UNIVERSITE VIRTUELLE DU BURKINA FASO
(UV-BF)

CITADEL

Master Fouilles de Données et Intelligence Artificielle
(FD & IA)

Projet de Groupe de Statistique & Probabilités

Armel SOUBEIGA, PhD
ML Engineer

@mail : armel.soubeiga@yahoo.fr

Projet : Analyse exploratoire et regroupement des conceptions d'une pièce d'avion

Durée : 15 jours – du 1er au 15 juillet 2025
Travail en groupe de 3 à 5 étudiants
Support à rendre : Un notebook Jupyter (Python ou R) et un court rapport synthétique (PDF ou Markdown intégré)

Contexte métier

Vous êtes data scientist dans une entreprise spécialisée dans la conception et fabrication de moteurs d'avion civils et militaires. Votre équipe travaille sur la conception d'une pièce critique du moteur, sujette à des itérations fréquentes afin d'optimiser sa performance. Chaque nouvelle conception est testée à travers une expérimentation de performance : deux indicateurs sont mesurés à cette occasion – notés γ_1 et γ_2.

Le département d'ingénierie souhaite mieux structurer les conceptions passées en identifiant des groupes homogènes de configurations, selon leurs caractéristiques techniques et leurs résultats expérimentaux.

Données disponibles

Vous disposez d'un ensemble de données tabulaire *dataset_anon.csv*. Dans ce dataset :

- Chaque ligne décrit les résultats d'expérimentation pour une itération de conception spécifique
- Colonne experiment_date : date à laquelle l'expérience a été réalisée
- Colonnes γ_1 et γ_2 indicateurs mesurés pendant l'expérience
- Colonnes feat_[A-Z] : caractéristique spécifique de la conception testée pendant l'expérience

Quelques informations clés sur les colonnes de features feat_[A-Z] :

- Les valeurs manquantes dans ces colonnes ne doivent pas être interprétées comme des données incomplètes. Elles indiquent plutôt que la feature n'est pas pertinente pour la conception en question, par exemple parce que la conception ne contient pas un sous-composant spécifique
- Les valeurs textuelles dans les colonnes catégorielles (par exemple "competition") n'ont pas de signification intrinsèque - ce sont des placeholders anonymisés pour les valeurs présentes dans le dataset original (source)

Objectif du projet

Votre mission est de mener une analyse exploratoire avancée et une classification automatique non supervisée des conceptions, à partir des données fournies.

1. Analyse exploratoire univariée et bivariée
   - Étudiez la distribution de chaque variable (qualitative et quantitative)
   - Étudiez les corrélations ou associations entre variables
   - Explorez les relations entre les caractéristiques de conception et les performances γ1, γ2

2. Préparation des données pour la classification automatique
   - Proposez un traitement cohérent des variables mixtes (ex : distance de Gower, ACM…)
   - Argumentez le choix de la méthode de transformation ou de distance

3. Réalisation d'une classification ascendante hiérarchique (CAH)
   - Proposez plusieurs critères d'agrégation (Ward, complet, simple…)
   - Visualisez les dendrogrammes

- Interprétez les groupes obtenus (ex : via les centres de classes ou graphiques de profil)

4. (Optionnel) Réalisation d'une analyse en composantes (ACP ou ACM)
   - Si pertinent, utilisez les coordonnées principales des individus comme base pour la CAH
   - Superposez les résultats de la classification sur les plans factoriels

5. Interprétation et conclusions
   - Décrivez et comparez les groupes obtenus
   - Analysez les relations avec $\gamma 1$ et $\gamma 2$
   - Évoquez l'usage potentiel des résultats pour le métier : nouvelles conceptions proches de groupes performants ? classes à écarter ?

Contraintes et livrables

- Le notebook doit être clairement commenté et reproductible
- Si vous manquez de temps : décrivez les étapes manquantes que vous auriez incluses
- Vous êtes libres d'utiliser Python (*pandas*, *seaborn*, *scipy*, *sklearn*, *prince*, *etc.*) ou R (*FactoMineR*, *cluster*, *tidyverse*, *etc.*)

Critères d'évaluation

Critère	Détail
Exploration des données	Clarté des graphes, qualité de l'analyse
Préparation et transformation	Justification des choix techniques
Méthodologie de classification	Choix des distances, méthode d'agrégation, découpage
Interprétation	Capacité à caractériser et justifier les groupes
Clarté du code / présentation	Bonne structure, commentaires utiles
Travail en groupe	Répartition claire, qualité globale