Projet d'étude

4 ModIA
UFs Analyse de données & Eléments de modélisation statistique
Intervenantes : Cathy Maugis-Rabusseau et Béatrice Laurent-Bonneau

2025-2026

Organisation du projet et documents à rendre

- Le projet sera réalisé par groupe de 3 étudiant-e-s. La constitution des groupes sera donnée lors de la première séance.
- 6 séances de 2h30 sont dédiées dans votre emploi du temps au travail du projet. L'une de nous sera présente lors de chacune de ces séances pour répondre à vos questions.
- Livrables : vous devrez rendre (en déposant sous Moodle) au plus tard le **mercredi 28 janvier 2026 minuit** les 2 documents suivants :
  1. un fichier quarto (*nom1-nom2-nom3-Rapport.qmd*) contenant les codes R et générant le rapport au format pdf.
  2. un rapport au format .pdf (*nom1-nom2-nom3-Rapport.pdf*) généré par la compilation du fichier .qmd précédent. Attention : le rapport est limité à 25 pages, figures incluses.
- Un dossier "ModeleRapport", disponible sur Moodle, vous donne un exemple avec des consignes pour la rédaction de votre rapport. Il est important d'en prendre connaissance dès la première séance !

Evaluation du projet

Pour chaque UF, la note de projet compte pour un tiers de la note finale de l'UF. Elle sera issue de l'évaluation des critères suivants :

Critère	**UF EMS**	**UF AD**
Utilisation pertinente d'indicateurs/graphiques pour l'exploration des données		x
Utilisation pertinente des méthodes factorielles		x
Utilisation pertinente de méthodes de clustering adaptées à la question traitée		x
Choix des modélisations ML et MLG adaptées à la question traitée	x	
Ecriture mathématique des modèles considérés en ML et MLG	x	
Choix des procédures de tests adaptées à la question traitée	x	
Utilisation de méthodes de sélection de variables	x	
Aller-retour exploration $\leftrightarrow$ modélisation	x	x
Analyse ($\neq$ lecture !) des résultats obtenus	x	x
Choix et rendu des graphiques illustratifs	x	x
Rédaction d'un document en quarto		x
Programmation en R	x	x
Rédaction générale du document	x	x
Bonus pour des choix originaux adaptés	x	x

Jeu de données étudié

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 :

— **age** : Age du membre de la salle de sport
— **gender** : Sexe du membre de la salle de sport (homme ou femme)
— **weight** : Poids du membre en kilogrammes
— **height** : Taille du membre en mètres
— **bpm_ave** : Fréquence cardiaque moyenne pendant les séances d'entraînement
— **duration** : Durée de chaque séance d'entraînement en heures
— **calories** : Total des calories brûlées au cours de chaque séance
— **type** : Type d'entraînement effectué (cardio, strenght, yoga, HIIT).
— **fat** : Pourcentage de graisse corporelle du membre
— **water** : Consommation quotidienne d'eau pendant les séances d'entraînement
— **freq** : Nombre de séances d'entraînement par semaine
— **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

Questions à aborder

Dans votre rapport final, vous devez avoir abordé par une/des méthodes adaptées les questions suivantes :

- Analyse descriptive du jeu de données :
  — Faites quelques statistiques descriptives du jeu de données. Préciserez la nature des variables, adapterez vos choix à la nature des variables et étudierez les corrélations / liaisons entre les variables
  — Visualisez les individus dans un espace de faible dimension à partir des variables quantitatives. Interprétez.
  — Visualisez les individus dans un espace de faible dimension à partir des variables qualitatives. Interprétez.
- Tests d'hypothèses :
  — Le nombre de calories brûlées suit-il une loi normale ?
  — Utilisez un test approprié afin de déterminer si le nombre de calories brûlées dépend du genre.
  — Reprenez la question précédente en vous concentrant sur chaque sport pris séparément.
  — Le type de sport pratiqué dépend-il du genre ? Le niveau d'expérience dépend-il du genre ?
- Clustering des individus :
  — Obtenez un clustering des individus à partir des variables quantitatives à l'aide d'une méthode hiérarchique, d'une méthode de type Kmeans et d'une modélisation par mélanges gaussiens respectivement.
  — Comparez les trois clusterings retenus à l'aide d'indicateurs et par une méthode factorielle. Interprétez ces clusterings à l'aide des variables du jeu de données.
- Etude du nombre de calories brulées :
  — Modélisez le nombre de calories brulées en fonction des autres variables quantitatives. Visualisez les résidus et commentez. En utilisant les mêmes variables, proposez un modèle pour corriger la tendance présente dans les résidus.
  — Modélisez le nombre de calories brulées en fonction des deux variables qualitatives : "level" et "type".
  — Modélisez le nombre de calories brulées en fonction de toutes les variables explicatives et utilisez une méthode de sélection de modèles que vous connaissez pour simplifier le modèle. Visualisez les résidus et commentez. Peut-on améliorer le modèle ?
- Etude du niveau d'expérience :
  — La variable "level" comporte 3 modalités : $1, 2$ et 3. Créez une variable binaire, que vous nommerez "levelBis" en regroupant les niveaux d'expérience 2 et 3 en une seule catégorie.
  — Modélisez la variable "levelBis" en fonction des autres variables (n'oubliez pas de supprimer la variable "level" du jeu de données !)
  — Utilisez une procédure de sélection de modèles pour simplifier le modèle précédent. Analysez les résultats obtenus avec ce modèle.