

# Projet de Machine learning



## Student Performance



Réalisé par:

**M. Ousseynou DIOP**

**M. Seman Giovanni Jocelyn GADO**

**M. Omar THIAM**

**M. Cheikhna Amala YATABARE**

Formatrice :

**Mme Mously DIAW**

# INTRODUCTION

## 1. CONTEXTE

- La réussite éducative est un enjeu majeur des sociétés modernes.
- L'éducation vise à développer des compétences critiques et analytiques, comme le souligne Albert Einstein :  
*"L'éducation n'est pas l'apprentissage des faits, mais la formation de l'esprit à penser".*
- De nombreux défis subsistent : conditions socio-économiques, motivations personnelles, environnemental et familial.
- Grâce aux technologies avancées et au Machine Learning, il est possible d'analyser ces facteurs en profondeur
- L'exploitation des données éducatives ouvre la voie à des solutions concrètes pour améliorer les stratégies d'apprentissage.

# INTRODUCTION

## 2. OBJECTIFS

L'objectif général de ce projet est de développer un modèle de régression capable de prédire la performance académique des étudiants à partir de données contextuelles et sociodémographiques.

Les objectifs spécifiques incluent :

- Identifier les variables ayant une influence significative sur la performance académique des étudiants.
- Construire un modèle prédictif robuste en utilisant des techniques de Machine Learning adaptées.
- Évaluer la performance et la précision des prédictions à l'aide de métriques fiables.
- Déployer ce modèle sur une plateforme pour une utilisation concrète par des parties prenantes éducatives.

# PLAN

## Introduction

---

**01** Présentation des données

**02** Analyse

**03** Prétraitement

**04** Modélisation

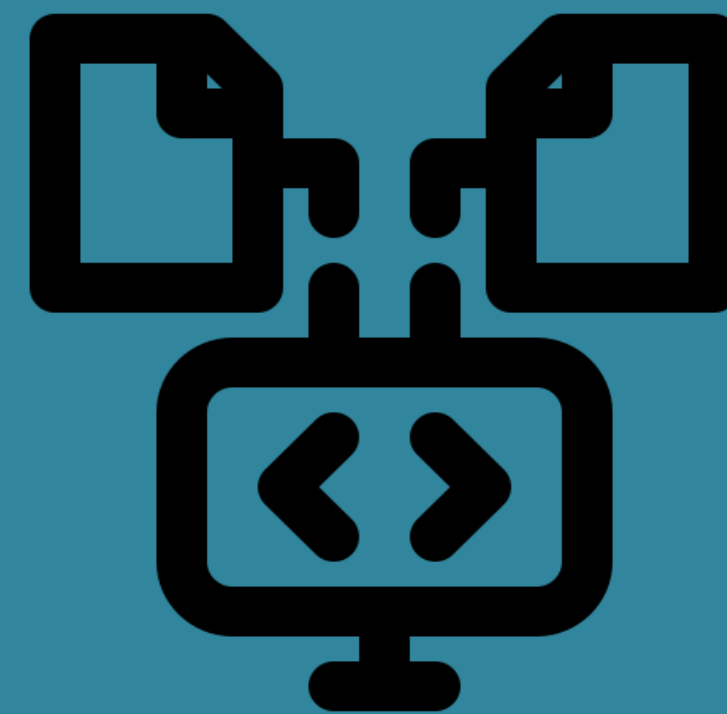
**05** Choix du modèle définitif

---

Déploiement et test de l'application de prédiction

1

# Présentation des données



# PRÉSENTATION DES DONNÉES

## SOURCE

Les données utilisées pour ce projet proviennent de la plateforme *Kaggle*, via le jeu de données intitulé "Student Performance Factors". Ce dataset, partagé par l'utilisateur Lainguyn123, vise à explorer les facteurs influençant la performance académique des étudiants. Il est accessible à l'adresse suivante : [Student Performance Factors](#)

- Nombre d'observations : 6607
- Nombre total de variables : 20
- 19 variables explicatives dont 13 qualitatives et 6 quantitatives
- Variable cible : le score à l'examen (*Exam\_Score*)

### Dataset statistics

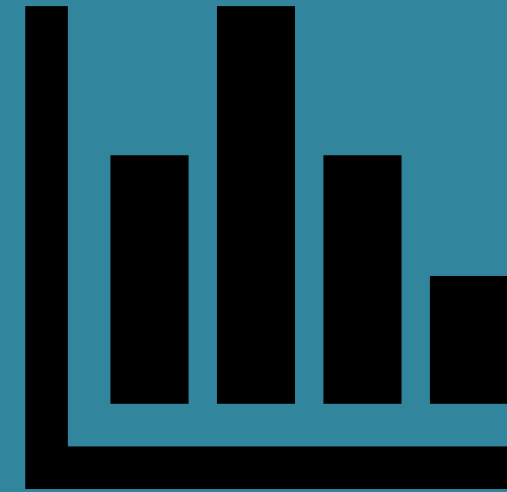
Number of variables	20
Number of observations	6607
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%

### Variable types

Numeric	6
Categorical	11
Boolean	3

**2**

**Analyse**



# Analyse

Exam score

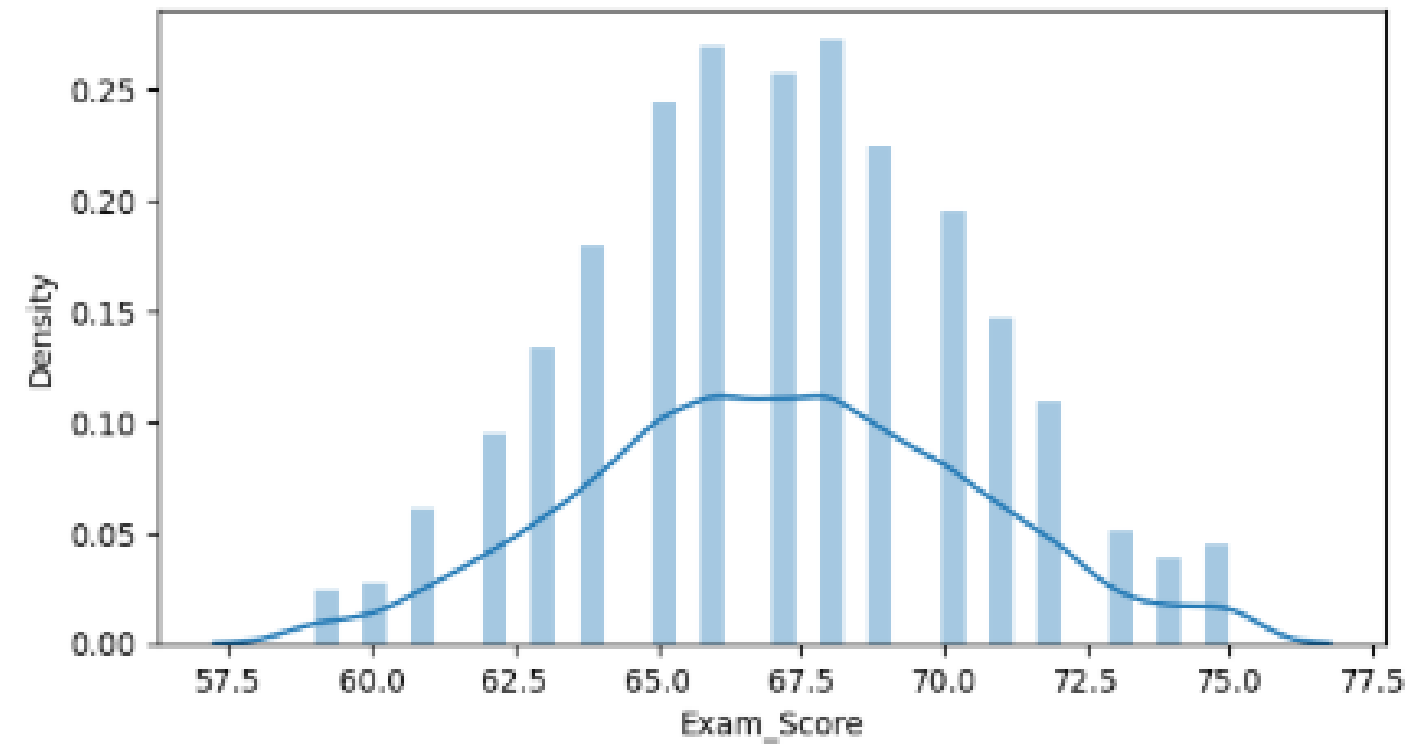


Diagramme circulaire pour Gender

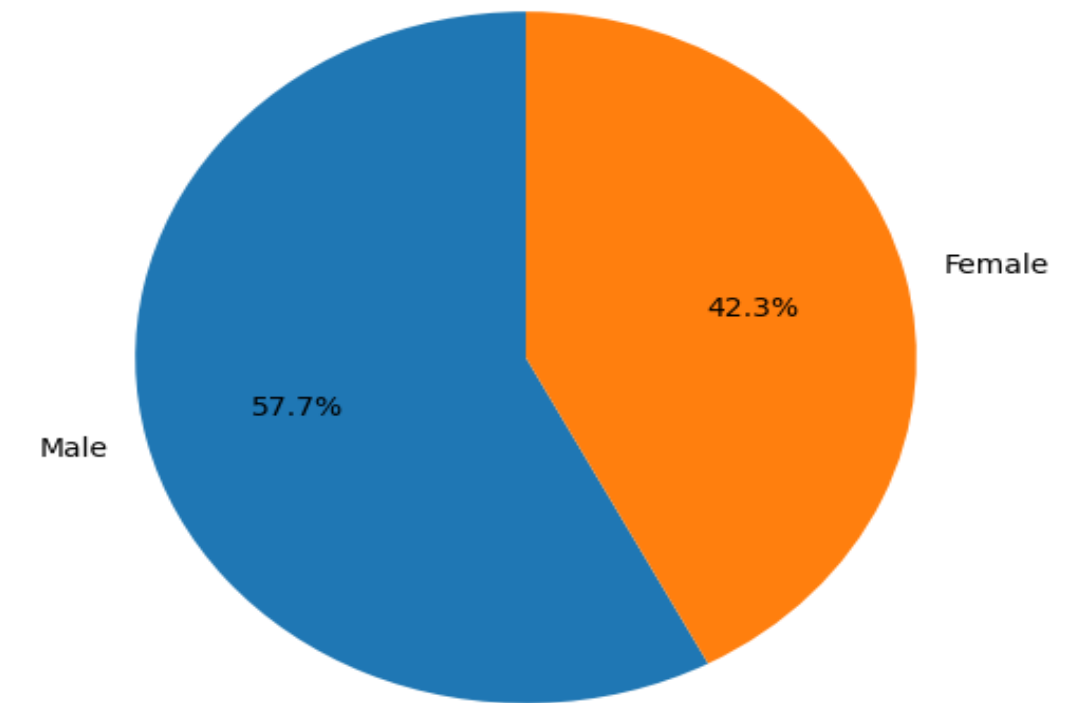


Diagramme en barres pour Parental\_Education\_Level

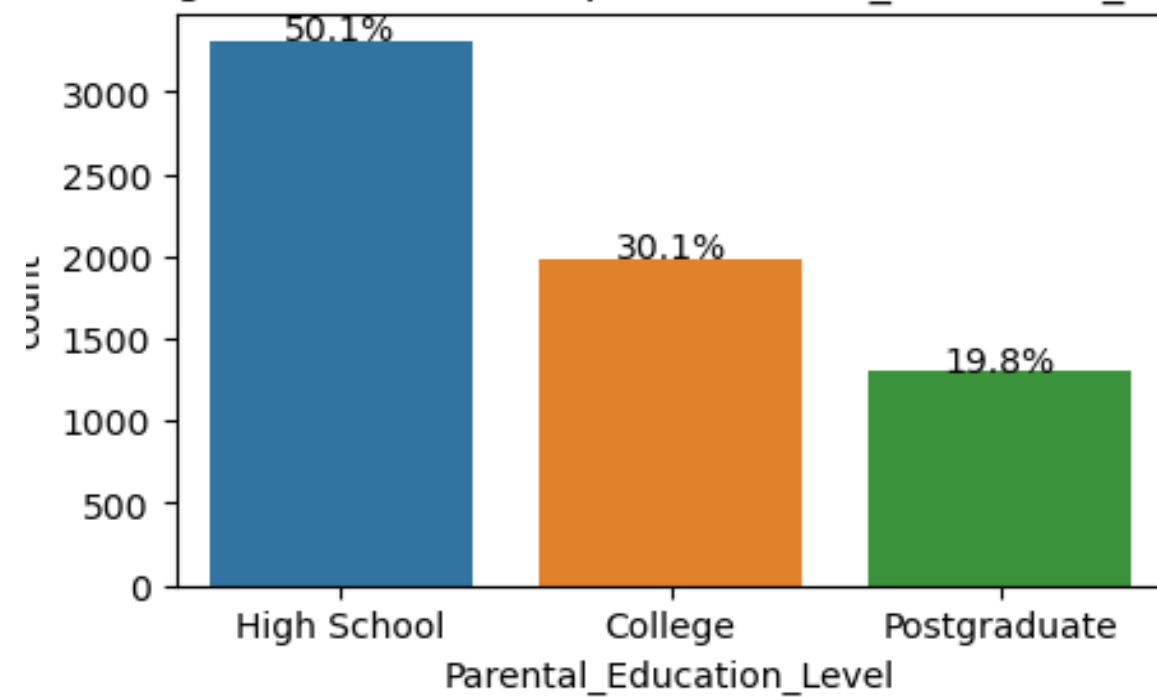
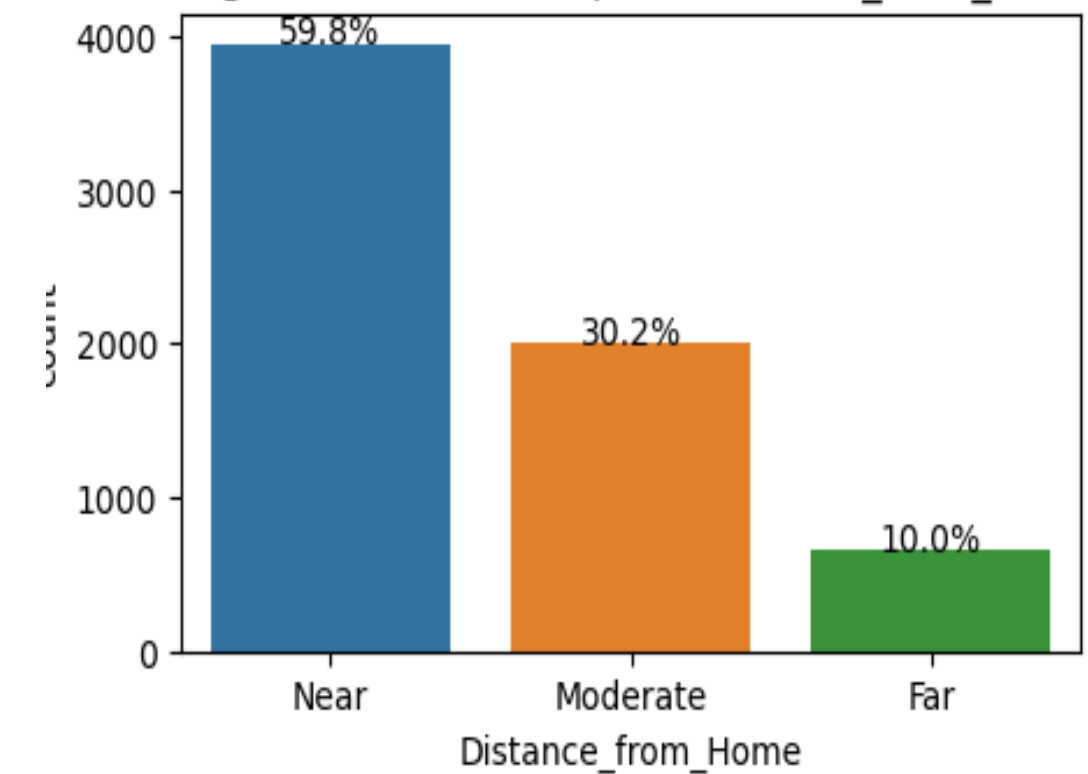


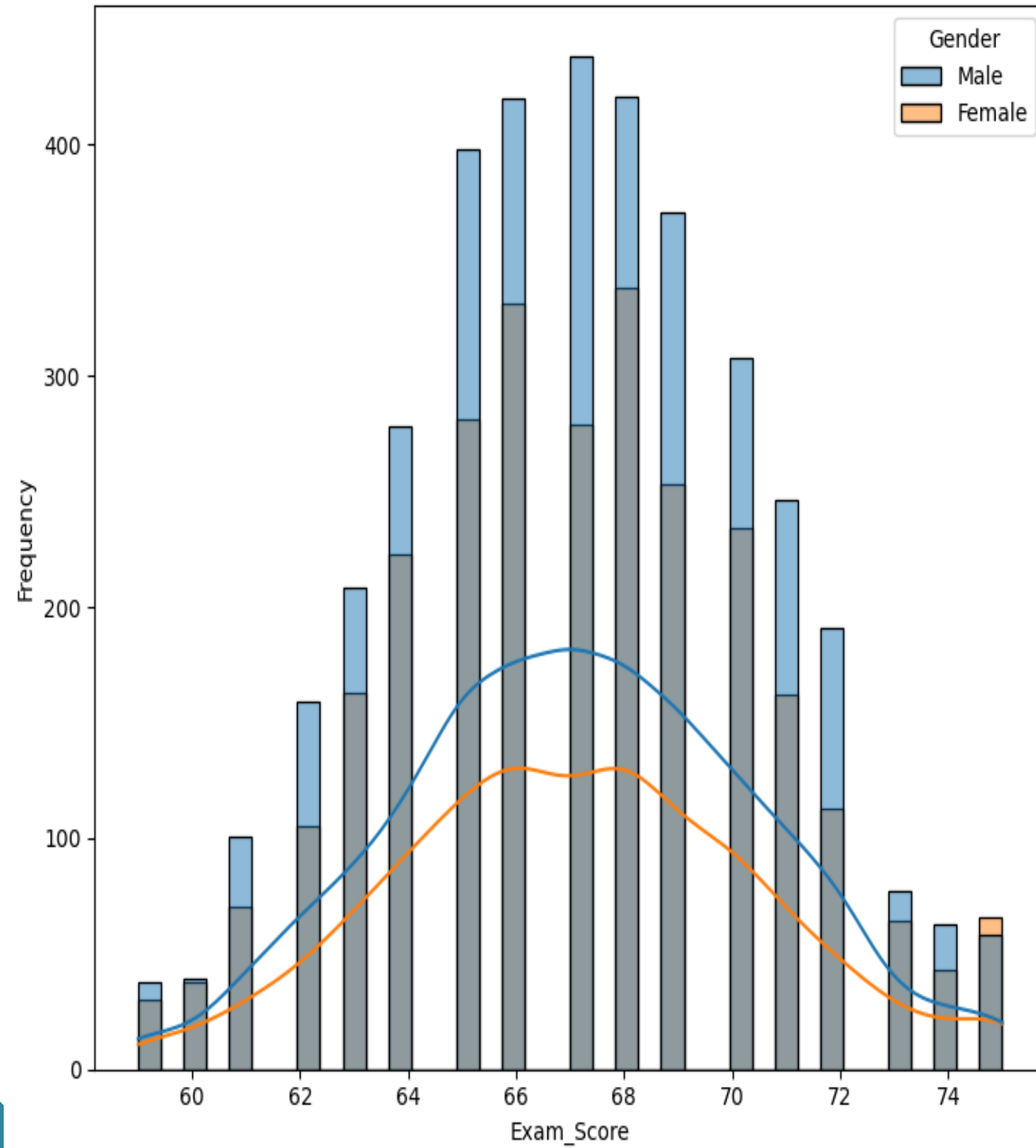
Diagramme en barres pour Distance\_from\_Home



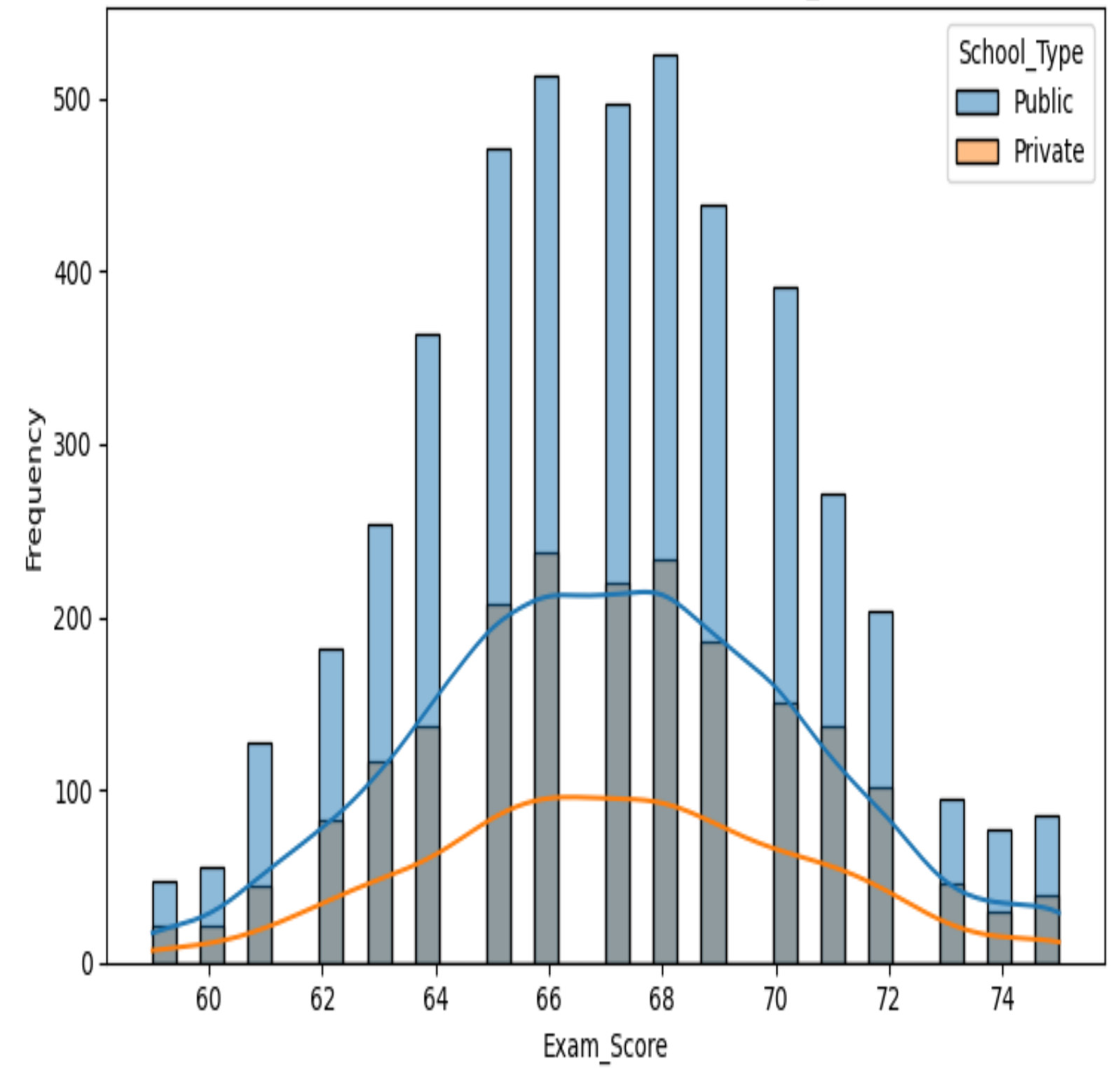


# Analyse

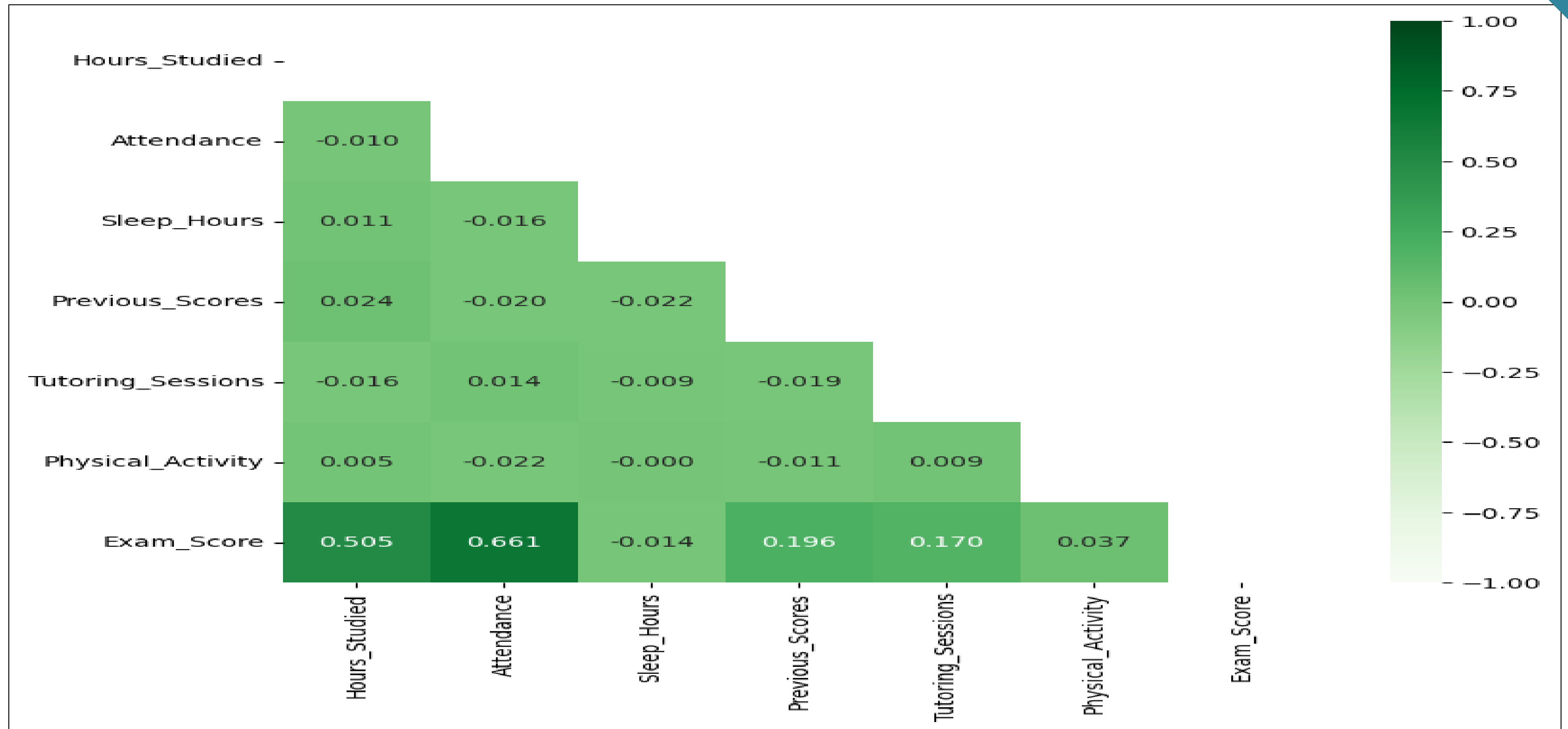
Distribution of exam Scores by Gender



Distribution of exam Scores by School\_Type



# Analyse



3

# Prétraitement



# Prétraitement

## 1. CORRECTION DES VALEURS MANQUANTES

Les variables avec des valeurs manquantes sont :

- *Teacher Quality* : renvoyant à la qualité de l'enseignement (1.18%)
- *Parental Education Level* : renvoyant au niveau d'éducation des parents (1.36%)
- *Distance from Home* : renvoyant à la distance entre la maison et l'école (1.01%)

Ces variables étant toutes qualitatives, nous avons choisi l'amputation par le mode de chacune d'elle.

## 2. CORRECTION DES VALEURS ABERRANTES

Une valeur aberrante, également connue sous le nom d'outlier en statistiques, est une observation qui se situe à une distance significative des autres valeurs dans un ensemble de données. Dans notre jeu de données, les valeurs aberrantes étant au dessus de troisième quartile ont été ramenées au troisième quartile et celles étant en dessous du premier quartile ont été ramenées au premier quartile.

# Prétraitement

## 3. ENCODAGE DES VARIABLES QUALITATIVES

Nos variables qualitatives comportent des variables ordinales et des variables binaires. Dans notre cas, la méthode d'encodage utilisée est le *label encoding*.

Le *Label Encoding* est particulièrement pertinent lorsque les catégories de la variable ont un ordre implicite ou explicite. Par exemple, dans le cas de la variable *Motivation Level* avec les catégories "Low", "Medium", et "High", il y a un ordre qui a été transformé en valeurs numériques comme 1, 2, et 3.

Celles binaires comme par exemple le sexe, ont été transformées en 0, 1.

Ce type d'encodage permet de conserver l'ordre des catégories, ce qui peut être important pour certains modèles qui prennent en compte cet ordre.

4

# Modélisation

$(x)$

# Modélisation

## 1. STANDARDISATION DES DONNÉES

La standardisation consiste à transformer les données de sorte que chaque variable ait une moyenne de 0 et un écart-type de 1. Cela s'effectue en soustrayant la moyenne de chaque variable et en la divisant par son écart-type.

Elle est indispensable car elle nous permet de ramener toutes les données à une même échelle afin de réduire l'impact des valeurs extrêmes et permettre ainsi au modèle de mieux généraliser sur de nouvelles données.

## 2. MODÈLES TESTÉS

- Régression linéaire multiple
- Régression Ridge
- Régression Lasso
- Elastic-Net
- KNN (K-Nearest Neighbors)
- Random Forest
- Gradient Boosting

# Modélisation

## 3. MÉTHODOLOGIE

Pour les différents modèles suscités, la méthodologie suivante a été appliquée :

- Initialisation du modèle
- Tests d'hypothèses si nécessaire
- Optimisation des paramètres avec la validation croisée

## 4. MÉTRIQUES D'ÉVALUATION

- Coefficient de détermination :  $R^2$
- Erreur quadratique moyenne : MSE



**5**

# **Choix du meilleur modèle**



# Choix du meilleur modèle



## 1. PRÉSENTATION DES RÉSULTATS

Les performances des différents modèles sont consignées dans le tableau suivant :

Modèle	CV R <sup>2</sup> Moyen	Train MSE	Test MSE	Train R <sup>2</sup>	Test R <sup>2</sup>	Commentaires
Ridge	0.9033	0.0955	0.0917	0.9045	0.9103	Très bonnes performances en validation croisée (R <sup>2</sup> élevé) et sur les données de test.
Lasso	0.8985	0.1003	0.0965	0.8997	0.9056	Performances légèrement inférieures à Ridge en validation croisée et sur le test, mais reste fiable.
Elastic Net	0.9029	0.0959	0.0922	0.9041	0.9098	Comparable à Ridge en validation croisée, mais Ridge est légèrement supérieur en généralisation.
KNN	-	0.1690	0.2410	0.8310	0.7642	Performances faibles sur le test (R <sup>2</sup> très bas).
Random Forest	0.8222	0.0233	0.1741	0.9767	0.8296	La validation croisée montre des performances acceptables, mais le modèle a surappris.
Boosting	0.8754	0.0917	0.1200	0.9083	0.8826	Résultats en validation croisée instables (R <sup>2</sup> moyen faible), mais bon R <sup>2</sup> sur le test.

# Choix du meilleur modèle

## 2. MEILLEUR MODÈLE

Le modèle Ridge se distingue par ses :

- Excellentes performances en validation croisée ( $R^2 = 0.9033$ ).
- Très bons résultats sur le test ( $R^2 = 0.9103$ ), garantissant ainsi une généralisation optimale.

Par suite, le modèle retenu est **la régression Ridge**.





# Déploiement et test de l'application de prédiction



Introduction



## Navigation

L'application est structurée autour de trois principales sections :

- **Introduction** : Un aperçu des données et du projet.
- **Modélisation** : Utilisation de modèles de régression pour prédire les performances des étudiants.
- **Visualisation** : Graphiques interactifs

Deploy



# Student Performance Analysis

Bienvenue dans l'application d'analyse des performances des étudiants !

Cette application vous permet d'explorer différentes techniques d'analyse, y compris l'introduction des données, la modélisation, et la visualisation des résultats.

Vous pouvez naviguer entre les différentes pages pour découvrir les insights cachés dans les données.

## Analyse et Prédiction des Performances Académiques

Exploiter le Machine Learning pour mieux comprendre et prédire la réussite éducative

### Contexte et Justification

La réussite éducative est un enjeu central des sociétés modernes. Dans un monde en constante évolution, il devient crucial de comprendre les facteurs influençant la performance académique des étudiants.

Prédiction

Navigation

L'application est structurée autour de trois principales sections :

- **Introduction** : Un aperçu des données et du projet.
- **Modélisation** : Utilisation de modèles de régression pour prédire les performances des étudiants.
- **Visualisation** : Graphiques interactifs pour explorer les résultats.
- **Prédiction** : Modèles de machine learning pour prédire les scores d'examen en fonction des données.

 Student Performance Data

Fork  

Troubles d'apprentissage

Non

Niveau d'éducation des parents

Aucun

Distance domicile-école (en km)

5

Genre

Masculin

Prédire

Le score prédit pour cet étudiant est : 70.66







# CONCLUSION



Comme dit beaucoup plus haut, l'éducation revêt d'une importance particulière. Notre travail consistait à prédire grâce aux outils de machine learning, le score à l'examen d'un apprenant en tenant en compte d'un certain nombre de ces caractéristiques notamment le nombre d'heurs d'études, le niveau d'éducation des parents, l'assiduité, etc. Par suite, les résultats nous ont permis de choisir comme meilleur modèle la régression Ridge car généralisant mieux sur les données. Ainsi, cet outil pourra servir de manière efficace quant au ciblage et à la bonne application des différentes politiques mises en place par les autorités éducatives.



**MERCI POUR VOTRE AIMABLE ATTENTION**







# BIBLIOGRAPHIE



- Cours du professeur
- Kaggle
- Scikit-learn