جامعة مولاي إسماعيل
UNIVERSITÉ MOULAY ISMAÏL

كلية العلوم والتقنيات
FACULTÉ DES SCIENCES ET TECHNIQUES

# PROJET DE DATA MINING

**Master :** Système d'Information décisionnels et Imagerie (TA)

**Réalisé par :** Ayoub Ourghite

**Année Universitaire :** 2025/2026

## Introduction

Ce projet de Data Mining a pour objectif d'analyser les performances académiques sur la discipline mathématique des étudiants et d'identifier les facteurs influençant leur réussite dans cette matière. Nous combinons des techniques d'analyse exploratoire, de règles d'association et de classification supervisée afin d'extraire des connaissances utiles et prédire la classe de performance e en mathématiques dans une optique d'aide à la décision.

## Problématique du projet

**La problématique de ce projet consiste à:**

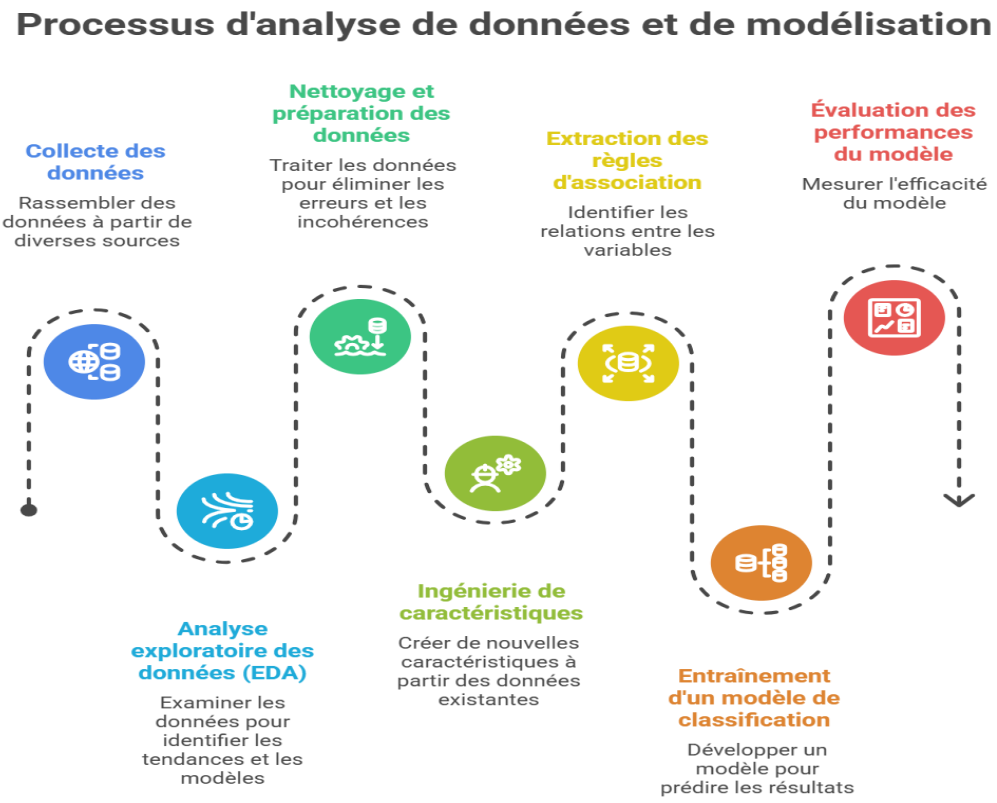
Analyser l'impact des variables sociodémographiques et éducatives sur la performance des étudiants en mathématiques, à identifier les relations significatives entre ces variables à travers l'extraction de règles d'association, et à développer un modèle de classification capable de prédire le niveau de performance des étudiants en mathématique (Low, Medium, High).

## Description du Dataset

Le dataset "Student Performance"contient des informations démographiques, sociales et académiques sur 1000 étudiants. Les principales variables sont :

- **gender :** sexe de l'étudiant
- **race/ethnicity :** Groupe ethnique auquel appartient l'étudiant (*group A, group B, etc.*)
- **parental level of education :** niveau d'éducation des parents
- **lunch :** type de repas (standard ou free/reduced)
- **test preparation course :** préparation au test (completed / none)
- **math score :** score en mathématiques
- **reading score :** score en lecture
- **writing score :** score en écriture

# Étapes du Projet de Data Mining




## 1. Collecte des données

La phase de collecte des données constitue la première étape du processus de Data Mining. Elle consiste à acquérir un jeu de données pertinent et fiable permettant d'analyser les performances académiques des étudiants en mathématiques.
Dans ce projet, le dataset comprend des informations relatives à 1000 étudiants, incluant des variables sociodémographiques (genre, origine ethnique), éducatives (niveau d'éducation des parents, type de repas, préparation au test) ainsi que les scores obtenus en mathématiques, lecture et écriture.
**Source :** https://www.kaggle.com/datasets/spscientist/students-performance-in-exams/data

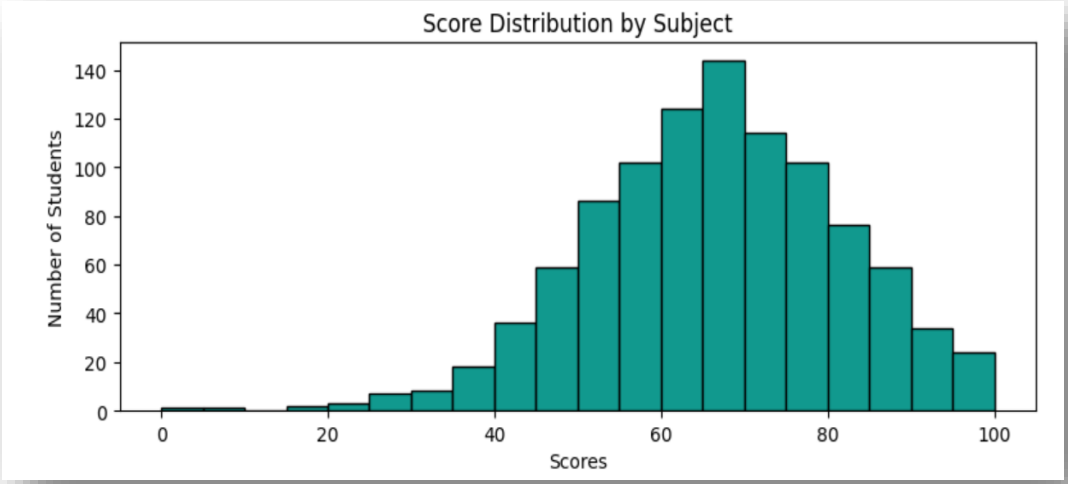
## 2. Analyse exploratoire des données (EDA)

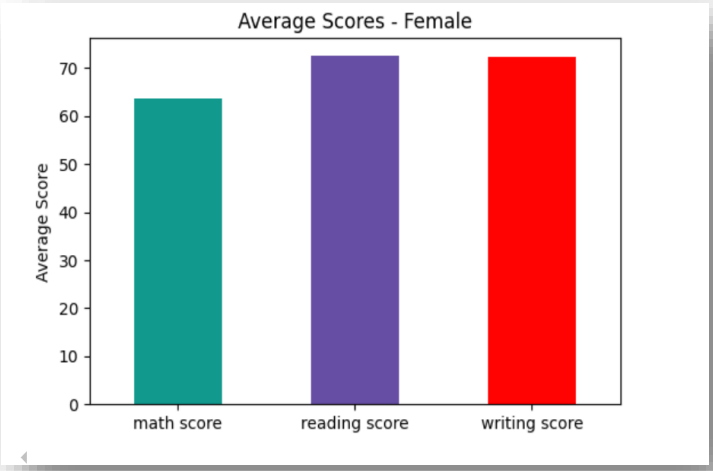
L'analyse exploratoire des données (EDA) a pour objectif de comprendre la structure du dataset et d'identifier les tendances principales avant la modélisation.

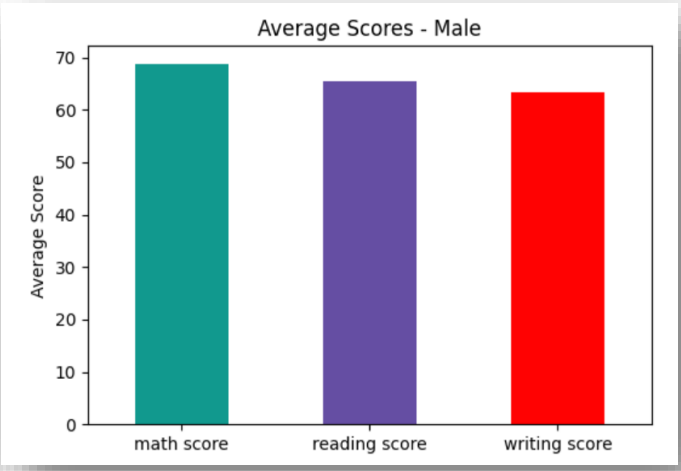
**Dans ce projet, l'EDA a permis de :**

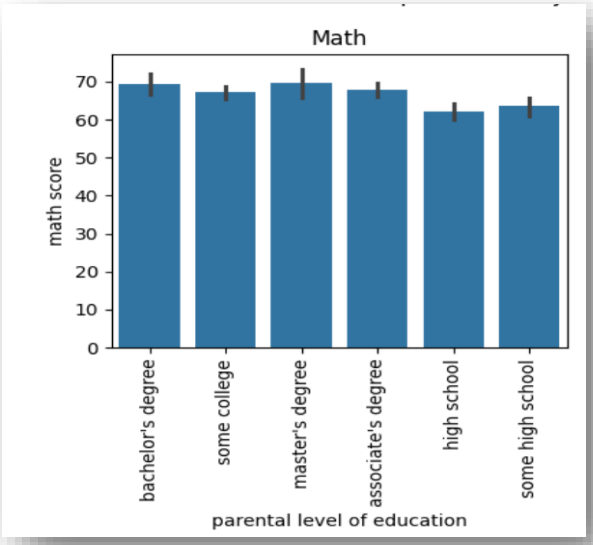
- Vérifier la taille du dataset et la détection de valeurs manquantes.
- Analyser la distribution des scores en mathématiques, lecture et écriture.
- Étudier la répartition des étudiants selon le genre, le type de repas (lunch) et la préparation au test.
- Observer la distribution de la variable cible ***math score***.
- Comparer les moyennes des scores selon certaines variables explicatives (genre, préparation au test, type de repas).
- Analyse de la corrélation entre les scores en Mathématiques, Lecture et Écriture en utilisant

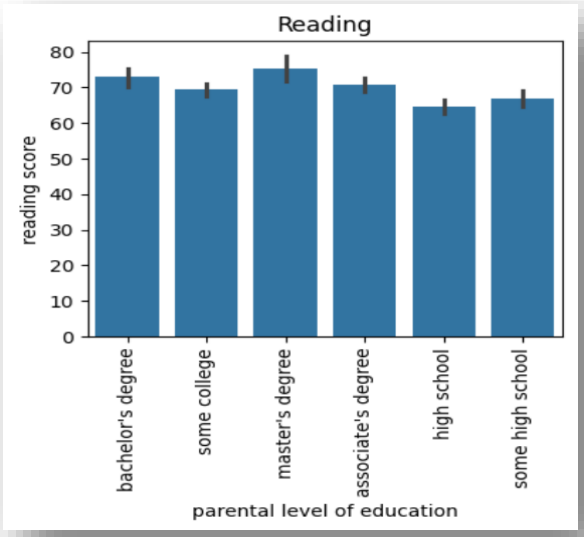
## ➢ Exemples des analyses univariée/multivariée

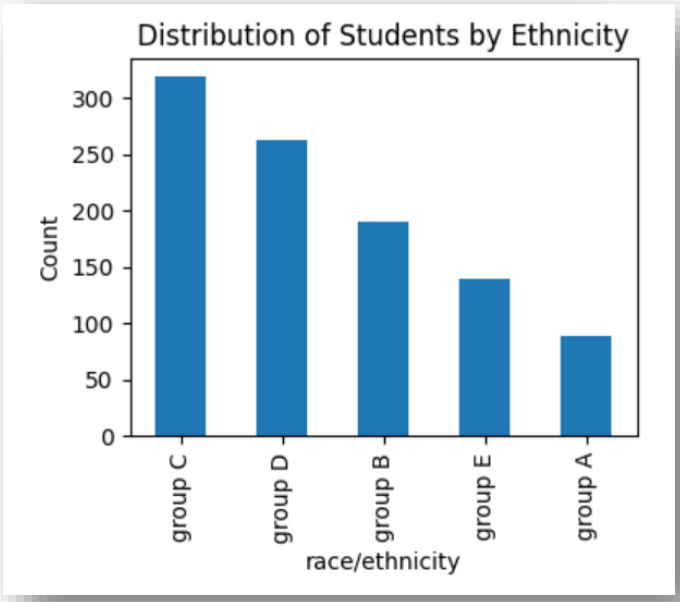


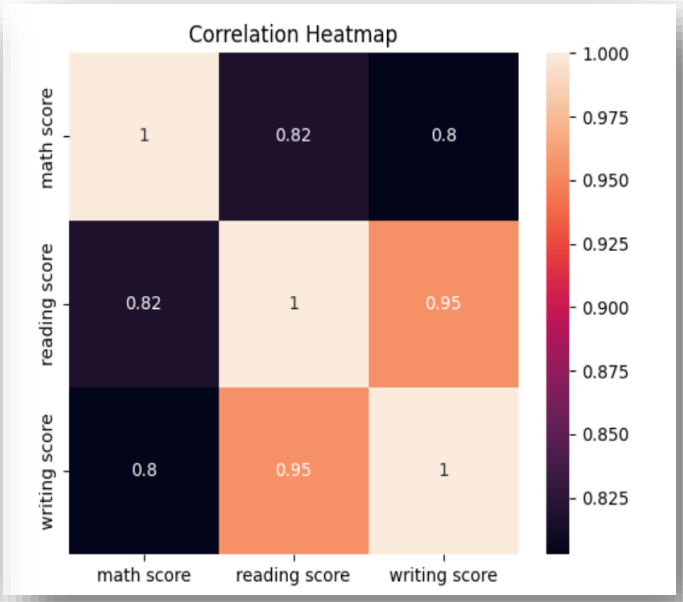






## Nettoyage et préparation des données

Cette étape consiste à assurer la qualité et la cohérence des données avant la modélisation.

Elle comprend :

- La vérification de l'absence de valeurs manquantes ou incohérentes.
- La sélection des variables pertinentes.
- La séparation des variables explicatives (X) et de la variable cible (y).
- La normalisation des variables numériques.

- L’encodage des variables catégorielles à l’aide d’un Pipeline afin de garantir un traitement correct des données lors de l’entraînement du modèle.

## 3. Feature Engineering (ingénierie de caractéristiques)

- Discrétisation des scores pour faciliter les règles d’association **Low, Medium, High**.
- La variable dérivée **math_cat** à partir du score en mathématiques (**math score**), est définie comme **variable cible** en le catégorisant en trois niveaux : **Low, Medium et High.**
- Création de la variable **avg_score** moyenne des trois scores (math, reading, writing).
- Transformation de cette moyenne en classes **Low, Medium, High**.

## 4. Extraction des règles d'association (Apriori)

Les règles d'association ont été extraites à l'aide de l'algorithme Apriori. Ces règles permettent d’identifier des relations fréquentes entre les variables.

➢ **Les métriques utilisées pour évaluer les règles :**

- **Support :** fréquence d'apparition de la règle dans le dataset.
- **Confidence :** probabilité que la conclusion soit vraie si la condition est vraie.
- **Lift :** mesure la force réelle de l'association (Lift > 1 indique une relation positive).

➢ **Interprétation de quelques règles :**

☑ **(Reading_High)=>(Math_High)**

- **Support = 0.33 :** ce qui signifie que 33 % des étudiants ont à la fois un score élevé en lecture et un score élévé en math.
- **Confidence =** 0.681 : 68 % des étudiants ayant un score élevé en reading ont aussi un score élevé en math.
- **Lift =** 1.74: relation forte positive.

☑ **(Reading_High) => (Lunch_standard)**

- **Support de 0.356** : ce qui signifie que 35,6 % des étudiants ont à la fois un score élevé en lecture et un lunch standard.
- **Confidence est de 0.731** : indiquant que 73,1 % des étudiants ayant un score élevé en lecture bénéficient d’un lunch standard.

- **Lift est de 1.13** : ce qui montre une association positive entre **Reading_High** et **Lunch_standard**

## 5. Entraînement d'un modèle de classification supervisée multiclasses (3 classes) Random Forest

Le modèle Random Forest est un algorithme d'ensemble basé sur plusieurs arbres de décision. Il permet d'améliorer la robustesse et la précision en réduisant le risque d'overfitting.

- La variable cible est performance en math **(math_cat)** => Low, Medium, High.
- Les données sont séparées en 80 % pour l'entraînement et 20 % pour le test.
- Les variables numériques sont standardisées et les variables catégorielles encodées via un Pipeline.

➢ **Paramètres principaux utilisés pour Random Forest** :

- **n_estimators =** 2000 (nombre d'arbres)
- **max_depth =** 10 (contrôle de la complexité)
- **class_weight =** balanced (gestion du déséquilibre des classes)

## 6. Évaluation des performances du modèle

L'évaluation du modèle a été réalisée à l'aide des métriques suivantes :

- **Confusion Matrix :** compare les prédictions aux valeurs réelles.
- **Accuracy :** proportion globale de bonnes prédictions.
- **Precision :** capacité du modèle à éviter les faux positifs.
- **Recall :** capacité à détecter correctement les vraies classes.
- **F1-score :** moyenne harmonique entre précision et rappel.

➢ Une bonne performance du modèle se traduit par des valeurs élevées d'accuracy, précision, recall et F1-score, ainsi qu'une matrice de confusion équilibrée.

➢ Le modèle est évalué aussi sur un nouvel étudiant en analysant la classe prédite et les probabilités associées.

```
Confusion matrix:
 [[21  9  0]
 [ 3 80  9]
 [ 0 15 63]]

Report:
                precision    recall  f1-score   support

           0       0.88      0.70      0.78        30
           1       0.77      0.87      0.82        92
           2       0.88      0.81      0.84        78

    accuracy                           0.82       200
   macro avg       0.84      0.79      0.81       200
weighted avg       0.83      0.82      0.82       200


Classe prédite (encodée): 2
Classe prédite (nom): High
Ordre des classes: [0 1 2]
Probabilités: [[0.00929988 0.43202313 0.55867699]]
```

## Importance des variables

L'importance des variables indique quelles caractéristiques influencent le plus la prédiction du modèle Random Forest, les scores académiques (reading, writing) étant les plus déterminants.

```
Importance of features
reading score                  0.320
writing score                  0.319
gender                         0.095
parental level of education    0.081
race/ethnicity                 0.077
lunch                          0.075
test preparation course        0.033
Name: Importance, dtype: float64
```

## Conclusion

Ce projet a démontré que les performances en mathématiques, en lecture et en écriture sont fortement corrélées et constituent les principaux facteurs prédictifs du niveau global des étudiants. Le modèle **Random Forest** a obtenu de bonnes performances, confirmant la pertinence des variables académiques dans la prédiction. Enfin, une mise en production du modèle permettrait son déploiement dans un contexte réel, afin d'automatiser la prédiction et de faciliter la prise de décision.