**Exercice semaine 4-data**

**Date limite de soumission des réponses: *Jeudi 13 avril 2023 à 12h***

**Discussion en ligne  : *Vendredi 14 avril 2023 à 16h***

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

**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 « pleuresie ».

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

|  |  |
| --- | --- |
| Variables | Explication et codes |
| **Les noms des variables sont explicites** | Si les codes ne sont pas précisés :  0 = absent , 1 = présent |
| age | Age en années |
| gender | 1=homme  2=femme |
| duration\_symptoms | Durée des symptômes en mois  1 : inférieur à 30 j  2 : de 30 à 60 j  3 : de 60 à 90 j  4 : supérieure à 90 j |
| HIV | 0=négatif  1= positif |
| weight | Poids en kg |
| height | Taille en m |
| **pleural\_proteins** | Taux de protéines dans le liquide pleural en g/l |
| **pleural\_CRP** | Taux de CRP dans le liquide pleural en mg/l |
| **pleural\_LDH** | Taux de LDH dans le liquide pleural en UI/l |
| **pleural\_glucose** | Taux de glucose dans le liquide pleural en g/l |
| **pleural\_TB** | Tuberculose pleurale  0= non  1=oui |

**Barème :** Toutes les questions sont notées sur 5 pts.

1. Importer la base de données « pleuresie » dans votre environnement de travail et nommer cette base « data » après importation. Montrer la preuve de la réussite de l’importation.

Cette table de données est disponible à l’adresse URL : <https://raw.githubusercontent.com/pefura/IFPERA/main/base_pleuresie.csv>

1. Calculer l’indice de masse corporelle (IMC) de l’ensemble de participants et créer une base de données « data\_1 » comprenant l’IMC et toutes les variables de « data » sauf le poids et la taille.

**Les données de data\_1 seront utilisées pour développer les modèles de classification.**

Dans la suite du devoir, il est question de développer les modèles de machine learning capables de distinguer les malades ayant une tuberculose pleurale de ceux qui n’ont pas de tuberculose pleurale.

1. Installer la bibliothèque (ou la librairie) «scikit-learn » et montrer la preuve de la réussite d’installation. Toutes les informations utiles pour l’utilisation de cette librairie sont disponibles à l’adresse : <https://scikit-learn.org/stable/>
2. Expliquer les bases du modèle de classification « d’arbre de décision » ou « decision tree ».
3. Expliquer les bases du modèle de classification « Random Forest».
4. Développer et Evaluer un modèle de « Random Forest» pour prédire le diagnostic de la tuberculose pleurale.
5. Développer et Evaluer un modèle de « Régression logistique» pour prédire le diagnostic de la tuberculose pleurale.
6. Comparer les deux modèles développés ci-dessus.
7. Faire une sélection des variables pertinentes expliquant le meilleur modèle.

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

**Good luck**