Formation Data Science (R et Hadoop)

Durée :	5 jours
Public :	Professionnels des bases de données, managers, analystes de données, data scientists et assistants à maîtrise d'ouvrage. Cette formation est très utile pour les professionnels chargés de gérer les prévisions et les tendances
Pré-requis :	Connaissances en matière de programmation et de statistiques sont utiles sans toutefois être obligatoires
Objectifs :	- Appliquer des techniques d'exploration des données pour améliorer la prise de décisions métier à partir de sources de données internes et externes - Prendre une longueur d'avance sur vos concurrents avec l'analyse des données structurées et non structurées - Prédire un résultat en utilisant des techniques d'apprentissage automatique supervisé
Sanction :	Attestation de fin de stage mentionnant le résultat des acquis
Taux de retour à l'emploi:	Aucune donnée disponible
Référence:	BUS100294-F
Note de satisfaction des participants:	Pas de données disponibles

Exploration et analyse des données avec R

Charger, interroger et manipuler des données avec R
Nettoyer les données brutes avant la modélisation
Réduire les dimensions avec l'analyse en composantes principales (ACP)
Développer les fonctionnalités de R avec les packages définis par l'utilisateur

Faciliter la pensée analytique avec la visualisation des données

Explorer les caractéristiques d'un ensemble de données à travers la visualisation
Représenter graphiquement la distribution des données avec des boîtes à moustaches, des histogrammes et des diagrammes de densité
Identifier les valeurs hors normes

Explorer les données non structurées pour les applications métier

Traitement préliminaire et préparation des données non structurées pour une analyse plus approfondie
Décrire un ensemble de documents avec une matrice termes-documents

Difficultés supplémentaires liées au Big Data

Examiner les architectures de MapReduce et Hadoop
Intégrer R et Hadoop à RHadoop

Estimer les valeurs avec les règles de régression linéaire et logistique

Modéliser la relation entre une variable de sortie et plusieurs variables d'entrée
Interpréter correctement les coefficients des données continues et qualitatives

Techniques de régression pour manipuler le Big Data

Traiter les ensembles de données volumineux avec RHadoop
Créer des modules de régression pour RHadoop

Identification automatique de chaque nouvel élément de données

Utiliser des arbres de décision pour prédire les valeurs cible
Appliquer des règles de probabilité pour prédire les résultats avec le modèle Naive Bayes
Combiner les variables de prédiction des arbres et les forêts aléatoires dans RHadoop

Évaluer les performances des modèles

Visualiser les performances des modèles avec une courbe ROC
Évaluer les modèles de classification avec des matrices de confusion

Identifier des groupes encore inconnus dans un ensemble de données

Segmenter le marché client avec l'algorithme K-Means
Trouver des similarités avec les mesures des distances
Créer des clusters en forme d'arbres et des mises en cluster hiérarchiques
Mettre en cluster les tweets et les fichiers texte pour mieux les comprendre

Mettre à jour les connexions avec l'analyse des associations

Identifier les connexions importantes avec l'analyse des réseaux sociaux
Comprendre l'utilisation des résultats de l'analyse des réseaux sociaux à des fins marketing

Définir et évaluer des règles d'association

Identifier les préférences réelles des clients à partir d'un ensemble de données transactionnelles pour améliorer l'expérience utilisateur
Calculer les indices de support et de confiance et le lift pour différencier les bonnes règles des mauvaises