**Exercice semaine 5-data**

**Date limite de soumission des réponses: *vendredi 28 avril 2023 à 12h***

**Discussion en ligne : *Samedi 29 avril 2023 à 18h***

**Thématiques abordées : Classification en 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. Vous devez recopier chaque question avec les points affectés et nommer votre fichier.***

La base des données de santé respiratoire dans la population générale camerounaise vous a été confiée et est disponible au format csv à l’adresse URL :

<https://raw.githubusercontent.com/pefura/IFPERA-data-science/main/ESR_EFR_GLOBAL_DAL_sauvegarde_race_neutral_other.csv>

Les variables d’intérêt sont consignées dans le tableau de codage suivant :

**Tableau de codage**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables** | **codes** | **commentaires** | **Nom de codes à donner après importation de la base** |
| SEXE\_REC | 1=masculin  2 = féminin |  | sex |
| Age | / | En années | age |
| I.3\_plus\_haut\_niveau\_education\_en\_tranches | 1= non scolarisé  2= primaire  3=secondaire  4= universitaire |  | education |
| TOUX\_CHRONIQUE\_REC | 1= oui  2 = non |  | chronic\_cough |
| dyspnea\_any | 0= non  1 = oui |  | dyspnea |
| III.1\_avezvousdéjaeula\_tuberculose | 1= oui  2 = non | Tuberculose pulmonaire | past\_tuberculosis |
| II\_siflements\_durant\_12\_derniers\_mois | 1= oui  2 = non |  | wheezing |
| CUISSON\_TYP\_COMBUST\_REC | 1 = exposés à la biomasse  2 = non exposé à la biomasse |  | biomass |
| TABAC\_RECC | 1 = fumeur  2 = non-fumeur | Tabagisme | smoking |
| SPO2 | / | Saturation transcutanée en oxygène en % | spo2 |
| Bmi\_copd | / | Indice de masse corporelle en kg/m2 | bmi |
| OLD | 0= non  1= oui |  | old |

**Barème :** 100 points

**Questions**

Dans ce projet, il vous ai demandé de prédire l’existence ou non de l’obstruction bronchique (**old**) chez les participants de cette étude. Vous devez exclusivement utiliser les variables d’intérêt indiquées dans le tableau de codage.

**Consignes :** Les 2 modèles de machine learning à entrainer sont : le KNN et le Random Forest.

1. Renommer les variables d’intérêt tel que spécifier dans la dernière colonne du tableau de codage et nommer la base obtenue « data » (5 pts).
2. Faire une analyse exploratoire des données (10 pts).
3. Définir un pipeline permettant de traiter simultanément toutes les colonnes des variables explicatives en appliquant un encodage one hot sur les variables qualitatives et une normalisation robuste sur les variables quantitatives (10pts).
4. Utiliser une technique de validation croisée pour choisir les hyperparamètres optimaux (**le score AUC-ROC sera choisi pour l’optimisation**) des différents modèles le cas échéant (10 pts)
5. Tracer les courbes d’apprentissage pour les différents modèles optimaux et interpréter ces courbes. Il est conseillé d’inclure les 2 courbes sur le même graphique (10 pts).
6. Définir une fonction d’évaluation vous permettant de calculer les différentes métriques de classification dans le test set (20 pts).
7. Fournir un tableau montrant les métriques de classification pertinentes pour tous les modèles testés (20 pts).
8. Prédire l’existence de l’obstruction bronchique chez les sujets dont les caractéristiques sont consignées dans le tableau unseen\_data disponible à l’adresse URL : https://raw.githubusercontent.com/pefura/IFPERA-data-science/main/unseen\_data\_old.csv (15 pts).

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

**Good luck**