PROJET DATA PIPELINE

PROMO 2028

Description du document

Titre	**Consignes projet data**
Objet	**Consignes pour la réalisation du projet**
Auteur	**Jean Noriot**
Responsable	**Alexandre Pereira De Almeida**
E-mail	**alexandre1.pereira-de-almeida@epitech.eu**
Mots-clés	**Data; Pipeline**
Promotion	2028
Date de mise à **jour**	**02 février 2026**
Version du modèle	**1.1**

Tableau des révisions

Date	Version	Auteur	Section(s)	Commentaires
28/01/2026	1.1	Alexandre Pereira	Toutes	Maj des dates

Table des matières

1. Projet de spé

1.1. Introduction

Ce projet se dÈroule sur 4 jours consÈcutifs ‡ partir du 02 février.
Il est rÈalisÈ par groupe de 3 ‡ 4 Ètudiants, avec pour objectif de concevoir un pipeline de traitement et de modÈlisation de donnÈes.

Vous travaillez pour un laboratoire de recherche botanique qui cherche à estimer la longueur des sépales de fleurs Iris à partir de leur largeur, dans une optique d'analyse automatisée des spécimens.

Les outils utilisés :
Docker, PostgreSQL, MLflow, et Flask ou FastAPI.

La soutenance du projet aura lieu le 06 février.

1.2. Les livrables

Chaque groupe devra réaliser trois livrables : un pipeline fonctionnel, un dossier technique et un support de présentation.

Le pipeline devra être entièrement dockerisé et orchestré à l'aide de Docker Compose.
Il devra inclure :
Un service de prétraitement des données (chargement, nettoyage, normalisation),
Une base de données PostgreSQL,
Un service de modélisation (entraînement, enregistrement dans MLflow),
Une API REST permettant de faire des prédictions.

Le dossier devra présenter :
Le schéma d'architecture du pipeline,
Les choix techniques (outils, librairies, design),
Les performances du modèle,
Une analyse critique et les pistes d'amélioration,
La répartition du travail dans l'équipe.

La présentation orale devra comprendre :
Une démonstration fonctionnelle du pipeline,
Un résumé clair de la démarche technique,
La mise en valeur de la collaboration dans l'équipe.

1.3. Données fournies

Le fichier iris.csv vous est fourni.
Il contient 150 observations de fleurs du genre Iris avec les colonnes suivantes :
sepal length
sepal width
petal length
petal.length
species

1.4. Architecture attendue du pipeline

Le pipeline doit comporter les étapes suivantes :

Prétraitement des données :
Chargement du fichier iris.csv.
Stockage des données dans PostgreSQL.

Modélisation :
Extraction des données depuis PostgreSQL.
Entraînement d'un modèle de régression supervisée (ex : RandomForestRegressor).
Enregistrement du modèle et des métriques dans MLflow.

Déploiement:
Développement d'une API REST avec Flask ou FastAPI.
Chargement du modèle entraîné.
Exposition d'une route /predict acceptant une largeur et retournant une longueur prédite.

Monitoring :
Utilisation de MLflow UI pour consulter les expériences.

1.5. Contraintes techniques

Chaque étape du pipeline doit être contenue dans un service Docker indépendant.
L'ensemble du projet doit être orchestré par Docker Compose.
Les conteneurs doivent communiquer entre eux (nom des services dans Compose).
L'API REST doit être testable (ex : via curl, Postman ou script Python).
Les expériences doivent être suivies avec MLflow (version, métriques, artefacts).
Le projet doit être facilement exécutable avec un README.md clair et une commande docker-compose up.

1.6. Ressources utiles

Docker : https://docs.docker.com/
MLflow : https://mlflow.org/docs/latest/index.html
PostgreSQL : https://www.postgresql.org/docs/
Flask : https://flask.palletsprojects.com/
FastAPI : https://fastapi.tiangolo.com/
Scikit-learn : https://scikit-learn.org/stable/

1.7. **Bonus – Interface utilisateur**

En complément du pipeline, les groupes les plus avancés peuvent développer une interface utilisateur permettant de soumettre une largeur de sépale et d'obtenir une prédiction directement depuis l'API Rest. Cette interface pourra également être déployée dans un conteneur Docker séparé.