**Exercice semaine 7 - data science**

**Date limite de soumission des réponses: *jeudi 04 mai 2023 à 12h***

**Discussion en ligne/présentiel : *Vendredi 05 mai 2023 de 16h à 18h***

**Thématiques abordées : Scikit-learn et modèles de régression de machine learning**

**Instructions :** Utilisez Python ou R pour traiter les questions. Aucune autre application logicielle n’est acceptée pour l’analyse des données. ***Il est souhaitable de fournir votre éditeur de codes annoté avec votre devoir.***

Le travail de cette semaine porte sur la base des données nommée «Cameroon\_lung\_function » et disponible à l’adresse URL :

<https://raw.githubusercontent.com/pefura/IFPERA/main/Cameroon_lung_function.csv>

**Les codages utilisés pour les variables sont indiqués dans le tableau ci-dessous** :

|  |  |
| --- | --- |
| Variables | Explication et codes |
| ID | Numéro identification des participants |
| sex | 1=homme  2=femme |
| age | Age en années |
| ethnicity | Groupe ethnique  1=bantou  2= Soudano-sahélien  3=Autre |
| height | Taille en cm |
| fvc | forced vital capacity (ou capacité vitale force) en L |
| fev | forced expiratory volume in 1s (ou volume expiratoire maximal à la 1ère seconde) en L |
| fef2575 | forced expiratory flow between 25% and 75% of forced vital capacity (ou débit expiratoire moyen médian) en L/S |
| fevfvc | Rapport fev/fvc |

**Barème :** Toutes les questions sont notées sur 10 pts.

**I. Partie 1**

1. Importer la base de données « Cameroon\_lung\_function » dans votre environnement de travail et nommer cette base « data » après importation. Montrer la preuve de la réussite de l’importation. Vous pourriez utiliser l’importation dans avec le logiciel R pour l’importation directe à partir de l’adresse URL.
2. Sélectionner la sous-base de données (que vous nommerez data\_male) contenant uniquement les données des **sujets de sexe masculin** et les variables **age, height et fev.**
3. Décrire numériquement et graphiquement les 3 variables de la base « data\_male ».
4. Décrire graphiquement la relation entre fev (variable expliquée) et l’âge d’une part et entre fev et la taille d’autre part. Quelles conclusions pouvez-vous tirer ?
5. Quelles sont les métriques utilisées pour comparer les modèles de régression en machine learning.

**II. Partie II**

Dans cette partie les données de data\_male seront utilisées pour développer les modèles de régression permettant de prédire **« fev ».**

1. Faire une analyse prédictive de « fev » avec les modèles suivants : K Neighbors Regressor, Gradient Boosting Regressor et Random Forest Regressor.
2. Quel est le modèle le plus performant des 3 modèles étudiés à la question 6.
3. Construire une application permettant de prédire fev médian ainsi que les limites inférieures et supérieures de fev (le modèle Gradient Boosting Regressor permet de calculer les limites supérieures et inférieures).

**The right way to learn data science is to do data science**

**Good luck**