Rapport Projet ML : Analyse de Performance Académique - (Kaggle)

1. Objectif

Effectuer **L'apprentissage machine** (Machine Learning) sur le **Student Performance Dataset (Kaggle)** afin d'analyser l'impact des habitudes d'études sur les performances des étudiants.

2. Tâches d'Apprentissage Machine Définies

Le projet vise à réaliser quatre tâches distinctes :

1. **Régression** : Prédire les notes finales (scores).
2. **Classification** : Distinguer le succès de l'échec des étudiants.
3. **Clustering** : Segmenter les profils d'étudiants en groupes (non supervisé).
4. **Méthodes d'Ensemble** : Combiner plusieurs modèles pour améliorer la robustesse des prédictions.

3. Description et Préparation des Données

Aperçu du Jeu de Données

- **Observations :** 1000 entrées.
- **Variables :** 8 colonnes.
- **Qualité des Données :** Le Notebook confirme l'absence de valeurs manquantes et de duplicatas.

Variables

Type	Variable	Exemples de Valeurs
Catégorielles	gender, race/ethnicity, lunch, test preparation course, parental level of education	Female/Male, Group A/Group B, Standard/Free Lunch, completed/none, High School, Master's Degree, Some College
Numériques	math score, reading score, writing score	Scores allant de 0 à 100

Stratégies d'Encodage

Pour préparer les données aux modèles de Machine Learning, différentes techniques d'encodage ont été appliquées aux variables catégorielles :

- **One-Hot Encoding (OHE)** pour les variables sans ordre (gender, lunch, test preparation course).
- **Ordinal Encoding (OE)** pour la variable parental level of education afin de respecter la hiérarchie des niveaux.
- **Frequency Encoding (FE)** pour la variable race/ethnicity (fréquence d'apparition du groupe).

4. Résultats des Modélisations

Le Notebook présente les résultats des premiers entraînements de modèles pour la classification et le clustering.

A. Classification

- **Modèle Utilisé : SVM, RandomForest Classifier (RFC)**.
- **Objectif :** Prédire le succès (vs échec) de l'étudiant.
- **Métriques Obtenues :**
    - **Accuracy (Précision globale)** : **0.74**
    - **Precision (Précision)** : 0.7068...
    - **Recall (Rappel)** : 0.4606...

B. Clustering

- **Modèle Utilisé : K-Means** (avec 2 clusters).
- **Objectif :** Regrouper les profils d'étudiants similaires.
- **Évaluation :** Le clustering est évalué en utilisant les métriques suivantes :
    - Score **Calinski-Harabasz**.
    - Score **Silhouette**.
    - Score **Davies-Bouldin**.