

Data Science

Data Science avec Python (Scikitlearn)

3 jours (21h00) | ★★★★★ 4,6/5 | BIG-PYT | Évaluation qualitative de fin de stage |
Formation délivrée en présentiel ou distanciel ⁽¹⁾

Formations Informatique > Big Data > Data Science



À l'issue de ce stage vous serez capable de :

- Utiliser Scikitlearn pour créer des modèles d'apprentissage machine
- Concevoir des expériences et interpréter les résultats des tests A/B
- Visualiser l'analyse de clustering et de régression en Python à l'aide de Matplotlib
- Produire des recommandations automatisées de produit ou de contenu avec des techniques de filtrage collaboratif
- Appliquer les meilleures pratiques en matière de nettoyage et de préparation de vos données avant l'analyse.

Niveau requis

Une expérience de codage ou de script préalable est nécessaire. Avoir au minimum des compétences en mathématiques de niveau secondaire.

Public concerné

Développeurs de logiciels, programmeurs, Data analysts, Data scientists.

Cette formation :

- Est animée par un consultant-formateur dont les compétences techniques, professionnelles et pédagogiques ont été validées par des diplômes et/ou testées et approuvées par l'éditeur et/ou par M2i Formation
- Bénéficie d'un suivi de son exécution par une feuille de présence émarginée par demi-journée par les stagiaires et le formateur.

(1) Modalité et moyens pédagogique :

Formation délivrée en présentiel ou distanciel * (e-learning, classe virtuelle, présentiel à distance). Le formateur alterne entre méthodes ** démonstrative, interrogative et active (via des travaux pratiques et/ou des mises en situation). La validation des acquis peut se faire via des études de cas, des quiz et/ou une certification.

Les moyens pédagogiques mis en oeuvre (variables suivant les formations) sont : ordinateurs Mac ou PC (sauf pour les cours de l'offre Management), connexion internet fibre, tableau blanc ou paperboard, vidéoprojecteur ou écran tactile interactif (pour le distanciel). Environnements de formation installés sur les postes de travail ou en ligne. Supports de cours et exercices.

* Nous consulter pour la faisabilité en distanciel. ** Ratio variable selon le cours suivi.

Programme

Jour 1

Introduction à la Data Science (la science des données)

- Qu'est-ce que la Data Science ?
 - Définition
 - Différence entre statistiques et Data Science
 - Champs d'application de la Data Science
 - Outils et algorithmes pour faire de la Data Science
- Les méthodes de Machine Learning (apprentissage automatique)
 - L'apprentissage supervisé
 - L'apprentissage non supervisé
 - L'apprentissage renforcé

Le langage de programmation Python - Bases

- Introduction à Python
- Les structures / propriétés / fonctions spéciales
- La Programmation Orientée Objet (POO)
- La bibliothèque standard

Le langage de programmation Python - Avancé

- Notions avancées de Python
- Respecter de bonnes pratiques de développement
- Manipuler les bibliothèques NumPy, SciPy, Pandas et Matplotlib

Préparation de données pour la Data Science

- Pourquoi la préparation des données est-elle si importante en Data Science ?
- Étapes de préparation des données (du nettoyage des données à la compréhension des données)

Jour 2

Exemples de travaux pratiques sur Scikitlearn (à titre indicatif)

- *Maîtriser les techniques d'apprentissage automatique et d'exploration de données avec Scikitlearn de Python*
 - *Extraction de caractéristiques et prétraitement*
 - *Réduction de la dimensionnalité avec l'Analyse des Composants Principaux (PCA)*
 - *Analyse de régression linéaire, multi-linéaire, polynomiale, logistique et régularisation*
 - *Classification supervisée*
 - *Classification non supervisée et régression avec arbres de décision*
 - *K-means clustering*
 - *Perceptron et le Support des Machines Vectorielles (SVM)*
 - *Réseaux de neurones artificiels*

Jour 3

Exemples de travaux pratiques sur Scikitlearn (à titre indicatif) - Suite

- *Maîtriser les techniques d'apprentissage automatique et d'exploration de données avec MLlib d'Apache Spark*

- Analyse en composantes principales
- Analyse de régression
- K-means clustering
- Train / Test et validation croisée
- Méthodes bayésiennes
- Arbres de décision et forêts aléatoires
- Régression multivariée
- Modèles multiniveaux
- Support des Machines Vectorielles (SVM)
- Apprentissage par renforcement
- Filtrage collaboratif
- K plus proche voisin
- Le compromis biais / variance
- Apprentissage d'ensemble
- Fréquence de terme / Fréquence inverse de documents
- Conception expérimentale et A/B testing

Cas d'usage particuliers de Data Science

- Algorithmes de classification / supervisés et non supervisés
 - Classification des itinéraires / chemins (basée sur les matrices origines / destinations temporelles)
 - Détection de mode de transport ou forme de mobilité à partir de jeux de données images
- Algorithmes de recommandation
 - Filtrage collaboratif pour les recommandations : basé sur le contenu, les éléments, l'utilisateur ou filtrage hybride
 - Clustering pour les recommandations
 - Deep Learning pour les recommandations

Modalités d'évaluation des acquis

L'évaluation des acquis se fait :

- En cours de formation, par des études de cas ou des travaux pratiques
- Et, en fin de formation, par un questionnaire d'auto-évaluation ou une certification (M2i ou éditeur)