EMGS GROUP
CONSULTING | RECRUITMENT | TRAINING

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

Fondamentaux de la Data Science

Code formation : EMFDS
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 le monde de la Data Science et les grandes familles de problèmes
- Savoir modéliser un problème de Data Science
- Créer ses premières variables
- Constituer sa boite à outils de data scientist

| Publics

Analyste, statisticien, architecte, développeur

| Pré-requis

- Connaissances de base en programmation ou scripting.
- Quelques souvenirs de statistiques sont un 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

Introduction au Big Data

- Qu'est-ce-que le Big Data ?
- L'écosystème technologique du Big Data

Introduction à la Data Science

- Le vocabulaire d'un problème de Data Science
- De l'analyse statistique au machine learning
- Overview des possibilités du machine learning

Modélisation d'un problème

- Input / ouput d'un problème de machine learning
- Mise en pratique « OCR» (Nous verrons comment modéliser le problème de la reconnaissance optique de caractère)

Identifier les familles d'algorithmes de machine learning

- Analyse supervisée
- Analyse non supervisée
- Classification / régression

Sous le capot des algorithmes : la régression linéaire

- Quelques rappels : fonction hypothèse, fonction convexe, optimisation
- La construction de la fonction de coût
- Méthode de minimisation : la descente de gradient

Sous le capot des algorithmes : la régression logistique

- Frontière de décision
- La construction d'une fonction de coût convexe pour la classification

La boite à outil du data scientist

- Introduction aux outils
- Introduction à python, pandas et scikit-learn

Cas pratique n°1 : « Prédire les survivants du Titanic »

- Exposé du problème
- Première manipulation en python

Jour 2

Rappels et révision du jour 1

Qu'est-ce qu'un bon modèle ?

- Cross-validation
- Les métriques d'évaluation : precision, recall, ROC, MAPE, etc.
- Overview des possibilités du machine learning

Les pièges du machine learning

- Overfitting ou sur-apprentissage
- Biais vs variance (Nous verrons comment modéliser le problème de la reconnaissance optique de caractère)
- La régularisation : régression Ridge et Lasso

Data cleaning

- Les types de données : catégorielles, continues, ordonnées, temporelles
- Détection des outliers statistiques, des valeurs aberrantes
- Stratégie pour les valeurs manquantes
- Mise en pratique : « Remplissage des valeurs manquantes»

Feature engineering

- Stratégies pour les variables non continues
- Détecter et créer des variables discriminantes

Cas pratique n°2 : « prédire les survivants du titanic »

- Identification et création des bonnes variables
- Réalisation d'un premier modèle
- Soumission sur Kaggle

Data visualisation

- La visualisation pour comprendre les données : histogramme, scatter plot, etc.
- La visualisation pour comprendre les algorithmes : train / test loss, feature importance, etc.

Introduction aux méthodes ensemblistes

- Le modèle de base : l'arbre de décision, ses avantages et ses limites
- Présentation des différentes stratégies ensemblistes : bagging, boosting, etc.
- Mise en pratique « OCR» (Utilisation d'une méthode ensembliste sur la base du précédent modèle)

Apprentissage semi-supervisé

- Les grandes classes d'algorithmes non supervisées : clustering, PCA, etc.
- Mise en pratique : « Détection d'anomalies dans les prises de paris»

(Nous verrons comment un algorithme non supervisé permet de détecter des fraudes dans les prises de paris)

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