

N° PFE : 7

A. U. 2022/2023

Filière : SMI

Module : PROJET DE FIN D'ÉTUDES DE LA LICENCE DE
MATHEMATIQUE ET INFORMATIQUE

MÉMOIRE

**Application de prédiction de la maladie diabète
en utilisant le machine learning**

Présenté par :

Noms& Prénoms

Krid Aboubakr

Douski Zineb

Parcours

Science de Données

Science de Données

Soutenu le 22 Juin 2023 devant la commission d'examen :

Pr. OUAHI HASSAN	Faculté des Sciences Appliquées – Ait Melloul	Examineur
Pr. ZEBBARA KHALID	Faculté des Sciences Appliquées – Ait Melloul	Encadrant

REMERCIEMENTS

En premier lieu, nous tenons à remercier le Dieu le tout puissant pour nous avoir accordé la force, la volonté, la patience, et le courage de bien mener ce travail.

En effet, la réalisation de ce projet n'aurait pas été possible sans l'intervention de certaines personnes que nous tenons à remercier très infiniment :

Pr. ZEBBARA KHALID, notre encadrante, pour sa disponibilité, ses remarques pertinentes et ses conseils avisés tout au long de la période de préparation de ce projet.

Sans oublier notre examinateur **Pr. OUAHI HASSAN** qui a accepté de juger notre travail.

Tous le cadre administratifs et le corps professoral de la Faculté des Sciences Appliquées Ait Melloul, pour la qualité de la formation dispensée ainsi que pour toutes les facilités qu'ils mettent à la disposition des étudiants afin de leurs garantir une carrière scientifique.

Et finalement, toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce modeste travail. Notamment, nos parents, nos frères et soeurs, nos collègues, pour leurs soutien constant, conseils, les bons moments que nous avons passé ensemble.

Table des matières

Introduction	1
I. Contexte Général.....	2
1. Sur ce projet	2
2. Prédiction du diabète.....	2
3. Objectif.....	2
4. Outils de développement.....	3
a) Python	3
b) Pandas	3
c) Scikit-learn.....	3
d) Tkinter	3
e) Jupyter-Notebook	3
f) MySQL	3
II. Intelligence Artificielle	4
1. Historique.....	4
2. Applications de l'intelligence Artificielle	5
3. Catégories de l'intelligence Artificielle	6
a) L'IA faible ou étroite (ou encore l'IA spécialisée).....	6
b) L'IA forte	6
c) L'apprentissage automatique (machine learning)	6
d) L'apprentissage profond (deep learning).....	7
e) L'IA cognitive	7
4. Machine Learning	7
a) Apprentissage supervisé	7
b) Apprentissage non supervisé	7
c) Apprentissage par renforcement.....	8
III. Models utilisé	9
1. DT(decision tree or arbre de decision):.....	9
2. SVM (Support Vector Machine):	11
3. GNB (Gaussian Naive Bayes Classifier).....	12
a) Avantages.....	13

b) Inconvénients	13
4. Ensemble modeling	13
a) Comment fonctionne le machine learning ensembliste ?	13
b) Boosting vs bagging vs forêt aléatoire	13
IV. Méthodologie	14
1. Dataset	14
2. Prétraitement des données	15
a) Nettoyage des données	15
b) Visualisations Des Données.....	16
3. Division des données.....	19
a) Séparer la colonne 'Outcome' de la data frame	19
b) Train_test_split()	19
4. Création et Entraînement du modèle	20
5. Prédiction et Évaluation du modèle.....	20
a) Accuracy score.....	21
b) Confusion matrix	22
6. Sauvgarde du modèle	23
V. Implémentation.....	24
1. Notion de securité.....	24
a) Login Page	24
b) Création de compte.....	26
c) Reset Password	28
2. Page d'aperçu.....	29
3. Formulaire de prédiction.....	30
4. Base de données diabete	33
a) Table de système de connxion	33
b) Table de Sauvegarde	34
VI. Conclusion	34
Bibliographie	35

LISTE DES FIGURES

Figure 1 : diabète dans le monde	2
Figure 2 : arbre de decisions 1	9
Figure 3 : arbre de decision 2.....	10
Figure 4 : Crédit : 2017, Julien Audiffren	11
Figure 5 : Crédit :2017, Haydar Ali Ismail.....	12
Figure 6 : Théorème de Bayes.....	12
Figure 7 : Description des caractéristiques de l'ensemble de données	14
Figure 8 : Dataset sous Forme Data Frame.....	15
Figure 9 : Retrouver les données manquantes.....	15
Figure 10 : Description des Données	16
Figure 11 : Visualisations des Colonnes	17
Figure 12 : Combinaison des Caractéristiques.....	18
Figure 13 : Target & data frame séparé	19
Figure 14 : train_test_data	20
Figure 15 : create models	20
Figure 16 : prediction pour évaluation	21
Figure 17 : accuracy score.....	21
Figure 18 : save model	23
Figure 19 : première connexion	24
Figure 20 : nouveau compte	26
Figure 21 : ResetPassword	28
Figure 22 : PreviewPage.....	29
Figure 23 : PredictionPage	31
Figure 24 : Prediction Result	32
Figure 25 : Signup Table	33
Figure 26 : Predict Table	34

ABREVIATIONS

PIMA	:	North American Indians who traditionally lived along the Gila and Salt rivers in Arizona, U.S.
NIDDK	:	National Institute of Diabetes and Digestive and Kidney Diseases
PID	:	Pima Indian Diabète
DT	:	Decision Tree
SVM	:	Support Vector Machine
NB	:	Naïve Bayes
IA	:	Intelligence Artificielle
SGBDR	:	Système de Gestion de Bases de Données Relationnelles

Introduction

Le machine learning, ou apprentissage automatique en français, est une branche de l'intelligence artificielle qui permet aux ordinateurs d'apprendre à partir de données, sans être explicitement programmés pour accomplir une tâche spécifique. Les algorithmes de machine learning utilisent des modèles mathématiques et statistiques pour détecter des schémas et des relations dans les données, puis les utilisent pour effectuer des prédictions ou des décisions.

Dans le domaine de la santé, le machine learning est de plus en plus utilisé pour améliorer les diagnostics, les traitements et la prévention des maladies. Les données de santé sont de plus en plus accessibles grâce aux avancées technologiques et aux enregistrements électroniques des dossiers médicaux. Ces données peuvent inclure des informations telles que les symptômes, les résultats de tests, les traitements et les résultats de traitements.

Comment le machine learning est-il utilisé pour la prédiction du diabète et quelles sont les implications de ces modèles de prédiction pour la prise en charge des patients diabétiques ?

I. Contexte Général

1. Sur ce projet

Dans le cadre du projet de fin d'études pour l'obtention de la licence en science mathématique et informatique.

Notre projet consiste à effectuer une classification binaire supervisée de maladies des diabètes à l'aide de l'apprentissage automatique, Plus précisément on utilise Les structures suivantes : le classificateurs SVM (support vector machine), DTC (Decision Tree Classifier), GNB (Gaussian Naive Bayes) et finalement la combinaison des dernier models à l'aide de L'ensemble modeling.

2. Prédiction du diabète

Le diabète est parmi les maladies les plus rependues à travers le monde. Il est considéré comme une maladie qui se propage comme une épidémie dans le Monde entier, elle peut atteindre toutes les générations (enfants, jeunes, les personnes Agées). Cette maladie peut entrainer des effets très graves en termes de défaillance d'organes et peut entrainer la mort aussi. (CEED, 2023)



Figure 1 : diabètes dans le monde

3. Objectif

L'objectif de ce projet est de proposer une nouvelle approche de prédiction du diabète qui donne des résultats utiles et efficaces. Ceci va aider à prédire si un patient donné va être un futur diabétique. Si c'est le cas le patient va essayer de prendre les précautions nécessaires (suivre un régime alimentaire adéquat, pratiquer du sport périodiquement, etc.) pour prévenir cette maladie chronique.

4. Outils de développement

a) Python

Python est un langage de programmation simple mais puissant avec d'excellentes fonctionnalités pour le traitement des données linguistiques. Python peut être téléchargé gratuitement sur : <http://www.python.org/>.

Nous avons choisi Python parce qu'il a une courbe d'apprentissage superficielle, sa syntaxe et sa sémantique sont transparentes, et il a une bonne fonctionnalité de gestion des chaînes. En tant que langage interprété, Python facilite l'exploration interactive. En tant que langage orienté objet, Python permet d'encapsuler et de réutiliser facilement les données et la méthode. En tant que langage dynamique, Python permet d'ajouter des attributs à des objets à la volée et de taper dynamiquement une variable, ce qui facilite le développement rapide. Python est livré avec une vaste bibliothèque standard, y compris des composants pour la programmation graphique, le traitement numérique et la connectivité.

Python est très utilisé dans l'industrie, la recherche scientifique et l'éducation dans le monde entier. Python est souvent loué pour la façon dont il facilite la productivité, la qualité et la maintenabilité des logiciels. (Souha, 2022)

b) Pandas

Pandas est une bibliothèque écrite pour le langage de programmation Python elle nous permet de manipuler et analyser les données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles. (wikipedia, 2023)

c) Scikit-learn

Scikit-learn (anciennement scikits. Learn) et également connu sous le nom de sklearn) est une bibliothèque d'apprentissage automatique pour le langage de programmation Python. Elle comporte des divers algorithmes de classification, de régression et de clustering, notamment les machines vectorielles de support, les forêts aléatoires, l'amplification de gradient, et est conçu pour interagir avec les bibliothèques numériques et scientifiques Python NumPy et SciPy. (wikipedia, 2023)

d) Tkinter

Tkinter (de l'anglais Tool kit interface) est la bibliothèque graphique libre d'origine pour le langage Python, permettant la création d'interfaces graphiques. (wikipedia, 2023)

e) Jupyter-Notebook

Les notebooks Jupyter sont des cahiers électroniques qui, dans le même document, peuvent rassembler du texte, des images, des formules mathématiques et du code informatique exécutable. Ils sont manipulables interactivement dans un navigateur web. (univ-paris, 2023)

f) MySQL

MySQL est un système de gestion de bases de données relationnelles (SGBDR). Il est distribué sous une double licence GPL(General Public License) et propriétaire. Il fait partie des logiciels de gestion de base de données les plus utilisés au monde. (ICTEA, 2023)

II. Intelligence Artificielle

C'est une discipline scientifique relative au traitement des connaissances et au raisonnement dans le but de permettre à une machine d'exécuter des fonctions normalement associées à l'être humain. L'intelligence artificielle tente de reproduire les processus cognitifs humains dans le but de réaliser des actions « intelligente ». Elle est comme « la construction des programmes informatiques qui s'adonnent à des tâches qui sont pour l'instant accomplies de façon plus satisfaisante par des êtres humains car elles demandent des processus mentaux de haut niveau tels que :

- L'apprentissage perceptuel.
- L'organisation de la mémoire et le raisonnement critique.

1. Historique

Il est intéressant de revenir sur les origines et l'historique de l'intelligence artificielle afin de bien comprendre ses premières orientations et ses perspectives pour l'avenir:

- **1950** : Alan M. Turing, mathématicien et théoricien précurseur de l'informatique, lance le concept d'intelligence artificielle.
- **1955 – 1956** : Lancement du premier programme d'intelligence artificielle par Allen Newell, John C. Shaw et Herbert A. Simon, Logic Theorist.
- **1957** : Modélisation des jeux d'échec.
- **1958** : John McCarthy invente le Lisp (list processing), langage de programmation interactif (développement au MIT).
- **1958** : Construction du premier réseau neuronal, le Perceptron, de Frank Rosenblatt, machine dite connexionniste.
- **1959** : Élaboration du premier GPS (général problem solver) -fin de la première période de l'intelligence artificielle.
- **1970** : Néoconnexionnisme.
- **1989** : DeepThought, supercalculateur d'IBM, deux millions de coups par seconde.
- **1990 – 1997** : Développement de Deep Blue rebaptisé DeeperBlue: conception d'un système de 256 processeurs fonctionnant en parallèle, chaque processeur peut calculer environ trois millions de coups par seconde.
- **2009** : Le MIT a lancé un projet visant à repenser la recherche en intelligence artificielle.
- **2011** : Watson, le superordinateur d'IBM remporte deux des trois manches du jeu télévisé Jeopardy! La performance a consisté pour cette intelligence artificielle à répondre à des questions de culture générale.
- **2013** : Humain Brain Project. Google ouvre un laboratoire de recherches dans les locaux de la NASA.
- **2014** : DeepKnowledge Ventures : nomme à son conseil d'administration VITAL, un algorithme capable d'élaborer ses décisions en analysant les bilans comptables des entreprises potentiellement intéressantes, le test clinique, la propriété intellectuelle et les précédents investissements.
- **2015** : Facebook Artificial Intelligence Research (FAIR). Google rend sa technologie d'intelligence artificielle TensorFlow accessible à tous. Développement d'une crainte que l'intelligence artificielle dépasse à terme les performances de l'intelligence humaine.
- **2016** : Amelia IPSoft un agent virtuel. Aussi, AlphaGo bat trois fois consécutives le champion du monde de jeu de go, Lee Se-Dol en cinq manches.

2. Applications de l'intelligence Artificielle

L'intelligence artificielle a déjà trouvé de multiples usages au sein de la société :

- ❖ **Les moteurs de recherche** : Tous les moteurs de recherche (y compris des agences de voyage) sont basés sur des systèmes intelligents d'extraction, d'analyse, et de classification de données pour produire le plus rapidement possible un résultat pertinent à la requête de l'utilisateur. C'est ainsi que Google a implémenté un système utilisant les techniques de machine learning pour son moteur de recherche en octobre 2015, intitulé RankBrain9. Ce système convertit de grandes quantités de texte en vecteurs mathématiques pour aider le système à deviner le sens des mots ou des phrases qu'il ne connaît pas et ainsi traiter les 15 % de requêtes jamais effectuées auparavant qu'il reçoit chaque jour.
- ❖ **Les moteurs de recommandation** : En s'appuyant sur les données issues de la navigation et des achats d'un utilisateur, des sites comme Amazon ou Netflix sont capables de lui proposer d'autres produits similaires qui pourraient l'intéresser. Ces technologies prédictives sont aussi utilisées pour les plateformes de publicité en ligne (Google, Criteo) pour proposer aux visiteurs des contenus d'annonceurs en rapport avec les pages qu'ils ont visité.
- ❖ **La traduction automatique** : Elle s'appuie sur des algorithmes de modélisation statistique du langage naturel. Ils intègrent les règles de construction de chaque langue.
- ❖ **Les assistants personnels (Siri, Cortana, Google Now...)** : Ils sont déployés sur les Smartphones qui s'appuient sur plusieurs briques technologiques: la reconnaissance vocale pour convertir le son en texte, le langage naturel pour comprendre le sens des mots, un moteur de recherche pour trouver réponse à la question et la synthèse vocale pour communiquer la réponse à l'utilisateur, la planification pour la gestion d'événements, etc.
- ❖ **Les agents conversationnels** : Ils sont utilisés dans les domaines du support client et du télémarketing et consistent en des fenêtres de chat qui s'ouvrent toutes seules sur un site web, ou de serveur vocal qui répond aux questions 24h sur 24. Ils utilisent le langage naturel et leur accès à de vastes bases de données leur permet de répondre aux questions les plus simples.
- ❖ **Les véhicules autonomes** : Si certains prototypes roulent déjà sur les routes au contact des autres véhicules, les voitures qui se garent toutes seules ou qui freinent par anticipation sont déjà une réalité. Le pilotage automatique des avions ou la gestion de trajectoire des véhicules spatiaux, ou encore les drones se basent aussi sur l'intelligence artificielle.
- ❖ **Les systèmes de navigation GPS** : Développé en 1968 par l'Institut de recherche de Stanford, cet algorithme permet d'optimiser le cheminement entre plusieurs points dans un réseau en se basant sur le coût du trajet ou la distance parcourue.
- ❖ **Les finances** : Elles sont gérées par des systèmes intelligents pour organiser leurs opérations, investir en bourse et gérer leurs biens, mais aussi pour repérer des transactions qui sortent de l'ordinaire. Les banques possèdent aussi des systèmes experts d'évaluation de risques liés à l'octroi de crédits (credit-scoring).
- ❖ **Le cyber sécurité** : Les acteurs du cyber sécurité ont adopté les techniques du machine learning afin de détecter des comportements anormaux dans les systèmes d'information, et de déceler les menaces persistantes pour éviter des opérations d'espionnage ou d'extraction de données10. Près de 300 paramètres (heures et IP des connexions et des machines, téléchargements, etc.) sont pris en compte pour établir le modèle d'analyse comportementale

dont la première phase d'apprentissage dure environ une semaine. Citons la jeune pousse lyonnaise Sentryo qui intègre des algorithmes de machine learning pour sécuriser les sites industriels critiques.

- ❖ **Les jeux vidéo :** Ils emploient des techniques d'intelligence artificielle pour donner vie aux personnages non joueurs ou encore pour créer des univers entiers à partir d'algorithmes. En 1997, DeepBlue, le superordinateur d'IBM avait battu Garry Kasparov, champion du monde d'échecs en titre. En 2016, c'est DeepMind, le programme d'intelligence artificielle de Google qui a annoncé la victoire de son programme AlphaGo contre le champion d'Europe de go en titre, Fan Hui13. Ce résultat prend appui sur la technologie des réseaux neuronaux, que nous décrivons précédemment. Deepmind avait déjà développé un système d'intelligence artificielle capable de déterminer l'action la plus judicieuse pour battre l'homme dans une vingtaine de jeux d'arcade.
- ❖ **La médecine :** L'ordinateur Watson d'IBM vise à compiler la plus large base de données dédiée à la santé dans le monde, portant sur 300 millions de patients. La branche santé d'IBM a acquis quatre sociétés médicales depuis sa création : Phytel (santé & population), Explorys (fichiers santé de cliniques), Merge Healthcare (imagerie médicale) et Truven (analytics médicales). Ce supercalculateur a notamment fait ses preuves dans un service d'oncologie pour des services personnalisés : analysant ADN, dossier du patient et autres publications et essais cliniques.
L'ordinateur propose un protocole adapté au patient. L'entreprise américaine Enlitic propose des technologies d'intelligence artificielle pour analyser l'imagerie médicale, qui permettraient de détecter des fractures de façon plus performante que les radiologues d'après les dires de la société. (Tawfik Beghriche, 2020).

3. Catégories de l'intelligence Artificielle

Il existe plusieurs catégories d'intelligence artificielle (IA), notamment :

a) L'IA faible ou étroite (ou encore l'IA spécialisée)

Elle est conçue pour effectuer une tâche spécifique de manière efficace, mais ne peut pas s'adapter ou apprendre de nouvelles tâches sans être reprogrammée. Les exemples d'IA faible sont les systèmes de reconnaissance vocale ou de recommandation de produits en ligne.

b) L'IA forte

L'IA forte est une catégorie d'intelligence artificielle qui vise à créer des systèmes capables de démontrer une intelligence équivalente ou supérieure à celle des êtres humains, en termes de capacités cognitives générales, telles que la compréhension, l'apprentissage, la perception, la résolution de problèmes, le raisonnement et la prise de décisions. Bien que ce type d'IA n'existe pas encore, il s'agit de l'objectif ultime de la recherche en intelligence artificielle.

c) L'apprentissage automatique (machine learning)

Il est une sous-catégorie de l'IA qui utilise des algorithmes pour apprendre des données et améliorer ses performances sans être explicitement programmé. Les exemples d'apprentissage automatique sont les réseaux de neurones, les arbres de décision et les algorithmes de clustering.

d) L'apprentissage profond (deep learning)

Il est une sous-catégorie de l'apprentissage automatique qui utilise des réseaux de neurones artificiels pour effectuer des tâches complexes, telles que la reconnaissance d'images ou de la parole. L'apprentissage profond est souvent utilisé pour les projets de traitement de données massives.

e) L'IA cognitive

L'IA cognitive est une catégorie d'intelligence artificielle qui vise à imiter les fonctions cognitives du cerveau humain, telles que la perception, l'attention, la mémoire, le raisonnement et la compréhension du langage naturel. Les systèmes d'IA cognitive sont capables de traiter des données complexes et de fournir des réponses précises et pertinentes. Les exemples d'IA cognitive incluent les systèmes de traitement de langage naturel, les chatbots avancés, les assistants virtuels et les robots sociaux.

REMARQUE 1

Ces catégories ne sont pas mutuellement exclusives et peuvent être combinées pour créer des systèmes d'IA plus avancés.

4. Machine Learning

Le machine learning utilise des algorithmes qui analysent les données pour identifier des schémas et des relations entre les variables, et qui peuvent ensuite être utilisés pour effectuer des prévisions ou prendre des décisions en temps réel. Le machine learning est un domaine en constante évolution, avec de nouvelles techniques et applications qui émergent régulièrement.

Il existe plusieurs types de machine learning, chacun avec ses propres caractéristiques et applications. Voici une brève description des principaux types de machine learning :

a) Apprentissage supervisé

Dans ce type d'apprentissage, le modèle est entraîné à partir d'un ensemble de données étiquetées, où chaque exemple est associé à une réponse connue. Le modèle utilise ces exemples pour apprendre à prédire des réponses similaires pour de nouvelles données. Les exemples d'apprentissage supervisé incluent la classification, la régression et la prédiction.

b) Apprentissage non supervisé

dans ce type d'apprentissage, le modèle est entraîné à partir d'un ensemble de données non étiquetées, où il n'y a pas de réponse connue. Le modèle utilise ces données pour découvrir des structures et des modèles cachés dans les données. Les exemples d'apprentissage non supervisé incluent la segmentation, la détection d'anomalies et la réduction de dimension.

c) **Apprentissage par renforcement**

Ici le modèle apprend à travers des interactions avec un environnement. Le modèle reçoit des récompenses ou des punitions pour ses actions, ce qui lui permet de découvrir les actions les plus appropriées pour atteindre un objectif. Les exemples d'apprentissage par renforcement incluent les jeux vidéo, les robots autonomes et les systèmes de recommandation.

REMARQUE 2

Chacun de ces types de machine learning présente des avantages et des limitations, et le choix du type approprié dépendra des données, de l'objectif et du contexte d'utilisation.

III. Models utilisé

1. DT(decision tree or arbre de decision):

Un arbre de décisions est un algorithme d'apprentissage supervisé non paramétrique, qui est utilisé à la fois pour les tâches de classification et régression. Il a une structure hiérarchique, une structure arborescente, qui se compose d'un noeud racine, de branches, de nœuds interne et de nœuds feuille.



Figure 2 : arbre de decisions 1

Comme vous pouvez le voir sur le diagramme ci-dessus, un arbre de décisions commence par un noeud racine, qui n'a pas de branches entrantes. Les branches sortant du noeud racine vont dans les noeuds internes, également connu comme nœuds de décision. Sur la base des caractéristiques disponibles, les deux types de noeud effectuent des évaluations sur des sous-ensembles homogènes, qui sont désignés comme nœuds feuille ou nœuds terminal. Les nœuds feuille représentent tous les résultats possibles au sein du fichier. A titre d' exemple, imaginons que vous essayez d'évaluer si vous devez ou non aller surfer, vous pouvez utiliser les règles de suivi de décision pour faire un choix :

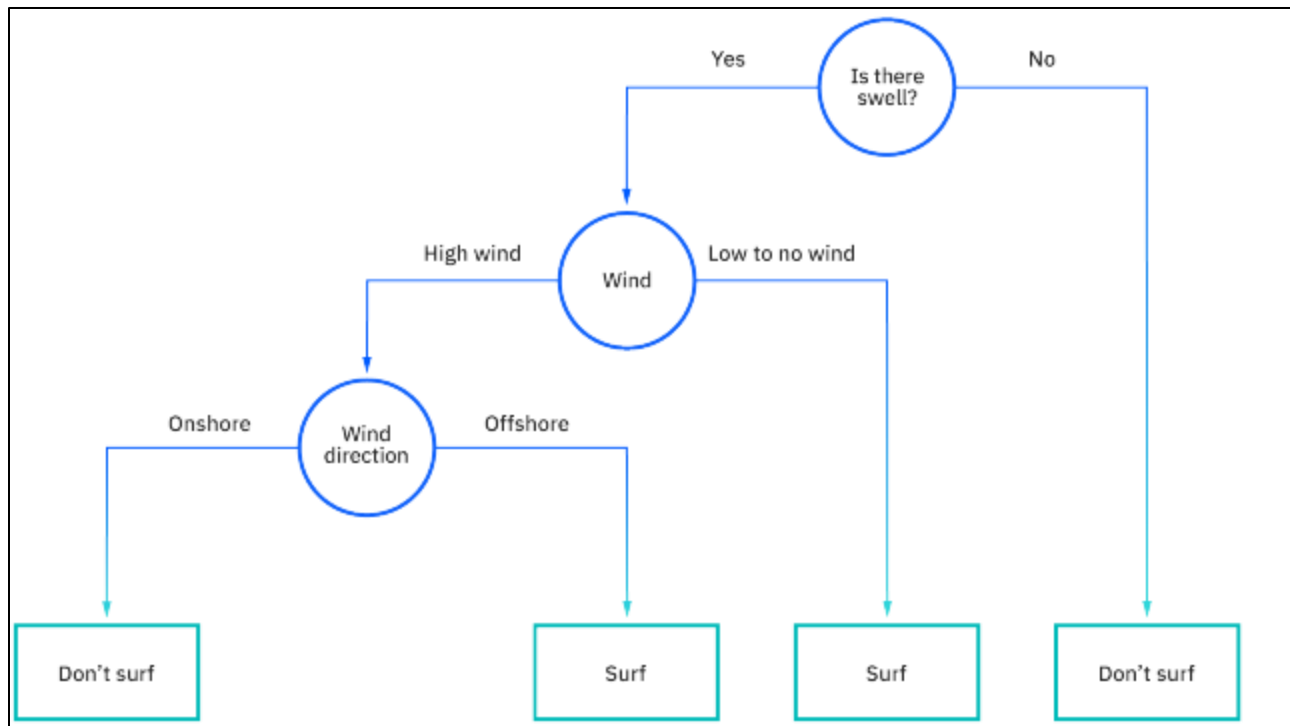


Figure 3 : arbre de décision 2

Ce type de structure en organigramme crée également une représentation facile à assimiler pour la prise de décision, permettant aux différents groupes d'une organisation de mieux comprendre pourquoi une décision a été prise.

L'apprentissage par arborescence de décisions utilise une stratégie de division et de conquête en effectuant une recherche gloutonne pour identifier les points de fractionnement optimaux au sein d'une arborescence. Ce processus de fractionnement est ensuite répété de manière descendante, récursive jusqu'à ce que tous, ou la majorité des enregistrements aient été classifiés sous des étiquettes de classe spécifiques. Que tous les points de données soient ou non classifiés comme des ensembles homogènes dépend majoritairement de la complexité de l'arbre de décisions. Les arbres plus petits sont plus facilement capables d'atteindre des nœuds feuille purs, c'est-à-dire des points de données dans une classe unique. Cependant, à mesure qu'une arborescence croît en taille, il devient de plus en plus difficile de maintenir cette pureté, et il en résulte généralement trop peu de données relevant d'un sous-arbre donné. Lorsque cela se produit, on parle de fragmentation des données, et cela peut souvent aboutir à un surajustement. Par conséquent, les arbres de décisions ont une préférence pour les petits arbres, ce qui est cohérent avec le principe de parcimonie d'Occam's Razor ; c'est-à-dire que « les entités ne doivent pas être multipliées au-delà de la nécessité ». Autrement dit, les arbres de décisions ne doivent ajouter complexité que si nécessaire, car l'explication la plus simple est souvent la meilleure. Pour réduire la complexité et éviter le surajustement, l'élagage est généralement employé ; il s'agit d'un traitement, qui supprime les branches qui se fractionnent sur des éléments de basse importance. L'ajustement du modèle peut alors être évalué par le traitement de validation croisée. Une autre façon pour les arbres de décisions de maintenir leur exactitude est de former un ensemble via un algorithme de forêt aléatoire ; ce discriminant prédit des résultats plus précis, en particulier lorsque les arbres individuels ne sont pas corrélés les uns avec les autres.

Types d'arbres de décisions

L'algorithme de Hunt, qui a été développé dans les années 1960 pour devenir le modèle d'apprentissage humain en psychologie, forme la base de nombreux algorithmes d'arbre de décisions populaires, tels que les suivants :

ID3 : Ross Quinlan est crédité dans le développement d'ID3, qui est l'abréviation de « Iterative Dichotomiser 3 ». Cet algorithme exploite l'entropie et le gain d'informations comme mesures pour évaluer les fractionnements candidats. Certaines des recherches de Quinlan sur cet algorithme à partir de 1986.

C4.5 : Cet algorithme est considéré comme une itération ultérieure d'ID3, également développée par Quinlan. Il peut utiliser des informations de gain ou de ratios de gain pour évaluer des points de fractionnement au sein des arbres de décisions.

CART : Le terme, CART, est une abréviation pour « arbres de classification et de régression » et a été introduit par Leo Breiman. Cet algorithme utilise typiquement l'impureté Gini pour identifier l'attribut idéal pour effectuer le fractionnement. L'impureté Gini mesure la fréquence à laquelle un attribut choisi au hasard est mal classé. Lors de l'évaluation à l'aide de l'impureté Gini, une valeur inférieure est plus idéale. (IBM, 2023)

2. SVM (Support Vector Machine):

SVM (Support Vector Machine ou Machine à vecteurs de support) : Les SVMs sont une famille d'algorithmes d'apprentissage automatique qui permettent de résoudre des problèmes tant de classification que de régression ou de détection d'anomalie. Ils sont connus pour leurs solides garanties théoriques, leur grande flexibilité ainsi que leur simplicité d'utilisation même sans grande connaissance de data mining.

Les SVMs ont été développés dans les années 1990. Comme le montre la figure ci-dessous, leur principe est simple : il ont pour but de séparer les données en classes à l'aide d'une frontière aussi « simple » que possible, de telle façon que la distance entre les différents groupes de données et la frontière qui les sépare soit maximale. Cette distance est aussi appelée « marge » et les SVMs sont ainsi qualifiés de « séparateurs à vaste marge », les « vecteurs de support » étant les données les plus proches de la frontière.

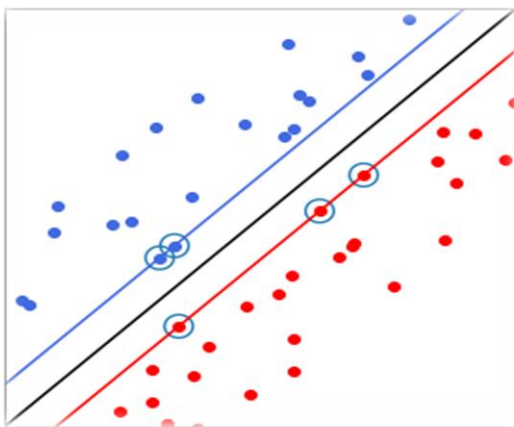


Figure 4 : Crédit : 2017, Julien Audiffren

Dans cet espace à deux dimensions, la « frontière » est la droite noire, les « vecteurs de support » sont les points entourés (les plus proches de la frontière) et la « marge » est la distance entre la frontière et les droites bleue et rouge.

Cette notion de frontière suppose que les données soient linéairement séparables, ce qui est rarement le cas. Pour y pallier, les SVMs reposent souvent sur l'utilisation de « noyaux ». Ces fonctions mathématiques permettent de séparer les données en les projetant dans un feature space (un espace vectoriel de plus grande dimension, voir figure ci-dessous). La technique de maximisation de marge permet, quant à elle, de garantir une meilleure robustesse face au bruit – et donc un modèle plus généralisable.

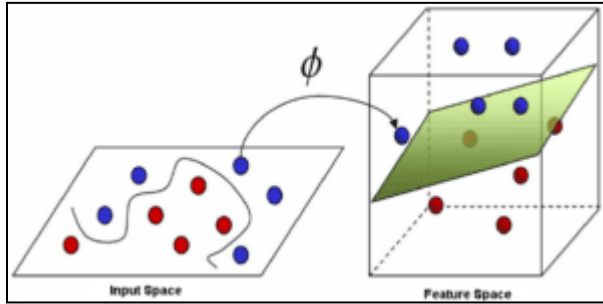


Figure 5 : Crédit :2017, Haydar Ali Ismail

Les SVMs permettent de projeter les données dans une espace de plus grande dimension via une fonction noyau pour les séparer linéairement.

Les SVMs sont utilisés dans une variété d'applications (bioinformatique, recherche d'informations, vision par ordinateur, finance, etc.) notamment parce qu'à la différence des réseaux de neurones, on peut les utiliser sans comprendre leur fonctionnement : il existe des jeux d'hyperparamètres par défaut – pour la classification, la régression ou la détection d'anomalie – qui fonctionnent dans l'immense majorité des cas. C'est un de leurs principaux avantages. Ces hyperparamètres sont, par ailleurs, en nombre très réduit : ils se limitent au choix de la technique de régularisation (de type lasso ou encore régularisation RKHS*, une méthode spécifique aux SVMs) et au choix du noyau (noyaux polynomiaux, Sobolev, RBF**...). Concernant les algorithmes SVMs, citons le kernel ridge regression pour la régression ou le one class SVM pour la détection d'anomalie.

Enfin, selon les données, la performance des SVMs est en général de même ordre voire supérieure à celle d'un réseau de neurones ou d'un modèle de mélanges gaussiens, à l'exception de certains cas notables comme la classification d'images. Il a aussi été montré qu'en utilisant un noyau RBF, les SVMs deviennent un « approximateur universel », c'est à dire qu'avec suffisamment de données, l'algorithme peut toujours trouver la meilleure frontière possible pour séparer deux classes (à condition que cette frontière existe). (DAP, 2023)

3. GNB (Gaussian Naive Classifier)

Naive Bayes Classifier est un algorithme populaire en Machine Learning. C'est un algorithme du Supervised Learning utilisé pour la classification. Il est particulièrement utile pour les problématiques de classification de texte.

Le naive Bayes classifieur se base sur le théorème de Bayes. Ce dernier est un classique de la théorie des probabilités. Ce théorème est fondé sur les probabilités conditionnelles.

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Figure 6 : Théorème de Bayes

a) Avantages

En se basant sur l'exemple de classification des fruits, on remarque plusieurs avantages pour cet algorithme :

le Naive Bayes Classifier est très rapide pour la classification : en effet les calculs de probabilités ne sont pas très coûteux.

La classification est possible même avec un petit jeu de données

b) Inconvénients

l'algorithme Naive Bayes Classifier suppose l'indépendance des variables : C'est une hypothèse forte et qui est violée dans la majorité des cas réels. (JDN, 2023)

4. Ensemble modeling

L'apprentissage ensembliste (ou ensemble learning) est une technique qui repose sur la combinaison de multiples algorithmes de machine learning pour accroître les performances du modèle d'apprentissage, et parvenir à un niveau de précision supérieur à celui qui serait réalisé si on utilisait un de ces algorithmes pris séparément.

Selon les cas de figure et le résultat désiré, il est possible d'utiliser un seul et même type d'algorithme pour produire un résultat homogène, ou encore des algorithmes différents pour obtenir un résultat hétérogène. Le concept sur lequel repose la méthode ensembliste est que l'on peut améliorer les performances des modèles apprenants en les combinant.

a) Comment fonctionne le machine learning ensembliste ?

Il existe deux grandes méthodes ensemblistes, la séquentielle et la parallèle. Avec la méthode d'ensemble learning séquentielle, les modèles sont entraînés à la suite, ce qui leur permet d'apprendre au fur et à mesure de leurs erreurs. Avec la méthode ensembliste parallèle, les modèles sont entraînés en simultané. Chaque modèle est entraîné sur une sous-échantillon de l'ensemble de données d'apprentissage. Pour déterminer le résultat final, on procède par un vote des résultats de chaque modèle pour une classification, ou par une moyenne pour une régression.

b) Boosting vs bagging vs forêt aléatoire

Pour combiner toutes les informations portées par chaque modèle, il convient de diminuer leur variance afin de réduire leur sensibilité aux données, et pour y parvenir il est possible d'employer diverses techniques en machine learning :

Le bagging (une méthode ensembliste parallèle), aussi appelé bootstrap aggregating, consiste à sous-échantillonner les données, en créant un data set pour chaque modèle. Pour déterminer le résultat final, on procède par un vote des résultats de chaque modèle pour une classification, ou par une moyenne pour une régression.

Le boosting (une méthode ensembliste séquentielle) combine les modèles classifieurs en les pondérant à chaque nouvelle prédiction, de façon à ce que les modèles ayant prédit correctement aient un poids plus important que les modèles incorrects.

La forêt aléatoire est une amélioration du bagging, qui est associé au concept de sous-espace aléatoire, et qui s'attache à créer de multiples arbres de décision, avec des modèles entraînés sur des sous-ensembles de données légèrement différents. Vu que les échantillons sont créés de manière aléatoire, la corrélation entre les arbres est réduite, et on obtient in fine à un meilleur résultat. Cette méthode est de nos jours très utilisée par les data scientists. (JDN, 2023)

IV. Méthodologie

1. Dataset

Cet ensemble de données provient initialement de l'Institut national du diabète et des maladies digestives et rénales. L'objectif de l'ensemble de données est de prédire de manière diagnostique si un patient est atteint de diabète, en fonction de certaines mesures diagnostiques incluses dans l'ensemble de données. Plusieurs contraintes ont été imposées sur la sélection de ces instances à partir d'une base de données plus large. En particulier, tous les patients ici sont des femmes âgées d'au moins 21 ans d'origine indienne Pima.

La principale motivation de l'utilisation de l'ensemble de données PIMA est la suivante : la plupart des gens dans le monde suivent le même style de vie, avec une plus grande dépendance à l'égard des aliments transformés et un déclin de l'activité physique. PID est une étude de cohorte à long terme depuis 1965 par le NIDDK en raison du risque maximal de diabète.

L'ensemble de données contient certains paramètres de diagnostic et des mesures grâce auxquels le patient peut être identifié avec presque tout type de maladie chronique ou de diabète avant le temps. Le PID est composé d'un total de 768 instances, dont 268 échantillons ont été identifiés comme diabétiques et 500 comme non-diabétiques. Les 8 attributs les plus influents qui ont contribué à la prédiction du diabète sont les suivants : plusieurs grossesses de la patiente, IMC, taux d'insuline, âge, tension artérielle, épaisseur de la peau, glycémie, etc.

Le dataset Pima peut être téléchargeable depuis ce lien : [https://www.kaggle.com/ uciml/ pima-indians-diabetes-database/data](https://www.kaggle.com/uciml/pima-indians-diabetes-database/data). (Souha, 2022)

le tableau ci-dessus présente les caractéristiques utiliser dans le dataset .

Caractéristiques "Features"	Intervalle
Grossesses : (nombre de fois enceinte).	[1 - 17]
Glucose : (Concentration de glucose plasmatique à 2 heures dans un test de tolérance au glucose par voie orale).	[0 - 199]
Pression artérielle : (Pression sanguine diastolique).	[0 – 122]
Épaisseur de la peau : (Épaisseur du pli cutané du triceps (mm)).	[0 – 99]
Insuline : (Insuline sérique de 2 heures (mu U/ml)).	[0 – 846]
IMC : (Indice de masse corporelle).	[0 - 67.1]
Fonction Pedigree Diabète.	[0.078 – 2.42]
Age	[21 – 81]
Résultat	0 / 1

Figure 7 : Description des caractéristiques de l'ensemble de données

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
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

Figure 8 : Dataset sous Forme Data Frame

2. Prétraitement des données

a) Nettoyage des données

Le nettoyage de données est l'étape cruciale de prétraitement dans la science de données, qui consiste à éliminer les données inutiles, manquantes, en double ou erronées. Cette étape garantit la qualité et l'intégrité des données avant de les utiliser pour des analyses et des modélisations. Le nettoyage de données peut impliquer des techniques telles que l'imputation de données manquantes, la suppression des valeurs aberrantes et la normalisation des données.

i. Les Valeurs Manquantes

data.info()				data.isnull().sum()	
<class 'pandas.core.frame.DataFrame'>					
RangeIndex: 768 entries, 0 to 767					
Data columns (total 9 columns):					
#	Column	Non-Null Count	Dtype		
---	-----	-----	----		
0	Pregnancies	768 non-null	int64	Pregnancies	0
1	Glucose	768 non-null	int64	Glucose	0
2	BloodPressure	768 non-null	int64	BloodPressure	0
3	SkinThickness	768 non-null	int64	SkinThickness	0
4	Insulin	768 non-null	int64	Insulin	0
5	BMI	768 non-null	float64	BMI	0
6	DiabetesPedigreeFunction	768 non-null	float64	DiabetesPedigreeFunction	0
7	Age	768 non-null	int64	Age	0
8	Outcome	768 non-null	int64	Outcome	0
dtypes: float64(2), int64(7)				dtype: int64	

Figure 9 : Retrouver les données manquantes

REMARQUE 3

Maintenant, nos données n'ont plus de valeurs manquantes, toutes les lignes sont remplies.

ii. La Réalité des valeurs

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

Figure 10 : Description des Données

REMARQUE 4

Comme nous le voyons dans les colonnes sur l'étiquette min, il n'y a pas de valeurs négatives, l'étiquette max semble également des valeurs logiques.

b) Visualisations Des Données

La visualisation des données est une étape clé de la data science qui permet de présenter les résultats d'analyses et de modélisations de manière compréhensible et conviviale. Les graphiques, tableaux et diagrammes sont utilisés pour représenter visuellement les données et les tendances, pour identifier les modèles et les relations entre les variables, et pour communiquer les résultats aux parties prenantes. Les visualisations peuvent également aider à explorer les données et à découvrir des informations cachées ou inattendues.

i. Visualisations Des Colonnes

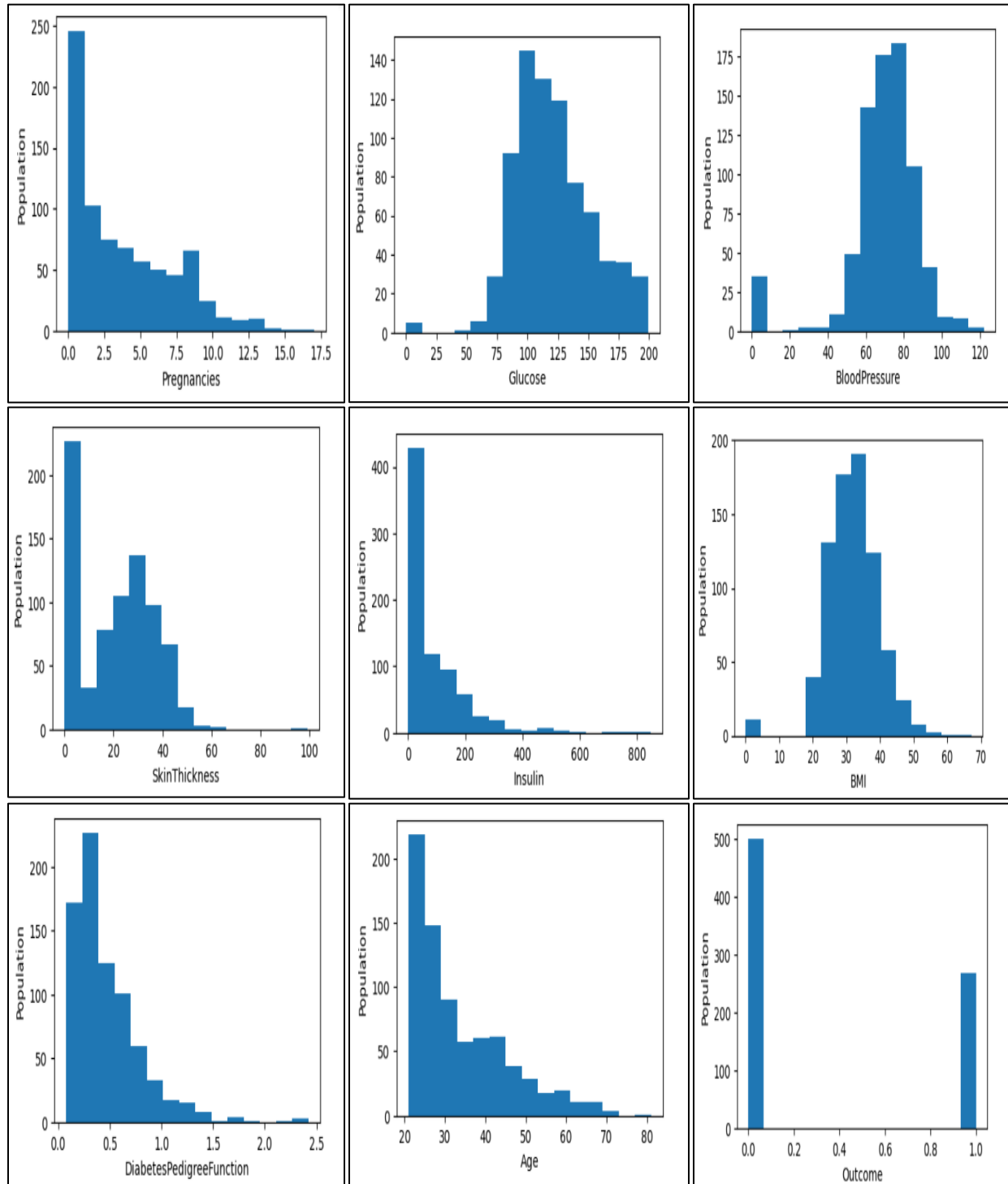


Figure 11 : Visualisations des Colonnes

ii. Liaison Entre les Caractéristiques

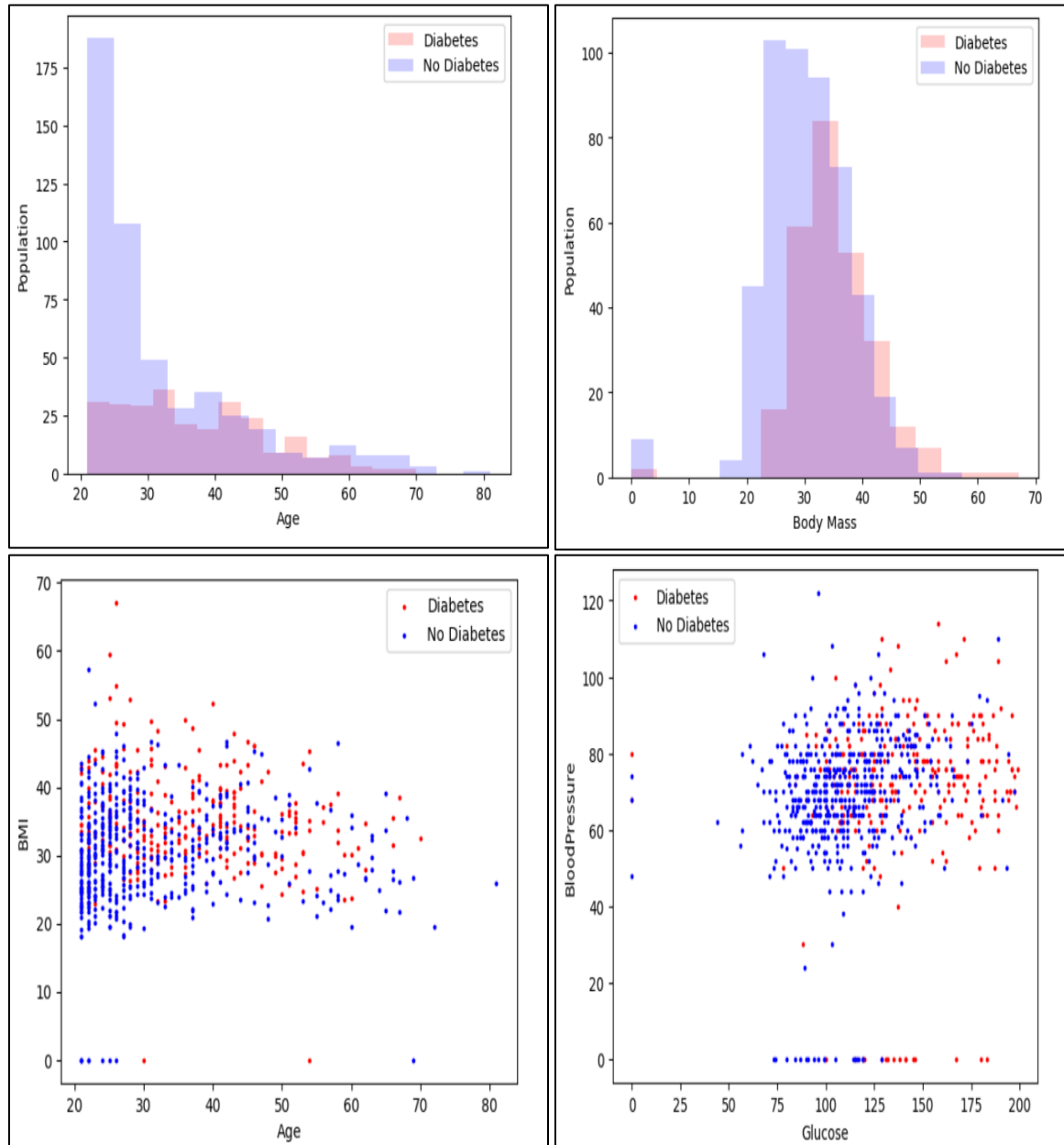


Figure 12 : Combinaison des Caractéristiques

iii. Avantages

La visualisation des données en data science offre plusieurs avantages, notamment une compréhension rapide des tendances et des modèles, une prise de décision éclairée, la détection d'erreurs, une communication plus efficace des résultats, une exploration des données plus facile et la possibilité d'optimiser les performances. Les visualisations permettent également de découvrir des informations cachées ou inattendues, ce qui peut conduire à des idées nouvelles et à des opportunités de création de valeur.

3. Division des données

Dans le domaine de la data science, la division des données est une étape important qui permet de séparer les données en différents ensembles afin d'entraîner, de valider et de tester les modèles de manière efficace. Cette étape permet d'évaluer la performance du modèle sur des données inconnues et d'éviter le surapprentissage, ce qui garantit une généralisation optimale du modèle. Selon les besoins spécifiques du projet, et la division des données peut être effectuée de manière aléatoire ou stratifiée.

a) Séparer la colonne 'Outcome' de la data frame

```
target = data['Outcome']
data.drop('Outcome',axis=1,inplace=True)
target
```

```
0      1
1      0
2      1
3      0
4      1
...
763    0
764    0
765    0
766    1
767    0
Name: Outcome, Length: 768, dtype: int64
```

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

768 rows x 8 columns

Figure 13 : Target & data frame séparé

b) Train_test_split()

La division des données en train-test split est une technique couramment utilisée en data science pour séparer les données en deux ensembles distincts : un ensemble d'entraînement (train) et un ensemble de test (test). L'ensemble d'entraînement est utilisé pour entraîner le modèle et ajuster ses paramètres, tandis que l'ensemble de test est utilisé pour évaluer la performance du modèle sur des données inconnues. Cette technique permet d'estimer la capacité de généralisation du modèle, c'est-à-dire sa capacité à prédire avec précision sur de nouvelles données.

```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(df, target, test_size=0.2, random_state=42)
print("X_train : ",X_train.shape)
print("X_test : ",X_test.shape)
print("y_train : ",y_train.shape)
print("y_test : ",y_test.shape)

X_train : (614, 8)
X_test : (154, 8)
y_train : (614,)
y_test : (154,)
```

Figure 14 : train_test_data

REMARQUE 5

Cette figure indique que les données sont prêtes à être formées .

4. Création et Entraînement du modèle

La création et l'entraînement d'un modèle de machine learning sont des étapes essentielles dans le processus de développement de solutions basées sur l'intelligence artificielle. Ces étapes consistent à construire un modèle mathématique capable d'apprendre à partir de données et de prendre des décisions et de faire des prédictions.

```
# Define a set of individual models
DT_model_1 = DecisionTreeClassifier()
DT_model_2 = DecisionTreeClassifier()
SVM_model_1 = SVC(probability=True)
SVM_model_2 = SVC(probability=True)
NB_model_1 = GaussianNB()
NB_model_2 = GaussianNB()

# Define the ensemble model
super_model = VotingClassifier(estimators=[('dt', DT_model_1), ('svm', SVM_model_1), ('gnb', NB_model_1)], voting='soft')

# Train the ensemble model on the training data
super_model.fit(X_train, y_train)
DT_model_2.fit(X_train, y_train)
SVM_model_2.fit(X_train, y_train)
NB_model_2.fit(X_train, y_train)
```

Figure 15 : create models

5. Prédiction et Évaluation du modèle

Une fois que le modèle de machine learning est créé et entraîné, il peut être utilisé pour effectuer des prédictions sur de nouvelles données. La prédiction consiste à fournir au modèle des données en entrée et à obtenir une estimation ou une réponse prévue en sortie. Cette étape est cruciale

pour utiliser le modèle afin de résoudre des problèmes spécifiques ou de prendre des décisions basées sur les prédictions.

```
# Generate predictions on the test data using the ensemble model
super_pred = super_model.predict(X_test)
DT_pred = DT_model_2.predict(X_test)
SVM_pred = SVM_model_2.predict(X_test)
NB_pred = NB_model_2.predict(X_test)
```

Figure 16 : prediction pour évaluation

Lorsque le modèle effectue des prédictions, il est important d'évaluer ses performances pour comprendre à quel point il est précis et fiable. L'évaluation du modèle permet de mesurer l'adéquation entre les prédictions du modèle et les valeurs réelles des données. Cela aide à déterminer si le modèle est capable de généraliser et de produire des résultats précis sur des données qu'il n'a pas encore vues.

a) Accuracy score

L'Accuracy score, ou taux de précision, est l'une des mesures d'évaluation couramment utilisées pour évaluer les performances d'un modèle de machine learning. Il mesure le pourcentage de prédictions correctes par rapport à l'ensemble des prédictions effectuées par le modèle.

Pour calculer l'Accuracy score, on compare les prédictions du modèle aux valeurs réelles des données et on détermine le nombre de prédictions correctes. Ce nombre est ensuite divisé par le nombre total de prédictions pour obtenir le taux de précision, exprimé en pourcentage.

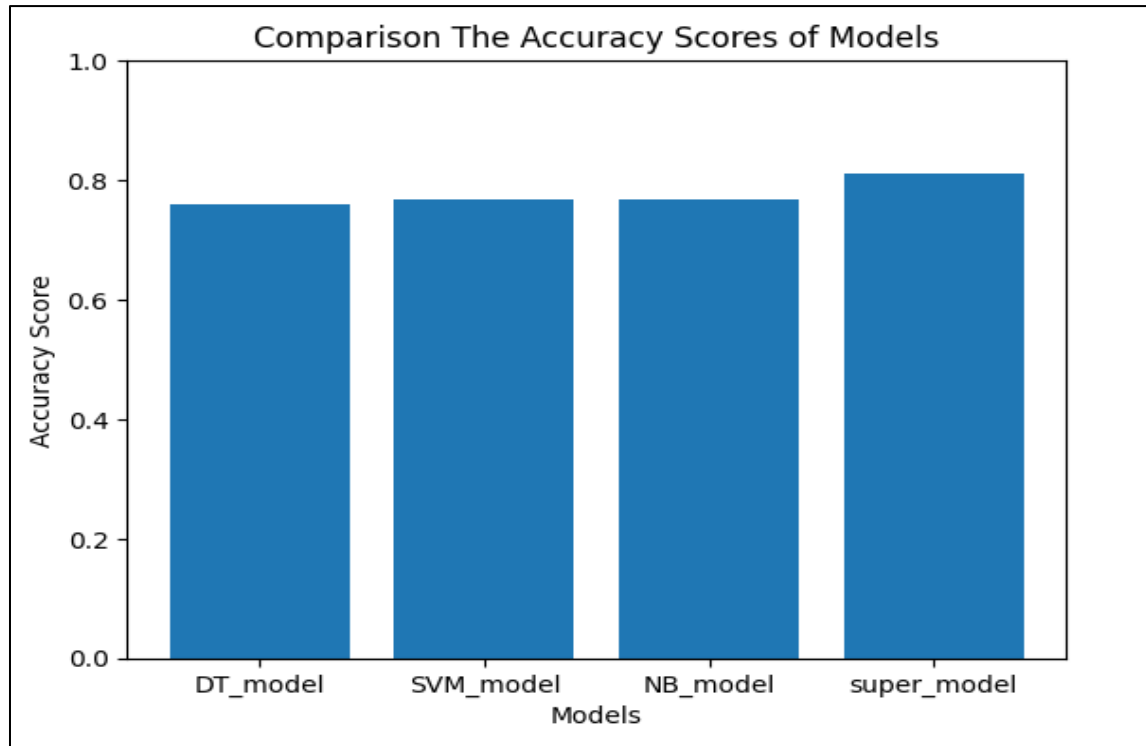
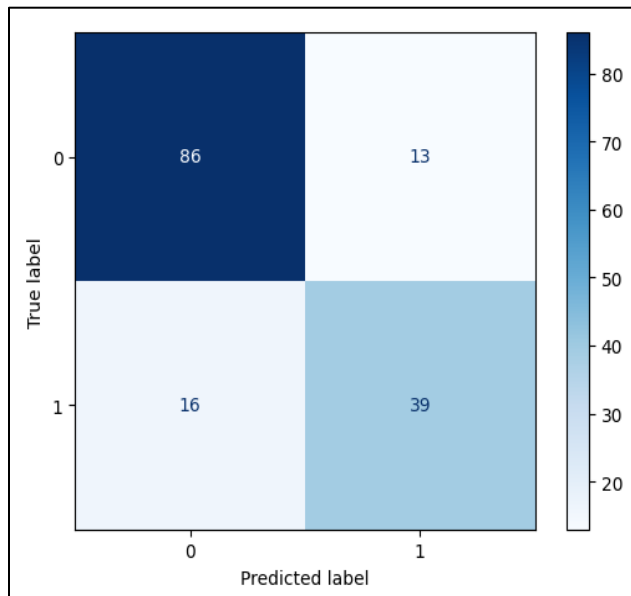


Figure 17 : accuracy score

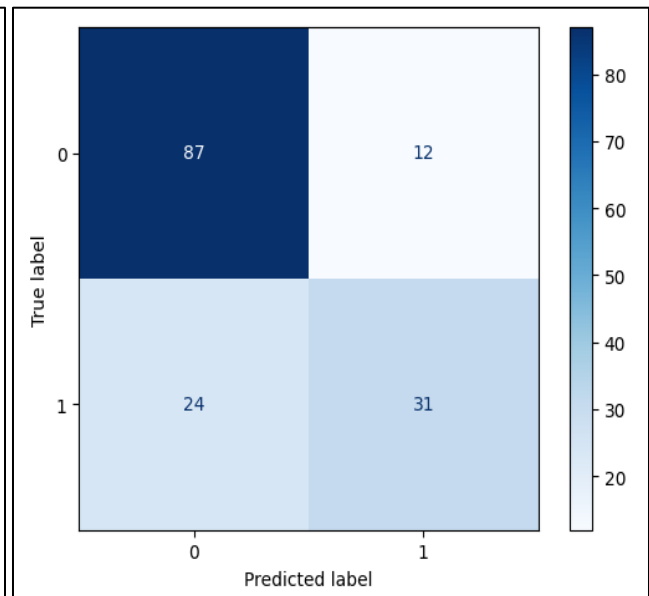
Dans l'ensemble, les résultats des Accuracy Scores suggèrent que le modèle "super_model" a les performances les plus élevées parmi les modèles évalués, suivi par les modèles "SVM" et "NB". Le modèle "DT" a obtenu le taux de précision le plus bas parmi les modèles évalués, mais il peut encore fournir des résultats acceptables .

b) Confusion matrix

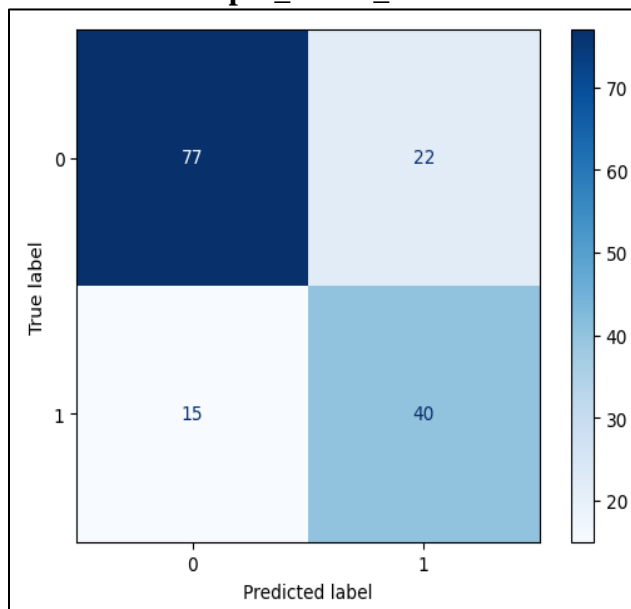
La Confusion Matrix, également appelée matrice de confusion, aussi est une représentation tabulaire utilisée pour évaluer les performances d'un modèle de machine learning en comparant les prédictions du modèle aux valeurs réelles des données. Elle est particulièrement utile lorsque le modèle effectue une classification supervisée avec plusieurs classes.



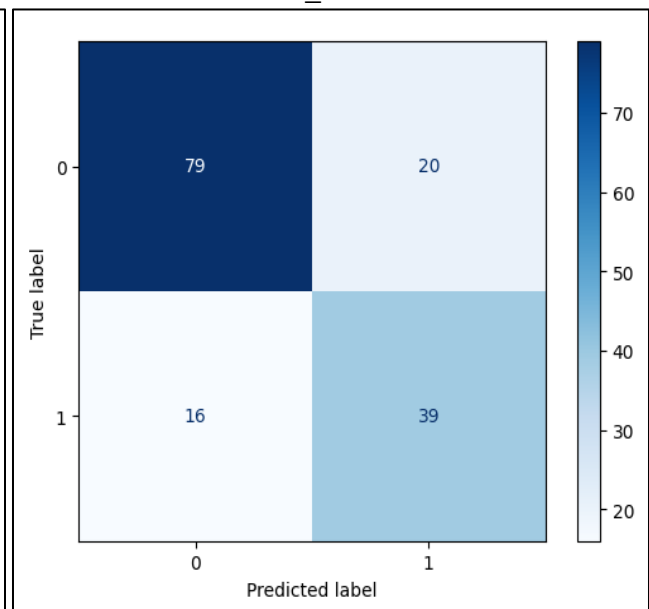
super_model_matrix



SVM_matrix



DT_matrix



NB_matrix

L'analyse de ces Confusion Matrix permet de comprendre comment les modèles ont performé en termes de prédictions correctes et incorrectes pour chaque classe. Il est essentiel de prendre en compte ces résultats pour évaluer la capacité du modèle à prédire avec précision les différentes classes du problème.

6. Sauvgarde du modèle

La sauvegarde d'un modèle de machine learning est une étape essentielle pour préserver notre travail et permettre une utilisation future. Cela nous permet de réutiliser, de partager et de déployer facilement nos modèles, tout en garantissant la reproductibilité et la disponibilité continue de nos résultats.

```
import pickle
with open(r'C:\Users\Aboubakr\Desktop\PFE\Application\super_model','wb') as file:
    pickle.dump(super_model,file)
```

Figure 18 : save model

V. Implémentation

1. Notion de sécurité

Dans le monde numérique d'aujourd'hui, la sécurité des systèmes et des données est d'une importance primordiale. Que ce soit dans le cadre d'une application, d'un site web ou d'un système d'information, il est crucial de mettre en place des mesures de sécurité adéquates pour protéger les informations sensibles et garantir l'intégrité des données.

L'un des éléments essentiels de la sécurité est le système de connexion, qui permet aux utilisateurs d'accéder de manière sécurisée à une plateforme ou à des fonctionnalités spécifiques. Le système de connexion joue un rôle clé dans l'authentification des utilisateurs, en vérifiant leur identité et en leur accordant les autorisations appropriées pour accéder aux ressources ou aux fonctionnalités.

a) Login Page

Notre page de connexion est la première interface que les utilisateurs rencontrent lorsqu'ils souhaitent accéder à notre application. Elle joue un rôle essentiel dans l'authentification des utilisateurs et garantit un accès sécurisé aux fonctionnalités et aux contenus réservés.

La conception de notre page de connexion est essentielle pour offrir une expérience utilisateur fluide tout en assurant la sécurité des informations sensibles. Voici une description des principaux éléments et fonctionnalités que l'on peut retrouver sur notre page de connexion

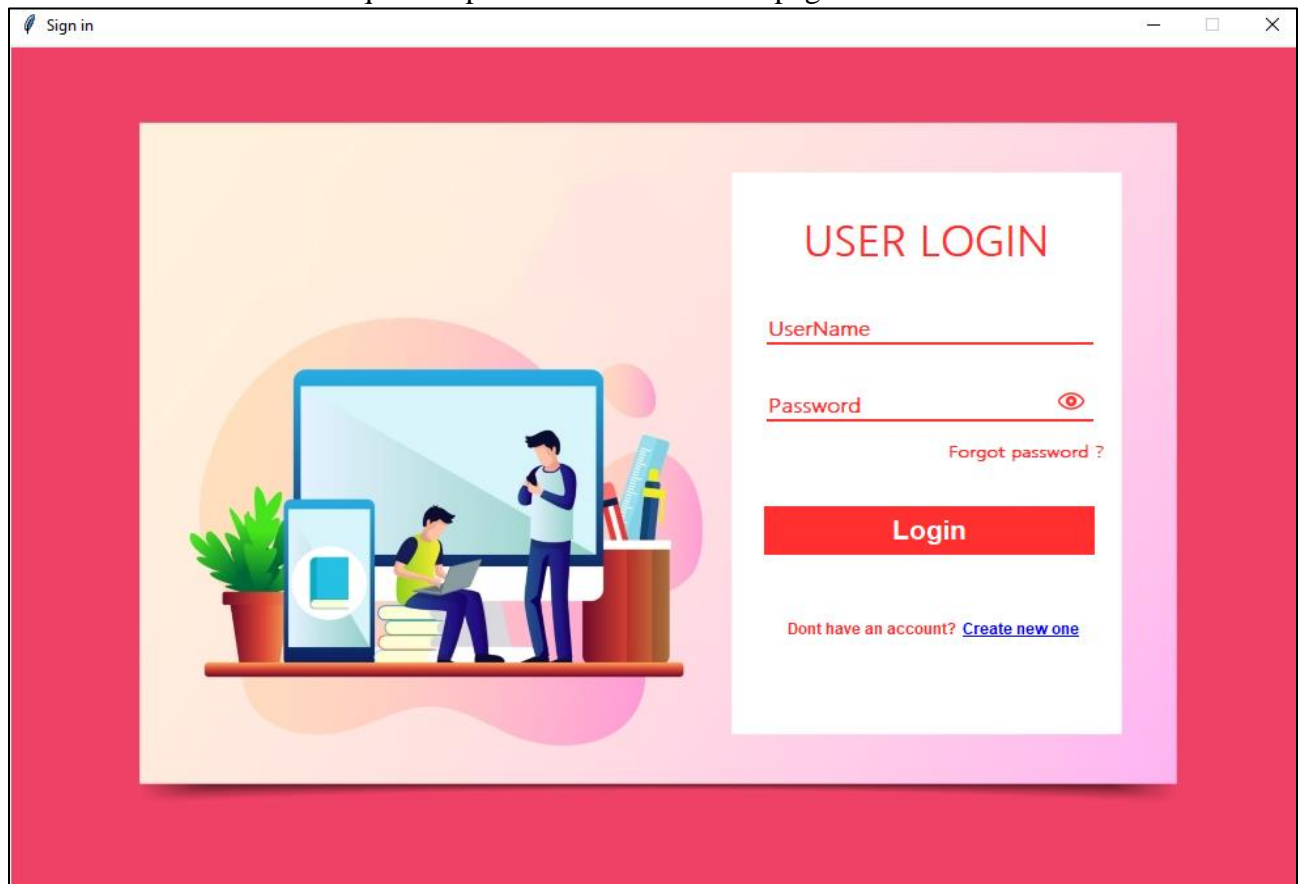


Figure 19 : première connexion

i. Formulaire de connexion

Le formulaire de connexion est l'élément central de notre page. Il comprend deux champs, l'un pour l'identifiant de l'utilisateur (UserName) et l'autre pour le mot de passe (Password). Les utilisateurs saisissent leurs informations d'identification dans ces champs pour accéder à leur compte.

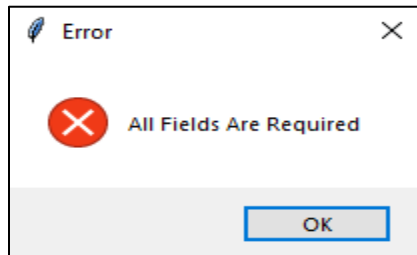
ii. Bouton de connexion

Un bouton de connexion bien visible permet aux utilisateurs de soumettre leurs informations d'identification et de se connecter à leur compte(Login). Ce bouton déclenche le processus d'authentification.

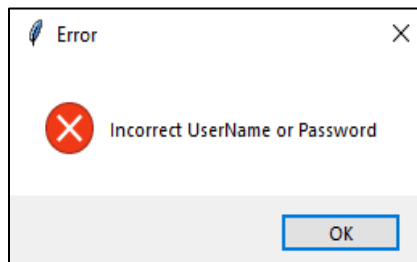
iii. Liens d'inscription et de récupération de mot de passe

Si les utilisateurs n'ont pas encore de compte, vous pouvez inclure un lien vers une page d'inscription où ils peuvent créer un nouveau compte(Dont have an account?). De plus, un lien vers une page de récupération de mot de passe peut être fourni pour aider les utilisateurs qui ont oublié leurs informations de connexion(Forgot Password).

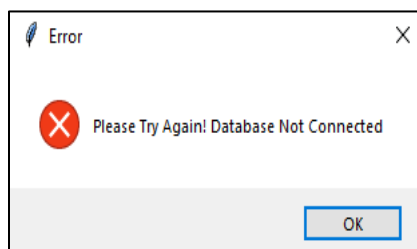
iv. Messages d'erreur et de succès



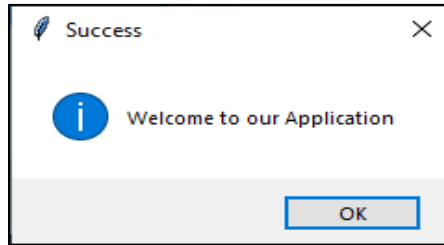
Les utilisateurs doit remplir les deux champs demandé.



Nécessite de insérer les correct informations de connexion.



Ce message indique que notre base de donnée n'est pas disponible pour le moment, donc doit mettre en marche la base de donnée MySQL.



Un message de bienvenue va être affiché lorsqu'ils se connectent avec succès.

b) Création de compte

Cette page de création de compte, également connue sous le nom de "Create An Account", est l'interface où les nouveaux utilisateurs peuvent s'inscrire pour créer un compte sur notre application. Cette page est un élément crucial de notre système d'inscription, car elle permet aux utilisateurs de fournir les informations nécessaires pour accéder à notre service.

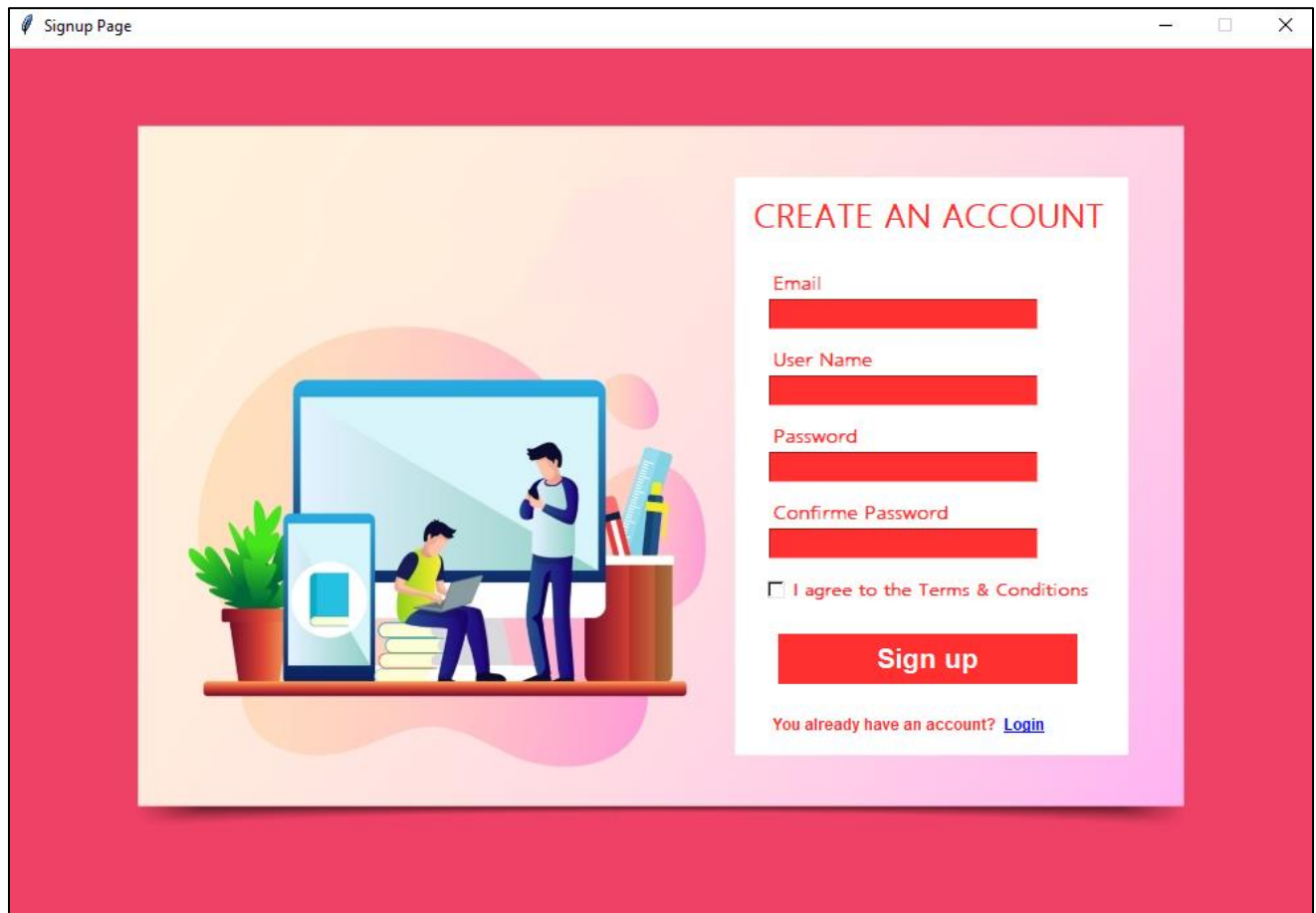
A screenshot of a web application's "Signup Page". The page has a pink background. On the left, there is a large illustration of two people working at a desk with a large monitor, a laptop, and a potted plant. On the right, there is a white box titled "CREATE AN ACCOUNT" in red. Inside this box, there are four red input fields for "Email", "User Name", "Password", and "Confirme Password". Below these fields is a checkbox labeled "I agree to the Terms & Conditions". At the bottom of the box is a red button labeled "Sign up". Below the button, there is a link that says "You already have an account? [Login](#)".

Figure 20 : nouveau compte

i. Formulaire d'inscription

Le formulaire d'inscription est l'élément central de cette page. Il comprend les champs où les utilisateurs peuvent saisir leurs informations personnelles, telles que leur adresse e-mail, leur user name ,leur mot de passe, confirmation de mot de passe . Ces informations sont nécessaires pour créer un compte utilisateur unique et sécurisé.

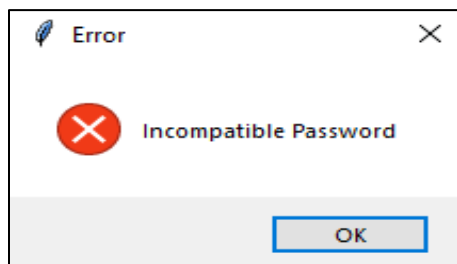
ii. Politique de confidentialité

Une case à cocher permet aux utilisateurs de prendre connaissance de nos pratiques en matière de confidentialité et des conditions auxquelles ils doivent adhérer en utilisant notre service.

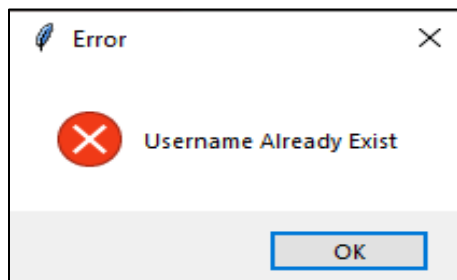
iii. Bouton de création de compte

Un bouton clairement visible, tel que "Sign up", permet aux utilisateurs de soumettre leurs informations d'inscription et de finaliser le processus de création de compte.

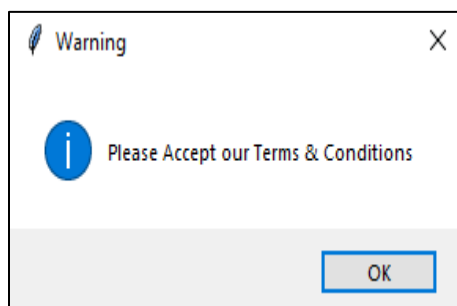
iv. Messages d'erreur et de succès



Il faut remplir les champs concerné par la même mot de passe .



L'utilisateur doit choisir une autre Nom de L'utilisateur car le nom qui a saisie déjà existe.



Une case à cocher doit être coché avant de cliquer sur le bouton de sign up .

c) Reset Password

La page de réinitialisation de mot de passe (Reset Password Page) est une fonctionnalité essentielle pour permettre aux utilisateurs de réinitialiser leur mot de passe en cas d'oubli ou de perte. Il comporte un formulaire où les utilisateurs peuvent entrer leur identifiant afin de démarrer le processus de réinitialisation.

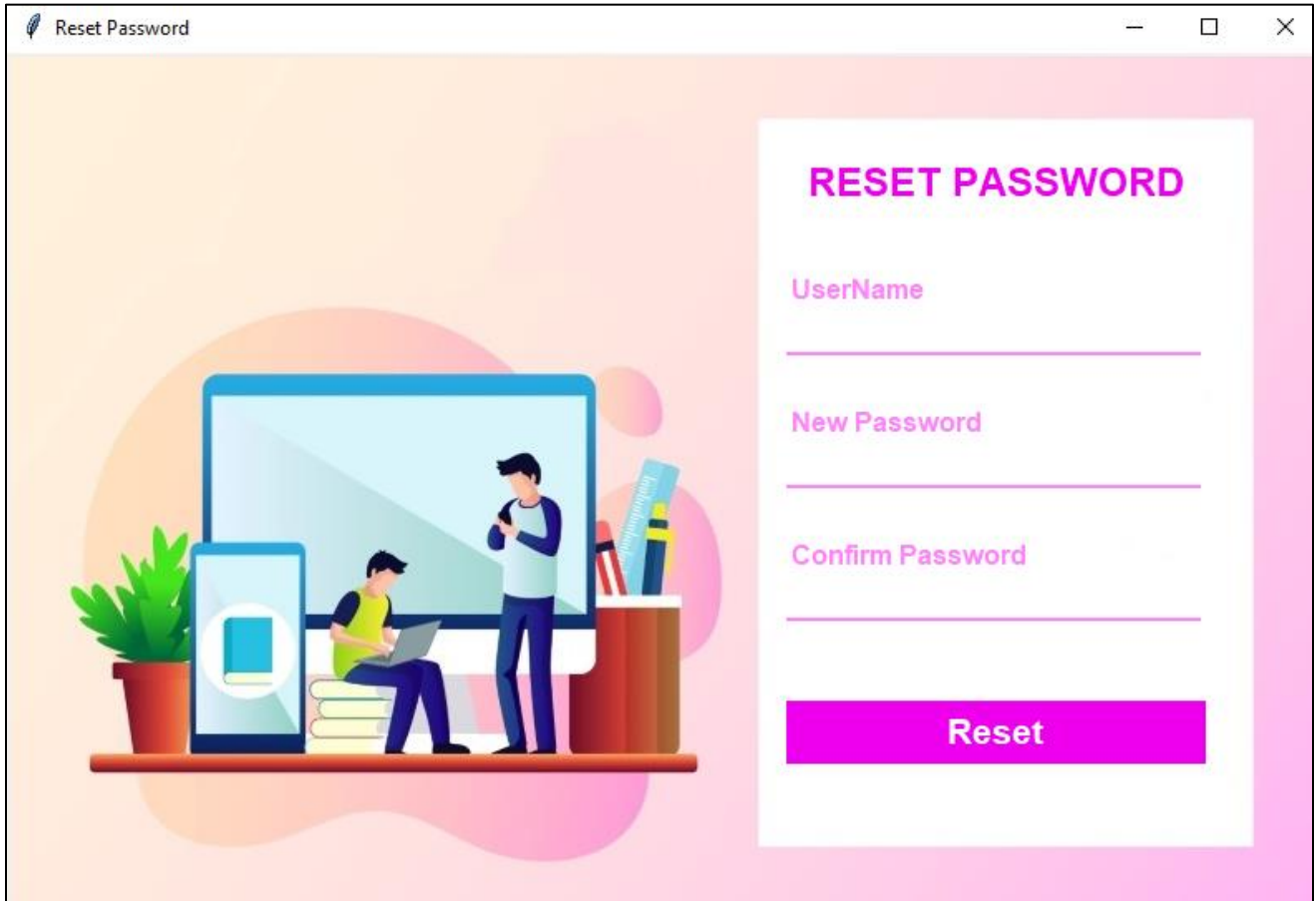
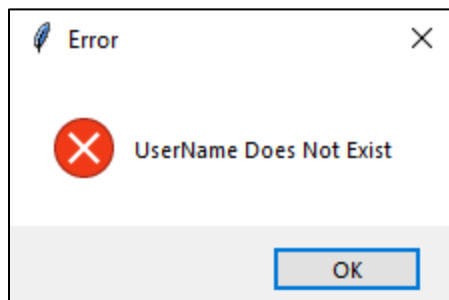


Figure 21 : ResetPassword

Voici quelque messages d'erreur et de succès :



L'utilisateur doit saisir son User Name correctement ou bien créer un nouveau compte.

2. Page d'aperçu

La page d'aperçu du jeu de données (Dataset Preview Page) est une fonctionnalité qui permet aux utilisateurs de visualiser et d'explorer les données contenues dans un jeu de données qui a utilisé et de les analyser plus en profondeur. Voici une description de cette page :

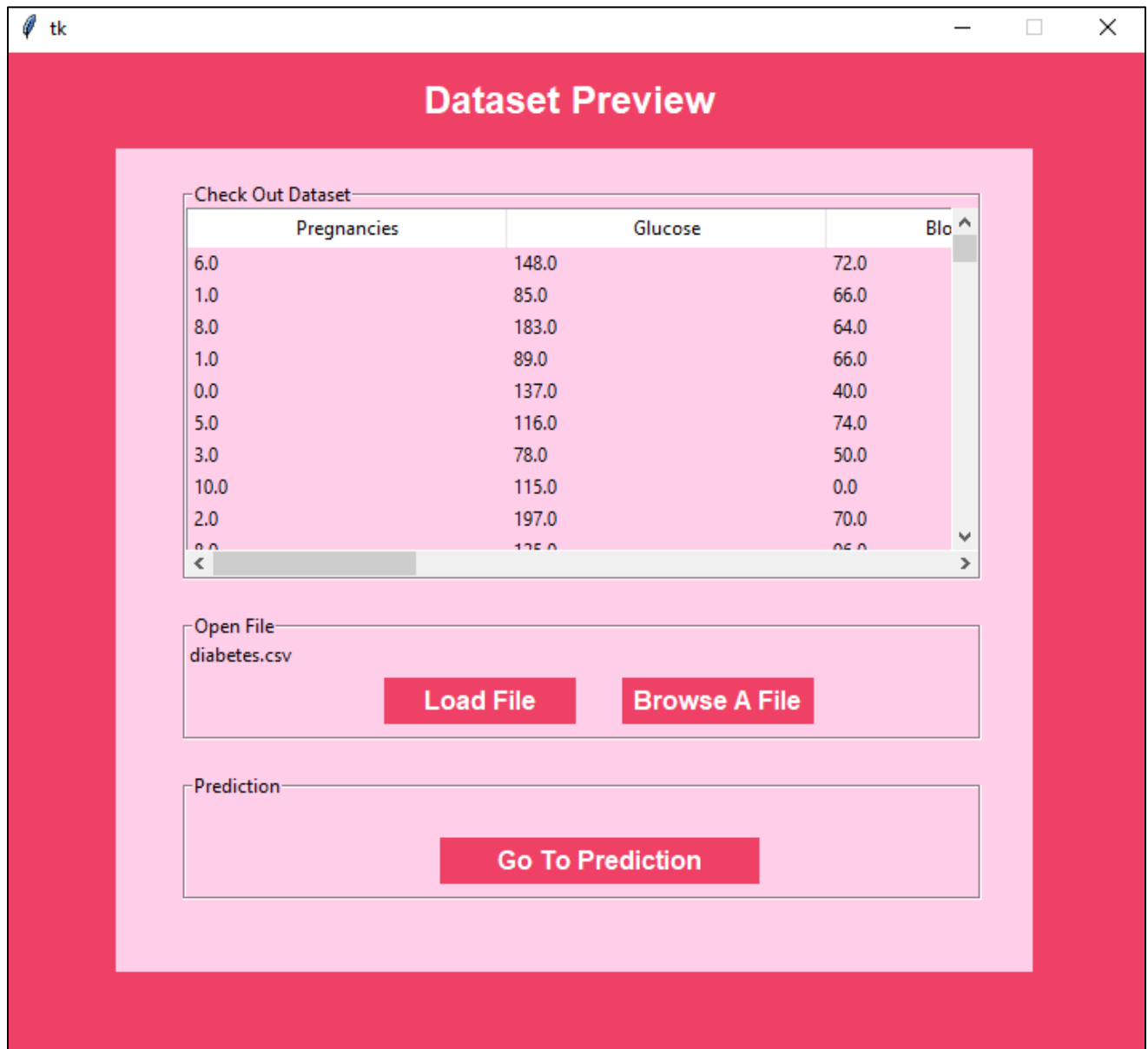


Figure 22 : PreviewPage

i. Tableau des données

Un champ sous forme tableau présente un aperçu des données contenues dans le jeu de données. Ce tableau peut inclure des colonnes représentant les différentes variables et attributs, ainsi que des lignes représentant les enregistrements individuels.

ii. Ouvrir un fichier

Le bouton "Browse a File" permet à l'utilisateur de parcourir les fichiers de son système d'exploitation pour sélectionner le fichier qu'il souhaite charger. Lorsque l'utilisateur clique sur ce bouton, une boîte de dialogue de sélection de fichier s'ouvre, lui permettant de naviguer dans les répertoires et de choisir un fichier contenant le jeu de données qui nous avons travaillé avec.

Le bouton "Load File" est utilisé pour charger le fichier sélectionné par l'utilisateur. Une fois que l'utilisateur a choisi un fichier en utilisant le bouton "Browse a File", il peut cliquer sur le bouton "Load File" pour lancer le processus de chargement du fichier dans le tableau des données .

iii. Bouton de prediction

Lorsque l'utilisateur clique sur ce bouton "Go To Prediction" l'application permet à l'utilisateur d'accéder aux page de prediction.

3. Formulaire de prédiction

Le formulaire de prédiction est un élément de notre interface graphique qui permet aux utilisateurs de saisir des informations pertinentes pour effectuer une prédiction . Voir une description de ce formulaire de prédiction :

The screenshot shows a web browser window titled "Diabetes Prediction". The main heading is "DIABETES PREDICTION" in bold black text. Below the heading, there is a form with eight input fields, each preceded by a label: "Pregnancies", "Glucose", "BP", "Skin TK", "Insulin", "BMI", "DiabetesPF", and "Age". Each input field is a simple text box. At the bottom of the form, there are two buttons: "Predict" and "Initialize", both in black text on a light blue background. The entire form is enclosed in a light blue border.

Figure 23 : PredictionPage

i. Formulaire de Prédiction

Lorsque vous accédez à cette page, vous trouverez une interface utilisateur conviviale avec huit champs de saisie où vous pourrez entrer les données nécessaires pour effectuer vos prédictions. Les champs peuvent varier en fonction du contexte spécifique de la prédiction, mais ils pourraient inclure que des informations telles que des valeurs numériques.

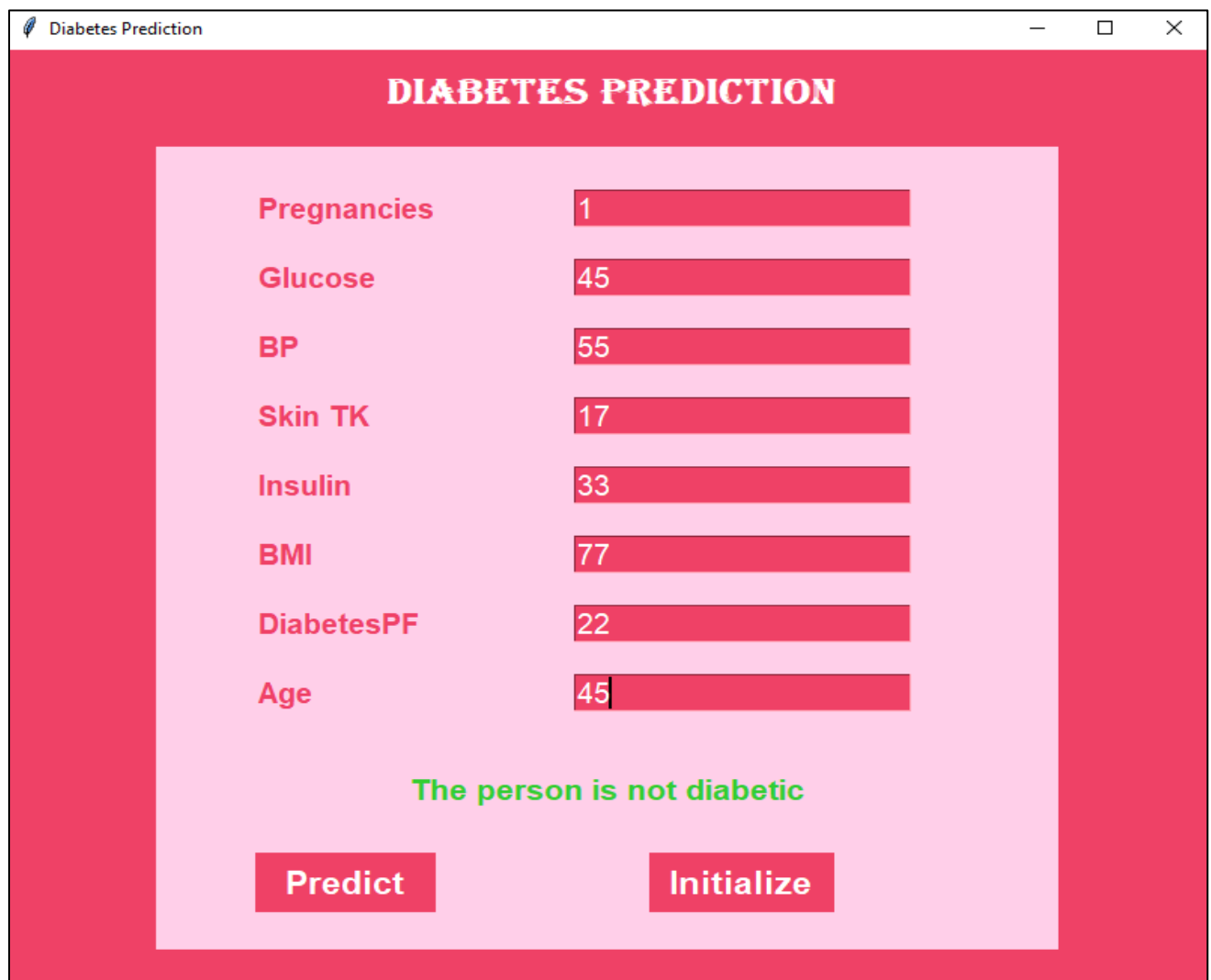
ii. Boutons

En plus des champs de saisie, nous verrons deux boutons situés à proximité. Le premier bouton est intitulé "Predict" (Prédire). Une fois que vous avez rempli tous les champs nécessaires, vous pouvez cliquer sur ce bouton pour déclencher le processus de prédiction. Ce processus utilisera un modèle prédictif spécifique pour analyser les données saisies et générer une prédiction en fonction des informations fournies.

Le deuxième bouton est appelé "Initialize" (Initialiser). Ce bouton vous permet de réinitialiser tous les champs de saisie à leurs valeurs par défaut. Il peut être utile lorsque vous souhaitez recommencer à zéro et saisir de nouvelles données sans avoir à supprimer chaque champ individuellement.

iii. Résultat de Prédiction

Une fois que vous cliquez sur le bouton "Predict" (Prédire), l'analyse est effectuée et le label situé en dessous des champs de saisie se met à jour automatiquement avec le texte approprié. Si la personne est diabétique, le texte "The Person is diabetic" s'affiche en rouge pour attirer l'attention sur cette condition. Dans le cas où la personne n'est pas diabétique, le texte "The person is not diabetic" apparaît en vert pour indiquer que la personne est exempte de cette maladie. Voici un exemple d'une Personne qui n'est pas malade :



The screenshot shows a web application window titled "Diabetes Prediction". The main heading is "DIABETES PREDICTION". Below it, there are eight input fields with labels and values: Pregnancies (1), Glucose (45), BP (55), Skin TK (17), Insulin (33), BMI (77), DiabetesPF (22), and Age (45). Below the input fields, the prediction result is displayed in green text: "The person is not diabetic". At the bottom, there are two buttons: "Predict" and "Initialize".

Parameter	Value
Pregnancies	1
Glucose	45
BP	55
Skin TK	17
Insulin	33
BMI	77
DiabetesPF	22
Age	45

The person is not diabetic

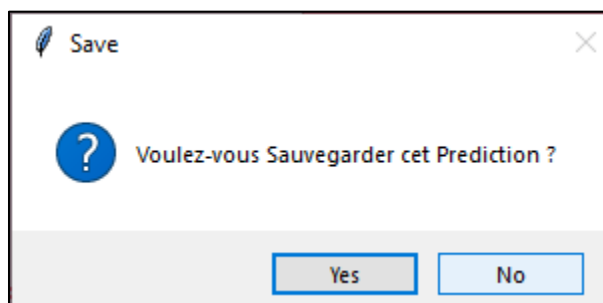
Predict **Initialize**

Figure 24 : Prediction Result

Cette conception visuelle avec des couleurs contrastées vise à fournir une indication rapide et claire du statut diabétique de la personne après la prédiction. Le label coloré facilite la lecture et permet aux utilisateurs de comprendre rapidement le résultat de la prédiction, sans avoir à examiner attentivement les chiffres ou les détails de la prédiction.

iv. Boîte de dialogue

Après avoir affiché les résultats de la prédiction sur la page, nous pouvons afficher une boîte de dialogue modale avec un message comme "Voulez-vous sauvegarder cette prédiction ?" accompagné de deux boutons, "Oui" et "Non".



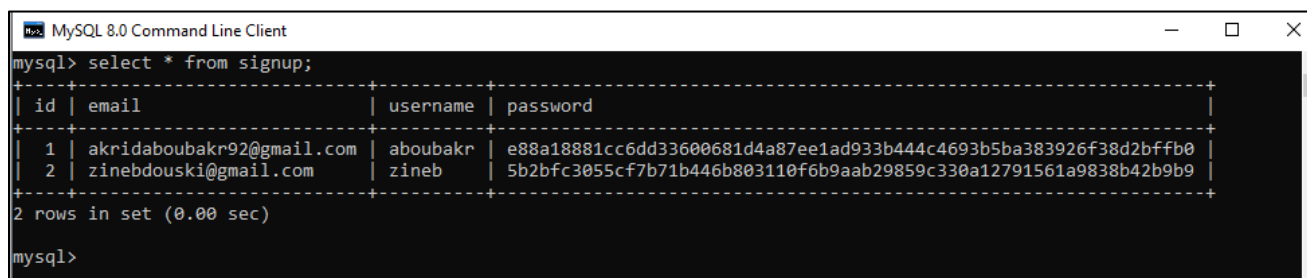
Si l'utilisateur clique sur le bouton "Oui", nous pouvons procéder à la sauvegarde de la prédiction dans notre base de données. Cela peut impliquer l'envoi des données pertinentes, telles que les champs saisis et le résultat de la prédiction, à une fonction qui les enregistrer dans notre base de données.

Si l'utilisateur clique sur le bouton "Non", vous pouvez simplement fermer la boîte de dialogue sans effectuer de sauvegarde et poursuivre normalement.

4. Base de données diabete

a) Table de système de connxion

La table "signup" est utilisée pour stocker les informations des utilisateurs qui se sont inscrits sur notre application. Elle contient des colonnes telles que "id" pour l'identifiant unique de l'utilisateur, "username" pour le nom d'utilisateur, "email" pour l'adresse e-mail, "password" pour le mot de passe idéalement stocké sous forme de hash pour des raisons de sécurité.



```
mysql> select * from signup;
```

id	email	username	password
1	akridaboubakr92@gmail.com	aboubakr	e88a18881cc6dd33600681d4a87ee1ad933b444c4693b5ba383926f38d2bffb0
2	zinebdouski@gmail.com	zineb	5b2bfc3055cf7b71b446b803110f6b9aab29859c330a12791561a9838b42b9b9

```
2 rows in set (0.00 sec)
```

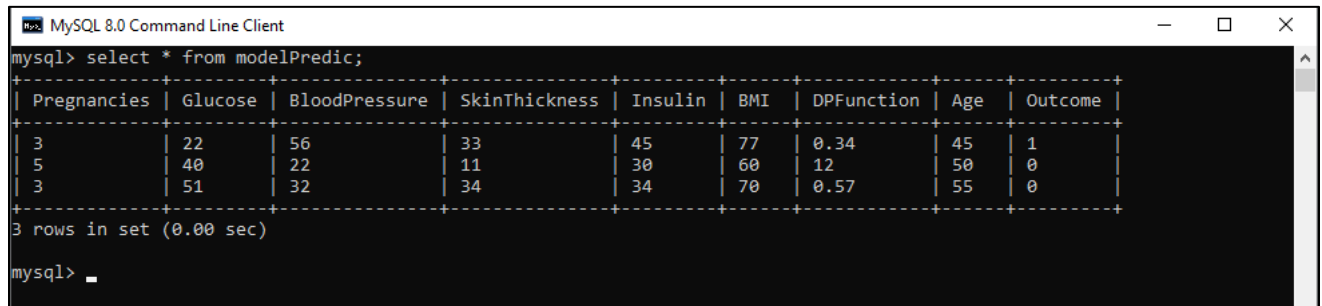
```
mysql>
```

Figure 25 : Signup Table

Cette table "signup" nous permet de gérer les comptes des utilisateurs et de stocker les informations nécessaires pour l'authentification et l'autorisation, afin de leur permettre de se connecter à notre application.

b) Table de Sauvegarde

La table "modelPerdic" est utilisée pour sauvegarder les nouvelles prédictions générées par notre modèle prédictif. Cette table inclure des colonnes telles que les huit champs de saisie avec les informations requises, "Outcome" pour le résultat de la prédiction.



The screenshot shows a MySQL 8.0 Command Line Client window. The command 'mysql> select * from modelPerdic;' has been executed. The output is a table with 9 columns: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DPFfunction, Age, and Outcome. There are 3 rows of data. Below the table, it says '3 rows in set (0.00 sec)'.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DPFunction	Age	Outcome
3	22	56	33	45	77	0.34	45	1
5	40	22	11	30	60	12	50	0
3	51	32	34	34	70	0.57	55	0

Figure 26 : Predict Table

En utilisant cette table "modelPerdic", nous pouvons enregistrer les prédictions générées par notre application, en les associant à l'utilisateur correspondant. Cela nous permet de conserver un historique des prédictions et de les utiliser à des fins d'analyse ultérieure, de suivi des performances du modèle ou de personnalisation de l'expérience utilisateur.

VI. Conclusion

A la rependre de la question précédente, il est clair que le machine learning joue un rôle crucial dans l'amélioration des soins de santé.

En développant des modèles de prédiction du diabète basés sur le machine learning, nous avons pu exploiter les vastes quantités de données de santé disponibles, notamment les symptômes, les résultats de tests et les antécédents médicaux, pour générer des prédictions précises. Ces modèles permettent une détection précoce du diabète, facilitant ainsi une intervention rapide et une gestion proactive de la maladie.

L'implémentation de ces modèles dans notre application desktop dédiée nous a permis de fournir une plateforme conviviale et pratique pour les professionnels de la santé. Ils peuvent désormais accéder rapidement et facilement aux prédictions du diabète, ce qui les aide à prendre des décisions éclairées concernant les traitements et les recommandations pour les patients diabétiques. Cela permet une personnalisation des soins en fonction des besoins individuels de chaque patient.

Bibliographie

CEED. (2023). Récupéré sur <https://ceed-diabete.org/fr/le-diabete/les-chiffres/>

DAP. (2023). Récupéré sur [https://dataanalyticspost.com/Lexique/svm/#:~:text=SVM%20\(Support%20Vector%20Machine%20ou%20de%20d%C3%A9tection%20d'anomalie.](https://dataanalyticspost.com/Lexique/svm/#:~:text=SVM%20(Support%20Vector%20Machine%20ou%20de%20d%C3%A9tection%20d'anomalie.)

IBM. (2023). Récupéré sur <https://www.ibm.com/fr-fr/topics/decision-trees>

ICTEA. (2023). Récupéré sur <https://www.ictea.com/cs/index.php?rp=%2Fknowledgebase%2F3500%2FiQue-es-MySQL.html&language=french>

JDN. (2023). Récupéré sur <https://www.journaldunet.fr/web-tech/guide-de-l-intelligence-artificielle/1501321-classification-naive-bayesienne-definition/>

Kaggle. (2023). Récupéré sur Kaggle: <https://www.kaggle.com/uciml/pima-indians-diabetes-database/data>

python. (2023). *python*. Récupéré sur <http://www.python.org/>

Souha, S. (2022). *Prédiabète : Un Système de Détection et prédiction de diabète*. Guelma: Université de 8 Mai 1945 – Guelma -.

Tawfik Beghriche, B. t. (2020). *Prédiction des maladies les plus fréquentes*. M'SILA: UNIVERSITÉ MOHAMED BOUDIAF - M'SILA.

univ-paris. (2023). Récupéré sur https://python.sdv.univ-paris-diderot.fr/18_jupyter/

wikipedia. (2023). Récupéré sur https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Accueil_principal