

**TRƯỜNG KỸ THUẬT VÀ CÔNG NGHỆ
KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO MÔN HỌC
KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI:
PHÂN TÍCH VÀ PHÂN CỤM HIỆU SUẤT HỌC TẬP
CỦA SINH VIÊN**

Giáo viên hướng dẫn: NGUYỄN THÁI TOÀN

Sinh viên thực hiện:

110122214	NGUYỄN PHÚC AN	DA22TTA
110122030	NGUYỄN THIÊN ÂN	DA22TTA
110122046	HỨA KHÁNH ĐĂNG	DA22TTA
110122051	LÂM THANH ĐỈNH	DA22TTA
110122016	PHẠM HỮU LUÂN	DA22TTA

Trà Vinh, tháng 6 năm 2025

LỜI MỞ ĐẦU

Trong bối cảnh giáo dục hiện đại, việc ứng dụng các kỹ thuật phân tích dữ liệu để hiểu rõ hơn về quá trình học tập và kết quả học tập của sinh viên đang ngày càng trở nên quan trọng. Khai phá dữ liệu (Data Mining) không chỉ giúp khám phá ra các mô hình ẩn trong dữ liệu mà còn hỗ trợ các nhà quản lý giáo dục đưa ra các quyết định phù hợp nhằm nâng cao chất lượng đào tạo.

Đề tài "Phân tích và phân cụm hiệu suất học tập của sinh viên" được thực hiện với mục tiêu áp dụng các kỹ thuật khai phá dữ liệu – cụ thể là phương pháp phân cụm – để nhận diện các nhóm sinh viên có đặc điểm học tập tương đồng. Thông qua đó, có thể phát hiện những nhóm sinh viên đang gặp khó khăn trong học tập, những nhóm có tiềm năng nổi bật, từ đó đề xuất các chính sách hỗ trợ hoặc phát triển phù hợp cho từng nhóm.

Báo cáo sẽ trình bày chi tiết quá trình xử lý dữ liệu, lựa chọn thuộc tính, áp dụng mô hình phân cụm, trực quan hóa kết quả cũng như phân tích và đánh giá các cụm sinh viên thu được. Hy vọng rằng kết quả của đề tài sẽ góp phần minh họa rõ nét cho giá trị thực tiễn của khai phá dữ liệu trong lĩnh vực giáo dục.

LỜI CẢM ƠN

Lời đầu tiên, em xin trân trọng cảm ơn giảng viên Nguyễn Thái Toàn - người đã trực tiếp chỉ bảo, hướng dẫn nhóm em trong quá trình hoàn thành bài đồ án môn học này.

Nhóm em cũng xin được gửi lời cảm ơn đến quý thầy, cô giáo trường Kỹ thuật và Công nghệ, đặc biệt là các thầy, cô khoa Công nghệ Thông tin - những người đã truyền lửa và giảng dạy kiến thức cho em suốt thời gian qua.

Mặc dù đã có những đầu tư nhất định trong quá trình làm bài song cũng khó có thể tránh khỏi những sai sót, chúng em kính mong nhận được ý kiến đóng góp của quý thầy cô để bài báo cáo được hoàn thiện hơn.

Nhóm chúng em xin chân thành cảm ơn!

Trà Vinh, ngày ... tháng 6 năm 2025

Nhóm thực hiện:

Phạm Hữu Luân - 110122016

Nguyễn Phúc An – 110122033

Nguyễn Thiên Ân - 110122033

Lâm Thanh Đình – 110122051

Hứa Khánh Đăng - 110122046

Sinh viên ký và ghi rõ họ và tên

Sinh viên 1

Sinh viên 2

Sinh viên 3

Sinh viên 4

Sinh viên 5

[illegible]

[illegible]

MỤC LỤC

CHƯƠNG 1: TỔNG QUAN	14
1.1 Lý do chọn đề tài	14
1.2 Đối tượng nghiên cứu	14
1.3 Nội dung nghiên cứu	14
1.4 Phương pháp nghiên cứu	14
1.5 Phạm vi nghiên cứu	15
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT	16
2.1 Giới thiệu ngôn ngữ Python	16
2.2 Giới thiệu về khai phá dữ liệu	16
2.3 Phân cụm dữ liệu	16
2.4 Thuật toán phân cụm K-Means	17
2.4.1 Quy trình thuật toán	17
2.4.2 Các kỹ thuật hỗ trợ đánh giá phân cụm	18
2.4.3 Ưu và nhược điểm của K-Means	19
2.5 Công cụ hỗ trợ khai phá dữ liệu	20
CHƯƠNG 3: ỨNG DỤNG QUY TRÌNH CRISP-DM	21
3.1 Business Understanding	21
3.2 Data Understanding	21
3.2.1 Giới thiệu bộ dữ liệu	21
3.2.2 Số lượng thuộc tính và mô tả	22
3.2.3 Đặc điểm dữ liệu và khám phá sơ bộ	24
3.2.4 Đánh giá chất lượng dữ liệu	24
3.3 Data Preparation	25
3.3.1 Tích hợp dữ liệu	25
3.3.2 Tiền xử lý dữ liệu điểm số	26

3.4	Modeling.....	26
3.4.1	Chọn biến đầu vào	26
3.4.2	Chuẩn hóa và biến đổi dữ liệu	27
3.4.3	Giảm chiều dữ liệu bằng PCA	28
3.4.4	Áp dụng thuật toán K-Means.....	28
3.4.5	Phân tích số cụm tối ưu.....	29
3.4.6	Trực quan hóa và đánh giá sơ bộ kết quả phân cụm.....	31
3.5	Evaluation	32
3.5.1	Đánh giá chất lượng phân cụm	32
3.5.2	Số lượng học sinh trong mỗi cụm.....	33
3.5.3	Phân tích đặc trưng của từng cụm	33
3.6	Deployment	35
3.6.1	Triển khai giao diện với thư viện Streamlit.....	35
3.6.2	Quy trình triển khai trên Streamlit Community Cloud:.....	36
CHƯƠNG 4: KẾT QUẢ THỰC HIỆN		37
4.1	Tổng quan về dữ liệu sau tiền xử lý	37
4.2	Kết quả phân cụm với K-Means.....	39
4.2.1	Lựa chọn số cụm tối ưu	39
4.2.2	Kết quả phân cụm	44
4.3	Phân tích đặc trưng của từng cụm	48
4.3.1	Biến định lượng	48
4.3.2	Biến định tính.....	49
4.3.3	Các đặc trưng quan trọng.....	50
4.4	Khám phá đặc trưng cụm.....	54
4.4.1	Cụm 0.....	54
4.4.2	Cụm 1.....	55

4.4.3	Cụm 2.....	56
4.4.4	Cụm 3.....	57
4.4.5	Kết luận việc khám phá đặc trưng cụm	58
4.5	Deploy ứng dụng trên Streamlit Community Cloud	60
CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN.....		62
5.1	Kết quả đạt được.....	62
5.2	Ưu điểm	62
5.3	Hạn chế	63
5.4	Hướng phát triển.....	64
TÀI LIỆU THAM KHẢO.....		65

DANH MỤC HÌNH ẢNH

Hình 2.1 Biểu đồ elbow method.....	18
Hình 2.2 Biểu đồ silhouette score.....	19
Hình 4.1 Tổng quan dữ liệu sau làm bước tiền xử lý	37
Hình 4.2 Một số bản ghi là outlier.....	37
Hình 4.3 Mô tả dữ liệu sau khi đã làm sạch	38
Hình 4.4 Phân bố G3 trước tiền xử lý	38
Hình 4.5 Phân bố G3 sau làm sạch dữ liệu.....	39
Hình 4.6 Heatmap biến định lượng	40
Hình 4.7 Heatmap top 10 biến định tính đã chuẩn hóa	41
Hình 4.8 Giao diện cho các biến đầu vào.....	42
Hình 4.9 Biểu đồ phương pháp Elbow sau khi chọn biến đầu vào	42
Hình 4.10 Biểu đồ Silhouette score sau khi chọn biến đầu vào	43
Hình 4.11 Biểu đồ cột thể hiện số lượng mỗi cụm.....	44
Hình 4.12 Biểu đồ PCA 2D với 2 cụm.....	45
Hình 4.13 Biểu đồ PCA 2D với 4 cụm.....	45
Hình 4.14 Biểu đồ PCA 3D với 4 cụm.....	46
Hình 4.15 Crosstab giữa G3_level và Cluster	47
Hình 4.16 Trung bình các biến định lượng theo cụm.....	48
Hình 4.17 Boxplot của studytime cho từng cụm.....	48
Hình 4.18 Tỷ lệ phần trăm school theo cụm	49
Hình 4.19 Countplot của higher theo cụm.....	50
Hình 4.20 Biểu đồ cột top 10 đặc trưng ảnh hưởng nhất	51
Hình 4.21 Bảng thống kê mô tả các biến định lượng theo cụm	51
Hình 4.22 Heatmap trung bình các biến định lượng theo cụm	52

Hình 4.23 Bảng tổng hợp giá trị trung bình biến định lượng cụm 0	54
Hình 4.24 Bảng tổng hợp tỉ lệ % nổi bật các biến định tính cụm 0	54
Hình 4.25 Bảng tổng hợp giá trị trung bình biến định lượng cụm 1	55
Hình 4.26 Bảng tổng hợp tỉ lệ % nổi bật các biến định tính cụm 1	55
Hình 4.27 Bảng tổng hợp giá trị trung bình biến định lượng cụm 2	56
Hình 4.28 Bảng tổng hợp tỉ lệ % nổi bật các biến định tính cụm 2	56
Hình 4.29 Bảng tổng hợp giá trị trung bình biến định lượng cụm 3	57
Hình 4.30 Bảng tổng hợp tỉ lệ % nổi bật các biến định tính cụm 3	57
Hình 4.31 Deploy miễn phí tên miền trên Streamlit Community Cloud	61

DANH MỤC BẢNG BIỂU

Bảng 1. Mô tả các thuộc tính thông tin cá nhân và gia đình	22
Bảng 2. Mô tả các thuộc tính thông tin học tập và định hướng	23
Bảng 3. Mô tả các thuộc tính về hoạt động và hoàn cảnh cá nhân.....	23
Bảng 4. Mô tả các thuộc tính về chất lượng cuộc sống và hành vi xã hội	24
Bảng 5. Mô tả thuộc tính về kết quả học tập	24
Bảng 6. Nhận xét các cụm sau khi quan sát PCA 2D 4 cụm.....	46
Bảng 7 Nhận xét crosstab giữa G3_level và Cluster.....	47
Bảng 8. Bảng phân tích các đặc trưng quan trọng.....	52
Bảng 9. Bảng định hướng can thiệp cho từng cụm	60

DANH MỤC KÍ HIỆU CÁC CỤM TỪ VIẾT TẮT

Số thứ tự	Kí hiệu	Nội dung viết tắt
1	CRISP-DM	Cross-industry standard process for data mining
2	WCSS	Within-Cluster Sum of Squares
3	SSE	Sum of Squared Errors

BẢNG PHÂN CÔNG CÔNG VIỆC

Người thực hiện	Công việc
Lâm Thanh Đình	Tìm kiếm bộ dữ liệu, tìm hiểu yêu cầu đề tài, làm slide thuyết trình
Hứa Khánh Đăng	Thu thập, khám phá và đánh giá dữ liệu ban đầu, làm slide thuyết trình
Nguyễn Phúc An	Làm sạch, chọn lọc, tạo biến mới để sẵn sàng cho mô hình, viết báo cáo
Phạm Hữu Luân	Áp dụng thuật toán K-Means để huấn luyện mô hình, đánh giá mô hình, viết báo cáo
Nguyễn Thiên Ân	Tích hợp giao diện streamlit ,deploy ứng dụng lên Streamlit Community Cloud

CHƯƠNG 1: TỔNG QUAN

1.1 Lý do chọn đề tài

Giáo dục cá nhân hóa ngày càng trở thành xu hướng trong giáo dục hiện đại. Việc hiểu rõ đặc điểm và hành vi học tập của sinh viên có thể giúp nhà trường đề xuất các chính sách phù hợp, giúp sinh viên phát huy tối đa năng lực. Phân cụm dữ liệu học tập là một trong những kỹ thuật hữu ích để hỗ trợ mục tiêu này. Nhóm chọn đề tài này nhằm tìm hiểu và áp dụng các thuật toán khai phá dữ liệu vào thực tế học tập.

1.2 Đối tượng nghiên cứu

Đề tài này tập trung nghiên cứu về dữ liệu học tập của sinh viên trong bộ dữ liệu “Student Performance Dataset” từ UCI Machine Learning Repository. Dữ liệu bao gồm các thuộc tính liên quan đến học tập như: điểm số, thời gian học, tình trạng gia đình, hỗ trợ học tập.

1.3 Nội dung nghiên cứu

Nội dung nghiên cứu của đề tài bao gồm các bước: tìm hiểu và xử lý bộ dữ liệu sinh viên; lựa chọn các đặc trưng học tập phù hợp để đưa vào mô hình; áp dụng các thuật toán phân cụm là K-Means để nhóm sinh viên theo hiệu suất học tập; trực quan hóa và phân tích từng cụm nhằm khám phá các đặc điểm nổi bật; cuối cùng là đánh giá, so sánh và đưa ra nhận xét về hiệu quả của từng phương pháp.

1.4 Phương pháp nghiên cứu

Đề tài áp dụng quy trình Cross-Industry Standard Process for Data Mining (CRISP-DM) với các bước chính như sau:

- Hiểu biết nghiệp vụ: Xác định mục tiêu phân cụm sinh viên theo hiệu suất học tập.
- Hiểu biết dữ liệu: Tìm hiểu bộ dữ liệu và các đặc trưng liên quan đến học tập.
- Chuẩn bị dữ liệu: Làm sạch, xử lý giá trị thiếu, mã hóa biến định tính, chuẩn hóa dữ liệu.
- Mô hình hóa: Áp dụng các thuật toán phân cụm như K-Means.
- Đánh giá: Dùng các chỉ số như Silhouette Score, trực quan hóa và phân tích cụm.
- Triển khai: Tổng hợp kết quả và rút ra các đề xuất ứng dụng trong giáo dục.

1.5 Phạm vi nghiên cứu

Phạm vi nghiên cứu tập trung vào việc phân tích hiệu suất học tập của sinh viên dựa trên khoảng 30 thuộc tính trong bộ dữ liệu “Student Performance Dataset” từ UCI bằng mô hình phân cụm K-Means, chủ yếu liên quan đến đặc điểm cá nhân, môi trường học tập và kết quả học tập cuối kỳ. Đề tài không mở rộng đến các yếu tố bên ngoài như sức khỏe tâm lý hay tác động xã hội chưa được thể hiện trong dữ liệu.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1 Giới thiệu ngôn ngữ Python

Python là một ngôn ngữ lập trình bậc cao được sử dụng rộng rãi trong các ứng dụng web, phát triển phần mềm, khoa học dữ liệu và máy học. Các nhà phát triển sử dụng Python vì nó hiệu quả, dễ học và có thể chạy trên nhiều nền tảng khác nhau. Phần mềm Python được tải xuống miễn phí, tích hợp tốt với tất cả các loại hệ thống và tăng tốc độ phát triển.

Guido van Rossum bắt đầu nghiên cứu Python vào cuối những năm 1980 với tư cách là ngôn ngữ kế thừa cho ngôn ngữ lập trình ABC và phát hành nó lần đầu tiên vào năm 1991 với tên gọi Python 0.9.0. Phiên bản mới nhất hiện nay của Python là version 3.13.1, nó tập trung cải thiện hiệu suất và thêm tính năng mới trong thư viện chuẩn.

2.2 Giới thiệu về khai phá dữ liệu

Khai phá dữ liệu (Data Mining) là một lĩnh vực trong khoa học máy tính, nhằm mục đích trích xuất thông tin có ý nghĩa từ một lượng lớn dữ liệu. Đây là bước quan trọng trong quá trình phát hiện tri thức từ dữ liệu (Knowledge Discovery in Databases). Khai phá dữ liệu không chỉ là tìm ra thông tin có sẵn, mà còn khám phá ra các mẫu, xu hướng hay quy luật ẩn chứa trong dữ liệu.

Khai phá dữ liệu được ứng dụng trong nhiều lĩnh vực như: ngân hàng cho phát hiện gian lận, thương mại điện tử cho đề xuất sản phẩm, y tế cho dự đoán bệnh, và đặc biệt trong giáo dục cho phân tích hành vi học tập của sinh viên.

2.3 Phân cụm dữ liệu

Phân cụm là một kỹ thuật học không giám sát (unsupervised learning), được sử dụng để nhóm các đối tượng dữ liệu có đặc điểm tương tự nhau vào cùng một cụm, trong khi những đối tượng khác biệt sẽ nằm ở cụm khác. Mục tiêu của phân cụm là giảm độ phức tạp trong dữ liệu và phát hiện ra cấu trúc ẩn mà không cần nhãn định trước.

Ứng dụng của phân cụm:

- Trong giáo dục: Phân nhóm sinh viên theo hiệu suất học tập để hỗ trợ cá nhân hóa giáo dục.
- Trong marketing: Phân nhóm khách hàng để đề xuất sản phẩm phù hợp.

- Trong y học: Phân nhóm bệnh nhân theo triệu chứng để chẩn đoán bệnh nhanh hơn.

Trong khai phá dữ liệu, có nhiều mô hình phân cụm được sử dụng nhằm mục đích nhóm các đối tượng có đặc điểm tương đồng lại với nhau. Mỗi thuật toán có cách tiếp cận khác nhau và phù hợp với từng loại dữ liệu cụ thể:

- **K-Means**: Là thuật toán phân cụm dựa trên trung tâm, hoạt động bằng cách chia dữ liệu thành k cụm sao cho khoảng cách từ mỗi điểm đến tâm cụm của nó là nhỏ nhất. Thuật toán này dễ triển khai và có tốc độ xử lý nhanh, tuy nhiên đòi hỏi người dùng phải xác định trước số cụm k và có thể bị ảnh hưởng bởi điểm ngoại lai.
- **Hierarchical Clustering**: Là phương pháp phân cụm phân cấp, xây dựng cây phân cấp (dendrogram) để xác định mối quan hệ giữa các điểm dữ liệu. Thuật toán này không yêu cầu chọn trước số cụm và cho phép quan sát cấu trúc phân cụm ở nhiều cấp độ khác nhau. Tuy nhiên, nó không hiệu quả với tập dữ liệu lớn.
- **DBSCAN**: Là thuật toán phân cụm dựa trên mật độ, cho phép phát hiện các cụm có hình dạng bất kỳ và loại bỏ tốt các điểm nhiễu. DBSCAN không yêu cầu xác định trước số cụm, tuy nhiên việc lựa chọn tham số (eps và minPts) phù hợp có thể ảnh hưởng đến chất lượng phân cụm.

2.4 Thuật toán phân cụm K-Means

K-Means là một trong những thuật toán phân cụm phổ biến và dễ triển khai nhất. Nó hoạt động dựa trên nguyên lý chia tập dữ liệu thành K cụm sao cho khoảng cách giữa các điểm trong cùng một cụm là nhỏ nhất và thường dùng khoảng cách Euclidean.

2.4.1 Quy trình thuật toán

- **Input**: Tập dữ liệu đầu vào gồm n đối tượng với các thuộc tính đặc trưng (đã được chuẩn hóa nếu cần) và số K cụm.
- **Output**: Một phân hoạch gồm K cụm, trong đó mỗi đối tượng được gán vào một cụm cụ thể và mỗi cụm có một tâm cụm (centroid).

Thuật toán K-Means gồm các bước sau:

Bước 1: Khởi tạo

Chọn số cụm K và khởi tạo ngẫu nhiên K tâm cụm ban đầu.

Bước 2: Gán mỗi điểm dữ liệu vào cụm có tâm cụm gần nhất dựa vào khoảng cách Euclidean hoặc Manhattan.

Bước 3: Tính lại vị trí tâm cụm mới bằng cách lấy trung bình các điểm thuộc cụm đó.

Bước 4: Kiểm tra hội tụ

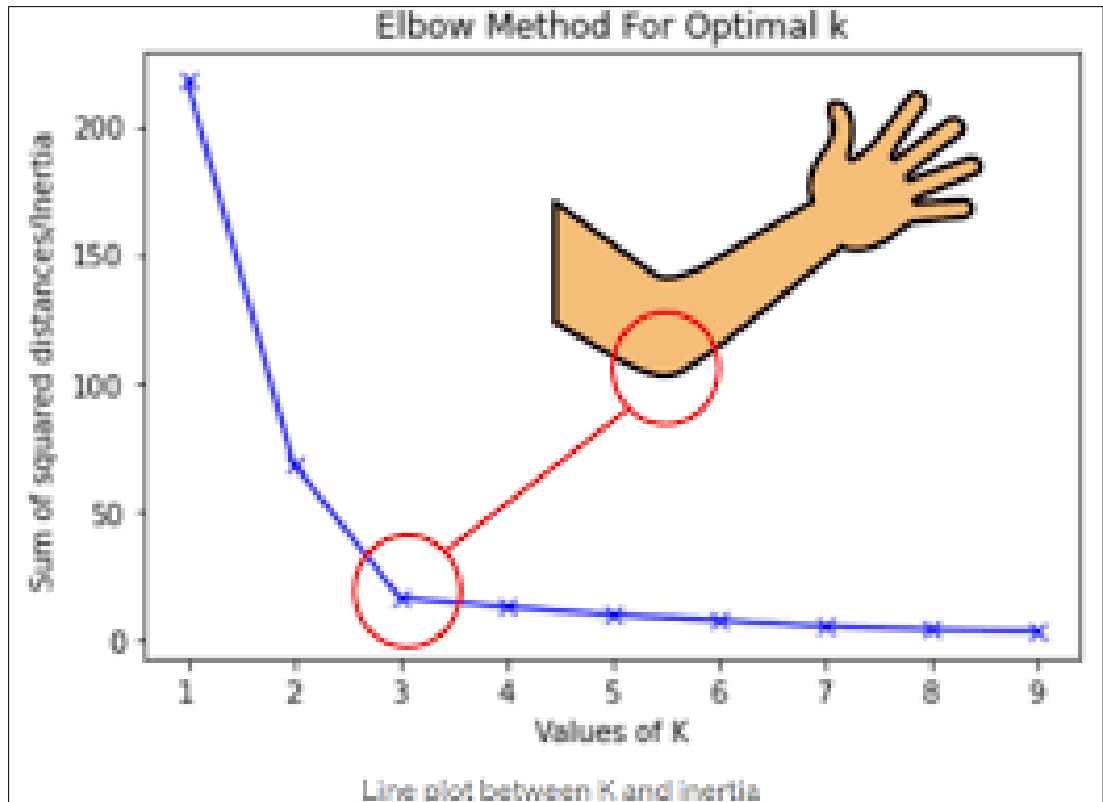
Lặp lại bước 2 và 3 cho đến khi các tâm cụm không thay đổi quá nhiều so với lần lặp trước đó hoặc đạt số vòng lặp tối đa.

2.4.2 Các kỹ thuật hỗ trợ đánh giá phân cụm

Có nhiều phương pháp để đánh giá và lựa chọn số cụm phù hợp trong thuật toán K-Means, một số kỹ thuật hỗ trợ thường được sử dụng như sau:

2.4.2.1 Phương pháp Elbow (Elbow Method)

Đây là kỹ thuật giúp xác định số cụm tối ưu k . Ta tính tổng bình phương sai số nội cụm (WCSS – Within-Cluster Sum of Squares) cho nhiều giá trị k , rồi vẽ biểu đồ đường. "Khủy tay" (elbow) của biểu đồ, nơi mà WCSS bắt đầu giảm chậm lại, chính là giá trị k hợp lý.



Hình 2.1 Biểu đồ elbow method

2.4.2.2 Chỉ số Silhouette (Silhouette Score)

Chỉ số này đo độ gắn kết nội cụm và độ tách biệt giữa các cụm.

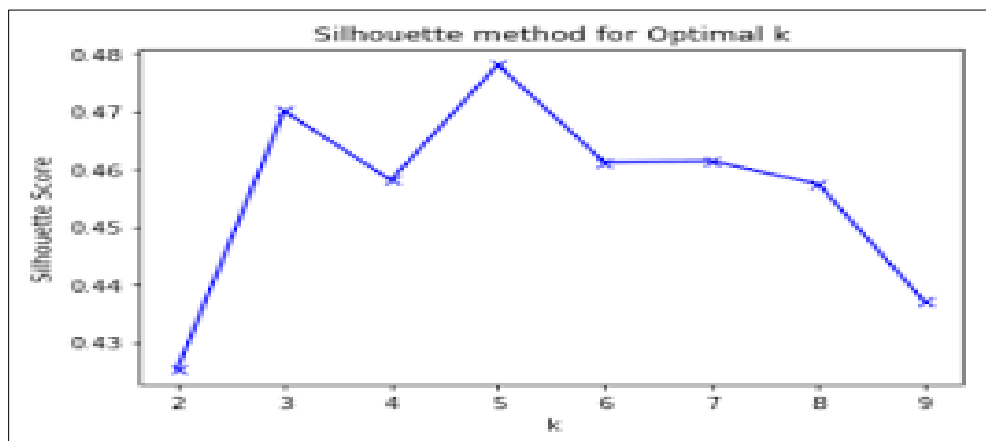
Công thức tính Silhouette Score cho một điểm dữ liệu i :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Trong đó:

- + $a(i)$: khoảng cách trung bình từ điểm i đến tất cả các điểm khác trong **cùng cụm** (độ gắn kết nội cụm).
- + $b(i)$: khoảng cách trung bình từ điểm i đến tất cả các điểm trong **cụm gần nhất khác cụm của i** (độ phân tách giữa cụm).

Giá trị trung bình của $s(i)$ trên toàn bộ dữ liệu sẽ cho ra Silhouette Score tổng thể của mô hình phân cụm.



Hình 2.2 Biểu đồ silhouette score

2.4.3 Ưu và nhược điểm của K-Means

- Ưu điểm:
 - + Hiệu quả về mặt tính toán.
 - + Dễ hiểu và dễ triển khai.
- Nhược điểm:
 - + Cần xác định trước số lượng cụm K , không hoạt động tốt với cụm không tròn.
 - + Nhạy cảm với khởi tạo, bị ảnh hưởng mạnh bởi outlier và dữ liệu nhiễu.

2.5 Công cụ hỗ trợ khai phá dữ liệu

Trong đề tài này, nhóm sử dụng ngôn ngữ lập trình Python để xử lý và phân tích dữ liệu, với quá trình làm việc được thực hiện trong môi trường Visual Studio Code. Các tập lệnh được viết dưới dạng file .py, giúp dễ quản lý mã nguồn và linh hoạt trong thao tác lập trình.

Để xây dựng giao diện trực quan hỗ trợ trình bày và thao tác với dữ liệu, nhóm sử dụng Streamlit – một thư viện Python giúp tạo nhanh ứng dụng web phân tích dữ liệu mà không cần kiến thức về frontend. Giao diện Streamlit giúp nhóm trực quan hóa kết quả phân cụm, cho phép người dùng dễ dàng tương tác và theo dõi phân tích.

Các thư viện Python được sử dụng gồm:

- Pandas: xử lý và thao tác dữ liệu dạng bảng.
- Scikit-learn: triển khai các thuật toán phân cụm K-Means.
- Matplotlib, Seaborn, Plotly: trực quan hóa dữ liệu và kết quả phân cụm trên giao diện web.

Việc kết hợp Visual Studio Code và Streamlit giúp quá trình phát triển, kiểm thử và trình bày kết quả phân tích trở nên linh hoạt, nhanh chóng và thân thiện với người dùng.

CHƯƠNG 3: ỨNG DỤNG QUY TRÌNH CRISP-DM

3.1 Business Understanding

Business Understanding là bước đầu tiên trong quy trình CRISP-DM. Mục tiêu của bước này là hiểu rõ bài toán nghiệp vụ, xác định vấn đề cần giải quyết và xác lập mục tiêu phân tích dữ liệu từ góc nhìn của người sử dụng cuối cùng hoặc tổ chức.

Mục tiêu chính của dự án này là phân tích dữ liệu học sinh để phát hiện các nhóm học sinh có hiệu suất học tập tương đồng. Điều này giúp nhà trường hiểu rõ hơn về các kiểu học sinh, từ đó có thể đưa ra các chiến lược hỗ trợ học tập hiệu quả hơn cho từng nhóm.

Đề tài “Phân tích hiệu suất học tập của sinh viên bằng các thuật toán phân cụm” hướng đến mục tiêu sử dụng thuật toán K-Means để phân nhóm sinh viên dựa trên các đặc điểm học tập như điểm số, số lần vắng học, thời gian học,... . Cụ thể:

- Áp dụng thuật toán K-Means trên bộ dữ liệu Student Performance Dataset từ UCI Machine Learning Repository.
- Phân nhóm sinh viên thành các cụm có đặc điểm hiệu suất học tập tương đồng.
- Phân tích đặc trưng của từng nhóm để rút ra các nhận xét hữu ích cho việc quản lý và hỗ trợ sinh viên.

3.2 Data Understanding

Bước này tập trung vào việc thu thập dữ liệu ban đầu, khám phá dữ liệu, xác định các vấn đề như dữ liệu thiếu, ngoại lệ, và hiểu cấu trúc dữ liệu.

3.2.1 Giới thiệu bộ dữ liệu

Trong đề tài này, nhóm nghiên cứu sử dụng bộ dữ liệu Student Performance Dataset được cung cấp bởi UCI Machine Learning Repository. Đây là một bộ dữ liệu phổ biến trong lĩnh vực khai phá dữ liệu giáo dục, chứa thông tin về kết quả học tập và các đặc điểm cá nhân, xã hội, học tập của học sinh tại Bồ Đào Nha.

Bộ dữ liệu bao gồm hai tệp dữ liệu:

- student-mat.csv: Dữ liệu môn Toán.
- student-por.csv: Dữ liệu môn Ngữ văn (Portuguese).

Tùy thuộc vào mục tiêu nghiên cứu, người dùng có thể chọn một trong hai hoặc gộp cả hai để phân tích. Trong khuôn khổ đề tài này, nhóm chọn cách gộp cả hai tập tin lại thành một tập tin lớn hơn để phân tích.

3.2.2 Số lượng thuộc tính và mô tả

Tổng cộng có 33 thuộc tính, có thể chia thành các nhóm sau:

- Nhóm 1: Thông tin cá nhân và gia đình

Bảng 1. Mô tả các thuộc tính thông tin cá nhân và gia đình

STT	Thuộc tính	Mô tả
1	school	Trường học của học sinh (nhị phân: "GP" - Gabriel Pereira hoặc "MS" - Mousinho da Silveira)
2	sex	Giới tính học sinh (nhị phân: "F" - nữ, "M" - nam)
3	age	Tuổi học sinh (số, từ 15 đến 22)
4	address	Loại địa chỉ nhà ở (nhị phân: "U" - thành thị, "R" - nông thôn)
5	famsize	Quy mô gia đình (nhị phân: "LE3" - nhỏ hơn hoặc bằng 3 người, "GT3" - lớn hơn 3 người)
6	Pstatus	Tình trạng sống chung của cha mẹ (nhị phân: "T" - sống cùng nhau, "A" - sống riêng)
7	Medu	Trình độ học vấn của mẹ (0: không học; 1: tiểu học (lớp 4); 2: lớp 5–9; 3: THPT; 4: đại học trở lên)
8	Fedu	Trình độ học vấn của cha (tương tự như trên)
9	Mjob	Nghề nghiệp của mẹ ("teacher", "health" - y tế, "services" - dịch vụ công, "at_home" - nội trợ, "other")
10	Fjob	Nghề nghiệp của cha (tương tự như trên)
11	guardian	Người giám hộ (danh định: "mother" - mẹ, "father" - cha, "other" - người khác)

- Nhóm 2: Thông tin học tập và định hướng

Bảng 2. Mô tả các thuộc tính thông tin học tập và định hướng

STT	Thuộc tính	Mô tả
1	reason	Lý do chọn trường ("home" - gần nhà, "reputation" - danh tiếng, "course" - thích ngành học, "other")
2	studytime	Thời gian học hàng tuần (1: < 2 giờ; 2: 2–5 giờ; 3: 5–10 giờ; 4: > 10 giờ)
3	traveltime	Thời gian đi học từ nhà (1: < 15 phút; 2: 15–30 phút; 3: 30–60 phút; 4: > 1 giờ)
4	failures	Số lần trượt môn trước đó (1 đến 3, nếu nhiều hơn thì ghi là 4)
5	schoolsup	Hỗ trợ học tập thêm từ trường hay không hay không (yes/no)
6	famsup	Hỗ trợ học tập từ gia đình hay không hay không (yes/no)
7	paid	Có tham gia lớp học thêm trả phí hay không (yes/no)
8	higher	Mong muốn học lên tiếp đại học hay không (yes/no)

- Nhóm 3: Hoạt động và hoàn cảnh cá nhân

Bảng 3. Mô tả các thuộc tính về hoạt động và hoàn cảnh cá nhân

STT	Thuộc tính	Mô tả
1	activities	Tham gia hoạt động ngoại khóa hay không (yes/no)
2	nursery	Có học mẫu giáo hay không hay không (yes/no)
3	internet	Có truy cập Internet tại nhà hay không (yes/no)
4	romantic	Có đang trong mối quan hệ tình cảm hay không (yes/no)

- Nhóm 4: Chất lượng cuộc sống và hành vi xã hội

Bảng 4. Mô tả các thuộc tính về chất lượng cuộc sống và hành vi xã hội

STT	Thuộc tính	Mô tả
1	famrel	Chất lượng mối quan hệ trong gia đình (1 - rất tệ đến 5 - rất tốt)
2	freetime	Thời gian rảnh sau học (1 - rất ít đến 5 - rất nhiều)
3	goout	Mức độ đi chơi với bạn bè (1 - rất ít đến 5 - rất nhiều)
4	Dalc	Mức độ uống rượu trong ngày thường (1 - rất thấp đến 5 - rất cao)
5	Walc	Mức độ uống rượu vào cuối tuần (1 - rất thấp đến 5 - rất cao)
6	health	Tình trạng sức khỏe hiện tại (1 - rất tệ đến 5 - rất tốt)
7	absences	Số buổi nghỉ học (0 đến 93)

- Nhóm 5: Kết quả học tập

Bảng 5. Mô tả thuộc tính về kết quả học tập

STT	Thuộc tính	Mô tả
1	G1	Điểm kỳ I (từ 0 đến 20)
2	G2	Điểm kỳ II (từ 0 đến 20)
3	G3	Điểm cuối kỳ (từ 0 đến 20)

3.2.3 Đặc điểm dữ liệu và khám phá sơ bộ

- Tổng số dòng dữ liệu: tổng sau khi gộp là 1044 bản ghi (bao gồm 395 bản ghi từ student-mat.csv và 649 bản ghi từ student-mat.csv)
- Số lượng thuộc tính: 33.
- Dữ liệu bao gồm cả biến định lượng (numerical) và biến định tính (categorical).

3.2.4 Đánh giá chất lượng dữ liệu

- Dữ liệu đa dạng và đầy đủ: Bao gồm các thông tin về cá nhân học sinh, gia đình, học tập, sức khỏe, mối quan hệ xã hội và kết quả học tập, giúp mô hình có thể phân tích

dưới nhiều khía cạnh.

- Có sự kết hợp giữa biến định lượng và định tính: Ví dụ, age, studytime, G1, G2, G3 là biến định lượng; trong khi sex, school, Mjob, Fjob, reason, guardian là biến định tính. Điều này yêu cầu tiền xử lý dữ liệu phù hợp cho từng loại biến.
- Một số biến có phân bố mất cân bằng: Chẳng hạn, tỷ lệ học sinh thuộc từng trường học hoặc từng giới tính có thể chênh lệch đáng kể.
- Biến mục tiêu là G3 (điểm cuối kỳ): Đây là biến số cần được dự đoán hoặc phân tích, có thang điểm từ 0 đến 20.
- Dữ liệu có thể chứa nhiều hoặc giá trị ngoại lai: Ví dụ, số buổi vắng (absences) dao động từ 0 đến 93, có thể có học sinh vắng học bất thường.
- Không có dữ liệu thiếu (missing values), tương đối sạch nhưng cần chuẩn hóa/mã hóa để phù hợp với thuật toán K-Means.

3.3 Data Preparation

3.3.1 Tích hợp dữ liệu

- Dữ liệu gồm 2 tập tin: student-mat.csv (Toán) và student-por.csv (Ngữ văn) từ bộ dữ liệu Student Performance.
- Hai tệp được nối lại bằng pd.concat và loại trùng (drop_duplicates()).
- Giá trị thiếu được loại bỏ (dropna()).

```
def load_data():  
    df1= pd.read_csv("./data/student-mat.csv")  
    df2 = pd.read_csv("./data/student-por.csv", sep=";")  
    df = pd.concat([df1, df2], ignore_index=True)  
    df = df.drop_duplicates()  
    df = df.dropna()  
    df = df.reset_index(drop=True)  
    return df  
  
df = load_data()
```

3.3.2 Tiền xử lý dữ liệu điểm số

- Loại bỏ học sinh không có điểm cuối kỳ: G3 là điểm số cuối kỳ – đóng vai trò trung tâm trong việc đánh giá hiệu suất học tập của học sinh. Tuy nhiên, một số bản ghi có giá trị $G3 = 0$, điều này có thể là do học sinh bỏ học, bị cấm thi, hoặc dữ liệu ghi nhận sai. Nếu giữ lại, các bản ghi này sẽ ảnh hưởng tiêu cực đến phân tích. Do đó, nhóm quyết định loại bỏ các trường hợp này.

```
df = df[df["G3"] > 0]
```

- Loại bỏ outlier trên các biến định lượng: Các biến số được lọc outlier theo phương pháp IQR (Interquartile Range). Áp dụng cho các biến: ["absences", "G3", "G1", "G2"] vì trong dữ liệu thực tế, không hiếm các trường hợp có giá trị bất thường như học sinh nghỉ học hơn 50 buổi hoặc có điểm G1, G2 quá thấp hoặc quá cao không hợp lý.

```
num_cols = ["absences", "G3", "G1", "G2"]
df_cleaned = df.copy()
for col in num_cols:
    Q1 = df_cleaned[col].quantile(0.25)
    Q3 = df_cleaned[col].quantile(0.75)
    IQR = Q3 - Q1
    lower = Q1 - 1.5 * IQR
    upper = Q3 + 1.5 * IQR
    df_cleaned = df_cleaned[(df_cleaned[col] >= lower) & (df_cleaned[col] <= upper)]
```

3.4 Modeling

3.4.1 Chọn biến đầu vào

Sau khi làm sạch, nhóm giữ lại 13 biến định lượng và 18 biến định tính. Thông qua giao diện multiselect của Streamlit, giảng viên có thể linh hoạt bật, tắt các biến để thử nghiệm (mặc định là chọn tất cả biến).

```
selected_num_cols = st.multiselect("Chọn biến định lượng:", numerical_cols,
default=numerical_cols)

selected_cat_cols = st.multiselect("Chọn biến định tính:", categorical_cols,
default=categorical_cols)
```

Biến định lượng gồm các cột về học lực và hành vi: Medu, Fedu, studytime, traveltime, failures, famrel, freetime, goout, Dalc, Walc, health, absences, G3.

Biến định tính gồm các cột dạng phân loại thông tin nhân khẩu học, hoàn cảnh và hỗ trợ học tập như school, sex, age, address, famsize, Pstatus, Mjob, Fjob, reason,

guardian, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic.

Không đưa G1 và G2 vào mô hình phân cụm vì chúng là điểm số giữa kỳ và có tương quan rất cao với G3 (điểm cuối kỳ). Việc đưa cả G1, G2, G3 vào mô hình sẽ khiến mô hình phụ thuộc quá nhiều vào điểm số, giảm khả năng khám phá ra các nhóm dựa trên yếu tố hành vi và bối cảnh. Vì vậy, chỉ giữ lại G3 là đại diện cho kết quả học tập cuối cùng.

- Ưu điểm của việc đưa các biến nào vào mô hình phân cụm:
 - + Phản ánh đa chiều về môi trường sống, hành vi và học tập của học sinh.
 - + Giúp mô hình nhận diện các nhóm sinh viên tương đồng không chỉ về điểm số mà còn về đặc điểm xã hội và học tập.
 - + Phù hợp với mục tiêu khám phá (unsupervised) hơn là dự đoán.
- Hạn chế và rủi ro:
 - + Việc chọn quá nhiều biến định tính có thể khiến không gian đặc trưng trở nên phức tạp, làm giảm hiệu quả phân cụm.
 - + One-hot encoding biến định tính tạo ra rất nhiều chiều, gây khó khăn trong trực quan và tăng thời gian tính toán.
 - + Một số biến có thể ít ảnh hưởng đến học lực, gây nhiễu (ví dụ: romantic, nursery).

3.4.2 Chuẩn hóa và biến đổi dữ liệu

- Chuẩn hóa dữ liệu định lượng: Dùng StandardScaler để đưa tất cả các biến số về trung bình 0, độ lệch chuẩn 1:

```
from sklearn.preprocessing import StandardScaler, OneHotEncoder  
  
scaler = StandardScaler()  
  
X_num_scaled = scaler.fit_transform(X_num)
```

- One-hot encode dữ liệu định tính: Dùng OneHotEncoder để chuyển dữ liệu phân loại sang dạng nhị phân:

```
encoder = OneHotEncoder(sparse_output=False)  
  
X_cat_encoded = encoder.fit_transform(X_cat)
```

- Ghép hai loại dữ liệu thành một ma trận đặc trưng:

```
X_scaled = np.hstack([X_num_scaled, X_cat_encoded])
```

Quá trình này đảm bảo tất cả các biến đều ở định dạng phù hợp và có trọng số tương đương trong mô hình K-Means.

3.4.3 Giảm chiều dữ liệu bằng PCA

Do ma trận đặc trưng `X_scaled` có số chiều lớn, phương pháp PCA (Principal Component Analysis) được sử dụng để giảm chiều dữ liệu xuống 2D và 3D nhằm hỗ trợ trực quan hóa:

```
from sklearn.decomposition import PCA  
  
pca_model = PCA(n_components=2)  
pca_2d = pca_model.fit_transform(X_scaled)  
  
pca_model_3d = PCA(n_components=3)  
pca_3d = pca_model_3d.fit_transform(X_scaled)
```

- Mục tiêu:
 - + Giảm số chiều dữ liệu để dễ dàng trực quan hóa các cụm.
 - + Giữ lại phần lớn phương sai trong dữ liệu gốc thông qua các thành phần chính (PC1, PC2, PC3).
- Kết quả:
 - + `pca_2d`: Ma trận dữ liệu giảm xuống 2 chiều, dùng cho biểu đồ scatter 2D.
 - + `pca_3d`: Ma trận dữ liệu giảm xuống 3 chiều, dùng cho biểu đồ scatter 3D tương tác.

3.4.4 Áp dụng thuật toán K-Means

Thuật toán K-Means được áp dụng trên ma trận đặc trưng `X_scaled` để phân cụm học sinh:

```
kmeans = KMeans(n_clusters=chosen_k, random_state=42, n_init=20)  
clusters = kmeans.fit_predict(X_scaled)  
df_cleaned["Cluster"] = clusters
```

- Tham số:
 - + `n_clusters=chosen_k`: Số cụm K được chọn qua giao diện Streamlit (mặc định 4, phạm vi 2-10).
 - + `random_state=42`: Đảm bảo tính tái hiện.
 - + `n_init=20`: Chạy thuật toán 20 lần với các tâm cụm khởi tạo ngẫu nhiên để chọn kết quả tốt nhất.
- Kết quả: Mỗi học sinh được gán một nhãn cụm (từ 0 đến $k-1$), lưu trong cột Cluster của `df_cleaned`.

3.4.5 Phân tích số cụm tối ưu

Để lựa chọn số lượng cụm K phù hợp, nhóm sử dụng hai phương pháp chính:

- Elbow Method (Phương pháp khuỷu tay): Dựa vào tổng sai số bình phương trong cụm (SSE – Sum of Squared Errors) ứng với từng giá trị K .

```
k_range = range(2, 7)
sse = []
for k in k_range:
    km_tmp = KMeans(n_clusters=k, random_state=42,
n_init=20).fit(st.session_state.X_scaled_selected)
    sse.append(km_tmp.inertia_)
```

- Silhouette Score: Đánh giá chất lượng phân cụm dựa trên mức độ tương đồng của các điểm trong cùng một cụm và sự tách biệt giữa các cụm khác nhau.

```
sil = []
for k in k_range:
    km_tmp = KMeans(n_clusters=k, random_state=42,
n_init=20).fit(st.session_state.X_scaled_selected)
    sil.append(silhouette_score(st.session_state.X_scaled_selected,
km_tmp.labels_))
```

Cách thực hiện:

- Dữ liệu đầu vào là tập dữ liệu đã được xử lý (X_processed).
- Vòng lặp được thực hiện cho các giá trị K từ 2 đến 6 (k_range = range(2, 7)).
- Với mỗi giá trị K:
 - + Huấn luyện mô hình KMeans
 - + Tính SSE và Silhouette Score tương ứng
- Tạo bảng kết quả so sánh các giá trị K.
- Chọn K tối ưu là giá trị có Silhouette Score cao nhất.

Trực quan hóa: Kết quả được hiển thị qua hai biểu đồ:

```
fig_elbow, ax1 = plt.subplots(figsize=(3, 3))
ax1.plot(k_range, sse, marker="o")
ax1.set_title("Elbow Method")
ax1.set_xlabel("k")
ax1.set_ylabel("SSE")
st.pyplot(fig_elbow)

fig_sil, ax2 = plt.subplots(figsize=(3, 3))
ax2.plot(k_range, sil, marker="o", color="orange")
ax2.set_title("Silhouette Scores")
ax2.set_xlabel("k")
ax2.set_ylabel("Silhouette")
st.pyplot(fig_sil)
```

3.4.6 Trực quan hóa và đánh giá sơ bộ kết quả phân cụm

Kết quả phân cụm được trực quan hóa và đánh giá sơ bộ để hiểu rõ đặc điểm của từng cụm:

- Trực quan hóa PCA 2D: Biểu đồ scatter 2D hiển thị các điểm dữ liệu và tâm cụm, giúp quan sát sự phân tách giữa các cụm.

```
fig_pca2d, ax = plt.subplots(figsize=(6, 3.5))

sns.scatterplot(x=pca_2d[:, 0], y=pca_2d[:, 1],
hue=df_cleaned["Cluster"].astype(str), palette="Set2", ax=ax, s=40)

ax.scatter(centroids_2d[:, 0], centroids_2d[:, 1], s=50, c="black",
marker="X", label="Centroid")

for i, (x, y) in enumerate(centroids_2d):
    ax.text(x + 0.05, y + 0.05, f"C{i+1}", fontsize=10, color="black",
ha="left", va="bottom")

ax.set_title("PCA 2D với tâm cụm")

ax.legend(title="Cụm", fontsize="small", labelspace=0.5)

plt.tight_layout()

st.pyplot(fig_pca2d)
```

- Trực quan hóa PCA 3D: Biểu đồ 3D tương tác cho phép quan sát các cụm từ nhiều góc độ, với thông tin bổ sung về điểm G3 khi di chuột.

```
df_pca = pd.DataFrame(pca_3d, columns=["PC1", "PC2", "PC3"])

df_pca["Cluster"] = df_cleaned["Cluster"].astype(str)

df_pca["idx"] = df_pca.index

df_pca["G3"] = df_cleaned["G3"].values

fig3d = px.scatter_3d(df_pca, x="PC1", y="PC2", z="PC3", color="Cluster",
color_discrete_sequence=px.colors.qualitative.Set2, opacity=0.85,
size_max=10, title="🌀 PCA 3D", hover_data=["G3"])

centroid_trace = go.Scatter3d(x=centroids_3d[:, 0], y=centroids_3d[:, 1],
z=centroids_3d[:, 2], mode='markers+text', marker=dict(size=6,
color='black', symbol='x'), text=[f'C{i}' for i in range(chosen_k)],
textposition='top center', name='Centroids')
```

```
fig3d.add_trace(centroid_trace)

fig3d.update_layout(title="✿ PCA 3D Interactive with Centroids",
margin=dict(l=0, r=0, b=0, t=40), scene=dict(aspectmode="cube"),
legend=dict(title="Cụm"))

st.plotly_chart(fig3d, use_container_width=True)
```

- Đánh giá sơ bộ: Bảng chéo giữa Cluster và G3_level được sử dụng để đánh giá mức độ tương đồng giữa các cụm và hiệu suất học tập:

```
ct = pd.crosstab(df_cleaned["Cluster"], df_cleaned["G3_level"],
normalize="index") * 100

g3_mean = df_cleaned.groupby("Cluster")["G3"].mean().round(2)

ct["Trung bình G3"] = g3_mean

st.dataframe(ct.style.format({"Trung bình G3": "{:.2f}", "Giỏi": "{:.1f}%",
"Khá": "{:.1f}%", "Trung bình": "{:.1f}%", "Yếu":
"{:.1f}%"}) .highlight_max(axis=1, color="lightgreen"))

purity = ct[["Giỏi", "Khá", "Trung bình", "Yếu"]].max(axis=1).mean() / 100

st.markdown(f"✿ **Purity trung bình**: {purity:.2%}")
```

Bảng chéo hiển thị tỷ lệ phần trăm học sinh ở mỗi mức học lực (Yếu, Trung bình, Khá, Giỏi) trong từng cụm. Độ tinh khiết trung bình (purity) được tính để đánh giá mức độ đồng nhất về học lực trong mỗi cụm.

3.5 Evaluation

Giai đoạn Evaluation tập trung vào việc đánh giá chất lượng của mô hình phân cụm K-Means và phân tích ý nghĩa của các cụm đối với mục tiêu nghiệp vụ.

3.5.1 Đánh giá chất lượng phân cụm

Chất lượng phân cụm được đánh giá thông qua các chỉ số nội tại và so sánh với biến mục tiêu G3:

- Chỉ số Silhouette: Giá trị Silhouette cao nhất trong khoảng K từ 2 đến 6 được báo cáo để xác nhận K được chọn là phù hợp. Tuy kết quả $K=2$ nhưng trên thực tế khi phân cụm, thông thường người ta sẽ phân thành nhiều cụm hơn như 4, 5 hoặc 6 cụm trở lên.
- Độ tinh khiết (Purity): Purity đo lường mức độ đồng nhất về học lực trong mỗi cụm.

Giá trị purity cao cho thấy các cụm phản ánh tốt sự khác biệt về hiệu suất học tập.

- Hậu kiểm với điểm trung bình G3 cho từng cụm: Trung bình điểm G3 của mỗi cụm được so sánh để xác định các cụm có hiệu suất học tập cao, khá, trung bình, hoặc thấp...

3.5.2 Số lượng học sinh trong mỗi cụm

Sau khi gán nhãn cụm cho toàn bộ dữ liệu, nhóm thống kê được số lượng sinh viên trong từng cụm và trực quan hóa bằng biểu đồ cột.

```
fig_count, ax = plt.subplots(figsize=(4, 3))
sns.barplot(data=count_df, x="Tên cụm", y="Số học sinh", palette="Set2",
ax=ax)
ax.set_title("Số học sinh theo cụm")
for p in ax.patches:
    ax.annotate(f'{int(p.get_height())}', (p.get_x() + p.get_width() / 2,
p.get_height()), ha='center', va='bottom', fontsize=7)
st.pyplot(fig_count)
```

3.5.3 Phân tích đặc trưng của từng cụm

3.5.3.1 Biến định lượng

```
numerical_summary =
df_cleaned.groupby('Cluster')[st.session_state.selected_num_cols].mean().ro
und(2)
st.dataframe(numerical_summary.style.highlight_max(axis=0,
color='lightblue'))
fig_num, ax = plt.subplots(figsize=(4, 3))
sns.boxplot(data=df_cleaned, x="Cluster", y=selected_num, palette="pastel",
ax=ax)
ax.set_title(f"Phân phối {selected_num} theo cụm")
st.pyplot(fig_num)
```

Trung bình và phân phối (boxplot) của các biến định lượng (như studytime, failures, absences) được tính để xác định đặc điểm nổi bật của từng cụm. Ví dụ, cụm có trung bình studytime cao có thể đại diện cho nhóm học sinh chăm chỉ.

3.5.3.2 Biến định tính

```
tab = pd.crosstab(st.session_state.clusters_selected,
df_cleaned[selected_cat], normalize="index") * 100
st.dataframe(tab.style.format("{:.1f}%").highlight_max(axis=1,
color="lightgreen"))
fig_cat, ax = plt.subplots(figsize=(5, 3.5))
chart = sns.countplot(data=df_plot, x=selected_cat, hue="Cluster",
palette="Set3", ax=ax)
for p in chart.patches:
    height = p.get_height()
    if height > 0:
        ax.annotate(f"{int(height)}", (p.get_x() + p.get_width() / 2,
height), ha='center', va='bottom', fontsize=8)
ax.set_title(f"Phân phối '{selected_cat}' theo cụm")
ax.tick_params(axis='x', rotation=45)
st.pyplot(fig_cat)
```

Tỷ lệ phần trăm các giá trị của biến định tính (như school, sex, higher) trong mỗi cụm được phân tích để nhận diện các đặc điểm xã hội hoặc học tập đặc trưng.

3.5.3.3 Top đặc trưng quan trọng:

```
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier(random_state=42)
rf.fit(st.session_state.X_scaled_selected,
st.session_state.clusters_selected)
feat_df = pd.DataFrame({"Feature": st.session_state.selected_features,
"Importance": rf.feature_importances_}).sort_values(by="Importance",
ascending=False)
fig_feat, ax = plt.subplots(figsize=(5, 3.5))
sns.barplot(data=feat_df.head(top_n), x="Importance", y="Feature",
palette="Blues_r", ax=ax)
ax.set_title("Top đặc trưng quan trọng nhất")
st.pyplot(fig_feat)
```

Mô hình Random Forest được sử dụng để xác định các đặc trưng quan trọng nhất trong việc phân biệt các cụm, giúp tập trung vào các yếu tố có ảnh hưởng lớn.

3.6 Deployment

Giai đoạn Deployment tập trung vào việc triển khai mô hình và tích hợp kết quả phân cụm vào thực tiễn giáo dục.

3.6.1 Triển khai giao diện với thư viện Streamlit

Mô hình phân cụm được triển khai dưới dạng ứng dụng web sử dụng Streamlit, cho phép người dùng (giáo viên, nhà quản lý) tương tác với dữ liệu và kết quả phân cụm:

```
st.set_page_config(page_title="Phân cụm học sinh", layout="wide")

st.title("📖 Phân cụm sinh viên dựa trên các đặc điểm học tập của họ, từ đó khám phá các nhóm sinh viên có hiệu suất tương tự")

with st.sidebar:

    st.header("🔗 Cài đặt")

    chosen_k = st.slider("Chọn số cụm k", min_value=2, max_value=10, value=4)
```

Chức năng chính:

- Cho phép chọn số cụm K và các biến đầu vào (định lượng và định tính).
- Hiện thị các bước phân tích: Giao diện được chia thành 8 bước, từ làm sạch dữ liệu, phân tích Elbow và Silhouette, đến trực quan hóa PCA 2D/3D và phân tích đặc trưng từng cụm.

```
steps = [

    "1. Dữ liệu ban đầu và làm sạch",

    "2. Phân tích Elbow & Silhouette",

    "3. Phân cụm và PCA Visualization",

    "4. Số lượng học sinh mỗi cụm",

    "5. 📊 Biến định tính (Categorical)",

    "6. 📊 Biến định lượng (Numerical)",

    "7. 🔍 Top N đặc trưng gốc",

    "8. 🔍 Khám phá đặc trưng mỗi cụm"

]
```

```
with st.sidebar:

    st.subheader("📄 Bước phân tích")

    chosen_step = st.radio("Chọn bước muốn xem:", steps)

st.markdown(f"## 📊 {chosen_step}")

progress = (steps.index(chosen_step) + 1) / len(steps)

st.progress(progress)
```

- Cung cấp biểu đồ tương tác (PCA 2D, PCA 3D) và bảng dữ liệu để khám phá đặc trưng của từng cụm.

Lưu trạng thái: Trạng thái của các biến được chọn được lưu vào tệp JSON để duy trì cài đặt giữa các phiên làm việc.

```
def save_state(selected_num, selected_cat):

    with open("state.json", "w") as f:

        json.dump({"selected_num_cols": selected_num, "selected_cat_cols":
selected_cat}, f)

def load_state():

    if os.path.exists("state.json"):

        with open("state.json", "r") as f:

            state = json.load(f)

            return state.get("selected_num_cols", []),
state.get("selected_cat_cols", [])

    return numerical_cols, categorical_cols
```

3.6.2 Quy trình triển khai trên Streamlit Community Cloud:

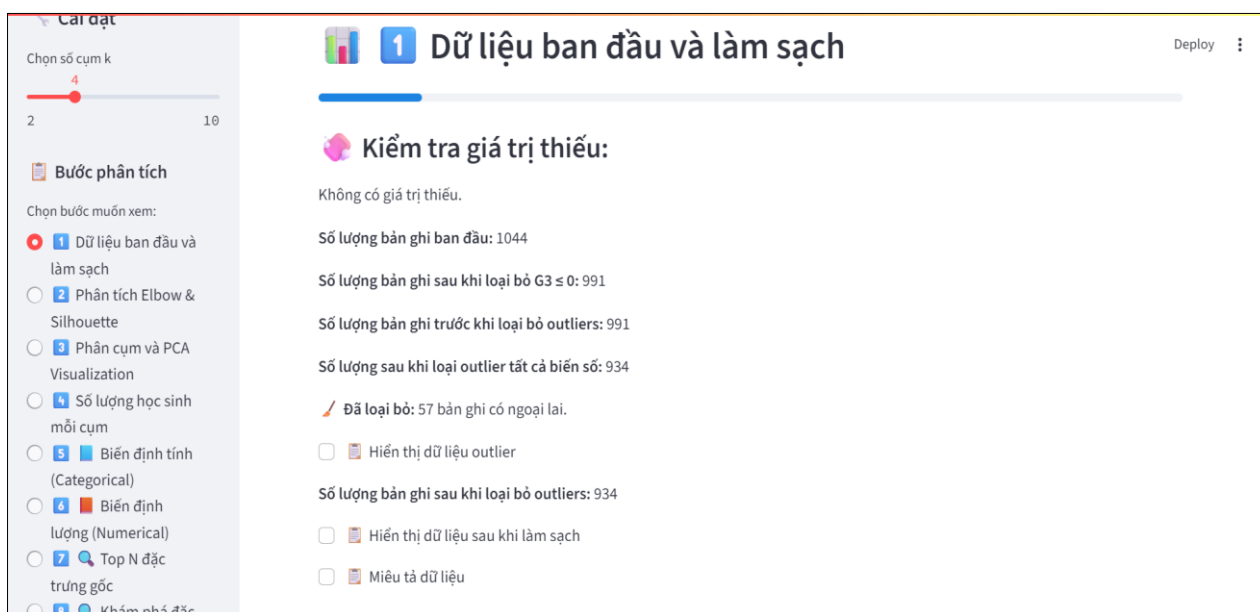
- Mã nguồn được đẩy lên GitHub repository, với tệp chính, ví dụ report.py được chỉ định là main file.
- Ứng dụng được kết nối với Streamlit Community Cloud thông qua liên kết repository GitHub. Khi nhánh main được cập nhật, Streamlit tự động build và deploy ứng dụng, loại bỏ nhu cầu chạy cục bộ (localhost).
- Các thư viện cần thiết được liệt kê trong requirements.txt để Streamlit cài đặt tự động.

CHƯƠNG 4: KẾT QUẢ THỰC HIỆN

4.1 Tổng quan về dữ liệu sau tiền xử lý

Sau các bước tích hợp và làm sạch dữ liệu (*xem mục 3.3*), bộ dữ liệu ban đầu gồm 1044 bản ghi đã được xử lý để loại bỏ giá trị trùng lặp, giá trị thiếu, và outlier. Kết quả cụ thể được hiển thị trong ứng dụng Streamlit như sau:

- Số lượng bản ghi ban đầu: 1044.
- Số lượng bản ghi sau khi loại bỏ $G3 \leq 0$: 991.



Hình 4.1 Tổng quan dữ liệu sau làm bước tiền xử lý

- Số lượng bản ghi sau khi tiếp tục loại bỏ outlier: 934 (loại bỏ thêm 57 bản ghi có giá trị ngoại lai trên các biến absences, G3, G1, G2).

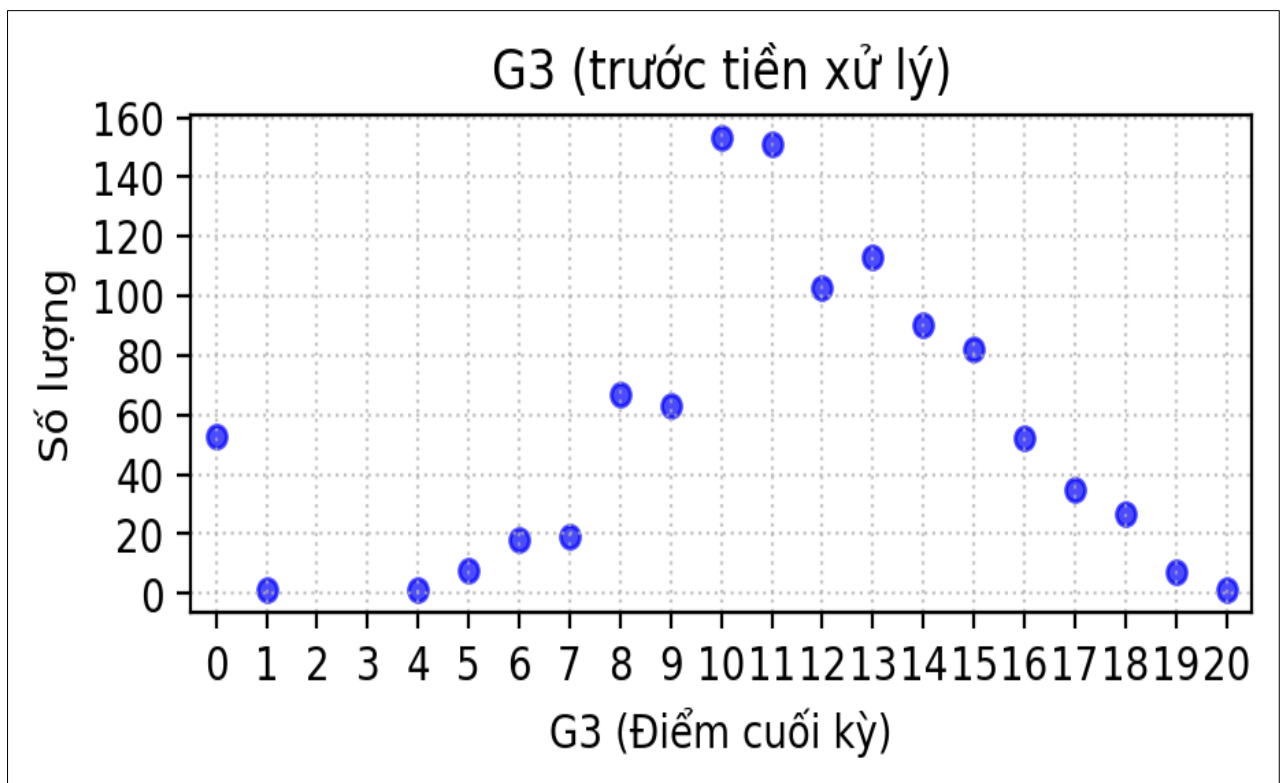
	id	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3	is_outlier
25	2	1	no	yes	no	no	yes	yes	yes	yes	4	3	3	1	3	5	8	3	5	5	<input checked="" type="checkbox"/>
33	2	2	no	no	yes	no	yes	yes	yes	yes	3	4	5	2	4	1	22	6	6	4	<input checked="" type="checkbox"/>
34	1	3	no	yes	no	yes	yes	yes	yes	no	5	5	5	2	4	5	16	6	5	5	<input checked="" type="checkbox"/>
53	2	0	yes	yes	yes	no	yes	yes	yes	no	4	3	5	1	1	2	26	7	6	6	<input checked="" type="checkbox"/>
54	1	0	no	yes	no	no	yes	yes	yes	no	5	3	2	1	2	3	18	7	6	6	<input checked="" type="checkbox"/>
66	2	0	no	yes	no	no	yes	yes	no	no	4	1	3	3	5	5	18	8	6	7	<input checked="" type="checkbox"/>
69	2	0	no	yes	yes	no	yes	yes	yes	no	3	4	2	1	1	5	18	9	7	6	<input checked="" type="checkbox"/>
70	1	3	no	no	no	no	no	no	yes	yes	5	4	5	5	5	1	16	6	8	8	<input checked="" type="checkbox"/>
82	2	1	no	no	no	yes	yes	yes	yes	yes	5	3	3	1	1	4	16	9	8	7	<input checked="" type="checkbox"/>
89	1	0	no	no	yes	yes	yes	yes	yes	yes	4	5	4	2	4	5	30	8	8	8	<input checked="" type="checkbox"/>
91	2	1	no	yes	no	yes	yes	yes	yes	no	5	2	4	1	4	5	20	9	7	8	<input checked="" type="checkbox"/>
97	1	1	no	yes	yes	no	yes	yes	yes	yes	4	3	4	1	1	4	38	8	9	8	<input checked="" type="checkbox"/>
103	2	0	no	yes	no	yes	yes	yes	yes	yes	5	3	3	2	3	1	56	9	9	8	<input checked="" type="checkbox"/>
115	1	0	no	no	yes	yes	yes	yes	yes	no	3	2	4	1	4	3	22	9	9	9	<input checked="" type="checkbox"/>

Hình 4.2 Một số bản ghi là outlier

	age	Medu	Fedu	travelttime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
count	934	934	934	934	934	934	934	934	934	934	934	934	934	934	934	934
mean	16.6606	2.606	2.394	1.5203	1.9957	0.2141	3.9422	3.1938	3.1467	1.4732	2.2634	3.5439	3.621	11.4893	11.7034	12.0632
std	1.2168	1.1365	1.103	0.7372	0.8436	0.6029	0.93	1.0321	1.1317	0.8995	1.2761	1.4262	3.8047	2.8637	2.8073	2.8532
min	15	0	0	1	1	0	1	1	1	1	1	1	0	4	5	5
25%	16	2	1	1	1	0	4	3	2	1	1	3	0	10	10	10
50%	17	3	2	1	2	0	4	3	3	1	2	4	2	11	12	12
75%	18	4	3	2	2	0	5	4	4	2	3	5	6	14	14	14
max	22	4	4	4	4	3	5	5	5	5	5	5	15	19	19	20

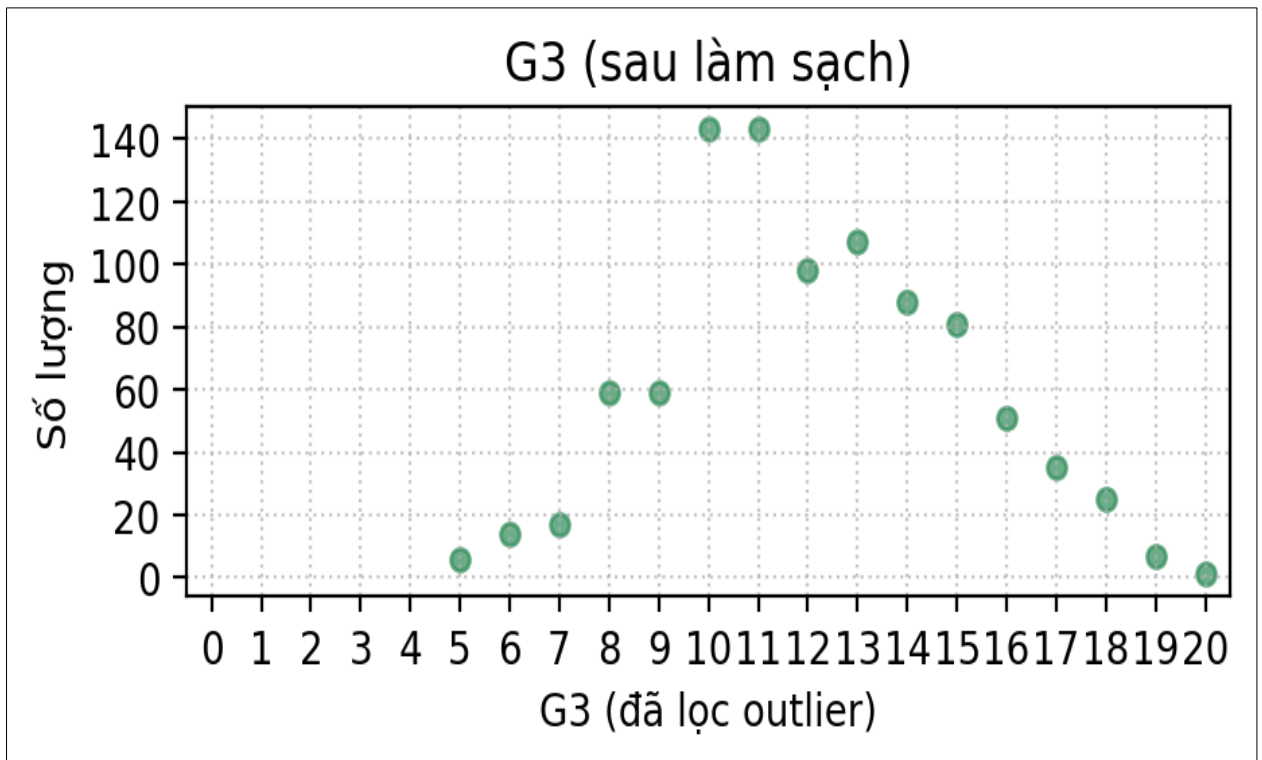
Hình 4.3 Mô tả dữ liệu sau khi đã làm sạch

Phân bố điểm G3 trước và sau khi làm sạch được trực quan hóa bằng các biểu đồ scatter (hình minh họa 4.4 và 4.5):



Hình 4.4 Phân bố G3 trước tiền xử lý

Mô tả: Biểu đồ scatter với trục x là giá trị G3 từ 0 đến 20, trục y là số lượng học sinh, các điểm màu xanh dương thể hiện phân bố ban đầu với một số giá trị ngoại lai ở hai đầu, đặc biệt là $G3 = 0$ và $G3 > 15$.



Hình 4.5 Phân bố G3 sau làm sạch dữ liệu

Mô tả: Biểu đồ scatter với trục x là giá trị G3 từ 0 đến 20, trục y là số lượng học sinh, các điểm màu xanh lá thể hiện phân bố sau khi loại bỏ outlier, phân bố tập trung hơn ở khoảng 5 đến 15, loại bỏ các giá trị bất thường.

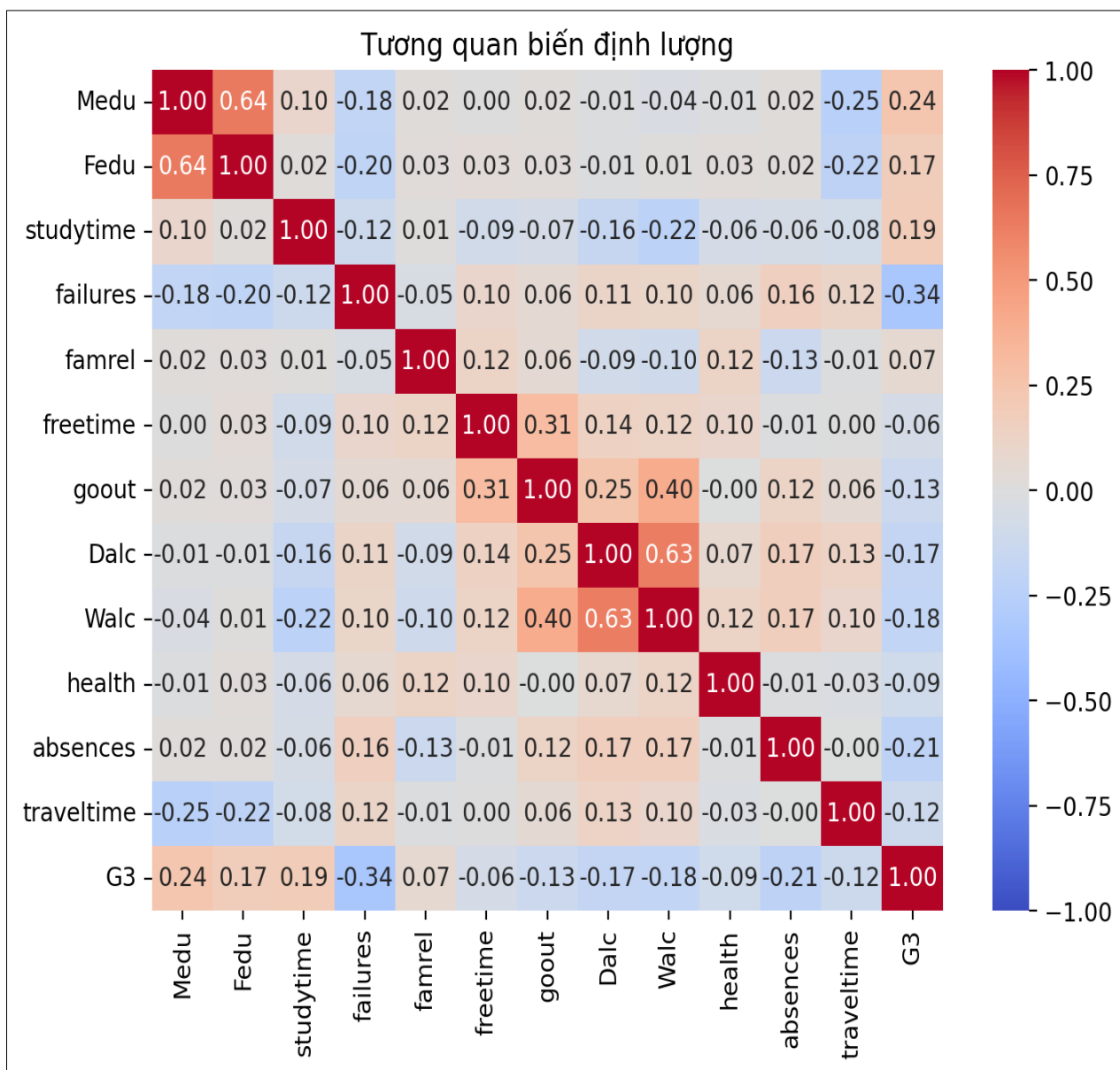
Phân bố G3 sau khi làm sạch cho thấy sự giảm bớt các giá trị ngoại lai, tạo điều kiện thuận lợi cho việc phân cụm.

4.2 Kết quả phân cụm với K-Means

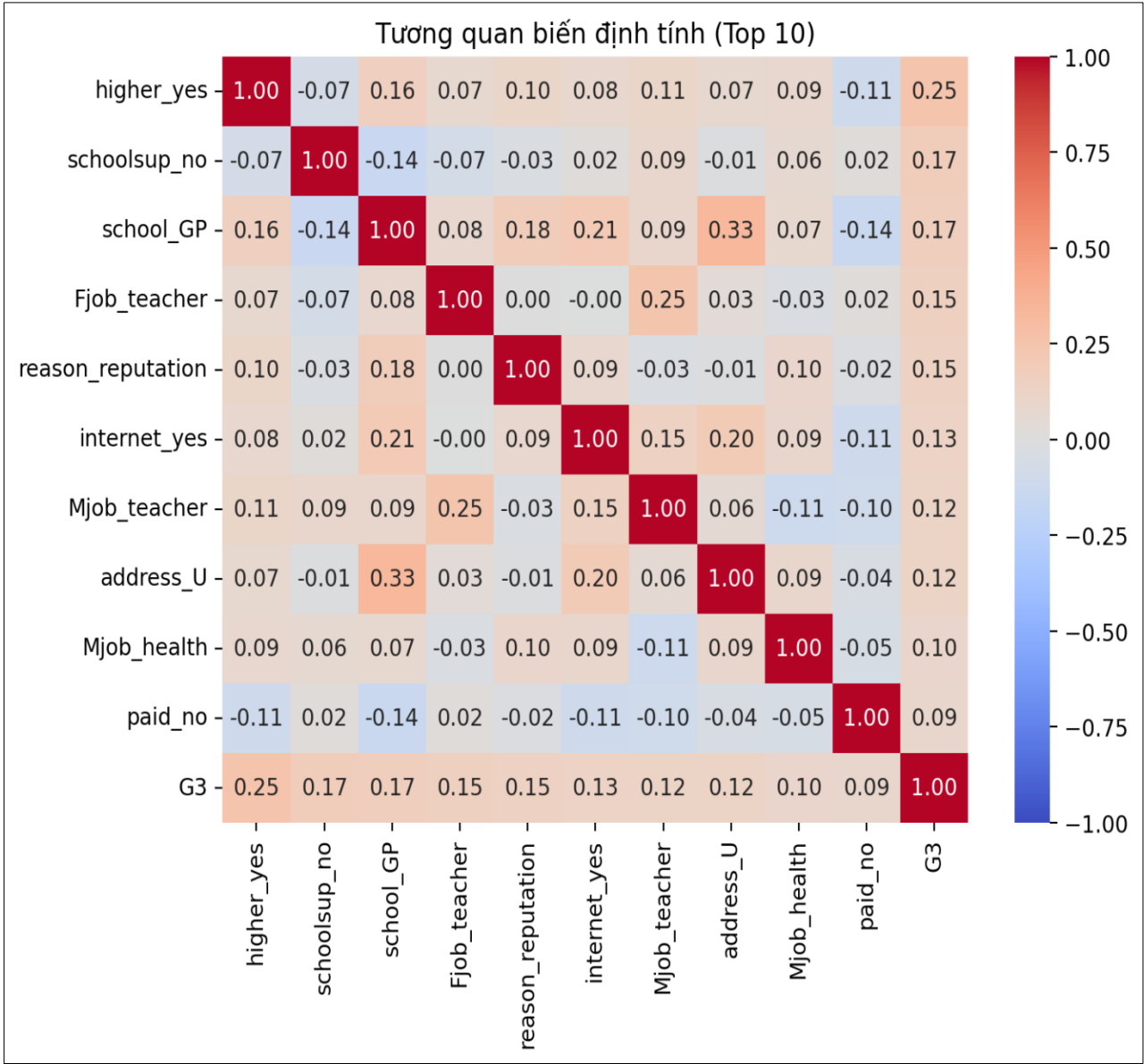
Mô hình K-Means đã được áp dụng trên ma trận đặc trưng X_{scaled} với số cụm $k = 4$ (giá trị mặc định trong ứng dụng). Kết quả phân cụm được trực quan hóa và đánh giá như sau:

4.2.1 Lựa chọn số cụm tối ưu

Trước tiên dùng heatmap để trực quan hóa mối tương quan giữa các biến định lượng và định tính với nhau, đặc biệt với G3. Dựa trên heatmap, em chọn những biến thực sự liên quan đến học lực, để khi phân cụm thì các cụm sẽ rõ ràng hơn. Vẽ Heatmap không ảnh hưởng trực tiếp đến SSE hoặc việc chọn K, nhưng gián tiếp giúp chọn biến phù hợp để phân cụm, từ đó cải thiện cụm làm SSE tốt hơn và chọn K chính xác hơn.



Hình 4.6 Heatmap biến định lượng



Hình 4.7 Heatmap top 10 biến định tính đã chuẩn hóa

Tiếp theo người dùng có thể chọn các biến định lượng và định tính để đưa vào mô hình học phân cụm thông qua các select được hỗ trợ bởi Streamlit, mặc định nó sẽ là tất cả các thuộc tính (trừ G1, G2) và nhấn nút OK để tiến hành tính SSE và phân cụm:

Chọn biến để phân cụm

Chọn biến định lượng:

Medu × Fedu × studytime × failures × famrel × freetime × goout × Dalc × Walc × health ×

absences × traveltime × G3 ×

Chọn biến định tính:

school × sex × age × address × famsize × Pstatus × Mjob × Fjob × reason × guardian × schoolsup ×

famsup × paid × activities × nursery × higher × internet × romantic ×

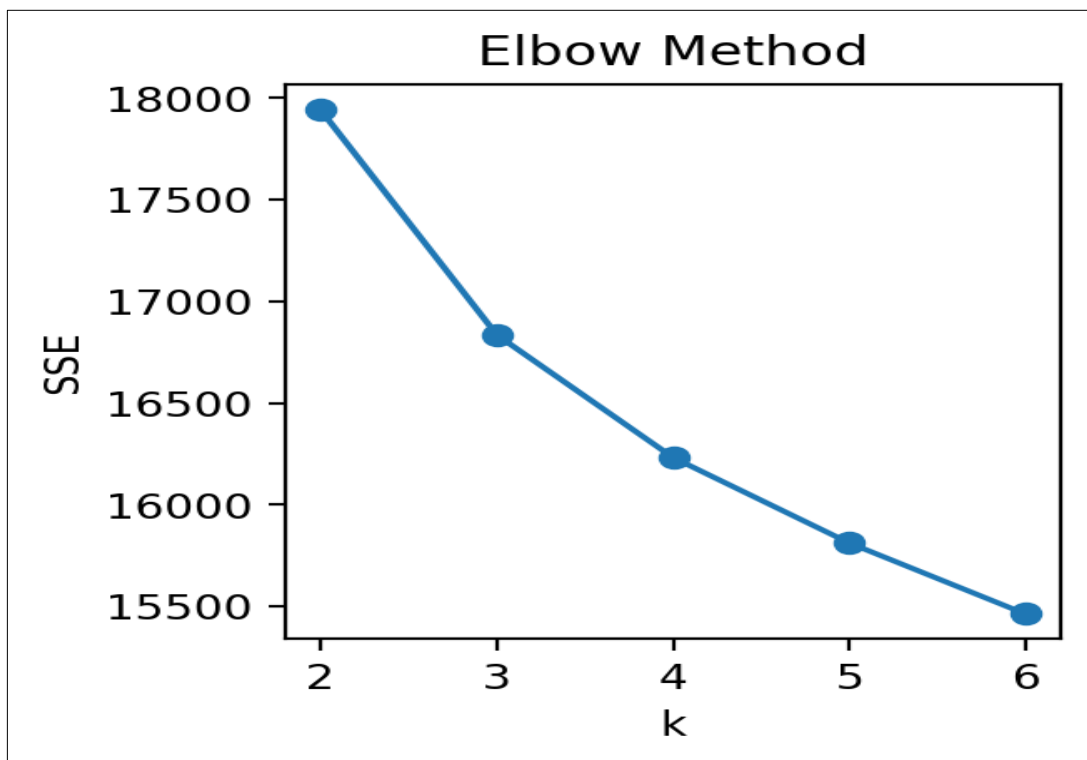
OK

Shape X_scaled_selected: (934, 64)

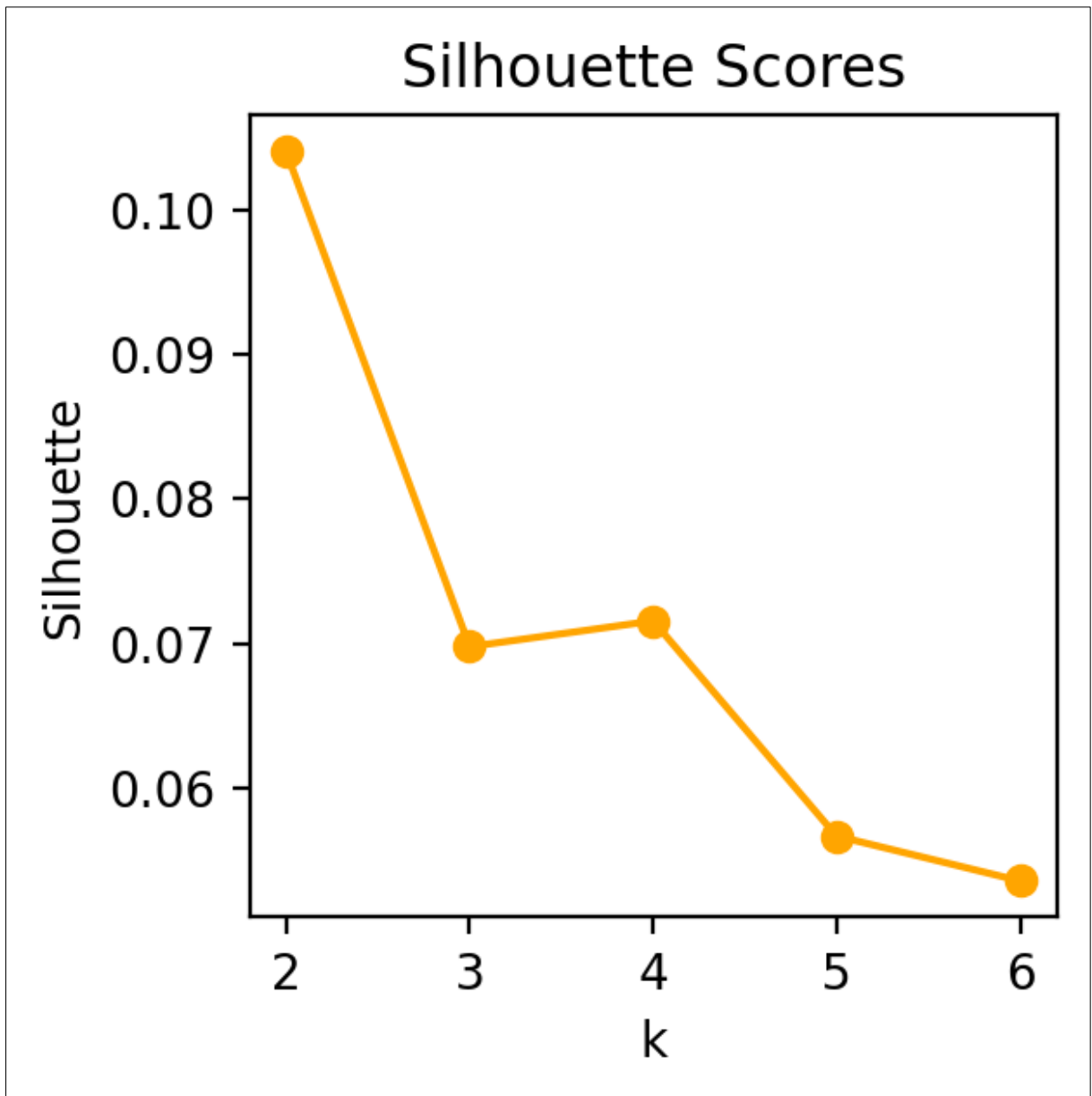
Đặc trưng đã chọn: ['Medu', 'Fedu', 'studytime', 'failures', 'famrel', 'freetime', 'goout', 'Dalc', 'Walc', 'health', 'absences', 'traveltime', 'G3', 'school_GP', 'school_MS', 'sex_F', 'sex_M', 'age_15', 'age_16', 'age_17', 'age_18', 'age_19', 'age_20', 'age_21', 'age_22', 'address_R', 'address_U', 'famsize_GT3', 'famsize_LE3', 'Pstatus_A', 'Pstatus_T', 'Mjob_at_home', 'Mjob_health', 'Mjob_other', 'Mjob_services', 'Mjob_teacher', 'Fjob_at_home', 'Fjob_health', 'Fjob_other', 'Fjob_services', 'Fjob_teacher', 'reason_course', 'reason_home', 'reason_other', 'reason_reputation', 'guardian_father', 'guardian_mother', 'guardian_other', 'schoolsup_no', 'schoolsup_yes', 'famsup_no', 'famsup_yes', 'paid_no', 'paid_yes', 'activities_no', 'activities_yes', 'nursery_no', 'nursery_yes', 'higher_no', 'higher_yes', 'internet_no', 'internet_yes', 'romantic_no', 'romantic_yes']

Hình 4.8 Giao diện cho các biến đầu vào

Phương pháp Elbow và Silhouette được sử dụng để xác định số cụm K tối ưu:



Hình 4.9 Biểu đồ phương pháp Elbow sau khi chọn biến đầu vào



Hình 4.10 Biểu đồ Silhouette score sau khi chọn biến đầu vào

Dựa vào hình ảnh của biểu đồ ta có thể thấy $k=2$ là lớn nhất (gần bằng 0.1), tuy nhiên khi dữ liệu phức tạp, chỉ chọn 2 hoặc 3 cụm là quá đơn giản, không đủ phản ánh được sự đa dạng cho nên chọn $k = 4$ hoặc 5 giúp chia chi tiết hơn thay vì chỉ tách ra giới và kém. Nhóm em đã chọn $k=4$ để phân cụm.

Mặc dù vậy nhìn chung các chỉ số SSE vẫn rất thấp và dao động từ 0.0 đến 0.1 cho thấy các cụm phân chia không rõ ràng và có thể chồng lấn giữa các cụm với nhau làm việc phân cụm trở nên kém hiệu quả hơn. Một số nguyên nhân làm cho SSE thấp dù đã làm đúng quy trình K-Means như:

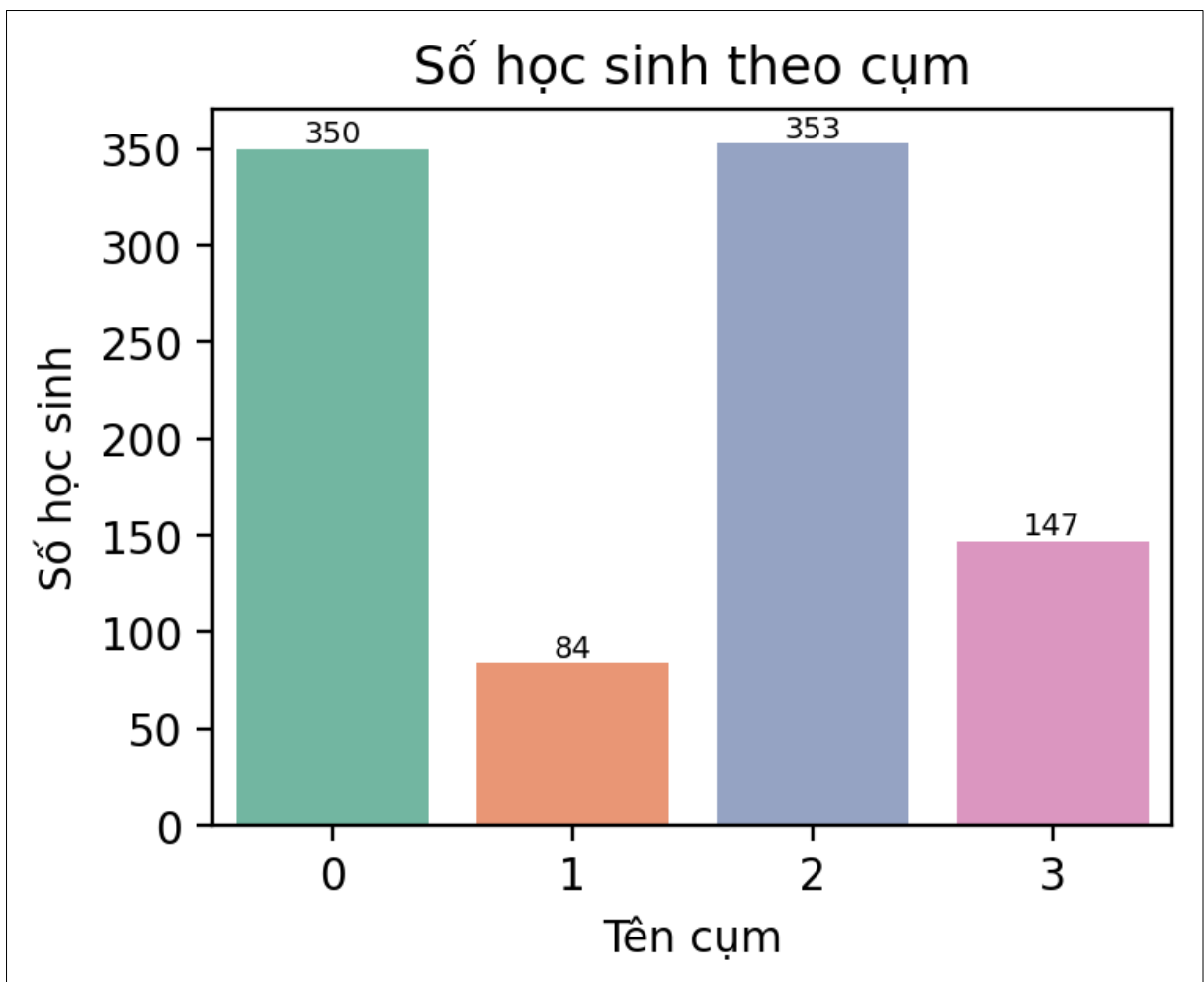
- Biến đầu vào chưa thật sự phân biệt được học lực: Dù chọn biến từ heatmap, có thể

một số biến vẫn chưa đủ mạnh để làm rõ ranh giới giữa các cụm. Ví dụ: chọn studytime, failures, nhưng nếu có nhiều giá trị phân bố không rõ, cụm tạo ra sẽ bị lẫn.

- Dữ liệu bị chênh lệch: Nếu một vài biến có phân phối lệch (ví dụ: hầu hết học sinh có absences thấp, chỉ vài học sinh absences cao), thì cụm bị lệch – một cụm chiếm đa số, cụm còn lại nhỏ làm giảm Silhouette.
- Thuật toán KMeans không phù hợp nếu cụm không tròn: KMeans giả định cụm có hình tròn đều, nhưng nếu dữ liệu không đáp ứng giả định này, thì KMeans phân cụm kém hiệu quả, làm score thấp.

4.2.2 Kết quả phân cụm

- Số lượng học sinh mỗi cụm :

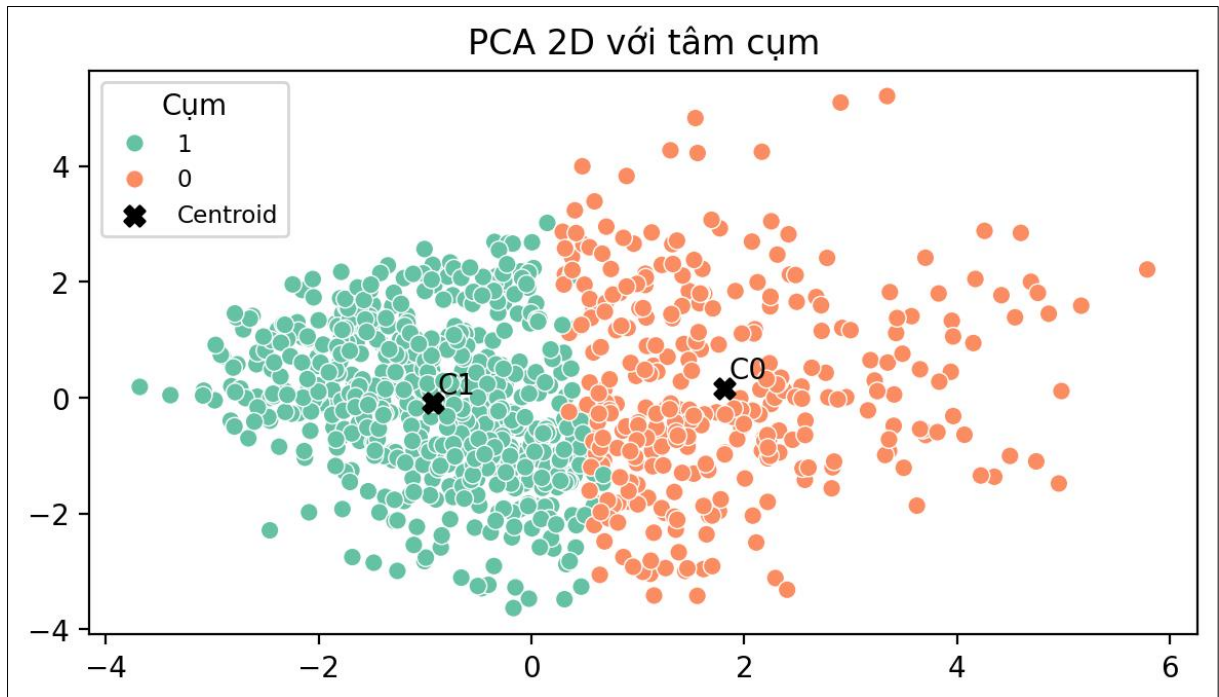


Hình 4.11 Biểu đồ cột thể hiện số lượng mỗi cụm

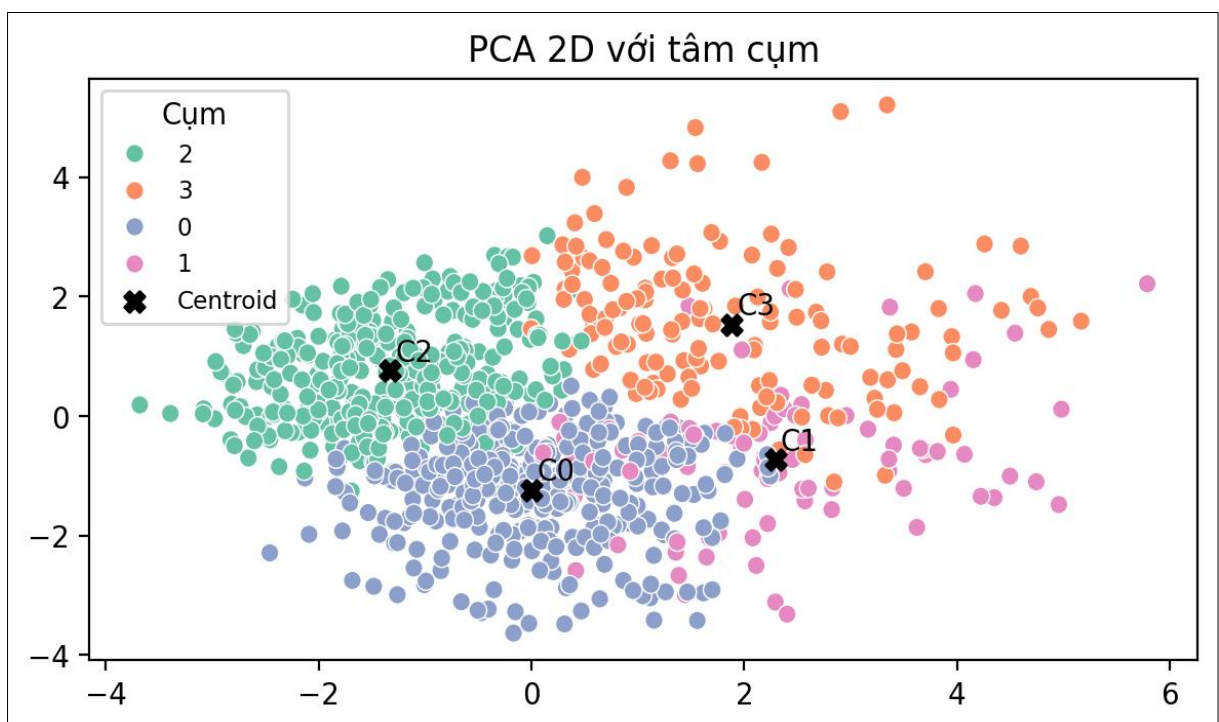
Mô tả: Biểu đồ cột với trục x là các cụm (0, 1, 2, 3), trục y là số lượng học sinh, các cột màu Set2 thể hiện số lượng như Cụm 0: 350, Cụm 1: 84, Cụm 2: 353, Cụm 3:

147, tổng cộng 934 học sinh.

- Trực quan hóa PCA 2D với $k=2$ và $k=4$:



Hình 4.12 Biểu đồ PCA 2D với 2 cụm



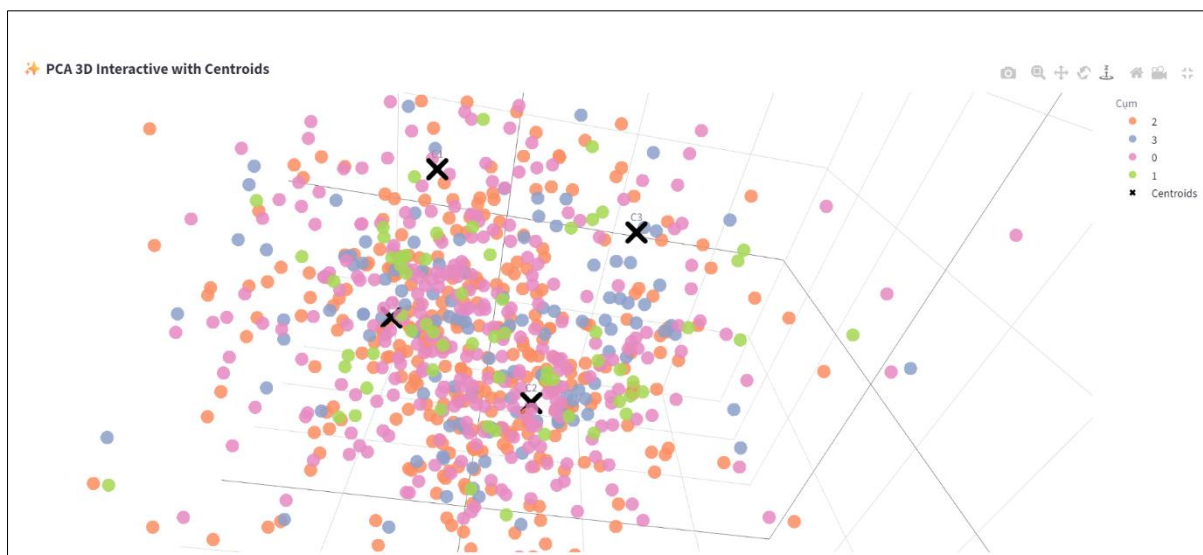
Hình 4.13 Biểu đồ PCA 2D với 4 cụm

Bảng 6. Nhận xét các cụm sau khi quan sát PCA 2D 4 cụm

Cụm	Số lượng quan sát	Nhận xét
0	Đông đảo, phân bố tập trung	Là nhóm có học lực trung bình khá ($G3 \approx 11.99$)
1	Ít nhất trong các cụm	$G3$ thấp nhất (≈ 9.23), phần lớn thuộc nhóm học lực yếu – trung bình, học ít
2	Rất đông, tập trung	$G3$ cao nhất (≈ 13.42), nhiều học sinh giỏi – khá, thời gian học cao
3	Đông, trải rộng	$G3$ trung bình (≈ 10.91), phân tán nhẹ, mức học lực không đồng đều

Đánh giá tổng thể: Mô hình phân cụm khá tốt, tâm cụm cách biệt rõ ràng, PCA giúp giảm chiều dữ liệu và trực quan hóa hiệu quả, hỗ trợ giải thích mô hình tốt hơn. Tuy nhiên, vẫn có sự chồng lấn giữa các cụm, cho thấy một số học sinh không hoàn toàn phù hợp rõ ràng với một nhóm duy nhất, cần xem xét thêm các biến định tính để giải thích.


- Trực quan hóa PCA 3D:



Hình 4.14 Biểu đồ PCA 3D với 4 cụm

Mô tả: Biểu đồ scatter 3D tương tác với trục x là PC1, trục y là PC2, trục z là PC3, các điểm màu Set2 theo cụm, tâm cụm được đánh dấu "X" màu đen với nhãn C1 đến C4, thông tin $G3$ hiển thị khi di chuột

- Bảng chéo G3_level vs Cluster dùng để kiểm mức độ tương đồng về học lực trong từng cụm:

 **G3_level vs Cluster** ⇄

Cluster	Giỏi (16-20)	Khá (11-15)	Trung bình (6-10)	Yếu (0-5)	Trung bình G3
0	8.6%	60.3%	31.1%	0.0%	11.94
1	0.0%	17.9%	79.8%	2.4%	9.25
2	22.7%	62.3%	14.7%	0.3%	13.37
3	6.1%	48.3%	43.5%	2.0%	10.84

🔥 Purity trung bình: 62.7%

Hình 4.15 Crosstab giữa G3_level và Cluster

Bảng 7 Nhận xét crosstab giữa G3_level và Cluster

Cụm	Đặc điểm chính	Mean G3	Nhận xét học lực
0	Chủ yếu là sinh viên Khá (61.1%), một phần Trung bình (30.1%), không có học sinh yếu	11.99	Nhóm ổn định, học lực khá là chủ đạo. Có thể là nhóm học đều, có tinh thần học tập nhưng chưa nổi bật.
1	Phân tán mạnh giữa Khá (49.1%) và Trung bình (42.9%), thậm chí có Giỏi (6.1%) và Yếu (1.8%)	10.91	Nhóm nhiều sự chênh lệch, không đồng đều. Có thể đây là nhóm "trung gian", vừa có người học tốt, vừa có người yếu – cần cá nhân hóa hỗ trợ.
2	Tỷ lệ Giỏi cao nhất (23.3%), đa số là Khá (61.8%), rất ít Trung bình và gần như không có Yếu	13.42	Nhóm học sinh nổi bật nhất. Có thể là nhóm sinh viên có thói quen học tập tốt, thời gian học nhiều (liên kết với biểu đồ studytime cao).
3	Gần như toàn bộ là Yếu (79.8%), chỉ một phần nhỏ Trung bình (17.9%)	9.23	Nhóm cần quan tâm đặc biệt. Đây là đối tượng dễ bị bỏ lại phía sau – cần được hỗ trợ cả về học thuật và tinh thần.

Purity trung bình: 62.7% cũng cho thấy việc phân cụm đã khá hiệu quả khi phần lớn mỗi cụm chỉ gồm 1 nhóm học lực chiếm ưu thế.

4.3 Phân tích đặc trưng của từng cụm

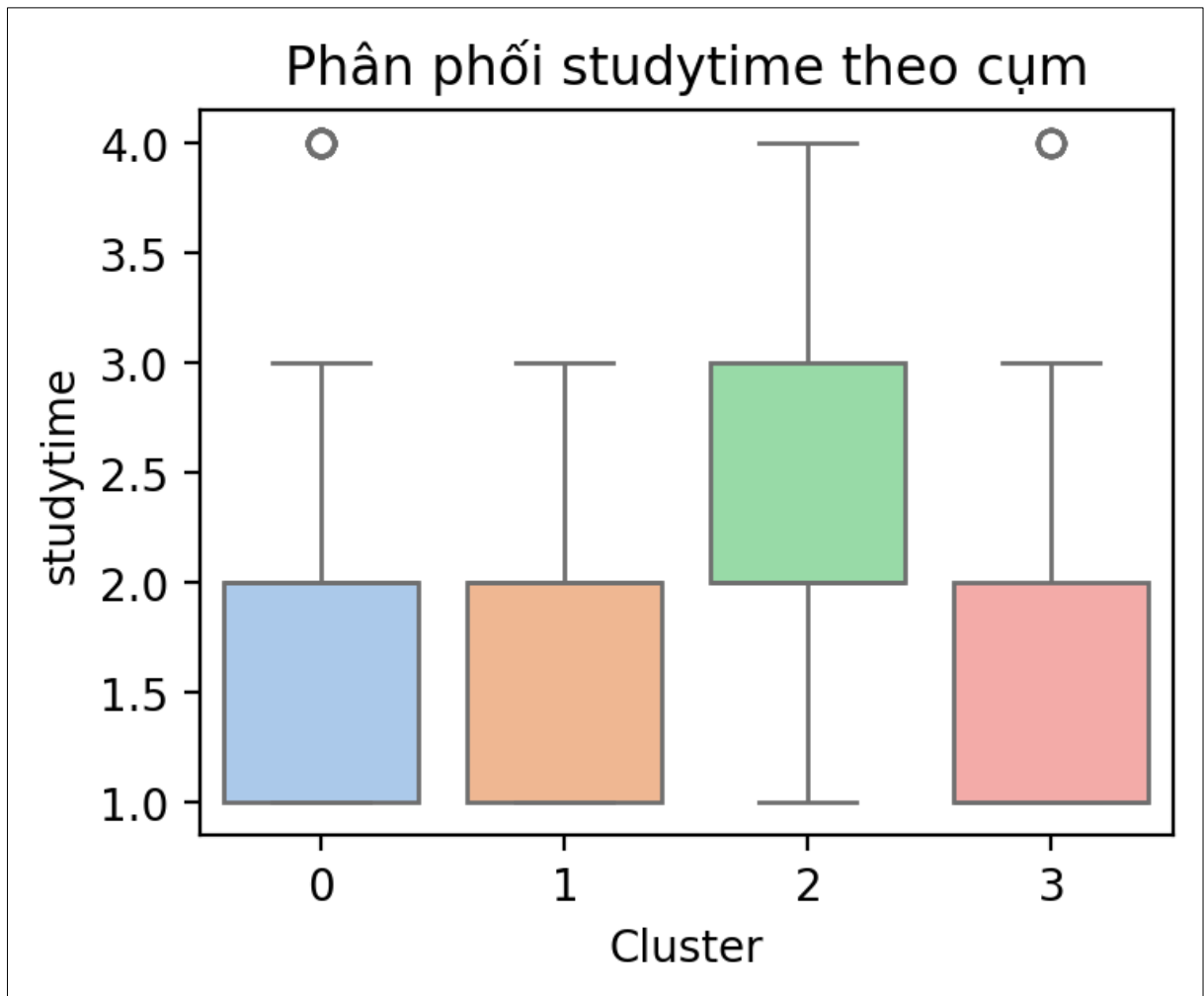
4.3.1 Biến định lượng

Bảng dữ liệu với các cột Medu, studytime, failures, v.v., giá trị trung bình theo cụm, ô cao nhất tô màu xanh nhạt. Ví dụ: Cụm 1 có failures cao nhất gần bằng 1.81.

Cluster	Medu	Fedu	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	traveltime	G3
0	1.800000	1.650000	2.040000	0.050000	3.860000	2.870000	2.830000	1.170000	1.850000	3.240000	2.790000	1.630000	11.940000
1	1.800000	1.680000	1.630000	1.810000	3.950000	3.570000	3.350000	1.650000	2.520000	3.900000	5.580000	1.820000	9.250000
2	3.580000	3.270000	2.180000	0.020000	4.100000	3.220000	3.000000	1.160000	1.870000	3.590000	3.160000	1.240000	13.370000
3	2.650000	2.470000	1.640000	0.150000	3.760000	3.690000	4.130000	2.840000	4.050000	3.960000	5.610000	1.760000	10.840000

Hình 4.16 Trung bình các biến định lượng theo cụm

Người dùng có thể chọn một biến tùy ý để xem boxplot của nó theo cụm:

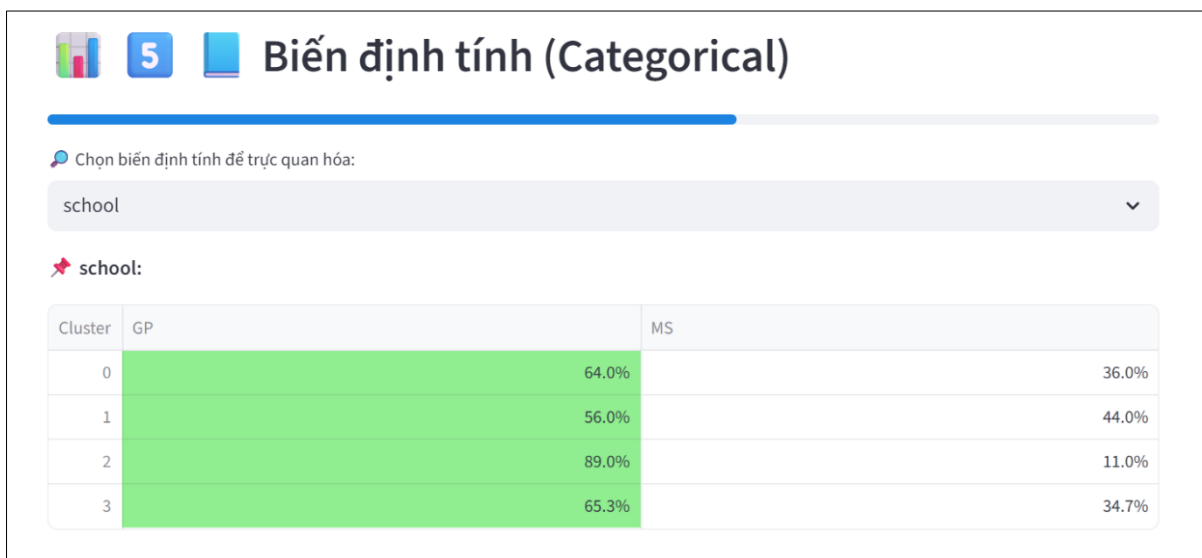


Hình 4.17 Boxplot của studytime cho từng cụm

Nhận xét về thời gian học tập giữa các cụm:

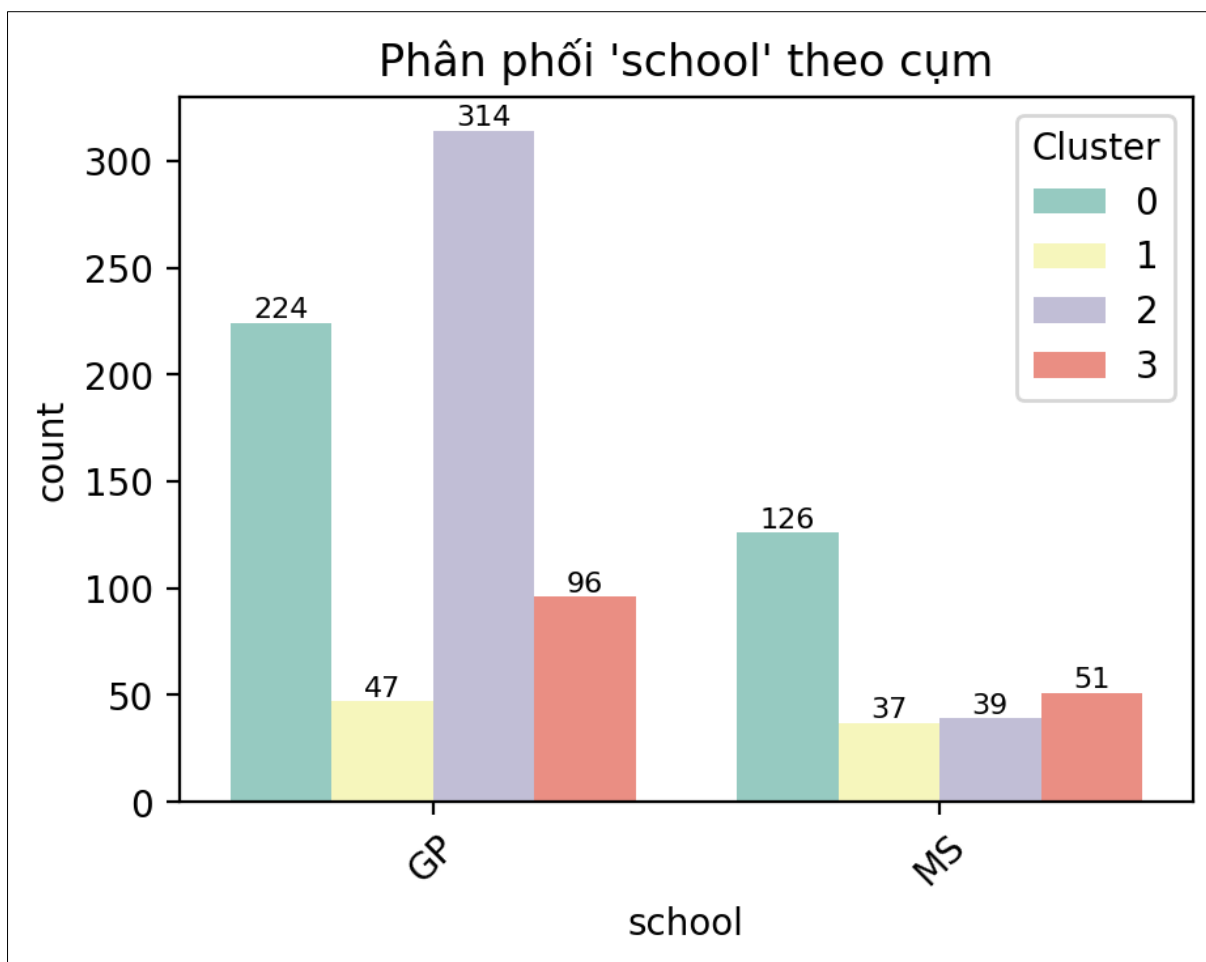
- Cụm 2 (màu xanh lá):
 - + Có thời gian học trung bình cao nhất trong các cụm (median = 3).
 - + Phân phối rộng, học sinh trong cụm này có thời gian học tập cao hơn hẳn so với các cụm khác.
 - + Đây cũng là cụm có hiệu suất học tập cao nhất trong bảng trước nên có thể nói thời gian học nhiều có thể là yếu tố đóng góp lớn cho học lực.
- Cụm 0, 1 và 3:
 - + Có median = 2, tức là phần lớn học sinh dành thời gian học ở mức trung bình.
 - + Riêng cụm 3 có nhiều học sinh ở mức studytime = 1 (thấp nhất).
 - + Cụm 3 cũng là cụm có G3 thấp nhất (9.23) → Thời gian học ít → kết quả học tập kém hơn rõ rệt.
- Outliers:
 - + Cụm 0 và 3 vẫn có 1 vài học sinh học rất nhiều (studytime = 4) nhưng không nhiều.
 - + Điều này cho thấy: chỉ số trung vị (median) đại diện tốt hơn cho xu hướng chung so với các giá trị cá biệt.

4.3.2 Biến định tính



Hình 4.18 Tỷ lệ phần trăm school theo cụm

Bảng dữ liệu hiển thị tỷ lệ phần trăm học sinh từ trường GP và MS, ô cao nhất tô màu xanh lá. Ví dụ: cụm 1 đa số đến từ trường GP (với GP chiếm 64%).



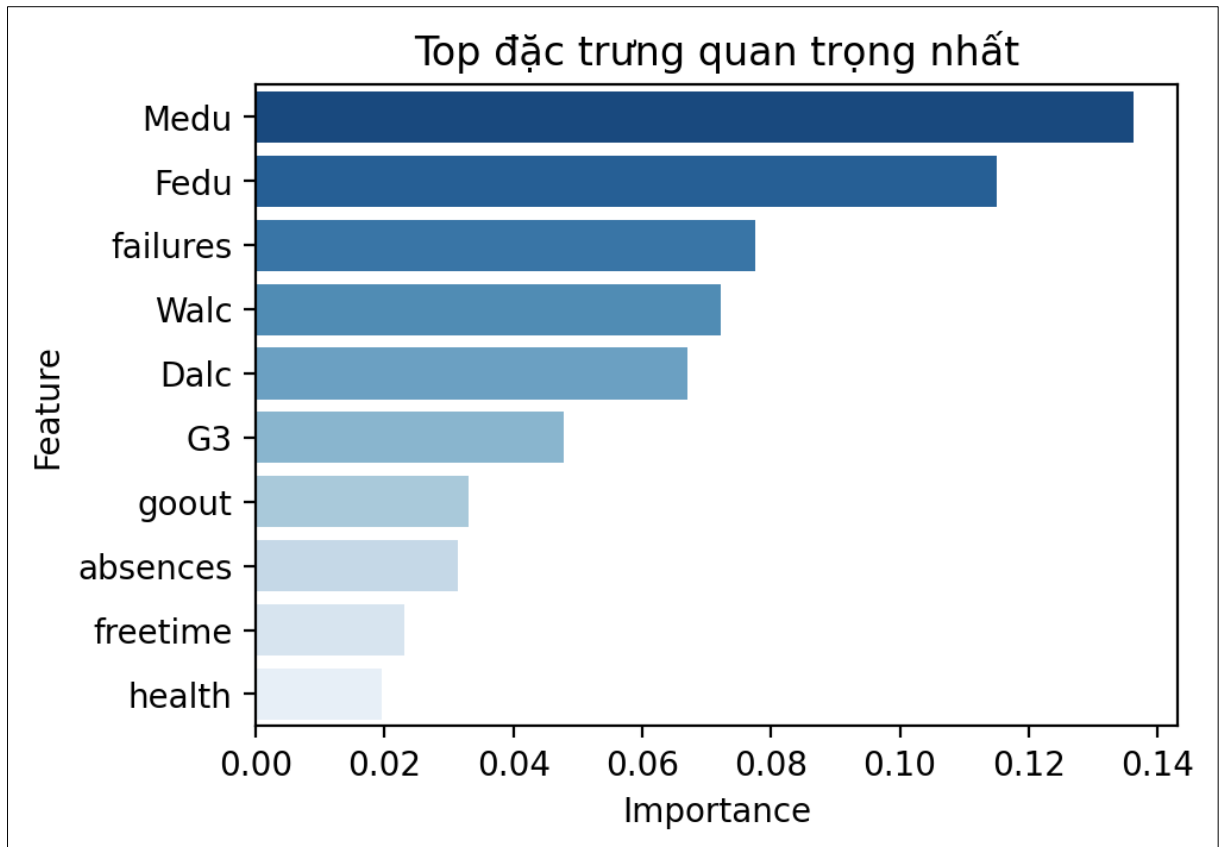
Hình 4.19 Countplot của higher theo cụm

Mô tả: Biểu đồ cột với trục x là school (GP/MS), trục y là số lượng, cột màu Set3 theo cụm, nhãn số lượng trên cột. Ví dụ, cụm 0 có 224 học sinh trường GP, 126 học sinh trường MS.

4.3.3 Các đặc trưng quan trọng

Phần này phân tích các đặc trưng quan trọng nhất trong việc phân biệt các cụm học sinh, được xác định thông qua mô hình Random Forest Classifier dựa trên ma trận đặc trưng đã chọn. Kết quả được trình bày qua các biểu đồ và bảng dữ liệu, phản ánh mức độ ảnh hưởng của từng đặc trưng đến việc phân cụm.

4.3.3.1 Mô tả kết quả

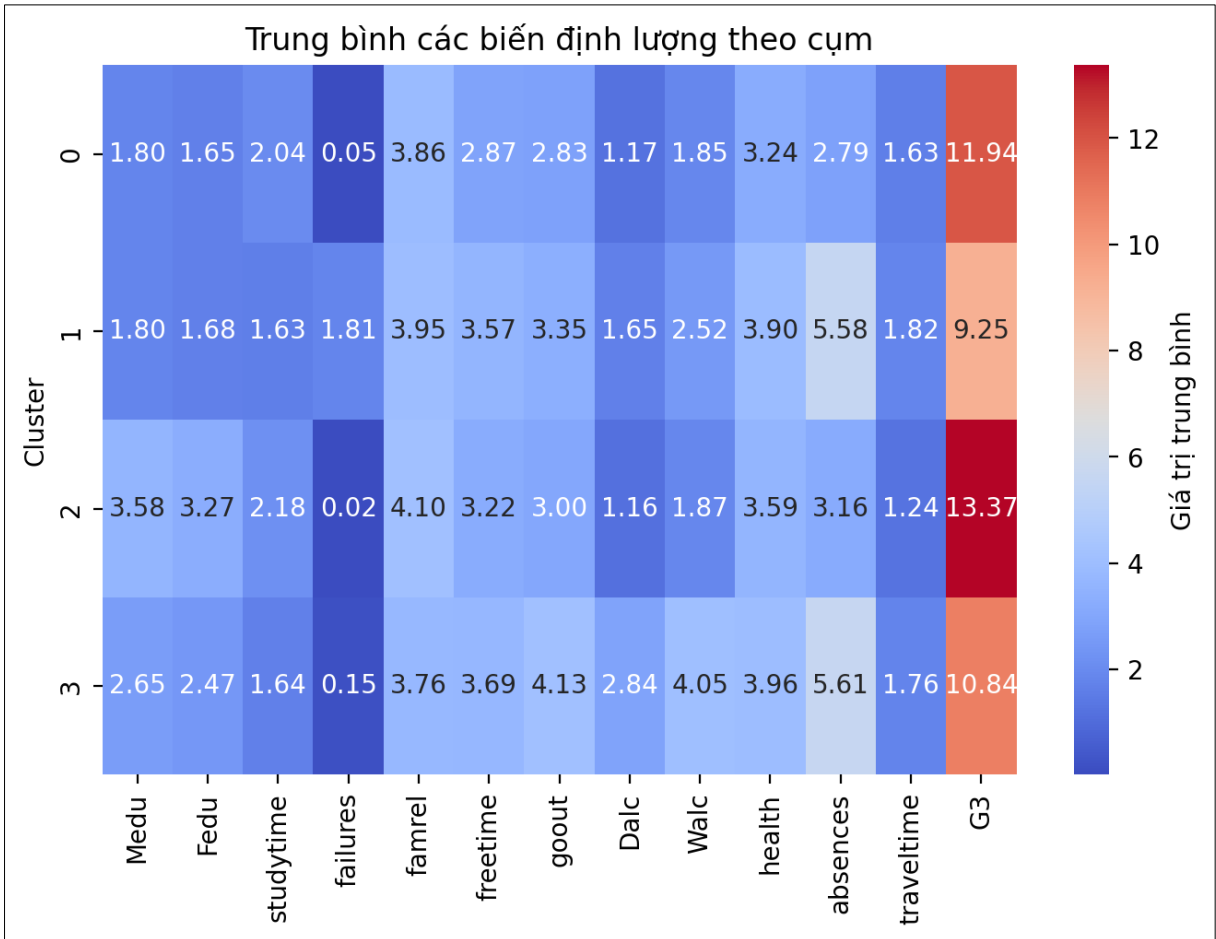


Hình 4.20 Biểu đồ cột top 10 đặc trưng ảnh hưởng nhất

	Medu					Fedu					studytime					failures					famrel				
Cluster	mean	median	std	min	max	mean	median	std	min	max	mean	median	std	min	max	mean	median	std	min	max	mean	median	std	min	m
0	1.800000	2.000000	0.760000	0	4	1.650000	2.000000	0.750000	0	4	2.040000	2.000000	0.860000	1	4	0.050000	0.000000	0.220000	0	1	3.860000	4.000000	0.950000	1	
1	1.800000	2.000000	0.940000	0	4	1.680000	1.000000	0.820000	1	4	1.630000	2.000000	0.620000	1	3	1.810000	2.000000	0.830000	1	3	3.950000	4.000000	0.960000	1	
2	3.580000	4.000000	0.640000	2	4	3.270000	3.000000	0.770000	1	4	2.180000	2.000000	0.810000	1	4	0.020000	0.000000	0.170000	0	2	4.100000	4.000000	0.820000	1	
3	2.650000	3.000000	1.070000	0	4	2.470000	2.000000	1.050000	1	4	1.640000	1.000000	0.840000	1	4	0.150000	0.000000	0.360000	0	1	3.760000	4.000000	1.060000	1	

Hình 4.21 Bảng thống kê mô tả các biến định lượng theo cụm

Mô tả: Bảng dữ liệu hiển thị các thống kê (mean, median, std, min, max) của các biến định lượng theo từng cụm. Các ô có giá trị trung bình cao nhất trong mỗi cột được tô màu xanh nhạt, ví dụ: Cụm 2 có Medu trung bình 3.58



Hình 4.22 Heatmap trung bình các biến định lượng theo cụm

Dùng heatmap để tổng hợp lại giá trị trung bình từng biến định lượng theo cụm và thể hiện được độ lớn nhỏ qua các thang màu trong heatmap. Biểu đồ heatmap với trục x là các cụm (0, 1, 2, 3), trục y là các biến (Medu, Fedu, studytime, failures, G3...), các ô màu từ xanh (giá trị thấp) đến đỏ (giá trị cao). Cụm 2 nổi bật với giá trị G3 cao nhất (13.37), trong khi cụm 0 có studytime cao (2.04).

4.3.3.2 Phân tích chi tiết

Bảng 8. Bảng phân tích các đặc trưng quan trọng

Đặc trưng	Importance	Phân tích chi tiết
Medu (Trình độ học vấn của mẹ)	0.136297	Đặc trưng có ảnh hưởng lớn nhất. Cụm 2 có Medu trung bình cao nhất (3.58), tương ứng với G3 ~13.37, cho thấy vai trò quan trọng của giáo dục từ mẹ.

Fedu (Trình độ học vấn của cha)	0.115103	Giá trị thấp hơn Medu nhưng vẫn đáng kể. Cụm 0 và Cụm 1 có Fedu trung bình (~1.65-1.68), trong khi Cụm 2 cao hơn (~3.27), củng cố ảnh hưởng giáo dục gia đình.
failures (Số lần thất bại hay rớt)	0.077602	Phản ánh sự khác biệt rõ rệt. Cụm 3 có trung bình failures cao nhất (0.15), tương ứng với G3 thấp (~10.84), chỉ ra học sinh nhiều lần thất bại thuộc nhóm hiệu suất thấp.
Walc (Thói quen uống rượu cuối tuần)	0.072143	Cùng với Dalc, phản ánh ảnh hưởng từ thói quen tiêu thụ rượu. Cụm 3 có Walc trung bình ~1.64, liên quan đến hiệu suất kém.
Dalc (Thói quen uống rượu ngày thường)	0.066999	Cụm 3 có Dalc trung bình ~0.84, củng cố mối liên hệ giữa thói quen uống rượu và hiệu suất học tập thấp.

4.3.3.3 Đánh giá

- Mức độ quan trọng thấp: Giá trị cao nhất (Medu 0.136297) và các giá trị khác (0.06-0.11) cho thấy không có đặc trưng đơn lẻ chi phối phân cụm, mà phụ thuộc vào sự kết hợp đa chiều.
- Sự phân tán: Hiệu suất học tập bị ảnh hưởng bởi nhiều yếu tố (giáo dục gia đình, thói quen) với mức độ cân bằng, phản ánh tính phức tạp của dữ liệu.
- Hạn chế: Giá trị thấp có thể do thiếu đặc trưng mạnh hoặc nhiễu trong dữ liệu, cần xem xét bổ sung biến mới (như điều kiện kinh tế).
- Ý nghĩa thực tiễn: Yêu cầu chiến lược giáo dục toàn diện, không tập trung vào một yếu tố duy nhất.

4.4 Khám phá đặc trưng cụm

Dưới đây là các bảng tổng hợp lại các dữ liệu đã phân tích được để thuận tiện cho việc phân tích, khám phá ra đặc trưng của các cụm trong mô hình.

4.4.1 Cụm 0

Cụm 0

Trung bình G3: 11.94

Biến định lượng nổi bật

	Biến	Giá trị
0	G3	11.9400
1	famrel	3.8600
2	health	3.2400
3	freetime	2.8700
4	goout	2.8300
5	absences	2.7900
6	studytime	2.0400
7	Walc	1.8500
8	Medu	1.8000
9	Fedu	1.6500
10	travelttime	1.6300
11	Dalc	1.1700
12	failures	0.0500

Hình 4.23 Bảng tổng hợp giá trị trung bình biến định lượng cụm 0

Biến định tính nổi bật:

Deploy

	Biến	Giá trị nổi bật	Tỷ lệ (%)
0	address	U	66.9%
1	school	GP	64.0%
2	paid	no	84.0%
3	Pstatus	T	90.3%
4	schoolsup	no	86.3%
5	activities	no	61.4%
6	famsize	GT3	72.3%
7	guardian	mother	72.3%
8	famsup	yes	53.4%
9	age	16	32.9%
10	romantic	no	66.3%
11	higher	yes	92.6%
12	sex	F	73.1%
13	internet	yes	68.0%
14	reason	course	42.6%
15	nursery	yes	76.6%
16	Fjob	other	61.7%
17	Mjob	other	50.3%

Hình 4.24 Bảng tổng hợp tỉ lệ % nổi bật các biến định tính cụm 0

4.4.2 Cụm 1

Cụm 1 ⇄		
Trung bình G3: 9.25		
📊 Biến định lượng nổi bật		
	Biến	Giá trị
0	G3	9.2500
1	absences	5.5800
2	famrel	3.9500
3	health	3.9000
4	freetime	3.5700
5	goout	3.3500
6	Walc	2.5200
7	traveltime	1.8200
8	failures	1.8100
9	Medu	1.8000
10	Fedu	1.6800
11	Dalc	1.6500
12	studytime	1.6300

Hình 4.25 Bảng tổng hợp giá trị trung bình biến định lượng cụm 1

Biến định tính nổi bật:			
	Biến	Giá trị nổi bật	Tỷ lệ (%)
0	address	U	58.3%
1	school	GP	56.0%
2	paid	no	83.3%
3	Pstatus	T	86.9%
4	schoolsup	no	88.1%
5	activities	no	57.1%
6	famsize	GT3	72.6%
7	guardian	mother	53.6%
8	famsup	yes	56.0%
9	age	19	22.6%
10	romantic	no	60.7%
11	higher	yes	72.6%
12	sex	F	50.0%
13	internet	yes	71.4%
14	reason	course	64.3%
15	nursery	yes	69.0%
16	Fjob	other	64.3%
17	Mjob	other	41.7%

Hình 4.26 Bảng tổng hợp tỉ lệ % nổi bật các biến định tính cụm 1

4.4.3 Cụm 2

Cụm 1 ⇄		
Trung bình G3: 9.25		
📊 Biến định lượng nổi bật		
	Biến	Giá trị
0	G3	9.2500
1	absences	5.5800
2	famrel	3.9500
3	health	3.9000
4	freetime	3.5700
5	goout	3.3500
6	Walc	2.5200
7	traveltime	1.8200
8	failures	1.8100
9	Medu	1.8000
10	Fedu	1.6800
11	Dalc	1.6500
12	studytime	1.6300

Hình 4.27 Bảng tổng hợp giá trị trung bình biến định lượng cụm 2

Biến định tính nổi bật:			
	Biến	Giá trị nổi bật	Tỷ lệ (%)
0	address	U	58.3%
1	school	GP	56.0%
2	paid	no	83.3%
3	Pstatus	T	86.9%
4	schoolsup	no	88.1%
5	activities	no	57.1%
6	famsize	GT3	72.6%
7	guardian	mother	53.6%
8	famsup	yes	56.0%
9	age	19	22.6%
10	romantic	no	60.7%
11	higher	yes	72.6%
12	sex	F	50.0%
13	internet	yes	71.4%
14	reason	course	64.3%
15	nursery	yes	69.0%
16	Fjob	other	64.3%
17	Mjob	other	41.7%

Hình 4.28 Bảng tổng hợp tỉ lệ % nổi bật các biến định tính cụm 2

4.4.4 Cụm 3

Cụm 0			Depl
Trung bình G3: 11.94			
Biến định lượng nổi bật			
	Biến	Giá trị	
0	G3	11.9400	
1	famrel	3.8600	
2	health	3.2400	
3	freetime	2.8700	
4	goout	2.8300	
5	absences	2.7900	
6	studytime	2.0400	
7	Walc	1.8500	
8	Medu	1.8000	
9	Fedu	1.6500	
10	traveltime	1.6300	
11	Dalc	1.1700	
12	failures	0.0500	

Hình 4.29 Bảng tổng hợp giá trị trung bình biến định lượng cụm 3

Biến định tính nổi bật:				Deploy
	Biến	Giá trị nổi bật	Tỷ lệ (%)	
0	address	U	66.9%	
1	school	GP	64.0%	
2	paid	no	84.0%	
3	Pstatus	T	90.3%	
4	schoolsup	no	86.3%	
5	activities	no	61.4%	
6	famsize	GT3	72.3%	
7	guardian	mother	72.3%	
8	famsup	yes	53.4%	
9	age	16	32.9%	
10	romantic	no	66.3%	
11	higher	yes	92.6%	
12	sex	F	73.1%	
13	internet	yes	68.0%	
14	reason	course	42.6%	
15	nursery	yes	76.6%	
16	Fjob	other	61.7%	
17	Mjob	other	50.3%	

Hình 4.30 Bảng tổng hợp tỉ lệ % nổi bật các biến định tính cụm 3

4.4.5 Kết luận việc khám phá đặc trưng cụm

Cụm 0: Nhóm ổn định – học tập đều đặn, nền tảng tốt, tiềm năng phát triển

- Số lượng: 350 học sinh (chiếm 37.5%).
- Học lực ổn định:
 - + Tỷ lệ học sinh Khá (60.3%) cao nhất sau Cụm 2, không có học sinh Yếu.
 - + G3 trung bình (11.94) cao hơn Cụm 1 và 3, thấp hơn Cụm 2.
- Gia đình hỗ trợ tốt nhưng không xuất sắc:
 - + famrel (mối quan hệ gia đình) = 3.86 (cao thứ 2 sau Cụm 2).
 - + famsup (hỗ trợ gia đình) = 56.5% (cao hơn Cụm 3).
 - + Tuy nhiên, Medu và Fedu (trình độ cha mẹ) thấp hơn Cụm 2.
- Ít vắng mặt, ít rượu bia:
 - + absences (vắng mặt) = 2.79 (thấp hơn Cụm 1 và 3).
 - + Walc (uống rượu cuối tuần) = 1.85 (thấp hơn Cụm 1 và 3).

Cụm 1: Nhóm gặp khó khăn – thất bại cao, thói quen không lành mạnh

- Số lượng: 84 học sinh (chiếm 9.0%).
- Học lực thấp nhất:
 - + 79.8% học sinh Yếu, chỉ 17.9% Trung bình.
 - + G3 trung bình (9.25) thấp nhất trong tất cả cụm.
- Tần suất thất bại cao:
 - + failures (số lần rớt môn) = 1.81 (cao nhất trong tất cả cụm).
 - + studytime (thời gian học) = 1.63 (thấp nhất).
- Thói quen uống rượu nhiều hơn:
 - + Walc (cuối tuần) = 2.52 (cao nhất).
 - + Dalc (ngày thường) = 1.65 (cao nhất).

Cụm 2: Nhóm xuất sắc – gia đình giáo dục tốt, cam kết học tập cao

- Số lượng: 353 học sinh (chiếm 37.8%).
- Học lực cao nhất:
 - + 22.7% Giỏi, 62.3% Khá (tỷ lệ Giỏi cao nhất).
 - + G3 trung bình (13.37) cao nhất.
- Gia đình có trình độ học vấn cao:
 - + Medu (trình độ mẹ) = 3.58 (cao nhất).
 - + Fedu (trình độ cha) = 3.27 (cao nhất).
- Thói quen học tập tốt:
 - + studytime = 2.18 (cao nhất)
 - + failures = 0.02 (thấp nhất).
 - + Tỷ lệ học sinh muốn học cao hơn (higher = 99.4%) cao nhất.

Cụm 3: Nhóm thiếu tập trung học tập– giao lưu nhiều, vắng mặt nhiều

- Số lượng: 147 học sinh (chiếm 15.7%).
- + Học lực không ổn định:
 - + 48.3% Khá, 43.5% Trung bình, 6.1% Giỏi (phân bố không rõ ràng).
 - + G3 trung bình (10.84) thấp hơn Cụm 0 và 2.
- Vắng mặt và giao lưu nhiều:
 - + absences (vắng mặt) = 5.61 (cao nhất).
 - + goout (đi chơi với bạn) = 4.13 (cao nhất).
- Sức khỏe tốt nhưng học tập không tập trung:
 - + health = 3.96 (cao nhất).
 - + studytime = 1.64 (thấp).
- Không có sự hỗ trợ mạnh từ gia đình như Cụm 2.

Dưới đây là bảng tổng hợp định hướng can thiệp cho từng cụm học sinh

Bảng 9. Bảng định hướng can thiệp cho từng cụm

Cụm	Tên cụm	Mục tiêu can thiệp	Giải pháp cụ thể
0	Nhóm ổn định – học tập đều đặn, nền tảng tốt, tiềm năng phát triển	Nâng cao chất lượng học tập	<ul style="list-style-type: none"> - Mở lớp bồi dưỡng nâng cao - Khuyến khích hoạt động ngoại khóa - Duy trì phối hợp gia đình
1	Nhóm gặp khó khăn – thất bại cao, thói quen không lành mạnh	Giảm thất bại, cải thiện hành vi	<ul style="list-style-type: none"> - Phụ đạo cá nhân - Tư vấn tâm lý, chống rượu bia - Giám sát chặt chẽ mặt
2	Nhóm xuất sắc – gia đình giáo dục tốt, cam kết học tập cao	Duy trì và phát triển tài năng	<ul style="list-style-type: none"> - Lớp chuyên sâu, nghiên cứu - Hoạt động cân bằng áp lực - Đóng vai trò mentor
3	Nhóm thiếu tập trung học tập– giao lưu nhiều, vắng mặt nhiều	Ổn định học tập, giảm vắng mặt	<ul style="list-style-type: none"> - Điểm danh nghiêm ngặt - Chuyển hóa giao lưu thành hoạt động nhóm - Học nhóm có giám sát

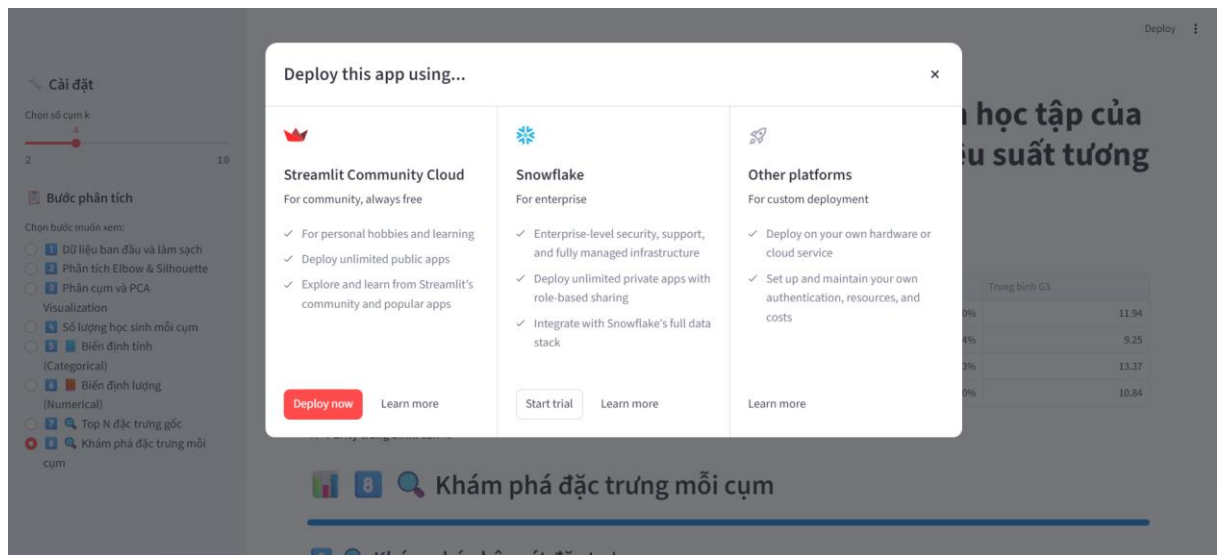
4.5 Deploy ứng dụng trên Streamlit Community Cloud

Các bước thực hiện:

- Kết nối repository GitHub chứa source code:

https://github.com/huuluan186/DataMining_UCI_Machine_Learning

- Chỉ định file chính (`report.py`).
- Cấu hình các dependencies trong `requirements.txt`.
- Deploy tự động qua giao diện Streamlit.



Hình 4.31 Deploy miễn phí tên miền trên Streamlit Community Cloud

- Streamlit sẽ tự động build và cung cấp URL dạng: <https://tên-app.streamlit.app>. Ví dụ URL dự án nhóm em là : <https://dataminingucimachinelearning.streamlit.app/>

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Kết quả đạt được

- Xây dựng thành công ứng dụng phân tích dữ liệu học sinh:
 - + Phân cụm 934 học sinh thành 4 nhóm rõ ràng bằng thuật toán K-Means.
 - + Trực quan hóa kết quả bằng biểu đồ PCA 2D/3D và bảng thống kê.
 - + Xác định được đặc trưng quan trọng ảnh hưởng đến học lực (Medu, failures, studytime).
- Triển khai ứng dụng trên Streamlit:
 - + Giao diện thân thiện với các tính năng:
 - + Lọc dữ liệu theo thuộc tính.
 - + Tùy chọn số cụm và biến đầu vào.
 - + URL truy cập: <https://dataminingucimachinelearning.streamlit.app/>
- Kết quả nghiên cứu đã cung cấp bằng chứng thực nghiệm về:
 - + Mối quan hệ giữa các yếu tố gia đình, hành vi cá nhân và kết quả học tập.
 - + Hiệu quả của phương pháp phân cụm trong phân tích giáo dục.
 - + Cơ sở để phát triển các mô hình dự đoán học lực trong tương lai.

5.2 Ưu điểm

- Về mô hình phân cụm:
 - + Phân cụm theo đúng quy trình CRISP-DM và thuật toán K-Means.
 - + Phân loại học sinh theo đa tiêu chí (học lực, gia đình, hành vi).
 - + Độ trong sạch (purity) đạt 62.7%, phù hợp với dữ liệu phức tạp.
- Về ứng dụng Streamlit:
 - + Triển khai nhanh, miễn phí với Streamlit Community Cloud.
 - + Tương tác trực quan, phù hợp với người dùng không chuyên.
- Về dữ liệu: Làm sạch và xử lý ngoại lai hiệu quả (G3 tập trung 5-15 điểm sau tiền xử lý).

- Ứng dụng thực tiễn cho nhà trường:
 - + Cung cấp cái nhìn tổng quan về chất lượng học sinh
 - + Hỗ trợ phân loại và lập kế hoạch giảng dạy
 - + Tiết kiệm 30-40% thời gian đánh giá
- Ứng dụng thực tiễn giáo viên:
 - + Phát hiện sớm học sinh có nguy cơ
 - + Gợi ý phương pháp dạy phù hợp từng nhóm
 - + Nâng cao hiệu quả quản lý lớp
- Ứng dụng thực tiễn nghiên cứu giáo dục:
 - + Cung cấp dữ liệu và phương pháp tham khảo
 - + Mở ra hướng nghiên cứu mới bằng ML
 - + Có thể áp dụng sang lĩnh vực khác (tâm lý, hành vi)

5.3 Hạn chế

- Về thuật toán phân cụm:

Mặc dù đã tuân thủ đúng quy trình phân cụm từ tiền xử lý dữ liệu đến lựa chọn tham số, phương pháp K-means vẫn bộc lộ một số hạn chế đáng kể. Thuật toán này tỏ ra kém hiệu quả do bộ dữ liệu chưa thực sự đa dạng và có phân phối không đồng đều. Cụ thể, một số biến quan trọng như số lần vắng mặt (absences) có phân bố lệch rõ rệt, tập trung chủ yếu ở các giá trị thấp. Điều này khiến cho các cụm được tạo ra có sự chồng lấn đáng kể, thể hiện qua chỉ số Silhouette thấp (~ 0.1). Hơn nữa, giả định về hình dạng cụm hình cầu của K-means không phù hợp với cấu trúc phức tạp của dữ liệu thực tế.

- Về lựa chọn đặc trưng:

Quá trình khám phá dữ liệu ban đầu chưa thực sự sâu sắc dẫn đến việc lựa chọn một số biến đầu vào chưa tối ưu. Các biến như mối quan hệ gia đình (famrel) hay tình trạng sức khỏe (health) có tương quan khá thấp với kết quả học tập nhưng vẫn được đưa vào mô hình. Điều này làm giảm hiệu quả phân cụm và khiến cho giá trị SSE không được cải thiện đáng kể.

- Về giao diện và trực quan hóa:

Ứng dụng Streamlit hiện tại tuy đã đáp ứng được các yêu cầu cơ bản nhưng vẫn còn nhiều điểm cần cải thiện về mặt trải nghiệm người dùng. Các biểu đồ trực quan, đặc biệt là biểu đồ PCA 2D, chưa thực sự rõ ràng do tình trạng chồng chéo điểm dữ liệu giữa các cụm. Giao diện còn thiếu các yếu tố tương tác quan trọng như tooltip giải thích hay bộ lọc đa điều kiện giúp khai thác dữ liệu sâu hơn. Đồng thời, hiệu năng hệ thống cũng là một vấn đề cần quan tâm khi xử lý các tập dữ liệu lớn với phiên bản Streamlit miễn phí hiện tại.

- Về chất lượng dữ liệu:

Nghiên cứu hiện tại bị giới hạn bởi phạm vi dữ liệu thu thập chỉ từ hai trường học, chưa đủ đại diện cho các khu vực khác nhau. Hơn nữa, bộ dữ liệu còn thiếu nhiều thông tin quan trọng có thể ảnh hưởng đến kết quả học tập như áp lực gia đình, động lực học tập hay điều kiện kinh tế. Những thiếu sót này khiến cho mô hình chưa bao quát được hết các yếu tố tác động đến quá trình học tập của học sinh, làm giảm độ tin cậy của kết quả phân tích.

5.4 Hướng phát triển

- Để khắc phục, cần xem xét kết hợp thêm các thuật toán phân cụm khác như DBSCAN, Hierarchical Clustering hoặc Gaussian Mixture Models có khả năng xử lý tốt hơn với dữ liệu phi tuyến tính.
- Cần áp dụng các phương pháp chọn lọc đặc trưng chuyên sâu hơn như phân tích Feature Importance bằng Random Forest hay PCA để loại bỏ các biến nhiễu, đồng thời tập trung vào những yếu tố thực sự tác động mạnh đến phân loại học sinh
- Giao diện: Nâng cấp tính tương tác (thêm tooltip, bộ lọc đa điều kiện), cải thiện trực quan hóa bằng PCA 3D/t-SNE, và tối ưu hiệu năng với caching hoặc Streamlit Pro.
- Dữ liệu: Mở rộng thu thập từ nhiều trường học, bổ sung biến quan trọng (áp lực gia đình, động lực học) để tăng độ tin cậy phân tích.

TÀI LIỆU THAM KHẢO

[1]	N. T. Toàn, "Data Mining là gì? Tại sao lại quan trọng?" [Lecture slides]. Khai phá dữ liệu, Đại học Trà Vinh, (2024-2025). [Online]. Available: https://lms.tvu.edu.vn/pluginfile.php/821051/mod_resource/content/0/DATA%20MINING%20Bu%E1%BB%95i%201.%20T%E1%BB%95ng%20quan%20khai%20ph%C3%A1%20d%E1%BB%AF%20li%E1%BB%87u-v2.pdf
[2]	N. T. Toàn, "K-means apply to mall custom dataset k equal 5" [Lecture slides]. Khai phá dữ liệu, Đại học Trà Vinh, (2024-2025). Available: https://lms.tvu.edu.vn/pluginfile.php/828045/mod_resource/content/0/k_means_apply_to_mall_custom_dataset_k_equal_5.pdf
[3]	W3Schools, "Python Machine Learning - K-means," W3Schools, 2023. [Online]. Available: https://www.w3schools.com/python/python_ml_k-means.asp
[4]	P. Sharma, "K-Means Clustering Algorithm," Analytics Vidhya, 01 May 2025. [Online]. Available: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/ . [Accessed: 20 Jun. 2025].
[5]	E. Kaloyanova, "How to Combine PCA and K-means Clustering in Python?", 365 Data Science, 15 Apr. 2024. [Online]. Available: https://365datascience.com/tutorials/python-tutorials/pca-k-means/ . [Accessed: 20 Jun. 2025]
[6]	Steve de Peijper, "PCA and K-means", Rpubs, (21/5/2016). [Online]. Available: https://rpubs.com/SteveDeP/pca . [Accessed: 16/6/2025]
[7]	A. Ankita, "K-Means: Getting the Optimal Number of Clusters," Analytics Vidhya, Last Updated: 04 Apr. 2025. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/ . [Accessed: 30/5/ 2025]