DATA MINING AVEC R
Traitement avancé des données

DURÉE : 3 JOURS
21 heures - Présentiel

PRÉ-REQUIS
Stage Environnement R ou niveau équivalent

OBJECTIFS
Comprendre la démarche du Data Mining et de la modélisation statistique
Choisir entre la régression et la classification en fonction du type de données
Évaluer les performances prédictives d'un algorithme

PUBLIC VISÉ
Utilisateurs finaux, data scientists, statisticiens, analystes type Data Miner, acteurs impliqués dans l'analyse/fouille des données

MOYENS PÉDAGOGIQUES
- Diagnostic pédagogique amont permettant de regrouper les apprenants par niveau homogène et d'assurer la parfaite adéquation entre vos besoins et le programme
- Organisation en petit groupe de 1 à 6 apprenants maximum garantissant une meilleure interactivité
- 1 poste informatique par apprenant
- Environnement confortable, calme et propice à la concentration (dans nos salles)
- Supports de cours et documentation individuels de qualité (livrets papier)
- Enchaînement de : ① théorie
  ② démonstration par l'exemple
  ③ mise en pratique sur exercices
- Visualisation et partage des connaissances transmises par projection audiovisuelle en appui
- A l'issue de la formation, tous les exercices et leurs corrigés vous sont remis

MOYENS D'ENCADREMENT
Consultant formateur spécialisé, validé par notre équipe tant sur la capacité pédagogique que la connaissance technique métier

MOYENS D'ÉVALUATION
Diagnostic préalable des connaissances individuelles à partir d'un questionnaire de positionnement
Évaluation de l'atteinte des objectifs par l'apprenant
Évaluation du transfert des acquis par le formateur

ORGANISATION - Inter ou Intra

INTER-ENTREPRISES
Prix et dates sur calendrier si programmé

PRIX INTRA-ENTREPRISE
Sur devis - Programme adaptable en intra

CONTENU PEDAGOGIQUE

Introduction
Démarche CRISP-DM pour le Data Mining
Architectures type pour la mise en production de modèles prédictifs
Les types de données dans R
Importation-exportation de données

Analyse en composantes
Analyse en Composantes Principales
Analyse Factorielle des Correspondances
Analyse des Correspondances Multiples
Analyse Factorielle pour Données Mixtes
Classification Hiérarchique sur Composantes Principales

La modélisation
Les étapes de construction d'un modèle statistique
Les algorithmes supervisés et non supervisés
Le choix entre la régression et la classification

Procédures d'évaluation de modèles
Séparation des jeux d'apprentissage, de validation et de test
Test de représentativité des données d'apprentissage
Mesures de performance des modèles de régression
Matrice de confusion, la courbe ROC et AUC pour les modèles de classification

Les algorithmes non supervisés
La classification hiérarchique
Le clustering non hiérarchique (KMeans)
Les algorithmes supervisés
Le principe de régression linéaire univariée
La régression multivariée
La régression polynomiale
La régression logistique
Méthode des plus proches voisins (KNN)
Les arbres de décisions
Approche Naïve Bayésienne
Introduction aux réseaux de neurones

Analyse de données textuelles
Collecte et prétraitement des données textuelles
Extraction d'entités primaires et d'entités nommées
Étiquetage grammatical, analyse syntaxique, analyse sémantique
Lemmatisation Représentation vectorielle des textes Pondération TF-IDF