CODE
7109

TECHNOLOGIES DU NUMÉRIQUE ET SYSTÈME D'INFORMATION / IA

NOUVEAU POSSIBLE EN INTRA

INTRODUCTION AU MACHINE LEARNING AVEC SCIKIT-LEARN

Le Machine Learning devient un levier majeur de valorisation des données. Aussi les équipes techniques doivent maîtriser les étapes clés, de la préparation des données à l'évaluation et à la mise en production des modèles prédictifs avec scikit-learn, la bibliothèque de référence de l'écosystème Python.

1 780 € HT | **2 JOURS** (14 H.) | **NOUS CONSULTER** POUR LES DATES DE SESSION

COMPÉTENCE PRINCIPALE VISÉE

Construire un modèle de Machine Learning adapté à un problème métier à partir de données préparées et structurées.

LES + DE LA FORMATION

60 % du temps est consacré à la pratique, avec des applications immédiates sous Python et un mini-projet complet permettant de construire, évaluer et livrer un modèle.

PUBLIC

- Débutants en Machine Learning,
- Data scientists

PRÉREQUIS

- Maîtriser les bases de Python
- Disposer de connaissances élémentaires en statistiques.

OBJECTIFS PÉDAGOGIQUES

- Préparer et explorer les données en construisant des pipelines reproductibles et en utilisant des méthodes d'analyse non supervisée.
- Construire et optimiser des modèles de régression et de classification en sélectionnant les algorithmes, métriques et hyperparamètres adaptés au problème métier.
- Évaluer, interpréter et livrer un modèle de Machine Learning en vérifiant sa performance, sa capacité de généralisation et son exploitabilité en production.

97,2%
de clients satisfaits*
* enquête réalisée auprès de nos clients en septembre 2025

CONTENU

JOUR 1 - FONDAMENTEAUX DU MACHINE LEARNING, PRISE EN MAIN DE SCIKIT-LEARN, PRÉPARATION ET EXPLORATION DES DONNÉES

PARTIE 1 - Fondamentaux du Machine Learning & Prise en Main

- Apprentissage supervisé et non supervisé,
  - positionnement
  - cas d'usage respectifs.
- Concepts fondamentaux :
  - features,
  - cible,
  - généralisation,
  - surapprentissage
  - sous-apprentissage.
- API scikit-learn : fit / predict / transform.
- Séparation apprentissage / test :
  - rôle du jeu de test,
  - protocole de découpage
  - fuite de données (data leakage).
- Premiers pas :
  - chargement d'un jeu de données,
  - exploration avec pandas
  - construction d'un premier modèle en quelques lignes de code.

TRAVAUX PRATIQUES : Première prise en main de scikit-learn

- Chargement et exploration d'un jeu de données avec pandas.
- Séparation apprentissage / test et entraînement d'un premier modèle.
- Production des premières prédictions et analyse d'un premier score.

PARTIE 2 - Préparation des Données & Pipelines

- Traitement des variables :
  - encodage des variables catégorielles,
  - normalisation des variables numériques
  - imputation des valeurs manquantes.
- Structuration des traitements :
  - Pipeline et ColumnTransformer pour enchaîner prétraitements et modèle.
- Fuite de données lors du prétraitement ajusté sur l'ensemble des données avant le découpage.

TRAVAUX PRATIQUES - Pipeline de préparation des données

- Construction d'un pipeline complet de prétraitement sur un jeu de données mixte intégrant variables numériques et catégorielles.
- Mise en évidence d'une fuite de données par comparaison des scores obtenus avec et sans pipeline.
- Intégration du modèle dans le pipeline et validation du fonctionnement de bout en bout.

PARTIE 3 - Exploration, Réduction de Dimension & Clustering

- Analyse en composantes principales (PCA) :
  - principe,
  - variance expliquée
  - application à la visualisation d'un jeu de données à plusieurs dimensions et comme prétraitement intégrable dans un pipeline.
- Clustering avec k-means comme méthode d'analyse de données,
  - choix du nombre de groupes
  - limites de la méthode.

- Interprétabilité des composantes et des groupes obtenus en contexte métier,
    - risques de surinterprétation.

DÉMONSTRATION GUIDÉE – Exploration d'un jeu de données

- Exécution commentée d'un notebook fourni :
    - réduction à deux composantes principales
    - visualisation.
- Application de k-means et caractérisation des groupes obtenus.
- Discussion sur la pertinence métier des clusters

JOUR 2 - MODÈLES SUPERVISÉS, MESURE DE LA PERFORMANCE, OPTIMISATION ET LIVRAISON D'UN MODÈLE

PARTIE 4 - Modèles Supervisés & Mesure de la Performance

- Modèles de régression :
    - modèles linéaires
    - modèles régularisés
- Modèles de classification — principes de fonctionnement, avantages, limites et cas d'usage.
    - k-NN,
    - arbres de décision,
    - forêts aléatoires,
    - gradient boosting
- Référence de comparaison :
    - utilisation d'un modèle trivial comme point de comparaison de toute performance annoncée.
- Métriques de régression :
    - MAE,
    - RMSE
    - coefficient de détermination.
- Métriques de classification :
    - accuracy,
    - précision,
    - rappel,
    - F1-score,
    - ROC-AUC
    - matrice de confusion.
- Données déséquilibrées :
    - pourquoi une accuracy élevée peut être trompeuse,
    - choix de la métrique en cas de déséquilibrage des classes,
    - découpage stratifié
    - ajustement du seuil de décision.
- Choix d'une approche :
    - critères de sélection d'un algorithme et d'une métrique adaptés à un problème métier.

TRAVAUX PRATIQUES - Comparaison de modèles de classification sur un jeu de données déséquilibré

- Établissement d'une baseline triviale, puis entraînement de plusieurs familles de modèles.
- Comparaison des résultats selon l'accuracy puis selon la précision, le rappel et la matrice de confusion : mise en évidence de l'illusion de performance.
- Ajustement du seuil de décision
    - sélection d'un modèle au regard des erreurs.

PARTIE 5 - Audit Fournisseurs & Comité de Validation (Go/No-Go)

- Validation croisée :

    - principes,
    - mise en œuvre
    - découpage stratifié pour une estimation fiable de la performance.
- Optimisation des hyperparamètres :
    - GridSearchCV appliqué au pipeline complet,
    - définition de la grille
    - choix de la métrique d'optimisation.
- Interprétation d'un modèle :
    - importances des variables d'une forêt aléatoire (feature_importances_),
    - sensibilité d'un modèle.

TRAVAUX PRATIQUES - Réglage et interprétation d'un modèle

- Optimisation par GridSearchCV sur un pipeline complet, prétraitements inclus.
- Comparaison de la performance en validation croisée et sur le jeu de test.
- Explicabilité d'un modèle de forêt aléatoire.

PARTIE 6 - Livraison d'un Modèle

- Sérialisation du pipeline :
    - enregistrement avec joblib de l'objet complet,
    - prétraitements inclus.
- Chargement d'un pipeline et prédiction :
    - utilisation du pipeline sérialisé sur de nouvelles données.
- Bonnes pratiques :
    - ce qu'il faut conserver et versionner aux côtés du modèle (données d'entraînement, versions des bibliothèques, métriques de référence)
    - pourquoi ces éléments conditionnent le suivi du modèle en production.

TRAVAUX PRATIQUES de synthèse – Mini-projet de bout en bout

- Exploration puis préparation des données au sein d'un pipeline unique.
- Modélisation, choix d'une métrique adaptée au problème et optimisation des hyperparamètres.
- Interprétation des résultats, sérialisation du pipeline retenu.

ÉQUIPE PÉDAGOGIQUE

L'équipe enseignante est composée de plusieurs enseignants-chercheurs de l'INSA Lyon, membres du laboratoire LIRIS, avec une activité de recherche reconnue en sciences des données. Cette équipe dispose d'une forte expertise technico-scientifique, obtenue via des nombreux projets de recherche partenariale, avec des grands groupes français mais également avec des start-up, PME ou ETI.

MOYENS ET MÉTHODES PÉDAGOGIQUES

Alternance d'échanges techniques, d'illustrations, de démonstrations commentées et de travaux pratiques Un support de cours sera remis à chacun des participants

PROCHAINE SESSION

L'ouverture de la session est conditionnée par un nombre minimum de participants. Nous consulter pour d'autres dates.

ÉVALUATION ET RÉSULTATS

Évaluation des acquis de la formation

Evaluation des acquis des apprenants par auto-examen. 94.3% des apprenants ont acquis la compétence principale visée. (sur 177 apprenants évalués sur cette thématique depuis 2020)

Évaluation de la satisfaction des participants en fin de formation (Niveau 1 KIRKPATRICK)

4.3 par les participants. (sur 199 participants ayant suivi une formation dans la thématique depuis 2020)

RENSEIGNEMENTS ET INSCRIPTION
Tel : +33 (0)4 72 43 83 93
Fax : +33 (0)4 72 44 34 24
mail : formation@insavalor.fr
Préinscription sur formation.insavalor.fr

Accueil des personnes en situation de handicap nécessitant un besoin spécifique d'accompagnement : nous contacter à l'inscription. Nos locaux sont accessibles aux personnes à mobilité réduite.

Actualisée le 29/09/2026