Projet de Fin de Module : Analyse de Données

Données à analyser

Les étudiants utiliseront le fichier suivant :
ITU_regional_global_Key_ICT_indicator_aggregates_Nov_2024.xlsx

L'analyse devra porter sur l'une des trois premières feuilles du fichier Excel :
- By dvpmnt status and spec. reg.
- By BDT region
- By urban-rural area

Chaque groupe doit choisir et extraire les données pertinentes dans un fichier .csv.

Organisation

- Travail en groupe de 3 étudiants
- Chaque groupe définit une variable cible (à justifier)

Objectifs pédagogiques

Le projet permet de mobiliser l'ensemble des techniques vues en cours :

1. Exploratory Data Analysis (EDA) :
   - Statistiques descriptives, visualisations, traitement des valeurs manquantes
2. Préparation des données :
   - Nettoyage, normalisation, transformation
3. Analyse en Composantes Principales (ACP/PCA)
4. Modélisation supervisée :
   - SVM (classification ou régression selon la variable cible)
   - Decision Tree, Random Forest, XGBOOST
5. Tuning des hyperparamètres (GridSearchCV) :
   - Application de GridSearchCV pour optimiser les performances des modèles
   - Évaluation des modèles avant/après tuning (courbes de validation, cross-validation)
6. Sélection des variables ("Best features") :
   - À l'aide de la Random Forest et de l'analyse des importances
7. Interprétation et conclusion :
   - Analyse critique des résultats
   - Recommandations ou pistes d'amélioration

Livrables attendus

- Un notebook Jupyter (.ipynb) bien structuré, documenté et commenté avec une courte note de synthèse présentant les choix et résultats principaux
- Son exportation HTML
- Le ou les fichiers .csv utilisés

Date de rendu

1/06/2025 à 23h59

Critères d'évaluation

L'évaluation du projet se fera selon les critères suivants :

Critère	Barème
Qualité du nettoyage et de la préparation des données	5 pts
Pertinence et justification de la variable cible	5 pts
Analyse exploratoire (EDA) et visualisation pertinente et bien illustrée	10 pts
Utilisation correcte de la PCA	5 pts
Modélisation SVM (classification/régression) avec évaluation	5 pts
Utilisation de Decision Tree, Random Forest, XGBOOST et analyse des features	10 pts
Application de GridSearchCV pour l'optimisation	10 pts
Clarté des interprétations et des conclusions	5 pts
Structure et lisibilité du notebook	5 pts
Qualité du rendu (html, csv, documentation)	5 pts