République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A. Mira de Béjaïa

Faculté des Sciences Exactes

Département d'Informatique

Mémoire de Fin d'Etude

En vue de l'obtention du diplôme de Master recherche en Informatique

Option : Intelligence Artificielle

Thème

Prédiction du Diabète géstationnel en utilisant les techniques de l'intelligence artificielle

Réalisé par :

- M[lle] Nabila YAYA
- M[lle] Melissa LACHI

Devant le jury composé de

Présidente :	*D[r] ALOUI Soraya*	M.C.A - Université de Béjaïa.
Examinatrice :	*D[r] BACHIRI Lina*	M.C.A - Université de Béjaïa.
Encadrante :	*D[r] EL BOUHISSI Houda*	M.C.A - Université de Béjaïa.
CO-Encadrant :	*P[r] AMROUN Kamal*	Professeur - Université de Béjaïa.

Promotion 2021 - 2022

Remerciements

En tout premier lieu, nous remercions Dieu le tout puissant,

de nous avoir donné la santé et la force pour pour dépasser toutes les difficultés et terminer ce travail,

Nous tenons à exprimer notre chaleur remerciements à notre encadrante

Madame Houda El Bouhissi. Nous le remercions de nous avoir encadré, orienté, aidé et conseillé, et pour sa disponibilité

Nous adressons nos sincères remerciements à Monsieur Kamel Amroun

de nous avoir aider et donner le principe pour entamer notre mémoire,

nous tenons à remercier toutes les personnes qui ont contribué par leurs paroles, leurs écrits durant nos recherches.

Dédicaces

A mon très chers grand père Rabah, la personne la plus cher à mon cœur,

je le remercie pour ses conseils tous les jours, pour ses sacrifices, et ses encouragements. il est un exemple pour moi par ses réussites,

ses qualités humaines et sa bénédiction,

Aucune dédicace ne peut exprimer la profondeur de mes sentiments, respects, et amour,

Que dieu vous protéger et vous accorder une bonne santé et longue vie,

A mon très cher père Moussa pour son soutien et ses encouragements,

et prières habituels qui m'a donné la force à résister malgré tous les difficultés,

Que dieu vous préserve de toute trahi,

Que dieu t'accorde une longue vie pleine de bonheur et une bonne santé,

A mes chères sœurs Katia, Hanane

A mes frères Saad et Saadi,

A tous mes amis,

A tous qui sont chéres à mon coeur,

Que dieu vous protège tous,

Nabila.

Dédicaces

Je dédie ce travail ,

A la mémoire de mon oncle LACHI Abdelkader qui m'a toujours encouragé dans mes études , que Dieu garde son âme dans son vaste paradis.

A toute ma famille, berceau de ma culture. Sans elle je ne serai pas ce que je suis aujourd'hui.

A mes chéres grands parents pour leurs prières tout au long de mes études.

A mes parents qui ont attendu avec patience les fruits de leur bonne education et de leur devouement .

A mes chères tantes Zineb, Samira, Sabiha, Radia et mes chères soeur zoulikha , katia , nawal pour leurs encouragements permanents, et leur soutien moral.

A mes chères cousin Ahmed , Hichem , Nassim, Faycel, adlen , said et cousines Wissem , Imen, wiza , sarah , hamida , melissa maria pour leurs amours et leurs encouragement.

A tous mes amis qui m'ont toujours encouragé , et à qui je souhaite plus de succès .

A tous qui sont chéres à mon coeur .

Melissa.

Table des matières

Table des figures

Liste des abréviations

IA	**I**ntelligence **A**rtificielle
BMI	**B**ody **M**asse **i**ndex
DNN	**D**eep **N**eural **N**etwork
ML	**M**achine **L**earning
SVM	**S**upport **V**ector **M**achine
RF	**R**andom **F**orest
OMS	**O**rganisation **M**ondiale de la **S**anté
AVC	**A**ccident **V**asculaire **C**érébral
KNN	**K**- **N**earest **N**eighbors
CNN	**C**onvolutional **N**eural **N**etwork
ROC	**R**eceiver **O**perating **C**haracteristic
OR	**O**apport de **C**otes
NB	**N**aive **B**ayes
DT	**D**ecision **T**ree
LR	**L** ogistic **R**egression
AUC	**A**rea **U**nder **C**urve
RNA	**R**éseau **N**eurones **A**rtificiels
AB	**A**ada **B**Boost

Chapitre 1

Introduction Générale

1.1 Introduction

L'intelligence artificielle est un processus d'imitation de l'intelligence humaine qui repose sur la création et l'application d'algorithmes exécutés dans un environnement informatique dynamique. Son but est de permettre à des ordinateurs de penser et d'agir comme des êtres humains et réaliser des tâches complexes.

L'IA est l'un des domaines qui est en cours de développement dans tous les secteurs tel que la santé qui est la plus en pointes.

L'apprentissage automatique (machine Learning en anglais), est un champ d'étude de l'intelligence artificielle qui peut être la meilleure solution de développer des systèmes de prévision fiables afin de diminuer la propagation de certaines maladies. Son application en médecine permettant à la machine d'analyser les données par elle-même et de fournir des estimations, dans le but d'aider les médecins de différentes spécialités à établir des diagnostiques de plus en plus précis, de prendre des décisions optimale, de choisir des traitements appropriés et de faire des prédictions dans le but de réduire les risques de complications des maladies sur la santé du patient.

Le domaine de recherche présenté dans ce mémoire est la prédiction du diabète. Le but est de permettre éventuellement aux médecins de démasquer plus aisément les patients à risque de développer le diabète.

1.2 Problématique

Comme le diabète est devenu une maladie de siècle au cours des dernières années, les médecines demandent aux personnes de prévenir face à cette maladie en suivant quelques consignes recommandées pour déminer les facteurs de risque au but d'éviter la survenu et les complications de cette dernière.

Cependant, comme le nombre des patients de cette maladie augmente, il devient de plus en plus difficile de la contrôler. Les chercheures ont mise en place plusieurs systèmes qui aident à réduire le risque d'infection avec une prédiction précoce et une connaissance des facteurs les plus importants qui la contrôlent.

1.3 Objectifs

L'objectif de ce mémoire est d'utiliser les 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 complication de cette maladie chronique sur la santé du patient.

En effet, dans notre travail on s'intéresse à donner des solutions pour la prédiction de cette maladie afin d'aider les personnes susceptibles d'atteindre le diabète de prédire cette maladie en utilisant les algorithmes d'apprentissage automatique.

1.4 Méthodologie du recherches

Étape de recherche et d'analyse : qui établit une analyse approfondie de l'état de l'art des différentes approches proposées par les chercheurs dans le cadre de prédiction du diabète et qui fait une comparaison selon les performances des méthodes à partir des taux positives et taux négatives.

Étape d'identification du problème et de la proposition d'une solution : qui permet de définir la problématique et la solution proposée.

Étape d'implémentions et d'expérimentation du système proposé : qui met en évidence le système proposé, son fonctionnement et son intérêt.

1.5 Organisation du mémoire

Le mémoire est organisé comme suit :

Le deuxième chapitre est consacré aux généralités sur la maladie du diabète. On présentera la définition de cette maladie, ses types, ses complication et son traitement.

Dans le troisième chapitre, nous élaborerons l'état de l'art qui représentera les travaux connexes les plus importants que nous synthétiserons, nous présenterons ceci dans un tableau qui contiendra les grandes lignes de chaque document synthétisé,par la suite, nous procéderons à une analyse comparative de ces travaux.

Le quatrième chapitre porte sur la description de notre approche.

Dans le cinquième chapitre, nous présentons les outils de programmation et l'implémentassions de notre système et les résultats d'exécution, ainsi que les logiciels choisis .

Enfin, ce mémoire est clôture par le chapitre six qui donne les conclusion et perspectives de ce travail .

Chapitre 2

Généralités

2.1 Introduction

Le diabète est une maladie qui empêche le corps d'utiliser correctement l'énergie fournie par les aliments ingères. Par ailleurs, la maladie survient lorsque le pancréas ne sécrète plus d'insuline ou lorsque le corps devient résistant à la quantité d'insuline produit.

Il existe principalement deux types de diabète : le type 1 appelé diabète insulinodépendant ou diabète juvénile. Ces symptômes sont notamment les suivantes : émission d'urine, soif excessives, faim constante, perte de poids, altération de la vision et la fatigue. Le type 2 appelé diabète non insulinodépendant ou diabète de l'adulte , les symptômes peuvent être similaire a ceux du diabète de type 1, mais ils sont souvent moins marques ou absents.

En outre, il existe un autre type de diabète appelé diabète gestationnel qui se d'enveloppe pendant la grossesse il est associé à un risque à long terme de diabète de type 2. Le sur poids , le manque d'exercice, les antécédent familiaux et le stress est augmenté le risque possible de diabète et le mauvais contrôle de dosage du sucre dans le sang peut entraîner des complications très grave (cécité, cataracte, thrombose , néphropathie...)

2.2 Définition du diabète

Le diabète est une maladie chronique caractérisé par un trouble de la régulation naturelle du taux de sucre dans le sang (glycémie). Le diabète est l'un des principaux tueurs au monde, avec l'hypertension artérielle et le tabagisme, selon l'Organisation Mondial de la Santé (OMS). Cette maladie constitue un

problème de santé publique majeure et malgré les efforts de prévention, la pandémie se poursuit.

Le diabète parmi les maladie les plus propagé au monde , en 2019, le diabète affecte plus de 463 millions de personnes dans le monde selon Fédération Internationale pour le diabète (9,3% de la population adulte). En 2021, le diabète affecte plus de 537 millions de personnes dans le monde (soit 1 personne sur 10) [44].

Lorsque nous mangeons, les aliments sont dégradé en glucose (sucre). Ce glucose fournit de l'énergie au corps afin qu'il puisse fonctionner correctement en puisant dans ses ressources. Pendant la digestion, le sang transporte le glucose dans tout le corps et vient alimenter les cellules. Cependant, pour que le sucre présent dans le sang puisse ensuite être transmis aux cellules, le corps a besoin d'insuline, une hormone sécrétée par le pancréas.L'insuline agit donc comme une cle permettant au glucose de passer du sang aux cellules de notre corps[42].

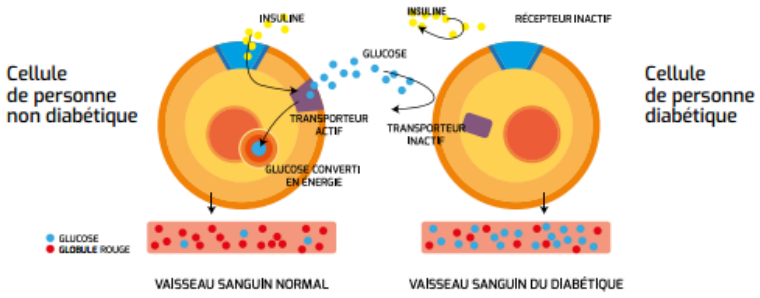

FIGURE 2.1 – Qu'est ce que le diabète [32]

2.3 Complications du Diabète

Le diabète est également associé à des complications à court terme (hypoglycémie),et des complications à long terme (L'hyperglycémie) apparaissent après plusieurs années en cas de diabète non ou mal équilibré et ces complications chroniques touchant de nombreuses parties de l'organisme.

Au nombre des complications possibles figurent maladie cardio-vasculaire, lésions nerveuses, lésions oculaires,lésions rénales, dommages aux pieds.

— **Maladie cardiovasculaire :** le diabète augmente considérablement le risque de divers problèmes cardiovasculaires, y compris la maladie coronarienne avec douleur thoracique (angine), crise cardiaque, accident vasculaire cérébral et rétrécissement des artères (athérosclérose).

— **Lésions nerveuses (neuropathie) :** l'excès de sucre peut endommager les parois des minuscules vaisseaux sanguins (capillaires) qui nourrissent vos nerfs, en particulier dans vos jambes. Cela peut provoquer des picotements, des engourdissements, des brûlures ou des douleurs qui commencent généralement au bout des orteils ou des doigts et se propagent progressivement vers le haut.

— **Lésions rénales (néphropathie) :** les reins contiennent des millions de petits amas de vaisseaux sanguins (glomérules) qui filtrent les déchets de votre sang. Le diabète peut endommager ce délicat système de filtrage. Des dommages graves peuvent entraîner une insuffisance rénale ou une maladie rénale terminale irréversible, qui peut nécessiter une dialyse ou une greffe de rein.

— **Lésions oculaires (rétinopathie) :** le diabète peut endommager les vaisseaux sanguins de la rétine (rétinopathie diabétique), pouvant entraîner la cécité. Le diabète augmente également le risque d'autres problèmes de vision graves, tels que la cataracte et le glaucome.

— **Dommages aux pieds :** des lésions nerveuses dans les pieds ou une mauvaise circulation sanguine dans les pieds augmentent le risque de diverses complications du pied. Sans traitement, les coupures et les cloques peuvent développer des infections graves, qui guérissent souvent mal. Ces infections peuvent finalement nécessiter l'amputation d'un orteil, d'un pied ou d'une jambe.

— **Maladies de la peau :** le diabète peut vous rendre plus vulnérable aux problèmes de peau, y compris les infections bactériennes et fongiques.

— **Déficience auditive :** Les problèmes auditifs sont plus fréquents chez les personnes atteintes de diabète. [7]

2.4 Diagnostic du diabète

La maladie de diabète est détecté le plus souvent lorsque les complication à longe terme s'exprime .Le diabète de type 1 est diagnostiqué en présence des symptômes qui apparaissent rapidement, par contre le diabète de type 2 est fréquemment identifié par hasard.

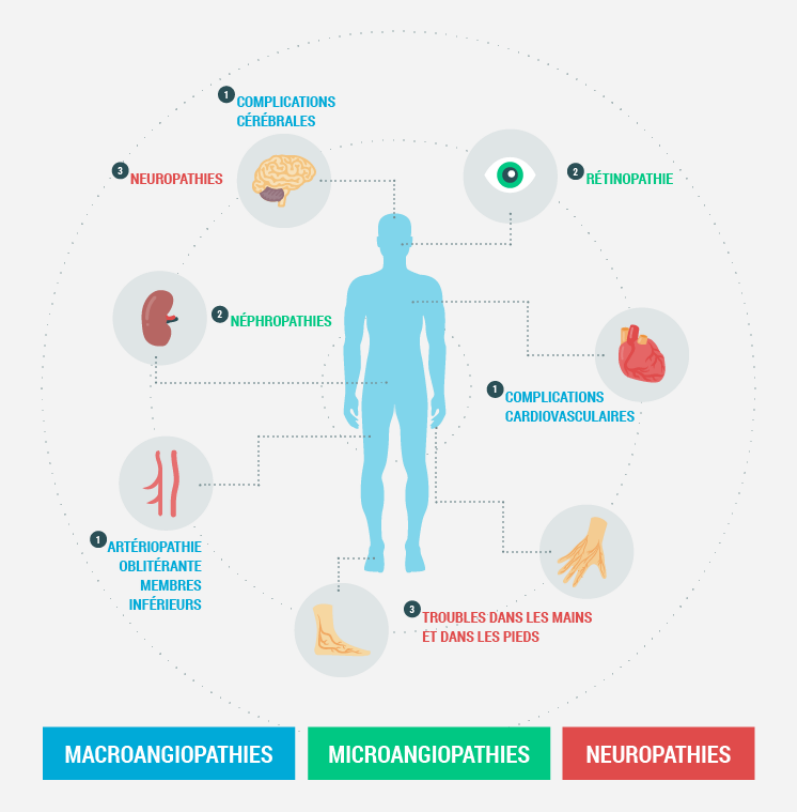

FIGURE 2.2 – Complications du diabète [17]

2.4.1 Dépistage du diabète

Toute personne ayant des membres de sa famille atteints de diabète de type 2 doit se faire dépister régulièrement car un risque héréditaire existe (si l'un des deux parents est diabétique de type 2, le risque héréditaire est de 40 % si les deux parents sont atteints, le risque monte à 70 %). Pour le diabète de type 1, le risque de transmission aux enfants est de 6 % si le père est diabétique, 2 ou 3 % si la mère l'est, et 30 % si les deux parents sont atteints de diabète ,les personnes en sur poids ou souffrant de troubles de la glycémie doivent également se plier au dépistage. Il en va de même pour les femmes ayant développé du diabète pendant leur grossesse (diabète gestationnel) ou ayant mis au monde un bébé de faible poids. Le dépistage est également recommandé aux personnes de plus de 65 ans [64].

2.4.2 Signes du diabète

Les signes énumérés ci-dessus ne sont pas des symptômes uniquement liés au diabète donc ce n'est pas facile de savoir par soi-même si l'on est diabétique ou non c'est pour cela la diagnostique du diabète se fait par un teste de prise du sang.Il existe deux façons pour tester et contrôler la glycémie :

Le test d'hémoglobine glyquée : permet de mesurer la glycémie a jeune au cours des 3 derniers mois.

L'auto-test : c'est un teste à faire chez soi qui mesure le glucose pour contrôler plusieurs fois par jour sa glycémie capillaire (sur une goutte de sang) à des moments précis. C'est ce qu'on appelle l'autosurveillance glycémique (ASG)[13].

FIGURE 2.3 – Glucomètre

2.4.3 Analyse des résultats de la glycémie

D'une manière générale, les résultats des dosages de glycémie peuvent être interprétés de la manière suivante :

— **Si à jeun, elle est inférieure à 1,10 g/l :** vous êtes dans les normes. Prochaine prise de sang dans 3 ou 4 ans, sauf si un risque est soupçonné entre-temps.

— **Si elle est supérieure ou égale à 1,10 g/l et inférieure à 1,25 g/l :** vous êtes en situation de prédiabète. Il est important de reprendre une activité physique et si nécessaire perdre du poids. Prochaine prise de sang dans un an .

— **Si elle est supérieur ou égale à 1,26 g /l et inférieure à 2 g/l :** votre médecin va vous prescrire un second dosage. "On compte une petite semaine entre les deux, dit le Pr Jean-Pierre Riveline,

endocrinologue. "Mais s'il y a une infection, on attend qu'elle soit guérie pour refaire la glycémie". Si cette seconde glycémie est à nouveau supérieure ou égale à 1,26 g/l, le diabète est confirmé.

— **Si elle est d'emblée supérieur ou égale à 2 g/l :** le diagnostic de diabète est posé.[39]

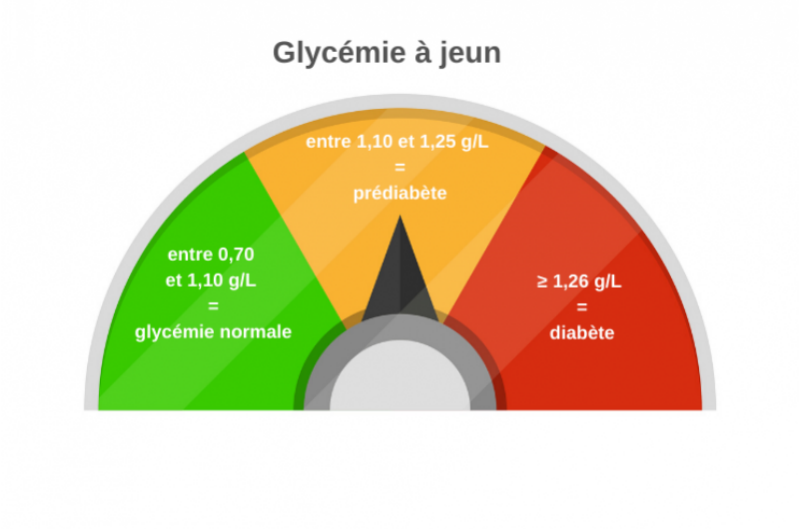

FIGURE 2.4 – Lecteur de la glycémie [14]

2.5 Classification du diabète

La classification du diabète selon l'organisation mondiale de la santé (OMS) est globalement classés en trois catégories :

— Diabète de type 1

— Diabète de type 2

— Diabète gestationnel

2.5.1 Diabète de type 1

Regroupe le diabète principalement attribuable à la destruction des cellules bêta du pancréas, qui s'accompagne d'une carence en insuline susceptible d'évoluer vers une acidocétose diabétique. Cette forme de diabète comprend les cas attribuables à un processus auto-immun et les cas dont la cause de la destruction des cellules bêta est inconnue [68].

2.5.1.1 Symptômes

Les symptômes du diabète de type 1 apparaissent lorsque la maladie est déjà avancée. Le plus souvent, ce sont :

— Une augmentation inhabituelle de la soif et de la faim.

— Un besoin fréquent d'uriner, ce qui peut entraîner des problèmes de pipi au lit chez un enfant jusque-là propre.

— Une fatigue anormale.

— Une mauvaise cicatrisation des blessures et des coupures.

— Une peau sèche sujette à démangeaisons.

— Ses infections fréquentes des gencives, de la vessie, du vagin, de la vulve ou du prépuce.[34]

2.5.1.2 Traitement

Le traitement du diabète de type 1 repose sur des injections sous-cutanées d'insuline, plusieurs fois par jour, pour compenser son défaut de production par l'organisme. On utilise aujourd'hui des analogues d'insuline humaine, produits par des bactéries génétiquement modifiées [26].

FIGURE 2.5 – Traitement du diabete de type 1 [8]

2.5.2 Diabète de type 2

Les personnes atteintes de diabète de type 2 sécrètent de l'insuline, mais cette hormone régule avec moins d'efficacité le taux de sucre dans leur sang. Ce taux, appelé glycémie, reste anormalement élevé

après un repas, ce qui est la définition du diabète. Petit à petit, le pancréas s'épuise à sécréter des quantités croissantes d'insuline. Également appelé diabète gras, ou diabète non insulinodépendant, le diabète de type 2 touche surtout les personnes en surpoids ou obèses, sédentaires, le plus souvent après 45 ans.

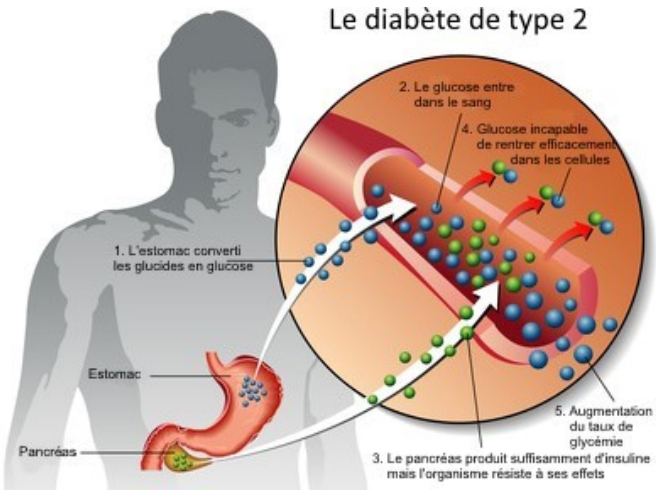

FIGURE 2.6 – Le diabète de type 2 [49]

2.5.2.1 Symptômes

Sont les mêmes que celles du type 1. Les symptômes du diabète de type 2 sont discrets et il est le plus souvent diagnostiqué à l'occasion d'une prise de sang. Les symptômes des complications du diabète de type 2 sont une difficulté à cicatriser, une perte de sensibilité au niveau des pieds, des troubles de la vision, une insuffisance rénale, un infarctus ou un AVC [35].

2.5.2.2 Traitement

Le traitement du diabète de type 2 repose sur :

— la diminution et le contrôle du poids par une alimentation équilibrée,

— une activité physique régulière,

— l'arrêt du tabac le cas échéant.

Si ces mesures ne se sont pas suffisantes, des médicaments antidiabétiques peuvent être prescrits, d'abord sous forme de comprimés puis, si nécessaire, en injections. L'objectif du traitement est de réduire le risque de complication en maintenant le taux sanguin de sucre dans des valeurs normales [37].

FIGURE 2.7 – Traitement du diabète type 2 [55]

2.5.3 Diabète gestationnel

Le diabète gestationnel, aussi appelé diabète de grossesse, est une situation d'intolérance au glucose qui apparaît chez des femmes enceintes qui auparavant n'avaient aucun antécédent diabétique. Chez certaines, la maladie, qui se caractérise par une intolérance au glucose entraînant une glycémie supérieure à 1,26 g/l de sang, va persister au-delà de la grossesse. Le diabète gestationnel concernerait environ une grossesse sur 20. La grossesse est en elle-même diabétogène, mais la résistance au glucose ne se révèle pathologique que dans une minorité des cas [20].

2.5.3.1 Le traitement

Généralement, une saine alimentation qui tient compte des portions et de la répartition des glucides (sucres) ainsi qu'une bonne hygiène de vie (gestion du stress, sommeil adéquat et activité physique) sont suffisantes pour gérer le diabète de grossesse.

Si les valeurs glycémie demeurent trop élevées, le médecin prescrira alors des injections d'insuline ou, dans certains cas, des antihypergycémiants oraux. L'insuline est tout à fait sécuritaire pendant la grossesse.

Valeurs cibles de glycémie pour la majorité des femmes enceintes :

— À jeun : inférieure à 5,3 mmol/L

— 1 heure après un repas : inférieure à 7,8 mmol/L

— 2 heures après un repas : inférieure à 6,7 mmol/L

Les valeurs cibles en cas de diabète de grossesse sont plus basses que celles des autres types de diabète[69].

2.5.3.2 Conséquences

La mère et son enfant courent des risques en cas de diabète gestationnel.

— **Pour l'enfant :** quand le poids à la naissance dépasse les quatre kilogrammes (macrosomie). Le nouveau-né présente une jaunisse, une hypoglycémie et une détresse respiratoire. Il peut développer un diabète de type 2 qui peut l'accompagner jusqu'à l'âge adulte. Dans les cas sévères, il y a le risque de la mort du fœtus in utero.

— **Pour la mère :** elle est exposée à des avortements spontanés, des infections urinaires, une hypertension artérielle avec des œdèmes. Elle est exposée aussi à un accouchement prématuré. Quelques semaines après l'accouchement, dans 90% des cas, le diabète gestationnel disparaît. Mais le risque de diabète (type 1 ou type 2) sur plusieurs mois voire plusieurs années, existe [19].

2.6 Prévention du diabète de type 2

Afin de prévenir le diabète de type 2, il existe plusieurs mesures préventives de base importantes qui suivent afin de parvenir à un mode de vie sain et de diminuer votre risque de développer cette maladie.

— Adoptez un régime alimentaire sain et équilibré.

— Combinez un bon équilibre d'activité physique.

— Réduisez votre stress

— Surveiller votre poids .

— Évitez le tabac.

— Faire le point sur ses facteurs de risque c'est le cas si :

  — Âge supérieur à 35 ans

  — Antécédent d'accouchement d'un gros bébé

  — Syndrome des ovaires polykystiques

  — Antécédents familiaux de premier degré de diabète de type 2

2.7 Machine Learning

Même s'il est actuellement dopé par les nouvelles technologies et de nouveaux usages,le Machine Learning n'est pas un domaine d'étude qui datent d'hier puisque certains ont été conçus dès 1950, le plus connu d'entre eux étant le Perceptron qui a été inventé par le psychologue américain Frank Rosenblatt.

Le machine learning est une technique de programmation informatique qui utilise des probabilités statistiques pour donner aux ordinateurs la capacité d'apprendre par eux-mêmes sans programmation explicite. L' objectif de base du machine learning est d'apprendre à apprendre aux ordinateurs – et par la suite, à agir et réagir – comme le font les humains, en améliorant leur mode d'apprentissage et leurs connaissances de façon autonome sur la durée. L'objectif ultime serait que les ordinateurs agissent et réagissent sans être explicitement programmés pour ces actions et réactions. Le machine learning utilise des programmes de développement qui s'ajustent chaque fois qu'ils sont exposés à différents types de données en entrée [60].

Le Machine Learning est le terme le plus utilisé pour désigner l'intelligence artificielle pourtant, ces deux notions ne sont pas équivalentes mais imbriquées. La figure suivante illustre les différents sous-ensembles qui constituent l'IA.

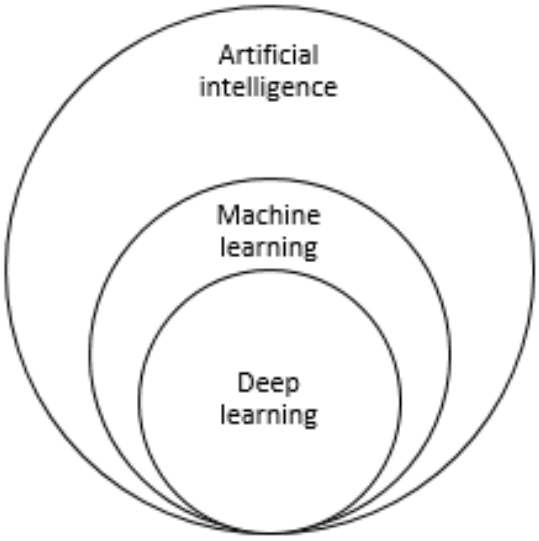

FIGURE 2.8 – Les sous ensembles de L'IA[5]

Le machine learning peut être classé en trois grandes catégories et sont détaillés en dessous :

2.7.1 Apprentissage supervisé

Le machine learning supervisé est une technologie élémentaire mais stricte. Les opérateurs présentent à l'ordinateur des exemples d'entrées et les sorties souhaitées, et l'ordinateur recherche des solutions pour obtenir ces sorties en fonction de ces entrées. Le but est que l'ordinateur apprenne la règle générale qui mappe les entrées et les sorties. Le machine learning supervisé peut être utilisé pour faire des prédictions sur des données indisponibles ou futures (on parle alors de "modélisation prédictive"). L'algorithme essaie de développer une fonction qui prédit avec précision la sortie à partir des variables d'entrée Le machine learning supervisé peut se subdiviser en deux types :

Classification : la variable de sortie est une catégorie.

Régression : la variable de sortie est une valeur spécifique. Les principaux algorithmes du machine learning supervisé sont les suivants : forêts aléatoires, arbres de décision, algorithme K-NN (k-Nearest Neighbors), régression linéaire, algorithme de Naïve Bayes, machine à vecteurs de support (SVM), régression logistique et boosting de gradient [60].

2.7.2 Apprentissage non supervisé

Dans le cadre du machine learning non supervisé, l'algorithme détermine lui-même la structure de l'entrée (aucune étiquette n'est appliquée à l'algorithme). Cette approche peut être un but en soi (qui permet de découvrir des structures enfouies dans les données) ou un moyen d'atteindre un certain but. Il existe deux types de machine learning non supervisé :

Clustering : l'objectif consiste à trouver des regroupements dans les données.

Association : l'objectif consiste à identifier les règles qui permettront de définir de grands groupes de données. Les principaux algorithmes du machine learning non supervisé sont les suivants : K-Means, clustering/regroupement hiérarchique et réduction de la dimensionnalité [60].

2.7.3 Apprentissage avec renforcement

L'apprentissage par renforcement ou Reinforcement Learning est une méthode de Machine Learning. Elle consiste à entraîner des modèles d'intelligence artificielle d'une manière bien spécifique.

L'agent IA doit apprendre à atteindre un objectif au sein d'un environnement incertain et potentiellement complexe. Pour y parvenir, l'ordinateur essaye toutes les façons possibles et apprend de ses erreurs.

À chaque tentative, l'IA reçoit une récompense ou une punition en fonction des actions effectuées.

Elle est programmée pour maximiser sa récompense, et tentera donc de trouver la méthode le lui permettant [31].

2.8 Apprentissage profond

Depuis 2006, l'apprentissage profond est apparu comme une nouvelle zone de recherche de l'apprentissage automatique. Au cours des dernières années, les techniques développées dans l'apprentissage profond ont déjà eu un impact sur les travaux de traitement des signaux et de l'information, y compris les aspects de l'apprentissage automatique et l'intelligence artificielle [23]. L'apprentissage profond (« deep learning ») est un ensemble de techniques d'apprentissage automatique qui a permis des avancées importantes en intelligence artificielle dans les dernières années. Dans l'apprentissage automatique, un programme analyse un ensemble de données afin de tirer des règles qui permettront de tirer des conclusions sur de nouvelles données. L'apprentissage profond est basé sur ce qui a été appelé, par analogie, des « réseaux de neurones artificiels », composés de milliers d'unités (les « neurones ») qui effectuent chacune de petites opérations simples. Les résultats d'une première couche de « neurones » servent d'entrée aux calculs d'une deuxième couche et ainsi de suite.

Le deep Learning est utilisé dans de nombreux domaines :

— reconnaissance d'image.

— traduction automatique.

— voiture autonome.

— diagnostic médical.

— recommandations personnalisées.

— modération automatique des réseaux sociaux.

— prédiction financière et trading automatisé.

— identification de pièces défectueuses.

— détection de malwares ou de fraudes.

— chatbots (agents conversationnels).

— exploration spatiale.

— robots intelligents.[23]

2.9 Conclusion

Dans ce chapitre nous avons présenté la maladie du diabète, ses différents types, les symptômes ainsi que le diagnostic et le traitement de chaque type et nous avons cité quelques préventions au but d'éviter la survenue de diabète, et enfin on a parler du Deep learning et le Machine Learning qui est le noyau de ce projet ainsi que ses différents types.

Dans le chapitre suivant, nous allons établi un état de l'art des principales approches relatives à la prédiction du diabète. Une étude comparative sera établie pour les principaux travaux déjà réalisés.

Chapitre 3

Etat de l'art

3.1 Introduction

Grâce à l'intelligence artificielle,la santé est devenue de plus en plus simple à traiter ce qui facilite le diagnostic aux médecins, cela encourage les chercheurs à encore développer des systèmes intelligents qui contribuent à une meilleure prise en charge des patients et à mettre en place des systèmes de prédictions du diabète.

La prédiction du diabète est le sujet le plus abordé par les chercheurs, tel qu'il existe plusieurs travaux surs, à partir de ce chapitre nous présenterons les principaux travaux relatifs basés sur le thème posé et présenter les méthodes utilisées dans chaque chapitre. Ci-dessous nous avons résumé quelques travaux prédicatifs sur le diabète. Dans ces études ils ont utilisé la précision comme un facteur de comparaison entre les algorithmes d'apprentissage utilisés.

Certaines recherches ont basé sur la prédiction du diabète afin de réduire le risque des complications de cette maladie sur la santé du patient, ou la prédisposition de la survenue de la maladie prochainement et de prévention du diabète type 2.

Pour classer une personne comme étant diabétique, il est nécessaire d'identifier les facteurs des risques qui peuvent mener à la maladie du diabète en utilisant des techniques Machine Learning. Multiples catégories ont été utilisés tel que Deep Learning ainsi que les algorithmes d'apprentissages supervisé et non supervisé.

Il existe plusieurs méthodes de prédiction du diabète, nous verrons quelques-unes dans la synthèse des travaux connexes que nous allons voir dans ce chapitre.

D'après certaines études, les résultats ont montré que les méthodes de Deep Learning comme réseaux

de neurones (CNN et KNN) et la méthode de classification Random forest (RF) ont donnés les taux de précision les plus élevés comparés à d'autres approches.

3.2 Travaux connexes

La prédiction du diabète englobe des multiples catégories concernant le traitement de cette maladie, c'est pourquoi les savants ont pensé à l'utilisation de Machine Learning pour un meilleur traitement.

L'objectif de Machine Learning est de rendre la machine capable de traiter une quantité volumineuse et inimaginable de données d'une manière rapide et d'effectuer des tâches extrêmement complexes et d'obtenir des résultats en temps réel, ce qui est difficiles à obtenir avec des algorithmes classiques.

Dans ce chapitre nous établerons des recherches de prédiction du diabète basées sur les techniques de Machine Learning.

Maniruzzaman et al [47] ont proposé l'approche : Régression logistique, cette approche est utilisé pour identifier les facteurs de risque de la maladie du diabète en fonction de p valeur et rapport de cotes (OR), ils ont adopté quatre classificateurs comme naïve Bayes (NB), arbre de décision (DT), Adaboost (AB),et forêt aléatoire (RF) pour prédire les patients diabétiques. Trois types de protocoles de partition (K2, K5 et K10) ont également adopté et répété ces protocoles dans 20 essais. Les performances de ces classificateurs sont évaluées à l'aide de la précision(ACC) et l'aire sous la courbe (AUC).Ces chercheurs ont utilisés le dataset diabetes réalisé en 2009-2012, délivré par l'Enquête nationale sur l'examen de la santé et de la nutrition. Le modèle LR démontre que 7 facteurs sur 14 comme l'âge, l'éducation, l'IMC, la TA systolique, la TA diastolique, le cholestérol direct(LDL) et le cholestérol total(HDL) sont les facteurs de risque du diabète. L'ACC global du système basé sur le ML est de 90,62%. La combinaison de la fonction basée sur LR, la sélection et le classificateur basé sur RF donnent 94,25% ACC et 0,95 AUC pour le protocole K10.

VijiyaKumar et al [67] ont proposé un cadre de travail pour objectif de développer un système qui peut effectuer une prédiction précoce du diabète pour un patient avec une plus grande précision en utilisant l'algorithme Random Forest dans la technique d'apprentissage automatique. L'algorithme Random Forest est souvent utilisés pour chaque classification et des tâches de régression. Le niveau de précision est supérieur par rapport aux autres algorithmes. Le modèle proposé donne les meilleurs résultats pour la prédiction du diabète et le résultat a montré que le système de prédiction est capable de prédire la maladie du diabète de manière efficace, efficiente et surtout, instantanément. Random Forest

Algorithm est developpé par Leo Bremen peut être un groupe d'arbres de classification ou de régression non abstraites constitués du choix aléatoire des échantillons de connaissances de coaching. Cette règle a pour habitude de réaliser la prédiction de maladie polygénique chez un patient. la précision obtenue pour Random Forest est supérieure à 90% (environ). Ceci est encore plus important que d'autres algorithmes d'apprentissage automatique pour la prédiction du diabète.

Woldemichael et al [18] (2018) ont suggéré les technique de data mining pour la prédiction de diabètes, en effet ils ont mis en place l'algorithme de backpropagation pour prédire si un patient est diabétique. De plus J48, NB et SVM qui ont été utilisés pour prédire le diabète. L'architecture de ce réseau de neurones est constitué comme suit : une couche d'entrée avec huit neurones, une couche cachée composée de six neurones et une couche de sortie. Ils ont amélioré ce modèle grace aux les schémas découverts avec ML car ces schémas ont fourni des résultats efficaces pour extraire les informations. À cette fin, des modèles de prédiction ont été développés à partir d'ensembles de données médicales diagnostiqués recueillies auprès de personnes atteintes de diabète.

Le SVM, NB, KNN et C4.5 DT étaient quatre algorithmes bien connus pour prévoir le DM sur les données de la population adulte. Les résultats des expériences ont révélé que l'algorithme C4.5 DT offrait une plus grande précision par rapport aux autres algorithmes.

Benbelouaer [22] a réaliser un système de prédiction et de prévention du diabète type 2 en utilisant les réseaux de neurones artificiels (la prédiction multicouche), Un réseau de neurones artificiels (RNA), (ou Artificiel Neural Network en anglais), est un système informatique matériel et / ou logiciel dont le fonctionnement est calqué sur celui des neurones du cerveau humain. Il s'agit là d'une variété de technologie Deep Learning (apprentissage profond), qui fait elle-même partie de la sous-catégorie d'intelligence artificielle du Machine Learning (apprentissage automatique). Lors de réalisation de ce système de prédiction ils ont utilisé le data set Pima Indian database, la précision obtenue pour l'application des réseaux de neurones artificiels MLP est de 83.33%.

Daanouni et al [10] ont appliqué et évalué quatre algorithmes de Machine Learning (Decision Tree, K-Nearest Neighbours, Artificial Neural Network et Deep Neural Network) pour prédire le diabète sucré.

Ces techniques ont été formées et testées sur la base de données Pima Indian. Les performances des algorithmes expérimentés ont été évalués après la suppression des données bruitées et utilisation des fonctionnalités de sélection avec les Composantes du voisinage d'analyse afin de réduire le nombre de fonctionnalités et atténuer la complexité de dimensionnalité en faveur d'accélérer le processus d'appren-

tissage et améliorer la compréhension des données. Différentes mesures de similarité ont été utilisées pour comparer les performances du modèle telles que la précision, la sensibilité et la spécificité. Les précisions obtenues pour ces 4 techniques sont : KNN : 87.36%, DT : 79.95%, ANN : 87.09%, DNN : 92.3%.

Nareshkumar et al [45] ont maintenu à trouver une solution pour la prédiction du diabète en proposant un système de recommandation de soins de santé (HRS) qui peut être conçu pour prédire l'état de santé en évaluant le mode de vie du patient, sa santé physique et sont aspects de santé mentale à l'aide de modèle d'apprentissage en profondeur proposé avec réseaux de neurones convolutifs (D-CNN). L'application de cette approche atteint une précision globale de 96,25 %. D-CNN s'avère plus efficace pour la prédiction du diabète que les autres approches d'apprentissage automatique (ML) dans l'analyse expérimentale.

Parisa et al [62] ont utilisé différents algorithmes de machine learning (Logistic Regression, Nearest Neighbor, Decision Tree, Random Forest, Support Vector Machine, Naive Bayesian, Neural Network and Gradient Boosting) , Le modèle basé sur l'algorithme de gradient boosting a montré une meilleur performance avec une précision de prédiction de 95,50 %. En utilisant un dataset clinique populaire obtenu a partir de 'Deryad digital' qui est un site web permettant de rendre les données brutes des articles publiés librement réutilisables pour l'analyse secondaire.

Talha et al [61] ont mise en œuvre les méthodes : Réseau de neurones artificiels (ANN), forêt aléatoire (RF) et des techniques de clustering K-means pour prédire par le diagnostic si un patient est atteint de diabète, via la méthode d'analyse en composantes principales. Les résultats indiquent une forte association du diabète avec l'indice de masse corporelle (IMC) et avec le niveau de glucose, le dataset utilisé pour cet étude est le Pima Indian Diabetes qui provient à l'origine de l'Institut national du diabète et des maladies digestives et rénale, la précision obtenue pour l'application des réseaux de neurones artificiels(ANN) est de 75.7%, contre 74.7% pour forêt aléatoire (RF).

Sidahmad et al [58] ont utilisé des algorithmes d'apprentissage automatique supervisé (K nearest neighbors, Decision Trees, Random Forest, Support Vector Machine, Naïves Bayes) dans le but de développer une application web pour la Prédiction précoce du diabète de type 2, afin de réduire le risque des complications de cette maladie sur la santé du patient en utilisant le dataset extrait de l'hôpital Frankfurt (Allemagne). Les performances des classifieurs ont été comparées en fonction du taux de précision et de la sensibilité du modèle. Les plus hauts taux de classifications obtenus par l'application de Random Forest et l'arbre de décision sont respectivement 91% et 87%, en appliquant les deux méthodes d'évaluation

train/test et validation croisée.

Nasir et al [56] ont maintenu six algorithmes d'apprentissage automatique pour prédire le diabète chez les patients. Ces algorithmes sont : Nearest Neighbours (KNN), Support Vector Classfier(SVM), Logistic Regression(LR), Decision Tree Classfier(DT), Gaussian Naive Bayes(NB) et Random Forest (RF). Tous ces algorithmes ont été appliqués à Dataset PIMA Indian comprenant 768 enregistrements et 9 attributs.

Dans ce travail, le principal paramètre d'évaluation entre les algorithmes est la précision de prédiction du diabète. Cette étude a conclu que SVM et KNN sont appropriés pour prédire l'état du diabète des patients, ces deux derniers ont atteint la précision la plus élevée qui est de 77%.

3.3 Analyse et comparaison

Dans les tableaux ci-dessous nous effectuerons une étude comparatives des approches proposées ci-dessus selon les 8 facteurs suivants :

- **Approche :** désigne l'approche de chaque papier.
- **Dataset :** indique les sources de données utilisée pour l'implémentation de l'approche pour la prédiction du diabètes.
- **Résultats :** les résultats de l'approche.
- **Techniques utilisées :** les technique utilisées pour prédire le diabète.
- **Avantages :** avantages de l'approche abordée.
- **Inconvénients :** inconvénients de l'approche abordée.

Approche	Dataset	Résultats(Accuracy)	Techniques utilisées	Avantages	Inconvénients
Md.Maniruzzaman, Md.Jahanur Rahman, Benojir Ahammed, Md.Menhazul Abedin (2020)	DataSet Diabetes	NB + K2 = 86.42% NB + K5 =86.61% NB + K10 =86.70% DT + K2 =89.90% DT + K5 =89.97% DT + K10 =89.65% AB + k2 =91.32% AB + K5 =92.72% AB + K10 =92.93% RF + k2 =93.12% RF + K5 =94.15% RF + K10 =94.20%.	-LR	-LR facile à Comprendre -Nécessite pas beaucoup de donnée pour un bon fonctionnement.	-Si le nombre d'observations est inférieur au nombre d'entités, la régression logistique ne doit pas être utilisée, sinon elle peut conduire à un sur-ajustement. -La principale limitation de la régression logistique est l'hypothèse de linéarité entre la variable dépendante et les variables indépendantes. [33]
K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Caroline,S (2019)	UCI machine learning reposit	Accuracy is more than 90%	-Random Forest (RF)	-Il permet d'obtenir une prédiction fiables grâce a son système d'arbres décisionnels. - Une gestion efficace de grands ensembles de données	-Difficilement interprétable, difficilement améliorable -Entrainement plus lent[48]
Fikirte Girma Woldemichael, Sumitra Menaria (2018)	PIMA indian dataset	BP = 83.11% jv48=78.26% naïve bayes=78.97% SVM=81.69%	-Backpropagation -J48 -NB classifier -SVM Classifier	BP algorithme utilisé pour calculer rapidement les dérivées et Il permet de pallier une déficience de l'algorithme perceptron qui est incapable de modifier les poids des couches cachées[3]	L'algorithme de rétropropagation est gourmand en temps de calculs.[3]
Benbelouaer ghada (2021)	Pima Indian database	BP=83,33 %	-MLP(multilayer Perceptron).	-Incrémentalité - Scalabilité (capacité à être mis en œuvre sur de grandes bases)	-Très difficile d'analyser et comprendre le fonctionnement. -Difficulté de choisir la structure.[15]
Sidahmed Amel, Rabhi Karima , (2020)	le datasets extrait du l'hôpital Frankfurt (Allemagne).	KNN = 82.5% SVM = 79.5% DT = 87.5% NB = 78.5% RF = 91%	-KNN (K-Nearest Neighbors) -L'arbre de decision (Decision tree) -SVM (support vector machin) -Random Forest (for^et aleatoire) -Naïve Bayes	-Entraînement facile et efficace des différents modèles grâce à des données déjà étiquetée.	-Très difficile d'analyser et comprendre le fonctionnement. -Difficulté de choisir la structure.[15]
Othmane Daanouni, Bouchaib Cherradi, Amal Tmiri, (2020)	Pima Indian dataset	KNN = 90% DT = 82.5% ANN = 87.5% DNN = 90%	-Decision Tree -K-Nearest Neighbours -Deep neural Learning(DNN) -Artificial Neural Network	-L'algorithme DNN donne la précision la plus élevé que d'autres algorithmes plus de 90% -Exécution efficace	-DNN nécessite une grande puissance de calcul -DNN est un Une technologie coûteuse à mettre en place -Il nécessite une vaste base de données[54]
Nareshkumar Mustary, Phani Kumar Singamsetty (2022)	3 critères sont : mode de vie, la santé physique et les aspects de santé mentale du patient	DCNN =96,25 %	- DCNN	-Réduire l'empreinte mémoire -Améliore les performances -L'avantage majeur des réseaux convolutifs est l'utilisation d'un poids unique associé aux signaux entrant dans tous les neurones d'un même noyau de convolution. [2]	Le comportement de CNN est floue. -Il a besoin des données volumineuses pour l'entraînement efficace.
Parisa Karimi Darabi, Mohammad Jafar Tarokh, (2020)	dataset clinique populaire obtenu a partir de 'Deryad digital'	LR = 0.953% DT = 0.953% RF = 0.953% SVM = 0.954% NB = 0.936% GBM = 0.955% KNN = 0.948% NN = 0.944%	-Logistic Regression -Nearest Neighbor -Decision Tree -Random Forest -Support Vector Machine(SVM) -Naive Bayesian -Neural Network and Gradient Boosting	Le modèle basé sur l'algorithme de gradient boosting a montré une meilleure performances qui est une technique en intelligence artificielle qui consiste à assembler un grand nombre d'algorithmes avec de faibles performances individuelles pour en créer un beaucoup plus efficace	-Modèle non explicite -Paramètres nombreux (taille de l'arbre, nombre d'itérations, paramètre de régularisation, etc.) -Danger du sur-apprentissage. [51]
Talha Mahboob Alama, Muhammad Atif Iqbala, Yasir Alia, Abdul Wahabb, Safdar Ijazb, Talha Imtiaz Baigb, Ayaz Hussainc, Muhammad Awais Malikb, Muhammad Mehdi Razab, Salman Ibrarb, Zunish Abbasd, (2019)	l'Institut national du diabète et des maladies digestives et rénales UCI ML	- RF = 74.7% -ANN = 75.7% -K-means = 73.6%	-Artificial neural network (ANN) -Random Forest (RF) -K- means clustering -L'objectif d'ANN est de convertir les données d'entrée en sortie signicative -RF permet d'obtenir une prédiction fiable, grâce à son système d'arbres décisionnel	-Méthode de random forest est un algorithme d'apprentissage automatique flexible, rapide et simple -ANN détecte des modèles complexes et apprend sur la base de ces modèles il est rapide en calcule	- RF : La difficulté de choix de l'architecture et des paramètres -ANN : admet Le problème d'initialisation et de codage. [51]
Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali Shah, (2018)	l'Institut national du diabète et des maladies digestives et rénales UCI ML	LR = 74% SVM = 77% NB = 74% DT , RF = 71% KNN = 77%	-Nearest Neighbours (KNN) -Support Vector Classier(SVM) -Logistic Regression(LR) -Decision Tree Classier(DT) -Gaussian Naive Bayes(NB) -Random Forest(RF)	-KNN : algorithme est simple et facile à mettre en œuvre. -SVM : Capacité à traiter des grandes dimensionnalités -NB :il n'a pas besoin de grand données d'entraînement pour estimer les paramètres nécessaires à la classification, -LR : facile à mettre en oeuvre, et très efficace -DT : facile à comprendre , à interpréter et rapide -RF : apprentissage automatique flexible, rapide et simple.	-L'algorithme KNN devient beaucoup plus lent à mesure quand le nombre d'exemples d'apprentissage augmente.[24] -Temps de calcul est grand quand K augmente.[52]

TABLE 3.1 – Étude comparative des travaux connexes

Les travaux étudiés tentent d'analyser les données et d'obtenir de meilleurs résultats en termes de précision, de rappel et d'exactitude, plusieurs approches ont été utilisées comme le Machine Learning et deep learning en tenant compte de plusieurs facteurs, dans ce cas la glycémie, l'age, l'indice de masse corporelle(BMI) et blood pressure ect...

Dans cette étude, des divers algorithmes de classification comme le retour propagation, svm, j48 et l'algorithme naïf bayes ont été abordés pour prédire la maladie du diabète. Cette étude aide à soutenir la médecine décision et aide à améliorer le traitement médical et le pronostic des patients.

Les avantages majeur de cette machine learning pour la prédiction du diabète sont : facile à utiliser et à comprendre, nécessite pas beaucoup de donnée pour un bon fonctionnement et sont entraînement est facile et efficace des différents modèles grâce à des données déjà étiquetée. Par contre, ces approches présentent plusieurs inconvénients comme le modèle est très difficile à analyser et comprendre son fonctionnement et la difficulté de choisir la structure.

Parmi les techniques utilisées le deep neural network a monté la précision la plus élevé parmi, le majeur avantage de cette méthode est l'exécution efficace, et l'amélioration des performances.par contre elle est un Une technologie coûteuse à mettre en place et nécessite une grande puissance de calcul.

Les recherches étudiées ont proposé des systèmes de recommandations et de prédictions, dans notre étude nous avons proposé un système qui fait prédire le diabètes, notre approche est inspirées de ces travaux connexes tel que la méthode de deep learning que nous avons utilisé.

3.4 Conclusion

Dans ce chapitre, nous avons établi un état de l'art sur la prédiction du diabète qui représente une étude comparative de tous les travaux connexes que nous avons abrégé, nous avons présenté ceci dans un tableau détaillé décrivant chaque approche des documents résumés, tout en suivant chaque travail par un bref paragraphe qui le résume. Dans le chapitre suivant, nous allons présenter notre approche et ses différentes étapes.

Chapitre 4

Approche proposée

4.1 Introduction

Dans nos jours, les systèmes de prédiction de diabètes et l'extraction de l'information utile d'une façon automatique sont très importantes et aident les médecins de diagnostiquer efficacement les patients diabétiques et de construire les bases de données des patients.

Le nombre de personnes souffrant de maladies diabètes est en augmentation. Un diagnostic précis à un stade précoce suivi d'un traitement ultérieur approprié peut permettre de réduire les risques des complications sur la santé du patient issues de la maladie de diabète ou de la prévenir.

Dans ce chapitre, nous présenterons en détail notre approche qu'on a utilisé au cours de notre projet ainsi que ses différentes étapes pour la prédiction du diabète à l'aide de détails médicaux.

4.2 Contribution

Notre projet consiste à une prédiction précoce du diabète de type 2, c'est-à-dire de permettre aux personnes de savoir s'ils ont le risque de développer un diabète avec un taux de prédiction bien défini. Pour atteindre cet objectif, il faut suivre plusieurs étapes qui doivent être effectuées pour obtenir de meilleurs résultats, ces étapes sont les suivantes : Collecte des données, prétraitement, Sélection de modèle, entraînement des données, évaluations des modèles.

La figure 4.1 donne un aperçu de l'approche proposée et les différentes étapes qui la composent :

importer les données depuis Kaggle sous forme d'un Dataset au format csv, prétraitement des donnée d'entrés(Exploration et visualisation de données, Nettoyage de données...), ensuite la sélection des

caractéristiques, division de donné en deux :testset et trainset, sélection de model(DNN,SVM,Random forest), entrainement, prédiction et calcul de l'accuracy, évaluation de model.

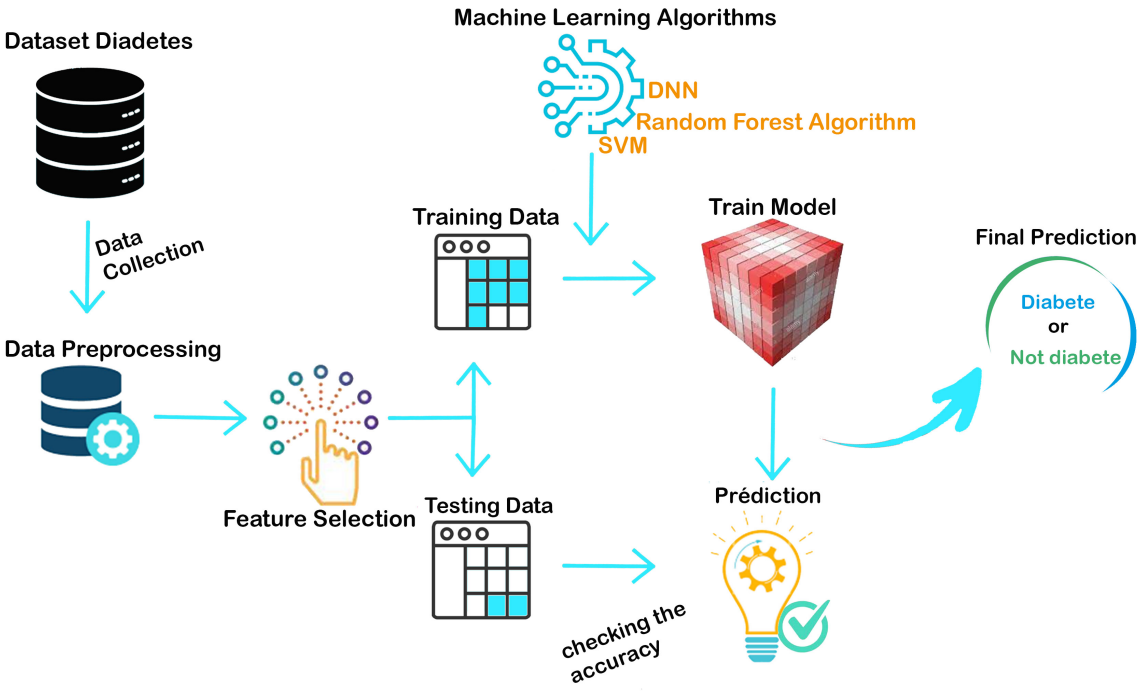

FIGURE 4.1 – Schéma global de l'approche

Nous détaillons ci-dessous chaque étape comme suit :

La première étape est consacrée à la collecte des données qui consiste à extraire activement l'information de la source.

La deuxième étape consiste au pré-traitement des données collectées,cette étape est composée de plusieurs sous étapes on cite : nettoyage des donnés aberrantes, analyse et visualisation des données..

La 3ème étape consiste à entraîner et tester les données, c'est une étape très importante pour parvenir un bon résultat de la prédiction.

La dernière étape est consacré à choisir les méthodes de modèles suivis d'une prédiction finale.

4.2.1 Collecte des données

La collecte des données est une étape initiale très importante pour la prédiction de diabète. Cela permet un bon traitement et évaluation de la proche choisie, le dataset diabetes est un ensemble de données,

extrait de l'hôpital de Frankfort, Allemagne, il se composent de plusieurs variables prédictives médicales.

L'objectif le plus important de la collecte de données est de s'assurer que les données complète et fiable en informations qui permet de mieux évaluer des résultats et de mieux anticiper les probabilités et les tendances à venir[63].

4.2.2 Prétraitement

Après la collecte des données, l'étape suivante est le prétraitement, cette dernière est très importante pour extraire un dataset parfait au but d'obtenir des résultats de qualité. La plupart des dataset peut y avoir des valeurs manquantes et/ou bruyantes et données incohérentes. Si la qualité des données est faible, aucun résultat de qualité ne peut être trouvé.

Il y a plusieurs étapes dans ce prétraitement, dont le nettoyage, Nettoyage, intégration, transformation, exploration et visualisation de données, sélection des fonctionnalités, sélection et évaluation de modèle.

Voici une description détaillée des étapes de prétraitement ci-dessus :

- **Exploration et visualisation de données :**

  La visualisation des données est définie comme l'exploration visuelle des données, qui aide à obtenir et connaîtrez des informations et caractéristiques approfondies et clairs sur l'ensemble de données et les variables. On remarque que :

— Le nombre des observations (2000 patients) dont 1316 diabétiques et 684 non diabétiques.

— les technique utilisées pour prédire le diabète.

— Les valeurs manquantes.

— La taille de dataset...

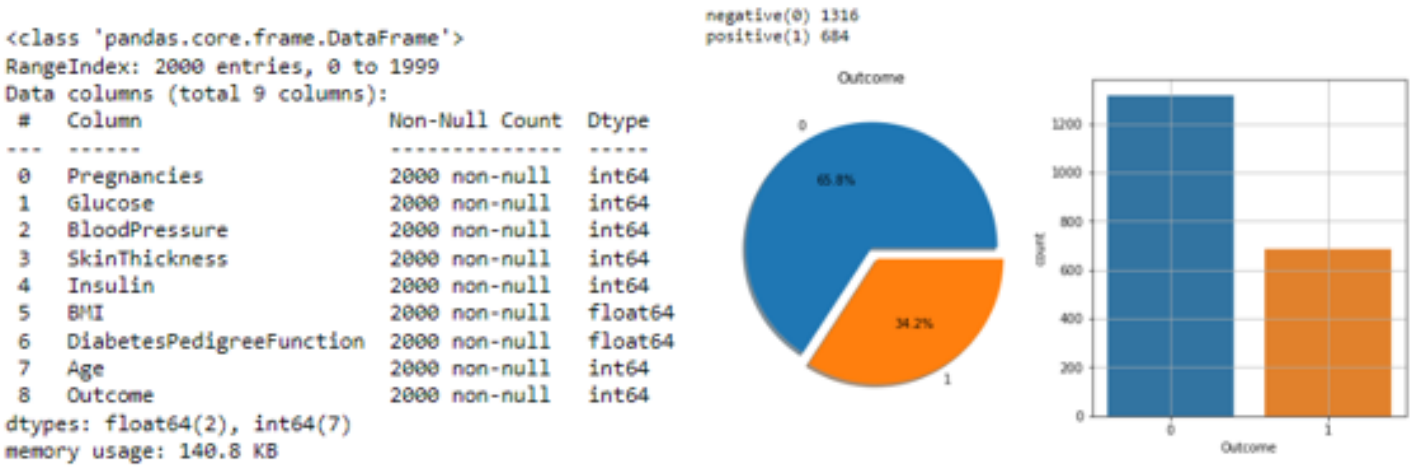

FIGURE 4.2 – Informations de la base de données

- **Nettoyage des données :**

  Le nettoyage des données consiste à remplir les valeurs manquantes et à supprimer les données bruyantes. Les données bruyantes contiennent des valeurs aberrantes qui serons supprimées pour résoudre l'incohérences [1].

  Dans notre jeu de données, Glucose,BloodPressure,SkinThickness,Insulin,BMI ont des valeurs nulles (0). Ainsi, tout le zéro les valeurs ont été remplacées, Cette étape permet de réduire le risque dans l'ensemble de données.

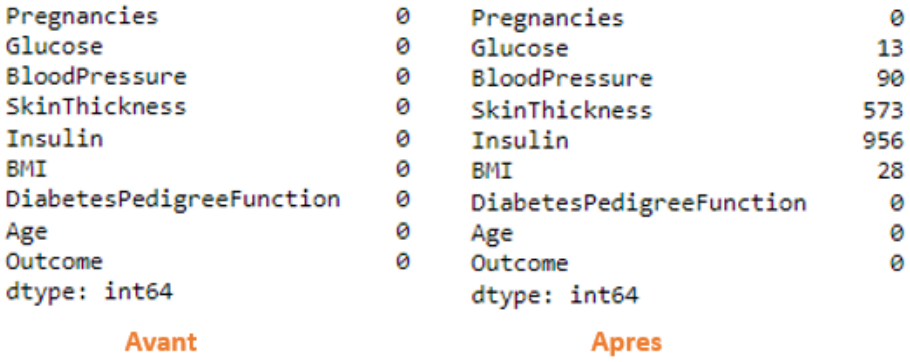

FIGURE 4.3 – Après et avant le remplacement des valeurs zéro

- **Importance des caractéristiques**

  La figure 4.4, montre l'importance des variables. On peut voir que la concentration de glucose et l'IMC jouent un rôle clé dans la détection du diabète a la plus haute importance parmi d'autres variables. L'indice de masse corporelle et l'âge sont respectivement le deuxième et troisième variables importantes, Cela signifie que la précision du modèle dépend principalement du glucose et de l'IMC, On peut en déduire que ces variables importantes jouent un rôle important dans la prédiction et indiquent si le patient est diabétique.

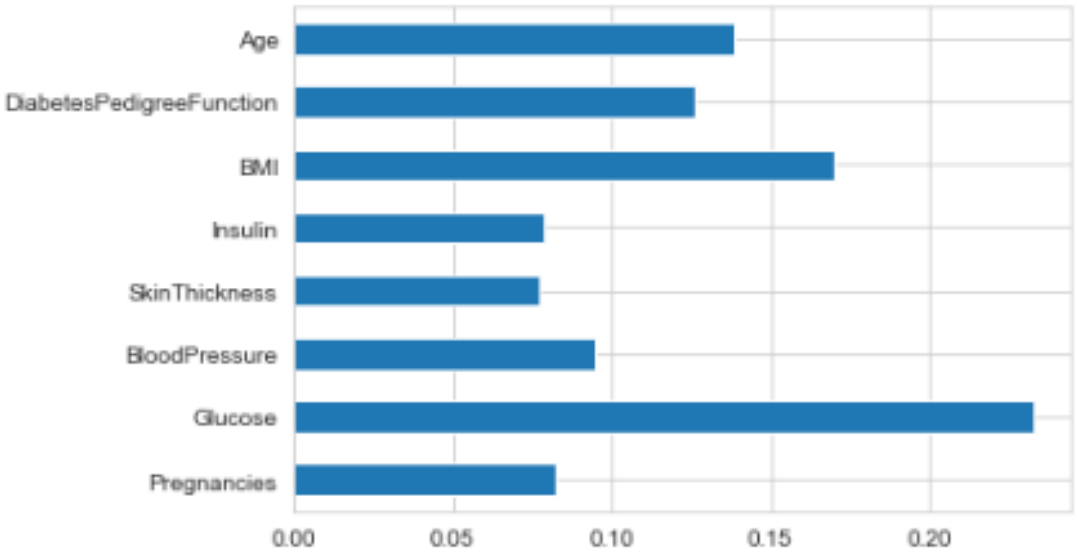

FIGURE 4.4 – Feature importance

- **Corrélation**

  Un bon ensemble de données est un ensemble dans lequel les caractéristiques sont fortement corrélées à la classe cible et sont fortement non corrélées les unes aux autres. Pour trouver les attributs non corrélés.

  Coefficient de corrélation : est un nombre qui indique la force de la relation entre deux variables. Il existe plusieurs types de coefficients de corrélation, mais le plus commun de tous est le coefficient de Pearson noté, défini par : r = Cov(X, Y) /XY[16].

  La figure 4.5 montre que l'Age, Glucose et BMI sont des caractéristiques importantes pour le diagnostic du la maladie de diabète et qui sont fortement non corrèles les uns aux autres.

  D'après la visualisation des données on constate qu'il n'y aucun point de données manquant ou nulle dans l'ensemble de données

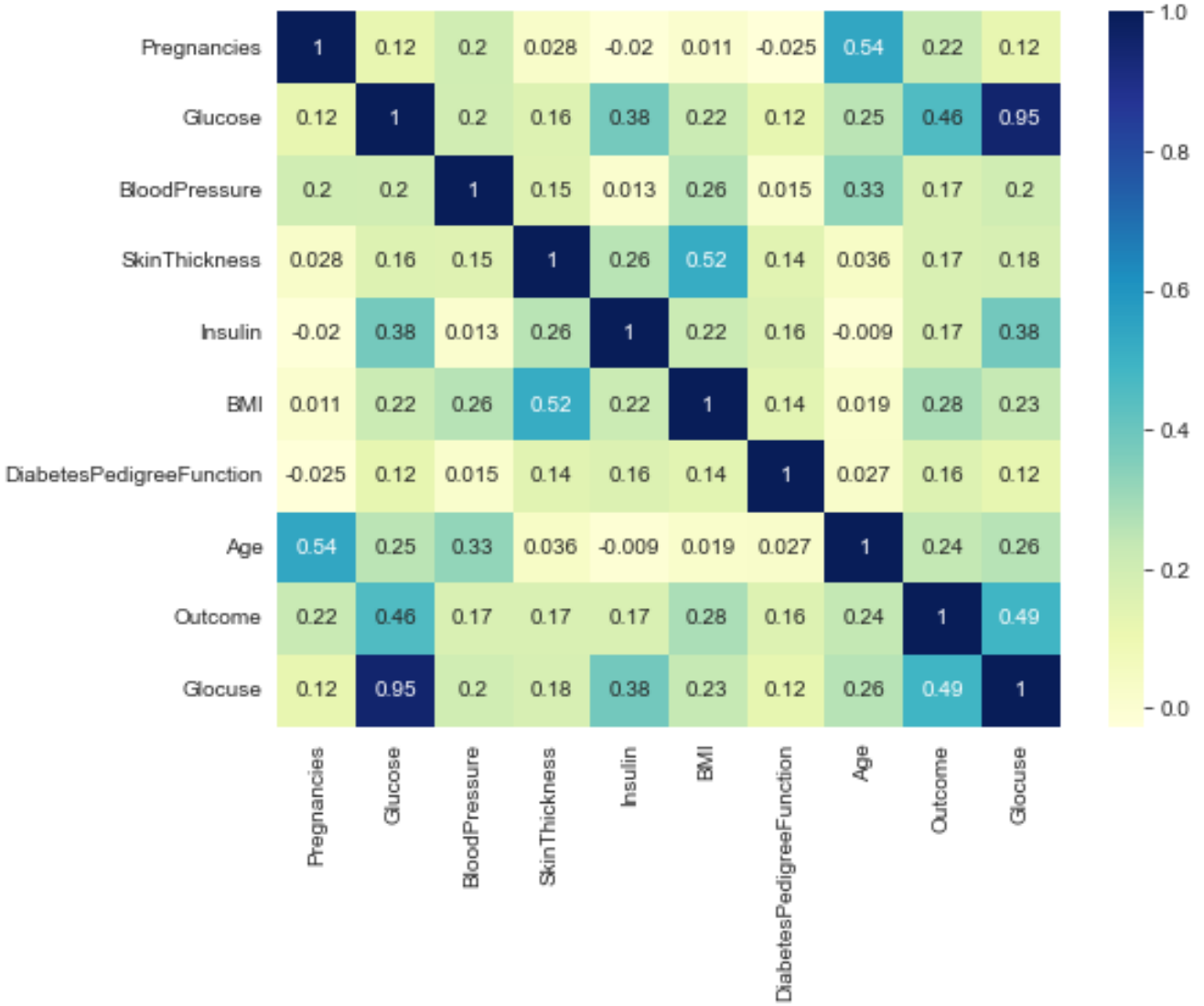

FIGURE 4.5 – Matrice de corrélation

- **Transformation des données**

  La transformation des données consiste en un lissage, une normalisation et une agrégation des données[9], La normalisation des données est une méthode de prétraitement des données qui permet de réduire et simplifier la complexité des modèles, La normalisation des données fait référence au décalage des valeurs des données afin qu'elles soient entre 0 et 1.

- **Proportion des variables**

  Dans la figure 4.6 représente la proportion des variables par rapport aux diabétiques et les non diabétiques.

  On remarque un décalage entre les variables pour les deux types, les valeurs sont normaux, après le prétraitement et la suppression des zéros.

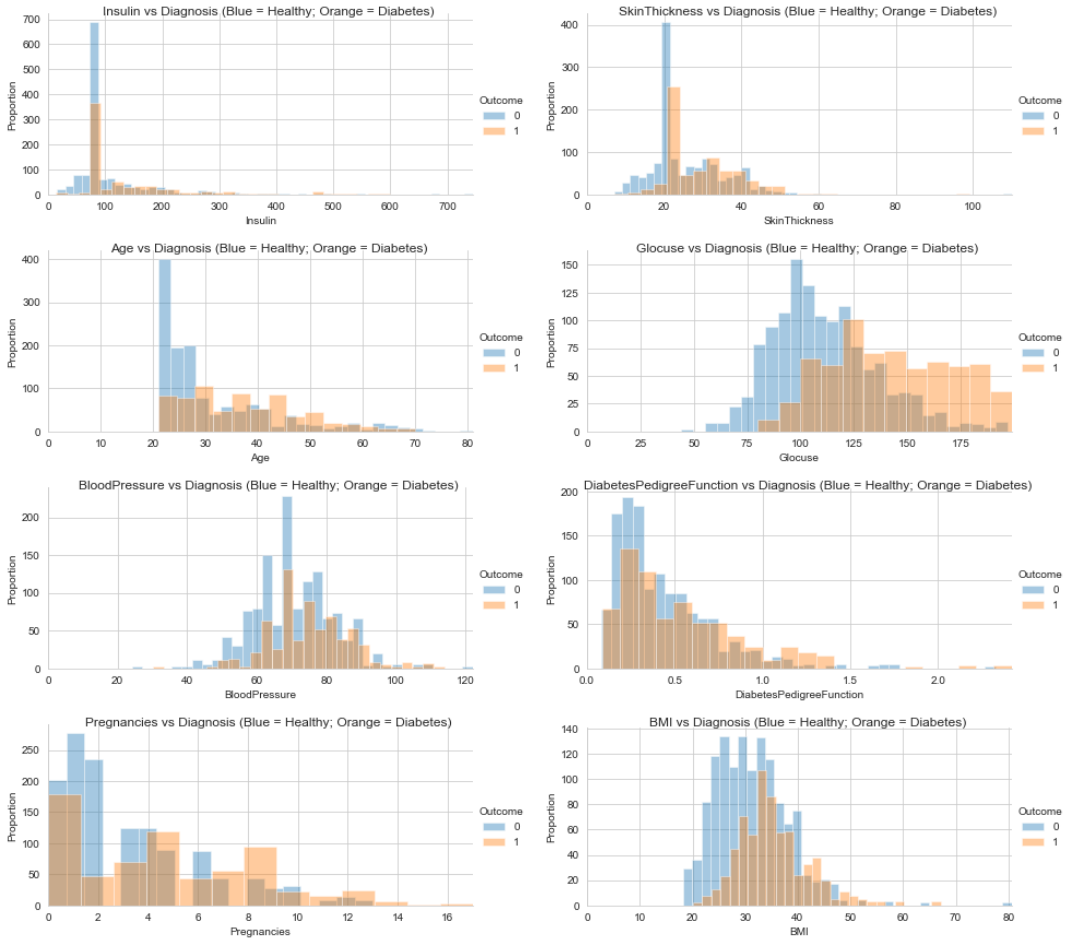

FIGURE 4.6 – Proportion des variables pour les personnes diabétiques et non diabétique

4.2.3 Entraînement et Test des données

Il est important de former et de tester un modèle pour parvenir à un bon résultat de la prédiction, 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 test le modèle et évaluer les performances des classifieurs sélectionnés. Si un modèle fonctionne mieux dans les deux ensembles de données, alors la précision attendue est meilleure.

FIGURE 4.7 – Répartition des données de train/test

‘

4.2.4 Sélection de modèle

Après la division de données, l'étape suivante est la sélection de modèle pour la prédiction du diabète. Cette étape est celle qui permet de mettre en œuvre divers algorithmes d'exploration de données. Trois modèles ont été utilisés pour la prédiction précoce du diabète sont les suivants :

4.2.4.1 Deep neural network

DNN sont devenus une solution prometteuse pour injecter de l'IA dans notre vie quotidienne à partir de voitures autonomes, de smartphones, de jeux, de drones, etc.

Les réseaux de neurones profonds (DNN) sont des versions améliorées de l'ANN conventionnel avec plusieurs couches. Les modèles DNN sont récemment devenus très populaires en raison de leurs excellentes performances pour apprendre non seulement le mappage entrée-sortie non linéaire, mais également la structure sous-jacente des vecteurs de données d'entrée[6].

Le réseau neuronal profond est un modèle très utilisé compte tenu ses résultats et la capacité de réaliser des tâches complexes, et le traitement efficace des données très volumineuse.

Une architecture simple de réseau neuronal profond est illustrée à la Fig. 4.8 Dans chaque couche de nœuds, la sortie dépend de la sortie de la couche précédente. Dans un réseau de neurones, les neurones de la couche de sortie n'ont le plus souvent pas de fonction d'activation.

Le réseau neuronal profond se caractérise de fonction d'activation sont nécessaires pour implémenter

des fonctionnalités de mappage complexes qui ne sont pas linéaires afin d'apporter la propriété de non-linéarité très nécessaire qui leur permet d'approximer n'importe quelle fonction.

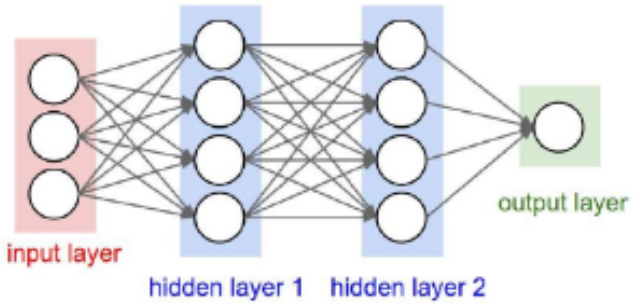

FIGURE 4.8 – Architecture Deep Neural Network [27]

4.2.4.2 SVM classifier

Support Vector Machine ou SVM est l'un des algorithmes d'apprentissage supervisé les plus populaires, utilisé pour les problèmes de classification et de régression. Cependant, il est principalement utilisé pour les problèmes de classification dans l'apprentissage automatique.

Le but de l'algorithme SVM est de créer la meilleure ligne ou limite de décision qui peut séparer l'espace à n dimensions en classes afin que nous puissions facilement mettre le nouveau point de données dans la bonne classe à l'avenir. Cette meilleure frontière de décision est appelée un hyperplan.

SVM choisit les points / vecteurs extrêmes qui aident à créer l'hyperplan. Ces cas extrêmes sont appelés vecteurs de support, et donc l'algorithme est appelé machine de vecteur de support. [21]

SVM est un classifier est une représentation des exemples sous forme de points, il peut efficacement classer les données non linéaires à l'aide de l'astuce du noyau.

4.2.4.3 Random forest classifier

RF est un classificateur basé sur ML en construisant des arbres de décision. La forêt aléatoire est un algorithme d'apprentissage automatique supervisé largement utilisé dans les problèmes de classification et de régression.

RF peut être utilisé dans plusieurs biomédecine la recherche[57],en particulier le diagnostic du diabète.

Les étapes de RF comme suit :

- Étape 1 : Dans la forêt aléatoire, un nombre n d'enregistrements aléatoires est extrait de l'ensemble de données contenant 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[38].

Le majeur avantage de ce classifier est qu'Il résout le problème de overfitting car la sortie est basée sur le vote à la majorité ou la moyenne.

4.3 Conclusion

Dans ce chapitre, nous avons présenté en détail notre approche de prédiction du diabète, en utilisant les 2 classifier de ML est deep neural network.

Notre approche est inspirée des recherches relatifs à la prédiction du diabète qui permet d'utiliser plusieurs métriques pour la prédiction.

Dans le chapitre suivant, nous procéderons à l'explication de tous les aspects liés à l'implémentation de notre approche.

Chapitre 5

Expérimentation et évaluation

5.1 Introduction

Dans le cadre d notre travail, nous avons traité notre jeu de donnée pour une meilleur prédiction. Ce traitement permet de corriger les valeurs aberrantes et visualisation des données de dataset pour prédire la survenue de la maladie du diabètes chez les patients en analysant 8 facteurs de risque comme le glucose, BMI, pregnancies.

Dans ce chapitre, nous introduirons les différents aspects liés à l'implémentation de notre approche que nous avons développée, puis nous définirons notre dataset avec une description de ses caractéristiques, ainsi que les performances de modèles choisi et nous terminerons par une conclusion.

5.2 Description du dataset

Dataset : est un ensemble de données contient des informations de différents types ou de même types, structuré des lignes et des colonnes, chaque colonne est caractérisée par un type spécifique de valeurs numériques, où chaque valeur est associée à une variable et à une observation, chaque dataset est généralement caractérisé d'une variable labellisée pour classifier les ligne de données qu'il contient. Un dataset peut être d'une extension csv, tsv, qu'on peut les importer avec les fonctions de pandas dans python.

La quantité de données que peut contenir un jeu de données est défère d'un à un autre petit(quelques caractéristiques et 100 lignes) , volumineux((plus de 1 000 caractéristiques et plus d'un million de lignes),la sélection des fonctionnalités dans le jeu de données est très essentielle à la création d'un modèle

performant pour plusieurs raisons. donc une évaluation de l'importance des caractéristiques est nécessaire, Les données contiennent presque toujours plus d'informations que nécessaire pour générer le modèle ou elles contiennent un type d'informations inapproprié, Toutefois, si certaines colonnes contiennent des données éparses, cela n'est pas très utile de les ajouter au modèle, les algorithmes qui peut évaluer les caractéristiques on trouve équivalent bayésien de Dirichlet (BDE), Entropie de Shannon[43].

En machine Learning ,il existe des algorithme qui nécessitent pas nécessite pas beaucoup de données pour un bon fonctionnement, on cite : la régression logistique, et d'autre qui nécessite une vaste base de données comme Deep neural network.

Le Dataset extrait de l'hôpital de Frankfort, Allemagne téléchargé sur kaggle[30],il se composent de plusieurs variables prédictives, il est au format CSV car il est plus pratique pour Python de traiter ce type de fichiers dans le domaine. La taille du Dataset est 62,06 kB et comporte 2000 patients diabétiques et non diabétiques.

Le Dataset est composé de neuf (9) colonnes :

- **Glucose :** Concentration plasmatique de glucose à 2 heures dans un test oral de tolérance au glucose.
- **Pregnancies :** Nombre de fois enceinte.
- **BloodPressure :** Pression artérielle diastolique (mm Hg).
- **SkinThickness :** Epaisseur de pli cutané du triceps (mm).
- **Insulin :** Insuline ssérique 2 heures (mu U/ ml).
- **BMI :** (ou IMC) Indice de masse corporelle (poids en kg / (taille en m)2).
- **DiabetesPedigreeFunction :** Fonction généalogique du diabète.
- **Âge :** l'âge en années.
- **Outcome :** variable de classe (0 ou 1) où 0 indique que le patient ne souffre pas de diabète et 1 indique que le patient est diabétique[29].

Variable	Description	Analyse de données
Glucose	Une valeur de 2 heure entre (140 et 200 mg)/dl (7.8 et 11.1 mmol/L) est appelé tolérance au glucose altère signifie que il y a un risque accru de développe le diabète au fil de temps. Un taux de glucose de 200 mg/dL(11.1 mmol/L) ou plus utilisé pour diagnostiquer le diabète.	Minimum : 0 Maximum : 199
Pregnancies	Nombre de fois enceinte	Minimum : 0 Maximum : 17
BloodPressure	Si un TA diastolique supérieur à 90 signifie une pression artérielle élevé (probabilité élevé de diabète) Un TA diastolique inférieur à 60 signifie une pression artérielle base (mois probabilité de diabète)	Minimum : 0 Maximum : 122
SkinThikness	Valeur estimée pour la graisse corporelle. épissure normale du pli cutané chez les femmes est de 23 mm. Une épissure plus élevée conduit à l'obésité et les chances de diabète augmente	Minimum : 0 Maximum : 110
Insulin	poids en kg / taille en m2) IMC de 18.5 'a 20 c'est normal IMC entre 25 et 30 situer dans une plage surpoids Et de 30 ou plus situer dans la fourchette d'obésité	Minimum : 0 Maximum : 80.6
DiabetePredigme Function	Fournit des informations sur les antécédentes chez les parents et la relation génétique avec les patients. Une fonction de pedigree plus élevée signifie que le patient plus susceptible de souffrir un diabète	Minimum : 0.078 Maximum : 2.42
Age	Age d'une personne en années	Minimum : 21 Maximum : 81
Outcome	Indique si une personne est diabétique ou non	0(non diabétique) :1316 1(diabétique) : 684

TABLE 5.1 – Description des variables d'ensemble de données [59]

5.3 Environnement de développement

- **Anaconda** Anaconda est une distribution libre et open source des langages de programmation Python et R appliqué au développement d'applications dédiées à la science des données et à l'apprentissage automatique, qui vise à simplifier la gestion des paquets et de déploiement[53].
- **Jupyter notebook**

  Jupyter notebook est le dernier environnement de développement interactif basé sur le Web pour les blocs-notes, le code et les données. Son interface flexible permet aux utilisateurs de configurer et d'organiser des flux de travail en science des données, en informatique scientifique, en journalisme informatique et en apprentissage automatique. Une conception modulaire invite les extensions à étendre et enrichir les fonctionnalités[28].

5.3.1 Langage de programmation

- **Python**

Python est un langage de programmation puissant et facile à apprendre. Il a des structures de données de haut niveau efficaces et une approche simple mais efficace pour programmation orientée objet. La syntaxe élégante et le typage dynamique de Python, ainsi que sa nature interprétée, en font un langage idéal pour des scripts et développement rapide d'applications dans de nombreux domaines sur la plupart des plates-formes[65].

5.3.2 Bibliothèques de Python

- **Pandas**

  pandas est une bibliothèque Python open source pour l'analyse de données hautement spécialisée. C'est actuellement le point de référence que tous les professionnels utilisant le langage Python doivent étudier à des fins statistiques d'analyse et de prise de décision.

  Cette bibliothèque a été conçue et développée principalement par Wes McKinney à partir de 2008. En 2012, Sien Chang, l'un de ses collègues, a été ajouté au développement. Ensemble, ils ont mis en place l'une des bibliothèques les plus utilisées de la communauté Python[46].

- **Tensorflow**

tensorflow est une bibliothèque open source créé à l'origine pour les tâches de calculs numériques lourds (Learning TensorFlow [Auteurs : Tom Hope, Yehezkel S. Resheff Itay Lieder]). Son application principale est machine Learning et deep Learning où un ordinateur apprend de l'expérience et le monde est compris sous la forme d'une hiérarchie de concepts, chaque concept définissant sa relation avec des concepts plus simples (Deep Learning Pipeline : Building A Deep Learning Model With TensorFlow [Authors : Hisham El-Amir, Mahmoud Hamdy])[50].

- **Numpy**

  Le terme NumPy est en fait l'abréviation de » Numerical Python « . Il s'agit d'une bibliothèque Open Source en langage Python. On utilise cet outil pour la programmation scientifique en Python, et notamment pour la programmation en Data Science, pour l'ingénierie, les mathématiques ou la science.

  Cette bibliothèque est très utile pour effectuer des opérations mathématiques et statistiques en Python. Elle fonctionne à merveille pour la multiplication de matrices ou de tableaux multidimensionnels. L'intégration avec C/C++ et Fortran est très facile[11].

- **Matplotlib** Matplotlib est une bibliothèque complète pour créer des visualisations statiques, animées et interactives en Python. Matplotlib rend les choses faciles faciles et les choses difficiles possibles[41].

- **Sklearn** Scikit-learn est une bibliothèque Python qui fournit une interface standard pour la mise en œuvre d'algorithmes d'apprentissage automatique. Il comprend d'autres fonctions auxiliaires qui font partie intégrante du pipeline d'apprentissage automatique, telles que les étapes de prétraitement des données, les techniques de rééchantillonnage des données, les paramètres d'évaluation et les interfaces de recherche pour régler/optimiser les performances d'un algorithme[4].

5.4 Description de l'outil

Dans cette partie nous allons présenter les résultats de notre approche ainsi que les différentes interfaces de notre application, et expliquer l'utilité de chacune d'entre elles.

5.4.1 Page d'accueil

C'est la première interface accessible par l'utilisateur.

FIGURE 5.1 – Page d'accueil de l'application

5.4.2 L'interface de prédiction

Le but de cette interface est de prédire si une personne a le risque d'être diabétique ou non avec un taux de prédiction, pour cela il doit remplir le formulaire ci-dessous qui contient les informations suivantes : no of Pregnancies, Glucose, BloodPressure, SkinThikness, Insulin, BMI, DiabetePedigreeFunction et Age.

Le remplissage de tout les champs du formulaire est obligatoire, sinon un message d'erreurs sera affiché.

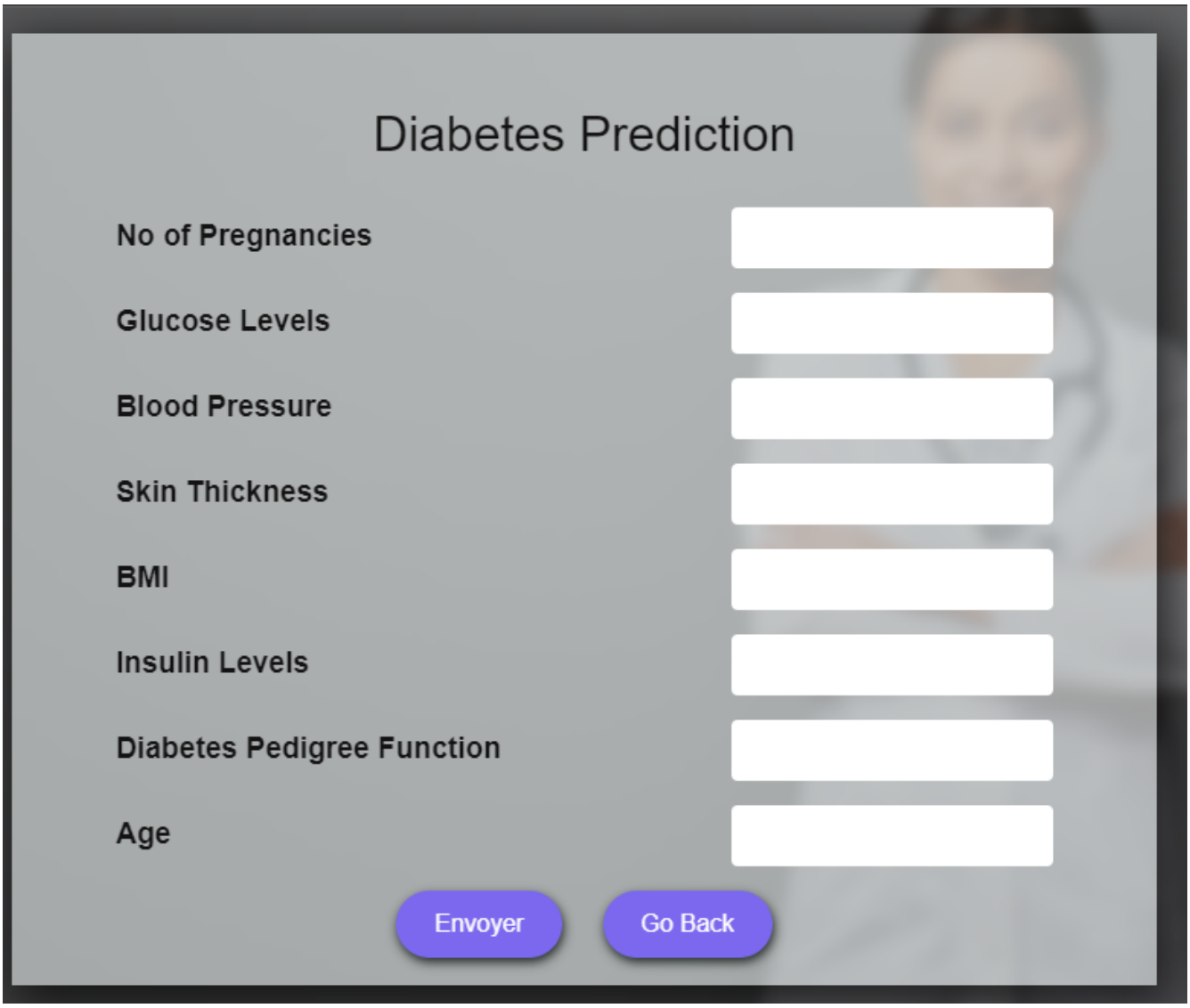

FIGURE 5.2 – L'interface de prédiction

5.4.3 Les résultats du diagnostic

Les deux figure ci-dessous présente les résultat possible que le patient peut avoir après la saisie des informations, la première figure montre que le patient est diabétique et doit consulter le médecin, et le 2ème capture montre que le patient n'est pas diabétique.

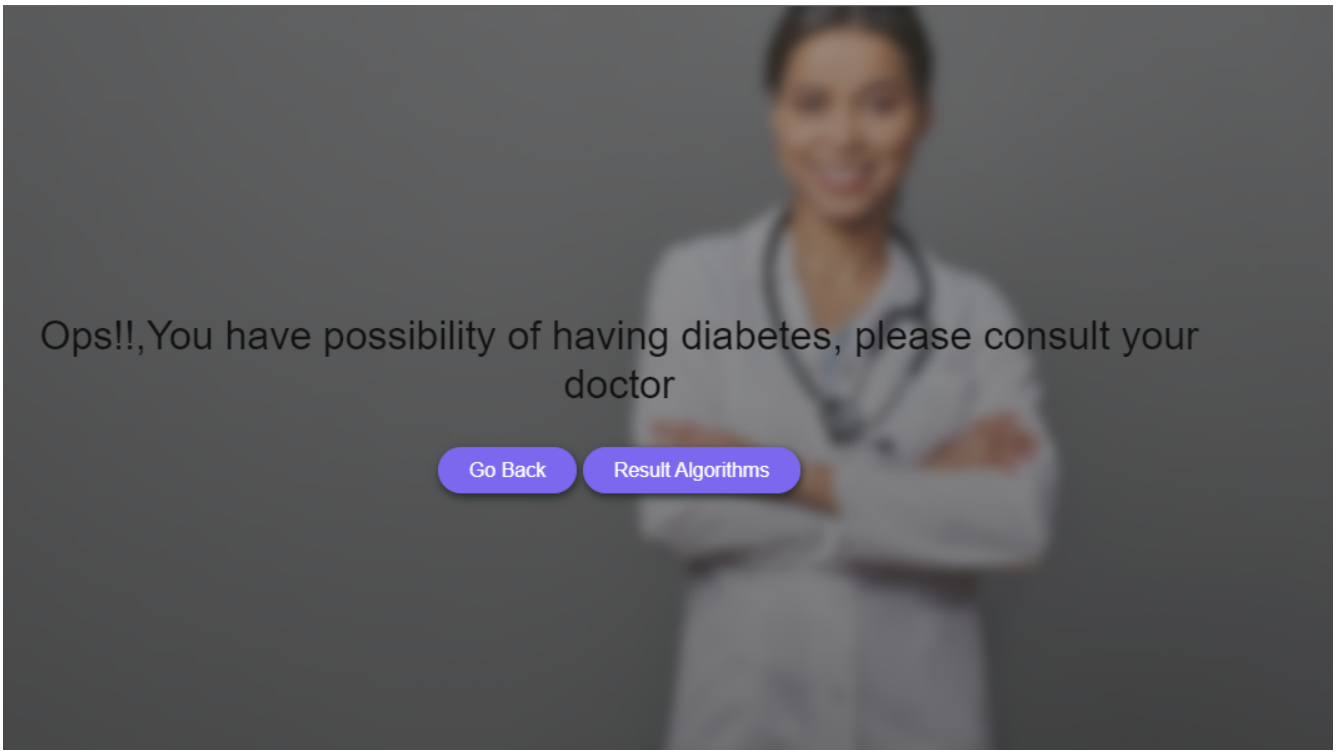

FIGURE 5.3 – Résultat du prédiction diabétique.

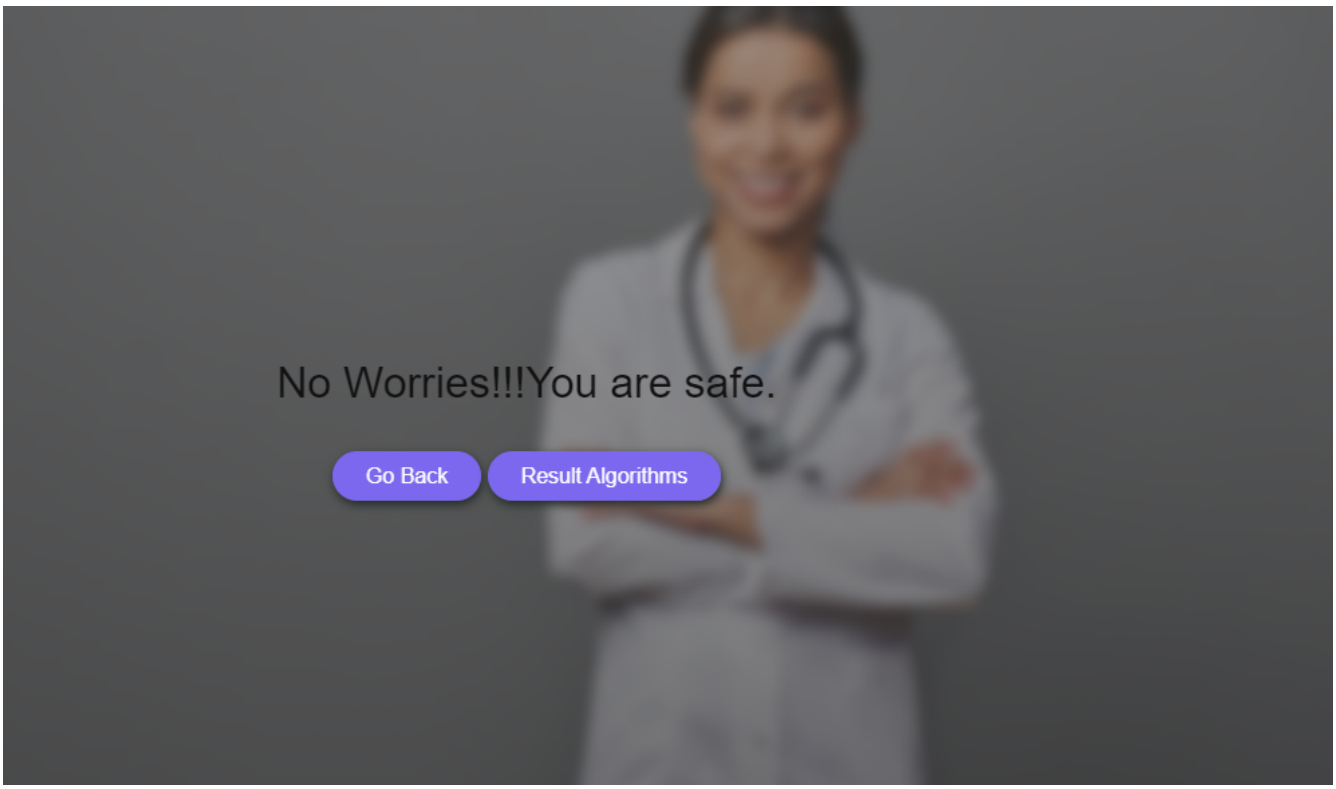

FIGURE 5.4 – Résultat du prédiction non diabétique

5.4.4 Résultats Des algorithmes

La figure 5.5 présente l'interface d'accueil des résultats des algorithmes SVM, DNN, RF , on cliquant sur les buttons RF,DNN, SVM les figures 10,11,12 s'ouvre en affichant les résultats de ces algorithmes appliqués sur les données comme la précision, le rappel, le f1-score ainsi que les graphe ROC respectivement.

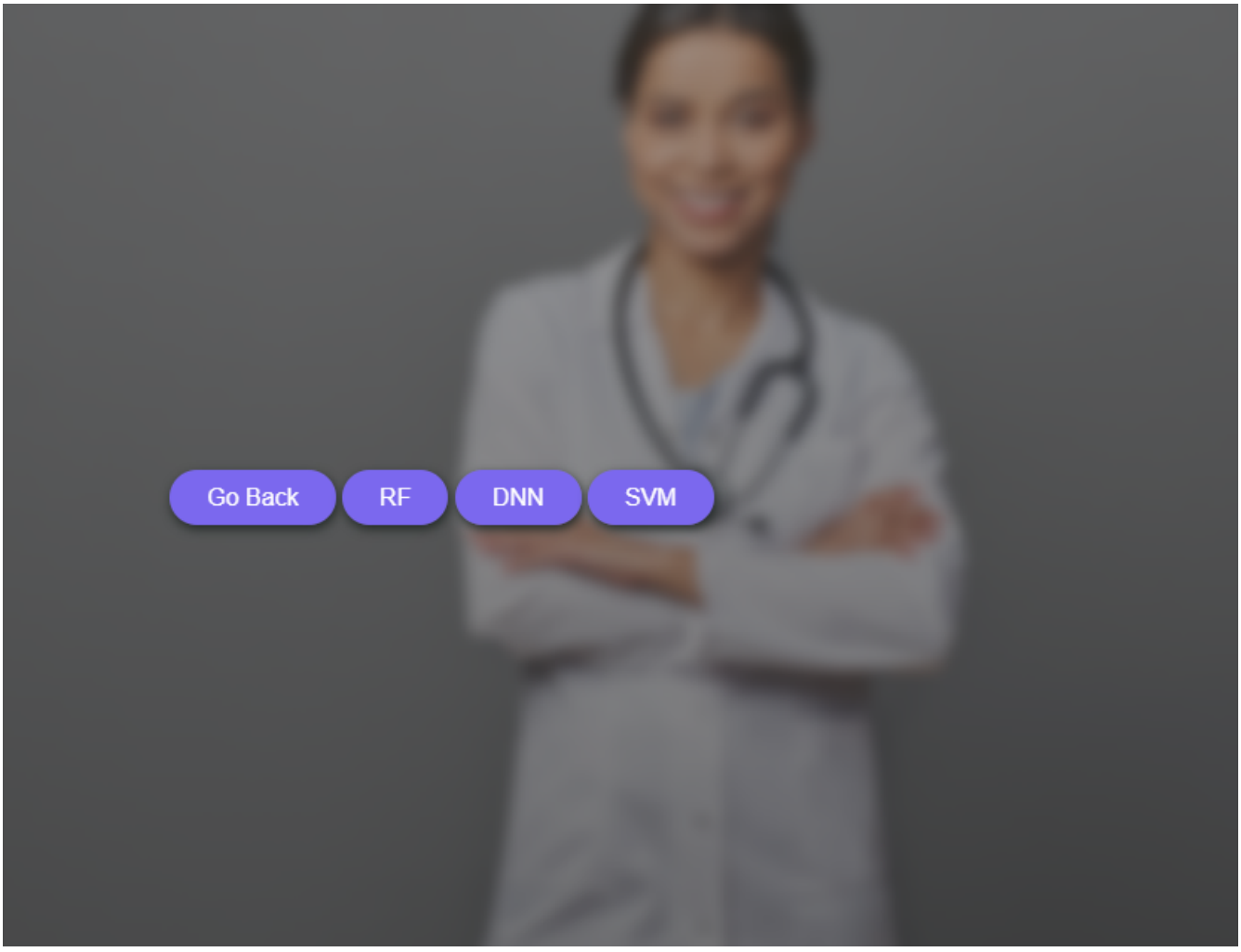

FIGURE 5.5 – Page des Résultats

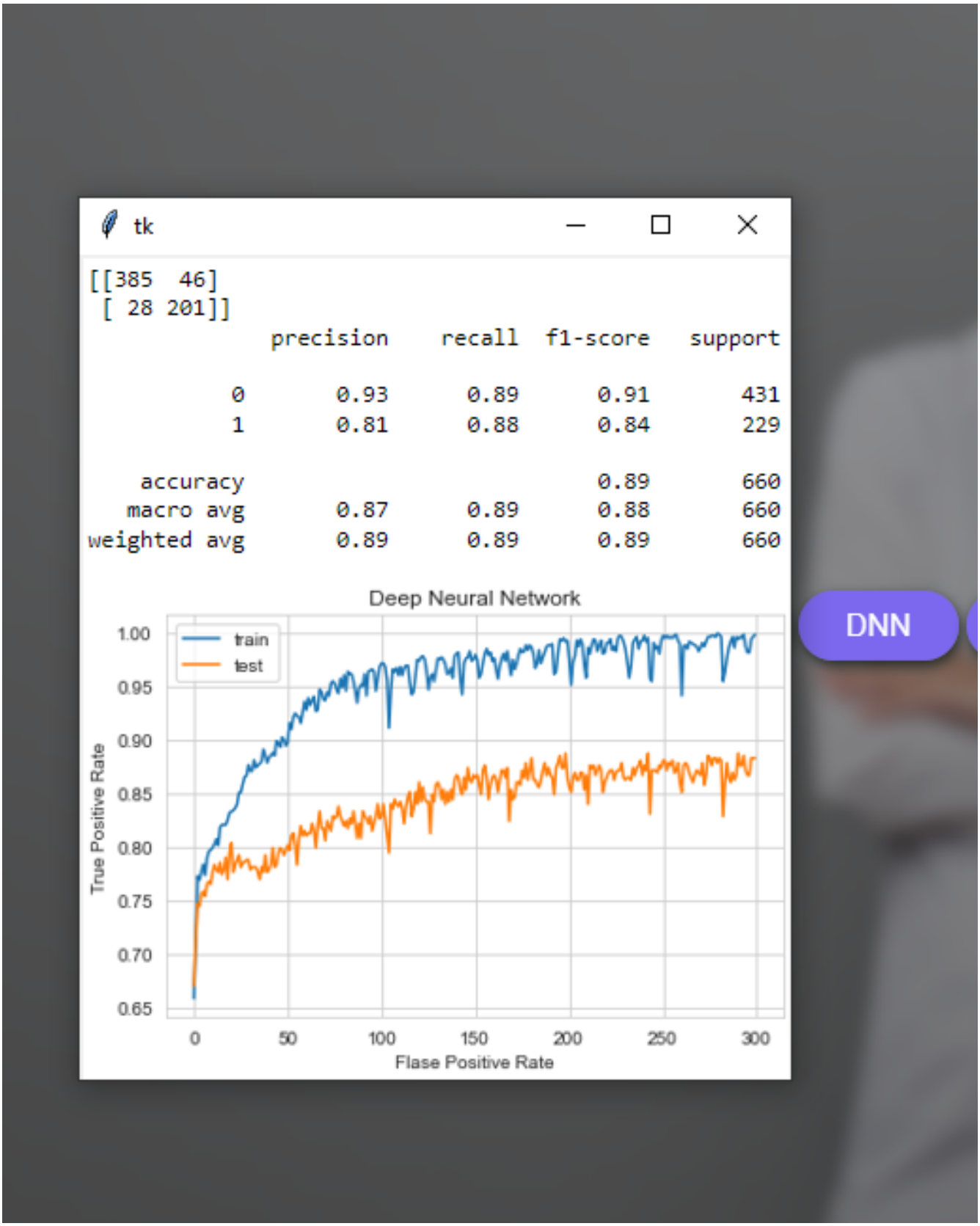

FIGURE 5.6 – interface DNN

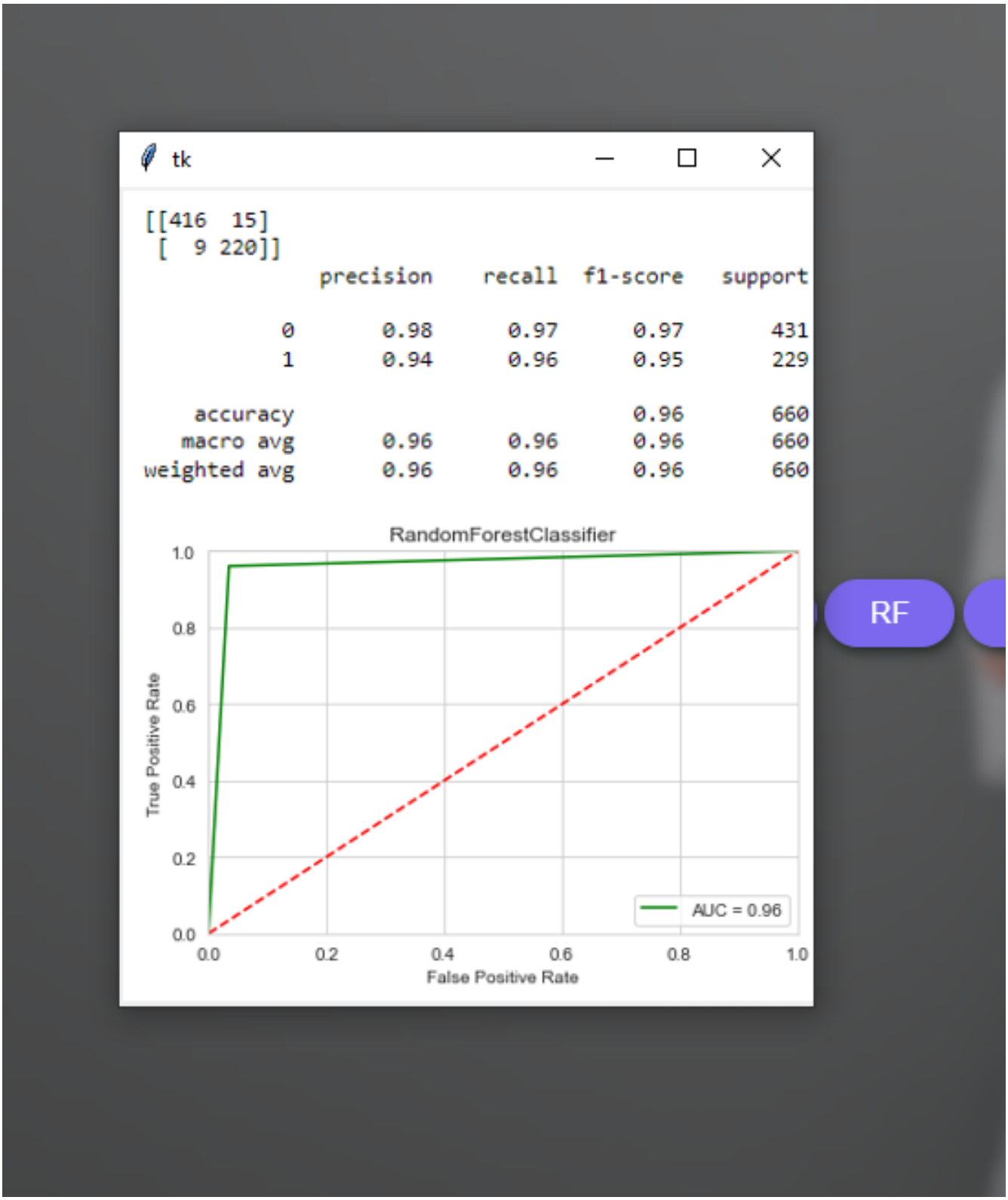

FIGURE 5.7 – interface RF

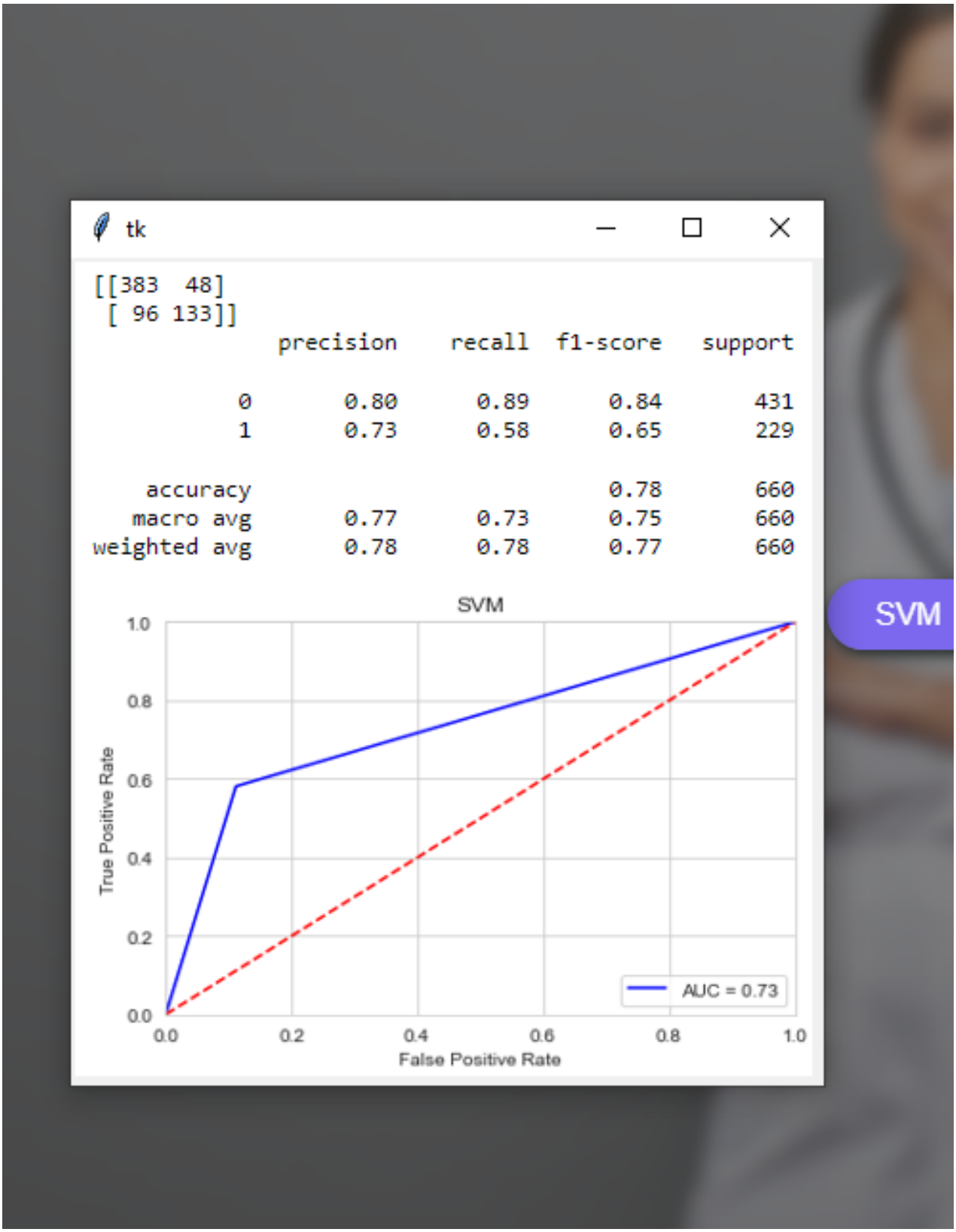

FIGURE 5.8 – interface SVM

5.5 Évaluation

Après la prédiction de diabètes, une évaluation est nécessaire pour bien déterminer les performances de l'approche choisie, les résultats de modèle ont été évaluer en analysant quelques critères à savoir les paramètres d'exactitude, précision et rappel.

L'évaluation des performances est une étape très nécessaire pour tester la qualité de modèle, afin d'assurer la fiabilité des résultats prédicatifs de modèle.

5.5.1 Accuracy

Il indique le pourcentage de bonnes prédictions. C'est un très bon indicateur parce qu'il est très simple à comprendre[66].

Accuracy=(Vrai Positif+Vrai négatif) / (Total) (1)

5.5.2 Rappel

Il se concentre uniquement sur les clients qui ont réellement résilié et donne une indication sur la part de faux négatifs. Les faux négatifs ce sont les clients qui résilient mais qui ne sont pas détectés par le score. Concrètement ce sont des clients que vous ne détectez pas et pour lesquels vous ne pourrez pas agir pour éviter leur départ.[66]

Recall=(Vrai Positif) /(Vrai positif+Faux négatif) (2)

5.5.3 Précision

C'est le 3ème indicateur vient compléter l'accuracy et le recall, l se concentre uniquement sur les clients pour lesquels le modèle a prédit une résiliation et donne une indication sur les faux positifs. Les faux positifs ce sont les clients pour lesquels le score a prédit une résiliation mais qui sont restés abonnés. C'est à dire que pour ces clients, vous engagerez surement des actions marketing pour les fidéliser mais ces actions n'étaient pas nécessaires puisqu'ils n'allaient pas résilier. Il faut limiter les faux positifs pour réduire le coût des campagnes.[66]

Precision=(Vrai Positif)/ (Vrai+Faux positif) (3)

5.5.4 ROC

La courbe ROC (Receiver Operating Characteristic) représente la sensibilité en fonction de 1 – spécificité pour toutes les valeurs seuils possibles du marqueur étudié. La sensibilité est la capacité du test à bien détecter les malades et la spécificité est la capacité du test à bien détecter les non-malades.[25]

Les figures 7, 8,9 montrent les courbes ROC des classifieurs SVM, RF et le DNN La courbe ROC du classifieurs RF couvre 96 % de la zone qui dépasse la ligne de base diagonale. La courbe ROC du classifieurs SVM couvre 73 % de la zone qui dépasse la ligne de base diagonale. La courbe ROC du l'Algorithme DNN couvre 89 % de la zone qui dépasse la ligne de base diagonale.

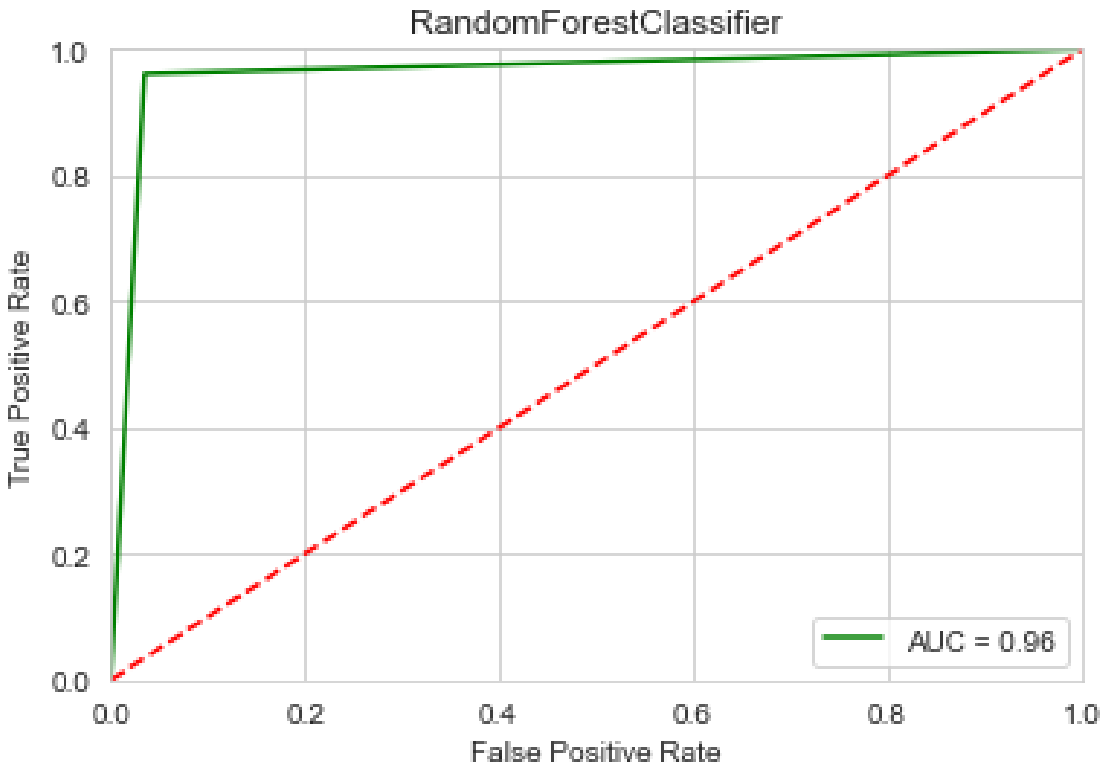

FIGURE 5.9 – La courbe ROC du classifieur Random Forest

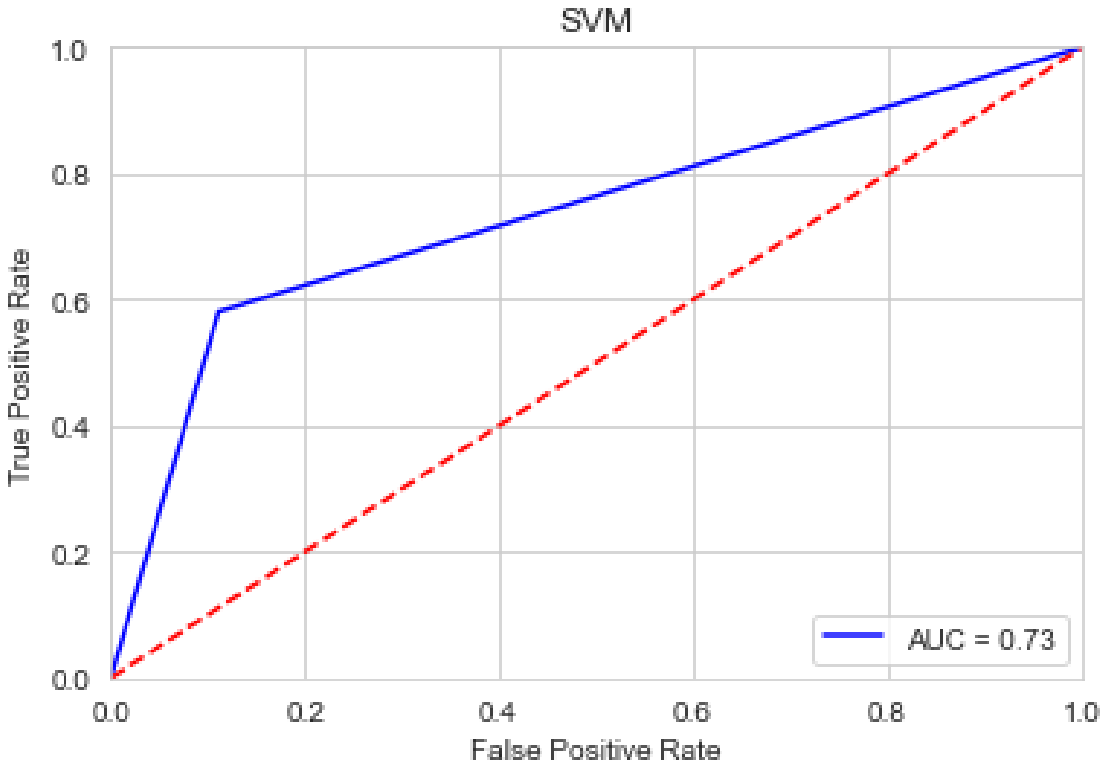

FIGURE 5.10 – La courbe ROC du classifieur SVM

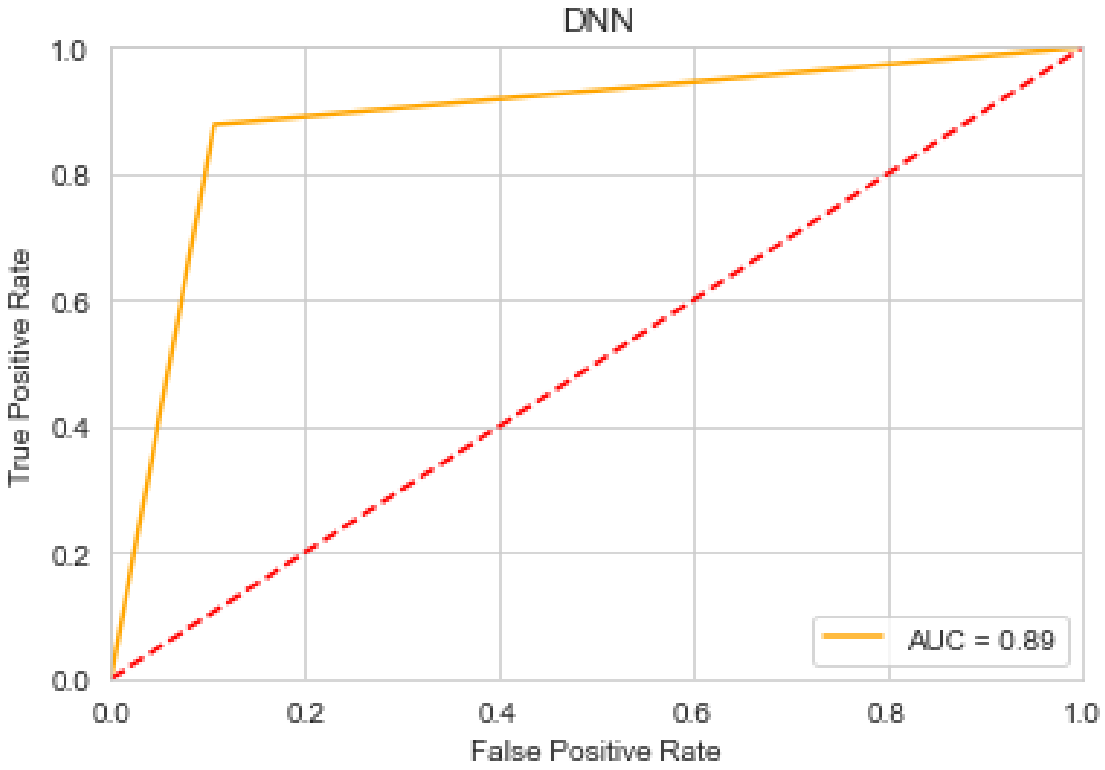

FIGURE 5.11 – La courbe ROC du classifieur Deep neural network

Le tableau ci-dessous représente la précision, le rappel et le F1-score pour les classifieurs RF , SVM, DNN.

Classifieur	Prédiction	Précision	Rappel	F1-score
DNN	Positif	0.81	0.88	0.84
	Négative	0.93	0.89	0.91
RF	Positif	0.94	0.96	0.95
	Négative	0.98	0.97	0.97
SVM	Positif	0.73	0.58	0.65
	Négative	0.80	0.89	0.84

TABLE 5.2 – Tableau de comparaison de la précision, du rappel et du F1 score des 3 classifieurs RF, DNN et SVM.

La figure 5.12 montre la comparaison des performances des algorithmes de classification en utilisant la courbe ROC.Plus une courbe a des valeurs élevées, plus l'aire sous la courbe est grande, moins le classifieur fait d'erreur.. Le classifieur RF atteint une précision moyenne de 96 % meilleur que DNN (89 %) et SVM (73 %).

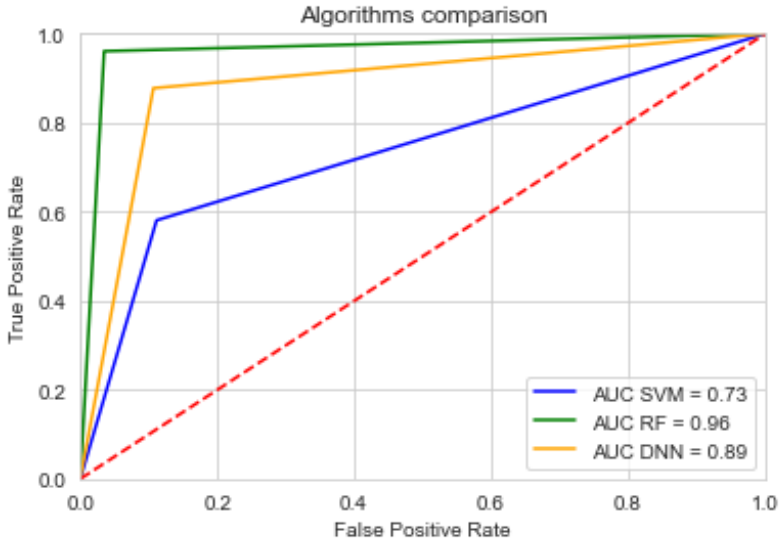

FIGURE 5.12 – Comparaison de la précision des classifieurs

Pour conclure, nous comprenons que pour l'ensemble de données, le classifieur RF fonctionne mieux que les classifieurs DNN et SVM. Il est important que nous devions considérer plusieurs ensembles de données pour avoir des meilleures précisions pour la prédiction du diabète des patients.

5.6 Conclusion

Dans ce chapitre, nous avons présenté l'essentiel de notre travail qui consiste à créer un système de prédiction du diabète chez les patients. Pour l'implémentation, nous avons utilisé des méthodes de classification les plus connues de machine learning : SVM, RF, Et le Deep neural natwork de Deep Learning.

Notre système s'intègre dans le domaine d'intelligence artificielle précisément (Machine Learning) et Deep learning. Car la précision de la classification augmente à chaque fois quand exécute l'algorithme de classification.

Dans le chapitre suivant, nous allons fini notre travail par une conclusion générale en résumant globalement notre étude.

Conclusion Générale

Ce travail a été réalisé dans le cadre de notre projet de fin de cycle Master en informatique option intelligence artificielle. Il a consisté en une approche du Machine Learning pour la prédiction précoce de la maladie du diabète chez les patients.

La prédiction du diabète est un sujet très étudié par les chercheurs au but de réaliser des systèmes prédictives afin d'aider les personnes de prédire s'il souffre de diabète Type 2, et de minimiser les risques de complication interviennent de diabètes.

L'analyse prédictive dans le domaine de la santé peut changer la façon dont comment les chercheurs et les praticiens médicaux obtiennent des informations partir de données médicales et prendre des décisions.

Le dataset qu'on a utilisé pour réaliser notre approche est extrait de l'hôpital Frankfort ,Allemagne télécharger sur le site Kaggle ,composé de 2000 patients diabétiques et non diabétiques,8 attributs ont été sélectionnées pour entraîner les modèles prédictives.

Machine Learning est la technique la plus utilisées pour les sujets de prédiction, cette technique englobe plusieurs méthodes d'apprentissage automatique pratiques pour avoir des résultats très satisfaisante, ses techniques sont faciles et simples à appliquer, Le deep Learning est tout comme machine Learning mais il nécessite une vaste base de données d'entrée pour donner des résultats fiables, dans notre travail on 'a utilisé deux classifieurs de ML sont : RF et SVM et le DNN de deep Learning.

Certaines limitations de notre travail est le manque d'un grand ensemble de données, et Pour construire un modèle avec une meilleur précision, généralement les méthodes de machine Learning ont besoin de milliers enregistrements avec zéro valeur manquante pour un bon entraînement de l'approche ce qui donnent des résultats prédicatifs fiables.

Nos perspectives futures sont d'améliorer les interfaces de notre application, de tester et de développer les algorithmes de prédiction du diabète proposés en utilisant un grand ensemble de données afin

d'améliorer les performances des classificateurs SVM, RF et DNN.

La réalisation de ce travail nous a permis d'enrichir beaucoup plus nos capacités, et de maîtriser mieux le fonctionnement des techniques machine Learning ainsi que deep Learning, à travers cette étude nous allons arriver d'apprendre de multiples connaissances au paravent, on considère ce travail le départ intéressant de nos challenges futur.

Bibliographie

[1] Nadzurah Zainal Abidin, Amelia Ritahani Ismail, and Nurul A Emran. Performance analysis of machine learning algorithms for missing value imputation. *International Journal of Advanced Computer Science and Applications*, 9(6), 2018.

[2] Gridi Adel. Un outil de deep learning pour les données textuelles.avantages de cnn (26p),université de l'arbi ben mhidi oum el bouaghi. 2020.

[3] Lemmou Amira-BellaKhdar Khaoukha-Ledjedel Adila. Avantages et inconvénients de retro propagation du gradient de l'erreur,université de m'sila algérie. 2011.

[4] Ekaba Bisong. Introduction to scikit-learn. In *Building machine learning and deep learning models on Google cloud platform*, pages 215–229. Springer, 2019.

[5] Anthony Gachagan Chigozie Enyinna Nwankpa, Winifred Ijomah. Activation functions : Comparison of trends in practice and research for deep learning. 2018.

[6] Hamed Chitsaz, Hamid Shaker, Hamidreza Zareipour, David Wood, and Nima Amjady. Short-term electricity load forecasting of buildings in microgrids. *Energy and Buildings*, 99 :50–60, 2015.

[7] Mayo clinic. Complications. https ://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444 : :text=Diabetes20dramatically20increases20the20risk,Nerve20damage20(neuropathy), Mise en ligne le 30 octobre 2020, consulté le 04 février 2022.

[8] clinique des oliviers. conséquences. https ://cliniquelesoliviers.net/fr/actualites/le-diab%C3%A8te-gestationnel-causes-cons%C3%A9quences-et-recommandations., Mise en ligne en 2017, consulté le 19 février 2022.

[9] MIT Critical Data. *Secondary analysis of electronic health records*. Springer Nature, 2016.

[10] Cherradi B. Tmiri A Daanouni, O. Diabetes diseases prediction using supervised machine learning and neighbourhood components analysis. in proceedings of the 3rd international conference on networking, information systems security (pp. 1-5). 2020.

[11] DataScientest. Numpy : la bibliothèque python la plus utilisée en data science. https ://datascientest.com/numpy, mise en ligne 20 avril 2022 , consulté le 14/08/2022.

[12] DataScientest. Découvrir l'apprentissage supervisé en 5 questions. https ://datascientest.com/apprentissage-supervise, Mise en ligne 31 mars 2022, consulté le 02 Avril 2022.

[13] La Fédération Française des Diabétiques. Comment surveiller ma glycémie ? https ://www.federationdesdiabetiques.org/diabete/glycemie., Mise en ligne le 01 Mai 2015, consulté le 04 février 2022.

[14] La Fédération Française des Diabétiques. le prédiabète. https ://www.federationdesdiabetiques.org/information/recherche-innovations-diabete/actualites/le-prediabete., Mise en ligne le 01 Mai 2015, consulté le 05 février 2022.

[15] Dr A. Djeffal. Classification réseaux de neurones.inconvénients(40p),université de biskra. 2017.

[16] M.like geeks Ebrahim. Python correlation matrix tutoriel. https ://likegeeks.com/python-correlationmatrix/, mise en ligne le 29 juillet 2020, consulté le 01 juillet 2022.

[17] Centre européen d'étude du Diabète.Diabètes et complications. Diabètes et complications à long terme. http ://ceed-diabete.org/fr/le-diabete/diabete-etcomplications/, Mise en ligne le 12 Mars 2019, consulté le 03 février 2022.

[18] Sumitra Menaria Fikirte Girma Woldemichael. Prediction of diabetes using data mining techniques. in 2018 2nd international conference on trends in electronics and informatics (icoei), ieee, 414–418. -196). 2018.

[19] Fitostic. Comment adapter l'insuline lente ? https ://fitostic.com/sante/comment-adapter-linsuline-lente/, Mise en ligne en 2022, consulté le 19 mars 2022.

[20] future sante. Diabète gestationnel : qu'est-ce que c'est ? https ://www.futura-sciences.com/sante/definitions/medecine-diabete-gestationnel-13631/., Mise en ligne en 2021, consulté le 18 février 2022.

[21] Sébastien Gavois. Ia : des nano-neurones pour « repenser l'architecture interne de l'électronique».les réseaux de neurones artificiels. https ://www.nextinpact.com/article/27283/105231-ia-nano-neurones-pour-repenser-larchitecture-interne-lelectronique, Mise en ligne le 2017, consulté le 18 juillet 2022.

[22] Benbelouaer ghada. Un système de prédiction et de prévision du diabète de type 2 , université mohamed el-bachir el-ibrahimi bordj bou arrérid(61). 2021.

[23] y. bengio Goodfellow, i. and a. courville. deep learning. mit press. 2016.

[24] Onel Harrison. Machine learning basics with the k-nearest neighbors algorithm. 2018.

[25] idbc. Tutoriel : Comment lire une courbe roc et interpréter son auc ? https ://www.idbc.fr/tutoriel-comment-lire-une-courbe-roc-et-interpreter-son-auc/, mise en ligne 2022, consulté le 15 juillet 2022.

[26] Inserm. L'insulinothérapie, traitement de référence. https ://www.inserm.fr/dossier/diabete-type-1/ : :text=Le%20traitement%20du%20diab%C3%A8te%20de,par%20des%20bact%C3%A9ries%20g%C3%A9n%C3%A9tiquement%20modifi%C3%A9es., Mise en ligne le 11 juillet 2017, consulté le 07 février 2022.

[27] java T point. Support vector machine algorithm. https ://www.javatpoint.com/machine-learning-support-vector-machine-algorithm, Mise en ligne le 2021, consulté le 15 juillet 2022.

[28] jupyter. Jupyterlab : A next-generation notebook interface). https ://jupyter.org, mise en ligne 2022,consulté le 11 aout 2022.

[29] Kaggle. diabetes. https ://www.kaggle.com/datasets/mathchi/diabetes-data-set, mise en ligne 2020,consulté le 15 avril 2022.

[30] Kaggle. diabetes. https ://www.kaggle.com/johndasilva/diabetes, mise en ligne le 2018,consulté le 14 mai 2022.

[31] lebigdata.fr. Qu'est-ce que le reinforcement learning ?

[32] leem les entreprises de médicaments. De quoi parle-t-on ? https ://www.leem.org/., Mise en ligne le 20 Aout 2012, consulté le 25 février 2022.

[33] Acervo Lima. Avantages et inconvénients de la régression logistique. https ://fr.acervolima.com/avantages-et-inconvenients-de-la-regression-logistique/, Mise en ligne septembre 2021, consulté le 05 Mai 2022.

[34] l'intelligence médicale au service de soin. quelles sont les symptômes du diabetes de type 1 ? https ://www.vidal.fr/maladies/metabolisme-diabete/diabete-type-1.html, Mise en ligne le jeudi 06 février 2020, consulté le 07 février 2022.

[35] l'intelligence médicale au service de soin. quelles sont les symptômes du diabetes de type 2 ? https ://www.vidal.fr/maladies/metabolisme-diabete/diabete-type-2.html : :text=Qu'est%2Dce%20que%20le,est%20la%20d%C3%A9finition%20du%20diab%C3%A8te., Mise en ligne le jeudi 06 février 2020, consulté le 14 février 2022.

[36] l'intelligence médicale au service de soin. qu'est-ce que le diabètede type 2 ? https ://www.vidal.fr/maladies/metabolisme-diabete/diabete-type-2.html : :text=Qu'est%2Dce%20que%20le,est%20la%20d%C3%A9finition%20du%20diab%C3%A8te, Mise en ligne le jeudi 06 février 2020, consulté le 14 février 2022.

[37] l'intelligence médicale au service de soin. quel est le traitement du diabètes de type 2 ? https ://www.vidal.fr/maladies/metabolisme-diabete/diabete-type-2/traitement.html : :text=Le%20traitement%20du%20diab%C3%A8te%20de%20type%202%20repose%20sur%20%3A,du%20tabac%20le%20cas%20%C3%A9ch%C3%A9ant., Mise en ligne le jeudi 06 février 2020, consulté le 16 février 2022.

[38] Gilles Louppe. Understanding random forests : From theory to practice. *arXiv preprint arXiv :1407.7502*, 2014.

[39] Santé Magazine. Comprendre et analyser les résultats de la glycémie. https ://www.santemagazine.fr/sante/maladies/maladies-endocriniennes-et-metaboliques/diabete/comment-depiste-t-on-un-diabete-334903., Mise en ligne le 10 Août 2022, consulté le 05 février 2022.

[40] Yusuke Mitari et Yuji Kaneda Masakazu Matusugu, Katsuhiko Mori. Subject independent facial expression recognition with robust face detection using a convolutional neural network.neural networks , p. 555–559. 2003.

[41] Matplotlib. Matplotlib : Visualization with python. https ://matplotlib.org/, mise en ligne le 2022, consulté le 14/08/2022.

[42] Medtronic. 'le diabète en quelque mots". https ://www.parlonsdiabete.com/parlons-diabete/le-diabete-en-quelques-mots., Mise en ligne le 05 Août 2012, consulté le 25 février 2022.

[43] Microsoft. Sélection des fonctionnalités (exploration de données). https ://docs.microsoft.com/fr-fr/analysis-services/data-mining/feature-selection-data-mining ?view=asallproducts-allversions, mise en ligne 2022,consulté le 05 mai 2022.

[44] Organisation mondiale de la Santé. Rapport mondial sur le diabète. 2016.

[45] Singamsetty P. K. Mustary, N. Prediction and recommendation system for diabetes using machine learning models. in handbook of research on applied intelligence for health and clinical informatics (pp. 316-327). igi global. 2022.

[46] Fabio Nelli. The pandas library—an introduction. In *Python Data Analytics*, pages 87–139. Springer, 2018.

[47] Rahman M. Ahammed B. Abedin M niruzzaman, M. Classification and prediction of diabetes disease using machine learning paradigm.health information science and systems, 8(1), 1-14. 2020.

[48] Szilard Pafka. Benchmarking random forest implementations. https ://www.r-bloggers.com/2015/05/benchmarking-random-forest-implementations/, Mise en ligne Mai 19 2015, consulté le 30 Mai 2022.

[49] Pilbox. Qu'est-ce que le diabète de type 2 ? https ://pilbox.de/en/content/36-le-diabete-de-type-2, Mise en ligne le 14 novembre 2016, consulté le 11 février 2022.

[50] Kolla Bhanu Prakash, Adarsha Ruwali, and GR Kanagachidambaresan. Introduction to tensorflow package. In *Programming with TensorFlow*, pages 1–4. Springer, 2021.

[51] Ricco Rakotomalala. Gradient boosting.inconvénients (16p),université lumière lyon 2. 2016.

[52] Ricco Rakotomalala. Svm support vector machine.inconvénients (35p),université lumière lyon 2. 2022.

[53] Bernadette M Randles, Irene V Pasquetto, Milena S Golshan, and Christine L Borgman. Using the jupyter notebook as a tool for open science : An empirical study. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–2. IEEE, 2017.

[54] Retengr. Deep learning : définition, applications, avantages et inconvénients. https ://www.retengr.com/2021/01/22/deep-learning-definitions-applications-avantages-inconvenients/, Mise en ligne juin 19 2021, consulté le 10 Mai 2022.

[55] Top Santé. Diabète de type 2 : 6 questions sur le traitement. https ://www.topsante.com/medecine/maladies-chroniques/diabete/diabete-de-type-2-6-questions-sur-le-traitement-631396., Mise en ligne le jeudi 16 février 2019, consulté le 18 février 2022.

[56] MA.Kamal N.Hamid W .Ali Shah M Sarwar. Prediction of diabetes using machine learning algorithms in healthcare.keywords-big data analytics ; predic-tive analytics ; machine learning ; healthcare.6p. 2018.

[57] Setu Shah, Xiao Luo, Saravanan Kanakasabai, Ricardo Tuason, and Gregory Klopper. Neural networks for mining the associations between diseases and symptoms in clinical notes. *Health information science and systems*, 7(1) :1–9, 2019.

[58] Rabhi Karima Sidahmad Amel. La prédiction du diabète en utilisant les algorithmes de machine learning , université amo de bouira ,71p. 2020.

[59] Rabhi Karima Sidahmed Amel. La prédiction du diabète en utilisant les algorithmes de machine learning. (2) :85.

[60] talend. Tout savoir sur le machine learning.

[61] Muhammad Atif Iqbala Yasir Alia Abdul Wahabb Safdar Ijazb Talha Imtiaz Baigb Ayaz Hussainc Muhammad Awais Malikb Muhammad Mehdi Razab Salman Ibrarb Zunish Abbasd Talha Mahboob Alama. A model for early prediction of diabetes. informatics in medicine unlocked, 16, 100204. 2019.

[62] M. J Tarokh. Type 2 diabetes prediction using machine learning algorithms. jorjani biomedicine journal, 8(3), 4-18. 2020.

[63] TechTarget. Collecte de données. https ://www.lemagit.fr/definition/Collecte-de-donnees : :text=La%20collecte%20des%20donn%C3%A9es%20permet,et%20les%20tendances%20%C3%A0%20venir, Mise en ligne le 2018, consulté le 15 juillet 2022.

[64] TopSanté. Qui est concerné par le dépistage du diabète ? https ://www.topsante.com/medecine/maladies-chroniques/diabete/commentsavoir-si-je-suis-diabetique-609768., Mise en ligne le 09 mars 2016, consulté le 04 février 2022.

[65] Guido Van Rossum and Fred L Drake. *An introduction to Python*. Network Theory Ltd. Bristol, 2003.

[66] Marie-Jeanne Vieille. Mesurer la performance d'un modèle : Accuracy, recall et precision. https ://www.lovelyanalytics.com/2020/05/26/accuracy-recall-precision/, mise en ligne 2022,consulté le 15 juillet 2022.

[67] Lavanya B. Nirmala I. Caroline S. S VijiyaKumar, K. Random forest algorithm for the prediction of diabetes. in 2019 ieee international conference on system, computation, automation and networking (icscan) (pp. 1-5). ieee. 2019.

[68] M.Sc. FRCPC Ronald Goldenberg M.D. FRCPC FACE Pamela Katz M.D. FRCPCs Zubin Punthakee, M.D. Le diabète de type 1. https ://guidelines.diabetes.ca/CDACPG/media/documents/French%202018%20CPG/03-Definition,-Classification-and-Diagnosis-FR.pdf., Mise en ligne en 2018, consulté le 06 février 2022.

[69] Équipe des professionnelles de la santé de Diabète Québec. Le traitement. https ://www.diabete.qc.ca/fr/comprendre-le-diabete/tout-sur-le-diabete/types-de-diabete/diabete-de-grossesse/., Mise en ligne Mai 2021, consulté le 18 février 2022.

Résumé

Le diabète est une maladie chronique due à un trouble du travail de pancréas, ce qui entraîne une concentration trop élevée de la glycémie dans le sang, çà peut affecter le fonctionnement de système corporel.

Le taux élevé de la glycémie dans le sang contribue à des complications, au fil du temps il peut endommager le cœur, les vaisseaux sanguins, les yeux, les reins et les nerfs ... etc. Ainsi il a cruellement besoin de développer un système capable de diagnostiquer efficacement les patients diabétiques à l'aide de détails médicaux.

Il existe plusieurs techniques de machine Learning pour l'analyse prédictive du diabète, cela peut aider les patients à prévenir cette maladie ou à la détection précoce afin d'éviter les complications.

Le résultat obtenu montre une forte relation du diabète avec les critères BMI et le glucose.

Dans notre étude de recherche, nous allons utiliser 3 techniques de machine Learning pour la prédiction du diabète qui sont DNN, Random forest et SVM, l'expérience a été appliqué à l'ensemble de données diabètes extrait de l'hôpital Frankfort.

La technique RF a fourni une meilleure précision de 96 % et peut être utile pour aider les professionnels de la santé dans le traitement.

Mots clés : Diabètes, ML, DNN, RF, SVM, Prédiction, Dataset

Abstract

Diabetes is a chronic disease due to a malfunction of the pancreas, which leads to too high a concentration of blood sugar in the blood, which can affect the functioning of the body system.

High blood sugar levels contribute to complications, over time it can damage the heart, blood vessels, eyes, kidneys and nerves...etc. So he badly needs to develop a system capable of effectively diagnosing diabetic patients using medical details.

There are several machine learning techniques for the predictive analysis of diabetes, this can help patients prevent this disease or early detection to avoid complications.

The result obtained shows a strong relationship of diabetes with the BMI criteria and glucose.

In our research study, we will use 3 machine learning techniques for diabetes prediction which are DNN, Random Forest and SVM, the experiment was applied to diabetes dataset extracted from Frankfurt hospital.

The RF technique provided better accuracy of 96% and may be useful to aid healthcare professionals in treatment.

Keywords : Diabetes, ML, DNN, RF, SVM, Prediction, Dataset