|  |
| --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT HƯNG YÊN**  **D:\_Data_2019_2020\logo_utehy.png**  **BÀI TẬP LỚN**  **MÔN: HỌC MÁY CƠ BẢN**  **DỰ ĐOÁN ĐIỂM TRUNG BÌNH GPA CỦA SINH VIÊN**  NGÀNH: KHOA HỌC MÁY TÍNH  CHUYÊN NGÀNH: TRÍ TUỆ NHÂN TẠO VÀ KHOA HỌC DỮ LIỆU  SINH VIÊN:VŨ HUY HOÀNG MÃ SINH VIÊN: 12423073  MÃ LỚP: 124231  GIẢNG VIÊN HƯỚNG DẪN: PGS. NGUYỄN VĂN HẬU  **HƯNG YÊN – 2025** |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT HƯNG YÊN**  **D:\_Data_2019_2020\logo_utehy.png**  **BÀI TẬP LỚN**  **MÔN: HỌC MÁY CƠ BẢN**  **DỰ ĐOÁN ĐIỂM TRUNG BÌNH GPA CỦA SINH VIÊN**  NGÀNH: KHOA HỌC MÁY TÍNH  CHUYÊN NGÀNH: TRÍ TUỆ NHÂN TẠO VÀ KHOA HỌC DỮ LIỆU  SINH VIÊN:VŨ HUY HOÀNG MÃ SINH VIÊN: 12423073  MÃ LỚP: 124231  GIẢNG VIÊN HƯỚNG DẪN: PGS. NGUYỄN VĂN HẬU  **HƯNG YÊN – 2025** |

Nhận xét của giảng viên hướng dẫn:

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

Ký và ghi họ tên

LỜI CAM ĐOAN

Em xin cam đoan bài tập lớn môn Học Máy Cơ Bản có tên là “Dự đoán điểm trung bình GPA của sinh viên” là kết quả thực hiện của bản thân em dưới sự hướng dẫn của thầy Nguyễn Văn Hậu và thầy Nguyễn Tuấn Anh. Những phần sử dụng tài liệu tham khảo trong bài tập lớn đã được nêu rõ trong phần tài liệu tham khảo. Các kết quả trình bày trong bài tập lớn và chương trình xây dựng được hoàn toàn là kết quả do bản thân em thực hiện.

Nếu vi phạm lời cam đoan này, em xin chịu hoàn toàn trách nhiệm trước khoa và nhà trường.

Hưng Yên, ngày 26 tháng 12 năm 2025

Sinh viên

LỜI CÁM ƠN

Để có thể hoàn thành bài tập lớn này, lời đầu tiên em xin phép gửi lời cảm ơn tới bộ môn Khoa học máy tính, Khoa Công nghệ thông tin – Trường Đại học Sư phạm Kỹ thuật Hưng Yên đã tạo điều kiện thuận lợi cho em thực hiện bài tập lớn môn học này.

Đặc biệt em xin chân thành cảm ơn thầy Nguyễn Văn Hậu và thầy Nguyễn Tuấn Anh đã rất tận tình hướng dẫn, chỉ bảo em trong suốt thời gian thực hiện bài tập lớn vừa qua.

Em cũng xin chân thành cảm ơn tất cả các Thầy, các Cô trong Trường đã tận tình giảng dạy, trang bị cho em những kiến thức cần thiết, quý báu để giúp em thực hiện được bài tập lớn này.

Mặc dù em đã có cố gắng, nhưng với trình độ còn hạn chế, trong quá trình thực hiện đề tài không tránh khỏi những thiếu sót. Em hi vọng sẽ nhận được những ý kiến nhận xét, góp ý của các Thầy giáo, cô giáo về những kết quả triển khai trong bài tập lớn.

Em xin trân trọng cảm ơn!

**MỤC LỤC**

[LỜI CAM ĐOAN 2](#_Toc218085593)

[LỜI CÁM ƠN 3](#_Toc218085594)

[DANH SÁCH TỪ VIẾT TẮT 8](#_Toc218085595)

[DANH SÁCH BẢNG 9](#_Toc218085596)

[DANH SÁCH HÌNH ẢNH 10](#_Toc218085597)

[CHƯƠNG 1: TỔNG QUAN 12](#_Toc218085598)

[1.1.Lí do chọn đề tài 12](#_Toc218085599)

[1.2.Mục tiêu của đề tài 13](#_Toc218085600)

[1.3. Giới hạn và phạm vi của đề bài 14](#_Toc218085601)

[1.4. Nội dung thực hiện 16](#_Toc218085602)

[1.5. Phương pháp tiếp cận 18](#_Toc218085603)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 21](#_Toc218085604)

[2.1. Trí tuệ nhân tạo AI (Artificial Intelligence) 21](#_Toc218085605)

[2.2. Học máy (Machine Learning) 21](#_Toc218085606)

[2.2.1. Học có giám sát (Supervised Learning) 21](#_Toc218085607)

[2.2.2. Học không giám sát (Unsupervised Learning) 22](#_Toc218085608)

[2.2.3. Học tăng cường (Reinforcement Learning) 24](#_Toc218085609)

[2.2.4. Bài toán hồi quy trong Trí tuệ nhân tạo 25](#_Toc218085610)

[2.2.5. Hàm mất mát và tối ưu hóa trong học máy (Regression) 25](#_Toc218085611)

[2.2.6. Overfitting và khả năng tổng quát hóa (Generalization) 31](#_Toc218085612)

[2.2.7. Vai trò của AI trong bài toán dự đoán GPA 32](#_Toc218085613)

[2.3. Các mô hình của bài toán 32](#_Toc218085614)

[2.3.1. Linear Regression 32](#_Toc218085615)

[2.3.2. Random Forest 35](#_Toc218085616)

[2.3.3. Gradient Boosting 38](#_Toc218085617)

[2.3.4. XGBoost 40](#_Toc218085618)

[CHƯƠNG 3: NỘI DUNG THỰC HIỆN 46](#_Toc218085619)

[3.1. Khảo sát và xác định yêu cầu 46](#_Toc218085620)

[3.1.1. Mô hình 46](#_Toc218085621)

[3.1.2. Xác định đầu vào và đầu ra của mô hình 46](#_Toc218085622)

[3.1.3. Xác định loại bài toán 46](#_Toc218085623)

[3.1.4. Xác định yêu cầu chức năng 47](#_Toc218085624)

[3.1.5. Xác định yêu cầu phi chức năng 47](#_Toc218085625)

[3.2. Thu thập và mô tả dữ liệu 47](#_Toc218085626)

[3.2.1. Thu thập dữ liệu 47](#_Toc218085627)

[3.2.2. Quy mô và cấu trúc dữ liệu 48](#_Toc218085628)

[3.2.3. Mô tả các thuộc tính dữ liệu 48](#_Toc218085629)

[3.2.4. Khảo sát sơ bộ dữ liệu 49](#_Toc218085630)

[3.3. Tiền xử lý dữ liệu 51](#_Toc218085631)

[3.3.1. Mục đích của tiền xử lý dữ liệu 51](#_Toc218085632)

[3.3.2. Kiểm tra và xử lý giá trị thiếu (Missing Values) 51](#_Toc218085633)

[3.3.3. Xử lý dữ liệu phân loại (Categorical Encoding) 52](#_Toc218085634)

[3.3.4. Chuẩn hóa dữ liệu số (Feature Scaling) 53](#_Toc218085635)

[3.4. Phân tích và khám phá dữ liệu (Exploratory Data Analysis – EDA) 54](#_Toc218085636)

[3.4.1. Mục đích của phân tích và khám phá dữ liệu 54](#_Toc218085637)

[3.4.2. Phân tích phân phối điểm thi (Exam Score) 54](#_Toc218085638)

[3.4.3. Phân bố GPA học kỳ trước 55](#_Toc218085639)

[3.4.4. Mối quan hệ giữa GPA học kỳ trước và điểm thi trung bình 56](#_Toc218085640)

[3.4.5. Mối quan hệ giữa giờ học mỗi ngày và điểm thi 57](#_Toc218085641)

[3.4.6. Mối quan hệ giữa áp lực học tập và điểm thi 57](#_Toc218085642)

[3.4.7. So sánh điểm thi giữa nhóm có và không có học gia sư 58](#_Toc218085643)

[3.4.8. Phân bố điểm thi theo thời gian sử dụng mạng xã hội 59](#_Toc218085644)

[3.4.9. Phân bố điểm thi theo tình trạng đi làm thêm 59](#_Toc218085645)

[3.4.10. Phân bố điểm thi theo phong cách học tập 60](#_Toc218085646)

[3.4.11. Mối quan hệ giữa tần suất tập thể dục và điểm thi 61](#_Toc218085647)

[3.4.12. Mối quan hệ giữa khả năng quản lý thời gian và điểm thi 62](#_Toc218085648)

[3.4.13. Mối quan hệ giữa mức độ hỗ trợ của bố mẹ và điểm thi 63](#_Toc218085649)

[3.3.14. Tổng kết phân tích và khám phá dữ liệu (EDA) 63](#_Toc218085650)

[3.5. Xây dựng mô hình dự đoán 64](#_Toc218085651)

[3.5.1. Mục tiêu xây dựng mô hình 64](#_Toc218085652)

[3.5.2. Xác định dữ liệu đầu vào và đầu ra 64](#_Toc218085653)

[3.5.3. Lí do lựa chọn các mô hình học máy 64](#_Toc218085654)

[3.5.4. Chiến lược huấn luyện mô hình 65](#_Toc218085655)

[3.5.5. Tiêu chí đánh giá mô hình 66](#_Toc218085656)

[3.6. Huấn luyện và đánh giá mô hình 66](#_Toc218085657)

[3.6.1. Quy trình huấn luyện mô hình 66](#_Toc218085658)

[3.6.2. Chuẩn bị dữ liệu 67](#_Toc218085659)

[3.6.3. Phân loại các thuộc tính số và thuộc tính phân loại 67](#_Toc218085660)

[3.6.4. Chia tập dữ liệu huấn luyện và kiểm tra 68](#_Toc218085661)

[3.6.5. Xây dựng quy trình tiền xử lý dữ liệu bằng ColumnTransformer 68](#_Toc218085662)

[3.6.6. Áp dụng tiền xử lý cho tập huấn luyện và tập kiểm tra 69](#_Toc218085663)

[3.6.7. Chuyển đổi dữ liệu dạng sparse sang dense 70](#_Toc218085664)

[3.6.8. Tái tạo tập dữ liệu sau khi mã hóa và chuẩn hóa đặc trưng 71](#_Toc218085665)

[3.6.9. Huấn luyện và đánh giá mô hình Linear Regression 72](#_Toc218085666)

[3.6.10. Huấn luyện và đánh giá mô hình Random Forest Regression 73](#_Toc218085667)

[3.6.11. Huấn luyện và đánh giá mô hình XGBoost Regression 74](#_Toc218085668)

[3.6.12. Huấn luyện và đánh giá mô hình Gradient Boosting Regression 76](#_Toc218085669)

[3.6.13. So sánh các kết quả của mô hình 77](#_Toc218085670)

[3.6.14. Trực quan hóa tỉ lệ chính xác của các mô hình 79](#_Toc218085671)

[3.6.14. Trực quan hóa toàn bộ Metric đánh giá của các mô hình 80](#_Toc218085672)

[3.7. Demo mô hình 81](#_Toc218085673)

[KẾT LUẬN 88](#_Toc218085674)

[TÀI LIỆU THAM KHẢO 91](#_Toc218085675)

DANH SÁCH TỪ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| **Từ viết tắt** | **Nghĩa tiếng Anh** | **Nghĩa tiếng Việt** |
| AI | Artificial Intelligence | Trí tuệ nhân tạo |
| GPA | Grade Point Average | Điểm trung bình học tập |
| EDA | Exploratory Data Analysis | Phân tích khám phá dữ liệu |
| XGBoost | Extreme Gradient Boosting | Thuật toán tăng cường gradient cải tiến |
| MAE | Mean Absolute Error | Sai số tuyệt đối trung bình |
| MSE | Mean Squared Error | Sai số bình phương trung bình |
| RMSE | Root Mean Squared Error | Căn sai số bình phương trung bình |
| R² | Coefficient of Determination | Hệ số xác định |
| CPU | Central Processing Unit | Bộ xử lý trung tâm |

DANH SÁCH BẢNG

[Bảng 2.1. Bảng so sánh các thuật toán trong mô hình 44](#_Toc218086446)

[Bảng 3.1. Bảng mô tả các thuộc tính dữ liệu 48](#_Toc218086447)

DANH SÁCH HÌNH ẢNH

[Hình 3.1. Thông tin tổng quan của dữ liệu 50](#_Toc218086814)

[Hình 3.2. 5 dòng đầu tiên trong dataset 50](#_Toc218086815)

[Hình 3.3. Thống kê tổng quát dữ liệu 51](#_Toc218086816)

[Hình 3.4. Kiểm tra dữ liệu bị thiếu 52](#_Toc218086817)

Hình [3.5. Kiểm tra các cột có dữ liệu chữ 53](#_Toc218086818)

[Hình 3.6. Dữ liệu sau khi đã chuẩn hóa dữ liệu số và dữ liệu chữ 54](#_Toc218086819)

[Hình 3.7. Biểu đồ phân bố dữ liệu điểm thi 55](#_Toc218086820)

[Hình 3.8. Biểu đồ phân bố dữ liệu điểm GPA học kì trước 56](#_Toc218086821)

[Hình 3.9. Biểu đồ trực quan hóa mối quan hệ điểm GPA học kì trước với điểm thi 57](#_Toc218086822)

[Hình 3.10. Biểu đồ trực quan hóa mối quan hệ giờ học mỗi ngày với điểm thi 57](#_Toc218086823)

[Hình 3.11. Biểu đồ trực quan hóa mối quan hệ giữa áp lực học tập với điểm thi 58](#_Toc218086824)

[Hình 3.12. Biểu đồ trực quan hóa mối quan hệ giữa việc học gia sư với điểm thi 59](#_Toc218086825)

[Hình 3.13. Biểu đồ trực quan hóa mối quan hệ giữa thời gian sử dụng mạng xã hội với điểm thi 59](#_Toc218086826)

[Hình 3.14. Biểu đồ trực quan hóa mối quan hệ tình trạng đi làm thêm với điểm thi 60](#_Toc218086827)

[Hình 3.15. Biểu đồ trực quan hóa mối quan hệ phong cách học tập với điểm thi 61](#_Toc218086828)

[Hình 3.16. Biểu đồ trực quan hóa mối quan hệ tần suất tập thể dục với điểm thi 62](#_Toc218086829)

[Hình 3.17. Biểu đồ trực quan hóa mối quan hệ mức độ quản lý thời gian với điểm thi 63](#_Toc218086830)

[Hình 3.18. Biểu đồ trực quan hóa mối quan hệ mức độ giúp đỡ của bố mẹ với điểm thi 63](#_Toc218086831)

[Hình 3.19. Thống kê các Metric đánh giá của mô hình 77](#_Toc218086832)

[Hình 3.20. Trực quan hóa tỉ lệ chính xác của các mô hình 79](#_Toc218086833)

[Hình 3.20. Trực quan hóa toàn bộ Metric đánh giá của các mô hình 80](#_Toc218086834)

[Hình 3.21. Giao diện demo dự đoán điểm GPA của sinh viên 81](#_Toc218086835)

[Hình 3.22. Hình ảnh dữ liệu đầu vào quan trọng (GPA kì trước) 82](#_Toc218086836)

[Hình 3.23. Hình ảnh kết quả dự đoán GPA bằng mô hình Random Forest 83](#_Toc218086837)

[Hình 3.24. Hình ảnh kết quả dự đoán GPA bằng mô hình XGBoost 84](#_Toc218086838)

[Hình 3.25. Hình ảnh kết quả dự đoán GPA bằng mô hình Linear Regression 85](#_Toc218086839)

[Hình 3.26. Hình ảnh kết quả dự đoán GPA bằng mô hình Gradient Boosting 86](#_Toc218086840)

[Hình 3.27. Hình ảnh kết quả thực tế điểm GPA của sinh viên 86](#_Toc218086841)

CHƯƠNG 1: TỔNG QUAN

## Lí do chọn đề tài

Trong bối cảnh giáo dục hiện đại, việc nâng cao chất lượng đào tạo và hỗ trợ sinh viên đạt được kết quả học tập tốt nhất là một trong những mục tiêu quan trọng hàng đầu của các cơ sở giáo dục. Điểm trung bình học tập (GPA) không chỉ phản ánh năng lực học tập của sinh viên mà còn ảnh hưởng trực tiếp đến cơ hội học bổng, xét tốt nghiệp, định hướng nghề nghiệp và khả năng cạnh tranh trên thị trường lao động sau này.

Tuy nhiên, kết quả học tập của sinh viên chịu tác động bởi nhiều yếu tố khác nhau như thói quen học tập, thời gian tự học, mức độ tham gia lớp học, điều kiện sinh hoạt, kỹ năng quản lý thời gian và các yếu tố cá nhân khác. Việc đánh giá và dự đoán GPA theo cách truyền thống thường mang tính chủ quan, chưa khai thác hiệu quả nguồn dữ liệu ngày càng phong phú trong môi trường giáo dục hiện nay.

Sự phát triển mạnh mẽ của Trí tuệ nhân tạo (Artificial Intelligence) và Học máy (Machine Learning) đã mở ra nhiều cơ hội mới trong việc phân tích dữ liệu giáo dục. Các mô hình học máy có khả năng học từ dữ liệu lịch sử, phát hiện các mối quan hệ tiềm ẩn giữa các yếu tố đầu vào và kết quả học tập, từ đó đưa ra dự đoán có độ chính xác cao hơn so với các phương pháp thống kê truyền thống.

Xuất phát từ những lý do trên, em đã chọn đề tài “Dự đoán điểm trung bình (GPA Prediction) của sinh viên bằng các thuật toán học máy” nhằm:

- Ứng dụng kiến thức Trí tuệ nhân tạo đã học vào một bài toán thực tế, gần gũi với môi trường học tập của sinh viên.

- So sánh và đánh giá hiệu quả của nhiều thuật toán khác nhau như Linear Regression, Random Forest, Gradient Descent và XGBoost trong bài toán dự đoán GPA.

- Phân tích mức độ ảnh hưởng của các yếu tố học tập và sinh hoạt đến kết quả học tập của sinh viên.

- Đề xuất một mô hình hỗ trợ dự đoán GPA, từ đó có thể giúp sinh viên sớm nhận biết nguy cơ học tập kém và điều chỉnh kế hoạch học tập phù hợp, đồng thời hỗ trợ giảng viên và nhà trường trong công tác tư vấn, quản lý và nâng cao chất lượng đào tạo.

Thông qua đề tài này, em không chỉ củng cố kiến thức lý thuyết về các thuật toán học máy mà còn rèn luyện kỹ năng xử lý dữ liệu, xây dựng mô hình, đánh giá kết quả và trình bày một báo cáo khoa học hoàn chỉnh.

## Mục tiêu của đề tài

Mục tiêu chính của đề tài là xây dựng và đánh giá các mô hình học máy nhằm dự đoán điểm trung bình học tập (GPA) của sinh viên dựa trên các yếu tố liên quan đến thói quen học tập, mức độ tham gia học tập và các đặc trưng cá nhân khác. Thông qua đó, đề tài hướng đến việc ứng dụng kiến thức Trí tuệ nhân tạo vào giải quyết một bài toán thực tiễn trong lĩnh vực giáo dục.

Cụ thể, đề tài hướng tới các mục tiêu sau:

**Mục tiêu tổng quát**

-Ứng dụng các thuật toán học máy để dự đoán điểm trung bình (GPA) của sinh viên một cách hiệu quả và có cơ sở khoa học.

-Đánh giá khả năng dự đoán của các mô hình khác nhau trong cùng một tập dữ liệu, từ đó lựa chọn mô hình phù hợp nhất cho bài toán.

**Mục tiêu cụ thể**

-Thu thập và tiền xử lý dữ liệu liên quan đến kết quả học tập và thói quen học tập của sinh viên (làm sạch dữ liệu, xử lý giá trị thiếu, chuẩn hóa và mã hóa dữ liệu).

-Xây dựng và huấn luyện các mô hình học máy bao gồm:

+Linear Regression

+Gradient Descent

+Random Forest

+XGBoost

-Đánh giá và so sánh hiệu quả của các mô hình thông qua các chỉ số đánh giá như MAE, MSE, RMSE và hệ số xác định R².

-Phân tích mức độ ảnh hưởng của các đặc trưng đầu vào đến điểm GPA, từ đó xác định những yếu tố quan trọng nhất tác động đến kết quả học tập của sinh viên.

-Kiểm tra khả năng tổng quát hóa của mô hình trên tập dữ liệu kiểm tra nhằm hạn chế hiện tượng overfitting.

-Đề xuất hướng ứng dụng thực tế của mô hình trong việc hỗ trợ sinh viên, giảng viên và nhà trường trong công tác tư vấn học tập và nâng cao chất lượng đào tạo

Việc chọn đề tài "Dự đoán điểm trung bình (GPA Prediction) của sinh viên" giúp em có cơ hội:

-Về mặt học thuật: Giúp củng cố kiến thức về các thuật toán học máy hồi quy, quy trình xây dựng mô hình và đánh giá hiệu quả mô hình trong môn Trí tuệ nhân tạo.

-Về mặt thực tiễn: Góp phần xây dựng công cụ hỗ trợ dự đoán kết quả học tập, giúp sinh viên sớm nhận biết nguy cơ học tập kém và có kế hoạch cải thiện phù hợp.

## 1.3. Giới hạn và phạm vi của đề bài

**Phạm vi của đề tài**

Đề tài tập trung nghiên cứu bài toán dự đoán điểm trung bình học tập (GPA) của sinh viên dựa trên dữ liệu học tập và các yếu tố liên quan đến thói quen học tập. Cụ thể:

-Đối tượng nghiên cứu là sinh viên trong một tập dữ liệu trên Kaggle có sẵn, bao gồm các thông tin về kết quả học tập và một số đặc trưng phản ánh hành vi, thói quen và mức độ tham gia học tập.

-Phạm vi dữ liệu được giới hạn trong một bộ dữ liệu cụ thể, không mở rộng sang nhiều trường học hoặc nhiều hệ thống giáo dục khác nhau.

-Đề tài chỉ xem xét bài toán hồi quy (Regression) nhằm dự đoán giá trị liên tục là điểm GPA, không nghiên cứu các bài toán phân loại (classification) như xếp loại học lực.

-Các thuật toán học máy được sử dụng trong đề tài bao gồm:

+Linear Regression

+Gradient Descent

+Random Forest

+XGBoost  
Các mô hình này được lựa chọn nhằm so sánh giữa phương pháp tuyến tính, phương pháp tối ưu hóa, và các mô hình ensemble hiện đại.

-Việc đánh giá mô hình được thực hiện thông qua các chỉ số phổ biến như MAE, MSE, RMSE và R².

**Giới hạn của đề tài**

Bên cạnh những đóng góp đạt được, đề tài vẫn tồn tại một số giới hạn nhất định:

-Dữ liệu sử dụng trong nghiên cứu có quy mô và phạm vi hạn chế, do đó kết quả dự đoán có thể chưa phản ánh đầy đủ cho toàn bộ sinh viên trong các môi trường giáo dục khác nhau.

-Một số yếu tố quan trọng ảnh hưởng đến GPA như phương pháp giảng dạy của giảng viên hoặc hoàn cảnh gia đình chưa được đưa vào mô hình do hạn chế về dữ liệu.

-Đề tài chưa đi sâu vào việc tối ưu hóa siêu tham số (hyperparameter tuning) ở mức độ nâng cao cho tất cả các mô hình, mà chủ yếu tập trung vào việc xây dựng và so sánh hiệu quả cơ bản giữa các thuật toán.

-Mô hình được xây dựng phục vụ cho mục đích học tập và nghiên cứu trong khuôn khổ môn học, chưa được triển khai thử nghiệm trong môi trường thực tế hoặc hệ thống quản lý đào tạo của nhà trường.

-Các thuật toán Deep Learning chưa được xem xét trong phạm vi đề tài do yêu cầu về dữ liệu lớn và thời gian huấn luyện.

## 1.4. Nội dung thực hiện

**Thu thập và tìm hiểu dữ liệu**

Trong bước đầu tiên, đề tài tiến hành thu thập bộ dữ liệu phục vụ cho bài toán dự đoán điểm trung bình (GPA) của sinh viên. Dữ liệu bao gồm các thuộc tính liên quan đến kết quả học tập và thói quen học tập như thời gian tự học, mức độ tham gia lớp học, điểm số các môn học, cũng như một số đặc trưng cá nhân của sinh viên.

Sau khi thu thập, dữ liệu được tiến hành khảo sát tổng quan nhằm:

-Nắm bắt cấu trúc dữ liệu, số lượng mẫu và số lượng thuộc tính.

-Xác định kiểu dữ liệu (số, phân loại).

-Phát hiện các vấn đề tiềm ẩn như dữ liệu bị thiếu, dữ liệu ngoại lai hoặc phân phối không đồng đều.

**Tiền xử lý dữ liệu**

Tiền xử lý dữ liệu là bước quan trọng nhằm đảm bảo chất lượng đầu vào cho các mô hình học máy. Các công việc chính được thực hiện bao gồm:

-Làm sạch dữ liệu: xử lý các giá trị thiếu, loại bỏ hoặc thay thế các giá trị không hợp lệ.

-Mã hóa dữ liệu phân loại bằng các kỹ thuật phù hợp (Label Encoding hoặc One-Hot Encoding).

-Chuẩn hóa dữ liệu số bằng các phương pháp như StandardScaler nhằm đưa các đặc trưng về cùng một thang đo, đặc biệt cần thiết đối với các mô hình tuyến tính và Gradient Descent.

-Chia dữ liệu thành các tập huấn luyện và kiểm tra (Train/Test split) để đánh giá khả năng tổng quát hóa của mô hình.

**Phân tích và khám phá dữ liệu (EDA)**

Trước khi xây dựng mô hình, dữ liệu được phân tích và trực quan hóa nhằm hiểu rõ mối quan hệ giữa các biến đầu vào và điểm GPA. Một số nội dung EDA được thực hiện bao gồm:

-Phân tích thống kê mô tả các thuộc tính (mean, median, variance,…).

-Trực quan hóa mối quan hệ giữa GPA và các yếu tố như thời gian học, mức độ tham gia lớp học.

-Phát hiện các đặc trưng có ảnh hưởng mạnh đến GPA, làm cơ sở cho việc lựa chọn mô hình và đánh giá kết quả.

**Xây dựng mô hình dự đoán**

Sau khi hoàn tất bước tiền xử lý và EDA, đề tài tiến hành xây dựng và huấn luyện các mô hình học máy để dự đoán GPA, bao gồm:

-Linear Regression: làm mô hình cơ sở (baseline) nhằm đánh giá mối quan hệ tuyến tính giữa các biến đầu vào và GPA.

-Gradient Descent: sử dụng phương pháp tối ưu hóa để huấn luyện mô hình hồi quy, giúp hiểu rõ quá trình tối ưu hàm mất mát.

-Random Forest Regression: mô hình ensemble dựa trên tập hợp nhiều cây quyết định, có khả năng xử lý tốt dữ liệu phi tuyến và giảm overfitting.

-XGBoost Regression: mô hình boosting mạnh mẽ, khai thác các mối quan hệ phức tạp trong dữ liệu và thường cho hiệu quả dự đoán cao.

**Đánh giá và so sánh mô hình**

Các mô hình sau khi huấn luyện được đánh giá trên tập dữ liệu kiểm tra thông qua các chỉ số:

-Mean Absolute Error (MAE)

-Mean Squared Error (MSE)

-Root Mean Squared Error (RMSE)

-Hệ số xác định R²

Kết quả đánh giá được tổng hợp và so sánh nhằm xác định mô hình có hiệu suất tốt nhất cho bài toán dự đoán GPA, đồng thời phân tích nguyên nhân dẫn đến sự khác biệt giữa các mô hình.

**Phân tích kết quả và rút ra nhận xét**

Dựa trên kết quả thu được, đề tài tiến hành:

-Phân tích mức độ chính xác và khả năng tổng quát hóa của từng mô hình.

-Nhận xét ưu và nhược điểm của mỗi thuật toán trong bối cảnh bài toán dự đoán GPA.

-Xác định các yếu tố có ảnh hưởng lớn nhất đến kết quả học tập của sinh viên.

**Tổng kết và đề xuất hướng phát triển**

Cuối cùng, đề tài tổng hợp các kết quả đạt được, đánh giá mức độ hoàn thành mục tiêu đề ra và đề xuất các hướng cải tiến, mở rộng trong tương lai như mở rộng dữ liệu, tối ưu mô hình hoặc áp dụng các phương pháp học sâu.

## 1.5. Phương pháp tiếp cận

Đề tài tiếp cận bài toán dự đoán điểm trung bình học tập (GPA) của sinh viên theo hướng học máy có giám sát (Supervised Learning), cụ thể là bài toán hồi quy (Regression). Phương pháp tiếp cận được xây dựng dựa trên quy trình của một bài toán Trí tuệ nhân tạo, bao gồm các bước từ xử lý dữ liệu đến xây dựng, đánh giá và so sánh mô hình.

**Tiếp cận theo hướng dữ liệu (Data-driven approach)**

Đề tài sử dụng dữ liệu lịch sử của sinh viên làm nền tảng cho việc huấn luyện mô hình. Thay vì xây dựng các quy tắc thủ công, mô hình học máy được thiết kế để tự động học các mối quan hệ giữa các yếu tố đầu vào (đặc trưng học tập) và giá trị đầu ra là điểm GPA.

Cách tiếp cận này cho phép:

-Khai thác hiệu quả dữ liệu có sẵn.

-Phát hiện các mối quan hệ phức tạp và phi tuyến giữa các yếu tố ảnh hưởng đến GPA.

-Nâng cao khả năng dự đoán so với các phương pháp truyền thống.

**Tiếp cận theo quy trình học máy**

Quy trình tiếp cận của đề tài bao gồm các bước chính sau:

Bước 1: Tiền xử lý dữ liệu  
Dữ liệu được làm sạch, chuẩn hóa và mã hóa nhằm đảm bảo phù hợp với yêu cầu của các thuật toán học máy.

Bước 2: Chia tập dữ liệu  
Dữ liệu được chia thành tập huấn luyện và tập kiểm tra để đánh giá khả năng tổng quát hóa của mô hình.

Bước 3: Xây dựng mô hình  
Các mô hình hồi quy khác nhau được huấn luyện trên cùng một tập dữ liệu nhằm đảm bảo tính công bằng khi so sánh.

Bước 4: Đánh giá mô hình  
Hiệu suất của mô hình được đo lường thông qua các chỉ số đánh giá phổ biến như MAE, MSE, RMSE và R².

Bước 5: So sánh và lựa chọn mô hình  
Kết quả từ các mô hình được so sánh để lựa chọn phương pháp phù hợp nhất cho bài toán dự đoán GPA.

**Tiếp cận so sánh nhiều thuật toán**

Đề tài không chỉ tập trung vào một mô hình duy nhất mà áp dụng nhiều thuật toán học máy khác nhau để đánh giá ưu nhược điểm của từng phương pháp:

-Linear Regression: mô hình tuyến tính cơ bản, dễ diễn giải, đóng vai trò mô hình tham chiếu (baseline).

-Gradient Descent: tiếp cận theo hướng tối ưu hóa hàm mất mát, giúp làm rõ quá trình học của mô hình hồi quy.

-Random Forest: mô hình ensemble dựa trên nhiều cây quyết định, có khả năng xử lý tốt dữ liệu phi tuyến và giảm overfitting.

-XGBoost: mô hình boosting tiên tiến, tối ưu hóa sai số theo từng vòng lặp và thường đạt hiệu suất cao trong các bài toán hồi quy.

Cách tiếp cận này cho phép đánh giá toàn diện hiệu quả của từng thuật toán trong cùng một bài toán.

**Tiếp cận đánh giá dựa trên các chỉ số định lượng**

Hiệu quả của các mô hình được đánh giá bằng các chỉ số định lượng khách quan, bao gồm:

-MAE (Mean Absolute Error)

-MSE (Mean Squared Error)

-RMSE (Root Mean Squared Error)

-R² (Coefficient of Determination)

Việc sử dụng nhiều chỉ số đánh giá giúp phản ánh đầy đủ độ chính xác, mức độ sai lệch và khả năng giải thích của mô hình.

**Tiếp cận hướng ứng dụng thực tế**

Ngoài mục tiêu học thuật, đề tài còn hướng tới khả năng ứng dụng thực tế. Mô hình dự đoán GPA có thể được sử dụng như một công cụ hỗ trợ:

-Giúp sinh viên nhận biết sớm nguy cơ học tập kém.

-Hỗ trợ giảng viên và nhà trường trong công tác tư vấn học tập.

-Góp phần nâng cao chất lượng đào tạo dựa trên phân tích dữ liệu.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## 2.1. Trí tuệ nhân tạo AI (Artificial Intelligence)

Trí tuệ nhân tạo (Artificial Intelligence – AI) là một lĩnh vực của khoa học máy tính nghiên cứu và phát triển các hệ thống có khả năng mô phỏng trí thông minh của con người. Các hệ thống AI có thể thực hiện những nhiệm vụ như học tập, suy luận, dự đoán, ra quyết định và giải quyết vấn đề dựa trên dữ liệu.

Trong những năm gần đây, AI được ứng dụng rộng rãi trong nhiều lĩnh vực như y tế, tài chính, giáo dục, giao thông và thương mại điện tử. Đặc biệt, trong lĩnh vực giáo dục, AI đóng vai trò quan trọng trong việc phân tích dữ liệu học tập, cá nhân hóa quá trình học và hỗ trợ ra quyết định.

## 2.2. Học máy (Machine Learning)

Học máy (Machine Learning – ML) là một nhánh quan trọng của Trí tuệ nhân tạo, tập trung vào việc xây dựng các mô hình cho phép máy tính tự động học từ dữ liệu mà không cần lập trình các quy tắc cụ thể.

Thay vì dựa trên các công thức cố định, mô hình học máy sử dụng dữ liệu huấn luyện để tìm ra các mối quan hệ tiềm ẩn giữa đầu vào và đầu ra, từ đó đưa ra dự đoán cho các dữ liệu mới.

Ba nhóm chính của học máy bao gồm:

-Học có giám sát (Supervised Learning)

-Học không giám sát (Unsupervised Learning)

-Học tăng cường (Reinforcement Learning)

Trong phạm vi đề tài này, bài toán dự đoán GPA thuộc nhóm học có giám sát.

### 2.2.1. Học có giám sát (Supervised Learning)

Học có giám sát là phương pháp học máy trong đó mô hình được huấn luyện trên tập dữ liệu đã có sẵn nhãn (label). Mỗi mẫu dữ liệu bao gồm:

-Tập đặc trưng đầu vào

-Giá trị đầu ra mong muốn

Mục tiêu của mô hình là học một hàm ánh xạ từ đến sao cho sai số dự đoán là nhỏ nhất.

Học có giám sát thường được chia thành hai dạng chính:

-Phân loại (Classification): đầu ra là các nhãn rời rạc.

-Hồi quy (Regression): đầu ra là giá trị liên tục

### 2.2.2. Học không giám sát (Unsupervised Learning)

Học không giám sát (Unsupervised Learning) là một phương pháp học máy trong đó mô hình được huấn luyện trên tập dữ liệu không có nhãn. Khác với học có giám sát, dữ liệu đầu vào chỉ bao gồm các đặc trưng mà không có giá trị đầu ra mong muốn .

Mục tiêu của học không giám sát là:

-Khám phá cấu trúc tiềm ẩn trong dữ liệu

-Tìm ra các mẫu (patterns), nhóm (clusters) hoặc mối quan hệ giữa các dữ liệu

-Giảm chiều dữ liệu để biểu diễn thông tin hiệu quả hơn

Thay vì “dự đoán”, học không giám sát tập trung vào việc khai phá dữ liệu

**Các dạng bài toán phổ biến của học không giám sát**

**a) Phân cụm (Clustering)**

Phân cụm là bài toán nhóm các dữ liệu tương tự nhau vào cùng một cụm sao cho:

-Các điểm trong cùng một cụm có độ tương đồng cao

-Các cụm khác nhau có độ khác biệt lớn

Một số thuật toán phân cụm phổ biến:

-K-means

-Hierarchical Clustering

-DBSCAN

Ví dụ:  
Trong lĩnh vực giáo dục, phân cụm có thể được dùng để nhóm sinh viên theo phong cách học tập, mức độ tham gia học tập hoặc hiệu suất học tập mà không cần biết trước GPA.

**b) Giảm chiều dữ liệu (Dimensionality Reduction)**

Giảm chiều dữ liệu là quá trình biến đổi dữ liệu từ không gian nhiều chiều sang không gian ít chiều hơn nhưng vẫn giữ được phần lớn thông tin quan trọng.

Một số phương pháp phổ biến:

-Principal Component Analysis (PCA)

-Autoencoder

Giảm chiều giúp:

-Giảm nhiễu dữ liệu

-Tăng hiệu suất mô hình

-Trực quan hóa dữ liệu dễ dàng hơn

**c) Phát hiện bất thường (Anomaly Detection)**

Phát hiện bất thường là bài toán tìm ra các điểm dữ liệu khác biệt rõ rệt so với phần lớn dữ liệu còn lại.

Ứng dụng phổ biến:

-Phát hiện gian lận

-Phát hiện lỗi hệ thống

-Phát hiện sinh viên có hành vi học tập bất thường

### 2.2.3. Học tăng cường (Reinforcement Learning)

Học tăng cường (Reinforcement Learning – RL) là một phương pháp học máy trong đó một tác nhân (agent) học cách hành động trong một môi trường (environment) thông qua việc tương tác và nhận phản hồi dưới dạng phần thưởng (reward) hoặc hình phạt (penalty).

Không giống như học có giám sát, học tăng cường không yêu cầu dữ liệu có nhãn sẵn. Thay vào đó, tác nhân học thông qua quá trình thử – sai (trial and error) nhằm tối đa hóa tổng phần thưởng tích lũy theo thời gian.

Một bài toán học tăng cường thường bao gồm các thành phần:

-Agent: tác nhân ra quyết định

-Environment: môi trường mà agent tương tác

-State (S): trạng thái hiện tại của môi trường

-Action (A): hành động mà agent có thể thực hiện

-Reward (R): phần thưởng nhận được sau mỗi hành động

-Policy (π): chiến lược lựa chọn hành động của agent

Mục tiêu của học tăng cường là tìm ra chính sách tối ưu giúp tối đa hóa tổng phần thưởng kỳ vọng.

Học tăng cường thường được ứng dụng trong các bài toán như:

-Điều khiển robot

-Trò chơi (cờ vua, cờ vây, game)

-Điều khiển tự động và hệ thống thông minh

Trong phạm vi đề tài dự đoán GPA, học tăng cường không được áp dụng, do bài toán không yêu cầu tương tác theo thời gian hay tối ưu chuỗi hành động, mà chỉ tập trung vào việc dự đoán giá trị GPA dựa trên dữ liệu có sẵn.

### 2.2.4. Bài toán hồi quy trong Trí tuệ nhân tạo

Hồi quy là bài toán dự đoán một biến liên tục dựa trên một hoặc nhiều biến đầu vào. Trong bài toán dự đoán GPA:

-Biến đầu ra là điểm GPA (giá trị liên tục).

-Các biến đầu vào là các yếu tố học tập và thói quen học tập của sinh viên.

Mục tiêu của mô hình hồi quy là tìm ra hàm số sao cho giá trị dự đoán gần nhất với giá trị thực tế của GPA.

### 2.2.5. Hàm mất mát và tối ưu hóa trong học máy (Regression)

**a. Hàm mất mát (Loss Function) là gì?**

Trong bài toán hồi quy, mô hình nhận đầu vào và dự đoán đầu ra (ở đây là GPA dự đoán). Vì dự đoán thường không trùng khớp hoàn toàn với giá trị thực tế (GPA thật), ta cần một đại lượng đo “mức sai lệch” giữa và . Đại lượng đó gọi là hàm mất mát.

Với một mẫu dữ liệu :

-Giá trị thực:

-Giá trị dự đoán:

-Sai số:

Hàm mất mát cho một mẫu thường là .  
Còn với cả tập dữ liệu gồm mẫu, ta dùng hàm mục tiêu (objective) là trung bình loss:

Trong đó là tập tham số của mô hình (ví dụ trọng số của Linear Regression).

Mục tiêu của huấn luyện: tìm để nhỏ nhất.

**b. MAE (Mean Absolute Error)**

Định nghĩa: MAE đo sai số trung bình theo trị tuyệt đối.

Ý nghĩa trực quan: MAE trả lời câu hỏi: “Trung bình mô hình dự đoán lệch bao nhiêu điểm GPA?”  
Ví dụ MAE = 0.25 nghĩa là dự đoán lệch trung bình khoảng 0.25 điểm GPA.

**Ưu điểm**

-Dễ hiểu, đơn vị giống GPA.

-Ít bị ảnh hưởng bởi outlier hơn MSE/RMSE.

**Nhược điểm**

-Do dùng trị tuyệt đối nên hàm loss không “mượt” tại điểm 0 (khó tối ưu bằng giải tích, nhưng vẫn tối ưu được bằng GD).

**c. MSE (Mean Squared Error)**

Định nghĩa: MSE lấy sai số bình phương rồi tính trung bình.

Tại sao bình phương sai số?  
Vì bình phương sẽ phạt rất nặng sai số lớn. Nếu mô hình dự đoán sai quá xa ở một số sinh viên, MSE sẽ tăng mạnh.

**Ưu điểm**

-Trơn và dễ tối ưu (đạo hàm đẹp).

-Rất phổ biến trong hồi quy, đặc biệt cho Linear Regression.

**Nhược điểm**

-Nhạy với outlier: vài dự đoán sai quá lớn có thể làm MSE tăng mạnh.

**d. RMSE (thường dùng kèm MSE)**

MSE có đơn vị “bình phương GPA”, hơi khó diễn giải. Vì vậy người ta hay dùng RMSE:

RMSE trở về cùng đơn vị với GPA, dễ hiểu hơn.  
Nhưng bản chất vẫn “phạt nặng lỗi lớn” như MSE.

**e. R² (Hệ số xác định) trong bài toán hồi quy**

R² (Coefficient of Determination) đo mức độ mà mô hình giải thích được sự biến thiên của biến mục tiêu (ở đây là GPA). Nói đơn giản:

R² cho biết mô hình dự đoán “tốt hơn” bao nhiêu so với việc luôn đoán một giá trị cố định là GPA trung bình.

Công thức R²

Gọi:

: GPA thực tế

: GPA dự đoán

: GPA trung bình trên tập đánh giá

Ta có:

Tổng bình phương sai số (Residual Sum of Squares):

Tổng bình phương toàn phần (Total Sum of Squares):

Khi đó:

**Ý nghĩa của R²**

: mô hình dự đoán hoàn hảo (khớp 100%).

: mô hình không tốt hơn việc đoán (đoán GPA trung bình).

: mô hình còn tệ hơn đoán trung bình (dự đoán sai nhiều).

Ví dụ diễn giải:

: mô hình giải thích được **80%** sự biến thiên của GPA (thường là rất tốt).

: chỉ giải thích được **30%**, mô hình còn yếu hoặc thiếu đặc trưng quan trọng.

**R² khác MAE/MSE/RMSE như thế nào?**

-MAE / MSE / RMSE: đo mức sai số dự đoán (càng nhỏ càng tốt).

-R²: đo mức độ giải thích và hiệu quả so với baseline (càng gần 1 càng tốt).

**Tối ưu hóa (Optimization) trong học máy**

**f. Khái niệm tối ưu hóa**

Trong học máy, tối ưu hóa (Optimization) là quá trình tìm ra tập tham số của mô hình sao cho hàm mất mát đạt giá trị nhỏ nhất. Các tham số này quyết định cách mô hình ánh xạ từ dữ liệu đầu vào đến giá trị dự đoán.

Gọi:

là tập tham số của mô hình

là hàm mục tiêu (objective function), thường là trung bình hàm mất mát trên toàn bộ dữ liệu

Bài toán huấn luyện mô hình được phát biểu dưới dạng:

Trong đó là tập tham số tối ưu giúp mô hình dự đoán tốt nhất.

**Ý nghĩa của các tham số trong Linear Regression**

Với mô hình hồi quy tuyến tính:

Trong đó:

: vector đặc trưng đầu vào (ví dụ: số giờ học, mức độ tham gia lớp học, điểm giữa kỳ…)

: **vector trọng số**, biểu thị mức độ ảnh hưởng của từng đặc trưng đến GPA

: bias (sai số tự do), giúp mô hình dịch chuyển đường hồi quy để phù hợp dữ liệu

: giá trị GPA dự đoán

Ý nghĩa thực tế:

-Trọng số lớn → đặc trưng có ảnh hưởng mạnh đến GPA

-Bias giúp mô hình không bị “ép” đi qua gốc tọa độ

**Hàm mục tiêu với MSE**

Khi sử dụng Mean Squared Error (MSE) làm hàm mất mát, hàm mục tiêu của mô hình hồi quy tuyến tính được viết như sau:

Trong đó:

: GPA thực tế của sinh viên thứ

: GPA dự đoán

: số lượng mẫu trong tập dữ liệu

**Ý nghĩa**:

-Thành phần là sai số dự đoán

-Bình phương sai số giúp phạt mạnh các dự đoán sai lệch lớn

-Lấy trung bình để đánh giá chất lượng mô hình trên toàn bộ tập dữ liệu

**Không gian tham số và điểm tối ưu**

Tập tham số có thể được xem là một điểm trong không gian tham số.  
Mỗi điểm trong không gian này tương ứng với một mô hình khác nhau và một giá trị loss khác nhau.

-Hàm có thể hình dung như một “bề mặt” (loss surface)

-Điểm có giá trị nhỏ nhất chính là điểm tối ưu

-Mục tiêu của tối ưu hóa là di chuyển từ điểm ban đầu đến điểm tối ưu này

Đối với Linear Regression + MSE:

-Hàm loss là lồi (convex)

-Do đó chỉ tồn tại một điểm tối ưu toàn cục (global minimum)

Điều này đảm bảo thuật toán tối ưu như Gradient Descent có thể hội tụ về nghiệm đúng.

**Vai trò của tối ưu hóa trong huấn luyện mô hình**

Quá trình tối ưu hóa đóng vai trò:

-Điều chỉnh các trọng số sao cho mô hình học được mối quan hệ giữa các yếu tố học tập và GPA

-Giảm dần sai số dự đoán qua từng vòng lặp huấn luyện

-Đảm bảo mô hình đạt hiệu suất tốt nhất trên dữ liệu huấn luyện và tổng quát hóa tốt trên dữ liệu mới

Trong bài toán dự đoán GPA, tối ưu hóa giúp mô hình:

-Dự đoán GPA gần với giá trị thực tế hơn

-Giảm sai lệch trung bình (MAE, RMSE)

-Tăng hệ số xác định

**Liên hệ với các mô hình trong đề tài**

-Linear Regression: tìm nghiệm tối ưu của để giảm MSE

-Gradient Descent: là phương pháp tối ưu hóa dùng để tìm

-XGBoost: tối ưu hàm loss bằng gradient boosting (sử dụng đạo hàm bậc nhất và bậc hai)

-Random Forest: không tối ưu theo gradient, mà giảm sai số thông qua trung bình nhiều cây quyết định

### 2.2.6. Overfitting và khả năng tổng quát hóa (Generalization)

**Khái niệm Overfitting**

Overfitting là hiện tượng xảy ra khi mô hình học quá chi tiết các đặc điểm của tập huấn luyện, bao gồm cả nhiễu (noise), dẫn đến việc:

-Sai số trên tập huấn luyện rất thấp

-Nhưng sai số trên dữ liệu mới (chưa từng thấy) lại cao

Nói cách khác, mô hình “ghi nhớ” dữ liệu huấn luyện thay vì học được quy luật tổng quát của bài toán.

Trong bài toán dự đoán GPA, overfitting có thể khiến mô hình:

-Dự đoán rất chính xác GPA của sinh viên trong tập huấn luyện

-Nhưng dự đoán sai đáng kể GPA của các sinh viên mới

**Nguyên nhân gây ra Overfitting**

Một số nguyên nhân phổ biến dẫn đến overfitting bao gồm:

-Mô hình quá phức tạp so với quy mô dữ liệu  
Ví dụ: sử dụng mô hình có quá nhiều tham số cho một tập dữ liệu nhỏ.

-Số lượng đặc trưng lớn nhưng dữ liệu không đủ đa dạng.

-Huấn luyện quá lâu khiến mô hình học cả nhiễu.

-Dữ liệu huấn luyện không đại diện cho dữ liệu thực tế.

Trong đề tài này, các mô hình như Random Forest và XGBoost có khả năng biểu diễn mạnh nên dễ xảy ra overfitting nếu không kiểm soát tốt.

**Khả năng tổng quát hóa (Generalization)**

Khả năng tổng quát hóa là khả năng của mô hình dự đoán chính xác trên dữ liệu mới, chưa xuất hiện trong quá trình huấn luyện.

Một mô hình được coi là tốt khi:

-Sai số trên tập huấn luyện và tập kiểm tra **không chênh lệch quá lớn**

-Các chỉ số đánh giá (MAE, RMSE, R²) trên tập kiểm tra đạt giá trị tốt

Khả năng tổng quát hóa là tiêu chí quan trọng hơn so với việc chỉ đạt độ chính xác cao trên tập huấn luyện.

### 2.2.7. Vai trò của AI trong bài toán dự đoán GPA

Việc áp dụng Trí tuệ nhân tạo vào bài toán dự đoán GPA giúp:

-Phân tích dữ liệu học tập một cách hệ thống và khách quan.

-Phát hiện các yếu tố quan trọng ảnh hưởng đến kết quả học tập.

-Nâng cao độ chính xác của dự đoán so với các phương pháp truyền thống.

-Hỗ trợ ra quyết định trong giáo dục dựa trên dữ liệu.

## 2.3. Các mô hình của bài toán

### 2.3.1. Linear Regression

**Khái niệm Linear Regression**

Linear Regression (Hồi quy tuyến tính) là một trong những thuật toán cơ bản và quan trọng nhất trong học máy có giám sát, được sử dụng để mô hình hóa mối quan hệ giữa một biến phụ thuộc liên tục và một hoặc nhiều biến độc lập.

Mục tiêu của Linear Regression là tìm ra một hàm tuyến tính mô tả mối quan hệ giữa các đặc trưng đầu vào và giá trị đầu ra sao cho sai số dự đoán là nhỏ nhất.

Trong bài toán dự đoán GPA, Linear Regression được sử dụng để mô hình hóa mối quan hệ giữa GPA và các yếu tố như thời gian học tập, mức độ tham gia lớp học, mức độ hỗ trợ của cha mẹ, tần suất tập thể dục, GPA kì trước, …

**Ý tưởng cốt lõi của Linear Regression**

Ý tưởng chính của Linear Regression là: Giả sử rằng giá trị đầu ra có thể được biểu diễn như một tổ hợp tuyến tính của các đặc trưng đầu vào, cộng với một sai số.

Nói cách khác, mỗi đặc trưng đóng góp một phần tuyến tính vào giá trị dự đoán cuối cùng.

Mô hình cố gắng tìm ra các trọng số phù hợp cho từng đặc trưng để đường (hoặc siêu phẳng) hồi quy “khớp” với dữ liệu nhất.

**Mô hình toán học**

**a) Hồi quy tuyến tính đơn (Simple Linear Regression)**

Với một biến đầu vào :

Trong đó:

: hệ số hồi quy (độ dốc của đường thẳng)

: hệ số chệch (bias)

: giá trị dự đoán

**b) Hồi quy tuyến tính đa biến (Multiple Linear Regression)**

Với nhiều biến đầu vào:

Hay viết đầy đủ:

Trong đó:

: đặc trưng thứ

: trọng số tương ứng, thể hiện mức độ ảnh hưởng của đặc trưng đó đến đầu ra

**Ý nghĩa của trọng số và bias**

**Trọng số** :

: đặc trưng tăng → GPA dự đoán tăng

: đặc trưng tăng → GPA dự đoán giảm

lớn → đặc trưng ảnh hưởng mạnh

**Bias** :

-Giúp mô hình dịch chuyển đường hồi quy

-Cho phép mô hình dự đoán linh hoạt hơn, không bị ép đi qua gốc tọa độ

**Nguyên lý học của Linear Regression**

Linear Regression học tham số bằng cách tối thiểu hóa hàm mất mát, thường là Mean Squared Error (MSE):

Quá trình huấn luyện tương đương với việc tìm bộ tham số sao cho tổng sai số bình phương là nhỏ nhất.

**Các phương pháp tìm nghiệm**

**Giải pháp đóng (Normal Equation)**:

Áp dụng khi số chiều không quá lớn.

**Gradient Descent**:

-Dùng khi dữ liệu lớn hoặc số đặc trưng nhiều

-Cập nhật tham số theo hướng giảm dần hàm mất mát

**Các giả định của Linear Regression**

Linear Regression hoạt động tốt khi các giả định sau được thỏa mãn:

-Mối quan hệ giữa đầu vào và đầu ra là tuyến tính

-Sai số có kỳ vọng bằng 0

-Phương sai của sai số là không đổi (homoscedasticity)

-Các quan sát độc lập với nhau

-Không có đa cộng tuyến nghiêm trọng giữa các đặc trưng

**Ưu điểm và nhược điểm**

**Ưu điểm**

-Đơn giản, dễ hiểu và dễ triển khai

-Dễ diễn giải kết quả (interpretability cao)

-Thích hợp làm mô hình baseline

**Nhược điểm**

-Khả năng biểu diễn hạn chế

-Không mô hình hóa tốt quan hệ phi tuyến

-Nhạy cảm với outlier

**Vai trò của Linear Regression trong đề tài**

Trong đề tài dự đoán GPA:

-Linear Regression được sử dụng làm mô hình cơ sở (baseline)

-Giúp đánh giá mức độ tuyến tính giữa các yếu tố học tập và GPA

-Là cơ sở so sánh với các mô hình phức tạp hơn như Random Forest và XGBoost

### 2.3.2. Random Forest

**Khái niệm Random Forest**

Random Forest là một thuật toán học máy thuộc nhóm ensemble learning, trong đó nhiều mô hình đơn giản (các cây quyết định – Decision Tree) được kết hợp lại để tạo thành một mô hình mạnh hơn.

Ý tưởng chính của Random Forest là: Thay vì dựa vào một cây quyết định duy nhất (dễ bị overfitting), mô hình xây dựng một “rừng” gồm nhiều cây quyết định độc lập, sau đó tổng hợp kết quả dự đoán của chúng để đưa ra kết quả cuối cùng.

Trong bài toán dự đoán GPA, Random Forest được sử dụng để mô hình hóa các mối quan hệ phi tuyến giữa các yếu tố học tập và điểm GPA của sinh viên.

**Ý tưởng cốt lõi của Random Forest**

Random Forest dựa trên hai ý tưởng chính:

**a) Bagging (Bootstrap Aggregating)**

-Từ tập dữ liệu ban đầu, tạo ra nhiều tập con khác nhau bằng cách lấy mẫu có hoàn lại (bootstrap sampling).

-Mỗi tập con được dùng để huấn luyện một cây quyết định riêng biệt.

-Nhờ đó, các cây có sự đa dạng và không giống hệt nhau.

**b) Ngẫu nhiên hóa đặc trưng (Feature Randomness)**

-Khi xây dựng mỗi nút trong cây, thuật toán chỉ xét một tập con ngẫu nhiên các đặc trưng thay vì toàn bộ đặc trưng.

-Điều này làm giảm sự phụ thuộc giữa các cây và tăng tính đa dạng của mô hình.

-Hai cơ chế trên giúp Random Forest giảm phương sai (variance) và hạn chế overfitting.

**Mô hình toán học và cơ chế dự đoán**

Giả sử có cây quyết định trong rừng.

-Với bài toán **hồi quy**, mỗi cây đưa ra một giá trị dự đoán .

-Dự đoán cuối cùng của Random Forest là **trung bình** các dự đoán:

Việc lấy trung bình giúp làm “mượt” sai số và tăng độ ổn định của mô hình.

**Xây dựng một cây quyết định trong Random Forest**

Mỗi cây quyết định được xây dựng theo các bước:

-Chọn ngẫu nhiên một tập dữ liệu con bằng bootstrap sampling.

-Tại mỗi nút:

+Chọn ngẫu nhiên một tập con các đặc trưng.

+Tìm phép chia (split) tốt nhất dựa trên tiêu chí giảm sai số (với hồi quy thường là MSE).

-Lặp lại cho đến khi đạt điều kiện dừng (độ sâu tối đa, số mẫu tối thiểu trong lá, v.v.).

**Các siêu tham số quan trọng**

Một số siêu tham số ảnh hưởng mạnh đến hiệu suất của Random Forest:

**-n\_estimators**: số lượng cây trong rừng

**-max\_depth**: độ sâu tối đa của mỗi cây

**-min\_samples\_split**: số mẫu tối thiểu để tiếp tục chia nút

**-min\_samples\_leaf**: số mẫu tối thiểu tại một nút lá

**-max\_features**: số đặc trưng được xét tại mỗi lần chia

Việc điều chỉnh các siêu tham số này giúp cân bằng giữa bias và variance

**Ưu điểm và nhược điểm của Random Forest**

**Ưu điểm**

-Giảm overfitting so với Decision Tree đơn lẻ

-Mô hình hóa tốt quan hệ phi tuyến

-Ít nhạy cảm với outlier

-Hoạt động tốt với dữ liệu nhiều đặc trưng

**Nhược điểm**

-Khó diễn giải hơn Linear Regression

-Tốn tài nguyên tính toán khi số cây lớn

-Kích thước mô hình lớn

### 2.3.3. Gradient Boosting

**Khái niệm Gradient Boosting**

Gradient Boosting là một thuật toán học máy thuộc nhóm ensemble learning, kết hợp nhiều mô hình yếu (weak learners), thường là cây quyết định nông (shallow decision trees), theo cách học nối tiếp (sequential learning) nhằm cải thiện dần độ chính xác của mô hình.

Khác với Random Forest (huấn luyện các cây độc lập), Gradient Boosting xây dựng các mô hình phụ thuộc lẫn nhau, trong đó mỗi mô hình mới được huấn luyện để sửa các lỗi mà mô hình trước đó mắc phải.

**Ý tưởng cốt lõi của Gradient Boosting**

Ý tưởng trung tâm của Gradient Boosting là:

-Mô hình mới sẽ học để dự đoán phần sai số (residual) của mô hình hiện tại, từ đó từng bước cải thiện chất lượng dự đoán tổng thể.

-Quá trình này được lặp lại nhiều lần, mỗi lần bổ sung thêm một mô hình yếu để giảm dần sai số.

**Mô hình toán học tổng quát**

Giả sử:

: giá trị GPA thực tế

: dự đoán của mô hình tại vòng lặp thứ

Mô hình Gradient Boosting có dạng:

Trong đó:

: mô hình yếu thứ (thường là cây quyết định)

: learning rate, kiểm soát mức độ đóng góp của mỗi mô hình mới

**Vai trò của hàm mất mát và Gradient**

Gradient Boosting hoạt động dựa trên gradient của hàm mất mát, do đó có tên gọi là “Gradient” Boosting.

Với hàm mất mát , tại mỗi vòng lặp, thuật toán:

-Tính gradient của loss theo dự đoán:

-Huấn luyện mô hình yếu để dự đoán các giá trị

-Cập nhật mô hình tổng hợp

Với bài toán hồi quy sử dụng MSE, gradient chính là residual:

**Quy trình huấn luyện Gradient Boosting**

Quy trình tổng quát gồm các bước:

-Khởi tạo mô hình ban đầu (thường là giá trị trung bình của )

-Tính sai số (residual) giữa dự đoán và giá trị thực

-Huấn luyện cây quyết định để dự đoán residual

-Cập nhật mô hình tổng thể

-Lặp lại cho đến khi đạt số vòng lặp mong muốn

**Các siêu tham số quan trọng**

Một số siêu tham số ảnh hưởng mạnh đến hiệu suất của Gradient Boosting:

**-n\_estimators**: số lượng mô hình yếu

**-learning\_rate**: tốc độ học

**-max\_depth**: độ sâu của cây quyết định

**-subsample**: tỷ lệ mẫu dùng để huấn luyện mỗi vòng (giúp giảm overfitting)

Việc điều chỉnh các siêu tham số này giúp cân bằng giữa bias và variance.

**Ưu điểm và nhược điểm của Gradient Boosting**

**Ưu điểm**

-Hiệu suất cao, đặc biệt với dữ liệu phức tạp

-Mô hình hóa tốt quan hệ phi tuyến

-Linh hoạt với nhiều hàm mất mát khác nhau

**Nhược điểm**

-Dễ overfitting nếu không kiểm soát tốt

-Thời gian huấn luyện lâu hơn

-Khó diễn giải hơn các mô hình tuyến tính

### 2.3.4. XGBoost

**Khái niệm XGBoost**

XGBoost (Extreme Gradient Boosting) là một thuật toán học máy thuộc nhóm ensemble learning, được phát triển dựa trên Gradient Boosting nhưng được tối ưu mạnh mẽ hơn về:

-Hiệu suất tính toán

-Khả năng kiểm soát overfitting

-Độ chính xác dự đoán

XGBoost hiện là một trong những thuật toán phổ biến và hiệu quả nhất trong các bài toán hồi quy và phân loại trên dữ liệu .

**Ý tưởng cốt lõi của XGBoost**

Ý tưởng trung tâm của XGBoost là: Xây dựng mô hình dự đoán bằng cách cộng dồn nhiều cây quyết định, trong đó mỗi cây mới được huấn luyện để giảm sai số của mô hình hiện tại, đồng thời kiểm soát độ phức tạp của mô hình thông qua regularization.

**So với Gradient Boosting truyền thống, XGBoost cải tiến ở các điểm:**

-Sử dụng gradient bậc nhất và bậc hai của hàm mất mát

-Thêm regularization trực tiếp vào hàm mục tiêu

-Tối ưu tốc độ huấn luyện và khả năng mở rộng

**Mô hình toán học tổng quát**

Giả sử mô hình tại vòng lặp thứ có dạng:

Trong đó:

-: cây quyết định thứ

-: vector đặc trưng của mẫu thứ

Mục tiêu của XGBoost là tối thiểu hóa hàm mục tiêu tổng quát:

**Hàm regularization trong XGBoost**

Điểm nổi bật của XGBoost là hàm regularization:

Trong đó:

-: số nút lá của cây

-: giá trị tại nút lá thứ

-: phạt số lượng nút lá (độ phức tạp cấu trúc)

-: phạt độ lớn trọng số (L2 regularization)

Regularization giúp:

-Hạn chế overfitting

-Ưu tiên mô hình đơn giản nhưng hiệu quả

**Gradient bậc nhất và bậc hai**

XGBoost sử dụng khai triển Taylor bậc hai của hàm mất mát:

Trong đó:

: gradient bậc nhất

: gradient bậc hai

Việc sử dụng gradient bậc hai giúp XGBoost:

-Hội tụ nhanh hơn

-Chọn phép chia cây tối ưu hơn

**Cơ chế xây dựng cây trong XGBoost**

Khi xây dựng cây:

-XGBoost đánh giá **độ lợi (gain)** của mỗi phép chia

-Chỉ thực hiện phép chia nếu gain >

-Điều này giúp tránh tạo các nhánh không cần thiết

Công thức gain thể hiện mức giảm loss khi chia một nút thành hai nút con.

**Các siêu tham số quan trọng**

Một số siêu tham số quan trọng của XGBoost:

-n\_estimators: số cây

-learning\_rate: tốc độ học

-max\_depth: độ sâu tối đa của cây

-subsample: tỷ lệ mẫu dùng mỗi vòng

-colsample\_bytree: tỷ lệ đặc trưng sử dụng

-lambda, gamma: regularization

Việc tinh chỉnh các tham số này giúp cân bằng giữa bias và variance.

**Ưu điểm và nhược điểm của XGBoost**

**Ưu điểm**

-Độ chính xác rất cao

-Kiểm soát overfitting tốt

-Hoạt động tốt với dữ liệu phức tạp

-Tối ưu hóa hiệu quả về mặt tính toán

**Nhược điểm**

-Nhiều siêu tham số, khó tinh chỉnh

-Khó diễn giải hơn Linear Regression

-Thời gian huấn luyện có thể dài nếu cấu hình chưa hợp lý

Bảng 2.1. Bảng so sánh các thuật toán trong mô hình

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tiêu chí** | **Linear Regression** | **Random Forest** | **Gradient Boosting** | **XGBoost** |
| Nhóm thuật toán | Hồi quy tuyến tính | Ensemble (Bagging) | Ensemble (Boosting) | Ensemble (Boosting nâng cao) |
| Kiểu học | Học có giám sát | Học có giám sát | Học có giám sát | Học có giám sát |
| Cách huấn luyện | Một mô hình duy nhất | Nhiều cây huấn luyện song song | Nhiều cây huấn luyện nối tiếp | Nhiều cây huấn luyện nối tiếp |
| Mối quan hệ mô hình hóa | Tuyến tính | Phi tuyến | Phi tuyến | Phi tuyến phức tạp |
| Khả năng học quan hệ phức tạp | Thấp | Tốt | Rất tốt | Rất tốt |
| Overfitting | Thấp (nếu mô hình đơn giản) | Ít hơn Decision Tree | Dễ nếu không kiểm soát | Kiểm soát tốt |
| Regularization | Không (mặc định) | Không trực tiếp | Hạn chế | Có (L1/L2, gamma) |
| Gradient của loss | Có thể dùng | Không | Gradient bậc nhất | Gradient bậc nhất & bậc hai |
| Độ chính xác dự đoán | Thấp – Trung bình | Trung bình – Cao | Cao | Rất cao |
| Khả năng diễn giải | Rất cao | Trung bình | Thấp | Thấp |
| Tốc độ huấn luyện | Rất nhanh | Trung bình | Chậm hơn RF | Nhanh (đã tối ưu) |
| Số siêu tham số | Ít | Trung bình | Nhiều | Rất nhiều |
| Nhạy cảm với outlier | Cao | Thấp | Trung bình | Thấp |
| Phù hợp làm baseline | Rất phù hợp | Không phù hợp | Không phù hợp | Không phù hợp |
| Phù hợp dữ liệu tabular | Trung bình | Tốt | Tốt | Rất tốt |

CHƯƠNG 3: NỘI DUNG THỰC HIỆN

## 3.1. Khảo sát và xác định yêu cầu

### 3.1.1. Mô hình

Trong bối cảnh giáo dục hiện nay, điểm trung bình học tập (GPA) là một trong những chỉ số quan trọng phản ánh năng lực và kết quả học tập của sinh viên. Việc dự đoán GPA dựa trên các dữ liệu liên quan đến quá trình học tập có thể giúp sinh viên, giảng viên và nhà trường sớm nhận diện tình trạng học tập, từ đó đưa ra các biện pháp hỗ trợ kịp thời.

Qua khảo sát bộ dữ liệu sử dụng trong đề tài, có thể nhận thấy dữ liệu bao gồm các thuộc tính phản ánh thói quen và quá trình học tập của sinh viên như thời gian học tập, mức độ tham gia lớp học, kết quả các bài kiểm tra, cùng một số đặc trưng cá nhân khác. Dữ liệu đã có sẵn nhãn là điểm GPA, phù hợp để xây dựng mô hình dự đoán.

Do đó, bài toán được xác định là dự đoán giá trị GPA của sinh viên dựa trên các đặc trưng học tập.

### 3.1.2. Xác định đầu vào và đầu ra của mô hình

Đầu vào (Input): Tập các đặc trưng phản ánh quá trình và thói quen học tập của sinh viên, bao gồm các biến số và biến phân loại liên quan đến hoạt động học tập.

Đầu ra (Output): Điểm trung bình học tập (GPA) của sinh viên, là một giá trị liên tục.

Việc xác định rõ đầu vào và đầu ra giúp định hướng đúng phương pháp và thuật toán học máy cần sử dụng.

### 3.1.3. Xác định loại bài toán

Dựa trên đặc điểm dữ liệu và mục tiêu nghiên cứu, bài toán thuộc nhóm:

-Học máy có giám sát (Supervised Learning) do dữ liệu đầu vào đã có sẵn nhãn GPA.

-Bài toán hồi quy (Regression) do giá trị cần dự đoán là một biến liên tục.

### 3.1.4. Xác định yêu cầu chức năng

Các yêu cầu chức năng của bài toán bao gồm:

-Xây dựng mô hình có khả năng dự đoán điểm GPA của sinh viên dựa trên dữ liệu đầu vào.

-Áp dụng và so sánh nhiều thuật toán học máy khác nhau nhằm đánh giá hiệu quả dự đoán.

-Đánh giá mô hình dựa trên các chỉ số phù hợp như MAE, RMSE và R².

-Lựa chọn mô hình có hiệu suất tốt nhất để phục vụ mục tiêu nghiên cứu.

### 3.1.5. Xác định yêu cầu phi chức năng

Bên cạnh các yêu cầu chức năng, bài toán còn đặt ra một số yêu cầu phi chức năng:

-Mô hình có thời gian huấn luyện và dự đoán hợp lý.

-Kết quả dự đoán và đánh giá dễ hiểu, dễ so sánh.

-Mô hình có khả năng mở rộng khi dữ liệu được bổ sung trong tương lai.

## 3.2. Thu thập và mô tả dữ liệu

### 3.2.1. Thu thập dữ liệu

Dữ liệu sử dụng trong đề tài được thu thập từ một bộ dữ liệu có sẵn trên Kaggle dưới dạng tệp CSV, bao gồm các thông tin liên quan đến quá trình học tập và thói quen học tập của sinh viên. Bộ dữ liệu này được sử dụng cho mục đích nghiên cứu và học tập, phù hợp với bài toán dự đoán điểm trung bình học tập (GPA).

Dữ liệu được lưu trữ và xử lý trong môi trường lập trình Python, sử dụng các thư viện phổ biến như Pandas và NumPy để hỗ trợ quá trình đọc, xử lý và phân tích dữ liệu.

### 3.2.2. Quy mô và cấu trúc dữ liệu

Bộ dữ liệu bao gồm 80000 bản ghi, trong đó mỗi bản ghi tương ứng với 20 cột của một sinh viên. Mỗi bản ghi chứa tập các thuộc tính phản ánh thông tin học tập và kết quả học tập của sinh viên.

Cấu trúc dữ liệu gồm:

Các biến đầu vào (Features): Các thuộc tính phản ánh hành vi, thói quen trong đời sống sinh hoạt và kết quả học tập của sinh viên.

Biến mục tiêu (Target): Điểm trung bình học tập (GPA).

Dữ liệu bao gồm cả: Biến số (Numerical features), Biến phân loại (Categorical features)

Điều này đòi hỏi quá trình tiền xử lý phù hợp trước khi đưa vào huấn luyện mô hình.

### 3.2.3. Mô tả các thuộc tính dữ liệu

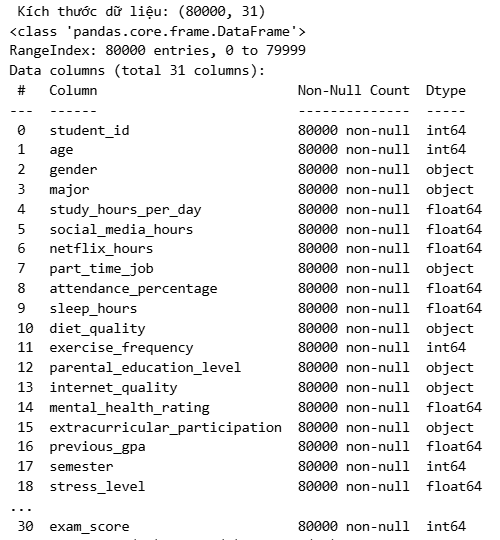
Bộ dữ liệu sử dụng trong đề tài bao gồm các thuộc tính phản ánh thông tin cá nhân, thói quen học tập, điều kiện sinh hoạt và kết quả học tập của sinh viên. Mỗi bản ghi tương ứng với một sinh viên trong một học kỳ cụ thể. Các thuộc tính được mô tả chi tiết trong Bảng dưới đây.

Bảng 3.1. Bảng mô tả các thuộc tính dữ liệu

|  |  |  |
| --- | --- | --- |
| **Tên thuộc tính** | **Kiểu dữ liệu** | **Mô tả** |
| study\_hours\_per\_day | Số thực | Số giờ học trung bình mỗi ngày |
| social\_media\_hours | Số thực | Thời gian sử dụng mạng xã hội mỗi ngày |
| part\_time\_job | Phân loại | Sinh viên có làm thêm hay không |
| attendance\_percentage | Số thực | Tỷ lệ tham gia lớp học (%) |
| sleep\_hours | Số thực | Số giờ ngủ trung bình mỗi ngày |
| diet\_quality | Phân loại | Chất lượng chế độ ăn uống |
| exercise\_frequency | Phân loại | Tần suất tập thể dục ( lần/tuần) |
| mental\_health\_rating | Số nguyên | Đánh giá tình trạng sức khỏe tinh thần |
| previous\_gpa | Số thực | GPA của học kỳ trước |
| stress\_level | Phân loại | Mức độ căng thẳng |
| dropout\_risk | Phân loại | Nguy cơ bỏ học |
| screen\_time | Số thực | Thời gian sử dụng thiết bị điện tử |
| study\_environment | Phân loại | Môi trường học tập |
| access\_to\_tutoring | Phân loại | Khả năng tiếp cận gia sư/hỗ trợ học tập |
| parental\_support\_level | Phân loại | Mức độ hỗ trợ từ gia đình |
| motivation\_level | Phân loại | Mức độ động lực học tập |
| exam\_anxiety\_score | Số nguyên | Mức độ lo lắng khi thi |
| learning\_style | Phân loại | Phong cách học tập |
| time\_management\_score | Số nguyên | Kỹ năng quản lý thời gian |
| exam\_score | Số thực | Điểm thi trung bình năm học |

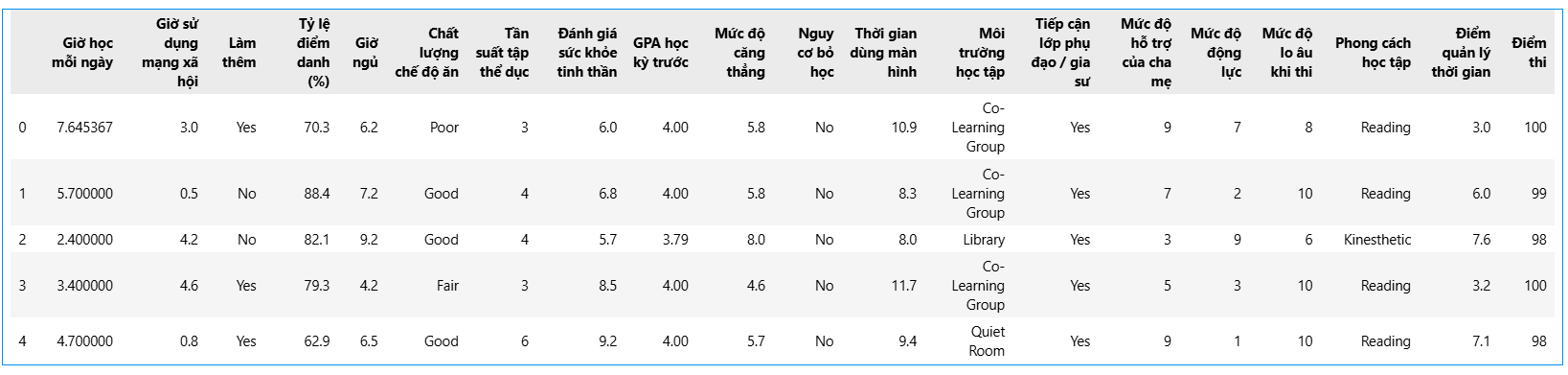
### 3.2.4. Khảo sát sơ bộ dữ liệu

**a. Kích thước dữ liệu**

****

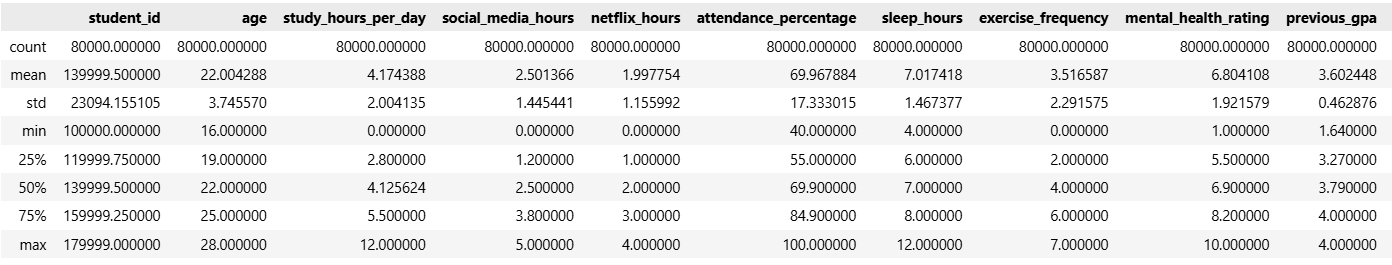
Hình 3.1. Thông tin tổng quan của dữ liệu

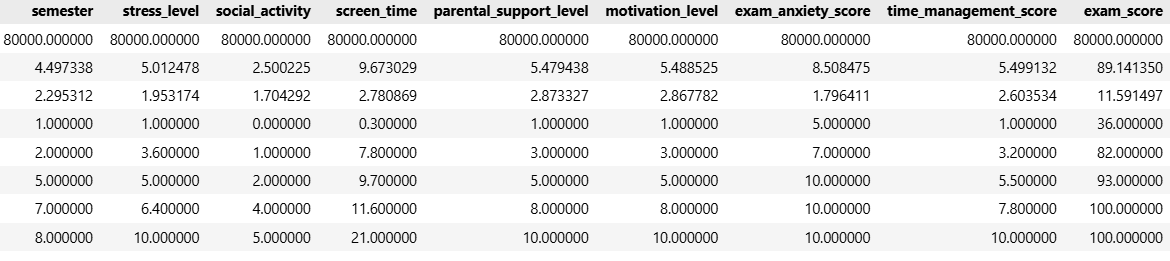
**b. Các feature trong dữ liệu**



Hình 3.2. 5 dòng đầu tiên trong dataset

**c. Thống kê dữ liệu**

****

**

Hình 3.3. Thống kê tổng quát dữ liệu

## 3.3. Tiền xử lý dữ liệu

### 3.3.1. Mục đích của tiền xử lý dữ liệu

Tiền xử lý dữ liệu là bước quan trọng nhằm đảm bảo dữ liệu đầu vào có chất lượng tốt, phù hợp với các thuật toán học máy. Dữ liệu thô ban đầu thường chứa các giá trị thiếu, sự khác biệt về thang đo, cũng như các thuộc tính phân loại chưa thể sử dụng trực tiếp cho mô hình.

Mục tiêu của bước tiền xử lý là:

-Làm sạch dữ liệu

-Chuẩn hóa dữ liệu về cùng thang đo

-Biến đổi dữ liệu về dạng phù hợp cho quá trình huấn luyện mô hình

-Nâng cao hiệu quả và độ ổn định của các thuật toán học máy

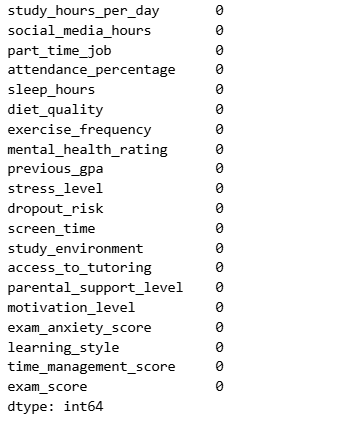
### 3.3.2. Kiểm tra và xử lý giá trị thiếu (Missing Values)

Trước tiên, dữ liệu được kiểm tra để xác định sự tồn tại của các giá trị thiếu trong từng thuộc tính. Việc xử lý giá trị thiếu được thực hiện như sau:

-Đối với các biến số (numerical features): Các giá trị thiếu được thay thế bằng giá trị trung bình hoặc trung vị của cột nhằm hạn chế ảnh hưởng của outlier.

-Đối với các biến phân loại (categorical features): Các giá trị thiếu được thay thế bằng giá trị xuất hiện nhiều nhất (mode) hoặc gán nhãn “Unknown”.

Cách tiếp cận này giúp giữ lại tối đa số lượng mẫu và tránh làm sai lệch phân phối dữ liệu.

****

Hình 3.4. Kiểm tra dữ liệu bị thiếu

### 3.3.3. Xử lý dữ liệu phân loại (Categorical Encoding)

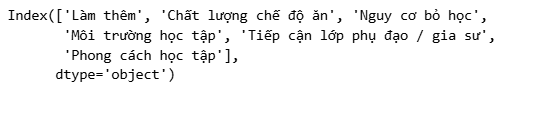
Bộ dữ liệu chứa nhiều thuộc tính dạng phân loại như: gender, major, diet\_quality, exercise\_frequency, stress\_level, learning\_style, …

Do các thuật toán học máy không thể xử lý trực tiếp dữ liệu dạng chuỗi, các thuộc tính này được mã hóa theo các phương pháp phù hợp:

-Label Encoding: Áp dụng cho các thuộc tính có tính thứ bậc (ordinal) như stress\_level, diet\_quality, motivation\_level.

-One-Hot Encoding: Áp dụng cho các thuộc tính không có thứ tự như gender, major, learning\_style.

Việc lựa chọn phương pháp mã hóa phù hợp giúp mô hình học được đúng ý nghĩa của từng thuộc tính.

****

3.5. Kiểm tra các cột có dữ liệu chữ

### 3.3.4. Chuẩn hóa dữ liệu số (Feature Scaling)

Các thuộc tính số trong dữ liệu như study\_hours\_per\_day, attendance\_percentage, sleep\_hours, exam\_score, previous\_gpa có thang đo khác nhau.

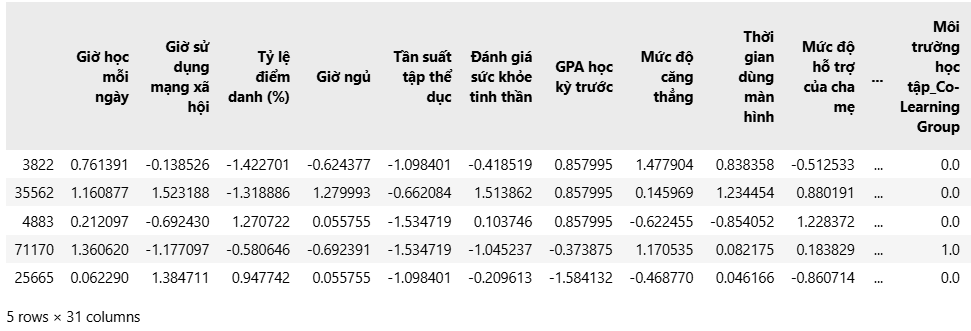
Để đảm bảo quá trình huấn luyện mô hình diễn ra hiệu quả, các thuộc tính số được chuẩn hóa bằng phương pháp StandardScaler, đưa dữ liệu về phân phối có trung bình bằng 0 và độ lệch chuẩn bằng 1.

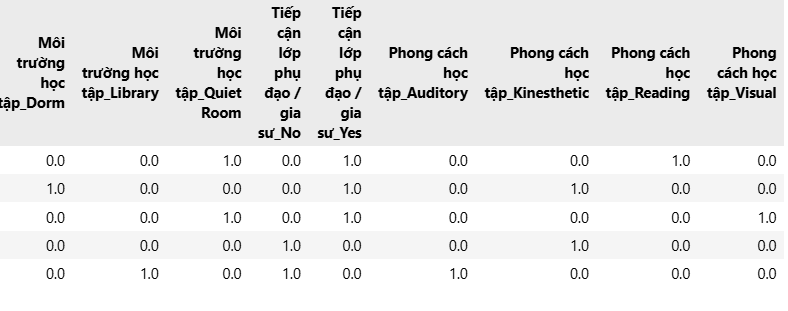
Việc chuẩn hóa đặc biệt quan trọng đối với các mô hình:

-Linear Regression

-Gradient Descent

Trong khi đó, các mô hình dựa trên cây như Random Forest và XGBoost ít bị ảnh hưởng bởi sự khác biệt thang đo, nhưng vẫn được sử dụng chung tập dữ liệu đã xử lý để đảm bảo tính thống nhất.





Hình 3.6. Dữ liệu sau khi đã chuẩn hóa dữ liệu số và dữ liệu chữ

## 3.4. Phân tích và khám phá dữ liệu (Exploratory Data Analysis – EDA)

### 3.4.1. Mục đích của phân tích và khám phá dữ liệu

Phân tích và khám phá dữ liệu (EDA) được thực hiện nhằm:

-Hiểu rõ đặc điểm và phân phối của dữ liệu

-Khám phá mối quan hệ giữa các thuộc tính đầu vào và điểm GPA

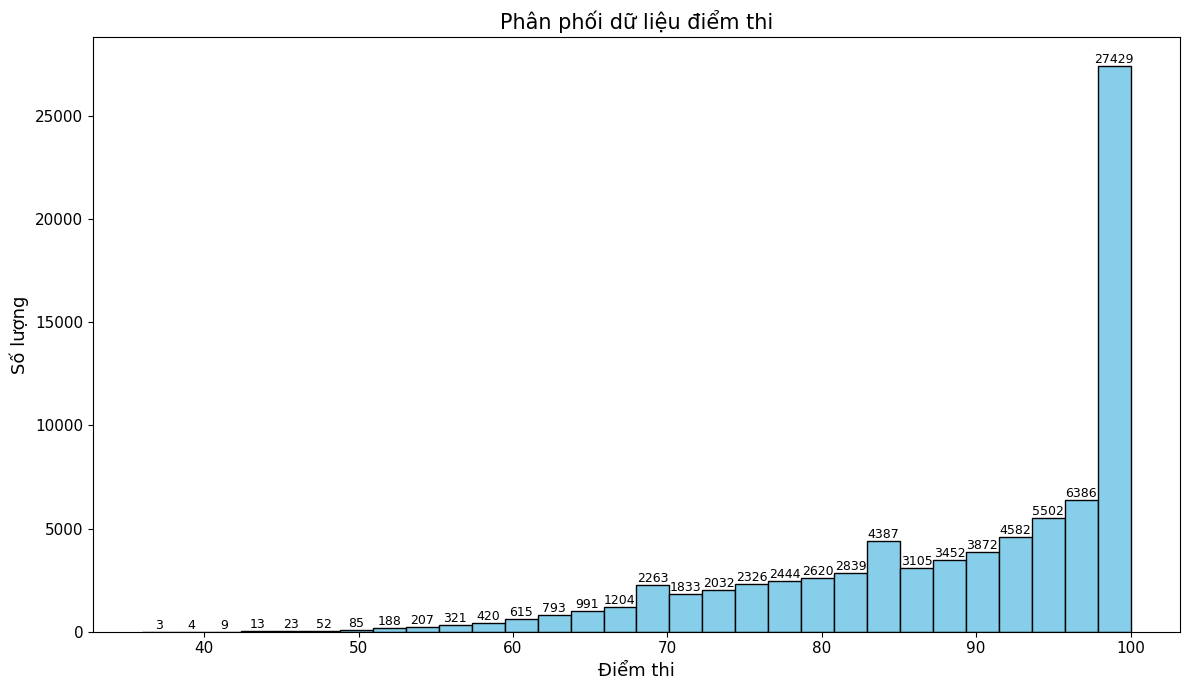
-Phát hiện các xu hướng, quy luật tiềm ẩn và các yếu tố có ảnh hưởng lớn đến kết quả học tập

-Hỗ trợ việc lựa chọn và đánh giá mô hình học máy phù hợp

EDA đóng vai trò quan trọng trong việc định hướng các bước xây dựng và huấn luyện mô hình ở giai đoạn sau.

### 3.4.2. Phân tích phân phối điểm thi (Exam Score)

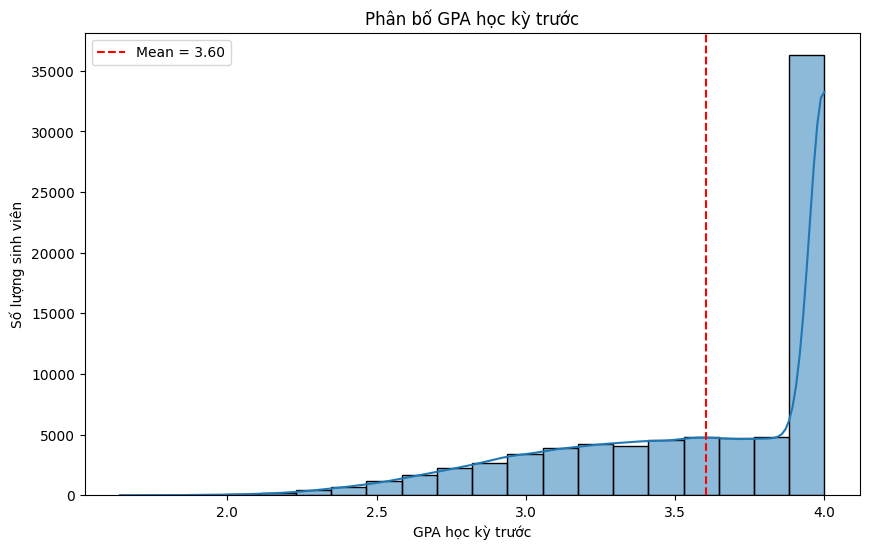
Biểu đồ cho thấy điểm thi của sinh viên tập trung chủ yếu ở mức cao, đặc biệt trong khoảng từ 80 đến 100 điểm. Số lượng sinh viên đạt điểm thấp là rất ít, cho thấy phân phối điểm bị lệch về phía điểm cao. Điều này phản ánh đa số sinh viên có kết quả thi tốt và điểm thi là một đặc trưng quan trọng ảnh hưởng đến GPA. Phân phối không chuẩn này gợi ý nên sử dụng các mô hình có khả năng xử lý quan hệ phi tuyến như Random Forest hoặc XGBoost.



Hình 3.7. Biểu đồ phân bố dữ liệu điểm thi

### 3.4.3. Phân bố GPA học kỳ trước

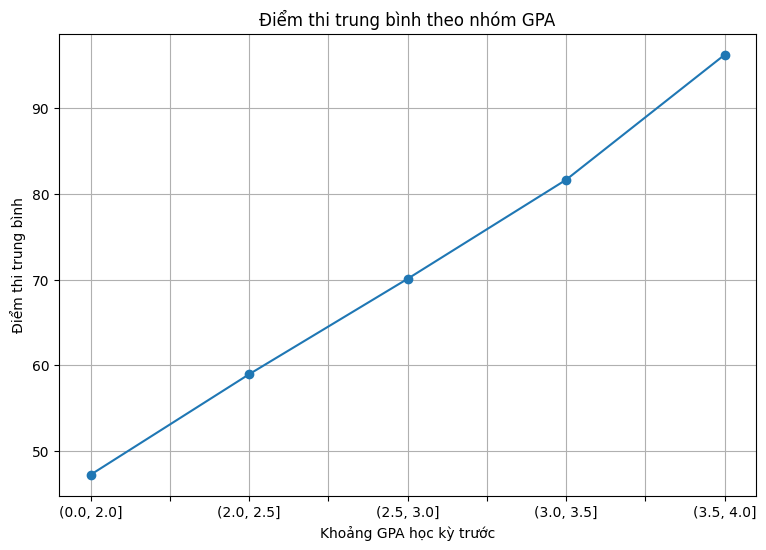
Biểu đồ cho thấy GPA học kỳ trước của sinh viên tập trung chủ yếu ở mức cao, đặc biệt trong khoảng từ 3.5 đến 4.0. Giá trị trung bình của GPA học kỳ trước vào khoảng 3.60, phản ánh mặt bằng học lực chung khá tốt. Phân phối GPA có xu hướng lệch về phía giá trị cao, cho thấy số lượng sinh viên có GPA thấp chiếm tỷ lệ nhỏ. Điều này gợi ý rằng GPA học kỳ trước là một biến quan trọng và có khả năng dự đoán mạnh đối với GPA hiện tại.



Hình 3.8. Biểu đồ phân bố dữ liệu điểm GPA học kì trước

### 3.4.4. Mối quan hệ giữa GPA học kỳ trước và điểm thi trung bình

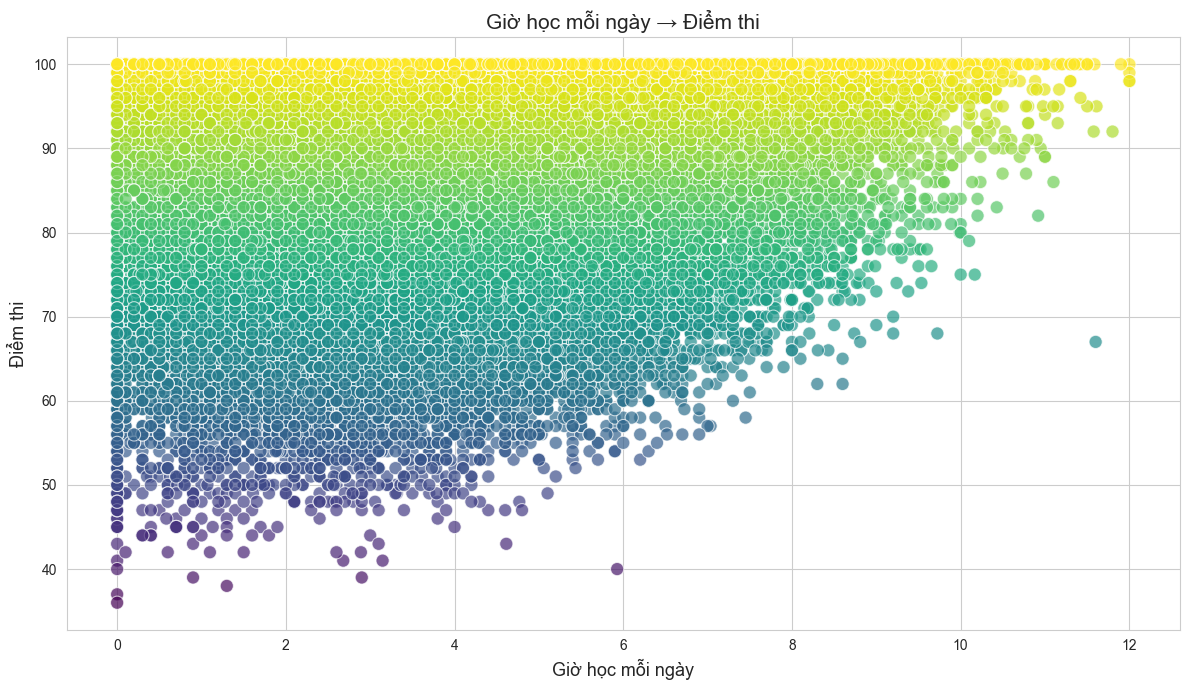
Biểu đồ cho thấy điểm thi trung bình tăng dần theo từng nhóm GPA học kỳ trước, thể hiện mối quan hệ đồng biến rõ rệt giữa hai biến. Sinh viên có GPA học kỳ trước càng cao thì điểm thi trung bình hiện tại càng lớn. Xu hướng tăng gần như tuyến tính cho thấy GPA học kỳ trước là một yếu tố dự báo rất mạnh đối với kết quả thi. Điều này khẳng định tầm quan trọng của biến GPA học kỳ trước trong mô hình dự đoán GPA hiện tại.



Hình 3.9. Biểu đồ trực quan hóa mối quan hệ điểm GPA học kì trước với điểm thi

### 3.4.5. Mối quan hệ giữa giờ học mỗi ngày và điểm thi

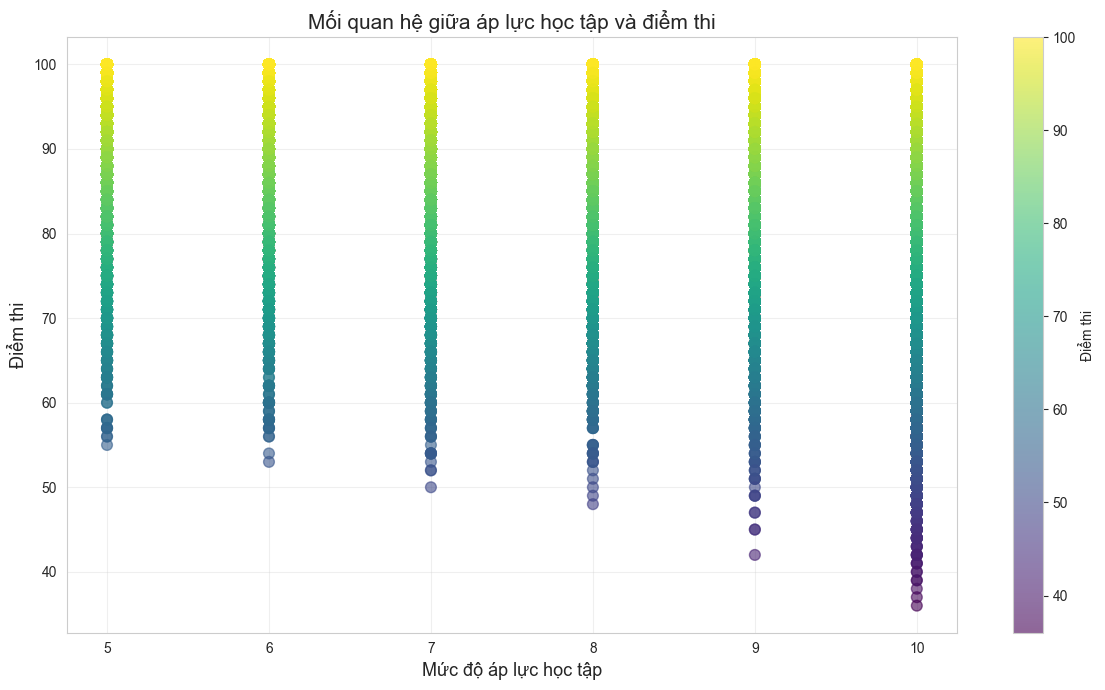
Biểu đồ cho thấy mối quan hệ đồng biến rõ rệt giữa số giờ học mỗi ngày và điểm thi. Khi thời gian học tăng, điểm thi có xu hướng cải thiện theo, đặc biệt rõ ở nhóm sinh viên học từ khoảng 6 giờ trở lên mỗi ngày. Tuy nhiên, ở cùng một mức giờ học vẫn tồn tại độ phân tán nhất định về điểm thi, cho thấy kết quả học tập còn chịu ảnh hưởng của các yếu tố khác ngoài thời gian học



Hình 3.10. Biểu đồ trực quan hóa mối quan hệ giờ học mỗi ngày với điểm thi

### 3.4.6. Mối quan hệ giữa áp lực học tập và điểm thi

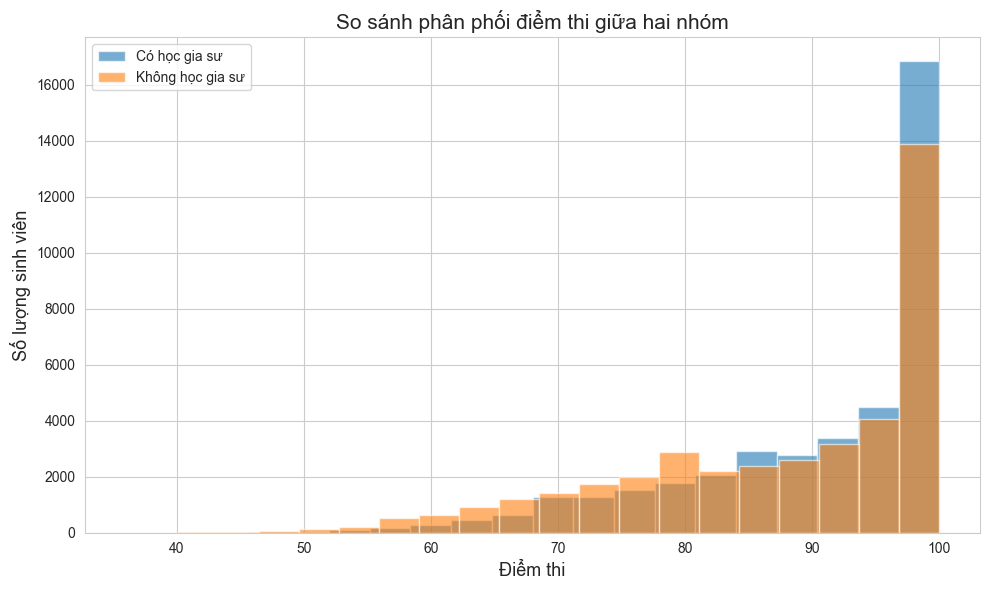
Biểu đồ cho thấy khi mức độ áp lực học tập tăng, điểm thi có xu hướng giảm nhẹ, đặc biệt rõ ở các mức áp lực cao. Ở mỗi mức áp lực vẫn tồn tại độ phân tán lớn về điểm thi, cho thấy áp lực không phải là yếu tố duy nhất quyết định kết quả học tập. Áp lực ở mức vừa phải vẫn cho phép sinh viên đạt điểm cao, nhưng áp lực quá cao có thể ảnh hưởng tiêu cực đến hiệu suất. Điều này cho thấy mối quan hệ giữa áp lực học tập và điểm thi mang tính phi tuyến.

**

Hình 3.11. Biểu đồ trực quan hóa mối quan hệ giữa áp lực học tập với điểm thi

### 3.4.7. So sánh điểm thi giữa nhóm có và không có học gia sư

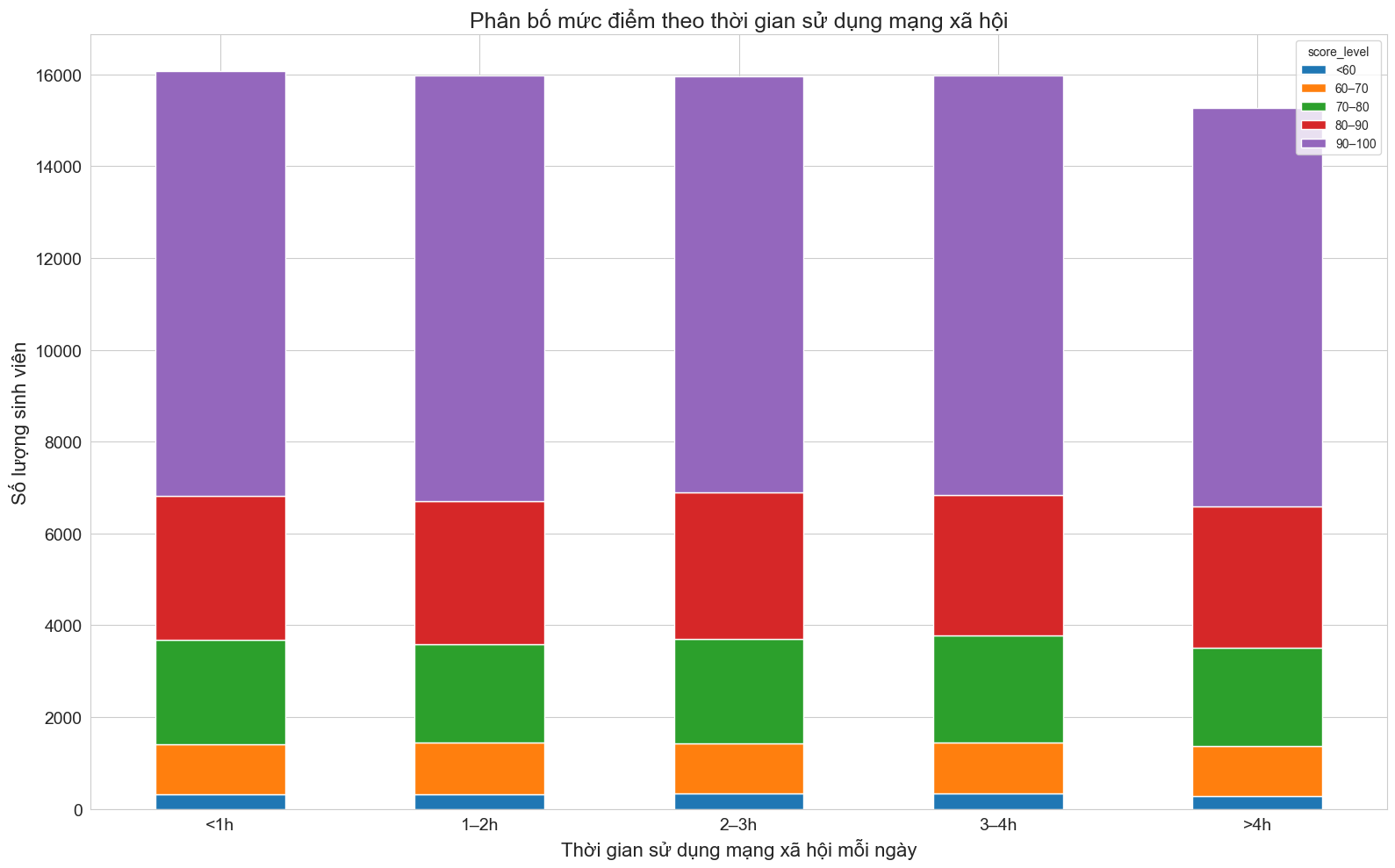
Biểu đồ cho thấy nhóm sinh viên có học gia sư có xu hướng đạt điểm thi cao hơn so với nhóm không học gia sư, đặc biệt ở vùng điểm từ 85 đến 100. Phân phối điểm của nhóm có gia sư tập trung nhiều hơn ở mức điểm cao, trong khi nhóm không học gia sư có phân bố trải rộng hơn về các mức điểm thấp. Điều này cho thấy việc có gia sư hỗ trợ học tập có ảnh hưởng tích cực đến kết quả thi.



Hình 3.12. Biểu đồ trực quan hóa mối quan hệ giữa việc học gia sư với điểm thi

### 3.4.8. Phân bố điểm thi theo thời gian sử dụng mạng xã hội

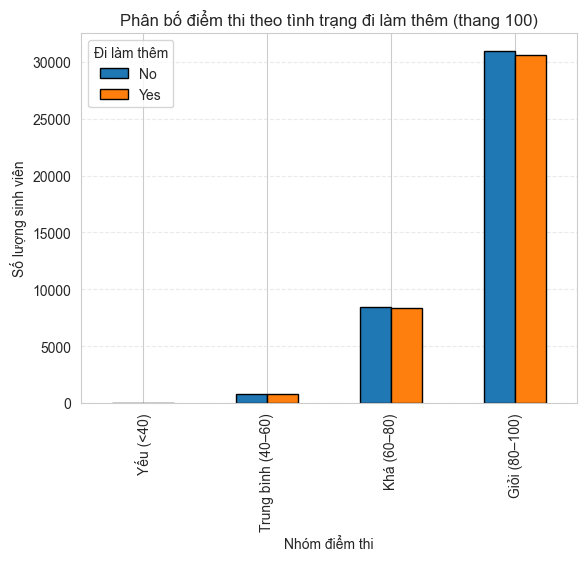
Biểu đồ cho thấy khi thời gian sử dụng mạng xã hội tăng, tỷ lệ sinh viên đạt điểm cao (90–100) có xu hướng giảm nhẹ, trong khi tỷ lệ sinh viên ở các mức điểm trung bình và thấp tăng lên. Nhóm sinh viên sử dụng mạng xã hội trên 4 giờ/ngày có ít sinh viên đạt điểm rất cao hơn so với các nhóm sử dụng ít hơn. Tuy nhiên, sự khác biệt giữa các nhóm không quá lớn, cho thấy mạng xã hội là yếu tố ảnh hưởng gián tiếp đến kết quả học tập.



Hình 3.13. Biểu đồ trực quan hóa mối quan hệ giữa thời gian sử dụng mạng xã hội với điểm thi

### 3.4.9. Phân bố điểm thi theo tình trạng đi làm thêm

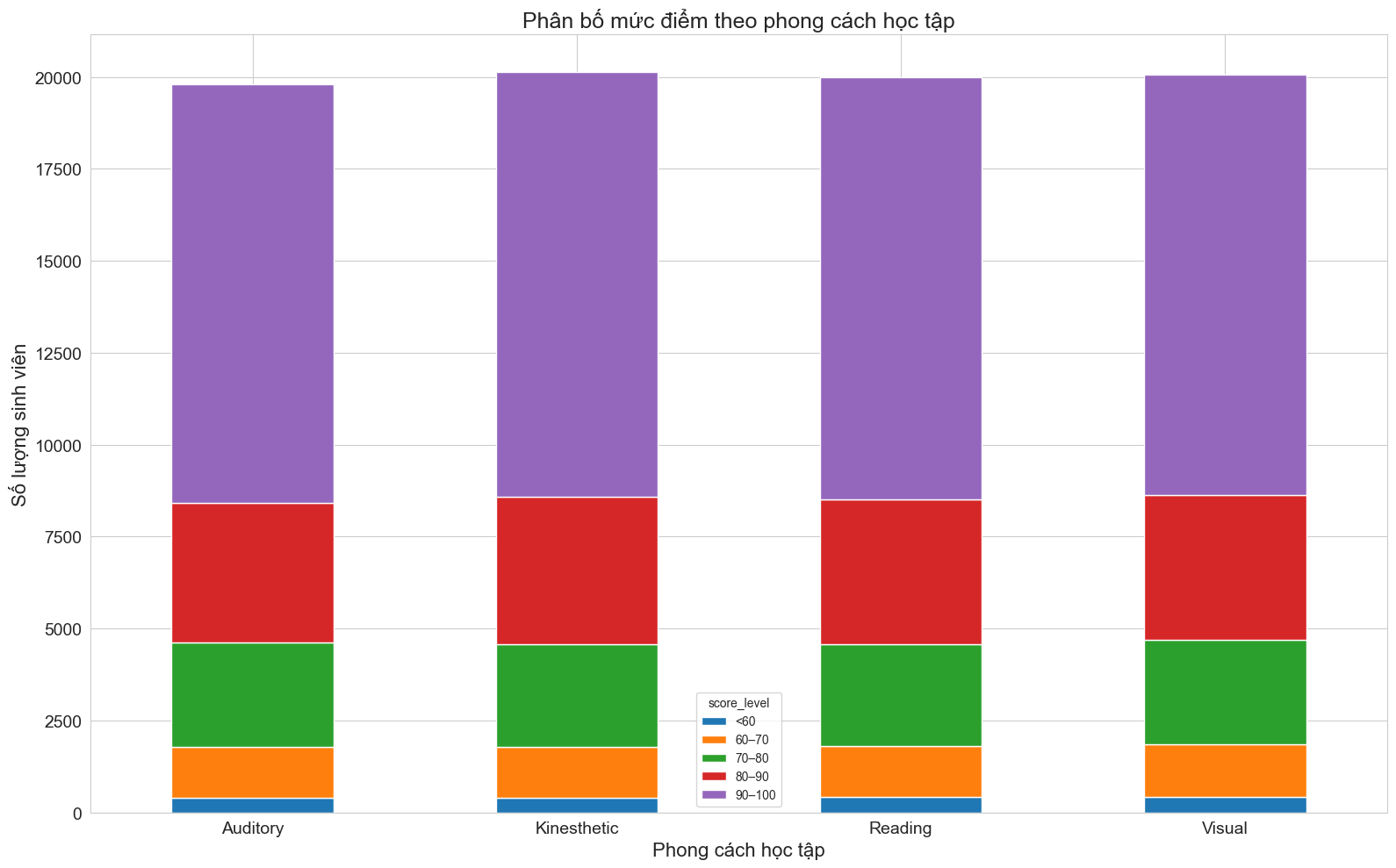
Biểu đồ cho thấy phân bố điểm thi giữa nhóm sinh viên đi làm thêm và không đi làm thêm khá tương đồng. Cả hai nhóm đều tập trung chủ yếu ở mức điểm Khá (60–80) và Giỏi (80–100), trong khi số lượng sinh viên có điểm thấp chiếm tỷ lệ rất nhỏ. Nhóm không đi làm thêm có xu hướng nhỉnh hơn nhẹ ở nhóm điểm Giỏi, cho thấy việc không phải đi làm thêm có thể giúp sinh viên tập trung học tập tốt hơn



Hình 3.14. Biểu đồ trực quan hóa mối quan hệ tình trạng đi làm thêm với điểm thi

### 3.4.10. Phân bố điểm thi theo phong cách học tập

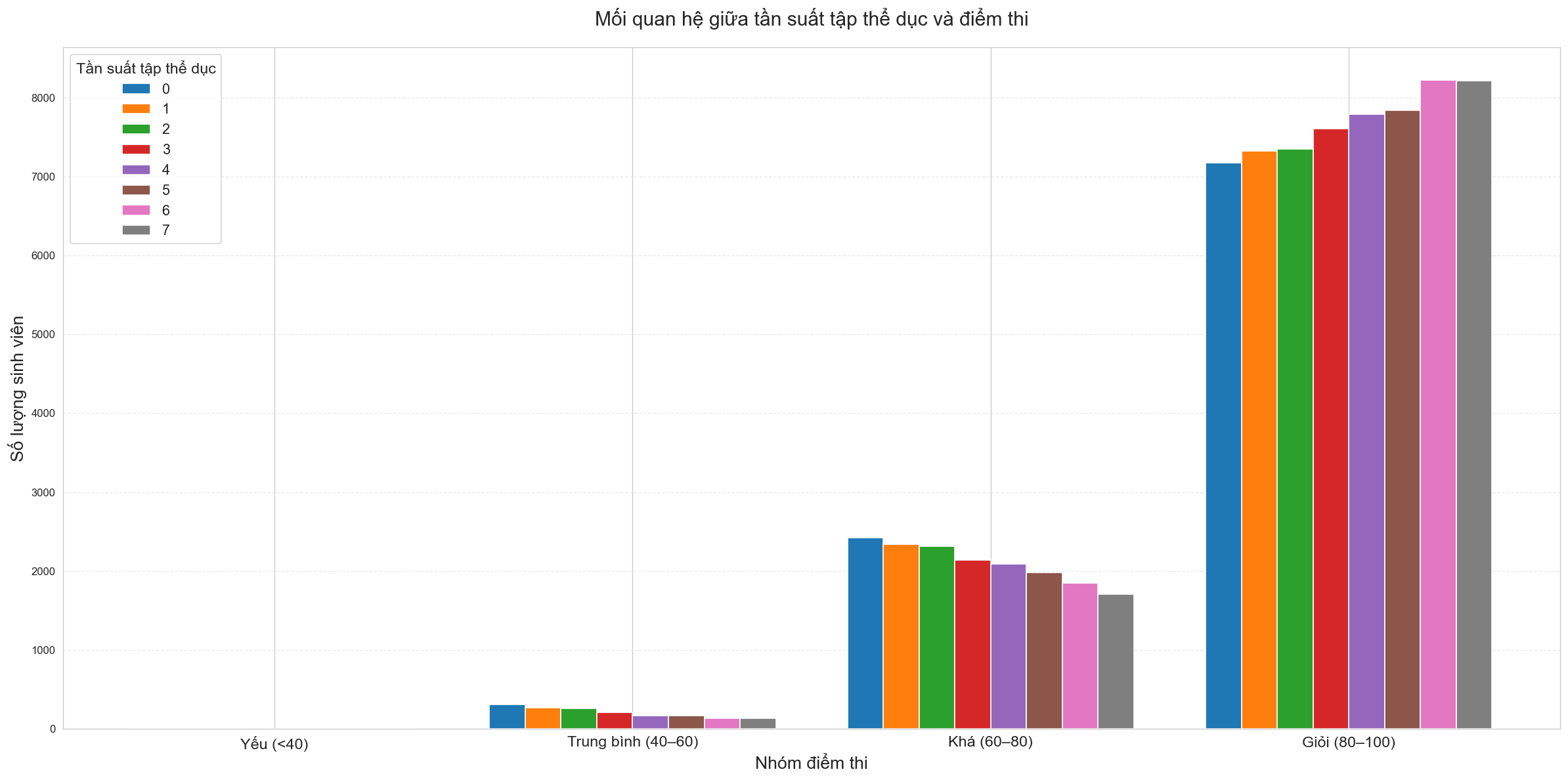
Biểu đồ cho thấy phân bố điểm thi giữa các phong cách học tập khá tương đồng, không có sự chênh lệch quá lớn giữa các nhóm Auditory, Kinesthetic, Reading và Visual. Ở tất cả các phong cách, số lượng sinh viên đạt điểm cao (80–100) chiếm tỷ lệ lớn nhất. Nhóm học theo phong cách Visual và Kinesthetic có xu hướng nhỉnh hơn nhẹ ở mức điểm rất cao (90–100), tuy nhiên sự khác biệt không đáng kể. Điều này cho thấy phong cách học tập không phải là yếu tố quyết định trực tiếp, mà cần được xem xét kết hợp với các yếu tố khác trong mô hình dự đoán GPA.



Hình 3.15. Biểu đồ trực quan hóa mối quan hệ phong cách học tập với điểm thi

### 3.4.11. Mối quan hệ giữa tần suất tập thể dục và điểm thi

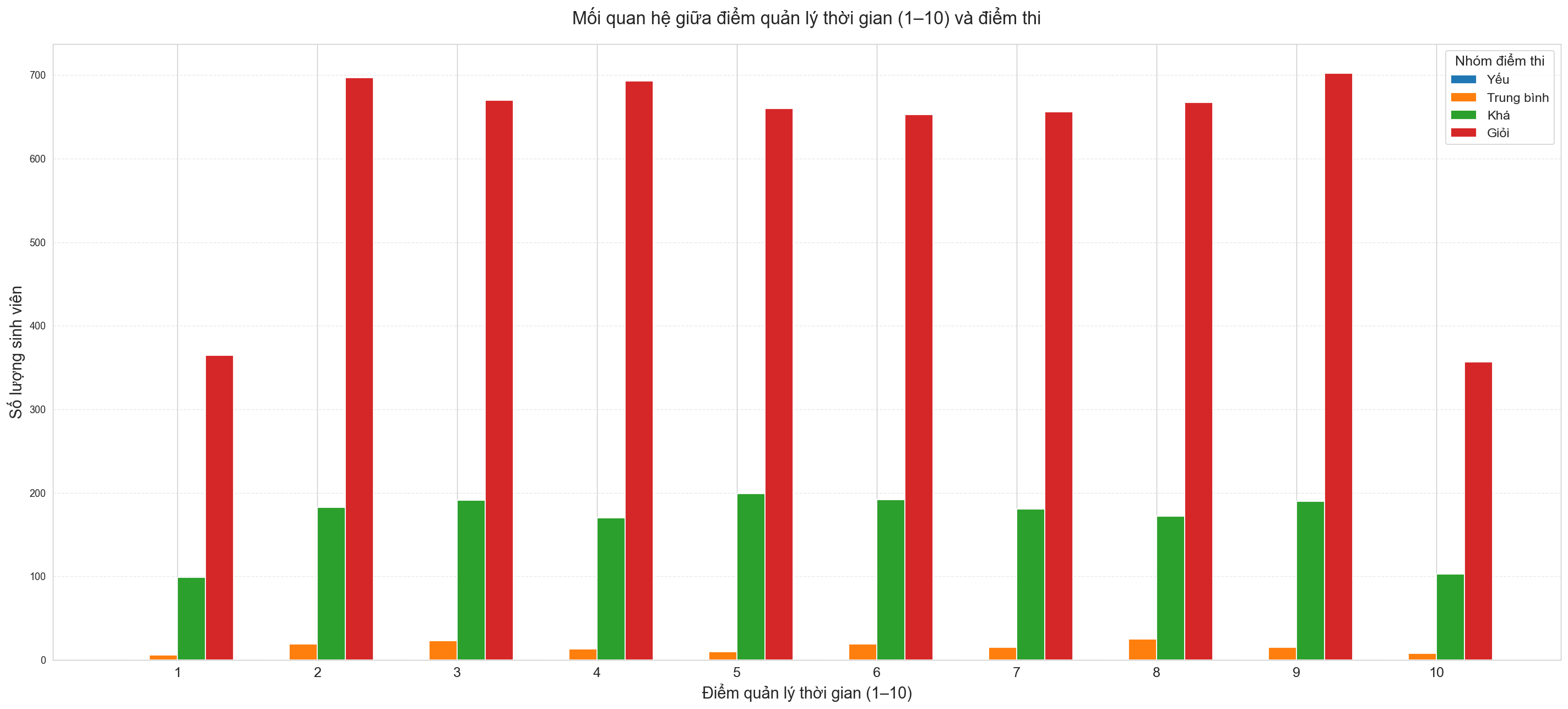
Biểu đồ cho thấy khi tần suất tập thể dục tăng, số lượng sinh viên đạt điểm Giỏi (80–100) có xu hướng tăng dần, trong khi số sinh viên ở nhóm điểm Trung bình giảm nhẹ. Nhóm sinh viên tập thể dục thường xuyên (từ 4–7 lần/tuần) có tỷ lệ đạt điểm cao cao hơn so với nhóm ít hoặc không tập thể dục. Điều này cho thấy việc duy trì thói quen vận động có ảnh hưởng tích cực đến hiệu quả học tập. Tuy nhiên, sự khác biệt giữa các mức tần suất không quá lớn, cho thấy tập thể dục là yếu tố hỗ trợ chứ không phải yếu tố quyết định duy nhất.



Hình 3.16. Biểu đồ trực quan hóa mối quan hệ tần suất tập thể dục với điểm thi

### 3.4.12. Mối quan hệ giữa khả năng quản lý thời gian và điểm thi

Biểu đồ cho thấy khi điểm quản lý thời gian tăng, số lượng sinh viên đạt điểm Giỏi (80–100) tăng rõ rệt, trong khi số sinh viên thuộc nhóm điểm thấp và trung bình chiếm tỷ lệ nhỏ. Những sinh viên có kỹ năng quản lý thời gian tốt (điểm từ 7–9) tập trung nhiều hơn ở nhóm điểm cao. Ở mức điểm quản lý thời gian quá thấp hoặc quá cao, số lượng sinh viên đạt điểm cao giảm nhẹ, cho thấy mối quan hệ không hoàn toàn tuyến tính. Điều này khẳng định quản lý thời gian là một trong những yếu tố quan trọng ảnh hưởng tích cực đến kết quả học tập.



Hình 3.17. Biểu đồ trực quan hóa mối quan hệ mức độ quản lý thời gian với điểm thi

### 3.4.13. Mối quan hệ giữa mức độ hỗ trợ của bố mẹ và điểm thi

Biểu đồ cho thấy tỷ lệ sinh viên đạt điểm Giỏi (80–100) chiếm ưu thế ở tất cả các mức độ hỗ trợ của bố mẹ, và không có sự chênh lệch quá lớn giữa các mức. Khi mức độ hỗ trợ của bố mẹ tăng, số lượng sinh viên đạt điểm Khá và Giỏi có xu hướng ổn định hoặc tăng nhẹ, trong khi nhóm điểm thấp chiếm tỷ lệ rất nhỏ. Điều này cho thấy sự hỗ trợ từ gia đình có ảnh hưởng tích cực nhưng không mang tính quyết định trực tiếp đến kết quả thi. Kết quả học tập vẫn phụ thuộc nhiều hơn vào các yếu tố cá nhân như thói quen học tập và quản lý thời gian.



Hình 3.18. Biểu đồ trực quan hóa mối quan hệ mức độ giúp đỡ của bố mẹ với điểm thi

### 3.3.14. Tổng kết phân tích và khám phá dữ liệu (EDA)

Kết quả EDA cho thấy điểm thi và GPA của sinh viên chịu ảnh hưởng bởi nhiều yếu tố khác nhau, trong đó các yếu tố học tập như điểm thi, GPA học kỳ trước, số giờ học mỗi ngày và khả năng quản lý thời gian có mối quan hệ tích cực và rõ rệt với kết quả học tập. Các yếu tố về thói quen và lối sống như tần suất tập thể dục có tác động hỗ trợ, trong khi áp lực học tập cao và thời gian sử dụng mạng xã hội nhiều có xu hướng ảnh hưởng tiêu cực đến điểm thi. Một số yếu tố xã hội như học gia sư và sự hỗ trợ từ gia đình cho thấy ảnh hưởng tích cực nhưng không mang tính quyết định. Ngoài ra, phân phối dữ liệu và các mối quan hệ giữa biến đầu vào và đầu ra không hoàn toàn tuyến tính, thể hiện qua độ phân tán và sự chồng lấn giữa các nhóm. Điều này cho thấy việc sử dụng các mô hình học máy phi tuyến như Random Forest, Gradient Boosting và XGBoost là phù hợp cho bài toán dự đoán GPA. EDA đã cung cấp cơ sở thực tiễn quan trọng cho việc lựa chọn mô hình và xây dựng quy trình huấn luyện ở các bước tiếp theo.

## 3.5. Xây dựng mô hình dự đoán

### 3.5.1. Mục tiêu xây dựng mô hình

Mục tiêu của bước xây dựng mô hình là thiết lập các mô hình học máy có khả năng dự đoán điểm trung bình học tập (GPA) của sinh viên dựa trên các đặc trưng đã được tiền xử lý và phân tích ở các bước trước. Các mô hình được lựa chọn nhằm so sánh hiệu quả giữa phương pháp tuyến tính và các phương pháp phi tuyến, từ đó tìm ra mô hình phù hợp nhất cho bài toán.

### 3.5.2. Xác định dữ liệu đầu vào và đầu ra

Tập đặc trưng đầu vào (X): Bao gồm các thuộc tính liên quan đến học tập, thói quen sinh hoạt, yếu tố tâm lý và môi trường học tập của sinh viên sau khi đã được xử lý (mã hóa, chuẩn hóa).

Biến mục tiêu (y): Điểm trung bình học tập GPA – là biến liên tục cần dự đoán.

### 3.5.3. Lí do lựa chọn các mô hình học máy

Dựa trên kết quả phân tích và khám phá dữ liệu (EDA), có thể nhận thấy mối quan hệ giữa các đặc trưng đầu vào và GPA không hoàn toàn tuyến tính, đồng thời dữ liệu chịu ảnh hưởng bởi nhiều yếu tố khác nhau như học tập, thói quen sinh hoạt và yếu tố tâm lý. Do đó, đề tài lựa chọn kết hợp cả mô hình tuyến tính và các mô hình phi tuyến nhằm đánh giá toàn diện hiệu quả dự đoán.

Linear Regression được sử dụng làm mô hình cơ sở (baseline) do tính đơn giản và khả năng diễn giải cao. Mô hình này giúp đánh giá mức độ phù hợp của giả định tuyến tính giữa các đặc trưng và GPA, đồng thời làm cơ sở so sánh với các mô hình phức tạp hơn. Nếu Linear Regression đạt kết quả tốt, điều này cho thấy mối quan hệ giữa dữ liệu và GPA có tính tuyến tính tương đối rõ ràng.

Random Forest Regression được lựa chọn nhằm khắc phục hạn chế của mô hình tuyến tính trong việc mô hình hóa các quan hệ phi tuyến. Nhờ cơ chế ensemble dựa trên nhiều cây quyết định được huấn luyện song song, Random Forest có khả năng giảm overfitting, xử lý tốt dữ liệu có nhiễu và khai thác hiệu quả các mối quan hệ phức tạp trong dữ liệu GPA.

Gradient Boosting Regression được sử dụng để tiếp cận bài toán theo hướng boosting, trong đó các mô hình yếu được huấn luyện nối tiếp nhằm dần dần sửa lỗi của mô hình trước. Mô hình này phù hợp với dữ liệu có cấu trúc phức tạp và cho phép cải thiện độ chính xác dự đoán thông qua việc học từ phần sai số (residual).

XGBoost Regression được lựa chọn như một mô hình nâng cao, kế thừa ưu điểm của Gradient Boosting và được bổ sung thêm các cơ chế regularization và tối ưu hóa hiệu quả. XGBoost có khả năng kiểm soát overfitting tốt hơn, hội tụ nhanh và thường cho hiệu suất cao trên dữ liệu dạng bảng (tabular data) như bài toán dự đoán GPA trong đề tài.

Việc sử dụng đồng thời bốn mô hình trên cho phép so sánh giữa các phương pháp từ đơn giản đến phức tạp, từ tuyến tính đến phi tuyến, qua đó đánh giá một cách toàn diện hiệu quả dự đoán GPA và lựa chọn mô hình phù hợp nhất cho bài toán.

### 3.5.4. Chiến lược huấn luyện mô hình

Dữ liệu sau tiền xử lý được chia thành:

-Tập huấn luyện (Training set): dùng để huấn luyện các mô hình.

-Tập kiểm tra (Test set): dùng để đánh giá hiệu suất mô hình trên dữ liệu chưa từng thấy.

Đối với các mô hình nhạy cảm với thang đo dữ liệu như Linear Regression và Gradient Descent, dữ liệu đầu vào được chuẩn hóa trước khi huấn luyện. Các mô hình dựa trên cây như Random Forest và XGBoost không yêu cầu chuẩn hóa nhưng vẫn sử dụng cùng tập dữ liệu đã xử lý để đảm bảo tính nhất quán.

### 3.5.5. Tiêu chí đánh giá mô hình

Để đánh giá hiệu quả dự đoán của các mô hình hồi quy, đề tài sử dụng các chỉ số đánh giá phổ biến trong học máy, bao gồm:

-MAE (Mean Absolute Error)  
-MSE (Mean Squared Error)

-RMSE (Root Mean Squared Error)  
-R² (Hệ số xác định)

Việc sử dụng đồng thời các chỉ số MAE, MSE, RMSE và R² cho phép đánh giá toàn diện hiệu quả mô hình, bao gồm cả mức độ sai lệch dự đoán và khả năng tổng quát hóa của mô hình trên dữ liệu kiểm tra.

## 3.6. Huấn luyện và đánh giá mô hình

### 3.6.1. Quy trình huấn luyện mô hình

Sau khi hoàn tất các bước tiền xử lý và xây dựng mô hình, dữ liệu được sử dụng để huấn luyện các mô hình học máy đã lựa chọn. Tập dữ liệu được chia thành tập huấn luyện (training set) và tập kiểm tra (test set) theo tỷ lệ phù hợp nhằm đảm bảo khả năng tổng quát hóa của mô hình.

Các mô hình được huấn luyện trên cùng một tập dữ liệu huấn luyện để đảm bảo tính công bằng trong quá trình so sánh. Đối với các mô hình nhạy cảm với thang đo dữ liệu như Linear Regression và Gradient Descent, dữ liệu đầu vào đã được chuẩn hóa trước khi huấn luyện. Trong khi đó, các mô hình dựa trên cây như Random Forest, Gradient Boosting và XGBoost không yêu cầu chuẩn hóa nhưng vẫn sử dụng cùng tập dữ liệu đã được tiền xử lý để đảm bảo tính nhất quán.

### 3.6.2. Chuẩn bị dữ liệu

Xác định biến đầu vào và biến mục tiêu

Biến mục tiêu (y):

Điểm thi – đại diện cho kết quả học tập của sinh viên.

Biến đầu vào (X):

Bao gồm tất cả các đặc trưng còn lại như: 'Giờ học mỗi ngày', 'Giờ sử dụng mạng xã hội', 'Làm thêm', 'Tỷ lệ điểm danh (%)', 'Giờ ngủ', 'Chất lượng chế độ ăn','Tần suất tập thể dục', 'Đánh giá sức khỏe tinh thần', 'GPA học kỳ trước', 'Mức độ căng thẳng', 'Nguy cơ bỏ học', 'Thời gian dùng màn hình', 'Môi trường học tập', 'Tiếp cận lớp phụ đạo / gia sư', 'Mức độ hỗ trợ của cha mẹ', 'Mức độ động lực', 'Mức độ lo âu khi thi', 'Phong cách học tập', 'Điểm quản lý thời gian'

### 3.6.3. Phân loại các thuộc tính số và thuộc tính phân loại

**numeric\_cols = X.select\_dtypes(include=["int64", "float64"]).columns**

**categorical\_cols = X.select\_dtypes(include=["object"]).columns**

Trong quá trình chuẩn bị dữ liệu cho mô hình học máy, các thuộc tính đầu vào được phân chia thành hai nhóm chính là thuộc tính số (numerical features) và thuộc tính phân loại (categorical features). Việc phân loại này nhằm áp dụng các phương pháp tiền xử lý phù hợp cho từng loại dữ liệu.

Các thuộc tính số được xác định bằng cách lựa chọn các cột có kiểu dữ liệu số nguyên và số thực (int64, float64). Nhóm thuộc tính này bao gồm các biến phản ánh trực tiếp giá trị định lượng như thời gian học tập, điểm thi, GPA học kỳ trước, số giờ ngủ hoặc các thang điểm đánh giá. Các thuộc tính số thường được sử dụng trực tiếp trong mô hình và có thể được chuẩn hóa để đảm bảo các thuật toán nhạy cảm với thang đo hoạt động hiệu quả.

Trong khi đó, các thuộc tính phân loại được xác định thông qua kiểu dữ liệu chuỗi (object). Nhóm này bao gồm các biến mô tả đặc điểm hoặc trạng thái của sinh viên như giới tính, ngành học, phong cách học tập, mức độ căng thẳng hoặc điều kiện học tập. Do các thuật toán học máy không thể xử lý trực tiếp dữ liệu dạng chuỗi, các thuộc tính phân loại cần được mã hóa (ví dụ: One-Hot Encoding hoặc Label Encoding) trước khi đưa vào huấn luyện mô hình.

Việc tách biệt rõ ràng hai nhóm thuộc tính giúp xây dựng quy trình tiền xử lý hợp lý, đảm bảo dữ liệu đầu vào phù hợp với yêu cầu của từng mô hình và góp phần nâng cao độ chính xác cũng như khả năng tổng quát hóa của mô hình dự đoán GPA.

### 3.6.4. Chia tập dữ liệu huấn luyện và kiểm tra

**X\_train, X\_test, y\_train, y\_test = train\_test\_split(**

**X, y, test\_size=0.2, random\_state=42)**

Sau khi hoàn tất quá trình tiền xử lý và xác định tập đặc trưng đầu vào cùng biến mục tiêu , dữ liệu được chia thành tập huấn luyện và tập kiểm tra bằng phương pháp train\_test\_split.

Trong đó:

-80% dữ liệu được sử dụng làm tập huấn luyện (X\_train, y\_train) để huấn luyện mô hình.

-20% dữ liệu còn lại được sử dụng làm tập kiểm tra (X\_test, y\_test) nhằm đánh giá hiệu suất mô hình trên dữ liệu chưa từng thấy.

Việc thiết lập tham số random\_state = 42 giúp đảm bảo quá trình chia dữ liệu có thể tái lập, từ đó tạo điều kiện thuận lợi cho việc so sánh kết quả giữa các mô hình khác nhau. Cách chia dữ liệu này giúp đánh giá khách quan khả năng tổng quát hóa (generalization) của mô hình và hạn chế hiện tượng overfitting.

### 3.6.5. Xây dựng quy trình tiền xử lý dữ liệu bằng ColumnTransformer

**preprocess = ColumnTransformer(**

**transformers=[**

**("num", StandardScaler(), numeric\_cols),**

**("cat", OneHotEncoder(handle\_unknown="ignore"), categorical\_cols)**

**]**

Để đảm bảo các thuộc tính đầu vào được xử lý phù hợp với từng loại dữ liệu, đề tài sử dụng ColumnTransformer nhằm xây dựng một quy trình tiền xử lý thống nhất cho cả dữ liệu số và dữ liệu phân loại.

Cụ thể, quy trình tiền xử lý bao gồm hai bước chính:

-Đối với các thuộc tính số (numeric features): Phương pháp StandardScaler được áp dụng để chuẩn hóa dữ liệu về phân phối có trung bình bằng 0 và độ lệch chuẩn bằng 1. Việc chuẩn hóa giúp các mô hình nhạy cảm với thang đo như Linear Regression và các thuật toán tối ưu dựa trên Gradient Descent hội tụ nhanh hơn và ổn định hơn.

-Đối với các thuộc tính phân loại (categorical features): Phương pháp One-Hot Encoding được sử dụng để chuyển các giá trị phân loại sang dạng vector nhị phân. Tham số handle\_unknown="ignore" giúp mô hình xử lý an toàn các giá trị phân loại mới xuất hiện trong tập kiểm tra mà không gây lỗi trong quá trình dự đoán.

Việc kết hợp hai phương pháp tiền xử lý này trong ColumnTransformer giúp đảm bảo rằng mỗi loại thuộc tính được xử lý đúng cách trong cùng một pipeline, tránh rò rỉ dữ liệu (data leakage) và đảm bảo tính nhất quán giữa quá trình huấn luyện và đánh giá mô hình.

### 3.6.6. Áp dụng tiền xử lý cho tập huấn luyện và tập kiểm tra

**X\_train\_processed = preprocess.fit\_transform(X\_train)**

**X\_test\_processed  = preprocess.transform(X\_test)**

Sau khi xây dựng quy trình tiền xử lý bằng ColumnTransformer, bước tiếp theo là áp dụng quy trình này cho dữ liệu huấn luyện và dữ liệu kiểm tra.

Quy trình tiền xử lý được huấn luyện (fit) trên tập huấn luyện thông qua phương thức fit\_transform. Ở bước này, các tham số cần thiết như trung bình và độ lệch chuẩn của các thuộc tính số (đối với StandardScaler) cũng như các danh mục xuất hiện trong các thuộc tính phân loại (đối với One-Hot Encoder) được học từ dữ liệu huấn luyện.

Sau đó, cùng quy trình tiền xử lý này được áp dụng (transform) cho tập kiểm tra mà không huấn luyện lại, nhằm đảm bảo rằng tập kiểm tra được xử lý theo đúng các tham số đã học từ tập huấn luyện. Cách làm này giúp tránh rò rỉ dữ liệu (data leakage) và đảm bảo tính khách quan trong quá trình đánh giá mô hình.

Kết quả của bước này là hai tập dữ liệu:

-X\_train\_processed: dữ liệu huấn luyện đã được chuẩn hóa và mã hóa, dùng để huấn luyện mô hình.

-X\_test\_processed: dữ liệu kiểm tra đã được xử lý tương tự, dùng để đánh giá mô hình.

### 3.6.7. Chuyển đổi dữ liệu dạng sparse sang dense

**import scipy.sparse**

**if scipy.sparse.issparse(X\_train\_processed):**

**X\_train\_processed = X\_train\_processed.toarray()**

**X\_test\_processed  = X\_test\_processed.toarray()**

Sau khi áp dụng các bước tiền xử lý như StandardScaler và One-Hot Encoding, dữ liệu đầu ra từ ColumnTransformer có thể ở dạng ma trận thưa (sparse matrix). Điều này thường xảy ra khi sử dụng One-Hot Encoding cho các thuộc tính phân loại có nhiều giá trị khác nhau.

Do một số mô hình và phép toán trong quá trình huấn luyện không hỗ trợ trực tiếp dữ liệu dạng sparse, dữ liệu cần được kiểm tra và chuyển đổi sang dạng ma trận đầy đủ (dense array) nếu cần thiết.

Đoạn mã thực hiện:

-Kiểm tra xem dữ liệu sau tiền xử lý có phải là sparse matrix hay không.

-Nếu đúng, dữ liệu được chuyển sang dạng numpy array bằng phương thức toarray().

Sau bước này:

-X\_train\_processed và X\_test\_processed đều ở dạng mảng số đầy đủ.

Dữ liệu sẵn sàng để đưa vào huấn luyện và đánh giá các mô hình học máy.

### 3.6.8. Tái tạo tập dữ liệu sau khi mã hóa và chuẩn hóa đặc trưng

**num\_features = list(numeric\_cols)**

**cat\_features = preprocess.named\_transformers\_["cat"] \**

**.get\_feature\_names\_out(categorical\_cols)**

**all\_features = num\_features + list(cat\_features)**

**df\_train\_encoded = pd.DataFrame(**

**X\_train\_processed,**

**columns=all\_features,**

**index=X\_train.index**

**)**

**df\_train\_encoded.head()**

Sau khi hoàn tất quá trình tiền xử lý và chuyển dữ liệu về dạng ma trận số, các đặc trưng đầu vào được tái tạo dưới dạng một DataFrame có nhãn cột rõ ràng nhằm thuận tiện cho việc kiểm tra, phân tích và diễn giải mô hình.

Trước hết, danh sách các đặc trưng số được lấy trực tiếp từ tập numeric\_cols. Đối với các đặc trưng phân loại, tên các đặc trưng mới sau khi áp dụng One-Hot Encoding được trích xuất thông qua phương thức get\_feature\_names\_out. Các tên này thể hiện đầy đủ thông tin về thuộc tính gốc và từng giá trị phân loại tương ứng.

Sau đó, toàn bộ các đặc trưng số và đặc trưng phân loại đã được mã hóa được kết hợp lại thành một danh sách đặc trưng đầy đủ. Danh sách này phản ánh chính xác cấu trúc dữ liệu đầu vào mà mô hình học máy sử dụng sau tiền xử lý.

Tập dữ liệu huấn luyện đã được mã hóa và chuẩn hóa được chuyển thành một DataFrame Pandas, trong đó:

-Mỗi cột tương ứng với một đặc trưng sau tiền xử lý.

-Mỗi dòng tương ứng với một mẫu dữ liệu trong tập huấn luyện.

-Chỉ số (index) được giữ nguyên để đảm bảo sự nhất quán với dữ liệu gốc.

Việc tái tạo DataFrame từ dữ liệu đã xử lý giúp dễ dàng kiểm tra dữ liệu, phân tích mức độ đóng góp của từng đặc trưng, cũng như hỗ trợ việc diễn giải kết quả mô hình ở các bước tiếp theo.

### 3.6.9. Huấn luyện và đánh giá mô hình Linear Regression

Để thiết lập mô hình cơ sở (baseline) cho bài toán dự đoán GPA, đề tài sử dụng Linear Regression – một mô hình hồi quy tuyến tính phổ biến và dễ diễn giải.

Cấu hình mô hình

Mô hình Linear Regression được sử dụng với cấu hình mặc định, không bao gồm các tham số điều chỉnh phức tạp. Việc sử dụng mô hình đơn giản này giúp đánh giá khả năng dự đoán GPA dựa trên giả định mối quan hệ tuyến tính giữa các đặc trưng đầu vào và biến mục tiêu.

**Quá trình huấn luyện và dự đoán**

Mô hình được huấn luyện trên tập dữ liệu đã tiền xử lý (X\_train\_processed, y\_train). Trước đó, các đặc trưng số đã được chuẩn hóa và các đặc trưng phân loại đã được mã hóa, đảm bảo dữ liệu đầu vào phù hợp với mô hình tuyến tính.

Sau khi huấn luyện, mô hình được sử dụng để dự đoán GPA trên tập kiểm tra (X\_test\_processed). Các giá trị dự đoán thu được được so sánh với giá trị thực tế nhằm đánh giá hiệu quả mô hình.

Đánh giá hiệu suất mô hình

Hiệu suất của mô hình Linear Regression được đánh giá thông qua các chỉ số hồi quy:

MAE (Mean Absolute Error): 3.1963517382350384

MSE (Mean Squared Error): 17.53946028778897

RMSE (Root Mean Squared Error): 4.188013883428393

R² (Hệ số xác định): 0.8704507853290462

Các chỉ số này giúp đánh giá mức độ phù hợp của giả định tuyến tính trong bài toán dự đoán GPA.

### 3.6.10. Huấn luyện và đánh giá mô hình Random Forest Regression

Sau khi hoàn tất quá trình tiền xử lý dữ liệu, mô hình Random Forest Regression được huấn luyện với các siêu tham số đã được tinh chỉnh nhằm kiểm soát độ phức tạp của mô hình và cải thiện khả năng tổng quát hóa.

Cấu hình mô hình

Mô hình Random Forest được khởi tạo với các tham số chính như sau:

-n\_estimators = 300: sử dụng 300 cây quyết định để đảm bảo tính ổn định và giảm phương sai của mô hình.

-max\_depth = 15: giới hạn độ sâu tối đa của mỗi cây nhằm tránh việc mô hình học quá chi tiết vào dữ liệu huấn luyện.

-min\_samples\_split = 10: yêu cầu tối thiểu 10 mẫu dữ liệu tại một nút để tiếp tục tách, giúp giảm nguy cơ overfitting.

-min\_samples\_leaf = 5: mỗi nút lá phải chứa ít nhất 5 mẫu, đảm bảo các dự đoán tại lá có tính đại diện cao hơn.

-random\_state = 42: đảm bảo khả năng tái lập kết quả.

-n\_jobs = -1: tận dụng toàn bộ tài nguyên CPU để tăng tốc quá trình huấn luyện.

Cấu hình này giúp cân bằng giữa khả năng học các mối quan hệ phức tạp và việc kiểm soát độ phức tạp của mô hình.

**Quá trình huấn luyện mô hình**

Mô hình được huấn luyện trên tập dữ liệu đã được tiền xử lý (X\_train\_processed, y\_train). Trong quá trình huấn luyện, mỗi cây quyết định được xây dựng trên một tập dữ liệu con khác nhau và tại mỗi nút chỉ xét một tập con ngẫu nhiên các đặc trưng, từ đó làm tăng tính đa dạng của mô hình và hạn chế hiện tượng overfitting.

**Dự đoán và đánh giá mô hình**

Sau khi huấn luyện, mô hình được sử dụng để dự đoán điểm GPA trên tập kiểm tra (X\_test\_processed). Kết quả dự đoán được đánh giá thông qua các chỉ số hồi quy:

MAE (Mean Absolute Error): 3.2394352334515184MSE

(Mean Squared Error): 17.487895829548048

RMSE (Root Mean Squared Error): 4.1818531573392255

R² (Hệ số xác định): 0.870831648534665.

Các chỉ số này cho phép đánh giá toàn diện hiệu quả dự đoán cũng như khả năng tổng quát hóa của mô hình Random Forest.

### 3.6.11. Huấn luyện và đánh giá mô hình XGBoost Regression

Sau khi huấn luyện Random Forest, đề tài tiếp tục sử dụng XGBoost Regression – một mô hình boosting nâng cao – nhằm cải thiện độ chính xác dự đoán GPA và kiểm soát tốt hơn hiện tượng overfitting.

Cấu hình mô hình XGBoost

Mô hình XGBoost được khởi tạo với các siêu tham số như sau:

-n\_estimators = 300: Số lượng cây quyết định được xây dựng trong mô hình. Số cây đủ lớn giúp mô hình học ổn định và giảm sai số.

-learning\_rate = 0.05: Hệ số học nhỏ giúp mỗi cây mới chỉ đóng góp một phần nhỏ vào mô hình tổng thể, từ đó giảm nguy cơ overfitting và tăng khả năng tổng quát hóa.

-max\_depth = 6: Giới hạn độ sâu của mỗi cây quyết định, giúp mô hình không học quá chi tiết vào dữ liệu huấn luyện.

-subsample = 0.8: Mỗi cây được huấn luyện trên 80% dữ liệu, giúp tăng tính ngẫu nhiên và giảm overfitting.

-colsample\_bytree = 0.8: Mỗi cây chỉ sử dụng 80% số đặc trưng, giúp mô hình đa dạng hơn và học tốt hơn các mối quan hệ phức tạp.

-objective = "reg:squarederror": Xác định đây là bài toán hồi quy với hàm mất mát là sai số bình phương.

-random\_state = 42: Đảm bảo khả năng tái lập kết quả.

-n\_jobs = -1: Tận dụng toàn bộ tài nguyên CPU để tăng tốc quá trình huấn luyện.

Cấu hình này giúp XGBoost đạt được sự cân bằng giữa độ chính xác và khả năng tổng quát hóa.

**Quá trình huấn luyện và dự đoán**

Mô hình XGBoost được huấn luyện trên tập dữ liệu đã tiền xử lý (X\_train\_processed, y\_train). Trong quá trình huấn luyện, các cây quyết định được xây dựng theo cơ chế boosting, trong đó mỗi cây mới tập trung học từ phần sai số (residual) của mô hình trước đó.

Sau khi huấn luyện, mô hình được sử dụng để dự đoán GPA trên tập kiểm tra (X\_test\_processed). Các giá trị dự đoán được so sánh với giá trị thực tế để đánh giá hiệu suất mô hình.

**Đánh giá hiệu suất mô hình**

Hiệu suất của mô hình XGBoost được đánh giá thông qua các chỉ số hồi quy:

MAE (Mean Absolute Error): 3.24058198928833

MSE (Mean Squared Error): 17.500734329223633

RMSE (Root Mean Squared Error): 4.183387900879338

R² (Hệ số xác định): 0.87073683738708

Các chỉ số này cho phép đánh giá toàn diện độ chính xác và khả năng tổng quát hóa của mô hình.

### 3.6.12. Huấn luyện và đánh giá mô hình Gradient Boosting Regression

Sau khi xây dựng và đánh giá các mô hình khác, đề tài tiếp tục sử dụng Gradient Boosting Regression với các siêu tham số đã được tinh chỉnh nhằm cải thiện độ chính xác dự đoán và khả năng tổng quát hóa.

Cấu hình mô hình

Mô hình Gradient Boosting được khởi tạo với các tham số như sau:

-n\_estimators = 300: Số lượng cây quyết định được huấn luyện nối tiếp. Số cây đủ lớn giúp mô hình học ổn định và giảm sai số.

-learning\_rate = 0.05: Hệ số học nhỏ giúp mỗi cây mới chỉ điều chỉnh mô hình tổng thể một lượng nhỏ, từ đó hạn chế overfitting.

max\_depth = 4: Giới hạn độ sâu của mỗi cây quyết định. Việc sử dụng cây nông giúp mô hình tập trung học các mối quan hệ quan trọng và tránh học nhiễu.

subsample = 0.8: Mỗi cây chỉ sử dụng 80% dữ liệu huấn luyện, giúp tăng tính ngẫu nhiên và cải thiện khả năng tổng quát hóa.

-min\_samples\_leaf = 5: Quy định số mẫu tối thiểu tại mỗi nút lá, đảm bảo mỗi quyết định được đưa ra dựa trên đủ dữ liệu đại diện.

-random\_state = 42: Đảm bảo khả năng tái lập kết quả huấn luyện.

Cấu hình này giúp Gradient Boosting đạt được sự cân bằng giữa độ chính xác và độ ổn định.

**Quá trình huấn luyện và dự đoán**

Mô hình được huấn luyện trên tập dữ liệu đã tiền xử lý (X\_train\_processed, y\_train). Trong quá trình huấn luyện, các cây quyết định được xây dựng theo cơ chế boosting, trong đó mỗi cây mới tập trung học từ phần sai số (residual) của mô hình trước.

Sau khi hoàn tất huấn luyện, mô hình được sử dụng để dự đoán GPA của sinh viên trên tập kiểm tra (X\_test\_processed). Các giá trị dự đoán được so sánh với giá trị thực tế để đánh giá hiệu quả mô hình**.**

**Đánh giá hiệu suất mô hình**

Hiệu suất của mô hình được đánh giá thông qua các chỉ số hồi quy:

-MAE (Mean Absolute Error): 3.2239551090962184

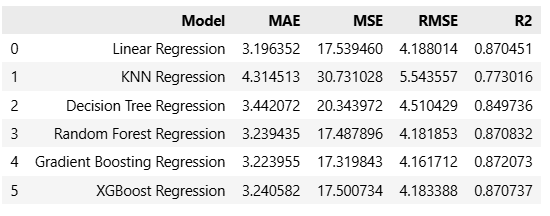
-MSE (Mean Squared Error): 17.319843066818894

-RMSE (Root Mean Squared Error): 4.161711554975776

-R² (Hệ số xác định): 0.8720729126943163

Các chỉ số này cho phép đánh giá toàn diện cả độ chính xác dự đoán và khả năng tổng quát hóa của mô hình.

### 3.6.13. So sánh các kết quả của mô hình

****

Hình 3.19. Thống kê các Metric đánh giá của mô hình

Dựa trên bảng kết quả, có thể nhận thấy các mô hình ensemble và boosting nhìn chung cho hiệu suất dự đoán tốt hơn so với các mô hình đơn giản.

-XGBoost Regression và Gradient Boosting Regression đạt hiệu suất tốt nhất, thể hiện qua MAE thấp, RMSE nhỏ và R² cao (~0.872). Điều này cho thấy các mô hình boosting có khả năng khai thác tốt các mối quan hệ phi tuyến trong dữ liệu GPA.

-Random Forest Regression cũng cho kết quả rất cạnh tranh, với các chỉ số gần tương đương XGBoost và Gradient Boosting. Mô hình này thể hiện sự ổn định và khả năng tổng quát hóa tốt nhờ cơ chế ensemble.

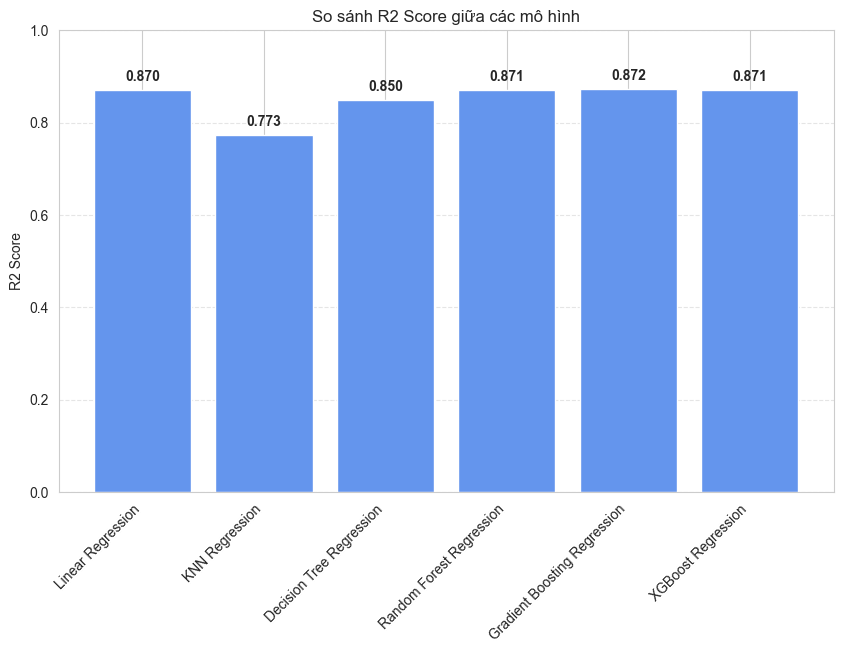
-Linear Regression đạt kết quả khá tốt (R² ≈ 0.87), cho thấy dữ liệu GPA vẫn tồn tại mối quan hệ tuyến tính tương đối rõ ràng. Tuy nhiên, hiệu suất của mô hình này vẫn thấp hơn nhẹ so với các mô hình phi tuyến.

-Decision Tree Regression cho kết quả kém hơn các mô hình ensemble do dễ bị overfitting khi chỉ sử dụng một cây quyết định duy nhất.

-KNN Regression có hiệu suất thấp nhất trong các mô hình được thử nghiệm, thể hiện qua MAE và RMSE cao nhất. Điều này cho thấy KNN không phù hợp với dữ liệu có số chiều lớn và nhiều đặc trưng đã được mã hóa.

Nhìn chung, kết quả thực nghiệm khẳng định rằng việc sử dụng các mô hình phi tuyến và ensemble là phù hợp cho bài toán dự đoán GPA, đặc biệt khi dữ liệu chịu ảnh hưởng đồng thời của nhiều yếu tố học tập và hành vi.

### 3.6.14. Trực quan hóa tỉ lệ chính xác của các mô hình

****

Hình 3.20. Trực quan hóa tỉ lệ chính xác của các mô hình

Biểu đồ thể hiện sự khác biệt về R² Score giữa các mô hình hồi quy được sử dụng trong bài toán dự đoán GPA. Có thể thấy rằng các mô hình ensemble và boosting đạt giá trị R² cao hơn so với các mô hình đơn giản.

-Gradient Boosting Regression đạt R² cao nhất (≈ 0.872), cho thấy khả năng giải thích biến thiên của GPA tốt nhất trong số các mô hình được thử nghiệm.

-Random Forest Regression và XGBoost Regression có R² rất gần nhau (≈ 0.871), thể hiện hiệu suất ổn định và khả năng tổng quát hóa tốt.

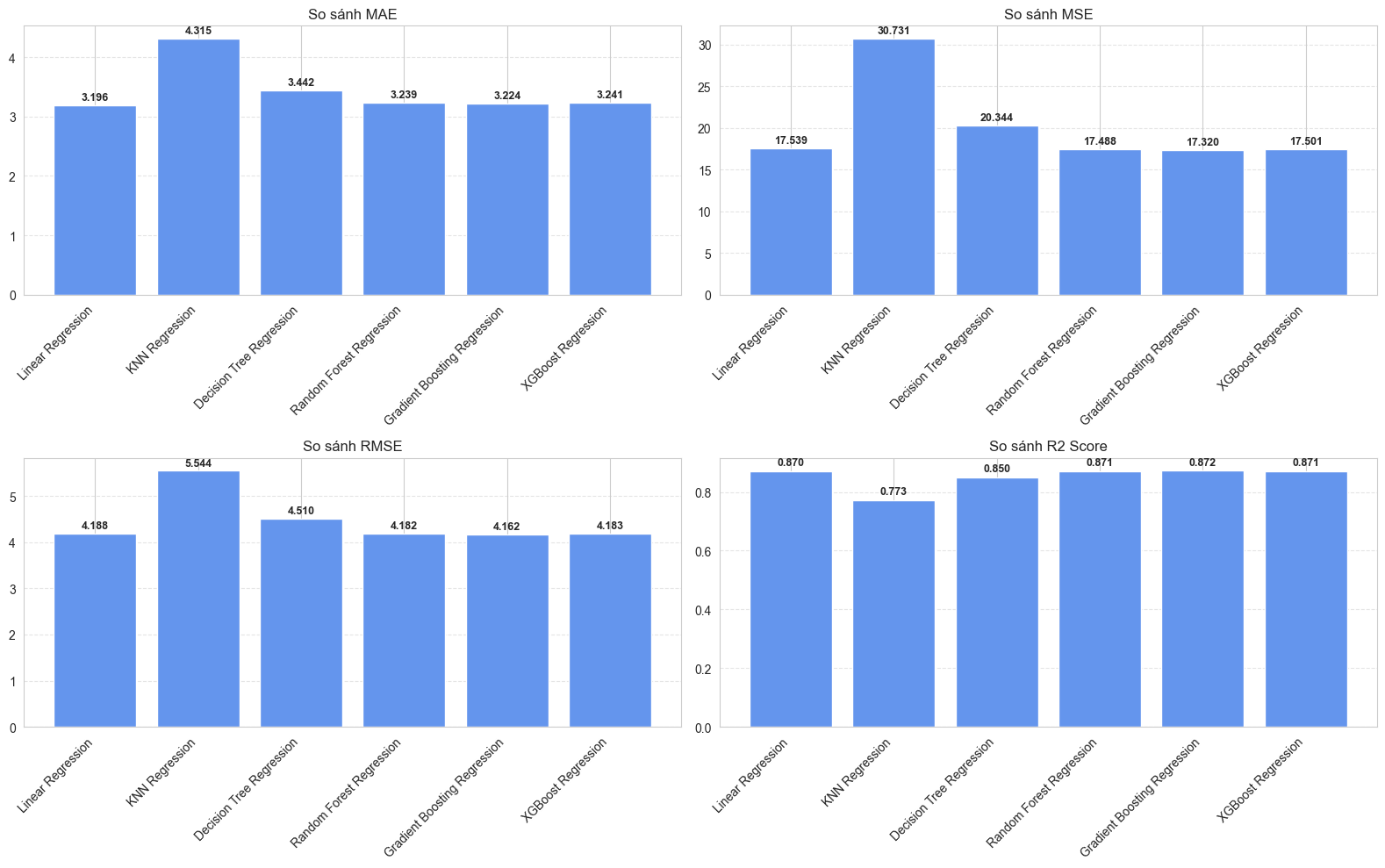
-Linear Regression đạt R² ≈ 0.870, cho thấy dữ liệu vẫn có mối quan hệ tuyến tính tương đối rõ ràng, tuy nhiên mô hình tuyến tính không khai thác được các mối quan hệ phi tuyến nhỏ trong dữ liệu.

-Decision Tree Regression cho kết quả thấp hơn các mô hình ensemble do chỉ sử dụng một cây quyết định, dễ bị overfitting.

-KNN Regression có R² thấp nhất (≈ 0.773), cho thấy mô hình này không phù hợp với dữ liệu có số chiều lớn và nhiều đặc trưng đã được mã hóa.

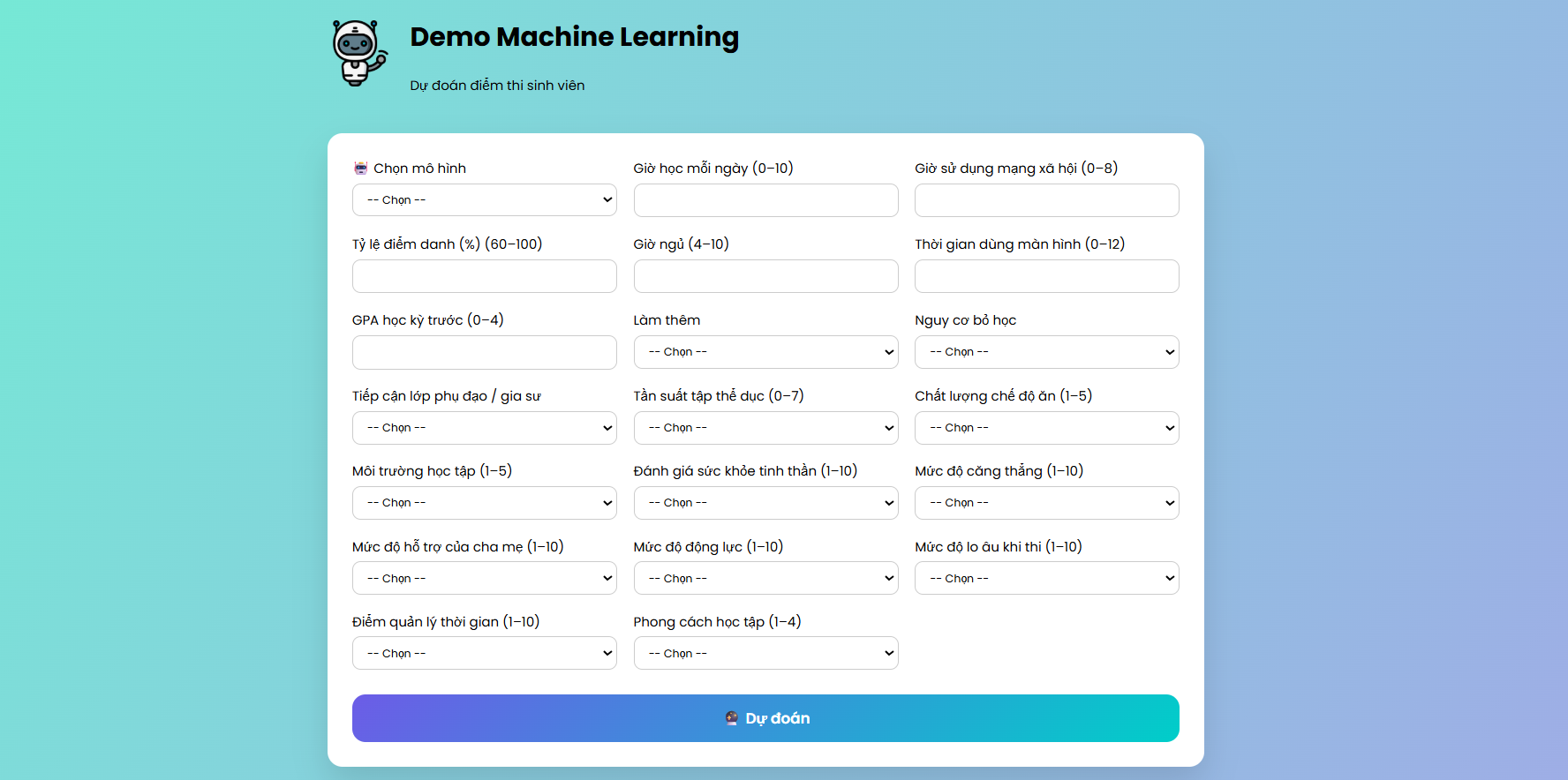
Nhìn chung, sự chênh lệch R² giữa các mô hình tốt nhất không quá lớn, điều này cho thấy bài toán dự đoán GPA có cấu trúc tương đối ổn định. Tuy nhiên, các mô hình boosting vẫn cho kết quả nhỉnh hơn nhờ khả năng học các mối quan hệ phi tuyến phức tạp.

### 3.6.14. Trực quan hóa toàn bộ Metric đánh giá của các mô hình

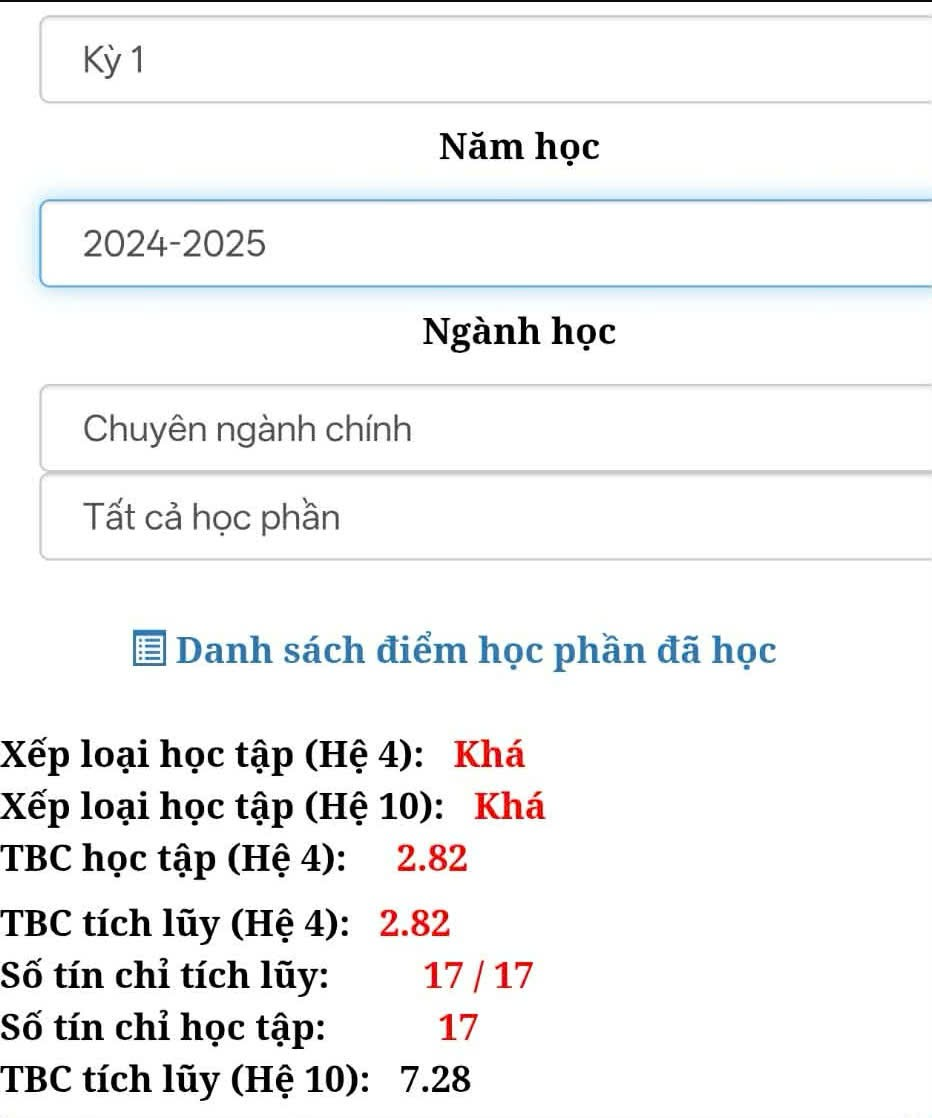
****

Hình 3.20. Trực quan hóa toàn bộ Metric đánh giá của các mô hình

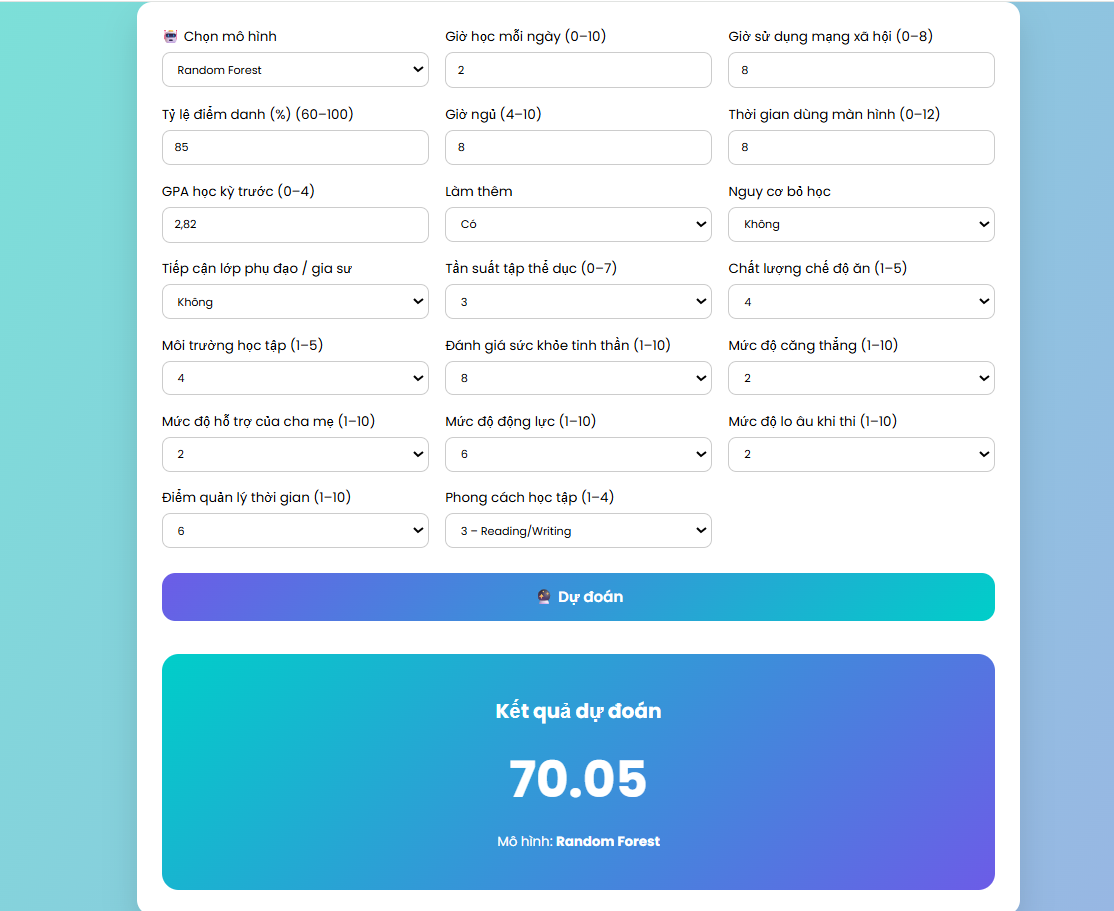
## 3.7. Demo mô hình

****

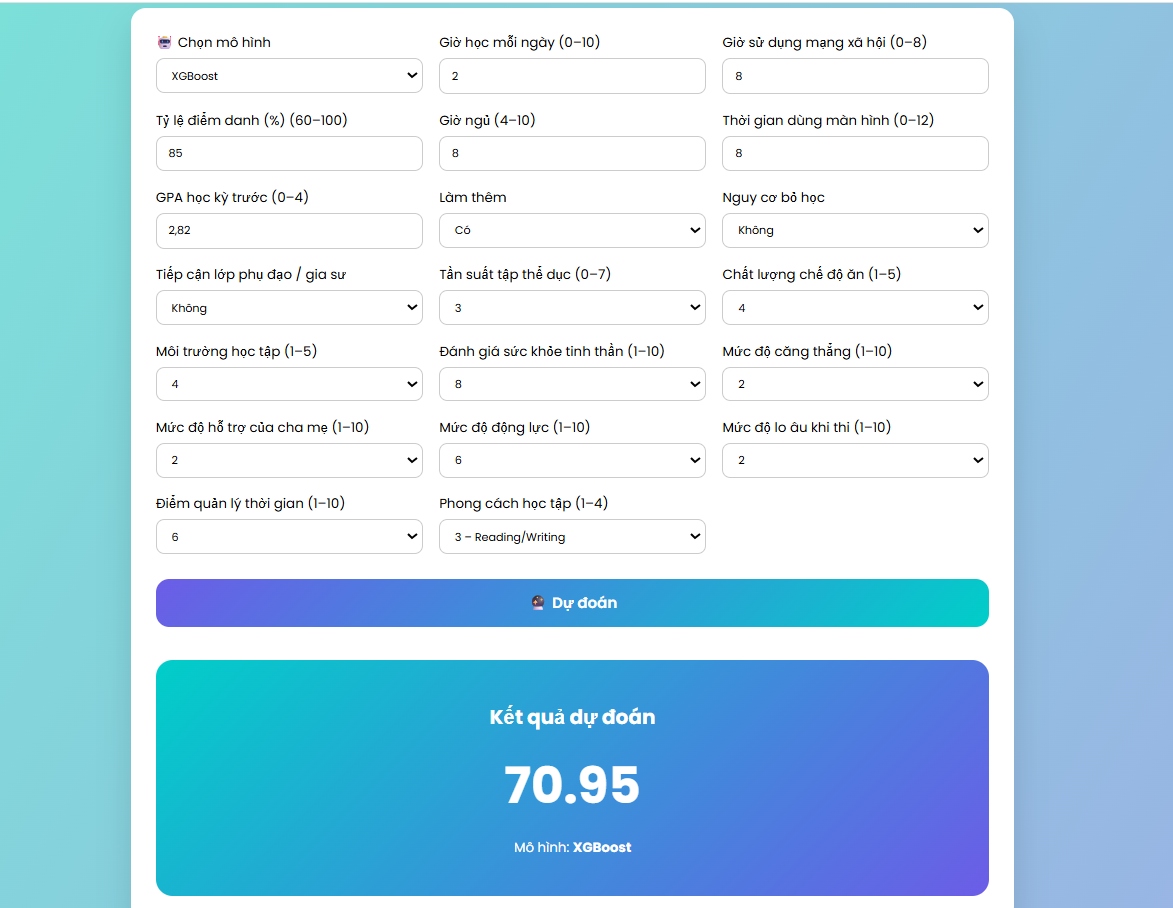
Hình 3.21. Giao diện demo dự đoán điểm GPA của sinh viên

****

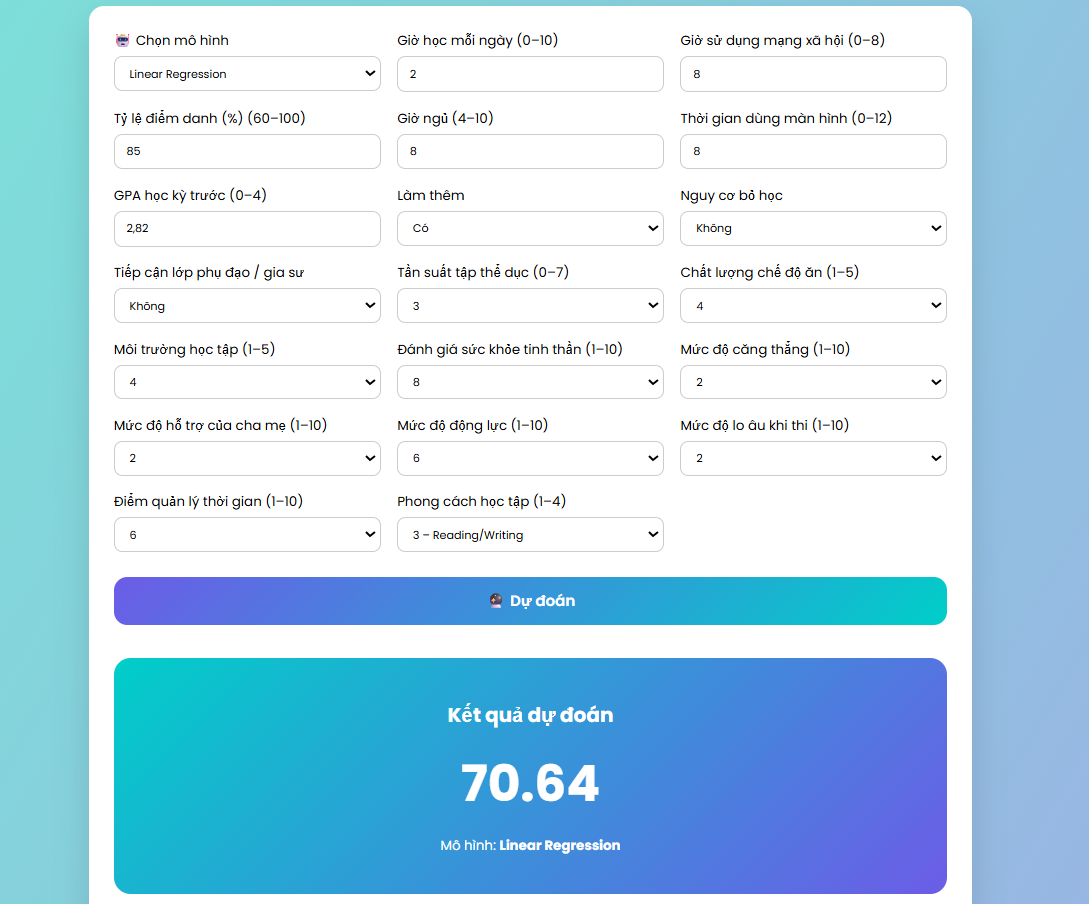
Hình 3.22. Hình ảnh dữ liệu đầu vào quan trọng (GPA kì trước)

****

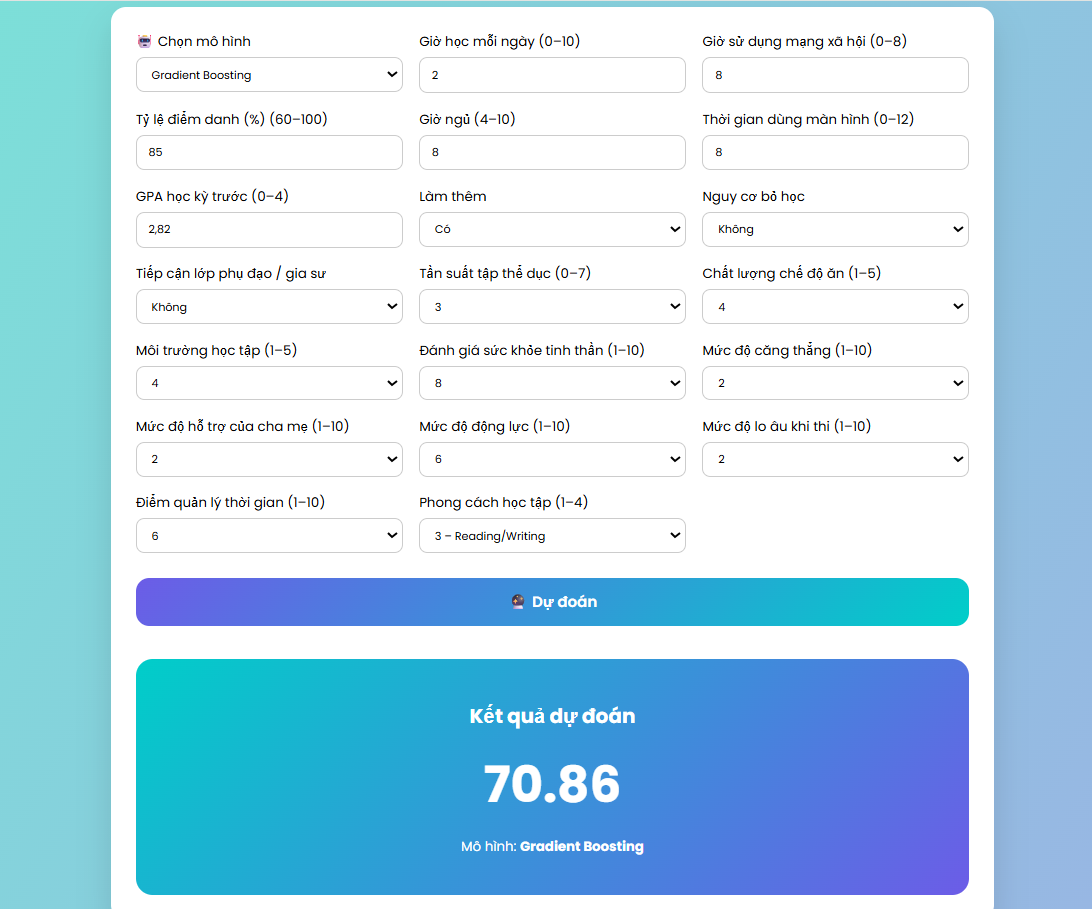
Hình 3.23. Hình ảnh kết quả dự đoán GPA bằng mô hình Random Forest



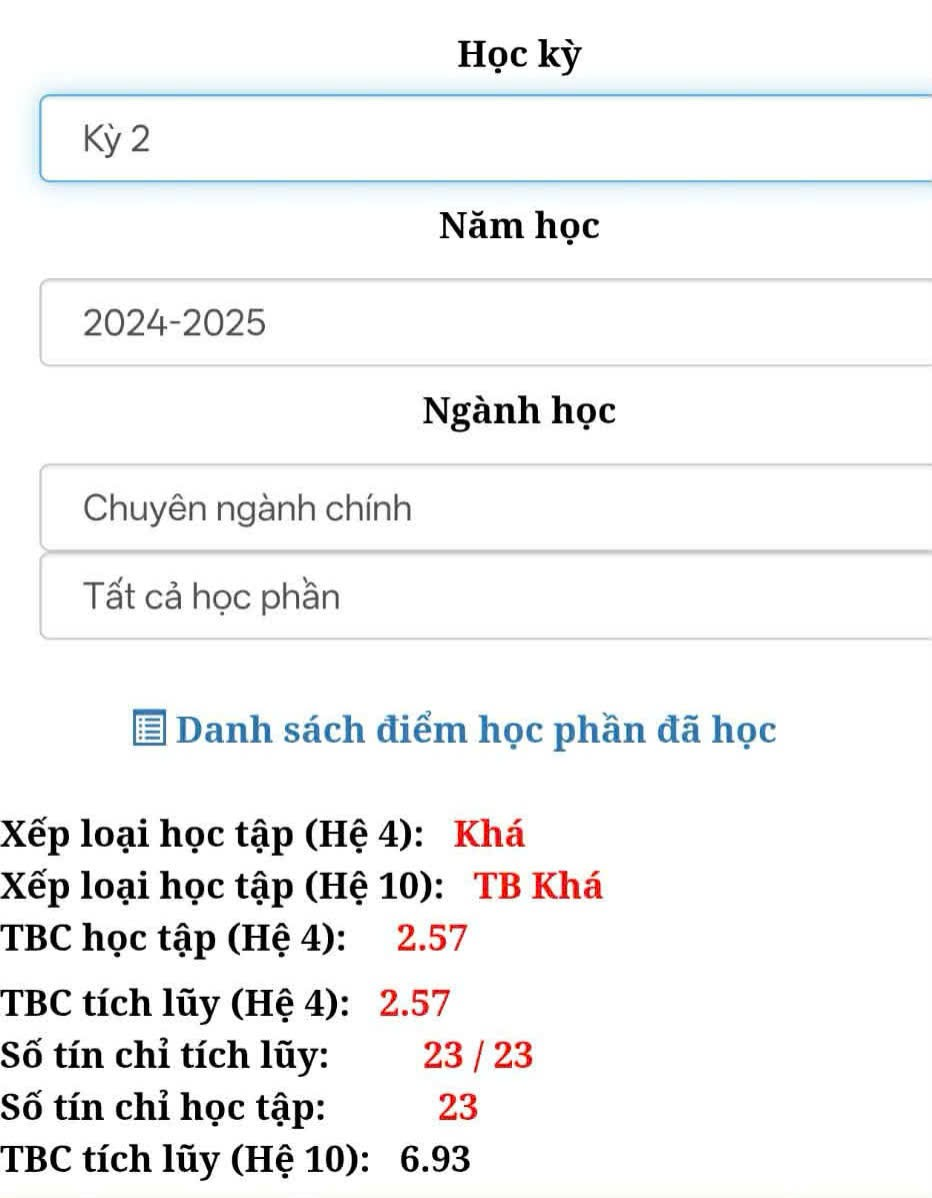
Hình 3.24. Hình ảnh kết quả dự đoán GPA bằng mô hình XGBoost

****

Hình 3.25. Hình ảnh kết quả dự đoán GPA bằng mô hình Linear Regression

****

Hình 3.26. Hình ảnh kết quả dự đoán GPA bằng mô hình Gradient Boosting



Hình 3.27. Hình ảnh kết quả thực tế điểm GPA của sinh viên

KẾT LUẬN

**KẾT QUẢ ĐẠT ĐƯỢC**

Sau quá trình nghiên cứu và thực nghiệm, đề tài đã đạt được các kết quả chính như sau:

-Đề tài đã xây dựng thành công bài toán dự đoán điểm trung bình học tập (GPA) của sinh viên dựa trên dữ liệu học tập, thói quen sinh hoạt và các yếu tố liên quan khác. Dữ liệu được xử lý đầy đủ thông qua các bước tiền xử lý, mã hóa và chuẩn hóa, đảm bảo phù hợp cho các thuật toán học máy.

-Em đã triển khai và huấn luyện nhiều mô hình học máy bao gồm Linear Regression, Random Forest Regression, Gradient Boosting Regression và XGBoost Regression. Các mô hình được huấn luyện và đánh giá thống nhất trên cùng tập dữ liệu nhằm đảm bảo tính công bằng trong so sánh.

-Kết quả đánh giá cho thấy các mô hình ensemble và boosting cho hiệu suất dự đoán tốt hơn so với mô hình tuyến tính. Trong đó, Gradient Boosting Regression và XGBoost Regression đạt các chỉ số MAE và RMSE thấp, cùng với giá trị R² cao (xấp xỉ 0.87), thể hiện khả năng dự đoán và tổng quát hóa tốt.

-Mô hình Linear Regression mặc dù đơn giản nhưng vẫn đạt kết quả tương đối cao, cho thấy giữa các đặc trưng đầu vào và GPA tồn tại mối quan hệ tuyến tính đáng kể. Điều này giúp mô hình đóng vai trò hiệu quả như một baseline để so sánh với các mô hình phức tạp hơn.

-Thông qua quá trình phân tích và trực quan hóa dữ liệu, đề tài đã xác định được các yếu tố có ảnh hưởng lớn đến GPA như GPA học kỳ trước, điểm thi, số giờ học mỗi ngày và áp lực học tập qua đó hỗ trợ việc giải thích kết quả mô hình.

Nhìn chung, đề tài đã hoàn thành đầy đủ các mục tiêu đề ra, xây dựng được mô hình dự đoán GPA có độ chính xác cao và cung cấp cơ sở thực tiễn cho việc ứng dụng học máy trong lĩnh vực giáo dục.

**HẠN CHẾ CỦA ĐỀ TÀI**

Mặc dù đề tài đã đạt được những kết quả khả quan, nghiên cứu vẫn tồn tại một số hạn chế nhất định:

-Phạm vi dữ liệu còn hạn chế, dữ liệu được thu thập từ một nguồn duy nhất và trong một khoảng thời gian nhất định, do đó khả năng tổng quát hóa của mô hình đối với các nhóm sinh viên khác có thể chưa cao.

-Chưa xem xét yếu tố thời gian, điểm GPA của sinh viên có thể thay đổi theo từng học kỳ, tuy nhiên đề tài chưa khai thác dữ liệu dạng chuỗi thời gian để phản ánh xu hướng học tập lâu dài.

-Quá trình tối ưu siêu tham số chưa toàn diện, các tham số của mô hình chủ yếu được lựa chọn dựa trên kinh nghiệm và thử nghiệm thủ công, chưa áp dụng đầy đủ các kỹ thuật tự động như GridSearchCV hoặc RandomizedSearchCV.

-Chưa phân tích sâu khả năng diễn giải mô hình, mặc dù các mô hình ensemble đạt hiệu suất cao, việc giải thích chi tiết mức độ đóng góp của từng đặc trưng đến GPA vẫn còn hạn chế.

-Chưa triển khai ứng dụng thực tế, mô hình mới dừng lại ở mức thử nghiệm trong môi trường nghiên cứu, chưa được tích hợp thành một hệ thống hỗ trợ học tập hoặc tư vấn học vụ cho sinh viên.

Những hạn chế trên là cơ sở để định hướng các hướng phát triển và cải tiến mô hình trong các nghiên cứu tiếp theo.

**HƯỚNG PHÁT TRIỂN CỦA ĐỀ TÀI**

Trong tương lai, đề tài có thể được mở rộng và phát triển theo các hướng sau:

Mở rộng và đa dạng hóa dữ liệu, thu thập dữ liệu từ nhiều trường đại học, nhiều khóa học và nhiều năm học khác nhau nhằm nâng cao khả năng tổng quát hóa của mô hình dự đoán GPA.

-Áp dụng các kỹ thuật tối ưu siêu tham số nâng cao như GridSearchCV, RandomizedSearchCV hoặc Bayesian Optimization để tìm ra bộ tham số tối ưu cho các mô hình học máy, từ đó cải thiện độ chính xác dự đoán.

Xem xét dữ liệu theo chuỗi thời gian, xây dựng các mô hình dự đoán GPA theo từng học kỳ nhằm phản ánh rõ hơn quá trình học tập lâu dài của sinh viên.

-Nghiên cứu các mô hình học sâu (Deep Learning) như mạng nơ-ron nhân tạo (Neural Network) để so sánh hiệu quả với các mô hình học máy truyền thống trong bài toán dự đoán GPA.

-Tăng cường khả năng diễn giải mô hình, sử dụng các phương pháp như feature importance, SHAP hoặc LIME để phân tích mức độ ảnh hưởng của từng đặc trưng đến kết quả dự đoán GPA.

-Triển khai ứng dụng thực tế, tích hợp mô hình dự đoán GPA vào hệ thống web hoặc phần mềm hỗ trợ học tập nhằm giúp sinh viên và nhà trường có công cụ theo dõi, tư vấn và cải thiện kết quả học tập.

Những hướng phát triển này không chỉ giúp nâng cao chất lượng mô hình mà còn mở rộng khả năng ứng dụng của đề tài trong thực tiễn giáo dục.

TÀI LIỆU THAM KHẢO