****

**UNIVERSITE INTERNATIONALE PRIVEE**

**D’EXCELLENCE**



**Département Informatique**

**Mémoire en vue de l’obtention du grade de licence en Informatique. Spécialité : Ingénierie du Logiciel et des données. Année Universitaire 2023 – 2024**

**Prédiction du diabète à l'aide de l'apprentissage automatique**

**Encadre par :**

M. Solemane

Coulibaly

**Présenté par :**

M. Djibrael Traoré

Soutenu le 13 Septembre 2024, devant les membres du jury composé de :

Pr Daouda TRAORE (Président du jury)

Dr Ismaël SISSOKO (Membre du jury)

Dr Solemane COULIBALY (Directeur de mémoire)

Promotion 2021 - 2024

# **Dédicace**

Je dédie ce travail à mon père ***Mamadou Traoré***. Vous m'avez toujours soutenu et encouragé dans mes études.

# **Remerciement**

Je tiens à exprimer ma profonde gratitude à toutes les personnes qui ont contribué à la réalisation de ce mémoire. Tout d'abord, je remercie chaleureusement mon directeur de recherche, (***Mr Soulemane Coulibaly***), pour son encadrement précieux, ses conseils avisés et son soutien indéfectible tout au long de ce projet.

Je souhaite également remercier l'ensemble du corps professoral et administratif de (***UIE***) pour leur accompagnement et leurs enseignements qui ont été déterminants dans l'aboutissement de ce travail.

Un merci spécial à mes collègues et amis, dont les échanges constructifs et les encouragements ont été une source constante de motivation.

Je suis particulièrement reconnaissant envers ma famille pour leur soutien inconditionnel, leur patience et leur compréhension durant ces mois de travail intense. Leur confiance en moi a été une force inestimable.

Enfin, je tiens à remercier tous les participants et professionnels de la santé qui ont contribué par leur expertise et leur disponibilité à enrichir ce mémoire, notamment dans le contexte de l'étude du diabète. Leur contribution a été essentielle à la réussite de ce projet.

À toutes et à tous, merci infiniment.

# **Liste des abréviations**

|  |  |
| --- | --- |
| LR | Logistique Régression |
| KNN | K-Nearest Neighbors |
| ROC | Receiver Operating Characteristic |
| NB | NaïveBayes |
| SVC | Support Vector Classifie |
| DT | Décision Tree |
| RF | Random Forest |

# **Liste des figures et liste des tableaux**

[Figure 1: Les différentes types du machine learning 10](#_Toc180159932)

[Figure 2: Random Forest schéma 18](#_Toc180159933)

[Figure 3: Données d'analyses 23](#_Toc180159934)

[Figure 4: La distribution en pourcentage et en nombre 25](#_Toc180159935)

[Figure5: Histogramme des colonnes 26](#_Toc180159936)

Figure 6 : Matrice de corrélation……………...……………………………………………….27

[Figure 7: Interface du teste 34](#_Toc180159937)

**Tableaux**

[Tableau 1: Description des variables d’ensemble de données 22](#_Toc176532607)

[Tableau 2: Evaluation des différents modèles 27](#_Toc176532608)

# **Sommaire**

[Dédicace I](#_Toc176539042)

[Remerciement II](#_Toc176539043)

[Liste des abréviations III](#_Toc176539044)

[Liste des figures et liste des tableaux IV](#_Toc176539045)

[Sommaire V](#_Toc176539046)

[Introduction générale 1](#_Toc176539047)

[Contexte et problématique 1](#_Toc176539048)

[Objectifs 2](#_Toc176539049)

[ Objectif général 2](#_Toc176539050)

[ Objectifs spécifiques 2](#_Toc176539051)

[Organisation du mémoire 2](#_Toc176539052)

[Chapitre 1 : Définition des concepts et étude de lien entre les concepts. 4](#_Toc176539053)

[Section 1 : Définition des concepts. 4](#_Toc176539054)

[Section 2 : Relation entre l'Analyse Prédictive et le Diabète. 6](#_Toc176539064)

[Chapitre 2 : Méthodologie de la recherche. 7](#_Toc176539069)

[Section 1 : Choix de la méthodologie à adapter. 7](#_Toc176539070)

[Section 2 : Techniques d’analyse prédictive. 8](#_Toc176539085)

[Chapitre 3 : Présentation et Analyse des données. 21](#_Toc176539123)

[Section 1 : Présentation des données. 21](#_Toc176539124)

[Section 2 : Analyse des données. 23](#_Toc176539128)

[Chapitre 4 : Réalisation de l’interface du teste. 29](#_Toc176539141)

[Section 1 : Recommandation. 29](#_Toc176539142)

[Section 2 : Analyse d’amélioration. 33](#_Toc176539148)

[ Conclusion générale et perspective 35](#_Toc176539155)

[ Bibliographie 36](#_Toc176539156)

# **Introduction générale**

# **Contexte et problématique**

À l'ère de la médecine moderne, l'analyse prédictive émerge comme un outil inestimable pour révolutionner la gestion des maladies chroniques telles que le diabète. Cette affection, caractérisée par une hyperglycémie chronique, impose un fardeau considérable sur les individus et les systèmes de santé du monde entier. La capacité de prédire avec précision les risques de développement du diabète et ses complications potentielles représente une avancée majeure dans la lutte contre cette épidémie silencieuse. En intégrant des techniques d'analyse prédictive, les professionnels de la santé peuvent désormais élaborer des stratégies personnalisées de prévention et de traitement, réduisant ainsi les risques de morbidité et de mortalité associés à cette maladie, offrant une perspective d'avenir où la technologie et la médecine s'entrelacent pour améliorer la santé et le bien-être des patients.

L’analyse prédictive est un outil puissant pour la gestion du diabète. Elle permet d'identifier les personnes à risque de développer la maladie grâce à l'analyse des données sur les facteurs de risque.

Une fois le diabète concerné, l'analyse prédictive aide à personnaliser le traitement. En étudiant la réponse du patient aux différents traitements, les médecins peuvent ajuster le plan de soins pour optimiser le contrôle de la glycémie et réduire les complications.

L'analyse prédictive est également précieuse pour la recherche sur le diabète. Elle aide à concevoir des essais cliniques plus ciblés et efficaces, en identifiant les sous-groupes de patients qui répondent mieux à certains traitements.

Cependant, l'utilisation de ces technologies soulève des questions éthiques sur la confidentialité des données et l'équité d'accès. Il est essentiel de mettre en place des politiques pour garantir une utilisation responsable de l'analyse prédictive dans les soins de santé.

# **Question principale**

Comment les algorithmes d’apprentissage automatique peuvent-ils améliorer la précision et l’efficacité de la prédiction du diabète par rapport aux méthodes traditionnelles ?

### Questions spécifiques

* Quels sont les facteurs de risque les plus pertinents dans la prédiction du diabète ?
* Quels algorithmes d’apprentissage automatique sont les plus performants pour la prédiction précoce du diabète ?
* Comment l’intégration de facteurs liés au mode de vie (alimentation, exercice physique) améliore-t-elle la prédiction du diabète à l’aide de modèles d’apprentissage automatique ?

# **Objectifs**

# **Objectif général**

Comment développer et implémenter un modèle d’analyse prédictive pour améliorer la détection précoce et la gestion du diabète, en utilisant des techniques de machine Learning et des données médicales pertinentes ?

# **Objectifs spécifiques**

1. Identifier les facteurs de risque prédictifs du diabète.
2. Développer et valider un modèle prédictif de détection précoce du diabète.
3. Améliorer l’analyse clinique du diabète ?

# **Organisation du mémoire**

Pour élaboration de ce mémoire, il a été organisé en quatre chapitres de la façon suivante :

Dans **le premier chapitre,** nous évoquerons la définition des différents concepts clés de notre étude, et le lien de relation entre ces concepts.

Dans **le deuxième chapitre,** tout d’abord, nous allons parler de la méthodologie de recherche notamment le choix de la méthodologie à adopter telle que l’approche quantitative, la collecte des données et le prétraitement des données.

Nous parlerons des Méthodes et des techniques d’analyse prédictive pour notre étude.

Dans **le troisième chapitre,** nous aborderons la présentation et l’analyse des données et aussi la provenance des données et la description des données.

Pour la partie analyse des données nous préparerons les données pour l’analyse et ensuite nous allons appliquer des modelés prédictifs.

Après avoir appliqué les modèles prédictifs nous allons à atteindre des comme : (Régression logistique, Arbre de décision, réseau de neurones).

Dans **le quatrième chapitre,** nous allons aborder les critiques et suggestions concernant l'utilisation de l'analyse prédictive pour la gestion du diabète. Nous soulignerons les limites des données utilisées et des modèles prédictifs, puis propose des améliorations, notamment en termes de collecte de données plus complètes et d'optimisation des algorithmes prédictifs. Enfin, nous ouvrirons des perspectives futures pour la recherche dans ce domaine, en mettant l'accent sur l'importance d'une approche éthique et responsable dans l'utilisation de ces technologies de pointe en santé.

Et enfin le document se termine par une conclusion générale.

**PREMIÈRE PARTIE : CADRE CONCEPTUEL DE L’ETUDE.**

# : Définition des concepts et étude de lien entre les concepts.

# **Section 1 : Définition des concepts.**

# **Introduction**

Dans un monde de plus en plus dominé par les données, l'analyse prédictive s'impose comme un outil essentiel pour anticiper les évolutions et prendre des décisions éclairées. Elle repose sur l'exploitation de données historiques et l'application de modèles statistiques et algorithmiques pour prévoir les résultats futurs. Grâce à cette approche, de nombreux secteurs, notamment celui de la santé, peuvent non seulement réagir face aux défis actuels, mais aussi prévoir les tendances et les évolutions à venir.

## **Analyse prédictive**

L’analyse prédictive (ou logique prédictive) est la technique analytique et statistique qui, en utilisant à la fois des données actuelles et historiques, permet de créer hypothèse et prédictions sur des événements futurs.

# **Machine Learning**

Le Machine Learning peut être défini comme étant une technologie d’intelligence artificielle permettant aux machines d’apprendre sans avoir été au préalablement programmées spécifiquement à cet effet. Le Machine Learning est explicitement lié au Big Data, étant donné que pour apprendre et se développer, les ordinateurs ont besoin de flux de données à analyser et sur lesquelles s’entraîner.  ***Source****:* [*https://ia-data-analytics.fr/machine-learning/*](https://ia-data-analytics.fr/machine-learning/)

# **Diabète**

Le diabète est une maladie chronique qui apparaît lorsque le pancréas ne produit pas suffisamment d’insuline ou que l’organisme n’utilise pas correctement l’insuline qu’il produit. L’insuline est une hormone qui régule la concentration de sucre dans le sang. L’hyperglycémie, ou concentration sanguine élevée de sucre, est un effet fréquent du diabète non contrôlé qui conduit avec le temps à des atteintes graves de nombreux systèmes organiques et plus particulièrement des nerfs et des vaisseaux sanguins.

En 2014, 8,5% de la population adulte (18 ans et plus) était diabétique. En 2015, le diabète a été la cause directe de 1,6 million de décès et en 2012 l'hyperglycémie avait causé 2,2 millions de décès supplémentaires. ***Source***: [OMS - Diabète](https://www.who.int/fr/news-room/fact-sheets/detail/diabetes). Publiée par l'OMS le 10 novembre 2021.

# **Types de diabète**

# **Diabètes de type\_1**

Le diabète de type-1 du a une absence d’insuline par le pancréas. Les Diabétiques de type-1 doivent s’injecter quotidiennement de l’insuline.

# **Diabètes de types\_2**

Le diabète de type 2 est dû à une résistance à l’insuline entraînant Une carence insulinique à terme. Les techniques d’exploration de données ont supplanté les méthodologies existantes en offrant une meilleure prédiction, une meilleure exactitude et une meilleure précision. De plus, l’apprentissage automatique est une technologie de l’intelligence artificielle qui apprend les relations entre les nœuds sans formation préalable.

# **Diabètes gestationnels**

Il se caractérise par une hyperglycémie, c’est-à-dire une élévation de la teneur en sucre du sang, avec des valeurs supérieures à la normale, mais inférieures à celles posant le diagnostic de diabète, apparaissant pendant la grossesse.

Les femmes ayant un diabète gestationnel ont un risque accru de complications pendant la grossesse et à l’accouchement. Leur risque ainsi que celui de leur enfant, d’avoir un diabète de type II à un stade ultérieur de leur vie augmente également.

# **Symptômes de la maladie**

Le diabète se manifeste par une soif excessive, une urination fréquente, une fatigue constante, une perte de poids inexpliquée et une vision floue. D'autres symptômes incluent une guérison lente des plaies, des infections fréquentes, ainsi qu'un engourdissement ou des picotements dans les mains et les pieds. La progression des symptômes peut être rapide dans le diabète de type 1, alors qu'elle est souvent plus lente et discrète dans le diabète de type 2. Le diabète gestationnel est souvent asymptomatique, mais peut être détecté par des tests de glycémie durant la grossesse.

## **Section 2 : Relation entre l'Analyse Prédictive et le Diabète.**

L'analyse prédictive joue un rôle de plus en plus important dans la gestion et la prévention du diabète. Nous allons vous montrer les détails :

# **Prédiction de la progression de la maladie**

Cette prédiction permet aux patients et aux professionnels de santé de prendre des mesures préventives pour ralentir la progression de la maladie.

### Détection précoce des complications

Une détection précoce permet d'intervenir plus rapidement et d'améliorer les chances de réussite du traitement.

# **Optimisation de la gestion du diabète**

Elle permet de prédire le niveau de glycémie, de personnaliser les doses d'insuline et de fournir des conseils nutritionnels et d'exercice adaptés.

# **Conclusion**

La conclusion de la définition des concepts d'analyse prédictive, de diabète et de machine Learning montre que l'intégration de ces technologies dans le domaine de la santé permet d'anticiper l'évolution des maladies chroniques comme le diabète. L'analyse prédictive, soutenue par la machine Learning, exploite les données pour identifier des schémas et des tendances, améliorant ainsi la prise de décision clinique. Cela permet de personnaliser les traitements et d'intervenir de manière proactive, réduisant ainsi les risques pour les patients. Ces approches innovantes transforment la gestion du diabète et d'autres pathologies, ouvrant la voie à une médecine plus précise et efficace.

# **: Méthodologie de la recherche.**

# **Section 1 : Choix de la méthodologie à adapter.**

# **Introduction**

L'introduction de la méthodologie de recherche présente les approches et les outils utilisés pour répondre aux questions de recherche et atteindre les objectifs fixés. Dans le cadre de cette étude, nous avons adopté une méthodologie mixte, combinant l'analyse quantitative et qualitative pour une compréhension approfondie du sujet. L'analyse des données, notamment par l'application des techniques de machine Learning, sera au cœur de l'investigation, permettant de générer des prédictions basées sur les données historiques des patients atteints de diabète. Cette méthodologie vise à explorer et valider l'efficacité de l'analyse prédictive dans le domaine de la santé.

# **Approche Quantitative**

L'analyse prédictive dans le domaine de la santé, plus précisément dans la gestion du diabète, peut apporter des avantages significatifs pour améliorer les résultats des patients et optimiser les ressources médicales.



### Collecte des données

Par collecte de données, on entend l'approche systématique qui consiste à réunir et à mesurer des informations en provenance de sources variées, afin d'obtenir une vue complète et précise d'un domaine d'intérêt. La collecte des données permet à une personne ou à une entreprise de répondre à des questions pertinentes, d'évaluer des résultats et de mieux anticiper les probabilités et les tendances à venir pour appliquer des techniques d’analyse prédictive dans le domaine de la santé, et en particulier pour le diabète.

# **Prétraitement des données**



Le prétraitement des données consiste à transformer les données brutes en un format plus adapté et plus pertinent pour l'analyse et l'apprentissage des modèles.

# **Nettoyage des données**

Le nettoyage des données consiste à identifier, corriger ou supprimer les erreurs, les valeurs manquantes, les doublons et les incohérences dans les données.

**Suppression des doublons** : Identifier et éliminer les lignes ou les enregistrements en double.

**Gestion des valeurs manquantes** : Imputation des valeurs manquantes ou suppression des lignes/colonnes contenant des valeurs manquantes excessives.

Correction des erreurs : Rectification des erreurs typographiques ou des erreurs de saisie.

# **Transformation des données**

La transformation des données vise à convertir les données brutes en un format approprié pour l'analyse.

### Division des données

Dans cette analyse notre donnée sera divisée en deux grandes parties à savoir :

80% des données sera utilisée pour l’entrainement de la modèle et le restant des 20% sera utilisée pour le teste de notre modèle.

**Données d’entrainement Données de teste**

**20%**

**80%**

# **Section 2 :** **Techniques d’analyse prédictive.**

# **Les différents types d’algorithme du machine Learning**

Le domaine de l’apprentissage automatique est souvent divisé en sous-domaines selon les types de problèmes abordés. On peut les classer approximativement comme suit :

### Apprentissage supervisé

A partir d’une donnée d’entrée, par exemple une photographie avec un panneau de signalisation, la tâche consiste à prédire le résultat ou l’étiquetage correct, par exemple le panneau de signalisation qui se trouve dans l’image (limitation de vitesse, panneau stop, etc.). Dans les cas les plus simples, les réponses se présentent sous la forme de oui/non (on les qualifie alors de problèmes de classification binaire).

### Apprentissage semi supervisé

L’apprentissage semi-supervisé est une technique d’apprentissage automatique qui utilise à la fois des données labellisées et des données non labellisées. Découvrez ci-dessous comment fonctionne cette méthode d’apprentissage.

# **Apprentissage non supervisé**

Il n’y a pas d’étiquetage ou de résultats corrects. La tâche consiste à découvrir la structure des données : par exemple, regrouper des éléments similaires pour former des « grappes », ou réduire les données à un petit nombre de « dimensions » importantes. La visualisation des données peut aussi être considérée comme un apprentissage non supervisé.

# **Apprentissage par renforcement**

Communément utilisé dans les situations où un agent d’IA, comme une voiture autonome, doit fonctionner dans un environnement et où un retour d’information sur les bons ou mauvais choix est disponible avec un certain retard. Également utilisé dans les jeux dont le résultat ne peut être décidé qu’à la fin de la partie.

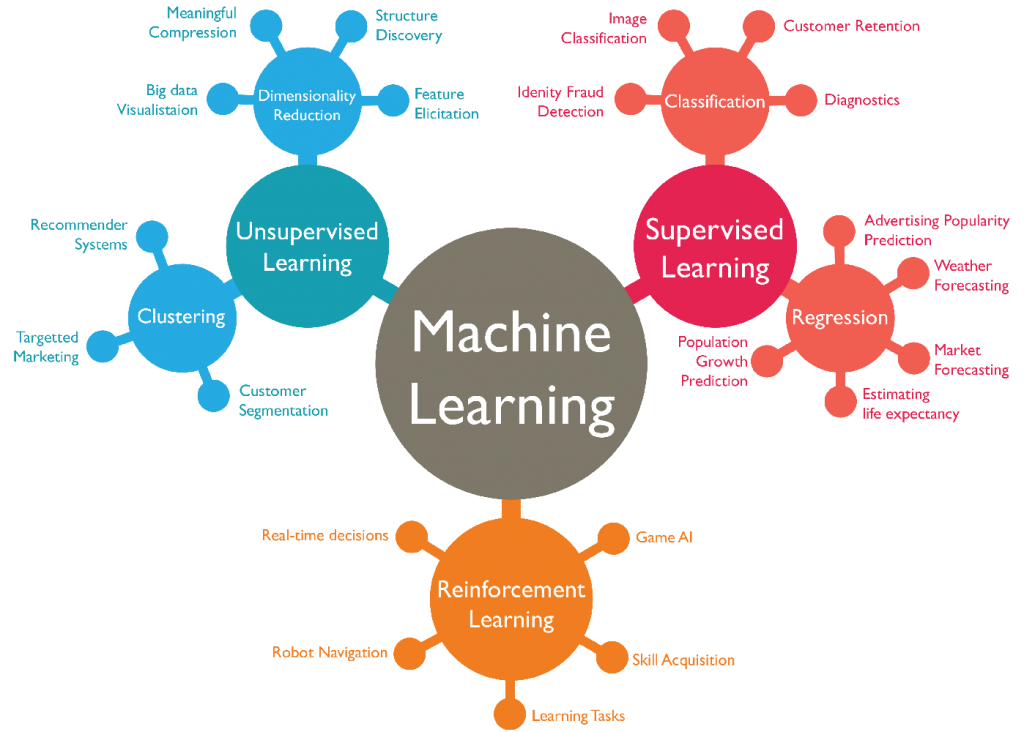


Figure 1: Les différentes types du machine learning

# **Algorithmes de prédiction utilisée**

Les algorithmes de prédiction sont des modèles mathématiques ou des méthodes de machine Learning utilisés pour prédire des résultats futurs en se basant sur des données passées. Ces algorithmes peuvent être appliqués à divers domaines, comme la médecine, la finance, le marketing, et bien d'autres. Ils permettent de faire des prévisions sur des événements, des comportements ou des tendances à partir de patterns identifiés dans les données existantes.

L'analyse prédictive est généralement réalisée selon trois grands types de algorithmes :

### Régression logistique

La régression logistique est une technique d'analyse de données qui utilise les mathématiques pour trouver les relations entre deux facteurs de données. Elle utilise ensuite cette relation pour prédire la valeur de l'un de ces facteurs en fonction de l'autre.

# **Avantages**

**Simplicité et interopérabilité**

La régression logistique est mathématiquement moins complexe que d'autres méthodes de machine Learning. Les coefficients permettent d'estimer l'effet de chaque variable indépendante sur la probabilité de l'événement d'intérêt, facilitant ainsi l'interprétation des résultats.

**Flexibilité**

La régression logistique peut gérer à la fois des variables numériques et catégorielles. Elle peut capturer des relations non linéaires entre les variables indépendantes et la variable dépendante binaire.

**Rapidité et efficacité**

Les modèles de régression logistique peuvent traiter de grands volumes de données rapidement, car ils nécessitent moins de ressources en termes de mémoire et de puissance de calcul.

**Probabilités prédites**

La régression logistique fournit des probabilités prédites pour chaque classe, ce qui permet de quantifier l'incertitude et de prendre des décisions éclairées en fonction du seuil choisi.

# **Inconvénients**

**Hypothèses restrictives**

La régression logistique repose sur des hypothèses comme l'absence de multi colinéarité et l'indépendance des erreurs. Si ces hypothèses ne sont pas respectées, les résultats peuvent être biaisés.

**Sensibilité aux données déséquilibrées**

Lorsque les classes sont très déséquilibrées dans les données d'entraînement, la régression logistique peut avoir tendance à prédire majoritairement la classe majoritaire, ce qui dégrade les performances sur la classe minoritaire.

**Limites pour les relations complexes**

Pour des relations très complexes entre les variables, la régression logistique peut avoir des difficultés à capturer toute la complexité, contrairement à des méthodes plus flexibles comme les arbres de décision ou les réseaux de neurones.

**Sensibilité aux valeurs aberrantes**

La régression logistique peut être sensible aux valeurs aberrantes dans les données d'entraînement, ce qui peut affecter la qualité des prédictions.

En résumé, la régression logistique offre une approche simple, rapide et interprétable pour la classification binaire, mais elle a aussi des limitations en termes d'hypothèses et de flexibilité pour capturer des relations complexes. Son choix dépendra du problème spécifique et des caractéristiques des données.

### K-Nearest Neighbors

L’algorithme K-Nearest Neighbors (KNN) est une méthode d'apprentissage automatique supervisée utilisée pour résoudre les problèmes de classification et de régression. Evelyn Fix et Joseph Hodges ont développé cet algorithme en 1951, qui a ensuite été développé par Thomas Cover.

# **Avantages**

**Simplicité et facilité d'utilisation**

KNN est très simple à comprendre et à mettre en œuvre. Il ne nécessite pas de construction de modèle complexe ou d'hypothèses préalables, ce qui en fait un bon choix pour les débutants.

**Polyvalence**

Cet algorithme peut être utilisé pour des tâches de classification et de régression, ainsi que pour la recherche d'informations, ce qui le rend adaptable à divers types de problèmes.

Pas de phase d'apprentissage

KNN ne nécessite pas de phase d'apprentissage explicite. Il fonctionne directement sur les données d'entraînement, ce qui peut être un avantage dans certaines situations.

# **Inconvénients**

**Sensibilité à la taille des données**

KNN peut devenir très lent à mesure que le volume de données augmente, car il doit calculer la distance entre l'échantillon à prédire et tous les points de données d'entraînement.

**Nécessité de mémoire**

L'algorithme doit conserver toutes les données d'entraînement en mémoire, ce qui peut poser des problèmes de stockage et de gestion des données.

**Choix du paramètre K**

Le choix du nombre de voisins (K) peut affecter considérablement les performances du modèle. Un K trop petit peut rendre le modèle sensible au bruit, tandis qu'un K trop grand peut diluer les informations pertinentes.



### Arbres de décision

Les arbres de décision sont des modèles de classification qui placent les données dans différentes catégories en fonction de variables distinctes. Cette méthode est surtout utile pour tenter de comprendre les décisions d'un individu. Le modèle ressemble à un arbre, où chaque branche représente un choix potentiel, et la feuille de la branche représente le résultat de la décision. Les arbres de décision sont généralement faciles à comprendre et fonctionnent bien lorsqu'il manque plusieurs variables dans un ensemble de données.

# **Avantages**

**Simplicité et interprétabilité**

Les arbres de décision sont faciles à comprendre et à interpréter. Leur structure graphique permet de visualiser les décisions et les conséquences, ce qui les rend accessibles même pour les non-experts.

**Flexibilité**

Ils peuvent traiter différents types de données, qu'elles soient numériques ou catégorielles, ce qui les rend polyvalents pour divers problèmes.

**Peu de préparation des données**

Les arbres de décision nécessitent peu de préparation des données, comme la normalisation ou le traitement des valeurs manquantes, ce qui simplifie leur utilisation.

**Rapidité d'exécution**

Ils sont généralement rapides à entraîner et à exécuter, ce qui les rend efficaces, même avec de grands ensembles de données.

# **Inconvénients**

**Sensibilité aux variations**

Les arbres de décision peuvent être instables ; de petits changements dans les données d'entrée peuvent entraîner des variations significatives dans la structure de l'arbre, affectant ainsi la fiabilité des prédictions.

**Risque de sur apprentissage**

Ils ont tendance à sur ajuster les données d'entraînement, surtout si l'arbre est trop complexe, ce qui peut réduire leur performance sur des données non vues.

**Limitations pour des calculs complexes**

Pour des problèmes très complexes avec de nombreuses variables, les arbres de décision peuvent devenir trop encombrants et difficiles à gérer, offrant un faux sentiment de sécurité.

**Biais potentiel**

Sans réglage approprié, les arbres de décision peuvent être biaisés, surtout si certaines classes dominent dans les données d’entraînement.

### Support Vector Classifie

SVC (Support Vector Classifier) : SVC est une implémentation spécifique de l'algorithme Support Vector Machine conçu spécifiquement pour les tâches de classification. En d’autres termes, SVC est un SVM utilisé pour la classification. Il cherche à trouver l’hyperplan qui sépare le mieux les points de données en différentes classes. Les termes « SVC » et « SVM » sont parfois utilisés de manière interchangeable, mais lorsque quelqu'un fait référence à un « SVC », il fait généralement référence à la variante de classification de l'algorithme.

# **Avantages**

**Précision élevée**

Les SVM sont connus pour leur grande précision dans la classification, surtout dans des espaces de grande dimension. Ils sont efficaces pour trouver des hyperplans qui maximisent la marge entre les classes.

**Efficacité avec des petits ensembles de données**

Ils fonctionnent bien avec des ensembles de données de taille modérée et peuvent être plus efficaces car ils utilisent un sous-ensemble de points d'entraînement (les vecteurs de support) pour effectuer des prédictions.

**Flexibilité avec les noyaux**

Les SVM peuvent utiliser des fonctions noyaux pour traiter des problèmes non linéaires, ce qui leur permet de s'adapter à des structures de données complexes.

**Robustesse aux d’outliers**.

Les SVM sont relativement robustes aux valeurs aberrantes, surtout si le paramètre de régularisation est bien ajusté, ce qui aide à contrôler l'impact des points extrêmes sur le modèle.

# **Inconvénients**

**Temps d'entraînement long**

Les SVM peuvent être lents à entraîner sur de grands ensembles de données, car le calcul de l'hyperplan optimal peut devenir complexe avec l'augmentation du nombre d'exemples.

**Sensibilité au bruit**

Les SVM peuvent être moins efficaces sur des ensembles de données contenant beaucoup de bruit ou de valeurs aberrantes, ce qui peut compliquer la séparation des classes.

**Difficulté à choisir les paramètres**

Le choix des paramètres, comme le type de noyau et les valeurs de régularisation, peut être délicat et nécessite souvent une optimisation minutieuse pour obtenir de bonnes performances.

**Problèmes avec les classes déséquilibrées**

Les SVM peuvent avoir des difficultés à traiter des jeux de données déséquilibrés, ce qui peut entraîner un biais vers la classe majoritaire

### Naïve Bayes

Naïve Bayes fait partie d'une famille d'algorithmes d'apprentissage génératif, ce qui signifie qu'il cherche à modéliser la distribution des entrées d'une classe ou d'une catégorie donnée. Contrairement aux classificateurs discriminatifs, comme la régression logistique, il n'apprend pas quelles caractéristiques sont les plus importantes pour différencier les classes.

# **Avantages**

**Simplicité**

Naïve Bayes est facile à comprendre et à mettre en œuvre. Sa structure simple permet une codification rapide et nécessite peu de paramètres à régler, ce qui le rend accessible aux débutants.

**Rapidité**

Il est très rapide à entraîner et à prédire, même avec de grands ensembles de données, car il repose sur des calculs de probabilités simples.

**Efficacité avec des données textuelles**

Naïve Bayes est particulièrement performant pour des tâches de classification de texte, comme le filtrage de spam et l'analyse de sentiments, où il donne souvent de bons résultats malgré ses hypothèses simplistes.

**Peu de données nécessaires**

Il fonctionne bien même avec des petits ensembles de données, ce qui le rend utile dans des situations où les données sont limitées.

# **Inconvénients**

**Hypothèse d'indépendance**

L'algorithme repose sur l'hypothèse que les caractéristiques sont indépendantes les unes des autres, ce qui est souvent irréaliste dans des contextes réels. Cette hypothèse peut conduire à des prédictions sous-optimales.

**Performances variables**

Naïve Bayes peut-être surpasser par des modèles plus sophistiqués lorsque les relations entre les caractéristiques sont complexes. Dans de tels cas, d'autres algorithmes peuvent offrir de meilleures performances.

**Sensibilité aux données déséquilibrées**

Il peut être moins efficace sur des ensembles de données déséquilibrés, où certaines classes sont beaucoup plus représentées que d'autres.

### Random Forest

Les forêts aléatoires ou forêts de décision aléatoires sont une méthode d'apprentissage d'ensemble pour la classification, la régression et d'autres tâches qui fonctionne en construisant une multitude de arbres de décision au moment de la formation. Pour les tâches de classification, la sortie de la forêt aléatoire est la classe sélectionnée par la plupart des arbres.

# **Avantages**

**Précision élevée**

Les forêts aléatoires combinent les prédictions de plusieurs arbres de décision, ce qui améliore la précision globale et réduit le risque de sur ajustement par rapport à un seul arbre.

**Robustesse aux données bruyantes**

Elles sont résilientes aux valeurs aberrantes et aux données bruyantes, car la combinaison des prédictions de plusieurs arbres permet d'atténuer l'impact des points de données problématiques.

**Capacité à gérer des données complexes**

Les forêts aléatoires peuvent traiter des ensembles de données avec de nombreuses caractéristiques (variables) et sont efficaces pour la classification et la régression, ce qui les rend polyvalentes.

**Importance des caractéristiques**

Cet algorithme permet d'évaluer l'importance des différentes caractéristiques dans le modèle, ce qui aide à comprendre quelles variables influencent le plus les prédictions.

# **Inconvénients**

**Manque d'interprétabilité**

Bien que les forêts aléatoires soient puissantes, elles sont souvent considérées comme des "boîtes noires", car il est difficile d'interpréter les résultats et de comprendre comment les décisions sont prises.

**Temps de calcul**

L'entraînement d'un grand nombre d'arbres peut être coûteux en temps de calcul et en ressources, surtout avec des ensembles de données très volumineux.

**Risque de sur ajustement avec trop d'arbres**

Si le nombre d'arbres est trop élevé, cela peut conduire à un sur ajustement, bien que cela soit généralement moins problématique qu'avec un seul arbre de décision.

**Sensibilité aux déséquilibres de classes**

Les forêts aléatoires peuvent avoir des difficultés avec des ensembles de données déséquilibrés, ce qui peut affecter leur capacité à prédire correctement les classes minoritaires.

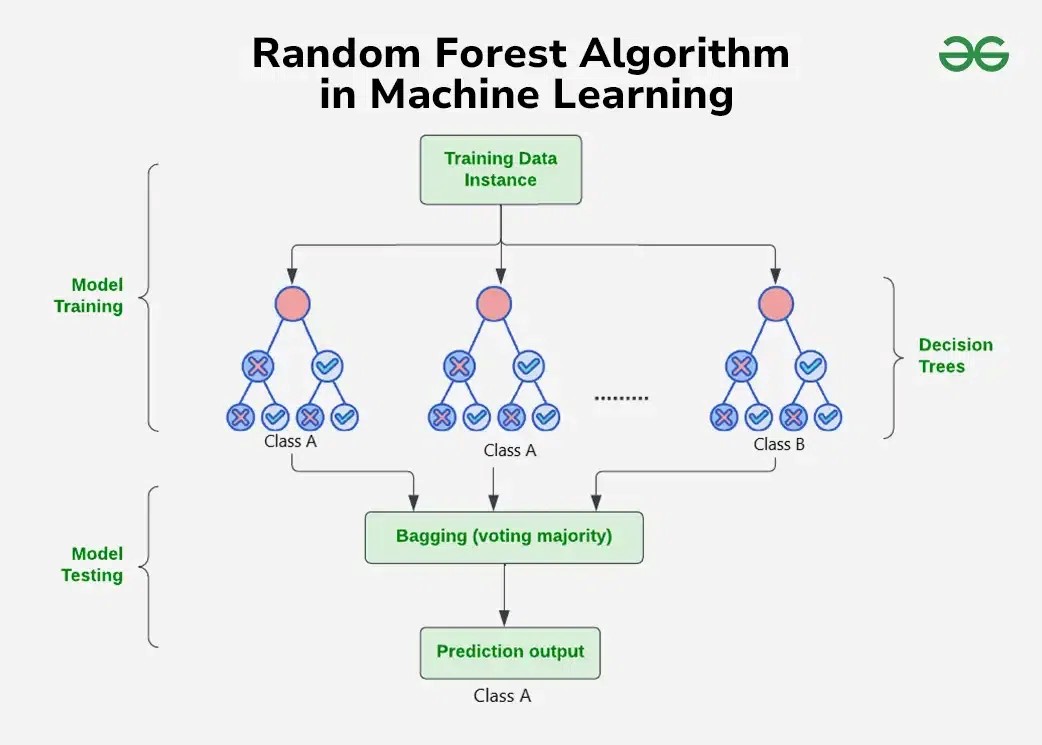


Figure 2: Random Forest schéma

# **Logiciels et Bibliothèques utilisés en machine Learning**

# **Logiciels**

# **Anaconda**

Anaconda est une distribution des langages de programmation python et R pour le calcul scientifique, qui vise à simplifier la gestion et le déploiement des paquets.

La distribution comprend des paquets de science des données adaptées à Windows, Linux et MacOs.

# **Python**

Python est un langage de programmation informatique souvent utilisé pour créer des sites Web et des logiciels automatiser des tâches et effectuer des analyses de données.

# **Bibliothèques**

# **Jupyter notebook**

Jupyter Notebook est une application de création de notebooks sous l'égide du projet Jupyter. Construit sur la puissance du format de notebook informatique jupyter Notebook offre de nouvelles façons rapides et interactives de prototyper et d'expliquer votre code explorer et de visualiser vos données et de partager vos idées avec d'autres.

# **Pandas**

Pandas est un outil d'analyse et de manipulation de données open source rapide puissant flexible et facile à utiliser construit sur le langage de programmation Python.

# **Numpy**

Numpy est une bibliothèque pour langage de programmation python destinée à manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableaux.

# **Matplotlib**

Matplotlib est une bibliothèque complète pour créer des visualisations statiques animées et interactives en Python.

# **Seaborn**

**Seaborn est une bibliothèque permettant de créer des graphiques statistiques en Python.**

**Elle s'appuie sur matplotlib et s'intègre étroitement aux structures de données pandas.**

# **Scikit-Learn (sklearn)**

Scikit-learn est une bibliothèque clé pour le langage de programmation Python généralement utilisé dans les projets d'apprentissage automatique.

Scikit-learn se concentre sur les outils d'apprentissage automatique notamment les algorithmes mathématiques statistiques et à usage général qui constituent la base de nombreuses technologies d'apprentissage automatique.

En tant qu'outil gratuit, Scikit-learn est extrêmement important dans de nombreux types de développement d'algorithmes pour l'apprentissage automatique et les technologies associées.

# **Conclusion**

La conclusion de la méthodologie de recherche souligne l'importance des choix méthodologiques dans la conduite de cette étude. L'approche mixte, combinant des méthodes quantitatives et qualitatives, ainsi que l'application de techniques de machine Learning, a permis d'analyser les données de manière approfondie et rigoureuse. La division des données en ensembles d'entraînement et de test a permis de valider l'efficacité du modèle prédictif, garantissant ainsi des résultats fiables et pertinents. En somme, cette méthodologie a démontré son adéquation pour répondre aux objectifs de recherche, tout en offrant une approche solide pour l'application des techniques d'analyse prédictive dans le domaine de la santé.

**SECONDE PARTIE : CADRE ANALYTIQUE DE LA RECHERCHE.**

# **:** **Présentation et Analyse des données.**

# **Section 1 : Présentation des données.**

# **Introduction**

La présentation et l'analyse des données constituent une étape clé dans la validation des hypothèses et des objectifs de recherche. Après avoir collecté et préparé les données, il est crucial de les analyser de manière rigoureuse pour en extraire des informations pertinentes. Cette section se concentre sur l'exploration des données, leur visualisation, ainsi que l'application de techniques d'analyse statistique et de machine Learning pour identifier des tendances, des corrélations et des patterns significatifs. L'objectif est de transformer les données brutes en connaissances exploitables, permettant ainsi de formuler des conclusions basées sur des faits et des prédictions fiables.

# **Source des Données**

La source de notre donnée provient du plateforme kaggle qui est une communauté de rechercher des données pour la data science. Notre jeu de données bien connu appelé "Pima Indians Diabetes Database". Ce jeu de données est couramment utilisé pour la recherche et l'analyse prédictive, en particulier dans les études sur le diabète. Il contient des informations médicales relatives à la santé des femmes Pima, une population amérindienne sujette à des taux élevés de diabète.

Le jeu de données comporte plusieurs variables comme les grossesses, la glycémie, la pression artérielle, l'épaisseur de la peau, l'insuline, l'indice de masse corporelle (IMC), l'âge et une fonction appelée "DiabetesPedigreeFunction". L'objectif est de prédire l'apparition du diabète (variable cible "Outcome").

Ce jeu de données est largement utilisé dans les projets d'apprentissage automatique et est disponible via plusieurs sources publiques, y compris la bibliothèque Scikit-learn et le référentiel de données UCI Machine Learning.

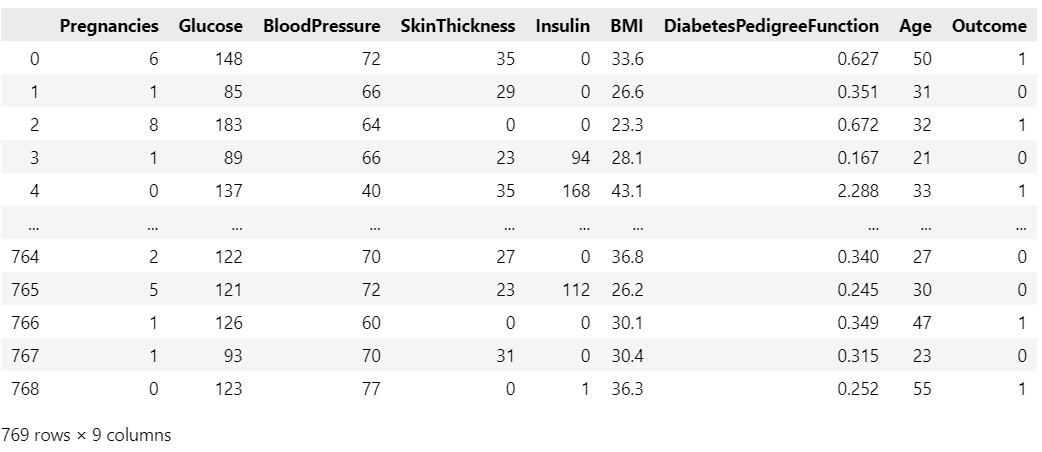


Figure 3: Données d'analyses

<https://www.kaggle.com/datasets/lara311/diabetes-dataset-using-many-medical-metrics>

# **Descriptions des données**

Tableau 1: Description des variables d’ensemble de données

|  |  |  |
| --- | --- | --- |
| **Numéro** | **Abréviation** | **Signification** |
| 1 | Pregnancies | Nombre de fois enceintes. |
| 2 | Glucose | Concentration de glucose plasmatique à 2 heures dans un test de tolérance au glucose par voie orale. |
| 3 | BloodPressure | BloodPressure Pression artérielle diastolique (mm Hg). |
| 4 | SkinThickness | Épaisseur du pli cutané du triceps (mm). |
| 5 | Insulin | Insuline sérique 2 heures (mu U / ml). |
| 6 | BMI | Indice de masse corporelle (poids en kg). |
| 7 | DiabetesPedigreeFunction | Fonction pedigree du diabète. |
| 8 | Age | Âge en années. |

La taille de l'ensemble de données comporte **769 lignes** et **9 colonnes**.

# **Section 2 :** **Analyse des données.**

### Préparation des données pour l’analyse

La préparation des données est une étape cruciale dans toute recherche basée sur l'analyse prédictive et le machine Learning. Cette phase implique plusieurs sous-étapes essentielles.

# **Téléchargement des données**

Cette étape consiste à recueillir les données nécessaires à partir de sources fiables, telles que des bases de données médicales, des archives de santé publique, ou des sources en ligne spécialisées dans le domaine du diabète.

*Voici le lien pour accéder au plateforme kaggle.*

[*https://www.kaggle.com/*](https://www.kaggle.com/)

# **Annotation des données**

Une fois les données collectées, elles doivent être annotées pour identifier et marquer les variables d'intérêt. Dans le cas du diabète, cela peut inclure des informations sur les caractéristiques démographiques des patients, leurs habitudes de vie, leurs antécédents médicaux, et les mesures biologiques liées à la glycémie.

# **Organisation des données**

Les données annotées doivent être organisées de manière structurée pour faciliter leur traitement et leur analyse ultérieure. Cela peut inclure la normalisation des valeurs, la gestion des données manquantes, et la répartition des ensembles de données en sous-ensembles dédiés à l'entraînement et au test du modèle.

# **Visualisation des données**

# **Répartition des données**

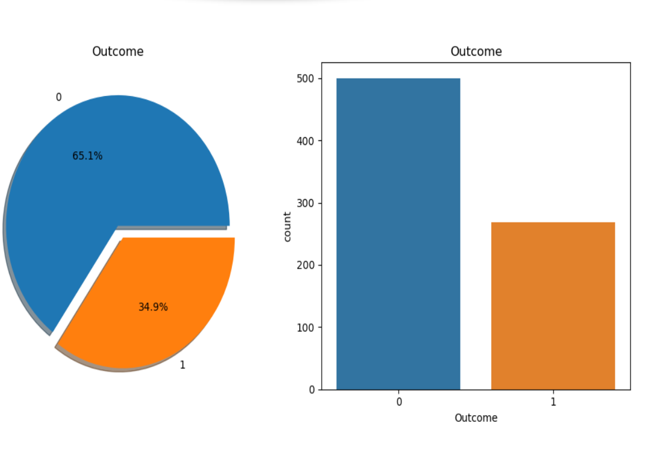


Figure 4: La distribution en pourcentage et en nombre

Cette image montre deux graphiques illustrant la distribution d'un résultat binaire (étiqueté '0' et '1'). Le graphique montre la répartition en pourcentage des deux résultats.

**1. Graphique en camembert (à gauche) - "Distribution du résultat"** :

Le graphique montre la répartition en pourcentage des deux résultats

La section bleue (étiquetée "0") représente 65,1 % des données.

La section orange (étiquetée "1") représente 34,9 % des données.

**2. Graphique en barres (à droite) - "Nombre de résultat"** :

Ce graphique montre le nombre d'occurrences de chaque résultat

La barre bleue correspond à près de 500 occurrences du résultat "0".

La barre orange correspond à environ 250 occurrences du résultat "1". Ces visualisations permettent de constater que le résultat "0" est plus fréquent que le résultat "1

# **Histogramme des colonnes**

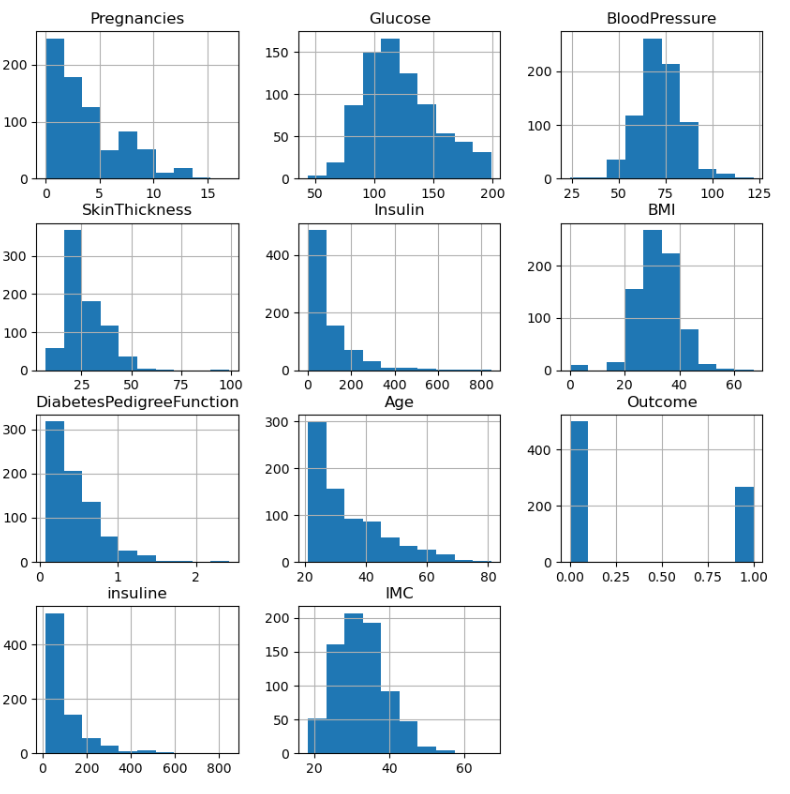
**

Figure5: Histogramme des colonnes

# **Matrice de corrélation**

Est une donnée tabulaire chaque ligne et colonne représente une variable et chaque valeur de cette matrice est le coefficient de corrélation entre les variables représentées par la ligne et la colonne correspondantes. La matrice de corrélation est une métrique d'analyse de données importante qui est calculée pour résumer les données afin de comprendre la relation entre diverses variables et prendre des décisions en conséquence. Les valeurs proches de +1 indiquent la présence d'une forte relation positive entre X et Y, tandis que celles proches de -1 indiquent

Une forte relation négative entre X et Y. Des valeurs proches de zéro signifient qu'il n'y a aucune relation entre X et Y

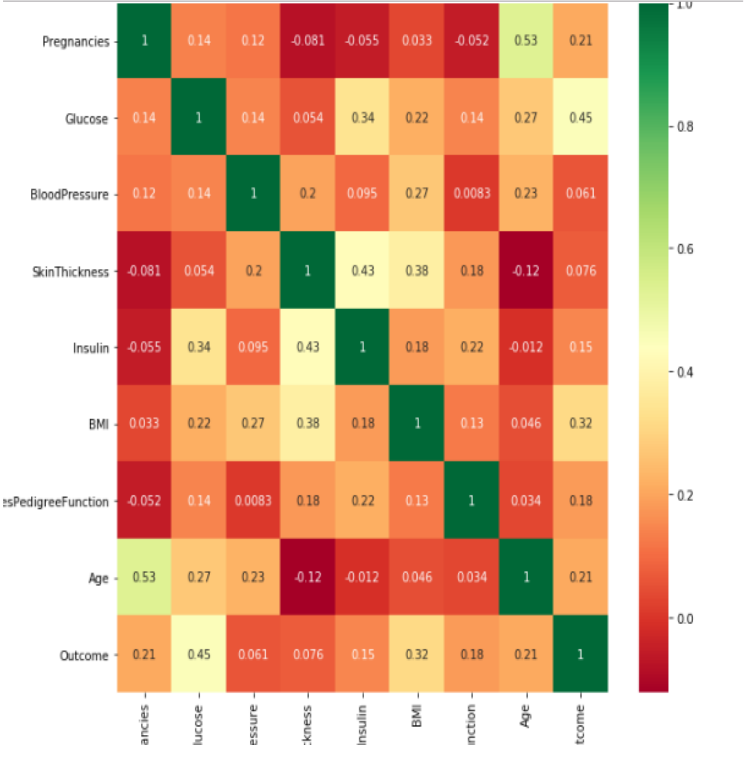


Figure 6 : Matrice de corrélation

# **Evaluation et comparaison des modèles**

L'évaluation des performances des modèles de machine Learning est cruciale pour déterminer lequel d'entre eux offre les meilleurs résultats pour une tâche spécifique. Dans cette étude comparative, on a examiné six modèles : **Régression Logistique (LR)**, **K-Nearest Neighbors (KNN)**, **Arbre de Décision (DT)**, **Support Vector Classifier (SVC)**, **Naïve Bayes (NB)**, et **Forêt d'Arbres Décisionnels (RF)**. Les critères d'évaluation utilisés incluent l'**Accuracy**, la **Précision**, le **Rappel (Recall)**, et le score **F1**.

Pour une étudié comparative entre nos différents modèles de prédiction nous constatons que le modèle ***SVM (Support Vector Machine)*** avec une précision de ***83.11%***. C’est modèle détient la meilleure performance comparativement aux autres modèles de prédiction.

# **Accuracy**

L'Accuracy mesure le pourcentage de prédictions correctes par rapport au nombre total de prédictions. Le modèle SVC se distingue avec la meilleure précision à 83,11%, suivi de RF (82,46%) et DT (79,22%). Les autres modèles, tels que LR, KNN, et NB, présentent des performances plus modestes avec des Accuracy variant entre 74% et 77%.

# **Précision**

La précision mesure la proportion de prédictions positives correctes parmi toutes les prédictions positives effectuées par le modèle. Le modèle KNN présente la meilleure précision (0,721), tandis que le modèle NB obtient la précision la plus faible (0,604). Les modèles SVC, LR, DT, et RF se situent entre ces deux extrêmes, montrant des performances raisonnables mais inférieures à KNN.

# **Rappel (Recall)**

Le rappel évalue la capacité d'un modèle à détecter toutes les occurrences positives réelles dans les données. Ici, le modèle RF surpasse les autres avec un score de rappel de 0,630, suivi de KNN (0,574) et DT (0,574). Les modèles SVC, LR, et NB obtiennent des scores plus faibles dans cette métrique, montrant une capacité réduite à identifier correctement tous les cas positifs.

# **Score F1**

Le score F1 représente un équilibre entre la précision et le rappel, ce qui en fait une mesure idéale pour évaluer les modèles lorsqu'il y a un déséquilibre entre les classes. Le modèle RF atteint le meilleur score F1 (0,630), suivi de KNN (0,639) et SVC (0,598). Bien que le modèle DT ait une bonne Accuracy, son score F1 (0,590) indique qu'il n'est pas aussi équilibré que les autres modèles.

**Tableau 2:** Evaluation des différents modèles

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **LR** | **KNN** | **DT** | **SVC** | **NB** | **RF** |
| **Accuracy** | **77,27%** | **74,67%** | **79,22%** | **83,11%** | **74,02%** | **82,46%** |
| **Precision** | **0,659** | **0,721** | **0,608** | **0,674** | **0,604** | **0,630** |
| **Recall** | **0,500** | **0,574** | **0,574** | **0,537** | **0,537** | **0,630** |
| **F1\_Score** | **0,568** | **0,639** | **0,590** | **0,598** | **0,569** | **0,630** |

# **Matrice de confusion**

Une matrice de confusion est un tableau utilisé pour évaluer la performance d'un modèle de classification. Elle compare les prédictions du modèle avec les résultats réels. La matrice comporte généralement quatre éléments :

**Vrais positifs (VP)** : Nombre de fois où le modèle a correctement prédit la classe positive.

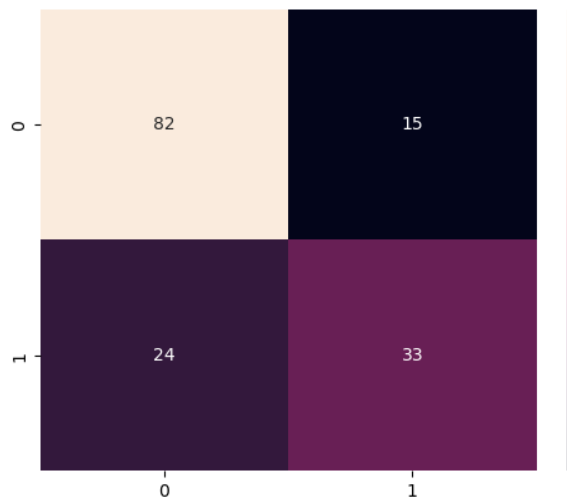
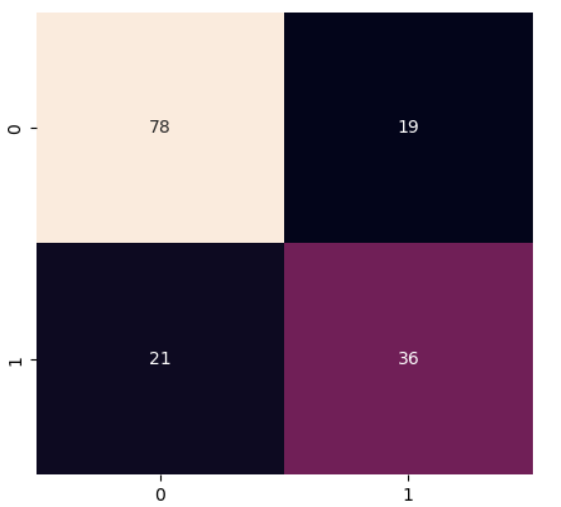
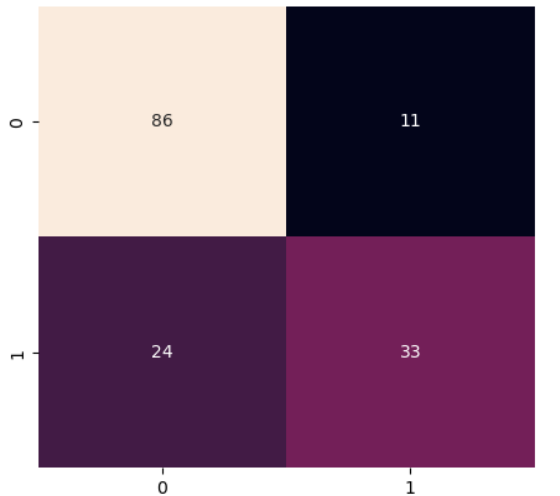
**Faux positifs (FP)** : Nombre de fois où le modèle a prédit la classe positive alors que c'était faux.

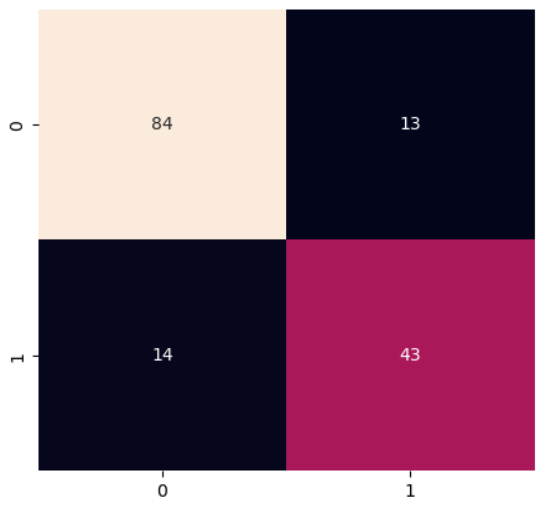
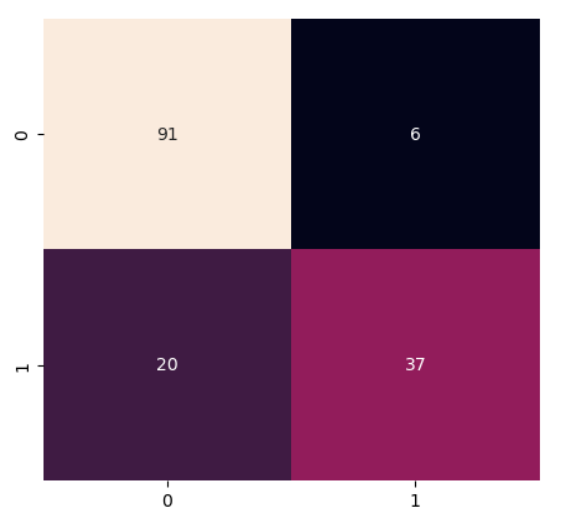
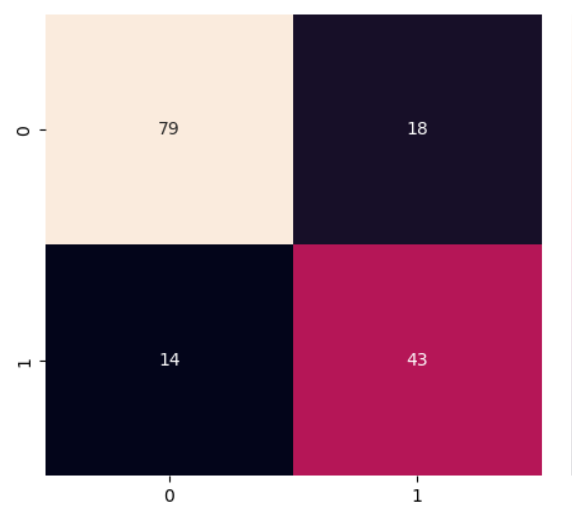
**Vrais négatifs (VN)** : Nombre de fois où le modèle a correctement prédit la classe négative.

**Faux négatifs (FN)** : Nombre de fois où le modèle a prédit la classe négative alors que c'était faux.

Ces éléments sont disposés dans une matrice qui permet de visualiser la précision et l'erreur du modèle.

**Régression logistique KNN DT**

******

**** SVCNB RF**

# **Analyse critique de l’étude**

### ****Problèmes de qualité des données****

De nombreuses études sur la prédiction du diabète utilisent des jeux de données standardisés tels que le **PIMA Indian Diabetes Dataset**. Bien que ces jeux de données soient populaires, ils présentent des limitations :

* **Taille limitée**
* **Manque de diversité ethnique**
* **Biais d'échantillonnage**

### ****Choix limité des algorithmes de machine learning****

Certaines études se limitent à l'utilisation de techniques traditionnelles telles que les **régressions logistiques** ou **arbres de décision.**

**Tandis qu’ils existent plusieurs types de modèle pour la prédiction du diabète.**

### ****Prise en compte limitée des variables contextuelles****

Les études existantes se concentrent souvent sur un nombre restreint de variables comme l’IMC, taux **de glucose,** sont **familiaux.**

### ****Évaluation limitée des modèles****

Les études antérieures n'évaluent souvent les performances de leurs modèles qu'à travers des métriques telles que l’exactitude, **sensibilité,** spécificité**.**

# **:** Réalisation de l’interface du teste.

# **Section 1 : Recommandation.**

# **Introduction**

La réalisation de l'interface de test est une étape cruciale dans le développement d'une solution d'analyse prédictive. Cette interface permet aux utilisateurs d'interagir avec le modèle prédictif en fournissant des données d'entrée et en visualisant les résultats des prédictions. Elle doit être conçue de manière intuitive et ergonomique afin de faciliter l'utilisation pour les non-experts en technologie, tout en offrant les fonctionnalités nécessaires pour tester et valider efficacement les performances du modèle. Cette section détaillera les choix de conception, les technologies employées, ainsi que les étapes de développement et de mise en place de l'interface utilisateur pour le test du modèle d'analyse prédictive appliqué à la gestion du diabète.

### Bibliothèque << Streamlit >>

Streamlit est une bibliothèque Python open-source qui permet de créer facilement des applications web interactives, principalement pour les data scientists et les ingénieurs en machine Learning. Elle se distingue par sa simplicité d'utilisation, permettant de transformer des scripts Python en applications web avec une interface utilisateur sans avoir à maîtriser des Framework web complexes comme Flask ou Django.

# **Fonctionnalités principales**

**Simplicité d'utilisation** : Streamlit permet d'intégrer des widgets interactifs tels que des boutons, des curseurs, des cases à cocher, etc., directement dans un script Python. Il suffit d'exécuter le script, et une interface web est automatiquement générée. Par exemple, une simple fonction st. Write () affiche du texte, des graphiques ou des tableaux dans l'interface web.

**Intégration facile avec les bibliothèques de data science** : Streamlit fonctionne de manière transparente avec des bibliothèques populaires comme Pandas, Matplotlib, Seaborn, Plotly, et TensorFlow. Cela permet de visualiser rapidement des données ou des modèles d'apprentissage automatique sans passer par des processus complexes.

**Réactivité** : Les applications créées avec Streamlit sont interactives et réactives. Chaque interaction avec l'application (par exemple, modifier une valeur de curseur) déclenche une nouvelle exécution du script Python, mettant automatiquement à jour l'affichage. Cela permet de visualiser instantanément les effets des changements sur les résultats.

**Déploiement facile** : Streamlit simplifie le déploiement d'applications. Les applications peuvent être hébergées localement ou déployées sur le cloud en utilisant des services comme Streamlit Cloud, Heroku, ou même sur des serveurs personnels. Cela facilite la diffusion des résultats d'une analyse ou d'un modèle à un public plus large.

En résumé, Streamlit est un outil puissant pour créer rapidement des interfaces interactives pour des projets de data science, sans nécessiter de compétences approfondies en développement web.

### Installation

Nous allons ouvrir un éditeur de code, par exemple Visual Studio Code, puis ouvrir le terminal et enfin écrire la commande ci-dessous.

**Pip install streamlit**

## **Présentation de l’interface pour le teste**

Cette interface de prévision du diabète est un outil destiné à estimer la probabilité de développement du diabète chez un individu. Nous vous présenterons les détaillée des différents champs et de leur importance dans le modèle de prédiction :

**Nombre de grossesses**

Indique le nombre de grossesses antérieures de l'individu.

Ce paramètre est important, car les femmes ayant eu plusieurs grossesses ont un risque accru de développer le diabète de type 2.

**Valeur de glucose**

Indique le taux de glucose sanguin mesuré.

Un taux élevé de glucose est un indicateur clé du diabète, car il montre que le corps a du mal à réguler le sucre dans le sang.

**Valeur de la pression artérielle**

Mesure la pression du sang contre les parois des artères.

Une pression artérielle élevée est souvent associée au diabète, car elle peut endommager les vaisseaux sanguins et réduire l'efficacité de l'insuline.

**Valeur de l'épaisseur de la peau**

Indique l'épaisseur du pli cutané (en mm) pris au niveau du triceps.

Cette mesure peut être utilisée pour estimer la quantité de graisse corporelle, ce qui est un facteur de risque pour le diabète.

**Niveau d'insuline**

Indique la concentration d'insuline dans le sang.

L'insuline est l'hormone qui régule le glucose sanguin, et des niveaux anormaux peuvent indiquer un diabète.

**Valeur IMC**

Indice de Masse Corporelle (IMC) de l'individu.

Un IMC élevé est un facteur de risque majeur pour le diabète de type 2.

Fonction de pedigree du diabète

Évalue la probabilité héréditaire de développer le diabète en se basant sur les antécédents familiaux.

Cette valeur quantifie l'impact de la génétique sur le risque de diabète.

**Entrez votre âge**

L'âge de l'individu.

Le risque de développer le diabète augmente avec l'âge.

**Bouton "PREDIRE"**

Ce bouton permet de lancer la prédiction basée sur les valeurs saisies dans les champs ci-dessus. Le modèle de prédiction utilise ces données pour estimer le risque de diabète, souvent à l'aide d'algorithmes de machine Learning qui ont été entraînés sur des ensembles de données cliniques.

**Utilité de l'interface**

Cette interface est utile pour les professionnels de santé et les patients à risque de diabète. Elle permet de surveiller les facteurs de risque et de prendre des mesures préventives ou de gérer le diabète existant plus efficacement.

**Déploiement du Modèle et Communication avec l'Interface Client**

**Intégration du Modèle** : Après avoir entraîné votre modèle prédictif (par exemple, un modèle de machine Learning pour prédire les risques de diabète), vous l'intégrez directement dans votre application Streamlit. Streamlit permet de charger des modèles enregistrés (par exemple avec Pickle ou Joblib) et de les utiliser pour des prédictions en temps réel. L'intégration se fait simplement via Python.

**Fonctionnement**

**Frontend (Streamlit)** : Le frontend de l'application est l'interface utilisateur créée par Streamlit, qui permet aux utilisateurs d'entrer des données (comme des variables cliniques pour la prédiction du diabète).

**Backend (Modèle)** : Le modèle, déployé avec le backend Python, est chargé et attend les entrées de l'utilisateur via l'interface Streamlit.

**Communication** : Lorsqu'un utilisateur soumet des données via l'interface Streamlit, celles-ci sont envoyées au modèle. Le modèle effectue ses prédictions, puis Streamlit affiche les résultats dans l'interface, sans avoir besoin de recharger la page.

**Déploiement**

**Localement** : Vous pouvez déployer votre application Streamlit localement pour développement et tests.



Figure 7 : Interface du teste

**Figure 1: Les différentes types du machine learningFigure 7: Interface du teste**

# **Section 2 : Analyse d’amélioration.**

# **Identification des limitations dans la gestion du diabète**

## **Faiblesse dans la détection précoce**

Les méthodes actuelles, telles que les tests de glycémie réguliers, manquent parfois de prédire le développement du diabète de type 2 ou les complications à un stade précoce. Cela entraîne des interventions tardives qui peuvent être moins efficaces.

L'absence de prise en compte de facteurs de risque combinés dans les systèmes de surveillance actuels (comme la génétique, les habitudes de vie, et les données cliniques continues).

### Gestion standardisée vs individualisée

La plupart des approches de gestion du diabète sont standardisées et ne tiennent pas compte des différences individuelles. Chaque patient a des besoins uniques en termes de traitement, de suivi, et de gestion de la maladie.

Les régimes de traitement actuels peuvent ne pas être optimisés pour des groupes spécifiques de patients, ce qui réduit l'efficacité des soins.

### Limitations technologiques et algorithmiques

Bien que certaines technologies (comme les capteurs de glucose en continu) aient amélioré la gestion du diabète, elles sont souvent limitées par des modèles algorithmiques simplistes qui n'intègrent pas de grandes quantités de données pour des prédictions personnalisées et précises.

Les modèles prédictifs actuels peuvent ne pas prendre en compte toutes les variables possibles et peuvent manquer de précision, ce qui entraîne des prédictions sous-optimales.

### Problèmes d'accès et d'adhérence

Dans de nombreuses régions, l'accès aux nouvelles technologies ou aux soins personnalisés est limité par des facteurs économiques ou logistiques.

Les systèmes de gestion actuels ne parviennent pas toujours à motiver ou à faciliter l'adhésion des patients à long terme, entraînant des abandons ou une mauvaise gestion de la maladie.

## **Conclusion**

La réalisation de l'interface de test a permis de créer un outil pratique et convivial pour évaluer les performances du modèle prédictif. En mettant l'accent sur l'ergonomie et la simplicité d'utilisation, cette interface facilite l'interaction entre les utilisateurs et le modèle, rendant accessible l'analyse prédictive même à des utilisateurs non spécialisés. Les tests effectués à travers cette interface ont permis de valider les prédictions du modèle en conditions réelles et d'identifier des pistes d'amélioration. Ainsi, cette interface se révèle être un élément clé dans l'évaluation et l'affinement du modèle, contribuant à l'efficacité globale du projet.

# **Conclusion générale et perspective**

Conclusion générale et perspective Le diabète est l'un des problèmes de santé majeurs dans le monde. Selon le rapport de l'OMS 2011, environ 346 millions de personnes dans le monde souffrent de diabète sucré. Un diagnostic plus précoce évite de nombreuses complications qui peuvent surviennent identification ici en Algérie la prévalence estimée à 14.4% d’après SANOFI qui est une partenaire de santé des patients algérienne. Une prédiction plus précoce évitera des complications de cette maladie. L’approche d’apprentissage automatique résoudre ce problème critique dans le but de cette étude pour construire un modèle capable de prédire si les personnes sont diabétiques des classificateurs. Dans ce mémoire nous avons choisi l’algorithme svc comme classificateur dans lequel plusieurs paramètres tels que le nombre de svc et les mesures de distances ou de similarités ont étaient testés dans cette étude. Les résultats ont montré que plus le paramètre svc augmente, plus les taux de précision, de rappel et de F-mesure augmentent également. Le meilleur score est enregistré au niveau de svc. Les distances euclidien et Minkowski comme des métriques les plus optimales et qui fonctionne mieux pour notre ensemble de données en raison de sa grande précision et score de rappel. En plus, une comparaison est accompagnée comportant l’algorithme svc contre différents algorithmes de classification d’apprentissage supervisé tel que : les arbres de décision, forêt aléatoire, machine à vecteurs de support, Naïves Bayes sur les données ‘Pima indian diabètes database’. Les résultats de performances ont montré clairement l’avance de l’algorithme SVC contre tous les autres algorithmes choisis dans cette étude. Pour les travaux futurs, plusieurs pistes peuvent être explorées. On peut appliquer la même expérimentation sur d’autres bases de données de diabète ou même de type différent pour confirmer les résultats obtenus. Améliorer l’algorithme svc pour avoir de meilleurs résultats en termes de précision et rappel. Une autre alternative est de tester la réduction de dimension sur le problème de détection de diabète.

# **Bibliographie**

* **Ouvrages généraux**

[1] American Diabetes Association. (2021). Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care*, 44(Suppl. 1), S15-S33. DOI: 10.2337/dc21-S002.

2. World Health Organisation. (2021). Global Report on Diabetes. Geneva: WHO Press.

3. International Diabetes Federation. (2021). *IDF Diabetes Atlas*, 10th Edition. Brussels : International Diabetes Federation.

4. Silver, N., & Paynter, H. (2020). *The Signal and the Noise: Why So Many Predictions Fail – but Some Don’t.* Penguin Books.

5. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can Machine Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data? *PLoS ONE*, 12(4), e0174944. DOI: 10.1371/journal.pone.0174944.

7. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future – Big Data, Machine Learning, and Clinical Medicine. *The New England Journal of Medicine*, 375(13), 1216-1219. DOI: 10.1056/NEJMp1606181.

8. Deo, R. C. (2015). Machine Learning in Medicine. *Circulation*, 132(20), 1920-1930. DOI: 10.1161/CIRCULATIONAHA.115.001593.

9. Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(3), 223-230. DOI: 10.1016/j.bpsc.2017.11.007.

10. Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling.* Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R.* Springer.

11. Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *BMC Medicine*, 13, 1. DOI: 10.1186/s12916-014-0241-z.

12.Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology.* John Wiley & Sons.

13. Creswell, J. W., & Creswell, J. D. (2017). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches.* SAGE Publications.

14. Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models.* CRC Press.

15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

16. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

18. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, 3(2), 2053951716679679. DOI: 10.1177/2053951716679679.

19. Price, W. N., & Cohen, I. G. (2019). Privacy in the Age of Medical Big Data. *Nature Medicine*, 25(1), 37-43. DOI : 10.1038/s41591-018-0272-7.

20.London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy Versus Explainability. *Hastings Center Report*, 49(1), 15-21. DOI : 10.1002/hast.973.

**21.** *Mayo Clinic - Articles mis à jour régulièrement, disponibles en ligne. Organisation Mondiale de la Santé (OMS) - Publications régulières et rapports sur le diabète, consultables sur leur site officiel.*

* **Webographie**

**22.** (geeksforgeeks).<https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>. (Consulté le 12 - 07 - 2024)

23. (Clevy).<https://blog.clevy.io/fr/le-machine-learning-decrypte-3-3-types-dapprentissage-et-limites/>. (Consulté le 13 - 07 - 2024)

**Table des matières**

[Dédicace I](#_Toc178243791)

[Remerciement II](#_Toc178243792)

[Liste des abréviations III](#_Toc178243793)

[Liste des figures et liste des tableaux IV](#_Toc178243794)

[Sommaire V](#_Toc178243795)

[Introduction générale 1](#_Toc178243796)

[Contexte et problématique 1](#_Toc178243797)

[ Question principale 1](#_Toc178243798)

[ Questions spécifiques 2](#_Toc178243799)

[Objectifs 2](#_Toc178243800)

[ Objectif général 2](#_Toc178243801)

[ Objectifs spécifiques 2](#_Toc178243802)

[Organisation du mémoire 2](#_Toc178243803)

[Chapitre 1 : Définition des concepts et étude de lien entre les concepts. 5](#_Toc178243804)

[Section 1 : Définition des concepts. 5](#_Toc178243805)

[1.1. Introduction 5](#_Toc178243806)

[1.2. Analyse prédictive 5](#_Toc178243807)

[1.3. Machine Learning 5](#_Toc178243808)

[1.4. Diabète 5](#_Toc178243809)

[2. Types de diabète 6](#_Toc178243810)

[2.1. Diabètes de type\_1 6](#_Toc178243811)

[2.2. Diabètes de types\_2 6](#_Toc178243812)

[2.3. Diabètes gestationnels 6](#_Toc178243813)

[2.4. Symptômes de la maladie 6](#_Toc178243814)

[Section 2 : Relation entre l'Analyse Prédictive et le Diabète. 7](#_Toc178243815)

[2.1. Prédiction de la progression de la maladie 7](#_Toc178243816)

[2.2. Détection précoce des complications 7](#_Toc178243817)

[2.3. Optimisation de la gestion du diabète 7](#_Toc178243818)

[3.1. Conclusion 7](#_Toc178243819)

[Chapitre 2 : Méthodologie de la recherche. 8](#_Toc178243820)

[Section 1 : Choix de la méthodologie à adapter. 8](#_Toc178243821)

[2.1. Introduction 8](#_Toc178243822)

[2.2. Approche Quantitative 8](#_Toc178243823)

[2.2.1. Collecte des données 8](#_Toc178243830)

[a) Prétraitement des données 8](#_Toc178243831)

[b) Nettoyage des données 8](#_Toc178243833)

[c) Transformation des données 9](#_Toc178243834)

[2.2.3. Division des données 9](#_Toc178243835)

[Section 2 : Techniques d’analyse prédictive. 9](#_Toc178243836)

[2.1. Les différents types d’algorithme du machine Learning 9](#_Toc178243837)

[2.1.1 Apprentissage supervisé 9](#_Toc178243838)

[2.1.2 Apprentissage semi supervisé 10](#_Toc178243839)

[2.1.3 Apprentissage non supervisé 10](#_Toc178243840)

[2.1.4 Apprentissage par renforcement 10](#_Toc178243841)

[2.2. Algorithmes de prédiction utilisée 11](#_Toc178243842)

[2.2.1 Régression logistique 11](#_Toc178243843)

[a. Avantages 11](#_Toc178243844)

[b. Inconvénients 12](#_Toc178243845)

[2.2.2 K-Nearest Neighbors 12](#_Toc178243846)

[a. Avantages 12](#_Toc178243847)

[b. Inconvénients 13](#_Toc178243848)

[2.2.3 Arbres de décision 13](#_Toc178243850)

[a. Avantages 13](#_Toc178243851)

[b. Inconvénients 14](#_Toc178243852)

[2.2.4 Support Vector Classifie 15](#_Toc178243853)

[a. Avantages 15](#_Toc178243854)

[b. Inconvénients 15](#_Toc178243855)

[2.2.5 Naïve Bayes 16](#_Toc178243856)

[a. Avantages 16](#_Toc178243857)

[b. Inconvénients 16](#_Toc178243858)

[2.2.6 Random Forest 17](#_Toc178243859)

[a. Avantages 17](#_Toc178243860)

[b. Inconvénients 18](#_Toc178243861)

[3. Logiciels et Bibliothèques utilisés en machine Learning 19](#_Toc178243862)

[3.1. Logiciels 19](#_Toc178243863)

[a. Anaconda 19](#_Toc178243864)

[b. Python 19](#_Toc178243865)

[3.2. Bibliothèques 19](#_Toc178243866)

[a. Jupyter notebook 19](#_Toc178243867)

[b. Pandas 19](#_Toc178243868)

[c. Numpy 19](#_Toc178243869)

[d. Matplotlib 19](#_Toc178243870)

[e. Seaborn 20](#_Toc178243871)

[f. Scikit-Learn (sklearn) 20](#_Toc178243872)

[3.2. Conclusion 20](#_Toc178243873)

[Chapitre 3 : Présentation et Analyse des données. 22](#_Toc178243874)

[Section 1 : Présentation des données. 22](#_Toc178243875)

[3.1. Introduction 22](#_Toc178243876)

[3.2. Source des Données 22](#_Toc178243877)

[3.3. Descriptions des données 23](#_Toc178243878)

[Section 2 : Analyse des données. 24](#_Toc178243879)

[3.4. Préparation des données pour l’analyse 24](#_Toc178243880)

[ Téléchargement des données 24](#_Toc178243881)

[ Annotation des données 24](#_Toc178243882)

[ Organisation des données 24](#_Toc178243883)

[3.5. Visualisation des données 25](#_Toc178243884)

[a. Répartition des données 25](#_Toc178243885)

[b. Histogramme des colonnes 26](#_Toc178243886)

[c. Matrice de corrélation 26](#_Toc178243887)

[4. Evaluation et comparaison des modèles 27](#_Toc178243888)

[a) Accuracy 28](#_Toc178243889)

[b) Précision 28](#_Toc178243890)

[c) Rappel (Recall) 28](#_Toc178243891)

[d) Score F1 28](#_Toc178243892)

[5. Matrice de confusion 29](#_Toc178243893)

[6. Analyse critique de l’étude 30](#_Toc178243894)

[6.1. Problèmes de qualité des données 30](#_Toc178243895)

[6.2. Choix limité des algorithmes de machine learning 30](#_Toc178243896)

[6.3. Prise en compte limitée des variables contextuelles 30](#_Toc178243897)

[6.4. Évaluation limitée des modèles 30](#_Toc178243898)

[Chapitre 4 : Réalisation de l’interface du teste. 31](#_Toc178243899)

[Section 1 : Recommandation. 31](#_Toc178243900)

[1.4. Introduction 31](#_Toc178243901)

[1.4.1. Bibliothèque << Streamlit >> 31](#_Toc178243902)

[a. Fonctionnalités principales 31](#_Toc178243903)

[1.4.2. Installation 32](#_Toc178243904)

[1.5. Présentation de l’interface pour le teste 32](#_Toc178243905)

[Section 2 : Analyse d’amélioration. 35](#_Toc178243906)

[2.1. Identification des limitations dans la gestion du diabète 35](#_Toc178243907)

[2.1.1 Faiblesse dans la détection précoce 35](#_Toc178243908)

[2.1.2 Gestion standardisée vs individualisée 35](#_Toc178243909)

[2.1.3 Limitations technologiques et algorithmiques 35](#_Toc178243910)

[2.1.4 Problèmes d'accès et d'adhérence 35](#_Toc178243911)

[2.2. Conclusion 36](#_Toc178243912)

[ Conclusion générale et perspective 37](#_Toc178243913)

[ Bibliographie 38](#_Toc178243914)