

PYTHON

Data Scientist avec Python

Plan de cours N° : 1040

Durée : 5 jours (35h)

PARTICIPANTS / PRE-REQUIS

Développeur, chef de projets proche du développement, ingénieur scientifique sachant coder.

Maîtriser l'algorithmique, avoir une appétence pour les mathématiques. La connaissance de Python et des statistiques est un plus.

OBJECTIFS PEDAGOGIQUES

Appréhender les concepts fondamentaux de la data science, du machine learning et des statistiques. Utiliser Python et ses bibliothèques pour manipuler des données, créer des modèles et visualiser des résultats. Créer et gérer des DataLake, DataMart et DataWarehouse. Utiliser Matplotlib pour afficher des données sous forme de graphiques 2D. Mettre en oeuvre des modèles de régression, de classification et de clustering. Utiliser MongoDB, Cassandra et autres bases de données NoSQL pour stocker et manipuler des données. Utiliser Numpy, SciPy et ScikitLearn pour améliorer la précision et l'efficacité des modèles. Comprendre et utiliser Hadoop et ses composants pour traiter de grandes quantités de données.

MOYENS PEDAGOGIQUES

Tour de table au début de chaque formation pour définir les objectifs de chaque participant,

Alternance entre apports théoriques (en moyenne 30%) et exercices pratiques (en moyenne 70%),

Utilisation de cas concrets issus de l'expérience professionnelle de nos formateurs,

Remise d'un support de cours,

Assistance post-formation d'une durée de 1 an sur le contenu de la formation via notre adresse mail dédiée formateurs@atp-formation.com

MOYENS PERMETTANT LE SUIVI DE L'EXECUTION ET DES RESULTATS

Positionnement préalable oral ou écrit,

Evaluation des acquis tout au long de la formation par des exercices de synthèse,

Attestation de stage remise à chaque apprenant, avec son niveau d'acquisition pour chaque objectif pédagogique,

Feuille de présence signée par demi-journée,

Questionnaire de satisfaction pour évaluer la qualité de l'enseignement,

En option: passage certification possible selon les thématiques.

MOYENS TECHNIQUES EN PRESENTIEL

Accueil des stagiaires dans une salle dédiée à la formation, équipée d'ordinateurs récents et performants, d'un vidéoprojecteur et d'un tableau blanc.

MOYENS TECHNIQUES DES CLASSES A DISTANCE

Grâce à un logiciel comme Teams, suivez une formation en temps réel et entièrement à distance. Lors de la classe en ligne, les apprenants interagissent et communiquent entre eux et avec le formateur.

Nous vous conseillons très fortement l'utilisation de votre webcam et de disposer d'un double écran.

Pour toute question avant et pendant le parcours, une assistance technique et pédagogique est à disposition par téléphone au 04.76.41.14.20.

ORGANISATION

Les cours ont lieu de 9h00-12h30 13h30-17h00 (adaptable à la demande).

PROFIL FORMATEUR

Nous recrutons méticuleusement nos formateurs selon 3 critères: expertise, pédagogie et agilité.

ACCESSIBILITE

Les personnes atteintes de handicap souhaitant suivre nos formations sont invitées à nous contacter directement, afin d'étudier ensemble les possibilités d'organisation.

MISE A JOUR

15/07/2024

Plan de cours N° : 1040

Durée : 5 jours (35h)

PYTHON

Data Scientist avec Python

Introduction aux Data Sciences

- Qu'est que la data science ?
- Qu'est-ce que Python ?
- Qu'est que le Machine Learning ?
- Apprentissage supervisé vs non supervisé
- Les statistiques
- La randomisation
- La loi normale

Introduction à Python pour les Data Science

- Les bases de Python
- Les listes en intention
- Les listes
- Les tuples
- Les dictionnaires
- Les modules et packages
- L'orienté objet
- Le module math
- Les expressions lambda
- Map, reduce et filter
- Le module CSV
- Les modules DB-API 2
- Anaconda

Introduction aux DataLake, DataMart et DataWarehouse

- Qu'est-ce qu'un DataLake ?
- Les différents types de DataLake
- Le Big Data
- Qu'est-ce qu'un DataWarehouse ?
- Qu'est qu'un DataMart ?
- Mise en place d'un DataMart
- Les fichiers
- Les bases de données SQL
- Les bases de données No-SQL

Python Package Installer

- Utilisation de PIP
- Installation de package PIP
- PyPi

Matplotlib

- Utilisation de la bibliothèque scientifique de graphes Matplotlib
- Affichage de données dans un graphique 2D
- Affichages de sous-graphes
- Affichage de polynômes et de sinusoïdales

Machine Learning

- Mise en place d'une machine learning supervisé
- Filtrage
- Qu'est qu'un modèle et un dataset
- Qu'est qu'une régression
- Les différents types de régression
- La régression linéaire
- Gestion du risque et des erreurs
- Quarter d'Ascombe
- Trouver le bon modèle
- La classification
- Loi normale, variance et écart type
- Apprentissage
- Mesure de la performance
- No Fee Lunch

La régression linéaire en Python

- Programmer une régression linéaire en Python
- Utilisation des expressions lambda et des listes en intention
- Afficher la régression avec Matplotlib
- L'erreur quadratique
- La variance
- Le risque

Siège social :

31 avenue du Granier
38240 MEYLAN

Agences :

170 rue de Chatagnon
38430 Moirans

Le Thélème

1501/1503 route des Dolines
06560 Valbonne

PYTHON

Data Scientist avec Python

Plan de cours N° : 1040

Durée : 5 jours (35h)

Le Big Data

- Qu'est-ce que Apache Hadoop ?
- Yarn
- Qu'est-ce que l'informatique distribué ?
- Installation et configuration de Hadoop
- HDFS
- Création d'un datanode
- Création d'un namenode distribué
- Manipulation de HDFS
- Hadoop comme DataLake
- Map Reduce
- Hive
- Hadoop comme DataMart
- Python HDFS

Les bases de données NoSql

- Les bases de données structurées
- SQL avec SQLite et Postgresql
- Les bases de données non ACID
- JSON
- MongoDB
- Cassandra, Redis, CouchDb
- MongoDB sur HDFS
- MongoDB comme DataMart
- PyMongo

Numpy et SciPy

- Les tableaux et les matrices
- L'algèbre linéaire avec Numpy
- La régression linéaire SciPy
- Le produit et la transposée
- L'inversion de matrice
- Les nombres complexes
- L'algèbre complexe
- Les transformées de Fourier
- Numpy et Matplotlib

ScikitLearn

- Régressions polynomiales
- Le machine Learning avec SKLearn
- La régression linéaire
- La création du modèle
- L'échantillonnage
- La randomisation
- L'apprentissage avec fit
- La prédition du modèle
- Les metrics
- Choix du modèle
- PreProcessing et Pipeline
- Régressions non polynomiales

Nearest Neighbors

- Algorithme des k plus proches voisins (k-NN)
- Modèle de classification
- K-NN avec SciKitLearn
- Choix du meilleur k
- Sérialisation du modèle
- Variance vs Erreurs
- Autres modèles : SVN, Random Forest

Pandas

- L'analyse des données avec Pandas
- Les Series
- Les DataFrames
- La théorie ensembliste avec Pandas
- L'importation des données CSV
- L'importation de données SQL
- L'importation de données MongoDB
- Pandas et SKLearn

Le Clustering

- Regroupement des données par clusterisation
- Les clusters SKLearn avec k-means
- Autres modèles de clusterisation : AffinityPropagation, MeanShift, ...
- L'apprentissage semi-supervisé

Siège social :

31 avenue du Granier
38240 MEYLAN

Agences :

170 rue de Chatagnon
38430 Moirans

Le Thélème

1501/1503 route des Dolines
06560 Valbonne

PYTHON

Data Scientist avec Python

Plan de cours N° : 1040

Durée : 5 jours (35h)

Jupyter

Présentation de Jupyter et Ipython

Installation

Utilisation de Jupyter avec Mathplotlib et Sklearn

Python Yield

La programmation efficace en Python

Le générateurs et itérateurs

Le Yield return

Le Yield avec Db-API 2, Pandas et Sklearn

Les réseaux neuronaux

Le perceptron

Les réseaux neuronaux

Les réseaux neuronaux supervisés

Les réseaux neuronaux semi-supervisés

Les réseaux neuronaux par Hadoop Yarn

Les heuristiques

Le deep learning