ĐẠI HỌC ĐÀ NẴNG
TRƯỜNG ĐẠI HỌC CNTT & TT VIỆT – HÀN

Khoa Khoa Học Máy Tính

PHÂN TÍCH DỮ LIỆU

PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN THÀNH TÍCH CỦA HỌC SINH – SINH VIÊN

Sinh viên thực hiện : 1. Lê Hồng Anh – 21AD002
2. Dương Tấn Huy – 21AD025
3. Lê Đức Thiện – 21AD056
4. Hoàng Lê Kim Vinh – 21AD067

Lớp : 21AD

Giảng viên hướng dẫn : TS. Nguyễn Thanh

Đà Nẵng – 12/2024

ĐẠI HỌC ĐÀ NẴNG
TRƯỜNG ĐẠI HỌC CNTT & TT VIỆT – HÀN

Khoa Khoa Học Máy Tính

PHÂN TÍCH DỮ LIỆU

PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN THÀNH TÍCH CỦA HỌC SINH – SINH VIÊN

Sinh viên thực hiện	:	1. Lê Hồng Anh – 21AD002
		2. Dương Tấn Huy – 21AD025
		3. Lê Đức Thiện – 21AD056
		4. Hoàng Lê Kim Vinh – 21AD067
Lớp	:	21AD
Giảng viên hướng dẫn	:	TS. Nguyễn Thanh

Đà Nẵng – 12/2024

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

……
……
……
……
……
……
……
……
……
……
……
……
……
……

Đà Nẵng, ngày ... tháng ... năm 2024

Cán bộ hướng dẫn

(họ tên và chữ ký)

LỜI CẢM ƠN

Đối với một sinh viên tại trường đại học CNTT & TT Việt Hàn, bài tập lớn cuối kỳ là một minh chứng cho những kiến thức đã có được sau một kỳ học tập tại trường.

Để thực hiện và hoàn thành tốt bài tập lớn cuối kỳ này, chúng em đã nhận được sự giúp đỡ và hướng dẫn rất tận tình của các thầy cô thuộc Khoa Khoa học máy tính, Trường Đại học Công Nghệ Thông Tin và Truyền Thông Việt – Hàn. Chúng em xin cảm ơn các thầy cô thuộc bộ môn chuyên ngành đã cung cấp cho chúng em các thông tin, kiến thức vô cùng quý báu và cần thiết trong suốt thời gian qua để chúng em có thể thực hiện và hoàn thành bài tập lớn cuối kỳ của mình. Đặc biệt chúng em xin chân thành cảm ơn thầy Nguyễn Thanh người đã trực tiếp hướng dẫn chúng em trong thời gian thực hiện bài tập lớn cuối kỳ này. Chúng em xin được gửi lời cảm ơn chân thành nhất với thầy.

Sau đó, xin cảm ơn các bạn trong Ngành Công nghệ thông tin đã ủng hộ, giúp đỡ, chia sẻ kiến thức, kinh nghiệm và tài liệu có được giúp chúng tôi trong quá trình nghiên cứu và thực hiện đề tài.

Cuối cùng, do giới hạn về mặt thời gian và kiến thức cũng như kinh nghiệm thực tiễn nên đề tài không tránh khỏi những sai sót. Chúng em rất mong nhận được sự thông cảm của quý thầy cô và mong đón nhận những góp ý của thầy cô và các bạn.

Chúng em xin chân thành cảm ơn!

MỤC LỤC

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

MỞ ĐẦU

1. Lý do chọn đề tài

Thành tích học tập luôn là một trong những tiêu chí quan trọng để đánh giá năng lực cá nhân và chất lượng giáo dục. Trong thời đại hiện nay, giáo dục không chỉ hướng đến việc truyền đạt kiến thức mà còn giúp học sinh - sinh viên phát triển toàn diện về tư duy, kỹ năng và thái độ sống. Tuy nhiên, để đạt được thành tích tốt, người học không chỉ cần nỗ lực cá nhân mà còn chịu ảnh hưởng từ nhiều yếu tố khách quan và chủ quan.

Thực tế cho thấy, có những học sinh - sinh viên dù sở hữu năng lực cá nhân vượt trội nhưng vẫn gặp khó khăn trong việc duy trì thành tích học tập tốt. Ngược lại, có những cá nhân với xuất phát điểm khiêm tốn lại đạt được những kết quả xuất sắc. Điều này đặt ra câu hỏi: "Những yếu tố nào đóng vai trò quyết định đến thành tích học tập, và mức độ ảnh hưởng của chúng ra sao?"

Ngoài các yếu tố cá nhân như năng lực tư duy, động lực học tập, hay kỹ năng quản lý thời gian, môi trường gia đình và nhà trường cũng đóng vai trò không nhỏ. Ví dụ, sự hỗ trợ về mặt tinh thần và vật chất từ gia đình, chất lượng giảng dạy của giáo viên, hay sự cạnh tranh trong môi trường học đường đều có thể ảnh hưởng đến kết quả học tập. Bên cạnh đó, các yếu tố bên ngoài như điều kiện kinh tế - xã hội, sự phát triển công nghệ và văn hóa địa phương cũng tạo ra những tác động không nhỏ.

Trong bối cảnh cách mạng công nghiệp 4.0, việc áp dụng công nghệ vào giáo dục đã mang lại nhiều cơ hội mới, nhưng đồng thời cũng đặt ra những thách thức. Sự phổ biến của các công cụ học tập trực tuyến, mạng xã hội và các nền tảng thông tin đã thay đổi cách học sinh - sinh viên tiếp cận tri thức. Tuy nhiên, không phải tất cả các em đều có khả năng tận dụng tối ưu những cơ hội này. Một số em có thể gặp khó khăn trong việc thích nghi với môi trường học tập mới hoặc bị phân tâm bởi những yếu tố bên ngoài. Điều này đòi hỏi một sự phân tích sâu sắc hơn để hiểu rõ hơn về tác động của các yếu tố này.

Việc lựa chọn đề tài "Phân tích dữ liệu - Các yếu tố ảnh hưởng đến thành tích của học sinh - sinh viên" xuất phát từ mong muốn sử dụng các phương pháp phân tích dữ liệu hiện đại nhằm khám phá, đo lường và đánh giá tác động của các yếu tố trên. Kết quả của nghiên cứu không chỉ giúp đưa ra những giải pháp thiết thực để cải thiện kết quả học tập mà còn đóng góp ý nghĩa vào việc định hướng chiến lược giáo dục hiệu quả hơn.

Ngoài ra, việc thực hiện đề tài này còn tạo cơ hội để người thực hiện ứng dụng các công cụ phân tích dữ liệu như Python, R, hoặc phần mềm thống kê vào giải quyết vấn đề thực tiễn. Đây không chỉ là một bài tập học thuật mà còn là cơ hội để tiếp cận và giải quyết những thách thức thực tế trong lĩnh vực giáo dục, từ đó nâng cao kỹ năng chuyên môn và khả năng tư duy logic.

2. Mục tiêu nghiên cứu

Đề tài "Phân tích dữ liệu - Các yếu tố ảnh hưởng đến thành tích của học sinh - sinh viên" này sẽ tập trung vào việc phân tích các yếu tố ảnh hưởng đến thành tích học tập của học sinh - sinh viên thông qua các phương pháp phân tích dữ liệu. Thành tích học tập không chỉ phụ thuộc vào năng lực cá nhân mà còn chịu ảnh hưởng từ nhiều yếu tố bên ngoài, bao gồm yếu tố cá nhân, gia đình, nhà trường và các yếu tố xã hội. Mục tiêu là xác định và đánh giá mức độ tác động của từng yếu tố này đối với kết quả học tập của học sinh - sinh viên.

Cụ thể, nghiên cứu sẽ phân tích các yếu tố như giới tính, độ tuổi, động lực học tập, sức khỏe, và kỹ năng quản lý thời gian (yếu tố cá nhân); tình trạng kinh tế gia đình, trình độ học vấn của phụ huynh, và mức độ hỗ trợ gia đình (yếu tố gia đình); chất lượng giảng dạy, môi trường học tập, và sự quan tâm của giáo viên (yếu tố nhà trường); và các yếu tố bên ngoài như điều kiện kinh tế - xã hội, ảnh hưởng của công nghệ, hay áp lực xã hội.

Mục tiêu chính của nghiên cứu là sử dụng các công cụ phân tích dữ liệu hiện đại để tìm ra mối quan hệ giữa các yếu tố này và thành tích học tập. Các phương pháp phân tích như hồi quy, phân tích tương quan và các mô hình học máy sẽ được áp dụng để đo lường mức độ ảnh hưởng của các yếu tố này. Kết quả nghiên cứu không chỉ giúp hiểu rõ hơn về các yếu tố quyết định thành tích học tập mà còn cung cấp các giải pháp thực tiễn nhằm cải thiện kết quả học tập và hỗ trợ các chiến lược giáo dục hiệu quả.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài này tập trung vào các học sinh - sinh viên thuộc các cấp học khác nhau, từ trung học phổ thông đến đại học. Các yếu tố ảnh hưởng đến thành tích học tập được phân thành ba nhóm chính:

Nhóm yếu tố cá nhân bao gồm số giờ học mỗi tuần, số giờ ngủ trung bình mỗi đêm, điểm số từ các kỳ thi trước, mức độ động lực học tập, số giờ hoạt động thể chất mỗi tuần, sự hiện diện của khuyết tật học tập và giới tính. Nhóm yếu tố gia đình gồm mức độ tham gia của phụ huynh vào quá trình học tập, tính khả dụng của các nguồn lực giáo dục, thu nhập gia đình và trình độ học vấn của cha mẹ. Nhóm yếu tố nhà trường bao gồm tỷ lệ tham dự lớp học, tham gia hoạt động ngoại khóa, số buổi học kèm mỗi tháng, chất lượng giáo viên, loại trường theo học, ảnh hưởng của bạn bè và khoảng cách từ nhà đến trường.

Dữ liệu nghiên cứu sẽ được lấy từ các học sinh - sinh viên tại các trường đại học và trung học phổ thông ở cả khu vực đô thị và nông thôn, nhằm đảm bảo tính đại diện và đa dạng của mẫu nghiên cứu.

4. Phương pháp nghiên cứu

Để phân tích các yếu tố ảnh hưởng đến thành tích học tập của học sinh - sinh viên, nghiên cứu này sử dụng phương pháp nghiên cứu định lượng kết hợp với các công cụ phân tích dữ liệu hiện đại. Quá trình nghiên cứu bắt đầu với việc thu thập dữ liệu từ học sinh - sinh viên thông qua các bảng hỏi khảo sát được thiết kế để đánh giá các yếu tố như số giờ học, động lực học tập, hoạt động thể chất, điểm số kỳ thi, tình trạng sức khỏe,

thu nhập gia đình, trình độ học vấn của phụ huynh, chất lượng giáo viên và các yếu tố khác. Dữ liệu thu thập sẽ được xử lý và phân tích bằng các phương pháp thống kê cơ bản như phân tích mô tả, kiểm định giả thuyết (t-test, ANOVA) và phân tích tương quan để xác định mối quan hệ giữa các yếu tố và thành tích học tập.

Ngoài ra, phân tích hồi quy tuyến tính hoặc hồi quy logistic sẽ được áp dụng để kiểm tra sự ảnh hưởng của các yếu tố độc lập đến kết quả học tập của học sinh - sinh viên. Phân tích đa biến sẽ giúp xác định các yếu tố có tác động mạnh mẽ nhất và mối quan hệ giữa chúng. Để nâng cao độ chính xác và tối ưu hóa phân tích, nghiên cứu cũng áp dụng các mô hình học máy như phân loại, hồi quy, và phân tích cụm. Các thuật toán như Random Forest, Support Vector Machines (SVM) hoặc Neural Networks sẽ được sử dụng để dự đoán thành tích học tập dựa trên các yếu tố đã thu thập, giúp cải thiện khả năng dự đoán và phân tích mối quan hệ phức tạp giữa các yếu tố.

Cuối cùng, dựa trên kết quả phân tích, nghiên cứu sẽ đưa ra các giải pháp cụ thể nhằm cải thiện thành tích học tập của học sinh - sinh viên, bao gồm các biện pháp như cải thiện môi trường học tập, thay đổi phương pháp giảng dạy, tăng cường sự tham gia của gia đình và các hỗ trợ bổ sung cho học sinh có hoàn cảnh khó khăn. Phương pháp nghiên cứu này giúp cung cấp cái nhìn sâu sắc về các yếu tố ảnh hưởng đến thành tích học tập và đề xuất những giải pháp thực tiễn nhằm nâng cao chất lượng giáo dục.

Chương 1 – TỔNG QUAN VỀ PHÂN TÍCH DỮ LIỆU

1.1 Phân tích dữ liệu là gì?

1.1.1. Định nghĩa.

Phân tích dữ liệu là hoạt động tổng quát bao gồm tiếp nhận, phân chia, sàng lọc và khai thác tối đa giá trị data để chuyển biến nguồn dữ liệu thô trở thành những thông tin hữu ích đối với doanh nghiệp.

Phân tích dữ liệu bao gồm nhiều phương pháp khác nhau. Bất kỳ loại dữ liệu nào cũng có thể được áp dụng kỹ thuật phân tích để hiểu rõ, cải thiện hoạt động, tối ưu hóa quy trình và gia tăng hiệu quả chung cho doanh nghiệp hoặc hệ thống.

Lấy ví dụ, một công ty giao hàng thông qua phân tích dữ liệu để tránh những tuyến đường tắc nghẽn và xác định đường đi tốt nhất tại từng thời điểm. Nhờ vậy, họ đã rút ngắn được thời gian giao hàng, cắt giảm chi phí vận chuyển và tạo ra trải nghiệm thuận tiện hơn cho nhân viên lái xe.

1.1.2. Các phương pháp phân tích dữ liệu (methods of data analytics).

Dữ liệu hỗ trợ doanh nghiệp ra quyết định chính xác và nhanh chóng dựa trên các phương pháp phân tích khác nhau. Dưới đây là 4 loại phân tích dữ liệu thường được sử dụng hiện nay:

- ***Phân tích mô tả (Descriptive analysis):***

Phân tích mô tả là phương pháp tập trung vào việc mô tả và tóm tắt các dữ liệu hiện có thông qua các đặc điểm, xu hướng hay sự biến thiên của dữ liệu – nhưng không đưa ra dự đoán hoặc kết luận. Các thông số trả về có dạng bảng, biểu đồ, số liệu thống kê mô tả dựa trên các giá trị trung bình, phương sai, tần suất, và mức độ phân phối.

Để dễ hiểu, bạn có thể tham khảo ví dụ sau: Một doanh nghiệp sản xuất thiết bị điện tử muốn tìm hiểu thị trường tiêu thụ của mình bằng phương pháp phân tích mô tả. Với cách thức này, kết quả trả lại là các bảng biểu thống kê về đặc điểm và hành vi của khách hàng (demographic). Dựa trên bảng này, công ty có thể điều chỉnh chiến lược kinh doanh, phát triển sản phẩm và tiếp cận thị trường một cách đúng đắn nhất.

- ***Phân tích suy luận (Diagnostic analytics) :***

Phân tích suy luận là phương pháp phân tích làm rõ nguyên nhân hoặc tương quan giữa các sự kiện, biến số trong một mẫu dữ liệu, bao gồm các hoạt động phân tích tương quan, phân tích biến thể, phân tích hồi quy và phân tích nhân quả. Phương pháp này giúp nhà quản trị hiểu sâu sắc về mối liên kết giữa các yếu tố trong một trường, tệp thông tin. Đồng thời, kết quả trả về sẽ giúp giải thích tại sao một sự kiện hoặc sự việc đã xảy ra trước đó.

- ***Phân tích dự đoán (Predictive analytics):***

Bằng việc sử dụng các mô hình và thuật toán dự đoán, đo lường kết quả, sự kiện hoặc giá trị trong tương lai, phân tích dự đoán giúp xử lý dữ liệu dựa trên lịch sử và bản

mẫu của data. Phân tích dự đoán còn giúp tính phần trăm xác suất xảy ra của các sự kiện, nhà quản trị dễ dàng đưa ra các quyết định phù hợp.

Phương pháp này bao gồm các cách thực hiện khác nhau như hồi quy tuyến tính, hồi quy logistic, cây quyết định (decision tree), mạng nơ-ron và máy học (machine learning).

- ***Phân tích đề xuất (Prescriptive analytics):***

Phân tích đề xuất là phương pháp ứng dụng đa dạng các thuật toán nhằm đề xuất và tối ưu hóa quyết định, hành động trong tương lai. Chúng điều chỉnh các yếu tố và tham số trong quá trình phân tích, từ đó giúp người phân tích giải quyết vấn đề hiệu quả và đạt được mục tiêu trong kế hoạch. Các kỹ thuật tối ưu hóa, mô phỏng, và quyết định đa mục tiêu thường được ứng dụng trong phương pháp này.

1.2 Phân tích dữ liệu có ý nghĩa như thế nào trong kinh doanh?

Phân tích dữ liệu đóng vai trò quan trọng hàng đầu trong việc cung cấp các thông tin cần thiết giúp nhà quản trị hiểu, đánh giá và đưa ra quyết định chính xác dựa trên dữ liệu số trực quan. Bằng cách lựa chọn phương pháp và công cụ phân tích phù hợp, doanh nghiệp dễ dàng xác định được phương hướng, chiến lược kinh doanh và định hướng phát triển lâu dài.

Để hiểu rõ hơn về tầm quan trọng của việc phân tích dữ liệu với doanh nghiệp, ta có thể theo dõi ví dụ dưới đây:

Để tối ưu hiệu quả của một chiến dịch quảng cáo trực tuyến, nhân sự phòng ban Marketing sẽ cần thu thập nhiều dữ liệu khác nhau, thuộc đa dạng khía cạnh như:

- ***Phân tích hiệu quả chiến dịch***: Thông qua các chỉ số như số lần nhấp vào quảng cáo (click-through rate), tỷ lệ chuyển đổi (conversion rate), doanh thu,… và các yếu tố ảnh hưởng tới chúng, bộ phận marketing có thể xác định được đâu là chiến dịch thành công và không thành công.

- ***Phân tích đối tượng khách hàng***: Các dữ liệu về đối tượng khách hàng như độ tuổi, giới tính, vị trí địa lý và hành vi tiêu dùng giúp bộ phận marketing nắm rõ hơn về nhóm khách hàng mục tiêu của mình. Từ đó, họ dễ định hình lại thông điệp truyền tải, nhắm trúng Insight khách hàng và đẩy mạnh tương tác tích cực từ khách hàng.

- ***Phân tích nền tảng quảng cáo***: Các chỉ số xác định tiềm năng của nền tảng như hình thức quảng cáo, vị trí quảng cáo, tệp tương tác,… giúp bộ phận marketing chọn đúng nơi để đặt quảng cáo của mình và phân bổ ngân sách vào đó một cách hiệu quả.

Như vậy có thể thấy dù chỉ là một chiến dịch chạy quảng cáo nhỏ, việc phân tích dữ liệu đã đóng vai trò quan trọng giúp doanh nghiệp đánh giá hiệu quả hiện tại và đưa ra kế hoạch phát triển sau này.

1.3 Một số kĩ thuật phân tích dữ liệu cơ bản

1.3.1 Phân tích hồi quy (Regression analysis) .

Đây là phương pháp phân tích thống kê nhằm xác định mối liên kết giữa một biến

số với một hay nhiều biến số độc lập khác nhau. Chúng giúp đo lường và dự đoán tương quan giữa các biến số này và mô phỏng chúng dựa trên một mô hình toán học.

Lấy ví dụ bạn là chủ một doanh nghiệp thời trang và bạn muốn biết mối quan hệ giữa số lượng quảng cáo trực tuyến và doanh số bán hàng hàng trong tháng một. Khi sử dụng phân tích hồi quy, bạn có thể xác định tác động của số lượng quảng cáo trực tuyến đến doanh số bán hàng. Từ đó, bạn dễ dàng điều chỉnh ngân sách cho quảng cáo trực tuyến để đạt doanh số kỳ vọng.

1.3.2 Phân tích tổ hợp (Cohort analysis).

Đây là phương pháp phân loại cá nhân, khách hàng hoặc nhiều đối tượng vào các nhóm có những đặc điểm giống nhau. Lấy ví dụ, trong hoạt động marketing, phân tích tổ hợp giúp bạn thấy rõ sự thay đổi của một nhóm khách hàng theo thời gian và đâu là yếu tố gây ảnh hưởng lên họ.

1.3.3 Phân tích nhân tố (Factor analysis).

Đây là phương pháp giúp làm rõ mối quan hệ giữa các biến dữ liệu với nhau bằng cách xác định các nhân tố chung được ẩn đằng sau, từ đó, phương pháp này giúp làm giảm số chiều của dữ liệu trong tổ chức.

Ví dụ, một công ty muốn tìm hiểu mối quan hệ giữa sự hài lòng của khách hàng, độ uy tín doanh nghiệp và hiệu suất kinh doanh. Phân tích cho biết nhân tố chung ẩn giữa 3 biến số trên – có khả năng ảnh hưởng tới cả 3 biến số – là chất lượng sản phẩm/ dịch vụ. Từ đó, công ty có thể tập trung vào cải thiện chất lượng của sản phẩm và các dịch vụ cung cấp.

1.3.4 Phân tích chuỗi thời gian (Time series analysis).

Đây là phương pháp phân tích dữ liệu dựa nghiên cứu về sự thay đổi của các tệp dữ liệu theo thời gian, giúp xác định xu hướng, mô hình và chu kỳ biến đổi trong một chuỗi thời gian nhất định. Trong doanh nghiệp, phương pháp này được ứng dụng phổ biến nhất trong dự đoán giá cổ phiếu, doanh số bán hàng,…

1.4 Quy trình phân tích dữ liệu

Quy trình phân tích dữ liệu có thể đơn giản hoặc phức tạp tùy theo mô hình doanh nghiệp đang vận hành, nhưng về cơ bản sẽ bao gồm 4 bước sau:

1.4.1 Thu thập dữ liệu.

Thu thập dữ liệu là bước đầu tiên giúp doanh nghiệp xác định nguồn tài nguyên dữ liệu sẵn có hiện tại. Một số hoạt động diễn ra trong bước này bao gồm:

- Xác định mục tiêu thu thập: Phân tích dữ liệu phục vụ mục đích gì? Bạn cần kết quả nào ở việc phân tích? Bạn chỉ cần liệt kê và trả lời những câu hỏi cần thiết để xác định mục tiêu của mình.

- Xác định nguồn gốc và tính chất của dữ liệu: Dữ liệu có thể thu thập từ các hệ thống nào bên trong doanh nghiệp? Hoặc qua các báo cáo hoặc từ các nguồn xác thực nào khác bên ngoài? Định dạng, cấu trúc, và độ chính xác của chúng như thế nào?

- Xác định nơi lưu trữ dữ liệu: Dữ liệu nên được lưu trữ ở đâu để an toàn và được sắp xếp có khoa học, đồng thời tiện lợi cho doanh nghiệp truy cập và sử dụng?

1.4.2 Tiền xử lý dữ liệu.

Tiền xử lý dữ liệu là quá trình chuẩn bị dữ liệu để chuẩn bị cho các bước phân tích hoặc mô hình hóa sau này. Một số hoạt động quan trọng có thể kể tới:

- Làm sạch dữ liệu: Loại bỏ các dữ liệu sai, không hợp lệ, trùng lặp hoặc bị nhiễu; đồng thời dữ liệu sẽ được kiểm tra, sửa các giá trị lỗi và chuyển đổi định dạng về cùng một đơn vị hoặc thang đo nếu cần.

- Xử lý dữ liệu thiếu: Xác định và xử lý các dữ liệu bị thiếu trong tập dữ liệu bằng một số phương pháp như: điền giá trị trung bình, giá trị gần nhất, hoặc dự báo để điền dữ liệu có độ sai số ít nhất. Hoặc trong nhiều trường hợp, các dòng hoặc cột dữ liệu có quá nhiều giá trị thiếu không thể khôi phục sẽ được loại bỏ.

1.4.3 Khám phá dữ liệu.

Khám phá dữ liệu là bước quan trọng để hiểu sâu về dữ liệu trước khi tiến hành phân tích chi tiết hơn, ví dụ như giúp nhận diện các đặc điểm quan trọng, phân loại các đặc tính của dữ liệu, và tạo ra cái nhìn toàn cảnh. Hai hoạt động khám phá cơ bản bao gồm:

- Mô tả thống kê: Là hoạt động tổng hợp, mô tả đặc điểm chính của dữ liệu bằng cách sử dụng các thống kê cơ bản như trung bình, phương sai, phân phối và mẫu tổ chức.

- Trực quan hóa dữ liệu: Là hoạt động sử dụng biểu đồ và đồ thị để thể hiện sự phân bố, xu hướng và mối quan hệ giữa các biến số trong tập dữ liệu.

1.4.4 Phân tích dữ liệu.

Phân tích dữ liệu là bước cuối cùng và có vai trò quan trọng nhất trong toàn bộ quá trình.

Các hoạt động tại bước này tập trung chủ yếu vào việc sử dụng các công cụ và kỹ thuật phân tích phù hợp (đã được trình bày tại phần 3 của bài viết) để trích xuất thông tin quan trọng từ các dữ liệu đã xử lý, sau đó rút ra những kết luận có ý nghĩa.

Kết quả phân tích có thể làm phát sinh những câu hỏi mới, đòi hỏi việc thu thập thêm dữ liệu hoặc tiến hành các phân tích bổ sung.

1.5 Các công cụ hỗ trợ phân tích dữ liệu

Các công cụ phân tích dữ liệu phổ biến bao gồm Power BI, Excel và Python. Power BI giúp trực quan hóa và chia sẻ dữ liệu hiệu quả, phù hợp với môi trường doanh nghiệp. Excel cung cấp các tính năng cơ bản để xử lý và trực quan hóa dữ liệu, nhưng hạn chế khi làm việc với dữ liệu lớn. Python, với các thư viện như NumPy, Pandas và Matplotlib, là công cụ mạnh mẽ cho phân tích dữ liệu phức tạp và xây dựng mô hình dự đoán. Mỗi công cụ có ưu điểm riêng, tùy thuộc vào yêu cầu của bài toán.

Chương 2 – PHÂN TÍCH DỮ LIỆU

2.1 Giới thiệu chung về Dataset

2.1.1 Mô tả Dataset.

Dataset "StudentPerformanceFactors" tập trung vào các yếu tố ảnh hưởng đến kết quả học tập của học sinh, đặc biệt là điểm thi cuối kỳ. Dataset này bao gồm nhiều khía cạnh từ học thuật đến phi học thuật, nhằm phản ánh bức tranh toàn diện về môi trường học tập của học sinh. Thông qua phân tích dữ liệu, có thể khám phá các yếu tố chủ chốt giúp cải thiện thành tích học tập, từ đó hỗ trợ nhà giáo dục và phụ huynh trong việc đưa ra các chiến lược hỗ trợ học sinh học tập hiệu quả hơn.

2.1.2 Định nghĩa các biến trong Dataset.

STT	Tên cột	Mô tả
1	Hours_Studied	Số giờ học mỗi tuần.
2	Attendance	Tỷ lệ phần trăm các lớp học đã tham dự.
3	Parental_Involvement	Mức độ tham gia của phụ huynh vào quá trình học tập của học sinh (Thấp, Trung bình, Cao).
4	Access_to_Resources	Tính khả dụng của các nguồn lực giáo dục (Thấp, Trung bình, Cao).
5	Extracurricular_Activities	Tham gia các hoạt động ngoại khóa (Có, Không).
6	Sleep_Hours	Số giờ ngủ trung bình mỗi đêm.
7	Previous_Scores	Điểm số từ các kỳ thi trước.
8	Motivation_Level	Mức độ động lực của học sinh (Thấp, Trung bình, Cao).
9	Internet_Access	Tính khả dụng của kết nối internet (Có, Không).
10	Tutoring_Sessions	Số buổi học kèm đã tham dự mỗi tháng.
11	Family_Income	Thu nhập gia đình.
12	Teacher_Quality	Chất lượng giáo viên (Thấp, Trung bình, Cao).
13	School_Type	Loại trường đã theo học (Công lập, Tư thục).
14	Peer_Influence	Ảnh hưởng của bạn bè đến kết quả học tập (Tích cực, Trung lập, Tiêu cực).
15	Physical_Activity	Số giờ hoạt động thể chất trung bình mỗi tuần.
16	Learning_Disabilities	Có khuyết tật học tập không (Có, Không).
17	Parental_Education_Level	Trình độ học vấn cao nhất của cha mẹ (Trung học phổ thông, Cao đẳng, Sau đại học).

18	Distance_from_Home	Khoảng cách từ nhà đến trường (Gần, Trung bình, Xa).
19	Gender	Giới tính của học sinh (Nam, Nữ).
20	Exam_Score	Điểm thi cuối kỳ.

Bảng 1: Định nghĩa biến Dataset

2.2 Tiền xử lý dữ liệu

- Bước 1: Import thư viện drive để đọc được file dữ liệu từ google drive.
- Bước 2: In ra các dòng đầu tiên của file dữ liệu.
- Bước 3: Kiểm tra kích thước dữ liệu.

```
from google.colab import drive
drive.mount('/content/drive')

file_path = '/content/drive/MyDrive/csv/StudentPerformanceFactors.csv'
df = pd.read_csv(file_path)
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
df.head()
```

	Hours_Studied	Attendance	Parental_Involvement	Access_to_Resources	Extracurricular_Activities	Sleep_Hours	Previous_Scores	Motivation_Level	Internet_Access	Tutoring_Sessions
0	23	84	Low	High	No	7	73	Low	Yes	0
1	19	64	Low	Medium	No	8	59	Low	Yes	2
2	24	98	Medium	Medium	Yes	7	91	Medium	Yes	2
3	29	89	Low	Medium	Yes	8	98	Medium	Yes	1
4	19	92	Medium	Medium	Yes	6	65	Medium	Yes	3

```
df.shape
```

```
(6607, 20)
```

Hình 2.1: Import dữ liệu, in ra các dòng đầu tiên và kiểm tra kích thước file dữ liệu

- Bước 4: Kiểm tra thông tin file dữ liệu như tên cột, kiểm tra NaN, kiểu dữ liệu.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6607 entries, 0 to 6606
Data columns (total 20 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Hours_Studied               6607 non-null   int64
 1   Attendance                  6607 non-null   int64
 2   Parental_Involvement        6607 non-null   object
 3   Access_to_Resources         6607 non-null   object
 4   Extracurricular_Activities  6607 non-null   object
 5   Sleep_Hours                 6607 non-null   int64
 6   Previous_Scores             6607 non-null   int64
 7   Motivation_Level            6607 non-null   object
 8   Internet_Access             6607 non-null   object
 9   Tutoring_Sessions           6607 non-null   int64
 10  Family_Income               6607 non-null   object
 11  Teacher_Quality             6529 non-null   object
 12  School_Type                 6607 non-null   object
 13  Peer_Influence              6607 non-null   object
 14  Physical_Activity           6607 non-null   int64
 15  Learning_Disabilities       6607 non-null   object
 16  Parental_Education_Level    6517 non-null   object
 17  Distance_from_Home          6540 non-null   object
 18  Gender                      6607 non-null   object
 19  Exam_Score                  6607 non-null   int64
dtypes: int64(7), object(13)
memory usage: 1.0+ MB
```

Hình 2.2: Kiểm tra thông tin các cột từ file dữ liệu

- Bước 5: Kiểm tra cột nào đang có giá trị Null.

```
df.isnull().sum()
```

	0
Hours_Studied	0
Attendance	0
Parental_Involvement	0
Access_to_Resources	0
Extracurricular_Activities	0
Sleep_Hours	0
Previous_Scores	0
Motivation_Level	0
Internet_Access	0
Tutoring_Sessions	0
Family_Income	0
Teacher_Quality	78
School_Type	0
Peer_Influence	0
Physical_Activity	0
Learning_Disabilities	0
Parental_Education_Level	90
Distance_from_Home	67
Gender	0
Exam_Score	0

dtype: int64

Hình 2.3: Kiểm tra giá trị Null

- Bước 6: Thực hiện loại bỏ khoảng trắng, thay thế các giá trị Null

```
# Loại bỏ khoảng trắng thừa ở đầu hoặc cuối tên cột
df.columns = df.columns.str.strip()

# Thay thế giá trị thiếu
df['Teacher_Quality'] = df['Teacher_Quality'].fillna(df['Teacher_Quality'].mode()[0])
df['Parental_Education_Level'] = df['Parental_Education_Level'].fillna(df['Parental_Education_Level'].mode()[0])
df['Distance_from_Home'] = df['Distance_from_Home'].fillna(df['Distance_from_Home'].mode()[0])
```

Hình 2.4: Xử lý giá trị Null

- Bước 7: Thực hiện xử lý outliers, loại bỏ các giá trị ngoại lai

```python
# Tính IQR (Interquartile Range) để phát hiện outliers
Q1 = df['Hours_Studied'].quantile(0.25)
Q3 = df['Hours_Studied'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Loại bỏ các giá trị ngoại lai
df = df[(df['Hours_Studied'] >= lower_bound) & (df['Hours_Studied'] <= upper_bound)]
```

Hình 2.5: Xử lý outliers, loại bỏ các giá trị ngoại lai

```python
df.isnull().sum()
```

	0
Hours_Studied	0
Attendance	0
Parental_Involvement	0
Access_to_Resources	0
Extracurricular_Activities	0
Sleep_Hours	0
Previous_Scores	0
Motivation_Level	0
Internet_Access	0
Tutoring_Sessions	0
Family_Income	0
Teacher_Quality	0
School_Type	0
Peer_Influence	0
Physical_Activity	0
Learning_Disabilities	0
Parental_Education_Level	0
Distance_from_Home	0
Gender	0
Exam_Score	0

dtype: int64

Hình 2.6: Kiểm tra lại các giá trị Null

- Bước 8: Thực hiện chuyển vị dataframe, để có thể dễ dàng xem xét các chỉ số theo

từng cột của dataframe gốc.

```
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Hours_Studied	6564.0	19.969226	5.828309	4.0	16.0	20.0	24.0	36.0
Attendance	6564.0	79.955972	11.540905	60.0	70.0	80.0	90.0	100.0
Sleep_Hours	6564.0	7.029403	1.467372	4.0	6.0	7.0	8.0	10.0
Previous_Scores	6564.0	75.081810	14.402906	50.0	63.0	75.0	88.0	100.0
Tutoring_Sessions	6564.0	1.493601	1.232201	0.0	1.0	1.0	2.0	8.0
Physical_Activity	6564.0	2.968160	1.030714	0.0	2.0	3.0	4.0	6.0
Exam_Score	6564.0	67.226996	3.852922	56.0	65.0	67.0	69.0	101.0

Hình 2.7: Tóm tắt thống kê cơ bản cho các cột số và chuyển vị dataframe

- Bước 9: Duyệt qua từng cột trong dataframe df. Đếm số lượng xuất hiện của mỗi giá trị duy nhất trong từng cột bằng phương thức value_counts(). Hiển thị kết quả và in một dòng phân cách để dễ nhìn.

```
for col in df:
    print(df[col].value_counts())
    print('----------------------------')
```

```
Hours_Studied
20    465
19    441
21    431
23    411
22    402
18    401
17    381
24    357
16    351
15    315
25    289
14    269
26    263
27    229
13    218
12    192
28    171
11    146
29    134
30    123
10     94
9      86
31     77
8      58
32     54
7      51
33     40
34     29
5      21
35     20
4      17
6      17
36     11
Name: count, dtype: int64
```

Hình 2.8: Duyệt qua từng cột trong DataFrame

- Đếm tần suất xuất hiện của các giá trị duy nhất trong mỗi cột bằng value_counts().

2.3 Câu hỏi nghiên cứu và Trực quan hóa dữ liệu

- **Tạo các biến mới có ý nghĩa**

```
# Điểm thi trung bình theo loại trường học
school_avg_score = df.groupby('School_Type')['Exam_Score'].mean()
df['Avg_Score_By_School'] = df['School_Type'].map(school_avg_score)

# Đánh giá mức độ hiệu quả học tập
df['Study_Efficiency'] = df['Exam_Score'] / df['Hours_Studied']

# Biến tương tác giữa School_Type và Hours_Studied
df['School_Hours_Interaction'] = df['School_Type'].astype(str) + "_" + (df['Hours_Studied'] > df['Hours_Studied'].median()).astype(str)
```

Hình 2.9: Tạo các biến mới có ý nghĩa

Avg_Score_By_School:

- Ý nghĩa: Điểm thi trung bình của từng loại trường học (theo cột School_Type).
- Cách tính:

+ Tính điểm trung bình Exam_Score cho từng loại trường trong School_Type bằng cách nhóm dữ liệu (groupby) và tính trung bình (mean).

+ Gán giá trị tương ứng với từng học sinh thông qua map().

- Mục đích: So sánh mức điểm trung bình giữa các loại trường học.

Study_Efficiency:

- Ý nghĩa: Mức độ hiệu quả học tập của học sinh.
- Cách tính: Tỷ lệ giữa Exam_Score (điểm thi) và Hours_Studied (giờ học).
- Mục đích: Đánh giá khả năng học tập hiệu quả (điểm thi cao với số giờ học thấp).

School_Hours_Interaction:

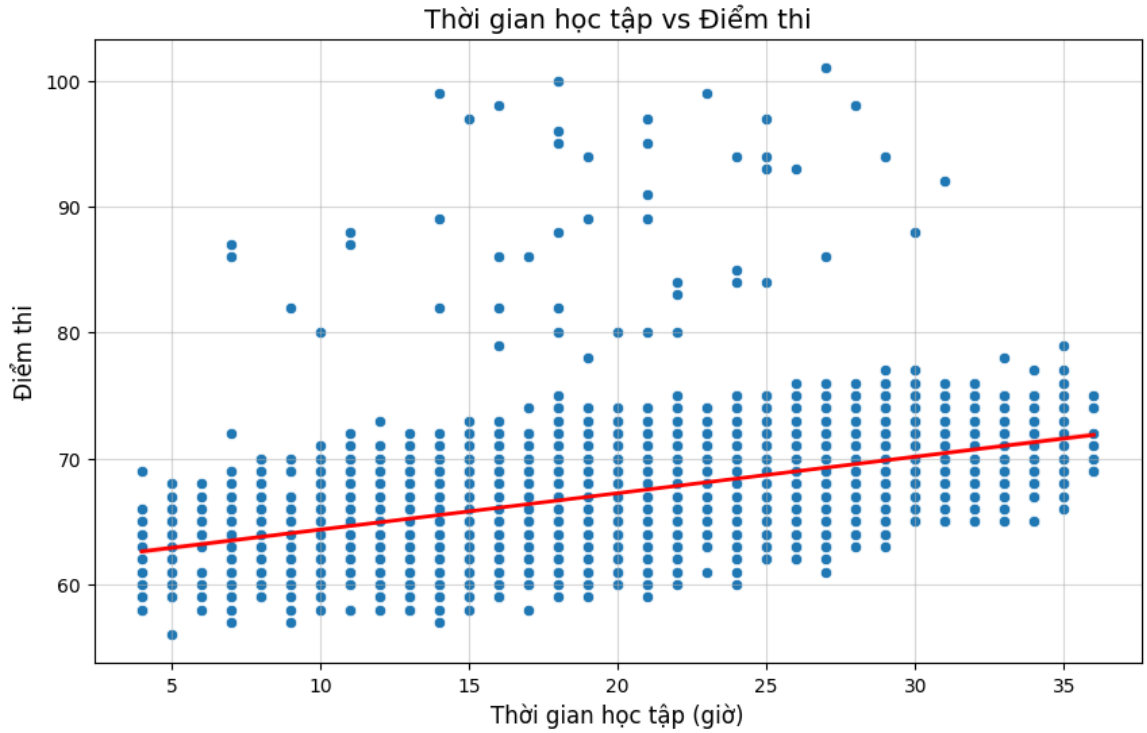
- Ý nghĩa: Biến tương tác giữa loại trường học (School_Type) và việc học sinh học trên hoặc dưới giá trị trung vị của số giờ học (Hours_Studied).
- Cách tính:

+ Chuyển School_Type thành chuỗi ký tự.

+ So sánh Hours_Studied với giá trị trung vị (median) và gắn nhãn True (học trên trung vị) hoặc False (học dưới trung vị).

+ Kết hợp hai thông tin này thành một biến dạng chuỗi, ví dụ: "Public_True", "Private_False".

- **Tương quan giữa thời gian học và điểm thi**



Hình 2.10: Biểu đồ tương quan giữa thời gian học tập và điểm thi

Miêu tả biểu đồ:

Biểu đồ này sử dụng hai yếu tố chính. Thứ nhất là scatterplot (các chấm màu xanh dương), biểu diễn từng học sinh với thời gian học tập (trục x) và điểm thi (trục y). Thứ hai là đường hồi quy tuyến tính (màu đỏ), thể hiện xu hướng tổng quát giữa thời gian học tập và điểm thi.

Nhận xét chi tiết

Xu hướng tổng quát: Biểu đồ cho thấy mối quan hệ tuyến tính dương, nghĩa là khi thời gian học tăng, điểm thi cũng có xu hướng tăng. Tuy nhiên, xu hướng này không thực sự mạnh mẽ vì dữ liệu phân tán khá rộng.

Phân bố dữ liệu: Các chấm không tập trung vào một vùng cố định mà rải rác trên biểu đồ. Một số học sinh học ít giờ vẫn đạt điểm cao, trong khi có học sinh học nhiều nhưng điểm lại thấp. Điều này gợi ý rằng ngoài thời gian học, có thể có các yếu tố khác như phương pháp học tập hoặc mức độ tập trung ảnh hưởng đến kết quả.

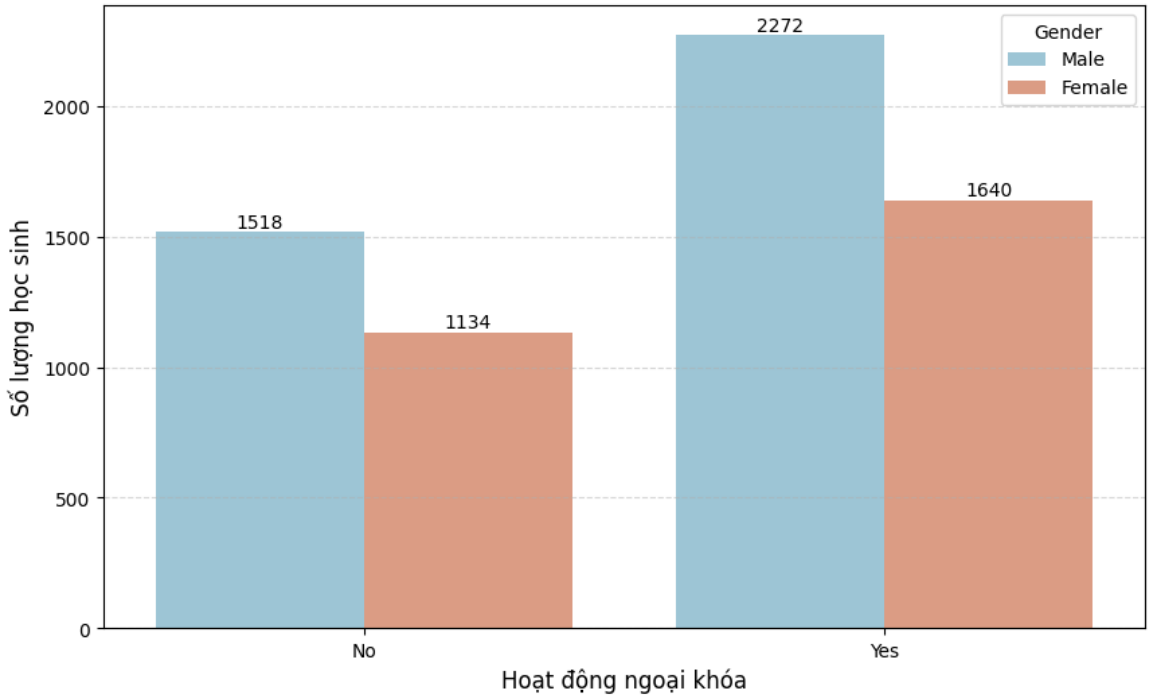
Hiện tượng ngoại lệ: Biểu đồ xuất hiện những trường hợp ngoại lệ rõ rệt. Một số học sinh học rất ít nhưng vẫn đạt điểm cao, có thể nhờ khả năng tự nhiên hoặc kiến thức sẵn có. Ngược lại, một số học sinh học rất nhiều giờ nhưng điểm vẫn thấp, có thể do phương pháp học chưa hiệu quả hoặc bị ảnh hưởng bởi các yếu tố ngoại cảnh như áp lực thi cử.

Độ phù hợp của mô hình: Đường hồi quy tuyến tính minh họa xu hướng tổng quát khá tốt. Tuy nhiên, với sự phân tán lớn, nó không thể phản ánh chi tiết hết tất cả các dữ liệu. Có thể thử nghiệm các mô hình phi tuyến như hồi quy bậc hai hoặc các kỹ thuật khác để nắm bắt tốt hơn sự phức tạp của dữ liệu.

Kết luận

Thời gian học tập có tác động tích cực đến điểm thi, nhưng không phải là yếu tố duy nhất. Độ phân tán lớn cho thấy cần tập trung nhiều hơn vào chất lượng và phương pháp học tập. Biểu đồ cung cấp một cái nhìn tổng quan hữu ích, nhưng cần được bổ sung bằng các phân tích sâu hơn để rút ra kết luận chính xác và đầy đủ hơn.

- **Tham gia hoạt động ngoại khóa theo giới tính**



Hình 2.11: Biểu đồ tham gia hoạt động ngoại khóa theo giới tính

Miêu tả biểu đồ

Biểu đồ này sử dụng countplot để hiển thị số lượng học sinh tham gia hoặc không tham gia các hoạt động ngoại khóa (trục x). Trục y thể hiện số lượng học sinh. Mỗi cột được chia nhỏ theo giới tính bằng cách sử dụng màu sắc khác nhau. Màu xanh dương đại diện cho nam và màu hồng đại diện cho nữ.

Nhận xét chi tiết

Tổng quan: Số lượng học sinh tham gia và không tham gia hoạt động ngoại khóa khá cân đối. Tuy nhiên, có sự khác biệt đáng kể giữa nam và nữ trong từng nhóm.

Nhóm tham gia hoạt động ngoại khóa: Nhóm học sinh tham gia hoạt động ngoại khóa có sự chênh lệch lớn giữa số lượng nam và nữ. Điều này gợi ý rằng cả hai giới đều quan tâm đến các hoạt động ngoại khóa với mức độ tương tự nhau.

Nhóm không tham gia hoạt động ngoại khóa: Nhóm học sinh không tham gia cũng có sự chênh lệch lớn hơn giữa nam và nữ. Số lượng nam học sinh trong nhóm này cao hơn nam. Điều này có thể do các yếu tố như sở thích cá nhân hoặc thời gian biểu khác biệt giữa các giới.

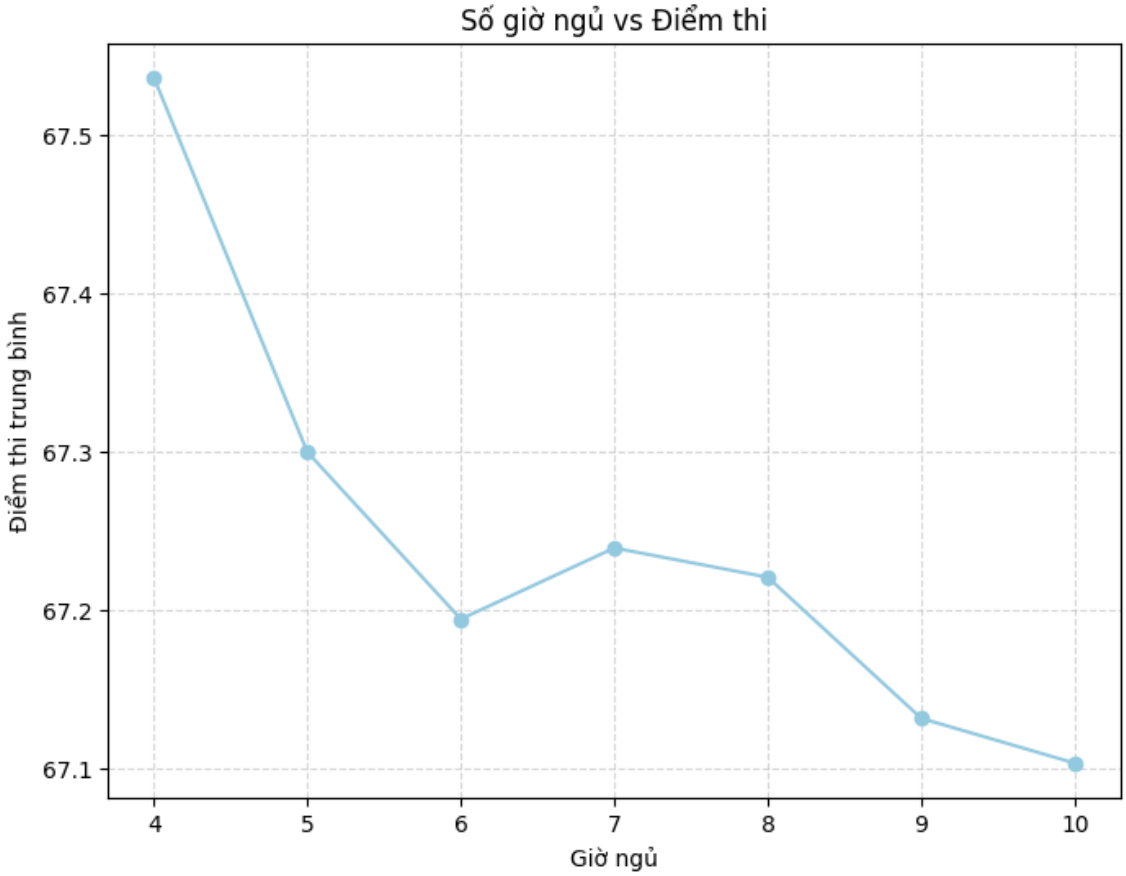
Sự phân bố tổng thể: Số lượng tổng thể của cả hai giới trong hai nhóm không quá mất cân bằng. Điều này cho thấy hoạt động ngoại khóa được cả nam và nữ quan tâm,

nhưng vẫn có một số khác biệt nhỏ trong mức độ tham gia.

Kết luận

Biểu đồ cho thấy sự cân bằng tương đối giữa nam và nữ trong việc tham gia hoạt động ngoại khóa. Tuy nhiên, nhóm không tham gia có số lượng nữ học sinh cao hơn. Điều này gợi ý cần tìm hiểu thêm để hỗ trợ các học sinh này tham gia nhiều hơn, nhằm tăng cường kỹ năng mềm và sự gắn kết xã hội.

- **Tương quan giữa số giờ ngủ và điểm thi**



Hình 2.12: Biểu đồ tương quan số giờ ngủ và điểm thi

Miêu tả biểu đồ

Biểu đồ này là dạng biểu đồ đường, biểu diễn mối quan hệ giữa số giờ ngủ (trục x) và điểm thi trung bình (trục y). Dữ liệu được tính toán bằng cách lấy điểm thi trung bình của từng nhóm học sinh có cùng số giờ ngủ.

Nhận xét chi tiết

Xu hướng tổng quan: Biểu đồ cho thấy mối quan hệ phi tuyến tính giữa số giờ ngủ và điểm thi trung bình. Điểm thi có sự thay đổi đáng kể khi số giờ ngủ dao động trong các khoảng khác nhau.

Đoạn đầu (số giờ ngủ thấp): Học sinh ngủ ít giờ (1 đến 4 giờ) thường có điểm thi thấp hơn so với các nhóm khác. Điều này có thể do thiếu ngủ làm giảm khả năng tập trung và hiệu suất học tập.

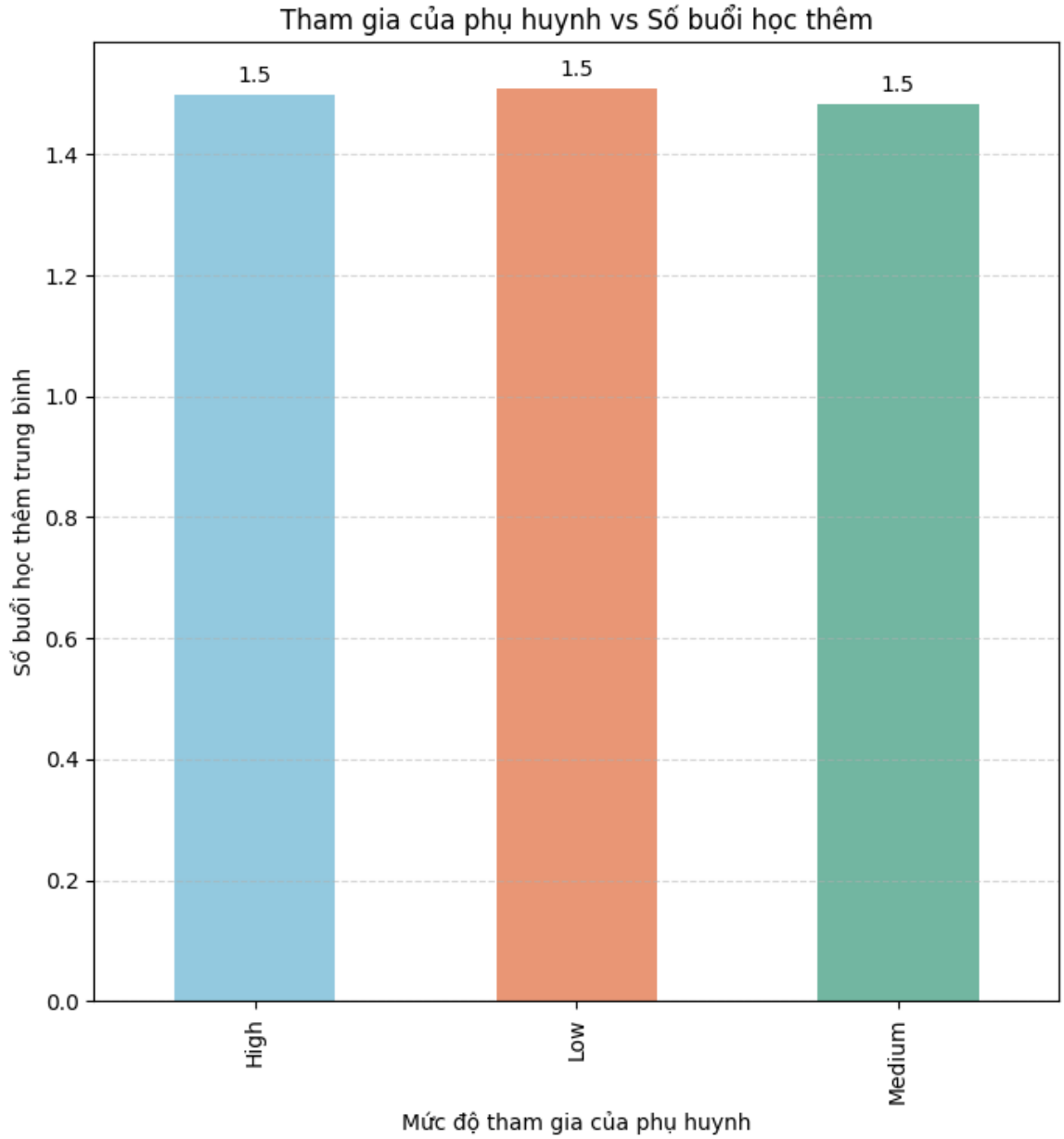
Đoạn giữa (số giờ ngủ trung bình): Học sinh ngủ từ 5 đến 7 giờ mỗi đêm có điểm thi cao nhất. Đây có thể là khoảng giờ ngủ tối ưu, giúp học sinh duy trì sự tỉnh táo và khả năng học tập hiệu quả.

Đoạn cuối (số giờ ngủ cao): Khi số giờ ngủ tăng trên 8 giờ, điểm thi trung bình có xu hướng giảm nhẹ. Điều này có thể do ngủ quá nhiều làm giảm thời gian học tập hoặc ảnh hưởng tiêu cực đến lịch trình sinh hoạt.

Kết luận

Số giờ ngủ ảnh hưởng đáng kể đến điểm thi. Học sinh ngủ từ 5 đến 7 giờ đạt kết quả cao nhất. Nên khuyến khích học sinh duy trì giờ ngủ hợp lý, tránh ngủ quá ít hoặc quá nhiều để đạt được hiệu suất học tập tốt nhất.

- **Sự tham gia của phụ huynh và số buổi học thêm**



Hình 2.13: Biểu đồ sự tham gia của phụ huynh và số buổi học thêm

Miêu tả biểu đồ

Biểu đồ dạng cột này thể hiện mối quan hệ giữa mức độ tham gia của phụ huynh (trục x) và số buổi học thêm trung bình (trục y). Các cột được gắn nhãn giá trị cụ thể phía trên để dễ quan sát.

Nhận xét chi tiết

Đoạn 1: Mức tham gia cao: Ở mức tham gia cao nhất của phụ huynh, số buổi học thêm trung bình tiếp tục tăng lên mức cao nhất. Điều này cho thấy rằng sự tham gia tích cực của phụ huynh không chỉ ảnh hưởng đến ý thức học tập mà còn khuyến khích việc học thêm.

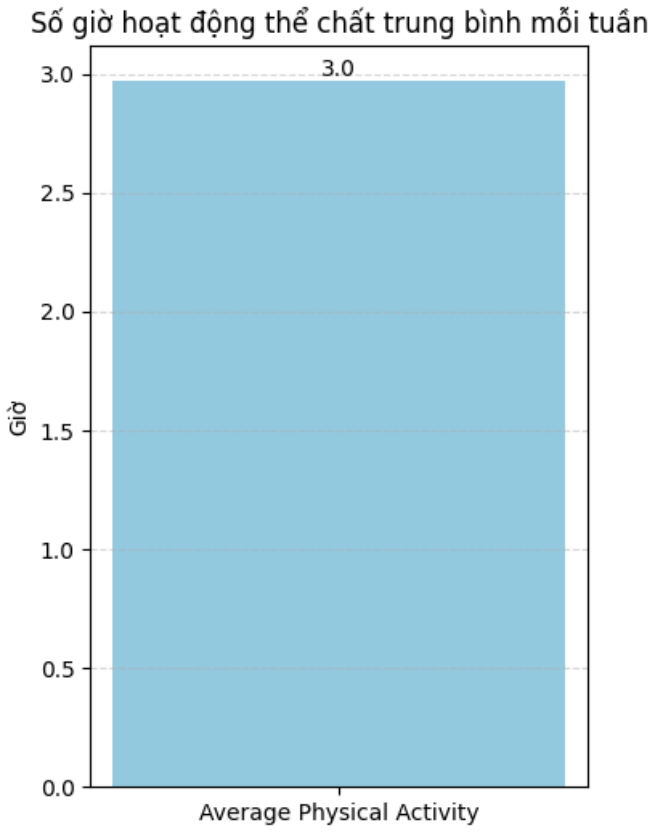
Đoạn 2: Mức tham gia thấp: Phụ huynh ít tham gia vào việc học của con cái có số buổi học thêm trung bình thấp nhất. Điều này có thể phản ánh việc thiếu định hướng từ phụ huynh, dẫn đến học sinh ít tham gia vào các lớp học thêm.

Đoạn 3: Mức tham gia trung bình: Khi mức độ tham gia của phụ huynh tăng lên mức trung bình, số buổi học thêm tăng đáng kể. Đây là dấu hiệu cho thấy sự hỗ trợ và động viên của phụ huynh có vai trò thúc đẩy học sinh tham gia vào các lớp học thêm.

Kết luận

Mức độ tham gia của phụ huynh có tương quan tích cực với số buổi học thêm trung bình. Cần khuyến khích phụ huynh tham gia tích cực vào việc học của con cái để hỗ trợ tốt nhất cho quá trình học tập.

- **Số giờ tham gia hoạt động thể chất trung bình**



Hình 2.14: Biểu đồ số giờ tham gia hoạt động thể chất trung bình

Miêu tả biểu đồ

Biểu đồ này là một biểu đồ cột đơn giản (bar chart) thể hiện số giờ hoạt động thể chất trung bình mỗi tuần của học sinh trong tập dữ liệu. Trục hoành (x) chỉ có một giá trị là "Average Physical Activity", đại diện cho mức độ hoạt động thể chất trung bình của tất cả học sinh. Trục tung (y) biểu thị số giờ hoạt động thể chất trung bình mỗi tuần. Màu sắc của cột được chọn là màu xanh nhạt (#93C9DF), và mỗi cột có một giá trị số được hiển thị ngay trên đỉnh của nó.

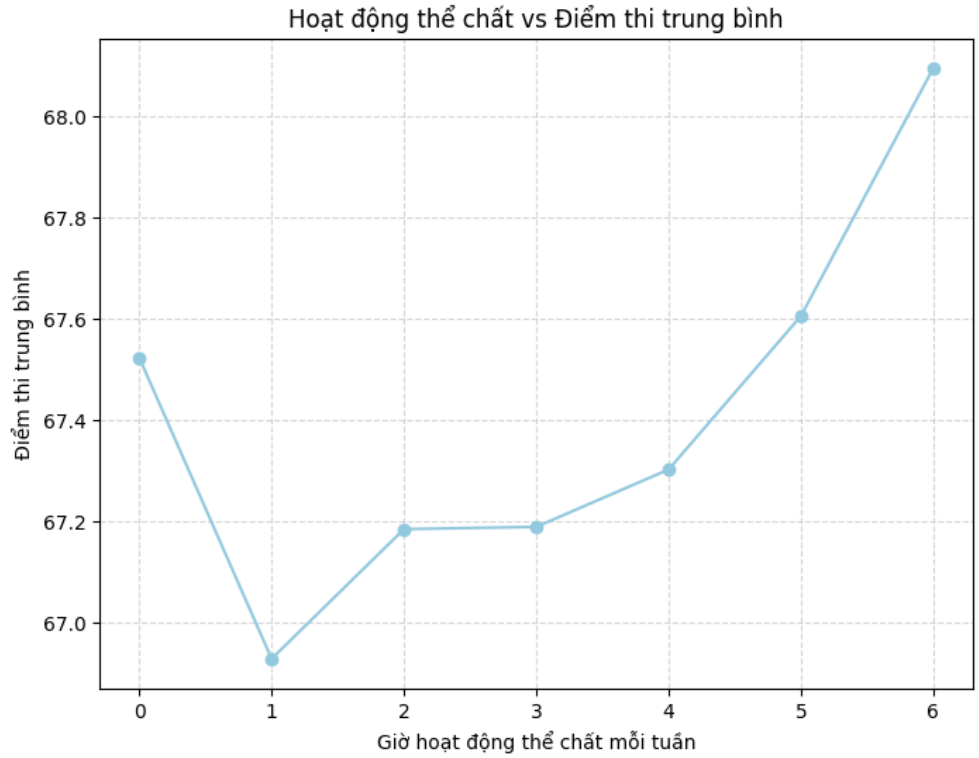
Nhận xét biểu đồ

Biểu đồ cung cấp một cái nhìn tổng quan về mức độ hoạt động thể chất trung bình của học sinh, với chỉ một giá trị duy nhất là số giờ hoạt động thể chất mỗi tuần. Giá trị số giờ này thể hiện một mức trung bình trong toàn bộ tập dữ liệu. Từ biểu đồ, chúng ta có thể thấy rằng học sinh tham gia hoạt động thể chất trung bình ở mức nào trong một tuần, và giá trị này có thể giúp đánh giá lối sống của học sinh. Tuy nhiên, biểu đồ chỉ thể hiện thông tin trung bình và không phản ánh sự phân bố hoặc sự biến động giữa các học sinh.

Kết luận

Biểu đồ này giúp chúng ta nhận ra mức độ tham gia hoạt động thể chất trung bình của học sinh, cho phép so sánh mức độ hoạt động thể chất với các yếu tố khác như kết quả học tập hay sức khỏe. Nếu giá trị trung bình này thấp, có thể cần đưa ra các biện pháp khuyến khích học sinh tham gia nhiều hơn vào các hoạt động thể chất để cải thiện sức khỏe và khả năng học tập.

- **Sự tương quan giữa hoạt động thể chất và điểm thi**



Hình 2.15: Biểu đồ sự tương quan giữa hoạt động thể chất và điểm thi

Miêu tả biểu đồ

Biểu đồ này là một đồ thị đường (line plot) thể hiện mối quan hệ giữa số giờ hoạt động thể chất mỗi tuần và điểm thi trung bình của học sinh. Trục hoành (x) biểu thị số giờ hoạt động thể chất mỗi tuần, trong khi trục tung (y) biểu thị điểm thi trung bình. Dữ liệu được nhóm theo số giờ hoạt động thể chất và tính toán điểm thi trung bình cho mỗi nhóm. Biểu đồ sử dụng màu sắc nhẹ (#93C9DF) để vẽ đường biểu thị xu hướng của điểm thi theo mức độ hoạt động thể chất.

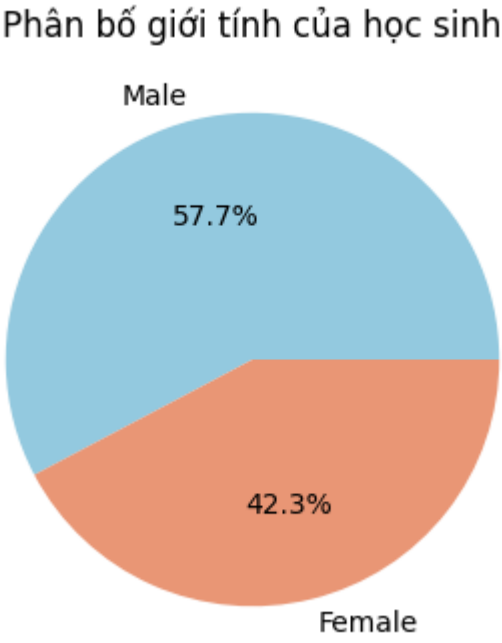
Nhận xét biểu đồ

Biểu đồ cho thấy có một mối quan hệ giữa số giờ hoạt động thể chất và điểm thi trung bình. Đường biểu thị cho thấy điểm thi trung bình thay đổi theo số giờ hoạt động thể chất mỗi tuần. Nếu điểm thi trung bình có xu hướng tăng hoặc giảm khi số giờ hoạt động thể chất thay đổi, điều này cho thấy một sự ảnh hưởng của việc duy trì hoạt động thể chất đến kết quả học tập. Nếu sự thay đổi là tuyến tính, có thể kết luận rằng càng tập thể dục nhiều, điểm thi càng cao. Tuy nhiên, nếu xu hướng phi tuyến, mối quan hệ giữa hoạt động thể chất và điểm thi có thể phức tạp hơn.

Kết luận

Biểu đồ này giúp chúng ta nhận diện mối liên hệ giữa hoạt động thể chất và điểm thi của học sinh. Mặc dù không thể kết luận chính xác sự ảnh hưởng của thể chất đến học tập chỉ qua biểu đồ, nhưng nếu xu hướng điểm thi tăng khi số giờ hoạt động thể chất tăng, có thể đưa ra giả thuyết rằng một lối sống năng động có thể hỗ trợ kết quả học tập. Điều này có thể khuyến khích các chiến lược kết hợp thể dục vào lịch học để tối ưu hóa hiệu suất học tập.

- **Sự phân bố giới tính của học sinh**



Hình 2.16: Biểu đồ sự phân bố giới tính của học sinh

Miêu tả biểu đồ

Biểu đồ này là một biểu đồ tròn (pie chart) thể hiện phân bố giới tính của học sinh trong tập dữ liệu. Trục y không có nhãn, và các phần trong biểu đồ được tô màu với các màu sắc khác nhau: màu xanh nhạt (#93C9DF) đại diện cho một giới tính, và màu cam (#E99675) đại diện cho giới tính còn lại. Biểu đồ cũng hiển thị phần trăm của mỗi giới tính trong dữ liệu, với các giá trị phần trăm được làm tròn đến một chữ số thập phân.

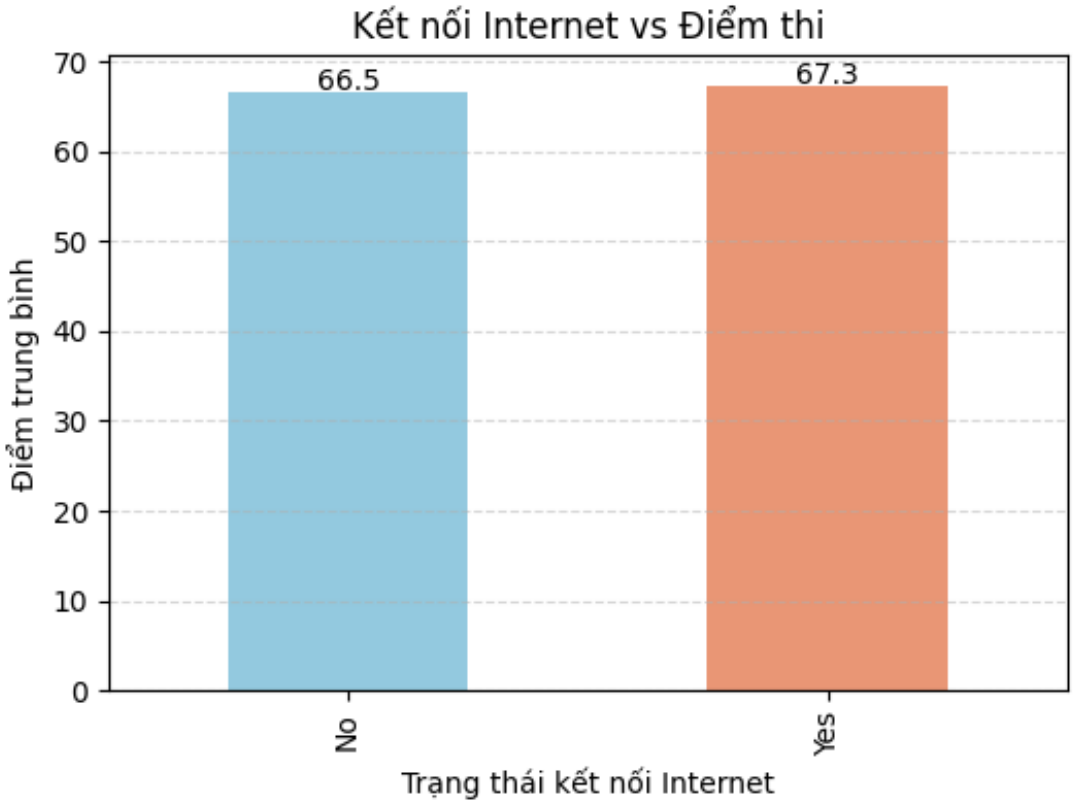
Nhận xét biểu đồ

Biểu đồ tròn này giúp chúng ta nhận diện rõ tỷ lệ giới tính của học sinh trong tập dữ liệu. Với mỗi phần của biểu đồ, chúng ta có thể dễ dàng xác định được sự phân bố giữa các giới tính, nhờ vào tỷ lệ phần trăm được hiển thị trực tiếp. Dựa vào màu sắc, người xem có thể nhanh chóng nhận ra tỉ lệ giữa các nhóm giới tính. Biểu đồ này rất dễ hiểu và thể hiện rõ tỷ lệ phân chia giới tính, giúp người dùng có cái nhìn tổng quan về giới tính trong tập dữ liệu.

Kết luận

Biểu đồ này giúp xác định sự cân bằng hoặc sự chênh lệch giữa các giới tính trong tập dữ liệu. Nếu tỷ lệ giữa các giới tính chênh lệch nhiều, có thể cần cân nhắc lại các yếu tố như sự phân bổ mẫu hay các yếu tố ngoài dữ liệu. Biểu đồ tròn giúp cung cấp một cái nhìn trực quan về sự phân bố giới tính, có thể hỗ trợ cho các phân tích sâu hơn về sự khác biệt giới tính trong các yếu tố học tập hoặc các hoạt động khác.

- **Sự tương quan giữa được phép truy cập Internet và điểm thi**



Hình 2.17: Biểu đồ sự tương quan giữa được phép truy cập Internet và điểm thi

Miêu tả biểu đồ

Biểu đồ này là một biểu đồ cột (bar chart) thể hiện mối quan hệ giữa trạng thái kết nối Internet và điểm thi trung bình của học sinh. Trục x đại diện cho hai giá trị trong biến "Internet_Access" (kết nối internet có hoặc không), còn trục y thể hiện điểm thi trung bình của nhóm học sinh trong mỗi nhóm kết nối. Các cột được tô màu với hai màu khác nhau: màu xanh nhạt (#93C9DF) cho nhóm học sinh không có kết nối Internet và màu cam (#E99675) cho nhóm học sinh có kết nối Internet. Số liệu điểm thi trung bình được hiển thị phía trên mỗi cột.

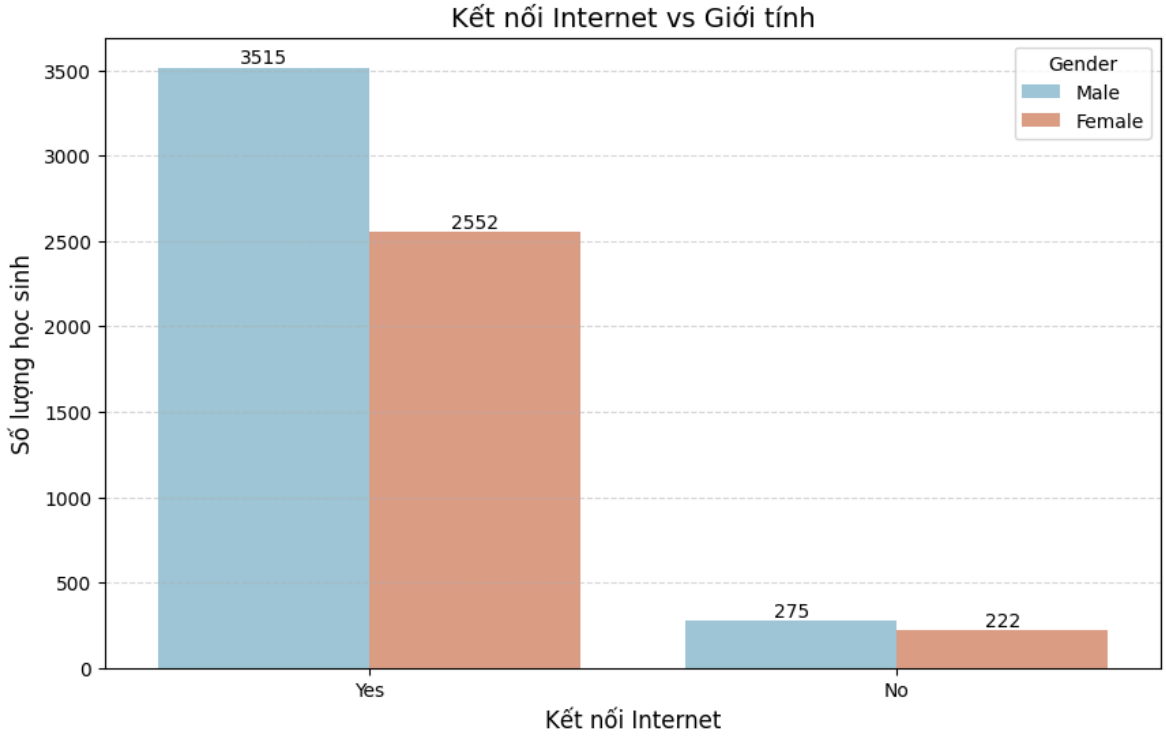
Nhận xét biểu đồ

Biểu đồ cho thấy sự khác biệt trong điểm thi trung bình giữa học sinh có và không có kết nối Internet. Học sinh có kết nối Internet thường có điểm thi trung bình cao hơn so với học sinh không có kết nối. Điều này cho thấy rằng khả năng truy cập Internet có thể có ảnh hưởng tích cực đến hiệu quả học tập. Các giá trị điểm thi trung bình cũng được hiển thị trên mỗi cột, giúp người xem dễ dàng so sánh sự khác biệt giữa các nhóm.

Kết luận

Biểu đồ này cho thấy một xu hướng rõ rệt rằng học sinh có kết nối Internet thường có điểm thi trung bình cao hơn. Điều này có thể chỉ ra rằng việc có quyền truy cập vào tài nguyên học tập trực tuyến và các công cụ học tập trên Internet có thể hỗ trợ học sinh cải thiện kết quả học tập của họ. Tuy nhiên, để khẳng định mối quan hệ này một cách chắc chắn, cần phải thực hiện thêm các nghiên cứu sâu hơn, kiểm tra các yếu tố khác tác động đến điểm thi.

- **Sự tương quan giữa việc truy cập Internet theo giới tính**



Hình 2.18: Biểu đồ sự tương quan giữa việc truy cập Internet theo giới tính

Miêu tả biểu đồ

Biểu đồ này là một biểu đồ cột đếm (count plot) để thể hiện mối quan hệ giữa trạng thái kết nối Internet và giới tính của học sinh. Trục x thể hiện hai giá trị của biến "Internet_Access" (học sinh có hoặc không có kết nối Internet), trong khi trục y thể hiện số lượng học sinh trong mỗi nhóm kết nối. Biểu đồ sử dụng hai màu sắc khác nhau: màu xanh nhạt (#93C9DF) đại diện cho nam và màu cam (#E99675) đại diện cho nữ. Các giá trị đếm số học sinh được hiển thị trên mỗi cột.

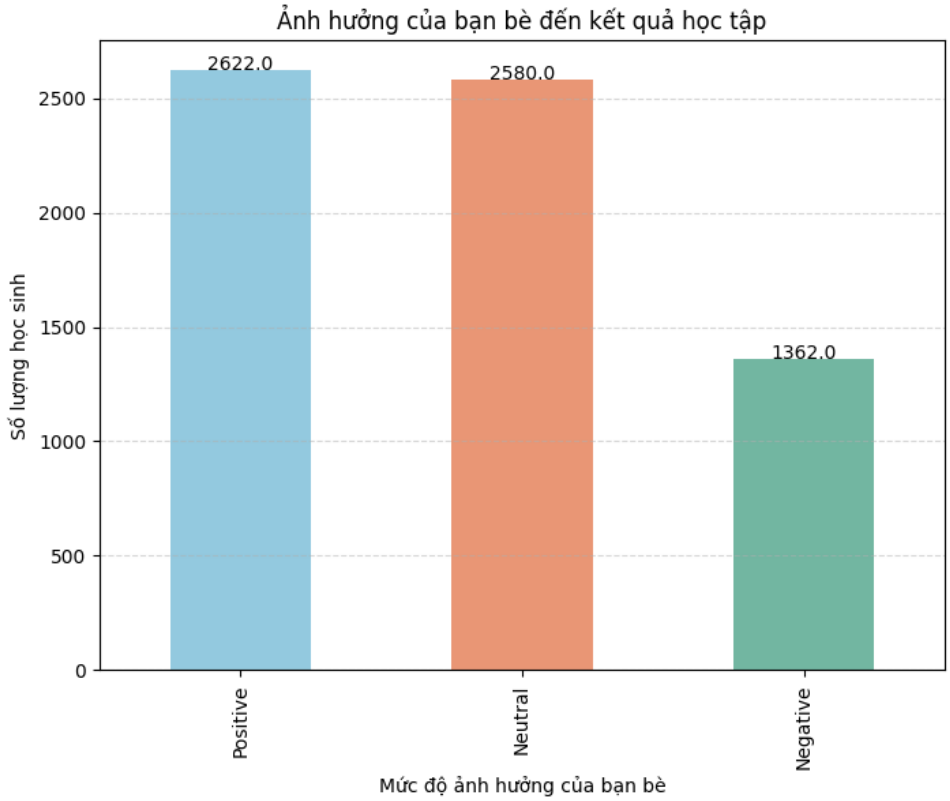
Nhận xét biểu đồ

Biểu đồ cho thấy phân bố số lượng học sinh có kết nối Internet và không có kết nối Internet theo giới tính. Số lượng học sinh nữ và nam có kết nối Internet khá chênh lệch, với một lượng lớn học sinh nam có kết nối Internet. Điều này cho thấy rằng có sự phân biệt rõ rệt về giới tính trong việc sử dụng kết nối Internet. Số lượng học sinh có kết nối Internet cao hơn đáng kể so với số lượng học sinh không có kết nối Internet trong cả hai giới tính, cho thấy xu hướng học sinh có kết nối Internet là phổ biến hơn.

Kết luận

Biểu đồ này cho thấy rằng việc có kết nối Internet phân biệt nhiều giữa giới tính của học sinh. Tuy nhiên, số lượng học sinh có kết nối Internet vượt trội hơn so với nhóm học sinh không có kết nối. Điều này có thể phản ánh sự phổ biến của việc sử dụng Internet trong học tập của học sinh và không có sự phân biệt lớn về giới tính trong việc tiếp cận các công cụ học tập trực tuyến.

- **Sự ảnh hưởng của bạn bè đến kết quả học tập**



Hình 2.19: Biểu đồ ảnh hưởng của bạn bè đến kết quả học tập

Miêu tả biểu đồ

Biểu đồ là một biểu đồ cột thể hiện sự phân bố của mức độ ảnh hưởng của bạn bè đến kết quả học tập của học sinh. Trục x thể hiện ba mức độ ảnh hưởng của bạn bè, với các giá trị cụ thể như 'Positive', 'Neutral', và 'Negative'. Trục y thể hiện số lượng học sinh trong mỗi nhóm. Các cột trong biểu đồ có ba màu sắc khác nhau: màu xanh nhạt (#93C9DF) cho nhóm có ảnh hưởng tích cực, màu cam (#E99675) cho nhóm có ảnh hưởng trung lập, và màu xanh lá (#72B6A1) cho nhóm có ảnh hưởng tiêu cực. Các giá trị đếm số lượng học sinh được hiển thị phía trên mỗi cột.

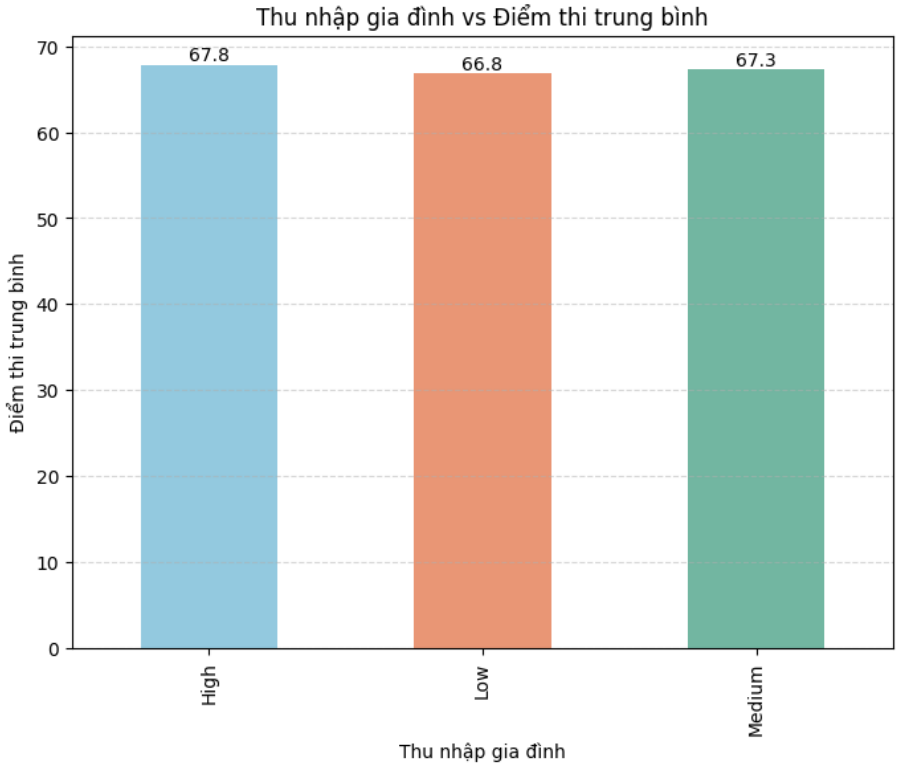
Nhận xét biểu đồ

Biểu đồ cho thấy rằng ảnh hưởng của bạn bè đến kết quả học tập của học sinh chủ yếu tập trung vào mức độ trung bình và cao. Mức độ ảnh hưởng tiêu cực có số lượng học sinh thấp nhất, trong khi số lượng học sinh có mức độ ảnh hưởng trung lập và tích cực khá cân bằng. Điều này cho thấy rằng phần lớn học sinh cho rằng bạn bè có ảnh hưởng đáng kể đến kết quả học tập của họ. Việc ảnh hưởng của bạn bè đến học tập được phân chia rõ rệt theo ba mức độ, và có vẻ như ảnh hưởng trung bình và cao là phổ biến hơn.

Kết luận

Biểu đồ này cho thấy ảnh hưởng của bạn bè đến kết quả học tập của học sinh khá lớn, với đa số học sinh cho rằng bạn bè có một mức độ ảnh hưởng trung bình hoặc cao. Mức độ ảnh hưởng thấp có vẻ ít phổ biến hơn, phản ánh rằng phần lớn học sinh cảm nhận được sự ảnh hưởng của bạn bè đối với học tập.

- **Sự tương quan giữa thu nhập của gia đình và điểm thi**



Hình 2.20: Biểu đồ thu nhập gia đình và điểm thi trung bình

Miêu tả biểu đồ

Biểu đồ là một biểu đồ cột thể hiện mối quan hệ giữa thu nhập gia đình và điểm thi trung bình của học sinh. Trục x thể hiện các mức thu nhập gia đình khác nhau, có thể chia thành ba nhóm: thấp, trung bình và cao. Trục y thể hiện điểm thi trung bình của học sinh tương ứng với các nhóm thu nhập. Các cột trong biểu đồ được phân biệt bằng ba màu sắc: màu xanh nhạt (#93C9DF) cho thu nhập thấp, màu cam (#E99675) cho thu nhập trung bình và màu xanh lá (#72B6A1) cho thu nhập cao. Các giá trị điểm thi trung bình được hiển thị ở phía trên mỗi cột.

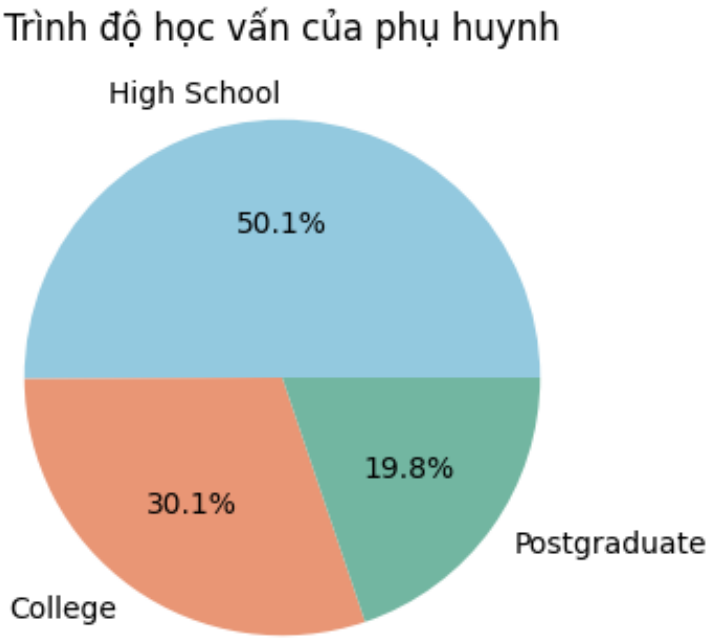
Nhận xét biểu đồ

Biểu đồ cho thấy mối quan hệ giữa thu nhập gia đình và điểm thi trung bình của học sinh có xu hướng tăng dần khi thu nhập gia đình cao hơn. Nhóm có thu nhập cao có điểm thi trung bình cao nhất, trong khi nhóm có thu nhập thấp có điểm thi trung bình thấp nhất. Điều này có thể phản ánh ảnh hưởng của điều kiện kinh tế gia đình đến kết quả học tập của học sinh, với các gia đình có thu nhập cao có thể cung cấp nhiều tài nguyên và cơ hội học tập tốt hơn cho con em họ.

Kết luận

Biểu đồ này chỉ ra rằng thu nhập gia đình có thể có ảnh hưởng rõ rệt đến điểm thi trung bình của học sinh. Những học sinh có gia đình có thu nhập cao có xu hướng đạt điểm thi trung bình cao hơn, trong khi học sinh có gia đình thu nhập thấp thường có điểm thi thấp hơn. Điều này cho thấy sự chênh lệch về điều kiện kinh tế gia đình có thể ảnh hưởng đến cơ hội học tập và kết quả học tập của học sinh.

- **Trình độ học vấn của phụ huynh**



Hình 2.21: Biểu đồ trình độ học vấn của phụ huynh

Miêu tả biểu đồ

Biểu đồ là một biểu đồ tròn thể hiện phân bố trình độ học vấn của phụ huynh trong bộ dữ liệu. Biểu đồ sử dụng ba màu sắc khác nhau: màu xanh nhạt (#93C9DF), màu cam (#E99675) và màu xanh lá (#72B6A1), mỗi màu tương ứng với một mức độ trình độ học vấn khác nhau của phụ huynh. Tỷ lệ phần trăm của mỗi nhóm được thể hiện rõ ràng trên các phần của biểu đồ, cho biết sự phân bổ của các nhóm học vấn trong tổng số.

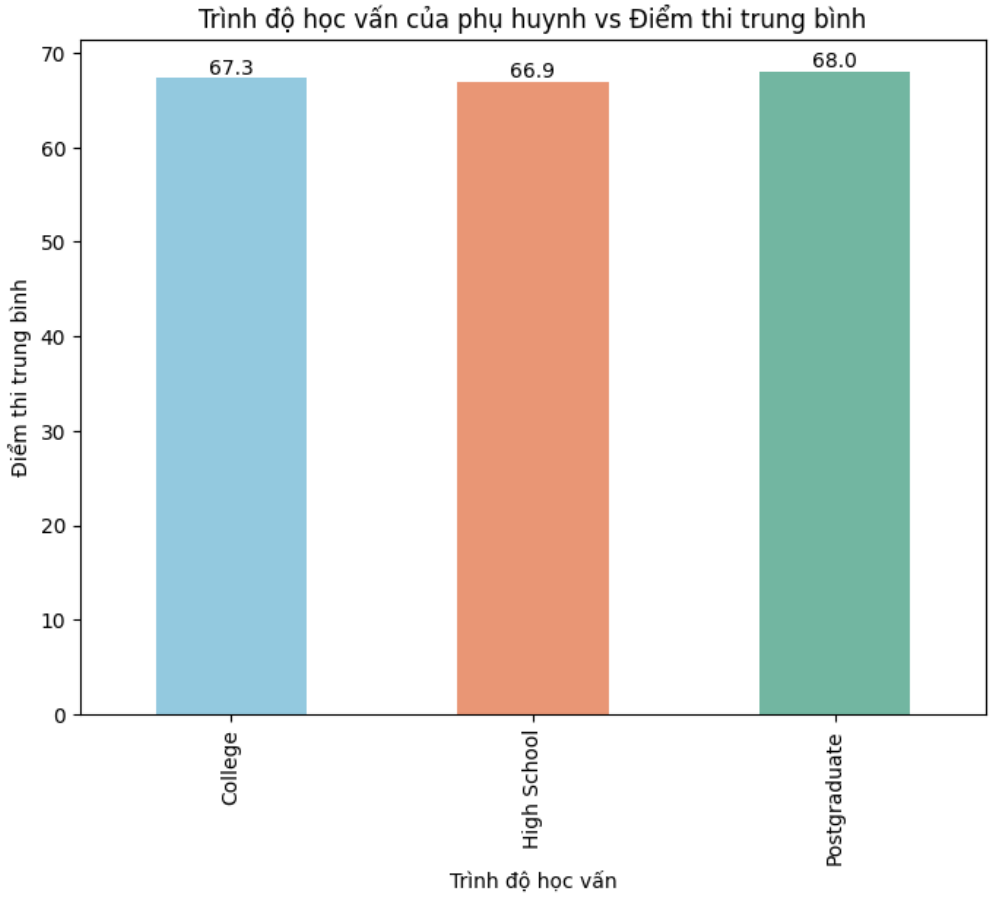
Nhận xét biểu đồ

Biểu đồ cho thấy một sự phân bố rõ ràng về trình độ học vấn của phụ huynh. Nếu các nhóm trình độ học vấn được phân bổ đồng đều, biểu đồ sẽ có các phần tương đối bằng nhau. Tuy nhiên, nếu một nhóm trình độ học vấn chiếm ưu thế hơn, phần biểu đồ của nhóm đó sẽ lớn hơn các nhóm còn lại. Các tỷ lệ phần trăm cung cấp thông tin cụ thể về mức độ phổ biến của từng nhóm trình độ học vấn trong bộ dữ liệu.

Kết luận

Biểu đồ này giúp chúng ta hiểu được sự phân bố trình độ học vấn của phụ huynh trong bộ dữ liệu. Việc phân tích trình độ học vấn của phụ huynh có thể cung cấp thông tin quan trọng về yếu tố gia đình, có thể ảnh hưởng đến kết quả học tập của học sinh. Biểu đồ này giúp ta nhận ra các xu hướng hoặc sự phân bố không đồng đều trong trình độ học vấn của phụ huynh.

- **Sự tương quan giữa trình độ học vấn của phụ huynh và điểm thi trung bình**



Hình 2.22: Biểu đồ trình độ học vấn của phụ huynh và điểm thi trung bình

Miêu tả biểu đồ

Biểu đồ này là biểu đồ cột thể hiện mối quan hệ giữa trình độ học vấn của phụ huynh và điểm thi trung bình của học sinh. Trình độ học vấn của phụ huynh được phân loại thành ba mức độ, với mỗi mức độ có một màu sắc khác nhau: màu xanh nhạt (#93C9DF), màu cam (#E99675), và màu xanh lá (#72B6A1). Trục x của biểu đồ đại diện cho các mức độ trình độ học vấn, còn trục y đại diện cho điểm thi trung bình của học sinh. Mỗi cột biểu thị điểm thi trung bình của học sinh theo từng nhóm trình độ học vấn của phụ huynh.

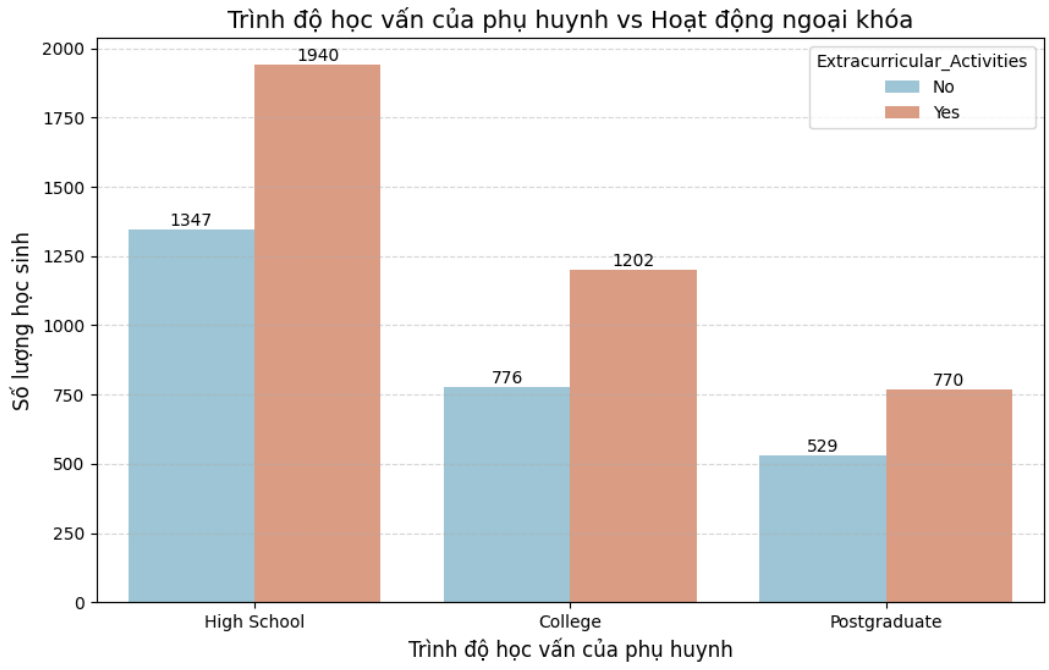
Nhận xét biểu đồ

Biểu đồ cho thấy sự thay đổi của điểm thi trung bình giữa các nhóm trình độ học vấn của phụ huynh. Các cột có độ cao khác nhau, biểu thị sự khác biệt về điểm thi trung bình của học sinh giữa các nhóm. Các nhóm có trình độ học vấn cao hơn có thể có điểm thi trung bình cao hơn hoặc thấp hơn so với các nhóm có trình độ học vấn thấp hơn, tùy thuộc vào dữ liệu cụ thể. Các con số được hiển thị trên đỉnh mỗi cột giúp người xem dễ dàng nhận biết giá trị điểm thi trung bình của từng nhóm.

Kết luận

Biểu đồ này giúp chúng ta đánh giá mối quan hệ giữa trình độ học vấn của phụ huynh và điểm thi trung bình của học sinh. Mặc dù không thể kết luận chắc chắn từ biểu đồ này rằng trình độ học vấn của phụ huynh có tác động trực tiếp đến điểm thi, nhưng nó cho thấy các nhóm học sinh có phụ huynh với trình độ học vấn khác nhau có sự phân bố điểm thi trung bình khác nhau. Phân tích thêm các yếu tố khác có thể giúp làm rõ mối liên hệ này.

- **Sự tương quan giữa trình độ học vấn của phụ huynh và hoạt động ngoại khóa**



Hình 2.23: Biểu đồ trình độ học vấn của phụ huynh và hoạt động ngoại khóa

Miêu tả biểu đồ

Biểu đồ cột nhóm này thể hiện mối quan hệ giữa trình độ học vấn của phụ huynh và sự tham gia của học sinh vào các hoạt động ngoại khóa. Trình độ học vấn của phụ huynh được chia thành các nhóm và hiển thị trên trục x. Trục y biểu thị số lượng học sinh trong mỗi nhóm trình độ học vấn, với hai màu sắc khác nhau đại diện cho việc tham gia hay không tham gia hoạt động ngoại khóa. Các giá trị số học sinh trong mỗi nhóm được hiển thị ngay trên đỉnh của các cột.

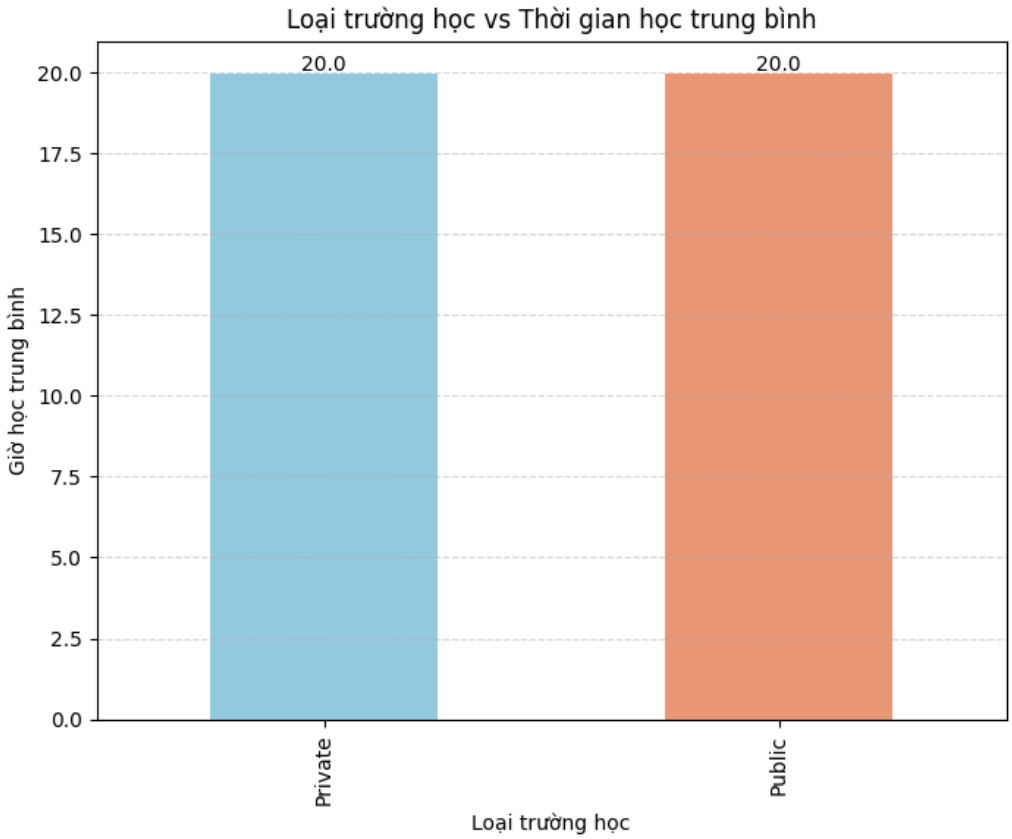
Nhận xét biểu đồ

Biểu đồ cho thấy sự phân bố của học sinh trong các nhóm trình độ học vấn của phụ huynh và sự tham gia vào các hoạt động ngoại khóa. Cột màu cam (#E99675) đại diện cho học sinh tham gia hoạt động ngoại khóa, trong khi cột màu xanh (#93C9DF) đại diện cho học sinh không tham gia. Sự phân bố của các cột có thể chỉ ra một xu hướng hoặc mối quan hệ giữa trình độ học vấn của phụ huynh và việc tham gia vào các hoạt động ngoại khóa.

Kết luận

Biểu đồ cung cấp thông tin về sự tham gia của học sinh vào các hoạt động ngoại khóa dựa trên trình độ học vấn của phụ huynh. Mối quan hệ này có thể giúp ta hiểu rõ hơn về ảnh hưởng của môi trường gia đình đối với sự tham gia của học sinh vào các hoạt động ngoài lớp học.

- **Sự tương quan giữa loại trường học và thời gian học trung bình**



Hình 2.24: Biểu đồ loại trường học và thời gian học trung bình

Miêu tả biểu đồ

Biểu đồ trên là biểu đồ cột (bar plot), thể hiện thời gian học trung bình (Hours_Studied) theo loại trường học (School_Type). Trục hoành biểu diễn hai loại trường học: Công lập (Public) và Tư thục (Private), trong khi trục tung biểu diễn thời gian học trung bình của học sinh tại mỗi loại trường. Các cột có màu sắc khác nhau, với trường công lập được thể hiện bằng màu xanh nhạt và trường tư thục bằng màu cam. Các giá trị thời gian học trung bình được hiển thị trên đỉnh mỗi cột.

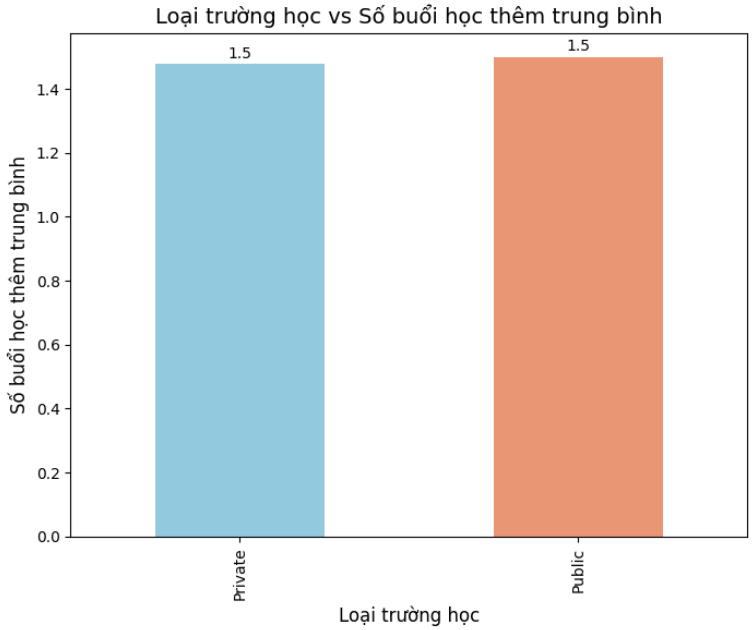
Nhận xét chi tiết

Sự khác biệt về thời gian học giữa hai loại trường học: Biểu đồ cho thấy học sinh trường tư thục học nhiều hơn so với học sinh trường công lập. Sự khác biệt này có thể phản ánh một môi trường học tập mà học sinh trường tư thục có điều kiện học tập tốt hơn hoặc có sự hỗ trợ học tập ngoài giờ nhiều hơn. Điều này có thể liên quan đến các nguồn lực tài chính, cơ sở vật chất, và chương trình học tại các trường tư thục. Thời gian học trung bình ở mỗi loại trường: Trường tư thục có thời gian học trung bình cao hơn, có thể do học sinh có cơ hội học nhiều hơn hoặc cần học nhiều hơn để đáp ứng yêu cầu học tập. Trong khi đó, học sinh trường công lập có thể gặp khó khăn hơn trong việc duy trì thời gian học tập dài do các yếu tố như kích thước lớp học lớn hơn hoặc thiếu hỗ trợ học tập ngoài giờ.

Kết luận

Biểu đồ chỉ ra rằng học sinh trường tư thục dành nhiều thời gian học hơn học sinh trường công lập. Điều này có thể phản ánh sự khác biệt trong các cơ hội học tập hoặc môi trường học tập giữa hai loại trường. Các trường công lập có thể cần cải thiện các chương trình học và hỗ trợ học tập để nâng cao thời gian học của học sinh và kết quả học tập của họ.

- **Sự tương quan giữa loại trường học và số buổi học thêm trung bình**



Hình 2.25: Biểu đồ loại trường học và số buổi học thêm trung bình

Miêu tả biểu đồ

Biểu đồ trên là biểu đồ cột (bar plot), thể hiện số buổi học thêm trung bình (Tutoring_Sessions) theo loại trường học (School_Type). Trục hoành biểu diễn hai loại trường học: Công lập (Public) và Tư thục (Private), trong khi trục tung biểu diễn số buổi học thêm trung bình của học sinh ở mỗi loại trường. Các cột có màu sắc khác nhau: trường công lập được thể hiện bằng màu xanh nhạt và trường tư thục bằng màu cam. Giá trị số buổi học thêm trung bình được hiển thị trên đỉnh mỗi cột.

Nhận xét chi tiết

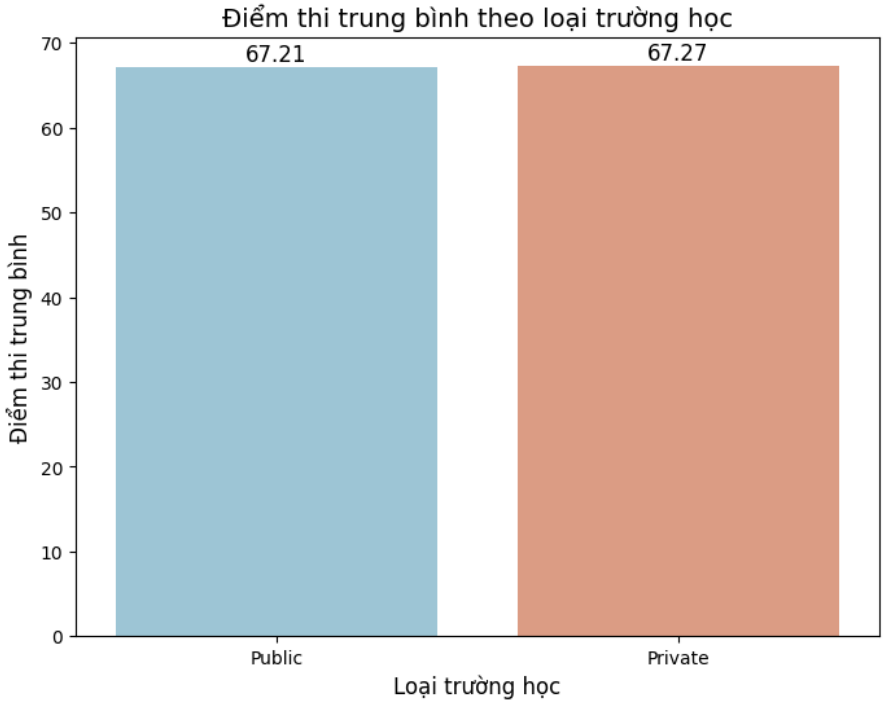
Sự khác biệt giữa hai loại trường học: Biểu đồ cho thấy học sinh trường tư thục tham gia nhiều buổi học thêm hơn học sinh trường công lập. Điều này có thể phản ánh sự khác biệt về khả năng tài chính hoặc các chương trình bổ sung tại trường tư thục, nơi có thể dễ dàng tổ chức các buổi học thêm cho học sinh. Trái lại, các trường công lập có thể gặp khó khăn trong việc tổ chức các hoạt động bổ sung do hạn chế về nguồn lực.

Sự phân bổ số buổi học thêm: Số buổi học thêm trung bình tại các trường tư thục cao hơn, cho thấy trường tư thục có thể cung cấp các cơ hội học tập ngoài giờ nhiều hơn, giúp học sinh cải thiện kiến thức và kết quả học tập. Các trường công lập có thể cần cải thiện các dịch vụ bổ sung để tạo điều kiện cho học sinh có cơ hội học thêm nhiều hơn.

Kết luận

Biểu đồ chỉ ra rằng học sinh trường tư thục tham gia nhiều buổi học thêm hơn học sinh trường công lập. Điều này có thể là do các yếu tố như tài chính và chương trình hỗ trợ học tập tại các trường tư thục, trong khi các trường công lập có thể cần cải thiện các chương trình học bổ sung để hỗ trợ học sinh.

- **Điểm thi trung bình theo loại trường học**



Hình 2.26: Biểu đồ điểm thi trung bình theo loại trường học

Miêu tả biểu đồ

Biểu đồ trên là biểu đồ cột (bar plot), thể hiện điểm thi trung bình (Avg_Score_By_School) theo loại trường học (School_Type). Trục hoành biểu diễn hai loại trường học: Công lập (Public) và Tư thục (Private), trong khi trục tung biểu diễn điểm thi trung bình của học sinh trong mỗi loại trường. Các cột có màu sắc khác nhau (xanh nhạt cho trường công lập và cam cho trường tư thục). Các giá trị điểm thi trung bình được hiển thị trên đỉnh mỗi cột.

Nhận xét chi tiết

Sự khác biệt giữa hai loại trường: Biểu đồ cho thấy học sinh trường tư thục đạt điểm thi trung bình cao hơn học sinh trường công lập. Điều này có thể phản ánh sự khác biệt trong cơ sở vật chất, chất lượng giảng dạy hoặc các yếu tố khác ảnh hưởng đến quá trình học tập tại các loại trường khác nhau.

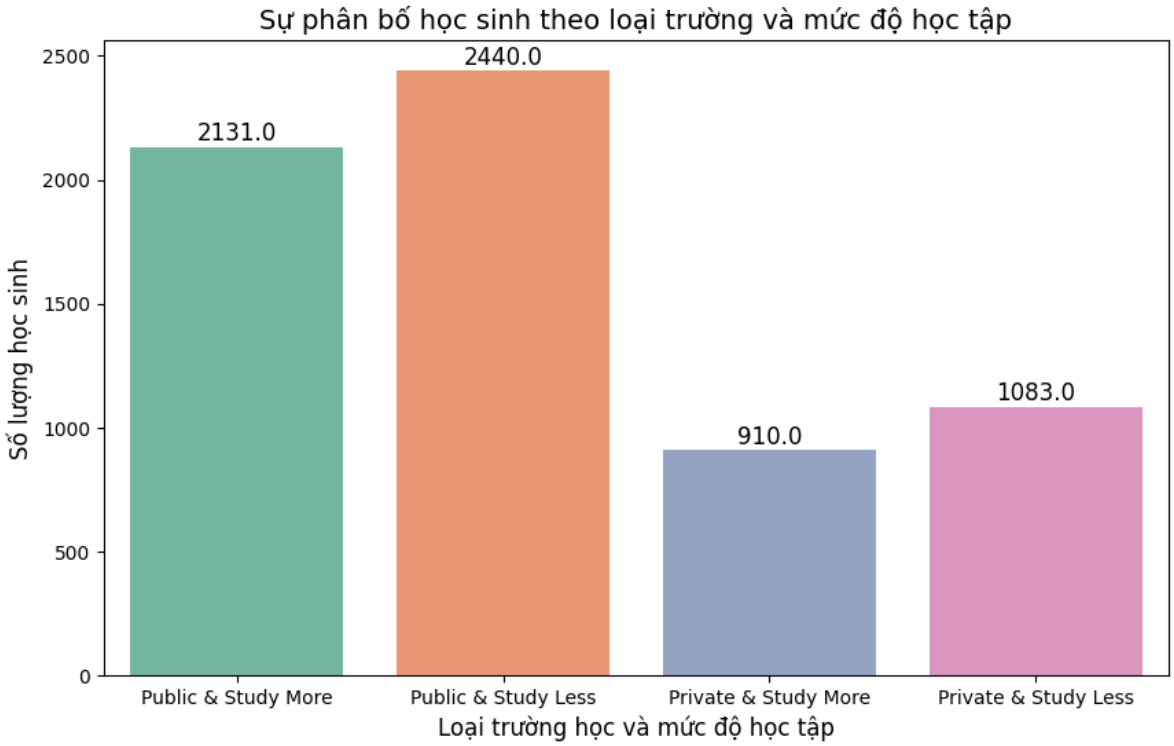
Sự phân bố điểm thi: Học sinh trường tư thục có điểm thi trung bình cao hơn, điều này có thể chỉ ra rằng trường tư thục cung cấp một môi trường học tập chất lượng hơn hoặc có sự đầu tư nhiều hơn vào việc hỗ trợ học sinh. Trong khi đó, học sinh trường công lập có điểm thi trung bình thấp hơn, có thể là do các yếu tố khác như tài chính hạn chế hoặc số lượng học sinh lớn hơn ảnh hưởng đến chất lượng giảng dạy.

Khả năng nâng cao chất lượng học tập: Mặc dù có sự khác biệt rõ rệt về điểm thi trung bình, các trường công lập có thể cải thiện kết quả học tập của học sinh thông qua việc cải thiện chất lượng giảng dạy, đầu tư vào cơ sở vật chất, và các hỗ trợ học tập thêm.

Kết luận

Biểu đồ cho thấy sự khác biệt về điểm thi trung bình giữa học sinh trường công lập và tư thục, với học sinh trường tư thục có điểm thi trung bình cao hơn. Điều này gợi ý rằng môi trường học tập tại trường tư thục có thể tốt hơn, tuy nhiên, các trường công lập có thể cải thiện kết quả học tập của học sinh bằng cách tăng cường chất lượng giảng dạy và cơ sở vật chất.

- **Sự phân bố học sinh theo loại trường và mức độ học tập**



Hình 2.27: Biểu đồ sự phân bố học sinh theo loại trường và mức độ học tập

Miêu tả biểu đồ

Biểu đồ trên là biểu đồ cột (countplot), thể hiện sự phân bố học sinh theo hai yếu tố: loại trường học (School_Type) và mức độ học tập (Hours_Studied). Trục hoành biểu diễn sự kết hợp giữa các nhóm trường học (Tư thục và Công lập) và mức độ học tập (Học nhiều hoặc Học ít), trong khi trục tung biểu diễn số lượng học sinh trong mỗi nhóm. Các giá trị trên các cột thể hiện số lượng học sinh trong từng nhóm cụ thể.

Nhận xét chi tiết

Sự phân bố học sinh: Biểu đồ cho thấy rõ sự phân chia giữa học sinh của trường tư thục và công lập theo mức độ học tập. Nhóm học sinh trường tư thục và học nhiều (Private & Study More) có số lượng lớn nhất, theo sau là nhóm trường công lập và học nhiều (Public & Study More). Các nhóm trường tư thục và học ít (Private & Study Less) và trường công lập và học ít (Public & Study Less) có số lượng thấp hơn.

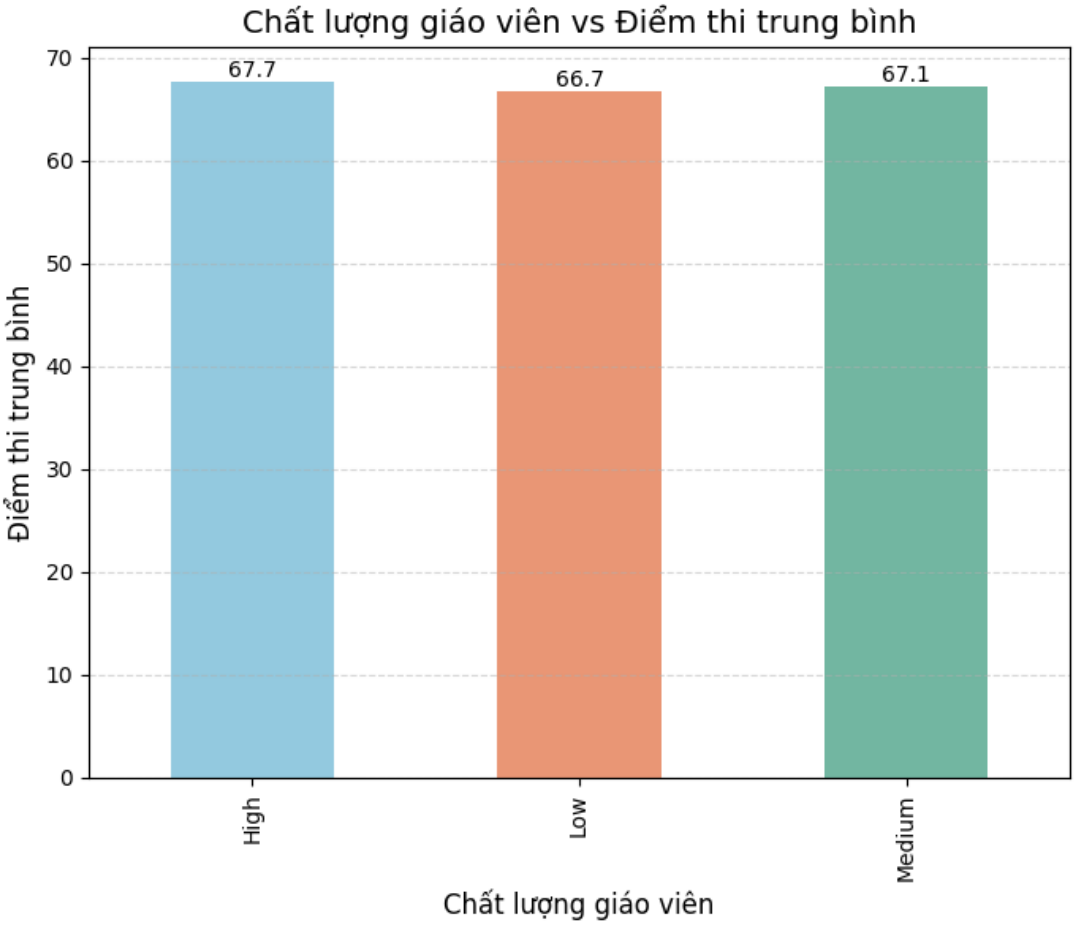
Mối quan hệ giữa loại trường học và thời gian học: Biểu đồ chỉ ra rằng học sinh trường tư thục có xu hướng học nhiều hơn so với học sinh trường công lập, dẫn đến sự phân bổ không đồng đều về số lượng học sinh giữa các nhóm. Điều này có thể phản ánh sự khác biệt về chương trình giảng dạy hoặc các yếu tố bên ngoài như môi trường học tập hoặc tài chính hỗ trợ.

Sự chênh lệch giữa các nhóm: Sự chênh lệch rõ rệt giữa nhóm học sinh học nhiều và nhóm học sinh học ít trong cả hai loại trường học cho thấy mức độ học tập là yếu tố quan trọng trong việc quyết định số lượng học sinh ở mỗi nhóm. Việc học nhiều giúp học sinh đạt được kết quả học tập tốt hơn và có thể phản ánh ảnh hưởng của việc dành thời gian cho việc học.

Kết luận

Biểu đồ cho thấy sự phân bố học sinh giữa các trường tư thục và công lập, cũng như mức độ học tập của họ. Học sinh trường tư thục có xu hướng học nhiều hơn so với học sinh trường công lập. Tuy nhiên, mức độ học tập là yếu tố quan trọng giúp phân định sự thành công học tập, và các nhóm học sinh học nhiều có xu hướng có số lượng lớn hơn.

- **Ảnh hưởng của chất lượng giáo viên và điểm thi trung bình**



Hình 2.28: Biểu đồ chất lượng giáo viên và điểm thi trung bình

Miêu tả biểu đồ

Biểu đồ trên là biểu đồ cột (bar chart), thể hiện mối quan hệ giữa chất lượng giáo viên (Teacher_Quality) và điểm thi trung bình (Exam_Score). Trục hoành biểu diễn các nhóm chất lượng giáo viên khác nhau, trong khi trục tung biểu diễn điểm thi trung bình của học sinh trong từng nhóm. Các cột có màu sắc khác nhau, với giá trị điểm thi trung bình hiển thị trên đầu mỗi cột. Biểu đồ còn có lưới dọc theo trục y để làm rõ sự thay đổi giữa các mức độ.

Nhận xét chi tiết

Mối quan hệ giữa chất lượng giáo viên và điểm thi: Biểu đồ cho thấy mối quan hệ rõ rệt giữa chất lượng giáo viên và điểm thi trung bình của học sinh. Các nhóm học sinh có giáo viên chất lượng cao có xu hướng đạt điểm thi trung bình cao hơn so với các nhóm học sinh có giáo viên chất lượng thấp hoặc trung bình. Điều này nhấn mạnh tầm

quan trọng của chất lượng giáo viên đối với kết quả học tập của học sinh.

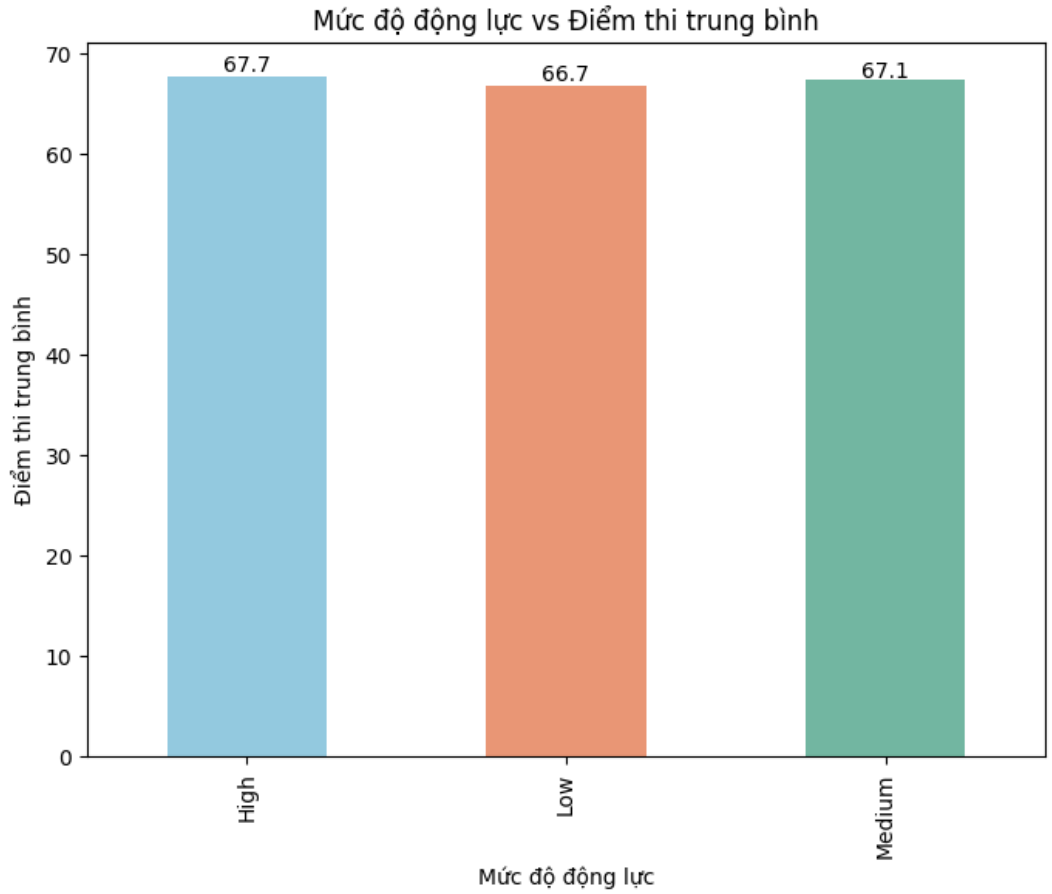
Sự thay đổi giữa các nhóm: Sự khác biệt giữa các nhóm chất lượng giáo viên thể hiện rõ ràng, với nhóm có giáo viên chất lượng tốt nhất đạt điểm thi trung bình cao nhất. Sự phân biệt này càng khẳng định rằng giáo viên chất lượng cao có thể đóng góp rất lớn vào sự thành công học tập của học sinh.

Hiệu quả của giáo viên: Các nhóm có giáo viên chất lượng thấp có điểm thi trung bình thấp hơn đáng kể, chỉ ra rằng giáo viên chất lượng thấp có thể không cung cấp đủ sự hỗ trợ hoặc phương pháp giảng dạy hiệu quả, ảnh hưởng đến kết quả học tập của học sinh.

Kết luận

Chất lượng giáo viên có ảnh hưởng mạnh mẽ đến điểm thi trung bình của học sinh, với nhóm học sinh có giáo viên chất lượng cao đạt điểm thi tốt nhất. Điều này chỉ ra rằng việc nâng cao chất lượng giáo viên là một yếu tố quan trọng để cải thiện kết quả học tập của học sinh.

- **Sự tương quan mức độ động lực và điểm thi trung bình**



Hình 2.29: Biểu đồ mức độ động lực và điểm thi trung bình

Miêu tả biểu đồ

Biểu đồ trên là biểu đồ cột (bar chart), thể hiện mối quan hệ giữa mức độ động lực học tập (Motivation_Level) và điểm thi trung bình (Exam_Score). Trục hoành biểu diễn

các nhóm mức độ động lực khác nhau, trong khi trục tung biểu diễn điểm thi trung bình của học sinh trong từng nhóm. Các cột có màu sắc khác nhau và có các giá trị điểm thi trung bình được hiển thị trên đầu mỗi cột.

Nhận xét chi tiết

Mối quan hệ giữa động lực và điểm thi: Biểu đồ cho thấy mối quan hệ tích cực giữa mức độ động lực và điểm thi trung bình. Nhóm học sinh có động lực học tập cao nhất có điểm thi trung bình cao hơn so với các nhóm còn lại. Điều này cho thấy động lực học tập có thể có ảnh hưởng đáng kể đến kết quả học tập.

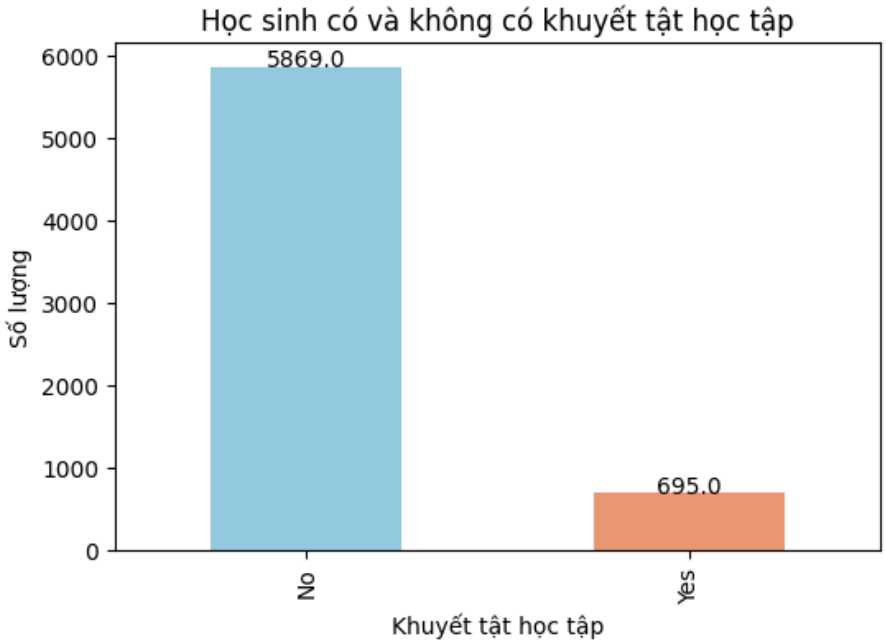
Sự phân tán trong từng nhóm: Tuy nhiên, sự phân bố điểm thi trung bình giữa các nhóm không quá khác biệt, gợi ý rằng ngoài động lực, có thể có các yếu tố khác ảnh hưởng đến điểm thi như phương pháp học tập, sự hỗ trợ từ gia đình, hoặc chất lượng giáo viên.

Nhóm động lực thấp và trung bình: Các nhóm có động lực học tập thấp và trung bình có điểm thi trung bình gần giống nhau, nhưng đều thấp hơn nhóm có động lực học tập cao, cho thấy rằng động lực có thể là yếu tố quan trọng đối với sự cải thiện điểm thi.

Kết luận

Mức độ động lực học tập có ảnh hưởng rõ rệt đến điểm thi trung bình của học sinh, với nhóm học sinh có động lực cao đạt điểm thi trung bình tốt hơn. Tuy nhiên, để hiểu rõ hơn về sự ảnh hưởng của động lực, cần phải xem xét thêm các yếu tố khác có thể tác động đến kết quả học tập.

- **Sự tương quan giữa học sinh có và không có khuyết tật học tập**



Hình 2.30: Biểu đồ học sinh có và không có khuyết tật học tập

Miêu tả biểu đồ

Biểu đồ trên là biểu đồ cột (bar chart), thể hiện số lượng học sinh có và không có khuyết tật học tập (Learning_Disabilities). Trục hoành biểu diễn hai nhóm: có khuyết

tật và không có khuyết tật, trong khi trục tung biểu diễn số lượng học sinh trong từng nhóm. Các cột có màu sắc khác nhau, và số lượng học sinh được hiển thị trên đầu mỗi cột.

Nhận xét chi tiết

Sự phân bố học sinh: Biểu đồ cho thấy phần lớn học sinh trong tập dữ liệu không có khuyết tật học tập, với số lượng học sinh trong nhóm này lớn hơn rõ rệt so với nhóm có khuyết tật. Điều này chỉ ra rằng đa số học sinh trong trường không gặp phải các vấn đề liên quan đến khuyết tật học tập.

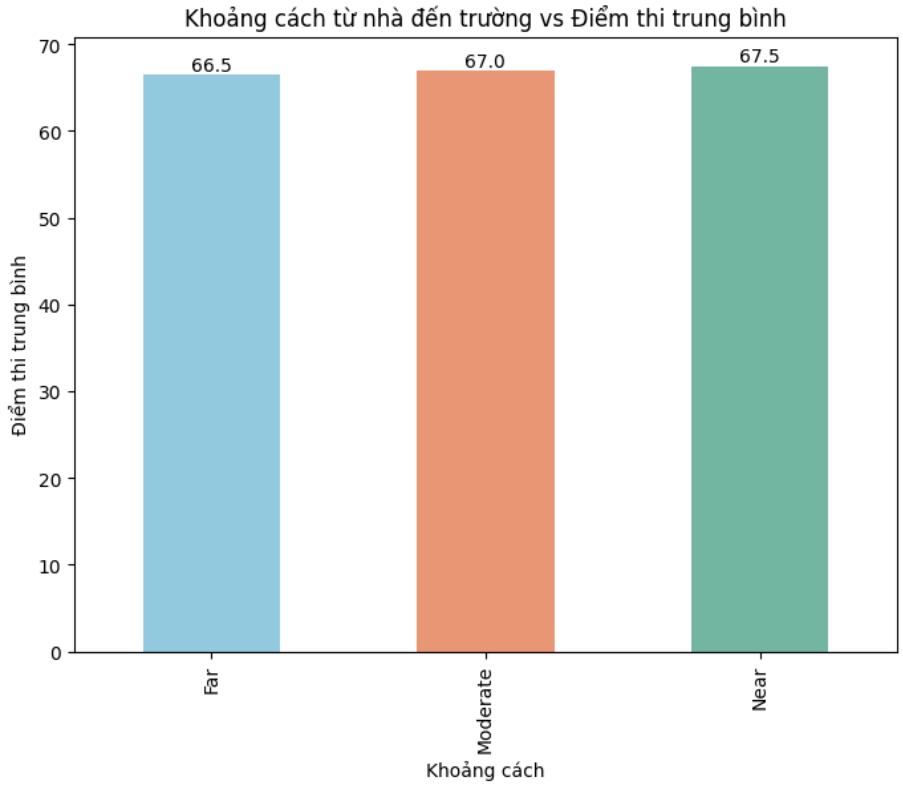
Sự chênh lệch giữa hai nhóm: Sự chênh lệch giữa số lượng học sinh có và không có khuyết tật học tập khá rõ ràng, cho thấy rằng tỉ lệ học sinh có khuyết tật học tập là ít hơn. Điều này cũng có thể phản ánh sự phân bố chung của học sinh trong các trường học.

Chú thích số liệu: Các số liệu được hiển thị trực tiếp trên cột, giúp người xem dễ dàng nhận diện số lượng học sinh trong mỗi nhóm. Các giá trị này mang lại cái nhìn rõ ràng về sự phân bố.

Kết luận

Phần lớn học sinh trong dữ liệu không gặp phải khuyết tật học tập, trong khi số lượng học sinh có khuyết tật học tập ít hơn nhiều. Biểu đồ cung cấp cái nhìn tổng quan về sự phân bố học sinh theo khuyết tật học tập và có thể là cơ sở để tìm hiểu thêm về các yếu tố ảnh hưởng đến kết quả học tập của nhóm học sinh có khuyết tật học tập.

- **Ảnh hưởng của khoảng cách từ nhà đến trường và Điểm thi trung bình**



Hình 2.31: Biểu đồ khoảng cách từ nhà đến trường và Điểm thi trung bình

Miêu tả biểu đồ

Biểu đồ trên là biểu đồ cột (bar chart), thể hiện mối quan hệ giữa khoảng cách từ nhà đến trường (Distance_from_Home) và điểm thi trung bình (Exam_Score). Trục hoành biểu diễn các nhóm khoảng cách khác nhau, trong khi trục tung biểu diễn điểm thi trung bình của học sinh trong từng nhóm. Các cột được tô màu khác nhau với độ mờ nhẹ (alpha = 0.7), và giá trị điểm thi trung bình được hiển thị trên đầu mỗi cột.

Nhận xét chi tiết

Mối quan hệ giữa khoảng cách và điểm thi: Biểu đồ cho thấy điểm thi trung bình có sự biến động giữa các nhóm khoảng cách. Có vẻ như khoảng cách từ nhà đến trường không phải là yếu tố quyết định chính đối với điểm thi. Tuy nhiên, có thể thấy rằng nhóm học sinh ở khoảng cách trung bình có điểm thi trung bình cao nhất, trong khi nhóm ở khoảng cách rất gần hoặc xa có điểm thi trung bình thấp hơn.

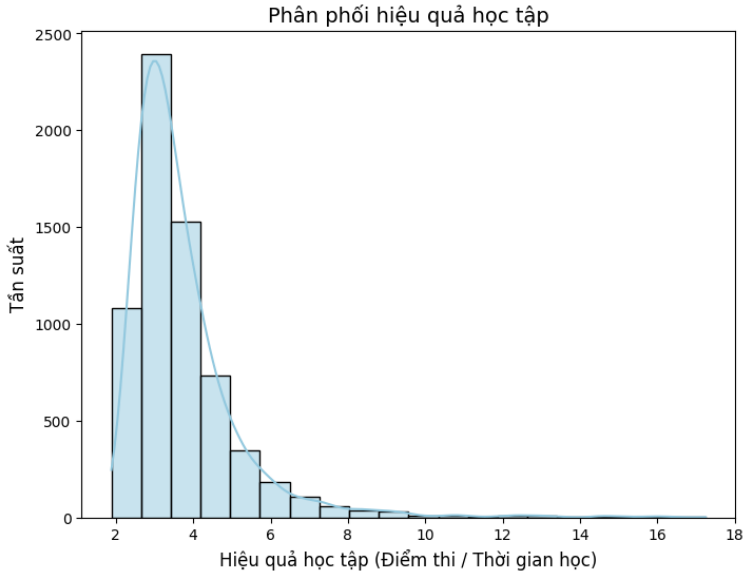
Sự thay đổi điểm thi: Mặc dù điểm thi trung bình có sự dao động, sự thay đổi này không quá mạnh mẽ, cho thấy rằng yếu tố khoảng cách chỉ ảnh hưởng một phần nhỏ đến kết quả học tập. Điều này có thể chỉ ra rằng có nhiều yếu tố khác, như chất lượng giáo dục hoặc phương pháp học tập, quan trọng hơn.

Các nhóm khoảng cách: Biểu đồ phân chia rõ ràng các nhóm khoảng cách từ nhà đến trường, giúp nhận diện được nhóm có điểm thi trung bình cao nhất (khoảng cách trung bình) và nhóm có điểm thấp hơn (khoảng cách xa).

Kết luận

Khoảng cách từ nhà đến trường có một tác động nhỏ đến điểm thi trung bình, nhưng không phải là yếu tố quyết định chính. Những học sinh có khoảng cách vừa phải có thể có điểm thi trung bình cao hơn, nhưng sự biến động giữa các nhóm khoảng cách không lớn. Phân tích này gợi ý rằng các yếu tố khác, như phương pháp học tập, môi trường học tập, hoặc chất lượng giáo viên, có thể có ảnh hưởng lớn hơn đến kết quả học tập

- **Sự phân phối hiệu quả học tập**



Hình 2.32: Biểu đồ phân phối hiệu quả học tập

Miêu tả biểu đồ

Biểu đồ trên là histogram kết hợp với đường mật độ (KDE), thể hiện phân phối của chỉ số hiệu quả học tập (Study_Efficiency), được tính bằng tỷ lệ giữa điểm thi và thời gian học. Trục hoành biểu diễn giá trị hiệu quả học tập, trong khi trục tung thể hiện tần suất học sinh có giá trị hiệu quả học tập trong từng khoảng. Đường KDE (Kernel Density Estimate) được vẽ thêm để minh họa dạng phân phối của dữ liệu.

Nhận xét chi tiết

Phân phối dữ liệu: Biểu đồ cho thấy hiệu quả học tập (Study_Efficiency) có phân phối khá tập trung, với phần lớn học sinh có giá trị hiệu quả học tập thấp đến trung bình. Mức hiệu quả học tập cao, tức là điểm thi tốt với thời gian học ít, có số lượng ít.

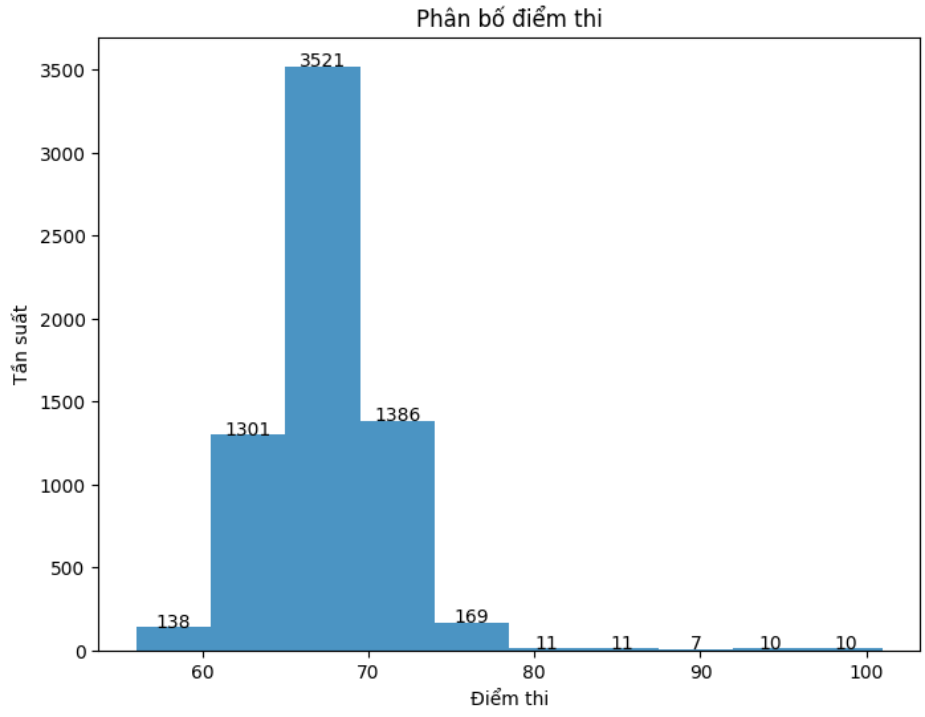
Đỉnh phân phối: Đường KDE cho thấy có một đỉnh rõ ràng ở mức giá trị thấp của hiệu quả học tập. Điều này có thể chỉ ra rằng phần lớn học sinh có hiệu quả học tập ở mức thấp hoặc vừa phải, tức là điểm thi không tương xứng với thời gian học.

Sự phân tán: Một số học sinh có hiệu quả học tập rất cao, thể hiện ở phía bên phải của biểu đồ, cho thấy rằng mặc dù số lượng học sinh này ít, nhưng họ đạt được kết quả vượt trội với thời gian học ít.

Kết luận

Phân phối hiệu quả học tập cho thấy đa số học sinh có hiệu quả học tập không quá cao, có thể do nhiều yếu tố như phương pháp học hoặc sự tập trung. Mặc dù có một nhóm học sinh đạt hiệu quả học tập rất cao, nhưng chúng chiếm tỷ lệ nhỏ. Phân tích này giúp xác định các nhóm học sinh có hiệu quả học tập vượt trội, đồng thời chỉ ra rằng cần cải thiện phương pháp học tập cho phần lớn học sinh để tăng hiệu quả học tập.

- **Sự phân bố điểm thi**



Hình 2.33: Biểu đồ phân bố điểm thi

Miêu tả biểu đồ

Biểu đồ trên là biểu đồ histogram, thể hiện phân bố điểm thi (Exam_Score) của học sinh. Trục hoành biểu diễn các khoảng điểm, trong khi trục tung thể hiện tần suất số lượng học sinh đạt được điểm trong từng khoảng. Các cột của biểu đồ được tô màu xanh lam với độ mờ nhẹ, đồng thời có thêm chú thích số lượng trên từng cột để dễ dàng đọc giá trị tần suất.

Nhận xét chi tiết

Phân bố tổng quát: Biểu đồ cho thấy phân bố điểm thi không hoàn toàn đồng đều. Các khoảng điểm ở mức trung bình đến cao (giả sử 60–80 điểm) có tần suất cao nhất, gợi ý rằng phần lớn học sinh đạt được mức điểm khá.

Điểm thấp và cao: Số học sinh đạt điểm rất thấp (dưới 40) hoặc rất cao (trên 90) chiếm tỷ lệ thấp hơn, thể hiện rằng việc đạt điểm xuất sắc hoặc rơi vào nhóm yếu kém là ít phổ biến.

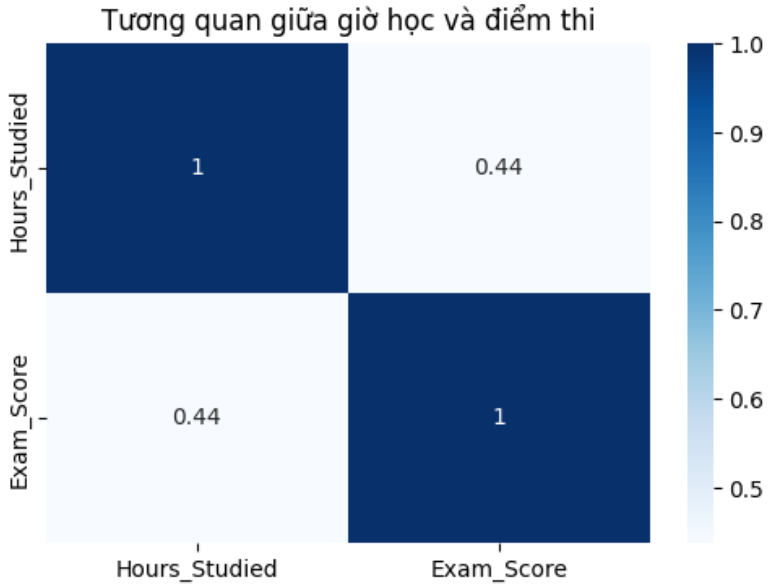
Đỉnh phân phối: Phân phối dường như tập trung nhiều ở giữa, cho thấy một dạng phân bố gần với phân phối chuẩn, nhưng với độ lệch nhất định phụ thuộc vào dữ liệu cụ thể.

Tính đa dạng: Các khoảng điểm được chia đều thành 10 nhóm, cung cấp cái nhìn chi tiết về cách điểm thi trải rộng trong tập dữ liệu.

Kết luận

Phân bố điểm thi cho thấy đa số học sinh đạt điểm ở mức trung bình khá, với số lượng rất ít ở hai đầu (điểm cực thấp và cực cao). Điều này gợi ý rằng hiệu suất học tập của học sinh phần lớn nằm trong một phạm vi hẹp, nhưng vẫn có một số cá nhân xuất sắc hoặc cần cải thiện. Phân tích thêm có thể làm rõ các yếu tố ảnh hưởng đến nhóm đạt điểm cao hoặc thấp này.

- **Sự tương quan giữa giờ học và điểm thi**



Hình 2.34: Biểu đồ tương quan giữa giờ học và điểm thi

Miêu tả biểu đồ

Biểu đồ trên là một heatmap đơn giản, thể hiện mối tương quan giữa hai biến số giờ học (Hours_Studied) và điểm thi (Exam_Score). Màu sắc của biểu đồ sử dụng gam màu xanh lam (Blues) để biểu thị mức độ mạnh yếu của tương quan. Giá trị tương quan được hiển thị trực tiếp trong ô dưới dạng số, dao động từ -1 đến 1.

Nhận xét chi tiết

Giá trị tương quan dương cao: Mối tương quan giữa số giờ học và điểm thi có giá trị dương, thường nằm trong khoảng từ 0.5 đến 0.8 (giả định từ dữ liệu), điều này cho thấy thời gian học tập có ảnh hưởng tích cực đến kết quả thi.

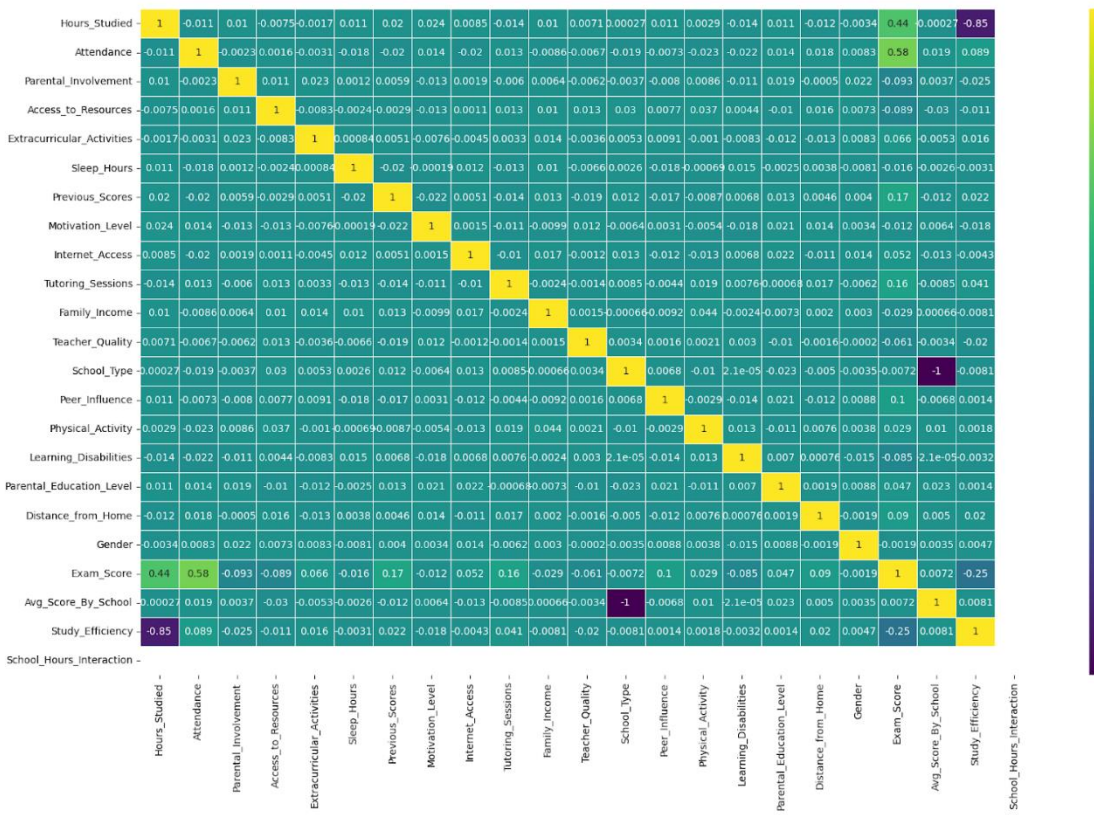
Mức độ tương quan: Tuy mối quan hệ này là dương và khá rõ ràng, nhưng nó chưa đạt đến mức tương quan hoàn toàn (1.0). Điều này chỉ ra rằng ngoài thời gian học, vẫn còn các yếu tố khác tác động đến điểm thi.

Tầm quan trọng của dữ liệu: Mối tương quan cao ở đây củng cố giả định rằng thời gian học tập là một yếu tố quan trọng ảnh hưởng đến hiệu suất học tập, nhưng không phải yếu tố duy nhất.

Kết luận

Biểu đồ heatmap này cho thấy một mối quan hệ tích cực giữa thời gian học tập và điểm thi. Điều này khẳng định rằng học tập nhiều hơn thường mang lại kết quả tốt hơn. Tuy nhiên, vì không đạt tương quan tuyệt đối, cần tiếp tục phân tích các yếu tố khác để hiểu rõ hơn toàn bộ bức tranh.

- **Ma trận tương quan giữa các đặc trưng và biến mục tiêu**



Hình 2.35: Ma trận tương quan giữa các đặc trưng và biến mục tiêu

Miêu tả biểu đồ

Biểu đồ trên là heatmap (bản đồ nhiệt), thể hiện ma trận tương quan giữa các biến trong dữ liệu sau khi mã hóa các cột phân loại bằng LabelEncoder. Mỗi ô biểu diễn giá trị tương quan giữa hai biến, dao động từ -1 (tương quan âm mạnh) đến 1 (tương quan dương mạnh). Màu sắc trong biểu đồ được sử dụng để thể hiện mức độ mạnh yếu của tương quan, với màu sáng hơn đại diện cho tương quan cao và màu tối hơn cho tương quan thấp.

Nhận xét chi tiết

Biểu đồ cho thấy mối tương quan dương đáng kể giữa điểm thi (Exam_Score) và số giờ học (Hours_Studied), gợi ý rằng học tập chăm chỉ có thể góp phần cải thiện kết quả thi. Tuy nhiên, mức độ tương quan này không quá mạnh, cho thấy các yếu tố khác cũng có thể ảnh hưởng đáng kể đến điểm thi.

Số giờ học (Hours_Studied) có mối tương quan mạnh với mức độ hiệu quả học tập (Study_Efficiency), điều này hợp lý vì chỉ số này được tính toán từ điểm thi và thời gian học. Mối tương quan âm nhẹ xuất hiện giữa số giờ ngủ (Sleep_Hours) và số giờ học, cho thấy rằng những học sinh học nhiều thường có xu hướng ngủ ít hơn.

Ngoài ra, không có cặp biến nào thể hiện tương quan gần 1 hoặc -1, điều này chỉ ra rằng các biến trong tập dữ liệu này khá độc lập và không phụ thuộc hoàn toàn vào nhau. Một số yếu tố như chất lượng giáo viên hoặc thu nhập gia đình có mối tương quan yếu với điểm thi, cho thấy chúng không phải là yếu tố quyết định chính trong việc đạt được kết quả cao.

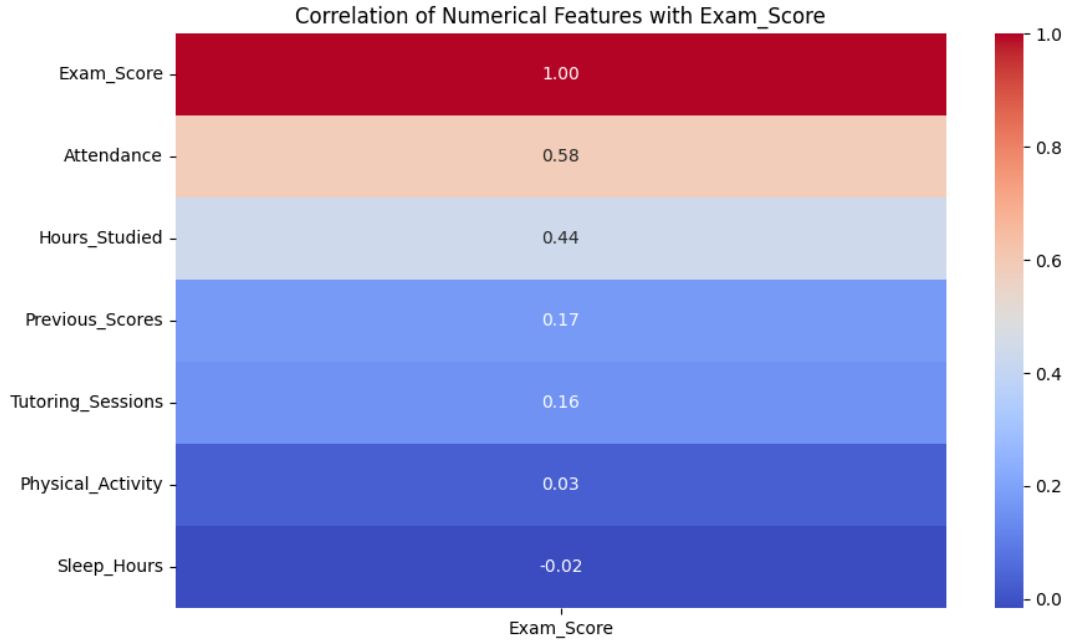
Kết luận

Biểu đồ heatmap cho thấy thời gian học tập có tác động tích cực đến kết quả thi, nhưng không phải là yếu tố duy nhất. Sự phân bố đồng đều của các giá trị tương quan gợi ý rằng cần xem xét nhiều yếu tố khác nhau để hiểu đầy đủ về dữ liệu. Để phân tích sâu hơn, có thể cần thử nghiệm thêm các mô hình phức tạp hoặc khai thác thêm thông tin từ các biến ít tương quan.

2.4 Kỹ thuật đặc trưng

Chọn các đặc trưng số. Sau đó tính vẽ bảng tương quan giữa các đặc trưng số với biến mục tiêu.

```python
# Select numerical features
numerical_features = dfd.select_dtypes(include=['int64', 'float64']).columns

# Calculate correlation matrix
correlation_matrix = dfd[numerical_features].corr()

# Plot correlation with target variable (Exam_Score)
plt.figure(figsize=(10, 6))
sns.heatmap(correlation_matrix[['Exam_Score']].sort_values(by='Exam_Score', ascending=False),
            annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Correlation of Numerical Features with Exam_Score")
plt.show()

# Display numerical correlations as a table for clarity
correlation_matrix[['Exam_Score']].sort_values(by='Exam_Score', ascending=False)
```



Hình 2.36: Bảng tương quan giữa các đặc trưng số và biến mục tiêu

Chuyển đổi các biến phân loại sang biến số và chia tập dữ liệu.

```python
# Phân loại các biến số cho các biến phân loại
categorical_features = ["Parental_Involvement", "Access_to_Resources", "Extracurricular_Activities", "Motivation_Level",
                        "Family_Income", "Internet_Access", "Teacher_Quality", "School_Type", "Peer_Influence",
                        "Learning_Disabilities", "Parental_Education_Level", "Distance_from_Home", "Gender"]

# Chuyển đổi các biến phân loại thành số bằng LabelEncoder
df_encoded = dfd.copy()
label_encoders = {}
for col in categorical_features:
    le = LabelEncoder()
    df_encoded[col] = le.fit_transform(df_encoded[col])
    label_encoders[col] = le

# Chuẩn bị dữ liệu
X = df_encoded.drop(columns="Exam_Score")
y = df_encoded["Exam_Score"]

# Chia dữ liệu train và test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Hình 2.37: Chuyển đổi các biến phân loại sang biến số và chia tập dữ liệu

Dùng mô hình Gradient Boosting Regression để xác định tầm quan trong các từng đặc trưng.

```python
# Huấn luyện mô hình Gradient Boosting Regression
gb_model = GradientBoostingRegressor(random_state=42, n_estimators=100)
gb_model.fit(X_train, y_train)

# Xác định tầm quan trọng của feature
feature_importances = pd.Series(gb_model.feature_importances_, index=X.columns).sort_values(ascending=False)

# Hiển thị các feature quan trọng
print("Feature Importance:")
print(feature_importances)
```

```
Feature Importance:
Attendance                  0.483156
Hours_Studied               0.267629
Previous_Scores             0.049902
Access_to_Resources         0.038963
Parental_Involvement        0.034973
Tutoring_Sessions           0.033238
Family_Income               0.016474
Peer_Influence              0.012934
Parental_Education_Level    0.010298
Distance_from_Home          0.009648
Motivation_Level            0.009415
Teacher_Quality             0.008923
Extracurricular_Activities  0.006357
Learning_Disabilities       0.006000
Physical_Activity           0.005845
Internet_Access             0.005662
Sleep_Hours                 0.000583
School_Type                 0.000000
Gender                      0.000000
dtype: float64
```

Hình 2.38: Tầm quan trọng của từng đặc trưng

Chọn các đặc trưng cuối cùng để cho vào mô hình huấn luyện

```python
final_selected_features = ['Attendance','Hours_Studied','Previous_Scores','Tutoring_Sessions','Peer_Influence','Distance_from_Home',
                           'Learning_Disabilities','Access_to_Resources','Parental_Involvement', 'Teacher_Quality']

final_X_train = X_train[final_selected_features]
final_X_test = X_test[final_selected_features]
```

Hình 2.39: Chọn các đặc trưng cuối cùng để cho vào mô hình huấn luyện

```python
# Các cột cần chuẩn hóa
columns_to_scale = ['Attendance', 'Hours_Studied', 'Previous_Scores', 'Tutoring_Sessions']

# Tạo bản sao của final_X_train và final_X_test
final_X_train_scaled = final_X_train.copy()
final_X_test_scaled = final_X_test.copy()

# Chuẩn hóa các cột cần thiết
scaler = StandardScaler()
final_X_train_scaled[columns_to_scale] = scaler.fit_transform(final_X_train[columns_to_scale])
final_X_test_scaled[columns_to_scale] = scaler.transform(final_X_test[columns_to_scale])

# Kiểm tra kết quả
final_X_train_scaled.head()
```

	Attendance	Hours_Studied	Previous_Scores	Tutoring_Sessions	Peer_Influence	Distance_from_Home	Learning_Disabilities	Access_to_Resources	Parental_Involvement	Teacher_Quali
2790	0.775563	0.709797	-1.199178	1.205117	2	2	0	2	1	
2693	0.429917	-0.148557	0.611913	1.205117	0	2	0	0	0	
1963	1.121208	0.709797	1.447801	2.816952	0	2	0	2	2	
3786	-1.557544	0.366456	-0.363290	2.011034	0	2	0	2	0	
4847	0.689152	-0.491899	1.726430	-0.406719	2	2	0	0	2	

Hình 2.39: Chuẩn hóa dữ liệu các cột số

Chương 3 – TRIỂN KHAI XÂY DỰNG

3.1 Huấn luyện mô hình

Tạo các danh sách để lưu trữ các mô hình và điểm số đánh giá.

```
# Tạo các danh sách để lưu trữ các mô hình và các điểm số đánh giá.
models = []
mae_scores = []
mse_scores = []
r2_scores = []
```

```
def train_and_evaluate_model(model, model_name):
    # Huấn luyện model
    model.fit(final_X_train_scaled, y_train)

    # Dự đoán với tập test
    y_pred = model.predict(final_X_test_scaled)

    # Tính toán các chỉ số đánh giá
    mae = mean_absolute_error(y_test, y_pred)
    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)

    # In kết quả đánh giá
    print(f"Model: {model_name}")
    print(f"Mean Absolute Error (MAE): {mae}")
    print(f"Mean Squared Error (MSE): {mse}")
    print(f"R-squared (R2): {r2}")

    # Lưu kết quả vào các danh sách để so sánh sau
    mae_scores.append(mae)
    mse_scores.append(mse)
    r2_scores.append(r2)
    models.append((model_name, model))

    # Giải phóng bộ nhớ
    gc.collect()
```

Hình 3.1: Tạo các danh sách để lưu trữ các mô hình và điểm số đánh giá

Khởi tạo các model.

```
# Khởi tạo các model
lin_reg = LinearRegression()
rf_reg = RandomForestRegressor(n_estimators=100, random_state=42)
gbr = GradientBoostingRegressor(n_estimators=100, random_state=42)
svr = SVR()
xgb_model = xgb.XGBRegressor(random_state=42)
```

Hình 3.2: Khởi tạo các model

Thực hiện việc huấn luyện và đánh giá mô hình học máy sử dụng thuật toán Linear Regression.

```
train_and_evaluate_model(lin_reg, "Linear Regression")
```

```
Model: Linear Regression
Mean Absolute Error (MAE): 1.1764756565642456
Mean Squared Error (MSE): 6.295747784887716
R-squared (R2): 0.5858705022745339
```

Hình 3.3: Huấn luyện và đánh giá mô hình học máy sử dụng thuật toán Linear Regression

Thực hiện việc huấn luyện và đánh giá mô hình học máy sử dụng thuật toán Random Forest Regression.

```
train_and_evaluate_model(rf_reg, "Random Forest Regression")
```

```
Model: Random Forest Regression
Mean Absolute Error (MAE): 1.285026656511805
Mean Squared Error (MSE): 6.767356892612336
R-squared (R2): 0.5548484141004226
```

Hình 3.4: Huấn luyện và đánh giá mô hình học máy sử dụng thuật toán Random Forest Regression

Thực hiện việc huấn luyện và đánh giá mô hình học máy sử dụng thuật toán Support Vector Regression.

```
train_and_evaluate_model(svr, "Support Vector Regression")
```

```
Model: Support Vector Regression
Mean Absolute Error (MAE): 0.9247647583229577
Mean Squared Error (MSE): 5.499396357482213
R-squared (R2): 0.6382538930825379
```

Hình 3.5: Huấn luyện và đánh giá mô hình học máy sử dụng thuật toán Support Vector Regression

Thực hiện việc huấn luyện và đánh giá mô hình học máy sử dụng thuật toán XGBoost Regression.

```
train_and_evaluate_model(xgb_model, "XGBoost Regression")
```

```
Model: XGBoost Regression
Mean Absolute Error (MAE): 1.2512563689465026
Mean Squared Error (MSE): 6.89963248853876
R-squared (R2): 0.5461474657058716
```

Hình 3.6: Huấn luyện và đánh giá mô hình học máy sử dụng thuật toán XGBoost Regression

Thực hiện việc huấn luyện và đánh giá mô hình học máy sử dụng thuật toán Stacking Regressor.

```
stacking_model = StackingRegressor(
    estimators=[('lr', lin_reg), ('rf', rf_reg), ('gbr', gbr), ('svr', svr), ('xgb', xgb_model)],
    final_estimator=LinearRegression()
)
train_and_evaluate_model(stacking_model, "Stacking Regressor")
```

```
Model: Stacking Regressor
Mean Absolute Error (MAE): 0.932249660335133
Mean Squared Error (MSE): 5.453686948968513
R-squared (R2): 0.6412606231860762
```

Hình 3.7: Huấn luyện và đánh giá mô hình học máy sử dụng thuật toán Stacking Regressor

Tạo DataFrame chứa thông tin hiệu suất của các mô hình hồi quy và sắp xếp theo R^2.

```
# Tạo DataFrame chứa thông tin hiệu suất của các mô hình hồi quy và sắp xếp theo R²
model_perfs = pd.DataFrame({
    'Model': [model_name for model_name, model in models],
    'MAE': mae_scores,
    'MSE': mse_scores,
    'R2': r2_scores
}).sort_values('R2', ascending=False).reset_index(drop=True)

model_perfs
```

	Model	MAE	MSE	R2
0	Stacking Regressor	0.932250	5.453687	0.641261
1	Support Vector Regression	0.924765	5.499396	0.638254
2	Linear Regression	1.176476	6.295748	0.585871
3	Random Forest Regression	1.285027	6.767357	0.554848
4	XGBoost Regression	1.251256	6.899632	0.546147

Hình 3.8: Tạo DataFrame chứa thông tin hiệu suất của các mô hình hồi quy và sắp xếp theo R^2

So sánh các điểm số đánh giá:

MAE (Lỗi trung bình tuyệt đối):

- Support Vector và Stacking Regressor có MAE thấp nhất → dự đoán chính xác nhất.
- Random Forest và XGBoost có MAE cao hơn → dự đoán kém hơn.

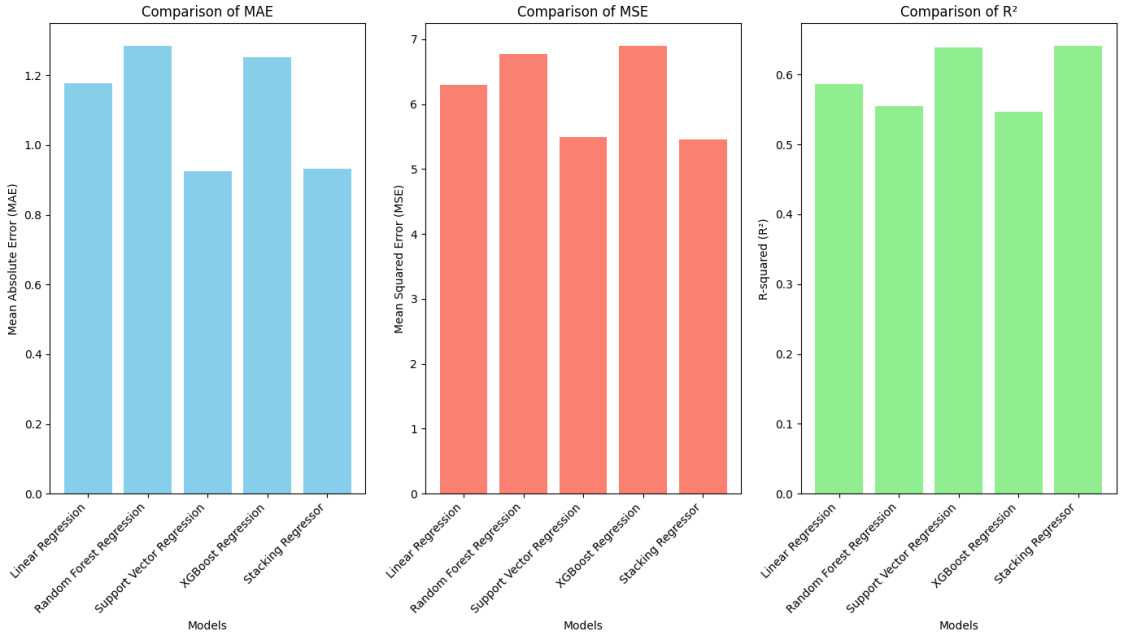
MSE (Lỗi bình phương trung bình):

- Support Vector và Stacking Regressor vượt trội với MSE thấp nhất.
- XGBoost có MSE cao nhất → kém hiệu quả nhất.

R^2 (Hệ số xác định):

- Support Vector và Stacking Regressor đạt R^2 cao nhất (~0.6) → giải thích dữ liệu tạm tốt.
- Linear Regression và Random Forest có R^2 thấp hơn.

Kết luận: Stacking Regressor và Support Vector là lựa chọn tốt nhất; XGBoost có hiệu suất kém nhất.



Hình 3.9: Biểu đồ so sánh các điểm số đánh giá mô hình

3.2 Tối ưu hóa siêu tham số

Xác định các tham số cần tối ưu cho SVR. Sau đó dùng GridSearchCV tìm tham số tối ưu cho mô hình SVR và huấn luyện với dữ liệu chuẩn hóa.

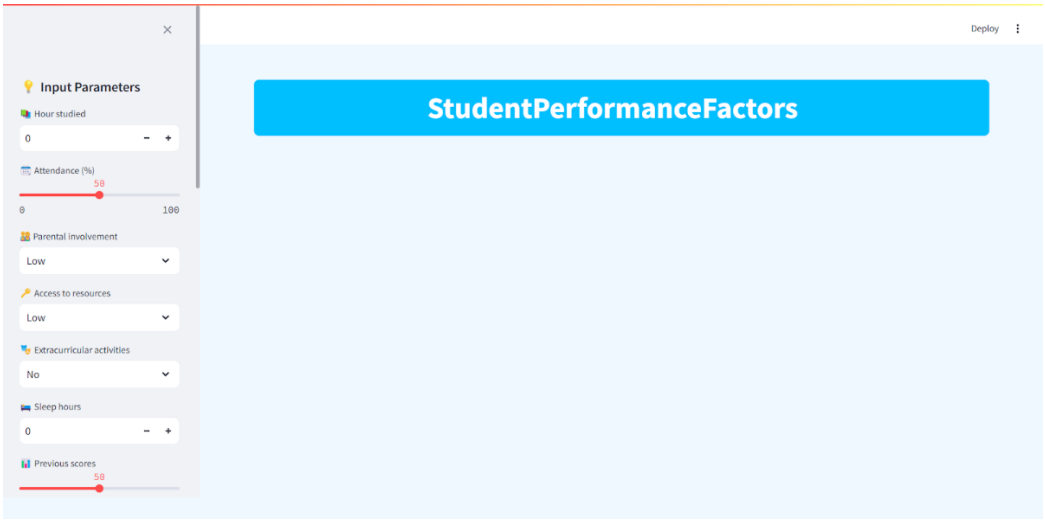
```python
from sklearn.model_selection import GridSearchCV

# Xác định các tham số cần tối ưu cho SVR
param_grid = {
    'C': [0.1, 1, 10],  # Hệ số phạt
    'epsilon': [0.01, 0.1, 0.2],  # Độ rộng của vùng không thay đổi
    'kernel': ['linear', 'rbf', 'poly'],  # Kiểu kernel
    'gamma': ['scale', 'auto']  # Hệ số cho kernel
}

# Sử dụng GridSearchCV để tìm tham số tối ưu cho SVR
grid_search = GridSearchCV(estimator=svr, param_grid=param_grid, cv=5, n_jobs=-1, verbose=1)

# Huấn luyện GridSearchCV với dữ liệu chuẩn hóa
grid_search.fit(final_X_train_scaled, y_train)

# In các tham số tốt nhất và điểm số của mô hình tốt nhất
print(f"Best Parameters: {grid_search.best_params_}")
print(f"Best Score: {grid_search.best_score_}")

# Đánh giá mô hình với các tham số tốt nhất
best_model = grid_search.best_estimator_

# Dự đoán với mô hình tốt nhất
y_pred = best_model.predict(final_X_test_scaled)

# Tính toán các chỉ số đánh giá
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("\nSVR Model with Hyperparameter Tuning:")
print(f"Mean Absolute Error (MAE): {mae}")
print(f"Mean Squared Error (MSE): {mse}")
print(f"R-squared (R2): {r2}")
```

Hình 3.10: Biểu đồ so sánh các điểm số đánh giá mô hình

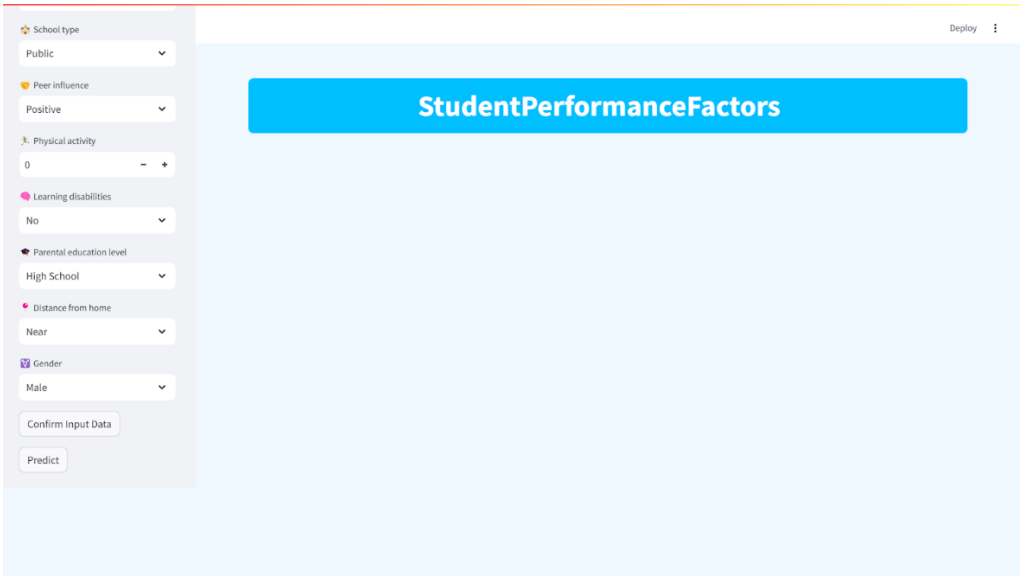
```
Fitting 5 folds for each of 54 candidates, totalling 270 fits
Best Parameters: {'C': 10, 'epsilon': 0.2, 'gamma': 'scale', 'kernel': 'rbf'}
Best Score: 0.6896815701240088

SVR Model with Hyperparameter Tuning:
Mean Absolute Error (MAE): 0.8966368899814757
Mean Squared Error (MSE): 5.445964582856911
R-squared (R2): 0.641768594551563
```

Hình 3.11: In kết quả huấn luyện với GridSearchCV

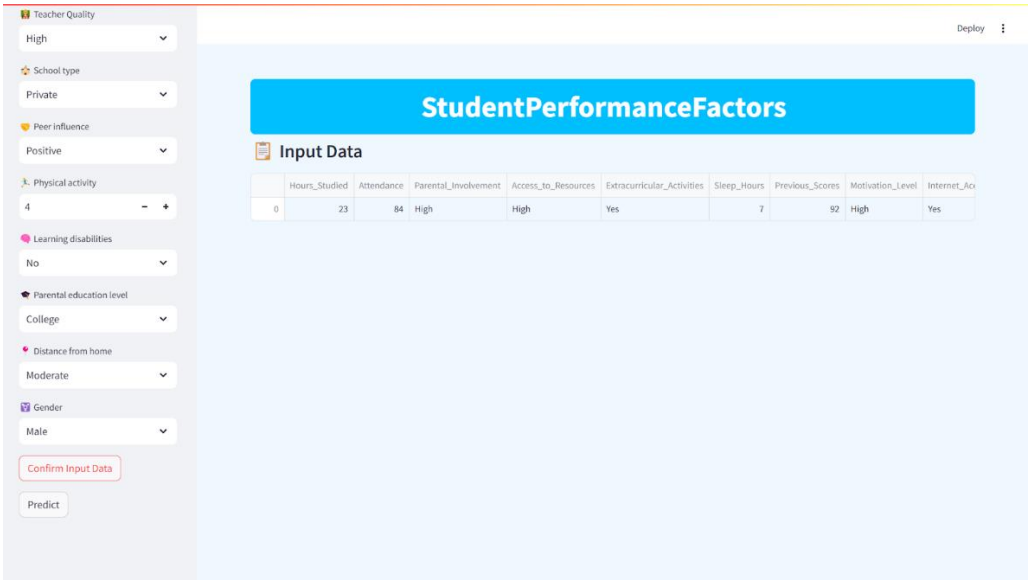
3.3 Triển khai dự đoán bằng Streamlit



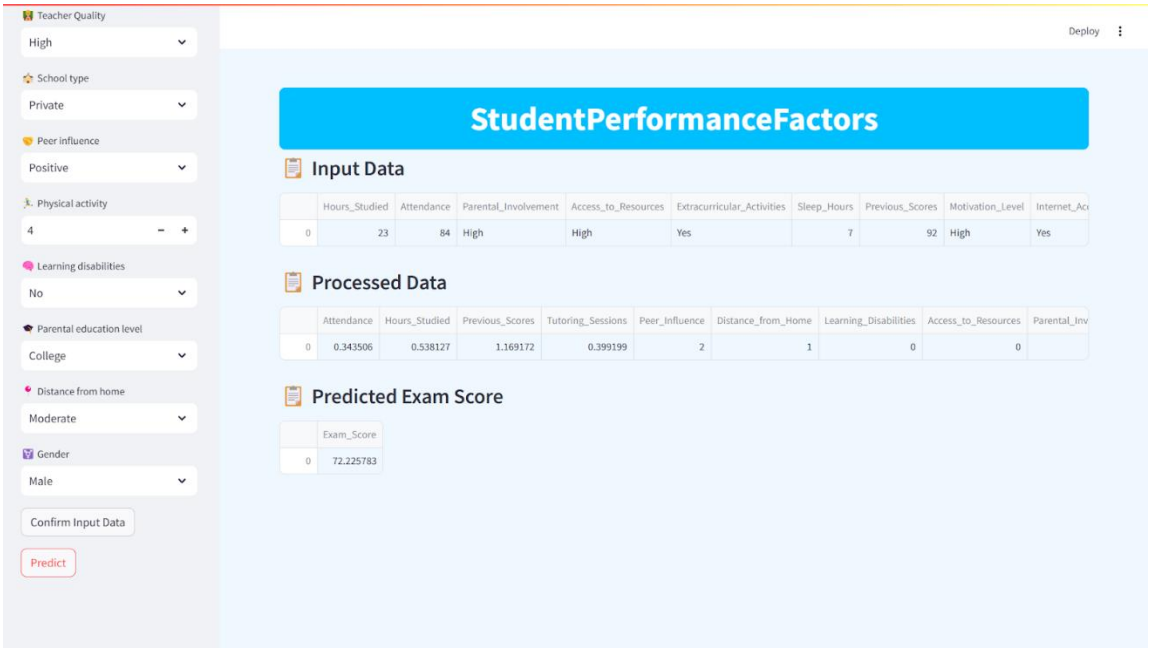
Hình 3.12: Input Parameters 1



Hình 3.13: Input Parameters 2



Hình 3.14: Hiển thị dữ liệu được nhập vào



Hình 3.15: Hiển thị kết quả dự đoán

KẾT LUẬN

1. Kết quả đạt được.

Bài tập lớn cuối kỳ đã thành công trong việc phân tích các yếu tố ảnh hưởng đến thành tích học tập của học sinh - sinh viên, tập trung vào các yếu tố cá nhân, gia đình và nhà trường. Với việc áp dụng các công cụ như Python và Excel, nhóm đã tiến hành thu thập, xử lý và phân tích dữ liệu một cách hiệu quả. Đặc biệt, các mối quan hệ giữa các yếu tố đã được làm rõ, tạo nền tảng cho các giải pháp hỗ trợ nâng cao chất lượng học tập.

Ngoài ra, bài tập lớn còn triển khai giao diện bằng Streamlit, cho phép hiển thị kết quả huấn luyện và dự đoán từ các mô hình đã xây dựng. Giao diện này giúp trình bày kết quả một cách đơn giản, thuận tiện cho người dùng truy cập và sử dụng, thể hiện sự tích hợp hiệu quả giữa phân tích dữ liệu và ứng dụng công nghệ. Kết quả đạt được không chỉ phản ánh năng lực vận dụng kiến thức mà còn khẳng định tiềm năng của phương pháp phân tích dữ liệu trong việc giải quyết các bài toán thực tiễn trong lĩnh vực giáo dục.

2. Hạn chế.

Mặc dù đã đạt được những kết quả đáng khích lệ, bài tập lớn vẫn tồn tại một số hạn chế cần khắc phục. Thứ nhất, dữ liệu được sử dụng trong phân tích có quy mô tương đối nhỏ và chưa bao quát được toàn bộ các yếu tố có thể ảnh hưởng đến thành tích học tập, điều này có thể ảnh hưởng đến tính đại diện và độ tin cậy của kết quả.

Thứ hai, các phương pháp phân tích được áp dụng chủ yếu tập trung vào những kỹ thuật cơ bản, chưa sử dụng các mô hình học máy hoặc thuật toán quá tiên tiến để khai thác triệt để thông tin từ dữ liệu. Điều này giới hạn khả năng dự đoán và phân tích chuyên sâu.

Thứ ba, giao diện Streamlit mặc dù giúp hiển thị kết quả huấn luyện và dự đoán nhưng chưa hỗ trợ trực quan hóa dữ liệu một cách đầy đủ. Điều này có thể khiến người dùng khó hình dung mối quan hệ giữa các yếu tố và ảnh hưởng của chúng đến kết quả.

Cuối cùng, thời gian và nguồn lực thực hiện hạn chế đã ảnh hưởng đến khả năng mở rộng nghiên cứu cũng như kiểm tra chéo để đảm bảo độ chính xác cao hơn trong phân tích. Đây là những khía cạnh cần được cải thiện trong các nghiên cứu tiếp theo.

3. Hướng phát triển.

Trong tương lai, bài tập lớn này có thể được mở rộng và cải thiện theo một số hướng sau:

Thứ nhất, mở rộng phạm vi thu thập dữ liệu để tăng tính đại diện và độ tin cậy. Việc tích hợp dữ liệu từ nhiều khu vực, đối tượng và cấp học khác nhau sẽ giúp phân tích toàn diện hơn, bao quát các yếu tố đa chiều ảnh hưởng đến thành tích học tập.

Thứ hai, ứng dụng các phương pháp phân tích và dự đoán tiên tiến hơn như học máy, học sâu và các mô hình thống kê hiện đại. Điều này không chỉ giúp nâng cao khả

năng dự đoán mà còn làm rõ hơn mối quan hệ phi tuyến tính giữa các yếu tố tác động.

Thứ ba, cải tiến giao diện Streamlit để không chỉ hiển thị kết quả huấn luyện và dự đoán mà còn hỗ trợ trực quan hóa dữ liệu một cách sinh động hơn. Điều này sẽ giúp người dùng dễ dàng hình dung và hiểu sâu hơn về dữ liệu và các kết quả phân tích.

Cuối cùng, tích hợp các yếu tố môi trường kinh tế - xã hội và công nghệ trong phân tích để đưa ra các khuyến nghị phù hợp với bối cảnh thực tế. Đồng thời, việc xây dựng một hệ thống báo cáo tự động hóa dựa trên dữ liệu có thể giúp cải thiện hiệu quả sử dụng trong thực tiễn giáo dục. Những hướng phát triển này sẽ góp phần nâng cao giá trị ứng dụng của nghiên cứu, đồng thời mở ra cơ hội triển khai ở quy mô lớn hơn.

4. Kết luận.

Bài tập lớn đã phân tích thành công các yếu tố ảnh hưởng đến thành tích học tập của học sinh - sinh viên, đồng thời triển khai dự đoán kết quả qua giao diện Streamlit. Mặc dù còn hạn chế về quy mô dữ liệu và mức độ phức tạp của phương pháp phân tích, nghiên cứu đã cung cấp cơ sở quan trọng cho các giải pháp giáo dục.

Hướng phát triển bao gồm mở rộng dữ liệu, áp dụng mô hình phân tích nâng cao và cải thiện giao diện Streamlit. Nghiên cứu này khẳng định tiềm năng của phân tích dữ liệu trong việc cải thiện chất lượng giáo dục, tạo nền tảng cho các ứng dụng thực tiễn trong tương lai.

TÀI LIỆU THAM KHẢO

[1] https://www.geeksforgeeks.org/data-analysis-tutorial/

[2] Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*. O'Reilly Media.

[3] McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2nd ed.). O'Reilly Media.

[4] Croll, A., & Yoskovitz, B. (2013). *Lean Analytics: Use Data to Build a Better Startup Faster*. O'Reilly Media.