La Cité Collégiale
Institut des Technologies, des Arts et de la Communication
Projet Final - UA 3

Cours	Apprentissage Machine Appliqué (IFM31103-0020-A2025)
Année académique	2025
Semestre	Automne
Instructeur	Paul Mvula
Annoncé	9 septembre 2025
Date de soumission 8 décembre 2025, 11:59 PM ET	

Projet Final

Objectif du Projet

En groupes de deux (2), les étudiant.e.s devront utiliser des bibliothèques Python pour appliquer des **techniques d'apprentissage supervisé (classification/régression), d'apprentissage non supervisé (clustering)** et **d'ensembles** sur un jeu de données réel (UCI ou Kaggle).

L'objectif est de :

1. Nettoyer et préparer les données.
2. Comparer plusieurs modèles de classification/régression.
3. Explorer des regroupements cachés avec des méthodes de clustering.
4. Améliorer les performances avec des méthodes d'ensembles.
5. Produire des visualisations claires et un rapport documenté.
6. Le projet se déroule sur 12 semaines et une présentation (10 minutes par groupe) est attendue en dernière semaine (le **10 décembre 2025**).

Idées de Projet

1. **Prédiction du départ des employés (HR Analytics – Kaggle)**
   - **Tâches :**
     - Classification supervisée pour prédire la rétention/départ.
     - Clustering pour regrouper les profils d'employés (types de carrières, satisfaction, performance).
     - Utiliser des ensembles (Random Forest, Gradient Boosting) pour comparer les performances.
   - **Concepts couverts :** Prétraitement, sélection de variables, classification, clustering, ensembles.

2. **Analyse des fraudes par carte de crédit (UCI – Credit Card Fraud Dataset)**
   - **Tâches :**
     - Classification binaire (fraude vs normal).
     - Clustering pour détecter des comportements suspects.

    - Combinaison avec bagging/boosting pour améliorer la détection.
- **Concepts couverts :** Données déséquilibrées, métriques adaptées (rappel, F1), ensembles.

3. **Analyse de survie de patients (Breast Cancer Wisconsin – UCI)**
    - **Tâches :**
        - Régression pour prédire la gravité/stade.
        - Classification binaire (bénin vs malin).
        - Clustering pour explorer des sous-groupes de patients.
        - Comparaison de modèles avec stacking (ensembles).
    - **Concepts couverts :** Classification, régression, clustering médical, ensembles.

4. **Prédiction de prix de l'immobilier (Housing Prices – Kaggle)**
    - **Tâches :**
        - Régression supervisée pour prédire les prix.
        - Clustering des quartiers ou types de maisons.
        - Ensembles (XGBoost, Random Forest) pour améliorer la performance.
    - **Concepts couverts :** Régression multiple, validation croisée, ensembles.

5. **Reconnaissance de chiffres manuscrits (MNIST – UCI)**
    - **Tâches :**
        - Classification multiclasse (chiffres 0–9).
        - Réduction dimensionnelle + clustering pour découvrir des regroupements cachés.
        - Ensembles de classifieurs (bagging/boosting) pour améliorer la précision.
    - **Concepts couverts :** Classification avancée, clustering non supervisé, ensembles.

6. **Analyse des plaintes des consommateurs (Consumer Complaints – Kaggle)**
    - **Tâches :**
        - Traitement de texte → classification des types de plaintes.
        - Clustering thématique des plaintes (NLP non supervisé).
        - Comparaison d'ensembles (boosting, stacking) sur des représentations TF-IDF.
    - **Concepts couverts :** NLP appliqué, classification, clustering, ensembles.

7. **Analyse des habitudes de consommation (Online Retail Dataset – UCI)**
    - **Tâches :**
        - Clustering des clients (segmentation RFM).
        - Classification de la probabilité de réachat.

    - Ensembles pour optimiser la prédiction des comportements futurs.
- **Concepts couverts :** Clustering (k-means, DBSCAN), classification, ensembles.

8. **Prédiction de la qualité des vins (Wine Quality – UCI)**
    - **Tâches :**
        - Régression pour prédire le score de qualité.
        - Classification multiclasse (mauvais, moyen, bon vin).
        - Clustering des vins selon les propriétés chimiques.
        - Comparaison d'ensembles pour déterminer la meilleure approche.
    - **Concepts couverts :** Régression, classification, clustering, validation croisée, ensembles.

9. **Analyse de performance académique (Student Performance Dataset – UCI)**
    - **Tâches :**
        - Régression pour prédire les notes finales.
        - Classification (succès vs échec).
        - Clustering des profils d'étudiants.
        - Ensembles pour combiner les prédictions.
    - **Concepts couverts :** Prétraitement, classification, régression, clustering, ensembles.

Données à Utiliser

Projet	Jeu de données	Lien	Concepts couverts
Prédiction du départ des employés	Employee Attrition & Performance	Kaggle	Classification, Clustering, Ensembles
Détection de fraude par carte de crédit	Credit Card Fraud Detection	Kaggle	Classification déséquilibrée, Clustering, Ensembles
Analyse de survie de patients	Breast Cancer Wisconsin (Diagnostic)	UCI	Classification binaire, Régression, Clustering, Ensembles
Prédiction des prix de l'immobilier	House Prices: Advanced Regression Techniques	Kaggle	Régression multiple, Clustering, Ensembles
Reconnaissance de chiffres manuscrits	MNIST Handwritten Digit Dataset	Kaggle	Classification multiclasse, Clustering, Ensembles
Analyse de plaintes consommateurs (NLP)	Consumer Complaints Dataset	Kaggle	NLP, Classification, Clustering, Ensembles
Segmentation des clients et prévision d'achats	Online Retail Data Set	UCI	Clustering, Classification, Ensembles
Analyse de la qualité des vins	Wine Quality Data Set	UCI	Régression, Classification multiclasse, Clustering, Ensembles
Performance académique des étudiants	Student Performance Data Set	UCI, Kaggle	Régression, Classification, Clustering, Ensembles

Chaque dataset contient plusieurs colonnes (numériques et/ou catégorielles). Les étudiants devront :

- Identifier les variables cibles (classification ou régression).
- Sélectionner des variables pertinentes.
- Justifier leurs choix en fonction des objectifs du projet.

Bibliothèques à Utiliser

Les bibliothèques suivantes sont obligatoires :

- **NumPy / Pandas** : manipulation et préparation des données.
- **Matplotlib / Seaborn / Plotly** : visualisations.
- **scikit-learn** : modèles de classification, régression, clustering, et ensembles.
- **XGBoost / LightGBM** : ensembles avancés.

Tâches à Réaliser

1. Chargement et préparation des données
    a. Nettoyage, encodage des variables catégorielles, normalisation.
    b. Gestion des valeurs manquantes.
    c. Analyse descriptive des données.
2. Analyse exploratoire et visualisation
    a. Visualisations des distributions et corrélations.
    b. Graphiques par classes, regroupements, ou tendances temporelles (si applicable).
3. Modélisation supervisée (classification/régression)
    a. Entraîner au moins deux modèles différents (ex. : régression linéaire/logistique, SVM, arbres de décision).
    b. Comparer les performances avec validation croisée et métriques adaptées (précision, rappel, RMSE, etc.).
4. Modélisation non supervisée (clustering)
    a. Appliquer au moins une méthode de clustering (k-means, DBSCAN, hiérarchique).
    b. Interpréter les résultats et comparer la qualité des clusters.
5. Méthodes d'ensembles
    a. Implémenter au moins un modèle d'ensemble (Random Forest, Gradient Boosting, Bagging).
    b. Comparer avec les modèles individuels.
6. Rapport et présentation
    a. Rédiger un rapport documenté (4–6 pages) décrivant : données, méthodes, résultats, interprétation.
    b. Préparer une présentation orale avec diapositives (10 minutes par groupe).

À soumettre

Dans un fichier compressé (.zip, .rar, .tar, .tar.gz), veuillez inclure:

1. Le notebook (.ipynb) ou script (.py)
2. Une présentation (PowerPoint) – à présenter en **10 minutes le 29 Mai**.
3. Un rapport (format .pdf – 5 pages maximum)

Évaluation (100%)

Critère	Description	Poids
Préparation et qualité des données	Nettoyage, gestion des valeurs manquantes, encodage, normalisation ; pertinence des choix faits (justifiés dans le rapport).	15 %
Qualité du code	Clarté, organisation, efficacité ; respect des bonnes pratiques (commentaires, modularité).	15 %
Analyse exploratoire et visualisations	Pertinence et qualité des statistiques descriptives et graphiques ; clarté des visualisations.	15 %
Modélisation supervisée	Mise en œuvre de plusieurs modèles de classification/régression ; usage de validation croisée ; choix des métriques adaptés ; justification des résultats.	20 %
Modélisation non supervisée (clustering)	Application correcte d'au moins une méthode ; interprétation des résultats ; évaluation des clusters (silhouette, Davies-Bouldin, etc.).	10 %
Techniques d'ensembles	Utilisation d'au moins une méthode (bagging, boosting, random forest, etc.) ; comparaison avec modèles de base.	10 %
Rapport écrit	Structure claire, synthèse pertinente, discussion critique (forces/limites, interprétation des résultats).	10 %
Présentation orale	Clarté et concision des diapositives ; qualité de l'explication ; respect du temps (10 min).	5 %