ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Nhận dạng

BÁO CÁO BÀI TẬP 1

House Price Prediction Using Ridge and Lasso Regression

Sinh viên thực hiện:
Trương Huỳnh Thúy An 22520033

Mục lục

1. Tổng quan và phân tích dữ liệu

Bộ dữ liệu dự đoán giá nhà bao gồm 500 mẫu với 12 cột được lấy trên kaggle

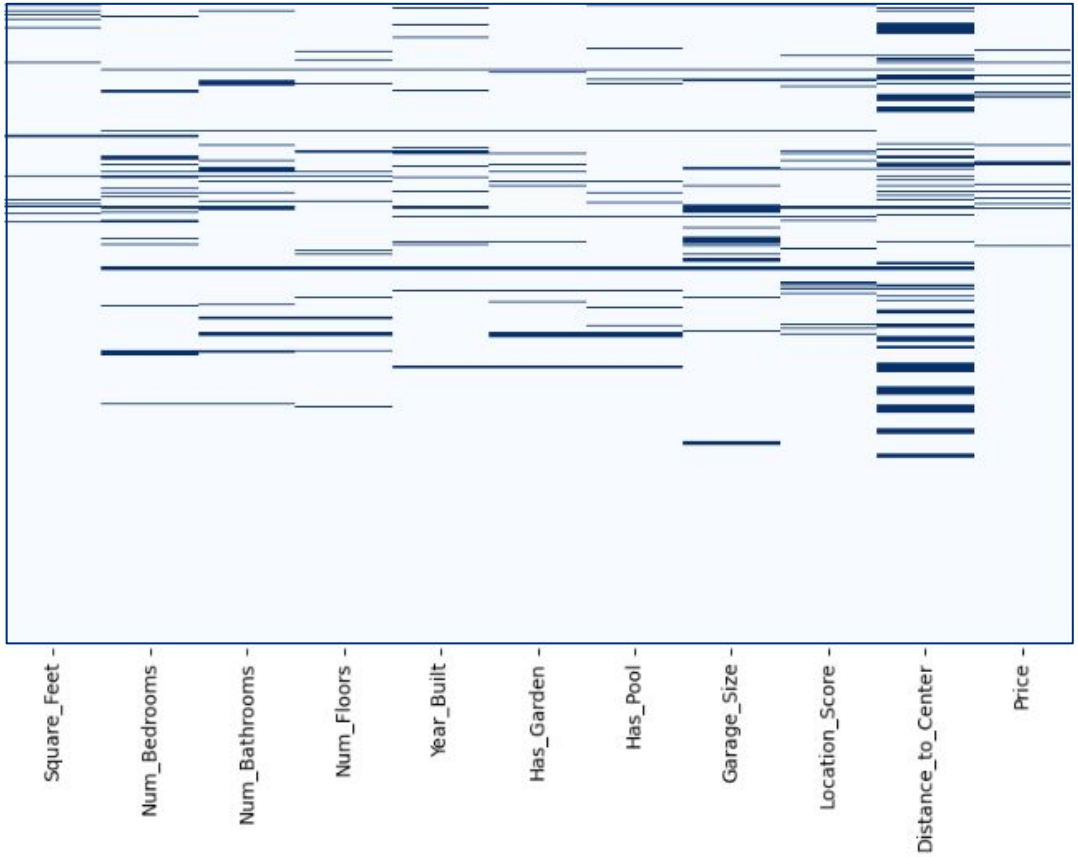
	ID	Square_Feet	Num_Bedrooms	Num_Bathrooms	Num_Floors	Year_Built	Has_Garden	Has_Pool	Garage_Size	Location_Score	Distance_to_Center	Price
0	1	143.635030	1.0	3.0	3.0	1967.0	1.0	1.0	48.0	8.297631	5.935734	602134.8167
1	2	287.678577	1.0	2.0	1.0	1949.0	0.0	1.0	37.0	6.061466	10.827392	591425.1354
2	3	NaN	1.0	3.0	2.0	1923.0	1.0	NaN	NaN	NaN	6.904599	464478.6969
3	4	199.664621	5.0	2.0	2.0	1918.0	0.0	0.0	17.0	2.070949	8.284019	583105.6560
4	5	89.004660	4.0	3.0	3.0	1999.0	1.0	0.0	34.0	1.523278	14.648277	619879.1425
...	...	...	...	...	...	...	...	...	...	...	...	...
495	496	138.338057	2.0	2.0	2.0	1967.0	1.0	0.0	16.0	4.296086	5.562583	488496.3507
496	497	195.914028	2.0	3.0	1.0	1977.0	0.0	1.0	45.0	7.406261	2.845105	657736.9217
497	498	69.433659	1.0	1.0	2.0	2004.0	0.0	0.0	18.0	8.629724	6.263264	405324.9502
498	499	293.598702	5.0	1.0	3.0	1940.0	1.0	0.0	41.0	5.318891	16.990684	773035.9680
499	500	296.552686	4.0	3.0	1.0	1988.0	1.0	1.0	20.0	7.894322	1.779794	864299.5002

500 rows × 12 columns

link dataset: https://www.kaggle.com/datasets/denkuznetz/housing-prices-regression/data

1. Tổng quan và phân tích dữ liệu

Heatmap of missing value



1. Tổng quan và phân tích dữ liệu

Số lượng dữ liệu thiếu của mỗi cột

	Missing Values	Percentage (%)
Square_Feet	14	2.8
Num_Bedrooms	34	6.8
Num_Bathrooms	34	6.8
Num_Floors	25	5.0
Year_Built	26	5.2
Has_Garden	22	4.4
Has_Pool	23	4.6
Garage_Size	37	7.4
Location_Score	28	5.6
Distance_to_Center	104	20.8
Price	18	3.6

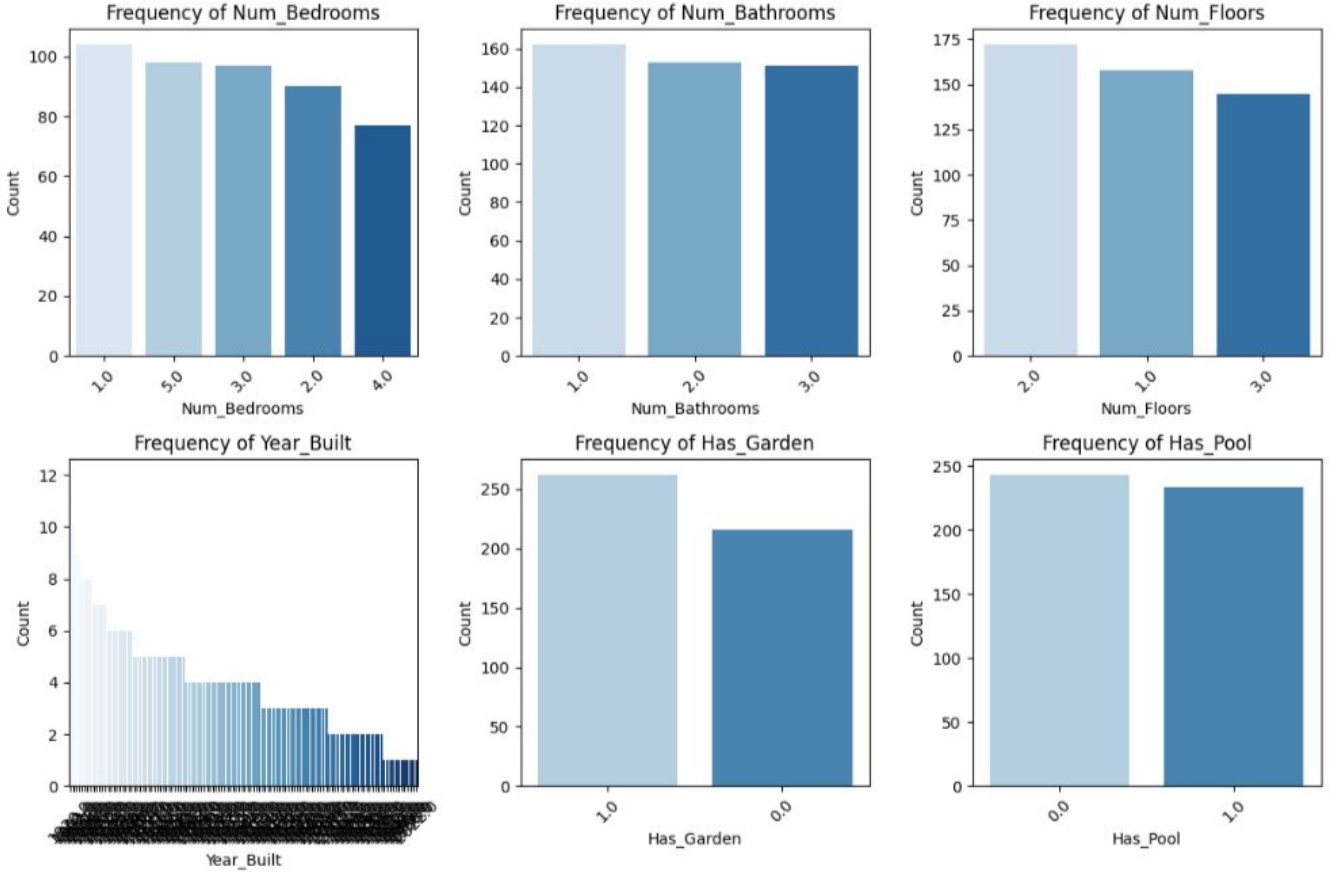
1. Tổng quan và phân tích dữ liệu

Bảng thống kê đối với các cột dữ liệu rời rạc

	Num_Bedrooms	Num_Bathrooms	Num_Floors	Year_Built	Has_Garden	Has_Pool
count	466.0	466.0	475.0	474.0	478.0	477.0
unique	5.0	3.0	3.0	120.0	2.0	2.0
top	1.0	1.0	2.0	1920.0	1.0	0.0
freq	104.0	162.0	172.0	12.0	262.0	243.0

1. Tổng quan và phân tích dữ liệu

Biểu đồ tần suất các cột dữ liệu rời rạc



1. Tổng quan và phân tích dữ liệu

Bảng thống kê đối với các cột dữ liệu liên tục

	Square_Feet	Garage_Size	Location_Score	Distance_to_Center	Price
count	486.000000	463.000000	472.000000	396.000000	482.000000
mean	174.378454	30.185745	5.198274	10.451424	582905.171008
std	74.735332	11.645933	2.866391	5.601969	123378.609003
min	51.265396	10.000000	0.004428	0.062818	276892.470100
25%	110.475759	20.000000	2.760650	5.904564	503205.143400
50%	177.929432	30.000000	5.272489	10.712509	576637.411800
75%	239.318092	40.500000	7.810841	15.179650	667431.212275
max	298.241199	49.000000	9.962623	19.927966	960678.274300

1. Tổng quan và phân tích dữ liệu

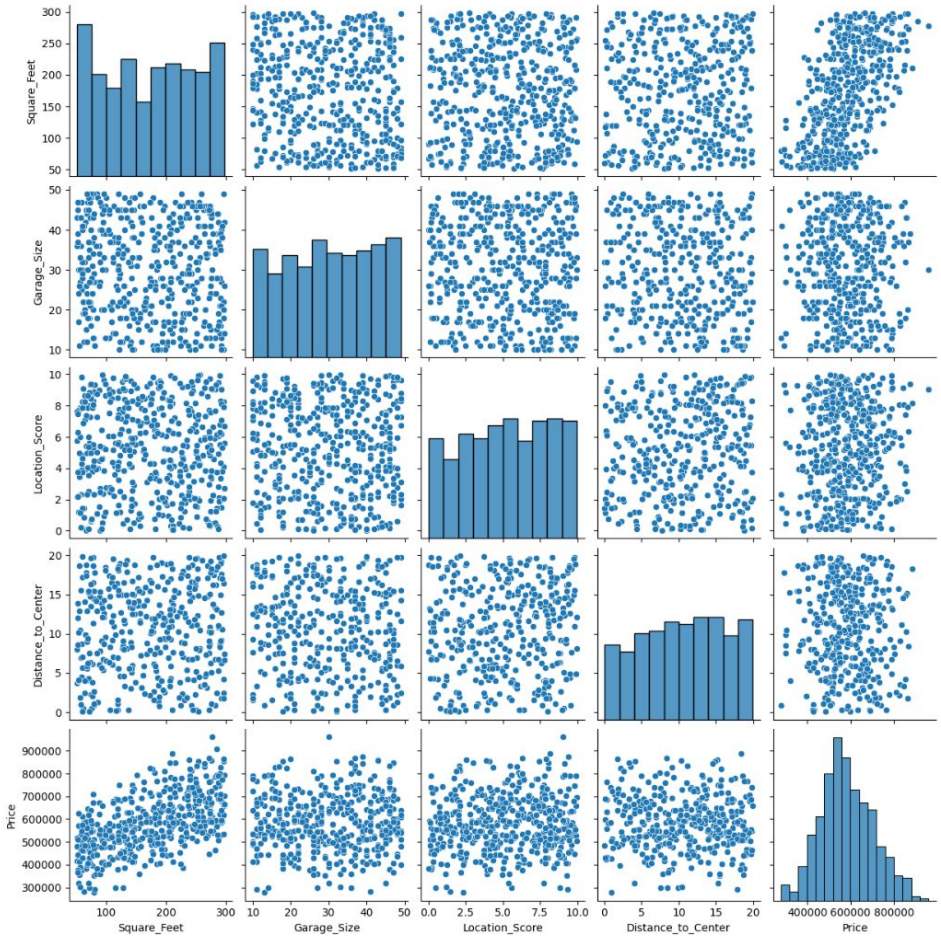
Square_Feet

Garage_Size

Location_Score

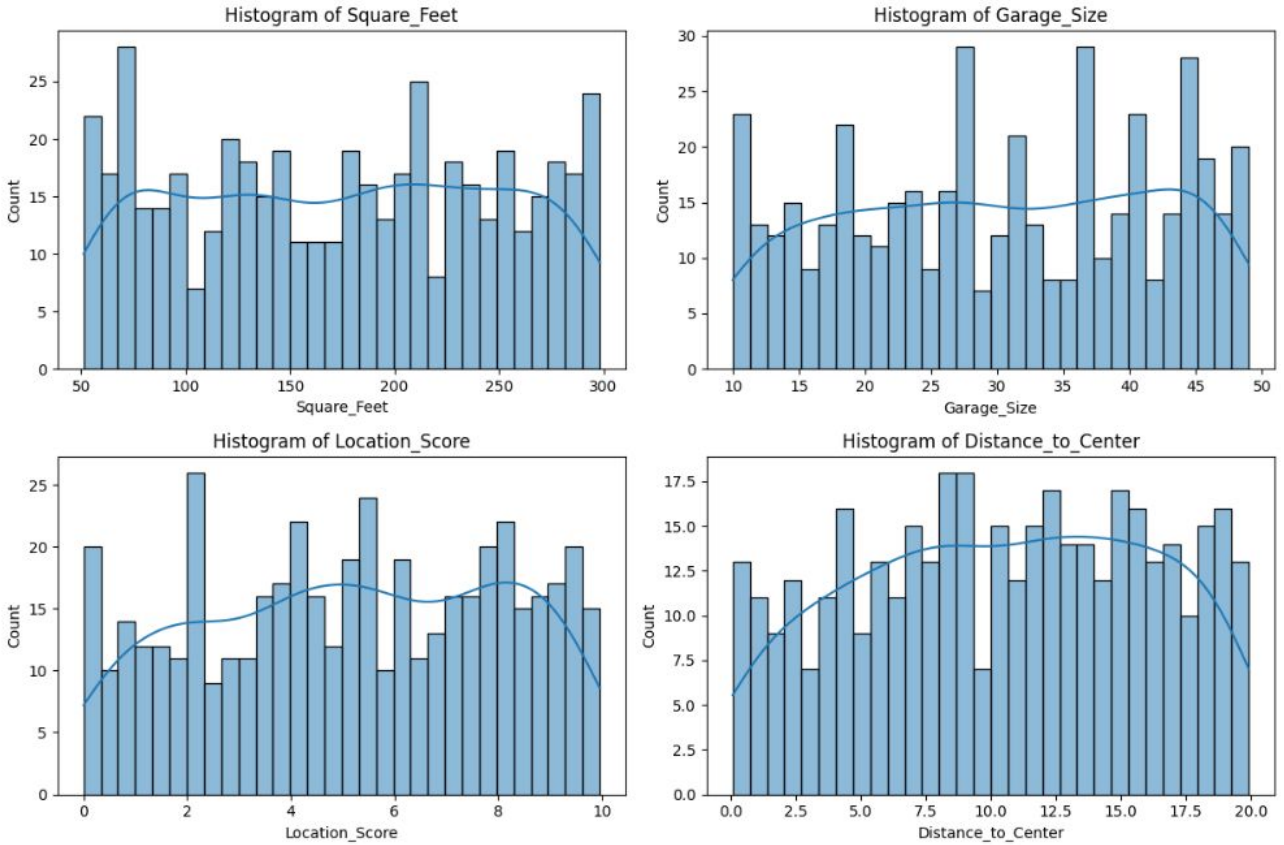
Distance_to_Center

Price

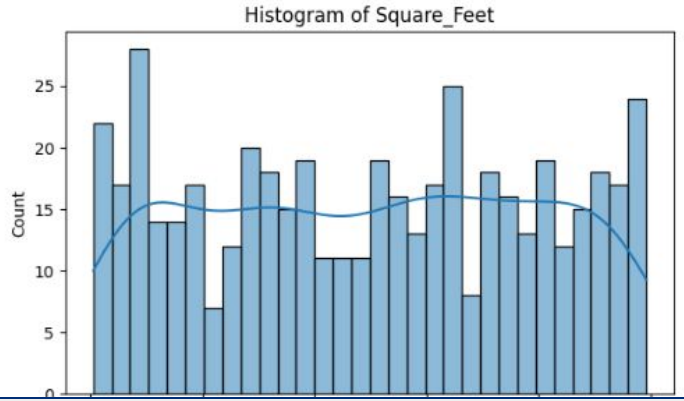


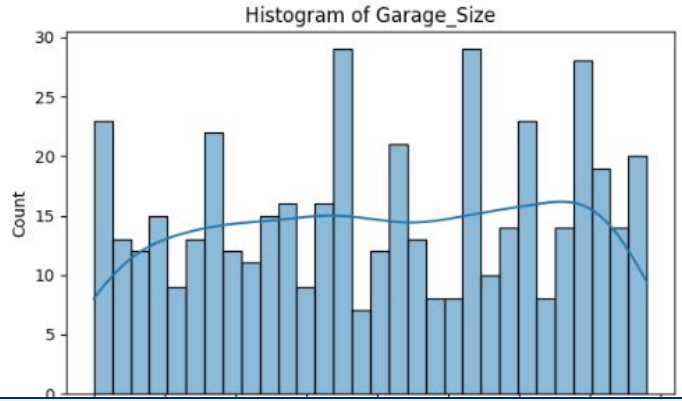
1. Tổng quan và phân tích dữ liệu

Biểu đồ tần suất các cột dữ liệu rời rạc

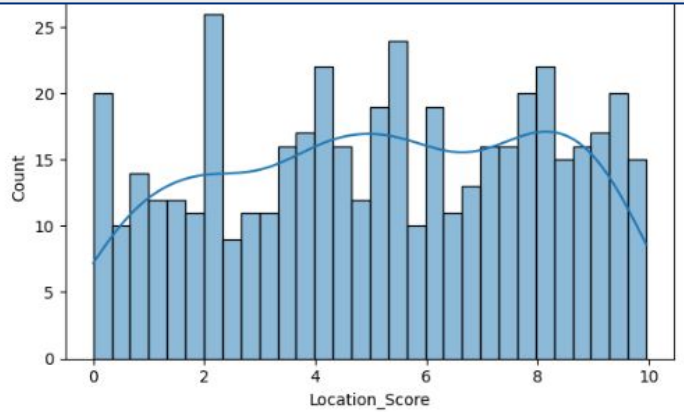


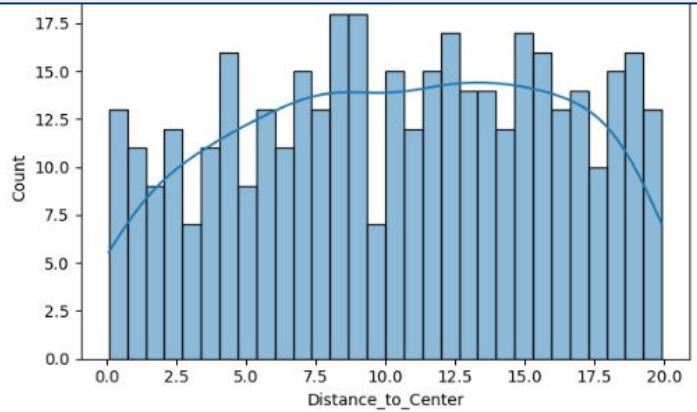
1. Tổng quan và phân tích dữ liệu





Standardization hay Normalization ?



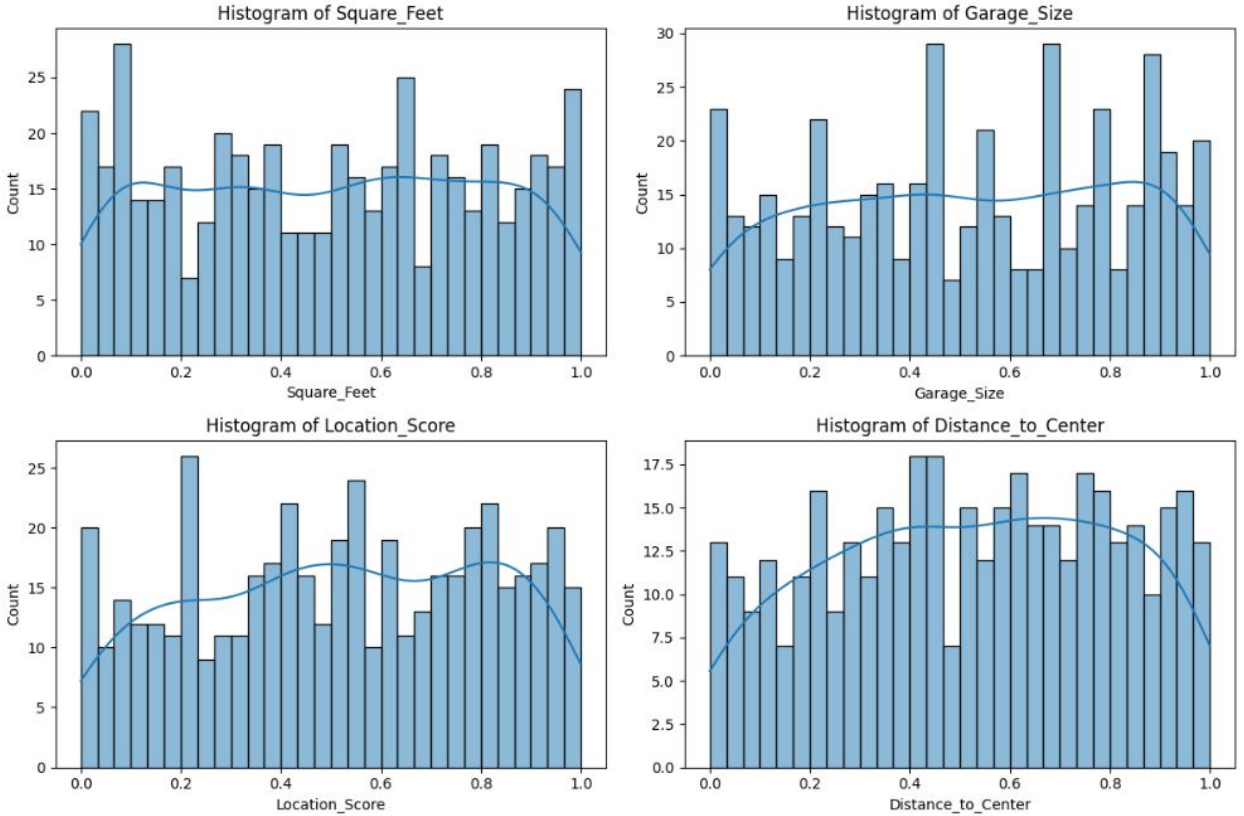


1. Tổng quan và phân tích dữ liệu

Vì các cột dữ liệu không tuân theo phân phối chuẩn (Gaussian Distribution) => dùng Normalization (Min-Max Scaling)

1. Tổng quan và phân tích dữ liệu

Biểu đồ tần suất các cột dữ liệu rời rạc đã được chuẩn hóa



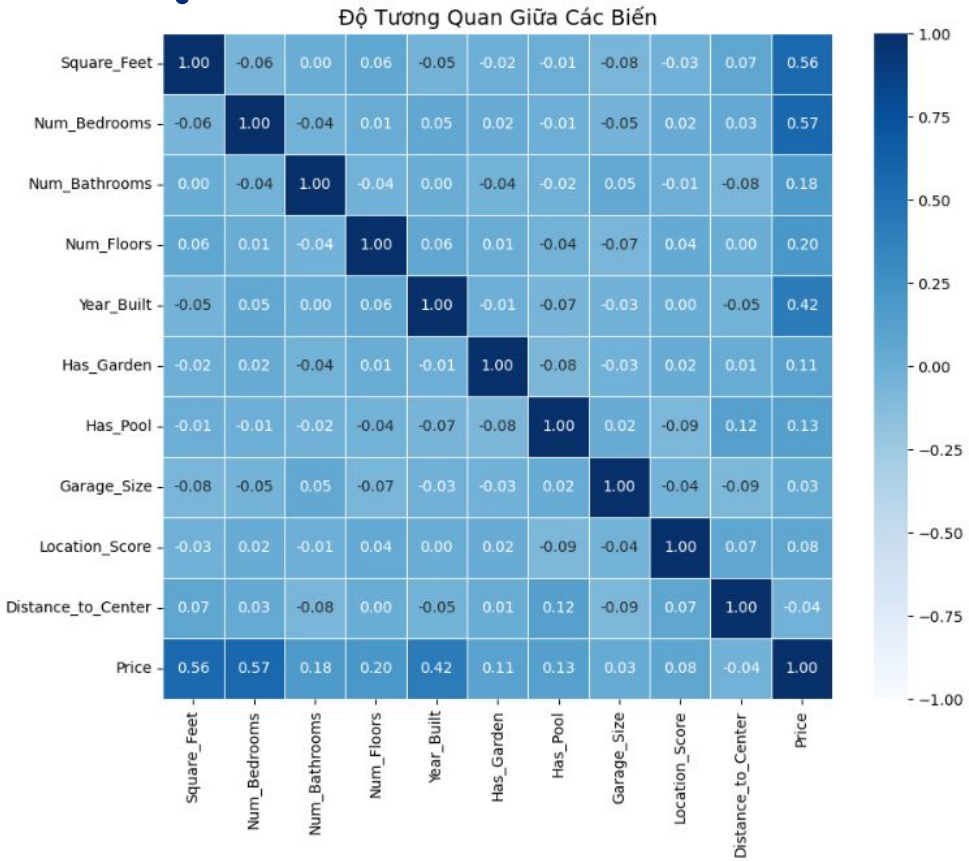
1. Tổng quan và phân tích dữ liệu

- Square_Feet, Num_Bedrooms, Year_Built có tương quan cao với Price.

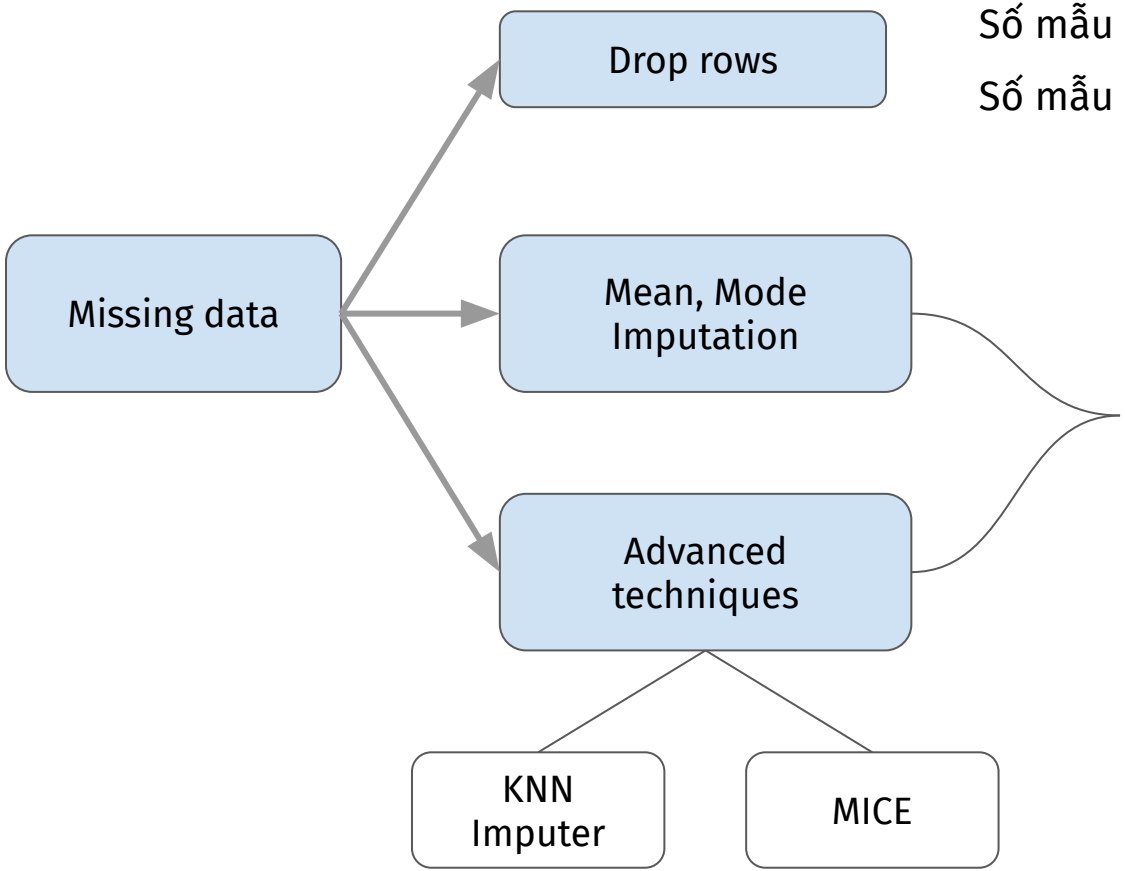
- Garage_Size và Distance_to_Center có độ tương quan thấp với Price, gần như không có tương quan.

- Các biến còn lại có tác động nhỏ đến Price, không phải là yếu tố quyết định chính.

Tất các các biến (trừ Price) có mức độ tương quan thấp với nhau, chứng tỏ chúng không ảnh hưởng đáng kể đến nhau trong mô hình dữ liệu này.



2. Xử lý dữ liệu

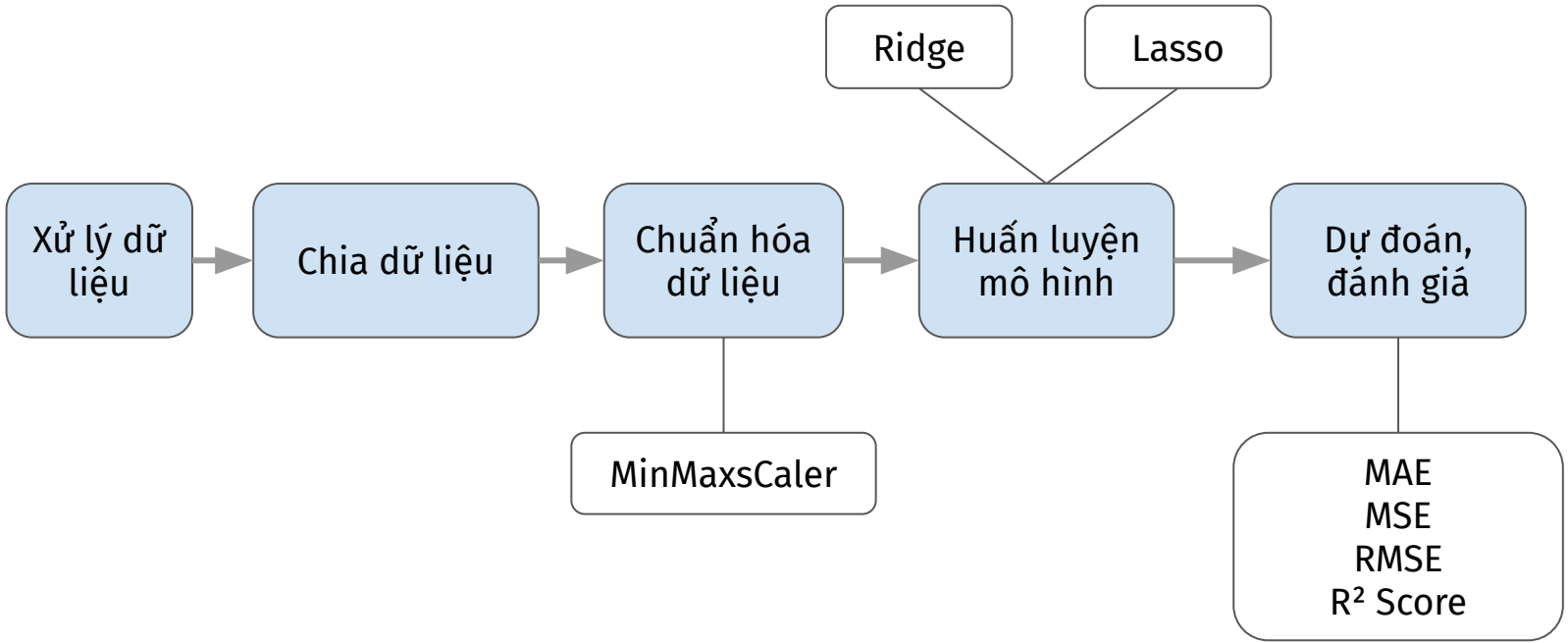


Số mẫu dữ liệu ban đầu: 500

Số mẫu dữ liệu sau khi xóa: 307

Vẫn giữ nguyên số lượng mẫu dữ liệu

3. Xây dựng mô hình

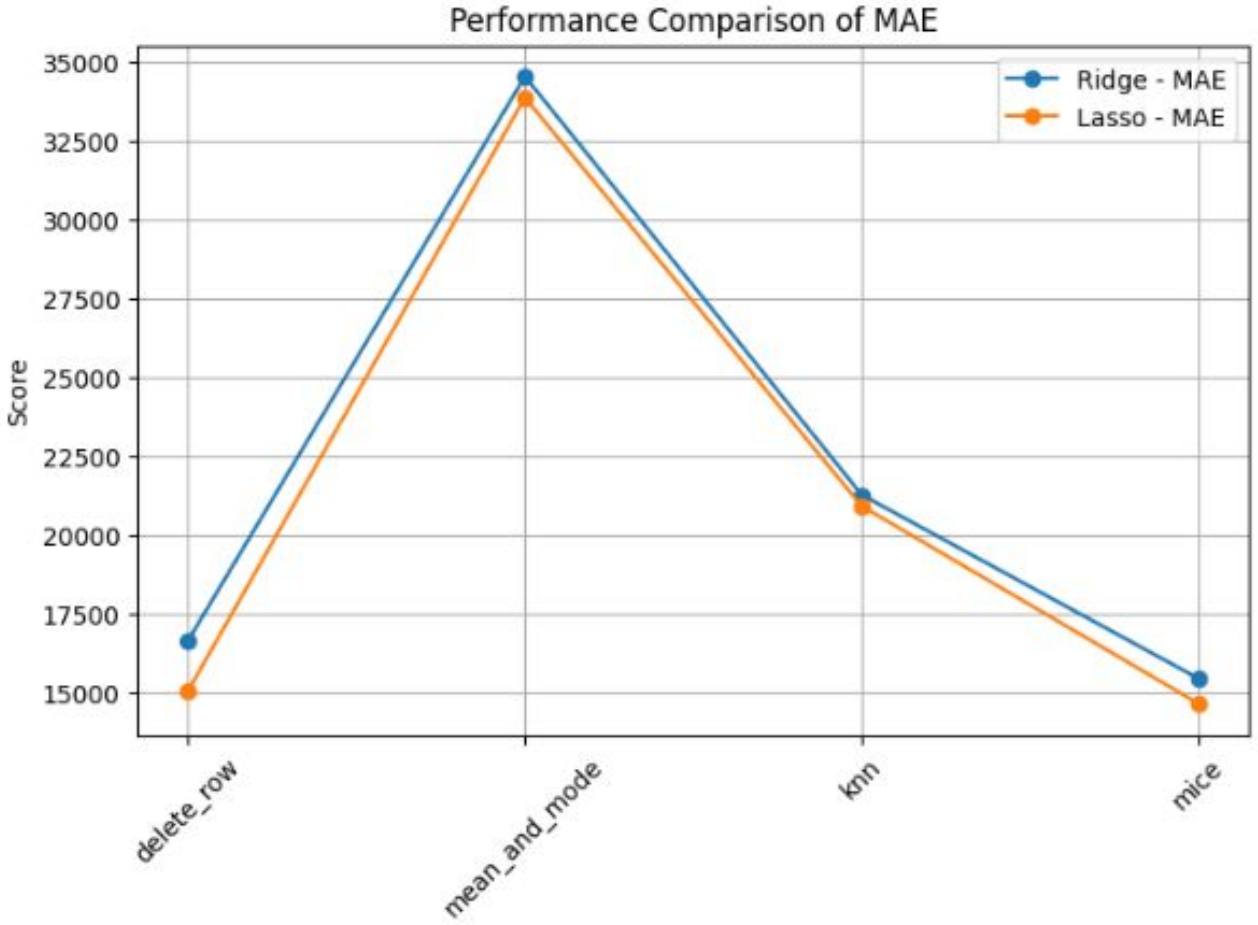


4. Kết quả

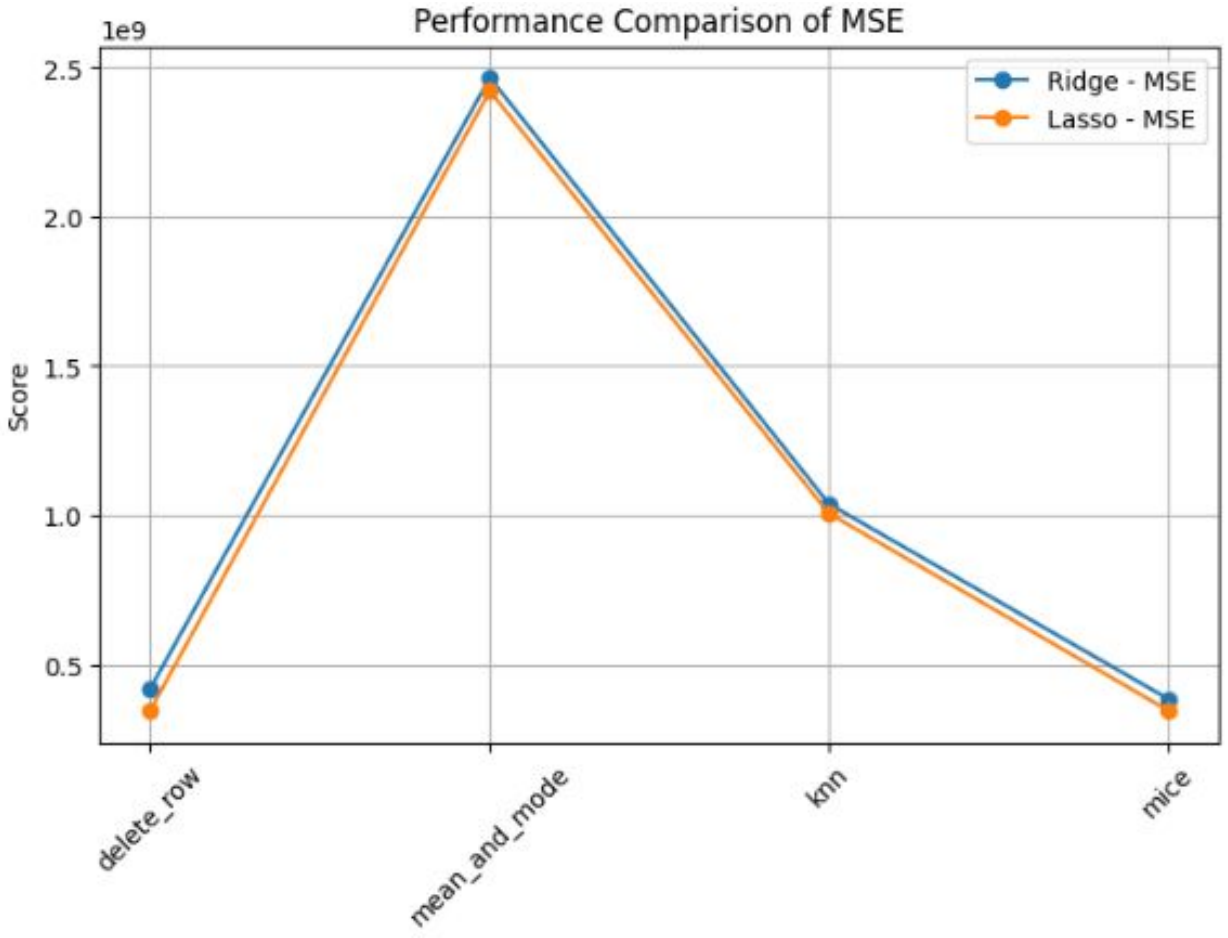
Detailed Results Table:

Method	Model	MAE	MSE	RMSE	R2 Score
delete_row	Ridge	16642.52	420449212.69	20504.86	0.9726
delete_row	Lasso	15047.67	348119568.2	18657.96	0.9773
mean_and_mode	Ridge	34515.73	2465672476.39	49655.54	0.8348
mean_and_mode	Lasso	33844.42	2420865642.47	49202.29	0.8378
knn	Ridge	21259.26	1037328703.06	32207.59	0.932
knn	Lasso	20902.34	1005611288.29	31711.37	0.9341
mice	Ridge	15437.51	384250313.26	19602.3	0.9747
mice	Lasso	14642.86	345225891.47	18580.26	0.9773

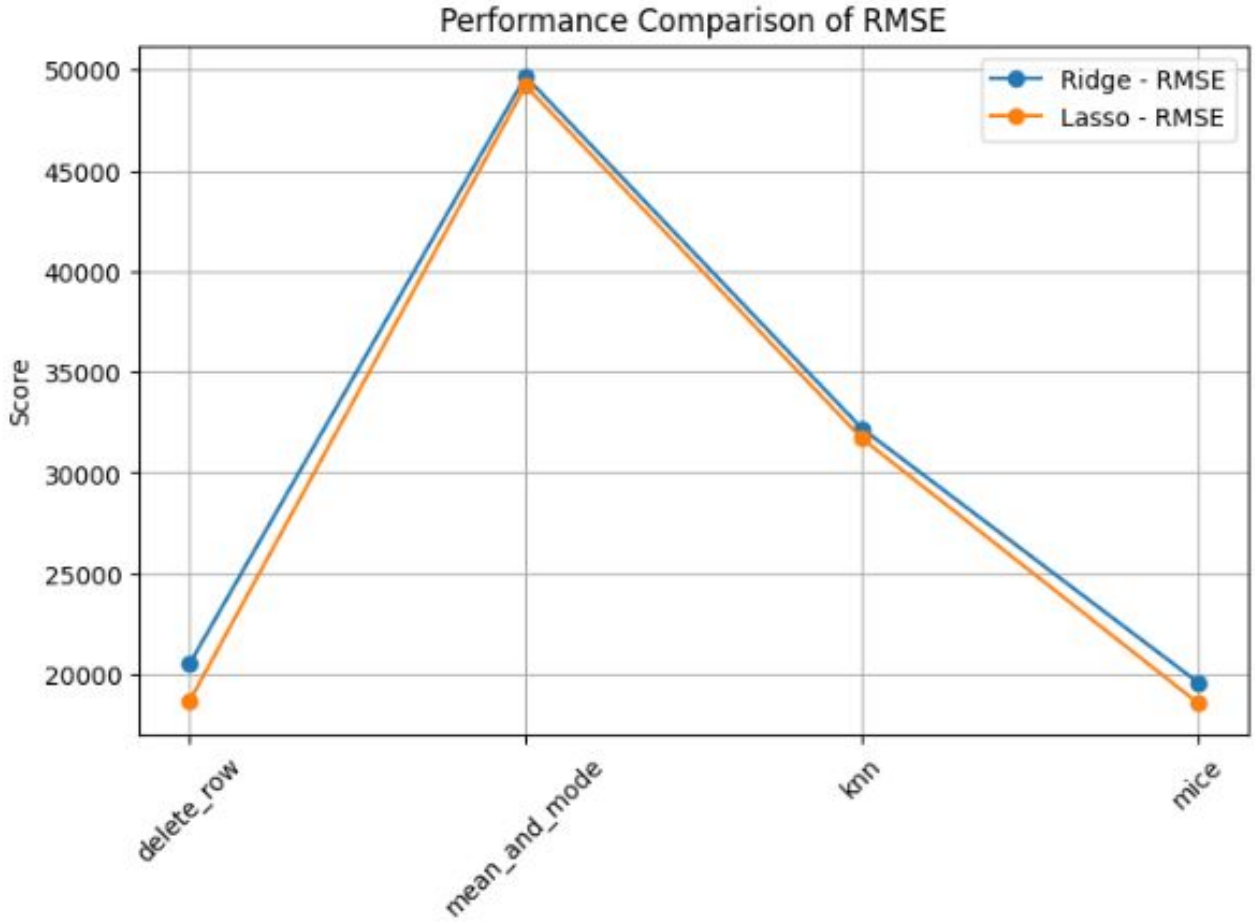
4. Kết quả


Performance Comparison of MAE
Ridge - MAE
Lasso - MAE
Score
35000
32500
30000
27500
25000
22500
20000
17500
15000
delete_row
mean_and_mode
knn
mice

4. Kết quả


Performance Comparison of MSE
1e9
2.5
2.0
1.5
1.0
0.5
Score
Ridge - MSE
Lasso - MSE
delete_row
mean_and_mode
knn
mice

4. Kết quả


Performance Comparison of RMSE
Ridge - RMSE
Lasso - RMSE
Score
50000
45000
40000
35000
30000
25000
20000
delete_row
mean_and_mode
knn
mice

4. Kết quả


Performance Comparison of R2 Score
Score
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
Ridge - R2 Score
Lasso - R2 Score
delete_row
mean_and_mode
knn
mice

Nhận xét kết quả

- **Phương pháp MICE cho kết quả dự đoán tốt nhất.** Cụ thể, MICE với mô hình **Lasso** đạt **MAE = 14,642.86**, **RMSE = 18,580.26**, và **R^2 Score = 0.9773**, cao nhất trong tất cả các phương pháp. Điều này chứng tỏ MICE có khả năng xử lý dữ liệu thiếu một cách hiệu quả, giúp mô hình học được thông tin quan trọng mà không làm giảm độ chính xác.

- **Phương pháp điền bằng Mean và Mode cho kết quả kém nhất.** Với **MAE lên đến 34,515.73 (Ridge) và 33,844.42 (Lasso)**, cùng với **R^2 Score giảm mạnh xuống còn khoảng 0.8348 - 0.8378**. Điều này có thể do việc thay thế giá trị bị thiếu bằng trung bình hoặc mode làm mất đi tính đa dạng của dữ liệu gốc, khiến mô hình không thể học được quan hệ thực sự giữa các biến.

Nhận xét kết quả

- Phương pháp KNN mặc dù cải thiện hơn so với Mean/Mode, nhưng vẫn không bằng MICE. Với **MAE = 21,259.26 (Ridge) và 20,902.34 (Lasso)** cùng **R^2 Score khoảng 0.932 - 0.9341**, phương pháp này vẫn giữ được mức độ chính xác nhất định, tuy nhiên không tối ưu bằng MICE.

Phương pháp xóa hàng chứa dữ liệu thiếu (delete_row) cho thấy kết quả khá tốt. Với **R^2 Score = 0.9726 (Ridge) và 0.9773 (Lasso)**, rất gần với phương pháp MICE. Tuy nhiên, dù phương pháp này có thể duy trì độ chính xác, nhưng nếu số lượng dữ liệu bị thiếu lớn, việc xóa quá nhiều hàng có thể làm mất đi thông tin quan trọng, gây ảnh hưởng đến chất lượng mô hình.

5. Kết luận

MICE là phương pháp tốt nhất, cho độ chính xác cao nhất với $R^2 \approx 0.9773$.

Mean/Mode là phương pháp kém nhất, làm giảm đáng kể hiệu suất mô hình.

KNN có hiệu suất trung bình, tốt hơn Mean/Mode nhưng kém MICE.

Xóa hàng dữ liệu thiếu có thể là một lựa chọn hợp lý, nhưng cần cân nhắc nếu dữ liệu bị thiếu quá nhiều.

👉 Dựa trên kết quả này, nên ưu tiên sử dụng **MICE** để xử lý dữ liệu thiếu trong bài toán này nhằm đảm bảo mô hình đạt hiệu suất cao nhất.