**Bài tập thực hành 04:**

Tiền xử lý, trực quan hóa dữ liệu

1. **Tiền xử lý dữ liệu là gì[[1]](#footnote-1)? Các ứng dụng của nó. Những lợi ích và thách thức của việc tiền xử lý dữ liệu.**

**Bài làm**

Tiền xử lý dữ liệu là quá trình biến đổi dữ liệu thô thành dữ liệu sạch và phù hợp để phục vụ cho các mục đích khác nhau, như phân tích, học máy, khai thác dữ liệu, hoặc trình bày. Tiền xử lý dữ liệu có nhiều ứng dụng trong các lĩnh vực liên quan đến cơ sở dữ liệu, như quản lý quan hệ khách hàng, quảng cáo trực tuyến, y tế, giáo dục, và nhiều hơn nữa. Tiền xử lý dữ liệu có nhiều lợi ích, như:

* Cải thiện chất lượng và độ tin cậy của dữ liệu, bằng cách loại bỏ hoặc sửa chữa các giá trị bị thiếu, sai lệch, nhiễu, hoặc không nhất quán.
* Tăng hiệu quả và độ chính xác của các thuật toán học máy và khai thác dữ liệu, bằng cách giảm kích thước và độ phức tạp của dữ liệu, và tăng tính liên quan và đồng nhất của các đặc trưng.
* Tạo ra các biểu diễn dữ liệu mới và phong phú hơn, bằng cách kết hợp, biến đổi, hoặc tạo ra các đặc trưng mới từ các đặc trưng ban đầu.
* Tăng khả năng hiểu và diễn giải dữ liệu, bằng cách sử dụng các kỹ thuật trực quan hóa, thống kê mô tả, hoặc phân tích thành phần chính.

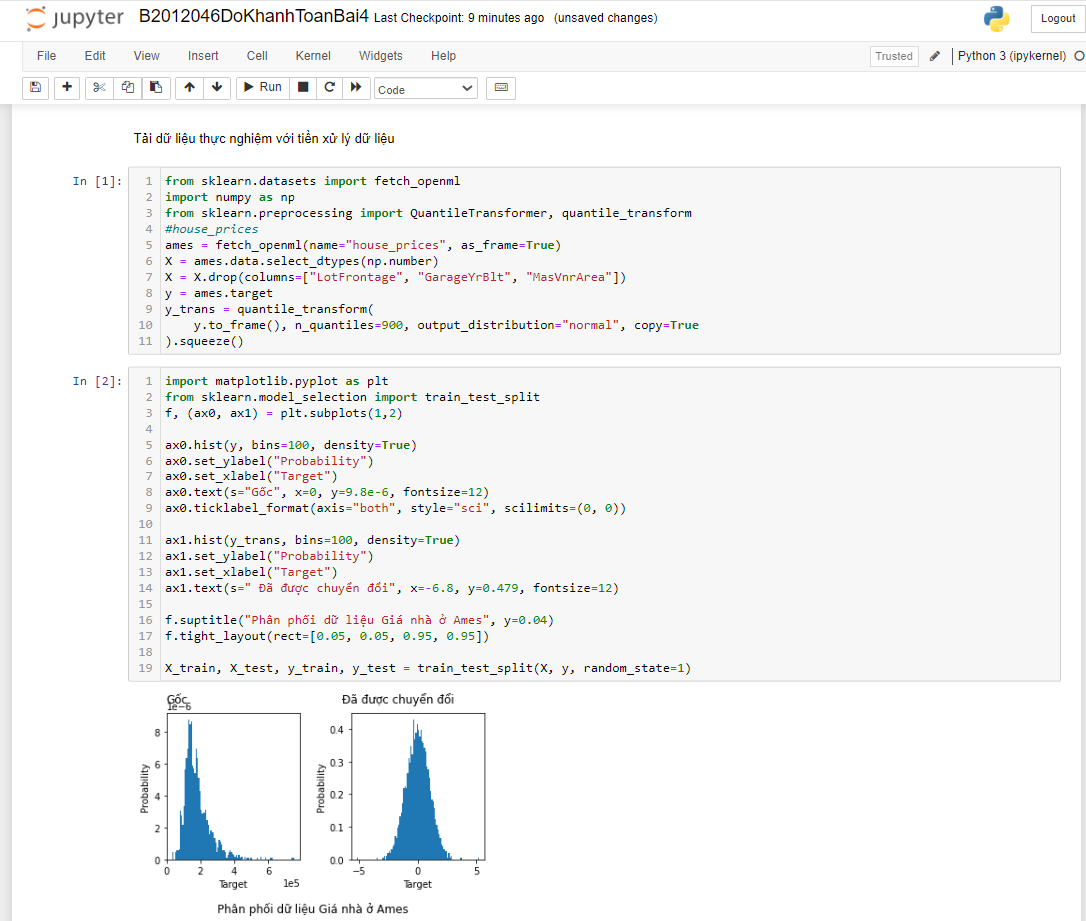
Tuy nhiên, tiền xử lý dữ liệu cũng gặp phải nhiều thách thức, như:

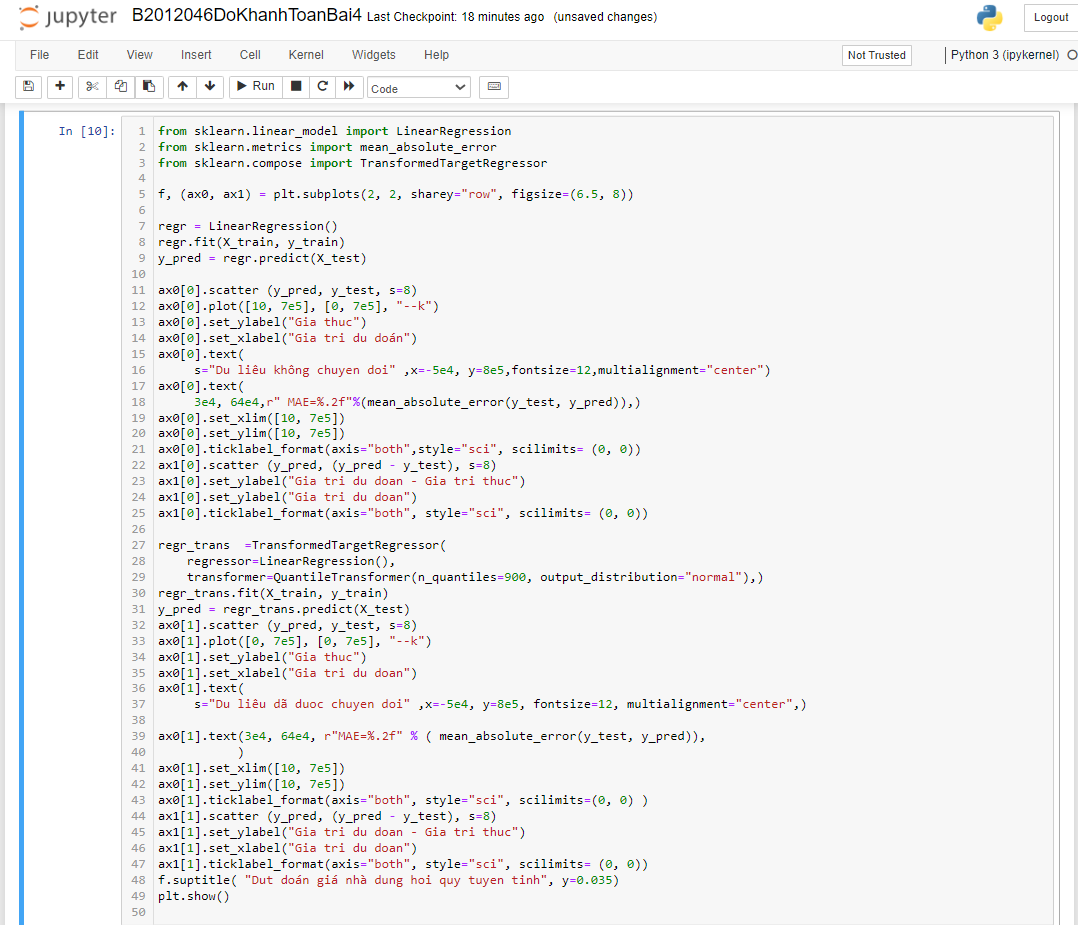
* Khó khăn trong việc xác định và xử lý các vấn đề về chất lượng dữ liệu, do sự thiếu sót của nguồn dữ liệu, công cụ thu thập dữ liệu, hoặc quy trình nhập dữ liệu.
* Khó khăn trong việc lựa chọn và áp dụng các kỹ thuật tiền xử lý dữ liệu phù hợp với từng loại dữ liệu và mục tiêu sử dụng, do sự đa dạng và phức tạp của các bộ dữ liệu và các bài toán trong thực tế.
* Khó khăn trong việc đánh giá hiệu quả và ảnh hưởng của các kỹ thuật tiền xử lý dữ liệu đến kết quả cuối cùng, do sự thiếu hụt của các tiêu chuẩn đo lường và kiểm tra.

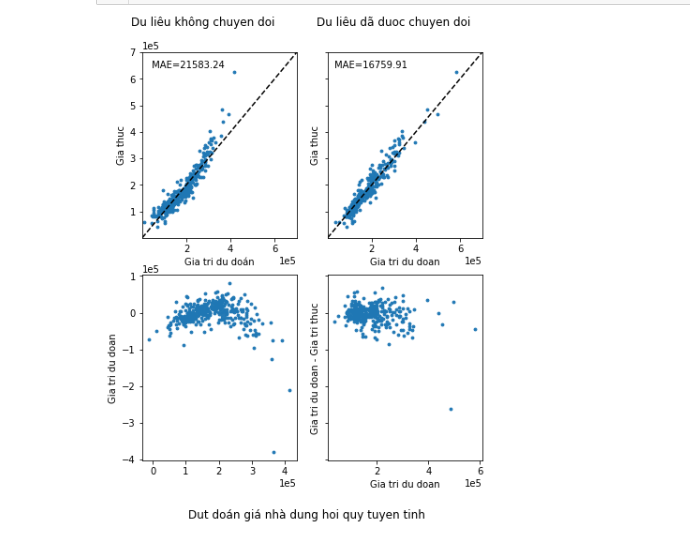
**Link tham khảo:**

1. [**https://www.vimentor.com/vi/lesson/tien-xu-ly-du-lieu-trong-linh-vuc-hoc-may-phan-1**](https://www.vimentor.com/vi/lesson/tien-xu-ly-du-lieu-trong-linh-vuc-hoc-may-phan-1)
2. [**https://websitehcm.com/tong-quan-ve-tien-xu-ly-du-lieu/**](https://websitehcm.com/tong-quan-ve-tien-xu-ly-du-lieu/)
3. [**https://vi.theastrologypage.com/data-preprocessing**](https://vi.theastrologypage.com/data-preprocessing)
4. **Thực hiện lại các hướng dẫn từ trang 84 đến 89.**

**Bài làm**

****

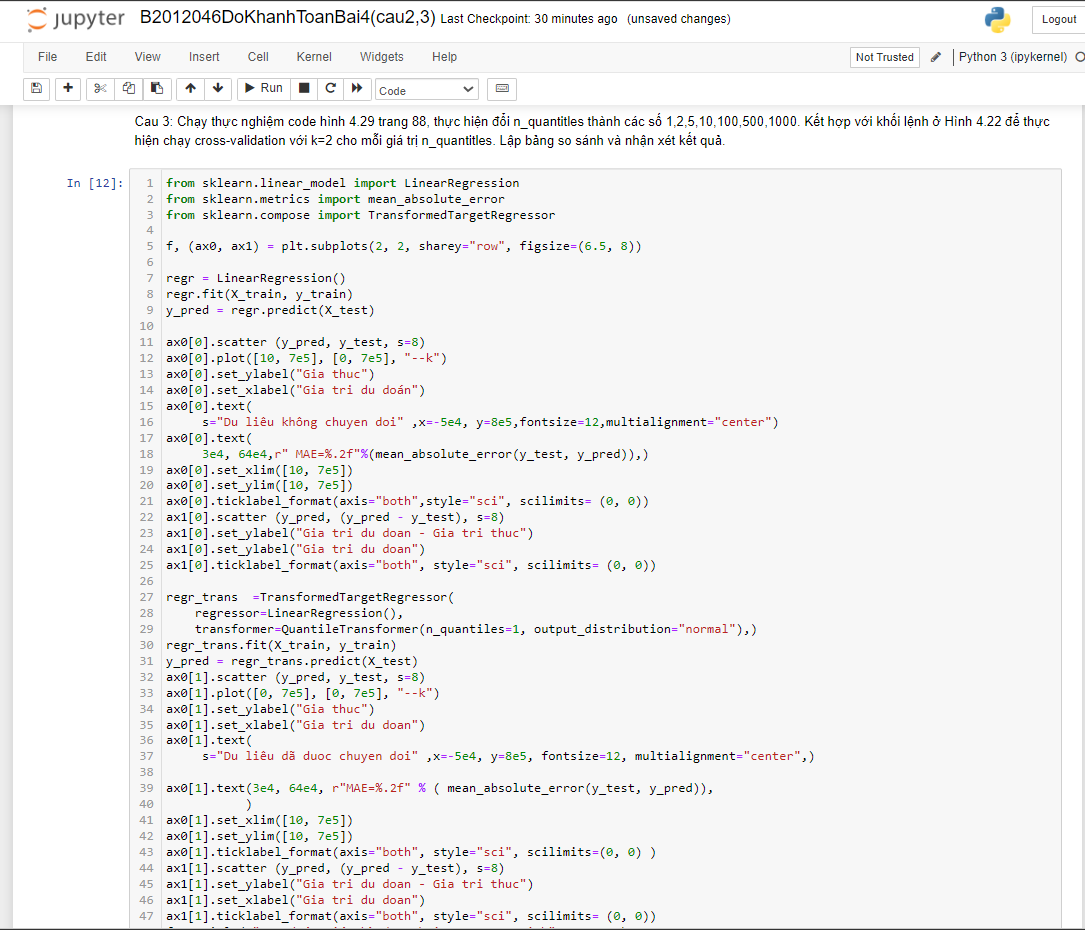
****

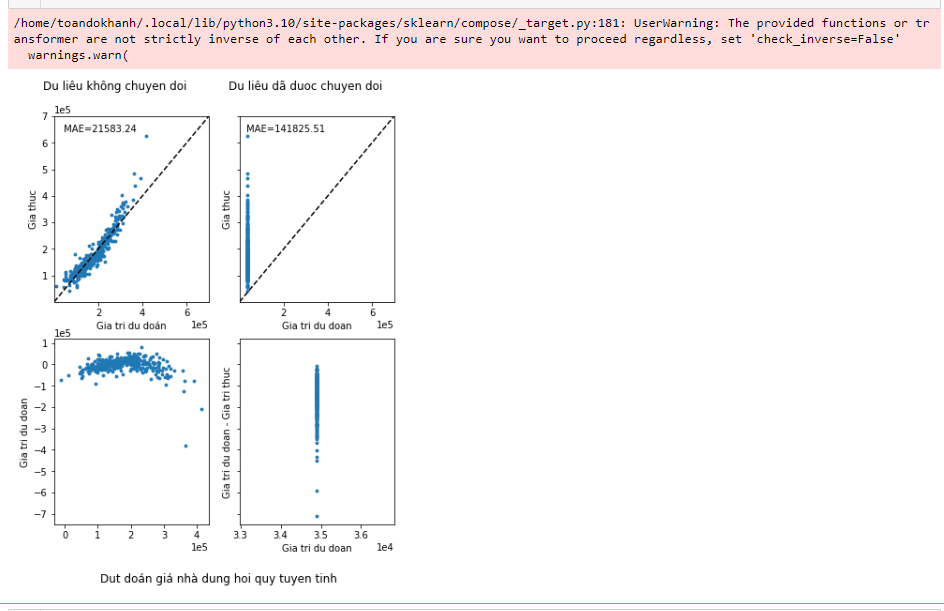
****

1. **Chạy thực nghiệm code hình 4.29 trang 88, thực hiện đổi n\_quantitles thành các số 1,2,5,10,100,500,1000. Kết hợp với khối lệnh ở Hình 4.22 để thực hiện chạy cross-validation với k=2 cho mỗi giá trị n\_quantitles. Lập bảng so sánh và nhận xét kết quả.**

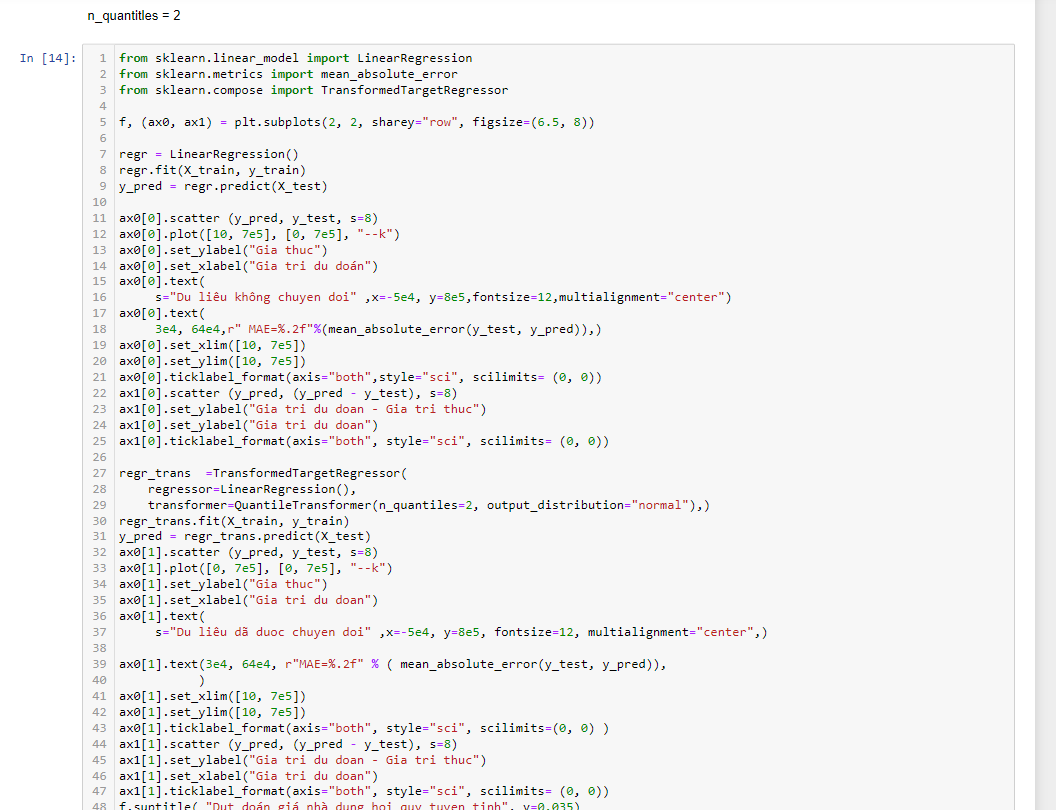
**Bài làm**

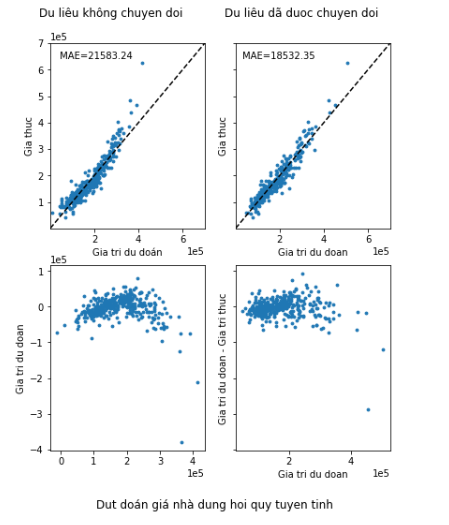
**N\_quantitles = 1**

****

****

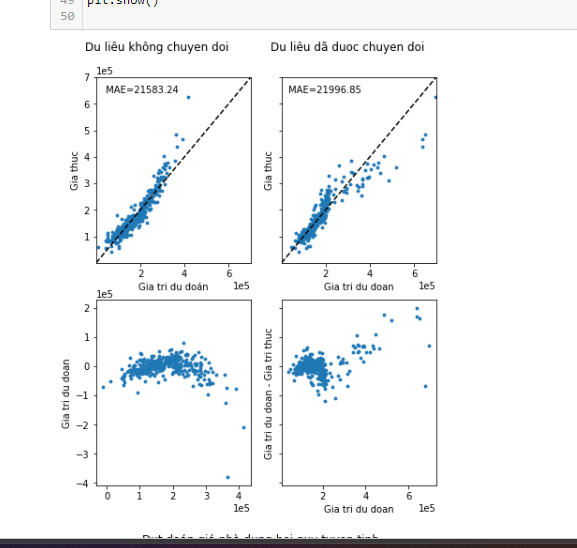
**N\_quantitles = 2**

****

****

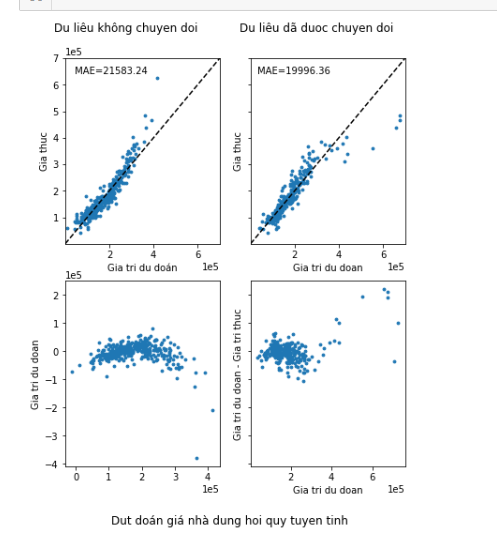
**N\_quantitles = 5**

****

****

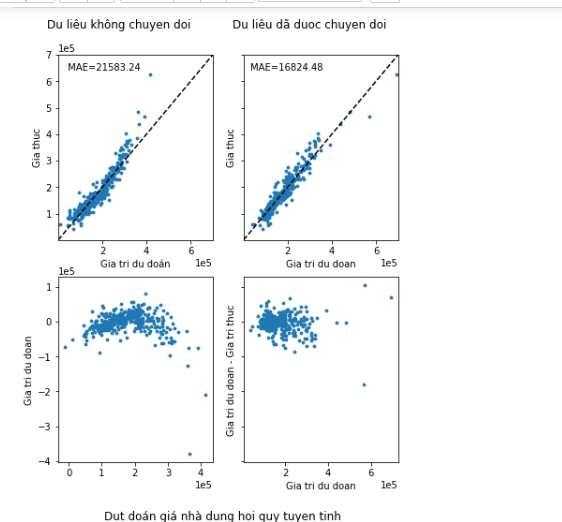
**N\_quantitles = 10**

****

****

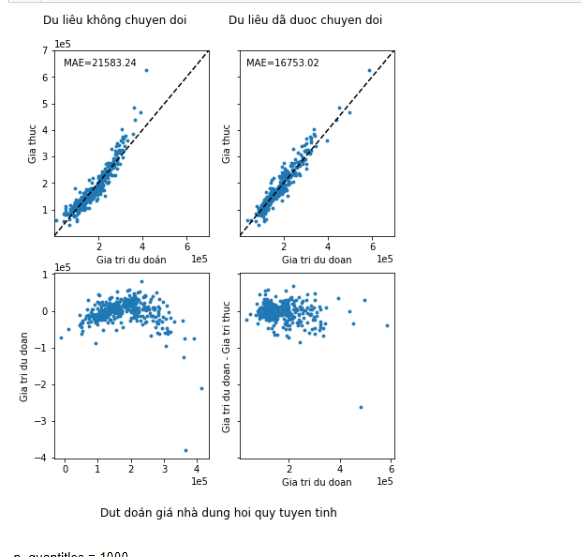
**N\_quantitles =100**

****

****

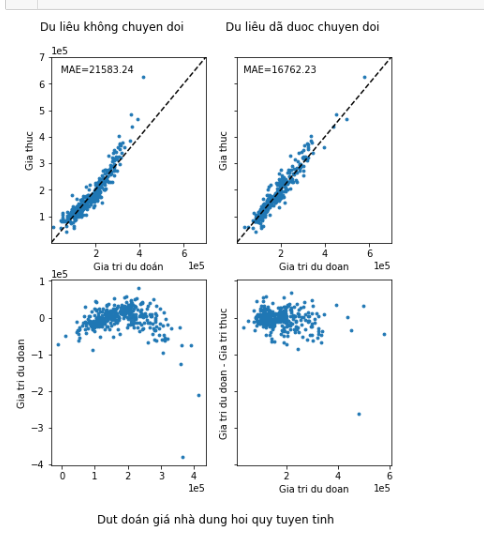
**N\_quantitles = 500**

****

****

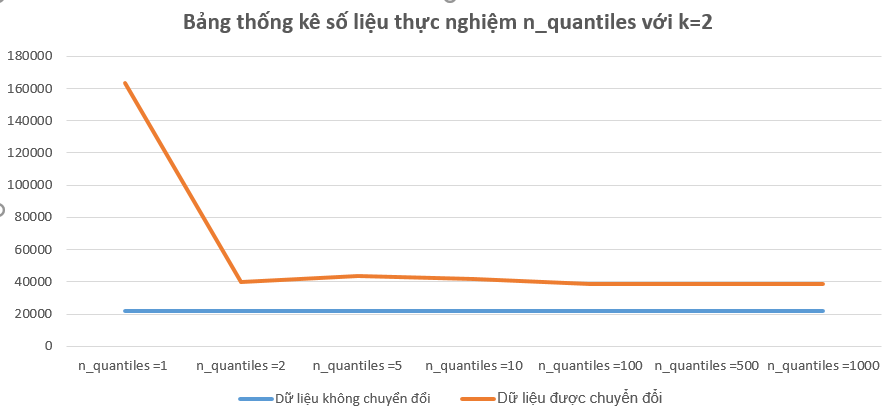
**N\_quantitles = 1000**

****

****

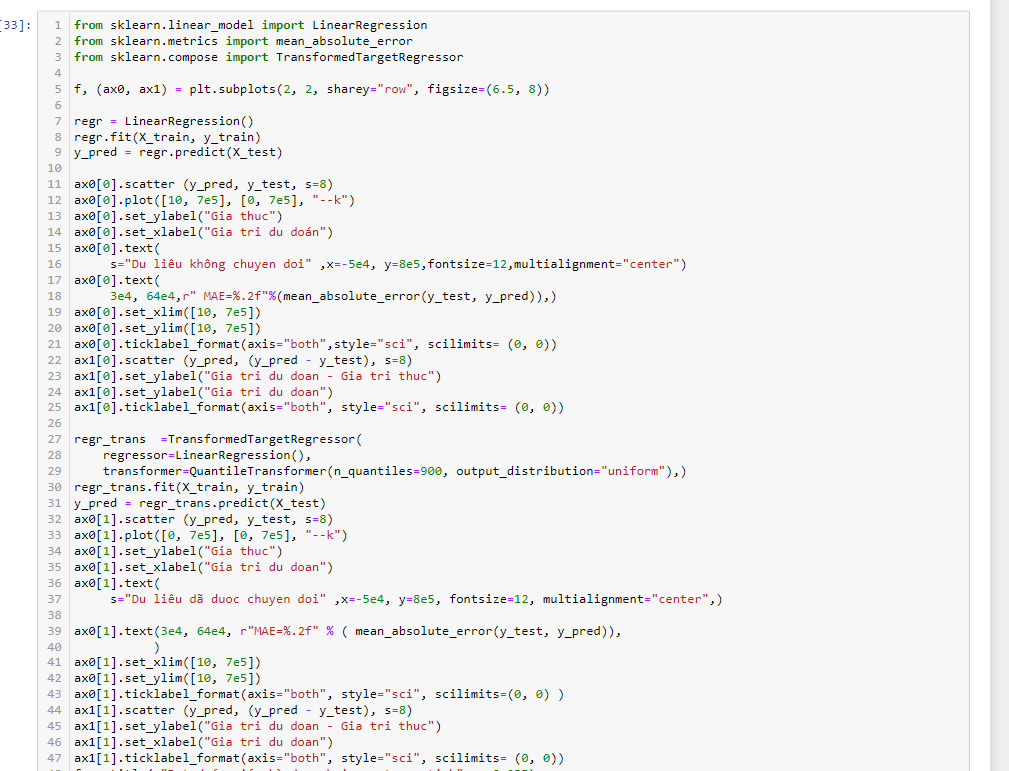
**Nhận xét:**

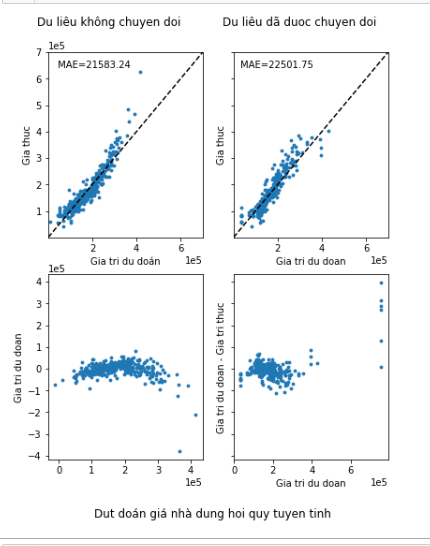
sau khi tôi chạy thực hiện và thu được kết quả như hình sau



1. **Chạy thực nghiệm code hình 4.29 trang 88, thực hiện đổi output\_distribution thành *uniform*. Kết hợp với khối lệnh ở Hình 4.22 để thực hiện chạy cross-validation với k=2 để thể hiện các độ đo accuracy, f1-score, recall. Lập bảng so sánh và nhận xét kết quả.**

**Bài làm**

****

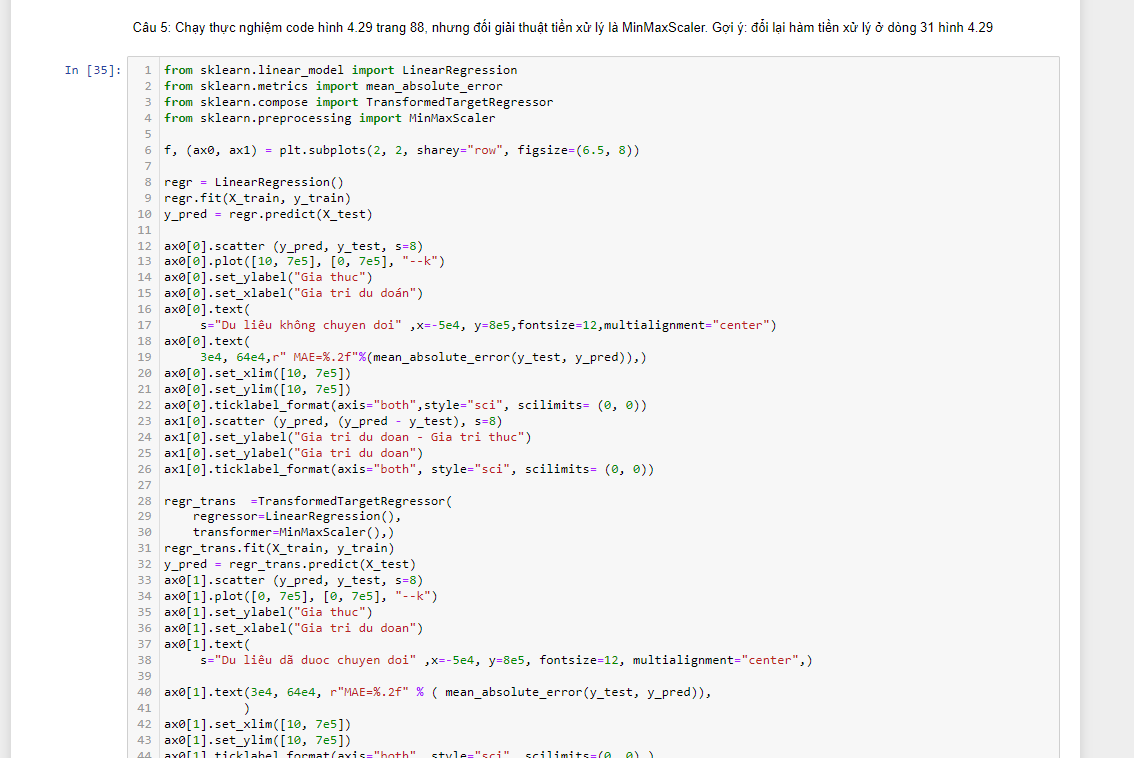
****

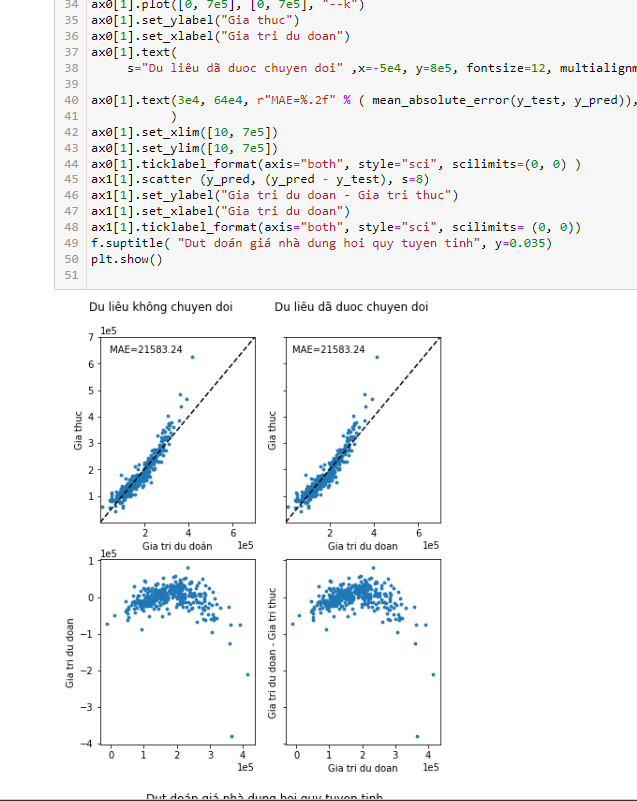
**Nhận xét:**

Sau khi chạy thử nghiệm, kết quả nhận thấy là như nhau,nhưng tốc độ xử lý về thời gan thu nhận được ở phần này khả quan hơn phần trước. Thực nghiệm cho thấy tốc độ xử lý của phần tiết kiệm được thời gian hơn.

1. **Chạy thực nghiệm code hình 4.29 trang 88, nhưng đối giải thuật tiền xử lý là MinMaxScaler[[2]](#footnote-2). Gợi ý: đổi lại hàm tiền xử lý ở dòng 31 hình 4.29**

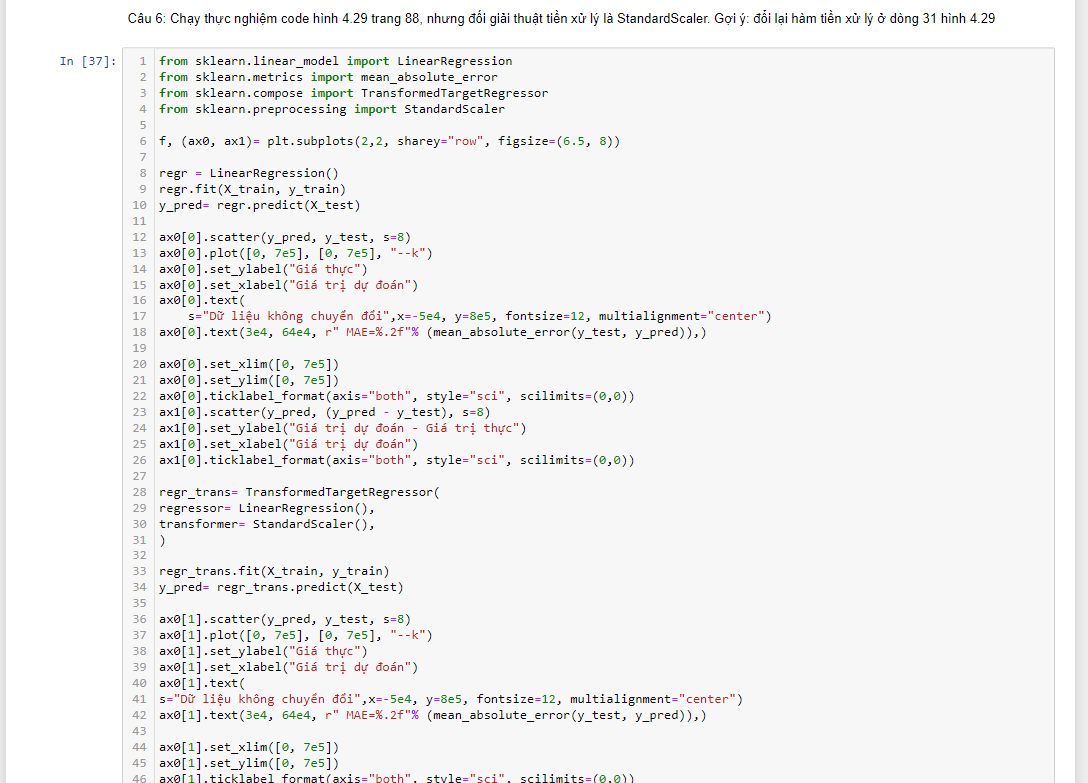
**Bài làm**

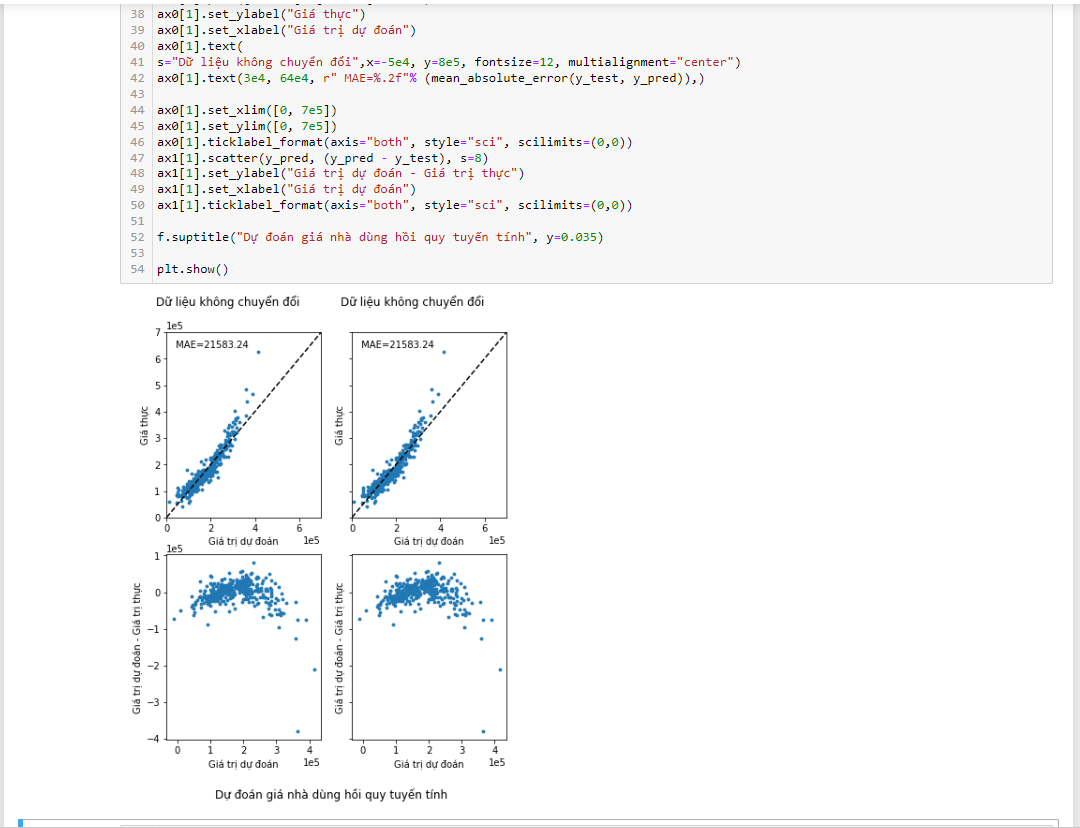
****

****

1. **Chạy thực nghiệm code hình 4.29 trang 88, nhưng đối giải thuật tiền xử lý là StandardScaler[[3]](#footnote-3). Gợi ý: đổi lại hàm tiền xử lý ở dòng 31 hình 4.29**

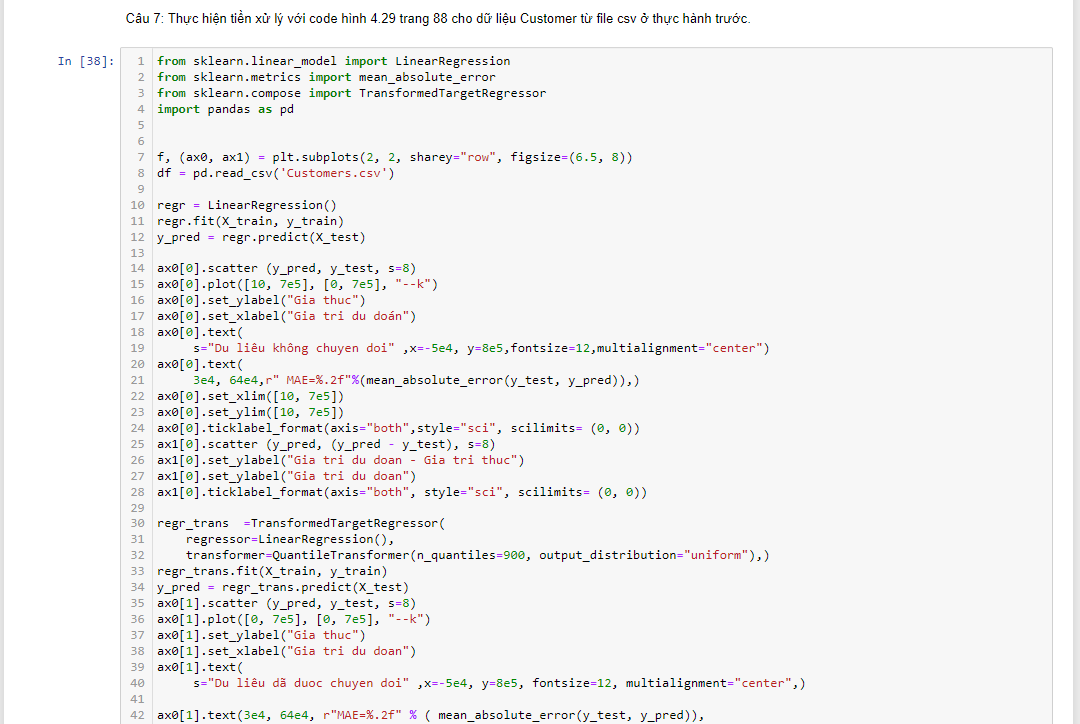
**Bài làm**

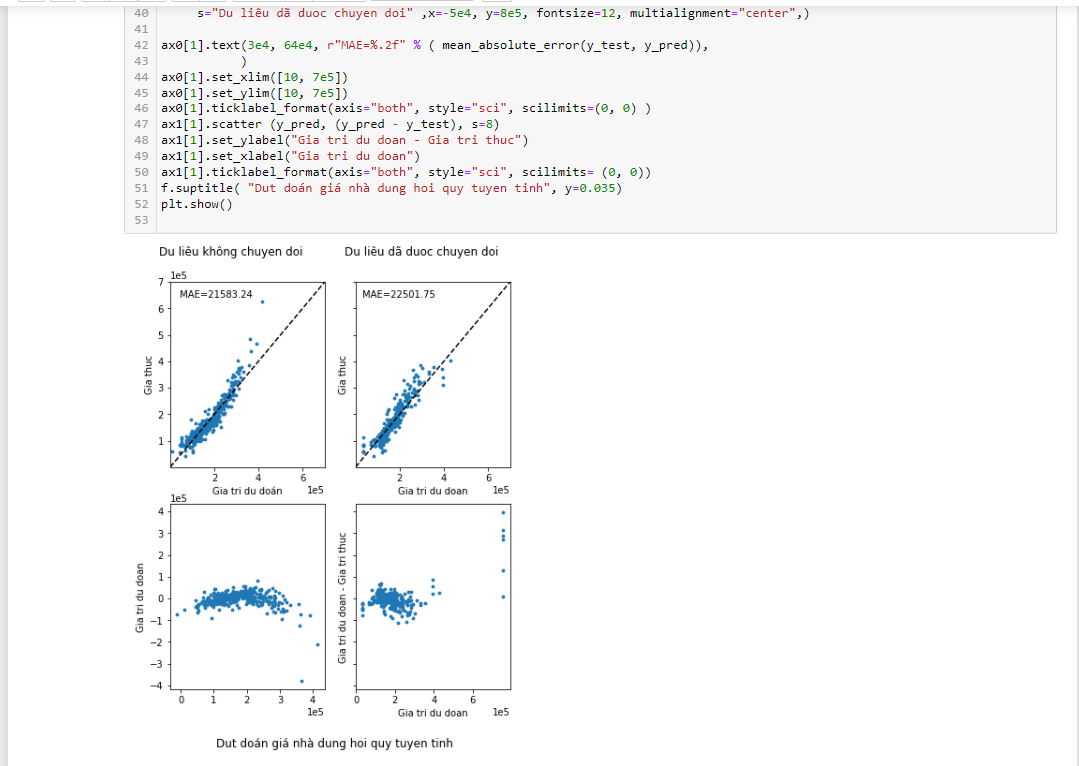
****

****

1. **Thực hiện tiền xử lý với code hình 4.29 trang 88 cho dữ liệu Customer từ file csv ở thực hành trước.**

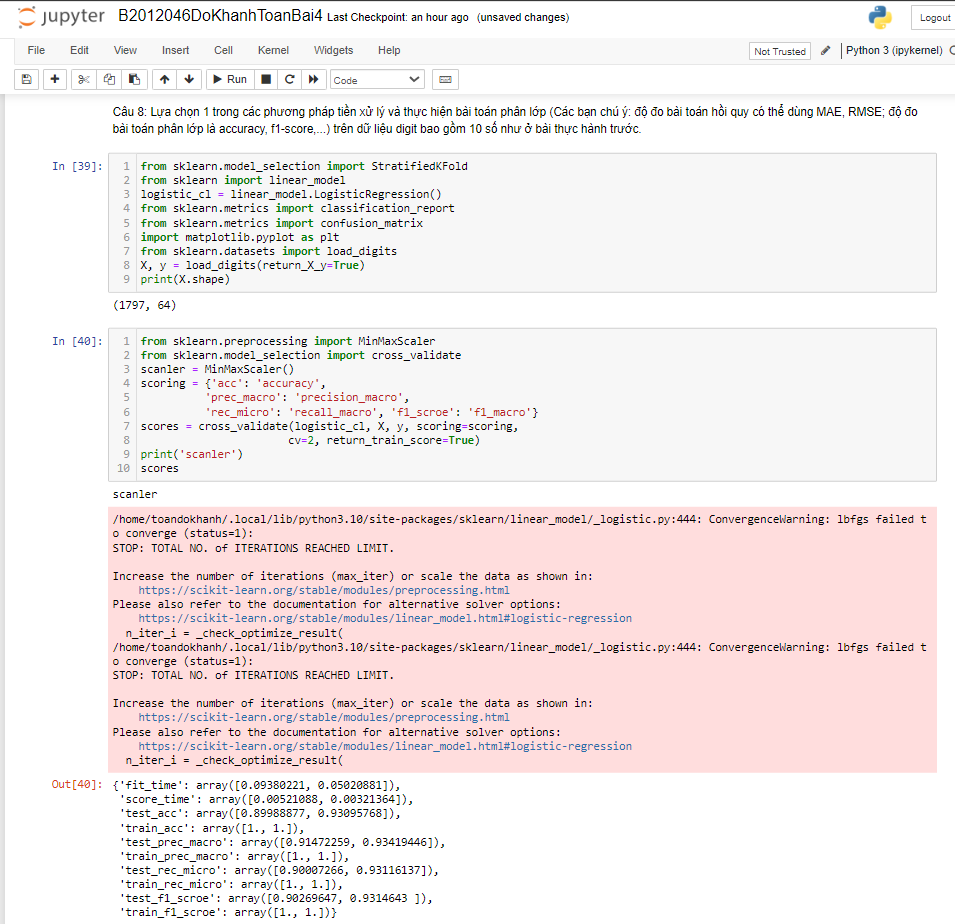
**Bài làm**

****

****

1. **Lựa chọn 1 trong các phương pháp tiền xử lý và thực hiện bài toán phân lớp (Các bạn chú ý: độ đo bài toán hồi quy có thể dùng MAE, RMSE; độ đo bài toán phân lớp là accuracy, f1-score,...) trên dữ liệu digit bao gồm 10 số như ở bài thực hành trước.**

**Bài làm**



**Chú ý:**

* Các bạn nộp file word: Quy tắc đặt tên file: **<mssv>-<hoten>-<bai><stt\_bai thực hành>.docx** nộp lên Classroom (VD: **B123456-NguyenVanA-bai1.docx**), kèm với các file khác được yêu cầu như phần câu hỏi đã nêu. **Ngoại trừ file word trả lời câu hỏi, các file còn lại các bạn nén vào 1 file zip**. File zip đặt tên như file word.
* Mỗi câu các bạn trả lời bằng hình hoặc dạng text tùy vào yêu cầu của câu hỏi và **TRẢ LỜI THEO ĐÚNG THỨ TỰ CÂU HỎI**. Nếu câu nào không trả lời được các bạn cứ để số thứ tự câu hỏi và bỏ trống phần trả lời.
* Các câu trả lời có tham khảo trên Internet phải trích dẫn link/nguồn.
* *Vi phạm 1 trong các điều sau đây bài thực hành sẽ bị 0 điểm:*
  + Đặt tên KHÔNG ĐÚNG quy tắc được yêu cầu.
  + Bài không đủ các thành phần (word, code+data (nếu có),...) đã được yêu cầu.
  + Bài không thực hiện đúng yêu cầu “**Ngoại trừ file word trả lời câu hỏi, các file còn lại các bạn nén vào 1 file .zip”**
  + Bị phát hiện copy, sao chép từ các bạn khác
  + Phần trả lời không ghi rõ trả lời cho câu nào
  + Thứ tự câu trả lời không đúng thứ tự câu hỏi

1. https://www.sciencedirect.com/topics/computer-science/regression-task [↑](#footnote-ref-1)
2. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html [↑](#footnote-ref-2)
3. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html [↑](#footnote-ref-3)