

# Data Scientist



## Informatique

Référence formation : 4-IT-DAS - Durée : 5 jours

### Objectifs

Savoir mettre en place un DataLake et un DataMart en SQL ou big data, puis une stratégie de Machine Learning en Python afin de créer le modèle le plus satisfaisant possible en le mesurant et en affichant les résultats, le tout en utilisant des algorithmes performants

### Pré-requis

Développeurs, chefs de projets proches 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

## Contenu pédagogique

### 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
- Les tuples
- Les dictionnaires
- Les modules et packages
- L'orienté objet
- Le module math
- Les expressions lambda
- Map, reduce et filter
- Le module CSV



### XXL Formation

34 rue Raymond Aron  
76130 Mont Saint Aignan

Tél : 02 35 12 25 55 – Fax : 02 35 12 25 56

N° siret : 485 050 611 00014 – N° d'agrément : 23.76.03752.76



- 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

## Mathplotlib

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

## Machine Learning

- Mise en place d'une machine learning supervisé
- 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 Mathplotlib



### XXL Formation

34 rue Raymond Aron  
76130 Mont Saint Aignan

Tél : 02 35 12 25 55 – Fax : 02 35 12 25 56

N° siret : 485 050 611 00014 – N° d'agrément : 23.76.03752.76



- L'erreur quadratique
- La variance
- Le risque

## Le Big Data

- Qu'est-ce que Apache Hadoop ?
- Qu'est-ce que l'informatique distribuée ?
- 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
- La régression linéaire



### XXL Formation

34 rue Raymond Aron  
76130 Mont Saint Aignan

Tél : 02 35 12 25 55 – Fax : 02 35 12 25 56

N° siret : 485 050 611 00014 – N° d'agrément : 23.76.03752.76



- La création du modèle
- L'échantillonnage
- La randomisation
- L'apprentissage avec fit
- La prédiction 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

## Jupyter

- Présentation de Jupyter et lpython
- Installation
- Utilisation de Jupyter avec Mathplotlib et Sklearn

## Python Yield



### XXL Formation

34 rue Raymond Aron  
76130 Mont Saint Aignan

Tél : 02 35 12 25 55 – Fax : 02 35 12 25 56

N° siret : 485 050 611 00014 – N° d'agrément : 23.76.03752.76



- La programmation efficace en Python
- Les 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

<b>Méthodes pédagogiques</b>	Présentation des concepts, démonstration, exécution, synthèse et exercices pratiques d'assimilation
<b>Modalités pédagogiques</b>	Présentiel - Distanciel - AFEST
<b>Moyens pédagogiques</b>	Formateur expert du domaine - 1 ordinateur, 1 support de cours version papier ou numérique, un bloc-note et un stylo par personne - vidéo projecteur - tableau blanc
<b>Modalités d'évaluation</b>	Positionnement préalable oral ou écrit - Evaluation formative tout au long de la formation - Evaluation sommative faite par le formateur ou à l'aide de la certification <b>NULL</b> : <a href="https://www.francecompetences.fr/recherche/NULL">https://www.francecompetences.fr/recherche/NULL</a>
<b>Durée journée de formation</b>	7h00
<b>Public concerné</b>	Salariés - Demandeur d'emploi - Reconversion professionnelle - Public en situation de handicap



### XXL Formation

34 rue Raymond Aron  
76130 Mont Saint Aignan

Tél : 02 35 12 25 55 – Fax : 02 35 12 25 56

N° siret : 485 050 611 00014 – N° d'agrément : 23.76.03752.76

