnostic plus précoce évite de nombreuses complications qui peuvent surviennent identification ici en Algérie la prévalence estimée à 14.4% d'après SANOFI qui est une partenaire de santé des patients algérienne. Une prédiction plus précoce évitera des complications de cette maladie. L'approche d'apprentissage automatique résoudre ce problème critique dans le but de cette étude pour construire un modèle capable de prédire si les personnes sont diabétiques des classificateurs.

Dans ce mémoire nous avons choisi l'algorithme KNN Comme classificateur dans lequel plusieurs paramètres tels que le nombre de voisin (k) et les mesures de distances ou de similarités ont étaient testés dans cette étude. Les résultats ont montré que plus le paramètre k n'augmente, plus le taux de mesures de précision, rappel et F mesure augmentent également. Le meilleur score est enregistré au niveau de k=11. Les distances euclidien et Minkowski comme des métriques les plus optimales et qui fonctionne mieux pour notre ensemble de données en raison de sa grande précision et score de rappel. En plus, une comparaison est accompagnée comportant l'algorithme KN contre différents algorithmes de classification d'apprentissage supervisé tel que : les arbres de décision, forêt aléatoire, machine à vecteurs de support, Naïves Bayes sur les données '**Pima indian diabetes database**'. Les résultats de performances ont montré clairement l'avance de l'algorithme KNN contre tous les autres algorithmes choisis dans cette étude.

Pour les travaux futurs, plusieurs pistes peuvent être explorées. On peut appliquer la même expérimentation sur d'autres bases de données de diabète ou même de type différents pour confirmer les résultats obtenus. Améliorer l'algorithme KNN pour avoir de meilleurs résultats en termes de précision et rappel. Une autre alternative est de tester la réduction de dimension sur le problème de détection de diabète

Références

[1] Diabète, épidémiologie, diagnostic, étiologie .Diabétologie - Pr. A. Grimaldi 1999 - 2000

[2] *Garnier-Delamare* ,dictionnaire illustré des termes de médecine,29 ième édition.page238.

[3] https://www.sanofi-diabete.fr/comprendre-diabete/qu-est-ce-que-le-diabete/diabete-qu-est-ce-que-c-est .

[4] https://www.diabetevaud.ch/wp-content/uploads/2016/08/Fiche-S1_1-Definition-du-diabete.pdf

[5] https://www.diabetes.org.uk/Preventing-Type-2-diabetes/Diabetes-risk-factors

[6] Le diabète de type 2 ou diabète non insulinodépendant (DNID), Professeur Serge HALIMI, Avril 2003, Corpus Médical – Faculté de Médecine de Grenoble.

[7] https://www.sanofi-diabete.fr/comprendre-diabete.

[8] guide pratique de diabétologie à l'usage du personnel infirmier, n. Marcoz, PIC, EHC, v1.

[9] http://campus.cerimes.fr/endocrinologie/enseignement/item233c/site/html/cours.pdf

[10] http://univ.ency-education.com/uploads/1/3/1/0/13102001/endocrino5an-complications_chroniques_diabete2018boudaoud.pdf

[11] http://www.reseau-diabenfant.org/l-histoire-du-diabete-et-de-son-traitement-85028.kjsp

[12] https://www.dinnosante.fr/fre/17/histoire-du-diabete#:~:text=C'est%204000%20ans%20avant,diab%C3%A8te%20%3A%20soif%20intense%20et%20amaigrissement.

[13] https://fr.wikipedia.org/wiki/Apprentissage_automatique

[14] Pensée Artificielle. Machine Learning pour débutant : Introduction au Machine Learning. [en ligne].Disponible sur :http ://penseeartificielle.fr/introduction-au-machine Learning/

[15] https://waytolearnx.com/2018/11/difference-entre-apprentissage-supervise-et-non-supervise.html

[16] Pensée Artificielle.Machine Learning pour débutant : Introduction au Machine Learning.[en ligne].Disponible sur :http ://penseeartificielle.fr/introduction-au-machinelearning/

[17] https://fr.wikipedia.org/wiki/Apprentissage_non_supervis%C3%A9

[18] Ilemona S.Atawodi.(2019).A Machine Learning Approach to Network Intrusion Detection System Using K Nearest Neighbor and Random Forest.Th`ese 68 Bibliographie de master : université de Southern Mississippi .52p.[en ligne].Disponible sur :https ://aquila.usm.edu/cgi/viewcontent.cgi ?article=1707&context=masters theses

[19] https://datascientest.com/knn

[20] https://mrmint.fr/introduction-k-nearest-neighbors

[21] Gupta, P.to wards datascience.(2017).Decision Trees in Machine Learning.[en ligne].Disponible sur :https ://towardsdatascience.com/decision-trees-in-machinelearning-641b9c4e8052

[22] Choudhury, A.Analytics India Magazine.(2019).Beginner's Guide To Decision Trees : Why Are They Crucial For Data Science Applications.[en ligne].Disponible sur : https ://analyticsindiamag.com/beginners-guide-to-decision-trees-why-arethey-crucial-for-data-science-applications/

[23] https://miro.medium.com/max/3316/1*UGsHP6GeQmLBeteRz80OPw.png

[24] Sharma, N.heartbeat.Understanding the Mathematics behind Support Vector Machines.[en ligne].Disponible sur : https ://heartbeat.fritz.ai/understanding-themathematics-behind-support-vector-machines-5e20243d64d5

[25] A. Al Jarullah, "Decision tree discovery for the diagnosis of type II diabetes." pp. 303-307.

[26] X.-H. Meng, Y.-X. Huang, D.-P. Rao, Q. Zhang, and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," The Kaohsiung journal of medical sciences, vol. 29, no. 2, pp. 93-99, 2013.

[27] Asha Gowda Karegowda, M.A. Jayaram, A.S. Manjunath(2012) 'Cascading K-means Clustering and KNearestNeighbor Classifier for Categorization of Diabetic Patients' IJEAT Vol.1 No.3 pp 147-151

[28] Y. Angeline Christobel, P.Sivaprakasam, ‡$1HZ&ODVVZLVHN1HDUHVW 1HLJKERU&.110HWKRGIRUWKH&ODVVLILFDWLRQRI'LDEHWHV'DWDVHW· IJEAT, Volume-2, Issue-3, February 2013, pp. 396-400, ISSN: 2249 – 8958.

[29] Kumari V. Anuja, Chitra R. (2013). Classification of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications. Vol. 3, pp. 1797-1801, ISSN: 2248-9622.

[30] Parashar A., Burse K., Rawat K. (2014). A Comparative Approach for Pima Indians Diabetes Diagnosis using LDA-Support Vector Machine and Feed Forward Neural Network. International Journal of Advanced Research in Computer Science and Software Engineering. Vol. 4, pp. 378-383, ISSN: 2277 128X.

[31] L. Li, "Diagnosis of Diabetes Using a Weight-Adjusted Voting Approach." pp. 320-324.

[32] L. Han, S. Luo, J. Yu, L. Pan, and S. Chen, "Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes," Biomedical and Health Informatics, IEEE Journal of, vol. 19, no. 2, pp. 728-734, 2015.

[33] S. Khanna, and S. Agarwal, "An Integrated Approach towards the prediction of Likelihood of Diabetes." pp. 294-298.

[34] Farahmandian M., Lotfi Y., Maleki I. (2015). Data Mining Algorithms Application in Diabetes Diseases Diagnosis: A Case Study. MAGNT Research Report. Vol. 3, PP. 989-997, ISSN. 1444-8939.

[35] https://www.kaggle.com/

[36] https://datascientest.com/top-10-des-librairies-python-pour-un-data-scientist