Student Performance Analyse :

Le projet présente une analyse complète des performances académiques des étudiants, en utilisant le dataset *Students Performance*. L'objectif est d'explorer les relations entre les différentes variables et les niveaux de performance en mathématiques, lecture et écriture.

1. Description du Dataset

Variable	Type	Description
gender	Catégoriel	Genre de l'étudiant (male/female)
race/ethnicity	Catégoriel	Groupe ethnique
parental level of education	Catégoriel	Niveau d'éducation des parents
lunch	Catégoriel	Type de déjeuner (standard ou free/reduced)
test preparation course	Catégoriel	Suivi ou non d'un cours de préparation
math score	Numérique	Score en mathématiques
reading score	Numérique	Score en lecture
writing score	Numérique	Score en écriture

2. Prétraitement

- Encodage des variables catégorielles avec LabelEncoder.
- Discrétisation des scores en **Low**, **Medium** et **High** pour math, lecture et écriture.
- Création de nouvelles features d'interaction si nécessaire.

Exemple :

```
df['math_level_num'] =
df['math_level'].map({'Low':1,'Medium':2,'High':3})
```

```
df['reading_level_num'] =
df['reading_level'].map({'Low':1,'Medium':2,'High':3})
df['math_reading_interaction'] = df['math_level_num'] *
df['reading_level_num']
```

3. Analyse Exploratoire

3.1 Distribution des niveaux de math par genre

Genre	Low (%)	Medium (%)	High (%)
Male	25%	50%	25%
Female	20%	55%	25%

> Observation : Les étudiantes ont tendance à obtenir un peu plus de scores moyens en mathématiques.

3.2 Distribution par type de lunch

Lunch Type	Low (%)	Medium (%)	High (%)
Standard	22%	52%	26%
Free/Reduced	30%	45%	25%

> Observation : Les étudiants avec lunch free/reduced ont légèrement plus de faibles scores en math.

3.3 Corrélation entre variables numériques

- Matrice de corrélation pour identifier les relations fortes entre les scores.

```
import seaborn as sns
import matplotlib.pyplot as plt

corr =
df[['math_level_num','reading_level_num','writing_level_num']].corr()
```

```
sns.heatmap(corr, annot=True, cmap="Blues")
plt.show()
```

> Observation : Les scores en math, lecture et écriture sont fortement corrélés, surtout entre lecture et écriture (~0.85).

4. Modélisation

- **Régression logistique** pour prédire le niveau en mathématiques.
- Split train/test 75/25 et traitement éventuel de déséquilibre.

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report

X = df.drop(['math_level','reading_level','writing_level'], axis=1)
y = df['math_level_num']

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.25, random_state=42)

model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

print(classification_report(y_test, y_pred))
```

> Observation : Le modèle montre une précision satisfaisante sur la prédiction des niveaux de mathématiques, mais des améliorations sont possibles avec des features supplémentaires.

5. Conclusions

- Les performances en math, lecture et écriture sont corrélées.

- Le genre et le type de lunch ont un impact modéré sur les scores.

- Les nouvelles features d'interaction peuvent améliorer la modélisation
- Des analyses supplémentaires peuvent inclure l'exploration des interactions complexes ou l'application d'algorithmes plus puissants (Random Forest, Gradient Boosting).