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

Université A/Mira de Béjaia
Faculté des Sciences Exactes
Département d'Informatique

MÉMOIRE DE MASTER RECHERCHE

En Informatique

Option

Intelligence Artificielle

Thème

Prédiction du diabète en utilisant les algorithmes de machine learning

Présenté par : Mlle. BOUAOUINA Sara

Mlle. IZOUAOUEN Akila

Soutenu le 18 Juin 2023 devant le jury composé de :

Présidente	**Mme S. ALOUI**	**Maître de conf. A**	**U. A/Mira Béjaia.**
Rapporteur	**Mr K. AMROUN**	**Professeur**	**U. A/Mira Béjaia.**
Co-Rapporteur	**Mr A. ZIANE**	**Doctorant**	**U. A/Mira Béjaia.**
Examinatrice	**Mme Z. TAHAKOURT**	**Maître de conf. B**	**U. A/Mira Béjaia.**
Examinatrice	**Mme D. BOULAHROUZ**	**Maître de conf. A**	**U. A/Mira Béjaia.**

Béjaia, Juin 2023.

⁕ *Remerciements* ⁕

Tout d'abord, nous souhaitons exprimer notre gratitude envers Dieu le Tout-Puissant pour son aide, en nous donnant la volonté, le courage et la patience nécessaires pour mener à bien ce modeste travail.

Nous tenons à remercier sincèrement notre promoteur, le Pr. AMROUN Kamal, d'avoir accepté de nous encadrer tout au long de ce projet.

Nos remerciements vont également à notre co-encadrant, M. ZIANE Amine, pour le temps précieux qu'il nous a accordé chaque fois que cela était nécessaire, pour ses remarques bienveillantes, son orientation et ses conseils précieux tout au long de notre travail.

Nous adressons notre reconnaissance aux membres du jury pour l'honneur qu'ils nous ont fait en évaluant ce travail.

Nous tenons à exprimer nos vifs remerciements au Professeur OUAIL Djamal Eddine ainsi que DR. TEBANI pour nous avoir accueillis au sein de leurs service de médecine interne pendant notre période de stage.

Nous souhaitons également remercier chaleureusement nos familles et nos amis(es) qui nous ont soutenus, encouragés et aidés tout au long de ces années.

Un grand merci à tous ceux qui ont contribué de près ou de loin à l'élaboration de ce modeste travail. Leur soutien et leur collaboration ont été précieux, et nous leur exprimons ici nos sincères reconnaissances.

⁕ *Dédicaces* ⁕

***J**e souhaiterais exprimer ma gratitude envers la vie elle-même, qui m'a donné l'opportunité de suivre cette voie et de vivre cette expérience enrichissante.*

***J**e dédie ce travail*

AUX deux chères personnes dans ma vie, pour leur sacrifice et leur amour éternel

***M**on cher papa et Ma chère maman.*

***A** mes chers frères 'Lyes' et 'Jugurtha', mes deux piliers qui me soutiennent face à tous les obstacles.*

***A** ma chère belle-soeur 'Zakia', ainsi que mes petites nièces 'Iline' et 'Milina', pour leurs encouragements constants et leur soutien moral.*

***A** mes grandes mères, mes oncles, mes tantes, mes cousins et toute la famille pour leur soutien inconditionnel.*

***A** ma chère copine et binôme 'Sarah', et à tous les jours de formation que nous avons vécus ensemble.*

***A** mes amis 'Dihia', 'Thafrara', 'Nina', 'Sarah', 'Ryma', 'Amira', 'Yanis', 'Kaci', 'Salim' et 'Massine' .*

***A** ma promotion intelligence artificielle 2023.*

***A** toutes personnes chères à moi.*

Akila Izouaouen

⁕ *Dédicaces* ⁕

***J**e dédie cet humble travail à toutes personnes ayant le succès comme objectif dans leur vie*

***A** la mémoire de mes chers grands parents, de mon oncle et tante qui sont dignes de ma gratitude et mon estime*

***A**ux prunelles de mes yeux à la personne qui rêvait de cette journée plus que moi*

***M**on cher Papa et ma chère Maman*

***M**erci pour toutes vos sacrifices, votre amour, votre tendresse, votre soutien et vos prières tout au long de mes études*

***A** mes chères soeurs Kenza, Zahra et Hassina*

***A** mon cher frère Amirouche et ma belle-soeur Sonia*

***A** mes soeurs du coeur mes « BIZBIZS » Ania et Liticia*

***A** mes petits neveux d'amours Ayoub, Raouf et Zakria*

***A** mes chers cousins Tahar, Youcef et Amel*

***A** ma chère copine et binôme Akila*

***A** tous mes amis*

***A** toute la promotion Master II Intelligence Artificielle*

***A** tous ceux qui me sont chères*

Sara Bouaouina

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

ML : Machine Learning (Apprentissage automatique)

DM : Diabetes Mellitus (Diabète sucré)

IMC Indice de masse corporelle

LDL : Low-Density Lipoprotein (Lipoprotéines de basse densité, mauvais cholésterol)

OMS : Organisation mondiale de la santé

CHU : Centre Hospitalier Universitaire

SVM : Support Vector Machines (Machines à vecteurs de support)

RF : Random Forest (Forêt aléatoire)

LR : Logistic Regression (Régression logistique)

KNN : K-Nearest Neighbors (k Plus proches voisins)

DT : Decision tree (arbre de decision)

GBM : Gradient Boosting Machines (Machines à gradient boosting)

XGBoost : eXtreme Gradient Boosting (Boosting de gradient extrême)

Introduction générale

Dernièrement, l'Intelligence Artificielle (IA) a connu un développement sans précédent. En tant que technologie de pointe, son objectif est de simuler l'intelligence humaine, en particulier la capacité de raisonnement, d'apprentissage et de prise de décision comme la prédiction.

Avec l'avènement des ordinateurs à haute performance et l'explosion des données numériques, l'IA ouvre la voie à de nombreuses opportunités passionnantes. Toutefois, il est important de rester vigilant et de s'assurer que son utilisation soit responsable et éthique afin de maximiser ses avantages pour l'humain.

Les avancées technologiques dans le domaine de l'Intelligence Artificielle, en particulier le machine learning, ont des répercussions dans plusieurs domaines, dont la médecine. Ces dernières ont ouvert de nouvelles perspectives pour la prédiction de maladies chroniques telles que le diabète.

0.1 Problématique

Au cours des dernières années, le diabète est devenu une maladie très répandue, touchant de plus en plus de personnes. Les coûts des soins liés à cette maladie sont élevés, ce qui nécessite une solution. Pour lutter contre son développement, il est important de prévenir la maladie en amont en mettant en place différents systèmes automatisés pour aider les médecins à prendre des décisions et à diagnostiquer les risques d'infection. Il est également crucial de se concentrer sur les méthodes et les techniques les plus efficaces en matière de prédiction.

0.2 Objectif

L'objectif de ce mémoire est de réaliser un système de prédiction de diabète performant en utilisant les algorithmes de machine learning, afin de réduire les risques de complications liés à cette maladie chronique sur la santé du patient.

En effet, nous nous intéressons dans ce travail à présenter une méthode fiable et efficace pour une détection précoce du diabète à partir de données médicales et de facteurs de risque

individuels. Cette méthode exploitera des techniques de machine learning pour créer un système prédictif robuste.

0.3 Organisation du mémoire

Ce travail est composé de quatre principaux chapitres comme suit :

Le premier chapitre aborde des généralités sur le diabète. On présentera la définition de la maladie, les différents types existants et leurs symptômes, les facteurs de risque et les complications qui en découlent ainsi que les causes et les traitements couramment utilisés pour chaque type.

Le deuxième chapitre donne un aperçu sur l'apprentissage automatique, ses types, ses algorithmes, ainsi qu'un état de l'art contenant un résumé des articles et une étude comparative entre les approches proposées et les travaux connexes.

Le troisième chapitre est dédié à la présentation de l'organisme d'accueil du CHU de Béjaia, ses différents services et l'objectif de notre stage au sein de cet établissement.

Le quatrième chapitre porte sur la description de notre approche, nous présentons aussi les outils de programmation et l'implémentassions de notre système et les résultats d'exécution, ainsi que les logiciels choisis.

Dans le dernier chapitre, nous fournissons une description détaillée de notre application web, en mettant l'accent à la fois sur ses différentes interfaces et sur les avantages qu'elle apporte à l'utilisateur.

En fin, ce travail se clôture par une conclusion générale qui récapitule les idées principales que nous avons développées ainsi que les perspectives pour de futures recherches.

Chapitre 1

Le diabète

1.1 Introduction

Le 14 novembre de chaque année, la communauté internationale célèbre la journée mondiale du diabète pour sensibiliser le public à la prévalence croissante de cette maladie et aux stratégies de prévention et de traitement.

Selon les statistiques mondiales, environ 463 millions de personnes vivent actuellement avec le diabète, un chiffre qui devrait atteindre les 700 millions d'ici 2045 si des mesures efficaces de prévention et de traitement ne sont pas mises en place. En Afrique en particulier, on dénombre actuellement 24 millions d'adultes atteints de diabète, avec une estimation de 55 millions d'ici 2045. Cette maladie non transmissible a causé 416 000 décès l'année dernière et devrait devenir l'une des principales causes de mortalité dans le monde.

Il est important de noter que le risque de décès prématuré associé au diabète continue d'augmenter plutôt que de diminuer. Par conséquent, une sensibilisation accrue à la maladie et une amélioration de l'accès aux soins de santé sont nécessaires pour prévenir et traiter efficacement le diabète. [18]

1.2 Définition du diabète sucré

Le diabète, souvent appelé "Diabète sucré", "Diabetes Mellitus DM" par les médecins, est un problème métabolique majeur se caractérisant par une augmentation anormale du taux de glucose dans le corps humain, également appelée hyperglycémie. Cette condition est causée par une absence ou une production insuffisante d'insuline par le pancréas.

Plus Clairement : Lorsque nous consommons des aliments, ces derniers sont dégradés en glucose (sucre), qui fournit de l'énergie au corps pour fonctionner correctement. Le glucose est transporté dans tout le corps par le sang pour alimenter les cellules. Cependant, pour que le glucose présent dans le sang puisse être absorbé par les cellules, le corps nécessite l'action de l'insuline, une hormone produite par le pancréas. L'insuline agit comme une clé en permettant au glucose de pénétrer dans les cellules de notre corps, voir la figure 1.1. [27]

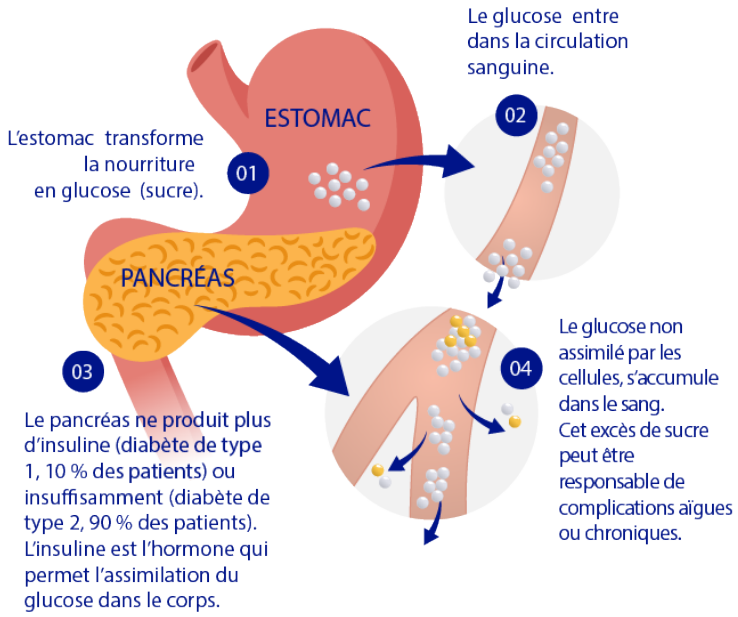


FIGURE 1.1 – Définition du diabète [12]

1.3 Historique et origine du diabète

Plus de 3500 ans en arrière, l'Égypte, la Chine et l'Inde ont fait les premières descriptions du diabète. À cette époque, le diabète était caractérisé par une soif excessive, une émission abondante d'urine et une perte de poids. Les anciens médecins avaient remarqué que l'urine des patients diabétiques avait un goût sucré. Le mot "diabète" a été utilisé pour la première fois en Grèce, il signifie "passer à travers" en référence à la polyurie (uriner excessivement) caractéristique de la maladie. [16]

Quelque progrès réalisé par la recherche sur la maladie :

1869 : découverte des îlots de Langerhans par l'étudiant allemand Paul Langerhans.

1889 : lien établi entre le pancréas et le diabète par les Allemands Oskar Minkowski et Josef Von Mering.

1921 : découverte de l'insuline par Frederick Grant Banting et Charles Best.

1922 : première injection d'insuline (extraite du pancréas de porc) sauvant un enfant de 14 ans.

1923 : prix Nobel décerné à Frederick Grant Banting pour cette grande avancée. Début de la production industrielle et commercialisation d'insuline par des laboratoires à partir du pancréas de boeuf et de porc.

1955 : Frederick Sanger décrit la structure chimique de l'insuline humaine.

1978-1982 : grâce aux progrès des technologies, l'insuline est produite par génie génétique.

A ce jour, les avancées technologiques ont stimulé la poursuite des recherches en vue de trouver des solutions plus efficaces pour soigner les personnes diabétiques.

1.4 Les facteurs de risque du diabète

Pour éviter l'apparition du diabète, il est nécessaire de connaître et de corriger certains facteurs de risque. Et parmi ces derniers on cite :

— Hérédité : si l'un de vos parents ou frères et soeurs ont le diabète, vous avez un risque plus élevé de développer la maladie.
— Surpoids ou obésité : un indice de masse corporelle (IMC) élevé est un facteur de risque important de diabète.
— Sédentarité : une activité physique insuffisante augmente le risque de diabète.
— Âge : Le risque de diabète augmente avec l'âge. Les personnes de plus de 45 ans sont plus susceptibles de développer la maladie.
— Hypertension artérielle : des niveaux élevés de pression artérielle augmentent le risque de diabète.
— Niveaux élevés de cholestérol : des niveaux élevés de cholestérol, en particulier de LDL (mauvais cholestérol), augmentent le risque de diabète.
— Syndrome des ovaires poly-kystiques (SOPK) : le SOPK est une affection courante chez les femmes qui peut augmenter le risque de diabète.
— Grossesse : les femmes enceintes qui ont développé un diabète gestationnel ont un risque accru de développer un diabète de type 2 plus tard dans leur vie.
— Tabagisme : fumer peut augmenter le risque de diabète. [28]

La figure 1.2 montre les différents facteurs de risque du diabète.

FIGURE 1.2 – Les facteurs de risque du diabète

1.5 Classification du diabète

1.5.1 Diabète type I

Le diabète de type 1, également connu sous le nom de diabète insulinodépendant ou diabète juvénile, représente environ 10% de tous les cas de diabète et affecte principalement les enfants et les adolescents, mais rarement les personnes âgées. Dans ce type, il y a une absence totale de production d'insuline, ce qui oblige les personnes atteintes à avoir des injections d'insuline quotidiennes ou à utiliser une pompe à insuline pour survivre [26] , Ils doivent aussi assurer une glycémie adéquate en effectuant des tests sanguins et suivre un régime spécial.[37]

1.5.1.1 Les symptômes du diabète type I

Les symptômes courant du diabète de type 1 comprennent :

— Polydipsie : soif accrue.
— Polyphagie : faim accrue.
— Polyurie : besoin fréquent d'uriner.
— Perte de poids.
— Manque d'énergie ou faiblesse.
— Trouble de vision.
— Sensation de picotement ou d'engourdissement dans les mains et les pieds. [20]

Ces symptômes peuvent apparaître rapidement, parfois en quelques semaines ou quelques mois, et nécessitent un diagnostic et un traitement immédiats.

1.5.1.2 Causes et traitements du diabète type I

Causes : Les causes de ce type ne sont pas clairement établies, mais les chercheurs proposent certains facteurs de risque comme :
— La prédisposition génétique
— La mauvaise alimentation
— Les facteurs environnementaux

Traitements : Les traitements du diabète type I sont :
— Multi injections.
— Pompe à insuline.
— Une bonne nutrition est importante.

La figure 1.3 représente des injections d'insuline.

FIGURE 1.3 – Injection d'insuline [17]

1.5.2 Diabète type II

Le diabète de type 2, également connu sous le nom de diabète non-insulinodépendant ou diabète de la maturité ou de l'adulte, représente environ 90% de tous les cas de diabète et affecte principalement les personnes adultes qui dépassent les quarantaines [26] . Cette anomalie résulte d'un dysfonctionnement des îlots pancréatiques, qui peut être causé par une perte dans la sécrétion ou l'utilisation de l'insuline. Cette perte est le résultat d'une combinaison de facteurs génétiques héritables et de facteurs environnementaux liés au mode de vie. [48]

1.5.2.1 Les symptômes du diabète type II

Les symptômes courant du diabète de type 2 comprennent :

— Soif intense et faim exagérée.
— La fatigue.
— Vision brouillée.
— Mictions fréquentes.
— Trouble du comportement ou maladies mentales.
— Cicatrisation lente.
— Mauvaise haleine.
— Mauvaise circulation sanguine.
— Problèmes de digestion.
— Sensation de picotement ou d'engourdissement dans les mains et les pieds. [20]

Les symptômes du diabète de type 2 peuvent se développer progressivement sur une longue période, parfois sans symptômes visibles. Certains des symptômes peuvent également être confondus avec le vieillissement ou le stress. Il est donc important de vérifier régulièrement le taux de sucre dans le sang par un médecin.

1.5.2.2 Causes et traitements du diabète type II

Causes : Les causes du type II type sont :
— La génétique et l'hérédité.
— Le surpoids et l'obésité.
— L'hypertension artérielle.
— Survenue d'un diabète sucré durant une grossesse.
— Taux élevé de cholestérol.
— Intolérance au glucose. [48]

Traitements : Les traitements de ce type sont :
— L'adoption d'une meilleure alimentation.
— Pratique régulière d'une activité physique.
— Prise de médicaments tels que des antidiabétiques oraux.

La figure 1.4 représente des comprimés antidiabétiques oraux.

FIGURE 1.4 – Comprimés antidiabétiques oraux [11]

1.5.3 Diabète gestationnel

Le diabète gestationnel, également connu sous le nom de diabète de grossesse, représente environ 3 à 20% des femmes enceintes, il survient généralement vers la fin du 6ème mois.

Le diabète gestationnel est causé par des changements hormonaux pendant la grossesse qui peuvent rendre l'organisme de la femme enceinte moins sensible à l'insuline, ce qui peut conduire à une augmentation des taux de glucose dans le sang. Ce type s'il n'est pas traiter, il peut représenter un danger sur elle ainsi que son bébé. Bien qu'il disparaisse souvent après l'accouchement, mais ces femmes sont exposées à un risque plus élevé de développer un diabète de type 2 plus tard dans leur vie. [16]

1.5.3.1 Les symptômes du diabète gestationnel

Le diabète gestationnel ne présente souvent aucun symptôme perceptible. Cependant, certains signes peuvent être observés, tels que :

— Soif accrue.
— Des urines plus abondantes et envie plus fréquente d'uriner.
— Maux de tête et fatigue.
— Prise de poids rapide et inexplicable
— Hypertension artérielle. [20]

C'est pourquoi il est nécessaire pour les femmes enceintes de dépister régulièrement le diabète gestationnel, même dans le cas ou aucun symptôme n'est présent.

1.5.3.2 Causes et traitements du diabète gestationnel

Causes : Les causes du diabète gestationnel sont :

— La résistance des cellules à l'action de l'insuline causée durant la grossesse par les hormones du placenta.
— Le surpoids ou l'obésité entre deux grossesses
— prédisposition génétique
— Grossesse multiple [15]

Traitements : Les traitements de ce dernier sont :

— Modifications de l'alimentation maternelle.
— Contrôle du poids et l'activité physique.
— Une bonne hygiène de vie.
— Insuline : Si les changements de régime alimentaire et l'exercice ne suffisent pas à maintenir le taux de glucose dans le sang dans une plage normale.

FIGURE 1.5 – Régime allimentaire pour diabète gestationnel [24]

1.6 Autres types de diabète

Mis à part le diabète de type 1, de type 2 et gestationnel, il existe d'autres types de diabète moins fréquents.

— Diabète secondaire à certaines maladies.
— Diabète secondaire à la prise des médicaments.
— Diabète de MODY (Maturity Onset Diabetes Of The Young).
— Diabète de LADA (Latent Autoimmune Diabetes In Adulte).
— Diabète néonatal. [15]

1.7 Complications du diabète

Lorsque le taux de sucre dans le sang n'est pas bien contrôlé, il peut entraîner des complications à court et à long terme.

— **Neuropathie diabétique :** Le diabète peut endommager les nerfs, ce qui peut engendrer une perte auditive, une perte de sensation ou une douleur dans les mains, les pieds et les jambes.
— **Rétinopathie diabétique :** Le diabète peut endommager les vaisseaux sanguins de la rétine, ce qui peut entraîner une perte de vision.
— **Néphropathie diabétique :** Le diabète peut endommager les reins, ce qui peut entraîner une insuffisance rénale.
— **Maladies cardiovasculaires :** Les personnes atteintes de diabète sont plus susceptibles de développer des maladies cardiovasculaires, telles que des maladies coronariennes, une hypertension artérielle et des accidents vasculaires cérébraux.
— **Infections cutanées :** Le diabète peut rendre la peau plus sujette aux infections, en particulier dans les pieds et les jambes.
— **Amputation :** Le diabète peut entraîner des lésions nerveuses et vasculaires dans les pieds et les jambes, augmentant ainsi le risque d'amputation. [30]

Voir la figure 1.6.

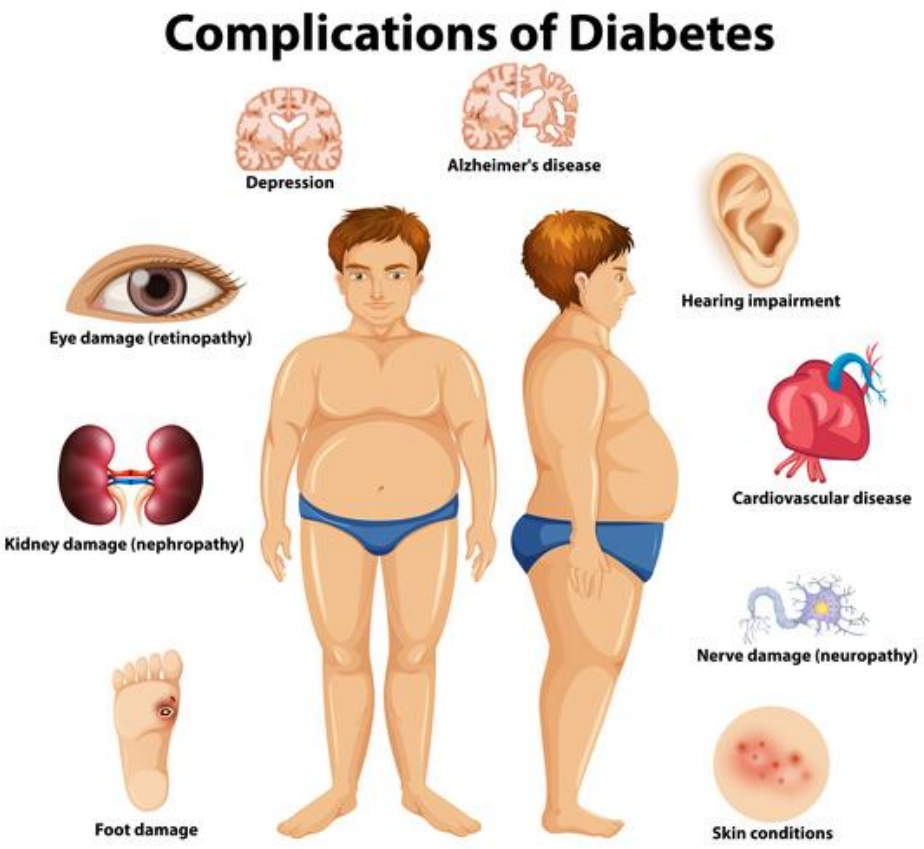


FIGURE 1.6 – Les complications du diabète [10]

1.8 Le prédiabète

Le prédiabète, également appelé intolérance au glucose, se caractérise par une élévation modérée du taux de sucre dans le sang qui ne dépasse pas le seuil diagnostique du diabète, mais qui est liée à une probabilité accrue de développer ultérieurement un diabète de type 2.

Il est défini, selon les critères proposés en 2006 par l'Organisation mondiale de la santé (OMS) par :

— Une hyperglycémie modérée à jeun : glycémie entre 1,10 g/l (6,1 mmol/l) et 1,25 g/l (6,9 mmol/l) après un jeûne de 8 heures et vérifiée à deux reprises.

— Une intolérance au glucose : glycémie (sur plasma veineux) comprise entre 1,4 g/l (7,8 mmol/l) et 1,99 g/l (11,0 mmol/l) 2 heures après une charge orale de 75 g de glucose. [47]

1.9 Diagnostic de diabète

Le diagnostic de diabète implique une évaluation de la glycémie (taux de glucose dans le sang) afin de déterminer si elle est dans la plage normale ou si elle indique un diabète. Il est généralement recommandé pour les personnes présentant des symptômes de diabète ainsi que pour celles ayant

des antécédents familiaux de diabète ou présentant les facteurs de risques cités précédemment. Les tests diagnostiques courants pour le diabète comprennent :

1. Test de glycémie à jeun : ce test mesure la glycémie après au moins 8 heures de jeûne. Si le taux de glycémie est supérieur ou égal à 126 mg/dl (7,0 mmol/l), un diagnostic de diabète peut être posé.
2. Test de tolérance au glucose oral : ce test mesure la glycémie avant et après une boisson sucrée contenant une quantité précise de glucose. Si la glycémie est supérieure ou égale à 200 mg/dl (11,1 mmol/l) après 2 heures, un diagnostic de diabète peut être posé.
3. Test aléatoire de glycémie : ce test mesure la glycémie à un moment aléatoire de la journée. Si le taux de glycémie est supérieur ou égal à 200 mg/dl (11,1 mmol/l) avec des symptômes de diabète tels que soif excessive, urination fréquente et fatigue, un diagnostic de diabète peut être posé
4. Le test d'hémoglobine glyquée (HbA1c) : permet de mesurer la glycémie à jeune au cours des 3 derniers mois. [48]

FIGURE 1.7 – Glucomète. [15]

1.10 Les résultats de la glycémie

Pour maintenir une glycémie normale, le corps a besoin d'un bon fonctionnement de deux hormones : l'insuline et le glucagon.

— L'insuline, produite par les cellules bêta du pancréas, agit comme une clé de régulation de la glycémie en stimulant la conversion du glucose excédentaire en glycogène, qui est ensuite stocké dans les muscles, les tissus adipeux et le foie en cas d'hyperglycémie.

— En outre, le glucagon, une autre hormone produite par les cellules alpha du pancréas en cas d'hypoglycémie, encourage la décomposition du glycogène en glucose en dehors des repas ou lorsqu'il y a une baisse d'énergie. [17]

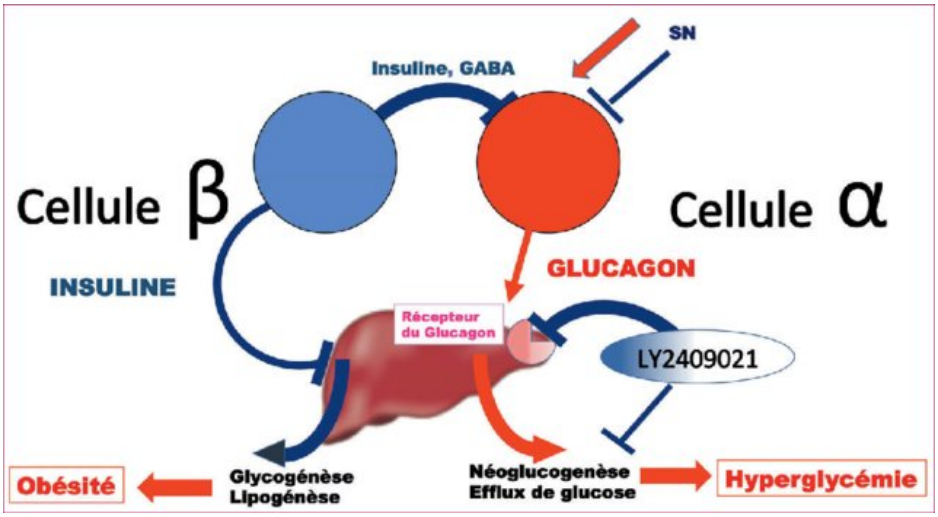


FIGURE 1.8 – Régulation du glucose par l'insuline et le glucagon. [33]

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

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

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

— Si elle est supérieure ou égale à 1,26 g /l et inférieure à 2 g/l : votre médecin va vous prescrire Un second dosage. Si cette seconde glycémie est à nouveau supérieure ou égale à 1,26 g/l, le diabète est confirmé.

— Si elle est d'emblée supérieure ou égale à 2 g/l : le diagnostic de diabète est posé. [30]

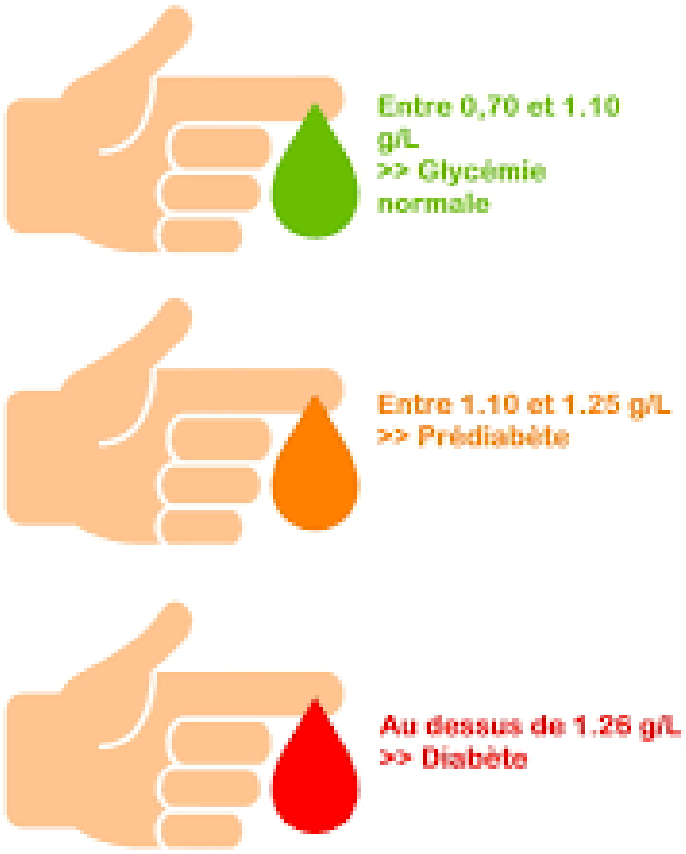


FIGURE 1.9 – Analyse de taux de glycémie. [2]

1.11 Conclusion

Au cours de ce chapitre, nous avons abordé divers aspects du diabète sucré tels que sa définition, ses origines, les différents types existants, les symptômes correspondants ainsi que les causes, traitements et diagnostic associés à chacun d'entre eux. Nous avons également examiné les complications potentielles liées à cette maladie et souligné l'importance de prendre en considération certains facteurs de risque afin de prévenir son apparition.

Dans le chapitre suivant, nous allons explorer les différentes approches de prédiction du diabète qui existent aujourd'hui. Nous ferons un état de l'art en examinant les méthodes les plus importantes et en les comparant entre elles.

Chapitre 2

L'apprentissage automatique et le diabète

2.1 Introduction

Le domaine de la prédiction du diabète par l'utilisation des algorithmes de machine learning a été proposé il y a plusieurs années, les premières recherches sur l'utilisation de techniques de machine learning pour prédire le diabète remontent au début des années 2000.

Depuis, de nombreuses études ont été menées dans ce domaine, montrant que ces algorithmes peuvent être très efficaces pour prédire le diabète. Ils peuvent utiliser une variété de techniques pour identifier les schémas et les relations dans les données des patients diabétiques et prédire leur évolution.

Actuellement, la prédiction du diabète à l'aide de techniques de ML est devenue un domaine de recherche et d'application en constante évolution, avec de nouvelles avancées technologiques et de nouvelles méthodes d'analyse des données qui permettent une meilleure compréhension et une meilleure prédiction de l'évolution de la maladie.

2.2 Machine learning

2.2.1 Définition de l'apprentissage automatique

L'apprentissage automatique (également connu sous le nom de machine learning, d'apprentissage artificiel ou d'apprentissage statistique) est un champ d'étude de l'intelligence artificielle qui repose sur des approches mathématiques et statistiques pour permettre aux ordinateurs d'apprendre à partir de données et d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune d'entre elles. L'apprentissage statistique consiste à créer un modèle dont l'erreur statistique moyenne est la plus faible possible.

Ce domaine est reconnu comme prometteur en raison de la capacité des algorithmes à apprendre des données et à utiliser ces connaissances pour des prédictions et des décisions ultérieures. Il existe un certain nombre d'approches d'apprentissage automatique et de modélisation statistique qui ont jusqu'à présent été impliquées dans divers aspects de la résolution des problèmes. [51]

L'apprentissage automatique se compose généralement de deux phases principales. La première phase, appelée phase d'apprentissage ou d'entraînement, consiste à estimer un modèle à partir de données finies, appelées observations, qui sont disponibles lors de la conception du système. La seconde phase, appelée mise en production, intervient une fois que le modèle a été déterminé et permet de soumettre de nouvelles données afin d'obtenir le résultat correspondant à la tâche souhaitée. Certains systèmes peuvent également poursuivre leur apprentissage une fois en production, pour peu qu'ils aient un moyen d'obtenir un retour sur la qualité des résultats produits. [21]

2.2.2 Types d'apprentissage automatique

On peut répartir le machine learning en trois grandes catégories qui sont décrites ci-dessous :

2.2.2.1 Aprrentissage supervisé

On parle d'apprentissage supervisé lorsque l'on dispose de données d'entraînement étiquetées, c'est à dire dont on connaît la sortie voulue. En notant les N entrées xi et les sorties cibles associées, on dispose de l'ensemble de données D = (x1 ; y1) ; . . . ; (xn ; yn), une fonction f(X), pour tout (xn ; yn) ϵ D et D : ensemble fini de données. On ait f(xn) = yn. L'objectif est d'entraîner le modèle choisi pour qu'il puisse prédire correctement la sortie pour des entrées non étiquetées.

L'apprentissage supervisé est généralement utilisé pour de la régression ou de la classification :

— La régression est utilisée lorsque la sortie à prédire peut prendre des valeurs continues, il s'agit d'une variable réelle.
— La classification est une tâche consistant à choisir une classe (valeur) parmi toutes celles possibles. [48]

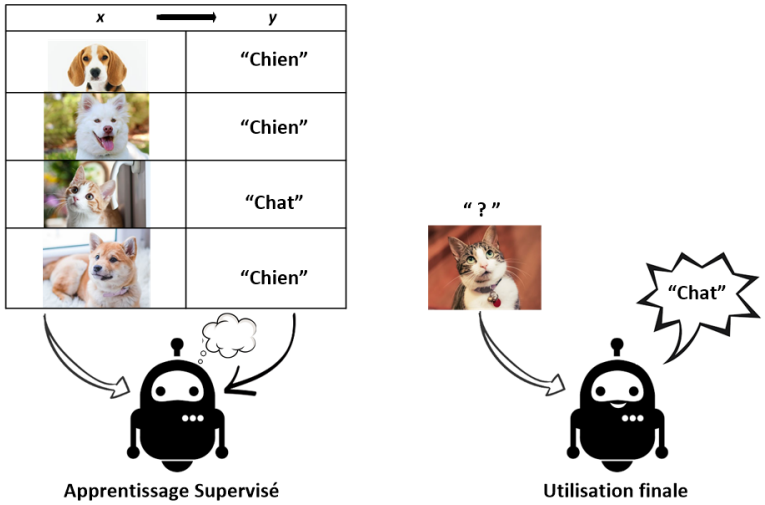


FIGURE 2.1 – Exemple d'apprentissage supervisé. [5]

2.2.2.2 Apprentissage non-supervisé

A la différence de l'apprentissage supervisé, le contexte non supervisé est celui ou l'algorithme doit opérer à partir d'exemples non étiquetés. Il doit extraire automatiquement les catégories à associer aux données qu'on lui soumet.

Les algorithmes d'apprentissage non supervisé sont principalement composés d'algorithmes de regroupement (ou clustering). Ces algorithmes cherchent à diviser les données d'entrée en un nombre prédéfini de groupes. Chaque élément d'un groupe doit avoir des caractéristiques similaires à celles des autres éléments du même groupe, mais différentes de celles des éléments des autres groupes. Ces algorithmes permettent donc de regrouper les entrées en familles pour les catégoriser automatiquement. Par exemple, un algorithme de regroupement peut être utilisé pour classer des patients en fonction de leurs symptômes, afin de prédire les réactions possibles à certains traitements. [48]

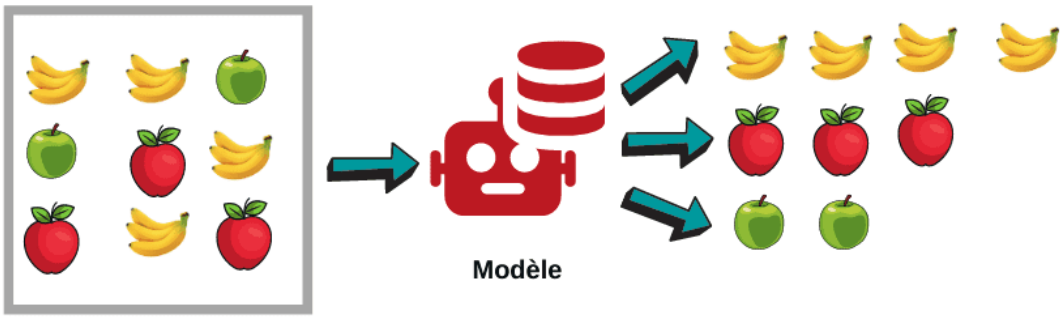


FIGURE 2.2 – Exemple d'apprentissage non-supervisé. [3]

2.2.2.3 Apprentissage par renforcement

Le renforcement learning ou l'apprentissage par renforcement est une méthode de machine learning dont l'objectif est de permettre à un agent (entité virtuelle : robot, programme, etc.), placé dans un environnement interactif (ses actions modifient l'état de l'environnement), de choisir des actions maximisant des récompenses quantitatives.

Au fil du temps, l'agent améliore sa stratégie d'action en utilisant les récompenses obtenues pour ajuster son comportement en fonction des récompenses fournies par l'environnement. L'apprentissage par renforcement est souvent utilisé pour résoudre des problèmes de prise de décision dans des environnements complexes et dynamiques, tels que la robotique, les jeux vidéo et la finance. [4]

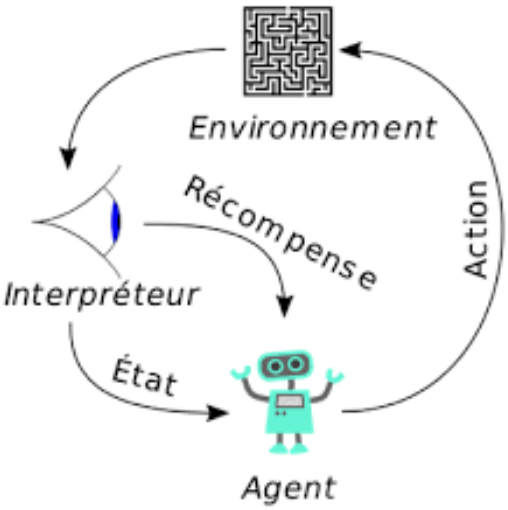


FIGURE 2.3 – Apprentissage par renforcement. [4]

2.2.3 Les algorithmes de machine learning

Il existe plusieurs méthodes de ML proposées dans la littérature, dont certaines sont citées dans la figure suivante.

2.2.3.1 K-plus proche voisins

L'algorithme des k plus proches voisins, également connu sous le nom de KNN ou k-NN (K nearest neighbors en anglais), est un algorithme d'apprentissage supervisé non paramétrique qui se sert de la proximité entre les points de données pour effectuer des prédictions ou des classifications sur un point de données individuel. Il est utilisé pour résoudre des problèmes de classification et de régression, mais il est souvent préféré pour la classification. [19] Cette préférence est dû principalement à :

— La simplicité algorithmique de la méthode comparée aux autres méthodes globales telles que les réseaux de neurones ou les algorithmes génétiques.
— La méthode KNN a démontré empiriquement une importante capacité de prédiction. [38]

Algorithme de construction de KNN

1. Sélectionnez le nombre K des voisins.
2. Pour chaque exemple de l'ensemble de données :
2.1. Calculez la distance entre l'exemple de requête et l'exemple actuel à partir des données.
2.2. Ajouter la distance et l'index de l'exemple à une collection ordonnée.
3. Trier cette collection de distances et d'indices du plus petit au plus grand (par ordre croissant) ordonnée par les distances.
4. Choisi les k premiers entrée de collections.

5. Attribuer l'exemple de requête à la classe où laquelle le nombre de k voisins est maximal (classe le plus fréquent). [34]

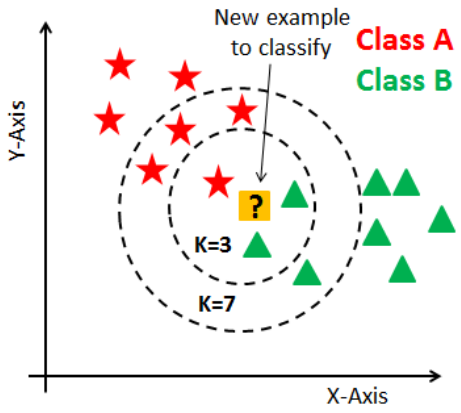


FIGURE 2.4 – Algorithme KNN. [34]

2.2.3.2 Machine a vecteur de support

Le SVM (Support Vector Machine) est l'un des algorithmes d'apprentissage automatique supervisé les plus populaires, utiliser pour les problèmes de classification et de régression, il est principalement utilise pour les problèmes de classification dans l'apprentissage automatique. Son but est de trouver la frontière optimale entre les points de données d'un ensemble de caractéristiques. En général, le SVM tente de trouver un hyperplan qui maximise la marge de séparation entre les deux classes, ce qui correspond à la meilleure ligne de séparation possible. [40]

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

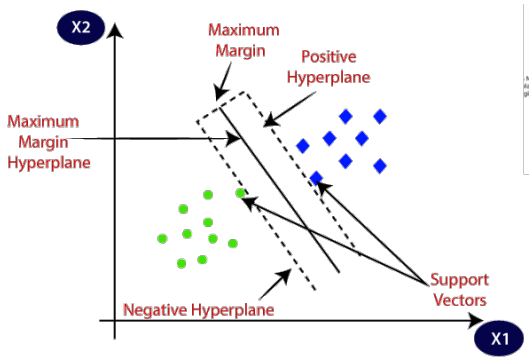


FIGURE 2.5 – Algorithme SVM [1]

2.2.3.3 Naïve Bayes

La classification bayésienne représente une méthode d'apprentissage supervisé ainsi qu'une méthode de classification statistique. La technique du classificateur Naive Bayes est basée sur le théorème de Bayes et est utilisée en particulier lorsque la dimensionalité des entrées est élevée.

La classification bayésienne fournit des algorithmes d'apprentissage pratiques et des connaissances antérieures, où les données observées peuvent être combinées. Elle calcule la probabilité hypothétique apparente. L'algorithme fonctionne comme suit. Le théorème de Bayes offre une façon de calculer la probabilité d'une hypothèse basée sur nos connaissances antérieures. Il fonctionne sur la base de la probabilité conditionnelle.

Il peut être représenté comme suit :

$$P(M/N) = \frac{P(M/N)*P(M)}{P(N)}$$

Ici, M et N sont deux événements et P(M—N) est la probabilité conditionnelle de M étant donné N. P(M) est la probabilité de M. P(N) est la probabilité de N. P (N—M) est la probabilité conditionnelle de N étant donné M. Le classificateur Naive Bayes est un prédicteur puissant et robuste. Cette technique peut être utile pour un très grand nombre d'ensembles de données [48] . Le classificateur bayésien Naive est rapide et incrémental et peut traiter des attributs discrets et continus. Il a d'excellentes performances et peut expliquer les décisions. [36]

2.2.3.4 Régression Logistique

La régression logistique est un modèle statistique utilisé pour étudier les relations entre une variable de sortie binaire (Y) et un ensemble de variables d'entrée (Xi). Il s'agit d'un modèle linéaire généralisé qui utilise une fonction logistique comme fonction de lien pour modéliser la probabilité que Y prenne la valeur 1 ou 0 en fonction des variables d'entrée.

Le modèle de régression logistique est utilisé pour prédire la probabilité qu'un événement se produise (valeur de 1) ou non (valeur de 0) en optimisant les coefficients de régression. Le résultat de la prédiction est une probabilité comprise entre 0 et 1. Lorsque la valeur prédite est supérieure à un seuil (généralement 0,5), l'événement est considéré comme susceptible de se produire, alors que lorsque cette valeur est inférieure au même seuil, il ne l'est pas.

Si l'entrée est $X = x1, x2, x3, ..., xn$, alors la régression logistique cherche à trouver une fonction h telle que nous puissions calculer :

$$\text{y =1 si h(X)} \geq \text{ seuil, 0 si h(X)} < \text{seuil}$$

La fonction h doit être une fonction sigmoïde, c'est-à-dire une fonction qui se situe entre 0 et 1, paramétrée par les coefficients de régression à optimiser. La fonction sigmoïde est définie sur R à valeurs dans [0,1] et s'écrit comme suit : $1/(1 + exp(-z))$, où z est une combinaison linéaire des variables d'entrée et des coefficients de régression.

Graphiquement, la fonction sigmoïde est une courbe en forme de S qui a des limites de 0 et 1 lorsque x tend respectivement vers $-\infty et + \infty$, passant par$y = 0,5$ en $x = 0$. [23]

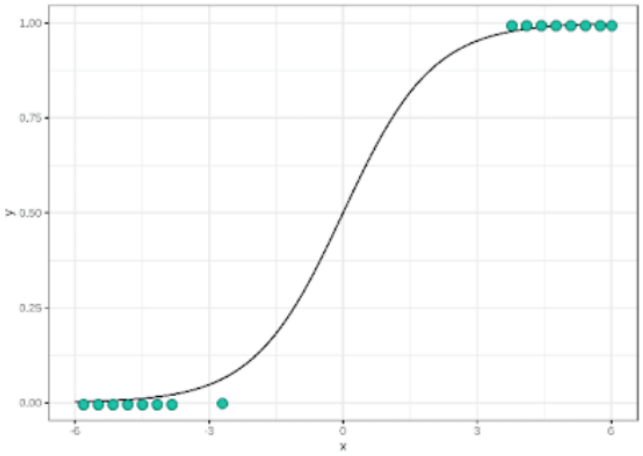


FIGURE 2.6 – La fonction segmoid [13]

2.2.3.5 Arbre de décisions

L'arbre de décisions est un algorithme de machine permet à une organisation ou une personne d'évaluer différentes actions possibles en fonction des bénéfices, des probabilités et des coûts. Il se base pour ce faire sur un ensemble de données exploitable.

Cet algorithme peut également alimenter une discussion formelle. Ce modèle très connu a donné naissance à des algorithmes puissants tels que XGBoost ou Random Forest (forêt d'arbres). Les arbres de décision sont le plus souvent constitués d'un noeud central à partir duquel peuvent être tirées plusieurs Data possibles. Les noeuds conduisent à d'autres noeuds qui à leur tour font ressortir plusieurs autres possibilités. [7] On obtient un schéma de la forme d'un arbre avec des branches multiples.
On distingue trois types de noeuds :

— Les noeuds de hasard
— Les noeuds de décision
— Les noeuds terminaux

Représenté par un cercle, le noeud de hasard met en évidence les probabilités de certaines Data. Le noeud de décision est représenté par un carré. Il illustre une décision qui doit être prise. Le noeud terminal permet d'avoir le résultat final d'un chemin sur les arbres de décision.

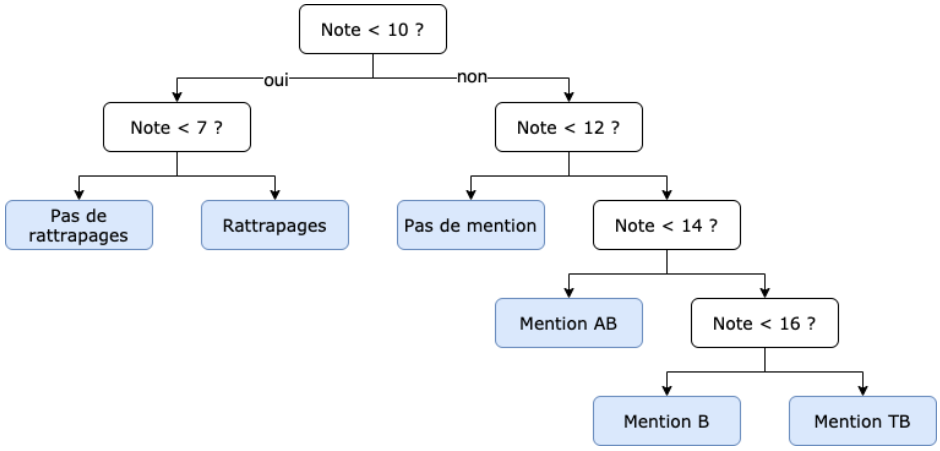


FIGURE 2.7 – Exemple arbre de décision [6]

2.2.3.6 Foret Aléatoire

Le random forest est composé de plusieurs arbres de décision, entrainés de manière indépendante sur des sous-ensembles du data set d'apprentissage (méthode de bagging). Chacun produit une estimation, et c'est la combinaison des résultats qui va donner la prédiction finale qui se traduit par une variance réduite. En somme, il s'agit de s'inspirer de différents avis, traitant un même problème, pour mieux l'appréhender. Chaque modèle est distribué de façon aléatoire en sous-ensembles d'arbres décisionnels.
On a 2 types :

— Random forest de la régression (Reposant sur un système de bagging) consiste schématiquement à calculer la moyenne des prévisions obtenues par l'ensemble des estimations des arbres décisionnels de la forêt aléatoire.
— Random forest de la classification, l'estimation finale consiste à choisir la catégorie de réponse la plus fréquente. Plutôt qu'utiliser tous les résultats obtenus, on procède à une sélection en recherchant la prévision qui revient le plus souvent. [14]

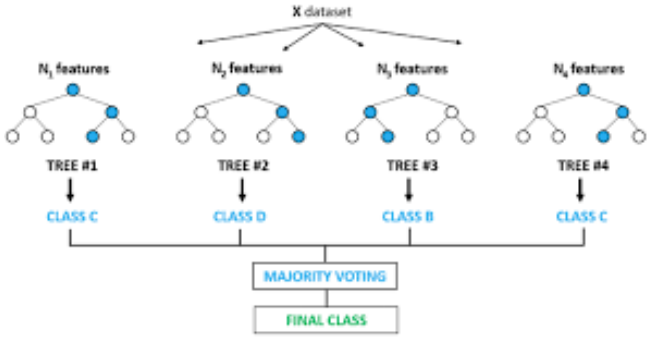


FIGURE 2.8 – Exemple de foret aléatoire [14]

2.2.3.7 Le Gradient Boosting Machine

Le Gradient Boosting Machine (GBM) est une technique d'apprentissage automatique qui fait partie de la famille des méthodes d'ensemble. Il s'agit d'un algorithme de renforcement qui combine plusieurs modèles d'apprentissage faibles pour former un modèle fort.

Le GBM fonctionne en itérant sur des arbres de décision simples appelés "arbres de renforcement". À chaque itération, un nouvel arbre est ajouté à l'ensemble pour corriger les erreurs de prédiction des arbres précédents. Le modèle final est une combinaison pondérée de ces arbres individuels.

L'entraînement du GBM se fait par minimisation d'une fonction de perte (par exemple, la régression logistique pour un problème de classification ou l'erreur quadratique moyenne pour un problème de régression). L'algorithme utilise des techniques de gradient pour optimiser cette fonction et ajuster les poids des arbres de manière itérative.

Le GBM est apprécié pour sa capacité à capturer des relations complexes entre les variables d'entrée et la variable cible. Il est également résistant au surajustement et peut gérer efficacement des ensembles de données de grande taille. Le GBM est largement utilisé dans des domaines tels que la prédiction, la classification, le ranking et la recommandation. [32]

2.2.3.8 EXtreme Gradient Boosting Algorithm

XGBoost est un nouvel algorithme d'apprentissage automatique qui est né en février 2014. Cet algorithme a suscité une grande attention en raison de son excellent pouvoir d'apprentissage et de sa vitesse d'entraînement efficace. L'algorithme XGBoost est une amélioration de l'arbre de décision de renforcement du gradient (GBDT) et peut être utilisé à la fois pour des problèmes de classification et de régression. Il convient de noter que XGBoost est également l'un des algorithmes d'arbres de renforcement, ce qui signifie qu'il intègre de nombreux classifieurs faibles pour former un classifieur fort. Le modèle d'arbre qu'il utilise est le modèle d'arbre de classification et de régression (CART).

L'idée de cet algorithme est d'ajouter continuellement des arbres et de diviser les caractéristiques de manière continue pour faire pousser un arbre. À chaque ajout d'un arbre, vous apprenez en fait une nouvelle fonction pour ajuster le résidu prédit précédemment. Lorsque nous avons fini de former k arbres, nous devons prédire le score d'un échantillon. En réalité, en fonction des caractéristiques de cet échantillon, un noeud feuille correspondant tombera dans chaque arbre, et chaque noeud feuille correspond à un score. [35]

Le score correspondant à chaque arbre doit être ajouté pour être la valeur prédite de l'échantillon. Plus précisément, le déroulement de l'algorithme est le suivant :

1. Avant de commencer à itérer sur le nouvel arbre, calculez les matrices de première et deuxième dérivées de la fonction de perte correspondant à chaque échantillon.

2. À chaque itération, ajoutez un nouvel arbre, et chaque arbre s'adapte au résidu de l'arbre précédent.
3. Calculez la valeur de gain de division de la fonction objective pour sélectionner le meilleur point de division, et utilisez l'algorithme glouton pour déterminer la meilleure structure de l'arbre.
4. Ajoutez un nouvel arbre au modèle et multipliez-le par un facteur pour éviter le surajustement. Lors de l'ajustement des résidus, une taille de pas ou un taux d'apprentissage est généralement utilisé pour contrôler l'optimisation, afin de réserver plus d'espace d'optimisation pour l'apprentissage ultérieur.
5. Après l'entraînement, un modèle de plusieurs arbres est obtenu, dans lequel chaque arbre a plusieurs noeuds feuilles.
6. Dans chaque arbre, l'échantillon tombe sur plusieurs noeuds feuilles en fonction des valeurs propres. La valeur prédite finale est le score du noeud feuille correspondant à chaque arbre multiplié par le poids de l'arbre.

En raison de ses performances élevées et de sa flexibilité, XGBoost est largement utilisé dans la compétition de science des données Kaggle et est appliqué avec succès dans divers domaines, y compris la finance, la recherche médicale, la recommandation de produits et la prédiction de la demande.

2.3 Etat de l'art

2.3.1 Les travaux de recherche portant sur l'utilisation des algorithmes de machine learning pour la prédiction du diabète

Les avancées en matière d'apprentissage automatique ont suscité un grand intérêt dans le domaine médical. Les chercheurs ont ainsi exploré la possibilité d'utiliser cette technologie pour améliorer les traitements dans des domaines où les algorithmes classiques montrent des limites.

Dans ce qui suit, nous allons examiner les recherches les plus récentes sur la prédiction du diabète, qui sont basées sur des techniques d'apprentissage automatique. Nous verrons comment ces dernières peuvent aider à prédire l'apparition du diabète chez les patients et améliorer la précision des résultats obtenus :

Article 1 : Rashi Rastogi et Al ont proposés les techniques de data mining pour la prédiction du diabète, ils ont appliqué quatre techniques d'exploration de données Naïve bayes, SVM, Régression logistique et Random forest.

Ces méthodes ont été testées sur un ensemble de données de 614 patients collectées sur Kaggle, après le prétraitement et l'analyse de données, les résultats ont montré que la régression logistique a obtenu une précision élevée de 82,46% et des résultats meilleurs par rapport aux

autres techniques. [45]

Article 2 : Muhammad Exell et Al ont mené une analyse comparative entre deux algorithmes, le KNN et le Naïve Bayes pour prédire le diabète à partir de plusieurs attributs de santé dans un ensemble de données en utilisant l'apprentissage supervisé.

Les résultats ont montré que le Naïve Bayes était supérieur au KNN, avec une valeur moyenne d'exactitude de 76,07%, une précision de 73,37% et un rappel de 71,37% pour Naïve Bayes, tandis que le KNN avait une valeur moyenne d'exactitude de 73,33%, une précision de 70,25% et un rappel de 69,37%. En conclusion, le Naïve Bayes est meilleur que le KNN pour prédire le diabète à partir de l'ensemble de données des Indiens Pima. [31]

Article 3 : Chollette C et Al ont proposé un cadre d'apprentissage automatique puissant pour améliorer la prédiction du diabète, ce cadre utilise des approches de prétraitement des données, la corrélation de Spearman et la régression polynomiale pour améliorer la performance des modèles RF, SVM et 2GDNN proposés dans l'article.

Les résultats montrent que ce cadre obtient des performances remarquables avec une précision de classification exceptionnelle de 97,931% et 100% sur l'ensemble de données indien PIMA et une précision de 97,33% sur l'ensemble de données LMCH. [40]

Article 4 : Aghila et Al ont présenté un nouveau cadre hybride d'apprentissage automatique de réseau de neuronal artificiel pour prédire le diabète en utilisant des techniques de régularisation et de prédiction personnalisées. Ils ont pris en considération l'ensemble de données médicales de pima indians diabetes et les normalisées avec un algorithme décisionnel qui fonctionne d'une manière cohérente pour tous les dégrées asymétrie, en fin ils ont obtenu des résultats plus précis que les modèles déjà existants pour cette population. [43]

Article 5 : Victor et Al ont utilisé le système de surveillance de facteurs de risque comportementaux (BRFSS) pour créer un système de prédiction de diabète permettant de repérer et de diagnostiquer la maladie à un stade précoce.

L'objectif principal de leur étude était de comparer plusieurs algorithmes de prédiction en utilisant différents modèles. La méthodologie de l'étude comportait six étapes : une étude de la littérature, l'acquisition des données, l'analyse exploratoire, le prétraitement des données, la sélection du modèle et l'évaluation du modèle. Les résultats ont montré que les réseaux de neurones et les machines à vecteurs de support étaient les algorithmes les plus performants pour la prédiction et le diagnostic précoces du diabète. Les chercheurs ont également constaté que la normalisation des données était une technique de prétraitement efficace pour améliorer les performances des algorithmes. [29]

Article 6 : R.Ranjitha et Al ont utilisé l'algorithme de rétropropagation d'un modèle de réseau de neurones artificiels sur un jeu de données PIMA indien qui contient un ensemble d'attributs sur les patients pour la prédiction précoce de la maladie de diabète.

Dans cette étude, ils ont commencé par le prétraitement des données à l'aide de l'outil Weka, où ils ont identifié les valeurs manquantes, suivi de l'identification et de la suppression des données aberrantes. Ensuite, ils ont sélectionné la fonction et trouvé un filtre de corrélation coefficient, et enfin effectué la normalisation des données. Deuxièmement, ils ont itéré le modèle de réseau de neurones artificiels avec une couche cachée, puis avec deux couches cachées. Finalement, ils ont obtenu la meilleure précision avec 250 données qui a atteint 99,20%. [44]

Article 7 : Aishawarya Mujumdar et Al ont utilisé une méthode pour développer un système fiable de prédiction du diabète en suivant plusieurs étapes :

Tout d'abord, ils ont collecté des données en analysant des big data et ont construit un ensemble de données. Ensuite, ils ont préparé les données en les mettant à l'échelle pour les normaliser. En deuxième lieu, ils ont effectué un prétraitement des données en triant les valeurs manquantes. En troisième lieu, ils ont regroupé les données à l'aide de l'algorithme K-means. En quatrième lieu, ils ont construit un modèle en utilisant plusieurs algorithmes d'apprentissage automatique, tels que le classificateur à vecteur de support, le classificateur de forêt aléatoire, le classificateur d'arbre de décision, le classificateur d'arbre supplémentaire, l'algorithme AdaBoost, le perceptron, l'algorithme d'analyse discriminante linéaire, la régression logistique, le k-voisin le plus proche, les bayes naïves gaussiennes, l'algorithme de regroupement et le classificateur à gradient de croissance. Enfin, ils ont évalué les modèles en utilisant la précision de classification et la matrice de confusion, obtenant des résultats satisfaisants. [39]

Article 8 : Jitranjan Sahoo et Al ont cherché de développer un système de prédiction de diabète, ils ont mené un travail expérimental impliquant l'utilisation de six algorithmes de classification de machine learning : régression logistique, naïve bayes, KNN, arbre de décision, RF et SVM. Ces algorithmes ont été évalués à l'aide de différentes mesures pour travailler avec l'ensemble de données PIMA sur le diabète en Inde.

Les résultats ont confirmé que le système conçu avait une précision de 79,17% en utilisant l'algorithme de classification par régression logistique. De plus, ce système peut être adapté pour prédire d'autres maladies. [46]

Article 9 : Anju Prabha et Al ont proposé dans leur travail un système de détection non invasif du diabète sucré basé sur le signal de photopléthysmographie du bracelet et les paramètres physiologiques de base. Ils ont utilisé XGBoost avec une sélection de fonctionnalités (FS) sur un ensemble de données comprenant 217 participants atteints de diabète, de prédiabète et de personnes non diabétiques. Ils ont effectué un prétraitement des données, notamment

le fenêtrage, le spectre de la transformée de Fourier discrète et le spectre de MEL, puis ils ont normalisé et classifié les données à l'aide de XGBoost. Pour améliorer l'efficacité et réduire la complexité, une technique hybride de sélection de caractéristiques a également été appliquée. [42]

Article 10 : Liyang Wang et Al Ont mené une étude dans laquelle ils ont utilisé l'algorithme d'apprentissage ensembliste XGBoost pour prédire le risque de diabète de type 2 sur un échantillon de 380 personnes. Ils ont comparé les résultats de XGBoost à ceux des machines à vecteurs de support, de l'algorithme random forest et de l'algorithme k-nearest neighbor.
Dans un premier temps, les chercheurs ont collecté des données expérimentales en utilisant un questionnaire. Ensuite, ils ont effectué une représentation des vecteurs de caractéristiques à partir de ces données. Ils ont appliqué XGBoost en réglant ses paramètres.Enfin, une analyse comparative a été réalisée pour évaluer les performances des différents algorithmes. Les résultats ont montré que XGBoost était le meilleur. [49]

Article 11 : Mingqi Li et Al ont réalisé une analyse comparative entre l'algorithme XGBoost et des modèles d'algorithmes traditionnels tels que SVM, KNN, NB, DT et LR. Cette analyse a démontré que XGBoost offre la meilleure précision parmi les six algorithmes.

Afin de trouver la solution optimale, ils ont amélioré l'algorithme XGBoost en le combinant avec d'autres algorithmes. Le principe de la combinaison XGBoost + LR consiste à former un modèle XGBoost avec les données, puis à fournir les instances des données d'entraînement au modèle XGBoost afin d'obtenir les noeuds feuilles correspondants, et enfin à utiliser ces noeuds feuilles comme caractéristiques d'entraînement.

En conclusion, on peut affirmer que XGBoost est plus efficace que certains algorithmes traditionnels. [35]

Article 12 : Zhongxian Xu et Al ont présenté un nouveau modèle de prédiction hybride composé de deux parties : un algorithme de forêt aléatoire amélioré pour une sélection optimale des caractéristiques, et l'algorithme XGBoost pour la classification. Afin de permettre une comparaison efficace avec des études antérieures, le même ensemble de données UCI Pima Indian Diabetes a été utilisé pour l'entraînement et le test du modèle.

Ce modèle combine de manière efficace le prétraitement des données avec l'algorithme de sélection amélioré des caractéristiques, en choisissant le sous-ensemble de caractéristiques optimal en fonction du prétraitement des données. Il améliore considérablement la précision de la classification et les performances du modèle, lui permettant de s'adapter à différents ensembles de données. Les résultats expérimentaux démontrent que le modèle atteint une meilleure précision de classification (93,75%) et présente de nombreux avantages par rapport aux méthodes existantes.

Dans un contexte de demande croissante d'analyse des données médicales, ce modèle propose

une grande aide aux chercheurs et aux médecins, leur permettant de prendre des décisions plus précises pour les patients. [50]

2.3.2 Analyse et comparaison

Une étude comparative des approches proposées, en utilisant les facteurs suivants :

1. **Approche :** l'approche proposée dans chaque article.
2. **Dataset :** indique les sources des données utilisées pour l'implémentation de l'approche pour la prédiction du diabète.
3. **Résultats :** les résultats de l'approche.
4. **Techniques utilisées :** les techniques utilisées pour prédire le diabète cité dans l'article.
5. **Avantages :** les avantages de l'approche abordée.
6. **Inconvénients :** les inconvénients de l'approche abordée.

Article 1 : Diabetes prediction model using data mining techniques, authors

— Approche : Rashi Rastogi et Mamta Bansal
— Dataset : données collecter sur Kaggle
— Résultats : dans le modèle de régression logistique la précision est élevée (82,46%) par rapport aux autres modèles, tandis que dans le SVM, la précision est faible (79,22%) par rapport aux autres Naïve Bayes : 79,22% RF : 81,81%
— Techniques utilisées : SVM, Régression logistique, Naïve bayes, classificateur de foret aléatoire
— Avantages : Les méthodes sont robustes aux données manquantes et aberrantes et aux variables non pertinentes La régression logistique est une méthode simple à mettre en oeuvre et interprétable
— Inconvénients : La régression logistique ne peut pas être utilisée pour la classification multiclasse, et peut avoir des difficultés avec des ensembles de données très grands. Les interprétations sont un peu limitées car les techniques de data mining ne fournissent pas toujours une explication claire des résultats.

Article 2 : Diabetes prediction using supervised machine learning

— Approche : Muhammad Exell Febrian, Fransiskus Xaverius Ferdinan, Gustian Paul Sendani, Kristien Margi Suryanigrum, Rezki Yunanda
— Dataset : Indiens Pima diabetes database
— Résultats : KNN précision (73,33%) Naïve bayes précision (76,07%) Techniques utilisées : KNN et Naïve bayes

— Avantages : Naïve bayes est un modèle rapide et évolutif, sa procédure est parallèle et il peut être utiliser pour la classification binaire et multiclasse
— Inconvénients : - KNN est sensible aux données bruyantes et aberrantes, problème de curse of dimensionality (la précision se réduit lorsque la dimension des données est élevée)

Article 3 : Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective

— Approche : Chollette C. Olisah, Lyndon Smith, Melvyn Smith
— Dataset : Indiens Pima diabetes dataset et LMCH (laboratory of medical city hospital) dataset
— Techniques utilisées : RF (forêt aléatoire), SVM (machine à vecteur de support) et 2GDNN (réseau neuronal profond à double croissance)
— Résultats : 2GDNN : Précision de 97,248% sur indiens PIMA et 97,333% sur LMCH
— Avantages : RF : facile à utiliser et à mettre en oeuvre, rapide et efficace, il peut traiter des données à grande échelle.
SVM : il est relativement plus facile à interpréter par rapport aux autres modèle de machine learning.
DNN : très performant pour des taches de classification et de prédiction complexe.
— Inconvénients : RF : problème de surajustement s'il est mal gérer.
SVM : sensible aux choix des hyperparamètres et il peut être lent à entrainer pour les grandes quantités de données.
DNN : Ils peuvent être complexe et difficile à mettre en oeuvre, ils ont souvent besoin d'un grand nombre de données d'entraînement pour fonctionner efficacement, ce qui peut être difficile à obtenir dans certains cas.

Article 4 : A novel hybride machine learning framework for the prediction of diabetes with context-customized regularized and prediction procedures

— Approche : Aghila Rajagopal, Sudan Jha, Ramachandran Alagarsamy, Shio Gai Quek, Ganeshsree Selvachandran
— Dataset : Pima Indians Diabetes (768 femmes de descendance indienne Pima, 500 sont normales et 268 diabétiques
— Résultats : Modèle hybride de réseau neuronal artificiel personnalisé à atteindre une précision de 80 à 81%.
— Techniques utilisées : -Modèle hybride de réseau neuronal artificiel personnalisé
— Avantage : -Le travail aborde véritablement le contexte de diabète.
- Ils ont utilisées tous les données collectées et ils n'ont rejeté aucune dans les 768.
-L'algorithme peut fonctionner dans des conditions moins idéales.
-La combinaison de plusieurs algorithmes de prédiction améliore la prédiction .

— Inconvénients : -L'absence de certaines données dans le dataset par contre le modèle hybride est besoin des données précises.
-La méthode de prédiction personnalisée rend le système complexe.

Article 5 : An assessment of machine learning models and algorithms for early prediction and diagnosis of diabets using health indicators

— Approche : Victor Chang, Meghana ashok Ganatra, Karl Hall, Lewis Golightly, Qianwen Ariel Xu.
— Dataset : : BRFSS (Un système de surveillance de la sante publique au États-Unis)
— Résultats : foret aléatoire 82.26%, naïve Bayes 70.56%, régression logistique 72.64%, KNN 80.55%, arbre de décision 81.02%
— Techniques utilisées : arbre de décision, régression logistique, foret aléatoire, KNN, naïve Bayes.
— Avantage : - La mise en oeuvre de plusieurs classificateurs d'ensemble et des techniques associe améliorerait les performances prédictives.
-Obtenir un système de diagnostic automatique plus efficace.
— Inconvénients : - Un système complexe.

Article 6 : Diabetes Prediction by Articial Neural Network

— Approche : R.Ranjitha, V.Agalya and K.Archana
— DataSet : PIMA (environs 768 patients de la tribu Pima située dans l'Arizona)
— Techniques utilisées : réseau neuronal artificiel (ANN) .
— Résultats : ANN 99%
— Avantage : -Avoir un système conçu pour détecter le diabète avec une précision très élevée.
-L'approche proposée peut l'appliquer sur d'autres maladies.
— Inconvénients : -Les ANN ne sont pas toujours compréhensible .
-La mise en oeuvre des ANN nécessite certaine l'expertise dans le domaine.

Article 7 : Diabetes prediction using Machine Learning Algorithms

— Approche : Aishawarya Mujumdar, Dr.Vaidehi
— DataSet : Des données collecter par l'analyse des big data (800 patients)
— Techniques utilisées : K-means, algorithmes d'apprentissage automatique qui comprennent : le classificateur a vecteur de support, le classificateur de foret aléatoire, classificateur d'arbre de décision, le classificateur d'arbre supplémentaire, l'algorithme Ada Boost, le perceptron, l'algorithme d'analyse discriminante linéaire, la régression logistique, le K-voisin le plus proche, les bayes naïves gaussiennes, algorithme de regroupement, classificateur a gradient de croissance.

- Résultats : les degrés de précision pour chaque algo : le classificateur a vecteur de support(90%), le classificateur de foret aléatoire (91%), classificateur d'arbre de décision (86%), le classificateur d'arbre supplémentaire(91%), l'algorithme Ada Boost(93%), le perceptron(76%), l'algorithme d'analyse discriminante linéaire (94%), la régression logistique(96%), le K-voisin le plus proche(90%), les bayes naïves gaussiennes(93%), algorithme de regroupement(90%), classificateur a gradient de croissance(93%).
- Avantage : - L'étude approfondie de l'utilisation des algorithmes d'apprentissage automatique.
  - Utilisation de plusieurs algo permis de comparer les performances.
- Inconvénients : - utilisations des algorithme d'apprentissage automatique sur un seul ensemble de données peut réduire la généralisation sur d'autres modèles.
  -Les couts et les difficultés de mise en oeuvre ces modèles dans l'environnement médical.

Article 8 : Diabetes Prediction Using Machine Learning Classification Algorithms

- Approche : Jitranjan Sahoo, Manoranjan Dash, Abhilash Pati
- Dataset : PIMA indiens dataset
- Techniques utilisées : régression logistique, naïve bayes, KNN, arbre de décision, RF et SVM
- Résultats : Précision : Régression logistique 79,17%, Naïve bayes 74,48%, KNN 74,48%, Arbre de décision 71,88%, RF 77,08%, SVM 76,56%.
- Avantages : Les algorithmes de classification sont relativement faciles à comprendre et à mettre en oeuvre, même pour les utilisateurs novices.
- Inconvénients : -Les algorithmes de classification ne peuvent pas prendre en compte tous les facteurs de risque potentiels, ce qui peut limiter leur précision.
  -Ils nécessitent des données de haute qualité pour produire des résultats précis, ce qui peut être difficile à obtenir pour certains ensembles de données.

Article 9 : Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier

- Approche : Anju Prabha, Jyoti Yadav, Asha Rani, Vijander Singh
- Dataset : ensembles de données de 217 participants
- Techniques utilisées : XGBoost .
- Résultats : XGBoost 99.93%.
- Avantages : - Une amélioration considérable de la précision par rapport aux autres techniques existantes a été observée.
  - Un système compétent capable de tester de manière fiable des données provenant de différentes populations a été développé.
- Inconvénients : - Au fil du temps, des facteurs physiologiques tels que la température peut

affecter le système.

Article 10 : Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model

- Approche : Liyang Wang, Xiaoya Wang, Angxuan Chen, Xian Jin and Huilian Che
- Dataset : ensembles de données de 380 personnes.
- Techniques utilisées : XGBoost, Random Forest, KNN, SVM.
- Résultats : XGBoost 89.09%, SVM 81.58%, Random Forest 78.95%, KNN 73.68%.
- Avantages : l'étude comparative démontrer l'efficacité du modèle proposé par rapport aux autres méthodes.
- Inconvénients : Un petit échantillon de données peut limiter la généralisation des prédictions.

Article 11 : Diabetes Prediction Based on XGBoost Algorithm

- Approche : Mingqi Li, Xiaoyang Fu, Dongdong Li.
- Dataset : 768 personnes de l'institut nationale du diabète et des maladies digestives et rénales
- Techniques utilisées : Comparaison entre l'algorithme XGBoost et les modèles d'algorithmes traditionnels tels que SVM, KNN, NB, DT, LR.
- Résultats : XGBoost : 81.2%, SVM : 66.5%, KNN : 71.9%, NB : 76.7%, DT : 70.3%, LR : 76.7%.
- Avantages : Gestion des caractéristiques et Robustesse aux données manquantes.
- Inconvénients : Complexité computationnelle : XGBoost peut être relativement complexe en termes de calculs et de ressources informatiques nécessaires.

Article 12 : A Risk Prediction Model for Type 2 Diabetes Based on Weighted Feature Selection of Random Forest and XGBoost Ensemble Classifier

- Approche : Zhongxian Xu, Zhiliang Wang.
- Dataset : Indiens Pima PIDD.
- Techniques utilisées : xgboost, foret aléatoire .
- Résultats : modèle proposé 93.75%.
- Avantages : - Prédictions plus robustes et plus fiables en prenant en compte différentes perspectives et méthodes de modélisation.
  -Précision accrue et capacité à traiter des ensembles de données complexes et à détecter des modèles cachés.
- Inconvénients : - Une validation insuffisante peut conduire à des prédictions incorrectes ou biaisées.

Au cours des dernières années, de nombreuses études sur la prédiction du diabète ont connu une croissance considérable grâce aux avancées dans le domaine de l'Intelligence Artificielle, en particulier l'apprentissage automatique et l'analyse de données.

Pour cela nous avons choisis douze études prédictives conçus pour prédire le diabète, ces études analysées visent a bien amélioré la précision et obtenir des résultats meilleurs, plus efficaces et plus précis en examinant les données à travers diverses approches telles que le Machine Learning, tout en basant sur des données médicales et plusieurs facteurs de risque liés au diabète comme l'âge, la glycémie, l'IMC, la pression artérielle etc... .

Dans ces travaux, plusieurs algorithmes de ML ont été abordés et examinés pour prédire le diabète dont la régression logistique, naïve bayes, KNN, arbre de décision, RF et SVM, GBM et XGBoost.

Les résultats ont montré que ces algorithmes peuvent aider à soutenir la prise de décision médicale et à améliorer le traitement et le pronostic des patients atteints cette maladie chronique.

Les algorithmes de ML possèdent de nombreux avantages tels que la capacité de traiter de grandes quantités de données et de détecter des modèles complexes, ainsi que la capacité d'apprendre et de s'adapter aux données en temps réel pour améliorer la précision des prévisions. Cependant, ils ont également des inconvénients tels que la nécessité d'un grand nombre de données pour entraîner les modèles, la complexité des modèles utilisés et la difficulté de comprendre comment les prévisions sont faites.

Malgré ces inconvénients, les algorithmes de Machine Learning restent une méthode prometteuse pour améliorer la prédiction et le traitement du diabète.

Dans le cadre de notre étude, nous avons développé un système de prédiction du diabète qui s'appuie sur une approche inspirée des travaux connexes, notamment ceux portant sur le Machine Learning.

2.4 Conclusion

Dans ce chapitre, nous avons expliqué les concepts nécessaires à propos de l'apprentissage automatique, y compris la définition et les types d'apprentissage, ainsi que leurs algorithmes. A la fin, nous avons rédigé l'état de l'art qui contient un résumé des articles et une étude comparative entre les approches proposées et les travaux connexes.

Ce chapitre nous aide à choisir notre approche qui sera la base de notre système de prédiction que nous allons programmer dans le chapitre quatre.

Chapitre 3

Présentation de l'établissement d'accueil

3.1 Introduction

Pour développer des systèmes de prédiction fiables pour le diabète, il est crucial d'avoir accès à des données précises et solides concernant les patients diabétiques.

C'est la raison pour laquelle nous avons choisi de mener notre étude au sein du service de médecine interne du centre hospitalo-universitaire de Bejaia Khellil Amrane. Ce service est doté d'une équipe médicale hautement compétente et spécialisée dans la prise en charge des patients diabétiques, avec une culture de collaboration et de mentorat, facilitant ainsi l'échange d'informations pour répondre à nos besoins.

3.2 Sa création

Le CHU de Bejaïa a été créé par le décret exécutif n° 09-319 du 17 Chaoual 1430 correspondant au 6 octobre 2009 complétant la liste des centres hospitalo-universitaires annexée au décret exécutif n° 97-467 du 2 Chaabane 1418 correspondant au 23 décembre 1997 fixant les règles de création, d'organisation et de fonctionnement des centres hospitalo-universitaires. [9] La liste des centres

FIGURE 3.1 – Centre hospitalo-universitaires Khellil Amrane de Bejaia. [8]

hospitalo-universitaires annexée au décret exécutif n°97- 467 du 2 chaabane 1418 correspondant au 2 décembre 1997 susvisé est complétée comme suit :

— Dénomination : CHU Bejaïa
— Siège : Hôpital khellil Amrane
— Consistances Physique :
Hôpital Khellil Amrane
Hôpital Frantz Fanon
Hôpital Targua Ouzemmour (Clinique Mère-Enfant) [9]

3.3 Présentation du CHU-Khellil Amrane

Le secteur sanitaire de Béjaïa comprend plusieurs structures de santé, parmi lesquelles l'hôpital Khellil Amrane.

CHU Khellil Amrane couvre sur une superficie de 460,65 Km2. Il assure une couverture sanitaire aux 240.258 habitants des sept (07) communes suivantes : Béjaïa, Oued-Ghir, Tichy, Tala hamza, Boukhlifa, Aokas et Tizi-Nberber.

Le secteur sanitaire est géré par la direction de l'hôpital Khellil Amrane, situé au chef-lieu de la commune de Béjaïa.

Il est doté d'un budget de fonctionnement et d'une autonomie de gestion. Jusqu'en 1991, date de l'inauguration et de l'entrée en fonction de l'EPH Khellil Amrane, le secteur sanitaire de Béjaïa n'était doté que de deux hôpitaux : Aokas et Frantz Fanon, hérités de la période coloniale. En 2011, l'hôpital Khellil Amrane est devenu le siège du Centre Hospitalo-universitaire (CHU) de Béjaïa. La création de ce dernier est faite suite à l'inauguration de la faculté de médecine.

Le centre hospitalo-universitaire est un établissement public à caractère administratif, doté de la personnalité morale et de l'autonomie financière. Il est créé par décret exécutif, sur proposition conjointe du ministre chargé de la santé et du ministre chargé de l'enseignement supérieur et de la recherche scientifique. Il est placé sous la tutelle administrative du ministre chargé de la santé.

La tutelle pédagogique est assurée par le ministre chargé de l'enseignement supérieur. Le CHU est chargé, en relation avec l'établissement d'enseignement et/ou de formation supérieure en sciences médicales concerné, des missions : de diagnostic, d'exploration, de soins, de prévention, de formation, d'études et de recherche. [9]

3.4 Services de CHU Khellil Amrane Bejaia

Le CHU Khellil Amrane de Bejaia est composé de plusieurs services médicaux tels que :

— Cellule d'accueil et d'orientation des cancéreux.
— Anesthésie réanimation.
— Chirurgie générale.

— Médecine interne.
— Bloc opératoire central.
— Laboratoire Central.
— Pédiatrie.
— Cardiologie.
— Neurochirurgie.
— Orthopédie traumatologie.
— Imagerie médicale.
— Urgences Medicaux Chirurgicale.

3.5 Service Médecine Interne

3.5.1 présentation

- Professeur, chef de service : Professeur OUAIL Djamel Eddine.
- Personnel médical : 01 maitre -assistante hospitalo-universitaire ,08 spécialistes assistants dans 05 disciplines, et 03 généralistes.
- Capacité litière : 32 lits
- Nombre d'unité : 2
- 01 Unité Homme : 14 Lits et une chambre à 2 Lits pour le pénitencier
- 01 unité Femme : 28 Lits.

FIGURE 3.2 – Service de médecine interne. [22]

3.6 Objectif du stage

Au cours de notre stage à l'hôpital Khellil Amrane pour collecter des données sur les patients diabétiques, nous avons travaillé en étroite collaboration avec l'équipe médicale du service de médecine interne pour identifier les patients atteints de diabète afin de former un nouveau dataset.

Nous avons commencé par prendre le temps de comprendre les protocoles et les procédures de collecte de données en place.

A l'aide des médecins et de toutes l'équipe du service, nous avons pu collecter des données sur les patients diabétiques et non-diabétiques et recueillir les informations nécessaires telles que l'âge, le sexe, l'hémoglobine glyquée, les traitements, les niveaux de glycémie, le poids et la taille, le taux de cholestérol, la tension artérielle, les antécédents familiaux et personnels, les grossesses pour les femmes et le tour de taille.

Nous avons suivi les protocoles de confidentialité et de protection des données pour nous assurer que toutes les informations recueillies étaient sécurisées.

3.7 Conclusion

Dans ce chapitre, nous avons présenté l'organisme d'accueil du CHU de Bejaia, ses différents services, en particulier le service de médecine interne.

Dans la fin de ce chapitre nous avons cité l'objectif de notre stage « la collection du dataset » qui est la porte du prochain chapitre.

Chapitre 4

L'approche proposée

4.1 Introduction

Ce chapitre est dédié à l'implémentation de notre approche de prédiction du diabète en exploitant les techniques d'apprentissage automatique. Tout d'abord, nous nous attelons à la tâche de collecte et description des données. Ensuite, nous consacrons une étape cruciale au prétraitement de ces données afin de les rendre adaptées à notre modèle. Une fois cette étape terminée, nous passons à l'entraînement de notre modèle en utilisant divers algorithmes d'apprentissage, que nous sélectionnons avec soin pour obtenir les meilleurs résultats.

Enfin, nous procédons à l'évaluation des performances de chaque modèle afin de déterminer leur efficacité et leur précision dans la prédiction du diabète.

4.2 Approche Proposée

Notre projet vise à réaliser une prédiction précoce du diabète, permettant aux individus de connaître leur risque de développer cette maladie, avec un taux de prédiction précis. Pour atteindre cet objectif, plusieurs étapes doivent être suivies afin d'obtenir des résultats optimaux.

La figure ci-dessous offre un aperçu de l'approche proposée et des différentes étapes qui la composent :

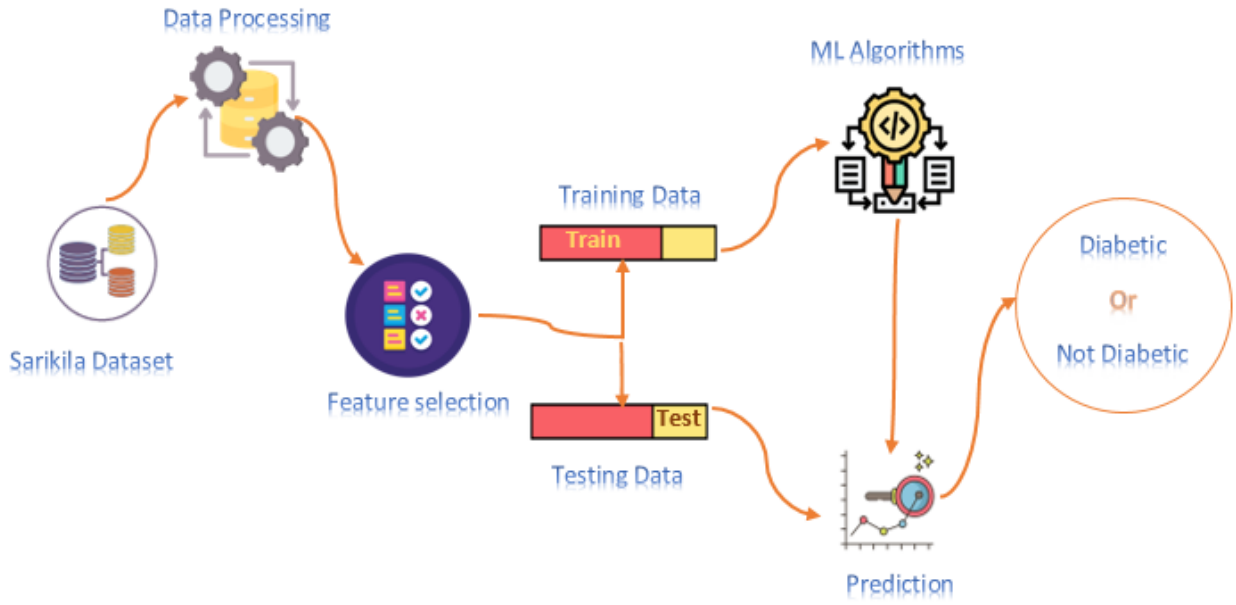


FIGURE 4.1 – Schéma de l'approche proposée

4.3 Outils et librairies utilisés

4.3.1 Google Collab

Colab (ou "Colaboratory") est une plateforme de développement et de collaboration en ligne proposée par Google. Il s'agit d'un environnement de notebook interactif qui permet d'écrire, d'exécuter et de partager du code Python.

Il est largement utilisé dans le domaine de l'apprentissage automatique (machine learning) et de l'analyse de données, car il fournit un accès gratuit à des ressources informatiques puissantes, y compris des GPU et des TPU, qui peuvent accélérer les calculs intensifs.

4.3.2 Python

Python est un langage de programmation polyvalent, utilisé dans de nombreux domaines, y compris la science des données. En ce qui concerne l'apprentissage automatique (machine learning), Python est particulièrement prisé en raison de sa vaste gamme de bibliothèques de haute qualité qui couvrent tous les types d'apprentissage disponibles.

Parmi ces bibliothèques, nous avons utilisé :

4.3.2.1 Matplotlib

Matplotlib est une bibliothèque populaire de visualisation de données en Python. Elle offre une grande flexibilité pour créer une grande variété de graphiques, de diagrammes et de visualisations interactives.

4.3.2.2 Pandas

Pandas est une bibliothèque Python largement utilisée pour la manipulation et l'analyse de données. L'un de ses principaux atouts réside dans sa fonctionnalité de nettoyage des données, qui permet de résoudre efficacement le problème du temps consacré au nettoyage des données dans un projet d'apprentissage automatique. En effet, de nombreux ensembles de données disponibles contiennent des champs vides ou nuls, ce qui peut avoir un impact significatif et négatif sur nos modèles.

4.3.2.3 Numpy

Numpy (Numerical Python) est une bibliothèque fondamentale en Python pour le calcul scientifique et numérique. Elle fournit des structures de données efficaces pour représenter et manipuler des tableaux multidimensionnels, ainsi que des fonctions mathématiques avancées pour effectuer des opérations numériques rapides et efficaces.

4.3.2.4 Seaborn

Seaborn est une bibliothèque Python de visualisation de données basée sur Matplotlib. Elle fournit une interface de haut niveau permettant de créer rapidement et facilement des graphiques statistiques attrayants et informatifs.

4.3.2.5 Sklearn

Également connu sous le nom de sklearn, est une bibliothèque Python populaire utilisée pour l'apprentissage automatique (machine learning). Elle offre une vaste gamme d'outils et d'algorithmes pour la modélisation, l'analyse et la manipulation des données. La bibliothèque offre une implémentation de nombreux algorithmes d'apprentissage automatique couramment utilisés.

4.4 Implémentation de l'approche

4.4.1 Collection de données

La collecte des données est une étape cruciale dans tout projet de recherche ou d'analyse de données. Cela permet de s'assurer que les données utilisées sont fiables, précises et représentatives de la population ciblée. Nous veillons à ce que toutes les informations soient recueillies de manière éthique, confidentielle et sécurisée.

Dans le cadre de notre projet, nous avons accordé une grande importance à la collecte de données. Nous avons fait un stage au CHU de Bejaia (les détails sont présentés dans le chapitre précédent) pour construire un nouveau dataset qui n'a jamais été utilisé auparavant, afin d'avoir un travail personnel et des résultats meilleurs à la fin de l'exécution. Nous sommes donc fiers de présenter un tableau de paramètres pour 590 personnes, comprenant à la fois des femmes et des hommes diabétiques et non diabétiques.

FICHE DES PATIENTS

Patient	Age	sexe	Poids	Taille	IMC	taux chol	TA	Antécédents familiaux	Antécédents personnels	TT	Glycémie à jeun	Outcome
01	43	F	58	1.65	21.30	1.2	150/70	/	Thromboembolique	82	0.95	0
02	40	F	76	1.62	28.96	1.45	130/70	Grand-mère et tente diab	Diab I- goitre - Behçet	95	1.25	1
03	50	F	71	1.65	26.08	1.46	130/50	Père diab	HTA – goitre - allergie	91	1.02	1
04	70	H	75	1.70	25.95	1.45	160/80	Mére diab	Diab I – HTA - Horton	107	1.05	1
05	68	H	47	1.67	16.81	1.36	170/70	/	HTA - Horton	81	1.10	0
589	21	F	75	1.64	27.89	1.58	110/70	Parents diab	Mastite	100	0.84	0
590	53	F	96	1.60	37.50	1.89	120/80	Père diab	Diab II - HTA	118	0.96	1

FIGURE 4.2 – La fiche de paramètre utilisée pour la collecte de données

4.4.1.1 Mise à jour de notre jeu de données

Les antécédents familiaux : ils peuvent être numérisés à l'aide d'une fonction mathématique. Pour chaque membre de la famille, on utilise un score de ligne. Le score est calculé comme suit :

- p représente le nombre de parents (père et mère) diabétiques
- g représente le nombre de grands-parents diabétiques (grand-père et grand-mère)
- f représente le nombre de frères, soeurs, fils et filles diabétiques
- o représente le nombre d'oncles, tantes, cousins et neveux diabétiques
- si aucun membre de la famille n'est diabétique, on ajoute 0,001

Le score S est calculé à l'aide de la formule suivante : [41]

$$S = (0,5 * p) + (0,25 * g) + (0,25 * f) + (0,125 * o) + 0,001$$

Les antécédents personnels : La numérisation des antécédents personnels dans un dataset pour diabète est une étape cruciale pour la prédiction et la prévention de cette maladie. Cependant, nécessite une approche multidisciplinaire impliquant des cliniciens, des experts en données et des scientifiques de la santé pour s'assurer que les données sont complètes, précises et fiables.

Dans notre cas, il est préférable de supprimer ce paramètre pour éviter la confusion ou l'incohérence des données.

Tension artérielle (TA) : la tension artérielle est présentée en deux chiffres, la pression systolique et la pression diastolique. Bien que certains ensembles de données n'utilisent que la pression systolique, nous avons choisi d'utiliser les deux chiffres en les séparant en deux variables distinctes. Cela nous permettra d'obtenir des informations plus complètes sur la tension artérielle des patients dans notre dataset et d'identifier plus facilement les patients à risque de diabète associé à l'hypertension.

4.4.2 Définition de l'enssemble de données

Au finale, Nous avons sélectionné l'ensemble de variables suivant :

Age : l'âge d'une personne correspond au nombre d'années écoulées depuis sa naissance (ans).

Poids : le poids d'une personne (kg).

Taille : la taille d'une personne (m).

IMC : (ou BMI) indice de masse corporelle (poids en kg / (taille en m) 2).

Taux de cholestérol : substance graisseuses dans le sang (g/l).

TA diastolique : pression artérielle diastolique (mm Hg).

TA systolique : pression artérielle systolique (mm Hg).

Diabete Predigme function : le risque de diabète en fonction des antécédents familiaux.

Tour de taille : le tour de taille (cm).

Glycémie à jeun : quantité de glucose dans le sang après avoir jeune au moins 8 heures (g/l).

Outcome : variable de classe (0 et 1) d'où la valeur 0 indique que le patient ne souffre pas de diabète tandis que la valeur 1 indique que le patient est diabétique .

Tableau 4.1 – Description des variables de l'ensemble de données

Variable	Description	Analyse de données
IMC	(poids en kg / taille en m2) IMC de 18.5 à 20 c'est normal, IMC entre 25 et 30 situer dans une plage surpoids Et de 30 ou plus situer dans la fourchette d'obésité	Minimum : 0 Maximum : 90
Taux de cholestérol	Une substance cireuse semblable à une graisse qui est produite par le foie et qui se trouve également dans certains aliments. Pour une bonne santé il est recommandé d'avoir un taux supérieur à 1.4 g/L et inférieur à 2.00 g/L	Minimum : 1.00 Maximum : 3.00
TA diastolique	Si un TA diastolique > 90 signifie une pression artérielle élevée (probabilité élevé de diabète) Un TA diastolique < 60 signifie une pression artérielle basse (moins probabilité de diabète)	Minimum : 60 Maximum : 90
TA systolique	Si un TA systolique > 120 signifie une pression artérielle élevée (probabilité élevé de diabète) Un TA systolique < 90 signifie une pression artérielle basse (moins probabilité de diabète)	Minimum : 90 Maximum : 120
Diabete Predigme function	Une fonction qui utilise les informations sur les antécédents familiaux et la relation génétique avec les patients pour évaluer leur risque de diabète. En effet, si la fonction de pedigree est élevée, cela signifie que le patient est plus susceptible de développer un diabète	Minimum : 0.001 Maximum : 2.001
Tour de taille	Une mesure de la circonférence de la région abdominale d'une personne, prise à un niveau horizontal juste au-dessus de l'os de la hanche. un tour de taille de plus de 94 cm chez les hommes et de plus de 80 cm chez les femmes est considéré comme un facteur de risque pour la santé.	Minimum : 50 Maximum : 200
Glycémie à jeun	C'est un taux de glucose sanguin supérieur ou égal à 1.26 g/L qui est considéré comme diagnostique pour le diabète, tandis qu'un taux compris entre 1.00 et 1.25 g/L est généralement considéré comme un prédiabète. Les niveaux de glycémie à jeun compris entre 0.70 et 0.99 g/L sont considérés comme normaux	Minimum : 0.60 Maximum : 3.00
Outcome	Variable de classe (0 et 1) d'où la valeur 0 indique que le patient ne souffre pas de diabète tandis que la valeur 1 indique que le patient est diabétique	0 = non-diabétique 1 = diabétique

4.4.3 Prétraitement des données

Après la collecte des données, l'étape suivante est le prétraitement, qui revêt une grande importance pour extraire un jeu de données de qualité optimale afin d'obtenir des résultats précis. En effet, la plupart des jeux de données peuvent comporter des valeurs manquantes, du bruit ou des incohérences. Ainsi, si la qualité des données n'est pas rigoureusement traitée, il est peu probable d'obtenir des résultats fiables et de haute qualité.

4.4.3.1 La visualisation des données

Il est important de visualiser le jeu de données pour rendre les informations plus compréhensibles et accessibles. Cela peut aider les analystes, les chercheurs et les décideurs à prendre des décisions éclairées basées sur les données. Pour ce faire, on utilise différentes instructions et diagrammes.

Pour afficher les premières lignes de notre tableau de données, nous avons utilisé la méthode **head()** du DataFrame df.

```
# The first 5 observation units of the data set were accessed.
df.head()
```

	age	sexe	poids	taille	IMC	TAsystolique	TAdiastolique	tauxdecholesterol	peedegreefnct	TT	glycemieajeun	outcome
0	29	F	50	1.63	18.82	120.0	60.0	NaN	0.626	81.0	NaN	0
1	41	H	89	1.85	26.00	120.0	70.0	1.09	0.501	96.0	2.01	1
2	42	H	77	1.60	30.08	NaN	NaN	NaN	0.001	102.0	NaN	0
3	48	F	65	1.65	23.88	NaN	NaN	NaN	0.501	91.0	0.90	0
4	48	F	75	1.60	29.30	110.0	70.0	1.43	0.001	107.0	1.70	1

FIGURE 4.3 – L'entête de jeu de données Sarikila

Pour visualiser les variables de notre ensemble de données, nous utilisons la bibliothèque *"pandas profiling"*, qui génère un rapport de profil à partir des données. Ce rapport fournit des informations globales et détaillées sur l'ensemble des données et les variables qu'il contient. Il aide à obtenir une compréhension approfondie des caractéristiques des données, en mettant en évidence les statistiques descriptives, les valeurs manquantes, les corrélations et d'autres aspects importants.

La sortie est enregistrée sous forme de rapport HTML (voir la Figure suivante)

Dataset statistics

Number of variables	12
Number of observations	589
Missing cells	210
Missing cells (%)	3.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	55.3 KiB
Average record size in memory	96.2 B

Variable types

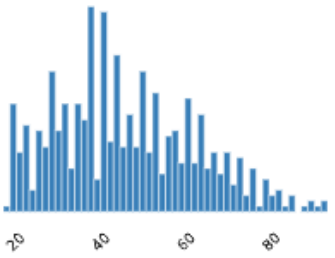
Numeric	10
Categorical	2

FIGURE 4.4 – Rapport HTML de l'ensemble de données

Visualisation des variables

age
Real number (ℝ)

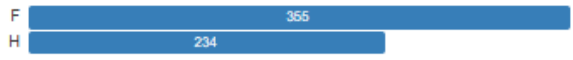
Distinct	72	Minimum	17
Distinct (%)	12.2%	Maximum	93
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	45.865874	Memory size	4.7 KiB



More details

sexe
Categorical

Distinct	2
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	4.7 KiB



More details

FIGURE 4.5 – Visualisation de la variable Age et Sexe

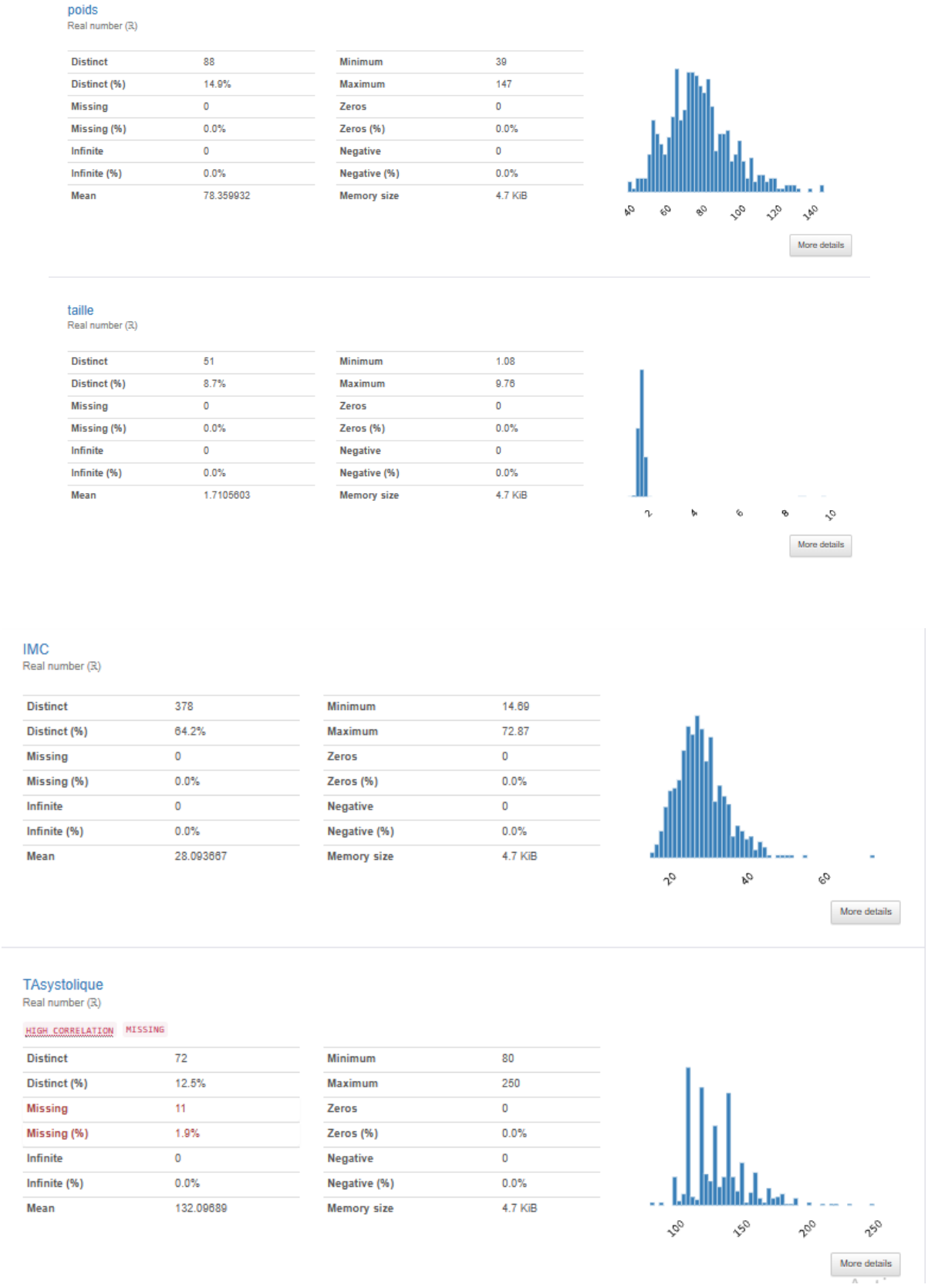


FIGURE 4.6 – Visualisation des variables : Poids, taille, IMC et TASystolique

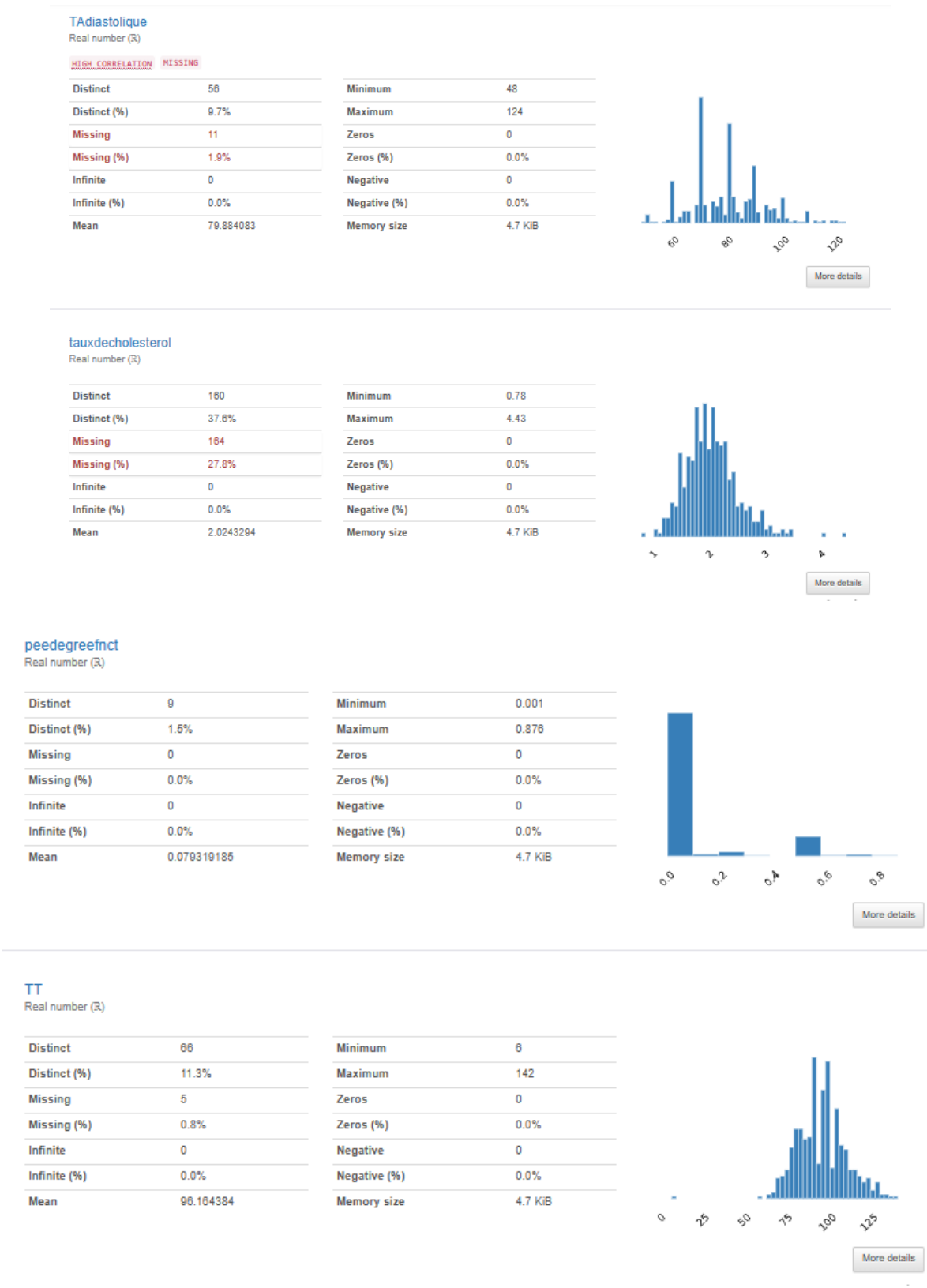


FIGURE 4.7 – Visualisation de variables : TADiastolique, taux de cholésterol, peedegreefnct et Tour de taille.

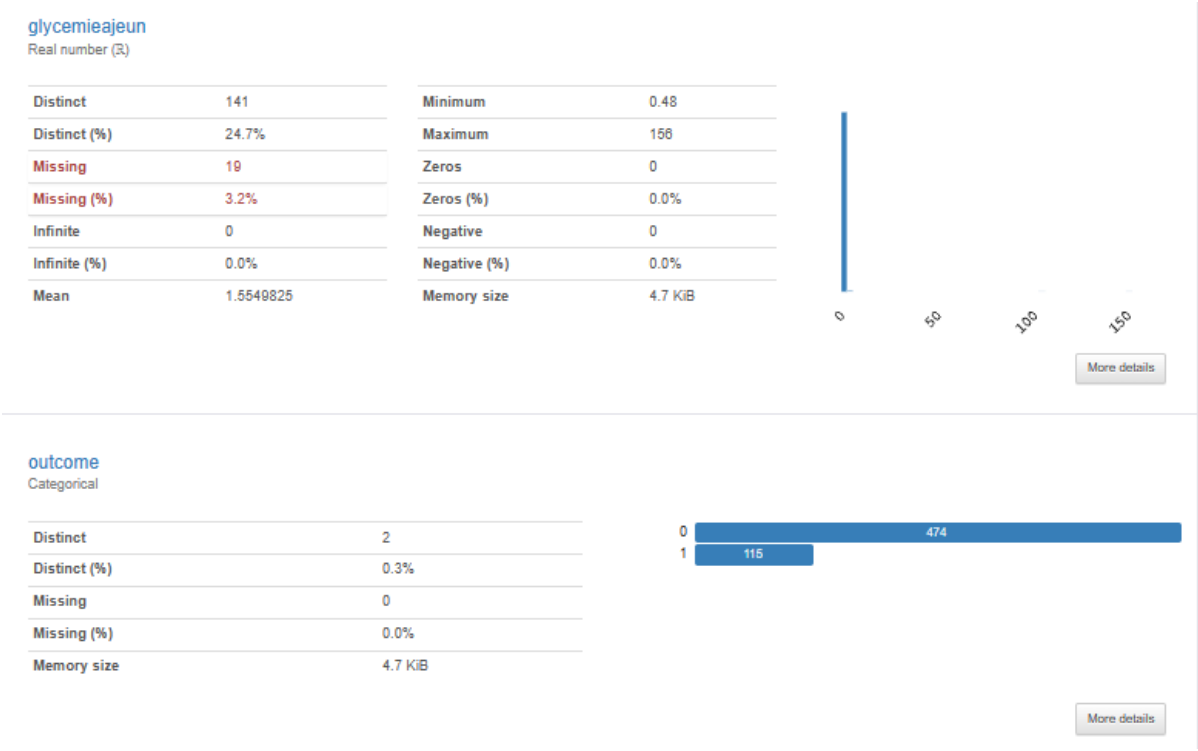


FIGURE 4.8 – Visualisation des variables : Glycémie à jeun et outcome.

Interprétation des figures

On a eu pour chaque variable :

- Le nombre des valeurs distinctes.
- Le pourcentage des valeurs unique.
- Le nombre et les pourcentages (%) des valeurs manquantes.
- La taille.
- La moyenne, minimum et maximum.
- Le nombre et le pourcentage (%) des valeurs nulles (zéro).
- Histogramme : est une représentation graphique qui permet de visualiser la distribution des données dans un dataset. Il est particulièrement utilisé pour les variables quantitatives continues. L'axe horizontal de l'histogramme représente les différentes plages de valeurs possibles pour la variable, tandis que l'axe vertical indique la fréquence ou la densité des observations

On observe que la variable booléenne (outcome) a 474 lignes où la valeur de l'outcome est égale à 0 et 115 lignes où la valeur est égale à 1. On peut également vérifier cela en utilisant le diagramme a cercle et le diagramme a barre.

Nous observons dans le diagramme à cercle un pourcentage de 80,5% de personnes en bonne

santé et un pourcentage de 19,5% de personnes diagnostiquées comme diabétiques. En revanche, le diagramme à barres fournit les valeurs chiffrées correspondantes.

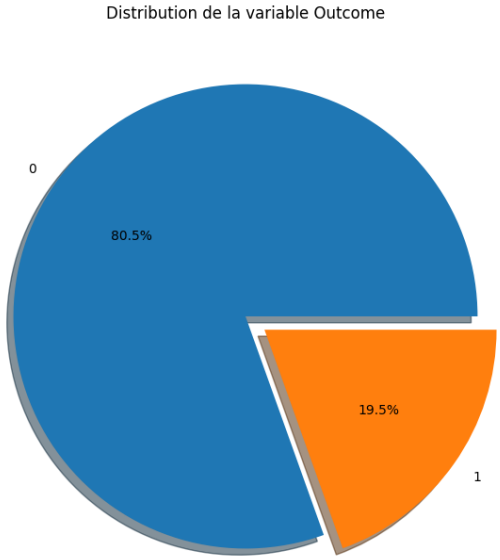


FIGURE 4.9 – Diagramme à cercle de Outcome

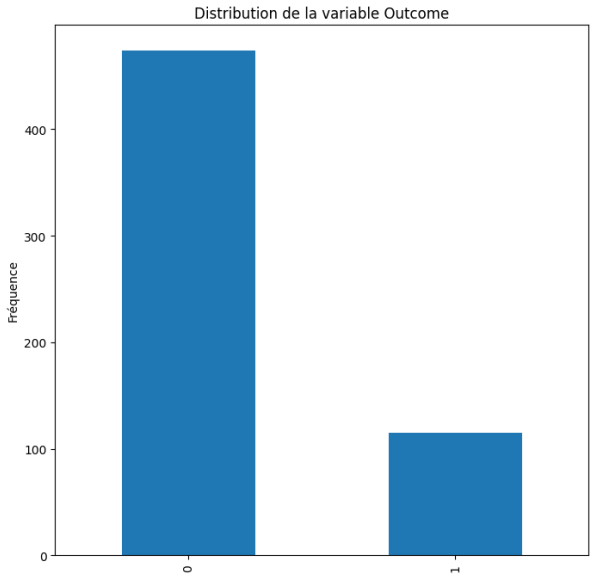


FIGURE 4.10 – Diagramme à barre de Outcome

Corrélation

La corrélation est une mesure de la liaison entre des variables, indiquant le degré de dépendance entre elles et permettant d'évaluer l'ajustement linéaire d'une variable par rapport à une autre. Pour quantifier cette corrélation, on calcule un coefficient de corrélation linéaire.

Le coefficient de corrélation linéaire est un nombre qui mesure la force et la direction de la relation linéaire entre deux variables. Le coefficient le plus couramment utilisé est le coefficient de corrélation de Pearson, noté r. Il est calculé de la manière suivante :

$$r = Cov(X, Y)/(\sigma X * \sigma Y)$$

Où $Cov(X, Y)$ représente la covariance entre les variables X et Y, et σX et σY sont les écarts types respectifs. La valeur du coefficient de corrélation linéaire varie entre -1 et +1 :

— +1 forte relation positive.
— -1 forte relation négative.
— 0 aucune corrélation.

Le tableau suivant montre la corrélation entre les différentes variables de l'ensemble de données :

	age	poids	taille	IMC	TAsystolique	TAdiastolique	tauxdecholesterol	peedegreefnct	TT	glycemieajeun	outcome
age	1.000000	-0.012887	0.008884	0.049656	0.388233	0.092055	0.197408	-0.041669	0.140289	0.093216	0.301994
poids	-0.012887	1.000000	0.015757	0.835765	0.170970	0.210689	0.088477	-0.094944	0.789407	-0.046960	0.112835
taille	0.008884	0.015757	1.000000	-0.098236	-0.154694	-0.112377	-0.083821	0.019793	0.007346	-0.009049	0.082977
IMC	0.049656	0.835765	-0.098236	1.000000	0.190162	0.194570	0.106630	-0.020881	0.719634	-0.030145	0.125097
TAsystolique	0.388233	0.170970	-0.154694	0.190162	1.000000	0.641556	0.214885	-0.181540	0.151468	-0.057182	0.089765
TAdiastolique	0.092055	0.210689	-0.112377	0.194570	0.641556	1.000000	0.206729	-0.223778	0.118765	-0.078857	-0.021475
tauxdecholesterol	0.197408	0.088477	-0.083821	0.106630	0.214885	0.206729	1.000000	-0.199630	0.111901	0.102751	0.066691
peedegreefnct	-0.041669	-0.094944	0.019793	-0.020881	-0.181540	-0.223778	-0.199630	1.000000	0.007898	0.111016	0.148142
TT	0.140289	0.789407	0.007346	0.719634	0.151468	0.118765	0.111901	0.007898	1.000000	-0.018457	0.197143
glycemieajeun	0.093216	-0.046960	-0.009049	-0.030145	-0.057182	-0.078857	0.102751	0.111016	-0.018457	1.000000	0.156800
outcome	0.301994	0.112835	0.082977	0.125097	0.089765	-0.021475	0.066691	0.148142	0.197143	0.156800	1.000000

FIGURE 4.11 – Table de correlation

Un autre outil qui représente la relation entre les variables est la matrice de corrélation ou chaque cellule remplit en couleur en fonction du coeficient de corrélation de la paire qu'elle représente (voir la figure suivante) :

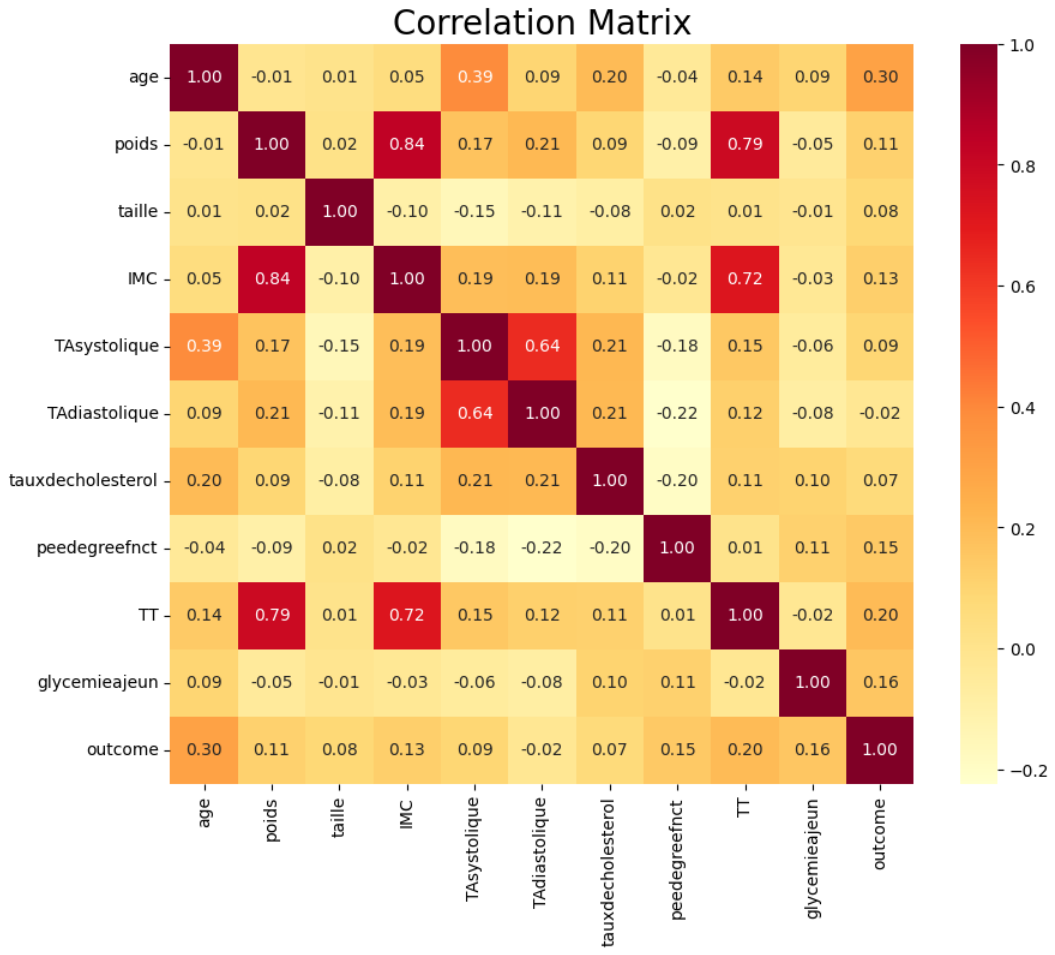


FIGURE 4.12 – Matrice de correlation

4.4.4 Nettoyage du dataset

Le nettoyage des données est une étape essentielle du prétraitement des données qui vise à identifier et à corriger les données altérées, inexactes ou non pertinentes. Cette étape joue un rôle crucial pour améliorer la cohérence, la fiabilité et la valeur des données, ce qui se traduit par des données de meilleure qualité. En conséquence, cela conduit à l'obtention de meilleurs modèles et résultats d'analyse.

D'après le diagramme des valeurs manquantes, il est observé la présence de valeurs manquantes dans certaines catégories ou barres.

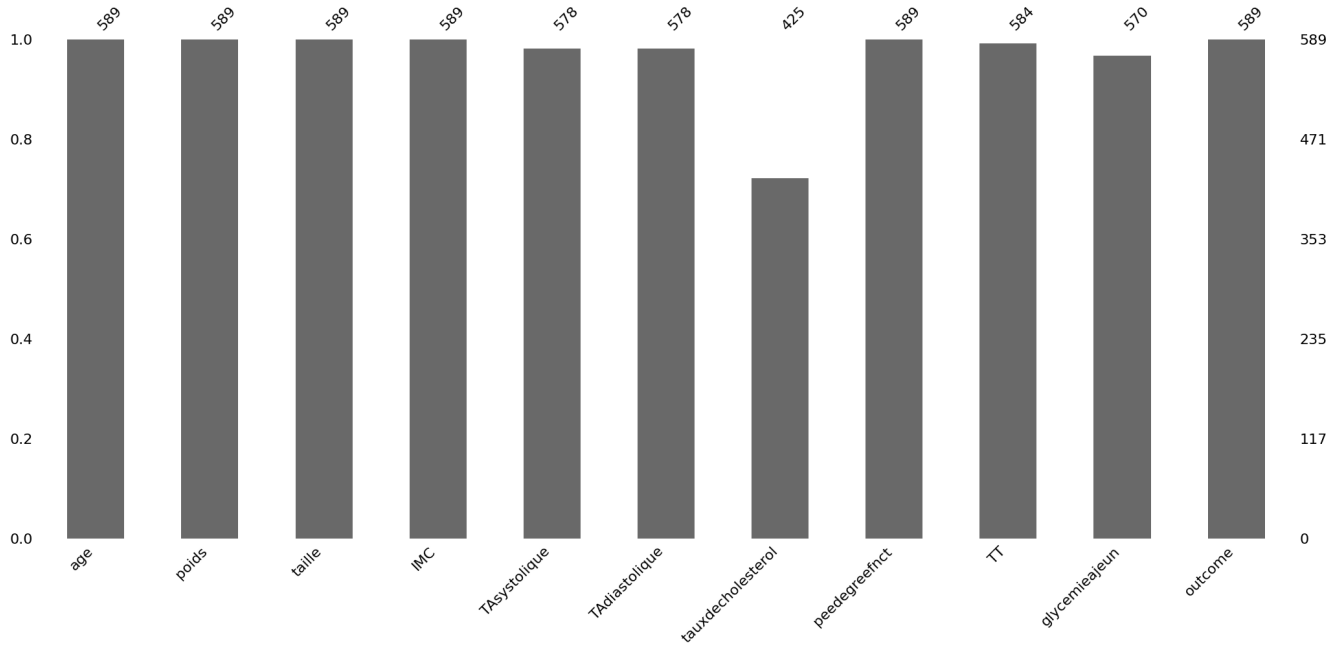


FIGURE 4.13 – Diagramme des valeurs manquantes

4.4.4.1 Élimination des valeurs manquantes

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

```
df.isnull().sum()

age                 0
poids               0
taille              0
IMC                 0
TAsystolique       11
TAdiastolique      11
tauxdecholesterol  164
peedegreefnct       0
TT                  5
glycemieajeun      19
outcome             0
dtype: int64
```

Avant

```
df.isnull().sum()

age                 0
poids               0
taille              0
IMC                 0
TAsystolique        0
TAdiastolique       0
tauxdecholesterol   0
peedegreefnct       0
TT                  0
glycemieajeun       0
outcome             0
dtype: int64
```

Après

FIGURE 4.14 – Les valeurs manquantes avant et après le nettoyage

La meilleure façon de traiter ces valeurs est soit en les supprimant, soit en les remplaçants par des valeurs appropriées (par exemple, la moyenne, la médiane ou la valeur précédente).

Dans notre travail, nous avons choisi de les remplacer par la moyenne. Pour ce faire, nous avons calculé la moyenne pour chaque colonne en prenant en compte les lignes avec un 'outcome' égal à 0 d'une part, et les lignes avec un 'outcome' égal à 1 d'autre part.

```
Moyenne pour outcome = 0:
age                 43.430380
poids               77.362869
taille               1.688671
IMC                 27.690253
TAsystolique       131.109677
TAdiastolique       80.025806
tauxdecholesterol    2.009333
peedegreefnct        0.065884
TT                  94.753715
glycemieajeun        0.933275
dtype: float64
```

FIGURE 4.15 – Le calcul de la moyenne pour outcome = 0

```
Moyenne pour outcome = 1:
age                 55.904348
poids               82.469565
taille               1.800783
IMC                 29.756435
TAsystolique       136.159292
TAdiastolique       79.300885
tauxdecholesterol    2.089000
peedegreefnct        0.134696
TT                 102.044248
glycemieajeun        4.014783
dtype: float64
```

FIGURE 4.16 – Le calcul de la moyenne pour outcome = 1

En utilisant cette méthode, nous avons pu remplacer les valeurs manquantes par les moyennes correspondantes, ce qui nous permet de conserver les informations globales tout en évitant d'introduire des biais dans les données.

4.4.4.2 Suppression des données redondantes

Il est toujours important de vérifier la présence de duplicatas dans un jeu de données afin d'éviter toute distorsion ou redondance indésirable lors de l'analyse. En cas d'existence de lignes identiques dans le seuil est recommandé de supprimer l'une d'entre elles.

Cependant, dans notre dataset, nous n'avons pas identifié de duplicatas. Cela signifie que chaque ligne du jeu de données est unique et qu'il n'y a pas de répétitions ou de doublons.

```
# Compter le nombre de duplicatas
nombre_duplicatas = df.duplicated().sum()

# Afficher le nombre de duplicatas
print("Nombre de duplicatas :", nombre_duplicatas)
```

```
Nombre de duplicatas : 0
```

FIGURE 4.17 – Code pour la suppression des données redondantes

4.4.4.3 Normalisation

La normalisation des variables est une technique largement utilisée en apprentissage automatique pour mettre à l'échelle les variables numériques dans un intervalle spécifique. L'idée générale est de transformer les données afin qu'elles soient comparables et comparables entre elles, ce qui peut être utile dans plusieurs contextes.

Dans notre code, on a effectué la normalisation des variables à l'aide de la classe MinMaxScaler de la bibliothèque scikit-learn (sklearn).

Après l'application de cette méthode, nous avons obtenu un jeu de données normalisé, tel qu'il est illustré dans la figure ci-dessous :

```
df
```

	age	poids	taille	IMC	TAsystolique	TAdiastolique	tauxdecholesterol	peedegreefnct	TT	glycemieajeun	outcome
0	0.315789	0.462963	0.088710	0.194397	0.235294	0.289474	0.084932	0.571429	0.661765	0.009838	1
1	0.328947	0.351852	0.059908	0.264524	0.300645	0.421392	0.336804	0.000000	0.705882	0.002915	0
2	0.407895	0.240741	0.065668	0.157958	0.300645	0.421392	0.336804	0.571429	0.625000	0.002701	0
3	0.407895	0.333333	0.059908	0.251117	0.176471	0.289474	0.178082	0.000000	0.742647	0.007845	1
4	0.434211	0.453704	0.064516	0.309900	0.300645	0.421392	0.115068	0.000000	0.764706	0.002701	0
...	...	...	...	...	...	...	...	...	...	...	...
584	0.868421	0.166667	0.048387	0.182365	0.411765	0.552632	0.408219	0.000000	0.602941	0.003665	0
585	0.881579	0.444444	0.050691	0.393778	0.376471	0.526316	0.405479	0.000000	0.720588	0.014853	1
586	0.947368	0.120370	0.054147	0.118769	0.811765	0.552632	0.610959	0.000000	0.529412	0.002701	0
587	0.973684	0.175926	0.054147	0.161739	0.529412	0.447368	0.421918	0.000000	0.602941	0.008745	1
588	0.986842	0.546296	0.056452	0.429873	0.470588	0.447368	0.238356	0.000000	0.904412	0.002958	0

589 rows × 11 columns

FIGURE 4.18 – Sarikila nettoyer et normaliser

4.4.5 Sélection et entrainement des modèles

4.4.5.1 Train/Test Split

Pour obtenir de bons résultats de prédiction, il est essentiel de former et de tester un modèle de manière adéquate. Cela implique de diviser l'ensemble de données en deux parties distinctes : une partie pour l'entraînement, sur laquelle le modèle apprend, et une partie pour les tests, où l'on évalue les performances des classifieurs sélectionnés. Si un modèle fonctionne bien dans les deux ensembles de données, cela indique une meilleure précision attendue.

Nous utilisons la méthode "train test split" importé de la bibliothèque *sklearn* pour effectuer le fractionnement train/test. "test size=0.2" à l'intérieur de la fonction indique le pourcentage des données qui doivent être conservées pour le test. C'est généralement autour 20% pour le test et le reste de 80% pour l'entrainement.

```
# Splitting X and Y
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.20, random_state=42, stratify=df['outcome'])
```

FIGURE 4.19 – Division de l'ensemble de données

Ce qui signifie 471 observations partie d'entrainement et 118 observations partie test comme le montre le tableau ci-dessous

TABLEAU 4.2 – Aperçu de la division de données

	X	Y
Test	(118, 7)	118
Train	(471, 7)	471

4.4.5.2 Sélection des modèles

La sélection du modèle est une phase cruciale et centrale de l'apprentissage automatique, où l'on choisit le modèle qui fonctionne le mieux pour l'ensemble de données parmi une collection de modèles candidats d'apprentissage automatique.

Les modèles que nous avons choisis pour la prédiction de diabète sont :

1. **K-plus proche voisins**

```
# K nearest neighbors Algorithm
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors = 24, metric = 'minkowski', p = 2)
knn.fit(X_train, Y_train)
```

2. Machine à vecteur de support

```
# Support Vector Classifier Algorithm
from sklearn.svm import SVC
svc = SVC(kernel = 'linear', random_state = 42)
svc.fit(X_train, Y_train)
```

3. Régression Logistique

```
# Logistic Regression Algorithm
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression(random_state = 42)
logreg.fit(X_train, Y_train)
```

4. Arbre de décisions

```
# Decision tree Algorithm
from sklearn.tree import DecisionTreeClassifier
dectree = DecisionTreeClassifier(criterion = 'entropy', random_state = 42)
dectree.fit(X_train, Y_train)
```

5. Foret Aléatoire

```
# Random forest Algorithm
from sklearn.ensemble import RandomForestClassifier
ranfor = RandomForestClassifier(n_estimators = 11, criterion = 'entropy', random_state = 42)
ranfor.fit(X_train, Y_train)
```

6. GBM

```
# Gradient Boosting Algorithm
from sklearn.ensemble import GradientBoostingClassifier
gbm = GradientBoostingClassifier(n_estimators=11, learning_rate=0.1, random_state=42)
gbm.fit(X_train, Y_train)
```

7. xgboost

```
# xgboost Algorithm
from sklearn.ensemble import XGBClassifier
xgboost = xgb.XGBClassifier(n_estimators=100, learning_rate=0.1, random_state=42)
xgboost.fit(X_train, Y_train)
```

4.4.5.3 Les précisions des modèles

Nous avons évalué la précision des modèles dans la méthode d'entraînement/test, nous avons importé la métrique "precision score" de la bibliothèque *"sklearn"*.

```
Logistic Regression: 79.66101694915254
K Nearest neighbors: 80.50847457627118
Support Vector Classifier: 80.50847457627118
Naive Bayes: 87.28813559322035
Decision tree: 94.0677966101695
Random Forest: 93.22033898305084
GBM: 91.52542372881356
XGBoost: 94.91525423728814
```

FIGURE 4.20 – Précisions des modèles

En utilisant la métrique précédente, nous avons ensuite tracé un diagramme à barres correspondant.

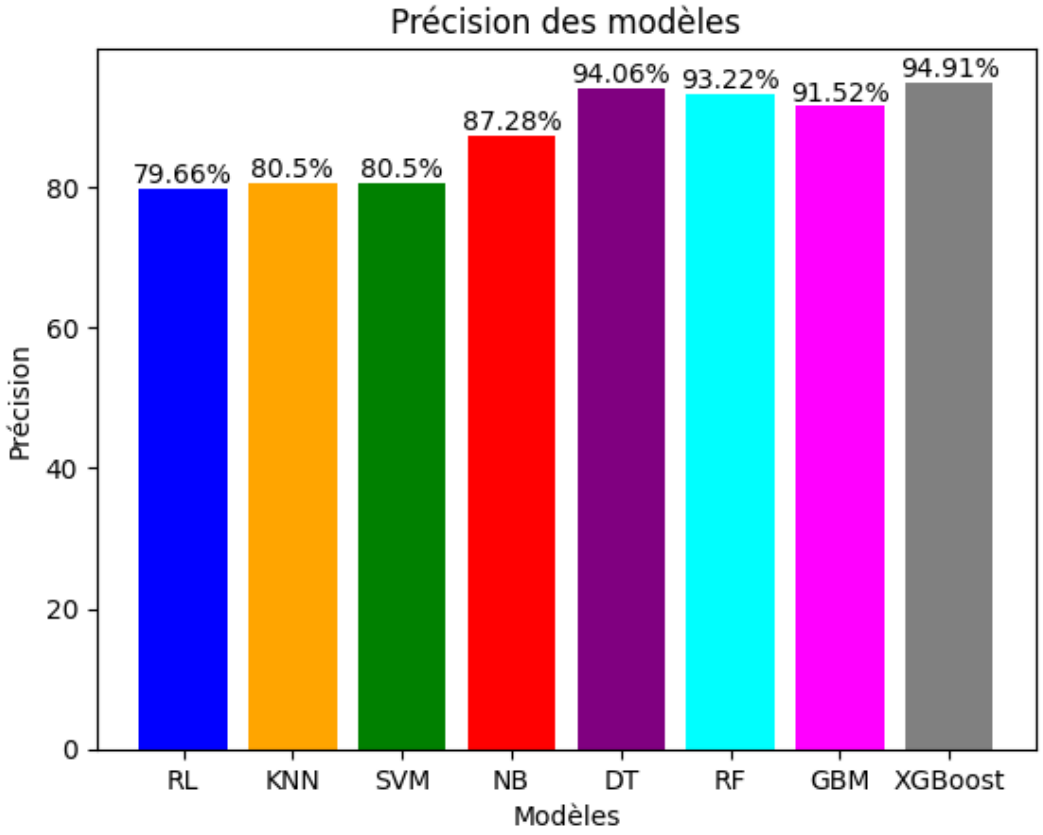


FIGURE 4.21 – Diagramme a barre des précisions des différents algorithmes

4.4.6 Evaluation des modèles

L'évaluation de la performance joue un rôle essentiel dans l'appréciation de la qualité d'un modèle et dans la garantie de la fiabilité des résultats prédictifs. Elle permet de tester l'efficacité du modèle en termes de capacité à fournir des prédictions précises.
Par conséquent, après avoir réalisé des prédictions de diabète, il est essentiel de procéder à une évaluation approfondie afin de mesurer les performances de l'approche choisie. Les résultats du modèle ont été évalués en analysant des critères tels que la précision, recall score, F1 score.

1. **La précision** : est une mesure de la capacité d'un modèle de classification à ne renvoyer que les instances pertinentes. Elle est définie comme le nombre de vrais positifs divisé par la somme des vrais positifs et des faux positifs.

$$Precision = (Vraispositifs)/(Vraispositifs + Fauxpositifs)$$

2. **Recall score** : également appelé sensibilité, mesure la capacité d'un modèle de classification à identifier toutes les instances pertinentes. Il est défini comme le nombre de vrais positifs divisé par la somme des vrais positifs et des faux négatifs.

$$recall = (Vraispositifs)/(Vraispositifs + Fauxnegatifs)$$

3. **F1 score** : Le score F1 fournit une mesure globale de la performance du modèle en tenant compte à la fois de la précision (capacité à renvoyer uniquement les instances pertinentes) et du rappel (capacité à identifier toutes les instances pertinentes). Il est souvent utilisé lorsque l'on souhaite équilibrer l'importance de la précision et du rappel dans l'évaluation d'un modèle de classification.

$$F1score = 2 * (Precision * recallscore)/(Precision + recallscore)$$

Le tableau ci-dessous représente les résultats des attributs d'évaluations pour les différents modèles :

Tableau 4.3 – Les résultats des attributs d'évaluations pour les différents modèles

	Précision	Recall Score	F1 Score
Régression logistique	0.79	0.80	0.71
KNN	0.80	0.72	0.89
SVM	0.81	0.80	0.72
Naïve bayes	0.87	0.39	0.55
Arbre de décision	0.94	0.78	0.84
Foret aléatoire	0.93	0.70	0.80
GBM	0.92	0.60	0.74
XGBoost	0.95	0.83	0.86

D'après le tableau, le modèle XGBoost a obtenu la meilleure précision, égale à 95%, ainsi que le meilleur score de rappel, qui est de 0,83. Cela signifie que sur l'ensemble des patients diabétiques,

83% d'entre eux sont correctement classés à l'aide des mesures de diagnostic médical.

Nous sélectionnons donc le modèle XGBoost comme le modèle le plus optimal et le plus performant pour notre jeu de données en raison de sa grande précision et de son score de rappel élevé. Ces résultats démontrent la capacité du modèle à prédire avec précision les cas positifs et à minimiser les faux négatifs.

4.4.7 Amélioration du modèle sélectionné

4.4.7.1 Importance des caractéristiques

L'importance des variables dans les modèles de machine learning est essentielle pour comprendre comment les caractéristiques d'un dataset contribuent à la prédiction d'un modèle. Cette évaluation permet d'identifier les caractéristiques les plus informatives ou prédictives, ce qui permet de réduire la dimensionnalité du dataset en éliminant les caractéristiques redondantes ou peu pertinentes, pour construire des modèles plus précis, plus interprétables et plus efficaces, tout en réduisant la complexité des données.

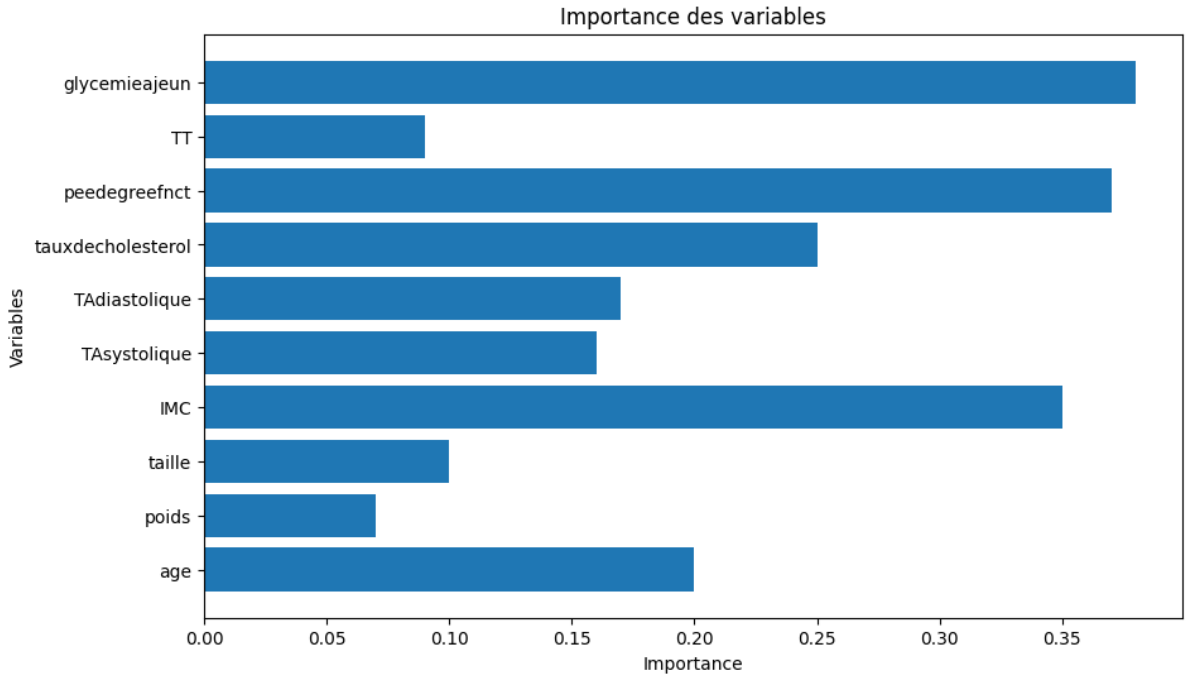


FIGURE 4.22 – Importance des caractéristiques de XGBoost

Cette figure représente l'importance des caractéristiques du dataset utilisées dans la prédiction. On peut observer que la glycémie à jeun, l'IMC et le peedegree jouent un rôle clé dans la détection du diabète, avec la plus haute importance parmi les autres variables. Le taux de cholestérol, l'âge, la tension artérielle systolique et diastolique sont d'une importance moyenne. En revanche, le tour de taille, le poids et la taille ont une importance moindre, ce qui indique qu'ils peuvent être considérés comme des données non pertinentes pour la prédiction du diabète. Il serait donc judicieux de les

supprimer du dataset afin de réduire la dimensionnalité et d'améliorer les performances du modèle XGBoost.

4.4.7.2 Sélection des variables

Nous avons sélectionné les variables qui ont une grande importance pour entraîner notre modèle :

```
# Sélection des variables
variables = ['age','IMC', 'TAdiastolique', 'TAsystolique', 'tauxdecholesterol', 'peedegreefnct', 'glycemieajeun']
X = df[variables].values
Y = df.iloc[:, -1].values  # Dernière colonne
```

FIGURE 4.23 – Sélection des variables

4.4.7.3 La nouvelle précision

Après l'implémentation du modèle XGBoost avec les caractéristiques sélectionnées, nous avons obtenu de meilleurs résultats avec une précision de 97%.

```
l'ancienne précision est : 94.91%
la nouvelle précision est :97.46%
```

FIGURE 4.24 – Nouvelle précision

4.5 Conclusion

En conclusion, après avoir réalisé une implémentation approfondie de notre approche de prédiction du diabète en utilisant des techniques d'apprentissage automatique, nous sommes parvenus à des résultats prometteurs. Parmi les différents algorithmes testés, XGBoost s'est démarqué en termes de performances élevées et de flexibilité, en faisant le meilleur choix pour prédire le diabète.

Ce dernier nous permettra de développer une application web intitulée "*Sarikila Prediction*" qui permet de prédire si une personne donnée est diabétique ou non.

Chapitre 5

Application

5.1 Introduction

Dans ce dernier chapitre, nous allons présenter en détail notre application web, ainsi que les interfaces graphiques que nous avons développées pour clarifier les performances du système. Nous mettrons en évidence les fonctionnalités essentielles et les résultats obtenus grâce à ces interfaces, afin d'offrir une vision globale des activités du système. Enfin, nous conclurons en récapitulant les principales contributions de notre application.

5.2 Outils utilisés

5.2.1 Flask

Flask est un petit framework web Python léger qui fournit des outils et des fonctionnalités utiles facilitant la création d'applications web en Python. Il offre aux développeurs une certaine flexibilité et constitue un cadre plus accessible, permettant de construire rapidement une application web en utilisant un seul fichier Python. Flask est également extensible et n'impose pas une structure de répertoire particulière ni la nécessité d'un code standard compliqué.

5.3 Déscription de l'application

"Sarikila Prediction" est une application web basée sur les prénoms Sara et Akila. Le nom de l'application est une fusion harmonieuse des deux prénoms, créant ainsi une identité unique. Son objectif principal est d'aider les utilisateurs à déterminer s'ils présentent un risque de développer le diabète.

L'application Sarikila Prediction utilise l'algorithme avancé XGBoost basé sur des données médicales et des études scientifiques pour évaluer les facteurs de risque individuels liés au diabète. Elle prend en compte des informations telles que l'âge, le poids, la taille, tour de taille, la glycémie à jeun, la tension artérielle, le taux de cholésterol et les antécédents familiaux diagnostiqués

diabétiques pour fournir une prédiction du risque de diabète.

En se basant sur les données fournies par les utilisateurs, Sarikila Prediction effectue une analyse complète et fournit des résultats clairs et faciles à comprendre. Sarikila Prediction se veut être un outil pratique et accessible pour promouvoir la sensibilisation et la prévention du diabète.

5.4 Démonstration de l'application

Ci-dessous nous fournissons nos interfaces d'application « *Sarikila prediction* » dans le but de permettre aux personnes de savoir s'ils ont le risque de développer un diabète.

5.4.1 La page d'accueil

Cette page fournit des liens hypertexte vers les différentes interfaces qui composent notre application Web.

Voici quelques-unes de ces interfaces :

1. **Prediction**
2. **A propos**
3. **Contact**

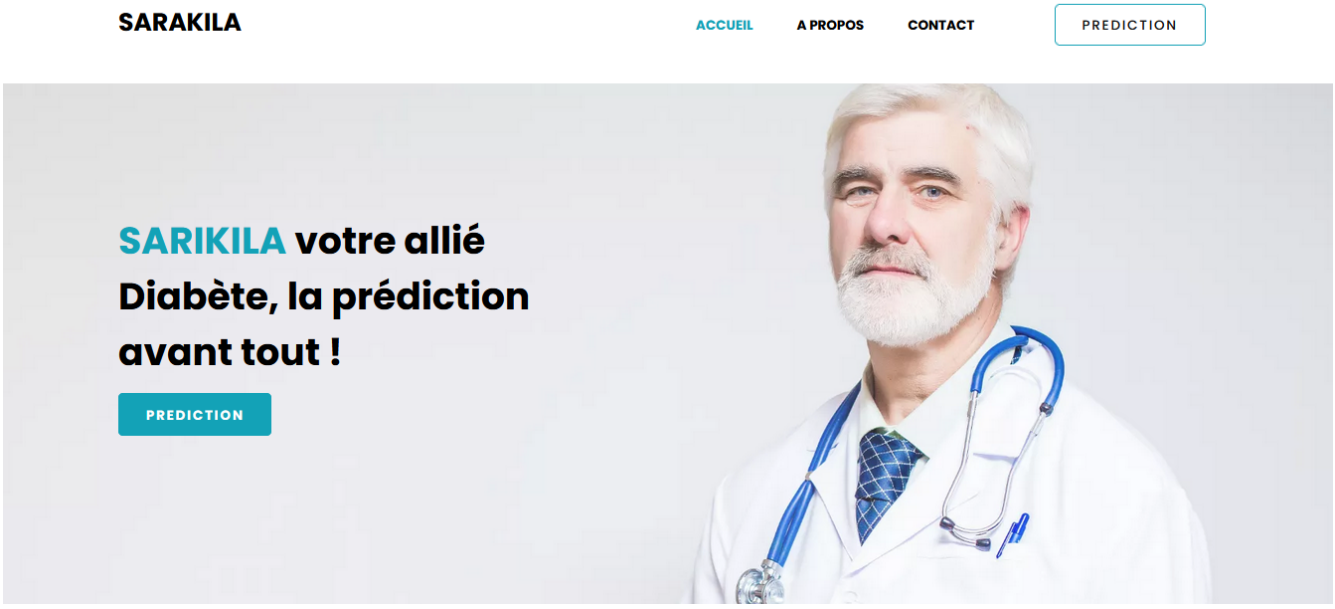


FIGURE 5.1 – Page d'accueil de l'application Sarikila Prediction

5.4.2 Interface de prédiction

Le but de cette interface est de prédire si une personne est diabétique ou non, en fournissant un taux de prédiction. Pour cela, l'utilisateur doit remplir le formulaire ci-dessous, qui contient les informations suivantes : Sexe (Femme, Homme), Age, Poids (en Kg), Taille et tour de taille (en cm), Glycémie à jeun, tension diastolique et systolique, taux de cholestérol ainsi que le nombre d'antécédents familiaux diagnostiqués diabétiques.

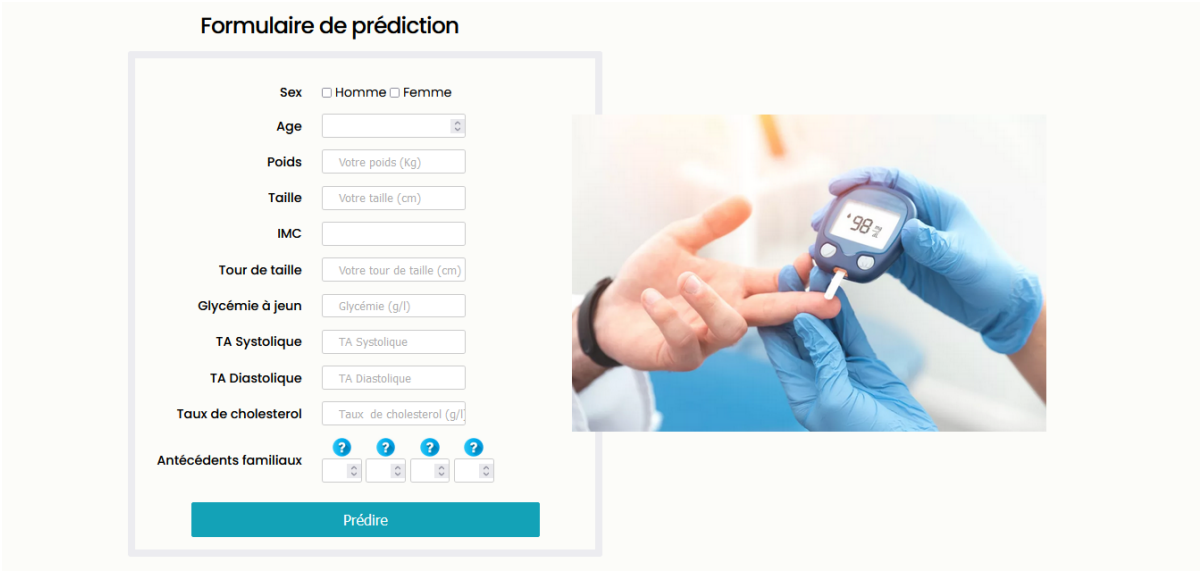


FIGURE 5.2 – Interface de prediction de Sarikila Prediction

Le remplissage de tous les champs du formulaire ci-dessus est obligatoire, sinon un message d'erreur sera affiché (voir la figure suivante)

FIGURE 5.3 – Message d'erreur

Une fois que toutes les informations ont été remplies, l'utilisateur doit cliquer sur le bouton "Prédire" pour que les entrées soient récupérées et utilisées par le modèle XGBoost. Le résultat de la prédiction sera affiché, selon les cas suivants :

1. Cas d'une prédiction non diabétique : L'application affichera un message indiquant que l'utilisateur n'est pas diabétique.

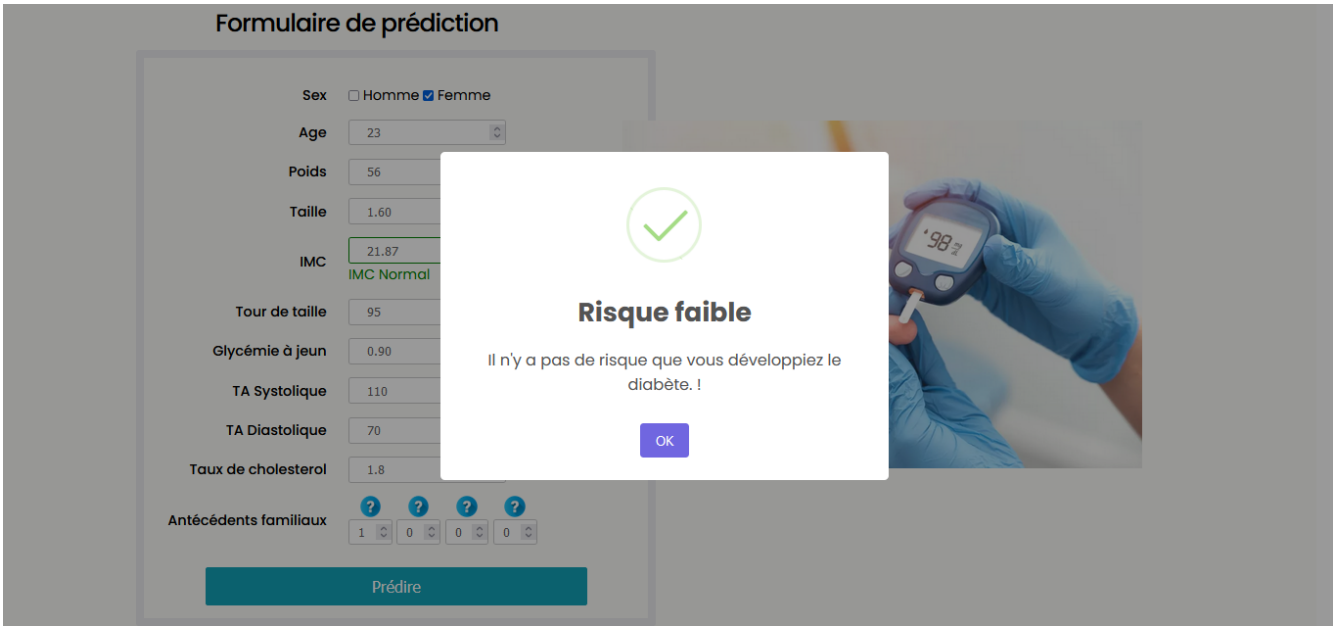


FIGURE 5.4 – Résultat de la prédiction non diabétique

2. Cas d'une prédiction diabétique : Il affiche que l'utilisateur est diabétique

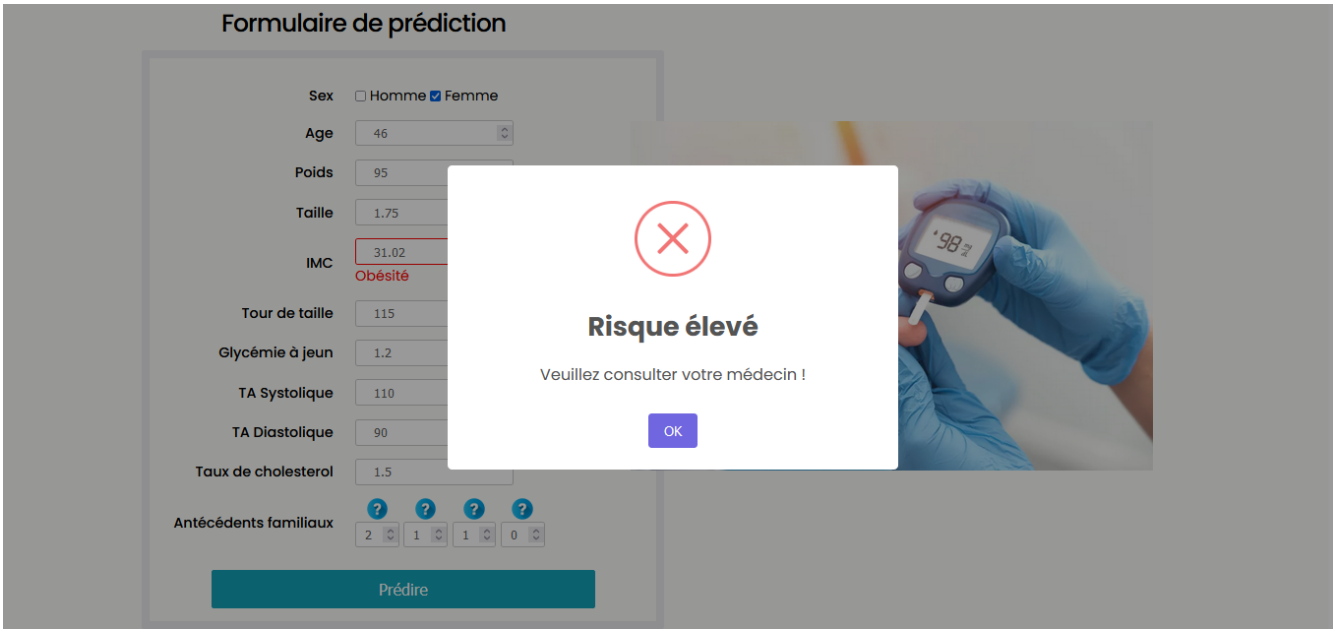


FIGURE 5.5 – Résultat de la prédiction diabétique

5.4.3 A propos diabète

Cette interface offre aux utilisateurs la possibilité d'accéder au *"chapitre 01"* de notre mémoire consacré au diabète. Ce chapitre fournit des informations générales essentielles sur le diabète, permettant ainsi de mieux comprendre cette condition de santé.

Nous croyons en l'importance d'une éducation approfondie sur le diabète, non seulement pour les personnes atteintes de cette condition, mais aussi pour leurs proches et le grand public. Par conséquent, nous avons créé ce chapitre pour offrir un accès facile à des informations claires et concises sur le diabète.

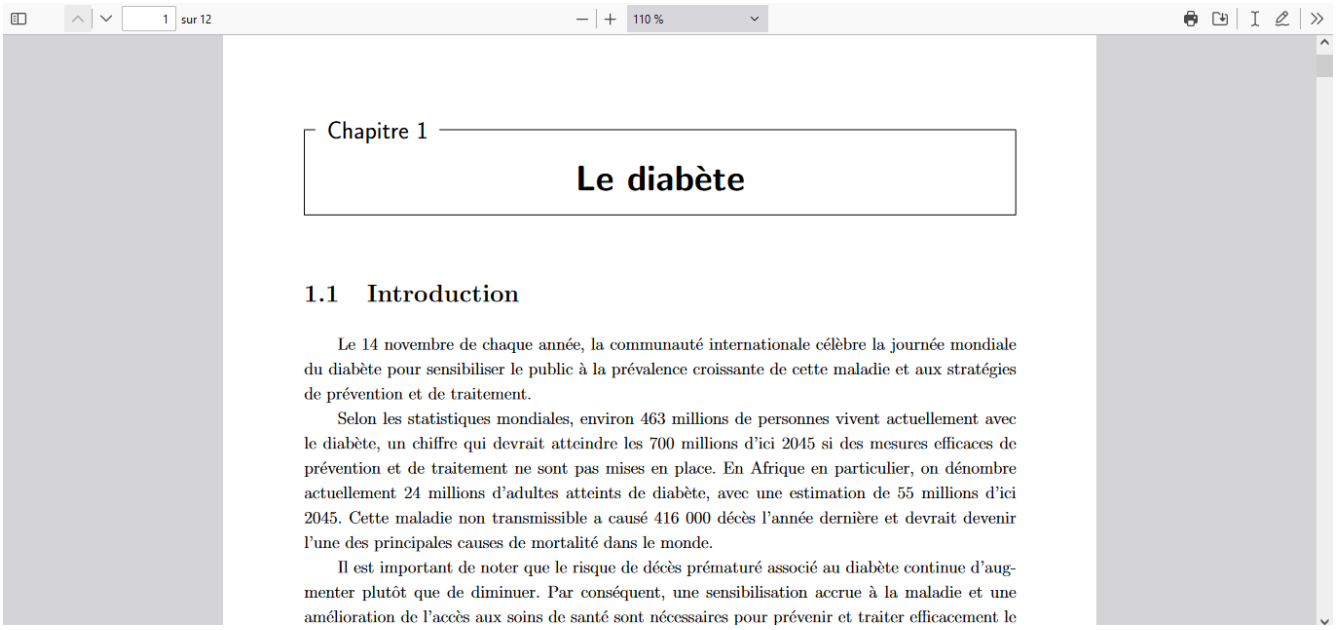


FIGURE 5.6 – Interface à propos le diabète

N'hésitez pas à explorer le *"chapitre 01"* et à revenir vers nous si vous avez des questions supplémentaires ou si vous avez besoin d'informations plus approfondies sur d'autres aspects du diabète.

5.4.4 Section Contact et Map

Cette interface offre aux utilisateurs la possibilité de contacter l'équipe de Sarikila, que ce soit pour poser des questions, faire des suggestions ou demander des renseignements supplémentaires. Nous sommes là pour vous accompagner dans votre parcours lié au diabète.

De plus, cette interface permet également aux utilisateurs de rechercher facilement les diabétologues à Bejaia sur Google Maps.

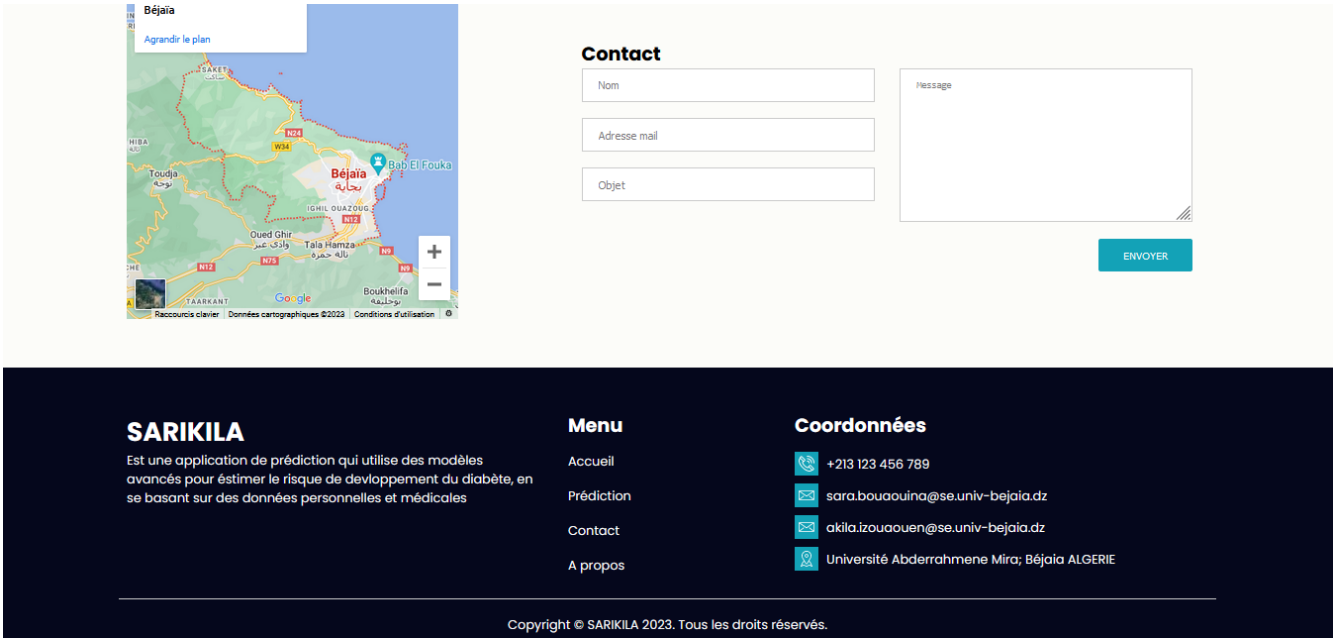


FIGURE 5.7 – Contact et Map

5.5 Conclusion

Dans ce dernier chapitre, nous avons consacré notre attention à présenter en détail notre application web Sarikila Prediction. Cette application innovante offre aux utilisateurs la possibilité de prédire s'ils présentent un risque de développer le diabète. Grâce à ses différentes interfaces intuitives et conviviales, les utilisateurs peuvent facilement accéder aux fonctionnalités de prédiction et obtenir des informations précieuses sur leur santé.

Pour clôturer notre travail, nous souhaitons maintenant présenter une conclusion générale qui résume globalement notre étude.

Conclusion et perspectives

Ce projet a été réalisé dans le cadre de notre mémoire de fin d'études en Master en informatique, option Intelligence artificielle. Son objectif était de développer une méthode basée sur le Machine Learning permettant de prédire de manière précoce la maladie du diabète.

Dans ce mémoire, nous avons mené à faire une comparaison entre huit algorithmes d'apprentissage automatique : la régression logistique, l'arbre de décision, Random Forest, Naïve Bayes, K plus proches voisins, machine à vecteurs de support, GBM et XGBoost. Les résultats expérimentaux obtenus sur notre ensemble de données collectés au sein du CHU Khellil Amrane de Bejaia démontrent que XGBoost est meilleur que les autres en terme de sa grande précision.

Nous avons développé une solution basée sur l'algorithme XGBoost qui permet d'appliquer le modèle à un large public via une application web. Cette solution vise à aider les personnes à prédire s'ils souffrent de diabète. En terme de perspective :

1. Amélioration de l'application : Améliorer les interfaces pour les rendre conviviales et intuitives, et développer une application Android pour la rendre accessible à un large public. Ajouter des fonctionnalités supplémentaires telles que des conseils de prévention, des régimes alimentaires équilibrés adaptés à chaque cas, et des recommandations personnalisées en fonction des résultats de prédiction.

2. Développement des algorithmes de prédiction du diabète en utilisant un ensemble de données vaste et riche en données démographiques et géographiques variées comprenant une diversité de personnes provenant de différentes régions et catégories, cela permettra de prendre en compte les variations et les relations entre les caractéristiques.

3. Évaluation de l'impact clinique : Considération de l'impact potentiel de notre modèle de prédiction dans un contexte clinique réel. C'est à dire on évalue comment le modèle peut être utilisé pour aider les professionnels de la santé dans la détection précoce du diabète, la prise de décision médicale ou la gestion des risques.

Bibliographie

[1] Algorithme svm. https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/. Consulté le 26/02/2023.

[2] Analyse de taux de glycémie. https://sante.journaldesfemmes.fr/fiches-maladies/2729847-prediabete-definition-symptomes-que-faire-taux-glycemie-traitement-fatigue-poi Consulté le 05/03/2023.

[3] Apprentissage non-superviser. https://brightcape.co/apprentissage-supervise-vs-non-supervise/. Consulté le 26/02/2023.

[4] Apprentissage par renforcement. https://datascientest.com. Consulté le 05/03/2023.

[5] Apprentissage superviser. https://machinelearnia.com/apprentissage-supervise-4-etapes/. Consulté le 26/02/2023.

[6] Arbre de décision. https://blent.ai/blog/a/arbres-de-decision-en-machine-learning. Consulté le 26/02/2023.

[7] Arbres de décisions. https://www.jedha.co/formation-ia/arbre-de-decision-random-forest.

[8] Chu. https://www.algerie360.com/cevital-reamenage-gratuitement-le-service-neurochirurgi Consulté le 15/03/2023.

[9] Chu création. https://www.chubejaia.dz/Historique#_CR. Consulté le 15/03/2023.

[10] Complication diabète. https://fr.vecteezy.com/art-vectoriel/298793-complications-du-diabete. Consulté le 05/03/2023.

[11] Comprimés antidiabètiques. https://pillintrip.com/fr/medicine/diamicron. Consulté le 26/02/2023.

[12] Définition. https://www.freestylediabete.fr/le-diabete. Consulté le 15/02/2023.

[13] Fonction segmoid. https://medium.com/@shiny_jay/logistic-regression-a9a8749e1e68. Consulté le 26/02/2023.

[14] Foret aléatoire. https://www.journaldunet.fr/web-tech/guide-de-l-intelligence-artificielle/1501905-random-forest-ou-foret-aleatoire.

[15] Glucomètre. https://www.federationdesdiabetiques.org/federation/actualites/du-pre-diabete-au-diabete-de-type-2-il-est-encore-temps-dagir. Consulté le 28/02/2023.

[16] Historique du diabète. https://www.freestylediabete.fr/le-diabete. Consulté le 15/02/2023.

[17] Injection d'insuline. https://treedir.com/2018/08/05/diabete-type-1-comprimes-dinsuline/. Consulté le 26/02/2023.

[18] Introduction. https://www.afro.who.int/fr/regional-director/speeches-messages/journee-mondiale-du-diabete-2022. Consulté le 15/02/2023.

[19] K-nearest neighbors. https://www.ibm.com/fr-fr/topics/knn. Consulté le 05/03/2023.

[20] Les symptômes du diabète type i. https://www.ameli.fr/assure/sante/themes/diabete/diabete-symptomesevolution/diagnosticdiabete. Consulté le 25/02/2023.

[21] Machine learning. https://en.wikipedia.org/wiki/Journal_of_Machine_Learning_Research?oldid=728817752. Consulté le 05/03/2023.

[22] Medecine interne. https://www.chubejaia.dz/medecineinterne. Consulté le 15/03/2023.

[23] Regression logistique. https://datascientest.com/Largressionlogistique,qu?est-cequec'est?

[24] Régime diabète gestationnel. https://www.passeportsante.net/famille/grossesse?doc=regime-diabete-gestationnel. Consulté le 28/02/2023.

[25] Svm. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm/. Consulté le 26/02/2023.

[26] *Classification pour diagnostic du diabète utilisant les algorithmes d ?apprentissage.* PhD thesis, Université Badji Moukhtar Annaba, 2020/2021.

[27] Un système de prédiction et de prévision du diabète de type 2. Mémoire de master, Université bordj bou arréridj, 2020/2021.

[28] Mohammed Ammar. *Reconnaissance Automatique Du Diabète Et Prédiction de la dose d'insuline.* PhD thesis.

[29] Victor Chang, Meghana Ashok Ganatra, Karl Hall, Lewis Golightly, and Qianwen Ariel Xu. An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. *Healthcare Analytics*, 2 :100118, 2022.

[30] Houda El Bouhissi, Rafa E Al-Qutaish, Amine Ziane, Kamal Amroun, Nabila Yaya, and Melissa Lachi. Towards diabetes mellitus prediction based on machine-learning. In *2023 International Conference on Smart Computing and Application (ICSCA)*, pages 1–6. IEEE, 2023.

[31] Muhammad Exell Febrian, Fransiskus Xaverius Ferdinan, Gustian Paul Sendani, Kristien Margi Suryanigrum, and Rezki Yunanda. Diabetes prediction using supervised machine learning. *Procedia Computer Science*, 216 :21–30, 2023.

[32] Jerome H Friedman. Greedy function approximation : a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[33] S Halimi and J Girard. Traitement du diabète de type 2. où en sommes-nous des voies agissant sur le glucagon ? *Médecine des maladies Métaboliques*, 12(1) :16–21, 2018.

[34] Hady W Lauw, Raymond Chi-Wing Wong, Alexandros Ntoulas, Ee-Peng Lim, See-Kiong Ng, and Sinno Jialin Pan. *Advances in Knowledge Discovery and Data Mining : 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11-14, 2020, Proceedings, Part II*, volume 12085. Springer Nature, 2020.

[35] Mingqi Li, Xiaoyang Fu, and Dongdong Li. Diabetes prediction based on xgboost algorithm. In *IOP conference series : materials science and engineering*, volume 768, page 072093. IOP Publishing, 2020.

[36] Olta Llaha and Amarildo Rista. Prediction and detection of diabetes using machine learning. In *RTA-CSIT*, pages 94–102, 2021.

[37] NasserEddine Mayou and Mohammed Belhachani. *Une application web pour la prédiction précoce du diabète basant sur les algorithmes d'apprentissage automatique*. PhD thesis, UNIVERSITE KASDI MERBAH OUARGLA.

[38] Rachid Mifdal. *Application des techniques d'apprentissage automatique pour la prédiction de la tendance des titres financiers*. PhD thesis, École de technologie supérieure, 2019.

[39] Aishwarya Mujumdar and V Vaidehi. Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, 165 :292–299, 2019.

[40] Chollette C Olisah, Lyndon Smith, and Melvyn Smith. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, 220 :106773, 2022.

[41] MEIGS PARENT.

[42] Anju Prabha, Jyoti Yadav, Asha Rani, and Vijander Singh. Design of intelligent diabetes mellitus detection system using hybrid feature selection based xgboost classifier. *Computers in Biology and Medicine*, 136 :104664, 2021.

[43] Aghila Rajagopal, Sudan Jha, Ramachandran Alagarsamy, Shio Gai Quek, and Ganeshsree Selvachandran. A novel hybrid machine learning framework for the prediction of diabetes with context-customized regularization and prediction procedures. *Mathematics and Computers in Simulation*, 198 :388–406, 2022.

[44] R Ranjitha, V Agalya, and K Archana. Diabetes prediction by artificial neural network. In *Inventive Communication and Computational Technologies : Proceedings of ICICCT 2021*, pages 1011–1019. Springer, 2022.

[45] Rashi Rastogi and Mamta Bansal. Diabetes prediction model using data mining techniques. *Measurement : Sensors*, 25 :100605, 2023.

[46] Jitranjan Sahoo, Manoranjan Dash, and Abhilash Pati. Diabetes prediction using machine learning classification algorithms. *International Research Journal of Engineering and Technology*, 7(8), 2020.

[47] André SCHEEN and Nicolas Paquot. Le diabete de type 2 : voyage au coeur d'une maladie complexe. *Revue Médicale de Liège*, 67(5-6), 2012.

[48] Amel Sidahmed and Karima Rabhi. *La prédiction du diabete en utilisant les algorithmes de machine learning*. PhD thesis, Université Akli Mohand Oulhadje-Bouira, 2020.

[49] Liyang Wang, Xiaoya Wang, Angxuan Chen, Xian Jin, and Huilian Che. Prediction of type 2 diabetes risk and its effect evaluation based on the xgboost model. In *Healthcare*, volume 8, page 247. MDPI, 2020.

[50] Zhongxian Xu and Zhiliang Wang. A risk prediction model for type 2 diabetes based on weighted feature selection of random forest and xgboost ensemble classifier. In *2019 eleventh international conference on advanced computational intelligence (ICACI)*, pages 278–283. IEEE, 2019.

[51] Tafa Zhilbert, N Pervetica, and B Karahoda. An intelligent system for diabetes prediction. In *Proceedings of the4th Mediterranean Conference on Embedded Computing (MECO)*, 2015.

RÉSUMÉ

Le diabète est une maladie chronique caractérisée par un dysfonctionnement du métabolisme du glucose dans le corps, résultant soit d'une production insuffisante d'insuline par le pancréas, soit d'une résistance des cellules à l'insuline produite. L'intelligence artificielle (IA) a connu une évolution majeure dans le domaine médical, offrant de nouvelles opportunités pour la compréhension, le diagnostic précis et le traitement amélioré de cette maladies. Divers modèles d'apprentissage automatique tels que KNN, SVM, arbre de décision, forêt aléatoire, GBM et XGBoost ont été appliqués sur un jeu de données de 590 entrées collecter lors d'un stage au CHU Khellil Amrane à Bejaïa, comprenant des informations sur les personnes diabétiques et non diabétiques telles que l'âge, le poids, la taille, les antécédents personnels et familiaux, le taux de cholestérol, etc. Après le nettoyage et le traitement des données, ainsi que l'entraînement et l'amélioration du modèle XGBoost, un résultat prometteur avec une précision de 97% a été obtenu pour notre système de prédiction. Ces résultats ont conduit à la création d'une application web innovante permettant aux individus de prédire leur risque de développer le diabète, offrant ainsi une approche préventive en fournissant des informations clés sur la santé actuelle et en encourageant la prise de mesures proactives pour réduire les risques associés au diabète.

Mots clés : Diabète ; Insuline ; Machine learning ; Xgboost ; Prédiction ; Dataset.

ABSTRACT

Diabetes is a chronic disease characterized by a malfunction in the body's glucose metabolism, resulting from either insufficient insulin production by the pancreas or cells' resistance to produced insulin. Artificial intelligence (AI) has undergone significant advancements in the medical field, providing new opportunities for understanding, accurate diagnosis, and improved treatment of this disease. Various machine learning models such as KNN, SVM, decision trees, random forest, GBM, and XGBoost were applied to a dataset of 590 entries collected during an internship at CHU Khellil Amrane in Bejaïa, which included information about both diabetic and non-diabetic individuals, such as age, weight, height, personal and family history, cholesterol levels, etc. After data cleaning and processing, as well as training and enhancing the XGBoost model, a promising result with 97% prediction accuracy was achieved for our prediction system. These results led to the development of an innovative web application that enables individuals to predict their risk of developing diabetes, offering a preventive approach by providing key information about their current health status and encouraging proactive measures to reduce diabetes-related risks.

Key words : Diabetes ; Insulin ; Machine learning ; XGBoost ; Prediction ; Dataset.