DR Grand Ouest
Formation Continue

Inserm
La science pour la santé
From science to health

Formation en STATISTIQUES « Méthodes statistiques pour les données massives – Big Data »

Principaux aspects des Big Data abordés : le volume des données et leur variété.
Principaux thèmes travaillés : Données de grande dimension, Données de nature hétérogène, Classification, Régression, Clustering

Dates et Lieu
Les jeu-ven 7-8 sept
ET le lun 18 sept 2023
(3 jours en 2j+1j)

En distanciel

Publics
Personnes souhaitant se mettre à jour sur les dernières méthodes statistiques pour les données massives (Big Data)

Pré-requis
Avoir des notions avancées en statistique (inférence, clustering, régression, classification), ainsi que des notions de R. Réfléchir à des problématiques, jeux de données ou documents susceptibles d'être utilisés en support lors de la formation.

Intervenant
Formateur de la Sté Arkesys

Support logiciels
Cette formation n'est pas dédiée à la pratique d'un logiciel particulier mais nous proposons de nous appuyer sur le logiciel R pour les exercices et les illustrations

OBJECTIFS :
- Réaliser une analyse de régression lorsque les données sont en grande dimension
- Réaliser un clustering en utilisant des modèles parcimonieux spécifiques à la grande dimension
- Réaliser une étude de classification sur des données de grande dimension
- Effectuer une sélection des variables pertinentes
- Effectuer ces analyses sous le logiciel R
- Extraire de l'information sur la problématique métier à partir des résultats de l'analyse

PROGRAMME :
Introduction au Big Data
> Les grandes évolutions de la statistique
> Les 3 V : Volume, Variété, Vélocité
> Les problèmes que cela engendre pour les techniques classiques

Big Data et Régressions
> Illustration de la problématique de la grande dimension
> Méthode de sélection de variables
> Méthodes de projection
> Méthodes de régularisation
> Comparaison de méthodes
> Régression sur données non quantitatives

Apprentissage non supervisé (clustering)
> Illustration de la problématique de la grande dimension
> Modèle de mélange parcimonieux
> Algorithme EM
> Sélection de modèles
> Sélection de variables
> Prise en compte de données hétérogènes

Apprentissage supervisé
> Illustration de la problématique de la grande dimension
> Sélection de variables
> Méthodes de projection
> Méthodes de régularisation
> Comparaison de méthodes
> Prise en compte de données hétérogènes

METHODES :
Explications théoriques suivies de pratiques guidées puis mises en autonomie, sur des thématiques et jeux de données proches des problématiques des participants

POUR Y PARTICIPER :
1/ S'inscrire sur SIRENE : https://www.sirene.inserm.fr/
(onglet « Agent formation », menu « Demander une formation », sous-menus « offre de formation continue & collective », domaine « TS4-Statistiques »)
2/ Compléter le questionnaire de pré-formation en ligne : Questionnaire Big Data

1/ & 2/ A FAIRE pour le 21/06/2023 au plus tard

CONTACT : Michèle HAYS – INSERM DR Grand Ouest – 02 40 35 86 80 – michele.hays@inserm.fr