TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

KHOA CÔNG NGHỆ THÔNG TIN

ĐẶNG HỒ THIÊN PHÚC – 52300145

TRẦN PHẠM ANH QUÂN – 52300150

BÙI NGỌC QUÝ – 52300153

HÀ NGỌC NHI – 52300141

VÕ VIỆT QUÂN – 52300151

DỰ ĐOÁN KẾT QUẢ HỌC TẬP

BÁO CÁO CUỐI KỲ

KHAI PHÁ DỮ LIỆU VÀ TRI THỨC

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

KHOA CÔNG NGHỆ THÔNG TIN

ĐẶNG HỒ THIÊN PHÚC – 52300145

TRẦN PHẠM ANH QUÂN – 52300150

BÙI NGỌC QUÝ – 52300153

HÀ NGỌC NHI – 52300141

VÕ VIỆT QUÂN – 52300151

DỰ ĐOÁN KẾT QUẢ HỌC TẬP

BÁO CÁO CUỐI KỲ

KHAI PHÁ DỮ LIỆU VÀ TRI THỨC

Người hướng dẫn

TS. Hoàng Anh

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025

LỜI CẢM ƠN

Lời đầu tiên, chúng em xin gửi lời tri ân sâu sắc đến thầy Hoàng Anh, người đã tận tâm truyền đạt cho chúng em những kiến thức quý báu về môn khai thác dữ liệu và khai phá tri thức. Nhờ sự hướng dẫn tận tình của thầy, chúng em không chỉ hiểu rõ hơn về cách thức và quy trình để tạo ra mô hình dự đoán mà còn tiếp thu được những phương pháp học tập hiệu quả cùng những tài liệu hữu ích, góp phần quan trọng vào quá trình hoàn thành bài báo cáo này.

Bên cạnh đó, chúng em cũng xin gửi lời cảm ơn chân thành đến khoa Công nghệ thông tin đã luôn tạo điều kiện thuận lợi để chúng em có cơ hội học tập và phát triển tư duy.

Chúng em ý thức được rằng, với vốn kiến thức và kinh nghiệm thực tiễn còn hạn chế, ài báo cáo này chắc chắn vẫn còn nhiều thiếu sót. Vì vậy, chúng em rất mong nhận được những ý kiến đóng góp quý báu từ thầy để có thể tiếp tục hoàn thiện và nâng cao hơn nữa kiến thức cũng như kỹ năng của mình trong những bài báo cáo sau.

Chúng em xin chân thành cảm ơn và kính chúc thầy dồi dào sức khỏe, luôn giữ vững nhiệt huyết để tiếp tục truyền lửa cho các thế hệ sinh viên trên con đường học tập và phát triển.

TP. Hồ Chí Minh, ngày 18 tháng 12 năm 2025.

Tác giả 1 *Tác giả 2* *Tác giả 3*

Đặng Hồ Thiên Phúc *Trần Phạm Anh Quân* *Bùi Ngọc Quý*

Tác giả 4 *Tác giả 5*

Hà Ngọc Nhi *Võ Việt Quân*

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH
TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của TS. Hoàng Anh. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong Dự án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Dự án của mình. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày 18 tháng 12 năm 2025.

Tác giả 1	*Tác giả 2*	*Tác giả 3*
Đặng Hồ Thiên Phúc	*Trần Phạm Anh Quân*	*Bùi Ngọc Quý*
Tác giả 4	*Tác giả 5*	
Hà Ngọc Nhi	*Võ Việt Quân*	

DỰ ĐOÁN KẾT QUẢ HỌC TẬP

TÓM TẮT

Trong bối cảnh giáo dục ngày càng được quan tâm và yêu cầu về chất lượng đào tạo không ngừng được nâng cao, việc theo dõi và đánh giá kết quả học tập của học sinh đóng vai trò quan trọng trong công tác quản lý giáo dục. Tuy nhiên, trên thực tế chất lượng học tập của học sinh có sự phân hóa rõ rệt và việc đánh giá hiện nay vẫn chủ yếu dựa trên các phương pháp truyền thống như điểm số và nhận xét của giáo viên, chưa đáp ứng tốt yêu cầu phát hiện sớm các học sinh có nguy cơ học tập kém.

Xuất phát từ thực tiễn đó, báo cáo sẽ tập trung nghiên cứu bài toán dự đoán kết quả học tập của học sinh dựa trên dữ liệu học tập sẵn có. Trên cơ sở khai thác và phân tích dữ liệu, các kỹ thuật học máy được áp dụng nhằm xây dựng các mô hình dự đoán phù hợp. Kết quả thực nghiệm cho thấy các mô hình có khả năng dự đoán tương đối chính xác kết quả học tập của học sinh, từ đó hỗ trợ hiệu quả cho giáo viên và nhà trường trong việc theo dõi, đánh giá và quản lý học tập.

Từ kết quả đạt được, đề tài đề xuất một số khuyến nghị nhằm giúp nhà trường và giáo viên có thể chủ động hơn trong việc phát hiện sớm học sinh gặp khó khăn trong học tập, đưa ra các biện pháp hỗ trợ kịp thời và góp phần nâng cao chất lượng giáo dục.

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU

DANH MỤC CÁC CHỮ VIẾT TẮT

R2	R-squared
MSE	Mean Squared Error
RMSE	Root Mean Squared Error

BẢNG PHÂN CÔNG CÔNG VIỆC

Mã số sinh viên	Họ và tên	Công việc phân công	Tỷ lệ hoàn thành
52300145	Đặng Hồ Thiên Phúc	Xây dựng mô hình Tìm giải pháp cho bài toán	100%
52300150	Trần Phạm Anh Quân	Phân tích dữ liệu Viết báo cáo	100%
52300141	Hà Ngọc Nhi	Tìm kiếm dữ liệu Viết báo cáo	100%
52300153	Bùi Ngọc Quý	Tìm kiếm dữ liệu Viết báo cáo	100%
52300151	Võ Việt Quân	Xây dựng mô hình Tìm giải pháp cho bài toán.	100%

CHƯƠNG 1. TỔNG QUAN ĐỀ TÀI

1.1 Bài toán cần giải quyết từ đề tài

Trong những năm gần đây, tình trạng học sinh có kết quả học tập sa sút, bị xếp loại học lực yếu hoặc có nguy cơ không đạt yêu cầu lên lớp đang có xu hướng gia tăng tại nhiều cơ sở giáo dục. Sự phát triển của chương trình học với số lượng môn học và yêu cầu ngày càng cao khiến không ít học sinh gặp khó khăn trong việc theo kịp tiến độ học tập. Thực tế cho thấy, nhiều học sinh chỉ được phát hiện gặp vấn đề về học lực khi kết quả học tập đã giảm sút nghiêm trọng, dẫn đến tâm lý chán nản, giảm động lực học tập và khó cải thiện trong thời gian ngắn.

Có nhiều nguyên nhân dẫn đến tình trạng trên, tuy nhiên việc đánh giá học sinh hiện nay chủ yếu vẫn dựa trên các phương pháp truyền thống như điểm số và sự theo dõi của giáo viên chủ nhiệm, điều này chưa đủ để phát hiện sớm các nguy cơ tiềm ẩn. Do đó, bài toán được đặt ra là làm thế nào có thể dự đoán trước kết quả học tập của học sinh, từ đó kịp thời can thiệp và xác định các nguyên nhân chính nhằm đề xuất những biện pháp hỗ trợ và khắc phục phù hợp.

1.2 Mặt thuận lợi và khó khăn của bài toán

1.2.1 Thuận lợi

- Hầu hết các trường học hiện nay đều lưu trữ hồ sơ học tập điện tử hoặc sổ điểm của học sinh, bao gồm điểm số, hạnh kiểm và số buổi nghỉ học. Đây là nguồn dữ liệu sẵn có, phản ánh quá trình học tập và tạo điều kiện thuận lợi cho việc thu thập thông tin phục vụ bài toán nghiên cứu.
- Bên cạnh đó, giáo viên với kinh nghiệm giảng dạy thực tiễn có thể xác định được các yếu tố ảnh hưởng đến kết quả học tập của học sinh.

1.2.2 Khó khăn

- Việc đánh giá kết quả học tập của học sinh hiện nay chủ yếu vẫn được thực hiện thủ công dựa trên điểm số và nhận xét của giáo viên, nên chưa có khả năng dự đoán trước kết quả học tập.

- Phương pháp này phụ thuộc nhiều vào đánh giá chủ quan và chưa khai thác được mối quan hệ giữa các yếu tố học tập khác nhau, dẫn đến khó khăn trong việc phát hiện sớm học sinh có nguy cơ học tập kém.

1.3 Phương pháp giải quyết

Bài toán sẽ được giải quyết theo hướng xây dựng và so sánh các mô hình học máy hồi quy nhằm dự đoán kết quả học tập của học sinh từ tập dữ liệu thu thập được. Thay vì sử dụng một mô hình duy nhất, bài toán sẽ được giải quyết theo hướng thực nghiệm để đánh giá hiệu quả của nhiều thuật toán khác nhau trên cùng một tập dữ liệu.

Phương pháp này tận dụng trực tiếp nguồn dữ liệu học tập đã được thu thập trước đó, bao gồm cả đặc trưng số và đặc trưng phân loại. Dữ liệu được tiền xử lý trước khi huấn luyện mô hình, gồm nhiều bước như scaling, biến đổi skewness–kurtosis, mã hóa dữ liệu và giảm chiều bằng SparsePCA được áp dụng có chọn lọc theo yêu cầu của từng mô hình cụ thể.

Quá trình thực nghiệm sẽ được thực hiện song song trên nhiều mô hình bao gồm các mô hình tuyến tính (Linear Regression, Ridge, Lasso) và các mô hình dựa trên cây quyết định và boosting (Random Forest, XGBoost, CatBoost). Mỗi mô hình được huấn luyện với cấu hình tiền xử lý phù hợp nhằm đảm bảo tính công bằng trong quá trình đánh giá.

Kết quả cuối cùng của các mô hình được so sánh thông qua kiểm định chéo K-Fold và các chỉ số đánh giá MSE, RMSE và R^2. Dựa trên kết quả so sánh, mô hình có hiệu suất và độ ổn định tốt nhất được lựa chọn để tiếp tục tối ưu và huấn luyện để cho ra kết quả tốt nhất.

CHƯƠNG 2. TỔNG QUAN VỀ TẬP DỮ LIỆU

2.1 Nguồn gốc của tập dữ liệu

- Nguồn gốc: Tập dữ liệu được trích xuất từ diễn đàn Kaggle (Đường dẫn)
- Độ tin cậy/tin dùng của các người dùng trên diễn đàn:
    - Về số lượt upvote: Tính đến 23/12/2025, tập dữ liệu này có 1131 lượt upvote.
    - Về đánh giá của Kaggle về tập dữ liệu: Trên 3 thang điểm là: Độ hoàn thiện (Completeness), độ tin cậy (Credibility) và tính tương thích (Compatibility). Cả 3 thang điểm này tập dữ liệu này đều đạt 100%. Chứng minh sự hợp lệ của tập dữ liệu.

2.2 Phân tích khám phá dữ liệu

2.2.1 Sơ lược về tập dữ liệu

- Bộ dữ liệu ghi lại thông tin của 6,607 học sinh với 20 biến số khác nhau, bao gồm các yếu tố về cá nhân, gia đình, môi trường học tập và kết quả thi cuối kỳ (Exam_Score).
    - Số lượng bản ghi: 6,607 dòng.
    - Số lượng đặc trưng: 20 đặc trưng (bao gồm 6 đặc trưng liên tục, 9 đặc trưng thứ bậc và 5 đặc trưng nhị phân).
    - Mục tiêu: Exam_Score (Điểm thi).
- Ở Bảng 2.1 bao gồm các đặc trưng và mô tả của tập dữ liệu. Các đặc trưng đầu vào được chia ra thành 3 yếu tố:
    - Yếu tố từ học sinh: Đánh giá năng lực, sức khỏe thể chất và tinh thần của học sinh.
    - Yếu tố từ gia đình: Đánh giá điều kiện kinh tế của gia đình, mức độ quan tâm của gia đình.
    - Yếu tố học đường: Đánh giá chất lượng giảng dạy của trường học của học sinh.

Yếu tố	**Tên đặc trưng**	**Mô tả**
Yếu tố từ học sinh	Hours_Studied	Số giờ học mỗi tuần
	Attendance	Tỷ lệ phần trăm tham gia lớp học
	Previous_Scores	Điểm số từ các kỳ thi trước
	Motivation_Level	Mức độ động lực của học sinh
	Sleep_Hours	Số giờ ngủ trung bình mỗi đêm
	Extracurricular_Activities	Tham gia các hoạt động ngoại khóa
	Tutoring_Sessions	Số buổi học thêm mỗi tháng
	Physical_Activity	Số giờ hoạt động thể chất trung bình mỗi tuần
	Learning_Disabilities	Có khuyết tật/khó khăn về học tập hay không (Có, Không)
	Gender	Giới tính của học sinh (Nam, Nữ).
Yếu tố gia đình	Parental_Involvement	Mức độ tham gia của phụ huynh vào việc học của học sinh
	Parental_Education_Level	Trình độ học vấn cao nhất của phụ huynh (Trung học, Đại học, Sau đại học)

	Family_Income	Mức thu nhập gia đình (Thấp, Trung bình, Cao)
	Access_to_Resources	Khả năng tiếp cận các nguồn tài liệu giáo dục
	Internet_Access	Có truy cập internet hay không
	Distance_from_Home	Khoảng cách từ nhà đến trường (Gần, Trung bình, Xa)
Yếu tố học đường	Parental_Involvement	Mức độ tham gia của phụ huynh vào việc học của học sinh
	Parental_Education_Level	Trình độ học vấn cao nhất của phụ huynh (Trung học, Đại học, Sau đại học)
	Family_Income	Mức thu nhập gia đình (Thấp, Trung bình, Cao)
	Exam_Score	Điểm thi cuối kỳ.

Bảng 2.1 Bảng mô tả các đặc trưng

- Ở Bảng 2.2, bảng biểu diễn kiểu đặc trưng và giá trị duy nhất của chúng. Có 3 kiểu đặc trưng trong tập dữ liệu này:
    - Liên tục.
    - Nhị phân: Thuộc về kiểu đặc trưng rời rạc, giá trị duy nhất bao gồm 2 giá trị.
    - Thứ bậc: Thuộc về kiểu đặc trưng rời rạc, giá trị duy nhất bao gồm từ 3 giá trị trở lên và các giá trị thể hiện mức độ khác nhau.

Tên đặc trưng	**Kiểu đặc trưng**	**Giá trị duy nhất (đối với đặc trưng rời rạc)**
Hours_Studied	Liên tục	
Attendance	Liên tục	
Parental_Involvement	Thứ bậc	[Low, Medium, High]
Access_to_Resources	Thứ bậc	[High, Medium, Low]
Extracurricular_Activities	Nhị phân	[No, Yes]
Sleep_Hours	Liên tục	
Previous_Scores	Liên tục	
Motivation_Level	Thứ bậc	[Low, Medium, High]
Internet_Access	Nhị phân	[Yes, No]
Tutoring_Sessions	Liên tục	
Family_Income	Thứ bậc	[Low, Medium, High]
Teacher_Quality	Thứ bậc	[Medium, High, Low]
School_Type	Nhị phân	[Public, Private]
Peer_Influence	Thứ bậc	[Positive, Negative, Neutral]
Physical_Activity	Liên tục	
Learning_Disabilities	Nhị phân	[No, Yes]

Parental_Education_Level	Thứ bậc	[High School, College, Postgraduate]
Distance_from_Home	Thứ bậc	[Near, Moderate, Far]
Gender	Nhị phân	[Male, Female]
Exam_Score	Liên tục	

Bảng 2.2 Bảng kiểu đặc trưng và giá trị duy nhất của các đặc trưng rời rạc

2.2.2 Các thông số kĩ thuật và thống kê dữ liệu

- Đánh giá chất lượng dữ liệu và dữ liệu thiếu (Missing Data): bộ dữ liệu có chất lượng tốt với tỷ lệ dữ liệu đầy đủ cao. Tuy nhiên, có 3 biến số xuất hiện giá trị thiếu là:
    - Parental_Education_Level với tỷ lệ thiếu 1.36%.
    - Teacher_Quality với tỷ lệ thiếu 1.18%.
    - Distance_from_Home với tỷ lệ thiếu 1.01%.
- Nhận xét: Tỷ lệ thiếu rất thấp (đều < 2%), do đó có thể xử lý dễ dàng bằng phương pháp xóa bỏ hoặc thay thế bằng giá trị yếu vị mà không làm sai lệch đáng kể kết quả phân tích.

Tên biến	**Data Type**	**Non-Null**	**Null Count**	**Null %**
Hours_Studied	int64	6607	0	0.00
Attendance	int64	6607	0	0.00
Parental_Involvement	object	6607	0	0.00
Access_to_Resources	object	6607	0	0.00
Extracurricular_Activities	object	6607	0	0.00
Sleep_Hours	int64	6607	0	0.00
Previous_Scores	int64	6607	0	0.00
Motivation_Level	object	6607	0	0.00
Internet_Access	object	6607	0	0.00
Tutoring_Sessions	int64	6607	0	0.00
Family_Income	object	6607	0	0.00
Teacher_Quality	object	6529	78	1.18
School_Type	object	6607	0	0.00
Peer_Influence	object	6607	0	0.00

Physical_Activity	int64	6607	0	0.00
Learning_Disabilities	object	6607	0	0.00
Parental_Education_Level	object	6517	90	1.36
Distance_from_Home	object	6540	67	01.01
Gender	object	6607	0	0.00
Exam_Score	int64	6607	0	0.00

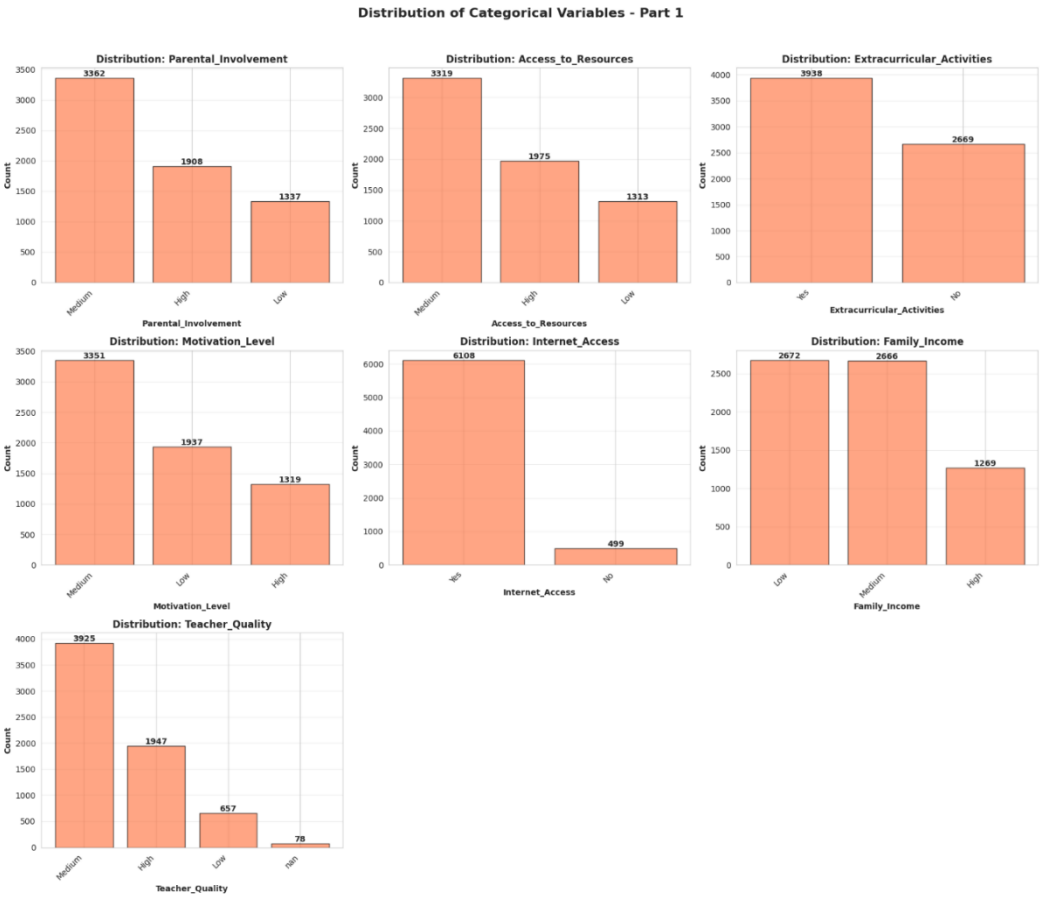
Bảng 2.3 Bảng đánh giá chất lượng của tập dữ liệu

2.2.3 Phân tích độc lập đặc trưng

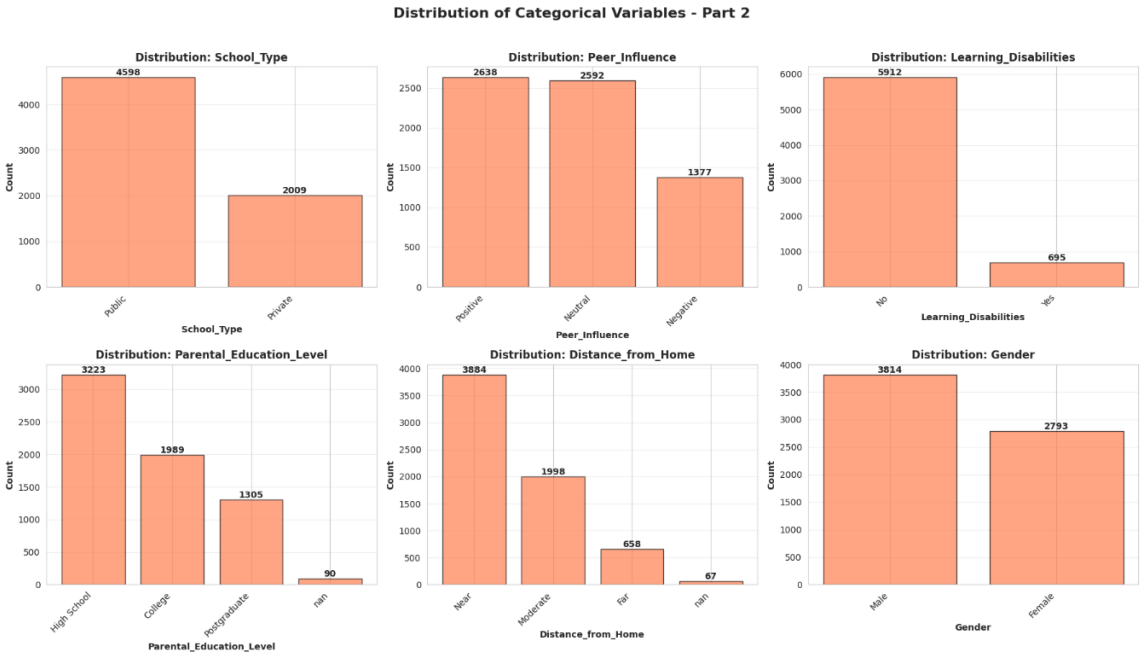
2.2.3.1 Phân tích độc lập đặc trưng rời rạc

- Parent_Involvement và Motivation_Level: Cả hai thuộc tính này có xu hướng phân phối tương tự nhau. Đa số học sinh ở mức Medium (trung bình), sau đó là High/Low. Điều này cho thấy môi trường học tập khá ổn định nhưng thiếu sự bứt phá về động lực cực cao.
- Internet Access: Đây là biến có sự mất cân bằng rõ rệt nhất. Đại đa số học sinh (hơn 6100) có quyền truy cập Internet, trong khi chỉ có khoảng 500 em là không. Điều này là một lợi thế lớn cho việc khai thác tài liệu trong học tập.
- Access_to_Resources và Extracurricular_Activities: Đa số học sinh có mức tiếp cận nguồn lực trung bình và có tham gia hoạt động ngoại khóa (gần 4000 em có tham gia so với khoảng 2600 em không tham gia).
- Teacher_Quality: đáng chú ý là chất lượng giáo viên được đánh giá chủ yếu ở mức Medium (hơn 3900), trong khi mức Low chỉ chiếm tỉ lệ rất nhỏ.
- Family_Income: Số lượng học sinh có gia đình ở mức thu nhập thấp (Low: 2672) và mức trung bình (Medium: 2666) gần như tương đương nhau. Chỉ có 1269 học sinh đến từ gia đình có thu nhập cao, chiếm chưa đến một nửa so với mỗi nhóm còn lại.

- School Type: Số lượng học sinh học trường công lập (Public) cao gấp hơn 2 lần so với trường tư thục (Private).
- Trình độ học vấn của phụ huynh: phần lớn phụ huynh dừng lại ở mức High School (Cấp 3). Số lượng phụ huynh có trình độ Postgraduate (Sau đại học) là thấp nhất. Có thể là biến số ảnh hưởng đến kỳ vọng và định hướng cho học sinh.
- Khoảng cách từ nhà đến trường: đa số học sinh sống ở mức Near (Gần), giúp giảm thiểu thời gian đi lại và có thể tăng thời gian tự học.
- Giới tính (Gender): Có sự chênh lệch nhẹ khi số lượng học sinh Nam (3814) nhiều hơn học sinh Nữ (2793).
- Khuyết tật học tập (Learning Disabilities): Chỉ một nhóm nhỏ học sinh (695 em) ghi nhận có khuyết tật, phần lớn còn lại là không.



Hình 2.1 Biểu đồ biểu diễn phân phối của 7 đặc trưng rời rạc đầu tiên



Hình 2.2 Biểu đồ biểu diễn phân phối của 6 đặc trưng rời rạc tiếp theo

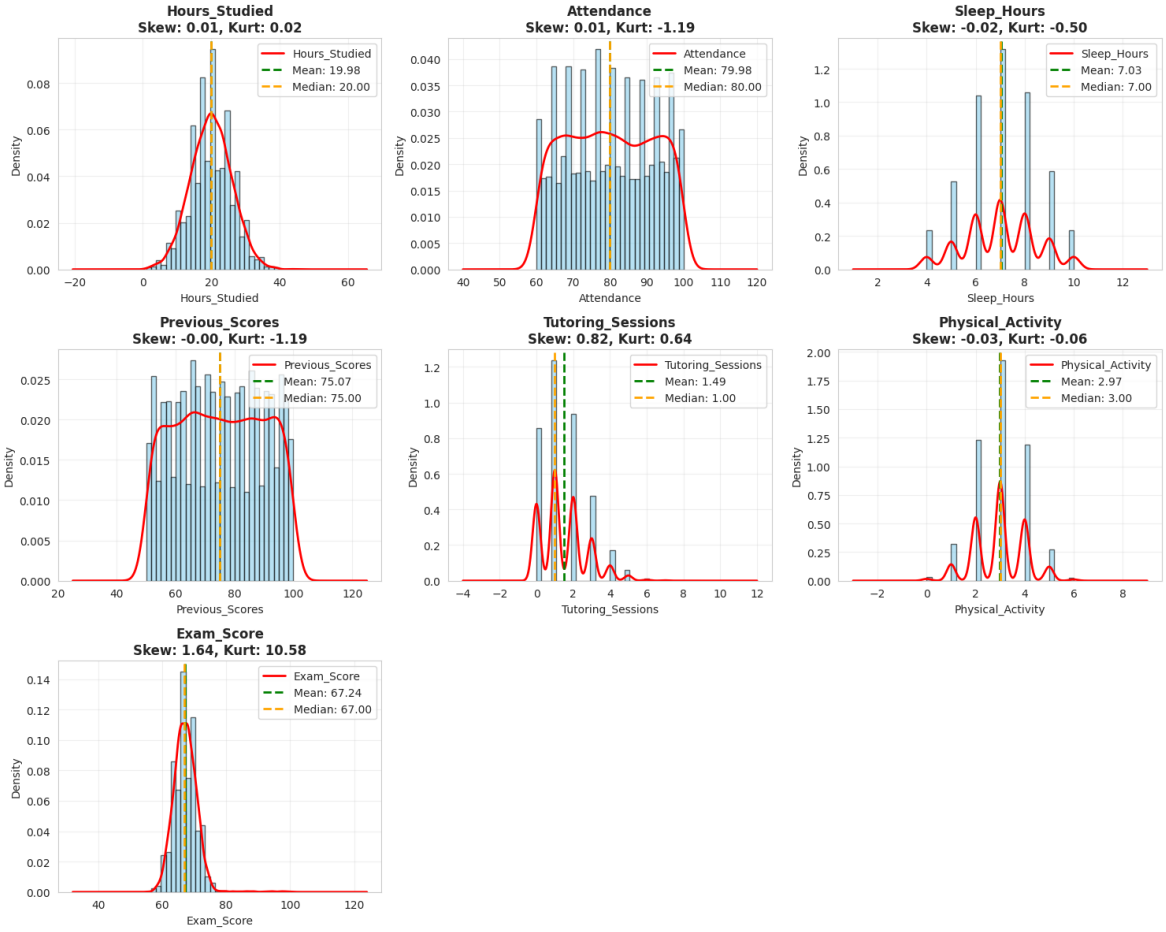
2.2.3.2 Phân tích độc lập đặc trưng liên tục

- Để phân tích được phân phối của các đặc trưng liên tục, cần làm rõ 2 khái niệm: Skewness và Kurtosis.
    - Độ lệch (Skewness) là chỉ số thống kê phản ánh mức độ mất cân đối của phân phối xác suất so với phân phối chuẩn. Một phân phối đối xứng hoàn hảo sẽ có độ lệch bằng 0. Giá trị Skewness dương cho thấy phân phối lệch phải nghĩa là tập trung ở giá trị thấp, đuôi kéo dài về phía giá trị cao, ngược lại, giá trị âm biểu thị sự lệch trái.
    - Độ nhọn (Kurtosis) là đại lượng đo lường mức độ tập trung của dữ liệu tại vùng trung tâm so với vùng đuôi. Nó cho biết độ dày của đuôi phân phối. Một phân phối có Kurtosis cao đồng nghĩa với việc tồn tại nhiều giá trị cực đoan nằm xa giá trị trung bình, trong khi Kurtosis thấp thể hiện sự phân tán đồng đều hơn và ít giá trị ngoại lai.
- Bảng 2.4 dưới đây nhận xét về phân phối của các đặc trưng liên tục:

Đặc trưng	**Phân tích phân phối**	**Độ lệch (Skewness)**	**Độ nhọn (Kurtosis)**	**Nhận xét**
Hours_Studied	Phân phối gần chuẩn (Normal-like)	0.013	0.018	Dữ liệu khá đối xứng, tập trung quanh trung bình (~20 giờ). Không có outliers đáng kể.
Attendance	Phân phối phẳng (Uniform-like)	0.014	-1.194	Tỷ lệ đi học trải đều từ 60%–100%. Kurtosis âm cho thấy dữ liệu không tập trung vào một mức cụ thể.

Sleep_Hours	Phân phối gần chuẩn	-0.024	-0.504	Phân phối đối xứng, học sinh thường ngủ quanh các mốc cố định (≈7 giờ). Ít biến động cực đoan.
Previous_Scores	Phân phối phẳng	-0.004	-1.191	Điểm số cũ trải đều từ 50–100, giúp mô hình học tốt trên toàn bộ dải điểm.
Tutoring_Sessions	Lệch phải vừa (Moderate positive skew)	0.816	0.644	Đa số học sinh học thêm ít buổi (0–2). Một số học sinh học rất nhiều → đuôi phải dài.
Physical_Activity	Phân phối đối xứng	-0.031	-0.059	Hoạt động thể chất ổn định, tập trung chủ yếu trong khoảng 2–4 giờ/tuần.
Exam_Score	Lệch phải mạnh, nhọn (Highly skewed & leptokurtic)	1.645	10.575	Biến mục tiêu. Dữ liệu tập trung mạnh quanh ~70 điểm, kurtosis rất cao → khả năng tồn tại outliers điểm cao.

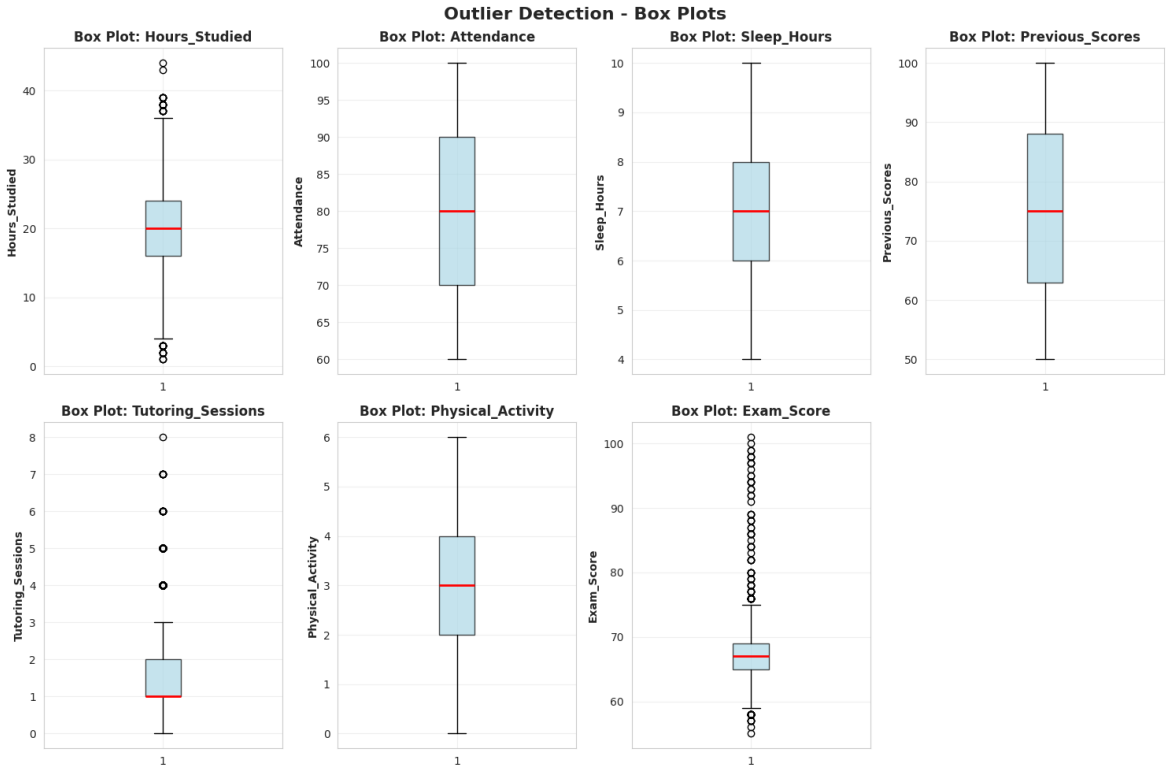
Bảng 2.4 Bảng nhận xét phân phối của các đặc trưng liên tục



Hình 2.3 Biểu diễn phân phối của các đặc trưng liên tục

2.2.4 Phân tích các đặc trưng có giá trị ngoại lai (Outliers):

- Hours_Studied: Xuất hiện ngoại lai ở cả hai phía (thấp và cao). Có những sinh viên học rất ít (dưới 5 giờ) và một số ít học cực nhiều (trên 35-40 giờ).
- Tutoring_Sessions: Có nhiều giá trị ngoại lai ở phía trên. Trong khi đa số chỉ học 0-2 buổi, có những cá biệt học từ 4 đến 8 buổi. Đây là nhóm dữ liệu cần lưu ý vì có thể là các trường hợp đặc biệt cần hỗ trợ.
- Exam_Score (Mục tiêu): Đây là đặc trưng có nhiều ngoại lai nhất, đặc biệt là ở phía trên (điểm cao). Điều này giải thích tại sao ở biểu đồ Histogram trước đó, Kurtosis của biến này lại rất cao (10.58). Có một nhóm sinh viên đạt điểm vượt trội (từ 75 đến trên 100) so với phần còn lại của lớp.



Hình 2.4 Biểu diễn vùng giá trị ngoại lai của các đặc trưng liên tục

2.2.5 Ma trận tương quan giữa các đặc trưng

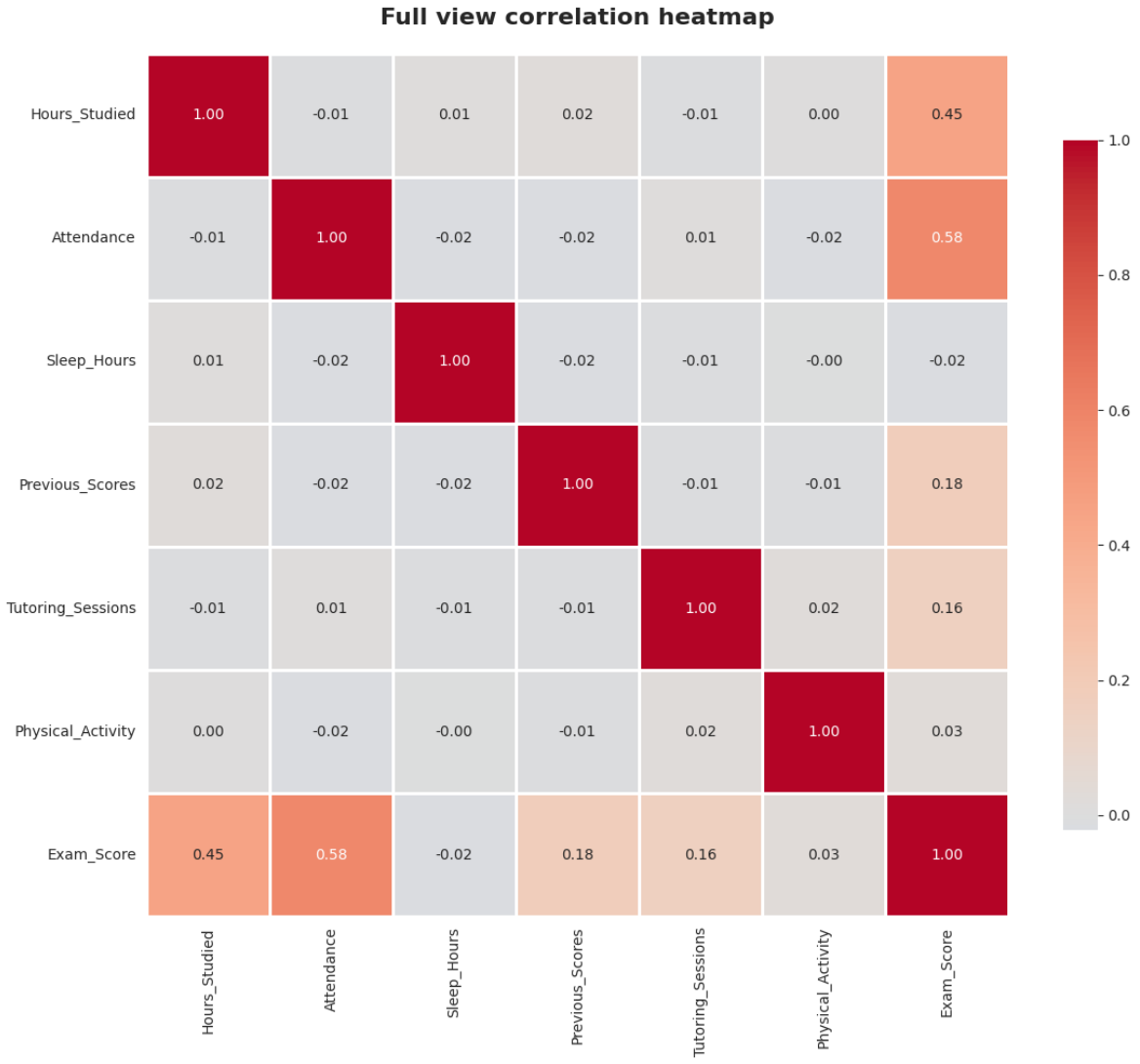
2.2.5.1 Khái niệm của ma trận tương quan

- Ma trận tương quan là công cụ thống kê dùng để đánh giá mức độ và chiều hướng của mối quan hệ tuyến tính giữa các cặp biến số định lượng.
- Đối với bài toán dự báo điểm số, việc phân tích ma trận tương quan có ý nghĩa quan trọng trong việc nhận diện hiện tượng đa cộng tuyến, giúp loại bỏ các biến độc lập có sự phụ thuộc lẫn nhau quá lớn, ví dụ như số giờ học và chi phí học tập. Đồng thời, công cụ này hỗ trợ chọn lọc những biến đầu vào có mối tương quan mạnh nhất với biến mục tiêu là kết quả học tập, từ đó nâng cao tính ổn định và độ chính xác cho các mô hình hồi quy.

2.2.5.2 Ma trận tương quan giữa các đặc trưng liên tục

- Tương quan mạnh nhất: Attendance (0.58): Tỷ lệ đi học là yếu tố có tác động tích cực nhất đến điểm thi.

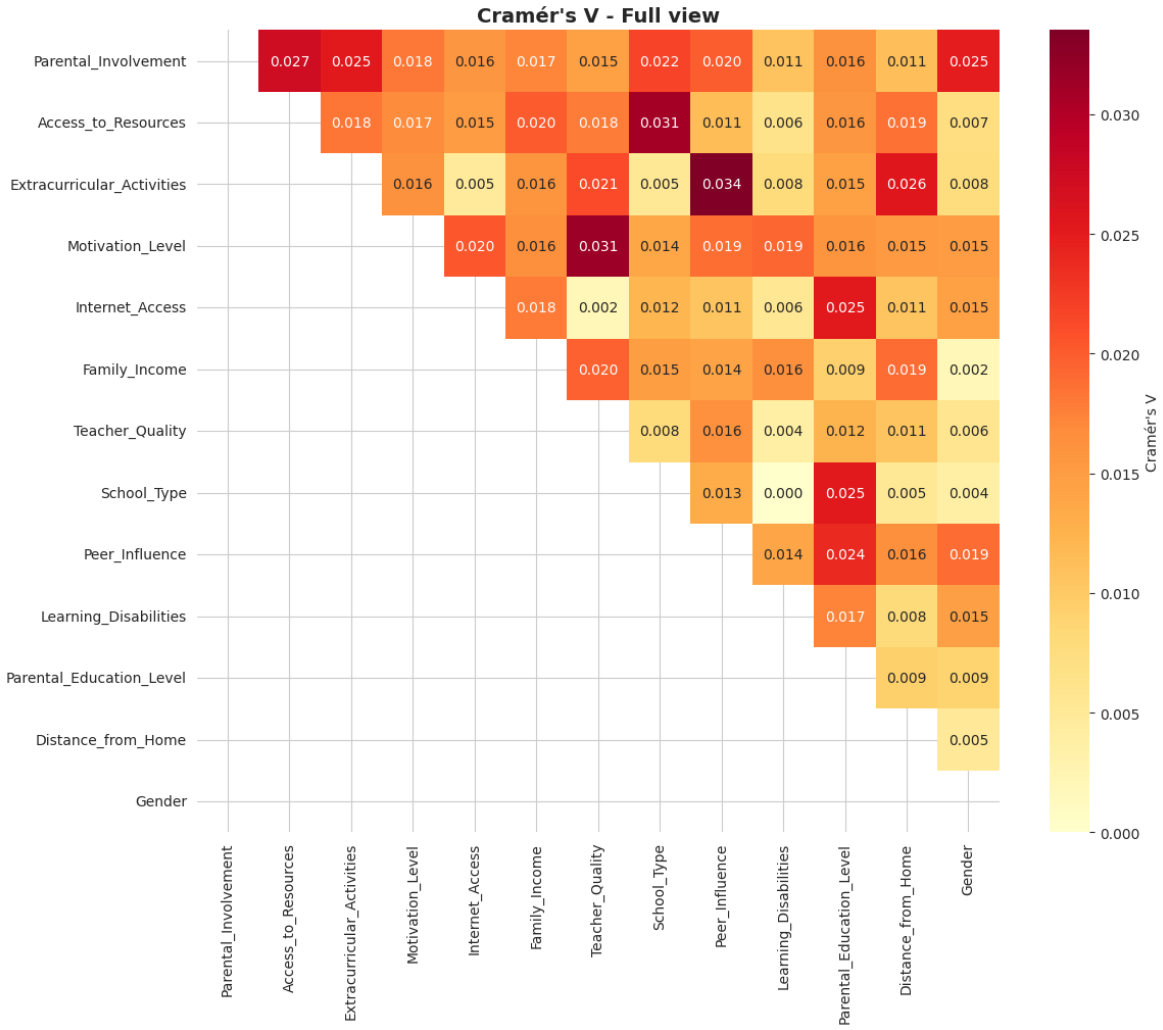
- Tương quan trung bình: Hours_Studied (0.45): Số giờ học có mối tương quan thuận rõ rệt với điểm số. Sinh viên dành nhiều thời gian học tập có xu hướng đạt kết quả cao hơn.
- Tương quan yếu: Previous_Scores (0.18) và Tutoring_Sessions (0.16): Đáng ngạc nhiên là điểm số cũ và việc học thêm chỉ có tác động nhẹ đến kết quả thi cuối cùng.
- Không có tương quan: Sleep_Hours (-0.02) và Physical_Activity (0.03): Hai biến này gần như không ảnh hưởng đến điểm số.
- Vấn đề đa cộng tuyến (Multicollinearity): Các biến độc lập gần như không tương quan với nhau (các giá trị quanh mức 0.00 đến 0.02). Đây là dấu hiệu rất tốt, cho thấy dữ liệu không bị chồng chéo thông tin



Hình 2.5 Ma trận tương quan giữa các đặc trưng liên tục

2.2.5.3 Ma trận tương quan giữa các đặc trưng rời rạc

- Mức độ liên kết cực thấp: tất cả các chỉ số đều nằm trong khoảng 0.000 đến 0.034. Điều này cho thấy các biến định tính như Gender, Parental_Involvement, Internet_Access, hay School_Type gần như độc lập hoàn toàn với nhau.
- Cặp có liên kết cao nhất (tương đối): Extracurricular_Activities và Peer_Influence (0.034) hoặc Motivation_Level và Teacher_Quality (0.031). Thể hiện được trường tư có xu hướng tài nguyên tốt hơn trường công và giáo viên chất lượng cao làm cho học sinh có động lực học cao hơn



Hình 2.6 Ma trận tương quan giữa các đặc trưng rời rạc

2.2.5.4 Ma trận tương quan giữa các đặc trưng liên tục và rời rạc

- Các yếu tố ảnh hưởng mạnh nhất đến Điểm thi (Exam_Score):
  - Access_to_Resources (0.170): Đây là biến có tương quan cao nhất với điểm số. Điều này khẳng định rằng việc tiếp cận đầy đủ tài liệu và công cụ học tập là yếu tố then chốt quyết định thành tích.
  - Parental_Involvement (0.158): Sự quan tâm của phụ huynh đứng thứ hai về mức độ ảnh hưởng đến điểm thi. Kết hợp với biểu đồ phân phối ở Part 1 (đa số ở mức Medium), việc cải thiện sự tham gia của cha mẹ có thể là mục tiêu quan trọng để nâng cao kết quả.

- Parental_Education_Level (0.105) và Peer_Influence (0.101): Trình độ học vấn của bố mẹ và ảnh hưởng từ bạn bè cũng có tác động đáng kể (Eta > 0.1) đến điểm thi.

- Các biến có tầm ảnh hưởng thấp: một số biến dù có số lượng lớn nhưng lại ít có tác động trực tiếp đến các chỉ số liên tục:
  - Gender (0.002): Giới tính hầu như không có sự tương quan với điểm thi hay các thói quen học tập khác. Điều này cho thấy môi trường giáo dục này có sự bình đẳng về kết quả giữa nam và nữ.
  - School_Type (0.009): Đáng ngạc nhiên là việc học trường Công hay Tư không tạo ra sự khác biệt lớn về điểm số trong tập dữ liệu này.
- Tương quan với các thói quen học tập khác:
  - Sự tham gia của phụ huynh và Tài liệu: Không chỉ ảnh hưởng đến điểm thi, hai đặc trưng này còn có tương quan nhẹ với số giờ học (Hours_Studied) và điểm số trước đó (Previous_Scores).
  - Physical_Activity: đặc trưng này có tương quan thấp với hầu hết các yếu tố phân loại, ngoại trừ một chút ảnh hưởng từ thu nhập gia đình (Family_Income - 0.045) và nguồn lực (Access_to_Resource - 0.040)

Hình 2.7 Ma trận tương quan giữa các đặc trưng liên tục và rời rạc

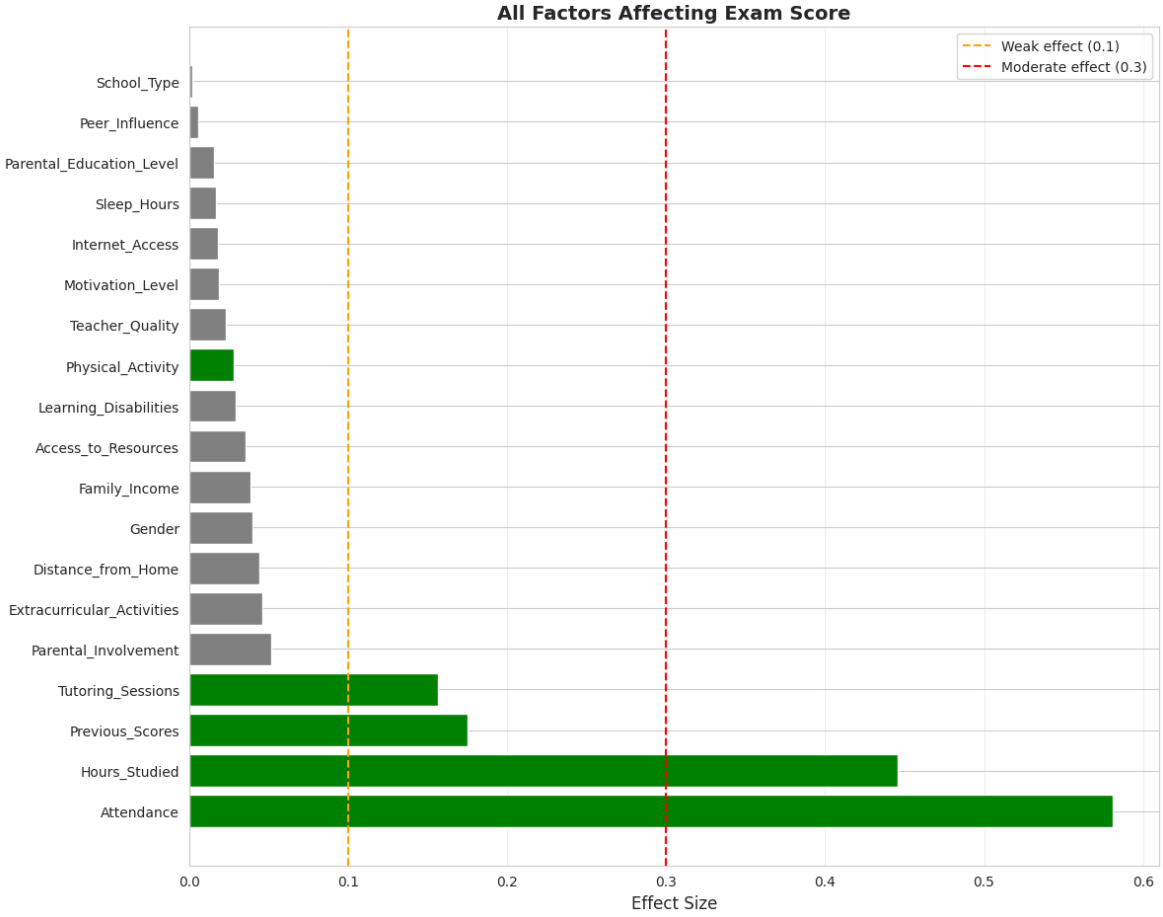
2.2.5.5 Sự tương quan giữa các đặc trưng đầu vào với mục tiêu

Biểu đồ ở Hình 2.8 cho thấy sự phân hóa rõ rệt thành 3 nhóm ảnh hưởng đến Exam_Score:

- Nhóm ảnh hưởng mạnh (Moderate effect > 0.3):
  - Attendance (0.58): Là yếu tố quyết định hàng đầu. Đi học đầy đủ có tác động mạnh mẽ nhất đến kết quả thi.
  - Hours_Studied (0.45): Xếp thứ hai, cho thấy nỗ lực cá nhân về mặt thời gian mang lại hiệu quả trực tiếp.
- Nhóm ảnh hưởng yếu (Weak Effect $0.1 - 0.2$):
  - Previous_Scores (0.17) và Tutoring_Sessions (0.14): Hai yếu tố này có tác động nhưng không quá lớn như kỳ vọng. Điều này cho thấy kết

quả quá khứ hoặc việc học thêm không quan trọng bằng sự chuyên cần cá nhân trong học tập.

- Nhóm ảnh hưởng không đáng kể (< 0.1):
  - Các yếu tố như Parental_Involvement, Gender, Family_Income, Internet_Access, và Physical_Activity đều có chỉ số thấp.
  - Cho thấy được kết quả thi trong tập dữ liệu này mang tính công bằng cao, không bị phụ thuộc vào điều kiện khác.



Hình 2.8 Biểu đồ hiển thị mức độ ảnh hưởng (tương quan) đến kết quả đầu ra của tập dữ liệu

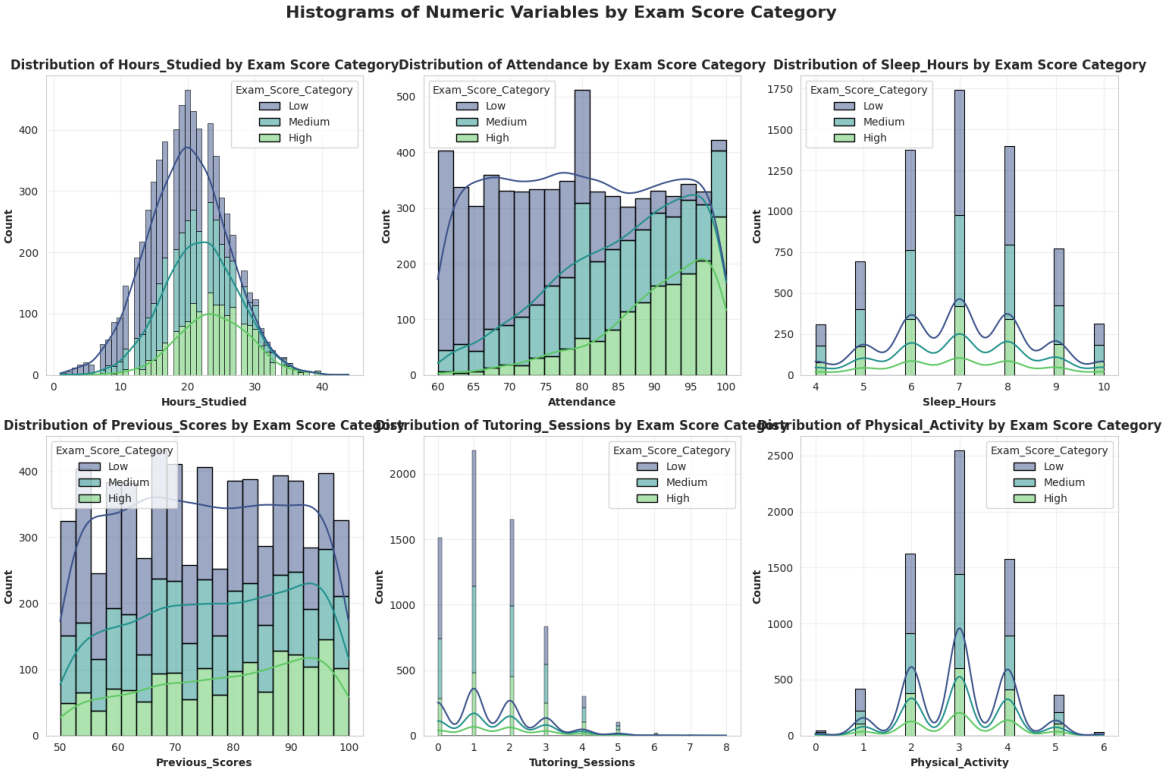
2.2.6 Phân tích phân phối theo nhóm ảnh hưởng đến điểm thi

- Nhóm nhân tố tác động mạnh:
    - Attendance:
        - Nhóm high (Xanh lá): Có sự tập trung cực kỳ cao ở dải 90% - 100%. Học sinh đi học gần như đầy đủ mới có cơ hội cao lọt vào nhóm điểm giỏi.
        - Nhóm Low (Xanh dương đậm): Phân bổ rộng nhưng chiếm ưu thế ở dải 60% - 75%. Khi tỷ lệ đi học giảm, mật độ nhóm điểm thấp tăng lên rõ rệt.
        - Đường mật độ (Density Curve): Đường màu xanh lá dốc đứng về phía bên phải, khẳng định Attendance là biến phân loại (classifier) tốt nhất cho mô hình.
    - Hours_Studied:
        - Sự dịch chuyển đỉnh: Đỉnh phân phối của nhóm Low nằm ở mức 15-20 giờ, trong khi nhóm High dịch chuyển hẳn sang mức 25-30 giờ.
        - Vùng giao thoa: Ở mức trung bình (20 giờ), cả 3 nhóm điểm đều xuất hiện, cho thấy nếu chỉ học ở mức trung bình thì kết quả sẽ phụ thuộc thêm vào các yếu tố khác.
        - Cận trên: Những học sinh học trên 35 giờ hầu hết đều thuộc nhóm Medium hoặc High, rất ít học sinh nhóm Low duy trì được cường độ này.
- Nhóm nhân tố nền tảng và hỗ trợ:
    - Previous_Scores:
        - Tính kế thừa: Nhóm điểm High hiện tại thường có nền tảng điểm cũ tốt (phân bổ nhiều ở dải 80-100).
        - Sự bứt phá: có một lượng học sinh nhóm High có điểm cũ chỉ từ 60-70.

- Đường mật độ: Các đường mật độ của 3 nhóm khá song song và không tách biệt quá xa.

o Tutoring_Sessions:

- Điểm bão hòa: Đa số học sinh không học thêm hoặc học rất ít (0-2 buổi) bất kể nhóm điểm nào.
- Hiệu quả biên: Nhóm High có mật độ cao hơn một chút ở các cột từ 3-5 buổi so với nhóm Low. Tuy nhiên, từ 6 buổi trở lên, số lượng dữ liệu rất thưa thớt (outliers), cho thấy học thêm quá nhiều không phải là xu hướng chung của nhóm dẫn đầu.



Hình 2.9 Biểu đồ phân tích phân phối theo nhóm ảnh hưởng đến điểm thi

2.2.7 Xác định các yếu tố ảnh hưởng đến điểm thi

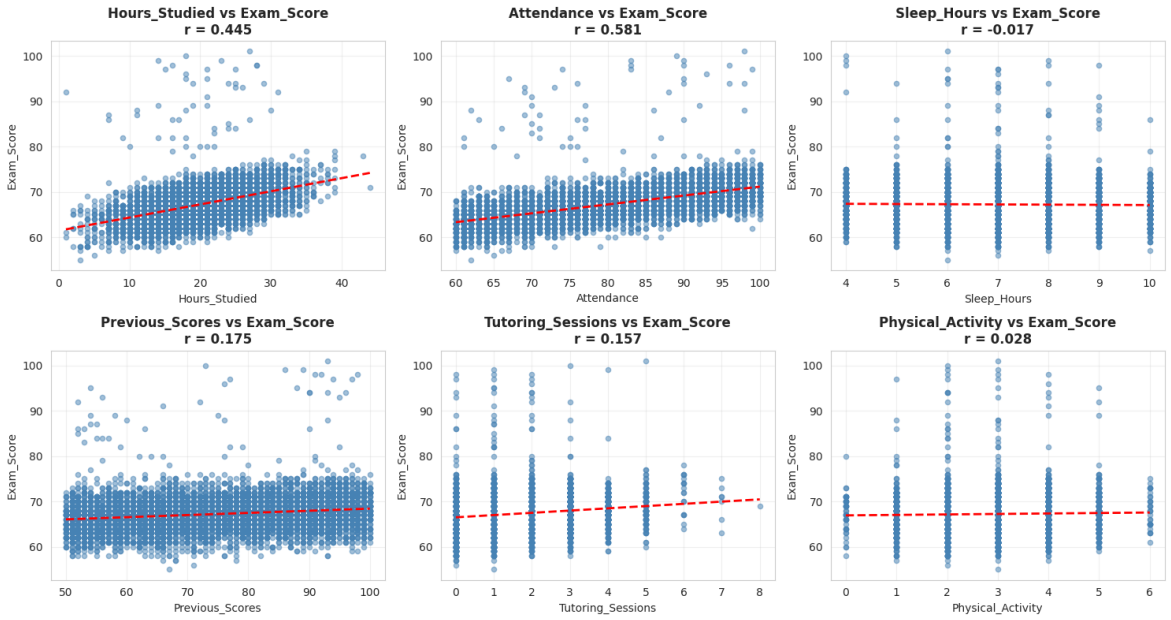
– Nhóm ảnh hưởng mạnh nhất:

o Attendance ($r = 0.581$): Đây là yếu tố có mối tương quan thuận mạnh nhất với điểm thi. Biểu đồ cho thấy khi tỷ lệ chuyên cần tăng lên, điểm

số có xu hướng tăng rõ rệt. Điều này chỉ ra rằng việc có mặt trên lớp là yếu tố tiên quyết để đạt kết quả cao.

- Hours_Studied ($r = 0.445$): Có tương quan thuận ở mức trung bình khá. Đường hồi quy (màu đỏ) dốc lên khá rõ. Dữ liệu có độ phân tán rộng, có những học sinh học ít nhưng điểm vẫn cao và ngược lại, cho thấy hiệu quả học tập cũng quan trọng không kém.

- Nhóm ảnh hưởng yếu:
  - Previous_Scores ($r = 0.175$): Tương quan thuận nhưng rất yếu. Khá bất ngờ vì thông thường năng lực học tập trong quá khứ phản ánh kết quả tương lai. Trong tập dữ liệu, điểm số cũ không đảm bảo chắc chắn cho điểm số mới.
  - Tutoring_Sessions ($r = 0.157$): Tương quan rất thấp. Việc đi học thêm có tác động tích cực nhưng không mang tính quyết định đến điểm số cuối cùng.
- Nhóm không có tương quan (Trung tính):
  - Physical_Activity ($r = 0.028$): Hệ số gần bằng 0. Hoạt động thể chất gần như không ảnh hưởng trực tiếp đến điểm thi.
  - Sleep_Hours ($r = -0.017$): Hệ số âm cực nhỏ, coi như không có tương quan. Biểu đồ cho thấy dù ngủ 4 tiếng hay 10 tiếng, dải điểm số vẫn trải dài tương tự nhau.



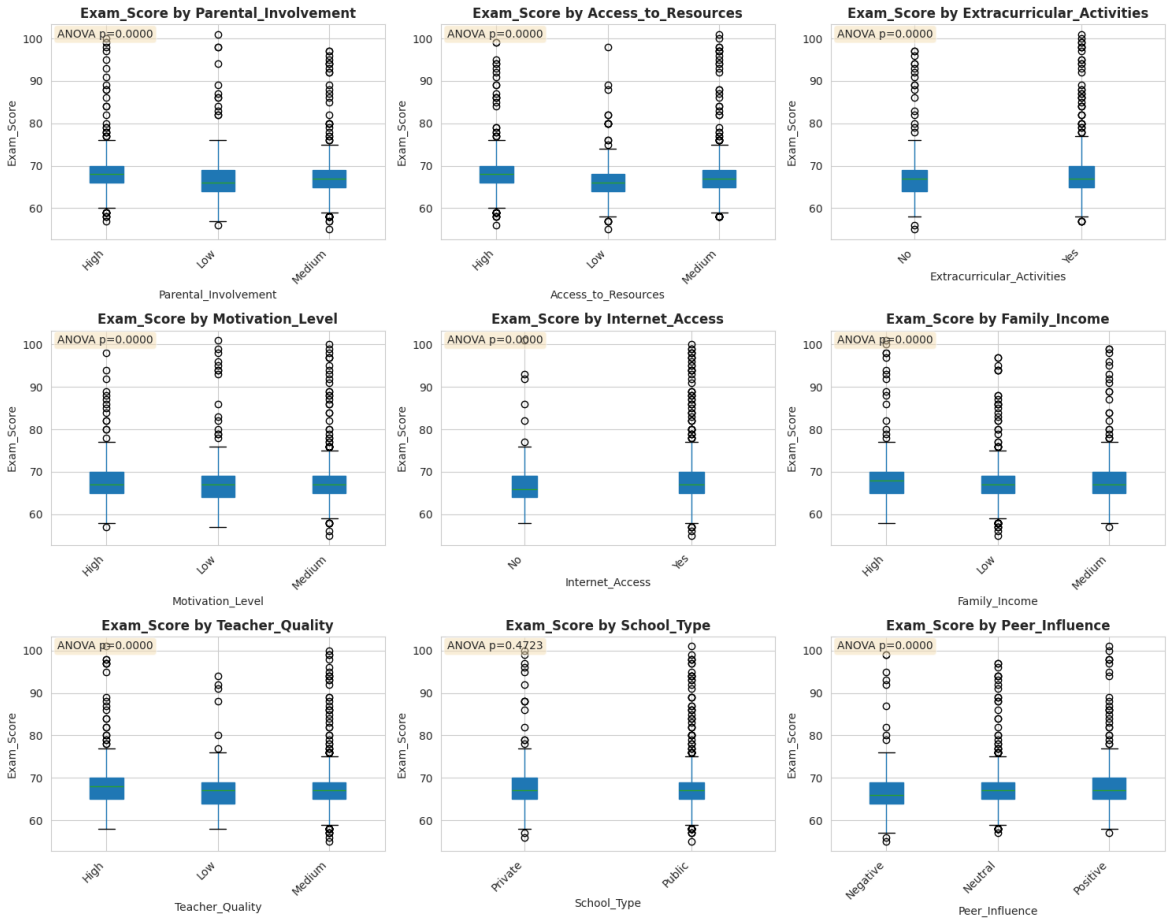
Hình 2.10 Biểu đồ tập trung xác định các yếu tố ảnh hưởng đến điểm thi

2.2.8 Phân tích giá trị Thống kê (ANOVA p-value)

- Hầu hết các biến (p = 0.0000): Có ý nghĩa thống kê cực kỳ lớn. Điều này khẳng định rằng sự khác biệt về điểm số giữa các nhóm (như High/Low/Medium) không phải do ngẫu nhiên mà thực sự bị ảnh hưởng bởi các yếu tố này.
- Đặc trưng duy nhất không có ý nghĩa (School_Type, p = 0.4723): Giá trị p > 0.05 cho thấy không có sự khác biệt đáng kể về điểm số giữa học sinh trường công (Public) và trường tư (Private).
- Nhóm ảnh hưởng tích cực rõ rệt:
    - Parental_Involvement và Teacher_Quality: Khi mức độ tham gia của phụ huynh hoặc chất lượng giáo viên ở mức High, trung vị (đường vạch xanh) của điểm số cao hơn rõ rệt so với mức Low.
    - Access_to_Resources: Những học sinh có khả năng tiếp cận nguồn lực học tập cao có phân phối điểm số tập trung ở mức cao hơn so với nhóm Low.
    - Internet_Access: Những học sinh có truy cập Internet (Yes) có điểm số ổn định và trung vị cao hơn nhóm không có (No).
    - Nhóm có sự khác biệt nhưng biên độ hẹp:

- Motivation_Level và Family_Income: Mặc dù p = 0.0000 (có sự khác biệt), nhưng khi nhìn vào hình dáng các hộp, sự chênh lệch về điểm số trung bình giữa mức High và Medium không quá lớn. Tuy nhiên, mức Low vẫn cho thấy điểm số thấp hơn hẳn.
- Peer_Influence: Ảnh hưởng tích cực từ bạn bè (Positive) giúp kéo điểm trung vị lên cao hơn so với ảnh hưởng tiêu cực (Negative).

– Nhóm biến không tác động (School_Type): Biểu đồ School_Type cho thấy hai hộp Private và Public gần như nằm ngang hàng nhau. Điều này cho thấy trong tập dữ liệu, môi trường trường học (công hay tư) không phải là yếu tố quyết định sự thành công của học sinh.



Hình 2.11 Đồ thị phân tích giá trị thống kê (ANOVA p-value) của các đặc trưng liên tục

CHƯƠNG 3. CHỌN LỌC MÔ HÌNH

3.1 Lý do thực hiện chọn lọc mô hình

- Trong quá trình giải quyết bài toán, việc chọn một mô hình phù hợp so với yêu cầu bài toán luôn là một điều cần thiết.
- Trong quá trình thảo luận về các lựa chọn về mô hình, nhóm đã có sự phân chia giữa việc chọn mô hình tuyến tính hoặc mô hình cây.
- Để giải quyết xung đột này, nhóm đã quyết định so sánh hiệu suấ của tất cả các mô hình ứng cử.

3.2 Danh sách mô hình ứng cử

3.2.1 Mô hình tuyến tính

3.2.1.1 Linear Regression (Hồi quy tuyến tính)

- Là mô hình hồi quy giải thích hành vi trung bình của biến phụ thuộc (Y) theo các biến giải thích (X) hay nói cách khác nó sẽ mô hình hóa mối quan hệ giữa biến đầu vào và biến đầu ra bằng một hàm tuyến tính
- Ưu điểm: Dễ triển khai và huấn luyện nhanh, tốn ít tài nguyên.
- Nhược điểm: Tuy nhiên dễ xảy ra tình trạng Overfitting và hiệu suất huấn luyện khá kém.

3.2.1.2 Ridge

- Là giải pháp kỹ thuật nhằm khắc phục hạn chế của hồi quy tuyến tính thường khi đối mặt với hiện tượng đa cộng tuyến. Cơ chế cốt lõi của phương pháp này là bổ sung một thành phần điều chỉnh dựa trên chuẩn L2 vào hàm tối ưu hóa.
- Hàm mất mát của Hồi quy Ridge được định nghĩa như sau:

$$J(\theta) = \sum_{i=1}^{m} (y_i - \hat{y_i})^2 - \lambda \sum_{j=1}^{n} (\theta_j^2)$$

Trong biểu thức trên, đại lượng $\lambda \sum_{j=1}^{n} (\theta_j^2)$ đóng vai trò kìm hãm sự gia tăng quá mức của các hệ số hồi quy. Trong bối cảnh dự báo điểm số, khi các môn

học có sự tương quan mạnh (ví dụ Toán và Lý), Ridge sẽ phân phối trọng số một cách hài hòa giữa các biến thay vì tập trung vào một biến duy nhất, giúp mô hình hoạt động ổn định hơn trước những biến động nhỏ của dữ liệu đầu vào.

- Ưu điểm: Lợi thế lớn nhất của Ridge là khả năng cân bằng giữa độ lệch (bias) và phương sai (variance). Bằng cách chấp nhận một lượng nhỏ độ lệch, mô hình giảm thiểu đáng kể phương sai, từ đó ngăn chặn hiệu quả hiện tượng quá khớp (overfitting), đặc biệt trong các trường hợp dữ liệu chứa nhiều nhiễu.
- Nhược điểm: Hạn chế căn bản của Ridge là không thực hiện được việc chọn lọc biến. Các hệ số hồi quy chỉ bị co về giá trị rất nhỏ tiệm cận 0 chứ không bao giờ bị triệt tiêu hoàn toàn. Điều này khiến mô hình cuối cùng vẫn bao gồm tất cả các biến đầu vào, gây khó khăn cho việc thông giải kết quả và làm tăng chi phí tính toán nếu số lượng đặc trưng quá lớn.

3.2.1.3 Lasso (Least Absolute Shrinkage Selection Operator)

- Tương tự như Ridge, Hồi quy Lasso cũng là một phương pháp hồi quy nhưng sử dụng chuẩn L1 làm thành phần điều chỉnh. Đặc tính toán học của chuẩn L1 cho phép Lasso có khả năng ép các hệ số của những biến ít quan trọng về đúng bằng 0.
- Hàm mục tiêu cần tối thiểu hóa của Lasso là:

$$J(\theta) = \sum_{i=1}^{m} (y_i - \hat{y_i})^2 - \lambda \sum_{j=1}^{n} |\theta_j|$$

- Thành phần phạt $\lambda \sum_{j=1}^{n} |\theta_j|$ cung cấp cho Lasso khả năng chọn lọc đặc trưng tự động. Trong bài toán này, với hàng loạt yếu tố ảnh hưởng tiềm năng, Lasso đóng vai trò như một bộ lọc, loại bỏ triệt để các biến nhiễu hoặc ít tác động, giúp cô lập được các nhân tố thực sự chi phối kết quả học tập của sinh viên.
- Ưu điểm: Tính năng sparsity là ưu điểm vượt trội của Lasso. Kết quả đầu ra là một mô hình tinh gọn, chỉ chứa các biến quan trọng nhất, giúp việc giải thích

ý nghĩa mô hình trở nên trực quan và dễ hiểu hơn nhiều so với các phương pháp khác.

- Nhược điểm: Tuy nhiên, Lasso bộc lộ điểm yếu khi xử lý các biến có tính tương quan cao. Trong nhóm các biến tương quan chặt chẽ, thuật toán có xu hướng chọn ngẫu nhiên một biến và loại bỏ các biến còn lại. Hành vi này có thể dẫn đến việc mất mát thông tin và khiến kết quả lựa chọn biến trở nên thiếu ổn định khi dữ liệu thay đổi.

3.2.2 Mô hình dạng cây

3.2.2.1 Random Forest

- Khái niệm: Random Forest là thuật toán tiêu biểu của phương pháp Ensemble Learning, hoạt động bằng cách kiến tạo một tập hợp lớn các Decision Trees riêng biệt. Kết quả cuối cùng được tổng hợp từ sự đồng thuận của toàn bộ các cây thành phần thông qua cơ chế biểu quyết hoặc lấy trung bình.
- Giá trị dự báo của Random Forest $\hat{y}$ được tính toán dựa trên trung bình kết quả của K cây quyết định $h_k(x)$:

$$\hat{y} = \frac{1}{K}\sum_{k=1}^{K} h_k(x)$$

- Đối với dữ liệu giáo dục, mối liên hệ giữa các yếu tố đầu vào và kết quả học tập thường rất phức tạp và phi tuyến tính. Random Forest, với cấu trúc tổ hợp của mình, có khả năng mô hình hóa các đường biên quyết định phức tạp này tốt hơn hẳn so với các mô hình tuyến tính đơn giản, từ đó nâng cao độ chính xác của dự báo.
- Ưu điểm: Random Forest nổi bật với độ chính xác cao và khả năng chịu lỗi tốt. Thuật toán này rất mạnh mẽ trong việc xử lý dữ liệu bị khuyết thiếu, dữ liệu nhiễu và không yêu cầu các giả định khắt khe về phân phối dữ liệu. Nó cũng giảm thiểu đáng kể nguy cơ quá khớp so với việc sử dụng một cây quyết định đơn lẻ.

- Nhược điểm: Nhược điểm chính của mô hình này là sự phức tạp trong cấu trúc dẫn đến thời gian huấn luyện và dự báo lâu hơn. Ngoài ra, tính chất "hộp đen" (black-box) của mô hình tổ hợp khiến việc giải thích tường tận cơ chế ra quyết định bên trong trở nên khó khăn hơn so với các mô hình có cấu trúc tường minh như hồi quy tuyến tính hay cây quyết định đơn.

3.2.2.2 XGB Regressor

- Khái niệm: XGBoost là một thuật toán học máy dựa trên cây quyết định (Decision Tree) sử dụng khung Gradient Boosting. Thuật toán được thiết kế để tối ưu hóa về tốc độ lẫn hiệu suất nhờ vào việc tận dụng tối đa tài nguyên hệ thống và các cải tiến về mặt toán học. Trong các thuật toán Boosting cũ, chỉ sử dụng đạo hàm bậc một (Gradient). XGBoost sử dụng khai triển Taylor bậc hai để xấp xỉ hàm mất mát.
- Nguyên lý hoạt động của thuật toán: XGBoost xây dựng các cây quyết định một cách tuần tự. Mỗi cây mới sẽ cố gắng dự báo sai số (residual) của các cây đứng trước nó. Dự báo cuối cùng là tổng trọng số kết quả của tất cả các cây.

$$Obj = \sum_{i}^{n} L(\hat{y}_i, y_i) + \sum_{t}^{n} \Omega(f_t)$$

Trong đó:

  - L: hàm mất mát (Loss function) đo lường sự khác biệt giữa thực tế và dự báo.
  - $\Omega(f_t)$*:* Phạt các cây quyết định quá phức tạp để tránh hiện tượng overfitting

- Để giảm thiểu sự mất mát (loss), XGBoost sử dụng hai giá trị quan trọng từ giải tích:
  - Gradient g_i*:* Đạo hàm bậc nhất (độ dốc). Gradient cho biết hướng mà dự đoán cần điều chỉnh để làm giảm sai số. Đạo hàm bậc nhất của hàm mất mát theo giá trị dự đoán:

$$g_i = \frac{\partial L}{\partial \widehat{y_i}} = \widehat{y_i} - y_i$$

Ý nghĩa của gradient:

- $g_i > 0$ → mô hình dự đoán quá cao (overestimate)
- $g_i < 0$ → mô hình dự đoán quá thấp (underestimate)

- Hessian h_i: Đạo hàm bậc hai (độ cong). Hessian cho biết gradient thay đổi nhanh như thế nào, tức là nó phản ánh độ cong của hàm mất mát (loss). Nếu Gradient cho biết hướng đi thì Hessian cho biết nên đi mạnh hay đi nhẹ theo hướng đó. Hessian là đạo hàm bậc hai của hàm mất mát theo giá trị dự đoán:

$$h_i = \frac{\partial^2 L}{\partial \widehat{y_i^2}} = 1$$

Với hàm Mean Squared Error (MSE): Hessian luôn bằng 1, đây là một hằng số, không phụ thuộc vào dữ liệu.

– XGBoost không tối ưu trực tiếp hàm loss gốc. Thay vào đó, sử dụng khai triển Taylor bậc hai để biểu diễn hàm loss thành một hàm bậc hai đơn giản và hiệu quả hơn:

$$L_{approx} = \sum_i \left(g_i . w + \frac{1}{2} h_i . w^2 \right) + const$$

Với w là giá trị dự đoán của một lá (leaf)

– Ưu điểm:

- Hiệu năng: Sử dụng xấp xỉ Taylor bậc hai giúp hội tụ nhanh hơn và đạt độ chính xác cao hơn so với Gradient Boosting truyền thống.
- Khả năng điều tiết (Regularization): Tích hợp cả L1 (Lasso) và L2 (Ridge), giúp mô hình kiểm soát tốt trọng số của các lá cây, từ đó giảm thiểu hiện tượng Overfitting.
- Xử lý dữ liệu thưa (Sparse Data): Thuật toán tự động học cách xử lý các giá trị thiếu (Missing values) hoặc dữ liệu có nhiều số 0 mà không cần can thiệp thủ công.

- Nhược điểm:
    - Tốn tài nguyên bộ nhớ: Do cơ chế lưu trữ dữ liệu trong các block để phục vụ tính toán song song, nên thường tốn nhiều RAM khi làm việc với tập dữ liệu lớn.
    - Tiền xử lý dữ liệu: Không hỗ trợ trực tiếp biến định tính (Categorical features). Bắt buộc phải dùng One-Hot Encoding hoặc Label Encoding trước khi đưa vào mô hình.
    - Nhạy cảm với tham số: Có rất nhiều tham số (hyperparameters) cần tinh chỉnh. Nếu không biết cách "tune", mô hình rất dễ bị Overfitting hoặc không đạt hiệu suất tối ưu.

3.2.2.3 CatBoosting Regressor

- Khái niệm:
    - CatBoost là một thuật toán Gradient Boosting nổi bật khi làm việc với các tập dữ liệu có các đặc trưng phân loại. Thuật toán này hỗ trợ các biến phân loại, làm cho nó đặc biệt phù hợp cho các tác vụ như phân loại, hồi quy, và xếp hạng, nơi mà dữ liệu phân loại thường xuyên xuất hiện.
    - Khác với các thuật toán boosting khác yêu cầu phải mã hóa các biến phân loại dưới dạng các giá trị số (ví dụ, sử dụng mã hóa nhãn hay mã hóa one-hot), CatBoost xử lý dữ liệu phân loại thông qua một phương pháp mới gọi là Mã hóa Mục tiêu Có Trật tự (Ordered Target Encoding).
- Nguyên lý hoạt động:
    - CatBoost được xây dựng dựa trên nguyên lý của Gradient Boosting, nơi mỗi mô hình mới sẽ sửa chữa các lỗi của các mô hình trước đó. Tuy nhiên, nó được cải tiến để tăng hiệu suất, đặc biệt là khi làm việc với dữ liệu phân loại.

- Ordered target encoding: Một trong những cải tiến quan trọng của CatBoost là khả năng xử lý các đặc trưng phân loại mà không cần mã hóa rõ ràng. CatBoost sử dụng kỹ thuật là mã hóa mục tiêu có trật tự (Ordered target encoding), giúp ngăn ngừa hiện tượng rò rỉ mục tiêu (target leakage) một vấn đề phổ biến khi mã hóa các đặc trưng phân loại.

– Mã hóa mục tiêu có trật tự trong CatBoost:
  - CatBoost giải quyết vấn đề bằng cách sắp xếp các điểm dữ liệu và chỉ sử dụng các điểm dữ liệu trong quá khứ để mã hóa các đặc trưng phân loại. Cách làm này giúp mô hình chỉ nhìn vào các quan sát đã xảy ra trước đó trong quá trình huấn luyện, tránh hiện tượng rò rỉ mục tiêu.
  - Đối với một đặc trưng phân loại xi và giá trị mục tiêu tương ứng yi , mã hóa sẽ được tính toán dựa trên trung bình chạy (running mean) chỉ sử dụng các giá trị mục tiêu từ các điểm dữ liệu trước đó:

$$Encoding(x_i) = \frac{\sum_{j<i} y_j}{\sum_{j<i} 1}$$

– Xử lý các giá trị bị thiếu:
  - CatBoost được tích hợp sẵn cơ chế xử lý giá trị bị thiếu (missing values) cho cả đặc trưng số và đặc trưng phân loại. Đối với các biến số, CatBoost áp dụng phương pháp chia nhánh dựa trên giá trị thiếu, trong đó các giá trị bị thiếu được xem như một nhóm đặc biệt và thuật toán sẽ tự động quyết định cách phân tách tối ưu tại mỗi nút của cây.
  - Đối với các biến phân loại có giá trị bị thiếu, CatBoost xử lý tương tự như các giá trị phân loại thông thường bằng cách coi giá trị thiếu là một danh mục riêng biệt. Quá trình mã hóa vẫn tuân theo cơ chế có trật tự (ordered encoding), chỉ sử dụng các dữ liệu trong quá khứ để học cách xử lý các giá trị này. Nhờ đó, CatBoost có thể xử lý hiệu quả dữ liệu không đầy đủ mà không cần các bước tiền xử lý phức tạp, đồng thời hạn chế rò rỉ thông tin và overfitting.

- Hàm mục tiêu trong CatBoost, tương tự như các thuật toán Gradient Boosting khác, được xây dựng từ hai thành phần chính: hàm mất mát (loss function) và thành phần điều chuẩn (regularization).
- Ưu điểm:
    - Xử lý trực tiếp dữ liệu phân loại: CatBoost hỗ trợ xử lý các biến phân loại một cách tự nhiên mà không cần mã hóa thủ công, giúp giảm thời gian tiền xử lý và cải thiện hiệu suất mô hình.
    - Khả năng chống overfitting tốt: Nhờ áp dụng các kỹ thuật như boosting có trật tự và mã hóa mục tiêu có trật tự, CatBoost thể hiện khả năng chống overfitting cao, đặc biệt hiệu quả trên các tập dữ liệu nhỏ.
    - Tốc độ huấn luyện nhanh: Thuật toán được tối ưu hóa về hiệu năng và hỗ trợ huấn luyện trên cả CPU và GPU.
    - Hiệu suất cao với dữ liệu hỗn hợp: CatBoost hoạt động hiệu quả trên các tập dữ liệu chứa đồng thời cả đặc trưng số và đặc trưng phân loại.
- Nhược điểm:
    - Cần tinh chỉnh tham số cẩn thận: tương tự các thuật toán boosting khác, CatBoost yêu cầu lựa chọn và tinh chỉnh các siêu tham số một cách hợp lý để đạt hiệu quả tối ưu.
    - Tiêu tốn nhiều bộ nhớ: Đối với các tập dữ liệu rất lớn yêu cầu dung lượng bộ nhớ đáng kể.
    - Chưa tối ưu cho dữ liệu thuần số: Mặc dù vượt trội trong xử lý dữ liệu phân loại, CatBoost có thể không nhanh bằng XGBoost khi làm việc với các tập dữ liệu chỉ bao gồm các đặc trưng số.

3.3 Tiền xử lý dữ liệu trước khi chọn lọc mô hình

3.3.1 Standard Scaling

- Về mặt toán học, phương pháp này thực hiện phép biến đổi Z-score cho từng đặc trưng liên tục theo công thức:

$$Z = \frac{x - \mu}{\sigma}$$

Trong đó:

- μ là giá trị trung bình
- σ là độ lệch chuẩn của biến đó trên tập dữ liệu huấn luyện.

– Kết quả của quá trình này là các đặc trưng mới có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1, giúp các mô hình tối ưu hóa hàm mất mát hiệu quả hơn và không bị chi phối bởi các biến có giá trị lớn so với phương pháp MinMaxScaler.

3.3.2 Skewness và kurtosis transformation

Trong nghiên cứu này, quá trình xử lý các biến liên tục không áp dụng một công thức cố định mà sử dụng một bộ biến đổi tùy chỉnh có tên ***SkewKurtTransformer***. Bộ biến đổi này tự động đánh giá và lựa chọn phương pháp tối ưu giữa việc giữ nguyên dữ liệu gốc, biến đổi Logarit hoặc biến đổi Yeo-Johnson.

3.3.2.1 Biến đổi Logarit

– Phương pháp đầu tiên được cân nhắc là biến đổi Logarit, đặc biệt hữu hiệu cho dữ liệu bị lệch phải mạnh.

– Công thức tổng quát của phương pháp này là:

$$x' = ln(x + c)$$

Trong đó: c là hằng số dịch chuyển được thêm vào để đảm bảo giá trị đầu vào của hàm logarit luôn dương.

3.3.2.2 Biến đổi Yeo-Johnson

– Phép biến đổi Yeo-Johnson được tích hợp như một giải pháp phức tạp và linh hoạt hơn với khả năng xử lý cả giá trị âm và dương. Phép biến đổi này được định nghĩa bởi hệ phương trình:

$$x_i^{(\lambda)} = \begin{cases} \dfrac{(x_i+1)^{\lambda}-1}{\lambda} & \text{nếu } \lambda \neq 0, x_i \geq 0 \\ ln(x_i+1) & \text{nếu } \lambda = 0, x_i \geq 0 \\ -\dfrac{(-x_i+1)^{2-\lambda}-1}{2-\lambda} & \text{nếu } \lambda \neq 2, x_i < 0 \\ -ln(-x_i+1) & \text{nếu } \lambda = 2, x_i < 0 \end{cases}$$

- Trong đó, tham số đóng vai trò quan trọng và được ước lượng thông qua phương pháp hợp lý cực đại Maximum Likelihood Estimation nhằm tìm ra giá trị giúp dữ liệu đạt được tính chuẩn hóa tốt nhất.

3.3.2.3 Lựa chọn phương pháp biến đổi phù hợp

Quyết định chọn phương pháp biến đổi nào dựa trên nguyên tắc tối thiểu hóa một hàm mục tiêu, kết hợp giữa hai chỉ số độ lệch và độ nhọn. Ngưỡng chấp nhận đối với giá trị tuyệt đối của hệ số độ lệch Skewness được thiết lập ở mức 0.5; nghĩa là nếu giá trị tuyệt đối của độ lệch nhỏ hơn mức này, phân phối được xem là chấp nhận được và không cần can thiệp. Mục tiêu cuối cùng là tìm ra phương pháp biến đổi giúp đưa phân phối dữ liệu về dạng gần chuẩn nhất có thể.

3.3.3 Các phương pháp mã hoá dữ liệu

3.3.3.1 One-Hot Encoding

- Khái niệm: Kỹ thuật One-Hot Encoding thực hiện việc biến đổi một biến phân loại có n giá trị khác nhau thành n biến nhị phân độc lập. Tại đây, mỗi quan sát sẽ được biểu diễn bằng một vectơ, trong đó chỉ có một phần tử duy nhất mang giá trị 1 tương ứng với loại của quan sát đó và các phần tử còn lại đều bằng 0. Ví dụ, thuộc tính "Giới tính" sẽ được tách thành hai trường dữ liệu riêng biệt là "Nam" và "Nữ".
- Ưu điểm: Lợi điểm lớn nhất của phương pháp này là loại bỏ hoàn toàn việc áp đặt thứ tự giả tạo lên dữ liệu. Các thuật toán sẽ không thể diễn giải sai lệch rằng giá trị này ưu việt hơn giá trị kia, đảm bảo tính khách quan tuyệt đối cho

các biến danh nghĩa vốn không có cấu trúc thứ bậc, chẳng hạn như quê quán hay dân tộc.

- Nhược điểm: Mặc dù vậy, hạn chế đáng kể của One-Hot Encoding là sự gia tăng đột biến về số chiều của không gian dữ liệu. Đối với các biến có số lượng giá trị đơn nhất lớn, kỹ thuật này sẽ tạo ra một lượng lớn các cột mới, dẫn đến hiện tượng curse of dimensionality. Điều này không chỉ làm tăng chi phí tính toán mà còn tạo ra các ma trận thưa thớt, gây khó khăn cho quá trình hội tụ của mô hình.

3.3.3.2 Ordinal Encoding

- Khái niệm: Trái ngược với One-Hot, Ordinal Encoding chuyển đổi các giá trị phân loại thành các số nguyên theo một trật tự nhất định. Phương pháp này thường được ưu tiên cho các dữ liệu có tính chất xếp hạng. Điển hình như biến xếp loại hạnh kiểm có thể được gán các giá trị tăng dần: 1 cho Yếu, 2 cho Trung bình, 3 cho Khá và 4 cho Tốt.
- Ưu điểm: Thế mạnh của Ordinal Encoding nằm ở khả năng bảo toàn thông tin về thứ bậc tự nhiên của dữ liệu, giúp mô hình nắm bắt được xu hướng tăng giảm của biến số. Đồng thời, kỹ thuật này duy trì nguyên vẹn kích thước của bộ dữ liệu gốc, không phát sinh thêm các đặc trưng mới, từ đó tối ưu hóa hiệu năng tính toán.
- Nhược điểm: Tuy nhiên, việc gán các số nguyên liên tiếp vô tình tạo ra giả định rằng khoảng cách giữa các mức độ là đồng nhất. Trong thực tế, sự chênh lệch về năng lực giữa mức Yếu và Trung bình có thể không tương đồng với khoảng cách giữa Khá và Giỏi. Sự áp đặt toán học này có thể dẫn đến việc các mô hình hồi quy ước lượng sai trọng số, làm giảm độ chính xác của dự báo.

3.3.4 SparsePCA

- Khái niệm:
    - Principal Component Analysis (PCA) là một phương pháp phân tích dữ liệu đa biến cổ điển, nhằm giảm chiều dữ liệu bằng cách tìm các thành

phần chính (principal components) là các tổ hợp tuyến tính của các biến đầu vào, sao cho chúng giải thích được phần lớn phương sai trong dữ liệu. Tuy nhiên, trong PCA truyền thống, các thành phần chính thường liên quan đến tất cả các biến, dẫn đến khó diễn giải, đặc biệt trong dữ liệu chiều cao (high-dimensional data).

- Sparse PCA (SPCA) là phần mở rộng của PCA, giới thiệu tính thưa thớt (sparsity) vào các vector loadings, nghĩa là nhiều hệ số trong tổ hợp tuyến tính bằng 0. Điều này giúp các thành phần chính chỉ phụ thuộc vào một tập con nhỏ các biến, tăng tính diễn giải và khả năng áp dụng thực tế.

– Nguyên lý hoạt động: SparsePCA giải quyết bài toán tối ưu hóa bằng cách thêm các thành phần phạt (penalty terms) vào hàm mục tiêu của PCA truyền thống.

- Cho ma trận dữ liệu $X \in R^{n \times p}$ (đã trung tâm hóa) và hiệp phương sai $\hat{\Sigma} = \frac{1}{n-1} X^T X$.
- Trong sparsePCA, thêm phạt L1 (Lasso) để tạo ra tính thưa thớt:

$$\min_{A,B} \sum_{i=1}^{n} |x_i - AB^T x_i|^2 + \lambda \sum_{j=1}^{k} |\beta_j|^2 + \sum_{j=1}^{k} \lambda_{1,j} |\beta_j|_1 \quad \text{s.t.} \quad A^T A = I_k$$

Trong đó:

- A: Ma trận hướng orthogonal;
- B: Ma trận loadings thưa thớt.
- Hồi quy: $\min_{A,B} \sum_{i=1}^{n} |x_i - AB^T x_i|^2$: đo sai lệch tái tạo dữ liệu.
- Phạt Ridge (L2): $\lambda \sum_{j=1}^{k} |\beta_j|^2$: ổn định mô hình với siêu tham số.
- Phạt Lasso (L1): $\sum_{j=1}^{k} \lambda_{1,j} |\beta_j|_1$: tạo thưa thớt với siêu tham số riêng $\lambda_{1,j}$
- Ràng buộc: $A^T A = I_k$: đảm bảo tính trực giao.
- k: số thành phần
- I_k: ma trận đơn vị.

- Ưu điểm:
    - Tăng tính diễn giải: Loadings thưa thớt giúp dễ xác định biến quan trọng, hữu ích trong lĩnh vực như sinh học hoặc tài chính.
    - Xử lý dữ liệu chiều cao tốt: Duy trì tính nhất quán khi $p >> n$, PCA truyền thống có thể thất bại.
    - Giảm nhiễu và chi phí: Tập trung vào ít biến hơn, giảm nhiễu và chi phí thực tế.
    - Linh hoạt: Có nhiều biến thể phù hợp với dữ liệu lớn.
- Nhược điểm:
    - Tối ưu hóa khó khăn: Vấn đề NP-hard, chỉ đạt giải xấp xỉ, có thể không phải tối ưu toàn cục.
    - Phụ thuộc siêu tham số: Cần điều chỉnh λ, $\lambda_{1,j}$, k qua cross-validation, tốn thời gian và tài nguyên tính toán.
    - Có thể mất thông tin: Nếu mức thưa thớt quá cao, có nguy cơ bỏ qua biến quan trọng, dẫn đến mất phương sai.
    - Tính toán phức tạp hơn: So với PCA thông thường, thuật toán đòi hỏi nhiều tài nguyên hơn, đặc biệt với dữ liệu rất lớn.

3.4 Thực nghiệm và kết quả chọn lọc mô hình

3.4.1 Thực nghiệm

3.4.1.1 K-Fold cross-validation

- Khái niệm:
    - K-fold cross-validation là phương pháp chia tập dữ liệu gốc thành k phần (folds) bằng nhau (hoặc gần bằng nhau). Mỗi fold sẽ lần lượt được sử dụng làm tập kiểm tra (validation set), trong khi các fold còn lại được dùng để huấn luyện mô hình. Quá trình này lặp lại k lần, và kết quả cuối cùng là trung bình của các chỉ số đánh giá từ k lần thử nghiệm.
    - K-fold cross-validation là một kỹ thuật phổ biến trong thống kê để đánh giá hiệu suất của mô hình dự đoán. Phương pháp này giúp giảm thiểu

rủi ro overfitting (quá khớp) và underfitting (không khớp), đồng thời cung cấp ước lượng ổn định hơn về khả năng tổng quát hóa của mô hình trên dữ liệu mới.

- Nguyên lý hoạt động:
  - Tập dữ liệu gốc (gồm N mẫu) được chia thành k fold không chồng chéo, mỗi fold có kích thước khoảng $\frac{N}{k}$.
  - Lặp lại k lần:
    - Chọn fold thứ i ($i = 1$ đến k) làm tập kiểm tra.
    - Sử dụng $k - 1$ fold còn lại làm tập huấn luyện.
    - Huấn luyện mô hình trên tập huấn luyện.
    - Đánh giá mô hình trên tập kiểm tra, tính toán các chỉ số như accuracy, precision, F1-score, hoặc mean squared error (MSE) tùy theo bài toán phân loại hay hồi quy.
  - Tính trung bình cộng của các chỉ số đánh giá từ k lần thử nghiệm:

$$Score_{average} = \frac{1}{k}\sum_{i=1}^{k} Score_i$$

- Ưu điểm:
  - Sử dụng tối đa dữ liệu: Mỗi mẫu dữ liệu đều được sử dụng để huấn luyện $k - 1$ lần và được sử dụng để kiểm tra đúng 1 lần. Điều này rất hữu ích khi tập dữ liệu nhỏ.
  - Giảm thiểu tính ngẫu nhiên: Đánh giá mô hình trên nhiều tập con khác nhau giúp giảm bớt sự sai lệch (bias) so với việc chỉ chia Train/Test một lần duy nhất (Hold-out method).
  - Đánh giá độ ổn định: Nếu kết quả giữa các fold chênh lệch nhau quá lớn, đó là dấu hiệu cho thấy mô hình đang bị Overfitting hoặc dữ liệu có quá nhiều nhiễu.
- Nhược điểm:

- Chi phí tính toán cao: Vì phải huấn luyện mô hình k lần, thời gian thực hiện sẽ tăng gấp k lần so với phương pháp thông thường.
- Không phù hợp với dữ liệu chuỗi thời gian (Time series): Đối với dữ liệu thời gian, việc xáo trộn ngẫu nhiên sẽ làm hỏng tính tuần tự.

3.4.1.2 Tiêu chí đánh giá

– R^2:

- Chỉ số này đại diện cho tỷ lệ phương sai của biến mục tiêu có thể được giải thích bởi các biến độc lập trong mô hình. Công thức được xác định là:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}_i\right)^2}$$

Trong đó:

- y_i là điểm thực tế
- $\hat{y}_i$ là điểm dự báo
- $\overline{y}_i$ là điểm trung bình.

- Trong bối cảnh dự báo điểm số, **R2** càng gần 1 chứng tỏ mô hình càng giải thích tốt sự biến động của điểm số dựa trên các yếu tố đầu vào như thời gian học hay chuyên cần.

– MAE:

- Đo lường độ lớn trung bình của các sai số dự báo mà không quan tâm đến chiều hướng âm hay dương.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Chỉ số MAE mang lại cái nhìn trực quan và dễ hiểu, nó cho biết trung bình mô hình dự đoán sai lệch bao nhiêu điểm so với thực tế. Ví dụ, MAE bằng 2.5 nghĩa là dự báo của mô hình thường chênh lệch khoảng 2.5 điểm so với điểm thi thật.

– RMSE:

- Tương tự như MAE nhưng RMSE bình phương sai số trước khi tính trung bình, do đó nó trừng phạt nặng hơn các sai số lớn.

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

- Trong bài toán này, RMSE đặc biệt hữu ích để phát hiện các trường hợp mô hình dự báo có sai số rất lớn đối với một số sinh viên cá biệt. Một giá trị RMSE thấp cho thấy mô hình không chỉ dự báo đúng trung bình mà còn ổn định, ít khi đưa ra các dự báo quá xa rời thực tế.

3.4.1.3 Thực hiện tiền xử lý dữ liệu với từng mô hình khác nhau

Đối với các loại mô hình có cấu trúc khác nhau, việc thực hiện bước tiền xử lý dữ liệu cũng sẽ cần phải khác nhau.

- Đối với các mô hình tuyến tính:
  - Những mô hình này có độ nhạy cảm với các giá trị liên tục cao. Cho nên, việc áp dụng StandardScaler và Skew/Kurt transform là việc rất cần thiết.
  - Ngoài ra, những mô hình này không nhận vào các giá trị rời rạc, bắt buộc phải mã hóa các giá trị rời rạc bằng One-hot encode và Ordinal encode.
  - SparsePCA được thiết kế tối ưu cho dữ liệu liên tục và phát huy hiệu quả cao nhất khi tồn tại mối tương quan tuyến tính mạnh giữa các đặc trưng. Nên với các mô hình tuyến tính, việc xử lý như sparsePCA sẽ giúp hiệu quả của mô hình.
- Ngược lại, đối với các mô hình dạng cây quyết định như Random Forest hay XGBoost:
  - Nghiên cứu thiết lập quy trình xử lý để bỏ qua bước chuẩn hóa này. Lý do là vì cấu trúc cây quyết định hoạt động dựa trên các quy tắc phân

chia giá trị tại các nút lá và không bị ảnh hưởng bởi sự chênh lệch về đơn vị đo hay độ lớn của biến số.

- Việc giữ nguyên dữ liệu gốc cho các mô hình cây giúp bảo toàn tính trực quan và khả năng giải thích của dữ liệu mà không làm giảm độ chính xác dự báo.
- Tuy nhiên, cả hai mô hình này đều không có bộ mã hóa riêng, việc mã hóa các giá trị rời rạc sẽ là quy trình tiền xử lý bắt buộc.
- Với sparsePCA, vì sparsePCA phù hợp hơn với loại dữ liệu liên tục, điều này không phù hợp với quy trình của các mô hình dạng cây.

– Cuối cùng, với mô hình CatBoost, do có bộ mã hóa riêng và là mô hình dạng cây, việc chuyển đổi cho cả giá trị liên tục, mã hóa giá trị rời rạc lẫn áp dụng sparsePCA đều không cần thiết.

Mô hình	Scaling	Skew/Kurt transform	Encoding	SparsePCA
Linear Regression	Có	Có	Có	Có
Lasso	Có	Có	Có	Có
Ridge	Có	Có	Có	Có
Random Forest Regressor	Không	Không	Có	Không
XGBoost Regressor	Không	Không	Có	Không
CatBoost Regressor	Không	Không	Không	Không

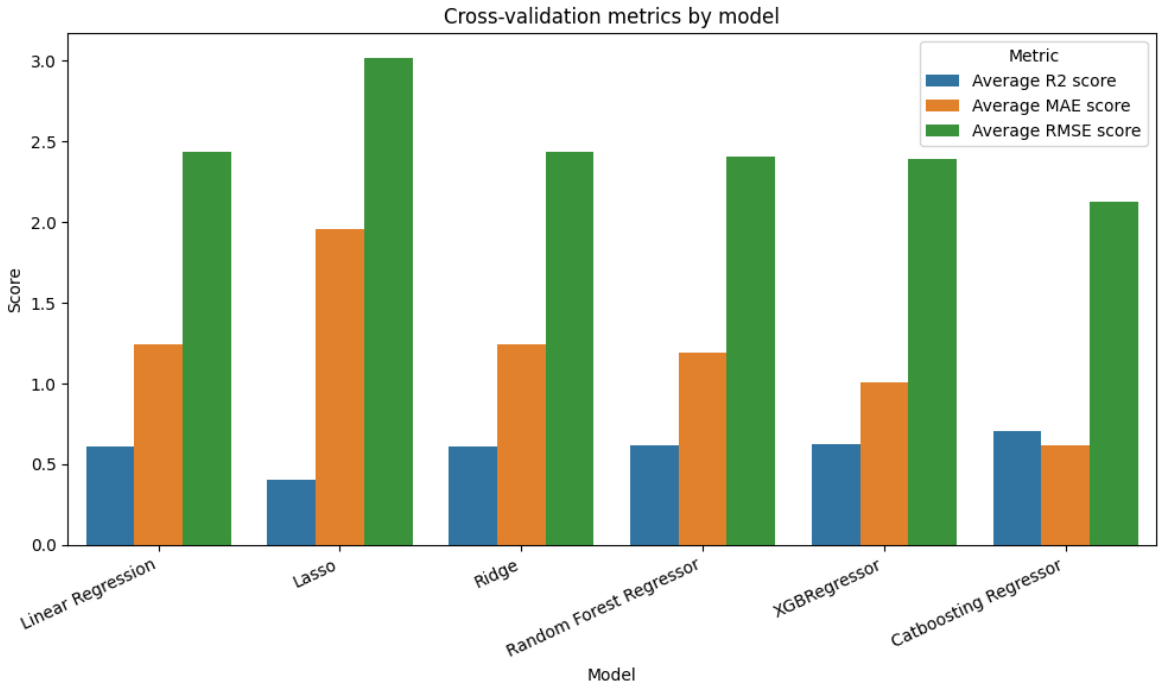
Bảng.3.1 Bảng tóm tắt bước tiền xử lý dữ liệu với từng mô hình

3.4.2 Kết quả chọn lọc mô hình

– Sau khi thực hiện KFold cross-validation với $k = 5$. Nhóm đã thu hoạch được đồ thị ở Hình 3.1.

– Lý do nhóm chọn giá trị $k = 5$ vì nó mang lại sự cân bằng giữa việc tính toán nhanh và sự ổn định trong đánh giá mô hình.

– Có thể thấy rõ, mô hình CatBoost Regressor có cả 3 tiêu chí vượt trội hơn các mô hình còn lại. Nhóm quyết định lựa chọn mô hình CatBoost Regressor cho việc giải quyết bài toán của để tài.



Hình 3.1 Kết quả chọn lọc mô hình dựa theo 3 chỉ tiêu

CHƯƠNG 4. HUẤN LUYỆN MÔ HÌNH

4.1 Các phương pháp hỗ trợ huấn luyện mô hình

4.1.1 Tối ưu siêu tham số

4.1.1.1 Giới thiệu

- Trong việc huấn luyện mô hình, có hai loại tham số: tham số mô hình và siêu tham số.
- Những tham số này nhiều hoặc ít đều đóng góp cho độ chính xác, tốc độ huấn luyện và mức độ cải thiện mất mát của mô hình. Nên việc tối ưu hóa chúng cũng quan trọng như tối ưu hóa các tham số mô hình.
- Trong các phương pháp tối ưu siêu tham số, bao gồm: Tìm kiếm theo lưới (Grid search), tìm kiếm ngẫu nhiên (Random search),… Nhưng trong số đó, có một phương pháp phức tạp hơn và bảo đảm về việc tối ưu hóa chính là phương pháp tối ưu Bayesian (Bayesian Optimization).

4.1.1.2 Phương pháp tối ưu Bayesian

- Phương pháp tối ưu này được dựa theo hướng tư duy của Bayes.
    - Thử nghiệm với các siêu tham số và cho ra kết quả.
    - Dựa theo kết quả để trích xuất ra thông tin.
    - Từ thông tin truy xuất ra được, tìm ra hướng đi đúng cho lần thử nghiệm sau.
- Khác với các phương pháp như tìm theo lưới hoặc tìm ngẫu nhiên, phương pháp tối ưu Bayesian này tối ưu dựa theo thông tin đã học được từ lần thử nghiệm trước đó, cho nên sự tối ưu hóa này sẽ dần đi đến cực tối ưu cục bộ (local optimum).

4.1.2 Phương pháp Dừng sớm (Early Stopping)

- Phương pháp dừng sớm (Early stopping) là phương pháp dừng quá trình huấn luyện trước khi mô hình bắt đầu học quá kỹ đặc điểm riêng của tập huấn luyện, tức là trước khi xảy ra hiện tượng overfitting.
- Phương pháp dừng sớm hoạt động bằng cách liên tục theo dõi hiệu suất của một hình trên tập kiểm thử (validation set) trong suất quá trình huấn luyện. Khi hiệu suất trên tập kiểm thử không còn cải thiện sau một số lượng chu kỳ huấn luyện (epoch) nhất định, quá trình huấn luyện sẽ được dừng lại. Mô hình tốt nhất, tức là mô hình có hiệu suất cao nhất trên tập kiểm thử, sẽ được lưu lại và sử dụng.

4.2 Huấn luyện mô hình

4.2.1 Chia tập dữ liệu

Sau qua nhiều quá trình huấn luyện mô hình, nhóm nhận ra tỷ lệ chia giữa các tập huấn luyện, tập kiểm thử và tập kiểm tra như sau giảm thiểu hiện tượng underfit của mô hình đáng kể. Hiện tượng underfit này đến từ việc số lượng dữ liệu của tập dữ liệu rất ít so với một bài toán hồi quy.

- Tập huấn luyện (training set): Được chia ra từ 70% từ tập dữ liệu.
- Tập kiểm thử (validation set): Được chia ra từ 20% từ tập dữ liệu.
- Tập kiểm tra (test set): Được chia ra từ 10% còn lại.

4.2.2 Tối ưu siêu tham số

- Trong mô hình CatBoost Regressor, các siêu tham số cần được tối ưu bao gồm:
  - ***depth:*** Độ sâu của cây quyết định. Cây sâu hơn có thể học các mẫu phức tạp hơn nhưng dễ bị quá khớp (overfitting).
  - ***iterations:*** Số lượng cây được xây dựng. Số lượng càng lớn mô hình càng mạnh, nhưng tốn thời gian và có thể gây overfitting nếu không dừng sớm.

- ***learning_rate:*** Tốc độ học. Kiểm soát mức độ cập nhật trọng số sau mỗi vòng lặp. Giá trị nhỏ thường cần nhiều iterations hơn nhưng cho kết quả chính xác hơn.
- ***l2_leaf_reg:*** Hệ số điều chuẩn L2. Giúp ngăn chặn overfitting bằng cách phạt các trọng số quá lớn.

– Trong quá trình tối ưu, nhóm lựa chọn cho tối ưu 10 vòng lặp. Với mỗi vòng lặp, nhóm quyết định thực hiện KFold cross-validation với $k = 3$ để có thể giảm thời gian tính toán mà không hy sinh quá nhiều sự ổn định trong việc huấn luyện.

– Ngoài ra, việc huấn luyện mô hình trong giai đoạn tối ưu hóa siêu tham số được thực hiện trên cả tập lớn bao gồm tập huấn luyện và tập kiểm thử, nhằm giúp tìm được siêu tham số phù hợp. Việc sử dụng KFold cross-validation cũng sẽ giảm thiểu khả năng bị rò rỉ dữ liệu (data leakage) cho mô hình tốt nhất.

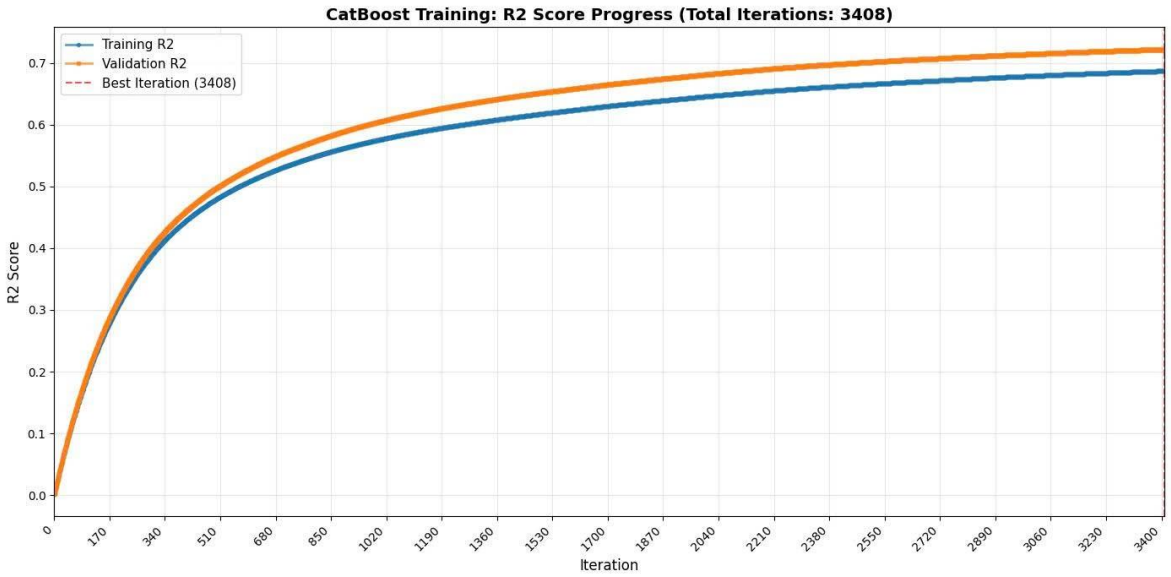
4.2.3 Huấn luyện mô hình tốt nhất

– Sau khi tối ưu hóa thành công, mô hình tốt nhất sẽ được huấn luyện dựa trên tập huấn luyện và kiểm thử ở tập kiểm thử.

– Trong quá trình huấn luyện và kiểm thử, nhóm quyết định thực thi phương pháp dừng sớm nếu mô hình sau 50 vòng lặp không có sự tiến triển gì.

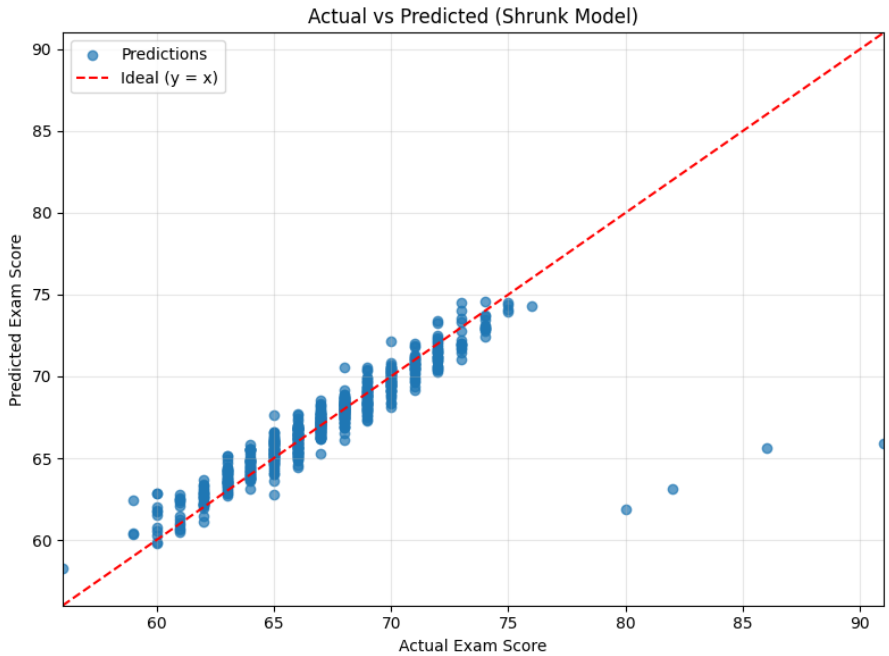
4.3 Kết quả cuối cùng

– Kết quả của mô hình trong suốt quá trình huấn luyện thể hiện rõ việc mô hình không bị hiện tượng underfit hoặc overfit. Ngoài ra, khi đánh giá bằng tiêu chí R2, mô hình đạt được giá trị 0.8.



Hình 4.1 Biểu đồ thể hiện tiêu chí R2 trong suốt quá trình huấn luyện ở tập huấn luyện và tập kiểm thử

– Tiến hành kiểm tra với tập kiểm tra, mô hình cho ra giá trị R2 không quá khác so với lúc huấn luyện mô hình, xấp xỉ 0.73. Điều này chứng minh việc mô hình đã được huấn luyện hoàn chỉnh, không bị hiện tượng underfit dù tập dữ liệu của đề tài không quá lớn.



Hình 4.2 So sánh giá trị thực và giá trị dự đoán của mô hình

CHƯƠNG 5. KẾT LUẬN

5.1 Tổng kết kết quả nghiên cứu

Nghiên cứu này đã giải quyết thành công bài toán dự báo kết quả học tập của học sinh thông qua việc áp dụng các kỹ thuật học máy tiên tiến trên tập dữ liệu giáo dục. Quá trình thực hiện đã tuân thủ chặt chẽ quy trình khai phá dữ liệu, từ bước thu thập, làm sạch, phân tích khám phá đến tiền xử lý và mô hình hóa.

Kết quả phân tích dữ liệu cho thấy các yếu tố như tỷ lệ chuyên cần và số giờ tự học đóng vai trò then chốt, có tác động mạnh mẽ nhất đến điểm số cuối kỳ. Ngược lại, các yếu tố về hoàn cảnh gia đình hay hoạt động thể chất có mức độ ảnh hưởng thấp hơn, cho thấy tính công bằng của hệ thống giáo dục được phản ánh qua dữ liệu.

Về mặt kỹ thuật, nghiên cứu đã tiến hành thực nghiệm so sánh sáu mô hình học máy khác nhau, bao gồm nhóm mô hình tuyến tính (Linear Regression, Ridge, Lasso) và nhóm mô hình cây quyết định (Random Forest, XGBoost, CatBoost). Quá trình đánh giá thông qua kỹ thuật kiểm định chéo K-Fold (với $k = 5$) đã chỉ ra sự vượt trội của các mô hình phi tuyến tính trong việc nắm bắt các mối quan hệ phức tạp của dữ liệu.

Đặc biệt, CatBoost Regressor đã được xác định là mô hình tối ưu nhất cho bài toán này. Với khả năng xử lý tốt cả dữ liệu số và dữ liệu phân loại mà không cần quá nhiều bước tiền xử lý phức tạp, CatBoost không chỉ đạt được độ chính xác cao nhất, thể hiện qua chỉ số R2 cao và RMSE thấp nhất trên tập kiểm thử mà còn duy trì được sự ổn định tốt, hạn chế hiện tượng overfitting nhờ cơ chế dừng sớm và tối ưu hóa siêu tham số bằng Bayesian Optimization.

5.2 Đóng góp của đề tài

Đề tài đã mang lại những đóng góp thiết thực cả về mặt lý luận và thực tiễn:

- Thứ nhất, nghiên cứu cung cấp một quy trình mẫu mực về việc ứng dụng khoa học dữ liệu trong giáo dục, từ việc xử lý các biến định tính phức tạp đến việc giải quyết các vấn đề về phân phối dữ liệu và đa cộng tuyến.

- Thứ hai, kết quả dự báo của mô hình có thể được sử dụng làm công cụ hỗ trợ đắc lực cho nhà trường và giáo viên. Hệ thống có khả năng đưa ra cảnh báo sớm về những học sinh có nguy cơ đạt kết quả thấp dựa trên các chỉ số hành vi học tập hiện tại, từ đó giúp các nhà giáo dục có cơ sở để can thiệp kịp thời và cá nhân hóa lộ trình hỗ trợ cho từng học sinh.

5.3 Hạn chế và hướng phát triển

Mặc dù đã đạt được những kết quả khả quan, đề tài vẫn còn tồn tại một số hạn chế nhất định cần được khắc phục trong tương lai:

- Tập dữ liệu hiện tại, dù có chất lượng tốt, vẫn bị giới hạn về quy mô và phạm vi. Việc mở rộng thu thập dữ liệu từ nhiều trường học khác nhau với đa dạng vùng miền sẽ giúp kiểm chứng tốt hơn khả năng tổng quát hóa của mô hình.
- Nghiên cứu hiện tại mới chỉ tập trung vào các yếu tố định lượng và định tính có cấu trúc. Trong tương lai, việc tích hợp thêm các dữ liệu phi cấu trúc như nhận xét văn bản của giáo viên hay dữ liệu hành vi từ các hệ thống học tập trực tuyến có thể giúp nâng cao hơn nữa độ chính xác của dự báo.
- Mô hình hiện tại hoạt động như một công cụ dự báo độc lập. Hướng phát triển tiếp theo là xây dựng một ứng dụng web hoặc tích hợp mô hình vào hệ thống quản lý học tập của nhà trường, giúp giáo viên và phụ huynh có thể tiếp cận kết quả dự báo một cách trực quan và dễ dàng hơn.

TÀI LIỆU THAM KHẢO

Tiếng Anh

Dorogush, A. V. (2018). CatBoost: gradient boosting with categorical features. *arXiv*.

Han, J. K. (2012). *Data Mining Concepts and Techniques (3rd Ed).* Waltham, MA: Morgan Kaufmann.