الجمهورية الجزائرية الديمقراطية الشعبية

**République Algérienne Démocratique et Populaire**

وزارة التعليم العالي و البحث العلمي

**Ministère de l'Enseignement Supérieur et de la Recherche Scientifique**

**N°Réf:……**

**Centre Universitaire Abd**

**Elhafid Boussouf Mila**

**Institut des Sciences et Technologie** **Département de Mathématiques et Informatique**

# Mémoire préparé en vue de l'obtention du diplôme de Master

En : **Informatique**

**Spécialité : Intelligence Artificiel et ses Applications (I2A)**

# *Thème*

## *Développement d'un modèle de prédiction et de diagnostic médical assisté par l'intelligence artificielle*

**Préparé par:**

**Mlle : BENDJEDDOU Meroua**

**Mlle : BOUKHECHE Marwa**

**Soutenue devant le jury:**

| | | |
|---|---|---|
| **Présidente** | **KHALFI Souheila** | **Grade MCA** |
| **Examinateur** | **ATTIA Mourad** | **Grade MAA** |
| **Encadrant** | **HADJI Atmane** | **Grade MAA** |

**Année Universitaire: 2024/2025**

# *Remerciements*

Tout d'abord, nous remercions Allah de nous avoir donné le courage, la santé, la force et la patience nécessaires pour mener à bien ce travail.

Nous sommes profondément reconnaissants à notre directeur de mémoire, Dr Hadji A. pour ses précieux conseils, son soutien et sa disponibilité tout au long de ce projet.

Enfin, nos familles méritent toute notre gratitude et nous tenons à leur exprimer notre profonde reconnaissance pour leur soutien indéfectible et leurs encouragements sans faille tout au long de nos études. Leur amour, leur patience et leurs sacrifices ont été une source constante de motivation pendant les moments difficiles, et nous n'aurions pas pu y arriver sans leur soutien.

Merci à tous ceux qui ont participé à ce long voyage. Vos contributions, aussi modestes soient-elles, ont été profondément appréciées et ont eu un impact significatif sur notre travail.

**BENDJEDDOU MEROUA****BOUKHECHE MARWA**

## Résumé

Le diabète constitue aujourd'hui un enjeu majeur de santé publique à l'échelle mondiale, affectant des centaines de millions de personnes. Une détection précoce et efficace est essentielle pour limiter les complications graves liées à cette maladie chronique. Dans ce contexte, l'intelligence artificielle, notamment l'apprentissage automatique et l'apprentissage profond, offre des solutions prometteuses pour prédire le risque de diabète à partir de données médicales.

Ce mémoire vise à concevoir, implémenter et comparer différents modèles intelligents pour la prédiction du diabète à l'aide de caractéristiques cliniques simples. Plusieurs approches ont été explorées, allant des algorithmes supervisés tels que Random Forest et Extra Trees, aux modèles de gradient boosting comme XGBoost, LightGBM et CatBoost, jusqu'aux réseaux de neurones, notamment le perceptron multicouche (MLP) et un réseau profond. Des techniques d'ensemble telles que le stacking, le voting et l'agrégation pondérée ont été utilisées pour améliorer les performances. Le traitement des données déséquilibrées, l'optimisation des hyperparamètres et l'explicabilité des modèles font également partie des contributions de ce travail.

Trois jeux de données (PIMA, Frankfurt et un jeu combiné) ont été utilisés pour entraîner et évaluer les modèles. Les résultats ont été analysés à l'aide de plusieurs métriques (accuracy, précision, rappel, F1-score, AUC-ROC) et visualisés sous forme de matrices de confusion et courbes ROC. Une application web interactive a été développée pour permettre aux professionnels de santé de prédire le risque de diabète de manière simple et rapide, avec génération automatique d'un rapport explicatif.

Ce travail illustre ainsi l'apport concret de l'IA à la médecine prédictive et souligne le potentiel des modèles intelligents dans l'amélioration du diagnostic précoce des maladies chroniques.



**Mots-clés** : Diabète, Intelligence Artificielle, Apprentissage Automatique, Apprentissage Profond, Prédiction, Stacking, Application Web, Médecine Prédictive.

## Abstract

Diabetes is currently a major global public health issue, affecting hundreds of millions of people. Early and effective detection is essential to reduce the serious complications associated with this chronic disease. In this context, artificial intelligence (AI), particularly machine learning and deep learning, offers promising solutions to predict the risk of diabetes based on medical data.

This thesis aims to design, implement, and compare various intelligent models for diabetes prediction using simple clinical features. Several approaches have been explored, ranging from supervised algorithms such as Random Forest and Extra Trees, to gradient boosting models like XGBoost, LightGBM, and CatBoost, as well as neural networks including the Multilayer Perceptron (MLP) and a deep neural network. Ensemble techniques such as stacking, voting, and weighted averaging have been used to improve performance. Handling imbalanced data, hyperparameter optimization, and model explainability are also among the contributions of this work.

Three datasets (PIMA, Frankfurt, and a combined dataset) were used to train and evaluate the models. The results were analyzed using various metrics (accuracy, precision, recall, F1-score, AUC-ROC) and visualized through confusion matrices and ROC curves. An interactive web application was developed to enable healthcare professionals to easily and quickly predict the risk of diabetes, with automatic generation of an explanatory report.

This work thus illustrates the concrete contribution of AI to predictive medicine and highlights the potential of intelligent models in improving early diagnosis of chronic diseases.

**Keywords**: Diabetes, Artificial Intelligence, Machine Learning, Deep Learning, Prediction, Stacking, Web Application, Predictive Medicine.

**الملخص**

يعد مرض السكري حالياً مشكلة صحية عامة عالمية كبيرة، حيث يصيب مئات الملايين من الأشخاص. ويعد الكشف المبكر والفعال عن هذا المرض أمراً ضرورياً للحد من المضاعفات الخطيرة المرتبطة بهذه الحالة المزمنة. وفي هذا السياق، يقدم الذكاء الاصطناعي (AI)، ولا سيما التعلم الآلي والتعلم العميق، حلولاً واعدة للتنبؤ بخطر الإصابة بمرض السكري استناداً إلى البيانات الطبية.

تهدف هذه الأطروحة إلى تصميم وتنفيذ ومقارنة نماذج ذكية مختلفة للتنبؤ بمرض السكري باستخدام ميزات سريرية بسيطة. تم استكشاف عدة مناهج، بدءًا من الخوارزميات الخاضعة للإشراف مثل Random Forest و Extra Trees، إلى نماذج تعزيز التدرج مثل XGBoost و LightGBM و CatBoost، بالإضافة إلى الشبكات العصبية بما في ذلك Multilayer Perceptron (MLP) والشبكة العصبية العميقة. تم استخدام تقنيات التجميع مثل التكديس والتصويت والمتوسط المرجح لتحسين الأداء. كما أن التعامل مع البيانات غير المتوازنة، وتحسين المعلمات الفائقة، وقابلية تفسير النموذج هي أيضًا من بين مساهمات هذا العمل.

تم استخدام ثلاث مجموعات بيانات (PIMA و Frankfurt ومجموعة بيانات مجمعة) لتدريب النماذج وتقييمها. تم تحليل النتائج باستخدام مقاييس مختلفة (الدقة، والإحكام، والاسترجاع، و F1-score، و AUC-ROC) وتم تصورها من خلال مصفوفات الارتباك ومنحنيات ROC. تم تطوير تطبيق ويب تفاعلي لتمكين المهنيين في مجال الرعاية الصحية من التنبؤ بسهولة وسرعة بخطر الإصابة بمرض السكري، مع إنشاء تقرير توضيحي تلقائيًا.

وبالتالي، يوضح هذا العمل المساهمة الملموسة للذكاء الاصطناعي في الطب التنبئي ويسلط الضوء على إمكانات النماذج الذكية في تحسين التشخيص المبكر للأمراض المزمنة.

الكلمات المفتاحية: السكري، الذكاء الاصطناعي، التعلم الآلي، التعلم العميق، التنبؤ، التكديس، تطبيق الويب، الطب التنبئي.

# Table des Matières

# Liste des Figures

# Liste des Tableaux

# Liste des sigles et acronymes

| | |
|---|---|
| CatBoost | Categorical Boosting |
| CNN | Convolutional Neural Network |
| DL | Deep Learning |
| DNN | Deep Neural Network |
| IA | Intelligence Artificielle |
| IAE | Intelligence Artificielle Explicable |
| LightGBM | Light Gradient Boosting Machine |
| Lime | Local Interpretable Model-Agnostic Explanations |
| MLP | Multi Layer Perceptron |
| RNN | Recurrent Neural Network |
| Shap | SHapley Additive Explanations |
| SMOTE | Synthetic Minority Over-sampling Technique |
| SVM | Support Vectors Machine |
| XGBoost | eXtreme Gradient Boosting |

Le diabète est aujourd'hui l'une des maladies chroniques les plus répandues dans le monde, représentant un défi majeur de santé publique. Selon l'Organisation mondiale de la santé (OMS), des centaines de millions de personnes vivent avec le diabète, souvent sans le savoir, ce qui retarde leur prise en charge et accroît le risque de complications graves telles que les maladies cardiovasculaires, l'insuffisance rénale ou la cécité. Face à cette situation alarmante, la capacité à détecter précocement les cas à risque est devenue une priorité dans les politiques de prévention.

Dans ce contexte, l'intelligence artificielle (IA), et plus spécifiquement l'apprentissage automatique (Machine Learning) et l'apprentissage profond (Deep Learning), offrent des opportunités prometteuses pour améliorer le diagnostic et la prédiction du diabète. Ces modèles peuvent apprendre à identifier les profils de patients à risque et à fournir des prédictions précises, rapides et automatisées.

Ce mémoire de master s'inscrit dans cette dynamique et vise à développer et comparer différents modèles intelligents capables de prédire le risque de diabète à partir de caractéristiques médicales simples, accessibles dans un contexte clinique réel. À travers l'usage d'algorithmes d'apprentissage supervisé : Random Forest, Extra Trees, des algorithmes de gradient boosting : XGBoost, LightGBM et CatBoost, des réseaux de neurones : MLP et un réseau neuronal profond, et en utilisant des techniques avancées telles que le stacking, le voting et l'ensemble pondéré, notre objectif est d'explorer l'apport de ces technologies à la médecine prédictive, tout en intégrant des techniques modernes de gestion des données déséquilibrées, d'optimisation des hyperparamètres et d'explicabilité des modèles.

La structure de ce mémoire est organisée comme suit :

Dans le premier chapitre, nous introduisons les fondements médicaux du diabète : sa définition, ses différents types, ses complications et ses facteurs de risque, ainsi que les principaux symptômes. Nous discutons également de l'émergence de l'intelligence artificielle en médecine, notamment dans la prédiction des maladies chroniques, en particulier du diabète.

Le deuxième chapitre est consacré à la revue des approches d'apprentissage automatique appliquées à la prédiction du diabète. Nous décrivons les méthodes supervisées, les approches avancées ainsi que les réseaux de neurones. Nous abordons les métriques d'évaluation (accuracy, précision, rappel, F1-score, AUC-ROC), etun tableau comparatif des travaux récents vient compléter ce chapitre.

Dans le troisième chapitre, nous présentons les algorithmes utilisés pour le développement de notre système de prédiction. Nous détaillons les jeux de données utilisés (PIMA, Frankfurt et une version combinée), ainsi que les étapes de prétraitement.

Le quatrième chapitre est dédié à l'implémentation des modèles, à leur évaluation et à la

comparaison de leurs performances. Nous y analysons les résultats obtenus à l'aide de différentes métriques et visualisations (matrices de confusion, courbes ROC), en comparant les performances entre les modèles classiques et avancés, ainsi qu'entre les différents jeux de données. Enfin, nous présentons une application web interactive que nous avons développée, permettant aux médecins d'entrer les données d'un patient, d'obtenir une prédiction personnalisée du risque de diabète et de générer un rapport PDF explicatif.

# Chapitre1:

## *Diabète et intelligence artificielle*

## 1.1. Introduction

Le diabète est une maladie chronique affectant la régulation de la glycémie, avec des formes variées et des complications graves en cas de retard de prise en charge. Sa prévention repose sur une détection précoce et une gestion personnalisée. L'intelligence artificielle (IA) offre des outils performants pour analyser des données médicales complexes et prédire le risque diabétique. Elle permet d'améliorer la précision diagnostique et l'efficacité des décisions cliniques. Ce chapitre présente successivement : les bases du diabète, les fondamentaux de l'IA en santé, les méthodes actuelles de prédiction par l'IA, ainsi que les avancées récentes telles que l'IA explicable, l'apprentissage profond, l'intégration multimodale, l'apprentissage fédéré et les dispositifs de surveillance en temps réel.

## 1.2. Présentation générale du diabète

### 1.2.1. Définition

Le diabète est une maladie chronique qui survient lorsque le pancréas ne produit pas suffisamment d'insuline, ou lorsque l'organisme n'utilise pas efficacement l'insuline produite. En l'absence de traitement adéquat, cette perturbation entraîne une hyperglycémie persistante, pouvant provoquer à long terme de graves complications [1].

## 1.3. Classification du diabète

La classification et le diagnostic du diabète sont complexes et ont fait l'objet de nombreux débats, consultations et révisions au fil des ans. Aujourd'hui, il est largement admis que le diabète comporte trois types principaux : le diabète de type 1, le diabète de type 2 et le diabète gestationnel (DG) [2].

### 1.3.1. Diabète de type 1

Le diabète de type 1 est une maladie auto-immune provoquant la destruction des cellules bêta du pancréas, responsables de la production d'insuline [3, 4]. Cette forme de diabète apparaît le plus souvent chez les enfants et les adolescents, mais peut également survenir à l'âge adulte. Des formes idiopathiques, plus courantes chez les personnes d'origine africaine ou asiatique, peuvent aussi être observées [5]. À long terme, un traitement insulinique est nécessaire pour tous les patients atteints de diabète de type 1 [6, 7].

### 1.3.2. Diabète de type 2

Le diabète de type 2 se caractérise par une résistance à l'insuline, associée initialement à une carence relative en sécrétion d'insuline [8]. Bien que les concentrations plasmatiques d'insuline soient souvent élevées, elles demeurent insuffisantes pour compenser le niveau de résistance à l'insuline, perturbant ainsi l'homéostasie glycémique [9, 10]. Au fil du temps, une détérioration progressive des cellules bêta pancréatiques surviennent, aggravant la déficience insulinique. Il a été démontré que les individus présentant à la fois une glycémie à jeun altérée et une intolérance au glucose ont déjà perdu environ 80 % de leur capacité de sécrétion pancréatique avant même que le diagnostic ne soit posé [11]. Cette figure 1.1 illustre le rôle de l'insuline dans la régulation de l'entrée du glucose dans les cellules et met en évidence les différences de fonctionnement selon l'état métabolique.

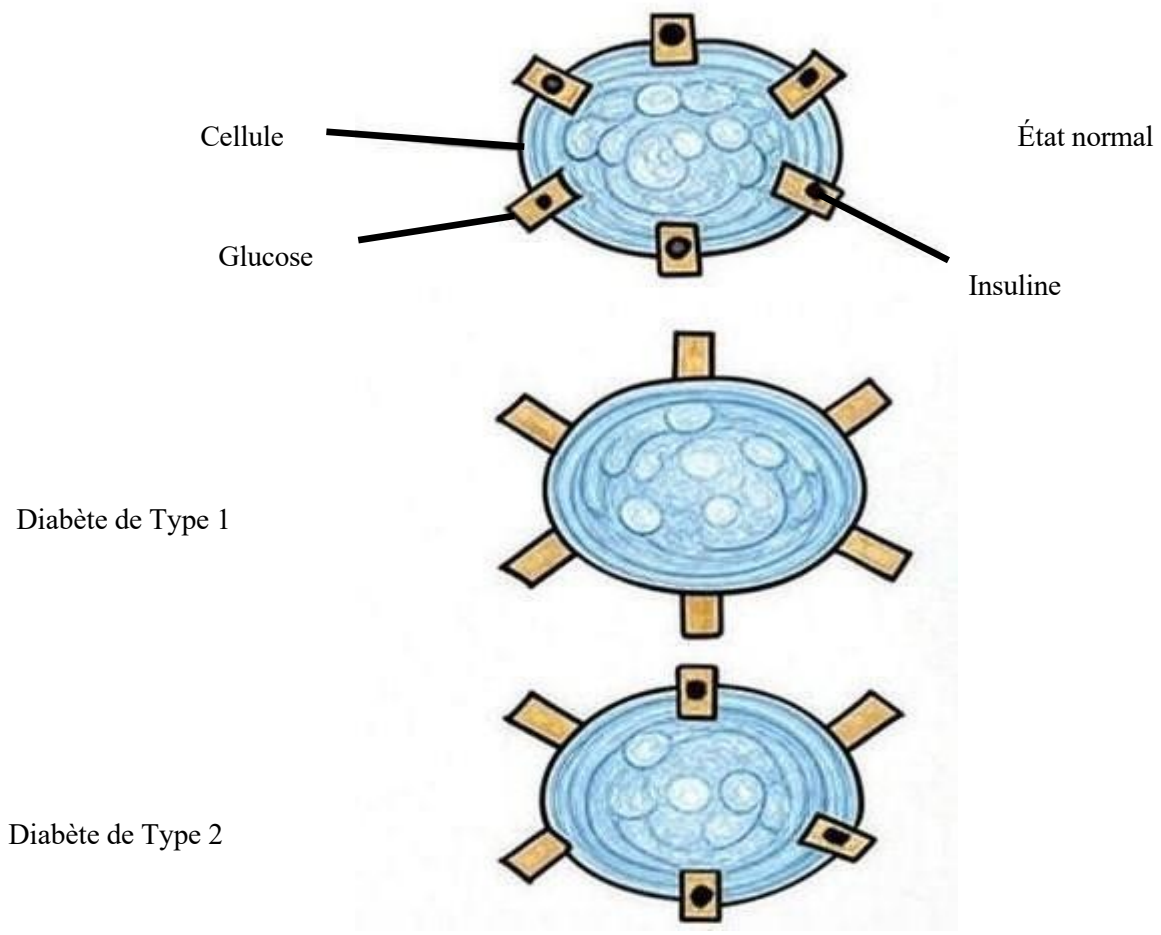


Figure 1.1: Fonctionnement de l'insuline

### 1.3.3. Diabète gestationnel

Le diabète gestationnel (GDM) est défini comme une intolérance au glucose diagnostiquée pour la première fois pendant la grossesse, le plus souvent au troisième trimestre [12]. Environ 8 à 9 % des grossesses sont compliquées par un GDM, ce taux pouvant doubler dans les populations à

haut risque de diabète de type 2. Il est recommandé de réaliser un test de tolérance au glucose (OGTT) environ six semaines après l'accouchement pour reclassifier l'état glycémique de la mère (diabète, intolérance au glucose, glycémie normale, etc.).

## 1.4. Complications du diabète

Un diabète mal contrôlé peut endommager presque tous les organes du corps, notamment lecœur, les vaisseaux sanguins, les reins, les yeux, le système nerveux, etc. L'hyperglycémieendommage au fil du temps les parois des minuscules artères sanguines qui apportent l'oxygèneet les nutriments à tous les tissus, ce qui affecte tous ces organes [13, 14].

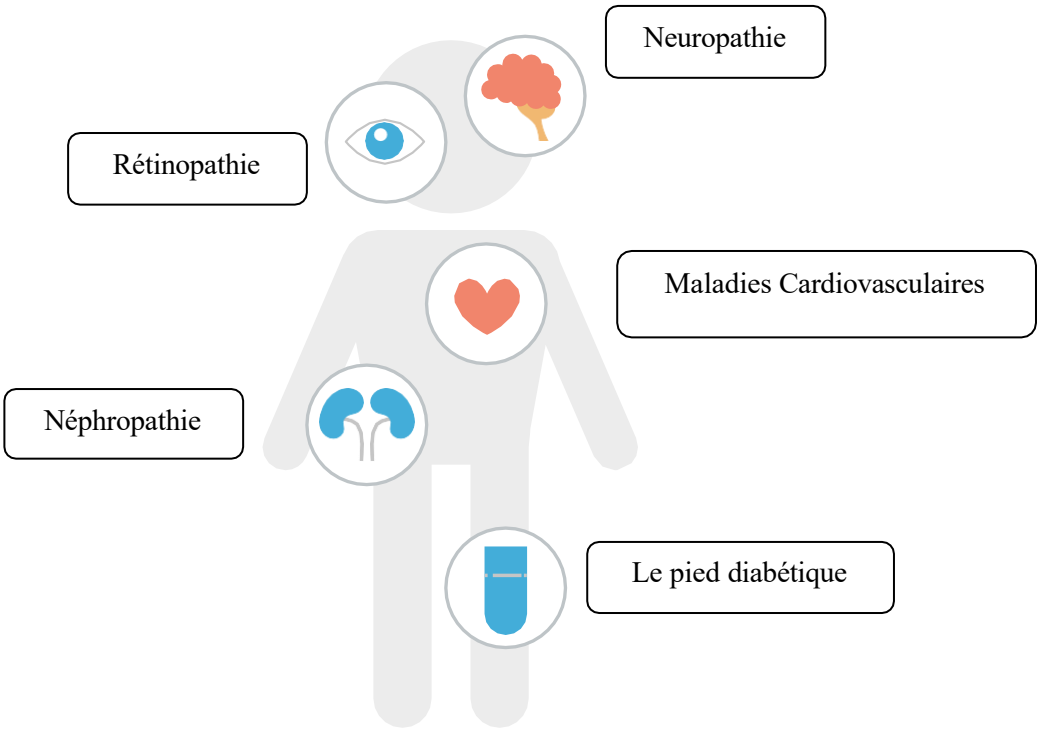


**Figure 1.2** : Principales complications de diabète

### 1.4.1. Maladies cardiovasculaires

Le diabète augmente considérablement le risque de maladies cardiovasculaires, les personnes diabétiques ayant 2 à 4 fois plus de risques d'en souffrir que la population générale. Une glycémie élevée favorise la coagulation et l'obstruction des vaisseaux, pouvant entraîner des crises cardiaques ou des AVC. Ce risque est renforcé par des facteurs comme l'âge, l'hérédité, l'hypertension, l'obésité ou le tabagisme. Les diabétiques de type 2 présentent souvent une prédisposition génétique, et environ un sur deux décède d'un accident cardiovasculaire.

### 1.4.2. Néphropathie

La néphropathie, terme issu du grec *nephros* signifiant rein, désigne une atteinte des reins souvent liée au diabète. Ce dernier endommage les petites artères rénales responsables de filtrer le

sang, ce qui peut entraîner une dégradation progressive de la fonction rénale, allant jusqu'à l'insuffisance rénale chronique ou permanente. L'hypertension est également un facteur aggravant de cette complication.

### 1.4.3. Troubles oculaires

Le diabète peut provoquer une dégradation progressive de la vision, allant jusqu'à la cataracte ou la cécité. Les troubles visuels sont parmi les complications les plus fréquentes, affectant presque tous les diabétiques de type 1 et environ 60 % de ceux de type 2. La rétine est généralement la zone la plus touchée, bien que d'autres parties de l'œil puissent aussi être affectées.

### 1.4.4. Neuropathie

La neuropathie désigne des troubles nerveux souvent douloureux, touchant 40 à 50 % des diabétiques de type 1 ou 2 dans les dix premières années de la maladie. Elle résulte d'une mauvaise circulation sanguine et d'un excès de glucose, qui endommagent les nerfs. Elle se manifeste par des picotements, une perte de sensation ou une gêne, débutant aux extrémités et progressant vers le haut. Elle peut également affecter les nerfs liés à la digestion, la pression artérielle et le rythme cardiaque.

### 1.4.5. Pied diabétique

Le pied diabétique est particulièrement vulnérable en raison de la neuropathie, d'une mauvaise circulation sanguine et d'un taux élevé de glycémie. Ces facteurs diminuent la capacité de cicatrisation et affaiblissent les défenses immunitaires, augmentant ainsi le risque d'infections et de complications.

## 1.5. Facteurs de risque et symptômes du diabète

Le développement du diabète résulte de l'interaction complexe entre des facteurs génétiques et environnementaux. Ces facteurs de risque peuvent être classés en deux grandes catégories : non modifiables et modifiables [15].

### 1.5.1. Facteurs de risque non modifiables

Ce sont des éléments sur lesquels il n'est pas possible d'agir, mais qui influencent significativement la probabilité de développer un diabète :

- **Âge** : Le risque de diabète augmente avec l'âge, en particulier après 45 ans, en raison d'une diminution progressive de la sensibilité à l'insuline et de la fonction pancréatique [14].
- **Antécédents familiaux** : Les personnes ayant un parent du premier degré atteint de diabète présentent un risque significativement accru de développer la maladie, ce qui suggère une forte composante héréditaire [15, 16].
- **Origine ethnique :** Certaines populations, notamment les personnes d'origine africaine, asiatique, hispanique et amérindienne, sont plus susceptibles de développer un diabète de

type 2, même à un IMC plus faible [15, 17].

### 1.5.2. Facteurs de risque modifiable

Ces facteurs sont liés au mode de vie ou à des conditions de santé sur lesquelles une intervention est possible :

- **Obésité et surcharge pondérale** : L'excès de masse grasse, en particulier abdominale, est l'un des principaux déterminants du diabète de type 2. Il favorise l'insulinorésistance et l'altération de la fonction des cellules bêta [18, 15].
- **Sédentarité** : Un faible niveau d'activité physique est associé à une diminution de la sensibilité à l'insuline. L'exercice régulier améliore la régulation glycémique et réduit significativement le risque de diabète [15].
- **Habitudes alimentaires** : Une alimentation riche en glucides raffinés, en acides gras saturés et pauvre en fibres augmente le risque de développer un diabète. À l'inverse, une alimentation équilibrée contribue à la prévention de la maladie [15].
- **Tabagisme** : Le tabac est un facteur de risque reconnu du diabète, car il altère la fonction des cellules bêta et augmente l'inflammation systémique [15].
- **Hypertension artérielle** : Une pression artérielle élevée est fréquemment associée à l'insulinorésistance et fait partie des composantes du syndrome métabolique, un précurseur fréquent du diabète [15].

### 1.5.3. Symptômes de diabète

Le diabète de type 1 et de type 2 peuvent avoir des symptômes similaires, mais il y a des différences [14].

### 1.5.4. Symptômes de type 1

Les symptômes du diabète de type 1 incluent une augmentation de la soif, des mictions fréquentes, une fatigue persistante, une perte de poids malgré une augmentation de l'appétit, ainsi qu'une vision floue. D'autres signes peuvent apparaître, tels que des blessures qui guérissent lentement, de l'irritabilité, des nausées et vomissements, une humeur changeante et des mictions nocturnes fréquentes.

### 1.5.5. Symptômes de type 2

Les symptômes du diabète de type 2 peuvent ressembler à ceux du type 1, notamment la soif accrue, la fréquence urinaire élevée, la fatigue et la vision floue. À cela s'ajoutent des infections fréquentes, une cicatrisation lente des plaies, des picotements ou engourdissements dans les mains ou les pieds, des douleurs ou sensations de brûlure dans les extrémités, ainsi que des démangeaisons fréquentes dans la région génitale.

Il est important de noter que certaines personnes atteintes de diabète de type 2 peuvent ne

présenter aucun symptôme au début de la maladie, ce qui souligne l'importance du dépistage régulier pour une détection précoce. Cela motive l'utilisation de systèmes de prédiction automatisés basés sur l'intelligence artificielle.

## 1.6. Importance de l'intelligence artificielle dans les applications médicales

L'intelligence artificielle (IA) est un domaine multidisciplinaire de l'informatique qui vise à concevoir des systèmes capables de simuler, voire de dépasser, certaines facultés cognitives humaines telles que l'apprentissage, le raisonnement, la perception et la prise de décision. Elle repose sur un ensemble de technologies telles que le traitement automatique du langage naturel, la reconnaissance vocale, la vision par ordinateur, la robotique, ainsi que l'apprentissage automatique (machine learning) et l'apprentissage profond (deep learning) [19]. On distingue généralement trois niveaux d'IA : l'IA faible, spécialisée dans des tâches précises ; l'IA forte, capable de reproduire une intelligence comparable à celle de l'humain ; et l'IA super-intelligente, encore hypothétique, censée dépasser largement les capacités humaines.

L'impact de l'IA est aujourd'hui considérable dans de nombreux secteurs, notamment celui de la santé, où elle joue un rôle central dans l'amélioration de la qualité des soins, l'accélération des diagnostics et l'optimisation des traitements [20]. En analysant automatiquement de vastes ensembles de données médicales telles que les images radiologiques, les signaux physiologiques ou les dossiers médicaux électroniques l'IA permet une prise de décision plus rapide et plus précise, souvent comparable, voire supérieure, à celle des experts humains. Elle soutient également la médecine personnalisée, en identifiant des modèles complexes associés aux profils génétiques ou aux réponses thérapeutiques des patients. Par ailleurs, elle facilite la détection précoce des maladies chroniques, améliore la gestion hospitalière et assure un suivi en temps réel des patients, participant ainsi à une médecine plus prédictive, préventive et efficiente.

### 1.6.1. Rôle de l'intelligence artificielle dans la médecine moderne

L'intégration de l'IA dans le domaine médical répond à des enjeux majeurs tels que l'amélioration des résultats cliniques, la réduction des coûts de santé et l'optimisation des ressources humaines. Face à la complexité croissante des pathologies et à la pression exercée sur les systèmes de santé, l'IA se présente comme une solution innovante pour renforcer la médecine préventive. Les technologies de dépistage précoce basées sur l'IA permettent par exemple d'identifier rapidement et à moindre coût les individus à risque, facilitant ainsi des interventions ciblées et efficaces [21–23].

Au cours de la dernière décennie, les progrès du machine learning (ML) et du deep learning (DL) ont profondément transformé l'intelligence artificielle médicale (AIM). Ces techniques ont permis la mise au point de modèles prédictifs avancés capables de diagnostiquer un large éventail

de maladies, d'évaluer la réponse aux traitements et de personnaliser les parcours de soins. Elles contribuent également à l'automatisation des flux de travail cliniques, au suivi de l'évolution des maladies, et à la prise de décisions thérapeutiques plus éclairées, entraînant ainsi une amélioration tangible de la qualité des soins [24–26].

### 1.6.2. Systèmes d'intelligence artificielle en santé

Les systèmes d'IA en santé exploitent la capacité des algorithmes à traiter et interpréter de grandes quantités de données en temps réel, apportant ainsi une aide précieuse à la décision clinique, au suivi des patients et à la personnalisation des soins [21]. Les domaines d'application couvrent plusieurs spécialités médicales telles que la radiologie, la génomique, l'analyse des dossiers médicaux électroniques, ainsi que l'épidémiologie et la santé publique [27, 28].

Cependant, malgré ces avancées, l'adoption de l'IA dans les systèmes de santé nécessite un encadrement rigoureux afin de garantir la protection des données sensibles des patients et de respecter les cadres éthiques et réglementaires en vigueur. Une utilisation responsable et transparente de ces technologies est indispensable pour assurer leur efficacité clinique, leur acceptabilité sociale et leur durabilité à long terme [26].

## 1.7. Méthodes actuelles de prédiction du diabète avec l'IA

La prédiction du diabète à l'aide de l'intelligence artificielle (IA) repose aujourd'hui sur une diversité de méthodes avancées tirant parti des technologies informatiques les plus récentes pour analyser la complexité de cette pathologie métabolique. L'IA, à travers l'utilisation d'algorithmes d'apprentissage automatique sophistiqués et de techniques poussées de fouille de données, permet d'extraire des connaissances précieuses et de mettre en évidence des patterns complexes à partir de vastes ensembles de données hétérogènes liées au diabète [26, 27].

Ces approches regroupent un large éventail de techniques innovantes, notamment les méthodes d'apprentissage en ensemble, les réseaux de neurones profonds, les machines à vecteurs de support, les réseaux bayésiens, ainsi que des modèles hybrides combinant plusieurs algorithmes afin d'optimiser la performance prédictive. L'intégration de données variées telles que les indicateurs cliniques, les facteurs génétiques, les habitudes de vie ou encore les données omiques permet de concevoir des modèles de prédiction complets, capables de refléter la nature multifactorielle de l'apparition et de l'évolution du diabète.

Par ailleurs, l'exploitation conjointe de sources de données hétérogènes à travers des techniques modernes d'intégration telles que la fusion de données, l'apprentissage multi-vue ou encore l'apprentissage par transfert renforce les capacités prédictives des modèles développés, en tirant parti de l'information complémentaire issue de chaque modalité [28-30].

## 1.8. Tendances actuelles de l'IA appliquée à la prédiction du diabète

### 1.8.1. IA explicable (IAE) dans la prédiction du diabète

L'intelligence artificielle explicable (ou Explainable AI, « XAI ») constitue un paradigme émergent visant à rendre compréhensibles les mécanismes de décision complexes des modèles d'intelligence artificielle, en particulier dans le cadre de la prédiction du diabète. Cette approche permet d'apporter une transparence accrue sur les facteurs influençant les prédictions, ce qui est crucial pour une évaluation fiable des risques et une interprétation clinique exploitable [31].

Dans ce contexte, les techniques d'IAE s'appuient sur des algorithmes avancés pour mettre en lumière les relations non triviales, les patterns sous-jacents et le poids relatif des variables dans les modèles prédictifs. Cela permet aux professionnels de santé, aux chercheurs et aux patients de mieux comprendre le raisonnement des systèmes d'IA. Cette transition des modèles "boîte noire" vers des systèmes interprétables apporte de multiples bénéfices, notamment l'identification de nouveaux biomarqueurs, la compréhension des mécanismes pathologiques, la détection de facteurs de risque modifiables et le soutien à la prise de décision clinique [32].

Les approches explicatives englobent une variété de méthodes, telles que les explications fondées sur des règles, l'analyse de l'importance des variables, les techniques d'interprétabilité locale et globale, ainsi que des méthodes dites agnostiques, applicables quel que soit le modèle utilisé. Par ailleurs, l'intégration des techniques d'IAE aux connaissances médicales et à l'expertise métier renforce la pertinence clinique et contextuelle des explications fournies, améliorant ainsi leur utilité pratique [31, 33].

Mais de grands défis persistent malgré les avancées de l'intelligence artificielle explicable (IAE) dans le domaine médical. Il s'agit notamment de trouver un équilibre entre la complexité des modèles et leur interprétabilité, tout en limitant les biais et en tenant compte des limites propres aux méthodes utilisées. La transparence des systèmes doit également être conciliée avec la protection des données personnelles. De plus, l'absence de critères d'évaluation standardisés rend difficile l'appréciation de la qualité des explications fournies. Ces enjeux nécessitent une collaboration étroite entre chercheurs, professionnels de santé et autorités réglementaires.

### 1.8.2. Méthodes d'apprentissage profond

Ces dernières années, le domaine de l'intelligence artificielle (IA) a connu des avancées remarquables et des tendances émergentes dans le domaine de la prédiction du diabète. Ces avancées ont eu un impact significatif sur le paysage de la recherche et présentent un grand potentiel pour révolutionner la gestion des maladies et améliorer les résultats de santé. Notamment, les méthodes d'apprentissage profond, telles que les réseaux neuronaux convolutifs (CNN) et les réseaux neuronaux récurrents (RNN), ont suscité une attention considérable et ont démontré leur

efficacité dans la capture de motifs complexes et de dépendances temporelles dans diverses modalités de données pertinentes pour la prédiction du diabète, telles que les images médicales, les séries temporelles et les dossiers de santé électroniques [34].

### 1.8.3. Intégration de sources de données multimodales

Une autre tendance importante est l'utilisation croissante de données multimodales, c'est-à-dire la combinaison de plusieurs types d'informations, comme les données cliniques, les informations génétiques, les habitudes de vie et les mesures physiologiques [35]. En regroupant ces différentes sources, les modèles d'IA deviennent plus complets et sont capables de mieux comprendre les causes et l'évolution du diabète. Cela permet d'obtenir des évaluations de risque plus précises et de proposer des interventions adaptées à chaque patient [36].

### 1.8.4. Apprentissage fédéré

De plus, l'utilisation de l'apprentissage fédéré (Federated Learning, FL) et des techniques de protection de la vie privée prend de plus en plus d'importance dans la prédiction du diabète. L'apprentissage fédéré permet à plusieurs établissements de collaborer pour entraîner des modèles d'IA sans avoir à partager directement les données des patients, ce qui garantit leur confidentialité[37]. Cette méthode est particulièrement utile dans le domaine de la santé, où la sécurité des données est essentielle. En combinant les connaissances issues de sources réparties, cette approche permet de créer des modèles de prédiction du diabète plus solides et applicables à différents contextes [38].

### 1.8.5. Surveillance en temps réel

On observe également une avancée vers la prédiction du diabète en temps réel et personnalisée, grâce à l'intégration des modèles d'IA avec des applications mobiles et des objets connectés. Cette évolution permet un suivi continu des paramètres de santé, facilitant les interventions rapides, la surveillance à distance et des soins adaptés à chaque individu. En combinant les données en temps réel avec l'intelligence artificielle, les personnes à risque ou atteintes de diabète peuvent mieux comprendre leur état et prendre des décisions éclairées pour gérer leur santé [39].

Ainsi, les récentes avancées et tendances en IA dans la prédiction du diabète montrent un fort potentiel pour transformer la gestion de cette maladie. L'intégration de l'apprentissage profond, des données multimodales, de l'IA explicable, des techniques de protection de la vie privée et de la surveillance en temps réel marque un changement important vers des modèles plus précis, interprétables et centrés sur le patient. Ces innovations offrent des perspectives prometteuses pour réduire l'impact du diabète, améliorer les résultats cliniques et la qualité de vie des personnes concernées [40].

## 1.9. Conclusion

Ce chapitre a permis de montrer les bases essentielles à une compréhension approfondie du diabète et des enjeux associés à sa prise en charge. Après avoir présenté la maladie, ses principales formes, ses complications, ses facteurs de risque et ses symptômes, nous avons introduit les concepts fondamentaux de l'intelligence artificielle (IA) et mis en évidence son rôle croissant dans le domaine de la santé, notamment dans les systèmes d'aide à la prédiction et au diagnostic du diabète. Ce cadre théorique solide constitue un socle indispensable pour l'exploration et le développement de modèles prédictifs avancés visant une détection plus précoce et plus précise du diabète.

# Chapitre2:

# *Apprentissage automatique pour la prédiction du diabète*

## 2.1 Introduction

L'apprentissage automatique joue un rôle essentiel dans la prédiction du diabète en permettant l'analyse automatique de données médicales pour identifier les facteurs de risque et prédire la présence ou l'apparition de la maladie. Grâce à sa capacité à modéliser des relations complexes, il améliore la précision des diagnostics et soutient les décisions cliniques. Ce chapitre introduit l'apprentissage automatique et décrit ses principales techniques appliquées à la prédiction du diabète, les méthodes d'apprentissage profond, les approches d'ensemble telles que le bagging, le boosting et le stacking, ainsi que les métriques utilisées pour évaluer les performances des modèles. Il se conclut par un aperçu des travaux récents dans ce domaine.

## 2.2 Apprentissage automatique « Machine Learning »

L'apprentissage automatiqueconsiste à utiliser des caractéristiques spécifiques d'un ensemble de données pour identifier des modèles. Ces modèles permettent ensuite d'analyser de nouvelles situations. Une fois entraînée, la machine peut appliquer ce qu'elle a appris à des cas similaires à venir [33], [41].

Cet outil de prédiction peut être intégré de façon dynamique dans le processus de décision clinique, afin d'adapter les soins en fonction de chaque patient, plutôt que de suivre un protocole rigide [35].Le processus de machine learning suit généralement sept étapes (Figure 2.1).

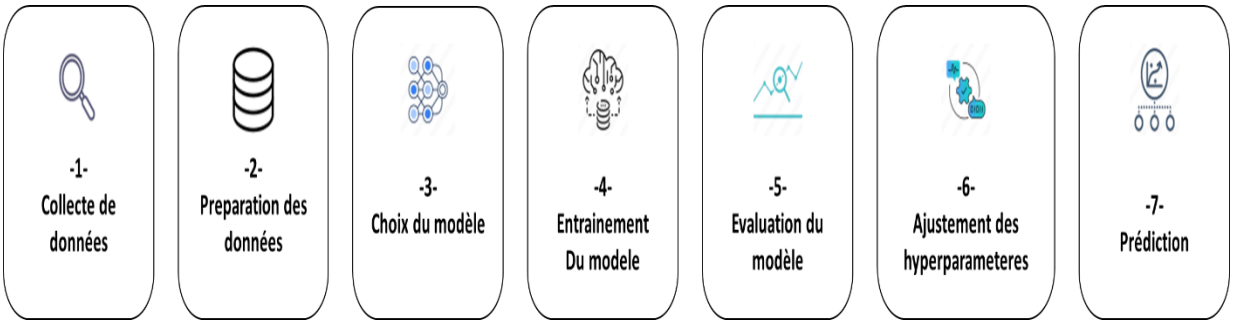


**Figure 2.1:** Étapes d'apprentissage automatique

## 2.3 Application de l'apprentissage automatique à la prédiction du diabète

L'apprentissage automatique est de plus en plus utilisé pour prédire le diabète en raison de sa capacité à traiter et à analyser d'importants volumes de données hétérogènes. Ces techniques permettent d'extraire des motifs pertinents à partir de données cliniques, biologiques et comportementales, afin de construire des modèles prédictifs précis et robustes. Elles suscitent un intérêt considérable dans le domaine médical, notamment en raison de leur potentiel à améliorer

le dépistage précoce et à soutenir la prise de décision clinique.

Dans la suite, nous présentons les différentes étapes et techniques clés du processus de modélisation en machine learning, à savoir : le prétraitement des données (data preprocessing), étape essentielle pour la qualité des prédictions ; les approches d'apprentissage supervisé et non supervisé, selon la disponibilité des étiquettes dans les données ; les techniques d'apprentissage profond (deep learning), utilisées notamment pour l'analyse de données complexes et massives ; les méthodes d'ensemble (ensemble methods) et l'empilement de modèles (model stacking), qui visent à améliorer les performances globales des prédictions en combinant plusieurs algorithmes.

### 2.3.1. Prétraitement des données (Data Preprocessing)

Avant d'utiliser des algorithmes d'apprentissage automatique, il est essentiel de préparer les données afin d'assurer leur qualité et leur pertinence. Cela passe par plusieurs techniques de prétraitement, comme le nettoyage des données, la normalisation, le traitement des valeurs manquantes, la gestion du déséquilibre entre les classes, la détection des valeurs aberrantes et l'encodage des variables [43].

- **Nettoyage des données :**Le nettoyage des données est une étape essentielle pour garantir la qualité des informations avant l'entraînement d'un modèle. Il consiste à gérer les valeurs manquantes (par suppression ou imputation), à détecter et corriger les valeurs aberrantes, à corriger les erreurs de saisie, à uniformiser les noms de catégories similaires, et à supprimer les doublons afin d'éviter les biais et améliorer la performance du modèle.
- **Normalisation :** La normalisation consiste à mettre les données à l'échelle pour qu'elles se situent dans une plage définie, ce qui facilite l'apprentissage des modèles et évite que certaines variables n'aient trop d'influence en raison de leur échelle. Parmi les méthodes courantes, on trouve : la normalisation Min-Max (plage 0 à 1), la standardisation ou Z-score (moyenne 0, écart-type 1), la normalisation par décimale (nombre fixe de décimales) et la normalisation par plage (ajustement selon une plage définie) [44], [45].

### 2.3.2. Gestion du déséquilibre des classes

Lorsque les classes d'un jeu de données sont déséquilibrées (par exemple, peu de cas positifs de diabète par rapport aux cas négatifs), les algorithmes d'apprentissage automatique peuvent favoriser la classe majoritaire. Cela nuit à la performance du modèle, notamment pour la détection des cas rares. Pour remédier à ce problème, plusieurs techniques sont utilisées :

1. **Sous-échantillonnage** : consiste à réduire le nombre d'exemples de la classe majoritaire. Bien qu'efficace pour équilibrer les données, cela peut entraîner une perte d'informations importantes [46].

2. **Sur-échantillonnage**: consiste à augmenter le nombre d'exemples de la classe minoritaire, soit en dupliquant les données existantes, soit en générant de nouvelles instances synthétiques à l'aide de méthodes comme [47], [48] :

- **SMOTE (SyntheticMinority Over-sampling Technique)**: génère de nouveaux exemples synthétiques en interpolant les caractéristiques entre des exemples proches de la classe minoritaire ;
- **ADASYN (Adaptive SyntheticSampling)** : variante de SMOTE qui génère davantage de données synthétiques dans les zones où la classe minoritaire est moins représentée, en fonction de la densité locale ;
- **SVMSMOTE** : combine SMOTE avec un classifieur SVM pour générer des exemples synthétiques à proximité de la frontière de décision, ce qui permet de mieux apprendre les cas difficiles à classer ;
- **KMeansSMOTE** : applique SMOTE de manière ciblée dans des groupes homogènes, créés par regroupement via l'algorithme K-means, pour générer des exemples synthétiques plus représentatifs ;
- **SMOTE-Tomek** : combine SMOTE avec la suppression des "liens Tomek", c'est-à-dire des paires d'exemples de classes différentes trop proches. Cette méthode permet à la fois de suréchantillonner et de nettoyer les données bruyantes.

## 2.4 Approches d'apprentissage supervisé

Les méthodes d'apprentissage supervisé sont couramment employées dans les travaux de prédiction du diabète. Elles reposent sur l'entraînement de modèles à partir de données étiquetées, c'est-à-dire pour lesquelles chaque exemple est associé à un diagnostic connu (comme la présence ou l'absence de diabète). Parmi les algorithmes les plus utilisés figurent la régression logistique, les arbres de décision, les machines à vecteurs de support (SVM) ainsi que les forêts aléatoires. Ces techniques ont permis de concevoir des modèles prédictifs exploitant diverses variables, notamment des indicateurs cliniques, des marqueurs génétiques, des habitudes de vie et d'autres facteurs pertinents [49]-[51]. Nous analyserons dans la suite les performances de ces algorithmes dans le cadre de la prédiction du diabète.

### 2.4.1 Régression Logistique

La régression logistique est une méthode d'apprentissage supervisé utilisée principalement pour la classification binaire. Contrairement à la régression linéaire qui prédit des valeurs continues, la régression logistique modélise la probabilité qu'un événement appartienne à une classe donnée (par exemple, malade ou non malade). Elle utilise la fonction logistique (ou sigmoïde) pour contraindre la sortie entre 0 et 1, ce qui permet d'interpréter cette sortie comme

une probabilité. Une valeur seuil (généralement 0,5) est ensuite appliquée pour transformer cette probabilité en prédiction catégorielle.

La fonction logistique est définie comme suit :

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}} \quad \ldots\ldots\ldots (1)$$

Où :

- $P(Y=1|X)$ : Probabilité que l'observation appartiennent à la classe 1 (la classe positive).
- $\beta 0$ : Terme de biais (intercept).
- $\beta 1, \beta 2, \ldots, \beta n$ : Coefficients du modèle (poids associés à chaque variable d'entrée).
- $X1, X2, \ldots, Xn$ : Les variables d'entrée (features).

Cette technique est largement utilisée dans des domaines comme la médecine, les sciences sociales ou la finance, en raison de sa simplicité, de son interopérabilité et de son efficacité pour des données linéairement séparables [52].

### 2.4.2 Arbre de decision

Un arbre de décision est un algorithme d'apprentissage supervisé utilisé à la fois pour des tâches de classification et de régression. Il repose sur un processus de partition récursive des données en sous-groupes homogènes à l'aide de tests logiques (conditions) sur les variables d'entrée.

Chaque nœud interne de l'arbre correspond à un test sur un attribut, chaque branche à une issue de ce test, et chaque feuille à une prédiction de classe (ou une valeur dans le cas d'une régression). Le modèle construit ainsi une structure hiérarchique facile à interpréter, qui permet d'identifier les facteurs les plus déterminants dans la prise de décision [53].

### 2.4.3 Machines à Vecteurs de Support (SVM)

La machine à vecteurs de support (SVM) est un algorithme d'apprentissage supervisé largement reconnu pour sa puissance, son élégance et son efficacité dans divers domaines. Utilisé aussi bien pour des tâches de classification que de régression, notamment dans des espaces de haute dimension, cet algorithme repose sur les travaux fondamentaux de Vapnik (1995, 1998) et s'appuie sur la théorie de l'apprentissage statistique, en particulier le principe de minimisation du risque structurel, garantissant une bonne capacité de généralisation.

Le principe de SVM consiste à trouver l'hyperplan optimal qui sépare au mieux les différentes classes dans un espace à n dimensions, en maximisant la marge entre elles. Cette frontière de décision est définie à partir de certains points clés appelés vecteurs de support. Bien que SVM soit principalement conçu pour la classification binaire, il peut également être adapté aux problèmes multi-classes en les divisant en plusieurs sous-problèmes binaires. De nombreuses études récentes ont montré que les SVM offrent une précision de classification

supérieure à celle de nombreux autres algorithmes, ce qui les rend particulièrement efficaces dans le contexte de la prédiction du diabète, où il s'agit souvent de distinguer entre présence et absence de la maladie [54].

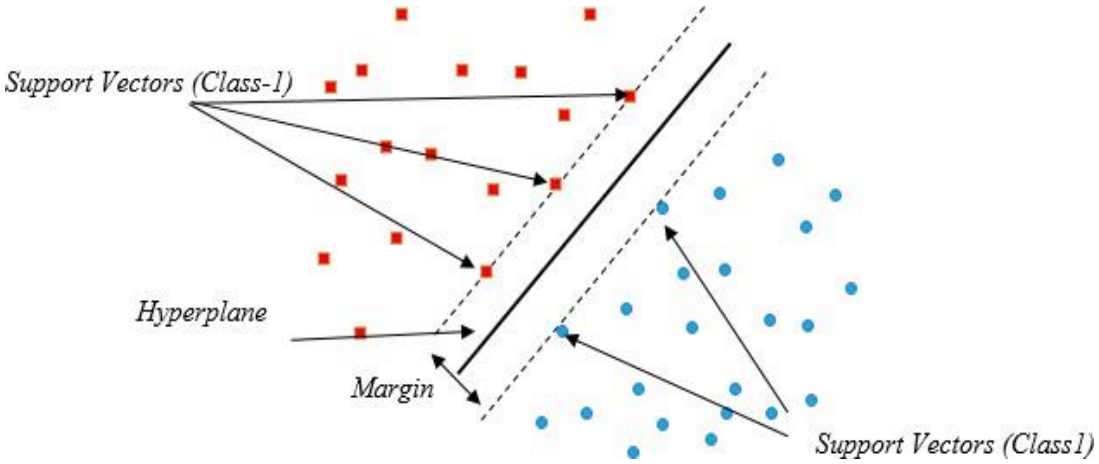


**Figure 2. 2:** Séparation parfait de deux classes avec un hyperplan

### 2.4.4 Forêts aléatoires (Random Forest)

Les forêts aléatoires sont une extension des arbres de décision, fondée sur l'agrégation de plusieurs arbres de décision. Elle se distingue par sa capacité à traiter efficacement différents types de données (nominales, numériques et binaires) et par ses bonnes performances générales, souvent difficiles à surpasser. Cette technique repose sur la construction de multiples arbres de décision entraînés sur des sous-ensembles aléatoires des données et des variables, puis sur l'agrégation de leurs prédictions, généralement par vote majoritaire dans les tâches de classification. Bien que principalement utilisée pour la classification, la forêt aléatoire peut également être appliquée à des tâches de régression. Son fonctionnement repose sur une logique simple : par exemple, si sept arbres donnent une prédiction sur une variable et que quatre d'entre eux s'accordent sur une classe donnée, celle-ci est retenue comme résultat final. Ce mécanisme basé sur le vote permet de construire des modèles fiables, tout en réduisant le risque de surapprentissage. Par ailleurs, en sélectionnant aléatoirement des sous-ensembles de caractéristiques à chaque division de l'arbre, la méthode assure une plus grande diversité entre les arbres, ce qui améliore la généralisation du modèle. Les hyperparamètres de la forêt aléatoire sont similaires à ceux utilisés dans les arbres de décision ou les méthodes de type « bagging».Contrairement aux arbres de décision classiques, les forêts aléatoires réduisent le risque de sur-apprentissage aux données d'entraînement, offrant ainsi une meilleure généralisation des prédictions [55].

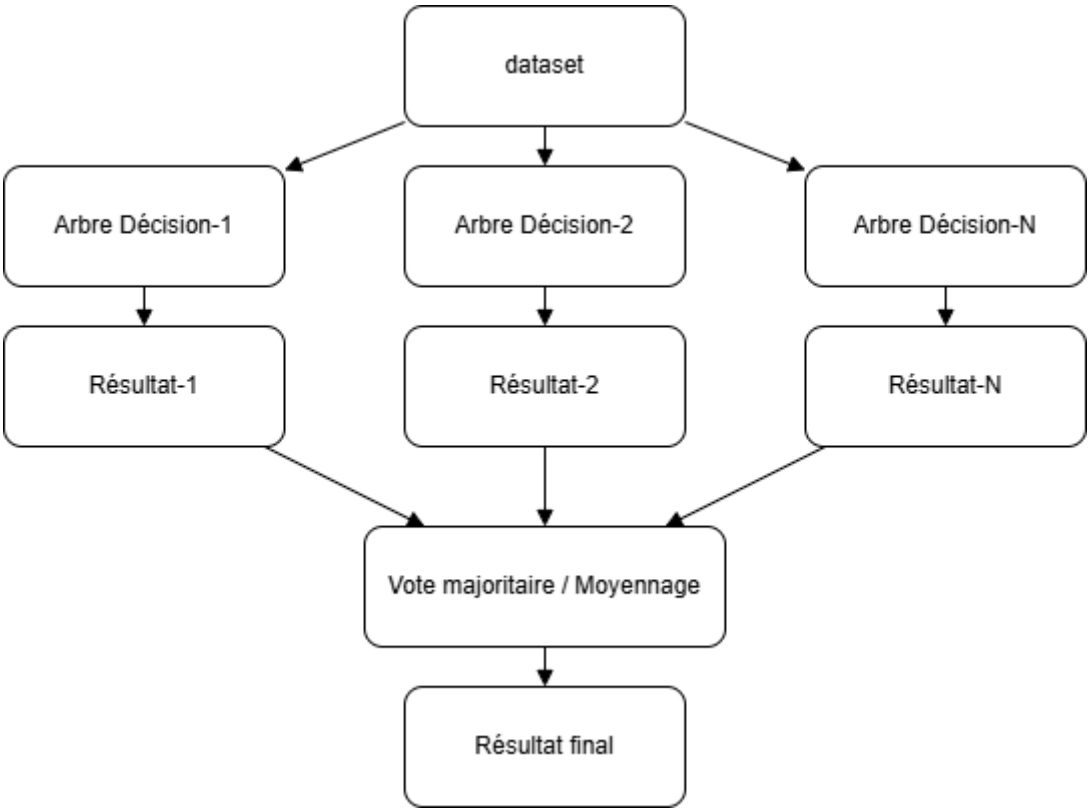


**Figure 2. 3:** Structure de l'algorithme Random Forest

## 2.5 Techniques d'apprentissage profond

L'apprentissage profond, qui constitue une branche de l'apprentissage automatique, s'est affirmé comme une approche particulièrement performante dans de nombreux domaines, notamment celui de la prédiction du diabète. Grâce à leur capacité à extraire automatiquement des représentations complexes à partir de données de grande dimension, les réseaux neuronaux profonds ont démontré une efficacité notable dans l'identification de relations non linéaires et de signaux subtils associés au diabète. En particulier, les perceptrons multicouches (MLP), les réseaux de neurones convolutifs (CNN) et les réseaux de neurones récurrents (RNN) ont été employés avec succès sur diverses sources de données, telles que les images médicales, les séries temporelles physiologiques et les dossiers médicaux électroniques (DME) [56], [57].

### 2.5.1 Réseaux neuronaux convolutifs (CNN)

Les réseaux de neurones convolutifs (CNN), développés pour la première fois par Yann LeCun en 1988, sont une catégorie spécialisée de réseaux de neurones conçus pour le traitement de données structurées en grille, comme les images [58]. Inspirés de processus biologiques, les CNN appartiennent à la famille des réseaux de neurones feed-forward et ont démontré une grande efficacité dans des domaines variés, notamment la reconnaissance et la classification d'images et de vidéos. Ils sont largement utilisés dans des applications telles que l'identification de visages, la détection d'objets, les panneaux de signalisation, ainsi que dans les systèmes de

conduite autonome. L'architecture d'un CNN repose sur une succession de blocs de traitement destinés à extraire des caractéristiques discriminantes permettant de différencier les classes d'images. Chaque bloc est généralement composé de plusieurs couches : la couche de convolution, qui applique des filtres appris pour extraire des motifs caractéristiques (comme des bords ou des textures) ; la couche de correction non linéaire (ReLU), qui introduit une non-linéarité essentielle à l'apprentissage ; la couche de pooling, qui réduit la dimensionnalité tout en conservant l'information pertinente ; la couche entièrement connectée, qui consolide les caractéristiques extraites pour la classification finale ; et enfin, la couche de perte, qui mesure l'écart entre les prédictions du modèle et les valeurs réelles [59]. Le terme "convolutif" fait référence à l'opération mathématique de convolution, qui consiste à appliquer un filtre glissant sur l'image d'entrée. Ce filtre, dont les paramètres sont appris durant l'entraînement, permet par exemple de détecter des motifs comme des angles, utiles pour la classification [58], [59].

### 2.5.2 Réseaux neuronaux récurrents (RNN)

Les réseaux de neurones récurrents (RNN) sont une classe particulière de réseaux neuronaux conçus pour traiter des données séquentielles ou temporelles, grâce à leur capacité à intégrer une mémoire interne. Contrairement aux réseaux neuronaux traditionnels, les RNN possèdent des connexions récurrentes, c'est-à-dire que la sortie d'une couche peut être réintroduite comme entrée dans cette même couche lors de l'étape suivante, formant ainsi des boucles de rétroaction [60]. Cette architecture leur permet de conserver l'information des étapes précédentes et de prendre en compte le contexte dans les séquences de données, ce qui les rend particulièrement adaptés aux tâches impliquant une dépendance temporelle ou un enchaînement logique. Les RNN sont utilisés dans divers domaines tels que la traduction automatique, le sous-titrage d'images, le traitement du langage naturel, la détection automatique des apnées du sommeil à partir de l'ECG nocturne, ou encore le traitement automatique de la parole. On les retrouve également dans des applications comme la reconnaissance vocale, la composition musicale, l'analyse de sentiments, l'analyse de séquences ADN et l'interprétation de séries temporelles issues de capteurs. La principale différence entre les RNN et d'autres types de réseaux réside dans leur capacité à traiter les données en tenant compte de leur ordre, souvent dans le temps, en se souvenant des observations précédentes afin de mieux comprendre les relations au sein de la séquence [61].

### 2.5.3 Multilayer Perceptron (MLP)

Le Multi-Layer Perceptron (MLP) est un type de réseau de neurones artificiels feedforward, composé d'au moins trois couches de neurones : une couche d'entrée, une ou plusieurs couches cachées, et une couche de sortie. Chaque neurone d'une couche est entièrement connecté à ceux de la couche suivante. L'apprentissage du MLP se fait généralement à l'aide de

la rétropropagation de l'erreur (backpropagation), qui ajuste les poids synaptiques pour minimiser l'erreur de prédiction. Les MLP sont capables de modéliser des fonctions non linéaires complexes, ce qui les rend particulièrement adaptés aux tâches de classification et de régression [62].

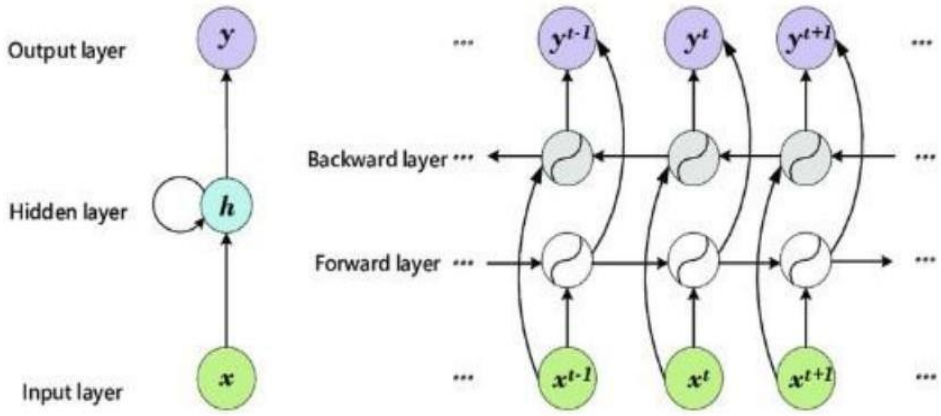


**Figure 2.4:**Architecture de réseau récurent

## 2.6 Méthodes d'ensemble et empilement de modèles

Les méthodes d'ensemble constituent une stratégie efficace pour améliorer la performance globale et la robustesse des modèles de prédiction du diabète. En combinant plusieurs modèles prédictifs, ces approches permettent de tirer parti de la diversité et des forces complémentaires de chaque modèle individuel. Des techniques populaires telles que le *bagging*, le *boosting* et le *stacking* sont couramment utilisées pour construire ces ensembles. Elles contribuent notamment à réduire le surapprentissage, à limiter le biais et à renforcer la capacité de généralisation des modèles [63], [64]. Dans le cadre de la prédiction du diabète, ces méthodes se sont révélées efficaces pour accroître à la fois la précision et la stabilité des résultats.

### 2.6.1 Bagging

Le bagging (BootstrapAggregating) est une technique d'ensemble qui consiste à entraîner plusieurs modèles (souvent des arbres de décision) de manière indépendante sur différents sous-ensembles des données d'entraînement, générés par échantillonnage aléatoire avec remise. Ainsi, certains échantillons peuvent être répétés dans un sous-ensemble donné, tandis que d'autres peuvent ne pas y figurer du tout. Les prédictions de ces modèles sont ensuite combinées, soit par vote majoritaire dans les tâches de classification, soit par moyenne dans les tâches de régression. L'objectif principal du bagging est de réduire la variance des modèles prédictifs et d'améliorer leur capacité de généralisation, en atténuant les effets du surajustement. Cette méthode est particulièrement efficace lorsqu'elle est appliquée à des modèles instables, c'est-à-dire sensibles

aux variations dans les données d'apprentissage [65].

### 2.6.2 Arbres extrêmement aléatoires (Extra Trees)

Extra Trees (ou ExtremelyRandomizedTrees) est une variante de l'algorithme Random Forest. Comme ce dernier, il construit un ensemble d'arbres de décision à partir de sous-échantillons aléatoires des données d'apprentissage et en sélectionnant aléatoirement un sous-ensemble de caractéristiques pour chaque division de nœud. Cependant, Extra Trees pousse plus loin la randomisation en sélectionnant les seuils de séparation de manière entièrement aléatoire, sans chercher à optimiser la qualité de la division. Cette randomisation accrue peut améliorer la capacité de généralisation du modèle tout en réduisant sa sensibilité au bruit présent dans les données [66].

### 2.6.3 Boosting

Le boosting est une méthode d'ensemble itérative qui consiste à construire une série de modèles faibles, généralement des arbres de décision peu profonds, en accordant progressivement plus d'importance aux exemples mal prédits lors des itérations précédentes. Contrairement au bagging, où les modèles sont entraînés indépendamment, le boosting forme les modèles de manière séquentielle, chaque nouveau modèle corrigeant les erreurs commises par ses prédécesseurs.

Dans ce processus, des poids sont attribués aux instances d'apprentissage : les exemples mal classés obtiennent des poids plus élevés, ce qui oblige les modèles suivants à se concentrer davantage sur ces cas difficiles. Chaque modèle faible contribue à la prédiction finale avec un poids proportionnel à sa performance, permettant ainsi de construire un modèle robuste à partir de modèles de base simples.

Le boosting permet de réduire le biais et d'améliorer significativement la précision globale du modèle en capturant des relations complexes dans les données [67], [68].

- **XGBoost (Extreme Gradient Boosting)**

XGBoost (Extreme Gradient Boosting) est une implémentation efficace et évolutive de l'algorithme de gradient boosting pour les arbres de décision. Il a été conçu pour optimiser à la fois la vitesse et les performances prédictives. XGBoost repose sur une formulation de l'apprentissage par boosting en minimisant une fonction objectif régulière, ce qui permet de contrôler le surapprentissage et de généraliser efficacement. Il intègre plusieurs optimisations techniques telles que la pruningalgorithm, la sparsityawareness, et une parallélisation efficace, le rendant adapté à des applications à grande échelle [69].

- **CatBoost (CategoricalBoosting)**

CatBoost est un algorithme de gradient boosting développé par Yandex, conçu pour gérer efficacement les variables catégorielles sans nécessiter de prétraitement complexe (comme

l'encodage one-hot). Il repose sur des obliviousdecisiontrees (arbres symétriques) et intègre des innovations telles que l'encodage basé sur les statistiques de cibles avec permutations aléatoires, ce qui réduit les risques de surapprentissage. CatBoost est également optimisé pour les performances sur CPU et GPU, et fonctionne efficacement sur des données structurées [70].

- **LightGBM (Light Gradient Boosting Machine)**

LightGBM est un algorithme de gradient boosting qui adopte une méthode innovante de construction des arbres de décision. Il améliore l'efficacité de l'apprentissage et les performances de prédiction en utilisant une technique fondée sur les histogrammes, qui consiste à regrouper les valeurs des variables continues. Cette approche permet de réduire significativement le temps d'entraînement et de traiter plus efficacement de grands ensembles de données, surpassant ainsi les méthodes traditionnelles de gradient boosting en termes de rapidité et de consommation mémoire [71].

### 2.6.4 Stacking

Le stacking,est une technique d'ensemble qui combine les prédictions de plusieurs modèles de base en les intégrant dans un modèle de niveau supérieur, appelé méta-modèle ou modèle d'agrégation. Contrairement au bagging et au boosting qui s'appuient sur des approches homogènes ou séquentielles, le stacking adopte une architecture hiérarchique : les prédictions générées par les modèles de premier niveau (souvent hétérogènes) sont utilisées comme variables d'entrée pour entraîner le méta-modèle. Ce dernier apprend à combiner ces prédictions de manière optimale afin d'améliorer la performance globale du système.

Le stacking permet ainsi de capitaliser sur les points forts de différents algorithmes, en fusionnant leur capacité prédictive pour obtenir des résultats plus robustes et précis [72]. En complément du bagging et du boosting, il offre des gains notables en termes de réduction de variance, d'amélioration de la précision et d'adaptabilité aux relations complexes présentes dans les données.Cependant, l'efficacité du stacking dépend fortement du choix des modèles de base, de la configuration du méta-modèle, ainsi que d'une gestion rigoureuse des hyperparamètres. Une validation croisée soignée est également indispensable pour prévenir le surapprentissage et garantir une généralisation fiable [73], [74].

## 2.7 Métriques d'évaluation et évaluation des performances

L'évaluation des modèles prédictifs appliqués à la détection du diabète repose sur l'utilisation de métriques adaptées permettant de quantifier leur performance. Parmi les mesures les plus couramment utilisées, on retrouve l'accuracy, la précision, le rappel, le score F1, l'AUC-ROC. Ces indicateurs permettent d'apprécier la capacité du modèle à faire des prédictions correctes, à détecter efficacement les cas positifs, et à maintenir un bon équilibre entre les faux

positifs et les faux négatifs.

- **Accuracy (Exactitude)**

L'accuracy mesure la proportion de prédictions correctes parmi l'ensemble des observations. Elle est efficace lorsque les classes sont équilibrées, mais peut être trompeuse en cas de déséquilibre de classes [75].

$$Accuracy = \frac{tp + tn}{(tp + tn + fp + fn)} \quad \text{..... (2)}$$

**TP** : Vrais positifs, **TN** : Vrais négatifs, **FP** : Faux positifs, **FN** : Faux négatifs

- **Précision (Precision)**

La précision indique la proportion de vrais positifs parmi toutes les prédictions positives. Elle est utile quand le coût des faux positifs est élevé (par exemple en diagnosticmédical) [76].

$$Precision = \frac{tp}{tp + fp} \quad \text{..... (3)}$$

- **Rappel (Recall )**

Le rappel mesure la capacité du modèle à identifier correctement tous les vrais positifs. Il est crucial lorsque le coût des faux négatifs est élevé (comme dans les dépistages de maladies graves) [77].

$$Recall = \frac{tp}{tp + fn} \quad \text{..... (4)}$$

- **Score F1 (F1-Score)**

Le F1-score est la moyenne harmonique entre la précision et le rappel, ce qui permet de trouver un équilibre entre les deux. Il est recommandé lorsque les données sont déséquilibrées[78].

$$F1\ score = 2 * \frac{precosion * recall}{precision + recall} \quad \text{..... (5)}$$

- **AUC-ROC (Area Under the Curve - Receiver Operating Characteristic)**

L'AUC-ROC est une métrique qui mesure la capacité du modèle à distinguer entre les classes. Elle trace la courbe ROC (taux de vrais positifs vs taux de faux positifs) et calcule l'aire sous cette courbe. Plus l'AUC est proche de 1, meilleure est la performance [79].

## 2.7 Travaux récents

De nombreuses études ont montré des performances encourageantes pour la prédiction du diabète.

**Tableau 2.1:** Travaux Connexes pour la prédiction du diabète .

| Papier | Dataset | Détails Dataset | Année | Algorithmes | Performances |
|---|---|---|---|---|---|
| Towards Transparent and Accurate Diabetes Prediction Using Machine Learning and Explainable Artificial Intelligence [80] | BRFSS (sous-ensemble) | Données propres de 70 692 réponses à l'enquête BRFSS2015. Cet ensemble de données compte 253 680 enregistrements et 21 variables non équilibrées. Split 80:20. PCA appliquée + SMOTE utilisé pour corriger le déséquilibre. | 2025 | Modèles d'ensemble | 92,5 %(Accuracy)<br>97,50%(AUC) |
| Machine Learning Algorithm-Based Prediction of Diabetes Among Female Population Using PIMA Dataset [81] | PIMA, dataset privé | PIMA : 768 patients, 9 attributs, 268 diabétiques, 500 non. Un dataset privé (application mobile, patients réels) est également utilisé. | 2024 | Random Forest (RF)<br>Decision Tree (DT)<br>Naïve Bayes (NB)<br>Régression Logistique (LR) | Random Forest :<br>80 %(Accuracy)<br>82 %(Precision)<br>88 %(Recall)<br>20 %(Taux d'erreur)<br>83 % (AUC) |
| A Proposed Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App [82] | PIMA | SMOTE utilisé pour équilibrer les classes. | 2024 | XGBoost<br>Naïve Bayes<br>Support Vector Machine (SVM)<br>K-Nearest Neighbors (KNN)<br>Decision Tree<br>Random Forest<br>Logistic Regression<br>Gradient Boosting<br>AdaBoost<br>Bagging (SMOTE a été utilisé pour équilibrer les classes). | **XGBoost** :<br>97.4% (Accuracy)<br>87 % **(**AUC)<br>96.3% (Précision)<br>97.8%(Rappel)<br>**NB, SVM, etc.** : performances moindres que XGBoost (non détaillées individuellement). |

| | | | | | |
|---|---|---|---|---|---|
| Prediction of gestational diabetes using deep learning and Bayesian optimization and traditional machine learning techniques [83] | Données cliniques collectées | Données collectées prospectivement par trois médecins spécialistes (489 patients, 73 variables). | 2023 | RNN-LSTM with Bayesian optimization | 98.00% (AUC) |
| A novel solution of deep learning for enhanced support vector machine for predicting the onset of type 2 diabetes [84] | FIMMG (Metmedica Italia) | ___ | 2023 | Support Vector Machine (SVM) algorithm+ Radial Base Function (RBF) along + Long Short-term Memory | 86.31% (Accuracy)<br>82.70% (AUC) |
| A novel hybrid machine learning framework for the prediction of diabetes regularization and prediction procedures [85] | PIMA | ___ | 2022 | Modèle hybride personnalisé de réseau neuronal artificiel | 80% (Accuracy) |
| An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators [86] | DiabetesBinaryHealthIndicatorsDataset (BRFSS probable) | ___ | 2022 | RF<br>DT<br>KNN<br>LR<br>NB | 82.26%<br>81.02%<br>80.55%<br>72.64%<br>70.56%<br>(Accuracy) |

| | | | | | |
|---|---|---|---|---|---|
| Machine learning based diabetes prediction and development of smart web application [87] | PIMA, dataset 2 (950 enregistrements, 19 attributs) | ___ | 2021 | Dataset1 : | |
| | | | | NB | 86.17% |
| | | | | DT | 96.81% |
| | | | | RF | 96.81% |
| | | | | SVM | 91.49% |
| | | | | LR | 84.04% |
| | | | | GB | 91.00% |
| | | | | KNN | 90.43% |
| | | | | Dataset2 : | |
| | | | | NB | 78.95% |
| | | | | DT | 76.32% |
| | | | | RF | 80.26% |
| | | | | SVM | 80.26% |
| | | | | LR | 77.63% |
| | | | | GB | 78.95% |

## 2.8 Conclusion

L'apprentissage automatique et l'apprentissage profond ont montré un fort potentiel dans la prédiction du diabète, en facilitant une analyse plus fine, rapide et automatisée des données médicales. Ce chapitre a présenté plusieurs techniques, allant des modèles supervisés classiques aux méthodes d'ensemble et aux réseaux neuronaux, toutes en offrant des performances encourageantes pour la détection précoce de la maladie. Une évaluation rigoureuse à l'aide de métriques appropriées a permis de comparer l'efficacité relative de ces approches. Ces résultats sont en accord avec plusieurs travaux connexes dans la littérature, qui confirment l'apport significatif de ces techniques dans le domaine médical, notamment pour le diagnostic assisté.

# Chapitre 3:

## *Approche proposée*

## 3.1 Introduction

Ce chapitre est consacré à la présentation de la méthodologie mise en œuvre dans le cadre du développement de notre application de détection et de prédiction du diabète. L'objectif est de fournir une vue d'ensemble claire et structurée des étapes suivies pour concevoir une solution basée sur l'intelligence artificielle, capable d'analyser efficacement les données médicales afin d'identifier les individus à risque. Nous débuterons par une description générale de notre approche, en justifiant les choix méthodologiques et les sources de données utilisées. Nous détaillerons ensuite les étapes de prétraitement nécessaires à la préparation des données, incluant le nettoyage, la normalisation et la gestion des valeurs manquantes. Des techniques d'exploration et de visualisation seront également appliquées afin de mieux comprendre la structure et les corrélations présentes dans les données. Par la suite, nous présenterons les principaux algorithmes d'apprentissage automatique retenus pour la tâche de prédiction, en exposant leurs principes de fonctionnement et leurs justifications d'utilisation. Enfin, nous décrirons les métriques d'évaluation adoptées pour mesurer et comparer les performances des modèles développés, en vue d'identifier celui offrant les meilleurs résultats.

## 3.2 Approche proposée

L'approche proposée de notre système de prédiction du diabète repose sur une chaîne modulaire et cohérente. En premier lieu, les données cliniques sont soigneusement chargées, nettoyées et prétraitées : les valeurs aberrantes (notamment les zéros dans certaines variables biologiques) sont corrigées, les données manquantes imputées, des variables dérivées sont créées, puis les données sont normalisées pour assurer une meilleure convergence des modèles. Plusieurs algorithmes d'apprentissage supervisé sont ensuite entraînés, parmi lesquels XGBoost, LightGBM, CatBoost, Random Forest, Extra Trees, SVM, ainsi qu'un réseau de neurones profond.

Le déséquilibre des classes est géré via la technique SMOTE, et chaque modèle bénéficie d'une validation rigoureuse par validation croisée et optimisation des hyperparamètres. Pour renforcer la robustesse des prédictions, nous mettons en œuvre une stratégie d'assemblage hybride : un stacking, un voting classifier classique, ainsi qu'un Weighted Ensemble manuel qui combine les prédictions des différents modèles selon des poids définis, permettant ainsi d'ajuster finement l'influence de chaque modèle dans la décision finale. L'évaluation des performances s'appuie sur des métriques exhaustives (accuracy, précision, rappel, F1-score, AUC-ROC), accompagnées de rapports détaillés, matrices de confusion et courbes ROC pour chaque modèle, y compris l'ensemble pondéré. Enfin, pour assurer l'interprétabilité essentielle dans le contexte clinique, nous exploitons des méthodes avancées d'explicabilité : SHAP pour visualiser l'importance globale et locale des variables explicatives, et LIME pour fournir des explications locales des prédictions

individuelles. Cette architecture hybride garantit ainsi une prédiction fiable, tout en offrant une transparence indispensable à la confiance et à l'adoption médicale. Cette architecture a été conçue pour être modulaire, extensible et facilement intégrable par meilleur modèle dans un application web.

**- Schéma global d'approche proposée**

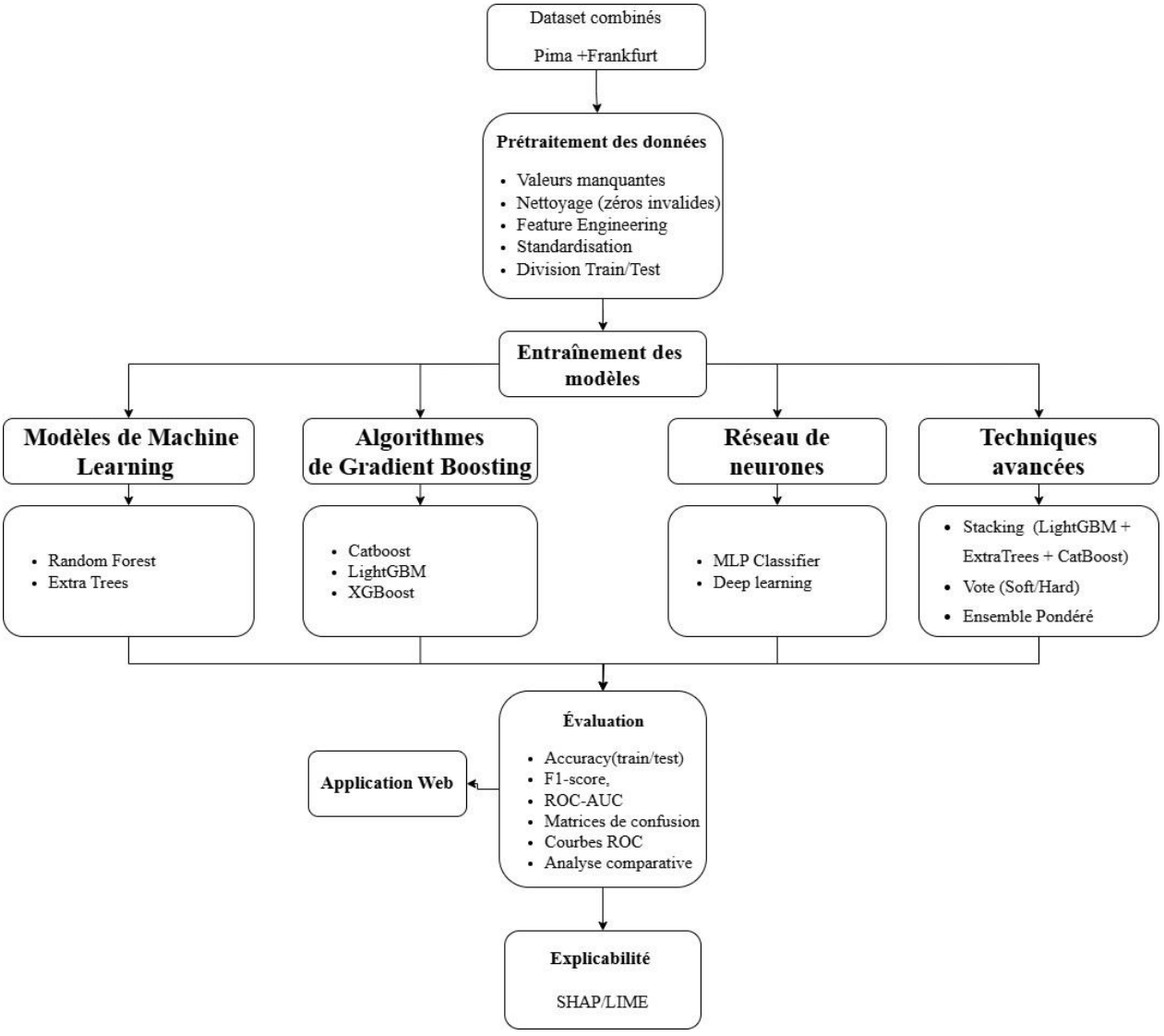


**Figure 3.1:** Schéma global d'approche proposée

## 3.3 Choix du Dataset

Dans le cadre de ce projet, deux datasets ont été exploités et combinéspour entraîner et évaluer les modèles d'apprentissage automatique :

### 3.6.1 Dataset PIMA

Sur le site Web de Kaggle, le dataset PIMA sur le diabète peut être trouvé11. Cet dataset Pima Indian Diabetes Database (PIDD), qui provient de l'Institut national du diabète et des maladies digestives et rénales, peut être utilisé pour prédire de manière diagnostique si un patient est diabétique ou non sur la base de certaines mesures de diagnostic fournies dans la collection. Il se

compose de nombreux paramètres médicaux et d'un paramètre dépendant à valeur binaire (Résultat). Une chose à noter est que tous les patients ici sont des femmes d'au moins 21 ans d'origine indienne Pima. Il y a 768 lignes et 9 colonnes dans cet ensemble de données, nous avons "Résultat" comme variable cible[88].

**Tableau 3.2 :** Description d'attribut de Dataset PIMA.

| N° d'attribut | Description d'attribut | Type de variable |
|---|---|---|
| **1** | Grossesses (Nombre de fois enceinte). | Entier |
| **2** | Glucose (Concentration de glucose plasmatique après une durée de 2 heures lors d'un test de tolérance au glucose oral). | Réel |
| **3** | Pression artérielle (Pression artérielle diastolique/systolique (mm Hg)). | Réel |
| **4** | Épaisseur du pli cutané du triceps | Réel |
| **5** | Insuline (Insuline sérique après une durée de 2 heures (mu U/ml)). | Réel |
| **6** | IMC (indice de masse corporelle : poids en kg/(taille en m)^2). | Réel |
| **7** | Fonction de pedigree du diabète. | Réel |
| **8** | Âge (années). | Entier |
| **9** | Résultat (avec diabète (1) ou non (0)). | Binaire |

### 3.6.2 Dataset Frankfurt

Le Dataset extrait de l'hôpital de Frankfort, Allemagne téléchargé sur kaggle , 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, identiques à celles du dataset PIMA[89].

### 3.6.3 Dataset combiné

Afin d'augmenter la quantité de données, de diversifier les profils de patients et d'améliorer la performance des modèles, les deux datasets précédents ont été fusionnés. Le dataset combiné regroupe ainsi les données cliniques de 2768 patients réparties sur 9 variables identiques.

**Tableau 3.3** : Capture de Dataset

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 6 | 148.0 | 72.0 | 35.0 | 154.23783 | 33.6 | 0.627 | 50 | 1 |
| **1** | 1 | 85.0 | 66.0 | 29.0 | 154.23783 | 26.6 | 0.351 | 31 | 0 |
| **2** | 8 | 183.0 | 64.0 | 29.289634 | 154.23783 | 23.3 | 0.672 | 32 | 1 |
| **3** | 1 | 89.0 | 66.0 | 23.0 | 94.0 | 28.1 | 0.167 | 21 | 0 |
| **4** | 0 | 137.0 | 40.0 | 35.0 | 168.0 | 43.1 | 2.288 | 33 | 1 |
| **0** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **2763** | 2 | 75.0 | 64.0 | 24.0 | 55.0 | 29.7 | 0.37 | 33 | 0 |
| **2764** | 8 | 179.0 | 72.0 | 42.0 | 130.0 | 32.7 | 0.719 | 36 | 1 |
| **2765** | 6 | 85.0 | 78.0 | 29.289634 | 154.23783 | 31.2 | 0.382 | 42 | 0 |
| **2766** | 0 | 129.0 | 110.0 | 46.0 | 130.0 | 67.1 | 0.319 | 26 | 1 |
| **2767** | 2 | 81.0 | 72.0 | 15.0 | 76.0 | 30.1 | 0.547 | 25 | 0 |

2768 rows × 13 columns

**Tableau 3.4:** Description des variables de dataset

| 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. | Frankfurt et Pima:<br>Minimum: 0<br>Maximum:199 |
| **Pregnancies** | Nombre de fois enceinte | Frankfurt et Pima:<br>Minimum: 0<br>Maximum:17 |
| **Blood Pressure** | 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). | Frankfurt et Pima:<br>Minimum: 0<br>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. | Frankfurt :<br>Minimum: 0<br>Maximum:110<br>Pima :<br>Minimum: 0<br>Maximum:99 |
| **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. | Frankfurt :<br>Minimum : 0<br>Maximum : 744<br>Pima :<br>Minimum : 0<br>Maximum : 846 |
| **BMI** | Indice de masse corporelle (poids en kg / (taille en m)^2), Un BMI inférieur à 18,5 indique une maigreur, entre 18,5 et 24,9 un poids normal, entre 25 et 29,9 un surpoids (risque modéré de diabète), et un IMC de 30 ou plus correspond à une obésité (risque élevé de diabète). | Frankfurt :<br>Minimum: 0<br>Maximum:80.60<br>Pima :<br>Minimum: 0<br>Maximum:67.10 |
| **DiabetePredigmeFunction** | 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. | Frankfurt et Pima:<br>Minimum:0.08<br>Maximum:2.42 |
| **Age** | Age d'une personne en années. | Frankfurt et Pima:<br>Minimum:21<br>Maximum:81 |
| **Outcome** | Indique si une personne est diabétique ou non. | Frankfurt et Pima:<br>0(nondiabétique)<br>1 (diabétique) |

### 3.6.4 Visualisation de dataset

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 le dataset et les variables.

**- Comptagede la Variable Cible**

Il visualise la distribution des classes (diabétique/non-diabétique),952 cas de patients diabétiques (Outcome = 1), et 1816 cas de patients non diabétiques (Outcome = 0).

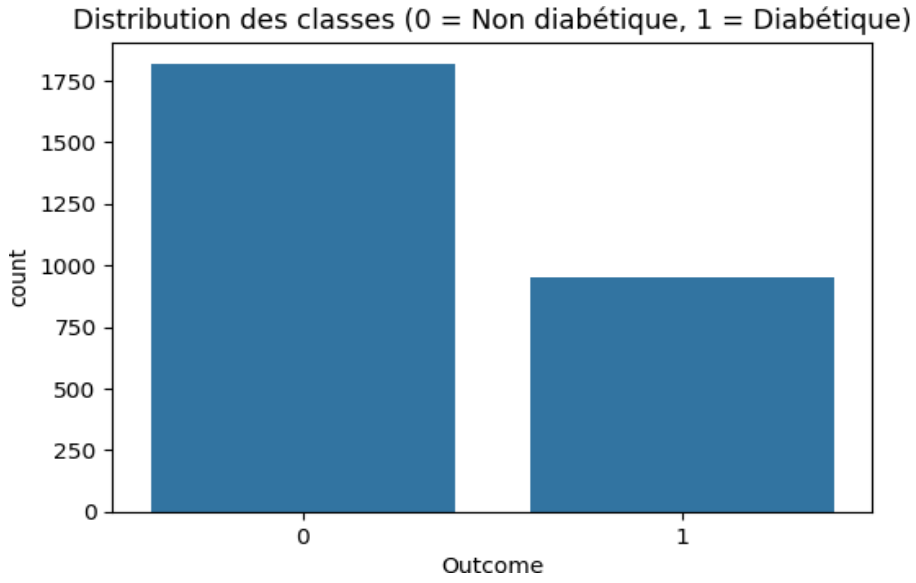


**Figure 3.2**: Comptage de la variable cible (Outcome)

## 3.4 Prétraitement des Données

Après le choix du dataset, l'étape suivante est le prétraitement, une phase cruciale pour garantir la qualité des résultats. La plupart des datasets comportent des valeurs manquantes, bruitées ou incohérentes. Une mauvaise qualité des données conduit inévitablement à de mauvaises performances du modèle, même si celui-ci est performant.

Le prétraitement regroupe plusieurs sous-étapes essentielles, voici une description détaillée des étapes effectuées :

- **Étape 1 : Gestion des valeurs aberrantes (zéro non valides)**

Les colonnes comme 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI' ne devraient pas avoir de zéros. Les zéros ont été considérés comme des **valeurs manquantes** et remplacés par NaN.

- **Étape 2 : Imputation des valeurs manquantes**

Les NaN ont été remplacés par la **moyenne de chaque colonne concernée**, afin de conserver une cohérence statistique tout en évitant de supprimer des lignes.

- **Étape 3 : Définition des colonnes potentiellement communes**

Une liste de colonnes a été définie pour ne conserver que celles présentes dans les deux datasets :

- **Étape 4 : Vérification des colonnes disponibles dans les datasets**

Les colonnes communes ont été vérifiées manuellement dans les datasets pour garantir une compatibilité parfaite.

- **Étape 5 : Filtrage des deux datasets selon les colonnes communes**

Chaque dataset a été réduit aux **colonnes communes uniquement**, pour permettre une fusion correcte.

- **Étape 6 : Concaténation des deux datasets en un seul dataset**

Les datasets**PIMA** et **Frankfurt** ont été concaténés pour créer un seul dataset combiné :

- **Étape 7 : Création de nouvelles variables à partir des existantes**

Des nouvelles variables ont été créées à partir des données existantes pour enrichir le jeu de données .

- **Étape 8 : Division entre caractéristiques (X) et cible (y)**

On sépare les caractéristiques (X) (toutes les colonnes sauf Outcome) de la variable cible (y), qui contient les étiquettes binaires indiquant si un patient est diabétique ou non.

**Caractéristiques (X):**(2768, 12)

**Cible (y):**(2768,)

- **Étape 9 : Standardisation des données (StandardScaler)**

Les données sont **normalisées** avec StandardScaler pour que chaque variable ait une **moyenne nulle** et **une variance de 1**. Cela est nécessaire pour certains algorithmes sensibles à l'échelle des données.

- **Étape 10 : Découpage en ensemble d'entraînement et test (80%/20%)**

Les données sont divisées en deux parties :

- **Un ensemble d'entraînement** : 80 % (2214, 12) pour apprendre les modèles.
- **Un ensemble de test** : 20 % (554, 12) pour évaluer leur performance sur des données inédites.

### 3.6.1 Sélection des Caractéristiques (Facteurs / Attributs)

La sélection des caractéristiques est une étape cruciale dans la conception d'un modèle d'apprentissage automatique. Elle permet de réduire la complexité du modèle, d'éviter le surapprentissage (overfitting), et d'améliorer l'interprétabilité et la performance globale du système.

Après la création des nouvelles variables, les caractéristiques finales sont :Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Glucose_BMI_Ratio, Age_BMI, Insulin_Glucose_Ratio, Pregnancies_Age_Ratio.

### 3.6.2 Analyse statistique et corrélation

Une analyse exploratoire des données a été réalisée pour identifier les relations entre les variables et la variable cible (Outcome).

**Matrice de Corrélation :**Une analyse exploratoire des données afind'identifier les relations linéaires entre les variables, y compris celles entre les variables explicatives et la variable cible (Outcome).

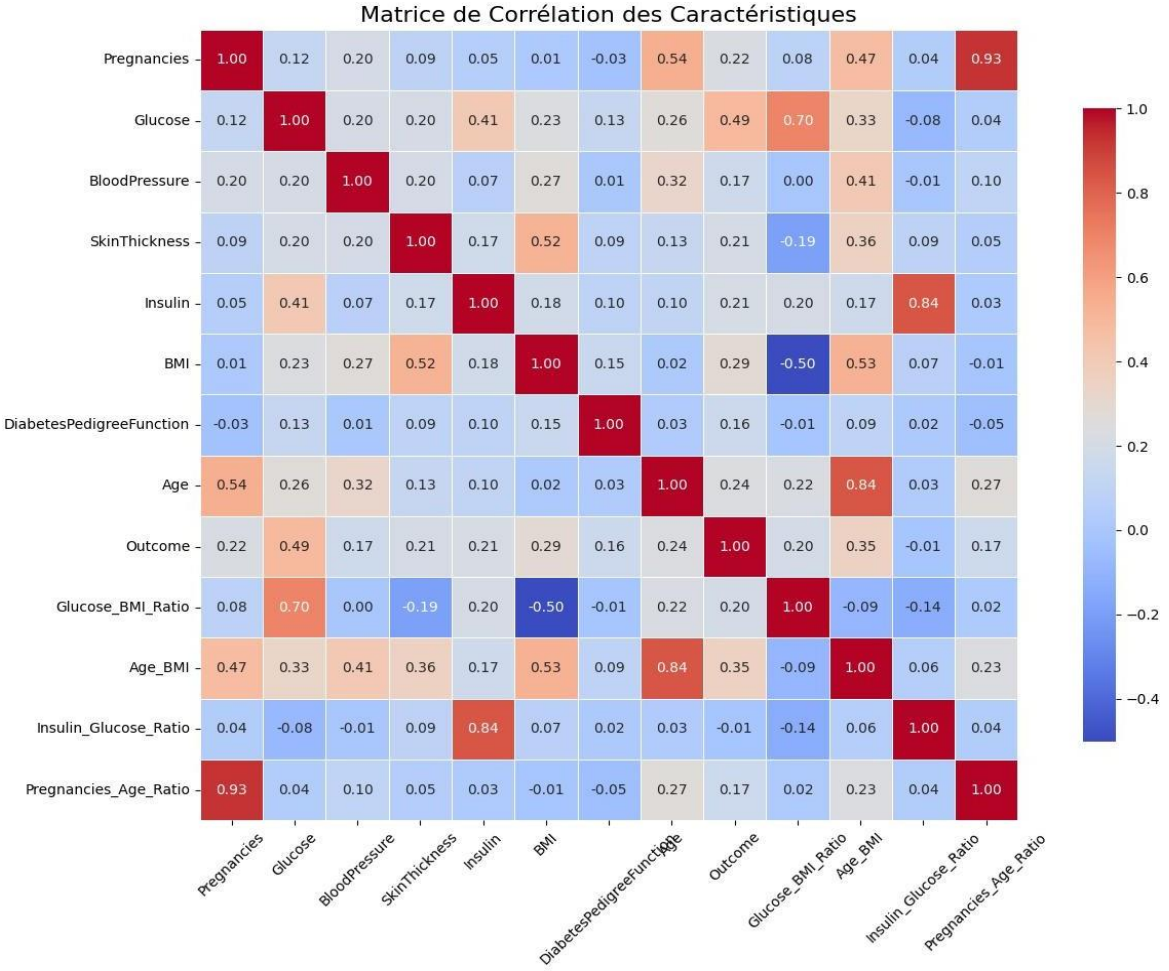


**Figure3.3:** Matrice de corrélation

## 3.5 Description des Algorithmes Utilisés

### 3.6.1 Modèles d'apprentissage supervisé

Ces modèles ont été entraînés sur des datasets(PIMA, Frankfurt) après le prétraitement :

- **Random Forest**

La forêt aléatoire est un ensemble d'arbres de décision combinés pour améliorer la précision et réduire le surapprentissage Le modèle ici comporte n_estimators=50 arbres, une profondeur maximale de 10 (max_depth=10), et des seuils de division fixés à min_samples_split=10 et min_samples_leaf=5 pour éviter la complexité excessive. Ce modèle sert de référence dans la comparaison globale grâce à sa robustesse et sa bonne interprétabilité

- **Extra Trees**

Le modèle Extra Trees (ExtremelyRandomizedTrees) est un algorithme de classification qui utilise plusieurs arbres de décision. Contrairement aux méthodes classiques, il choisit les seuils de division de manière totalement aléatoire, ce qui rend chaque arbre encore plus différent des autres. Cela permet de réduire le risque que le modèle apprenne trop les détails du jeu d'entraînement (surapprentissage).

Ce modèle construit 100 arbres pour prendre des décisions plus précises et stables. Il utilise aussi un paramètre appelé class_weight='balanced', qui aide à mieux gérer les cas où l'une des classes (par exemple : patients diabétiques) est moins représentée que l'autre. Grâce à cela, il fonctionne bien même sur des données compliquées ou avec beaucoup de bruit.

### 3.6.2 Algorithmes de Gradient Boosting

- **XGBoost**

XGBoost (Extreme Gradient Boosting) est un algorithme d'ensemble basé sur le boosting itératif des arbres de décision. Il inclut des mécanismes de régularisation pour améliorer la généralisation. Dans ce code, XGBoost est combiné avec SMOTE pour gérer le déséquilibre des classes. En outre, GridSearchCV est utilisé pour rechercher les meilleurs hyperparamètres (comme learning_rate, max_depth, gamma, etc.) via une validation croisée stratifiée. Cela permet de trouver la meilleure combinaison de paramètres pour maximiser la performance du modèle.

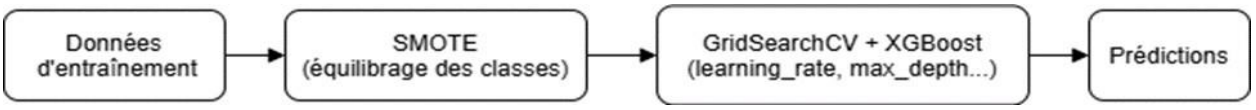


**Figure 3.4:** Architecture de XGBoost

- **LightGBM**

LightGBM est un algorithme de boosting optimisé pour la vitesse d'exécution et l'efficacité mémoire. Il Utilise une méthode de croissance par feuilles plutôt que par niveaux., ce qui permet une meilleure. Ce modèle est configuré avec n_estimators=100, learning_rate=0.1, max_depth=6, et class_weight='balanced' pour gérer le déséquilibre des classes. LightGBM est particulièrement adapté aux grands ensembles de données.

- **CatBoost**

CatBoost est un algorithme de boosting spécialement conçu pour traiter facilement les variables catégorielles, mais il fonctionne également bien sur des données numériques et une imputation automatique des valeurs manquantes. Il inclut également des techniques avancées pour réduire le surapprentissage. Le modèle est configuré avec iterations=500, depth=6,

learning_rate=0.05, et auto_class_weights='Balanced' pour équilibrer les classes. CatBoost est souvent utilisé dans des scénarios complexes avec de nombreuses variables catégorielles. Le modèle a été sauvegardé au format .pkl ainsi que le scaler associé pour une utilisation sur application web.

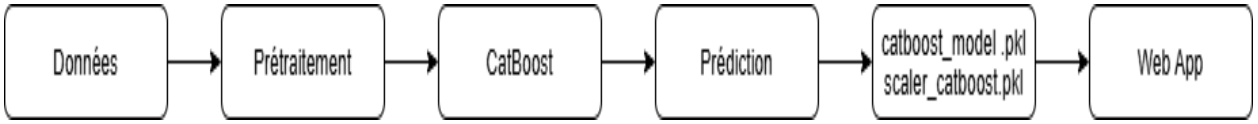


**Figure 3.5:**Architecture de CatBoost

### 3.6.1 Réseau de neurones

- **MLP Classifier**

Le MLP est un réseau de neurones artificiel de type perceptron multicouche (MLP), utilisé pour faire de la classification supervisée. Il fonctionne en apprenant à partir des exemples d'entraînement (X_train_scaled, y_train). Le réseau est composé de couches de neurones : une couche d'entrée, une couche cachée de 100 neurones, et une couche de sortie. Chaque neurone reçoit des données, applique une fonction d'activation ReLU pour introduire de la non-linéarité, et transmet le résultat aux neurones suivants. Il a été entraîné avec l'algorithme d'optimisation Adam pendant 200 époques maximum (itérations) ou jusqu'à ce que le modèle converge. Il a été standardisé avec le même Standard Scaler que les autres modèles.

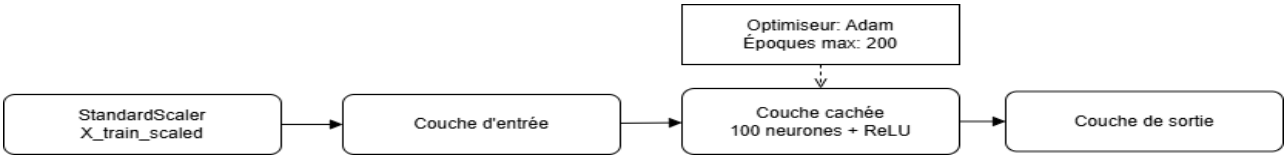


**Figure 3.6:** Architecture de MLP Classifier

- **DNN**

Ce modèle est un réseau neuronal profond implémenté à l'aide de TensorFlow/Keras, comprenant trois couches cachées avec dropout pour limiter le surapprentissage. Les couches sont composées respectivement de 128, 64 et 32 neurones, suivies d'une couche de sortie avec une activation sigmoïde pour la classification binaire. Il est entraîné avec EarlyStopping pour arrêter prématurément l'apprentissage si la performance ne s'améliore plus L'apprentissage a été stabilisé par l'utilisation d'un earlystopping (patience = 25 époques),et une régularisation des poids de classe via class_weight. Des mesures ont également été prises pour garantir la reproductibilité (désactivation du GPU, opérations déterministes).

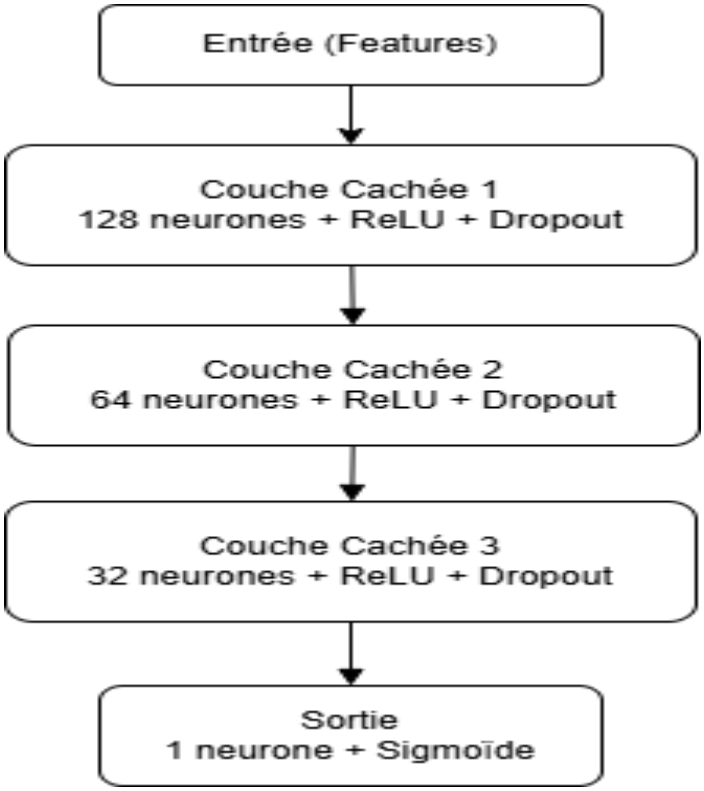


**Figure 3.7 :** Architecture de Deep learning

### 3.6.2 Techniques avancées

- **Stacking(LightGBM + ExtraTrees + CatBoost)**

Le stacking consiste à combiner plusieurs modèles en utilisant un méta-classifieur qui apprend à les agréger de manière optimale. Ici, nous avons empilé LightGBM, ExtraTrees etCatBoost, avec une régression logistique comme méta-modèle. La validation croisée (cv=5) a été utilisée pour éviter le surapprentissage, et les prédictions des modèles de base ont été conservées (passthrough=True). Cette approche permet de tirer parti des forces individuelles de chaque modèle tout en corrigeant leurs erreurs.

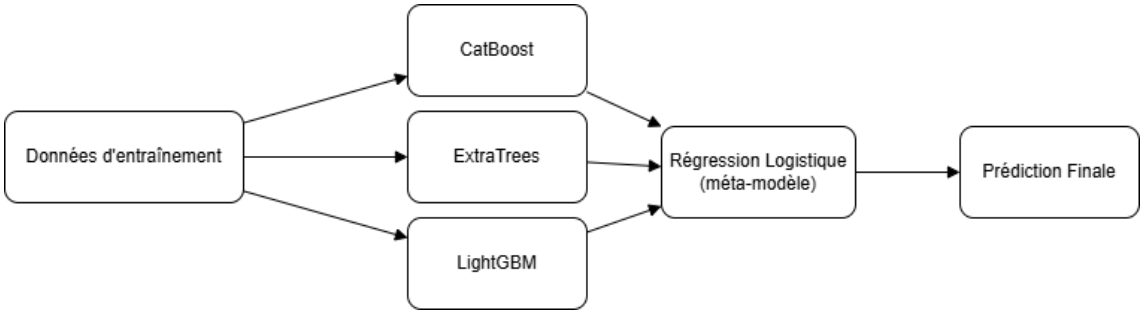


**Figure 3.8:** Architecture de Stacking(CatBoost+ ExtraTrees + LightGBM)

- **Voting**

Combine les prédictions de plusieurs modèles soit en moyenne des probabilités (soft voting)

soit par majorité (hard voting ) avec CatBoost, XGBoost, LightGBM, et ExtraTrees.

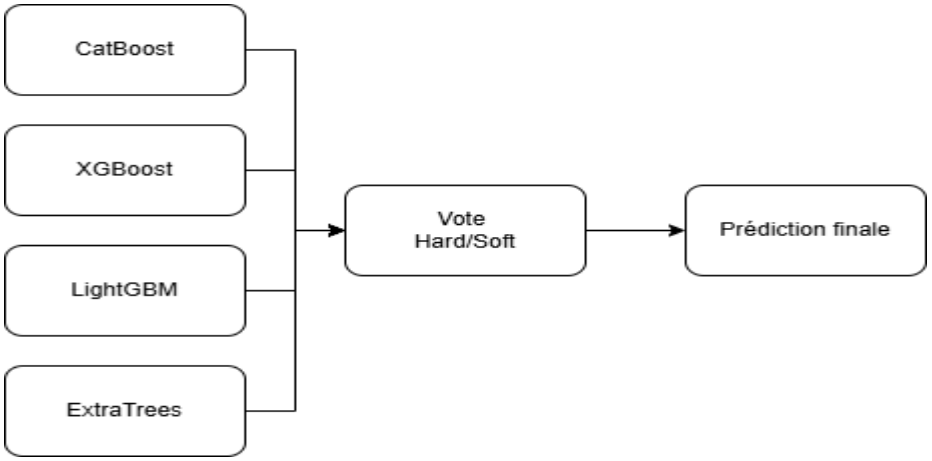


**Figure 3.9:** Architecture de Vote

Il s'agit d'un vote souple (soft voting) combinant quatre modèles (CatBoost, XGBoost, LightGBM et ExtraTrees). Chaque modèle vote en fonction de ses probabilités prédites, et la moyenne pondérée donne la classe finale.

Contrairement au vote souple, le vote dur (hard voting) choisit la classe majoritaire parmi les prédictions de chaque modèle individuel. Cela signifie que seules les étiquettes finales, et non les probabilités, sont prises en compte.

- **Ensemble Pondéré**

Ce modèle combine manuellement les prédictions de CatBoost, XGBoost, LightGBM et ExtraTrees avec des poids attribués à chaque modèle (0.3, 0.3, 0.2, 0.2 respectivement) basés sur leurs performances sur les données de validation. Pour chaque instance, les votes des modèles sont additionnés selon leurs poids et la classe majoritaire est sélectionnée Cette approche personnalisée permet de donner plus d'importance à certains modèles jugés plus fiables.

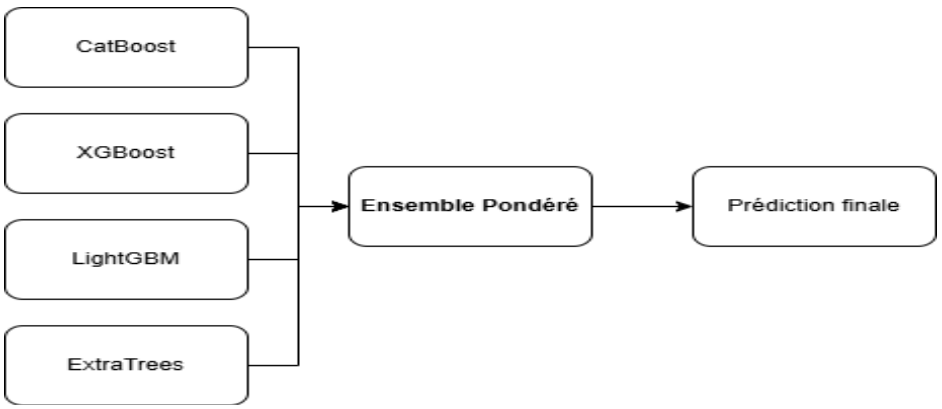


**Figure 3.10:** Architecture d'ensemble pondéré

## 3.6 Paramétrage des Algorithmes

Le paramétrage des algorithmes consiste à configurer les hyperparamètres des modèles d'apprentissage automatique afin d'obtenir les meilleures performances possibles. On utilise plusieurs techniques pour le réglage fin des modèles, notamment l'utilisation de valeurs optimisées via GridSearchCV ou RandomizedSearchCV , la gestion du déséquilibre des classes avec SMOTE , ainsi que des paramètres personnalisés adaptés à chaque modèle.

Le paramétrage des hyperparamètres joue un rôle essentiel dans la performance finale des modèles. Plusieurs stratégies ont été mises en œuvre :

### 3.6.1 Hyper paramètres spécifiques pour chaque modèle

Les principaux paramètres optimisés incluent :

- **Nombre d'arbres (n_estimators) : contrôle la complexité du modèle.**
- **Profondeur maximale (max_depth) : limite la profondeur des arbres pour éviter le surapprentissage.**
- **Learning rate : vitesse d'apprentissage pour les modèles boostés.**
- **Paramètres du noyau SVM (C, gamma) : influence de la régularisation et de la courbure du noyau.**

### 3.6.2 Validation croisée et recherche d'hyperparamètres

Certains modèles ont bénéficié de recherches automatisées :

- **GridSearchCV**: teste toutes les combinaisons possibles d'un ensemble de valeurs.
- **RandomizedSearchCV** : explore aléatoirement l'espace des hyperparamètres pour gagner du temps.

Exemple : Recherche aléatoire pour Stacking (Random Forest + SVM)

### 3.6.3 Gestion du déséquilibre des classes

Plusieurs approches ont été utilisées :

- **SMOTE** : génération synthétique d'échantillons pour la classe minoritaire.
- **class_weight='balanced'** : ajustement automatique des poids.

**Tableau 3.5:** Récapitulation des techniques utilisées

| Modèl | Hyperparamètresprincipaux | Val. croisée | Équilibrageutilisé |
|---|---|---|---|
| **Random Forest** | Arbres = 50, Profondeur = 10 | Non | Aucun |
| **XGBoost** | Arbres = 100-150, Learning rate = 0.01–0.1 | Oui | SMOTE |
| **LightGBM** | Arbres = 100, Profondeur = 6 | Non | class_weight='balanced'. |
| **CatBoost** | Itérations = 500, Learning rate = 0.05 | Non | auto_class_weights='Balance' |
| **MLP** | 1 couche cachée | Non | Aucun |

| | | | |
|---|---|---|---|
| **Deep Learning** | 3 couches cachées<br>Dropout = 0.4 | Non | class_weight |
| **Extra Trees** | Arbres = 100 | Non | class_weight='balanced' |
| **Stacking** | CatBoost+ ExtraTrees<br>+ LightGBM | Oui | Combiné<br>(modèleséquilibrés) |
| **Voting (Soft/Hard)** | Combinaison de 4 modèles | Non | Chaquemodèlegèrel'équilibrage |
| **Ensemble pondéré** | Combinaisonmanuelle | Non | Chaquemodèlegèrel'équilibrage |

## 3.7 Conclusion

En résumé, ce chapitre a exposé de manière détaillée la méthodologie adoptée pour la mise en œuvre de notre système de prédiction du diabète. Nous avons décrit le choix des jeux de données, les étapes de prétraitement, la sélection des caractéristiques pertinentes ainsi que les techniques de gestion du déséquilibre des classes. Plusieurs algorithmes d'apprentissage supervisé ont été entraînés et optimisés, incluant des méthodes avancées comme XGBoost, CatBoost, et un réseau de neurones profond. Afin d'améliorer la robustesse des prédictions, des approches d'assemblage telles que le stacking et les votes pondérés ont été intégrées. Le chapitre suivant sera consacré à l'analyse comparative des résultats obtenus, en évaluant les performances de chaque modèle selon différents indicateurs.

# Chapitre 4 :

# *Implémentation et expérimentation*

## 4.1 Introduction

Ce chapitre présente en détail l'implémentation des différents modèles prédictifs utilisés dans notre étude, ainsi que l'environnement de développement adopté. Il expose les résultats expérimentaux issus de l'entraînement et de l'évaluation des différents algorithmes, en s'appuyant sur des métriques standards telles que la précision, le rappel, le score F1 et la courbe AUC-ROC. Une analyse comparative rigoureuse des performances permet d'identifier les modèles les plus pertinents selon les objectifs de notre étude. Cette analyse éclaire également le choix du modèle final retenu pour être intégré dans l'application web DiaRisk, en tenant compte à la fois de la robustesse, de la généralisation et de l'efficacité computationnelle.

## 4.2 Environnement de développement

### 4.2.1 Langage de programmation

- **Python**

Python est un langage de programmation interprété, multi-paradigme et multiplateformes, réputé pour sa syntaxe claire et sa vaste bibliothèque standard. Nous l'avons choisi pour sa facilité d'utilisation, son écosystème riche pour l'analyse de données (Pandas, NumPy, Scikit-learn) et le deeplearning (TensorFlow/Keras), ainsi que pour sa grande communauté de support. [90]

### 4.2.2 Bibliothèques de Python

- **Numpy**

NumPy est une bibliothèque spécialisée dans la manipulation de tableaux multidimensionnels (ndarray) et propose de nombreuses fonctions optimisées pour les calculs numériques, notamment en flottant. Elle peut être importée directement dans l'environnement courant (fromnumpy import *) ou de manière abrégée (import numpy as np). NumPy est principalement utilisée pour le traitement de vecteurs et de matrices. Les tableaux qu'elle gère sont homogènes, c'est-à-dire constitués d'éléments de même type. La bibliothèque offre également un large éventail de routines facilitant l'accès rapide aux données, leur manipulation ainsi que la réalisation de divers calculs [91].


**Figure 4.1:** Logo de NumPy

- **Pandas**

Pandas est une bibliothèque Python dédiée à la manipulation et à l'analyse de données. Elle permet de gérer aisément des tableaux de données (DataFrames) comportant des étiquettes pour les

variables (colonnes) et les individus (lignes). Pandas facilite également la lecture et l'écriture de ces structures depuis ou vers des fichiers tabulés. De plus, elle permet de générer facilement des représentations graphiques à partir des DataFrames en s'appuyant sur la bibliothèque matplotlib [92].

- **Matplotlib**

Matplotlib est une bibliothèque Python open-source, créée en 2002 par le neurobiologiste John Hunter dans le but de visualiser les signaux électriques du cerveau chez les personnes épileptiques. Son objectif était de reproduire les capacités graphiques de MATLAB en utilisant Python. Cette bibliothèque est particulièrement utile pour les utilisateurs de Python et de NumPy. Elle est fréquemment utilisée sur des serveurs d'application web, dans des shells et dans des scripts Python. Grâce à ses APIs, Matplotlib permet également d'intégrer des graphiques dans des applications à interface graphique [93].

- **Scikit-learn**

Scikit-learn est une bibliothèque open source incontournable dans l'écosystème Python, dédiée à l'analyse de données et à l'apprentissage automatique (machine learning). Elle offre un large éventail d'algorithmes pour diverses tâches de décision, telles que la classification, la régression et le regroupement de données (clustering) [94].

- **Seaborn**

Seabornest une bibliothèque de visualisation de données Python basée sur matplotlib. Elle fournit une interface de haut niveau pour dessiner des graphiques statistiques attrayants et informatifs.[95]

### 4.2.3 Outilsutilisées

- **Visual Studio Code**

visual Studio Code est un éditeur de code source performant, compatible avec Windows, macOS et Linux. Il prend en charge des langages tels que JavaScript, TypeScript et Node.js, et dispose d'un large écosystème d'extensions pour de nombreux autres langages de programmation [96].

- **Jupyter notebook**

Jupyter Notebook est une application web open-source conçue pour créer et partager des documents. Anciennement connu sous le nom de Python Notebook, il constitue un environnement de calcul interactif accessible via un navigateur [97].

- **Google Colab**

Nous avons choisi d'exploiter les capacités de calcul en infonuagique offertes par la plateforme "Google Colab". Celle-ci permet d'accéder à des ressources de calcul accéléré telles que les GPU, TPU et CPU. Son principal atout est la gratuité de ses services, bien que les sessions en ligne soient limitées dans le temps. Par ailleurs, "Google Colab" s'intègre naturellement à

l'environnement "Jupyter Notebook", facilitant ainsi l'exécution de diverses routines de programmation en Python. [98].

## 4.3 Évaluation des performances

Les expériences menées ont permis d'évaluer les performances de plusieurs modèles d'apprentissage automatique appliqués à la prédiction du diabète. Chaque algorithme a été entraîné sur les données prétraitées et optimisé via une validation croisée, en tenant compte du déséquilibre des classes. Les résultats présentés dans les tableaux suivants comparent les performances selon différents critères d'évaluation standard. Cette analyse permet de mesurer l'efficacité de chaque approche et de guider le choix du modèle le plus adapté.

- **Modèles de Machine Learning**

**Tableau 4.1:** Évaluation des performances des Modèles Random Forest et ExtraTrees

| Modèle | Accuracy (Train) | Accuracy (Test) | Précision (Diabetic) | Recall (Diabetic) | F1-Score (Diabetic) | AUC-ROC |
|---|---|---|---|---|---|---|
| **Random Forest** | 98.33 % | 93.86 % | 92.43 % | 89.53 % | 90.96 % | 98.73 % |
| **Extra Trees** | 100% | 99.28% | 98.95% | 98.95% | 98.95% | 99.98% |

- **Algorithmes de Gradient Boosting**

**Tableau 4.2:** Évaluation des performances des Modèles XGBoost , LightGBM et CatBoost

| Modèle | Accuracy (Train) | Accuracy (Test) | Précision (Diabetic) | Recall (Diabetic) | F1-Score (Diabetic) | AUC-ROC |
|---|---|---|---|---|---|---|
| **XGBoost** | 99.77 % | 98.01 % | 95.45 % | 98.95 % | 97.17 % | 99.89% |
| **LightGBM** | 99.68 % | 98.38 % | 95.50 % | 100% | 97.70 % | 99.95% |
| **CatBoost** | 100% | 99.28 % | 98.45 % | 99.48 % | 98.96 % | 99.97% |

- **Reseaux de neurones**

**Tableau4.3:** Évaluation des performances des Modèles MLP et Deep Learning

| Modèle | Accuracy (Train) | Accuracy (Test) | Précision (Diabetic) | Recall (Diabetic) | F1-Score (Diabetic) | AUC-ROC |
|---|---|---|---|---|---|---|
| **MLP (Perceptron Multi-Couches)** | 87.44 % | 84.12 % | 77.84 % | 75.39 % | 76.60 % | 91.95 % |
| **Deep Learning (Keras)** | 97.52% | 94.04% | 89.50% | 93.72% | 91.56% | 98.68% |

### Techniques avancées

**Tableau 4.4:**Évaluation des performances des Modèles Stacking , Voting et Ensemble Pondéré

| Méthode | Accuracy (Train) | Accuracy (Test) | Précision (Diabetic) | Recall (Diabetic) | F1-Score (Diabetic) | AUC-ROC |
|---|---|---|---|---|---|---|
| **Stacking (LightGBM +ExtraTrees+Cat Boost)** | 100% | 99.10 % | 98.44 % | 98.95 % | 98.69 % | 99.99% |
| **Soft voting** | 99.82% | 98.01% | 97.87% | 96.34% | 97.10% | 99.90% |
| **Hard voting** | 99.91% | 98.74 % | 98.42 | 97.91% | 98.16% | ___ |
| **Ensemble Pondéré** | 99.95% | 99.28% | 98.45% | 99.48% | 98.96% | 99.32% |

## 4.4 Courbes ROC et matrices de confusion

### 4.4.1 Modèles de Machine Learning

#### Random Forest

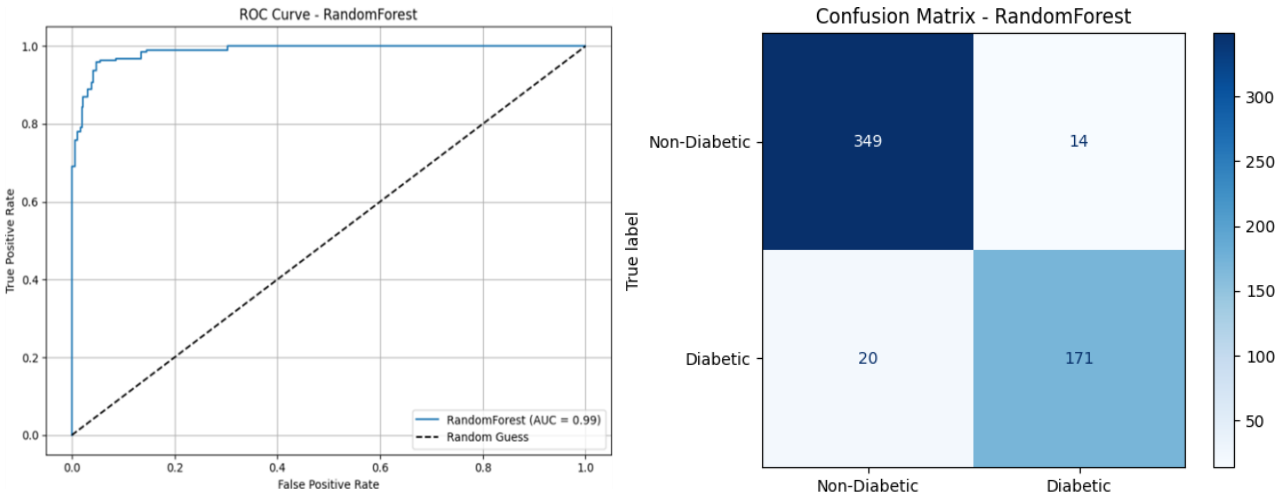


**Figure 4.2**: Évaluation Random Forest ROC et Matrice de Confusion

La figure 4.2 montre que le modèle Random Forset est très bon avec une courbe ROC (AUC = 0.99) et peu d'erreurs dans la matrice de confusion : 14 non-diabétiques prédits diabétiques et 20 diabétiques prédits non-diabétiques.

### Extra Trees

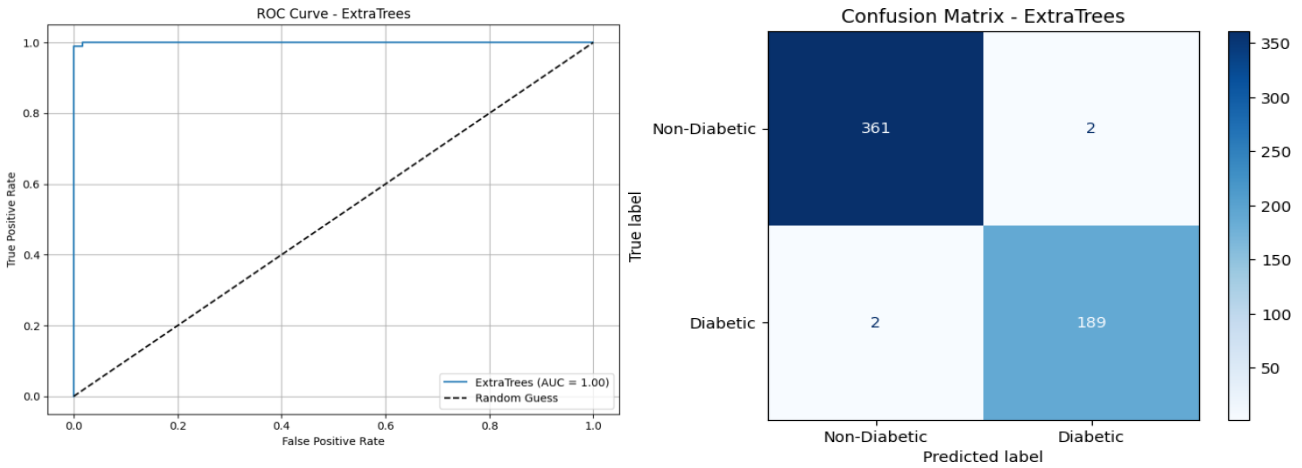


**Figure 4.3 :** Évaluation Extra Trees ROC et Matrice de Confusion

La figure 4.3 montre que le modèle Extra Trees est très bon avec une courbe ROC (AUC = 1.00) et peu d'erreurs dans la matrice de confusion : 2 non-diabétiques prédits diabétiques et 2 diabétiques prédits non-diabétiques

## 4.4.2 Algorithmes de Gradient Boosting

### XGBoost

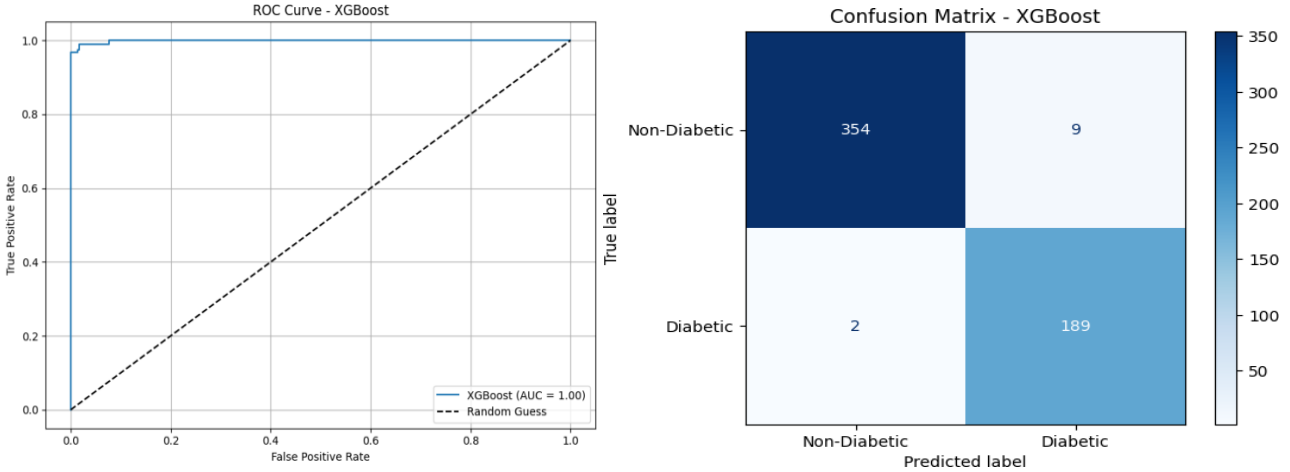


**Figure 4.4:** Évaluation XGBoost ROC et Matrice de Confusion

La figure 4.4 montre que le modèle XGBoost est très bon avec une courbe ROC (AUC = 1.00) et peu d'erreurs dans la matrice de confusion : 9 non-diabétiques prédits diabétiques et 2 diabétiques prédits non-diabétiques.

**LightGBM**

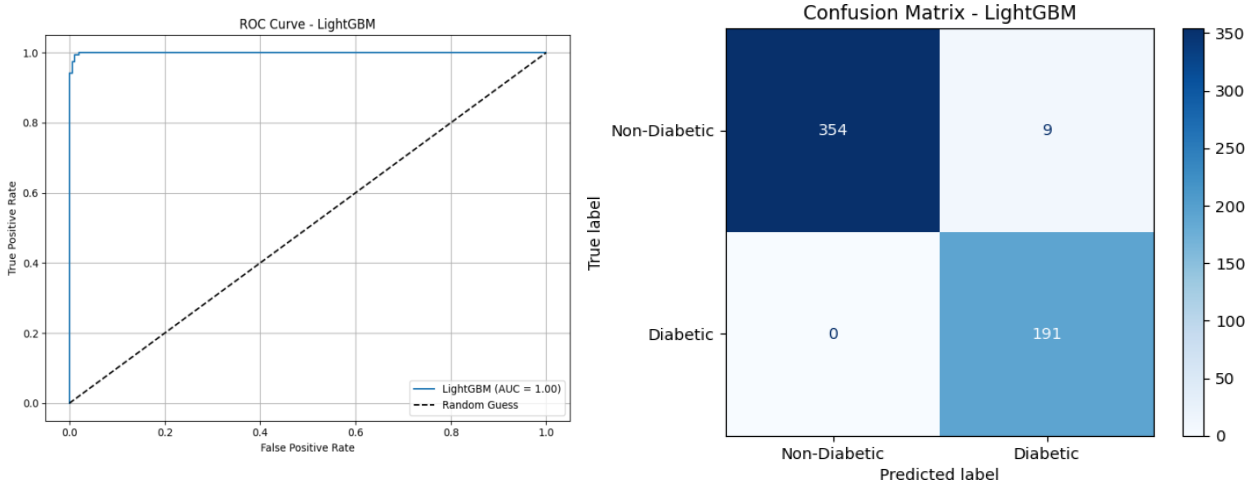


**Figure 4.5:**Évaluation LightGBM ROC et Matrice de Confusion

La figure 4.5 montre que le modèle LightGBM est très bon avec une courbe ROC (AUC = 1.00) et peu d'erreurs dans la matrice de confusion : 9 non-diabétiques prédits diabétiques et 0 diabétiques prédits non-diabétiques.

**CatBoost**

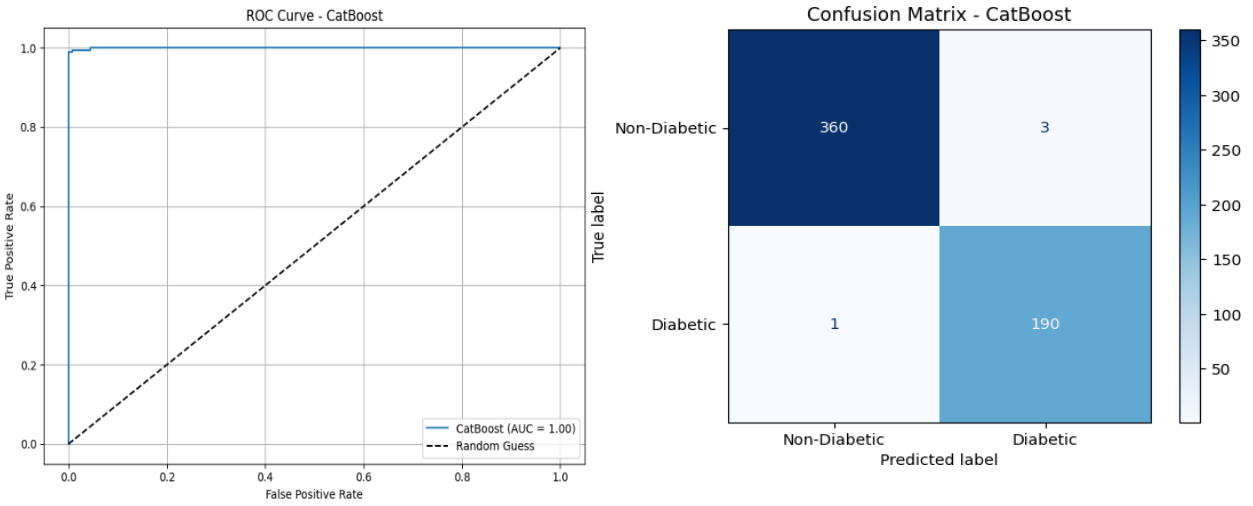


**Figure 4.6:**Évaluation CatBoost ROC et Matrice de Confusion

La figure 4.6 montre que le modèle CatBoost est très bon avec une courbe ROC (AUC = 1.00) et peu d'erreurs dans la matrice de confusion: 3 non-diabétiques prédits diabétiques et 1 diabétiques prédits non-diabétiques.

### 4.4.3 Réseau de neurones

**MLPClassifier**

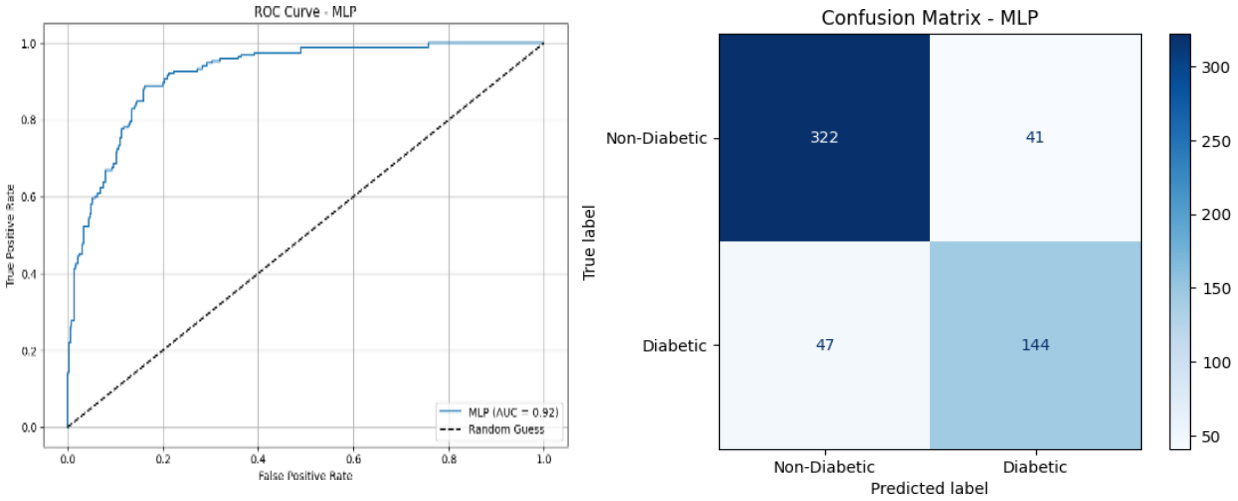


**Figure 4.7:**Évaluation MLP ROC et Matrice de Confusion

La figure 4.7 montre que le modèle MLP est très bon avec une courbe ROC (AUC = 0.92) et peu d'erreurs dans la matrice de confusion : 41 non-diabétiques prédits diabétiques et 47 diabétiques prédits non-diabétiques.

**Deep Learning**

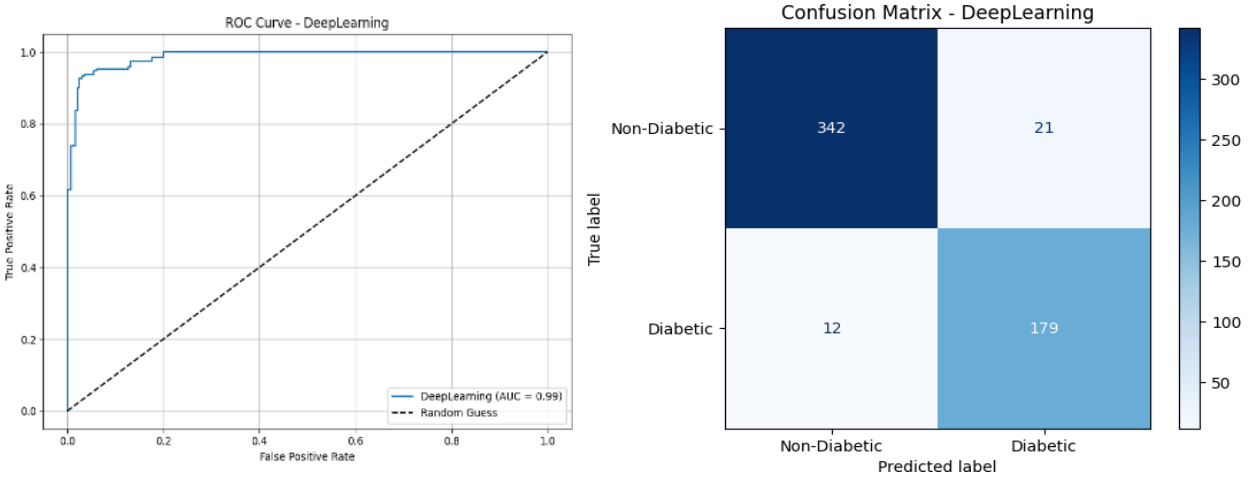


**Figure 4.8:** Évaluation Deep Learning ROC et Matrice de Confusion

La figure 4.8 montre que le modèle est très bon avec une courbe ROC (AUC = 0.99) et peu d'erreurs dans la matrice de confusion : 20 non-diabétiques prédits diabétiques et 12 diabétiques prédits non-diabétiques.

### 4.4.4 Techniques avancées

**Stacking(LightGBM + ExtraTrees + CatBoost)**

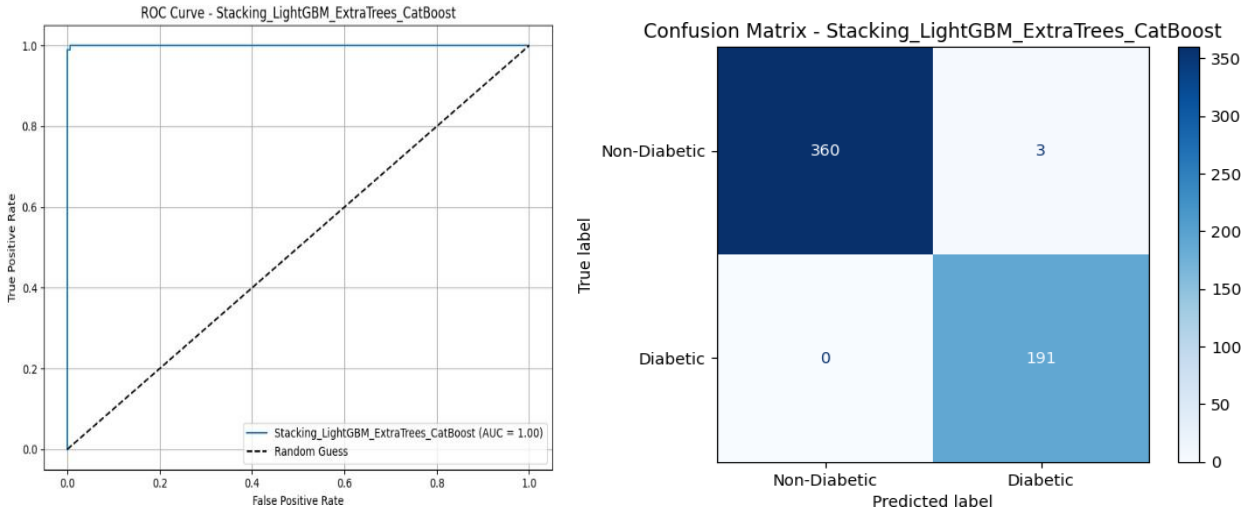


**Figure 4.9:**Évaluation Stacking ROC et Matrice de Confusion

La figure 4.9 montre que le modèle Stacking est très bon avec une courbe ROC (AUC = 1.00) et peu d'erreurs dans la matrice de confusion : 3 non-diabétiques prédits diabétiques et 0 diabétiques prédits non-diabétiques.

**Ensemble Pondéré**

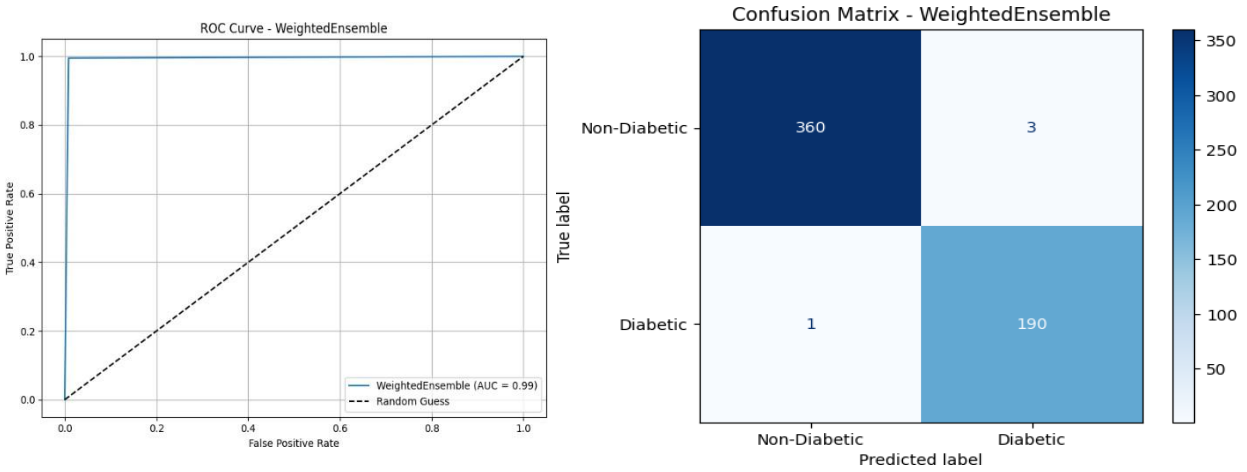


**Figure 4.10:**Évaluation Ensemble Pondéré ROC et Matrice de Confusion

La figure 4.10 montre que le modèle Ensemble Pondéré est très bon avec une courbe ROC (AUC = 0.99) et peu d'erreurs dans la matrice de confusion : 3 non-diabétiques prédits diabétiques et 1 diabétiques prédits non-diabétiques.

**Soft voting**

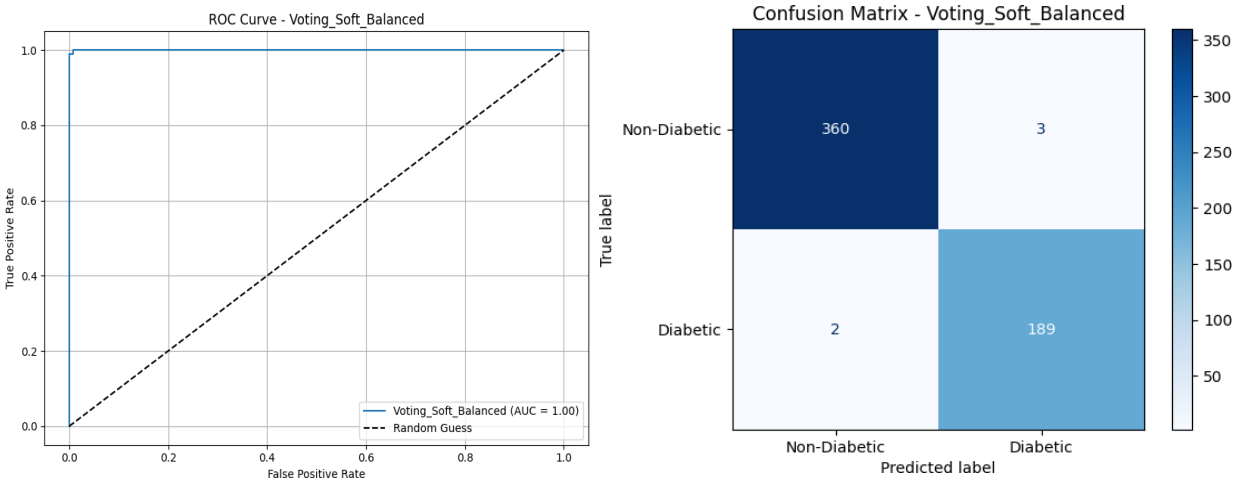


**Figure 4.11:**Évaluation de Soft Voting ROC et Matrice de Confusion

La figure 4.11 montre que le modèle Soft Voting est très bon avec une courbe ROC (AUC = 1.00) et peu d'erreurs dans la matrice de confusion : 3 non-diabétiques prédits diabétiques et 2 diabétiques prédits non-diabétiques.

**Hard voting**

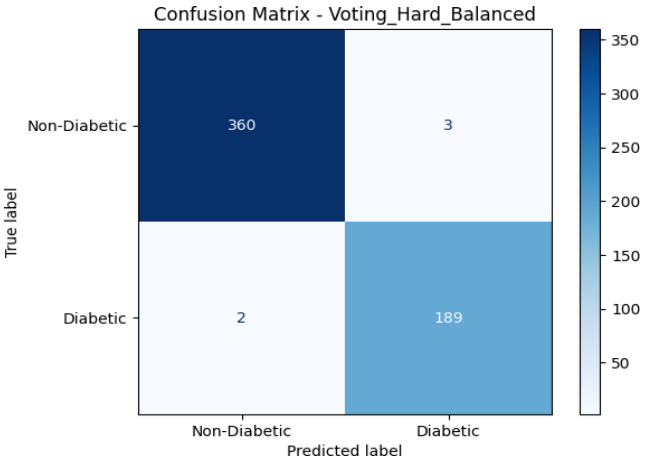


**Figure 4.12 :** Évaluation de Hard Voting ROC et Matrice de Matrice de Confusion

La figure 4.12 montre que le modèle Hard Voting est très bon avec peu d'erreurs dans la matrice de confusion : 3 non-diabétiques prédits diabétiques et 2 diabétiques prédits non-diabétiques.

## 4.5 Analyse et discussion des résultats

### 4.5.1 Comparaison des performances selon les algorithmes

Les résultats obtenus montrent que certains modèles sont bien meilleurs que d'autres pour prédire le diabète. Le MLP (Multi-Layer Perceptron) donne les moins bons résultats avec une précision de 0.8744 et un rappel de 0.7539, ce qui signifie qu'il se trompe assez souvent. En revanche, les modèles comme CatBoost, ExtraTrees, LightGBM et XGBoost donnent de très bons résultats, avec des précisions supérieures à 0.95 et des rappels très élevés, ce qui montre qu'ils arrivent à bien détecter les cas de diabète.

Parmi eux, CatBoost est le meilleur modèle avec une précision de 0. 9977, un rappel de 0.9948 et un score AUC-ROC presque parfait (0.9998). On va utiliser cette model dans notre application web. Il est suivi de près par ExtraTrees et les modèles combinés comme Voting_Soft_Balanced et Stacking XGBoost + ExtraTrees + CatBoost, qui offrent aussi de très bonnes performances.

Un point positif est que ces modèles gardent une bonne précision aussi bien sur les données d'entraînement que sur les données de test, ce qui veut dire qu'ils ne font pas de surapprentissage. Leurs scores AUC-ROC très élevés prouvent qu'ils distinguent très bien les personnes diabétiques des non-diabétiques.En revanche, le modèle DeepLearning, même s'il est bon (F1-score de 0.9156), est un peu moins performant que les modèles de type boosting (comme CatBoost). De plus, les modèles combinés sont plus complexes à utiliser dans la réalité, car ils demandent plus de calcul et sont plus difficiles à mettre en place.

En conclusion, les modèles de type boosting (comme CatBoost ou LightGBM) sont les plus efficaces pour ce problème. Ils sont à la fois précis, rapides et fiables pour détecter le diabète dans ce dataset.

### 4.5.2 Comparaison entre pima et frunkfurt et datasetscombinees

**Tableau 4.5:** Comparaison entre pima et Frankfurt et datasets combinées

| Métrique (Test Set) | Pima (Meilleur Modèle ) | Frankfurt (Meilleur Modèle ) | Combinaison (Meilleur Modèle) |
|---|---|---|---|
| **Accuracy** | 0.7597 | 0.9850 | 0.9928 |
| **F1 Score (Diabetic)** | 0.6833 | 0.9781 | 0.9896 |
| **AUC-ROC** | 0.8259 | 0.9993 | 0.9999 |
| **Taille (Test Set)** | 154 | 400 | 554 |

L'analyse des résultats montre que le dataset combiné est le plus performant. Grâce à un volume de données plus important, il permet d'obtenir la meilleure précision, le meilleur F1 Score et la meilleure AUC-ROC parmi les trois jeux de données. Cette richesse en données améliore

l'apprentissage des modèles ce qui les aide à mieux détecter le diabète.

## 4.6 Shape /Lime

### 4.6.1 SHAP

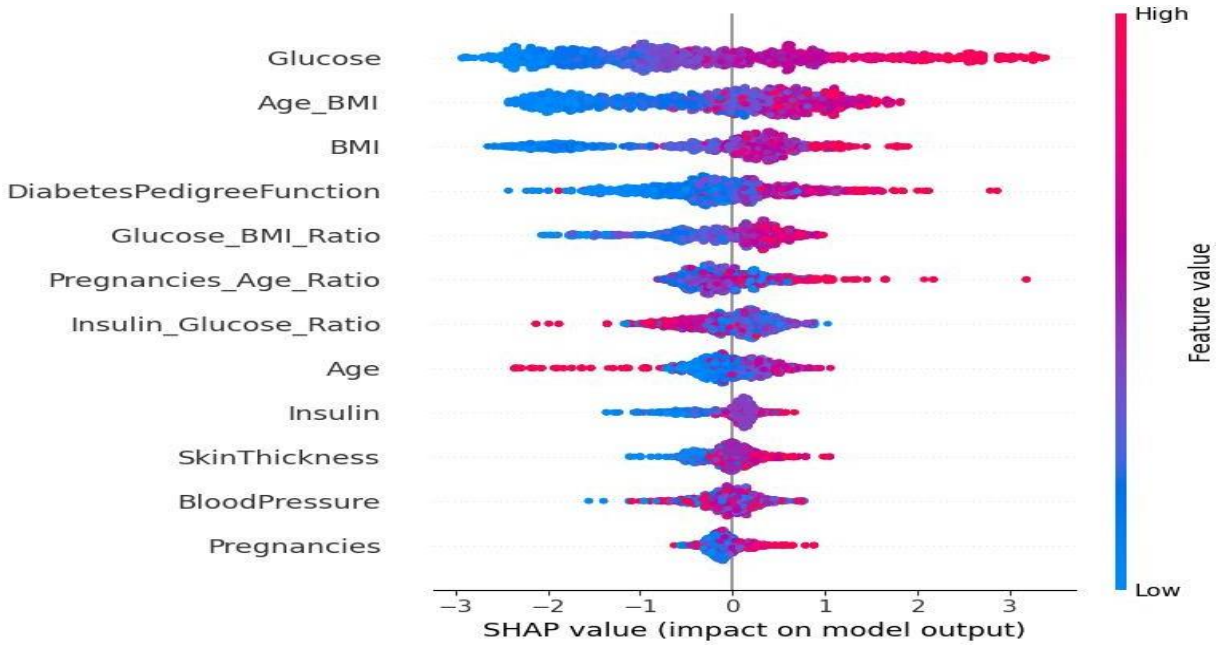


**Figure 4.13:** Importance globale des caractéristiques pour CatBoost

Le graphique récapitulatif SHAP (Figure 4.13) montre comment chaque caractéristique influence les prédictions du modèle CatBoost. Chaque point représente une observation, où sa position horizontale indique l'impact de la caractéristique sur la prédiction (vers le haut ou vers le bas), et sa couleur (bleu à rouge) reflète la valeur de la caractéristique (basse à élevée). Les caractéristiques sont classées par leur importance moyenne sur les prédictions. Les trois caractéristiques les plus influentes sont Glucose , Age_BMI et BMI . Pour ces variables : Des valeurs élevées (points rouges) poussent généralement la prédiction vers la droite, ce qui augmente la probabilité prédite. À l'inverse, des valeurs basses (points bleus) réduisent la prédiction en la déplaçant vers la gauche, diminuant ainsi cette probabilité. Cela suggère une relation positive entre ces caractéristiques et la variable cible: plus leurs valeurs augmentent, plus la prédiction est forte.

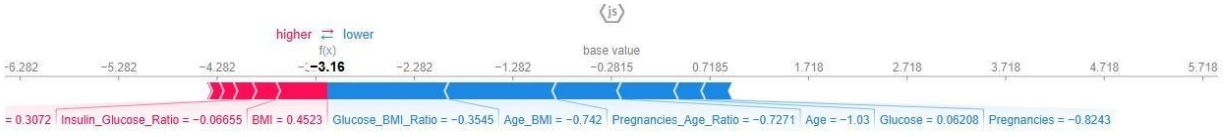


**Figure 4.14:**Explication individuelle d'une prédiction

Ce graphique (Figure 4.14) explique la prédiction pour une seule instance. Il part de la prédiction du modèle et montre comment chaque caractéristique (rouge pour pousser vers le haut, bleu vers le bas) contribue à arriver au résultat final prédit pour ce patient.

### 4.6.2 LIME

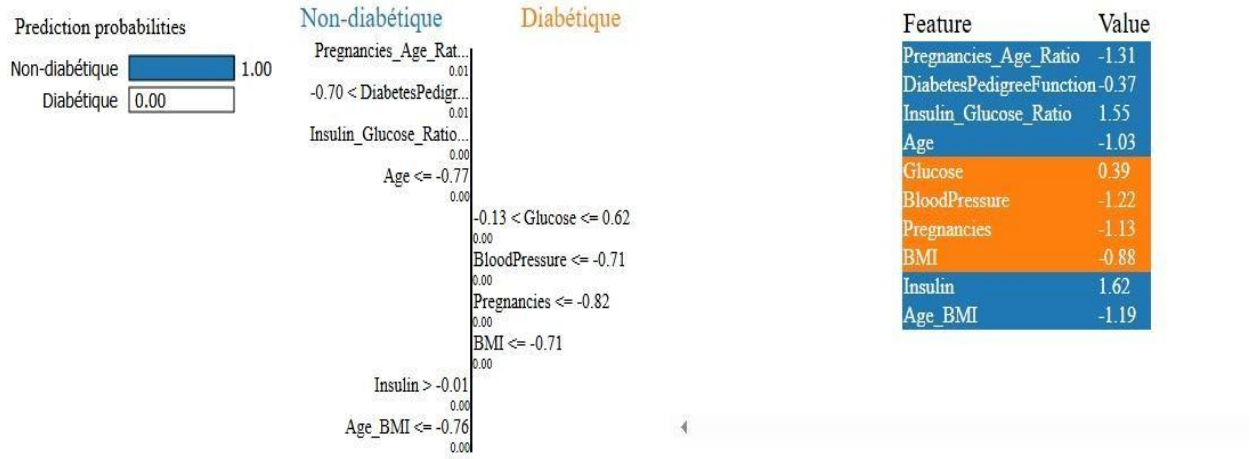


**Figure 4.15:** Explication individuelle avec LIME

Cette visualisation LIME [Figure 4.15] explique pourquoi le modèle a prédit "Non-diabétique" (probabilité 1.00) pour ce patient. Le graphique montre comment la valeur de chaque caractéristique pousse la prédiction : vers "Non-diabétique" à gauche ou vers "Diabétique" à droite. Ici, un faible Pregnancies_Age_Ratio et un haut Insulin_Glucose_Ratio sont des facteurs clés qui ont fortement influencé la décision vers le "Non-diabétique".

## 4.7 Application web

### 4.7.1 Explication Générale de l'Application

L'application web DiaRisk, développée avec Dash (Python) qui permet aux professionnels de santé évaluer le risque de diabète des patients à partir de données cliniques. Elle utilise le modèle CatBoost qui est entraîné pour prédire un risque (élevé/faible) avec probabilité, fournit des considérations cliniques personnalisées, et on peut génère des rapports PDF. Son interface est multilingue (anglais, français, arabe), propose des thèmes clair/sombre, et la mise en page est entièrement réactive, s'adaptant à différentes tailles d'écran pour une utilisation sur ordinateurs de bureau, tablettes et mobiles.

### 4.7.2 Développent d'application

L'application est développée en Python et utilise le frameworkDash pour construire l'interface web interactive.

- **Backend**

**Langage :** Python

**Framework Web :** Dash (utilisant Flask)

Bibliothèques principales :

**Pandas et NumPy :** Pour la manipulation et le traitement des données (préparation des données pour le modèle).

**CatBoost et Scikit-learn :** Pour le modèle d'évaluation du risque (modèle CatBoost spécifique) et les outils de prétraitement des données (notamment StandardScaler de Scikit-learn) utilisés avant la prédiction.

**Joblib :** Pour le chargement efficace du modèle et du scaler sauvegardés au format .pkl.

**ReportLab :** Pour la génération de rapports PDF personnalisés.

**ArabicReshaper et Python-Bidi :** Bibliothèques Python pour le support correct de l'affichage du texte arabe (gestion de la mise en forme des caractères et de la direction du texte) dans les PDF générés par ReportLab.

**Stockage de l'historique :** Fichiers CSV locaux (predictions_log.csv) pour stocker un journal des prédictions passées.

- **Frontend**

**Framework:**Dash (génère dynamiquement l'interface utilisateur en HTML, CSS et JavaScript).

**Composants d'Interface Utilisateur:** DashBootstrap Components (DBC) pour fournir des composants UI stylisés, réactifs et basés sur Bootstrap (cartes, boutons, inputs, etc.).

**Visualisation Graphique:**Plotly (intégré via le composant dcc.Graph de Dash) pour créer les graphiques interactifs comparant les entrées patient aux plages de référence.

**Icônes :** Bibliothèque d'icônes Font Awesome

**Logique Côté Client:** Callbacks clientside (exécutés directement en JavaScript dans le navigateur) pour des interactions rapides comme le calcul de l'âge à partir de la date de naissance ou la mise à jour des classes CSS du corps du document.

**Gestion de l'état UI (thème, langue, état de la sidebar):** Composants dcc.Store de Dash, utilisant le stockage local du navigateur pour persister ces paramètres entre les sessions.

**Styles CSS:** Utilisation de thèmes Bootstrap externes (BOOTSTRAP et DARKLY), ainsi que du CSS personnalisé intégré dans app.index_string pour des styles spécifiques et des animations.

### 4.7.2 Diagramme de l'Application

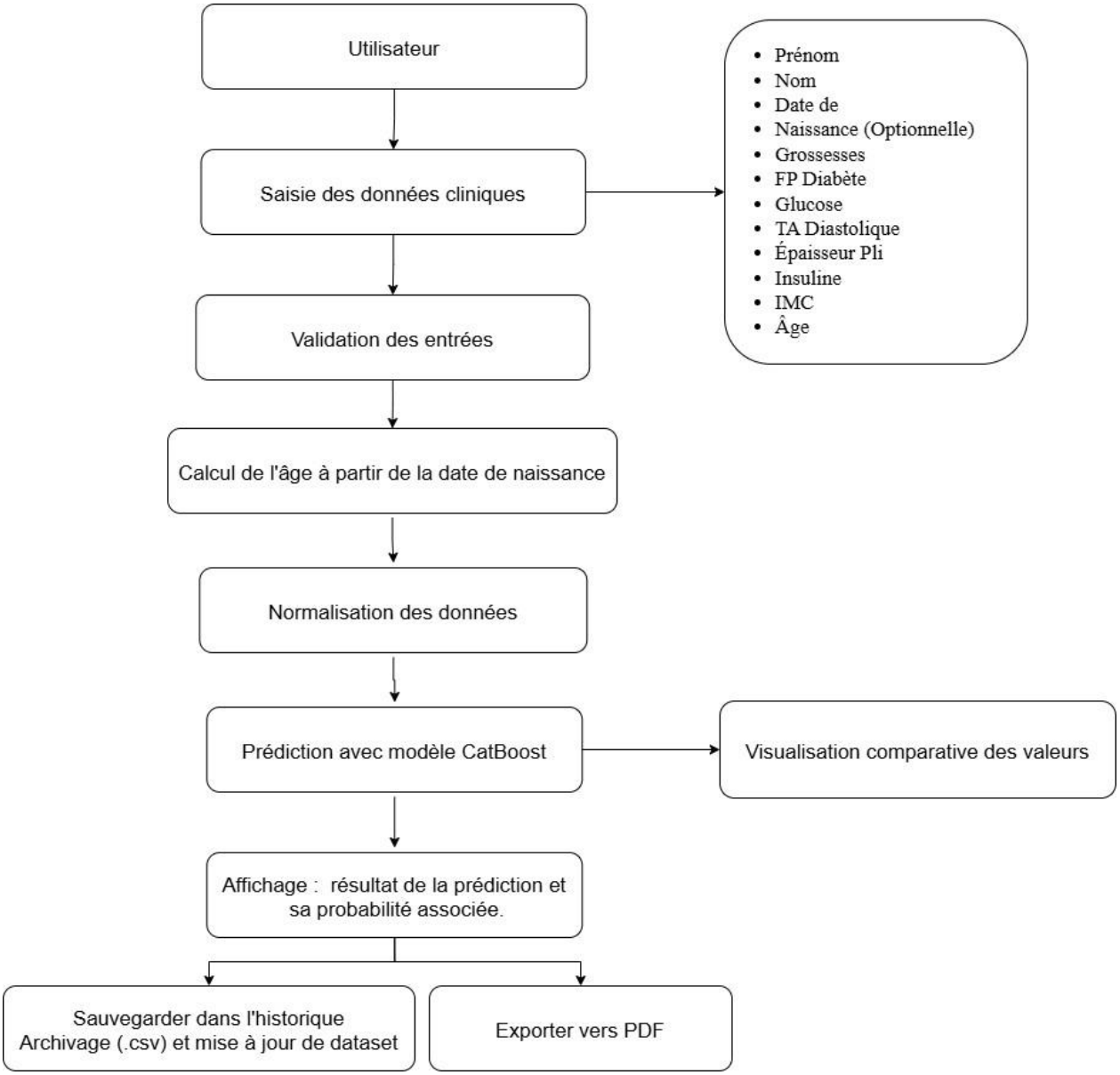


**Figure 4.16:** Diagramme de l'application DiaRisk réalisée

- **Page d'accueil**

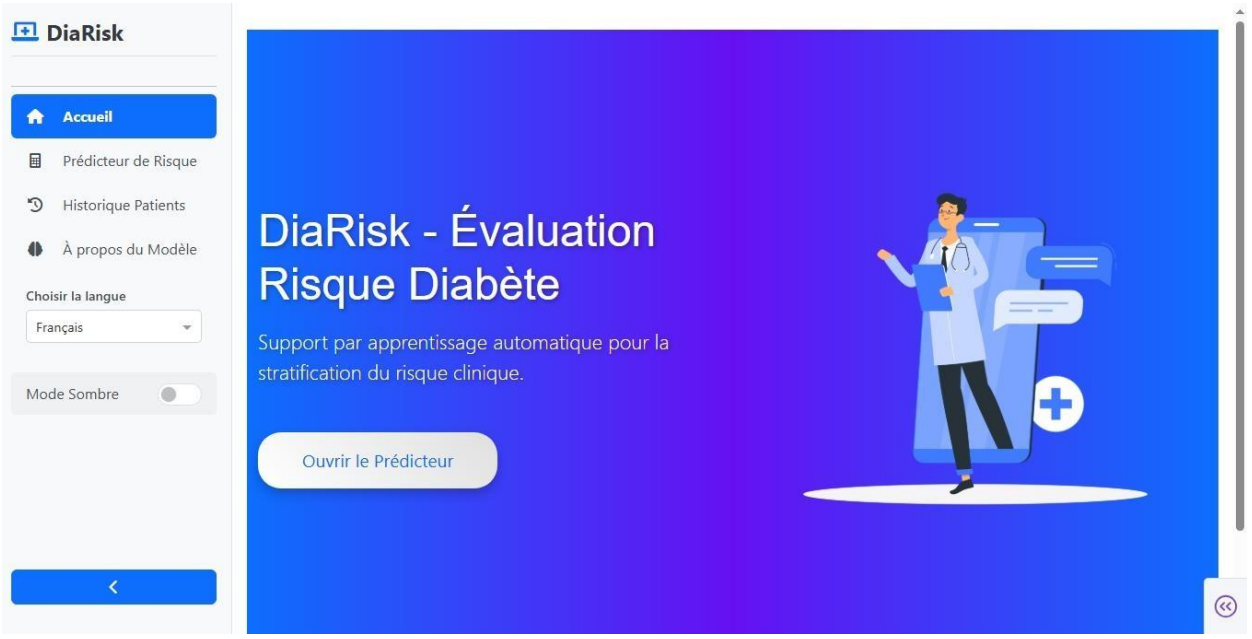


**Figure 4.17:** Page d'accueil

La page d'accueil présente l'application DiaRisk - Diabetes Risk Assessment et son rôle de support à l'évaluation clinique du risque diabétique. Elle permet à l'utilisateur d'accéder au la page de prédiction via un bouton dédié.

- **Page de prédiction**

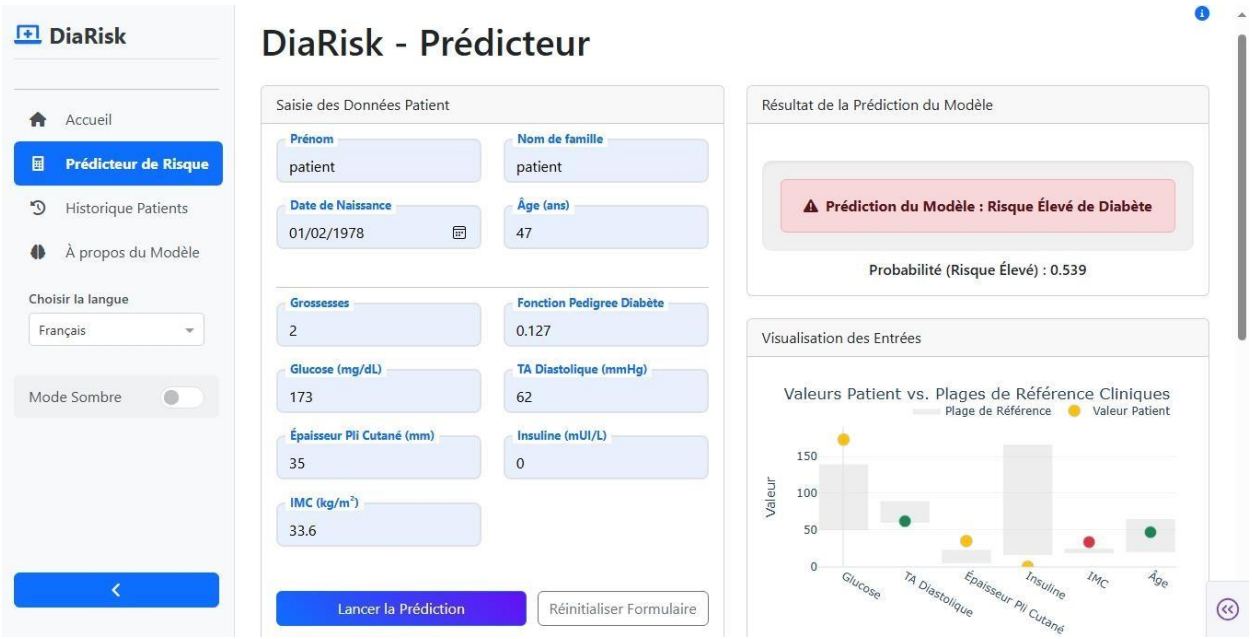


**Figure 4.18**: Page de prédiction

La page de prédiction permet aux professionnels de santé de saisir les données cliniques d'un patient (Prénom, Nom, Date de Naissance (Optionnelle), Grossesses, FP Diabète, Glucose, TA Diastolique, Épaisseur Pli, Insuline, IMC, Âge) pour évaluer le risque de diabète via le modèle entraîné CatBoost via bouton "Lancer la Prédiction. Elle affiche le résultat prédit (faible/élevé), la probabilité associée, des considérations cliniques basées sur les inputs, et une visualisation graphique des valeurs patient et permet de générer un rapport PDF complet des résultats via bouton "Enregistrer Rapport (PDF) et on peut efface toutes les données saisies dans les champs d'entrée et réinitialise les résultats via bouton "Réinitialiser Formulaire".

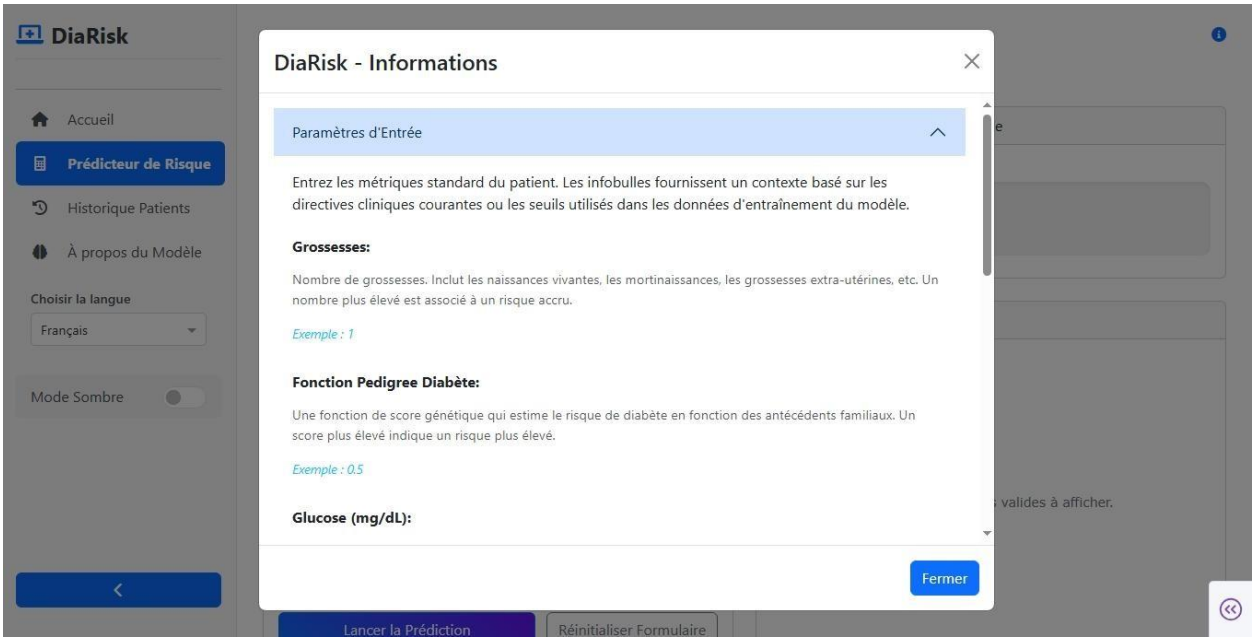


**Figure 4.19:** Page d' information

Sur la page de prédiction, des infobulles s'affichent sur les étiquettes des champs de saisie. Elles expliquent chaque valeur en fournissant un contexte clinique rapide, tel que les plages normales, les seuils critiques, les unités et des exemples pour aider à la saisie et à l'interprétation.

**DiaRisk - Rapport d'Analyse Médicale**

Rapport généré le : 2025-06-03 12:29:26

***Données d'Entrée Patient***

| Paramètre | Valeur | Unité |
|---|---|---|
| Prénom | patient | |
| Nom | patient | |
| Date de Naissance | 1978-01-02 | |
| Grossesses | 2 | |
| FP Diabète | 0.13 | |
| Glucose | 173 | mg/dL |
| TA Diastolique | 62 | mmHg |
| Épaisseur Pli Cutané | 35 | mm |
| Insuline | 0 | mUI/L |
| IMC | 33.60 | kg/m² |
| Âge | 47 | ans |

***Résultat Prédiction Modèle***

**Prédiction du Modèle : Risque Élevé de Diabète**

Probabilité (Risque Élevé) : 0.539

***Considérations Cliniques Basées sur les Entrées***

- **Glucose (173.0 mg/dL, soit env. 9.6 mmol/L)** : Tolérance au glucose altérée (prédiabète). Risque accru de diabète.
- **TA Diastolique (62 mmHg)** : Pression artérielle diastolique dans la plage normale.
- **Épaisseur Pli Cutané (35 mm)** : Épaisseur élevée (>23 mm), indicateur d'adiposité excessive.
- **Insuline (0 mUI/L)** : Valeur d'insuline basse (2h post-glucose), nécessite une évaluation clinique.
- **IMC (33.6 kg/m²)** : Obésité (risque élevé de diabète).
- **Âge (47 ans)**: Âge adulte.

Cette application ne remplace pas un avis médical. Destinée aux professionnels de santé qualifiés.

**Figure 4.20:**Rapport de prédiction délivré en Pdf

Après une prédiction réussie, l'application permet de générer un rapport PDF téléchargeable. Ce document résume les données patient saisies, le résultat de la prédiction du modèle, les considérations cliniques pertinentes, et est formaté dans la langue sélectionnée.

- **Page Historique des Patients**


**Figure 4.21:** Page Historique des patients

La page "Historique des Patients" permet de consulter un tableau affichant les résultats des évaluations de risque enregistrées précédemment. Ces informations sont stockées localement dans

un fichier CSV (predictions_log.csv). Pour chaque entrée, le tableau affiche les données d'entrée du patient, le résultat de la prédiction et sa probabilité associée.

- **Page À propos du Modèle**

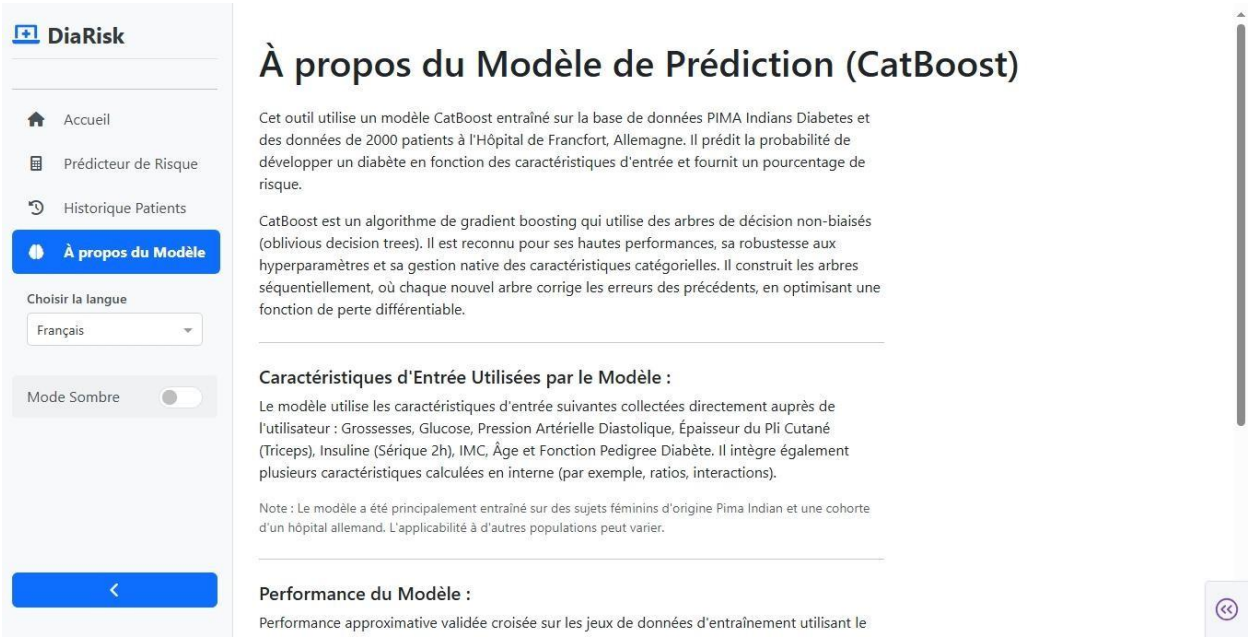


**Figure 4.22:** Page À propos du Modèle

La page "À propos du Modèle" détaille le modèle CatBoost utilisé pour la prédiction : ses données d'entraînement, les 12 caractéristiques d'entrée qu'il utilise, ses performances approximatives (Accuracy, AUC), et les notes importantes concernant son applicabilité et son utilisation clinique.

## 4.8 Conclusion

En conclusion, les expérimentations ont montré que les modèles de type Gradient Boosting, en particulier CatBoost et LightGBM, surpassent largement les approches classiques et les réseaux de neurones simples dans la détection du diabète. Parmi eux, CatBoost se distingue comme le modèle le plus performant, avec une précision, un rappel et une AUC-ROC remarquablement élevés.

Par ailleurs, la combinaison des datasets Pima et Frankfurt s'est révélée bénéfique, améliorant considérablement la qualité des prédictions grâce à une plus grande diversité et quantité de données.

Enfin, l'intégration du meilleur modèle dans l'application **DiaRisk** permet une évaluation fiable du risque de diabète, tout en offrant une interface simple et accessible aux professionnels de santé. Cette solution allie performance algorithmique et utilité clinique, répondant efficacement aux objectifs fixés dans ce projet.

# Conclusion générale

Au terme de ce mémoire de master, il est clair que l'apprentissage automatique joue un rôle crucial dans la prédiction et la détection précoce du diabète. Face à l'augmentation constante du nombre de cas à travers le monde, l'utilisation de techniques avancées d'intelligence artificielle s'impose comme une solution efficace et prometteuse pour améliorer la prise en charge médicale.

Au terme de ce mémoire de master, il apparaît clairement que l'apprentissage automatique constitue un levier puissant dans la prédiction et la détection précoce du diabète. Face à la progression alarmante de cette maladie chronique à l'échelle mondiale, l'intégration de techniques avancées d'intelligence artificielle se révèle non seulement pertinente, mais aussi prometteuse pour améliorer la prise en charge des patients et soutenir les décisions cliniques.

Ce travail a permis d'expérimenter et de comparer plusieurs algorithmes d'apprentissage supervisé, notamment Random Forest, ExtraTrees, ainsi que des modèles de boosting tels que XGBoost, LightGBM et CatBoost, sans oublier les réseaux de neurones artificiels. L'évaluation rigoureuse de leurs performances à l'aide d'indicateurs comme la précision, le rappel, le F1-score et l'AUC-ROC a mis en évidence la supériorité des modèles d'ensemble (Stacking, Voting soft/hard, Ensemble pondéré), en particulier dans des contextes de déséquilibre de classes. Ces modèles se sont distingués par leur robustesse, leur capacité de généralisation et leur aptitude à produire des prédictions fiables.

Un apport notable de ce travail réside également dans l'intégration de méthodes d'explicabilité telles que SHAP et LIME, qui rendent les décisions des modèles plus compréhensibles pour les professionnels de santé. Cette transparence est essentielle pour favoriser l'acceptation et l'intégration des outils d'intelligence artificielle en contexte clinique, où la confiance et la lisibilité des résultats sont cruciales.

Par ailleurs, la mise en œuvre de l'application web DiaRisk, développée avec Dash (Python), constitue une concrétisation fonctionnelle de ce mémoire. Cette application propose une interface intuitive destinée aux médecins, permettant d'évaluer le risque diabétique à partir des meilleurs modèles entraînés. Elle intègre des fonctionnalités avancées telles que la gestion multilingue, la génération de rapports PDF, le stockage local de l'historique des patients, ainsi que l'affichage des explications locales des prédictions. Ces caractéristiques en font un outil pratique, adaptable et potentiellement déployable dans un cadre médical réel.

**Perspectives**

Ce travail ouvre plusieurs axes de recherche et de développement pour l'avenir :

- Intégration de données en temps réel : connecter l'application à des dispositifs médicaux (capteurs de glycémie, montres intelligentes) permettrait d'affiner les prédictions et de fournir un suivi personnalisé et dynamique.

- Utilisation de modèles de Deep Learning plus complexes : explorer des architectures plus avancées, telles que les réseaux neuronaux convolutifs (CNN) ou récurrents (LSTM/GRU), pourrait améliorer la détection de patterns non linéaires dans des jeux de données plus riches.
- Extension à d'autres pathologies chroniques : l'approche proposée peut être adaptée à la prédiction d'autres maladies telles que les maladies cardiovasculaires, l'hypertension ou certains types de cancer, élargissant ainsi l'impact clinique de l'outil.
- Amélioration de l'interface utilisateur et de l'ergonomie : en collaboration avec des professionnels de santé, une version plus ergonomique et certifiée pourrait être développée, afin de répondre aux standards des dispositifs médicaux numériques.
- Validation clinique à grande échelle : une étude prospective impliquant des professionnels de santé et des patients permettrait d'évaluer l'efficacité et l'acceptabilité de l'outil en situation réelle.

Ce travail se distingue par sa rigueur scientifique, son innovation technologique et sa pertinence clinique. Il démontre le potentiel transformateur de l'intelligence artificielle dans le domaine médical, en contribuant à une médecine plus prédictive, personnalisée et accessible.

## Bibliographies.


**[1]** Organisation mondiale de la Santé. (2024). *Diabète – Fiche d'information*. https://www.who.int/news-room/fact-sheets/detail/diabetes

**[2]** DeFronzo, R. A. (1997). Pathogenesis of type 2 diabetes: Metabolic and molecular implications for identifying diabetes genes. *Diabetes Reviews*, *5*, 178–269.

**[3]** Pihoker, C., Gilliam, L. K., Hampe, C. S., & Lernmark, Å. (2005). Autoantibodies in diabetes. *Diabetes*, *54*(suppl_2), S52–S61.

**[4]** Atkinson, M. A., & Eisenbarth, G. S. (2001). Type I diabetes: New perspectives on disease pathogenesis and treatment. *The Lancet*, *358*, 221–229.

**[5]** Skylar, J. S., Bakris, G. L., Bonifacio, E., Darsow, T., Eckel, R. H., Groop, L., & Groop, P.-H. (2017). Differentiation of diabetes by pathophysiology, natural history, and prognosis. *Diabetes*, *66*, 241–255.

**[6]** Michels, A., & Gottlieb, P. (2015, March 4). Pathogenesis of Type 1A Diabetes. In L. J. De Groot et al. (Eds.), *Endotext* [Internet]. MDText.com, Inc. https://www.ncbi.nlm.nih.gov/books/NBK166097/

**[7]** Yau, M., Maclaren, N. K., & Sperling, M. (2015, March 4). Etiology and Pathogenesis of Diabetes Mellitus. In L. J. De Groot et al. (Eds.), *Endotext* [Internet]. MDText.com, Inc. https://www.ncbi.nlm.nih.gov/books/NBK279115/

**[8]** DeFronzo, R. A. (1988). Lilly Lecture. The triumvirate: Beta-cell, muscle, liver: A collusion responsible for NIDDM. *Diabetes*, *37*, 667–687.

**[9]** DeFronzo, R. A. (2004). Pathogenesis of type 2 diabetes mellitus. *Medical Clinics of North America*, *88*, 787–835.

**[10]** Abdul-Ghani, M., & DeFronzo, R. A. (2008). Mitochondrial dysfunction, insulin resistance, and type 2 diabetes mellitus. *Current Diabetes Reports*, *8*, 173–178.

**[11]** DeFronzo, R. A. (2009). Banting Lecture. From the triumvirate to the ominous octet: A new paradigm for the treatment of type 2 diabetes mellitus. *Diabetes*, *58*(4), 773–795.

**[12]** DeSisto, C. L., Kim, S. Y., & Sharma, A. J. (2014). Prevalence estimates of gestational diabetes mellitus in the United States, Pregnancy Risk Assessment Monitoring System (PRAMS), 2007–2010. *Preventing Chronic Disease*, *11*, 130415. https://doi.org/10.5888/pcd11.130415

**[13]** Sahnine, N., & Yahiaoui, Y. (2018). *Analyse des moyens à mettre en œuvre pour lutter contre le diabète : Cas CHU l'hôpital Belloua Tizi-Ouzou* [Mémoire de Master, Université Mouloud Mammeri de Tizi-Ouzou].

**[14]** Lamdjadani, A. K., & Bouazza, A. (2017). *Étude épidémiologique sur les facteurs de risque associés au diabète de type 2* [Mémoire de Master, Université Abdelhamid Ibn Badis de Mostaganem].

**[15]** Cersosimo, E., & DeFronzo, R. A. (2006). Insulin resistance and endothelial dysfunction: The road map for cardiovascular diseases. *Diabetes/Metabolism Research and Reviews*, *22*, 423–436.

**[16]** van Tilburg, J., van Haeften, T. W., Pearson, P., & Wijmenga, C. (2001). Defining the genetic contribution of type 2 diabetes mellitus. *Journal of Medical Genetics*, *38*, 569–578.

**[17]** DeFronzo, R. A. (1997). Pathogenesis of type 2 diabetes: Metabolic and molecular implications for identifying diabetes genes. *Diabetes Reviews*, *5*, 178–269.

**[18]** Miyazaki, Y., Mahankali, A., Matsuda, M., Mahankali, S., Hardies, J., Cusi, K., Mandarino, L. J., & DeFronzo, R. A. (2002). Effect of pioglitazone on abdominal fat distribution and insulin sensitivity in type 2 diabetic patients. *Journal of Clinical Endocrinology & Metabolism*, *87*, 2784–2791.

**[19]** Dosilovic, F. K., Brcic, M., & Hlupic, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 210–215). IEEE. https://doi.org/10.23919/MIPRO.2018.8400040

**[20]** Krueger, C., et al. (2020). Machine learning and artificial intelligence: Definitions, applications, and future directions. *Current Reviews in Musculoskeletal Medicine*, *13*(1), 69–76.

**[21]** Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94

**[22]** Reddy, S., Fox, J., & Purohit, M. P. (2019). Artificial intelligence-enabled healthcare delivery. *Journal of the Royal Society of Medicine*, *112*(1), 22–28. https://doi.org/10.1177/0141076818815510

**[23]** Castagno, S., & Khalifa, M. (2020). Perceptions of artificial intelligence among healthcare staff: A qualitative survey study. *Frontiers in Artificial Intelligence*, *3*, 578983. https://doi.org/10.3389/frai.2020.578983

**[24]** Laï, M. C., Brian, M., & Mamzer, M. F. (2020). Perceptions of artificial intelligence in healthcare: Findings from a qualitative survey study among actors in France. *Journal of Translational Medicine*, *18*(1), 14. https://doi.org/10.1186/s12967-019-02204-y

**[25]** Secinaro, S., Calandra, D., Secinaro, A., Muthurangu, V., & Biancone, P. (2021). The role of artificial intelligence in healthcare: A structured literature review. *BMC Medical Informatics and Decision Making*, *21*(1), 125. https://doi.org/10.1186/s12911-021-01488-9

**[26]** Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, *20*(1), 310. https://doi.org/10.1186/s12911-020-01332-6

**[27]** Xu, J. J. (2014). Knowledge discovery and data mining. In *Computer Handbook: Information Systems and Information Technology* (3rd ed., pp. 19–1–19–22). CRC Press. https://doi.org/10.1201/b16768

**[28]** Bukhari, M. M., Alkhamees, B. F., Hussain, S., Gumaei, A., Assiri, A., & Ullah, S. S. (2021). An improved artificial neural network model for effective diabetes prediction. *Complexity*, *2021*, 5525271. https://doi.org/10.1155/2021/5525271

**[29]** Ramesh, J., Aburukba, R., & Sagahyroon, A. (2021). A remote healthcare monitoring framework for diabetes prediction using machine learning. *Healthcare Technology Letters*, *8*(3), 45–57. https://doi.org/10.1049/htl2.12010

**[30]** Rufo, D. D., Debelee, T. G., Ibenthal, A., & Negera, W. G. (2021). Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM). *Diagnostics*, *11*(9), 1714. https://doi.org/10.3390/diagnostics11091714

**[31]** Patil, R., Tamane, S., Rawandale, S. A., & Patil, K. (2022). A modified mayfly-SVM approach for early detection of type 2 diabetes mellitus. *International Journal of Electrical and Computer Engineering*, *12*(1), 524–533. https://doi.org/10.11591/ijece.v12i1.pp524-533

**[32]** Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

**[33]** Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5), e1424. https://doi.org/10.1002/widm.1424

**[34]** Vidhya, K., & Shanmugalakshmi, R. (2020). Deep learning based big medical data analytic model for diabetes complication prediction. *Journal of Ambient Intelligence and Humanized Computing*, *11*(11), 5691–5702. https://doi.org/10.1007/s12652-020-01930-2

**[35]** Esteva, M., Xu, W., Simone, N., Gupta, A., & Jah, M. (2020). Modeling data curation to scientific inquiry: A case study for multimodal data integration. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (pp. 235–242). https://doi.org/10.1145/3383583.3398539

**[36]** Cai, Q., Wang, H., Li, Z., & Liu, X. (2019). A survey on multimodal data-driven smart healthcare systems: Approaches and applications. *IEEE Access*, *7*, 133583–133599. https://doi.org/10.1109/ACCESS.2019.2941419.

**[37]** Alakwaa, F. M., Chaudhary, K., & Garmire, L. X. (2018). Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data. *Journal of Proteome Research*, *17*(1), 337–347. https://doi.org/10.1021/acs.jproteome.7b00564

**[38]** Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, *380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

**[39]** Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., ... & Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface*, *15*(141), 20170387. https://doi.org/10.1098/rsif.2017.0387

**[40]** Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*, 44–56. https://doi.org/10.1038/s41591-018-0300-7

**[41]** Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, *375*(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

**[42]** Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, *19*(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

**[43]** Deo, R. C. (2015). Machine learning in medicine. *Circulation*, *132*(20), 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

**[44]** Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230–243. https://doi.org/10.1136/svn-2017-000101

**[45]** Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, *319*(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391

**[46]** Liu, Y., Chen, P. C., Krause, J., & Peng, L. (2019). How to read articles that use machine learning: users' guides to the medical literature. *JAMA*, *322*(18), 1806–1816. https://doi.org/10.1001/jama.2019.13434

**[47]** Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, *17*, Article 195. https://doi.org/10.1186/s12916-019-1426-2

**[48]** Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., ... & Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, *25*, 1337–1340. https://doi.org/10.1038/s41591-019-0548-6

**[49]** Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*, 24–29. https://doi.org/10.1038/s41591-018-0316-z

**[50]** Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

**[51]** Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, *29*(2), 102–127. https://doi.org/10.1016/j.zemedi.2018.11.002

**[52]** Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005

**[53]** LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

**[54]** Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

**[55]** Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

**[56]** Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

**[57]** He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://doi.org/10.1109/CVPR.2016.90

**[58]** Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1409.1556

**[59]** Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9. https://doi.org/10.1109/CVPR.2015.7298594

**[60]** Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

**[61]** Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Fourth International Conference on 3D Vision (3DV)*, 565–571. https://doi.org/10.1109/3DV.2016.79

**[62]** Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965

**[63]** Zhang, Y., & Yang, Q. (2017). A survey on multi-task learning. *arXiv preprint*, arXiv:1707.08114. https://arxiv.org/abs/1707.08114

**[64]** Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*(1), 41–75. https://doi.org/10.1023/A:1007379606734

**[65]** Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*, *5*, 8869–8879. https://doi.org/10.1109/ACCESS.2017.2694446

**[66]** Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint*, arXiv:1711.05225. https://arxiv.org/abs/1711.05225

**[67]** Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, *316*(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

**[68]** Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

**[69]** Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., ... & van der Laak, J. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports*, *6*, Article 26286. https://doi.org/10.1038/srep26286

**[70]** Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., ... & Corrado, G. S. (2017). Detecting cancer metastases on gigapixel pathology images. *arXiv preprint*, arXiv:1703.02442. https://arxiv.org/abs/1703.02442

**[71]** Wang, P., Berzin, T. M., Brown, J. R. G., Bharadwaj, S., Becq, A., Xiao, X., ... & Liu, X. (2018). Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut*, *68*(10), 1813–1819. https://doi.org/10.1136/gutjnl-2018-317500

**[72]** Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W., & Baldi, P. (2018). Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology*, *155*(4), 1069–1078.e8. https://doi.org/10.1053/j.gastro.2018.06.037

**[73]** Jin, Y., Xu, S., Zhang, J., Wu, Y., Wang, Y., & Wei, J. (2020). Artificial intelligence in lung cancer: Diagnosis, treatment and prognosis. *Journal of Thoracic Disease*, *12*(11), 6630–6640. https://doi.org/10.21037/jtd-20-2587

**[74]** Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., ... & Corrado, G. C. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, *25*(6), 954–961. https://doi.org/10.1038/s41591-019-0447-x

**[75]** Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., ... & Thomas, L. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, *29*(8), 1836–1842. https://doi.org/10.1093/annonc/mdy166

**[76]** De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., ... & Suleyman, M. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, *24*, 1342–1350. https://doi.org/10.1038/s41591-018-0107-6

**[77]** Liang, H., Tsui, B. Y., Ni, H., Valentim, C. C., Baxter, S. L., Liu, G., ... & Duong, D. Q. (2019). Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature Medicine*, *25*, 433–438. https://doi.org/10.1038/s41591-018-0335-9

**[78]** Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine Learning for Healthcare Conference*, 359–380. https://proceedings.mlr.press/v106/tonekaboni19a.html

**[79]** Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? *arXiv preprint*, arXiv:1712.09923. https://arxiv.org/abs/1712.09923

**[80]** P. B. Khokhar, V. Pentangelo, F. Palomba, and C. Gravino, "Towards Transparent and

Accurate Diabetes Prediction Using Machine Learning and Explainable Artificial Intelligence," arXiv preprint arXiv:2501.18071, Jan. 2025.

**[81]** H. F. El-Sofany, S. A. El-Seoud, O. H. Karam, Y. M. A. El-Latif, and I. A. T. F. Taj-Eddin,

"A proposed technique using machine learning for the prediction of diabetes disease through a mobile app," International Journal of Intelligent Systems, vol. 2024, Article ID 6688934, 2024, doi: 10.1155/2024/6688934.

**[82]** A. Ahmed, J. Khan, M. Arsalan, K. Ahmed, A. A. Shahat, A. Alhalmi, and S. Naaz,

"Machine Learning Algorithm-Based Prediction of Diabetes Among Female Population Using PIMA Dataset," Healthcare, vol. 13, no. 1, p. 37, Dec. 2024, doi: 10.3390/healthcare13010037.

**[83]** B. Kurt et al., "Prediction of gestational diabetes using deep learning and Bayesian optimization and traditional machine learning techniques," Med. Biol. Eng. Comput., no.

0123456789, 2023, doi: 10.1007/s11517-023-02800-7.

**[84]** M. Shrestha et al., "A novel solution of deep learning for enhanced support vector machine for predicting the onset of type 2 diabetes," Multimed. Tools Appl., vol. 82, no. 4, pp.

6221–6241, 2023, doi: 10.1007/s11042-022-13582-9.

**[85]** A. Rajagopal, S. Jha, R. Alagarsamy, S. G. Quek, and G. Selvachandran, "A novel hybrid machine learning framework for the prediction of diabetes with context-customized regularization and prediction procedures," Math. Comput. Simul., vol. 198, pp. 388–406, Aug.2022, doi: 10.1016/j.matcom.2022.03.003.

**[86]** V. Chang, M. A. Ganatra, K. Hall, L. Golightly, and Q. A. Xu, "An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators," Healthc. Anal., vol. 2, no. October, p. 100118, Nov. 2022, doi: 10.1016/j.health.2022.100118.

**[87]** N. Ahmed et al., "Machine learning based diabetes prediction and development of smart web application," Int. J. Cogn. Comput. Eng., vol. 2, no. March, pp. 229–241, Jun. 2021, doi: 10.1016/j.ijcce.2021.12.001.

**[88]** Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988).Pima Indians Diabetes Database. UCI Machine Learning Repository via Kaggle.https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

**[89]** da Silva, J. (2023). Diabetes Dataset. Kaggle. https://www.kaggle.com/datasets/johndasilva/diabetes

**[90]** Python Software Foundation. (n.d.). *Python programming language*. https://www.python.org/

**[91]** NetApp. Définition de python . Disponible sur : Girard, Gabriel. "IFT211/IFT776 Programmation scientifique en Python.

**[92]** NetApp. Définition de la beb pandas . Disponible sur : http://www.python-simple.com/python-pandas/panda-intro.php.

**[93]** NetApp. Définition de la beb matplotlib. Disponible sur : https://datascientest.com/matplotlib-tout-savoir.

**[94]** ActiveState (s.d.). What is scikit-learn in python? de https://www.activestate.com/resources/quick-reads/what-is-scikit-learn-in-python/.

**[95]** Waskom, M. L. (2024). *seaborn: statistical data visualization — seaborn 0.13.2 documentation*. https://seaborn.pydata.org/

**[96]** Microsoft Corporation (s.d.). Visual studio code documentation. de https://code.visualstudio.com/Docs.

**[97]** DataScientest (s.d.a). Jupyter notebook : Tout savoir. De https://datascientest.com/jupyter-notebook-tout-savoir.

**[98]** Bisong, E. (2019). Google Colaboratory. In Building Machine Learning and Deep Learning Models on Google Cloud Platform (pp. 59–64). Apress. https://doi.org/10.1007/978-1-4842-4470-8_7