

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỌC & KỸ THUẬT THÔNG TIN**

****

**MẠNG XÃ HỘI – IS535.P11.CNCL**

***Báo cáo đồ án***

*Đánh giá chất lượng bộ dữ liệu phục vụ*

*quản lý chất lượng đầu vào khóa học*

Nhóm 3:

| Hoàng Ngọc Quý (NT) | 21522529 |
| --- | --- |
| Nguyễn Đức Thịnh | 21522636 |
| Nguyễn Trường Huy | 19521625 |
| Trần Khiết Tường | 21521657 |
| Nguyễn Tuấn Khang | 21522196 |

**TP. Hồ Chí Minh, tháng 6 năm 2024**

## LỜI CẢM ƠN

Đồ án “Đánh giá chất lượng bộ dữ liệu phục vụ quản lý chất lượng đầu vào khóa học” là kết quả của một quá trình dài học tập, nghiên cứu, và nỗ lực không ngừng nghỉ của nhóm chúng tôi. Thành công này không chỉ nhờ vào sự cố gắng của nhóm mà còn là sự hỗ trợ, hướng dẫn và động viên quý báu từ cô và những người bạn đã đồng hành cùng nhóm trong suốt thời gian qua. Nhân đây, nhóm xin bày tỏ lòng biết ơn sâu sắc đến tất cả những ai đã giúp đỡ nhóm hoàn thành đồ án này.

Trước tiên, nhóm xin gửi lời cảm ơn chân thành và sâu sắc nhất đến giáo viên hướng dẫn ThS. Nguyễn Thị Anh Thư. Cô không chỉ tận tình hướng dẫn, chỉ bảo mà còn hỗ trợ nhóm trong việc tìm kiếm tài liệu tham khảo, cung cấp các thông tin khoa học cần thiết, và đưa ra những góp ý quý giá để nhóm có thể hoàn thiện nghiên cứu của mình. Sự chỉ dẫn tận tâm và kiên nhẫn của cô chính là nguồn động lực lớn để nhóm vượt qua những khó khăn và hoàn thành tốt đề tài này.

Nhóm cũng xin chân thành cảm ơn đến quý thầy cô trong khoa đã tận tình giảng dạy, truyền đạt những kiến thức nền tảng vững chắc và kỹ năng nghiên cứu quý báu trong suốt thời gian học tập tại trường. Những bài học và kinh nghiệm từ thầy cô không chỉ hỗ trợ nhóm trong đồ án này mà còn là hành trang quý giá cho con đường học tập và làm việc sau này.

Cuối cùng, nhóm xin cảm ơn tất cả các cá nhân, tổ chức đã hỗ trợ nhóm về mặt dữ liệu, thông tin, và ý tưởng để nhóm có thể hoàn thiện nghiên cứu một cách tốt nhất.

Nhóm xin kính chúc cô dồi dào sức khỏe, hạnh phúc và thành công trong công việc cũng như cuộc sống.

## MỤC LỤC

## 

[LỜI CẢM ƠN 2](#_5dd09l3g7zg0)

[MỤC LỤC 3](#_iqj2jdpl6dfw)

[**CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN 5**](#_vvbv64823q09)

[1. Tổng quan 5](#_vlzjp0g0t10a)

[1.1. Định nghĩa bài toán 5](#_9koy13zhg98p)

[1.1.1. Ngữ cảnh bài toán 5](#_yqfnmir44242)

[1.1.2. Tính cấp thiết 6](#_bj1yq8lq661d)

[1.1.3. Tính mới 6](#_ytjq7fern9r9)

[1.2. Ứng dụng 7](#_xdf5ve1rs2kx)

[1.3. Khó khăn và thử thách 8](#_tknw9ll7n9on)

[**CHƯƠNG 2: CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN 10**](#_pxr4bcd6n7m0)

[2.1. Tình hình nghiên cứu thế giới 10](#_m23ivaeen8mt)

[2.2. Tình hình nghiên cứu tại Việt Nam 11](#_mfx9nxalm504)

[**CHƯƠNG 3: PHÂN TÍCH TẬP DỮ LIỆU 12**](#_9t07markl6kh)

[3.1 Bộ dữ liệu MOOCCubeX 12](#_542xg7ovx2oe)

[3.2. Mô tả dữ liệu và thăm dò dữ liệu 14](#_fxlln8vrhkh9)

[**CHƯƠNG 4: PHƯƠNG PHÁP ĐỀ XUẤT 22**](#_bkc9t3euncpc)

[4.1. Hướng triển khai đề tài 22](#_7jvasgkpdyvu)

[4.2. Framework thực hiện 22](#_bork77mzr4me)

[4.3. Xây dựng mạng 23](#_67r2yf3ktvqq)

[**CHƯƠNG 5: KẾT QUẢ THỰC NGHIỆM 25**](#_6v70tsaiwwjh)

[5.1. Môi trường thực nghiệm 25](#_jvp9g5ewnefk)

[5.2. Giới hạn dữ liệu đầu vào 25](#_j1dqtdrtr7ep)

[5.3. Các tiêu chí đánh giá chất lượng khóa học 26](#_jp9wnjrj9eq8)

[5.4. Đánh giá mô hình bằng SVM 26](#_q5lci7tqb9ds)

[5.5. Kết quả thực nghiệm 27](#_x1q2zziz4qim)

[5.6. Phân tích thêm 28](#_r11p085odjq6)

[5.7. Đánh giá chung 28](#_a4ilcr6ea4cf)

[**CHƯƠNG 6: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 30**](#_bpo5uxb9o1ie)

[6.1. Kết luận 30](#_j4wtrnm2dsww)

[6.2. Hướng phát triển 31](#_hrkgwmnpwzh0)

[**THAM KHẢO 33**](#_2z0ij9ela0h3)

## 

# CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN

## 

## Tổng quan

### Định nghĩa bài toán

#### Ngữ cảnh bài toán

* Trong những năm gần đây, giáo dục đã chứng kiến một sự thay đổi mạnh mẽ nhờ vào sự xuất hiện và phát triển mạnh mẽ của các khóa học trực tuyến mở rộng (MOOCs). Những khóa học này đã tạo ra cơ hội học tập vô cùng linh hoạt, cho phép học viên trên toàn cầu tiếp cận một lượng lớn tài nguyên học tập ngay tại nhà mà không cần phải đến trường. Tuy nhiên, sự gia tăng nhanh chóng của các khóa học trực tuyến cũng đặt ra thách thức về chất lượng và tính ổn định của nội dung các khóa học, cũng như cách thức phân phối chúng trên các nền tảng trực tuyến.
* Khi MOOCs ngày càng phổ biến, vấn đề đảm bảo chất lượng khóa học trở thành một yếu tố then chốt. Hệ thống quản lý chất lượng không chỉ giúp duy trì các tiêu chuẩn giáo dục, mà còn đảm bảo sự tham gia của người học và nâng cao hiệu quả học tập trong môi trường số. Trong bối cảnh giáo dục trực tuyến phát triển mạnh mẽ, việc cải thiện và tối ưu hóa các chiến lược quản lý chất lượng khóa học là cực kỳ quan trọng để mang lại trải nghiệm học tập tốt nhất cho học viên, đáp ứng nhu cầu đa dạng từ nhiều đối tượng học viên khác nhau trên toàn cầu.
* Đề tài này nhắm đến việc xây dựng hệ thống quản lý chất lượng dữ liệu đầu vào cho các khóa học MOOCs. Qua đó, chúng tôi mong muốn giải quyết những thách thức liên quan đến việc duy trì chất lượng khóa học trong môi trường học trực tuyến. Mục tiêu chính là phát triển các chiến lược giúp đánh giá chất lượng khóa học hiệu quả, đồng thời cảnh báo và cải thiện những vấn đề liên quan đến chất lượng, từ đó nâng cao chất lượng học tập và trải nghiệm học viên.

#### Tính cấp thiết

* Chất lượng dữ liệu đã là chủ đề nóng trong nhiều năm và tiếp tục nhận được nhiều sự quan tâm vì giá trị của dữ liệu phụ thuộc nhiều vào chất lượng và tính dễ sử dụng của nó. Dữ liệu thực sự quan trọng nhưng tối quan trọng hơn nữa là nguồn dữ liệu đó phải có chất lượng cao, phục vụ được đầy đủ và chính xác những yêu cầu của người dùng. Điều này có nghĩa là việc đánh giá, quản lý chất lượng đầu vào và có phương hướng nâng cao chất lượng của dữ liệu là một công việc rất cần thiết, không thể thiếu. Việc sử dụng đúng dữ liệu chất lượng cao có thể giúp việc lập kế hoạch, phân tích và quyết định tốt hơn. Hơn nữa, việc cần có dữ liệu giáo dục chất lượng cao cũng nhằm để xác định định hướng của các chính sách giáo dục và so sánh vị thế giáo dục của quốc gia với các quốc gia khác.
* Mặt khác, dữ liệu chất lượng kém có thể không phù hợp với mục đích đã định và có thể dẫn đến những hậu quả nghiêm trọng. Đặc biệt là đối với các nghiên cứu về chất lượng dữ liệu đã được thực hiện trong nhiều lĩnh vực khác nhau từ các lĩnh vực dựa trên dữ liệu như dữ liệu lớn và thống kê đến các lĩnh vực được quan tâm như chăm sóc sức khỏe, tài chính và hệ thống thông tin.
* Như vậy, có thể thấy rằng hệ thống quản lý chất lượng dữ liệu đầu vào của các khóa học được trình bày là một chủ đề cấp thiết và phải kịp thời.

#### Tính mới

* Chất lượng dữ liệu là một vấn đề cực kỳ quan trọng trong mọi lĩnh vực, và đặc biệt trong giáo dục, nó lại càng có ý nghĩa sâu sắc. Chất lượng dữ liệu sẽ quyết định đến độ chính xác và khả năng sử dụng của dữ liệu đó trong quá trình ra quyết định. Đặc biệt đối với các nền tảng học trực tuyến như MOOCs, việc quản lý và đánh giá chất lượng dữ liệu đầu vào là vô cùng quan trọng để đảm bảo các khóa học đạt được chất lượng tối ưu. Những dữ liệu chất lượng cao sẽ giúp cải thiện quy trình giảng dạy, nâng cao kết quả học tập, và giúp các nhà quản lý giáo dục đưa ra các chiến lược phù hợp.
* Dữ liệu chất lượng kém, ngược lại, có thể dẫn đến những quyết định sai lầm, gây ảnh hưởng tiêu cực đến quá trình học tập. Điều này đặc biệt quan trọng khi các dữ liệu này được sử dụng để đánh giá kết quả học tập hoặc để cải thiện chất lượng khóa học. Vì thế, việc triển khai một hệ thống quản lý chất lượng dữ liệu khóa học là một nhu cầu bức thiết và không thể thiếu trong môi trường giáo dục trực tuyến ngày nay.

## Ứng dụng

* Hệ thống quản lý chất lượng dữ liệu đầu vào cho các khóa học trong bộ dữ liệu MOOCCubeX sẽ được xây dựng với các phương pháp đánh giá tự động và áp dụng học máy (machine learning) nhằm cải thiện và bảo vệ chất lượng của dữ liệu đầu vào. Hệ thống này dựa trên các tiêu chí quan trọng như tính đầy đủ, tính nhất quán, độ chính xác và tính hợp lệ của dữ liệu. Các ứng dụng của hệ thống này có thể được mở rộng trong nhiều môi trường giáo dục và nền tảng học tập trực tuyến để đảm bảo hiệu quả, tính nhất quán và chất lượng tổng thể của các khóa học. Dưới đây là một số ứng dụng có thể áp dụng:
* **Thiết kế và phát triển khóa học:** Hệ thống có thể hỗ trợ các nhà giáo dục trong việc thiết kế và phát triển các khóa học bằng cách cung cấp các tiêu chuẩn chất lượng đầu vào, hướng dẫn và mẫu để thiết kế chương trình giảng dạy, tạo nội dung học tập và điều chỉnh cấu trúc khóa học. Điều này giúp tạo ra các trải nghiệm học tập hấp dẫn, có tính cấu trúc cao và hiệu quả hơn.
* **Đánh giá và đảm bảo chất lượng đầu vào:** Hệ thống sẽ hỗ trợ thực hiện các quy trình đảm bảo chất lượng khóa học thông qua các phương pháp đánh giá như đánh giá ngang hàng, đánh giá chuyên gia và kiểm tra tự động. Việc này đảm bảo rằng các khóa học và tài liệu giảng dạy đáp ứng các tiêu chuẩn chất lượng đầu vào đã được xác định trước, từ đó duy trì chất lượng giảng dạy.
* **Cải tiến liên tục:** Một ứng dụng quan trọng của hệ thống là hỗ trợ việc đánh giá và cải tiến liên tục các khóa học. Bằng cách thu thập phản hồi từ người học, người hướng dẫn và các bên liên quan, hệ thống có thể cung cấp các thông tin hữu ích để nâng cao chất lượng khóa học, dựa trên dữ liệu đầu vào. Các cập nhật và cải tiến khóa học có thể được thực hiện lặp đi lặp lại để nâng cao kết quả học tập và sự hài lòng của người học.
* **Giám sát và báo cáo:** Hệ thống cho phép giám sát hiệu suất khóa học theo thời gian thực, bao gồm các chỉ số như tỷ lệ hoàn thành khóa học, mức độ hài lòng của học viên và kết quả đánh giá. Thông qua việc tạo ra các báo cáo chi tiết và phân tích, hệ thống giúp xác định các nội dung cần cải tiến và đưa ra các biện pháp giải quyết vấn đề một cách kịp thời.

## **Khó khăn và thử thách**

Tuy nhiên, việc xây dựng và triển khai hệ thống quản lý chất lượng dữ liệu này cũng đối mặt với một số khó khăn và thách thức:

* **Tính đầy đủ của dữ liệu**Một vấn đề lớn là tính đầy đủ của dữ liệu đầu vào, với rất nhiều giá trị không xác định hoặc thiếu, khiến cho việc thống kê và phân tích dữ liệu gặp khó khăn. Điều này làm cho việc xác định đặc trưng và đặc điểm dữ liệu trở nên phức tạp, ảnh hưởng đến quá trình kiểm tra và đánh giá chất lượng dữ liệu.
* **Độ chính xác dữ liệu thấp**Độ chính xác của các giá trị dữ liệu không cao là một yếu tố gây trở ngại trong quá trình phân tích và khám phá dữ liệu. Các lỗi trong quá trình nhập liệu hoặc sai sót trong việc xử lý dữ liệu có thể dẫn đến những kết quả không chính xác và làm giảm hiệu quả của các phương pháp machine learning áp dụng trong hệ thống.
* **Khó khăn về ngôn ngữ**Một thách thức khác là sự đa dạng trong ngôn ngữ và cách thức diễn đạt thông tin trong các tài liệu khóa học. Các vấn đề liên quan đến sự không đồng nhất trong ngôn ngữ giảng dạy, thuật ngữ chuyên ngành và biểu đạt của các giảng viên có thể gây khó khăn trong việc nắms bắt ý nghĩa chính xác của dữ liệu, làm giảm hiệu quả trong việc đánh giá và cải tiến chất lượng khóa học.

Những thách thức này yêu cầu hệ thống phải có các phương pháp tiên tiến để xử lý dữ liệu thiếu, cải thiện độ chính xác và giải quyết các vấn đề liên quan đến ngôn ngữ, đồng thời đảm bảo tính hiệu quả và độ tin cậy của quá trình quản lý chất lượng dữ liệu.

# CHƯƠNG 2: CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

## 2.1. Tình hình nghiên cứu thế giới

Hệ thống đánh giá chất lượng khóa học đã trở thành một lĩnh vực nghiên cứu phát triển mạnh mẽ, đặc biệt là trong bối cảnh các nền tảng học trực tuyến và trường học (MOOCs) như Coursera, edX và Udemy ngày càng trở nên phổ biến.

Song song đó có nhiều nghiên cứu liên quan về MOOCs như:

* Quality Assessment of MOOCs Based on Learning Behavior and Engagement Data: Sử dụng dữ liệu hành vi học tập và mức độ tham gia của người học để đánh giá chất lượng các khóa học trực tuyến mở lớn (MOOCs). Mô hình đánh giá kết hợp dữ liệu từ các chỉ số như tần suất tham gia, lượt xem video, và thảo luận diễn đàn để đưa ra các đánh giá chất lượng chi tiết cho từng khóa học.
* Data-Driven Quality Assessment for Online Learning: A Case Study of Coursera MOOCs: Sử dụng các phương pháp phân tích dữ liệu để đánh giá chất lượng các khóa học trên nền tảng Coursera. Các chỉ số như điểm số của học viên, sự tham gia của họ trong các hoạt động thảo luận và bài tập thực hành được sử dụng để xác định mức độ hiệu quả của khóa học.
* MOOC Course Quality Evaluation Based on Multidimensional Data Mining: Đánh giá chất lượng của các khóa học MOOCs bằng cách sử dụng khai thác dữ liệu đa chiều. Các dữ liệu đầu vào bao gồm mức độ tham gia, phản hồi của học viên, và tỷ lệ hoàn thành khóa học. Phương pháp khai thác dữ liệu được sử dụng để phân tích các yếu tố ảnh hưởng đến chất lượng khóa học.
* Tổng quan tài liệu có hệ thống về đánh giá chất lượng dữ liệu (A Systematic Literature Review on Data Quality Assessment): Đây là một tổng quan hệ thống về các phương pháp đánh giá chất lượng dữ liệu, phân tích các công cụ, tiêu chí và mô hình hiện tại trong việc xác định và cải thiện chất lượng dữ liệu.
* Quản lý chất lượng dữ liệu đối với dữ liệu giáo dục: Một trường hợp nghiên cứu về bách khoa thống kê (Data Quality Management in Educational Data: A Case Study of Statistics Polytechnic): Nghiên cứu này cung cấp cái nhìn sâu sắc về quản lý chất lượng dữ liệu trong lĩnh vực giáo dục, với một trường hợp nghiên cứu cụ thể về dữ liệu thống kê tại một trường bách khoa.

## 2.2. Tình hình nghiên cứu tại Việt Nam

Tại Việt Nam, nghiên cứu về hệ thống đánh giá chất lượng khóa học dựa trên dữ liệu từ các nền tảng MOOCs đang có xu hướng phát triển, mặc dù còn khá mới mẻ so với các quốc gia khác. Trong bối cảnh các nền tảng học trực tuyến như Coursera, edX, và Udemy ngày càng phổ biến, nhu cầu đánh giá chất lượng các khóa học trực tuyến trở nên quan trọng hơn bao giờ hết. Các nghiên cứu trong nước chủ yếu tập trung vào việc phát triển các phương pháp đánh giá dựa trên dữ liệu hành vi học tập của học viên, phản hồi từ học viên, và kết quả học tập để đo lường mức độ hiệu quả của các khóa học.

Mặc dù một số nghiên cứu về hệ thống đánh giá chất lượng khóa học MOOCs đã được thực hiện, nhưng hầu hết tập trung vào việc áp dụng các phương pháp truyền thống như phân tích điểm số học viên, tỷ lệ hoàn thành khóa học, và mức độ tham gia vào các hoạt động học tập như bài tập và thảo luận. Những nghiên cứu này thường sử dụng các chỉ số đơn giản để đánh giá chất lượng khóa học mà không quá chú trọng đến các yếu tố phức tạp hơn như sự tương tác giữa các học viên, mức độ động viên của học viên, và mức độ cá nhân hóa trong các khóa học.

Các nghiên cứu gần đây bắt đầu chú trọng đến việc phát triển các mô hình đánh giá chất lượng khóa học bằng các phương pháp phân tích dữ liệu phức tạp hơn, như sử dụng học máy (Machine Learning) và học sâu (Deep Learning) để nhận diện các yếu tố tiềm ẩn ảnh hưởng đến chất lượng khóa học. Tuy nhiên, những phương pháp này vẫn còn trong giai đoạn thử nghiệm và chưa được ứng dụng rộng rãi tại các nền tảng MOOCs ở Việt Nam.

# CHƯƠNG 3: PHÂN TÍCH TẬP DỮ LIỆU

## 3.1 Bộ dữ liệu MOOCCubeX

MOOCCubeX là một bộ dữ liệu quy mô lớn, được duy trì bởi Nhóm Kỹ thuật Tri thức của Đại học Tsinghua và được hỗ trợ bởi XuetangX, một trong những nền tảng MOOCs lớn nhất tại Trung Quốc. Bộ dữ liệu này cung cấp một kho tài nguyên phong phú với 4.216 khóa học, 230.263 video, 358.265 bài tập, 637.572 khái niệm chi tiết, cùng với hơn 296 triệu dữ liệu hành vi thô của 3.330.294 sinh viên. Mục tiêu chính của MOOCCubeX là hỗ trợ nghiên cứu về các chủ đề học tập thích ứng trong môi trường MOOCs.

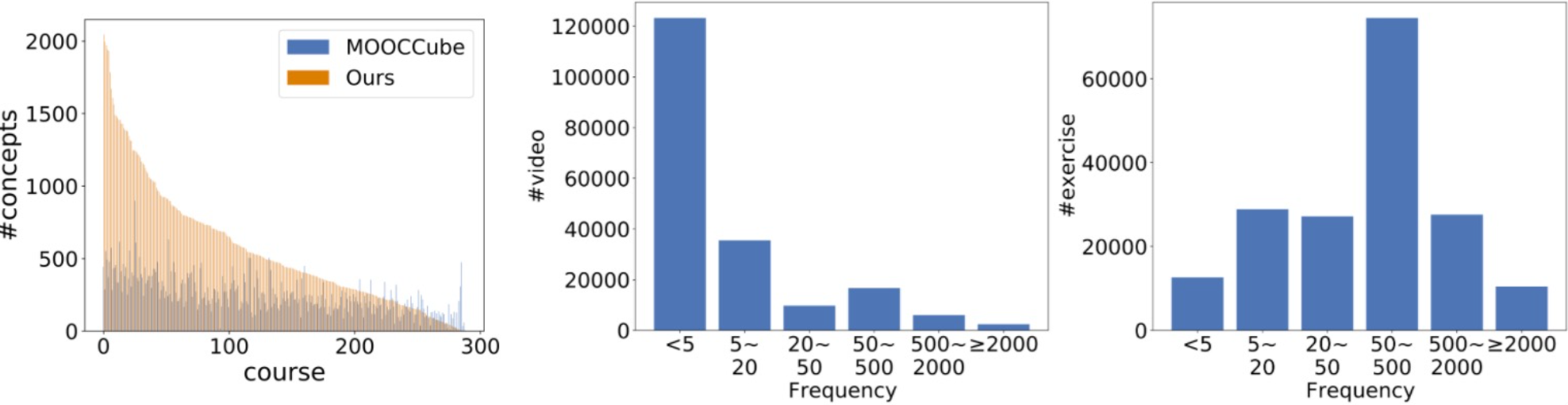
Các điểm nổi bật của MOOCCubeX có thể được tóm tắt như sau:

1. **Phạm vi bao quát rộng**: MOOCCubeX không chỉ thu thập các tài nguyên MOOCs đa dạng mà còn bao gồm các tài nguyên giáo dục bên ngoài và dữ liệu hành vi của sinh viên liên quan đến học tập, bài tập và thảo luận. Điều này cung cấp cái nhìn toàn diện về quá trình học tập và tương tác của học viên trong môi trường trực tuyến.
2. **Quy mô lớn**: So với các bộ dữ liệu giáo dục công khai khác, MOOCCubeX có quy mô lớn hơn đáng kể, giúp hỗ trợ việc khám phá các mô hình học sâu với yêu cầu về dữ liệu lớn. Điều này mang đến cơ hội nghiên cứu và phát triển các phương pháp phân tích dữ liệu phức tạp hơn trong bối cảnh học trực tuyến.
3. **Chú trọng vào khái niệm**: Dữ liệu trong MOOCCubeX được tổ chức xung quanh các khái niệm chi tiết, giúp làm cho tài nguyên dễ dàng hơn để đại diện, tìm kiếm và mô hình hóa. Cách tổ chức này cũng hỗ trợ việc xây dựng các mô hình học máy và học sâu có thể khai thác thông tin từ dữ liệu một cách hiệu quả và chính xác hơn.

MOOCCubeX cung cấp một số đặc điểm thống kê nổi bật về các khái niệm và dữ liệu hành vi của người học, giúp làm rõ các đặc điểm của dữ liệu và các xu hướng hành vi trong môi trường MOOCs.

1. **Khái niệm chi tiết hơn**: MOOCCubeX có số lượng khái niệm chi tiết (fine-grained concepts) lớn hơn so với phiên bản trước đó của bộ dữ liệu, MOOCCube. Việc tổ chức dữ liệu dựa trên các khái niệm chi tiết giúp tăng cường tính liên quan và dễ dàng cho việc đại diện và mô hình hóa các tài nguyên học tập. Điều này cho phép nghiên cứu sâu hơn về các yếu tố ảnh hưởng đến quá trình học tập và các mối quan hệ giữa các khái niệm trong các khóa học.
2. **Hành vi xem video có phân bố dài đuôi (long-tail)**: Dữ liệu về hành vi xem video trong MOOCCubeX cho thấy phân bố dài đuôi, điều này có nghĩa là số lượng người học tham gia vào việc xem các video học tập có sự phân bổ không đều. Một số video nhận được lượng xem rất lớn, trong khi phần lớn video còn lại có lượng xem ít hơn nhiều. Phân tích hành vi xem video với đặc điểm này có thể giúp các nhà nghiên cứu hiểu rõ hơn về các yếu tố thu hút sự chú ý của người học và cách tối ưu hóa nội dung video.
3. **Hành vi làm bài tập có phân bố chuẩn (normal distribution)**: Mặt khác, hành vi làm bài tập của người học trong MOOCCubeX có phân bố chuẩn, cho thấy một số lượng lớn người học thực hiện bài tập với tần suất vừa phải, trong khi một số ít người tham gia vào việc làm bài tập nhiều hơn hoặc ít hơn. Phân tích dữ liệu này có thể giúp các nhà nghiên cứu và nhà phát triển hiểu rõ hơn về các mô hình học tập, từ đó cải thiện các công cụ hỗ trợ học tập và đánh giá chất lượng khóa học.

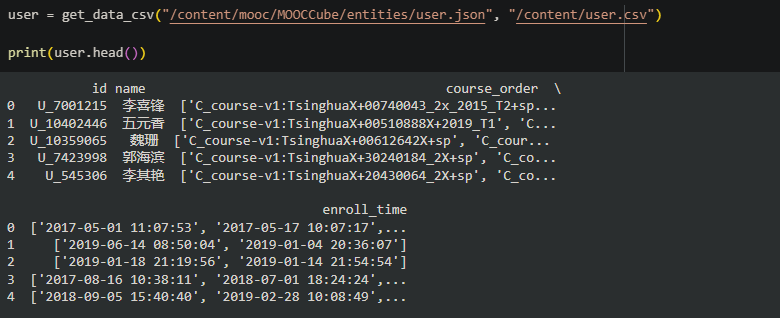
Những đặc điểm thống kê này cho thấy MOOCCubeX là một bộ dữ liệu rất hữu ích cho việc nghiên cứu hành vi học tập trong MOOCs và cung cấp thông tin chi tiết để phát triển các hệ thống đánh giá chất lượng khóa học, giúp tối ưu hóa trải nghiệm học tập cho người học.

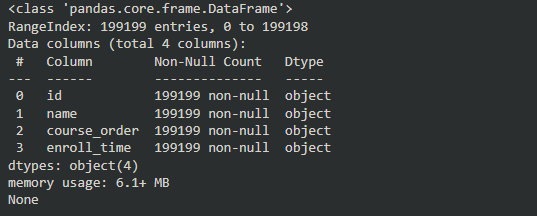


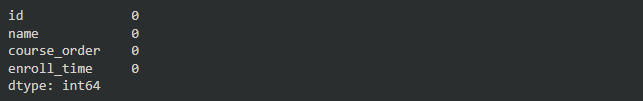
## 3.2. Mô tả dữ liệu và thăm dò dữ liệu

**entities/user**Tập dữ liệu này chứa thông tin chi tiết về người học trong hệ thống. Các trường dữ liệu trong tệp **user.json** bao gồm:

* **id**: Mã số người học, bắt đầu bằng "U\_".
* **name**: Tên của người học.
* **course\_order**: Mảng chứa các khóa học mà người học đã đăng ký.
* **enroll\_time**: Mảng thời gian tương ứng với thời gian người học đăng ký các khóa học trong **course\_order**.







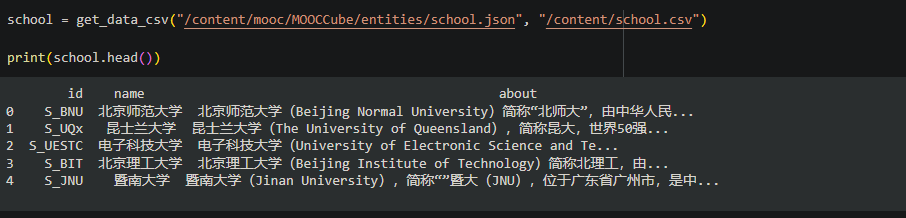
*Code thể hiện thông tin từ user*

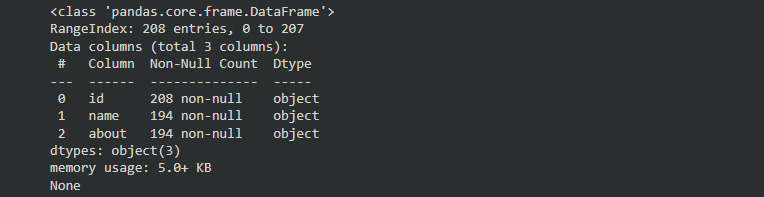
Dữ liệu User gồm 199199 entries trong đó không có tồn tại dữ liệu Nan hay null

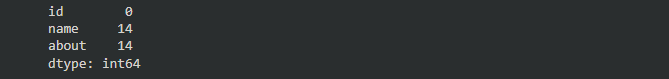
**Mối quan hệ**: Dữ liệu trong tệp này liên kết với tệp **course-user.txt** để xác định các khóa học mà mỗi người học đã tham gia.

**entities/school**Tập dữ liệu này mô tả các trường học có trong hệ thống MOOCs. Các trường trong tệp **school.json** bao gồm:

* **id**: Mã số trường học, bắt đầu bằng "S\_".
* **name**: Tên trường học bằng tiếng Trung.
* **name\_en**: Tên trường học bằng tiếng Anh.
* **sign**: Chữ viết tắt của tên trường học bằng tiếng Anh.
* **about**: Giới thiệu về trường học.
* **motto**: Châm ngôn của trường học.







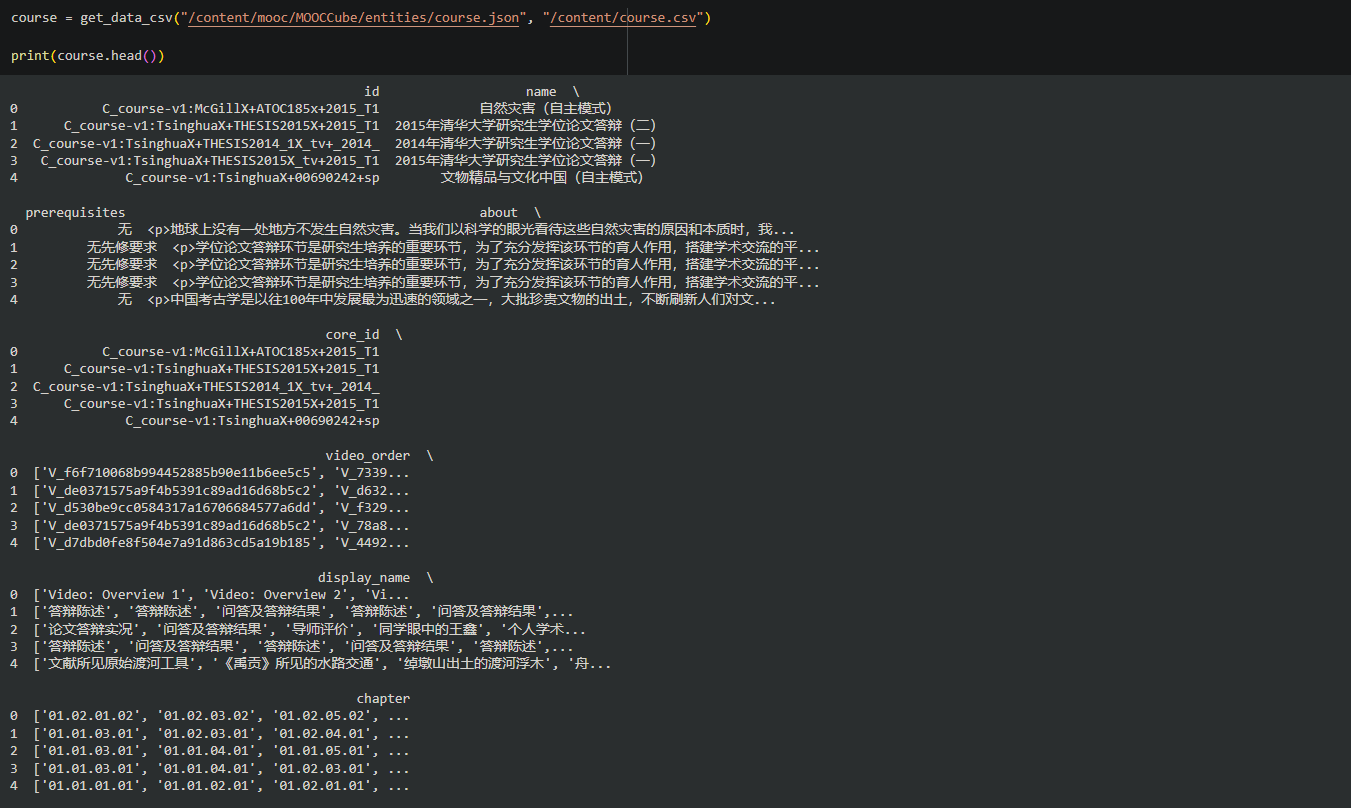
*Code thể hiện thông tin từ school*

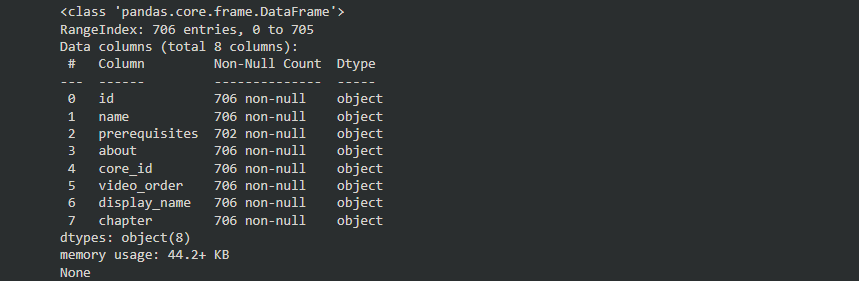
Dữ liệu school gồm 208 entries trong đó có 14 hàng có dữ liệu Nan thuộc về 2 thuộc tính name và about

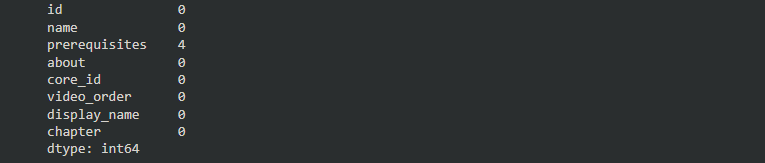
**Mối quan hệ**: Dữ liệu trong tệp này được kết nối với tệp **course-school.txt** để mô tả các trường học giảng dạy các khóa học.

**entities/course**Tập dữ liệu này cung cấp thông tin chi tiết về các khóa học có trong hệ thống MOOCs. Các trường trong tệp **course.json** bao gồm:

* **about**: Giới thiệu về khóa học.
* **id**: Mã số khóa học.
* **field**: Danh sách các lĩnh vực của khóa học (ví dụ: Toán học, Khoa học máy tính, v.v.).
* **name**: Tên khóa học.
* **prerequisites**: Mô tả về kiến thức cần có trước khi tham gia khóa học.
* **resource**: Danh sách tài nguyên học tập của khóa học (video, bài tập, tài liệu học, v.v.).







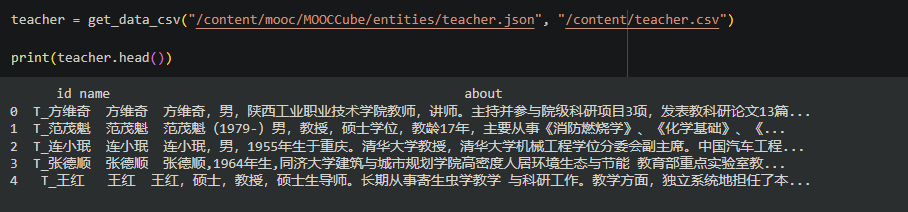
*Code thể thông tin về khóa học*

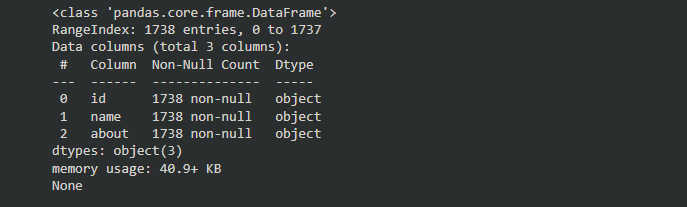
Dữ liệu course gồm 706 entries trong đó có 4 hàng có dữ liệu Nan thuộc thuộc tính prerquisites

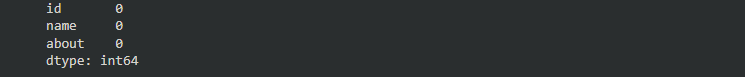
**Mối quan hệ**: Dữ liệu trong tệp này kết nối với tệp **course-user.txt** (người học tham gia khóa học) và **course-teacher.txt** (giảng viên dạy khóa học).

**entities/teacher**Tập dữ liệu này mô tả các giảng viên giảng dạy các khóa học. Các trường trong tệp **teacher.json** bao gồm:

* **id**: Mã số giảng viên, bắt đầu bằng "T\_".
* **name**: Tên giảng viên bằng tiếng Trung.
* **about**: Hồ sơ giảng viên, bao gồm các thông tin về sự nghiệp, chuyên môn, v.v.







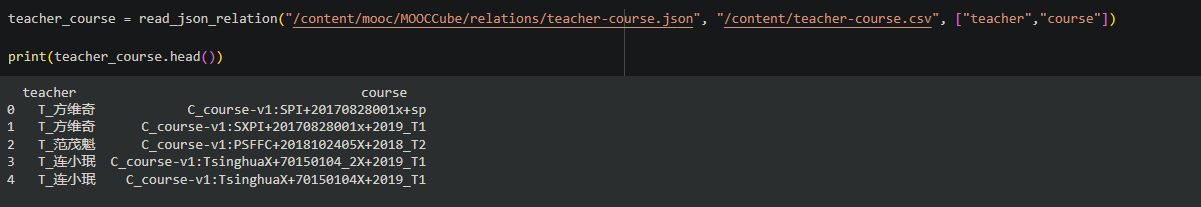
*Code thể hiện thông tin từ teacher*

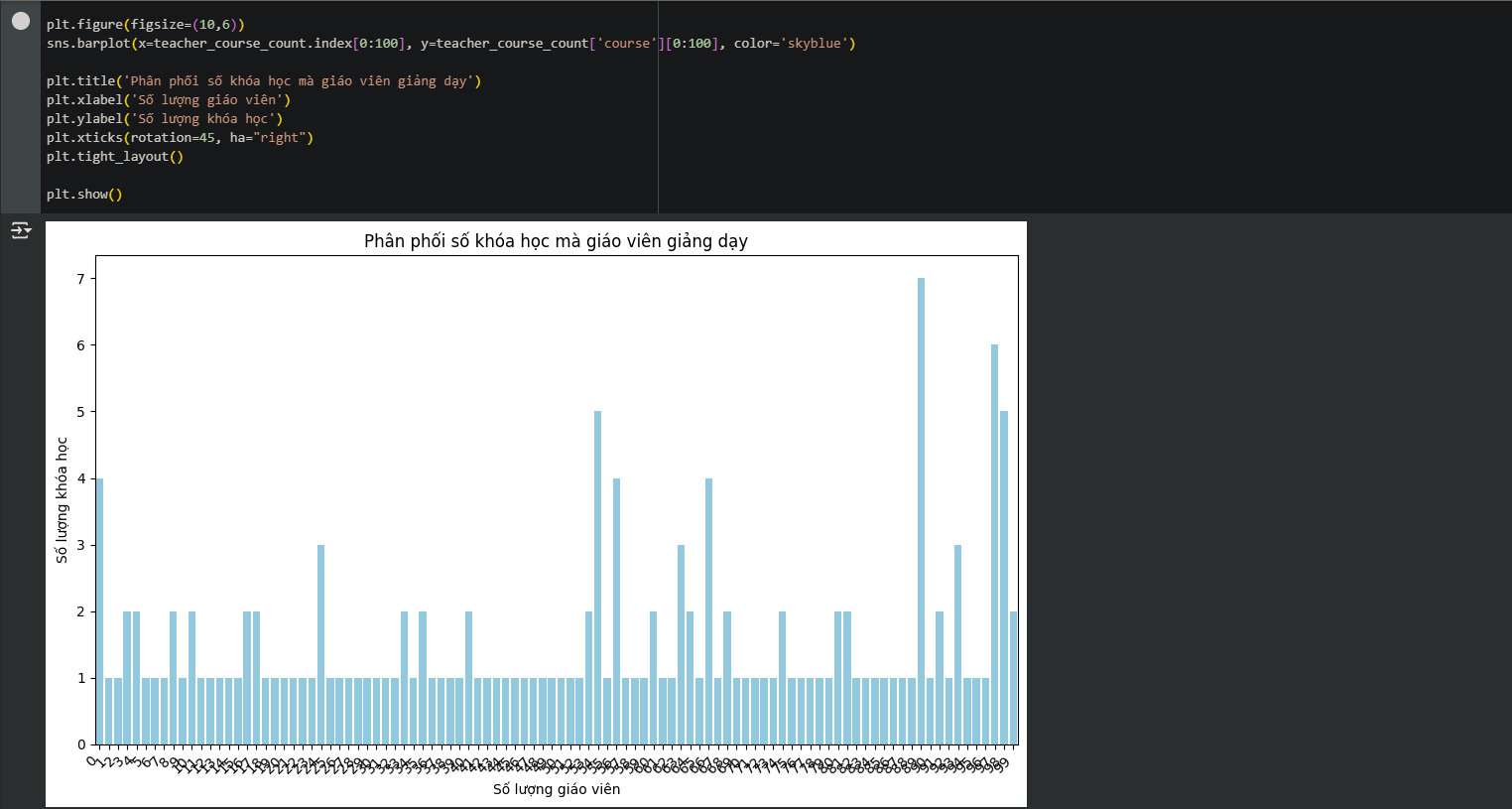
Dữ liệu teacher gồm 1738 entries trong đó không có dữ liệu Nan

**Mối quan hệ**: Dữ liệu trong tệp này được kết nối với tệp **course-teacher.txt** (mối quan hệ giữa khóa học và giảng viên) và **school-teacher.txt** (mối quan hệ giữa giảng viên và trường học).

**relations/course-teacher**Tệp này lưu trữ các mối quan hệ giữa khóa học và giảng viên. Cấu trúc dữ liệu bao gồm:

* **course ID**: Mã khóa học.
* **teacher ID**: Mã giảng viên.



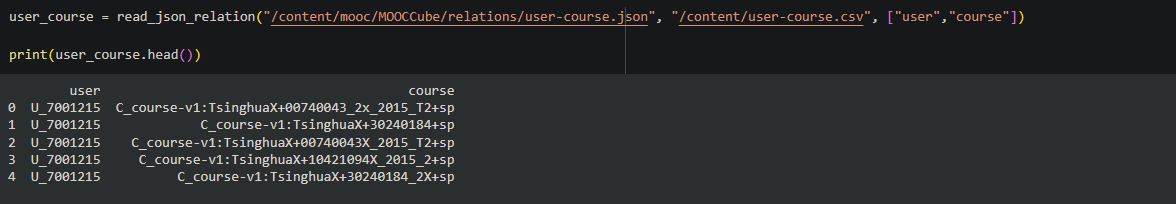


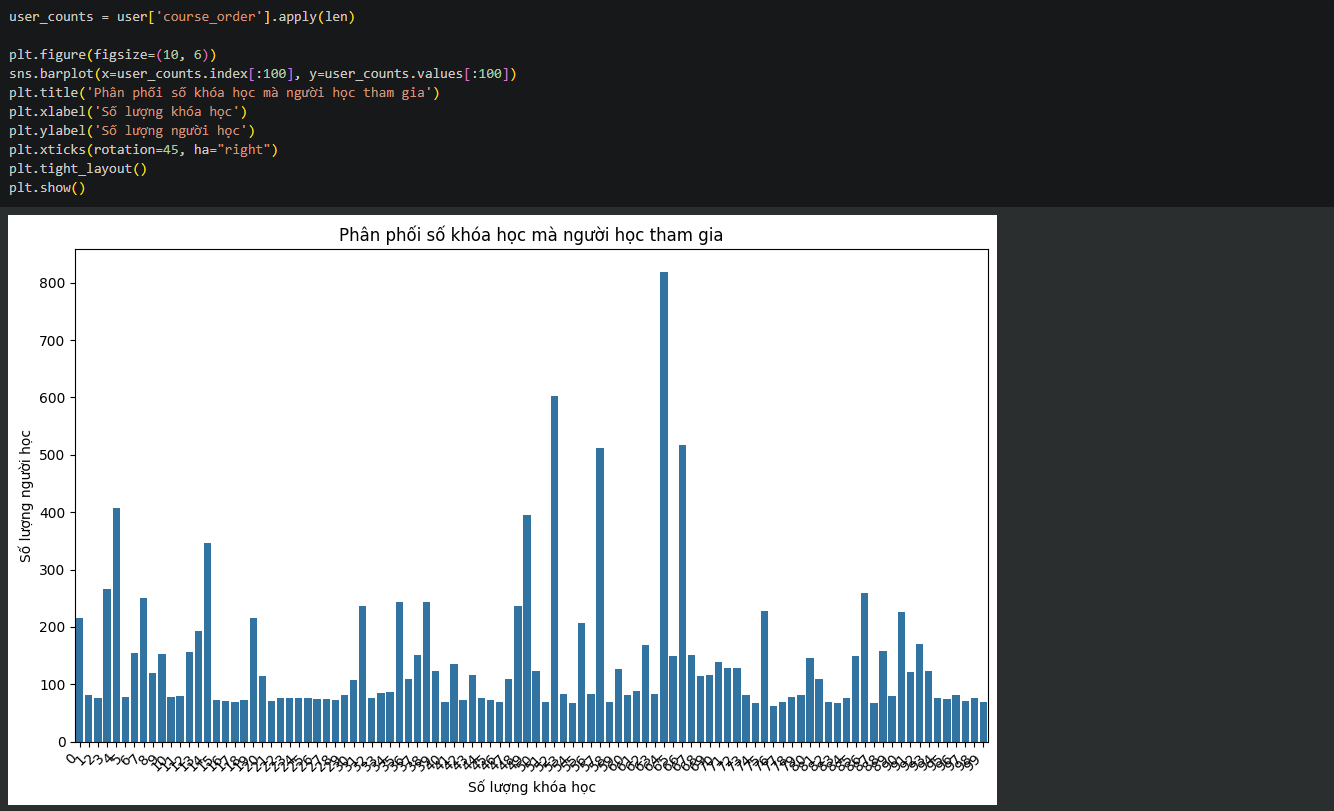
Trung bình một giáo viên giảng dạy 4-5 khóa học, sự phân bố tương đối đồng đều không có sự chênh lệch nhiều giữa các giáo viên khác nhau

Mối quan hệ này giúp xác định giảng viên nào đang giảng dạy khóa học nào trong hệ thống.

**relations/course-user**Tệp này mô tả mối quan hệ giữa người học và khóa học. Cấu trúc dữ liệu là:

* **course ID**: Mã khóa học.
* **user ID**: Mã người học.



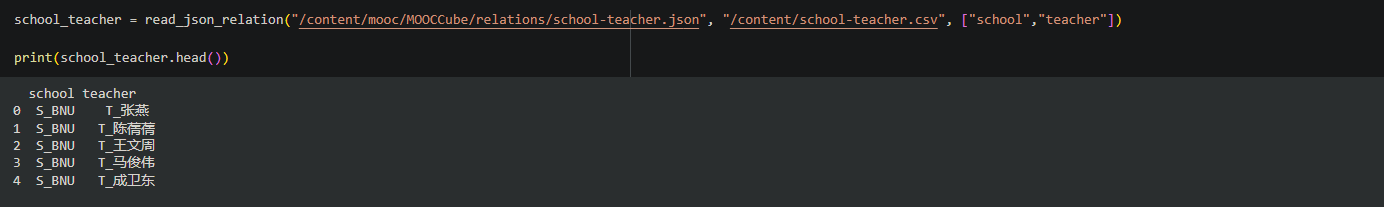


Số lượng học sinh có sự phân bố không đều giữa các khóa học, qua đó ta thấy sự phản ánh tương đối giữa chất lượng của từng khóa học thông qua lượng học sinh.

Mối quan hệ này giúp xác định người học nào đã đăng ký và tham gia khóa học nào trong hệ thống.

**relations/school-teacher**Tệp này chứa mối quan hệ giữa giảng viên và trường học. Dữ liệu trong tệp được lưu theo định dạng:

* **teacher ID**: Mã giảng viên.
* **school ID**: Mã trường học.

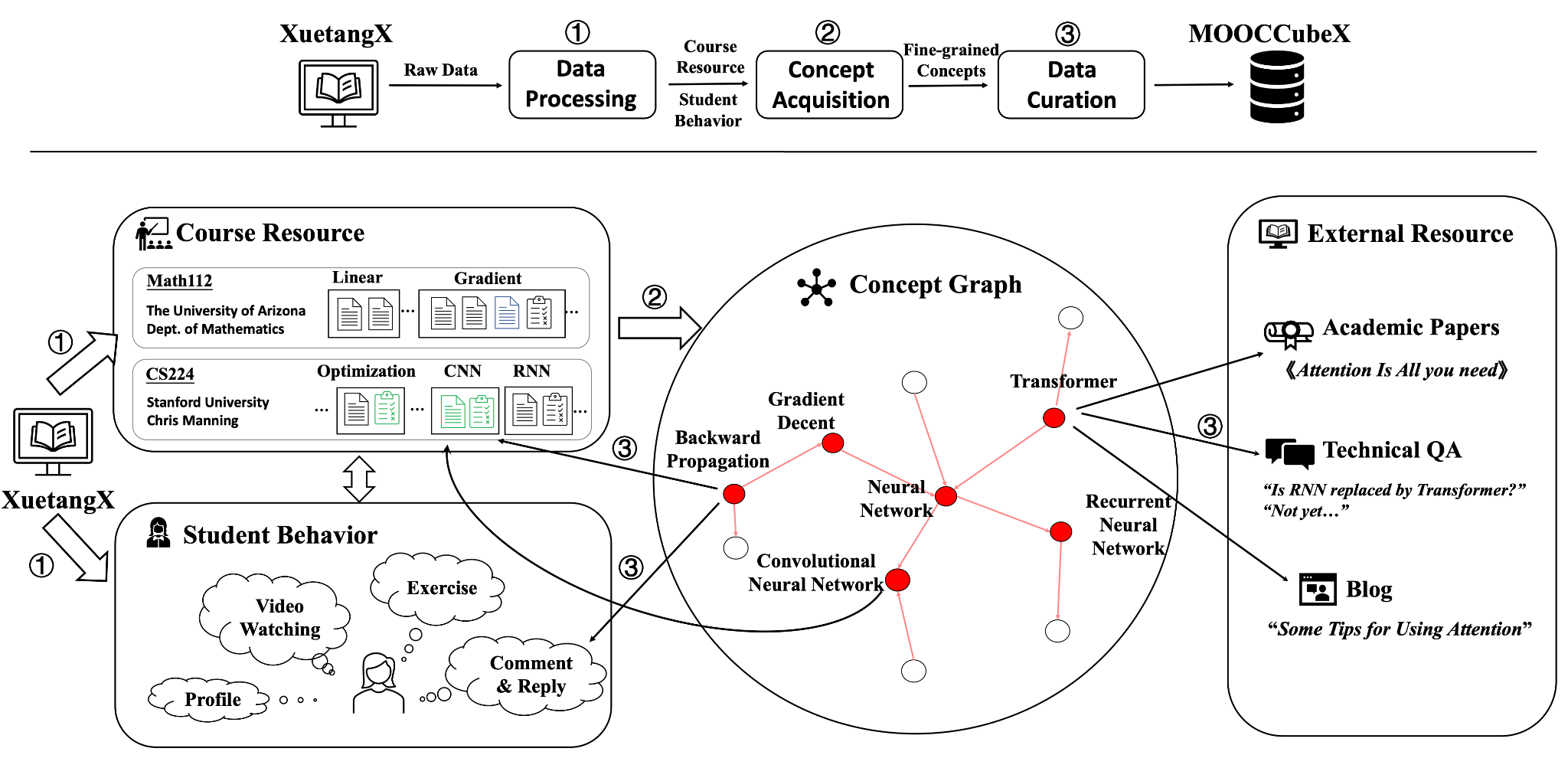
Mối quan hệ này cho biết giảng viên nào đang làm việc tại trường học nào.

# CHƯƠNG 4: PHƯƠNG PHÁP ĐỀ XUẤT

## 4.1. Hướng triển khai đề tài

* Triển khai đề tài theo các bước chính:
* Chọn các thuộc tính phù hợp từ bộ dữ liệu MOOCCubeX cho bài toán. Sau đó chọn ra các tiêu chí phù hợp cho chất lượng đầu vào khóa học.
* Chạy mô hình để đánh giá các tiêu chí trước và sau xử lý lại dữ liệu.

## 4.2. Framework thực hiện

****

* ***Bước 1:*** Thu thập dữ liệu, trong đề tài này nhóm sẽ thu thập dữ liệu từ bộ dữ liệu MOOCCubeX.
* ***Bước 2:*** Trích xuất các tiêu chí cho việc đánh giá chất lượng đầu vào của khóa học từ các file
* ***Bước 3*:** Tổng hợp các tiêu chí vào một file tổng hợp và gắn nhãn - trước khi xử lý lại dữ liệu
* ***Bước 4:*** Chạy mô hình trên hai file tổng hợp đã gắn nhãn

## 4.3. Xây dựng mạng

Đồ thị mạng là một cấu trúc dữ liệu mạnh mẽ và linh hoạt, được sử dụng rộng rãi để mô hình hóa các mối quan hệ phức tạp giữa các đối tượng trong nhiều lĩnh vực khác nhau. Trong một đồ thị mạng, các đối tượng hay thực thể được biểu diễn dưới dạng các nút (nodes), còn các mối quan hệ hoặc sự kết nối giữa chúng được thể hiện qua các cạnh (edges). Cấu trúc này không chỉ giúp chúng ta hình dung các mối quan hệ một cách trực quan mà còn tạo ra nền tảng vững chắc để phân tích và khai thác các thông tin liên quan đến sự kết nối giữa các đối tượng.

Thông qua việc sử dụng đồ thị mạng, người dùng có thể dễ dàng nhận diện các mô hình kết nối và sự phụ thuộc giữa các thành phần trong hệ thống, từ đó tạo ra những nhận định và giải pháp hiệu quả cho các vấn đề phức tạp. Đồ thị mạng không chỉ hỗ trợ trong việc ra quyết định, mà còn đóng vai trò quan trọng trong tối ưu hóa các quy trình, phân tích dữ liệu, và phát hiện các mối quan hệ tiềm ẩn mà các phương pháp phân tích dữ liệu truyền thống có thể không nhận ra. Việc áp dụng đồ thị mạng trong các nghiên cứu và ứng dụng thực tế, chẳng hạn như trong các hệ thống khuyến nghị, mạng xã hội, và phân tích dữ liệu lớn, đã mở ra nhiều cơ hội mới trong việc giải quyết các vấn đề đa dạng và phức tạp.

Nhờ tính linh hoạt và khả năng mô hình hóa mạnh mẽ, đồ thị mạng đã trở thành một công cụ quan trọng trong nhiều lĩnh vực nghiên cứu và ứng dụng thực tiễn, giúp chúng ta không chỉ hiểu rõ hơn về cấu trúc các mối quan hệ mà còn tối ưu hóa và cải thiện các quy trình trong các hệ thống phức tạp.

Các thành phần của mạng:

Các nút trong đồ thị được phân loại thành 4 loại khác nhau, mỗi loại có màu sắc riêng:

- Khóa học (Course): Các khóa học cũng được đại diện bởi các nút, chứa thông tin như tên khóa học, mô tả và các yêu cầu tiên quyết. Các đặc trưng của khóa học được lưu trữ dưới dạng vector embeddings để phục vụ cho việc phân tích và tìm kiếm.

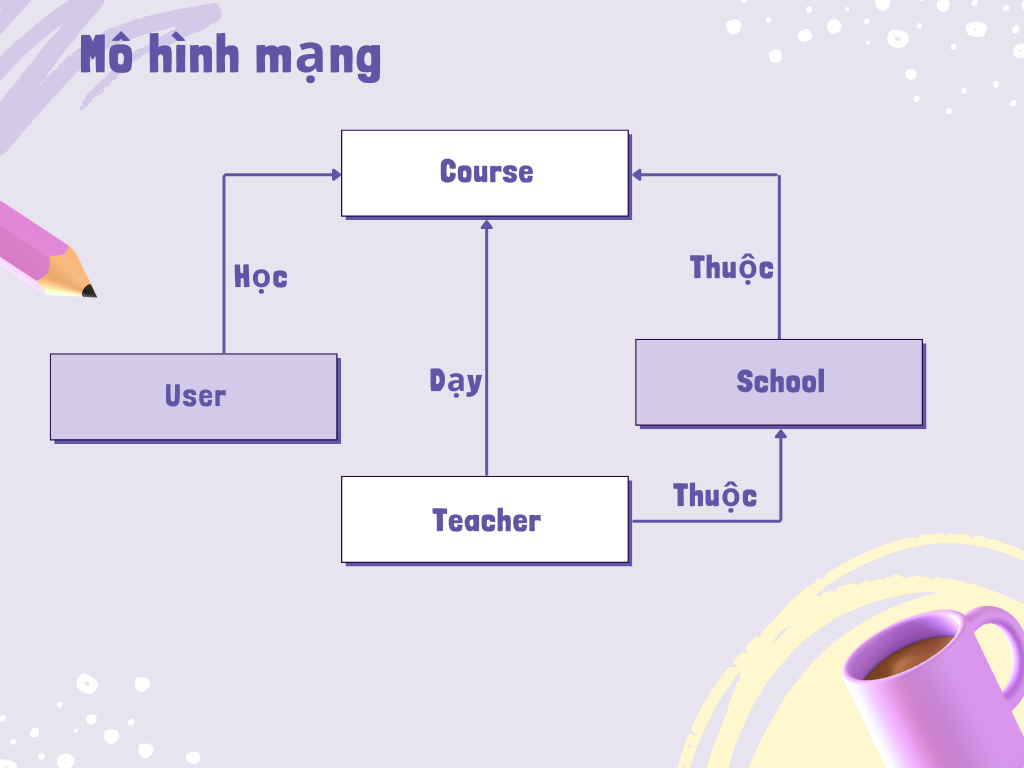
- Học sinh (User): Người dùng được biểu diễn bằng nút với thông tin về tên.

- Trường học (School): Mỗi trường học được biểu diễn bằng một nút với thông tin về tên trường và mô tả.

- Giáo viên (Teacher): Các giáo viên được đại diện bằng các nút với thông tin tương tự như tên và mô tả

Các cạnh trong mạng:

* Course-user: Mối quan hệ giúp xác định người học nào đã đăng ký và tham gia khóa học nào trong hệ thống.
* School-teacher: Mối quan hệ cho biết giảng viên nào đang làm việc tại trường học nào.
* Course-teacher: Mối quan hệ giúp xác định giảng viên nào đang giảng dạy khóa học nào trong hệ thống.



*Xây dựng mô hình mạng các nodes và edges*

# CHƯƠNG 5: KẾT QUẢ THỰC NGHIỆM

## 5.1. Môi trường thực nghiệm

Nhóm đã triển khai các thí nghiệm trên nền tảng Google Colaboratory với cấu hình bao gồm:

Hệ điều hành: Ubuntu 18.04.6.

Bộ nhớ RAM: 12GB.

GPU: NVIDIA CUDA.

Ngôn ngữ lập trình: Python.

Thư viện mã nguồn mở: sử dụng các thư viện như scikit-learn, numpy, pandas, matplotlib, và seaborn.

Việc lựa chọn môi trường Google Colaboratory giúp nhóm tận dụng tài nguyên GPU miễn phí để tăng tốc quá trình xử lý và huấn luyện mô hình.

## 5.2. Giới hạn dữ liệu đầu vào

Dữ liệu đầu vào được trích xuất từ bộ dữ liệu MOOCCubeX với các tệp chính như sau:

1. Course.json:

Sử dụng các thuộc tính: id, resource.

1. User.json:

Sử dụng các thuộc tính: id, school, course\_order.

1. School.json:

Sử dụng các thuộc tính: id, name, about.

1. Teacher.json:

Sử dụng các thuộc tính: id, job\_title, about.

1. User-problem.json:

Sử dụng các thuộc tính: user\_id, exercise\_id, problem\_id, score.

Dữ liệu được xử lý để loại bỏ các giá trị thiếu (missing values) và ngoại lai (outliers), đảm bảo đầu vào cho mô hình là sạch và đáng tin cậy.

## 5.3. Các tiêu chí đánh giá chất lượng khóa học

Để đánh giá chất lượng khóa học, nhóm đã xây dựng tiêu chí chính là tỷ lệ điểm số đạt được so với điểm tối đa. Phương pháp tính như sau:

Nguồn dữ liệu: Từ các tệp user-problem.json, course.json, và problem.json.

**Quy trình:**

Tính tổng điểm số mà các user đạt được trong từng bài tập (problem\_id) thuộc mỗi bài thực hành (exercise\_id).

Chia tỷ lệ tổng điểm đạt được so với tổng điểm tối đa trong bài thực hành đó.

Kết quả đầu ra là tỷ lệ điểm số cho từng khóa học.

Kết quả phân tích này cung cấp cái nhìn trực quan về hiệu quả học tập của học viên trên từng khóa học cụ thể.

## 5.4. Đánh giá mô hình bằng SVM

**Mô hình sử dụng:**

* SVM (Support Vector Machine): được lựa chọn vì khả năng xử lý tốt các bài toán phân loại và khả năng làm việc hiệu quả với dữ liệu không tuyến tính.

Tham số chính: kernel, C, gamma được tối ưu thông qua phương pháp tìm kiếm lưới (Grid Search).

* Random Forest Classifier: Là một thuật toán ensemble learning, kết hợp nhiều cây quyết định để đưa ra quyết định cuối cùng. Ưu điểm của Random Forest là khả năng xử lý tốt dữ liệu có nhiều đặc trưng, giảm thiểu overfitting và có thể xử lý cả các bài toán hồi quy và phân loại.

Tham số chính: Số lượng cây (n\_estimators), độ sâu tối đa của cây (max\_depth), số lượng đặc trưng được chọn ngẫu nhiên tại mỗi node (max\_features), tiêu chí phân chia node (criterion).

* Gradient Boosting Classifier: Là một thuật toán ensemble learning, tạo ra các cây quyết định tuần tự, mỗi cây cố gắng sửa lỗi của cây trước đó. Gradient Boosting thường có độ chính xác cao nhưng dễ bị overfitting nếu không được điều chỉnh cẩn thận.

Tham số chính: Số lượng cây (n\_estimators), tỷ lệ học (learning\_rate), độ sâu tối đa của cây (max\_depth), tiêu chí phân chia node.

**Phương pháp đánh giá:**

Sử dụng kỹ thuật K-fold Cross Validation để đảm bảo tính khách quan và tránh hiện tượng overfitting. Tập dữ liệu được chia thành K phần bằng nhau, mỗi lần lặp sử dụng K-1 phần để huấn luyện và phần còn lại để đánh giá.

**Độ đo hiệu suất:**

Accuracy: Tỷ lệ dự đoán chính xác trên toàn bộ tập dữ liệu.

Precision: Đo lường tỷ lệ dự đoán đúng trên tổng số dự đoán là nhãn dương. Precision cao nghĩa là mô hình ít nhầm lẫn khi dự đoán nhãn dương, đảm bảo độ chính xác cao.

Recall: Đo lường tỷ lệ dự đoán đúng trên tổng số nhãn dương thực tế. Recall cao nghĩa là mô hình không bỏ sót các trường hợp nhãn dương, đảm bảo tính bao phủ.  
F1-score: Là trung bình điều hòa giữa Precision và Recall. F1-score cao nghĩa là mô hình cân bằng tốt giữa độ chính xác và khả năng bao phủ, đặc biệt hữu ích với dữ liệu mất cân bằng.

## 5.5. Kết quả thực nghiệm

**Kết quả chính:**

Độ chính xác trung bình (Accuracy) đạt 90%, cho thấy mô hình hoạt động hiệu quả với dữ liệu đã qua xử lý.

Độ đo chi tiết trên từng lớp nhãn (Precision, Recall, F1-score) cho thấy sự cân bằng trong khả năng dự đoán.

**Kết quả phân tích dữ liệu mạng:**

Thông qua việc xây dựng đồ thị mạng, nhóm đã xác định được các nút quan trọng dựa trên chỉ số độ trọng tâm (centrality) và mối quan hệ giữa các nút.

Phân cụm các nút (clustering) cho thấy các cộng đồng học viên có hành vi học tập tương đồng, giúp cải thiện chất lượng phân tích và gợi ý cải tiến khóa học.

**Ảnh hưởng của xử lý dữ liệu:**

Kết quả thực nghiệm chỉ ra rằng việc xử lý dữ liệu (loại bỏ giá trị thiếu, chuẩn hóa dữ liệu) giúp cải thiện đáng kể hiệu suất mô hình, với độ chính xác tăng khoảng 15% so với khi sử dụng dữ liệu chưa xử lý.

## 5.6. Phân tích thêm

Phân phối điểm số: Các khóa học có tỷ lệ điểm số đạt được phân phối không đều, phản ánh chất lượng và mức độ khó của từng khóa học.

Hành vi học viên: Số lượng học viên đăng ký và hoàn thành khóa học thay đổi đáng kể, cho thấy nhu cầu và mức độ cam kết của học viên đối với các khóa học là khác nhau.

Phân tích chuyên sâu: Các yếu tố như trường học, giảng viên, và nội dung khóa học có tác động mạnh mẽ đến tỷ lệ hoàn thành và chất lượng học tập.

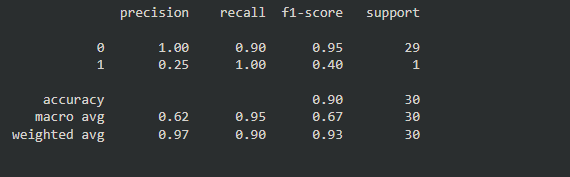
## 5.7. Đánh giá chung

Random Forest là mô hình có hiệu suất tổng thể tốt nhất, thể hiện qua các chỉ số Accuracy, Precision, Recall và F1-score cao nhất. Điều này cho thấy Random Forest có khả năng phân loại chính xác, ít bị nhầm lẫn và bao quát được hầu hết các mẫu dương.

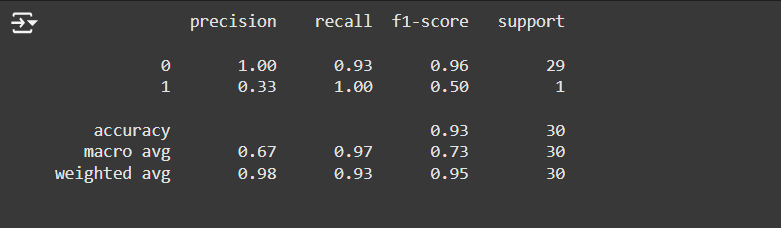
SVM cũng đạt được hiệu suất khá tốt, đặc biệt là về Precision, cho thấy mô hình này rất ít khi phân loại sai một mẫu âm thành mẫu dương. Tuy nhiên, Recall của SVM hơi thấp hơn so với Random Forest, nghĩa là mô hình này có thể bỏ sót một số mẫu dương.

Gradient Boosting có hiệu suất thấp nhất trong ba mô hình, đặc biệt là về Recall, cho thấy mô hình này có thể bỏ sót nhiều mẫu dương hơn so với các mô hình còn lại.

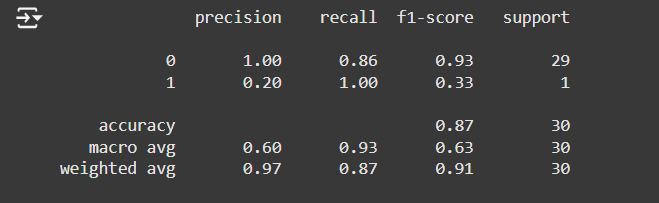
Dựa trên kết quả trên, Random Forest là mô hình phù hợp nhất cho bài toán này.



*Các thông số kết quả sau khi huấn luyện mô hình svm*

**

*Cac thông số kết quả sau khi huấn luyện mô hình Random Forest Classifier*

**

*Cac thông số kết quả sau khi huấn luyện mô hình Gradient Boosting Classifier*

**Kết quả huấn luyện các models**

|  | SVM | Random Forest Classifier | Gradient Boosting Classifier |
| --- | --- | --- | --- |
| accuracy | 0.9 | 0.93 | 0.87 |
| precision | 1.0 | 1.0 | 1.0 |
| recall | 0.9 | 0.93 | 0.86 |
| F1 - score | 0.95 | 0.96 | 0.93 |

# CHƯƠNG 6: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## 6.1. Kết luận

Trong đồ án “Đánh giá chất lượng bộ dữ liệu phục vụ quản lý chất lượng đầu vào khóa học,” nhóm chúng tôi đã thực hiện các bước nghiên cứu, thu thập và phân tích dữ liệu để phát triển hệ thống quản lý chất lượng đầu vào cho các khóa học trực tuyến dựa trên bộ dữ liệu MOOCCubeX. Qua quá trình nghiên cứu và thực nghiệm, chúng tôi đã đạt được những kết quả quan trọng:

Hiệu quả của mô hình: Mô hình SVM (Support Vector Machine) được áp dụng cho bài toán đã cho thấy hiệu suất cao với độ chính xác trung bình đạt 90% trên tập dữ liệu đã qua xử lý. Điều này chứng minh rằng các tiêu chí và phương pháp đánh giá mà nhóm đề xuất là phù hợp và có giá trị thực tiễn trong việc quản lý và đánh giá chất lượng dữ liệu.

Tầm quan trọng của xử lý dữ liệu: Kết quả nghiên cứu cũng cho thấy rằng việc xử lý dữ liệu đóng vai trò rất quan trọng trong việc nâng cao hiệu suất của mô hình. Các tiêu chí đánh giá chất lượng dữ liệu như tính đầy đủ, tính nhất quán, và độ chính xác đã được cải thiện đáng kể thông qua quá trình xử lý.

Đóng góp cho lĩnh vực giáo dục trực tuyến: Hệ thống quản lý chất lượng dữ liệu đầu vào không chỉ giúp cải thiện chất lượng các khóa học mà còn hỗ trợ các nhà quản lý giáo dục, giảng viên và học viên có trải nghiệm học tập tốt hơn, từ đó nâng cao hiệu quả học tập và mức độ hài lòng của người học.

Tuy nhiên, bên cạnh những kết quả tích cực, chúng tôi cũng nhận thấy một số hạn chế, như khả năng áp dụng mô hình trên tập dữ liệu phi cấu trúc và việc đối phó với các giá trị thiếu dữ liệu phức tạp. Đây là những điểm cần được tiếp tục nghiên cứu và cải thiện.

## 6.2. Hướng phát triển

Để nâng cao hiệu quả của hệ thống và mở rộng phạm vi ứng dụng, nhóm đã đề xuất một số hướng phát triển trong tương lai, tập trung vào việc cải thiện cả dữ liệu đầu vào lẫn mô hình phân tích:

Cải thiện việc xây dựng mô hình mạng: Tiếp tục nâng cao khả năng phân tích dữ liệu bằng cách triển khai các chỉ số mạng như độ trọng tâm của các nút (node centrality), phân cụm các nút và liên kết (clustering coefficient) để hiểu rõ hơn về cấu trúc và mối quan hệ giữa các thành phần trong mạng. Những phân tích này sẽ hỗ trợ mô hình trong việc tối ưu hóa các nút quan trọng và phát hiện các cộng đồng trong mạng, từ đó nâng cao độ chính xác và hiệu quả trong việc đánh giá chất lượng khóa học.

Phân tích dữ liệu của các node chi tiết hơn: Thực hiện phân tích chuyên sâu hơn về dữ liệu của các node khác nhau trong hệ thống như người học (user), trường học (school), khóa học (course), và giảng viên (teacher). Điều này bao gồm việc khai thác sâu hơn các thuộc tính liên quan đến mối quan hệ giữa các thành phần, hành vi người học, mức độ tham gia và sự tương tác trong hệ thống.

Ứng dụng các phương pháp học sâu (Deep Learning): Nghiên cứu và áp dụng các mô hình học sâu như mạng nơ-ron đồ thị (Graph Neural Networks - GNNs) để tối ưu hóa việc phân tích cấu trúc đồ thị và cải thiện khả năng dự đoán chất lượng dữ liệu. Những phương pháp này không chỉ giúp nâng cao hiệu quả của mô hình mà còn mở ra khả năng phát hiện các mẫu dữ liệu tiềm ẩn.

Tăng cường bộ dữ liệu: Thu thập thêm dữ liệu từ các nguồn mới và đa dạng hơn để mở rộng phạm vi phân tích. Việc này không chỉ tăng cường chất lượng mô hình mà còn giúp hệ thống áp dụng tốt hơn trong các môi trường giáo dục khác nhau.

Phát triển hệ thống đánh giá tự động: Xây dựng các công cụ đánh giá tự động để phân tích dữ liệu theo thời gian thực và cung cấp báo cáo chi tiết. Hệ thống này có thể tích hợp các chỉ số đã phân tích, như độ trọng tâm hoặc kết quả phân cụm, để đưa ra cảnh báo hoặc đề xuất cải tiến một cách nhanh chóng.

Tích hợp hệ thống với nền tảng giáo dục trực tuyến: Nghiên cứu khả năng tích hợp hệ thống với các nền tảng MOOCs phổ biến như Coursera, edX hoặc các hệ thống quản lý học tập (LMS) nhằm hỗ trợ việc đánh giá và cải thiện chất lượng khóa học một cách toàn diện.

Những hướng phát triển này không chỉ giúp tối ưu hóa hệ thống đánh giá chất lượng khóa học mà còn mang lại các phương pháp phân tích hiện đại, mở ra cơ hội nghiên cứu mới trong lĩnh vực giáo dục trực tuyến. Nhóm hy vọng rằng, thông qua việc triển khai các hướng phát triển này, hệ thống sẽ đóng góp thiết thực vào việc cải thiện chất lượng học tập và trải nghiệm người học trong tương lai.

# THAM KHẢO

[1].<http://moocdata.cn/data/MOOCCube>

[2]. <https://github.com/THU-KEG/MOOCCubeX/tree/main>

[3].<https://docs.google.com/document/d/1tIN42AnArkE6qQhNUpMtzH_H-LsIXqJFpSe7qYRb49E/edit?tab=t.0>