IS353.P12 - Mạng Xã hội

Báo cáo tiến độ

ĐÁNH GIÁ ĐỘ HIỆU QUẢ CỦA KHÓA HỌC

Nhóm 2

Thành viên nhóm 2:

1. Nguyễn Trần Vũ Quang	-	21521347
2. Nguyễn Hoàng Phúc	-	21522473
3. Nguyễn Thị Mai Trinh	-	21522718
4. Lê Công Hoài Nam	-	22520910
5. Lê Thanh Tài	-	22521276
6. Nguyễn Văn Thắng	-	22520005

NỘI DUNG CHÍNH

- GIỚI THIỆU VỀ BÀI TOÁN
- PHƯƠNG PHÁP GÁN NHÃN
- XÂY DỰNG MÔ HÌNH
- KẾT LUẬN

I. GIỚI THIỆU VỀ BÀI TOÁN

1.1. Phát biểu về bài toán

Input: Thông tin khóa học (Bao gồm: course, user, user-problem, problem, video, user-video, comment, course-comment)

Output:

- Kết quả đánh giá khóa học dựa theo 04 tiêu chí (1-5).
- Kết quả đánh giá Độ hiệu quả khóa học (0-2).
- Mô hình dự đoán Độ hiệu quả khóa học

1.2. Mục tiêu đề tài

- Phát triển một mô hình đánh giá độ hiệu quả cho các khóa học trực tuyến dựa trên thông tin cơ bản.
- Đánh giá các khóa học dựa trên nhiều tiêu chí khác nhau.
- Cung cấp đánh giá chi tiết và tổng hợp về độ hiệu quả các khóa học (0-2 sao).

TC 1: Điểm trung bình các khóa học

TC 2: Thói quen học trên video

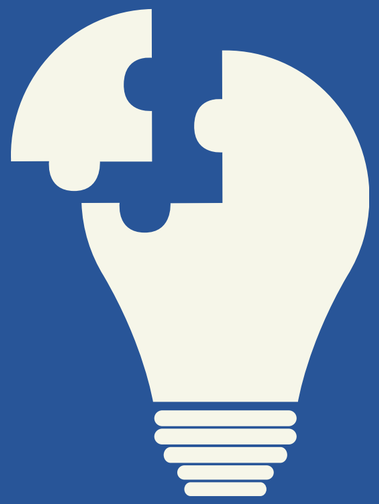
TC 3: Số lần thử và nộp các bài tập

TC 4: Mức độ hài lòng của các học viên

1.3. Ngữ cảnh

- Hiện nay, có nhiều khóa học trực tuyến được cung cấp trên các nền tảng MOOCs với số lượng lớn. Việc đánh giá chất lượng thủ công thông qua các phương pháp truyền thống như khảo sát học viên, phỏng vấn, hay thử nghiệm tốn kém rất nhiều nguồn lực (thời gian, chi phí, nhân sự).
- Vì vậy, việc phát triển một hệ thống tự động để đánh giá các khóa học trên nền tảng trực tuyến sẽ giúp các nhà quản lý dễ dàng theo dõi chất lượng giảng dạy và cải thiện trải nghiệm học viên.

II. PHƯƠNG PHÁP GÁN NHÃN



2.1. Bộ dữ liệu

- Nguồn dữ liệu: Bộ MOOCCubeX, phát triển bởi AMiner, cung cấp dữ liệu từ các khóa học trực tuyến (MOOC) nhằm phân tích hành vi và mô hình học tập, hỗ trợ cá nhân hóa học tập và cải tiến hệ thống khuyến nghị.
- Mô tả: Bộ dữ liệu chứa thông tin chi tiết về khóa học, người dùng, bài tập, video bài giảng, giúp phân tích hành vi, tỷ lệ hoàn thành, và đánh giá hiệu quả học tập. Nhóm sử dụng 8 bảng dữ liệu từ nguồn.

2.1. Bộ dữ liệu

Bộ dữ liệu MOOCCubeX cung cấp thông tin phong phú về khóa học, người học, bài tập và video, hỗ trợ phân tích hành vi học tập, tỷ lệ hoàn thành và đánh giá hiệu quả học tập.

Để xây dựng bài toán, nhóm chúng em sử dụng 8 bảng dữ liệu trong bộ dữ liệu MooccubeX, bao gồm:

STT	Biến	Giải thích	Định dạng	Miền giá trị
1	id	Id của người dùng	String	Bắt đầu với U_
2	name	Tên	String	
3	gender	Giới tính	Byte	2000,1,2
4	school	Trường học	String	
5	year_of_birth	Năm sinh	Integer	
6	course_order	Mảng số nguyên chứa id các khóa học đã đăng kí	Array	
7	enroll_time	Mảng chứa thời gian đăng ký khóa học	Array	Mỗi phần tử là một chuỗi theo định dạng DateTime

B1: Mô tả bảng User

STT	Biến	Giải thích	Định dạng	Miền giá trị
1	log_id	Id của bản ghi	String	
2	problem_id	Id của câu hỏi	String	Bắt đầu bằng “Pm_”
3	user_id	Id của người dùng	String	Bắt đầu bằng “U_”
4	is_correct	- 'true' nếu người dùng đã đưa ra bài làm đúng hoàn toàn - 'false' nếu ngược lại	Boolean	True, False
5	attempts	Số lần người dùng nộp bài tập	Integer	>= 1
6	score	Điểm số của người dùng đạt được khi làm bài tập	Float	>= 0 hoặc null
7	submit_time	Thời điểm người dùng nộp bài	DateTime	

B2: Mô tả bẳng user_problem

STT	Biến	Giải thích	Định dạng	Miền giá trị
1	id	Id khóa học	String	Bắt đầu bằng “C_”
2	name	Tên khóa học	String	
3	prerequisites	Những kiến thức, kĩ năng cần thiết mà người học cần phải nắm vững trước khi đăng kí khóa học mới	String	
4	about	Giới thiệu về khóa học	String	
5	resource	Danh sách tài nguyên của khóa học	Array	
6	field	Mô tả phạm vi, chủ đề hoặc ngành học chính mà khóa học thuộc về	Array	

B3: Mô tả bảng course

STT	Biến	Giải thích	Định dạng	Miền giá trị
1	problem_id	Id vấn đề	String	Bắt đầu bằng “Pm_”
2	title	Tiêu đề của câu hỏi	String	
3	content	Mô tả câu hỏi	String	
4	option	Mảng chứa 4 đáp án	Array	A, B, C, D
5	answer	Câu trả lời đúng	Char	A, B, C, D
6	score	Số điểm của câu hỏi	Int	
7	type	Loại câu hỏi	Int	
8	typetext	Loại câu hỏi	String	
9	location	Vị trí chương của câu hỏi	String	
10	context_id	Mảng chứa id liên quan đến vấn đề, cho biết vấn đề này thuộc về một khoá học, một bài giảng, hoặc một kỳ thi cụ thể nào đó.	Array	
11	exercise_id	Id bài tập	String	Bắt đầu bằng “Ex_”
12	language	Ngôn ngữ mô tả câu hỏi	String	Chinese

B4: Mô tả bảng problem

STT	Biến	Giải thích	Định dạng	Miền giá trị
1	seq	Một mảng các dữ liệu có cấu trúc, mỗi phần tử bao gồm video_id và một mảng segment. Mỗi segment bao gồm: start_point, end_point, speed, local_start_time	Array	
2	user_id	Id của người dùng	String	Bắt đầu bằng “U_”

B5: Mô tả bảng user_video

STT	Biến	Giải thích	Định dạng
1	ccid	Chứa mã phân loại dùng để phân loại video theo một tiêu chí cụ thể nào đó, chẳng hạn như thể loại, nguồn gốc, hoặc trạng thái	String
2	name	Tên của video	String
3	start	Mảng các số chỉ thời điểm bắt đầu của một sự kiện trong video	Array
4	end	Mảng các số chỉ thời điểm kết thúc của một sự kiện trong video	Array
5	text	Mảng chứa nội dung văn bản liên quan đến đoạn video, chẳng hạn như lời thoại, mô tả nội dung, hoặc chú thích cho một cảnh cụ thể	Array

B6: Mô tả bảng video

STT	Biến	Giải thích	Định dạng	Miền giá trị
1	course_id	Id khóa học	String	Bắt đầu bằng “C_”
2	comment_id	Id của bình luận	String	Bắt đầu bằng “Cm_”

B7: Mô tả bảng course_comment

STT	Biến	Giải thích	Định dạng	Miền giá trị
1	id	Id của comment	String	Bắt đầu bằng “Cm_”
2	user_id	Id của người dùng	String	Bắt đầu bằng “U_”
3	text	Nội dung của bình luận	String	
4	create_time	Thời điểm bình luận	DateTime	

B8: Mô tả bảng comment

2.2. Các tiêu chí gán nhãn

Điểm trung bình khóa học

Thói quen học qua video

Nỗ lực làm bài tập

Mức độ hài lòng của học viên

Tiêu chí 1 : Điểm trung bình các bài tập

Bảng	user_problem, problem, exercise_problem
Mô tả	Tiêu chí này nhằm tính toán điểm trung bình đạt được của học viên trên tất cả các bài tập trong khóa học, từ đó đánh giá mức độ khó và hiệu quả của khóa học.

Tiêu chí 1 : Điểm trung bình các bài tập

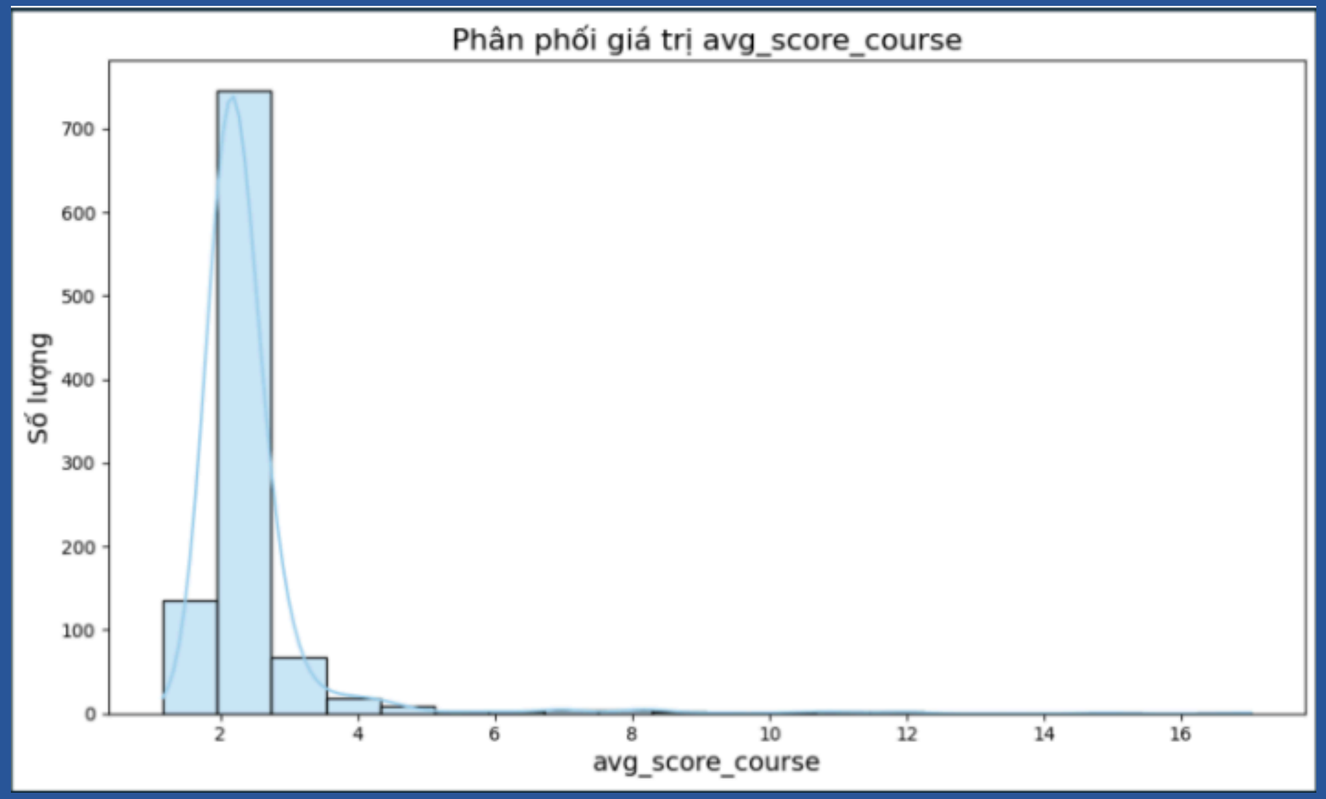
Mục tiêu:

- Phân loại điểm trung bình học viên qua bài tập.
- Đánh giá độ khó và hiệu quả của khóa học qua các trường avg_score_user, avg_score_problem, avg_score_exercise, avg_score_course.

Xử lý dữ liệu:

- Lọc các trường cần thiết: user_id, score, is_correct, attempts.
- Đọc và xử lý 10 file JSON, xử lý lỗi khi đọc file.
- Loại bỏ dữ liệu thiếu user_id hoặc score null.
- Điền giá trị mặc định cho is_correct, attempts, và score nếu thiếu.

Tiêu chí 1 : Điểm trung bình các bài tập



Biểu đồ phân phối giá trị avg score course

- Tính toán avg_score_user:
- Tính toán avg score problem:
- Tính toán avg score exercise:
- Tính toán avg score course:

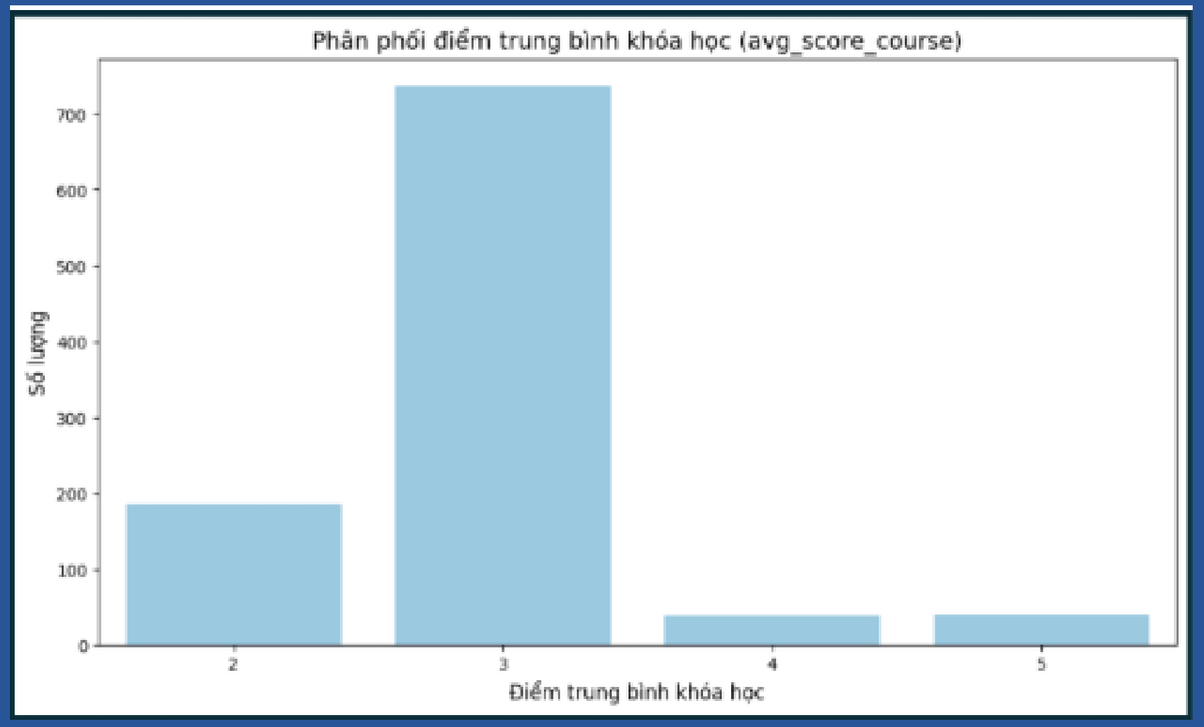
Tiêu chí 1 : Điểm trung bình các bài tập

- Từng bước thực hiện:
  - Tính điểm trung bình cho từng học viên (user).
  - Tính điểm trung bình cho từng bài tập (problem).
  - Tính điểm trung bình cho từng bài thi (exercise).
  - Tính điểm trung bình chung của toàn khóa học (course).

  Lọc bỏ các cột không cần thiết sau khi sử dụng.
- Kết quả: Điểm trung bình của từng bài tập, bài thi, và khóa học, phân phối điểm trung bình khóa học được hiển thị dưới dạng biểu đồ.

Tiêu chí 1 : Điểm trung bình các bài tập

Sau khi hoàn thành việc tính toán, chúng ta sẽ có được điểm trung bình của từng exercise và điểm trung bình của từng course.



Biểu đồ phân phối điểm trung bình khóa học(avg_score_course)

Tiêu chí 2 : Thói quen học qua video

Bảng	user-video.json, course-video.csv
Mô tả	Tính điểm trung bình đạt được của học viên trên tất cả các bài tập để đánh giá mức độ khó và hiệu quả.

Tiêu chí 2 : Thói quen học qua video

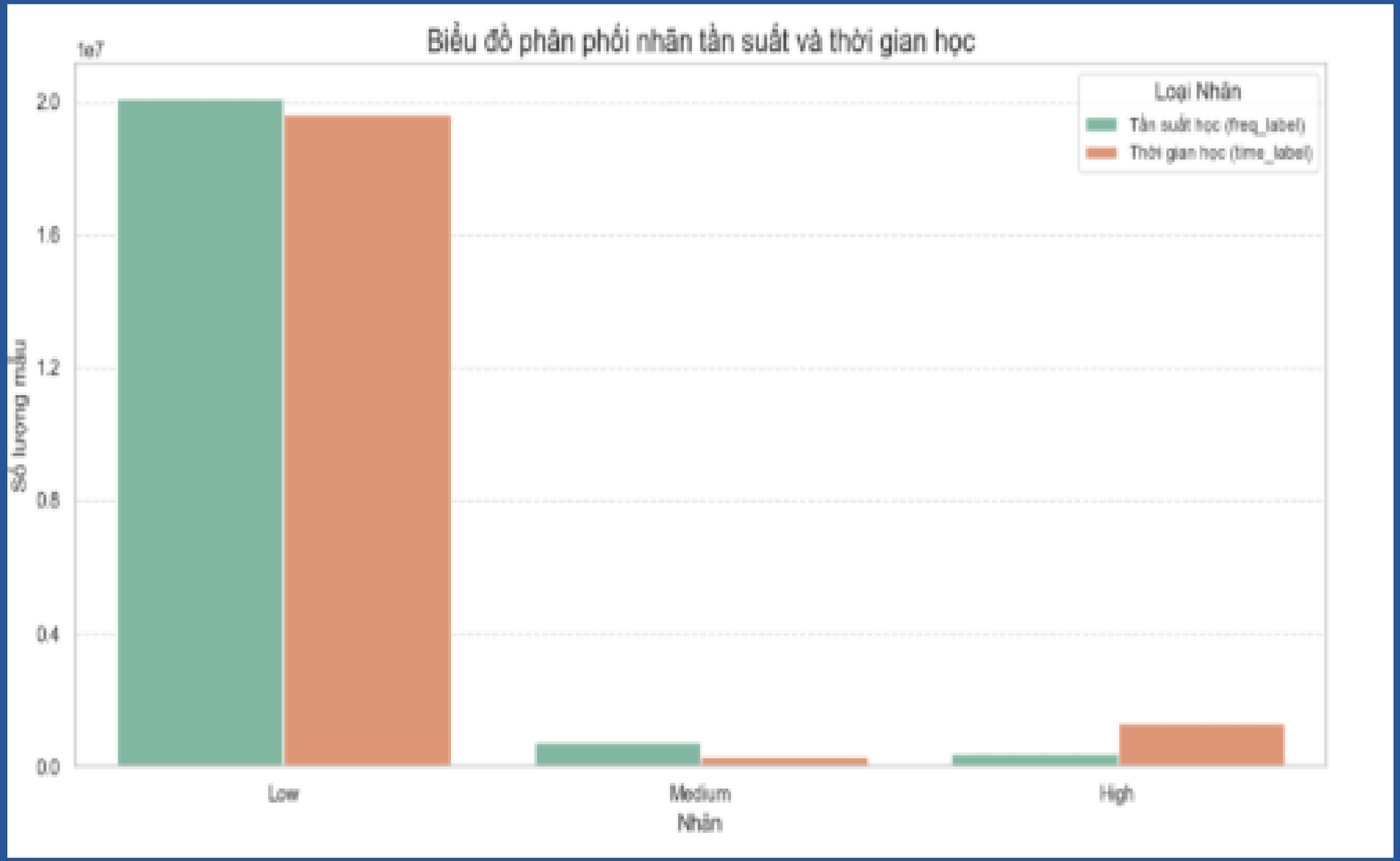
Mục tiêu:

- Phân loại video học tập theo tần suất học và thời gian học.
- Hiểu hành vi người dùng và phân tích cách họ tương tác với video.

Quy trình thực hiện:

- Tính thời gian học và số lần học cho từng video, lưu vào các cột total_study_time và frequency.
- Gán nhãn:

- Thời gian học (time_label): Low (< 30 phút), Medium (30-60 phút), High (> 60 phút).
- Tần suất học (freq_label): Low (< 2 lần), Medium (2-5 lần), High (> 5 lần).


Biểu đồ phân phối nhãn tần suất và thời gian học
1e7
2.0
1.6
1.2
0.8
0.4
0.0
Số lượng mẫu
Loại Nhãn
Tần suất học (freq_label)
Thời gian học (time_label)
Low
Medium
High
Nhãn

Tiêu chí 2 : Thói quen học qua video

- Kết quả:

  Dữ liệu đã được gán nhãn với hai tiêu chí time_label và freq_label.

  Phân tích: Biểu đồ phân phối cho thấy phần lớn các mẫu thuộc nhãn Low ở cả hai tiêu chí, cho thấy sự tương quan giữa tần suất và thời gian học.

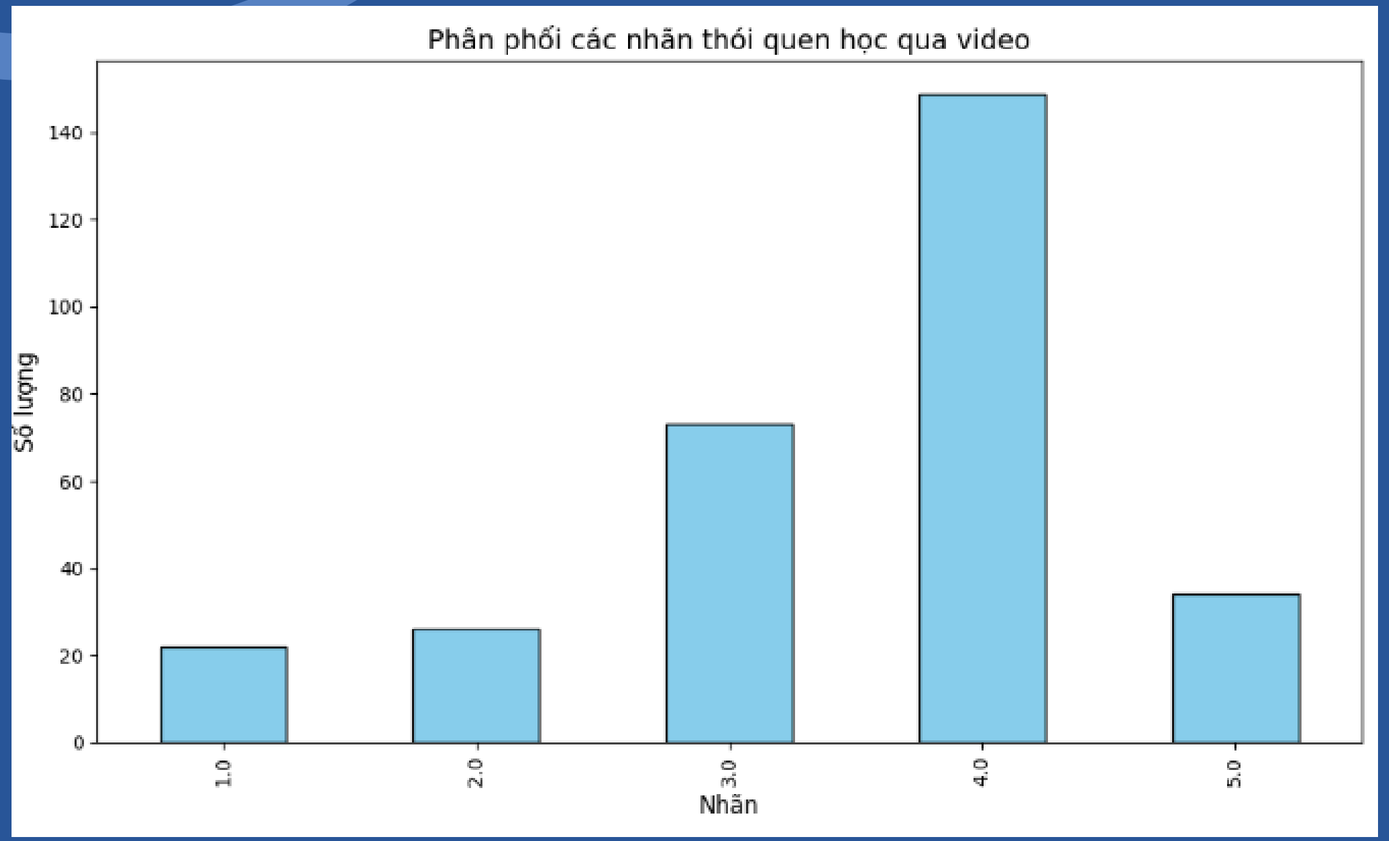
- Kết luận:

  Người học có tần suất thấp thường dành ít thời gian học.

  Phân phối nhãn cung cấp góc nhìn toàn diện về hành vi học tập và giúp tối ưu hóa nội dung video cho các nhóm người dùng khác nhau.

Tiêu chí 2 : Thói quen học qua video

Gán một nhãn mới chỉ thói quen học qua video



Tiêu chí 3 : Số lần thử và nộp bài tập

Bảng	user_problem, course_user
Mô tả	Đo lường mức độ nỗ lực của học viên thông qua số lần thử và nộp bài tập, chỉ ra tính thử thách của khóa học.

Tiêu chí 3 : Số lần thử và nộp bài tập

- Mục tiêu:

  Đánh giá mức độ thử thách của khóa học dựa trên số lần thử (attempts) và nộp bài (submit_time) của học viên.

  Hiểu hành vi học tập và cải thiện thiết kế bài tập.

- Xử lý dữ liệu:

  Dữ liệu gồm: attempts, submit_time, score, is_correct.

  Tính toán số lần thử trung bình (average_attempts) để đánh giá độ khó của bài tập.

  Số lần thử cao phản ánh bài tập khó, có thể yêu cầu điều chỉnh nội dung hoặc hướng dẫn.

log_id	problem_i	user_id	is_correct	attempts	score	submit_time	challenge_index
3.2E+14	Pm_60619	U_320205	0	1		10/8/2020 21:03	2
3.2E+14	Pm_60619	U_320205	1	1		10/8/2020 20:59	2
3.2E+14	Pm_60619	U_320205	1	1		10/8/2020 21:02	2
3.2E+14	Pm_60619	U_320205	1	1		10/8/2020 21:02	2
3.2E+14	Pm_60619	U_320205	0	1		11/27/2020 20:54	2
3.2E+14	Pm_60620	U_320205	1	1		11/27/2020 21:13	2
3.2E+14	Pm_60620	U_320205	1	1		11/27/2020 21:13	2
3.2E+14	Pm_60620	U_320205	1	1		11/27/2020 21:13	2
3.2E+14	Pm_60620	U_320205	1	1		11/27/2020 21:14	2
3.2E+14	Pm_60620	U_320205	1	1		11/27/2020 21:14	2
3.2E+14	Pm_60620	U_320205	1	1		11/27/2020 21:14	2
3.2E+14	Pm_60620	U_320205	1	1		11/27/2020 21:14	2
3.2E+14	Pm_60620	U_320205	1	1		11/27/2020 21:11	2

- Xây dựng chỉ số thử thách:

Challenge Index: Tỉ lệ số lần thử và nộp bài so với mức trung bình của khóa học.

- Gán nhãn độ thử thách (1-5) dựa trên giá trị của Challenge Index:
  - 1 (Rất dễ): Challenge Index ≤ 0.5 × average_attempts.
  - 2 (Dễ): 0.5 × average_attempts < Challenge Index ≤ average_attempts.
  - 3 (Trung bình): average_attempts < Challenge Index ≤ 1.5 × average_attempts.
  - 4 (Khó): 1.5 × average_attempts < Challenge Index ≤ 2 × average_attempts.
  - 5 (Rất khó): Challenge Index > 2 × average_attempts.

Tiêu chí 3 : Số lần thử và nộp bài tập

- Kết quả:

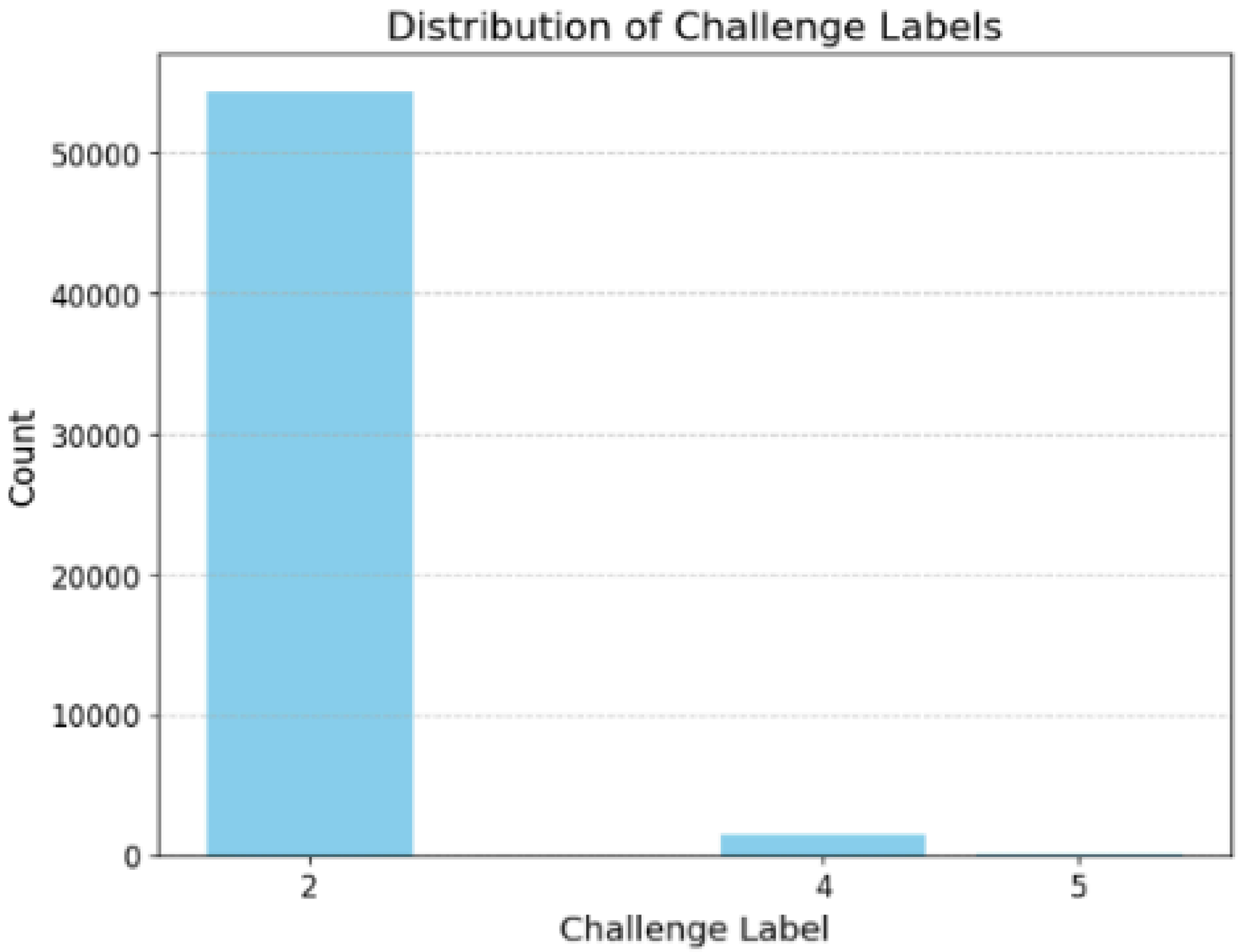
  Gán nhãn thử thách cho từng bài tập và tạo sơ đồ phân phối nhãn.

  Biểu đồ cho thấy phần lớn bài tập được gán nhãn "2 (Dễ)", phản ánh độ thử thách vừa phải.

- Nhận xét:

  Phân phối nhãn chủ yếu tập trung ở mức dễ và trung bình, cho thấy khóa học được thiết kế phù hợp với trình độ học viên.

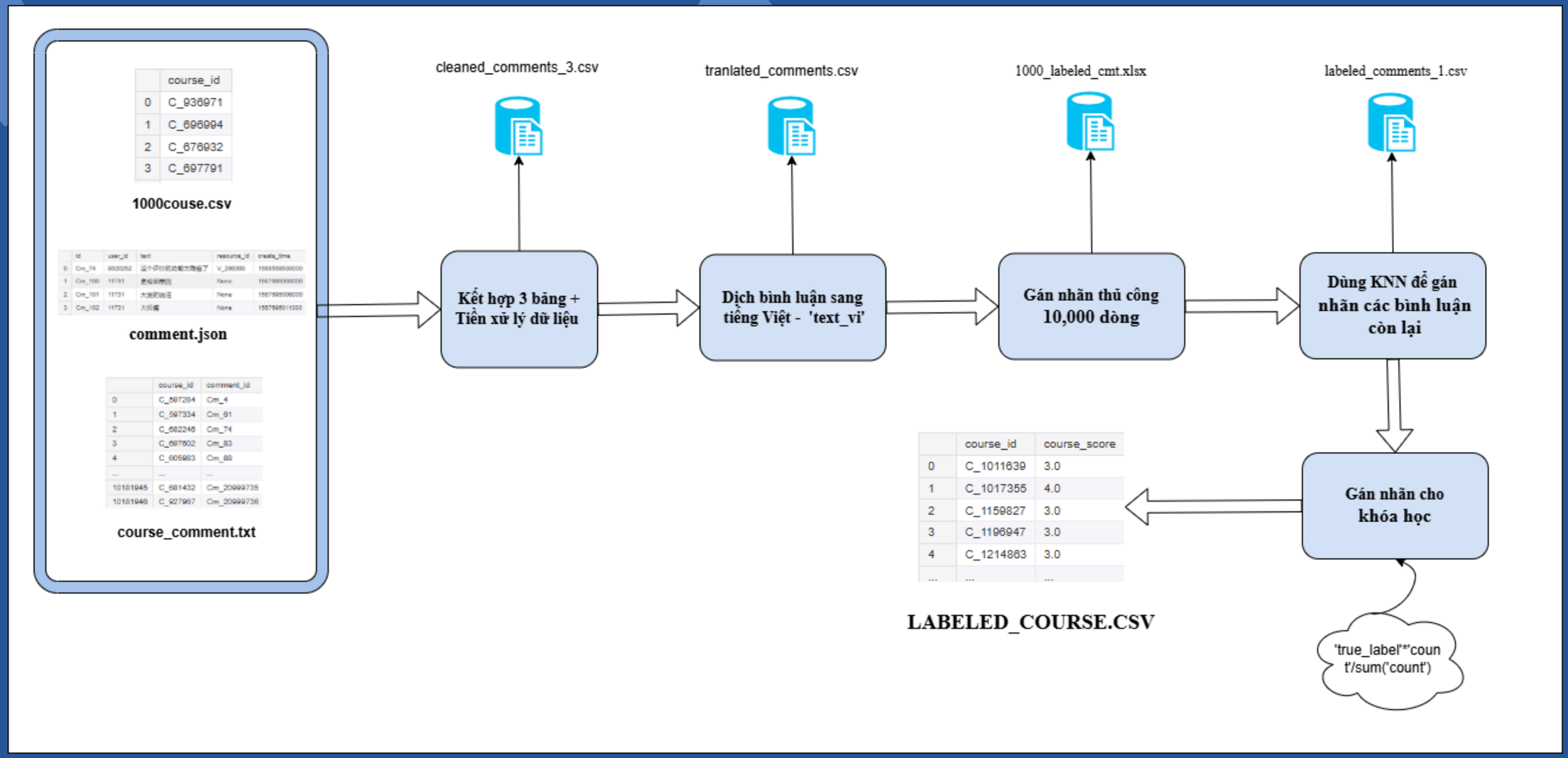
  Các bài tập có Challenge Index cao có thể cần điều chỉnh để giảm độ khó hoặc tăng tính rõ ràng trong hướng dẫn.


Distribution of Challenge Labels
50000
40000
30000
20000
10000
0
Count
2
4
5
Challenge Label

Tiêu chí 4 : Mức độ hài lòng của học viên

Bảng	course.json, course-comments.txt, comments.json
Mô tả	Phân tích cảm xúc và nội dung bình luận để đo lường mức độ hài lòng và giá trị của khóa học từ góc nhìn học viên

Tiêu chí 4 : Mức độ hài lòng của học viên



Tiêu chí 4 : Mức độ hài lòng của học viên

Tiền xử lý dữ liệu:

- Lấy thông tin của 1000 khóa học phổ biến nhất từ bộ dữ liệu.
- Kết hợp 3 bảng dữ liệu, loại bỏ thông tin thừa.
- Tiền xử lý văn bản bình luận (loại bỏ dấu câu, ký tự đặc biệt, khoảng trắng thừa).
- Giảm số lượng bình luận (tối đa 300 bình luận mỗi khóa học).

Dịch bình luận:

- Dịch các bình luận sang tiếng Việt bằng công cụ GoogleTranslator().
- Tiến hành tiền xử lý thêm cho dữ liệu Tiếng Việt.

Tiêu chí 4 : Mức độ hài lòng của học viên

Bộ dữ liệu cuối cùng có 144,787 dòng dữ liệu, được xử dụng để gán nhãn cảm xúc cho bình luận.

	course_id	text	filtered_text	count	text_vi
0	C_948419	期待更新	期待更新	1.0	Mong nhận được thông tin cập nhật
1	C_948419	老师快快更自我隔离期间特别需要您	老师快快自我隔离期间特别需要	1.0	Thưa thầy, xin vui lòng cập nhật. Chúng tôi cả...
2	C_948419	涨知识谢谢老师	涨知识谢谢老师	1.0	Kiến thức được nâng cao, cảm ơn thầy
3	C_948419	为什么后期不再使用异化这一概念与之对应的概念是什么	后期不再使用异化这一概念对应概念	1.0	Tại sao khái niệm tha hóa không còn được sử dụ...
4	C_948419	特别有感染力想一直听下去	特别感染力想一直听下去	1.0	Nó dễ lây lan đến mức tôi muốn tiếp tục nghe nó.

Dữ liệu sau khi được làm sạch và dịch thuật

Tiêu chí 4 : Mức độ hài lòng của học viên

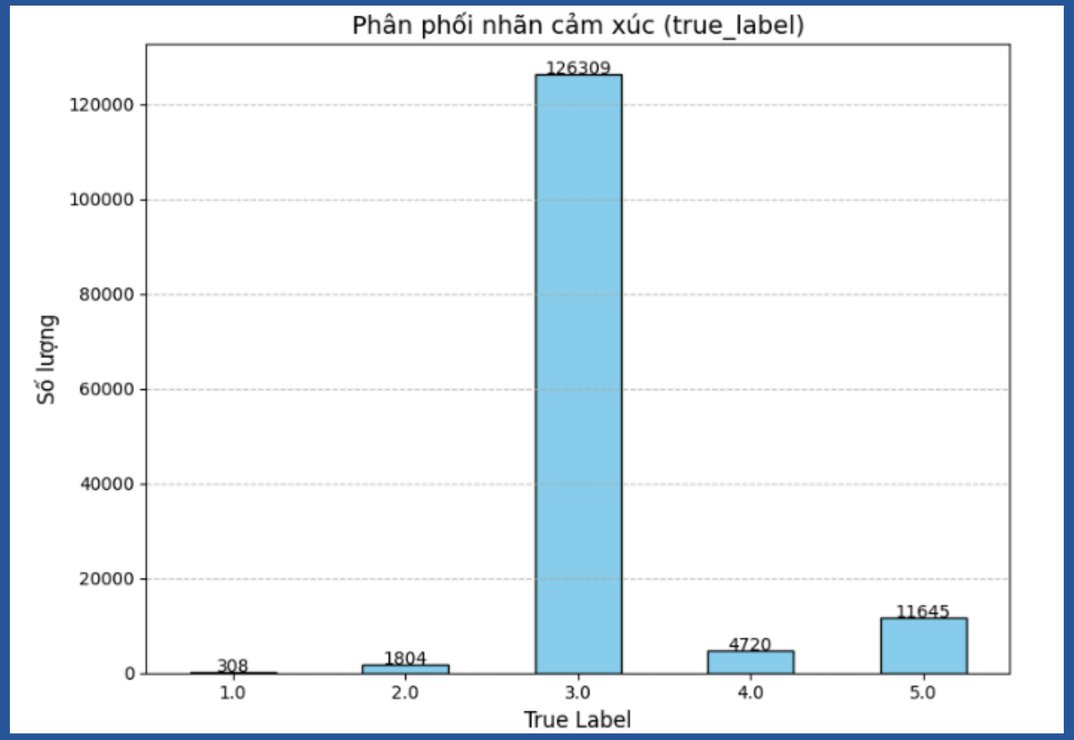
Quy tắc gán nhãn:

- 1: Cảm xúc tiêu cực, ví dụ: 'Thầy này cứ mắc lỗi hoài”
- 2: Cảm xúc không tích cực, ví dụ: “bị tắc”, “không thể vào”,...
- 3: Trung tính, không bày tỏ cảm xúc. Ví dụ: 'Đăng nhập', 'Hãy bắt đầu.'
- 4: Cảm xúc tích cực, ví dụ: “Mong được bắt đầu”
- 5: Cảm xúc rất tích cực, ví dụ: 'Nói rất hay', 'Dễ hiểu'.

Tiêu chí 4 : Mức độ hài lòng của học viên

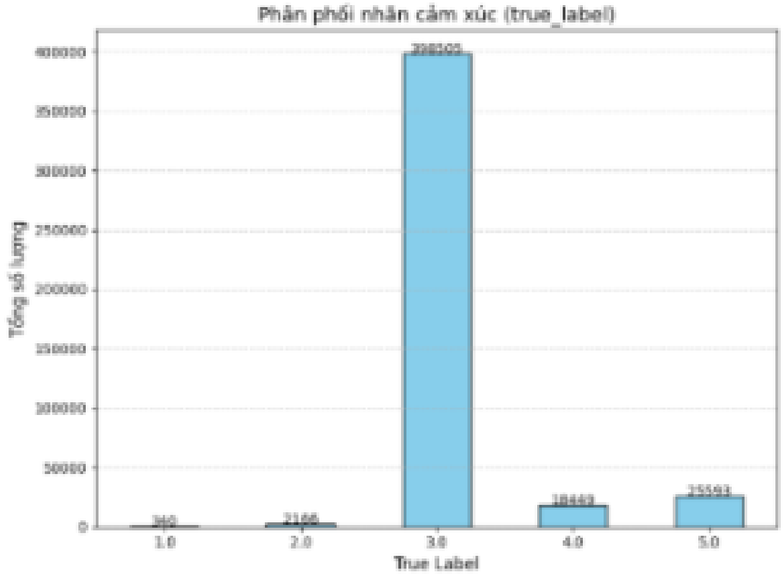
Sau đó, ta thu được bộ dữ liệu đã gán nhãn:

- Dữ liệu ban đầu phân bố chủ yếu ở nhãn 3 (trung tính), gây mất cân bằng.
- Giảm giá trị ‘count’ của nhãn 3 về 1 để giảm ảnh hưởng và cân bằng phân phối dữ liệu.

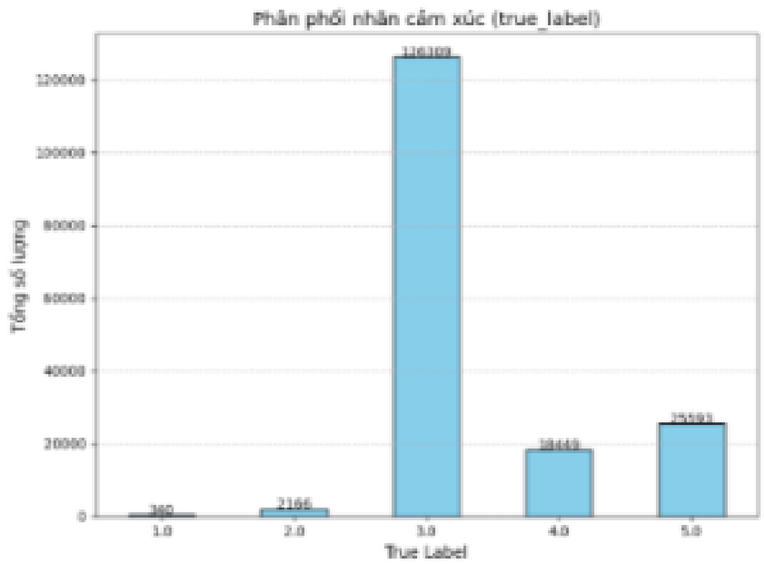


Tiêu chí 4 : Mức độ hài lòng của học viên

Cải thiện 1 phần tình trạng mất cân bằng dữ liệu



Hình 3.1.4.a.2: Phân phối nhãn theo (1)



Hình 3.1.4.a.3: Phân phối nhãn theo (1) đã giảm count của nhãn 3

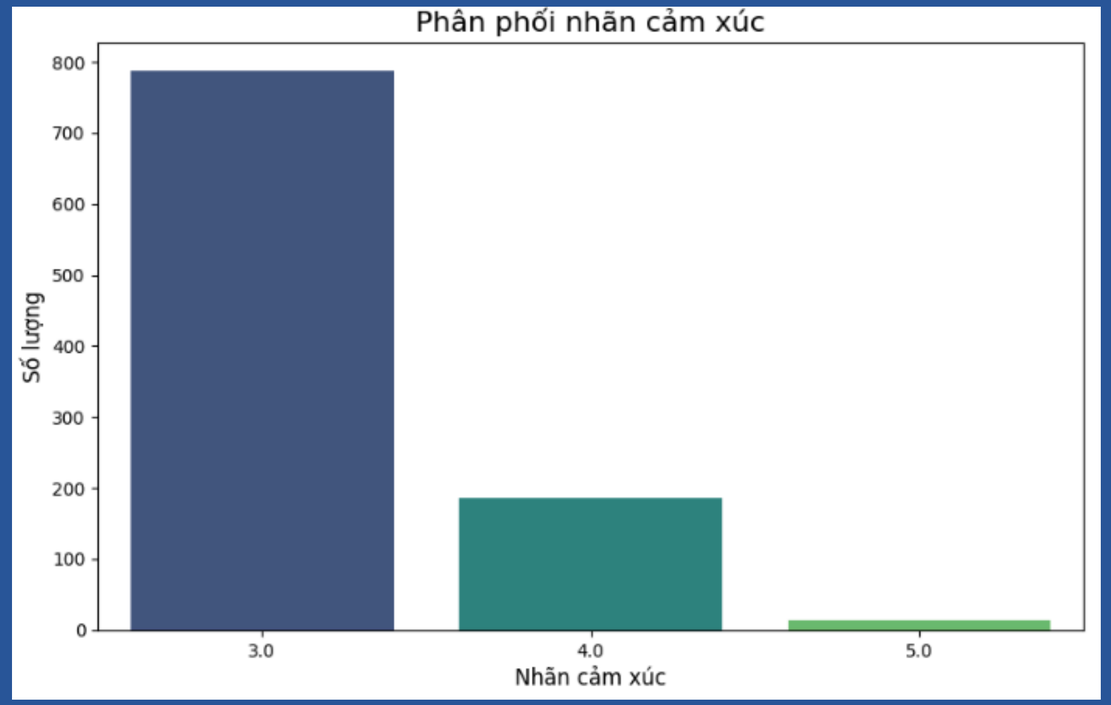
Tiêu chí 4 : Mức độ hài lòng của học viên

- Gán nhãn cảm xúc cho khóa học:

  ‘sentiment_index’ = (‘true_label’*’count’)/sum(‘count’)

Kết quả:

- Bộ dữ liệu đã gán nhãn gồm 986 khóa học, phân phối cảm xúc chủ yếu ở mức 3 sao, phù hợp với thực tế rằng phần lớn bình luận mang tính góp ý chuyên môn.

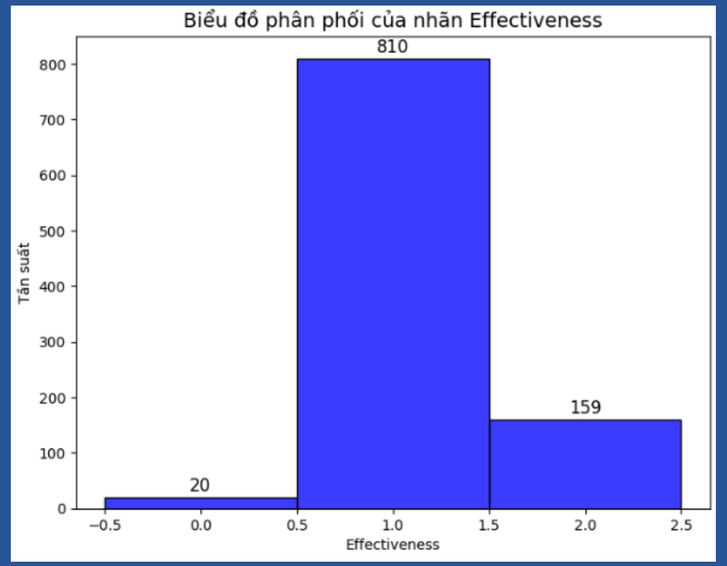


Tổng hợp các tiêu chí

Nhóm gán nhãn độ hiệu quả cho 1,000 khóa học dựa trên 4 tiêu chí, với trọng tâm là "Mức độ hài lòng của học viên".

Bộ dữ liệu thu được cho thấy sự mất cân bằng giữa các nhãn, trong đó 980 khóa học có hiệu quả từ trung bình đến cao, còn 20 khóa học có hiệu quả thấp.

1. avg_score	2. video_habit	3. challenge_index	4. sentiment_index	effectiveness
3	0	2	4	2
3	0	2	4	2
2	0	2	4	1
3	0	4	3	1
3	4	2	3	1



IV. Xây dựng mô hình

Độ hiệu quả của khóa học

- Tổng quan bài toán
- Mô hình và thuật toán
- Giải pháp xử lý mất cân bằng dữ liệu
- Độ đo đánh giá
- Kết quả thực nghiệm

Tổng quan bài toán

Mục tiêu là xây dựng mô hình dự đoán chất lượng khóa học từ thông tin cơ bản (tên, giảng viên, trường, mô tả), giúp giảng viên và quản lý điều chỉnh kịp thời. Sử dụng 4 mô hình máy học (Logistic Regression, Random Forest, SVM, XGBoost) kết hợp SMOTE để cải thiện hiệu suất dự đoán.

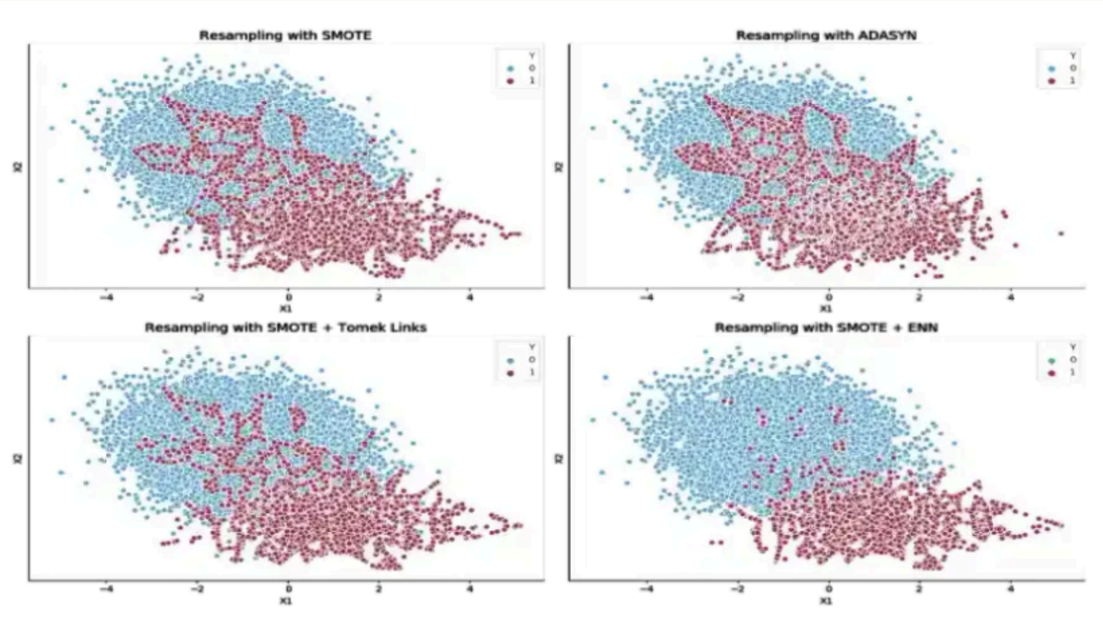
Bài toán

- Input: Thông tin cơ bản của khóa học, cụ thể ở đây là tổ hợp ‘name’ + ‘about’
- Output: Độ hiệu quả của khóa học

Lựa chọn mô hình

- Logistic Regression
- Random Forest
- SVM (Support Vector Machine)
- XGBoost (Extreme Gradient Boosting)

Giải pháp xử lý mất cân bằng dữ liệu



Kỹ thuật SMOTE (Synthetic Minority Oversampling Technique)

- SMOTE là kỹ thuật oversampling giúp cân bằng dữ liệu bằng cách tạo ra các mẫu tổng hợp cho lớp thiểu số. SMOTE cải thiện hiệu suất mô hình mà không tạo dữ liệu lặp lại và dễ dàng tích hợp vào pipeline xử lý dữ liệu.
- Khi kết hợp với các mô hình máy học giúp cải thiện hiệu suất trên lớp thiểu số, mặc dù hiệu quả phụ thuộc vào từng thuật toán và chất lượng dữ liệu.

Độ đo đánh giá

Accuracy: $$Accuracy = \frac{TP+TN}{TN+TP+FN+FP}$$

Precision: $$Precision = \frac{TP}{TP+FP}$$

Recall: $$Recall = \frac{TP}{TP+FN}$$

Weighted F1: $$WeightedF1 = \frac{F1_1 \times w_1 + F1_2 \times w_2 + \ldots + F1_n \times w_n}{w_1 + w_2 + \ldots + w_n}$$

Kết quả ✔ thực nghiệm

Mô hình	Accuracy	Weighted precision avg	Weighted recall avg	Weighted f1-score avg
Logistic Regression	0.8332	0.83	0.82	0.75
Random Forest	0.7980	0.70	0.80	0.73
SVM	0.8131	0.73	0.81	0.74
XGBoost	0.8182	0,76	0,82	0.76
LogisticRegression…class_weight='balanced'...) + SMOTE	0.7626	0.70	0.76	0.72
RandomForestClassifier(class_weight='balanced'...) + SMOTE	0.7879	0.70	0.79	0.73
SVC(class_weight='balanced'...) + SMOTE	0.7626	0.69	0.76	0.72
XGBClassifier(scale_pos_weight=1…) + SMOTE	0.7828	0.71	0.78	0.74

KHÓ KHĂN & THÁCH THỨC

- Bộ dữ liệu MoocCubeX quá lớn cần tiền xử lý, giảm kích thước dữ liệu mà đảm bảo chất lượng dữ liệu
- Tự xây dựng nhãn dữ liệu
- Khối lượng dữ liệu không đồng nhất
- Cần dịch về ngôn ngữ thông thạo hơn (Tiếng Việt)
- Dữ liệu mất cân bằng

KẾT LUẬN

Chúng tôi đã đánh giá hiệu quả khóa học trên bộ dữ liệu MoocCubeX dựa trên các tiêu chí với 1,000 khóa học được gán nhãn.

Mô hình dự đoán SVM kết hợp SMOTE cho kết quả tốt nhất, đạt Accuracy 78.28% và Weighted F1-score 0.74, nhưng vẫn gặp khó khăn với nhãn 0 và 2 do dữ liệu mất cân bằng.

Bài toán có tiềm năng phát triển hệ thống tự động tính độ hiệu quả của khóa học, hỗ trợ quản lý giáo dục hiệu quả hơn.

Cảm ơn cô và các ban đã chú ý lắng nghe!