



République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Cheikh Larbi Tebessi- Tébessa

Faculté des Sciences Exactes et Sciences de la Nature et de la Vie

Département des Mathématiques et de l'Informatique

Mémoire de Master

Domaine : Mathématiques et Informatique

Filière : Informatique

Option : Système d'Information (SI)

Thème

Un Système pour la prédiction du Diabète

Présenté par : Mesbahi Salsabil

Dirigé par : Dr. Ghrieb Nawel

Devant le jury :

Président : Dr. Menassel Yahia

Superviseur : Dr. Ghrieb Nawel

Examineur : Dr. Tag Samir

Année Universitaire : 2023/2024

Remerciements

Je tiens à remercier Allah, le Tout-Puissant, pour m'avoir accordé la subsistance, la guidance, le savoir et l'opportunité de participer à cette recherche.

Mes sincères remerciements et ma gratitude vont à **Dr. Ghrieb Nawel**, qui a été une source d'aide et de soutien tout au long de mon travail.

Je suis également immensément reconnaissant envers **Dr. Tag Samir** et **Dr. Menassel Yahia** pour avoir aimablement accepté d'apporter à ce travail leurs précieux commentaires.

Dédicaces

Au nom de Dieu, le Plus Miséricordieux, le Plus Compatissant.

Ce travail est dédié :

À mes parents, en particulier à ma mère,

À ma chère sœur,

À ma famille,

À mes amis.

Résumé

Un système de prédiction du diabète est un outil informatique conçu pour évaluer le risque qu'une personne développe le diabète en se basant sur plusieurs critères médicaux. Ces prédictions permettent aux professionnels de la santé d'identifier les individus à risque et de mettre en place des mesures préventives appropriées. Ce système repose sur des techniques d'apprentissage automatique, contribuant ainsi à améliorer la détection précoce du diabète et à réduire les complications liées à cette maladie chronique. Dans cette étude, nous avons utilisé les modèles d'apprentissage automatique Random Forest, DecisionTree, LogisticRegression, KNN et Gradient Boosting. Parmi ces modèles, Random Forest s'est révélé être le plus performant, avec une précision de 94 % et une exactitude de 87 %.

Mots-Clés : Apprentissage Automatique, Diabète, Prédiction.

Abstract

A diabetes prediction system is a computer tool designed to predict the risk of a person developing diabetes based on several medical criteria. These predictions assist healthcare professionals in identifying individuals at high risk of diabetes so that they can implement appropriate preventive measures. This system is based on machine learning and thus contributes to improving the early management of diabetes and reducing complications associated with this chronic disease. In this study, we used the following machine learning models: Random Forest, DecisionTree, LogisticRegression, KNN, and Gradient Boosting. We found that Random Forest is the best model with an accuracy of 87% and a precision of 94%.

Keywords: Machine Learning, Diabetes, Prediction.

ملخص

نظام تنبؤ بالسكري هو أداة حاسوبية مصممة لتنبؤ بمخاطر تطوير شخص ما للسكري من خلال بعض المعايير الطبية. تساعد هذه التنبؤات المحترفين الصحيين في تحديد الأفراد ذوي المخاطر العالية للإصابة بالسكري حتى يتمكنوا من اتخاذ التدابير الوقائية المناسبة. يقوم هذا النظام على التعلم الآلي وبالتالي يساهم في تحسين الرعاية المبكرة للسكري وتقليل المضاعفات المرتبطة بهذا المرض المزمن. في هذه الدراسة، استخدمنا النماذج التالية للتعلم الآلي: Random Forest، DecisionTree، LogisticRegression، KNN. وهو أفضل النماذج بدقة تبلغ 94% ودقة بنسبة 87% لـ Random Forest وجدنا أن Gradient Boosting.

الكلمات الرئيسية: التعلم الآلي، السكري، التنبؤ

Table des matières

Introduction Générale.....	1
Chapitre 1 : Le Diabète.....	3
1. Introduction.....	4
2. L'épidémiologie du diabète	4
2.1 Prévalence du diabète sucré dans le monde	4
2.2 Prévalence du diabète sucré dans l'Algérie	5
3. La maladie du diabète.....	6
4. Types du diabète.....	6
4.1 Diabète de type 1 (DT1).....	6
4.2 Diabète de type 2 (DT2).....	7
4.3 Diabète Gestationnel	7
5. Symptômes du diabète.....	7
6. Complications du diabète.....	7
6.1 Maladies cardiovasculaires.....	8
6.2 Néphropathie	8
6.3 Troubles oculaires.....	8
6.4 Neuropathie	8
6.5 Sensibilité aux infections	9
7. Facteurs de risque.....	9
8. Causes du diabète	9
9. Diagnostic du diabète	10
10. Traitement du diabète	10
11. Réduction des risques du diabète.....	10
12. Conclusion	11
Chapitre 2 : L'Apprentissage Automatique	12

1.	Introduction.....	13
2.	Définition	13
3.	Types d'apprentissage automatique.....	13
3.1	Apprentissage supervisé.....	13
3.2	Apprentissage non supervisé	14
3.3	Apprentissage par renforcement	14
4.	Algorithmes de l'apprentissage automatique	15
4.1	DecisionTree (Arbre de Décision)	15
4.2	Random Forest (Forêt Aléatoire).....	16
4.3	KNeighbors (KNN).....	17
4.4	K-means	19
5.	Travaux de recherche sur la prédiction du diabète par les méthodes d'apprentissage automatique	20
6.	Synthèse.....	23
7.	Conclusion	24
	Chapitre 3: Description du Système Développé et Analyse des Résultats	25
1.	Introduction.....	26
2.	Outils et bibliothèques utilisées.....	26
2.1	Anaconda.....	26
2.2	Jupyter.....	27
2.3	Python	28
2.4	Modules de développement.....	28
2.4.1	Pandas	28
2.4.2	Numpy.....	29
2.4.3	Scikit-learn	30
2.4.4	Matplotlib	30

2.4.5	Seaborn.....	31
2.4.6	Tkinter.....	31
2.4.7	Joblib.....	31
3.	Plan de développement du système.....	32
4.	Le système de prédiction du diabète proposé.....	33
4.1	Description du Dataset.....	33
4.2	Exploration et Visualisation des données.....	34
4.3	Corrélation des données.....	35
4.4	Prétraitement des données.....	36
4.5	Répartition des données.....	38
4.6	Entraînement des modèles.....	39
4.7	Evaluation des modèles.....	39
4.8	Sélection du modèle.....	43
5.	Interfaces du système développé.....	44
6.	Conclusion.....	46
	Conclusion générale.....	48
	Références bibliographiques.....	49

Table des figures :

Figure 1.1: Nombre de personnes atteintes de diabète dans le monde et par région de la FID en 2021-2045 (20-79 ans) [1].....	5
Figure 1.2: Diagramme de la physiopathologie du diabète [4].....	6
Figure 1.3: Diagramme des complications du diabète [7].....	8
Figure 2.1: Exemple de l'apprentissage supervisé [12].	14
Figure 2.2: Exemple de l'apprentissage non supervisé [12].....	14
Figure 2.3: Exemple de l'apprentissage par renforcement [12].....	15
Figure 2.4: Structure d'un arbre de décision [13].	16
Figure 2.5: Vote majoritaire des arbres de décisions pour le Random Forest [11].....	17
Figure 2.6 Classification par la méthode KNN [11].	18
Figure 2.7: Exemple sur K-means [14].	20
Figure 3.1: Logo d'Anaconda [26].	26
Figure 3.2: Interface d'Anaconda [25].....	27
Figure 3.3: Logo de Jupyter [27].....	28
Figure 3.4: Logo de Python [29].....	28
Figure 3.5: Logo de Pandas [32].....	29
Figure 3.6: Logo de Numpy [32].	29
Figure 3.7: Logo de Scikit-learn [33].....	30
Figure 3.8: Logo de Matplotlib [35].....	30
Figure 3.9: Logo de Seaborn [37].	31
Figure 3.10: Logo de Joblib [39].....	31

Figure 3.11: Plan de développement du système de prédiction.	32
Figure 3.12: Dataset PIMA.	34
Figure 3.13: Répartition des données.	34
Figure 3.14: Distribution des variables.	35
Figure 3.15: Matrice de Corrélation.	36
Figure 3.16: Visualisation des valeurs nuls.	37
Figure 3.17 : Code de nettoyage des valeurs nulles.	37
Figure 3.18 : Dataset après le nettoyage des données.	38
Figure 3.19: Répartition des données après l'équilibrage du dataset.	38
Figure 3.20: Aperçu de la répartition des données.	39
Figure 3.21 Entraînement des modèles.	39
Figure 3.22: Evaluation de modèle de Random Forest.	41
Figure 3.23: Evaluation de modèle de DecisionTree.	42
Figure 3.24: Evaluation de modèle de KNN.	42
Figure 3.25: Evaluation de modèle de Logistic Regression	43
Figure 3.26: Evaluation de modèle de Gradient Boosting.	43
Figure 3.27: Interface du système.	45
Figure 3.28: Résultat de prédiction d'une personne diabétique.	45
Figure 3.29: Résultat de prédiction d'une personne non diabétique.	46

Liste des tableaux:

Tableau 2.1: Travaux connexes.....	23
Tableau 3.1: Description des variables du dataset.	33
Tableau 3.2: Métriques de performance des modèles.	44

Introduction Générale

Introduction Générale

Le diabète, une maladie complexe, constitue un problème de santé majeur à l'échelle mondiale. Sa prévalence ne cesse de croître, alimentée par des facteurs, tels que le vieillissement de la population, l'urbanisation, les modes de vie sédentaires et les régimes alimentaires déséquilibrés.

Cette augmentation exponentielle engendre des coûts sociaux et économiques considérables, tout en impactant la qualité de vie des personnes touchées. Face à cette épidémie, la prévention et la gestion efficace du diabète deviennent des priorités cruciales. Les systèmes de prédiction du diabète émergent comme des outils prometteurs pour identifier précocement les individus à risque de développer la maladie. Ces systèmes utilisent une variété de données pour évaluer la probabilité qu'un individu développe le diabète dans un avenir proche.

Le rôle de l'Intelligence Artificielle (IA) dans ce contexte est crucial. Les techniques d'apprentissage automatique et de modélisation statistique permettent d'analyser de vastes ensembles de données et d'identifier des schémas complexes et des relations non linéaires entre les facteurs de risque et le développement du diabète. En utilisant ces informations, les systèmes de prédiction peuvent générer des modèles personnalisés qui aident les professionnels de la santé à prendre des décisions éclairées sur la prévention et la gestion du diabète chez les individus.

L'IA offre également la possibilité d'améliorer continuellement les performances des systèmes de prédiction en intégrant de nouvelles données et en ajustant les algorithmes en fonction des retours d'expérience. Cela permet une adaptation dynamique aux changements dans les caractéristiques de la population et dans les connaissances médicales, améliorant ainsi l'efficacité et la précision des prédictions.

Pour sauver des vies grâce à la détection précoce du diabète, il est crucial de savoir si une personne est diabétique ou non. Dans cette optique, notre objectif principal consiste à créer un système robuste et efficace pour la prédiction du diabète. Ce système intégrera des techniques avancées d'apprentissage automatique pour permettre une détection précoce, une intervention préventive ciblée, ainsi qu'une gestion personnalisée de la maladie.

Introduction générale

Ce travail est structuré en trois chapitres, comme suit :

Dans le premier chapitre, nous avons présenté des informations sur le diabète : l'épidémiologie du diabète, ses types, ses symptômes, ses complications, ses facteurs de risque, ses causes, son diagnostic, son traitement, ainsi que la réduction du diabète.

Dans le second chapitre nous abordons les différents concepts de l'apprentissage automatique et ses types et nous présentés les différents algorithmes d'apprentissage automatique utilisés pour détecter l'apparition précoce du diabète.

Enfin, le dernier chapitre est consacré à la description des outils et des modules utilisés pour le développement du système. Ce chapitre présente également, les différentes phases de réalisation du système ses interfaces, ainsi que les résultats obtenus.

Chapitre 1

Le Diabète

1. Introduction

Le diabète est parmi les maladies les plus répandues à travers le monde. Actuellement on estime à 150 millions le nombre de personnes atteintes de diabète dans le monde. Malgré les efforts de recherche entrepris depuis plusieurs décennies, cette maladie ne bénéficie encore que de traitements substitutifs aux contraintes quotidiennes.

Dans ce chapitre nous présentons l'épidémiologie du diabète, la maladie du diabète sucré, ses types, les différentes complications du diabète ainsi que les méthodes de prévention contre les complications du diabète et les facteurs de risque associés.

2. L'épidémiologie du diabète

En 2014, le nombre de personnes atteintes de diabète dans le monde s'élevait à 422 millions, comparé à seulement 108 millions en 1980. En 1990, les premières prévisions de l'Organisation Mondiale de la Santé (OMS) et de la Fédération Internationale du Diabète (IDF) craignaient que ce nombre atteigne 240 millions d'ici 2025.

2.1 Prévalence du diabète sucré dans le monde

En 2019, le nombre de personnes atteintes de diabète dans le monde a dépassé les 463 millions, dont 59 millions en Europe. En 2021, ce nombre a encore augmenté pour atteindre plus de 537 millions à l'échelle mondiale, soit environ une personne sur dix, incluant 61 millions en Europe. De plus, en 2021, le diabète a entraîné le décès de 6,7 millions de personnes, marquant une augmentation de 2,5 millions par rapport à 2019 (4,2 millions de décès). En 2021, 81 % des adultes diabétiques vivaient dans des pays à revenu faible ou intermédiaire, contre 79 % en 2019.

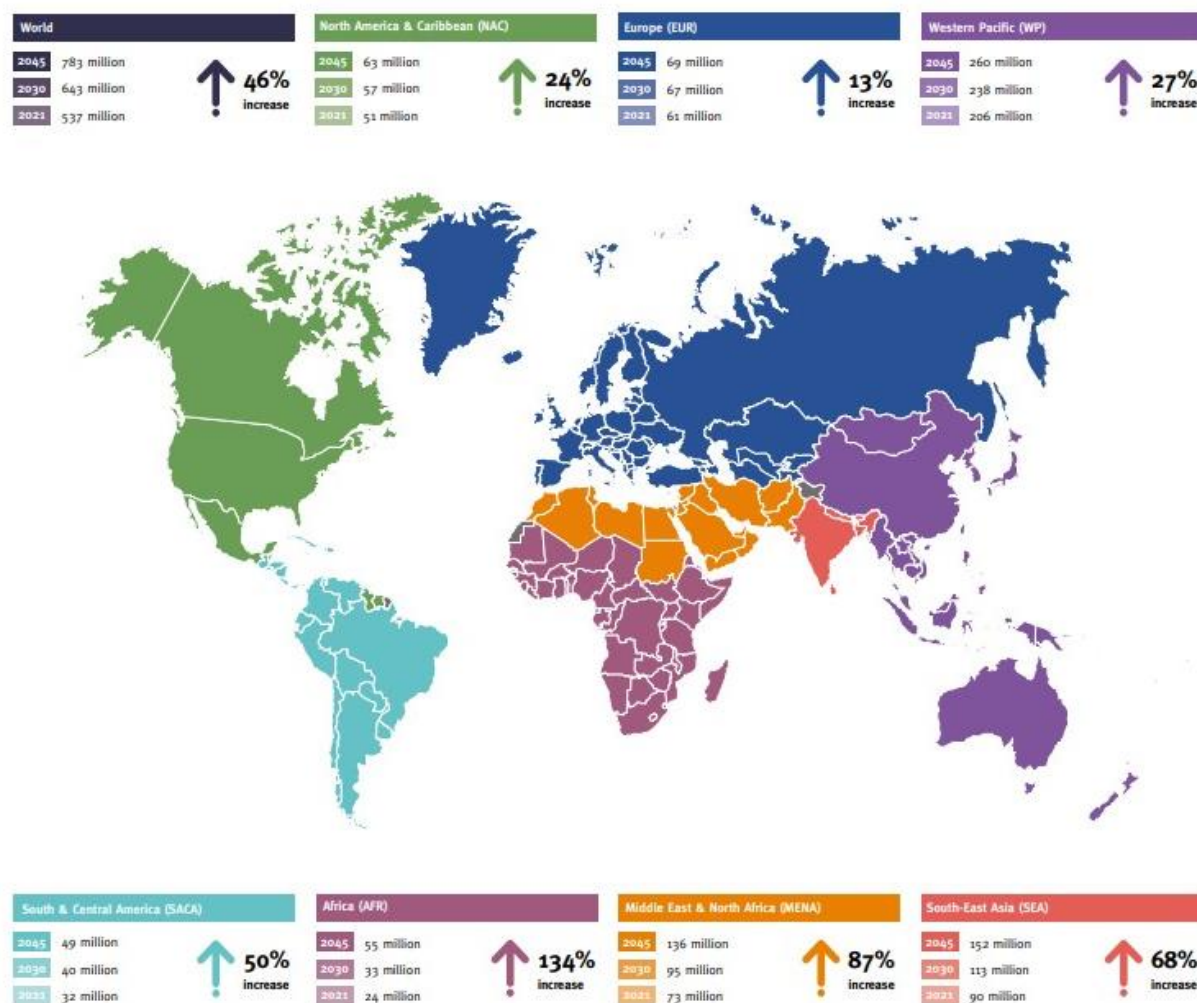


Figure 2.1: Nombre de personnes atteintes de diabète dans le monde et par région de la FID en 2021-2045 (20-79 ans) [1].

2.2 Prévalence du diabète sucré dans l'Algérie

En Algérie, le diabète est la deuxième maladie chronique la plus répandue après l'hypertension. En 2018, environ 14,4 % de la population âgée de 18 à 69 ans, soit environ 4 millions de personnes, étaient estimés être atteints de diabète. Le diabète de type 2 représente environ 90 % des cas diagnostiqués en Algérie, avec des facteurs de risque tels qu'une alimentation peu saine, un manque d'activité physique et une prédisposition génétique. Sensibiliser et gérer efficacement le diabète en Algérie sont des enjeux cruciaux pour prévenir et traiter efficacement cette maladie, afin d'améliorer la santé et le bien-être de la population [2].

3. La maladie du diabète

Le diabète sucré résulte d'une diminution de la production d'insuline, combinée à une résistance variable des tissus périphériques à cette hormone, ce qui entraîne une augmentation du taux de sucre dans le sang (hyperglycémie). Les premiers symptômes incluent une soif excessive, une augmentation de la production d'urine et des problèmes de vision floue. Les complications à long terme peuvent inclure des problèmes vasculaires, des troubles nerveux périphériques, des affections rénales et une susceptibilité accrue aux infections. Le diagnostic repose sur la mesure du taux de sucre dans le sang. Le traitement implique généralement un régime alimentaire spécifique, de l'exercice régulier et l'utilisation de médicaments antidiabétiques, tels que l'insuline ou des médicaments non insulino-dépendants. Un contrôle strict de la glycémie peut aider à prévenir ou retarder ces complications, bien que les maladies cardiaques restent la principale cause de décès chez les personnes atteintes de diabète[3].

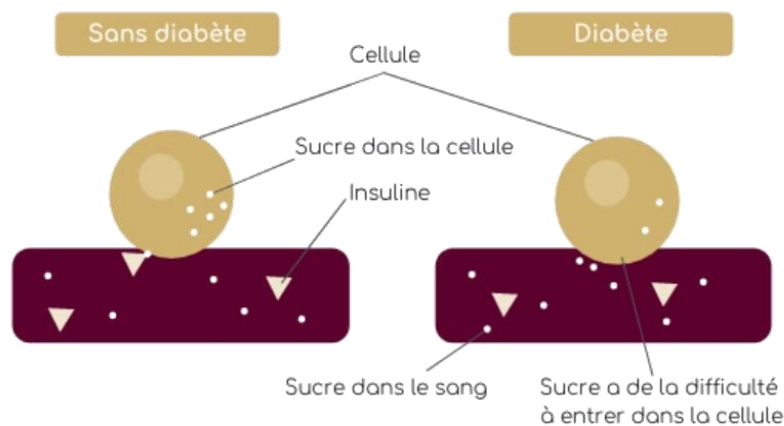


Figure 1.2: Diagramme de la physiopathologie du diabète [4].

4. Types du diabète

4.1 Diabète de type 1 (DT1)

Autrefois désigné sous le nom de diabète insulino-dépendant (DID) ou diabète juvénile en raison de sa prévalence chez les enfants et les jeunes adultes de moins de 35 ans, bien qu'il puisse également affecter les personnes plus âgées, ce type de diabète est rare, touchant environ 0,2 à 0,5% de la population diabétique dans son ensemble, avec une incidence comprise entre 10 et 15%. Dans le cas du DID, comme son nom l'indique, les personnes atteintes de cette maladie deviennent dépendantes d'un apport externe en insuline car leur organisme cesse de la produire. Par conséquent, elles doivent s'injecter plusieurs fois par jour des doses précises d'insuline pour compenser ce manque, car un simple régime alimentaire ne suffit pas à contrôler la maladie [5].

4.2 Diabète de type 2 (DT2)

Autrefois désigné sous le nom de diabète non insulino-dépendant (DNID) et parfois qualifié de "diabète gras" en raison de son association étroite avec l'obésité, ce type de diabète représente la forme la plus courante, constituant entre 85 et 90% de tous les cas de diabète dans le monde. Il se développe progressivement et est souvent déclenché par des habitudes alimentaires inadéquates et un manque d'activité physique. Il survient généralement chez les individus âgés de plus de 40 ans [5].

4.3 Diabète Gestationnel

Environ 4 % des femmes enceintes développent un diabète durant leur grossesse, une condition connue sous le nom de diabète gestationnel. Ce dernier survient plus fréquemment dans le cas de Présence d'obésité, des antécédents familiaux de diabète et dans certaines origines ethniques. Lorsque le diabète gestationnel, n'est pas détecté et traité dans les temps, le risque de complications augmente pour la mère et le fœtus, et peut même conduire au décès du fœtus [5].

5. Symptômes du diabète

Les symptômes les plus fréquents du diabète sont :

- Accroissement de la sensation de soif.
- Augmentation des besoins d'uriner.
- Accroissement de l'appétit.
- Vision floue.
- Sensation de fatigue accrue.
- Nausées.
- Réduction de la capacité à maintenir l'effort pendant l'exercice physique [6].

6. Complications du diabète

Le diabète peut entraîner des complications affectant pratiquement toutes les parties du corps. Avec le temps, l'hyperglycémie affaiblit les parois des petits vaisseaux sanguins, impactant ainsi le cœur, les vaisseaux sanguins, les reins, les yeux, le système nerveux, et plus encore. La figure 1.2 présente les différentes complications du diabète [5].

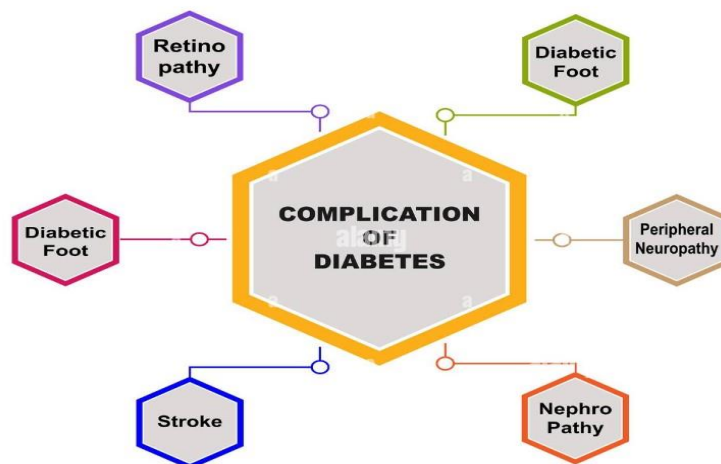


Figure 1.3: Diagramme des complications du diabète [7].

6.1 Maladies cardiovasculaires

Le diabète augmente le risque de maladies cardiovasculaires, telles que les infarctus et les AVC, en raison de l'effet de l'hyperglycémie sur la coagulation sanguine. Les facteurs de risque incluent l'âge, l'hérédité, l'hypertension, l'obésité et le tabagisme, et les diabétiques de type 2 présentent un risque accru dès le départ.

6.2 Néphropathie

Le diabète peut endommager les petits vaisseaux sanguins des reins, entraînant une détérioration progressive allant de l'insuffisance rénale à une maladie rénale irréversible, souvent exacerbée par l'hypertension.

6.3 Troubles oculaires

Le diabète peut progressivement affecter la vision, entraînant des cataractes et même la perte de vue. Les problèmes oculaires sont fréquents chez les diabétiques, touchant la rétine et d'autres parties de l'œil.

6.4 Neuropathie

Les complications nerveuses, appelées neuropathies, sont courantes chez les diabétiques et peuvent provoquer des sensations de picotements, de perte de sensibilité et de douleur, souvent commençant dans les extrémités et progressant le long des membres. La neuropathie peut également affecter les nerfs responsables de la digestion, de la pression artérielle et du rythme cardiaque.

6.5 Sensibilité aux infections

Les fluctuations de la glycémie et la fatigue associée au diabète rendent les patients plus susceptibles d'infections cutanées, gingivales et respiratoires. De plus, le diabète peut ralentir la cicatrisation des plaies, augmentant le risque d'infections chroniques, en particulier aux pieds, ce qui peut parfois nécessiter une amputation en cas de gangrène [5].

7. Facteurs de risque

Plusieurs facteurs contribuent au risque de développer le diabète chez une personne, et ces facteurs peuvent également influencer la probabilité de complications liées à la maladie ainsi que la gestion de celle-ci.

En général, les hommes ont une prévalence plus élevée de diabète par rapport aux femmes. Toutefois, cette disparité entre hommes et femmes en termes de taux de diabète est influencée par divers facteurs socioéconomiques tels que le revenu, le niveau d'éducation et le statut professionnel.

Les populations défavorisées sur le plan socioéconomique présentent un risque accru de développer le diabète et sont plus vulnérables à certains facteurs de risque tels que l'obésité, le tabagisme et l'hypertension artérielle.

Les facteurs de risque du diabète de type 1 restent mal compris, incluant des éléments tels que l'âge, la génétique et les influences environnementales. Prévenir le diabète de type 1 demeure impossible. Plusieurs facteurs de risque jouent un rôle dans le développement du diabète de type 2, parmi lesquels figurent :

- Le prédiabète.
- L'âge avancé.
- Un taux élevé de cholestérol.
- La sédentarité.
- L'hypertension artérielle.
- Le surpoids ou l'obésité [8].

8. Causes du diabète

- Les antécédents familiaux.

- Les disparités culturelles et sociales.

- Les variations d'accès aux soins [8].

9. Diagnostic du diabète

Les indicateurs diagnostiques basés sur une analyse sanguine à jeun sont les suivants :

- Une glycémie normale à jeun est comprise entre 70 et 110 mg/dl.
- Une glycémie à jeun située entre 100 mg/dl et 125 mg/dl indique une intolérance glucidique ou un état de "prédiabète".
- Un diagnostic de diabète est posé si la glycémie à jeun est égale ou supérieure à 126 mg/dl, ou si elle atteint 200 mg/dl à n'importe quel moment [9].

10. Traitement du diabète

Alimentation et activité physique sont primordiales pour les patients atteints de diabète, qu'il soit de type 1 ou de type 2. Dans le cas du diabète de type 1, l'administration d'insuline est essentielle. Pour le diabète de type 2, différents médicaments sont utilisés, tels que les anti hyperglycémiant oraux, les injections non insuliniques comme les inhibiteurs du récepteur du GLP-1, ainsi que l'insuline seule ou en combinaison.

En prévention des complications, il est souvent recommandé d'utiliser des inhibiteurs du système rénine-angiotensine-aldostérone (inhibiteurs de l'ECA ou bloqueurs des récepteurs de l'angiotensine II) et des statines. L'éducation thérapeutique du patient, l'adoption d'un régime adapté, l'exercice physique, la gestion du poids et le suivi régulier de la glycémie sont des éléments clés dans le traitement de tous les patients diabétiques. Certains patients atteints de diabète de type 2 peuvent même éviter ou interrompre leur traitement médicamenteux en maintenant un mode de vie sain axé sur l'alimentation et l'exercice. Pour plus de détails, il est conseillé de consulter les directives spécifiques sur le traitement médicamenteux du diabète [3].

11. Réduction des risques du diabète

Adopter des habitudes de vie saines peut avoir plusieurs bénéfices, notamment :

- Prévenir ou retarder les complications liées au diabète.
- Réduire le risque de prédiabète et de diabète de type 2.
- Contribuer à la prévention d'autres maladies chroniques et à l'amélioration générale de la qualité de vie.

Les experts reconnaissent que notre capacité à modifier notre mode de vie et nos habitudes alimentaires est influencée par divers facteurs tels que l'âge, le sexe, la culture, le revenu, l'éducation, l'emploi, les soutiens sociaux, ainsi que notre environnement quotidien. Il existe de nombreuses façons de cultiver des habitudes de vie saines, telles que :

- Éviter le tabagisme.
- Maintenir un poids santé.
- Limiter la consommation d'alcool.
- S'assurer de suffisamment de sommeil et de repos.
- Adopter une alimentation variée et équilibrée.
- Faire de l'exercice physique régulièrement.
- Contrôler sa pression artérielle, son taux de cholestérol et sa glycémie.

Il est également essentiel de se soumettre régulièrement à des tests de dépistage du diabète et de signaler tout signe ou symptôme à votre professionnel de santé. Des examens médicaux réguliers pour mesurer la pression artérielle, le taux de cholestérol et la glycémie sont également recommandés. Le dépistage précoce permet souvent une prise en charge plus efficace des problèmes de santé [8].

12. Conclusion

Dans ce chapitre, nous avons présenté des informations sur le diabète, y compris l'épidémiologie du diabète, sa définition, ses différents types, ses symptômes, ses complications, ainsi que ses facteurs de risque, ses causes, ses méthodes de diagnostic, son traitement et la réduction du diabète.

Chapitre 2

L'Apprentissage Automatique

1. Introduction

L'apprentissage automatique est un domaine de recherche consacré à l'étude et à l'amélioration des systèmes informatiques capables d'apprendre et de s'adapter à partir des données et de l'expérience. Dans le domaine de la santé, l'apprentissage automatique est utilisé pour développer des systèmes capables de prédire et de détecter les maladies, de diagnostiquer les patients et de fournir des outils aidant les professionnels de la santé à prendre des décisions plus éclairées. Les technologies d'apprentissage automatique peuvent analyser les données médicales, prédire quel traitement sera le plus efficace pour un patient donné et même diagnostiquer des maladies.

Dans ce chapitre, nous présenterons l'apprentissage automatique et nous explorerons ses différents types. Ensuite, nous examinerons les principaux algorithmes utilisés en apprentissage automatique. Nous examinerons dans ce chapitre également quelques travaux de recherche sur la prédiction du diabète et nous présenterons une synthèse de ces approches.

2. Définition

L'apprentissage automatique, une composante de l'Intelligence Artificielle (IA), a la capacité d'acquérir des connaissances à partir de données. En l'absence d'instructions précises, il est capable d'identifier des schémas, d'évaluer des situations, et de s'améliorer continuellement en utilisant des données annotées, des algorithmes, et des modèles statistiques. Les données sont annotées avec des étiquettes informatives qui fournissent un contexte permettant aux algorithmes d'apprentissage automatique d'apprendre à partir de ces données [10].

3. Types d'apprentissage automatique

3.1 Apprentissage supervisé

Dans le cadre de l'apprentissage supervisé, les données d'entraînement fournies à l'algorithme incluent les solutions recherchées, désignées sous différents noms tels que cibles, étiquettes ou « labels ». Ainsi, la machine doit apprendre le processus nécessaire pour produire la sortie désirée à partir de l'entrée. La figure 2.1 présente un exemple de l'apprentissage supervisé [11].

Deux types principaux de problèmes se présentent dans ce contexte

- **La régression** : Dans ce cas, la valeur cible à prédire est de nature continue.
- **La classification** : Ici, la valeur cible à prédire est de nature discrète.

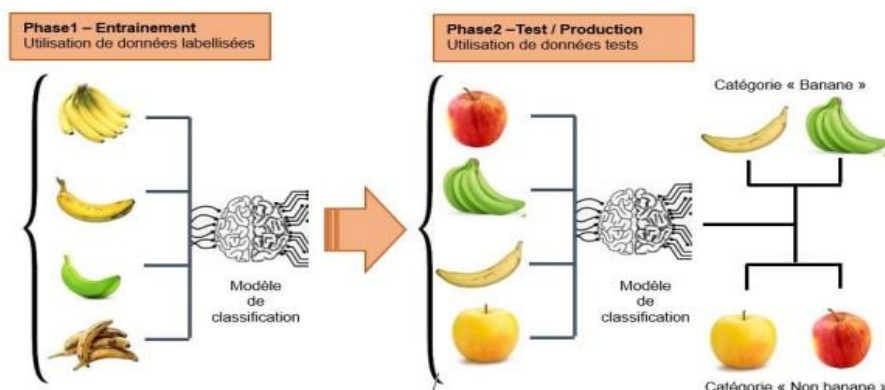


Figure 2.1: Exemple de l'apprentissage supervisé [12].

3.2 Apprentissage non supervisé

Dans l'apprentissage non supervisé, les données sont représentées par une matrice X , mais sans variable cible définie. Lorsque des étiquettes sont présentes, la machine apprend à découvrir les structures dans ces données. En conséquence, elle peut regrouper les données en clusters (via le Clustering), détecter des anomalies, ou encore réduire la dimension des données. Ces résultats proviennent de la machine qui étiquette les solutions en identifiant des motifs communs dans les données [11]. La figure 2.2 présente un exemple de l'apprentissage supervisé.

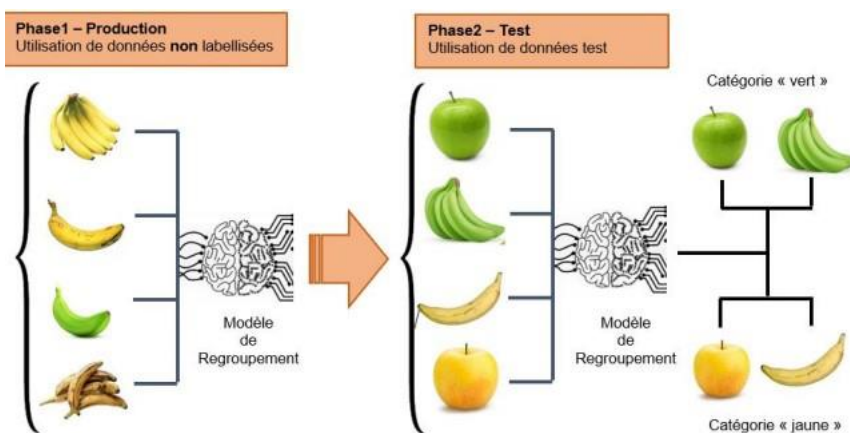


Figure 2.2: Exemple de l'apprentissage non supervisé [12].

3.3 Apprentissage par renforcement

Ce type d'apprentissage, appelé "reinforcement learning", implique un agent qui observe son environnement et prend des actions pour obtenir des récompenses ou des pénalités. Il apprend ensuite la meilleure stratégie, appelée "politique", pour maximiser ses récompenses au fil du temps. L'apprentissage par renforcement est en général, utilisé dans la programmation des robots pour l'apprentissage des mouvements [11]. La figure 2.3 présente un exemple de l'apprentissage par renforcement.

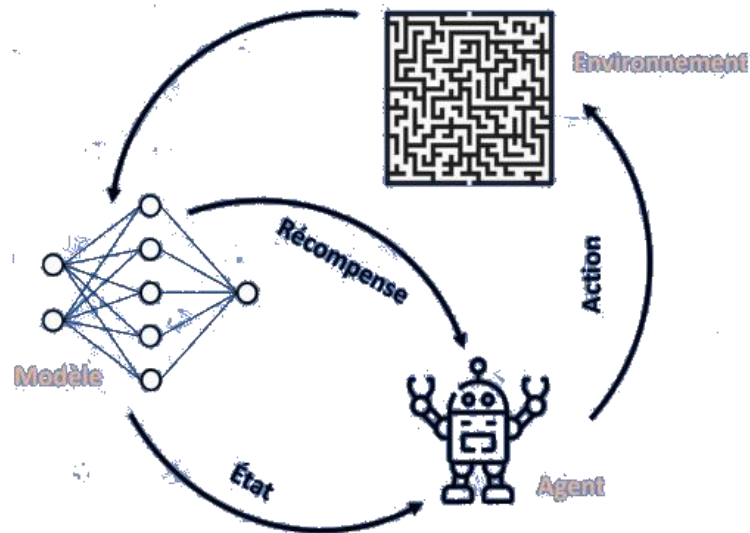


Figure 2.3: Exemple de l'apprentissage par renforcement [12].

4. Algorithmes de l'apprentissage automatique

4.1 DecisionTree (Arbre de Décision)

Un arbre de décision représente visuellement un processus pour aider à la prise de décisions et à la résolution de problèmes.

Un arbre de décision organise les choix et les alternatives possibles sous forme d'une structure en arborescence pour résoudre un problème ou une situation donnée. Il comprend :

- Des nœuds représentant les décisions à prendre pour progresser vers la meilleure solution.
- Des branches illustrant les différentes alternatives associées à chaque décision.
- Des feuilles qui représentent les solutions fournies par l'arbre [13]. La figure 2.4 présente la structure de l'arbre de décision.

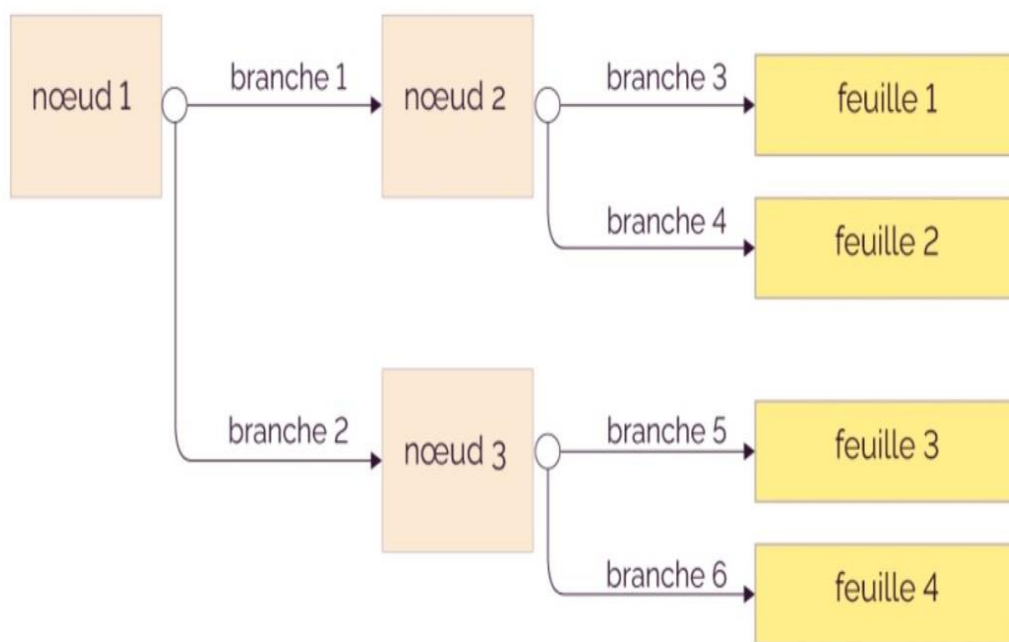


Figure 2.4: Structure d'un arbre de décision [13].

4.2 Random Forest (Forêt Aléatoire)

Les forêts aléatoires (RF) sont des ensembles de classificateurs qui créent plusieurs arbres de décision. Chaque arbre est construit à partir d'un échantillon bootstrap de l'ensemble d'entraînement, utilisant une sélection aléatoire des nœuds. Pour classer une instance, les RF combinent les classifications des arbres individuels en attribuant la classe avec le plus de votes à cette instance. Cette approche protège contre le surapprentissage, souvent observé avec les arbres de décision, tout en offrant des performances élevées, une robustesse remarquable et des temps de calcul raisonnables. Le seul paramètre à ajuster est le nombre de variables disponibles pour le fractionnement à chaque nœud.

Pour formaliser les forêts aléatoires, il est d'abord nécessaire de définir le concept de Bootstrap Aggregating (Bagging) : cette méthode implique un échantillonnage répété avec remplacement d'un ensemble de données. Plutôt que d'estimer une statistique une seule fois sur l'ensemble des données, elle est estimée plusieurs fois sur des échantillons (avec remplacement) de l'échantillon d'origine, permettant ainsi d'obtenir un vecteur d'estimations. La variance, la valeur attendue, la distribution empirique et d'autres statistiques pertinentes peuvent ensuite être calculées à partir de ces estimations.

La formalisation de cet algorithme consiste à construire plusieurs arbres de décision en utilisant des échantillons rééchantillonnés (bootstrapés) de l'ensemble d'entraînement, chaque arbre présentant une variance élevée. L'agrégation de ces

arbres permet de réduire la variance. Pour éviter la corrélation entre les arbres, un sous-ensemble aléatoire de variables est généralement choisi pour chaque arbre, souvent autour de \sqrt{N} , où N est le nombre total de variables. L'importance des variables est évaluée en utilisant l'index de Gini, qui mesure leur impact sur le déroulement des arbres de décision.

Ensuite, une fonction d'agrégation est définie pour prédire une nouvelle instance : dans le cas de la classification, la classe majoritaire prédite par les arbres de décision est choisie, tandis que dans le cas de la régression, la moyenne des résultats prédits par chaque arbre est calculée.

$$G(x) = \text{Vote majoritaire } (G_1(x), \dots, G_B(x))$$

Sachant que B est le nombre d'échantillons obtenus aléatoirement à partir de l'ensemble d'entraînement [11]. Le fonctionnement d'un algorithme de Random Forest peut être illustré à travers la figure 2.5.

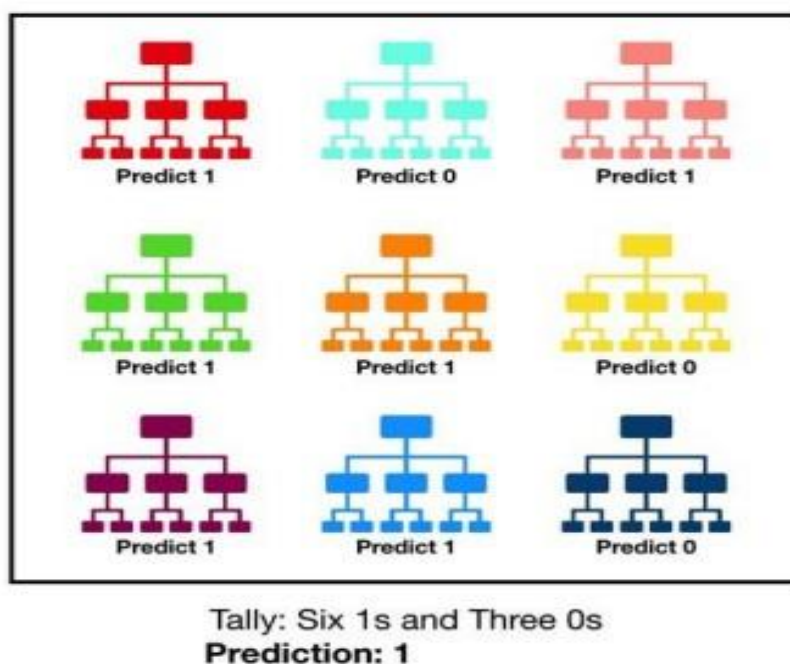


Figure 2.5: Vote majoritaire des arbres de décisions pour le Random Forest [11].

4.3 KNeighbors (KNN)

L'algorithme des K plus proches voisins (KNN) est fondamental mais crucial en apprentissage automatique. Il est applicable à la fois pour les tâches de régression et de classification. Son attrait réside dans son caractère intuitif et sa rapidité de calcul relativement faible.

Le fonctionnement de l'algorithme est simple : pour une tâche de classification donnée, avec deux étiquettes comme "rouge" et "bleu", et un point d'entrée noir, KNN identifie les K points les plus proches et détermine leur couleur majoritaire. Si la majorité est "rouge", le point noir est classé comme "rouge". Les K voisins sont sélectionnés en utilisant une mesure de distance, généralement la distance euclidienne pour les valeurs réelles. D'autres métriques comme la distance de Manhattan, la distance de Minkowski et la distance Euclidienne peuvent également être utilisées [11].

La distance Euclidienne entre deux points p et q :

$$D(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2}$$

Le processus d'apprentissage de la méthode KNN est illustré sur la figure 2.6.

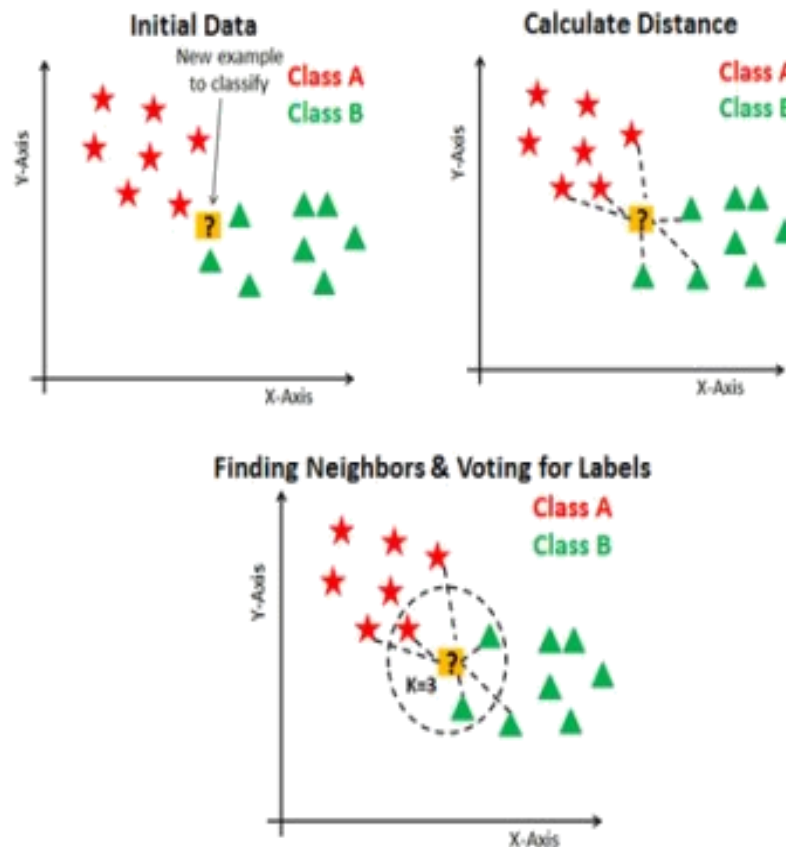


Figure 2.6 Classification par la méthode KNN [11].

4.4 K-means

K-means, également appelé K-moyennes, est l'un des algorithmes de clustering les plus utilisés en analyse de données. Son objectif est de regrouper des données similaires en clusters, basé sur un ensemble de descripteurs caractérisant ces données.

La similarité entre deux données est évaluée en fonction de la distance entre leurs descripteurs. Ainsi, des données très similaires ont des descripteurs très proches. Le problème de partitionnement des données est alors formulé comme la recherche de K "prototypes de données" autour desquels les autres données peuvent être regroupées. Ces prototypes sont appelés centroïdes.

L'algorithme associe chaque donnée à son centroïde le plus proche pour former des clusters. Les centroïdes sont ensuite déplacés vers la moyenne des descripteurs de leur groupe, ce qui donne son nom à l'algorithme (K-moyennes ou K-means).

K-means initialise ses centroïdes en choisissant aléatoirement des données dans le jeu de données, puis itère plusieurs fois en regroupant les données autour des centroïdes les plus proches et en ajustant les centroïdes vers les moyennes des descripteurs de leur groupe.

Une fois que l'algorithme a convergé, c'est-à-dire qu'il a trouvé un découpage stable du jeu de données, il est considéré comme terminé.

K-means est apprécié pour sa simplicité, sa rapidité et sa facilité de compréhension, bien qu'il ne soit pas idéal pour détecter des groupes avec des formes complexes.

L'exemple donné utilise le jeu de données "Iris", décrivant des fleurs en termes de longueurs et de largeurs de pétales et de sépales, pour illustrer le fonctionnement de l'algorithme. Chaque fleur est représentée par un point, coloré selon son groupe d'appartenance, tandis que les croix représentent les centroïdes et les traits délimitent les frontières entre les clusters [14]. La figure 2.7 présente un exemple sur k-means.

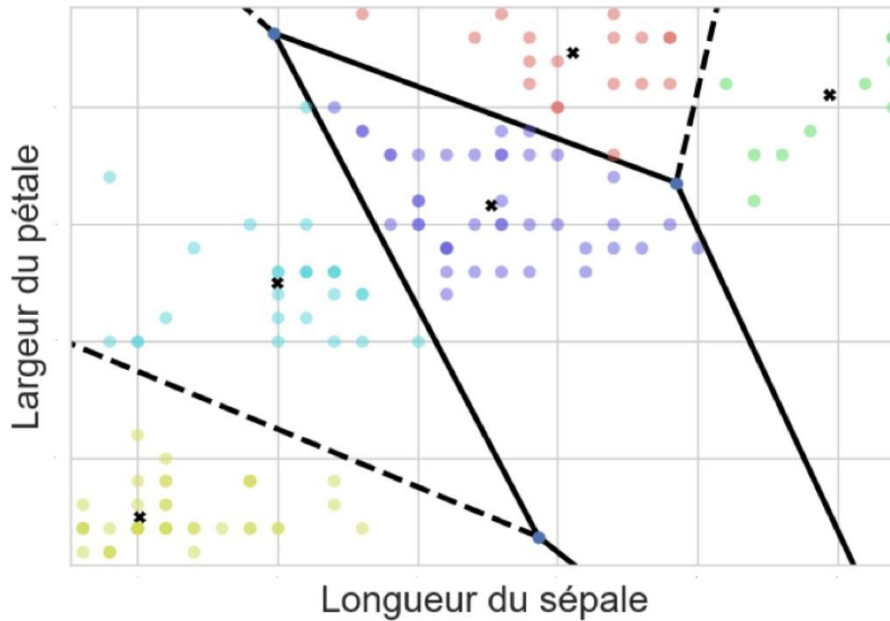


Figure 2.7 : Exemple sur K-means [14].

5. Travaux de recherche sur la prédiction du diabète par les méthodes d'apprentissage automatique

Ces dernières années, de nombreuses études ont été menées sur la prédiction du diabète avec les méthodes de l'apprentissage automatique. Dans cette section, nous présenterons quelques-uns des travaux que nous avons examinés au cours de notre recherche.

- RishabBothra [15], a utilisé des différents algorithmes d'apprentissage automatique ont été appliqués au jeu de données pour effectuer la classification. Parmi ces algorithmes, le Random Forest a obtenu la plus haute précision avec 90%. Les autres résultats sont les suivants : Logistic Regression 73%, XGBoost 88%, SVM 74%, KNN 89%. Une comparaison des précisions des différents algorithmes a été réalisée, ainsi qu'une analyse des matrices de confusion afin de minimiser autant que possible les faux négatifs. De plus, il est envisageable d'étendre la recherche pour déterminer si une personne non diabétique est susceptible de développer un diabète dans les prochaines années.
- Debadri Dutta et Al. [16], ont conclu que l'algorithme Random Forest est le plus adapté pour prédire le diabète, avec une précision d'environ 84%. Ils ont également constaté que pour prévenir le diabète, il est recommandé de maintenir un taux de glucose bas et d'adopter une alimentation équilibrée en vieillissant. De plus, ils ont observé que les personnes nées dans des familles ayant des antécédents de diabète devraient vraiment prendre soin d'elles-mêmes.

- Aishwarya Jakka et Vakula Rani J [17], ont utilisés différents modèles de classification en apprentissage automatique sont évalués pour le diagnostic du diabète. La régression logistique se distingue avec la plus haute précision, à 77,6 %, surpassant les autres techniques, telles que le KNN avec 73,43 %, l'arbre de décision avec 70,31 %, le Naive Bayes avec 75,52 %, le SVM avec 65,63 %, et le Random Forest avec 74,30 %.

- Gudluri Saranya et Sagar DhanrajPande [19], ont appliqués divers algorithmes d'apprentissage automatique sur la base de données des Indiens Pima pour prédire la présence du diabète. Les algorithmes utilisés comprennent le SVM, l'arbre de décision, la forêt aléatoire, KNN, la régression linéaire, la régression logistique, le Naive Bayes et XGBoost. Les résultats expérimentaux ont révélé que la forêt aléatoire a surpassé les autres avec une précision de 91,10%. Voici un aperçu des précisions obtenues avec chaque algorithme : SVM : 83,11%, Arbre de décision : 87,01%, Forêt aléatoire : 91,10%, KNN : 79,22%, Régression linéaire : 80,52%, Régression logistique : 84,41, Naive Bayes : 81,81%, XGBoost : 75,32%.

- Talha Mahboob Alama & al [19], dans cette recherche, le diabète est prédit à partir des attributs significatifs, et la relation entre ces attributs est également caractérisée. Ils ont utilisés divers outils pour sélectionner les attributs significatifs, ainsi que pour le clustering, la prédiction et l'extraction de règles d'association pour le diabète. La sélection des attributs significatifs a été réalisée par la méthode d'analyse en composantes principales. Les techniques de réseau de neurones artificiels (ANN), de forêt aléatoire (RF) et de clustering K-means ont été utilisées pour prédire le diabète. La technique ANN a atteint la meilleure précision avec 75,7%.

- Farooqal, N. A., & al [20], ils ont utilisés différentes techniques d'apprentissage automatique, à savoir l'arbre de décision, les k-plus proches voisins, la forêt aléatoire et la machine à vecteurs de support, pour prédire les performances de différentes techniques de classification. Ainsi ils ont conclu que la performance de la forêt aléatoire est 84,05%, et elle est supérieure à celle des autres techniques de classification.

- Rajagopal, A, & Al [21], cette recherche présente un modèle hybride personnalisé combinant un réseau de neurones artificiels et des algorithmes génétiques pour prédire efficacement la maladie du diabète. Ce modèle utilise une technique de normalisation innovante pour prétraiter les données médicales, identifie l'importance des variables influençant la prédiction, et applique une méthode de régularisation asymétrique adaptée aux caractéristiques du jeu de données. Les

résultats montrent une précision de prédiction de 80% sur le jeu de données des Indiens Pima provenant du référentiel UCI Machine Learning.

- Nguyen, B. P. & [22] dans cet article, ils ont appliqués un modèle d'apprentissage large et profond qui combine la puissance d'un modèle linéaire généralisé avec diverses caractéristiques et un réseau neuronal à propagation avant profond pour améliorer la prédiction du début du diabète de type 2 (T2DM).
- Das, P., & Nanda, S. [23], dans cet article, ils ont utilisés un classifieur de machine à apprentissage extrême avec régression ridge et un algorithme d'optimisation des vecteurs de poids basé sur la méthode des lucioles. La base de données des diabétiques de la tribu PIMA indienne est utilisée pour l'entraînement et les tests du modèle. Les précisions maximales atteintes sont de 93,4 %, la sensibilité de 97,5 % et la spécificité de 85,72 %. Les résultats du modèle sont comparés à deux méthodes populaires, la machine à vecteur de support (SVM) et la machine à apprentissage extrême (ELM), démontrant que la méthode proposée surpasse SVM et ELM.
- Vidhya, K., & Shanmugalakshmi, R. [24], dans cette recherche, le modèle proposé suit les étapes de collecte des données, de pré-entraînement, d'extraction des caractéristiques, de Réseau de Croyance Profonde (RCP), de processus de validation, et de classification pour prédire les complications diabétiques. Le processus d'entraînement ainsi que celui de test délimite la prévalence du risque avec une précision de 81,20%. Ce modèle de prédiction réaliste sera très utile pour gérer efficacement le diabète.

Auteurs	Année	Méthodes	Accuracy
Farooqual, N. A. & al [20]	2018	KNN, RF, SVM	RF= 84,05%
Talha Mahboob Alama & al [19]	2018	RF, K-means, ANN	ANN=75,7%
Debadri Dutta et Al. [16]	2018	RF	RF=84%
Aishwarya Jakka et Vakula Rani J [17]	2019	Logistic Regression, KNN, DT, NB, SVM, RF	Logistic Regression=77,6%
Nguyen, B. P., & [22]	2019	Wide and Deep Learning	84,28%
Vidhya, K., & Shanmugalakshmi, R.[24]	2020	Deep Bielf Network	81,2%

RishabBothra [15]	2021	Logistic Regression, XGBoost, SVM, KNN, RF	RF= 90%
Das, P., & Nanda, S.[23]	2021	Classifieur de machine à apprentissage extrême , Régression ridge et un algorithme d'optimisation des vecteurs de poids basé sur la méthode des lucioles	93,4%
Rajagopal, A, & Al [21]	2022	Modèle hybride personnalisé combinant un réseau neurones artificiels	80%
Gudluri Saranya et Sagar DhanrajPande [18]	2023	DT, RF, KNN, XGBoost, NB, Logistic Regression, Linear Regression, SVM	RF=91,1%

Tableau 1.1: Travaux connexes.

6. Synthèse

À partir de l'étude que nous avons menée sur la prédiction du diabète nous avons constaté que :

- **Variabilité des résultats** : Les différentes études rapportent des précisions différentes pour les mêmes algorithmes. Cela souligne l'importance de prendre en compte divers facteurs tels que la qualité des données, les paramètres de l'algorithme, et la méthodologie de l'étude.
- **Choix des algorithmes** : Chaque étude utilise une gamme différente d'algorithmes d'apprentissage automatique. Bien que Random Forest soit souvent mis en avant, d'autres algorithmes comme la régression logistique et SVM montrent également des performances compétitives dans certaines études.
- **Importance des caractéristiques** : Certaines études soulignent l'importance de certaines caractéristiques, telles que le taux de glucose et les antécédents familiaux, dans la prédiction du diabète. Cela met en évidence l'importance de la sélection et de l'ingénierie des caractéristiques dans la construction de modèles efficaces.

- **Considérations cliniques** : Plusieurs études font des recommandations cliniques basées sur leurs résultats, telles que le maintien d'un taux bas de glucose et une alimentation équilibrée pour prévenir le diabète. Cela montre comment les résultats de l'apprentissage automatique peuvent être traduits en recommandations pratiques pour la santé.

7. Conclusion

Dans ce chapitre, nous avons présenté les principaux algorithmes d'apprentissage automatique qui sont les plus utilisés pour la détection des maladies afin de réduire les risques de complications pour la santé des patients. L'objectif principal étant d'appliquer ces différents algorithmes de classification (K-Nearest Neighbors, Decision Trees, Random Forest, Logistic Regression et Gradient Boosting) dans le contexte de notre étude.

Chapitre 3

Description du Système Développé et Analyse des Résultats

1. Introduction

Dans ce chapitre, nous allons tout d'abord présenter les différents outils, les bibliothèques, et les langages de programmation utilisés pour le développement de notre système de prédiction. Ensuite, nous décrirons l'ensemble de données utilisées dans la prédiction et nous présenterons les différentes étapes que nous avons suivi pour la création des modèles de prédiction. Enfin, nous présenterons une évaluation des modèles d'apprentissage utilisés afin de choisir le modèle le plus approprié à utiliser pour la prédiction du diabète au sein de notre système.

2. Outils et bibliothèques utilisées

Pour l'implémentation de notre système, nous avons installé la distribution Anaconda pour le système d'exploitation Windows 10, et nous avons utilisé l'ensemble d'outils et les bibliothèques que nous présentons dans cette partie.

2.1 Anaconda

Anaconda, une distribution libre et open source de Python et R, est conçu pour simplifier le développement d'applications dédiées à la science des données et à l'apprentissage automatique. Elle facilite la gestion des paquets et leur déploiement grâce au système conda. Utilisée par plus de 6 millions de personnes, Anaconda offre une installation comprenant plus de 250 paquets populaires pour la science des données, compatibles avec Windows, Linux et MacOS. De plus, plus de 7 500 paquets open-source supplémentaires peuvent être installés à partir de PyPI ou via le gestionnaire de paquets et les environnements virtuels de conda [25]. La figure 3.1 présente le logo d'Anaconda, et la figure 3.2 présente leur interface.



Figure 3.1: Logo d'Anaconda [26].

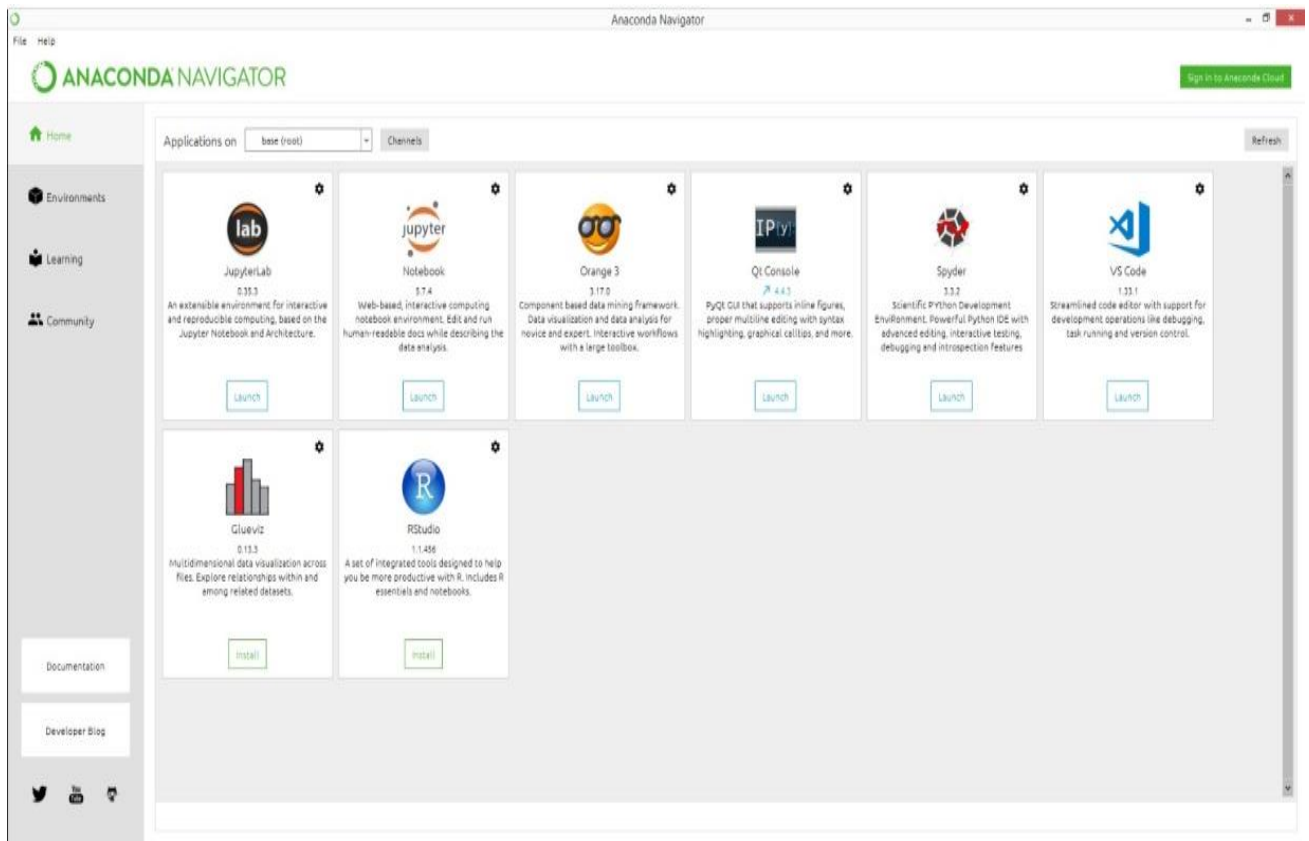


Figure 3.2 : Interface d'Anaconda [25].

2.2 Jupyter

Jupyter, une plateforme web, qui permet la programmation dans plus de 40 langages, incluant Python, Julia, Ruby, R, et Scala2. Ce projet collaboratif vise à développer des logiciels libres, des formats ouverts, et des services pour l'informatique interactive. Émergeant du projet IPython, Jupyter facilite la création de calepins ou notebooks, fusionnant texte formaté grâce à Markdown et code source, ainsi que les résultats d'exécution. Ces calepins sont largement utilisés en science des données pour explorer et analyser des données [27]. La figure 3.3 présente le logo de Jupyter.



Figure 3.3: Logo de Jupyter [27].

2.3 Python

Python est un langage de programmation polyvalent, adapté à différents styles de programmation, comme l'impératif, le fonctionnel et l'orienté objet. Il se distingue par son typage dynamique fort, sa gestion automatique de la mémoire et son système d'exceptions. Ce langage est également populaire dans l'éducation, car sa syntaxe claire et séparée des détails techniques facilite l'initiation aux concepts de base de la programmation. Selon l'Index TIOBE, sa popularité ne cesse de croître, notamment en raison de son efficacité pour l'apprentissage automatique [28]. La figure 3.4 présente le logo de Python.



Figure 3.4: Logo de Python [29].

2.4 Modules de développement

2.4.1 Pandas

Pandas, une bibliothèque conçue pour Python, vise à simplifier la manipulation et l'analyse de données. Elle offre diverses fonctionnalités pour travailler avec des tableaux numériques et des séries temporelles. La figure 3.5 présente le logo de Pandas. Les structures de données principales qu'elle gère incluent :

- Séries : elles stockent des données unidimensionnelles associées à un index.
- DataFrames : ces structures bidimensionnelles organisent les données en lignes et colonnes.
- Panels : pour représenter les données sur trois dimensions.
- Panels4D ou DataFrames avec MultiIndex : ces derniers permettent de gérer des données sur plus de trois dimensions, également appelées hypercubes [30].

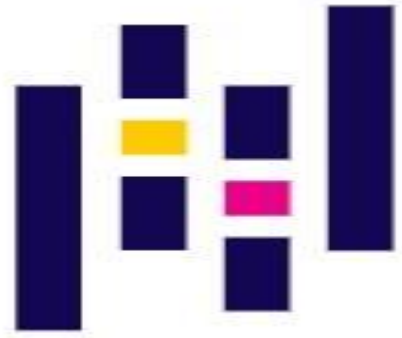


Figure 3.5: Logo de Pandas [31].

2.4.2 Numpy

NumPy est une bibliothèque conçue pour le langage de programmation Python, se concentre sur la manipulation de matrices et de tableaux multidimensionnels, accompagnée de fonctions mathématiques pour les opérer. Cette bibliothèque open source offre une gamme étendue de fonctionnalités, notamment la création et la sauvegarde de tableaux à partir de fichiers, ainsi que la manipulation de vecteurs, matrices et polynômes. En tant que fondement de SciPy, un ensemble de bibliothèques Python dédiées au calcul scientifique, NumPy joue un rôle essentiel dans ce regroupement [32]. La figure 3.6 présente le logo de Numpy.



Figure 3.6 : Logo de Numpy [32].

2.4.3 Scikit-learn

Scikit-learn est une bibliothèque Python open source dédiée à l'apprentissage automatique. Cette bibliothèque offre une multitude d'algorithmes prêts à l'emploi dans son cadre, facilitant ainsi le travail des data scientists. Elle propose une gamme variée de fonctionnalités, notamment l'estimation de forêts aléatoires, de régressions logistiques, d'algorithmes de classification et de machines à vecteurs de support. Scikit-learn est conçue pour s'intégrer harmonieusement avec d'autres bibliothèques Python open source telles que NumPy et SciPy [33]. La figure 3.7 présente le logo de Scikit-learn.



Figure 3.7: Logo de Scikit-learn [33].

2.4.4 Matplotlib

Matplotlib est une librairie de Python, est conçue pour créer et présenter des données graphiquement. Elle s'intègre aisément avec NumPy et SciPy pour manipuler les données scientifiques. De plus, elle offre une API orientée objet pour incorporer des graphiques dans diverses applications, en utilisant des outils d'interface graphique comme Tkinter, wxPython, Qt ou GTK [34]. La figure 3.8 présente le logo de Matplotlib.



Figure 3.8: Logo de Matplotlib [35].

2.4.5 Seaborn

Seaborn est une bibliothèque permettant de créer des graphiques statistiques en Python. Elle se base sur matplotlib et s'intègre étroitement avec les structures de données pandas. Elle aide à explorer et à comprendre les données. Ses fonctions de traçage opèrent sur des dataframes et des tableaux contenant des ensembles de données entiers et effectuent internement le mappage sémantique nécessaire et l'agrégation statistique pour produire des graphiques [36]. La figure 3.9 présente le logo de Seaborn.



Figure 3.9: Logo de Seaborn [37].

2.4.6 Tkinter

Tkinter en Anglais Tool kit interface est la bibliothèque graphique libre d'origine pour le langage Python, permettant la création d'interfaces graphiques. Elle vient d'une adaptation de la bibliothèque graphique Tk écrite pour Tcl [38].

2.4.7 Joblib

Joblib est une bibliothèque open-source pour Python qui simplifie le traitement parallèle, la mise en cache des résultats et la distribution de tâches. Elle permet aux développeurs d'accélérer les calculs intensifs en parallélisant les opérations, ce qui réduit significativement le temps d'exécution global [39]. La figure 3.10 présente le logo de Joblib.



Figure 3.10: Logo de Joblib [39].

3. Plan de développement du système

Pour le développement d'un système pour la prédiction du diabète, nous avons suivi l'ensemble des étapes présenté sur la figure 3.11.

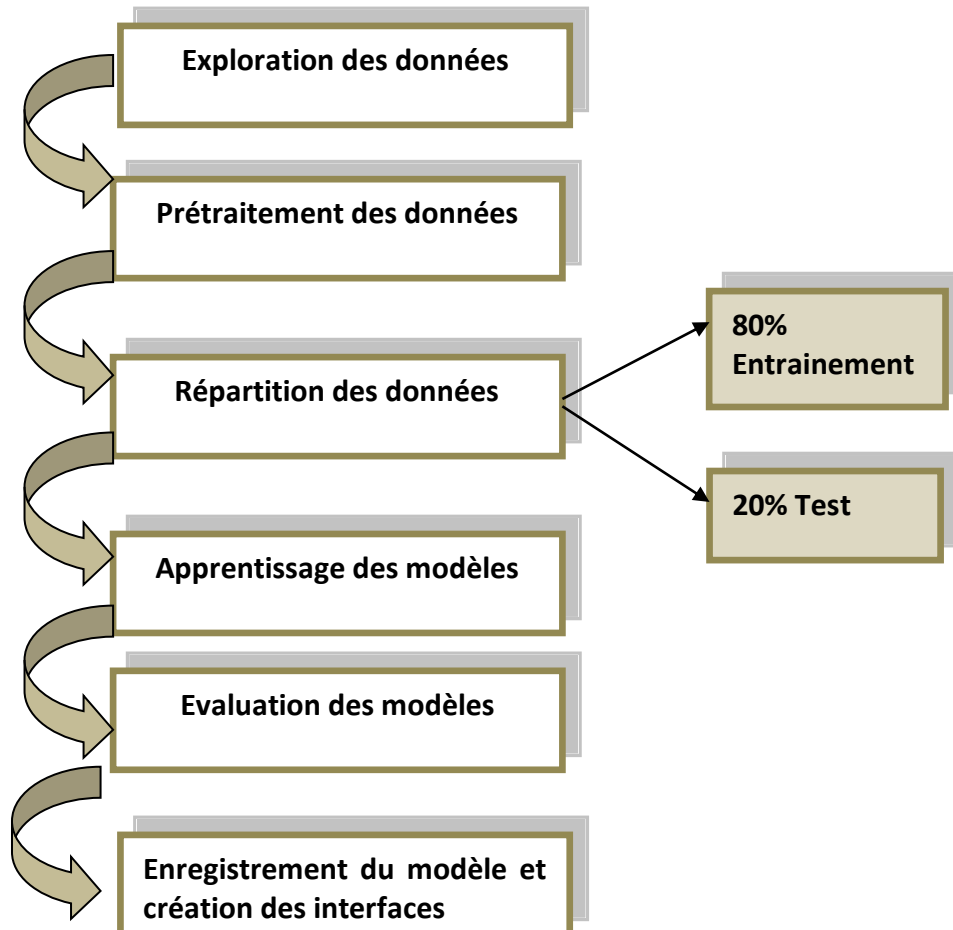


Figure 3.11 :Plan de développement du système de prédiction.

- **Exploration des données** : Dans cette étude nous avons utilisé des données médicales provenant de 768 patients qui sont regroupés dans le datasetPIMA (IndiansDiabetesDataset) [40].
- **Prétraitement des données** : Cette étape consiste au nettoyage du dataset et à la préparation des données pour pouvoir les utiliser dans l'apprentissage des modèles.
- **Répartition des données** : Une fois le prétraitement effectué, nous avons divisées les données en deux ensembles distincts : l'ensemble d'apprentissage 80% et l'ensemble de test 20%. L'ensemble d'apprentissage est utilisé pour entraîner le modèle, tandis que l'ensemble de test est utilisé pour évaluer les performances du modèle.

- **Apprentissage des modèles :** Dans cette étape, nous avons utilisés les cinq algorithmes pour l'apprentissage automatique des modèles : Decision Tree, Random Forest, KNN, Gradient Boosting et Logistic Regression.
- **Evaluation des modèles:** Une fois que les modèles ont été entraînés, ils sont évalués en fonction d'un ensemble de métriques qui permettent l'évaluation des performances de chaque modèle.
- **Enregistrement du modèle et création des interfaces :** Après l'évaluation des modèles nous avons opté pour le modèle qui a donné les meilleurs résultats. Pour la création de notre système de prédiction nous avons enregistré le meilleur modèle obtenu et nous avons procédé à la création des interfaces du système.

4. Le système de prédiction du diabète proposé

4.1 Description du Dataset

Dans ce travail nous avons utilisé le dataset PIMA, qui comprend des cas réels de 768 patients. Parmi ces patients, 268 sont diabétiques et les 500 restants ne le sont pas. La figure 3.12 présente les caractéristiques principales du dataset PIMA.

Caractéristique	Définition	Intervalle
Pregnancies	C'est le nombre de fois que la patiente a été enceinte.	[0...17]
Glucose	Niveau de glucose dans le sang.	[0...199 mg/dl]
BloodPressure	Mesure de la pression artérielle.	[0...122 mmHg]
SkinThickness	Mesure de l'épaisseur du pli cutané.	[0...99 mm]
Insulin	Niveau d'insuline sérique.	[0...846 μU/mL]
BMI	Mesure de la corpulence d'une personne, calculée en fonction de sa taille et de son poids.	[0...67.1 kg/m ²]
DiabetesPedigreeFunction	Est une mesure utilisée dans le contexte de la génétique du diabète pour évaluer la prédisposition génétique à cette maladie.	[0.078...2.42]
Age	C'est l'âge du patient.	[21...81 ans]

Table 3.1: Description des variables du dataset.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1
10	4	110	92	0	0	37.6	0.191	30	0
11	10	168	74	0	0	38.0	0.537	34	1
12	10	139	80	0	0	27.1	1.441	57	0
13	1	189	60	23	846	30.1	0.398	59	1

Figure 3.12 :Dataset PIMA.

4.2 Exploration et Visualisation des données

La visualisation des données consiste en une exploration visuelle et interactive de jeux de données. Elle permet de révéler des informations qui étaient auparavant difficiles à percevoir.

- **Visualisation des données**

La Figure 3.13 présente les deux classes contenues dans la base de données : "1" pour les patients non malades et "0" pour les patients malades. Comme, il apparaît sur la figure on constate que les deux classes ne sont pas équilibrées. Sur un total de 768 patients, 268 qui ont une maladie du diabète et 500 qui n'ont pas.

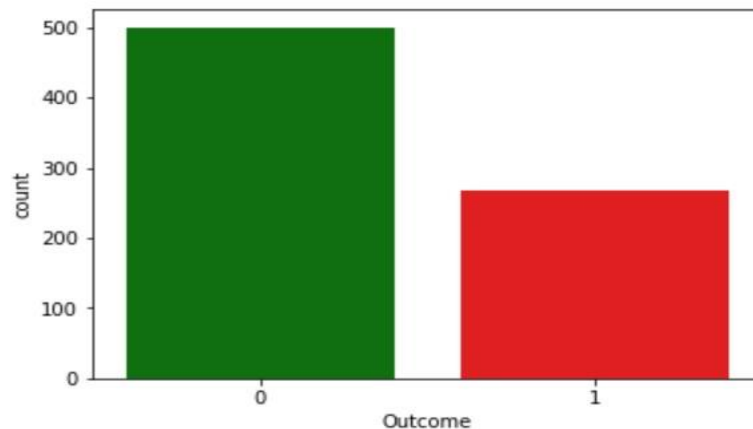


Figure 3.13 :Répartition des données.

▪ **Distribution des variables**

La distribution des variables fait référence à la manière dont les valeurs d'une variable (ou d'un ensemble de variables) sont réparties dans un ensemble de données.

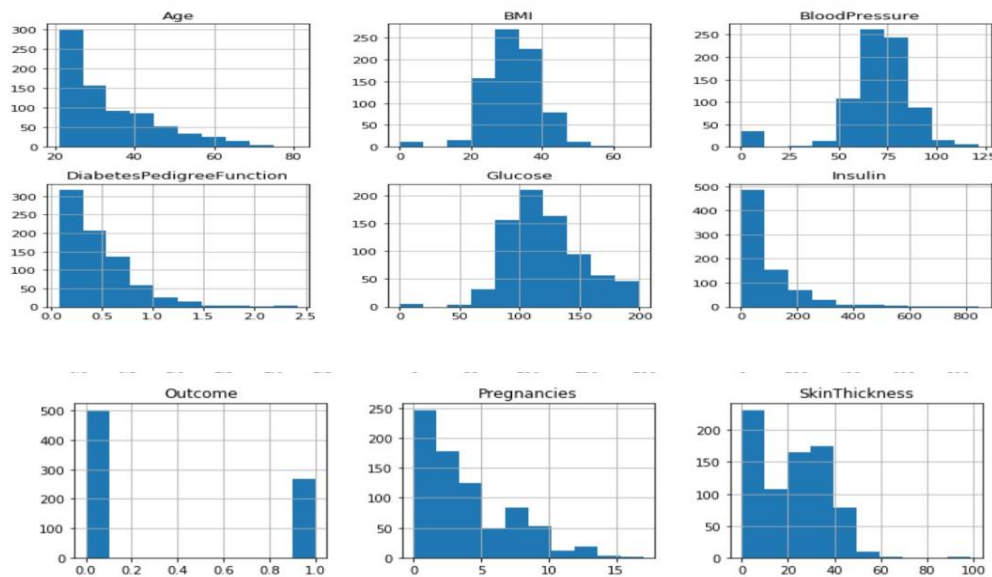


Figure 3.14 : Distribution des variables.

4.3 Corrélation des données

L'une des étapes clés pour améliorer les données consiste à déterminer la corrélation entre les variables. La corrélation est utilisée pour mesurer la dépendance entre deux variables différentes.

La matrice de corrélation est une matrice symétrique qui contient les coefficients de corrélation (ou d'autres mesures de corrélation) entre toutes les paires de variables dans un ensemble de données. Chaque ligne et chaque colonne de la matrice représentent une variable différente, et chaque cellule indique la corrélation entre les variables correspondantes.

Les coefficients de corrélation dans la matrice peuvent varier de -1 à +1 :

- La valeur +1 dans la matrice indique une corrélation positive parfaite : lorsque la valeur d'une variable augmente, la valeur de l'autre variable augmente également dans une proportion constante.
- La valeur -1 indique une corrélation négative parfaite : lorsque la valeur d'une variable augmente, la valeur de l'autre variable diminue dans une proportion constante.
- Une valeur proche de 0 indique une absence de corrélation linéaire entre les variables.

Pour mieux comprendre les dépendances entre les différentes variables du Dataset, nous avons créé la matrice de corrélation sur Figure 3.15, qui montre la relation entre les indicateurs et l'étendue des relations entre eux.

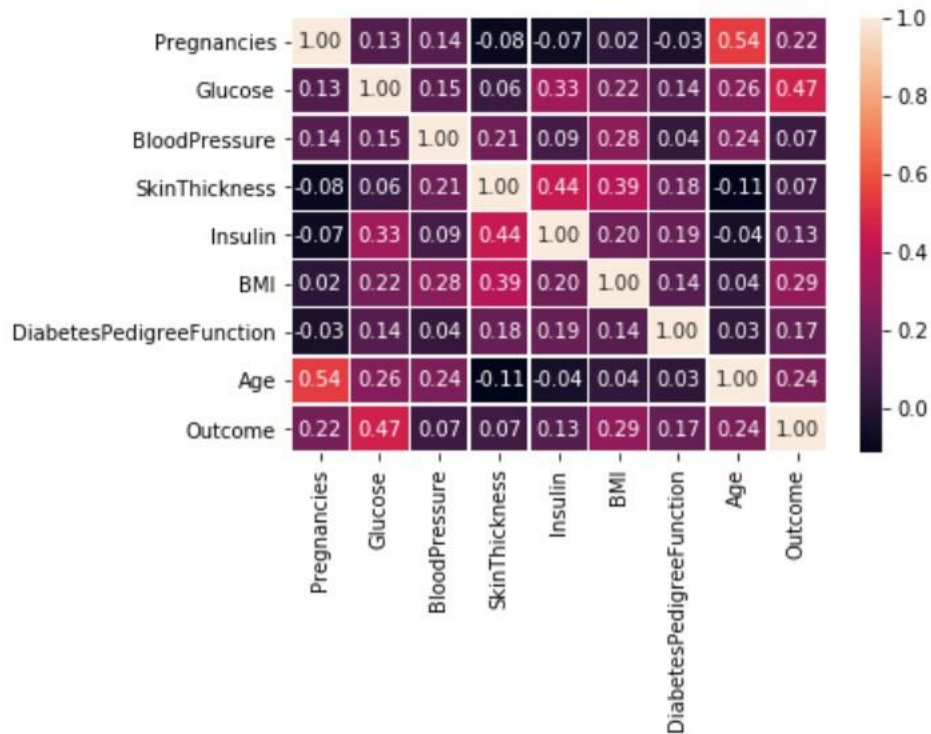


Figure 3.15 : Matrice de Corrélation.

Comme nous pouvons le constater sur la matrice de corrélation :

- Glucose est la variable la plus corrélée avec le diabète (0.47) et devrait donc être une caractéristique clé dans les modèles prédictifs.
- BMI, Age, et Pregnancies montrent aussi des corrélations modérées avec le diabète, indiquant qu'elles sont également importantes.
- Il y a une corrélation modérée à forte entre certaines variables indépendantes, tels que : Insulin et Glucose (0.54), ce qui est logique car les niveaux d'insuline influencent directement la concentration de glucose dans le sang.
- Même si certaines variables ont une faible corrélation avec le diabète individuellement tels que : BloodPressure (0.07), SkinThickness (0.08), elles peuvent encore améliorer la précision prédictive du modèle lorsqu'elles sont combinées avec d'autres variables.

À partir de l'analyse effectuée sur cette matrice, nous n'avons éliminé aucune variable dans cette étude, et le jeu de données prétraité final comprenait les 9 variables.

4.4 Prétraitement des données

Le prétraitement des données consiste à corriger les enregistrements contenant des valeurs corrompues ou non valides, dans le but d'améliorer la qualité des données. Dans le cadre de

notre travail, nous avons appliqué deux techniques pour le traitement des données afin d'améliorer les résultats de prédiction :

- **Nettoyage des données**

Afin de nettoyer les données, nous avons éliminé les valeurs nulles par le remplacement les valeurs nulles de certaines caractéristiques (telles que le Glucose, Blood Pressure, SkinThickness et l'Insulin) par leurs moyennes.

Comme nous pouvons le constater la figure 3.16, présente le nombre de valeurs nulle des variables Glucose, BloodPressure, Insulin et SkinThickness dans le dataset. En fait l'ensemble de ces variables médicales, ne peuvent avoir de valeurs égales à zéro (0), car cela les rendrait non représentatives dans leurs plages de valeurs médicales. De même, L'BMI (Indice de Masse Corporelle) est une caractéristique calculée qui dépend du poids, et ce dernier ne peut pas être nul.

```
print(donnees[donnees['Glucose']==0].shape)
print(donnees[donnees['BloodPressure']==0].shape)
print(donnees[donnees['SkinThickness']==0].shape)
print(donnees[donnees['Insulin']==0].shape)
print(donnees[donnees['BMI']==0].shape)
```

```
(5, 9)
(35, 9)
(227, 9)
(374, 9)
(11, 9)
```

Figure 3.16 : Visualisation des valeurs nuls.

La figure 3.17 présente l'opération de nettoyage de valeurs nulles dans le dataset.

```
donnees['Glucose']=donnees['Glucose'].replace(0,donnees['Glucose'].mean())
donnees['BloodPressure']=donnees['BloodPressure'].replace(0,donnees['BloodPressure'].mean())
donnees['SkinThickness']=donnees['SkinThickness'].replace(0,donnees['SkinThickness'].mean())
donnees['Insulin']=donnees['Insulin'].replace(0,donnees['Insulin'].mean())
donnees['BMI']=donnees['BMI'].replace(0,donnees['BMI'].mean())
```

Figure 3.17 : Code de Nettoyage des valeurs nulles.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.000000	35.000000	79.799479	33.600000	0.627	50	1
1	1	85.0	66.000000	29.000000	79.799479	26.600000	0.351	31	0
2	8	183.0	64.000000	20.536458	79.799479	23.300000	0.672	32	1
3	1	89.0	66.000000	23.000000	94.000000	28.100000	0.167	21	0
4	0	137.0	40.000000	35.000000	168.000000	43.100000	2.288	33	1
5	5	116.0	74.000000	20.536458	79.799479	25.600000	0.201	30	0
6	3	78.0	50.000000	32.000000	88.000000	31.000000	0.248	26	1
7	10	115.0	69.105469	20.536458	79.799479	35.300000	0.134	29	0
8	2	197.0	70.000000	45.000000	543.000000	30.500000	0.158	53	1
9	8	125.0	96.000000	20.536458	79.799479	31.992578	0.232	54	1
10	4	110.0	92.000000	20.536458	79.799479	37.600000	0.191	30	0
11	10	168.0	74.000000	20.536458	79.799479	38.000000	0.537	34	1
12	10	139.0	80.000000	20.536458	79.799479	27.100000	1.441	57	0
13	1	189.0	60.000000	23.000000	846.000000	30.100000	0.398	59	1

Figure 3.18 : Dataset après le nettoyage des données.

- **Équilibrage du dataset:** Cette étape vise à équilibrer les deux classes de décision (Outcome). La figure suivante montre l'équilibrage des données du dataset.

```
rm=RandomOverSampler(random_state=41)
x_res,y_res=rm.fit_resample(x,y)

print('old donnees shape{}'.format(Counter(y)))
print('old donnees shape{}'.format(Counter(y_res)))

old donnees shapeCounter({0: 500, 1: 268})
old donnees shapeCounter({1: 500, 0: 500})
```

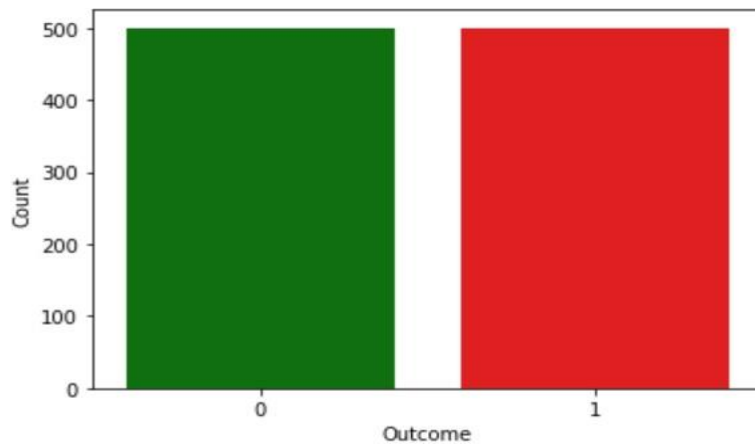


Figure 3.19: Répartition des données après l'équilibrage du Dataset.

4.5 Répartition des données

Nous avons effectué une répartition de 80% pour l'ensemble d'entraînement et de 20% pour l'ensemble de test, comme le montre la figure 3.20. Cette répartition est couramment utilisée pour évaluer les performances des modèles d'apprentissage automatique.



```
x_train, x_test, y_train, y_test = train_test_split(x_res, y_res, test_size=.2, random_state=41, shuffle=True)
```

Figure 3.20 : Aperçu de la répartition des données.

4.6 Entraînement des modèles

Après la division des données, l'étape suivante consiste à choisir un modèle pour prédire le diabète. Dans cette étude, quatre modèles ont été utilisés pour la prédiction précoce du diabète, à savoir: Forêt Aléatoire (Random Forest), Arbre de Décision (DecisionTree), K-Plus Proche Voisin (KNN), Gradient Boosting et la Régression Logistique (LogisticRegression). Ces modèles sont entraînés sur l'ensemble de données d'entraînement (Figure 3.21).

```
def cal(model):
    model.fit(x_train,y_train)
    pre=model.predict(x_test)
    accuracy=accuracy_score(pre,y_test)
    precision=precision_score(pre,y_test)
    recall=recall_score(pre,y_test)
    f1=f1_score(pre,y_test)
    resultat_n01.append(accuracy)
    resultat_n02.append(precision)
    resultat_n03.append(recall)
    resultat_n04.append(f1)
    sns.heatmap(confusion_matrix(pre,y_test), annot=True)
    print(model)
    print('accuracy is :',accuracy,'precision is:',precision,'recall is :',recall, 'f1 is:', f1)
cal(modele_n01)
```

Figure 3.21 Code d'Entraînement des modèles.

4.7 Evaluation des modèles

Après l'entraînement des différents modèles, dans cette étape nous avons exécuté chaque modèle sur l'ensemble de données de test (20% du Dataset) et nous avons comparé les résultats de classification obtenus aux résultats attendus. Cette phase implique, en fait, l'évaluation de la qualité de prédiction des différents modèles à travers la création de la

matrice de confusion et la mesure des métriques de performances des différents modèles utilisés.

➤ Matrice de confusion

La matrice de confusion est un outil essentiel pour évaluer la qualité du modèle et sa capacité de prédiction. Elle offre une vue détaillée des performances de classification en comparant les prédictions du modèle aux valeurs réelles des données.

➤ Mesures d'évaluation des performances

Pour évaluer la performance du modèle de classification et obtenir des indications sur sa capacité à prédire correctement les classes, il est crucial de considérer les mesures suivantes : **Précision, Rappel, F-mesure (F1-score) et Exactitude (Accuracy)**. Ces mesures doivent être analysées dans le contexte spécifique du problème et de l'ensemble de données pour fournir une évaluation précise de la performance du modèle.

- **Précision (Precision) :** La précision mesure la proportion d'instances positives parmi les prédictions positives faites par le modèle.

Formule de calcul :

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

Tel que :

TP est le nombre de vrais positifs (observations correctement prédites comme positives),

FP est le nombre de faux positifs (observations incorrectement prédites comme positives).

- **Rappel (Recall) :** Le rappel mesure la proportion d'instances positives que le modèle parvient à identifier parmi toutes les instances réellement positives.

Formule de calcul :

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

- **F-mesure (F1-score) :** La F-mesure combine la précision et le rappel en une seule métrique qui représente la moyenne des deux. Cette métrique fournit une mesure globale de la performance du modèle en tenant compte à la fois des vrais positifs et des faux positifs.

Formule de calcul :

$$\text{F1-score} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}$$

- **Exactitude (Accuracy)** : L'exactitude mesure la proportion d'observations correctes parmi toutes les prédictions.

Formule de calcul :

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP}$$

Les figures suivantes montrent les résultats de test des différents modèles déjà entraînés, à savoir les modèles Random Forest, DecisionTree, KNN, LogisticRegression et Gradient Boosting. Sur chaque figure l'évaluation de chaque modèle est effectuée à travers la visualisation de la matrice de confusion et des métriques de performances de chaque modèle.

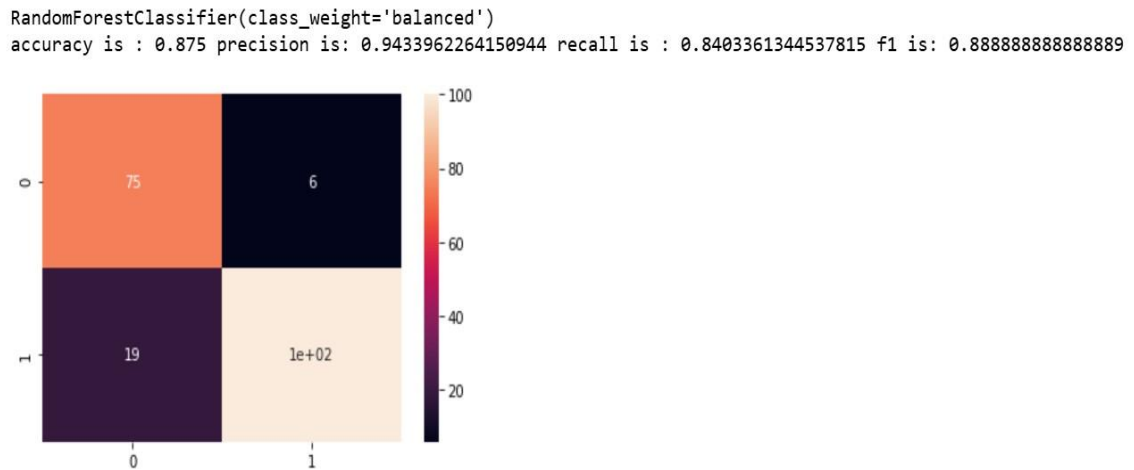


Figure 3.22 : Evaluation du modèle Random Forest.

DecisionTreeClassifier()
accuracy is : 0.825 precision is: 0.8867924528301887 recall is : 0.8034188034188035 f1 is: 0.8430493273542601

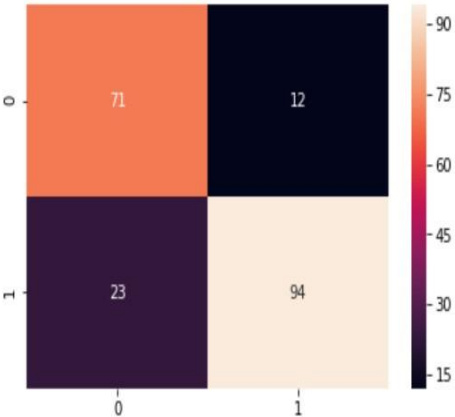


Figure 3.23 : Evaluation du modèle DecisionTree.

KNeighborsClassifier()
accuracy is : 0.76 precision is: 0.8207547169811321 recall is : 0.75 f1 is: 0.7837837837837837

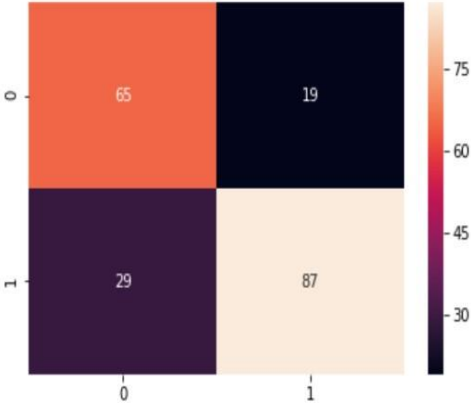


Figure 3.24 : Evaluation du modèle KNN.

```
LogisticRegression()
accuracy is : 0.735 precision is: 0.6792452830188679 recall is : 0.7912087912087912 f1 is: 0.7309644670050762
```

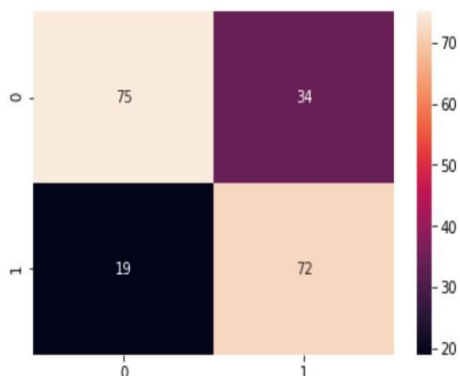


Figure 3.25 : Evaluation du modèle Logistic Regression.

```
GradientBoostingClassifier(n_estimators=1000)
accuracy is : 0.86 precision is: 0.9150943396226415 recall is : 0.8362068965517241 f1 is: 0.8738738738738739
```



Figure 3.26 : Evaluation du modèle Gradient Boosting.

4.8 Sélection du modèle

Afin de sélectionner le modèle que nous allons utiliser au sein de notre système de prédiction, nous avons comparé les différents modèles sur la base des critères clés suivantes : l'exactitude (Accuracy), le rappel, la précision, et le F1-score. Le tableau ci-dessous illustre les résultats des différents algorithmes utilisés.

L'Accuracy est l'une des mesures de performance les plus importantes pour la classification. Cette mesure permet d'estimer plus précisément la qualité du système proposé. Comme nous pouvons le constater, les meilleurs résultats sont obtenus par l'algorithme Random Forest.

D'après la table 3.2, le modèle Random Forest atteint une précision de 0,94 et une Exactitude (accuracy) de 0,87. Il convient de souligner que ces valeurs sont remarquables et indiquent une capacité de prédiction précise. Cela signifie que le modèle est capable de classer correctement presque tous les cas de maladies de diabète dans notre jeu de données.

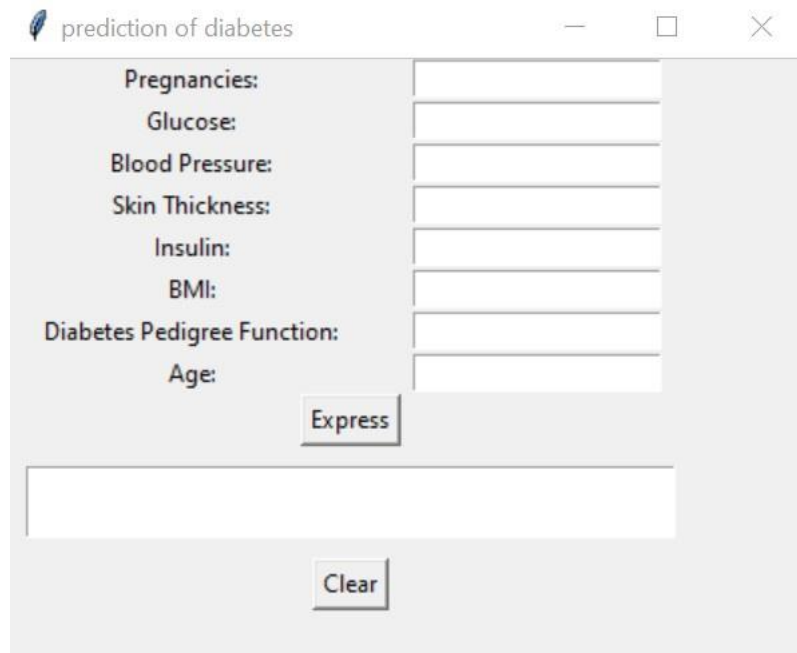
Algorithmes	Accuracy	Precision	Recall	F1-Score
Random Forest	0.875	0.943	0.840	0.888
DecisionTree	0.825	0.886	0.803	0.843
KNN	0.76	0.82	0.75	0.78
LogisticRegression	0.735	0.679	0.791	0.730
Gradient Boosting	0.86	0.91	0.83	0.87

Table 3.2: Métriques de performance des modèles.

Ces résultats nous amènent à choisir le modèle Random Forest pour la prédiction du diabète au sein du système de prédiction que nous avons développé.

5. Interfaces du système développé

Le système de prédiction que nous avons développé, comporte une fenêtre principale (Figure 3.27) qui permet à l'utilisateur de remplir un formulaire. Ce formulaire contient des informations sur le patient dont nous prévoyons de prédire l'état.



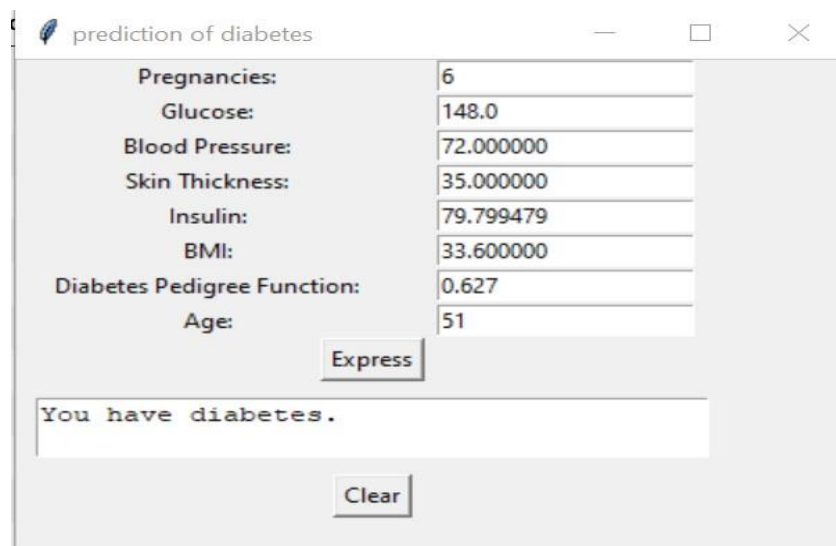
Pregnancies:	
Glucose:	
Blood Pressure:	
Skin Thickness:	
Insulin:	
BMI:	
Diabetes Pedigree Function:	
Age:	

Express

Clear

Figure 3.27 : Interface du système.

Pour introduire les informations nécessaires à la prédiction du diabète, l'utilisateur remplit un formulaire. En cliquant sur le bouton 'Express' la prédiction est lancée. Pour effacer les données du formulaire, il suffit de cliquer sur le bouton "Clear". Les résultats de la prédiction sont affichés, comme illustré sur les figures 3.28 et 3.29.



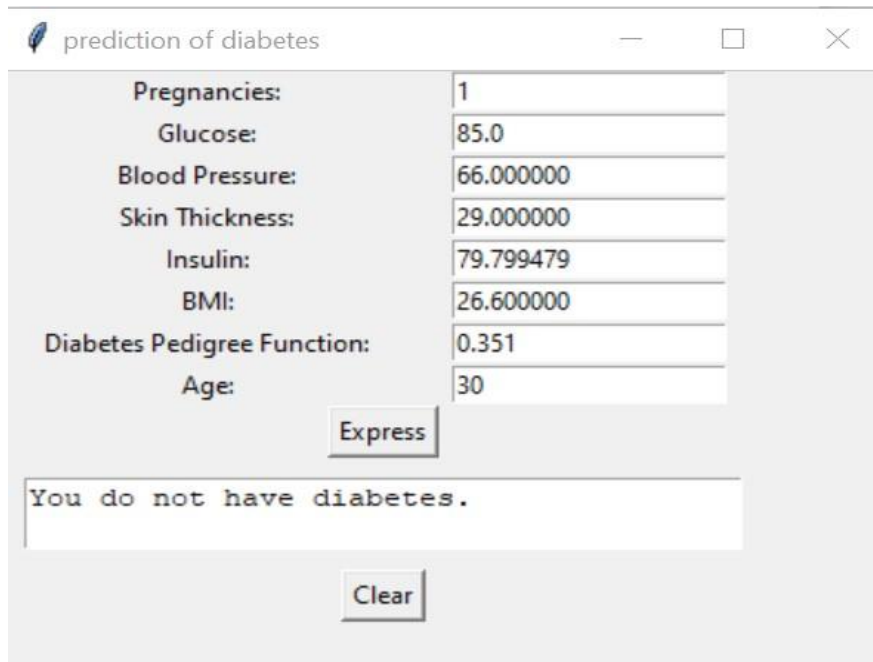
Pregnancies:	6
Glucose:	148.0
Blood Pressure:	72.000000
Skin Thickness:	35.000000
Insulin:	79.799479
BMI:	33.600000
Diabetes Pedigree Function:	0.627
Age:	51

Express

Clear

You have diabetes.

Figure 3.28 :Résultat de prédiction d'une personne diabétique.



Pregnancies:	1
Glucose:	85.0
Blood Pressure:	66.000000
Skin Thickness:	29.000000
Insulin:	79.799479
BMI:	26.600000
Diabetes Pedigree Function:	0.351
Age:	30

Express

You do not have diabetes.

Clear

Figure 3.29 :Résultat de prédiction d'une personne non diabétique.

Grâce à notre modèle d'apprentissage basé sur l'algorithme Random Forest, nous sommes en mesure d'identifier les individus à risque d'avoir le diabète avant même l'apparition des symptômes. Cette approche permet de mettre en place des mesures préventives, comme des ajustements du mode de vie ou un traitement médical, afin de diminuer le risque de complications graves.

6. Conclusion

Dans ce chapitre, nous avons présenté les différentes phases que nous avons suivi pour la réalisation de notre système de prédiction. Ce système utilise le modèle de Random Forest, qui représente le meilleur modèle de prédiction parmi les modèles que nous avons implémenté. A la fin du chapitre, nous avons présenté quelques exemples de prédiction fournis par notre système dans les cas des patients diabétique et non diabétique.

Conclusion Générale

Conclusion générale

La mise en place d'un système de prédiction du diabète représente une avancée significative dans le domaine de la santé préventive et de la médecine personnalisée. En exploitant les données cliniques, ces systèmes permettent d'identifier précocement les individus à risque élevé de développer le diabète, offrant ainsi plusieurs avantages majeurs pour la gestion de la santé publique et individuelle.

Dans notre étude, nous avons développé un système de prédiction du diabète basé sur le modèle Random Forest, reconnu pour sa haute précision de 94% et avec exactitude de 87%. Cette prédiction précoce facilite une intervention préventive ciblée. En identifiant les personnes susceptibles de développer le diabète, les professionnels de la santé peuvent recommander des ajustements du mode de vie tels que des modifications alimentaires, une augmentation de l'activité physique, voire des programmes de gestion du poids, afin de réduire le risque de maladie. Ces interventions précoces sont non seulement efficaces mais aussi économiquement avantageuses, car elles peuvent potentiellement diminuer les coûts associés aux soins de santé liés aux complications du diabète.

Les systèmes de prédiction du diabète contribuent à personnaliser les stratégies de prévention et de gestion, optimisant ainsi les résultats cliniques et améliorant la qualité de vie des patients. Cette avancée prometteuse dans le domaine de la santé publique transforme l'approche de la prévention et de la gestion des maladies chroniques. Une intégration appropriée de ces systèmes dans les pratiques cliniques pourrait considérablement améliorer les résultats de santé et réduire la charge globale de la maladie, marquant ainsi une étape significative vers une médecine plus proactive et personnalisée.

Références bibliographiques

- [1] International Diabetes Federation. (2021). *IDF Diabetes Atlas, 10th Edition, 2021* [PDF]. [https://diabetesatlas.org/idfawp/resource-files/2021/07/IDF Atlas 10th Edition 2021.pdf](https://diabetesatlas.org/idfawp/resource-files/2021/07/IDF_Atlas_10th_Edition_2021.pdf)
- [2] Masson, E. (s. d.-d). *BAROMÈTRE Algérie : enquête nationale sur la prise en charge de spersonnes diabétiques*. EM-Consulte. <https://www.emconsulte.com/article/1285339/barometre-algerie-enquete-nationale-sur-la-prise-e>
- [3] Brutsaert, E. F. (2023, 5 octobre). *Diabète sucré*. Édition Professionnelle du Manuel MSD. <https://www.msmanuals.com/fr/professional/troubles-endocriniens-et-m%C3%A9taboliques/diab%C3%A8te-sucr%C3%A9-troubles-du-m%C3%A9tabolisme-glucidique/diab%C3%A8te-sucr%C3%A9>
- [4] Lactalis Ingredients. (2021, 27 mai). *Quels rôles peuvent jouer les protéines sur le diabète de type II ? - Lactalis Ingredients*. <https://www.lactalisingredients.com/fr/news/blog/quels-roles-peuvent-jouer-les-proteines-sur-le-diabete-de-type-ii/>
- [5] Sahnine, N., & Yahiaoui, Y. (2018). *Analyse des moyens à mettre en œuvre pour lutter contre le diabète: Cas CHU l'hôpital belloua Tizi-Ouzou* (Doctoral dissertation, Université Moul d Mammeri).
- [6] Brutsaert, E. F., & Msd, M. (2023, 5 octobre). *Diabète sucré (DS)*. Manuels MSD Pour le Grand Public, https://www.msmanuals.com/fr/accueil/troubles-hormonaux-et-m%C3%A9taboliques/diab%C3%A8te-sucr%C3%A9-ds-et-troubles-du-m%C3%A9tabolisme-de-la-glyc%C3%A9mie/diab%C3%A8te-sucr%C3%A9-ds#Symt%C3%B4mes_v772851_fr
- [7] Limited, A. (s. d.-b). *Complication diabète Banque d'images vectorielles - Alamy*. Alamy. <https://www.alamyimages.fr/photos-images/complication-diab%C3%A8te.html?imgt=8&sortBy=relevant>
- [8] De la Santé Publique du Canada, A. (2023, 28 décembre). *Diabète : Prévention et facteurs de risque*. Canada.ca. <https://www.canada.ca/fr/sante-publique/services/maladies-chroniques/diabete/prevention-facteurs-risque.html>
- [9] Hypnotized. (s. d.-b). *Le diabète est une maladie chronique*. Association Belge du Diabète. <https://www.diabete.be/le-diabete-2/diabete-10#gsc.tab=0>
- [10] *Qu'est-ce que l'apprentissage automatique ? Le guide ultime par Acronis*. (2021, 24 juin). [Vidéo]. Acronis. <https://www.acronis.com/fr-fr/blog/posts/machine-learning/>

- [11] HAMOUR, M., BENHAMDINE, N. (2020). Prédiction du Churn Rate Par le Machine Learning dans le secteur des M&A Application au sein de KPMG
- [12] L, Gimazane. Les différents algorithmes de l'IA Suite de "Les différents types d'IA"[Présentation PowerPoint]. <https://dane.daneteach.fr/wp-content/uploads/Les-differents-algorithmes-de-IIA.pdf>
- [13] *Construire un arbre de décision - myMaxicours.* (s. d.). myMaxicours. <https://www.maxicours.com/se/cours/construire-un-arbre-de-decision/>
- [14] Data Analytics Post. (2024, 8 février). *K-Means - Data Analytics Post.* <https://dataanalyticspost.com/Lexique/k-means/>
- [15] Bothra, R. (2021). Diabetes Prediction Using Machine Learning Algorithms. *International Journal of Engineering Applied Sciences and Technology*, 6(5), 2455-2143.
- [16] Dutta, D., Paul, D., & Ghosh, P. (2018, November). Analysing feature importances for diabetes prediction using machine learning. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 924-928). IEEE.
- [17] Jakka, A., & Vakula Rani, J. (2019). Performance evaluation of machine learning models for diabetes prediction. *J. Innov. Technol. Explor. Eng.(IJITEE)*, 8(11), 10-35940.
- [18] Saranya, G., & Pande, S. D. (2024). Enhancing Diabetes Prediction with Data Preprocessing and various Machine Learning Algorithms. *EAI Endorsed Transactions on Internet of Things*, 10.
- [19] Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., ... & Abbas, Z. (2019). A model for early prediction of diabetes. *Informatics in Medicine Unlocked*, 16, 100204.
- [20] Farooq, N. A., & Ritika, A. T. (2018). Prediction model for diabetes mellitus using machine learning technique. *Int. J. Comput. Sci. Eng*, 6(03).
- [21] Rajagopal, A., Jha, S., Alagarsamy, R., Quek, S. G., & Selvachandran, G. (2022). 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.
- [22] Nguyen, B. P., Pham, H. N., Tran, H., Nghiem, N., Nguyen, Q. H., Do, T. T., ... & Simpson, C. R. (2019). Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. *Computer methods and programs in biomedicine*, 182, 105055.

[23] Das, P., & Nanda, S. (2021). An Improved Ridge Regression-Based Extreme Learning Machine for the Prediction of Diabetes. In *Proceedings of International Conference on Communication, Circuits, and Systems: IC3S 2020* (pp. 541-547). Springer Singapore.

[24] Vidhya, K., & Shanmugalakshmi, R. (2020). Deep learning based big medical data analytic model for diabetes complication prediction. *Journal of Ambient Intelligence and Humanized Computing*, 11(11), 5691-5702.

[25] Contributeurs aux projets Wikimedia. (2022, 31 octobre). *Anaconda (distribution Python)*. <https://fr.wikipedia.org/wiki/Anaconda> (distribution Python)

[26] Logo Anaconda Python, HD PNG Download, Transparent PNG Image –PNGItem. (s.d-b).PNGItem.com.https://www.pngitem.com/middle/loiwoRo_logo-anaconda-python-hd-png-download/

[27] Contributeurs aux projets Wikimedia. (2024a, février 15). *Jupyter*. <https://fr.wikipedia.org/wiki/Jupyter>

[28] Contributeurs aux projets Wikimedia. (2024c, mai 15). *Python (langage)*. <https://fr.wikipedia.org/wiki/Python> (langage)

[29] Fichier : Python-logo-notext.svg — Wikipédia. (s. d.). <https://fr.m.wikipedia.org/wiki/Fichier:Python-logo-notext.svg>

[30] Contributeurs aux projets Wikimedia. (2024c, avril 26). *Pandas*. <https://fr.wikipedia.org/wiki/Pandas>

[31] IcePanel Technologies Inc. (s. d.). *Pandas SVG and transparent PNG icons | TechIcons*. TechIcons. <https://techicons.dev/icons/pandas>

[32] Contributeurs aux projets Wikimedia. (2024b, février 12). *NumPy*. <https://fr.wikipedia.org/wiki/NumPy>

[33] Contributeurs aux projets Wikimedia. (2024d, avril 23). *Scikit-learn*. <https://fr.wikipedia.org/wiki/Scikit-learn>

[34] Contributeurs aux projets Wikimedia. (2024a, février 12). *Matplotlib*. <https://fr.wikipedia.org/wiki/Matplotlib>

[35] File :Mpl screenshot figures and code.png — Wikimedia Commons. (2016, 26 septembre). https://commons.wikimedia.org/wiki/File:Mpl_screenshot_figures_and_code.png?uselang=fr

[36] *An introduction to seaborn — seaborn 0.13.2 documentation.* (s. d.-b). <https://seaborn.pydata.org/tutorial/introduction.html>

[37] *Citing and logo — seaborn 0.13.2 documentation.* (s. d.). <https://seaborn.pydata.org/citing.html>

[38] *Contributeurs aux projets Wikimedia.* (2024h, mai 12). *Tkinter.* <https://fr.wikipedia.org/wiki/Tkinter>

[39] Robert, J. (2023c, novembre 9). *Joblib : Quelle est cette bibliothèque Python ? Comment l'utiliser ?* Formation Data Science | DataScientest.com. <https://datascientest.com/joblib-tout-savoir>

[40] Plotly. (s. d.). *datasets/diabetes.csv at master · plotly/datasets.* GitHub. <https://github.com/plotly/datasets/blob/master>