

Objectif du TP

ESGI a pour ambition de mettre en place un accompagnement personnalisé pour chaque étudiant de la filière AL. Pour ce faire, elle a besoin de connaître les types d'étudiants présents dans cette filière.

Données disponibles

Fichier fourni : `ML_5AL_students_dataset.csv`

Nombre de lignes : 50 000

Chaque ligne représente un étudiant.

Variables principales :

Colonne	Description
---------	-------------

-----	-----
-------	-------

`Age`	Âge de l'étudiant
-------	-------------------

`Gender`	Sexe
----------	------

`Academic_Year`	Année académique (de 2005 à 2024)
-----------------	-----------------------------------

`Region`	Région géographique
----------	---------------------

`Residence_Type`	Type de résidence (Urbain, Rural...)
------------------	--------------------------------------

`Student_Income_EUR`	Revenu mensuel de l'étudiant
----------------------	------------------------------

`Parent_Income_EUR`	Revenu mensuel des parents
---------------------	----------------------------

`Grade_Math` , `Grade_Programming` , etc.	Notes dans les différentes matières
---	-------------------------------------

`Average_Grade`	Moyenne calculée sur toutes les notes
-----------------	---------------------------------------

`Status`	Cible à prédire : `Success` ou `Failure`
----------	--

1. Préparation des données et Analyse exploratoire des données (EDA) (3 pts)

- Compréhension de la problématique métier et annonce de la stratégie d'étude
- Présenter la donnée : dimension et variables et leur type
- Faire toutes les étapes de la préparation des données
- Faire quelques graphiques avec des variables quali, des variables quanti et graphique quali x quanti
- Faire une table de corrélation

- - Maitrise de la préparation, de l'analyse et de la visualisation des données

2. Entraînement des modèles (2 pts)

- Entraîner, évaluer et visualiser au moins ****3 modèles différents****
- Comparer leurs performances
- Choisir un modèle

3. Interprétation : Analyse des clusters (3 points)

- Identifier et nommer chaque cluster en fonction des caractéristiques les plus importantes
- Quelle est la répartition du sexe et de l'âge dans les différents clusters
- Dans quel cluster se trouvent les étudiants ayant le plus réussi
- Dans quel cluster se trouvent les étudiant dont les parents gagnent le moins
- Est-ce qu'il y a un lien entre les régions d'habitations et les clusters
- Proposez une stratégie d'accompagnement pour chaque cluster

Livrables attendus

1. ****Notebook Jupyter**** avec :

- Code propre et commenté
- Visualisations utiles
- Explications des choix et interprétations

2. récapitulatif :

- Méthodologie
- Résultats
- Recommandations finales