République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique
Université Echahid Cheikh Larbi Tébessi - Tébessa
Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie
Département : Mathématiques et Informatique
LAMIS
Laboratoire de Mathématiques, Informatique et Systèmes

Thèse

En vue de l'obtention du diplôme de

DOCTORAT LMD

en Informatique

Spécialité : Systèmes d'information

Thème

Approche basée IA pour un système de prédiction du diabète.

Présenté Par :
SAMET Sarra

Devant le jury :

Mr. Bendjenna Hakim	*Pr*	*Univérsité Larbi Tébessi*	*Président*
Mr. Siam Abderrahim	*Pr*	*Univérsité de Khenchela*	*Examinateur*
Mr. Laimeche Lakhdar	*MCA*	*Univérsité Larbi Tébessi*	*Examinateur*
Mr. Laouar Med Ridda	*Pr*	*Univérsité Larbi Tébessi*	*Encadrant*
Mr. Bendib Issam	*MCA*	*Univérsité Larbi Tébessi*	*Co-encadant*

Date de soutenance : 29 Mai 2024

Résumé

Cette thèse explore le potentiel de l'intelligence artificielle dans la prédiction précoce du diabète, en mettant en lumière des contributions innovantes tout en utilisant différentes bases de données.

Les deux premières contributions utilisent "la base de données des Indiens Pima". Où un modèle d'ensemble basé sur les trois meilleures méthodes obtenues de classification par apprentissage automatique supervisé atteint une exactitude de 90,62%, surpassant d'autres méthodes de pointe. La contribution se poursuit en se concentrant sur le traitement des valeurs manquantes, introduisant une approche novatrice d'imputation. Le modèle Random Forest, avec une exactitude de 92%, démontre l'efficacité de cette méthode dans la gestion des données incomplètes.

Les contributions suivantes exploitent le jeu de données "Early stage diabetes risk prediction dataset". Où dans la troisième contribution, une évaluation de sept techniques majeures de classification révèle que l'algorithme XGBoost surpasse ses homologues avec un score F1 remarquable de 94,74% et une exactitude de 96,15%. Pour la quatrième contribution, on aborde le défi crucial de la prédiction précoce du diabète en équilibrant les données, sélectionnant soigneusement les caractéristiques et appliquant neuf techniques d'apprentissage automatique supervisé. L'algorithme Extra Trees se distingue avec une exactitude exceptionnelle de 97,95%, surpassant significativement d'autres modèles référencés de la littérature mettant en évidence l'efficacité des approches développées dans la prédiction du diabète, avec des précisions remarquables.

La dernière contribution se base sur "l'ensemble de données des Centres de contrôle et de prévention des maladies U.S." pour améliorer la performance des modèles de prédiction basés sur des données d'enquête. L'approche d'apprentissage automatique présentée, en particulier avec le modèle Random Forest, démontre une performance exceptionnelle sur un ensemble de données de test plus vaste, anticipant la validité du système dans un contexte réel.

En conclusion, cette thèse propose une approche performante basée sur l'IA pour la prédiction précoce du diabète tout en exploitant différents jeux de données. Les résultats exceptionnels obtenus soulignent l'efficacité des modèles développés, ouvrant la voie à des interventions médicales plus ciblées et préventives, ainsi qu'à une amélioration significative des systèmes existants de prédiction du diabète.

Les Mots clés *: Prédiction du diabète, stade précoce, analyse de données, analyse prédictive, intelligence artificielle, apprentissage automatique supervisé.*

Abstract

This thesis explores the potential of artificial intelligence in early diabetes prediction, highlighting innovative contributions while utilizing various databases.

The first two contributions use the "Pima Indian Diabetes Database". An ensemble model based on the three best methods obtained from supervised machine learning classification achieves an accuracy of 90.62%, surpassing other state-of-the-art methods. The contribution continues by focusing on handling missing values, introducing an innovative imputation approach. The Random Forest model, with an accuracy of 92%, demonstrates the effectiveness of this method in managing incomplete data.

The subsequent contributions leverage the "Early Stage Diabetes Risk Prediction Dataset". In the third contribution, an evaluation of seven major classification techniques reveals that the XGBoost algorithm outperforms its counterparts with a remarkable F1 score of 94.74% and an accuracy of 96.15%. For the fourth contribution, the crucial challenge of early diabetes prediction is addressed by balancing data, carefully selecting features, and applying nine supervised machine learning techniques. The Extra Trees algorithm stands out with an exceptional accuracy of 97.95%, significantly surpassing other referenced models in the literature, highlighting the efficiency of the developed approaches in diabetes prediction, with remarkable precision.

The last contribution is based on the "U.S. Centers for Disease Control and Prevention dataset" to enhance the performance of prediction models based on survey data. The presented machine learning approach, especially with the Random Forest model, demonstrates outstanding performance on a larger test dataset, anticipating the system's validity in a real-world context.

In conclusion, this thesis proposes a performance-driven approach based on AI for early diabetes prediction while exploiting different datasets. The exceptional results obtained underscore the efficiency of the developed models, paving the way for more targeted and preventive medical interventions, as well as significant improvements to existing diabetes prediction systems.

Keywords: *Diabetes prediction, early stage, data analysis, predictive analysis, artificial intelligence, supervised machine learning.*

مـلـخص

تستكشف هذه الأطروحة إمكانيات الذكاء الاصطناعي في توقع مرض السكري في مراحله المبكرة، مسلطة الضوء على الإسهامات الجديدة مع استخدام مجموعة متنوعة من قواعد البيانات.

يعتمد الإسهام الأول و الثاني على "قاعدة بيانات الهنود بيما لمرض السكري". يحقق النموذج التجميعي الذي يستند إلى أفضل ثلاث طرق تم الحصول عليها من تصنيف التعلم الآلي المشرف دقة بنسبة 90.62٪، متفوقاً على طرق أخرى رائدة. ينتقل الإسهام للتركيز على التعامل مع القيم المفقودة من قاعدة البيانات، وتقديم نهج ابتكاري لتعويضها. يُظهر نموذج الغابات العشوائية، بدقة تبلغ 92٪، فعالية هذا النهج في إدارة البيانات غير الكاملة.

تستغل الإسهامات التالية "مجموعة بيانات توقع مخاطر مرض السكري في المرحلة المبكرة". الإسهام الثالث، يكشف تقييم لسبع تقنيات رئيسية للتصنيف أن خوارزمية XGBoost تتفوق على نظرائها بنتيجة ملحوظة من F1-score بنسبة 94.74٪ ودقة بنسبة 96.15٪. في الإسهام الرابع، يُعالج التحدي الحاسم لتوقع مرض السكري في المراحل المبكرة من خلال تحقيق توازن في البيانات واختيار مدروس للميزات، وتطبيق تسع تقنيات للتعلم الآلي المشرف. يتألق Extra Trees بدقة استثنائية تبلغ 97.95٪، متفوقا بشكل كبير على نماذج أخرى مشار إليها في الدراسات السابقة، مسلطة الضوء على كفاءة النهج المطورة في توقع مرض السكري بدقة ملحوظة.

يعتمد الإسهام الأخير على "مجموعة بيانات مراكز مراقبة ووقاية الأمراض الأمريكية" لتحسين أداء نماذج التوقعات القائمة على بيانات الاستطلاع. يُظهر النهج في التعلم الآلي المقدم، خاصةً مع نموذج الغابات العشوائية، أداءً استثنائيًا على مجموعة بيانات اختبار واسعة، متوقعًا صحة النظام في سياق العالم الحقيقي.

في الختام، تقترح هذه الرسالة نهجًا يعتمد على الأداء والذكاء الاصطناعي لتوقع مرض السكري في مراحله المبكرة مع استغلال مجموعات بيانات متنوعة. تؤكد النتائج الاستثنائية المحققة على فعالية النماذج المطورة، وتفتح الباب أمام تدخلات طبية أكثر توجيهًا ووقائية، فضلاً عن تحسينات كبيرة في أنظمة توقع مرض السكري الحالية.

<u>الكلمات المفتاحية</u>: *توقع مرض السكري، المرحلة المبكرة، تحليل البيانات، تحليل تنبؤي، الذكاء الاصطناعي، التعلم الآلي المشرف.*

Remerciments

*Tout d'abord, Alhamduli « **Allah** » le tout-puissant qui nous a guidé sur le droit chemin tout au long du travail et m'a inspiré les bons pas et les justes réflexes. Sans sa miséricorde, ce travail n'aurait pas abouti.*

*L'encadrement scientifique de ce travail a été assuré par **Pr. Mohamed Ridda LAOUAR**, professeur à la faculté des Sciences Exactes et des Sciences de la Nature et de la Vie, Université Tébessa qui a su guider et orienter ce travail. Je veux lui adresser tous mes remerciements pour la confiance qu'il m'a accordée en acceptant d'encadrer ce travail doctoral, pour sa disponibilité, ses encouragements, ses critiques objectives et ses conseils avisés tout au long de la préparation de cette thèse.*

*Je souhaite exprimer toute ma gratitude à mon co-directeur de thèse **Dr. Issam BENDIB** pour les encouragements qu'il m'a toujours prodigués.*

*Je tiens aussi à remercier **Monsieur Sean B. EOM**, Professeur à l'université de Southeast Missouri State pour les conseils et pour sa collaboration aux travaux de recherche.*

***Monsieur Abdellatif GAHMOUSSE**. aucun remerciement ne saurait exprimer mon respect et considération pour les orientations que vous avez consenties pour mes premiers pas, merci pour votre encouragement et gentillesse*

*Je souhaite exprimer ma sincère reconnaissance envers les membres du jury: **Pr. Abderrahim SIAM, Dr. Lakhdar LAIMECHE, Pr. Hakim BENDJENNA** qui ont accepté de lire et d'évaluer ce travail pour leur attention et l'intérêt porté à cette thèse et qui ont eu l'obligeance d'accepter de juger ce travail. Comme je présente mes sincères remerciements à tous **les membres de LAMIS** qui m'ont accueilli pendant cette thèse et qui m'ont permis de travailler dans une ambiance exceptionnelle.*

A nos parents

Pour l'enfance merveilleuse qu'ils nous ont offerte ainsi que pour leurs encouragements.

Pour leurs soutiens et leurs aides.

Avec tout notre amour.

A tous mes amis

Pour leurs bonnes humeurs, leurs gentillesses et pour tous nos fous rires partagés.

DEDICACE

À mes chers parents,

Que nulle dédicace ne puisse exprimer ce que je leur dois, pour leur bienveillance, leur affection et leur soutien…Trésors de bonté, de générosité et de tendresse, en témoignage de mon profond amour et ma grande reconnaissance « Que Dieu vous garde ».

À l'âme de ma chère grand-mère,

Tu restes à jamais dans mon cœur, source d'amour et d'inspiration infinie.

À mes tantes adorées,

Votre amour et votre soutien sont des cadeaux précieux que je chérirai toujours. Merci d'être des piliers solides dans ma vie.

À mes professeurs

Qui doivent voir dans ce travail la fierté d'un savoir bien acquis.

À tous mes amies

(Rimka, Marocco, Kouki, Chou…)

Pour leur aide et leur soutien moral durant l'élaboration du travail de fin d'études.

À toute ma Famille

À tous ceux dont l'oubli du nom n'est guère celui du cœur…

Table des matières

Chapitre 2 L'intelligence Artificielle

Chapitre 3 État de l'art sur la prédiction du diabète par l'IA

Chapitre 5 Propositions et évaluation (partie 2)

Liste de Figures

Liste de Tableaux

Introduction Générale

1. Introduction

À l'ère du numérique, la grande quantité de données cliniques disponibles a permis aux techniques d'apprentissage automatique de donner de bons résultats en matière de diagnostic et de pronostic médicaux. Avec les progrès des techniques d'apprentissage automatique et de l'intelligence artificielle, les chercheurs et les médecins ont commencé à adopter et à appliquer l'estimation du risque de maladie.

L'exploration de données a permis d'obtenir des connaissances à partir d'une grande quantité de données. Grâce à des techniques statistiques, mathématiques et d'intelligence artificielle, il nous permet d'examiner de vastes modèles et de comparer les différences dans de grands ensembles de données. Le processus d'exploration de données est utilisé pour identifier des modèles cachés dans le comportement des données ou pour prédire les tendances futures probables. L'exploration de données a été appliquée avec succès dans un certain nombre de domaines, notamment la médecine clinique, la recherche épidémiologique, la génétique et la protéomique.

2. Problématique

Le domaine de la santé est unique en son genre, avec des attentes élevées en matière de prévention et de gestion des maladies telles que le diabète qui nécessite des outils de prédiction précis pour aider à la prise de décision médicale. Actuellement, l'évaluation du risque de diabète repose souvent sur des méthodes traditionnelles, qui peuvent être limitées par leur subjectivité, leur manque de sensibilité, et leur incapacité à anticiper les fluctuations individuelles. Les avancées récentes dans le domaine de l'intelligence artificielle (IA) offrent des solutions prometteuses pour la prédiction du diabète, avec une précision accrue. Cependant, l'adoption de ces technologies dans le secteur de la santé demeure relativement lente par rapport à d'autres domaines, en raison des défis spécifiques associés à la confidentialité des données et à la sécurité des patients. Face à la prévalence croissante du diabète dans le monde et à ses répercussions significatives sur la santé publique, il est impératif de développer des outils de prédiction précis et fiables. L'I.A offre un potentiel considérable pour améliorer la détection précoce, la gestion et la prévention du diabète. Cependant, cette approche suscite des questions cruciales liées à la performance des modèles d'IA, à la disponibilité de données de qualité, à la généralisation des résultats, à la transparence des décisions.

Les principaux problèmes que nous essayons de résoudre notamment le prétraitement de données multidimensionnelles et hétérogènes, la conception de modèles d'IA avancés pour la prédiction du diabète, l'amélioration de la précision des modèles de prédiction, l'explicabilité des résultats, ainsi que l'évaluation comparative avec les méthodes traditionnelles de prédiction du diabète.

3. Objectif

Cette thèse s'attache à explorer de manière approfondie la conception et l'optimisation de modèles d'IA pour la prédiction du diabète à partir de données médicales diverses. En définitive, cette recherche vise à fournir une base solide pour l'application pratique de l'IA dans le domaine de la santé pour améliorer la prévention, la gestion et le traitement du diabète. En fin de compte, cette recherche vise à fournir des contributions significatives pour l'amélioration de la prévention, de la gestion et de la compréhension du diabète grâce à l'IA.

4. Motivations

Selon l'organisation mondiale de la santé (OMS) les maladies non transmissibles (MNT), également appelées maladies chroniques, sont des affections de longue durée. Elles évoluent en général lentement. Les quatre principales MNT sont les maladies cardiovasculaires (MCV), les cancers, les maladies respiratoires chroniques et le diabète. Les MNT sont la première cause de décès dans le monde. Elles entrainent des conséquences sociales et économiques désastreuses dans les pays en particulier dans les populations pauvres et vulnérables. Elles sont responsables d'absentéisme, d'incapacités et de décès prématurés qui entrainent une perte de productivité. Elles nécessitent une prise en charge coûteuse qui augmente les dépenses de santé. Elles constituent ainsi un obstacle au développement dans les pays à revenu faible ou intermédiaire.

L'OMS affirme que le diabète est une réelle menace sanitaire au niveau mondial, qui ne dépend pas du statut socio-économique et qui ne connaît pas de frontières. Les personnes vivant avec le diabète sont à risque de développer un certain nombre de complications graves et potentiellement mortelles. En Algérie et dans le monde la situation est tout à fait dramatique. En effet, le diabète touche environ 500 millions de personnes dans le monde. Il figure parmi les 10 principales causes de décès. Cependant, une identification et un traitement rapides et adéquats de l'affection peuvent contribuer à réduire la gravité et le coût des conséquences, ainsi que la mortalité.

Selon l'Atlas 2019 du diabète de la Fédération Internationale du Diabète (FID), une grossesse sur six est affectée par l'hyperglycémie. Autre source d'inquiétude, le pourcentage de personnes vivant avec un diabète non diagnostiqué (principalement de type 2) reste très élevé, dépassant actuellement 50 %. Ce chiffre souligne l'urgente nécessité de diagnostiquer les personnes atteintes de diabète sans le savoir et de fournir des soins appropriés à toutes les personnes atteintes de diabète le plus rapidement possible.

La plupart des gens ignorent qu'ils présentent des symptômes de diabète à un stade précoce. Elle augmente la soif et l'appétit et favorise les mictions fréquentes. Le diabète a la capacité d'altérer, voire de détruire le système de santé en raison de sa forte prévalence et de ses complications, qui peuvent affecter de nombreuses composantes de l'organisme. Chez l'homme, un certain nombre de facteurs de risque peuvent augmenter la probabilité de contracter le diabète.

Il est crucial de souligner que chez certaines personnes souffrant de diabète de type 2, la maladie peut débuter sans aucun symptôme apparent. C'est pourquoi il est primordial de réaliser des dépistages réguliers afin de détecter la maladie à un stade précoce. Cette situation nous encourage à automatiser cette tâche en utilisant un système de prédiction précoce de la maladie.

L'utilisation de la technologie dans le secteur médical est l'une des étapes les plus importantes à franchir. Le diabète et d'autres maladies humaines graves sont détectés à un stade précoce grâce à l'utilisation d'algorithmes d'apprentissage automatique.

5. Contributions

Le référentiel d'apprentissage automatique de l'UCI donne accès à la base de données sur le diabète des Indiens Pima, qui est utilisée dans les expériences 1 et 2.

- Contribution -1- : où six méthodes de classification par apprentissage automatique supervisé et un modèle hybride basé sur les trois meilleurs résultats sont utilisés pour détecter le diabète à un stade précoce. Tous les modèles sont évalués sur la base d'une série de mesures. Il ressort que le modèle d'ensemble, qui a obtenu une exactitude de 90,62 %, est plus performant que d'autres méthodes de pointe.

- Contribution -2- : Le traitement des valeurs manquantes dans les ensembles de données est essentiel pour garantir la précision des modèles d'apprentissage automatique. Une approche novatrice a été adoptée (le mélange de techniques d'imputation des valeurs manquantes), utilisant des valeurs anticipées basées sur un modèle de régression linéaire construit à partir des données non manquantes et pour les caractéristiques avec quelques enregistrements manquants, une imputation moyenne/médiane par colonne a été choisie. Six algorithmes d'apprentissage automatique supervisé ont été évalués pour prédire le diabète à partir de la base de données des Indiens Pima, avec un accent particulier sur le Random Forest. Grâce à une combinaison de techniques d'imputation, ce modèle a atteint une exactitude de 92%, surpassant les approches antérieures et soulignant l'efficacité de cette approche novatrice dans la gestion des valeurs manquantes.

L'ensemble de données pour la prédiction du risque de diabète à un stade précoce « Early stage diabetes risk prediction dataset » est utilisé pour la contribution -3- et -4- :

- Contribution -3- : Ma contribution significative en utilisant ce jeu de donnée réside dans l'évaluation et la comparaison de 7 techniques majeures de classification par apprentissage automatique pour la prédiction du diabète. Mon travail a permis de mettre en lumière les performances exceptionnelles de l'algorithme XGBoost, qui a affiché un score F1 remarquable de 94,74% dépassant ainsi ses homologues. Les résultats obtenus par tous les algorithmes ont été globalement positifs. Notamment, la comparaison du modèle XGBoost avec des études antérieures a révélé une

exactitude supérieure, atteignant 96,15 %, soulignant ainsi la pertinence et l'efficacité de notre approche dans le domaine de la prédiction du diabète.

- Contribution -4- : Ma contribution à cette étude repose sur la réponse à un défi crucial dans le domaine médical : la prédiction précoce du diabète à l'aide de modèles d'apprentissage automatique. En mettant l'accent sur l'identification des individus à risque. Nous avons approché ce problème en équilibrant les données, en sélectionnant soigneusement les caractéristiques pertinentes via deux algorithmes de sélection, et en appliquant neuf techniques d'apprentissage automatique supervisé, incluant des approches traditionnelles, ainsi que les algorithmes de bagging et de boosting. Les résultats ont été évalués avec des critères de performance variés tels que l'exactitude, le score F1, le rappel et la précision. Notamment, l'algorithme Extra Trees s'est distingué avec une exactitude exceptionnelle de 97,95 % dans la prédiction précoce du diabète, surpassant significativement d'autres algorithmes référencés dans la littérature. Ces résultats soulignent le potentiel remarquable de notre modèle pour anticiper le diabète, ouvrant ainsi la voie à des interventions médicales plus ciblées et préventives.

L'ensemble de données des Centres de contrôle et de prévention des maladies U.S. est utilisé dans la contribution 5 :

- Contribution -5- : Il existe plusieurs modèles de prédiction du diabète. Cependant, en raison de la complexité de la causalité du diabète, la performance de prédiction des modèles basés sur des données d'enquête doit être améliorée. En outre, si de nombreux facteurs de risque du diabète ont été découverts, tels que l'obésité et l'âge, d'autres restent à trouver. C'est pourquoi l'ensemble de données du « Behavioral Risk Factor Surveillance System » est utilisé. Les méthodes existantes de prédiction de la maladie diabétique utilisent un tout petit ensemble de données. L'objectif de notre système est d'opérer sur un plus grand ensemble de données afin d'améliorer l'efficacité globale du système. Dans ce travail, une approche d'apprentissage automatique pour prédire le diabète précoce a été présentée. Nous avons découvert que le Random_forest est exceptionnellement performant sur l'ensemble de données de test. Il a presque anticipé tous les exemples de l'ensemble de test correctement, démontrant ainsi leur validité dans un contexte réel.

6. L'organisation du manuscrit

Cette thèse est constituée d'une introduction générale suivie de cinq chapitres. Elle représente une avancée significative dans le domaine de la prédiction précoce du diabète en exploitant des méthodes innovantes basées sur l'intelligence artificielle.

Le chapitre initial présente les concepts liés au contexte médical du diabète, en abordant l'épidémiologie de cette maladie, sa définition ainsi qu'une exploration de ses divers types, causes, critères de diagnostic, symptômes, complications possibles et les facteurs de risque inhérents au diabète.

Le deuxième chapitre se focalise sur les bases de l'intelligence artificielle (IA) ainsi que sur ses subdivisions et leurs interactions. Il examine également les diverses catégories d'IA et leurs applications variées. En outre, nous plongerons en détail dans l'impact de l'IA dans le domaine médical, en mettant l'accent sur son utilisation pour améliorer les diagnostics et les traitements. Enfin, nous aborderons également l'importance des sources de données pertinentes dans le domaine de la santé et leur rôle essentiel dans la mise en œuvre de l'IA en médecine.

Au sein du troisième chapitre qui est un état de l'art en matière de prédiction du diabète par l'IA une analyse minutieuse des études pertinentes sera menée, avec une attention particulière portée à la comparaison des approches méthodologiques, des jeux de données mobilisés et des résultats obtenus.

Les quatrième et cinquième chapitres sont dédiés à nos propres apports, tout en mettant en avant cinq contributions de grande envergure. Chacune de ces contributions contribuera à enrichir la connaissance des meilleures méthodes de prédiction du diabète grâce à l'apprentissage automatique.

En conclusion, nous terminons ce manuscrit en abordant les travaux accomplis et en évoquant nos orientations futures.

Chapitre 1

Contexte Médical du Diabète

1. Introduction

Le diabète est « l'un des principaux tueurs au monde », avec l'hypertension artérielle et le tabagisme, selon l'Organisation Mondiale de la Santé (OMS). Cette maladie constitue un problème de santé publique majeur et malgré les efforts de prévention, la pandémie se poursuit.

Dans ce chapitre nous présentons tout d'abord l'épidémiologie du diabète puis nous donnons une définition du diabète, ses différents types, ses causes, ses critères de diagnostique. Ensuite, nous exposerons ses symptômes, ainsi que les différentes complications du diabète et aussi les facteurs de risque du diabète. Finalement un point sur le diabète et le COVID-19.

2. L'épidémiologie du diabète

En 2014, le diabète affectait 422 millions de personnes au niveau mondial, alors qu'il ne concernait que 108 millions de patients dans le monde en 1980 et que les premières prévisions de l'Organisation Mondiale de la Santé (OMS) et de l'International Diabetes Federation (IDF) s'inquiétaient en 1990 du risque de voir le diabète affecter 240 millions de personnes en 2025…

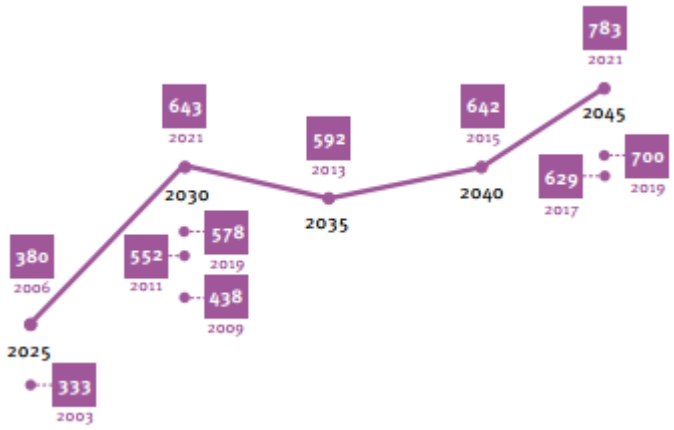

FIGURE 1.1 : PROJECTIONS DE PREVALENCE MONDIALE DU DIABETE DANS LE GROUPE D'AGE DE 20 A 79 ANS (MILLIONS) [1].

2.1. Prévalence du diabète sucré dans le monde

En 2019, le diabète affecte plus de 463 millions de personnes dans le monde, dont 59 millions en Europe. En 2021, le diabète affecte plus de 537 millions de personnes dans le monde (soit 1 personne sur 10), dont 61 millions en Europe. De plus, 6,7 millions de personnes sont décédées en 2021 en raison de leur diabète, soit une augmentation de 2,5 millions par rapport à 2019 (4,2 millions de décès) ! En 2021, 81 % des adultes diabétiques vivent dans des pays à revenu faible ou intermédiaire (contre 79 % en 2019).

Les prévisions actuelles de ces deux organismes sont très préoccupantes : ils annoncent 643 millions de patients diabétiques pour 2030 et 784 millions pour 2045 (source : Atlas 2021 de la International Diabetes Federation[1]).

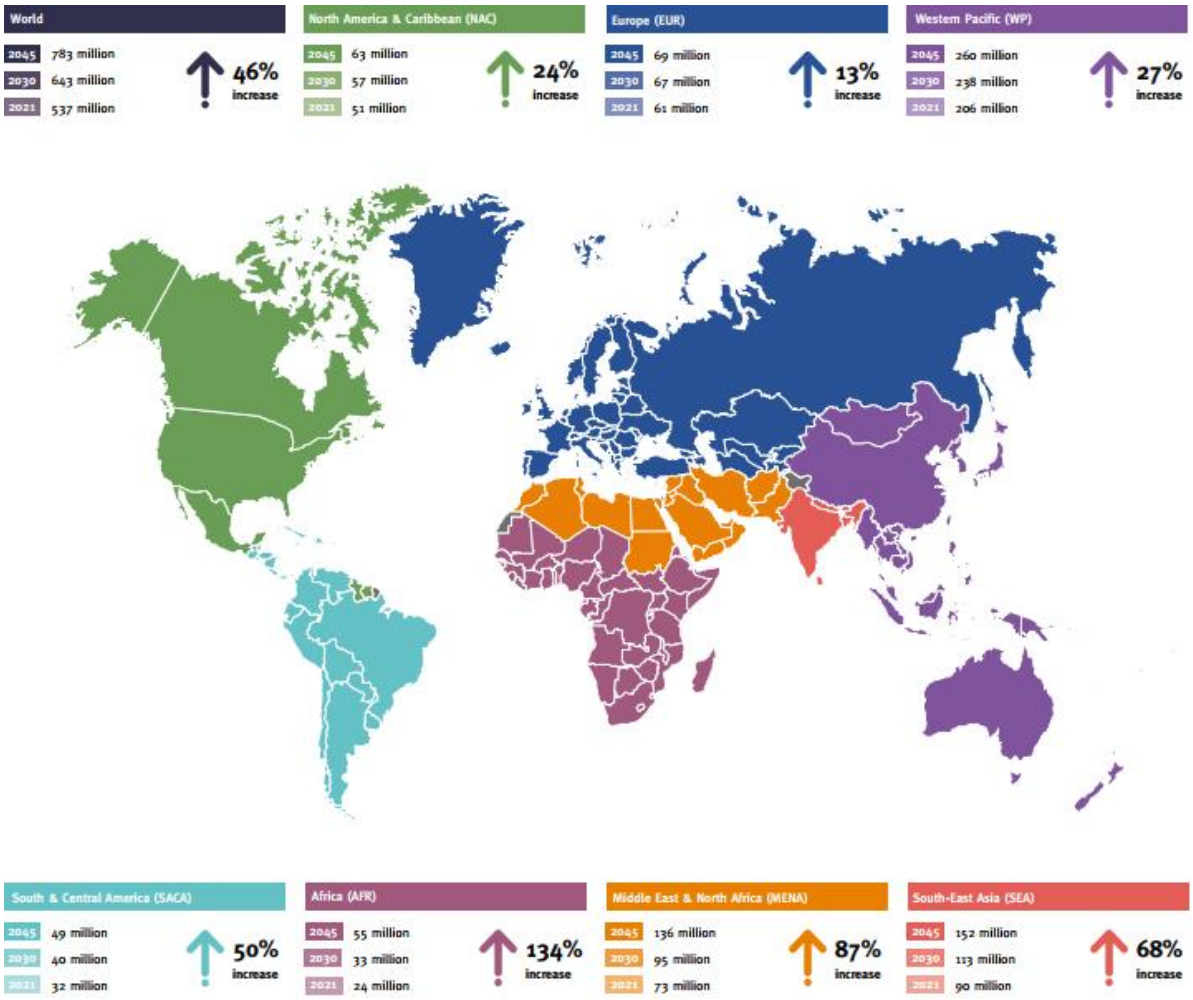

FIGURE 1.2 : NOMBRE DE PERSONNES ATTEINTES DE DIABETE DANS LE MONDE ET PAR REGION DE LA FID EN 2021-2045 (20-79 ANS) [1].

2.2. Prévalence du diabète sucré en Algérie

En Algérie, le diabète est la deuxième maladie chronique la plus courante après l'hypertension. En 2018, environ 14,4 % de la population âgée de 18 à 69 ans étaient estimés avoir le diabète, ce qui équivaut à environ 4 millions de personnes. Le diabète de type 2 est la forme la plus répandue en Algérie, représentant environ 90 % des cas diagnostiqués. Les facteurs de risque pour le diabète en Algérie comprennent une alimentation malsaine, un manque d'activité physique et une prédisposition génétique. Il est important de sensibiliser

[1] https://ceed-diabete.org/fr/le-diabete/les-chiffres/

au diabète et de gérer efficacement la maladie en Algérie, où une prévention et un traitement efficaces sont nécessaires pour améliorer la santé et le bien-être de la population [2].

3. Le diabète

Le diabète sucré, plus simplement appelé diabète, est une maladie grave, à long terme (ou «chronique») qui survient lorsque le taux de glycémie d'une personne est élevé parce que son organisme ne peut pas produire d'insuline, qu'il n'en produit pas suffisamment ou qu'il ne peut pas utiliser efficacement l'insuline qu'il produit [3]–[5].

L'insuline est une hormone essentielle sécrétée dans le pancréas. Elle permet au glucose de quitter la circulation sanguine et d'entrer dans les cellules de l'organisme, où il est converti en énergie. L'insuline est également essentielle au métabolisme des protéines et des graisses. Un manque d'insuline, ou l'incapacité des cellules à répondre à un manque d'insuline, entraîne des taux élevés de glucose sanguin (hyperglycémie), ce qui constitue un indicateur clinique du diabète. Si elle demeure non contrôlée de façon prolongée, l'hyperglycémie peut provoquer des lésions au niveau de divers organes et conduire au développement de complications de santé invalidantes, voire mortelles, telles que des maladies cardiovasculaires, une neuropathie, une néphropathie et des maladies oculaires pouvant déboucher sur une rétinopathie et la cécité [6]–[9]. En revanche, une gestion appropriée du diabète permettra de retarder ou de prévenir ces complications graves.

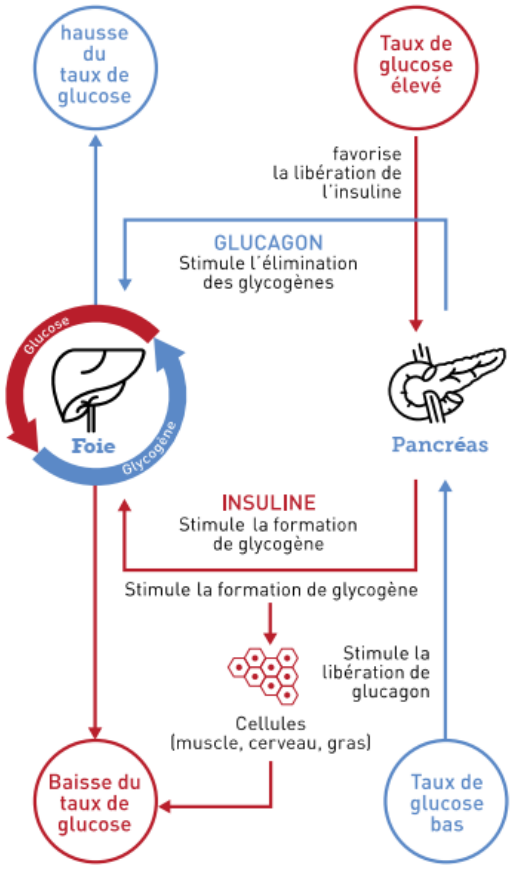

FIGURE 1.3 : REGULATION DE LA GLYCEMIE DANS LE CORPS HUMAIN [10].

4. Les types de diabète

La classification et le diagnostic du diabète sont complexes et ont fait l'objet de nombreux débats, consultations et révisions au fil des ans. Il est aujourd'hui largement reconnu que le diabète comporte trois types principaux : le diabète de type 1, le diabète de type 2, le diabète gestationnel (DG) [10].

4.1. Le diabète de type 1

Le diabète de type 1 est la principale cause de diabète chez les enfants, mais il peut survenir à tout âge. À ce jour, il ne peut pas être prévenu. Les personnes présentant un diabète de type 1 peuvent mener une vie saine et épanouissante, à la seule condition qu'elles bénéficient d'un apport interrompu d'insuline, qu'elles soient soutenues, et qu'elles aient accès à l'éducation sur le diabète ainsi qu'à un équipement de test de la glycémie [1].

4.2. Le diabète de type 2

Le diabète de type 2 constitue la forme de diabète la plus représentée (environ 90 % des cas) dans le monde. L'éducation sur le diabète, le soutien et un mode de vie sain, combinés à des médicaments le cas échéant, permettent de prendre en charge le diabète de type 2 de façon efficace. Des données probantes indiquent que le diabète de type 2 peut être prévenu et de plus en plus de données montrent que certaines personnes vivant avec le diabète de type 2 peuvent être en rémission [1], [10].

4.2.1. Différence entre diabète de type 1 et de type 2

Plusieurs caractéristiques nous permettent de distinguer le diabète de type 1 de celui de type 2, telle que la fréquence, l'âge, les causes, les signes révélateurs et autres qui sont regroupées dans le tableau ci-dessous :

TABLEAU 1.1 : DIFFERENCE ENTRE LES DIABETES DE TYPE 1 ET 2 [11].

Type de diabète	*Type 1*	*Type 2*
Fréquence	15%	85%
Age de début	<20 ans	>45 ans
Facteur héréditaire	Faible	Fort
Obésité	Non	Oui
Signes auto-immuns	Oui	Non
Insolino-sécrétion	Nulle	Carence relative
Insulino-résistance	Non	Oui

4.2.2. Le diabète sucré

Le diabète sucré est le même que le diabète. Le terme "diabète sucré" est souvent utilisé pour faire référence à la forme la plus courante de la maladie, le diabète de type 2. Cependant, le terme "diabète" est généralement utilisé pour englober toutes les formes de la maladie [2], [10], [11].

4.2.3. Le prédiabète

Le terme « prédiabète » est de plus en plus utilisé pour désigner les personnes présentant une intolérance au glucose et/ou une anomalie de la glycémie à jeun. Cela indique un risque de développement futur du diabète de type 2 et de complications liées au diabète [1].

4.3. Le diabète gestationnel

Le diabète gestationnel est la troisième forme principale de diabète qui survient lorsque les femmes enceintes développent des taux de glycémie élevés sans antécédents de diabète. Il est défini comme tout degré d'intolérance au glucose apparaissant ou reconnu pour la première fois pendant la grossesse [12].

4.4. Autres types spécifiques de diabète

Une grande variété d'affections relativement peu courantes sont classés comme « autres types particuliers ». Ces affections sont surtout des formes de diabète définies génétiquement ou associées à d'autres maladies ou à l'usage de médicaments [11], [12].

5. Les causes du diabète sucré

Plusieurs facteurs d'ordre génétique et environnemental concourent à l'apparition du diabète sucré, mais dont la nature diffère entre le type 1 et le type 2, d'où la nécessité de distinction entre ces deux derniers.

5.1. Les causes du diabète de type 1

Le diabète de type 1 est une affection auto-immune, c'est-à-dire que les cellules du pancréas qui fabrique l'insuline β sont progressivement détruites par le système immunitaire. Jusqu'à ce jour, les chercheurs ont cerné deux principaux facteurs qui expliquent cette affection : la génétique et l'environnement.

5.1.1. Le facteur génétique

L'existence d'un terrain génétique favorise l'apparition du diabète de type 1. Il y a une forte probabilité de développer un diabète de type 1 lorsque les parents sont eux même diabétiques. Cependant, il est rare que d'autres membres de la famille aient le diabète ; la situation se produit dans moins d'une famille sur deux.

5.1.2. Les facteurs environnementaux

Plusieurs facteurs externes contribuent au déclenchement du diabète de type 1, à savoir : l'infection virale ou bactérienne qui perturberait le système de reconnaissance qui protège nos organes de l'action destructrice de l'immunité, il y a aussi la nature de l'alimentation pendant la petite enfance (l'allaitement maternel semble réduire le risque de diabète chez l'enfant) ou le stress psychologique qui favorise le déclenchement d'un diabète de type 1. Enfin, certaines maladies touchant le pancréas (inflammation, kyste, cancer, etc.) peuvent indirectement provoquer un diabète de type 1.

5.2. Les causes du diabète de type 2

L'obésité est l'une des principales causes de la résistance à l'insuline. En outre, des facteurs génétiques entrent probablement en jeux dans l'apparition du diabète de type 2. Des chercheurs ont constaté que des antécédents familiaux de diabète augmentent le risque de survenue de cette affection.

D'autres facteurs de risque contribuent à l'apparition du diabète de type 2, entre autres :

- Age supérieur à 45 ans ;
- Avoir de forts antécédents familiaux ;
- Les descendances de famille ;
- Être en puberté : les changements des taux hormonaux pendant la puberté causent une insulinorésistance et une baisse de l'action de l'insuline ;
- Avoir le syndrome des ovaires poly kystique : il s'agit d'un trouble qui comporte de nombreux symptômes, dont l'absence de menstruation, une croissance des cheveux anormale et le gain de poids ;
- L'accouchement d'un bébé d'un poids élève ;
- Des antécédents d'un diabète gestationnel ;
- L'usage de certains médicaments ;
- Des désordres mentaux ;
- Un prédiabète ou une anomalie de la glycémie à jeun.

6. Critères de diagnostique

Les critères de diagnostic du diabète ont fait l'objet de débats et de mises à jour au fil des décennies, mais les critères actuels de l'Organisation mondiale de la santé (OMS) préconisent d'observer l'élévation des taux de glucose dans le sang pour diagnostiquer le diabète.

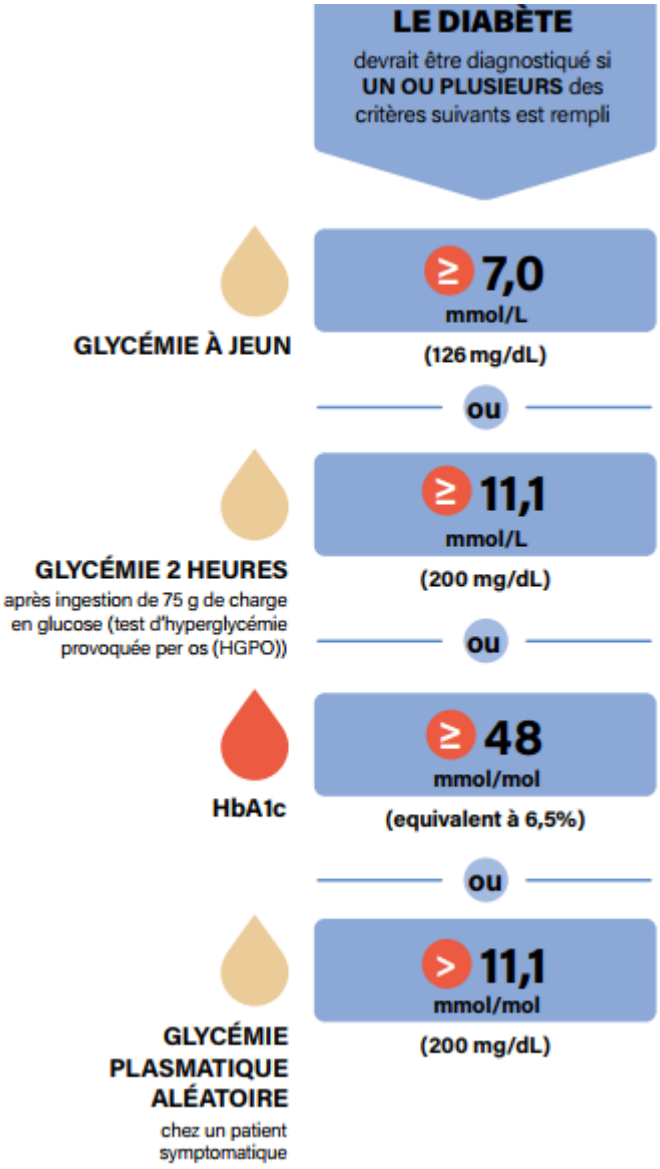

FIGURE 1.4 : CRITERES DE DIAGNOSTIC DU DIABETE [13].

Dans la (Figure 1.4) être à jeun se définit comme l'absence d'apport calorique pendant au moins 8 heures. Le test d'HbA1c doit être effectué en laboratoire à l'aide d'une méthode certifiée NGSP et standardisée au test DCCT (Diabetes Control and Complications Trial) [13].

7. Les symptômes de diabète

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

7.1. Les symptômes de type 1

- Augmentation de la soif
- Besoin fréquent d'uriner
- Fatigue

- Perte de poids malgré une augmentation de l'appétit
- Vision floue
- Blessures qui guérissent lentement
- Irritabilité
- Nausées et vomissements
- Humeur changeante
- Mictions nocturnes fréquentes

7.2. Les symptômes de type 2

- Augmentation de la soif
- Besoin fréquent d'uriner
- Fatigue
- Vision floue
- Blessures qui guérissent lentement
- Infections fréquentes
- Picotements ou engourdissements dans les mains ou les pieds
- Douleur ou sensation de brûlure dans les pieds, les jambes ou les mains
- Démangeaisons fréquentes dans la région génitale

Il est important de noter que certaines personnes atteintes de diabète de type 2 peuvent ne présenter aucun symptôme au début de la maladie, ce qui rend encore plus important le dépistage régulier pour détecter la maladie à un stade précoce. Cela nous motive pour informatiser cette tache en utilisant un system de prédiction précoce de la maladie [1], [13].

8. Les complications du diabète

Un diabète mal contrôlé peut endommager presque tous les organes du corps, notamment le cœur, les vaisseaux sanguins, les reins, les yeux, le système nerveux, etc. L'hyperglycémie endommage au fil du temps les parois des minuscules artères sanguines qui apportent l'oxygène et les nutriments à tous les tissus, ce qui affecte tous ces organes [11], [14].

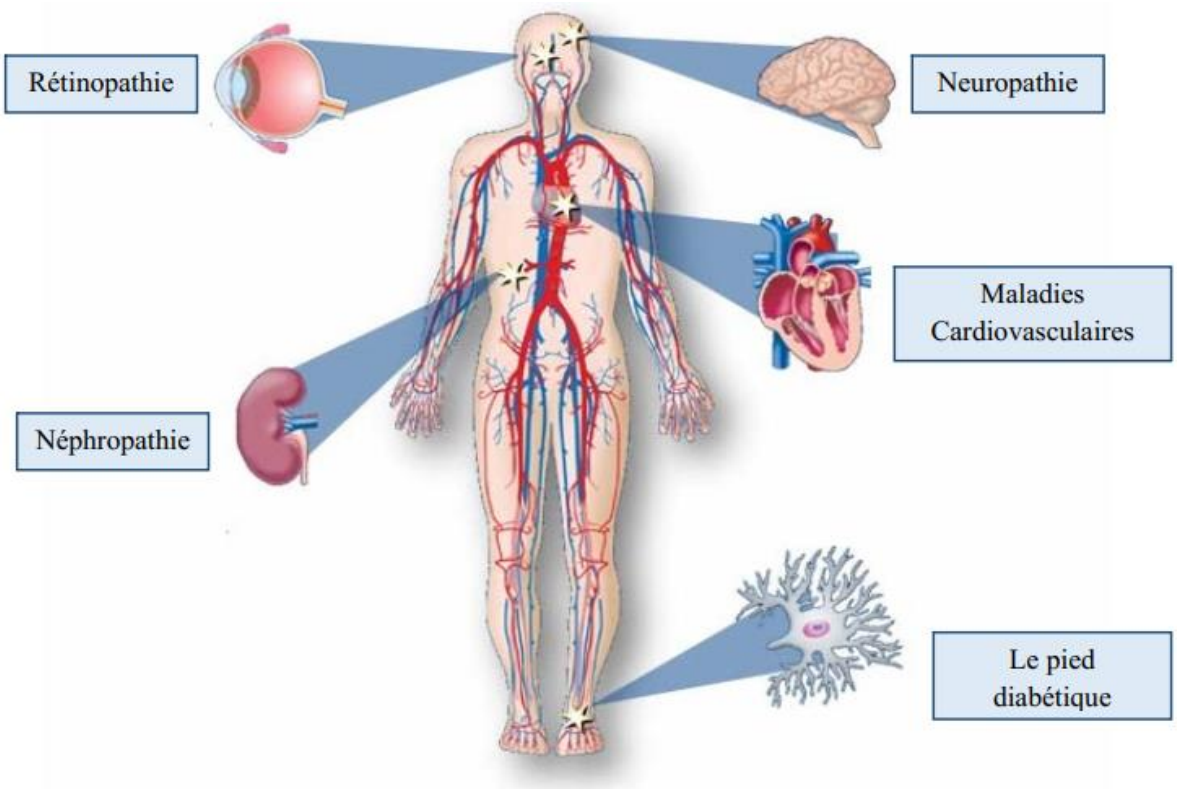

FIGURE 1.5 : LES PRINCIPALES COMPLICATIONS DE DIABETES.

8.1. Maladies cardiovasculaires

Le diabète joue un rôle dans le développement des maladies cardiovasculaires. Les diabétiques ont 2 à 4 fois plus de risques que la population générale de les développer. La coagulation du sang est favorisée par une teneur élevée en glucose dans le sang. Le risque d'obstruction des vaisseaux sanguins près du cœur (crise cardiaque) ou dans le cerveau (accident vasculaire cérébral) augmente avec le temps. Les risques sont également accrus par l'âge, la génétique, l'hypertension, l'obésité et le tabagisme. Les diabétiques de type 2 ont souvent un profil génétique qui les expose à un risque accru de la maladie. Un diabétique sur deux meurt d'une crise cardiaque ou d'un accident vasculaire cérébral.

8.2. Néphropathie

Le mot néphropathie est dérivé du mot grec nephros, qui signifie rein. Le tissu rénal est constitué d'un réseau de petites artères sanguines qui agissent comme un filtre, éliminant les toxines et les déchets de la circulation sanguine. Comme le diabète provoque des difficultés vasculaires, les petites artères des reins peuvent être compromises, ce qui entraîne une détérioration progressive des reins et diverses maladies allant de l'insuffisance rénale à l'insuffisance rénale permanente. Il convient de noter que l'hypertension joue un rôle dans la néphropathie.

8.3. Troubles oculaires

Le diabète peut entraîner une dégénérescence progressive de la vue. Il peut également entraîner l'apparition de cataractes et éventuellement la cécité. Les problèmes de vision sont la conséquence la plus répandue du diabète. Ils touchent presque toutes les personnes atteintes de diabète de type 1 et 60 % des diabétiques de type 2. La rétine est la zone de l'œil la plus souvent touchée, mais d'autres parties peuvent l'être également.

8.4. Neuropathie

La neuropathie est un terme utilisé pour décrire un groupe de troubles qui endommagent les nerfs et peuvent être extrêmement douloureux, quelle qu'en soit la cause. Elle se développe chez 40 à 50 % des personnes atteintes de diabète de type 1 ou de type 2 au cours des dix premières années de leur diabète. Une mauvaise circulation sanguine (et donc un apport insuffisant d'oxygène aux nerfs) et un taux de glucose élevé dégradent la structure des nerfs, ce qui entraîne la neuropathie. En général, la personne ressent des picotements, une perte de sensation et une gêne qui commence au bout des orteils ou des doigts et progresse vers le haut des membres affectés. Les nerfs qui régissent la digestion, la pression sanguine et le rythme cardiaque peuvent tous être affectés par la neuropathie.

8.5. Le pied diabétique

Le pied diabétique présente un risque plus élevé de problèmes en raison d'une perte de sensation (neuropathie), d'une mauvaise circulation sanguine (vascularisation) et d'un niveau élevé de glycémie, qui réduit la capacité de cicatrisation et la capacité du système immunitaire à combattre les infections.

8.6. Sensibilité aux infections

Les diabétiques sont plus sensibles aux infections récurrentes, qui peuvent être difficiles à traiter, en raison du pic de glycémie et de l'épuisement produits par la maladie. Les infections de la peau, de la bouche et des voies respiratoires en sont des exemples. Le diabète peut également entraver le processus de guérison, entraînant des infections persistantes des plaies. La maladie la plus fréquente est l'infection du pied. Elles peuvent s'accompagner d'ulcères et nécessiter l'amputation du pied en cas de gangrène, ce qui est en partie imputable à la neuropathie.

9. Les facteurs de risque du diabète

Plusieurs facteurs de risque peuvent contribuer au développement de cette maladie. Les facteurs de risque du diabète peuvent être divisés en deux catégories [14] :

9.1. Les facteurs de risque non modifiables

- L'âge : le risque de développer le diabète augmente avec l'âge.
- L'hérédité : le risque de diabète est plus élevé chez les personnes ayant des antécédents familiaux de la maladie.
- Le groupe ethnique : les personnes d'origine africaine, asiatique, hispanique ou amérindienne sont plus susceptibles de développer le diabète.

9.2. Les facteurs de risque modifiables

- L'obésité : l'excès de poids augmente le risque de diabète de type 2.
- L'inactivité physique : le manque d'activité physique peut augmenter le risque de diabète.
- L'alimentation : une alimentation riche en graisses et en sucres peut augmenter le risque de diabète de type 2.
- Le tabagisme : les fumeurs ont un risque accru de développer le diabète.
- L'hypertension artérielle : une pression artérielle élevée peut augmenter le risque de diabète.

9.3. Le diabète et le Covid-19

La détection ou la prédiction précoce du diabète peut être utile pour lutter contre le COVID-19 en permettant l'identification de personnes à risque de développer une forme grave de la maladie. Les personnes souffrant de diabète ont plus de chances de présenter des complications sévères lorsqu'elles contractent le COVID-19. Par conséquent, en détectant le diabète à un stade précoce chez les personnes présentant des facteurs de risque, il est possible de prendre des mesures préventives pour réduire leur risque de complications graves en cas d'infection.

De plus, la détection précoce du diabète peut également aider à prévenir l'apparition de la maladie chez les personnes présentant un risque élevé. Des recherches ont prouvé que des changements de mode de vie tels que l'adoption d'une alimentation saine, la pratique régulière d'exercice physique et la perte de poids peuvent contribuer à prévenir ou à retarder l'apparition du diabète de type 2 chez les personnes à risque [15], [16].

En identifiant les personnes à risque élevé de diabète et en leur fournissant des conseils sur les changements de mode de vie à apporter, il est possible de réduire leur risque de développer le diabète de type 2 et par conséquent de minimiser leur risque de complications liées au COVID-19. En somme, la détection précoce du diabète peut aider efficacement à lutter contre le COVID-19 en permettant une identification rapide et une prise en charge adéquate des personnes présentant un risque élevé [3], [17].

10. Conclusion

Dans ce chapitre ont été abordées des notions de généralité en rapport avec le diabète. Après avoir présenté de façon concise l'épidémiologie du diabète, sa définition, en passant par ces types ainsi que quelques causes, critères de diagnostique, symptômes et complications. L'objectif qui sous-tend la présentation de ces éléments, c'est de mettre en lumière l'importance de lutter contre cette maladie. Notre travail va s'orienter vers la prédiction du diabète en utilisant l'I.A qui est le sujet de notre suivant chapitre.

Chapitre 2

L'intelligence artificielle

1. Introduction

Les origines de l'intelligence artificielle (IA) remontent aux années 1950, lorsque les chercheurs se sont lancés dans la quête de machines capables d'imiter la cognition humaine. Depuis lors, l'IA a connu de multiples phases de progression et de régression, au gré des percées technologiques, des changements de paradigmes de recherche et de l'évolution des attentes de la société. Ces époques englobent les premiers systèmes experts et basés sur des règles, l'avènement des algorithmes d'apprentissage automatique et des réseaux neuronaux dans les années 1980 et 1990, l'hiver de l'IA au début des années 2000 et la résurgence actuelle de l'IA catalysée par l'apprentissage profond, le big data et l'informatique en nuage (cloud computing). Actuellement, l'IA est en train de remodeler diverses industries, notamment les soins de santé, la finance, la fabrication et les transports, avec la promesse de transformer la société d'une manière qui était autrefois considérée comme de la simple science-fiction.

Ce chapitre se concentre sur les fondements de l'intelligence artificielle (IA), ses sous-domaines et leur relation entre eux, ainsi que sur les différents types d'IA et leurs applications. En outre, nous explorerons spécifiquement l'impact de l'IA dans le domaine de la médecine, en mettant l'accent sur son utilisation pour améliorer les diagnostics et les traitements. Enfin, nous aborderons également les sources de données pertinentes dans le domaine de la santé et leur rôle crucial dans la mise en œuvre de l'IA en médecine.

2. L'intelligence artificielle « IA »

L'intelligence artificielle (IA) est un domaine interdisciplinaire qui s'efforce de concevoir des machines intelligentes capables d'émuler ou de dépasser les capacités cognitives humaines. Le spectre de l'IA englobe diverses applications telles que le traitement du langage naturel, la reconnaissance vocale, la prise de décision, l'apprentissage automatique, la vision par ordinateur et la robotique [18]. Les systèmes d'IA sont classés en trois catégories : l'IA faible, l'IA forte et l'IA super-intelligente. L'IA faible fonctionne dans le cadre de tâches et de limites spécifiques, tandis que l'IA forte peut raisonner et comprendre à l'égal de la cognition humaine. L'IA super-intelligente reste un concept théorique qui surpasse l'intelligence humaine par des ordres de grandeur. L'impact de l'IA s'étend à de nombreux domaines, tels que les soins de santé, la finance, la fabrication, la sécurité et les transports, ce qui entraîne de nouvelles avancées en matière de technologie et d'innovation [19].

La résurgence de l'intelligence artificielle (IA) peut être attribuée à plusieurs facteurs, notamment l'émergence de nouveaux algorithmes et la disponibilité accrue des ressources informatiques, ainsi que l'accessibilité généralisée des données provenant de divers aspects de la vie quotidienne, y compris les soins de santé [20]. L'efficacité récente de l'IA est largement attribuée à l'utilisation d'algorithmes adaptatifs qui peuvent établir des règles de prise de décision basées sur des données observées, plutôt que sur des règles prédéterminées introduites par les humains. Cette adaptabilité, bien que prometteuse, représente également un défi pour les professionnels de la santé qui peuvent hésiter à adopter cette approche [21], [22].

TABLEAU 2.1 : LES SOUS-DOMAINES DE L'INTELLIGENCE ARTIFICIELLE.

Sous-champ	Description
Apprentissage automatique « ML »	Utilisation d'algorithmes et de modèles statistiques pour permettre aux machines d'apprendre et de faire des prédictions.
Apprentissage profond « DL »	Sous-domaine de l'apprentissage automatique qui utilise des réseaux neuronaux artificiels pour apprendre et résoudre des tâches complexes.
Traitement du langage naturel (Natural Language Processing « NLP »)	Étude de la manière dont les ordinateurs peuvent comprendre, interpréter et générer du langage humain.
La fouille de données « Data Mining »	Un processus de découverte de nouveaux motifs, relations et tendances à partir de données pour générer des connaissances. Cela implique la collecte, la sélection, la préparation et l'analyse des données en utilisant des méthodes d'apprentissage automatique et des statistiques. Les résultats de la fouille de données peuvent être descriptifs ou prédictifs, représentant les motifs et relations dans les données ou les tendances et conditions futures.
Robotique	La conception, la construction et l'exploitation de robots pour accomplir des tâches de manière autonome ou avec l'aide de l'homme.
Vision par ordinateur (Computer Vision « CV »)	Capacité des machines à interpréter et à comprendre les informations visuelles du monde.

Les modèles d'apprentissage automatique sont capables d'analyser de multiples facteurs de risque, notamment l'âge, le sexe, les antécédents familiaux, les habitudes de vie et les informations cliniques, afin d'élaborer des modèles prédictifs de risque de diabète personnalisés. Ces derniers peuvent guider les prestataires de soins de santé dans l'élaboration de stratégies de prévention personnalisées pour les patients à haut risque, telles que les interventions diététiques et l'administration de médicaments. En outre, l'IA peut aider les cliniciens à établir un diagnostic précis, un pronostic et à choisir un traitement en fonction de paramètres spécifiques au patient [23], [24].

Dans l'ensemble, l'IA a le potentiel de transformer les soins du diabète en fournissant aux patients et aux cliniciens des informations individualisées et fondées sur des données. Bien que des recherches supplémentaires soient nécessaires pour authentifier la précision et

l'efficacité des outils de prédiction et de gestion du diabète basés sur l'IA, les perspectives d'amélioration des résultats et de réduction des dépenses de santé en font un domaine encourageant à explorer davantage [25], [26].

2.1. Les types d'I.A

En général, il existe trois types d'intelligence artificielle [27], [28]:

2.1.1. Intelligence Artificielle Étroite/Faible

Elle est conçue pour exécuter une tâche spécifique, telle que la reconnaissance de la parole ou le jeu d'échecs. Cette dernière désigne les systèmes d'IA avec des capacités limitées conçus pour résoudre un seul problème ou exécuter une seule tâche de manière efficace. Elle opère dans un domaine particulier et ne peut pas fonctionner au-delà de son champ d'application.

2.1.2. Intelligence Artificielle Générale

Elle est destinée à raisonner et à penser comme les humains, gérer diverses tâches et opérer dans différents domaines tels que le traitement du langage, le traitement de l'image et la raison. Cependant, les chercheurs travaillent toujours sur le développement de ce type d'IA.

2.1.3. Intelligence Artificielle Supérieure/Forte

Elle fait référence à une forme hypothétique d'IA qui est plus intelligente que les humains y compris la prise de décision et l'intelligence émotionnelle, et peut accomplir des tâches telles que la création d'art meilleur ou la construction de relations émotionnelles. Elle peut résoudre des problèmes complexes et prendre des décisions sans intervention humaine. Son développement est encore un sujet de discussion parmi les experts et les chercheurs, et elle n'est pas encore disponible.

Actuellement, nous travaillons avec l'Intelligence Artificielle Étroite et l'Intelligence Artificielle Générale. On prédit que la future Intelligence Artificielle, c'est-à-dire l'Intelligence Artificielle forte, surpassera l'intelligence humaine [28].

3. L'apprentissage automatique « Machine Learning »

« L'apprentissage automatique est la discipline donnant aux ordinateurs la capacité d'apprendre sans qu'ils soient explicitement programmés. » Arthur Samuel, 1959

L'apprentissage automatique (machine learning « ML ») implique l'utilisation d'attributs distinctifs dans le but de discerner des modèles qui peuvent ensuite être utilisés pour l'analyse d'une situation donnée. Par la suite, la machine est capable d'assimiler et de mettre en œuvre ces connaissances acquises pour des scénarios comparables à venir [29], [30].

Plusieurs domaines ont bénéficié de l'application réussie de techniques basées sur l'apprentissage automatique. Ces domaines comprennent, mais ne se limitent pas à la reconnaissance de motifs, la vision par ordinateur, l'ingénierie spatiale, la finance, le

divertissement, la biologie computationnelle et les applications biomédicales et médicales [31].

Cet outil de pronostic et de prédiction peut être mis en œuvre de manière dynamique dans la prise de décision clinique afin de personnaliser les soins aux patients, plutôt que d'adhérer à un algorithme statique [32].

De manière générale, tous les problèmes d'apprentissage automatique peuvent être classés en deux grandes catégories distinctes :

3.1. Les étapes de ML

Le processus d'apprentissage automatique peut être décomposé en 7 étapes[2], comme illustré dans la (Figure 2.1) Afin d'illustrer l'importance et la fonction de chaque étape, nous utiliserons un exemple d'un modèle simple.

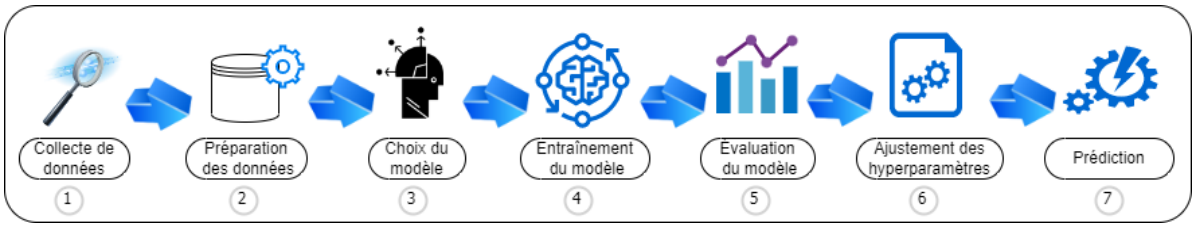

FIGURE 2.1 : ÉTAPES D'APPRENTISSAGE AUTOMATIQUE.

3.1.1. Étape 1 : Collecte de données / Rassemblement de données

Cette phase initiale revêt une importance considérable, car l'efficacité et la performance du modèle sont étroitement liées à la qualité et à l'abondance des données accumulées. Le socle fondamental du parcours de l'apprentissage automatique réside dans l'acquisition méticuleuse de données pertinentes, qui servent de base à l'entraînement du modèle. Une prudence particulière doit être exercée lors de cette étape afin d'éviter des erreurs critiques, telles que la sélection de caractéristiques inappropriées ou la restriction excessive des entrées de l'ensemble de données, car de telles erreurs ont le potentiel de rendre le modèle totalement inefficace.

3.1.2. Étape 2 : Préparation des données

Après la collecte des données d'entraînement, nous passons à la phase suivante de l'apprentissage automatique : la préparation des données, où les données sont chargées dans un référentiel approprié, puis préparées pour être utilisées dans l'entraînement de l'apprentissage automatique. Ici, les données sont toutes d'abord

[2] https://medium.com/dataseries/7-steps-to-machine-learning-how-to-prepare-for-an-automated-future-78c7918cb35d (access: 20/08/2023)

regroupées, puis l'ordre est aléatoire car l'ordre des données ne devrait pas affecter ce qui est appris.

Dans cette étape, nous traitons les données collectées à l'étape 1 et les préparons pour l'entraînement. Nous pouvons nettoyer les données en supprimant les doublons, corriger les erreurs, traiter les valeurs manquantes, effectuer des conversions de types de données, et ainsi de suite. Nous pouvons également visualiser les données, car cela nous aidera à identifier d'éventuelles relations pertinentes entre les différentes caractéristiques, comment en tirer parti, et aussi à montrer s'il existe des déséquilibres de données.

Un autre élément majeur de la préparation des données consiste à diviser les ensembles de données en 2 parties. La plus grande partie (~80%) sera utilisée pour entraîner le modèle, tandis que la plus petite partie (-20%) sera utilisée pour évaluer les performances du modèle entraîné. Cela est important car l'utilisation des mêmes ensembles de données pour l'entraînement et l'évaluation ne donnerait pas une évaluation équitable des performances du modèle dans des scénarios réels.

3.1.3. Étape 3 : Choix du modèle

Une fois les étapes centrées sur les données terminées, notre prochaine démarche consiste à sélectionner délibérément un modèle approprié. Dans le domaine de l'apprentissage automatique, une pléthore de modèles minutieusement développés par des scientifiques des données compétents attendent notre considération, chacun étant adapté pour répondre à des objectifs distincts. Diverses classes de modèles excellent dans la capture des schémas complexes inhérents à des ensembles de données variés, assurant des performances optimales du modèle en accord avec la nature intrinsèque des données. Il est impératif de reconnaître que différents modèles sont spécifiquement conçus pour répondre à des domaines spécifiques ; par exemple, certains modèles se révèlent particulièrement doués pour l'analyse de texte, tandis que d'autres se vantent d'une expertise supérieure en reconnaissance et interprétation d'images.

3.1.4. Étape 4 : Entraînement du modèle

L'essence même de l'entreprise d'apprentissage automatique réside dans le domaine de l'entraînement du modèle. Cette étape capitale englobe la majeure partie du processus d'apprentissage, exigeant une combinaison de patience, d'expérimentation inlassable et d'une compréhension profonde du domaine cible où le modèle sera déployé.

La phase d'entraînement promet des récompenses substantielles lorsque le modèle commence à manifester sa compétence dans son rôle désigné. Le processus d'entraînement implique l'initialisation de valeurs aléatoires pour les paramètres clés du modèle, tels que X et Y. Par la suite, des prédictions sont générées à l'aide de ces valeurs, qui sont ensuite comparées à la sortie attendue du modèle. Des ajustements sont effectués de manière itérative afin d'aligner les valeurs sur les prédictions souhaitées établies lors des itérations précédentes.

Ce processus cyclique de mise à jour et de perfectionnement constitue une étape d'entraînement, semblable à la progression d'un enfant dans la maîtrise de l'art de faire du vélo. Analogiquement aux premières rencontres de l'enfant avec des chutes et des hésitations, le modèle peut rencontrer des revers dans ses premières étapes d'entraînement. Cependant, à chaque étape d'entraînement, le modèle développe progressivement une compréhension accrue et une adaptabilité, semblable à la capacité améliorée de l'enfant à réagir avec aisance aux situations diverses rencontrées lors de la navigation sur le terrain du vélo.

3.1.5. Étape 5 : Évaluation du modèle

Après l'entraînement du modèle, il est impératif de le soumettre à une évaluation rigoureuse afin d'évaluer son efficacité dans des scénarios réels. Pour ce faire, le sous-ensemble d'évaluation de l'ensemble de données, spécifiquement désigné à cet effet, est utilisé pour scruter les performances du modèle. Cette phase critique expose le modèle à des situations inédites qui n'ont pas été rencontrées lors de sa phase d'entraînement.

L'évaluation revêt une importance capitale. Elle permet aux scientifiques des données de déterminer si les objectifs prédéfinis ont été atteints avec succès. Des résultats insatisfaisants nécessitent une réexamination des étapes précédentes afin de mettre en évidence les causes profondes sous-tendant les performances médiocres du modèle, suivies de mesures correctives. Des procédures d'évaluation insuffisantes ou défectueuses risquent d'entraver la capacité du modèle à remplir son objectif prévu.

3.1.6. Étape 6 : Ajustement des hyperparamètres

Une fois la phase d'évaluation terminée, le potentiel d'amélioration du processus d'entraînement émerge par le biais du réglage fin des hyperparamètres. Les paramètres implicitement supposés utilisés lors de la phase d'entraînement sont examinés, ainsi que l'inclusion de facteurs critiques tels que le taux d'apprentissage, qui régit l'amplitude du déplacement de la ligne lors de chaque étape itérative, guidée par les informations recueillies lors des étapes d'entraînement précédentes. La sélection et l'ajustement de ces valeurs ont un impact significatif tant sur l'exactitude du modèle d'entraînement que sur la durée du processus d'entraînement.

3.1.7. Étape 7 : Prédiction

La phase finale du processus d'apprentissage automatique consiste en le pronostic des résultats (prédiction), marquant ainsi le stade où le modèle est considéré prêt pour un déploiement pratique.

Ce moment crucial marque le sommet de la valeur intrinsèque de l'apprentissage automatique, car c'est ici que nous exploitons les compétences de notre modèle pour anticiper et envisager les résultats souhaités.

3.2. L'apprentissage supervisé

On utilise un ensemble de données connu appelé ensemble de données d'apprentissage pour effectuer des prévisions avec un algorithme d'apprentissage automatique supervisé.

L'apprentissage supervisé offre aux utilisateurs la possibilité de définir les classes désirées en fonction des besoins spécifiques de l'application. Cela implique l'utilisation d'une base de données d'apprentissage où chaque donnée est manuellement étiquetée avec sa classe correspondante [33]. Cependant, pour obtenir un classifieur performant, il est nécessaire d'étiqueter un nombre significatif de données pour chaque classe, ce qui peut être coûteux en termes de ressources humaines et de temps [34], [35].

L'objectif de l'apprentissage supervisé est de générer des règles de comportement à partir d'une base de données comprenant des exemples de cas préalablement étiquetés. Cette base de données consiste généralement en un ensemble de couples entrée/sortie {(X, Y)}. L'objectif est d'apprendre à prédire la sortie Y pour toute nouvelle entrée X [36], [37].

En ce qui suit les catégories d'apprentissage supervisé, avantages et inconvénients de l'apprentissage automatique supervisé, et ses applications.

3.2.1. Régression

La régression est une technique d'apprentissage supervisé qui prédit des réponses continues ou des valeurs numériques. Les modèles de régression sont utilisés pour estimer des variables continues telles que les salaires, les prix et les poids, ce qui en fait un outil crucial pour l'apprentissage automatique et une technique statistique largement utilisée. La méthode utilise des relations apprises entre les caractéristiques des données et les valeurs continues observées pour faire des prédictions [38]. Il existe plusieurs algorithmes de régression bien connus, tels que l'algorithme de régression linéaire simple, l'algorithme de régression multivariée, l'algorithme de l'arbre de décision, l'algorithme de régression Lasso, et d'autres [37].

3.2.2. Classification

Les algorithmes de classification sont utilisés pour résoudre des problèmes dans lesquels la variable de sortie est catégorique, telle que "Oui" ou "Non", "Homme" ou "Femme", "Rouge" ou "Bleu", etc. Ces algorithmes prédisent les catégories présentes dans un ensemble de données. Des exemples d'applications concrètes d'algorithmes de classification comprennent la détection de spam, le filtrage des e-mails, et d'autres. Les algorithmes de classification populaires incluent l'algorithme de forêt aléatoire, l'algorithme d'arbre de décision, l'algorithme de régression logistique, l'algorithme de machine à vecteurs de support, entre autres [39]–[41].

3.2.3. Avantages et inconvénients de l'apprentissage automatique supervisé

TABLEAU 2.2: LES AVANTAGES ET INCONVÉNIENTS DE LA ML SUPERVISÉ.

Avantages	Inconvénients
Donne des prédictions précises	Ne peut pas résoudre des tâches complexes
Utile pour la classification	Peut prédire une sortie incorrecte si les données de test sont différentes des données d'entraînement
Peut traiter des données de grande taille	Nécessite beaucoup de temps de calcul pour entraîner l'algorithme
Permet de comprendre clairement la relation entre les variables d'entrée et de sortie.	Peut suradapter «overfit» les données, ce qui entraîne une mauvaise généralisation.
Permet la création de modèles capables d'apprendre et de s'améliorer au fil du temps avec de nouvelles données.	Nécessite une grande quantité de données étiquetées pour être efficace.

3.2.4. Applications de l'apprentissage supervisé

L'apprentissage supervisé est une technique couramment utilisée dans divers domaines pour résoudre des problèmes de prédiction et de classification. Parmi les domaines d'application, on peut citer la médecine, la finance, la sécurité et la reconnaissance de la parole [42]–[44].

- En médecine, l'apprentissage supervisé est utilisé pour le diagnostic de maladies et la prédiction de résultats de traitements.
- Dans le domaine de la finance, l'apprentissage supervisé peut être utilisé pour la détection de fraudes et la prédiction des tendances du marché.
- La reconnaissance vocale utilise également l'apprentissage supervisé pour améliorer la précision de la transcription de la parole.
- Enfin, la segmentation d'image est une autre application courante de l'apprentissage supervisé, où l'algorithme est utilisé pour diviser une image en différentes parties en fonction des caractéristiques identifiées.

3.3. L'apprentissage non supervisé

L'apprentissage non supervisé consiste à explorer des problèmes sans préjuger des résultats attendus. Cette approche permet de découvrir des structures à partir des données, même en l'absence de connaissance sur l'effet des variables utilisées [45]. Donc dans l'apprentissage non supervisé, l'algorithme doit apprendre à partir de données brutes sans étiquettes ni annotations fournies par un expert. L'objectif est de découvrir les motifs ou la structure cachée des données et de regrouper des exemples similaires en clusters [33]. Il existe diverses techniques d'apprentissage non supervisé, telles que le clustering (qui

regroupe des données similaires), la réduction de dimensionnalité (qui diminue le nombre de variables), la détection d'anomalies (qui identifie des données aberrantes) et l'association de règles (qui met en évidence des relations entre les variables) [46].

3.3.1. Avantages et inconvénients de l'apprentissage automatique non-supervisé

TABLEAU 2.3: LES AVANTAGES ET INCONVÉNIENTS DE LA ML NON-SUPERVISÉ.

Avantages	Inconvénients
Peut découvrir des modèles inattendus et des relations cachées dans les données	Peut être difficile de comprendre et d'interpréter les résultats
Nécessite peu ou pas d'étiquetage manuel des données, ce qui permet de gagner du temps et de l'argent	Peut être difficile de déterminer le nombre optimal de groupes ou de clusters à utiliser
Peut être utilisé pour la segmentation de marché, l'analyse des réseaux sociaux, la détection d'anomalies et la recommandation de produits	Peut avoir du mal à gérer les données de grande taille ou de haute dimension
Peut être utilisé pour la réduction de dimension, permettant de visualiser et d'analyser des données complexes	Les résultats peuvent être biaisés en raison de la sélection aléatoire des données
Peut être utilisé pour prétraiter les données avant l'apprentissage supervisé	Les algorithmes non supervisés ne peuvent pas évaluer la qualité des résultats de la même manière que les algorithmes supervisés.

3.3.2. Applications de l'apprentissage non-supervisé

Voici quelques exemples d'applications de l'apprentissage non supervisé :

- **Clustering :** regroupement de données similaires pour identifier des motifs ou des structures. Exemples : segmentation de clients, regroupement de documents, analyse d'expression génétique
- **Réduction de la dimensionnalité :** réduction du nombre de variables ou de fonctionnalités dans un ensemble de données tout en conservant des informations importantes. Exemples : compression d'images, systèmes de recommandation, détection d'anomalies
- **Détection d'anomalies :** identification d'événements ou de points de données rares ou inhabituels. Exemples : détection de fraudes, détection d'intrusions, détection de flambées de maladies

- **Modèles génératifs :** création de nouveaux points de données similaires aux données d'entraînement. Exemples : synthèse d'images et de discours, traduction de langues, augmentation de données
- **Apprentissage de règles d'association :** découverte de relations ou de dépendances entre les variables d'un ensemble de données. Exemples : analyse du panier d'achat, systèmes de recommandation, analyse du comportement des clients

L'apprentissage non supervisé est particulièrement utile dans l'analyse exploratoire des données et pour découvrir des motifs ou des idées cachées dans les données lorsque le résultat ou l'étiquette est inconnu [47]–[50].

3.4. Differences majeurs entre apprentissage supervisé et non-supervisé

TABLEAU 2.4 : DIFFÉRENCES ENTRE L'APPRENTISSAGE SUPERVISÉ ET NON SUPERVISÉ.

	Apprentissage supervisé	Apprentissage non supervisé
Définition	Algorithme qui utilise un ensemble de données étiquetées pour faire des prédictions	Algorithme qui cherche à découvrir des structures à partir de données non étiquetées
Objectif	Prédire des résultats futurs (prédiction)	Identifier des structures et des modèles cachés (analyse)
Utilisation	Dans les situations où la réponse est connue	Dans les situations où la réponse n'est pas connue
Données d'entrée	Des données étiquetées avec des résultats connus	Des données non étiquetées sans résultats connus

4. L'apprentissage profond « Deep Learning »

Dans le domaine de l'apprentissage automatique, le Deep Learning (DL) est un sous-domaine qui se concentre sur le développement d'algorithmes inspirés par la structure et la fonction du cerveau, en particulier les réseaux de neurones artificiels [51]. Ces concepts sont utilisés pour apprendre aux ordinateurs à effectuer des tâches qui sont naturelles pour les humains, telles que la classification d'images, de textes ou de sons. Le DL a gagné en popularité en raison de sa capacité à atteindre des taux de précision élevés, et il implique l'entraînement de modèles informatiques à l'aide de grands ensembles de données étiquetées et d'architectures de réseau de neurones [52].

5. La science des données « Data Science »

La Data Science (DS) désigne le processus d'analyse et d'extraction d'informations pertinentes à partir de données, notamment en identifiant des motifs cachés. Un « Data

Scientist » utilise souvent l'apprentissage automatique pour prédire des événements futurs en se basant sur les données [53].

La relation entre ces disciplines peut être représentée dans une image montrant l'apprentissage automatique comme un sous-ensemble de l'intelligence artificielle. Cependant, la Data Science est une discipline distincte qui se croise avec l'IA et l'apprentissage automatique sans en être un sous-ensemble [54].

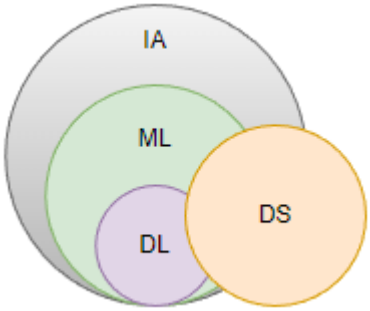

FIGURE 2.2 : RELATION ENTRE AI, ML, DL, DS.

6. Data Mining

- La fouille de données (Data Mining « DM ») désigne le processus de découverte de connaissances, d'identification de motifs et d'extraction de règles à partir d'ensembles de données [53]. Le modèle joue un rôle central dans le processus de fouille de données, en tant que composant principal. Par analogie, on pourrait soutenir que la fouille de données aurait pu être plus précisément décrite comme ‘‘l'extraction de connaissances à partir des données’’, car son objectif principal est d'extraire des informations précieuses et des connaissances à partir des ensembles de données disponibles [55]–[57].

- La découverte de connaissances (Knowledge discovery « KD ») est une procédure systématique qui vise à révéler de nouvelles idées et une meilleure compréhension dans un domaine d'application spécifique. Elle englobe plusieurs étapes, l'une d'entre elles étant l'extraction de données (DM), chacune étant axée sur la réalisation d'un objectif de découverte spécifique. Ces objectifs sont atteints en appliquant des méthodes de découverte appropriées adaptées à la tâche en cours [58], [59].

- La découverte de connaissances dans les bases de données (Knowledge discovery in databases « KDD ») fait référence au processus d'application de la procédure de découverte de connaissances spécifiquement aux bases de données [60].

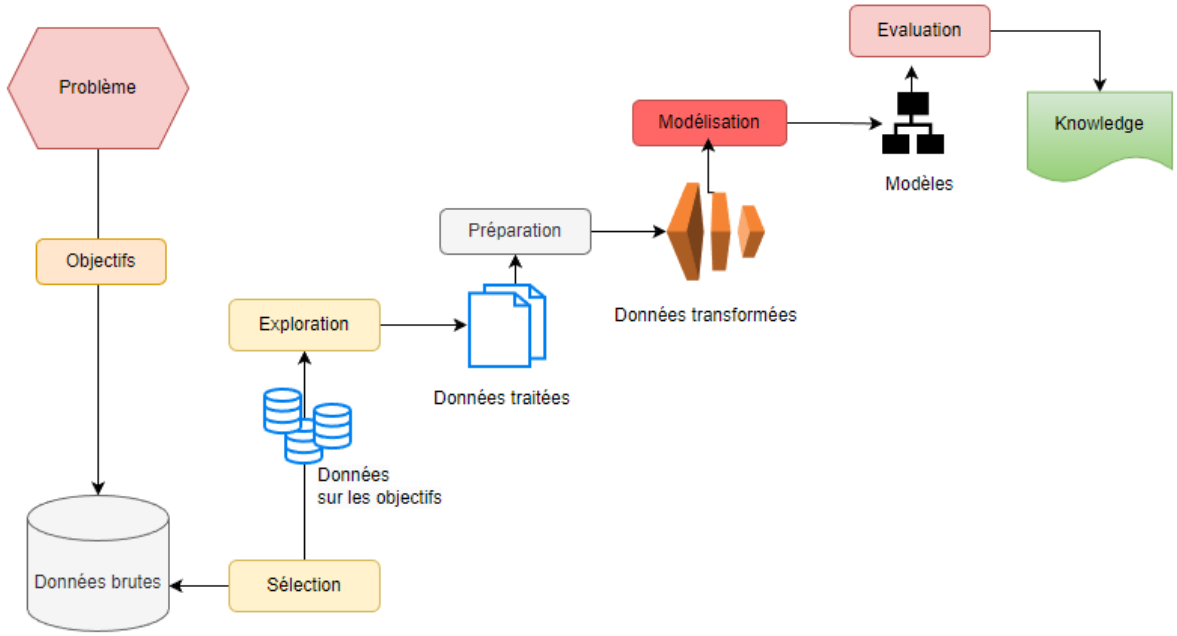

FIGURE 2.3 : DATA MINING (KDD) PROCESSUS.

6.1. Types de fouille de données

La fouille de données a la capacité de s'appliquer à divers types de données. Les bases de données, les entrepôts de données, les données transactionnelles et d'autres formats de données sont parmi les formes les plus couramment utilisées dans les applications de fouille de données [61], [62].

6.1.1. Les bases de données

Ils stockent des informations structurées dans des tableaux relationnels, sont une source de données couramment exploitée pour la fouille de données. Par exemple, dans le domaine de la fouille de données, il est possible d'analyser les données des clients pour prédire le risque de crédit des nouveaux clients en se basant sur leur âge [63], [64]. En utilisant un système de fouille de données, des modèles et des algorithmes peuvent être appliqués aux données des clients afin d'identifier des schémas et des tendances significatifs. En examinant l'âge des clients existants et en le comparant à leur historique de crédit, il est possible de développer un modèle prédictif qui évalue le risque de crédit des nouveaux clients en fonction de leur âge. Ce modèle peut être utilisé par les institutions financières et les prêteurs pour prendre des décisions éclairées lors de l'octroi de crédit aux nouveaux clients [31], [65], [66].

6.1.2. Les entrepôts de données

Les « Data Warehouse » centralisent de vastes quantités de données provenant de différentes sources, offrent également un terrain propice à l'exploration et à l'analyse [61].

6.1.3. Les données transactionnelles

Elles enregistrent les interactions entre les utilisateurs et les systèmes, sont souvent utilisées pour comprendre les schémas d'achat, les habitudes de consommation

et les préférences des clients. Dans une base de données transactionnelle, il peut exister des tables supplémentaires qui stockent des détails supplémentaires liés aux transactions. Ces détails peuvent inclure des descriptions d'articles, des informations sur le vendeur ou la succursale concernée, ainsi que d'autres informations pertinentes [61], [62].

6.1.4. Autres formats de données

En outre, il existe d'autres formats de données tels que les données textuelles, les données temporelles, les données géospatiales, les données multimédias, etc., qui peuvent être explorés et analysés à l'aide de techniques de fouille de données. L'objectif principal de la fouille de données est d'extraire des informations pertinentes, des tendances cachées et des motifs significatifs à partir de ces différentes formes de données, afin de prendre des décisions éclairées, de prédire des comportements futurs ou de générer de nouvelles connaissances.

Ainsi, la fouille de données s'avère être une méthode essentielle pour exploiter la richesse d'informations contenue dans ces divers types de données et pour en tirer des avantages stratégiques dans de nombreux domaines d'application [61], [62].

6.2. Perspectives de la fouille de données

La fouille de données a intégré de nombreuses techniques provenant d'autres domaines. En ceux que suivent plusieurs disciplines qui influencent fortement le développement des méthodes de fouille de données [62], [65], [67].

6.2.1. La statistique

La statistique est le domaine qui se concentre sur la collecte, l'analyse, l'interprétation et la présentation des données. Les modèles statistiques sont des fonctions mathématiques utilisées pour représenter les données et comprendre les caractéristiques des objets d'une classe spécifique. Les tâches de fouille de données peuvent intégrer des modèles statistiques. Par exemple, les statistiques peuvent être utilisées pour modéliser le bruit et traiter les valeurs de données manquantes. De plus, les méthodes statistiques peuvent être utilisées pour valider les résultats de la fouille de données par le biais de tests d'hypothèses statistiques. Ce processus implique de prendre des décisions statistiques basées sur des données expérimentales [68].

6.2.2. Machine Learning

Le sous-domaine de l'intelligence artificielle qui a été décrit dans la section 3.

6.2.3. Les systèmes de bases de données et les entrepôts de données

Les systèmes de bases de données et les entrepôts de données servent de référentiels de stockage pour les ensembles de données, qu'ils soient petits ou grands, comme mentionné précédemment. Cependant, les entrepôts de données offrent une perspective multidimensionnelle des données. Par conséquent, nous pouvons exploiter

ces deux environnements de données pour les traiter efficacement en utilisant la technologie de la fouille de données [68].

7. Phases de la fouille de données

Le processus de découverte des connaissances (KD) comprend généralement une série d'étapes bien définies qui guident l'ensemble du processus. Les données servent d'entrée, tandis que la sortie souhaitée est l'information précieuse répondant aux besoins des utilisateurs. Le processus global de solution pour la fouille de données peut être divisé en trois phases principales : la définition du problème, la fouille des données et la mise en œuvre des actions, qui implique le déploiement de la solution [63], [68].

7.1. Identification du problème

Le processus de fouille de données ne commence pas par les données, mais par l'identification d'un problème à résoudre. Il s'agit de formuler le problème, de le traduire en une ou plusieurs questions auxquelles la fouille de données peut répondre, et de comprendre comment les résultats de la fouille de données seront utilisés pour résoudre le problème, ce qui nous permet de déterminer l'approche la plus appropriée pour la fouille de données. Il est important de mentionner que la fouille de données n'est pas toujours l'outil adéquat pour résoudre les problèmes. Par exemple, si l'objectif est d'améliorer la productivité de l'équipe de vente en réduisant le temps nécessaire pour recueillir certaines informations, la meilleure suggestion pourrait être la mise en place d'un nouveau système de report basé sur des requêtes multidimensionnelles en temps réel plutôt que la fouille de données. Comprendre le problème nous permet de choisir l'approche analytique appropriée pour identifier la meilleure méthode de fouille à utiliser. En réalité, il existe deux catégories d'approches de fouille de données : les méthodes de découverte (descriptives) et les méthodes prédictives [68].

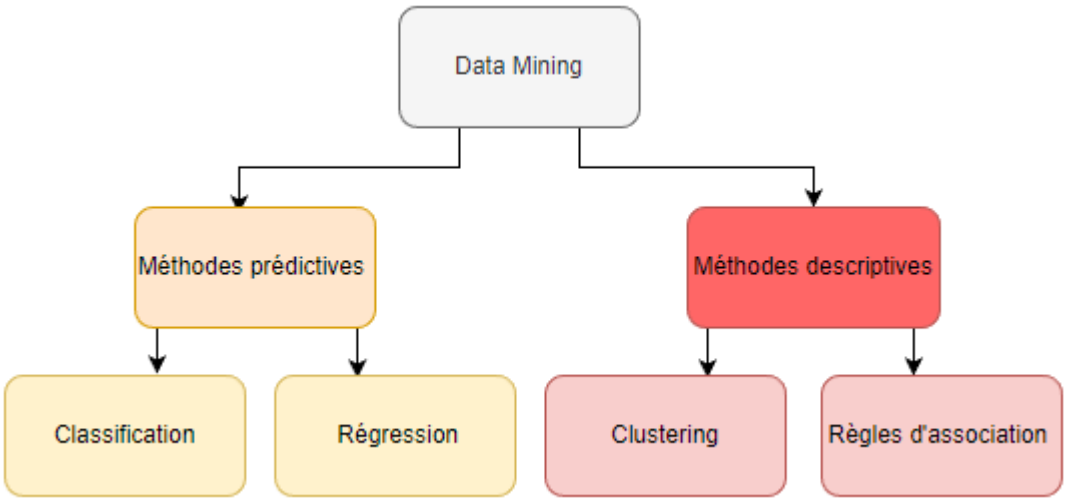

FIGURE 2.4: TÂCHES DU DATA MINING.

- **Méthodes prédictives :** effectuer une prédiction sur les valeurs des données en utilisant des résultats connus obtenus à partir de différentes données (classification, régression).
- **Méthodes de découverte (descriptives) :** sont des techniques de fouille de données qui permettent de trouver des motifs dans les données. L'objectif est de découvrir les relations qui sont inhérentes aux données (clustering, résumé, règles d'association et découverte de séquences).
- **Combinaison de différentes méthodes :** il existe des situations où il est plus approprié de combiner plusieurs techniques de fouille de données. Par exemple, si l'objectif est d'améliorer l'efficacité d'une campagne marketing pour une promotion de rentrée des classes à venir, une approche pourrait consister à effectuer une segmentation des clients en utilisant des techniques de regroupement pour identifier un groupe spécifique de clients. Ensuite, une analyse du panier d'achat peut être réalisée en utilisant des techniques d'association, en se concentrant uniquement sur les transactions du groupe cible identifié. Cette analyse peut aider à identifier les affinités entre les produits, qui peuvent ensuite servir de base à la stratégie promotionnelle. Il y a aussi d'autres exemples concernant le domaine des diagnostics médicaux [69].

7.2. Construire et déployer le data mining

Le processus de construction et de mise en œuvre de solutions de fouille de données implique une série de neuf étapes importantes, qui peuvent être hautement itératives par nature.

7.2.1. Préparation des données

La préparation des données implique de résoudre différents problèmes couramment présents dans les ensembles de données et les relations de bases de données, tels que les incohérences, les valeurs nulles, les valeurs extrêmes et le bruit. Pour atténuer ces problèmes, des routines de nettoyage des données sont utilisées pour combler les valeurs nulles, détecter les valeurs aberrantes et résoudre les incohérences. Ne pas nettoyer les données peut entraîner des confusions lors du processus d'analyse. Bien que certains algorithmes puissent intégrer des routines de nettoyage, ils ne sont pas toujours robustes, il est donc préférable de nettoyer les données en amont [59], [65].

7.2.2. Intégration de données (Collection)

La combinaison de données provenant de différentes sources peut réduire les redondances et les incohérences dans le jeu de données final si le processus d'intégration est réalisé avec soin. Cela peut améliorer la vitesse et la précision du processus de fouille de données ultérieur [59], [70].

7.2.3. Sélection et exploration des données

Pendant le processus de fouille de données, il est important d'identifier les sources de données disponibles et d'extraire et de sélectionner soigneusement les données pertinentes pour l'analyse. Cela implique d'évaluer la quantité et la qualité des données afin de garantir le développement de modèles robustes [70]. Les données elles-mêmes se composent d'une série d'échantillons, chacun étant décrit par un ensemble de variables. Cette étape peut être réalisée à travers une série d'étapes, telles que l'analyse des métadonnées associées à chaque variable pour recueillir des informations sur les types de données, les valeurs potentielles, la source d'origine, le format et d'autres caractéristiques. Il existe différents types de variables utilisées pour décrire les échantillons [69]:

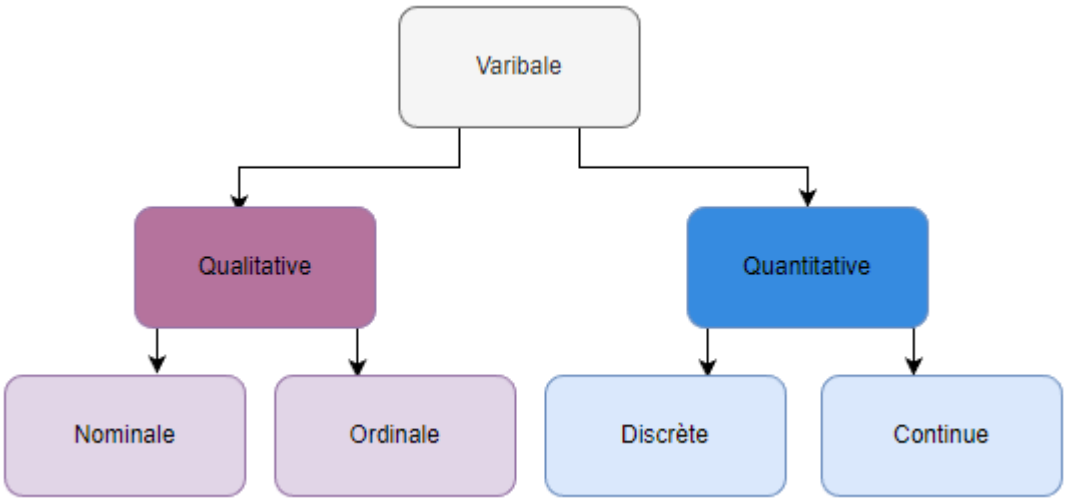

FIGURE 2.5 : TYPE DE VARIABLE.

- Quantitative : Peut inclure :
  - • Discrète : Par exemple : nombre de personnes, nombre de véhicules...etc.
  - • Continue : Par exemple : salaire, longueur, avantages...etc.
- Qualitative : Peut inclure :
  - • Nominale : Nomme l'objet auquel ils font référence sans pouvoir établir un ordre (état civil, genre, couleur, race...etc).
  - • Ordinale : Un ordre peut être établi dans ses valeurs (élevé, moyen, faible).

7.2.4. Nettoyage des données (Data Cleaning)

Au cours des phases précédentes, les données peuvent rencontrer différents problèmes tels que l'incomplétude, le bruit, les valeurs manquantes et les valeurs aberrantes. Cette étape particulière joue un rôle crucial en abordant et en rectifiant ces problèmes, garantissant ainsi que les données sont soigneusement nettoyées et affinées [67], [71], [72].

7.2.4.1. Problèmes de nettoyage des données

La présence de divers problèmes de nettoyage des données pose des obstacles importants lorsqu'il s'agit de garantir l'intégrité et la fiabilité des ensembles de données. Dans le domaine de l'exploration de données et de l'analyse, il est primordial de faire face à de nombreux problèmes, tels que des données incomplètes, des données bruitées, des valeurs manquantes et des valeurs aberrantes [71]. Surmonter ces défis nécessite la mise en œuvre de méthodes de nettoyage des données résilientes et efficaces, capables de renforcer la précision et l'authenticité des analyses ultérieures. En comprenant et en abordant de manière efficace le large éventail de problèmes de nettoyage des données, nous pouvons établir des bases solides pour extraire des informations précieuses et prendre des décisions éclairées basées sur des données fiables [69].

Les valeurs manquantes se produisent lorsque de nombreuses tuples ne comportent pas de valeurs enregistrées pour plusieurs attributs. Il existe plusieurs techniques disponibles pour gérer ce problème [72] :

- Une approche consiste à **ignorer les tuples avec des valeurs manquantes**. Cependant, cette méthode peut ne pas être très efficace lorsqu'il s'agit de plusieurs attributs ayant des valeurs manquantes.
- Une autre option consiste à **remplir manuellement les valeurs manquantes**. Cependant, cette approche peut ne pas être réalisable lorsqu'il s'agit d'un grand ensemble de données avec de nombreuses valeurs manquantes.
- Utiliser **une constante globale** pour remplir les valeurs manquantes est une autre méthode, mais elle n'est pas infaillible. Parfois, le programme de fouille de données peut interpréter à tort ces constantes globales comme des concepts intéressants.
- Alternativement, on peut utiliser **une mesure de tendance centrale pour l'attribut**. La moyenne convient aux distributions de données normales (symétriques), tandis que la médiane est plus appropriée pour les distributions de données asymétriques.
- **Les données bruitées** font référence à des erreurs aléatoires ou des variations dans une variable qui peuvent corrompre les données et entraver l'interprétation précise par le système. Cela peut avoir un effet néfaste sur les résultats de toute analyse de données.

7.2.5. Transformation de données (Data Transformation)

À ce stade, les données sont converties ou transformées en une représentation et un format adapté à l'exploitation minière. Par exemple, si l'algorithme sélectionné nécessite des données numériques en entrée mais que les données disponibles sont catégorielles, il devient crucial de convertir ces données dans le format approprié, c'est-

à-dire en données numériques. Cela garantit la compatibilité et permet à l'algorithme de traiter efficacement les données.

Dans le domaine de l'extraction de données, les stratégies de transformation des données servent de boussole guide, illuminant le chemin vers des informations précieuses [70], [73].

- **Smoothing (Lissage):** Éliminer ou réduire la présence de bruit dans les données.
- **Construction d'attributs:** Introduire des attributs supplémentaires qui aident dans le processus d'exploration de données.
- **Agrégation:** Effectuer des opérations de synthèse ou d'agrégation sur les données, comme dans la création de « data warehouses ».
- **Normalisation:** Ajuster les valeurs des données à une plage spécifique, telle que [0, 1] ou [-1, 1], en utilisant des techniques telles que le Minimum-Maximum, le score Z ou l'échelle décimale. Ces méthodes de normalisation sont couramment utilisées pour accélérer la phase d'apprentissage [63].

7.2.6. Choix de l'analyste de données

Durant cette phase, nous sélectionnons l'analyste de données qui sera chargé de créer et d'appliquer des techniques ou des modèles d'exploration de données pour découvrir des motifs ou des règles. Divers algorithmes d'apprentissage automatique sont utilisés pour construire différents modèles de prédiction. Parmi ces modèles, le plus optimal est choisi pour être déployé [65].

7.2.7. Phase de présentation

Pendant cette étape, le modèle généré (la connaissance découverte) est développé, sauvegardé et examiné à l'aide de visualisations appropriées choisies par l'analyste pour évaluer la qualité du modèle et valider sa pertinence. À cette étape, l'analyste de données doit comprendre l'impact de l'ajustement des différents paramètres sur le modèle résultant et acquérir une compréhension de la manière d'interpréter efficacement le modèle de données [67], [70].

7.2.8. Évaluation des modèles

Avant que les modèles puissent être déployés et utilisés pour prendre des décisions au sein d'une organisation, il est important de les évaluer pleinement et de prouver qu'ils conviennent à l'objectif. Cette phase englobe toutes les tâches d'évaluation nécessaires pour démontrer qu'un modèle de prédiction sera en mesure de faire des prédictions précises après son déploiement et qu'il ne souffre pas de surajustement « overfitting » ou de sous-ajustement « underfitting » [36], [74], [75].

7.2.9. Deployer la solution (Knowledge discovery)

Les modèles d'apprentissage automatique sont construits dans l'intention de servir des objectifs spécifiques au sein d'une organisation. Les résultats obtenus grâce à l'exploration de données peuvent être mis en œuvre dans une série de systèmes de diverses manières [65], [67].

8. Deep learning vs machine learning?

Le choix entre le Deep Learning et le Machine Learning dépend de la nature précise de la tâche à accomplir, de la disponibilité des données pertinentes et des ressources informatiques à disposition [19], [34], [54].

TABLEAU 2.5 : APPRENTISSAGE PROFOND PAR RAPPORT À APPRENTISSAGE MACHINE.

	Apprentissage Profond	**Apprentissage Automatique**
Définition	Sous-domaine de l'apprentissage automatique	Domaine plus large de l'apprentissage à partir des données
Architecture	Réseaux de neurones profonds	Diverses techniques et algorithmes d'apprentissage
Représentation	Apprentissage hiérarchique des caractéristiques	Ingénierie manuelle des caractéristiques
Données requises	Grandes quantités de données étiquetées	Peut fonctionner avec de plus petits ensembles de données étiquetées
Performance	Performances élevées sur des tâches complexes	Performance dépendante de la qualité des caractéristiques et des modèles
Interprétabilité	Faible interprétabilité en raison de modèles complexes	Plus grande interprétabilité, modèles plus transparents
Exigences matérielles	Calcul intensif, bénéficie des GPU	Moins exigeant en termes de matériel, peut fonctionner sur des CPU standard
Domaines d'application	Vision par ordinateur, traitement du langage naturel, reconnaissance vocale	Analyse prédictive, classification, régression, etc.

L'apprentissage automatique est un terme plus large englobant diverses techniques pour l'apprentissage à partir des données, tandis que l'apprentissage profond est une approche plus spécifique basée sur des réseaux de neurones profonds. Les deux domaines ont des

applications étendues dans de nombreux domaines tels que la reconnaissance d'images, la traduction automatique, la reconnaissance vocale, la recommandation personnalisée et bien d'autres.

9. Relation entre l'apprentissage automatique « ML » et la fouille de données « DM »

Il est évident qu'il existe de nombreux parallèles entre la fouille de données et l'apprentissage automatique. Dans le contexte des tâches de classification et de regroupement, le domaine de l'apprentissage automatique se concentre principalement sur l'évaluation de l'exactitude des modèles. En revanche, la recherche en fouille de données met non seulement l'accent sur l'exactitude, mais accorde également une importance significative à l'efficacité et à la scalabilité des techniques de fouille lorsqu'elles sont appliquées à des ensembles de données étendus. De plus, la fouille de données aborde également les défis liés à la gestion de types de données complexes et s'efforce d'explorer des méthodes nouvelles et alternatives pour l'analyse [64], [76], [77].

- L'apprentissage automatique : est associé à l'examen, à la création et à l'avancement d'algorithmes qui permettent aux ordinateurs d'acquérir des connaissances sans avoir besoin d'une programmation explicite.
- La fouille de données : consiste à extraire des connaissances ou à découvrir des motifs inconnus et intéressants à partir de données non structurées. Des algorithmes d'apprentissage automatique sont utilisés dans le processus de fouille de données.

10. L'intelligence artificielle dans la médecine

Avec autant d'argent en jeu, autant de consommateurs mécontents et la vie et la santé humaine en jeu, il semble qu'il y ait un grand besoin de résolveurs de problèmes dans le domaine de la santé. L'une des idées pour améliorer les soins de santé est de mettre davantage l'accent sur la prévention et moins sur le traitement. Tous les problèmes de santé ne peuvent pas être évités, mais dans de nombreux cas, une intervention précoce peut conduire à de meilleurs résultats en matière de santé et à une réduction des coûts. L'un des éléments clés des soins de santé préventifs est le dépistage des maladies. Grâce à ce dépistage, les maladies peuvent être diagnostiquées à un stade précoce et traitable [18]. Cependant, il n'est pas possible pour tout le monde de se faire dépister pour toutes les maladies possibles. Une meilleure solution consisterait à disposer d'un moyen peu coûteux, évolutif et fiable de mesurer le risque de maladie et d'en prédire l'apparition. Les prestataires de soins de santé ou les assureurs pourraient alors utiliser ce prédicteur pour identifier les personnes à haut risque et leur recommander des tests ou d'autres interventions [20], [78], [79].

Au cours des sept dernières années, le domaine de l'AIM « AI in Medicine » a connu une évolution spectaculaire. Avec l'avènement de la ML et de la DL, les applications de l'AIM se sont développées de manière exponentielle, offrant des possibilités sans précédent

pour la médecine personnalisée par opposition aux approches basées uniquement sur des algorithmes. Les modèles prédictifs offrent désormais une approche innovante pour le diagnostic de diverses maladies, l'anticipation de la réponse thérapeutique et, potentiellement, la réalisation d'une médecine préventive à l'avenir. L'intégration de l'IA dans la médecine promet d'améliorer la précision des diagnostics, de rationaliser le flux de travail des prestataires et les opérations cliniques, de faciliter un meilleur suivi des maladies et des thérapies, et d'optimiser la précision des procédures, améliorant ainsi les résultats globaux pour les patients [80]–[82].

10.1 Systèmes d'intelligence artificielle en santé

L'IA a connu un regain de popularité ces dernières années dans le secteur de la santé en raison de sa capacité à traiter et à analyser des quantités massives de données en temps réel. Les avantages de l'utilisation des systèmes d'IA dans les soins de santé sont multiples : aide à la décision clinique, suivi des patients, détection précoce des maladies et soins individualisés [20]. Les applications de l'IA dans le secteur des soins de santé couvrent un large éventail de spécialités, telles que la radiologie, la génomique, la gestion des dossiers médicaux électroniques et l'analyse des données de surveillance de la santé publique [67], [83]. Néanmoins, la mise en œuvre de l'IA dans les soins de santé doit être soigneusement étudiée en ce qui concerne la confidentialité des données et les préoccupations réglementaires afin de garantir une mise en œuvre éthique et efficace de cette technologie [82].

11. Les sources de données dans la santé

La prolifération des sources de données de santé est évidente, englobant les données acquises et numérisées lors d'hospitalisations ou de consultations médicales, telles que les diagnostics, les traitements, les examens médicaux et l'imagerie, les données relatives à la consommation de soins de santé et les données provenant d'appareils connectés médicaux et non médicaux, tels que les tensiomètres, les pèse-personnes, les « smartwatches » et les téléphones intelligents. Il est évident que le potentiel de nouveauté de ces sources de données ne provient pas uniquement de leur volume massif, mais de la capacité d'accéder et de croiser des données qui étaient auparavant inaccessibles, et des moyens d'analyser et de traiter ces données grâce à l'intelligence artificielle. Néanmoins, les méthodes médicales et épidémiologiques conventionnelles peuvent également être utilisées pour analyser ces données [66], [68], [84].

L'utilisation de grandes quantités de données à des fins de santé publique souligne l'importance de l'individu dans la promotion et la sauvegarde de la santé de la population. Néanmoins, cette mobilisation comporte des risques inhérents, notamment en ce qui concerne l'exposition d'informations personnelles relatives au comportement, à l'état de santé et à la vie privée des individus [70].

L'industrie de la santé a toujours disposé d'une abondance de données. Avec la multitude de composants et d'acteurs impliqués, les prestataires de soins et les compagnies d'assurance ont une pléthore de facteurs à surveiller et à évaluer. Ces données accumulées servent plusieurs objectifs cruciaux, tels que la surveillance et la gestion des coûts et des opérations

des établissements de santé. Plus important encore, ces données capturent et documentent l'état de santé des individus, à la fois à petite et à grande échelle. La valeur des données dans le domaine de la santé est inestimable, en particulier dans le contexte de l'amélioration des systèmes de santé. Bien que notre recherche se concentre sur l'utilisation de données médicales pour la prédiction du diabète, de nombreux autres aspects de la santé pourraient en bénéficier et subir une transformation complète grâce à l'utilisation intelligente des données [79].

Gagner l'accès aux données de santé peut être une tâche difficile, car des règlements de confidentialité stricts et des intérêts commerciaux imposent plusieurs obstacles qui doivent être surmontés avant que les données ne puissent être partagées. Cela peut entraver le progrès des chercheurs indépendants qui n'ont pas d'affiliation avec des compagnies d'assurance ou des institutions de santé. Établir des partenariats académiques avec ces organisations est une étape essentielle pour obtenir l'accès aux données. L'auteur de cet ouvrage a vécu cela de première main, tout comme son superviseur. Malgré ces défis, les avantages potentiels de comprendre et d'utiliser les données médicales pour améliorer les soins de santé dépassent de loin les difficultés rencontrées pour accéder aux données [70].

11.1. Données sous forme d'images

La majorité des données d'imagerie médicale sont produites à partir de techniques de radiologie telles que les radiographies, les échographies, les scanners CT, les scans PET et les IRM, qui produisent des représentations graphiques des tissus internes. Généralement, ces images sont examinées par un spécialiste du diagnostic visuel et ensuite transmises au médecin et au patient. Cependant, l'utilisation de données d'images à des fins d'analyse informatique représente un défi de taille. La création d'algorithmes capables d'extraire des significations importantes à partir d'images numériques est un domaine de recherche actif et en évolution dans les domaines de la vision par ordinateur et de l'apprentissage automatique. Pour exploiter pleinement le potentiel des données d'imagerie médicale, il est impératif d'utiliser des techniques provenant de ces domaines [66], [84].

11.2. Données en texte libre « FreeText »

Une partie importante des données médicales se trouve dans des champs de texte non structurés. Ces données non structurées peuvent prendre la forme de notes prises par des infirmières ou des médecins concernant différents aspects de la visite d'un patient. Ces notes peuvent contenir des phrases incomplètes, des paragraphes détaillés, ou des abréviations particulières à un infirmier, ainsi que des fautes de frappe et des gribouillis illisibles. Malgré leur nature chaotique, les données médicales non structurées peuvent contenir des informations essentielles sur la santé d'un patient et son expérience avec le système de santé, en faisant une source de données cruciale. Cependant, les données non structurées peuvent être difficiles à utiliser sur le plan informatique. Bien que les humains puissent interpréter et comprendre les données textuelles non structurées, la reconnaissance de l'écriture manuscrite et l'extraction de la signification des données requièrent des techniques avancées de vision par ordinateur et de traitement du langage naturel. Bien que le travail actuel n'aborde pas cette question, le défi d'incorporer des données médicales non structurées dans les méthodes d'apprentissage automatique est significatif [66], [69].

11.3. Données structurées

Les données structurées en matière de santé font référence aux données organisées dans un format prédéfini, tel qu'une base de données ou un tableur, et qui peuvent être analysées aisément à l'aide de techniques statistiques standard. En comparaison, les données non structurées sont plus compliquées à traiter et à classifier, car elles sont plus complexes et hétérogènes [66], [68], [84].

On peut trouver plusieurs types de données structurées en santé, dont voici quelques exemples :

- **Les données démographiques** qui comprennent des informations sur le nom, l'âge, le genre et l'adresse du patient, les données cliniques qui comprennent des informations sur les antécédents médicaux du patient tels que les diagnostics antérieurs, les traitements et les procédures.
- **Les signes vitaux** qui incluent des données sur la fréquence cardiaque, la pression artérielle, la fréquence respiratoire et la température.
- **Les résultats des tests de laboratoire** qui incluent des données sur les tests sanguins, les tests d'urine et autres tests de diagnostic, les données sur les médicaments qui comprennent des informations sur les médicaments prescrits au patient, telles que la posologie, la fréquence et la durée.
- Et enfin **les données de réclamation** qui incluent des informations sur les services médicaux fournis au patient et les coûts correspondants.

Les données organisées ou structurées sont essentielles pour les applications d'intelligence artificielle dans le domaine de la santé, car elles sont faciles à traiter et à analyser à l'aide d'algorithmes d'apprentissage automatique. Les modèles d'IA peuvent ainsi tirer parti de ces données pour prédire les résultats des patients, repérer les risques sanitaires potentiels et recommander des plans de traitement personnalisés [66], [68], [84].

11.4. Variables simples

Une grande partie des données de santé se compose de variables numériques et catégorielles simples, qui incluent des facteurs démographiques tels que l'âge, le sexe et l'ethnicité, ainsi que des indicateurs de santé tels que la taille, le poids, la pression artérielle et le pouls. Ces variables simples peuvent être analysées à l'aide de techniques statistiques conventionnelles telles que la régression linéaire ou logistique. Cependant, se limiter à l'analyse de ces variables de base reviendrait à passer à côté d'informations précieuses pouvant être obtenues à partir de sources de données plus complexes [84].

11.5. Capture de données

Il existe plusieurs systèmes pour capturer les données médicales. Dans les systèmes de santé modernes, les outils dossiers de santé électroniques (Electronic Health Record ou « EHR ») sont utilisés pour stocker de manière systématique et numérique une grande variété de données, incluant les données démographiques et antécédents médicaux des

patients, les résultats de laboratoire, les examens physiques, les images radiologiques, et bien plus encore [20]. Les EHR facilitent également l'accès aux données et leur visualisation, permettant aux médecins et aux patients de mieux s'informer. Bien que ces dossiers ne captent que les événements qui se produisent dans un établissement ou un ensemble d'établissements particuliers, ils fournissent un compte rendu vivant de l'état de santé individuel au cours des visites à l'hôpital et dans d'autres établissements de soins. Les demandes de remboursement d'assurance maladie représentent un autre réservoir riche en données de santé [19], [32]. Les données de remboursement se concentrent sur les personnes inscrites dans le plan d'assurance et leurs interactions avec le système de santé. Ces documents comprennent généralement des informations démographiques de base sur les patients, ainsi que des diagnostics, des procédures, des médicaments, des visites à l'hôpital et en ambulatoire, ainsi que les coûts associés. En raison de leur nature centrée sur le patient et de leur capacité à capturer l'activité médicale d'un individu dans une variété d'hôpitaux, de cliniques et de pharmacies, les données de remboursement « Claims Data » fournissent un portrait plutôt complet de l'historique médical et de l'état de santé actuel. Cependant, les données de remboursement manquent souvent de détails fins et d'informations cliniques riches présentes dans les EHR [68], [84].

12. Conclusion

L'avènement de la révolution numérique dans le domaine de la santé a ouvert une nouvelle ère de collecte, stockage et analyse de données massives. Ce déluge de données a donné naissance à une multitude de techniques basées sur l'IA qui contribuent à la création d'un écosystème de santé plus interconnecté et érudit. La confluence de données massives et de méthodologies liées à l'IA est profondément entrelacée et se renforce mutuellement, car les algorithmes d'apprentissage automatique nécessitent des quantités considérables de données pour acquérir des connaissances et sans l'aide d'analyses basées sur l'IA, les ensembles de données massives ont une valeur limitée.

En conclusion de ce chapitre, nous anticipons avec enthousiasme le prochain chapitre qui portera sur un état de l'art approfondi de l'utilisation des techniques d'intelligence artificielle dans le domaine de la prédiction du diabète. Ce panorama exhaustif des avancées actuelles nous permettra de mieux comprendre les différentes approches existantes et de situer notre propre recherche dans ce contexte dynamique et en constante évolution.

Chapitre 3

État de l’art sur la prédiction du diabète par l’IA

1. Introduction

L'utilisation de l'intelligence artificielle (IA) dans le domaine de la prédiction du diabète a connu des progrès significatifs, montrant le potentiel d'améliorer les résultats pour les patients et de réduire les dépenses de santé. Le diabète étant une maladie chronique qui touche une vaste population dans le monde entier, une identification et une intervention précoces sont nécessaires pour une prise en charge efficace. Les algorithmes d'IA sont capables de traiter des données substantielles provenant de diverses sources, telles que les dossiers médicaux électroniques et les données de santé générées par les patients, afin de prédire la probabilité de développer un diabète et d'identifier les personnes à haut risque.

Le diabète est une affection persistante qui touche un nombre important de personnes dans le monde et constitue une cause majeure de maladie et de décès. L'identification et l'intervention précoces sont essentielles à la bonne gestion de la maladie et à la prévention de toute complication associée. Ces derniers temps, l'intérêt pour l'application de l'intelligence artificielle (IA) dans le secteur des soins de santé, en particulier pour la prédiction du diabète, s'est considérablement accru. Des algorithmes d'IA ont été utilisés dans diverses études de recherche pour prédire le risque de diabète, identifier les personnes à haut risque et développer des stratégies de prévention personnalisées.

La prédiction du diabète est un domaine de recherche essentiel qui vise à identifier les facteurs prédictifs de cette maladie chronique et à développer des outils de diagnostic précoce. Ce chapitre est un état de l'art en matière de prédiction du diabète par l'IA. Dans ce chapitre, nous examinerons attentivement les études sélectionnées qui se sont penchées sur la prédiction du diabète. Nous nous concentrerons également sur la comparaison des méthodologies utilisées, des ensembles de données recueillis et des résultats obtenus. L'objectif est de mettre en évidence les forces et les faiblesses des différentes approches afin de mieux comprendre les progrès réalisés dans ce domaine. Tout ça pour nous aider à identifier les lacunes et les limites des approches existantes, mettant en évidence les domaines où des recherches supplémentaires sont nécessaires pour améliorer la prédiction du diabète. Cette analyse approfondie nous permettra d'acquérir une vision globale de l'état actuel des recherches et d'orienter les futures études dans ce domaine prometteur.

2. Méthodes actuelles de prédiction du diabète avec l'IA

Le paysage actuel de la prédiction du diabète avec l'intelligence artificielle (IA) est caractérisé par une multitude de méthodologies sophistiquées qui exploitent des techniques informatiques de pointe pour élucider les subtilités de ce trouble métabolique complexe. En exploitant le potentiel de l'IA, comprenant des algorithmes d'apprentissage automatique de pointe et des stratégies avancées d'exploration de données, il est possible d'extraire des informations inestimables et de découvrir des schémas complexes à partir de vastes ensembles de données hétérogènes liées au diabète [85], [86]. Ces méthodes englobent un large éventail d'approches innovantes, comprenant l'apprentissage en ensemble, les architectures d'apprentissage profond, les machines à vecteurs de support, les réseaux bayésiens et les modèles hybrides qui amalgament plusieurs algorithmes pour atteindre des

performances prédictives améliorées. L'intégration de caractéristiques diverses, telles que les variables cliniques, les marqueurs génétiques, les données liées au mode de vie et les données omiques, facilite le développement de modèles prédictifs complets capables de capturer la nature multifacette de l'étiologie et de la progression du diabète. De plus, la fusion de sources de données disparates grâce à des stratégies innovantes d'intégration de données, telles que la fusion de données, l'apprentissage multi-vue et l'apprentissage par transfert, permet d'exploiter des informations complémentaires et d'améliorer les capacités prédictives des modèles [87]–[89].

3. Avancées et tendances récentes en matière d'IA pour la prédiction du diabète

3.1. L'IA explicative (IAE) dans la prédiction du diabète

L'intelligence artificielle explicative (Explainable AI « XAI ») dans le contexte de la prédiction du diabète représente un paradigme novateur qui vise à élucider les processus de prise de décision complexes des modèles d'intelligence artificielle, permettant ainsi une compréhension globale des facteurs sous-jacents contribuant à leurs prédictions. Dans le domaine de la prédiction du diabète, où une évaluation précise des risques et des informations exploitables revêt une importance primordiale, l'intégration des techniques d'IAE joue un rôle essentiel dans l'amélioration de la transparence, de l'interprétabilité et de la fiabilité [90].

En exploitant des algorithmes et des méthodologies sophistiqués, les techniques d'IAE s'efforcent de dévoiler les relations complexes, les schémas latents et l'importance des caractéristiques au sein des modèles d'IA, permettant ainsi aux cliniciens, aux chercheurs et aux patients de comprendre le raisonnement derrière les prédictions. Ce changement de paradigme, passant des modèles opaques à des systèmes interprétables, offre de nombreux avantages, notamment l'identification de nouveaux biomarqueurs, l'élucidation des mécanismes de la maladie, l'identification de facteurs de risque modifiables et l'amélioration de la prise de décision clinique [91].

Les approches d'IAE englobent une gamme diverse de méthodes, notamment les explications basées sur des règles, l'analyse de l'importance des caractéristiques, les techniques d'interprétabilité locale et globale, ainsi que les méthodes agnostiques vis-à-vis des modèles pouvant être appliquées à différents modèles d'IA. De plus, la fusion de l'IAE avec l'expertise métier et les connaissances médicales renforce la pertinence contextuelle des explications, garantissant ainsi leur utilité et leur applicabilité clinique [90]–[92]. Néanmoins, l'intégration de l'IAE dans la prédiction du diabète n'est pas sans défis.

3.1.1. Les défis existants

Ces défis comprennent l'équilibre entre la complexité du modèle et son interprétabilité, la prise en compte des biais potentiels et des limites des techniques d'IAE elles-mêmes, la conciliation entre la nécessité de transparence et les préoccupations liées à la confidentialité des données, et l'établissement de métriques

d'évaluation normalisées pour évaluer la qualité et la compréhensibilité des explications. Pour surmonter ces défis, des collaborations interdisciplinaires entre les chercheurs en IA, les professionnels de la santé et les organismes de réglementation sont nécessaires pour développer des cadres solides, des lignes directrices et des meilleures pratiques pour le déploiement éthique de l'IAE dans la prédiction du diabète.

En adoptant le paradigme de l'IAE, le domaine de la prédiction du diabète peut favoriser une compréhension approfondie des mécanismes sous-jacents de la maladie, permettre aux patients de prendre des décisions éclairées, faciliter les interventions personnalisées et instaurer la confiance dans la fiabilité et la responsabilité des prédictions basées sur l'IA [92], [93].

En fin de compte, l'intégration de l'IAE dans la prédiction du diabète ouvre la voie à un avenir où les systèmes d'IA agissent en tant que partenaires de confiance dans le domaine de la santé, renforçant l'expertise humaine et révolutionnant la gestion et la prévention de ce défi mondial de santé.

3.2. Les méthodes d'apprentissage profond

Ces dernières années, le domaine de l'intelligence artificielle (IA) a connu des avancées remarquables et des tendances émergentes dans le domaine de la prédiction du diabète. Ces avancées ont eu un impact significatif sur le paysage de la recherche et présentent un grand potentiel pour révolutionner la gestion des maladies et améliorer les résultats de santé. Notamment, les méthodes d'apprentissage profond, telles que les réseaux neuronaux convolutifs (CNN) et les réseaux neuronaux récurrents (RNN), ont suscité une attention considérable et ont démontré leur efficacité dans la capture de motifs complexes et de dépendances temporelles dans diverses modalités de données pertinentes pour la prédiction du diabète, telles que les images médicales, les séries temporelles et les dossiers de santé électroniques [94]–[97].

3.3. L'intégration de sources de données multimodales

Une autre tendance marquante est l'intégration de sources de données multimodales, où la convergence de différents types de données, notamment les données cliniques, les informations génétiques, les facteurs liés au mode de vie et les mesures physiologiques, est devenue de plus en plus répandue [98], [99]. En exploitant la puissance synergique de ces sources de données multimodales, des modèles d'IA complets peuvent capturer l'étiologie multifacette et la progression du diabète, permettant ainsi des évaluations de risques plus précises et des interventions personnalisées [100], [101].

3.4. L'apprentissage fédéré

De plus, le déploiement de l'apprentissage fédéré (Federated Learning « FL ») et des techniques de préservation de la vie privée a gagné une importance dans le contexte de la prédiction du diabète. L'apprentissage fédéré offre une approche distribuée pour l'entraînement des modèles d'IA, permettant un apprentissage collaboratif entre plusieurs institutions sans partage de données, préservant ainsi la confidentialité des patients [102],

[103]. Cette approche revêt une grande importance dans les établissements de santé où la sécurité et la confidentialité des données sont primordiales. En agrégeant les connaissances à partir de sources de données décentralisées, l'apprentissage fédéré facilite le développement de modèles de prédiction du diabète robustes et généralisables [104], [105].

3.5. La surveillance en temps réel

En outre, on observe une évolution vers une prédiction en temps réel et personnalisée du diabète, en combinant les modèles d'IA avec des applications mobiles et des dispositifs portables. Cette tendance permet une surveillance continue des paramètres de santé des individus, facilitant des interventions rapides, une surveillance à distance et des stratégies de soins adaptées. La combinaison des prédictions basées sur l'IA avec des données en temps réel donne aux individus à risque de diabète ou déjà diagnostiqués la possibilité de prendre des décisions éclairées et d'adopter des interventions personnalisées, améliorant ainsi les résultats en matière de santé [106].

Donc, les récentes avancées et tendances émergentes dans l'IA pour la prédiction du diabète démontrent le potentiel transformateur de ces méthodologies dans la gestion des maladies, les interventions personnalisées et les stratégies de santé publique. L'intégration de l'apprentissage profond, des données multimodales, de l'IA explicative, des techniques de préservation de la vie privée et de la surveillance en temps réel représentent un changement de paradigme dans le domaine de la santé, favorisant des modèles de prédiction du diabète plus précis, interprétables et centrés sur le patient. Ces avancées promettent d'alléger le fardeau du diabète, d'améliorer les résultats en matière de santé et, en fin de compte, d'améliorer la qualité de vie globale des personnes atteintes de ce trouble métabolique chronique [107], [108].

4. Application de l'apprentissage automatique à la prédiction du diabète

Ces dernières années, l'application des techniques d'apprentissage automatique dans le domaine de la prédiction du diabète a suscité une attention considérable. Les algorithmes d'apprentissage automatique ont démontré des capacités remarquables à exploiter des ensembles de données à grande échelle pour extraire des informations précieuses et développer des modèles prédictifs précis. Dans cette section, nous explorons les différentes applications de l'apprentissage automatique dans la prédiction du diabète, en examinant différentes méthodologies tous en abordant les métriques de performance.

4.1. Prétraitement des données (Data Preprocessing)

Avant d'appliquer des algorithmes d'apprentissage automatique, il est crucial de prétraiter et d'ingénierie les données pour garantir leur qualité et leur pertinence. Il existe différentes techniques de prétraitement des données, notamment le nettoyage des données, la normalisation, la gestion des valeurs manquantes, la gestion du déséquilibre des classes, la détection des valeurs aberrantes et l’encodage des variables [109].

4.1.1. Nettoyage des données

Le nettoyage des données vise à identifier et à corriger les erreurs, les incohérences et les valeurs aberrantes dans les données avant de les utiliser pour former un modèle. Voici quelques étapes courantes de nettoyage des données[3] :

4.1.1.1. Gestion des valeurs manquantes

Les données peuvent contenir des valeurs manquantes, ce qui peut affecter la performance du modèle. Vous pouvez choisir de supprimer les lignes ou les colonnes contenant des valeurs manquantes, ou bien les remplacer par des valeurs appropriées comme la moyenne, la médiane ou une valeur prédéfinie.

4.1.1.2. Détection et traitement des valeurs aberrantes

Les valeurs aberrantes sont des valeurs qui diffèrent considérablement du reste des données. Elles peuvent résulter d'erreurs de saisie ou d'autres anomalies. Il est important de les détecter et de décider si elles doivent être corrigées ou supprimées.

4.1.1.3. Correction des erreurs de saisie

Les erreurs de saisie sont courantes dans les données. Par exemple, des valeurs mal formatées ou des unités incorrectes. Il est important de les corriger pour garantir l'exactitude des données.

4.1.1.4. Normalisation des noms de catégories

Si vos données contiennent des catégories similaires écrites de différentes manières, vous pouvez normaliser ces noms pour les regrouper correctement.

4.1.1.5. Élimination des doublons

Les données peuvent contenir des enregistrements en double, ce qui peut biaiser les résultats. Il est recommandé de les supprimer pour éviter toute répétition inutile.

4.1.2. Normalisation

La normalisation[4] est le processus de mise à l'échelle des données afin qu'elles se situent dans une plage spécifique. Cela peut aider les algorithmes d'apprentissage automatique à converger plus rapidement et à éviter que certaines fonctionnalités ne dominent les autres en raison de leur échelle différente.

[3] https://datascience.eu/fr/apprentissage-automatique/nettoyage-des-donnees/(accès:08/08/2023)
[4] https://dataaspirant.com/data-normalization-techniques/(accès:08/08/2023)

4.1.2.1. Normalisation Min-Max

Cette méthode met les données à l'échelle dans une plage spécifique, généralement entre 0 et 1. Elle se calcule en soustrayant la valeur minimale et en divisant par la différence entre la valeur maximale et la valeur minimale.

4.1.2.2. Normalisation Z-score (Standardisation)

Cette méthode transforme les données pour avoir une moyenne de 0 et un écart type de 1. Elle se calcule en soustrayant la moyenne et en divisant par l'écart type.

4.1.2.3. Normalisation par décimale

Dans cette méthode, vous normalisez les données pour qu'elles aient un nombre fixe de décimales.

4.1.2.4. Normalisation par plage

Cette méthode met les données à l'échelle dans une plage spécifique en utilisant une formule particulière pour ajuster les valeurs.

4.1.3. L'encodage des variables

En apprentissage automatique, les variables catégoriques sont des caractéristiques non numériques qui représentent différents groupes ou catégories. Ces variables ne peuvent pas être utilisées directement dans la plupart des algorithmes d'apprentissage automatique car elles nécessitent des entrées numériques. Par conséquent, l'encodage des variables catégoriques est une étape de prétraitement cruciale pour convertir ces variables en un format numérique pouvant être efficacement utilisé par les modèles d'apprentissage automatique. Il existe plusieurs méthodes courantes pour l'encodage des variables catégoriques [5] (le choix de la méthode d'encodage dépend de la nature des données, du problème à résoudre et de l'algorithme prévoyé d'être utilisé):

4.1.3.1. Encodage par étiquetage « Label Encoding »

« LabelEncoder » est une classe de la bibliothèque scikit-learn en Python qui est utilisée pour encoder des étiquettes catégoriques (étiquettes textuelles) en valeurs numériques. « LabelEncoder » attribue un entier unique à chaque étiquette unique, convertissant ainsi les données catégoriques en un format numérique pouvant être utilisé par les algorithmes d'apprentissage automatique.

4.1.3.2. Encodage one-hot (OneHotEncoder)

[5] https://towardsdatascience.com/6-ways-to-encode-features-for-machine-learning-algorithms-21593f6238b0 (accès:08/08/2023)

Cette technique est utilisée pour convertir des variables catégorielles en vecteurs binaires. Chaque catégorie est transformée en une colonne binaire distincte. Cela est utile lorsque la variable catégorielle n'a pas de relation d'ordre. (get_dummies : Cette fonction de la bibliothèque Pandas est utilisée pour effectuer un encodage one-hot. Elle crée un nouveau tableau de données « DataFrame » avec des colonnes binaires pour chaque catégorie, en se basant sur une colonne catégorielle dans le « DataFrame » d'origine.).

4.1.3.3. OrdinalEncoder

Similaire à « LabelEncoder », il est utilisé pour encoder des variables catégorielles, mais il est spécifiquement conçu pour les catégories ordinales (catégories avec un ordre significatif).

4.1.3.4. Encodage cible (Target Encoding)

Cette technique consiste à remplacer les valeurs catégorielles par la moyenne (ou une autre agrégation) de la variable cible pour chaque catégorie. Elle est utile lorsque vous souhaitez intégrer des informations sur la cible dans l'encodage.

4.1.3.5. Encodage par fréquence (Frequency Encoding)

L'encodage par fréquence remplace chaque catégorie par la fréquence de cette catégorie dans l'ensemble de données. Il peut être utile lorsque la fréquence des catégories est liée à la variable cible.

4.1.3.6. Encodage binaire (Binary Encoding)

L'encodage binaire est un compromis entre le codage par étiquetage et l'encodage one-hot. Il convertit d'abord les catégories en entiers ordinaux, puis représente ces entiers au format binaire. Cette méthode réduit la dimensionnalité par rapport à l'encodage one-hot tout en préservant certains avantages de l'encodage one-hot.

4.1.3.7. Embedding

Couramment utilisées en traitement du langage naturel et en apprentissage profond, les techniques d'embedding représentent des variables catégorielles sous forme de vecteurs denses dans un espace continu, permettant aux modèles d'apprendre des relations entre les catégories.

4.1.3.8. DictVectorizer

Il s'agit d'une autre technique de scikit-learn pour convertir des variables catégorielles représentées sous forme de dictionnaires (telles que des structures de type JSON) en une matrice creuse de caractéristiques.

4.1.4. La gestion du déséquilibre des classes

La gestion du déséquilibre des classes est un aspect crucial lors de la création de modèles d'apprentissage automatique pour des problèmes où les classes ne sont pas représentées de manière équilibrée dans l'ensemble de données. Dans de tels cas, le modèle peut avoir tendance à favoriser la classe majoritaire, ce qui peut entraîner une performance médiocre pour la classe minoritaire. Pour surmonter ce problème, diverses méthodes de prétraitement peuvent être utilisées pour équilibrer les classes avant l'entraînement du modèle (Il est important de noter que le choix de la méthode dépend du problème spécifique et des données en question. Il peut être nécessaire d'expérimenter plusieurs approches pour déterminer celle qui fonctionne le mieux pour un cas d'utilisation spécifique).

4.1.4.1. Sous-échantillonnage « Undersampling »

Cette méthode implique la réduction du nombre d'échantillons dans la classe majoritaire afin de l'équilibrer avec la classe minoritaire. Cela peut aider à éviter la prédominance de la classe majoritaire. Cependant, le sous-échantillonnage[6] peut entraîner une perte d'informations potentiellement importantes.

4.1.4.2. Sur-échantillonnage « Oversampling »

Dans cette méthode, de nouveaux échantillons sont générés pour la classe minoritaire afin d'augmenter sa taille et de l'équilibrer avec la classe majoritaire. Cela peut se faire en dupliquant des échantillons existants ou en générant de nouvelles données synthétiques à l'aide de techniques comme SMOTE (Synthetic Minority Over-sampling Technique).

Le module imblearn.over_sampling de la bibliothèque imbalanced-learn de Python fournit diverses techniques de suréchantillonnage[7] de la classe minoritaire dans les ensembles de données déséquilibrés. Certaines des techniques de suréchantillonnage couramment utilisées et disponibles dans imblearn.over_sampling sont les suivantes[8] :

- RandomOverSampler : Cette technique duplique aléatoirement les instances de la classe minoritaire jusqu'à ce qu'elle atteigne un équilibre avec la classe majoritaire.
- SMOTE (Synthetic Minority Over-sampling Technique) : SMOTE génère des échantillons synthétiques pour la classe minoritaire en interpolant les vecteurs de caractéristiques entre les instances existantes de la classe minoritaire. Cette méthode crée des points de données synthétiques en considérant les k plus proches voisins de chaque instance de la classe minoritaire.

[6] https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html(accès:08/08/2023)
[7] https://towardsdatascience.com/class-imbalance-strategies-a-visual-guide-with-code-8bc8fae71e1a
[8] https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html

- ADASYN (échantillonnage synthétique adaptatif) : ADASYN est une extension de SMOTE qui introduit une approche basée sur la densité pour générer des échantillons synthétiques. Il s'agit de générer davantage d'échantillons synthétiques dans les régions de l'espace des caractéristiques où le déséquilibre entre les classes est le plus important.
- SVMSMOTE (Support Vector Machine SMOTE) : SVMSMOTE combine SMOTE avec un classifieur SVM pour générer des échantillons synthétiques. Il sélectionne les échantillons proches de la limite de décision pour créer des instances synthétiques plus informatives.
- KMeansSMOTE : KMeansSMOTE combine le regroupement K-means et SMOTE pour générer des échantillons synthétiques. Il regroupe les instances des classes minoritaires et applique SMOTE indépendamment dans chaque groupe.
- SMOTE-Tomek : SMOTE-Tomek est une combinaison de SMOTE et de liens Tomek. Il applique d'abord SMOTE pour suréchantillonner la classe minoritaire et supprime ensuite les liens Tomek, qui sont des paires d'instances de différentes classes proches les unes des autres et considérées comme bruyantes.

Le choix de la méthode de suréchantillonnage appropriée dépend de l'ensemble de données spécifique et des caractéristiques du problème à résoudre.

4.1.4.3. Pondération des classes

L'ajout de poids aux classes lors de l'entraînement du modèle peut aider à mettre davantage l'accent sur la classe minoritaire. Cela peut être réalisé en attribuant des poids différents aux différentes classes dans la fonction de perte du modèle.

4.1.4.4. Ensemble d'ensembles

Cette approche combine plusieurs modèles formés sur différents sous-ensembles de données, chacun équilibré différemment. Les prédictions de chaque modèle sont ensuite combinées pour produire une prédiction finale.

4.1.4.5. Méthodes basées sur les coûts

Ces méthodes ajustent les coûts associés aux erreurs de classification des différentes classes, de manière à pénaliser davantage les erreurs sur la classe minoritaire.

4.1.4.6. Utilisation de métriques d'évaluation appropriées

Plutôt que de se concentrer uniquement sur l'exactitude, il est souvent préférable d'utiliser des métriques telles que la précision, le rappel (taux de vrais positifs) et le F1-score qui tiennent compte du déséquilibre des classes.

4.1.4.7. Approches de transfert d'apprentissage

Dans certaines situations, il peut être bénéfique d'entraîner un modèle sur une tâche similaire avec des classes équilibrées, puis de transférer ce modèle sur le problème de déséquilibre des classes.

4.2. Sélection des caractéristiques « Feature Selection» et réduction de la dimensionnalité

Les techniques de sélection des caractéristiques et de réduction de la dimensionnalité jouent un rôle crucial dans l'identification des caractéristiques les plus informatives et pertinentes pour les modèles de prédiction du diabète. Des méthodes telles que l'analyse en composantes principales (Principal Component Analysis « PCA»), l'élimination récursive des caractéristiques (Recursive Feature Elimination « RFE ») et la régularisation lasso, qui permettent de réduire la dimensionnalité de l'espace d'entrée et d'améliorer les performances et l'interprétabilité du modèle [110], [111].

4.2.1. La PCA

L'Analyse en Composantes Principales, également connue sous le nom de (Principal Component Analysis « PCA ») en anglais, est une technique statistique sophistiquée largement utilisée dans le domaine de l'intelligence artificielle et de l'analyse de données. Elle représente un outil puissant de réduction de dimensionnalité en identifiant les combinaisons linéaires optimales, appelées composantes principales, au sein d'un ensemble de données donné. La PCA s'appuie sur des principes mathématiques avancés tels que l'algèbre linéaire, la théorie des valeurs propres et la décomposition en valeurs singulières [112].

Le processus de PCA consiste à transformer un ensemble de variables corrélées en un ensemble de variables non corrélées, représentées par les composantes principales. Ces composantes principales sont classées par ordre d'importance décroissante, la première composante capturant la variance la plus élevée des données, suivie de la deuxième composante, et ainsi de suite. Ainsi, la PCA permet de réduire la dimensionnalité des données tout en préservant les informations essentielles contenues dans l'ensemble initial de variables [110].

La PCA s'appuie sur des concepts mathématiques complexes, tels que les vecteurs propres et les valeurs propres. Les vecteurs propres représentent des directions dans l'espace des variables qui maximisent la variance des données, tandis que les valeurs propres correspondent aux quantités d'informations expliquées par chaque composante principale. En sélectionnant un sous-ensemble de composantes principales, il est possible de réduire efficacement la dimensionnalité des données tout en conservant

une partie substantielle de la variance et des motifs structurels présents dans l'ensemble de données d'origine [110], [113].

4.2.1.1. Avantages de la PCA

L'application de la PCA présente de nombreux avantages dans le contexte de la prédiction du diabète. En réduisant la dimensionnalité des données, la PCA permet de surmonter le fléau de la dimensionnalité, où un grand nombre de variables peut entraîner un surajustement et une capacité de généralisation réduite. De plus, la PCA peut identifier les combinaisons linéaires de variables les plus informatives pour la prédiction du diabète, facilitant ainsi la sélection de caractéristiques pertinentes. De plus, elle permet la visualisation des données dans un espace de dimension réduite, améliorant ainsi l'interprétation et la compréhension des relations entre les variables [113].

Cependant, il est important de noter que l'application de la PCA nécessite une préparation minutieuse des données, notamment la gestion des valeurs manquantes, la normalisation des variables et la prise en compte des corrélations non linéaires. De plus, l'interprétation des composantes principales peut parfois être complexe, notamment lorsque les variables d'origine présentent des corrélations fortes ou lorsque les contributions des variables à chaque composante principale sont équilibrées.

4.2.2. RFE

Recursive Feature Elimination (RFE), également connue sous le nom d'Élimination de Caractéristiques Récursive en français, est une technique avancée de sélection de caractéristiques qui a suscité beaucoup d'attention dans le domaine de l'apprentissage automatique et de l'analyse de données. RFE est une approche algorithmique puissante qui élimine systématiquement les caractéristiques non pertinentes ou redondantes d'un ensemble de données donné, dans le but d'améliorer les performances et l'interprétabilité du modèle [114].

RFE fonctionne selon un processus récursif en ajustant itérativement des modèles et en éliminant les caractéristiques ayant la plus faible importance ou contribution à la tâche de prédiction. Cette procédure itérative consiste à entraîner un modèle, évaluer l'importance de chaque caractéristique selon un critère prédéfini et éliminer la caractéristique la moins importante. Le processus est répété jusqu'à ce qu'un nombre spécifié de caractéristiques ou un seuil prédéterminé soit atteint [115].

Au cœur de RFE se trouve le concept de classement des caractéristiques, où les caractéristiques se voient attribuer des scores ou des classements d'importance en fonction de leur contribution aux performances du modèle. Ce classement est généralement dérivé des coefficients, des poids ou des mesures d'importance obtenus à partir de l'algorithme d'apprentissage automatique sous-jacent. En éliminant

itérativement les caractéristiques ayant des classements inférieurs, RFE se concentre progressivement sur les caractéristiques les plus informatives et discriminantes, réduisant ainsi la dimensionnalité des données tout en préservant les informations cruciales nécessaires aux prédictions précises [116], [117].

4.2.2.1. Avantage de RFE

L'utilisation de RFE présente plusieurs avantages dans le contexte de la sélection de caractéristiques pour la prédiction du diabète. En évaluant systématiquement l'importance des caractéristiques, RFE permet d'identifier les biomarqueurs pertinents, les variables cliniques ou les marqueurs génétiques qui contribuent significativement à la prédiction du diabète. De plus, RFE aide à réduire la dimensionnalité des données, à aborder le problème de la dimensionnalité et à atténuer le risque de surajustement. En se concentrant sur les caractéristiques les plus informatives, RFE améliore l'interprétabilité du modèle et sa capacité de généralisation, facilitant ainsi une meilleure compréhension des mécanismes sous-jacents au diabète [118].

Il est crucial de prendre en compte plusieurs facteurs pour une application réussie de RFE. Cela inclut la sélection soignée du critère d'évaluation, du modèle d'apprentissage automatique sous-jacent et du nombre optimal de caractéristiques à conserver. De plus, la présence de caractéristiques corrélées peut affecter la performance de RFE, car l'élimination d'une caractéristique peut impacter la capacité prédictive des autres caractéristiques associées. Une attention particulière doit donc être accordée à ces considérations pour garantir les meilleurs résultats lors de l'utilisation de la technique RFE.

4.2.3. La régularisation Lasso

La régularisation Lasso, également connue sous le nom de (Least Absolute Shrinkage and Selection Operator), est une technique avancée de régularisation qui a acquis une grande popularité dans le domaine de l'apprentissage automatique et de l'analyse des données. Elle offre une approche élégante pour gérer le problème de la sélection de variables dans les modèles prédictifs, en introduisant une pénalité qui favorise la sparsité et la parcimonie des coefficients [119].

La régularisation Lasso repose sur des principes mathématiques sophistiqués, tels que la norme L1 « L1 Regularization », qui mesure la somme des valeurs absolues des coefficients du modèle. Cette pénalité L1 contraint les coefficients à être nuls pour certaines variables, ce qui a pour effet d'effectuer une sélection automatique des caractéristiques les plus importantes [115]. Ainsi, la régularisation Lasso permet d'obtenir des modèles plus simples et plus interprétables en éliminant les caractéristiques moins significatives et en se concentrant sur les caractéristiques les plus prédictives [120].

4.2.3.1. Avantages de la régularisation Lasso

L'avantage majeur de la régularisation Lasso réside dans sa capacité à réaliser une sélection de variables automatique et à gérer efficacement les problèmes de surajustement « overfitting » et de multi-collinéarité. En favorisant la parcimonie des coefficients, la régularisation Lasso peut réduire la complexité du modèle, améliorer la généralisation des prédictions et faciliter l'interprétation des résultats. De plus, la régularisation Lasso est particulièrement utile lorsque les données contiennent un grand nombre de variables potentiellement redondantes ou non informatives.

Cependant, il est important de noter que le choix du paramètre de pénalité (alpha) dans la régularisation Lasso est crucial pour trouver le bon équilibre entre la réduction de la complexité du modèle et la préservation de l'information prédictive. De plus, la régularisation Lasso peut introduire un biais de sélection des variables, où certaines caractéristiques importantes peuvent être négligées ou sous-estimées en raison de la pénalité L1. Par conséquent, une analyse approfondie de l'impact de la régularisation Lasso sur les caractéristiques sélectionnées et sur les performances prédictives du modèle est nécessaire pour garantir des résultats fiables et interprétables [114], [115], [119].

4.2.4. D'autres techniques de sélection des caractéristiques

Le (Tableau 3.1) présente d'autres techniques couramment utilisées pour la sélection de caractéristiques en apprentissage automatique [121], [122] :

TABLEAU 3.1 : TECHNIQUES COURAMMENT UTILISÉES POUR LA SÉLECTION DE CARACTÉRISTIQUES EN ML.

Technique	**Description**
« Score-based methods »	Méthodes basées sur les scores: • Utiliser des techniques statistiques telles que *le test du chi carré « chi-square»*, *l'analyse de variance* (ANOVA) ou *les coefficients de corrélation* pour évaluer l'importance de chaque caractéristique. • Classer les caractéristiques en fonction de leurs scores et sélectionner les meilleures.
« Model-based methods »	Méthodes basées sur les modèles : • Entraîner un modèle d'apprentissage automatique sur l'ensemble des données. • Évaluer l'importance de chaque caractéristique en utilisant les coefficients ou poids du modèle. • Supprimer les caractéristiques ayant une faible importance selon l'évaluation du modèle.
« Tree-based feature selection »	Sélection basée sur les arbres de décision :

	• Évaluer l'importance des caractéristiques en utilisant *des arbres de décision* basés sur la manière dont ils divisent les nœuds de l'arbre.
« Optimization-based selection »	Sélection basée sur l'optimisation : • Utiliser des algorithmes d'optimisation tels que *les algorithmes génétiques* ou *l'optimisation par essaim de particules* « particle *swarm optimization* » pour trouver le sous-ensemble optimal de caractéristiques qui maximise les performances du modèle.
« Univariate Feature Selection»	Sélection univariée de caractéristiques : • Sélectionner les caractéristiques en fonction de tests statistiques univariés entre chaque caractéristique et la variable cible, tels que *SelectKBest* et *SelectPercentile*.
« Recursive Feature Addition »	Ajout récursif de caractéristiques : • Commencer avec un ensemble vide de caractéristiques et ajouter itérativement une caractéristique à la fois, en sélectionnant celle qui apporte la meilleure amélioration des performances.
« Stability Selection»	Sélection par stabilité : • Utiliser une approche basée sur un ensemble en entraînant plusieurs modèles sur des sous-ensembles d'échantillons générés aléatoirement et en sélectionnant les caractéristiques qui sont choisies de manière cohérente parmi ces modèles.
« Mutual Information»	Information mutuelle : • Mesurer la dépendance entre les variables et évaluer la pertinence des caractéristiques.

Lors de l'application des méthodes de sélection de caractéristiques, il faut tenir compte de la nature du problème et de la taille de l'ensemble de données. Une sélection appropriée des caractéristiques peut considérablement améliorer l'efficacité et les performances des modèles d'apprentissage automatique en éliminant les informations redondantes ou moins utiles [123], [124].

4.2.5. Envisagement de la sélection des caractéristiques

La sélection des caractéristiques[9] doit être envisagée dans les scénarios suivants :

- **Données à haute dimension :** Lorsque l'on traite des ensembles de données comportant un grand nombre de caractéristiques par rapport au nombre d'instances (données à haute dimension), la sélection des caractéristiques devient cruciale. Les données à haute dimension peuvent conduire à un surajustement, à une augmentation des coûts de calcul et à une diminution de l'interprétabilité du modèle. En sélectionnant les caractéristiques pertinentes, nous pouvons réduire la dimensionnalité et améliorer l'efficacité du modèle.

- **Caractéristiques non pertinentes ou redondantes :** Si l'ensemble de données contient des caractéristiques non pertinentes ou redondantes pour la tâche de prédiction, leur inclusion peut avoir un impact négatif sur les performances du modèle. La sélection des caractéristiques permet d'exclure ces caractéristiques moins utiles, ce qui permet d'obtenir un modèle plus concis et plus informatif.

- **Efficacité informatique :** Dans les environnements à grande échelle ou à ressources limitées, la sélection des caractéristiques peut être utilisée pour réduire la charge de calcul. Moins de caractéristiques signifie des temps de formation et d'inférence plus rapides.

- **Interprétabilité du modèle :** Dans certaines applications, l'interprétabilité du modèle est cruciale. En sélectionnant un ensemble plus restreint de caractéristiques, le modèle obtenu devient plus facile à interpréter et à expliquer aux parties prenantes.

- **Réduction du bruit :** Certaines caractéristiques peuvent contenir du bruit ou des erreurs susceptibles d'affecter négativement les performances du modèle. La sélection des caractéristiques peut contribuer à réduire l'impact des caractéristiques bruyantes.

- **Multicollinéarité :** Lorsque les caractéristiques sont fortement corrélées « multicollinearity », les coefficients du modèle peuvent être instables et il est difficile d'interpréter les contributions individuelles des caractéristiques. La sélection des caractéristiques peut contribuer à résoudre ce problème en ne conservant qu'une caractéristique représentative des groupes corrélés.

Toutefois, dans certains cas, la sélection des caractéristiques peut ne pas être nécessaire, voire bénéfique :

- **Petit ensemble de données :** Si nous disposons d'un petit ensemble de données, la suppression de caractéristiques peut entraîner une perte d'informations et le risque de

[9] https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning (accès: 08/08/2023)

surajustement peut être moins important. Dans ce cas, la sélection des caractéristiques doit être abordée avec prudence.

- **Importance des caractéristiques :** Certains algorithmes d'apprentissage automatique, tels que les modèles basés sur des arbres (par exemple, Random Forest, Gradient Boosting), fournissent intrinsèquement des scores d'importance des caractéristiques. Si l'algorithme peut gérer l'analyse de l'importance des caractéristiques, nous n'aurons peut-être pas besoin d'une étape distincte de sélection des caractéristiques. NB : (L'importance peut également être évaluée en examinant les corrélations entre les caractéristiques et la variable cible)

- **Connaissance du domaine :** Dans certains cas, la connaissance du domaine peut guider la sélection des caractéristiques pertinentes. Si nous somme certain que toutes les caractéristiques sont significatives et pertinentes pour la tâche de prédiction, la sélection des caractéristiques peut ne pas être nécessaire.

4.3. Approches d'apprentissage supervisé

Les méthodes d'apprentissage supervisé sont largement utilisées dans les tâches de prédiction du diabète. Ces approches consistent à entraîner un modèle sur des données étiquetées, où chaque instance est associée à un résultat connu (par exemple, un diagnostic de diabète). Des algorithmes d'apprentissage supervisé populaires tels que la régression logistique, les arbres de décision, les machines à vecteurs de support (SVM) et les forêts aléatoires ont été utilisés pour développer des modèles prédictifs basés sur des variables cliniques, des marqueurs génétiques, des facteurs de mode de vie et d'autres caractéristiques pertinentes [125]–[127]. Nous verrons les performances de ces algorithmes dans le contexte de la prédiction du diabète.

4.4. Techniques d'apprentissage profond

L'apprentissage profond, une sous-catégorie de l'apprentissage automatique, s'est imposé comme une approche puissante dans divers domaines, y compris la prédiction du diabète. Les réseaux neuronaux profonds, avec leur capacité à apprendre des schémas complexes à partir de données de grande dimension, ont donné des résultats prometteurs dans la capture de relations complexes et la détection d'indicateurs subtils du diabète. Les réseaux neuronaux convolutifs (CNN) et les réseaux neuronaux récurrents (RNN) ont été appliqués à différentes modalités de données, telles que les images médicales, les données de séries chronologiques et les dossiers médicaux électroniques (DME) [128], [129]. Nous explorons les applications des techniques d'apprentissage profond dans la prédiction du diabète et discutons des défis liés à l'interprétabilité des modèles et aux exigences en matière de données.

4.5. Méthodes d'ensemble et empilement de modèles

Les méthodes d'ensemble combinent plusieurs modèles prédictifs pour améliorer les performances globales et renforcer la robustesse de la prédiction du diabète. Des techniques

telles que le bagging, le boosting et le stacking sont utilisées pour créer des ensembles qui tirent parti de la diversité et des forces complémentaires des modèles individuels. Les méthodes d'ensemble peuvent atténuer le surajustement, réduire le biais et améliorer la capacité de généralisation des modèles de prédiction du diabète [130], [131]. Nous verrons de l'application des méthodes d'ensemble dans la prédiction du diabète ainsi que leur efficacité pour améliorer la précision et la stabilité prédictives.

4.5.1. Le bagging

Le bagging consiste à former plusieurs apprenants (par exemple, des arbres de décision) indépendamment sur différents sous-ensembles des données de formation, puis à combiner leurs prédictions par le biais d'un vote (dans la classification) ou d'un calcul de moyenne (dans la régression).

Chaque apprenant est formé sur un sous-ensemble de données de formation échantillonné de manière aléatoire, avec remplacement. Cela signifie que certaines instances peuvent apparaître plusieurs fois dans un sous-ensemble, tandis que d'autres peuvent ne pas apparaître du tout.

Le bagging vise à réduire la variance et à améliorer la généralisation du modèle en établissant une moyenne des erreurs et en réduisant le surajustement. Il fonctionne bien avec les modèles instables qui peuvent être sensibles à de petits changements dans les données d'apprentissage [132].

4.5.2. Le boosting

Le boosting, quant à lui, est une technique d'apprentissage itératif dans laquelle des modèles faibles successifs sont construits en accordant une attention accrue aux échantillons mal prédits par les modèles précédents. Les modèles faibles sont ensuite combinés pour former un modèle fort capable de capturer des relations complexes dans les données. Chaque modèle faible contribue à la prédiction finale en fonction de sa performance et une attention particulière est accordée aux échantillons difficiles à prédire.

Le boosting consiste à former plusieurs apprenants (par exemple, des arbres de décision) de manière séquentielle, chaque apprenant corrigeant les erreurs de l'apprenant précédent. Il attribue des poids aux instances d'apprentissage, en donnant plus de poids aux instances qui ont été mal classées par les apprenants précédents. Cet accent mis sur les instances difficiles à classer permet d'améliorer l’exactitude et la précision du modèle. Chaque apprenant est formé sur l'ensemble de données modifié, où les instances sont repondérées en fonction de leur performance de classification par les apprenants précédents.

Le boosting vise à réduire les biais et à améliorer l’exactitude du modèle en se concentrant sur les instances mal classées et en construisant un apprenant fort à partir d'apprenants faibles[133], [134].

4.5.3. Le stacking

Le stacking, également connu sous le nom de (stacked generalization), est une approche qui combine les prédictions de plusieurs modèles en utilisant un modèle de niveau supérieur, souvent appelé méta-modèle ou modèle d'agrégation. Les prédictions des modèles de niveau inférieur sont utilisées comme caractéristiques d'entrée pour le modèle de niveau supérieur, qui apprend à agréger ces prédictions pour produire la prédiction finale. Le stacking permet de tirer parti des forces de chaque modèle de niveau inférieur et d'obtenir des prédictions plus précises et plus fiables [135].

Ces techniques d'ensemble, le bagging, le boosting et le stacking, offrent des avantages significatifs en termes d'amélioration des performances de prédiction et de réduction de la variance. Elles permettent également de prendre en compte une plus grande diversité de modèles et de capturer des relations plus complexes dans les données. Cependant, elles nécessitent une attention particulière lors de la sélection des modèles de base, de la gestion des hyperparamètres et de l'évaluation des performances pour éviter le surapprentissage et maximiser les bénéfices de l'ensemble [136], [137].

4.6. Métriques d'évaluation et évaluation des performances

L'évaluation des modèles prédictifs dans la prédiction du diabète nécessite des métriques d'évaluation appropriées pour mesurer leurs performances. Les métriques courantes incluent : exactitude « accuracy », sensibilité « sensitivity», spécificité «specificity », precision, rappel « recall », F1-score, MCC « matthews correlation coefficient » (voir Tableau 4.6 pour les formules).

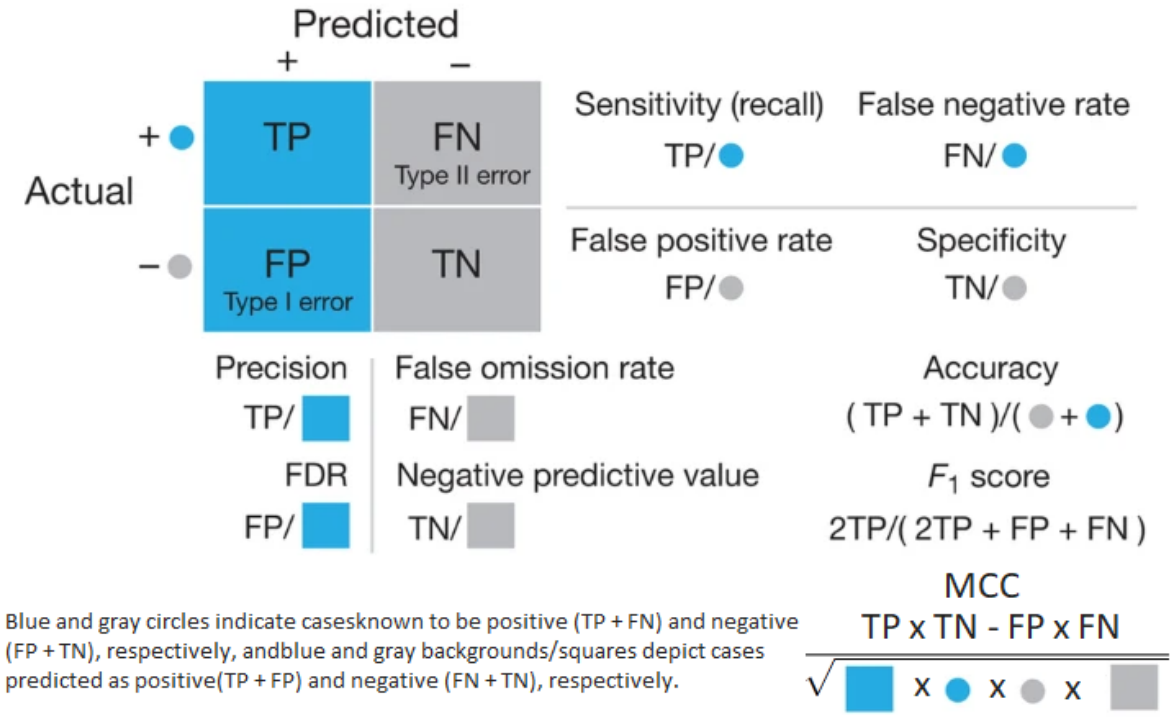

FIGURE 3.1 : LES MÉTRIQUES D'ÉVALUATION LES PLUS UTILISÉES DES MODÈLES PRÉDICTIFS.

Dans la (Figure 3.1), les équations de chaque mesure sont exprimées graphiquement en termes de valeurs dans la matrice de confusion. Les nombres de prévisions correctes et incorrectes effectuées à partir de données connues sont affichés dans le CM.

Ces métriques ont une grande importance pour évaluer l'efficacité des différentes approches d'apprentissage automatique tous en n'oubliant pas l'importance des techniques de validation et de validation croisée « cross-validation » comme la validation croisée k-fold, pour garantir la fiabilité et la généralisabilité des modèles.

5. Analyse comparative des études existantes

Dans les sections qui suivent, nous examinerons en détail les études récentes pertinentes sur la prédiction du diabète. Nous nous attarderons sur ces recherches afin de mieux comprendre les avancées dans le domaine de la prédiction de cette maladie chronique. En analysant attentivement les études sélectionnées, nous serons en mesure d'explorer les différentes approches, méthodologies et ensembles de données utilisés pour prédire le diabète. En examinant ces travaux de recherche, nous pourrons identifier les facteurs prédictifs les plus prometteurs et évaluer l'efficacité des différentes méthodes de prédiction. En somme, cette exploration approfondie des études récentes nous permettra de mettre en évidence les avancées significatives dans le domaine de la prédiction du diabète et de jeter les bases pour des recherches futures visant à améliorer les outils de diagnostic précoce.

Pour faciliter une compréhension complète, des tableaux résumant les informations clés de chaque étude sélectionnée seront présentés, permettant une analyse systématique et comparative.

5.1. Traveaux sur ML pour la prédiction du diabète

Nous présentons une synthèse des travaux récents portant sur l'utilisation de l'apprentissage automatique pour la prédiction du diabète. L'objectif est de fournir un aperçu global des études menées dans ce domaine et de résumer leurs principales caractéristiques dans un tableau. Les travaux de recherche examinés mettent en évidence l'application de diverses techniques de ML (Tableau 3.2).

TABLEAU 3.2 : RÉSUMÉ DES PAPIERS SUR LA PRÉDICTION DU DIABÈTE AVEC ML.

Papier	Dataset/Nombre	Algorithmes	Valeurs
Prediction of gestational diabetes in the first trimester using machine learning-based methods [138].	1443 femmes ont été incluses, et 86 femmes (5,96%) ont été diagnostiquées comme ayant un diabète gestationnel avec 11 attributs.	**Random Forest** Logistic Regression Support Vector Machine Deep Neural Network	**78.10%** 74.00% 75.60% 74.50% (AUC)
Population-centric risk prediction modeling for gestational diabetes	Les données de 909 grossesses de l'étude de cohorte mère-enfant la plus	**CatBoost Gradient Boosting**	**82.00%** (AUC)

mellitus: A machine learning approach [139].	profondément phénotypée de Singapour, Growing Up in Singapore Towards healthy Outcomes (GUSTO). (5fold stratified CV)		
A comparison of machine learning algorithms for diabetes prediction [140].	PIMA (L'ensemble de données contient des informations sur 768 patients et leurs 9 attributs uniques correspondants avec 268 échantillons sont classés diabétiques et 500 non-diabétiques.). (Utilisation séparée de la méthode de validation croisée K-fold et la méthode de division train/test à 85%)	DT(K-fold) DT(Division) RF(K-fold) RF(Division) NB(K-fold) NB(Division) LR(K-fold) LR(Division) KNN(K-fold) KNN(Division) k=7 SVM(K-fold) SVM(Division) **Artificial Neural Network** (2 hidden layers)	74.24% 73.14% 74.96% 77.14% 75.53% 78.28% 76.82% 78.85% 75.10% 79.42% 76.82% 77.71% **88.60%** (Accuracy)
A novel hybrid machine learning framework for the prediction of diabetes with context-customized regularization and prediction procedures [141].	PIMA (6-fold nested cross validation)	**Modèle hybride personnalisé de réseau neuronal artificiel**	**80%** (Accuracy)
Machine learning based diabetes prediction and development of smart web application [142].	(Deux datasets: 1ere PIMA+2eme (Nbr d'enregistrements 950, Nbr d'attributs 19) Division 80:20)	<u>Dataset1 :</u> NB **DT** **RF** SVM LR GB KNN <u>Dataset2 :</u> NB DT **RF** **SVM** LR GB KNN	(Accuracy) 86.17% **96.81%** **96.81%** 91.49% 84.04% 91.00% 90.43% 78.95% 76.32% **80.26%** **80.26%** 77.63% 78.95% 75.00%
Predicting the onset of type 2 diabetes using	L'ensemble de données comprend les dossiers	**Wide Deep Classifier**	**84.28%**

wide and deep learning with electronic health records [143].	médicaux électroniques de 9948 patients, dont 1904 ont été diagnostiqués avec un DT2. (70% entrainement & validation après faire 10 fold CV+30% test)		(Accuracy)
An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators [144].	L'ensemble de données est un sous-ensemble des données du Behavioral Risk Factor Surveillance System (BRFSS). Données propres de 70 692 réponses à l'enquête BRFSS2015. Cet ensemble de données compte 253 680 enregistrements et vingt-et-une variables de caractéristiques non équilibrées. (80:20 training & testing) A Principal Component Analysis (PCA) a été appliquée pour réduire la dimensionnalité de l'ensemble de données. Synthetic Minority Oversampling Techniques (SMOTE) ont été utilisées pour stabiliser le déséquilibre de la variable de sortie.	**RF** DT KNN LR NB	**82.26%** 81.02% 80.55% 72.64% 70.56% (Accuracy)
Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus [145].	SDHD :Sylhet Diabetes Hospital Dataset. Connais aussi comme Early-stage diabetes risk prediction (520 dossiers de patients+17 attributs) provenant du dépôt d'apprentissage automatique de l'Université de Californie, Irvine (UCI) de l'Hôpital	**C4.5 DT**	**73.5%** (Accuracy)

	du diabète de Sylhet.		
A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques [146].	PIMA (60 :10 après faire 10fold CV sur trainset)	**RF** SVM DL	**83.67%** 65.38% 76.81% (Accuracy)
Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques [147].	PIMA avec SMOTE	RF LR SVM KNN GB NB **Voting Classifier (3 meilleurs algorithmes)**	80.70% 76.70% 77.20% 77.50% 77.90% 73.90% **81.50%** (Accuracy)
A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques [148].	PIMA (90 :10 train,test)	**LR (Hyper-parameter tuning)**	**83.00%** (Accuracy)
Primary Stage of Diabetes Prediction using Machine Learning Approaches Minhaz [149].	SDHD (75 :25 train test)	**LR**	**92%** (Accuracy)
A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM [146].	Saudi Arabian dataset 897 records -avec et sans (SMOTE, bagging , stacking)	<u>Experiment1 :</u> SVM **K-NN** DT Experiment2 : SVM **K-NN** DT <u>Experiment3 : (Bagging ensemble)</u> SVM Bagging **K-NN Bagging** DT Bagging	<u>Sans sampling:</u> 85.84% **85.95%** 84.50% <u>Avec sampling:</u> 90.83% **93.11%** 90.56% <u>Avec sampling:</u> 90.70% **94.34%** 94.12%

		Experiment4 : (Stacking ensemble) **Stacking**(bagging KNN, bagging DT, KNN)	Avec sampling: **94.48%** (Accuracy)

5.2. Traveaux sur les avancées et tendances récentes d'IA pour la prédiction du diabète

Nous abordons les avancées et les tendances récentes dans le domaine de l'intelligence artificielle (IA) pour la prédiction du diabète. L'objectif est de fournir un aperçu des travaux de recherche les plus récents et de résumer leurs principales caractéristiques dans le (Tableau 3.3) suivant :

TABLEAU 3.3 : RÉSUMÉ DES PAPIERS SUR LES TENDANCES D'IA POUR LA PRÉDICTION DU DIABÈTE.

Papier	Dataset/Nombre	Algorithmes	Valeurs
An Ensemble Deep Learning Method for Diabetes Mellitus [150].	PIMA	**Esemble Multi-layer perceptron (EB-MLP) hybrid classifier**	**92.30%** (Accuracy)
Diabetes Mellitus Prediction Using Ensemble Learning Approach with Hyper parameterization [151].	PIMA	NB + RF **NB + RF + XGBOOST**	82.50% **85.60%** (Accuracy)
An Improved Ridge Regression-Based Extreme Learning Machine for the Prediction of Diabetes [152].	PIMA	**Ridge Extreme learning machine (RELM) with Fire Fly (FF) Algorithm**	**93.44%** (Accuracy)
Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers [153].	PIMA	KNN DT RF NB AdaBoost XGBoost MLP **AdaBoost + XGBoost**	92.60% 91.20% 93.90% 87.90% 94.10% 94.60% 90.20% **95.00%** (Accuracy)
Deep learning for predicting the onset of type 2 diabetes: enhanced ensemble classifier using modified t-SNE [154].	3 jeux de données : PIMA, Polarity(2000), et Luzhou(6000) Avec t-SNE embedding technique.	**Wide and Deep Algorithm**	**85.34%** (Accuracy)

A hybrid super ensemble learning model for the early-stage prediction of diabetes risk [155].	3 jeux de données : 1)Early-stage diabetes risk prediction(520 patients+17 attributs) 2)PIMA 3)Diabetes 130-US hospitals dataset(100,000 patients+50 attributs) (70:30 train test sur la base de la technique d'exclusion. Ensuite, la technique de validation croisée 10-fold sur trainset) Feature selection technique (Chi-square) Hyper-parameter settings (GridSearch)	**Super Learner Model avec 4 base-learners (logistic regression, decision tree, random forest, gradient boosting) & meta learner (SVM)**	1èr: **99.60%** 2ème: **92.00%** 3ème: **98.00%** (Accuracy)
Type-2 Diabetes Mellitus Diagnosis from Time Series Clinical Data Using Deep Learning Models [156].	King Abdul- lah International Research Centre Diabetes (KAIMRCD) qui comprend plus de 14 000 données de patients.	**Long Short-Term Memory (LSTM) & Gated-Recurrent Unit (GRU)**	**97.00%** (Accuracy)
Exploratory Study on Direct Prediction of Diabetes Using Deep Residual Networks [157].	Retinal images (comprend 8924 images de bonne qualité des yeux gauche et droit de 2336 sujets) Division 80:20%	**Deep Residual Network**	**75.80%** (F1-score)
Deep learning based big medical data analytic model for diabetes complication prediction [94].	Les données sont collectées à partir de différents référentiels diabétiques. Le référentiel diabétique comprend 50 000 ensembles de données.	**Deep Belief Network**	**80.99%** (Accuracy)
Prediction of gestational diabetes using deep learning and Bayesian optimization and traditional machine learning techniques [158].	Les données ont été collectées de manière prospective par trois médecins spécialistes de la santé publique (489 patients & 73 variables)	**RNN-LSTM with Bayesian optimization**	**98.00%** (AUC)
A novel solution of deep learning for enhanced support vector machine for predicting the onset of type 2 diabetes [159].	L'ensemble de données de la Federazione Italiana Medici DI Medicina Generale (FIMMG) est obtenu auprès de Metmedica Italia (NMI).	**Support Vector Machine (SVM) algorithm+ Radial Base Function (RBF) along + Long Short-term Memory**	86.31% (Accuracy) 82.70% (AUC)

	1862 features (Division 70 :30)	**Layer (LSTM)**	
Deep convolutional neural network for diabetes mellitus prediction [160].	PIMA (SMOTE utilisée)	**DCNN classifier**	**86.29%** (Accuracy)

5.3. Défis actuels dans la prédiction du diabète avec l'IA

Le domaine de la prédiction du diabète avec l'intelligence artificielle (IA) englobe une pléthore de méthodes actuelles et présente une multitude de défis qui exigent une attention méticuleuse. L'utilisation des techniques d'IA, telles que l'apprentissage automatique et l'exploration de données, offre une immense promesse en révélant des informations précieuses et des schémas à partir de vastes ensembles de données liées au diabète. Ces méthodes facilitent le développement de modèles prédictifs capables de discerner des relations complexes et d'identifier des facteurs de risque significatifs qui contribuent à l'apparition et à la progression de ce trouble métabolique chronique. Cependant, malgré les avancées considérables réalisées jusqu'à présent, plusieurs défis redoutables persistent dans le domaine de la prédiction du diabète avec l'IA. Ces défis englobent :

- des dimensions complexes, notamment la sélection et la curation de jeux de données diversifiés et représentatifs, (comme PIMA a que les données des femmes aussi il existe des jeux de données qui ne prennent pas en considération « lifestyle » du patient).

- l'optimisation des techniques de sélection de caractéristiques et de réduction de la dimensionnalité

- l'établissement de modèles prédictifs robustes et interprétables

- l'intégration de sources de données hétérogènes

- l'atténuation des biais algorithmiques

- ainsi que la validation et la généralisabilité des modèles proposés.

Pour relever ces défis, il est nécessaire de comprendre de manière exhaustive les interactions complexes entre les différentes subtilités méthodologiques, les avancées technologiques et les connaissances propres au domaine. De plus, la nature dynamique du diabète, caractérisée par son étiologie multifactorielle et ses manifestations cliniques évolutives, exige le raffinement et l'adaptation continus des approches prédictives basées sur l'IA pour prendre en compte de nouveaux facteurs de risque, les comorbidités émergentes et les évolutions du paysage des soins de santé.

6. Conclusion

Contemplez, une ère nouvelle de la médecine pointe à l'horizon, où la suprématie de l'IA et des données massives pourrait conduire à une efficacité accrue des traitements, une rentabilité accrue grâce à l'automatisation de certaines tâches, ainsi qu'à la découverte de nouveaux chemins pour la prestation de soins et la prévention. Dans cette optique, des algorithmes autonomes traitant des problèmes de plus en plus complexes peuvent offrir un soutien précieux à la prise de décision des professionnels de la santé, y compris ceux liés au diagnostic et au pronostic. De plus, lorsqu'ils sont intégrés à des robots de soins ou d'autres outils tels que des applications mobiles, ces algorithmes peuvent même établir un "rapport" direct avec les patients.

Néanmoins, l'avènement de ces technologies est accompagné de changements qui semblent rompre avec le système de soins et de recherche en santé traditionnel. Cette transformation dans le domaine de la santé présente alors des défis sur lesquels il est nécessaire de se pencher dedans. Les prochains chapitres nous permettront d'observer comment l'apprentissage automatique peut générer des résultats efficaces.

Chapitre 4

Propositions et évaluation (partie 1)

1. Introduction

Le domaine de ML est en constante évolution, avec de nouvelles avancées et propositions qui voient le jour régulièrement. Ce chapitre, intitulé "Propositions et évaluation (partie 1)", se penche sur la plateforme et le langage de programmation avec les bibliothèques qui sous-tendent nos travaux. Ainsi que nous explorerons en détail deux contributions majeures menées sur le jeu de données PIMA, chacune apportant une contribution significative à l'état de l'art sur ce jeu de données et par rapport à la prédiction de diabète.

2. Plateforme et langage de programmation

2.1. Google Colab

Nous avons décidé de tirer parti des potentialités de calcul en infonuagique grâce à la plateforme "Google Colab". Cette plateforme se présente comme une ouverture vers des calculs accélérés via des unités de traitement graphique (GPU), des unités de traitement tensoriel (TPU) et même des unités centrales de traitement (CPU). L'un de ses avantages prépondérants réside dans sa gratuité, bien qu'elle propose des sessions en ligne de durée limitée. "Google Colab" s'intègre harmonieusement dans l'écosystème de l'environnement "Jupyter Notebook", offrant ainsi une facilité accrue pour la mise en œuvre de diverses routines de programmation au moyen du langage Python.

2.2. Bibliothèques Python

Des modèles de détection précis et efficaces sont possibles grâce à l'environnement de configuration pour l'apprentissage automatique de la prédiction du diabète à l'aide de Python. Les chercheurs et les praticiens peuvent créer un cadre fiable pour la prédiction du diabète en utilisant Python, un langage de programmation flexible, et son énorme écosystème de progiciels d'apprentissage automatique. L'installation de Python est la première étape, suivie de l'installation des paquets nécessaires tels que scikit-learn, pandas et matplotlib, qui offrent des fonctions cruciales pour la manipulation, l'analyse et la visualisation des données. La création d'un environnement virtuel garantit une configuration réglementée et isolée, ce qui permet une meilleure gestion des dépendances. Pour assurer la qualité des données et améliorer les performances du modèle, le jeu de données doit être correctement préparé, y compris les procédures de prétraitement telles que la gestion des valeurs manquantes et la mise à l'échelle des caractéristiques « features ». Le (Tableau 4.1) présente les librairies utilisées[10] :

TABLEAU 4.1 : LIBRAIRIES UTILISÉES.

Librairie	Description	Caractéristiques	Avantages
Scikit Learn	Pour développement	-Aide à l'exploration de	-Une documentation

[10] https://mobiskill.fr/blog/conseils-emploi-tech/les-bibliotheque-python-a-utiliser-pour-le-machine-learning/(accès: 20/08/2023)

	ML en python.	données et à leurs analyses. -Il fournit des modèles et des algorithmes pour la classification, la régression, le clustering, la réduction dimensionnelle, et le prétraitement.	facilement compréhensible est fournie. -Les paramètres de tout algorithme spécifique peuvent être modifiés lors de l'appel d'objets.
NumPy	Une bibliothèque Python pour le calcul scientifique largement utilisée dans le domaine de ML.	-Il fournit un objet tableau (array) multidimensionnel hautes performances, ainsi que des outils pour travailler avec ces tableaux. -Utilisé pour stocker et manipuler de grandes quantités de données utilisées comme entrées dans des modèles d'apprentissage automatique. -Ainsi que pour effectuer des opérations mathématiques sur ces données afin de les préparer à être utilisées dans des algorithmes de ML.	-L'utilisation de NumPy offre : -Des performances optimisées. -Des fonctionnalités mathématiques avancées. -Une intégration avec d'autres bibliothèques scientifiques et facilite le traitement des données en Python.
Matplotlib	Bibliothèque complète pour créer des visualisations statiques, animées et interactives en Python.	-Offre une excellente visualisation des données statiques : diagrammes à barres, diagrammes dispersés, graphiques, etc. -Construit des modèles ML fiables : plusieurs tracés permettent une analyse approfondie des données, ce qui garantit en outre que les développeurs disposent de suffisamment de données pertinentes pour créer des modèles ML fiables.	- Ne nécessite que quelques lignes de code pour générer un tracé pour les ensembles de données. -Personnalisable et extensible grâce à de nombreuses fonctionnalités et options de configuration. -Large gamme d'outils de traçage : à l'aide de la bibliothèque Matplotlib, il est possible de tracer divers graphiques 2D, diagrammes 3D, histogrammes, graphiques d'erreurs, histogrammes et graphiques. Il permet aux experts d'effectuer une analyse détaillée des données.

Pandas	Pandas est un outil d'analyse et de manipulation de données open source rapide, construit sur Python.	Aider à de multiples tâches de traitement de données : regroupement par syntaxe, combinaison de données avec d'autres blocs de données, calcul de la corrélation des colonnes, fourniture de calculs de fenêtre glissante, etc.	-Puissant, flexible et facile à utiliser (Manipulation facile des ensembles de données : utile aux professionnels qui souhaitent gérer (structurer, trier, remodeler, filtrer) de grands ensembles de données avec facilité.) --Il est considéré comme l'une des normes modernes de représentation des données multidimensionnelles. -Capacités d'exploration de données : c'est un bon outil pour vérifier les corrélations dans les données et peut être utilisé pour nettoyer les données avant d'appliquer un modèle.
Seaborn	-Basée sur Matplotlib (qui se concentre sur le traçage et la visualisation des données) mais présente les structures de données de Pandas. -Graphiques plus esthétiques.	-Il offre des fonctionnalités intégrées pour tracer des graphiques de distribution, des graphiques de densité & de régression, des diagrammes en boîte, des matrices de corrélation, etc. -Ces fonctionnalités permettent d'explorer et d'analyser les données.	-Facile. -Offre des fonctionnalités avancées de visualisation de données statistiques. -Personnalisation des graphiques. -Visualisation des modèles statistiques. -Documentation et communauté active.

3. Contribution -1- (Modèle d'ensemble : Stacking)

La démarche proposée est la suivante :

• Étape I : Rassemblez l'ensemble de données sur le diabète et analysez-le.

• Étape II : Prétraiter les données.

• Étape III : Divisez les données en deux ensembles : formation et test.

• Étape IV : Développer un modèle d'ensemble

─ Niveau 0 : Stack KNN, SVM et Decision Tree dans la couche de base.

─ Niveau 1 : Créer un modèle de régression logistique dans la méta-couche en fonction des résultats de la couche de base.

• Étape V : À l'aide des équations (1) et (2), trouvez le résultat prédictif à l'aide de l'ensemble de tests sur le modèle.

Ensemble_Model.fit(train_x_data,train_y_data)…(1)
Predictive_Outcome=model.predict(test_x_data)…(2)

• Étape VI : déterminer l'exactitude (accuracy) du modèle.

3.1. Spécification du jeu de données

Sur le site Web de Kaggle, l'ensemble de données PIMA sur le diabète peut être trouvé[11]. Cet ensemble de données Pima Indian Diabetes Database (PIDD), qui provient de l'Institut national du diabète et des maladies digestives et rénales, peut être utilisé pour prédire de manière diagnostique si un patient est diabétique ou non sur la base de certaines mesures de diagnostic fournies dans la collection. Il se compose de nombreux paramètres médicaux et d'un paramètre dépendant à valeur binaire (Résultat). Une chose à noter est que tous les patients ici sont des femmes d'au moins 21 ans d'origine indienne Pima. Il y a 768 lignes et 9 colonnes dans cet ensemble de données, nous avons "Résultat" comme variable cible. Il y a neuf attributs dans chaque ligne, tels que :

TABLEAU 4.2 : DESCRIPTION ET TYPE DES ATTRIBUTS DU JEU DE DONNÉES.

N° d'attribut	Description d'attribut	Type de variable
1	Grossesses (Nombre de fois enceinte).	Entier
2	Glucose (Concentration de glucose plasmatique après une durée de 2 heures lors d'un test de tolérance au glucose oral).	Réel
3	Pression artérielle (Pression artérielle diastolique/systolique (mm Hg)).	Réel
4	Épaisseur du pli cutané du triceps[12]	Réel
5	Insuline (Insuline sérique après une durée de 2 heures (mu U/ml)).	Réel
6	IMC (indice de masse corporelle : poids en kg/(taille en m)^2).	Réel
7	Fonction de pedigree du diabète.	Réel
8	Âge (années).	Entier
9	Résultat (avec diabète (1) ou non (0)).	Binaire

[11] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (accessed Jul. 20, 2023)
[12] La mesure normale est compris entre 3 et 5mm, au-delà il y a trop de graisse.

Ces colonnes représentent des conditions médicales spécifiques, et voici un aperçu du jeu de données :

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns

FIGURE 4.1 : CAPTURE DE PIMA [161].

3.2. Visualisation des données

Nos données sont enregistrées dans un fichier CSV, qui doit être importé dans le notebook sous forme de tableau de données à l'aide du module Python Pandas. Après avoir importé les données, nous pouvons effectuer de nombreuses analyses. Pour tracer différents graphiques, nous devons charger le module Matplotlib, qui inclut toutes les techniques de tracé de graphiques.

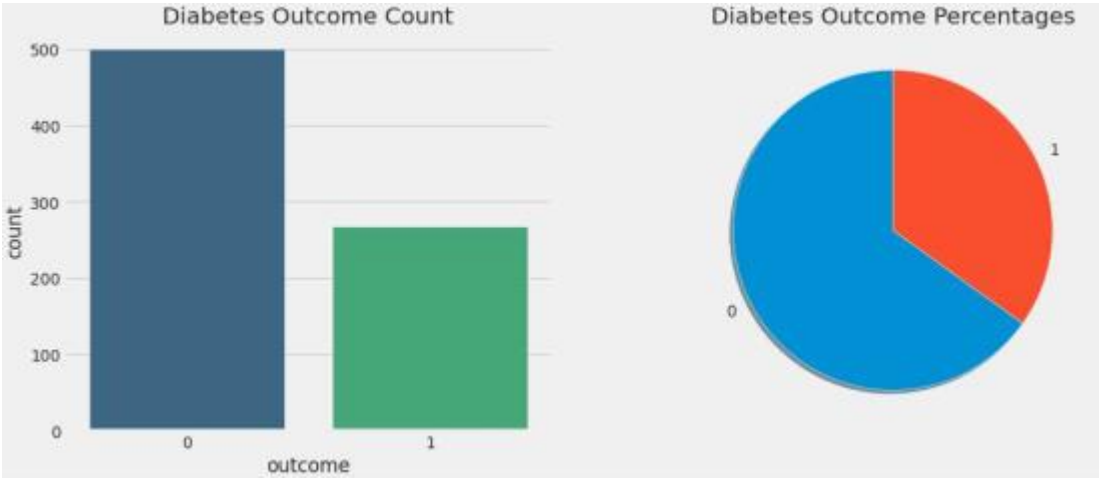

FIGURE 4.2 : NOMBRE ET POURCENTAGE DE L'ATTRIBUT « OUTCOME » DANS PIMA [161].

La (Figure 4.2) montre la répartition de la variable de résultat : il y a 500 cas de non-diabétiques et 286 cas de diabétiques. Au total, il y a 786 cas. Les non-diabétiques sont presque deux fois plus nombreux que les patients diabétiques.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

FIGURE 4.3 : DESCRIPTION DE PIMA [161].

Sur la (Figure 4.3), tous les paramètres sont calculés à l'aide de la méthode "df.describe()" afin d'obtenir la tendance centrale des différents champs du jeu de données, notamment la moyenne, la médiane et le mode.

La fonction describe( ) liste : le nombre total d'observations, la moyenne, l'écart-type, la valeur minimale, le premier quartile (Q1), la médiane (Q2), le troisième quartile (Q3) et la valeur maximale.

Le décompte (count) nous indique combien de lignes non vides il y a dans une caractéristique, la valeur de l'écart-type (std) indique l'écart type de la caractéristique. Et grâce aux percentiles/quartiles pour chaque caractéristique, nous pouvons identifier les valeurs aberrantes. Il est normal de trouver 0 grossesse et nous remarquons qu'il n'y a pas de valeurs nulles pour l'âge et la DiabetesPedigreeFunction.

Comme nous avons un nombre très restreint de variables, nous pouvons avoir une compréhension globale de nos variables et de leur relation. Nous pouvons voir la distribution des données sous forme d'histogrammes et même tracer une courbe de distribution de probabilité [161].

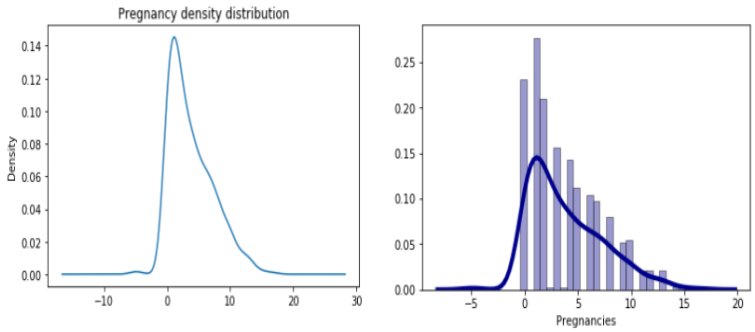

FIGURE 4.4 : DISTRIBUTION DE LA DENSITÉ ET HISTOGRAMME DE «PREGNANCIES» [161].

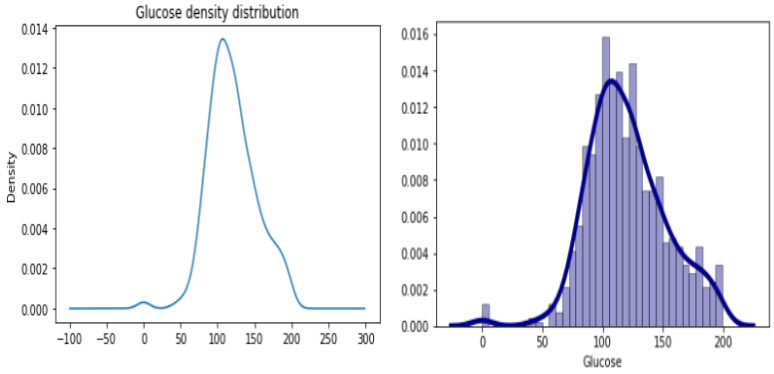

FIGURE 4.5 : DISTRIBUTION DE LA DENSITÉ ET HISTOGRAMME DE «GLUCOSE» [161].

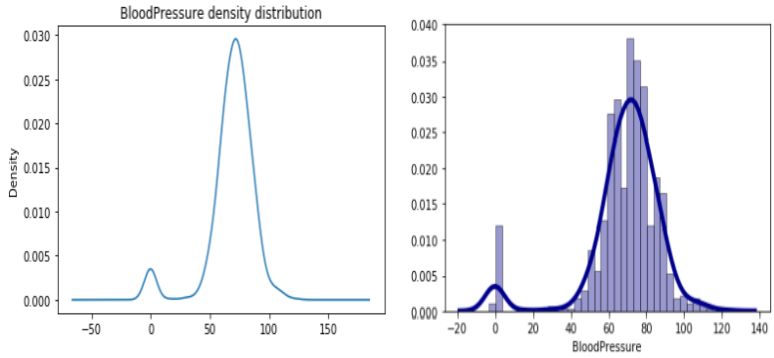

FIGURE 4.6 : DISTRIBUTION DE LA DENSITÉ ET HISTOGRAMME DE «BLOODPRESSURE» [161].

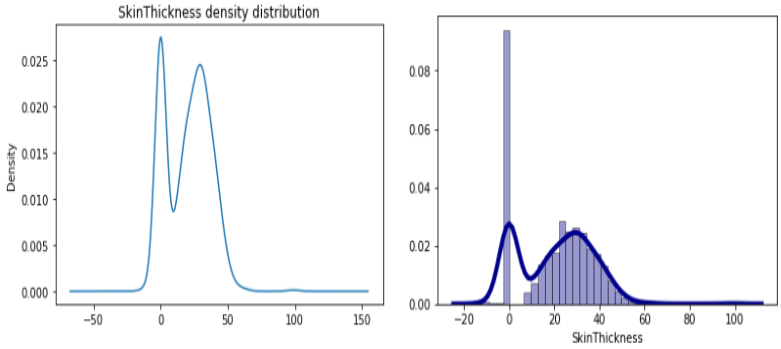

FIGURE 4.7 : DISTRIBUTION DE LA DENSITÉ ET HISTOGRAMME DE «SKINTHICKNESS» [161].

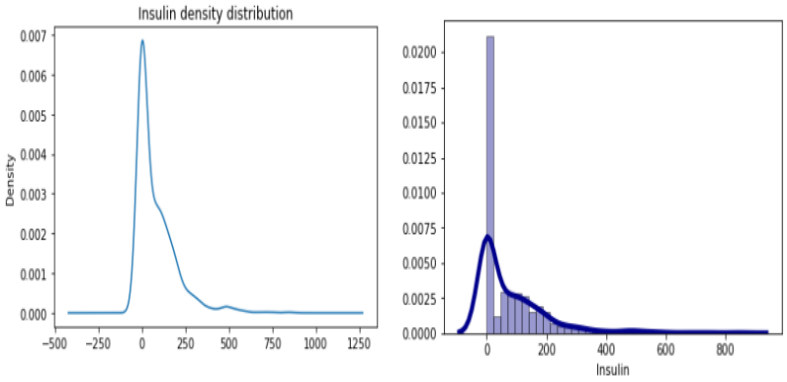

FIGURE 4.8 : DISTRIBUTION DE LA DENSITÉ ET HISTOGRAMME DE «INSULIN» [161].

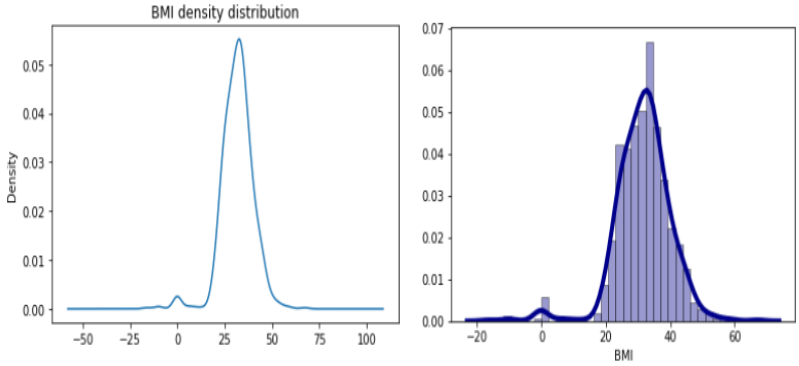

FIGURE 4.9 : DISTRIBUTION DE LA DENSITÉ ET HISTOGRAMME DE «BMI» [161].

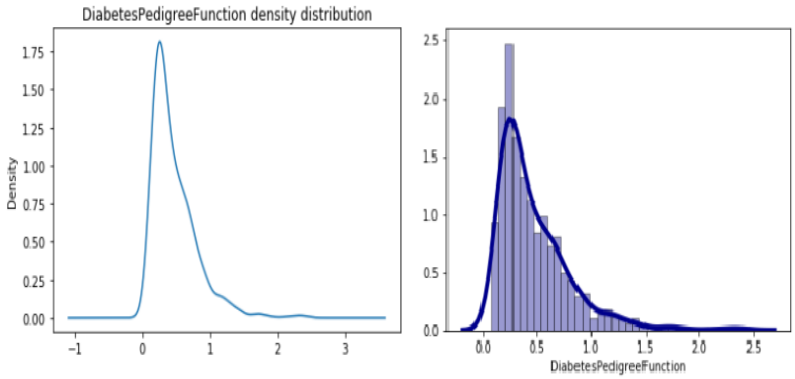

FIGURE 4.10 : DISTRIBUTION DE LA DENSITÉ ET HISTOGRAMME DE «DIABETESPEDIGREEFUNCTION» [161].

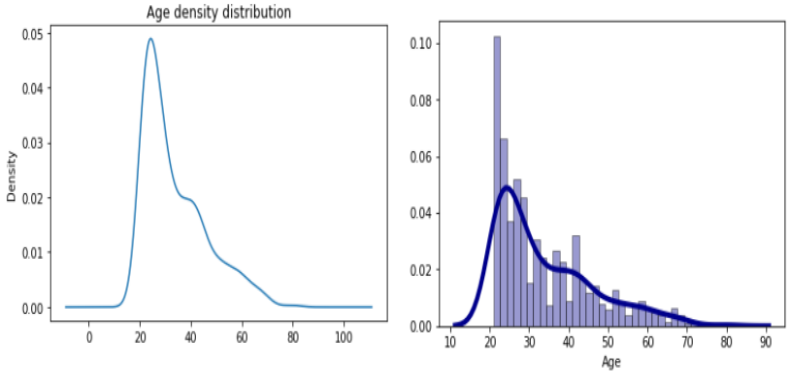

FIGURE 4.11 : DISTRIBUTION DE LA DENSITÉ ET HISTOGRAMME DE «AGE» [161].

Pour trouver la corrélation entre différents champs, nous utilisons la fonction corr() et la représentons à l'aide de la fonction heatmap() de seaborn. Nous pouvons observer l'association entre les champs dans la carte thermique ci-dessous.

Les corrélations sont plus fortes dans les zones plus claires et vice versa dans les endroits plus sombres ou ayant une connexion minimale.

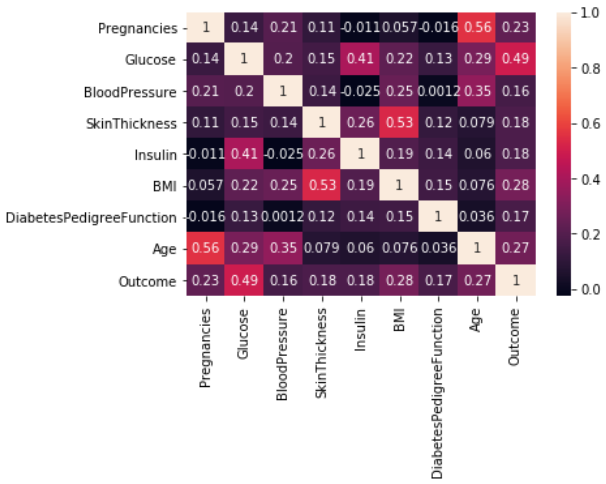

FIGURE 4.12 : CORRÉLATION PIMA[161].

À partir de la carte thermique « heatmap» ci-dessus, nous pouvons déduire que "Glucose" et "Outcome" ont un coefficient de corrélation de 0,49. Nous observons également une corrélation marquée entre "Âge" et "Grossesses", c'est-à-dire 0,56, ce qui s'explique par le fait que l'âge d'une femme augmente, le nombre de grossesses qu'elle a tend à augmenter. Nous pouvons également constater que "IMC" et "épaisseur de la peau" présentent une forte corrélation entre eux. Il existe également une corrélation entre "insuline" et "glucose" [161].

3.3. Prétraitement des données

L'insuline compte le plus grand nombre de zéros (366 enregistrements), suivi de l'épaisseur de la peau (220 enregistrements). La méthode la plus simple pour traiter ces entrées de valeur zéro est de les supprimer, cependant cela supprime un pourcentage important des enregistrements du jeu de données et rend impossible l'obtention d'informations utiles à partir des données restantes (Figure 4.13).

```
Glucose                     5
BloodPressure              34
SkinThickness             220
Insulin                   366
BMI                        10
DiabetesPedigreeFunction    0
Age                         0
```

FIGURE 4.13 : NOMBRE D'ENREGISTREMENTS AVEC DES VALEURS ZÉRO DANS CHAQUE COLONNE [161].

La science de la fouille de données rencontre un problème avec les données manquantes. Cela se produit lorsqu'il n'y a pas de valeurs correspondantes pour une instance donnée, ou lorsque les valeurs sont non pertinentes ou collectées de manière incorrecte lorsque les données sont fournies. L'exactitude et les performances d'une base de données

peuvent être affectées par des valeurs manquantes. Dans le cas du diabète des Indiens Pima, différentes caractéristiques ont des valeurs nulles dans les données, il n'est donc pas possible qu'un patient ait une pression artérielle de 0 ou une concentration de glucose plasmatique de 0 dans son corps. Pour obtenir de bons résultats de classification, il est nécessaire de remplir les valeurs manquantes, sinon des résultats de classification inexacts en découleront.

Ainsi, les données manquantes sont un problème qui doit être traité avec soin. Il existe différentes stratégies pour gérer les données manquantes, telles que : supprimer les observations à données manquantes si nous avons assez de données, ou utiliser la moyenne/la médiane de la distribution du même paramètre pour compléter les données manquantes d'un paramètre.

La technique choisie pour gérer les valeurs nulles est de les substituer par les valeurs moyennes ou médianes du même champ [161].

```
# Replacing the values having 0 with Mean/Median based on distribution
diabetes.Glucose = diabetes.Glucose.replace(0,diabetes.Glucose.mean())
diabetes.BloodPressure = diabetes.BloodPressure.replace(0,diabetes.BloodPressure.median())
diabetes.SkinThickness = diabetes.SkinThickness.replace(0,diabetes.SkinThickness.mean())
diabetes.Insulin = diabetes.Insulin.replace(0,diabetes.Insulin.median())
diabetes.BMI = diabetes.BMI.replace(0,diabetes.BMI.mean())
```

FIGURE 4.14 : REMPLACEMENT DES VALEURS MANQUANTES [161].

3.4. Approches d'apprentissage automatique utilisées

Les méthodologies les plus couramment utilisées dans l'apprentissage automatique sont l'apprentissage supervisé, qui entraîne les algorithmes avec des données d'entrée et de sortie d'instances étiquetées par l'homme, et l'apprentissage non supervisé, qui ne donne pas de données marquées à un algorithme et lui demande de trouver un sens à ses propres données.

Il peut être difficile de choisir l'approche d'apprentissage optimale pour la prédiction des maladies, car elle dépend de la taille de l'ensemble de données et de l'accès de l'utilisateur. Dans la majorité des études, des méthodologies d'apprentissage automatique supervisé (SML) sont utilisées, ainsi qu'une modélisation prédictive simple et directe. La mise en œuvre de ces modèles dans la pratique clinique peut sans aucun doute contribuer à la fourniture de meilleurs services de santé et améliorer la prise de décision des spécialistes. Nous avons choisi la méthode supervisée car nous connaissons déjà les résultats de l'ensembles de données dont nous disposons.

Le théorème "No Free Lunch" est un principe de l'apprentissage automatique. En un mot, aucun algorithme d'apprentissage automatique ne fonctionne de manière optimale pour chaque problème et il est d'une importance cruciale pour l'apprentissage supervisé (c'est-à-dire la modélisation prédictive). Par exemple, on ne peut pas affirmer que les réseaux neuronaux sont toujours meilleurs que les arbres de décision ou que les arbres de décision sont toujours meilleurs que les réseaux neuronaux. De nombreux facteurs entrent en jeu,

notamment la taille et la structure de votre ensemble de données. Par conséquent, vous devez essayer plusieurs méthodes pour votre problème tout en évaluant les performances et en sélectionnant le vainqueur à l'aide d'un "ensemble de test" de données.

Nous utiliserons plusieurs techniques d'apprentissage automatique supervisé pour la prédiction (les algorithmes : NB, KNN(n=5), SVM, DT(random_state=10, max_depth=12), RF(n_estimators = 11, criterion = 'entropy', random_state = 42) et LR (voir section 5.3 pour les définitions)). Dans la méthode proposée, l'ensemble de données a été divisé en deux groupes.

Nous allons nous efforcer de rendre ces modèles plus précis. Afin d'améliorer encore la précision, un modèle d'ensemble hybride est développé, dans lequel trois algorithmes sont combinés avec l'exactitude (accuracy) maximale et alimentent un modèle différent [161].

3.4.1. Le modèle hybride

Les modèles hybrides sont une collection de divers algorithmes de ML. Ils peuvent être utilisés pour tirer parti de l'efficacité de plusieurs modèles sur certains ensembles de données, ce qui augmentera la précision globale du modèle de prédiction.

L'empilement a été utilisé pour créer notre modèle hybride. Le modèle comporte deux couches : la couche de base (niveau 0) et la couche méta (niveau 1). Dans la couche de base, nous utiliserons des algorithmes qui se sont avérés précis lors des tests.

3.5. Résultats expérimentaux de la contribution -1-

Nous utilisons Jupyter Notebook, un logiciel libre pour l'apprentissage automatique. Nous devons importer le module Sklearn, qui comprend les algorithmes et les fonctions nécessaires à l'apprentissage automatique.

Après avoir appliqué tous les algorithmes et leur fait une validation croisée de 10-folds, nous avons évalué notre modèle pour en vérifier son exactitude, comme le montre le (Tableau 4.3). Il a été observé que l'exactitude du modèle est de 77,27 % après la mise en œuvre de la méthode Naïve Bayes, et après la mise en œuvre de l'algorithme Random Forest, il a été déterminé que l'exactitude du modèle est de 83,76 %. En outre, il a été découvert que l'exactitude du modèle était de 78,57 % après l'utilisation de la technique de régression logistique. Après avoir utilisé la méthode KNN, il a été déterminé que le meilleur « accuracy » du modèle est à n = 5, soit 88,31%. En ce qui concerne l'algorithme Support Vector Machine, il a été constaté que l' exactitude du modèle est de 87,01 % et après l'application de l'algorithme pour l'arbre de décision de 85,71 % [161].

Une fois toutes les techniques testées, nous avons découvert que KNN, SVM et l'arbre de décision fournissaient la meilleure exactitude (Figure 4.15). Notre problème étant basé sur la classification, nous avons utilisé la régression logistique dans la métacouche, comme indiqué ci-dessous :

```
from sklearn.ensemble import StackingClassifier # Importing library
def get_stacking(): # Function to make the base layer(level 0) for
                                    the model
    level0 = list()
    level0.append(('knn', KNeighborsClassifier()))
    level0.append(('svm', svm.SVC()))
    level0.append(('dt', DecisionTreeClassifier()))
    level1 = LogisticRegression()
    mod = StackingClassifier(estimators=level0, final_estimator=
                                        level1, cv=10)
    return mod
HybridModel = get_stacking() # Making the Hybrid Model
HybridModel.fit(X_train,y_train) # Fitting training data to the
                                    Hybrid Model
```

Nous avons ensuite obtenu un accuracy de 90,62% pour le modèle hybride.

TABLEAU 4.3 : LES MESURES DES MODÈLES [161].

Modèle	Precision	Recall	F1-score	Accuracy
Modèle Hybride	**0.91**	**0.91**	**0.90**	**90.62%**
NB	0.77	0.77	0.77	77.27 %
LR	0.78	0.79	0.78	78.57%
DT	0.86	0.86	0.86	85.71%
KNN (n=5)	0.88	0.88	0.88	88.31%
SVM	0.88	0.87	0.87	87.01%
RF	0.84	0.84	0.84	83.76%

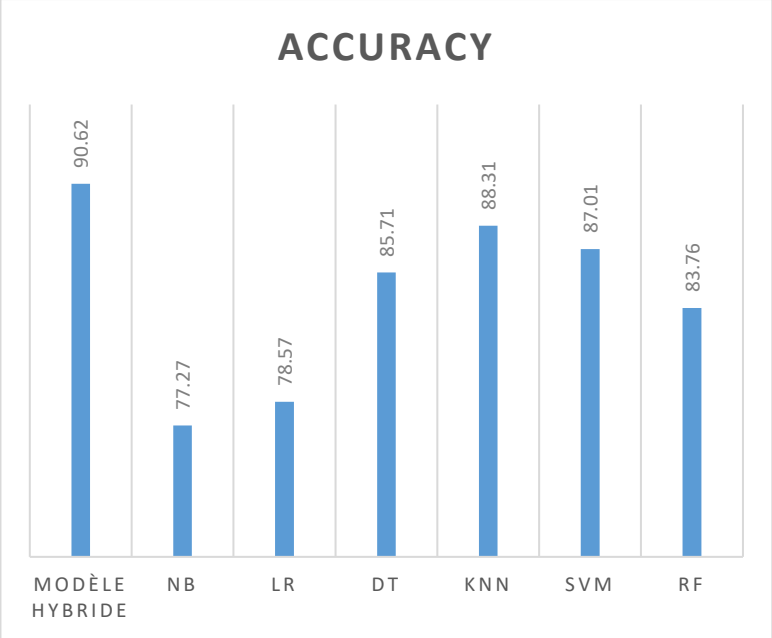

FIGURE 4.15 : L'EXACTITUDE DE CHAQUE CLASSIFIEUR.

3.6. Discussion de la contribution -1-

Les résultats ont été comparés aux travaux antérieurs et sont présentés dans le (Tableau 4.4). L'ensemble de données sur le diabète des Indiens Pima a été utilisé pour toutes les évaluations.

TABLEAU 4.4 : LA COMPARAISON DES PERFORMANCES.

Approche	Modèle	Accuracy
Notre travail [161]	**Hybrid classifier**	**90.62%**
Ramezani et al. 2018 [162]	Hybrid classifier	88.05%
Harleen Kaur et al. 2019 [163]	KNN	88.00%
Nonso Nnamoko et al. 2020 [164]	C4.5	89.50%
Shekharesh Barik et al. 2021 [165]	XGBoost method	74.10%

4. Contribution -2-

Le développement de nos modèles prédictifs comprend les phases montrées dans la (Figure 4.16) ci-dessous. Le jeu de donnée utilisé est PIMA (section 4.1.).

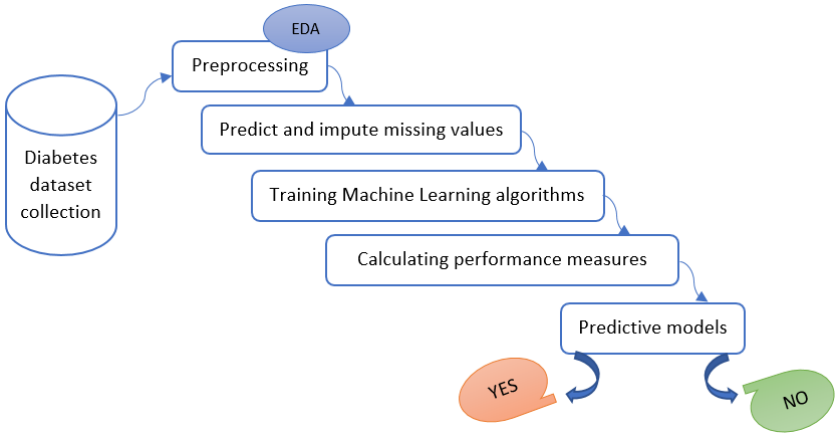

FIGURE 4.16 : PHASES DE DÉVELOPPEMENT DES MODÈLES PRÉDICTIFS [166].

4.1. Travaux connexes des techniques d'imputation

Dans notre recherche, nous avons proposé une nouvelle méthode pour traiter les valeurs manquantes (le mélange de techniques d'imputation des valeurs manquantes). Grâce à cette approche, notre modèle a obtenu de meilleures performances que les travaux antérieurs réalisés avec la PIDD. Nous avons donc passé en revue les techniques d'imputation utilisées dans les études antérieures (Tableau 4.5).

Dans les travaux d'Islam & Jahan (2017) [167], seules les observations présentant des valeurs impossibles pour plusieurs attributs ont été supprimées pour que les ensembles de données restent gérables. Il restait 755 observations. Cependant, dans les études de Zou et al. (2018) [168] et Lai et al. (2019) [169], les 786 enregistrements originaux de diabétiques ont été réduits à 392 après que les enregistrements manquants ont été supprimés.

Ramezani et al. (2018) [162] ont mis au point le classifieur hybride "Logistic Adaptive Network Based Fuzzy Inference System" (LANFIS). Il combine la régression logistique avec un réseau adaptatif basé sur un cadre d'induction floue. Grâce à cela, les enregistrements contenant des valeurs manquantes dans l'ensemble de données sur le diabète Pima ont pu être traités, sans s'appuyer sur des attributs triviaux.

Dans l'étude de Syed & Khan (2020) [170], les colonnes d'attributs de l'ensemble de données PIMA qui avaient des valeurs manquantes ont été complétées par la valeur moyenne ou médiane de la colonne correspondante. Toutefois, dans les recherches de Alam et al. (2019) [171] et Arora et al. (2021) [172], la valeur médiane de l'attribut a été utilisée pour remplacer tous les zéros. En revanche, Khanam & Foo (2021) [173]; Taz et al. (2021) [174] et Barik et al. (2021) [165] ont examiné les valeurs d'attribut manquantes. Les valeurs manquantes ont ensuite été imputées à la valeur moyenne de chaque attribut.

TABLEAU 4.5 : ÉTUDES ANTÉRIEURES ET LEURS TECHNIQUES D'IMPUTATION.

Année	Études antérieures	Techniques d'imputation
2018	Zou et al. [168]	Les lignes avec des valeurs manquantes sont éliminées
2018	Ramezani et al. [162]	L'utilisation d'algorithmes prenant en charge les valeurs manquantes (LANFIS)
2019	Alam et al. [171]	Imputer les valeurs manquantes pour variables continues (avec médiane)
2019	Lai et al. [169]	Les lignes avec des valeurs manquantes sont éliminées
2020	Syed & Khan [170]	Imputer les valeurs manquantes pour variables continues (avec moyenne ou médiane)
2021	Taz et al. [174]	Imputer les valeurs manquantes pour variables continues (avec moyenne)
2021	Khanam & Foo [173]	Imputer les valeurs manquantes pour variables continues (avec moyenne)
2021	Arora et al. [172]	Imputer les valeurs manquantes pour variables continues (avec médiane)
2021	Barik et al. [165]	Imputer les valeurs manquantes pour variables continues (avec moyenne)

4.2. Analyse des données exploratoires et prétraitement des données (EDA)

Il est possible que les données du monde réel contiennent des informations incohérentes ou des valeurs manquantes. Lorsque la qualité des données est médiocre, il est difficile de tirer des conclusions sur la qualité. Avant d'obtenir des résultats de qualité, les données doivent être prétraitées. Lors du prétraitement, les données sont nettoyées, intégrées, converties, réduites et discrétisées. Il s'agit du processus de conversion des données brutes en un format compréhensible. L'efficacité en termes de temps, de coût et de qualité doit être améliorée dans les processus d'exploration et d'analyse des données.

Avant d'appliquer les calculs de ML à l'assortiment d'informations, les auteurs examinent d'abord les données exploratoires. Les données incohérentes sont traitées au cours de cette étape afin de produire des résultats plus fiables et plus précis. Dans cet ensemble de données, certaines valeurs sont manquantes. Quelques variables, telles que le taux de glucose, l'épaisseur de la peau, la tension artérielle et l'âge, ont donc été imputées comme données manquantes car leurs valeurs ne peuvent pas être nulles. En utilisant l'analyse exploratoire des données (Exploratory Data Analysis « EDA »), une manière de disséquer les ensembles de données pour résumer leurs propriétés fondamentales, en

utilisant régulièrement des méthodes visuelles, nous avons effectué les tâches suivantes [166]:

- Recherche des types de données (Tableau 4.2).
- Etudier la variable cible.
- Analyser la distribution des données pour avoir une idée plus précise de l'appropriation des avantages de chaque variable (Figure 4.17)
- Identifier les données manquantes (Figure 4.18 et Figure 4.19). Caractéristiques : Glucose, pression sanguine, épaisseur de la peau, insuline, IMC).
- Identification des valeurs aberrantes “outliers”.
- Détermination de la corrélation entre les caractéristiques (Figure 4.20).
- Détermination de l'importance des caractéristiques.

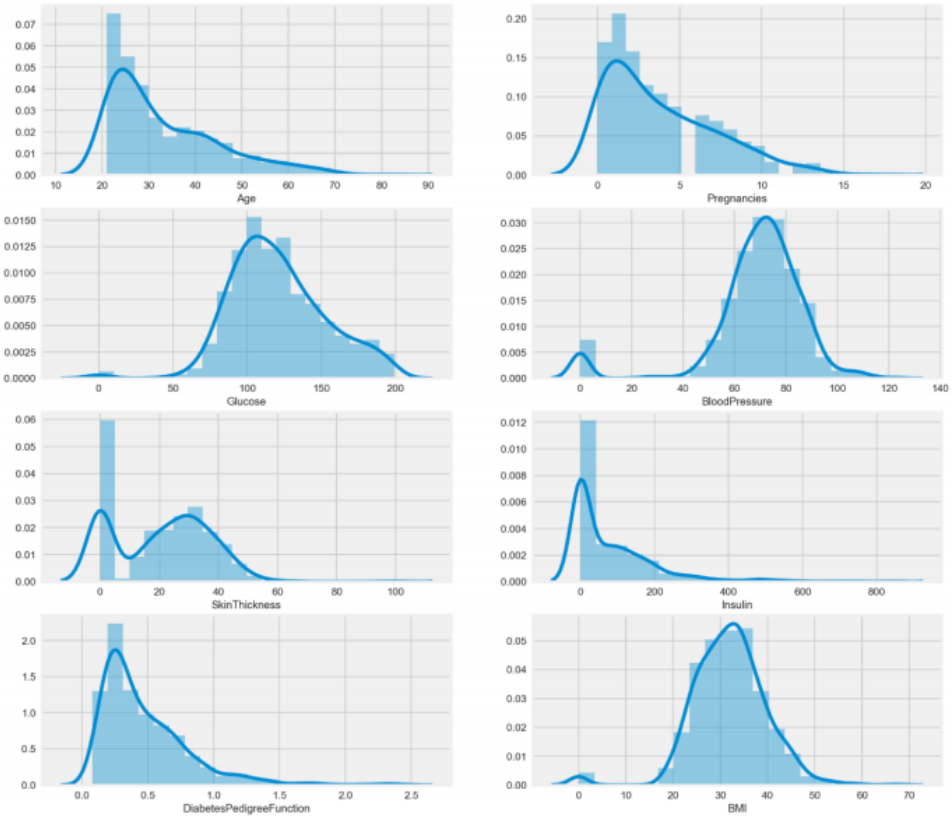

FIGURE 4.17 : DISTRIBUTION DES DONNÉES [166].

```
1 diabetes.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
has_diabetes	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

FIGURE 4.18 : DONNÉES MANQUANTES [166].

```
Total number of 0 values in Pregnancies = 111
Total number of 0 values in Glucose = 5
Total number of 0 values in BloodPressure = 35
Total number of 0 values in SkinThickness = 227
Total number of 0 values in Insulin = 374
Total number of 0 values in BMI = 11
Total number of 0 values in DiabetesPedigreeFunction = 0
Total number of 0 values in Age = 0
Total number of 0 values in Outcome = 500
```

FIGURE 4.19 : NOMBRE TOTAL DE VALEURS NULLES [166].

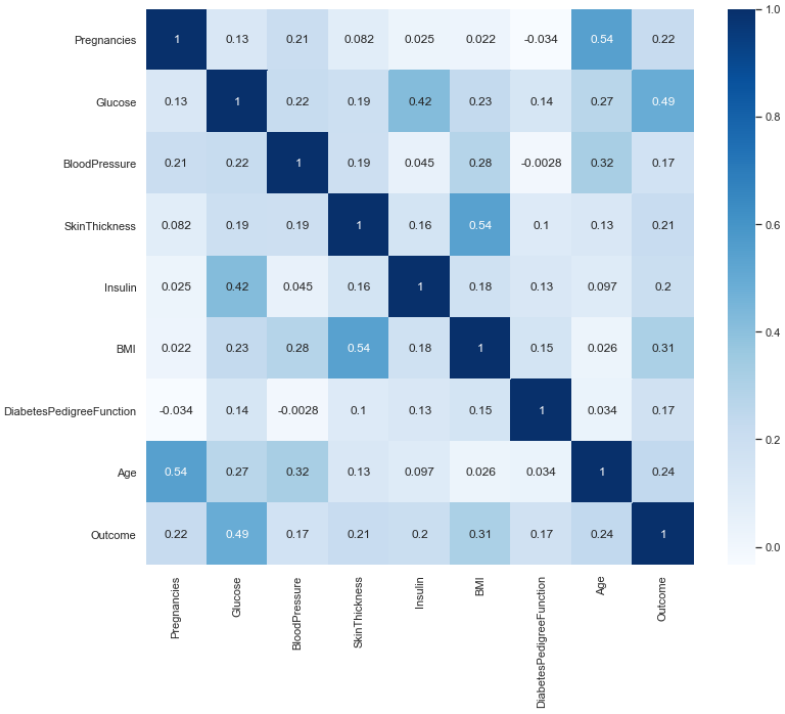

FIGURE 4.20 : CORRÉLATION ENTRE LES ATTRIBUTS [166].

4.2.1. Prédire et imputer les valeurs manquantes

Différents attributs de l'ensemble de données sur le diabète des Indiens Pima contiennent des valeurs nulles, et il n'y a donc pas de patients dont la tension artérielle

ou le taux de glucose plasmatique sont nuls. Des résultats de classification inexacts apparaîtront si les valeurs manquantes ne sont pas correctement imputées.

Il existe plusieurs techniques pour traiter les valeurs manquantes dans un ensemble de données :

- Les lignes comportant des valeurs manquantes sont éliminées (cette stratégie n'est recommandée que lorsque l'ensemble de données comporte un nombre suffisant d'échantillons).
- Imputer les valeurs manquantes pour les variables continues.
- Imputation des valeurs manquantes pour les variables catégorielles.
- Autres méthodes d'imputation.
- Utilisation d'algorithmes prenant en charge les valeurs manquantes.
- Prédiction des valeurs manquantes.

Dans ce travail, il y a 5 caractéristiques avec des données manquantes qui doivent être traitées avec soin. Dans la (Figure 4.18), les auteurs remarquent que l'insuline contient le plus grand nombre d'enregistrements (374), suivie par l'épaisseur de la peau (227), tandis que le glucose, la tension artérielle et l'IMC en contiennent le moins. On propose donc une "combinaison d'imputation de données" représentée par le mélange de deux techniques (imputation des valeurs manquantes pour les variables continues + prédiction des valeurs manquantes). La technique choisie pour gérer les quelques valeurs nulles est l'imputation des valeurs manquantes pour les variables continues". Cette approche les remplace par les valeurs moyennes/médianes de la même colonne (champ) pour le glucose, la pression artérielle et l'IMC.

Pour les attributs Insuline et Epaisseur de la peau qui ont un nombre élevé de zéros, un algorithme de régression ML, comme SVM, Régression linéaire,... est nécessaire pour prédire une variable continue. D'autres caractéristiques, celles qui n'incluent pas les zéros, peuvent être utilisées pour prédire les données manquantes.

Les valeurs de chaque colonne sont choisies au hasard parmi les données non manquantes. L'étape suivante consiste à essayer d'anticiper ce que les valeurs auraient dû être si elles avaient été mesurées avec précision. Pour ce faire, on peut utiliser un modèle de régression linéaire. Ainsi, en profitant des algorithmes de régression linéaire, il serait facile d'anticiper les valeurs nulles en utilisant des caractéristiques qui n'ont pas de valeurs manquantes. En utilisant d'autres caractéristiques accessibles, les auteurs utiliseront la régression linéaire pour remplacer les valeurs nulles dans les caractéristiques "Insulin" et " SkinThickness".

Lorsqu'un modèle est formé avec toutes les colonnes restantes, il peut parfois prédire les valeurs manquantes mieux qu'un modèle construit avec seulement certaines des colonnes, mais lorsque certaines des colonnes sont fortement corrélées, un modèle formé avec seulement quelques colonnes fortement corrélées peut offrir un meilleur

résultat. Comme le montre la (Figure 4.20), aucune corrélation forte ne peut être utilisée [166].

4.3. Modélisation

Cette section passe en revue les différentes méthodes de classification d'apprentissage supervisé qui ont été utilisées dans la construction de modèles prédictifs de diabète. En apprentissage automatique, la division en ensembles d'entraînement et de test est utilisée pour mesurer les performances d'un algorithme. Cela peut être utilisé pour résoudre des problèmes de régression/classification, ainsi que pour les algorithmes d'apprentissage supervisé mis en œuvre. Le jeu de données est divisé en deux sous-groupes à l'aide de cette méthode, puis réadopté. Le jeu de données d'entraînement est le composant initial utilisé pour ajuster le modèle. L'entraînement du modèle sur le deuxième sous-ensemble est inutile. À la place, il est utilisé comme valeurs d'entrée pour les modèles, et des prévisions sont créées et comparées avec les valeurs anticipées. Cela s'appelle : jeu de données de test.

Suite à la préparation des données, on peut dire que les auteurs sont confrontés à des données déséquilibrées. Cela comprend (65,1 %) de non-diabétiques et (34,9 %) de diabétiques. Lors du développement de modèles d'apprentissage automatique, il est nécessaire de prendre en compte les données déséquilibrées. Un déséquilibre de classe provoque un biais qui pousse le classifieur d'apprentissage automatique à prédire la classe majoritaire. Personne ne souhaite que le modèle de prédiction néglige la population minoritaire. La suréchantillonnage des instances de la classe minoritaire est l'une des solutions. Avant d'ajuster un modèle, il suffit de dupliquer les enregistrements de la classe minoritaire dans le jeu de données d'entraînement. Cela peut équilibrer la distribution des classes mais n'ajoute aucune nouvelle information au modèle. C'est pourquoi les auteurs ont importé et utilisé SMOTE (Synthetic Minority Oversampling Technique) de imblearn.over_sampling qui est une forme d'augmentation de données pour la population minoritaire. Les modèles de prédiction qui ont été construits comprenaient l'entraînement et le test pour chacun des six algorithmes. Les données ont été utilisées pour entraîner les modèles, qui ont ensuite été évalués pour déterminer à quel point ils prédisaient le résultat. Chaque modèle a été formé et testé en divisant le jeu de données en données d'entraînement et données de test. Quatre-vingt-cinq pour cent des données ont été utilisés pour l'entraînement des modèles, avec 15 % conservés pour tester les modèles. Dans cette étude, les six algorithmes suivants ont été utilisés avec les paramètres des classifieurs choisis s'ils existent [166].

4.3.1. Random Forest

En tant que mélange de prédicteurs arborescents, la technique de la forêt aléatoire est polyvalente, rapide et simple. La plupart du temps, la forêt aléatoire donne de bons résultats. Ses performances sont difficiles à améliorer, de plus elle peut contenir une variété de types de données (données nominales, numériques et binaires). En utilisant la technique RF, plusieurs arbres de décision sont construits et agrégés pour obtenir des résultats précis. Outre la classification, elle a également été utilisée pour la régression. En apprentissage automatique, la classification est une tâche clé. Ces

hyperparamètres sont identiques à ceux d'un arbre de décision / d'un classifieur de mise en sac. Le principe sous-jacent de la forêt aléatoire est la superposition d'arbres aléatoires, il est également facilement compréhensible. Supposons que sept arbres aléatoires aient fourni des informations sur une variable, et que quatre d'entre eux soient d'accord tandis que les trois autres sont en désaccord. En s'appuyant sur les probabilités, des modèles d'apprentissage automatique sont construits, ainsi que sur un vote majoritaire.

Dans les collections de données volumineuses en RF, un sous-ensemble aléatoire de caractéristiques produit des résultats précis, et plus d'arbres aléatoires peuvent être créés en établissant un seuil aléatoire pour toutes les caractéristiques plutôt qu'en sélectionnant le meilleur seuil. La méthode aborde également le problème de surajustement.

```
rfc = RandomForestClassifier (class_weight='balanced', random_state=42,
n_estimators=450, min_samples_leaf=2)
```

4.3.2. Machine à Vecteurs de Support

Élégante, puissante et largement utilisée, la machine à vecteurs de support (SVM) est l'un des algorithmes SML couramment utilisés. Elle est utilisée en classification ML ainsi qu'en régression (dans un espace de dimension supérieure). Vapnick en 1995 et 1998 a posé les bases des SVM, qui est une méthode générique basée sur les limites de risque garanties de la théorie de l'apprentissage statistique, souvent appelée principe de minimisation du risque structurel. Selon plusieurs études récentes, les SVM sont capables de donner une meilleure exactitude de classification que d'autres algorithmes. Pour prédire le diabète, les auteurs ont utilisé SVM, qui convient bien aux problèmes de classification binaire.

```
svc = SVC(class_weight='balanced', random_state=42)
```

4.3.3. K Plus Proches Voisins

Le classifieur K Plus Proches Voisins (K-Nearest Neighbors « KNN ») se caractérise par sa méthode non paramétrique. Lors de l'attribution des étiquettes de classe des exemples de test, il prend en compte la fonction de distance, les exemples d'entraînement ainsi que le nombre de voisins les plus proches. La distance euclidienne est une solution générique pour le calcul des distances. Le vote majoritaire des étiquettes d'un nombre prédéfini de voisins détermine les étiquettes de classe des exemples de test. L'approche du voisin le plus proche peut être utilisée pour trouver des points de données anonymes en utilisant des points de données familiers. Conceptuellement simple, on peut dire que KNN est un apprentissage paresseux, où le voisin le plus proche est désigné par "K". Cette méthode est conçue pour identifier K exemples dans les données d'entraînement qui ressemblent à un autre exemple nouveau avec une grande précision.

```
knn = KNeighborsClassifier(n_neighbors=4, weights='distance',
metric='euclidean', n_jobs=5)
```

4.3.4. Naïve Bayes

Le Naïve Bayes « NB » repose sur la probabilité d'occurrence d'événements en fonction de la connaissance de facteurs qui peuvent y être liés. Pour les grands ensembles de données, le Naïve Bayes est l'algorithme de classification le plus rapide et le plus simple. On peut trouver l'utilisation du classifieur NB dans de nombreuses applications, dont l'analyse de sentiments et la classification de texte. Il est possible de prévoir des classes inconnues à l'aide du théorème de Bayes sur la probabilité. Le Naïve Bayes est une méthode simple et basique, facile à mettre en œuvre. Par conséquent, lorsque les données sont peu nombreuses, elle peut surpasser des modèles plus compliqués. Le modèle de classification attribue des étiquettes de classe aux cas en fonction des valeurs des caractéristiques, et les étiquettes de classe sont sélectionnées dans une plage restreinte d'étiquettes disponibles. En supposant que l'élément à catégoriser existe, on détermine que la catégorie à classer a la plus forte probabilité de se produire. La prédiction diabétique peut bénéficier de ce type de prédiction probabiliste de la classe la plus probable.

4.3.5. Arbre de Décision

Les approches d'apprentissage supervisé comprennent l'algorithme de l'arbre de décision (Decision Tree « DT »). Avec cette approche, les utilisateurs peuvent traiter des informations à la fois continues et catégorielles. Les arbres de décision divisent les populations, et le diviseur détermine comment la population est divisée. Le diviseur peut séparer deux ou plusieurs sous-groupes du même type. Après avoir terminé un travail, il crée un arbre, et sa justesse dépend fortement des jugements. En ce qui concerne la classification et l'application de la régression à un arbre, les méthodes sont quelque peu différentes. Pour éviter le surajustement, des règles ou des contraintes sur la construction et la taille des arbres peuvent être décrites.

Comparés à certaines autres méthodes de classification supervisées, comme la classification de maximum de vraisemblance, les arbres de décision présentent plusieurs avantages. Ils traitent également des relations non linéaires entre les caractéristiques et les classes, et ils permettent les valeurs manquantes. Ils peuvent également traiter des données d'entrée catégorielles et numériques.

```
dt = DecisionTreeClassifier(criterion = 'entropy', max_depth = 28,
min_samples_leaf = 1, random_state = 42)
```

4.3.6. Régression Logistique

Pour classer des données sous supervision, la régression logistique (Logistic Regression « LR ») est une technique efficace. Seule une variable dichotomique peut être modélisée, ce qui reflète généralement la non-occurrence ou l'occurrence d'un événement dans ce type de régression. La LR aide à déterminer si une nouvelle instance est susceptible d'appartenir à une classe particulière. Les probabilités vont de zéro à un.

Un seuil doit être appliqué à la LR pour l'utiliser comme classifieur binaire. Par exemple, une instance d'entrée classée comme "classe A" a une probabilité supérieure à 0,50 ; sinon, elle est classée comme "classe B".

```
lr = LogisticRegression(class_weight='balanced', random_state=42)
```

4.4. Calcul des Mesures de Performance

C'est ici que l'analyse des performances se produit. Dans cette étude, on a utilisé à la fois la matrice de confusion (CM) pour évaluer les performances des 6 classifieurs d'apprentissage automatique et certaines mesures d'évaluation (Tableau 4.6 et Tableau 4.7), telles que le coefficient de corrélation de Matthews (Matthews Correlation Coefficient MCC), le score F1 (F1-Score), la spécificité (specificity) et la sensibilité (sensitivity). Une matrice de confusion (Confusion Matrix CM) en apprentissage automatique est utilisée pour évaluer les performances du processus de classification. Il y a deux colonnes dans la CM : l'une indique la classification réelle et l'autre la prédit. L'accuracy de classification est également l'un des mesures d'évaluation des performances. Il mesure l'exactitude des résultats prédits en se basant sur les données d'entraînement [166]. Comme le montrent la (Figure 4.21), la capacité de prévision du classifieur est abordée par la CM.

4.5. Modèles Prédictifs

Après avoir terminé avec les classifieurs et enregistré les modèles obtenus, il pourrait être possible de prédire le risque futur de diabète d'un patient en fournissant ses informations (attributs nécessaires "entrées"). Le modèle prédira si :

- Non : La personne a un faible risque de diabète (indiqué par 0).
- Oui : La personne a un risque de diabète et devrait consulter un médecin dès que possible (indiqué par 1).

4.6. Résultats expérimentaux de la contribution -2-

Dans cette partie, les auteurs présentent les résultats expérimentaux qui ont été obtenus après l'entraînement des six algorithmes. Afin de trouver les meilleurs hyperparamètres pour nos modèles d'apprentissage automatique, nous avons utilisé GridSearchCV 10 folds pour obtenir des performances optimales. GridSearchCV effectue une recherche exhaustive dans une grille d'hyperparamètres spécifiée, évaluant les performances du modèle pour chaque combinaison à l'aide d'une validation croisée. Ces résultats d'expérimentation visent à évaluer les performances des classifieurs.

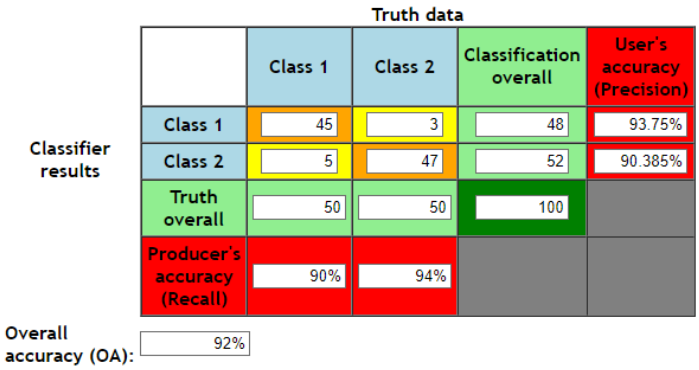

FIGURE 4.21 : MATRICE DE CONFUSION DU RF [166].

Les performances des modèles sont évaluées en termes de MCC, de score F1, d'accuracy, de taux de faux négatifs, de taux de fausse découverte, de taux de faux positifs, de valeur prédictive négative, de précision, de spécificité et de sensibilité. Toutes les formules de calcul sont présentées dans le (Tableau 4.6). Les termes "faux positif" (FP), "vrai positif" (TP), "faux négatif" (FN) et "vrai négatif" (TN) sont utilisés pour calculer ces facteurs de mesure de la classification.

TABLEAU 4.6 : MESURE AVEC SA DÉRIVATION [166].

Mesures	Formules de calcul
Sensitivity	TPR=TP / (TP+FN)
Specificity	SPC=TN / (FP+TN)
Precision	PPV=TP / (TP+FP)
Negative Predictive Value	NPV=TN / (TN+FN)
False Positive Rate	FPR=FP / (FP+TN)
False Discovery Rate	FDR=FP / (FP+TP)
False Negative Rate	FNR=FN / (FN+TP)
Accuracy	ACC=(TP+TN) / (P+N)
F1 Score	F1=2TP / (2TP+FP+FN)
Matthews Correlation Coefficient	TN*TP – FN*FP / sqrt ((FP+TP)*(FN+TP)*(FP+TN)*(FN+TN))

Le (Tableau 4.7) compare toutes les techniques de classification en fonction d'un grand nombre de mesures. Le modèle RF est manifestement le meilleur pour anticiper le risque de diabète.

TABLEAU 4.7 : PERFORMANCES DES MODÈLES [166].

Modèle	TPR	SPC	PPV	NPV	FPR	FDR	FNR	ACC	F1	MCC
LR	88.00	82.00	83.02	87.23	18.00	16.98	12.00	85.00%	85.44	70.13
DT	80.00	82.00	81.63	80.39	18.00	18.37	20.00	81.00%	80.81	62.01
NB	86.00	88.00	87.76	86.27	12.00	12.24	14.00	87.00%	87.00	74.01
KNN(n=4)	82.00	92.00	91.11	83.64	08.00	08.89	18.00	87.00%	86.32	74.37
SVM	84.00	90.00	89.36	84.91	10.00	10.64	16.00	87.00%	86.60	74.13
RF	**90.00**	**94.00**	**93.75**	**90.38**	**06.00**	**06.25**	**10.00**	**92.00%**	**91.84**	**84.07**

La (Figure 4.22) montre un examen de divers classifieurs ML selon leur taux d'exactitude:

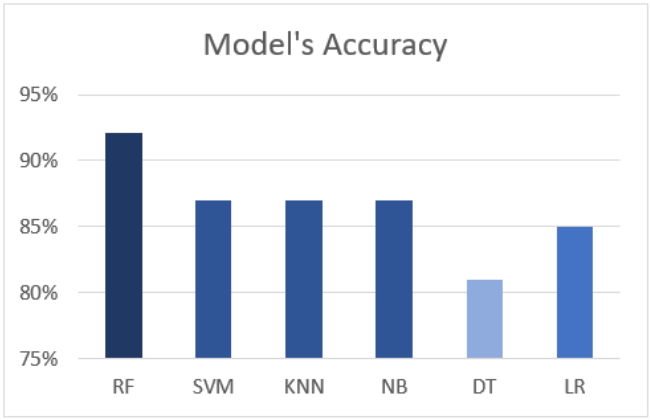

FIGURE 4.22 : EXACTITUDE DE SIX MODÈLES ML SUPERVISÉS [166].

4.7. Discussion de la contribution -2-

Dans cette étude, l'ensemble de données des Indiens PIMA a été exploré. Dans le cadre de cette recherche, le jeu de données contenait huit attributs indépendants et une classe d'attribut dépendant, qui ont été entraînés pour prédire les maladies diabétiques.

Il arrive parfois qu'un certain élément manque aux données du monde réel pour diverses raisons, notamment des données défectueuses, l'incapacité de charger les informations ou une extraction insuffisante. L'une des tâches les plus difficiles pour les analystes est de gérer les valeurs manquantes, car prendre la décision appropriée sur la manière de les gérer conduit à des modèles de données plus robustes. Pour garantir la précision du modèle, il est nécessaire de prendre soin du prétraitement des données. Par conséquent, avant de modéliser la tâche de prédiction, il est conseillé de rechercher et de remplacer les valeurs manquantes pour chaque colonne dans les données d'entrée. Cela s'appelle l'imputation des données manquantes, ou simplement l'imputation. Calculer une valeur statistique pour chaque colonne (comme une moyenne) et remplacer toutes les valeurs manquantes de cette colonne par la statistique est une stratégie courante pour l'imputation des données. C'est une méthode courante car la statistique est simple à calculer

à l'aide du jeu de données d'entraînement, et elle donne souvent de bons résultats. Une technique d'apprentissage automatique peut également être utilisée pour prévoir les valeurs manquantes. Nous avons proposé une nouvelle manière de traiter les valeurs manquantes (mélange de techniques d'imputation des données manquantes). Avec cette approche, notre modèle a obtenu une meilleure précision que les travaux précédents réalisés avec les données sur le diabète de Pima.

Contrairement à la plupart des études dans ce domaine comme Taz et al. (2021) [174], Khanam & Foo (2021) [173], Arora et al. (2021) [172] et Barik et al. (2021) [165], nous n'avons pas utilisé l'approche de remplacement des données manquantes par la moyenne ou la médiane pour toutes les caractéristiques, car cela n'est pas recommandé lorsque le nombre de valeurs manquantes est élevé. Nous n'avons pas non plus supprimé les lignes ayant des valeurs manquantes comme Zou et al. (2018) [168] et Lai et al. (2019) [169], car cela n'est pas recommandé lorsque l'ensemble de données a un nombre insuffisant d'échantillons. Ainsi, selon notre conceptualisation, on peut en déduire que le nombre de valeurs manquantes pour chaque attribut est un facteur important pour déterminer quelle technique utiliser pour résoudre le problème. Et la combinaison de techniques d'imputation permettrait d'obtenir un ensemble de données bien préparé avant d'utiliser des algorithmes d'apprentissage automatique.

Pour améliorer la prédiction du diabète chez les patients, les auteurs ont utilisé six algorithmes d'apprentissage automatique : RF, SVM, KNN, NB, DT et LR. Et ils les ont utilisés après un bref résumé de l'importance du diagnostic précoce des patients diabétiques pour réduire les décès. Lors de l'évaluation des algorithmes, les auteurs ont utilisé différentes mesures.

La (Figure 4.22) représente les performances des six méthodes d'apprentissage automatique supervisé pour la prédiction du diabète. Les scores d'exactitudes moyens pour la classification PIDD se situaient entre 81 et 92 pour cent. Dans ce cas, la forêt aléatoire a surpassé les autres approches de catégorisation, avec une exactitude de 92%. Les autres classifieurs ont également pu atteindre des exactitudes élevées de plus de 80%. En revanche, DT a eu les moins bonnes performances parmi les algorithmes de classification. Le classifieur RF a obtenu les résultats suivants : exactitude de 92 %, précision de 93,75 %, spécificité de 94 %, rappel de 90 %, score F1 de 91,84 % et MCC de 84,07 %. Le classifieur DT a obtenu les moins bonnes performances, avec les résultats suivants : exactitude de 81 %, précision de 81,63 %, spécificité de 82 %, rappel de 80 %, score F1 de 80,81 % et MCC de 62,01 %. SVM, KNN et NB ont obtenu la même précision de 87 %. Le classifieur LR a obtenu les résultats suivants : exactitude de 85 %, précision de 83,02 %, spécificité de 82 %, rappel de 88 %, score F1 de 85,44 % et MCC de 70,13 %. Le (Tableau 4.7) fournit des informations détaillées sur chaque statistique qui a été utilisée dans notre étude, où une comparaison a été réalisée pour déterminer quel algorithme d'apprentissage automatique fait les prédictions les plus précises [166].

Les résultats ont été comparés à d'autres études, comme le montre le (Tableau 4.8) ci-dessous, et il semble que la plupart des « accuracies » obtenues pour nos classifieurs étaient meilleures que les précisions des autres études.

TABLEAU 4.8 : COMPARAISON DES PERFORMANCES DES MÉTHODES IMPLÉMENTÉES [166].

Travail	Modèle	Accuracy
Islam and Jahan (2017)	Random Forest	74.83%
	Naïve Bayes	75.76%
	Logistic Regression	78.01%
	Support Vector Machine	77.08%
Zou et al. (2018)	Random Forest	76.04%
Mujumdar and Vaidehi (2019)	Random Forest	72.00%
	Gaussian Naïve Bayes	67.00%
	Logistic Regression	76.00%
Tigga and Garg (2020)	Random Forest	75.00%
	Naïve Bayes	68.90%
	Logistic Regression	74.40%
	Support Vector Machine	74.40%
	Decision Tree	69.70%
	KNN	70.80%
Nnamoko and Korkontzelos (2020)	Random Forest	75.50%
	Naïve Bayes	77.00%
	Support Vector Machine -RBF	77.70%
Khanam and Foo (2021)	Random Forest	77.34%
	Naïve Bayes	78.28%
	Logistic Regression	78.85%
	Support Vector Machine	77.71%
	Decision Tree	73.14%
	KNN	79.42%
Ramesh et al. (2021)	Naïve Bayes	75.50%
	Logistic Regression	76.40%
	Support Vector Machine	83.20%
	KNN	79.80%
Barik et al. (2021)	Random Forest	71.90%
Arora et al. (2021)	Random Forest	73.85%

5. Conclusion

En conclusion de ce chapitre, nous avons parcouru un voyage captivant à travers les différents aspects de notre recherche axée sur le jeu de données PIMA. Les deux expérimentations que nous avons menées ont représenté des jalons importants de notre recherche. Chacune d'entre elles a apporté sa contribution unique. En exploitant de nouvelles approches, en testant des hypothèses novatrices et en fournissant des résultats significatifs. Ces expériences ont contribué à enrichir la littérature existante et à ouvrir de nouvelles perspectives de recherche dans ce domaine crucial.

Dans la pratique clinique, les modèles prédictifs d'apprentissage automatique peuvent mettre l'accent sur de meilleures lignes directrices pour la prise de décisions concernant le traitement individuel des patients. La détection précoce et les thérapies appropriées sont les seuls moyens de réduire les taux de mortalité causés par les maladies chroniques. En utilisant les modèles de prédiction par apprentissage automatique présentés dans la contribution -1- , nous avons pu détecter une plus grande précision en utilisant un modèle hybride qui a atteint un accuracy de 90,62 %.

De plus dans la contribution -2-, les résultats ont démontré l'efficacité du système, avec une exactitude de 92% en utilisant le classifieur Random Forest. L'amélioration de la classification a été obtenue comme un effet immédiat de la "Combinaison d'Imputation de Données". Selon les modèles de prédiction, parmi les recherches antérieures dans le domaine, la forêt aléatoire est plus précise. Dans cette optique, nous espérons intégrer ce modèle dans un système capable de prédire d'autres maladies dangereuses. Il y a de la place pour le progrès à l'avenir en ce qui concerne l'automatisation de l'analyse du diabète ou de toute autre maladie.

Dans les chapitres suivants, nous approfondirons davantage nos découvertes, tout en continuant à explorer de nouvelles opportunités de recherche dans ce domaine passionnant à l’aide d’autres jeux de données et techniques.

Chapitre 5

Propositions et évaluation (partie 2)

1. Introduction

Dans ce chapitre nous allons travailler avec 2 autres jeux de données tous en présentant 3 contributions majors. Chacune de ces contributions contribuera à notre compréhension des meilleures pratiques en matière de prédiction du diabète grâce à l'apprentissage automatique. En fin de compte, elles renforceront notre confiance dans la création de modèles prédictifs du diabète hautement performants pour améliorer les soins de santé et la qualité de vie des patients.

2. Jeux de données additionnels

2.1. Ensemble de données pour la prédiction du risque de diabète à un stade précoce

L'ensemble de données pour la prédiction du risque de diabète à un stade précoce « Early stage diabetes risk prediction dataset » a été obtenu auprès du référentiel de « UC Irvine Machine Learning Repository » et est disponible en libre accès[13], donné le 7/11/2020. Cet ensemble de données contient des informations sur les signes et symptômes des patients nouvellement diabétiques ou en passe de le devenir. Preuve de la méticulosité des processus de collecte de données utilisés dans la recherche scientifique, les données ont été recueillies de manière experte grâce à l'utilisation habile de questionnaires directs auprès de patients de l'hôpital du diabète de Sylhet, au Bangladesh, et de résultats de diagnostic qui ont été approuvées par un médecin.

L'ensemble de données comporte (cinq cent vingt lignes sur dix-sept colonnes) 16 caractéristiques uniques à valeur numérique et une seule variable d'objectif « target » appelée "classe", qui est soigneusement étiquetée comme "testé négatif" ou "testé positif" pour le diabète (Figure 5.1).

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 520 entries, 0 to 519
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Age                 520 non-null    int64 
 1   Gender              520 non-null    object
 2   Polyuria            520 non-null    object
 3   Polydipsia          520 non-null    object
 4   sudden weight loss  520 non-null    object
 5   weakness            520 non-null    object
 6   Polyphagia          520 non-null    object
 7   Genital thrush      520 non-null    object
 8   visual blurring     520 non-null    object
 9   Itching             520 non-null    object
 10  Irritability        520 non-null    object
 11  delayed healing     520 non-null    object
 12  partial paresis     520 non-null    object
 13  muscle stiffness    520 non-null    object
 14  Alopecia            520 non-null    object
 15  Obesity             520 non-null    object
 16  class               520 non-null    object
dtypes: int64(1), object(16)
memory usage: 69.2+ KB
```

No null values are found

FIGURE 5.1 : INFORMATIONS SUR L'ENSEMBLE DE DONNÉES DE L'HÔPITAL POUR DIABÉTIQUES DE SYLHET.

[13] https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset#.

2.2. Ensemble de données des Centres de contrôle et de prévention des maladies U.S.

L'ensemble de données provient d'une enquête téléphonique annuelle sur la santé réalisée par les Centers for Disease Control and Prevention[14] (CDC) afin de découvrir divers facteurs de risque. Cette enquête recueille des données sur les citoyens américains concernant leurs comportements à risque en matière de santé, les problèmes de santé chroniques et l'utilisation de services préventifs.

D'après la (Figure 5.2), l'ensemble de données comporte 253680 entrées et les colonnes de données sont au nombre de 22 au total. L'attribut diabetes_binary est la cible qui signifie être diabétique ou non.

```
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_binary       253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
```

FIGURE 5.2 : INFORMATIONS SUR L'ENSEMBLE DES DONNÉES DE CDC.

3. Contribution -3-

Le « Early stage diabetes risk prediction dataset » est utilisé dans cette contribution (voir section 2.1).

[14] https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system)

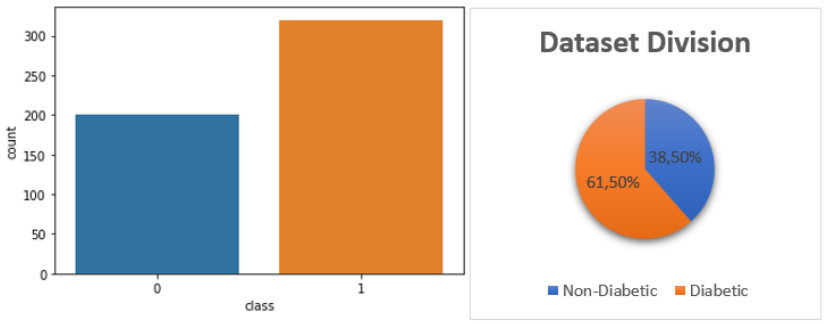

FIGURE 5.3 : NOMBRE ET POURCENTAGES DES DIABÉTIQUES ET NON-DIABÉTIQUES [21].

L'ensemble de données, qui contient un total de 16 caractéristiques et une variable cible nommée classe, comprend les attributs suivants :

1. L'âge « Age » : Âge en années allant de (20 ans ~ 65 ans) ;
2. Sexe « Gender » : Homme / Femme ;
3. polyurie « Polyuria » : Oui / Non ;
4. Polydipsie « Polydipsia » : Oui / Non ;
5. Perte de poids soudaine « Sudden weight loss » : Oui / Non ;
6. Faiblesse « Weakness » : Oui/ Non ;
7. polyphagie « Polyphagia » : Oui/ Non ;
8. Muguet génital « Genital thrush » : Oui/ Non ;
9. Trouble de la vision « Visual blurring » : Oui/ Non ;
10. Démangeaisons « Itching » : Oui/Non ;
11. Irritabilité « Irritability » : Oui/Non ;
12. Retard de cicatrisation « Delayed healing » : Oui/Non ;
13. Parésie partielle « Parésie partielle » : Oui/ Non ;
14. Raideur musculaire « Muscle stiffness » : Oui/ Non ;
15. Alopécie « Alopecia » : Oui/ Non ;
16. Obésité « Obesity » : Oui/ Non ;

Classe « Class » : Positive / Négative ;

3.1. Prétraitement des données

Avant d'utiliser les algorithmes de ML sur la collection de données, nous devons utiliser Pandas et NumPy pour analyser les données exploratoires. Cette phase traite les

données incohérentes afin de fournir des résultats plus fiables et plus précis. L'analyse des données exploratoires nous a permis d'effectuer les opérations suivantes :

3.1.1. Nettoyage des données

Nous allons améliorer la casse et le formatage des noms de colonnes. Pour effectuer des manipulations dans les « data frames », il est préférable d'avoir des 1 et des 0 au lieu de Oui et Non. Nous remplaçons donc toutes les chaînes "Oui" par 1 et toutes les chaînes "Non" par "0". Cette étape est cruciale pour permettre à l'algorithme d'apprentissage de comprendre les étiquettes et de les utiliser dans le processus d'apprentissage. LabelEncoder est utilisé pour convertir l'ensemble de données au format numérique [21].

-Sexe : Femme(0), Homme(1)

-tous : Non(0), Oui(1)

3.1.2. Vérification des valeurs manquantes

Comme le montre la (Figure 5.4), il n'y a pas de valeurs manquantes:

```
Age                 0
Gender              0
Polyuria            0
Polydipsia          0
sudden weight loss  0
weakness            0
Polyphagia          0
Genital thrush      0
visual blurring     0
Itching             0
Irritability        0
delayed healing     0
partial paresis     0
muscle stiffness    0
Alopecia            0
Obesity             0
class               0
```

FIGURE 5.4 : NOMBRE DE VALEURS MANQUANTES POUR CHAQUE ATTRIBUT [21].

3.1.3. Distribution des données

3.1.3.1. Distribution de fréquence en utilisant la tranche d'âge

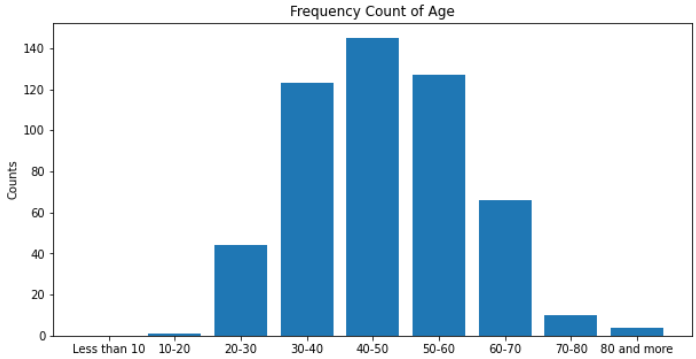

FIGURE 5.5 : NOMBRE DE FRÉQUENCES DE L'ÂGE [21].

On constate que la tranche d'âge des 40-50 ans a la plus forte prévalence de diabète, suivie par les 50-60 ans et les 30-40 ans. De même, les individus de moins de 20 ans ont les revenus les plus faibles, suivis par les personnes âgées de plus de 80 ans (Figure 5.5).

3.1.3.2. Sexe

Notre ensemble de données compte 328 points de données pour la classe 1 (hommes) et 192 points de données pour la classe 0 (femmes). Il y a plus d'hommes que de femmes (Figure 5.6).

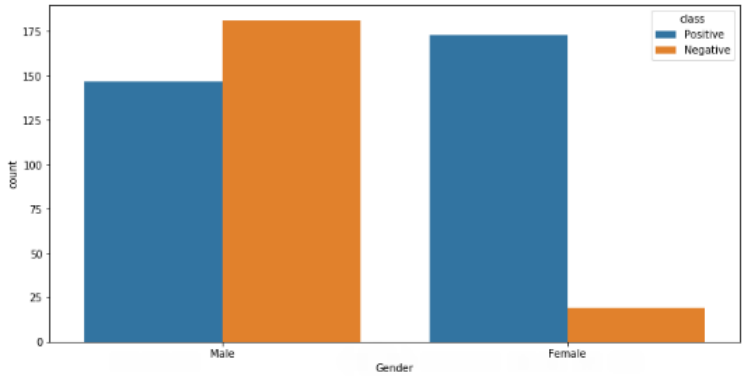

FIGURE 5.6 : DISTRIBUTION PAR SEXE [21].

3.1.3.3. Distribution de la polyurie

La polyurie se définit comme l'émission fréquente de grandes quantités d'urine (plus de 3 litres par jour, alors que la production quotidienne normale d'urine est de 1 à 2 litres chez l'adulte). La polyurie est le plus souvent due à un diabète sucré non contrôlé, qui provoque une diurèse osmotique, c'est-à-dire lorsque les niveaux de glucose sont si élevés que le glucose est excrété dans l'urine. L'eau suit passivement les concentrations de glucose, ce qui entraîne un débit urinaire anormalement élevé. Les causes les plus courantes en l'absence de diabète sucré sont une diminution de la sécrétion d'aldostérone due à une tumeur corticosurrénalienne et une polydipsie primaire (consommation excessive de liquides) [21].

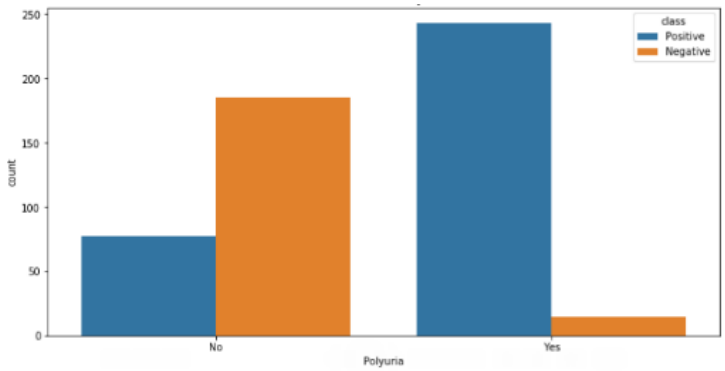

FIGURE 5.7 : DISTRIBUTION DE LA POLYURIE [21].

3.1.3.4. Distribution de la polydipsie

La polydipsie est un terme désignant une soif excessive, qui est l'un des premiers signes du diabète. Elle s'accompagne fréquemment d'une sécheresse buccale temporaire ou durable. Cependant, si vous avez constamment soif ou si votre soif est plus forte que d'habitude et persiste même après avoir bu, cela peut indiquer que quelque chose ne va pas dans votre corps. La soif excessive est l'un des "trois grands" signes du diabète sucré (polyurie, polydipsie, polyphagie) et peut être causée par un taux élevé de sucre dans le sang (hyperglycémie) [21].

La polydipsie (soif accrue) et la polyurie (besoin accru d'uriner) sont souvent associées.

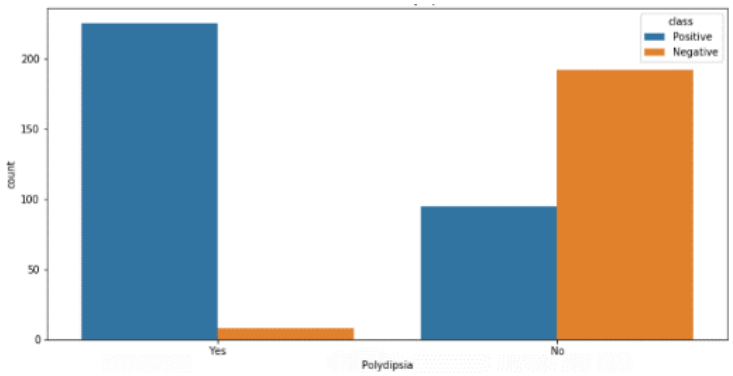

FIGURE 5.8 : DISTRIBUTION DE LA POLYDIPSIE [21].

3.1.3.5. Distribution de la perte de poids soudaine

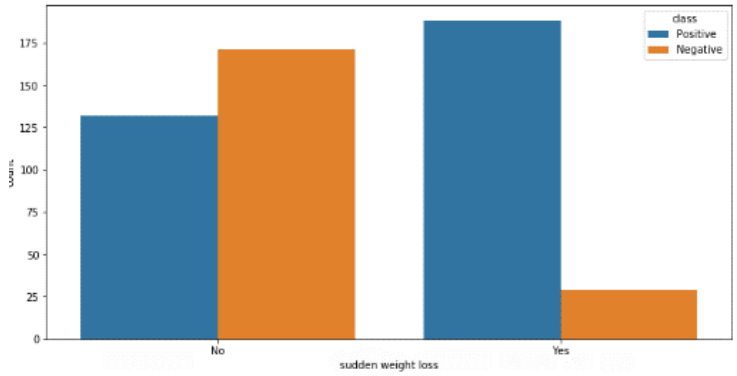

FIGURE 5.9 : DISTRIBUTION DE LA PERTE DE POIDS SOUDAINE [21].

3.1.3.6. Distribution de la faiblesse

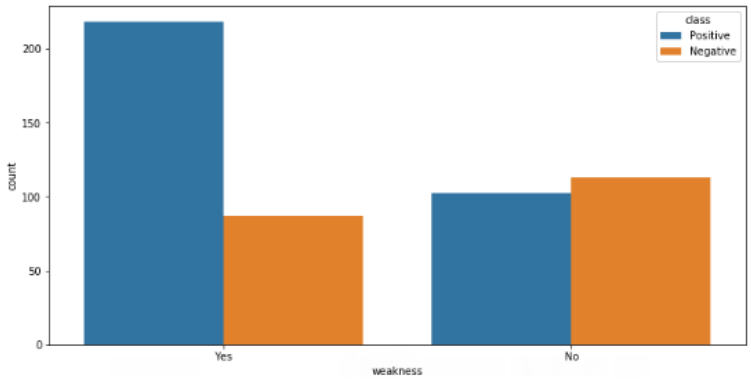

FIGURE 5.10 : DISTRIBUTION DE LA FAIBLESSE [21].

3.1.3.7. Répartition de la polyphagie

Le terme médical désignant une faim excessive ou extrême est la polyphagie, également connue sous le nom d'hyperphagie. Ce n'est pas la même chose que d'avoir un appétit accru après avoir fait de l'exercice ou d'autres formes d'activité physique. La polyphagie ne disparaîtra pas si vous mangez plus, même si votre niveau de faim revient à la normale après avoir mangé [21].

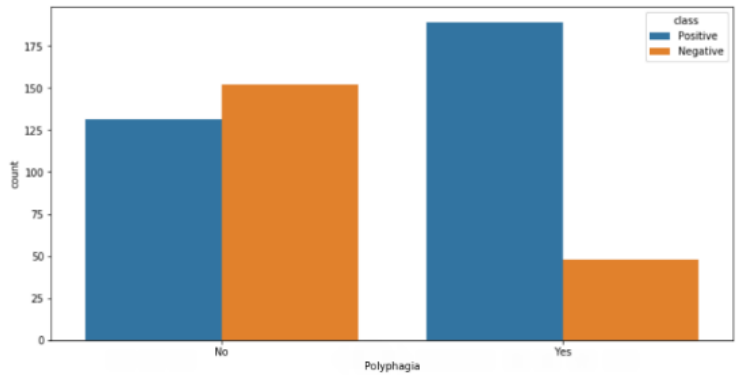

FIGURE 5.11 : DISTRIBUTION DE LA POLYPHAGIE [21].

3.1.3.8. Distribution du muguet génital

Le muguet (également connu sous le nom de candidose) est une infection à levures courante causée par Candida. Elle affecte principalement la zone vaginale, mais peut également toucher le pénis, et peut être irritante et douloureuse. De nombreux types de levures et de bactéries vivent naturellement dans la zone vaginale et causent rarement des problèmes. Le Candida est un champignon ressemblant à une levure qui se développe dans des environnements chauds et humides comme la bouche, les intestins, la zone vaginale et le prépuce du pénis. Le muguet est causé par une surabondance de Candida [21].

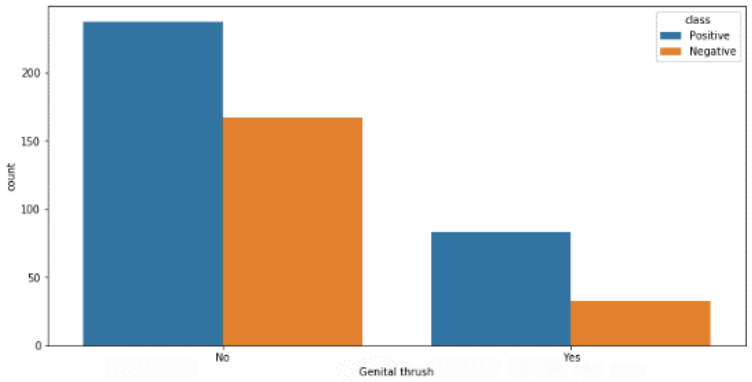

FIGURE 5.12 : RÉPARTITION DU MUGUET GÉNITAL [21].

3.1.3.9. Répartition des troubles visuels

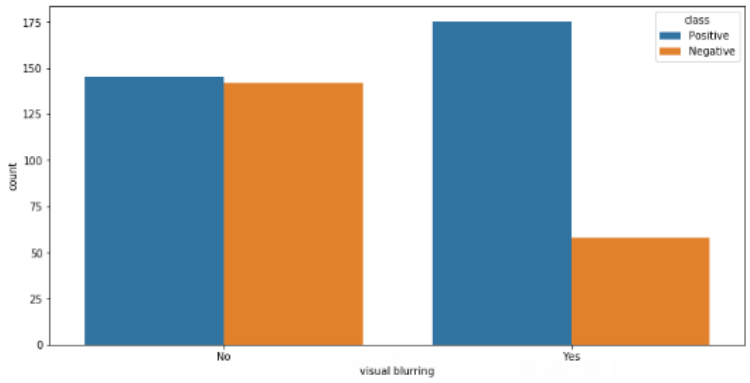

FIGURE 5.13 : DISTRIBUTION DU FLOU VISUEL [21].

3.1.3.10. Démangeaisons

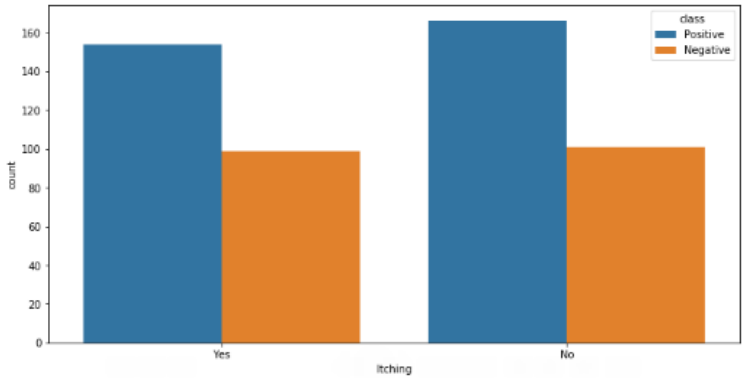

FIGURE 5.14 : DISTRIBUTION DES DÉMANGEAISONS [21].

3.1.3.11. Irritabilité

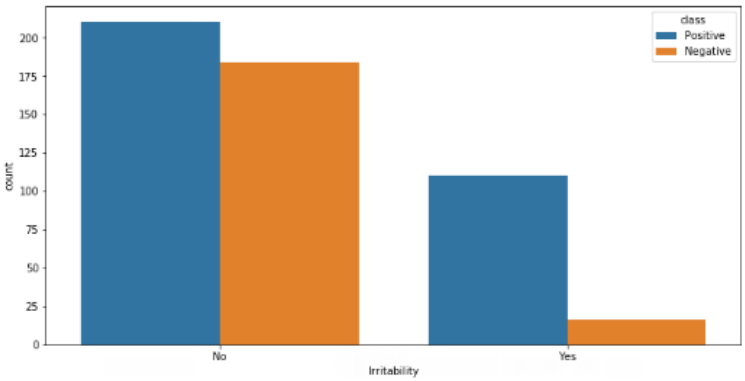

FIGURE 5.15 : DISTRIBUTION DE L'IRRITABILITÉ [21].

3.1.3.12. Retard de cicatrisation

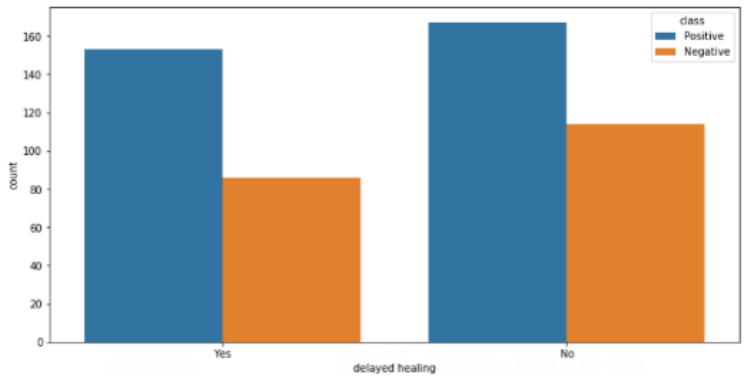

FIGURE 5.16 : DISTRIBUTION DE RETARD DE CICATRISATION [21].

3.1.3.13. Parésie partielle

L'affaiblissement d'un muscle ou d'un groupe de muscles est appelé parésie. On parle aussi de paralysie partielle ou légère. Contrairement à la paralysie, la

parésie permet de bouger les muscles. Ces mouvements sont simplement moins puissants que d'habitude [21].

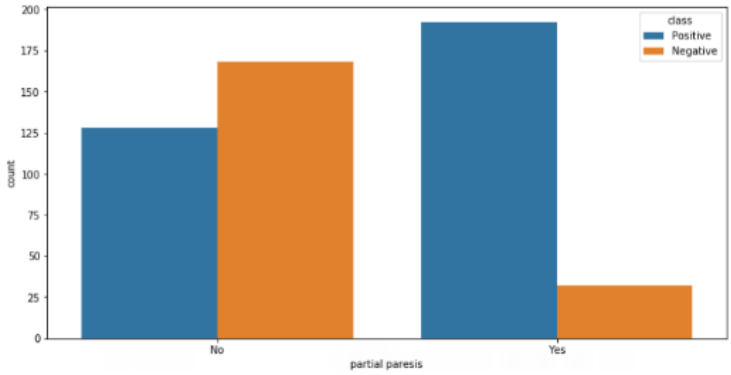

FIGURE 5.17 : DISTRIBUTION DE LA PARÉSIE PARTIELLE [21].

3.1.3.14. Raideur musculaire

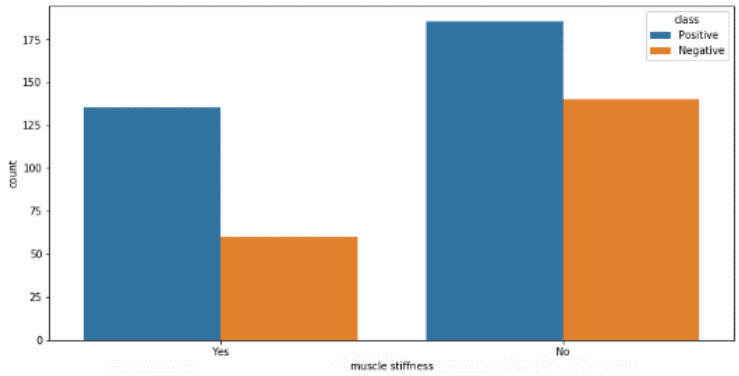

FIGURE 5.18 : DISTRIBUTION DE LA RAIDEUR MUSCULAIRE [21].

3.1.3.15. Alopécie

Plaques chauves soudaines qui peuvent se chevaucher. L'alopécie areata est une maladie dans laquelle le système immunitaire attaque les follicules pileux, ce qui peut être déclenché par un stress extrême. La perte de cheveux est le symptôme le plus courant [21].

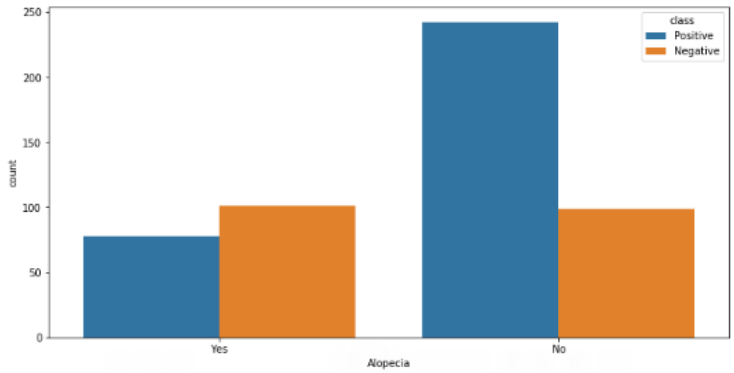

FIGURE 5.19 : DISTRIBUTION DE L'ALOPÉCIE [21].

3.1.3.16. Obésité

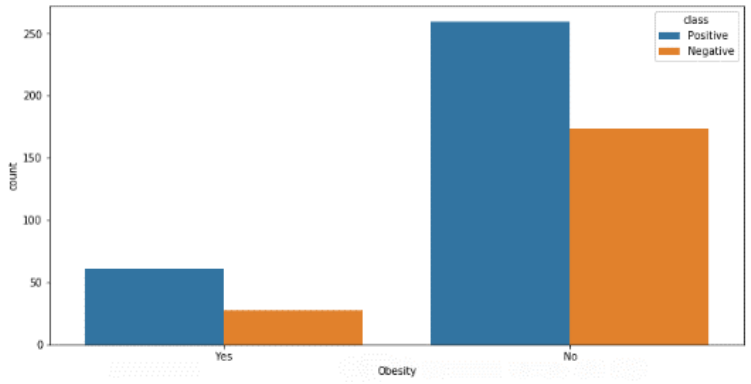

FIGURE 5.20 : DISTRIBUTION DE L'OBÉSITÉ [21].

3.1.4. Analyse de corrélation des caractéristiques par rapport à la classe cible

Nous aimerions examiner les données pour voir s'il existe un lien entre les caractéristiques et la classe d'étiquettes que nous recherchons. Comme le montre la figure ci-dessous, il n'y a pas de corrélation élevée qui pourrait être prise en considération.

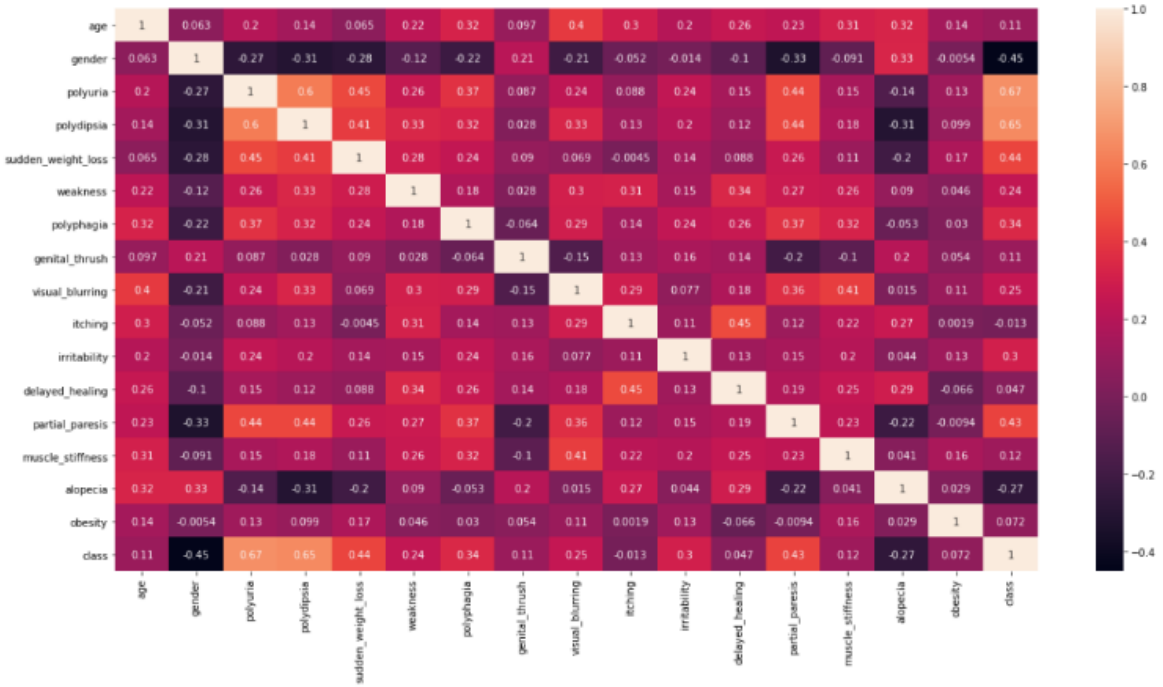

FIGURE 5.21 : CORRÉLATION ENTRE LES ATTRIBUTS [21].

3.2. Modélisation

Les modèles de prédiction du risque de diabète sont développés sur la base des données disponibles afin de prévenir l'apparition du diabète à l'avenir.

- Random Forest `(n_estimators = 10, criterion = 'entropy', random_state = 0)`,
- Support Vector Machine `(C=1.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=3, gamma='scale', kernel='linear',max_iter=-1, probability=True, random_state=0, shrinking=True, tol=0.001,verbose=False)`,
- K-Nearest Neighbors `(algorithm='auto', leaf_size=30, metric='minkowski', metric_params=None, n_jobs=None, n_neighbors=5, p=2, weights='uniform')`,
- Naïve Bayes `(priors=None, var_smoothing=1e-09)`,
- Decision Tree (*criterion = 'entropy', max_depth = 5, random_state = 0*), Logistic Regression `(C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, L1_ratio=None, max_iter=100, multi_class='auto', n_jobs=None, penalty='L2', random_state=None, solver='newton-cg', tol=0.0001, verbose=0,warm_start=False)` (Voir section 5.3 du chapitre 4 pour les définitions)
- XGBoost `(base_score=0.5,booster='gbtree', colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1, gamma=0, importance_type='gain', learning_rate=0.300000012, max_delta_step=0, max_depth=6, min_child_weight=1, missing=nan, monotone_constraints='()', n_estimators=100, n_jobs=8, num_parallel_tree=1, objective='binary:logistic', random_state=0, reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1)`

ont tous été utilisés pour modéliser le diabète [21].

3.2.1. XGBoost (Extreme Gradient Boosting)

Extreme Gradient Boosting est l'un des algorithmes de boosting qui a une vitesse d'exécution plus rapide et une meilleure performance du modèle par rapport aux autres algorithmes de boosting. Il utilise une approche dans laquelle de nouveaux modèles sont créés pour prédire les erreurs des modèles précédents, après quoi ils sont ajoutés les uns aux autres pour obtenir le modèle final. Un algorithme de descente de gradient est utilisé pour minimiser la perte lors de l'ajout de nouveaux modèles. Cet algorithme est utilisé pour les problèmes de classification et de régression [74].

Le test d'entraînement divisé est une méthode permettant de déterminer les performances d'un algorithme d'apprentissage automatique. Il peut être utilisé pour les problèmes de classification et de régression, ainsi que pour toute approche d'apprentissage supervisé. Dans le cadre de cette procédure, un ensemble de données est

adopté et divisé en deux sous-ensembles. L'ensemble de données de formation est le composant initial, qui est utilisé pour ajuster le modèle. Le second sous-ensemble n'est pas utilisé pour former le modèle ; il est alimenté par l'élément d'entrée de l'ensemble de données et les prédictions sont générées et comparées aux valeurs attendues. L'ensemble de données de test est la deuxième collection de données. Ici, la formation représente 80 % et le test 20 % [21].

3.3. Résultats expérimentaux et discussion de la contribution -3-

Après avoir entraîné nos sept algorithmes sur un ensemble de données de patients, nous présentons les résultats expérimentaux. Ces résultats sont destinés à évaluer les performances du classifieur. Ici nous avons utilisé une matrice de confusion ainsi que plusieurs mesures d'évaluation telles que la précision, la sensibilité, la spécificité, le score F1 et le MCC pour évaluer les performances des algorithmes d'apprentissage automatique aussi nous avons utilisé GridSearchCV 10-folds qui est un outil polyvalent pour l'ajustement des hyperparamètres dans scikit-learn.

L'exactitude (accuracy) d'un classifieur est une mesure de l'efficacité avec laquelle il prédit des événements sur la base de données d'apprentissage. Comme le montrent la figure 22, la capacité de prédiction du classifieur est représentée par la matrice de confusion [21].

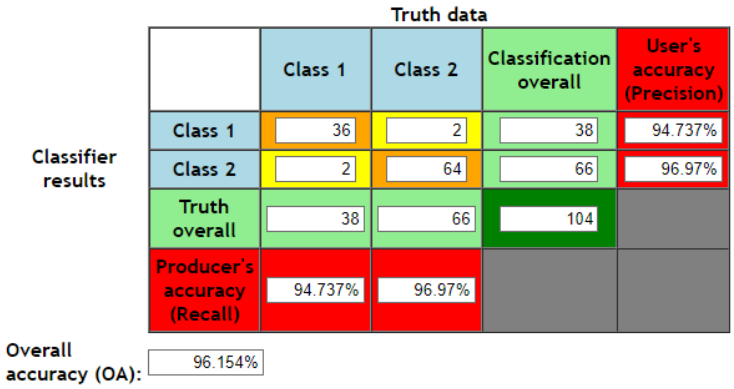

FIGURE 5.22 : MATRICE DE CONFUSION DE XGBOOST [21].

Le (Tableau 5.1) présente une comparaison de toutes les méthodes de catégorisation sur la base de divers paramètres. Selon ce tableau, XGBoost est la méthode la plus précise, suivie de RF. Par conséquent, le classifieur d'apprentissage automatique XGBoost est plus précis que les autres classifieurs pour prédire le risque de diabète. Les performances de chaque classifieur sont représentées sur un graphique à la (Figure 5.23) à l'aide de différentes mesures [21].

TABLEAU 5.1 : COMPARAISON DES PERFORMANCES DES MÉTHODES MISES EN ŒUVRE [21].

Modèle	TPR	SPC	PPV	ACC	F1	MCC
RF	0.9231	0.9692	0.9474	0.9519	0.9351	0.8971
SVM	0.8750	0.9531	0.9211	0.9231	0.8974	0.8367
KNN(n=5)	0.7600	1.0000	1.0000	0.8846	0.8636	0.7886

NB	0.8649	0.9104	0.8421	0.8942	0.8533	0.7708
DT	0.8372	0.9672	0.9474	0.9135	0.8889	0.8227
XGBoost	0.9474	0.9697	0.9474	0.9615	0.9474	0.9171
LR	0.8974	0.9538	0.9211	0.9327	0.9091	0.8559

L'algorithme XGBoost produit les meilleurs résultats, comme le montrent la (Figure 22 et 23). Par conséquent, le classifieur d'apprentissage automatique RF est plus performant que les autres classifieurs pour prédire la probabilité d'une maladie diabétique [21].

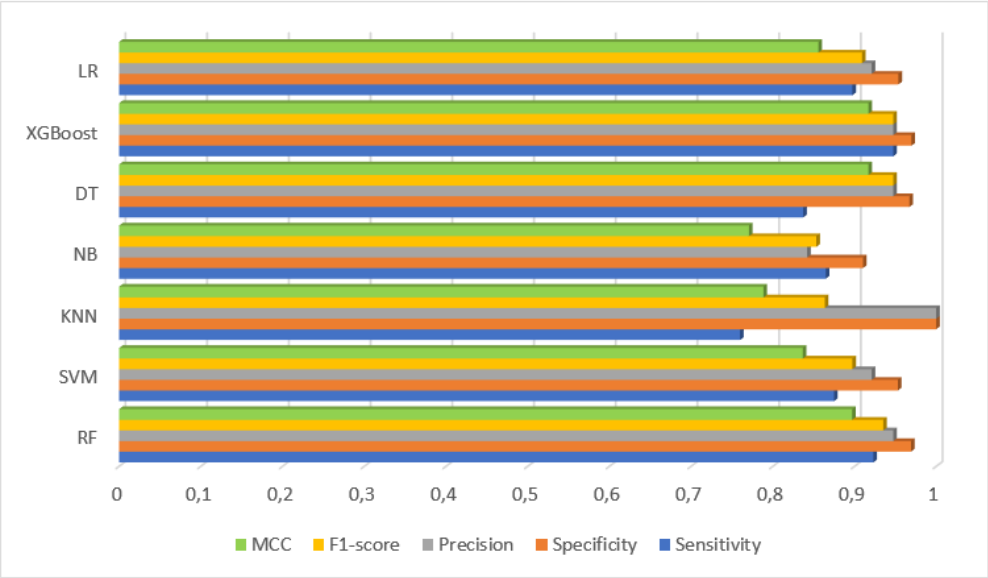

FIGURE 5.23 : PERFORMANCES DES TECHNIQUES DE CLASSIFICATION EN TERMES DE SENSIBILITÉ, DE SPÉCIFICITÉ, DE PRÉCISION, DE F1 & DE MCC [21].

La visualisation de ces « accuracies » dans la (Figure 5.24), qui présente une comparaison des algorithmes d'apprentissage automatique basée sur leurs exactitudes, nous aide à mieux comprendre les variations entre eux :

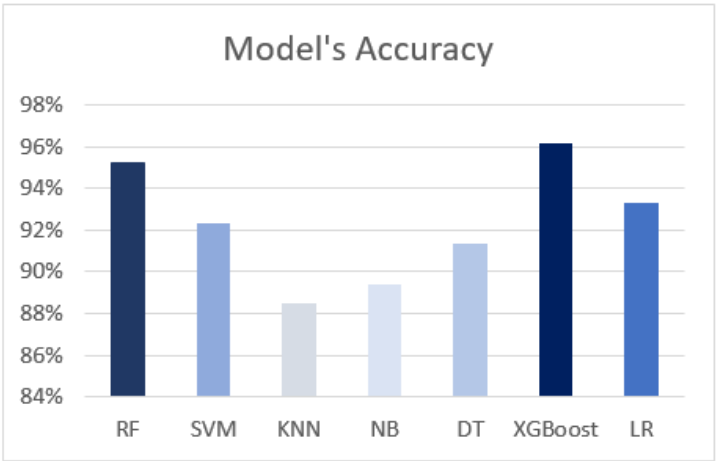

FIGURE 5.24 : EXACTITUDE DE CLASSIFICATION DE CHAQUE MODÈLE D'APPRENTISSAGE AUTOMATIQUE UTILISÉ [21].

Les résultats ont été comparés aux travaux existants et l'on constate que l'exactitude trouvée pour la régression logistique (93,27 %) est meilleure que celle de l'étude 2021 de Minhaz Uddin Emon et al. [175] (LR ~92 %), et que toutes les exactitudes trouvées sont également plus élevées que celles de Md. Faisal Faruque et al [145].

4. Experimentation -4-

Cette expérience utilise le « Early stage diabetes risk prediction dataset » (voir section 2.1). Cet ensemble de données assemblé comporte une bonne quantité de points de données pour les classifications positives et négatives, ce qui en fait un ensemble équilibré.

Cet équilibre garantit non seulement la validité de nos résultats, mais inspire également confiance dans la fiabilité de nos analyses, ce qui favorise la recherche de conclusions perspicaces et d'idées fondées sur des données probantes.

La méthodologie suivie est présentée dans la (Figure 5.25) :

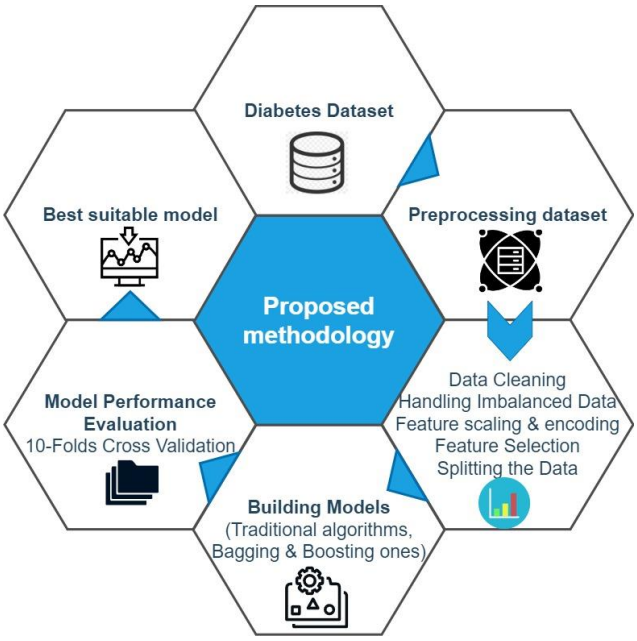

FIGURE 5.25 : MÉTHODOLOGIE DE LA CONTRIBUTION 4 [176].

4.1. Visualisation des données

Nous avons visualisé tous les attributs et voici un instantané de la distribution de l'âge (Figure 5.26).

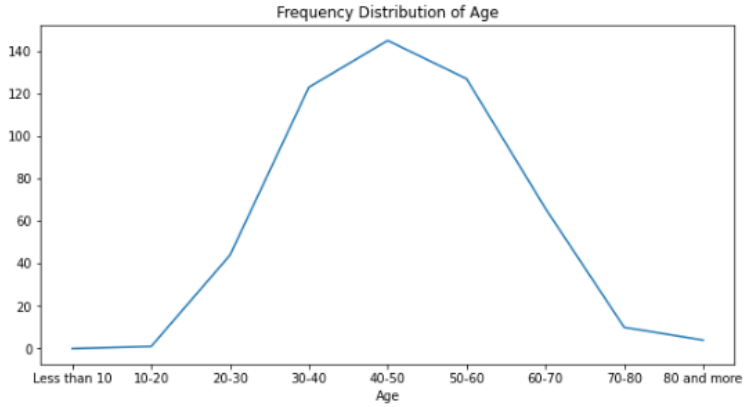

FIGURE 5.26 : DISTRIBUTION DE L'ÂGE [176].

Sur la base de nos observations, nous avons constaté que la tranche d'âge comprise entre 40 et 50 ans présente la prévalence la plus élevée de diabète. Cette information vitale met en lumière une fenêtre démographique critique dans laquelle l'incidence du diabète est plus prononcée.

4.2. Gestion du déséquilibre des classes et mise à l'échelle/encodage des caractéristiques

Lorsque l'on travaille avec des ensembles de données où une classe est nettement plus abondante que les autres, la gestion du déséquilibre des classes est un facteur crucial à prendre en compte.

Après avoir réalisé les diagrammes de comptage « countplots » pour chaque attribut (voir section 3.1.3), nous avons découvert que les caractéristiques variaient fortement entre les classes positives et négatives. À l'exception du sexe, où les femmes de la classe positive sont plus nombreuses que celles de la classe négative (Figure 5.6). Selon l'étude, les femmes ne sont pas plus susceptibles que les hommes de contracter le diabète. Cette situation semble donc déséquilibrée en faveur de la classe des femmes. Lorsque les femmes sont sous-représentées dans un ensemble de données par rapport aux hommes, les prédictions du modèle peuvent être faussées et les classes sous-représentées peuvent obtenir de moins bons résultats.

Différentes méthodes de suréchantillonnage de la classe minoritaire dans les ensembles de données déséquilibrés sont fournies par le module imblearn.over_sampling du paquet Python imbalanced-learn. Comme le montre la (Figure 5.27), nous avons choisi la méthode RandomOverSampler dans laquelle les instances de la classe minoritaire sont dupliquées de manière aléatoire jusqu'à ce qu'elles soient équilibrées avec celles de la classe majoritaire.

-Codage des variables catégorielles

La majorité des algorithmes d'apprentissage automatique ne peuvent pas utiliser ces variables directement car ils ont besoin d'entrées numériques.

Par conséquent, le codage des variables catégorielles constitue une étape essentielle du prétraitement. On utilise le codage par étiquette, qui est une approche simple dans laquelle une étiquette entière distincte est attribuée à chaque catégorie distincte [176].

```
from imblearn.over_sampling import RandomOverSampler
from collections import Counter


os=RandomOverSampler(0.9)
X_train_os,y_train_os=os.fit_resample(X_train_ns,y_train_ns)
print("The number of classes before fit {}".format(Counter(y_train_ns)))
print("The number of classes after fit {}".format(Counter(y_train_os)))
```

```
The number of classes before fit Counter({1: 139, 0: 14})
The number of classes after fit Counter({1: 139, 0: 125})
```

FIGURE 5.27 : SUR-ÉCHANTILLONNAGE DE LA CLASSE MINORITAIRE FÉMININE.

Le nombre d'échantillons dans la classe minoritaire sera augmenté jusqu'à ce qu'il atteigne 90% du nombre d'échantillons dans la classe majoritaire dans ce cas, où le ratio est fixé à RandomOverSampler(0.9).

4.3. Sélection des caractéristiques « Feature selection »

Lorsque les caractéristiques sont fortement corrélées, les coefficients du modèle peuvent être instables et il est difficile de comprendre les contributions relatives des différentes caractéristiques. Ce problème peut être résolu par la sélection des caractéristiques en conservant une seule caractéristique représentative des groupements associés. Nous avons découvert qu'il y avait une multicolinéarité, l'étape suivante est donc la sélection des caractéristiques (Figure 5.28).

Polyuria	class	0.665922
class	Polyuria	0.665922
	Polydipsia	0.648734
Polydipsia	class	0.648734
Polyuria	Polydipsia	0.598609
Polydipsia	Polyuria	0.598609
Itching	delayed healing	0.453316
delayed healing	Itching	0.453316
Gender	class	0.449233
class	Gender	0.449233
sudden weight loss	Polyuria	0.447207
Polyuria	sudden weight loss	0.447207
Polydipsia	partial paresis	0.442249
partial paresis	Polydipsia	0.442249
	Polyuria	0.441664

FIGURE 5.28 : CARACTÉRISTIQUES PRÉSENTANT LE COEFFICIENT DE CORRÉLATION LE PLUS ÉLEVÉ.

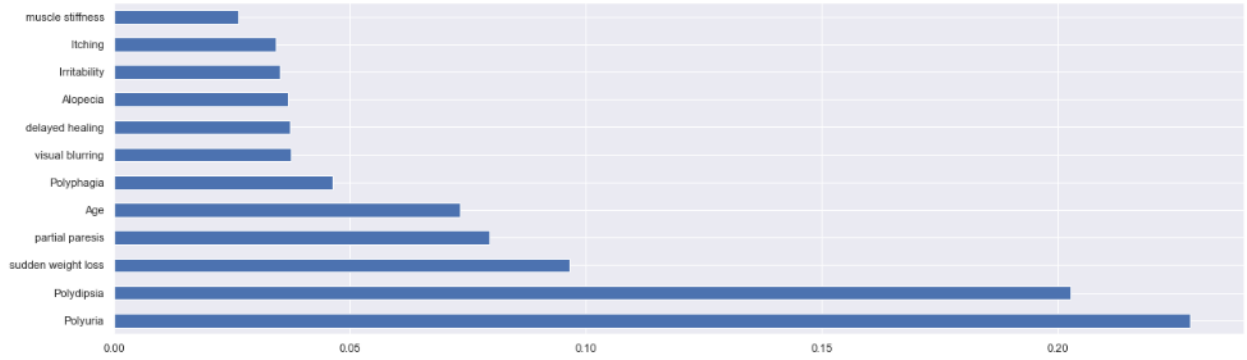

FIGURE 5.29 : LES 12 CARACTÉRISTIQUES LES PLUS IMPORTANTES EN UTILISANT SELECTKBEST.

Afin d'améliorer les performances du modèle et de réduire le surajustement, la sélection des caractéristiques est une phase cruciale de l'apprentissage automatique qui implique le choix d'un sous-ensemble de caractéristiques pertinentes à partir de l'ensemble de caractéristiques original, en particulier ici lorsque nous avons constaté la présence d'une multicolinéarité.

Nous avons utilisé deux algorithmes pour cette tâche. D'après la (Figure 5.29 et 5.30), nous constatons que les algorithmes sont plus efficaces que les algorithmes traditionnels. Depuis la (Figure 5.30), nous constatons que les deux algorithmes sont d'accord sur la majorité des caractéristiques, à l'exception de l'âge et de la polyphagie.

Les caractéristiques les plus significatives, selon le classifieur Extra Trees, sont la polyurie, la polydipsie, la perte de poids soudaine, la parésie partielle et l'âge. SelectKBest répertorie les symptômes suivants : polydipsie, polyurie, perte de poids soudaine, parésie partielle et polyphagie [176].

Feature_Scores	Feature_name
195.995811	Polydipsia
195.076582	Polyuria
106.234470	sudden weight loss
100.558154	partial paresis
72.614199	Polyphagia
54.466137	Irritability
39.959810	Age
35.551460	Alopecia
14.641500	Genital thrush
13.102942	muscle stiffness
12.556742	visual blurring
8.921300	weakness

FIGURE 5.30 : LES 12 CARACTÉRISTIQUES LES PLUS IMPORTANTES À L'AIDE D'EXTRA TREES [176].

4.4. Ensemble de données Train/Test_Split

Nous avons choisi une taille de test de 0,2, ce qui signifie que 20 % des données seront utilisées pour les tests et 80 % pour la formation.

- La matrice des caractéristiques d'apprentissage Xtrain contient 504 instances.
- La matrice des caractéristiques de test Xtest contient 127 instances.

4.5. Approches d'apprentissage automatique

Dans de nombreux domaines, y compris celui de la santé, on observe une augmentation de la fiabilité et de la précision des méthodes d'apprentissage automatique. Nous examinons plusieurs de ces algorithmes dans cette section.

Pour la prédiction, nous utiliserons neuf techniques d'apprentissage automatique supervisé. Nous avons testé certains algorithmes d'apprentissage automatique individuels ainsi que les algorithmes de bagging et de boosting, qui sont deux techniques d'apprentissage d'ensemble permettant d'améliorer les performances et la robustesse des modèles d'apprentissage automatique.

4.5.1. Les options choisies en matière d'algorithme

Nous avons utilisé :

- Algorithmes ML simples (K-voisins les plus proches KNN, Decision Tree DT et Régression Logistique LR).

- Les algorithmes de boosting construisent une séquence d'apprenants faibles dans laquelle chaque apprenant corrige les erreurs du précédent, ce qui permet d'obtenir un modèle final plus solide (XGBoost, Gradient Boost, Light GBM, AdaBoost).
- Plusieurs apprenants (arbres) sont formés indépendamment et combinés par le biais d'un vote ou d'une moyenne dans les algorithmes de regroupement « bagging algorithms » (Random Forest RF et Extra Trees).

Les performances de chaque classifieur sont présentées dans le (Tableau 5.2).

4.5.1.1. Renforcement du gradient (Gradient Boosting)

Il s'agit d'une approche d'apprentissage automatique qui permet de transformer les apprenants faibles en apprenants forts. Breiman Leo est à l'origine du concept de gradient boosting. Il se compose de trois éléments : l'apprenant faible, le modèle additif et la fonction de perte. Contrairement à la méthode d'apprentissage par bagging, qui crée des modèles séparément, le gradient boosting crée des modèles de manière séquentielle par itération afin de réduire l'erreur des modèles appris précédemment [177].

4.5.1.2. Arbres extrêmement aléatoires (Extra Trees)

Extra Trees est une extension de l'algorithme Random Forest. Il fonctionne en construisant un grand nombre d'arbres de décision, chacun formé sur un sous-ensemble aléatoire des données d'apprentissage et utilisant des sous-ensembles de caractéristiques aléatoires pour diviser les nœuds. Toutefois, contrairement aux forêts aléatoires, les arbres supplémentaires randomisent davantage le processus de division des nœuds en choisissant les seuils de division de manière aléatoire, sans essayer de trouver le meilleur seuil. Ce niveau supplémentaire d'aléatoire peut conduire à une meilleure généralisation et à une moindre sensibilité au bruit dans les données [87].

4.5.1.3. LightGBM (Light Gradient Boosting Machine)

LightGBM est un cadre de renforcement du gradient qui utilise une approche unique pour construire des arbres de décision. Il vise à optimiser l'efficacité de la formation et les performances prédictives en utilisant des techniques basées sur l'histogramme pour regrouper les valeurs des caractéristiques continues. Cette approche peut conduire à des temps de formation plus rapides et à la capacité de traiter des ensembles de données plus importants par rapport aux méthodes traditionnelles de renforcement du gradient [178].

4.5.1.4. AdaBoost (Adaptive Boosting)

AdaBoost est un algorithme de renforcement « boosting » populaire. Il commence par former un apprenant faible (par exemple, une souche de décision) sur

l'ensemble de données original. Les instances qui ont été mal classées reçoivent des poids plus élevés et un nouvel apprenant faible est formé sur cet ensemble de données modifié. Ce processus est répété plusieurs fois, chaque apprenant corrigeant les erreurs du précédent. La prédiction finale est obtenue en agrégeant les prédictions pondérées de tous les apprenants [179], [180].

4.6. Résultats expérimentaux de la contribution -4-

Cette section présente un résumé détaillé des résultats pour prédire le diabète. Les mesures de performance d'un algorithme d'apprentissage automatique nous montrent à quel point il fonctionne bien sur un certain ensemble de données. Par conséquent, nous pouvons évaluer les performances de plusieurs algorithmes en comparant leurs résultats. Divers indicateurs de qualité, tels que la précision, le rappel, le score F1 et l'exactitude, sont fréquemment utilisés. La matrice de confusion, qui résume les performances du modèle et compte le nombre de prédictions exactes et inexactes, peut être utilisée pour générer toutes ces mesures. Dans nos expériences, nous utilisons la validation croisée interne 10 fois (Tableau 5.2) [176].

TABLEAU 5.2 : ÉVALUATION DES MÉTRIQUES DES MODÈLES [176].

Modèle	Recall	Precision	Specificity	MCC	F1 score	Acc	Std. Deviation	10-Fold Mean Accuracy
Decision Tree	0.9333	0.9825	0.9851	0.9219	0.9573	96.06%	2.3967	96.9915%
Random Forest	0.9655	0.9825	0.9855	0.9525	0.9739	97.64%	2.0113	97.9414%
XGBoost	0.9500	1.0	1.0	0.9536	0.9744	97.64%	2.5727	97.4677%
Gradient Boost	0.9500	1.0	1.0	0.9536	0.9744	97.64%	3.7638	95.8829%
Extra Tree	**0.9655**	**0.9825**	**0.9855**	**0.9525**	**0.9739**	**97.64%**	**2.6335**	**97.9489%**
Light GBM	0.9661	1.0	1.0	0.9688	0.9828	98.43%	2.3736	97.6264%
AdaBoost	0.8793	0.8947	0.9130	0.7935	0.8870	89.76%	4.2412	87.4900%
K-Nearest Neighbor	0.8308	0.9474	0.9516	0.7863	0.8852	88.98%	7.0089	88.9236%
Logistic Regression	0.8448	0.8596	0.8841	0.7300	0.8522	86.61%	2.9894	89.3874%

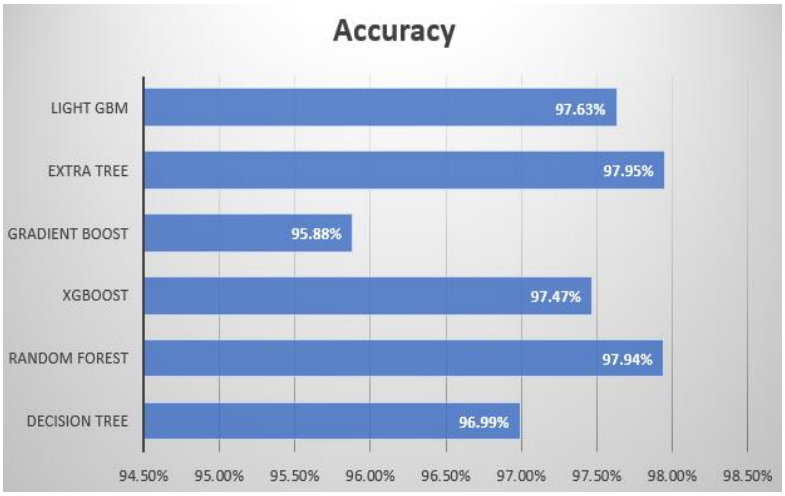

FIGURE 5.31 : COMPARAISON EXPÉRIMENTALE DES MEILLEURS MODÈLES EN TERMES D'EXACTITUDE [176].

La (Figure 5.31) illustre une comparaison des performances des meilleurs modèles en termes d'exactitude. Extra Tree a surpassé les autres modèles avec une exactitude de 97,95%.

4.7. Discussion de la contribution -4-

Après avoir évalué l'efficacité de chaque algorithme. Il a été découvert que l'exactitude du modèle Extra Tree est significativement plus élevée que celle des modèles de boosting et des modèles ML traditionnels. Extra Tree est donc considéré comme la meilleure technique d'apprentissage automatique supervisé avec une supérieure exactitude de 97,95 %.

Le (Tableau 5.3) présente les résultats des travaux antérieurs. Pour toutes les évaluations, l'ensemble de données de prédiction du risque de diabète au stade précoce a été utilisé. Nos modèles ont donc amélioré la précision de la prédiction précoce du diabète.

TABLEAU 5.3 : PERFORMANCE DES MODÈLES DES ARTICLES CONNEXES [176].

Travail	Modèle	Accuracy
D.Rani et al., 2021 [181]	LR RF	91.81% 97.82%
M. Banchhor, and P. Singh, 2021 [182]	RF	96.88%
B. Kumar Sahu, and N. Ghosh, 2022 [183]	Bagging Tree Boosted Tree Fine KNN	97.70% 96.20% 93.10%
L.Akter, and A. Ferdib, 2022 [184]	XGBoost	94.23%

5. Contribution -5-

Nous visons à utiliser trois algorithmes d'apprentissage automatique qui, au moment de la rédaction du présent document, sont les méthodes les plus fréquemment utilisées dans la littérature pour construire une classification binaire capable de prédire si une personne est atteinte de diabète ou non. Enfin, les résultats sont évalués en termes de performance et d'évolutivité.

Les hôpitaux disposent d'une grande quantité d'informations sur les patients. Ces données peuvent offrir beaucoup plus d'informations si elles sont récupérées avec succès, ce qui permet de prédire des maladies des mois, voire des années à l'avance. Dans l'apprentissage automatique, des modèles sont créés sur la base des données collectées et traitées

La (Figure 5.32) présente l'ensemble de la méthodologie proposée dans ce travail en utilisant l'ensemble de données des Centres de contrôle et de prévention des maladies U.S. (voir section 2.2) :

FIGURE 5.32 : MÉTHODOLOGIE PROPOSÉE [185].

L'ensemble de données choisi est d'abord chargé et, afin d'effectuer les analyses appropriées, un prétraitement et une analyse exploratoire des données sont nécessaires. La phase suivante consiste à appliquer des algorithmes d'apprentissage automatique pour créer des modèles de prédiction qui doivent être évalués à l'aide de diverses mesures de performance. Tout cela permet de choisir le modèle le plus approprié à la tâche et de prendre des décisions intelligentes [185].

5.1. Prétraitement et EDA de l'ensemble des données

Pour commencer, nous avons établi un prétraitement des données qui comprend les éléments suivants : Nettoyage des données (suppression des valeurs nulles si elles existent, élimination des doublons, gestion des valeurs manquantes, exploration et visualisation des données), traitement du déséquilibre par suréchantillonnage, analyse de corrélation, normalisation des données, encodage des variables catégorielles et sélection des fonctionnalités. Les données sont divisées en caractéristiques (X) et en cible (Y). Les ensembles de données prétraitées sont sauvegardés dans des fichiers pour faciliter la mémorisation des données et l'exécution du code. Les données n'étant pas équilibrées, un suréchantillonnage a été utilisé pour les équilibrer. Et pour faciliter le traitement, une mise à l'échelle standard est utilisée pour les X-caractéristiques [185].

Après la suppression des doublons, nous obtenons un jeu de données de dimensions (229474, 22) :

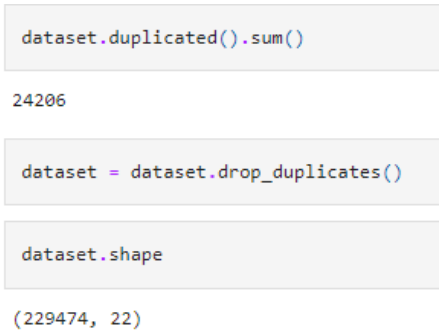

FIGURE 5.33 : ENSEMBLE DE DONNÉES APRÈS ÉLIMINATION DES DOUBLONS [185].

Exploration & Visualisation de la distribution pour les données catégorielles (Figure 5.34) avec utilisation de graphiques à barres (Bar Charts) : qui sont utiles pour comparer les fréquences de différentes catégories. Ceux-ci peuvent également fournir un premier aperçu de la sélection ultérieure des fonctionnalités, même si certains liens sont déjà clairement visibles ici. Par exemple, avec « HighBP » et « HighChol », il existe une nette dépendance entre eux et le statut diabétique.

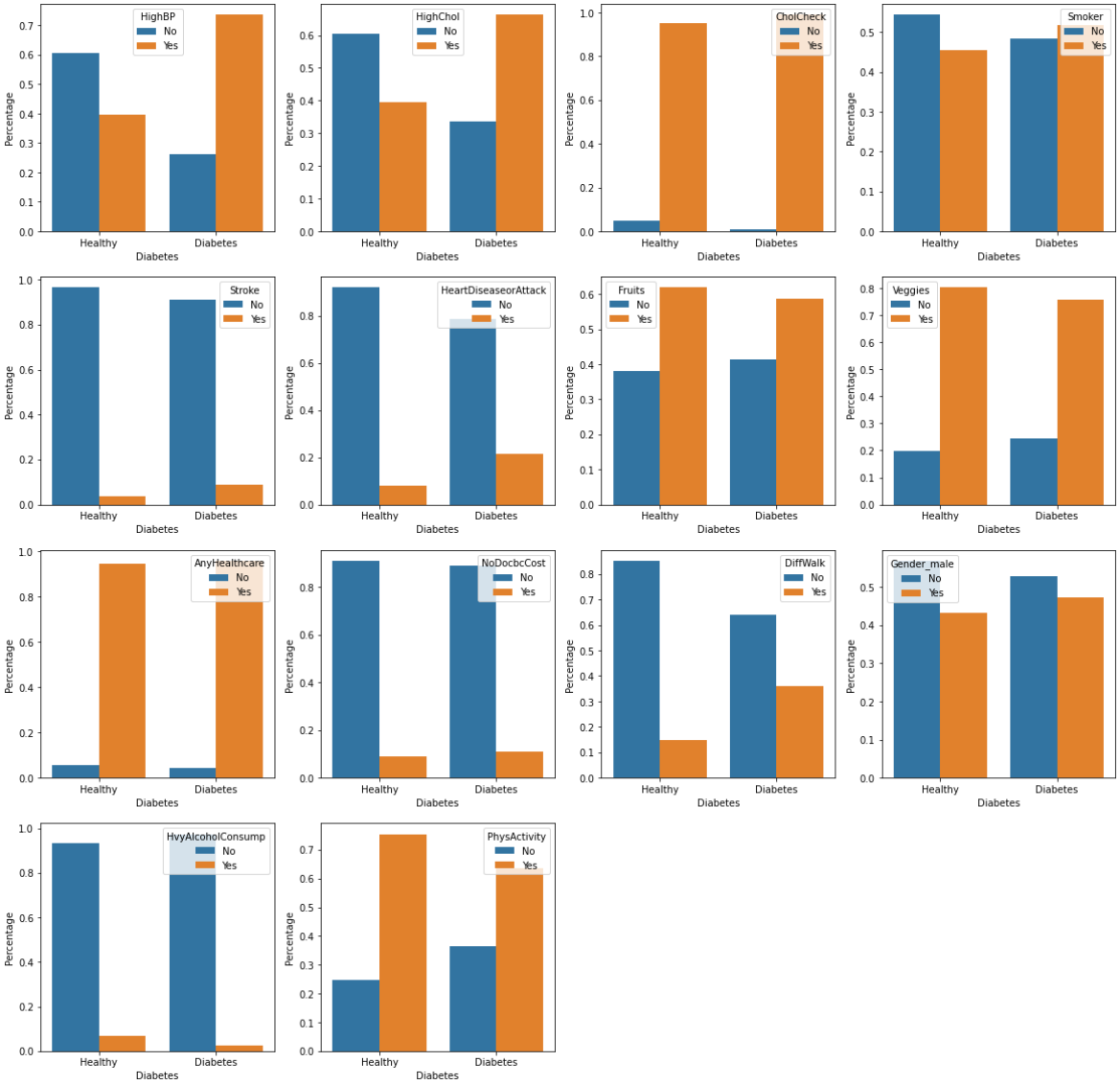

FIGURE 5.34 : DISTRIBUTION DES DONNÉES CATÉGORIELLES.

La fonction boxplot() de la bibliothèque seaborn basée sur matplotlib est utilisée ici pour la visualisation des caractéristiques numériques selon le statut diabétique (Figure 5.35). Après l'utilisation de boîtes à moustaches (Box Plots) qui sont utiles pour identifier les valeurs aberrantes et comprendre la répartition des données on constate qu'il existe des valeurs aberrantes évidentes pour les patients MentHlth, PhysHlth & IMC qui doivent être traitées (qui seront supprimées de l'ensemble de données). Il existe également un fort déséquilibre dans les données concernant le statut du diabète.

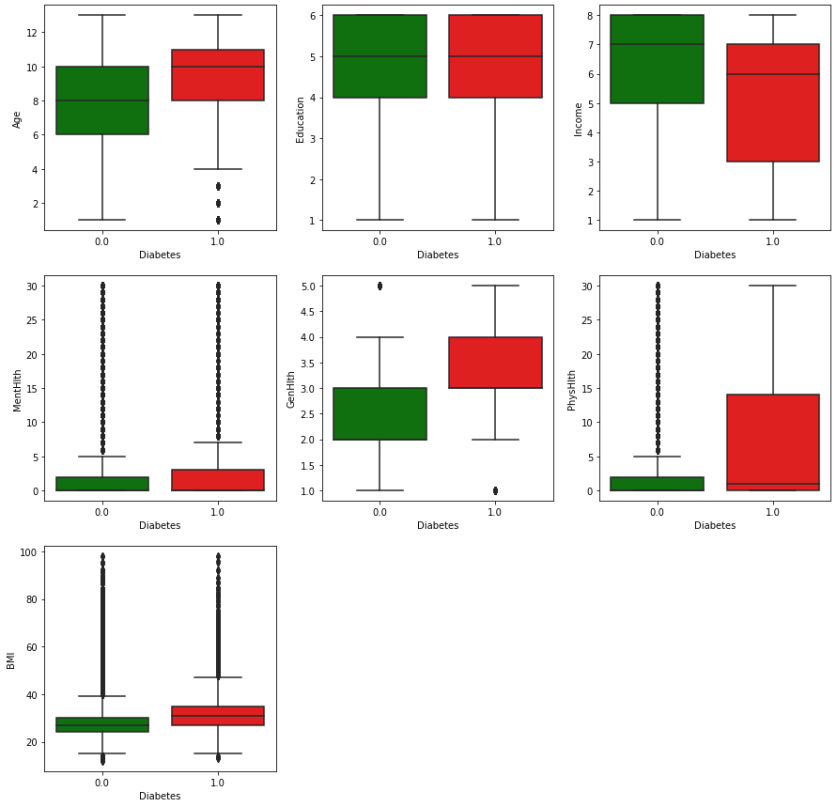

FIGURE 5.35 : DISTRIBUTION DES DONNÉES NUMÉRIQUES.

Le traitement des ensembles de données déséquilibrés « Handling imbalanced datasets » par suréchantillonnage « Oversampling ». Il s'agit d'une approche d'exploration et d'analyse des données qui permet de modifier des classes de données inégales afin de générer des ensembles de données équilibrés :

```
from imblearn.over_sampling import RandomOverSampler
sm = RandomOverSampler()
X_over_sampled , y_over_sampled = sm.fit_resample(X , y )
print("X_balanced shape is " , X_over_sampled.shape )
print("y_balanced shape is " , y_over_sampled.shape )
```

```
X_balanced shape is  (388754, 21)
y_balanced shape is  (388754,)
```

FIGURE 5.36 : SURMONTER LES DONNÉES DÉSÉQUILIBRÉES [185].

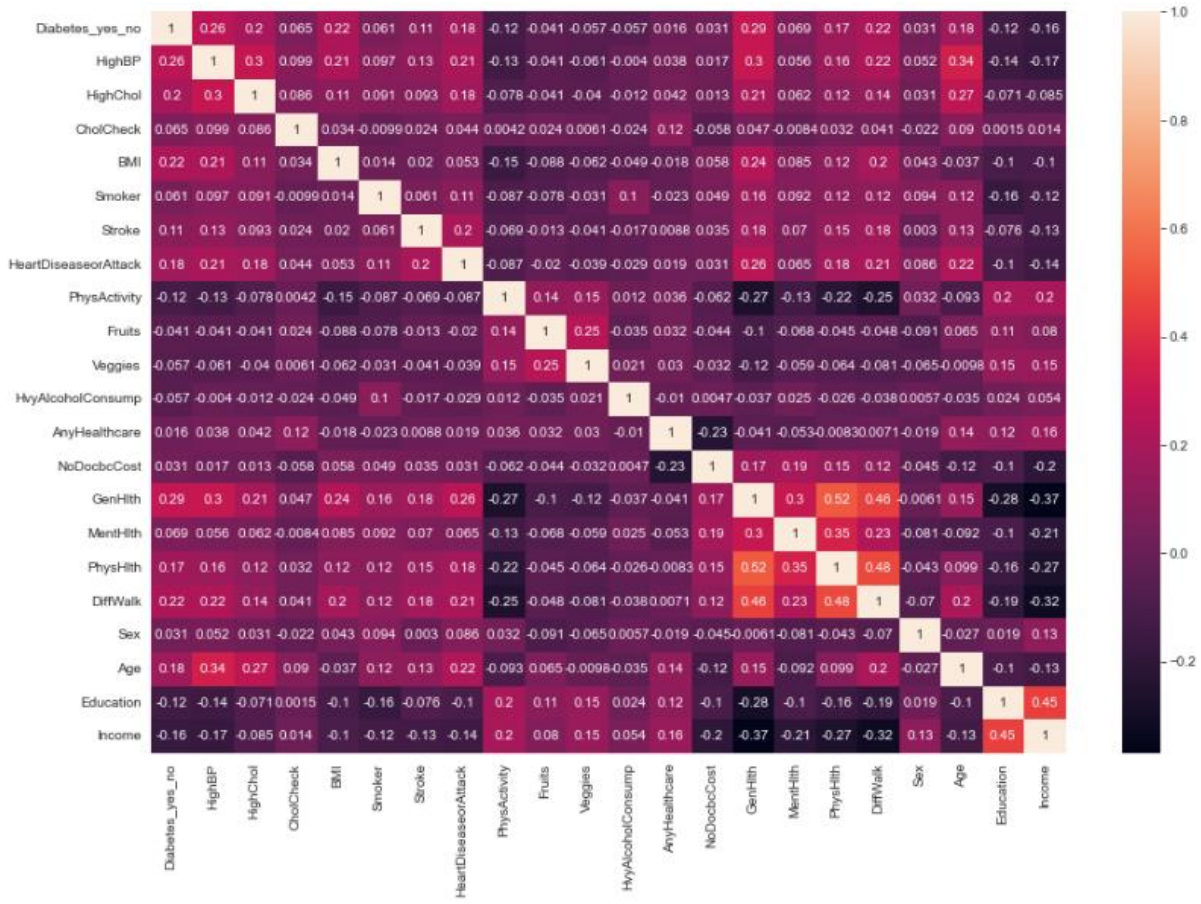

FIGURE 5.37 : GRAPHIQUE DE CORRÉLATION [185].

Après avoir vérifié la corrélation, nous pouvons conclure qu'il existe des faibles corrélations entre les attributs, ils sont donc relativement indépendants.

- ✓ Interprétabilité améliorée.
- ✓ Multicolinéarité réduite.
- ✓ Facilité de traitement et d'analyse.
- ✓ Meilleure généralisation.

Nous avons effectuez la standardisation des données avant la sélection des fonctionnalités pour garantir que toutes les fonctionnalités sont à une échelle comparable en utilisant StandardScaler(). Cela permet d'éviter que certaines caractéristiques ne dominent le modèle simplement parce qu'elles ont des valeurs numériques plus élevées (Figure 5.38).

```
from sklearn.preprocessing import StandardScaler
scalar = StandardScaler()
X_train = scalar.fit_transform(X_train)
X_test = scalar.fit_transform(X_test)
```

FIGURE 5.38 : STANDARDISATION DES DONNÉES.

Ensuite le codage de variables catégorielles qui est le processus de conversion de caractéristiques catégorielles en une représentation numérique pouvant être utilisée par les algorithmes d'apprentissage automatique. Il existe différentes techniques (encodage à chaud, encodage d'étiquettes, encodage cible et encodage ordinal) ici l'encodage ordinal est appliqué.

Concernant la phase de sélection de caractéristiques la téchnique (Recursive Feature Elimination with Cross-Validation « RFECV ») est appliquée :

```
rfc1 = RandomForestClassifier()

rfecv1 = RFECV(estimator=rfc1, cv=skf5, scoring='recall',
min_features_to_select=11 )
```

Pour effectuer RFE, nous avons d'abord besoin d'un estimateur. Ici, nous avons choisi RF - la condition d'arrêt est d'atteindre 11 entités.

La valeur 'rappel' (recall) a été choisie (joue un rôle déterminant dans la classification des maladies) et Skf5 signifie l'utilisation de StratifiedKFold.

Dans le DataFrame (Figure 5.39), les fonctionnalités sélectionnées sont marquées avec une « valeur sélectionnée » de « True » et une valeur « Classement » de « 1 ».

Columns	Feature_names	Selected	RFE_ranking
0	HighBP	True	1
1	HighChol	True	1
2	CholCheck	False	11
3	BMI	True	1
4	Smoker	True	1
5	Stroke	False	8
6	HeartDiseaseorAttack	False	6
7	PhysActivity	False	4
8	Fruits	True	1
9	Veggies	False	2
10	HvyAlcoholConsump	False	10
11	AnyHealthcare	False	9
12	NoDocbcCost	False	7
13	GenHlth	True	1
14	MentHlth	True	1
15	PhysHlth	True	1
16	DiffWalk	False	5
17	Gender_male	False	3
18	Age	True	1
19	Education	True	1
20	Income	True	1

FIGURE 5.39 : LES FONCTIONNALITÉS LES PLUS IMPORTANTES SÉLECTIONNÉES AVEC RFECV.

5.2. Modélisation

Des approches d'apprentissage automatique ont été appliquées dans ce travail pour prédire le pronostic du diabète précoce, et les conclusions attendues ont été obtenues. Nous avons divisé l'ensemble de données en 70 % pour la formation et 30 % pour le test, ce qui a permis d'obtenir respectivement 272127 et 116627 enregistrements. Avec réglage des hyperparamètres par RandomizedSearchCV 5-Folds.

En raison de sa grande capacité de classification, l'arbre de décision est une approche d'apprentissage automatique très répandue dans la profession médicale. Les forêts aléatoires sont également très utilisées car elles peuvent générer un grand nombre d'arbres de décision [185].

5.3. Résultats et discussion de la contribution -5-

Le diabète est la maladie la plus courante qui affecte les personnes. Il existe plusieurs symptômes et indicateurs du diabète qu'il convient de rechercher avant de procéder à une évaluation clinique. Bien que les nouveaux indicateurs identifiés soient faciles à repérer dans un manuel, la prédiction précise du diabète reste une difficulté importante. Pour résoudre ce problème, nous avons appliqué trois algorithmes d'apprentissage automatique supervisé à l'ensemble des données du CDC afin d'étudier la prédiction du diabète. L'ensemble de données a été divisé en 70 % de formation et 30 % de validation. Les mêmes données de formation et de validation ont été fournies à tous les algorithmes [185]. La figure suivante montre un exemple de matrice de confusion de RF:

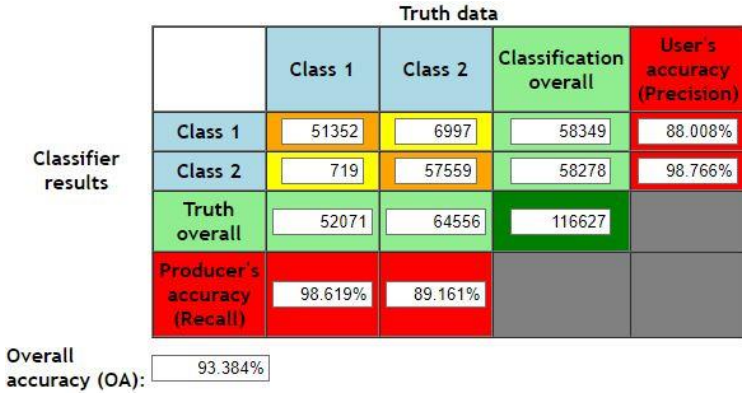

FIGURE 5.40 : MATRICE DE CONFUSION DE LA CLASSIFICATION BINAIRE DE RF [185].

La figure suivante montre la comparaison des performances de notre modèle en termes d'exactitude, de sensibilité, de précision et de score F1 après avoir fait une validation croisée de 5-folds. Comme on peut le constater, les meilleurs résultats ont été obtenus avec le classifieur RF, suivi du classifieur DT. LR s'est avéré le moins performant dans cette situation [185].

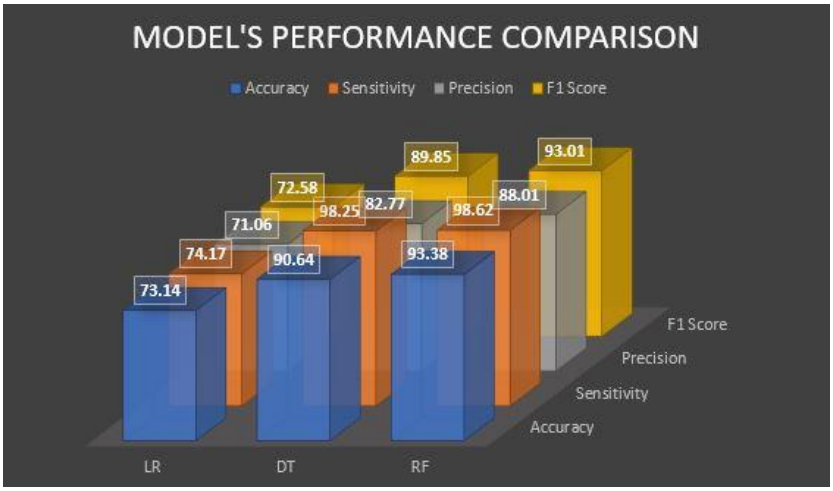

FIGURE 5.41 : GRAPHIQUE DE COMPARAISON DES PERFORMANCES DES MODÈLES [185].

6. Conclusion

Les approches d'apprentissage automatique et d'exploration de données sont utiles pour le diagnostic des maladies. La capacité à détecter le diabète à un stade précoce est essentielle pour que le patient reçoive un traitement adéquat. Le diagnostic précoce du diabète est l'un des problèmes les plus difficiles à résoudre dans le secteur des soins de santé. Dans le cadre de nos recherches, nous avons mis au point des méthodes permettant de prédire avec précision l'évolution du diabète. Cette thèse utilise l'apprentissage automatique pour prédire le diabète. Sept algorithmes d'apprentissage supervisé sont utilisés pour créer des modèles permettant de déterminer si le patient est diabétique ou non dans la contribution -3-. Tous les algorithmes ont produit de bons résultats. L'exactitude est une mesure de la perfection d'un algorithme. Une comparaison du modèle de prédiction XGBoost avec des études antérieures dans le domaine montre qu'il est plus précis (96,15 %) que les études antérieures.

Après un prétraitement bien réalisé et l'entraînement de plusieurs classifieurs avec toutes ces données, ceux-ci ont été évalués à l'aide de divers critères de performance, tels que l'exactitude, le score F1, le rappel et la précision. Avec une exactitude de 97,95 %, la méthode de classification Extra Trees donne de bons résultats dans la contribution -4-. Les performances de notre modèle démontrent un potentiel remarquable de prédiction du diabète par rapport à d'autres algorithmes de ML.

Dans la contribution -5- trois algorithmes d'apprentissage automatique supervisé ont été mis au point, et le plus performant a été retenu. Il s'agit du RF, qui a atteint une exactitude de 93,38 %, une sensibilité de 98,62 %, une précision de 88,01 % et un score F1 de 93,01 % lors des expériences. La principale contribution de cette expérimentation réside dans les stratégies utilisées pour prétraiter les données, ce qui nous permet de valider l'importance cruciale du prétraitement des données. Les erreurs, les redondances, les valeurs manquantes et les incohérences affectant toutes l'intégrité de l'ensemble, nous devons les traiter toutes pour obtenir un résultat plus précis. En outre, un vaste ensemble de données a été choisi pour accroître l'efficacité globale du système. Enfin, les algorithmes utilisés ont permis de créer des modèles prédictifs du diabète très performants.

Conclusion Générale

1. Conclusion

La prévalence du diabète de type 2 augmente à l'échelle mondiale, principalement en raison des processus d'urbanisation, de l'adoption de nouvelles habitudes alimentaires et des changements dans le mode de vie. Il est donc essentiel de pouvoir identifier rapidement le diabète, en particulier dans ses premiers stades.

Dans le diagnostic des maladies, les techniques d'apprentissage automatique (ML) et d'exploration de données sont importantes. Le diabète doit être détecté tôt pour que le patient puisse recevoir un traitement adéquat. Dans cette étude de thèse doctorale, les auteurs ont construit un mécanisme permettant de prédire de manière fiable le diabète. Prédire une maladie à ses débuts est un problème typique pour les scientifiques et les médecins, quelle que soit la nature de la maladie. Cela est principalement dû au fait que les pays défavorisés et en développement en sont peu conscients. Ignorer les conditions de contradiction, le pronostic précoce de la maladie et un traitement efficace peuvent sauver la vie d'une personne.

Cette thèse met en avant des avancées significatives dans la prédiction précoce du diabète à travers diverses contributions utilisant des ensembles de données variés.

- Contribution -1- (Base de données des Indiens Pima) : Utilisation de six méthodes de classification et d'un modèle hybride pour détecter le diabète à un stade précoce. Le modèle d'ensemble atteint une exactitude de 90,62 %, surpassant d'autres méthodes, démontrant ainsi son efficacité dans la détection précoce du diabète.
- Contribution -2- (Base de données des Indiens Pima) : L'utilisation de techniques innovantes de traitement des valeurs manquantes, notamment le mélange de techniques d'imputation, conduit à un modèle Random Forest avec une exactitude de 92%, dépassant les approches existantes.
- Contribution -3- (Ensemble de données Early stage diabetes risk prediction) : L'évaluation de 7 techniques majeures révèle les performances exceptionnelles de XGBoost avec un score F1 de 94,74% et une exactitude de 96,15 %. Ces résultats soulignent la pertinence et l'efficacité de l'approche dans la prédiction du diabète.
- Contribution -4- (Ensemble de données Early stage diabetes risk prediction) : L'algorithme Extra Trees atteint une exactitude exceptionnelle de 97,95 %, surpassant significativement d'autres modèles. Cela ouvre des perspectives pour des interventions médicales ciblées et préventives.
- Contribution -5- (Ensemble de données des Centres de contrôle et de prévention des maladies U.S.) : L'utilisation d'un ensemble de données plus vaste, notamment le Behavioral Risk Factor Surveillance System, améliore l'efficacité globale du système. Le modèle Random Forest se distingue en anticipant presque tous les exemples de l'ensemble de test correctement avec une exactitude de 93.38%.

En résumé, la thèse explore différentes approches, modèles, et ensembles de données, mettant en évidence des avancées significatives dans la prédiction précoce du diabète à l'aide de l'apprentissage automatique.

2. Perspectives futures

Pour nos perspectives futures, on vise d'ajouter et d'accentuer nos recherches sur les aspects suivants :

- L'introduction également davantage de modèles d'apprentissage automatique et d'apprentissage profond pour obtenir de meilleurs résultats. Nous améliorerons les algorithmes à l'avenir pour accroître l'efficacité et les performances du système.
- À l'avenir, avec l'aide d'un hôpital ou d'un institut médical, nous créerons un ensemble de données sur le diabète et viserons de meilleurs résultats.
- Nous essaierons également de travailler sur certaines fonctionnalités pour aider à lutter contre le diabète de manière encore plus efficace.
- Étendre notre travail à d'autres maladies chronique en relation directe avec le diabète comme la Maladie Rénale Chronique dont on a déjà commencer le travail sur (papier : Prédiction et Stadification de la Maladie Rénale Chronique à l'aide de l'Algorithme de Forêt Aléatoire Optimisé [189]) pour obtenir un système de diagnostic complet.

L'amalgamation d'une expertise pluridisciplinaire, regroupant des professionnels de la santé, des scientifiques des données et des spécialistes du domaine, est impérative pour surmonter ces défis et exploiter pleinement le potentiel de l'IA dans la prédiction du diabète. En surmontant ces obstacles, le domaine peut exploiter le pouvoir transformateur de l'IA pour favoriser des interventions de santé personnalisées et préventives, révolutionnant ainsi la gestion et la prévention du diabète à l'échelle mondiale. Une enquête ultérieure utilisant un ensemble de données élargi ainsi que la construction d'une boîte à outils automatisée devraient permettre d'accélérer considérablement l'identification rapide du diabète sucré, au bénéfice des professionnels de la santé et des personnes en quête d'une évaluation médicale.

Productions scientifiques

Revue Scientifique internationale

- S. Samet, M. R. Laouar, I. Bendib, and S. Eom, "Analysis and Prediction of Diabetes Disease Using Machine Learning Methods," *International Journal of Decision Support System Technology*, vol.14, no. 1, pp. 1–19, Jan. 2022, doi: 10.4018/IJDSST.303943.

Communications internationales

- S. Samet, M. R. Laouar, and I. Bendib, "Analysis and Prediction of Diabetes Disease using Machine Learning," in *International Conference on Software Engineering and New Technologies (ICSENT)*, July 2021.
- S. Samet, M. R. Laouar, and I. Bendib, "Diabetes mellitus early stage risk prediction using machine learning algorithms," in *2021 International Conference on Networking and Advanced Systems (ICNAS)*, Oct. 2021, pp. 1–6, doi:10.1109/ICNAS53565.2021.9628955.
- S. Samet, M. R. Laouar, and I. Bendib, "Use of Machine Learning Techniques to Predict Diabetes at an Early Stage," in *2021 International Conference on Networking and Advanced Systems (ICNAS)*, Oct. 2021, pp. 1–6, doi: 10.1109/ICNAS53565.2021.9628903.
- S. Samet, M. R. Laouar, and I. Bendib, "Predicting and Staging Chronic Kidney Disease using Optimized Random Forest Algorithm," in *International Conference on Information Systems and Advanced Technologies (ICISAT)*, Dec. 2021, pp. 1–8, doi: 10.1109/ICISAT54145.2021.9678441.
- S. Samet, M. R. Laouar, and I. Bendib, "Comparative Analysis of Diabetes Mellitus Predictive Machine Learning Classifiers," in *International Conference on Computing and Information Technology*, 2022, pp. 302–317, doi: 10.1007/978-3-031-25344-7_27.
- S. Samet and M. R. Laouar, "Building Risk Prediction Models for Diabetes Decision Support System," in *International Conference on Decision Support System Technology*, 2023, vol. 30, no. 4, pp. 171–181, doi: 10.1007/978-3-031-32534-2_13.

Communication nationale

- S. Samet and M. R. Laouar, "Enhancing early-stage risk prediction of diabetes mellitus through feature selection bagging and boosting techniques," in *National Conference on Artificial Intelligence (NCAI)*, 19-20 Dec.2023, pp. 1-6.

Activités scientifiques

- ✓ **Stage de perfectionnement à l'étranger courte durée :** Università degli studi di Milano, Février 2023.
- ✓ **Atelier :** SAMET Sarra. ChatGPT pour la recherche scientifique : Semaine universitaire de l'intelligence artificielle au service de la communauté, Tébessa du 16 au 19 avril, 2023.

Références bibliographiques

Références bibliographiques

[1] IDF, IDF Diabetes Atlas 10th edition, vol. 102, no. 2. 2021.

[2] M. Belhadj et al., "BAROMÈTRE Algérie : enquête nationale sur la prise en charge des personnes diabétiques," Médecine des Mal. Métaboliques, vol. 13, no. 2, pp. 188–194, Mar. 2019, doi: 10.1016/S1957-2557(19)30055-0.

[3] L. Chaves and G. Marques, "Data mining techniques for early diagnosis of diabetes: A comparative study," Appl. Sci., vol. 11, no. 5, pp. 1–12, 2021, doi: 10.3390/app11052218.

[4] X. Wei, F. Jiang, F. Wei, J. Zhang, W. Liao, and S. Cheng, "An ensemble model for diabetes diagnosis in large-scale and imbalanced dataset," ACM Int. Conf. Comput. Front. 2017, CF 2017, pp. 71–78, 2017, doi: 10.1145/3075564.3075576.

[5] S. Mishra, S. Hanchate, and Z. Saquib, "Diabetic retinopathy detection using deep learning," Proc. Int. Conf. Smart Technol. Comput. Electr. Electron. ICSTCEE 2020, pp. 515–520, 2020, doi: 10.1109/ICSTCEE49637.2020.9277506.

[6] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," in 2019 1st International Informatics and Software Engineering Conference (UBMYK), Nov. 2019, pp. 1–4, doi: 10.1109/UBMYK48245.2019.8965556.

[7] H. Cheng, J. Zhu, P. Li, and H. Xu, "Combining knowledge extension with convolution neural network for diabetes prediction," Eng. Appl. Artif. Intell., vol. 125, no. November 2022, 2023, doi: 10.1016/j.engappai.2023.106658.

[8] F. R. Liza et al., "An Ensemble Approach of Supervised Learning Algorithms and Artificial Neural Network for Early Prediction of Diabetes," 2021 3rd Int. Conf. Sustain. Technol. Ind. 4.0, STI 2021, vol. 0, pp. 18–19, 2021, doi: 10.1109/STI53101.2021.9732413.

[9] Z. Xu and Z. Wang, "A Risk prediction model for type 2 diabetes based on weighted feature selection of random forest and xgboost ensemble classifier," 11th Int. Conf. Adv. Comput. Intell. ICACI 2019, pp. 278–283, 2019, doi: 10.1109/ICACI.2019.8778622.

[10] M. De Bois, M. A. El Yacoubi, and M. Ammi, "Enhancing the Interpretability of Deep Models in Healthcare through Attention: Application to Glucose Forecasting for Diabetic People," Int. J. Pattern Recognit. Artif. Intell., vol. 35, no. 12, 2021, doi: 10.1142/S0218001421600065.

[11] N. Sahnine and Y. Yahiaoui, "Analyse des moyens à mettre en œuvre pour lutter contre le diabète : Cas CHU l'hôpital belloua Tizi- Ouzou," Mouloud Mammeri de Tizi-Ouzou, 2018.

[12] H. Shi, "Exploratory analysis of the hypertext structure linked to diabetes," Sorbonne Université, 2020.

[13] FID, L'atlas du diabète de la fid 9ème édition. 2019.

[14] A. K. Lamdjadani and A. Bouazza, "Étude epidemiologique sur les facteurs de risque associes au diabete de type 2," Abdelhamid Ibn Badis, Mostaganem, 2017.

[15] G. S. Chakraborty, D. Singh, M. Rakhra, S. Batra, and A. Singh, "Covid-19 and Diabetes Risk Prediction for Diabetic Patient using Advance Machine Learning Techniques and Fuzzy Inference System," Proc. 5th Int. Conf. Contemp. Comput. Informatics, IC3I 2022, pp. 1212–1219, 2022, doi: 10.1109/IC3I56241.2022.10073256.

[16] S. A. Jothi and J. A. Samath, "Enhanced Feed Forward Neural Network with Adam Optimization Model (Efnnao) for Predicting the Type 2 Diabetes Using Internet of Things," Proc. 5th Int. Conf. Contemp. Comput. Informatics, IC3I 2022, pp. 412–416, 2022, doi: 10.1109/IC3I56241.2022.10073047.

[17] S. Gündoğdu, "Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique," Multimed. Tools Appl., 2023, doi: 10.1007/s11042-023-15165-8.

[18] F. K. Dosilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," 2018 41st Int.

Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2018 - Proc., pp. 210–215, 2018, doi: 10.23919/MIPRO.2018.8400040.

[19] C. Krueger et al., “Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions,” Curr. Rev. Musculoskelet. Med., vol. 13, no. 1, pp. 69–76, 2020.

[20] T. Davenport and R. Kalakota, “The potential for artificial intelligence in healthcare,” Futur. Healthc. J., vol. 6, no. 2, pp. 94–98, 2019, doi: 10.7861/futurehosp.6-2-94.

[21] S. Samet, M. R. Laouar, and I. Bendib, “Use of Machine Learning Techniques to Predict Diabetes at an Early Stage,” in 2021 International Conference on Networking and Advanced Systems (ICNAS), Oct. 2021, pp. 1–6, doi: 10.1109/ICNAS53565.2021.9628903.

[22] S. M. Hasan Mahmud, M. A. Hossin, M. Razu Ahmed, S. R. H. Noori, and M. N. I. Sarkar, “Machine learning based unified framework for diabetes prediction,” ACM Int. Conf. Proceeding Ser., pp. 46–50, 2018, doi: 10.1145/3297730.3297737.

[23] P. N. Thotad, G. R. Bharamagoudar, and B. S. Anami, “Diabetes disease detection and classification on Indian demographic and health survey data using machine learning methods,” Diabetes Metab. Syndr. Clin. Res. Rev., vol. 17, no. 1, p. 102690, 2023, doi: 10.1016/j.dsx.2022.102690.

[24] F. M. Okikiola, O. S. Adewale, and O. O. Obe, “A DIABETES PREDICTION CLASSIFIER MODEL USING NAIVE BAYES ALGORITHM,” FUDMA J. Sci., vol. 7, no. 1, pp. 253–260, Feb. 2023, doi: 10.33003/fjs-2023-0701-1301.

[25] J. Hou, Y. Sang, Y. Liu, and L. Lu, “Feature Selection and Prediction Model for Type 2 Diabetes in the Chinese Population with Machine Learning,” ACM Int. Conf. Proceeding Ser., 2020, doi: 10.1145/3424978.3425085.

[26] H. M. Deberneh and I. Kim, “Prediction of type 2 diabetes based on machine learning algorithm,” Int. J. Environ. Res. Public Health, vol. 18, no. 6, pp. 9–11, 2021, doi: 10.3390/ijerph18063317.

[27] M. E. Castel, “The Road to Artificial Super-intelligence: Has International Law a Role to Play?,” Can. J. Law Technol., vol. 14, no. 1, p. 1, 2016, [Online]. Available: https://ojs.library.dal.ca/CJLT/article/download/7211/6256.

[28] A. Kaplan and M. Haenlein, “Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence,” Bus. Horiz., vol. 62, no. 1, pp. 15–25, Jan. 2019, doi: 10.1016/j.bushor.2018.08.004.

[29] J. Alzubi, A. Nayyar, and A. Kumar, “Machine Learning from Theory to Algorithms: An Overview,” J. Phys. Conf. Ser., vol. 1142, no. 1, pp. 0–15, 2018, doi: 10.1088/1742-6596/1142/1/012012.

[30] S. Ray, “Introduction to Machine Learning and Different types of Machine Learning Algorithms,” Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019, pp. 35–39, 2019.

[31] S. J. Maceachern and N. D. Forkert, “Machine learning for precision medicine,” Genome, vol. 64, no. 4, pp. 416–425, 2021, doi: 10.1139/gen-2020-0131.

[32] V. Kaul, S. Enslin, and S. A. Gross, “History of artificial intelligence in medicine,” Gastrointest. Endosc., vol. 92, no. 4, pp. 807–812, 2020, doi: 10.1016/j.gie.2020.06.040.

[33] V. S. Kadam, “Regression Techniques in Machine Learning &Applications: A Review,” Int. J. Res. Appl. Sci. Eng. Technol., vol. 8, no. 10, pp. 826–830, 2020, doi: 10.22214/ijraset.2020.32019.

[34] S. H. Shetty, S. Shetty, C. Singh, and A. Rao, “Supervised machine learning: Algorithms and applications,” Fundam. Methods Mach. Deep Learn. Algorithms, Tools, Appl., no. February 2022, pp. 1–16, 2022, doi: 10.1002/9781119821908.ch1.

[35] T. Chauhan, S. Rawat, S. Malik, and P. Singh, “Supervised and Unsupervised Machine Learning based Review on Diabetes Care,” 2021 7th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2021, pp. 581–585, 2021, doi: 10.1109/ICACCS51430.2021.9442021.

[36] A. E. Mohamed, “Comparative Study of Four Supervised Machine Learning Techniques for Classification,” Int. J. Appl. Sci. Technol., vol. 7, no. 2, pp. 5–18, 2017, [Online]. Available:

www.ijastnet.com.

[37] A. Haldorai and U. Kandaswamy, “Supervised machine learning techniques in intelligent network handovers,” EAI/Springer Innov. Commun. Comput., pp. 135–154, 2019, doi: 10.1007/978-3-030-15416-5_7.

[38] D. Maulud and A. M. Abdulazeez, “A Review on Linear Regression Comprehensive in Machine Learning,” J. Appl. Sci. Technol. Trends, vol. 1, no. 4, pp. 140–147, Dec. 2020, doi: 10.38094/jastt1457.

[39] F. Fabris, J. P. de Magalhães, and A. A. Freitas, “A review of supervised machine learning applied to ageing research,” Biogerontology, vol. 18, no. 2, pp. 171–188, 2017, doi: 10.1007/s10522-017-9683-y.

[40] Y. Bao et al., “Using Machine Learning and Natural Language Processing to Review and Classify the Medical Literature on Cancer Susceptibility Genes,” JCO Clin. Cancer Informatics, no. 3, pp. 1–9, 2019, doi: 10.1200/cci.19.00042.

[41] V. Nasteski, “An overview of the supervised machine learning methods,” Horizons.B, vol. 4, no. December 2017, pp. 51–62, 2017, doi: 10.20544/horizons.b.04.1.17.p05.

[42] N. Burkart and M. F. Huber, “A survey on the explainability of supervised machine learning,” J. Artif. Intell. Res., vol. 70, pp. 245–317, 2021, doi: 10.1613/JAIR.1.12228.

[43] A. Chowdhury, J. Rosenthal, J. Waring, and R. Umeton, “Applying self-supervised learning to medicine: Review of the state of the art and medical implementations,” Informatics, vol. 8, no. 3, pp. 1–29, 2021, doi: 10.3390/informatics8030059.

[44] T. Jiang, J. L. Gradus, and A. J. Rosellini, “Supervised Machine Learning: A Brief Primer,” Behav. Ther., vol. 51, no. 5, pp. 675–687, 2020, doi: 10.1016/j.beth.2020.05.002.

[45] M. Usama et al., “Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges,” IEEE Access, vol. 7, pp. 65579–65615, 2019, doi: 10.1109/ACCESS.2019.2916648.

[46] R. Saravanan and P. Sujatha, “Algorithms : A perspective of supervised learning approaches in data classification,” 2018 Second Int. Conf. Intell. Comput. Control Syst., no. Iciccs, pp. 945–949, 2018, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8663155.

[47] A. I. Károly, R. Fullér, and P. Galambos, “Unsupervised clustering for deep learning: A tutorial survey,” Acta Polytech. Hungarica, vol. 15, no. 8, pp. 29–53, 2018, doi: 10.12700/APH.15.8.2018.8.2.

[48] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, “A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science,” 2020, pp. 3–21.

[49] M. Khanum, T. Mahboob, W. Imtiaz, H. Abdul Ghafoor, and R. Sehar, “A Survey on Unsupervised Machine Learning Algorithms for Automation, Classification and Maintenance,” Int. J. Comput. Appl., vol. 119, no. 13, pp. 34–39, 2015, doi: 10.5120/21131-4058.

[50] C. M. Parlett-Pelleriti, E. Stevens, D. Dixon, and E. J. Linstead, “Applications of Unsupervised Machine Learning in Autism Spectrum Disorder Research: a Review,” Rev. J. Autism Dev. Disord., no. 0123456789, 2022, doi: 10.1007/s40489-021-00299-y.

[51] B. M. K. P, S. P. R, N. R K, and A. K, “Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier,” Int. J. Cogn. Comput. Eng., vol. 1, no. July, pp. 55–61, 2020, doi: 10.1016/j.ijcce.2020.10.002.

[52] S. Y. Rhee, J. M. Sung, S. Kim, I. J. Cho, S. E. Lee, and H. J. Chang, “Development and validation of a deep learning based diabetes prediction system using a nationwide population-based cohort,” Diabetes Metab. J., vol. 45, no. 4, pp. 515–525, 2021, doi: 10.4093/DMJ.2020.0081.

[53] L. N. Sanchez-Pinto, Y. Luo, and M. M. Churpek, “Big Data and Data Science in Critical Care,” Chest, vol. 154, no. 5, pp. 1239–1248, 2018, doi: 10.1016/j.chest.2018.04.037.

[54] S. Nosratabadi et al., “Data science in economics: Comprehensive review of advanced machine learning and deep learning methods,” Mathematics, vol. 8, no. 10, pp. 1–25, 2020, doi: 10.3390/math8101799.

[55] D. Shetty, K. Rit, S. Shaikh, and N. Patil, “Diabetes disease prediction using data mining,” Proc. 2017 Int. Conf. Innov. Information, Embed. Commun. Syst. ICIIECS 2017, vol. 2018-Janua, pp. 1–5, 2018, doi: 10.1109/ICIIECS.2017.8276012.

[56] R. Syed, R. K. Gupta, and N. Pathik, “An Advance Tree Adaptive Data Classification for the Diabetes Disease Prediction,” 2018 Int. Conf. Recent Innov. Electr. Electron. Commun. Eng. ICRIEECE 2018, pp. 1793–1798, 2018, doi: 10.1109/ICRIEECE44171.2018.9009180.

[57] S. S. Reddy, N. Sethi, and R. Rajender, “Safe Prediction of Diabetes Mellitus Using Weighted Conglomeration of Mining Schemes,” Proc. 4th Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2020, no. Dm, pp. 1213–1220, 2020, doi: 10.1109/ICECA49313.2020.9297390.

[58] J. R. Saura, “Using Data Sciences in Digital Marketing: Framework, methods, and performance metrics,” J. Innov. Knowl., vol. 6, no. 2, pp. 92–102, 2021, doi: 10.1016/j.jik.2020.08.001.

[59] M. A. Rahman Khan, M. Rahman, J. Us Salehin, M. S. Islam, and M. F. Rabbi, “Efficient Data Mining Techniques for Heart Disease Prediction and Comparative Analysis of Classification Algorithms,” Asian J. Res. Comput. Sci., vol. 12, no. 2, pp. 57–68, 2021, doi: 10.9734/ajrcos/2021/v12i230281.

[60] H. D. Prasetyo, P. A. Hogantara, and I. N. Isnainiyah, “A Web-Based Diabetes Prediction Application Using XGBoost Algorithm,” J. Comput. Appl. Informatics, vol. 5, no. 2, p. 59, 2021.

[61] L. Da Chen, T. Sakaguchi, and M. N. Frolick, “Data mining methods, applications, and tools,” Inf. Syst. Manag., vol. 17, no. 1, pp. 65–70, 2000, doi: 10.1201/1078/43190.17.1.20000101/31216.9.

[62] M. Rousset, Advances in Knowledge Discovery and Data Mining, vol. 40, no. 1. 1998.

[63] S. Agarwal, “Data Mining: Data Mining Concepts and Techniques,” in 2013 International Conference on Machine Intelligence and Research Advancement, Dec. 2013, pp. 203–207, doi: 10.1109/ICMIRA.2013.45.

[64] N. M. Ball, IN ASTRONOMY, vol. 19, no. 7. 2010.

[65] A. Rotondo and F. Quilligan, “Evolution Paths for Knowledge Discovery and Data Mining Process Models,” SN Comput. Sci., vol. 1, no. 2, pp. 1–19, 2020, doi: 10.1007/s42979-020-0117-6.

[66] B. N. Lakshmi and G. H. Raghunandhan, “A conceptual overview of data mining,” in 2011 National Conference on Innovations in Emerging Technology, Feb. 2011, pp. 27–32, doi: 10.1109/NCOIET.2011.5738828.

[67] J. J. Xu, “Knowledge discovery and data mining,” Comput. Handbook, Third Ed. Inf. Syst. Inf. Technol., no. February, pp. 19-1-19–22, 2014, doi: 10.1201/b16768.

[68] D. L. . Delen and D. Olson, “Advanced data mining techniques,” Choice Rev. Online, vol. 45, no. 12, pp. 45-6838-45–6838, Aug. 2008, doi: 10.5860/CHOICE.45-6838.

[69] C. C. Aggarwal, Data Mining. Cham: Springer International Publishing, 2015.

[70] W. B. Mulatu, M. F. Bedasa, and G. K. Terefa, “Prediction of Wheat Rust Diseases Using Data Mining Application,” OALib, vol. 07, no. 09, pp. 1–27, 2020, doi: 10.4236/oalib.1106717.

[71] M. O. S. Escobar, R. L. Espinosa, J. M. M. Espinosa, J. J. Noguez Monroy, and G. V. Solar, “Applying process mining to support management of predictive analytics/data mining projects in a decision making center,” 2019 6th Int. Conf. Syst. Informatics, ICSAI 2019, no. Icsai, pp. 1527–1533, 2019, doi: 10.1109/ICSAI48974.2019.9010135.

[72] P. C. Sen, M. Hajra, and M. Ghosh, Emerging Technology in Modelling and Graphics, vol. 937. 2020.

[73] F. G. Woldemichael and S. Menaria, “Prediction of Diabetes Using Data Mining Techniques,” Proc. 2nd Int. Conf. Trends Electron. Informatics, ICOEI 2018, no. Icoei, pp. 414–418, 2018, doi: 10.1109/ICOEI.2018.8553959.

[74] S. Srivatsan and T. Santhanam, “EARLY ONSET DETECTION OF DIABETES USING FEATURE SELECTION AND BOOSTING TECHNIQUES,” pp. 2474–2485, 2021, doi: 10.21917/ijsc.2021.0344.

[75] N. Kühl, R. Hirt, L. Baier, B. Schmitz, and G. Satzger, “How to conduct rigorous supervised machine learning in information systems research: The supervised machine learning report card,” Commun. Assoc. Inf. Syst., vol. 48, no. December, pp. 589–615, 2021, doi: 10.17705/1CAIS.04845.

[76] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research,” Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, 2017, doi: 10.1016/j.csbj.2016.12.005.

[77] A. Dogan and D. Birant, “Machine learning and data mining in manufacturing,” Expert Syst. Appl., vol. 166, no. February 2019, p. 114060, 2021, doi: 10.1016/j.eswa.2020.114060.

[78] S. Reddy, J. Fox, and M. P. Purohit, “Artificial intelligence-enabled healthcare delivery,” J. R. Soc. Med., vol. 112, no. 1, pp. 22–28, 2019, doi: 10.1177/0141076818815510.

[79] S. Castagno and M. Khalifa, “Perceptions of Artificial Intelligence Among Healthcare Staff: A Qualitative Survey Study,” Front. Artif. Intell., vol. 3, no. October, pp. 1–7, 2020, doi: 10.3389/frai.2020.578983.

[80] M. C. Laï, M. Brian, and M. F. Mamzer, “Perceptions of artificial intelligence in healthcare: Findings from a qualitative survey study among actors in France,” J. Transl. Med., vol. 18, no. 1, pp. 1–13, 2020, doi: 10.1186/s12967-019-02204-y.

[81] S. Secinaro, D. Calandra, A. Secinaro, V. Muthurangu, and P. Biancone, “The role of artificial intelligence in healthcare: a structured literature review,” BMC Med. Inform. Decis. Mak., vol. 21, no. 1, pp. 1–23, 2021, doi: 10.1186/s12911-021-01488-9.

[82] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, “Explainability for artificial intelligence in healthcare: a multidisciplinary perspective,” BMC Med. Inform. Decis. Mak., vol. 20, no. 1, pp. 1–9, 2020, doi: 10.1186/s12911-020-01332-6.

[83] M. van der Schaar et al., “How artificial intelligence and machine learning can help healthcare systems respond to COVID-19,” Mach. Learn., vol. 110, no. 1, pp. 1–14, 2021, doi: 10.1007/s10994-020-05928-x.

[84] O. Niakšu, “CRISP Data Mining Methodology Extension for Medical Domain,” Balt. J. Mod. Comput., vol. 3, no. 2, pp. 92–109, 2015.

[85] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, “Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications,” J. Healthc. Eng., vol. 2021, 2021, doi: 10.1155/2021/9930985.

[86] M. M. Bukhari, B. F. Alkhamees, S. Hussain, A. Gumaei, A. Assiri, and S. S. Ullah, “An Improved Artificial Neural Network Model for Effective Diabetes Prediction,” Complexity, vol. 2021, 2021, doi: 10.1155/2021/5525271.

[87] J. Ramesh, R. Aburukba, and A. Sagahyroon, “A remote healthcare monitoring framework for diabetes prediction using machine learning,” Healthc. Technol. Lett., vol. 8, no. 3, pp. 45–57, 2021, doi: 10.1049/htl2.12010.

[88] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, “Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM),” Diagnostics, vol. 11, no. 9, p. 1714, 2021, doi: 10.3390/diagnostics11091714.

[89] R. Patil, S. Tamane, S. A. Rawandale, and K. Patil, “A modified mayfly-SVM approach for early detection of type 2 diabetes mellitus,” Int. J. Electr. Comput. Eng., vol. 12, no. 1, pp. 524–533, 2022, doi: 10.11591/ijece.v12i1.pp524-533.

[90] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, “XAI—Explainable artificial intelligence,” Sci. Robot., vol. 4, no. 37, Dec. 2019, doi: 10.1126/scirobotics.aay7120.

[91] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, “Explainable artificial intelligence: an analytical review,” WIREs Data Min. Knowl. Discov., vol. 11, no. 5, pp. 1–13, Sep. 2021, doi: 10.1002/widm.1424.

[92] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, “Application of

explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," Comput. Methods Programs Biomed., vol. 226, p. 107161, Nov. 2022, doi: 10.1016/j.cmpb.2022.107161.

[93] A. Das and P. Rad, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey," pp. 1–24, 2020, [Online]. Available: http://arxiv.org/abs/2006.11371.

[94] K. Vidhya and R. Shanmugalakshmi, "Deep learning based big medical data analytic model for diabetes complication prediction," J. Ambient Intell. Humaniz. Comput., vol. 11, no. 11, pp. 5691–5702, Nov. 2020, doi: 10.1007/s12652-020-01930-2.

[95] B. Ljubic et al., "Predicting complications of diabetes mellitus using advanced machine learning algorithms," J. Am. Med. Informatics Assoc., vol. 27, no. 9, pp. 1343–1351, 2020, doi: 10.1093/jamia/ocaa120.

[96] O. Ghafki, L. Elachaak, F. Elouaai, and M. Bouhorma, "Deep learning approach as new tool for type 2 diabetes detection," ACM Int. Conf. Proceeding Ser., vol. Part F1481, pp. 3–6, 2019, doi: 10.1145/3320326.3320359.

[97] Z. Zhang, "Deep-Learning-Based Early Detection of Diabetic Retinopathy on Fundus Photography Using EfficientNet," ACM Int. Conf. Proceeding Ser., pp. 70–74, 2020, doi: 10.1145/3390557.3394303.

[98] M. Esteva, W. Xu, N. Simone, A. Gupta, and M. Jah, "Modeling Data Curation to Scientific Inquiry: A Case Study for Multimodal Data Integration," in Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Aug. 2020, pp. 235–242, doi: 10.1145/3383583.3398539.

[99] K. Zhou, B. S. Kottoori, S. A. Munj, Z. Zhang, S. Draghici, and S. Arslanturk, "Integration of Multimodal Data from Disparate Sources for Identifying Disease Subtypes," Biology (Basel)., vol. 11, no. 3, p. 360, Feb. 2022, doi: 10.3390/biology11030360.

[100] Q. Cai, H. Wang, Z. Li, and X. Liu, "A Survey on Multimodal Data-Driven Smart Healthcare Systems: Approaches and Applications," IEEE Access, vol. 7, pp. 133583–133599, 2019, doi: 10.1109/ACCESS.2019.2941419.

[101] U. Akhtar, J. W. Lee, H. S. Muhammad Bilal, T. Ali, W. A. Khan, and S. Lee, "The Impact of Big Data In Healthcare Analytics," in 2020 International Conference on Information Networking (ICOIN), Jan. 2020, vol. 2020-Janua, pp. 61–63, doi: 10.1109/ICOIN48656.2020.9016588.

[102] D. C. Nguyen et al., Federated Learning for Smart Healthcare: A Survey, vol. 55, no. 3. 2022.

[103] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated Learning for Healthcare Informatics," J. Healthc. Informatics Res., vol. 5, no. 1, pp. 1–19, Mar. 2021, doi: 10.1007/s41666-020-00082-4.

[104] R. S. Antunes, C. André da Costa, A. Küderle, I. A. Yari, and B. Eskofier, "Federated Learning for Healthcare: Systematic Review and Architecture Proposal," ACM Trans. Intell. Syst. Technol., vol. 13, no. 4, pp. 1–23, Aug. 2022, doi: 10.1145/3501813.

[105] N. Rieke et al., "The future of digital health with federated learning," npj Digit. Med., vol. 3, no. 1, p. 119, Sep. 2020, doi: 10.1038/s41746-020-00323-1.

[106] J. C. Pickup, M. Ford Holloway, and K. Samsi, "Real-Time Continuous Glucose Monitoring in Type 1 Diabetes: A Qualitative Framework Analysis of Patient Narratives," Diabetes Care, vol. 38, no. 4, pp. 544–550, Apr. 2015, doi: 10.2337/dc14-1855.

[107] M. Skevofilakas et al., "A Communication and Information Technology Infrastructure for Real Time Monitoring and Management of Type 1 Diabetes Patients," in 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Aug. 2007, vol. 2, pp. 3685–3688, doi: 10.1109/IEMBS.2007.4353131.

[108] G. Alfian, M. Syafrudin, M. Ijaz, M. Syaekhoni, N. Fitriyani, and J. Rhee, "A Personalized Healthcare Monitoring System for Diabetic Patients by Utilizing BLE-Based Sensors and Real-Time Data Processing," Sensors, vol. 18, no. 7, p. 2183, Jul. 2018, doi: 10.3390/s18072183.

[109] Y. Wu et al., “Novel binary logistic regression model based on feature transformation of XGBoost for type 2 Diabetes Mellitus prediction in healthcare systems,” Futur. Gener. Comput. Syst., vol. 129, pp. 1–12, 2022, doi: 10.1016/j.future.2021.11.003.

[110] S. Sivaranjani, S. Ananya, J. Aravinth, and R. Karthika, “Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction,” 2021 7th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2021, pp. 141–146, 2021, doi: 10.1109/ICACCS51430.2021.9441935.

[111] H. Kaur and V. Kumari, “Predictive modelling and analytics for diabetes using a machine learning approach,” Appl. Comput. Informatics, vol. 18, no. 1/2, pp. 90–100, Mar. 2022, doi: 10.1016/j.aci.2018.12.004.

[112] H. K. Al-Tammie and N. J. Al-Anber, “A Comparative Dimensionality Reduction Study in Diabetes Patients using LDA and PCA,” in 2022 3rd Information Technology To Enhance e-learning and Other Application (IT-ELA), Dec. 2022, pp. 170–175, doi: 10.1109/IT-ELA57378.2022.10107946.

[113] G. T. Reddy et al., “Analysis of Dimensionality Reduction Techniques on Big Data,” IEEE Access, vol. 8, pp. 54776–54788, 2020, doi: 10.1109/ACCESS.2020.2980942.

[114] J. Cruz, W. Mamani, C. Romero, and F. Pineda, “Selection of Characteristics by Hybrid Method: RFE, Ridge, Lasso, and Bayesian for the Power Forecast for a Photovoltaic System,” SN Comput. Sci., vol. 2, no. 3, pp. 1–14, 2021, doi: 10.1007/s42979-021-00584-x.

[115] M. Hamada, J. J. Tanimu, M. Hassan, H. A. Kakudi, and P. Robert, “Evaluation of Recursive Feature Elimination and LASSO Regularization-based optimized feature selection approaches for cervical cancer prediction,” Proc. - 2021 IEEE 14th Int. Symp. Embed. Multicore/Many-Core Syst. MCSoC 2021, pp. 333–339, 2021, doi: 10.1109/MCSoC51149.2021.00056.

[116] A. Al Mamun, W. Duan, and A. M. Mondal, “Pan-cancer Feature Selection and Classification Reveals Important Long Non-coding RNAs,” Proc. - 2020 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2020, pp. 2417–2424, 2020, doi: 10.1109/BIBM49941.2020.9313332.

[117] S. Mehta and K. S. Patnaik, “Improved prediction of software defects using ensemble machine learning techniques,” Neural Comput. Appl., vol. 33, no. 16, pp. 10551–10562, 2021, doi: 10.1007/s00521-021-05811-3.

[118] Alifah, T. Siswantining, D. Sarwinda, and A. Bustamam, “RFE and Chi-Square Based Feature Selection Approach for Detection of Diabetic Retinopathy,” in Proceedings of the International Joint Conference on Science and Engineering (IJCSE 2020), 2020, vol. 196, no. Ijcse, pp. 380–386, doi: 10.2991/aer.k.201124.069.

[119] J. Huo, C. Li, H. Wang, and H. Li, “LASSO Based Similarity Learning of Near-Infrared Spectra for Quality Control,” Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS, vol. 2020-Octob, pp. 424–427, 2020, doi: 10.1109/ICSESS49938.2020.9237682.

[120] R. Houari, A. Bounceur, M.-T. Kechadi, A.-K. Tari, and R. Euler, “Dimensionality reduction in data mining: A Copula approach,” Expert Syst. Appl., vol. 64, pp. 247–260, Dec. 2016, doi: 10.1016/j.eswa.2016.07.041.

[121] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O’Sullivan, “A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction,” Front. Bioinforma., vol. 2, no. June, pp. 1–17, Jun. 2022, doi: 10.3389/fbinf.2022.927312.

[122] C. C. Aggarwal, Data Classification. Chapman and Hall/CRC, 2014.

[123] Y. Bouchlaghem, Y. Akhiat, and S. Amjad, “Feature Selection: A Review and Comparative Study,” E3S Web Conf., vol. 351, p. 01046, May 2022, doi: 10.1051/e3sconf/202235101046.

[124] A. Jovic, K. Brkic, and N. Bogunovic, “A review of feature selection methods with applications,” in 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), May 2015, pp. 1200–1205, doi: 10.1109/MIPRO.2015.7160458.

[125] J. S, B. N, S. P, S. K. K, and V. Mani Nageshwar, “Diabetes Prediction Using Machine Learning Algorithms,” in 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), Mar. 2022, no. 05, pp. 46–51, doi: 10.1109/ICACCS54159.2022.9785073.

[126] M. Rady, K. Moussa, M. Mostafa, A. Elbasry, Z. Ezzat, and W. Medhat, “Diabetes Prediction Using Machine Learning: A Comparative Study,” in 2021 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES), Oct. 2021, pp. 279–282, doi: 10.1109/NILES53778.2021.9600091.

[127] S. BUYRUKOĞLU and A. AKBAŞ, “Machine Learning based Early Prediction of Type 2 Diabetes: A New Hybrid Feature Selection Approach using Correlation Matrix with Heatmap and SFS,” Balk. J. Electr. Comput. Eng., vol. 10, no. 2, pp. 110–117, Apr. 2022, doi: 10.17694/bajece.973129.

[128] P. K. Singh, W. Pawłowski, S. Tanwar, N. Kumar, and J. J. P. C. Rodrigues, Lecture Notes in Networks and Systems 121 Proceedings of First International Conference on Computing , Communications , and Cyber-Security, no. Ic4s. 2019.

[129] A. R and N. R, “Diabetes Mellitus Prediction and Severity Level Estimation Using OWDANN Algorithm,” Comput. Intell. Neurosci., vol. 2021, p. 5573179, 2021, doi: 10.1155/2021/5573179.

[130] V. Shorewala, “Early detection of coronary heart disease using ensemble techniques,” Informatics Med. Unlocked, vol. 26, no. June, p. 100655, 2021, doi: 10.1016/j.imu.2021.100655.

[131] Y. Joo, S. Lee, H. Kim, P. Kim, S. Hwang, and C. Choi, “Efficient healthcare service based on Stacking Ensemble,” ACM Int. Conf. Proceeding Ser., pp. 1–5, 2020, doi: 10.1145/3440943.3444727.

[132] Y. Joo, S. Lee, H. Kim, P. Kim, S. Hwang, and C. Choi, “Efficient healthcare service based on Stacking Ensemble,” in Proceedings of the 2020 ACM International Conference on Intelligent Computing and its Emerging Applications, Dec. 2020, vol. 27, no. 2, pp. 1–5, doi: 10.1145/3440943.3444727.

[133] A. Erekat, G. Servis, S. C. Madathil, and M. T. Khasawneh, “Efficient operating room planning using an ensemble learning approach to predict surgery cancellations,” IISE Trans. Healthc. Syst. Eng., vol. 10, no. 1, pp. 18–32, Jan. 2020, doi: 10.1080/24725579.2019.1641576.

[134] A. V. Kelarev, A. Stranieri, J. L. Yearwood, and H. F. Jelinek, “Empirical Study of Decision Trees and Ensemble Classifiers for Monitoring of Diabetes Patients in Pervasive Healthcare,” in 2012 15th International Conference on Network-Based Information Systems, Sep. 2012, pp. 441–446, doi: 10.1109/NBiS.2012.20.

[135] S. N. Mohanty, G. Nalinipriya, O. P. Jena, and A. Sarkar, “Ensemble Learning Method for Enhancing Healthcare Classification,” in Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering, 2020, no. March 2020, pp. 1–389, doi: 10.18178/wcse.2020.02.024.

[136] P. S. Mung and S. Phyu, “Ensemble learning method for enhancing healthcare classification,” WCSE 2020 2020 10th Int. Work. Comput. Sci. Eng., pp. 652–656, 2020, doi: 10.18178/wcse.2020.02.024.

[137] K. Yu and X. Xie, “Predicting Hospital Readmission: A Joint Ensemble-Learning Model,” IEEE J. Biomed. Heal. Informatics, vol. 24, no. 2, pp. 447–456, Feb. 2020, doi: 10.1109/JBHI.2019.2938995.

[138] J. J. Khanam and S. Y. Foo, “A comparison of machine learning algorithms for diabetes prediction,” ICT Express, vol. 7, no. 4, pp. 432–439, 2021, doi: 10.1016/j.icte.2021.02.004.

[139] M. Kumar et al., “Population-centric risk prediction modeling for gestational diabetes mellitus: A machine learning approach,” Diabetes Res. Clin. Pract., vol. 185, no. February, p. 109237, Mar. 2022, doi: 10.1016/j.diabres.2022.109237.

[140] S. M. Lee et al., “Prediction of gestational diabetes in the first trimester using machine learning-based methods,” Am. J. Obstet. Gynecol., vol. 224, no. 2, pp. S252–S253, Feb. 2021, doi: 10.1016/j.ajog.2020.12.412.

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

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

[143] B. P. Nguyen et al., “Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records,” Comput. Methods Programs Biomed., vol. 182, p. 105055, Dec. 2019, doi: 10.1016/j.cmpb.2019.105055.

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

[145] M. F. Faruque, Asaduzzaman, and I. H. Sarker, “Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus,” 2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE 2019, pp. 7–9, 2019, doi: 10.1109/ECACE.2019.8679365.

[146] M. Gollapalli et al., “A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM,” Comput. Biol. Med., vol. 147, no. June, p. 105757, 2022, doi: 10.1016/j.compbiomed.2022.105757.

[147] Z. Mushtaq, M. F. Ramzan, S. Ali, S. Baseer, A. Samad, and M. Husnain, “Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques,” Mob. Inf. Syst., vol. 2022, 2022, doi: 10.1155/2022/6521532.

[148] R. Krishnamoorthi et al., “A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques,” J. Healthc. Eng., vol. 2022, 2022, doi: 10.1155/2022/1684017.

[149] M. U. Emon, M. S. Keya, M. S. Kaiser, M. A. Islam, T. Tanha, and M. S. Zulfiker, “Primary Stage of Diabetes Prediction using Machine Learning Approaches,” in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Mar. 2021, pp. 364–367, doi: 10.1109/ICAIS50930.2021.9395968.

[150] N. K. Kumar, D. Vigneswari, R. J. Reynold, J. Josy, and J. C. Prince, “An Ensemble Deep Learning Method for Diabetes Mellitus,” vol. 1, 2021, pp. 75–83.

[151] R. Srivastava and R. K. Dwivedi, “Diabetes Mellitus Prediction Using Ensemble Learning Approach with Hyperparameterization,” in Lecture Notes in Networks and Systems, vol. 154, 2022, pp. 487–494.

[152] P. Das and S. Nanda, “An Improved Ridge Regression-Based Extreme Learning Machine for the Prediction of Diabetes,” 2021, pp. 541–547.

[153] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, “Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers,” IEEE Access, vol. 8, pp. 76516–76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[154] M. Pokharel et al., “Deep learning for predicting the onset of type 2 diabetes: enhanced ensemble classifier using modified t-SNE,” Multimed. Tools Appl., vol. 81, no. 19, pp. 27837–27852, Aug. 2022, doi: 10.1007/s11042-022-12950-9.

[155] A. Doğru, S. Buyrukoğlu, and M. Arı, “A hybrid super ensemble learning model for the early-stage prediction of diabetes risk,” Med. Biol. Eng. Comput., vol. 61, no. 3, pp. 785–797, Mar. 2023, doi: 10.1007/s11517-022-02749-z.

[156] Z. Alhassan, A. S. McGough, R. Alshammari, T. Daghstani, D. Budgen, and N. Al Moubayed, “Type-2 Diabetes Mellitus Diagnosis from Time Series Clinical Data Using Deep Learning Models,” in Icann 2018, 2018, pp. 468–478.

[157] S. Abbasi-Sureshjani, B. Dashtbozorg, B. M. ter Haar Romeny, and F. Fleuret, “Exploratory Study on Direct Prediction of Diabetes Using Deep Residual Networks,” 2018, pp. 797–802.

[158] B. Kurt et al., “Prediction of gestational diabetes using deep learning and Bayesian optimization and traditional machine learning techniques,” Med. Biol. Eng. Comput., no. 0123456789, 2023, doi: 10.1007/s11517-023-02800-7.

[159] M. Shrestha et al., “A novel solution of deep learning for enhanced support vector machine for predicting the onset of type 2 diabetes,” Multimed. Tools Appl., vol. 82, no. 4, pp. 6221–6241, 2023, doi: 10.1007/s11042-022-13582-9.

[160] S. A. Alex, J. J. V. Nayahi, H. Shine, and V. Gopirekha, “Deep convolutional neural network for

diabetes mellitus prediction," Neural Comput. Appl., vol. 34, no. 2, pp. 1319–1327, 2022, doi: 10.1007/s00521-021-06431-7.

[161] S. Samet, M. R. Laouar, and I. Bendib, "Diabetes mellitus early stage risk prediction using machine learning algorithms," in 2021 International Conference on Networking and Advanced Systems (ICNAS), Oct. 2021, pp. 1–6, doi: 10.1109/ICNAS53565.2021.9628955.

[162] R. Ramezani, M. Maadi, and S. M. Khatami, "A novel hybrid intelligent system with missing value imputation for diabetes diagnosis," Alexandria Eng. J., vol. 57, no. 3, pp. 1883–1891, 2018, doi: 10.1016/j.aej.2017.03.043.

[163] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," Appl. Comput. Informatics, 2019, doi: 10.1016/j.aci.2018.12.004.

[164] N. Nnamoko and I. Korkontzelos, "Efficient treatment of outliers and class imbalance for diabetes prediction," Artif. Intell. Med., vol. 104, no. December 2018, p. 101815, 2020, doi: 10.1016/j.artmed.2020.101815.

[165] S. Barik, S. Mohanty, S. Mohanty, and D. Singh, Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques, vol. 153. Springer Singapore, 2021.

[166] S. Samet, M. R. Laouar, I. Bendib, and S. Eom, "Analysis and Prediction of Diabetes Disease Using Machine Learning Methods," Int. J. Decis. Support Syst. Technol., vol. 14, no. 1, pp. 1–19, Jan. 2022, doi: 10.4018/IJDSST.303943.

[167] M. Aminul and N. Jahan, "Prediction of Onset Diabetes using Machine Learning Techniques," Int. J. Comput. Appl., vol. 180, no. 5, pp. 7–11, 2017, doi: 10.5120/ijca2017916020.

[168] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," Front. Genet., vol. 9, no. November, pp. 1–10, 2018, doi: 10.3389/fgene.2018.00515.

[169] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," BMC Endocr. Disord., vol. 19, no. 1, pp. 1–9, 2019, doi: 10.1186/s12902-019-0436-6.

[170] A. H. Syed and T. Khan, "Machine learning-based application for predicting risk of type 2 diabetes mellitus (t2dm) in saudi arabia: A retrospective cross-sectional study," IEEE Access, vol. 8, pp. 199539–199561, 2020, doi: 10.1109/ACCESS.2020.3035026.

[171] T. Mahboob Alam et al., "A model for early prediction of diabetes," Informatics Med. Unlocked, vol. 16, no. July, p. 100204, 2019, doi: 10.1016/j.imu.2019.100204.

[172] R. Arora, G. Kaur, and P. Gulati, Feature Selection and Hyperparameter Tuning in Diabetes Mellitus Prediction. Springer Singapore, 2021.

[173] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, no. xxxx, 2021, doi: 10.1016/j.icte.2021.02.004.

[174] N. H. Taz, A. Islam, and I. Mahmud, "A Comparative Analysis of Ensemble Based Machine Learning Techniques for Diabetes Identification," pp. 1–6, 2021, doi: 10.1109/icrest51555.2021.9331036.

[175] M. U. Emon, M. S. Keya, M. S. Kaiser, T. Tanha, M. S. Zulfiker, and others, "Primary Stage of Diabetes Prediction using Machine Learning Approaches," in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 364–367, doi: 10.1109/ICAIS50930.2021.9395968.

[176] S. Samet and M. R. Laouar, "Enhancing early-stage risk prediction of diabetes mellitus through feature selection bagging and boosting techniques," in National Conference on Artificial Intelligence: From Theory to Practice (NCAI'2023), 2023, p. 6.

[177] F. Nusrat, B. Uzbas, and Ö. K. Baykan, "Gradient Boosting Classification kullanarak Diabetes Mellitus Tahmini," Eur. J. Sci. Technol., no. September, pp. 268–272, 2020, doi: 10.31590/ejosat.803504.

[178] S. Ghane, N. Bhorade, N. Chitre, B. Poyekar, R. Mote, and P. Topale, "Diabetes Prediction using

Feature Extraction and Machine Learning Models," pp. 1652–1657, 2021, doi: 10.1109/icesc51422.2021.9532818.

[179] M. S. Tahsin, M. Jobayer, M. B. U. Antor, M. Islam, F. F. Raisa, and M. A. H. Shaikat, "Predictive Analysis Brief Study of Early-Stage Diabetes Using Multiple Classifier Models," 2022 IEEE 12th Annu. Comput. Commun. Work. Conf. CCWC 2022, pp. 203–207, 2022, doi: 10.1109/CCWC54503.2022.9720736.

[180] R. P. Alluri, "Diabetes Prediction Using Ensemble Techniques," vol. 16, no. 5, pp. 410–415, 2021.

[181] D. V. V. Rani, D. G. Vasavi, and D. K. R. . K. Kumar, "Significance Of Multilayer Perceptron Model For Early Detection Of Diabetes Over Ml Methods," J. Univ. Shanghai Sci. Technol., vol. 23, no. 08, pp. 148–160, 2021, doi: 10.51201/jusst/21/08358.

[182] M. Banchhor and P. Singh, "Comparative study of ensemble learning algorithms on early stage diabetes risk prediction," 2021 2nd Int. Conf. Emerg. Technol. INCET 2021, pp. 1–6, 2021, doi: 10.1109/INCET51464.2021.9456263.

[183] B. Kumar Sahu and N. Ghosh, Early Stage Prediction of Diabetes Using Machine Learning Techniques, vol. 302. Springer Singapore, 2022.

[184] L. Akter and A.-I. Ferdib, "Diabetes Mellitus Prediction and Feature Importance Score Finding Using Extreme Gradient Boosting," in Lecture Notes in Networks and Systems, vol. 322, Springer International Publishing, 2022, pp. 643–654.

[185] S. Samet and R. M. Laouar, "Building Risk Prediction Models for Diabetes Decision Support System," in International Conference on Decision Support System Technology, 2023, vol. 30, no. 4, pp. 171–181, doi: 10.1007/978-3-031-32534-2_13.

[186] E. Adua et al., "Predictive model and feature importance for early detection of type II diabetes mellitus," Transl. Med. Commun., vol. 6, no. 1, pp. 1–15, 2021, doi: 10.1186/s41231-021-00096-z.

[187] S. Padhy, S. Dash, S. Routray, S. Ahmad, J. Nazeer, and A. Alam, "IoT-Based Hybrid Ensemble Machine Learning Model for Efficient Diabetes Mellitus Prediction," Comput. Intell. Neurosci., vol. 2022, no. iii, pp. 1–11, 2022, doi: 10.1155/2022/2389636.

[188] A. Al-Zebari and A. Sengur, "Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection," 1st Int. Informatics Softw. Eng. Conf. Innov. Technol. Digit. Transform. IISEC 2019 - Proc., pp. 2–5, 2019, doi: 10.1109/UBMYK48245.2019.8965542.

[189] S. Samet, M. R. Laouar, and I. Bendib, "Predicting and Staging Chronic Kidney Disease using Optimized Random Forest Algorithm," in 2021 International Conference on Information Systems and Advanced Technologies (ICISAT), Dec. 2021, pp. 1–8, doi: 10.1109/ICISAT54145.2021.9678441.