

Formation **Python** pour la **Data Science**

Utiliser la programmation Python pour l'analyse de données, la visualisation et le Machine Learning

Découvrez la puissance de Python pour la science des données avec notre cours intensif de 3 jours. Cette formation vous permet d'acquérir les compétences nécessaires pour analyser, visualiser et modéliser les données de manière efficace. De l'expérience pratique avec les bibliothèques de base de Python à la construction de modèles d'apprentissage automatique, vous acquerez l'expertise nécessaire pour transformer vos données en informations exploitables et relever les multiples défis de la data science.

Durée

3 jours

Objectifs pédagogiques

- ◆ Comprendre les principes fondamentaux de la programmation Python et son application à la manipulation et à l'analyse des données
- ◆ Appliquer des techniques de nettoyage et de prétraitement des données pour préparer les ensembles de données à l'analyse
- ◆ Créer et interpréter des visualisations de données à l'aide de bibliothèques Python
- ◆ Développer des modèles de Machine Learning basiques pour l'analyse prédictive
- ◆ Évaluer les performances des modèles d'apprentissage automatique

Public

Développeurs, data analysts, data scientists...

Prérequis

Compréhension de base des concepts de programmation.
Familiarité avec les concepts mathématiques fondamentaux (statistiques et algèbre).

Programme de formation

Introduction à Python pour la science des données

Les bases de la programmation Python

Introduction à la syntaxe de Python, aux types de données (chaînes, entiers, flottants), aux structures de contrôle (instructions if, boucles), aux fonctions et aux modules.

Vue d'ensemble des environnements de développement Python (Jupyter Notebooks, PyCharm).

Configuration de l'environnement Python

Installation de Python et des bibliothèques essentielles : Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn.

Environnements virtuels et gestion des paquets avec pip.

Exemples d'activités pratiques :

Mise en place d'un environnement de développement Python.

Écrire des scripts Python de base pour effectuer des manipulations de données simples.

Manipulation de données avec Pandas

Comprendre Pandas pour l'analyse de données

Exploration des objets Series et DataFrame. Indexation, sélection, filtrage et tri des données. Fusionner, joindre et concaténer des données.

Techniques de nettoyage des données

Gestion des valeurs manquantes, des données dupliquées et des types de données incorrects. Traitement des données textuelles et encodage des données catégorielles.

Exemples d'activités pratiques :

Nettoyage d'un ensemble de données réelles à l'aide de Pandas.

Fusion de plusieurs ensembles de données en un seul DataFrame pour l'analyse.

Visualisation des données

Introduction à Matplotlib et Seaborn

Création de graphiques de base : graphiques linéaires, diagrammes à barres et histogrammes.

Visualisations avancées : diagrammes de dispersion, diagrammes en boîte et cartes thermiques.

Analyse visuelle des données

Utiliser les visualisations pour comprendre les distributions et les relations entre les données. Personnaliser les graphiques avec des styles, des couleurs et des annotations.

Exemples d'activités pratiques :

Visualisation des distributions de données et des relations dans un ensemble de données. Conception de visualisations personnalisées pour communiquer des informations sur les données.

Les bases de l'apprentissage automatique (machine learning)

Apprentissage automatique avec Scikit-learn

Vue d'ensemble de l'apprentissage supervisé et non supervisé.

Modèles de régression et de classification : Régression linéaire, régression logistique, k-voisins les plus proches.

Métriques de régression et classification et techniques d'évaluation : R^2 -score, erreur moyenne absolue, précision, rappel, validation croisée...

Feature engineering et sélection des modèles

Prétraitement des données pour l'apprentissage automatique.

Sélection et mise au point des modèles de machine learning.

Exemples d'activités pratiques :

Construction et évaluation d'un modèle de régression linéaire pour prédire des valeurs numériques.

Développement d'un modèle de classification pour catégoriser les points de données.

Machine Learning avancé et meilleures pratiques

Techniques avancées d'apprentissage automatique

Arbres de décision et forêts aléatoires.

Surajustement, sous-ajustement et régularisation du modèle.

Introduction aux réseaux neuronaux et aux concepts d'apprentissage profond.

Gestion de projet en science des données

Meilleures pratiques pour la collaboration dans les projets de science des données.

Considérations éthiques en science des données

Aborder la confidentialité des données, la sécurité et l'utilisation éthique des données.
Comprendre l'impact des données et des modèles biaisés.

Exemple d'activités pratiques :

Mise en œuvre d'un modèle d'arbre de décision pour améliorer les prédictions.

Moyens et méthodes pédagogiques

- ♦ La formation alterne entre présentations des concepts théoriques et mises en application à travers d'ateliers et exercices pratiques (hors formation de type séminaire).
- ♦ Les participants bénéficient des retours d'expérience terrains du formateur ou de la formatrice
- ♦ Un support de cours numérique est fourni aux stagiaires

Modalités d'évaluation

- ♦ **En amont de la session de formation**, un questionnaire d'auto-positionnement est remis aux participants, afin qu'ils situent leurs connaissances et compétences déjà acquises par rapport au thème de la formation.
- ♦ **En cours de formation**, l'évaluation se fait sous forme d'ateliers, exercices et travaux pratiques de validation, de retour d'observation et/ou de partage d'expérience, en cohérence avec les objectifs pédagogiques visés.
- ♦ **En fin de session**, le formateur évalue les compétences et connaissances acquises par les apprenants grâce à un questionnaire reprenant les mêmes éléments que l'auto-positionnement, permettant ainsi une analyse détaillée de leur progression.