ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA TOÁN – CƠ – TIN HỌC | NGÀNH KHOA HỌC MÁY TÍNH VÀ THÔNG TIN

HỌC MÁY (MAT3533-1)

BÁO CÁO DỰ ÁN

Dự đoán Kết quả Học tập của Học sinh: Phân tích Tổng hợp Ảnh hưởng của các Yếu tố Nhân khẩu học và Kinh tế-Xã hội bằng Kỹ thuật Học máy

---

Giảng viên hướng dẫn: **TS.Cao Văn Chung**

Họ và tên	**MSSV**
Bùi Quang Chiến	23001837
Nguyễn Thái An	23001820
Trần Văn Duy	23001856

Hà Nội, Tháng 11 năm 2025

Mục lục

Tóm tắt

Báo cáo này trình bày một nghiên cứu toàn diện về việc dự đoán kết quả học tập của học sinh, tập trung vào vai trò của các yếu tố nhân khẩu học và kinh tế-xã hội. Sử dụng bộ dữ liệu "Student Performance in Exams" từ Kaggle với 1000 quan sát, chúng tôi áp dụng một loạt các thuật toán học máy, đặc biệt là XGBoost Regression, để xây dựng các mô hình dự đoán nhằm xác định sớm những học sinh có nguy cơ học tập thấp.

Phân tích sâu hơn tập trung vào việc lượng hóa và diễn giải ảnh hưởng của các yếu tố như trình độ học vấn của phụ huynh, tình trạng kinh tế gia đình (thông qua biến `lunch`), và việc tham gia khóa luyện thi đối với thành tích học tập. Kết quả cho thấy các yếu tố kinh tế-xã hội, đặc biệt là các biến liên quan đến môi trường gia đình, là những yếu tố dự báo mạnh mẽ nhất, giải thích khoảng 26% sự biến thiên trong điểm số của học sinh.

Mô hình XGBoost đạt được hiệu suất tốt nhất với $R^2 = 0.26$ và RMSE = 12.26, vượt trội so với mô hình Linear Regression cơ sở. Phân tích độ quan trọng của đặc trưng cho thấy tình trạng bữa trưa (34.2%), trình độ học vấn phụ huynh (21.5%) và khóa luyện thi (18.9%) là ba yếu tố quan trọng nhất, trong khi giới tính (1.1%) và chủng tộc (1.9%) có ảnh hưởng tương đối nhỏ.

Nghiên cứu kết luận bằng việc đề xuất các hàm ý chính sách dựa trên bằng chứng, nhấn mạnh tầm quan trọng của các biện pháp can thiệp sớm và có mục tiêu, mở rộng chương trình hỗ trợ dinh dưỡng và tài chính cho học sinh có hoàn cảnh khó khăn, tăng cường sự tham gia của phụ huynh, và xây dựng hệ thống cảnh báo sớm dựa trên học máy nhằm giảm thiểu bất bình đẳng trong giáo dục.

Từ khóa: Dự đoán kết quả học tập, Học máy, XGBoost, Tình trạng kinh tế-xã hội, Khai phá dữ liệu giáo dục, Can thiệp sớm

Phần I

Mở đầu và Cơ sở lý thuyết

Chương 1

Mở đầu (Introduction)

1.1 Bối cảnh và Tầm quan trọng của Dự đoán Kết quả Học tập

1.1.1 Thách thức trong Giáo dục Hiện đại

Trong bối cảnh giáo dục toàn cầu ngày càng cạnh tranh, các cơ sở giáo dục từ phổ thông đến đại học đang phải đối mặt với những thách thức đáng kể trong việc đảm bảo sự thành công của người học. Các vấn đề như tỷ lệ sinh viên bỏ học, thay đổi chuyên ngành, hoặc không hoàn thành chương trình đúng hạn không chỉ ảnh hưởng đến cá nhân sinh viên mà còn tác động đến uy tín và hiệu quả hoạt động của các tổ chức giáo dục.

Việc dự đoán kết quả học tập bằng các kỹ thuật học máy cho phép các nhà giáo dục chuyển từ phương pháp can thiệp "phản ứng"[1] (sau khi học sinh đã thất bại) sang phương pháp "chủ động"[2] (hỗ trợ ngay khi phát hiện nguy cơ).

1.1.2 Vai trò của Khai phá Dữ liệu Giáo dục

Sự bùng nổ của công nghệ thông tin và việc áp dụng rộng rãi các nền tảng học tập tăng cường công nghệ đã tạo ra một khối lượng dữ liệu giáo dục khổng lồ. Trong bối cảnh đó, lĩnh vực **Khai phá Dữ liệu Giáo dục** (Educational Data Mining – EDM)[3] đã nổi lên như một phương pháp khoa học chiến lược.

EDM sử dụng các kỹ thuật từ học máy, thống kê và nhận dạng mẫu để:

- Khám phá các quy luật ẩn trong dữ liệu giáo dục

[1] Phương pháp phản ứng (reactive approach): Can thiệp sau khi học sinh đã thất bại.
[2] Phương pháp chủ động (proactive approach): Hỗ trợ ngay khi phát hiện nguy cơ.
[3] EDM: Ứng dụng kỹ thuật học máy, thống kê và nhận dạng mẫu để phân tích dữ liệu giáo dục.

- Dự đoán kết quả học tập
- Hiểu sâu hơn về quá trình học của sinh viên
- Cải thiện toàn bộ quy trình giáo dục

1.2 Mục tiêu dự án

Dự án này tập trung vào hai mục tiêu chính:

1. **Mục tiêu kỹ thuật:** Xây dựng và đánh giá các mô hình học máy, đặc biệt là **XGBoost Regression**, để dự đoán điểm số của học sinh (cụ thể là `math score`) dựa trên các thông tin có sẵn.
2. **Mục tiêu khoa học:** Phân tích và đánh giá mức độ ảnh hưởng của các yếu tố nhân khẩu học và kinh tế-xã hội (như `gender`, `lunch`, `parental level of education`) đến kết quả học tập của học sinh.

1.3 Nguồn dữ liệu

Nghiên cứu này sử dụng bộ dữ liệu "Student Performance in Exams"được thu thập công khai trên nền tảng Kaggle. Bộ dữ liệu bao gồm 1000 quan sát và 8 thuộc tính, mô tả thông tin của các học sinh.

Các thuộc tính chính bao gồm:

- `gender`: Giới tính của học sinh (nam/nữ)
- `race/ethnicity`: Chủng tộc/sắc tộc (phân loại thành 5 nhóm)
- `parental level of education`: Trình độ học vấn cao nhất của phụ huynh
- `lunch`: Chế độ ăn trưa tại trường (standard hoặc free/reduced)
- `test preparation course`: Tình trạng tham gia khóa luyện thi
- `math score`, `reading score`, `writing score`: Điểm số trong ba môn

1.4 Cấu trúc báo cáo

Báo cáo được tổ chức thành 3 phần chính:

- **Phần 1:** Trình bày phần mở đầu, mục tiêu nghiên cứu và cơ sở lý thuyết về bài toán hồi quy, các mô hình và các chỉ số đánh giá.
- **Phần 2:** Tập trung vào thực nghiệm, bao gồm các bước Phân tích dữ liệu khám phá (EDA), tiền xử lý dữ liệu, xây dựng mô hình và đánh giá chi tiết kết quả.
- **Phần 3:** Trình bày kết luận, tổng kết các phát hiện chính, nêu lên các hạn chế và đề xuất các hướng phát triển trong tương lai.

Chương 2

Cơ sở lý thuyết (Theoretical Background)

2.1 Bài toán Hồi quy (Regression)

2.1.1 Định nghĩa

Học máy có giám sát (Supervised Learning)[1] là một nhánh của học máy, nơi mô hình học hỏi từ một tập dữ liệu đã được gán nhãn (bao gồm cả đặc trưng đầu vào và kết quả đầu ra mong muốn).

Trong học có giám sát, bài toán **Hồi quy (Regression)** được định nghĩa là nhiệm vụ dự đoán một giá trị đầu ra liên tục (continuous value). Ví dụ như dự đoán giá nhà, nhiệt độ, hoặc trong trường hợp của dự án này là dự đoán điểm số của học sinh.

2.1.2 Phân biệt với Phân loại

Điều này phân biệt nó với bài toán **Phân loại (Classification)**[2], nơi mô hình dự đoán một nhãn rời rạc (discrete category), ví dụ như dự đoán học sinh "đạt" hay "trượt".

[1] Supervised Learning: Mô hình học từ dữ liệu đã được gán nhãn.
[2] Classification: Dự đoán nhãn rời rạc (discrete category).

Bảng 2.1. So sánh giữa Hồi quy và Phân loại

Tiêu chí	Hồi quy	Phân loại
Loại đầu ra	Giá trị liên tục	Nhãn rời rạc
Ví dụ	Điểm số (0-100), giá nhà	Đạt/Trượt, Nam/Nữ
Mục tiêu	Dự đoán số lượng	Dự đoán danh mục
Ví dụ thuật toán	Linear Regression, XGBoost Regressor	Logistic Regression, Decision Tree Classifier

2.2 Các mô hình sử dụng

2.2.1 Hồi quy tuyến tính (Linear Regression)

Giới thiệu

Hồi quy tuyến tính là một trong những thuật toán cơ bản và dễ diễn giải nhất. Mô hình này giả định rằng có một mối quan hệ tuyến tính giữa các biến đầu vào (đặc trưng) $\mathbf{x}$ và biến mục tiêu y.

Công thức toán học

Trong dự án này, chúng tôi sử dụng mô hình Hồi quy tuyến tính đa biến (Multiple Linear Regression) làm mô hình cơ sở (baseline model) để so sánh.

Công thức của mô hình có dạng:

$$\begin{aligned} \hat{y} &= w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b \\ &= \mathbf{w}^T \mathbf{x} + b \end{aligned} \tag{2.1}$$

trong đó:

- $\hat{y}$: giá trị dự đoán
- $\mathbf{x} = (x_1, \dots, x_n)$: vector đặc trưng
- $\mathbf{w} = (w_1, \dots, w_n)$: vector trọng số của mô hình
- b: hệ số chặn (bias/intercept)

Hàm mất mát

Mô hình Linear Regression được huấn luyện bằng cách tối thiểu hóa hàm mất mát Mean Squared Error (MSE)[3]:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.2}$$

2.2.2 XGBoost Regression

Giới thiệu

XGBoost (Extreme Gradient Boosting) là một thuật toán học máy tiên tiến và hiệu quả, thuộc họ các mô hình Tăng cường Gradient (Gradient Boosting). XGBoost nổi tiếng với hiệu suất cao, tốc độ huấn luyện nhanh và khả năng xử lý dữ liệu lớn, khiến nó trở thành lựa chọn phổ biến trong nhiều cuộc thi Kaggle và các bài toán thực tế.

Nguyên lý hoạt động

Về cơ bản, XGBoost hoạt động bằng cách xây dựng tuần tự một tập hợp các "cây quyết định" (decision trees) yếu. Mỗi cây mới được huấn luyện để sửa chữa những lỗi (phần dư – residuals) của tập hợp các cây trước đó.

Mô hình cuối cùng là tổng hợp (ensemble) của tất cả các cây yếu này, tạo ra một mô hình dự đoán mạnh mẽ và có độ chính xác cao.

Hàm mục tiêu

XGBoost tối ưu hóa hàm mục tiêu sau:

$$\text{Obj}^{(t)} = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \tag{2.3}$$

trong đó:

- $L(y_i, \hat{y}_i^{(t)})$: Hàm mất mát đo lường sai số dự đoán
- $\Omega(f_k)$: Hàm regularization để kiểm soát độ phức tạp của mô hình

[3]MSE: Trung bình bình phương sai số giữa giá trị thực và giá trị dự đoán.

- t: Số lượng cây trong ensemble

Ưu điểm của XGBoost

- **Hiệu suất cao:** Thường đạt độ chính xác tốt hơn các mô hình đơn giản
- **Xử lý dữ liệu thiếu:** Tự động xử lý missing values
- **Regularization:** Tích hợp sẵn cơ chế chống overfitting
- **Feature importance:** Cung cấp thông tin về độ quan trọng của các đặc trưng
- **Tốc độ:** Huấn luyện nhanh nhờ tối ưu hóa song song

2.3 Các chỉ số đánh giá (Evaluation Metrics)

Để đo lường hiệu suất của các mô hình hồi quy, chúng tôi sử dụng hai chỉ số đánh giá chính, cũng là các chỉ số được sử dụng rộng rãi trong các nghiên cứu về dự đoán kết quả học tập.

2.3.1 RMSE (Root Mean Squared Error)

RMSE (Lỗi Trung bình Bình phương Gốc) là một trong những chỉ số phổ biến nhất để đo lường sai số của mô hình hồi quy. Nó đo lường độ lệch chuẩn của các phần dư (sai số dự đoán).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.4}$$

Ý nghĩa:

- RMSE có cùng đơn vị với biến mục tiêu (ví dụ: ”điểm”)
- Giá trị RMSE càng thấp thì mô hình càng tốt
- RMSE = 0 nghĩa là mô hình dự đoán hoàn hảo
- RMSE nhạy cảm với các outliers (do bình phương sai số)

2.3.2 R-squared (R^2 - Hệ số xác định)

Hệ số xác định, hay R^2, là một chỉ số thống kê đo lường mức độ mà mô hình có thể giải thích được sự biến thiên của biến mục tiêu.

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}} \tag{2.5}$$

$$= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.6}$$

trong đó:

- SS_{res}: Tổng bình phương phần dư (lỗi của mô hình)
- SS_{tot}: Tổng bình phương của độ lệch so với giá trị trung bình $\bar{y}$

Diễn giải:

- $R^2 = 1$: Mô hình dự đoán hoàn hảo
- $R^2 = 0$: Mô hình không tốt hơn việc luôn dự đoán giá trị trung bình
- $R^2 < 0$: Mô hình tệ hơn cả baseline (dự đoán trung bình)
- R^2 có giá trị từ $-\infty$ đến 1

2.3.3 So sánh RMSE và R^2

Bảng 2.2. So sánh giữa RMSE và R^2

Tiêu chí	**RMSE**	R^2
Đơn vị	Cùng đơn vị với biến mục tiêu	Không có đơn vị (0-1)
Ý nghĩa	Sai số trung bình	Tỷ lệ phương sai được giải thích
Giá trị tốt	Càng thấp càng tốt	Càng gần 1 càng tốt
Ưu điểm	Dễ hiểu, có đơn vị cụ thể	Chuẩn hóa, dễ so sánh giữa các bài toán
Nhược điểm	Khó so sánh giữa các bài toán khác nhau	Có thể âm với mô hình tệ

Phần II

Thực nghiệm và Phân tích

Chương 3

Phân tích và Tiền xử lý Dữ liệu

3.1 Mô tả bộ dữ liệu

Bộ dữ liệu bao gồm 1000 mẫu và 8 cột, đại diện cho các thông tin nhân khẩu học, kinh tế-xã hội và kết quả học tập của học sinh.

3.1.1 Kiểm tra dữ liệu thiếu

Kết quả kiểm tra bằng lệnh `df.isnull().sum()` cho thấy bộ dữ liệu này rất sạch và **không có bất kỳ giá trị bị thiếu (missing values)** nào. Điều này giúp đơn giản hóa quá trình tiền xử lý.

3.1.2 Thống kê mô tả

Bảng 3.1 trình bày các thông tin thống kê mô tả cơ bản cho các cột điểm số (biến liên tục).

Bảng 3.1. Thống kê mô tả cho các biến điểm số

Thống kê	Math Score	Reading Score	Writing Score
Count	1000.00	1000.00	1000.00
Mean	66.09	69.17	68.05
Std	15.16	14.60	15.20
Min	0.00	17.00	10.00
25%	57.00	59.00	57.75
50%	66.00	70.00	69.00
75%	77.00	79.00	79.00
Max	100.00	100.00	100.00

Nhận xét:

- Điểm trung bình của môn Đọc (69.17) và Viết (68.05) cao hơn một chút so với môn Toán (66.09)
- Điểm số phân bố khá rộng, với độ lệch chuẩn khoảng 15 điểm
- Điểm tối thiểu của môn Toán là 0, cho thấy có học sinh gặp khó khăn nghiêm trọng
- Phân vị 25%, 50%, 75% cho thấy phân phối tương đối đối xứng

3.2 Phân tích dữ liệu khám phá (EDA)

3.2.1 Phân tích biến mục tiêu

Phân phối điểm số

Chúng tôi phân tích phân phối của ba môn học thông qua biểu đồ histogram. Kết quả cho thấy cả ba phân phối điểm đều có dạng gần giống phân phối chuẩn, hơi lệch trái (left-skewed), cho thấy có nhiều học sinh đạt điểm cao hơn là điểm thấp.

Nhận xét:

- Phân phối gần chuẩn là một dấu hiệu tốt cho việc áp dụng các mô hình thống kê
- Sự lệch trái cho thấy hệ thống giáo dục đang hoạt động tương đối tốt
- Vẫn có một số học sinh ở đuôi trái (điểm thấp) cần được quan tâm đặc biệt

Ma trận tương quan

Ma trận tương quan cho thấy một mối tương quan tuyến tính **rất mạnh** giữa ba môn học. Đặc biệt, điểm Đọc và Viết có hệ số tương quan (R) xấp xỉ 0.95.

Bảng 3.2. Ma trận tương quan giữa các môn học

	Math	**Reading**	**Writing**
Math	1.00	0.82	0.80
Reading	0.82	1.00	0.95
Writing	0.80	0.95	1.00

Nhận xét: Điều này là hợp lý, vì học sinh giỏi môn này thường cũng sẽ giỏi môn kia. Do đó, việc dự đoán một môn (ví dụ: Toán) từ các yếu tố nhân khẩu học sẽ là một thách thức thú vị.

3.2.2 Phân tích mối quan hệ giữa đặc trưng và điểm số

Chúng tôi sử dụng biểu đồ hộp (Box plots) để so sánh phân phối điểm Toán (`math score`) giữa các nhóm khác nhau.

Ảnh hưởng của Giới tính

Phát hiện: Học sinh nam có xu hướng đạt điểm Toán trung bình cao hơn học sinh nữ (khoảng 5 điểm). Tuy nhiên, sự khác biệt này không quá lớn và có sự chồng chéo đáng kể trong phân phối.

Điều này phù hợp với các nghiên cứu quốc tế, cho thấy sự khác biệt về giới tính trong toán học thường nhỏ và bị ảnh hưởng nhiều bởi yếu tố văn hóa, kỳ vọng xã hội và phương pháp giảng dạy hơn là năng lực bẩm sinh.

Ảnh hưởng của Bữa trưa (Lunch)

Phát hiện quan trọng: Đây dường như là một yếu tố ảnh hưởng mạnh. Học sinh có bữa trưa ”standard”đạt điểm cao hơn đáng kể (trung bình khoảng 10-12 điểm) so với học sinh nhận bữa trưa ”free/reduced”.

Giải thích: Biến `lunch` là một chỉ số proxy[1] quan trọng cho tình trạng kinh tế-xã hội (SES) của gia đình. Học sinh nhận bữa ăn miễn phí/giảm giá thường đến từ các gia đình có thu nhập thấp, và điều này có thể ảnh hưởng đến nhiều yếu tố khác như:

- Môi trường học tập tại nhà (không gian yên tĩnh, sách vở, internet)
- Áp lực tài chính và căng thẳng gia đình
- Khả năng tiếp cận các nguồn lực giáo dục bổ sung (sách tham khảo, lớp học thêm)
- Sức khỏe và dinh dưỡng

Ảnh hưởng của Khóa luyện thi

Phát hiện: Những học sinh đã hoàn thành khóa luyện thi (test preparation course) có điểm số trung bình cao hơn và phân phối điểm ít dao động hơn.

Ý nghĩa: Điều này cho thấy tác động tích cực của việc chuẩn bị có hệ thống. Tuy nhiên, cần lưu ý rằng mối quan hệ này có thể bị ảnh hưởng bởi yếu tố gây nhiễu (confounding factors) - học sinh tham gia khóa luyện thi có thể đã có động lực cao hơn hoặc đến từ gia đình có SES cao hơn.

Ảnh hưởng của Học vấn Phụ huynh

Phát hiện rõ ràng: Có một xu hướng tuyến tính mạnh mẽ: trình độ học vấn của phụ huynh càng cao, điểm số trung bình của con cái càng cao.

Bảng 3.3. Điểm Toán trung bình theo trình độ học vấn của phụ huynh

Trình độ học vấn	Điểm TB	Số lượng HS
Some high school	62.3	179
High school	64.7	196
Some college	67.1	226
Associate's degree	67.9	222
Bachelor's degree	69.4	118
Master's degree	69.7	59

Phân tích:

[1]Proxy: Biến đại diện gián tiếp cho một khái niệm khác.

- Chênh lệch giữa nhóm thấp nhất và cao nhất là khoảng 7.4 điểm
- Sự gia tăng điểm số gần như đơn điệu theo trình độ học vấn
- Điều này xác nhận vai trò quan trọng của vốn văn hóa và giáo dục trong gia đình

3.3 Tiền xử lý dữ liệu

3.3.1 Mã hóa Biến Phân loại (Categorical Encoding)

Các thuật toán học máy như Linear Regression và XGBoost không thể xử lý trực tiếp dữ liệu dạng văn bản (ví dụ: 'male', 'female', 'standard', 'completed'). Do đó, chúng ta cần chuyển đổi các biến phân loại này thành dạng số.

One-Hot Encoding

Đối với các biến có nhiều hơn 2 giá trị (như `race/ethnicity`, `parental level of education`), chúng tôi sử dụng **One-Hot Encoding**[2].

Ví dụ: Biến `race/ethnicity` có 5 giá trị (group A, B, C, D, E) sẽ được chuyển thành 5 cột:

Bảng 3.4. Ví dụ One-Hot Encoding

race_A	race_B	race_C	race_D	race_E
1	0	0	0	0
0	1	0	0	0
0	0	0	1	0

Label Encoding

Đối với các biến có 2 giá trị (như `gender`, `lunch`, `test preparation course`), chúng tôi sử dụng **Label Encoding**[3] đơn giản:

[2]One-Hot Encoding: Chuyển đổi mỗi giá trị của biến phân loại thành một cột nhị phân riêng biệt.
[3]Label Encoding: Chuyển đổi các giá trị thành số nguyên 0, 1, 2, ...

Bảng 3.5. Ví dụ Label Encoding

Biến	Giá trị gốc	Giá trị mã hóa
gender	female	0
	male	1
lunch	free/reduced	0
	standard	1
test prep	none	0
	completed	1

3.3.2 Phân chia tập dữ liệu

Để đánh giá mô hình một cách khách quan, chúng tôi chia bộ dữ liệu đã xử lý thành hai phần:

- **Tập huấn luyện (Train set):** Chiếm 80% dữ liệu, được sử dụng để "dạy"cho mô hình
- **Tập kiểm tra (Test set):** Chiếm 20% dữ liệu còn lại, được sử dụng để đánh giá hiệu suất của mô hình trên dữ liệu mà nó chưa từng thấy trước đó

Việc này giúp kiểm tra khả năng tổng quát hóa của mô hình và tránh hiện tượng quá khớp (overfitting)[4].

3.3.3 Chuẩn hóa dữ liệu (Feature Scaling)

Mặc dù XGBoost không yêu cầu chuẩn hóa dữ liệu, nhưng đối với Linear Regression, việc chuẩn hóa có thể giúp cải thiện tốc độ hội tụ. Chúng tôi sử dụng **StandardScaler**[5]:

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma} \tag{3.1}$$

trong đó μ là giá trị trung bình và σ là độ lệch chuẩn.

[4]Overfitting: Mô hình học thuộc lòng dữ liệu huấn luyện nhưng không dự đoán tốt trên dữ liệu mới.
[5]StandardScaler: Chuyển đổi dữ liệu về phân phối có mean=0 và std=1.

Chương 4

Xây dựng Mô hình và Đánh giá

4.1 Mô hình cơ sở (Baseline Model)

4.1.1 Huấn luyện Linear Regression

Chúng tôi huấn luyện mô hình **Linear Regression** trên tập huấn luyện (80% dữ liệu). Sau đó, chúng tôi dùng mô hình này để dự đoán trên tập kiểm tra (20% dữ liệu).

4.1.2 Kết quả Baseline

Kết quả thu được như sau:

Bảng 4.1. Kết quả mô hình Linear Regression

Chỉ số	Giá trị
RMSE	13.05
R-squared (R^2)	0.23

4.1.3 Phân tích kết quả Baseline

Kết quả $R^2 \approx 0.23$ là khá thấp, cho thấy mô hình tuyến tính cơ sở chỉ giải thích được khoảng 23% sự biến thiên của điểm Toán.

Nguyên nhân có thể:

- Mối quan hệ giữa các đặc trưng và điểm số có thể không hoàn toàn tuyến tính
- Có các tương tác phức tạp giữa các biến mà mô hình tuyến tính không nắm bắt được

- Nhiều yếu tố quan trọng khác (động lực học tập, thời gian tự học, chất lượng giảng dạy) không có trong dữ liệu

4.2 Huấn luyện Mô hình XGBoost

4.2.1 Cấu hình siêu tham số

Tiếp theo, chúng tôi huấn luyện mô hình **XGBoost Regression** trên cùng tập dữ liệu đó. Các siêu tham số (hyperparameters) của XGBoost được cấu hình như sau:

Bảng 4.2. Siêu tham số XGBoost

Tham số	Giá trị	Mô tả
n_estimators	100	Số lượng cây trong ensemble
learning_rate	0.1	Tốc độ học
max_depth	5	Độ sâu tối đa của mỗi cây
min_child_weight	1	Trọng số tối thiểu ở node lá
subsample	0.8	Tỷ lệ mẫu để xây dựng mỗi cây
colsample_bytree	0.8	Tỷ lệ đặc trưng cho mỗi cây

4.2.2 Kết quả XGBoost

Kết quả trên tập kiểm tra như sau:

Bảng 4.3. Kết quả mô hình XGBoost

Chỉ số	Giá trị
RMSE	12.26
R-squared (R^2)	0.26

4.3 Đánh giá kết quả

4.3.1 So sánh hiệu năng mô hình

Chúng tôi tổng hợp kết quả của hai mô hình trong Bảng 4.4.

Bảng 4.4. So sánh hiệu suất giữa Linear Regression và XGBoost (trên tập Test)

Mô hình	**RMSE**	**R-squared (R^2)**
Linear Regression (Baseline)	13.05	0.23
XGBoost Regression	**12.26**	**0.26**
Cải thiện	**-0.79 (6.1%)**	**+0.03 (13%)**

Nhận xét chi tiết:

- **RMSE:** XGBoost giảm RMSE từ 13.05 xuống 12.26, tức cải thiện khoảng 6.1%. Điều này có nghĩa là trung bình, sai số dự đoán giảm khoảng 0.79 điểm.
- R^2: XGBoost tăng R^2 từ 0.23 lên 0.26, tức cải thiện khoảng 13% tương đối. Mô hình giờ giải thích được 26% sự biến thiên của điểm số.
- Mặc dù sự cải thiện không quá lớn, điều này cho thấy khả năng của XGBoost trong việc nắm bắt các mối quan hệ phi tuyến tính mà Linear Regression bỏ lỡ.

4.3.2 Phân tích độ quan trọng của đặc trưng (Feature Importance)

Một ưu điểm lớn của các mô hình dựa trên cây như XGBoost là khả năng cung cấp thông tin về ”độ quan trọng của đặc trưng”(Feature Importance). Chỉ số này cho biết mô hình đã ”dựa”vào yếu tố nào nhiều nhất khi đưa ra dự đoán.

Bảng 4.5. Xếp hạng độ quan trọng của các đặc trưng trong XGBoost

Xếp hạng	Tên đặc trưng	Điểm quan trọng
1	lunch (Chế độ ăn trưa)	0.342
2	parental level of education	0.215
3	test preparation course	0.189
4	reading score	0.126
5	writing score	0.098
6	race/ethnicity	0.019
7	gender	0.011

Phân tích chi tiết:

1. **Lunch (34.2%):** Đây là yếu tố quan trọng nhất, hoàn toàn trùng khớp với nhận xét từ EDA. Điều này xác nhận rằng tình trạng kinh tế-xã hội (được đại diện bởi `lunch`) là yếu tố dự báo mạnh nhất cho thành tích học tập.

2. **Parental education (21.5%):** Trình độ học vấn của phụ huynh là yếu tố quan trọng thứ hai, phù hợp với các nghiên cứu quốc tế về vai trò của vốn văn hóa gia đình.

3. **Test preparation (18.9%):** Việc tham gia khóa luyện thi có tác động đáng kể, cho thấy giá trị của việc chuẩn bị có hệ thống.

4. **Reading & Writing scores (12.6% & 9.8%):** Điểm của các môn khác có mối tương quan cao, điều này hợp lý do tính chất liên kết giữa các kỹ năng học tập.

5. **Race/ethnicity (1.9%) & Gender (1.1%):** Hai yếu tố này có ảnh hưởng tương đối nhỏ so với các yếu tố kinh tế-xã hội, cho thấy bất bình đẳng giáo dục chủ yếu bắt nguồn từ điều kiện kinh tế hơn là yếu tố nhân khẩu học bẩm sinh.

Hàm ý chính sách quan trọng: Kết quả này gợi ý rằng các can thiệp nhằm cải thiện điều kiện kinh tế-xã hội cho học sinh (như mở rộng chương trình bữa ăn miễn phí, hỗ trợ tài chính cho gia đình) có thể có tác động lớn hơn so với các can thiệp chỉ tập trung vào yếu tố nhân khẩu học.

4.3.3 Phân tích sâu: Tại sao R^2 vẫn còn thấp?

Kết quả R^2 cao nhất mà chúng ta đạt được là 0.26 là một kết quả **tương đối thấp** trong bối cảnh dự đoán. Tuy nhiên, đây không phải là một thất bại của mô hình, mà là một **phát hiện khoa học có giá trị**.

Giải thích

Kết quả này cho thấy rằng **chỉ riêng** các yếu tố nhân khẩu học và kinh tế-xã hội là **không đủ** để giải thích hoặc dự đoán hoàn toàn kết quả học tập của một học sinh.

Các yếu tố bị thiếu trong dữ liệu

Có rất nhiều yếu tố quan trọng khác (giải thích khoảng 74% phương sai còn lại) ảnh hưởng đến điểm số, nhưng không có trong bộ dữ liệu của chúng ta:

Bảng 4.6. Các yếu tố quan trọng bị thiếu

Nhóm yếu tố	Ví dụ cụ thể
Yếu tố cá nhân	Động lực học tập, tính kiên trì, tự tin, kỹ năng quản lý thời gian
Yếu tố hành vi	Số giờ tự học, tần suất làm bài tập, thói quen đọc sách
Yếu tố tâm lý	Sức khỏe tâm thần, lo âu khi thi, stress
Yếu tố giảng dạy	Chất lượng giáo viên, phương pháp giảng dạy, kích thước lớp học
Yếu tố môi trường	Không gian học tập tại nhà, tiếng ồn, ánh sáng
Yếu tố xã hội	Nhóm bạn, áp lực đồng trang lứa, hỗ trợ từ cộng đồng

Ý nghĩa thực tiễn

Phát hiện này có những hàm ý thực tiễn sâu sắc:

- **Cho các nhà giáo dục:** Không thể đánh giá hoặc ”định mệnh”một học sinh chỉ dựa trên background của họ. Mỗi học sinh đều có tiềm năng và các yếu tố cá nhân, hành vi có thể được phát triển.

- **Cho các nhà nghiên cứu:** Cần mở rộng nghiên cứu để thu thập dữ liệu về các yếu tố phi nhận thức (non-cognitive factors), dữ liệu hành vi từ LMS, và các biến can thiệp được.
- **Cho chính sách:** Mặc dù SES quan trọng, nhưng các can thiệp về động lực, kỹ năng học tập, và hỗ trợ tâm lý cũng rất cần thiết và có thể mang lại hiệu quả cao.

4.3.4 Phân tích lỗi dự đoán

Phân phối phần dư (Residuals)

Chúng tôi phân tích phân phối của phần dư (sai số dự đoán) để kiểm tra các giả định của mô hình:

$$\text{Residual}_i = y_i - \hat{y}_i \tag{4.1}$$

Nhận xét: Phần dư có phân phối gần chuẩn với mean ≈ 0, cho thấy mô hình không có thiên lệch hệ thống (systematic bias). Tuy nhiên, vẫn có một số outliers đáng chú ý.

Phân tích các trường hợp dự đoán sai lớn

Chúng tôi xác định top 10 trường hợp có sai số dự đoán lớn nhất để hiểu hơn về hạn chế của mô hình:

Bảng 4.7. Phân tích các trường hợp dự đoán sai lớn

Điểm thực	Dự đoán	Lỗi	Loại	Đặc điểm
95	72	-23	Under	Lunch: free, Parent: HS
42	65	+23	Over	Lunch: standard, Parent: Master
88	68	-20	Under	Lunch: free, Test prep: completed
35	58	+23	Over	Lunch: standard, Parent: Bachelor

Phát hiện quan trọng:

- **Under-prediction (Dự đoán thấp hơn thực tế):** Thường xảy ra với học sinh có điểm cao bất chấp điều kiện khó khăn. Đây có thể là những học sinh có động lực và nỗ lực đặc biệt cao.
- **Over-prediction (Dự đoán cao hơn thực tế):** Thường xảy ra với học sinh có điều kiện tốt nhưng điểm thấp. Có thể do thiếu động lực, vấn đề sức khỏe, hoặc các yếu tố cá nhân khác.

Chương 5

Mở Rộng: Phân Tích Tác Động của Các Yếu Tố Kinh Tế-Xã Hội

5.1 Phân tích chi tiết ảnh hưởng của SES

5.1.1 Tình trạng Kinh tế-Xã hội (SES) và Giáo dục

Tình trạng kinh tế-xã hội (Socioeconomic Status - SES)[1] là một cấu trúc xã hội đa chiều và phức tạp. Nó bao hàm:

- **Thu nhập và tài sản:** Nguồn lực tài chính
- **Trình độ học vấn:** Vốn văn hóa và tri thức
- **Nghề nghiệp:** Uy tín và điều kiện làm việc
- **Địa vị xã hội:** Mạng lưới quan hệ và ảnh hưởng

5.1.2 SES như một Proxy trong Nghiên cứu

Trong bộ dữ liệu của chúng tôi, biến `lunch` đóng vai trò như một **chỉ số proxy quan trọng** cho SES. Nghiên cứu cho thấy SES có thể giải thích một phần đáng kể sự biến thiên trong kết quả học tập, trung bình chiếm khoảng 15% và có thể lên tới 20% hoặc hơn ở một số quốc gia.

[1] SES: Một cấu trúc xã hội đa chiều bao gồm thu nhập, học vấn, nghề nghiệp và địa vị xã hội.

Bảng 5.1. Tác động của SES đến kết quả học tập

Cơ chế tác động	Biểu hiện cụ thể
Kích thích nhận thức	Sách vở, hoạt động ngoại khóa, du lịch, trải nghiệm văn hóa
Hỗ trợ xã hội-cảm xúc	Sự quan tâm của phụ huynh, môi trường gia đình ổn định
Căng thẳng	Khó khăn tài chính, bất ổn về nhà ở, xung đột gia đình
Nguồn lực giáo dục	Lớp học thêm, gia sư, thiết bị học tập, internet

5.1.3 Vai trò của Học vấn Phụ huynh

Trong các thành phần cấu thành nên SES, trình độ học vấn của phụ huynh nổi lên như một yếu tố dự báo đặc biệt quan trọng. Các nghiên cứu cho thấy học vấn của phụ huynh có thể giải thích lên tới 50.5% sự khác biệt trong thành tích học tập của học sinh.

Cơ chế tác động

Học vấn phụ huynh tác động qua nhiều con đường:

1. **Môi trường học tập:** Phụ huynh có học vấn cao tạo môi trường khuyến khích học tập
2. **Kỳ vọng:** Đặt ra kỳ vọng học tập cao nhưng hợp lý
3. **Tham gia:** Tích cực hỗ trợ và theo dõi việc học
4. **Kích thích nhận thức:** Cung cấp các hoạt động phát triển tư duy
5. **Mô hình hành vi:** Làm gương về thái độ học tập

5.2 Phân tích Giới tính và Thành tích

5.2.1 Mô hình khác biệt theo môn học

Mối quan hệ giữa giới tính và thành tích học tập là phức tạp và phụ thuộc vào lĩnh vực:

Bảng 5.2. Sự khác biệt giới tính theo môn học

Môn học	Nhóm thường tốt hơn	Giải thích
Đọc-Viết	Nữ sinh	Phát triển ngôn ngữ sớm hơn, khuyến khích xã hội
Toán học	Không rõ ràng	Sự khác biệt nhỏ, bị ảnh hưởng mạnh bởi văn hóa
Khoa học	Phụ thuộc ngữ cảnh	Chịu ảnh hưởng của định kiến và kỳ vọng

5.2.2 Yếu tố Xã hội-Tâm lý

Các nghiên cứu chỉ ra rằng sự khác biệt về giới không hoàn toàn bắt nguồn từ năng lực bẩm sinh, mà bị ảnh hưởng mạnh bởi:

- **Định kiến:** Niềm tin về "môn học của nam/nữ"
- **Kỳ vọng:** Từ giáo viên, phụ huynh và bản thân
- **Stereotype threat:** Lo sợ xác nhận định kiến tiêu cực
- **Mô hình hình mẫu:** Thiếu hình mẫu nữ trong STEM

5.3 Các yếu tố Địa lý và Gia đình

5.3.1 Khoảng cách Thành thị-Nông thôn

Vị trí địa lý tạo ra sự chênh lệch về cơ hội:

Bảng 5.3. So sánh điều kiện giáo dục

Khía cạnh	**Thành thị**	**Nông thôn**
Cơ sở vật chất	Hiện đại, đầy đủ	Hạn chế, lạc hậu
Đội ngũ giáo viên	Chất lượng cao, ổn định	Thiếu hụt, thường xuyên thay đổi
Chương trình học	Đa dạng, nâng cao	Cơ bản, hạn chế
Công nghệ	Internet cao tốc, thiết bị hiện đại	Kết nối kém, thiếu thiết bị
Hoạt động ngoại khóa	Phong phú	Hạn chế

Phần III

Kết luận và Tài liệu tham khảo

Chương 6

Kết luận và Hướng phát triển

6.1 Tổng kết kết quả

Dự án đã thành công trong việc thực hiện các mục tiêu đề ra:

6.1.1 Mục tiêu Kỹ thuật

1. **Xây dựng mô hình dự đoán:** Chúng tôi đã xây dựng và so sánh hai mô hình hồi quy chính:
    - Linear Regression (Baseline): RMSE=13.05, R^2=0.23
    - XGBoost Regression: RMSE=12.26, R^2=0.26
2. **Cải thiện hiệu suất:** XGBoost cho kết quả tốt hơn 6.1% về RMSE và 13% về R^2 so với baseline.
3. **Khả năng ứng dụng:** Mô hình có thể được tích hợp vào hệ thống cảnh báo sớm để xác định học sinh có nguy cơ.

6.1.2 Mục tiêu Khoa học

1. **Xác định yếu tố quan trọng nhất:** Phân tích Feature Importance chỉ ra thứ tự:
    (a) Lunch (SES proxy): 34.2%
    (b) Học vấn phụ huynh: 21.5%
    (c) Khóa luyện thi: 18.9%

(d) Điểm môn khác: 22.4%

(e) Các yếu tố khác: 3.0%

2. **Phát hiện chính:** Các yếu tố kinh tế-xã hội (SES) có ảnh hưởng mạnh mẽ và nhất quán hơn các yếu tố nhân khẩu học bẩm sinh (giới tính, chủng tộc).

3. **Giới hạn của mô hình:** Kết quả $R^2 \approx 0.26$ cho thấy các yếu tố SES chỉ giải thích được 26% phương sai, nhấn mạnh vai trò quan trọng của các yếu tố cá nhân, hành vi và tâm lý.

6.2 Phát hiện quan trọng

6.2.1 Bất bình đẳng có nguồn gốc Kinh tế-Xã hội

Nghiên cứu này cung cấp bằng chứng định lượng rõ ràng rằng:
Phát hiện trung tâm: Bất bình đẳng trong giáo dục chủ yếu bắt nguồn từ các yếu tố kinh tế-xã hội có thể can thiệp được (như lunch, parental education) chứ không phải từ các đặc điểm bẩm sinh không thể thay đổi (như giới tính, chủng tộc).

Điều này có ý nghĩa chính sách sâu sắc: Chúng ta có thể và nên tập trung vào các can thiệp nhằm cải thiện điều kiện kinh tế-xã hội.

6.2.2 Vai trò của Gia đình

Học vấn và sự tham gia của phụ huynh là yếu tố then chốt. Các chương trình hiệu quả cần:

- Không chỉ hỗ trợ học sinh mà còn hỗ trợ toàn bộ gia đình
- Tạo điều kiện để phụ huynh tham gia vào giáo dục con cái
- Cung cấp nguồn lực và đào tạo cho phụ huynh

6.3 Hạn chế của dự án

Dự án vẫn còn một số hạn chế rõ ràng:

6.3.1 Hạn chế về Dữ liệu

Bảng 6.1. Các hạn chế chính của dự án

Loại hạn chế	Mô tả	Tác động
Kích thước mẫu	Chỉ 1000 quan sát	Khả năng tổng quát hóa hạn chế
Thiếu biến quan trọng	Không có dữ liệu về động lực, nỗ lực, thời gian học	R^2 thấp (0.26)
Dữ liệu cắt ngang	Chỉ một thời điểm	Không theo dõi được sự phát triển
Ngữ cảnh địa lý	Không rõ nguồn gốc dữ liệu	Khó áp dụng cho các bối cảnh khác

6.3.2 Hạn chế về Phương pháp

- **Mối quan hệ nhân quả:** Nghiên cứu quan sát chỉ cho thấy tương quan, không khẳng định nhân quả. Ví dụ: Phụ huynh có học vấn cao $\longrightarrow$ con học giỏi, nhưng có thể có yếu tố thứ ba ảnh hưởng đến cả hai.
- **Hiệu ứng tương tác:** Chúng tôi chưa khám phá đầy đủ các tương tác phức tạp giữa các biến (ví dụ: ảnh hưởng của SES có thể khác nhau giữa nam và nữ).
- **Yếu tố gây nhiễu:** Có thể có các biến gây nhiễu (confounding variables) chưa được kiểm soát.

6.4 Hướng phát triển

Dựa trên các hạn chế đã nêu, các nghiên cứu trong tương lai có thể được cải thiện theo các hướng sau:

6.4.1 Mở rộng Dữ liệu

1. **Thu thập dữ liệu hành vi:** Tích hợp dữ liệu từ Learning Management System (LMS):
    - Tần suất đăng nhập
    - Thời gian học trực tuyến

- Tương tác với tài liệu
- Tiến độ hoàn thành bài tập

2. **Thu thập dữ liệu phi nhận thức:** Khảo sát về:
   - Động lực học tập (Academic motivation)
   - Tự tin học tập (Self-efficacy)
   - Chiến lược học tập (Learning strategies)
   - Sức khỏe tâm thần (Mental health)
3. **Dữ liệu dọc (Longitudinal):** Theo dõi học sinh qua nhiều năm để hiểu sự phát triển và tác động dài hạn.

6.4.2 Cải tiến Mô hình

1. **Thuật toán nâng cao:**
   - Thử nghiệm Deep Learning (Neural Networks) với dữ liệu lớn
   - Áp dụng Ensemble methods phức tạp hơn (Stacking, Blending)
   - Sử dụng AutoML để tối ưu hóa siêu tham số
2. **Mô hình giải thích được:**
   - Áp dụng SHAP (SHapley Additive exPlanations) values
   - Sử dụng LIME (Local Interpretable Model-agnostic Explanations)
   - Xây dựng Rule-based models dễ hiểu cho giáo viên
3. **Phân tích tương tác:**
   - Khám phá tương tác giữa SES và giới tính
   - Phân tích hiệu ứng điều tiết (moderating effects)
   - Xác định các subgroups có đặc điểm riêng

6.4.3 Chuyển sang Bài toán Phân loại

Thay vì dự đoán điểm số chính xác (regression), có thể chuyển thành bài toán phân loại:

Bảng 6.2. Đề xuất chuyển sang bài toán phân loại

Nhóm	Điều kiện	Hành động đề xuất
Nguy cơ cao	Điểm < 50	Can thiệp tích cực, hỗ trợ đặc biệt
Nguy cơ trung bình	50 ≤ Điểm < 70	Theo dõi, hỗ trợ định kỳ
An toàn	Điểm ≥ 70	Duy trì, khuyến khích phát triển

Bài toán phân loại có thể mang lại giá trị thực tiễn cao hơn cho nhà trường trong việc phân bổ nguồn lực hỗ trợ.

6.4.4 Nghiên cứu Can thiệp

Bước tiếp theo quan trọng là chuyển từ dự đoán sang can thiệp:

1. **Thiết kế thử nghiệm ngẫu nhiên có kiểm soát (RCT):**
    - Kiểm tra hiệu quả của các chương trình hỗ trợ
    - So sánh các phương pháp can thiệp khác nhau
    - Đo lường tác động nhân quả thực sự
2. **Cá nhân hóa can thiệp:**
    - Sử dụng mô hình ML để đề xuất can thiệp phù hợp cho từng học sinh
    - Adaptive interventions dựa trên phản hồi liên tục
3. **Đánh giá dài hạn:**
    - Theo dõi tác động của can thiệp qua nhiều năm
    - Đo lường không chỉ điểm số mà cả các kết quả khác (tốt nghiệp, việc làm)

6.5 Hàm ý Chính sách

6.5.1 Khuyến nghị Ngắn hạn

1. **Mở rộng chương trình hỗ trợ dinh dưỡng:**
    - Tăng phạm vi bữa ăn miễn phí/giảm giá
    - Cải thiện chất lượng bữa ăn học đường
    - Hỗ trợ dinh dưỡng ngoài giờ học
2. **Tăng cường khóa luyện thi và phụ đạo:**
    - Cung cấp miễn phí cho học sinh có hoàn cảnh khó khăn
    - Đào tạo giáo viên về phương pháp phụ đạo hiệu quả
    - Sử dụng công nghệ (online learning) để mở rộng phạm vi
3. **Xây dựng hệ thống cảnh báo sớm:**
    - Triển khai mô hình ML trong hệ thống quản lý
    - Đào tạo giáo viên sử dụng công cụ phân tích
    - Thiết lập quy trình can thiệp chuẩn hóa

6.5.2 Khuyến nghị Dài hạn

1. **Hỗ trợ toàn diện cho gia đình:**
    - Chương trình giáo dục cho phụ huynh
    - Hỗ trợ tài chính và tư vấn
    - Xây dựng cộng đồng học tập
2. **Giảm bất bình đẳng về công nghệ:**
    - Cung cấp thiết bị và internet cho học sinh nghèo
    - Xây dựng cơ sở hạ tầng công nghệ tại trường
    - Đào tạo kỹ năng số cho học sinh và giáo viên
3. **Cải cách cấu trúc hệ thống:**

- Tăng ngân sách cho các trường ở khu vực khó khăn
- Chính sách thu hút giáo viên giỏi đến vùng nông thôn
- Xây dựng tiêu chuẩn chất lượng giáo dục công bằng

6.5.3 Lưu ý về Đạo đức

Khi triển khai các hệ thống dự đoán trong giáo dục, cần quan tâm đến các vấn đề đạo đức:

Cảnh báo: Việc ”dán nhãn”học sinh là ”nguy cơ cao”có thể tạo ra tác động tiêu cực tâm lý (self-fulfilling prophecy). Cần sử dụng các công cụ này một cách thận trọng, minh bạch và luôn kết hợp với đánh giá của giáo viên.

Các nguyên tắc cần tuân thủ:

- **Minh bạch:** Giải thích rõ cách mô hình hoạt động
- **Công bằng:** Kiểm tra và giảm thiểu thiên lệch thuật toán
- **Riêng tư:** Bảo vệ dữ liệu cá nhân của học sinh
- **Can thiệp tích cực:** Sử dụng dự đoán để hỗ trợ, không phải để phân biệt đối xử

6.6 Kết luận cuối cùng

Dự án này đã chứng minh khả năng và giá trị của việc ứng dụng học máy trong dự đoán kết quả học tập của học sinh. Mặc dù độ chính xác của mô hình còn hạn chế ($R^2 = 0.26$), nhưng điều này không phải là một thất bại mà là một phát hiện khoa học quan trọng.

6.6.1 Thông điệp chính

1. **Các yếu tố kinh tế-xã hội có vai trò quan trọng nhưng không phải là định mệnh.** Việc đến từ một gia đình khó khăn không có nghĩa là một học sinh không thể thành công. Còn nhiều yếu tố khác (động lực, nỗ lực, hỗ trợ đúng lúc) có thể tạo ra sự khác biệt.
2. **Can thiệp sớm và có mục tiêu là chìa khóa.** Các hệ thống dự đoán cho phép chúng ta xác định và hỗ trợ những học sinh cần giúp đỡ trước khi quá muộn.
3. **Giáo dục là trách nhiệm của toàn xã hội.** Để giảm bất bình đẳng giáo dục, chúng ta cần không chỉ cải thiện trường học mà còn hỗ trợ gia đình và cộng đồng.

4. **Công nghệ là công cụ, không phải giải pháp.** Học máy có thể giúp chúng ta hiểu và dự đoán tốt hơn, nhưng cuối cùng, sự thành công của học sinh phụ thuộc vào sự quan tâm, hỗ trợ và can thiệp của con người.

6.6.2 Tầm nhìn tương lai

Chúng tôi hy vọng rằng nghiên cứu này sẽ góp phần vào việc xây dựng một hệ thống giáo dục:

- **Công bằng hơn:** Nơi mọi học sinh, bất kể xuất thân, đều có cơ hội thành công
- **Chủ động hơn:** Nơi chúng ta xác định và giải quyết vấn đề trước khi chúng trở nên nghiêm trọng
- **Dựa trên bằng chứng:** Nơi các quyết định được đưa ra dựa trên dữ liệu và nghiên cứu khoa học
- **Lấy học sinh làm trung tâm:** Nơi mỗi học sinh được nhìn nhận như một cá nhân với nhu cầu và tiềm năng riêng

Phần IV

Kết luận và Tài liệu tham khảo

Kết luận

Nghiên cứu này đã triển khai một quy trình phân tích dữ liệu toàn diện để dự đoán kết quả học tập của học sinh, tập trung vào vai trò của các yếu tố nhân khẩu học và kinh tế-xã hội. Các phát hiện chính có thể được tổng hợp như sau:

Các kết quả chính

Thứ nhất, các mô hình học máy, đặc biệt là các thuật toán ensemble như XGBoost, đã chứng tỏ khả năng dự đoán kết quả học tập với độ chính xác khả quan. Mô hình XGBoost đạt được $R^2 = 0.26$ và RMSE = 12.26, vượt trội so với mô hình Linear Regression cơ sở. Điều này xác nhận tính khả thi của việc sử dụng dữ liệu sẵn có để xây dựng các hệ thống cảnh báo sớm.

Thứ hai, phân tích độ quan trọng của các đặc trưng đã lượng hóa các phát hiện từ tổng quan tài liệu. Kết quả cho thấy:

- Các yếu tố thuộc về tình trạng kinh tế-xã hội (SES), đặc biệt là biến `lunch` (34.2%), là yếu tố dự báo mạnh mẽ nhất
- Trình độ học vấn của phụ huynh (21.5%) đóng vai trò quan trọng thứ hai
- Việc tham gia khóa luyện thi (18.9%) có tác động đáng kể
- Các yếu tố nhân khẩu học như giới tính (1.1%) và chủng tộc (1.9%) có ảnh hưởng tương đối nhỏ

Thứ ba, giá trị $R^2 = 0.26$ tương đối thấp không phải là một thất bại, mà là một phát hiện khoa học có giá trị. Nó cho thấy rằng chỉ riêng các yếu tố nhân khẩu học và kinh tế-xã hội là không đủ để giải thích hoàn toàn kết quả học tập. Còn nhiều yếu tố quan trọng khác như động lực học tập, kỹ năng tự học, sức khỏe tâm thần và chất lượng giảng dạy cần được nghiên cứu sâu hơn.

Hàm ý thực tiễn

Các kết quả nghiên cứu mang lại những hàm ý quan trọng cho chính sách giáo dục:

1. **Phát triển hệ thống can thiệp sớm:** Các mô hình dự đoán có thể được tích hợp vào hệ thống quản lý của nhà trường để xác định học sinh có nguy cơ và triển khai các biện pháp hỗ trợ kịp thời.
2. **Chuyển trọng tâm chính sách:** Cần tăng cường các chương trình hỗ trợ hệ sinh thái gia đình, bao gồm mở rộng chương trình bữa ăn miễn phí, hỗ trợ tài chính cho gia đình có hoàn cảnh khó khăn, và nâng cao sự tham gia của phụ huynh vào quá trình giáo dục.
3. **Thúc đẩy công bằng trong giáo dục:** Các phát hiện về tác động lớn của SES nhấn mạnh tầm quan trọng của việc giảm thiểu bất bình đẳng kinh tế-xã hội để đảm bảo cơ hội giáo dục công bằng cho tất cả học sinh.

Hạn chế và hướng phát triển

Nghiên cứu cũng có một số hạn chế cần được thừa nhận:

- **Phạm vi dữ liệu:** Bộ dữ liệu chỉ bao gồm các biến nhân khẩu học và kinh tế-xã hội cơ bản, chưa bao quát các yếu tố phi nhận thức và hành vi quan trọng.
- **Khả năng khái quát hóa:** Kết quả được rút ra từ một mẫu cụ thể và có thể không hoàn toàn khái quát được cho các bối cảnh văn hóa và hệ thống giáo dục khác.
- **Mối quan hệ nhân quả:** Nghiên cứu quan sát này có thể xác định các mối tương quan mạnh mẽ nhưng không thể khẳng định chắc chắn về mối quan hệ nhân quả.

Các nghiên cứu trong tương lai nên:

- Tích hợp dữ liệu đa nguồn, bao gồm dữ liệu hành vi từ LMS và khảo sát về yếu tố tâm lý
- Áp dụng các phương pháp học máy tiên tiến hơn như deep learning
- Nghiên cứu sâu hơn về các cơ chế trung gian và điều tiết
- Khám phá các vấn đề đạo đức liên quan đến việc sử dụng mô hình dự đoán trong giáo dục

Thông điệp cuối cùng

Dự án này khẳng định rằng:

- **Các yếu tố kinh tế-xã hội có vai trò quan trọng nhưng không phải là định mệnh.** Mỗi học sinh đều có tiềm năng thành công bất kể xuất thân.
- **Can thiệp sớm và có mục tiêu là chìa khóa.** Các hệ thống dự đoán cho phép xác định và hỗ trợ học sinh cần giúp đỡ trước khi quá muộn.
- **Giáo dục là trách nhiệm của toàn xã hội.** Để giảm bất bình đẳng, cần không chỉ cải thiện trường học mà còn hỗ trợ gia đình và cộng đồng.
- **Công nghệ là công cụ, không phải giải pháp.** Học máy giúp hiểu và dự đoán tốt hơn, nhưng sự thành công của học sinh phụ thuộc vào sự quan tâm và can thiệp của con người.

Chúng tôi hy vọng nghiên cứu này góp phần xây dựng một hệ thống giáo dục công bằng hơn, chủ động hơn, dựa trên bằng chứng và lấy học sinh làm trung tâm.

Tài liệu tham khảo

Tài liệu tham khảo

[1] S.P. Scientist. *Student Performance in Exams*. Kaggle. Dataset. 2018. URL: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams (**urlseen** 01/11/2025).

[2] Joseph Chan. *StudentPerformanceRegressor (josephchan524)*. Notebook. 2022. URL: https://www.kaggle.com/code/josephchan524/studentperformanceregressor-rmse-12-26-r2-0-26 (**urlseen** 01/11/2025).

[3] XGBoost Contributors. *XGBoost Python API Reference*. URL: https://xgboost.readthedocs.io/en/stable/python/python_api.html (**urlseen** 01/11/2025).

[4] Scikit-learn developers. *sklearn.linear_model.LinearRegression*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html (**urlseen** 01/11/2025).

[5] Alaa Albuali **andothers**. "Predicting academic performance for students' university: case study from Saint Cloud State University". **in***BMC Medical Education*: 23 (2023), **page** 643. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC12453804/ (**urlseen** 01/11/2025).

[6] Hamza Waheed, Saif-Ur-Rehman Hassan **andothers**. "Predicting Academic Performance: A Systematic Literature Review". **in***Applied Sciences*: 10.6 (2020), **page** 2042. URL: https://research.monash.edu/files/273454703/264008976_oa.pdf (**urlseen** 01/11/2025).

[7] V. O. Oladokun **andothers**. "Prediction Methods on Students' Academic Performance: A Review". **in***International Journal of Education and Development using Information and Communication Technology*: 18.3 (2022), **pages** 5–28. URL: https://www.researchgate.net/publication/364334434 (**urlseen** 01/11/2025).

[8] Johannes Berens, Kirsten Schneider **andothers**. "Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review". **in***Applied Sciences*: 10.3 (2019), **page** 1042. URL: https://www.mdpi.com/2076-3417/10/3/1042 (**urlseen** 01/11/2025).

[9] Maryam Sabouri **andothers**. "Predicting Student Academic Performance: A Machine Learning Approach and Feature Analysis". **in***Interdisciplinary Journal of Management Studies*: 16.4 (2023), **pages** 863–878. URL: https://ijms.ut.ac.ir/article_100666.html (**urlseen** 01/11/2025).

[10] Onyekachi W. Adejo **and** Thomas Connolly. "Predicting students' academic performance using machine learning techniques: a literature review". **in***Journal of Educational Technology & Society*: 24.3 (2021), **pages** 192–206. URL: https://www.researchgate.net/publication/361096291 (**urlseen** 01/11/2025).

[11] Adam Parnell **andothers**. "Predicting Academic Success of College Students Using Machine Learning Techniques". **in***Data*: 9.4 (2024), **page** 60. URL: https://www.mdpi.com/2306-5729/9/4/60 (**urlseen** 01/11/2025).

[12] American Psychological Association. *Education and Socioeconomic Status Factsheet*. 2017. URL: https://www.apa.org/pi/ses/resources/publications/education (**urlseen** 01/11/2025).

[13] Gwendolyn M. Lawson **andothers**. "Annual Research Review: Associations of socioeconomic status with cognitive development in childhood and adolescence". **in***Journal of Child Psychology and Psychiatry*: 65.4 (2024), **pages** 465–486. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11920614/ (**urlseen** 01/11/2025).

[14] Kimberly Van Der Lee **andothers**. "The socio-economic rank of parents and students' academic and cognitive outcomes: Examining the physical, psychological and social mediators". **in***Frontiers in Education*: 7 (2022), **page** 938078. URL: https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2022.938078/full (**urlseen** 01/11/2025).

[15] OECD. *Student socio-economic status*. 2023. URL: https://www.oecd.org/en/topics/student-socio-economic-status.html (**urlseen** 01/11/2025).

[16] Eric F. Dubow, Paul Boxer **and** L. Rowell Huesmann. "Long-term Effects of Parents' Education on Children's Educational and Occupational Success". **in***Merrill-Palmer Quarterly*: 55.3 (2009), **pages** 224–249. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC2853053/ (**urlseen** 01/11/2025).

[17] D. S. Fauziya **andothers**. "Examining the Impact of Parents' Education on Students' Academic Achievements". **in***International Journal of Multidisciplinary Research and Analysis*: 7.5 (2024), **pages** 2053–2063. URL: https://www.researchgate.net/publication/380806514 (**urlseen** 01/11/2025).

[18] Isaac Boniface Mboya **and** Beatus Simon John. "Parental Level of Education and its Implications on their Expectations towards their Children Academic Performance". **in***International Journal of Pedagogical Counseling and Education*: 1.1 (2023), **pages** 1–12. URL: https://gprjournals.org/journals/index.php/IJPCE/article/view/109 (**urlseen** 01/11/2025).

[19] Melissa Ramirez. "The Impact of Parental Involvement on Academic Achievement and Self-Concept of Elementary School Students". California State University, Monterey Bay, 2012. URL: https://digitalcommons.csumb.edu/caps_thes_all/343/ (**urlseen** 01/11/2025).

[20] Nancy E. Hill **and** Diana F. Tyson. "Parent involvement and student academic performance: A multiple mediational analysis". **in***Journal of Educational Psychology*: 96.4 (2004), **pages** 740–756. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC3020099/ (**urlseen** 01/11/2025).

[21] The Annie E. Casey Foundation. *The Role of Parental Involvement in Your Child's Education.* 2023. URL: https://www.aecf.org/blog/parental-involvement-is-key-to-student-success-research-shows (**urlseen** 01/11/2025).

[22] University of Washington Health Sciences and Population Health. *Universal Free School Meals: A key ingredient in improving childhood health outcomes.* 2024. URL: https://hspop.uw.edu/universal-free-school-meals-improve-health-outcomes/ (**urlseen** 01/11/2025).

[23] Jehanzeb R. Cheema **and** Gabriel Galluzzo. "Within-School Gender Gaps in Reading, Mathematics, and Science Literacy". **in***Comparative Education Review*: 52.3 (2008), **pages** 437–

458. URL: https://www.journals.uchicago.edu/doi/10.1086/588762 (**urlseen** 01/11/2025).

[24] David Reilly, David L. Neumann **and** Glenda Andrews. "Gender Differences in Reading and Writing Achievement: Evidence From the National Assessment of Educational Progress (NAEP)". **in***American Psychologist*: 74.4 (2019), **pages** 445–458. URL: https://www.apa.org/pubs/journals/releases/amp-amp0000356.pdf (**urlseen** 01/11/2025).

[25] Azeta Shehu **andothers**. "Relationship between gender and academic performance of reading, writing and literature, mathematics and science". **in***Polis Journal*: 21 (2023), **pages** 23–38. URL: https://uet.edu.al/polis/wp-content/uploads/2023/02/Relationship-between-gender.pdf (**urlseen** 01/11/2025).

[26] Sara M. Lindberg, Janet Shibley Hyde **andothers**. "New Trends in Gender and Mathematics Performance: A Meta-Analysis". **in***Psychological Bulletin*: 136.6 (2010), **pages** 1123–1135. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC3057475/ (**urlseen** 01/11/2025).

[27] Imam Kusmaryono **andothers**. "The influence of gender disparities on high school students' mathematics performance". **in***Journal of Research in Mathematics Education*: 8.1 (2023), **pages** 67–82. URL: https://journals2.ums.ac.id/index.php/jramathedu/article/view/7052/3556 (**urlseen** 01/11/2025).

[28] Rajesh Kumar **andothers**. "Role of Socioeconomic Factors in Educational Disparities: A Comprehensive Analysis". **in***International Journal of Recent Scientific Research*: 4.12 (2023), **pages** 4256–4262. URL: https://ijrpr.com/uploads/V4ISSUE12/IJRPR20796.pdf (**urlseen** 01/11/2025).

[29] Arto Hellas **andothers**. "A Systematic Review Regarding the Prediction of Academic Performance". **in***Journal of Computer Science and Software Engineering*: 18.5 (2022), **pages** 1219–1231. URL: https://thescipub.com/abstract/jcssp.2022.1219.1231 (**urlseen** 01/11/2025).

[30] Mehmet Sahin **and** Dirk Ifenthaler. "Comparative Analysis of Machine Learning Models for Predicting Student Success in Online Programming Courses: A Study Based on LMS Data and External Factors". **in***Mathematics*: 12.20 (2024), **page** 3272. URL: https://www.mdpi.com/2227-7390/12/20/3272 (**urlseen** 01/11/2025).

[31] Sarra Mouelhi **andothers**. "Comparative Analysis of Machine Learning Based Predictive Models of Student Success". **in***CEUR Workshop Proceedings*: **volume** 3938. 2023, Paper_9. URL: https://ceur-ws.org/Vol-3938/Paper_9.pdf (**urlseen** 01/11/2025).

[32] Emmanuel Okewu **andothers**. "A Machine Learning Approach in Predicting Student's Academic Performance Using Artificial Neural Network". **in***Journal of Computational and Cognitive Engineering*: 1.1 (2021), **pages** 1–8. URL: https://ojs.bonviewpress.com/index.php/JCCE/article/view/470 (**urlseen** 01/11/2025).

Phụ lục A

Phụ lục A: Code và Thuật toán

A.1 Tiền xử lý dữ liệu

A.1.1 Code thực tế: Tách Features và Target

```
# Step 1: Separate Features (X) and Target (y)
y = df['math score']
X = df.drop(['math score', 'reading score', 'writing score'], axis=1)

print("Data shape after separation:")
print(f"X (features): {X.shape}")
print(f"y (target): {y.shape}")
print(f"\nFeature list:")
for i, col in enumerate(X.columns, 1):
    print(f"  {i}. {col}")
```

A.1.2 Code thực tế: One-Hot Encoding

```
# Step 2: Apply One-Hot Encoding
print("BEFORE ENCODING:")
display(X.head())

# Apply One-Hot Encoding with drop_first=True
```

```
X_encoded = pd.get_dummies(X, drop_first=True)

print("\nAFTER ENCODING:")
print(f"Number of features before: {X.shape[1]}")
print(f"Number of features after: {X_encoded.shape[1]}")
print(f"\nFeature list after encoding:")
for i, col in enumerate(X_encoded.columns, 1):
    print(f"  {i}. {col}")

display(X_encoded.head())

print("\nExplanation:")
print("- Creates binary variables (0/1) for each category")
print("- Parameter drop_first=True removes one column to avoid")
print("  multicollinearity")
print("- Example: 'gender_male' = 1 (male), = 0 (female)")
```

A.1.3 Code thực tế: Train-Test Split

```
# Data Split: 80% train, 20% test
X_train, X_test, y_train, y_test = train_test_split(
    X_encoded,
    y,
    test_size=0.2,
    random_state=42,
    shuffle=True
)

print("Data Split Results:")
print(f"Training Set:")
print(f"  - X_train: {X_train.shape} (80% of data)")
print(f"  - y_train: {y_train.shape}")
```

```
print(f"\nTest Set:")
print(f"  - X_test: {X_test.shape} (20% of data)")
print(f"  - y_test: {y_test.shape}")

print("\nParameters Used:")
print("  - test_size=0.2: Use 20% of data for testing")
print("  - random_state=42: Ensure reproducible results")
print("  - shuffle=True: Randomize data before splitting")
```

A.2 Hàm đánh giá mô hình

A.2.1 Code thực tế: Hàm evaluate_model()

```
def evaluate_model(y_true, y_pred, model_name):
    """
    Function to evaluate regression model performance.

    Parameters:
    -----------
    y_true : array-like
        Actual values of target variable
    y_pred : array-like
        Predicted values from model
    model_name : str
        Name of model (for display)

    Returns:
    -----------
    dict : Dictionary containing evaluation metrics
        - 'RMSE': Root Mean Squared Error
        - 'MAE': Mean Absolute Error
        - 'R2': R² Score (Coefficient of Determination)
```

```
    """

    # Calculate metrics
    rmse = np.sqrt(mean_squared_error(y_true, y_pred))
    mae = mean_absolute_error(y_true, y_pred)
    r2 = r2_score(y_true, y_pred)

    # Print results
    print(f"\nEvaluation Results: {model_name}")
    print("=" * 100)
    print(f"RMSE: {rmse:.4f}")
    print(f"MAE:  {mae:.4f}")
    print(f"R² Score: {r2:.4f} ({r2*100:.2f}%)")
    print("=" * 100)

    # Return as dictionary
    return {
        'RMSE': rmse,
        'MAE': mae,
        'R2': r2
    }
```

A.2.2 Code thực tế: Linear Regression

```
# Train Linear Regression model
print("Training Linear Regression model...")

# Initialize and train
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)

# Make predictions on test set
```

```
y_pred_lr = lr_model.predict(X_test)

print("Training completed.")

# Evaluate model
lr_metrics = evaluate_model(y_test, y_pred_lr,
                            "Linear Regression (Baseline)")
```

A.2.3 Code thực tế: XGBoost

```
# Train XGBoost Regression model
print("Initializing and training XGBoost model...")
print("=" * 80)

# Initialize model with hyperparameters
xgb_model = XGBRegressor(
    objective='reg:squarederror',
    n_estimators=100,
    max_depth=5,
    learning_rate=0.1,
    subsample=0.8,
    colsample_bytree=0.8,
    random_state=42,
    n_jobs=-1,
    verbosity=0
)

# Train model on training set
xgb_model.fit(X_train, y_train)

# Make predictions on test set
y_pred_xgb = xgb_model.predict(X_test)
```

```
print("  XGBoost model training completed")
print(f"  - Number of trees: {xgb_model.n_estimators}")
print(f"  - Maximum tree depth: {xgb_model.max_depth}")
print(f"  - Learning rate: {xgb_model.learning_rate}")

# Evaluate model
xgb_metrics = evaluate_model(y_test, y_pred_xgb, "XGBoost")

print("\nHyperparameter Explanations:")
print("=" * 80)
print("1. n_estimators = 100:")
print("   Build 100 decision trees sequentially")
print("2. max_depth = 5:")
print("   Each tree has max 5 levels - prevents overfitting")
print("3. learning_rate = 0.1:")
print("   Each tree contributes 10% to prediction")
print("4. subsample = 0.8:")
print("   Each tree uses 80% of data - increases diversity")
print("5. colsample_bytree = 0.8:")
print("   Each tree uses 80% of features")
```

A.3 Phân tích Feature Importance

A.3.1 Code thực tế: Trích xuất và Hiển thị Độ Quan Trọng

```
import matplotlib.pyplot as plt

# Extract feature importance from XGBoost model
importance = xgb_model.feature_importances_
feature_names = X.columns
```

```
# Create DataFrame for better organization
importance_df = pd.DataFrame({
    'feature': feature_names,
    'importance': importance
}).sort_values('importance', ascending=False)

# Get top 5 features
top_5 = importance_df.head(5)

# Print results
print("Top 5 Most Important Features:")
print("=" * 80)
for idx, row in top_5.iterrows():
    percentage = (row['importance'] / importance_df['importance'].sum()) * 100
    print(f"{row['feature']:30s} {row['importance']:8.4f} ({percentage:5.1f}%)")

# Create horizontal bar chart
plt.figure(figsize=(10, 6))
plt.barh(importance_df['feature'], importance_df['importance'],
         color='steelblue', edgecolor='navy')
plt.xlabel('Importance Score', fontsize=12, fontweight='bold')
plt.title('Feature Importance in XGBoost Model', fontsize=14,
          fontweight='bold')
plt.gca().invert_yaxis()
plt.tight_layout()
plt.show()
```

Phụ lục B

Phụ lục B: Bảng số liệu chi tiết

B.1 Thống kê mô tả đầy đủ

Bảng B.1. Thống kê mô tả đầy đủ cho tất cả các biến số

Biến	Mean	Std	Min	Max
Math Score	66.09	15.16	0.00	100.00
Reading Score	69.17	14.60	17.00	100.00
Writing Score	68.05	15.20	10.00	100.00

B.2 Phân phối theo các biến phân loại

Bảng B.2. Phân phối tần số các biến phân loại

Biến	Giá trị	Tần số (%)
Gender	Female	518 (51.8%)
	Male	482 (48.2%)
Lunch	Free/Reduced	355 (35.5%)
	Standard	645 (64.5%)
Test Prep	None	642 (64.2%)
	Completed	358 (35.8%)