Jedha

Data Essentials - dse-ft-101

Groupe “Room 4” - Facteurs de réussite scolaire

Group members

- Asma RHALMI
- Anthony GIACOBI
- Thomas DIMEK
- Albert ROMANO
- Olivier CHARDAC

Pitch

Problématique du client :

Les habitudes de vie d’un étudiant ont-elles un impact sur sa note d’examen final ?

Objectif du projet :

Est-il possible de prédire le score final d’un étudiant à l’examen à partir de ses habitudes de vie ?

EDA - Statistics

df.**info** :

- Lignes/colonnes : 1000/16
- 91 valeurs manquantes dans “parental_education_level”

```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   student_id                      1000 non-null   object
 1   age                             1000 non-null   int64
 2   gender                          1000 non-null   object
 3   study_hours_per_day             1000 non-null   float64
 4   social_media_hours              1000 non-null   float64
 5   netflix_hours                   1000 non-null   float64
 6   part_time_job                   1000 non-null   object
 7   attendance_percentage           1000 non-null   float64
 8   sleep_hours                     1000 non-null   float64
 9   diet_quality                    1000 non-null   object
 10  exercise_frequency              1000 non-null   int64
 11  parental_education_level        909 non-null    object
 12  internet_quality                1000 non-null   object
 13  mental_health_rating            1000 non-null   int64
 14  extracurricular_participation   1000 non-null   object
 15  exam_score                      1000 non-null   float64
dtypes: float64(6), int64(3), object(7)
```

EDA - Valeurs aberrantes (3 sigma)

- Outliers détectés : 8/1000 lignes
- Négligeables en tant que valeurs

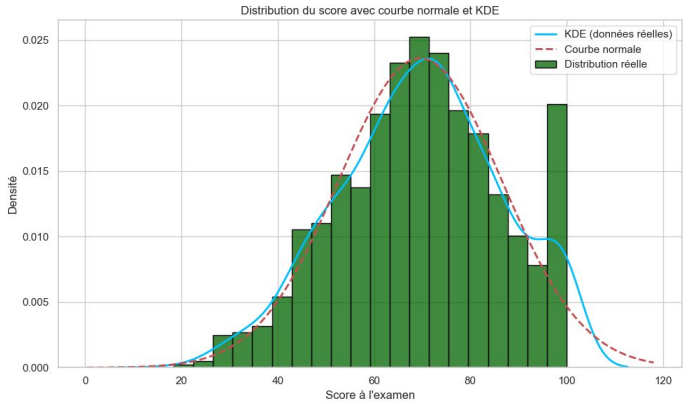
```
Colonne : 'study_hours_per_day'
Indexes des outliers : [455, 797]
Valeurs des outliers : [8.3, 8.2]

Colonne : 'social_media_hours'
Indexes des outliers : [145, 361, 735]
Valeurs des outliers : [6.2, 6.1, 7.2]

Colonne : 'netflix_hours'
Indexes des outliers : [556, 822]
Valeurs des outliers : [5.4, 5.3]

Colonne : 'exam_score'
Indexes des outliers : [265]
Valeurs des outliers : [18.4]
```

EDA - Analyse de la distribution

- Méthode Shapiro
  - La méthode de Shapiro sert à tester si une variable suit une distribution normale (loi normale)
  - Distribution non “normale” des numériques

	age	study_hours_per_day	attendance_percentage	sleep_hours	exercise_frequency	mental_health_rating	exam_score	Media_hours
0	9.248605e-01	0.997378	9.826074e-01	0.997267	9.139217e-01	9.381751e-01	9.869195e-01	0.997326
1	6.177718e-22	0.106471	1.502940e-09	0.088776	2.263751e-23	5.841297e-20	8.675028e-08	0.097718

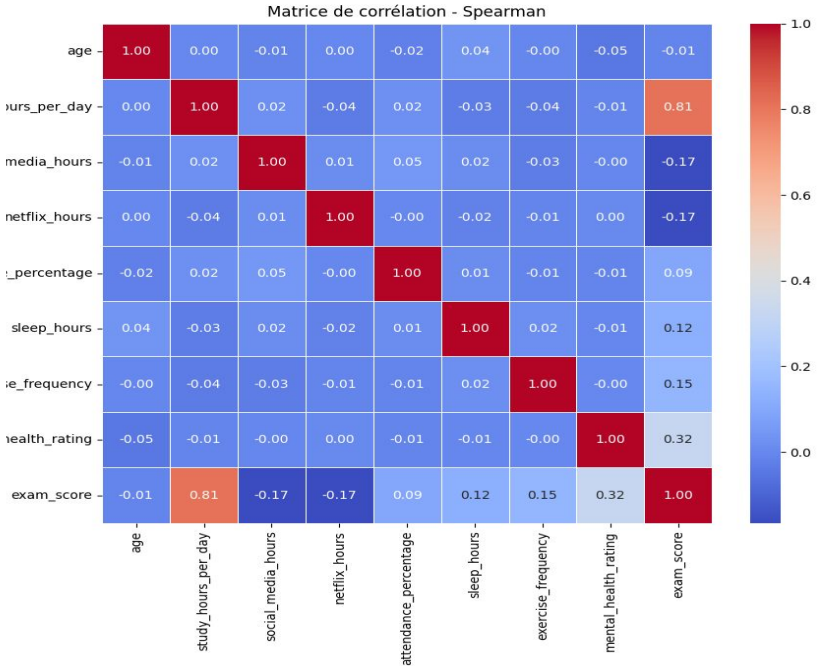
- Graphique



EDA - Analyse de la corrélation

Test des variables numériques

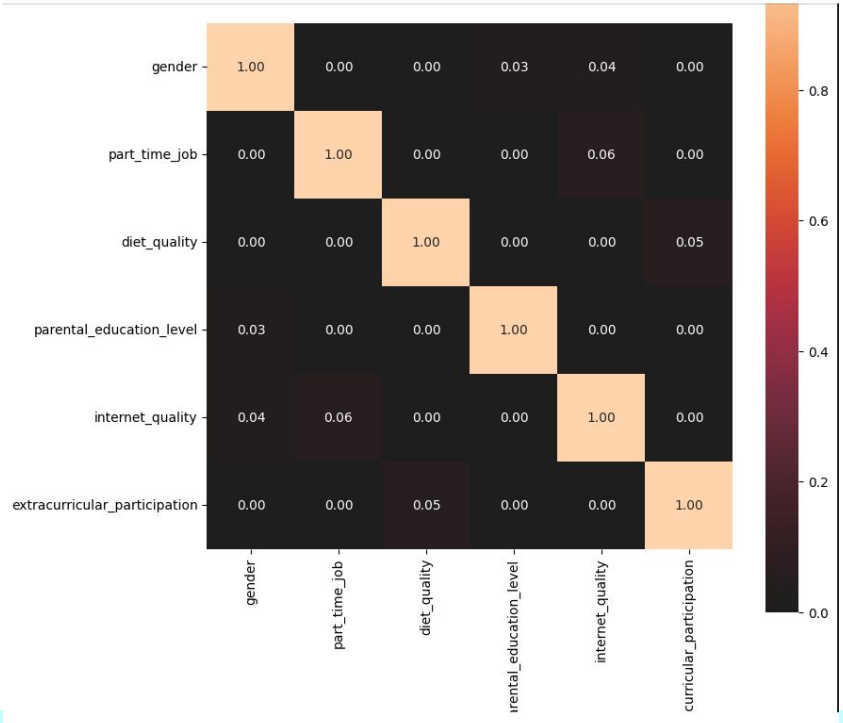
- Test de Spearman et Kendall
- Retenu : Spearman



EDA - Analyse de la corrélation

Test des variables catégorielles

- Test de Dython



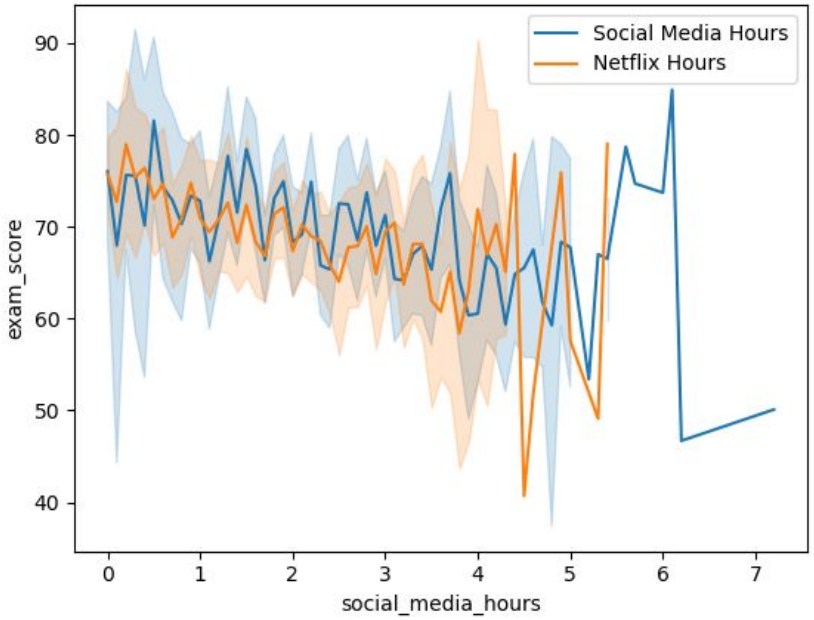
EDA - Test de la colinéarité numérique

Test avec Variable Inflation Factor :

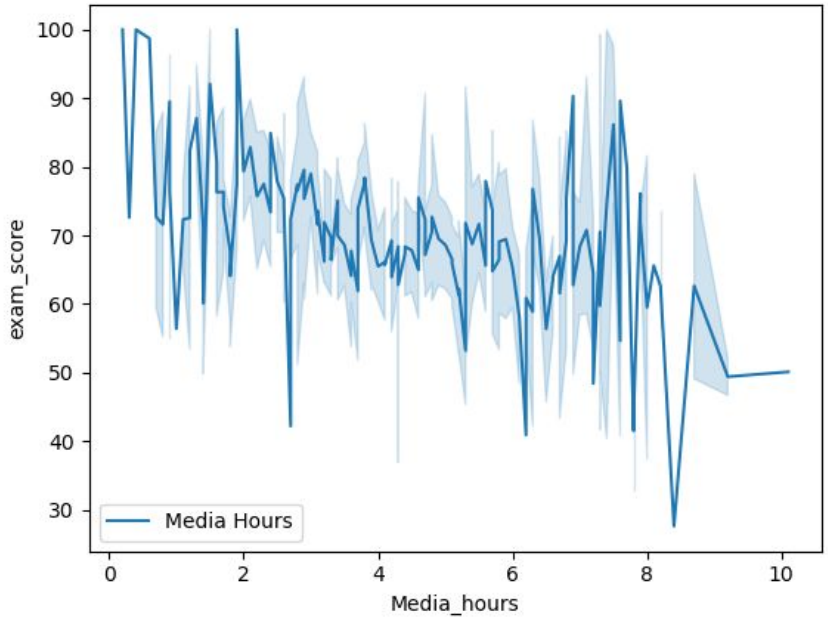
	Variable	VIF
0	const	207.930307
1	age	1.004321
2	study_hours_per_day	1.003660
3	social_media_hours	1.003940
4	netflix_hours	1.001235
5	attendance_percentage	1.003613
6	sleep_hours	1.003199
7	exercise_frequency	1.002675
8	mental_health_rating	1.002534

EDA - Impact des médias sur le score à l'examen

Superposition de Netflix et réseaux

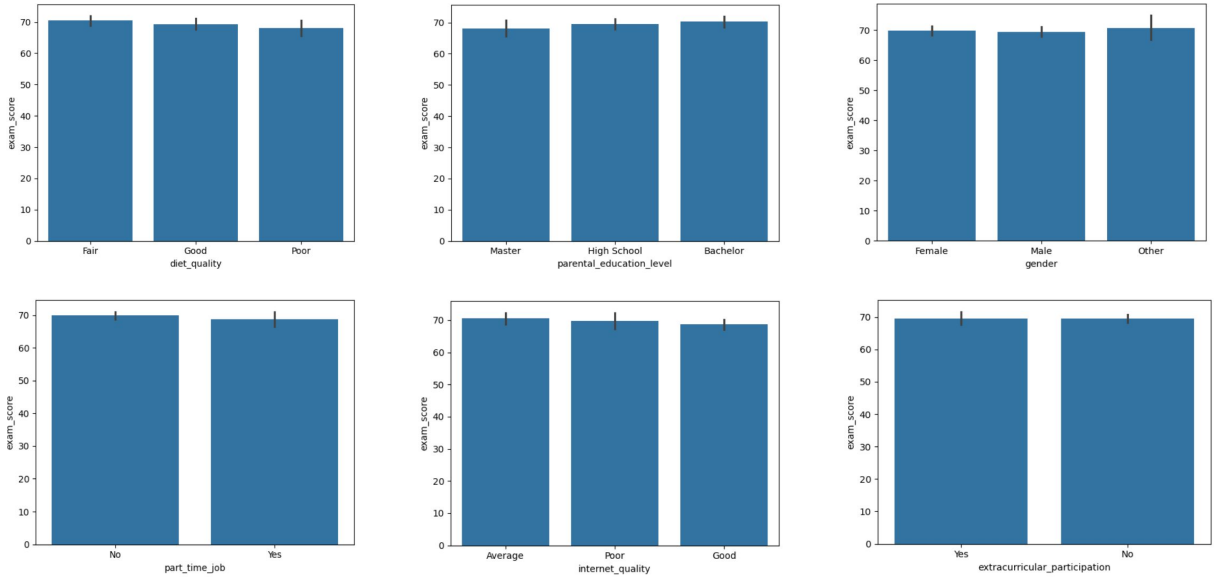


Fusion de Netflix et réseaux

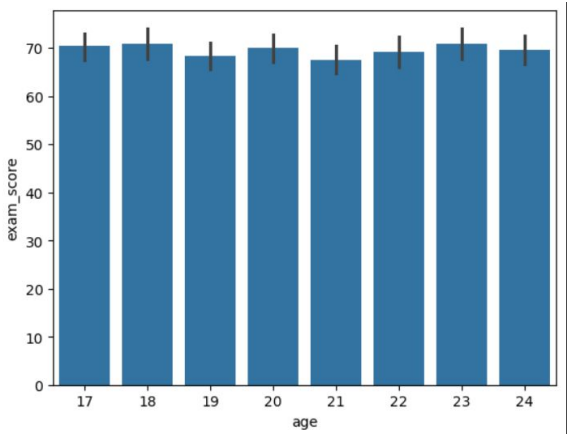


EDA - Variables non retenues

- Variables Catégorielles



- Variables Numériques



EDA - Data Frame

- Jeu de donnée final en vue de l'entraînement à la prédiction
  - cible : “exam_score”

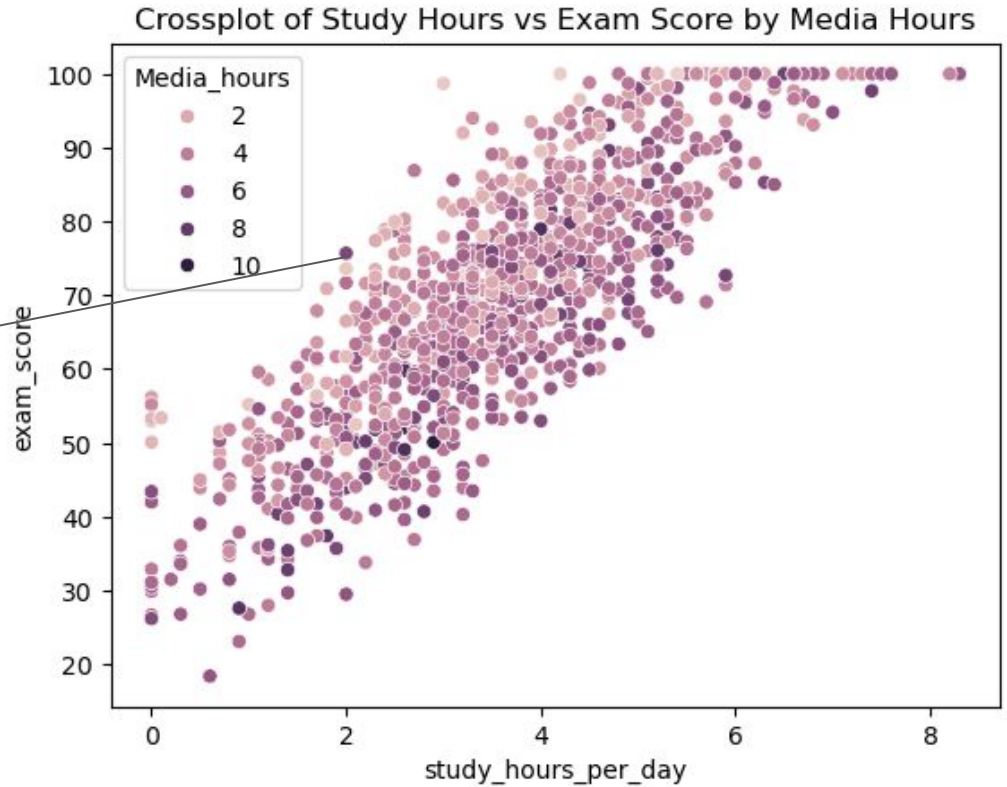
```
study_hours_per_day
attendance_percentage
sleep_hours
exercise_frequency
mental_health_rating
Media_hours
```

- Jeu de donnée à prédire

```
exam_score
```

EDA - Exemple

Jérôme passe 8h sur les médias et ne révise que 2h mais arrive à un score de 75/100 à l'examen



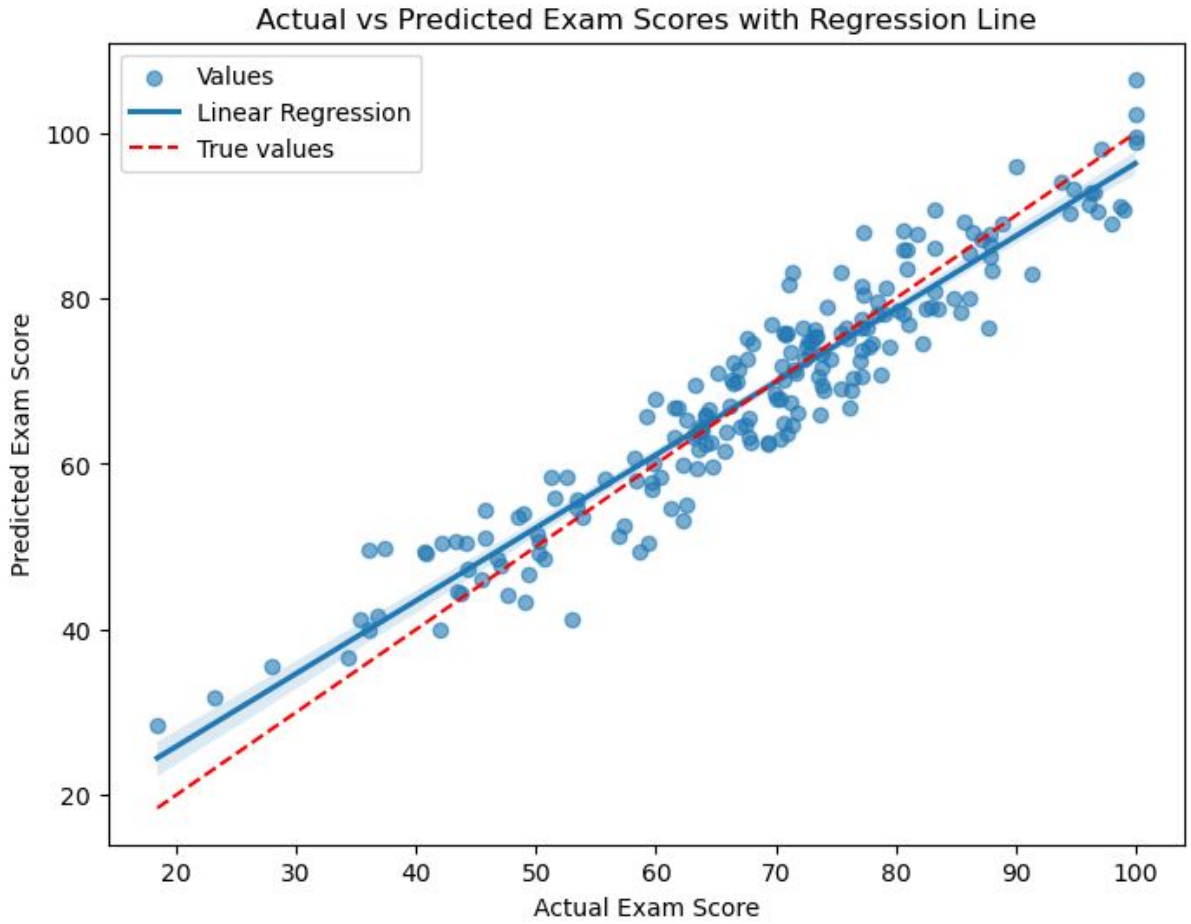
ML - Modèle 1

- Régression Linéaire

- Score d’apprentissage : 0.9006

  Score de test : 0.8996

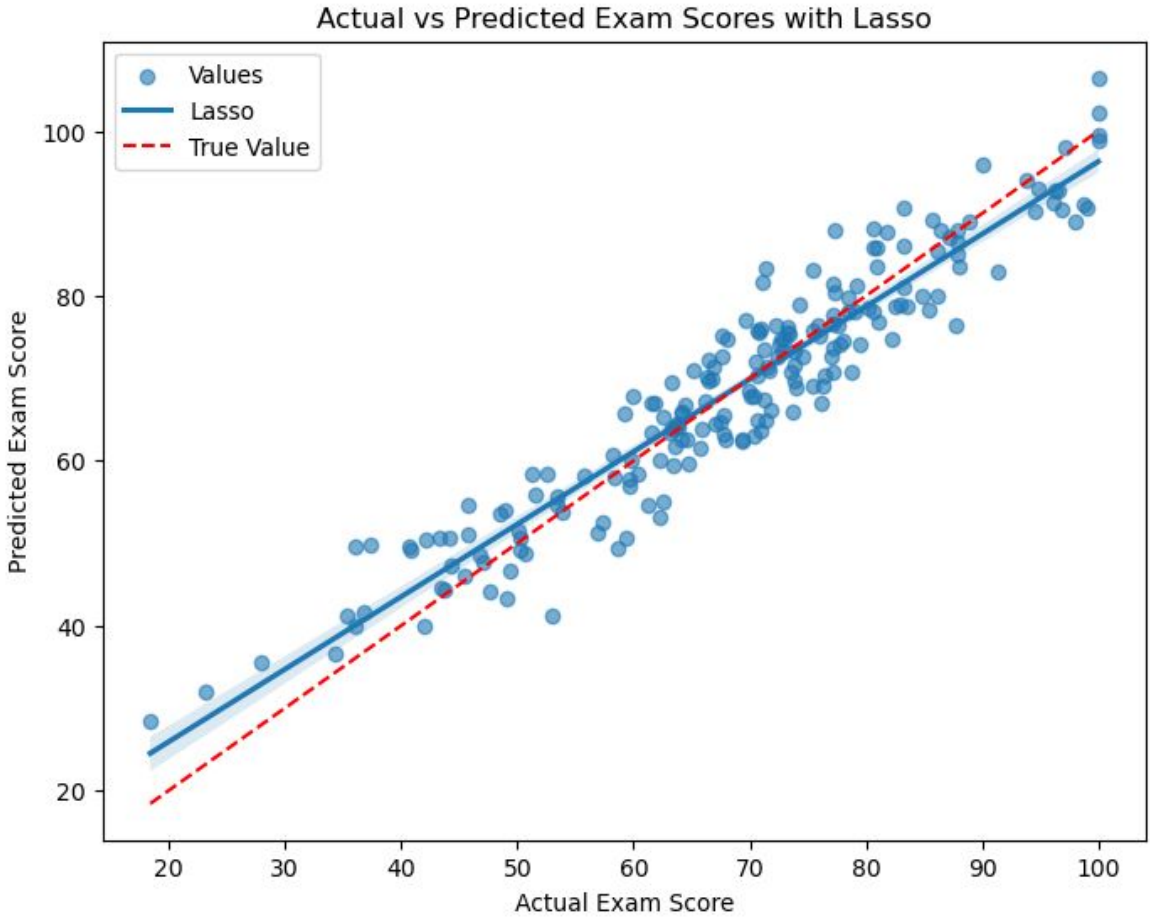
- Pourcentage d’erreur : 6.811%



ML - Modèle 2

- Lasso
- Score d’apprentissage : 0.9006

  Score de test : 0.8995
- Pourcentage d’erreur : 6.818%



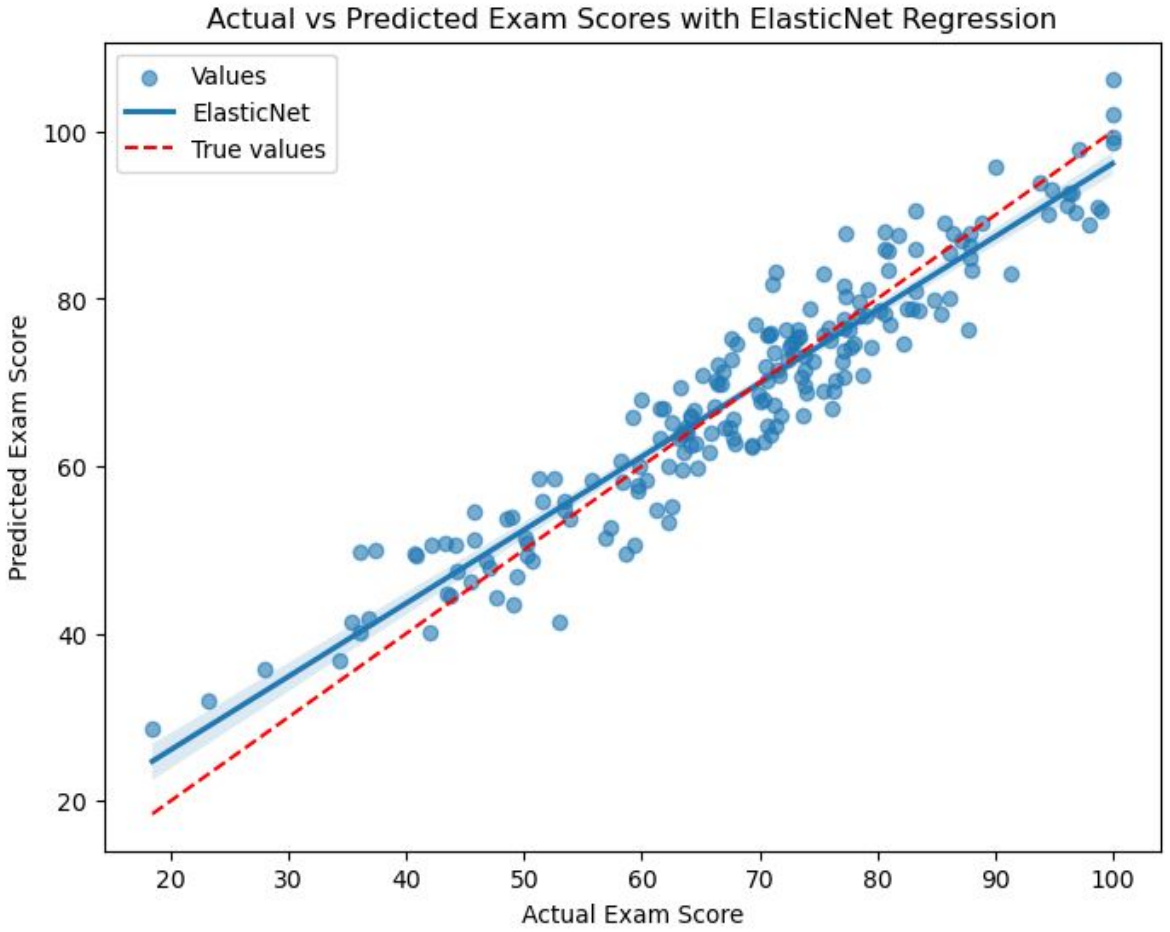
ML - Modèle 3

- Elastic net

- Score d’apprentissage : 0.9005

  Score de test : 0.8993

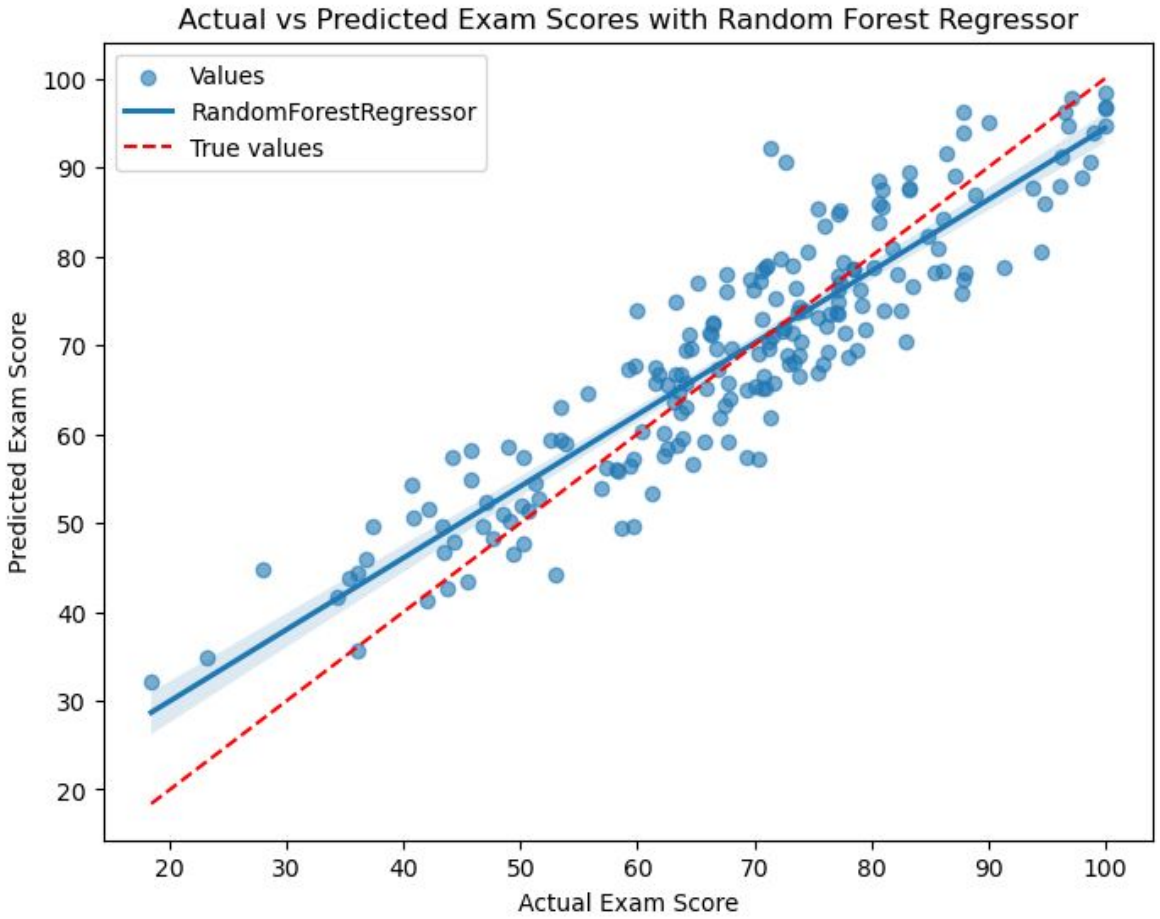
- Pourcentage d’erreur : 6.840%



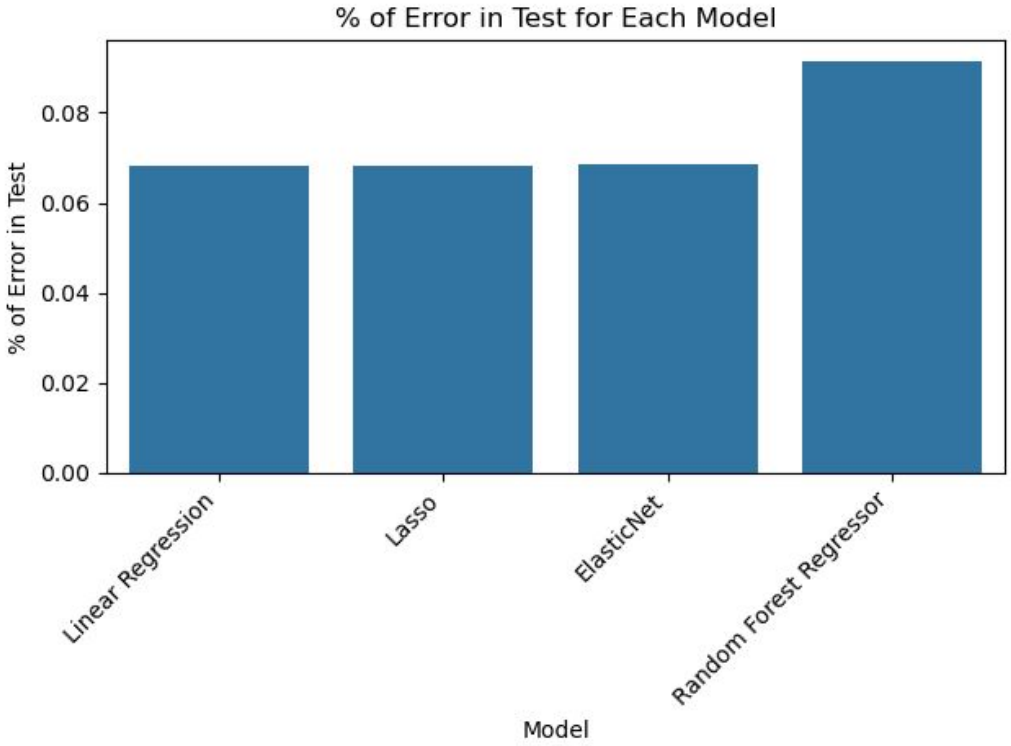
ML - Modèle 4

- RandomForestRegressor
- Score d’apprentissage : 0.8901

  Score de test : 0.8187
- Pourcentage d’erreur : 9.141%



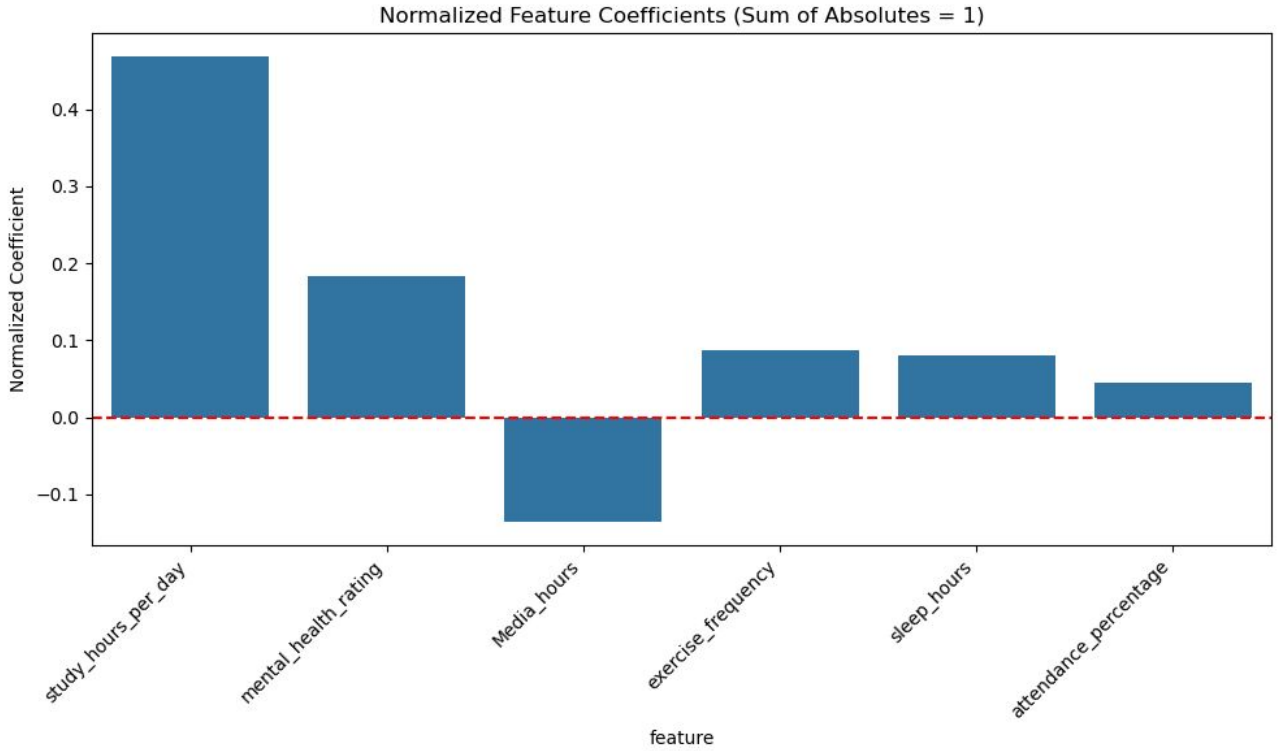
ML - Comparaison des modèles



- Les 3 premiers modèles ont une performance équivalente

- Le dernier est légèrement moins performant

ML - Importance des facteurs

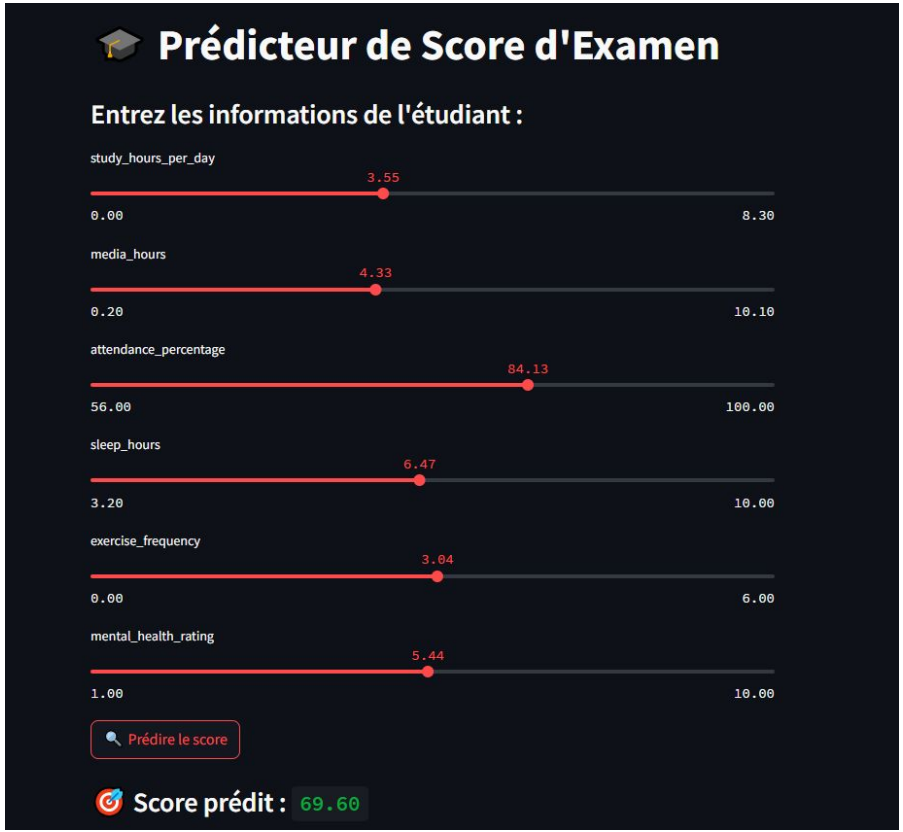


What’s next ?

Exemple de produit final fonctionnel :

[Lien 1]

[Lien 2]



What’s next ?

- Jeu de données semblant synthétique

- À tester sur un jeu de données réel

Jedha

Any questions ?