Analyse de données **2025-2026**
Préorientation MIC 3ème année

PROJET D'ANALYSE DE DONNEES

Deadline : 16 janvier 2026

Description du projet

Dans ce projet, on étudie un jeu de données qui fournit un aperçu détaillé des routines d'exercice, des attributs physiques et des mesures de la condition physique de 973 membres d'une salle de sport. Ce jeu de données comprend les variables suivantes :

— **gender** : Sexe du membre de la salle de sport (homme ou femme)
— **weight** : Poids du membre en kilogrammes
— **height** : Taille du membre en mètres
— **duration** : Durée de chaque séance d'entraînement en heures
— **calories** : Total des calories brûlées au cours de chaque séance
— **fat** : Pourcentage de graisse corporelle du membre
— **water** : Consommation quotidienne d'eau pendant les séances d'entraînement
— **level** : Niveau d'expérience, de débutant (1) à expert (3)
— **bmi** : Indice de masse corporelle (IMC), calculé à partir de la taille et du poids

Dans ce projet, vous répondrez aux questions suivantes :

- Décrivez l'ensemble du jeu de données en précisant la nature des variables.
- Faites une analyse uni-dimensionnelle et bi-dimensionnelle du jeu de données. Certaines variables sont-elles liées ? Une attention particulière sera portée sur le choix des représentations, et sur l'interprétation des résultats présentés.
- Visualisez les individus dans un espace de plus faible dimension à partir des variables quantitatives à l'aide d'une analyse en composantes principales. Interprétez le lien entre les méta-variables et les variables initiales. Vous pouvez exploiter les variables qualitatives pour l'interprétation.
- Clustering basé sur les variables quantitatives :
  - ▶ Faites un clustering des individus avec une méthode de classification hiérarchique
  - ▶ Faites un clustering des individus avec une méthode de type Kmeans
  - ▶ Comparez les deux classifications retenues, issues des deux questions précédentes
  - ▶ Interprétez ces classifications vis-à-vis des variables qualitatives du jeu de données.

Consignes

Vous rendrez par **binôme d'un même groupe** (ou trinôme, un seul possible par groupe d'effectif impair)
— un rapport au format **pdf** de 20 pages maximum tout compris
— le fichier Quarto qui l'a généré.

Les deux documents devront être intitulés **gpX-Nom1-Nom2** (ou **gpX-Nom1-Nom2-Nom3**) où **X** est à remplacer par la lettre du groupe de TD (A à E). Ils seront déposés sur la page moodle du cours (aucun retour par mail ne sera accepté).

Remarques : Gardez en tête que vous devez rendre un travail synthétique et clair qui nous permet d'évaluer les compétences listées ci-après. Toute sortie (table, figure, ...) doit être commentée. Au vu du nombre de pages limité, faites des choix pertinents et travaillez la mise en forme de votre rapport.

Modalités d'évaluation

Vous serez évalués sur la présentation et la rédaction du rapport, sur la pertinence des choix des représentations (à argumenter) ainsi que sur l'interprétation des différentes sorties obtenues (graphiques ou autres). Vous serez également évalués sur la manipulation de R et de Quarto. Plus précisément, vous serez évalués sur les compétences suivantes.

Compétences transversales

- Rédaction : Savoir mener un argumentaire clair et concis. Savoir justifier un raisonnement. Penser à définir toutes notations utilisées.
- Modélisation : Savoir modéliser une situation :
  - Identifier la nature des variables (qualitative nominale/ordinale, quantitative discrète/continue)
  - Définir un modèle statistique
- Logiciel R : Savoir mener l'étude d'un jeu de données grâce à R
- Quarto : Savoir rédiger un rapport en Quarto pour une analyse reproductible

Statistiques descriptives unidimensionnelle et bidimensionnelle

- Maîtriser les définitions des indicateurs usuels de statistique descriptive (moyenne, mode, variance, quantiles, fonction de répartition empirique, covariance, corrélation...)
- Savoir choisir les indicateurs et représentations adaptés aux données
- Savoir mener une interprétation des graphiques usuels de statistique descriptive (histogrammes, boxplots, barplots, diagramme en secteur, matrice de corrélation, mosaicplot,...)

Analyse en composantes principales (ACP)

- Maîtriser le vocabulaire de l'ACP : inertie, inertie axiale, axes principaux, composantes principales, plan factoriel
- Maîtriser les spécificités de l'ACP centrée et l'ACP centrée réduite
- Maîtriser le principe de l'ACP :
  - Diagonalisation de la matrice ΓM
  - Lien entre les valeurs propres et inerties axiales
  - Lien entre les vecteurs propres et les axes factoriels
- Maîtriser la définition des graphiques issus de l'ACP :
  - Projection des individus sur un plan factoriel
  - Corrélations des variables avec les composantes principales
- Savoir mener une interprétation des graphiques issus de l'ACP :
  - Interprétation individuelle de chaque graphique
  - Interprétation croisée des différents graphiques

Classification non supervisée (clustering)

- Connaître et savoir appliquer les différentes méthodes de clustering (Kmeans, DBSCAN, CAH) et leurs variantes
- Savoir calibrer les paramètres et choisir le nombre de classes d'une méthode de clustering, à l'aide de différents critères
- Savoir interpréter les classes données par une méthode de clustering
- Savoir comparer des clusterings