

TD – Pratique Analyse exploratoire

Objectif du dataset

L'objectif de cette analyse est de prédire si un patient est atteint de diabète (**Outcome**) en fonction de ses caractéristiques médicales et de santé. En construisant un modèle de classification sur ce dataset, nous cherchons à identifier les patients à risque en se basant sur des facteurs tels que la pression artérielle, le taux de glucose, l'indice de masse corporelle (BMI), etc.

Description du dataset

- **Pregnancies** : Nombre de grossesses (applicable aux femmes uniquement).
- **Glucose** : Niveau de concentration de glucose dans le sang.
- **BloodPressure** : Pression artérielle diastolique (mm Hg).
- **SkinThickness** : Épaisseur du pli cutané tricipital (mm).
- **Insulin** : Niveau d'insuline sérique (mu U/ml).
- **BMI** : Indice de masse corporelle (poids en kg divisé par la taille en m²).
- **DiabetesPedigreeFunction** : Score basé sur les antécédents familiaux de diabète.
- **Age** : Âge du patient (en années).
- **Outcome** : Indique si le patient est diagnostiqué avec le diabète (1) ou non (0). C'est la variable cible.

1. Exploration des Données

- Chargez le jeu de données "**diabetes_dataset.csv**".
- Affichez 10 lignes du jeu de données.

- Affichez les informations générales
- Affichez les statistiques descriptives des variables. Quelles sont vos interprétations
- Vérifier s'il y a des valeurs manquantes et les traiter en les remplaçant par le mode des valeurs existantes (pourquoi ? et la moyenne ?)
- Vérifier s'il y a des valeurs aberrantes (outliers) dans une colonne (ex. BMI) en utilisant la méthode interquartile
- Traiter en les remplaçant par la médiane des valeurs existantes
- Normalisez les données numériques (StandardScaler).

2. Visualisation et Analyse Exploratoire des Données (EDA)

- Créez **trois visualisations** pour comprendre les relations entre les variables (exemple : histogramme du taux de glucose en fonction de **Outcome**, boxplot de **BMI**, heatmap des corrélations).