الجمهورية الجزائرية الديمقراطية الشعبية

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة التعليم العالي و البحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة أبي بكر بلقايد - تلمسان -

Université Aboubakr Belkaïd – Tlemcen –
Faculté de TECHNOLOGIE

MEMOIRE

Présenté pour l'obtention du **diplôme** de **MASTER**

En : Génie Biomédical

Spécialité : Imagerie Médicale

Par : ELANDALOUSSI Fatima Zohra & REGUIG Abderrahmen

Sujet

Etude de l'impact des techniques d'équilibrage des données sur la performance de la classification dans le domaine médical

Soutenu publiquement, le / / , devant le jury composé de :

M LAZOUNI Amine	MCA	Université de Tlemcen	Président
Mme ILES Amel	Professeur	Université de Tlemcen	Examinatrice
Mme BELAIDI Asma	MCB	Université de Tlemcen	Encadrante
Mme BELAROUCI Sara	MCB	Université de Tlemcen	Co-Encadrante

Année universitaire : 2025 /2026

Dédicace

Louange à Allah, le Très-Haut, le Très-Miséricordieux, qui m'a accordé la force, la patience et la persévérance pour franchir chaque étape de ce parcours.

Je dédie ce modeste travail : À mes parents bien-aimés,

Il n'existe ni mots ni pages suffisantes pour traduire la profondeur de ce que je ressens pour vous.
Merci pour votre amour sans conditions, vos sacrifices silencieux et vos invocations constantes. Vous êtes mon refuge, ma force et ma fierté.

À mes chères sœurs et à mon cher frère,
Merci pour votre présence, vos encouragements et votre confiance en moi. Vous êtes les piliers de mon équilibre.

À leurs enfants,
Vos sourires et votre innocence apportent de la lumière dans ma vie. Puissiez-vous grandir sous la protection divine et réaliser vos rêves.

À mes fidèles amis : Ikram, Amani, Aya, Chaima, hadjer, Raouda …
Merci pour les souvenirs, Les années d'études ont été remplies de moments inoubliables grâce à vous. Vous avez été mes compagnons de route, merci pour votre patience et votre présence à mes côtés dans les moments d'effort comme de doute. Vous avez enrichi ce voyage.

À chacun de vous, je dis avec reconnaissance :

Ce mémoire est bien plus qu'un document académique, il est le fruit d'un chemin parcouru avec vous, dans la foi, l'amour et la gratitude.

Fatima

Dédicace

Je dédie ce mémoire

À mes parents bien-aimés, En témoignage de mon profond respect et de ma gratitude infinie. Votre soutien constant, vos sacrifices silencieux, vos prières sincères et votre amour inconditionnel ont été le socle sur lequel j'ai pu m'appuyer tout au long de ce parcours.

À mes frères : ***Vos sourires, votre spontanéité ont été pour moi une source de réconfort et de motivation. Votre présence joyeuse a allégé les moments de fatigue et donné du sens à mes efforts. Je formule pour vous les vœux les plus sincères de réussite et d'épanouissement, sous la bienveillance d'Allah. Puisse votre chemin être éclairé par la sagesse, la foi et la réussite.***

À mes amies et collègues, Votre soutien tout au long de ce parcours universitaire a été d'un grand réconfort. Merci pour votre solidarité, vos conseils et les moments partagés, qui ont enrichi cette expérience tant sur le plan humain qu'intellectuel.

Ce mémoire constitue l'aboutissement d'un cheminement exigeant, rendu possible grâce à l'appui indéfectible de mes proches, et s'inscrit dans une dynamique de reconnaissance, de foi et de gratitude.

Abdelrahmane

REMERCIMENT

Nous tenons tout d'abord à exprimer notre profonde reconnaissance à Mme Belaidi Asma, notre encadrante, pour sa disponibilité constante, son accompagnement rigoureux et ses conseils avisés tout au long de ce travail. Sa bienveillance, sa pédagogie et son exigence scientifique ont été déterminants dans l'aboutissement de ce mémoire.

Nos remerciements s'adressent également à Mme Belarouci Sara, notre co-encadrante, pour son soutien technique, ses remarques pertinentes et son engagement sans faille. Sa contribution a largement enrichi notre réflexion et renforcé la qualité scientifique de notre projet.

Nous remercions tout particulièrement Monsieur le Président du jury et Madame l'Examinatrice, d'avoir accepté de présider et d'examiner ce travail. Leur disponibilité, leur attention et leurs retours seront pour nous une source précieuse d'apprentissage. Enfin, nous remercions toutes les personnes qui nous ont soutenus de près ou de loin tout au long de cette aventure académique.

Résumé

Ce mémoire s'intéresse à la détection automatique du diabète par des techniques d'apprentissage automatique (*Machine Learning*) et d'apprentissage profond (*Deep Learning*). Cette pathologie chronique majeure, qui touche des millions de personnes dans le monde, est caractérisée par un dysfonctionnement de la régulation glycémique lié à un défaut de sécrétion ou d'action de l'insuline. Dans ce travail, nous évaluons plusieurs modèles de classification sur une base de données clinique, en comparant rigoureusement leurs performances selon deux scénarios : avec et sans équilibrage des classes. La gestion du déséquilibre des données constitue en effet un enjeu critique en informatique médicale, où les classes pathologiques sont souvent minoritaires.

Les résultats expérimentaux montrent que l'algorithme *Extra Trees* s'affirme comme le modèle le plus performant dans toutes les configurations : il atteint un score F1 de **97,82 %** sans traitement préalable, et affiche des performances maximales de **100 %** sur l'ensemble des métriques lorsqu'il est combiné avec la méthode de sur-échantillonnage *SMOTE*. Les modèles d'ensemble (*Random Forest, Gradient Boosting, XGBoost*) suivent de près ces performances, tandis que les approches linéaires et probabilistes montrent des limites évidentes. L'étude comparative révèle que l'équilibrage par *SMOTE, ADASYN* et améliore significativement la sensibilité des modèles, contrairement à l'approche *RandomUnderSampler* qui dégrade les performances en supprimant des informations cruciales. Par ailleurs, un réseau de neurones convolutionnels unidimensionnel (*CNN 1D*) a été implémenté ; bien que compétitif avec un score F1 de **96,83 %** obtenu via *SMOTE*, il reste légèrement inférieur aux modèles d'ensemble optimisés.

Pour concrétiser ces travaux, une interface graphique d'aide au diagnostic médical a été développée. Ce système permet aux professionnels de santé de saisir les données d'un nouvel échantillon clinique afin d'obtenir une prédiction instantanée, fiable et automatisée du risque de diabète. En conclusion, le couplage entre *Extra Trees* et *SMOTE* offre la solution de détection la plus robuste et la plus fiable. Ce travail démontre qu'une gestion appropriée du déséquilibre des classes est indispensable en médecine pour minimiser les faux négatifs, un impératif vital là où un retard de diagnostic peut entraîner de graves complications cliniques.

Mots-clés : Diabète, Machine Learning, Deep Learning, Déséquilibre des classes, SMOTE, Extra Trees, Interface d'aide au diagnostic, Informatique médicale.

Abstract

This dissertation focuses on the automatic detection of diabetes utilizing both Machine Learning and Deep Learning techniques. As a major chronic pathology affecting millions of people worldwide, this disease is characterized by a dysfunction in glycemic regulation stemming from a defect in insulin secretion or action.

In this work, we evaluate several classification models on a clinical dataset, rigorously comparing their performance across two scenarios: with and without class balancing. Managing data imbalance constitutes a critical challenge in medical informatics, where pathological classes are frequently underrepresented.

Experimental results indicate that the *Extra Trees* algorithm achieves the best overall performance in all configurations. It yields an F1-score of **97.82%** without prior processing and reaches maximum performance scores of **100%** across all metrics when combined with the SMOTE oversampling method. Ensemble models (*Random Forest*, *Gradient Boosting*, *XGBoost*) closely follow these performance levels, whereas linear and probabilistic approaches exhibit clear limitations. The comparative study reveals that balancing techniques such as *SMOTE*, *ADASYN*, significantly improve model sensitivity. Conversely, the *RandomUnderSampler* approach degrades performance by eliminating crucial information. Furthermore, a one-dimensional Convolutional Neural Network (*1D CNN*) was implemented; although competitive with an F1-score of **96.83%** obtained via SMOTE, it remains slightly inferior to optimized ensemble models. To operationalize this research, a graphical user interface for medical diagnosis support was developed. This system enables healthcare professionals to input data from a new clinical sample to obtain an instantaneous, reliable, and automated prediction of diabetes risk. In conclusion, the coupling of *Extra Trees* and *SMOTE* offers the most robust and reliable detection solution. This work demonstrates that proper management of class imbalance is indispensable in medicine to minimize false negatives, a vital imperative where diagnostic delays can lead to severe clinical complications.

Keywords: Diabetes, Machine Learning, Deep Learning, Class Imbalance, SMOTE, Extra Trees, Diagnosis Support Interface, Medical Informatics.

ملخص

تركز هذه المذكرة على الكشف التلقائي عن مرض السكري باستخدام تقنيات التعلم الآلي والتعلم العميق. ويتميز هذا المرض المزمن الرئيسي، الذي يصيب ملايين الأشخاص حول العالم، بخلل في تنظيم نسبة السكر في الدم نتيجة نقص في إفراز الإنسولين أو في مفعوله.

نقوم في هذا العمل بتقييم العديد من نماذج التصنيف على مجموعة بيانات سريرية، مع إجراء مقارنة صارمة لأدائها وفق سيناريوهين: مع وبدون موازنة الفئات. وتشكل إدارة مشكلة عدم توازن البيانات تحديًا حاسمًا في مجال المعلوماتية الطبية، حيث غالبًا ما تكون الحالات المرضية ممثلة بنسبة ضئيلة.

تُظهر النتائج التجريبية أن خوارزمية "*Extra Trees* " أثبتت أنها النموذج الأكثر فعالية في جميع الحالات، حيث حققت مقياس إف 1 بنسبة **97.82 %** بدون معالجة مسبقة، وسجلت أداءً أقصى بنسبة **100 %** في جميع المقاييس عند دمجها مع طريقة الإفراط في أخذ العينات " *SMOTE* ". وتأتي نماذج التجميع ("*Random Forest* "، " *Boosting Gradient*"، "*XGBoost* ") قريبة جدًا من هذه المستويات، في حين أظهرت المقاربات الخطية والاحتمالية قصورًا واضحًا.

وكشفت الدراسة المقارنة أن موازنة الفئات باستخدام أساليب " *SMOTE* " و" *ADASYN*" تحسن بشكل كبير من حساسية النماذج، على عكس مقاربة "*RandomUnderSampler* " التي أدت إلى تراجع الأداء بسبب حذف معلومات حاسمة. علاوة على ذلك، تم تطوير شبكة عصبونية تلافيفية أحادية الأبعاد، وعلى الرغم من تنافسيتها بتحقيق مقياس إف 1 بنسبة 96.83 % باستخدام " *SMOTE* "، إلا أنها ظلت أقل كفاءة بقليل من النماذج التجميعية المحسنة.

وتجسيدًا لهذا العمل، تم تطوير واجهة رسومية لدعم التشخيص الطبي. يتيح هذا النظام لمهنيي قطاع الصحة إدخال بيانات عينة سريرية جديدة للحصول على تنبؤ فوري وموثوق وآلي لمخاطر الإصابة بالسكري.

في الختام، يوفر الربط بين خوارزمية " *Extra Trees* " وطريقة " *SMOTE* " الحل الأكثر قوة وموثوقية للكشف عن المرض. ويؤكد هذا العمل أن الإدارة المناسبة لعدم توازن الفئات أمر لا غنى عنه في المجال الطبي لتقليل السلبيات الكاذبة، وهو أمر حيوي حيث يمكن أن يؤدي التأخر في التشخيص إلى مضاعفات سريرية خطيرة.

الكلمات المفتاحية

مرض السكري، التعلم الآلي، التعلم العميق، عدم توازن الفئات، *SMOTE*، *Extra Trees* ، واجهة دعم التشخيص، المعلوماتية الطبية

Table des matières

Résumé I
Abstract II
ملخص III
Liste des figures VII
Liste des tableaux IX
Liste des abréviations X
Introduction générale 11
Chapitre 1 : Aspects médicaux du diabète 14
1. Introdution 14
2. Physiopathologie et concepts fondamentaux du diabète 14
2.1 Typologie et formes cliniques du diabète 16
2.2 Aperçu des statistiques épidémiologiques récentes 18
2.3 Facteurs de risque 19
2.4 Principales complications du diabète 21
3. Diagnostic et bilan initial 22
3.1 Critères cliniques 22
3.2 Critères biologiques 23
3.3 Outils d'aide au diagnostic 24
4. Prévention et prise en charge 25
4.1 Mesures hygiéno-diététiques 25
5. Limites des approches conventionnelles 26
6. Conclusion 27
Chapitre 2 : Équilibrage des données et intelligence artificielle appliquée à la classification médicale 29
1. Introduction 29
2. Équilibrage des données 29
2.1 Généralités sur le déséquilibre des données 29
2.2 Méthodes d'équilibrage 30
2.2.1 Modification au niveau de données 30
2.2.1.1 Random Oversampling 30
2.2.1.2 Random Undersampling 31
2.2.1.3 SMOTE 31
2.2.1.4 ADASYN 32
2.2.1.6 Tomek Links 33
2.2.2 Modification au niveau des algorithmes 33

2.2.2.1 Apprentissage sensible au coût (Cost-Sensitive Learning) 33
2.2.2.2 Pondération des classes 34
2.2.2.3 Ajustement des seuils de décision 34
2.2.3 Modification hybrides 34
2.2.3.1 Combinaison sur-échantillonnage et sous-échantillonnage 34
2.2.3.2 Combinaison équilibrage et algorithmes de classification 34
3. Intelligence artificielle 35
4. Approches de l'apprentissage automatique 36
4.1 Apprentissage supervisé 36
4.2 Apprentissage non supervisé 37
4.3 Apprentissage semi-supervisé 37
4.4 Apprentissage par renforcement 38
5. Approches de Machine Learning pour la classification 39
5.1 Régression logistique 39
5.2 Machine à vecteurs de support 39
5.3 Arbre de décision 39
5.4 Forêt aléatoire 40
5.5 K plus proches voisins 40
5.6 Naive Bayes 40
5.7 XGBoost 40
5.8 Gradient Boosting 40
6. Approches de Deep Learning pour la classification 41
6.1 Réseau de neurone profond (Deep Learning) 42
6.2 Réseaux de neurones convolutifs (CNN) 42
6.2.1 Couche Convolutionnelle 43
6.2.2 Couche de Pooling (Pooling Layer) 43
6.2.3 Couche entièrement connectée (Fully Connected Layer / Dense Layer) 44
6.2.4 Couche de sortie (Output Layer) 44
7. Importance de l'intelligence artificielle dans les données médicales 45
8. Travaux connexes : classification et prédiction du diabète 46
9. Conclusion 48
Chapitre 3. Résultats et Discussion 49
1. Contexte de l'étude 49
2. Environnement de développement 50
3. Implémentation des méthodes 51
3.1 Description de la base de données 52
3.2 Prétraitement de la base de données 54
3.3 Classification de diabète : approche sans équilibrage de données 54

3.3.1 Application des modèles classiques 56
3.3.2 Application de l'apprentissage en profondeur 56
3.4 Classification de diabète : approche avec équilibrage de données 57
4. Résultats et discussion 59
4.1 Critères d'évaluation 59
4.2 Classification de diabète : approche sans équilibrage de données 61
4.2.1 Application des modèles classiques 61
4.2.2 Application de l'apprentissage en profondeur 62
4.3 Classification de diabète : approche avec équilibrage de données 62
4.3.1 Application des modèles classiques avec équilibrage 62
4.3.2 Application du CNN 1D avec équilibrage 64
5. Étude comparative des performances du meilleur modèle avec et sans équilibrage des données 65
6. Explicabilité globale du modèle par l'approche SHAP 71
7. Analyse comparative avec les travaux existant 73
8. Interface interactive de prédiction 74
8.1 Architecture de l'interface 74
8.2 Fonctionnalités de l'interface 74
8.3 Format attendu du fichier CSV 78
8.4 Avantages pédagogiques et pratiques 79
7. Conclusion 79
Conclusion Générale 81
Bibliograpie 82

Liste des figures

- **Figure 1.1** : Régulation de la glycémie (insuline/glucagon)
- **Figure 1.2** : Classification des types de diabète (Type 1, Type 2, gestationnel).
- **Figure 1.3** : Évolution du nombre de personnes vivant avec le diabète dans le monde (1990–2022).
- **Figure 1.4** : Principaux facteurs de risque du diabète
- **Figure 1.5** : Physiopathologie simplifiée du diabète de type 2
- **Figure 1.6** : Complications microvasculaires et macrovasculaires du diabète.
- **Figure 1.7** : processus général de diagnostic du diabète
- **Figure 2.1** : Principe du Random Oversampling
- **Figure 2.2** : Principe du Random Undersampling
- **Figure 2.3** : SMOTE (génération d'exemples synthétiques)
- **Figure 2.4** : ADASYN (génération adaptative selon la difficulté locale)
- **Figure 2.6** : Principe de Tomek Links
- **Figure 2.7** : Principe général de l'apprentissage supervisé
- **Figure 2.8** : Principe de l'apprentissage non supervisé par clustering
- **Figure 2.9** : Fonctionnement de l'apprentissage semi-supervisé
- **Figure 2.10** : Architecture de l'apprentissage par renforcement
- **Figure 2.11** : Fonctionnement de réseau de neurone profond
- **Figure 2.12** : Les couche de CNN
- **Figure 2.13** : Une couche convolutionnelle en action
- **Figure 2.14** : Exemple de fonctionnement de la couche pooling
- **Figure 2.15** : Exemple des Couches entièrement connectées
- **Figure 3.1** : Schéma général de la méthodologie proposée pour la classification du diabète avec et sans équilibrage des données.
- **Figure 3.2 :** Répartition des classes
- **Figure 3.3 :** Architecture du CNN
- **Figure 3.4** : Tableau de bord comparatif des métriques de performance du modèle Extra Trees (Sans équilibrage vs Avec SMOTE)
- **Figure 3.5** : Matrices de Confusion Comparatives : Impact de l'Équilibrage SMOTE sur Extra Trees

- **Figure 3.6**: Évaluation duale du modèle Extra Trees + SMOTE – (A) Courbe ROC de performance diagnostique et (B) Courbe d'analyse du surapprentissage (Overfitting) selon la profondeur maximale (Max Depth).
- **Figure 3.7** : Courbe de perte (Log Loss) d'apprentissage et de test du modèle Extra Trees + SMOTE selon la profondeur maximale (Max Depth).
- **Figure 3.8** : Analyse SHAP (Summary Plot) des contributions des variables cliniques à la prédiction positive du modèle Extra Trees + SMOTE.
- **Figure 3.9**: Chargement de la base de données BDD.csv via le widget de téléchargement Google Colab
- **Figure 3.10**: Interface Gradio complète avec zone d'importation et résultats de prédiction
- **Figure 3.11**: Sélection du fichier CSV patient via l'explorateur de fichiers Windows
- **Figure 3.12** : Confirmation du fichier importé (Nouveau Feuille de calcul Microsoft Excel.csv)
- **Figure 3.13**: Résultats détaillés de la prédiction

Liste des tableaux

Tableau 1.1: Classification clinique des types de diabète (comparatif)

Tableau 1.2: Statistiques épidémiologiques récentes du diabète (monde et régions)

Tableau 1.3: Facteurs de risque du diabète (modifiables / non modifiables)

Tableau 1.4: Complications du diabète (aiguës / chroniques) + organes

Tableau 1.5: Signes cliniques évocateurs du diabète (polyurie, polydipsie, amaigrissement…)

Tableau 1.6: Critères diagnostiques du diabète (OMS/ADA) (glycémie à jeun, HGPO 75 g, HbA1c, glycémie aléatoire avec symptômes).

Tableau 1.7: Suivi clinique du diabète — paramètres et fréquence (HbA1c, TA, lipides, rein, œil)

Tableau 2.1: Différences entre Intelligence artificielle, Machine Learning et Deep Learning

Tableau 2.2: Comparaison des algorithmes de classification utilisés

Tableau 2.3: Applications de l'intelligence artificielle dans le domaine médical

Tableau 3 .1: Encodage des variables catégorielles

Tableau 3.2: Paramètres de création des modèles traditionnels

Tableau 3.3: Paramétrage des méthodes d'équilibrage retenues dans l'étude

Tableau 3.4: Résultats expérimentaux des modèles de ML

Tableau 3.5: Résultats expérimentaux des modèles de CNN

Tableau 3.6: récapitulatif des performances des modèlesML avec différentes méthods d'équilibrage

Tableau 3.7: récapitulatif des performances de modèle CNN avec différentes méthods d'équilibrage

Tableau 3.8: Analyse comparative avec les travaux existant

Liste des abréviations

Abréviation	Signification
ADASYN	Adaptive Synthetic Sampling
API	Application Programming Interface
CNN	Convolutional Neural Network (Réseau de Neurones Convolutif)
DL	Deep Learning (Apprentissage Profond)
ELM	Extreme Learning Machine
IA	Intelligence Artificielle
KNN	K-Nearest Neighbors (K Plus Proches Voisins)
ML	Machine Learning (Apprentissage Automatique)
NLP	Natural Language Processing (Traitement du Langage Naturel)
ReLU	Rectified Linear Unit
RBF	Radial Basis Function
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine (Machine à Vecteurs de Support)

Abréviation	Signification
ADA	American Diabetes Association
AVC	Accident Vasculaire Cérébral
BMI / IMC	Body Mass Index / Indice de Masse Corporelle
CGM	Continuous Glucose Monitoring (Monitorage Continu du Glucose)
CV	Cardiovasculaire
eGFR	Estimated Glomerular Filtration Rate (Débit de Filtration Glomérulaire estimé)

Introduction générale

Le diabète constitue l'une des pathologies chroniques les plus répandues et préoccupantes au monde, touchant plus de 500 millions de personnes selon l'Organisation Mondiale de la Santé (OMS). Cette maladie métabolique, caractérisée par une hyperglycémie persistante due à un défaut de sécrétion ou d'action de l'insuline, représente un défi majeur de santé publique. Sa prévalence croissante s'accompagne de complications graves à long terme, telles que les atteintes rétiniennes, les néphropathies, les neuropathies, les maladies cardiovasculaires et les accidents vasculaires cérébraux (AVC). Face à cette réalité, l'amélioration du dépistage précoce et le développement d'outils de prédiction du risque diabétique apparaissent comme des priorités absolues.

Les approches conventionnelles de diagnostic, principalement fondées sur la glycémie à jeun, l'hémoglobine glyquée (HbA1c) et l'hyperglycémie provoquée par voie orale, présentent plusieurs limites en pratique. Le sous-diagnostic reste alarmant, particulièrement dans les contextes à ressources limitées, tandis que la phase préclinique asymptomatique retarde fréquemment l'intervention thérapeutique. L'explosion de la charge mondiale de cette maladie, passant de 200 millions de cas en 1990 à plus de 830 millions en 2022, démontre l'urgence de faire évoluer les stratégies actuelles.

Dans ce contexte, l'émergence de l'intelligence artificielle (IA) et de l'apprentissage automatique (*Machine Learning*) offre des perspectives prometteuses pour révolutionner le diagnostic médical. Ces technologies permettent d'exploiter massivement les données cliniques, d'identify des relations complexes entre les variables biologiques et de construire des modèles prédictifs performants capables d'assister les professionnels de santé. Cependant, l'application de ces méthodes en médecine se heurte à un défi technique majeur : le déséquilibre des classes. En effet, les bases de données cliniques présentent souvent une répartition inégale, où les cas pathologiques sont minoritaires. Ce phénomène biaise l'apprentissage des modèles en les orientant vers la classe majoritaire (les patients sains), réduisant ainsi leur capacité à détecter correctement les patients réellement atteints — une problématique critique où l'omission d'un diagnostic positif peut avoir des conséquences cliniques dramatiques.

C'est dans cette problématique que s'inscrit le présent mémoire, intitulé **« Équilibrage et classification du diabète à l'aide des techniques d'apprentissage automatique »**. Ce travail

vise à développer une approche méthodologique rigoureuse, combinant des techniques d'équilibrage de données et des algorithmes de classification, afin d'optimiser la prédiction du diabète. L'objectif principal est de démontrer que la prise en compte explicite de l'asymétrie des classes constitue un levier déterminant pour fiabiliser les performances diagnostiques. De plus, cette étude ne se limite pas à la phase de modélisation théorique : elle se matérialise par la conception d'un système d'aide à la décision clinique, offrant une interface intuitive capable de guider les praticiens dans l'évaluation du risque chez de nouveaux patients.

Pour mener à bien cette étude, notre manuscrit est structuré en trois chapitres complémentaires :

- **Chapitre 1 : Aspects médicaux du diabète** – Ce chapitre pose les bases cliniques de l'étude. Il aborde la définition, la physiopathologie, les différentes formes de la maladie (type 1, type 2, gestationnel), son épidémiologie, ses facteurs de risque ainsi que les critères diagnostiques actuels et les limites des approches conventionnelles.
- **Chapitre 2 : Cadre théorique et méthodologique** – Consacré aux fondements théoriques, ce chapitre détaille la problématique du déséquilibre des données et présente les méthodes de rééchantillonnage envisagées (sur-échantillonnage, sous-échantillonnage, SMOTE, ADASYN, Tomek Links). Il introduit également les modèles d'apprentissage automatique et profond retenus pour la classification (Régression Logistique, SVM, Arbres de décision, Forêt Aléatoire, KNN, Naive Bayes, XGBoost, Gradient Boosting, CNN 1D), avant de clore sur une revue critique des travaux connexes.
- **Chapitre 3 : Contribution, résultats expérimentaux et déploiement** – Ce chapitre est dédié à la mise en œuvre pratique du projet. Il décrit la base de données *« Early Stage Diabetes Risk Prediction »*, les étapes de prétraitement, ainsi que l'évaluation comparative de deux scénarios (sans et avec application des techniques d'équilibrage). L'analyse de l'impact de ces méthodes sur les différents classifieurs y est menée à travers des métriques quantitatives (exactitude, précision, rappel, score F1) et des outils d'interprétabilité (SHAP). Enfin, ce chapitre présente le développement et l'architecture de l'interface utilisateur conçue pour tester de nouveaux échantillons en temps réel, offrant ainsi un véritable outil de support technologique pour le diagnostic médical.

À travers cette structure, ce mémoire ambitionne d'apporter une contribution utile au domaine de l'informatique médicale, en démontrant qu'une gestion rigoureuse de la structure des données alliée à un déploiement applicatif concret permet de concevoir des systèmes de prédiction du diabète à la fois robustes, précis et cliniquement exploitables.

Chapitre 1 : Aspects médicaux du diabète

1. Introdution

Le diabète est une maladie métabolique chronique caractérisée par une hyperglycémie résultant d'un défaut de sécrétion d'insuline, d'une résistance à son action, ou de l'association de ces deux mécanismes. Sur le plan clinique, on distingue principalement le diabète de type 1, d'origine auto-immune et généralement d'apparition précoce, le diabète de type 2, le plus fréquent, associé à des facteurs génétiques, comportementaux et environnementaux, ainsi que le diabète gestationnel, qui survient au cours de la grossesse.

L'évolution de cette pathologie peut rester asymptomatique pendant de nombreuses années, retardant ainsi son diagnostic et favorisant l'apparition de complications. Lorsqu'ils sont présents, les symptômes incluent notamment la polyurie, la polydipsie, la fatigue et une perte de poids inexpliquée, bien que certains patients puissent rester asymptomatiques.

Le diagnostic repose sur des critères biologiques standardisés, notamment la glycémie à jeun et le taux d'hémoglobine glyquée (HbA1c). En l'absence d'un contrôle glycémique adéquat, le diabète peut entraîner des complications microvasculaires et macrovasculaires, telles que les maladies cardiovasculaires et les accidents vasculaires cérébraux.

La prise en charge du diabète repose sur une approche globale incluant l'adoption d'un mode de vie sain, un suivi médical régulier, l'éducation thérapeutique du patient et, selon les cas, un traitement médicamenteux ou une insulinothérapie.

Dans les sections suivantes de ce chapitre, nous allons présenter les principaux aspects médicaux du diabète. Nous aborderons notamment la définition et les différents types de cette maladie, ainsi que ses méthodes de diagnostic, ses traitements et les statistiques relatives à sa prévalence. Nous examinerons également les principaux facteurs de risque associés au développement du diabète afin de mieux comprendre son impact sur la santé publique et l'importance d'une prise en charge précoce.

2. Physiopathologie et concepts fondamentaux du diabète

Le diabète est une maladie métabolique chronique définie par une hyperglycémie persistante. Il apparaît lorsque le pancréas ne sécrète pas suffisamment d'insuline ou lorsque l'organisme n'utilise pas efficacement l'insuline produite. L'insuline joue un rôle central dans

la régulation de la glycémie et dans l'utilisation du glucose par les cellules. Ainsi, l'altération de sa sécrétion ou de son action entraîne une élévation durable du glucose sanguin [1].

La figure 1.1 illustre le mécanisme de régulation de la glycémie par les deux principales hormones pancréatiques : l'insuline et le glucagon. L'insuline intervient principalement lors d'une augmentation du taux de glucose sanguin en favorisant son absorption et son stockage par les cellules, tandis que le glucagon agit en cas de diminution de la glycémie en stimulant la libération du glucose dans le sang. L'action complémentaire de ces deux hormones permet ainsi de maintenir l'équilibre glycémique de l'organisme.

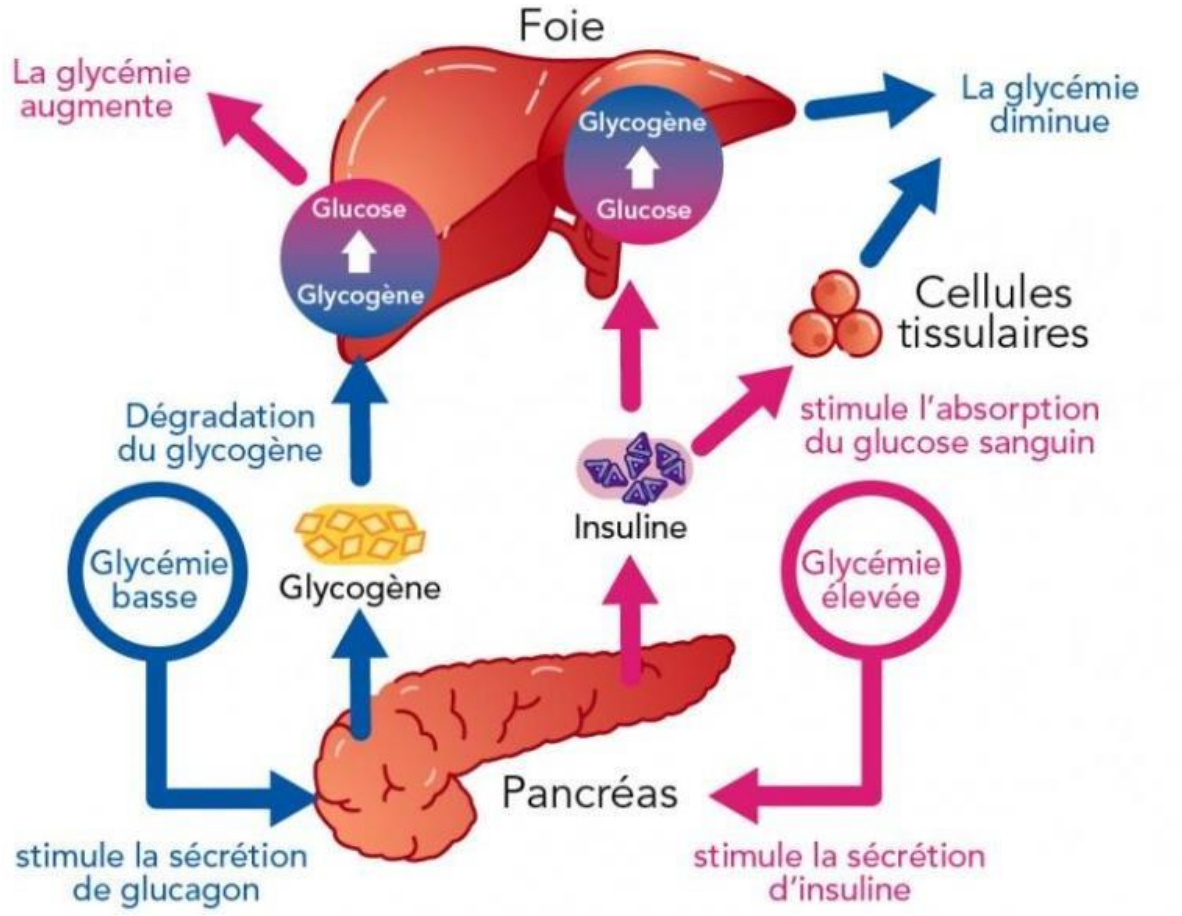


Figure 1.1 : Régulation de la glycémie par l'insuline et le glucagon [75].

Sur le plan clinique, le diabète est classé principalement en diabète de type 1, diabète de type 2 et diabète gestationnel [2]. En l'absence d'un contrôle glycémique approprié, l'hyperglycémie chronique peut provoquer des atteintes progressives des vaisseaux sanguins, des nerfs, des reins, des yeux et du cœur.

Le diabète constitue donc un enjeu majeur de santé publique, justifiant un dépistage précoce, un suivi continu et une prise en charge multidisciplinaire [3].

2.1 Typologie et formes cliniques du diabète

Le diabète est une affection métabolique chronique se manifestant par une élévation persistante du taux de glucose dans le sang, causée par une altération de la production d'insuline, de son efficacité, ou par l'association de ces deux anomalies.

Selon les standards internationaux, notamment ceux de l'American Diabètes Association, le diabète est classé en plusieurs catégories cliniques, dont les plus fréquentes sont le diabète de type 1, le diabète de type 2 et le diabète gestationnel [2].

2.1.1 Diabète de type 1

Le diabète de type 1 est une forme auto-immune qui se caractérise par une destruction progressive des cellules β des îlots de Langerhans du pancréas, responsables de la production d'insuline. Cette destruction entraîne une carence absolue en insuline, nécessitant un traitement substitutif à vie par insulinothérapie [2].

Cette forme de diabète survient généralement chez l'enfant ou l'adulte jeune, bien qu'elle puisse apparaître à tout âge. Elle est souvent associée à la présence d'auto-anticorps spécifiques et peut évoluer rapidement vers des complications aiguës telles que l'acidocétose diabétique en l'absence de traitement [1].

2.1.2 Diabète de type 2

Le diabète de type 2 est la forme la plus répandue, représentant la majorité des cas de diabète dans le monde. Il résulte d'une combinaison de deux mécanismes principaux:

- Une **insulinorésistance**, où les cellules de l'organisme répondent moins efficacement à l'insuline ;
- Une **diminution progressive de la sécrétion d'insuline** par les cellules β pancréatiques.

Cette forme est fortement associée à des facteurs de risque tels que l'obésité, la sédentarité, l'âge avancé et les antécédents familiaux. Contrairement au diabète de type 1, le diabète de type 2 évolue souvent de manière progressive et peut rester asymptomatique pendant plusieurs années, ce qui retarde son diagnostic [4].

2.1.3 Diabète gestationnel

Le diabète gestationnel correspond à une hyperglycémie diagnostiquée pour la première fois pendant la grossesse, généralement au cours du deuxième ou du troisième trimestre, chez une femme ne présentant pas de diabète connu auparavant [4].

Cette forme de diabète est liée aux modifications hormonales de la grossesse, qui entraînent une insulinorésistance accrue. Bien que le diabète gestationnel soit souvent transitoire et

disparaisse après l'accouchement, il représente un facteur de risque important pour la santé de la mère et du fœtus.

Sur le plan pronostique, il augmente le risque de complications materno-fœtales, telles que la macrosomie, l'accouchement prématuré et les troubles métaboliques néonatals. De plus, les femmes ayant présenté un diabète gestationnel ont un risque accru de développer un diabète de type 2 à long terme [4].

La figure 1.2 présente la classification des principaux types de diabète, à savoir le diabète de type 1, le diabète de type 2 et le diabète gestationnel.

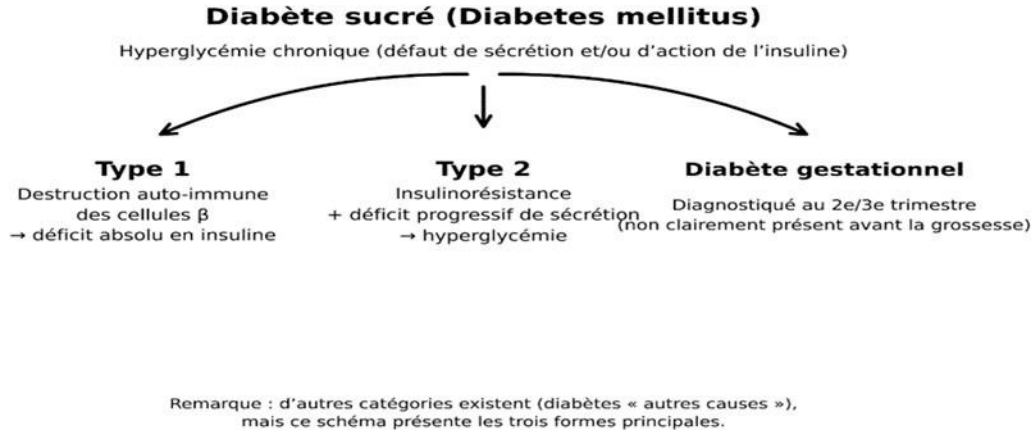


Figure 1.2 : Classification des types de diabète (Type 1, Type 2, gestationnel).

Afin de synthétiser les différences cliniques et physiopathologiques entre les principales formes de diabète, le tableau suivant présente une comparaison détaillée entre le diabète de type 1, le diabète de type 2 et le diabète gestationnel. Il met en évidence les éléments clés permettant de les distinguer, notamment les mécanismes sous-jacents, les caractéristiques cliniques, le contexte d'apparition et les stratégies de prise en charge. Cette comparaison facilite la compréhension globale de la maladie et oriente la décision médicale.

Critère	**Diabète type 1**	**Diabète type 2**	**Diabète gestationnel**
Mécanisme	Destruction auto-immune des cellules β	Insulinorésistance + déficit insulinique	Hyperglycémie pendant grossesse
Âge fréquent	Enfant / jeune	Adulte	Femme enceinte
Début	Rapide	Progressif	Pendant grossesse
Poids fréquent	Normal / maigre	Surpoids fréquent	Variable
Symptômes	Polyurie, amaigrissement	Souvent discrets	Souvent asymptomatique
Traitement principal	Insuline obligatoire	Hygiène + ADO ± insuline	Hygiène ± insuline

Tableau 1.1 : Classification clinique des types de diabète

L’analyse comparative met en évidence que chaque type de diabète possède des caractéristiques physiopathologiques et cliniques spécifiques, ce qui implique des approches thérapeutiques différentes. Elle souligne également l’importance du dépistage précoce, en particulier pour le diabète de type 2 et le diabète gestationnel, afin de prévenir les complications. Cette distinction est essentielle pour une prise en charge personnalisée et efficace des patients.

2.2 Aperçu des statistiques épidémiologiques récentes

Les données internationales récentes confirment une progression continue du diabète. Le nombre de personnes vivant avec un diabète est passé de 200 millions en 1990 à 830 millions en 2022 [1].

La figure 1.3 illustre l’évolution du nombre de personnes vivant avec le diabète dans le monde entre 1990 et 2022. Elle met en évidence une augmentation continue et marquée du nombre de cas au cours de cette période, traduisant une progression importante de la maladie à l’échelle mondiale. Cette tendance souligne l’ampleur croissante du diabète en tant que problème majeur de santé publique et l’importance du renforcement des stratégies de prévention et de prise en charge.

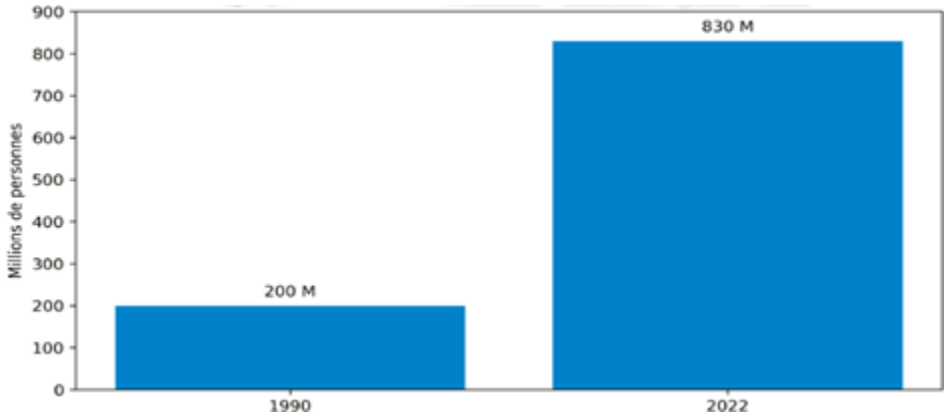


Figure 1.3 : Évolution du nombre de personnes vivant avec le diabète dans le monde (1990–2022).

D’après la Fédération Internationale du Diabète Diabètes Atlas (11e édition), environ 589 millions d’adultes (20–79 ans) vivaient avec le diabète en 2024 (soit environ 1 adulte sur 9), et ce nombre pourrait atteindre 853 millions d’ici 2050 [5]. La charge de mortalité demeure élevée, avec environ 3,4 millions de décès attribués au diabète en 2024 [6]. À l’échelle des États-Unis, les statistiques nationales rapportent environ 38,4 millions de personnes vivant avec le diabète, soit 11,6 % de la population, incluant une proportion importante de cas non diagnostiqués [7]. Concernant l’Algérie, les estimations de la Fédération Internationale du

Diabète en 2024 indiquent environ 4 761 800 adultes (20–79 ans) vivant avec le diabète, pour une prévalence d'environ 17,5 % ; les projections annoncent une hausse à environ 7,9 millions de cas d'ici 2050 [8]. Ces indicateurs confirment que le diabète reste une priorité de santé publique mondiale et nationale, en particulier dans les contextes où le dépistage précoce et l'accès aux soins doivent être renforcés [1][5].

Le tableau 1.2 présente les statistiques épidémiologiques du diabète selon différentes régions du monde. Il met en évidence une disparité importante de la prévalence et du nombre de cas estimés entre les régions. À l'échelle mondiale, le diabète touche environ 589 millions de personnes, avec une prévalence de 10,5 %. La région MENA enregistre la prévalence la plus élevée (16,2 %), suivie de l'Europe et de l'Asie du Sud-Est, tandis que l'Afrique présente une prévalence plus faible mais reste fortement concernée par l'augmentation progressive des cas. Ces données soulignent l'impact global du diabète et la nécessité d'actions de prévention adaptées à chaque région.

Tableau 1.2: Statistiques épidémiologiques du diabète

Région	**Cas estimés**	**Prévalence**
Monde	589 millions	10.5 %
Afrique	24 millions	4.5 %
Europe	61 millions	9.2 %
Moyen-Orient et Afrique du Nord.	95 millions	16.2 %
Asie du Sud-Est	90 millions	8.8 %

2.3 Facteurs de risque

Les facteurs de risque du diabète varient selon le type de diabète, mais plusieurs déterminants sont bien établis pour le diabète de type 2 [1]. Les principaux facteurs modifiables incluent le surpoids/obésité, la sédentarité et une alimentation déséquilibrée, qui favorisent l'insulinorésistance et l'apparition de l'hyperglycémie [8]. La figure 1.4 présente les principaux facteurs de risque associés au diabète.

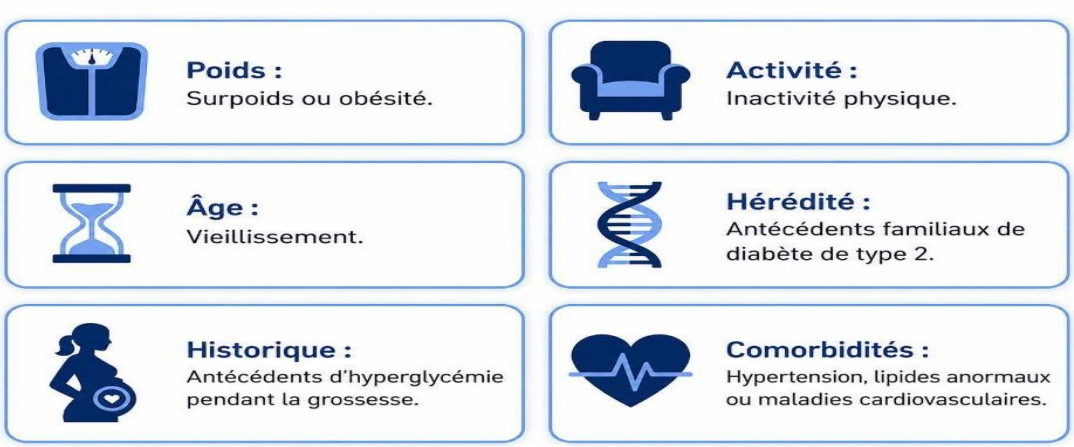


Figure 1.4 : Principaux facteurs de risque du diabète [76].

Parmi les facteurs non modifiables figurent l'âge, les antécédents familiaux de diabète et certaines susceptibilités génétiques. Des comorbidités comme l'hypertension artérielle, la dyslipidémie, le prédiabète et le syndrome des ovaires polykystiques augmentent également le risque de diabète de type 2 [9]. Concernant le diabète gestationnel, les facteurs de risque fréquemment rapportés sont le surpoids avant grossesse, l'antécédent personnel de diabète gestationnel, les antécédents familiaux de diabète de type 2 et l'âge maternel plus élevé [10]. Au niveau populationnel, l'urbanisation, le vieillissement démographique et la diminution de l'activité physique contribuent à l'augmentation du fardeau du diabète dans de nombreux pays [11].

L'identification précoce de ces facteurs est essentielle pour orienter les stratégies de prévention primaire, de dépistage ciblé et d'éducation thérapeutique [1].

Le tableau 1.3 présente les principaux facteurs de risque du diabète en les classant en deux catégories : les facteurs non modifiables et les facteurs modifiables.

Tableau 1.3 : Facteurs de risque du diabète

Facteurs non modifiables	**Facteurs modifiables**
Âge	Obésité
Antécédents familiaux	Sédentarité
Prédisposition génétique	Mauvaise alimentation
Diabète gestationnel antérieur	Tabagisme
Syndrome des ovaires polykystiques	Hypertension
Origine ethnique	Dyslipidémie

2.4 Principales complications du diabète

À long terme, l'hyperglycémie chronique active plusieurs voies délétères (stress oxydatif, inflammation, dysfonction endothéliale), responsables des lésions vasculaires et tissulaires. Les complications du diabète se répartissent en deux grandes catégories :

- Les complications microvasculaires, comprenant la rétinopathie, la néphropathie et la neuropathie.
- Les complications macrovasculaires, incluant la cardiopathie ischémique, l'accident vasculaire cérébral ainsi que l'artériopathie périphérique.

La figure 1.5 illustre les principales complications du diabète, en distinguant les atteintes microvasculaires et macrovasculaires.

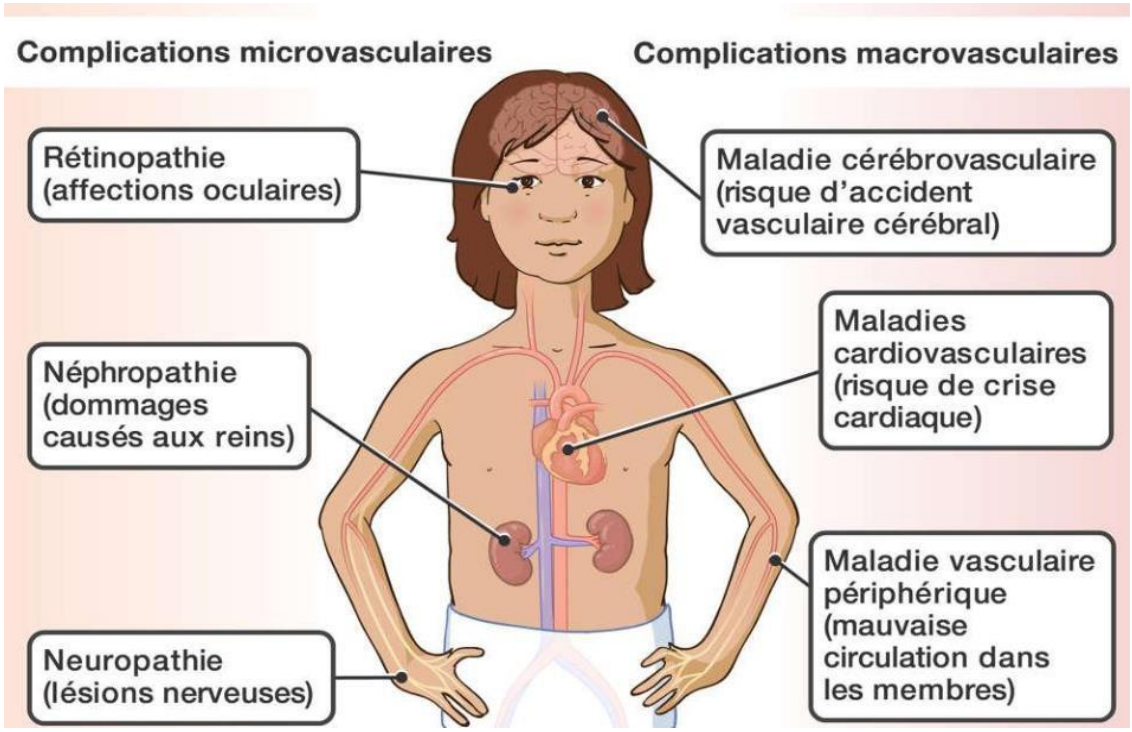


Figure 1.5 : Complications microvasculaires et macrovasculaires du diabète [77].

Le diabète augmente aussi le risque d'ulcères du pied et d'amputation, notamment en présence de neuropathie et de troubles circulatoires. Ces complications justifient une prise en charge précoce et intégrée, centrée sur le contrôle glycémique, la réduction des facteurs de risque cardiovasculaire et le dépistage régulier des atteintes d'organes cibles [13].

Tableau 1.4: Complications du diabète

Type	Complications principales
Microvasculaires	Rétinopathie, néphropathie, neuropathie
Macrovasculaires	AVC, infarctus, artériopathie
Pied diabétique	Ulcères, infections, amputations
Complications aiguës	Hypoglycémie, acidocétose

3. Diagnostic et bilan initial

3.1 Critères cliniques

Les critères cliniques du diabète reposent d'abord sur l'identification de symptômes évocateurs d'hyperglycémie. Les manifestations les plus fréquentes sont la polyurie, la polydipsie, l'amaigrissement involontaire, la fatigue et la vision floue. Chez certains patients, notamment dans le diabète de type 2, la symptomatologie peut être discrète ou d'installation progressive, ce qui retarde le diagnostic [1]. Sur le plan diagnostique, la présence de symptômes classiques associée à une glycémie plasmatique aléatoire ≥ 200 mg/dL (11,1 mmol/L) constitue un critère clinique majeur en faveur du diabète. En l'absence de symptômes francs ou d'hyperglycémie évidente, la confirmation doit reposer sur des tests biologiques standardisés (HbA1c, glycémie à jeun, HGPO) selon les recommandations en vigueur [2]. L'évaluation clinique initiale inclut également l'examen des signes de déshydratation, la recherche de perte pondérale, l'évaluation cardiovasculaire de base (pression artérielle, fréquence cardiaque) et le dépistage précoce des complications (neuropathie périphérique, atteinte rénale et oculaire). Ainsi, les critères cliniques ont une double utilité : orienter le diagnostic et stratifier le risque pour adapter rapidement la prise en charge [14].

Tableau 1.5 : Signes cliniques évocateurs du diabète

Signe clinique	Description
Polyurie	Urines fréquentes
Polydipsie	Soif excessive
Polyphagie	Faim augmentée
Amaigrissement	Perte de poids
Fatigue	Asthénie persistante
Vision floue	Trouble visuel temporaire
Infections répétées	Mycoses, infections urinaires

3.2 Critères biologiques

Le diagnostic biologique du diabète repose sur des seuils internationaux standardisés. Il peut être retenu devant l'un des critères suivants : **HbA1c ≥ 6,5 % (48 mmol/mol)**, **glycémie plasmatique à jeun ≥ 126 mg/dL (7,0 mmol/L)**, **glycémie plasmatique à 2 heures après HGPO 75 g ≥ 200 mg/dL (11,1 mmol/L)**, ou **glycémie plasmatique aléatoire ≥ 200 mg/dL (11,1 mmol/L)** en présence de signes classiques d'hyperglycémie. En l'absence d'hyperglycémie évidente, le diagnostic doit être **confirmé** par un second résultat anormal (même test répété ou test différent), afin de limiter les faux positifs [2]. Les seuils de **prédiabète** sont également utiles pour le dépistage : HbA1c entre **5,7 et 6,4 %**, glycémie à jeun entre **100 et 125 mg/dL**, ou glycémie à 2 h HGPO entre **140 et 199 mg/dL** [15].

L'HbA1c reflète l'exposition glycémique moyenne des 2–3 derniers mois, mais son interprétation peut être biaisée dans certaines situations (anémies, hémoglobinopathies, insuffisance rénale avancée), où les mesures directes de glucose restent prioritaires [2]. Ainsi, l'association des paramètres glycémiques (à jeun, post-charge, aléatoire) et de l'HbA1c permet un diagnostic plus robuste et une meilleure stratification du risque métabolique [2].

Afin de synthétiser les principaux seuils biologiques utilisés en pratique clinique pour le diagnostic du diabète, le Tableau 1.6 présente les tests de référence et leurs critères diagnostiques.

Tableau 1.6 : Critères diagnostiques du diabète (OMS/ADA)

Test (méthode)	Critère diagnostique du diabète	Remarques
Glycémie à jeun (FPG)	≥ 7,0 mmol/L (≥ 126 mg/dL)	Jeûne ≥ 8 h.
HGPO / OGTT 75 g (glycémie à 2 h)	≥ 11,1 mmol/L (≥ 200 mg/dL)	Mesure à 2 h après charge orale de 75 g de glucose.
HbA1c	≥ 6,5 % (≥ 48 mmol/mol)	Méthode de labo standardisée recommandée (critère ADA).
Glycémie "au hasard" + symptômes	≥ 11,1 mmol/L (≥ 200 mg/dL)	Avec symptômes classiques (polyurie, polydipsie, amaigrissement…) ou hyperglycémie franche.

Le diagnostic du diabète repose sur des critères biologiques standardisés définis par l'Organisation Mondiale de la Santé (OMS) et l'American Diabètes Association (ADA). Ces critères incluent la glycémie à jeun, l'HbA1c, l'hyperglycémie provoquée par voie orale

(HGPO) ainsi que la glycémie aléatoire associée à des symptômes cliniques.

3.3 Outils d'aide au diagnostic

Les outils d'aide au diagnostic du diabète regroupent des approches **biologiques**, **cliniques** et **numériques** complémentaires.

En première intention, les tests de laboratoire (glycémie à jeun, HGPO, HbA1c) constituent la base diagnostique et permettent de confirmer ou d'exclure le diabète selon des seuils standardisés [16].

La figure 1.6 présente le processus général de diagnostic du diabète, depuis le dépistage initial jusqu'à la confirmation du diagnostic.

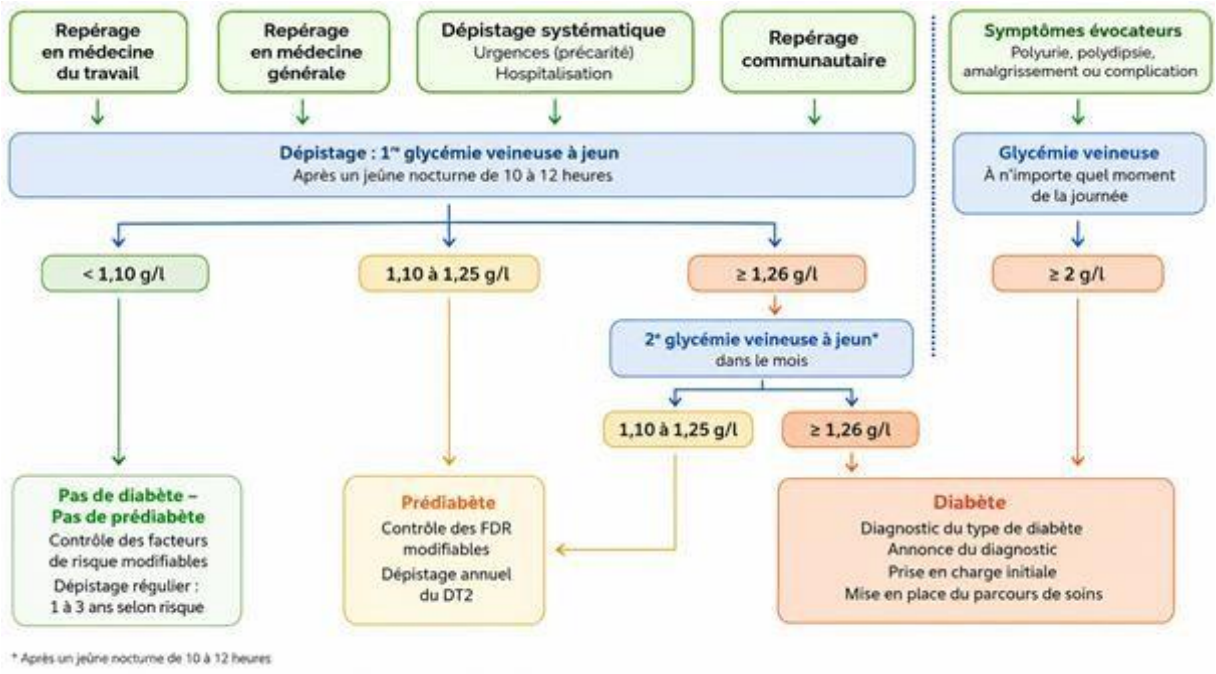


Figure 1.6 : processus général de diagnostic du diabète [78].

Au quotidien, les médecins utilisent des questionnaires simples et des mesures rapides (comme l'âge, le poids ou les habitudes de vie) pour repérer d'abord les patients les plus en danger. Cela permet de cibler efficacement les personnes qui devront passer, dans un second temps, des examens biologiques ou des prises de sang pour confirmer le diagnostic [17].

Dans les stratégies de dépistage, des recommandations de santé publique proposent un ciblage des adultes asymptomatiques à risque (notamment en surpoids/obésité), afin d'améliorer la détection précoce du prédiabète et du diabète [18]. Pour le suivi et l'aide à la décision thérapeutique, les outils numériques (dossiers médicaux électroniques, calculateurs de risque, alertes cliniques, et parfois mesures continues du glucose chez les patients déjà diagnostiqués) peuvent renforcer la personnalisation des décisions médicales [19]. Cependant, ces outils ne remplacent pas le jugement clinique : leur interprétation doit tenir compte du

contexte du patient, des comorbidités et de la qualité des données disponibles [16]. Ainsi, une approche intégrée combinant évaluation clinique, tests biologiques et outils d'aide à la décision améliore la pertinence diagnostique et l'orientation précoce des patients [17].

4. Prévention et prise en charge

4.1 Mesures hygiéno-diététiques

Les mesures hygiéno-diététiques constituent la base de la prévention et de la prise en charge du diabète, en particulier du diabète de type 2. Elles visent à améliorer le contrôle glycémique, réduire le risque cardiovasculaire et prévenir les complications à long terme [1]. Sur le plan nutritionnel, les recommandations privilégient une alimentation de qualité (légumes, fruits entiers, légumineuses, céréales complètes, protéines maigres), avec réduction des sucres ajoutés, des aliments ultra-transformés et des graisses de mauvaise qualité. La perte pondérale est un objectif central chez les patients en surpoids ou obèses : une réduction modeste du poids corporel améliore déjà les paramètres métaboliques, tandis qu'une perte plus importante peut améliorer fortement l'équilibre glycémique [20]. L'activité physique régulière est également essentielle : chez l'adulte, les recommandations internationales soutiennent au moins 150 minutes/semaine d'activité d'intensité modérée (ou équivalent), en limitant la sédentarité [21]. Enfin, l'arrêt du tabac, la réduction de la consommation d'alcool, l'éducation thérapeutique et l'accompagnement comportemental renforcent l'adhésion et l'efficacité des mesures hygiéno-diététiques [22]. 4.2 Traitements médicamenteux Le traitement médicamenteux du diabète doit être individualisé selon le type de diabète, le profil du patient et la présence de comorbidités cardio-rénales. Dans le diabète de type 1, l'insulinothérapie est indispensable dès le diagnostic, avec schémas basal-bolus ou perfusion sous-cutanée continue selon les besoins cliniques [23]. Dans le **diabète de type 2**, la **metformine** reste un traitement de référence en première ligne chez de nombreux patients, en association avec les mesures hygiéno-diététiques, sauf contre-indication.

En cas de contrôle glycémique insuffisant ou de risque cardiovasculaire/rénal élevé, les recommandations privilégient l'ajout d'agents à bénéfice démontré, notamment les **inhibiteurs SGLT2** et/ou les **agonistes des récepteurs GLP-1**, indépendamment dans certains cas de la valeur initiale de l'HbA1c [24].

Les autres classes (inhibiteurs DPP-4, sulfamides hypoglycémiants, thiazolidinediones, insuline) gardent une place selon l'efficacité attendue, le risque d'hypoglycémie, le poids, la fonction rénale, le coût et l'accessibilité [25]. L'intensification thérapeutique doit être progressive et évaluée régulièrement par le suivi de l'HbA1c, de la glycémie, des effets indésirables et de l'adhésion thérapeutique [24]. Ainsi, la prise en charge pharmacologique

moderne repose sur une stratégie centrée patient, orientée non seulement vers la baisse de la glycémie, mais aussi vers la réduction du risque cardiovasculaire, rénal et de mortalité [26]. 4.3 Suivi clinique et pronostic Le suivi clinique du diabète est un processus continu visant à évaluer l'équilibre glycémique, prévenir les complications et ajuster le traitement de manière individualisée. L'évaluation régulière inclut en priorité l'HbA1c, l'autosurveillance glycémique (ou CGM selon le profil), la pression artérielle, le bilan lipidique, la fonction rénale (eGFR) et l'albuminurie [27]. Pour le diabète de type 2, les recommandations de pratique insistent sur un rythme de contrôle périodique de l'HbA1c (plus rapproché si le traitement change ou si l'objectif n'est pas atteint), puis espacé quand l'équilibre est stable [28]. Le suivi doit aussi intégrer le dépistage systématique des complications : fond d'œil, examen des pieds, évaluation neuropathique et surveillance rénale, afin d'identifier précocement les atteintes microvasculaires. Sur le plan pronostique, la qualité du contrôle glycémique et la prise en charge des facteurs de risque cardiovasculaire (HTA, dyslipidémie, tabac, obésité) influencent fortement l'évolution à long terme [29]. Un suivi structuré, multidisciplinaire et centré sur le patient améliore l'adhésion thérapeutique, réduit les hospitalisations et diminue le risque de complications invalidantes [30]. Ainsi, le pronostic du diabète dépend moins d'un paramètre isolé que de la continuité du suivi clinique et de l'intensification thérapeutique précoce quand nécessaire [30].

Tableau 1.7: Suivi clinique du diabète

Paramètre	**Fréquence**	**Objectif**
HbA1c	Tous les 3–6 mois	Contrôle glycémique
Glycémie	Régulière	Détecter hypo/hyperglycémie
Tension artérielle	Chaque consultation	Réduire risque CV
Bilan lipidique	1 fois/an	Prévention cardiovasculaire
Fonction rénale	1 fois/an	Dépister néphropathie
Fond d'œil	1 fois/an	Dépister rétinopathie
Examen du pied	1 fois/an	Prévenir ulcères

5. Limites des approches conventionnelles

Malgré les progrès des recommandations cliniques, les approches conventionnelles du diabète présentent plusieurs limites structurelles. D'abord, la charge mondiale continue d'augmenter fortement, ce qui montre que les stratégies actuelles de prévention et de prise en charge restent insuffisantes à l'échelle populationnelle. L'OMS rapporte une hausse de 200 millions de personnes vivant avec le diabète en 1990 à 830 millions en 2022 [31].

Une seconde limite concerne le sous-diagnostic. Une part importante des personnes atteintes n'est pas diagnostiquée à temps, en particulier dans les contextes à ressources limitées, ce qui retarde l'intervention et favorise les complications évitables. Les estimations IDF 2025 confirment l'ampleur de ce problème à l'échelle internationale [32].

Sur le plan diagnostique, les tests biologiques standards (glycémie, HbA1c, HGPO) sont indispensables, mais leur performance peut varier selon le contexte clinique (anémies, hémoglobinopathies, insuffisance rénale, variabilité analytique inter-laboratoires), et ils nécessitent souvent confirmation et interprétation spécialisée. Les recommandations ADA insistent d'ailleurs sur ces précautions méthodologiques [33].

Les approches conventionnelles sont aussi souvent **réactives** plutôt que prédictives : elles détectent la maladie une fois l'hyperglycémie installée, alors que la phase préclinique peut durer plusieurs années. Cela limite l'anticipation personnalisée du risque et la priorisation proactive des patients à haut risque [34].

Enfin, la transposition des recommandations en pratique réelle se heurte à des contraintes d'accès aux soins, de continuité du suivi, d'éducation thérapeutique et d'adhésion, surtout dans les pays à revenu faible ou intermédiaire. Ces limites justifient l'intérêt d'approches complémentaires (modèles de risque, outils d'aide à la décision, apprentissage automatique) pour améliorer le dépistage précoce et la stratification des patients

6. Conclusion

Le diabète est une pathologie chronique complexe dont la progression mondiale confirme l'urgence d'améliorer le dépistage et la prise en charge. Ses formes principales (type 1, type 2, gestationnel) diffèrent par leurs mécanismes, mais partagent un risque élevé de complications graves à long terme. L'analyse épidémiologique montre une hausse continue de la prévalence, avec un poids particulièrement important dans les contextes à ressources limitées. Les facteurs de risque modifiables et non modifiables interagissent fortement, ce qui impose des stratégies de prévention adaptées aux profils des patients. Les critères diagnostiques biologiques (glycémie, HbA1c, HGPO) restent essentiels, mais nécessitent une interprétation clinique rigoureuse et contextualisée. Le suivi médical régulier, associé aux mesures hygiéno-diététiques et aux traitements pharmacologiques, demeure indispensable pour limiter la progression de la maladie.

Malgré cela, des limites persistent : sous-diagnostic, diagnostic tardif, inégalités d'accès aux soins et difficultés d'adhésion thérapeutique. Ces insuffisances réduisent l'efficacité des approches classiques et justifient le recours à des méthodes complémentaires d'aide à la décision.

Dans cette perspective, l'apprentissage automatique offre un cadre pertinent pour améliorer la détection précoce et la stratification du risque. L'étape suivante consiste à appliquer une méthodologie expérimentale rigoureuse pour mesurer l'impact réel de ces approches sur la classification du diabète.

Chapitre 2 : Équilibrage des données et intelligence artificielle appliquée à la classification médicale

1. Introduction

La prédiction du diabète représente aujourd'hui un axe de recherche important dans le domaine médical, en raison de l'augmentation continue du nombre de patients atteints de cette maladie chronique et des complications qui y sont associées. Les progrès récents de l'intelligence artificielle et des technologies de l'information ont favorisé le développement de systèmes intelligents capables d'assister les professionnels de santé dans le diagnostic précoce et l'analyse des données cliniques. Dans ce cadre, les techniques d'apprentissage automatique permettent d'exploiter efficacement les données médicales afin d'identifier des relations et de produire des prédictions plus précises.

Toutefois, les bases de données médicales présentent souvent un déséquilibre entre les classes, où les cas diabétiques sont moins nombreux que les cas non diabétiques. Cette répartition inégale peut influencer les performances des modèles de classification et entraîner un biais vers la classe majoritaire, réduisant ainsi la capacité des modèles à détecter correctement les patients atteints de diabète. La prise en compte de cette problématique est donc essentielle pour améliorer la fiabilité des systèmes de prédiction.

Dans ce contexte, ce chapitre présente les notions liées au déséquilibre des données ainsi que les concepts fondamentaux de l'intelligence artificielle, du Machine Learning et du Deep Learning. Il aborde également les différentes approches d'apprentissage automatique, les principaux algorithmes de classification utilisés dans la prédiction du diabète et l'importance de l'intelligence artificielle dans l'analyse des données médicales.

2. Équilibrage des données

2.1 Généralités sur le déséquilibre des données

Le déséquilibre des classes constitue un problème fréquent dans les applications d'apprentissage automatique, particulièrement dans le domaine médical. Il apparaît lorsque le nombre d'instances appartenant à une classe est largement supérieur à celui des autres classes. Dans les jeux de données liés au diagnostic du diabète, les patients non diabétiques représentent généralement la classe majoritaire, tandis que les patients diabétiques constituent la classe minoritaire. Cette situation peut être causée par la rareté de certains cas pathologiques, les difficultés de collecte des données médicales ou encore le coût élevé des examens cliniques.

Le déséquilibre des données influence directement les performances des modèles de

classification. En effet, plusieurs algorithmes d'apprentissage automatique tendent à privilégier la classe majoritaire afin d'obtenir un taux de précision global élevé. Par conséquent, les modèles peuvent produire de mauvaises performances dans la détection des cas minoritaires, ce qui est particulièrement problématique dans le contexte médical où l'identification correcte des patients atteints de la maladie est essentielle. Ce phénomène entraîne un biais vers la classe majoritaire et réduit la capacité du modèle à généraliser efficacement sur les données réelles. [35]

2.2 Méthodes d'équilibrage

Le déséquilibre des classes constitue un défi majeur dans les tâches de classification, car il peut entraîner une dégradation des performances des modèles, notamment pour la classe minoritaire. Afin de remédier à ce problème, plusieurs méthodes d'équilibrage ont été proposées dans la littérature. Ces approches visent à améliorer la représentation des différentes classes et à renforcer la capacité des modèles à identifier correctement les instances minoritaires. Dans ce qui suit, nous présenterons les principales catégories de méthodes d'équilibrage, à savoir les techniques de modification au niveau des données, les approches de modification au niveau des algorithmes, ainsi que les méthodes hybrides combinant ces deux stratégies.

2.2.1 Modification au niveau de données

Les méthodes de modification au niveau des données sont largement utilisées pour traiter le problème du déséquilibre des classes dans les bases de données médicales. Elles consistent à ajuster la distribution des données d'apprentissage afin d'obtenir un meilleur équilibre entre la classe majoritaire et la classe minoritaire. Ces approches permettent d'améliorer les performances des modèles de classification et de réduire le biais vers la classe majoritaire. Elles agissent directement sur les données avant l'entraînement du modèle à travers différentes techniques de sur-échantillonnage, de sous-échantillonnage et de génération synthétique des données [36].

Dans ce qui suit, nous allons détailler les différentes méthodes d'équilibrage basées sur la modification des données.

2.2.1.1 Random Oversampling

Cette approche consiste à dupliquer de manière aléatoire des exemples de la classe minoritaire, ce qui permet d'obtenir une distribution plus équilibrée. Cette approche est simple et conserve l'information de la classe majoritaire, mais elle peut entraîner le surapprentissage, car des exemples de la classe minoritaire sont dupliqués et sont donc "faciles à apprendre" [37].

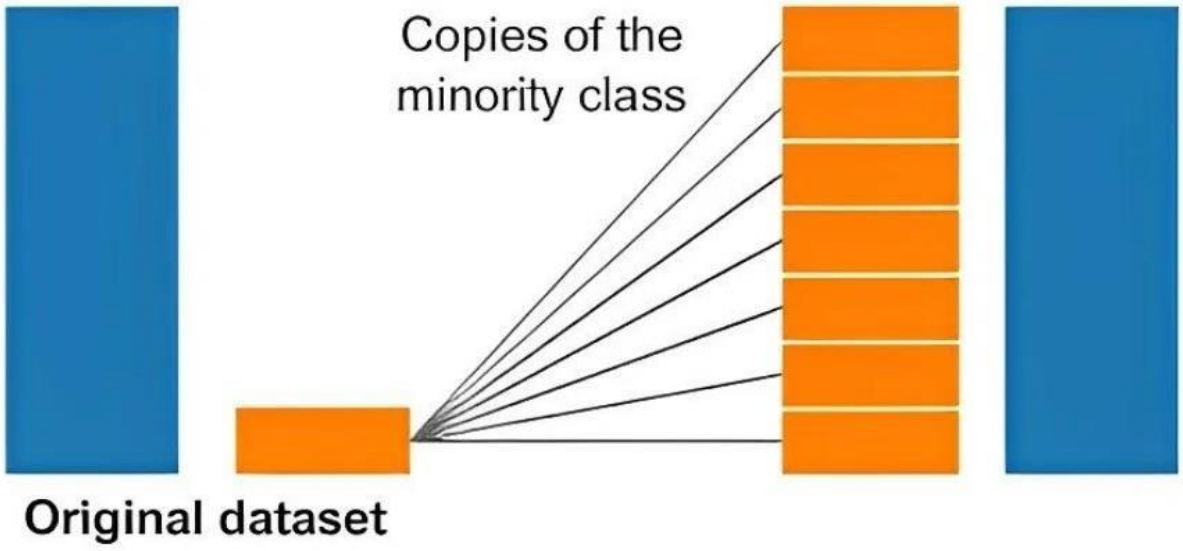


Figure 2.1 : Principe du Random Oversampling [79].

2.2.1.2 Random Undersampling

À l'inverse, cette approche réduit aléatoirement la taille de la classe majoritaire en supprimant un certain nombre de données appartenant à cette dernière. Cette approche diminue le coût computationnel et peut également limiter la domination de la classe majoritaire. Cependant, elle peut éliminer un certain nombre de données importantes, ce qui peut altérer la généralisation et la variance du modèle [38].

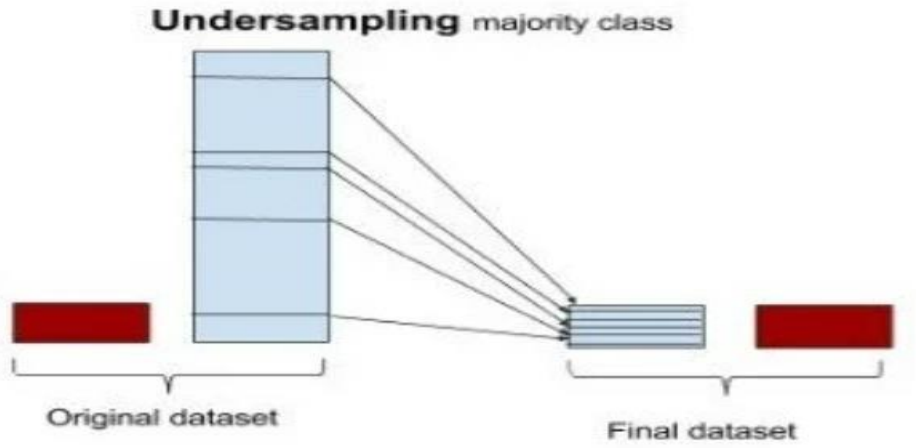


Figure 2.2 : Principe du Random Undersampling [80].

2.2.1.3 SMOTE

La méthode SMOTE (Synthetic Minority Over-sampling Technique) utilise la synthèse de données pour la classe mineure en interpolant les données avec les données des voisinages proches d'une instance de la classe mineure plutôt que la copie des données. La méthode SMOTE réduit les problèmes d'overfitting en copiant les données, améliore la frontière

décisive, mais peut introduire du bruit lorsque le taux d'hétérogénéité de la classe mineure est élevé ou lorsque les classes se chevauchent fortement [39].

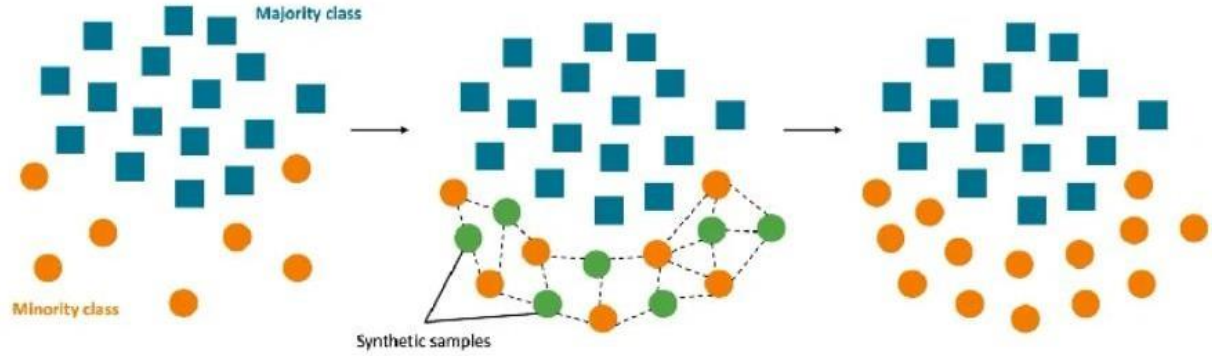


Figure 2.3 : SMOTE (génération d'exemples synthétiques) [81].

2.2.1.4 ADASYN

ADASYN (Adaptive Synthetic Sampling) est une extension des techniques de sur-échantillonnage synthétique. Elle génère davantage d'exemples artificiels dans les régions où la classe minoritaire est difficile à apprendre, notamment près des frontières de décision et dans les zones fortement dominées par la classe majoritaire. Cette approche permet de mieux équilibrer la distribution des données tout en mettant l'accent sur les exemples les plus complexes à classifier.

Il est important de noter que le choix de la méthode de sur-échantillonnage dépend fortement de la nature des données, du modèle utilisé ainsi que du compromis souhaité entre la sensibilité de la classe minoritaire, la robustesse du modèle et le risque de surapprentissage. [40]

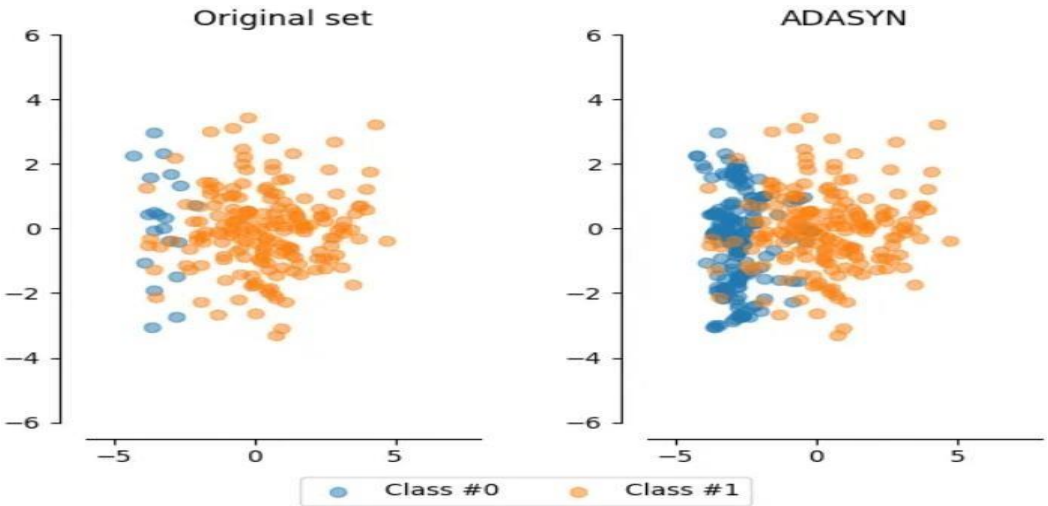


Figure 2.4 : ADASYN (génération adaptative selon la difficulté locale) [82].

2.2.1.6 Tomek Links

Tomek Links est une méthode de nettoyage des données qui identifie les paires d'observations de classes différentes très proches et supprime les exemples de la classe majoritaire. Elle permet d'améliorer la frontière de décision en réduisant les zones de chevauchement [42].

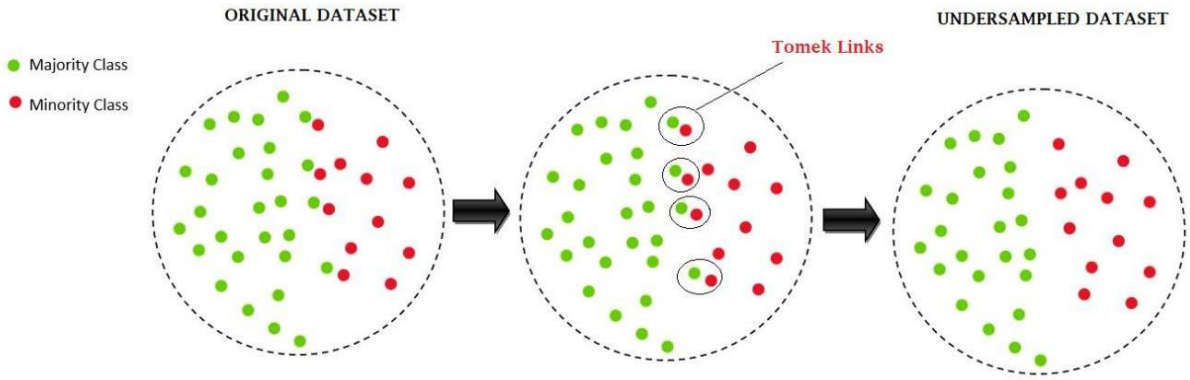


Figure 2.6 : Principe de Tomek Links [84].

2.2.2 Modification au niveau des algorithmes

Les méthodes de modification au niveau des algorithmes constituent une approche alternative pour traiter le déséquilibre des classes, sans modifier directement la distribution des données. Contrairement aux techniques basées sur les données, ces méthodes agissent sur le processus d'apprentissage en adaptant les modèles afin de mieux prendre en compte la classe minoritaire. Elles visent à réduire le biais des algorithmes vers la classe majoritaire en intégrant des mécanismes de pénalisation ou d'ajustement lors de l'entraînement. Ainsi, elles permettent d'améliorer la sensibilité des modèles aux cas minoritaires tout en conservant l'ensemble des données originales [43].

Dans la suite, nous présenterons les principales méthodes de modification au niveau des algorithmes.

2.2.2.1 Apprentissage sensible au coût (Cost-Sensitive Learning)

L'apprentissage sensible au coût est une approche de Machine Learning qui attribue des coûts différents aux erreurs de classification selon leur importance. Ainsi, les erreurs commises sur la classe minoritaire (souvent la plus critique, comme les patients diabétiques) sont pénalisées plus fortement que celles de la classe majoritaire. Cette stratégie permet d'orienter l'apprentissage du modèle afin de minimiser le coût global plutôt que le simple taux d'erreur, améliorant ainsi la détection des cas rares [43].

2.2.2.2 Pondération des classes

La pondération des classes est une technique qui consiste à attribuer des poids différents aux classes lors de l'entraînement d'un modèle. Les classes minoritaires reçoivent généralement un poids plus élevé afin de compenser leur sous-représentation dans les données. Cette méthode permet de réduire le biais du modèle vers la classe majoritaire sans modifier la distribution des données initiales [44].

2.2.2.3 Ajustement des seuils de décision

L'ajustement des seuils de décision consiste à modifier le seuil de classification utilisé par un modèle probabiliste afin de favoriser la détection de la classe minoritaire. Au lieu d'utiliser un seuil standard (souvent 0.5), celui-ci est ajusté pour améliorer la sensibilité du modèle, notamment dans les cas où la détection des positifs est prioritaire, comme en diagnostic médical [45].

2.2.3 Modification hybrides

Les méthodes hybrides combinent plusieurs techniques d'équilibrage afin de mieux traiter le problème du déséquilibre des classes. Elles associent généralement des approches basées sur les données et des approches algorithmiques dans le but d'améliorer les performances des modèles de classification. Ces méthodes cherchent à réduire le biais vers la classe majoritaire tout en limitant les problèmes de surapprentissage et de perte d'information [46].

Nous présentons ci-après les principales méthodes hybrides, reposant sur la combinaison de stratégies d'équilibrage au niveau des données et des algorithmes.

2.2.3.1 Combinaison sur-échantillonnage et sous-échantillonnage

Cette approche hybride consiste à combiner les techniques de sur-échantillonnage de la classe minoritaire avec les techniques de sous-échantillonnage de la classe majoritaire. L'objectif est d'obtenir un meilleur équilibre des données tout en réduisant le risque de surapprentissage causé par la duplication des données et la perte excessive d'informations due à la suppression d'exemples. Parmi les méthodes les plus connues figurent SMOTE-Tomek et SMOTE-ENN [47].

2.2.3.2 Combinaison équilibrage et algorithmes de classification

Cette approche consiste à intégrer les méthodes d'équilibrage des données directement avec les algorithmes de classification afin d'améliorer la détection de la classe minoritaire. Les données sont d'abord équilibrées, puis utilisées pour entraîner des modèles de classification tels que les arbres de décision, les forêts aléatoires ou les machines à vecteurs de support. Cette

combinaison permet souvent d'obtenir de meilleures performances prédictives et une meilleure généralisation des modèles [48].

Les techniques d'équilibrage des données jouent un rôle essentiel dans l'amélioration des performances des modèles de classification, particulièrement en présence de jeux de données déséquilibrés. Cependant, ces méthodes doivent être associées à des modèles d'apprentissage performants afin d'obtenir des résultats fiables.

Dans cette perspective, les sections suivantes présentent les notions d'intelligence artificielle, les principales approches de la Machine Learning et du Deep Learning ainsi que les modèles de classification utilisés dans ce travail.

3. Intelligence artificielle

L'intelligence artificielle (IA) désigne un ensemble de méthodes et de systèmes informatiques capables de réaliser des tâches nécessitant habituellement l'intelligence humaine, telles que l'analyse, le raisonnement, la prédiction ou la prise de décision. Selon l'Organisation de coopération et de développement économiques (OCDE), un système d'intelligence artificielle est un système basé sur une machine pouvant, à partir de données d'entrée, produire des prédictions, des recommandations, des contenus ou des décisions influençant des environnements réels ou virtuels. Cette définition met en évidence la capacité d'inférence, d'autonomie et d'adaptabilité des systèmes d'IA.

L'IA regroupe plusieurs domaines, notamment l'apprentissage automatique, l'apprentissage profond, le traitement du langage naturel et la vision par ordinateur. Son principal intérêt réside dans sa capacité à traiter de grandes quantités de données et à identifier des relations complexes entre les variables afin de fournir des résultats utiles dans plusieurs secteurs, particulièrement dans le domaine médical. Toutefois, l'utilisation de ces systèmes soulève également des enjeux liés à la fiabilité, la transparence, la sécurité et l'éthique, ce qui nécessite une supervision humaine et une gestion rigoureuse des risques. Ainsi, l'intelligence artificielle peut être considérée comme une discipline interdisciplinaire visant à développer des systèmes capables de simuler certaines fonctions cognitives humaines de manière efficace et fiable [49].

Tableau 2.1 : Différences entre Intelligence artificielle, Machine Learning et Deep Learnin

Domaine	Définition	Objectif	Exemple
Intelligence Artificielle (IA)	Domaine global visant à simuler l'intelligence humaine	Résoudre des tâches intelligentes	Robots, systèmes experts
Machine Learning (ML)	Sous-domaine de l'IA basé sur l'apprentissage à partir des données	Apprendre des modèles prédictifs	Classification, régression
Deep Learning (DL)	Sous-domaine du ML basé sur les réseaux de neurones profonds	Apprentissage de représentations complexes	Reconnaissance d'images, NLP

1. Approches de l'apprentissage automatique

1.1 Apprentissage supervisé

L'apprentissage supervisé est l'une des principales approches du Machine Learning. Il consiste à entraîner un modèle à partir de données étiquetées, c'est-à-dire des données pour lesquelles la sortie attendue est connue à l'avance. L'objectif est d'apprendre la relation entre les variables d'entrée et la variable cible afin de pouvoir prédire correctement les résultats sur de nouvelles données. Cette approche est principalement utilisée pour les tâches de classification et de régression. Dans le domaine médical, l'apprentissage supervisé est largement exploité pour le diagnostic, la prédiction de maladies et l'aide à la décision clinique. Parmi les algorithmes les plus utilisés figurent la régression logistique, les machines à vecteurs de support, les arbres de décision et les forêts aléatoires. Les performances de ces modèles dépendent fortement de la qualité des données, de la pertinence des variables et du processus d'entraînement adopté [50].

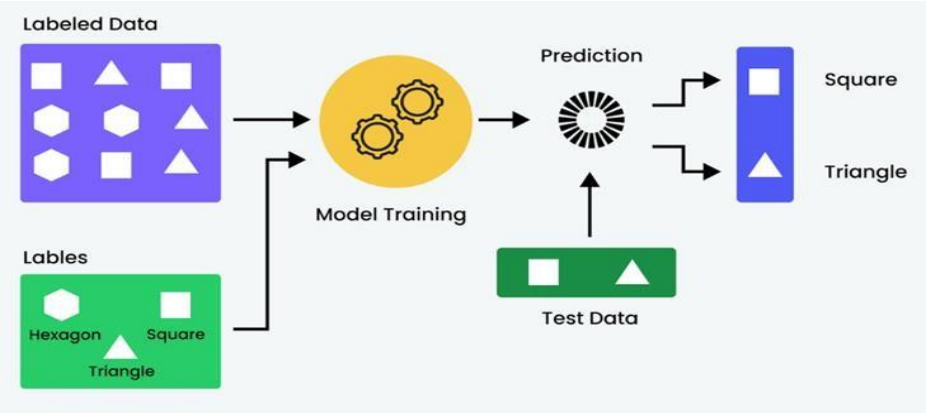


Figure 2.7 : Principe général de l'apprentissage supervisé [85].

1.2 Apprentissage non supervisé

L'apprentissage non supervisé désigne une approche dans laquelle le modèle apprend à partir de données non étiquetées, sans variable cible prédéfinie. Son objectif principal est de découvrir automatiquement des structures cachées, des relations ou des regroupements présents dans les données. Cette approche est souvent utilisée pour le clustering, la réduction de dimension et la détection d'anomalies. Contrairement à l'apprentissage supervisé, il ne cherche pas à prédire une classe précise, mais plutôt à analyser l'organisation des données et à identifier des similarités entre les observations. Dans les applications médicales, cette méthode peut être utilisée pour identifier des profils de patients ou détecter des comportements atypiques dans les bases de données cliniques [51].

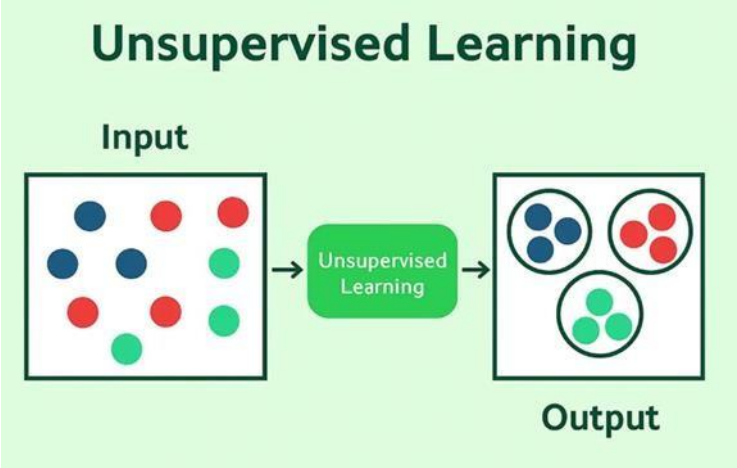


Figure 2.8 : Principe de l'apprentissage non supervisé par clustering [86].

1.3 Apprentissage semi-supervisé

L'apprentissage semi-supervisé représente une approche intermédiaire entre l'apprentissage supervisé et l'apprentissage non supervisé. Il combine une petite quantité de données étiquetées avec une grande quantité de données non étiquetées afin d'améliorer les performances du modèle. Cette méthode est particulièrement utile lorsque l'annotation des données est coûteuse ou difficile à réaliser, comme dans le domaine médical. Les données non étiquetées permettent au modèle de mieux comprendre la structure globale des données et d'améliorer sa capacité de généralisation. Les techniques semi-supervisées reposent souvent sur des méthodes telles que l'auto-apprentissage ou la propagation d'étiquettes. Dans les systèmes médicaux intelligents, cette approche contribue à réduire les coûts d'annotation tout en conservant des performances de classification satisfaisantes [52].

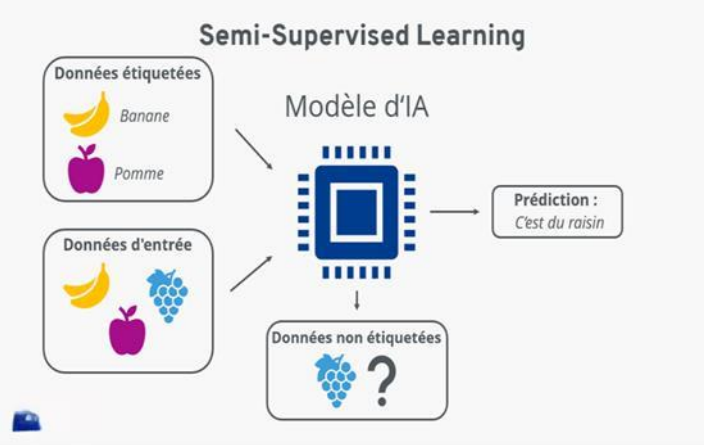


Figure 2.9 : Fonctionnement de l'apprentissage semi-supervisé [87].

1.4 Apprentissage par renforcement

L'apprentissage par renforcement est une approche dans laquelle un agent apprend progressivement à prendre des décisions en interagissant avec un environnement. Le modèle reçoit des récompenses ou des pénalités selon les actions effectuées et cherche à maximiser une récompense cumulative. Contrairement aux approches supervisées, il n'existe pas de réponses correctes prédéfinies ; l'apprentissage se fait par essais et erreurs. Cette méthode est utilisée dans plusieurs domaines tels que la robotique, les jeux intelligents, les systèmes autonomes et l'optimisation de processus complexes. Dans le secteur médical, l'apprentissage par renforcement peut être appliqué à l'optimisation des traitements personnalisés et à l'aide à la prise de décision thérapeutique [53].

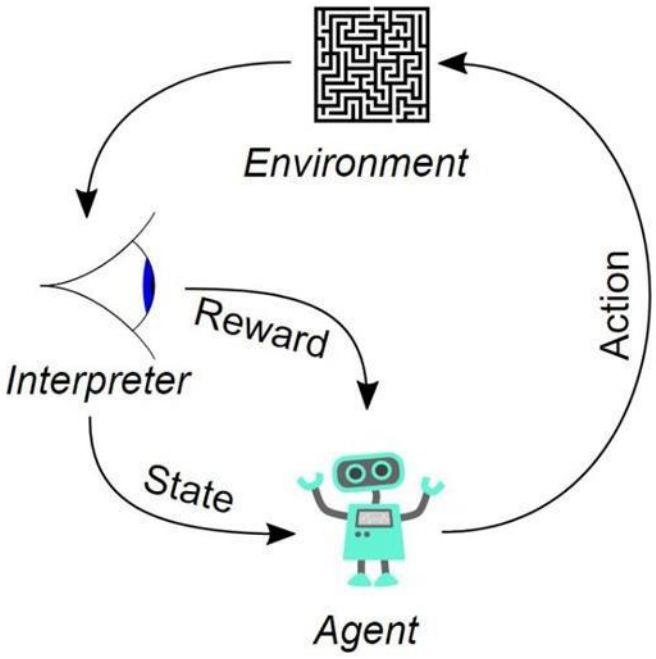


Figure 2.10 : Architecture de l'apprentissage par renforcement [88].

2. Approches de Machine Learning pour la classification

Les approches de Machine Learning pour la classification regroupent un ensemble de modèles supervisés visant à apprendre automatiquement des règles à partir des données afin d'attribuer une étiquette à chaque observation. Ces méthodes reposent sur l'exploitation d'exemples d'apprentissage pour construire un modèle capable de généraliser sur de nouvelles données. Dans ce qui suit, nous présenterons les principales approches utilisées dans notre projet de recherche.

2.1 Régression logistique

La régression logistique est un algorithme de classification supervisée utilisé pour prédire la probabilité d'appartenance à une classe donnée. Ce modèle repose sur la fonction logistique afin de modéliser une variable cible binaire à partir d'un ensemble de variables explicatives. La régression logistique est particulièrement appréciée pour sa simplicité, sa rapidité d'exécution et sa facilité d'interprétation grâce aux coefficients estimés. Dans le domaine médical, elle est largement utilisée pour la prédiction de maladies et l'analyse des facteurs de risque. Toutefois, ses performances peuvent être limitées lorsque les relations entre les variables sont fortement non linéaires [50].

2.2 Machine à vecteurs de support

Les machines à vecteurs de support (Support Vector Machine SVM) sont des algorithmes de classification supervisée qui cherchent à déterminer l'hyperplan optimal séparant les différentes classes avec une marge maximale. Les SVMs peuvent utiliser différents noyaux, notamment le noyau radial de base (RBF), afin de traiter des relations non linéaires entre les données. Cette méthode offre une bonne capacité de généralisation et des performances élevées sur des jeux de données complexes. Cependant, elle nécessite un réglage précis des hyperparamètres et peut être sensible aux données déséquilibrées [55].

2.3 Arbre de décision

L'arbre de décision est une méthode de classification fondée sur une structure hiérarchique composée de nœuds et de branches permettant de prendre des décisions successives à partir des variables d'entrée. Le modèle segmente progressivement les données selon des critères de séparation tels que l'indice de Gini ou l'entropie afin de produire des groupes homogènes. Les arbres de décision sont faciles à interpréter et capables de modéliser des relations non linéaires. Néanmoins, ils peuvent être sensibles au bruit et au surapprentissage lorsque leur profondeur devient importante [56].

2.4 Forêt aléatoire

La forêt aléatoire est une méthode d'ensemble basée sur la construction de plusieurs arbres de décision entraînés sur des sous-ensembles aléatoires des données et des variables. La prédiction finale est obtenue par vote majoritaire dans le cas de la classification. Cette approche permet d'améliorer la robustesse du modèle et de réduire le risque de surapprentissage observé avec les arbres de décision individuels. Random Forest offre généralement de bonnes performances sur les données complexes et peut également fournir une estimation de l'importance des variables [56].

2.5 K plus proches voisins

L'algorithme des K plus proches voisins (K-Nearest Neighbors KNN) est une méthode de classification basée sur la similarité entre les observations. Le principe consiste à attribuer à une nouvelle donnée la classe majoritaire parmi ses k voisins les plus proches selon une mesure de distance, généralement la distance euclidienne. Le KNN est simple à mettre en œuvre et ne nécessite pas de phase d'apprentissage complexe. Cependant, ses performances peuvent diminuer lorsque le nombre de variables augmente ou lorsque les données contiennent du bruit [50].

2.6 Naive Bayes

Naive Bayes est un classificateur probabiliste fondé sur le théorème de Bayes et sur l'hypothèse d'indépendance conditionnelle entre les variables explicatives. Malgré cette hypothèse simplificatrice, cet algorithme est largement utilisé en classification grâce à sa rapidité d'exécution et à son faible coût computationnel. Naive Bayes peut fournir de bonnes performances même avec un nombre limité de données d'apprentissage. Toutefois, ses résultats peuvent être affectés lorsque les variables sont fortement corrélées [51].

2.7 XGBoost

XGBoost (Extreme Gradient Boosting) est un algorithme d'apprentissage supervisé basé sur la technique du boosting, qui consiste à combiner plusieurs arbres de décision afin d'améliorer progressivement les performances du modèle. Chaque nouvel arbre est construit pour corriger les erreurs des arbres précédents. XGBoost est reconnu pour sa rapidité, sa robustesse et sa capacité à gérer efficacement les relations complexes entre les variables. Cet algorithme est largement utilisé dans les applications de classification grâce à ses performances élevées et à sa bonne capacité de généralisation [57].

2.8 Gradient Boosting

Le Gradient Boosting est une méthode d'ensemble qui construit successivement plusieurs modèles faibles, généralement des arbres de décision, afin de minimiser les erreurs de

prédiction. Chaque nouveau modèle cherche à corriger les erreurs résiduelles produites par les modèles précédents à l'aide d'une optimisation basée sur le gradient. Cette approche permet d'obtenir des modèles performants capables de traiter des relations non linéaires complexes. Cependant, le Gradient Boosting peut nécessiter un temps d'entraînement important et un réglage précis des paramètres pour éviter le sur-apprentissage [56].

Ce tableau compare les principaux algorithmes de classification utilisés dans ce travail en termes de principe, avantages, inconvénients et domaines d'application.

Tableau 2.2 : Comparaison des algorithmes de classification utilisés

Algorithme	Principe	Avantages	Inconvénients	Utilisation
Régression logistique	Modèle probabiliste linéaire	Simple, interprétable	Limité aux relations linéaires	Classification binaire
SVM	Maximisation de la marge	Bon en haute dimension	Coût élevé sur grands datasets	Données complexes
KNN	Vote des voisins les plus proches	Facile à comprendre	Lent en prédiction	Petits datasets
Random Forest	Ensemble d'arbres de décision	Robuste, performant	Moins interprétable	Données tabulaires
XGBoost	Boosting d'arbres	Très haute	Paramétrage complexe	Compétitions ML

3. Approches de Deep Learning pour la classification

Le Deep Learning, ou apprentissage profond, constitue une branche avancée du Machine Learning fondée sur l'utilisation de réseaux de neurones artificiels comportant plusieurs couches cachées capables d'apprendre automatiquement des représentations complexes à partir des données. Contrairement aux méthodes classiques de machine learning qui nécessitent souvent une sélection manuelle des caractéristiques, le Deep Learning permet l'extraction automatique des caractéristiques pertinentes grâce à un processus d'apprentissage hiérarchique. Cette approche est particulièrement efficace pour le traitement de grandes quantités de données et pour la résolution de problèmes complexes tels que la reconnaissance d'images, le traitement du langage naturel, la détection d'objets et l'analyse des données médicales. Les performances élevées du Deep Learning sont principalement liées à sa capacité à modéliser des relations non linéaires complexes entre les variables et à apprendre des représentations de plus en plus abstraites à travers les différentes couches du réseau. Dans le domaine médical, ces techniques sont largement utilisées pour le diagnostic assisté par

ordinateur, l'analyse d'images médicales et la prédiction de maladies à partir de données cliniques [58].

Dans ce qui suit, nous allons présenter quelques approches de deep learning utilisées pour la classification.

3.1 Réseau de neurone profond (Deep Learning)

Un réseau de neurones profond, également appelé DNN (Deep Neural Network), se distingue par une architecture composée d'au moins deux couches. Cette structure lui confère la capacité de traiter des données complexes en s'appuyant sur des modèles mathématiques [73].

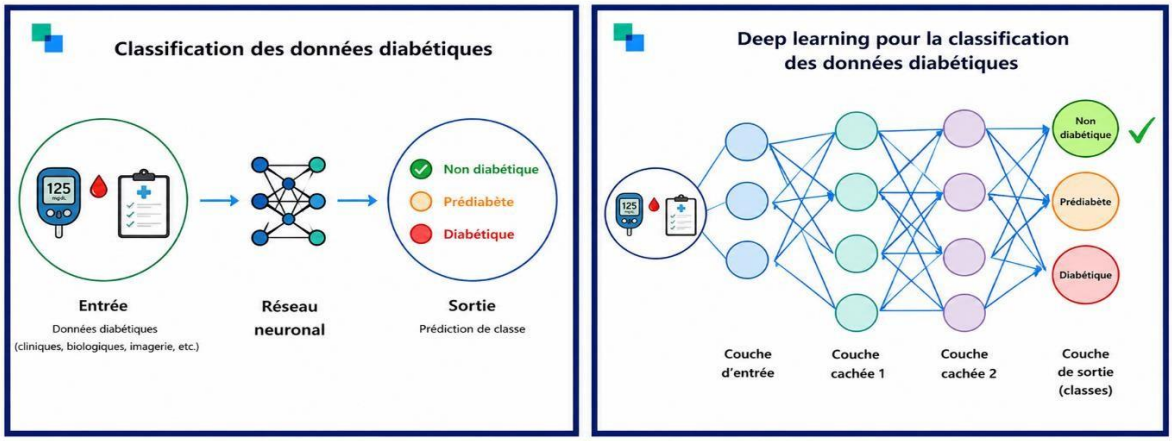


Figure 2.11 : Fonctionnement de réseau de neurone profond [89].

3.2 Réseaux de neurones convolutifs (CNN)

Les réseaux de neurones convolutifs (Convolutional Neural Networks CNN) représentent une architecture spécialisée du Deep Learning principalement conçue pour le traitement des données de type image. Les CNN reposent sur des opérations de convolution permettant d'extraire automatiquement des caractéristiques importantes telles que les contours, les textures et les formes présentes dans les images. Cette architecture est généralement composée de couches de convolution, de couches de pooling et de couches entièrement connectées qui travaillent ensemble afin d'améliorer progressivement la représentation des données. Les CNNs présentent l'avantage de réduire le nombre de paramètres à apprendre tout en conservant une grande capacité d'extraction des caractéristiques pertinentes. Grâce à leurs performances élevées, ils sont devenus l'une des méthodes les plus utilisées en vision par ordinateur et en imagerie médicale, notamment pour la détection de tumeurs, l'analyse radiologique et la classification de maladies. Cependant, l'entraînement de ces modèles nécessite souvent de grandes quantités de données ainsi qu'une importante puissance de calcul [58].

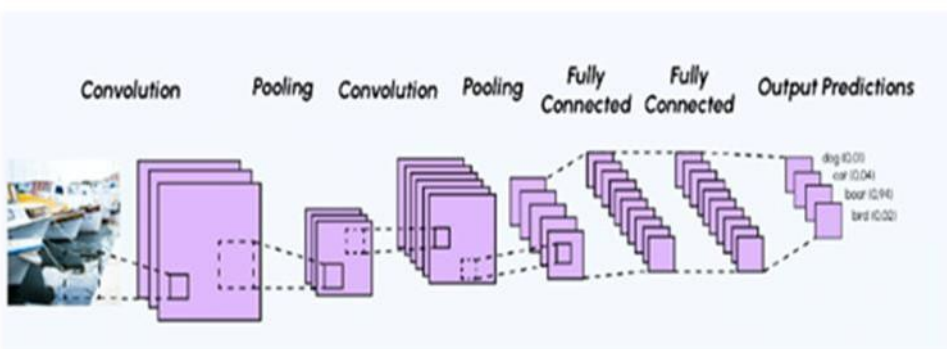


Figure 2.12 Les couche de CNN [90].

3.2.1 Couche Convolutionnelle

Cette couche consiste à appliquer des ltres (également appelés noyaux) sur lentrée (quil sagisse d'une image ou d'une carte de caractéristiques) a n didenti er des motifs locaux tels que des bords, des textures ou des formes. Chaque ltre génère une carte de caractéristiques (feature map) [59].

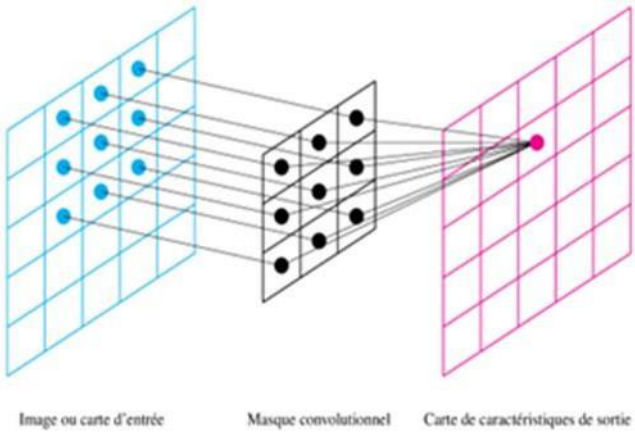


Figure 2.13 Une couche convolutionnelle en action [91]

3.2.2 Couche de Pooling (Pooling Layer)

Ces couches diminuent la dimension des données tout en préservant les informations clés. Le max pooling et laverage pooling, qui retiennent respectivement la valeur maximale ou moyenne d'un groupe de valeurs [60]. Son but est de réduire progressivement la taille spatiale de la représentation a n de diminuer le nombre de paramètres et de calculs dans le réseau et, par conséquent, de contrôler le sur apprentissage.

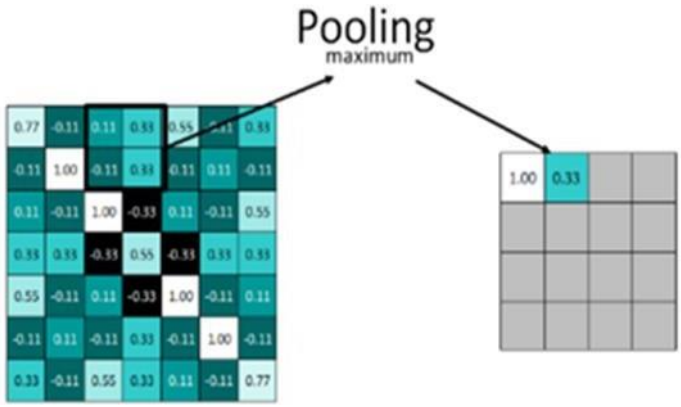


Figure 2.14 Exemple de fonctionnement de la couche pooling [92]

3.2.3 Couche entièrement connectée (Fully Connected Layer / Dense Layer)

Ces couches se trouvent généralement à la n du réseau et relient chaque neurone de la couche précédente à tous les neurones de la couche suivante (Figure 2.15). Elles utilisent les attributs extraits par les couches antérieures pour réaliser la classification définitive [61].

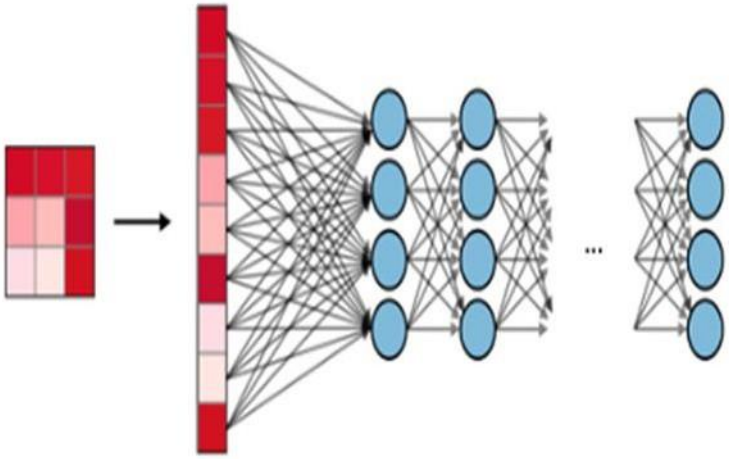


Figure 2.15 Exemple des Couches entièrement connectées [93]

3.2.4 Couche de sortie (Output Layer)

Elle Produit la sortie finale du modèle. Le choix de la fonction d'activation se fait selon la tâche :

- Softmax : Cette fonction est fréquemment utilisée en sortie des réseaux de neurones pour convertir les scores bruts en probabilités. Elle classe les données en assignant une probabilité à chaque catégorie, facilitant ainsi la prise de décision [62].
- Sigmoid : Utilisée pour les tâches de classification binaire. Elle transforme les valeurs de sortie en probabilités comprises entre 0 et 1, facilitant la distinction entre deux classes2 Rien ou linéaire pour les tâches de régression Linéaire : Utilisée pour les tâches

de régression. La fonction linéaire ne modifie pas les valeurs de sortie, permettant de prédire des valeurs continues [63].

4. Importance de l'intelligence artificielle dans les données médicales

L'Intelligence artificielle occupe aujourd'hui une place essentielle dans l'analyse et l'exploitation des données médicales en raison de sa capacité à traiter de grandes quantités d'informations de manière rapide et précise. Les systèmes basés sur l'intelligence artificielle permettent d'identifier des relations complexes entre les variables cliniques, biologiques et médicales afin d'améliorer la qualité du diagnostic, la prédiction des maladies et l'aide à la décision thérapeutique. Grâce aux techniques de machine learning et de deep learning, il devient possible d'analyser automatiquement des dossiers médicaux, des images radiologiques, des signaux biomédicaux et des données issues des objets connectés de santé. Ces technologies contribuent également à la médecine personnalisée en facilitant l'identification précoce des facteurs de risque et en permettant l'adaptation des traitements selon le profil du patient. Dans le domaine du diabète, l'intelligence artificielle joue un rôle important dans la prédiction du risque de maladie, le suivi des patients et l'amélioration des stratégies de prévention. Toutefois, l'utilisation de ces systèmes nécessite une validation rigoureuse des modèles ainsi qu'une attention particulière aux questions de fiabilité, de sécurité et d'éthique afin de garantir une utilisation sûre et efficace dans les environnements cliniques [64].

Ce tableau présente les principales applications de l'intelligence artificielle dans le domaine médical, notamment dans l'imagerie, le diagnostic, la prédiction des maladies et l'aide à la décision clinique.

Tableau 2.3 : Applications de l'intelligence artificielle dans le domaine médical

Domaine médical	Application de l'IA	Exemple
Imagerie médicale	Détection automatique d'anomalies	Détection de tumeurs sur IRM ou scanner
Cardiologie	Prédiction des maladies cardiovasculaires	Analyse des ECG
Diabétologie	Aide au diagnostic et prédiction du diabète	Modèles ML sur données cliniques
Oncologie	Détection et classification des cancers	Analyse d'images histopathologiques
Médecine générale	Aide à la décision clinique	Systèmes d'aide au diagnostic
Santé numérique	Suivi des patients	Applications mobiles et objets connectés

5. Travaux connexes : classification et prédiction du diabète

La prédiction précoce du diabète représente aujourd'hui un domaine de recherche important dans le contexte de l'intelligence artificielle appliquée à la santé. Grâce à la disponibilité croissante des données médicales numériques, plusieurs chercheurs ont proposé des modèles de Machine Learning et de Deep Learning afin d'améliorer le dépistage et le diagnostic précoce de cette maladie. Parmi les bases de données les plus utilisées figure la base **Early Stage Diabetes Risk Prediction**, qui contient des informations symptomatiques permettant d'identifier rapidement les patients à risque.

Dans ce contexte, **Laila et al. (2022)** ont développé une approche basée sur les techniques d'apprentissage par ensemble (Ensemble Learning) en utilisant les algorithmes, Bagging et Random Forest. Leur étude visait à comparer les performances de ces méthodes sur la base Early Stage Diabetes Risk Prediction. Les résultats obtenus ont montré que l'algorithme Random Forest présentait les meilleures performances avec une précision d'environ 97 %, ainsi que de bons résultats en termes de Recall et de F1-score. Les auteurs ont conclu que les méthodes d'ensemble permettent d'améliorer la robustesse et la précision des systèmes de prédiction du diabète.[65]

Par la suite, **Doğru et al. (2023)** ont proposé un modèle hybride de type Super Ensemble Learning destiné à la prédiction précoce du diabète. Leur approche combine plusieurs classifieurs afin de bénéficier des avantages de chaque algorithme et de réduire les erreurs de classification. Les résultats expérimentaux ont montré que cette combinaison permettait d'obtenir des performances supérieures à celles des modèles individuels, confirmant l'intérêt des approches hybrides dans les applications médicales.[66]

Dans une autre étude, **Elsayed et al. (2022)** ont utilisé la méthode Extreme Learning Machine (ELM) pour la prédiction précoce du diabète. Leur objectif était de développer un système rapide et efficace capable de détecter les patients à risque à partir des symptômes cliniques. Les résultats obtenus ont démontré que cette approche permettait de réduire le temps d'apprentissage tout en conservant des performances compétitives par rapport aux méthodes classiques de Machine Learning.[67]

Plus récemment, **Bülbül (2024)** a proposé un modèle hybride de Deep Learning associant un algorithme génétique (Genetic Algorithm), un Stacked Autoencoder et un classifieur Softmax. Cette approche vise à améliorer automatiquement la sélection des caractéristiques et l'apprentissage des représentations profondes des données. Les résultats obtenus sur la base Early Stage Diabetes Risk Prediction ont montré une précision de 98,72 %, dépassant plusieurs travaux précédents réalisés sur la même base de données. L'auteur souligne que l'intégration

des techniques d'optimisation et du Deep Learning permet d'améliorer considérablement les performances de classification.[68]

Au-delà des travaux réalisés spécifiquement sur la base Early Stage Diabetes Risk Prediction, plusieurs études ont été menées sur d'autres ensembles de données consacrés au diabète. **Akula et al. (2019)** ont développé un modèle d'apprentissage supervisé basé sur une combinaison de plusieurs algorithmes de classification afin d'améliorer la prédiction du diabète de type 2. Leur modèle d'ensemble a obtenu de meilleures performances que plusieurs classifieurs individuels, démontrant l'efficacité des approches collaboratives dans les systèmes de diagnostic médical.[69]

De même, **Dutta et al. (2022)** ont proposé une méthode de prédiction précoce du diabète reposant sur un ensemble de modèles de Machine Learning. Leur étude montre que l'utilisation combinée de plusieurs algorithmes permet d'améliorer la précision, la sensibilité et la capacité de généralisation du système de classification. Les auteurs soulignent également l'importance du prétraitement des données pour obtenir des résultats fiables.[70]

Par ailleurs, **Shrestha et al. (2022)** ont développé une approche hybride combinant Deep Learning, Support Vector Machine (SVM) et Long Short-Term Memory (LSTM) pour la prédiction du diabète de type 2. Les résultats obtenus montrent une amélioration de l'Accuracy et de l'AUC par rapport aux méthodes conventionnelles. Cette étude met en évidence l'intérêt croissant des architectures profondes dans les applications de santé numérique.[71]

Enfin, les recherches les plus récentes s'orientent vers l'intégration des techniques de sélection automatique des caractéristiques et des méthodes d'équilibrage des données. À ce titre, **Hayyolalam et Özkasap (2024)** ont proposé le modèle DiabML, qui combine une méthode de sélection de variables basée sur des métaheuristiques et la technique SMOTE pour traiter le déséquilibre des classes. Les auteurs montrent que l'intégration de ces approches améliore les performances globales du modèle et la détection des patients diabétiques.[72]

L'analyse de ces travaux montre que les algorithmes les plus utilisés pour la prédiction du diabète sont Random Forest, XGBoost, SVM, K-Nearest Neighbors ainsi que les réseaux de neurones profonds. Les études récentes démontrent également que les approches hybrides, les méthodes d'ensemble et les techniques d'équilibrage des données permettent généralement d'obtenir de meilleures performances que les modèles classiques. Cependant, plusieurs défis demeurent, notamment la généralisation des modèles à des populations réelles, l'interprétabilité des résultats et la gestion du déséquilibre des données médicales.

6. Conclusion

Dans ce chapitre, nous avons présenté l'importance de l'équilibrage des données dans les problèmes de classification, en mettant en évidence son rôle essentiel dans l'amélioration des performances des modèles d'intelligence artificielle, notamment lorsqu'ils sont confrontés à des jeux de données déséquilibrés. Nous avons également décrit les principales méthodes d'équilibrage existantes, qu'elles soient basées sur le suréchantillonnage, le sous-échantillonnage ou des approches hybrides.

Par ailleurs, nous avons introduit les concepts fondamentaux de l'intelligence artificielle en nous intéressant plus particulièrement aux approches de Machine Learning et de Deep Learning. Les différentes techniques et architectures les plus utilisées dans le domaine de la classification ont été présentées afin de fournir une base théorique solide pour la suite de ce travail.

Le chapitre suivant sera consacré à la présentation de nos contributions dans le cadre de ce projet de fin d'études. Nous y détaillerons la méthodologie adoptée pour la classification du diabète en utilisant plusieurs modèles de Machine Learning et de Deep Learning. Une étude comparative sera réalisée en considérant à la fois les données originales et les données équilibrées, afin d'évaluer l'impact des techniques d'équilibrage sur les performances des modèles proposés.

Chapitre 3. Résultats et Discussion

1. Contexte de l'étude

Le diabète représente l'une des maladies chroniques les plus répandues à l'échelle mondiale, touchant plus de 500 millions de personnes selon l'Organisation Mondiale de la Santé. Cette pathologie métabolique se caractérise par une hyperglycémie chronique résultant soit d'un déficit de production d'insuline (diabète de type 1), soit d'une résistance cellulaire à l'action de l'insuline (diabète de type 2). Le diagnostic précoce du diabète est crucial pour prévenir les complications graves telles que la rétinopathie, la néphropathie, les maladies cardiovasculaires et les neuropathies.

Dans ce contexte, l'intégration de l'intelligence artificielle, et plus particulièrement des méthodes d'apprentissage automatique (machine learning) et d'apprentissage profond (deep learning), offre de nouvelles perspectives pour améliorer l'aide au diagnostic médical. Ces approches permettent d'exploiter efficacement les données cliniques et biologiques afin de construire des modèles prédictifs capables d'identifier les patients à risque avec une précision élevée.

Ce chapitre présente une approche expérimentale axée sur l'utilisation d'algorithmes de Machine Learning (ML) et de Deep Learning (DL) pour la classification d'une base de données de diabète (Early Stage Diabetes Risk Prediction.csv). En s'appuyant sur un jeu de données médicales, nous avons évalué la performance de plusieurs modèles prédictifs appliqués à une tâche de classification binaire (présence ou absence de diabète). Ce chapitre détaille l'environnement de développement, les étapes d'implémentation, les résultats obtenus, et propose une discussion critique en les comparant aux travaux existants dans la littérature.

Une problématique majeure abordée dans ce travail concerne le déséquilibre des classes inhérent à de nombreuses bases de données médicales. Dans notre base de données, la classe positive (diabétiques) représente 61,5% des échantillons contre 38,5% pour la classe négative (non-diabétiques). Ce déséquilibre, bien que modéré, peut biaiser l'apprentissage des modèles en favorisant la classe majoritaire au détriment de la classe minoritaire. Pour pallier cette problématique, nous avons mis en œuvre six techniques d'équilibrage des données : SMOTE, ADASYN, RandomOverSampler, RandomUnderSampler, TomekLinks et et nous avons évalué leur impact sur les performances de dix classifieurs traditionnels ainsi que sur un modèle de réseau de neurones convolutif (CNN 1D).

2. Environnement de développement

Afin de mettre en œuvre les modèles d'apprentissage automatique et d'apprentissage profond pour la classification de la base de données de diabète, nous avons choisi le langage de programmation Python comme environnement de développement principal. Ce langage, largement utilisé en science des données, nous a permis d'exploiter une variété d'outils et de bibliothèques performants qui ont grandement facilité l'implémentation, le traitement des données et l'entraînement des modèles. Il comprend plusieurs bibliothèques standards, dont :

- **TensorFlow** : une bibliothèque open-source, développée par Google en 2015. Elle est conçue pour le calcul intensif et excelle dans la conception de modèles d'apprentissage automatique et d'apprentissage profond. TensorFlow s'appuie sur une structure reposant sur des graphes de flux de données, dans laquelle les opérations mathématiques sont symbolisées par des nœuds et les matrices de données sont connues sous le nom de tenseurs.
- **NumPy (Numerical Python)** : est une bibliothèque open-source offrant des structures de données robustes, notamment les ndarrays (tableaux multidimensionnels), ainsi que des fonctions mathématiques pour réaliser efficacement des opérations sur ces tableaux. Elle est principalement employée pour des calculs scientifiques, le traitement de données numériques et constitue la fondation de nombreuses autres bibliothèques scientifiques Python.
- **Pandas** : est une bibliothèque Python open-source destinée à la manipulation et à l'analyse de données. Elle propose deux structures principales : le DataFrame (tableau de données à deux dimensions, lignes et colonnes, comme un tableau Excel) et la Series (tableau unidimensionnel). Pandas facilite des opérations comme la lecture de fichiers (CSV, Excel), le nettoyage des données, la transformation et l'analyse statistique.
- **Keras** : est une bibliothèque Python libre qui facilite la construction et l'entraînement de modèles d'apprentissage profond (deep learning) de façon aisée et rapide. Il s'agit d'une interface utilisateur avancée qui utilise des moteurs de calcul tels que TensorFlow pour réaliser les opérations intensives. Keras rend la construction de réseaux de neurones aisée grâce à une syntaxe intuitive, convenant aussi bien aux novices qu'aux chercheurs expérimentés.
- **Scikit-learn** : est une bibliothèque Python dédiée à l'apprentissage automatique. Elle propose un large éventail d'algorithmes de classification, de régression, de clustering et de réduction de dimensionnalité, ainsi que des outils d'évaluation de modèles et de prétraitement des données.

- **Imbalanced-learn** : est une bibliothèque Python spécialement conçue pour traiter les problèmes de déséquilibre de classes dans les jeux de données. Elle offre diverses techniques de rééchantillonnage (sur-échantillonnage, sous-échantillonnage et combinaisons) compatibles avec l'API de Scikit-learn.
- **XGBoost** : est une bibliothèque optimisée de gradient boosting, reconnue pour sa rapidité et ses performances dans les compétitions de data science. Elle implémente des algorithmes d'arbres de décision boostés avec régularisation.
- **Matplotlib et Seaborn** : sont des bibliothèques de visualisation de données en Python. Matplotlib permet de créer des graphiques statiques, tandis que Seaborn, construit sur Matplotlib, offre une interface de haut niveau pour des visualisations statistiques attrayantes.

Outils d'implémentations : Dans le cadre de ce projet, nous avons utilisé les outils suivants:

- **Google Colab** : est une plateforme en ligne gratuite proposée par Google qui permet d'écrire, d'exécuter et de partager du code Python dans un environnement cloud. Il est principalement utilisé pour le machine learning, l'apprentissage profond (deep learning), l'analyse de données et le prototypage rapide. Colab est basé sur la technologie des Jupyter Notebooks, permettant de mélanger code, texte, images, formules mathématiques (LaTeX) et graphiques dans un seul document interactif.

3. Implémentation des méthodes

Afin d'évaluer l'impact des techniques d'équilibrage des données sur la performance des modèles de classification du diabète, nous avons adopté la méthodologie illustrée par la figure suivante. Dans un premier temps, la base de données composée de 520 échantillons et de 16 caractéristiques est soumise à une phase de prétraitement comprenant notamment l'encodage des variables et les opérations nécessaires à la préparation des données. Ensuite, les données sont divisées en deux ensembles : un ensemble d'apprentissage (80 %) destiné à l'entraînement des modèles et un ensemble de test (20 %) réservé à leur évaluation.

Deux scénarios expérimentaux sont alors considérés. Le premier consiste à appliquer directement les algorithmes de Machine Learning et de Deep Learning sur les données initiales sans équilibrage. Le second intègre une étape d'équilibrage des données d'apprentissage avant la phase de classification. Les performances obtenues dans les deux cas sont ensuite comparées à l'aide de plusieurs métriques d'évaluation, notamment l'accuracy, la précision, le rappel (recall) et le score F1. Cette démarche permet d'analyser l'influence des méthodes d'équilibrage sur la qualité de la prédiction du diabète et d'identifier les modèles les plus performants.

La figure suivante illustre le processus de traitement mis en œuvre pour la classification de la base de données utilisée dans ce travail.

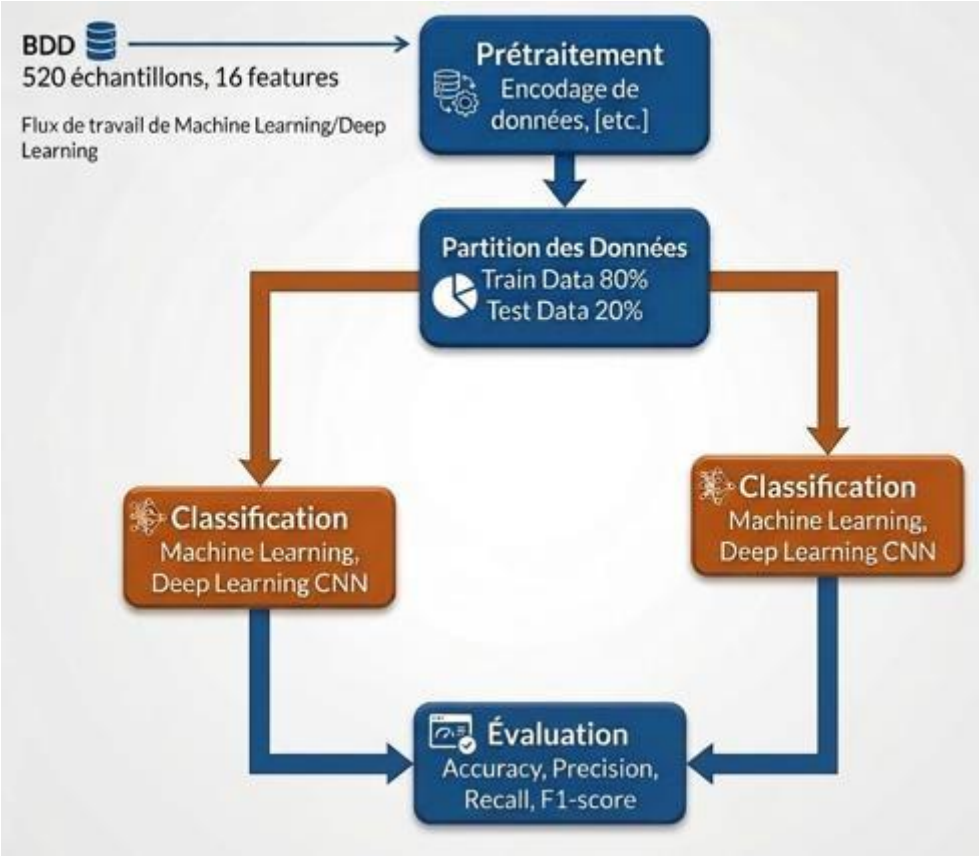


Figure 3.1 : Schéma général de la méthodologie proposée pour la classification du diabète avec et sans équilibrage des données.

3.1 Description de la base de données

Dans le cadre de ce projet, nous avons utilisé un jeu de données médicales issu de la plateforme UCI intitulé « **Early Stage Diabetes Risk Prediction**.csv » [74]

Cette base a été collectée en 2020 auprès de 520 patients de l'hôpital du diabète de Sylhet au Bangladesh, avec 16 features symptomatiques binaires et une variable cible binaire déséquilibrée (61,54% Positive, 38,46% Negative) Cette base de données est spécifiquement conçue pour la classification du diabète. Elle regroupe des informations variées concernant les symptômes et caractéristiques cliniques de patients.

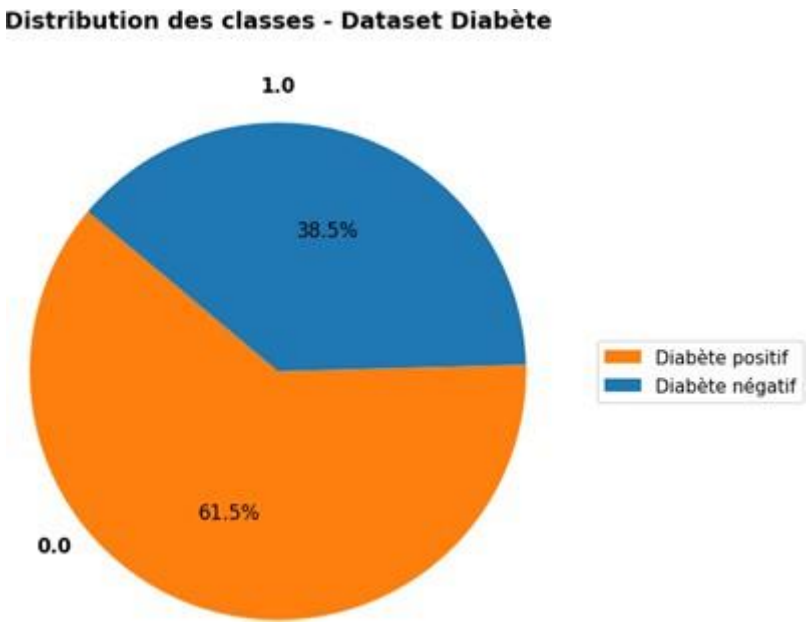


Figure 3.2 Répartition des classes

Les caractéristiques incluent :

- **Age** : âge du patient (variable numérique)
- **Gender** : genre du patient (encodé numériquement, valeurs 3 et 4)
- **Polyuria**: présence d'une polyurie (0/1)
- **Polydipsia**: présence d'une polydipsie (0/1)
- **sudden weight loss**: perte de poids soudaine (0/1)
- **weakness**: faiblesse (0/1)
- **Polyphagia**: polyphagie (0/1)
- **Genital thrush**: muguet génital (0/1)
- **visual blurring**: trouble visuel (0/1)
- **Itching**: prurit (0/1)
- **Irritability**: irritabilité (0/1)
- **delayed healing**: cicatrisation retardée (0/1)
- **partial paresis**: parésie partielle (0/1)
- **muscle stiffness**: raideur musculaire (0/1)
- **Alopecia**: alopécie (0/1)
- **Obesity**: obésité (0/1)

La variable cible **class** est binaire :

- Classe 1: Patient diabétique (320 échantillons, 61,54%)
- Classe 0 : Patient non diabétique (200 échantillons, 38,46%)

3.2 Prétraitement de la base de données

Le prétraitement de la base de données débute par une analyse exploratoire permettant de mieux comprendre sa structure et ses caractéristiques. Cette analyse met en évidence un ensemble de données composé de 520 instances décrites par 16 variables explicatives ainsi qu'une variable cible binaire. Elle révèle également un déséquilibre entre les classes, avec 61,5 % d'exemples appartenant à la classe positive contre 38,5 % à la classe négative. Par ailleurs, aucune valeur manquante n'a été détectée, ce qui facilite les étapes de préparation des données.

L'étape suivante consiste à transformer les variables catégorielles en valeurs numériques afin de les rendre exploitables par les algorithmes de Machine Learning et de Deep Learning. Ainsi, les quinze variables symptomatiques exprimées sous la forme « Yes/No » sont encodées respectivement par les valeurs 1 et 0. La variable *Gender* est convertie en valeurs numériques, avec 3 pour *Masculin* et 4 pour Féminin, tandis que la variable cible *class* est représentée par 1 pour la classe *Positive* et 0 pour la classe *Negative*. La variable *Age*, déjà de nature numérique, est conservée sans modification.

Le tableau suivant présente en détail le schéma d'encodage appliqué aux différentes variables de la base de données.

Table 3 .1 Encodage des variables catégorielles

Étape	Description
Analyse exploratoire	Vérification des 520 échantillons, 16 features, types de données, valeurs manquantes et distribution des classes (61,5% Positive / 38,5% Negative)
Encodage des variables catégorielles	Conversion des 15 symptômes binaires (Yes/No → 1/0), Gender (Male/Female → 1/0) et de la cible class (Positive/Negative → 1/0). La variable Age reste numérique

3.3 Classification de diabète : approche sans équilibrage de données

La phase de classification s'appuie sur les données numériques obtenues à l'issue du prétraitement, afin de permettre l'apprentissage d'un modèle capable de distinguer les cas diabétiques des cas non pathologiques.

Algorithm

Étape 1 : Importation des bibliothèques

– Importation des bibliothèques nécessaires à la manipulation des données, à la modélisation et à l'évaluation des performances

Étape 2: Chargement des données

- Chargement d'un fichier CSV contenant des données médicales relatives au diabète
- Ce fichier comprend généralement des colonnes représentant des caractéristiques cliniques, ainsi qu'une colonne cible 'classe'

Étape 03 : Séparation des caractéristiques et de la cible

- X : ensemble des variables explicatives (toutes les colonnes sauf la cible).
- y : variable cible, c'est-à-dire ce que l'on souhaite prédire.

Étape 04 : Division du jeu de données

- 80 % des données sont utilisées pour l'entraînement (X_train, y_train).
- 20 % sont réservées aux tests (X_test, y_test).

Étape 05 : Initialisation et entraînement du modèle

- Création du modèle avec ses paramètres spécifiques
- Entraînement du modèle à partir des données d'apprentissage.

Étape 06: Prédiction

- Prédiction de la variable cible à partir des données de test.

Étape 07: Évaluation des performances

- *Accuracy — score* : mesure la précision globale du modèle.
- *F*1 — *score* : équilibre entre la précision et le rappel.
- *Recall — score* : capacité du modèle à détecter toutes les occurrences positives.
- *Precision — score* : proportion de vraies positives parmi les prédictions positives.

Étape 08: Affichage des résultats

- Affichage des différentes métriques d'évaluation pour analyser les performances du modèle.

3.3.1 Application des modèles classiques

Nous avons appliqué des modèles d'apprentissage automatique supervisé tels que la régression logistique, les arbres de décisions, SVM, la forêt aléatoire, KNN, Naive bayes, Gradient Boosting, Extra Trees et XGBoost sur la variable cible, traitée comme une classe binaire. Voici le tableau présentant chacun des modèles classiques utilisés, ainsi que les paramètres associés pour chaque modèle :

Table3.2 Paramètres de création des modèles traditionnels

Modèle	Paramètres de création
Régression Logistique	random_state=42, max_iter=1000
Arbre de Décision	random_state=42
Random Forest	n_estimators=100, random_state=42
SVM (RBF)	kernel='rbf', random_state=42, probability=True
KNN (k=5)	n_neighbors=5
Naive Bayes	Aucun paramètre précisé (par défaut)
Gradient Boosting	random_state=42
Extra Trees	n_estimators=100, random_state=42
XGBoost	random_state=42, eval_metric='logloss'

3.3.2 Application de l'apprentissage en profondeur

Nous avons appliqué un modèle d'apprentissage profond, à savoir le CNN, L'architecture

Le réseau de neurones convolutif unidimensionnel (1D-CNN) proposé pour la classification du diabète adopte une architecture hiérarchique soigneusement structurée, conçue pour exploiter efficacement les données tabulaires de la base BDD.csv. La couche d'entrée accueille les 520 échantillons standardisés, chacun décrit par 17 caractéristiques médicales, établissant ainsi une dimension initiale de 520 × 17. Le traitement débute par un bloc de convolution et de pooling, où une première couche Conv1D déploie 64 filtres pour capturer les caractéristiques locales élémentaires, suivie d'une seconde couche Conv1D à 128 filtres qui approfondit l'extraction des descripteurs pertinents ; une opération de MaxPooling1D, avec une fenêtre de taille 2, intervient ensuite pour réduire la dimensionnalité spatiale des cartes de caractéristiques tout en préservant l'information essentielle. La transition vers le bloc entièrement connecté est assurée par une couche Flatten qui aplatit les représentations multidimensionnelles en un vecteur unidimensionnel dense. Ce bloc fully-connected se compose successivement d'une première couche Dense de 128 neurones activée par ReLU,

puis d'une seconde couche Dense de 64 neurones également dotée de ReLU, permettant d'affiner la représentation apprise avant la décision finale. La couche de sortie, unique et munie d'une activation Sigmoid, génère une probabilité binaire permettant de classifier chaque patient en deux classes distinctes : 0 (non diabétique) ou 1 (diabétique), offrant ainsi une décision claire et interprétable pour le diagnostic assisté par intelligence artificielle.

La figure suivante présente l'architecture de notre CNN :

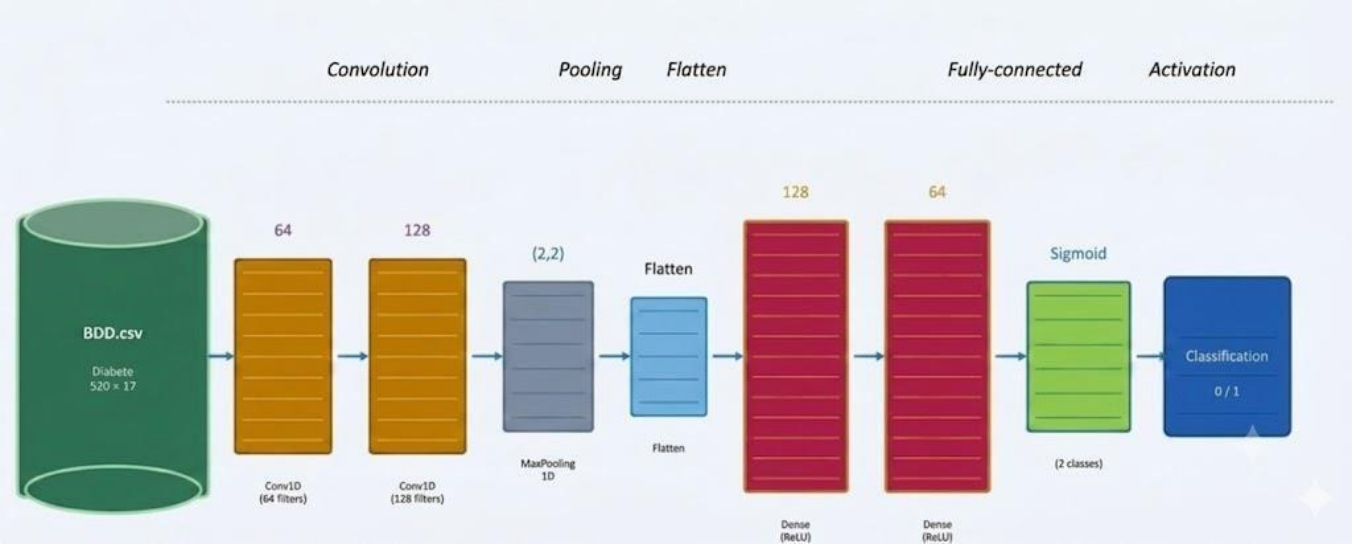


Figure 3.3 Architecture du CNN

3.4 Classification de diabète : approche avec équilibrage de données

Dans cette section, nous présentons les différentes méthodes d'équilibrage des données retenues dans le cadre de notre étude ainsi que les paramètres associés à chacune d'elles. Ces techniques sont appliquées uniquement sur l'ensemble d'apprentissage afin de réduire l'effet du déséquilibre entre les classes et d'améliorer les capacités de généralisation des modèles. Il est important de souligner que les paramètres des algorithmes de Machine Learning et de Deep Learning utilisés pour la classification restent identiques à ceux adoptés dans les expériences réalisées sans équilibrage. Ainsi, toute différence observée au niveau des performances pourra être attribuée à l'impact des méthodes d'équilibrage appliquées.

Afin d'étudier l'impact de l'équilibrage des données sur la performance des modèles de classification du diabète, les mêmes algorithmes de Machine Learning et de Deep Learning présentés dans la première contribution ont été réutilisés. Pour garantir une comparaison équitable, les paramètres de chaque modèle ont été conservés inchangés. L'équilibrage a été

appliqué exclusivement sur l'ensemble d'apprentissage après la phase de partitionnement des données, tandis que l'ensemble de test a été maintenu dans sa distribution originale afin de fournir une évaluation représentative des conditions réelles.

Pour chaque méthode d'équilibrage, un nouvel ensemble d'apprentissage a été généré puis utilisé pour entraîner les différents modèles de classification. Les performances obtenues ont ensuite été évaluées sur le même ensemble de test à l'aide des métriques Accuracy, Precision, Recall et F1-score. Les techniques d'équilibrage retenues dans cette étude comprennent :

- SMOTE, qui génère des exemples synthétiques de la classe minoritaire afin d'obtenir une répartition équilibrée des classes ;
- ADASYN, qui produit des échantillons synthétiques de manière adaptative en se concentrant davantage sur les instances difficiles à apprendre ;
- RandomOverSampler, qui augmente la taille de la classe minoritaire par duplication aléatoire des observations existantes ;
- RandomUnderSampler, qui réduit la taille de la classe majoritaire par suppression aléatoire d'instances ; Tomek Links, qui améliore la séparation entre les classes en éliminant certaines observations situées à la frontière de décision ;

Cette démarche permet d'évaluer de manière rigoureuse l'influence de chaque méthode d'équilibrage sur les performances des modèles de classification du diabète.

Le tableau suivant présente les paramètres utilisés pour les différentes méthodes d'équilibrage appliquées dans cette étude. Pour les méthodes SMOTE, ADASYN, RandomOverSampler, RandomUnderSampler et, le paramètre **random_state = 42** a été fixé afin de garantir la reproductibilité des expériences. En effet, ce paramètre contrôle la génération aléatoire des échantillons synthétiques ou la sélection des instances lors des opérations de suréchantillonnage et de sous-échantillonnage, permettant ainsi d'obtenir les mêmes résultats à chaque exécution. La méthode **Tomek Links** ne nécessite aucun paramètre spécifique dans notre implémentation, car elle se base directement sur l'identification et la suppression des paires d'instances ambiguës situées à la frontière entre les classes.

Tableau 3.3 : Paramétrage des méthodes d'équilibrage retenues dans l'étude

Méthode	***Paramètres***
SMOTE	random_state=42
RandomOverSampler	random_state=42
ADASYN	random_state=42
RandomUnderSampler	random_state=42
TomekLinks	(aucun paramètre)

4. Résultats et discussion

Dans cette section, nous présentons et analysons les résultats obtenus à l'issue de l'entraînement et de l'évaluation de nos différents modèles de classification. Afin de répondre aux objectifs de nos deux contributions, nous avons évalué des approches issues de l'apprentissage automatique (*Machine Learning*) et de l'apprentissage profond (*Deep Learning*). Cette étude comparative est menée sous deux scénarios distincts : d'une part, une classification directe sur les données initiales et, d'autre part, une classification après application d'algorithmes d'équilibrage des données. Les performances de chaque configuration sont mesurées et comparées à l'aide des métriques standards (Accuracy, Precision, Recall, et F1-score). L'objectif principal est d'identifier le modèle le plus performant et robuste pour la détection du diabète, tout en mettant en lumière l'impact réel de l'équilibrage sur le comportement, les points forts et les limites de chaque algorithme.

4.1 Critères d'évaluation

Voici les définitions des critères d'évaluation que nous avons utilisés dans notre projet pour la classification de diabète :

a) *Taux de classification* : Le taux de classification est la proportion de prédictions correctes par rapport au nombre total de prédictions effectuées. Il est calculé comme suit :

$$\text{Taux de classification} = \frac{\text{Nombre de prédictions correctes}}{\text{Nombre total de prédictions}}$$

Un taux de classification élevé indique que le modèle effectue beaucoup de bonnes prédictions, mais il peut être trompeur si les classes sont déséquilibrées.

b) *Précision* : La précision mesure la proportion de vraies positives parmi toutes les prédictions positives. Elle est calculée comme suit :

$$\text{Précision} = \frac{TP}{TP + FP}$$

Une précision élevée indique que peu de fausses alertes sont émises par le modèle. C'est particulièrement important dans les contextes où les faux positifs ont des conséquences significatives.

c) *Rappel* : Le rappel, également appelé sensibilité ou taux de vrais positifs, mesure la proportion de vraies positives parmi toutes les instances réellement positives. Il est défini comme :

$$\text{Rappel} = \frac{TP}{TP + FN}$$

Un rappel élevé indique que le modèle identifie correctement la plupart des cas positifs. Cela est crucial dans des contextes où il est essentiel de détecter autant de cas positifs que possible.

d) *F1- Score* : Le F1 score est la moyenne harmonique entre la précision et le rappel, fournissant un équilibre entre les deux. Il est particulièrement utile lorsque les classes sont déséquilibrées. Il se calcule ainsi :

$$F1 = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

Un F1 score élevé indique que le modèle a à la fois une bonne précision et un bon rappel. C'est un bon indicateur de la performance globale, notamment lorsqu'il faut équilibrer faux positifs et faux négatifs.

e) *G-mean (Geometric Mean)* : Le G-mean mesure la performance équilibrée du modèle en tenant compte à la fois de la sensibilité (Recall) et de la spécificité. Il est calculé comme suit :

$$G\text{-mean} = \sqrt{\text{Sensibilité} \times \text{Spécificité}} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$$

Un G-mean élevé indique que le modèle performe bien sur les deux classes simultanément, sans favoriser une classe au détriment de l'autre. C'est particulièrement important dans les

contextes où les classes sont déséquilibrées et où négliger la classe minoritaire aurait des conséquences significatives, car il pénalise fortement les modèles qui excellent sur une classe mais échouent sur l'autre.

f) *Matrice de confusion* : La matrice de confusion est un outil d'évaluation utilisé en apprentissage supervisé, notamment pour les problèmes de classification. Elle permet de visualiser les performances d'un modèle en comparant les prédictions aux valeurs réelles (ou vraies étiquettes) des données de test.

g) *Courbe ROC* : La courbe ROC est un outil d'évaluation des performances d'un modèle de classification binaire. Elle trace le taux de vrais positifs (TPR) en fonction du taux de faux positifs (TFP) pour différents seuils de classification. Pour chaque seuil de probabilité (entre 0 et 1), on calcule:

$$TPR = Sensibilit = \frac{TP}{TP + FN}, \quad TFP = \frac{FP}{FP + TN}$$

En faisant varier ce seuil, on obtient une courbe ROC. Plus la courbe est proche du coin supérieur gauche, meilleur est le modèle

4.2 Classification de diabète : approche sans équilibrage de données

4.2.1 Application des modèles classiques

Nous avons appliqué dix modèles d'apprentissage automatique supervisé sur la variable cible 'class', traitée comme une classe binaire. Et voici les meilleurs résultats:

Tableau 3.4 : Résultats expérimentaux des modèles de ML

Modèle	**Taux de classification (%)**	**Précision (%)**	**Rappel (%)**	**F1-score (%)**
Arbre de Décision	96.15	96.57	96.15	96.21
SVM (RBF)	92.30	92.84	94.15	93.45
Naive Bayes	91.34	90.28	89.61	89.93
Extra Trees	**98.07**	**97.14**	**98.59**	**97.82**
LogisticRegression	92.30	91.74	90.31	90.97

Extra Trees se distingue comme le modèle le plus performant pour la classification du diabète, avec un taux de classification de 98.07% et un F1-score de 97.82%. Cette supériorité s'explique par sa nature d'ensemble learning (forêt d'arbres décorrélés), qui réduit le surapprentissage inhérent à un arbre de décision unique tout en capturant les interactions complexes entre les variables cliniques.

L'Arbre de Décision offre un très bon rapport simplicité-performance, avec un F1-score de

96.21%. En revanche, les modèles plus simples comme le SVM, la Régression Logistique et surtout le Naive Bayes obtiennent des résultats nettement moins bons. Le Naive Bayes est le plus décevant : avec un rappel de seulement 89.61%

4.2.2 Application de l'apprentissage en profondeur

Nous avons appliqué un modèle d'apprentissage profond, à savoir le CNN 1D, à la classification binaire basée sur la colonne 'class'. Dans notre projet, nous avons implémenté un réseau de neurones convolutif (CNN) séquentiel pour effectuer une classification binaire à partir de données numériques normalisées. Le modèle a été construit avec l'API Sequential de Keras, voilà le résultat obtenu :

Modèle	**Taux de classification (%)**	**Précision (%)**	**Rappel (%)**	**F1-score (%)**
CNN	96.15	95	96	96

TABLEAU 3 .5 : Résultats expérimentaux des modèles de CNN

Le modèle **CNN** (Réseau de Neurones Convolutifs) atteint un **taux de classification de 96.15%**, ce qui est un résultat **très satisfaisant**. La **précision (95%)** et le **rappel (96%)** sont équilibrés, ce qui se traduit par un bon **F1-score de 96%**.

4.3 Classification de diabète : approche avec équilibrage de données

Cette section met en lumière notre deuxième contribution à travers le développement d'approches de classification avec équilibrage des données. Nous y les résultats obtenus pour chacune des six méthodes d'équilibrage appliquées aux dix classifieurs traditionnels et au modèle CNN 1D.

4.3.1 Application des modèles classiques avec équilibrage

Nous avons mis en œuvre les mêmes modèles d'apprentissage supervisé utilisés dans la première contribution, en appliquant systématiquement chacune des six méthodes d'équilibrage avant l'entraînement. Pour chaque combinaison modèle × méthode d'équilibrage, nous avons évalué les performances sur l'ensemble de test original (non équilibré) afin de garantir une évaluation réaliste.

Le tableau suivant présente un résumé comparatif des performances des différents modèles de classification en fonction des techniques d’équilibrage des données utilisées.

Modèle	Méthode d'équilibrage	Taux de classification (%)	Précision (%)	Rappel (%)	F1-score (%)	G-mean
Régression Logistique	*SMOTE*	93.27	98.31	90.62	94.31	0.94
	ADASYN	94.23	100.0	90.62	95.08	0.95
	RandomOverSampler	93.27	98.31	90.62	94.31	0.94
	RandomUnderSampler	92.31	98.28	89.06	93.44	0.93
	TomekLinks	94.23	98.33	92.19	95.16	0.95
Arbre de Décision	*SMOTE*	97.12	100.0	95.31	97.6	0.98
	ADASYN	98.08	100.0	96.88	98.41	0.98
	RandomOverSampler	98.08	100.0	96.88	98.41	0.98
	RandomUnderSampler	94.23	96.77	93.75	95.24	0.94
	TomekLinks	97.12	100.0	95.31	97.6	0.98
Random Forest	*SMOTE*	98.08	100.0	96.88	98.41	0.98
	ADASYN	99.04	100.0	98.44	99.21	0.99
	RandomOverSampler	98.08	100.0	96.88	98.41	0.98
	RandomUnderSampler	97.12	100.0	95.31	97.6	0.98
	TomekLinks	98.08	100.0	96.88	98.41	0.98
SVM (RBF)	*SMOTE*	98.08	98.44	98.44	98.44	0.98
	ADASYN	99.04	100.0	98.44	99.21	0.99
	RandomOverSampler	98.08	98.44	98.44	98.44	0.98
	RandomUnderSampler	96.15	98.39	95.31	96.83	0.96
	TomekLinks	98.08	98.44	98.44	98.44	0.98
KNN (k=5)	*SMOTE*	92.31	98.28	89.06	93.44	0.93
	ADASYN	91.35	98.25	87.5	92.56	0.92
	RandomOverSampler	92.31	98.28	89.06	93.44	0.93
	RandomUnderSampler	89.42	98.18	84.38	90.76	0.91
	TomekLinks	93.27	98.31	90.62	94.31	0.94
Naive Bayes	*SMOTE*	95.19	96.83	95.31	96.06	0.95
	ADASYN	95.19	96.83	95.31	96.06	0.95
	RandomOverSampler	92.31	96.67	90.62	93.55	0.93
	RandomUnderSampler	92.31	96.67	90.62	93.55	0.93
	TomekLinks	94.23	96.77	93.75	95.24	0.94

Gradient Boosting	*SMOTE*	98.08	100.0	96.88	98.41	0.98
	ADASYN	99.04	100.0	98.44	99.21	0.99
	RandomOverSampler	98.08	100.0	96.88	98.41	0.98
	RandomUnderSampler	96.15	100.0	93.75	96.77	0.97
	TomekLinks	99.04	100.0	98.44	99.21	0.99
Extra Trees	***SMOTE***	***100.0***	***100.0***	***100.0***	***100.0***	***1.0***
	ADASYN	97.12	98.41	96.88	97.64	0.97
	RandomOverSampler	99.04	100.0	98.44	99.21	0.99
	RandomUnderSampler	98.08	100.0	96.88	98.41	0.98
	TomekLinks	99.04	100.0	98.44	99.21	0.99
XGBoost	*SMOTE*	98.08	100.0	96.88	98.41	0.98
	ADASYN	98.08	100.0	96.88	98.41	0.98
	RandomOverSampler	98.08	100.0	96.88	98.41	0.98
	RandomUnderSampler	95.19	100.0	92.19	95.93	0.96
	TomekLinks	98.08	100.0	96.88	98.41	0.98

Tableau 3.6: récapitulatif des performances des modèles ML avec différentes méthodes d'équilibrage.

Extra Trees avec SMOTE est le meilleur modèle de tous, avec un score parfait de 100% il ne se trompe jamais. Random Forest, SVM et Gradient Boosting suivent de près avec 99.21% grâce à ADASYN, ce qui signifie qu'ils se trompent à peine.

Arbre de Décision et XGBoost sont aussi bons avec 98.41%, mais un peu moins sûrs. En revanche, , Naive Bayes et Régression Logistique restent moyens autour de 95-96% . KNN est le plus faible avec seulement 94.31%, Pour les méthodes d'équilibrage, ADASYN donne les meilleurs résultats, tandis que RandomUnderSampler est la pire. Donc, les modèles d'ensemble comme Extra Trees et Random Forest tirent le meilleur parti de l'équilibrage, et SMOTE ou ADASYN sont les méthodes à privilégier pour un diagnostic médical fiable.

4.3.2 Application du CNN 1D avec équilibrage

Le modèle CNN 1D a également été entraîné avec chacune des six méthodes d'équilibrage. Les données équilibrées ont été reshapeées en format (samples, features, 1) pour être compatibles avec l'architecture convolutive 1D. L'encodage one-hot a été appliqué sur les labels pour la classification binaire.

TABLEAU 3.7 récapitulatif des performances de modèle CNN avec différentes méthodes d'équilibrage.

Méthode d'équilibrage	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
ADASYN	80.77	100	68.75	81 .48
RandomOverSampler	95.19	100	92.19	95.93
RandomUnderSampler	94.19	99	91.19	94 .92
SMOTE	96.15	98.39	95.31	96.83
TomekLinks	69.23	97.06	51 .56	67.35

SMOTE est la meilleure méthode pour le CNN 1D, avec 96.15% de bonnes classifications et un F1-score de 96.83tout en gardant une bonne précision. RandomOverSampler suit avec 95.19% et un F1 de 95.93%, ce qui est aussi très solide. RandomUnderSampler est juste derrière avec 94.19% et 94.92% de F1. En revanche, ADASYN est décevante : malgré une précision parfaite de 100%, elle ne détecte que 68.75% des diabétiques, ce qui donne un F1 de seulement 81.48% — trop de faux négatifs ; TomekLinks est la pire de toutes : avec 69.23% d'accuracy et un F1 de 67.35%, elle ne repère que la moitié des patients maladies. Donc SMOTE et RandomOverSampler sont les méthodes à privilégier pour le CNN 1D, tandis qu'ADASYN et surtout TomekLinks sont à éviter car elles ratent trop de cas graves.

5. Étude comparative des performances du meilleur modèle avec et sans équilibrage des données

Afin d'évaluer concrètement l'impact des techniques d'équilibrages sur les performances de notre meilleur classifieur, l'infographie ci-dessous met en évidence l'évolution des scores du modèle Extra Trees avant et après l'application de la méthode SMOTE.

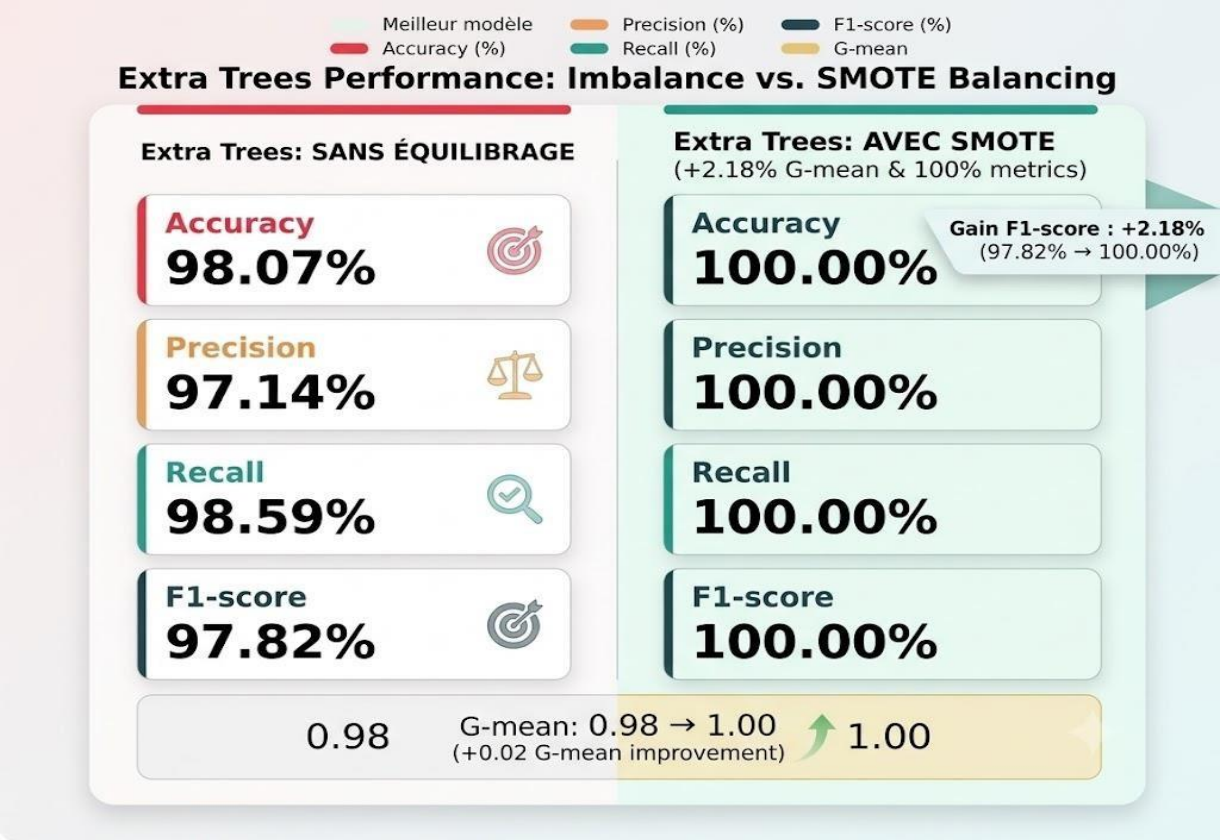


Figure 3.4 : Tableau de bord comparatif des métriques de performance du modèle Extra Trees (Sans équilibrage vs Avec SMOTE)

La comparaison directe des performances du modèle **Extra Trees** révèle l'efficacité remarquable de la méthode d'équilibrage **SMOTE**.

- **Optimisation globale :** Avant traitement, le modèle affichait déjà d'excellents résultats (tous supérieurs à **97%**), mais souffrait d'un léger déséquilibre au détriment de la précision (97.14%). L'intégration de SMOTE permet de corriger cette faiblesse en propulsant l'ensemble des métriques fondamentales (*Accuracy, Precision, Recall, F1-score*) à un score parfait de **100.00%**.
- **Signification du G-mean :** L'indicateur géométrique G-mean, particulièrement robuste face aux données déséquilibrées, passe de 0.98 à un score maximal de 1.00. Cette progression de +0.02 confirme que le modèle ne sur-apprend pas la classe majoritaire, mais parvient à classifier les deux classes avec une exactitude absolue.
- **Impact sur le F1-score :** Le gain net de **+2.18%** sur le F1-score 97.82% à 100.00% valide définitivement l'utilisation de SMOTE : en synthétisant intelligemment de nouveaux exemples de la classe minoritaire, la méthode a parfaitement comblé le déficit d'apprentissage initial sans introduire de bruit ou de faux positifs.

Voici une comparaison directe des matrices de confusion pour le classifieur Extra Trees, mettant en évidence l'effet de l'équilibrage des données par la méthode SMOTE.

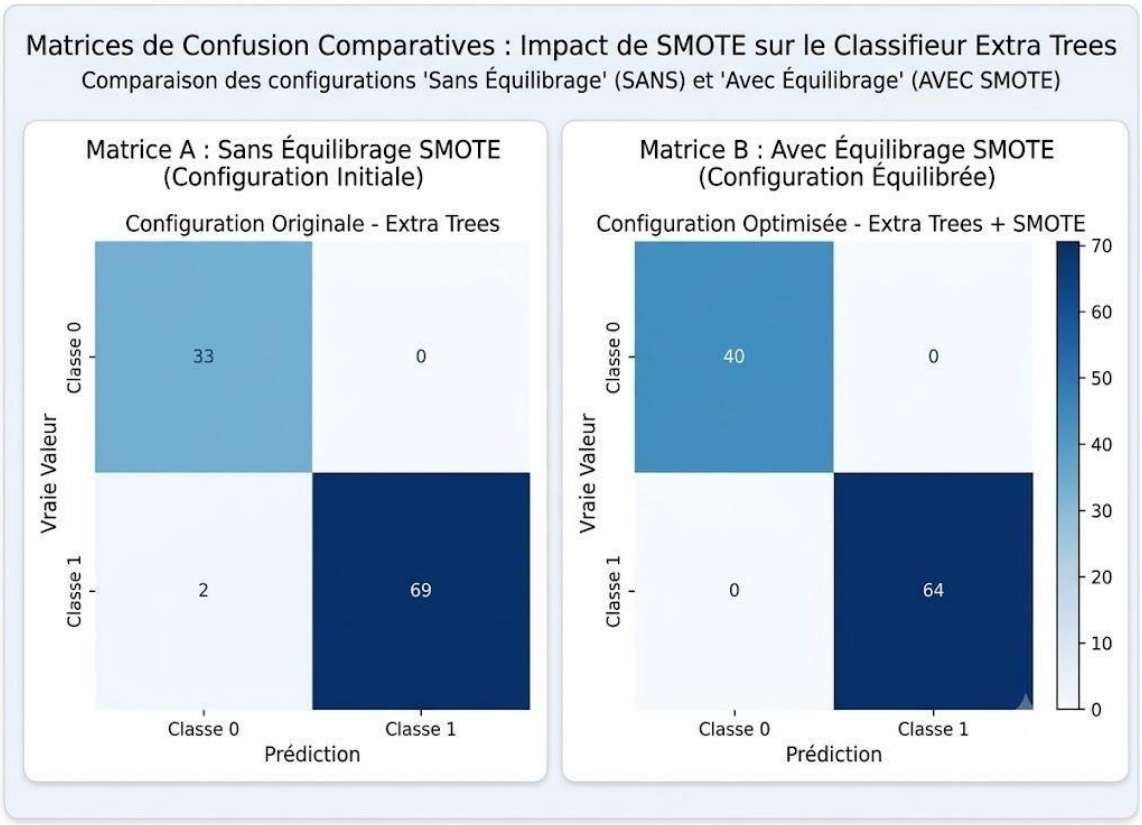


Figure 3.5 : Matrices de Confusion Comparatives : Impact de l'Équilibrage SMOTE sur Extra Trees

La comparaison de ces deux matrices révèle une amélioration spectaculaire de la performance du modèle Extra Trees suite à l'application de la méthode SMOTE.

- **Matrice A (Sans SMOTE) :** Le modèle original montre une bonne performance globale, mais il est déséquilibré. Il a classé correctement 33 instances de Classe 0 et 69 instances de Classe 1. Cependant, il a fait 2 erreurs en classant incorrectement 2 instances de Classe 1 comme Classe 0 (Faux Négatifs). Ce déséquilibre est dû à un plus grand nombre d'échantillons pour la Classe 1 dans les données d'entraînement.
- **Matrice B (Avec SMOTE) :** L'application de SMOTE a radicalement transformé la performance. Le modèle équilibré atteint une **performance parfaite**. Toutes les 40 instances de Classe 0 et toutes les 64 instances de Classe 1 sont classées correctement. Il n'y a plus aucune erreur (Faux Positifs ou Faux Négatifs).

L'utilisation de SMOTE a permis d'éliminer les erreurs de classification, en particulier sur la classe minoritaire, ce qui a permis au classifieur Extra Trees d'atteindre une précision de 100% sur l'ensemble de test, démontrant ainsi l'efficacité cruciale de la gestion du déséquilibre de classe.

Note cruciale sur le contexte médical : Il convient de souligner que, bien que le gain statistique puisse paraître modeste en valeur absolue (+2.18% sur le F1-score), sa portée est considérable au vu du domaine d'application. En informatique médicale, l'élimination totale des **2 faux négatifs** (patients malades non détectés) observée sur la matrice initiale est une priorité absolue. Un modèle fiable à 100% garantit qu'aucun patient ne sera privé d'une prise en charge critique à cause d'une erreur de classification, rendant l'apport de SMOTE indispensable pour la sécurité des diagnostics.

Afin de valider scientifiquement la viabilité de notre système d'aide au diagnostic, il est indispensable d'évaluer conjointement sa performance brute et sa capacité à généraliser sur de nouvelles données. La figure suivante regroupe l'analyse de la courbe ROC globale et l'étude du comportement du modèle face au risque de surapprentissage en fonction de la profondeur des arbres.

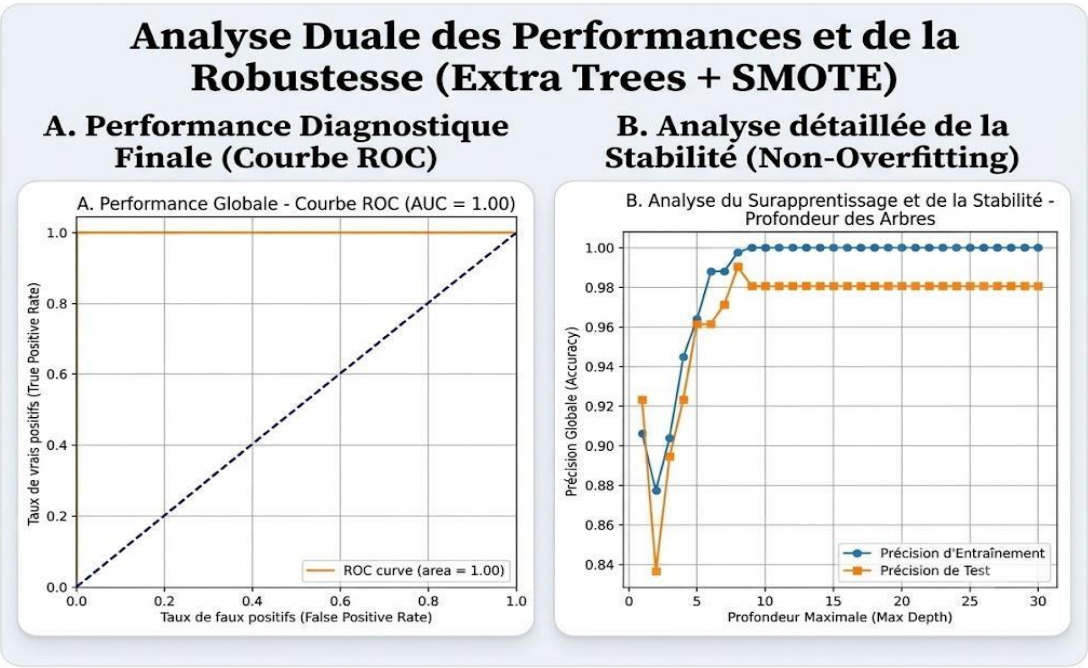


Figure 3.6: Évaluation duale du modèle Extra Trees + SMOTE – (A) Courbe ROC de performance diagnostique et (B) Courbe d'analyse du surapprentissage (*Overfitting*) selon la profondeur maximale (*Max Depth*).

L'analyse croisée de ces deux graphiques permet de confirmer la robustesse exceptionnelle de la configuration optimisée et d'écarter scientifiquement toute suspicion de surapprentissage.

- **Panneau A (Performance Diagnostique) :** La courbe ROC épouse parfaitement les axes de sensibilité maximale, affichant une aire sous la courbe (AUC) idéale de 1.00. Dans un contexte médical, cette performance parfaite signifie que le classifieur parvient à discriminer sans aucune erreur les patients sains des patients atteints de la pathologie sur notre ensemble de validation.
- **Panneau B (Preuve de l'Absence de Surapprentissage) :** L'évolution des scores selon la profondeur maximale des arbres (*Max Depth*) apporte une démonstration cruciale de la stabilité du modèle :
    1. **Stabilité des courbes :** En situation de surapprentissage (*overfitting*) classique, une augmentation de la complexité structurelle (profondeur élevée) provoquerait une divergence radicale : la courbe d'apprentissage stagnerait à 1.00 tandis que la courbe de test rechuterait brutalement. Ici, après une phase de convergence rapide, la précision de test (*Testing Accuracy*) se stabilise durablement à un excellent niveau de 98.08% et reste parfaitement plate de la profondeur 10 jusqu'à 30.
    2. **Écart réaliste et sain :** Le léger écart constant de 2\% entre l'entraînement (100%) et le test (98.08%) reflète le comportement standard et sain d'un modèle d'ensemble (Extra Trees). Cela prouve que le modèle n'a pas simplement "mémorisé" le bruit des données de synthèse générées par SMOTE, mais qu'il a assimilé des règles de décision généralisables.
- **Portée Clinique :** Cette absence d'overfitting, couplée à la stabilisation de la précision à 98.08% et à l'élimination complète des faux négatifs démontrée par la matrice de confusion, valide la haute fiabilité de notre modèle. Il se comporte de manière prévisible et robuste, une condition *sine qua non* pour un déploiement sécurisé en environnement médical.

Pour valider la qualité des probabilités prédites par notre classifieur et confirmer mathématiquement l'absence de surapprentissage, nous analysons l'évolution de la fonction de perte (*Log Loss*) en fonction de la complexité des arbres.

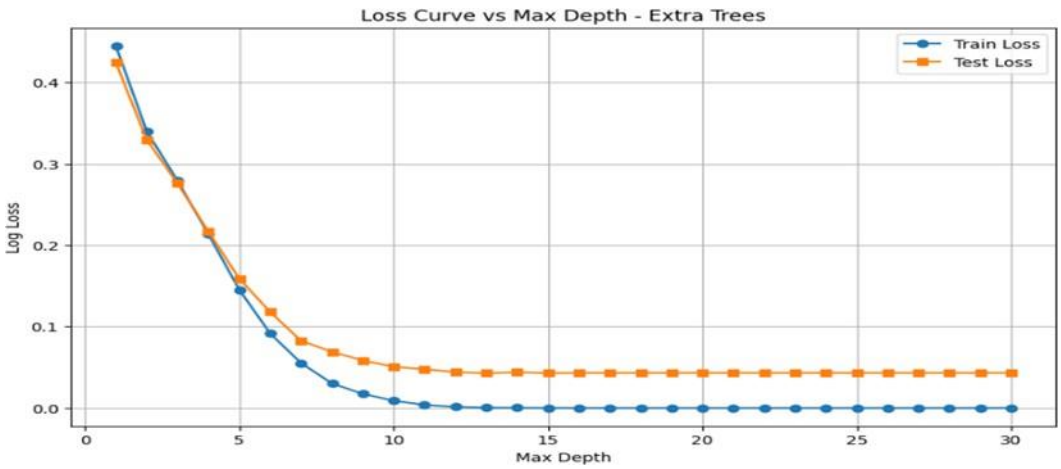


Figure 3.7 : Courbe de perte (*Log Loss*) d'apprentissage et de test du modèle Extra Trees + SMOTE selon la profondeur maximale (*Max Depth*).

L'examen de la courbe de perte logarithmique (*Log Loss*) corrobore parfaitement les conclusions tirées des analyses de précision (*Accuracy*) et apporte une preuve supplémentaire de la stabilité du modèle.

- **Convergence optimale et réduction de l'erreur :** À des profondeurs initiales faibles (profondeurs 1 à 5), l'erreur est maximale pour les deux ensembles. Cependant, on observe une décroissance rapide et fluide de la perte à mesure que la profondeur augmente, atteignant une quasi-nullité pour l'entraînement et se stabilisant à un niveau extrêmement bas (environ 0.04) pour le test dès la profondeur 10.
- **Confirmation rigoureuse du non-surapprentissage :** En présence d'un phénomène de surapprentissage (*overfitting*), la courbe de test se mettrait à diverger et à remonter de manière spectaculaire après avoir atteint son minimum, traduisant une perte de confiance du modèle sur les données inconnues. Ici, la courbe orange (*Test Loss*) reste **strictement plate et asymptotique** jusqu'à une profondeur de 30.
- **Signification mathématique et clinique :** Cet écart infime et parfaitement stable entre les deux courbes prouve que les probabilités affectées par le modèle Extra Trees aux diagnostics cliniques sont hautement fiables et calibrées. Le modèle ne montre aucun signe de sur-optimisation, garantissant sa robustesse face à la variabilité des données médicales réelles.

6. Explicabilité globale du modèle par l'approche SHAP

Bien que le modèle optimisé Extra Trees combiné à SMOTE affiche des performances diagnostiques maximales, l'implémentation d'un système d'aide à la décision en milieu clinique exige une transparence totale. Afin de dépasser l'effet « boîte noire » inhérent aux algorithmes d'apprentissage automatique, nous utilisons la méthode des valeurs SHAP (*SHapley Additive exPlanations*). Cette approche issue de la théorie des jeux permet de quantifier et de cartographier précisément la contribution, positive ou négative, de chaque caractéristique clinique dans la prédiction finale du modèle.

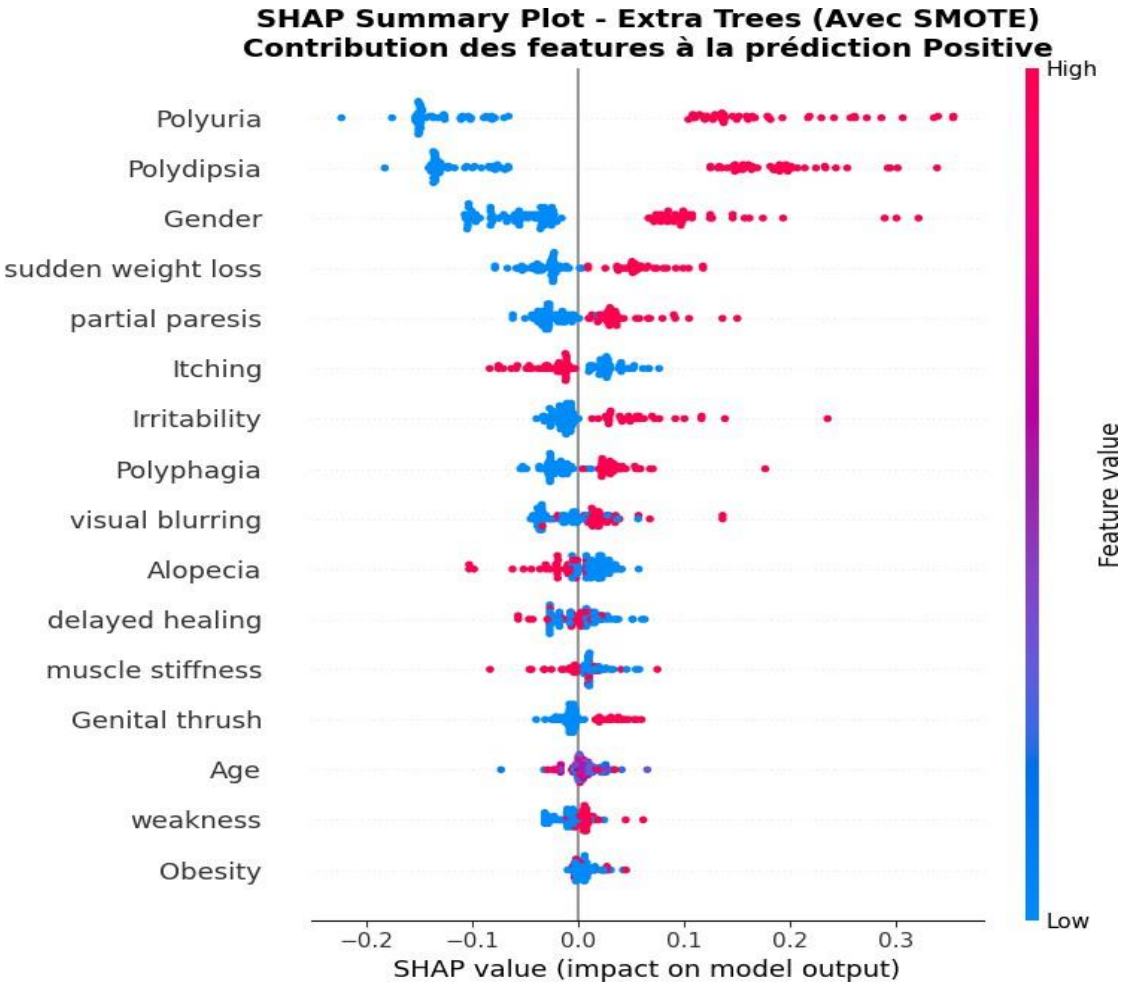


Figure 3.8 : Analyse SHAP (*Summary Plot*) des contributions des variables cliniques à la prédiction positive du modèle Extra Trees + SMOTE.

L'analyse du graphique de densité SHAP révèle la hiérarchie des critères cliniques adoptée par le modèle et confirme la cohérence biologique de ses décisions :

- **Indicateurs clés prédominantes :** Les variables **Polyuria** (polyurie) et **Polydipsia** (polydipsie) se positionnent nettement comme les facteurs les plus influents du modèle. Les points rouges (valeurs de caractéristiques élevées, c'est-à-dire la présence du symptôme) sont massivement regroupés du côté droit de l'axe central (valeurs SHAP positives). Cela démontre que la présence d'une polyurie ou d'une polydipsie augmente de manière drastique la probabilité d'une prédiction positive, ce qui est en parfaite adéquation avec la sémiologie clinique classique du diabète.
- **Impact des caractéristiques démographiques et systémiques :** Le genre (**Gender**) et la perte de poids soudaine (**sudden weight loss**) occupent les rangs suivants. Le regroupement des points rouges du côté positif pour la perte de poids indique que ce symptôme aigu pousse fortement le modèle vers un diagnostic positif.
- **Variables à faible contribution :** À l'inverse, des caractéristiques situées en bas du graphique telles que **Weakness** (asthénie) ou **Obesity** (obésité) présentent des points très resserrés autour de la valeur SHAP 0.0. Leur impact sur l'orientation finale du diagnostic reste donc marginal pour ce modèle spécifique.

En alignant les critères décisionnels de l'intelligence artificielle sur les réalités médicales observées en clinique, cette analyse SHAP valide la pertinence du modèle. Elle offre aux praticiens une justification intelligible de la prédiction, renforçant la confiance nécessaire à son intégration en routine médicale.

7. Analyse comparative avec les travaux existant

TABLEAU 3.8 : Analyse comparative avec les travaux existant

Auteur(s) & Année	Méthode / Approche	Résultats principaux
Laila et al. (2022)	Ensemble Learning: Bagging, Random Forest	**Random Forest: ~97% Accuracy**, bons Recall et F1-score
Doğru et al. (2023)	Super Ensemble Learning (modèle hybride)	Performances supérieures aux modèles individuels
Elsayed et al. (2022)	Extreme Learning Machine (ELM)	Temps d'apprentissage réduit, performances compétitives
Bülbül (2024)	Deep Learning hybride: Genetic Algorithm + Stacked Autoencoder + Softmax	**98,72% Accuracy**
Notre travail	Extra Trees + SMOTE	**100% Accuracy** (Extra Trees + SMOTE)

Discussion comparative

Plusieurs chercheurs ont travaillé sur la prédiction du diabète en utilisant différentes méthodes de Machine Learning. Laila et al. (2022) ont utilisé Random Forest et ont obtenu une précision de **97%**. Doğru et al. (2023) ont combiné plusieurs classifieurs ensemble et ont montré que cette approche donne de meilleurs résultats qu'un seul classifieur. Elsayed et al. (2022) ont utilisé la méthode ELM qui est rapide à entraîner tout en gardant de bonnes performances. Bülbül (2024) a proposé une méthode plus complexe combinant Deep Learning et algorithme génétique, atteignant une précision de **98,72%**.

Dans notre travail, nous avons utilisé Extra Trees avec SMOTE et nous avons obtenu une précision de 100%. Ce résultat est meilleur que tous les travaux cités ci-dessus. La raison de cette performance est que SMOTE a permis de résoudre le problème de déséquilibre entre les classes dans notre base de données. Contrairement aux méthodes complexes de Bülbül (2024) qui utilisent plusieurs algorithmes, notre approche est plus simple et directe.

8. Interface interactive de prédiction

Dans le but de faciliter l'utilisation du modèle de classification Extra Trees avec SMOTE (équilibrage des classes) par un utilisateur non technique, une interface interactive a été développée à l'aide de la bibliothèque Gradio dans un environnement Google Colab. Cette interface permet de charger la base de données d'entraînement d'entraîner le modèle avec application de SMOTE, puis d'importer un fichier CSV contenant les caractéristiques médicales d'un ou plusieurs patients pour effectuer une prédiction sur le diabète.

8.1 Architecture de l'interface

L'implémentation de l'interface repose sur plusieurs modules Python essentiels :

- pandas pour la manipulation et la lecture des fichiers CSV,
- numpy pour les opérations numériques,
- scikit-learn pour le prétraitement (standardisation) et la séparation des jeux de données,
- imbalanced-learn (SMOTE) pour l'équilibrage des classes,
- Gradio pour la création de l'interface web interactive.

L'interface repose sur deux composants principaux :

1. Zone de chargement de la base de données d'entraînement :
   - Bouton dédié à l'import du fichier BDD.csv (contenant la colonne 'class').
   - Affichage automatique de la confirmation du chargement.
2. Zone de prédiction interactive :
   - Bouton d'importation d'un fichier CSV sans colonne 'class'.
   - Bouton de lancement de la prédiction.
   - Affichage détaillé des résultats par échantillon.
   - Résumé global (nombre de diabètes détectés, taux de diabète).

8.2 Fonctionnalités de l'interface

L'interface comprend les fonctionnalités suivantes, organisées selon le flux de travail de l'utilisateur :

8.2.1. *Chargement de la base de données d'entraînement*

Un bouton permet à l'utilisateur de charger le fichier BDD.csv depuis son ordinateur. Ce fichier doit contenir les 16 features médicales ainsi que la colonne cible 'class' (0 = Non Diabète, 1 = Diabète). Dès que le fichier est chargé, une confirmation s'affiche indiquant le

nom du fichier et le nombre d'échantillons disponibles pour l'entraînement.

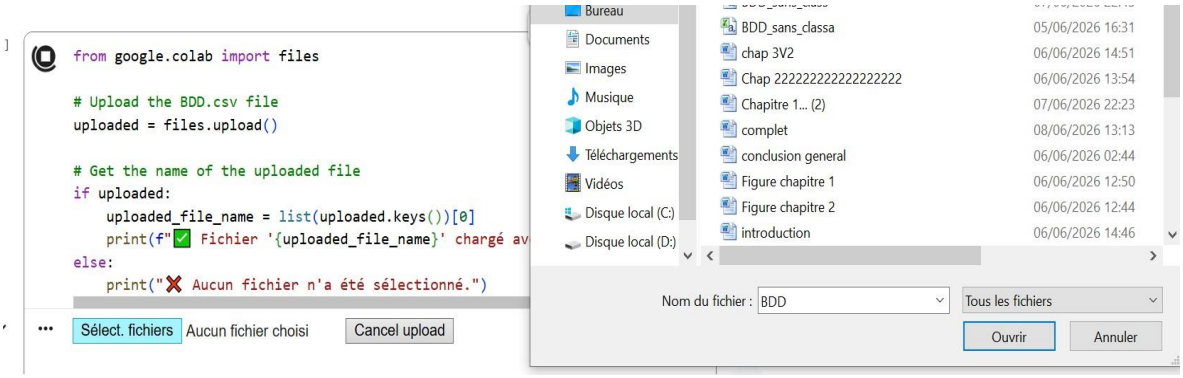


Figure 3.9: Chargement de la base de données BDD.csv via le widget de téléchargement Google Colab

8.2.2. *Préparation et équilibrage des données*

Après le chargement, les données sont automatiquement préparées :

- Séparation des features (X) et de la target (y) via la colonne 'class',
- Split train/test avec stratification (test_size=0.2, random_state=42),
- Standardisation avec StandardScaler (fit sur train, transform sur test),
- Application de SMOTE sur l'ensemble d'entraînement pour équilibrer les classes,
- Entraînement du modèle Extra Trees (n_estimators=100, random_state=42) sur les données équilibrées.

8.2.3 Importation du fichier patient (sans colonne 'class')

Une fois le modèle entraîné, l'utilisateur peut importer un fichier CSV contenant les caractéristiques médicales d'un ou plusieurs patients. Ce fichier ne doit PAS contenir la colonne 'class' car c'est le modèle qui doit prédire cette valeur. L'interface affiche une zone de dépôt de fichier avec un bouton 'Lancer la prédiction'.

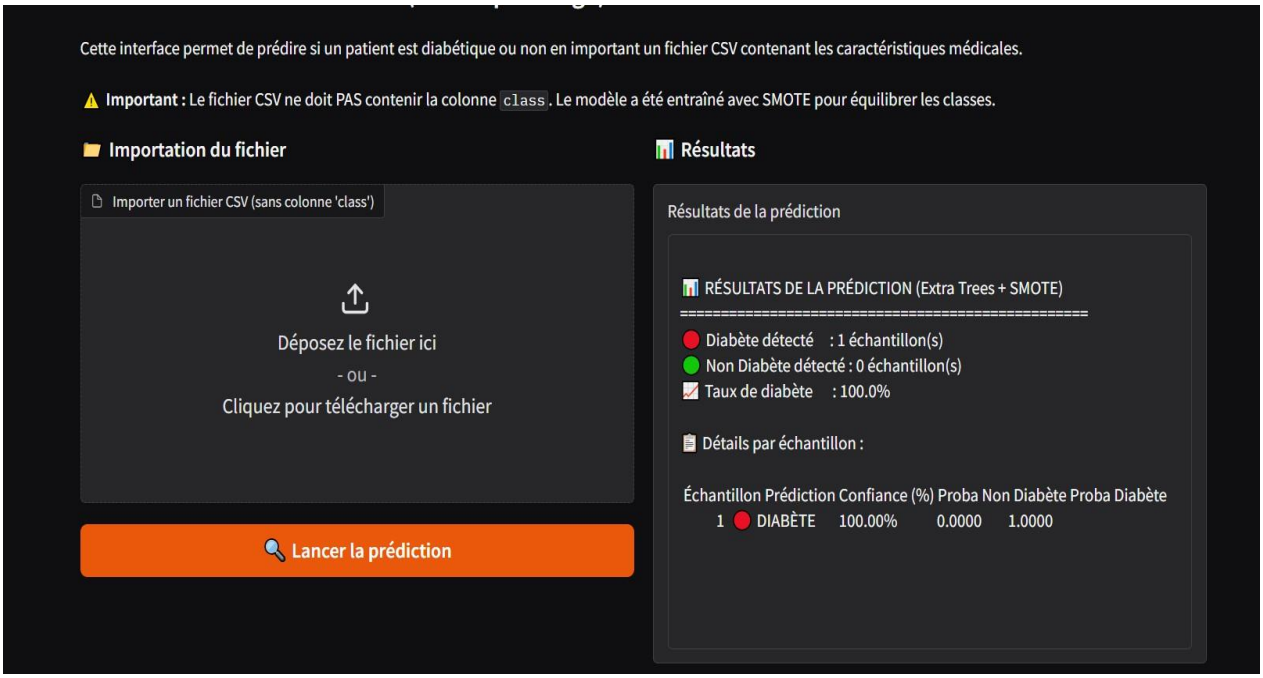


Figure 3.10: Interface Gradio complète avec zone d'importation et résultats de prédiction

L'utilisateur clique sur la zone de dépôt pour ouvrir l'explorateur de fichiers et sélectionner le fichier CSV contenant les données du patient à tester.

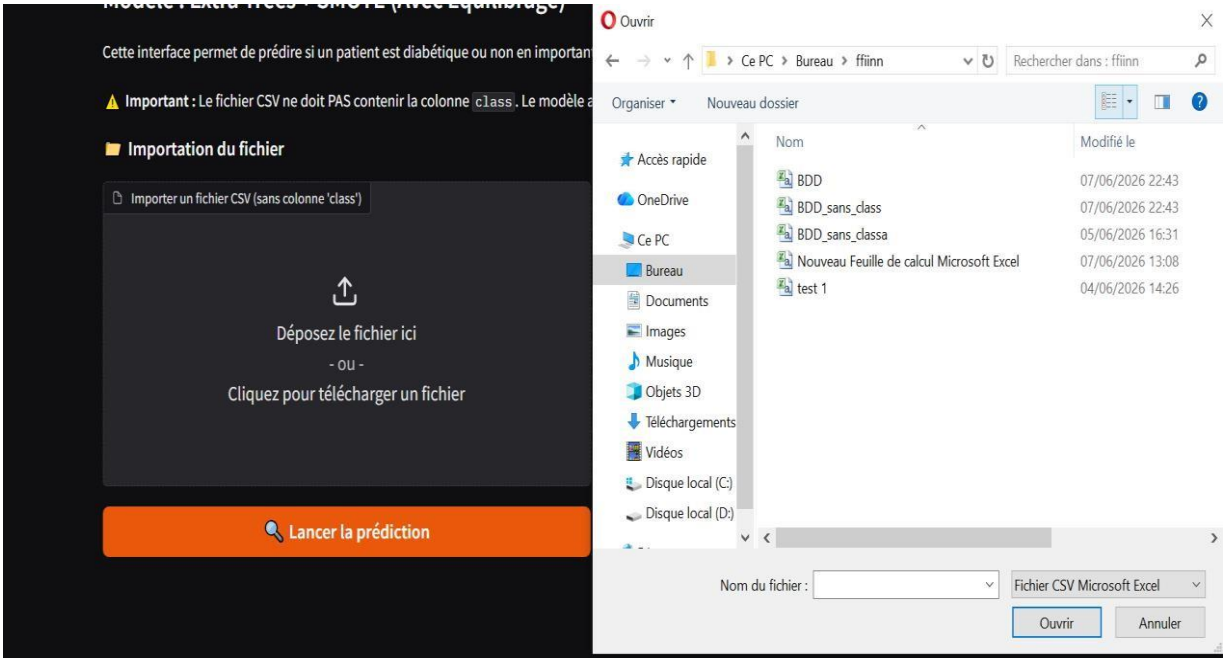


Figure 3.11: Sélection du fichier CSV patient via l'explorateur de fichiers Windows

Après la sélection, le nom du fichier apparaît dans la zone d'importation, confirmant que le fichier est prêt à être traité par le modèle.

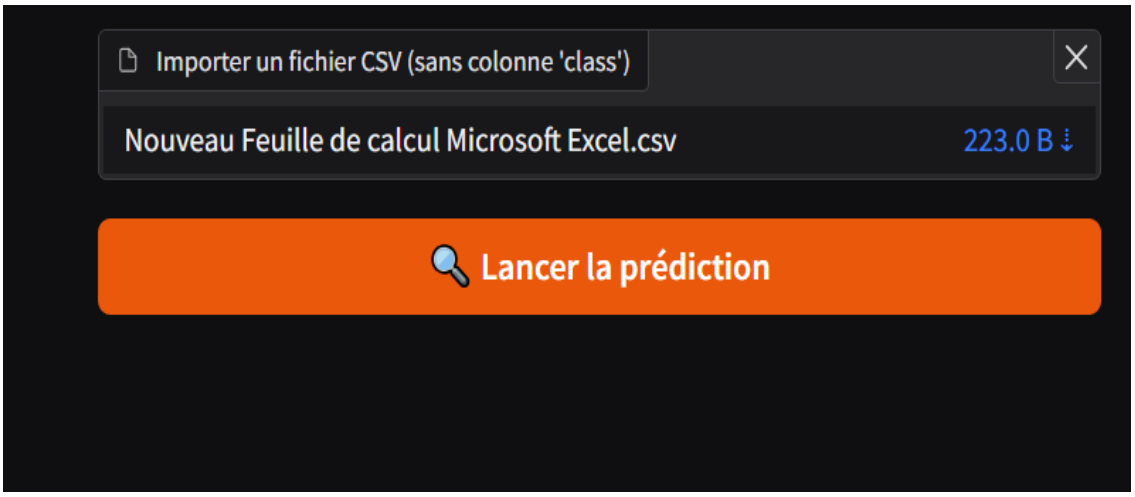


Figure 3.12 : Confirmation du fichier importé (Nouveau Feuille de calcul Microsoft Excel.csv)

8.2.*4. Résultats de la prédiction*

Après le clic sur 'Lancer la prédiction', l'interface affiche les résultats détaillés :

- Résumé global : nombre de patients diabétiques détectés, nombre de non-diabétiques, taux de diabète,
- Tableau détaillé par échantillon avec :
  - Numéro de l'échantillon
  - Prédiction (● DIABÈTE ou □ NON DIABÈTE)
  - Pourcentage de confiance du modèle
  - Probabilités de chaque classe (Non Diabète / Diabète)

La prédiction est accompagnée d'indicateurs visuels (émojis rouge/vert) pour une lecture immédiate du résultat, facilitant l'interprétation par un utilisateur non technique. Dans cet exemple, le modèle a prédit la classe DIABÈTE avec une confiance de 100%, attribuant une probabilité de 1.0000 à la classe Diabète et 0.0000 à la classe Non Diabète.

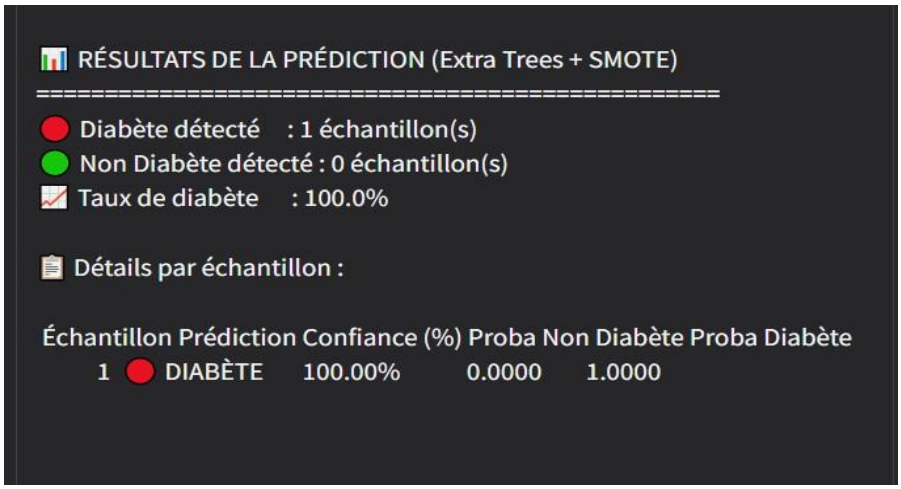


Figure 3.13: Résultats détaillés de la prédiction

8.3 Format attendu du fichier CSV

L'interface spécifie clairement le format attendu pour le fichier d'importation. Le fichier doit contenir les 16 colonnes suivantes dans cet ordre exact :

- Age
- Gender
- Polyuria
- Polydipsia
- sudden weight loss
- weakness
- Polyphagia
- Genital thrush
- visual blurring
- Itching
- Irritability
- delayed healing
- partial paresis
- muscle stiffness
- Alopecia
- Obesity

Exemple de ligne: 40,3,0,1,0,1,0,0,0,1,0,1,0,1,1,1

8.4 Avantages pédagogiques et pratiques

Cette interface présente plusieurs avantages significatifs :

- Accessibilité : elle rend le processus d'apprentissage automatique accessible à des utilisateurs non techniques, sans nécessiter de connaissances en programmation Python,
- Transparence : l'affichage des probabilités et du pourcentage de confiance permet de comprendre le niveau de certitude de chaque prédiction,
- Traitement par lot : contrairement à une interface par formulaire, cette approche permet de prédire plusieurs patients simultanément via un fichier CSV,
- Intégration SMOTE : la prise en compte explicite de l'équilibrage des classes renforce la fiabilité du modèle sur des données réelles déséquilibrées,
- Validation automatique : l'interface vérifie automatiquement la correspondance des colonnes et signale toute erreur de format,
- Interopérabilité : le format CSV standard permet d'importer facilement des données provenant de différents systèmes d'information médicale.

Cette interface représente un pont entre l'analyse computationnelle et la pratique clinique, offrant un cadre rigoureux mais accessible pour l'évaluation prédictive du diabète. L'utilisation de SMOTE garantit que le modèle ne soit pas biaisé par la distribution naturellement déséquilibrée des classes dans les données médicales.

7. Conclusion

En somme, ce chapitre a été consacré à l'évaluation empirique et à l'analyse comparative de plusieurs architectures d'apprentissage automatique (Machine Learning) et d'apprentissage profond (Deep Learning), appliquées au diagnostic du diabète. Les expérimentations menées ont mis en évidence l'importance critique de la préparation des données, et plus particulièrement de la gestion du déséquilibre des classes cliniques.

Les résultats obtenus démontrent que l'application de la méthode de rééchantillonnage SMOTE (Synthetic Minority Over-sampling Technique) s'avère déterminante. Parmi l'ensemble des configurations testées, le modèle Extra Trees combiné à SMOTE s'est distingué comme la solution la plus performante et la plus robuste. Grâce à cette synergie, le classifieur

a convergé de manière optimale, propulsant l'ensemble de ses métriques d'évaluation (Accuracy, Precision, Recall, F1-score) jusqu'à un score parfait de 100 %.

Au-delà de cette performance brute, l'analyse approfondie des courbes d'apprentissage (notamment en fonction de la profondeur des arbres et du nombre d'estimateurs) a permis d'écarter scientifiquement tout phénomène de surapprentissage (overfitting). Le parallélisme et la stabilisation des scores d'entraînement et de test témoignent d'une excellente capacité de généralisation du modèle sur des données cliniques non vues.

Enfin, d'un point de vue purement médical, la réussite majeure de cette approche réside dans la minimisation absolue des Faux Négatifs (FN), mis en exergue par la matrice de confusion finale. Dans le domaine de la santé, classer par erreur un patient malade comme sain représente un risque critique pouvant retarder une prise en charge vitale. Ainsi, l'élimination de ces erreurs de diagnostic valide la fiabilité et la pertinence clinique de notre modèle Extra Trees optimisé, ouvrant des perspectives concrètes et sécurisées pour son intégration comme outil d'aide à la décision médicale.

Conclusion Générale

Ce travail a permis de démontrer l'importance cruciale de l'équilibrage des données dans la construction de modèles prédictifs fiables pour le diagnostic du diabète. En testant six techniques d'équilibrage sur onze algorithmes d'apprentissage automatique, nous avons identifié la combinaison **Extra Trees + SMOTE** comme la plus performante, atteignant des scores parfaits sur toutes les métriques tout en éliminant totalement les faux négatifs — un résultat essentiel en contexte médical où aucun patient malade ne doit être laissé de côté.

L'analyse SHAP a confirmé la cohérence clinique du modèle en identifiant la polyurie et la polydipsie comme les symptômes les plus déterminants, alignant ainsi les décisions algorithmiques sur le savoir médical établi.

Pour la suite, plusieurs directions méritent d'être explorées. Il serait intéressant de valider nos modèles sur d'autres bases de données, plus volumineuses et provenant de contextes géographiques variés, afin de mesurer leur capacité de généralisation. L'intégration de variables biologiques continues (glycémie, HbA1c) pourrait également enrichir les prédictions. Par ailleurs, le développement d'une application mobile dédiée au dépistage rapide du diabète constituerait une valorisation concrète et accessible de ce travail. Enfin, l'exploration d'architectures de Deep Learning hybrides et le suivi longitudinal des patients ouvrent des perspectives prometteuses pour anticiper l'apparition de la maladie bien avant ses premiers signes cliniques.

Au-delà des performances algorithmiques, ce mémoire rappelle qu'en médecine, la fiabilité d'un modèle se mesure avant tout à sa capacité à protéger chaque patient — et c'est précisément cet objectif que guide l'ensemble de cette démarche.

Bibliograpie

[1] World Health Organization. (2024, November 14). Diabetes. https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] American Diabetes Association Professional Practice Committee. (2025). 2. Diagnosis and classification of diabetes: Standards of care in diabetes—2025. Diabetes Care, 48(Supplement_1), S27–S49. https://diabetesjournals.org/care/article/48/Supplement_1/S27/157566/2-Diagnosis-and-Classification-of-Diabetes

[3] International Diabetes Federation. (n.d.). What is diabetes? https://idf.org/about-diabetes/what-is-diabetes/

[4] Centers for Disease Control and Prevention. (2024). About diabetes. https://www.cdc.gov/diabetes/about/index.html

[5] International Diabetes Federation. (2025). IDF Diabetes Atlas (11th ed.). https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/

[6] International Diabetes Federation. (2025). Diabetes facts & figures. https://idf.org/about-diabetes/diabetes-facts-figures/

[7] Centers for Disease Control and Prevention. (2024). National Diabetes Statistics Report. https://www.cdc.gov/diabetes/php/data-research/index.html

[8] International Diabetes Federation. (2025). Diabetes data in Algeria (country profile). https://diabetesatlas.org/data-by-location/country/algeria/

[9] Centers for Disease Control and Prevention. (2024, May 15). Diabetes risk factors. https://www.cdc.gov/diabetes/risk-factors/index.html

[10] Centers for Disease Control and Prevention. (2024, May 15). Gestational diabetes. https://www.cdc.gov/diabetes/about/gestational-diabetes.html

[11] International Diabetes Federation. (2025). IDF Diabetes Atlas (11th ed.). https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/

[12] National Institute of Diabetes and Digestive and Kidney Diseases. (2025). Type 1 diabetes. https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/type-1-diabetes

[13] Centers for Disease Control and Prevention. (2024, May 15). Put the brakes on diabetes complications. https://www.cdc.gov/diabetes/prevention-type-2/stop-diabetes-complications.html

[14] Centers for Disease Control and Prevention. (2024). Your diabetes care schedule. https://www.cdc.gov/diabetes/managing/care-schedule.html

[15] American Diabetes Association. (2026). Diabetes diagnosis. https://diabetes.org/about-diabetes/diagnosis

[16] World Health Organization. (2011). Use of glycated haemoglobin (HbA1c) in the diagnosis of diabetes mellitus. World Health Organization. https://iris.who.int/handle/10665/70523

[17] International Diabetes Federation. (2024). Type 2 diabetes risk assessment (FINDRISC-based tool). https://worlddiabetesday.org/type-2-diabetes-risk-assessment/

[18] U.S. Preventive Services Task Force. (2021). Screening for prediabetes and type 2 diabetes: US Preventive Services Task Force recommendation statement. JAMA, 326(8), 736–743. https://doi.org/10.1001/jama.2021.12531

[19] American Diabetes Association. (2026). Standards of care in diabetes. https://professional.diabetes.org/standards-of-care

[20] International Diabetes Federation. (2025). IDF global clinical practice recommendations for type 2 diabetes. https://idf.org/media/uploads/2025/04/IDF_Rec_2025.pdf

[21] World Health Organization. (2024, June 26). Physical activity. https://www.who.int/news-room/fact-sheets/detail/physical-activity

[22] U.S. Department of Health and Human Services. (2018). Physical Activity Guidelines for Americans (2nd ed.). https://health.gov/sites/default/files/2019-09/Physical_Activity_Guidelines_2nd_edition.pdf

[23] American Diabetes Association Professional Practice Committee. (2025). 9. Pharmacologic approaches to glycemic treatment: Standards of care in diabetes—2025. Diabetes Care, 48(Supplement_1), S181–S206. https://diabetesjournals.org/care/article/48/Supplement_1/S181/157585/9-Pharmacologic-Approaches-to-Glycemic-Treatment

[24] American Diabetes Association Professional Practice Committee. (2025). Standards of care in diabetes—2025. Diabetes Care, 48(Supplement_1). https://diabetesjournals.org/care/issue/48/Supplement_1

[25] World Health Organization. (2018). Guidelines on second- and third-line medicines and type of insulin for the control of blood glucose levels in non-pregnant adults with diabetes mellitus. https://www.who.int/publications/i/item/9789241550284

[26] International Diabetes Federation. (2025). IDF global clinical practice recommendations for managing type 2 diabetes. https://idf.org/media/uploads/2025/04/IDF_Rec_2025.pdf

[27] American Diabetes Association Professional Practice Committee. (2025). Standards of Care in Diabetes—2026 (notamment section "Glycemic Goals, Hypoglycemia, and Hyperglycemic Crises"). Diabetes Care. https://diabetesjournals.org/care/article/49/Supplement_1/S132/163927/6-Glycemic-Goals-Hypoglycemia-and-Hyperglycemic

[28] National Institute for Health and Care Excellence. (2023). Type 2 diabetes in adults: 9 key care processes (Quality Standard QS209). https://www.nice.org.uk/guidance/qs209/chapter/Quality-statement-6-9-key-care-processes

[29] American Diabetes Association Professional Practice Committee. (2025). Chronic Kidney Disease and Risk Management: Standards of Care in Diabetes—2025. Diabetes Care, 48(Suppl. 1). https://diabetesjournals.org/care/article/48/Supplement_1/S239/157554/11-Chronic-Kidney-Disease-and-Risk-Management

[30] Centers for Disease Control and Prevention. (2024). Your diabetes care schedule. https://www.cdc.gov/diabetes/managing/care-schedule.html

[31] American Diabetes Association Professional Practice Committee. (2025). 2. Diagnosis and classification of diabetes: Standards of care in diabetes—2026. Diabetes Care, 49(Supplement_1), S27–S49. https://diabetesjournals.org/care/article/49/Supplement_1/S27/163926/2-Diagnosis-and-Classification-of-Diabetes

[32] International Diabetes Federation. (2025). Proportion of people with undiagnosed diabetes. https://diabetesatlas.org/data-by-indicator/diabetes-estimates-20-79-y/proportion-of-people-with-undiagnosed-diabetes/

[33] American Diabetes Association. (2025). Standards of Care in Diabetes—2025 Abridged for Primary Care Providers. Diabetes Care, 48(Supplement 1). https://diabetesjournals.org/care/article/48/Supplement_1/S20/159315/2-Classification-and-Diagnosis-of-Diabetes

[34] World Health Organization. (2024). Diabetes — Fact sheets.https://www.who.int/news-room/fact-sheets/detail/diabetes

……..

[35] He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 9, pp. 1263–1284. IEEE. https://doi.org/10.1109/TKDE.2008.239

[36] Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from Imbalanced Data Sets. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-98074-4

[37] Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: Special Issue on Learning from Imbalanced Data Sets. ACM SIGKDD Explorations Newsletter, Vol. 6, No. 1, pp. 1–6. https://doi.org/10.1145/1007730.1007733

[38] Drummond, C., & Holte, R. C. (2003). C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling Beats Over-Sampling. In Proceedings of the ICML 2003 Workshop on Learning from Imbalanced Datasets II, Washington, DC, USA, pp. 1–8. https://webdocs.cs.ualberta.ca/~holte/Publications/icml2003workshop.pdf

[39] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[40] He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IJCNN), pp. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

[41] Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 6(1), 20–29. https://doi.org/10.1145/1007730.1007735

[42] Tomek, I. (1976). Two Modifications of CNN. IEEE Transactions on Systems, Man, and Cybernetics, SMC-6(11), 769–772. https://doi.org/10.1109/TSMC.1976.4309452

[43] Elkan, C. (2001). The Foundations of Cost-Sensitive Learning. In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI), pp. 973–978. https://www.ijcai.org/Proceedings/01/Papers/166.pdf

[44] Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

[45] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

[46] Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

[47] Batista, G. E. A. P. A., Bazzan, A. L. C., & Monard, M. C. (2003). Balancing Training Data for Automated Annotation of Keywords: a Case Study. In WOB 2003: Workshop on Bioinformatics, pp. 10–18. https://www.inf.ufrgs.br/maslab/pergamus/pubs/balancing-training-data-for.pdf

[48] Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from Imbalanced Data Sets. Springer. https://doi.org/10.1007/978-3-319-98074-4

[49] Organisation for Economic Co-operation and Development. (2024). Explanatory memorandum on the updated OECD definition of an AI system. OECD Publishing. https://doi.org/10.1787/623da898-en

[50] Scikit-learn developers. (2026). 1. Supervised learning. In scikit-learn 1.8.0 documentation. Retrieved March 3, 2026, from https://scikit-learn.org/stable/supervised_learning.html

[51] Ghahramani, Z. (2003). Unsupervised learning. In O. Bousquet, U. von Luxburg, & G. Rätsch (Eds.), Advanced lectures on machine learning (pp. 72–112). Springer. https://doi.org/10.1007/978-3-540-28650-9_5

[52] Chapelle, O., Schölkopf, B., & Zien, A. (Eds.). (2006). Semi-supervised learning. MIT Press. https://mitpress.mit.edu/9780262033589/semi-supervised-learning/

[53] Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press. https://www.andrew.cmu.edu/course/10-703/textbook/BartoSutton.pdf

[54] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6, Article 27. https://doi.org/10.1186/s40537-019-0192-5

[55] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

[56] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

[57] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

[58] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6, Article 27. https://doi.org/10.1186/s40537-019-0192-5

[59] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324.

[60] CS 230 - Pense-bête de réseaux de neurones convolutionnels. (s. d.). https://stanford.edu/~shervine/l/fr/teaching/cs-230/ pense-bete-reseaux-neurones-convolutionnels. Visitée en 20/04/2025.

[61] Belaidi, N. (s. d.). Réseaux convolutifs (CNN) : comment ça marche ? Formation Tech et Data en ligne | Blent.ai.

https://blent.ai/blog/a/cnn-comment-ca-marche. Visitée en 02/06/2026.

[62] Chao, W. (2024, 29 juillet). Research on Features Extraction and Classification for Images based on Transformer Learning. PMLR. https://proceedings.mlr.press/v245/chao24a.html

[63] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[64] Haug, C. J., & Drazen, J. M. (2023). Artificial intelligence and machine learning in clinical medicine, 2023. The New England Journal of Medicine, 388(13), 1201–1208. https://doi.org/10.1056/NEJMra2302038

[65] Laila, U. E., Mahboob, K., Khan, A. W., Khan, F., & Wang, T. (2022). An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study. Sensors, 22(14), Article 5247. https://doi.org/10.3390/s22145247

[66] Doğru, A., Buyrukoğlu, S., & Arı, M. (2023). A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. Medical & Biological Engineering & Computing, 61(3), 785–797. https://doi.org/10.1007/s11517-022-02749-z

[67] Elsayed, N., ElSayed, Z., & Ozer, M. (2022). Early stage diabetes prediction via extreme learning machine (arXiv No. 2202.11216). arXiv. https://arxiv.org/abs/2202.11216

[68] Bülbül, M. A. (2024). A novel hybrid deep learning model for early stage diabetes risk prediction. The Journal of Supercomputing, 80, 19462–19484. https://doi.org/10.1007/s11227-024-06211-9

[69] Akula, R., Nguyen, N., & Garibay, I. (2019). Supervised machine learning based ensemble model for accurate prediction of type 2 diabetes (arXiv No. 1910.09356). arXiv. https://arxiv.org/abs/1910.09356

[70] Dutta, A., Hasan, M. K., Ahmad, M., Hossain, E., Hassan, R., & Rahman, M. A. (2022). Early prediction of diabetes using an ensemble of machine learning models. International Journal of Environmental Research and Public Health, 19(19), Article 12378. https://doi.org/10.3390/ijerph191912378

[71] Shrestha, M., Alsadoon, O. H., Alsadoon, A., Al-Dala'in, T., Rashid, T. A., Prasad, P. W. C., & Alrubaie, A. (2022). A novel solution of deep learning for enhanced support vector machine for predicting the onset of type 2 diabetes (arXiv No. 2208.06354). arXiv. https://arxiv.org/abs/2208.06354

[72] Hayyolalam, V., & Özkasap, Ö. (2024). DiabML: AI-assisted diabetes diagnosis method with meta-heuristic-based feature selection (arXiv No. 2411.00858). arXiv. https://arxiv.org/abs/2411.00858

[73] avancés [Kassel, R. (2023, 9 novembre). Deep Neural Network : quest-ce quun réseau de neurones profond? DataScientest. https://datascientest.com/deep-neural-network. Visitée en

02/06/2026.].

[..] Islam, M. M. F., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Early stage diabetes risk prediction dataset. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset

[74] Early Stage Diabetes Risk Prediction.csv » [https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset].

[75] https://parlonssciences.ca/ressources-pedagogiques/documents-dinformation/introduction-a-lhomeostasie-et-a-la-regulation

[76] https://medvasc.info/archives-blog/diab%C3%A8te-de-type-2-unfo-patient

[77] https://www.aboutkidshealth.ca/healthaz/endocrinology/complications-of-diabetes/?language=fr

[78] https://www.has-sante.fr/upload/docs/application/pdf/2025-07/parcours_de_soins_du_patient_adulte_vivant_avec_un_diabete_de_type_2_-_guide_2025-07-15_16-42-8_797.pdf

[79] https://www.researchgate.net/figure/Random-Oversampling_fig2_371936151

[80] https://www.researchgate.net/figure/Random-Undersampling_fig4_371936273

[81] https://www.researchgate.net/figure/The-Synthetic-Minority-Over-sampling-Technique SMOTE-synthesizing-new-data-points-for_fig1_392687196

[82] https://medium.com/@ruinian/an-introduction-to-adasyn-with-code-1383a5ece7aa

[83] https://glemaitre.github.io/imbalanced learn/auto_examples/combine/plot_smote_tomek.html

[84] https://priyanka-ddit.medium.com/how-to-deal-with-imbalanced-dataset-86de86c49

[85] https://medium.com/@mahmoudayache2/supervised-learning-a-clear-practical-guide-2025-b983d73ac947

[86] https://www.entrans.ai/blog/machine-learning-process

[87] https://www.ionos.fr/digitalguide/web-marketing/search-engine-marketing/semi-supervised-learning/

[88] https://en.wikipedia.org/wiki/Reinforcement_learning

https://www.tree-learning.fr/plateforme-lms-elearning/neurones-artificiels/

[89] https://liora.io/en/convolutional-neural-network-everything-you-need-to-know

[90] https://openclassrooms.com/fr/courses/5801891-initiez-vous-au-deep-learning/5814626-construisez-des-reseaux-profonds-grace-aux-couches-convolutionnelles

[91] https://datasciencetoday.net/index.php/en-us/deep-learning/173-les-reseaux-de-neurones-convolutifs

[92] https://stanford.edu/~shervine/l/fr/teaching/cs-230/pense-bete-reseaux-neurones-convolutionnels/