MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE ABBES LAGHROUR KHENCHELA
FACULTE DES SCIENCES ET DE LA TECHNOLOGIE

Département de Mathématiques et informatique

N° de série :……

Mémoire de fin d'études

Pour l'obtention du diplôme de Master (L.M.D)

Spécialité : informatique

Option : sécurité et Technologies web

THEME

Une approche de prédiction médicale basée sur les données cliniques utilisant des algorithmes d'apprentissage automatique

Réalisés par :

Souad BOUTEARI
Alla Eddine RICHE

Dirigé par :

Dr. Hichem RAHAB

Membre de jury:

Dr. Dalal BARDOU	Président
Dr. Hichem RAHAB	Rapporteur
Mr. Nabil AZIZI	Examinateur

Année universities: 2020 - 2021

Abstract

In this work, we have designed four models to predict diabetes in order to reduce the risk and the occurrence of complications of this disease on the health of the patient. To design these models, we used four machine learning algorithms, i.e. K nearest neighbors KNN, Decision trees DT, Support Vector Machine SVM, and Logistic Regression LR. The performance of the obtained models was tested according to the accuracy of each model. The highest accuracy rates were obtained in the decision tree model in both the split method (Train / Test Split) and k_fold cross validation splitting model.

Keywords: machine Learning, K nearest neighbors, Decision trees, Support vector machine, Logistic Regression, diabetes prediction, medical prediction.

Résumé

Dans ce mémoire, nous avons conçu quatre modèles pour prédire le diabète afin de réduire le risque et la survenue de complications de cette maladie sur la santé du patient. Pour concevoir ces modèles, nous avons utilisé des algorithmes d'apprentissage automatique, à savoir ; K voisins les plus proches, les Arbres de décision, Les séparateurs à Vaste Marge SVM, et la régression logistique. La comparaison des modèles a été faite selon le taux de classification (Accuracy). Les taux de classification les plus élevés ont été obtenus dans le modèle d'arbre de décision dans les deux méthodes de division ; Train/Test Split et la validation croisée k_fold.

Mots clés : apprentissage automatique**,** K voisins les plus proches, Arbres de décision, Les séparateurs à Vaste Marge, la régression logistique, prédiction de diabète, prédiction médicale**.**

ملخص

في هذه المذكرة قمنا بتصميم أربع نماذج للتنبؤ بمرض السكري، وذلك لتقليل مخاطر ومضاعفات هذا المرض على صحة المريض. لتصميم النماذج قمنا باستعمال خوارزميات التعلم الآلي والمتمثلة في: ك-الجار الأقرب، أشجار القرار، المقسمات ذات الهامش الأكبر والإنحدار اللوجستي. تمت مقارنة النماذج حسب نسبة الأداء (Accuracy). أعلى نسبة أداء تم التحصل عليها باستخدام نموذج أشجار القرار وذلك باعتماد طريقتين للتقسيم: التدريب|الاختبار والمصادقة المتداخلة.

الكلمات المفتاحية: التعلم الآلي، ك -الجار الأقرب، أشجار القرار، المقسمات ذات الهامش الأكبر ،الإنحدار اللوجستي، التنبؤ بمرض السكري، للتنبؤ الطبي.

Remerciements

Nous remercions et exprimons notre grande gratitude au grand Dieu qui nous a donné la force et la volonté de faire ce travail.

Tout d'abord, ce travail ne pourrait exister sans aide et l'encadrement de Dr.Hichem Rahab, nous le remercions pour la qualité de son encadrement exceptionnelle, pour sa patience, sa précision et sa volonté lors de notre préparation de ce mémoire.

Mes sincères remerciements aux membres du jury d'avoir accepté de juger l'œuvre.

Nous remercions également tous nos professeurs pour leur générosité et leur grande patience, malgré leurs responsabilités

Nous tenons également à remercier tous ceux qui nous ont aidés de près ou de loin pour développer ce travail.

A tous ceux dont le soutien nous a été utile et nécessaire.

Je dédie ce travail à :

A ma mère,

A mon père,

A mes frères et sœurs,

A toute ma famille,

A tous ceux que j'aime,

Sans oublier tous les professeurs que ce soit du primaire, du moyen, du secondaire ou de l'enseignement supérieur.

Qu'ils trouvent ici l'expression de toute ma reconnaissance.

BOUTEARI Souad

Je dédie ce mémoire,

A mes très chers parents

Qui veillent sans cesse sur moi avec leurs prières et leurs recommandations. Que Dieu le tout puissant les protège et leur réserve une longue et meilleure vie.

A mes très chers frères et sœurs.

A toute ma famille

A mes chères amies

RICH Alla Eddine

Table de matière

Table de figures

Table de tableaux

Introduction générale

Depuis la découverte des ordinateurs, de nombreuses activités de la vie quotidienne ont été simplifiées. Aujourd'hui, les gens peuvent facilement traiter l'information à l'aide de logiciels et de réseaux informatiques. Compte tenu de son évolution, Internet contribue à faire de ce monde un meilleur endroit où on peut vivre avec une plate-forme unique d'innovation, de créativité et d'opportunités économiques. Cette technologie est importante car elle contribue à améliorer la qualité de vie des personnes du monde entier.

L'intelligence artificielle est la capacité d'une machine d'agir par lui-même et non explicitement programmé pour reproduire de la parole ou des tâches qui sont généralement des activités humaines. Aujourd'hui, on trouve l'intelligence artificielle et l'informatique dans les réseaux sociaux, les transports, et notamment dans le secteur médical. L'application de l'intelligence artificielle en médecine permet aux machines d'analyser des données et fournir des estimations dans le but de prédire de nombreuses maladies pour que les médecins puissent intervenir le plus rapidement possible pour réduire les risques de complications des maladies sur la santé du patient et lutter contre la propagation des maladies.

L'apprentissage automatique (également appelé apprentissage machine) est une branche de l'intelligence artificielle liée à la conception et au développement d'algorithmes permettant à l'ordinateur (une machine au sens large) d'apprendre à exécuter des tâches très complexes sans avoir été explicitement programmé et il s'améliore automatiquement avec l'expérience.

Dans ce mémoire, nous avons étudié la prédiction (l'estimation) médicale par apprentissage automatique via l'utilisation des algorithmes d'apprentissage supervisé pour la prédiction du diabète qui est un dysfonctionnement du système de régulation de la glycémie. On vise à réduire les risques des complications de cette maladie chronique sur la santé du patient.

Pour résoudre cette problématique nous avons définit le diagnostic médical comme un processus de classification. Cette formulation nous a permet l'utilisation des ordinateurs avec des capacités de calculs importantes pour effectuer cette tâche de prédiction. Cependant, la décision du médecin reste le facteur le plus important dans le diagnostic. Les systèmes de classement sont très utiles, ils réduisent les erreurs dues à la fatigue et au temps de diagnostic.

La méthode utilisée dans ce travail consiste à appliquer différents algorithmes d'apprentissage supervisé, à savoir ; K voisins les plus proches, les arbres de décision, la régression logistique, et les Séparateurs à vaste Marge SVM (Support Vector Machine), avec des données extraites de l'hôpital qui sont des données cliniques des patients. L'algorithme qui a abouti à la meilleure classification des patients en termes de taux de classification et de sensibilité du modèle est l'arbre de décision dans notre expérimentation.

Ce travail a été organisé en trois chapitres principaux comme suit :

Dans le premier chapitre, on va présenter un aperçu général sur la maladie du diabète, sa définition et ses différents types, symptômes, diagnostic et traitement de la maladie et quelques précautions pour se protéger du diabète. Ensuite, nous définissons l'insuline et ses différents types.

Dans le deuxième chapitre, nous avons essayé de démarrer une étude théorique sur l'apprentissage automatique et les algorithmes utilisés. Nous nous sommes également intéressés par les algorithmes d'apprentissage supervisé afin de les utiliser dans le dernier chapitre pour prédire le diabète.

Le dernier chapitre présente d'abord une étude technique dans laquelle nous détaillerons la base de données utilisée et l'environnement logiciel adopté pour construire notre modèle de prédiction. Ensuite, on va présenter les différentes techniques utilisées et la partie d'implémentation (test/split, k_folde, matrice de confusion prédiction) et les captures d'écran des algorithmes utilisés et la comparaison enter ces différents algorithmes.

Enfin, ce travail se termine par une conclusion générale qui résume les principales idées que nous avons apportées et les perspectives pour des futurs travaux.

CHAPITRE 1 :

Généralités sur le diabète

Dans le premier chapitre, nous présentons quelques concepts préliminaires liés au diabète, qui est une partie essentielle de sa prédiction dans les chapitres suivants.

1.1 Introduction

Le diabète est l'une des maladies les plus courantes dans le monde. Actuellement, on estime que 150 millions de personnes dans le monde souffrent de diabète. Malgré des décennies d'efforts de recherche et l'espoir de traitements radicaux voire préventifs, cette maladie continue de ne bénéficier que d'alternatives aux restrictions quotidiennes, faisant participer activement le patient à son travail de traitement. Dans ce chapitre, nous fournissons d'abord une introduction au diabète et à l'insuline et à ses types, puis nous donnons les deux complications ainsi que la façon de diagnostiquer le diabète et les différents types de diabète et comment prévenir ces complications ou maladies.

1.2 Définition de diabète

Le diabète sucre, plus simplement appelé diabète connue aussi sous le nom d'une maladie silencieuse. L'organisation mondiale de la santé (OMS) définit le diabète comme une maladie grave, à long terme (ou chronique), qui survient lorsque le taux de glycémie d'une personne est élevée parce que son organisme ne peut pas produire assez d'insuline, qu'il n'en produit pas ou qu'il ne peut pas utiliser efficacement l'insuline qu'il produit [1][2].

Plus d'explication sur le diabète

Lorsque nous mangeons, les aliments sont dégradés en glucose (sucre). Ce glucose fournit de l'énergie au corps afin qu'il puisse fonctionner correctement en puisant dans ses ressources. Pendant la digestion, le sang transporte le glucose dans tout le corps et vient alimenter les cellules. Cependant, pour que le sucre présent dans le sang puisse ensuite être transmis aux cellules, le corps a besoin d'insuline, une hormone sécrétée par le pancréas. L'insuline agit donc comme une clé permettant au glucose de passer du sang aux cellules de notre corps. Si le glucose reste dans le sang, la glycémie augmente. À long terme, cela peut entraîner le dysfonctionnement et la détérioration de nombreux organes comme les yeux et les reins [3].

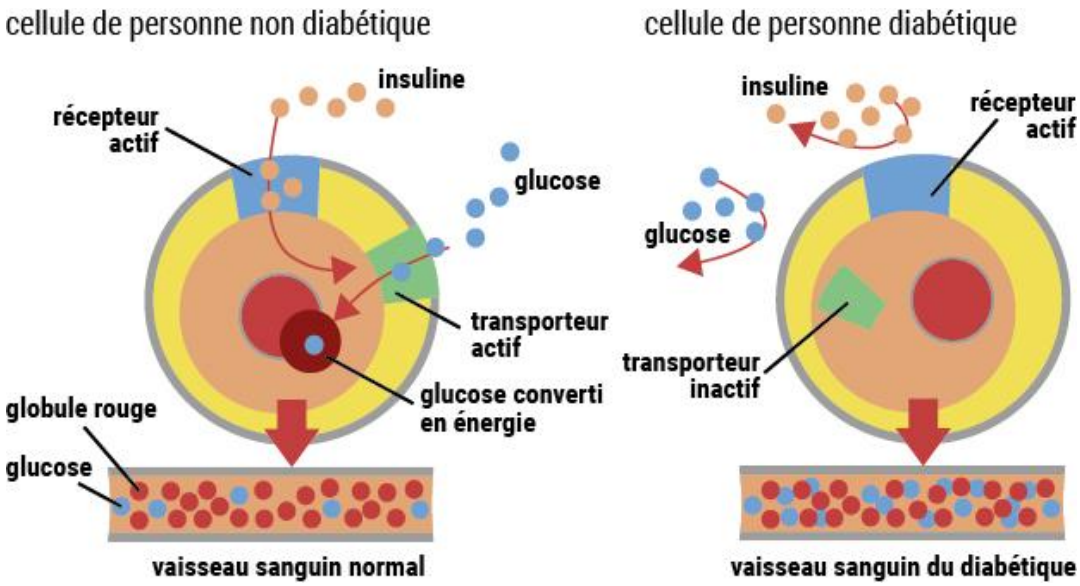

Figure 1:Insulinorésistance et insulinopénie[3]

1.3 Définition L'insuline

L'insuline est une hormone polypeptidique qui a un effet régulateur sur le métabolisme du glucose. Une insuline insuffisante conduit au diabète. L'insuline est fabriquée à partir de cellules bêta du pancréas, dans les îlots de Langerhans et transportée dans le sang. L'insuline permet au corps d'utiliser le glucose comme énergie [15].

1.3.1 Les différents types d'insuline :

Tous les types d'insuline produisent le même effet. Ils imitent les augmentations et les diminutions naturelles des niveaux d'insuline dans le corps pendant la journée. La composition des différents types d'insuline affecte la rapidité et la durée de leur action:[15]

- **Insuline à action rapide**: ce type d'insuline commence à agir environ 15 minutes après l'injection. Ses effets peuvent durer entre trois et quatre heures. Il est souvent utilisé avant un repas.
- **Insuline à action brève**: vous injectez cette insuline avant un repas. Il commence à agir 30 à 60 minutes après l'injection et dure cinq à huit heures.
- **Insuline à action intermédiaire**: ce type d'insuline commence à agir une à deux heures après l'injection et ses effets peuvent durer de 14 à 16 heures.
- **Insuline à action prolongée**: cette insuline peut ne commencer à agir que deux heures environ après l'injection. Ses effets peuvent durer jusqu'à 24 heures ou plus.

1.4 Diabètes et complications

Quel qu'en soit le type de diabète, ce dernier peut entraîner des complications à court terme (hypoglycémie, malaise...), et des complications à long terme (L'hyperglycémie) en cas de mauvais contrôle de la glycémie [3].

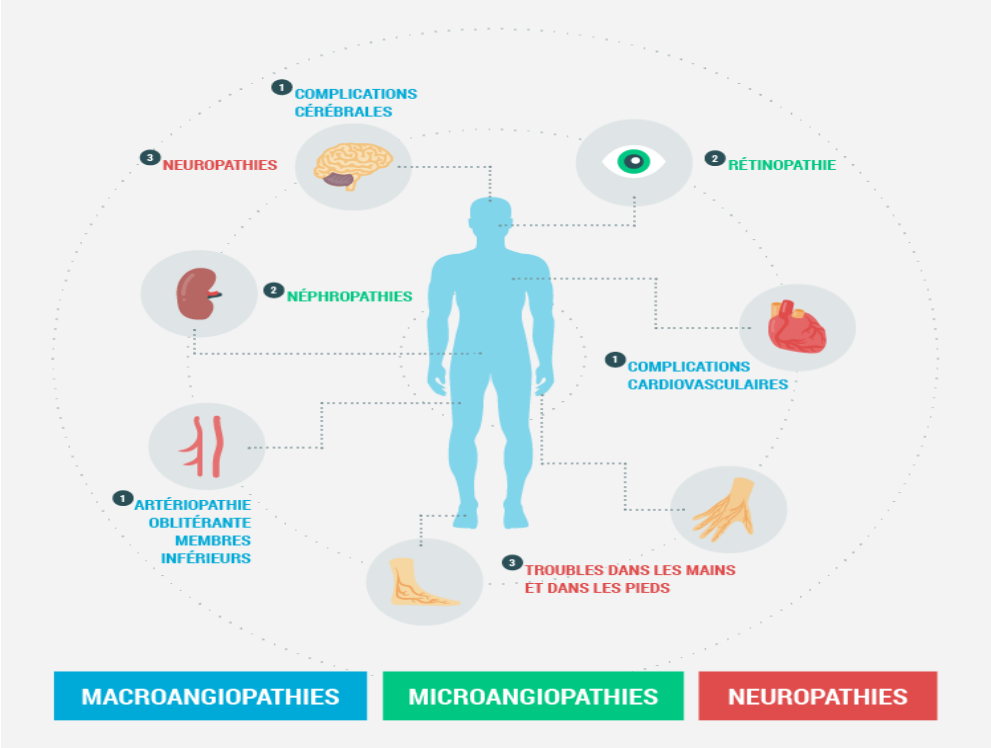

Figure 2: Diabètes et complications à long terme [4]

Il existe trois types de complications [5] :

1.4.1 Complications métaboliques :

On a d'abord le coma acidocetosique, le coma Hypoglycémique, et le coma hyperosmolairess.

- ***Acidocétose*** **:** C'est une complication grave du diabète, pouvant provoquer un coma. Elle survient lorsque l'organisme n'a plus assez d'insuline, et que le sang devient trop acide à Cause de la présence d'acétone.
- ***Coma hypoglycémique*** **:** Perte de conscience, relative à un manque de sucre dans le sang. Il peut être provoqué par une forte dose d'insuline ou après un effort violent. Il se traite par injection de glucagon.
- ***Comahyperosmolaire*** : concerne surtout le sujet âgé, le plus souvent diabétique de type 2, Non insulinodépendant. A l'occasion d'un déficit hydrique, des troubles de la conscience s'installent alors qu'apparaît une importante déshydratation.

1.4.2 Complications chroniques :

- ***Micro angiopathie*** **:** Observée après de longues périodes d'hyperglycémies, elle correspond à une atteinte des petits vaisseaux. Elle est à l'origine de lésions rétiniennes et rénales.
- ***Néphropathie*** **:** C'est l'atteinte des petits vaisseaux sanguins rénaux. L'excès de sucre va s'y déposer et perturber les flux sanguins donc le travail de filtration du rein. Cela aboutit à l'insuffisance rénale : le sang n'est plus « nettoyé » correctement.
- ***Neuropathie*** **:** C'est l'atteinte des nerfs périphériques et sensitifs. Elle peut se traduire par des douleurs aux membres inférieurs, des troubles de la sensibilité ou de la vidange gastrique. Elle est liée à l'hyperglycémie chronique.
- ***Rétinopathie*** **:** C'est l'atteinte des petits vaisseaux sanguins de la rétine. L'excès de sucre dans le sang abîme la paroi de ces vaisseaux, les rend perméable et des lésions apparaissent sur la rétine. Un traitement au laser peut stopper l'évolution de ces lésions.
- ***Macro angiopathie*** **:** Observée après de longues périodes d'hyperglycémie, elle correspond à une atteinte des gros vaisseaux. C'est l'équivalent de l'athérosclérose, mais ce phénomène est plus fréquent et plus étendu chez les personnes diabétiques mal équilibrées.

1.4.3 Complications infectieuses:

Le diabète favorise l'éclosion d'infections bactériennes et mycosiques, ce qui conduit à des gangrènes nécessitant l'amputation des membres

1.5 Diagnostic du diabète

Cette maladie silencieuse et indolore est détectée le plus souvent lorsque les complications à long terme s'expriment. Cette découverte peut notamment être brutale dans le cas de diabète de type 1 (pas de sécrétion d'insuline), allant jusqu'au coma diabétique.

Il est clair qu'il n'est pas toujours facile de savoir par soi-même si l'on est diabétique ou non, si vous constatez des symptômes multiples et/ou aigus, de vous adresser à un professionnel de la santé. En effet, c'est le taux de glucose dans le sang qui constitue le signe le plus manifeste de diabète. Le test sanguin sera effectué deux fois. Si le taux de glycémie est trop élevé dans les deux mesures, vous souffrez de diabète. Mais quand parle-t-on d'un taux de glycémie trop

élevé ? Le tableau ci-dessous donne un aperçu des valeurs de référence générales dans une situation normale avant et après un repas, ainsi que des valeurs pouvant indiquer une hypoglycémie ou une hyperglycémie. Ces valeurs ne fournissent qu'une indication générale ; elles peuvent varier d'une personne à l'autre et dépendent de la situation [7].

Tableau 1:Valeurs de référence lors d'une analyse à jeun Situation normale / Diabète

	Avant le repas	Valeurs de référence
Situation normale	2 heures après le repas	Entre 70 et 110 mg/dl Moins de 180 mg/dl
Diabète	2 heures après le repas	Plus de 126 mg/dl Plus de 200 mg/dl

Plus le taux de glucose dans le sang est élevé et plus cette augmentation dure longtemps, plus les symptômes seront nombreux et plus le risque de problèmes de santé graves sera élevé.

1.5.1 Qui est concerné par le dépistage du diabète ?

Toute personne ayant des membres de sa famille atteints de diabète de type 2 doit se faire dépister régulièrement car un risque héréditaire existe (si l'un des deux parents est diabétique de type 2, le risque héréditaire est de 40 % ; si les deux parents sont atteints, le risque monte à 70%). Pour le diabète de type 1, le risque de transmission aux enfants est de 6 % si le père est diabétique, 2 ou 3 % si la mère l'est, et 30 % si les deux parents sont atteints de diabète. Les personnes en surpoids ou souffrant de troubles de la glycémie doivent également se plier au dépistage.
Il en va de même pour les femmes ayant développé du diabète pendant leur grossesse (diabète gestationnel) ou ayant mis au monde un bébé de faible poids. Le dépistage est également recommandé aux personnes de plus de 65 ans [6].

1.5.2 Comment savoir si l'on est diabétique ?

La diagnostique du diabète se fait par un test de prise du sang mesurant la glycémie ou le taux de sucre sanguin, qui varie selon les apports alimentaires .il existe deux méthode de test : [3]

1. **Teste en laboratoire d'analyses médicales :** pour mesurer sa glycémie à jeun tous les 3 mois, son hémoglobine glyquée (HbA1c), c'est le meilleur indice de surveillance du diabète et de l'efficacité des traitements antidiabétiques. Ce marqueur permet de vérifier que le diabète est équilibré [8].

Figure 3:test-hémoglobine glyquée (HbA1c) [8]

2. **Auto- teste:** un lecteur de glycémie pour contrôler plusieurs fois par jour sur une goutte de sang à des moments précis, c'est ce qu'on appelle l'auto-surveillance. Le glucomètre permet à une personne de connaître le niveau de glycémie, généralement de petite taille, il s'agit d'un appareil de mesure transportable que le patient peut utiliser lui-même à domicile, au travail, etc. Même si son fonctionnement est assez simple, quelques précautions sont à prendre avant, durant, et après usage du lecteur de glycémie. [9]

Figure 4: Glucomètre [9]

1.5.3 Décoder et comprendre les résultats de la glycémie

Un résultat de test de glycémie correspond au taux de sucre dans le sang à un instant donné. La glycémie varie dans le temps, le résultat d'un test ne sera pas identique d'un jour à l'autre à la même heure et à différents moments de la même journée.

De même que deux mesures de la glycémie réalisées consécutivement ne seront pas forcément strictement identiques. Le dépistage du diabète peut être fait à tout moment de la journée, mais il est primordial de tester la glycémie après un minimum de 8 heures de jeûne. On considère qu'il y a présence de diabète si la glycémie est supérieure ou égale à 1,26 g/l a jeun et supérieure ou égale à 2 g/l après le repas. [3]

1.5.4 À quelle fréquence dois-je contrôler ma glycémie ?

Avant toute vérification, il est nécessaire de connaître vos objectifs glycémiques : à jeun et 2 heures après les repas (post- prandial). Votre médecin déterminera avec vous ces objectifs ainsi que la fréquence de mesure. Il n'existe pas de règle universelle. Toutefois, la Haute autorité de santé (HAS) recommande :

- **Pour le diabète de type 1 :** au moins quatre tests par jour. Les objectifs glycémiques sont fixés entre 70 et 120 mg/dl avant le repas et < 160 mg/dl en post-prandial.
- **Pour le diabète de type 2 :** dans tous les cas, les objectifs glycémiques sont fixés entre 70 et 120 mg/dl avant les repas et 180mg/dl en post-prandial. Selon le type de traitement, la fréquence est variable.
- **Pour le diabète gestationnel :** les objectifs sont stricts : à jeun < 0,95 g/l et < 1,20 g/l en postprandial. [10]

1.6 Classification du diabète

Selon l'Organisation Mondial de la Santé, la classification du diabète dépend principalement de cette classification étiologie et caractéristiques physiopathologiques en trois types : [3]

- ***Diabète de type 1***
- ***Diabète de type 2***
- ***Diabète gestationnel***

1.6.1 Diabète de type 1 (5-10% des patients)

Ce type de diabète apparaît en général chez le sujet jeune mais peut se développer à tout âge. L'étiologie exacte reste inconnue mais une pathologie auto-immune détruisant les cellules béta du pancréas est souvent évoquée, ainsi que des facteurs environnementaux et certains virus ou bactéries. Le pancréas ne produit plus du tout ou pas assez d'insuline ce qui provoque les symptômes classiques d'hyperglycémie [21].

1.6.2 Symptômes majeurs du diabète de type 1

Les principaux symptômes révélateurs, également appelés les signes cardinaux du diabète, sont présents dès le début de la maladie :

- **une polyurie importante**, c'est-à-dire une augmentation du volume des urines (2 ou 3 l/j, alors que chez un individu normal, il est compris entre 0,8 et 1,5 litre), imposant des mictions fréquentes
- **une polydipsie**, c'est-à-dire une soif excessive associée à une consommation très importante de liquide, comme si la boisson n'assouvissait pas la soif
- **une polyphagie**, c'est-à-dire une faim excessive avec une absence de sensation de satiété, malgré une absorption importante de nourriture
- **un amaigrissement**, quand le diabète est déjà installé depuis quelques semaines
- **une fatigue inexpliquée.**

Les signes majeurs du diabète de type 1 se manifestent souvent à l'occasion d'un épisode fébrile ou d'une infection virale, parfois lors d'un stress aigu [12].

Le traitement du diabète de type 1

Le traitement du diabète de type 1 a pour objectif de contrôler la glycémie. Il repose sur l'apport d'insuline**,** qui n'est plus fabriquée par le pancréas en quantité suffisante. L'enfant ou l'adolescent doit le suivre quotidiennement et toute sa vie, sans interruption. Le choix du type d'insuline (insulinothérapie) dépend de l'objectif défini avec le médecin de l'enfant pour le contrôle de la glycémie. Cet objectif est individuel et il est fonction de chaque situation personnelle. Il peut être modifié au cours de l'évolution du diabète. [13]

1.6.3 Diabète de type 2 (90-95% des patients)

Il peut apparaître à tout âge mais se développe en général chez les adultes d'âge moyen ou les personnes âgées pouvant déjà souffrir d'un syndrome métabolique (surpoids, obésité, dyslipidémie, hypertension...). L'étiologie est inconnue mais il apparaît plus fréquemment chez certaines ethnies ou après un diabète gestationnel. Le pancréas est en général encore fonctionnel (au moins au début) mais une production insuffisante d'insuline est observée ainsi qu'une résistance des cellules à l'action de celle-ci [21].

1.7 Les symptômes du diabète de type 2

L'hyperglycémie chronique est le plus souvent asymptomatique. Si le diabète est très déséquilibré, des symptômes peuvent apparaître et seront les signes d'une insulino - nécessitante, imposant un bilan médical rapide **:**

- Soif importante
- Envie d'uriner très fréquente (c'est le syndrome polyuro-polydipsique)
- Fatigue (asthénie) majeure
- Amaigrissement**. [**11]

Le traitement du diabète de type 2

Le traitement de référence du diabète de type 2 est l'optimisation des habitudes de vie une perte de poids si nécessaire, une activité physique régulière et une alimentation équilibrée peuvent être suffisants pour contrôler la glycémie dans un premier temps. En seconde intention, des antidiabétiques oraux et /ou injectables sont prescrits pour contrôler la glycémie. Lorsque le diabète évolue, il peut nécessiter la mise en place d'un traitement par insuline [10].

1.7.1 Diabète gestationnel (14% des femmes enceintes)

Ce diabète apparaît lors d'une grossesse. Il se développe une intolérance au glucose due à une sécrétion insuffisante d'insuline dans le cadre d'une résistance à l'action de celle-ci augmentée durant la grossesse. Ce diabète est en général asymptomatique d'où l'importance du dépistage chez la femme enceinte [21].

1.8 Symptômes du diabète gestationnel durant la grossesse

Le diabète gestationnel se déclare généralement à partir du sixième mois de grossesse et concerne un peu moins de 10 % des femmes enceintes. Certaines femmes enceintes ont un risque accru de développer cette forme particulière de diabète sucré, lorsqu'elles présentent un ou plusieurs facteurs de risque : [12]

- un surpoids ou une obésité
- une prise de poids importante au cours de la grossesse
- un âge supérieur à 35 ans
- des antécédents familiaux de diabète sucré
- des antécédents personnels de diabète gestationnel au cours d'une précédente grossesse

1.8.1 Le traitement du diabète gestationnel

Il faut absolument équilibrer un diabète**,** qu'il soit insulinodépendant ou non. La solution principale est la modification du régime alimentaire ou du mode de vie**.** Cela suffit généralement à maintenir la glycémie à un taux normal.

La future maman devra :

- Surveiller ses apports en glucides
- Manger moins d'aliments riches en gras saturés (beurre, crème...)
- Ne pas prendre trop de poids pendant la grossesse
- Augmenter son activité physique en faisant de l'exercice régulièrement
- Parfois prendre de l'insuline (en petite injection dans le ventre, la cuisse ou le bras) sous contrôle médical (et sous prescription médicale),

Les femmes concernées sont généralement prises en charge par une équipe de soignants : médecin, infirmière, sage femme, diététicienne.[14]

1.9 Prévention des complications du diabète

Pour parvenir à lutter efficacement contre le diabète, il est important de sensibiliser le public et les milieux professionnels aux facteurs de risque et aux symptômes. La prévention est assurée par des bilans régulière : [5]

- de l'état nutritionnel du malade
- de la glycémie
- du retentissement sur l'œil
- des retentissements cardio-vasculaires (surveillance de la TA, des troubles circulatoires, de la peau
- surveillance rénale et notamment de la glycosurie
- surveillance neurologique
- Tout diabétique doit être pris en charge par une diététicienne qui établit un régime adapté à sa pathologie

1.10 Conclusion

Le diabète présente une source d'inquiétude grandissante dans le domaine de la santé publique. Il pose un vrai problème dans les pays développés et aussi celle en développement et sera dû à l'accroissement démographique, au vieillissement de la population, à des régimes alimentaires déséquilibrés, à l'obésité et à un mode de vie sédentaire. Comme nous avons indiqué en introduction, il n'y a pas un traitement radical pour cette maladie mais le traitement a pour but d'éviter la survenue des complications dégénératives et métaboliques aigues. Dans ce chapitre nous avons présenté la maladie du diabète, leur différent types, les symptômes ainsi que le diagnostic et le traitement de la maladie et à la fin nous avons cité quelques préventions pour évité cette maladie

CHAPITRE 2 :

L'apprentissage automatique

Dans ce deuxième chapitre, nous avons essayé de démarrer une étude théorique sur l'apprentissage automatique et les algorithmes utilisés. Nous nous sommes également concentrés sur les algorithmes d'apprentissage supervisé afin de les utiliser dans le dernier chapitre pour prédire le diabète.

2 Chapitre 2 : L'apprentissage automatique

2.1 Introduction

L'apprentissage automatique (ou machine Learning) est un branche de l'intelligence artificielle performante dédié à la résolution de problèmes divers, qui peuvent aller du filtrage d'une collection de photos aux défis mondiaux les plus urgents (en termes de santé et d'environnement, par exemple). Dans ce chapitre, nous fournissons d'abord une définition de l'intelligence artificielle qui est à l'origine de l'apprentissage automatique, puis nous donnons une définition de l'apprentissage automatique avec une mention de ses types et les algorithmes spécifiques qu'il utilise. On va expliqués dans ce chapitre les algorithmes d'apprentissage supervisé et ceux de l'apprentissage non supervisé.

2.2 Définition de l'intelligence artificielle

L'intelligence artificielle (IA) est le nouveau terme que nous entendons à chaque fois ces dernières années. L'intelligence artificielle détermine généralement la capacité d'une machine d'agir en soi et non explicitement programmé pour reproduire des activités et tâches qui sont généralement liées au comportement humain. L'apprentissage automatique est une discipline de l'intelligence artificielle qui s'efforce de trouver un moyen de créer des programmes informatiques qui s'améliorent automatiquement avec l'expérience [3].

2.3 Apprentissage automatique

La définition de l'apprentissage automatique selon Wikipedia (septembre 2020) est : « L'apprentissage automatique (en anglais machine Learning, littéralement « apprentissage machine ») ou apprentissage statistique 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 d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. Plus largement, il concerne la conception, l'analyse, l'optimisation, le développement et l'implémentation de telles méthodes. » [16].

Figure 5:l'apprentissage automatique [17]

2.4 Les types d'apprentissage automatique

Il existe plusieurs façons d'apprendre automatiquement à partir des données dépendamment des problèmes à résoudre et des données disponibles. La **Figure 6** donne un sommaire des types d'apprentissage automatique les plus connus [18].

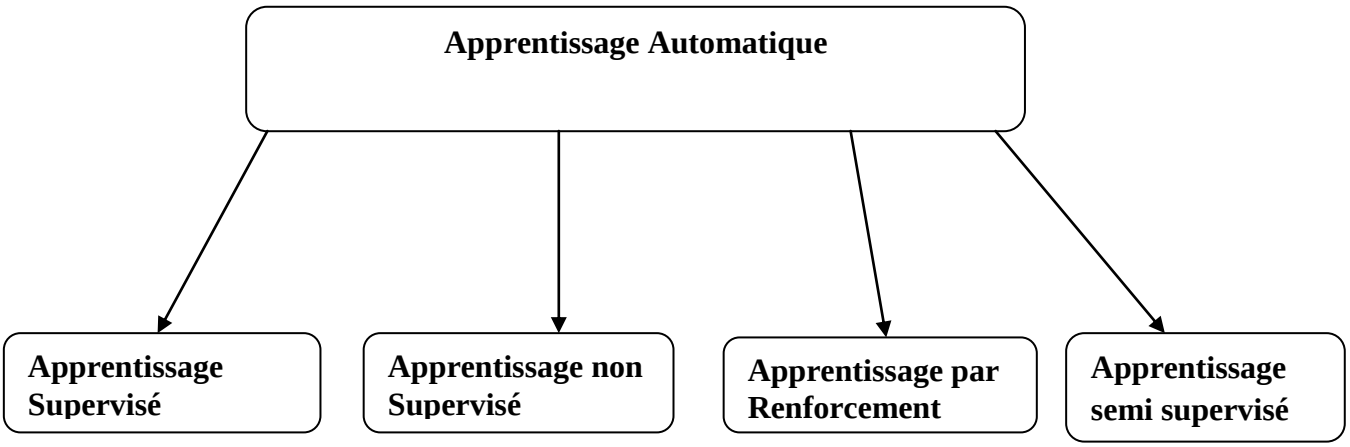

Figure 6: Les grandes classes d'apprentissage automatique [18]

2.4.1 Apprentissage Supervisé

L'algorithme est entrainé en utilisant une base de données d'apprentissage contenant des exemples de cas réels traités et validés. L'objectif est de trouver des corrélations entre les données d'entrée (variables explicatives) et les données de sorties (variables à prédire), pour ensuite inférer la connaissance extraite sur des entrées avec des sorties inconnues. Chaque

exemple, appelé aussi instance, est un couple d'entrée-sortie (x_n, y_n), $n \in [1, N]$ avec $x_n \in X$ et $y_n \in Y$ et Où :

X est l'ensemble d'attributs (discrets ou continues).

Y est l'ensemble des valeurs de sortie (la variable cible ou dépendante) [18].

Plus clairement : [3]

$D_{donn\,ée}$=$\{\{x_1, y_1\}, \ldots\ldots, \{x_n, y_n\}\}$ un ensemble de données fini. On ait $f(x_n) = y_n$

Phase 1 : l'ensemble d'apprentissage ou base d'apprentissage.

$D_{entr\,ée}$= $\{\{x_1, y_1\}, \ldots\ldots, \{x_i, y_i\}\}$ et $D_{entr\,ée} \in D_{donn\,ée}$

i : indice de donnée

x : donnée et y : classe ou étiquète de donnée.

Phase 2: La création de modèle ou la fonction de prédiction l'algorithme d'apprentissage reçoit$D_{(x_i,y_i)}$entrée et construit un modèle Ou bien

Une fonction de prédiction $f(x_i) = y_i$

Phase 3 : phase de test on test la qualité de modèle sur un ensemble de variables étiquetées qu'on désigne par :

$$D_{test} = \{(x_{i+1}, y_{i+1}); \ldots\ldots; (x_n, y_n)\}$$

$$D_{test} \in D_{donnée}$$

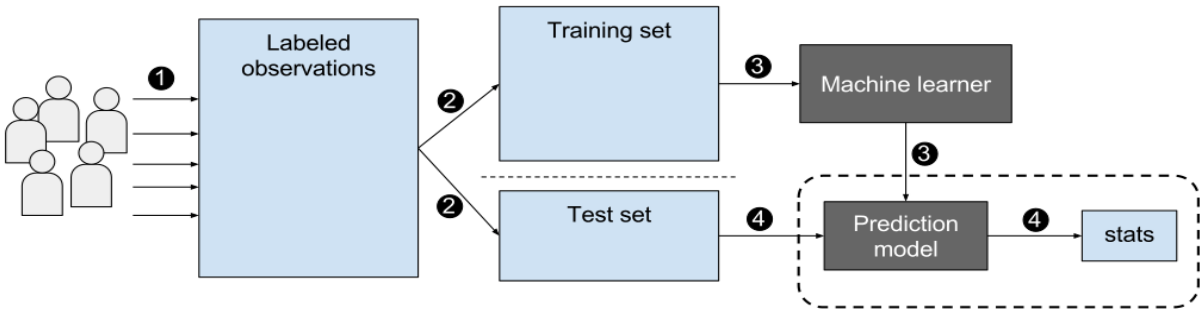

Figure 7: Workflow d'un apprentissage supervisé [3]

En apprentissage supervisé, on distingue entre deux types de tâches :

- **la classification** : quand la variable cible (à prédire) est discrète,$Y = \{1, \ldots . I\}$ Ce qui revient à attribuer une classe (ou étiquette) à chaque entrée. C'est le cas si on cherche à prédire la tendance d'un mouvement futur d'un actif (haut, neutre, bas).
- **la régression** : Un modèle de régression permet de prédire une valeur quantitative. Cela signifie que l'ensemble des valeurs de sortie Y qu'on essai d'estimer avec la fonction f est un ensemble de réels.

2.4.2 Apprentissage non Supervisé

L'apprentissage non-supervisé (ou classification automatique). Quand le système ou l'opérateur ne disposent que d'exemples, mais non d'étiquettes, et que le nombre de classes et leur nature n'ont pas été prédéterminés, on parle d'apprentissage non supervisé ou clustering. Aucun expert n'est requis. L'algorithme doit découvrir par lui-même la structure plus ou moins cachée des données. Le partitionnement de données, data clustering en anglais, est un algorithme d'apprentissage non supervisé. Le système doit ici dans l'espace de description (la somme des données) cibler les données selon leurs attributs disponibles, pour les classer en groupe homogènes d'exemples. La similarité est généralement calculée selon une fonction de distance entre paires d'exemples. [19]

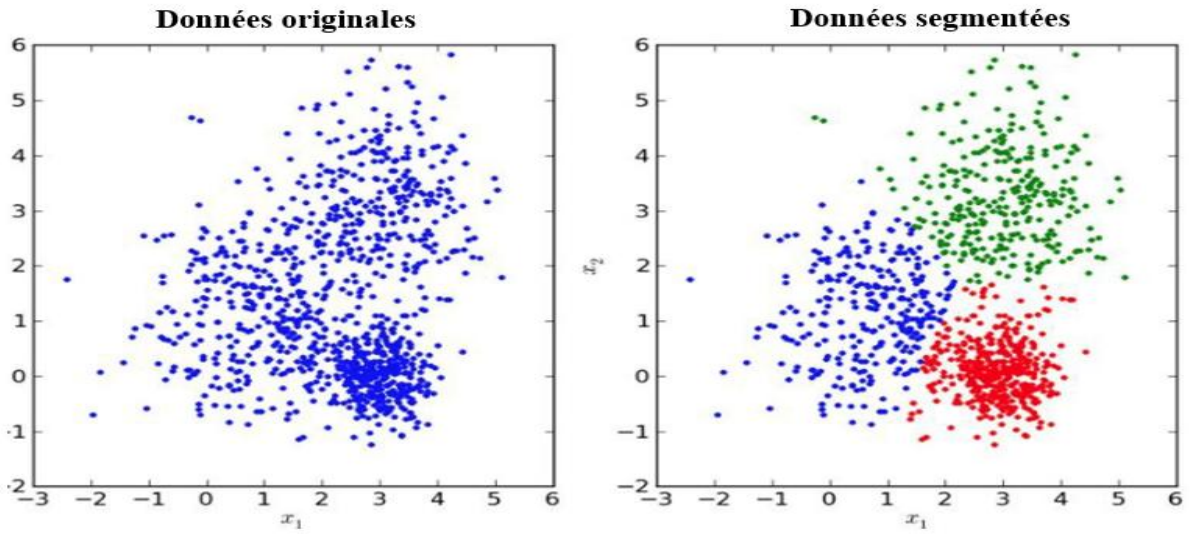

Figure 8: Exemple d'apprentissage non supervisé [18]

2.4.3 L'apprentissage par renforcement

L'apprentissage se fait sans supervision, par interaction avec l'environnement (principe d'essai / erreur) et, en observant le résultat des actions prises. Chaque action de la séquence est associée à une récompense. Le but est de déterminer la stratégie comportementale optimale afin de maximiser la récompense totale. Pour cela, un simple retour des résultats est nécessaire pour apprendre comment la machine doit agir. Ceci est appelé le signal de renforcement. Il peut être très avantageux pour la prévision financière à haute fréquence où l'environnement est dynamique et en conséquence, il est difficile de trouver ou d'automatiser manuellement des stratégies efficaces [18].

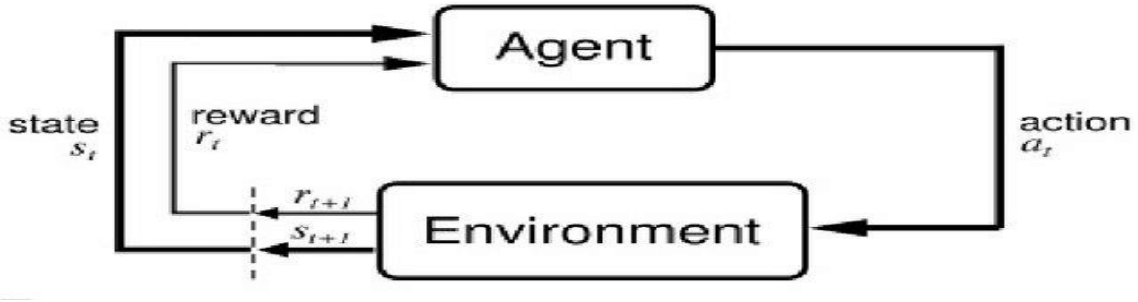

Figure 9: L'apprentissage par renforcement [20]

2.4.4 Apprentissage semi supervisé

Il s'agit d'un mixe entre l'apprentissage supervisé et non supervisé en utilisant des données. L'avantage d'utiliser cette approche réside dans le fait que l'étiquetage de données peut être couteux et prend souvent beaucoup de temps. En plus, il pourra entrainer un biais humain dans les données étiquetées. Dans ce cas, l'apprentissage semi-supervisé, qui ne nécessite que quelques étiquettes, est très pratique. Et le fait d'inclure un grand nombre de données non étiquetées au cours du processus d'entrainement a tendance à améliorer la performance du modèle final tout en réduisant le temps et les coûts consacrés à la construction de données étiquetées et non-étiquetées pour le même ensemble de données [18].

2.5 Les algorithmes de l'apprentissage automatique utilisés

2.5.1.1 K plus proche voisins(KNN)

C'est un algorithme utilisé dans des structures d'apprentissage automatique, souvent appelé algorithme de machine learning. L'idée est d'utiliser un grand nombre de données afin "d'apprendre à la machine" à résoudre un certain type de problème. Nourrie par un grand nombre d'exemples, elle va apprendre et devenir de plus en plus performante (l'algorithme de k plus proches voisins ne nécessite pas de phase d'apprentissage à proprement parler, il faut juste stocker le jeu de données d'apprentissage) [16].

KNN est un algorithme qui ne fait aucune hypothèse sur la structure des données et de la distribution, ce qui signifie qu'il s'agit d'un algorithme non paramétrique. Il est également appelé 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 classifiassions 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 a la même classe que la classe la plus proche des nombre des points K voisins [3].

Principe de l'algorithme : [16]

On suppose que l'ensemble $\boldsymbol{E}$ contient $\boldsymbol{n}$ données labellisées et une autre donnée n'appartenant pas a $\boldsymbol{E}$ qui ne possède pas de label. Soit $\boldsymbol{d}$ une fonction qui renvoie la distance (qui reste à choisir) entre la donnée u et une donnée quelconque appartenant à $\boldsymbol{E}$. Soit un entier $\boldsymbol{K}$ inferieur ou égal à $\boldsymbol{n}$.

Le principe de l'algorithme de k-plus proches voisins est le suivant:

- On calcule les distances entre la donnée u et chaque donnée appartenant à E à l'aide de la fonction d.
- On retient les Kdonnées du jeu de données E les plus proches de u.
- On attribue à u la classe qui est la plus fréquente parmi les k données les plus proches.

Distance des *k* plus proches voisins [22]

Mesures souvent utilisées pour la distance $dist(x_i, x_j)$

- **La distance euclidienne (valeurs continues)**

  La distance euclidienne calcule la racine carrée de la différence entre les cordonnées d'une paire d'objets (points ou classes). Si on considère deux points A et B, de cordonnées respectives (XA,A) et (XB,YB) [15] [17], la distance euclidienne est données par :

- **La distance de *Manhattan* (valeurs continues)**

Si on considère encore deux points A et B, de cordonnées respectives (XA,A) et (XB,YB) , la distance de Manhattan est définie par : $DistAB=|XB-XA|+|YB-YA|$.

- **La distance de *Hamming* (valeurs discrètes)**

  C'est la distance entre deux point données par la différence maximale entre leur cordonnées. Maintenant, on considère deux points A et B, de cordonnées respectives $(X_1, X_2, \dots, X_n)$et $(Y_1, Y_2, \dots, Y_n)$, la distance de Tchebychev est définie par :
  $DistAB=$ max$i\in$ [0,](|$Xi-Yi$|)

2.5.2 Arbre de décision

Un arbre de décision est un outil de classification et prédiction pour représentant un ensemble de choix sous la forme graphique d'un arbre. Les différentes décisions possibles sont situées aux extrémités des branches (les « feuilles » de l'arbre), et sont atteints en fonction de décisions prises à chaque étape. L'arbre de décision est un outil utilisé dans des domaines variés tels que la sécurité, la fouille de données, la médecine, etc. Il a l'avantage d'être lisible et rapide à exécuter. Il s'agit de plus d'une représentation calculable automatiquement par des algorithmes d'apprentissage supervisé. [21]

2.5.3 Séparateurs à Vaste Marge (SVM)

Les Séparateurs à Vaste Marge (SVM) souvent traduit par l'appellation de Séparateur à Vaste Marge sont une classe d'algorithmes d'apprentissage initialement définis pour la discrimination ; c'est-à-dire la prévision d'une variable qualitative initialement binaire. Ils ont

été ensuite généralisés à la prévision d'une variable quantitative. Dans le cas de la discrimination d'une variable dichotomique, ils sont basés sur la recherche de l'hyperplan de marge optimale qui, lorsque c'est possible, classe ou sépare correctement les données tout en étant le plus éloigné possible de toutes les observations. Le principe est donc de trouver un classifier, ou une fonction de discrimination, dont la capacité de généralisation (qualité de prévision) est la plus grande possible.

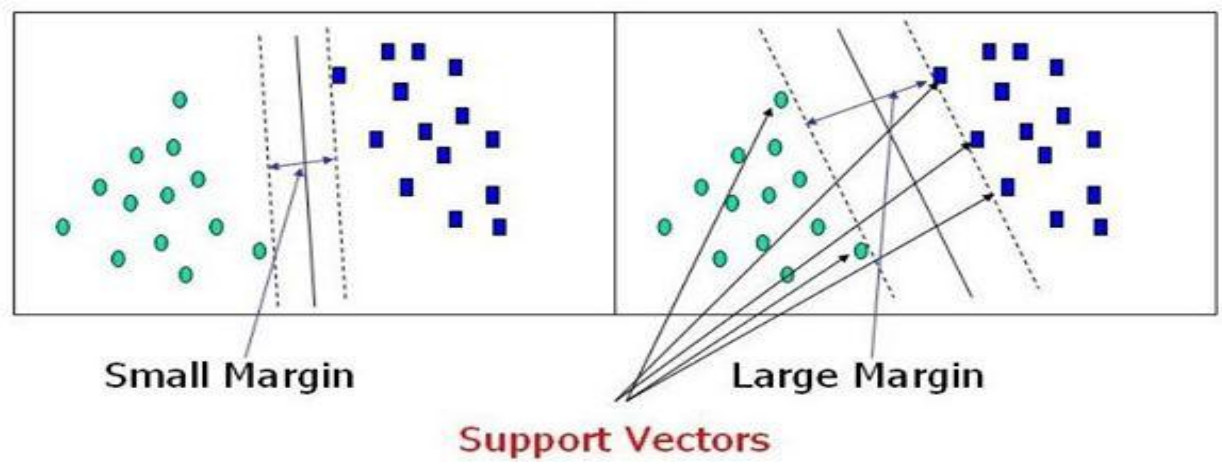

Figure 10:L'apprentissage par renforcement [20]

2.5.4 Naïves Bayes

Naïve bayes fait partie des algorithmes d'apprentissage automatique supervise qui sont principalement utilisés pour la classification. C'est un classificateur probabiliste simple basé sur l'application de théorème de bayes et qui aide à construire des modèles d'apprentissage automatique rapides qui peuvent faire des prédictions rapides. Naïve bayes, dans l'algorithme, se réfère à l'hypothèse naïve que l'algorithme fait, qui est que chaque fonctionnalité est indépendante des autres fonctionnalités [3].

2.5.5 Réseaux de neurones

Le terme réseau de neurones est une référence à la neurobiologie. Originalement, ce concept est inspiré du fonctionnement des neurones du cerveau humain, apprend essentiellement de l'expérience. Un réseau de neurones est une organisation hiérarchique de neurones connectés entre eux. Ces derniers transmettent un message ou un signal à d'autres neurones en fonction des paramètres d'entrés reçus et forment un réseau complexe

Un réseau de neurones est en général composé d'une succession de couches (Layer en anglais). Ce modèle comporte trois ensembles de règles : multiplication, sommation et activation. Les données d'entrées (Input data en anglais) sont consommées par les neurones de la première couche cachée (Hidden layers en anglais). Chaque couche peut avoir un ou plusieurs neurones. La connexion entre deux neurones de couches successives aurait un poids associé qui définit l'influence de l'entrée sur la sortie du neurone suivant et éventuellement sur la sortie finale globale. Sur chaque neurone artificiel, chaque valeur d'entrée est multipliée par le poids correspondant [22].

$$y = F(\sum_{i=0}^{m} w_i \, . \, x_i + b)$$

− w_i représentent les poids.

− x_i représentent les données d'entrées.

− b représente le biais.

− F représente la fonction d'activation.

− y représente la valeur de sortie

2.5.6 Régression

Les méthodes de régression s'appliquent lorsque le résultat que l'on cherche à estimer est une valeur continue. En machine learning ML, la régression est un outil important de l'apprentissage supervisé pour la modélisation et l'analyse des données. Elle est notamment utilisée en statistique et en économie [22].

2.5.6.1 Régression linéaire

On appelle modèle de régression tout modèle capable à établir une relation linéaire entre une variable, dite expliqué ou dépendante, et une ou plusieurs variables, dite explicatives ou variables indépendantes.

2.5.6.2 Régression logistique

La régression logistique a été développée par le statisticien David Cox en 1958. Le modèle logistique est un modèle statistique qui utilise une fonction logistique, il est également

utilisé dans les problèmes de classification. La régression logistique ne nécessite pas de relation linéaire entre les variables dépendantes et indépendantes, mais elle nécessite des échantillons de grande taille afin d'avoir plus de précision lors de l'estimation de la vraisemblance.

2.5.7 Les méthodes par ensemble

Ces techniques sont des méta-algorithmes qui consistent à combiner plusieurs modèles uniques de base, comme les arbres de décision, dans un même modèle prédictif. L'objectif est d'améliorer la généralisation et la robustesse de nos modèles. En effet, statistiquement parlant, la moyenne d'un ensemble d'échantillons est plus fiable que celle d'un seul échantillon. Les méthodes d'ensemble peuvent être divisées en deux catégories : les méthodes d'ensemble parallèles et les méthodes d'ensemble séquentielles [22].

2.5.7.1 Méthodes d'ensemble parallèles (Bagging)

Pour ces méthodes, les modèles de base (exemple : les arbres de décision) sont générés de façon indépendante et en parallèle (par exemple, les forêts aléatoires). La motivation derrière ces méthodes est que l'erreur de prédiction peut être réduite de manière significative en réduisant la variance [22].

2.5.7.2 Méthodes d'ensemble séquentielles (Boosting)

Pour ces méthodes, les classifieurs de base sont générés de manière séquentielle (par exemple AdaBoost) et dépendante, contrairement aux méthodes parallèles. À chaque fois qu'un classifieur de base est entrainé, les instances mal classées précédemment sont pondérées avec un poids plus élevé, dans le but que lors des prochaines itérations, les nouveaux modèles corrigent les erreurs des modèles précédents, ce qui devra améliorer la performance globale [22].

2.6 Apprentissage non supervisé

Dans l'apprentissage non supervisé il n'y a pas de valeurs de sortie, il s'agit de trouver des structures cachées à partir d'un ensemble de données qui doivent être regroupé d'où le terme «clustering ». Le but de ce type d'apprentissage est de séparer les données en groupes ou en catégories [22].

Le clustering est une technique d'apprentissage automatique non supervisé, utilisé pour le regroupement des données non étiquetées dans de nombreux domaines. Si on dispose d'un nombre fini de points de données et on cherche à les classer dans des groupe de sorte que chaque groupe contient des points de données ayant des propriétés et/ou caractéristiques similaires. Le problème principal qui se pose dans ces algorithmes c'est le choix des propriétés à prendre en compte au cours du regroupement. L'un des algorithmes de clustering les plus utilisés est le « *K*-Means ».[22]

2.6.1 *K*-Means :

C'est l'algorithme de classification le plus connu. Son principe est simple, facile à comprendre et à implémenter dans un code. Tout d'abord, on sélectionne un certain nombre de groupes puis, aléatoirement, on initialise le centre associée à chaque groupe. Il est préférable de commencer par analyser globalement les données présentes et essayer d'identifier des groupes distincts afin de mieux déterminer le nombre de classes à utiliser [22].

2.6.2 T-distributed stochastic neighbor embedding (T-SNE):

C'est une technique linéaire non supervisée, développée par Laurens Van der Martens et Geoffrey Hinton en 2008. T-SNE est une méthode de réduction de dimension, elle transforme la représentation de données multidimensionnelles en deux ou trois dimensions et donne, par conséquent, une idée sur la façon dont les données sont disposées dans un espace de grande dimension. T-SNE trouve des modèles à partir des données en identifiant des groupes (clusters) contenant les données qui partagent des caractéristiques similaires [22].

2.7 Conclusion

Dans ce chapitre, nous avons expliqué les algorithmes d'apprentissage automatique. C'est un outil puissant pour faire des prédictions et analyser et décrire les données dans divers types des problèmes. Parmi ces problèmes celui que nous étudierons dans chapitre 3. Cela nous aide à détecter l'apparition précoce du diabète, c'est a dire à prédire une maladie, ce qui peut aidera à réduire les risques et des complications de cette maladie sur la santé du patient. Dans la suite de cette étude, l'objectif principal est d'appliquer les algorithmes ; K voisins les

plus proches, les arbres de décision, les Séparateurs à Vaste Marge, et la régression logistique pour la classification et la prédiction de diabète.

CHAPITRE 3

Prédiction du diabète par Algorithmes l'apprentissage automatique

Dans ce troisième chapitre, on va appliquer les algorithmes d'apprentissage automatique (supervisé) pour prédire le diabète. On va déterminer le meilleur algorithme en termes de performances pour ce type de problème.

3 Chapitre 3 : prédiction du diabète par algorithmes d'apprentissage automatique

3.1 Introduction

Dans ce dernier chapitre de notre mémoire nous présentons notre contribution dans la prédiction médicale sur le cas du diabète. Nous présentons d'abord une étude de l'origine des données utilisées, le Dataset Pima, et nous décrivons ses caractéristiques. On va également présenter et identifié tous les outils et packages utilisés dans l'implémentation des algorithmes. Quatre algorithmes d'apprentissage automatique ont été appliqués. Nous avons également choisi le meilleur algorithme parmi les algorithmes appliqués. Une prédiction est faite avec les modèles entraînés avec des données entrées pour lesquelles les classes sont connues à l'avance. Il est nécessaire de savoir comment relier les entrées aux sorties afin que les sorties puissent être prédites à l'avenir pour toute nouvelle entrée.

3.2 Définition de Dataset utilisée

C'est un ensemble de données sur le diabète, extrait de l'Institut national de diabète et de diagnostique de maladies rénales indiens Pima (the National Institute of Diabetes and Digestive and Kidney Diseases). L'objectif de l'ensemble de données est de prédire de façon diagnostique si un patient est atteint ou non du diabète, en fonction de certaines données cliniques incluses dans l'ensemble de données. L'ensemble de données se compose de plusieurs variables prédictives médicales et d'une variable cible (classe) `<Outcome>`. La valeur '0' indique un diagnostic négatif (il n'y a pas de diabète) pour la valeur '1' il est question d'un cas positif (il y a du diabète).

Les variables dans un data set :

- Pregnancies : La condition d'être enceinte
- Glucose : Concentration de glucose plasmatique à 2 heures dans un test de tolérance au glucose par voie orale.
- BloodPressure : Pression artérielle diastolique (mm Hg).
- SkinThickness : Epaisseur de pli cutané du triceps (mm).

- Insulin : Hormone polypeptidique qui régule le métabolisme des glucides (insuline sérique de 2 heures (mu U/ml).
- BMI : indice de masse corporelle (poids en kg/(taille en m) ^{2}).
- DiabetesPedigreeFunction : Fonction généalogique du diabète.
- Age : age du patient
- Outcome : variable de classe (0 ou 1) ou 0 indique que le patient ne souffre pas de diabète et 1 indique que le patient est diabétique.

3.2.1 Description du Dataset

Tableau 2:Description des variables d'ensemble dataset

Variable	**Description de variable**	**Type de variable**
Pregnancies	Nombre de fois enceinte	int64
Glucose	Le glucose est un sucre simple de formule moléculaire $C_6H_{12}O_6$. Le glucose est le monosaccharide le plus abondant, une sous-catégorie de glucides.	int64
BloodPressure	Si un TA diastolique >90 signifie une pression artérielle élevé (probabilité élevé de diabète) Un TA diastolique < 60 signifie une pression artérielle base (mois probabilité de diabète)	int64
SkinThickness	Nous avons mesuré l'épaisseur de la peau chez 66 patients atteints de DSID âgés de 24 à 38 ans et recherché si elle était corrélée au contrôle glycémique à long terme et à la présence de certains	int64

Insulin	Hormone polypeptidique qui régule le métabolisme des glucides (insuline sérique de 2 heures (mu U/ml)).	int64
BMI	(poids en kg / taille en m2) IMC de 18.5 à 20 c'est normal IMC entre 25 et 30 situer dans une plage surpoids Et de 30 ou plus situer dans la fourchette d'obésité	float64
DiabetesPedigreeFunction	DiabetesPedigreeFunction : fonction généalogique du diabète (une fonction qui évalue la probabilité de diabète en fonction des antécédents familiaux) Âge : âge (années) Résultat : variable de classe (0 si non diabétique, 1 si diabétique)	float64
Age	âge du patient	Int64
Outcome	caractéristique cible 0 = négative (non diabétique) 1= positive (diabétique)	Int64

3.3 Définition des outils utilisés

3.3.1 Googlecolab (Colaboratoire Google)

Colaboratory, souvent raccourci en "Colab", est un produit de Google Research. Colab permet à n'importe qui d'écrire et d'exécuter le code Python de son choix par le biais du navigateur. C'est un environnement particulièrement adapté à l'apprentissage automatique, à l'analyse de données et à l'éducation. En termes plus techniques, Colab est un service hébergé

de notebooks Jupyter qui ne nécessite aucune configuration et permet d'accéder gratuitement à des ressources informatiques, dont des GPU [23].

3.3.2 Python

Python est un langage de programmation multi-paradigme et le langage de programmation dominant dans la science des données (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 orant une pléthore de librairies de très grande qualité, couvrant tous les types d'apprentissages disponibles qui combine la facilité d'utilisation et d'apprentissage avec la puissance des librairies qu'elles possèdent.
Parmi ces bibliothèques, nous avons utilisé :

Numpy

NumPy est une bibliothèque pour langage de programmation Python, destinée à manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableaux. Plus précisément, cette bibliothèque logicielle libre et open source fournit de multiples fonctions permettant notamment de créer directement un tableau depuis un fichier ou au contraire de sauvegarder un tableau dans un fichier, et manipuler des vecteurs, matrices et polynômes [24].

Pandas

Pandas est aussi une bibliothèque Python fournissant des structures de données rapides, flexibles et expressives conçues pour rendre le travail avec des données structurées (tabulaires, multidimensionnelles, potentiellement hétérogènes) possède une fonctionnalité importante nettoyage des 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 négatif énorme sur notre modèle [25].

Matplotlib

Matplotlib peut être utilisé pour créer des graphiques. La bibliothèque est généralement utilisée comme suit :

1. Appelez une fonction de traçage avec des données (par exemple .plot()).
2. Appelez de nombreuses fonctions pour configurer les propriétés du tracé (par exemple, les étiquettes et les couleurs).
3. Rendre l'intrigue visible (par exemple .show()).[3]

Scikit-learn

Scikit-learn (Sklearn) est la bibliothèque la plus utile et la plus robuste pour l'apprentissage automatique en Python. Il fournit une sélection d'outils efficaces pour l'apprentissage automatique et la modélisation statistique, y compris la classification, la régression, le clustering et la réduction de la dimensionnalité via une interface de consistance en Python. Cette bibliothèque, qui est en grande partie écrite en Python, est construite sur la base de NumPy, SciPy et Matplotlib [26].

3.4 visualisations de données

La visualisation permet de comprendre la composition de données afin d'obtenir les meilleurs résultats des algorithmes d'apprentissage automatique. Les visualisations de données fournissent des informations clés sur des ensembles de données complexes de manière significative et intuitive, elle aide à voir des choses n'étaient pas évidentes. La visualisation facilité la transmission des informations de façon universelle et facilite le partage d'idées avec les autres.

3.4.1 Charger et affiche le fichier dataset.CSV

```
#pour Charger le fiche CSV
df = pd.read_csv("/content/diabetes-dataset.csv")
x = np.array(df.drop(['Outcome'], 1))
y = np.array(df['Outcome'])
peek = df.head()
#pour affiche le fiche csv
print(peek )
```

Figure 11:Algorithme pour charger et affiche dataset (capture d'écran)

3.4.2 Statistiques descriptives

Les statistiques descriptives constituent une bonne idée de ce à quoi ressemble chaque attribut. Fonction **`describe ()`** sur l'objet pandas répertorie 8 propriétés statistiques pour chaque attribut. Ce sont :Compter, moyenne, écart-type, valeur minimum, 25e centile, 50e centile (médiane), 75e centile, valeur maximum.

```
import numpy as np
import pandas as pd
from google.colab import files
df = pd.read_csv("/content/diabetes-dataset.csv")
x = np.array(df.drop(['Outcome'], 1))
y = np.array(df['Outcome'])
#pour les information dataset le Statistiques descriptives
description = df.describe()
print(description)
```

Figure 12:Algorithme Statistiques descriptives pour dataset

Résultats :

	Pregnancies	Glucose	...	Age	Outcome
count	2000.000000	2000.000000	...	2000.000000	2000.000000
mean	3.703500	121.182500	...	33.090500	0.342000
std	3.306063	32.068636	...	11.786423	0.474498
min	0.000000	0.000000	...	21.000000	0.000000
25%	1.000000	99.000000	...	24.000000	0.000000
50%	3.000000	117.000000	...	29.000000	0.000000
75%	6.000000	141.000000	...	40.000000	1.000000
max	17.000000	199.000000	...	81.000000	1.000000

Figure 13: statistiques descriptives du Dataset

3.4.3 Histogrammes

Les Histogrammes sont un moyen rapide d'avoir une idée de la distribution de chaque attribut dans le Dataset. La fonction **`hist ()`** nous permet d'avoir faire une idée rapide sur les attributs, si un attribut est gaussien et asymétrique, ou même une a une distribution exponentielle.

```
import numpy as np
import pandas as pd
from matplotlib import pyplot
from google.colab import files
df = pd.read_csv("/content/diabetes-dataset.csv")
x = np.array(df.drop(['Outcome'], 1))
y = np.array(df['Outcome'])
#  Histogramme
df.hist()
pyplot.show()
```

Figure 14:Algorithme Histogrammes pour Dataset

Résultats :

Nous pouvons voir que peut-être les attributs Grossesses, âge et Insuline peuvent avoir un effet exponentiel Distribution. Nous pouvons également voir que les attributs Glucose, et BMI et Pression artérielle peuvent avoir un Distribution gaussienne ou presque gaussienne. Ceciest intéressant car de nombreux apprentissages automatiques les techniques supposent une distribution univariée gaussienne sur les variables d'entrée.

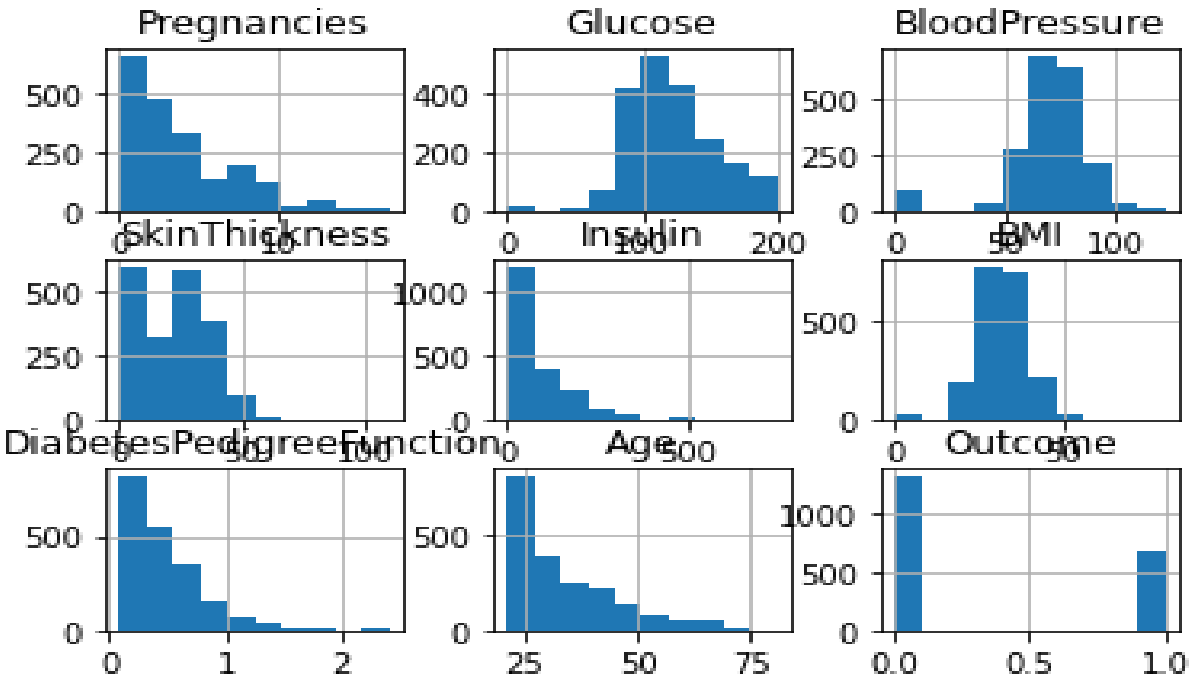

Figure 15: Résultats d'Algorithme Histogrammes

3.4.4 Diagrammes de densité

Les tracés de densité sont un autre moyen d'avoir une idée rapide de la distribution de chaque attribut. Les tracés ressemblent à un histogramme abstrait avec une courbe lisse tracée en haut de chaque casier,

Résultats d'Algorithme Diagrammes de densité :

```
import numpy as np
import pandas as pd
from matplotlib import pyplot
from google.colab import files
df = pd.read_csv("/content/diabetes-dataset.csv")
x = np.array(df.drop(['Outcome'], 1))
y = np.array(df['Outcome'])
#  Diagrammes de densité
df.plot(kind='density', subplots=True, layout=(3,3), sharex=False)
pyplot.show()
```

Figure 16:Code de Diagrammes de densité

Résultats :

Nous pouvons voir que la distribution pour chaque attribut est plus claire que les histogrammes.

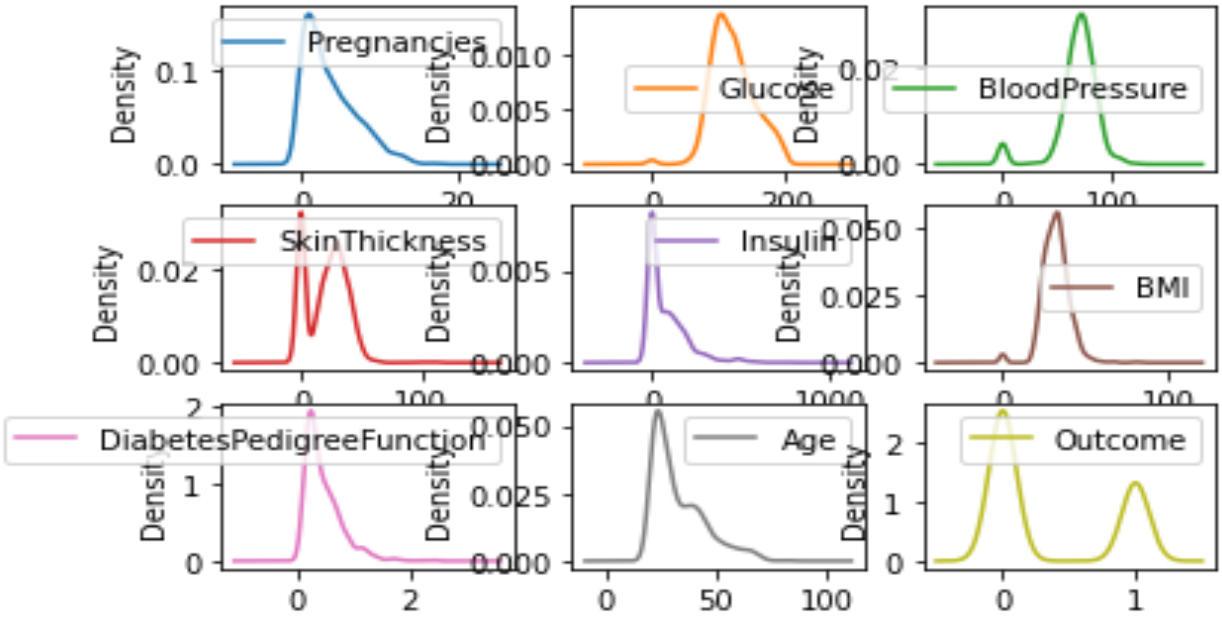

Figure 17:Résultat d'Algorithme Diagrammes de densité

3.5 Algorithme utilise

Dans cette étude, nous avons utilisé quatre algorithmes d'apprentissage supervisé, ces algorithmes sont:

- Les arbres de décision,
- Séparateurs à Vaste Marge (SVM),
- Régression logistique,
- K plus proche voisins (KNN).

3.6 Implémentation

Nous présenterons dans cette section les algorithmes d'apprentissage supervisé utilisés dans la prédiction du diabète. Nous adoptons le taux de classification (Accuracy) comme un facteur de comparaison entre les algorithmes d'apprentissage supervisé.

3.6.1 La division (Train/Test Split)

- **Méthode 01 : Train/Test Split**

Avant d'utiliser l'algorithme, nous avons divisé les données de Dataset en deux parties ; la première partie d'entraînement pour entraîner le modèle et la deuxième partie pour tester le modèle le modèle et évaluer ses performances.

Dans la première étape, nous avons divisé les données comme suit : Utilisez 90 % des données pour apprentissage et 10% pour les tests.

Dans la deuxième étape, nous avons réparti les données comme suit : 80 % des données utilisées pour apprentissage et 20% pour les tests.

- **Méthode 02 : Validation croisée (k_plis ou fold)**

Dons cette partie, nous avons divisé l'ensemble de données en k-parties (étape 1 k = 10, étape 2 k = 5, et étape 3 k=15). Chaque division des données est appelée un fold. Toutes les données aient une chance d'entrainement et de test. Par ce que le processus est répété de sorte que chaque pli (fold) de l'ensemble de données ait une chance d'appartenir à l'ensemble de test retenu. Le résultat de précision de chaque étape des algorithmes selon chaque méthode sont calculés.

3.6.2 Matrice de confusion

Dans cette partie, la matrice de confusion est utilisée. C'est une matrice bien connue dans le domaine de l'apprentissage automatique, utilisée pour tester les performances des algorithmes. Ce tableau contient des informations et des détails sur les évaluations réelles (évaluées par des humains) et les évaluations prédictives attendues par le classificateur.

Chaque colonne du tableau représente la catégorie attendue et chaque ligne représente la catégorie réelle.

Tableau 3:Matrice de confusion

	positive	négative
Positive	TP True Positive	FP False Positive
négatif	FN False négatif	TN True Négative

La signification de TP, TN, FP et FN est comme suit :

- **TP**: True Positive.
- **TN** : True Negative.
- **FN** : False négatif.
- **FP** : False Positive.

TP : signifie qu'une personne est réellement diabétique et elle a été prédite qu'elle est diabétique.

TN : signifie qu'une personne est réellement non diabétique et elle a été prédite qu'elle est non diabétique.

FP : signifie qu'une personne est réellement non diabétique et elle a été prédite qu'elle est diabétique.

FN: signifie qu'une personne est réellement diabétique et elle a été prédite qu'elle est non diabétique.

Pour vous donner une idée rapide de la précision du modèle on utilise un certain nombre de mesures.

- **Précision** : capacité du modèle de classification à ne renvoyer que des cas Lié, défini comme le nombre de vrais positifs divisé par le nombre de vrais positifs Positif plus le nombre de faux positifs.

$$précision = \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FP}}$$

- **Rappel** : (recall) est la capacité du modèle de classification identifier tous les cas pertinents, définis comme le nombre de vrais positifs Divisé par le nombre de vrais positifs plus le nombre de faux négatifs.

$$\text{recall} = \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FN}}$$

- **Score F1 :** une échelle unique qui combine rappel et précision avec La moyenne harmonique, en tenant compte des deux échelles dans l'équation suivante.

$$\mathbf{F1} = 2 * \frac{\mathbf{Précision} * \text{recall}}{\mathbf{Précision} + \text{recall}}$$

3.6.3 La prédiction

Dans cette partie, la prédiction se fait en saisissant les données du patient, ces donnée comporte : Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age.

Et puis il apparaît que le patient souffre de diabète ou non, de sorte que les sorties sont un message négatif, signifiant que le patient ne souffre pas de diabète, ou positif, lorsque le patient souffre du diabète

3.7 Arbre de décision

3.7.1 Méthode 01 : Division Train/Test

```
import numpy as np
import pandas as pd
from sklearn import model_selection
from sklearn.tree import DecisionTreeClassifier
from sklearn import model_selection
from io import StringIO
from google.colab import files
df = pd.read_csv("/content/diabetes-dataset.csv")
x = np.array(df.drop(['Outcome'], 1))
y = np.array(df['Outcome'])
#Split into datset  Train 80% and Test 10%
x_train, x_test, y_train, y_test = model_selection.train_test_split(x, y, test_size=0.1)
#used the DecisionTree model
Tree_model = DecisionTreeClassifier()
#the fit 'training function
Tree_model =Tree_model.fit(x_train, y_train)
#calculating the accuracy of the DecisionTree algorthm
print("decision tree model accuracy :",Tree_model.score(x_test, y_test) )
```

Figure 18:Algorithme Arbre de décision Train/Test Split

Résultat 1 :

Résultats de cet algorithme (les arbres de décision) : nous avons divisé les données comme suit : 90 % des données pour l'entraînement et 10% pour le test.

```
decision tree model accuracy : 1.0
```

Résultat 2 :

Résultats de cet algorithme (arbre de décision) : nous avons réparti les données comme suit : 80 % des données utilisées pour entraînement et 20% pour les tests.

```
decision tree model accuracy : 0.9725
```

Donc la meilleure division des données est : 90 % des données pour entraînement et 10% pour les tests.

3.7.2 Méthode 02 : Validation crois (k_fold)

```
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from google.colab import files
from sklearn.linear_model import LogisticRegression
df = pd.read_csv("/content/diabetes-dataset.csv")
x = np.array(df.drop(['Outcome'], 1))
y = np.array(df['Outcome'])
kfold = KFold(n_splits=10, random_state=7)
Tree_model= DecisionTreeClassifier()
results = cross_val_score(Tree_model, x, y, cv=kfold)
print("kfold n_splits=10 ")
print("Accuracy:" ,results.mean())
```

Figure 19:Algorithme Arbre de décision k_fold

Résultat : k-fold
K=10

```
kfold n_splits=10
Accuracy: 0.9914999999999999
```

K=5

```
kfold n_splits=5
Accuracy: 0.992
```

K=15

```
kfold n_splits=15
Accuracy: 0.9909924063891071
```

Matrice de confusion

```
from sklearn.metrics import classification_report, confusion_matrix
print("Matrice de confusion")
print(confusion_matrix(y_test,y_pred))
print(classification_report(y_test,y_pred))
```

Figure 20:Algorithme Matrice de confusion

Résultat :

	positive	négative
Positive	**125**	**0**
négative	**0**	**75**

	precision	recall	f1-score	support
0	1.00	1.00	1.00	125
1	1.00	1.00	1.00	75
accuracy			1.00	200
macro avg	1.00	1.00	1.00	200
weighted avg	1.00	1.00	1.00	200

Figure 21: Matrice de confusion Arbres de décision

La prédiction

```
#the predicion of model  decision tree  from a new dataset
li=[]
print('Please enter your patient data....')
x=input("Donner le taux Pregnancies :")
li.append(x)
x=input("Donner le taux Glucose :")
li.append(x)
x=input("Donner le taux BloodPressure :")
li.append(x)
x=input("Donner le taux SkinThickness :")
li.append(x)
x=input("Donner le taux Insulin :")
li.append(x)
x=input("Donner le taux BMI:")
li.append(x)
x=input("Donner le taux DiabetesPedigreeFunction :")
li.append(x)
x=input("Donner le taux Age :")
li.append(x)
new_patiant=Tree_model.predict(np.array([li]))
if  new_patiant== 0:
  print(" result the predicion :", new_patiant ," negative the patient dose not have diabetes ")
elif new_patiant == 1:
  print("result the predicion:", new_patiant , "positive the patient has diabetes ")
```

Figure 22:Algorithme Arbre de décision La prédiction

Résultats de prédiction

Pour exécution de cet algorithme, on demande les données d'un patient. Les données du patient ont été saisies afin de savoir si ce patient souffre diabète.

```
Please enter your patient data....
Donner le taux Pregnancies : 7
Donner le taux Glucose : 152
Donner le taux BloodPressure : 88
Donner le taux SkinThickness : 44
Donner le taux Insulin : 0
Donner le taux BMI:50
Donner le taux DiabetesPedigreeFunction : 0.337
Donner le taux Age : 36
```

L'algorithme estime que la personne a le diabète :

```
Résultat de predicion:[1] positive, the patient has diabetes.
```

3.8 Séparateurs à Vaste Marge (SVM)

3.8.1 Méthode 01 : Train/Test Split

```
import numpy as np
import pandas as pd
from sklearn import preprocessing, neighbors, svm
from sklearn.svm import SVC
from sklearn import svm
from sklearn import model_selection
from io import StringIO
from google.colab import files
df = pd.read_csv("/content/diabetes-dataset.csv")
x = np.array(df.drop(['Outcome'], 1))
y = np.array(df['Outcome'])
#Split into datset  Train 90% and Test 10%
x_train, x_test, y_train, y_test = model_selection.train_test_split(x, y, test_size=0.1)
### used the Support Vector Machine
SVM_model = svm.SVC()
#the fit 'training function
SVM_model.fit(x_train, y_train)
#calculating the accuracy of the Support Vector Machine algorthm
print(" Support Vector Machine Modele Accuracy:", SVM_model.score(x_test, y_test))
```

Figure 23:Algorithme séparateurs à vaste marge Train/Test Split

Résultats de l'algorithme séparateurs à vaste marge (SVM) : nous avons divisé les données comme suit : 90 % des données pour entraînement et 10% pour les tests.

```
Support Vector Machine Modele Accuracy: 0.775
```

Résultat 2 :

Résultat de cet algorithme séparateurs à vaste marge(SVM) nous avons réparti les données comme suit : 80 % des données utilisées pour entraînement et 20% pour test.

```
Support Vector Machine Modele Accuracy: 0.7475
```

3.8.2 Méthode 02 : Validation crois (k_fold)

```
df = pd.read_csv("/content/diabetes-dataset.csv")
x = np.array(df.drop(['Outcome'], 1))
y = np.array(df['Outcome'])
kfold = KFold(n_splits=10, random_state=7)
SVM_model = svm.SVC()
results = cross_val_score(SVM_model, x, y, cv=kfold)
print("kfold n_splits=10 ")
print("Accuracy:" ,results.mean())
```

Figure 24:Algorithme séparateurs à Vaste Marge k_folde

Résultat : k_folde

K=10

```
kfold n_splits=10
Accuracy: 0.7685
```

K=5

```
kfold n_splits=5
Accuracy: 0.7705
```

K=15

```
kfold n_splits=15
Accuracy: 0.770504619758351
```

Matrice de confusion (SVM)

	positive	négative
positive	**100**	**29**
négative	**55**	**16**

	precision	recall	f1-score	support
0	0.65	0.78	0.70	129
1	0.36	0.23	0.28	71
accuracy			0.58	200
macro avg	0.50	0.50	0.49	200
weighted avg	0.54	0.58	0.55	200

Figure 25:Résultats de matrice de confusion d'algorithme (svm)

Résultats de prédiction (SVM)

Pour exécution de cet algorithme, on demande les données de cet patient. Les données du patient ont été saisies afin que ce patient soit testé :

```
Please enter your patient data....
Donner le taux Pregnancies : 4
Donner le taux Glucose : 115
Donner le taux BloodPressure : 72
Donner le taux SkinThickness : 0
Donner le taux Insulin : 0
Donner le taux BMI:28.9
Donner le taux DiabetesPedigreeFunction : 0.376
Donner le taux Age : 46
```

Le résultat de cet algorithme est incorrect, en fait la personne a le diabète, mais le résultat de algorithme est que la personne n' a pas de diabète.

```
Result the predicion : [0]  negative, the patient does  not
have diabetes.
```

3.9 Régression logistique

3.9.1 Méthode 01 : Train/Test Split

```
import numpy as np
import pandas as pd
from sklearn import model_selection
from sklearn.linear_model import LogisticRegression
from sklearn import model_selection
from io import StringIO
from google.colab import files
df = pd.read_csv("/content/diabetes-dataset.csv")
x = np.array(df.drop(['Outcome'], 1))
y = np.array(df['Outcome'])
#Split into datset  Train 90% and Test 10%
x_train, x_test, y_train, y_test = model_selection.train_test_split(x, y, test_size=0.1)
# used the LogisticRegression
LR_model= LogisticRegression()
#the fit 'training function
LR_model.fit(x_train, y_train)
#calculating the accuracy of the LogisticRegression  algorthm
print("logistic regression accuracy", LR_model.score(x_test, y_test))
```

Figure 26:Algorithme régression logistique Train/Test Split

Résultat 1 :

Résultats de l'algorithme **régression logistique**: nous avons divisé les données comme suit : 90 % des données pour entraînement et 10% pour les tests.

```
logistic regression accuracy 0.77
```

Résultat 2 :

Résultat de cet algorithme **régression logistique** nous avons réparti les données comme suit : 80 % des données utilisées pour entraînement et 20% pour les tests.

```
logistic regression accuracy 0.8
```

3.9.2 Méthode 02 : Validation crois (k_fold)

```
df = pd.read_csv("/content/diabetes-dataset.csv")
x = np.array(df.drop(['Outcome'], 1))
y = np.array(df['Outcome'])
kfold = KFold(n_splits=10, random_state=7)
LR_model= LogisticRegression()
results = cross_val_score(LR_model, x, y, cv=kfold)
print("kfold n_splits=10 ")
print("Accuracy:" ,results.mean())
```

Figure 27:Algorithme Régression logistique k_folde

Résultat : k_fold

K=10

```
kfold n_splits=10
Accuracy: 0.7755000000000001
```

K=5

```
kfold n_splits=5
Accuracy: 0.7779999999999999
```

K=15

```
kfold n_splits=15
Accuracy: 0.7775072008379156
```

Matrice de confusion (Régression logistique)

	positive	négative
positive	**205**	**64**
négative	**95**	**36**

	precision	recall	f1-score	support
0	0.68	0.76	0.72	269
1	0.36	0.27	0.31	131
accuracy			0.60	400
macro avg	0.52	0.52	0.52	400
weighted avg	0.58	0.60	0.59	400

Figure 28: Matrice de confusion d'Algorithme logistique Régression

3.10 K plus proche voisins (kNN)

3.10.1 Méthode 01 : Train/Test Split

```
import numpy as np
import pandas as pd
from sklearn.neighbors import KNeighborsClassifier
from sklearn import model_selection
from google.colab import files
df = pd.read_csv("/content/diabetes-dataset.csv")
x = np.array(df.drop(['Outcome'], 1))
y = np.array(df['Outcome'])
#Split into datset  Train 90% and Test 10%
x_train, x_test, y_train, y_test = model_selection.train_test_split(x, y, test_size=0.1)
###  used the K Nearest Neighbor
KNN_model = neighbors.KNeighborsClassifier()
#the fit 'training function
KNN_model.fit(x_train, y_train)
#calculating the accuracy of the K Nearest Neighbor
print("K Nearest Neighbor model Accuracy",KNN model.score(x test, y test))
```

Figure 29 : Algorithme K plus proche voisins Train/Test Split

Résultats 1 :

Résultats de l'algorithme K plus proche voisins (kNN) : nous avons divisé les données comme suit : 90 % des données pour entraînement et 10% pour les tests.

```
K Nearest Neighbor model Accuracy 0.795
```

Résultats 2 :

Résultats de l'algorithme K plus proche voisins (kNN) nous avons réparti les données comme suit : 80 % des données utilisées pour entraînement et 20% pour les tests.

```
K Nearest Neighbor model Accuracy 0.8
```

3.10.2 Méthode 02 : Validation crois (k_folde)

```
df = pd.read_csv("/content/diabetes-dataset.csv")
x = np.array(df.drop(['Outcome'], 1))
y = np.array(df['Outcome'])
kfold = KFold(n_splits=10, random_state=7)
KNN_model= neighbors.KNeighborsClassifier()
results = cross_val_score(KNN_model, x, y, cv=kfold)
print("kfold n_splits=10 ")
print("Accuracy:" ,results.mean())
```

Figure 30:Algorithme K plus proche voisins k_folde

Résultats

K=10

```
kfold n_splits=10
Accuracy: 0.805
```

K=5

```
kfold n_splits=5
Accuracy: 0.8275
```

K=15

```
kfold n_splits=15
Accuracy: 0.7965024501552389
```

Matrice de confusion(KNN)

	positive	négative
positive	**104**	**14**
négative	**27**	**55**

	precision	recall	f1-score	support
0	0.79	0.88	0.84	118
1	0.80	0.67	0.73	82
accuracy			0.80	200
macro avg	0.80	0.78	0.78	200
weighted avg	0.80	0.80	0.79	200

Figure 31:Matrice de confusion d'Algorithme KNN

3.11 Comparaison enter les algorithmes

Tableau 4: comparaison de performance enter les quatre algorithmes

	10%-90%			**20%-80%**		
	Précision	Rappel	Accuracy	Précision	Rappel	Accuracy
DT	1.0	1.0	1.0	0.99	0.99	0.972
SVM	0.65	0.78	0.775	0.79	0.88	0.747
LR	0.68	0.76	0.77	0.80	0.88	0.8
KNN	0.80	0.88	0.79	0.83	0.88	0.8

Dans cette étude, les quatre algorithmes sont utilisés dans l'étude de la prédiction du diabète et leurs performances sont comparées. Deux méthodes ont été utilisées, la division (Train/Test Split) et k_fold, et à partir du tableau ci-dessus, le modèle d'arbre de décision a obtenu la meilleure taux de classification (Accuracy) dans les deux méthodes.
Nous sélectionnons le modèle d'arbre de décision comme le modèle le plus optimal et qui fonctionne mieux pour notre ensemble de données en raison de son importante précision, rappel et taux de classification.

Tableau 5 : Validation croisée (k_plis ou fold)

	15-fold			**10-fold**			**5-fold**		
	Précision	Rappel	Accuracy	Précision	Rappel	Accuracy	P	R	ACC
DT	0.96	0.99	0.990	0.98	0.99	0.991	0.96	0.99	0.992
SVM	0.77	0.92	0.771	0.78	0.92	0.769	0.77	0.92	0.770
LR	0.81	0.90	0.777	0.81	0.91	0.775	0.81	0.91	0.777
KNN	0.84	0.88	0.796	0.84	0.88	0.805	0.84	0.88	0.827

3.12 Conclusion

Dans ce chapitre, nous avons présenté et expliqué les données sur le diabète (définition et visualisations de données) et définir tous les logiciels et bibliothèques utilisées. Nous avons également appliqué des algorithmes d'apprentissage à savoir ; K voisins les plus proches(KNN), les Arbres de décision séparateurs à Vaste Marge (SVM), et la régression logistique afin de classifier et de prédire le diabète. Le taux de classification chacun des algorithmes utilisés a été calculé afin de choisir le meilleur algorithme. D'après notre étude comparative, nous avons trouvé que le meilleur modèle est l'arbre de décision.

Conclusion générale

Le diabète reste l'une des maladies que nous rencontrons fréquemment et qui provoque une augmentation de la glycémie. En fait, il est primordial de consacrer un effort pour mieux comprendre et reconnaître son mécanisme et ses causes. La prédiction du diabète fait partie des applications et problématiques rencontrées fréquemment dans le domaine médical. Mais une approche d'apprentissage automatique peut aider à résoudre ce problème. Le but de cette étude est de construire un modèle prédictif pour un problème critique, à savoir, le diagnostic automatique de diabète en utilisant des algorithmes d'apprentissage automatique.

Dans ce mémoire nous avons étudie la prédiction médicale par apprentissage automatique. Nous avons-nous concentrés sur l'étude le cas de diabète.

En premier chapitre, nous avons présenté la maladie du diabète, ses différent types, les symptômes ainsi que le diagnostic et le traitement de la maladie et à la fin nous avons cité quelques préventions pour se protéger contre cette maladie.

Dans le deuxième chapitre, nous avons présenté les algorithmes d'apprentissage automatique. L'apprentissage automatique, ou machine learning, constitue un outil puissant pour faire les analyses et décrire les données dans divers types des problèmes.

Le troisième chapitre explique les données sur le diabète (définition et visualisations de données) et décrit tous les outils et bibliothèques que nous avons utilisé dans ce travail. Nous avons appliqué quatre algorithmes d'apprentissage automatique à savoir ; K voisins les plus proches, Arbres de décision, Machine à vecteurs de support, et régression logistique. Afin de classifier et prédire le diabète, la précision de chacun des algorithmes utilisés a été calculée afin de choisir le meilleur algorithme. Après la comparaison, nous avons trouvé que le meilleur modèle est obtenu avec les arbres de décision.

Perspectives

Nous suggérons, comme suite de ce mémoire, l'exploration d'autres algorithmes qu'on n pas utiliser dans ce travaille et aussi l'apprentissage profond (Deep learning). Aussi nous proposons l'utilisation de plus importantes Datasets pour valider mieux les modèles propo

Bibliographie

[1]. International Diabète Fédération. L'ATLAS DU DIABÈTEDE LA FID9ème Édition 2019.176p.

[2]. Organisation mondiale de la santé. (2016).Rapport mondial sur le diabète.88p

[3]. Sidahmed Amel, Rabhi Karima,La prédiction du diabète en utilisant les algorithmes de machine learning, Université AMO de Bouira, année2019/2020,136p

[4]. CEED : Centre européen d'étude du Diabète.et complications.[en Ligne].Disponible sur : http : //ceed-diabete.org/fr/le-diabète/diabète-et complications/ (Consulté 25/04/2021)

[5]. Ammar Mohammed, *Reconnaissance Automatique Du Diabète Et Prédiction de la dose d'insuline***,**UNIVERSITE ABOU BAKR BELKAID–TLEMCEN, année 2008-2009,134p.

[6]. TopSante. Maladies chroniques. [en ligne].Disponible sur :https : //www.topsante.com/medecine/maladies-chroniques/diabete/commentsavoir-si-je-suis-diabétique-609768 (consulté 26/04/2021 / 10 :40).

[7]. ACCU-CHEK , AI-JE DEABETE .[en ligne]. Disponible sur : https://www.accu-chek.be/fr/lessentiel-sur-le-diabète/comment-savoir-si-vous-etes-diabetique (consulté 26/04/2021 / 12 :27).

[8]. SANTE.JOURNAL DES FEMMES. .[en ligne]. Disponible sur : https://sante.journaldesfemmes.fr/fiches-anatomie-et-examens/2517769-hemoglobine-glyquee-norme-prise-de-sang/(Consulté 26/04/2021).

[9]. GlucoGuide.top,[en ligne]. Disponible sur :https://www.glucoguide.top/comment-utiliser-un-glucometre/#A_quoi_ressemble_un_lecteur_de_glycemie/(Consulté 26/04/2021).

[10]. Federation des Française Diabetiques, [en ligne]. Disponible sur : http://www.federationdesdiabetiques.org/diabete/glycemie(Consulté 26/04/2021)

[11]. Diabète de type 2 (diabète sucré) - Définition, symptômes et traitements - Doctissimo https://www.doctissimo.fr/ /html/sante/encyclopedie/sa_1290_diab_02.htm

[12]. [12]ooreka SANTE .[en ligne]. Disponible sur : https://diabete.ooreka.fr/755423/rubrique/755679/symptomes-et-diagnostic-du-diabete (Consulté 27/04/2021 / 23 :25)

[13]. ameli.fr,l'Assurance Maladie, [en ligne]. Disponible sur : https://www.ameli.fr/assure/sante/themes/diabete-type-1-enfant-adolescent/traitement

[14]. aufeminin. en ligne]. Disponible sur :https://www.aufeminin.com/grossesse/diabete-gestationnel-risques-et-traitements-du-diabete-gestationnel-s653193.html

[15]. oraquaptail.com.Définition L'insuline ,[en ligne]. Disponible sur :https://www.aquaportail.com/definition-2890-insuline.html(Consulté 3/05/2021 /11 :48)

[16]. A.LOUGHANI + JNB, Algorithme des k-plus proches voisins,acdémielille,page 136R

[17]. Kaspersky daily, [en ligne]. Disponible sur :https://media.kasperskydaily.com/wp-content/uploads/sites/93/2016/11/06094159/machine-learning-featured.jpg

[18]. Rachid MIFDAL, Application des techniques d'apprentissage automatique pourLa prédiction de la tendance des titres financiers,ÉCOLE DE TECHNOLOGIE SUPÉRIEUREUNIVERSITÉ DU QUÉBEC,2 NOVEMBRE 2019,132p

[19]. Le Parisien, SENS a GENT DICTIONNAIRE, APPRENTISSAGE AUTOMATIQUE, [enligne]. Disponiblesur :http://dictionnaire.sensagent.leparisien.fr/Apprentissage%20automatique/fr-fr/(Consulté 7/05/2021 /11 :48)

[20]. researchgate,https://www.researchgate.net/figure/Principe-de-lapprentissage-par-renforcement-Whiteson-2010-I2-Definition-de_fig1_343268354

[21]. Recommandations pour la pratique clinique, https://www.recodiab.ch/ RPC1_types.pdf

[22]. GUENNINECHE Amel.Prédiction des propriétés des matériaux par apprentissage automatique.UNIVERSITE ABOU-BEKR BELKAID – TLEMCEN. 29/06/2019.58P.

[23]. Colaboratory,[enligne].Disponiblesur : https://research.google.com/colaboratory/faq.html?source=post_page(Consulté 8/06/2021 /11 :48)

[24]. Wikipedia, [enligne]. Disponible sur : https://fr.wikipedia.org/wiki/NumPy(Consulté 8/06/2021 /)

[25]. Fonctions Pandas,[enligne].Disponible sur :https://moncoachdata.com/blog/pandas-et-numpy-pour-la-data-science/https://moncoachdata.com/blog/pandas-et-numpy-pour-la-data-science/

[26]. Tutorial point, [enligne]. Disponible sur : https://www.tutorialspoint.com/scikit_learn/index.htm(Consulté 8/06/2021 /).