Analyse de données **2024-2025**
Préorientation MIC 3ème année

PROJET D'ANALYSE DE DONNEES

Deadline : 17 janvier 2025

Description du projet

Un biologiste vous demande de faire une analyse descriptive du jeu de données **DataProjet3MIC-2425.txt** disponible sur la page moodle du cours. Dans ce jeu de données, on observe pour $G = 542$ gènes d'une plante modèle des différences d'expression à différents temps, différents traitements et pour deux réplicats biologiques (répétition biologique de l'expérience). Ainsi, on mesure Y_{gtsr} la différence d'expression du gène g entre le traitement $t \in \{T1, T2, T3\}$ et un état sans traitement de référence, au temps $s \in \{1h, 2h, 3h, 4h, 5h, 6h\}$ pour le réplicat $r \in \{R1, R2\}$. Ces mesures sont organisées dans les colonnes Tt_sh_Rr.

A l'aide d'un test d'analyse différentielle, on a pour chaque traitement t la décision si le gène est sur-exprimé ("Sur" - forte différence d'expression positive), sous-exprimé ("Sous" - forte différence d'expression négative) ou non-exprimé ("Non" - pas de différence d'expression significative) à 6h. Ces informations sont dans les colonnes ExpT1, ExpT2 et ExpT3.

A noter que le traitement $T3$ est une combinaison des traitements $T1$ et $T2$.

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.
- Analyse des Tt_sH_Rr :
  - ▶ Menez une analyse en composantes principales où les Tt_sH_Rr sont les individus décrits par les gènes.
  - ▶ Faites une classification non supervisée (clustering) de ces données afin de regrouper les Tt_sH_Rr en plusieurs classes homogènes.
- Analyse des gènes :
  - ▶ Préliminairement, construisez un jeu de données *DataExpMoy* contenant la moyenne des expressions sur les réplicats de chaque gène, pour chaque traitement et chaque heure. *DataExpMoy* est donc une matrice de taille $G \times 18$. Vous pourrez utiliser les variables ExpT1, ExpT2 et ExpT3 pour commenter vos résultats des questions suivantes.
  - ▶ Menez une analyse en composantes principales pour les gènes à partir du jeu de données *DataExpMoy*.
  - ▶ Faites une classification non supervisée (clustering) des gènes à partir de leur expression (*DataExpMoy*) afin d'obtenir des classes de gènes homogènes (ayant la même évolution d'expression).
  - ▶ Faites une classification non supervisée (clustering) des gènes à partir des variables ExpT1, ExpT2 et ExpT3. Comparez avec les résultats de la question précédente.

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 Rmarkdown 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.

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 RMarkdown. 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
- RMarkdown : Savoir rédiger un rapport en RMarkdown 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