TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

fit@hcmus

TOÁN ỨNG DỤNG VÀ THỐNG KÊ

ĐỒ ÁN 3

LINEAR REGRESSION

Sinh viên thực hiện: Chang Jar bao
Mã số sinh viên: 23127323
Lớp: 23CLC01

MỤC LỤC

I. Ý tưởng thực hiện:

1. Tổng quan về đồ án:

Đồ án này tập trung vào việc phân tích các yếu tố ảnh hưởng đến thành tích học tập của sinh viên, được đo lường thông qua chỉ số Academic Student Performance Index (Performance Index). Mục tiêu là tìm ra mối quan hệ giữa các thói quen học tập, sinh hoạt và mức độ tham gia hoạt động ngoại khóa của sinh viên với kết quả học tập tổng thể.

Bộ dữ liệu được cung cấp với 10000 dòng dữ liệu với 6 thuộc tính:

- **Hours Studied:** Tổng số giờ học tập của mỗi sinh viên.
- **Previous Scores:** Điểm số các bài kiểm tra trước đó.
- **Extracurricular Activities:** Mức độ tham gia hoạt động ngoại khóa (0 – Không, 1 – Có).
- **Sleep Hours:** Số giờ ngủ trung bình mỗi ngày.
- **Sample Question Papers Practiced:** Số bài kiểm tra mẫu đã luyện tập.
- **Performance Index:** Chỉ số thể hiện thành tích học tập, nằm trong khoảng [10, 100].

2. Ý tưởng thực hiện:

Trong đồ án này, mô hình được chọn để huấn luyện và dự đoán dữ liệu là mô hình hồi quy tuyến tính hay còn gọi là Linear Regression. Sau đây sẽ là các bước thực hiện với mô hình:

Bước 1: Tiền xử lý dữ liệu. Xem xét các dữ liệu có đầy đủ không, đánh giá dữ liệu và chọn ra các mô hình phù hợp với từng yêu cầu (sử dụng các biểu đồ để thấy sự trực quan)
Bước 2: Xây dựng mô hình (sử dụng hồi quy tuyến tính)
Bước 3: Đánh giá mô hình. Em sử dụng độ đo MSE (Mean Squared Error) để đo độ sai lệch trung bình giữa giá trị dự đoán và giá trị thực tế.

II. Chi tiết thực hiện:

1. Các hàm và thư viện sử dụng:

1.1.Các hàm sử dụng:

Tên hàm	Dữ liệu vào	Kết quả	Công dụng
fit	Ma trận dữ liệu đặc trưng (X), ma trận dữ liệu mục tiêu (y)	Lưu kết quả $\{w_i\}$ vào lớp	Sử dụng phương pháp Bình phương tối thiểu thông thường để tìm các tham số tối ưu $\{w_i\}$
get_params	Không có	Các tham số $\{w_i\}$	Lấy các tham số $\{w_i\}$
predict	Ma trận dữ liệu đặc trưng (X)	Ma trận dữ liệu mục tiêu dự đoán (y_pred)	Dùng để dự đoán mô hình dựa trên các tham số $\{w_i\}$ được huấn luyện và tạo ra trước đó
MSE	Ma trận dữ liệu thực tế (y_test) và ma trận dữ liệu dự đoán (y_pred)	Đô đo MSE	Tính toán độ sai lệnh MSE giữa giá trị thực và dự đoán

kFold_cross_valid	Mô hình để huấn luyện (model), Ma trận dữ liệu đặc trưng (X), ma trận dữ liệu mục tiêu (y), số k (fold_indices), là tên model hoặc đặc trưng (name)	Danh sách gồm tên và độ đo MSE trung bình qua k fold	Cho mô hình huấn luyện qua phương pháp k-fold cross validation và tính toán ra độ đo trung bình MSE cho mô hình được huấn luyện đó

1.2. Các thư viện sử dụng

- **Thư viên pandas:** Đọc file csv và xử lý dữ liệu (như hàm corr() – Tính độ tương quan)
- **Thư viện sklearn:**
  + Sử dụng hàm KFold để chia tệp dữ liệu thành các phần train và set với (k – 1) train và 1 test.
  + Sử dụng hàm StandardScaler[4] để chuẩn hoá dữ liệu
  + Sử dụng hàm Pipeline[5]
- **Thư viện seaborn:** Sử dụng để trực quan hoá dữ liệu
- **Thư viện matplotlib:** Sử dụng để vẽ các biểu đồ phân tích các số liệu

2. Khám phá và phân tích dữ liệu:

Tiền xử lý: xem xét các giá trị có null hay lặp lại không

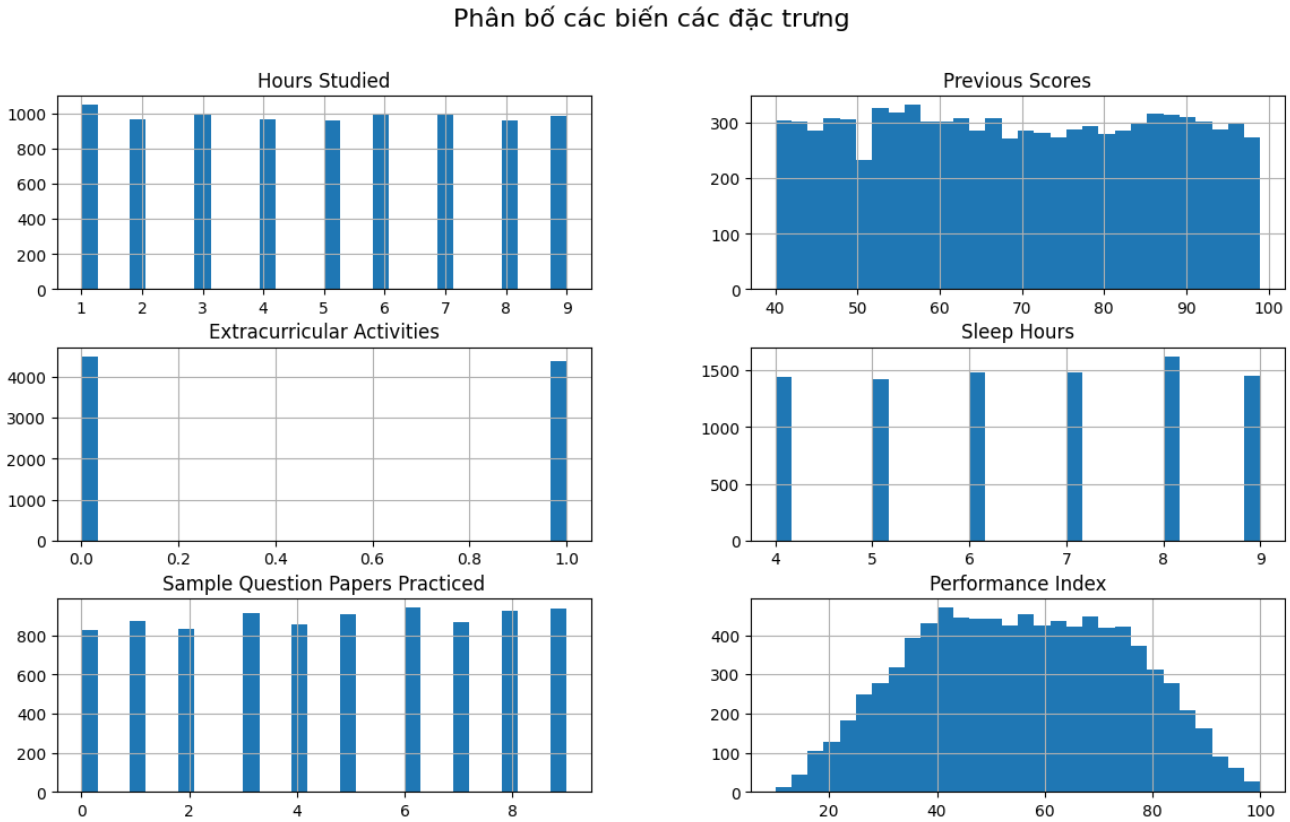
```
Số lượng giá trị null trong tập test:
Hours Studied                     0
Previous Scores                   0
Extracurricular Activities        0
Sleep Hours                       0
Sample Question Papers Practiced  0
Performance Index                 0
dtype: int64


Số lượng giá trị null trong tập train:
Hours Studied                     0
Previous Scores                   0
Extracurricular Activities        0
Sleep Hours                       0
Sample Question Papers Practiced  0
Performance Index                 0
dtype: int64


Giá trị lặp lại :
1 giá trị trong tập test và 103 giá trị trong tập train
```

Vậy ta sẽ xoá đi các dữ liệu bị lặp lại trong 2 tập trên. Nên tập train sẽ còn lại 8897 dòng dữ liệu trong tập train và 999 dòng dữ liệu trong tập test

2.1.Phân bố số lượng và giá trị của các đặc trưng

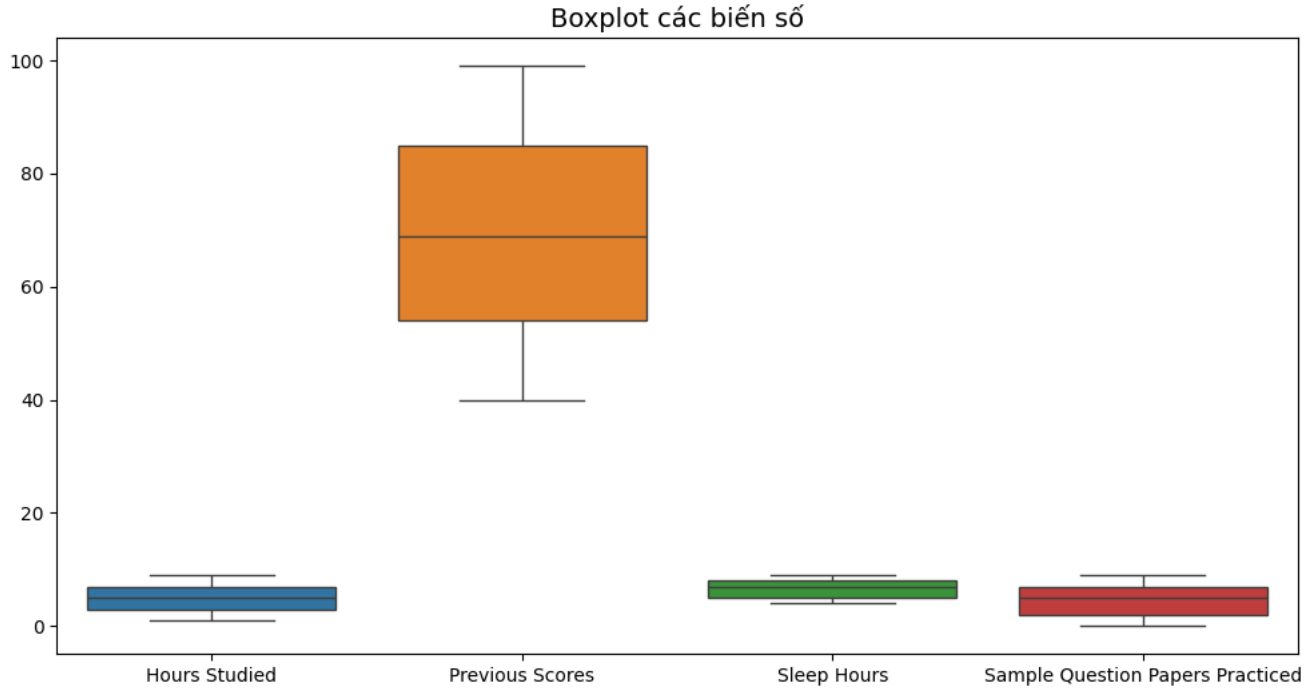


Trục hoành thể hiện các giá trị của biến đặc trưng và trục tung thể hiện số lượng của xuất hiện của giá trị đó. Dễ đàng nhận thấy:

- **Hours Studied:** Là biến rời rạc. Phân bố khá đồng đều từ mốc 1 giờ đến 9 giờ
- **Previous Scoures:** Là biến liên tục. Phân bố khá đều từ 40 điểm đến khoảng gần 100 điểm. Các phân bố đa số từ 250 biến trở lên
- **Extracurricular Activities:** Là biến nhị phân (0 – không tham gia, 1 – có tham gia). Phân bố gần như cân bằng ở 2 giá trị 0 và 1
- **Sleep Hours:** Là biến rời rạc. Phân bố khá đồng đều từ mốc 4 giờ đến 9 giờ
- **Sample Question Papers Practiced:** Là biến rời rạc. Phân bố khá đồng đều với nhau, tuy nhiên vẫn tập trung nhiều hơn ở khoảng từ 3 bài đến 6 bài kiểm tra
- **Performance Index:** Là biến liên tục. Phân bố theo hình chuông, gần giống với phân phối chuẩn. Đây là biến mục tiêu tốt cho mô hình hồi quy tuyến tính bởi nó gần như đã chuẩn hoá.

2.2.Phân bố box plot của các biến đặc trưng:

Ta sẽ không xét biến Extracurricular Activities vì chúng chỉ có 2 giá trị 1 và 0



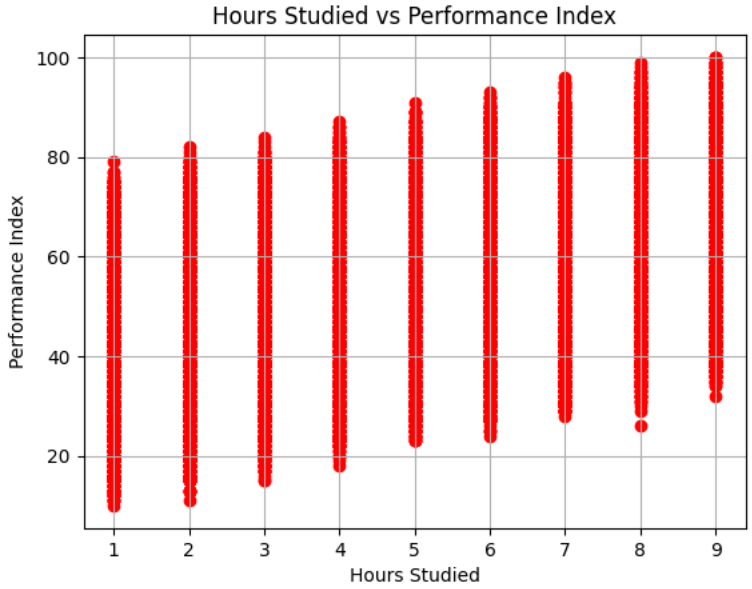
- **Hours Studied:** box hẹp (từ 3 đến 7), và whisker nhỏ (từ 1 đến 9). Tuy nhiên median thì ở giá trí trung bình, có nghĩa đặc trưng này phân bố khá đồng đều. Các giá trị cụ thể như sau:
  + Q1: 3.00
  + Median: 5.00
  + Q3: 7.00
  + Lower Whisker: 1.00
  + Upper Whisker: 9.00
- **Previous Scores:** box rộng (từ 54 đến 85), và whisker dài (từ 40 đến 99). Điều này thể hiện đặc trưng có độ biến thiên lớn. Các giá trị cụ thể như sau:
  + Q1: 54.00
  + Median: 69.00
  + Q3: 85.00
  + Lower Whisker: 40.00
  + Upper Whisker: 99.00
- **Sleep Hours:** box hẹp (từ 5 đến 8), và whisker nhỏ (từ 4 đến 9). Median = 7 có nghĩa đa số học sinh ngủ từ 7 tiếng trờ lên, điều này làm dữ liệu bị lệch về một phía. Các giá trị cụ thể như sau:
  + Q1: 5.00
  + Median: 7.00
  + Q3: 8.00
  + Lower Whisker: 4.00
  + Upper Whisker: 9.00
- **Sample Question Papers Practice**: box hẹp (từ 2 đến 7), và whisker nhỏ (từ 0 đến 9). Median khá gần vị trí trung tâm, có nghĩa mô hình này tập trung nhiều ở các giá trị trung tâm (từ 3 đến 6) hoặc phân bố tương đối đồng đều). Các giá trị cụ thể như sau:
  + Q1: 2.00
  + Median: 5.00
  + Q3: 7.00

+ Lower Whisker: 0.00
+ Upper Whisker: 9.00

Các phân bố trên không tìm thấy các outlier (điểm rời rạc).

2.3.Phân bố scatter plot của từng đặc trưng với biến mục tiêu

- **Hours Studied:**

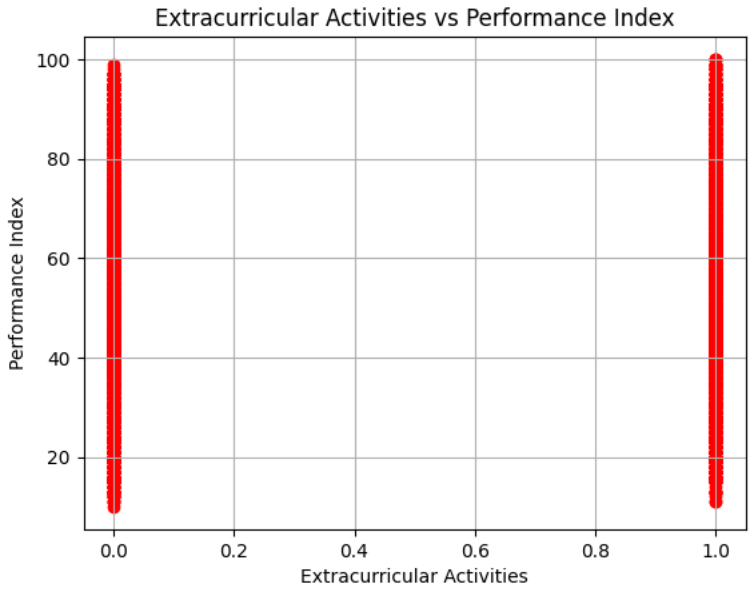


Nhận thấy Performance index có tăng dần theo Hours studied. Tuy nhiên đây không thể hiện được sự tuyến tính, các dữ liệu trùng nhau trên từng cột (từ mốc 1 đến 9), mức độ tương quan sẽ không quá tốt

- **Previous Scores:**

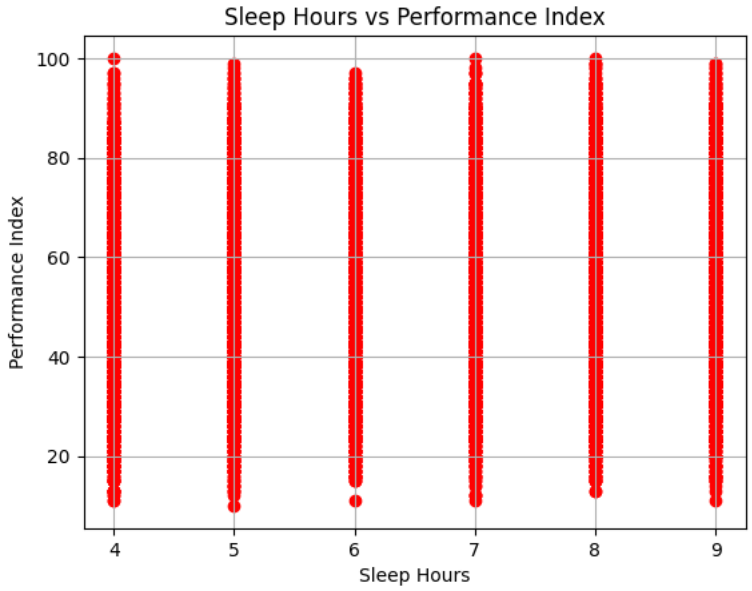
Nhận thấy Performance index tăng dần theo Previous scores và các điểm có xu hương thẳng và phân bố khá đều. Điều này có nghĩa mô hình hồi quy tuyến tính sẽ áp dụng hiệu quả đối với đặc trưng này.

- **Extracurricular Activities:**



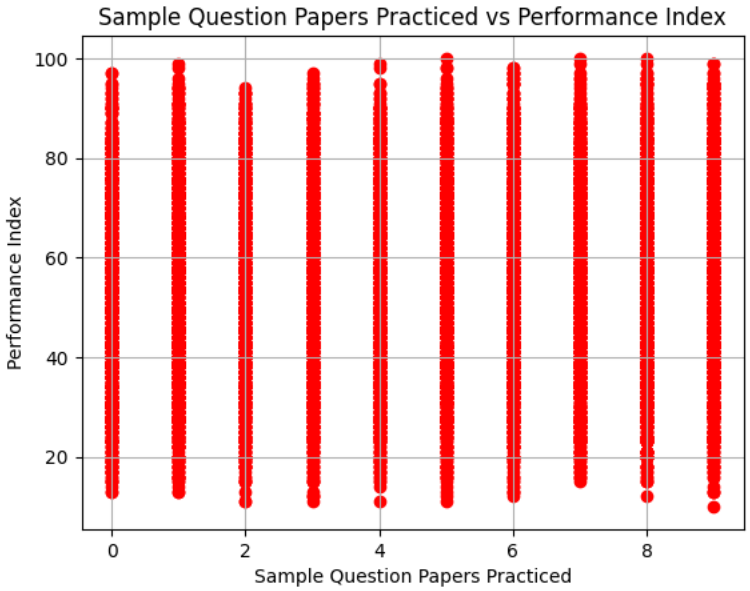
Nhận thấy phân bố ở hai cột 0 và 1 gần như giống. Đặc trưng này không thể hiện sự tuyến tính, cũng như không hề thể hiện sự tăng giảm, mức độ tương quan sẽ cực kì kém.

- **Sleep Hours:**



Nhận thấy mô hình không thể hiện được sự tuyến tính, các dữ liệu trùng nhau trên từng cột (từ mốc 4 đến 9), hơn nữa Performance index tăng giảm hỗn loạn qua từng cột. Điều này có nghĩa sự tương quan giữa đặc trưng này với biến muc tiêu sẽ khá kém

- **Sample Question Papers Practice:**



Nhận thấy mô hình không thể hiện được sự tuyến tính, các dữ liệu trùng nhau trên từng cột (từ mốc 0 đến 9), hơn nữa Performance index tăng giảm hỗn loạn qua từng cột. Điều này có nghĩa sự tương quan giữa đặc trưng này với biến muc tiêu sẽ khá kém.

3. Các mô hình hồi quy tuyến tính

Các mô hình sẽ được chia 90% là tập train và 10% còn lại là tập test.

3.1.Yêu cầu 2a: Hồi quy tuyến tính đa biến dùng toàn bộ 5 đặc trưng. Các bước thực hiện như sau:

Bước 1: Sử dụng tập train để huấn luyện mô hình

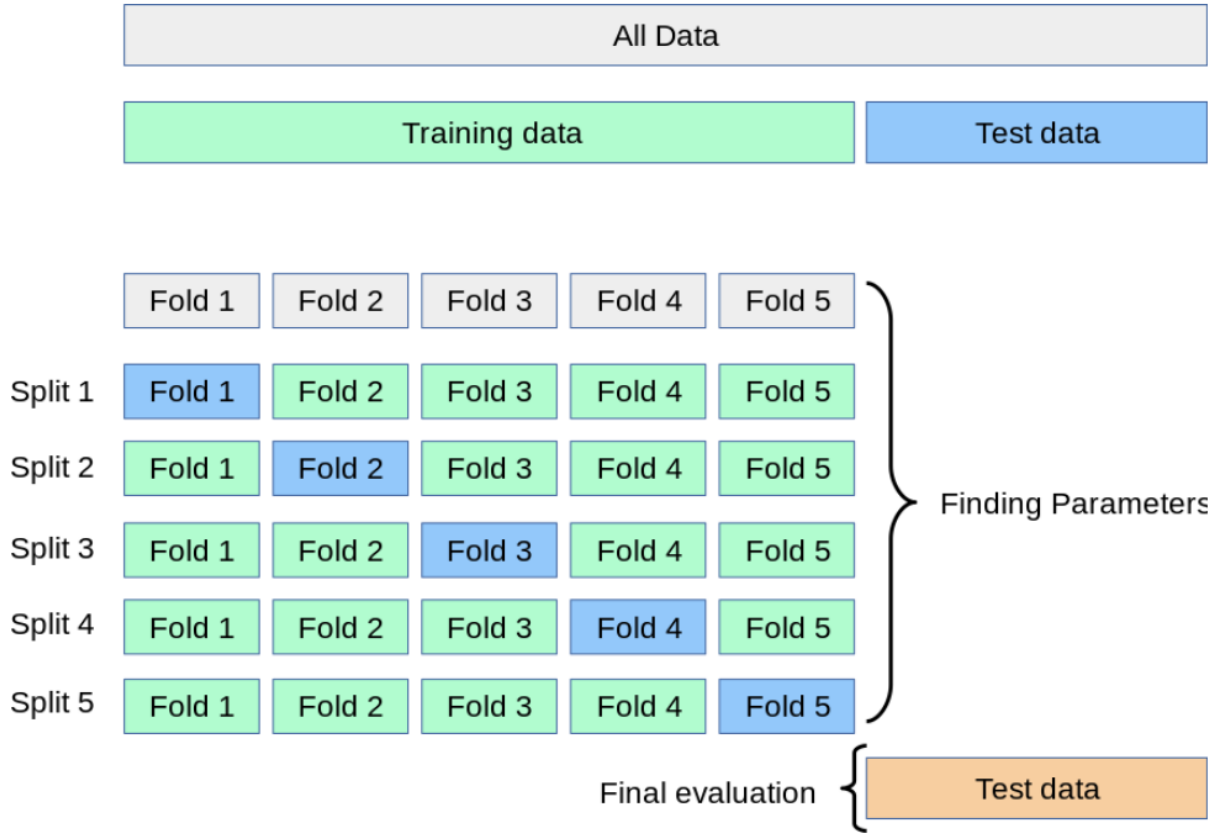
Bước 2: Sử dụng công thức hồi quy được huấn luyện từ trước áp dụng cho tập test

Bước 3: Tính độ đo MSE theo công thức[1]:

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

3.2.Yêu cầu 2b

Bước 1: Chia tập train thành k tập con đều nhau (k-fold) với k – 1 tập sẽ là tập train k tập sẽ thành tập test. Ở đay em sử dụng k-fold Cross Validation, ý tưởng như sau[2]:



Bước 2: Huấn luyện và tính độ đo MSE như yêu cầu 2a nhưng với 1 biến đặc trưng duy nhất, và ta sẽ huấn luyện k lần. Sau đó kết quả độ đo MSE cuối cùng sẽ là trung bình các độ đo trong k lần huấn luyện

Bước 3: Lấy đặc trưng tốt nhất (MSE nhỏ nhất) để huấn luyện trên toàn tập train lớn và trả kết quả 1 lần nữa

3.3.Yêu cầu 2c

Bước 1: Tạo mô hình, ở đây em chỉ tạo ba mô hình đó là:

Lấy 2 đặc trưng tốt nhất cho mô hình

- Em sử dụng corr (hệ số tương quan) và độ đo MSE trước đó để xác định nên ưu tiên đặc trưng nào hơn. Corr càng gần 1 chứng tỏ sự tương quan giữa đặc trưng đó với biến mục tiêu càng lớn và ngược lại. Vậy nên chọn hai dặc trưng có corr cao nhất và MSE nhỏ nhất

- Heatmap hệ số tương quan:

- Bảng hiển thị corr và mse của từng đặc trưng:

	Feature	Correlation	MSE
0	Previous Scores	0.914630	60.142226
1	Hours Studied	0.370113	317.507258
2	Sleep Hours	0.046501	367.292337
3	Sample Question Papers Practiced	0.040509	367.481439
4	Extracurricular Activities	0.026788	367.666941

⇨ **Vậy chọn Previous Scores và Hours Studied làm hai đặc trưng cho mô hình này**

Chuẩn hoá mô hình

- Em sử dụng hàm StandardScaler để đưa mô hình về dạng chuẩn hoá. Sau đó sử dụng các đặc trưng ấy đi huấn luyện

Cộng hai đặc trưng kém nhất

- Em sử dụng corr (hệ số tương quan) và độ đo MSE trước đó để xác định nên chọn đặc trưng nào. Vậy nên chọn hai dặc trưng có corr thấp nhất và MSE lớn nhất nhất. Các hình ảnh quan sát như hình ở mô hình đầu tiên
    - ⇨ **Vậy chọn Sample Question Papers Practiced và Extracurricular Activities**

Bước 2: Sử dụng phương pháp K-fold Cross Validation và huấn luyện từng mô hình. Sau đó kết quả độ đo MSE cuối cùng sẽ là trung bình các độ đo trong k lần huấn luyện

Bước 3: Sử dụng mô hình tốt nhất (MSE nhỏ nhất) để huấn luyện và dự đoán trên tập test

III. Kết quả và phân tích:.

Lưu ý:

+ Khi $w_i > 0$ nghĩa là mối quan hệ giữa X_i và y tích cực, khi X_i tăng 1 thì y sẽ tăng một lượng bằng w_i (trong trương hợp các đặc trưng khác không thay đổi). Khi $w_i < 0$ nghĩa là mối quan hệ giữa X_i và y tiêu cực, khi X_i tăng 1 thì y sẽ giảm một lượng bằng w_i (trong trương hợp các đặc trưng khác không thay đổi). Tuy nhiên ở đồ án này em có sử dụng intercept[3] nên các giá trị w_i ($i > 0$) tương ứng các đặc trưng sẽ lớn hơn 0

+ Độ đo MSE càng lớn có nghĩa sai số giữa dự đoán và thực tế càng lớn.

1. Yêu cầu 2a

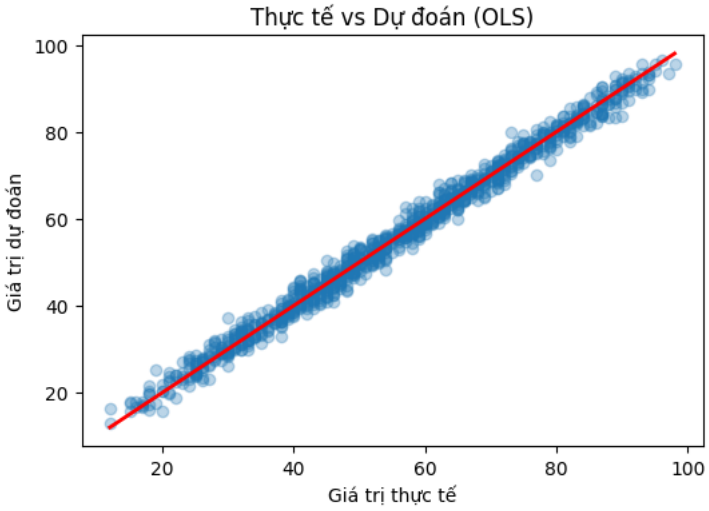
1.1.Công thức hồi quy

$$\text{Student Performance} = -33.969 + 2.852X_1 + 1.018X_2 + 0.604X_3 + 0.474X_4 + 0.192X_5$$

Với X_1, X_2, X_3, X_4, X_5 lần lượt là Hours Studied, Previous Scores, Extracurricular Activities, Sleep Hours, Sample Question Papers Practiced

Từ công thức ta nhận ra, hai đặc trưng Hours Studied và Previous Scores (tức X_1 và X_2) có khả năng dự đoán hiệu quả. Trong khi đó ba đặc trưng còn lại thì làm dự đoán bị nhiễu, gây ra sai số.

1.2.Biểu đồ dự đoán so với thực tế



Nhận thấy các điểm phân tán tạo thành tạo thành một đường thẳng hồi quy. Điều này có nghĩa mô hình đang dự đoán khá chính xác, tuy nhiên vẫn còn khá nhiều sai số (các điểm phân tán không nằm trên hoặc nằm gần đường chéo)

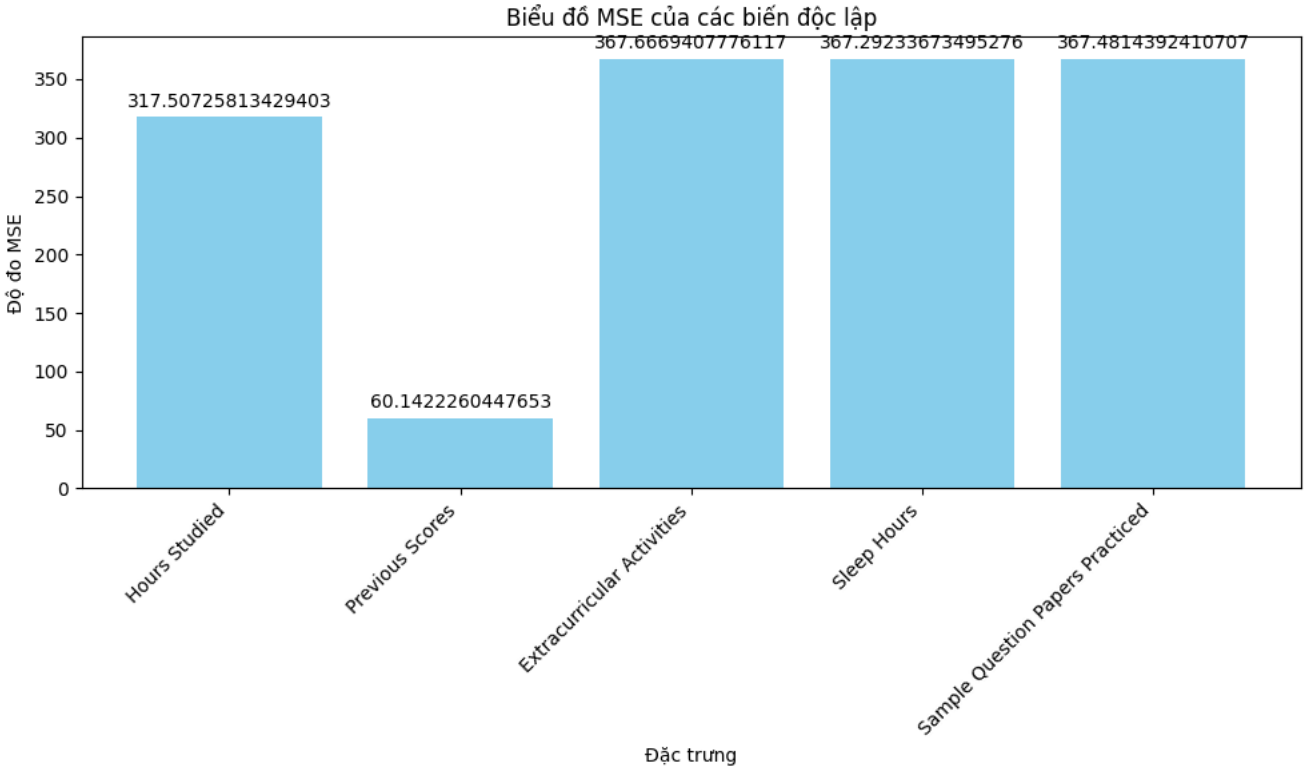
1.3.Độ đo MSE

MSE = 4.093 là một con số khá cho một mô hình dự đoán, cho thấy mô hình dự đoán khá tốt, chỉ lệch khoảng 2 điểm

⇨ **Nhận xét: Mô hình này dự đoán khá chính xác và tốt**

2. Yêu cầu 2b

2.1.So sánh độ sai số của từng đặc trưng đối với từng mô hình



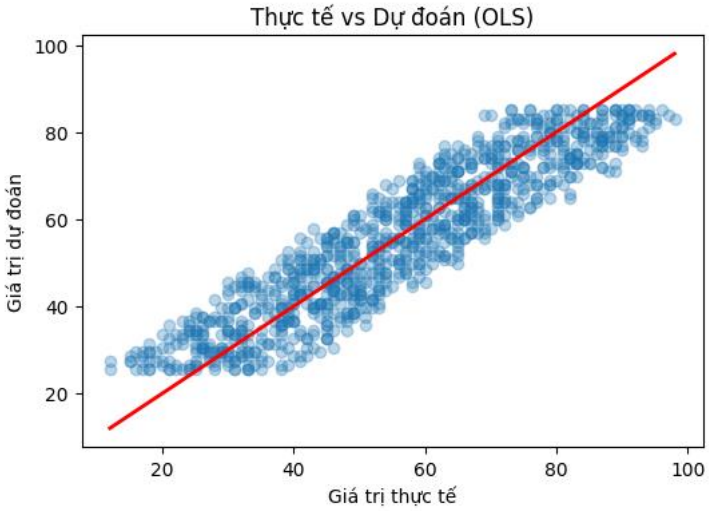
Ta nhận thấy đặc trưng Previous Scores có độ sai số thấp nhất. Ngược lại, 4 thuộc tính còn lại có độ sai số tương đối cao, khổng thể dự đoán được mô hình. Nên đặc trưng tốt nhất trong mô hình ta sử dụng sẽ là Previous Scores.

2.2.Công thức hồi quy cho đặc trưng tốt nhất

$$\text{Student Performance} = -14.989 + 1.011\text{Previous Scores}$$

Nhận thấy hệ số hồi quy của đặc trưng Previous Scores có vẻ nhỏ hơn so với ở yêu cầu 2a. Điều này là do ở yêu cầu 2a, cả 5 đặc trưng đều được sử dụng, dẫn đến việc đặc trưng Previous Scores phải "gánh" các đặc trưng khác sao cho tổng hệ số hồi quy gần bằng 1. Vậy nên, hệ số w sẽ phải lớn hơn để cân bằng được điều đó.

2.3.Biểu đồ dự đoán so với thực tế cho đặc trưng tốt nhất



Nhận thấy các điểm phân tán theo đường thẳng hồi quy, tuy nhiên các điểm vẫn còn nằm ngoài đường thẳng khá nhiều so với biểu đồ ở yêu cầu 2a.

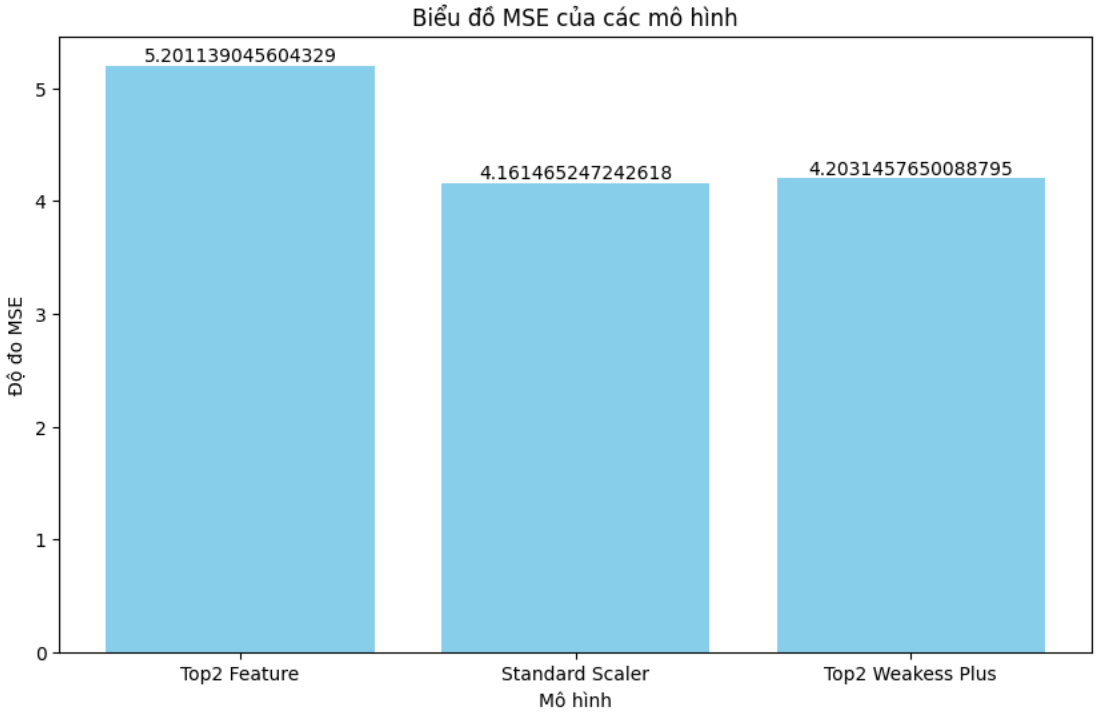
2.4.Độ đo MSE cho đặc trưng tốt nhất

MSE = 58.888 là một con số tương đối lớn so với mô hình dự đoán, dự đoán lệch khoảng 7 điểm so với thực tế. Điều này là do sự vắng mặt có các biến đặc trưng có góp phần trong việc quyết định thành tích học tập (Performance index) của học sinh.

⇨ **Nhận xét: Mô hình này có thể dự đoán được nhưng sai số sẽ khá cao.**

3. Yêu cầu 2c

3.1.So sánh độ sai số của từng mô hình



Nhận thấy các mô hình trên cho ra độ đo trung bình MSE khá tốt, dự đoán chỉ cách biệt khoảng 2 điểm so với thực tế. Hơn nữa có thể nhận thấy, khi chỉ dùng hai đặc trưng tốt nhất thì mô hình vẫn gần như tốt bằng việc dùng cả 5 đặc trưng (5,201 so với 4.093, nếu quy ra thì dự đoán hơn kém nhau chưa đến 1 điểm), điều này có thể dẫn đến một kết luận rằng đặc trưng Previous Hours và Hours Studied là hai đặc trưng góp phần mạnh nhất để dự đoán mô hình, trong khi ba biến đặc trưng còn lại mang lại giá trị khá ít trong việc dự đoán thành tích học tập của học sinh.

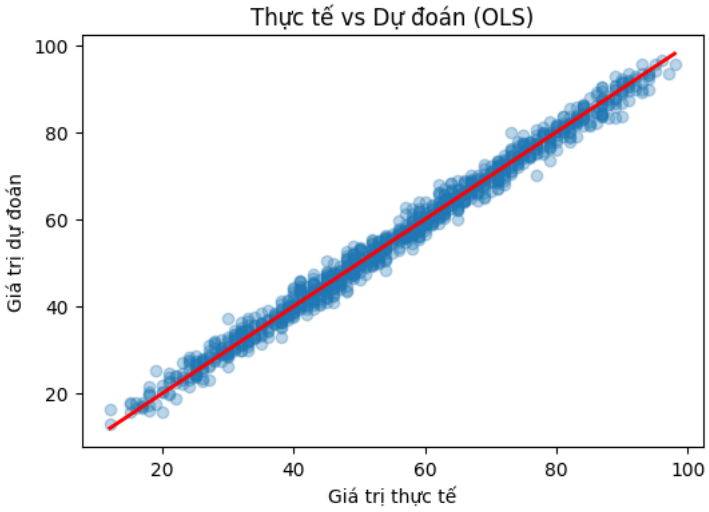
3.2.Công thức hồi quy của mô hình tốt nhất

$$\text{Student Performance} = 55.136 + 7.400X_1 + 17.679X_2 + 0.302X_3 + 0.803X_4 + 0.551X_5$$

Với X_1, X_2, X_3, X_4, X_5 lần lượt là Hours Studied, Previous Scores, Extracurricular Activities, Sleep Hours, Sample Question Papers Practiced

Có thể dễ dàng nhận thấy điểm trung bình của học sinh = intercept = 55.136 (tức là w_0 của công thức hồi quy). Điều này khá hợp lý do đây phần lớn học sinh sẽ đạt được số điểm trung bình như này trên thang 100 (đúng theo phân phối chuẩn). Hơn nữa do các dữ liệu đã được chuẩn hoá nên lưu ý về w_i không còn chính xác. Các w_i ở đây chỉ thể hiện mức độ ảnh hưởng của biến đặc trưng đó với biến mục tiêu, với w_i càng lớn thì sự ảnh hưởng càng lớn.

3.3.Biểu đồ dự đoán so với thực tế của mô hình tốt nhất



Không có sự khác biệt đáng kể so với mô hình ở yêu cầu 2a. Mô hình này dự đoán khá chính xác, các sai số cũng khá nhỏ, có thể chấp nhận được.

⇨ **Nhận xét: Mô hình chuẩn hoá đưa ra kết quả khá tương đồng so với mô hình xét cả 5 đặc trưng, tuy nhiên sẽ thể hiện rõ hơn về sự phân phối trung bình của dữ liệu (intercept có ý nghĩa hơn).**

IV. Tài liệu tham khảo:

1. Wikipedia, https://vi.wikipedia.org/wiki/Sai_số_toàn_phương_trung_bình
2. Scikit-learn, https://scikit-learn.org/stable/modules/cross_validation.html
3. theanalysisfactor, https://www.theanalysisfactor.com/interpreting-the-intercept-in-a-regression-model/
4. Scikit-learn, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html#sklearn.preprocessing.StandardScaler
5. Scikit-learn, https://scikit-learn.org/stable/api/sklearn.pipeline.html#module-sklearn.pipeline

Ngoài ra, còn có sự hỗ trợ của trí tuệ nhân tạo về các cách vẽ biểu đồ (heatmap, boxplot) trong quá trình sử dụng và tham khảo source code của NgocTien1010, github