**Rapport Livrable 3**

# I - Présentation des données

## Présentation des données

Le fichier vue\_voyelles\_m2105.csv contient plusieurs séries statistiques extraites de la base de données des étudiants en DUT Informatique :

* La population étudiée correspond aux étudiants de 1ère année de DUT, identifiés par leur code\_nip.
* La 1ʳᵉ variable est le prénom de l’étudiant.
* La 2ᵉ est la note obtenue au module M2105, qui porte sur les statistiques (au semestre 2). Cette variable est celle qu’on souhaite expliquer.
* Les autres variables sont les notes de modules issus du semestre 1 et début S2 :  
  + m1101 : Algorithmique et programmation
  + m1102 : Mathématiques discrètes
  + m1201 : Projet tutoré
  + m2101 : Système

En tout, la vue contient 5 variables explicatives et un total de plus de 20 individus (étudiants).

## Problématique

En utilisant ces données, on va essayer de répondre à la problématique suivante :

*Est- ce qu' avoir une voyelle dans son prénom permet d’avoir une meilleure moyenne.*

# II - Importation et traitement des données

## (a) Importation des données en Python

Les données sont importées en Python sous forme de DataFrame à l’aide de la commande suivante :

import pandas as pd

df = pd.read\_csv("Colleges.csv")

## (b) Mise en forme

* Suppression des valeurs manquantes (NaN) :

df.dropna(inplace=True)

* Conversion en array pour traitement numérique :

import numpy as np

data\_array = df.to\_numpy()

## (c) Normalisation

On conserve uniquement les colonnes contenant des données numériques. Ces données sont ensuite centrées-réduites :

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

data\_scaled = scaler.fit\_transform(df.select\_dtypes(include=np.number))

# III - Représentations graphiques

Nous visualisons certaines variables par des diagrammes en bâtons :

• Notes au brevet

* Les notes moyennes sont majoritairement en dessous de 10.
* Certains collèges ont des moyennes très faibles (inférieures à 7 !).

• Pourcentage de filles en 3e

* La plupart des collèges se situent autour de la parité (45–55%).
* Quelques établissements présentent un fort déséquilibre : jusqu’à 2/3 de filles ou de garçons.

*(À compléter avec d’autres diagrammes pertinents)*

# IV - Exploration par la matrice des corrélations

## (a) Démarche

On calcule la matrice des coefficients de corrélation de Pearson entre toutes les variables numériques.

correlation\_matrix = df.corr(numeric\_only=True)

## (b) Matrice obtenue

La matrice ainsi obtenue permet de repérer les variables les plus corrélées à la note moyenne au brevet.

# V - Régression linéaire multiple

## (a) Méthodologie

On utilise la régression linéaire multiple pour estimer la note moyenne au brevet (variable dépendante) à partir de certaines variables explicatives.

## (b) Sélection des variables explicatives

Les variables ayant les coefficients de corrélation les plus élevés (en valeur absolue) avec la note au brevet sont :

* IPS moyen (variable 5) : 0.617
* Écart-type de l’IPS (variable 6) : 0.606
* Valeur ajoutée (variable 1) : 0.486
* % de filles (variable 4) : 0.031

Nous choisissons donc ces quatre variables comme explicatives.

## (c) Lien avec la problématique

Les coefficients obtenus lors de la régression nous indiqueront l'influence réelle de chaque variable sur la note au brevet, et si cette influence est suffisante pour faire de la prédiction.

## (d) Régression en Python

from sklearn.linear\_model import LinearRegression

X = df[['IPS', 'EcartTypeIPS', 'ValeurAjoutee', 'PourcentageFilles']]

y = df['NoteBrevet']

model = LinearRegression()

model.fit(X, y)

coefficients = model.coef\_

intercept = model.intercept\_

## (e) Interprétation des paramètres

* Le signe de chaque coefficient indique le sens de l’influence.
* Les variables sont centrées-réduites, donc les coefficients peuvent être directement comparés pour estimer leur importance relative.

## (f) Coefficient de corrélation multiple

Le coefficient de détermination (R²) permet d’estimer la qualité de la prédiction :

r2 = model.score(X, y)

# VI - Conclusions

## (a) Réponse à la problématique

Certaines variables (notamment l’IPS et la valeur ajoutée) présentent une corrélation significative avec les performances au brevet. Il est donc possible de modéliser une partie de ces performances.

## (b) Argumentation basée sur la régression

Les résultats montrent que :

* L’IPS a un fort impact positif.
* L’écart-type de l’IPS est également significatif (hétérogénéité sociale).
* La valeur ajoutée joue un rôle notable.
* Le pourcentage de filles a un impact faible.

## (c) Interprétations personnelles

La réussite scolaire semble fortement influencée par le contexte social des établissements. Cela souligne le besoin d’actions ciblées dans les établissements défavorisés pour réduire les inégalités.

Souhaitez-vous que je vous génère une version PDF ou LaTeX de ce rapport ?