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

Project 03: Linear Regression

Môn học: Toán Ứng Dụng Và Thống Kê Cho Công Nghệ Thông Tin

MTH00057 - 23CLC04

Sinh viên thực hiện:

Đinh Xuân Khương - 23127398

Giảng viên hướng dẫn:

Nguyễn Ngọc Toàn

Trần Hà Sơn

Ngày 14 tháng 8 năm 2025

Mục lục

1 Tổng quan về đồ án

1.1 Mục tiêu

Mục tiêu của đồ án này chính là tìm ra các yếu tố có thể ảnh hưởng đến thành tích học tập của học sinh (Performance Index) và xây dựng mô hình dự đoán sử dụng OLS Linear Regression (Hồi quy tuyến tính sử dụng phương pháp bình phương tối thiểu) để khớp với dữ liệu, từ đó có thể đưa ra các đánh giá thông qua thử nghiệm với nhiều mô hình. Các yếu tố ảnh hưởng có thể là: Hours studied, Previous Scores, Extracurricular Activities, Sleep Hours, Sample Question Papers Practiced.

1.2 Input và output

1.2.1 Input

Dữ liệu thành tích sinh viên (Student Performance) có 10000 dòng và 6 cột. Ý nghĩa và kiểu dữ liệu của từng cột được thể hiện ở bảng sau:

SST	Thuộc tính	Mô tả	Kiểu dữ liệu
1	Hours Studied	Tổng số giờ học của mỗi sinh viên	Integer
2	Previous Scores	Điểm số học sinh đạt được trong các bài kiểm tra trước	Integer
3	Extracurricular Activities	Sinh viên có tham gia hoạt động ngoại khoá không (Có hoặc Không)	Interger: 1 → Có và 0 → Không
4	Sleep Hours	Số giờ ngủ trung bình mỗi ngày của sinh viên	Integer
5	Sample Question Papers Practiced	Số bài kiểm tra mẫu mà học sinh đã luyện	Integer
6	Performance Index	Thước đo thành tích tổng thể cho mỗi sinh viên. Chỉ số thể hiện thành tích học tập, nằm trong đoạn [10, 100]. Chỉ số này tỉ lệ thuận với thành tích.	Float

Bảng 1: Bảng mô tả thông tin dữ liệu của từng thuộc tính.

Bộ dữ liệu được chia ngẫu nhiên thành 2 tập với tỉ lệ 9:1. Trong đó, 9 phần cho tập huấn luyện, 1 phần cho tập kiểm tra.

- **p03.train.csv**: Chứa 9000 mẫu dùng để huấn luyện mô hình.
- **p03.test.csv**: Chứa 1000 mẫu dùng để kiểm tra và đánh giá mô hình.

1.2.2 Output

Các trọng số của mô hình với nhiều cách chọn thuộc tính khác nhau. Từ đó có những đánh giá và kết luận.

1.3 Thư viện sử dụng

1.3.1 Pandas

- **Mục đích:** Xử lý và phân tích dữ liệu dạng bảng (DataFrame).
- **Lý do sử dụng:**
  - Đọc/ghi dữ liệu từ nhiều định dạng như `.csv`, ...
  - Thao tác dữ liệu dễ dàng: lọc, gộp, nhóm, tính toán thống kê.
  - Quản lý dữ liệu có nhãn (columns, index) giúp code dễ đọc và trực quan.

1.3.2 Numpy

- **Mục đích:** Xử lý tính toán số học hiệu năng cao với mảng (array) và ma trận.
- **Lý do sử dụng:**
  - Cung cấp các phép toán vector, ma trận nhanh hơn khi sử dụng vòng lặp trong Python thuần.
  - Hỗ trợ các hàm toán học nâng cao: np.linalg.inv(), ...
  - Pandas DataFrame nội bộ cũng dựa trên NumPy để lưu trữ dữ liệu.

1.3.3 Matplotlib

- **Mục đích:** Vẽ biểu đồ 2D cơ bản đến nâng cao.
- **Lý do sử dụng:**
  - Tạo các loại biểu đồ: line chart, bar chart, scatter plot, histogram...
  - Tùy chỉnh màu sắc, tiêu đề, nhãn trục, chú thích.

1.3.4 Seaborn

- **Mục đích:** Vẽ biểu đồ thống kê đẹp và trực quan hơn, xây dựng dựa trên Matplotlib.
- **Lý do sử dụng:**
  - Cung cấp sẵn nhiều biểu đồ thống kê: heatmap, boxplot, ...
  - Hỗ trợ trực tiếp với Pandas DataFrame (truyền tên cột thay vì mảng).
  - Tích hợp style đẹp mắt mặc định (theme) giúp biểu đồ chuyên nghiệp.

1.4 Mức độ hoàn thành

STT	Yêu cầu	Mức độ hoàn thành
1	Thực hiện phân tích khám phá dữ liệu	100%
2.a	Sử dụng 5 đặc trưng xây dựng mô hình dự đoán chỉ số thành tích bằng mô hình hồi quy tuyến tính	100%
2.b	Sử dụng duy nhất 1 đặc trưng, tìm mô hình cho kết quả tốt nhất	100%
2.c	Tự xây dựng và thiết kế mô hình (3 mô hình), tìm mô hình cho kết quả tốt nhất	100%

Bảng 2: Bảng đánh giá mức độ hoàn thành.

2 Chi tiết thực hiện

2.1 Khám phá và phân tích dữ liệu (EDA)

Lưu ý: Các số liệu và đánh giá này được thực hiện trên tập `p03.train.csv`.

2.1.1 Khám phá và xử lý dữ liệu

Khi sử dụng hàm `info()`:

Bảng 3: Thông số dữ liệu

#	Column	Non-Null Count	Dtype
0	Hours Studied	9000 non-null	int64
1	Previous Scores	9000 non-null	int64
2	Extracurricular Activities	9000 non-null	int64
3	Sleep Hours	9000 non-null	int64
4	Sample Question Papers Practiced	9000 non-null	int64
5	Performance Index	9000 non-null	float64

Theo như bảng thông số dữ liệu, ta thấy không có hàng hay cột nào bị thiếu dữ liệu.
Vì dữ liệu này được thống kê trên nhiều sinh viên/học sinh khác nhau nên **không cần loại bỏ các hàng trùng nhau** vì có thể tồn tại các sinh viên có phần thể hiện đặc trưng giống nhau. Từ đó ta không cần xử lý dữ liệu.

2.1.2 Phân tích dữ liệu

Khi sử dụng hàm `describe()`:

Bảng 4: Thống kê dữ liệu

	HS	PS	EA	SH	SQPP	PI
count	9000	9000	9000	9000	9000	9000
mean	4.976	69.396	0.494	6.536	4.591	55.136
std	2.595	17.370	0.500	1.696	2.865	19.188
min	1.000	40.000	0.000	4.000	0.000	10.000
25%	3.000	54.000	0.000	5.000	2.000	40.000
50%	5.000	69.000	0.000	7.000	5.000	55.000
75%	7.000	85.000	1.000	8.000	7.000	70.000
max	9.000	99.000	1.000	9.000	9.000	100.000

Trong đó:

- HS: Hours Studied
- PS: Previous Scores
- EA: Extracurricular Activities
- SH: Sleep Hours

- SQPP: Sample Question Papers Practiced
- PI: Performance Index

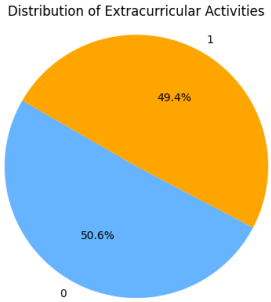
Phân tích phân bố dữ liệu

- **Hours Studied**: Phân bố gần như đều (uniform-like) từ 1-9 giờ, với đa số học sinh học 3-7 giờ/ngày. Độ lệch chuẩn cao cho thấy sự đa dạng trong thói quen học của học sinh/sinh viên.
- **Previous Scores**: Phân bố rộng, gần normal với mean cao (69%), nhưng độ lệch chuẩn khá lớn (17.37) so với trung bình, chỉ ra khoảng cách lớn giữa học sinh giỏi và kém.
- **Extracurricular Activities**: Phân bố nhị phân cân bằng (khoảng 49% có tham gia và 51% không tham gia).
- **Sleep Hours**: Phân bố hẹp, đa số ngủ 5-8 giờ (healthy range), độ lệch chuẩn thấp cho thấy ít biến thiên, từ đó có thể kết luận các học sinh vẫn chú trọng đến sức khỏe, ít tình trạng thức khuy và thiếu ngủ.
- **Sample Question Papers Practiced**: Phân bố rộng từ 0-9, mean 4.59 cho thấy trung bình luyện $\tilde{5}$ đề, nhưng std cao (2.86) phản ánh sự khác biệt lớn trong phong cách ôn tập của học sinh/sinh viên.
- **Performance Index**: Phân bố rộng nhất (std 19.19), mean trung bình (55%), cho thấy hiệu suất đa dạng, có tiềm năng cải thiện. 50% học sinh ở mức 40-70.

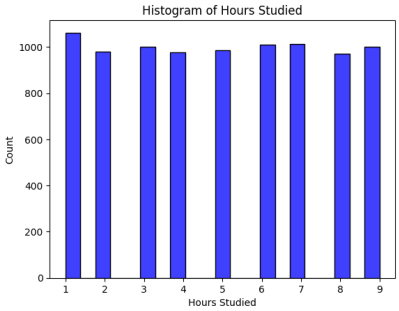
Kết luận

- Tổng quan, ta có thể kết luận rằng dữ liệu có tính cân bằng tốt (không skew mạnh), phù hợp cho mô hình học máy như hồi quy tuyến tính. Không có outliers rõ ràng (min/max hợp lý).
- Performance Index có std tương đương Previous Scores (19.19 vs 17.37), gợi ý mối liên hệ mạnh giữa điểm trước và hiệu suất hiện tại, từ đó cho thấy học sinh có nền tảng tốt thường duy trì hiệu suất cao.
- Phân phối của Extracurricular Activities tương đối cân bằng giúp tránh bias trong tập dữ liệu.

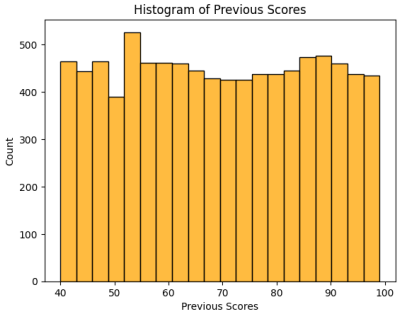
Histogram và Pie chart



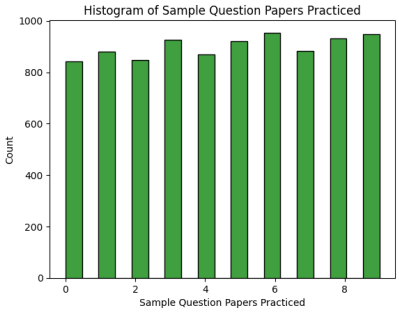
(a) Extracurricular Activities



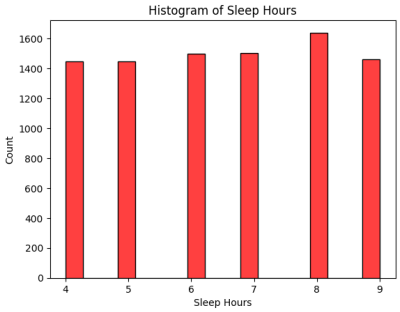
(b) Hour Studied



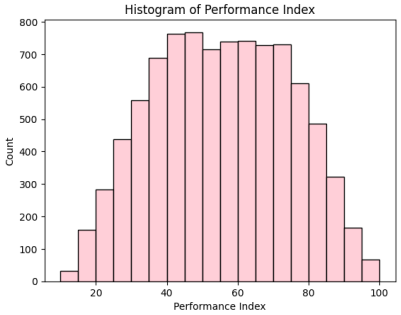
(c) Previous scores



(d) SQ Papers Practiced



(e) Sleep Hours



(f) Performance Index

- **(a) Extracurricular Activities – Pie chart**
  - Tỉ lệ 0/1 gần như cân bằng ($\sim 50.6\%$ vs $\sim 49.4\%$).
  - Ý nghĩa: không có mất cân bằng lớp (class imbalance) — tốt khi dùng biến này làm đặc trưng nhị phân trong mô hình; không cần cân bằng lại dữ liệu theo biến này.
- **(b) Hours Studied – Histogram**
  - Giá trị rời rạc 1–9 với số lượng mỗi mức khá đồng đều $\rightarrow$ phân phối gần "uniform".
  - Ý nghĩa khả dĩ: dữ liệu có vẻ được lấy/giả lập để phủ đều các mức giờ học (ít thiên lệch). Khi gần như đồng đều, riêng biến này khó tạo ra hình dạng "chuẩn" trong histogram.
- **(c) Previous Scores – Histogram**
  - Miền 40–100, phân phối khá phẳng, có dao động nhẹ theo từng bin. Không thấy lệch mạnh về hai đầu.

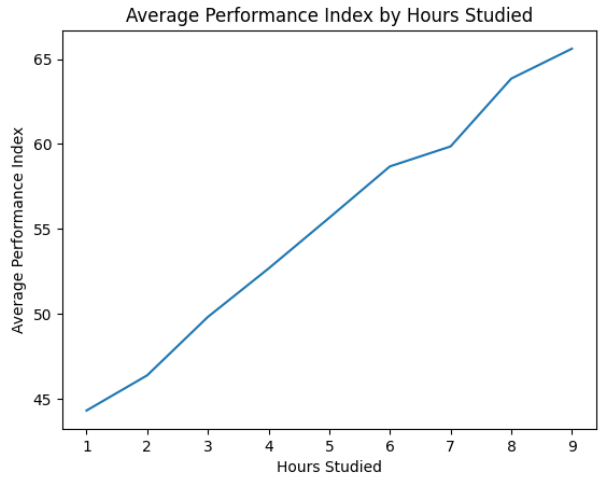
  - Ý nghĩa: đây không phải phân phối chuẩn; nếu dùng trong hồi quy, nên chuẩn hoá/chuẩn tính (standardize) trước; đồng thời cần kiểm tra tuyến tính với chỉ số kết quả.

- **(d) SQ Papers Practiced – Histogram**
  - Biến rời rạc 0–9; các cột dao động quanh một mức tương đương, có xu hướng hơi tăng về phía cao hơn.
  - Giải thích: có thể nhiều học viên làm 7–10 đề hơn 0–3 đề (hành vi học tập khá tích cực) nhưng không chênh lệch quá nhiều.

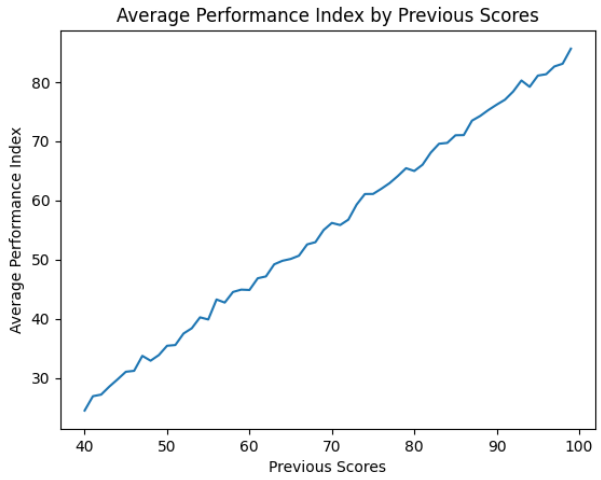
- **(e) Sleep Hours – Histogram**
  - Giá trị rời rạc 4–9; đỉnh có vẻ nằm quanh 6–8 giờ; 4,5 và 9 giờ ít hơn.
  - Ý nghĩa: Có thể tồn tại quan hệ dạng "đỉnh tối ưu" (quá ít hoặc quá nhiều ngủ đều kém).

- **(f) Performance Index – Histogram**
  - Phân phối gần chuẩn/bell-shaped, tâm khoảng 55–60, hai đuôi thu dần về ~ 20 và ~ 95.
  - Ý nghĩa: Rất giống một đại lượng tổng hợp (linear combination + noise) $\rightarrow$ thích hợp cho mô hình hoá hồi quy; nhiều kiểm định thống kê giả định chuẩn có thể áp dụng.

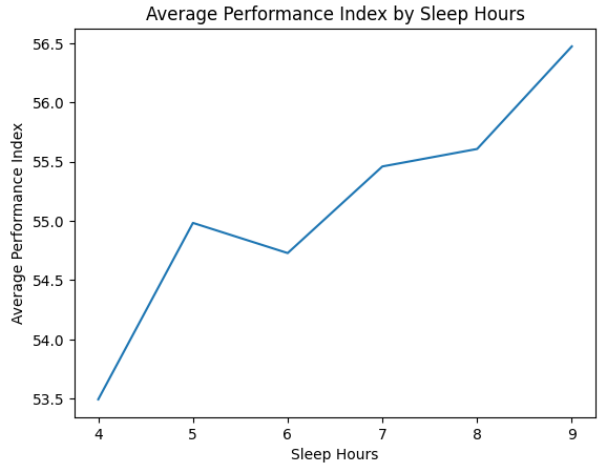
Kiểm tra quan hệ tuyến tính với trung bình Performance Index



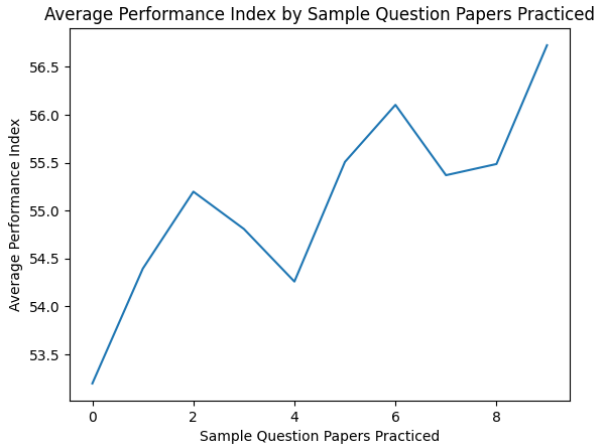
(a) Hours Studied



(b) Previous Scores



(c) Sleep Hours



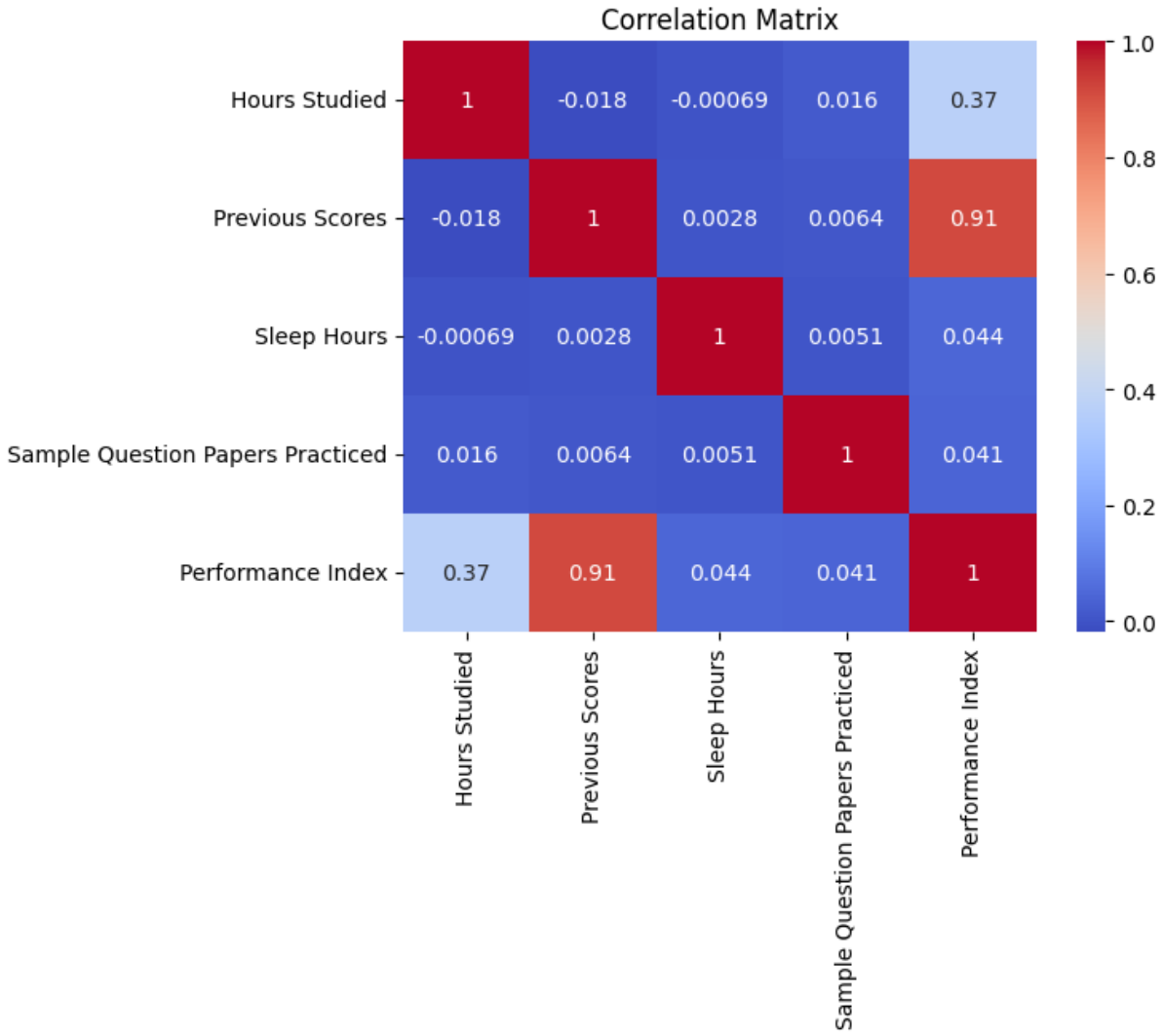
(d) Sample Question Papers Practiced

Hình 2: Bộ 4 ảnh chia 2 hàng, mỗi hàng 2 ảnh

Đánh giá

- Nhìn chung, Hours Studied và previous Scores có mối quan hệ tuyến tính tăng khá tốt với performance index.
- Trong khi đó, Sleep Hours và Sample Question Papers Practiced không cho chúng ta thấy được mối quan hệ rõ ràng, Performance Index có xuất hiện những khoảng giảm ở một số vị trí. Từ đó, cho thấy không phải ngủ đủ giấc, hay làm đề nhiều thì Performance Index sẽ tốt. Mà cách học, thời gian học và nền tảng vẫn sẽ chiếm chủ yếu cho hiệu quả làm bài.

Ma trận tương quan (correlation matrix)



Hình 3: correlation matrix

- **Tổng quan**: Ma trận hiển thị mối quan hệ tuyến tính giữa các biến: Hours Studied, Previous Scores, Sleep Hours, Sample Question Papers Practiced, và Performance Index. Đa số tương quan yếu, ngoại trừ một số cặp liên quan đến Performance Index.
- **Hours Studied và Performance Index**: Tương quan dương trung bình (0.37), cho thấy học nhiều giờ hơn có thể cải thiện hiệu suất.
- **Previous Scores và Performance Index**: Tương quan dương rất mạnh (0.91), là yếu tố dự báo tốt nhất cho hiệu suất.

- **Sleep Hours và Performance Index**: Tương quan dương yếu (0.044), giấc ngủ có ảnh hưởng tích cực nhẹ.
- **Sample Question Papers Practiced và Performance Index**: Tương quan dương yếu (0.041), luyện tập bài mẫu có lợi ích nhỏ.
- Các đặc trưng còn lại thì không có tương quan lớn, da số là độc lập với nhau.
- **Ý nghĩa tổng thể**: Previous Scores và Hours Studied là các yếu tố chính ảnh hưởng đến Performance Index. Không có đa cộng tuyến giữa các biến độc lập. Có thể được dùng để xây dựng mô hình ở câu `2c`.
- **Hạn chế**: Ma trận tương quan chỉ đo lường tương quan tuyến tính, không phản ánh kết quả nhân quả. Có thể cần nghiên cứu thêm về mối quan hệ phi tuyến tính trong tương lai.

2.2 Cấu trúc chương trình

2.2.1 Class OLSLinearRegression

Class này được tái sử dụng trong tài liệu lab04 mà thầy đã hướng dẫn ở lớp:

Hàm	Input	Output
`fit(X, y)`	- `X`: `np.ndarray` — Dữ liệu đầu vào (ma trận đặc trưng) - `y`: `np.ndarray` — Dữ liệu đầu ra (vector mục tiêu)	`self`: Trả về chính đối tượng của class
`get_params()`	Không có	`self.w`: `np.ndarray` — Vector tham số tối ưu
`predict(X)`	`X`: `np.ndarray` — Dữ liệu đầu vào	`np.ndarray`: Dự đoán đầu ra

Chức năng chi tiết:

- `fit`: Huấn luyện mô hình bằng phương pháp Bình phương tối thiểu (Ordinary Least Squares) sử dụng Moore–Penrose pseudoinverse để tìm vector trọng số **w** (Weights) tối ưu.
- `get_params`: Lấy ra vector trọng số **w** (weights) sau khi huấn luyện.
- `predict`: Sử dụng **w** để dự đoán giá trị đầu ra tương ứng với dữ liệu đầu vào.

Chú ý: Trong hồi quy tuyến tính, nghiệm của **OLS**:

$$\mathbf{w} = (X^\top X)^{-1} X^\top y$$

Khi $X^\top X$ không khả nghịch (hoặc X không vuông), ta dùng **pseudoinverse**:

$$\mathbf{w} = X^+ y$$

Trong đó $X^+ = \text{pinv}(X)$ và dùng để xấp xỉ cho $(X^\top X)^{-1} X^\top$.

2.2.2 Các hàm hỗ trợ

Hàm `df_to_numpy()`:

Input:

- `dataframe` (`pd.DataFrame`): Dữ liệu đầu vào dưới dạng `pandas DataFrame`.

Output:

- `np.ndarray`: Mảng `NumPy` chứa dữ liệu từ `DataFrame`.

Chức năng: Chuyển đổi một `pandas DataFrame` sang mảng `NumPy` để thuận tiện cho các phép nhân ma trận.

Hàm `cal_mse()`:

Input:

- `y` (`np.ndarray`): Giá trị thực tế của biến mục tiêu.
- `y_predicted` (`np.ndarray`): Giá trị dự đoán của mô hình.

Output:

- `float`: Giá trị Mean Squared Error (MSE).

Chức năng: Tính toán sai số trung bình bình phương (MSE) giữa giá trị thực tế và giá trị dự đoán, dùng để đánh giá độ chính xác của mô hình với công thức:

$$MSE = \frac{1}{n}\sum_{n}^{i=1}(y_i - \hat{y}_i)^2$$

MSE được ưa chuộng khi làm với dữ liệu có giá trị đầu ra nhỏ (khoảng 100 giống trong dataset này) vì sai số càng lớn, khi hàm MSE trả về giá trị càng lớn, và khi sai số càng nhỏ thì hàm MSE trả về giá trị càng nhỏ. Từ đó có thể dễ dàng so sánh với các mô hình khác.

Hàm `plot_compare()`**:**

Input:

- `y_test` (`np.ndarray`): Giá trị thực tế của tập kiểm tra, dạng (n_samples,).
- `y_predicted` (`np.ndarray`): Giá trị dự đoán tương ứng, dạng (n_samples,).

Output:

- `None`: Hàm tạo và hiển thị biểu đồ so sánh, không trả về giá trị.

Chức năng: Vẽ biểu đồ đường so sánh trực quan giữa giá trị thực tế và giá trị dự đoán. Để tránh biểu đồ bị rối khi dữ liệu lớn, hàm chỉ hiển thị mỗi 20 điểm dữ liệu. Hai đường biểu diễn được đặt trên cùng một hệ trục để dễ dàng đánh giá độ khớp giữa dự đoán và thực tế. Hàm này sẽ được sử dụng để đánh giá mô hình.

Hàm `add_bias()`**:**

Input:

- `X` (`np.ndarray`): Ma trận đặc trưng đầu vào có kích thước $(m\_samples, n\_features)$.

Output:

- `np.ndarray`: Ma trận đặc trưng mới có kích thước $(m\_samples, n\_features + 1)$, trong đó cột đầu tiên toàn giá trị 1 (bias term) và các cột còn lại là dữ liệu gốc.

Chức năng: Thêm một cột giá trị 1 vào đầu ma trận đặc trưng để biểu diễn bias (intercept) khi xây dựng các mô hình như hồi quy tuyến tính.

2.2.3 Hàm K-Fold Cross Validation

Input:

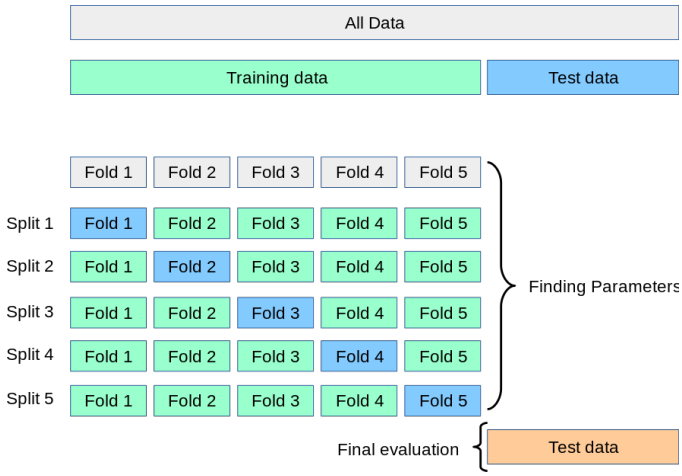
- `X` (`np.ndarray`): Ma trận đặc trưng .
- `y` (`np.ndarray`): Vector giá trị mục tiêu.
- `model` (`object`): Đối tượng mô hình cần đánh giá. Mô hình đã cài đặt 2 hàm `fit(X, y)` và `predict(X)`.
- `input_indices` (`np.ndarray`): Mảng chỉ số xác định thứ tự các mẫu để chia fold (phiên bản đã được xáo trộn).
- `k` (`int`, mặc định = 5): Số lượng fold trong cross-validation.

Output:

- `float`: Giá trị trung bình Mean Squared Error (MSE) của mô hình trên tất cả các fold.

Chức năng: Thực hiện *k-fold cross-validation* để đánh giá hiệu năng mô hình. Hàm chia dữ liệu thành k tập con có kích thước gần bằng nhau, lần lượt dùng $k-1$ tập để huấn luyện và tập còn lại để kiểm tra. Lặp quá trình này k lần và trả về MSE trung bình.

Cách hoạt động chi tiết:



Hình 4: K-fold Cross-Validation Flow

1. Tính kích thước mỗi fold: chia đều số mẫu thành k phần, nếu dư mẫu thì phân bổ đều vào các fold đầu.

2. Duyệt qua từng fold:
   (a) Xác định *val_idx* (chỉ số mẫu cho tập validation) và *train_idx* (chỉ số mẫu cho tập huấn luyện).
   (b) Lấy dữ liệu huấn luyện và kiểm tra từ `X` và `y` dựa trên các chỉ số này.
   (c) Huấn luyện mô hình bằng `model.fit(X_train, y_train)`.
   (d) Dự đoán đầu ra trên tập validation: $\hat{y} =$ `model.predict(X_val)`.
   (e) Tính MSE của fold hiện tại và lưu lại.
3. Sau khi chạy hết k fold, tính trung bình tất cả các MSE đã lưu để ra kết quả cuối.

2.3 Xây dựng và thiết kế mô hình dự đoán

2.3.1 Mô hình sử dụng toàn bộ 5 đặc trưng

Mô hình được khớp khi sử dụng toàn bộ 5 đặc trưng để tìm các trọng số và hệ số tự do tương ứng. Sau khi tính toán, ta có được **mô hình hồi quy** với phần trọng số được làm tròn đến 3 chữ số thập phân như dưới đây:

$$\text{Student Performance} = -33.969 + 2.852 \cdot \text{HS} + 1.018 \cdot \text{PS} + 0.604 \cdot \text{EA} + 0.474 \cdot \text{SH} + 0.192 \cdot \text{SQPP}$$

Trong đó:

- HS: Hours Studied
- PS: Previous Scores
- EA: Extracurricular Activities
- SH: Sleep Hours
- SQPP: Sample Question Papers Practiced

Các nhận xét và đánh giá sẽ nằm ở: Kết quả và kết luận

2.3.2 Các mô hình chỉ sử dụng 1 đặc trưng duy nhất

Ở phần này, mô hình sử dụng duy nhất một đặc trưng, tìm trọng số và hệ số tự do tương ứng. Từ đó đánh giá 5 mô hình sử dụng 5 đặc trưng riêng biệt, và tìm ra mô hình tốt nhất. **Với các bước thực hiện:**

1. Xáo trộn dữ liệu một lần duy nhất
2. Đối với mỗi đặc trưng, chạy K-fold với số lượng fold mặc định là 5, tìm ra được Mean MSE của từng mô hình.
3. Chọn mô hình với Mean MSE thấp nhất, vì đây là mô hình khớp dữ liệu nhất
4. Huấn luyện lại mô hình trên toàn bộ tập dữ liệu.
5. Thử nghiệm với và đánh giá với tập `p03.test.csv`.

Mô hình tuyến tính tốt nhất tìm được:

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

Các nhận xét và đánh giá sẽ nằm ở: Kết quả và kết luận

2.3.3 Tự thiết kế mô hình

Mô hình 1: Sử dụng 2 đặc trưng Previous Score (Standardized) và Hours Studied (Rescaled)

Giống như những đánh giá ở phần EDA, ta dễ dàng thấy được có hai đặc trưng có mối quan hệ tuyến tính tốt nhất với Performance Index chính là **Previous Scores** và **Hours Studied**.
Tuy nhiên: **Previous Scores** có phân phối khá phẳng. Không thấy lệch mạnh về hai đầu. Và khi làm việc với mô hình hồi quy, ta nên chuẩn hóa đặc trưng này để có thể làm việc ổn định hơn. Đồng thời cần rescale lại Hours Studied để tránh sự phụ thuộc quá nhiều vào trọng số của Previous Scores.

Mô hình 2: Sử dụng 3 đặc trưng Previous Scores, Hours Studied (Squared), Extracurricular Activities

Ở mô hình 2, ta vẫn tiếp tục sử dụng đặc trưng **Previous Scores**. Vì hệ số tương quan giữa Performance Index với **Hours Studied** (HS) là 0.37, ta thử **bình phương Hours Studied** để giảm sự phụ thuộc trọng số của **Hours Studied**, và thử sử dụng thêm một đặc trưng nhị phân **Extracurricular Activities** (EA) để thêm một số thành phần gây nhiễu. Nếu mô hình này hoạt động tốt thì có vẻ phương pháp bình phương và gây nhiễu sẽ tối ưu hơn trong tập dữ liệu này.

Mô hình 3: Sử dụng 4 đặc trưng Previous Scores (Standardized), Hours Studied (Rescaled), Sleep Hours (Rescaled), Sample Question Papers Practiced (Rescaled)

Sau khi kiểm tra MSE thử mô hình 1 và 2 ở trước, ta thấy mô hình có vẻ hoạt động tốt với dữ liệu đã chuẩn hóa và rescaled so với mô hình sử dụng đặc trưng bình phương. Vì vậy, mô hình 3 sẽ sử dụng **Previous Scores** đã chuẩn hóa cùng với (**Hours Studied**, **Sleep Hours** (SH) và **Sample Question Papers Practiced** (SQPP)) đã rescale, vì theo như ma trận tương quan thì SH và SQPP cũng có một chút tương quan nhẹ.

Kết quả:

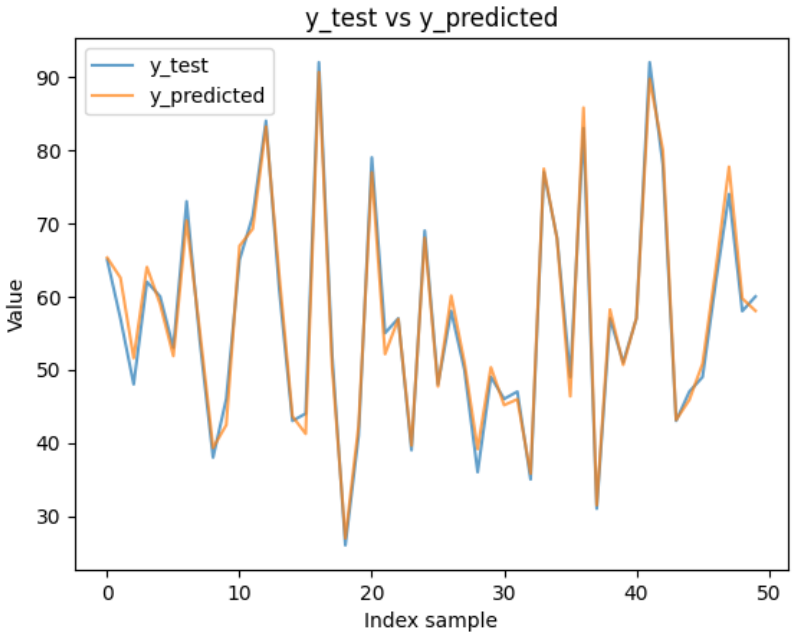
Mô hình tốt nhất tìm được trong 3 mô hình là **mô hình 3**:

$$\text{Student Performance} = 36.983 + 17.682 \cdot \frac{PS - 69.396}{17.369} + 25.673 \cdot \frac{HS}{9} + 4.229 \cdot \frac{SH}{9} + 1.739 \cdot \frac{SQPP}{9}$$

với MSE: **4.209**

3 Kết quả và kết luận

3.1 Mô hình sử dụng toàn bộ 5 đặc trưng



Hình 5: So sánh giữa kết quả dự đoán và kết quả thực tế trong tập kiểm tra

Mô hình:

$$\text{Student Performance} = -33.969 + 2.852 \cdot \text{HS} + 1.018 \cdot \text{PS} + 0.604 \cdot \text{EA} + 0.474 \cdot \text{SH} + 0.192 \cdot \text{SQPP}$$

có **MSE** trên tập kiểm tra là: **4.092**

Nhận xét:

- Kết quả dự đoán của mô hình có xu hướng bám sát khá tốt với giá trị thực tế trên tập kiểm tra, thể hiện qua đường dự đoán (màu cam) gần trùng với đường giá trị thật (màu xanh) ở hầu hết các mẫu. Sai số trung bình bình phương (MSE) đạt 4.092, cho thấy mức độ chênh lệch giữa dự đoán và thực tế là tương đối thấp, phù hợp với một mô hình hồi quy tuyến tính đơn giản.

- Hệ số của các biến cho thấy HS (Hours Studied) có ảnh hưởng lớn nhất đến kết quả học tập, tiếp theo là PS (Previous Scores), EA (Extracurricular Activities), SH (Sleep Hours) và cuối

cùng là SQPP (SSample Question Papers Practiced). Dấu dương của các hệ số chỉ ra rằng khi giá trị của các đặc trưng này tăng, điểm dự đoán cũng tăng, phản ánh mối quan hệ tích cực giữa các yếu tố và thành tích học tập.

- Nhìn chung, mô hình không chỉ có độ chính xác tốt trên tập kiểm tra mà còn thể hiện ý nghĩa logic trong mối quan hệ giữa đặc trưng và biến mục tiêu, điều này giúp tăng độ tin cậy khi áp dụng vào thực tế.

3.2 Các mô hình chỉ sử dụng 1 đặc trưng duy nhất

3.2.1 So sánh giữa 5 đặc trưng

Ta có bảng kết quả Mean MSE sau khi chạy K-fold trong các đặc trung riêng lẻ như sau:

STT	**Mô hình với 1 đặc trưng**	**MSE**
1	Hours Studied	318.038
2	Previous Scores	60.079
3	Extracurricular Activities	368.112
4	Sleep Hours	367.599
5	Sample Question Papers Practiced	367.656

Bảng 5: Kết quả MSE của mô hình với từng đặc trưng riêng lẻ

Nhật xét:

Đúng như phần đánh giá trong Ma trận tương quan, các đặc trưng như **Hours Studied** (0.37), **Previous Scores** (0.91) sẽ có sai số nhỏ hơn, do có độ tương quan với **Performance Index** lớn hơn. Ngược lại, 3 đặc trưng còn lại có sai số lớn cho thấy các hệ số tương quan trong ma trận đánh giá chính xác sự phụ thuộc tuyến tính giữa các đặc trưng.
Bảng kết quả Mean MSE cho thấy sự khác biệt rõ rệt về khả năng dự đoán của các mô hình khi chỉ sử dụng một đặc trưng:

- **Previous Scores** đạt MSE **60.079**, thấp vượt trội so với các đặc trưng khác. Điều này cho thấy điểm số trước đó của học sinh là yếu tố dự đoán mạnh nhất đối với thành tích hiện tại. Mối quan hệ giữa biến này và biến mục tiêu có thể mang tính tuyến tính và ổn định, giúp mô hình đạt sai số nhỏ.

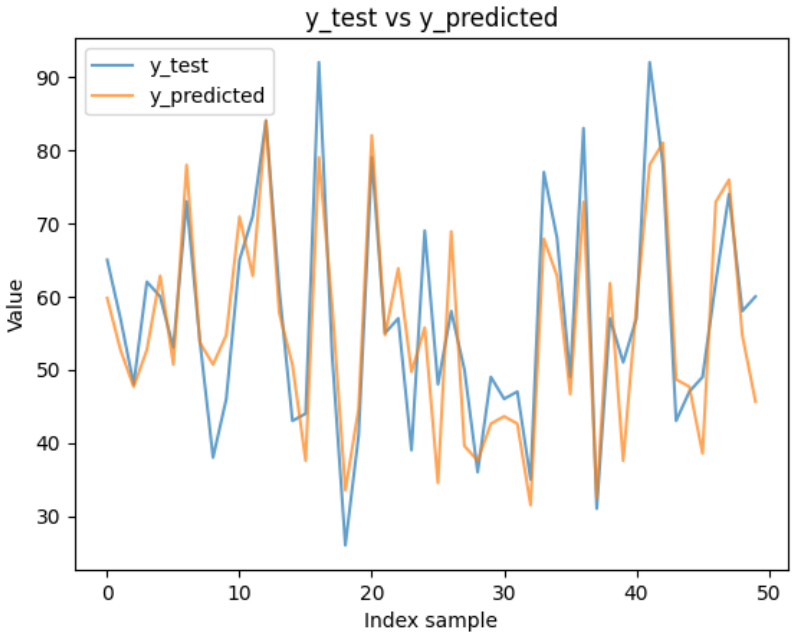
- **Hours Studied** có MSE **318.038**, đứng thứ hai. Mặc dù vẫn có ảnh hưởng đến kết quả học tập, nhưng mức MSE cao hơn nhiều so với Previous Scores, cho thấy số giờ học tuy quan trọng nhưng chưa đủ mạnh để dự đoán chính xác nếu không kết hợp thêm thông tin khác.
- **Extracurricular Activities**, **Sleep Hours**, và **Sample Question Papers Practiced** đều có MSE khoảng **368**, cho thấy khi chỉ dùng riêng lẻ, các yếu tố này gần như không có khả năng dự đoán tốt. Nguyên nhân có thể là mối liên hệ với kết quả học tập yếu, phi tuyến, hoặc phụ thuộc nhiều vào các biến khác.

Kết luận:

- **Previous Scores** là đặc trưng đơn lẻ mạnh nhất.
- **Hours Studied** có ảnh hưởng trung bình, vẫn hữu ích nhưng cần bổ sung thêm thông tin.
- Ba đặc trưng còn lại hầu như không hiệu quả khi dùng đơn lẻ, nhưng có thể đóng vai trò hỗ trợ trong mô hình tổng hợp nhiều đặc trưng.

3.2.2 Đặc trưng cho ra mô hình tốt nhất

Sau khi khớp dữ liệu với toàn bộ tập huấn luyện với đặc trưng **Previous Scores**, ta có:



Hình 6: Mô hình sử dụng đặc trưng Previous Scores

Mô hình:

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

với giá trị MSE trên tập kiểm tra là: **58.888**

Nhận xét:

- Mô hình hồi quy tuyến tính sử dụng duy nhất đặc trưng **Previous Scores** cho thấy khả năng dự đoán khá.
- Đường dự đoán bám sát với đường giá trị thực tế ở phần lớn các mẫu trong tập kiểm tra, mặc dù vẫn có những điểm sai lệch cục bộ.
- Giá trị **MSE = 58.888** là tương đối thấp so với các mô hình đơn biến khác, khẳng định **Previous Scores** là yếu tố dự đoán mạnh mẽ đối với kết quả học tập.
- Hệ số dương 1.011 trong phương trình hồi quy cho thấy mối quan hệ tuyến tính thuận: khi điểm số trước đó tăng, điểm dự đoán cũng tăng gần như tương ứng.
- Điều này phản ánh tính hợp lý về mặt thực tiễn và củng cố niềm tin vào tính hữu ích của biến này trong các mô hình dự đoán.

3.3 Tự thiết kế mô hình

3.3.1 3 mô hình tự thiết kế

Ta có bảng kết quả Mean MSE sau khi chạy K-fold trong 3 mô hình như sau:

STT	Mô hình	MSE
1	Sử dụng 2 đặc trưng Previous Score (Standardized) và Hours Studied (Rescaled)	5.199
2	Sử dụng 3 đặc trưng Previous Scores, Hours Studied (Squared), Extracurricular Activities	7.695
3	Sử dụng 4 đặc trưng Previous Scores (Standardized), Hours Studied (Rescaled), Sleep Hours (Rescaled), Sample Question Papers Practiced (Rescaled)	4.256

Bảng 6: Kết quả MSE của các mô hình khác nhau

Nhận xét

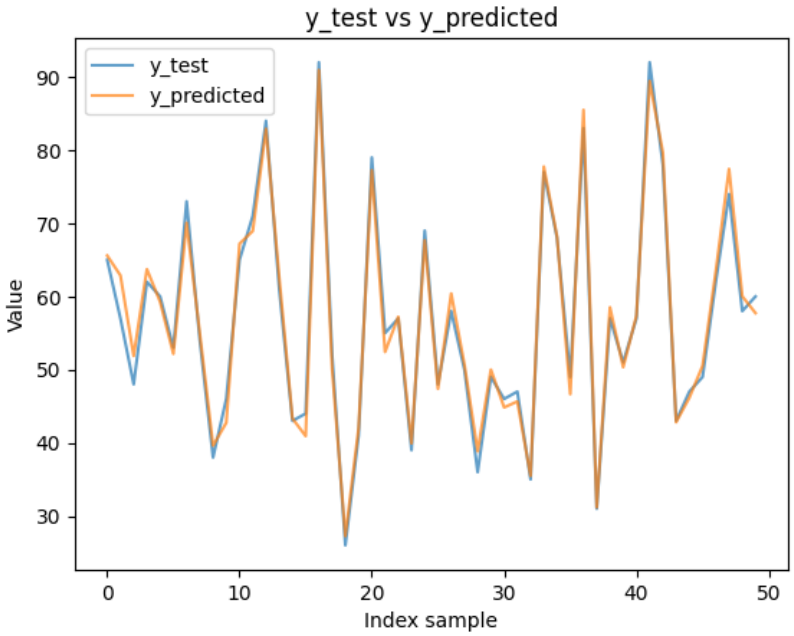
Bảng kết quả Mean MSE cho thấy hiệu quả dự đoán thay đổi đáng kể khi thay đổi số lượng và loại đặc trưng sử dụng:

- **Mô hình 1** (2 đặc trưng: Previous Score (Standardized) và Hours Studied (Rescaled)) đạt **MSE = 5.199**. Đây là kết quả khá tốt, cho thấy việc kết hợp hai biến quan trọng này giúp mô hình dự đoán tương đối chính xác, giảm sai số đáng kể so với mô hình đơn biến.
- **Mô hình 2** (3 đặc trưng: Previous Scores, Hours Studied (Squared), Extracurricular Activities) có **MSE = 7.695**, cao hơn mô hình 1. Điều này cho thấy việc bổ sung biến Extracurricular Activities và biến Hours Studied dạng bình phương không cải thiện, thậm chí làm giảm độ chính xác, có thể do nhiễu hoặc quan hệ phi tuyến chưa được mô hình tuyến tính khai thác tốt.
- **Mô hình 3** (4 đặc trưng: Previous Scores (Standardized), Hours Studied (Rescaled), Sleep Hours (Rescaled), Sample Question Papers Practiced (Rescaled)) đạt **MSE = 4.256** — thấp nhất trong 3 mô hình. Điều này chứng tỏ khi chọn lọc và chuẩn hóa hợp lý các đặc trưng có ý nghĩa, mô hình có thể cải thiện hiệu quả dự đoán rõ rệt.

Kết luận: Mô hình 3 cho kết quả tốt nhất, tiếp đến là mô hình 1, và cuối cùng là mô hình 2. Việc lựa chọn đặc trưng phù hợp và xử lý dữ liệu đúng cách (chuẩn hóa, scale) đóng vai trò then chốt trong việc giảm MSE và nâng cao độ chính xác dự đoán. **Tuy nhiên**, 3 mô hình này vẫn có sai số cao hơn mô hình sử dụng toàn bộ 5 đặc trưng được thực hiện ở câu 2a, từ đó cần tìm hiểu thêm các mô hình tối ưu hơn trong tương lai.

3.3.2 Mô hình tốt nhất

Chọn mô hình 3 và khớp lại trên toàn bộ tập huấn luyện:



Hình 7: Thể hiện của mô hình 3 (Mô hình tốt nhất)

Với công thức truy hồi:

$$\text{Student Performance} = 36.983 + 17.682 \cdot \frac{PS - 69.396}{17.369} + 25.673 \cdot \frac{HS}{9} + 4.229 \cdot \frac{SH}{9} + 1.739 \cdot \frac{SQPP}{9}$$

Với **MSE** sau khi huấn luyện trên toàn bộ tập train: **4.209**

Đánh giá mô hình:

- Mô hình 3, được lựa chọn là mô hình tốt nhất, thể hiện khả năng dự đoán rất sát so với giá trị thực tế.
- Đường dự đoán (màu cam) gần như trùng khớp với đường giá trị thật (màu xanh) trên hầu hết các điểm dữ liệu trong tập kiểm tra, chỉ xuất hiện sai lệch nhỏ tại một số mẫu. Điều này cho thấy mô hình đã học được mối quan hệ tuyến tính quan trọng giữa các đặc trưng và biến mục tiêu.
- Việc tiền xử lý dữ liệu hợp lý đã giúp giảm chênh lệch giữa các đặc trưng, đảm bảo các hệ số hồi quy phản ánh đúng tầm quan trọng tương đối của từng biến.

Kết luận: Mô hình 3 là lựa chọn tối ưu nhờ:

- Kết hợp hợp lý các đặc trưng quan trọng.

- Tiền xử lý dữ liệu phù hợp (chuẩn hóa và scale).
- Sai số dự đoán nhỏ và đường dự đoán khớp sát với dữ liệu thực tế.

Điều này chứng tỏ mô hình không chỉ phù hợp về mặt thống kê mà còn hợp lý về mặt ý nghĩa thực tiễn.

3.4 Kết luận

Chưa thể tìm được mô hình tối ưu, mô hình hồi quy tuyến tính OLS hoạt động tốt nhất khi sử dụng toàn bộ đặc trưng (MSE thấp nhất $\tilde{4}.09$), chứng tỏ kết hợp đa yếu tố giúp dự đoán chính xác hơn. Previous Scores và Hours Studied là các yếu tố then chốt, phản ánh thực tế: nền tảng kiến thức và nỗ lực học tập quyết định thành tích.

Hạn chế: Chỉ đo lường tương quan tuyến tính; có thể tồn tại quan hệ phi tuyến (ví dụ: giấc ngủ tối ưu ở mức trung bình). MSE của mô hình đơn giản vẫn cao ở một số trường hợp, gợi ý cần mô hình phức tạp hơn (như hồi quy phi tuyến hoặc machine learning nâng cao) trong tương lai.

Bài làm hoàn thành 100% yêu cầu, sử dụng thư viện phù hợp (Pandas, NumPy, Matplotlib, Seaborn) và phương pháp đánh giá chuẩn (K-fold cross-validation, MSE). Kết quả có ý nghĩa thực tiễn, hỗ trợ cải thiện thành tích học tập bằng cách tập trung vào thời gian học và điểm số trước đó.

4 Tài liệu tham khảo

1. Quy'Blog, Các phương pháp scale dữ liệu trong machine learning, url, [11/08/2025]
2. Scikit-learn, Cross-validation: evaluating estimator performance, url, [11/08/2025]
3. Jakia Mun, Student Performance: Regression and EDA, url, [11/08/2025]
4. StatQuest with Josh Starmer, Machine Learning Fundamentals: Cross Validation, url, [10/08/2025]
5. NeuralNine, Matplotlib Full Python Course - Data Science Fundamentals, url,[11/08/2025]

5 Lời cảm ơn

Em xin cảm ơn thầy *Nguyễn Ngọc Toàn* và *Trần Hà Sơn* đã giúp đỡ, giải đáp thắc mắc của em trong quá trình học cũng như trong quá trình làm đồ án này.
Đồ án này có sự giúp đỡ chatGPT và Grok trong việc viết Latex, tìm kiếm thông tin các hàm để trực quan hóa dữ liệu.