Các bước trong code:

1. Import các thư viện numpy và pandas để sử dụng trong xử lý dữ liệu.
2. Load dữ liệu từ file nhaphoc1.csv vào dataframe df\_nhaphoc bằng phương thức read\_csv () của pandas.
3. Chọn các đặc trưng (features) và biến phụ thuộc (target) từ dataframe df\_nhaphoc.
4. Áp dụng phương pháp giảm kích thước dữ liệu bằng cách sử dụng phân rã giá trị đơn lẻ (SVD) để chiếu dữ liệu đó vào không gian có chiều thấp hơn. Sử dụng PCA để thực hiện việc giảm kích thước.
5. Chia dữ liệu thành tập huấn luyện và tập kiểm tra bằng phương thức train\_test\_split() của sklearn.
6. Sử dụng hồi quy tuyến tính để huấn luyện mô hình trên tập huấn luyện.
7. Sử dụng mô hình huấn luyện để dự đoán trên tập kiểm tra.
8. Tính toán các hệ số và sai số của mô hình.
9. Vẽ đồ thị để biểu diễn dữ liệu và đánh giá mức độ lỗi của mô hình.

Lưu ý rằng, giá trị của "y" trong code này là điểm GRE của sinh viên, và các thuộc tính được sử dụng để dự đoán điểm GRE bao gồm TOEFL Score, University Rating, SOP, LOR, CGPA, và Chance of Admit.

1. **Serial No**.: Đây là số thứ tự của mỗi bản ghi trong tập dữ liệu. Nó không có ý nghĩa trong việc dự đoán khả năng nhập học.
2. **GRE Score:** Điểm số của bài kiểm tra GRE (Graduate Record Examination). Điểm số này thường được sử dụng để đánh giá khả năng tiếp thu và thành công trong chương trình học.
3. **TOEFL Score**: Điểm số của bài kiểm tra TOEFL (Test of English as a Foreign Language). Điểm số này thường được sử dụng để đánh giá khả năng sử dụng tiếng Anh trong học tập.
4. **University Rating**: Đây là một chỉ số đánh giá chất lượng của trường đại học mà sinh viên đã tốt nghiệp hoặc đang học tập. Chỉ số này được xác định bằng cách khảo sát các chuyên gia trong lĩnh vực giáo dục.
5. **Statement of Purpose** (SOP): Đây là một bài luận viết bởi sinh viên để giới thiệu về bản thân, kinh nghiệm học tập và các mục tiêu nghề nghiệp. Bài luận này thường được sử dụng để đánh giá sự nghiêm túc và quyết tâm của sinh viên.
6. **Letter of Recommendation** (LOR): Đây là thư giới thiệu của người giáo viên, giáo sư hoặc người có thẩm quyền khác về kinh nghiệm học tập và khả năng của sinh viên. Thư giới thiệu này thường được sử dụng để đánh giá khả năng học tập và tiềm năng nghiên cứu của sinh viên.
7. **CGPA:** Điểm trung bình tích luỹ của sinh viên trong quá trình học tập ở trường đại học.
8. **Research:** Chỉ số này cho biết liệu sinh viên có kinh nghiệm nghiên cứu khoa học hay không. Kinh nghiệm này thường được xem là một yếu tố quan trọng trong việc xét tuyển vào chương trình tiến sĩ.
9. **Chance of Admit**: Đây là xác suất của sinh viên được nhận vào chương trình tiến sĩ hoặc thạc sĩ. Thuộc tính này thường được sử dụng để dự đoán khả năng nhập học của một sinh viên vào chương trình nghiên cứu sau đại học.

Trong code, biến phụ thuộc và biến độc lập được lưu trữ trong cùng một dataframe df\_nhaphoc ban đầu. Tuy nhiên, khi tiến hành huấn luyện mô hình hồi quy tuyến tính, biến phụ thuộc là GRE Score được lưu trữ trong biến y và các đặc trưng là TOEFL Score, University Rating, SOP, LOR, CGPA, và Chance of Admit được lưu trữ trong biến X.

Do đó, mặc dù biến phụ thuộc và biến độc lập có cùng một giá trị trong dataframe ban đầu, nhưng trong quá trình huấn luyện mô hình hồi quy tuyến tính, chúng được phân tách thành hai biến riêng biệt để đảm bảo tính chính xác của mô hình.

Phương trình lỗi được biểu diễn bởi công thức sau:

RSS = Σ(yi - ŷi)^2

Trong đó:

* RSS (Residual Sum of Squares) là tổng bình phương sai số giữa giá trị dự đoán ŷi và giá trị thực tế yi của biến phụ thuộc trên tất cả các mẫu trong tập huấn luyện.
* yi là giá trị thực tế của biến phụ thuộc (GRE Score) của mẫu i trong tập huấn luyện.
* ŷi là giá trị dự đoán của biến phụ thuộc (GRE Score) của mẫu i trong tập huấn luyện, được tính bởi mô hình hồi quy tuyến tính cho các giá trị đặc trưng tương ứng của mẫu i.

Trong code, RSS được tính bằng cách sử dụng phương thức .residues\_ của model, tức là:

RSS = np.sum((y\_pred - y) \*\* 2)

Ở đây, y\_pred là giá trị dự đoán của biến phụ thuộc (GRE Score) được tính bởi mô hình hồi quy tuyến tính cho các giá trị đặc trưng tương ứng của các mẫu trong tập huấn luyện, và y là giá trị thực tế của biến phụ thuộc (GRE Score) của các mẫu trong tập huấn luyện.