**Exercice semaine 8 - data science**

**Date limite de soumission des réponses: *jeudi 11 mai 2023 à 12h***

**Discussion en ligne/présentiel : *Vendredi 12 mai 2023 de 16h à 18h***

**Thématiques abordées : modèles de régression de machine learning et mise en production**

**Instructions :** Utilisez Python 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 sauf les questions 7 et 8 qui sont notées sur 20 pts.

**I. Partie 1 - Développement des modèles**

Nous voulons prédire le rapport fev/fvc chez les sujets de sexe féminin en fonction de l’âge et de la taille.

1. Sélectionner la sous-base de données (que vous nommerez dataset) contenant uniquement les données des **sujets de sexe féminin** et les variables **age, height, fev fvc.**
2. Calculer fev/fvc pour tous les sujets de sexe féminin et inclure cette variable dans dataset.
3. Décrire graphiquement la relation entre fev/fvc (variable expliquée) et l’âge d’une part et entre fev/fvc et la taille d’autre part. Quelles conclusions pouvez-vous tirer ?
4. Comparer le gradient boosting et le réseau de neurones artificiels pour la prédiction du rapport fev/fvc.

**II. Partie I2- Mise en production des modèles**

Dans cette partie nous décidons d’utiliser le gradient boosting pour la prédiction de l’espérance et des limites inférieure et supérieure (IC à 90%) du fev/fvc.

1. Quelles sont les contributions relatives de chaque variable explicative ?
2. Définir ou construire une fonction de prédiction de l’espérance et des limites inférieure et supérieure de fev./fvc.
3. Mettre en production à l’aide de la librairie « streamlit » le modèle de gradient boosting permettant de faire les prédictions ci-dessus. Il s’agit de développer une application avec interface utilisateur permettant d’obtenir directement les prédictions souhaitées en fonction des caractéristiques fournies. Le lien de l’application développée devrait être fourni.

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

**Good luck**