Travaux Pratiques – Science des Données

Projet intégrateur : De la donnée brute au déploiement

1. Contexte et objectif du TP

Ce travail pratique a pour objectif de vous faire appliquer **l'ensemble du cycle de vie d'un projet de Science des Données**, à partir de **données réelles**, jusqu'au **déploiement d'une application fonctionnelle avec réentraînement automatique du modèle**.

Les données devront être issues de plateformes ouvertes telles que :

- **Zindi Africa** (fortement recommandé – problématiques africaines)
- **Kaggle**
- **UCI Machine Learning Repository**
- **Etc, …..**

Les thématiques doivent correspondre à des **problèmes réels de la vie courante**, de préférence en lien avec le contexte africain :
santé, agriculture, transport, énergie, finance, ports, éducation, environnement, sécurité, etc.

2. Travail demandé

Chaque groupe devra réaliser un **projet complet de Data Science**, structuré autour des phases suivantes :

2.1 Problématique métier

- Présenter clairement le **contexte métier** du problème traité.
- Identifier les **acteurs concernés** et les enjeux.
- Formuler l'objectif du projet sous forme d'un **problème Data Science** (prédiction, classification, clustering, etc.).
- Justifier l'intérêt du projet et son impact potentiel.

2.2 Analyse exploratoire des données (EDA)

Cette phase devra être **clairement documentée et illustrée** :

- Description du dataset (source, taille, variables)
- Analyse de la qualité des données (valeurs manquantes, doublons, incohérences)
- Statistiques descriptives
- Visualisations pertinentes (distributions, corrélations, tendances)
- Premières hypothèses et enseignements métiers issus des données

2.3 Modélisation

- Choix des algorithmes (justification)
- Prétraitement et feature engineering
- Séparation des données (train / test)
- Entraînement des modèles
- Évaluation des performances (métriques adaptées)
- Comparaison de plusieurs modèles si possible
- Sélection du modèle final

2.4 Déploiement et automatisation

- Déploiement du modèle sous forme d'une **application** (API ou application web)
- Mise en place d'un **processus d'entraînement automatique** (retraining) :
  - script Python
  - pipeline planifié (cron, Airflow, ou équivalent)
- Conteneurisation avec **Docker** (optionnel mais recommandé)

3. Livrables obligatoires

Chaque groupe devra fournir :

1. **Un dépôt GitHub public** contenant :
   - le code source complet (EDA, modèles, API, scripts d'entraînement)
   - un README clair expliquant l'installation et l'exécution du projet
2. **Un rapport écrit complet** (PDF) incluant :
   - la problématique métier
   - l'EDA détaillée
   - la modélisation
   - le déploiement et l'automatisation
   - les résultats et limites du projet
3. **Une présentation PowerPoint (20 slides maximum)** résumant :
   - le contexte
   - la méthodologie
   - les résultats
   - la démonstration de l'application

4. Organisation et modalités

- **Travail en groupe de 3 étudiants**
- **Un seul dépôt et une seule soumission par groupe**
- La soumission se fait **uniquement par le chef de groupe**
- **Date limite : La veille du dernier cours à 23h59**

5. Critères d'évaluation

- Pertinence de la problématique métier
- Qualité de l'EDA et des analyses
- Justesse et rigueur de la modélisation
- Fonctionnalité du déploiement
- Qualité du code et de la documentation
- Clarté du rapport et de la présentation