

NHẬN DIỆN CHÂN DUNG THÍ SINH TRONG KỶ NGUYÊN SỐ QUA MÔ HÌNH PHÂN CỤM K-MEANS

Tô Vĩnh Bảo ⁽¹⁾

(1) Viện Công nghệ Số, Trường Đại học Thủ Dầu Một

Email tác giả: baotv@tdmu.edu.vn

Ngày nhận bài: 18/5/2026; Chấp nhận đăng: 23/6/2025

DOI: 10.37550/tdmu.EJS/V2026.03.764

Tóm tắt

Trong bối cảnh cạnh tranh giáo dục đại học, việc hiểu rõ hành vi thí sinh là yếu tố sống còn. Nghiên cứu ứng dụng khai phá dữ liệu trên 17.663 hồ sơ thực tế tại Trường Đại học Thủ Dầu Một để phân đoạn hành vi người học. Quy trình thực hiện qua ba giai đoạn: (1) Áp dụng K-Means kết hợp hệ số Silhouette xác định cấu trúc bốn phân đoạn tối ưu; (2) Sử dụng PCA để giảm chiều và trực quan hóa không gian dữ liệu; (3) Kiểm chứng bằng Random Forest Classifier và phân tích Feature Importance.

Kết quả thực nghiệm cho thấy các phân đoạn được hình thành có mức độ tách biệt cao trong không gian đặc trưng. Bước đánh giá nội bộ bằng Random Forest đạt độ chính xác 99,28% trong việc tái nhận diện nhãn cụm, cho thấy cấu trúc phân đoạn có tính nhất quán trên tập dữ liệu nghiên cứu. Nghiên cứu định danh thành công 4 nhóm đặc trưng: (0) Tin tưởng nội bộ & Chủ động số; (1) Tiềm năng thụ động; (2) Tương tác đa kênh tích cực; (3) Tập trung nội dung số. Phân tích trọng số thuộc tính khẳng định Website TDMU (23,36%) và sự giới thiệu từ người thân (13,67%) là hai nhân tố then chốt định hình hành vi. Những phát hiện này cung cấp cơ sở khoa học giúp nhà trường chuyển đổi từ chiến lược tư vấn đại trà sang cá nhân hóa, tối ưu hóa nguồn lực truyền thông trong kỷ nguyên số.

Từ khóa: Hành vi thí sinh, khai phá dữ liệu giáo dục, PCA, K-Means, Random Forest.

Abstract

IDENTIFYING DIGITAL-AGE CONTESTANT PROFILES THROUGH THE K-MEANS CLUSTERING MODEL

In the context of competitive higher education, understanding student behavior is crucial. This study applies data mining to 17,663 real-world records at Thu Dau Mot University to segment student behavior. The process is carried out in three stages: (1) Applying K-Means combined with Silhouette coefficients to determine the optimal 4-segment structure; (2) Using PCA to reduce dimensionality and visualize the data space; (3) Verification using Random Forest Classifier and Feature Importance analysis.

Experimental results show that the formed segments have a high degree of separation in the feature space. The internal evaluation step using Random Forest achieved 99.28% accuracy in re-identifying cluster labels, showing that the segment structure is consistent across the research dataset. The study successfully identified 4 feature groups: (0) Internal Trust & Digital Proactiveness; (1) Passive Potential; (2) Positive Multi-Channel Interaction; (3) Focus on digital content. Attribute weight

analysis confirms TDMU Website (23.36%) and referrals from relatives (13.67%) are two key factors shaping behavior. These findings provide a scientific basis to help the university shift from a mass counseling strategy to a personalized one, optimizing communication resources in the digital age.

Keywords: *Candidate behavior, Educational data mining, K-Means, Random Forest, PCA*

1. Đặt vấn đề

Trong kỷ nguyên chuyển đổi số, ngành giáo dục đại học đang đối mặt với những thách thức chưa từng có trong công tác tuyển sinh. Sự bùng nổ của các kênh truyền thông đa phương tiện đã thay đổi căn bản cách thức thí sinh tiếp cận và tương tác với các cơ sở giáo dục. Thí sinh ngày nay không còn phụ thuộc duy nhất vào các phương thức truyền thống mà chủ động khai phá thông tin qua hệ thống website, mạng xã hội, livestream và các ngày hội tư vấn trực tuyến. Sự đa dạng này tạo ra một khối lượng dữ liệu hành vi khổng lồ, hay còn gọi là “dấu vết số”, chứa đựng những thông tin giá trị về nhu cầu và tâm lý của người học tiềm năng.

Tuy nhiên, thực tế tại nhiều cơ sở giáo dục đại học hiện nay, trong đó có Trường Đại học Thủ Dầu Một, công tác quản trị dữ liệu tuyển sinh vẫn chủ yếu dừng lại ở mức độ thống kê mô tả cơ bản. Việc áp dụng một chiến lược truyền thông “đại trà” cho toàn bộ hàng chục ngàn thí sinh không chỉ gây lãng phí nguồn lực mà còn làm giảm khả năng tương tác cá nhân hóa – yếu tố then chốt quyết định sự thành công trong việc thu hút người học. Thách thức đặt ra là làm thế nào để phân loại được các nhóm thí sinh có hành vi tương đồng từ tập dữ liệu đa kênh phức tạp, từ đó đưa ra các kịch bản tư vấn phù hợp cho từng đối tượng cụ thể.

Khai phá dữ liệu giáo dục (Educational Data Mining - EDM) đã nổi lên như một giải pháp đột phá nhằm giải quyết bài toán này. Trong đó, phân cụm dữ liệu (Clustering) là một kỹ thuật quan trọng giúp nhận diện các cấu trúc tiềm ẩn trong dữ liệu mà không cần dán nhãn trước. Thuật toán K-Means với ưu điểm về tốc độ xử lý và tính hiệu quả trên tập dữ liệu lớn đã được ứng dụng rộng rãi trong nhiều lĩnh vực nghiên cứu hành vi khách hàng. Tuy nhiên, một hạn chế lớn của các nghiên cứu phân đoạn hành vi hiện nay là thiếu đi bước kiểm chứng tính bền vững của các phân đoạn. Việc chỉ dừng lại ở kết quả phân cụm có thể dẫn đến những nhận định cảm tính nếu không có sự đối chiếu với các mô hình học máy có giám sát.

Xuất phát từ thực trạng đó, nghiên cứu này được thực hiện với mục tiêu: “Nhận diện chân dung thí sinh kỹ nguyên số qua mô hình phân cụm K-Means”. Nghiên cứu sử dụng tập dữ liệu thực chứng gồm 17.663 bản ghi để trả lời các câu hỏi: (1) Làm thế nào để xác định số lượng phân đoạn thí sinh tối ưu trong không gian dữ liệu đa chiều? (2) Những đặc trưng hành vi nào đóng vai trò quyết định trong việc định hình chân dung thí sinh? và (3) Độ tin cậy của các phân đoạn này khi được kiểm chứng bằng mô hình học máy là bao nhiêu?

Tuy nhiên, việc phân đoạn hành vi thí sinh trong môi trường tuyển sinh số đặt ra nhiều thách thức về mặt phân tích dữ liệu. Dữ liệu hành vi tuyển sinh thường được thu thập từ nhiều điểm chạm khác nhau như website, mạng xã hội, livestream, ngày hội tuyển sinh, tư vấn trực tiếp và giới thiệu từ người thân. Đặc trưng của dữ liệu này là tính đa chiều, mức độ tương quan giữa các biến cao và tồn tại nhiều tín hiệu nhiễu, khiến các phương pháp phân cụm truyền thống gặp khó khăn trong việc nhận diện cấu trúc tiềm

ẩn của dữ liệu.

Trong bối cảnh đó, việc kết hợp giữa K-Means, PCA và Random Forest được xem là một hướng tiếp cận phù hợp. PCA giúp giảm chiều dữ liệu và hạn chế ảnh hưởng của hiện tượng đa cộng tuyến giữa các biến hành vi; K-Means hỗ trợ khám phá các nhóm đối tượng có đặc điểm tương đồng mà không cần dữ liệu gán nhãn; trong khi Random Forest được sử dụng như một công cụ đánh giá mức độ nhất quán của cấu trúc phân cụm và phân tích tầm quan trọng của thuộc tính. Cách tiếp cận lai này cho phép nâng cao khả năng diễn giải kết quả và hỗ trợ xây dựng các chiến lược tư vấn tuyển sinh cá nhân hóa.

2. Tổng quan tình hình nghiên cứu

2.1. Phân khúc người học trong giáo dục đại học

Phân khúc người học (Student Segmentation) từ lâu đã được công nhận là một chiến lược quan trọng trong quản trị giáo dục. Theo các nghiên cứu kinh điển, thay vì coi sinh viên là một khối đồng nhất, việc chia nhỏ thị trường mục tiêu dựa trên nhân khẩu học, tâm lý học và hành vi giúp các trường đại học tối ưu hóa các nỗ lực chiêu sinh. Tuy nhiên, trong môi trường số hiện nay, các biến số truyền thống như địa lý hay kết quả học tập không còn đủ sức mạnh để giải thích các quyết định phức tạp của thí sinh. Các nghiên cứu của Baker và Yacef (2009), Baker và Inventado (2014) đã mở ra hướng đi mới về khai phá dữ liệu giáo dục (EDM), tập trung vào việc trích xuất các tri thức ẩn từ cơ sở dữ liệu lớn của nhà trường để dự báo và phân loại hành vi. Xu hướng này đặc biệt phù hợp trong bối cảnh môi trường giáo dục số phát triển mạnh, nơi dữ liệu tương tác đa kênh ngày càng trở thành nguồn tài nguyên chiến lược cho quản trị đại học hiện đại.

2.2. Thuật toán K-Means và bài toán phân cụm hành vi

K-Means hiện là một trong những thuật toán học máy không giám sát phổ biến nhất trong các nghiên cứu về phân đoạn khách hàng và thí sinh. Ưu điểm của K-Means nằm ở khả năng xử lý các tập dữ liệu có quy mô lớn với tốc độ hội tụ nhanh. Nhiều nghiên cứu trước đây đã chứng minh K-Means có khả năng nhóm các đối tượng có điểm chạm (touch-points) tương đồng thành các cụm có ý nghĩa thực tiễn. Tuy nhiên, thách thức lớn nhất của thuật toán này là việc xác định số lượng cụm k tối ưu và tính nhạy cảm với dữ liệu nhiễu. Để giải quyết vấn đề này, các nghiên cứu hiện đại thường kết hợp với phương pháp Hệ số *Silhouette* để đo lường mức độ gắn kết bên trong cụm và sự tách biệt giữa các cụm, đảm bảo tính khách quan cho mô hình (Rousseeuw, 1987).

2.3. Sự kết hợp giữa giảm chiều dữ liệu (PCA) và học máy có giám sát

Một xu hướng nổi bật trong các nghiên cứu EDM gần đây là việc sử dụng phương pháp lai (hybrid). Do dữ liệu hành vi thí sinh thường bao gồm rất nhiều kênh tương tác (đa kênh), hiện tượng đa cộng tuyến và đa chiều thường gây khó khăn cho việc phân cụm. Kỹ thuật phân tích thành phần chính (PCA) đã được chứng minh là công cụ hiệu quả để giảm số lượng biến số nhưng vẫn giữ lại phần lớn thông tin cốt lõi, giúp việc phân cụm trở nên chính xác và dễ trực quan hóa hơn (Jolliffe & Cadima, 2016).

Bên cạnh đó, việc sử dụng các mô hình học máy có giám sát như Random Forest Classifier để kiểm chứng lại kết quả của học máy không giám sát là một bước tiến mới về mặt phương pháp luận. Random Forest, với cấu trúc gồm hàng trăm cây quyết định

độc lập, không chỉ giúp đánh giá độ tin cậy của các cụm đã phân loại mà còn cung cấp khả năng phân tích độ quan trọng thuộc tính (Feature Importance). Điều này giúp các nhà quản trị giáo dục hiểu rõ tại sao một thí sinh lại thuộc về một phân khúc nhất định, từ đó có cơ sở thực chứng để xây dựng các chiến lược tác động hiệu quả (Breiman, 2001).

Ngoài các công trình nền tảng, nhiều nghiên cứu gần đây trong lĩnh vực khai phá dữ liệu đã nhấn mạnh vai trò của dữ liệu lớn và học máy trong việc tối ưu hóa các hoạt động giáo dục. Han et al. (2012) đã hệ thống hóa các kỹ thuật khai phá dữ liệu quan trọng, bao gồm phân cụm, phân loại, luật kết hợp và trực quan hóa dữ liệu. Bên cạnh đó, Witten et al. (2016) tập trung trình bày các phương pháp học máy và khai phá dữ liệu dưới góc độ ứng dụng thực tiễn, hỗ trợ hiệu quả cho việc phân tích và khám phá tri thức từ dữ liệu. Đặc biệt, Pedregosa et al. (2011) đã giới thiệu thư viện Scikit-learn như một nền tảng mã nguồn mở mạnh mẽ, góp phần phổ biến việc triển khai các thuật toán học máy trong nghiên cứu và ứng dụng khai phá dữ liệu.

Các nghiên cứu gần đây cho thấy xu hướng ứng dụng học máy và khai phá dữ liệu trong giáo dục đang chuyển dịch từ các mô hình dự báo truyền thống sang các phương pháp có khả năng giải thích và hỗ trợ ra quyết định. Alvarez-Garcia et al. (2024) đã áp dụng các kỹ thuật phân tích phân cụm có khả năng giải thích để xây dựng chân dung người học dựa trên bộ dữ liệu PISA 2022, qua đó cung cấp cơ sở cho việc nhận diện các nhóm học sinh có đặc điểm học tập tương đồng. Bên cạnh đó, Susnjak (2024) nhấn mạnh vai trò của trí tuệ nhân tạo có khả năng giải thích (Explainable Artificial Intelligence – XAI) và phân tích dự báo trong việc cá nhân hóa các can thiệp giáo dục, góp phần nâng cao hiệu quả hỗ trợ người học. Đồng thời, Matz et al. (2023) đã chứng minh rằng dữ liệu tương tác số và dữ liệu hành vi đa nguồn có thể được khai thác hiệu quả để nhận diện đặc điểm cá nhân cũng như dự báo khả năng duy trì học tập của người học. Các kết quả này cho thấy tiềm năng của việc kết hợp các kỹ thuật phân cụm, giảm chiều và học máy trong việc xây dựng chân dung người học dựa trên dữ liệu đa kênh.

Tuy nhiên, phần lớn các nghiên cứu tập trung vào dữ liệu học tập trực tuyến hoặc thương mại điện tử giáo dục, trong khi các nghiên cứu khai thác dữ liệu hành vi tuyển sinh đa kênh tại Việt Nam vẫn còn hạn chế, khoảng trống nghiên cứu hiện nay nằm ở việc thiếu các mô hình phân đoạn hành vi thí sinh dựa trên dữ liệu tuyển sinh đa kênh quy mô lớn và được kiểm chứng bằng các kỹ thuật học máy hỗ trợ diễn giải kết quả. Nghiên cứu này hướng tới lấp đầy khoảng trống đó, thông qua việc xây dựng quy trình phân tích tích hợp K-Means, PCA và Random Forest trên tập dữ liệu gồm 17.663 hồ sơ thực tế.

3. Phương pháp nghiên cứu

Quy trình nghiên cứu được triển khai theo hướng tiếp cận khai phá dữ liệu (Data Mining) kết hợp học máy không giám sát và có giám sát, nhằm xây dựng hệ thống phân đoạn hành vi thí sinh một cách khách quan.

3.1. Thu thập và tiền xử lý dữ liệu

Nghiên cứu sử dụng tập dữ liệu thực chứng quy mô lớn từ hệ thống quản lý tuyển sinh của Trường Đại học Thủ Dầu Một (TDMU) giai đoạn 2023-2025, bao gồm 17.663 bản ghi. Tập dữ liệu chứa thông tin đa chiều về phương thức xét tuyển, địa điểm

thường trú và đặc biệt là các biến định lượng về hành vi tương tác trên đa kênh (Website, Facebook, YouTube, Ngày hội tuyển sinh, Tư vấn trực tiếp, Bạn bè người thân giới thiệu, Phương tiện truyền thông đại chúng, Livestream tư vấn,...).

Quy trình tiền xử lý được thực hiện qua các bước:

- **Làm sạch:** Xử lý các giá trị thiếu bằng phương pháp điền giá trị phổ biến (Mode) cho biến định danh.

- **Mã hóa dữ liệu (Label Encoding):** Các thuộc tính định tính như phương thức tuyển sinh, ngành học và nơi thường trú được chuyển đổi sang dạng số thông qua kỹ thuật Label Encoding nhằm thuận tiện cho quá trình xử lý dữ liệu trên tập mẫu lớn gồm 17.663 bản ghi. Việc lựa chọn phương pháp này xuất phát từ yêu cầu tối ưu hiệu năng tính toán và hạn chế hiện tượng gia tăng số chiều dữ liệu khi áp dụng One-Hot Encoding đối với các thuộc tính có nhiều mức giá trị. Tuy nhiên, nghiên cứu nhận thức rằng Label Encoding có thể tạo ra quan hệ thứ bậc giả định giữa các giá trị danh nghĩa khi thuật toán K-Means sử dụng khoảng cách Euclidean. Vì vậy, kết quả phân cụm trong nghiên cứu cần được xem xét trong phạm vi của giả định này.

- **Chuẩn hóa (Standardization):** Áp dụng kỹ thuật StandardScaler để đưa các thuộc tính về cùng một thang đo với kỳ vọng bằng 0 và độ lệch chuẩn bằng 1, loại bỏ sai lệch do đơn vị đo lường khác nhau giữa các kênh tương tác.

3.2. Thuật toán Phân cụm K-Means

Thuật toán K-Means được lựa chọn để thực hiện phân đoạn thí sinh nhờ tốc độ hội tụ nhanh và khả năng xử lý tập dữ liệu quy mô lớn hiệu quả. Mục tiêu của thuật toán là phân chia n quan sát vào k cụm sao cho tổng bình phương khoảng cách giữa các điểm dữ liệu đến tâm cụm (**centroid**) là nhỏ nhất.

Hàm mục tiêu (Tổng bình phương sai số trong cụm - **WCSS**) được định nghĩa như sau:

$$J = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2$$

Trong đó:

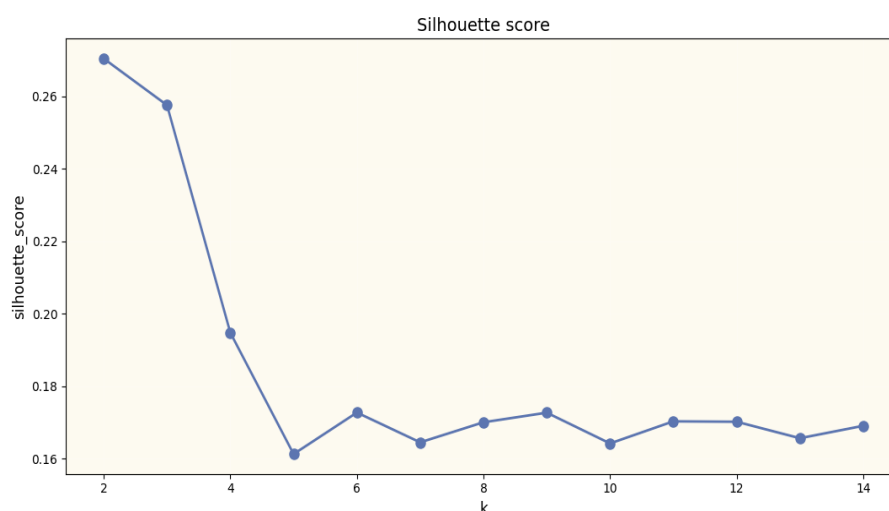
- k : Số lượng cụm cần phân chia.
- C_i : Tập hợp các điểm dữ liệu thuộc cụm i .
- μ_i : Tâm cụm (giá trị trung bình) của các điểm trong C_i .
- $\|x - \mu_i\|^2$: Khoảng cách Euclidean giữa điểm dữ liệu x và tâm cụm μ_i .

Quy trình thực hiện bao gồm 4 bước lặp:

- Khởi tạo: Chọn ngẫu nhiên k điểm làm tâm cụm ban đầu.
- Gán cụm: Gán mỗi thí sinh vào cụm có tâm gần nhất dựa trên khoảng cách Euclidean.
- Cập nhật: Tính toán lại tâm cụm μ_i bằng cách lấy trung bình cộng tọa độ của tất cả thí sinh trong cụm đó.

- Hội tụ: Lặp lại bước 2 và 3 cho đến khi vị trí các tâm cụm không còn thay đổi đáng kể.

Để xác định số lượng cụm tối ưu (k), nghiên cứu đã triển khai phương pháp hệ số *Silhouette* (Silhouette Coefficient).



Hình 1. Biểu đồ biến thiên hệ số *Silhouette*

Nguồn: Tác giả, 2026

Bảng 1. Kết quả đánh giá hệ số *Silhouette* theo số lượng cụm

Số cụm (k)	Silhouette Score
2	0,270
3	0,258
4	0,195
5	0,161
6	0,173
7	0,164
8	0,170
9	0,173
10	0,164
11	0,170
12	0,170
13	0,166
14	0,169

Nguồn: Tác giả, 2026

Để xác định số lượng cụm tối ưu, nghiên cứu sử dụng hệ số *Silhouette* trên dải giá trị từ ($k = 2$) đến ($k = 14$). Kết quả được trình bày trong Bảng 1 cho thấy giá trị *Silhouette* đạt cao nhất tại ($k = 2$) (0,270) và giảm dần khi số cụm tăng lên. Mặc dù ($k = 2$) và ($k = 3$) cho giá trị *Silhouette* cao hơn, cấu trúc cụm thu được chưa phản ánh đầy đủ sự đa dạng trong hành vi tiếp cận thông tin của thí sinh. Trong khi đó, tại ($k = 4$), hệ số *Silhouette* vẫn duy trì ở mức chấp nhận được (0,195) và các cụm hình thành có ý nghĩa quản trị rõ ràng, cho phép nhận diện bốn nhóm hành vi khác biệt gồm: (0) Tin

tưởng nội bộ và chủ động số; (1) Tiềm năng thụ động; (2) Tương tác đa kênh tích cực; và (3) Tập trung nội dung số. Do đó, nghiên cứu lựa chọn ($k = 4$) như một điểm cân bằng giữa tiêu chí thống kê và khả năng diễn giải thực tiễn trong bối cảnh quản trị tuyển sinh của Trường Đại học Thủ Dầu Một.

3.3. Phân tích thành phần chính (PCA)

Nhằm khắc phục hiện tượng đa chiều và loại bỏ tính đa cộng tuyến giữa các kênh thông tin, nghiên cứu áp dụng kỹ thuật phân tích thành phần chính (PCA). Phương pháp này giúp:

- Nén không gian dữ liệu gốc xuống còn 2 thành phần chính (PCA 1 và PCA 2) nhưng vẫn giải thích được hơn 42% tổng phương sai của dữ liệu.
- Trực quan hóa cấu trúc phân tán của thí sinh trên mặt phẳng hai chiều, làm cơ sở để nhận diện các hình thái tương tác đặc trưng.

Trong nghiên cứu này, PCA chỉ được sử dụng cho mục đích giảm chiều và trực quan hóa dữ liệu sau khi quá trình phân cụm K-Means hoàn tất. Thuật toán K-Means và mô hình Random Forest được huấn luyện trên tập thuộc tính đã chuẩn hóa nhưng chưa qua biến đổi PCA. Việc sử dụng PCA nhằm hỗ trợ quan sát trực quan mức độ phân tách giữa các cụm trên không gian hai chiều và không tham gia trực tiếp vào quá trình tạo nhãn cụm.

3.4. Kiểm chứng mô hình và đánh giá thuộc tính

- Để đánh giá mức độ tách biệt của các phân đoạn trong không gian đặc trưng, nghiên cứu sử dụng Random Forest Classifier nhằm tái dự đoán các nhãn cụm do K-Means tạo ra. Kết quả này được xem là một hình thức đánh giá nội bộ (internal validation), không phải là kiểm chứng độc lập trên dữ liệu bên ngoài. Kiểm chứng bằng thuật toán giám sát: Sử dụng mô hình Random Forest Classifier để dự đoán nhãn cụm đã phân loại. Các chỉ số Accuracy, Precision, Recall và F1-Score được tính toán để đo lường mức độ tách biệt của các cụm.

- Ma trận nhầm lẫn (Confusion Matrix): Được sử dụng để kiểm tra sự nhầm lẫn giữa các phân đoạn, đảm bảo mỗi cụm có một “biên giới hành vi” sắc nét.

- Độ quan trọng thuộc tính (Feature Importance): Sử dụng thuật toán Random Forest để xác định các biến số (features) đóng vai trò quyết định trong việc hình thành nên chân dung của từng phân đoạn, giúp giải thích bản chất tương tác của người học.

3.5. Công cụ thực hiện

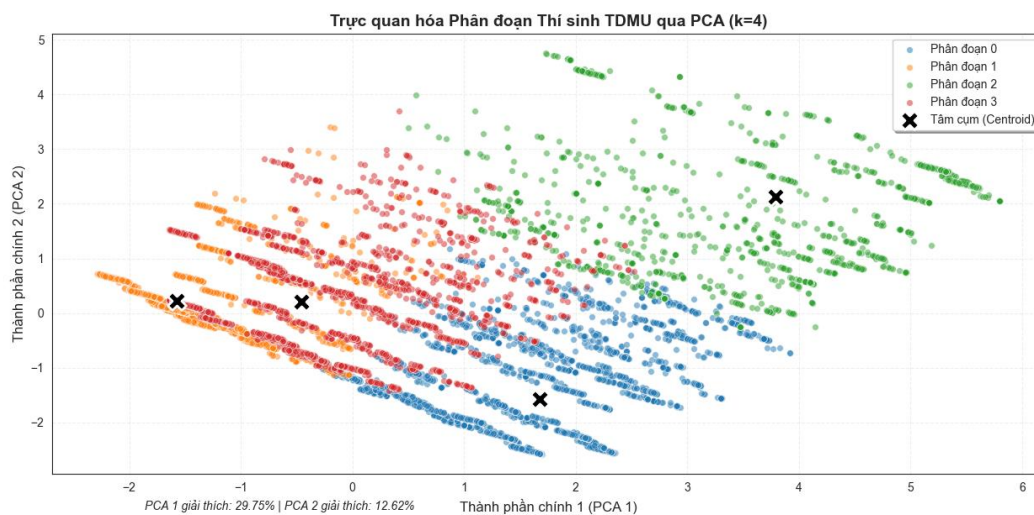
Toàn bộ quy trình phân tích và huấn luyện mô hình được thực hiện trên ngôn ngữ lập trình *Python 3.14*, sử dụng các thư viện chuyên dụng: *Scikit-learn* cho các thuật toán học máy, *Pandas* cho xử lý dữ liệu và *Seaborn/Matplotlib* cho trực quan hóa kết quả nghiên cứu.

4. Kết quả, thảo luận

4.1. Kết quả phân đoạn hành vi thí sinh

Thông qua việc triển khai thuật toán K-Means kết hợp với kỹ thuật giảm chiều dữ liệu PCA, nghiên cứu đã xác định được cấu trúc phân đoạn gồm 4 nhóm thí sinh riêng biệt. Kết quả trực quan hóa trên không gian 2D (Hình 2) cho thấy các cụm dữ liệu

có xu hướng hình thành các dải hành vi song song, minh chứng cho việc thí sinh không tương tác ngẫu nhiên mà tuân theo các kịch bản tiếp cận thông tin có quy luật. Thành phần chính PCA 1 giải thích được 29,75% biến thiên, đóng vai trò là trục phân loại chủ đạo giữa nhóm tương tác đa kênh và nhóm tương tác đơn lẻ.



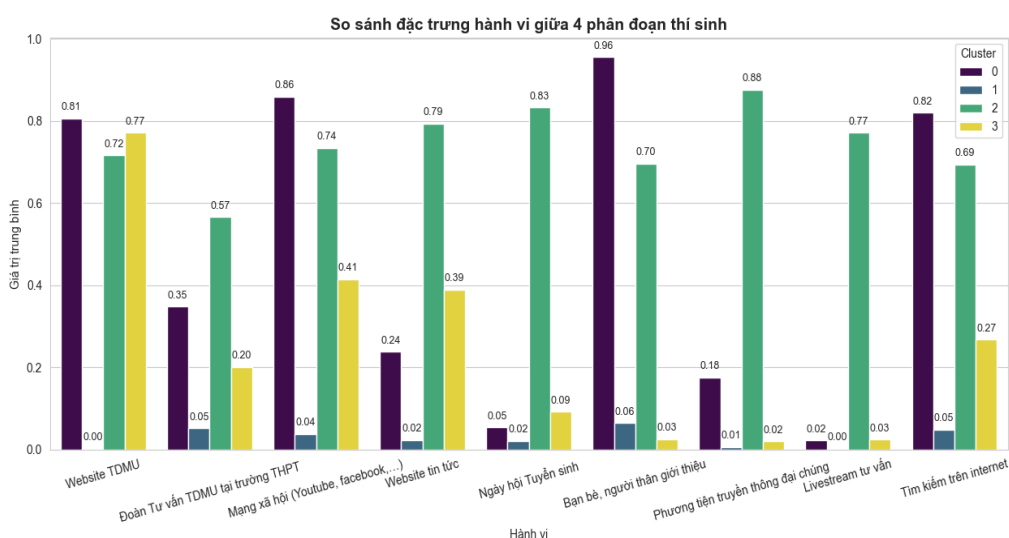
Hình 2. Biểu đồ trực quan hóa phân đoạn thí sinh TDMU

Nguồn: Tác giả, 2026

Sự phân bố các điểm dữ liệu theo các dải song song trong biểu đồ PCA là minh chứng cho việc các thuộc tính đầu vào (như Phương thức tuyển sinh, Kênh tư vấn...) không tồn tại độc lập mà có sự liên kết chặt chẽ theo từng nhóm đối tượng. Việc các tâm cụm (dấu X) nằm ở vị trí trung tâm của các mật độ điểm tương ứng khẳng định thuật toán K-Means đã hội tụ tốt và xác định đúng các “điểm neo” đại diện cho chân dung thí sinh TDMU.

4.2. Định danh và đặc điểm chân dung các phân đoạn

Dựa trên chỉ số giá trị trung bình của các thuộc tính hành vi quan trọng (Hình 3), nghiên cứu phân tích chân dung 4 phân đoạn thí sinh như sau:



Hình 3. Biểu đồ cột so sánh đặc trưng hành vi giữa 4 phân đoạn thí sinh

Nguồn: Tác giả, 2026

• Phân đoạn 0 - Nhóm “Tin tưởng nội bộ & Chủ động số”: Đây là nhóm thí sinh có số lượng điểm tương tác rất cao, đặc trưng bởi sự kết hợp giữa niềm tin truyền thống và hành vi số hóa. Nhóm này đạt điểm tuyệt đối trong việc chịu ảnh hưởng từ “Bạn bè, người thân giới thiệu” (0,96), đồng thời sở hữu mức độ tương tác vượt trội trên “Mạng xã hội” (0,86), “Tìm kiếm trên Internet” (0,82) và “Website TDMU” (0,81). Kết quả này phản ánh xu hướng tâm lý học hành vi: thí sinh đưa ra quyết định dựa trên sự bảo chứng từ người thân, sau đó chủ động sử dụng toàn bộ hệ sinh thái số của nhà trường để thẩm định và củng cố niềm tin đó.

• Phân đoạn 1 - Nhóm “Tiềm năng thụ động”: Phân đoạn này có đồ thị hành vi gần như đi ngang và tiệm cận mức 0,00 ở tất cả các thuộc tính (chỉ số cao nhất chỉ là 0,06 ở mục người thân giới thiệu). Đây là tệp dữ liệu đại diện cho nhóm thí sinh chưa phát sinh bất kỳ “dấu vết số” nào trên hệ thống truyền thông của nhà trường. Họ là nhóm thí sinh mới chỉ dừng ở mức độ nhận biết thương hiệu, chưa phát sinh hành vi tìm hiểu sâu hoặc tương tác trực tiếp với hệ thống tư vấn.

• Phân đoạn 2 - Nhóm “Tương tác đa kênh tích cực”: Đây là nhóm thí sinh có mức tương tác đa kênh cao nhất, khi ghi nhận chỉ số tương tác cao đồng đều ở cả hai mặt trận trực tuyến lẫn trực tiếp. Nhóm này dẫn đầu tại các điểm chạm mang tính trải nghiệm thực tế như “Ngày hội Tuyển sinh” (0,88) và “Livestream tư vấn” (0,77). Song song đó, họ thu thập kiến thức kiểm chứng qua “Website tin tức” (0,79), “Đoàn tư vấn TDMU tại trường THPT” (0,57) và duy trì mức theo dõi cao tại “Website TDMU” (0,72). Nhóm này không phụ thuộc vào ý kiến người thân (chỉ 0,70, thấp hơn cụm 0) mà tự mình trải nghiệm đa kênh để đưa ra lựa chọn.

• Phân đoạn 3 - Nhóm “Tập trung nội dung số”: Nhóm này phản ánh chân dung của những công dân số thế hệ mới với hành vi mang tính độc lập rất cao. Thí sinh cụm 3 khai thác thông tin cực kỳ mạnh mẽ trên “Website TDMU” (0,77) và “Mạng xã hội” (0,41), nhưng lại hầu như không chịu ảnh hưởng với các tác động ngoại cảnh như “Bạn bè, người thân giới thiệu” (chỉ 0,03), “Ngày hội Tuyển sinh” (0,09) hay “Livestream tư vấn” (0,03). Nhóm này có xu hướng tự nghiên cứu nội dung số một cách âm thầm và độc lập trước khi thực hiện các bước xét tuyển.

4.3. Kiểm chứng độ tin cậy và ý nghĩa các thuộc tính

Bảng 2. Kết quả kiểm chứng mô hình

Cụm	Precision	Recall	F1-score	Support
0	0,99	1,00	0,99	1224
1	1,00	1,00	1,00	2326
2	1,00	0,97	0,99	584
3	0,99	0,99	0,99	1165
Accuracy			0,99	5299
Macro avg	0,99	0,99	0,99	5299
Weighted avg	0,99	0,99	0,99	5299

Nguồn: Tác giả, 2026

Để đánh giá tính ổn định và sự tách biệt của 4 phân đoạn thí sinh đã xác lập, nghiên cứu thực hiện kiểm chứng chéo thông qua thuật toán học máy có giám sát

Random Forest. Bằng cách sử dụng các đặc trưng hành vi ban đầu để dự đoán nhãn cụm từ K-Means, kết quả thực nghiệm như sau:

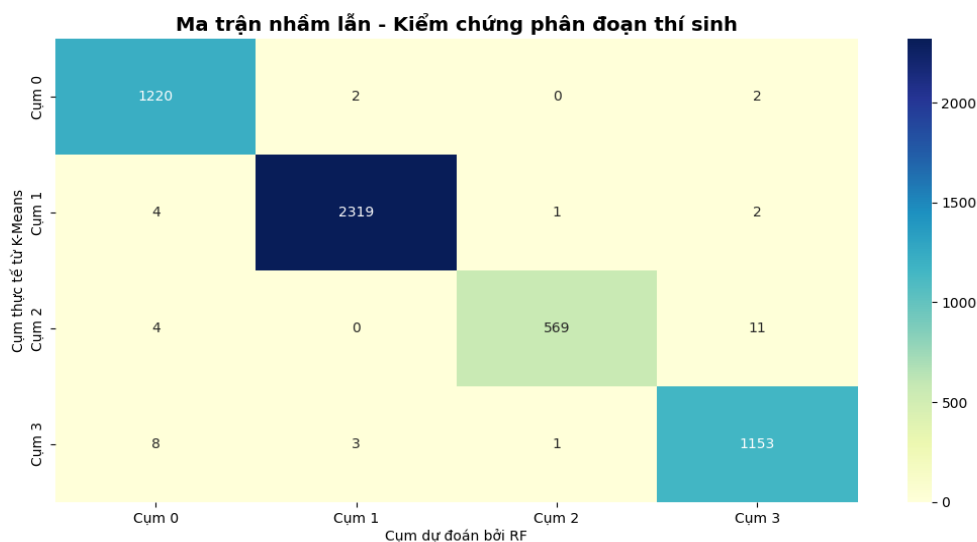
- Mô hình đạt độ chính xác tổng thể (Accuracy) lên đến 99,28%. Các chỉ số Precision, Recall và F1-score cho cả 4 cụm đều dao động ở mức từ 0,97 đến 1,00. Kết quả cho thấy các thuộc tính hành vi được sử dụng có khả năng phân biệt tốt giữa các cụm đã hình thành “dấu vết số” (Website, Mạng xã hội, Kênh tư vấn...) mang tính đặc trưng rất cao, cho phép phân loại thí sinh vào các nhóm hành vi một cách gần như tuyệt đối.

- Hiệu suất phân loại đồng đều: Cụm 1 đạt các chỉ số Precision và Recall cao nhất, cho thấy nhóm này có đặc trưng hành vi tương đối rõ ràng trong phạm vi dữ liệu nghiên cứu. Các cụm còn lại (0, 2, 3) chỉ có sự sai lệch rất nhỏ (dưới 3%), gợi ý rằng cấu trúc 4 phân đoạn có mức độ ổn định tương đối cao trên tập dữ liệu khảo sát.

Ma trận nhầm lẫn (Hình 4) cung cấp cái nhìn chi tiết về sự tách biệt giữa các phân đoạn:

- Đường chéo chính chiếm ưu thế: Hầu hết các mẫu dữ liệu đều nằm trên đường chéo chính (ví dụ: 2319 mẫu cụm 1 được dự đoán đúng hoàn toàn). Sự tập trung này minh chứng cho việc thuật toán K-Means đã tìm ra được các ranh giới hành vi rất sắc nét giữa các nhóm thí sinh tại TDMU.

- Sai số không đáng kể: Chỉ có một số lượng rất ít các trường hợp bị nhầm lẫn, chẳng hạn như 11 thí sinh từ Cụm 2 bị hệ thống dự đoán nhầm sang Cụm 3. Sự nhầm lẫn nhỏ này thường xuất phát từ những thí sinh có hành vi lai (hybrid), sử dụng đồng thời nhiều kênh tương tác tương đồng nhau.



Hình 4. Biểu đồ ma trận nhầm lẫn – kiểm chứng phân đoạn thí sinh

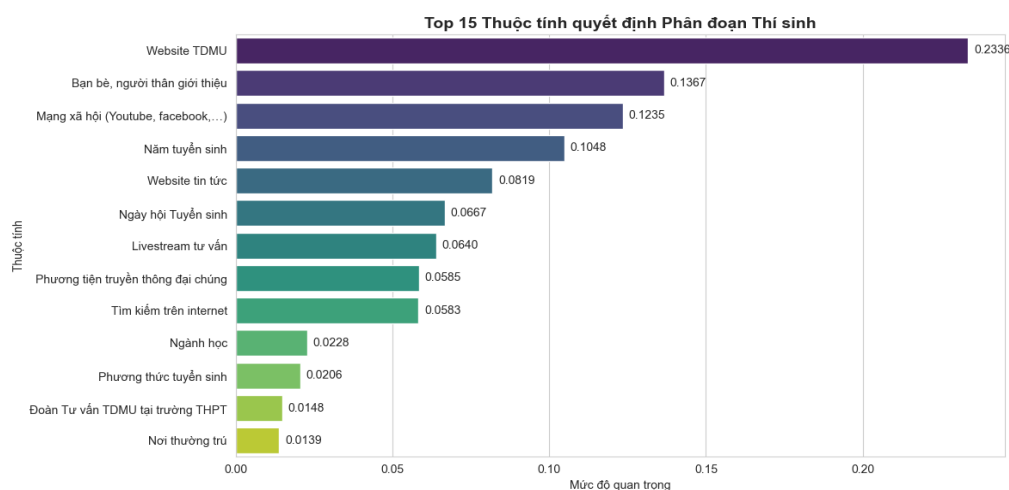
Nguồn: Tác giả, 2026

Kết quả kiểm chứng đạt 99,28% là một kết quả đáng chú ý trong bối cảnh nghiên cứu hành vi người học, nó cho thấy:

- Quy trình tiền xử lý và kỹ thuật giảm chiều PCA đã loại bỏ hiệu quả các nhiễu thông tin.

- Các kênh thông tin truyền thông của TDMU gợi ý sự hình thành những nhóm đối tượng thí sinh có đặc thù riêng biệt.

- Với độ chính xác này, nhà trường có thể cân nhắc sử dụng mô hình Random Forest để dự đoán phân đoạn cho các hồ sơ thí sinh mới ngay khi họ phát sinh tương tác đầu tiên, từ đó thực hiện chiến dịch cá nhân hóa tư vấn kịp thời với khả năng nhận diện tương đối chính xác các nhóm hành vi trên các dữ liệu có đặc điểm tương đồng.



Hình 5. Biểu đồ so sánh thuộc tính quyết định phân đoạn thí sinh

Nguồn: Tác giả, 2026

Kết quả từ mô hình Random Forest (Hình 5) đã làm sáng tỏ các biến số đóng vai trò then chốt trong việc định hình chân dung các nhóm thí sinh tại Đại học Thủ Dầu Một. Các phát hiện quan trọng bao gồm:

- Thuộc tính “Website TDMU” đứng đầu với mức độ quan trọng vượt trội (0,2336). Điều này khẳng định trang web chính thức của nhà trường là điểm chạm (touch-point) cốt lõi nhất, nơi thí sinh ở tất cả các cụm tập trung khai phá thông tin chính thống. Theo sau đó là các kênh “Mạng xã hội” (0,1235) và “Website tin tức” (0,0819), cho thấy hành vi của thí sinh hiện nay mang đậm đặc tính của kỷ nguyên số.

- Truyền thông truyền thống và sự tin tưởng: Yếu tố “Bạn bè, người thân giới thiệu” xếp thứ hai với trọng số 0,1367. Kết quả này chỉ ra rằng, bên cạnh các kênh kỹ thuật số, niềm tin dựa trên sự truyền miệng vẫn là nhân tố quyết định để phân biệt các nhóm thí sinh có xu hướng cộng đồng cao.

- Thời gian và phương thức tư vấn: “Năm tuyển sinh” (0,1048) cho thấy có sự thay đổi rõ rệt trong hành vi thí sinh qua từng năm. Đồng thời, sự xuất hiện của “Ngày hội Tuyển sinh” (0,0667) và “Livestream tư vấn” (0,0640) với trọng số tương đương nhau minh chứng cho sự tồn tại song song giữa hai phương thức tương tác: trực tiếp và trực tuyến.

- Các yếu tố nhân khẩu học và định hướng ngành nghề: Đáng chú ý, các yếu tố như “Ngành học” (0,0228), “Phương thức tuyển sinh” (0,0206) và “Nơi thường trú” (0,0139) có trọng số khá thấp. Điều này gợi ý rằng hành vi tiếp cận thông tin quan trọng hơn là đặc điểm lý lịch cá nhân trong việc phân đoạn thí sinh tại TDMU.

Từ kết quả phân tích độ quan trọng của thuộc tính, nghiên cứu đề xuất các định hướng chiến lược và kịch bản tư vấn theo phân đoạn:

- Với tệp thí sinh tập trung nội dung số (Cụm 3), hệ thống Website TDMU cần tích hợp, nâng cấp các công cụ tự động hóa như chatbot AI, công tra cứu điểm chuẩn, tài liệu học bổng số trực quan để họ tự khai phá mà không cần sự can thiệp của tư vấn viên trực tiếp.

- Kích hoạt mạng lưới đại sứ (Cụm 0): Tận dụng yếu tố “Bạn bè, người thân giới thiệu” bằng cách xây dựng các chương trình đại sứ sinh viên hoặc mạng lưới cựu sinh viên để tăng cường uy tín truyền miệng.

- Chiến dịch đa kênh tích hợp (Cụm 2): Kết hợp linh hoạt giữa Livestream và Ngày hội trực tiếp, vì cả hai đều là những biến số có giá trị phân loại thí sinh hiệu quả tương đương nhau. Nhà trường chỉ nên điều hướng nguồn lực, gửi thư mời đích danh và tổ chức các minigame, talkshow chuyên sâu hướng thẳng vào nhóm đối tượng này để tránh lãng phí ngân sách tổ chức đại trà cho các nhóm còn lại.

- Đối với tệp dữ liệu tĩnh ở Cụm 1, cần sử dụng các chiến dịch tin nhắn SMS hoặc email marketing với nội dung khơi gợi, khảo sát ngắn nhằm định hướng họ thực hiện “điểm chạm đầu tiên” trên website hoặc mạng xã hội của trường.

4.4. Hạn chế nghiên cứu

Mặc dù nghiên cứu sử dụng tập dữ liệu lớn và đa dạng, vẫn tồn tại một số hạn chế. Thứ nhất, dữ liệu được thu thập từ một cơ sở giáo dục duy nhất nên khả năng khái quát hóa cho các trường đại học khác còn hạn chế. Thứ hai, việc đánh giá mô hình chủ yếu dựa trên dữ liệu nội bộ và chưa thực hiện kiểm chứng ngoài mẫu theo từng năm tuyển sinh hoặc trên các nhóm thí sinh mới. Thứ ba, nghiên cứu sử dụng phương pháp Label Encoding để mã hóa các biến định tính, trong khi chưa tiến hành so sánh với các phương pháp mã hóa khác như One-Hot Encoding hoặc các thuật toán xử lý dữ liệu hỗn hợp như K-Prototypes. Việc đánh giá ảnh hưởng của các kỹ thuật mã hóa khác nhau đến chất lượng phân cụm có thể giúp phát hiện những cấu trúc phân đoạn khác biệt và nâng cao độ tin cậy của kết quả. Đây là những hướng nghiên cứu cần được tiếp tục xem xét trong tương lai.

5. Kết luận

Nghiên cứu đã xây dựng một quy trình phân tích dữ liệu tuyển sinh dựa trên sự kết hợp giữa K-Means, PCA và Random Forest nhằm nhận diện các nhóm hành vi thí sinh trong môi trường giáo dục số. Kết quả cho thấy dữ liệu tương tác đa kênh có khả năng hình thành các phân đoạn hành vi tương đối rõ ràng, đồng thời Random Forest đạt độ chính xác 99,28% trong việc tái nhận diện các nhãn cụm trên cùng tập dữ liệu nghiên cứu.

Kết quả phân tích xác định bốn nhóm thí sinh có đặc trưng tương tác khác biệt, trong đó website nội bộ và mạng xã hội giữ vai trò nổi bật, song niềm tin truyền miệng vẫn là yếu tố phân biệt quan trọng. Phát hiện này mở ra khả năng chuyển đổi mô hình quản trị tuyển sinh từ đại trà sang cá nhân hóa, giúp tối ưu hóa nguồn lực truyền thông.

Tuy nhiên, kết quả cần được xem xét trong bối cảnh dữ liệu của một cơ sở giáo dục cụ thể. Các nghiên cứu tiếp theo có thể mở rộng phạm vi khảo sát, bổ sung dữ liệu theo chuỗi thời gian hoặc thực hiện đánh giá ngoài mẫu để kiểm tra tính ổn định của các phân đoạn trong những bối cảnh khác nhau.

TÀI LIỆU THAM KHẢO

- [1] Alvarez-Garcia, M., Arenas-Parra, M., & Ibar-Alonso, R. (2024). Uncovering student profiles: An explainable cluster analysis approach to PISA 2022. *Computers & Education*, 223, 105166. <https://doi.org/10.1016/j.compedu.2024.105166>
- [2] Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3–17.
- [3] Baker, R. S., & Inventado, P. S. (2014). Educational Data Mining and Learning Analytics. In J. A. Larusson, & B. White, *Learning Analytics: From Research to Practice* (pp. 61-75). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-3305-7_4
- [4] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- [5] Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques (3rd ed.)*. Waltham, MA 02451, USA: Morgan Kaufmann.
- [6] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065). <https://doi.org/10.1098/rsta.2015.0202>
- [7] Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., Dinu, A., & Stachl, C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports*, 13(1). <https://doi.org/10.1038/s41598-023-32484-w>
- [8] Pedregosa, Fabian et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- [9] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
- [10] Susnjak, T. (2024). Beyond predictive learning analytics modelling and onto explainable artificial intelligence with prescriptive analytics and ChatGPT. *International Journal of Artificial Intelligence in Education*, 34(2), 452–482. <https://doi.org/10.1007/s40593-023-00336-3>
- [11] Witten, Ian H., Frank, Eibe, Hall, Mark A., & Pal, Christopher J. (2016). *Data mining: Practical machine learning tools and techniques (4th ed.)*. Waltham, MA 02451, USA: Morgan Kaufmann.