**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.

**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.