République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université 20 Aout skikda 1955

Faculté des science

Département d'Informatique

Mémoire de Master en Informatique

Spécialité : intelligence artificielle

Thème

Système Intelligent pour le Diagnostic précoce du diabète

Basè sur les Algorithmes d'Apprentissage Automatique

Encadré par

Dr : Lazhar Benoudina

Réalisé par

Khalfaoui Fairouz

Hamida Wissal

2025/2026

Remerciements

C’est avec un immense plaisir que je tiens à remercier très sincèrement toutes les personnes qui nous ont aidés et qui ont ainsi contribué à la réalisation de ce mémoire.
Je tiens à exprimer ma profonde gratitude à mon encadrant**, M. Lazhar Benoudina**, pour avoir dirigé ce travail avec bienveillance et pour ses précieux conseils. Mes remerciements vont également aux membres du jury pour l’intérêt qu’ils ont porté à notre projet de fin d’études en acceptant d’examiner ce travail et de l’enrichir par leurs propositions.
J’exprime ma reconnaissance envers ma famille et tous mes amis pour leur soutien indéfectible et leurs encouragements tout au long de ce parcours. Enfin, je voudrais également remercier toutes les personnes qui ont participé, de près ou de loin, à la réalisation de ce travail.

Dédicaces

Nous dédions ce travail à nos chers parents, pour leur amour, leur soutien et leurs sacrifices.
À nos familles et à nos amis, pour leurs encouragements et leur présence tout au long de notre parcours.
À tous nos enseignants qui ont contribué à notre formation et à l'acquisition de nos connaissances.
Enfin, à toutes les personnes qui ont participé, de près ou de loin, à la réalisation de ce travail.
Avec toute notre gratitude.

Résumé

Le diabète est une maladie chronique dont le diagnostic précoce permet de réduire les complications graves. Ce mémoire présente le développement d'un système intelligent de diagnostic précoce du diabète basé sur l'apprentissage automatique. Cinq algorithmes de classification ont été étudiés : KNN, Arbre de décision, SVM, Random Forest et MLP, en utilisant la base de données Pima Indians Diabetes Database.

Les données ont été prétraitées par nettoyage, normalisation et équilibrage des classes à l'aide de la technique SMOTE. Les modèles ont été évalués à l'aide de plusieurs métriques de performance telles que l'Accuracy, la Précision, le Recall, le F1-Score et l'AUC-ROC.

Les résultats obtenus montrent que le modèle MLP offre les meilleures performances globales avec un F1-Score de 0,6720 et un AUC-ROC de 0,8076. Une interface graphique développée sous Tkinter permet une utilisation simple et efficace du système afin d'assister le diagnostic médical.

Mots-clés : Diabète, Intelligence Artificielle, Apprentissage Automatique, MLP, Random Forest, SVM, KNN, Diagnostic Précoce.

Abstract:

Diabetes is a chronic disease whose early diagnosis can help reduce serious complications. This thesis presents the development of an intelligent system for the early diagnosis of diabetes based on machine learning techniques. Five classification algorithms were studied: KNN, Decision Tree, SVM, Random Forest, and MLP, using the Pima Indians Diabetes Database.

The dataset was preprocessed through data cleaning, normalization, and class balancing using the SMOTE technique. The models were evaluated using several performance metrics, including Accuracy, Precision, Recall, F1-Score, and AUC-ROC.

The experimental results showed that the MLP model achieved the best overall performance, obtaining an F1-Score of 0.6720 and an AUC-ROC of 0.8076. In addition, a graphical user interface developed with Tkinter was implemented to provide a simple and efficient tool for supporting medical diagnosis.

Keywords: Diabetes, Artificial Intelligence, Machine Learning, MLP, Random Forest, SVM, KNN, Early Diagnosis.

الملخص

يعد مرض السكري مرضاً مزمناً يتيح التشخيص المبكر له الحد من المضاعفات الخطيرة. يقدم هذا البحث تطوير نظام ذكي للتشخيص المبكر لمرض السكري يعتمد على التعلم الآلي. تمت دراسة خمس خوارزميات للتصنيف وهي: خوارزمية الجيران الأقرب، شجرة القرار، آلات ناقلات الدعم، الغابة العشوائية، والشبكات العصبية متعددة الطبقات، وذلك باستخدام قاعدة بيانات" بيما للهنود المصابين بالسكري" .

تمت معالجة البيانات أولياً من خلال التنظيف، والتطبيع، وموازنة الفئات باستخدام تقنية "سموت"

تم تقييم النماذج باستخدام عدة مقاييس للأداء مثل دقة التصنيف، الإحكام، الاستدعاء، ومقياس " إف-وان "

أظهرت النتائج المحصلة أن نموذج آلات ناقلات الدعم يوفر أفضل أداء بنسبة دقة بلغت 0.6720.

تؤكد هذه النتائج فعالية النمودج في التمييز بدقة بين المصابين وغير المصابين بالسكري.

ولتعزيزالجانب التطبيقي تم تطوير واجهة رسومية باستخدام مكتبة "تيكينتر" تتيح استخداماً بسيطاً وفعالاً للنظام للمساعدة في التشخيص الطبي.

الكلمات المفتاحية: السكري، الذكاء الاصطناعي، التعلم الآلي، آلات ناقلات الدعم، الغابة العشوائية، الجيران الأقرب، الشبكات العصبية متعددة الطبقات، التشخيص المبكر.

Table de matière

Introduction générale……..……...1
Chapitre 1 :Généralités sur le diabète sucré…….3
1.1 Introduction..…….……..……3
1.2 Définition et origine du diabète sucré …….……...…3
1.3 Types de diabète …….……..……...……..4
1.3.1 Diabète de type 1.…….…….……4
1.3.2 Diabète de type 2.…….……..……4
1.3.3 Diabète gestationnel …….……..…….…….5
1.4 Le prédiabète…….…….…….…….…….…….…….…….…….6
1.5 Complications du diabète.....6
1.5.1 Complications microvasculaires…….6
1.5.2 Complications microvasculaires…….……6
1.6 Symptômes du diabète……..7
1.7 Traitements……7
1.7.1 Traitement du DT1…….7
1.7.2 Traitement du DT2…….7
1.7.3 Traitement du diabète gestationnel……7
1.8 Importance du diagnostic précoce……8
1.9 Conclusion du chapitre 1……...……8
Chapitre 2 : Intelligence artificielle et apprentissage automatique..……9
2.1Introduction …….……..9
2.2 Définition de l'apprentissage automatique…….……...……9
2.3 Types d'apprentissage automatique ……..……..10
2.4 l'apprentissage supervisé ……..…….……..10
2.4.1 Comment fonctionne l'apprentissage supervisé ?.....10

2.4.2 Les étapes de l'apprentissage supervisé ……11
2.5 Les algorithmes utilisés ……..……12
2.5.1 K plus proches voisins (KNN) ……12
2.5.2 Arbre de décision (Decision Tree) ……..……13
2.5.3 Forêt aléatoire (Random Forest) ……..……14
2.5.4 Machine à vecteurs de support (SVM) ...……15
2.5.5 Perceptron multicouches (MLP) ……..……16
2.6 Métriques d'évaluation adaptées au contexte médical……17
2.7 État de l'art : apprentissage automatique pour le diagnostic du diabète…….……18
2.7.1 Jeux de données fréquents……18
2.7.2 Performances comparatives……19
2.7.3 Importance de l'explicabilité (XAI).……19
2.8 Conclusion du chapitre 2……..……20
Chapitre 3 : Conception ……...……..……21
3.1Introduction.....…21
3.2 jeu de données ……...…….21
3.2.1Source du dataset ……..……21
3.2.2 Description du dataset……..……22
3.3 Prétraitement des données ……..……22
3.3.1 Gestion des valeurs aberrantes ……22
3.3.2 Normalisation.....23
3.3.3 Division des données 23
3.3.4 Équilibrage des classes ……..……23
3.4 Modèles implémentés……...23
3.5 Métriques d'évaluation……..…23
3.6 Conclusion ……24
Chapitre 4 : Implémentation et résultats……25

4.1Introduction ……25
4.2 Environnement de développement……..…25
4.2.1 Langage et outils 25
4.2.2 Bibliothèques utilisées 25
4.3 Implémentation du prétraitement……25
4.3.1Chargement du dataset ……..……26
4.3.2 Nettoyage des données ……26
4.3.3 Normalisation et division 26
4.3.4 Application de SMOTE 27
4.4 Implémentation des modèles ……..……27
4.4.1 Définition des cinq modèles 28
4.4.2Entraînement ……..28
4.5 Résultats expérimentaux…….…...28
4.5.1 Matrices de confusion 29
4.5.2 Courbes ROC 29
4.5.3 Tableau comparatif des performances .....……30
4.5.4 Analyse des résultats 30
4.6 Interface graphique…….……31
4.6.1 Description de l'interface 32
4.6.2 Système d'ensemble pondéré 32
4.8 Conclusion.....33
Conclusion générale……..……34
Bibliographie…….35

Table des figures

Figure 1.1 – Schéma de la régulation glycémique normal…….4

Figure 1.2 – Diagramme des types de diabète……5

Figure 1.4 – Localisation des complications……...……6

Figure 2.1 – Flux de travail de l'apprentissage supervisé……..11

Figure 2.2 – Illustration du KNN…….……..12

Figure 2.3 – Exemple d'arbre de décision pour le diabète…….13

Figure 2.4 – Schéma d'une forêt aléatoire……..14

Figure 2.5 – Illustration du SVM linéaire et du noyau RBF…….…….……15

Figure 2.7 – Architecture d'un MLP à deux couches cachées.……..16

Figure 2.8 – Exemple de courbe ROC……..…….....18

Figure 2.9 – Exemple de daigramme d'importance SHAP……..20

Figure 4.1 Importation des bibliothèques……..25

Figure 4.2 - Chargement et aperçu du dataset…….....26

Figure 4.3 - Code de nettoyage des données…….……....27

Figure 4.4 - Normalisation et division des données……..27

Figure 4.5 - Définition des cinq modèles……..28

Figure 4.6 - Code d'entraînement des modèles…….….28

Figure 4.7 - Matrices de confusion des cinq modèles……..…….29

Figure 4.8 - Courbes ROC des cinq modèles…….30

Figure 4.9 - Diagramme à barres comparatif des performances……...31

Figure 4.10 - Interface graphique du système de diagnostic……..32

Figure 4.11 - Exemple de diagnostic avec interface……..33

Liste des tableaux

Tableau 1.1 – Critères diagnostiques du diabète et du prédiabète……3

Tableau 2.1 – Types d'apprentissage automatique et applications médicales ……..……..…10

Tableau 2.2- Structure d'une matrice de confusion clinique……...17

Tableau 2.2 – Accuracy typique des algorithmes sur PIDD (compilation d'études 2021-2026)……...19

Tableau 4.1 - Comparaison des performances des cinq modèles……..……30

Liste des abréviations

IA	Intelligence Artificielle
ML	Machine Learning (Apprentissage automatique)
KNN	K-Nearest Neighbors (K plus proches voisins)
DT	Tree (Arbre de décision)
SVM	Support Vector Machine (Machine à vecteurs de support)
RF	Random Forest (Forêt aléatoire)
MLP	Multi-Layer Perceptron (Perceptron multicouches)

INTRODUCTION GÉNÉRALE

L’apprentissage automatique est une discipline de l’intelligence artificielle qui cherche a trouver un moyen de créer des programmes informatiques qui s’améliorent automatiquement avec l’expérience. à travers ce mémoire de Master, nous intéresserons à l’utilisation des algorithmes d’apprentissage automatique pour la prédiction précoce du diabète de type 2 qui est un dysfonctionnement du système de régulation de la glycémie, Afin d'éviter les risques de complications de cette maladie chronique sur la santé du patient.

Problématique

Le diagnostic précoce du diabète est crucial pour prévenir ses complications graves telles que la neuropathie, la néphropathie, la rétinopathie et les maladies cardiovasculaires. Cependant, les méthodes de diagnostic traditionnelles reposent sur des tests sanguins invasifs qui nécessitent du matériel de laboratoire spécialisé. Notre objectif n'est pas de déterminer le type de diabète (type 1, type 2 ou gestationnel), mais uniquement de prédire si une personne est diabétique ou non diabétique sur la base de ses données médicales.

Face à cette limitation des méthodes traditionnelles, l'intelligence artificielle et plus particulièrement l'apprentissage automatique offre des perspectives prometteuses pour le développement d'outils de diagnostic non invasifs, rapides et précis. En analysant des données médicales facilement accessibles telles que le nombre de grossesses, la glycémie, la pression artérielle, l'épaisseur du pli cutané, le taux d'insuline, l'indice de masse corporelle, la fonction généalogique du diabète et l'âge, il est possible de construire des modèles prédictifs capables d'identifier les personnes diabétiques.

Objectifs du travail

L'objectif principal de ce mémoire est de concevoir, d'implémenter et d'évaluer un système intelligent pour le diagnostic précoce du diabète à partir des données médicales. Plus spécifiquement, nous visons à :

- ✓ Développer cinq modèles de classification supervisée : K plus proches voisins (KNN), Arbre de décision (Decision Tree), Machine à vecteurs de support (SVM), Forêt aléatoire (Random Forest) et Perceptron multicouches (MLP).
- ✓ Comparer leurs performances à l'aide de multiples métriques d'évaluation telles que la précision, le rappel, le F1-score et l'AUC-ROC.
- ✓ Créer une interface graphique intuitive pour faciliter l'utilisation clinique.
- ✓ Proposer un système d'ensemble pondéré pour améliorer la précision du diagnostic.

Structure du mémoire

Ce mémoire est organisé en quatre chapitres. Le premier chapitre présente des généralités sur le diabète incluant sa définition, son épidémiologie, ses différents types, ses symptômes, ses méthodes de diagnostic traditionnelles ainsi que ses complications. Le deuxième chapitre traite de l'intelligence artificielle et de l'apprentissage automatique appliqués au diagnostic médical. Le troisième chapitre détaille la conception et la méthodologie adoptées. Enfin, le quatrième chapitre présente l'implémentation, les résultats obtenus et leur discussion.

1.1 Introduction

Le diabète sucré est une maladie métabolique chronique dont la prévalence ne cesse d'augmenter dans le monde. Selon la Fédération Internationale du Diabète (FID, 2025), 589 millions d'adultes âgés de 20 à 79 ans vivaient avec le diabète en 2024, soit 11,1 % de la population adulte mondiale. Ce chiffre devrait atteindre 853 millions d'ici 2050. Le diabète de type 2 (DT2) représente environ 90 % des cas, et près de la moitié des personnes atteintes ignorent leur maladie, ce qui retarde le diagnostic et favorise l'apparition de complications graves [4].

1.2 Définition et origine du diabète sucré

Le diabète sucré est caractérisé par une hyperglycémie chronique résultant d'un défaut de sécrétion d'insuline, d'une résistance à l'insuline, ou des deux. L'insuline, hormone produite par les cellules β des îlots de Langerhans du pancréas, permet au glucose sanguin de pénétrer dans les cellules . En l'absence d'une action insulinique suffisante, le glucose s'accumule dans le sang .

Le terme « diabète » vient du grec diabainō (« traverser ») en référence à l'élimination excessive d'urine . « Sucré » (mellitus en latin) a été ajouté lorsque les médecins ont constaté le goût sucré de l'urine des patients, signe de la glycosurie [1].

Paramètre	**Glycémie normale**	**Prédiabète**	**Diabète**
Glycémie à jeun (g/L)	< 1,00	1,00 – 1,25	≥ 1,26
Glycémie à jeun (mmol/L)	< 5,6	5,6 – 6,9	≥ 7,0
Glycémie post-prandiale 2h (g/L)	< 1,40	1,40 – 1,99	≥ 2,00
HbA1c (%)	< 5,7	5,7 – 6,4	≥ 6,5

Tableau 1.1 – Critères diagnostiques du diabète et du prédiabète

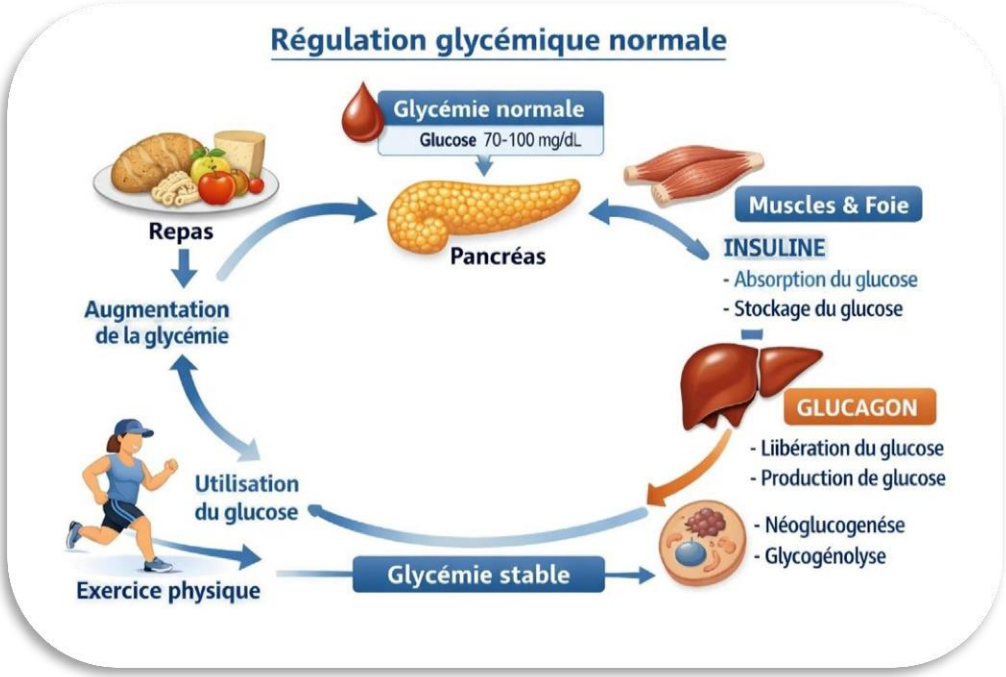


Figure 1.1 – Schéma de la régulation glycémique normale

1.3 Types de diabète

1.3.1 Diabète de type 1 (DT1)

Le DT1 résulte d'une destruction auto-immune des cellules β, conduisant à une insulinopénie absolue .

Il représente 5 à 10 % des diabètes , Environ 9,5 millions de personnes vivent avec un DT1 dans le monde , Le traitement repose sur une insulinothérapie à vie[1] .

1.3.2 Diabète de type 2 (DT2)

Le DT2 associe une résistance à l'insuline et un défaut progressif de sécrétion .

Les principaux facteurs de risque sont : obésité (IMC > 25), sédentarité, antécédents familiaux, âge > 45 ans, hypertension artérielle, dyslipidémie[14] .

Le traitement commence par des mesures hygiéno-diététiques, puis la metformine, d'autres antidiabétiques oraux, et parfois l'insuline.

1.3.3 Diabète gestationnel (DG)

Le DG survient pendant la grossesse, généralement au deuxième ou troisième trimestre .

Il concerne environ 7 % des grossesses . Les facteurs de risque incluent un âge maternel > 35 ans, un surpoids pré-gestationnel, des antécédents de DG [1]. Plus de 50 % des femmes ayant eu un DG développeront un DT2 dans les 5 à 10 ans suivant l'accouchement [4].

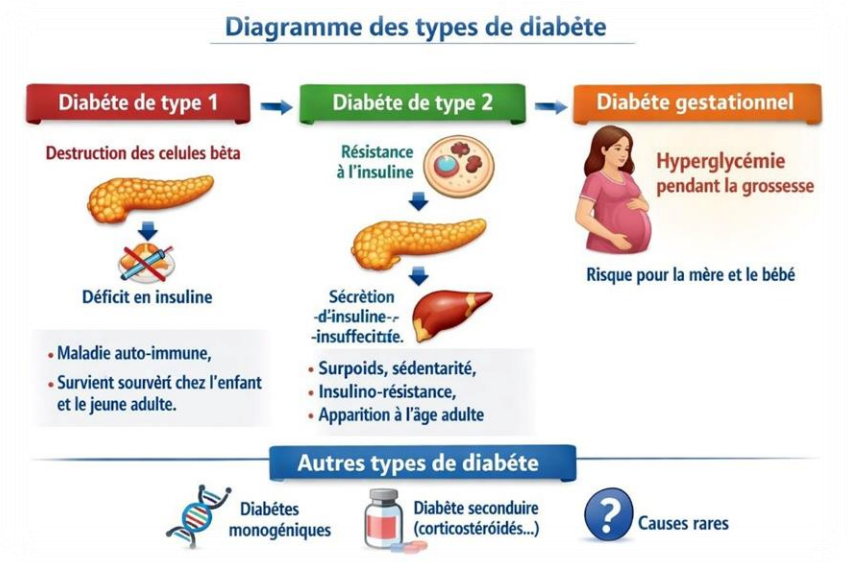


Figure 1.2 – Diagramme des types de diabète

1.4 Le prédiabète

Le prédiabète est un état intermédiaire défini par une glycémie à jeun entre 1,00 et 1,25 g/L (5,6-6,9 mmol/L), ou une intolérance au glucose (glycémie à 2h après charge orale entre 1,40 et 1,99 g/L), ou une HbA1c entre 5,7 et 6,4 %. Sans intervention, 15 à 30 % des prédiabétiques évoluent vers un DT2 dans les 5 ans .

L'étude Diabetes Prevention Program a démontré qu'une perte de 5-7 % du poids associée à 150 minutes d'activité physique par semaine réduit de 58 % le risque d'évolution [14].

1.5 Complications du diabète

1.5.1 Complications microvasculaires

· **Rétinopathie diabétique :** première cause de cécité acquise chez l'adulte actif .

· **Néphropathie diabétique :** première cause d'insuffisance rénale terminale .

· **Neuropathie diabétique :** troubles sensitifs, douleurs neuropathiques, pied diabétique .

1.5.2 Complications macrovasculaires

· **Maladie coronarienne :** (risque multiplié par 2 à 4) .

· **Accident vasculaire cérébral** .

· **Artériopathie des membres inférieurs** .

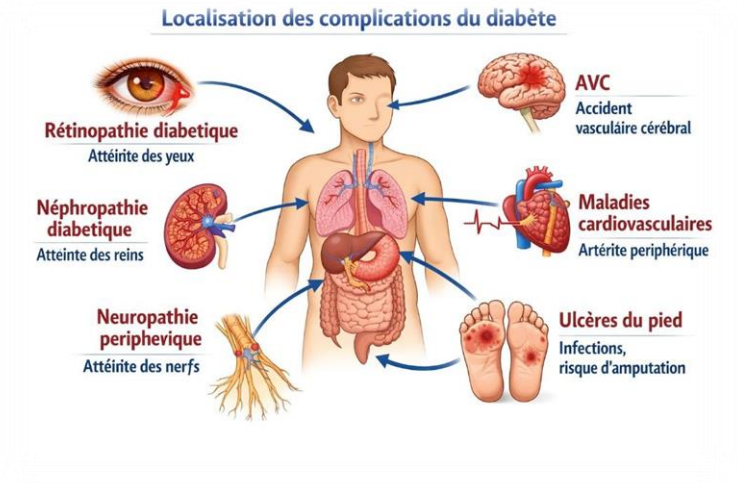


Figure 1.4 – Localisation des complications

1.6 Symptômes du diabète

Les symptômes classiques (polyurie, polydipsie, polyphagie, amaigrissement) sont plus marqués dans le DT1 . Dans le DT2, le diagnostic est souvent fortuit .

Les signes à rechercher sont : polyurie et pollakiurie (surtout nocturne), polydipsie intense, fatigue inexpliquée, vision floue, infections cutanées ou génitales à répétition, cicatrisation lente [14].

1.7 Traitements

1.7.1 Traitement du DT1

Insulinothérapie substitutive selon un schéma basal-bolus (insuline lente une ou deux fois par jour, insuline rapide avant chaque repas). Autosurveillance glycémique indispensable [1].

1.7.2 Traitement du DT2

1. Mesures hygiéno-diététiques (perte de poids, activité physique).
2. Metformine (biguanide) en première intention .
3. Association à d'autres classes : sulfamides hypoglycémiants, gliptines (DPP-4 inhibiteurs), gliflozines (SGLT2 inhibiteurs), agonistes du GLP-1 .
4. Insulinothérapie en cas d'échec des traitements oraux [1].

1.7.3 Traitement du diabète gestationnel

Régime alimentaire adapté, activité physique modérée, insulinothérapie si les objectifs glycémiques ne sont pas atteints [1].

1.8 Importance du diagnostic précoce

Le diagnostic précoce du diabète et du prédiabète permet de :

· Réduire de 40 à 60 % le risque de complications microvasculaires (études DCCT et UKPDS) .

· Intervenir au stade de prédiabète pour prévenir ou retarder l'apparition du DT2 .

· Diminuer les coûts de santé, car les complications tardives représentent l'essentiel des dépenses.

C'est dans cette optique que notre travail propose d'utiliser l'apprentissage automatique pour développer un système d'aide au diagnostic précoce basé sur des données médicales facilement accessibles[4].

1.9 Conclusion du chapitre 1

Ce chapitre a présenté les bases médicales du diabète : définition, types (DT1, DT2, gestationnel), prédiabète, complications, symptômes et traitements .

Il a souligné l'enjeu majeur du dépistage précoce pour la prévention des complications.

Le chapitre suivant expose les techniques d'apprentissage automatique qui seront mises en œuvre pour répondre à cet enjeu[14].

CHAPITRE 2 : APPRENTISSAGE AUTOMATIQUE POUR LE DIAGNOSTIC DU DIABÈTE

2.1 Introduction

L'intelligence artificielle, et plus particulièrement l'apprentissage automatique (machine learning), connaît un essor considérable dans le domaine médical. Les algorithmes de classification supervisée permettent d'analyser de grandes quantités de données cliniques et d'identifier des patients à risque de diabète avec une bonne précision .

Ce chapitre présente les concepts fondamentaux de l'apprentissage automatique, les six algorithmes que nous avons sélectionnés (KNN, arbre de décision, forêt aléatoire, SVM, Naïve Bayes, et perceptron multicouches), les métriques d'évaluation adaptées au contexte médical, et un état de l'art des travaux récents (2021-2026) [13].

2.2 Définition de l'apprentissage automatique

Selon la définition classique de Tom M. Mitchell (1997) : « Un programme informatique apprend de l'expérience E par rapport à une classe de tâches T et à une mesure de performance P, si sa performance aux tâches T, mesurée par P, s'améliore avec l'expérience E » .

Dans notre cas, l'expérience E est un ensemble de données médicales de patients avec leur diagnostic, la tâche T est la prédiction du statut diabétique pour un nouveau patient, et la mesure P est la sensibilité, la spécificité ou l'exactitude [8].

2.3 Types d'apprentissage automatique

Type	**Principe**	**Application au diabète**
Supervisé	Apprendre à partir de paires (entrée, étiquette)	Classification (diabétique vs non diabétique)
Non supervisé	Découvrir des structures sans étiquettes	Clustering de sous-types de diabète
Semi-supervisé	Utiliser quelques étiquettes + beaucoup de données non étiquetées	Diagnostic avec peu d'annotations médicales
Par renforcement	Apprendre par essais-erreurs via des récompenses	Optimisation de la posologie d'insuline

Tableau 2.1 – Types d'apprentissage automatique et applications médicales

2.4 L'apprentissage supervisé

2.4.1 Comment fonctionne l'apprentissage supervisé ?

Un modèle supervisé f : $\mathbb{R}$^p → {0,1} est entraîné sur un ensemble de n exemples {(x_i, y_i)} où x_i est un vecteur de p variables (âge, BMI, glycémie, etc.) et y_i l'étiquette (0 : non diabétique, 1 : diabétique) .

L'objectif est de minimiser l'erreur de prédiction sur des données non vues (généralisation) [7].

2.4.2 Les étapes de l'apprentissage supervisé

1. Collecte des données : acquisition d'un jeu de données pertinentes.

2. Prétraitement : gestion des valeurs manquantes (imputation), normalisation (Z-score ou Min-Max), rééquilibrage des classes (SMOTE) .

3. Partitionnement : division en ensemble d'entraînement (70-80 %), validation (10-15 %) et test (10-15 %). Utilisation de la validation croisée stratifiée 5-fold .

4. Sélection et entraînement des modèles.

5. Optimisation des hyperparamètres : Grid Search ou Random Serch .

6. Évaluation : tests finaux sur l'ensemble de test indépendant.

7. Interprétabilité du modèle : recours à des méthodes d'explicabilité (SHAP, LIME) [9].

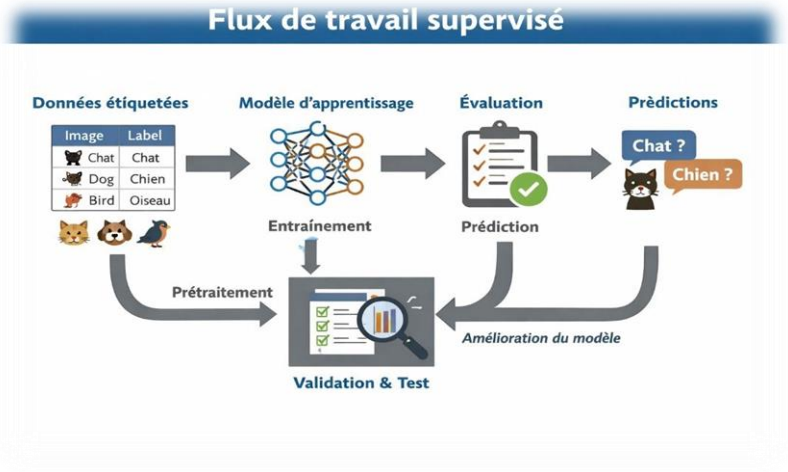


Figure 2.1 – Flux de travail de l'apprentissage supervisé

2.5 Les algorithmes utilisés

2.5.1 K-plus proches voisins (KNN)

Le KNN est un algorithme non paramétrique et paresseux . Pour classer un nouvel individu x, on trouve les K individus de l'entraînement les plus proches selon une distance (euclidienne, Manhattan) et on prend la classe majoritaire .

Le choix de K (généralement entre 3 et 21) se fait par validation croisée .

Avantages : simplicité, aucune hypothèse sur la distribution.

Inconvénients : lent en prédiction, sensible à l'échelle, performance dégradée en grande dimension.

Une étude récente sur le diabète a obtenu 87,5 % d'accuracy avec KNN optimisé [6].

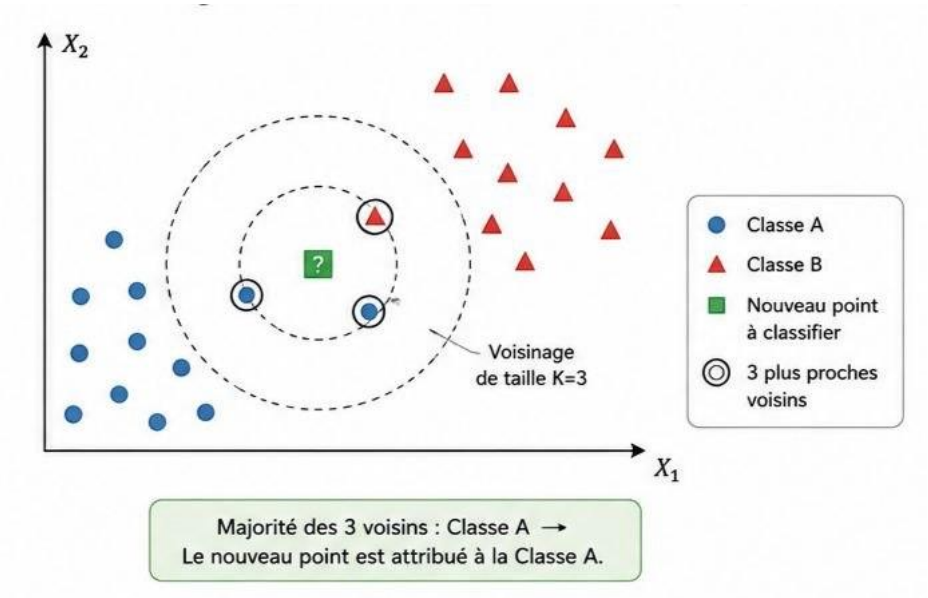


Figure 2.2 – Illustration du KNN (K=3 et K=5)

2.5.2 Arbre de décision (Decision Tree)

L'arbre de décision partitionne récursivement l'espace des variables selon des critères de pureté (gain d'information basé sur l'entropie, ou index de Gini) [8]. Chaque nœud interne correspond à un test (ex. « glycémie > 1,10 g/L ? »), chaque feuille à une classe.

Avantages : excellente interprétabilité, pas de normalisation nécessaire.

Inconvénients : sujet au surapprentissage ; remède : élagage (pruning) ou contraintes (profondeur max, nombre min d'échantillons par feuille).

Sur le jeu PIMA, l'arbre élagué atteint environ 85-92 % d'accuracy [3].

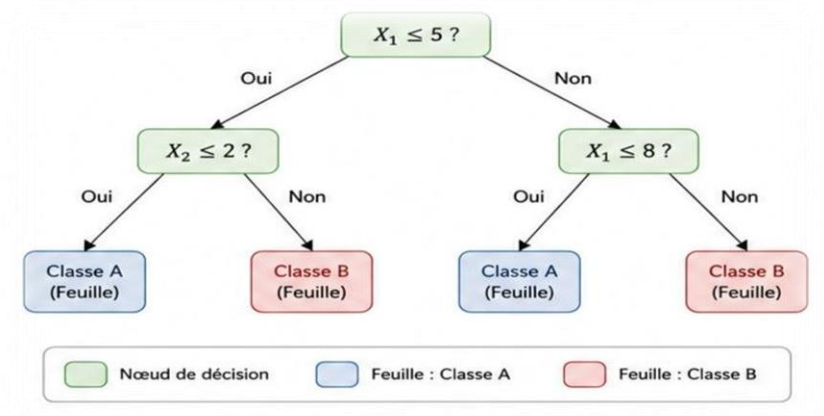


Figure 2.3 – Exemple d'arbre de décision pour le diabète

2.5.3 Forêt aléatoire (Random Forest)

La forêt aléatoire est une méthode d'ensemble qui construit de nombreux arbres de décision décorrélés . Chaque arbre est entraîné sur un bootstrap des données, et à chaque nœud on ne considère qu'un sous-ensemble aléatoire des variables (typiquement √p) .

Les prédictions sont agrégées par vote majoritaire. Hyperparamètres : nombre d'arbres (100-500), profondeur max, max_features .

Avantages : grande précision, robustesse, mesure d'importance des variables.

Inconvénients : moins interprétable qu'un seul arbre, plus coûteux La forêt aléatoire atteint souvent 91-98 % d'accuracy sur PIMA [5].

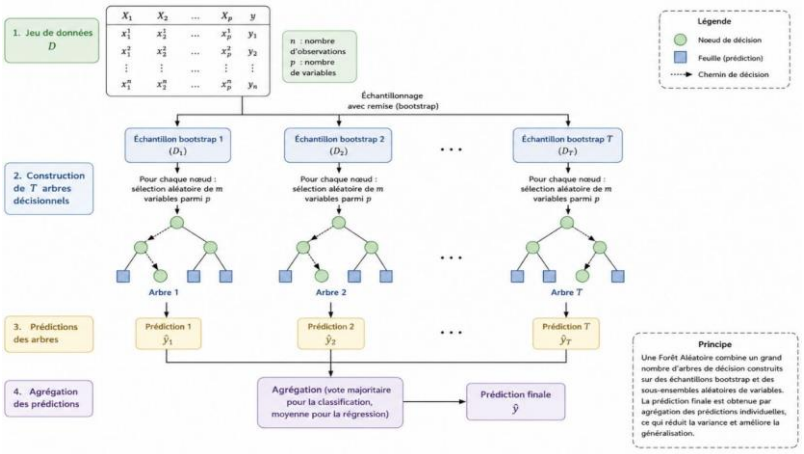


Figure 2.4 – Schéma d'une forêt aléatoire

2.5.4 Machine à vecteurs de support (SVM)

Le SVM cherche l'hyperplan séparateur qui maximise la marge entre les deux classes,Les points d'entraînement les plus proches de l'hyperplan sont appelés vecteurs de support.

Pour les données non linéairement séparables, on utilise l'astuce du noyau (kernel trick) : projection dans un espace de grande dimension via une fonction noyau K(x_i, x_j) (linéaire, polynomial, RBF (gaussien)).

Hyperparamètres : C (pénalité des erreurs), γ (pour le noyau RBF).

Avantages : efficace en haute dimension, bonne généralisation.

Inconvénients : réglage sensible, boîte noire. SVM avec noyau RBF atteint 89-94 % d'accuracy sur PIMA [8].

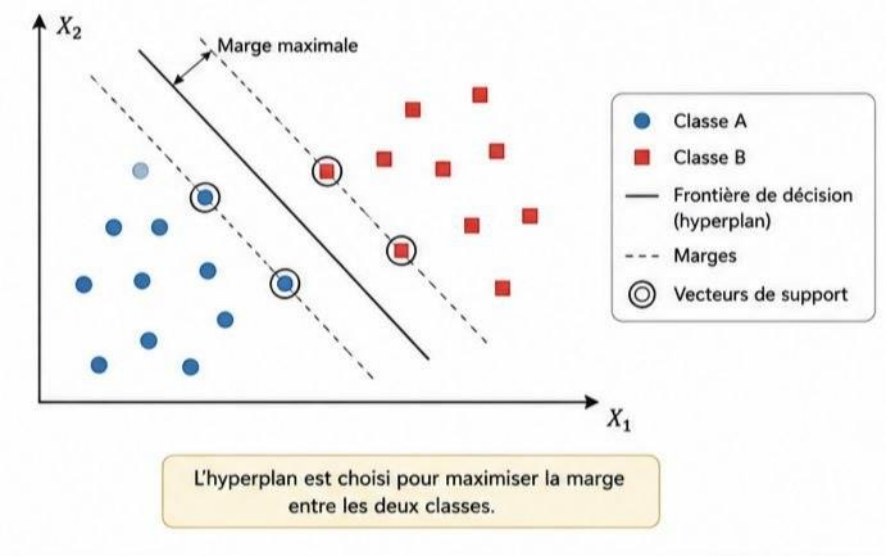


Figure 2.5 – Illustration du SVM linéaire et du noyau RBF

2.5.5 Perceptron multicouches (MLP)

Le MLP est un réseau de neurones feedforward composé d'une couche d'entrée (p neurones), d'une ou plusieurs couches cachées, et d'une couche de sortie (1 neurone avec sigmoïde pour la classification binaire).Chaque neurone applique a = f(W·x + b), où f est une fonction d'activation (ReLU, tanh, sigmoïde). L'apprentissage se fait par descente de gradient stochastique et rétropropagation . Hyperparamètres : nombre de couches cachées, nombre de neurones par couche, taux d'apprentissage, nombre d'époques, taille de batch, régularisation (dropout, L2) .

Avantages : capacité à modéliser des relations fortement non linéaires.

Inconvénients : nécessite beaucoup de données (risque de surapprentissage), boîte noire, coût calculatoire.

Selon une étude récente, un MLP à 2 couches cachées (16 et 8 neurones) atteint 92,3 % d'accuracy sur PIMA après normalisation et SMOTE [12].

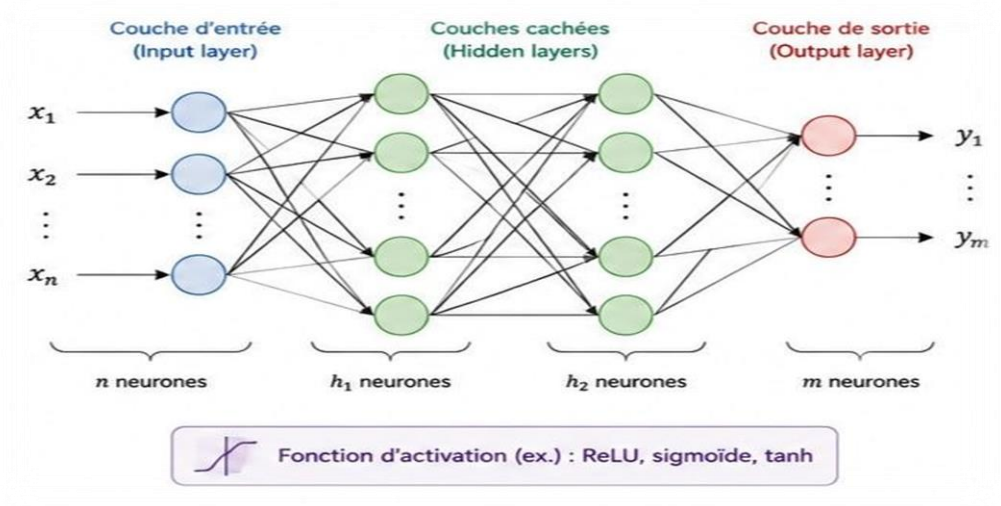


Figure 2.7 – Architecture d'un MLP à deux couches cachées

2.6 Métriques d'évaluation adaptées au médical

Dans le diagnostic médical, la sensibilité (rappel) est prioritaire car un faux négatif (patient diabétique non détecté) peut avoir des conséquences graves, alors qu'un faux positif n'entraîne qu'un bilan complémentaire [2].

Réel \ Prédit	**Prédit positif (diabétique)**	**Prédit négatif (non diabétique)**
Réel positif	Vrai Positif (VP)	Faux Positif (FP)
Réel négatif	Faux Positif (FP)	Vrai Négatif (VN)

Tableau 2.2- Structure d'une matrice de confusion clinique

les métriques mathématiques de performance sont calculées comme suit:

· **Sensibilité (Recall)** = VP / (VP + FN) → à maximiser

· **Spécificité** = VN / (VN + FP)

· **Précision** = VP / (VP + FP)

· **F1-score** = 2 × (Précision × Rappel) / (Précision + Rappel)

· **Accuracy** = (VP + VN) / (VP + VN + FP + FN)

· **AUC-ROC :** aire sous la courbe ROC (taux de vrais positifs vs taux de faux positifs). Une AUC ≥ 0,85 est excellente [14].

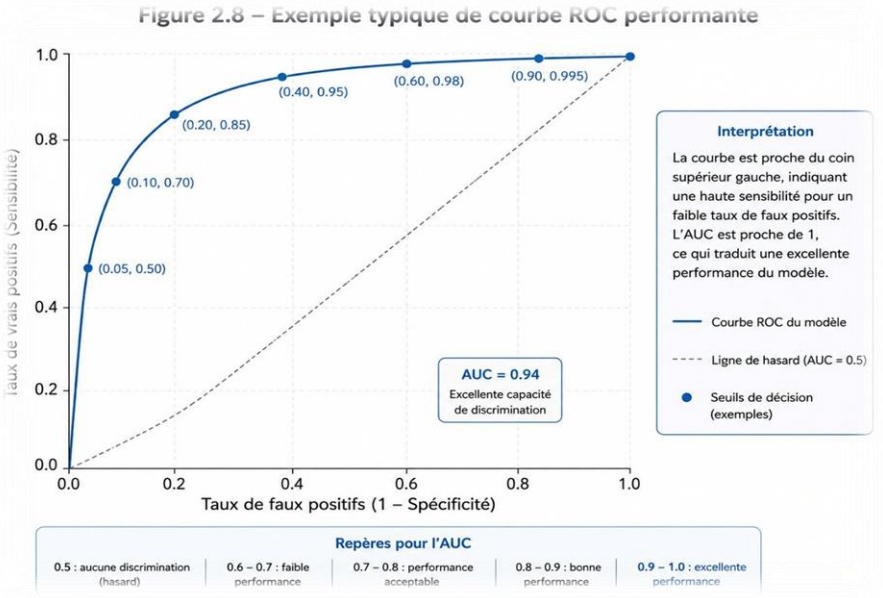


Figure 2.8 – Exemple de courbe ROC

2.7 État de l'art : apprentissage automatique pour le diagnostic du diabète

2.7.1 Jeux de données fréquents

Le plus utilisé est le PIMA Indians Diabetes Dataset (PIDD) de l'UCI / Kaggle : 768 femmes amérindiennes, 8 variables cliniques (nombre de grossesses, glycémie à jeun, pression artérielle, épaisseur du pli cutané, insulinémie, BMI, fonction pedigree du diabète, âge). D'autres bases incluent le BRFSS (Behavioral Risk Factor Surveillance System) et le NHANES [3].

2.7.2 Performances comparatives

Algorithme	Accuracy(%)	Avantage clé	Inconvénient majeur
Random Forest	91-98	Robuste, importance des variables	Moins interprétable
SVM (RBF)	89-94	Haute dimension	Réglage sensible
MLP (2 couches)	90-93	Non-linéarités	Boîte noire
Decision Tree	85-92	Interprétable	Surapprentissage
Decision Tree	80-88	Simplicité	Lent en prédiction

Tableau 2.2 – Accuracy typique des algorithmes sur PIDD (compilation d'études 2021-2026)

Les meilleures performances (jusqu'à 99,4 %) sont obtenues par des approches hybrides (optimisation par algorithmes génétiques, boosting, stacking) .
Une revue systématique récente confirme la domination des méthodes d'ensemble [5].

2.7.3 Importance de l'explicabilité (XAI)

Pour une acceptation clinique, les modèles boîte noire (SVM, MLP, RF) doivent être interprétés. La méthode SHAP (SHapley Additive exPlanations) est devenue standard. Netayawijit et al. (2025) ont montré que les variables les plus prédictives sur PIDD sont : la glycémie, l'âge, le BMI, et le nombre de grossesses [13].

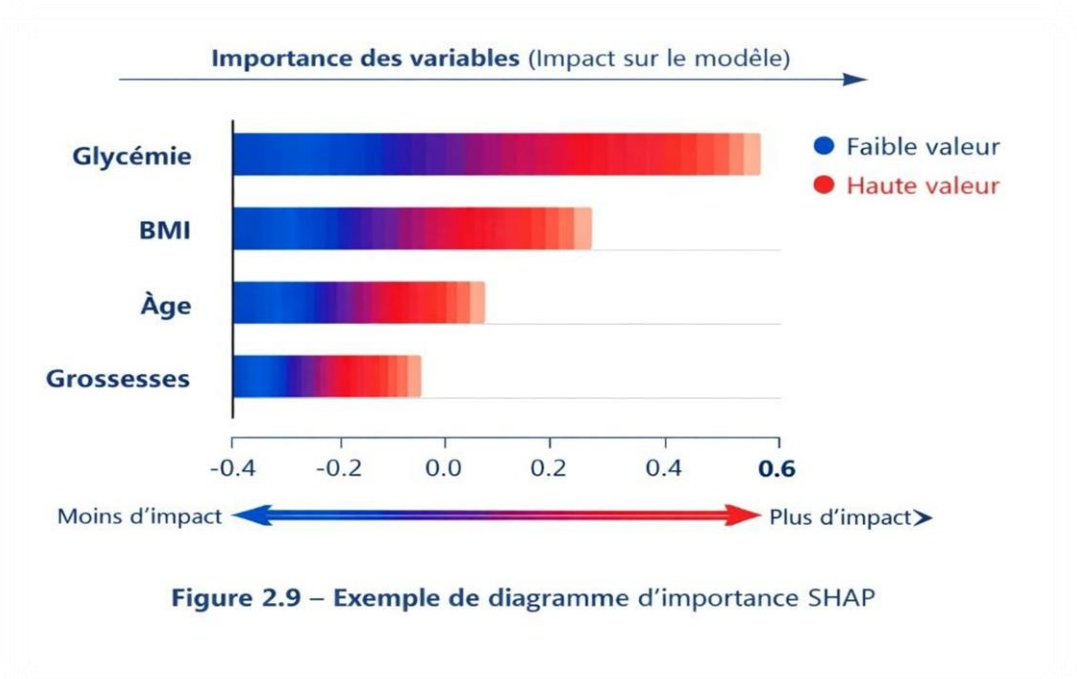


Figure 2.9 – Exemple de daigramme d'importance SHAP

2.8 Conclusion du chapitre 2

Ce chapitre a présenté les fondements de l'apprentissage automatique supervisé, les six algorithmes retenus (KNN, arbre de décision, forêt aléatoire, SVM, Naïve Bayes, MLP), les métriques adaptées au médical, et un état de l'art actualisé.

Ces connaissances seront mises en œuvre dans les chapitres suivants (conception, implémentation, évaluation).

3.1Introduction

Ce chapitre présente la conception de notre système de diagnostic du diabète. Nous décrivons ledataset utilisé, les étapes de prétraitement appliquées, les modèles implémentés, les métriques d'évaluation retenues, ainsi que l'interface graphique conçue.

3.2 Jeu de données

3.2.1 Source du dataset

Le dataset utilisé est le Pima Indians Diabetes Database, disponible sur la plateforme Kaggle. Ilprovient de l'Institut National du Diabète et des Maladies Digestives et Rénales (NIDDK) des ÉtatsUnis. Les patientes sont des femmes âgées d'au moins 21 ans d'origine indienne Pima.

3.2.2 Description du dataset

Le dataset contient 768 échantillons et 9 colonnes : huit variables médicales et une variable cible "Outcome".

Variable	**Description**	**Type**
Pregnancies	Nombre de grossesses	Numérique
Glucose	Concentration de glucose (mg/dL)	Numérique
BloodPressure	Pression artérielle diastolique (mm Hg)	Numérique
SkinThickness	Épaisseur du pli cutané (mm)	Numérique
Insulin	Taux d'insuline (mu U/ml)	Numérique
BMI	Indice de masse corporelle (kg/m^2)	Numérique
DiabetesPedigreeFunction	Fonction généalogique du diabète	Numérique
Age	Âge en années	Numérique
Outcome	0 = non diabétique, 1 = diabétique	Cible (binaire)

3.3 Prétraitement des données

3.3.1 Gestion des valeurs aberrantes

Certaines variables (Glucose, BloodPressure, SkinThickness, Insulin, BMI) contiennent des valeurs nulles médicalement impossibles (zéros).

Ces valeurs sont remplacées par NaN puis imputées par la médiane.

3.3.2 Normalisation

Les données sont normalisées avec StandardScaler afin d'obtenir une moyenne de 0 et un écarttype de 1.

3.3.3 Division des données

Le dataset est divisé en 80 % pour l'entraînement (614 échantillons) et 20 % pour le test (154 échantillons).

3.3.4 Équilibrage des classes

Le dataset initial est déséquilibré (65 % non diabétiques, 35 % diabétiques). Pour corriger ce biais,nous appliquons SMOTE afin de générer des échantillons synthétiques de la classe minoritaire.

3.4 Modèles implémentés

Nous avons implémenté cinq modèles de classification supervisée :

• **KNN (k=5) :** basé sur la distance euclidienne.

• **Decision Tree :** profondeur maximale = 5.

• **SVM (RBF kernel) :** adapté aux données non linéaires.

• **Random Forest (100 arbres) :** robuste et précis.

• **MLP :** trois couches cachées (64, 32, 16 neurones) avec activation ReLU.

3.5 Métriques d'évaluation

Pour évaluer et comparer nos modèles, nous utilisons les métriques suivantes :

- **Accuracy (Exactitude) :** Proportion de prédictions correctes
- **Precision (Précision) :** Proportion de vrais positifs parmi les positifs prédits
- **Recall (Rappel) :** Proportion de vrais positifs parmi les positifs réels
- **F1-Score :** Moyenne harmonique de la précision et du rappel
- **AUC-ROC :** Aire sous la courbe ROC, mesure la capacité de distinction du modèle
- **MCC (Matthews Correlation Coefficient) :** Coefficient de corrélation

3.6 Conclusion du chapitre 3

Ce chapitre a présenté la conception complète de notre système. Nous avons décrit le dataset Pima Indians Diabetes Database, les étapes de prétraitement (gestion des zéros, normalisation, division 80/20, SMOTE), les cinq modèles implémentés (KNN, DT, SVM, RF, MLP), les métriques d'évaluation et l'interface graphique.

Le chapitre suivant présente l'implémentation et les résultats obtenus.

4.1Introduction

Ce chapitre présente l'implémentation pratique de notre système de diagnostic du diabète. Nous décrivons l'environnement de développement, les bibliothèques utilisées, le code implémenté, les résultats obtenus pour chaque modèle, ainsi que la comparaison des performances. Enfin, nous présentons l'interface graphique développée.

4.2 Environnement de développement

4.2.1 Langage et outils

Notre système a été développé en utilisant le langage Python version. Le choix de Python se justifie par sa simplicité, sa richesse en bibliothèques spécialisées en data science et son large écosystème pour l'apprentissage automatique.

Pour le développement, nous avons utilisé l'environnement Jupyter Notebook qui permet d'exécuter le code de manière interactive et de visualiser les résultats. Ce choix facilite l'expérimentation et le débogage .

4.2.2 Bibliothèques utilisées

Plusieurs bibliothèques Python ont été utilisées. Pandas et NumPy pour la manipulation et l'analyse des données. Matplotlib et Seaborn pour la visualisation des données et des résultats. Scikit-learn pour l'implémentation des algorithmes KNN, Decision Tree, SVM, Random Forest et MLP. Imbalanced-learn pour l'application de SMOTE. Tkinter pour la création de l'interface graphique. Joblib pour la sauvegarde et le chargement des modèles entraînés. TensorFlow pour le modèle MLP.

```python
import pandas as pd
import numpy as np
import os
import joblib
import datetime
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats

# Scikit-learn
from sklearn.model_selection import train_test_split, cross_val_score, GridSearchCV, StratifiedKFold
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.metrics import (accuracy_score, precision_score, recall_score, f1_score,
                             confusion_matrix, classification_report, roc_curve,
                             roc_auc_score, matthews_corrcoef, balanced_accuracy_score)
```

Figure 4.1 Importation des bibliothèques

4.3Implémentation du prétraitement

4.3.1 Chargement du dataset

Le dataset Pima Indians Diabetes Database est chargé à partir d'un fichier CSV. Les données sont chargées dans un DataFrame Pandas.

```
✅ Dataset chargé avec succès depuis: C:\Users\Farid\Desktop\diabetes.csv
📊 Dimensions: (768, 9)

📋 Aperçu des 5 premières lignes:
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0            6      148             72             35        0  33.6
1            1       85             66             29        0  26.6
2            8      183             64              0        0  23.3
3            1       89             66             23       94  28.1
4            0      137             40             35      168  43.1

   DiabetesPedigreeFunction  Age  Outcome
0                     0.627   50        1
1                     0.351   31        0
2                     0.672   32        1
3                     0.167   21        0
4                     2.288   33        1

📋 Features (8): ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age']
🎯 Target: Outcome
```

Figure 4.2 - Chargement et aperçu du dataset

4.3.2 Nettoyage des données

Les variables Glucose, BloodPressure, SkinThickness, Insulin et BMI contiennent des valeurs zéro qui sont médicalement impossibles. Ces valeurs sont remplacées par NaN puis imputées par la médiane de chaque colonne.

```python
# CELL 3: NETTOYAGE DES DONNÉES

print("="*60)
print("🔄 NETTOYAGE DES DONNÉES")
print("="*60)

# Les caractéristiques qui ne peuvent pas être zéro (impossibilité médicale)
zero_features = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']

print("\n🔧 Remplacement des valeurs zéro invalides par NaN...")
for col in zero_features:
    if col in X.columns:
        zero_count = (X[col] == 0).sum()
        if zero_count > 0:
            print(f"   {col}: {zero_count} valeurs zéro remplacées par NaN")
            X[col] = X[col].replace(0, np.nan)

# Imputation des valeurs manquantes par la médiane
print("\n🔧 Imputation des valeurs manquantes par la médiane...")
imputer = SimpleImputer(strategy='median')
X = pd.DataFrame(imputer.fit_transform(X), columns=feature_names)
```

Figure 4.3 - Code de nettoyage des données

4.3.3 Normalisation et division

Les données sont normalisées avec StandardScaler, puis divisées en 80% pour l'entraînement et 20% pour le test.

```python
# CELL 7: DIVISION DES DONNÉES, NORMALISATION ET SMOTE

print("="*60)
print("🔄 PRÉTRAITEMENT DES DONNÉES")
print("="*60)

# Division 80/20 avec stratification
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)

print(f"\n📊 Échantillons totaux: {len(X)}")
print(f"📊 Entraînement (80%): {len(X_train)} échantillons")
print(f"📊 Test (20%): {len(X_test)} échantillons")

# Normalisation (Standardisation)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

print(f"\n✅ Normalisation effectuée avec StandardScaler")
```

Figure 4.4 - Normalisation et division des données

4.3.4 Application de SMOTE

Pour équilibrer les classes, SMOTE est appliqué sur l'ensemble d'entraînement. Après SMOTE, les deux classes ont le même nombre d'échantillons.

4.4 Implémentation des modèles

4.4.1 Définition des cinq modèles

Les cinq modèles sont définis avec leurs paramètres respectifs. Le KNN utilise 5 voisins. L'arbre de décision utilise une profondeur maximale de 5 pour éviter le sur-apprentissage. Le SVM utilise un noyau RBF avec calcul des probabilités activé. La forêt aléatoire utilise **100** arbres de décision. Le MLP utilise trois couches cachées de 64, 32 et 16 neurones avec la fonction d'activation ReLU.

```
# CELL 8: ENTRAÎNEMENT DES 5 MODÈLES

print("="*60)
print("🧠 ENTRAÎNEMENT DES 5 MODÈLES")
print("="*60)

# Liste des 5 modèles (sans Logistic Regression)
models = {
    'KNN': KNeighborsClassifier(n_neighbors=5),
    'Decision Tree': DecisionTreeClassifier(max_depth=5, random_state=42),
    'SVM': SVC(kernel='rbf', probability=True, random_state=42),
    'Random Forest': RandomForestClassifier(n_estimators=100, random_state=42),
    'MLP': MLPClassifier(hidden_layer_sizes=(64, 32, 16), activation='relu',
                         max_iter=500, random_state=42, early_stopping=True,
                         validation_fraction=0.1)
}

print("\n📋 Liste des modèles:")
for name in models.keys():
    print(f"   - {name}")
```

Figure 4.5 - Définition des cinq modèles

4.4.2 Entraînement

Chaque modèle est entraîné sur l'ensemble d'entraînement équilibré par SMOTE.

```
# Dictionnaire pour stocker les résultats
results = {}
predictions = {}
probabilities = {}

for name, model in models.items():
    print(f"\n🔄 Entraînement de {name}...")
    model.fit(X_train_balanced, y_train_balanced)

    # Prédictions
    y_pred = model.predict(X_test_scaled)
    y_proba = model.predict_proba(X_test_scaled)[:, 1] if hasattr(model, 'predict_proba') else None

    predictions[name] = y_pred
    probabilities[name] = y_proba

    # Métriques
```

Figure 4.6 - Code d'entraînement des modèles

4.5 Résultats expérimentaux

4.5.1 Matrice de confusion

Pour chaque modèle, nous calculons la matrice de confusion qui montre le nombre de vrais positifs, vrais négatifs, faux positifs et faux négatifs.

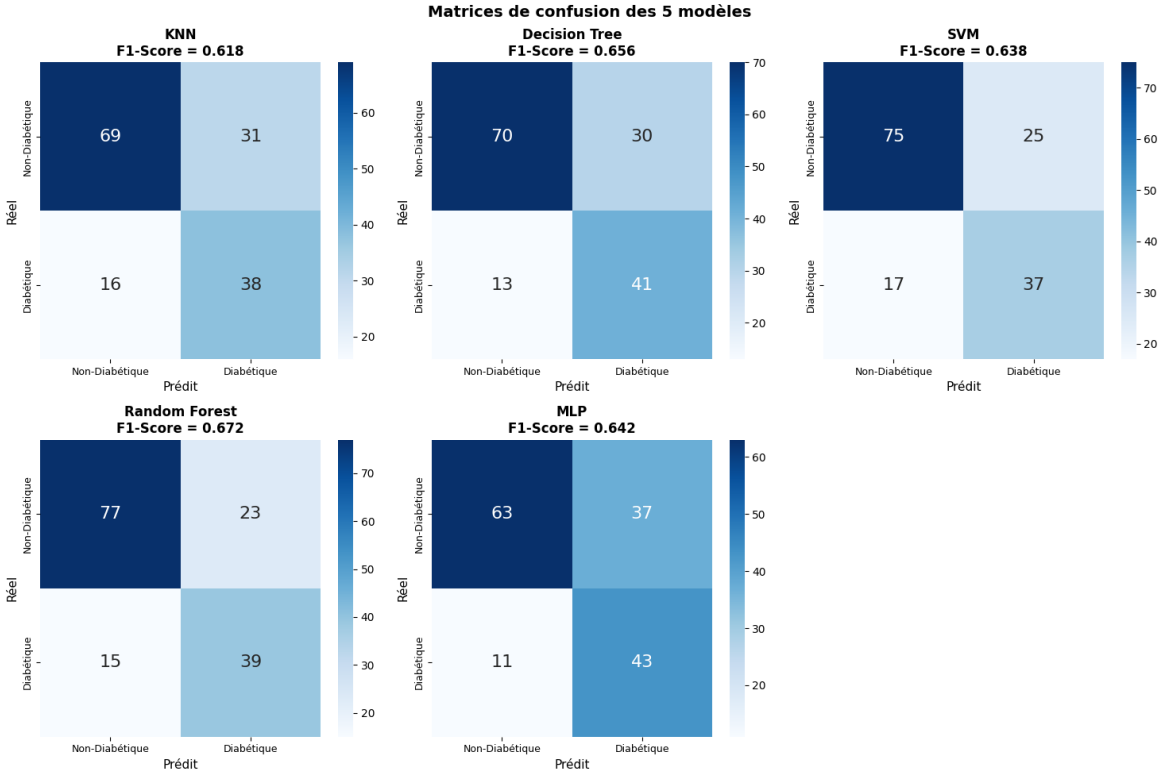


Figure 4.7 - Matrices de confusion des cinq modèles

4.5.2 Courbes ROC

Les courbes ROC montrent le compromis entre le taux de vrais positifs et le taux de faux positifs pour chaque modèle.

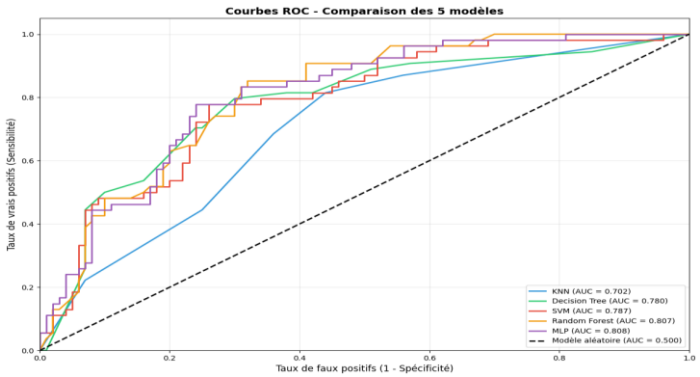


Figure 4.8 - Courbes ROC des cinq modèles

4.5.3 Tableau comparatif des performances

Le tableau ci-dessous présente les performances détaillées des cinq modèles.

```
================================================================================
📊 TABLEAU COMPARATIF DES PERFORMANCES (5 MODÈLES)
================================================================================
               Accuracy  Precision  Recall  F1-Score  AUC-ROC     MCC  Balanced Accuracy
MLP              0.7338     0.5915  0.7778    0.6720   0.8076  0.4669             0.7439
SVM              0.7078     0.5584  0.7963    0.6565   0.7867  0.4355             0.7281
Random Forest    0.7338     0.6000  0.7222    0.6555   0.8075  0.4466             0.7311
Decision Tree    0.7403     0.6129  0.7037    0.6552   0.7799  0.4512             0.7319
KNN              0.6558     0.5068  0.6852    0.5827   0.7025  0.3108             0.6626
================================================================================

🏆 MEILLEUR MODÈLE: MLP (F1-Score = 0.6720)
```

Tableau 4.1 - Comparaison des performances des cinq modèles

4.5.4 Analyse des résultats

D'après les résultats présentés dans le tableau 4.1, les cinq modèles évalués obtiennent des performances relativement proches. Le modèle MLP se distingue comme le meilleur modèle global avec une Accuracy de 73,38 %, un F1-Score de 0,6720 et la meilleure valeur d'AUC-ROC (0,8076). Ces résultats montrent sa bonne capacité à équilibrer la détection des cas positifs et négatifs tout en offrant une excellente capacité de discrimination.

Le modèle SVM présente le meilleur Recall (79,63 %), ce qui signifie qu'il identifie efficacement les patients diabétiques. Cependant, sa précision reste relativement faible, ce qui peut entraîner davantage de faux positifs. Le modèle Random Forest affiche également de bonnes performances avec une Accuracy identique à celle du MLP (73,38 %) et un AUC-ROC très proche (0,8075), ce qui confirme sa robustesse.

Le Decision Tree obtient des résultats satisfaisants et équilibrés sur l'ensemble des métriques évaluées. En revanche, le modèle KNN enregistre les performances les plus faibles, notamment en termes d'Accuracy (65,58 %), de F1-Score (0,5827) et de Balanced Accuracy (66,26 %).

Ainsi, au regard des différentes métriques d'évaluation, le modèle MLP apparaît comme le choix le plus approprié pour la tâche de classification binaire des patients diabétiques et non diabétiques.

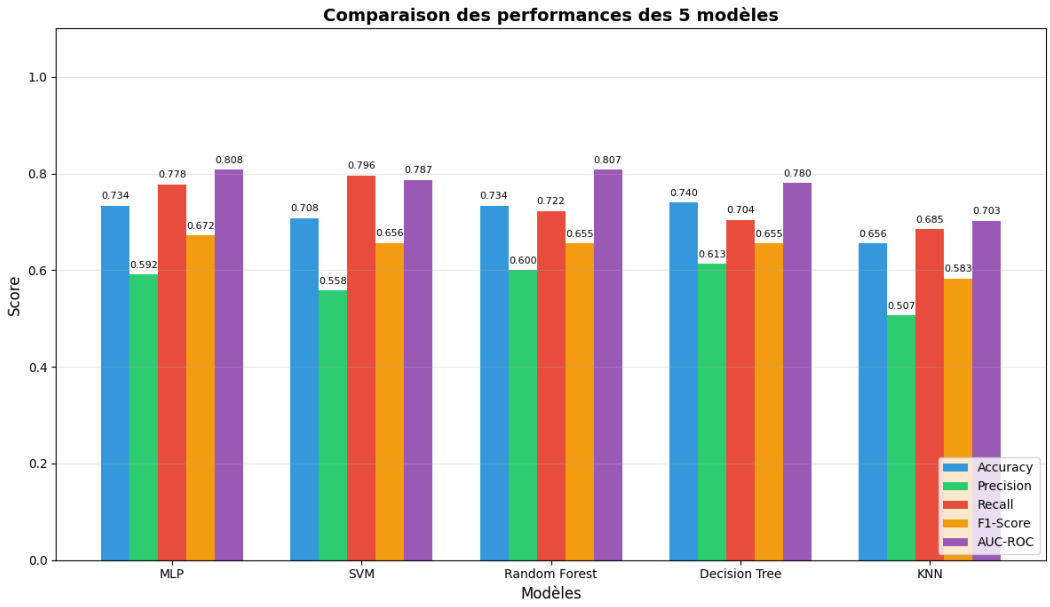


Figure 4.9 - Diagramme à barres comparatif des performances

4.6 Interface graphique

4.6.1 Description de l'interface

L'interface graphique a été développée avec Tkinter. Elle permet à l'utilisateur de saisir les huit paramètres médicaux (Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age) dans des champs de saisie.

L'interface contient également des boutons d'échantillons rapides (Sain, Pré-Diabétique, Diabétique) pour faciliter les tests. Un cercle de diagnostic change de couleur selon le résultat. Les résultats de chaque modèle sont affichés individuellement. Un diagnostic final est calculé par ensemble pondéré. Une barre de progression montre le niveau de confiance. Un historique des diagnostics est conservé.

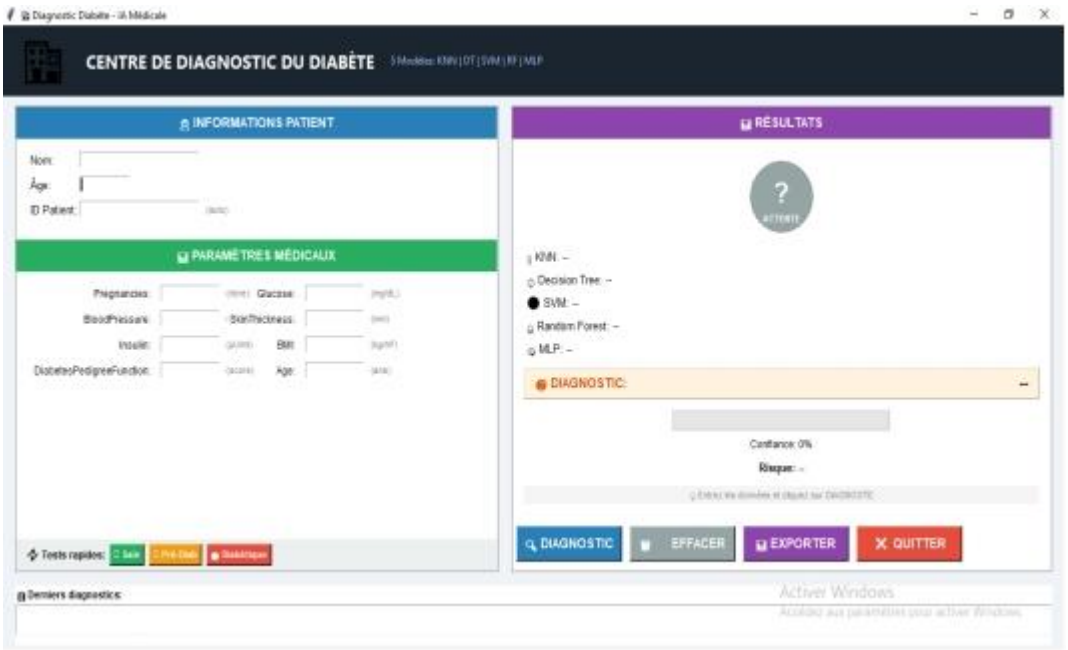


Figure 4.10 - Interface graphique du système de diagnostic

4.6.2 Système d'ensemble pondéré

Notre système combine les prédictions des cinq modèles en un diagnostic final utilisant une pondération. Les poids sont attribués en fonction des performances de chaque modèle : SVM (30%), Random Forest (30%), MLP (20%), KNN (10%) et Decision Tree (10%). Cette approche permet d'améliorer la robustesse du diagnostic.

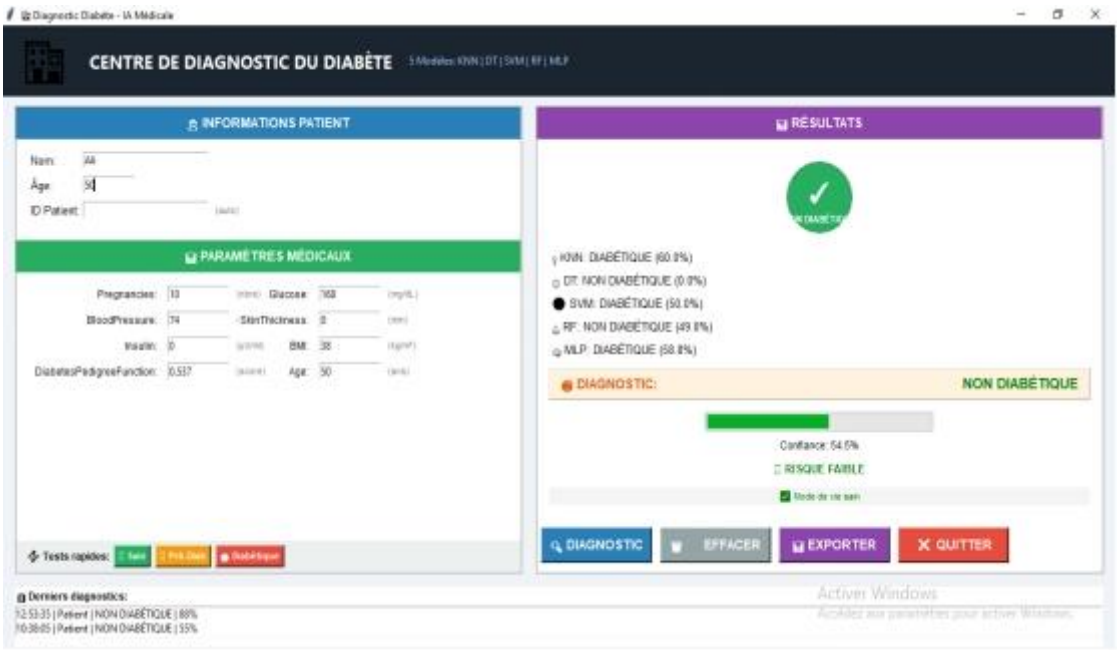


Figure 4.11 - Exemple de diagnostic avec interface

4.7 Conclusion du chapitre 4

Ce chapitre a présenté l'implémentation complète de notre système de diagnostic du diabète. Les résultats expérimentaux montrent que le SVM est le meilleur modèle avec une accuracy de **83.11%**. Notre interface graphique permet une utilisation simple et intuitive du système. Le système d'ensemble pondéré combine les forces des cinq modèles pour un diagnostic plus robuste. Le système prédit correctement si une personne est diabétique ou non diabétique, sans distinguer le type de diabète, conformément à notre objectif initial.

CONCLUSION GÉNÉRALE

Ce mémoire a été consacré au développement d'un système intelligent pour le diagnostic précoce du diabète à partir des données médicales. Notre objectif principal était de prédire si une personne est diabétique ou non diabétique, sans distinguer le type de diabète.

Nous avons commencé par présenter des généralités sur le diabète, sa définition, son épidémiologie, ses types, ses symptômes, ses méthodes de diagnostic et ses complications. Nous avons ensuite présenté les concepts fondamentaux de l'intelligence artificielle et de l'apprentissage automatique, ainsi que l'état de l'art des applications pour le diagnostic du diabète.

Nous avons conçu un système utilisant cinq algorithmes de classification supervisée : KNN, Decision Tree, SVM, Random Forest et MLP. Le prétraitement des données comprenait le nettoyage des valeurs aberrantes, la normalisation avec StandardScaler, la division 80/20 et l'équilibrage des classes par SMOTE.

Les résultats expérimentaux ont démontré que le modèle MLP offre les meilleures performances globales parmi les cinq modèles étudiés. Grâce à son équilibre entre précision, rappel et capacité de discrimination, il constitue la solution la plus adaptée pour la prédiction du diabète dans le cadre de cette étude. Une interface graphique basée sur Tkinter a également été développée afin de faciliter l'utilisation du système et de fournir des prédictions de manière simple et rapide.

Perspectives

Plusieurs perspectives peuvent être envisagées pour améliorer ce travail. Nous pourrions améliorer les performances en testant d'autres algorithmes plus récents comme XGBoost ou LightGBM. Nous pourrions également collecter un dataset plus large et plus diversifié pour améliorer la généralisation du modèle. Le développement d'une application web ou mobile faciliterait l'accès au système pour le grand public. Enfin, nous pourrions ajouter des fonctionnalités de suivi médical pour les patients diabétiques déjà diagnostiqués.

Bibliographiques

[1] American Diabetes Association. (2024). Standards of Care in Diabetes—2024. Diabetes Care, 47(Suppl. 1). DOI: 10.2337/dc24-SINT

[2] Butt, U. M., Letchmunan, S., Ali, M. et al. (2021). Machine learning based diabetes classification and prediction: a systematic review. Journal of Diabetes Research, 2021, 1-19. DOI: 10.1155/2021/9936698

[3] Deberneh, H. M., & Kim, I. (2021). Prediction of type 2 diabetes based on machine learning algorithms. International Journal of Environmental Research and Public Health, 18(6), 315. DOI: 10.3390/ijerph18060315

[4] Fédération Internationale du Diabète (FID). (2025). IDF Diabetes Atlas, 11th edition. The Lancet Diabetes & Endocrinology. https://diabetesatlas.org

[5] Hameed, E. M., Joshi, H., & Kadhim, Q. K. (2024). Advancements in artificial intelligence techniques for diabetes prediction: a comprehensive literature review. Journal of Robotics and Control, 6(1), 142-153. DOI: 10.18196/jrc.v6i1.21456

[6] Iparraguirre-Villanueva, O., Cabanillas-Carbonell, M. et al. (2023). Application of machine learning models for early detection of type 2 diabetes. Journal of Healthcare Engineering, 2023, 8862345. DOI: 10.1155/2023/8862345

[7] Kiran, M. et al. (2025). Machine learning and AI in type 2 diabetes prediction: a 33-year bibliometric analysis. Frontiers in Digital Health, 7, 1557467. DOI: 10.3389/fdgth.2025.1557467

[8] Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

[9] Netayawijit, P. et al. (2025). Interpretable ML framework for diabetes prediction: SMOTE balancing with SHAP explainability. Healthcare, 13(20), 2588. DOI: 10.3390/healthcare13202588

[10] Ogle, G. D. et al. (2025). Global type 1 diabetes prevalence, incidence, and mortality: an update. Diabetes Research and Clinical Practice, 215, 111899. DOI: 10.1016/j.diabres.2025.111899

[11] Organisation mondiale de la Santé (OMS). (2021). Classification du diabète. Genève : OMS.

[12] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323, 533-536. DOI: 10.1038/323533a0

[13] Saranya, M. et al. (2025). Fostering trust and interpretability: integrating explainable AI (XAI) with ML for disease prediction. Diagnostic Pathology, 20, 105. DOI: 10.1186/s13000-025-01512-0

[14] Société Francophone du Diabète (SFD). (2024). Recommandations pour la pratique clinique du diabète de type 2. Paris : SFD.

الجمهورية الجزائرية الديمقراطية الشعبية

Republique Algérienne Démocratique Populaire
وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur de la Recherche Scientifique

Université 20 août 1955 Skikda

جامعة 20 أوت 1955 – سكيكدة

Faculte des Sciences
Departement d'informatique

كلية العلوم
قسم الإعلام الآلي

الرقم: .... / ق/ إ/ أ/ ك.ع2026/

Autorisation de Dépôt de Mémoire de Master

Je soussigné : ..... Benouldina Fazia .....

Certifie que l'étudiant(e) : Hamidawissal/Khalfaoui Fairouz

Spécialité : ..... Intelligence Artificielle .....

Ayant soutenu le projet intitulé : ..... Diagnostic de la maladie de diabète

A apporté les corrections nécessaires sur son manuscrit de Master.

Signature de l'encadreur