EMGS GROUP
CONSULTING | RECRUITMENT | TRAINING

Contact
☎ 01 75 43 05 58
✉ training@emgsgroup.com

Data Science : niveau avancé

Code formation : EMDSA
Durée : 3 jours – 21h de cours
Format : Inter-entreprise*

3 jours
21 heures de cours

**Cette formation est également disponible en « Intra-entreprise », nous contacter pour plus d'infos.*

| Description

Surfant sur la vague du Big Data, le data scientist joue un rôle clé dans la valorisation de données. Au-delà des paillettes, quel est son rôle, ses outils, sa méthodologie, ses « tips and tricks » ? Venez le découvrir au travers de cette initiation à la Data Science délivrée par des data scientists renommés qui vous apporteront l'expérience des compétitions de Data Science et leurs riches retours d'expérience des modèles réels qu'ils mettent en place chez leurs clients.

| Objectifs pédagogiques

- Découvrir et utiliser la puissance prédictive des modèles ensemblistes
- Savoir effectuer un "feature engineering" performant
- Appréhender les techniques de text-mining et de deep-learning à travers des exemples concrets
- Enrichir sa boite à outils de data scientist

| Publics

Analyste, statisticien, architecte, développeur, data scientist

| Pré-requis

- Connaissances de base en programmation ou en scripting
- Avoir suivi la formation "Fondamentaux de la Data Science" (DSDFX) serait en plus

| Méthode pédagogique

Formation avec apports théoriques, échanges sur les contextes des participants et retours d'expérience pratique du formateur, complétés de travaux pratiques et de mises en situation.

| Programme détaillé

Jour 1

Rappel des fondamentaux

- Ecosystème Big Data et Data Science
- Comment modéliser un problème de data science ?
- Les différentes familles d'algorithmes (supervisé : classification/régression, non supervisé)
- Les algorithmes classiques
- Comment évaluer la performance ?
- Sur apprentissage et compromis biais/variance

Modèles ensemblistes

- Rappels
- Pourquoi ça fonctionne ? Raisons théoriques
- Introduction au stacking
- Architecture et promesses du stacking
- Feature weighted stacking
- Mise en application

Introduction au text mining

- Un modèle de représentation : le bag of words
- Normalisations usuelles
- Stemming, lemmatization
- Distances (Levenshtein, Hamming, Jaro-Winkler)
- Word2Vec

Jour 2

Feature engineering avancé

- Normalisation
- Qu'est-ce que la normalisation ?
- Quand l'utiliser ?
- Réduction de dimension (ACP, TSNE, LSA, etc.)
- Transformation et interactions entre variables
- Traitement des variables catégorielles à haute dimensionnalité
- Création de variables extraites d'arbres (Facebook Trick)

Réseaux de neurones et deep learning

- L'origine : le perceptron
- Les réseaux de neurones
- Deep learning
- Objectif : s'affranchir du feature engineering manuel

- Convolution
- Réseaux récurrents
- Cas concret : reconnaissance de chiffres

Apprentissage semi-supervisé

Jour 3

Rappels et révisions

- Synthèse des points abordés en journées 1 et 2
- Approfondissement des sujets sélectionnés avec l'intervenant

Mise en pratique

- Le dernier jour est entièrement consacré à des mises en pratique

Sélection et participation à une compétition

- Le formateur sélectionnera une compétition en cours sur Kaggle qui sera démarrée en jour 3 par l'ensemble des participants

| Programme détaillé

Jour 3

- **Rappels et révisions**
  - Synthèse des points abordés en journées 1 et 2
  - Approfondissement des sujets sélectionnés avec l'intervenant

- **Mise en pratique**
  - Le dernier jour est entièrement consacré à des mises en pratique

- **Sélection et participation à une compétition**
  - Le formateur sélectionnera une compétition en cours sur Kaggle ou datascience.net qui sera démarrée en jour 3 par l'ensemble des participants