

# Data classification et Machine Learning

Référence : PYCB050

Durée : 2 jours

Certification : **Aucune**

## CONNAISSANCES PREALABLES

- Connaissances des principes du BigData, et des architectures techniques mises en oeuvre.

## PROFIL DES STAGIAIRES

- Architectes. • Chefs de projet. • Data scientists. • Développeurs.

## OBJECTIFS

- Comprendre et mettre en oeuvre l'apprentissage automatique, les techniques de classification de données, les apports des réseaux de neurones et du Deep Learning. • Savoir définir les étapes de préparation des données.

## CERTIFICATION PREPAREE

Aucune

## METHODES PEDAGOGIQUES

- Mise à disposition d'un poste de travail par stagiaire
- Remise d'une documentation pédagogique papier ou numérique pendant le stage
- La formation est constituée d'apports théoriques, d'exercices pratiques, de réflexions et de retours d'expérience
- Le suivi de cette formation donne lieu à la signature d'une feuille d'émargement

## FORMATEUR

Consultant-Formateur expert Bigdata

## METHODE D'EVALUATION DES ACQUIS

- Auto-évaluation des acquis par le stagiaire via un questionnaire
- Attestation de fin de stage adressée avec la facture

## CONTENU DU COURS

### Introduction

- Zoom sur les données : format, volumes, structures,... et les requêtes, attentes des utilisateurs
- Etapes de la préparation des données
- Définitions, présentation du data munging
- Le rôle du data scientist

### Gouvernance des données

- Qualité des données
- Transformation de l'information en donnée
- Qualification et enrichissement
- Sécurisation et étanchéité des lacs de données
- Flux de données et organisation dans l'entreprise
- De la donnée maître à la donnée de travail
- MDM
- Mise en oeuvre pratique des différentes phases : nettoyage, enrichissement, organisation des données

### Traitements statistiques de base

- Introduction aux calculs statistiques
- Paramétrisation des fonctions
- Applications aux fermes de calculs distribués
- Problématiques induites
- Approximations
- Précision des estimations

### Data Mining

- Besoin, apports et enjeu
- Extraction et organisation des classes de données
- Analyse factorielle

### Machine Learning

- Apprentissage automatique
- Définition, les attentes par rapport au Machine Learning
- Les valeurs d'observation, et les variables cibles
- Ingénierie des variables
- Les méthodes : apprentissage supervisé et non supervisé

- Classification des données
- Algorithmes : régression linéaire, k-moyennes, k-voisins, classification naïve bayésienne, arbres de décision, forêts aléatoires, etc
- Création de jeux d'essai, entraînement et construction de modèles
- Prévisions à partir de données réelles
- Mesure de l'efficacité des algorithmes
- Courbes ROC
- Parallélisation des algorithmes
- Choix automatique

## IA

- Introduction aux réseaux de neurones
- Réseaux de neurones à convolution

- Modèles de CNN
- Les types de couches : convolution, pooling et pertes
- L'approche du Deep Learning
- Deeplearning4j sur Spark

## Les risques et écueils

- Importance de la préparation des données
- L'écueil du "surapprentissage"

## Visualisation des données

- L'intérêt de la visualisation
- Outils disponibles
- Exemples de visualisation avec R et Python