



**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP. HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ ĐIỆN TỬ**

**----🙣🕮🙡----**

**Học Máy**

**Đề tài:**

**Natural Language Processing with Disaster Tweets**

**NHÓM 14**

***Giảng viên hướng dẫn*: Phạm Việt Thành**

***Sinh viên thực hiện*: Nguyễn Duy Tùng 21085061**

**21001201**

**21136041**

**21135711**

**21071741**

***Lớp*: DHIOT17A**

*TP. Hồ Chí Minh, tháng 11, năm 2024*

**Tên đề tài: Natural Language Processing with Disaster Tweets**

***Lớp*: DHIOT17A Nhóm: 14**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **MSSV** | **Họ và Tên** | **Nhiệm vụ Công việc trong nhóm** | **Điểm tự đánh giá  của nhóm** | **Điểm đánh giá  của giáo viên** |
| 1 | 21001211 | Nguyễn Huy Thái | Huấn luyện và đánh giá mô hình, mã hóa các cột danh mục. Thực hiện powerpoint, word | 10 |  |
| 2 | 21001201 | La Kỳ Nguyên | Xử lý dữ liệu bị thiếu, chuẩn hóa dữ liệu. Thực hiện powerpoint,word | 10 |  |
| 3 | 21136041 | Đinh Quốc Huy | Trực quan hóa dữ liệu,Thực hiện powerpoint,word | 8 |  |
| 4 | 21071741 | Lê Thị Hoàng Anh | Xác định tỉ lệ rỗng các cột dữ liệu. Thực hiện powerpoint,word | 8 |  |
| 5 | 21135711 | Đỗ Mạnh Dũng | Chuyển đổi các cột từ dạng số sang danh mục. Thực hiện powerpoint | 8 |  |

MỤC LỤC

[BÁO CÁO KẾT QUẢ BÀI TẬP LỚN 4](#_Toc181033927)

[1. Mục tiêu đề tài: 4](#_Toc181033928)

[2. Tổng quan về đề tài 4](#_Toc181033929)

[3. Các nội dung đã thực hiện 5](#_Toc181033930)

[3.1. Phân tích và khám phá dữ liệu 5](#_Toc181033931)

[3.1.1. Xác định các đặc trưng có trong tập dữ liệu: 5](#_Toc181033932)

[3.1.2. Xử lý dữ liệu bị thiếu: 8](#_Toc181033933)

[3.2. Trực quan hóa dữ liệu: 12](#_Toc181033934)

[3.3. Mã hóa dữ liệu 21](#_Toc181033935)

[3.4. Chuẩn hóa dữ liệu: 22](#_Toc181033936)

[3.5. Mô hình hóa: 23](#_Toc181033937)

[3.6. Đánh giá: 23](#_Toc181033938)

[4. Các kết quả đã đạt được (chưa đạt được) 26](#_Toc181033939)

[4.1. Kết quả đạt được 26](#_Toc181033940)

[4.2. Các kết quả chưa đạt được 26](#_Toc181033941)

[5. Những khó khăn, vướng mắc gặp phải trong quá trình thực hiện 27](#_Toc181033942)

[6. Những kinh nghiệm thu được sau khi hoàn thành bài tập lớn 28](#_Toc181033943)

[7. Kết luận (và tự đánh giá theo thang điểm từ 0 đến 10) 29](#_Toc181033944)

[8. Tài liệu tham khảo 30](#_Toc181033945)

BÁO CÁO KẾT QUẢ BÀI TẬP LỚN

# Mục tiêu đề tài:

Phân loại tin tức về thảm họa tự động: xây dựng mô hình xử lý ngôn ngữ tự nhiên (Natural Language Processing hay NLP) nhằm phân loại các dòng Tweets (nền tảng mạng xã hội) thành hai nhóm chính là tin có nội dung liên quan đế thảm họa và tin không liên quan, từ đó giúp nhận diện sớm các thông tin hữu ích trong các tình huống khẩn cấp

Hiểu rõ hơn về cách sử dụng các kỹ thuật NLP: Tìm hiểu và ứng dụng các kỹ thuật tiên tiến trong NLP như stemming, loại bỏ từ dừng, xử lý các đặc trưng và trọng số của từ (như IDF – Inverse Documents Frequency), đồng thời tối ưu các tham số mô hình để cải thiện độ chính xác.

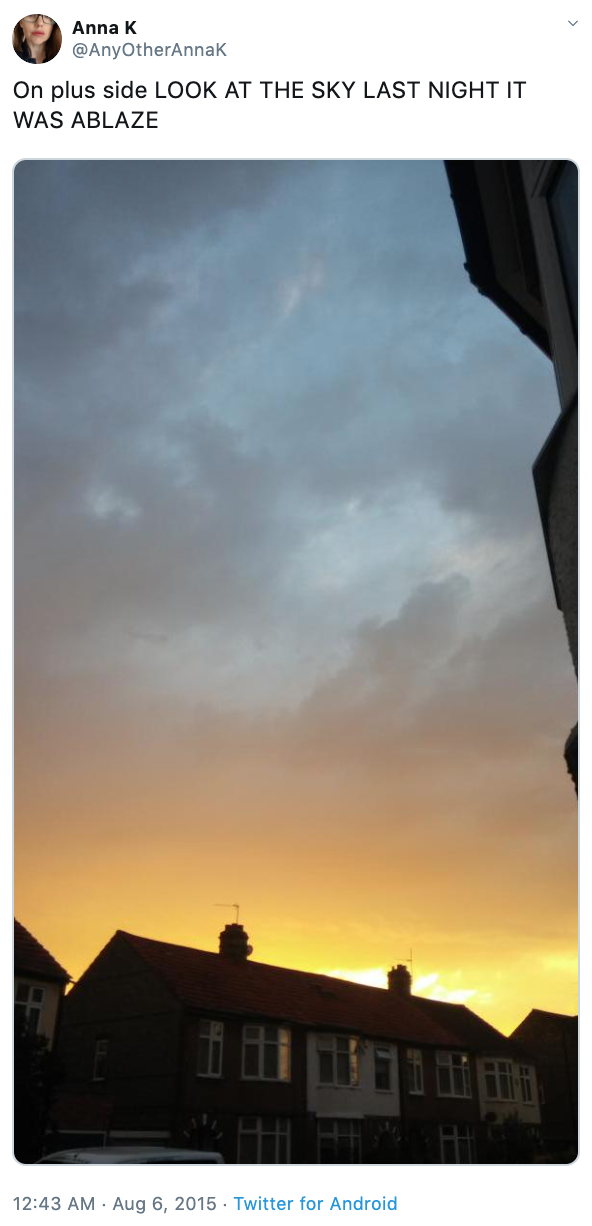
# Tổng quan về đề tài

Natural Language Processing – xử lý ngôn ngữ tự nhiên: là một phân ngành học của máy tính và đặc biệt là trí tuệ nhân tạo. Nó chủ yếu liên quan đến việc cung cấp cho máy tính khả năng xử lý dữ liệu được mã hóa bằng ngôn ngữ tự nhiên và do đó có liên quan chặt chẽ đến việc truy xuất dữ liệu, biễu diễn kiến thức và ngôn ngữ học tính toán, một phân ngành của ngôn ngữ học. Thông thường, dữ liệu được thu thập trong các tập hợp văn bản, sử dụng các phương pháp tiếp cận dựa trên quy tắc, thống kê hoặc dựa trên nơ-ron trong học máy và học sâu. Các nhiệm vụ chính trong xử lý ngôn ngữ tự nhiên là nhận dạng giọng nói, phân loại văn bản, hiểu ngôn ngữ tự nhiên và tạo ngôn ngữ tự nhiên.

Disaster Tweets: tweets là các mẩu tin nhỏ trong chức năng đọc, nhắn và cập nhật trong mạng xã hội Twitter bây giờ được gọi là X, còn disaster dịch theo nghĩa tiếng anh là thảm họa.

Mô tả: Twitter (hiện tại là X) đã trở thành một kênh giao tiếp quan trọng trong các tình huống khẩn cấp. Với sự phổ biến rộng rãi của điện thoại thông minh, mọi người có thể thông báo về một tình huống khẩn cấp mà họ đang chứng kiến theo thời gian thực. Vì lý do này, ngày càng có nhiều cơ quan quan tâm đến việc giám sát Twitter một cách tự động (ví dụ như các tổ chức cứu trợ thảm họa và các cơ quan tin tức)

Tuy nhiên, không phải lúc nào cũng rõ ràng rằng lời nói của một người có thực sự là đang thông báo về một thảm họa hay không. Hãy xem xét ví dụ sau đây:



Tác giả sử dụng từ “ABLAZE” một cách rõ ràng, nhưng ý của họ lại là ẩn dụ. Điều này rất dễ nhận ra đối với con người, đặc biệt khi có hình ảnh minh họa. Nhưng đối với máy móc thì lại khó nhận biết hơn.

Trong đề tài này nhóm sẽ xây dựng mô hình máy học để dự đoán xem các Tweets có liên quan đến thảm họa không

Tập dữ liệu đã được cuộc thi thu thập nhóm sẽ tận dụng tập dữ liệu có sãn để xây dựng mô hình

Để thực hiện được mô hình máy học thì nhóm sẽ liệt kê quy trình thực hiện xử lý NLP của nhóm sử dụng cho đề tài

1. Tokenization: tách văn bản từ tập dữ liệu
2. Stop Words Removal: loại bỏ các từ thông dụng không có giá trị để lọc nhiễu
3. Steming: rút gọn từ về dạng gốc
4. TF – IDF (term frequency – inverse document frequency): tạo ra các đặc trưng số dựa trên tần suất từ để xác định mức độ quan trọng của từ
5. Feature Engineering: xây dựng các đặc trưng từ văn bản để làm đầu vào cho mô hình
6. Classification: sử dụng mô hình học máy để phân loại văn bản, trong đề tài của nhóm sử dụng LinearRegression (hồi quy tuyến tính) và dùng F1-score đáng giá hiệu quả (dựa trên yêu cầu đề tài)

Các Đánh giá (Evaluation)

F1 được tính theo công thức sau

*F1 =*  với *precision =*  và recall =

Chú thích:

* True Positive [TP] = dự đoán của bạn là 1 và giá trị thực tế cũng là 1 - bạn dự đoán là tích cực và điều đó đúng
* False Positive [FP] = dự đoán của bạn là 1 và giá trị thực tế cũng là 0 - bạn dự đoán là tích cực và điều đó sai
* False Nagative [FN] = dự đoán của bạn là 0 và giá trị thực tế cũng là 1 - bạn dự đoán là tiêu cực và điều đó sai

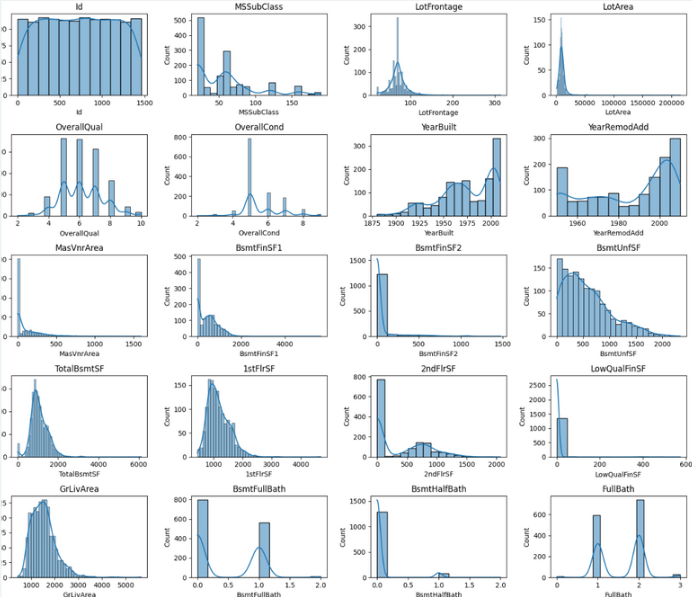
# Các nội dung đã thực hiện

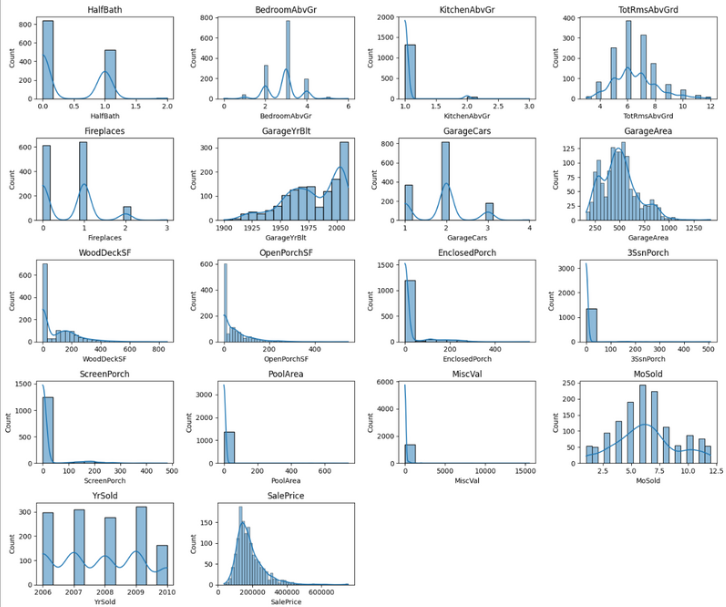
## Phân tích và khám phá dữ liệu

### Xác định các đặc trưng có trong tập dữ liệu:

### Xử lý dữ liệu bị thiếu:

## Trực quan hóa dữ liệu:

****

****

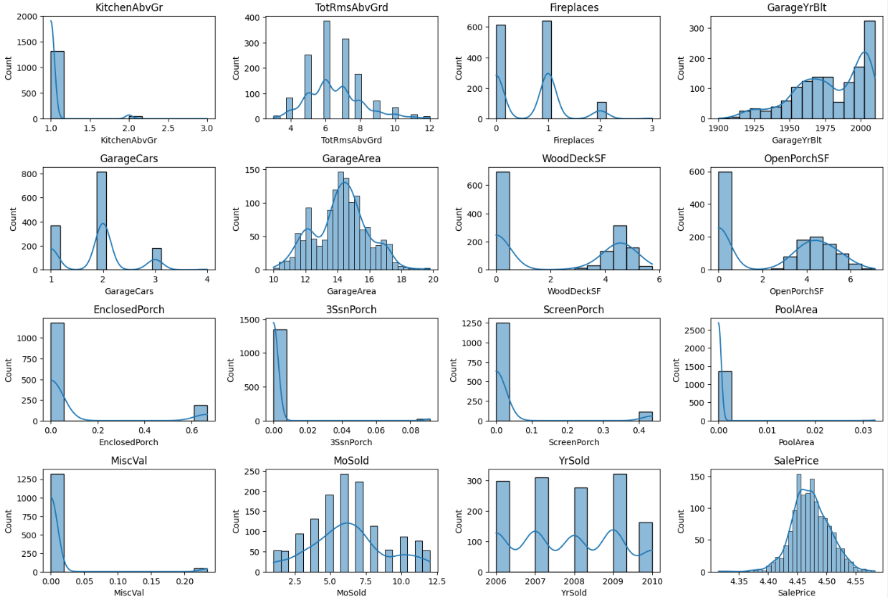
Có 37 cột dữ liệu số, bao gồm các loại sau:

* + Diện tích: Chỉ ra diện tích theo feet vuông
  + Thời gian: Các biến liên quan đến thời gian như khi nào ngôi nhà được xây dựng hoặc bán.
  + Chất lượng: Các biến định tính được đánh giá từ 1-10.
  + Phòng và tiện nghi: Các biến liên quan đến số lượng phòng và phòng tắm trong ngôi nhà.

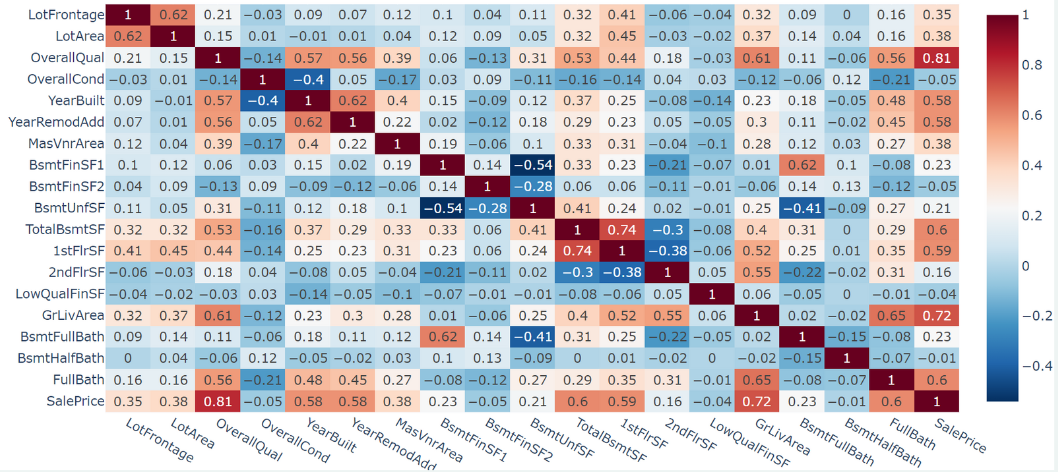
Phần lớn các biến liên quan đến không gian thực tế của căn hộ đều có phân phối lệch — điều này hợp lý vì mọi người có xu hướng sống trong các ngôi nhà hoặc căn hộ nhỏ hơn.

*Sử dụng phép biến đổi box-cox để giảm độ lệch. Ta có được biểu đồ phân phối như sau:*





Chúng ta có thể thấy rằng phép biến đổi boxcox đã làm cho dữ liệu gần với phân phối chuẩn hơn (normal distribution), đặc biệt là các biến liên quan đến không gian vật lý thực tế của căn hộ.

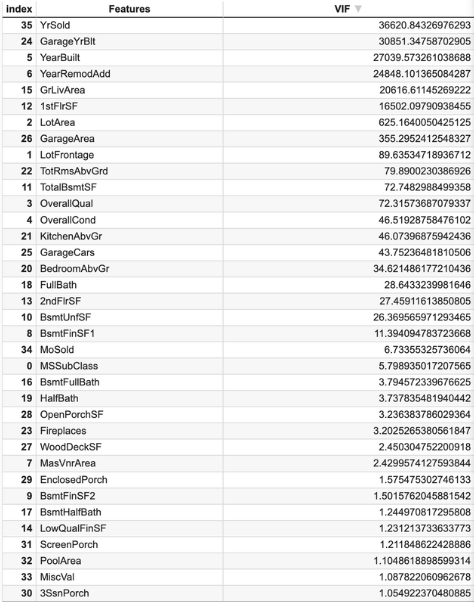




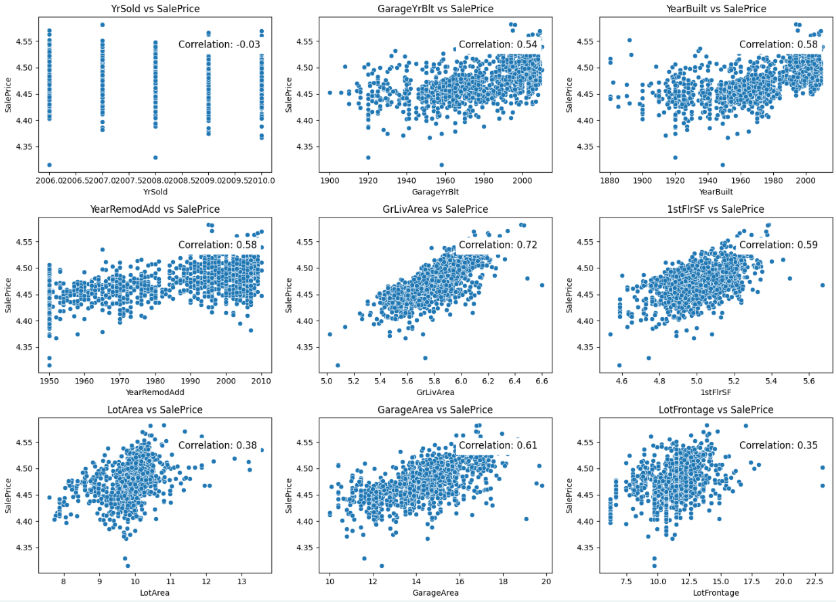
Chúng ta sẽ tính Variance Inflation Factor (VIF) để xem xét tính đa cộng tuyến (multicollinearity) giữa các biến độc lập.

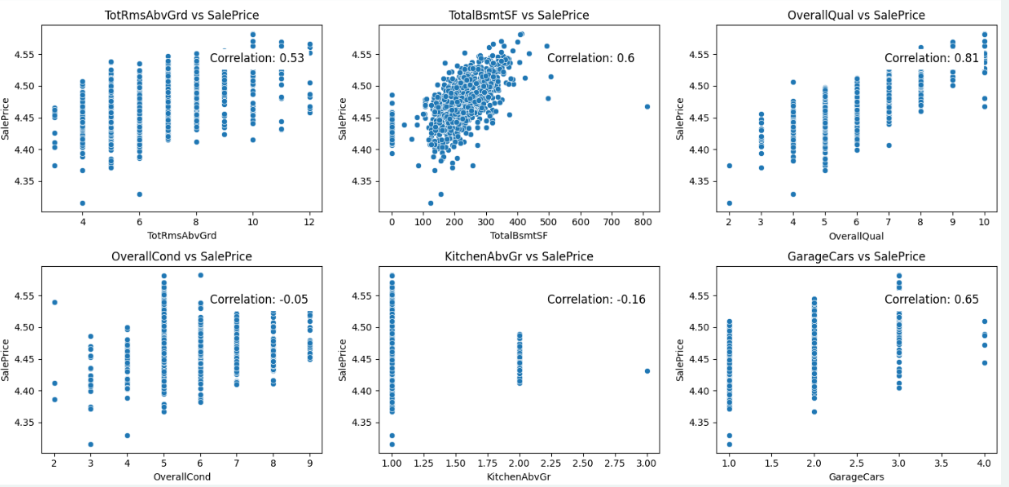
VIF là một chỉ số thống kê được sử dụng để phát hiện và đo lường mức độ đa cộng tuyến giữa các biến độc lập trong một mô hình hồi quy tuyến tính. Đa cộng tuyến xảy ra khi các biến độc lập trong mô hình có mối tương quan cao với nhau, gây ra các vấn đề trong việc ước lượng chính xác hệ số hồi quy và làm giảm độ tin cậy của các ước lượng.

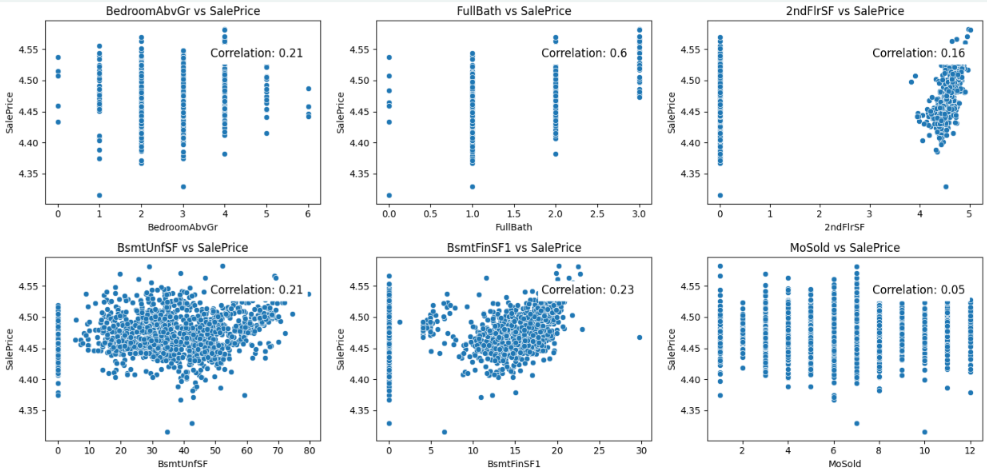
Bảng dưới là VIF của các biến với giá bán:



*Ta tiến hành phân tích các đặc trưng có VIF cao (VIF>5).*



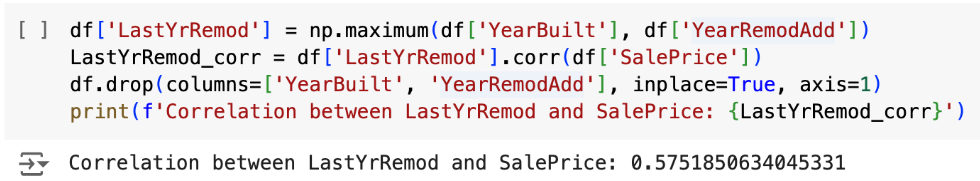




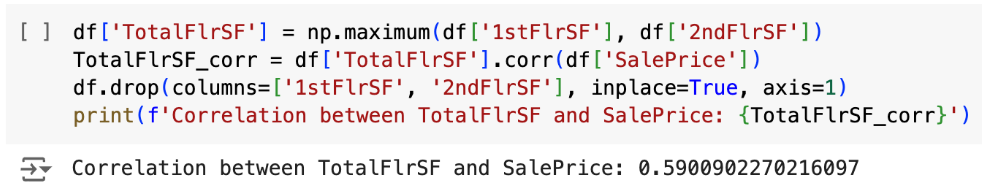
Dựa trên các phân tích ở kết quả trên thì ta có thể bỏ các cột sau:

* GarageCars: Thay thế vào cột này ta có GarageArea
* LotFrontage: tương quan thấp
* KitchenAbvGr, BedroomAbvGr: Thay thế hai cột này ta có TotRmsAbvGrd
* BsmtFinSF1, BsmtFinSF2:Thay thế vào cột này ta có BsmtUnfSF
* YrSold: tương quan rất thấp
* YearBuilt, YearRemodAdd: có thể gộp 2 cột này thành 1
* 1stFlrSF, 2ndFlrSF: có thể gộp 2 cột này thành 1

Gộp 2 cột YearBuilt, YearRemodAdd thành cột LastYrRemod:

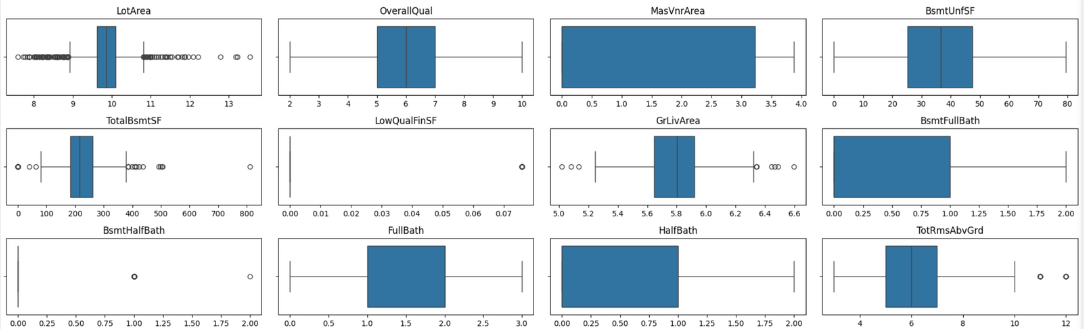
****

Gộp 2 cột 1stFlrSF, 2ndFlrSF thành cột TotalFlrSF:

****

Sự tương quan của đặc tính mới dường như giống hệt với các đặc tính cũ, thậm chí còn tốt hơn

*Tiếp theo, chúng ta tìm xem có đặc trưng nào có điểm ngoại lai (outliers) không.*

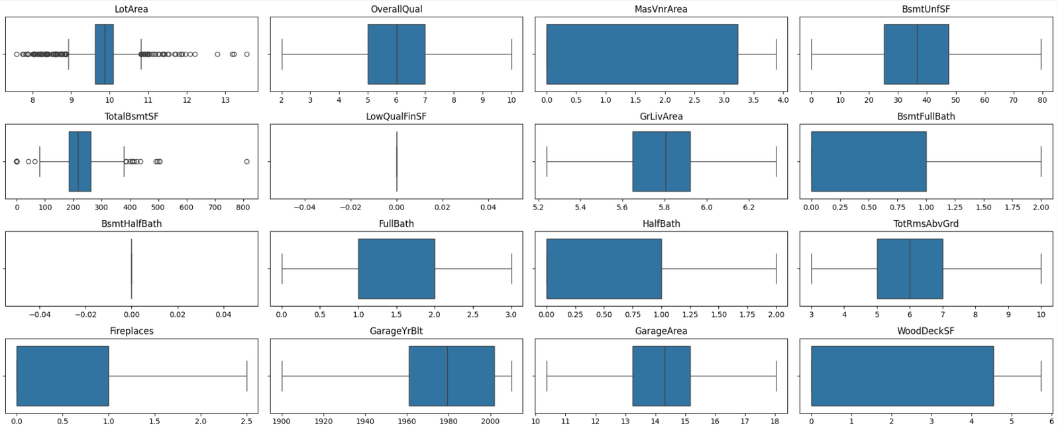


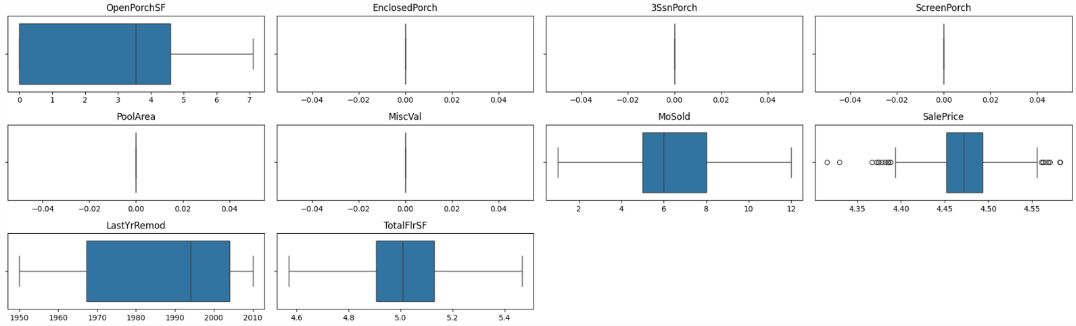


**Nhận xét:** sau khi phân tích các đặc trưng trên ta thấy có nhiều cột có giá trị outliers ta có thể xử lý bằng các phương pháp:

* Loại bỏ chúng
* Thay thế các giá trị nhỏ hơn bằng cách tăng các giá trị giới hạn trên hoặc giới hạn dưới.

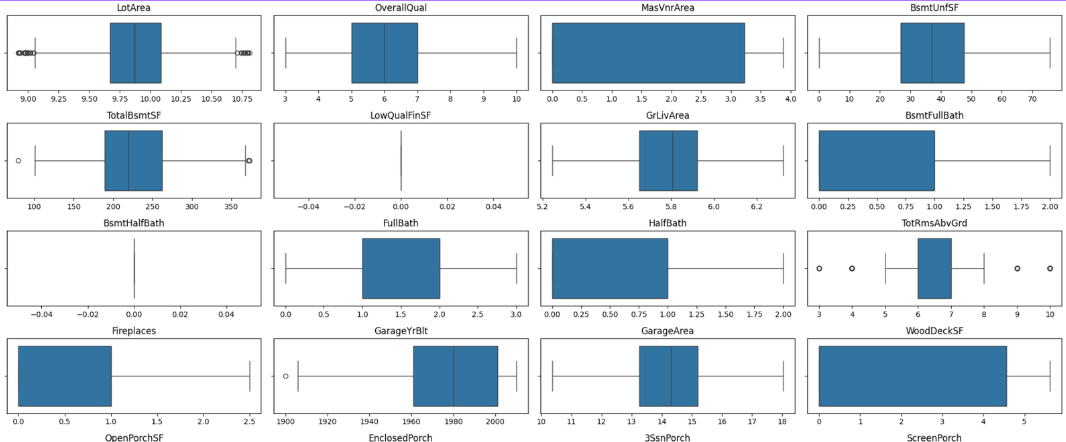
Một số cột có số lượng nhỏ giá trị outliers, chúng ta sẽ thử thay thế các giá trị bằng giới hạn trên và giới hạn dưới và xem cách này có hiệu quả hay không.

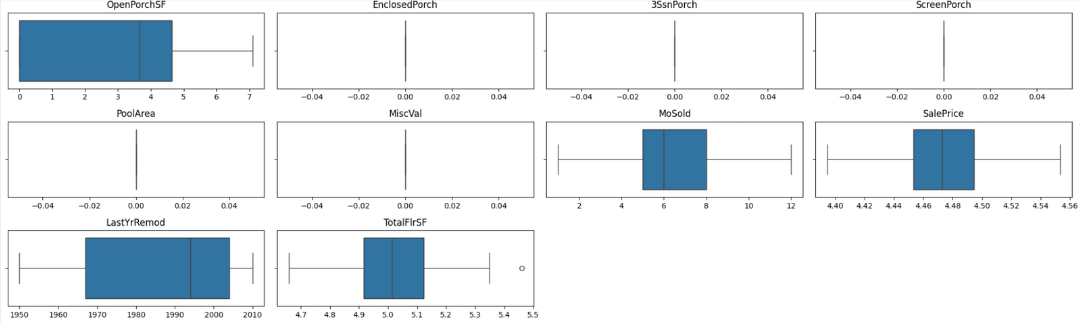




**Nhận xét:** Vẫn còn outliers ở các đặc trưng: LotArea, TotalBsmtSF, SalePrice.

Loại bỏ outliers ở các đặc trưng này.





## Mã hóa dữ liệu

Mã hóa biến phân loại là bước tiền xử lý dữ liệu, đặc biệt khi làm việc với các mô hình chỉ chấp nhận đầu vào dạng số.

Các biến phân loại (Categorical Variables), như tên gọi, biểu diễn dữ liệu theo các danh mục hoặc nhóm riêng biệt (ví dụ: màu sắc, loại nhà, thành phố...).

Các mô hình Machine Learning thường không thể trực tiếp xử lý các giá trị dạng chuỗi hoặc nhãn này, do đó chúng ta cần mã hóa chúng thành dạng số.

Ở mô hình này ta sử dụng hai phương pháp chính để mã hóa biến phân loại:

* Mã hóa thứ tự (Ordinal Encoding)
* Mã hóa danh nghĩa (Nominal Encoding)

Mã hóa thứ tự đối với các đặc trưng:

|  |  |
| --- | --- |
| * ExterQual | * HeatingQC |
| * ExterCond | * KitchenQual |
| * BsmtQual | * FireplaceQu |
| * BsmtCond | * GarageQual |
| * BsmtExposure | * GarageCond |
| * BsmtFinType1 | * PoolQC |
| * BsmtFinType2 | * Fence |

Các giá trị dạng chuỗi của các cột sẽ được mã hóa về dạng số theo thứ tự sau:

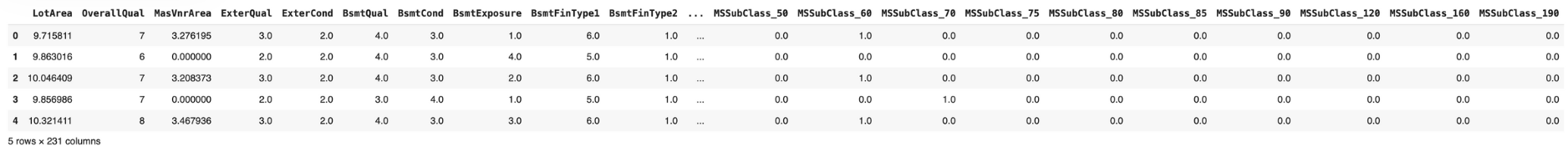
A screen shot of a computer code

Description automatically generated

Kết quả sau khi mã hóa:



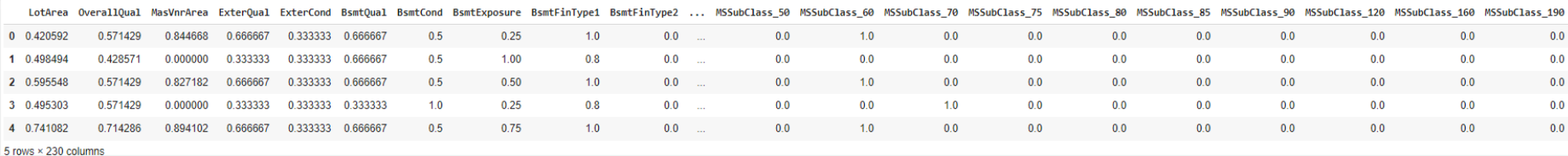
Thực hiện mã hóa danh nghĩa đối với các cột dữ liệu còn lại:



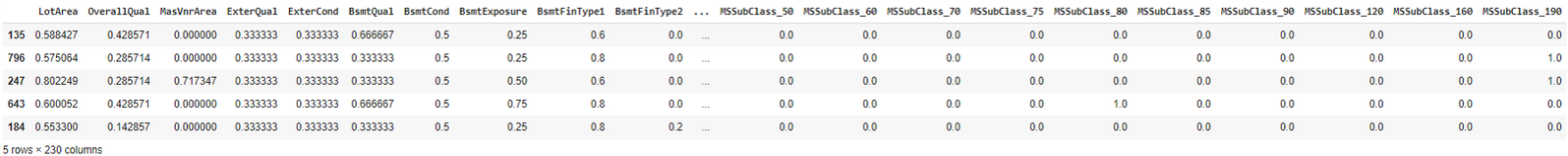
## Chuẩn hóa dữ liệu:

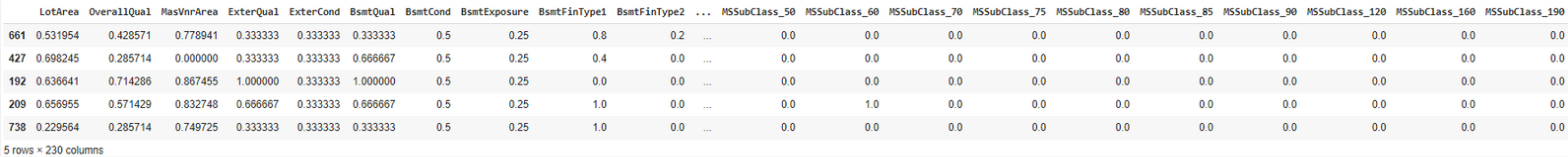
Bước này rất quan trọng để mô hình có thể hội tụ nhanh hơn và ngăn chặn mô hình bị nghiêng về các đặc trưng có độ lớn cao hơn.

Chuẩn hóa dữ liệu bằng cách đưa tất cả các giá trị về khoảng 0 đến 1:

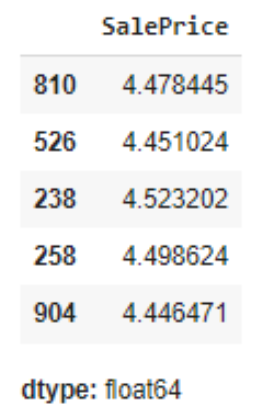


Chia tập dữ liệu đầu vào thành 2 tập **train** và **test**:





Tập y\_train,y\_test:



## Mô hình hóa:

Huấn luyện bằng nhiều mô hình khác nhau:

Linear Regression Model (LR): Huấn luyện sử dụng mô hình hồi quy tuyến tính với một đặc trưng duy nhất là OverallQual.

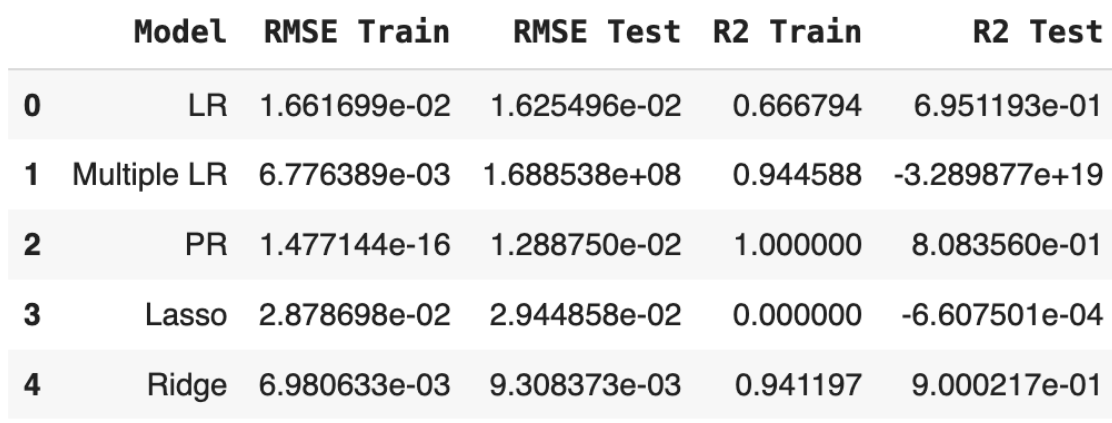
Multivariable Linear Regression (Multiple LR): Huấn luyện sử dụng hồi quy tuyến tính với tất cả các đặc trưng.

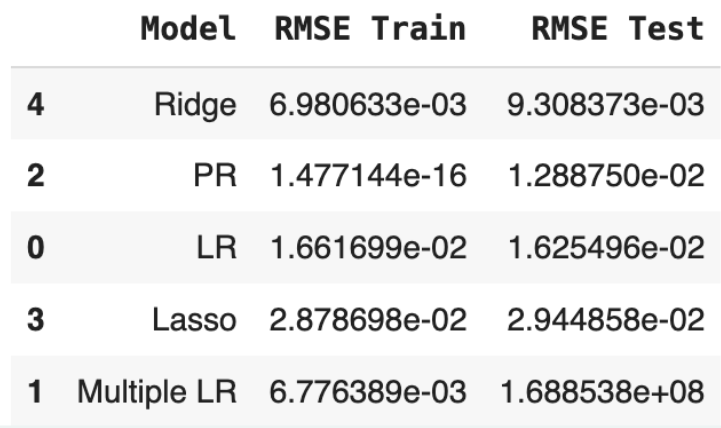
Polynomial Regression (PR):Huấn luyện sử dụng mô hình hồi quy tuyến tính với các đặc trưng được biến đổi thành các đặc trưng đa thức.

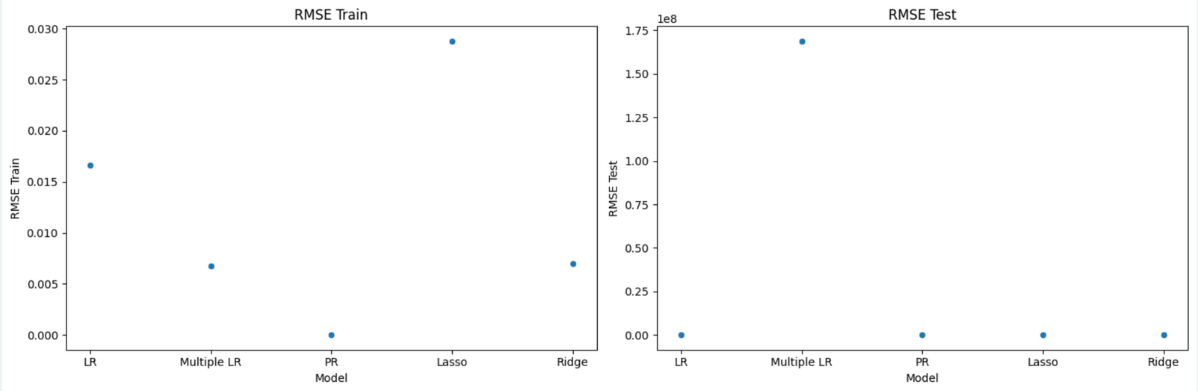
Lasso Regression (Lasso).

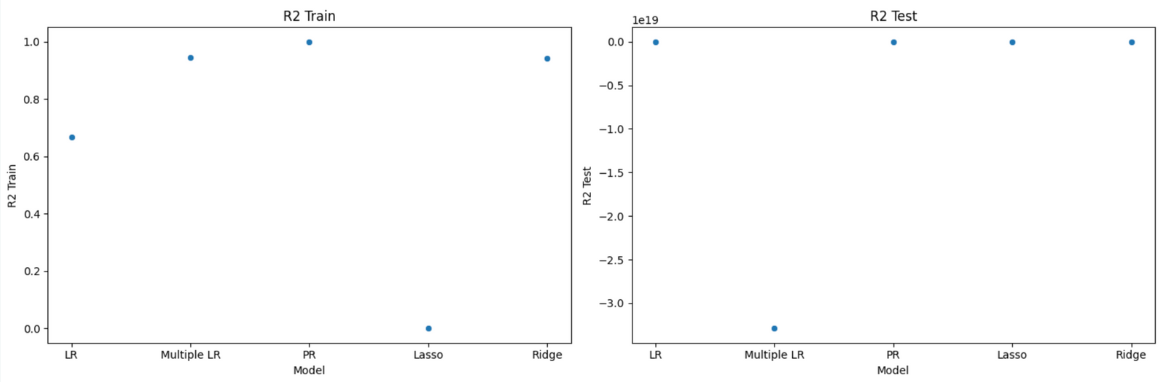
Ridge Regression (Ridge).

## Đánh giá:



****

****

****

Linear Regression (LR):

RMSE Train: 0.0166, RMSE Test: 0.0163. Mô hình có sai số khá thấp ở cả tập huấn luyện và kiểm tra.

R² Train: 0.667, R² Test: 0.695. R² cho thấy mô hình này giải thích được khoảng 66.7% biến thiên trong dữ liệu huấn luyện và 69.5% trong dữ liệu kiểm tra, cho thấy mô hình khá ổn định.

Multiple Linear Regression (Multiple LR):

RMSE Train: 0.0068, RMSE Test: 168,853,800. Đây là sự chênh lệch lớn giữa RMSE trên tập huấn luyện và kiểm tra, cho thấy mô hình có khả năng bị overfitting.

R² Train: 0.944, R² Test: -3.289877e+19. R² âm rất lớn ở tập kiểm tra là dấu hiệu rõ ràng của việc mô hình không dự đoán tốt, có thể do có outliers hoặc biến dự đoán quá nhiều làm mất tính ổn định của mô hình.

Polynomial Regression (PR):

RMSE Train: 1.477144e-16 (gần bằng 0), RMSE Test: 0.0129. RMSE gần bằng 0 ở tập huấn luyện cho thấy mô hình hoàn toàn khớp với dữ liệu huấn luyện.

R² Train: 1.0, R² Test: 0.808. R² của tập huấn luyện là 1.0, cho thấy mô hình hoàn toàn khớp, nhưng có sự giảm nhẹ ở tập kiểm tra (R² = 0.808), cho thấy mô hình có thể hơi overfitting nhưng vẫn dự đoán tốt trên tập kiểm tra.

Lasso Regression:

RMSE Train: 0.0288, RMSE Test: 0.0294. Sai số tương đối cân bằng giữa hai tập.

R² Train: 0.0, R² Test: -0.0007. R² cho thấy mô hình hầu như không giải thích được biến thiên trong dữ liệu. Điều này có thể là do regularization quá mạnh khiến cho các hệ số bị triệt tiêu về 0, làm giảm tính dự đoán của mô hình.

Ridge Regression:

RMSE Train: 0.0070, RMSE Test: 0.0093. Sai số nhỏ và khá cân bằng giữa hai tập.

R² Train: 0.941, R² Test: 0.900. R² cao và ổn định giữa hai tập, cho thấy Ridge Regression có hiệu quả tốt trong việc dự đoán và không bị overfitting.

*Kết luận:*

Ridge Regression cho thấy hiệu suất tốt nhất với R² cao và RMSE nhỏ, ổn định trên cả hai tập.

Polynomial Regression cũng có hiệu quả tốt, nhưng có dấu hiệu nhẹ của overfitting.

Multiple Linear Regression cho thấy vấn đề nghiêm trọng về overfitting, dẫn đến sai số rất lớn trên tập kiểm tra.

Lasso Regression dường như bị underfitting, có thể do regularization quá mạnh.

Vì vậy, Ridge Regression là mô hình tốt nhất trong các mô hình trên.

# Các kết quả đã đạt được (chưa đạt được)

## Kết quả đạt được

**Phân chia công việc hợp lý**

Nhóm đã tổ chức phân công công việc rõ ràng, mỗi thành viên phụ trách một phần nhất định như xử lý dữ liệu, xây dựng mô hình, đánh giá kết quả, và viết báo cáo. Điều này giúp tận dụng được điểm mạnh của từng thành viên và đảm bảo tiến độ công việc.

**Nâng cao kiến thức và kỹ năng**

Qua dự án này, cả nhóm đã có cơ hội thực hành các kỹ thuật hồi quy, xử lý dữ liệu nâng cao như loại bỏ outliers và xử lý dữ liệu thiếu, cũng như học cách sử dụng công cụ như Python, các thư viện machine learning, và công cụ phân tích dữ liệu.

**Kết quả mô hình khả quan**

Nhờ sự hợp tác và đóng góp của các thành viên, nhóm đã xây dựng được một mô hình dự đoán giá nhà đạt độ chính xác tương đối cao. Ridge Regression là mô hình nổi bật nhất trong số các mô hình thử nghiệm, giúp nhóm đưa ra dự đoán khá sát với giá trị thực tế.

**Tăng cường kỹ năng giao tiếp và báo cáo**

Nhóm đã học cách trình bày và báo cáo kết quả khoa học rõ ràng, từ cách giải thích các chỉ số đánh giá đến việc so sánh hiệu suất của các mô hình khác nhau. Điều này giúp đảm bảo rằng kết quả nghiên cứu dễ hiểu và có thể áp dụng vào thực tiễn.

## Các kết quả chưa đạt được

**Khó khăn trong xử lý và tích hợp ý kiến**

Trong quá trình làm việc nhóm, đã có những lúc ý kiến của các thành viên khác nhau về cách xử lý dữ liệu, cách chọn mô hình hoặc đánh giá kết quả. Điều này đôi khi gây mất thời gian và ảnh hưởng đến tiến độ công việc.

**Thiếu kinh nghiệm trong tối ưu hóa mô hình**

Việc xử lý dữ liệu là một công việc tương đối mới mẻ với các thành viên trong nhóm nên các thành viên còn thiếu kinh nghiệm trong việc tinh chỉnh các tham số của mô hình (hyperparameter tuning), dẫn đến việc một số mô hình chưa đạt hiệu quả cao nhất. Điều này cũng có thể là nguyên nhân làm tăng sai số dự đoán.

**Chưa tối ưu hóa hiệu quả xử lý dữ liệu lớn**

Một số công đoạn xử lý dữ liệu tốn nhiều thời gian do lượng dữ liệu lớn, và nhóm chưa có kinh nghiệm tối ưu hóa hiệu quả tính toán trong những trường hợp này. Điều này dẫn đến việc mô hình đôi khi chạy chậm và gây trở ngại trong việc kiểm tra kết quả nhanh chóng.

# Những khó khăn, vướng mắc gặp phải trong quá trình thực hiện

**Khó khăn về xử lý dữ liệu**

Dữ liệu thiếu và nhiễu: Trong dữ liệu về giá nhà, thường có nhiều giá trị bị thiếu (missing values) và dữ liệu không hợp lệ (outliers). Việc xác định và xử lý những giá trị này là một thách thức lớn, vì có thể ảnh hưởng trực tiếp đến độ chính xác của mô hình.

Biến đổi dữ liệu: Một số biến số có độ lệch lớn (skewed) hoặc không tuân theo phân phối chuẩn, gây khó khăn khi áp dụng các mô hình hồi quy. Nhóm cần thực hiện các phép biến đổi như log transformation, Box-Cox để cải thiện sự phân phối của các biến, nhưng điều này yêu cầu hiểu biết sâu về phương pháp và thời gian thử nghiệm.

**Khó khăn trong việc phối hợp làm việc nhóm**

Khác biệt về kỹ năng và kinh nghiệm: Mỗi thành viên trong nhóm có mức độ hiểu biết và kinh nghiệm khác nhau về machine learning và lập trình, điều này đôi khi gây khó khăn trong việc phân chia công việc và hoàn thành đúng tiến độ.

Giao tiếp và thống nhất ý kiến: Việc thống nhất ý kiến khi gặp những lựa chọn quan trọng như chọn mô hình, cách xử lý dữ liệu hoặc cách đánh giá kết quả là một thách thức. Đôi khi, việc bất đồng ý kiến có thể làm chậm tiến độ công việc và ảnh hưởng đến tinh thần làm việc của nhóm.

**Hạn chế về kiến thức chuyên môn**

Thiếu kiến thức về hồi quy và phân tích thống kê: Một số thành viên có thể chưa nắm vững các khái niệm về hồi quy, phân phối chuẩn, outliers, hoặc kỹ thuật biến đổi dữ liệu, dẫn đến khó khăn trong việc lựa chọn phương pháp phù hợp và phân tích kết quả.

Hiểu biết về đặc thù của dữ liệu giá nhà: Để xây dựng mô hình dự đoán giá nhà tốt, nhóm cần có kiến thức về thị trường bất động sản, như các yếu tố ảnh hưởng đến giá nhà, nhằm chọn ra các đặc trưng (features) quan trọng. Việc thiếu hiểu biết này có thể làm giảm tính chính xác của mô hình.

# Những kinh nghiệm thu được sau khi hoàn thành bài tập lớn

**Kinh nghiệm về xử lý và phân tích dữ liệu**

Kỹ năng xử lý dữ liệu thiếu và bất thường: Nhóm đã học được cách xác định và xử lý các giá trị thiếu (missing values) và outliers, chẳng hạn như sử dụng trung vị (median) để điền dữ liệu thiếu khi dữ liệu có sự phân phối lệch. Kỹ năng này rất quan trọng để đảm bảo dữ liệu sạch, giúp cải thiện độ chính xác của mô hình.

Biến đổi và chuẩn hóa dữ liệu: Kinh nghiệm trong việc áp dụng các phép biến đổi như log hoặc Box-Cox để làm cho dữ liệu tuân theo phân phối chuẩn. Điều này giúp nhóm hiểu hơn về sự ảnh hưởng của phân phối dữ liệu đến mô hình và hiệu suất dự đoán.

**Kinh nghiệm về xây dựng và tối ưu hóa mô hình**

Hiểu rõ ưu và nhược điểm của các mô hình hồi quy: Nhóm đã hiểu sâu hơn về các mô hình hồi quy như Linear Regression, Polynomial Regression, Lasso, và Ridge Regression. Qua việc thử nghiệm và đánh giá các mô hình này, nhóm nắm bắt được khi nào mô hình bị overfitting hoặc underfitting, và biết cách điều chỉnh tham số để tối ưu hóa mô hình.

**3. Kinh nghiệm làm việc nhóm và quản lý dự án**

Phân chia công việc và tận dụng kỹ năng từng thành viên: Qua đề tài này, nhóm học được cách phân chia công việc hiệu quả dựa trên điểm mạnh và kỹ năng của từng thành viên. Điều này không chỉ giúp hoàn thành công việc đúng tiến độ mà còn cải thiện chất lượng của từng phần công việc.

Giao tiếp và hợp tác: Nhóm đã rút ra bài học về tầm quan trọng của giao tiếp liên tục và rõ ràng trong việc thống nhất ý kiến và giải quyết xung đột. Việc học cách lắng nghe và tôn trọng ý kiến của từng thành viên giúp cải thiện không khí làm việc và tăng cường tinh thần làm việc nhóm.

**Kinh nghiệm sử dụng công cụ và thư viện**

Sử dụng thư viện và công cụ lập trình: Nhóm đã làm quen và thành thạo hơn trong việc sử dụng các thư viện như Pandas, Numpy, Scikit-learn để xử lý dữ liệu và xây dựng mô hình. Nhóm cũng học cách sử dụng Matplotlib và Seaborn để trực quan hóa dữ liệu, giúp hiểu rõ hơn về các đặc tính của dữ liệu.

Làm quen với môi trường làm việc và tài nguyên hạn chế: Khi phải làm việc với tài nguyên phần cứng giới hạn hoặc không gian lưu trữ dữ liệu hạn chế, nhóm đã học cách tối ưu hóa mã và sử dụng tài nguyên hợp lý, đồng thời chuẩn bị sẵn sàng cho các hạn chế khi làm việc với dữ liệu lớn.

**Hiểu biết sâu hơn về dữ liệu bất động sản**

Xác định yếu tố ảnh hưởng đến giá nhà: Nhóm đã hiểu rõ hơn về các yếu tố quan trọng ảnh hưởng đến giá nhà, như diện tích, chất lượng xây dựng, và năm xây dựng. Kiến thức này giúp nhóm xác định và chọn lọc các đặc trưng (features) có giá trị để cải thiện độ chính xác của mô hình.

Hiểu về tính phi tuyến và sự phức tạp của dữ liệu giá nhà: Qua việc phân tích dữ liệu giá nhà, nhóm nhận thấy tính phi tuyến của một số biến số, và hiểu rằng không phải mọi quan hệ đều có thể được giải thích bằng hồi quy tuyến tính. Điều này giúp nhóm có cách nhìn sâu sắc hơn khi chọn mô hình và thực hiện phân tích.

# Kết luận (và tự đánh giá theo thang điểm từ 0 đến 10)

Qua quá trình nghiên cứu và thực hiện đề tài dự đoán giá nhà bằng các phương pháp hồi quy, nhóm đã có được những hiểu biết sâu sắc hơn về quy trình phân tích và xử lý dữ liệu, cũng như áp dụng thành công các mô hình học máy để giải quyết bài toán thực tế. Việc so sánh các phương pháp hồi quy tuyến tính, hồi quy đa biến, hồi quy bậc hai, Lasso và Ridge giúp nhóm hiểu rõ hơn về ưu, nhược điểm của từng mô hình, đặc biệt trong việc tránh overfitting và underfitting.

Kết quả thu được cho thấy một số mô hình có khả năng dự đoán tốt với độ chính xác cao, nhưng cũng có những mô hình chưa đạt kết quả như mong đợi do một số yếu tố như outliers, đa cộng tuyến và phân phối dữ liệu không chuẩn. Nhóm đã tìm ra các phương pháp xử lý hiệu quả, bao gồm điều chỉnh đặc trưng, biến đổi dữ liệu và áp dụng kỹ thuật Regularization, giúp nâng cao chất lượng mô hình.

Nhìn chung, nhóm đã hoàn thành tốt các mục tiêu đề ra, tích lũy được nhiều kinh nghiệm bổ ích và có khả năng tự tin hơn khi thực hiện các dự án phân tích dữ liệu trong tương lai.

**Nhóm tự đánh giá**

Hiểu biết về dữ liệu và xử lý dữ liệu: 9/10

Nhóm đã có sự chuẩn bị tốt trong việc phân tích và làm sạch dữ liệu, xử lý các giá trị ngoại lai, lấp đầy dữ liệu thiếu và thực hiện các biến đổi cần thiết, giúp mô hình hoạt động ổn định hơn.

Kỹ năng xây dựng và tối ưu hóa mô hình: 9/10

Các mô hình đã được tối ưu và so sánh kỹ lưỡng, đặc biệt là việc áp dụng Box-Cox để điều chỉnh phân phối và sử dụng Regularization giúp giảm thiểu overfitting, cho thấy nhóm có sự hiểu biết sâu sắc trong tối ưu hóa mô hình.

Hiểu biết về lý thuyết hồi quy và học máy: 10/10

Nhóm nắm vững lý thuyết và các chỉ số đánh giá hiệu quả mô hình như RMSE và R-squared, giúp giải thích kết quả dự đoán rõ ràng và thuyết phục hơn.

Kỹ năng làm việc nhóm và phân chia công việc: 10/10

• Phân chia công việc hiệu quả, đảm bảo tiến độ và phối hợp nhịp nhàng là một trong những điểm mạnh của nhóm. Mỗi thành viên đều đóng góp tích cực vào từng phần của đề tài.

Khả năng trình bày và báo cáo: 9/10

Báo cáo và trình bày được chuẩn bị kỹ càng, có hệ thống, giúp người xem dễ hiểu và theo dõi quá trình thực hiện. Các biểu đồ, hình ảnh minh họa được sử dụng phù hợp, tạo ra báo cáo hấp dẫn.

**Tổng điểm tự đánh giá: 9.5/10**

Nhóm tự tin rằng những nỗ lực và kết quả đạt được xứng đáng với thang điểm 9.5/10. Đề tài này không chỉ giúp nhóm nắm vững kiến thức về hồi quy mà còn cải thiện kỹ năng làm việc nhóm và khả năng giải quyết các bài toán thực tế. Nhóm tin tưởng rằng những kinh nghiệm này sẽ là nền tảng vững chắc cho các dự án trong tương lai.

# Tài liệu tham khảo

Papiu, A. (2023). *Regularized Linear Models*. Retrieved from https://www.kaggle.com/code/apapiu/regularized-linear-models

Gusthema;Kin. (2023). *House Prices Prediction using TFDF*. Retrieved from https://www.kaggle.com/code/gusthema/house-prices-prediction-using-tfdf