# Programme de cours
# Machine Learning avancé avec Python

| Durée : | 2 jours soit 14 heures | Référence : | IF-PyML2 |
|---|---|---|---|

## Public visé :

Data scientist, data engineer, data analyst, chercheur, ingénieur R&D, statisticien, et toute personne travaillant dans la data et ayant une première expérience des modèles prédictifs

## Pré-requis :

- Avoir suivi une formation Machine Learning niveau 1, ou un MOOC en ligne (par exemple le MOOC d'Andrew Ng sur Coursera), ou avoir une première expérience dans la création de modèles prédictif et leur évaluation
- Expérience de Python et de scikit-learn

## Objectifs pédagogiques :

- Savoir choisir les bons algorithmes d'apprentissage en fonction du problème à résoudre (ensembles d'arbres de décision, modélisation linéaire / non linéaire, régularisation)
- Créer des modèles prédictifs qui peuvent se mettre à jour en continu, et ingérer de gros volumes de données (**Online Learning**)
- Trouver les meilleurs compromis entre temps de calcul et qualité des prédictions
- Comprendre et mettre en pratique la technique du **Boosting**, utilisée dans les meilleures solutions aux compétitions de Machine Learning
- Optimiser ses modèles prédictifs, grâce aux **techniques intelligentes d'optimisation** du choix d'hyperparamètres (“AutoML”) et aux architectures complexes d'ensembles de modèles (**Stacking**)

## Compétences acquises à l’issue de la formation :

- Savoir choisir les bons algorithmes d’apprentissage en fonction du problème à résoudre (ensembles d’arbres de décision, modélisation linéaire / non linéaire, régularisation)
- Créer des modèles prédictifs qui peuvent se mettre à jour en continu, et ingérer de gros volumes de données (Online Learning)
- Trouver les meilleurs compromis entre temps de calcul et qualité des prédictions
- Comprendre et mettre en pratique la technique du Boosting, utilisée dans les meilleures solutions aux compétitions de Machine Learning
- Optimiser ses modèles prédictifs, grâce aux techniques intelligentes d’optimisation du choix d’hyperparamètres (“AutoML”) et aux architectures complexes d’ensembles de modèles (Stacking)

## Modalités pédagogiques :

Session dispensée en présentiel ou téléprésentiel, selon la modalité inter-entreprises ou intra-entreprises sur mesure.
La formation est animée par un(e) formateur(trice) durant toute la durée de la session et présentant une suite de modules théoriques clôturés par des ateliers pratiques validant l'acquisition des connaissances. Les ateliers peuvent être accompagnés de Quizz.
L'animateur(trice) présente la partie théorique à l'aide de support de présentation, d'animation réalisée sur un environnement de démonstration.
En présentiel comme en téléprésentiel, l'animateur(trice) accompagne les participants durant la réalisation des ateliers.

## Moyens et supports pédagogiques :

**Cadre présentiel**
Salles de formation équipées et accessibles aux personnes à mobilité réduite.
- Un poste de travail par participant
- Un support de cours numérique ou papier (au choix)
- Un bloc-notes + stylo
- Vidéoprojection sur tableau blanc
- Connexion Internet
- Accès extranet pour partage de documents et émargement électronique

**Cadre téléprésentiel**
Session dispensée via notre solution iClassroom s'appuyant sur Microsoft Teams.
- Un compte Office 365 par participant
- Un poste virtuel par participant
- Un support numérique (PDF ou Web)
- Accès extranet pour partage de documents et émargement électronique

## Informations sur l'accessibilité :

# Programme de cours
# Machine Learning avancé avec Python

## Description / Contenu

**Synthèse :**
Approfondissez votre connaissance du Machine Learning pour rendre vos modèles plus performants et maîtrisez les meilleurs algorithmes actuels.

**Description :**
Apprenez à utiliser les techniques actuelles de modélisation prédictive les plus performantes, employées par les meilleurs compétiteurs dans les challenges de Machine Learning. Au travers de cette formation, vous mettrez en pratique la théorie sur divers types de données structurées — y compris sur de très gros volumes (plusieurs Go) — au travers de challenges Kaggle, en utilisant les librairies Python pandas, scikit-learn, XGBoost et Hyperopt. À la fin des 2 jours, vous disposerez de connaissances avancées et pratiques vous permettant de sélectionner les meilleurs algorithmes pour vos problèmes de ML, d'optimiser vos modèles de façon intelligente, et de les mettre à jour en continu.
La formation est principalement destinée aux développeurs et ingénieurs informaticiens expérimentés en Machine Learning. Elle sera également d'intérêt aux statisticiens et data scientists souhaitant approfondir et mettre en pratique leurs connaissances de Machine Learning avec les outils Python.
Si vous êtes débutant, consultez notre formation Machine Learning niveau 1.
Si votre priorité est de développer des applications perceptives — qui "comprennent" l'image ou le son, par exemple — ou si vous êtes déjà à l'aise avec les sujets abordés ici, notre formation Deep Learning est faite pour vous.

**JOUR 1 :**
**Module 1 : Rappels et/ou explications des principaux algorithmes de Machine Learning : ce contenu sera adapté en séance en fonction des connaissances des participants et de leurs attentes :**

- Rappels théoriques et description des principaux hyper-paramètres de :
  - Régression linéaire, polynomiale et logistique
  - K-plus proches voisins (KNN)
  - Machines à vecteur de support (SVM)
  - Arbres de décision, forêts aléatoires
  - Réseaux de neurones
- Avantages et inconvénients : comment et pourquoi sélectionner un type de modèl
- Méthodologie projet : workflow complet et best-practices
- Ecueils à éviter (et comment les éviter) : fuite de données (data leak), surapprentissage (overfitting)
- Mise en pratique avec scikit-learn

**Module 2 : Boosting**

- Principe du boosting, classe d'algorithmes souvent plus performants que les random forests (XGBoost, CatBoost, LightGBM…)
- Description de leurs principaux paramètres, délicats à prendre en main et importants à maîtriser
- Principe des techniques avancées d'**optimisation intelligente des hyper-paramètre**
- Mise en pratique avec Hyperopt

**JOUR 2 :**
**Module 3 : Apprentissage sur gros volumes de données et Online Learning**

- Présentation de l'algorithme de **descente de gradient (stochastique, mini-batch)** ; intuition de ses principaux paramètres
- Mise à jour de modèles en **flux continu et apprentissage hors-mémoire**
- Mise en pratique avec scikit-learn sur un jeu de données qui ne tient pas en mémoire vive
- Considérations supplémentaires : hashing trick, dilemme exploration/exploitation

**Module 4 : Stacking**

- Principe du **stacking** et état de l'art des architectures d'ensembles de modèles prédictifs
- Explication des systèmes les plus performants sur les concours Kaggle sur des données structurées
- Exercice final mettant en pratique une architecture de stacking utilisant les types de modèles vus dans la formation