DỰ ĐOÁN KẾT QUẢ HỌC TẬP CỦA HỌC SINH

(Linear Regression, Lasso Regression, Decision Tree, Random forest)

Trình bày bởi: Đoàn Thị Thu Linh

GVHD: PGS.TS Nguyễn Văn Hậu

BỐ CỤC

1. Tổng quan về đề tài
2. Khám phá dữ liệu
3. Tiền xử lý dữ liệu
4. Mô hình học máy
5. Đánh giá mô hình
6. Tổng kết

1 TỔNG QUAN VỀ ĐỀ TÀI

- Bài toán được xây dựng dưới dạng bài toán hồi quy (regression), với mục tiêu dự đoán điểm cuối kỳ của học sinh môn Toán tại thời điểm kết thúc năm học.

- Link Kaggle: https://www.kaggle.com/datasets/rxshark/uci-student-performance-dataset-by-paulo-cortez/code

- Bộ dữ liệu gồm:
  - Số lượng mẫu: 395 học sinh
  - Số lượng đặc trưng: 33 đặc trưng đầu vào
  - Biến mục tiêu: G3_ Điểm cuối kỳ môn Toán

2 KHÁM PHÁ DỮ LIỆU

index	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no	yes
1	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no	no
2	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3	yes	no	yes	no	yes
3	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes	yes	yes	yes
4	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes	yes	no	yes
5	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2	0	no	yes	yes	yes	yes
6	GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	no	no	no	no	yes
7	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0	yes	yes	no	no	yes
8	GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	0	no	yes	yes	no	yes
9	GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2	0	no	yes	yes	yes	yes
10	GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother	1	2	0	no	yes	yes	no	yes
11	GP	F	15	U	GT3	T	2	1	services	other	reputation	father	3	3	0	no	yes	no	yes	yes
12	GP	M	15	U	LE3	T	4	4	health	services	course	father	1	1	0	no	yes	yes	yes	yes
13	GP	M	15	U	GT3	T	4	3	teacher	other	course	mother	2	2	0	no	yes	yes	no	yes
14	GP	M	15	U	GT3	A	2	2	other	other	home	other	1	3	0	no	yes	no	no	yes
15	GP	F	16	U	GT3	T	4	4	health	other	home	mother	1	1	0	no	yes	no	no	yes
16	GP	F	16	U	GT3	T	4	4	services	services	reputation	mother	1	3	0	no	yes	yes	yes	yes
17	GP	F	16	U	GT3	T	3	3	other	other	reputation	mother	3	2	0	yes	yes	no	yes	yes
18	GP	M	17	U	GT3	T	3	2	services	services	course	mother	1	1	3	no	yes	no	yes	yes
19	GP	M	16	U	LE3	T	4	3	health	other	home	father	1	1	0	no	no	yes	yes	yes

1. Thông tin của học sinh
- age: Tuổi
- sex: Giới tính
- address: Nơi cư trú (thành thị / nông thôn)
- Pstatus: Tình trạng gia đình
- …..

2. Thông tin gia đình
- Medu: Trình độ học vấn của mẹ
- Fedu: Trình độ học vấn của cha
- Mjob: Nghề nghiệp của mẹ
- Fjob: Nghề nghiệp của cha
- famsup: Mức độ hỗ trợ học tập từ gia đình
- …..

3. Thông tin liên quan đến học tập và hành vi
- studytime: Thời gian học tập hàng tuần
- failures: Số lần trượt môn trước đó
- freetime: Thời gian rảnh sau giờ học
- goout: Mức độ đi chơi với bạn bè
- health: Tình trạng sức khỏe hiện tại
- …..

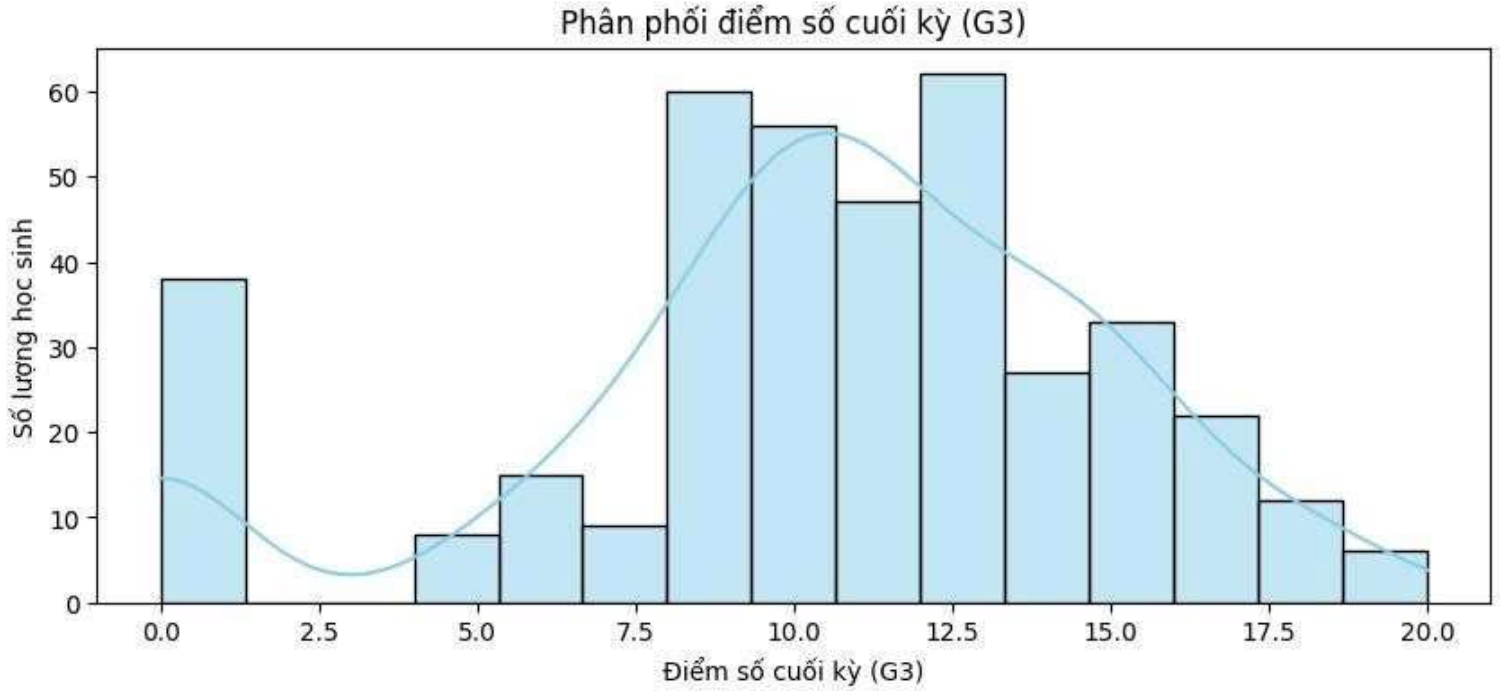
4. Kết quả học tập trong năm học
- G1: Điểm học kỳ thứ nhất
- G2: Điểm học kỳ thứ hai
- absences: Số buổi nghỉ học

2 KHÁM PHÁ DỮ LIỆU

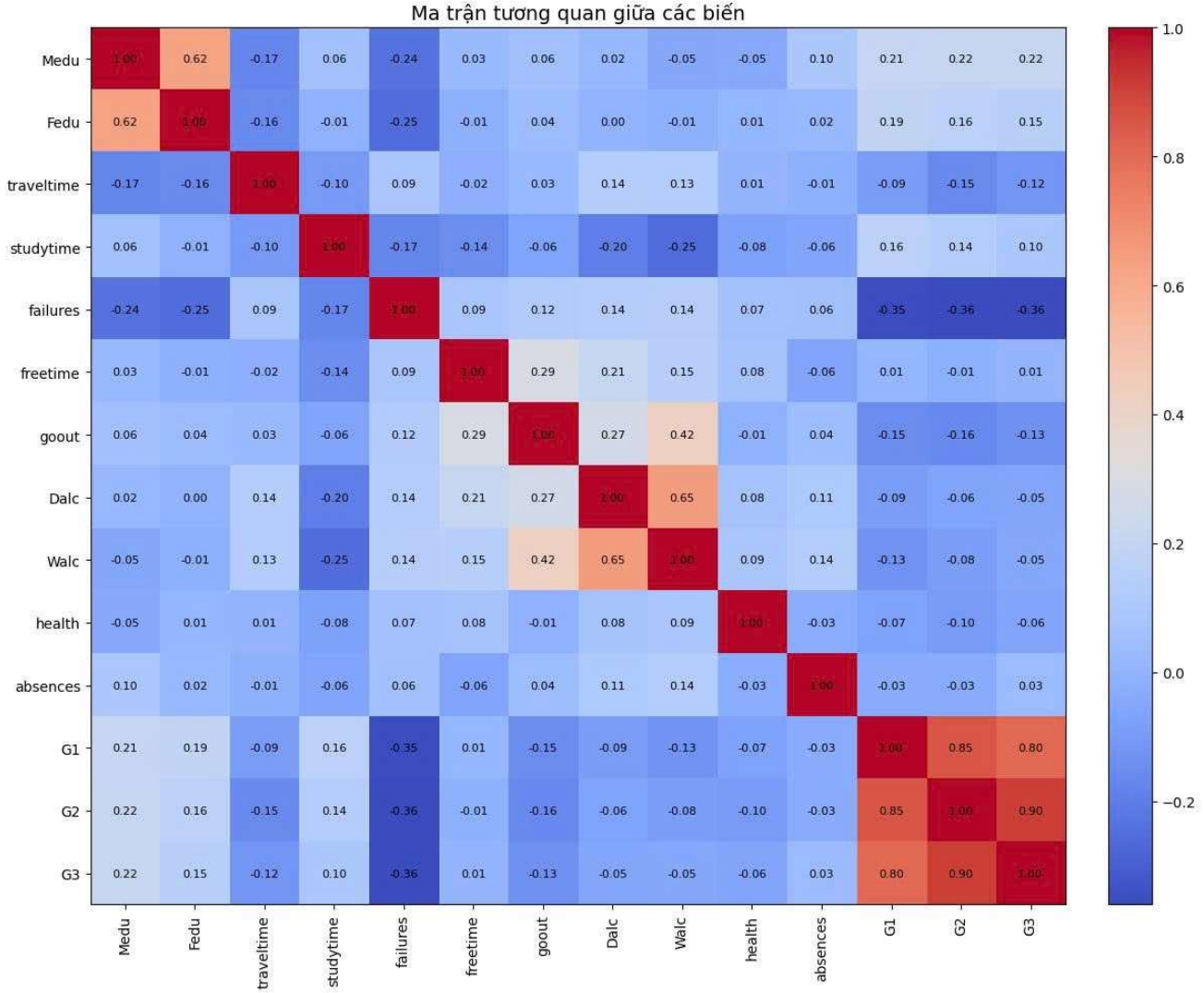
- df.head()
- df.info()
- df.describe()
- ….

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000
mean	16.696203	2.749367	2.521519	1.448101	2.035443	0.334177	3.944304	3.235443	3.108861	1.481013	2.291139	3.554430	5.708861	10.908861	10.713924	10.415190
std	1.276043	1.094735	1.088201	0.697505	0.839240	0.743651	0.896659	0.998862	1.113278	0.890741	1.287897	1.390303	8.003096	3.319195	3.761505	4.581443
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	3.000000	0.000000	0.000000
25%	16.000000	2.000000	2.000000	1.000000	1.000000	0.000000	4.000000	3.000000	2.000000	1.000000	1.000000	3.000000	0.000000	8.000000	9.000000	8.000000
50%	17.000000	3.000000	2.000000	1.000000	2.000000	0.000000	4.000000	3.000000	3.000000	1.000000	2.000000	4.000000	4.000000	11.000000	11.000000	11.000000
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000	5.000000	4.000000	4.000000	2.000000	3.000000	5.000000	8.000000	13.000000	13.000000	14.000000
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	75.000000	19.000000	19.000000	20.000000

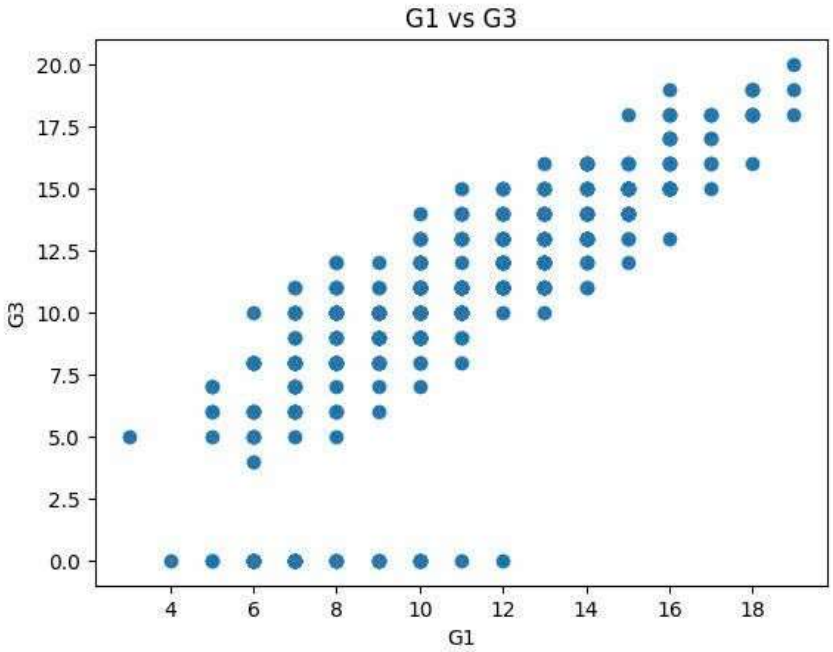
2 KHÁM PHÁ DỮ LIỆU

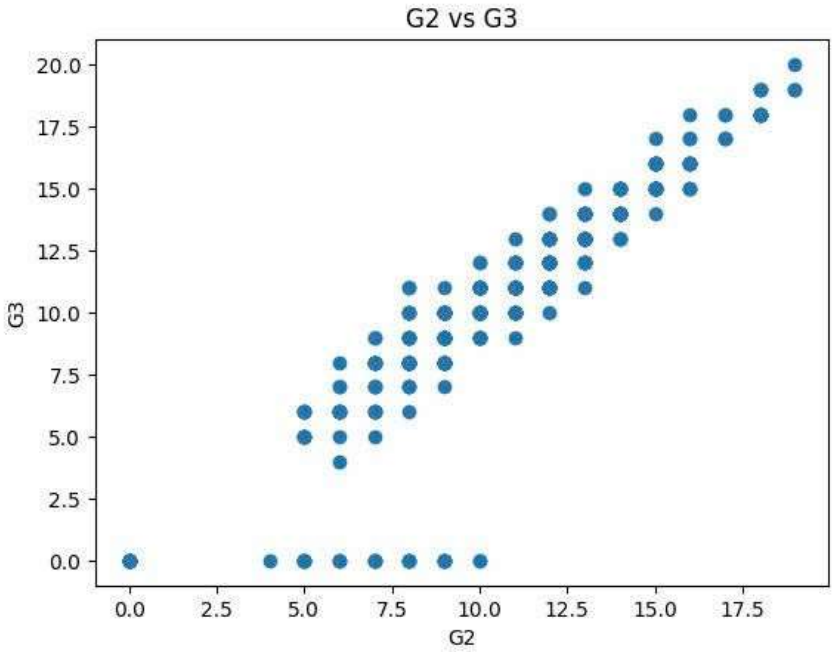


2 KHÁM PHÁ DỮ LIỆU

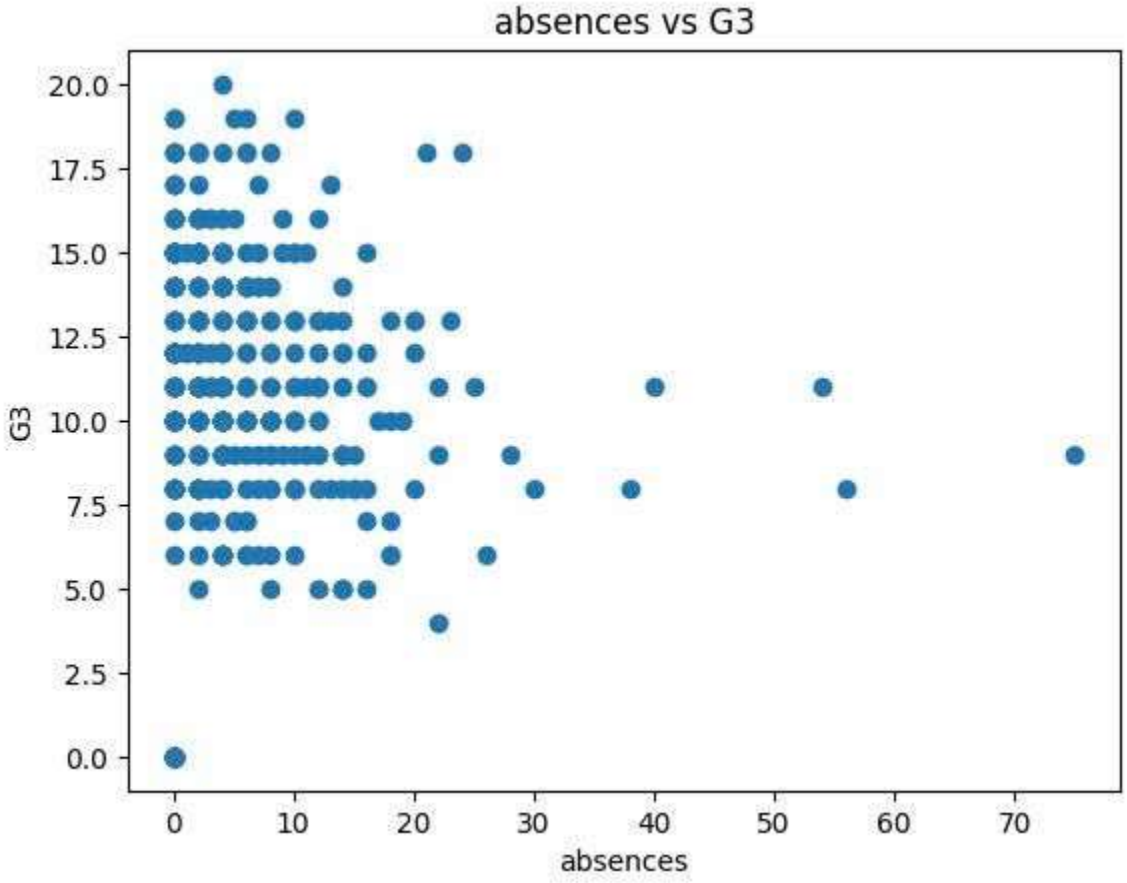


2 KHÁM PHÁ DỮ LIỆU

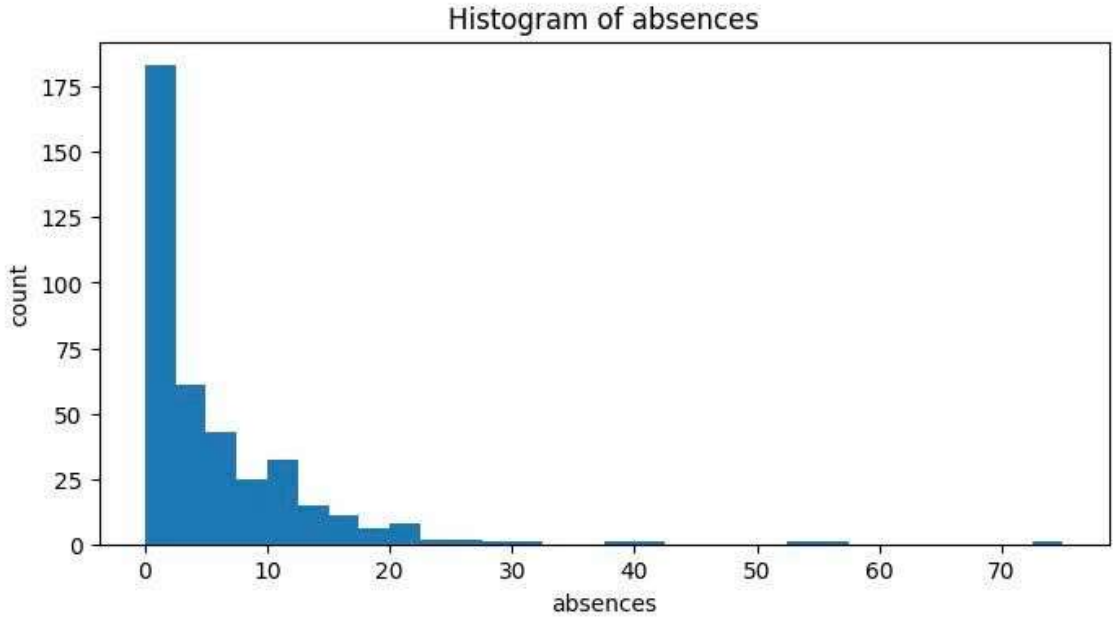


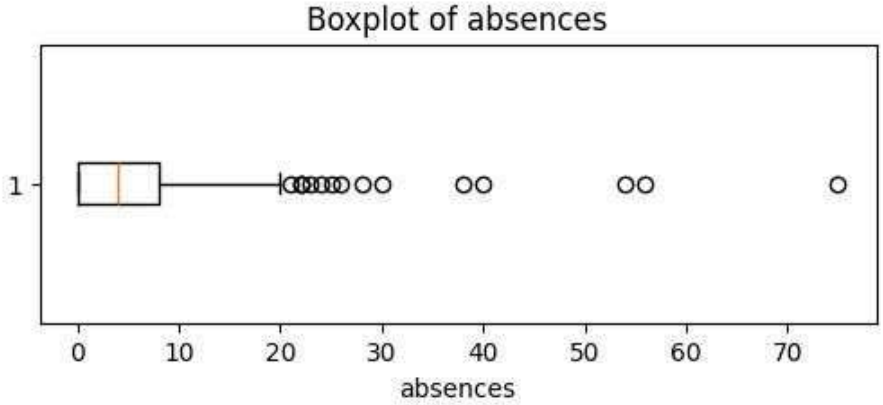


2 KHÁM PHÁ DỮ LIỆU

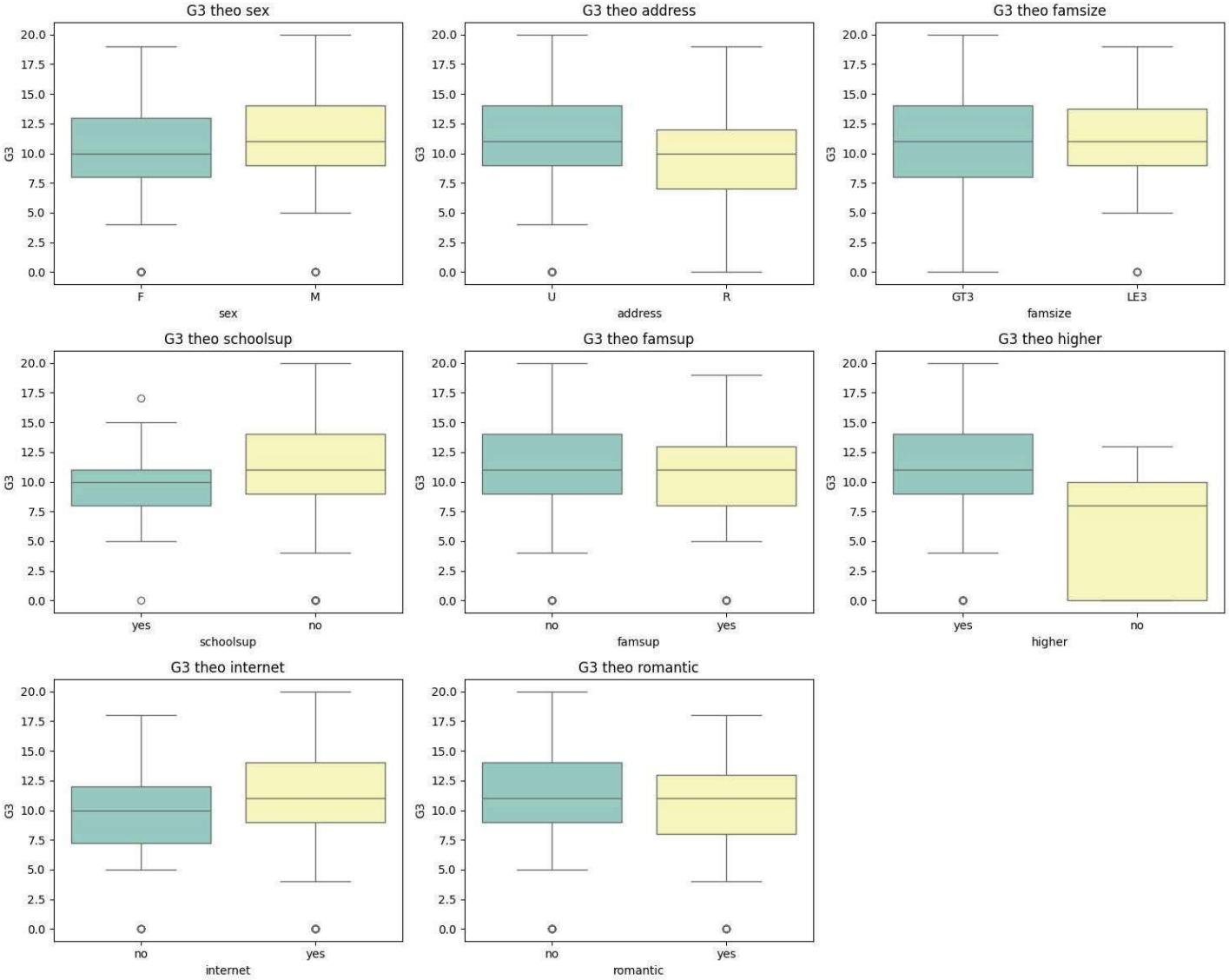


2 KHÁM PHÁ DỮ LIỆU

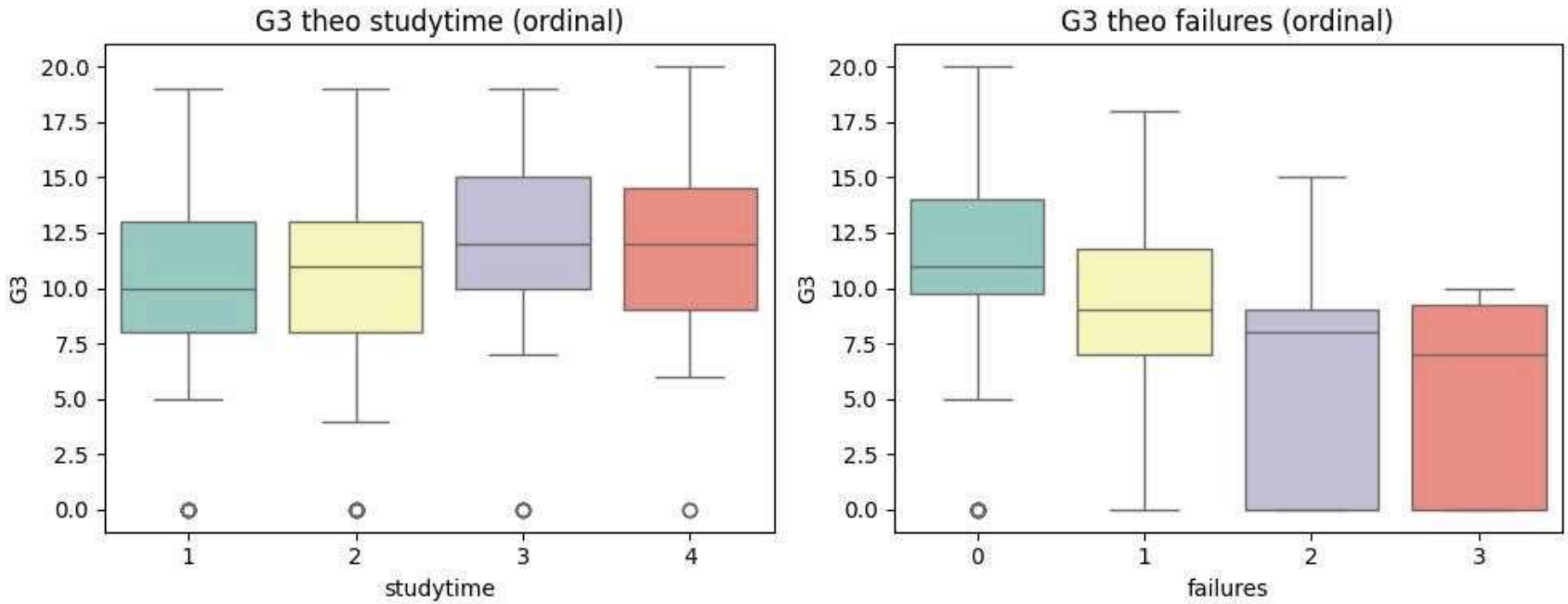




2 KHÁM PHÁ DỮ LIỆU



2 KHÁM PHÁ DỮ LIỆU



3 TIỀN XỬ LÝ DỮ LIỆU

- Bước 1: Kiểm tra và làm sạch dữ liệu

```
print(df.isnull().sum())
```

```
sex           0
address       0
famsize       0
Medu          0
Fedu          0
Mjob          0
Fjob          0
guardian      0
traveltime    0
studytime     0
failures      0
schoolsup     0
famsup        0
paid          0
activities    0
higher        0
internet      0
romantic      0
freetime      0
goout         0
Dalc          0
Walc          0
health        0
absences      0
G1            0
G2            0
G3            0
dtype: int64
```

3 TIỀN XỬ LÝ DỮ LIỆU

- Bước 2: Phân loại các biến theo kiểu dữ liệu:
  - Xử lý biến số (Numeric features): StandardScaler
  - Xử lý biến phân loại (Categorical features): One-Hot Encoding

3 TIỀN XỬ LÝ DỮ LIỆU

- Bước 3: Xử lý outlier
  - Biến absences có phân phối lệch phải và chứa các giá trị ngoại lai lớn.
  - Áp dụng phép biến đổi:

$$absences\_log = log(1 + absences)$$

- Bước 4: Tách biến input và biến target
  - Mô hình A:
    - (X): toàn bộ các đặc trưng (gồm cả G1,G2) ngoại trừ G3.
    - (y): G3 – điểm cuối kỳ.
  - Mô hình B:
    - (X): toàn bộ các đặc trưng ngoại trừ G1,G2, G3.
    - (y): G3 – điểm cuối kỳ.

3 TIỀN XỬ LÝ DỮ LIỆU

- Bước 5. Chia tập dữ liệu huấn luyện và kiểm tra
  - 80% tập huấn luyện (training set)
  - 20% tập kiểm tra (test set)

```
print("Train:", X_train_t.shape)
print("Test:", X_test_t.shape)
```

```
Train: (316, 46)
Test: (79, 46)
```

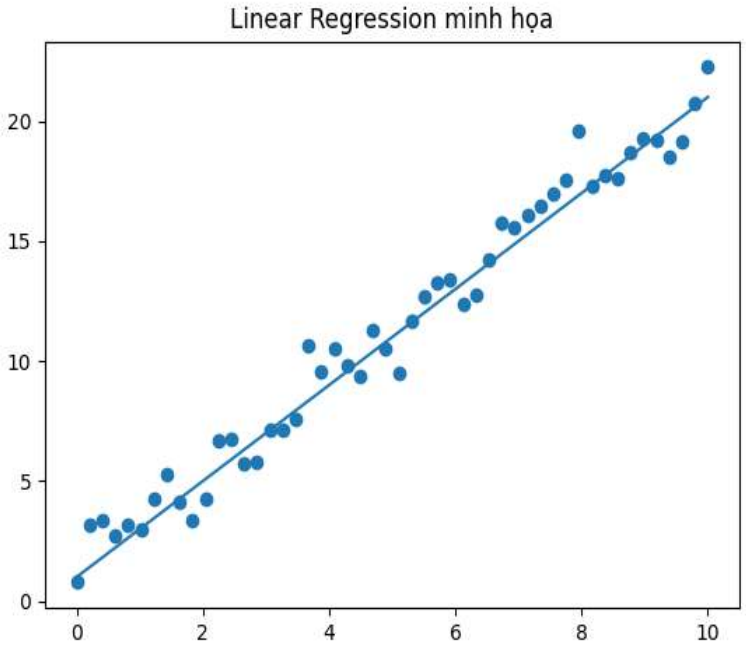
```
print("Train (B1):", X_train_b1_t.shape)
print("Test  (B1):", X_test_b1_t.shape)
```

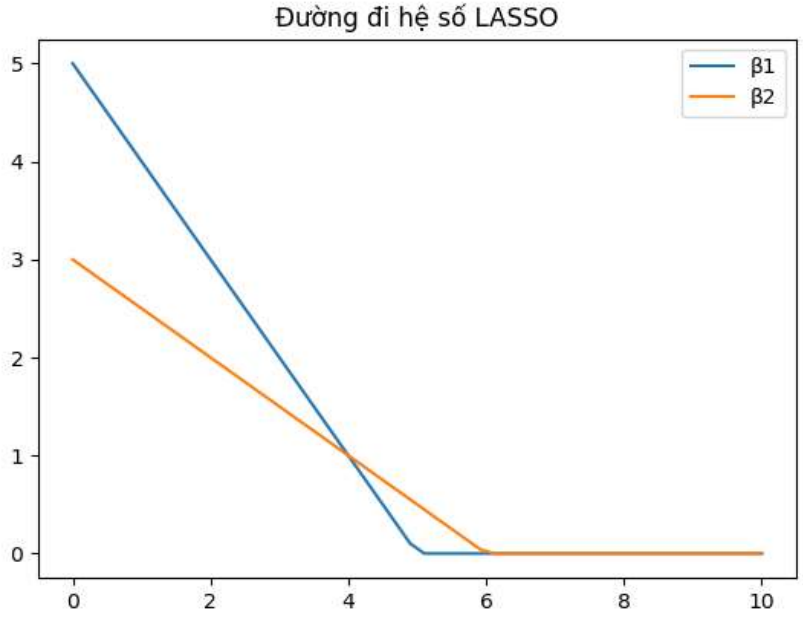
```
Train (B1): (316, 44)
Test  (B1): (79, 44)
```

4 MÔ HÌNH HỌC MÁY

- **Linear Regression**: là mô hình học máy có giám sát dùng để mô hình hóa mối quan hệ tuyến tính giữa biến độc lập X và biến phụ thuộc y.

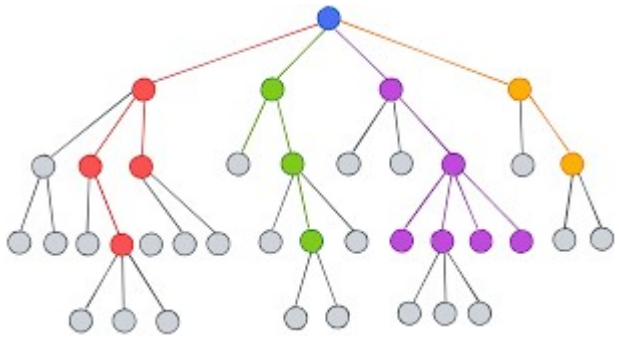
- **Lasso Regression**: Linear Regression kết hợp điều chuẩn L1. Mô hình này sử dụng phương pháp co rút (shrinkage), là quá trình thu nhỏ các giá trị dữ liệu về phía một điểm trung tâm làm giá trị trung bình.





Decision Tree

- Nguyên lý:
  - Mô hình chia dữ liệu thành các nhánh dựa trên đặc trưng quan trọng, gọi là “nút”
  - Mỗi nút kiểm tra một điều kiện, phân loại hoặc dự đoán kết quả cho dữ liệu.
  - Quá trình này lặp lại cho đến khi đạt điều kiện dừng (lá cây).



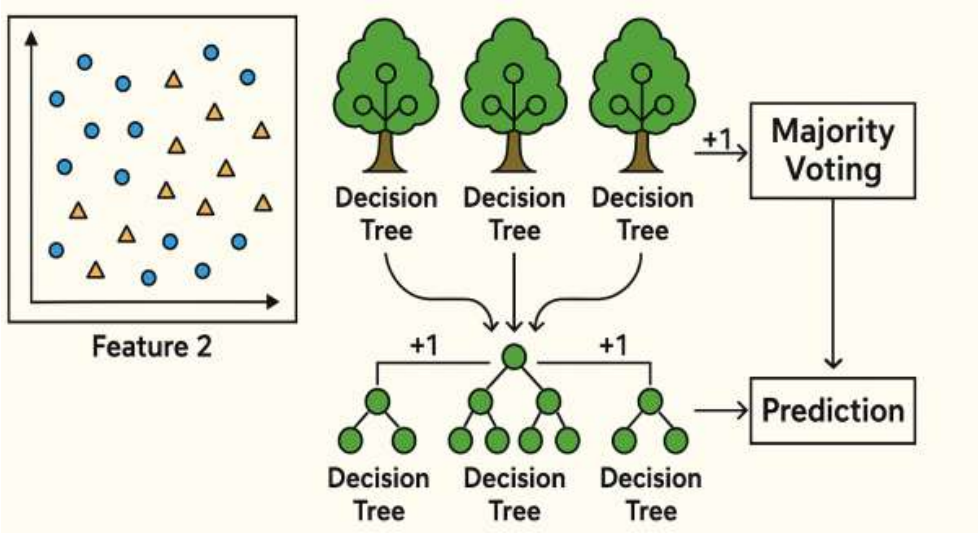
=> Dễ overfitting nếu cây quá sâu.

Ensemble Learning

- Khái niệm: Kết hợp nhiều mô hình cơ bản (base models) để tạo ra mô hình mạnh hơn, giảm bias hoặc variance.
- Các nhóm chính:
  - Bagging – Giảm variance
  - Boosting – Giảm bias
  - Stacking – Giảm bias, dùng meta-model để học cách kết hợp kết quả các base models.

Random Forest

- Nguyên lý:
  - Tạo nhiều cây quyết định từ các tập con dữ liệu và tập con đặc trưng
  - Kết quả dự đoán là trung bình hoặc đa số của các cây.
  - Giảm nguy cơ overfitting so với cây đơn lẻ.



- Ưu điểm:
  - Dự đoán ổn định, chính xác.
  - Giảm overfitting

5 ĐÁNH GIÁ MÔ HÌNH

- Các chỉ số đánh giá mô hình:

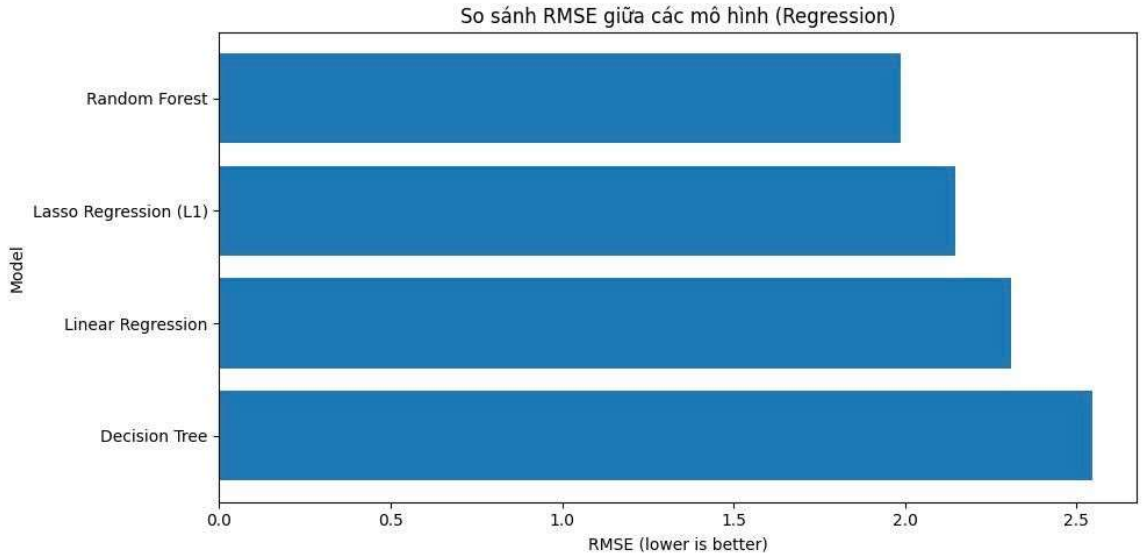
$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

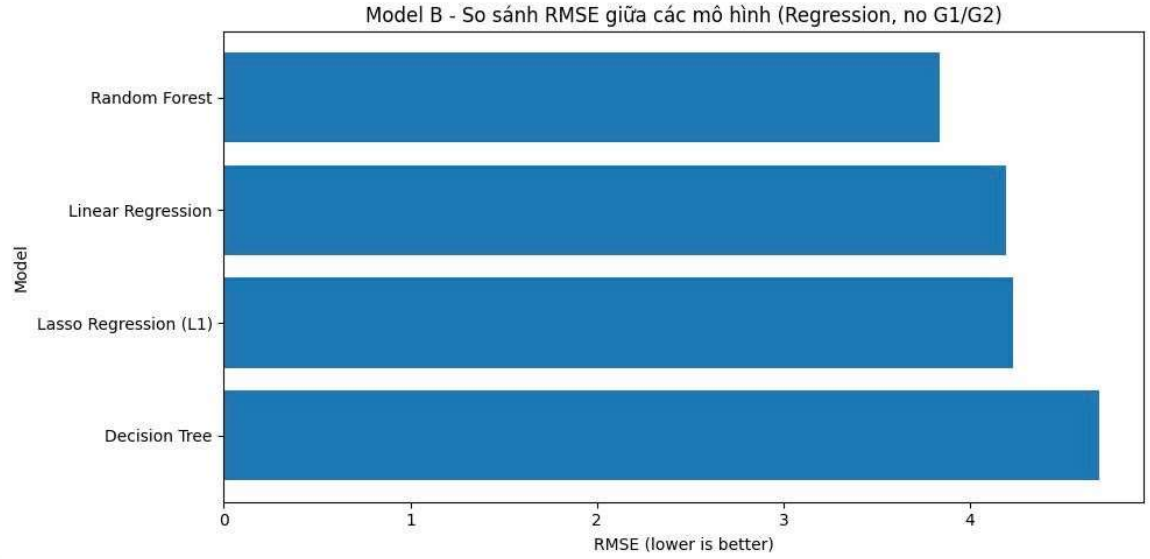
$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

5 ĐÁNH GIÁ MÔ HÌNH

- Mô hình A

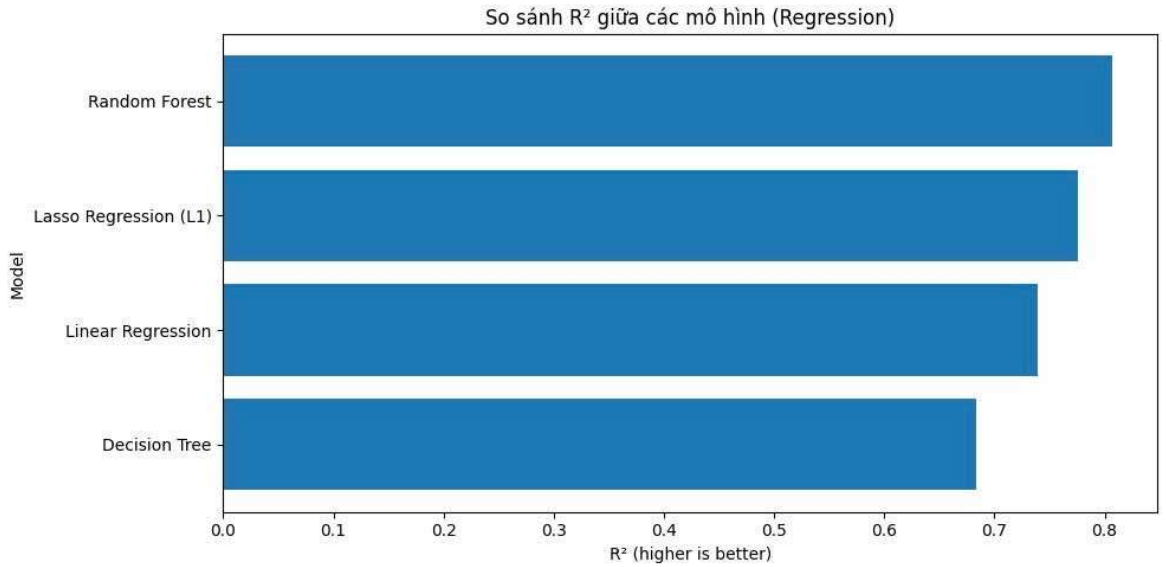


- Mô hình B

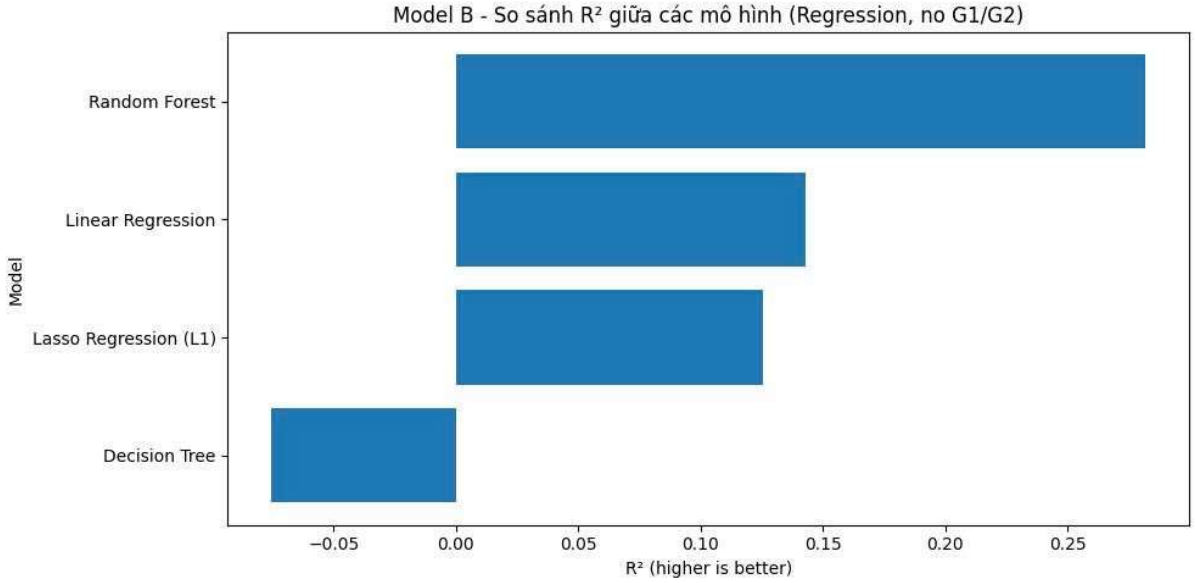


5 ĐÁNH GIÁ MÔ HÌNH

- Mô hình A



- Mô hình B



6 TỔNG KẾT

- Mô hình A

	Model	MAE	RMSE	R2
2	Random Forest	1.212380	1.987859	0.807287
3	Lasso Regression (L1)	1.487812	2.146620	0.775276
0	Linear Regression	1.691241	2.311294	0.739475
1	Decision Tree	1.354430	2.548268	0.683314

- Mô hình B

	Model	MAE	RMSE	R2
2	Random Forest	2.979038	3.837305	0.281889
0	Linear Regression	3.383695	4.191797	0.143081
3	Lasso Regression (L1)	3.419416	4.233886	0.125786
1	Decision Tree	3.620253	4.695810	-0.075376

6 TỔNG KẾT

- Mô hình A

```
import matplotlib.pyplot as plt

best_model_name = results_df.iloc[0]["Model"]
print("Best model by RMSE:", best_model_name)
```

```
Best model by RMSE: Random Forest
```

- Mô hình B

```
best_model_name_b1 = results_df_b1.iloc[0]["Model"]
print("Model B - Best model by RMSE:", best_model_name_b1)
```

```
Model B - Best model by RMSE: Random Forest
```

THANK YOU

FOR LISTENING!