**Exercice semaine 5-data**

**Date limite de soumission des réponses: *Jeudi 20 avril 2023 à 12h***

**Discussion en ligne : *Vendredi 21 avril 2023 à 16h***

**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 copier chaque question avec les points affectés.***

La base des données de la tuberculose au Cameroun de 2012 à 2015 vous a été fournie à l’adresse URL : <https://raw.githubusercontent.com/pefura/IFPERA/main/cohort_TB_death.csv>

Les codes utilisés pour les variables de cette base de données sont fournis ci-dessous :

**CODES:** GENDER (sexe du participant, 1= male, 2= female); AGE in years (age du participant); weight (poids en kg, en classes) ; TB\_FORM (forme de la tuberculose, 1= EPTB, 2= SNPTB, 3= SPPTB); VIH (test du VIH, 1=positive, 2=negative, 3=unknown); TYPEPATIE(type de patient, 1=new cases, 2=retreatment cases, 3=others, 9=unknown); HOSPINIT ( hospitalisation en phase de début de traitement, 1= yes, 2=no, 9=unknown); FOLLOW\_UP\_DURATION in days (durée de suivi); status( statut à la fin du suivi, 0=alive, 1=died, 2=unknown); status\_rec (statut à la fin du suivi regroupé en deux modalités, 0=alive/unknown, 1=died)

La base des données est nommée data par la suite. Vous devez faire attention au choix de la variable dépendante. La variable FOLLOW\_UP\_DURATION n’est pas importante pour le développement des modèles.

**Barème :** 100 points

**Questions**

**Consignes :** Les 4 modèles de machine learning à entrainer sont : la régression logistique, le KNN, le Random Forest et la machine à vecteurs de support (SVM).

1. Faire une analyse exploratoire des données (10 pts).
2. 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).
3. Utiliser une technique de validation croisée pour choisir les hyperparamètres optimaux (le score F1 sera choisi pour l’optimisation) des différents modèles le cas échéant (10 pts)
4. Tracer les courbes d’apprentissage pour les différents modèles optimaux et interpréter ces courbes. Il est conseillé d’inclure les 4 courbes dans le même graphique (20 pts).
5. Définir une fonction d’évaluation vous permettant de calculer les différentes métriques dans le test set (20 pts).
6. Fournir un tableau montrant les métriques de classification pertinentes pour tous les modèles testés (20 pts).
7. Prédire le décès chez les sujets dont les caractéristiques sont consignées dans le tableau ci-dessous (10 pts).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| GENDER | AGE | weight | TB\_FORM | VIH | TYPEPATI | HOSPINIT |
| male | 65 | missing | 3 | 2 | 1 | 2 |
| male | 36 | missing | 2 | 1 | 1 | 1 |
| female | 32 | 80 Kg | 2 | 1 | 2 | 2 |
| female | 32 | missing | 1 | 1 | 1 | 1 |
| female | 30 | 60 Kg | 2 | 1 | 2 | 2 |
| female | 30 | < 50 Kg | 2 | 1 | 1 | 2 |
| female | 25 | missing | 3 | 2 | 1 | 2 |
| male | 22 | 43 Kg | 1 | 2 | 2 | 1 |

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

**Good luck**