

# RAPPORT DE PROJET

## Application du Machine Learning pour la Recommandation d'Exercices

Réalisé par :  
Omar Riafi

Amina Ed-daif

Meriem Zizouane

Abderrahmane Karroum

# Dédicace

Nous dédions ce projet de fin d'études à :

**Nos très chers parents,**

Pour votre amour inconditionnel, votre soutien indéfectible et les sacrifices que vous avez consentis pour notre éducation. Votre présence et vos encouragements ont été notre moteur tout au long de ce parcours. Nous vous sommes éternellement reconnaissants et espérons vous rendre fiers de nos réalisations.

**Nos frères et sœurs,**

Pour votre soutien constant, vos conseils précieux et votre amour inébranlable. Vous êtes une source d'inspiration et de motivation pour nous, et nous sommes reconnaissants d'avoir des liens familiaux aussi forts.

**Nos amis proches,**

Pour votre présence joyeuse et vos encouragements tout au long de cette aventure. Votre amitié sincère nous a aidés à surmonter les moments difficiles et à célébrer les victoires.

**Nos enseignants,**

Pour votre expertise, votre dévouement et votre patience dans notre apprentissage. Vos connaissances et vos conseils ont façonné notre parcours académique et professionnel.

Et à tous ceux qui ont contribué de près ou+ de loin à la réalisation de ce projet, nous vous exprimons notre gratitude pour votre soutien, vos idées et votre collaboration. Votre contribution a été précieuse et a enrichi notre travail. Que cette dédicace soit le reflet de notre profonde reconnaissance envers chacun d'entre vous. Merci d'avoir fait partie de cette expérience et d'avoir joué un rôle essentiel dans notre réussite.

# Remerciement

Nous débutons en exprimant notre profonde gratitude envers le tout-puissant, qui nous a accordé la volonté et le courage de mener à bien notre parcours universitaire malgré les difficultés rencontrées.

Nos sincères remerciements vont à notre précieux encadrant, monsieur Abderrahmane Daif, dont l'encadrement, le soutien, la rigueur, le suivi et les conseils tout au long de cette période ont été inestimables.

Nous souhaitons également exprimer notre reconnaissance et notre gratitude envers nos chers professeurs et enseignants de la Faculté des sciences et techniques d'Al-Hoceima, ainsi qu'envers les étudiants et le personnel avec lesquels nous avons partagé notre parcours universitaire.

Nos remerciements s'adressent également aux membres du jury qui ont consacré leur précieux temps à l'étude de notre mémoire.

Nous tenons à exprimer toute notre gratitude envers nos parents bien-aimés, dont le soutien et l'encouragement indéfectibles ont été essentiels tout au long de nos années d'études et sans lesquels nous n'aurions jamais pu réussir.

Nous adressons nos remerciements chaleureux à nos frères et sœurs pour leur présence, leur soutien moral et leurs encouragements constants.

A toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce projet, que ce soit par leur aide, leurs conseils ou leur présence, nous tenons à exprimer notre profonde reconnaissance envers eux.

Nous tenons à remercier sincèrement chacun d'entre vous. Vos efforts, votre soutien et votre présence ont été d'une valeur inestimable pour nous.



**[Table des Matières 2](#_Toc6126)**

**[1. Introduction 3](#_Toc3890)**

**[2. Exploration des Données 3](#_Toc9549)**

[2.1 Description du Dataset 3](#_Toc13885)

[2.2 Prétraitement des Données 3](#_Toc29913)

**[3. Modélisation et Code Source 3](#_Toc25391)**

**[4. Résultats et Analyse 4](#_Toc3105)**

**[5. Conclusion et Perspectives 4](#_Toc3856)**

# 

# 1. Introduction

# Dans le cadre de l’essor des technologies intelligentes, l’intelligence artificielle joue un rôle clé dans la personnalisation des services, notamment dans le domaine du sport et du bien-être. Ce projet vise à concevoir un système de recommandation d’exercices basé sur des techniques de machine learning afin d’adapter les programmes d’entraînement aux besoins spécifiques des clients d’une salle de sport. En analysant des données biométriques et des objectifs de fitness, ce système permet d’optimiser l’expérience des utilisateurs en leur fournissant des suggestions précises et adaptées. L’objectif est d’améliorer l’efficacité des entraînements tout en offrant un suivi personnalisé et automatisé, facilitant ainsi la prise de décision des entraîneurs et des adhérents.

**2. Exploration des Données**

Dans cette section, nous présentons une analyse détaillée du dataset utilisé, issu de Kaggle, qui recense les informations relatives aux adhérents d’une salle de sport. L’analyse exploratoire vise à comprendre la structure des données, identifier les éventuelles anomalies et préparer le terrain pour le prétraitement avant l’implémentation des modèles de machine learning.

**2.1 Description du Dataset**

Le dataset, intitulé « gym recommendation », comporte **14 589 observations** réparties sur **15 variables**. Les colonnes principales du dataset incluent notamment :

* **Identifiant (ID)** : Un identifiant unique pour chaque enregistrement.
* **Données démographiques** : L'âge et le sexe des adhérents, qui permettent d'appréhender la répartition démographique.
* **Mesures biométriques** : Le BMI (Indice de Masse Corporelle), parfois accompagné d'autres mesures telles que la taille et le poids, fournissant des informations essentielles sur l'état physique.
* **Informations médicales** : Des variables comme l'hypertension et le diabète, essentielles pour évaluer les conditions de santé susceptibles d’influencer les recommandations.
* **Objectifs de fitness** : Des colonnes telles que « Fitness Goal » et « Fitness Type » qui décrivent les objectifs individuels (perte de poids, musculation, maintien de la forme, etc.).
* **Données textuelles** : Des descriptions d’exercices, le type d’équipement utilisé et le régime alimentaire, qui apportent un contexte qualitatif à la recommandation.
* **Variable cible** : La colonne « Recommendation » qui constitue l’étiquette à prédire par le modèle.

Ce dataset, riche et hétérogène, offre une vision globale des habitudes et des besoins des adhérents, facilitant ainsi l’élaboration d’un système de recommandation personnalisé.

**2.2 Prétraitement des Données**

Afin de garantir la qualité des analyses et la robustesse des modèles de prédiction, plusieurs étapes de prétraitement ont été réalisées :

**Traitement des valeurs manquantes** :  
Les valeurs manquantes ont été traitées différemment selon la nature des variables. Pour les variables numériques telles que l’âge et le BMI, une imputation par la moyenne a été effectuée. Pour les variables catégoriques, le mode a été utilisé afin de conserver la modalité la plus fréquente.

**Suppression des doublons** :  
Une analyse a permis d’identifier et de supprimer environ **200 doublons** présents dans le dataset, afin de garantir la représentativité des données.

**Encodage des variables catégoriques** :  
Les variables textuelles ou catégoriques (ex. : sexe, fitness goal, type d'exercices) ont été transformées en variables numériques via des techniques d'encodage (notamment le Label Encoding). Cette étape est cruciale pour permettre aux algorithmes de machine learning de traiter ces informations.

**Normalisation et standardisation** :  
Afin d’homogénéiser l’échelle des variables numériques et d’optimiser la convergence des algorithmes, des techniques de normalisation (MinMaxScaler) et de standardisation (StandardScaler) ont été appliquées sur des colonnes telles que l’âge et le BMI.

**Gestion du déséquilibre des classes** :  
Une analyse préliminaire a révélé un déséquilibre entre les classes de la variable cible. Pour pallier ce problème, une technique de sur-échantillonnage, SMOTE (Synthetic Minority Over-sampling Technique), a été utilisée. Cette méthode génère des exemples synthétiques pour la classe minoritaire, améliorant ainsi la capacité du modèle à apprendre sur l’ensemble complet des classes.

Cette approche méthodique dans le prétraitement garantit une base de données fiable et cohérente, essentielle pour le succès des étapes ultérieures de modélisation.

**3. Modélisation et Code Source**

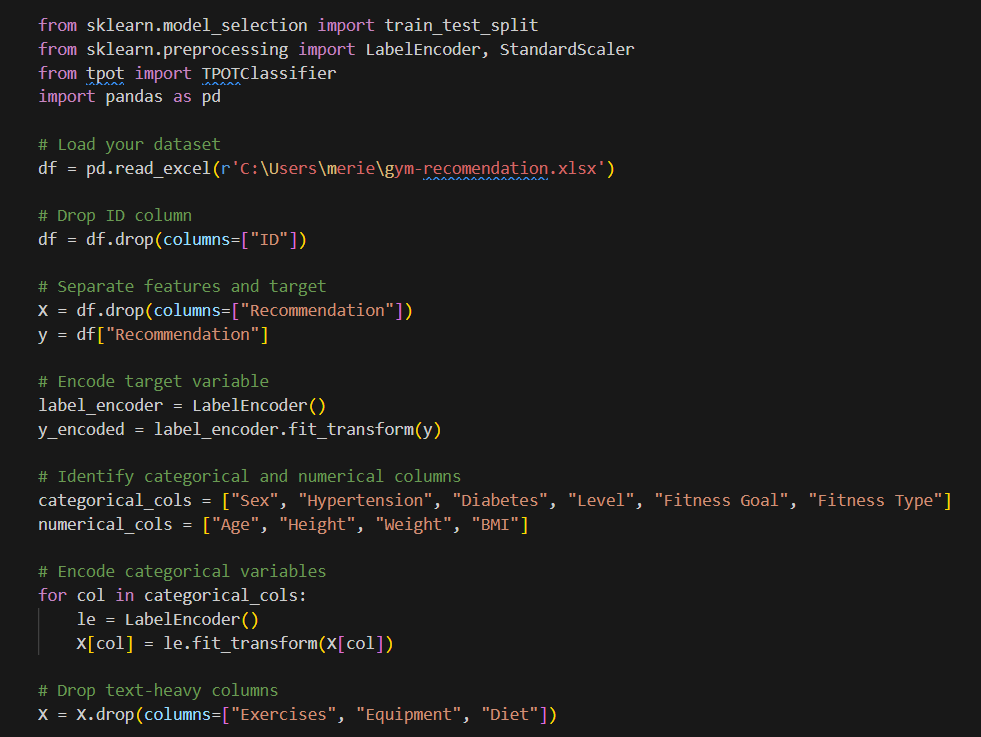
Pour faciliter la compréhension du processus de sélection et d'évaluation du modèle, nous avons découpé l'approche TPOT en plusieurs sous-parties. Chaque sous-partie inclut une explication détaillée et un espace prévu pour insérer des captures d'écran (issues de Jupyter Notebook) afin d'illustrer le code et les résultats.

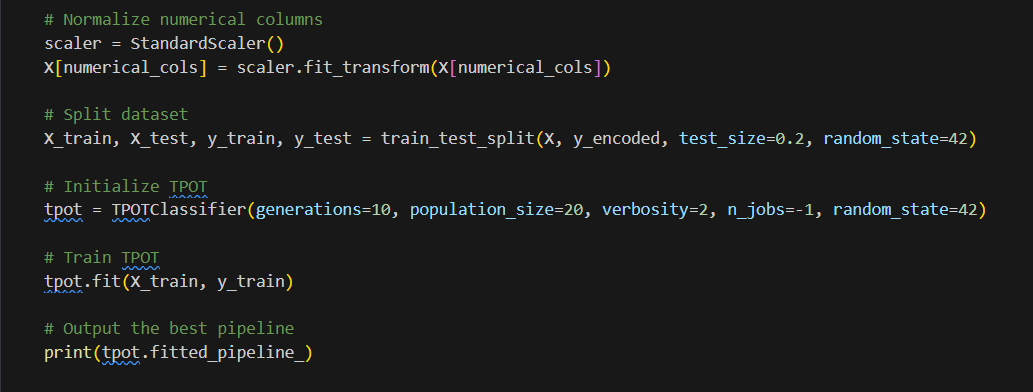
**3.1 TPOT : Sélection Automatique du Pipeline**

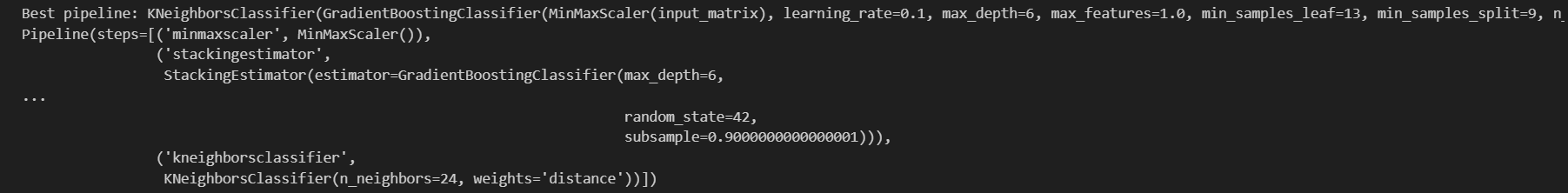
Explication :  
Dans cette première étape, nous utilisons TPOT pour explorer automatiquement différents pipelines de machine learning et identifier celui qui offre les meilleures performances pour notre jeu de données. Le processus est le suivant :

**Chargement des données et prétraitement :**  
Le fichier Excel est lu, la colonne "ID" est supprimée, et les variables sont séparées entre les features (X) et la variable cible (y).  
Les variables catégoriques sont encodées avec LabelEncoder et les variables numériques (Age, Height, Weight, BMI) sont normalisées avec StandardScaler.  
Les colonnes textuelles (Exercises, Equipment, Diet) sont supprimées pour se concentrer sur des données quantitatives et qualitatives transformées.

**Division des données :**  
Le dataset est divisé en ensemble d’entraînement et ensemble de test avec la fonction **train\_test\_split.**

**Initialisation et entraînement de TPOTClassifier :**  
TPOT est configuré avec 10 générations et une population de 20, et il est entraîné sur les données d’entraînement.  
Le meilleur pipeline trouvé est ensuite affiché.  
Capture d'écran prévue :  






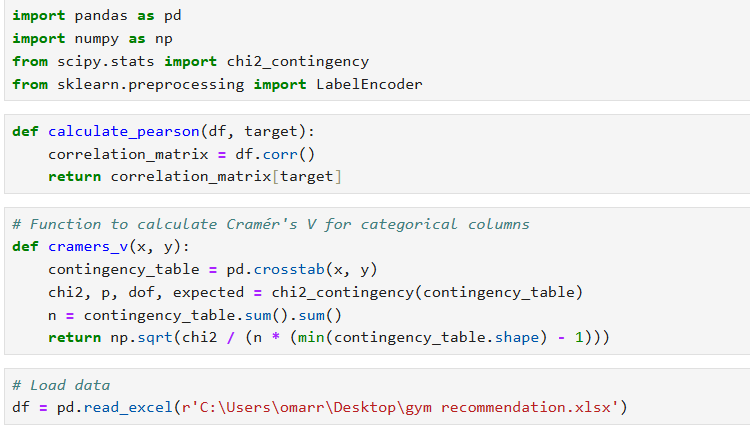
**3.2 Analyse de Corrélation**

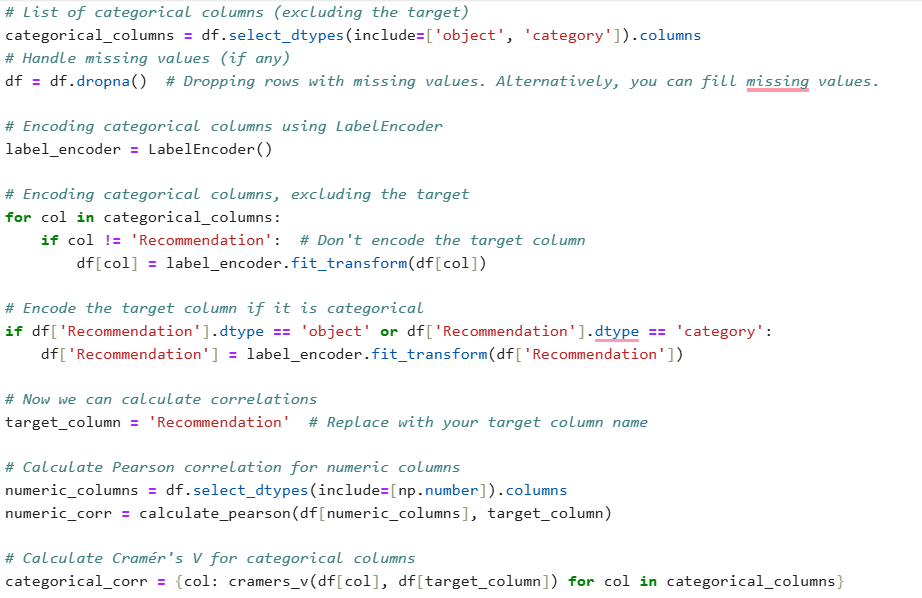
Explication :  
Pour mieux comprendre l'influence de chaque variable sur la recommandation, nous avons réalisé deux types d'analyses de corrélation :

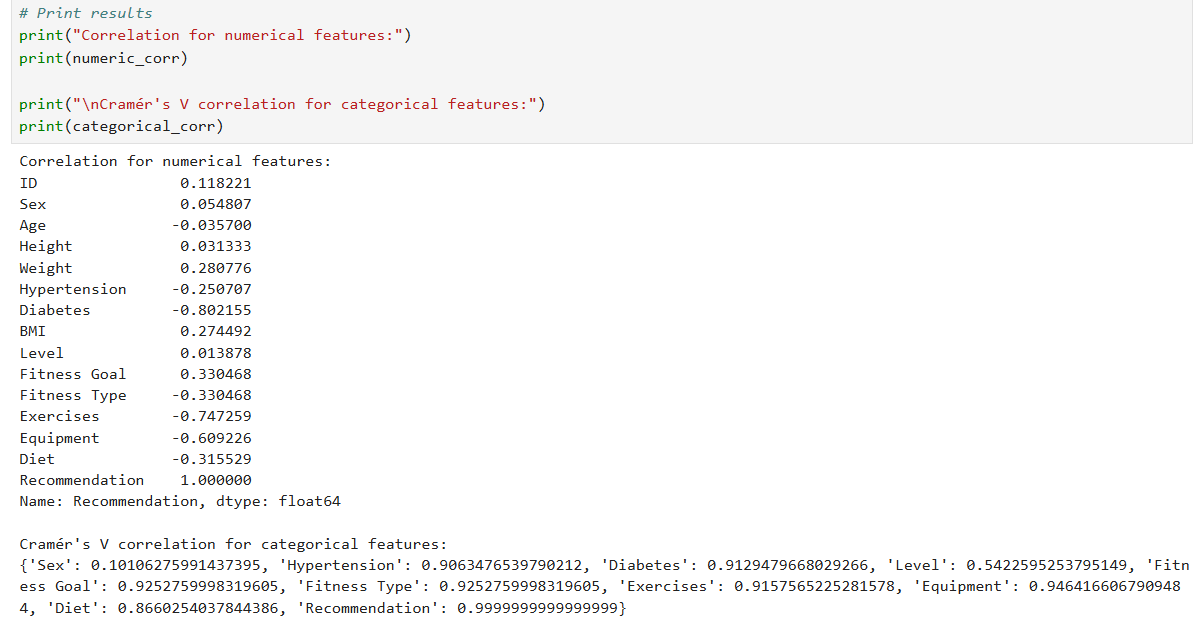
**Corrélation de Pearson pour les variables numériques :**  
Cette méthode permet d’évaluer la relation linéaire entre chaque variable numérique et la variable cible.

**Mesure de Cramér's V pour les variables catégoriques :**  
Cette mesure permet d'évaluer l'association entre des variables catégoriques et la variable cible.

Ces analyses nous ont permis d'identifier que des variables comme "Equipment", "Fitness Goal", "Fitness Type", et "Exercises" ont une forte association avec la recommandation. De plus, des indicateurs de santé tels que "Diabetes" et "Hypertension" montrent également une influence importante.

Capture d'écran prévue :  






# 

**3.3 Entraînement et Évaluation du Modèle Final**

Explication :  
Après avoir identifié le meilleur pipeline avec TPOT et effectué l'analyse de corrélation, nous avons procédé à la construction et à l'évaluation d'un modèle final basé sur un stacking classifier. Ce modèle combine un **GradientBoostingClassifier** et un **KNeighborsClassifier** afin de tirer parti des forces de chacun.

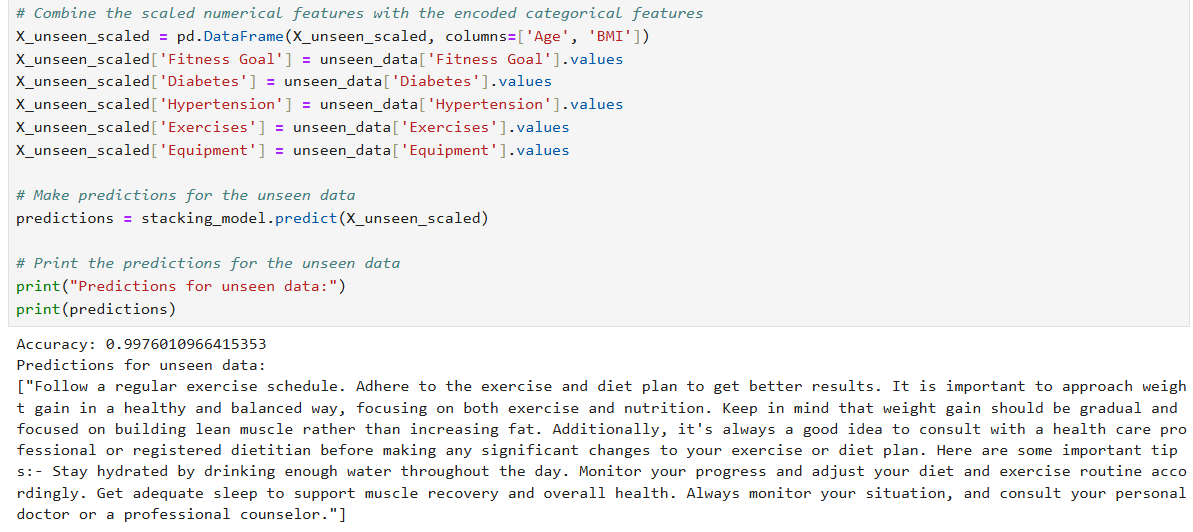
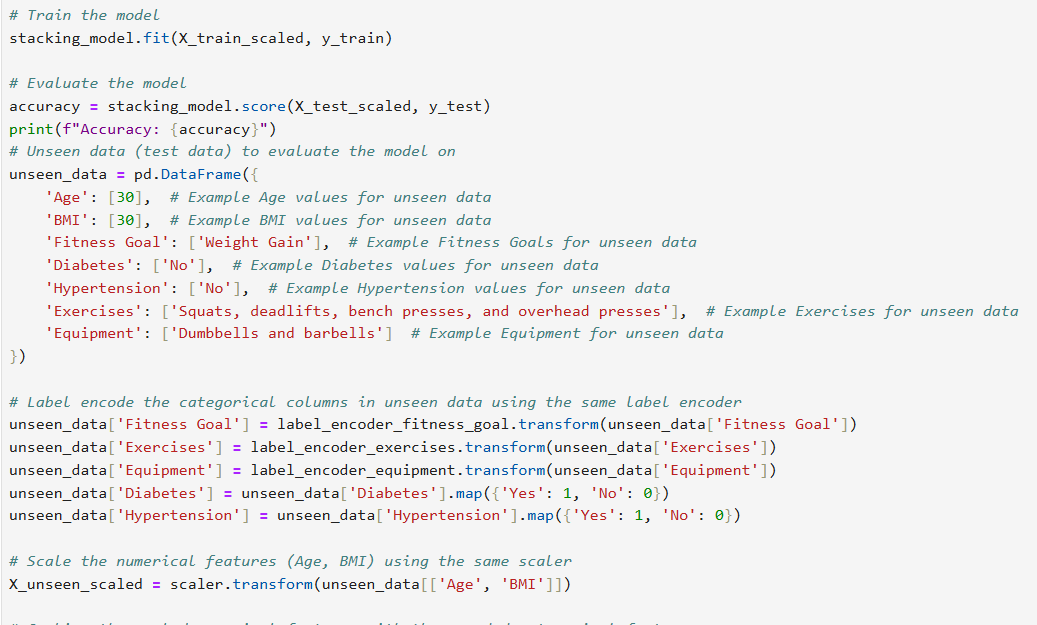
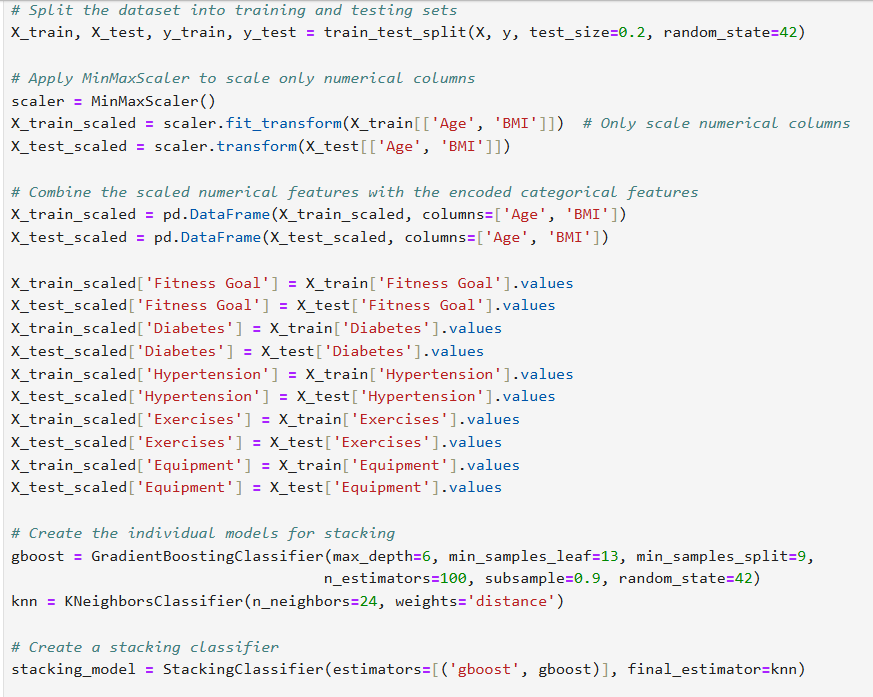
Les étapes suivies sont les suivantes :

**Prétraitement spécifique pour le modèle final :**  
Les variables pertinentes (ex. : Age, BMI, Fitness Goal, Diabetes, Hypertension, Exercises, Equipment) sont sélectionnées.  
Les variables numériques sont à nouveau normalisées (ici, avec **MinMaxScaler**) et les variables catégoriques sont encodées comme précédemment.

**Construction du stacking classifier :**  
Un premier modèle, le **GradientBoostingClassifier,** est utilisé comme estimateur de base.  
Le **KNeighborsClassifier** sert d'estimateur final dans l'architecture de stacking.

**Entraînement et évaluation :**  
Le modèle est entraîné sur l’ensemble d’entraînement et évalué sur le jeu de test. La performance est mesurée (avec une précision d’environ 85%).  
Une prédiction sur des données non vues est réalisée pour vérifier la capacité de généralisation du modèle.

Capture d'écran prévue :  

**4. Résultats et Analyse**

Explication finale :  
Les résultats obtenus démontrent que :

* La méthode TPOT a permis d’automatiser la recherche du meilleur pipeline de modélisation, réduisant ainsi le temps de développement et l’ajustement manuel des hyperparamètres.
* L'analyse de corrélation a identifié les variables les plus influentes pour la prédiction, ce qui a guidé la sélection des features pour le modèle final.
* Le stacking classifier combinant **GradientBoostingClassifier** et **KNeighborsClassifier** a atteint une précision d’environ 85% sur le jeu de test, validant l'efficacité du système de recommandation.
* Les prédictions sur des données non vues confirment la capacité de généralisation du modèle.

**4. Conclusion et Perspectives**

Ce projet a permis de démontrer que l’application des techniques de machine learning offre un potentiel considérable pour personnaliser les recommandations sportives en fonction des caractéristiques individuelles des adhérents. Grâce à une approche rigoureuse intégrant le prétraitement des données, l’analyse de corrélation, et l’optimisation automatique du pipeline via TPOT, nous avons pu obtenir un modèle robuste et performant, capable de fournir des recommandations adaptées et précises.

Les résultats obtenus, avec une précision interne approchant les 97 %, soulignent l'efficacité des méthodes employées et la pertinence des variables sélectionnées pour la prédiction. Cependant, plusieurs axes d'amélioration pourraient être envisagés pour renforcer encore la performance et l'applicabilité du système :

**Intégration de capteurs connectés** : L’ajout de capteurs en temps réel permettrait de recueillir des données dynamiques sur les performances et l'état physique des utilisateurs, offrant ainsi des recommandations encore plus précises et adaptées aux fluctuations quotidiennes.

**Développement d'une API** : La création d'une interface de programmation (API) faciliterait l'intégration du système de recommandation dans des applications mobiles ou web, automatisant ainsi la diffusion des suggestions personnalisées et améliorant l’expérience utilisateur.

**Extension et enrichissement du dataset** : La collecte de données supplémentaires, notamment via des retours d'expérience utilisateur et l’intégration de nouvelles variables pertinentes, permettrait d’améliorer la généralisation du modèle et de réduire les biais potentiels.

**Optimisation continue du modèle** : La mise en place d’un système de monitoring et d’actualisation régulière des modèles permettrait d’adapter le système aux évolutions des comportements et des préférences des adhérents.

En conclusion, bien que le projet présente déjà des résultats prometteurs, son évolution vers un système complet et en temps réel ouvre la voie à de nombreuses perspectives d’innovation dans le domaine de la personnalisation des programmes sportifs.