Modélisation avancée des données : de l'analyse au projet IA

Date et durée

Code formation : DATA006FR
Durée : 3 jours
Nombre d'heures : 21 heures

Description

À l'heure où l'IA génère plus de données qu'elle n'en analyse, savoir **transformer un fichier brut en un modèle fiable** est devenu un avantage compétitif crucial. Cette formation intensive de trois jours combine la modélisation supervisée et non supervisée avec l'art du "prompting" pour accélérer le nettoyage des données, le **feature engineering** et l'interprétation des résultats.
En binômes, vous traverserez l'intégralité du cycle de vie d'un projet de modélisation : **l'import, l'audit qualité, l'entraînement, la validation, le packaging et le déploiement**. Chaque étape sera illustrée par un cas réel (prédiction de départ de salarié, scoring de fraude, segmentation client) et un template Python/Excel que vous pourrez réutiliser dès votre retour en entreprise.
À l'issue de la formation, vous repartirez avec un projet abouti, une **check-list éthique conforme au RGPD et à l'AI Act**, ainsi qu'une feuille de route pour industrialiser votre modèle. Ces éléments vous garantissent une adoption rapide et responsable de l'IA dans votre entreprise.

Objectifs

À l'issue de cette formation en modélisation de données, vous atteindrez les objectifs de compétences suivants :

- identifier les enjeux métiers et les opportunités d'une démarche de modélisation de données ;
- manipuler les plates-formes d'analyse et les bibliothèques de modélisation les plus courantes ;
- choisir, entraîner et évaluer des algorithmes supervisés et non supervisés adaptés à vos données ;
- accélérer le nettoyage, l'enrichissement et la documentation des données grâce à des outils d'IA générative ;
- respecter les principes éthiques et réglementaires tout au long du cycle de vie des données ;
- planifier, exécuter et présenter un mini-projet de modélisation complet, du cadrage à la restitution des résultats.

Points forts

- **Une approche opérationnelle :** la formation se concentre sur l'intégration de l'IA générative pour optimiser et automatiser les tâches de modélisation.
- **Des outils de pointe :** vous utiliserez des outils et langages populaires comme Python, PySpark, Databricks et les modèles d'IA générative les plus performants (GPT-4 et Claude).
- **Un programme complet :** le programme couvre l'ensemble du cycle de vie de la donnée, de la préparation éthique à la mise en œuvre d'un projet complet.
- **Un projet concret :** la formation se conclut par un projet collaboratif qui vous permet d'appliquer immédiatement toutes les compétences acquises dans un contexte professionnel.

Modalités d'évaluation

Travaux Pratiques

Pré-requis

Suivre cette formation nécessite les prérequis suivants :

- **Compétences en manipulation des données** : maîtrise d’Excel ou d’un tableur équivalent, capacité à lire des rapports d’activité et à construire ou interpréter des KPI.
- **Notions de base en statistiques** : compréhension de la corrélation, des indicateurs de tendance centrale (moyenne, médiane) et du principe d’une régression simple.

Public

Cette formation s'adresse à un public de professionnels souhaitant monter en compétences sur la modélisation de données et l'IA, incluant :

- Les **cadres et responsables métiers** qui souhaitent comprendre les enjeux de la donnée.
- Les **responsables de la donnée ou du digital** en quête d’une expertise technique.
- Les **analystes et chargés d’études** qui cherchent à approfondir leurs compétences.
- Les **chefs de projets SI ou innovation** qui ont besoin de maîtriser les cycles de vie des projets IA.
- Les **professionnels en reconversion** vers les métiers de la donnée.

Programme

Module 1 : préparer les données avec l'IA générative

- Les bases de la modélisation des données.
- La distinction entre l'IA classique et l'IA générative.
- Les 5V du Big Data et leurs implications.
- Le panorama des outils d'IA générative (LLM, GPT, Claude, Gemini...).

Démo

- L'exploration automatique d’un jeu de données via GPT.

Travaux pratiques

- Explorer et visualiser un jeu de données avec Python et GPT.

Module 2 : nettoyer et explorer les données

- Les étapes clés de la préparation des données (le nettoyage, la gestion des valeurs manquantes, l'encodage).
- L'utilisation de l'IA générative pour créer des scripts de nettoyage et proposer des visualisations.

Travaux pratiques

- Nettoyer un jeu de données et le préparer pour la modélisation.

Module 3 : construire, évaluer et interpréter des modèles

- La présentation des méthodes de modélisation (les régressions, les forêts aléatoires).
- Les techniques d'évaluation des modèles (la validation croisée, les métriques de performance, les courbes ROC).
- L'utilisation de l'IA générative pour créer du code de modélisation et interpréter les résultats.

Travaux pratiques

- Construire un modèle et interpréter ses résultats avec l'aide de l'IA.

Module 4 : gérer la scalabilité et les grands jeux de données

- L'architecture et la logique des plateformes distribuées comme Spark.
- Les cas d'usage de l'IA pour la gestion des grands jeux de données.
- L'utilisation de l'IA générative pour créer des scripts PySpark, optimiser les pipelines et résumer des transformations.

Travaux pratiques

- Créer un pipeline de modélisation sur un grand jeu de données.

Module 5 : mettre en place la gouvernance et l'éthique

- Les réglementations clés (RGPD) et les enjeux éthiques (les biais, l'auditabilité).
- La présentation des outils d'explicabilité (SHAP, LIME).
- L'utilisation de l'IA générative pour produire de la documentation, des notes d'éthique et des chartes internes.

Travaux pratiques

- Auditer un modèle IA avec un rapport explicatif généré par GPT.

Module 6 : mener un projet complet de modélisation

- Les étapes clés d'un projet (la structuration, la préparation, la modélisation et la documentation).
- La constitution d'équipes pour un projet concret (la prédiction de départ de salarié, la segmentation client, etc.).

Travaux pratiques

- Structurer et mener un projet de modélisation de A à Z.
- Remettre le livrable final **:** le dossier d'analyse augmenté, incluant le code, la visualisation et le rapport avec l'appui de l'IA générative.

GPT est une marque déposée d'OpenAI, Inc.
Claude est une marque déposée d'Anthropic PBC.
Copilot est une marque déposée de Microsoft Corporation.
Gemini est une marque déposée de Google LLC.
Nous ne sommes ni approuvés ni affiliés à aucune de ces sociétés. Les outils cités le sont à titre d'exemple et d'illustration.