ĐẠI HỌC QUỐC GIA
THÀNH PHỐ HỒ CHÍ MINH

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

MTH00057 - Toán ứng dụng và thống kê cho Công nghệ thông tin

BÁO CÁO ĐỒ ÁN 3

Linear Regression

Họ tên	**MSSV**
Nguyễn Lê Hồ Anh Khoa	23127211

Giảng viên hướng dẫn

Nguyễn Văn Quang Huy
Trần Hà Sơn
Nguyễn Đình Thúc
Nguyễn Ngọc Toàn

Ngày 13 tháng 8 năm 2025

Mục lục

1 Thông tin sinh viên

Họ và tên: Nguyễn Lê Hồ Anh Khoa. MSSV: 23127211. Lớp: 23CLC09

2 Đánh giá

2.1 Bảng tự đánh giá các yêu cầu đã hoàn thành

Bảng 1: Bảng tự đánh giá đồ án

STT	**Yêu cầu**	**Mức độ hoàn thành**
1	Thực hiện phân tích khám phá dữ liệu	100%
2	Sử dụng toàn bộ 5 đặc trưng để xây dựng mô hình	100%
3	Xây dựng mô hình sử dụng duy nhất 1 đặc trưng, tìm mô hình cho kết quả tốt nhất	100%
4	Sinh viên tự xây dựng/thiết kế mô hình, tìm mô hình cho kết quả tốt nhất	100%
	Tổng cộng	**100%**

2.2 Đánh giá tổng thể mức độ hoàn thành của bài nộp

Bài nộp đã hoàn thành đầy đủ các yêu cầu đề ra trong bài tập. Tất cả các yêu cầu đều đã được cài đặt và kiểm thử thành công. Tổng thể, bài nộp đã hoàn thành 100% các yêu cầu đề ra.

3 Ý tưởng thực hiện

Mục tiêu của đề tài là xây dựng mô hình dự đoán Student Performance (Performance Index) từ các đặc trưng hành vi – học tập, đồng thời so sánh nhiều phương án mô hình hóa để chọn mô hình gọn nhẹ, dễ diễn giải và có sai số thấp.

3.1 Yêu cầu 1: Thực hiện phân tích khám phá dữ liệu

Để thực hiện phân tích khám phá dữ liệu, chúng ta sẽ sử dụng các hàm thống kê mô tả và trực quan hóa dữ liệu để hiểu rõ hơn về các đặc trưng của dữ liệu. Các bước thực hiện được tham khảo từ tài liệu [1] và [2] bao gồm:

- Tải dữ liệu từ file CSV.
- Thống kê nhanh số dòng, số cột, đặc trưng và kiểu dữ liệu của từng đặc trưng.
- Thống kê số lượng giá trị bị thiếu (null), duy nhất (unique) trong từng đặc trưng.

- Thống kê số lượng bộ dữ liệu bị trùng
- Làm sạch dữ liệu: loại bỏ hoặc thay thế các giá trị thiếu, xử lý ngoại lệ, loại bỏ các bản ghi trùng nhau
- Phân tích đơn biến nhằm nắm bắt phân phối của từng biến, phát hiện ngoại lệ và xác định biến cần biến đổi hoặc chuẩn hóa.
- Phân tích hai biến nhằm tìm hiểu mối quan hệ giữa các biến với biến mục tiêu (Performance Index).
- Phân tích tương tác và mối quan hệ phức tạp giữa các biến nhằm phát hiện các mẫu và mối liên hệ không tuyến tính.
- Trực quan hóa mối quan hệ giữa các đặc trưng và biến mục tiêu (Performance Index) bằng biểu đồ scatter plot, box plot, heatmap.

3.2 Yêu cầu 2a: Sử dụng toàn bộ 5 đặc trưng để xây dựng mô hình

Để xây dựng mô hình dự đoán Student Performance (Performance Index), chúng ta sẽ sử dụng toàn bộ 5 đặc trưng đã được phân tích ở trên. Các bước thực hiện bao gồm:

- Huấn luyện 1 lần duy nhất cho 5 đặc trưng trên cho toàn bộ tập huấn luyện.
- Thể hiện công thức cho mô hình hồi quy tuyến tính với 5 đặc trưng.
- Đánh giá mô hình: sử dụng chỉ số MSE để đánh giá hiệu suất của mô hình trên tập kiểm tra.

3.3 Yêu cầu 2b: Xây dựng mô hình sử dụng duy nhất 1 đặc trưng, tìm mô hình cho kết quả tốt nhất

Để tìm mô hình tốt nhất sử dụng duy nhất 1 đặc trưng, chúng ta sẽ thực hiện các bước sau:

- Thử nghiệm trên toàn bộ (5) đặc trưng đề bài cung cấp.
- Xáo trộn dữ liệu
- Sử dụng k-fold cross-validation (k=5) để đánh giá mô hình.
- Báo cáo kết quả MSE từ k-fold cross-validation của từng mô hình.
- Chọn mô hình có kết quả tốt nhất dựa trên chỉ số MSE.
- Tìm công thức hồi quy tuyến tính cho mô hình tốt nhất.
- Đánh giá mô hình: sử dụng chỉ số MSE để đánh giá hiệu suất của mô hình trên tập kiểm tra.

3.4 Yêu cầu 2c: Sinh viên tự xây dựng/thiết kế mô hình, tìm mô hình cho kết quả tốt nhất

Để tự xây dựng/thiết kế mô hình, ta sẽ dựa vào các đặc trưng của dữ liệu đã được khám phá ở Yêu cầu 1. Ý tưởng là sử dụng các đặc trưng có mối quan hệ mạnh với biến mục tiêu (Performance Index) để xây dựng mô hình. Các mô hình được xây dựng là:

3.4.1 Mô hình 1:

Dựa trên phân tích Count Plot và heatmap, tìm ra các đặc trưng có ảnh hưởng lớn nhất đến biến mục tiêu (Performance Index). Sử dụng các đặc trưng này để xây dựng mô hình hồi quy tuyến tính đơn giản:

3.4.2 Mô hình 2:

Từ phân tích Count Plot và heatmap, loại bỏ đặc trưng có ảnh hưởng nhỏ nhất và thêm vào bình phương của đặc trưng ảnh hưởng lớn nhất để tạo ra mô hình có độ chính xác cao hơn:

3.4.3 Mô hình 3:

Sử dụng đặc trưng tương tác giữa các đặc trưng để tạo ra mô hình có độ chính xác cao hơn. Cụ thể, ta sẽ sử dụng tích của các đặc trưng có mối quan hệ mạnh với biến mục tiêu (Performance Index) để tạo ra mô hình hồi quy tuyến tính.

4 Chi tiết thực hiện

4.1 Các thư viện cần thiết

Trong đồ án này, các thư viện chính được sử dụng bao gồm:

- **pandas**: Thư viện mạnh mẽ để thao tác và phân tích dữ liệu dạng bảng (DataFrame). Hỗ trợ đọc/ghi dữ liệu từ nhiều định dạng (CSV, Excel, SQL, v.v.), xử lý dữ liệu thiếu, và thực hiện các thao tác nhóm, lọc, sắp xếp.
- **numpy**: Thư viện xử lý mảng số học hiệu năng cao, cung cấp các phép tính toán vector hóa và đại số tuyến tính, hỗ trợ tối ưu tốc độ xử lý dữ liệu.
- **matplotlib**: Thư viện vẽ đồ thị 2D, được dùng để hiển thị biểu đồ, trực quan hóa kết quả phân tích dữ liệu.
- **seaborn**: Thư viện trực quan hóa dữ liệu dựa trên matplotlib, cung cấp các biểu đồ thống kê đẹp mắt và dễ tùy chỉnh, được dùng cho các biểu đồ như countplot, heatmap, pairplot.
- **IPython.display**: Cung cấp các hàm hỗ trợ hiển thị đối tượng trực tiếp trong môi trường Jupyter Notebook, đặc biệt hữu ích để hiển thị bảng dữ liệu đẹp mắt.
- **typing** (`Dict`, `Any`): Cung cấp các kiểu dữ liệu mạnh mẽ giúp mã dễ đọc và bảo trì.

4.2 Yêu cầu 1: Phân tích khám phá dữ liệu (EDA)

4.2.1 Các hàm chính

- **ds_overview(df)**: Trả về bảng tổng quan dữ liệu, bao gồm số dòng và số cột của DataFrame.
- **dtypes_table(df)**: Trả về tên cột và kiểu dữ liệu của từng cột trong DataFrame.
- **uniques_table(df)**: Thống kê số lượng giá trị duy nhất của mỗi cột, sắp xếp theo thứ tự giảm dần.
- **missing_table(df)**: Thống kê số lượng và tỷ lệ phần trăm giá trị thiếu (NaN) trong mỗi cột.
- **duplicates_table(df)**: Thống kê số lượng bản ghi trùng lặp trong DataFrame.
- **sample_values(df, k)**: Lấy k giá trị không null đầu tiên của mỗi cột, hiển thị dưới dạng bảng.
- **head_tail(df, n)**: Trả về n dòng đầu và n dòng cuối của DataFrame.
- **remove_duplicates(X_df, y_series)**: Loại bỏ các bản ghi trùng lặp trong DataFrame và Series, trả về DataFrame và Series đã loại bỏ trùng lặp.
- **eda_overview(X, y)**: Thực hiện tổng hợp EDA ở mức metadata cho cả tập đặc trưng `X` và biến mục tiêu `y`.
- **plot_histograms(df, bins, exclude)**: Vẽ histogram cho các cột số, loại trừ các cột chỉ định.
- **plot_countplot(df, col)**: Vẽ countplot cho một cột phân loại.
- **plot_scatter_vs_target(df, features, target)**: Vẽ scatter plot giữa từng đặc trưng số và biến mục tiêu.
- **plot_correlation_heatmap(df)**: Vẽ heatmap hiển thị ma trận tương quan giữa các biến số.
- **plot_pairwise_product_correlation_heatmap(df, target)**: Vẽ heatmap tương quan giữa tất cả các tích cặp đặc trưng và biến mục tiêu.
- **plot_boxplots(df, features, target)**: Vẽ boxplot của từng đặc trưng so với biến mục tiêu, tự động binning nếu là dữ liệu số.
- **plot_pairplot(df)**: Vẽ pairplot giữa tất cả các đặc trưng và biến mục tiêu.
- **plot_all_charts(df, target)**: Hàm tổng hợp, lần lượt gọi tất cả các hàm vẽ biểu đồ trên để thực hiện EDA toàn diện.

4.2.2 Khám phá tổng quan dữ liệu

	Property	Value
0	Number of rows	9000
1	Number of columns	6

Bảng 2: Tổng quan dữ liệu cho thấy dữ liệu có 9000 dòng và 6 cột.

	column	dtype	description
0	Hours Studied	int64	Số giờ học tập
1	Previous Scores	int64	Điểm số trước đó
2	Extracurricular Activities	int64	Hoạt động ngoại khóa
3	Sleep Hours	int64	Số giờ ngủ
4	Sample Question Papers Practiced	int64	Số lượng đề thi đã làm
5	Performance Index	float64	Chỉ số hiệu suất

Bảng 3: Kiểu dữ liệu các cột (các đặc trưng)

	column	nunique
0	Performance Index	91
1	Previous Scores	60
2	Sample Question Papers Practiced	10
3	Hours Studied	9
4	Sleep Hours	6
5	Extracurricular Activities	2

Bảng 4: Số lượng giá trị duy nhất của mỗi cột

	column	missing_count	missing_%
0	Hours Studied	0	0.0
1	Previous Scores	0	0.0
2	Extracurricular Activities	0	0.0
3	Sleep Hours	0	0.0
4	Sample Question Papers Practiced	0	0.0
5	Performance Index	0	0.0

Bảng 5: Số lượng và tỷ lệ phần trăm giá trị thiếu

n_duplicates	pct_duplicates
103	1.144

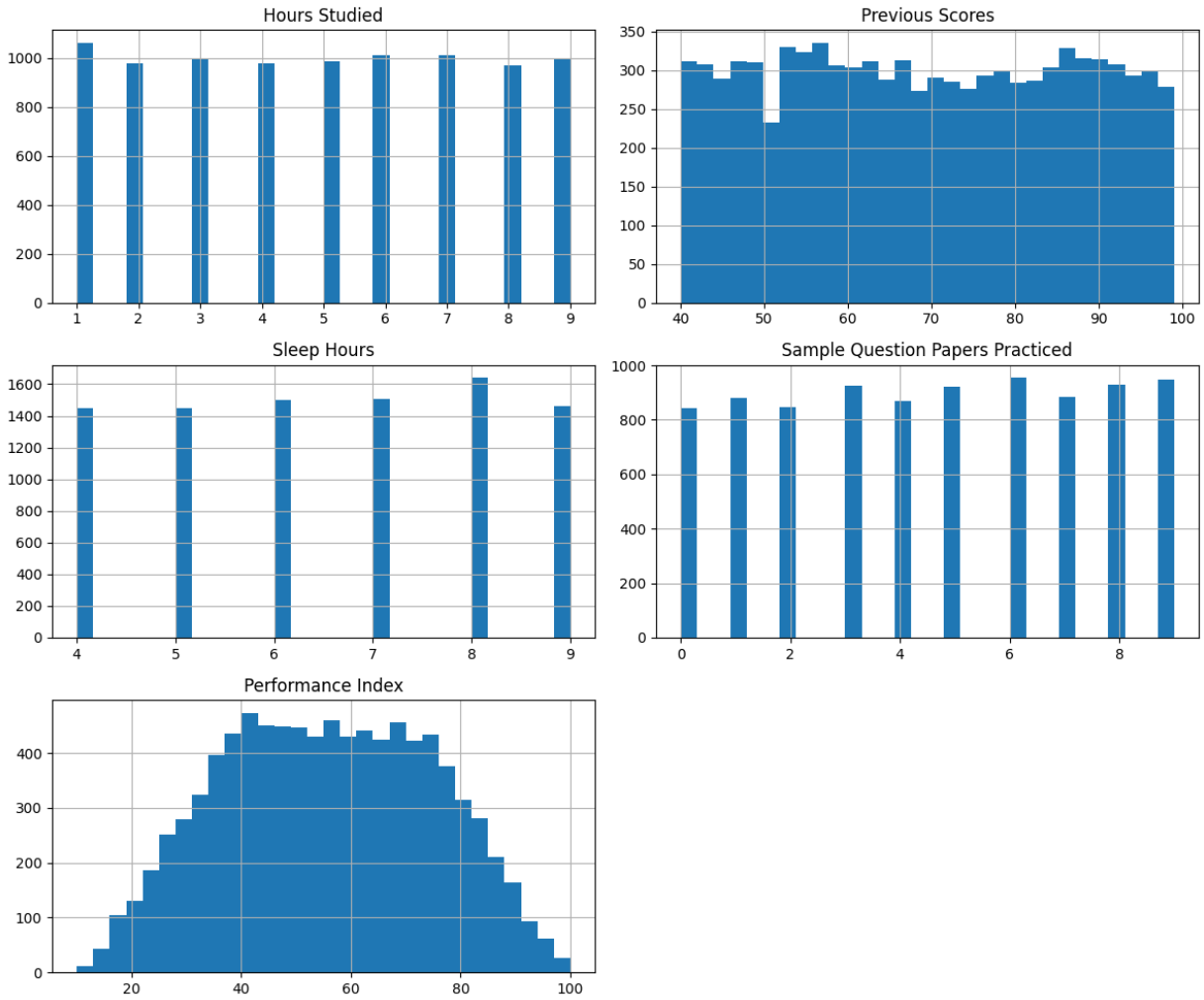
Bảng 6: Số lượng và tỷ lệ phần trăm dòng trùng lặp

Phân tích cho thấy có 103 dòng trùng lặp trong tập dữ liệu, chiếm 1,144% tổng số dòng. Mặc dù tỷ lệ này không cao, nhưng để đảm bảo chất lượng phân tích và độ chính

xác của mô hình, các bản ghi trùng lặp này đã được loại bỏ trước khi tiến hành các phân tích tiếp theo. Việc loại bỏ dữ liệu trùng lặp giúp tránh tình trạng các mẫu trùng lặp được đánh trọng số quá cao trong quá trình huấn luyện mô hình, từ đó nâng cao tính đại diện và độ tin cậy của mô hình hồi quy tuyến tính được xây dựng.

4.2.3 Phân tích đơn biến

Để thực hiện phân tích đơn biến, chúng ta sẽ xem xét từng đặc trưng một cách riêng biệt và lựa chọn các biểu đồ phù hợp để trực quan hóa phân phối của chúng. Các biểu đồ được sử dụng ở phần này bao gồm histogram, countplot:



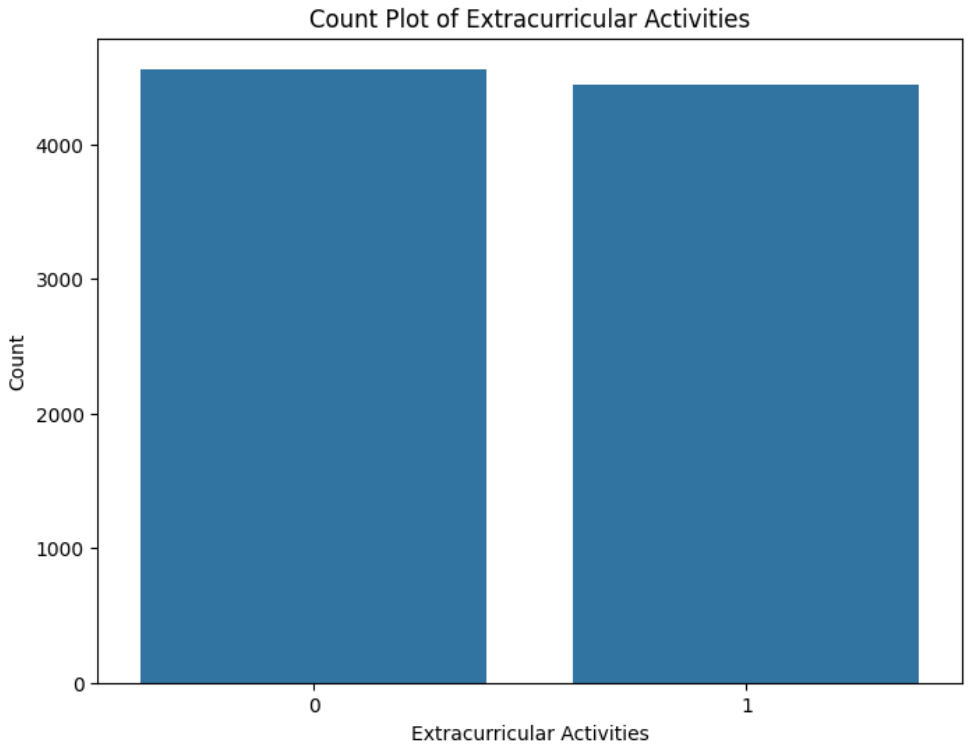
Hình 1: Biểu đồ phân phối của các đặc trưng số

Phân phối của Hours Studied, Previous Scores, Sleep Hours và Sample Question Papers Practiced đều cho thấy sự phân bố khá đồng đều, ngoại trừ Performance Index có dạng phân phối chuẩn rõ rệt.

Điều này cho thấy thói quen học tập và thời gian ngủ của sinh viên nhìn chung ổn định, không tạo ra những giá trị ngoại lệ đáng kể về thành tích. Vì vậy, nhiều khả năng còn tồn tại các yếu tố khác ảnh hưởng đến sự biến động của Performance Index.

Dạng phân phối chuẩn của Performance Index cũng phản ánh rằng phần lớn sinh viên

đạt mức thành tích trung bình, trong khi chỉ một số ít vượt trội hoặc tụt lại so với mặt bằng chung.



Hình 2: Biểu đồ countplot cho các đặc trưng phân loại

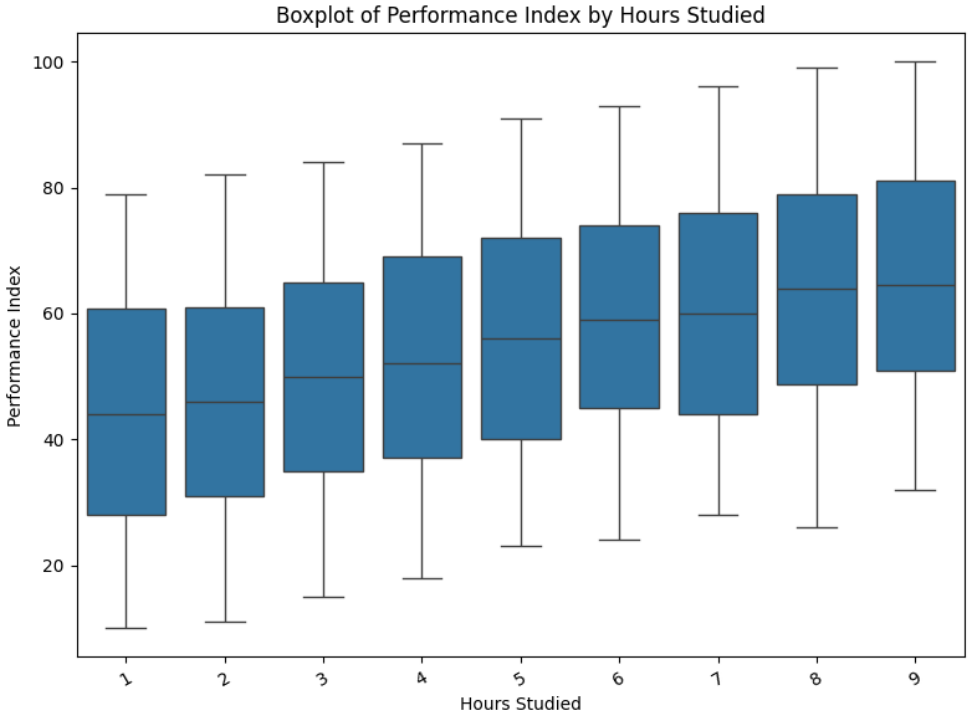
Kết quả phân tích cho thấy tỷ lệ sinh viên tham gia và không tham gia các hoạt động ngoại khóa gần như tương đương nhau.

Điều này phản ánh rằng các hoạt động ngoại khóa hiện nay đã đạt được mức độ tiếp cận tốt và thu hút được sự quan tâm của một bộ phận đáng kể sinh viên.

Tuy nhiên, vẫn tồn tại tiềm năng để mở rộng mức độ tham gia, thông qua việc đa dạng hóa loại hình hoạt động, nâng cao tính hấp dẫn hoặc đẩy mạnh công tác truyền thông nhằm khuyến khích nhiều sinh viên hơn tích cực tham gia.

4.2.4 Phân tích hai biến

Để hiểu rõ hơn về mối quan hệ giữa các đặc trưng và biến mục tiêu (Performance Index), chúng ta sẽ sử dụng scatter plot, heatmap và boxplot.

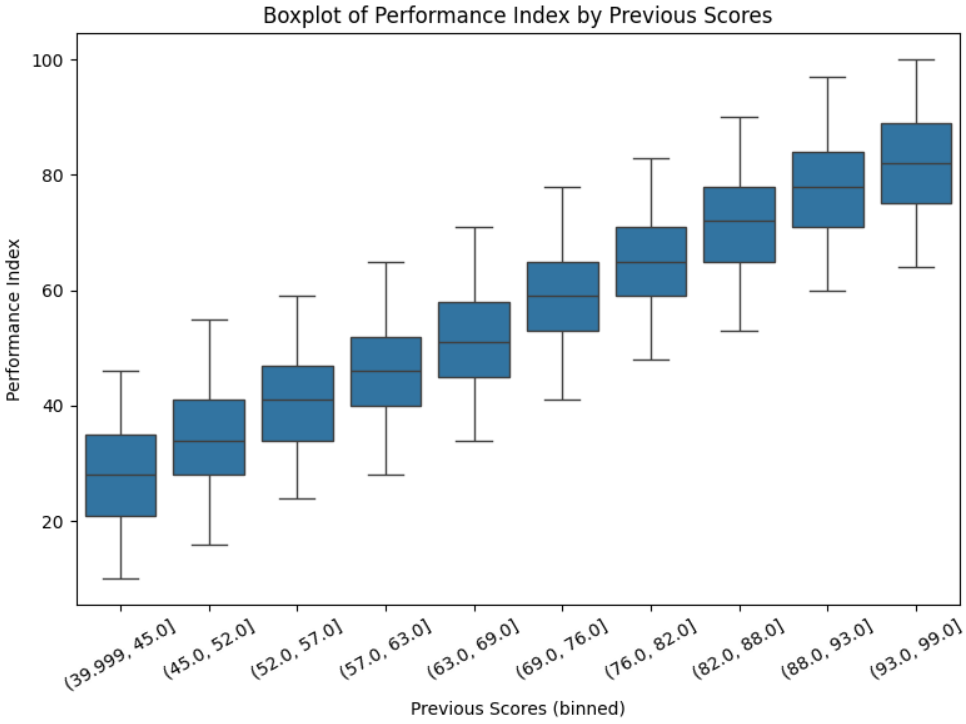


Hình 3: Biểu đồ boxplot giữa Performance Index và Hours Studied

Biểu đồ cho thấy trung vị điểm số tăng dần từ khoảng 45 (1 giờ học) lên 65 (9 giờ học), phản ánh mối quan hệ dương giữa thời gian học và thành tích.

Toàn bộ phân vị (Q1, Q3) dịch chuyển lên trên, nhưng độ phân tán lớn cho thấy kết quả vẫn chịu ảnh hưởng của nhiều yếu tố khác ngoài giờ học.

Ngoại lệ xuất hiện ở cả điểm rất thấp và rất cao ở mọi mức giờ học. Kết quả này phù hợp với hệ số tương quan dương vừa phải giữa Hours Studied và Performance Index, cho thấy nên giữ biến này và xem xét thêm các tương tác khi xây dựng mô hình.

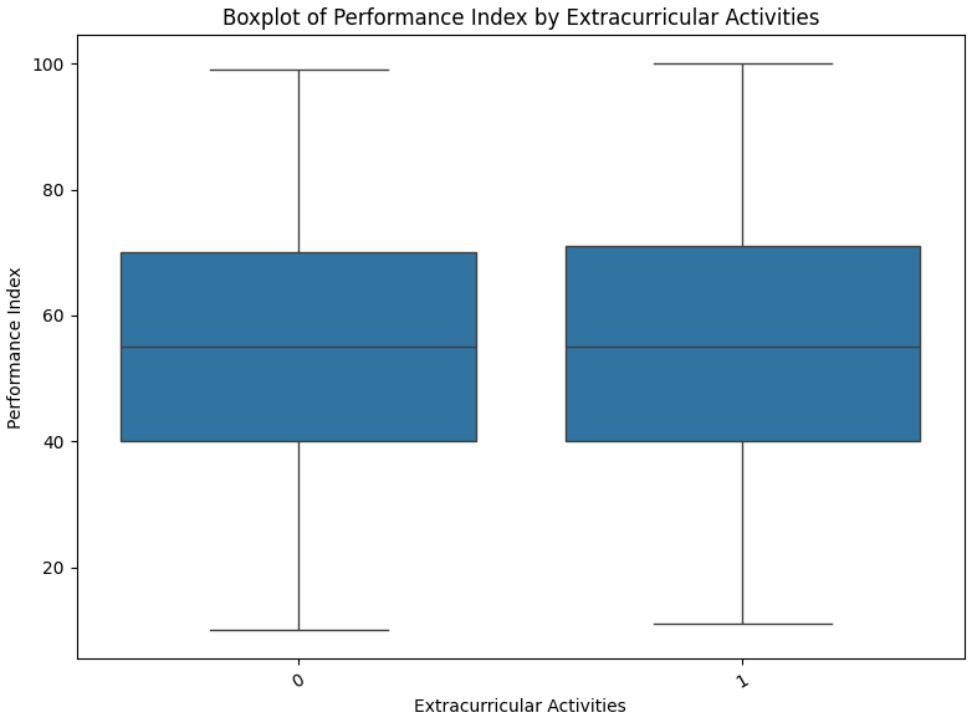


Hình 4: Biểu đồ boxplot giữa Performance Index và Previous Scores

Biểu đồ cho thấy trung vị của Performance Index tăng dần từ khoảng 25 (Previous Scores $\tilde{4}0$) lên khoảng 85 (Previous Scores $\tilde{9}5$), phản ánh mối quan hệ dương mạnh giữa điểm số trước đây và thành tích hiện tại.

Toàn bộ các phân vị (Q1, Q3) dịch chuyển lên trên khi Previous Scores tăng, đồng thời độ phân tán giảm ở nhóm điểm cao cho thấy sự ổn định hơn về kết quả ở các học sinh có nền tảng tốt.

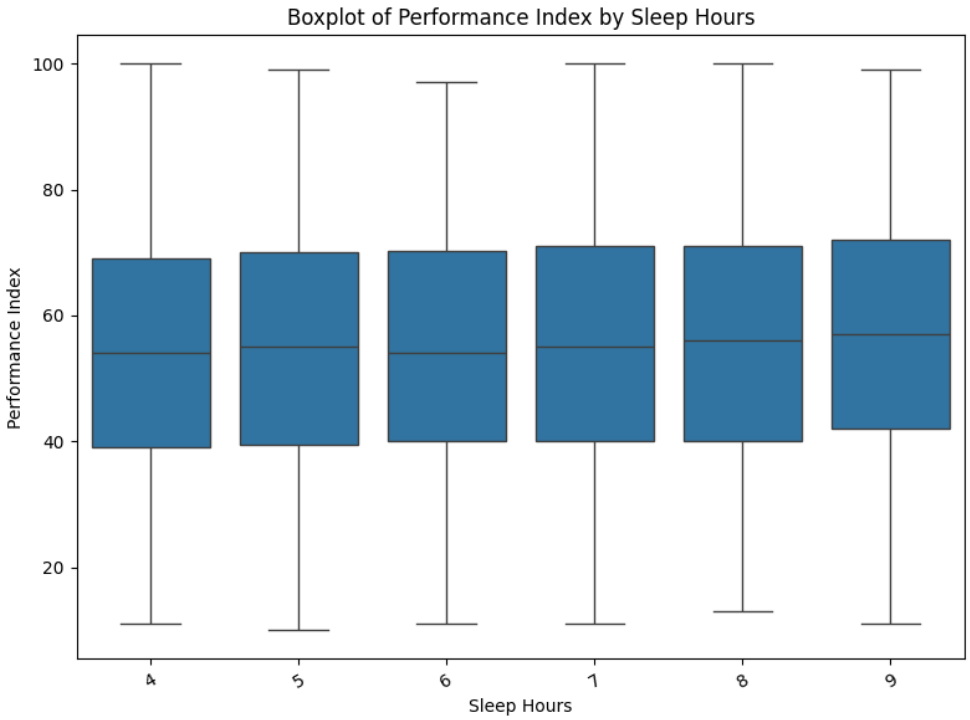
Ngoại lệ vẫn xuất hiện ở cả mức rất thấp và rất cao trong mọi nhóm điểm, cho thấy rằng mặc dù Previous Scores là yếu tố quan trọng, vẫn tồn tại các yếu tố khác ảnh hưởng đến Performance Index. Kết quả này phù hợp với hệ số tương quan cao giữa Previous Scores và Performance Index, cho thấy biến này nên được giữ và có thể tạo thêm các biến tương tác hoặc biến phi tuyến khi xây dựng mô hình.



Hình 5: Biểu đồ boxplot giữa Performance Index và Extracurricular Activities

Biểu đồ cho thấy trung vị Performance Index của hai nhóm sinh viên tham gia và không tham gia hoạt động ngoại khóa gần như tương đương ($\tilde{5}5$), phản ánh tác động hạn chế của yếu tố này đến thành tích học tập.

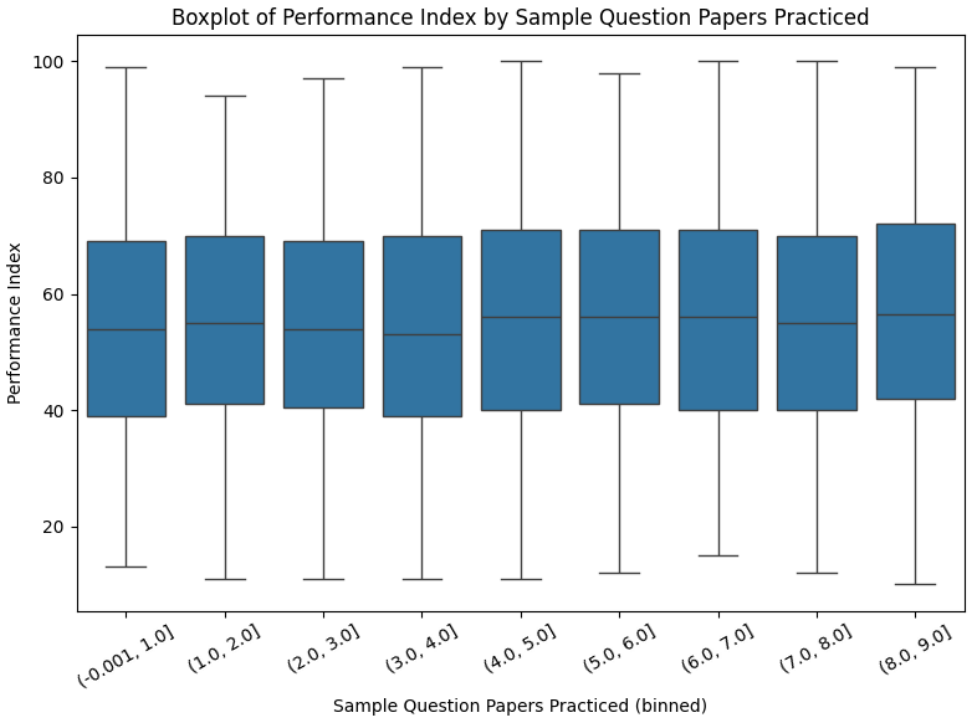
Phân vị của hai nhóm gần trùng nhau, cho thấy phân bố kết quả học tập tương đồng. Độ phân tán rộng cùng sự xuất hiện của các ngoại lệ ở cả hai nhóm cho thấy rằng thành tích chịu ảnh hưởng chủ yếu từ các yếu tố khác, không phải hoạt động ngoại khóa.



Hình 6: Biểu đồ boxplot giữa Performance Index và Sleep Hours

Biểu đồ cho thấy trung vị Performance Index dao động quanh mức 54 – 56 ở mọi mức giờ ngủ, cho thấy giấc ngủ không có mối quan hệ rõ rệt với thành tích học tập.

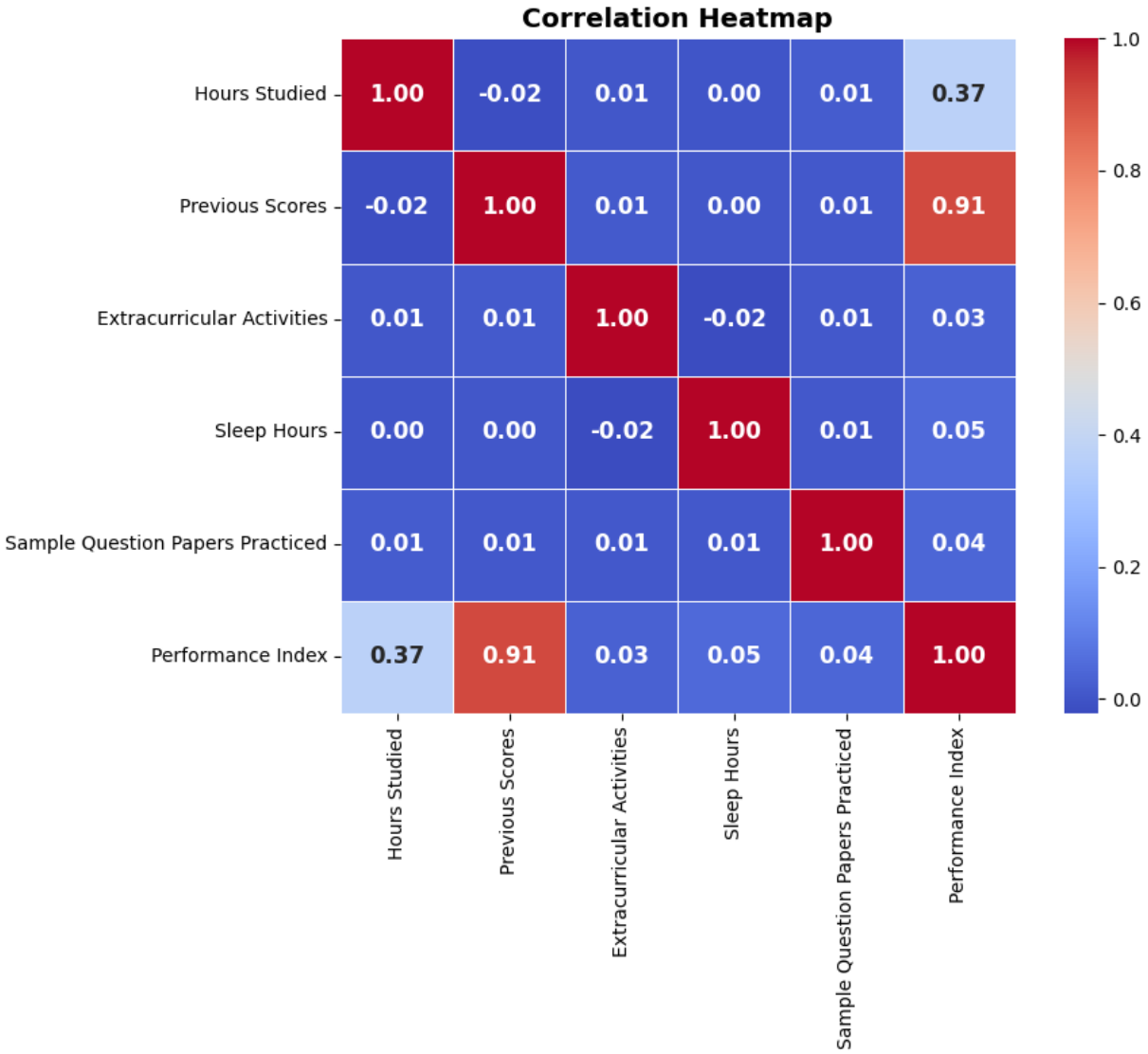
Các phân vị Q1 và Q3 gần như giữ nguyên, phản ánh phân bố kết quả tương đối ổn định giữa các nhóm. Độ phân tán rộng cùng sự xuất hiện của ngoại lệ ở cả hai đầu cho thấy các yếu tố khác ngoài giấc ngủ đóng vai trò quyết định hơn trong hiệu suất học tập.



Hình 7: Biểu đồ boxplot giữa Performance Index và Sample Question Papers Practiced

Biểu đồ cho thấy trung vị Performance Index duy trì quanh mức 54 – 56 ở hầu hết các nhóm số lượng đề thi thử đã làm, cho thấy mối quan hệ yếu giữa biến này và thành tích học tập.

Khoảng phân vị Q1 – Q3 tương đối ổn định, trong khi độ phân tán rộng và sự xuất hiện của ngoại lệ ở cả hai đầu cho thấy yếu tố này chỉ đóng vai trò hỗ trợ, không phải là yếu tố quyết định chính đến Performance Index.

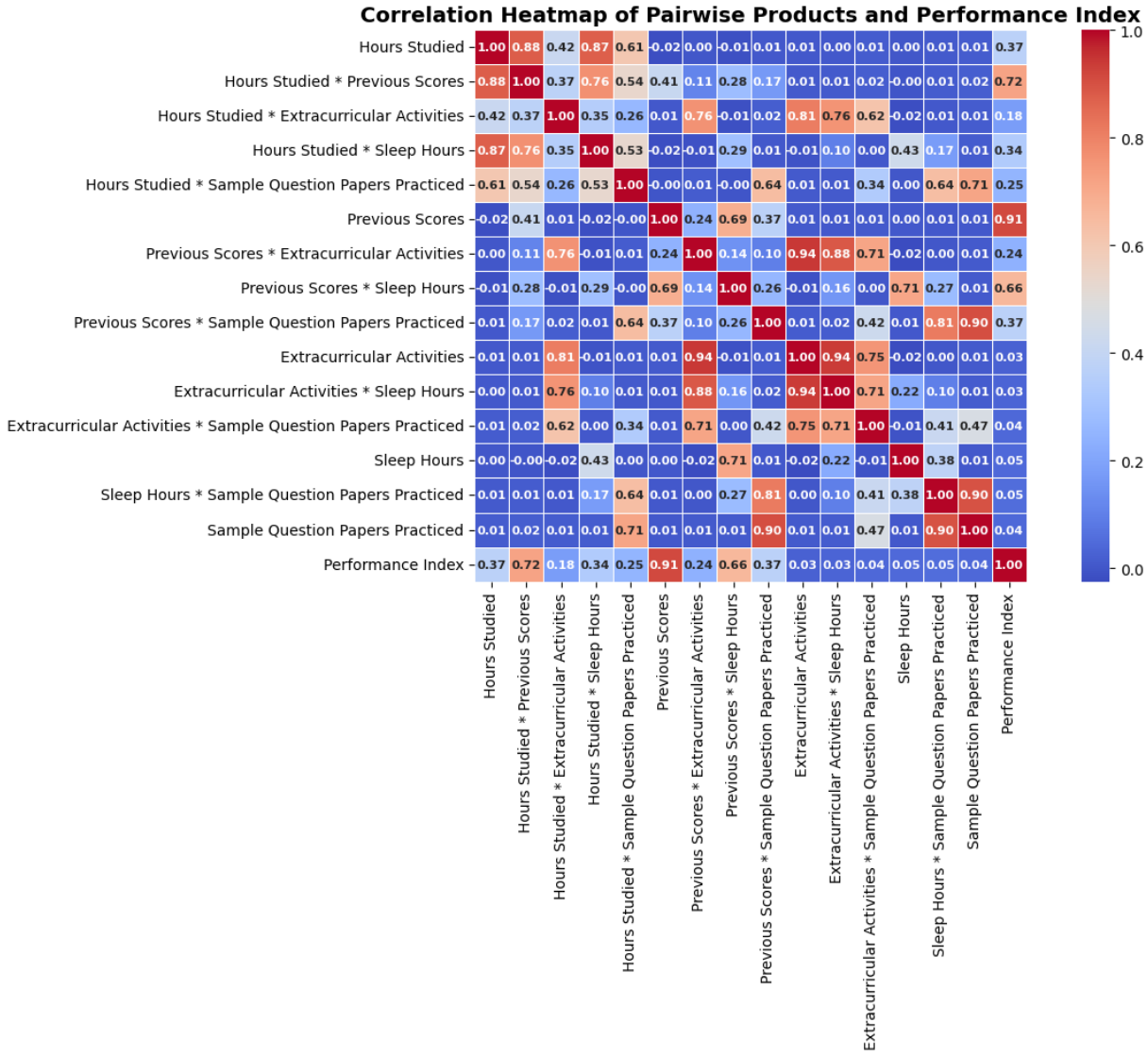


Hình 8: Ma trận tương quan giữa các đặc trưng

Điểm số trước đây là yếu tố có ảnh hưởng lớn nhất đến Performance Index, nhấn mạnh tầm quan trọng của việc duy trì thành tích học tập tốt theo thời gian.

Số giờ học cũng có tác động tích cực đến Performance Index, nhưng mức độ ảnh hưởng không mạnh bằng thành tích học tập trước đó.

Các yếu tố khác như hoạt động ngoại khóa, giấc ngủ và luyện tập đề mẫu hầu như không tạo ra tác động đáng kể đến kết quả học tập. Trong đó, hoạt động ngoại khóa là biến có tương quan thấp nhất với Performance Index.

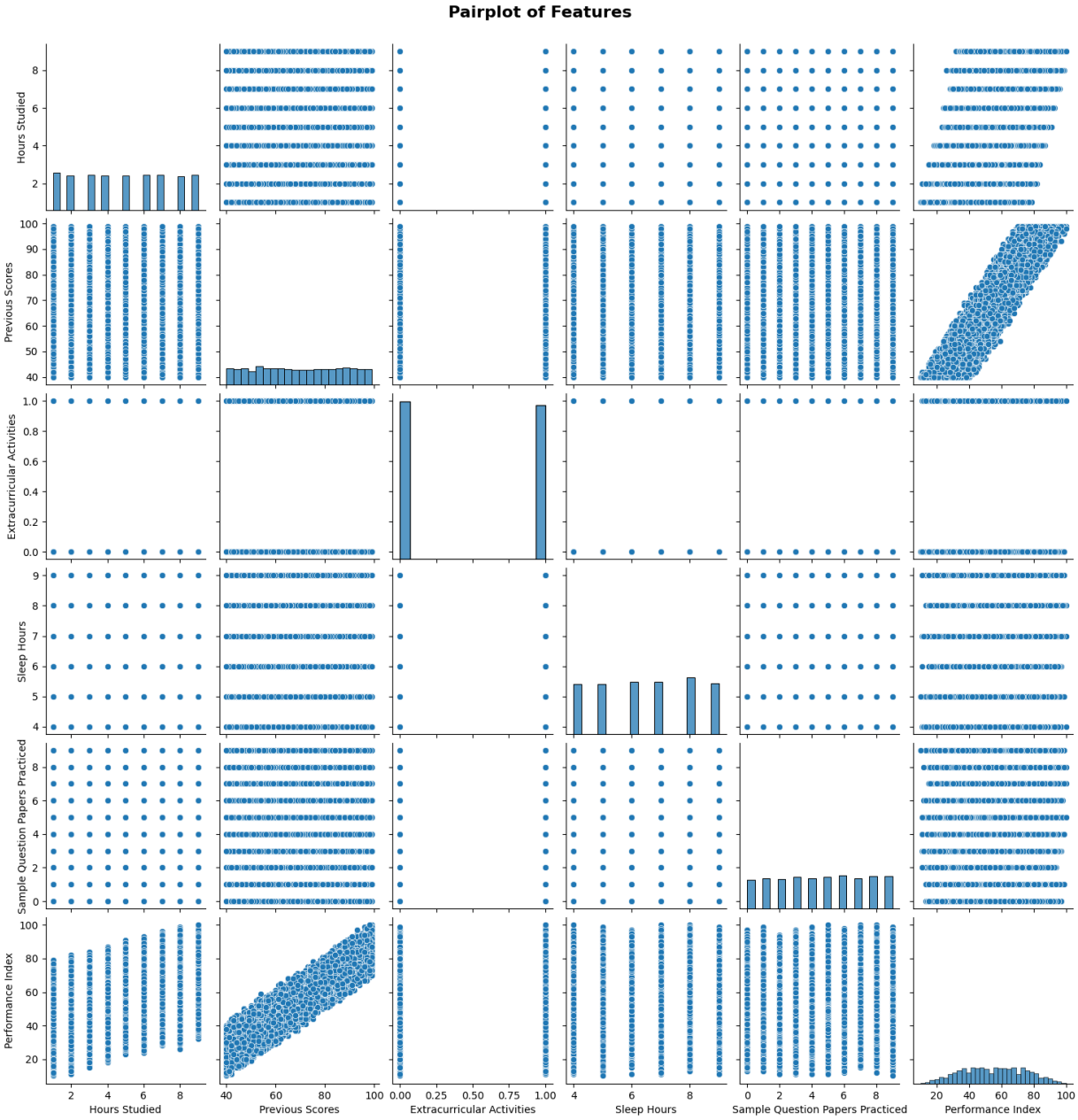


Hình 9: Ma trận tương quan giữa các biến tương tác theo cặp

Ma trận tương quan mở rộng cho thấy ngoài các biến gốc, một số biến tương tác (pairwise products) có mức tương quan cao hơn với Performance Index so với khi xét riêng lẻ:

- Previous Scores vẫn duy trì mức tương quan rất cao với Performance Index (0.91), tiếp tục khẳng định vai trò là yếu tố chính.
- Tương tác Previous Scores × Hours Studied đạt mức tương quan 0.72, cho thấy việc duy trì điểm số tốt kết hợp với học tập chăm chỉ có thể ảnh hưởng đến thành tích hiện tại.
- Hours Studied × Sleep Hours có tương quan 0.34, cao hơn so với xét riêng từng biến, cho thấy rằng sự cân bằng giữa thời gian học và ngủ có thể đóng góp tích cực.

4.2.5 Phân tích nhiều biến



Hình 10: Ma trận biểu đồ phân tán

Mối quan hệ mạnh: Mối quan hệ đáng kể và mạnh duy nhất được quan sát là giữa Previous Scores và Performance Index. Điều này nhất quán trên tất cả các biểu đồ và khẳng định rằng thành tích trong quá khứ là chỉ báo đáng tin cậy nhất cho thành tích trong tương lai trong bộ dữ liệu này.

Mối quan hệ yếu: Các yếu tố như Hours Studied, Extracurricular Activities, Sleep Hours và Sample Question Papers Practiced đều có mối quan hệ yếu với Performance Index. Điều này cho thấy rằng mặc dù chúng có thể ảnh hưởng đến thành tích, nhưng không phải là yếu tố quyết định chính.

Phân phối của các đặc trưng: Phân phối của các biến, đặc biệt là Performance

Index và Previous Scores, cho thấy phạm vi giá trị rộng, phản ánh sự đa dạng trong thành tích học tập và điểm số trước đó của học sinh. Performance Index có xu hướng phân phối gần chuẩn, cho thấy phần lớn học sinh tập trung quanh mức trung bình, với số lượng ít hơn ở mức điểm rất cao hoặc rất thấp.

Nhìn chung, phân tích EDA cho thấy Previous Scores là yếu tố quan trọng nhất ảnh hưởng đến Performance Index, trong khi các yếu tố khác như Hours Studied và Sleep Hours có tác động phụ trợ. Các hoạt động ngoại khóa và luyện tập đề mẫu không có mối liên hệ rõ ràng với thành tích học tập.

4.3 Yêu cầu 2a: Sử dụng toàn bộ 5 đặc trưng để xây dựng mô hình hồi quy tuyến tính

4.3.1 Các hàm hỗ trợ

- **train_linear_regression(X, y):** Hàm này huấn luyện mô hình hồi quy tuyến tính sử dụng công thức giải tích (Normal Equation). Đầu vào là ma trận đặc trưng X và vector mục tiêu y. Các bước thực hiện bao gồm:

  1. Tạo ma trận mở rộng bằng cách thêm một cột các giá trị 1 để xử lý hệ số tự do:

  $$X_{\text{aug}} = [X|\mathbf{1}] \tag{1}$$

  2. Áp dụng công thức Normal Equation để tìm các hệ số tối ưu:

  $$w_{\text{full}} = (X_{\text{aug}}^T X_{\text{aug}})^{-1} X_{\text{aug}}^T y \tag{2}$$

  3. Tách vector kết quả thành trọng số w cho các đặc trưng và hệ số tự do b:

  $$w = w_{\text{full}}[:-1], \quad b = w_{\text{full}}[-1] \tag{3}$$

- **mean_squared_error(y_true, y_pred):** Hàm tính sai số bình phương trung bình (MSE) giữa giá trị thực tế và dự đoán. Cài đặt bằng numpy thông qua lệnh `np.mean((y_true - y_pred)**2)`. MSE luôn không âm và càng nhỏ thì mô hình càng chính xác.

4.3.2 Huấn luyện mô hình

Sau khi cài đặt các hàm cần thiết, tiến hành huấn luyện mô hình hồi quy tuyến tính sử dụng toàn bộ 5 đặc trưng:

1. Chuyển đổi dữ liệu X_train và y_train thành mảng numpy để thực hiện các phép toán hiệu quả
2. Huấn luyện mô hình bằng hàm `train_linear_regression` để tìm vector trọng số w và hệ số tự do b
3. Xây dựng công thức hồi quy tuyến tính dựa trên các trọng số tìm được
4. Thực hiện dự đoán trên tập kiểm tra và đánh giá hiệu suất bằng MSE

4.4 Yêu cầu 2b: Xây dựng mô hình sử dụng duy nhất 1 đặc trưng, tìm mô hình cho kết quả tốt nhất

4.4.1 Các hàm hỗ trợ

- **manual_shuffle(X_df, y_series):** Hàm này trộn ngẫu nhiên các hàng của DataFrame đặc trưng và Series mục tiêu theo cùng một thứ tự để đảm bảo tính tương thích:
  1. Tạo mảng chỉ số từ 0 đến số lượng mẫu: `idx = np.arange(len(X_df))`
  2. Xáo trộn ngẫu nhiên chỉ số bằng numpy: `np.random.shuffle(idx)`
  3. Áp dụng chỉ số đã xáo trộn để tái sắp xếp dữ liệu và đặt lại chỉ mục
- **cross_validate_features(X, y, feature_names, k=5):** Thực hiện k-fold cross-validation cho từng đặc trưng riêng biệt:
  1. Chia dữ liệu thành k phần bằng nhau
  2. Với mỗi đặc trưng, lần lượt sử dụng $k-1$ phần làm tập huấn luyện và 1 phần làm tập kiểm định
  3. Huấn luyện mô hình hồi quy tuyến tính một chiều và tính MSE trên tập kiểm định
  4. Tính trung bình MSE qua k lần lặp cho mỗi đặc trưng
- **display_cv_table(results):** Hiển thị kết quả cross-validation dưới dạng bảng tóm tắt với các cột: STT, tên đặc trưng và MSE trung bình.
- **display_fold_detail_table(results, k=5):** Hiển thị bảng chi tiết MSE cho từng fold và từng đặc trưng, với hàng là tên đặc trưng, cột là số fold và MSE trung bình.

4.4.2 Đánh giá và lựa chọn đặc trưng tối ưu

Để tìm đặc trưng đơn lẻ tốt nhất, chúng tôi tiến hành quá trình đánh giá và lựa chọn thông qua 3 bước chính:

1. **Đánh giá đặc trưng bằng k-fold cross-validation:**
   - Trộn ngẫu nhiên dữ liệu huấn luyện bằng hàm `manual_shuffle`
   - Thực hiện cross-validation với $k = 5$ cho từng đặc trưng
   - Hiển thị bảng tổng hợp MSE và chi tiết từng fold để so sánh hiệu năng
   - Xác định đặc trưng có MSE trung bình thấp nhất (`Previous Scores`)
2. **Huấn luyện mô hình tốt nhất:**
   - Chọn đặc trưng tốt nhất từ tập huấn luyện bằng MSE
   - Huấn luyện mô hình hồi quy tuyến tính (mô hình của đặc trưng tốt nhất) trên toàn bộ tập huấn luyện

- Xây dựng công thức hồi quy tuyến tính dạng $y = wx + b$ với đặc trưng đã chọn

3. **Đánh giá trên tập kiểm tra:**

   - Áp dụng mô hình đã huấn luyện để dự đoán kết quả trên tập kiểm tra
   - Tính MSE để đánh giá hiệu năng cuối cùng

4.5 Yêu cầu 2c: Sinh viên tự xây dựng/thiết kế mô hình, tìm mô hình cho kết quả tốt nhất

4.5.1 Quy ước ký hiệu

Các biến sau đây sẽ được sử dụng để thể hiện các đặc trưng trong bộ dữ liệu:

Ký hiệu	Mô tả đặc trưng
F_1	Hours Studied (Số giờ học tập)
F_2	Previous Scores (Điểm số trước đó)
F_3	Extracurricular Activities (Hoạt động ngoại khóa)
F_4	Sleep Hours (Số giờ ngủ)
F_5	Sample Question Papers Practiced (Số lượng đề thi đã làm)

Bảng 7: Quy ước ký hiệu cho các đặc trưng trong mô hình

4.5.2 Thiết kế và lựa chọn mô hình

Dựa trên kết quả phân tích khám phá dữ liệu (EDA) từ yêu cầu 1, tôi đã thiết kế ba mô hình hồi quy tuyến tính với các lý do sau:

Mô hình 1: Sử dụng 2 đặc trưng quan trọng nhất (F_1 và F_2)

$$\text{Student Performance} = \alpha_0 + \alpha_1 F_1 + \alpha_2 F_2$$

- Từ ma trận tương quan (Hình 8), **Previous Scores** (F_2) có hệ số tương quan cao nhất với Performance Index (0.91), trong khi **Hours Studied** (F_1) có hệ số tương quan đứng thứ hai (0.37).
- Biểu đồ boxplot (Hình 3 và 4) cho thấy trung vị của Performance Index tăng rõ rệt theo cả hai đặc trưng này, đặc biệt là Previous Scores với mức tăng từ khoảng 25 (Previous Scores $\tilde{4}0$) lên khoảng 85 (Previous Scores $\tilde{9}5$).
- Mô hình này áp dụng nguyên tắc tiết kiệm (parsimony) bằng cách sử dụng số lượng tối thiểu các đặc trưng có tác động mạnh nhất, tránh overfitting.[3]

Mô hình 2: Loại bỏ F_3, thêm bình phương của F_2

$$\text{Student Performance} = \alpha_0 + \alpha_1 F_1 + \alpha_2 F_2 + \alpha_3 F_4 + \alpha_4 F_5 + \alpha_5 F_2^2$$

- Ma trận tương quan (Hình 8) cho thấy **Extracurricular Activities** (F_3) có hệ số tương quan thấp nhất (0.03) với Performance Index.

- Biểu đồ boxplot (Hình 5) xác nhận rằng trung vị Performance Index gần như giống nhau ($\tilde{5}5$) giữa nhóm tham gia và không tham gia hoạt động ngoại khóa nên loại bỏ.
- Biểu đồ phân tán (Hình 10) có thể thấy mối quan hệ tuyến tính giữa Previous Scores và Performance Index. Ngoài ra, thêm biến bình phương của F_2 (F_2^2) vào mô hình giúp kiểm tra mối quan hệ phi tuyến trong mô hình. Việc này giúp mô hình nắm bắt tốt hơn mối quan hệ phi tuyến nếu có tồn tại.
- Sleep Hours (F_4) và Sample Question Papers Practiced (F_5) được giữ lại vì chúng có thể đóng góp nhỏ nhưng có ý nghĩa khi kết hợp với các đặc trưng khác.

Mô hình 3: Sử dụng đặc trưng tương tác ($F_1 \times F_2$)

$$\text{Student Performance} = \alpha_0 + \alpha_1(F_1 \times F_2)$$

- Ma trận tương quan mở rộng (Hình 9) cho thấy tương tác **Hours Studied** $\times$ **Previous Scores** có hệ số tương quan cao (0.72) với Performance Index, cao hơn đáng kể so với Hours Studied đơn thuần (0.37).
- Điều này cho thấy rằng tác động của thời gian học có thể phụ thuộc vào điểm số trước đó: học sinh có nền tảng tốt có thể tận dụng hiệu quả hơn thời gian học.
- Mô hình tương tác này đơn giản với chỉ một biến dự đoán, nhưng vẫn có khả năng nắm bắt hiệu ứng kết hợp của hai đặc trưng quan trọng nhất.

Ba mô hình này thể hiện các phương pháp tiếp cận khác nhau dựa trên các hiểu biết từ EDA: từ mô hình đơn giản với các đặc trưng quan trọng nhất (Mô hình 1), mô hình mở rộng với biến phi tuyến (Mô hình 2), đến mô hình tập trung vào hiệu ứng tương tác (Mô hình 3). Việc so sánh hiệu suất của ba mô hình này sẽ giúp xác định phương pháp tiếp cận hiệu quả nhất cho bài toán dự đoán Performance Index.

4.5.3 Các hàm hỗ trợ

- **create_model1, create_model2, create_model3:** Ba hàm này tạo ra các DataFrame đặc trưng phù hợp với mỗi mô hình đề xuất:
  - `create_model1`: Chọn hai đặc trưng quan trọng nhất là Hours Studied và Previous Scores.
  - `create_model2`: Thêm đặc trưng bình phương của Previous Scores và loại bỏ Extracurricular Activities.
  - `create_model3`: Tạo đặc trưng tương tác giữa Hours Studied và Previous Scores.
- **define_models:** Định nghĩa danh sách các mô hình cần đánh giá với tên, hàm tạo đặc trưng và mô tả.
- **cross_validate_custom_model:** Thực hiện k-fold cross-validation cho một mô hình tùy chỉnh:
  - Chia dữ liệu thành k fold (mặc định $k = 5$)
  - Với mỗi fold, tạo tập validation và tập training

  - Áp dụng hàm tạo đặc trưng cho cả hai tập dữ liệu
  - Huấn luyện mô hình và tính MSE trên tập validation
  - Tính trung bình MSE qua tất cả các fold
- **evaluate_models:** Đánh giá tất cả mô hình trên cùng một dữ liệu đã xáo trộn:
  - Trộn dữ liệu một lần duy nhất bằng `manual_shuffle`
  - Đánh giá từng mô hình bằng `cross_validate_custom_model`
  - Xác định mô hình tốt nhất dựa trên MSE trung bình thấp nhất
- **display_model_results:** Hiển thị kết quả đánh giá các mô hình:
  - In thông báo về mô hình tốt nhất và MSE trung bình
  - Tạo bảng tổng kết với STT, tên mô hình và MSE
  - Hiển thị bảng chi tiết MSE từng fold cho từng mô hình
- **train_best_model:** Huấn luyện mô hình tốt nhất trên toàn bộ tập huấn luyện:
  - Lấy tên và hàm tạo đặc trưng của mô hình tốt nhất
  - Áp dụng biến đổi đặc trưng cho tập huấn luyện và kiểm tra
  - Huấn luyện mô hình và trả về thông tin chi tiết
- **print_regression_formula:** In công thức hồi quy với các trọng số đã làm tròn.
- **evaluate_on_test_set:** Đánh giá mô hình trên tập kiểm tra và in MSE.

5 Kết quả

Các biến sau đây sẽ được sử dụng để thể hiện các đặc trưng trong bộ dữ liệu:

Ký hiệu	Mô tả đặc trưng
F_1	Hours Studied (Số giờ học tập)
F_2	Previous Scores (Điểm số trước đó)
F_3	Extracurricular Activities (Hoạt động ngoại khóa)
F_4	Sleep Hours (Số giờ ngủ)
F_5	Sample Question Papers Practiced (Số lượng đề thi đã làm)

Bảng 8: Quy ước ký hiệu cho các đặc trưng trong mô hình

5.1 Yêu cầu 1: Phân tích khám phá dữ liệu

Đã thực hiện phân tích khám phá dữ liệu theo các bước đã nêu trong phần ý tưởng thực hiện. Các kết quả thu được bao gồm:

- Đã khám phá tổng quan, kiểm tra kiểu dữ liệu, giá trị duy nhất, giá trị thiếu của các đặc trưng trong bộ dữ liệu.
- Đã phân tích phân phối của các thuộc tính đầu vào và đầu ra bằng biểu đồ histogram, boxplot.
- Đã xác định mối tương quan giữa các thuộc tính bằng ma trận hệ số tương quan và biểu đồ heatmap.

5.2 Yêu cầu 2a: Mô hình với 5 đặc trưng

Đã xây dựng mô hình hồi quy tuyến tính với 5 đặc trưng. Công thức mô hình được thể hiện như sau:

$$\text{Student Performance} = (2.852)*F_1+(1.018)*F_2+(0.606)*F_3+(0.473)*F_4+(0.192)*F_5+(-33.961)$$

Giá trị MSE trên tập kiểm tra là 4.0928, cho thấy mô hình có hiệu suất tốt.

5.3 Yêu cầu 2b: Mô hình với 1 đặc trưng

Để xác định đặc trưng đơn lẻ tối ưu, đã thực hiện đánh giá từng đặc trưng bằng k-fold cross-validation với $k = 5$. Kết quả như sau:

Bảng 9: MSE từng fold cho từng đặc trưng

Đặc trưng	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Avg MSE**
Hours Studied	318.7034	333.1837	315.7927	308.1350	311.8824	317.5394
Previous Scores	62.4644	59.3010	59.7150	59.8759	59.3884	60.1490
Extracurricular Activities	373.5794	380.8945	375.7137	350.6797	357.4178	367.6570
Sleep Hours	374.2224	380.3985	374.2243	348.6933	358.1061	367.1289
Sample Question Papers Practiced	373.4499	380.4286	374.6270	349.2898	358.7206	367.3032

Bảng 10: MSE trung bình của các mô hình 1 đặc trưng

STT	**Mô hình với 1 đặc trưng**	**MSE**
1	Hours Studied	317.5394
2	Previous Scores	60.1490
3	Extracurricular Activities	367.6570
4	Sleep Hours	367.1289
5	Sample Question Papers Practiced	367.3032

Đặc trưng **Previous Scores** đạt MSE thấp nhất (60.1490), tốt hơn đáng kể so với các đặc trưng khác. Sau khi huấn luyện mô hình với đặc trưng này trên toàn bộ tập huấn luyện, thu được công thức hồi quy:

$$\text{Student Performance} = (1.011) * F_2 + (-15.015)$$

MSE trên tập kiểm tra với mô hình đặc trưng tốt nhất là 58.8835.

5.3.1 Yêu cầu 2c: Sinh viên tự xây dựng/thiết kế mô hình, tìm mô hình cho kết quả tốt nhất

Trong phần này, 3 mô hình đã được xây dựng dựa trên các đặc trưng đã phân tích ở Yêu cầu 1. Kết quả MSE của từng mô hình như sau:

Bảng 11: MSE từng fold cho từng mô hình

Mô hình	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Avg MSE**
F1 + F2	5.3100	5.1861	4.9536	5.2806	5.3860	5.2233
F1 + F2 + F4 + F5 + F2^2	4.3016	4.2324	4.0961	4.3237	4.4268	4.2761
F1*F2	184.7292	183.6679	169.1893	178.7960	179.7410	179.2247

Bảng 12: MSE trung bình của các mô hình tự thiết kế

STT	**Mô hình**	**MSE**
1	F1 + F2	5.2233
2	F1 + F2 + F4 + F5 + F2^2	4.2761
3	F1*F2	179.2247

Mô hình tốt nhất là **F1 + F2 + F4 + F5 + F2^2** với MSE trung bình thấp nhất (4.2761). Đây là mô hình mở rộng sử dụng hầu hết các đặc trưng (ngoại trừ F3 - Extracurricular Activities) và thêm thành phần phi tuyến F2^2 (bình phương của Previous Scores).

Sau khi huấn luyện lại trên toàn bộ tập huấn luyện, thu được công thức hồi quy:

$$\text{Student Performance} = (2.852)*F_1+(1.029)*F_2+(0.470)*F_4+(0.193)*F_5+(-0.000)*F_2^2+(-34.003)$$

MSE trên tập kiểm tra với mô hình tự thiết kế tốt nhất là 4.2076, cải thiện đáng kể so với mô hình sử dụng một đặc trưng (60.1490) và gần với hiệu suất của mô hình đầy đủ 5 đặc trưng (4.0928).

Kết quả này cho thấy việc loại bỏ đặc trưng không liên quan (Extracurricular Activities) đã giúp tối ưu hiệu suất mô hình hồi quy tuyến tính, với MSE chỉ tăng nhẹ từ 4.0928 lên 4.2076. Sự hy sinh nhỏ về độ chính xác này đổi lại mang đến một mô hình đơn giản hơn, dễ huấn luyện và dự đoán nhanh hơn, đặc biệt hữu ích khi xử lý dữ liệu lớn.

Đáng chú ý, hệ số của thành phần phi tuyến (Previous Scores^2) xấp xỉ 0, cho thấy mối quan hệ giữa Previous Scores và Performance Index là tuyến tính, không tồn tại mối quan hệ phi tuyến đáng kể. Điều này cũng được xác nhận bởi mô hình tương tác đơn thuần (F1*F2) có hiệu suất kém (MSE=179.2247), cho thấy mối quan hệ giữa các đặc trưng và biến mục tiêu phần lớn là tuyến tính.

5.4 Kết luận

Từ kết quả thu được của các mô hình đã xây dựng và đánh giá, có thể rút ra một số kết luận quan trọng:

- **Tính phù hợp của hồi quy tuyến tính:** MSE thấp của các mô hình cho thấy hồi quy tuyến tính là phương pháp phù hợp cho bài toán dự đoán Performance Index. Điều này được khẳng định thêm khi hệ số của thành phần phi tuyến (Previous Scores^2) xấp xỉ 0.

- **Tầm quan trọng của việc lựa chọn đặc trưng:** Khả năng dự đoán của mô hình phụ thuộc mạnh vào việc lựa chọn đúng đặc trưng quan trọng. Previous Scores có ảnh hưởng lớn nhất (MSE=60.1043 khi dùng đơn lẻ), tiếp theo là Hours Studied.

- **Giảm thiểu đặc trưng không liên quan:** Việc loại bỏ Extracurricular Activities (F3) chỉ làm giảm nhẹ độ chính xác của mô hình, nhưng lại tăng tính hiệu quả của việc giảm chiều dữ liệu.

- **Mối quan hệ tuyến tính giữa các biến:** Hệ số gần bằng 0 của thành phần phi tuyến chứng tỏ mối quan hệ giữa các đặc trưng và biến mục tiêu phần lớn là tuyến tính. Mô hình tương tác đơn thuần (F1*F2) có hiệu suất kém (MSE=179.5046) cũng xác nhận điều này.

- **Hiệu quả của cross-validation:** Phương pháp k-fold cross-validation cho thấy độ ổn định của các mô hình qua các fold khác nhau, đặc biệt là mô hình F1 + F2 + F4 + F5 + F2^2 có độ dao động MSE thấp (từ 3.8981 đến 4.4493).

- **Đánh đổi giữa độ chính xác và hiệu suất:** Mặc dù loại bỏ một số đặc trưng có thể làm tăng nhẹ MSE (giảm độ chính xác), nhưng mô hình đơn giản hơn sẽ có thời gian huấn luyện và dự đoán nhanh hơn, đặc biệt quan trọng khi làm việc với bộ dữ liệu lớn. Ví dụ, mô hình chỉ sử dụng F1 + F2 có MSE = 5.2233, cao hơn một chút so với mô hình đầy đủ nhưng đơn giản hơn nhiều và vẫn cho kết quả dự đoán chấp nhận được.

Kết quả nghiên cứu cho thấy mô hình hồi quy tuyến tính với bốn đặc trưng chính (Hours Studied, Previous Scores, Sleep Hours, và Sample Question Papers Practiced) đạt hiệu suất tối ưu trong việc dự đoán Performance Index của học sinh. Mô hình này vừa đơn giản, dễ giải thích, vừa có độ chính xác cao, phù hợp cho các ứng dụng thực tiễn trong lĩnh vực giáo dục.

Trong trường hợp yêu cầu tốc độ xử lý cao với dữ liệu lớn, có thể cân nhắc sử dụng mô hình chỉ với hai đặc trưng F1 + F2, chấp nhận giảm nhẹ độ chính xác để đổi lấy hiệu suất tính toán tốt hơn. Sự đánh đổi này là quan trọng trong các ứng dụng thực tế khi thời gian phản hồi là yếu tố then chốt.

Nghiên cứu này cũng khẳng định tầm quan trọng của phân tích khám phá dữ liệu (EDA) trước khi xây dựng mô hình, giúp hiểu rõ đặc tính của dữ liệu và đưa ra quyết định sáng suốt về việc lựa chọn đặc trưng và thiết kế mô hình phù hợp.

Tài liệu

[1] Tiep Vu Huu, *Mục đích của EDA*, 2021, https://machinelearningcoban.com/tabml_book/ch_data_processing/eda_purpose.html

[2] geeksforgeeks, *What is Exploratory Data Analysis?*, Aug, 06, 2021, https://www.geeksforgeeks.org/data-analysis/what-is-exploratory-data-analysis/

[3] Avadhoot Tavhare, *Occam's Razor in Machine Learning: The Principle of Parsimony*, May 21, 2024, https://medium.com/%40qjbqvwzmg/occams-razor-in-machine-learning-the-principle-of-