Évaluation finale — Data Science avec Python

Contexte général

Vous êtes data scientist au sein d'une organisation (entreprise, ONG, institution publique, startup, etc.) confrontée à une problématique business réelle (ex : prédiction du churn client, détection de fraude, segmentation de clients, prévision de ventes, analyse de satisfaction, optimisation de campagnes marketing, etc.).

Votre mission est de concevoir une solution complète de Data Science, depuis l'exploration des données jusqu'à la mise à disposition d'un modèle exploitable via une API.

Vous travaillez sur des données open source de votre choix (Kaggle, data.gouv.fr, UCI, World Bank, OpenDataSoft, etc.).

Objectifs pédagogiques

- Comprendre une problématique métier
- Explorer, nettoyer et préparer des données réelles
- Construire des modèles de Machine Learning
- Comparer modèles supervisés et non supervisés
- Justifier vos choix techniques
- Mettre en production un modèle de manière simplifiée
- Communiquer clairement vos résultats

Travail attendu

1. Choix de la problématique et des données

Vous devez définir une thématique business claire, expliquer le contexte métier, présenter le jeu de données choisi (source, volume, variables) et formuler un objectif Data Science précis.

2. Analyse exploratoire des données (EDA)

À réaliser dans un notebook Jupyter (.ipynb) avec analyse de la structure, statistiques descriptives, valeurs manquantes, visualisations, corrélations, outliers et hypothèses.

3. Nettoyage et préprocessing

Gestion des valeurs manquantes, encodage, normalisation si nécessaire, avec justification claire de chaque choix.

4. Modélisation

Un modèle supervisé et un modèle non supervisé sont attendus, avec justification des algorithmes et des métriques.

5. Sauvegarde du modèle

Le modèle doit être sauvegardé au format .pkl et pouvoir être rechargé.

6. Création d'une API

Exposer le modèle via une API (Flask ou FastAPI) avec un endpoint /predict.

7. Déploiement (optionnel)

GitHub, README clair et Dockerfile valorisés en bonus.

8. Analyse business et impact

Expliquer l'impact métier, les limites et les perspectives.

Livrables attendus

- eda.ipynb (notebook exploratoire)
- dossier API
- model.pkl
- GitHub / Docker (optionnel)

Barème (/20)

- Problématique business : 2 pts
- Qualité EDA : 3 pts
- Cleaning + preprocessing : 3 pts
- Modèle supervisé : 3 pts
- Modèle non supervisé : 2 pts
- Métriques et interprétation : 2 pts
- Modèle .pkl fonctionnel : 1 pt
- API fonctionnelle : 3 pts
- Analyse business : 1 pt
- Qualité globale du document: 1 pt

Bonus

- Docker fonctionnel : +2