

**Ecole Supérieure de Management, Télécommunication
et d'Informatique Oujda**

Machine Learning

Projet d'application

A. Objectif du projet

Vous devez développer un programme complet s'appuyant sur les techniques d'apprentissage automatique et de traitement de données abordées dans le cours. Ce projet consiste à monter un dataset et préciser un objectif et en suite mettre plusieurs pipelines pour atteindre cet objectif. Vous avez le choix entre une classification ou une régression.

Quelques exemples de problématique :

- Prédiction des performances des élèves
- Reconnaissance de style musical
- Prédiction des défaillances d'entreprise
- Analyse prédictive dans le domaine de la santé pour réduire le risque d'échec du traitement.
- Reconnaissance d'objets
- Prédiction d'évolution du cours de la bourse
-

Vous devriez mettre plusieurs pipelines en précisant :

- Prétraitement des données du dataset choisi en justifiant vos choix
- Analyse exploratoire des données
- Mettre en œuvre plusieurs algorithmes d'apprentissage pour la résolution du problème
- Justification du choix des algorithmes
- Établir une grille d'évaluation des algorithmes avec les différentes métriques
- Faire la comparaison avec une autre approche
- C'est à vous de choisir une problématique à résoudre, et le dataset à utiliser

B. Organisation

- Groupes de 2 ou 3 étudiants
- Livrable du travail dans un référentiel GitHub (publique) contenant :
 - Code source du projet (Jupyter-Nootbook, ...)
 - Dataset utilisé.
 - Un fichier README contenant : la liste des étudiants du groupe, les exigences et les bibliothèques nécessaires pour l'exécution de vos codes, ...
 - Rapport de 7 à 9 pages :
 - Définition d'objectif du projet
 - Source des données
 - Analyse exploratoire des données
 - Description de la phase de pré-traitement des données
 - L'approche (l'algorithme) utilisée pour la résolution du problème
 - Analyse des résultats ...
 - **Présentation (7 à 10 minutes) : le 16 Mai 2025**

• **Liens utiles :**

- UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/index.php>
- Kaggle: <https://www.kaggle.com/>
- Google Dataset Search: <https://datasetsearch.research.google.com/>
- Portail des données US : <https://catalog.data.gov/dataset>
- <https://github.com/awesomedata/awesome-public-datasets>
- <https://github.com/EpistasisLab/pmlb/tree/master/datasets>
- Date limite de soumission : **15 Mai 2025**
- NB : les projets avec des codes copier-coller ils ne seront absolument pas pris en compte.

C. Critères d'évaluation

1. Analyse et traitement des données

- Analyse exploratoire des données : Profondeur de l'analyse pour comprendre les caractéristiques, les tendances, et les anomalies dans les données.
- Visualisation des données : Clarté et pertinence des visualisations utilisées pour illustrer les insights des données.

- Prétraitement des données : Les techniques de nettoyage, de normalisation, de codage, et de sélection des caractéristiques sont-elles appropriées et bien justifiées ?

2. Qualité technique et exactitude des modèles :

- Exactitude des prédictions : la précision, le rappel, le F1-score, ou tout autre métrique pertinente selon le type de problème (classification ou régression).
- Pertinence du modèle : Le modèle choisit-il est adapté au problème ?
- Optimisation et validation : optimisation des hyperparamètres / La validation croisée

3. Structuration du code et utilisation des outils

- Clarté et organisation du code
- Bonne utilisation des bibliothèques de ML

4. Clarté et organisation du rapport et de la présentation.