AAS – licence 3 – UFR MIM

Projet : Analyse, nettoyage et modélisation d'un jeu de données réel

1. Contexte

Les jeux de données réels contiennent souvent des valeurs manquantes, du bruit, des incohérences ou des distributions atypiques. L'objectif de ce projet est de conduire une analyse complète : préparation des données, exploration, sélection d'un modèle de Machine Learning et interprétation des résultats.

Vous devez choisir l'un des jeux de données proposés dans la liste suivante :

- Jeux de données "Apprentissage automatique" de data.gouv.fr
- Consommation d'électricité par région
- Consommation d'électricité par département
- Fréquentation des parkings vélo
- Enseignants titulaires de l'enseignement supérieur (national)

2. Objectifs pédagogiques

- Appliquer des techniques de nettoyage de données.
- Réaliser une analyse exploratoire rigoureuse (EDA).
- Construire un pipeline de Machine Learning : préparation → entraînement → évaluation.
- Comparer plusieurs algorithmes et justifier les choix.
- Rédiger une synthèse claire des résultats.

3. Travail demandé

Étape 1 : Choix et compréhension du dataset

1. Présenter brièvement le jeu de données (origine, variables, taille, objectif).
2. Définir la tâche : classification ou/et régression.

Étape 2 : Nettoyage / Data Cleaning

Effectuer notamment (si besoin) :

- Détection et traitement des valeurs manquantes.
- Gestion des doublons.
- Standardisation des formats.
- Détection des outliers.
- Encodage des variables catégorielles.
- Normalisation ou standardisation.

Étape 3 : Analyse Exploratoire (EDA)

Inclure au minimum :

- Statistiques descriptives.
- Distribution des variables cibles et explicatives.
- Corrélations et heatmaps.
- Visualisations pertinentes.
- Analyse des relations importantes.

Étape 4 : Préparation et séparation des données

- Split train/test.
- Optionnel : validation croisée.

Étape 5 : Construction d'un modèle de ML

Tester au moins deux modèles (régression ou classification).

Étape 6 : Évaluation

Régression : RMSE, MAE, R^2

Classification : Accuracy, F1-score, Matrice de confusion, ROC-AUC

Étape 7 : Interprétation

- Feature importance

- Analyse des erreurs

- Recommandations ou conclusions pratiques

4. Livrables attendus

- Notebook Python propre et commenté (et version Python, version des librairies, etc.).
- Rapport final de 3 pages maximum.

5. Barème indicatif

Total : 30 points.

Partie	Points
Compréhension du dataset	2 pts
Nettoyage des données	4 pts
EDA	4 pts
Préparation & Pipeline	3 pts
Modélisation (2 modèles min.)	5 pts
Évaluation & comparaison	5 pts
Interprétation & clarté du notebook	3 pts
Rapport final	4 pts
Total	**30 pts**