**Evaluation fin de module 1 data science**

**Année académique : 2022-2023**

**Date limite de soumission des réponses: *lundi 29 mai 2023 à 10h***

**Barème : 100 pts**

La base des données des cas de tuberculose suivi à l’hôpital Jamot de Yaoundé de 2001 à 2021 est fournie à l’adresse URL : <https://raw.githubusercontent.com/pefura/IFPERA-data-science/main/base_TB_20years.csv>

Le tableau de codage est indiqué ci-dessous. La colonne « Remarques » indique les modalités de regroupement de certaines classes des variables.

Les variables d’intérêt pour ce devoir sont : sexcode, agecode, weight, typpat, Diag, TBsite, ISSTrait, RVIH, lieusuivi.

**Tableau de codage**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables** | **signification** | **codes** | **Remarques** |
| idcode | Numéro d’identification | / |  |
| sexcode | sexe | 1 = Homme  2= Femme |  |
| agecode | Age en années | / |  |
| weight | Poids en kg | / |  |
| adcode | Lieu de résidence | 1= Yaoundé  2= hors Yaoundé |  |
| gprisq | Groupes à risque | 1 Réfugié  2 Détenus  3 Déplacés internes  4 Personnels de santé  5 diabete  6 Non renseigné  7 aucun  ------------- |  |
| datdb | date de debut du traitemet | / |  |
| prot | protocole therapeutique | 1 4RHEZ/2RH  2 6RHEZ  3 2RHEZ/10RH  4 12RHEZ  5 2RHEZ/1RHEZ/5RHE  6 RHZ  7 mdr autre |  |
| typpat | Type de patient | 1 Nouveau cas  2 CAS à retraiter  3 AUTRE |  |
| casretraite | Type de cas de retraitement | 1 rechute  2 reprise apres abandon  3 echec  ------------- |  |
| Diag | Modalités du diagnostic | B bactériologiquement prouvée  C Cliniquement prouvée |  |
| TBsite | Localisation de la tuberculose | 1 pulmonaire  2 extrapulmonaire  3 multifocale |  |
| expulm | Site de localisation de la tuberculose extrapulmonaire | 1 pleurale  2 ganglionnaire  3 péritonéale  4 neuromeningé  5 mal de pott  6 articulaire  7 pericardite  8 non renseigné  9 image de miliaire pulmonaire  10 autre |  |
| micros | Résultats examen microscopique | 1 1+  2 2+  3 3+  4 Negatif  5 non fait |  |
| V39 | Date résultats microscopie | / |  |
| testmol | Résultats tests moléculaires (PCR) | 1 PCR TBLAMP POSITF  2 PCR XPERT POSITIF  3 Non fait |  |
| ISSTrait | Issue du traitement | 1 gueri  2 traitement terminé  3 echec  4 perdus de vue  5 décédé  6 transfert out  7 non precisé  8 MDR (multirésistant) | **Succès du traitement =** gueri + traitement termnié |
| dateissus | Date issue de traitement | / |  |
| RVIH | Test de VIH | CP CONNU POSITIF  NP Non connu positif  NG Negatif  NF Non fait  ID INDETERMINE | **VIH positif =** CONNU POSITIF + Non connu positif  **VIH negatif =** Negatif  **VIH non connu =** Non fait + Indeterminé |
| lieusuivi | Lieu suivi initial | 1 AMBULATOIRE  2 Hospitalisé  3 NON RENSEIGNE |  |

**Question :**

Il vous est demandé de développer un modèle de machine learning pour la prédiction de la probabilité d’être perdu de vue (la variable dépendante devrait être créé à partir de la variable « ISSTrait ») au cours du traitement de la tuberculose.

Les modèles de Random Forest et de régression logistique seront testés et le meilleur modèle sera choisi.

L’AUC et le score F1 seront utilisés pour le choix du modèle le plus performant.

Vous devriez décrire les étapes nécessaires et justifier les méthodes utilisées.