ĐẠI HỌC QUỐC GIA TPHCM

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

Toán ứng dụng và thống kê cho Công nghệ thông tin

Đồ án: Linear Regression

Sinh viên:

Đinh Nguyễn Gia Bảo
(22127027)

Giảng viên:

GV. Phan Thị Phương Uyên

GV. Vũ Quốc Hoàng

GV. Nguyễn Văn Quang Huy

GV. Nguyễn Ngọc Toàn

Ngày 17 tháng 8 năm 2024

MỤC LỤC

I/ Tổng quan

Mục tiêu của đồ án là tìm hiểu các yếu tố ảnh hưởng đến thành tích học tập của sinh viên (Academic Student Performance Index). Các yếu tố ảnh hưởng có thể là số giờ học tập/nghiên cứu, hoạt động ngoại khóa, số giờ ngủ, số bài kiểm tra mẫu đã luyện tập…

Thông tin sinh viên:

- Họ và tên: Đinh Nguyễn Gia Bảo
- MSSV: 22127027

Bảng đánh giá

STT	Yêu cầu	Đánh giá	Ghi chú
1.1	Sử dụng thống kê để phân tích dữ các đặc trưng.	100%	
1.2	Sử dụng biểu đồ (bar, box, heatmap, scatter, line...) để phân tích/quan sát các đặc trưng.	100%	
2	Xây dựng mô hình hồi quy tuyến tính	100%	
2a	Sử dụng toàn bộ 5 đặc trưng	100%	
2b	sử dụng duy nhất 1 đặc trưng, tìm mô hình cho kết quả tốt nhất	100%	
2c	Xây dựng mô hình sử dụng duy nhất 1 đặc trưng, tìm mô hình cho kết quả tốt nhất	100%	

3	Báo cáo	100%	

Ngôn ngữ lập trình và môi trường lập trình:

- Python (Jupyter Notebook)
- Visual Studio Code (VS Code)

Thư viện sử dụng:

- **Numpy (imported as np):** Được sử dụng để làm việc với các mảng và ma trận, rất quan trọng cho các thao tác xử lý hình ảnh.
- **matplotlib.pyplot (imported as plt):** Được sử dụng để trực quan hóa và hiển thị các hình ảnh đã được xử lý trong một bố cục lưới.
- **pandas (imported as pd):** Được sử dụng để xử lý và phân tích dữ liệu, bao gồm đọc và ghi dữ liệu từ các định dạng khác nhau như CSV, Excel, và SQL, cùng với các thao tác xử lý dữ liệu bảng như lọc, nhóm và tổng hợp dữ liệu.
- **seaborn (imported as sns):** Được sử dụng để trực quan hóa dữ liệu một cách trực quan và dễ hiểu, hỗ trợ nhiều loại biểu đồ như phân tán, hộp, heatmap, và giúp làm nổi bật các mối quan hệ trong dữ liệu với giao diện dễ sử dụng.
- **sklearn.linear_model.LinearRegression:** Được sử dụng để xây dựng mô hình hồi quy tuyến tính, cho phép dự đoán giá trị dựa trên các đặc trưng đầu vào và đánh giá hiệu suất của mô hình hồi quy tuyến tính.
- **sklearn.metrics.mean_absolute_error:** Được sử dụng để đánh giá hiệu suất của các mô hình hồi quy (sai số tuyệt đối trung bình MAE).
- **sklearn.model_selection.cross_val_score và sklearn.model_selection.KFold:** Được sử dụng để thực hiện k-fold cross-validation, giúp đánh giá hiệu suất của mô hình bằng cách chia dữ liệu thành nhiều tập con để huấn luyện và kiểm tra mô hình.
- **sklearn.preprocessing và StandardScaler**: Được sử dụng để chuẩn hóa dữ liệu, điều chỉnh dữ liệu sao cho có trung bình bằng 0 và độ lệch chuẩn bằng 1, giúp cải thiện hiệu suất của các mô hình học máy.

II/ Phân tích dữ liệu

Ý tưởng

Bước 1: Đọc Dữ Liệu:

- Đọc dữ liệu từ file CSV.
- Hiển thị những dòng đầu tiên và thông tin tổng quan.
- Thực hiện thống kê mô tả.

Bước 2: Kiểm Tra Giá Trị Thiếu:

- Đếm số lượng giá trị thiếu trong từng cột.
- Phân Tích Mối Quan Hệ:

Bước 3: Tính toán ma trận tương quan giữa các biến.

Bước 4: Trực quan dữ liệu bằng cách vẽ

- Biểu đồ phân phối của biến Performance Index.
- Biểu đồ phân tán giữa Performance Index và các yếu tố khác.
- Biểu đồ cột cho từng đặc trưng, bao gồm phân loại Previous Scores thành nhóm.
- Heatmap ma trận tương quan.

Tổng quan dữ liệu

Số lượng entries: 9000
Số cột: 6

- Hours Studied (Số giờ học)
- Previous Scores (Điểm số trước đó)
- Extracurricular Activities (Hoạt động ngoại khóa)
- Sleep Hours (Số giờ ngủ)
- Sample Question Papers Practiced (Số bài tập mẫu đã thực hành)
- Performance Index (Chỉ số hiệu suất)

Thống kê mô tả

Hours Studied

- Trung bình: 4.98 giờ
- Khoảng giá trị: từ 1 đến 9 giờ

Previous Scores:

- Trung bình: 69.40 điểm
- Khoảng giá trị: từ 40 đến 99 điểm

Extracurricular Activities:

- Trung bình: 0.49 (biến nhị phân, 0 hoặc 1)
- Khoảng giá trị: 0 hoặc 1

Sleep Hours:

- Trung bình: 6.54 giờ
- Khoảng giá trị: từ 4 đến 9 giờ

Sample Question Papers Practiced:

- Trung bình: 4.59 bài
- Khoảng giá trị: từ 0 đến 9 bài

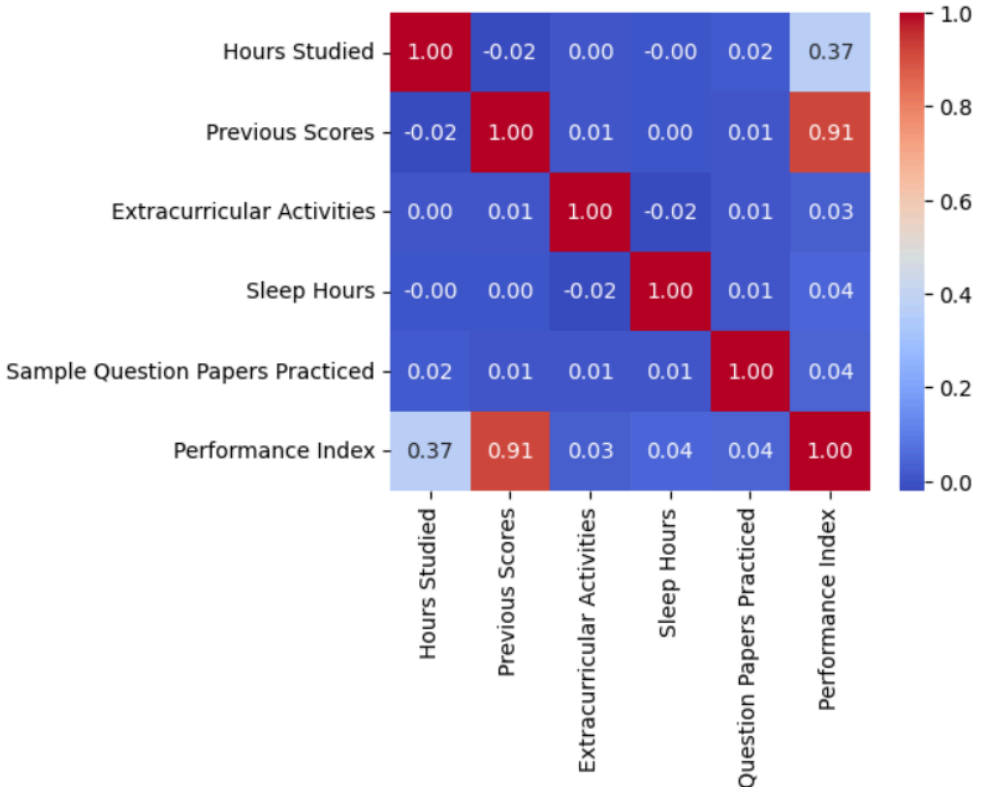
Performance Index:

- Trung bình: 55.14
- Khoảng giá trị: từ 10 đến 100

(Không có bất cứ giá trị thiếu nào trong các cột)

Phân Tích Mối Quan Hệ Giữa Các Biến

Sơ đồ heat map (Ma trận tương quan)



Nh

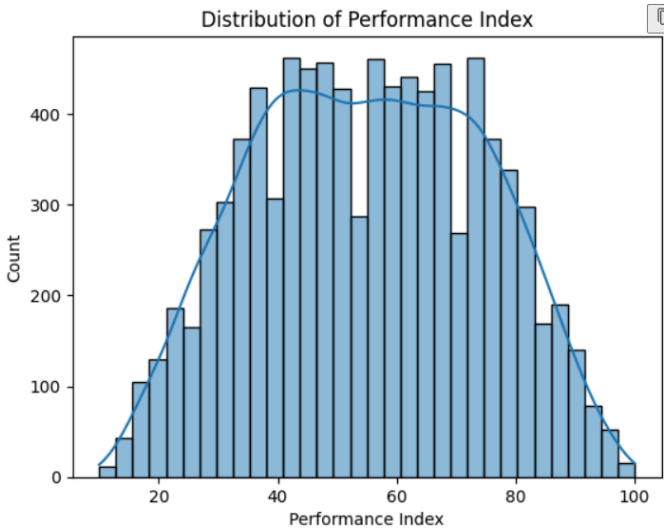
Nhận xét

Performance Index:

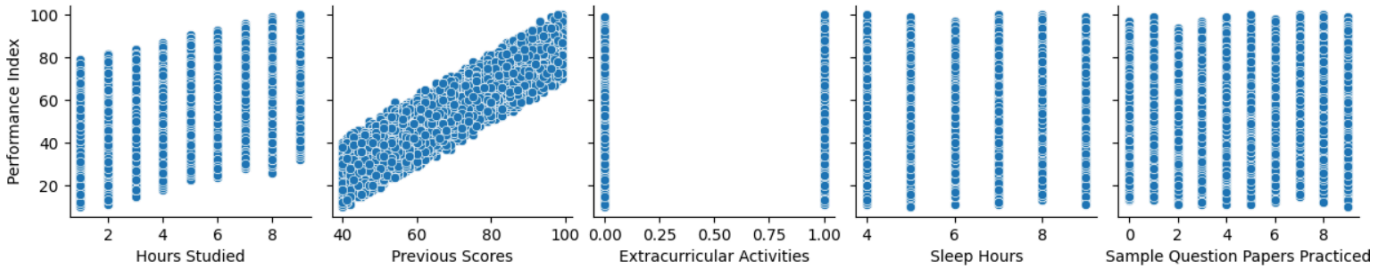
- Có mối tương quan mạnh với Previous Scores (0.915).
- Có mối tương quan nhẹ với Hours Studied (0.369)
- Tương quan với các yếu tố khác như Extracurricular Activities, Sleep Hours, và Sample Question Papers Practiced là rất thấp (dưới 0.1).

Một số biểu đồ khác:

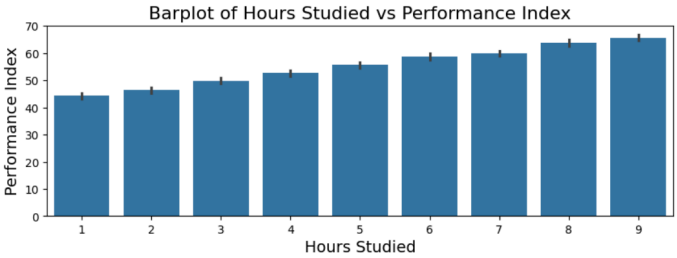
Biểu đồ phân phối dữ liệu

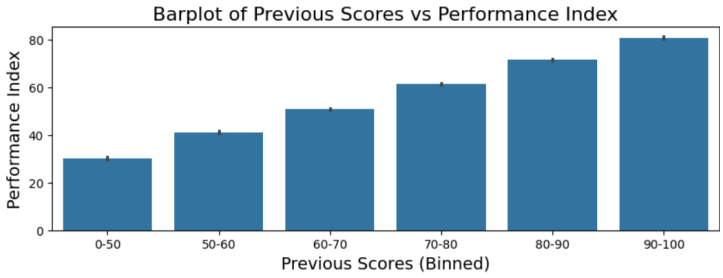


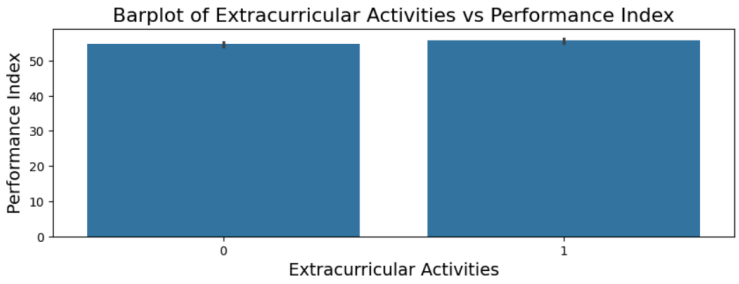
Biểu đồ phân tán dữ liệu

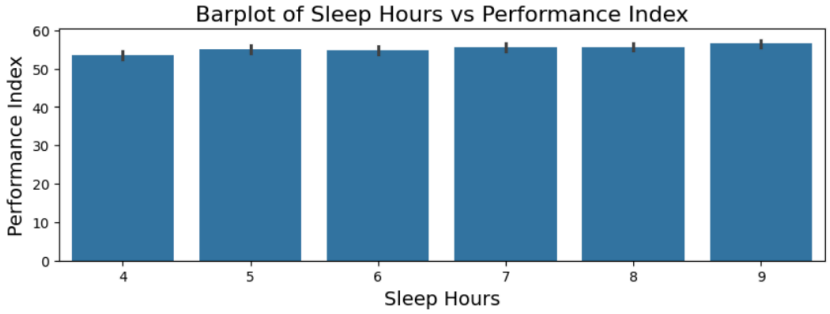


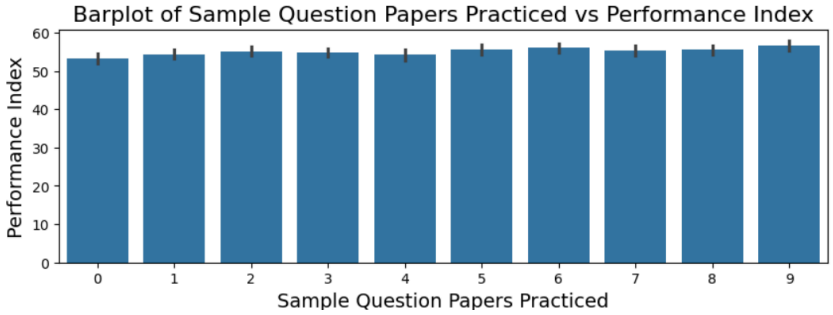
Biểu đồ cột











Nhận xét:

- Previous Scores có mối tương quan rất mạnh với Performance Index, cho thấy đây là yếu tố quan trọng nhất trong việc dự đoán chỉ số hiệu suất.
- Các yếu tố khác như Hours Studied và Sleep Hours có mối tương quan nhẹ với Performance Index.
- Extracurricular Activities và Sample Question Papers Practiced không có mối tương quan đáng kể với Performance Index, cho thấy chúng có thể ít ảnh hưởng đến hiệu suất hơn so với các yếu tố khác.

Kết luận: Previous Scores là yếu tố quan trọng nhất để tập trung vào khi xây dựng mô hình dự đoán chỉ số hiệu suất.

III/ Xây dựng mô hình hồi quy tuyến tính

Sử dụng toàn bộ đặc trưng

Ý tưởng

Xây dựng và đánh giá một mô hình hồi quy tuyến tính trên dữ liệu huấn luyện **(train.csv)**, sau đó sử dụng mô hình này để dự đoán và đánh giá trên tập kiểm tra **(test.csv)**

Công việc cụ thể:

- Tạo và huấn luyện mô hình hồi quy tuyến tính.
- Dự đoán giá trị trên tập kiểm tra.
- Trực quan hóa kết quả dự đoán so với giá trị thực tế.
- Đánh giá hiệu suất mô hình bằng cách tính toán sai số tuyệt đối trung bình (MAE).
- In ra công thức hồi quy của mô hình.

Các bước thực hiện

- **Tạo Mô Hình Hồi Quy Tuyến Tính:** Khởi tạo đối tượng LinearRegression từ thư viện sklearn.

- **Huấn Luyện Mô Hình:** Sử dụng phương thức fit của mô hình để huấn luyện với dữ liệu huấn luyện (X_train, y_train).
- **Dự Đoán Trên Tập Kiểm Tra:** Dự đoán giá trị trên tập kiểm tra (X_test) bằng cách sử dụng phương thức predict.
- **Trực Quan Hóa Kết Quả:**
    - **Biểu Đồ Phân Tán:** Vẽ biểu đồ phân tán của giá trị thực tế so với giá trị dự đoán bằng cách sử dụng hàm plot_true_vs_predictions.
    - **Đường Hồi Quy:** Vẽ đường hồi quy trên cùng biểu đồ phân tán bằng cách sử dụng hàm plot_regression_line.
- **Đánh Giá Mô Hình:** Tính toán và in ra sai số tuyệt đối trung bình (MAE) bằng cách sử dụng hàm printMAE.
- **In Ra Công Thức Hồi Quy:** Tạo công thức hồi quy từ hệ số và hệ số chệch của mô hình bằng cách sử dụng hàm create_regression_formula.

Biểu đồ phân tán

Vẽ đường hồi quy

Đánh giá và nhận xét

Kết quả MAE:

MAE = 1.596 cho thấy trung bình, dự đoán của mô hình sai lệch khoảng 1.596 so với giá trị thực tế. Giá trị này cần được đánh giá trong ngữ cảnh của dữ liệu cụ thể để xác định xem mô hình có đạt yêu cầu không.

Công Thức Hồi Quy:

Student Performance = −33.969 + 2.852×Hours Studied + 1.018×Previous Scores + 0.604×Extracurricular Activities + 0.474×Sleep Hours + 0.192×Sample Question Papers Practiced

- Hệ số chệch (-33.969) là giá trị của Student Performance khi tất cả các đặc trưng đều bằng 0. Giá trị này có thể không có ý nghĩa thực tiễn trong một số ngữ cảnh vì các đặc trưng không phải lúc nào cũng có giá trị bằng 0 trong thực tế.
- Hours Studied (Số Giờ Học Tập): Hệ số là 2.852. Điều này có nghĩa là mỗi giờ học tập thêm có thể làm tăng Student Performance khoảng 2.852 điểm, nếu các yếu tố khác giữ nguyên.
- Previous Scores (Điểm Số Trước Đó): Hệ số là 1.018. Mỗi điểm số trước đó cao hơn có thể làm tăng Student Performance khoảng 1.018 điểm, nếu các yếu tố khác giữ nguyên.
- Extracurricular Activities (Hoạt Động Ngoại Khóa): Hệ số là 0.604. Mỗi hoạt động ngoại khóa thêm có thể làm tăng Student Performance khoảng 0.604 điểm, nếu các yếu tố khác giữ nguyên.
- Sleep Hours (Số Giờ Ngủ): Hệ số là 0.474. Mỗi giờ ngủ thêm có thể làm tăng Student Performance khoảng 0.474 điểm, nếu các yếu tố khác giữ nguyên.
- Sample Question Papers Practiced (Số Đề Tập): Hệ số là 0.192. Mỗi đề tập thêm có thể làm tăng Student Performance khoảng 0.192 điểm, nếu các yếu tố khác giữ nguyên.

Ý Nghĩa Các Hệ Số:

- Các hệ số dương cho thấy rằng tất cả các đặc trưng đều có ảnh hưởng tích cực đến Student Performance. Điều này đồng nghĩa với việc khi các giá trị của các đặc trưng tăng lên, Student Performance có xu hướng tăng lên.
- Hệ số cao nhất là của Hours Studied (2.852), cho thấy đây có thể là yếu tố quan trọng nhất ảnh hưởng đến hiệu suất học tập.
- Hệ số thấp nhất là của Sample Question Papers Practiced (0.192), cho thấy ảnh hưởng của số lượng đề tập đến Student Performance là nhỏ hơn so với các yếu tố khác.

Kết luận: Mô hình này được huấn luyện và đánh giá hiệu quả trên tập kiểm tra. Công thức hồi quy cung cấp thông tin về ảnh hưởng của các đặc trưng và MAE giúp đánh giá độ chính xác của mô hình.

Sử dụng duy nhất 1 đặc trưng

Ý tưởng thực hiện:

- Xác định đặc trưng tốt nhất trong số các đặc trưng đầu vào bằng cách sử dụng k-fold Cross Validation và sau đó xây dựng mô hình hồi quy tuyến tính dựa trên đặc trưng tốt nhất để dự đoán hiệu suất học tập.
- Đánh giá dựa trên so sánh các đặc trưng dựa trên chỉ số R^2 từ Cross Validation và đánh giá mô hình hồi quy dựa trên Mean Absolute Error (MAE).

Các bước thực hiện

1. Khởi Tạo Mô Hình và K-Fold Cross Validation:
    - Tạo đối tượng KFold với n_splits=5 để chia dữ liệu thành 5 phần.
    - Khởi tạo mô hình hồi quy tuyến tính LinearRegression.

2. Tính Toán R^2 Score cho Từng Đặc Trưng:
    - Đối với mỗi đặc trưng, chọn một đặc trưng duy nhất từ tập dữ liệu và tính toán R^2 score bằng cách sử dụng cross_val_score.
    - Lưu kết quả trung bình R^2 score của từng đặc trưng trong một từ điển.

3. Xác Định Đặc Trưng Tốt Nhất:
    - Xác định đặc trưng tốt nhất dựa trên giá trị R^2 score cao nhất.
    - In ra kết quả R^2 scores của từng đặc trưng và đặc trưng tốt nhất.

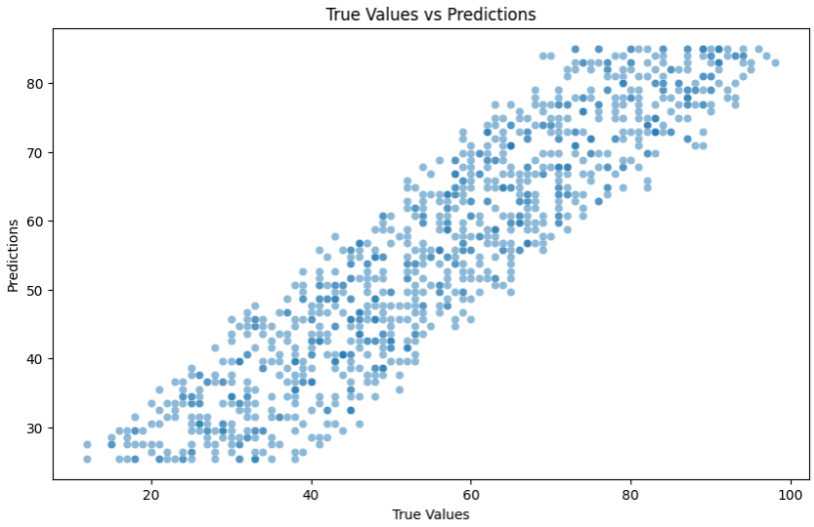
4. Huấn Luyện Mô Hình Với Đặc Trưng Tốt Nhất:
    - Huấn luyện lại mô hình hồi quy tuyến tính với đặc trưng tốt nhất.

- Dự đoán giá trị trên tập kiểm tra.

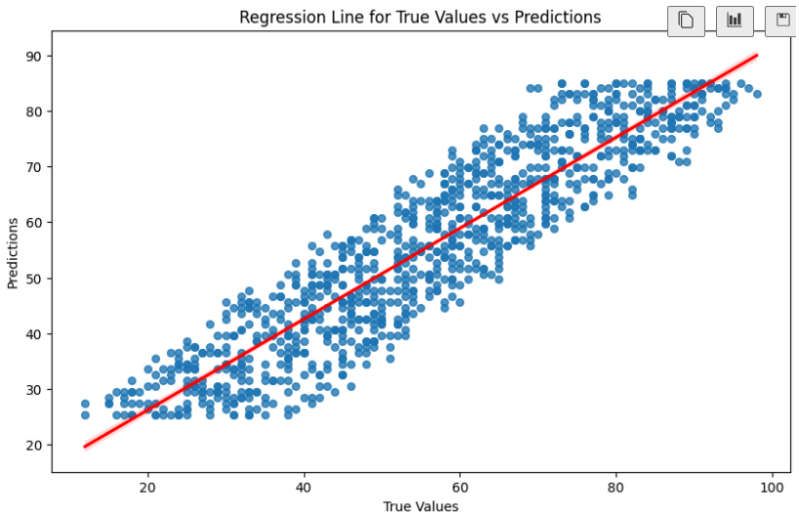
5. Đánh Giá Mô Hình:
   - Tính toán MAE để đánh giá độ chính xác của mô hình.
   - In ra công thức hồi quy dựa trên mô hình đã huấn luyện.

Ảnh demo

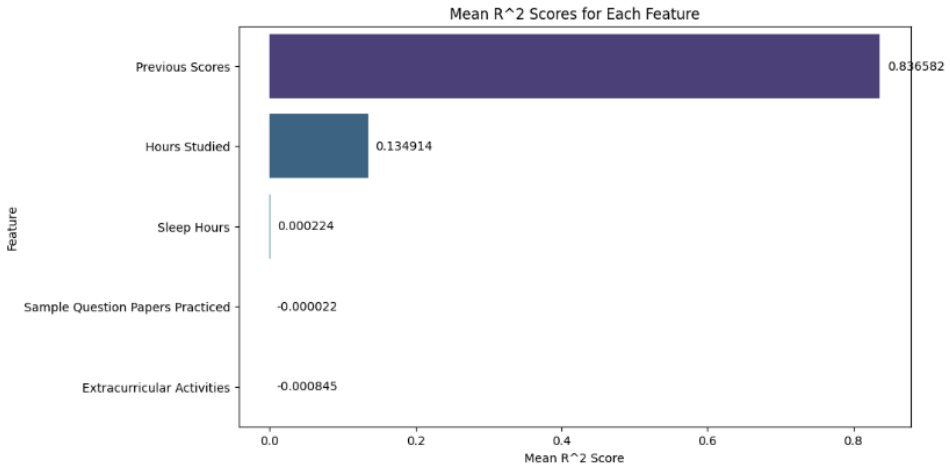
Biểu đồ phân tán dữ liệu giữa thực tế và dự đoán bởi mô hình



Đường hồi quy tuyến tính được vẽ



Biểu đồ cột cho từng đặc trưng



STT	Mô hình với đặc trưng	MAE
1	Hours Studied	0.134914
2	Previous Scores	0.836582
3	Extracurricular Activities	-0.000845
4	Sleep Hours	0.000224
5	Sample Question Papers Practiced	-0.000022

Đánh giá và nhận xét

Đặc Trưng Tốt Nhất:

- Kết Quả: Đặc trưng tốt nhất là Previous Scores với Mean R^2 score là 0.8366.
- Nhận Xét: Previous Scores có ảnh hưởng mạnh mẽ nhất đến Student Performance so với các đặc trưng khác. Điều này cho thấy rằng điểm số trước đó là yếu tố quan trọng nhất trong việc dự đoán hiệu suất học tập.

Mean Absolute Error (MAE):

- Kết Quả: MAE là 6.544.
- Nhận Xét: MAE cho thấy trung bình có sai lệch khoảng 6.544 điểm giữa giá trị dự đoán và giá trị thực tế. MAE cung cấp cái nhìn về độ chính xác của mô hình, và giá trị này có thể được coi là khá cao hoặc thấp tùy thuộc vào ngữ cảnh của dữ liệu và mức độ chính xác mong đợi.

Công Thức Hồi Quy: Student Performance = − 14.989 + 1.011 × Previous Scores

Nhận Xét:

- -14.989: Là giá trị Student Performance khi Previous Scores bằng 0. Trong thực tế, giá trị này có thể không có ý nghĩa đặc biệt nếu Previous Scores không bao giờ là 0.
- 1.011: Mỗi điểm số trước đó cao hơn sẽ làm tăng Student Performance khoảng 1.011 điểm. Hệ số này cho thấy mối quan hệ tích cực mạnh mẽ giữa điểm số trước đó và hiệu suất học tập

Kết luận: Previous Scores là đặc trưng tốt nhất với ảnh hưởng rõ ràng nhất đến hiệu suất học tập. Mức độ sai lệch (MAE) dự đoán không quá thấp, cho thấy có thể có các yếu tố khác ảnh hưởng đến hiệu suất học tập mà mô hình chưa lường trước.

Xây dựng/Thiết kế mô hình hồi quy tuyến tính

Tại phần này, chúng ta sẽ tập trung vào xây dựng và thiết kế 3 mô hình hồi quy tuyến tính khác nhau, tập trung vào 2 đặc trưng có sự ảnh hưởng nhiều nhất đến mô hình là “Hours Studied” và “Previous Scores”. Sau đó dựa trên kết quả thu được để chọn ra mô hình tốt nhất và tiến hành tái huấn luyện và đánh giá.

Cụ thể ý tưởng:

- Mô hình 1: Sử dụng hai đặc trưng "Hours Studied" và "Previous Scores" để dự đoán "Performance Index".
- Mô hình 2: Sử dụng hai đặc trưng sau khi chuẩn hóa để dự đoán "Performance Index".
- Mô hình 3: Tạo các đặc trưng mới bằng cách tổng hợp và bình phương các đặc trưng ban đầu và sử dụng chúng để dự đoán "Performance Index".

Mô hình tốt nhất: So sánh các mô hình dựa trên điểm số MAE để chọn mô hình tốt nhất. Sau đó, đánh giá mô hình tốt nhất trên tập kiểm tra bằng cách tính MAE và in ra công thức hồi quy.

Các bước thực hiện:

Mô Hình 1:

- Tạo mô hình hồi quy tuyến tính sử dụng "Hours Studied" và "Previous Scores".
- Tính toán điểm số MAE bằng k-fold cross-validation với k = 5.
- In kết quả MAE cho mô hình.

Mô Hình 2:

- Chuẩn hóa đặc trưng "Hours Studied" và "Previous Scores" bằng StandardScaler.
- Tạo mô hình hồi quy tuyến tính với các đặc trưng đã chuẩn hóa.
- Tính toán điểm số MAE bằng k-fold cross-validation với k = 5.
- In kết quả MAE cho mô hình.

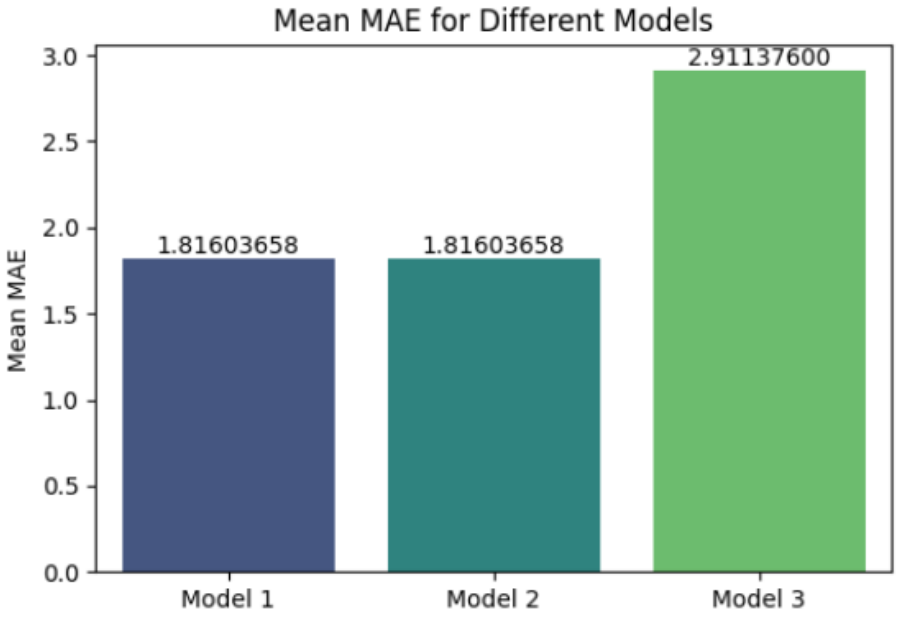
Mô Hình 3:

- Tạo các đặc trưng mới từ "Hours Studied" và "Previous Scores".
- Tạo mô hình hồi quy tuyến tính với các đặc trưng mới.
- Tính toán điểm số MAE bằng k-fold cross-validation với k = 5.
- In kết quả MAE cho mô hình.

Tìm mô hình tốt nhất:

- So sánh MAE của các mô hình.
- Chọn mô hình tốt nhất dựa trên điểm số R^2 cao nhất.
- Huấn luyện lại mô hình tốt nhất trên toàn bộ tập huấn luyện.
- Dự đoán trên tập kiểm tra.
- Đánh giá mô hình tốt nhất bằng MAE và in ra công thức hồi quy.

Ảnh kết quả MAE thu được trên từng mô hình



STT	Mô hình	MAE
1	Model 1	1.81603658
2	Model 2	1.81603658
3	Model 3	2.91137600

Đánh giá và nhận xét

Model 1: Hồi quy tuyến tính với hai đặc trưng "Hours Studied" và "Previous Scores".

- Mean MAE: 1.81603658

Model 2: Hồi quy tuyến tính với các đặc trưng đã chuẩn hóa.

- Mean MAE: 1.81603658

Model 3: Hồi quy tuyến tính với các đặc trưng biến đổi (Tổng và bình phương của "Previous Scores").

- Mean MAE: 2.91137600

Kết luận: Mô hình tốt nhất là Model 2 với MAE trung bình là 1.81603658, bằng với Model 1, nhưng được chọn vì sự nhất quán trong hiệu suất và các yếu tố chuẩn hóa có thể mang lại sự ổn định hơn trong các điều kiện khác nhau.

Ảnh biểu đồ phân tán cho mô hình tốt nhất (Model 2)

Vẽ đường tuyến tính hồi quy cho mô hình tốt nhất

Đánh giá

Mô hình tốt nhất: Mô hình 2

- Mean MAE: 1.81603658
- Công thức hồi quy: Student Performance=55.136+7.409×Hours Studied+17.688×Previous Scores

Nhận xét:

- Model 2 đã thể hiện hiệu suất tốt nhất về mặt MAE so với các mô hình khác. Điều này cho thấy việc chuẩn hóa các đặc trưng có thể giúp cải thiện độ chính xác của mô hình hồi quy tuyến tính.
- Model 1 và Model 2 có MAE tương đương, nhưng Model 2 được ưu tiên vì chuẩn hóa có thể cung cấp sự ổn định hơn trong điều kiện thực tế.
- Model 3, mặc dù sáng tạo với các đặc trưng biến đổi, không hoạt động tốt như các mô hình còn lại, với MAE cao hơn đáng kể.

IV/ Trích dẫn tham khảo

[1] Hồi quy tuyến tính, scikit-learn LinearRegression [Website]

[2] KFold Cross Validation, scikit-learn KFold [Website]

[3] KFold Cross Validation, scikit-learn cross_val_score

[4] Ideas Reference, Đặng Ngọc Tiến [Github 2022]

[5] Vẽ đồ thi bằng thư viện Seaborn [Website, Turotial]