Analyse de données et Data mining

Intervenant(s) : Philippe Collart et Catherine Dehon

Orientations concernées :

- Statistiques appliquées à l'Epidémiologie
- Statistiques appliquées au Contrôle de qualité et validation des méthodes analytiques

Nombre d'heures (applications et exercices compris) : 32

Remarque : ce nombre ne tient pas compte d'éventuels exercices à préparer à domicile, ni de la préparation de l'examen.

Langue : Français

Contenu du module

Méthodes exploratoires

- Détection des valeurs aberrantes
- Analyses en composantes principales (ACP)
- Analyses factorielles des correspondances (ACOBI)
- Analyses factorielle des correspondances multiples (AFCM)
- Analyses factorielles des données mixtes

Méthodes de partitionnement – Méthodes de clustering

- Introduction à la validation croisée et aux problèmes de sur-ajustement
- Méthodes de clustering non hiérarchique (k-means, nuées dynamiques, etc.)
- Méthodes de clustering hiérarchique
- Arbre de décision
- Random Forest

Méthodes de régression

- Régression linéaire multiple
- Régression logistique
- Méthodes de régularisation (Lasso, Ridge, Elasticnet, PLS, etc.)

Séances d'exercices avec R (Rcmdr, factomineR, etc.)

Pré-requis : Manipulations des objets R, automatisation des données, sélection

Evaluation :

Travail personnel proposé par le candidat au jury de délibération, en lien avec sa pratique professionnelle et avec les matières enseignées (dans l'ensemble de l'UE 4)

Horaire et lieu : A déterminer