Université Assane SECK de Ziguinchor

UFR Sciences et Technologies

Département Informatique

Mémoire de fin d'études

Pour l'obtention du diplôme de master

Mention : Informatique

Spécialité : Réseaux et Systèmes

Sujet :

Intégration de l'Intelligence Artificielle aux données diabétiques pour une médecine de plus en plus personnalisée

Présenté par : M. El Hadji NDIAYE DIALLO

Le 04/03/2023

Sous la direction de : Dr. Madiop DIOUF et Dr. Thierno Ahmadou DIALLO

Sous la supervision du Pr. Youssou DIENG

Membres du jury :

M. Youssou DIENG	Professeur Assimilé	Président	UASZ
M. Ibrahima DIOP	Professeur Assimilé	Rapporteur	UASZ
M. El. Malick NDOYE	Maitre Conférence	Rapporteur	UASZ
M. Madiop DIOUF	Maitre Conférence Assimilé	Encadrant	USSEIN
M. Thierno Ahmadou DIALLO	Maitre Conférence	Co-encadrant	UASZ

Année Universitaire 2021/2022

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

REMERCIEMENTS

Alhamdoulilah, je remercie ***Allah le Tout-Puissant*** *de m'avoir donné la force, la patience et le courage d'accomplir ce travail.*

Mes chaleureux remerciements à mon encadrant ***Dr Madiop DIOUF*** *et* ***Dr Thierno Ahmadou DIALLO****, pour leurs soutiens, leurs disponibilités, leurs encouragements et les nombreuses discussions qui m'ont permis d'y voir plus clair.*

Nous souhaitons adresser nos remerciements les plus sincères au corps professoral, pédagogique, administratif et les intervenants externes de l'Université Assane SECK de ZIGUINCHOR, pour la richesse et la qualité de leur enseignement et qui déploient de grands efforts pour assurer à leurs étudiants une formation actualisée spécialement le département informatique.

Mes vifs remerciements aux membres du jury, enseignants-chercheurs à l'université Assane SECK de Ziguinchor, pour l'intérêt qu'ils ont porté à mon travail, le temps qu'ils ont bien voulu consacrer à l'évaluation de ce mémoire et de l'enrichir par leurs propositions.

Un grand merci à ma mère et mon père, pour leur amour, leurs conseils ainsi que leur soutien inconditionnel, à la fois moral et économique, qui m'a permis de réaliser les études que je voulais et par conséquent ce mémoire.

Un grand merci à ma famille à savoir mes frères et sœurs ainsi que mes oncles et tantes pour leurs soutiens.

Je voudrais exprimer ma reconnaissance envers les amis et collègues dont Mamadou KANE, Moustapha THIAM, Ousseynou SOUARE, Khamad THIAO, Amadou THIAM, Seynabou FAYE et Ramatoulaye FAYE qui m'ont apporté leur soutien moral et intellectuel tout au long de ma démarche.

Je remercie spécialement un ancien étudiant en la personne de Birame NDOYE qui m'a accompagné durant tout le travail en me prodiguant documents et conseils.

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

DEDICACES

À mon défunt camarade et ancien Alpha SANE que le tout puissant l'accueille dans son Paradis et que son âme repose en paix

À mon défunt homonyme El Hadji NDIAYE que le tout puissant l'accueille dans son Paradis et que son âme repose en paix

À ma source de joie, ceux qui ont toujours veillé sur mon bonheur, qui ont sacrifié pour me voir réussir et qui m'ont comblé tant d'amour et de tendresse, mes chers parents. Ils ont été toujours présents à mes côtés par leurs sacrifices et leurs prières. Que Dieu leur procure une longue vie avec une bonne santé.

À vous, mes adorables frères et sœurs, ceux qui ont partagé avec moi tous les moments d'émotion lors de la réalisation de ce travail, merci pour votre support.

À mes chers amis avec qui j'ai partagé les meilleurs et les moments les plus agréables tout au long de mon parcours universitaire et tous mes collègues de ma promo du département de l'informatique.

À tous ceux qui, par un mot, m'ont donné la force de continuer…

À tous ceux qui m'aiment et que j'aime…

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

RESUME

Le Diabète est considéré comme la maladie la plus meurtrière et chronique qui provoque une augmentation du glucose. La maladie polygénique est celle où la glande exocrine ne fabriquant pas de l'agent hypoglycémiant et selon la Fédération Internationale des Maladies Polygéniques 382 millions d'individus vivent avec une maladie polygénique dans le monde. D'ici 2035, ce chiffre doublera pour atteindre 592 millions. Le diabète sucré peut être une maladie due à l'augmentation du taux de glucose dans le sang. De nombreuses difficultés peuvent survenir si le diabète n'est pas traité et n'est pas identifié par le médecin. Ainsi, l'Intelligence Artificielle (IA) qui est devenue le nouveau terme que l'on entend tous les jours ces dernières années est définie comme la capacité d'une machine d'agir par elle-même et qui n'est pas explicitement programmée pour reproduire des actions ou des fonctions qui sont généralement celles des êtres humains. Aujourd'hui, on la retrouve dans nos machines informatiques, les réseaux sociaux, les transports et dans le secteur médical etc... De ce fait, l'apprentissage automatique est l'une des disciplines de l'intelligence artificielle qui cherche à trouver un moyen de créer des programmes informatiques qui s'améliorent automatiquement avec l'expérience. Dans ce travail, nous nous intéressons à l'utilisation des algorithmes d'apprentissage automatique pour la prédiction du diabète, afin de réduire les risques de complications de cette maladie chronique sur la santé du patient. Pour atteindre cet objectif, nous avons utilisé des algorithmes d'apprentissage automatique tels que le Random Forest RF, la Régression Logistique RL, le *K-Nearest Neighbors KNN* et les Réseaux de Neurone ANN. Les données ont été extraites de Kaggle qui est une plateforme web appartenant à Google qui fonctionne comme une communauté pour les scientifiques et les développeurs de données. Les performances des classifieurs ont été comparées en fonction du taux de précision.

Mots clés : Random Forest (RF), Régression Logistique (RL), Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), Apprentissage Automatique, Apprentissage Profond.

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

ABSTRACT

Diabetes is considered the most deadly and chronic disease that causes an increase in glucose. Polygenic disease is where the exocrine gland does not produce the hypoglycemic agent and according to the International Federation of Polygenic Diseases 382 million people live with polygenic disease in the world. By 2035, this number will double to 592 million. Diabetes mellitus can be a disease caused by increased levels of glucose in the blood. Many difficulties can arise if diabetes is not treated and not identified by the doctor. Thus, artificial intelligence (AI), which has become the new term we hear every day in recent years, generally defines the ability of a machine to act on its own and is not explicitly programmed to reproduce actions or functions that are generally those of human beings. Today, we find it in our computing machines, social networks, transportation and in the medical sector etc... Therefore, machine learning is one of the disciplines of artificial intelligence that seeks to find a way to create computer programs that automatically improve with experience. In this work, we are interested in using machine learning algorithms for the prediction of diabetes, in order to reduce the risk of complications of this chronic disease on the health of the patient. To achieve this goal, we used machine learning algorithms such as Random Forest RF, Logistic Regression RL, K-Nearest Neighbors KNN and Neural Networks ANN. The data were extracted from Kaggle which is a web platform owned by Google that operates as a community for data scientists and developers. The performances of the classifiers were compared according to the accuracy rate.

Keywords: Random Forest (RF), Logistic regression (LR), Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), Machine Learning, Deep Learning.

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

Table des matières

Listes des Figures VIII
Liste des Tableaux IX
LISTE DES ABREVIATIONS X
Introduction Générale 1
Chapitre I : Généralités et justificatif du sujet 3
I.1 Contexte et Problématique 4
I.2 Objectif et contribution 5
I.3 Méthodologie de recherche 5
I.4 Enjeux et Défis de l'intelligence artificielle 6
I.4.1 Défis de l'intelligence artificielle 6
I.4.2 Enjeux de l'intelligence artificielle 7
a. Enjeux économiques 7
b. Enjeux sociétales 7
I.5 Etat de l'art sur la prédiction du diabète 8
I.6 Conclusion 11
Chapitre II : Etat de l'art de l'apprentissage automatique 12
II.1 Introduction 13
II.2 Définition de l'apprentissage automatique 13
II.3 Domaine d'application de l'apprentissage automatique 14
II.3.1 L'apprentissage automatique dans le domaine de la santé 14
II.3.2 L'apprentissage automatique dans le domaine de l'industrie 15
II.3.3 L'apprentissage automatique dans le domaine du transport 16
II.3.4 L'apprentissage automatique le domaine de l'agriculture 16
II.4 Méthodologie d'un projet de machine learning 17
II.4.1 Définition des objectifs 17
II.4.2 Définition d'ensemble de données utilisé et description des variables 18
II.4.3 Le nettoyage et la normalisation des données 18
II.4.4 Choisir un modèle 18
II.4.5 La Séparation des données train /test 18
II.4.6 Evaluations des modèles 18
II.5 Les types d'apprentissage automatique 19
II.5.1 Apprentissage Supervisé 19

II.5.2 Apprentissage non supervisé 20
II.5.3 Apprentissage par renforcement 21
II.6 Quelques algorithmes d'apprentissage automatique 21
II.6.1 K nearest Neighbors (KNN) 21
II.6.2 Decision Trees (Arbre de décision) 24
II.6.3 Random Forest (forêts aléatoires) 25
II.6.4 Régression Logistique 27
II.6.5 Réseaux de Neurone ANN 29
II.7 Conclusion 31
Chapitre III : Expérimentation 32
III.1 Introduction 33
III.2 Outils et environnement de développement 33
III.2.1 Kaggle 33
III.2.2 Langage de programmation 33
III.2.3 Éditeur de code 35
III.3 Analyse exploratoire des données 36
III.4 Expérimentation du page web de prédiction 38
III.4.1 Environnement de développement: Pycharm 38
III.4.2 Serveurs et base de données 39
a. Serveur web : serveur local de Django 39
b. Base de données : SQLite 39
III.4.3 Technologies utilisées pour la partie Front-End 40
a. HTML 40
b. CSS 3 41
III.4.4 Technologies utilisées pour la partie Back-End 42
a. Langage de programmation : python (version 3.9.7) 42
b. Framework Django 42
III.5 Conclusion 43
Chapitre IV: Proposition de la solution et Implémentation 44
IV.1 Introduction 45
IV.2 Les étapes à suivre dans ce travail 45
IV.2.1 Chargement des données 45
IV.2.2 Vérification et l'affichage des informations de la base 45
IV.2.3 Normalisation des données 46
IV.3 Sélection de modèle 47
IV.4 Train/Test Split 47

IV.5 Random Forest (Forêt aléatoire) 48
IV.6 Régression Logistique (RL) 50
IV.7 Artificial Neural Network (ANN) 52
IV.8 K-Nearest Neighbors (KNN) 54
IV.9 L'arborescence de l'application 57
IV.9.1 L'application principale du projet 57
IV.9.2 Les autres Paramètres 59
IV.10 Présentation de l'application 60
IV.10.1 Page d'accueil 61
IV.10.2 Page de prédiction 61
IV.10.3 Conclusion 62
Conclusion Générale et Perspectives 63
Bibliographie et Webographie 64

Listes des Figures

Figure II. 1 : Le processus typique de l'apprentissage automatique 14
Figure II. 2 : l'apprentissage automatique dans le domaine de la santé 15
Figure II. 3 : L'apprentissage automatique dans le domaine de l'industrie 15
Figure II. 4 : l'apprentissage automatique dans le domaine du Transport 16
Figure II. 5 : L'apprentissage automatique dans le domaine de l'agriculture 17
Figure II. 6 : Diagramme de processus d'apprentissage supervisé 20
Figure II. 7 : Exemple simple sur KNN 22
Figure II. 8 : Structure de l'algorithme random forest 26
Figure II. 9 : Architecture réseaux de neurone 30

Figure III. 1 : Bibliothèque Python 34
Figure III. 2 : Éditeur de code Google Colab 36
Figure III. 3 : Environnement de développement PyCharm 38
Figure III. 4: Base de données SQLite 40
Figure III. 5: Image de HTML5 41
Figure III. 6 : Image de CSS3 41
Figure III. 7 : Framework Django 43

Figure IV. 1 : Aperçu de l'ensemble des données 45
Figure IV. 2 : Les informations de la base de données 46
Figure IV. 3: La normalisation de la base de données 46
Figure IV. 4 : code de la normalisation 47
Figure IV. 5: Répartition des données Train/Test 47
Figure IV. 6 : Fractionnement de l'ensemble de données 48
Figure IV. 7 : Evolution des tests avec Random Forest 49
Figure IV. 8 : Evolution des tests avec Régression Logistique 51
Figure IV. 9 : Représentation entre Random Forest et Régression Logistique 51
Figure IV. 10 : Evolution des tests avec Artificial Neural Network 53
Figure IV. 11 : Représentation entre Random Forest, RL et ANN 53
Figure IV. 12 : Evolution des tests avec K-nearest neighbors 55
Figure IV. 13 : Représentation graphique 56
Figure IV. 14 : Application principale du projet 57
Figure IV. 15 : fichier urls.py 58
Figure IV. 16 : fichier views.py 59
Figure IV. 17: Autres paramètres 59
Figure IV. 18 : fichier home.html 60
Figure IV. 19 : fichier predict.html 60
Figure IV. 20: Page de connexion 61
Figure IV. 21 : Page d'accueil 61
Figure IV. 22: Page de prédiction 62
Figure IV. 23: Affichage résultats 62

Liste des Tableaux

Tableau II. 1 : Comparaison entre apprentissage supervisé et non supervisé 21
Tableau II. 2:Avantage et Inconvénients KNN 24
Tableau II. 3: Avantage et Inconvénient Décision Tree 24
Tableau II. 4: Avantage et Inconvénient RL 29
Tableau II. 5: Avantage et Inconvénient ANN 30

Tableau IV. 1: Résultats des différents tests avec random Forest 48
Tableau IV. 2: Résultats des différents tests avec régression logistique 50
Tableau IV. 3: Résultats des différents tests avec artificial neural network 52
Tableau IV. 4: Résultats des différents tests avec k-nearest neighbors 54
Tableau IV. 5: Résultats des évaluations pour les différents modèles 55

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

LISTE DES ABREVIATIONS

OMS Organisation mondiale de la santé

IA Intelligence Artificielle

RF Random Forest

RL Régression Logistique

ANN Artificial Neural Network

KNN K-Nearest Neighbors

TP True Positive

FP False Positive

FN False Négative

ASM Mesure de Sélection d'Attribut

SVM Support Vector Machine ou Machine à vecteurs de support

EANN Evolutifs Artificial Neural Network

AUC Area Under the ROC Curve

ONEIROS Open-ended Neuro-Electronic Intelligent Robot Operating System

HTML HyperText Markup Language

CSS Cascading Style Sheets

DM Diabetes Mellitus

DL Deep Learning

ML Machine Learning

Wi poids synaptiques

xi signaux d'entrées

UCI Union Cycliste Internationale

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

MFWC Manitoba Farm Women's Conference

GPU Graphics Processing Unit

BSD Système Base de Données

IMC Indice de Masse Corporelle

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

Introduction Générale

Le diabète est une maladie grave qui se développe largement dans le monde. La gravité de la maladie réside dans les complications qui surviennent lorsque le patient néglige de vérifier s'il est diabétique ou ne reçoit pas les soins appropriés. Les complications les plus courantes du diabète sont les maladies cardiaques, les accidents vasculaires cérébraux, les maladies rénales sont les causes de décès [1]. La prévalence mondiale du diabète chez les adultes âgés de plus de 18 ans était de 8,5 % en 2014 selon l'Organisation Mondiale de la Santé (OMS) [2]. En parallèle en 2030, l'OMS prévoit que le diabète sera la septième cause de décès [2]. La prédiction du diabète est l'un des principaux sujets de recherche sur la santé les plus importants. Aujourd'hui, les modèles informatiques permettant de prédire le diabète peuvent considérablement aider à la prise de décision et aider à l'autogestion de la maladie [3]. Par conséquent, l'apprentissage automatique est une discipline de l'intelligence artificielle qui cherche à trouver un moyen de créer des programmes informatiques qui s'améliorent automatiquement avec l'expérience. Aujourd'hui, on la retrouve dans nos machines informatiques, les réseaux sociaux, les transports et dans le secteur médical. L'application de l'IA en médecine permet de prédire de nombreuses maladies facilitant aux médecins d'intervenir le plus rapidement possible afin de réduire les risques et de lutter efficacement contre les dangers sanitaires. Ainsi, quand à l'apprentissage automatique c'est une discipline de l'intelligence artificielle qui cherche à trouver un moyen de créer des programmes informatiques qui s'améliorent automatiquement avec l'expérience.

Dans ce mémoire, nous nous intéressons à l'utilisation des algorithmes d'apprentissage automatique pour la prédiction du diabète qui est un dysfonctionnement du système de régulation de la glycémie, afin de réduire les risques de complications de cette maladie chronique sur la santé du patient à travers des données collectées sur kaggle.

Ce travail sera organisé en quatre principaux chapitres comme suit :

- **Le premier chapitre** présentera un aperçu général du sujet. Nous parlerons, d'abord, du contexte et de la problématique, puis, nous terminerons par donner l'objectif et une contribution.
- **Le deuxième chapitre** présentera 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 du diabète.

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

- **Le troisième chapitre** présentera une étude technique dans laquelle nous définissons l'environnement logiciel utilisé pour la partie expérimentale de notre projet.
- **Le quatrième chapitre** présentera, d'abord, la partie de proposition de la solution et de l'implémentation, ensuite, les résultats sont présentés, comparés et interprétés. Nous terminerons par une représentation des interfaces d'application d'apprentissage dans la prédiction du diabète.

En fin, nous terminerons par une conclusion générale qui fera la synthèse des idées fondamentales que nous avons développées et les perspectives.

Chapitre I : Généralités et justificatif du sujet

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

I.1 Contexte et Problématique

La population diabétique mondiale ne cesse d'augmenter, en 2021, le nombre était de 537 millions.
Ce chiffre doublera en 2025.

Cette épidémie est liée à plusieurs facteurs dont le vieillissement de la population, les régimes hypercaloriques, l'obésité et les changements de mode de vie dominés par la sédentarité. Il existe une extrême hétérogénéité de la prévalence du diabète d'un pays à l'autre. Le monde est en pleine transition épidémiologique et le diabète pose un vrai problème de santé publique par le biais des complications chroniques dominées par les complications cardiovasculaires, le pied diabétique, l'insuffisance rénale chronique et la rétinopathie.

Selon une enquête de l'institut national de santé publique en 2020, près de 4 millions de personnes étaient identifiées diabétiques par l'assurance maladie positionnant le diabète à la quatrième place dans les maladies chroniques non transmissibles. Actuellement, la sensibilisation de la population par l'identification des facteurs de risque qui peuvent être à l'origine du diabète, fait l'objectif des différents acteurs dans le domaine de la santé publique.

Cependant, la reconnaissance et l'identification de ces facteurs repose généralement sur des études faites sur une grande population. Ces études sont regroupées sous forme de bases de données informatisées dans les hôpitaux et les instituts médicaux.

Chaque fois la taille de ses bases de données médicales augmente, l'analyse visuelle et l'exploitation de ses données devient très complexe pour les experts humains. Pour cette raison, des techniques dites intelligentes d'extraction et d'analyse de données ont été utilisées. L'utilisation de systèmes de classification pour le diagnostic médical est en développement progressif. Il n'y a aucun doute que l'évaluation des données du patient et les décisions des experts sont les facteurs les plus importants dans le diagnostic. Mais, les systèmes experts et les différentes techniques d'intelligence artificielle ont prouvé dans les dernières années leurs efficacités d'aider les experts dans le domaine médical.

Le diabète est connu depuis l'antiquité comme un trouble avec ruine de miel. La déficience de la prise en charge provoque des complications et limite l'activité de la personne malade et conduit à la mort. Il est un problème majeur de santé publique à l'échelle mondiale. Son évolution est silencieuse et insidieuse jusqu'à l'apparition de complications lourdes de conséquences en termes de morbidité et de mortalité.

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!
Devant cette situation inquiétante, il est devenu indispensable d'étudier et de comprendre ce phénomène pour en trouver des solutions techniques performantes et efficaces.

I.2 Objectif et contribution

Notre objectif est de concevoir et de valider un modèle de comparaison de la performance des algorithmes d'IA sur les données diabétiques et un outil logiciel autonome performant permettant aux patients et au personnel médical (médecins ou infirmières) de prendre des décisions rapides en identifiant par anticipation et de prédire si le patient est positif ou non. Essentiellement, nous voulons répondre à la question suivante :

Comment peut-on concevoir et valider un modèle de prédiction autonome, performant, permettant de bien prédire l'état d'un patient?

Pour la comparaison des algorithmes, nous nous concentrerons sur l'utilisation d'algorithmes d'apprentissage automatique 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 supervisés tels que : la Régression Logistique, les forêts aléatoires ou Random Forest, l'algorithme Artificial Neural Network, et l'algorithme de k-nearest neighbors sur la base de données "Pima Indian Diabète Database". Les résultats de performance sont exprimés quantitativement en termes de précision.

I.3 Méthodologie de recherche

Notre démarche globale s'articule autour des points suivants :

- Survoler et comprendre les différentes définitions de contexte afin d'adopter la plus pertinente dans le cadre de cette recherche, cela permettra de comprendre le terrain de notre travail et le scénario qu'il faut prendre en compte dans la première question de notre objectif.
- Effectuer une revue de la littérature portant sur les modèles de représentation du contexte. Cette revue nous permettra de prouver notre choix.
- Effectuer une revue des algorithmes utilisés pour l'apprentissage automatique, comme les algorithmes du raisonnement, de la discrétisation et de la sélection des attributs pertinents, pour les comprendre et les améliorer.

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

- Comparer plusieurs algorithmes utilisés lors des phases de prétraitement et de traitement de l'apprentissage automatique, pour choisir les plus performants selon une métrique de classification.
- Sélectionner une combinaison d'ensemble des algorithmes dans l'apprentissage automatique, de façon ordonnée et performante selon notre test, afin que nous puissions proposer un modèle prédictif pouvant être généralisé et appliqué à n'importe quel système de prédiction.

I.4 Enjeux et Défis de l'intelligence artificielle

I.4.1 Défis de l'intelligence artificielle

➢ **Défis Sociétale**

Les acteurs du développement doivent donc s'engager activement dans un dialogue avec les autorités chargées de la protection des données et les responsables du traitement de ces données et la société civile [41]

Il est crucial d'analyser la portée et l'impact des applications d'IA et les défis qui y sont associés. Toutefois, les recherches précédentes n'ont examiné les applications de l'IA que de manière isolée et fragmentaire. Dans [42], les chercheurs ont synthétisé les publications scientifiques sur l'IA en vue de dresser un portrait d'ensemble des utilisations de l'IA dans le secteur public.

➢ **Défis Technologiques :**

L'intelligence artificielle promet une demande croissante de solutions. En plus du prestige de la marque, les entreprises doivent apporter des résultats réels, une expérience singulière et plus de valeur ajoutée seront privilégiées [43].

À une époque où les chercheurs, les développeurs et les scientifiques effectuent des percées dans le domaine de l'intelligence artificielle à une vitesse incroyable et ce, dans divers domaines. Certains concepts juridiques devront inévitablement être adaptés pour faire face aux défis que ces percées amèneront. Ainsi, l'essentiel est d'être conscient des risques juridiques associés aux importantes percées dans le domaine de l'intelligence artificielle et de prendre des décisions éclairées dans le cadre de la gestion du développement et de l'utilisation de l'intelligence artificielle. [44]

I.4.2 Enjeux de l'intelligence artificielle

Les principaux enjeux de l'IA restent dans l'économie et la société qui seront accompagnés par un changement de mentalité de la masse sociale afin qu'elle comprenne le meilleur de l'IA.

a. Enjeux économiques

De nos jours, le modèle économique est en plein changement suite à l'évolution du numérique et des technologies qui en découlent, à savoir l'intelligence artificielle (IA), le big data, les biotechnologies, la robotique ou encore l'internet des objets. Depuis lors, les progrès dans ce domaine ont été spectaculaires et incessants. La disponibilité croissante de grands ensembles de données est générée dans tous les domaines (agriculture, éducation, santé humaine, commerce, communications) et portée par les progrès continus de la puissance de calcul et du développement d'algorithmes mais aussi des techniques d'apprentissage automatique améliorées [45].

Ainsi, l'IA ouvre une nouvelle ère de disruption et de croissance, où l'intelligence humaine est renforcée par la rapidité d'exécution et la précision. Dès à présent, toute entreprise désirant intégrer l'intelligence artificielle va devoir relever ses défis pour pouvoir adopter l'IA et décrire les mesures qu'elle pourra prendre pour poursuivre le déploiement de l'IA, de l'adoption à l'expansion dans sa structure.

b. Enjeux sociétales

- Voir les compétences : ces informations n'apparaissent pas sur les CV classiques. L'intelligence artificielle permet d'en savoir plus sur la personnalité du candidat, sa trajectoire professionnelle, son potentiel, ses aptitudes à travailler en équipe, à innover, etc. Elle offre en outre une profondeur d'analyse des profils, en quelques secondes.
- Prendre la meilleure décision : l'IA est en mesure de fournir au recruteur plus d'informations pour une plus grande personnalisation dans l'approche. Elle nous fournit toutes les cartes pour rédiger un message impactant. Aussi l'IA par sa profondeur d'analyse, est capable de déterminer si un talent déjà en poste est à l'écoute du marché.
- Créer plus d'engagement : trouver un candidat qualifié, d'accord mais comment être sûr qu'il sera motivé au quotidien dans l'entreprise ? La profondeur d'analyse évoquée précédemment permet d'avoir une idée générale du potentiel du candidat ainsi que de sa réussite future une fois en poste. Et le temps dégagé grâce à l'IA est utilisé en entretien pour approfondir ces questions avec le candidat [46][47].

I.5 Etat de l'art sur la prédiction du diabète

Les techniques d'apprentissage automatique ont été appliquées dans plusieurs domaines, notamment dans le domaine médical. Nous citons quelques travaux qui utilisent les algorithmes d'apprentissage automatique pour résoudre des problèmes médicaux.

Fikirte Girma et al ont proposé «Prediction of diabète à l'aide de techniques d'exploration de données» [22]. Cette étude est destinée à prédire l'exploitation du DM (Diabetes Mellitus) par la règle de propagation de Back. Selon les résultats de ce travail, la précision de la propagation arrière dans la prédiction de polygénique est meilleure que celle des formules SVM, J48 et Naïve Bayes.

Terry Jacob Mathew et al ont proposé l'étude «Analysis of Supervised Learning Techniques for Cost Effective Disease Prediction»; prédiction rentable des maladies à l'aide de paramètres non cliniques [23]. Cet article utilise différents algorithmes, qui sont Naïve Bayes qui a donné une précision de 80,37% tandis que les arbres REP ont enregistré un maximum de la régression logistique donnant une précision de 77%.

Butwall & Kumar [24] ont proposé une méthodologie basée sur le classificateur Random Forest (RF) pour envisager le comportement du diabète en accord avec des paramètres particuliers du style de vie, incluant l'activité physique et les états émotionnels, en particulier les diabétiques âgés. Dans ce travail de recherche, le classificateur Random Forest a été utilisé avec différents paramètres de test sur la base de données de diabétique indiens Pima de l'UCI Machine Learning Lab. Il a été constaté que la RF est efficace dans le diagnostic du diabète sucré lorsque la personne fournit les valeurs d'attributs requises.

Dewangan & Agrawal [25] ont pris diverses méthodes de classification pour les regrouper afin de donner un nouveau modèle de classification hybride dans le but de trouver de meilleures performances. Des techniques d'apprentissage automatique telles que C4.5, la forêt aléatoire, et le Perceptron multicouche ont été entraînées sur l'ensemble des données sur le diabète collectées dans le dépôt de l'UCI. Le but principal de ce travail est la détection du diabète sucré et la classification des données en tant que diabétiques ou non.

Devi & Shyla [26] ont exploré la prédiction précoce du diabète en utilisant diverses techniques d'apprentissage automatique telles que Naïve Bayes, Perceptron multicouche, Random Forest, l'arbre aléatoire et le J48 modifié. L'ensemble de données a été pris de l'ensemble de données indiennes PIMA pour déterminer la précision de la technique d'apprentissage technique

automatique dans la prédiction. L'analyse a prouvé que le classificateur J48 modifié fournit la plus grande précision la plus élevée par rapport aux autres techniques.

Dans le travail de recherche [27], différents algorithmes de classification tels que Naïve bayes, Multi Layer Perceptron, J48, Random Forest, et la régression ont été appliqués pour décrire le résultat. Le site de recherche mené vise à extraire des connaissances à partir d'une donnée et à générer des résultats complets et intelligents. Dans cette étude, les auteurs se sont concentrés sur les données de patients diabétiques. L'objectif de la comparaison de l'algorithme sur le même ensemble de données est d'analyser et de prédire les résultats.

Aishwarya et Vaidehi [28] ont utilisé plusieurs algorithmes d'apprentissage automatique tels que Support Vector Machines, Random Forest Classifier, Decision Tree Classifier, Extra TreenClassifier, Adaboost algorithm, Perceptron, Linear Discriminant Analysis algorithm, Logistic Regression, K-NN, Gaussian Naïve Bayes, Algorithme de mise en sac et Gradient Boost Classifier. Ils ont utilisé deux ensembles de données différents - le PIMA indienne et un autre ensemble de données sur le diabète pour tester les différents modèles. La régression logistique leur a donné une valeur de précision de 96 %. D'autre part, Tejas et Pramila ont choisi deux algorithmes, la régression logistique et la SVM pour construire un modèle de prédiction du diabète. Le prétraitement des données a été effectué pour obtenir de meilleurs résultats. Ils ont constaté que SVM fonctionnait mieux avec une précision de 79 %.

Yuvaraj et Sripreethaa [29] ont conçu un modèle de prédiction du diabète en utilisant trois algorithmes d'apprentissage automatique différents : Random Forest, Decision Tree et Naïve Bayes, dans des clusters basés sur Hadoop. Ils ont utilisé des techniques de prétraitement sur l'ensemble de données. Les résultats ont montré que le taux de précision le plus élevé de 94 % a été obtenu avec l'algorithme Random Forest.

Deepti et Dilip [30] ont utilisé les algorithmes Decision Tree, SVM et Naïve Bayes. Une validation croisée à dix volets a été utilisée pour améliorer les performances. La précision la plus élevée a été obtenue par le Naïve Bayes, avec une précision de 76,30 %. Ces deux articles ont utilisé l'ensemble de données Pima Indian Diabètes.

Sajida et al. Dans [31] discute le rôle des méthodes d'apprentissage automatique d'ensemble Adaboost et Bagging [32] en utilisant J48 comme base pour classer le diabète sucré et les patients comme diabétiques ou non diabétiques, en fonction des facteurs de risque du diabète sur la base des facteurs de risque du diabète. Les résultats obtenus après l'expérience prouvent

que la technique d'apprentissage machine d'ensemble Adaboost surpasse les performances de l'arbre de décision J48.

Dalakleidi et al [33] ont mis en œuvre les réseaux neuronaux réseaux neuronaux évolutifs (EANN), un algorithme basé sur la méthode bayésienne, des arbres de décision et une régression logistique pour prédire le développement du diabète et la prédiction de l'une de ses complications du diabète, à savoir les maladies cardiovasculaires. La plus grande précision la plus élevée est obtenue par le modèle EANN, avec une précision de 80,20 % et une aire sous la courbe (AUC) de 0,849. Le modèle a pu prédire la complication avec une précision de 92,86% et une AUC de 0,739 également.

Atiquzzaman et al. [34] ont proposé un cadre de prédiction pour le diabète sucré en utilisant l'apprentissage profond, où le sur ajustement est réduit en utilisant la méthode d'exclusion. Il y a deux couches entièrement connectées chacune suivie d'une couche d'exclusion. La décision est trouvée de la couche de sortie avec un seul nœud. Le système est appliqué à l'ensemble de données sur le diabète des Indiens Pima. La plus grande précision obtenue par le système est de 88,41 %.

Zhu et al. [35] Ont proposé un système utilisant des classificateurs multiples et ont amélioré la précision de la prédiction de maladies complexes comme le diabète. Ils ont proposé un schéma de vote dynamique pondéré pour ce système. Le système est testé sur des ensembles de données T2DM et le jeu de données sur le diabète des Indiens Pima. La plus haute précision maximale obtenue par le système est de 93,45% en utilisant MFWC avec k=10 sur le jeu de données du diabète indien Pima.

Mukesh Kumari et al. [36] ont utilisé des techniques d'exploration de données pour prédire le diabète sucré. Ils extraient des connaissances à partir de l'ensemble de données et la description compréhensible des modèles. Le système a obtenu la plus grande précision de 99,51 % en utilisant le réseau bayésien.

Santhanam et al. [37] ont proposé un système pour prédire le diagnostic du diabète en utilisant K-Means, algorithme génétique, et SVM (Support Vector Machine). Le système a suivi les étapes suivantes. Première étape, mise à jour toutes les valeurs manquantes avec la moyenne. Deuxième étape, l'ensemble de données nettoyées est regroupé en utilisant K-Means pour éliminer les données aberrantes et inutiles et sélectionne la caractéristique optimale en utilisant l'algorithme génétique pour réduire les caractéristiques. La plus grande précision du système est de 98,82% en utilisant le SVM.

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

Vijayashree et al. [38] ont proposé un système qui utilise l'élimination récursive des caractéristiques et l'analyse pour la prédiction du diabète. Ils classifient le diabète en utilisant des réseaux neuronaux profonds et des réseaux neuronaux artificiels. En utilisant le réseau neuronal profond, leur précision était de 82,67 %, et en utilisant le réseau neuronal artificiel leur précision était de 78.62%.

Goncalves et al. [39] ont présenté un système permettant de prédire le diabète en utilisant la méthode hiérarchique Neuro-Fuzzy BSP. Ils proposent un nouveau modèle de partitionnement de l'espace binaire (BSP) neuro-flou hiérarchique, consacré à l'analyse des modèles. Ils ont trouvé 80,08 % dans l'ensemble de formation et 78,26 % dans l'ensemble de test. Han et al. [40] ont introduit le modèle K-means par paire et sous contrainte de taille. Par paire et contrainte par la taille pour dépister la population à haut risque de diabète sucré.

Compte tenu des solutions déjà proposées nous allons montrer notre démarche et l'importance de notre proposition. Celle-ci consiste à faire une étude et une comparaison de quelques algorithmes d'apprentissage automatique pour la prédiction du diabète en utilisant des données collectées sur kaggle. Avec le framework Django nous avons essayé de réaliser une page web permettant de mieux tester nos modèles.

I.6 Conclusion

En résumé, dans ce chapitre nous avons essayé d'abord de dégager le contexte et la problématique justifiant le sujet mais aussi donné les objectifs et les méthodes de recherche utilisés. Ensuite, nous avons introduit un état de l'art sur l'application des algorithmes d'apprentissage automatique sur le diabète et parlé des défis et enjeux de l'intelligence artificielle.

Chapitre II : Etat de l’art de l’apprentissage automatique

II.1 Introduction

Dans ce chapitre, nous allons d'abord aborder l'apprentissage automatique pour lequel nous introduisons les principaux types ainsi que les algorithmes utilisés, ensuite nous enchaînons par un état de l'art sur quelques algorithmes de classification appliqués dans la prédiction du diabète. Cette application apporte un grand avantage à partir duquel, on peut réduire les risques de complications de cette maladie sur la santé d'un patient.

II.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 d'effectuer des tâches extrêmement complexes afin d'obtenir des résultats en temps réel qu'il est difficile d'obtenir avec des algorithmes classiques. [4]

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. [5]

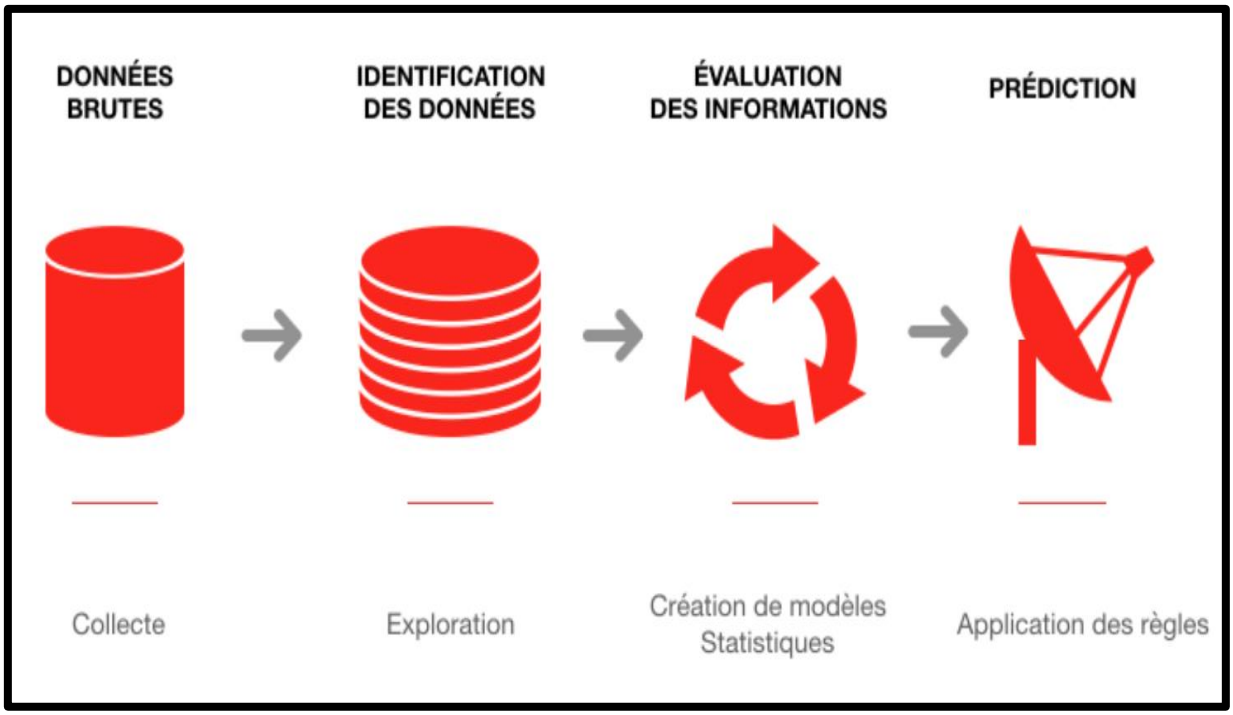

Figure II. 1 : Le processus typique de l'apprentissage automatique

II.3 Domaine d'application de l'apprentissage automatique

Depuis quelques décennies l'apprentissage automatique est devenu active et attractive au cœur des services de l'homme. Cette omniprésence visant à représenter la capacité de réaction de ce dernier dans plusieurs domaines n'est plus à démontrer.

II.3.1 L'apprentissage automatique dans le domaine de la santé

Les préoccupations majeures des professionnels de la santé restent la gestion des maladies et des patients. On assiste à un renforcement des liens, des patients sont désormais acteurs de leurs propres santés. Ils posent des diagnostics rapides et précis, des traitements optimisés.

Figure II. 2 : l'apprentissage automatique dans le domaine de la santé

II.3.2 L'apprentissage automatique dans le domaine de l'industrie

La quatrième révolution industrielle est marquée par l'utilisation de capteurs communicants, le secteur industriel devient intelligent, on assiste à un nouveau système de productivité, l'entreprise numérique est devenue une réalité [6]. Les données sont générées, traitées, analysées en continu, la maintenance machine est désormais prédictive ceci va booster la production en plus des assistant virtuels qui alertent en cas de présence de personne n'ayant pas accès au système.

Figure II. 3 : L'apprentissage automatique dans le domaine de l'industrie

II.3.3 L'apprentissage automatique dans le domaine du transport

L'apprentissage automatique occupe une place importante dans le secteur du transport avec l'avènement des voitures autonomes, le transport a subi d'énormes bouleversements. Dès lors qu'on assiste à une réduction drastique des accidents de la route et de la diminution de la congestion des parkings avec plus d'économie de consommation d'énergie.

Figure II. 4 : L'apprentissage automatique dans le domaine du transport

II.3.4 L'apprentissage automatique le domaine de l'agriculture

La naissance d'une nouvelle forme d'agriculture facilitée par l'IA a donné naissance à de nouvelles machines qui ont tendance à supplanter la présence humaine dans les champs [7]. Avec l'amélioration de la précision des technologies informatique cognitive, qui via la reconnaissance d'images permet de distinguer une plante mur va révolutionner l'agriculture ancienne. La tendance est que plusieurs startups fournissent des solutions innovantes pour tirer le meilleur de l'IA dans l'agriculture qui est utilisée dans de nombreux pays [8]. Plusieurs des exemples sont à noter :

Le développement d'algorithmes d'apprentissage profond capable de détecter l'état du sol et d'offrir plus de rendements ;

L'arrivé des « robots cueilleurs » dotés des systèmes d'IA avec des caméras dotées de capteurs fournissent des images en temps réel pour un champ donné avec plus de tâches réussi que les humains [9] ;

Des modèles d'apprentissage automatique sont utilisés pour contrer la main-d'œuvre analyse prédictive pour la suivie de l'évolution des rendements [10].

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

Figure II. 5 : L'apprentissage automatique dans le domaine de l'agriculture

II.4 Méthodologie d'un projet de machine learning

Un projet de Machine Learning s'avère très différent 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
- Définition d'ensemble de données utilisé et description des variables
- Le nettoyage et la normalisation des données
- Choisir un modèle
- La Séparation des données train /test
- Evaluations des modèles

II.4.1 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.

II.4.2 Définition d'ensemble de données utilisées et description des variables

Une fois que nous avons décidé 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.

II.4.3 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.

II.4.4 Choisir un modèle

L'étape suivante du flux de travail consiste à choisir un modèle. Les chercheurs et les data scientistes 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.

II.4.5 La Séparation des données train /test

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

II.4.6 Evaluations 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

$$\textbf{Précision} = \frac{\textbf{TP}}{\textbf{TP+FP}} \quad \textbf{……..(1)}$$

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

$$\textbf{Rappel} = \frac{\textbf{TP}}{\textbf{TP+FN}} \quad \textbf{……. … (2)}$$

- **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.

$$\textbf{Score F1} = \frac{\textbf{2}*(\textbf{Rappel}*\textbf{Précision})}{\textbf{Rappel+Précision}} \quad \textbf{……. (3)}$$

II.5 Les types d'apprentissage automatique

On distingue usuellement au moins trois types d'apprentissage machine : l'apprentissage par renforcement, l'apprentissage supervisé et l'apprentissage non supervisé.

II.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) [11].

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 3 montre le diagramme de processus d'apprentissage supervisé.

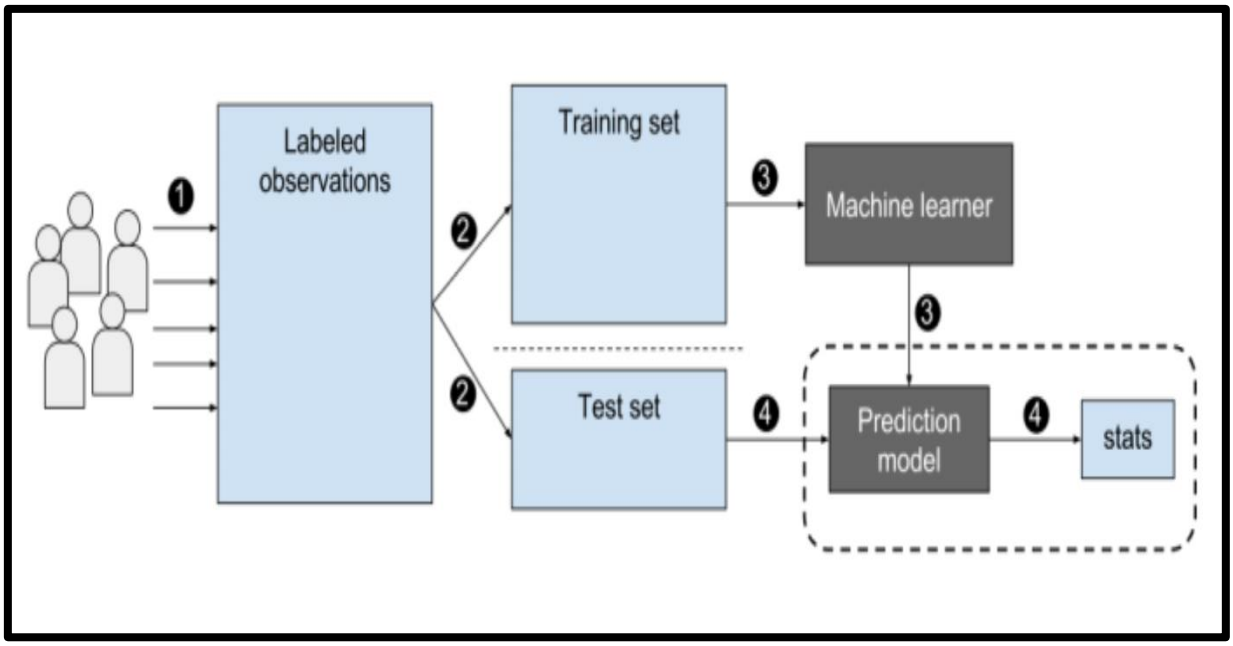

Figure II. 6 : Diagramme de processus d'apprentissage supervisé

Le diagramme ci-dessus comporte trois parties 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 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.

II.5.2 Apprentissage non supervisé

L'apprentissage non supervisé ne contient pas la variable de sortie correspondante comme 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 [12].

Cette discipline est connue dans ce type d'apprentissage 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 grande dimensionnalité. Une solution évidente face à ce problème c'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 [11].

II.5.3 Apprentissage par renforcement

L'apprentissage par renforcement suppose qu'un agent (entité qui agit de façon autonome) reçoit des récompenses ou des punitions en fonction des actions qu'il exécute. Il s'agit alors d'établir automatiquement, à partir des retours d'expérience, des stratégies d'action des agents qui maximisent l'espérance de récompenses. Ces techniques développées depuis la fin des années 1950 ont fait leurs preuves à la fois dans le domaine des jeux et dans celui de la robotique. [13]

Tableau II. 1 : Comparaison entre apprentissage supervisé et non supervisé

	Apprentissage Supervisé	**Apprentissage non Supervisé**
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 données
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

II.6 Quelques algorithmes d'apprentissage automatique

II.6.1 K nearest Neighbors (KNN)

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

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

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 des données.

KNN fonctionne par classification ou prédiction sur la base d'un nombre fixe (K) de points de données les plus proches de points d'entrée. Cela signifie que pour une valeur choisie de K, un point d'entrée serait classé ou devrait appartenir à la même classe que la classe la plus proche des nombre des points K voisins. [14]

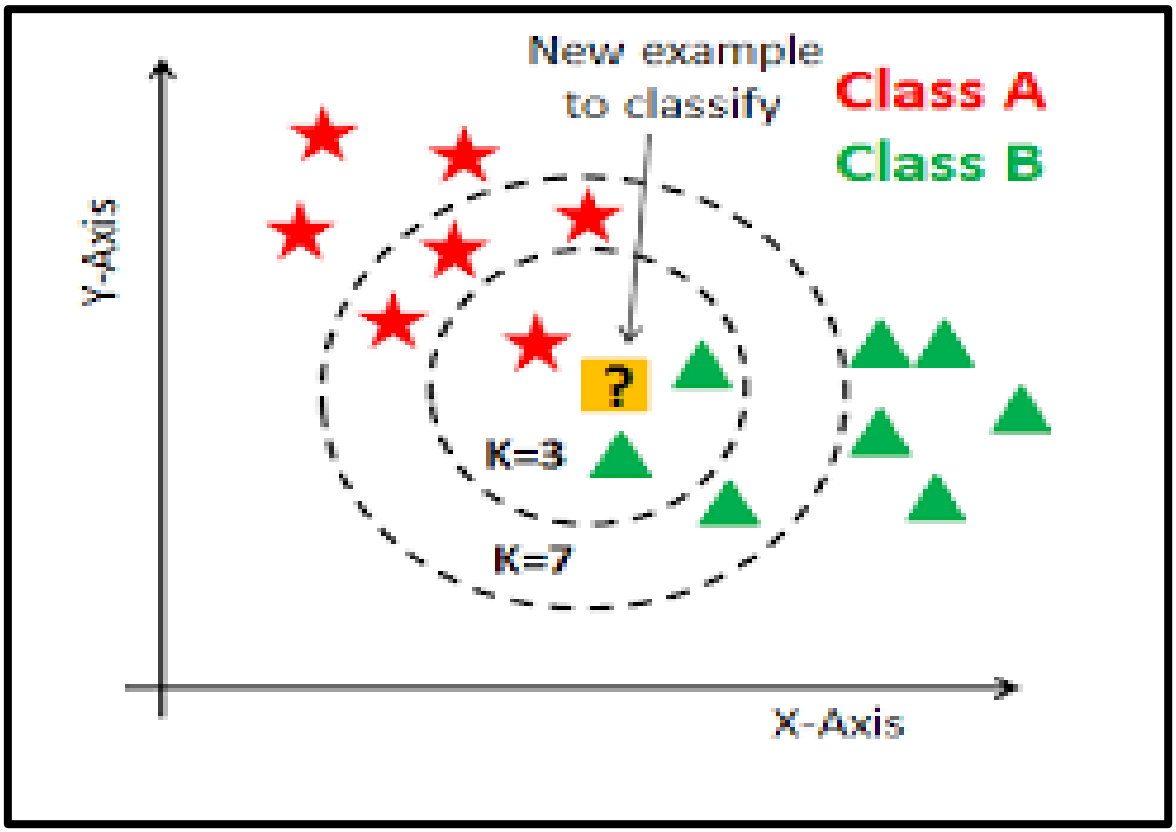

Figure II. 7 : Exemple simple sur KNN

Dans le schéma ci-dessus, nous avons une donnée non classée et tous les autres données sont classées (étoile et triangle) chacun avec leur classe (classe A et B).

- Si k=3 les données les plus proche du nouvelle donnée sont à l'intérieure du premier cercle, et la classe la plus prédominante c'est triangle (Classe B) car 2 triangles et seulement 1 étoile donc la donnée non classée sera classée comme triangle (Classe B).
- Si k=7 les données les plus proches de la nouvelle donnée sont à l'intérieure du deuxième cercle, et la classe la plus prédominante c'est l'étoile (Classe A) car on a 4 étoiles et 3 triangles donc la donnée non classée sera classée comme étoile (Classe A).

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

La distance entre le point non classée et les plus proches voisins est mesuré en utilisant différents méthode comme : la distance euclidienne, la distance de Manhattan, la distance de Minkowski, celle de Jaccard, la distance de Hamming etc., la fonction de distance est choisi en fonction de type de données qu'il manipule. Pour les données de même type la distance euclidienne est le bon candidat, et pour les données qui ne sont pas de même type la distance de Manhattan est la bonne mesure pour l'utiliser. [15][16]

- Distance euclidienne : $\boldsymbol{De(x; y)} = \sqrt{\sum nj = 1(xj - yj)2}$ (4)

- Distance Manhattan : $\boldsymbol{Dm(x; y)} = \sum_{i=1}^{k} |xi - yi|$ (5)

Algorithme de construction de KNN

for i ← 1 to m do

for j ← 1 to n do

Calculer la distance euclidienne dij entre x test i et x train j en utilisant l'équation (4)

dj ← dij

End

Calculer la classe z test i deuxième exemple qui vaut la classe de son ppv :

Trier les distances dj selon un ordre croissant pour j = 1, . . ., n

Récupérer en même temps les indices IndVoisins avant le tri des dj

Récupérer les classes des K premiers ppv à partir des indices IndVoisins et en trouver la classe majoritaire :

Ck ← 0 (k = 1, . . ., K)

for k ← 1 to K do

ind_voisink ← IndVoisinsk

h ← z train ind_voisink

End

Tableau II. 2:Avantage et Inconvénients KNN

Avantage KNN	**Inconvénient KNN**
Simple à implémenter	le choix de la valeur de k (le nombre de voisins le plus proche)
Gérer naturellement les cas multi classes	Le coût de calcul est ´élevé (pour chaque instance de l'ensemble de données on a besoin de calculer la distance)
Peut être utilisé pour la classification et la régression	Stockage de données
	Sensible aux fonctionnalités non pertinentes

II.6.2 Decision Trees (Arbre de décision)

Decision Trees ou arbre de décision est un algorithme parmi les algorithmes d'apprentissage supervisé le plus utilisé et le plus pratique, qui est adapté pour résoudre tout type de problèmes (classifications ou régressions) telle-que :

- 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.
- C'est un algorithme non-paramétrique qui signifie qu'il n'y a pas d'hypothèse sous-jacente sur la distribution des données. [17,18]

Le principal problème qui se pose lors de la construction d'un arbre de décision est comment choisir ou sélectionner le meilleur attribut pour le nœud racine et qui sépare mieux l'ensemble de données? Pour résoudre ce problème, il existe une technique appelée Mesure de sélection d'attribut ou ASM qui contient deux mesures principales et populaires qui sont : Indice de Gini et Gain d'information.

Algorithme de construction d'un arbre de décision

1. lire les données

2. sélectionner le meilleur attribut (nœud racine)

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.

Tableau II. 3: Avantage et Inconvénient Décision Tree

Avantages Decision Tree	**Inconvénients Decision Tree**
Faciles à expliquer et comprendre.	Il faut souvent plus de temps pour former le modèle
Fonctionne avec des données catégorielles et numériques	L'arbre devient plus complexe à mesure qu'il s'approfondit
Peu coûteux en termes de calcul.	Un petit changement dans les données peut entraîner un changement global de la structure de l'arbre de décision

II.6.3 Random Forest (forêts aléatoires)

Random Forest ou forêts aléatoires est un algorithme d'apprentissage supervisé très populaire. Il est également utilisé pour les problèmes de régression ou de classification. Basé sur un ensemble des algorithmes d'apprentissage, qui est un processus de combinaison de plusieurs algorithmes pour résoudre un problème complexe et améliorer les performances du modèle. C'est un algorithme qui crée de nombreux arbres de décision (c'est la raison pour laquelle il est appelé une forêt) sur divers sous-ensembles de l'ensemble de données. Elle prend la prédiction de chaque arbre et sur la base des votes majoritaires des prédictions, et elle prédit le résultat final. [19] La figure suivante explique le fonctionnement et la structure d'algorithme.

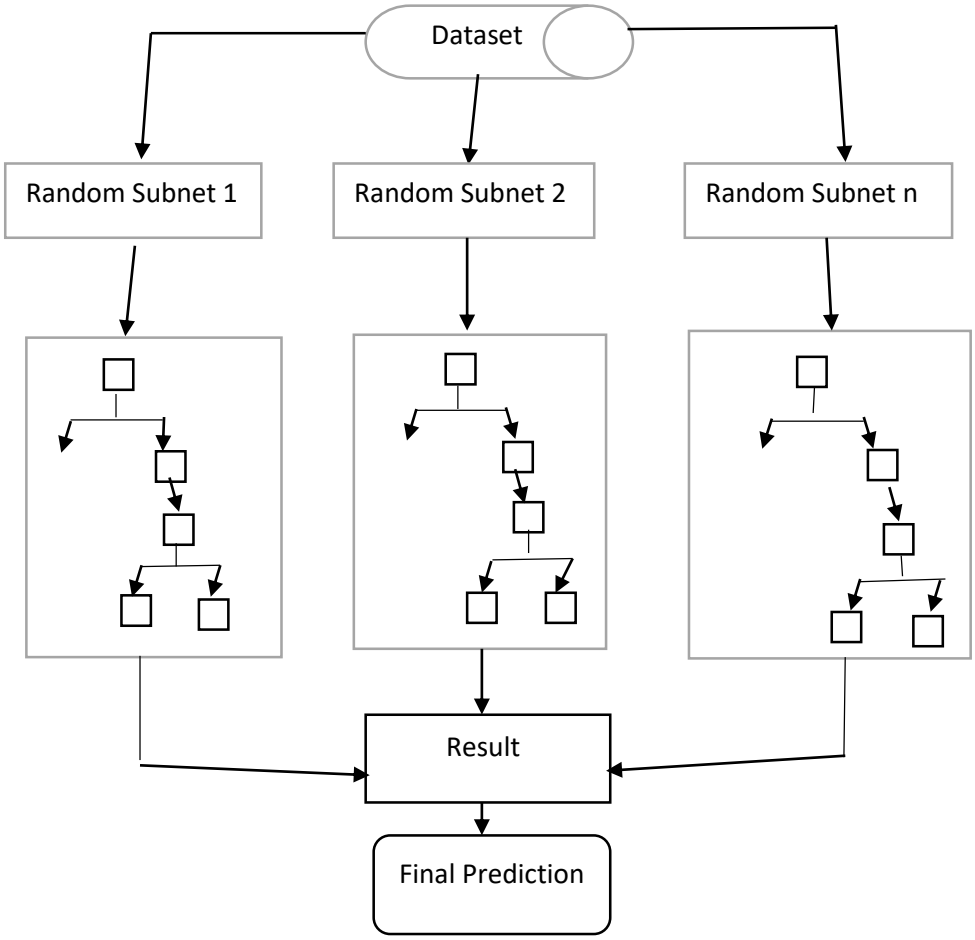

Figure II. 8 : Structure de l'algorithme random forest

Algorithme de construction Random Forest

Condition préalable : Un ensemble d'apprentissage S := (x1, y1), . . . (xn, yn), des caractéristiques F, et un nombre d'arbres dans la forêt B.

function RandomForest(S, F)

H ← ∅

for i ∈ 1, . . ., B do

S (i) ← A « Un échantillon de S »

hi ← RandomizedTreeLearn(S (i), F)

H ← H ∪ {hi}

end for

return H

end function

function RandomizedTreeLearn(S, F)

À chaque nœud :

f ← très petit sous-ensemble de F

return le meilleur résultat

Avantage de Random forest

1. Il s'agit de l'un des algorithmes d'apprentissage les plus précis et disponibles. Pour de nombreux ensembles de données, il produit un classificateur très précis.

2. Il fonctionne efficacement sur de grandes bases de données.

3. Il dispose d'une méthode efficace pour estimer les données manquantes et maintient la précision lorsqu'une grande partie des données sont manquantes.

Inconvénient de Random forest

Le principal inconvénient de l'algorithme random forest est qu'un grand nombre d'arbres peut rendre l'algorithme trop lent et inefficace pour les prédictions en temps réel. En général, ces algorithmes sont rapides à entraîner, mais assez lents à créer des prédictions une fois qu'ils sont formés. Une prévision plus précise nécessite plus d'arbres, ce qui entraîne un modèle plus lent. [20]

II.6.4 Régression Logistique

En épidémiologie, plusieurs modèles d'analyse multi variée sont couramment utilisés : régression linéaire multiple, régression logistique, régression de Poisson, modèle de Cox, etc. Effectuer une régression, c'est tenter de réduire les données d'un phénomène complexe en une loi mathématique simplificatrice. La fonction logistique (qui a donné son nom au modèle) possède des caractéristiques mathématiques expliquant son emploi dans un modèle d'analyse de données épidémiologiques : elle varie de 0 à 1 comme la probabilité de survenue d'un événement. La régression logistique peut être uni variée mais son intérêt réside dans son utilisation multi variée puisqu'elle permet, alors, d'estimer la force de l'association entre la variable dépendante et chacune des variables explicatives, tout en tenant compte de l'effet

simultané de l'ensemble des autres variables explicatives intégrées dans le modèle. L'association ainsi estimée est dite « ajustée » sur l'ensemble des autres facteurs.

Algorithme de construction de Régression Logistique

La réalisation pratique d'un modèle de régression logistique comporte plusieurs étapes :

1. La qualité d'une régression logistique repose, avant tout, sur le choix des variables explicatives que l'on est susceptible d'intégrer au modèle. Ce choix est fondé sur la pertinence clinique et sur la connaissance de facteurs de confusion avérés ou supposés. C'est pourquoi, une recherche bibliographique approfondie est au préalable obligatoire.

2. Il est nécessaire ensuite d'étudier chacune de ces variables : analyse de la distribution des variables qualitatives selon leurs différentes modalités et s'il y a lieu, regroupement de ces dernières ; étude de l'existence d'une relation linéaire entre chacune des variables quantitatives explicatives et la variable dépendante. Si, pour une variable, cette condition n'est pas vérifiée, on procédera à la transformation de celle-ci en une variable ordinale en créant des classes dont le choix repose sur des critères cliniques et statistiques.

3. On procède ensuite à l'analyse des liaisons entre chacune des variables explicatives et la variable dépendante : on réalise une analyse uni variée ; les odds-ratios calculés sont bruts. Deux catégories de variables explicatives pourront être intégrées dans un modèle de départ : celles pour lesquelles l'association avec la variable dépendante est suffisamment forte sans toutefois être trop stricte afin de ne pas omettre d'éventuels facteurs de confusion (p-value inférieure ou égale à 0,20, et non pas 0,05, seuil habituellement retenu) et celles qui ont un intérêt clinique avéré en dehors de tout critère d'association (elles sont rares : ce sont des variables dites « forcées »).

4. Plusieurs stratégies sont possibles pour parvenir à un modèle final qui devra porter le maximum d'informations tout en ayant un nombre limité de variables afin de faciliter l'interprétation : les plus employées sont les procédures dites « pas à pas descendantes ou pas à pas ascendantes ». La déclinaison des modèles permettra de rechercher les phénomènes d'interaction ou de confusion qu'il faudra prendre en compte lors de l'interprétation. Certaines variables seront impérativement conservées dans le modèle : la variable explicative d'intérêt principal et les facteurs de confusion.

5. En fin d'analyse, plusieurs modèles finaux peuvent s'avérer satisfaisants sur un plan statistique. Parmi ceux-ci, on retiendra le modèle le plus adéquat avec le phénomène constaté : des tests d'adéquation permettent de guider le statisticien.

Tableau II. 4: Avantage et Inconvénient RL

Avantages régression logistique	Inconvénients régression logistique
Elle est plus facile à mettre en œuvre, à interpréter et très efficace à former	Il construit des frontières linéaires
Très rapide pour classer les enregistrements inconnus	La principale limitation de la régression logistique est l'hypothèse de linéarité entre la variable dépendante et les variables indépendantes
Bonne précision pour de nombreux ensembles de données simples et fonctionne bien lorsque l'ensemble de données est linéairement séparable	La régression logistique nécessite une multi colinéarité moyenne ou nulle entre les variables indépendantes
La régression logistique est moins encline au sur-ajustement	Il est difficile d'obtenir des relations complexes en utilisant la régression logistique

II.6.5 Réseaux de Neurone ANN

Le neurone artificiel est un processeur élémentaire. Il reçoit un nombre variable d'entrées en provenance de neurones appartenant à un niveau situé en amont. A chacune des entrées est associé un poids w représentatif de la force de la connexion. Chaque processeur élémentaire est doté d'une sortie unique, qui se ramifie, ensuite pour alimenter un nombre variable de neurones appartenant à un niveau situé en aval. A chaque connexion est associé un poids. [21]

Un neurone se compose généralement d'une entrée formée des variables sur lesquelles opèrent ce neurone et une sortie représentant la valeur de la fonction réalisée (fonction d'activation). La sortie du neurone est une fonction non linéaire d'une combinaison des entrées xi (signaux d'entrées) pondérées par les paramètres wi (poids synaptiques). [21]

Une ANN implique généralement un grand nombre de processeurs fonctionnant en parallèle et disposés en étages. Le premier niveau reçoit les informations d'entrée brutes, comme les nerfs optiques dans le traitement visuel humain. Chaque niveau successif reçoit la sortie du niveau précédent, plutôt que l'entrée brute de la même manière que les neurones plus éloignés du nerf optique reçoit les signaux de ceux qui en sont plus proches. Le dernier niveau produit la sortie du système. Pour qu'un réseau de neurones artificiel puisse être utilisé, on déclenche un nœud

avec une donnée d'entrée et ce nœud va déclencher d'autres nœuds auxquels il est connecté. On organise les réseaux neuraux avec une couche de neurones d'entrées et de sorties bien définie. On définit aussi des liens directs entre les nœuds pour savoir où se dirige l'information c'est-à-dire comment elle se propage. Pour finir, on assigne des nombres différents sur nos connexions que l'on appelle "poids" (ou "weight" pour les anglophones) pour que certaines connexions soient plus forte que d'autres comme les vrais neurones. Enfin, on a une couche de neurones cachés (hidden layer**)** entre ceux qui définissent les entrées et les sorties. Ces neurones vont ainsi traiter les données.

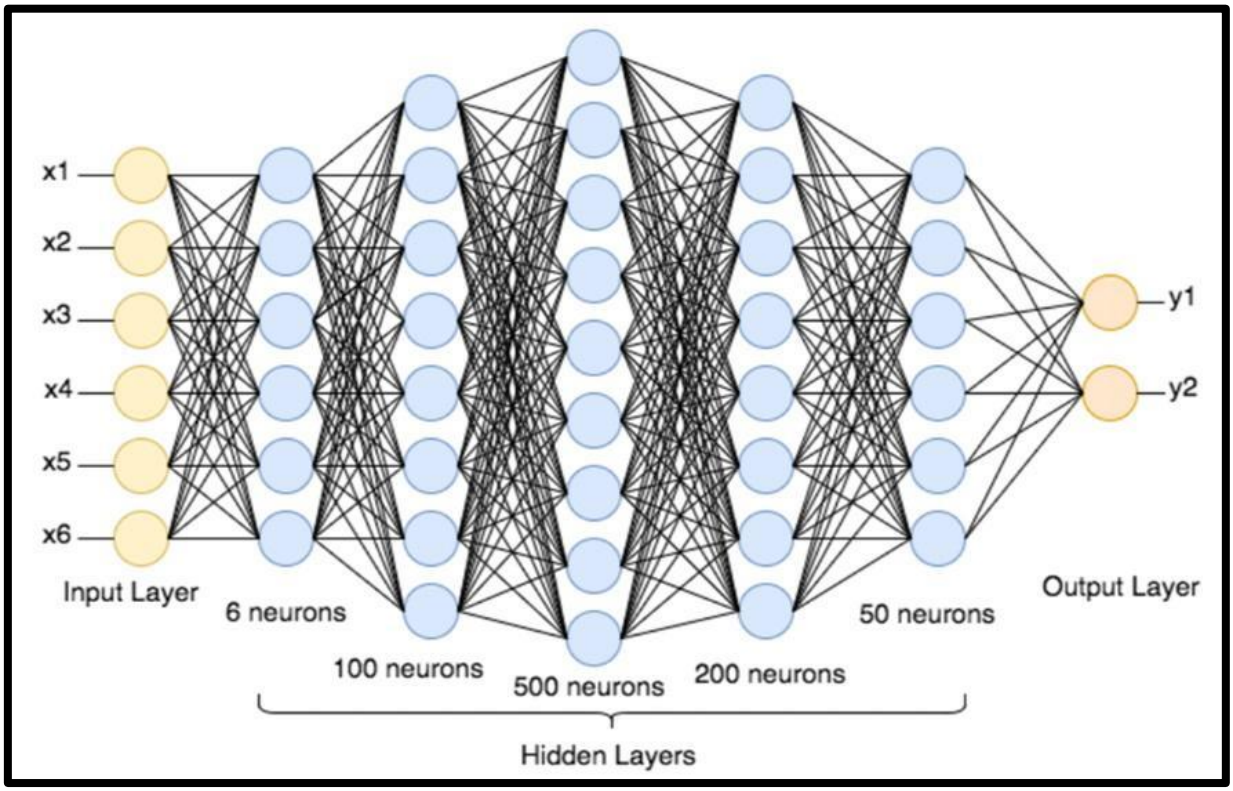

Figure II. 9 : Architecture réseaux de neurone

Tableau II. 5: Avantage et Inconvénient ANN

Avantages ANN	**Inconvénients ANN**
Les capacités de traitement parallèle signifient que le réseau peut effectuer plus d'une tâche à la fois	L'absence de règles pour déterminer la structure de réseau appropriée
La capacité d'apprendre et de modéliser des relations complexes non linéaires permet de modéliser les relations réelles entre l'entrée et la Sortie	L'exigence de processeurs ayant des capacités de traitement parallèle rend les réseaux neuronaux dépendants du matériel

La capacité à produire des résultats avec des connaissances incomplètes, la perte de performance étant basée sur l'importance des informations manquantes	Le réseau fonctionne avec des informations numériques, c'est pourquoi tous les problèmes doivent être traduits en valeurs numériques avant de pouvoir être présentés à l'ANN
La capacité à produire des résultats avec des connaissances incomplètes, la perte de performance étant basée sur l'importance des informations Manquantes	Le manque d'explication des solutions est l'un des plus grands inconvénients des ANN

II.7 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, non supervisé ou par renforcement. Des algorithmes d'apprentissage automatiques notamment les classificateurs ont été clairement montrés avec leur définition et concept.

Chapitre III : Expérimentation

III.1 Introduction

Dans ce chapitre, nous présentons la partie expérimentale de notre projet dans laquelle nous définissons l'environnement logiciel et matériel utilisé. 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 descriptives avec leurs types ainsi que les abréviations avec leurs significations. Ensuite, nous allons décrire les différentes étapes de prétraitement appliquées sur cette base de données.

III.2 Outils et environnement de développement

III.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. [48]

III.2.2 Langage de programmation

Python est un langage de programmation interprété, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions. [49]

Le langage Python est placé sous une licence libre proche de la licence BSD 7 et fonctionne sur la plupart des plates-formes informatiques, des smartphones aux ordinateurs centraux 8, de Windows à Unix avec notamment GNU/Linux en passant par MacOs, ou encore Android, Ios, et peut aussi être traduit en Java. Il est conçu pour optimiser la productivité des programmeurs en offrant des outils de haut niveau et une syntaxe simple à utiliser. [50]

Figure III. 1 : Bibliothèque Python

Les bibliothèques et package utilisé :

- **Keras** : est une bibliothèque open source écrite en Python (sous licence MIT) basée principalement sur les travaux du développeur de Google François Chollet dans le cadre du projet ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System). Une première version du logiciel multiplateforme a été publiée le 28 mars 2015. Le but de cette bibliothèque est de permettre la constitution rapide de réseaux neuronaux. Dans ce cadre, Keras ne fonctionne pas comme un framework propre mais comme une interface de programmation applicative (API) pour l'accès et la programmation de différents frameworks d'apprentissage automatique. TensorFlow fait notamment partie des frameworks pris en charge par Keras. [51]
- TensorFlow : est une bibliothèque open source de Machine Learning, créée par Google, permettant de développer et d'exécuter des applications d'apprentissage automatique et d'apprentissage en profondeur. [52]
- Scikit-learn : est une bibliothèque libre Python destinée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs2 notamment dans le monde académique par des instituts français d'enseignement supérieur et de recherche comme Inria3. Elle propose dans son framework de nombreuses bibliothèques d'algorithmes à implémenter clé en main, à disposition des data scientists. Elle comprend notamment des fonctions pour estimer des forêts aléatoires, des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support. Elle est conçue pour

s'harmoniser avec d'autres bibliothèques libres Python, notamment NumPy et SciPy. [53]

- NumPy : est une extension du 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. [54]
- Matplotlib : est une bibliothèque complète pour créer des visualisations statiques, animées et interactives en Python.
- 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: 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 négatif énorme sur notre modèle.

III.2.3 Éditeur de code

Pour éditer le code de ce système, nous avons utilisé Google colab ou Colaboratory est un service Cloud, offert par Google (gratuit), basé sur Jupyter Notebook et destiné à la formation et à la recherche dans l'apprentissage automatique. Cette plateforme permet d'entraîner des modèles de Machine Learning directement dans le cloud. Sans donc avoir besoin d'installer quoi que ce soit sur notre ordinateur à l'exception d'un navigateur. Cool, n'est-ce pas ? Avant de présenter ce magnifique service, nous rappellerons ce qu'est un Jupyter Notebook.

Figure III. 2 : Éditeur de code Google Colab

III.3 Analyse exploratoire des données

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 de ses variables prédictives.

Pima indian diabetes database est un ensemble de données provenant 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. Cette base de données se compose 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 figure ci-dessous donne un aperçu sur les premiers enregistrements de l'ensemble de données utilisées.

	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressu	SkinThicknes	Insulin	BMI	DiabetesPed	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1

Figure III.3: Aperçu de l'ensemble de données

Une brève description des variables est résumée dans le tableau ci-dessous.

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

Numéro	Abréviation	Description
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 (mmHg)
4	SkinThickness	Epaisseur du pli cutané du triceps (mm)
5	Insulin	Insuline sérique 2 heures (mu U/ml)
6	BMI	Indice de masse corporelle (poids en kg)
7	DiabetePedigreeFunction	Fonction pedigree du diabète
8	Âge	Age en année
9	Outcome	Variable de classe (0 ou 1)

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

L'ensemble de données contient 768 lignes et 9 colonnes. La variable "Résultat" est la colonne que nous allons prédire qui signifie 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).

III.4 Expérimentation du page web de prédiction

III.4.1 Environnement de développement: Pycharm

PyCharm est un environnement de développement intégré utilisé pour programmer en Python. Il permet l'analyse de code et contient un débogueur graphique. Il permet également la gestion des tests unitaires, l'intégration de logiciel de gestion de versions, et supporte le développement web avec Django. Développé par l'entreprise tchèque JetBrains, c'est un logiciel multiplateforme qui fonctionne sous Windows, Mac OS X et GNU/Linux. Il est décliné en édition professionnelle, diffusé sous licence propriétaire, et en édition communautaire diffusée sous licence Apache.

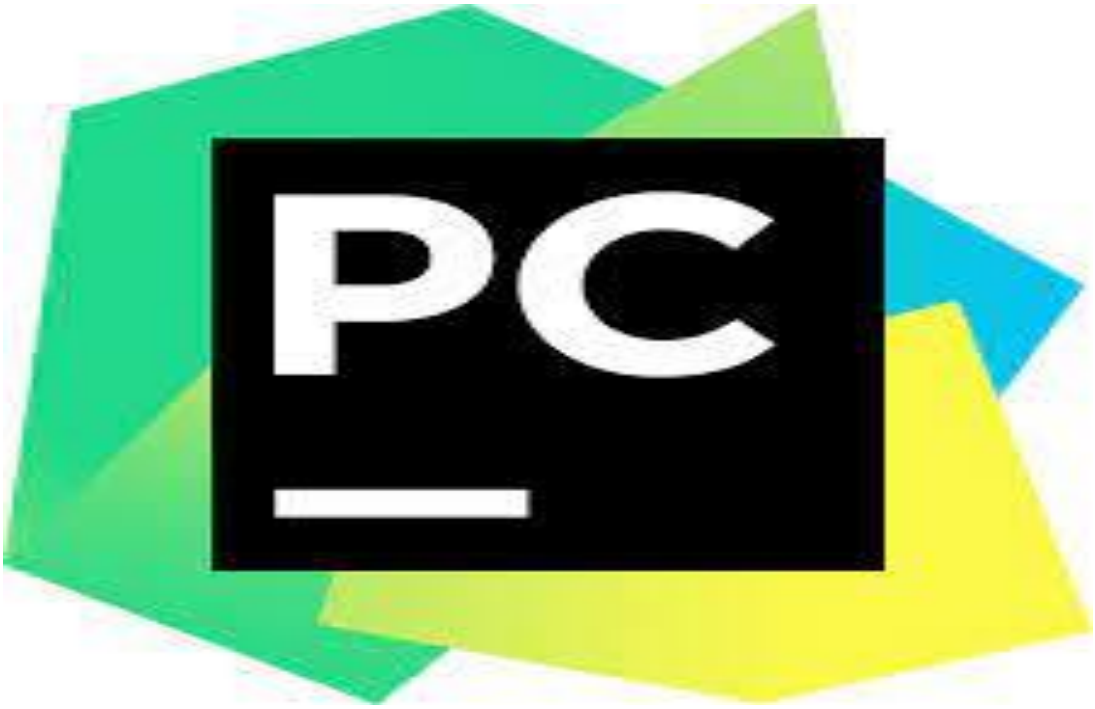

Figure III. 3 : Environnement de développement PyCharm

III.4.2 Serveurs et base de données

a. Serveur web : serveur local de Django

Une application web nécessite un serveur web local pour permettre de lancer l'exécution du code en vue d'avoir un résultat sur le compte rendu du site. On peut citer WampServer comme Exemple de serveur web. Cependant, pour gérer l'exécution de ces applications, Django utilise un serveur web local rapide et performant qui permettra d'exécuter les applications de Django.

b. Base de données : SQLite

Django est un Framework extrêmement souple qui embarque par défaut un fichier de base de données SQLite qui nous convient largement pour la phase de développement car ce type de configuration fonctionne parfaitement pour des applications peu gourmandes en ressource. Toutefois, utiliser un autre système de stockage peut aider à augmenter les performances en production.

Django tente d'activer autant de fonctionnalités que possible sur tous les types de base de données. Cependant, tous les types de base de données ne sont pas semblables, et nous avons dû prendre des décisions de conception sur les fonctionnalités à activer et les hypothèses sur lesquelles nous pouvions nous baser en toute sécurité. C'est ainsi qu'on a choisi de travailler avec SQLite qui répond naturellement à tous les critères définis.

SQLite est une bibliothèque en langage C qui implémente un petit, rapide, autonome, haute fiabilité, et complet, moteur de base de données SQL. SQLite est le moteur de base de données le plus utilisé au monde [55]. SQLite est intégré à tous les téléphones mobiles et à la plupart des ordinateurs et est intégré à d'innombrables autres applications que les gens utilisent chaque jour.

Le format de fichier SQLite est stable, multiplateforme et rétro compatible et les développeurs s'engagent à le maintenir ainsi tout au long jusqu'à l'an 2050. Les fichiers de base de données SQLite sont couramment utilisés comme conteneurs pour transférer un contenu riche entre les systèmes et comme format d'archivage à long terme pour les données. Il existe plus de 1milliard de bases de données SQLite en cours d'utilisation [55].

Figure III. 4: Base de données SQLite

III.4.3 Technologies utilisées pour la partie Front-End

a. HTML

Le HyperText Markup Language, généralement abrégé HTML ou dans sa dernière version HTML5, est le langage de balisage conçu pour représenter les pages web. Ce langage permet :

- D'écrire de l'hypertexte, d'où son nom ;
- De structurer sémantiquement des pages ;
- De mettre en forme le contenu d'une page ;
- De créer des formulaires de saisie ;
- D'inclure des ressources multimédias dont des images, des vidéos, et des programmes informatiques ;
- De créer des documents interopérables avec des équipements très variés de manière conforme aux exigences de l'accessibilité du web.

Figure III. 5: Image de HTML5

b. CSS 3

Les feuilles de style en cascade 1, généralement appelées CSS de l'anglais Cascading Style Sheets, forment un langage informatique qui décrit la présentation des documents HTML et XML. Les standards définissant CSS sont publiés par le World Wide Web Consortium (W3C). Introduit au milieu des années 1990, CSS devient couramment utilisé dans la conception de sites web et bien pris en charge par les navigateurs web dans les années 2000.

Figure III. 6 : Image de CSS3

III.4.4 Technologies utilisées pour la partie Back-End

a. Langage de programmation : python (version 3.9.7)

Python est un langage de programmation interprété, multi paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions ; il est ainsi similaire à Perl, Ruby, Schème, Small talk et Tcl.

Le langage Python est placé sous une licence libre proche de la licence BSD 4 et fonctionne sur la plupart des plates-formes informatiques, des smartphones aux ordinateurs centraux, de Windows à Unix avec notamment GNU/Linux en passant par MacOs, ou encore Android, Ios, et peut aussi être traduit en Java ou .NET. Il est conçu pour optimiser la productivité des programmeurs en offrant des outils de haut niveau et une syntaxe simple à utiliser.

Il est également apprécié par certains pédagogues qui y trouvent un langage où la syntaxe, clairement séparée des mécanismes de bas niveau, permet une initiation aisée aux concepts de base de la programmation.

Figure III. 7 : Logo Python

b. Framework Django

Django est un Framework Web Python de haut niveau qui encourage un développement rapide et une conception propre et pragmatique. Conçu par des développeurs expérimentés, il prend en charge une grande partie des tracas du développement Web, on peut donc se concentrer sur l'écriture de l'application sans avoir à réinventer la roue. C'est gratuit et open source.

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

- **Très rapide :** Django a été conçue pour aider les développeurs à faire passer les applications de la conception à la réalisation le plus rapidement possible.
- **Sécurisé de manière rassurante :** Django prend la sécurité au sérieux et aide les développeurs à éviter de nombreuses erreurs de sécurité courantes.
- **Extrêmement évolutif :** Certains des sites les plus fréquentés du Web tirent parti de la capacité de Django à évoluer rapidement et de manière flexible.

Figure III. 8 : Framework Django

III.5 Conclusion

Dans ce chapitre, nous avons évoqué et traité l'expérimentation. D'abord, on a expliqué l'environnement logiciel et matériel utilisé du système et celle de la page web de prédiction en détaillant les technologies utilisées tant côté Front-End que côté Back-End.

Cependant, pour nous assurer de la sécurité et de l'ergonomie du système, on a réalisé un ensemble de tests qui feront l'objet du prochain chapitre.

Chapitre IV: Proposition de la solution et Implémentation

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

IV.1 Introduction

Dans ce dernier chapitre, nous allons d'abord faire l'implémentation de nos données sur les quatre algorithmes à savoir RF, RL, KNN et ANN, ensuite faire une comparaison en termes de performance des différents algorithmes. A la fin, c'est la partie application où nous fournissons des interfaces graphiques importantes d´enveloppées sur Django.

IV.2 Les étapes à suivre dans ce travail

- Chargement des données
- Vérification et l'affichage des informations de la base
- Normalisation

IV.2.1 Chargement des données

La base utilisée est «Pima Indians Diabetes DataBase»

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

Figure IV. 1 : Aperçu de l'ensemble des données

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

IV.2.2 Vérification et l'affichage des informations de la base

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Figure IV. 2 : Les informations de la base de données

IV.2.3 Normalisation des données

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

La normalisation des données fait référence au décalage des valeurs de vos données afin qu'elles se situent entre 0 et 1. La normalisation des données, dans ce contexte, est utilisée comme technique de mise à l'échelle pour établir la moyenne et l'écart type à 0 et 1, respectivement. [56]

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

Figure IV. 3: La normalisation de la base de données

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

Figure IV. 4 : code de la normalisation

IV.3 Sélection de modèle

La sélection de modèle est une phase très importante et le cœur de l'apprentissage automatique où on sélectionne le modèle qui fonctionne mieux pour l'ensemble de données parmi une collection de modèles d'apprentissage automatique candidats.

Les modèles utilisés pour la prédiction de diabète sont :

1. Random Forest (forêt aléatoire RF)
2. Régression Logistique (RL)
3. Artificial Neural Network (ANN)
4. K-nearest neighbors (KNN)

Utilisé pour donner la capacité au modèle de prédire les données hors échantillons et évite le problème de sur-ajustement (overfitting) qui correspond à l'incapacité de modèle de le généraliser sur des données de test car il est appris par cœur sur les données d'entraînement. Les deux méthodes sont :

IV.4 Train/Test Split

Cette méthode consiste à diviser l'ensemble de données en deux parties : partie d'entraînement sur lequel le modèle fait son apprentissage et partie de test sur lequel on a testé le modèle et évaluer sa performance.

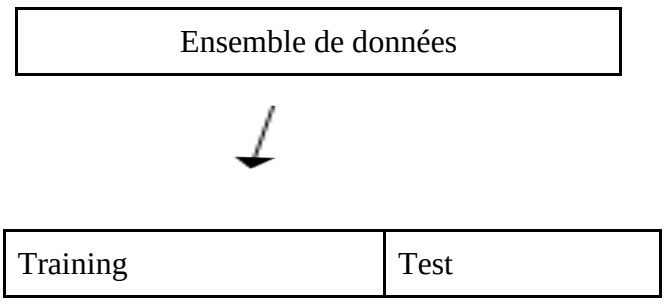

Figure IV. 5: Répartition des données Train/Test

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

Nous utilisons la méthode train test split importé de la bibliothèque sklearn pour effectuer le fractionnement train/test, test size=0.25 à l'intérieur de la fonction indique le pourcentage des données qui doivent être conservées pour le test. C'est généralement autour de 25% pour le test et de 75% pour l'entraînement, ce qui signifie 1600 observations parties d'entraînement et 400 observations parties test. Le code standard pour fractionner les données dans la figure suivant :

```
y = data["Outcome"]
x = data.drop("Outcome", axis = 1)
x_train, x_test, y_train, y_test  = train_test_split(x,y, test_size=0.25,random_state=42)
```

Figure IV. 6 : Fractionnement de l'ensemble de données

Retourne à 04 variables x_train et y_train pour l'entraînement et x_test et y_test pour le test.

IV.5 Random Forest (Forêt aléatoire)

Dans notre travail proposé, nous avons appliqué l'algorithme forêt aléatoire dénommé Random Forest sur les données de l'échantillon d'apprentissage automatique et de test. Nous avons obtenu des résultats pour différents tests. Les résultats obtenus des expérimentations effectuées sont communiqués dans le tableau ci-dessous:

Tableau IV. 1: Résultats des différents tests avec random Forest

Modèle	Test 1	Test 2	Test 3
Random Forest (RF)	0,65	0,7489	0,81

D'après les résultats montrés dans le tableau on constate que les résultats de performance varient entre 0.65 et 0.81 pour les mesures précises. On remarque aussi que plus les tests se multiplient, plus on note que la précision tend vers 100 ce qui montre une bonne maîtrise au niveau d'entrainement. Cela est plus explicite dans la figure ci-dessous.

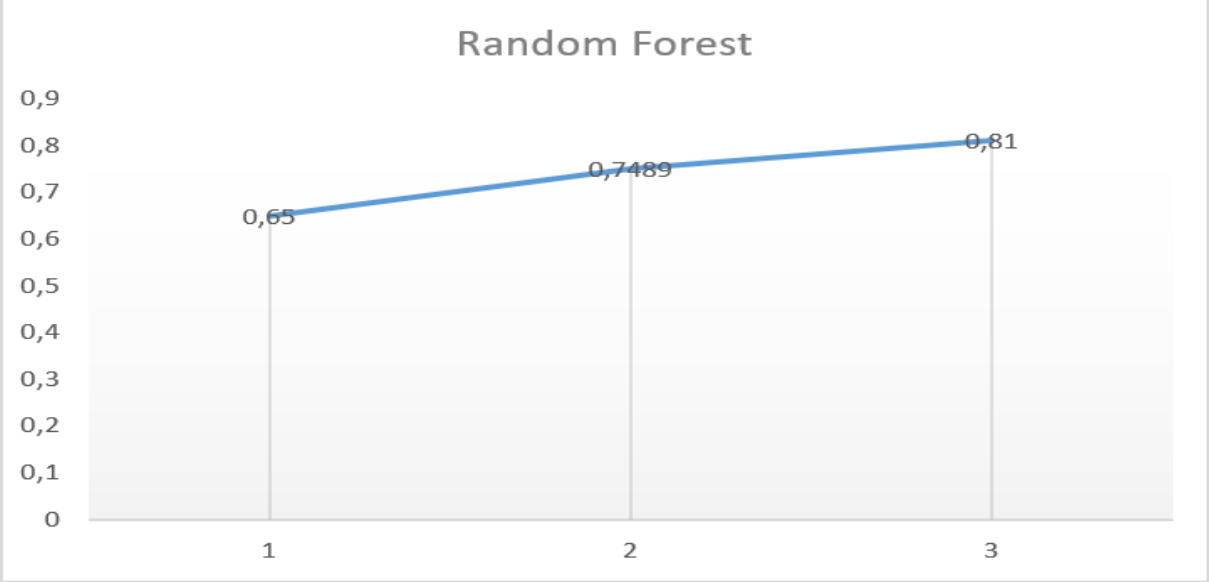

Figure IV. 7 : Evolution des tests avec Random Forest

Dans la figure, on constate clairement que le meilleur résultat est capturé lors du dernier test, cela est justifié par le fait que plus que les tests augmentent plus le modèle est entraîné plus il est efficace et donne des résultats satisfaisants.

Le graphe représente le taux de précision pour le modèle Random Forest en fonction de nombres d'itérations. On illustre que la meilleure itération est l'itération 3 avec un taux de précision égale à 81%.

Nous allons ci-dessous donner le pseudo code concernant ce modèle.

Pseudocode de Random Forest :

Étape 1 : Dans la forêt aléatoire, un nombre n d'enregistrements aléatoires est extrait de l'ensemble de données ayant un nombre k d'enregistrements.

Étape 2 : Des arbres de décision individuels sont construits pour chaque échantillon.

Étape 3 : Chaque arbre de décision générera une sortie.

Étape 4 : Le résultat final est considéré sur la base du vote à la majorité ou de la moyenne pour la classification et la régression respectivement.

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

Pseudocode de prédiction de forêt aléatoire :

1. Prenons les **fonctionnalités de test** et utilisons les règles de chaque arbre de décision créé au hasard pour prédire le résultat et stockons le résultat prédit (cible)

2. Calculez les **votes** pour chaque cible prévue.

3. Considérez la cible prédite à **vote élevé comme la prédiction finale** de l'algorithme de forêt aléatoire.

IV.6 Régression Logistique (RL)

Ensuite dans le travail proposé, nous avons aussi appliqué l'algorithme de la régression logistique sur les mêmes données de l'échantillon d'apprentissage automatique et de test. Nous avons obtenu des résultats pour différents tests. Les résultats obtenus des expérimentations effectuées sont communiqués dans le tableau ci-dessous:

Tableau IV. 2: Résultats des différents tests avec régression logistique

Modèle	Test 1	Test 2	Test 3
Régression Logistique (RL)	0,74	0,58	0,65

D'après les résultats montrés dans le tableau on constate que les résultats de performance varient entre 0.58 et 0.74 pour les mesures précises. On remarque aussi que plus les tests se multiplient plus on note parfois une augmentation et aussi une diminution au fur et à mesure que les tests se multiplient. Cela est plus explicite dans la figure ci-dessous.

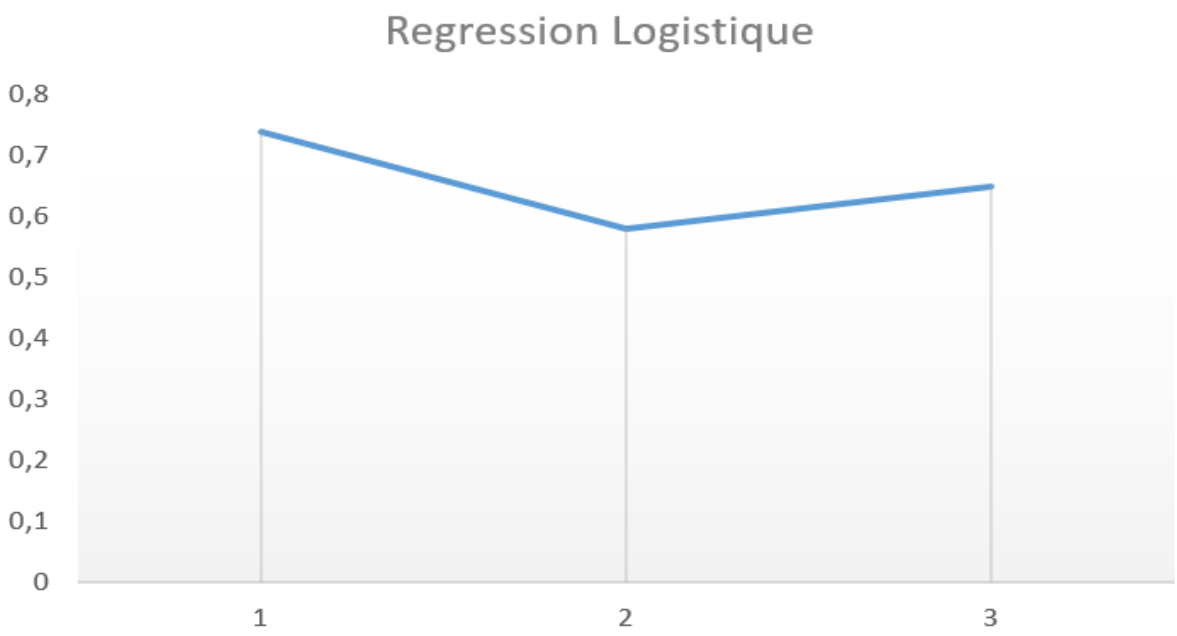

Figure IV. 8 : Evolution des tests avec Régression Logistique

Dans la figure, on constate clairement que le meilleur résultat est capturé lors du premier test. Cette divergence de performance entre l'apprentissage et la phase test est connue sous le nom de sur apprentissage.

Le graphe représente le taux de précision pour le modèle Régression Logistique en fonction du nombre d'itération. On illustre que la meilleure itération est l'itération 1 avec un taux de précision égale à 74%.

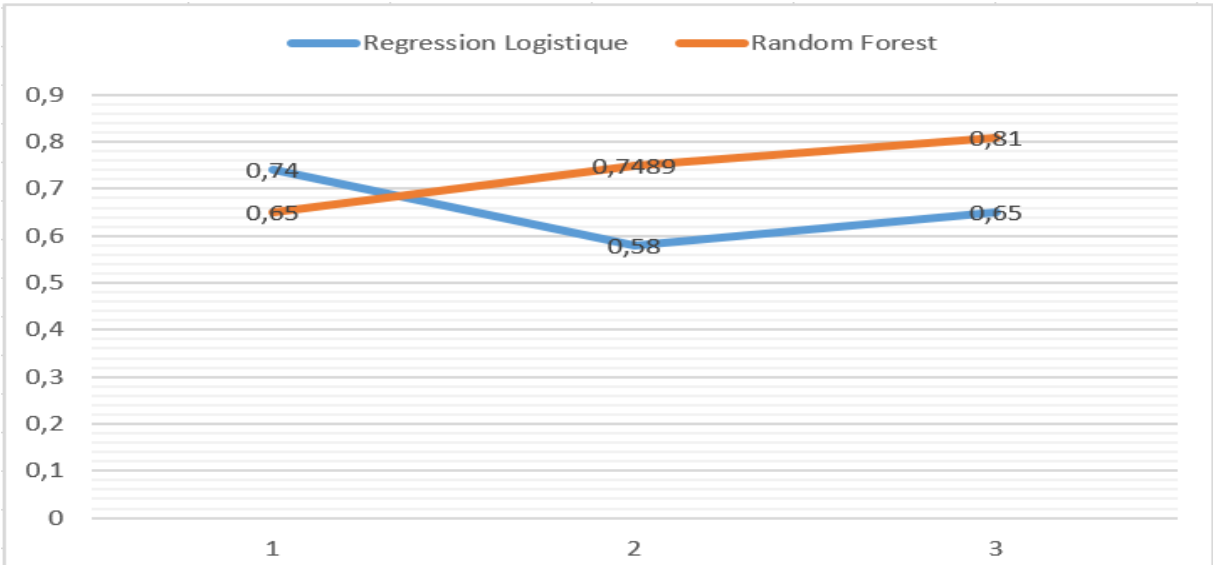

Figure IV. 9 : Représentation entre Random Forest et Régression Logistique

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

Le graphe ci-dessus représente une comparaison entre le modèle Random Forest et le modèle Régression Logistique par le taux de précision en fonction du nombre des itérations. On illustre que la meilleure itération pour le modèle Random forest est l'itération 3 avec un taux de précision de 81%. Le modèle Random forest est meilleur que le modèle Régression Logistique dans toutes les itérations sauf l'itération 1 avec une différence de taux de précision égale à 0.09%.

IV.7 Artificial Neural Network (ANN)

Dans ce travail proposé, nous avons aussi appliqué l'algorithme Artificial neural network sur les mêmes données de l'échantillon d'apprentissage automatique et de test. Nous avons obtenu des résultats pour différents tests. On note aussi que pour ce type d'apprentissage, nous avons utilisé une fonction d'activation de type sigmoïde pour la couche de sortie et une fonction d'activation de type relu pour la couche d'entrée ainsi que pour la couche cachée. Les résultats obtenus des expérimentations effectuées sont communiqués dans le tableau ci-dessous:

Tableau IV. 3: Résultats des différents tests avec Artificial neural network

Modèle	Test 1	Test 2	Test 3
Artificial Neural Network (ANN)	0,30	0,31	0,47

D'après les résultats montrés dans le tableau on constate que les résultats de performance varient entre 0.30 et 0.47 pour les mesures précises. On remarque aussi que plus les tests se multiplient, plus on note parfois une augmentation mais de façon légère. On peut noter que cela se justifie par le fait de la non complexité des données vue que ANN est considéré comme un algorithme de deep learning ou dans ce domaine les résultats sont meilleurs que si les données sont assez complexe. Cela est plus explicite dans la figure ci-dessous.

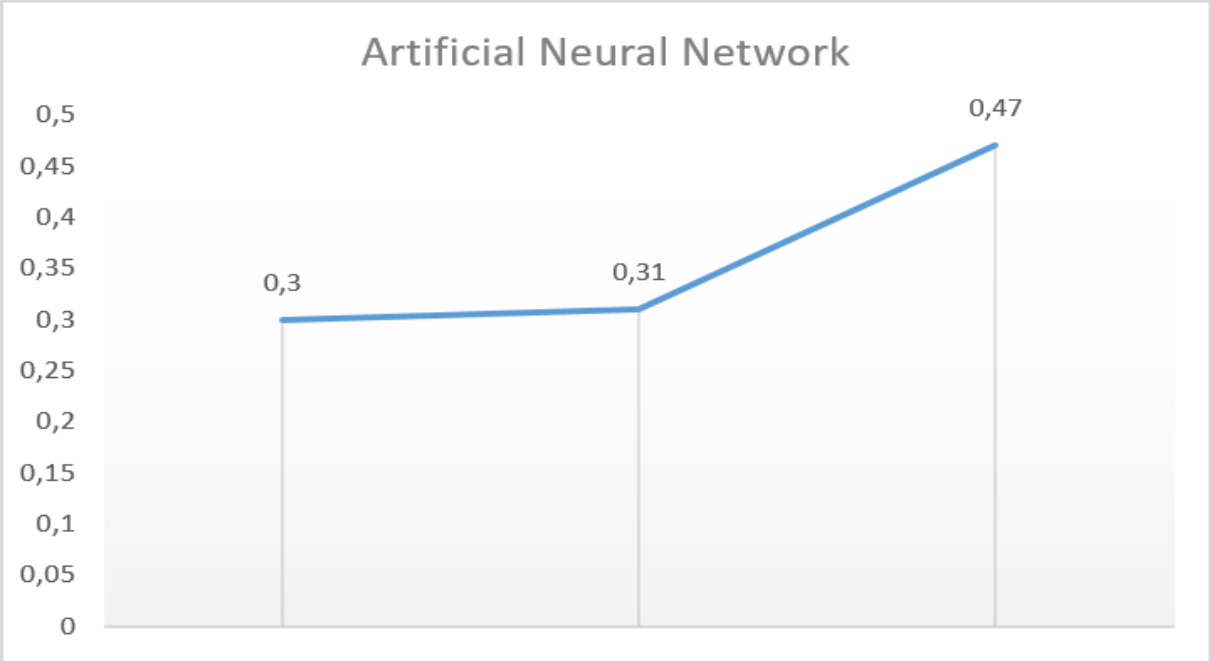

Figure IV. 10 : Evolution des tests avec Artificial Neural Network

Dans la figure, on constate clairement que le meilleur résultat est capturé lors du dernier test, cela est justifié par le fait que plus que les tests augmentent plus le modèle est entraîné plus il est efficace et donne des résultats satisfaisants.

Le graphe représente le taux de précision pour le modèle Artificial Neural Network en fonction du nombre d'itération. On illustre que la meilleure itération est l'itération 3 avec un taux de précision égale à 47%.

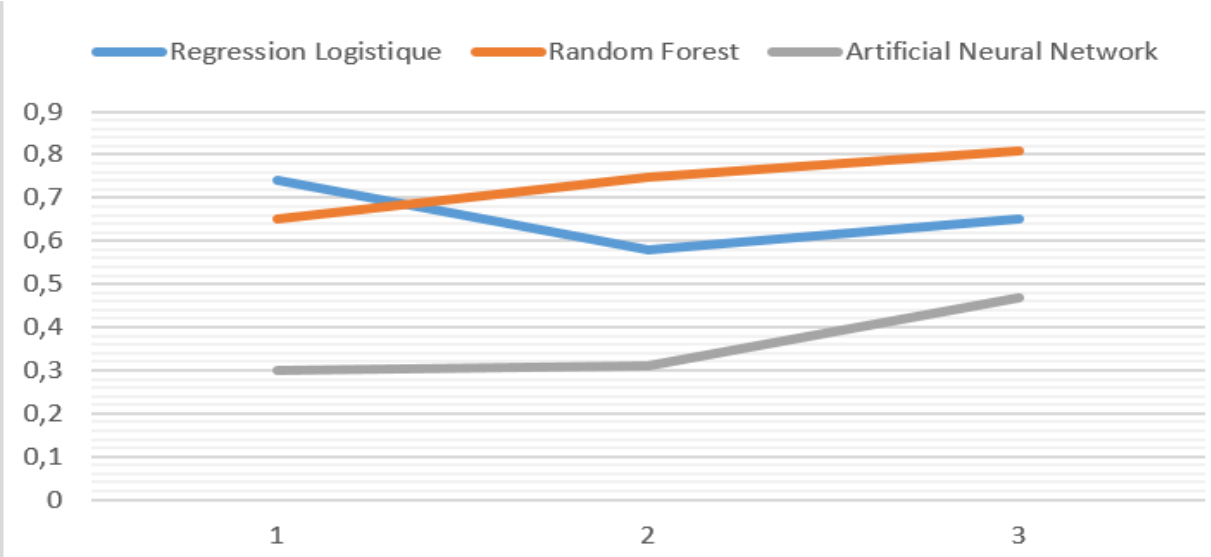

Figure IV. 11 : Représentation entre Random Forest, RL et ANN

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

Le graphe représente une comparaison entre trois modèles qui sont Random Forest, Régression Logistique et Artificial Neural Network par le taux de précision en fonction du nombre d'itération. On illustre que le modèle Random Forest est mieux que le modèle Régression Logistique sauf dans l'itération 1, et sa meilleure itération est l'itération 3 avec un taux de précision égale à 81%. Les modèles Random forest et Régression Logistique sont mieux que le modèle Artificial Neural Network dans toutes les itérations. Ci-dessous le pseudo code du modèle ANN:

Etape 1 : Définir les paramètres d'initialisation (en entrée)

Etape 2 : Sélectionner le nombre de couches cachées (Taille des couches cachées)

Etape 3 : Choisir comment diviser les échantillons et du mode de division de l'échantillon

Etape 4 : Définir la rétro propagation à gradient conjugué en fonction d'entraînement du réseau sélectionnée

Etape 5 : Choisir une fonction de performance

Etape 6 : Entraînement du réseau

Etape 7 : Tester le réseau (sortie = réseau (entrées))

Etape 8 : Prédire afin de calculer la performance du réseau

IV.8 K-Nearest Neighbors (KNN)

Dans le même travail proposé, nous avons aussi appliqué l'algorithme de K nearest neighbors sur les mêmes données de l'échantillon d'apprentissage automatique et de test tout en faisant varier le K par pas de deux et nous avons obtenu des résultats pour différents tests. Les résultats obtenus des expérimentations effectuées sont communiqués dans le tableau ci-dessous:

Tableau IV. 4: Résultats des différents tests avec k-nearest neighbors

Modèle	Test 1 K=1	Test 2 K=3	Test 3 K=5
K-nearest neighbors KNN	0,62	0,68	0,74

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

D'après les résultats montrés dans le tableau on constate que les résultats de performance varient entre 0.62 et 0.74 pour toutes les mesures de précision. On remarque aussi, que plus le paramètre k augmente, plus le taux de mesures de précision augmente également. Cela est plus explicite dans la figure ci-dessous.

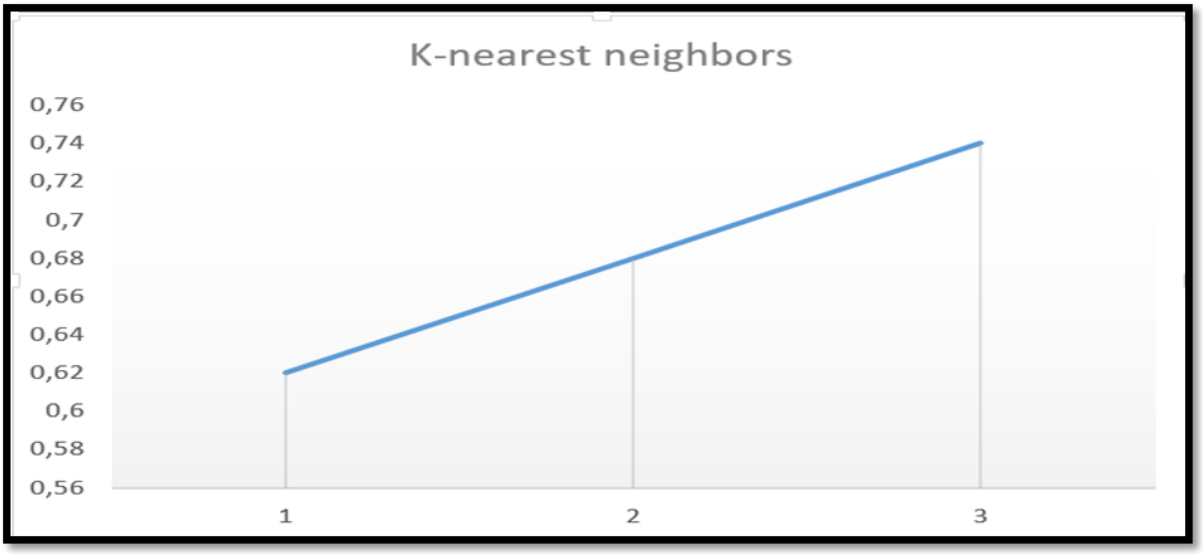

Figure IV. 12 : Evolution des tests avec K-nearest neighbors

Dans la figure, on constate clairement que le meilleur résultat est capturé lors du dernier test. Ce qui montre encore que la performance est proportionnelle à la variation du K. Le graphe représente le taux de précision pour le modèle K-nearest neighbors en fonction des valeurs de K. On illustre que la meilleure itération est l'itération avec k=5 avec un taux de précision égale à 74%.

Evaluation des modèles

Tableau IV. 5: Résultats des évaluations pour les différents modèles

Modèles	**Test 1**	**Test 2**	**Test 3**
Random Forest	**0,65**	**0,7489**	**0,81**
Régression Logistique	**0,74**	**0,58**	**0,65**
Artificial Neural Network	**0,30**	**0,31**	**0,47**
K-nearest neighbors KNN	**0,62**	**0,68**	**0,74**

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

D'après le tableau ci-dessus le modèle Random forest a obtenu la meilleure précision qui est égale à 81% classé à l'aide de mesure de diagnostiques médicales. Nous sélectionnons le modèle Random forest comme le modèle le plus optimal et qui fonctionne mieux pour notre ensemble de données en raison de sa grande précision.

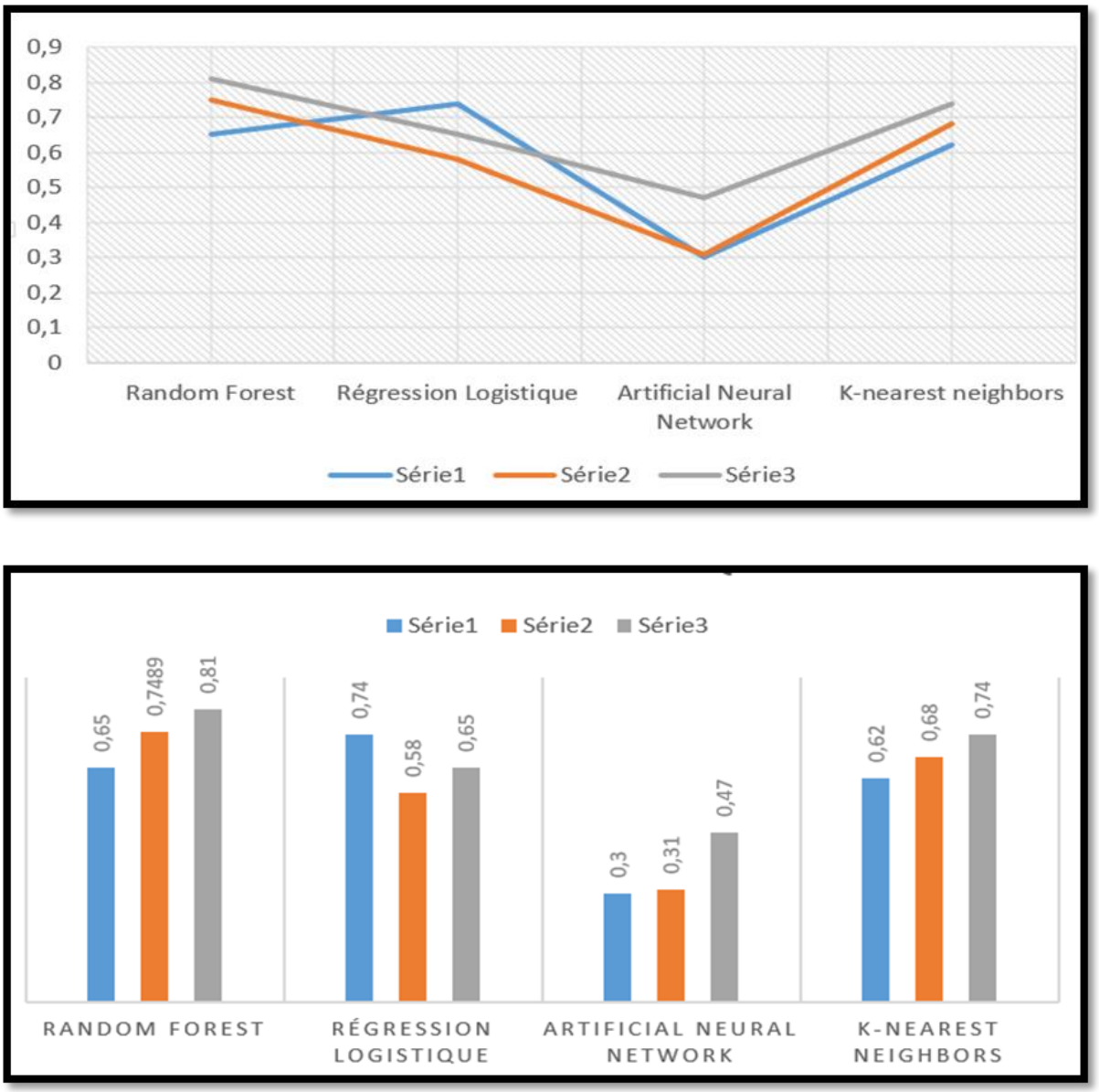

Figure IV. 13 : Représentation graphique

Les graphes représentent les résultats entre quatre modèles qui sont Random Forest, Régression Logistique, le K nearest neighbors et Artificial Neural Network par le taux de précision en fonction du nombre d'itération. On illustre que le modèle Random Forest est mieux que le modèle Régression Logistique sauf dans l'itération 1, et sa meilleure itération est l'itération 3 avec un taux de précision égale à 81% , on note aussi presque un faible résultat avec ANN au niveau des différents itérations mais une progression sur chaque itération avec

l'algorithme de KNN. Les modèles Random Forest, Régression Logistique et K nearest neighbors sont mieux que le modèle Artificial Neural Network. Ainsi, ANN qui est un algorithme de deep learning est plus intéressant avec d'autres types de données telles que le traitement d'image.

IV.9 L'arborescence de l'application

IV.9.1 L'application principale du projet

A travers les différentes phases d'étude et de conception par lesquelles nous sommes passés, nous avons pu mettre en place un système informatique pour la prédiction du diabète à travers les algorithmes d'apprentissage automatique et les données obtenues. Cependant, avant de présenter quelques interfaces de l'application, nous présentons d'abord l'application principale du projet qui permet de configurer le projet.

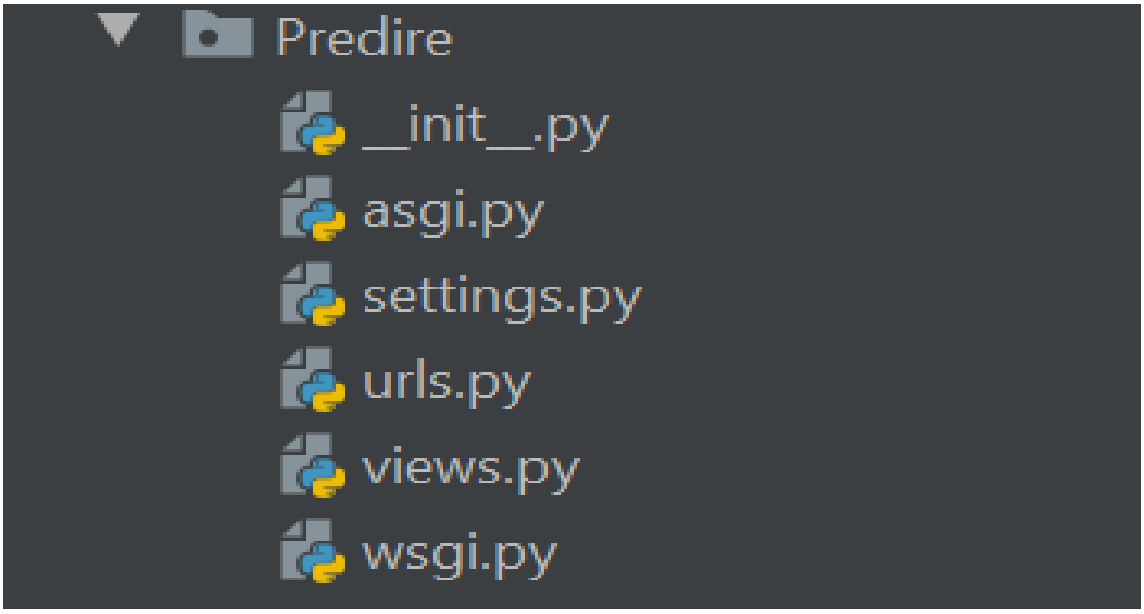

Figure IV. 14 : Application principale du projet

- **__init__.py** est un fichier vide qui demande à Python de traiter ce répertoire comme un package Python.
- **settings.py** contient tous les paramètres du site Web, y compris l'enregistrement des applications que nous créons, l'emplacement de nos fichiers statiques, les détails de configuration de la base de données, etc.
- **urls.py** définit les mappages URL-à-afficher du site. Bien que cela puisse contenir tout le code de mappage d'URL, il est plus courant de déléguer certains des mappages à des applications particulières, comme on le voit plus tard.

```python
from django.contrib import admin
from django.urls import path
from . import views

urlpatterns = [
    path("admin/", admin.site.urls),
    path("", views.home),
    path("predict/", views.predict),
    path("predict/result", views.result),
]
```

Figure IV. 15 : fichier urls.py

- **wsgi.py** est utilisé pour aider notre application Django à communiquer avec le serveur Web.
- **asgi.py** est une norme permettant aux applications Web et aux serveurs asynchrones Python de communiquer entre eux. ASGI est le successeur asynchrone de WSGI et fournit une norme pour les applications Python asynchrones et synchrones (alors que WSGI a fourni une norme pour les applications synchrones uniquement). Il est rétro compatible avec WSGI et prend en charge plusieurs serveurs et cadres d'application.
- **Views.py** est une fonction Python acceptant une requête Web et renvoyant une réponse Web. Cette réponse peut contenir le contenu HTML d'une page Web, une redirection, une erreur 404, un document XML, une image… ou vraiment n'importe quoi d'autre. La vue elle-même contient la logique nécessaire pour renvoyer une réponse.

```python
from django.shortcuts import render
import pandas as pd
import seaborn as sn
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score


def home(request):
    return render(request, 'home.html')


def predict(request):
    return render(request, "predict.html")
```

Figure IV. 16 : fichier views.py

IV.9.2 Les autres Paramètres

Sur la figure ci-dessous la liste allant d'account à voir constituent des paramètres de l'application. Ces paramètres sont des packages regroupant un ensemble de fonctionnalités communes à une tâche.

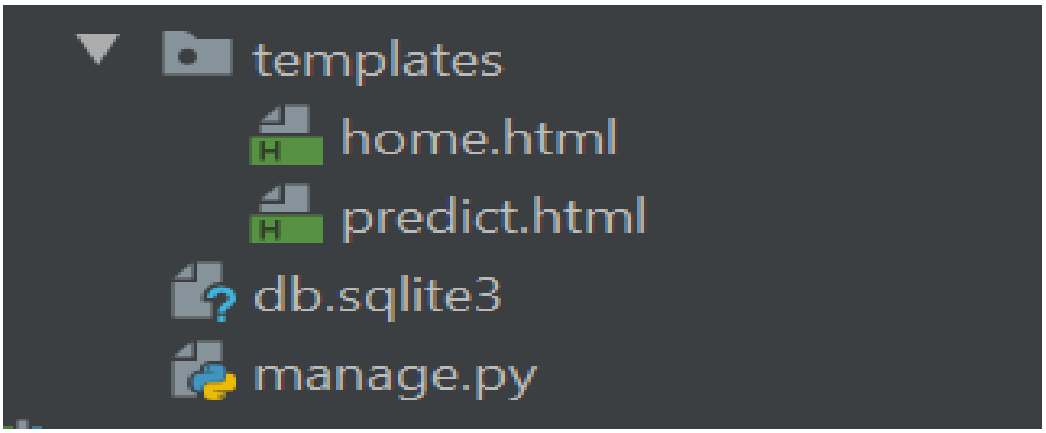

Figure IV. 17: Autres paramètres

On remarque que par rapport à une application principale un autre paramètre dénommé Template a les fichiers **home.html**, **predict.html** mais aussi db.sqlite ou sont stocké les données et le **manage.py**.

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

- **home.html:** est le fichier qui permet de gérer la page d'accueil**.**

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <title>HOME</title>
    <style type = text/css>
        div{
            color:white;
        }
        h1{
            color: white;
            font-family: arial, sans-serif;
            font-size: 60px;
            font-weight: bold;
            margin-top: 200px;
        }
        h2{
            color:whitesmoke;
            font-family: arial, sans-serif;
            font-size: 15px;
```

Figure IV. 18 : fichier home.html

- **predict.html:** est le fichier qui permet de gérer le formulaire de la page de prédiction**.**

```
</head>
<body>
    <u>Bienvenue</u>
    <div align = 'center' class="main">
        <h1>
            <u>Veuillez remplir les informations SVP:</u>
        </h1>
        <form action="result">
            <table>
                <tr>
                    <td align="right">Pregnancies:</td>
                    <td align="left"><input type="text" name="n1"></td>
                </tr>
                <tr>
                    <td align="right">Glucose:</td>
                    <td align="left"><input type="text" name="n2"></td>
                </tr>
                <tr>
                    <td align="right">Blood Pressure:</td>
```

Figure IV. 19 : fichier predict.html

IV.10 Présentation de l'application

Pour accéder à la page d'accueil il suffit de taper la commande suivante **"python manage.py runserver"** et de cliquer sur l'url proposé comme l'indique la figure ci-dessous.

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

```
(base) C:\Users\pc\PycharmProjects\Predire\Predire>python manage.py runserver
Watching for file changes with StatReloader
Performing system checks...

System check identified no issues (0 silenced).
November 14, 2022 - 15:41:52
Django version 4.1.1, using settings 'Predire.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CTRL-BREAK.
```

Figure IV. 20: Page de connexion

IV.10.1 Page d'accueil

Le menu d'accueil ou page d'accueil est la première page visible après avoir cliqué sur l'url. Il est composé d'une seule information ainsi que le bouton de fonctionnalité permettant aux utilisateurs de pouvoir accéder à la page de prédiction.

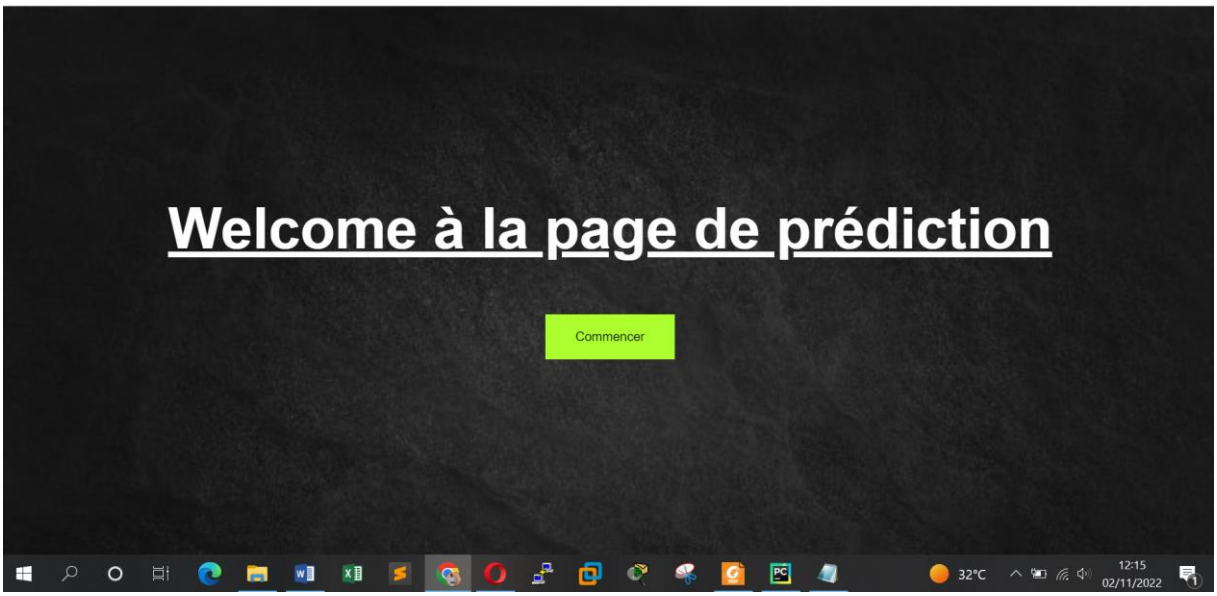

Figure IV. 21 : Page d'accueil

IV.10.2 Page de prédiction

La fonctionnalité de prédiction est réservée aux utilisateurs (médecin). Ainsi, l'utilisateur une fois sur la page il suffit juste de remplir le formulaire en face de lui selon les informations demandées et le modèle en question fait le travail et donne le résultat final en bas de page.

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

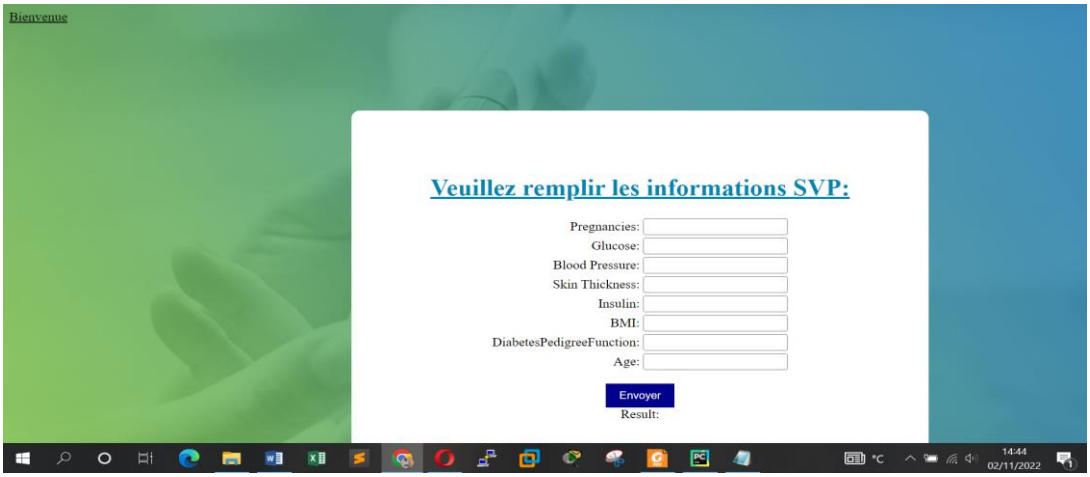

Figure IV. 22: Page de prédiction

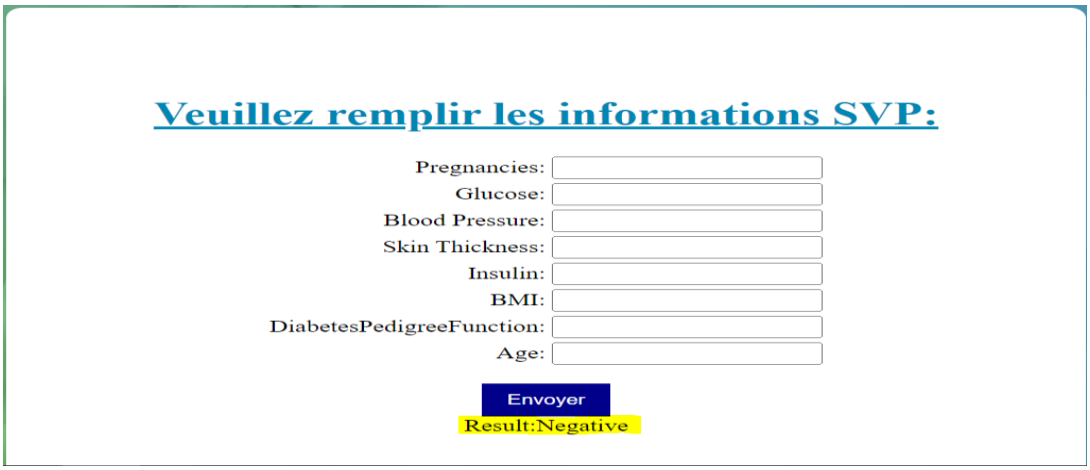

Figure IV. 23: Affichage résultats

IV.10.3 Conclusion

Tout compte fait, ce chapitre nous a permis de conduire la partie implémentation de notre projet. En effet, on a présenté la stratégie suivie pour le processus de classification ainsi que la base de données utilisée Pima indian diabetes database dans ce travail. Nous avons choisi 3 algorithmes de classification pour la partie prédiction de diabètes dans laquelle nous avons comparé leur comportement. Rappelons que ces algorithmes de classification sont : la Régression Logistique, Random Forest et l'algorithme ANN. Les résultats de performances ont montré clairement l'avance de l'algorithme Random Forest contre tous les autres algorithmes choisis dans cette étude. A la fin, on a développé une application web qui nous permet de prédire si une personne donnée est diabétique ou pas à partir de ces informations médicales.

Conclusion Générale et Perspectives

Dans le cadre de notre mémoire, nous avons effectué une comparaison entre quatre algorithmes d'apprentissage automatique à savoir : Random forest, K nearest neighbors, Régression logistique et Artificial neural network, les résultats expérimentaux obtenus pour l'ensemble de données 'Pima indian diabetes database' montre que Random forest est meilleure que les autres algorithmes en termes de sa grande précision. Sur la base des algorithmes nous avons besoin d'un moyen pour rendre les modèles applicatifs pour tout le monde, nous avons développé une solution basée sur une application web dans le but d'aider les personnes à prédire s'il souffre de diabète ou non.

Pour les travaux futurs, plusieurs pistes peuvent être explorées. Nous envisagerons de proposer des capteurs a insulines sur des patients et d'interconnecté avec un réseau IoT, appliquer la même expérimentation sur d'autres bases de données de diabète ou même de type différents pour confirmer les résultats obtenus. Améliorer les différents algorithmes utilisés pour avoir de meilleurs résultats en termes de précision, mais aussi la construction d'une application Android parallèle avec notre application web qui permet d'aider les personnes diabétiques de suivre leur situation médicale.

Bibliographie et Webographie

[1] P. Kasemthaweesab and W. Kurutach, “Association analysis of Diabetes Mellitus (DM) with complication states based on association rules,” in *Proc. 7th IEEE Conference on Industrial Electronics and Applications,* July 2012, pp. 1453-1457

[2] http://www.who.int/mediacentre/factsheets/fs312/en/ consulté le 25/05/2022

[3] K. Zarko Gianni, E. Litsa, K. Mitsis, P. Y. Wu, C. D. Kaddi, C. W. Cheng, and K. S. Nikita, “A review of emerging technologies for the management of diabetes mellitus,” *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 12, pp. 2735-2749, 2015

[4] https://fr.wikipedia.org/wiki/Apprentissage_automatique 15/06/2022

[5] https://fr.wikipedia.org/wiki/Apprentissage_automatique consulté le 15/06/2022

[6]SIEMENS,[Enligne]Available:https://new.siemens.com/fr/fr/entreprise/stories/industriedu-futur/l-intelligence-artificielle-dans-l-industrie.html

[7]SciencesTech,[Enligne].Available:https://www.rts.ch/info/sciences-tech/12585675-intelligence-artificielle-et-agriculture-lhumain-ne-sera-pas-remplace-de-sitot.html

[8]«agenceecofine,»[Enligne].Available:https://www.agenceecofin.com/entreprendre/1103-86076-au-maroc-l-intelligence-artificielle-pour-optimiser-l-agriculture

[9]CScience,[Enligne].Available:https://www.cscience.ca/2021/09/21/sami-4-0-le-robotcueilleur-au-service-de-la-recolte-au-quebec/][REUSSIRMachinisme,[Enligne].Available:https://www.reussir.fr/machinisme/kubotainvestit-dans-les-robots-volants-cueilleurs-de-fruits

[10]L'EXPRESS,[Enligne].Available:https://www.lexpressmontcalm.com/article/2021/06/15/lindustrie-4-0-s-invite-dans-le-monde-de-l-agriculture

[11] Pensée artificielle. Machine Learning pour débutant : Introduction au Machine Learning. [En ligne].Disponible sur : http : //penseeartificielle.fr/introduction-au-machine Learning/

[12]https://waytolearnx.com/2018/11/differenceentreapprentissagsuperviseetnonsupervise.html consulté le 12/08/2022

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

[13]https://www.universalis.fr/encyclopedie/apprentissage-profond-deep-learning/1 differentstypesdapprentissagemachine/#:~:text=On%20distingue%20usuellement%20au%20moins, et%20l'apprentissage%20non%20supervis%C3%A9 consulté le 25/10/2022

[14] Ilemona S.Atawodi. (2019).A Machine Learning Approach to Network Intrusion Detection System Using K Nearest Neighbors and Random Forest. Thèse de master : universitédeSouthernMississippi.52p[enligne].Disponiblesur:http://aquila.usm.edu/cgi/viewcontent.cgi ?article=1707&context=masters thèses

[15] Benzaki, Y.Mr Mint. (2018).Introduction à l'algorithme K Nearest Neighbors (KNN). [En ligne].Disponible sur : https : //mrmint.fr/introduction-k-nearest-neighbors consulté le 25/11/2022

[16]Harrison,O.towardsdatascience.(2018).MachineLearningBasicswiththeKNearestNeighbor Algorithm.[Enligne].Disponiblesur:https://towardsdatascience.com/machine-learning-basics-with-the-k-nearestneighbors-algorithm-6a6e71d01761

[17]Gupta,P.towardsdatascience(2017).DecisionTreesinMachineLearning.[enligne].Disponiblesur:https ://towardsdatascience.com/decision-trees-in-machinelearning-641b9c4e8052

[18] Shrivastav, A.to wards datascience.Almost Everything You Need To Know About DecisionTrees(WithCode).[enligne].Disponiblesur://towardsdatascience.com/almost-everything-you-need-to-know-aboutdecision-trees-with-code-dc026172a284

[19]JavaTpoint.RandomForestAlgorithm.[Enligne].Disponiblesur:https://www.javatpoint.com/machine-learning-random-forest-algorithm

[20]Tutorialspoint.ClassificationAlgorithmsRandomForest.[enligne].Disponiblesur:https://www.tutorialspoint.com/machine learning with python /machine learning with python classification algorithms random forest.htm

[21] Mémoire : un système de prévision la cherche électrique

[22]FikirteGirmaWoldeMichael,SumitraMenaria,"PredictionofDiabetesusingDataMiningTechniques, Dept.of ComputerScience and Engineering'', Sumitra.Menaria@paruluniversity.ac.in, Proceedings of the 2nd International conference on Trends in Electronics and Informatics (ICOEI 2018)

[23] Terry Jacob Mathew, Elizabeth Sherly," Analysis Supervised Learning Techniques for Cost Effective Disease Prediction using Non-Clinical Parameters", sherly@iitmk.ac.in,IIITM-KTechno park, Trivndrum, July 05-07,2018

[24] M., Butwal, and S., Kumar, "A Data Méninge Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier," International Journal of Computer Applications, vol. 120, pp. 0975–8887, 2015

[25] A.K., Dewangan, and P., Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," International Journal of Engineering and Applied Sciences, vol. 2, 2015

[26] M.R., Devi, and J.M. Shyla, "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus," International Journal of Applied Engineering Research, vol. 11, pp. 727–730, 2016

[27] S., Hina, A., Shaikh, and S., Abul Sattar, "Analyzing Diabetes Datasets using Data Mining," Journal of Basic & Applied Sciences, vol. 13, pp. 466–471, 2017

[28] A.Mujumdar, V. Vaidehi Prédiction du diabète à l'aide d'algorithmes d'apprentissage automatique Conférence internationale sur les tendances récentes en informatique avancée », 2019, ICRTAC (2019)

[29] N. Yuvaraj, KR SriPreethaa Prédiction du diabète dans les systèmes de santé à l'aide d'algorithmes d'apprentissage automatique sur le cluster Hadoop Calcul de cluster. , 22 (2017), p. 1 – 9

[30] D. Sisodia, DS Sisodia Prédiction du diabète à l'aide d'algorithmes de classification Process Comput. Sci. , 132 (2018), p. 1578 – 1585

[31] erveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K., 2016. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Computer Science 82, 115–121. doi:10.1016/j.procs.2016.04.016

[32] Nai-Arun, N., Sittidech, P., 2014. Ensemble Learning Model for Diabetes Classification. AdvancedMaterialsResearch931- 932, 1427–1431. doi:10.4028/www.scientific.net/AMR.931-932.1427

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

[33] K. Dalakleidi, K. Zarkogianni, A. Thanopoulou, and K. Nikita, “Comparative assessment of statistical and machine learning techniques towards estimating the risk of developing type 2diabetes and cardiovascular complications,” *Expert Systems*, 2017

[34] A. Ashiquzzaman, A. K. Tushar, M. Islam, J.-M. Kim et al. ``Reduction of overfitting in diabetes prediction using deep learning neural network," *Ari preprint* arXiv:1707.08386, 2017

[35] J. Zhu, Q. Xie, K. Zheng. “An Improved Early Detection Method of Type-2 Diabetes MellitusUsingMultipleClassifier Systems”. *Information Sciences*, volume 292, pages 1-14, 2015.

[36] M. Kumari, Dr. R. Vohra, and A. Arora, “Prediction of Diabètes using Bayesian Network*,”* *International Journal of Computer Science and Information Technologies*, vol. 5, pp. 5174-5178, 2014

[37] T. Santhanam and M.S Padmavathi, “Application of KMeans and Gentil Algorithms for Dimension Reduction by Integrating SNM for Diabetes Diagnosis,” *Procedia Computer Science*, vol. 47, pp. 76-83, 2015

[38] J. Vijayashree and J. Jayashree, “ An Expert System for the Diagnosis of Diabetic Patients using Deep Neural Networks and Recursive Feature Elimination,” *International Journal of Civil Engineering and Technology*, vol. 8, pp. 633-641, Déc. 2017

[39] L. B. Goncalves and M. M. Bernardes, “Inverted Hierarchical Neuro-Fuzzy BSP System: A Novel NeuroFuzzy Model for Pattern Classification and Rule Extraction in Databases,” *in IEEE Transactions on Systems, Man, and Cybernetics*, vol. 36, no. 2, pp. 236-248, Mar. 2006

[40] L. Han, S. Luo, H. Wang, L. Pan, X. Ma and T. Zhang, "An Intelligible Risk Stratification Model Based on Pairwise and Size Constrained Kmeans," *in IEEE Journal of Biomedical and HealthInformatics*, vol.21, no.5, pp.1288-1296, Sept. 2017

[41]ISARTA Infos, [Enligne].Available: https://isarta.com/infos/intelligence-artificielle-quels-sont les-défis-des-organisations/

[42]UniversitédeLAVA[Enligne].Available:https://www.administrationnumerique.chaire.ulaval.ca/recherches/les-applications-et-les-defis-de-lintelligenceartificielle-dans-le-secteur-public

Le succès n'est pas final, l'échec n'est pas fatal. C'est le courage de continuer qui compte !!!!

[43]DGITALWEEK,[Enligne].Available:https://mydigitalweek.com/ia-conversationnel-securiteles-tendances-de-lia-pour-2021/

[44]JeanSébastienDesroches,[Enligne].Available:https://www.lavery.ca/fr/publications/nospublications/3009-intelligence-artificielle-la-delicate-interaction-entre-les-defisjuridiques-et-technologiques.html

[45] S.MAROUANE, «https://revues.imist.ma/index.php/DOREG/article/view/21207»A. P. R. Bruno le Maire Cédric O, STRATÉGIE NATIONALE POUR L'INTELLIGENCE ARTIFICIELLE PRÉSENTATION DU VOLET ÉCONOMIQUE

[46][Enligne].Available:https://isarta.com/infos/intelligence-artificielle-quels-sont-les-defis-des organisations/

[47]LesEchosStart,[Enligne].Available:https://start.lesechos.fr/innovationsstartups/techfutur/comprendre-les-enjeux-de-lintelligence-artificielle-en-4-points-1177670

[48] https://www.kaggle.com/ consulté le 22/07/2022

[49] https://fr.wikipedia.org/wiki/Python_(langage) consulté le 05/09/2022

[50]https://m.facebook.com/Nkfondation/photos/a.2307858662815735/2787456238189306/?type=3&refsrc=deprecated&_rdr consulté le 11/07/2022

[51]https://www.ionos.fr/digitalguide/web-marketing/search-engine-marketing/quest-ceque-keras/ consulté le 18/05/2022

[52] https://www.lebigdata.fr/tensorflow-definition-tout-savoir consulté le 02/11/2022

[53] https://fr.wikipedia.org/wiki/Scikit-learn consulté le 02/10/2022

[54]https://fr.wikipedia.org/wiki/NumPy#:~:text=NumPy%20est%20une%20biblioth%C3%A8que%20pour,math%C3%A9matiques%20op%C3%A9rant%20sur%20ces%20taleaux.&text=NumPy%20est%20la%20base%20de,Python%20autour%20du%20calcul%20scientifique consulté le 25/08/2022

[55] Site de Wikipédia sqlite https://fr.wikipedia.org/wiki/SQLite consulté le 03/10/2022

[56] https://www.sisense.com/glossary/data -standardization/ consulté le 25/06/2022