Data Analytics avec R
modélisation et représentation des données

Cours Pratique de 4 jours - 28h
Réf : DTA - Prix 2024 : 2 860€ HT

Le Big Data Analytics suppose la maîtrise de techniques fondamentales de traitement des données : méthodes statistiques, classifications, régressions, ACP... Ce stage pratique vous montrera, à partir de données concrètes, comment utiliser ces techniques pour construire puis évaluer des modèles à l'aide du langage R.

OBJECTIFS PÉDAGOGIQUES

À l'issue de la formation l'apprenant sera en mesure de :

Comprendre le principe de la modélisation statistique

Choisir entre la régression et la classification en fonction du type de données

Évaluer les performances prédictives d'un algorithme

Créer des sélections et des classements dans de grands volumes de données pour dégager des tendances

LE PROGRAMME

dernière mise à jour : 07/2021

1) Rappels au langage R

- Les types de données dans R.
- Importation-exportation de données.
- Techniques pour tracer des courbes et des graphiques.

Mise en situation : Prise en main des scripts et Notebooks.

2) 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.

Travaux pratiques : Mise en œuvre de la diminution du nombre des variables et identification des facteurs sous-jacents des dimensions associées à une variabilité importante.

3) La modélisation

- Les étapes de construction d'un modèle.
- Les algorithmes supervisés et non supervisés.
- Le choix entre la régression et la classification.

Travaux pratiques : Mise en place d'échantillonnage de jeux de donnes. Effectuer des tests d'évaluations sur plusieurs modèles fournis.

4) Procédures d'évaluation de modèles

- Les techniques de ré-échantillonnage en jeu d'apprentissage, de validation et de test.
- Test de représentativité des données d'apprentissage.
- Mesures de performance des modèles prédictifs.

PARTICIPANTS

Responsables Infocentre (Datamining, Marketing, Qualité...), utilisateurs et gestionnaires métiers de bases de données.

PRÉREQUIS

Connaissances de base en statistiques et en R, ou avoir suivi les stages "Statistiques, maîtriser les fondamentaux" (Réf. STA) et "Environnement R, traitement de données et analyse ... " (Réf. TDA).

COMPÉTENCES DU FORMATEUR

Les experts qui animent la formation sont des spécialistes des matières abordées. Ils ont été validés par nos équipes pédagogiques tant sur le plan des connaissances métiers que sur celui de la pédagogie, et ce pour chaque cours qu'ils enseignent. Ils ont au minimum cinq à dix années d'expérience dans leur domaine et occupent ou ont occupé des postes à responsabilité en entreprise.

MODALITÉS D'ÉVALUATION

Le formateur évalue la progression pédagogique du participant tout au long de la formation au moyen de QCM, mises en situation, travaux pratiques…
Le participant complète également un test de positionnement en amont et en aval pour valider les compétences acquises.

MOYENS PÉDAGOGIQUES ET TECHNIQUES

• Les moyens pédagogiques et les méthodes d'enseignement utilisés sont principalement : aides audiovisuelles, documentation et support de cours, exercices pratiques d'application et corrigés des exercices pour les stages pratiques, études de cas ou présentation de cas réels pour les séminaires de formation.
• À l'issue de chaque stage ou séminaire, ORSYS fournit aux participants un questionnaire d'évaluation du cours qui est ensuite analysé par nos équipes pédagogiques.
• Une feuille d'émargement par demi-journée de présence est fournie en fin de formation ainsi qu'une attestation de fin de formation si le stagiaire a bien assisté à la totalité de la session.

MODALITÉS ET DÉLAIS D'ACCÈS

L'inscription doit être finalisée 24 heures avant le début de la formation.

ACCESSIBILITÉ AUX PERSONNES HANDICAPÉES

Vous avez un besoin spécifique d'accessibilité ? Contactez Mme FOSSE, référente handicap, à l'adresse suivante psh-accueil@orsys.fr pour étudier au mieux votre demande et sa faisabilité.

- Matrice de confusion, de coût et la courbe ROC et AUC.
Travaux pratiques : Mise en place d'échantillonnage de jeux de donnes. Effectuer des tests d'évaluations sur plusieurs modèles fournis.

5) Les algorithmes non supervisés
- Le clustering hiérarchique.
- Le clustering non hiérarchique.
- Les approches mixtes.
Travaux pratiques : Traitements de clustering non supervisés sur plusieurs jeux de données.

6) Les algorithmes supervisés
- Le principe de régression linéaire univariée.
- La régression multivariée.
- La régression polynomiale.
- La régression régularisée.
- Le Naive Bayes.
- La régression logistique.
Travaux pratiques : Mise en œuvre des régressions et des classifications sur plusieurs types de données.

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

LES DATES

CLASSE À DISTANCE	PARIS
2024 : 18 juin, 01 oct., 03 déc.	2024 : 11 juin, 24 sept., 26 nov.