10 juin-14 juin
09 déc.-13 déc.
18 nov.-22 nov.
16 sept.-20 sept.

Nantes / Rennes : 2890 € HT
Brest / Le Mans : 2890 € HT
Certification : NON

Data Scientist

Mise à jour nov. 2023

Durée 5 jours (35 heures)

« Délai d'accès maximum 1 mois »

Décisionnel

OBJECTIFS PROFESSIONNELS

- 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

PARTICIPANTS

-

PRE-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

MOYENS PEDAGOGIQUES

- Réflexion de groupe et apports théoriques du formateur
- Travail d'échange avec les participants sous forme de
- Utilisation de cas concrets issus de l'expérience professionnelle
- Validation des acquis par des questionnaires, des tests d'évaluation, des mises en situation et des jeux pédagogiques.
- Remise d'un support de cours.

MODALITES D'EVALUATION

- Feuille de présence signée en demi-journée,
- Evaluation des acquis tout au long de la formation,
- Questionnaire de satisfaction,
- Positionnement préalable oral ou écrit,
- Evaluation formative tout au long de la formation,
- Evaluation sommative faite par le formateur ou à l'aide des certifications disponibles,
- Sanction finale : Certificat de réalisation, certification éligible au RS selon l'obtention du résultat par le stagiaire

MOYENS TECHNIQUES EN PRESENTIEL

- Accueil des stagiaires dans une salle dédiée à la formation, équipée d'ordinateurs, d'un vidéo projecteur d'un tableau blanc et de paperboard. Nous préconisons 8 personnes maximum par action de formation en présentiel

MOYENS TECHNIQUES DES CLASSES EN CAS DE FORMATION DISTANCIELLE

- A l'aide d'un logiciel comme Teams, Zoom etc... un micro et éventuellement une caméra pour l'apprenant,
- suivez une formation uniquement synchrone en temps réel et entièrement à distance. Lors de la classe en ligne, les apprenants interagissent et comuniquent entre eux et avec le formateur.
- Les formations en distanciel sont organisées en Inter-Entreprise comme en Intra-Entreprise.
- L'accès à l'environnement d'apprentissage (support de cours, labs) ainsi qu'aux preuves de suivi et d'assiduité (émargement, évaluation) est assuré. Nous préconisons 4 personnes maximum par action de formation en classe à distance

ORGANISATION

- Les cours ont lieu de 9h à 12h30 et de 14h à 17h30.

PROFIL FORMATEUR

- Nos formateurs sont des experts dans leurs domaines d'intervention
- Leur expérience de terrain et leurs qualités pédagogiques constituent un gage de qualité.

Décisionnel

A L'ATTENTION DES PERSONNES EN SITUATION DE HANDICAP

- Les personnes atteintes de handicap souhaitant suivre cette formation sont invitées à nous contacter directement, afin d'étudier ensemble les possibilités de suivre la formation.

Programme de formation

Introduction aux Data Sciences (02h15)

- 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 (03h00)

- 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
- Les modules DB-API 2 Anaconda

Introduction aux DataLake, DataMart et DataWharehouse (02h30)

- Qu'est-ce qu'un DataLake ?
- Les différents types de DataLake
- Le Big Data
- Qu'est-ce qu'un DataWharehouse ?
- 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 (00h30)

- Utilisation de PIP
- Installation de package PIP PyPi

MathPlotLib (01h15)

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

Machine Learning (03h30)

- 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 (01h45)

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

Le Big Data (03h30)

- Qu'est-ce que Apache Hadoop ?
- 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 (02h15)

- 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 (02h15)

- 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 Mathplotlib

ScikitLearn (03h00)

- Régressions polynomiales
- La régression linéaire
- 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 (02h00)

- 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 (02h00)

- 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 (01h15)

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

Jupyter (00h45)

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

Python Yield (01h15)

- La programmation efficace en Python
- Le générateurs et itérateurs
- Le Yield return
- Le Yield avec Db-API 2, Pandas et Sklearn

Les réseaux neuronaux (02h00)

- 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