FACULTY OF INFORMATION TECHNOLOGY

Phân tích dữ liệu video trên YouTube

- Giảng viên hướng dẫn: ThS. Nguyễn Thị Hoài Linh
- Sinh viên thực hiện:

1. Nguyễn Tú Như
2. Nguyễn Thúy Vy
3. Nguyễn Thị Chúc Ngọc

Khái quát

- **Mục tiêu**:
  - Tìm ra đặc trưng ảnh hưởng sự thịnh hành.
  - So sánh hiệu quả áp dụng các mô hình máy học đối với video YouTube.
- **Phạm vi**: 1000 video từ US, CA, MX, BR, AR x 500 bình luận/ video.
- **Phương pháp luận**:
  - Nghiên cứu định lượng.
  - Nghiên cứu định tính.

Xử lí dữ liệu

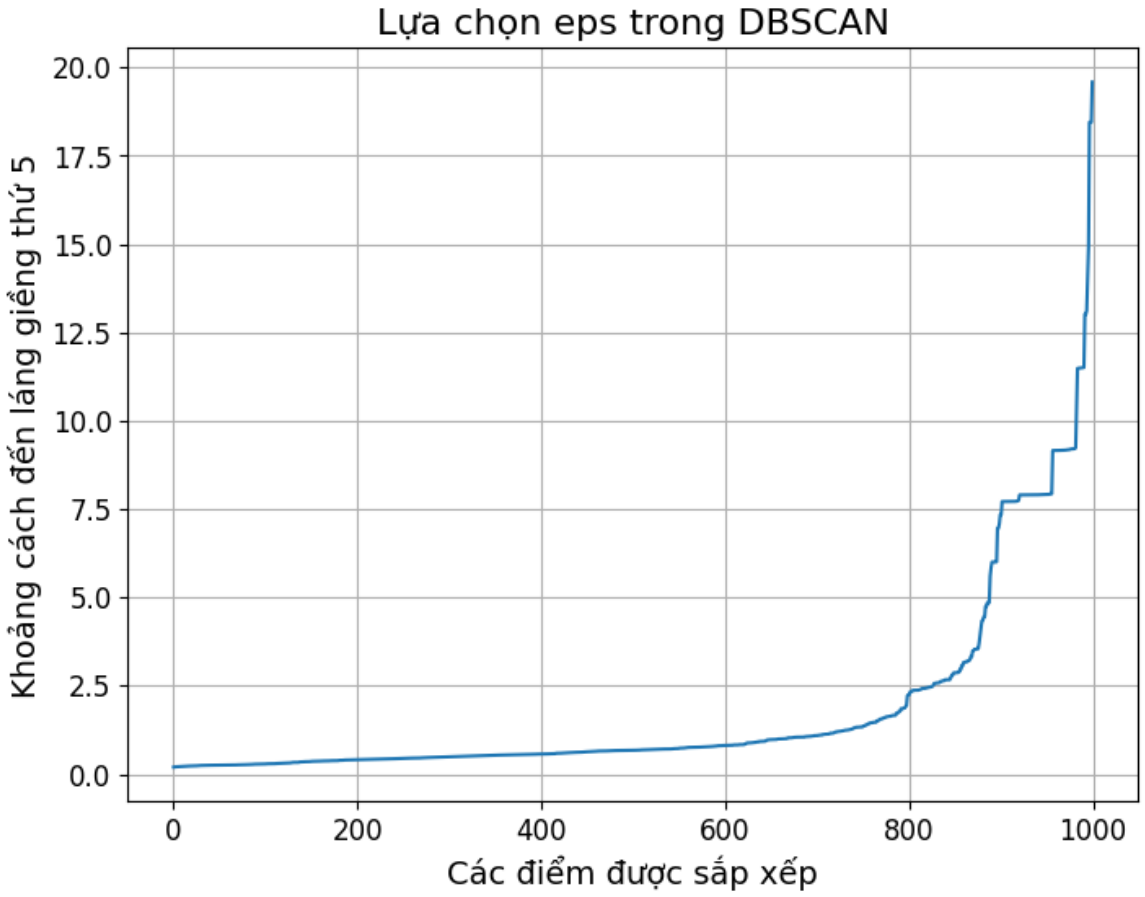
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	video_id	title	description	published_at	channel_id	channel_titl	category_i	tags	duration	definition	caption	view_cour	like_coun	comment_count	
2	w6vCF0Dk	OUR FIRST	It's the...	2025-04-14T09:0(	UCooVYzDxd	Brawl Stars	20	['brawl sta	PT1M46S	hd	TRUE	7615076	262668	17614	
3	sR7rMP4G	KIOSK	Why does	2025-04-14T23:3(	UC7_YxT-KID	Markiplier	20	['markiplie	PT50M32S	hd	FALSE	717425	57123	3014	
4	J75GuCvhI	1,000,000	GROX	2025-04-13T18:0(	UCK5vUVoJs	Grox	20	['grox', 'gro	PT30M11S	hd	FALSE	6746646	312363	48126	
5	FOixb26tjv	6 GAYS VS	Follow if you	2025-04-15T00:01	UCt_DaLB_N	LARRAY	22	['LARRAY']	PT31M11S	hd	FALSE	320794	29934	2148	
6	IlpxO4KRV	Eddington	SUBSCRIBE:	2025-04-14T13:01	UCuPivVjnfN	A24	1	['a24', 'a24	PT1M4S	hd	TRUE	487490	15338	1192	
7	0hm108LC	LIVE: Blue	Watch live as	2025-04-14T14:3(	UC52X5wxOI	Associated P	25	['associate	PT2H14M4	hd	FALSE	410204	3750	1447	
8	MBQBY9b(	I Tested Ev	I tested tech	2025-04-14T11:1(	UCMiJRAwDI	Mrwhosethet	28	['youtuber'	PT33M30S	hd	FALSE	2212611	91826	5219	
9	hx-0OtBsw	Rory McIlr(	2025 Masters	2025-04-14T01:0(	UCja8sZ2T4y	CBS Sports	17	['CBS', 'CB	PT26M34S	sd	FALSE	783900	8076	1426	
10	JwtDtF3DI	Trapped in	Can Salish &	2025-04-12T14:0(	UCKaCalz5N	Jordan Matte	24	['salish ma	PT32M30S	hd	FALSE	11216715	158796	19583	
11	hqGjzCatL	Rory McIlr(	Every single	2025-04-14T02:2(	UCSZxaHGo1	The Masters	17	['Masters',	PT16M27S	hd	FALSE	1042174	10617	689	
12	T4l0KBAn6	Here, Tom(	Around every	2025-04-14T14:5(	UC2t5bjwHd	League of Leg	20	['league', 'l	PT3M18S	hd	TRUE	1838234	71193	3690	
13	IGB9uuetr:	McIlroy's M	Rory McIlroy	2025-04-14T00:0(	UCSZxaHGo1	The Masters	17	['Masters',	PT6M45S	hd	FALSE	1003301	11095	1213	
14	E9tUpO0B	Stadium G	Forge your	2025-04-14T16:0(	UClOf1XXinv	PlayOverwat(	20	['Overwatc	PT2M58S	hd	FALSE	406884	24852	2590	
15	I1o0UHvh(	Gayle King	Watch Lauren	2025-04-14T14:3(	UC-SJ6nODI	CBS Morning	25	['Blue Orig	PT4M52S	hd	TRUE	188966	1640	1405	

- **Nguồn**: API Youtube V3
- ID video • Tiêu đề • Mô tả • ID kênh • Tên kênh
- Thời gian phát hành Thẻ • ID thể loại • Thời lượng
- Độ phân giải • Phụ đề
- Lượt xem • Lượt thích • Lượt bình luận

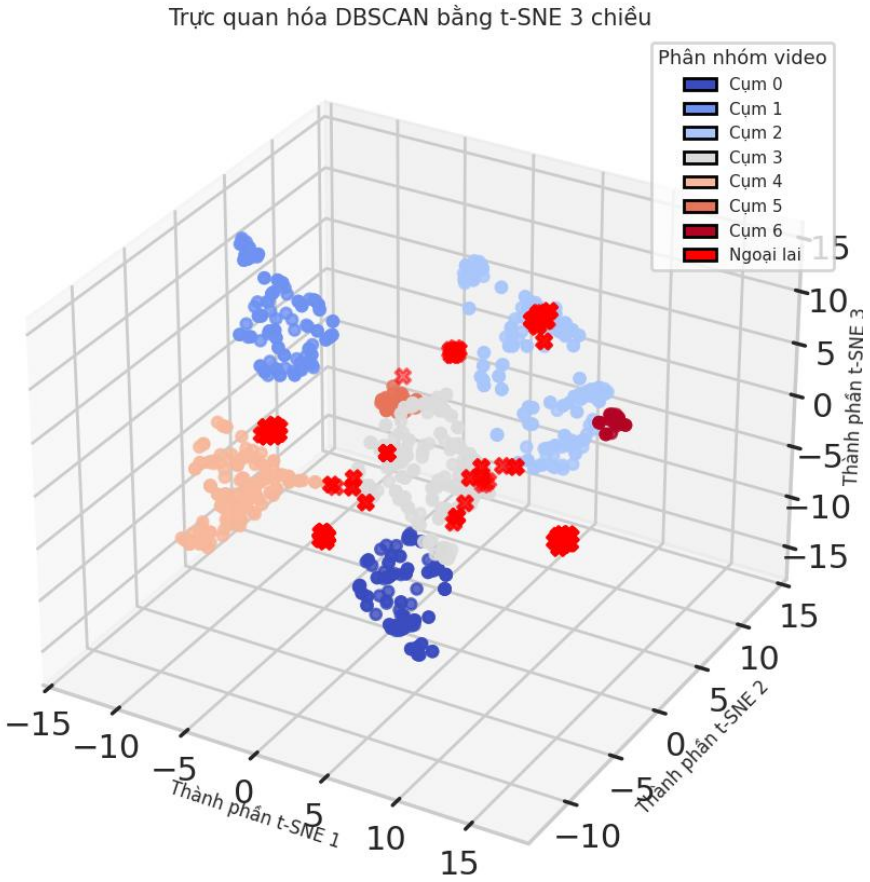
Xử lí dữ liệu

- Dữ liệu thiếu ở description, không có trùng lặp
- **Tạo trường dữ liệu mới**:
  - Giá trị cảm xúc bình luận
  - Tạo cột Tên thể loại
  - Encode ID thể loại, Độ phân giải, Phụ đề
  - Tính thời gian: lúc thịnh hành – lúc phát hành
- **Chuẩn hóa**:
  - Chuyển kiểu dữ liệu
  - Thời lượng: chuỗi (ví dụ: PT15M48S) sang số thực (phút)

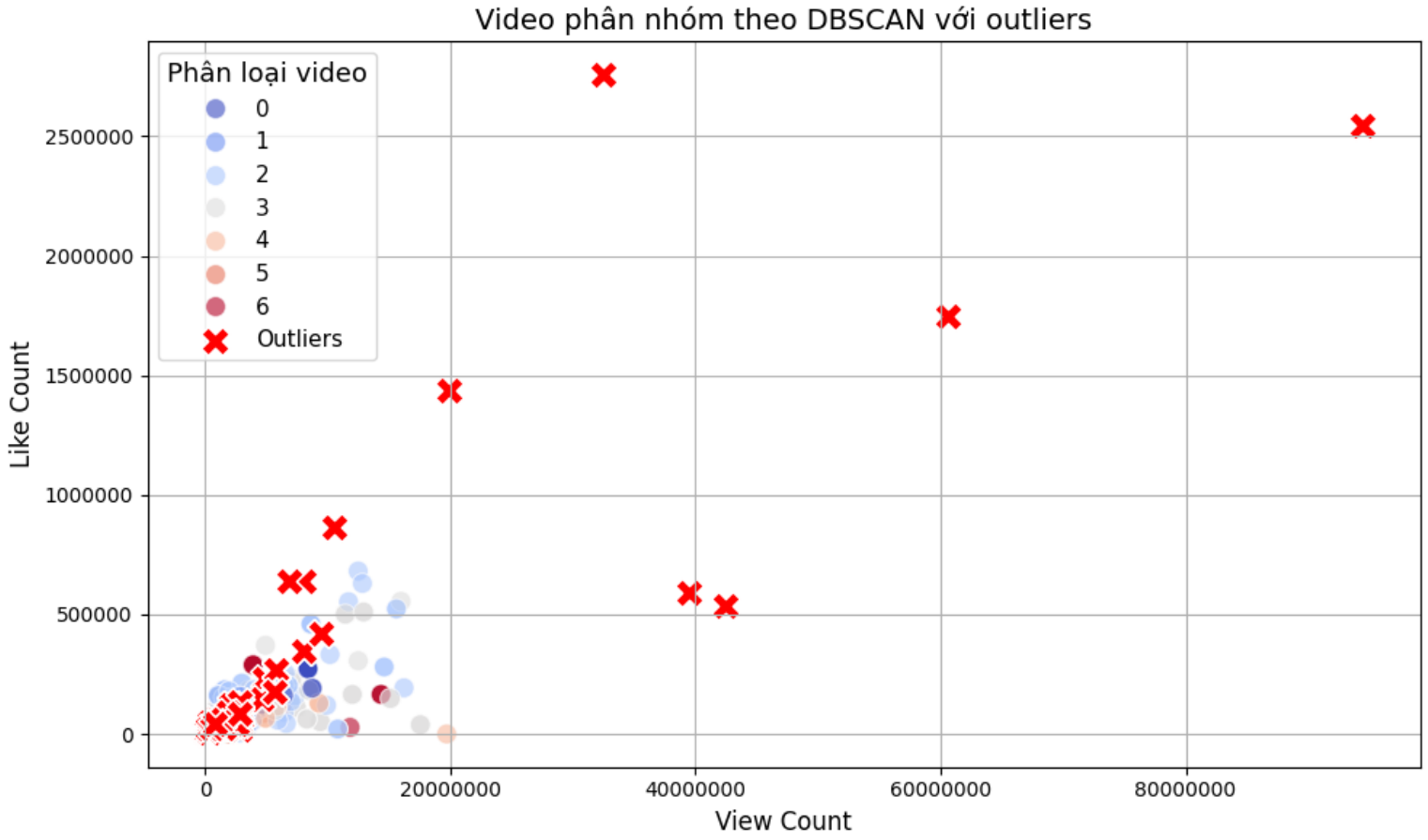
Xử lí dữ liệu


Lựa chọn eps trong DBSCAN
Khoảng cách đến láng giềng thứ 5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
200
400
600
800
1000
Các điểm được sắp xếp

Xử lí dữ liệu


Trực quan hóa DBSCAN bằng t-SNE 3 chiều
Phân nhóm video
Cụm 0
Cụm 1
Cụm 2
Cụm 3
Cụm 4
Cụm 5
Cụm 6
Ngoại lai
Thành phần t-SNE 1
Thành phần t-SNE 2
Thành phần t-SNE 3

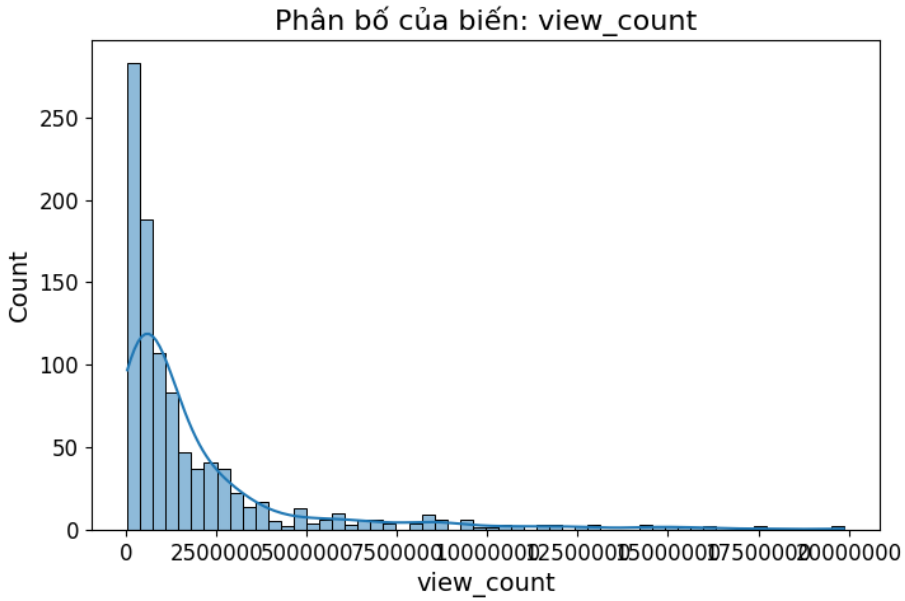
Xử lí dữ liệu

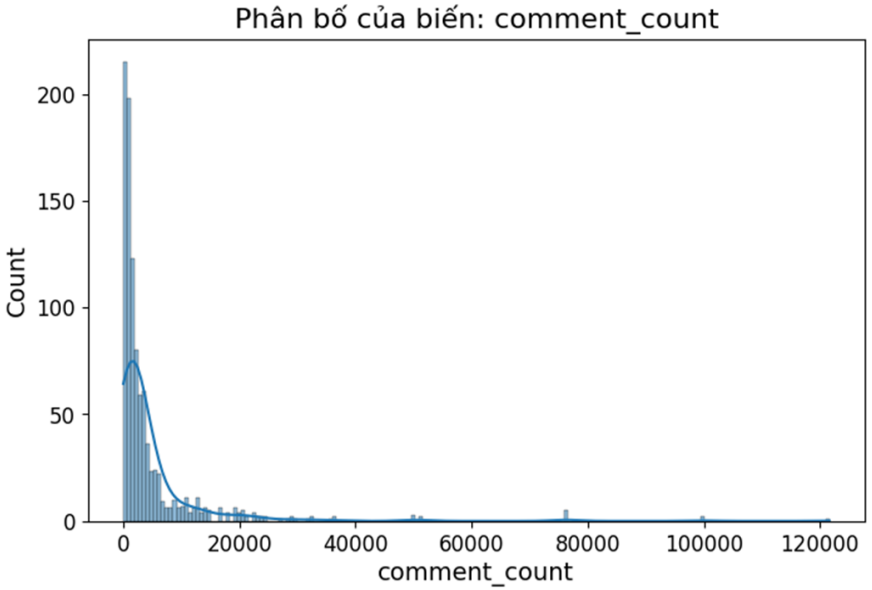

Video phân nhóm theo DBSCAN với outliers
Phân loại video
0
1
2
3
4
5
6
Outliers
Like Count
2500000
2000000
1500000
1000000
500000
0
0
20000000
40000000
60000000
80000000
View Count

Xử lí dữ liệu

- **Phương pháp**: Dùng thuật toán DBSCAN với eps=2.5, min_samples=20.
- **Kết quả**:
  - 127 video (Cụm -1) xác định là nhiễu.
  - Các cụm còn lại (993 video) có đặc điểm riêng về lượt xem, thích, bình luận, thời lượng và thể loại (ví dụ: Trò chơi, Âm nhạc...).
- **Đặc điểm nhiễu**: Lượt xem/thích/bình luận rất cao, thời lượng ngắn, phát hành lâu.

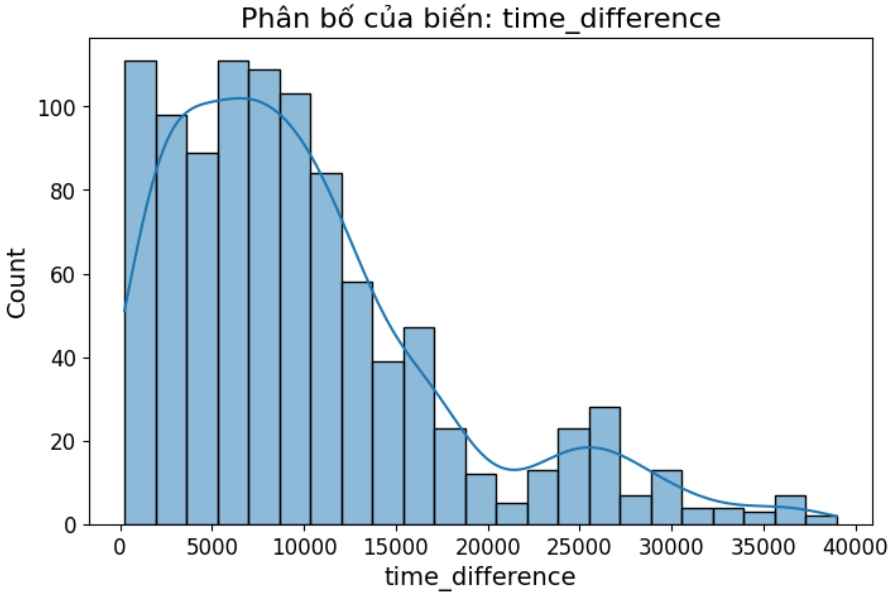
Phương pháp – EDA đơn biến

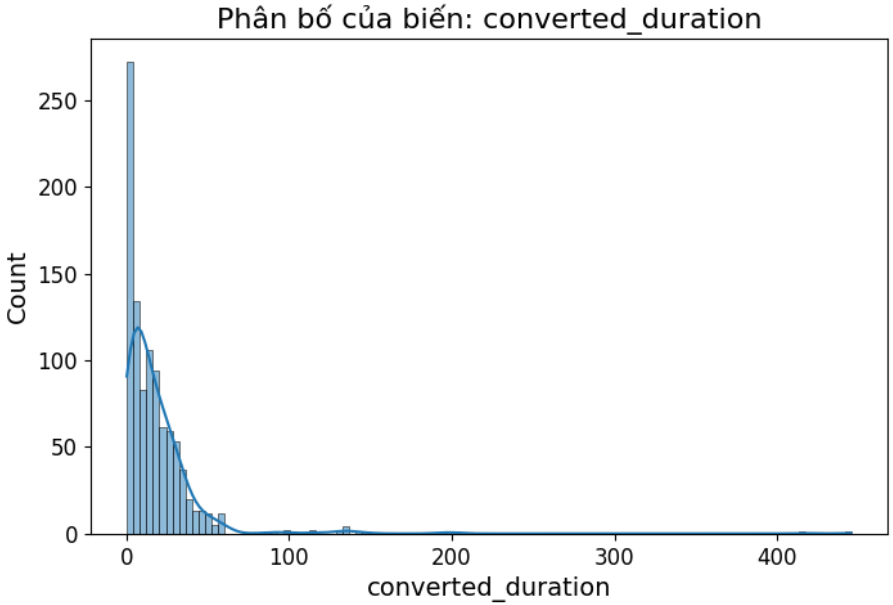




- **Lượt xem, Bình luận, Thích**: Phân phối lệch phải, trung bình cao hơn trung vị, độ lệch chuẩn lớn, nhiều video nổi bật với số lượng tương tác rất cao.

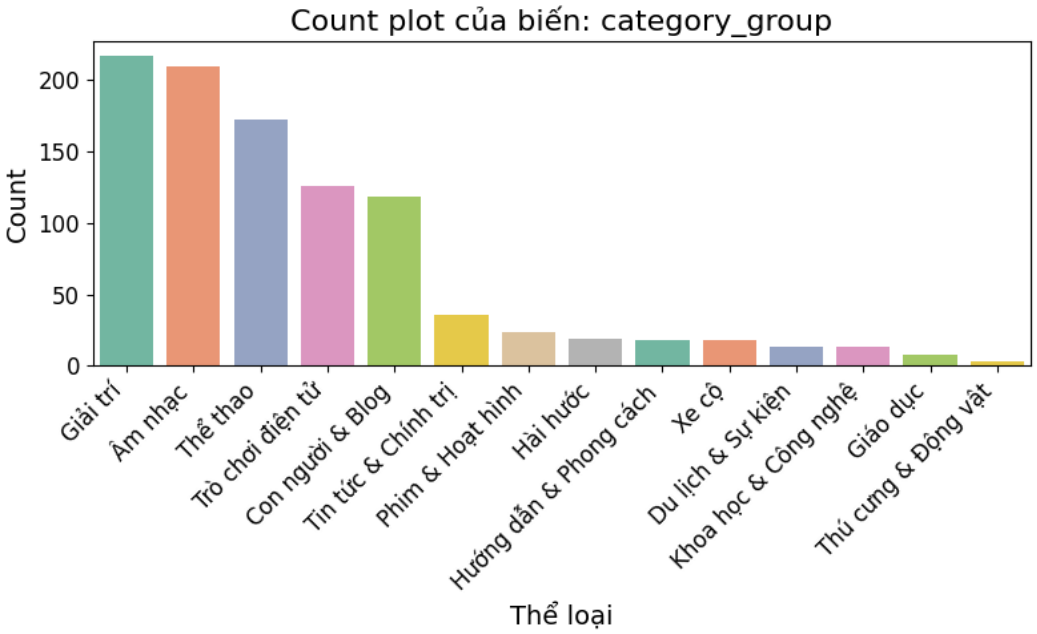
Phương pháp – EDA đơn biến

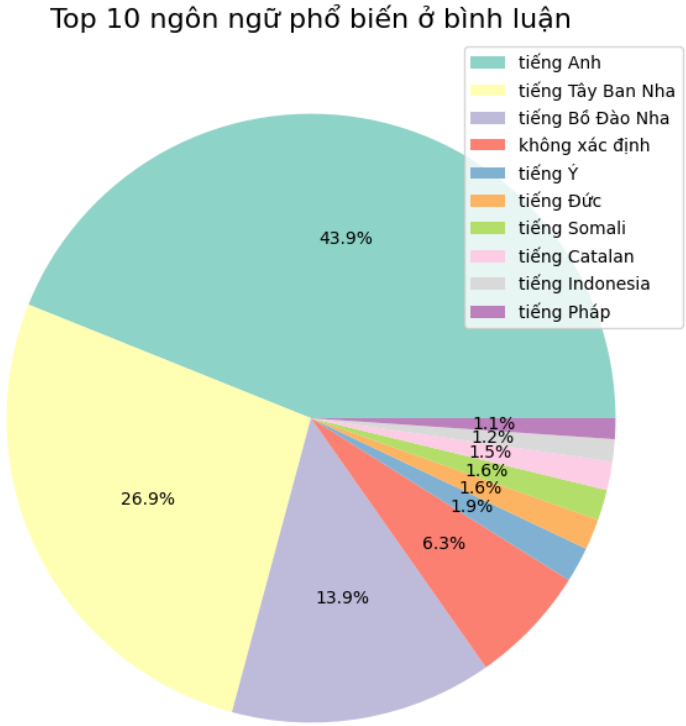




- **Thời lượng video**: Phân phối lệch phải, đa số video ngắn, một số ít video rất dài.
- **Thời gian đăng tải**: Phân phối phức tạp, có cả video mới và video cũ.

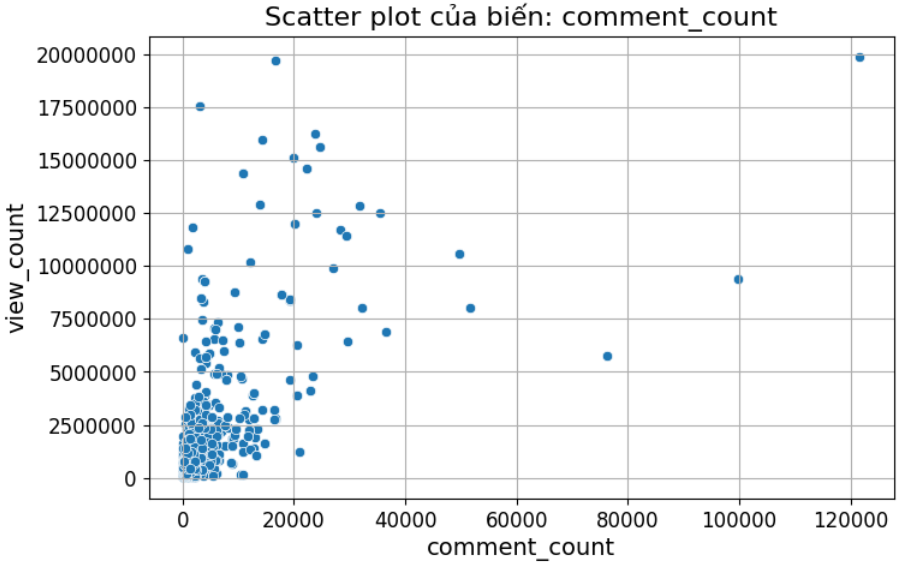
Phương pháp – EDA đơn biến

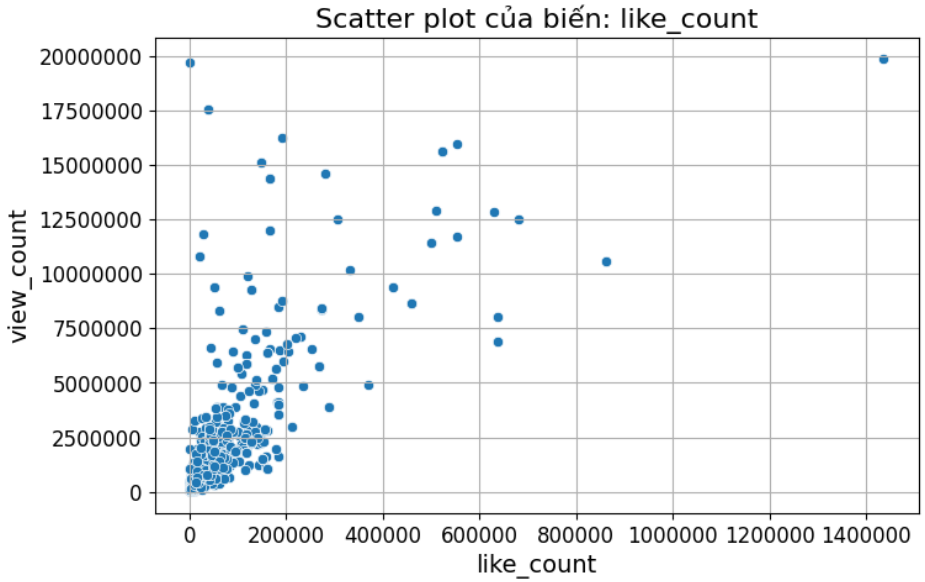




- **Phụ đề**: Đa số video không có phụ đề.
- **Thể loại**: "Giải trí", "Âm nhạc", "Thể thao" phổ biến nhất.
- **Ngôn ngữ bình luận**: Tiếng Anh chiếm ưu thế.

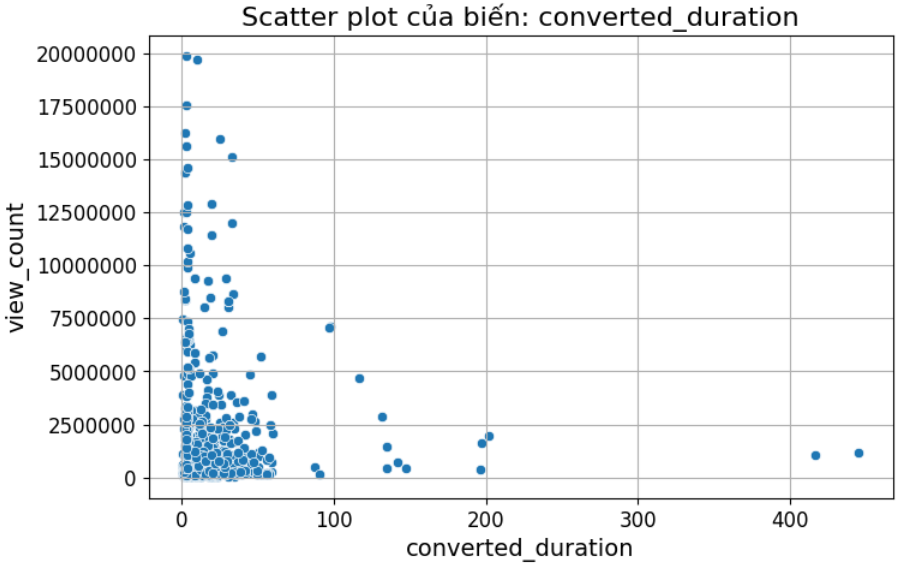
Phương pháp – EDA 2 biến

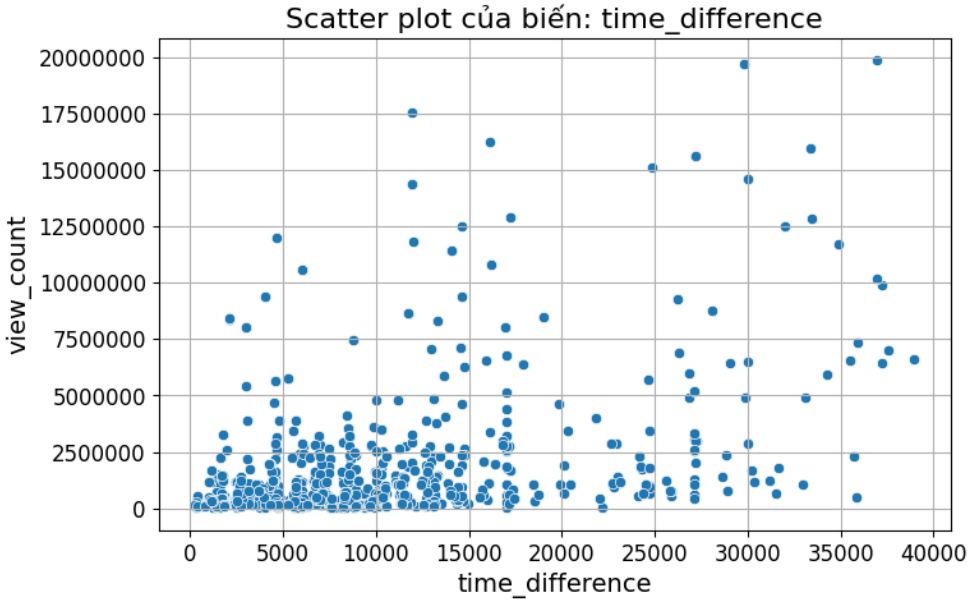




- **Bình luận**: Tương quan yếu, không quyết định lượt xem.
- **Thích**: Tương quan dương, thích nhiều có xu hướng xem nhiều.

Phương pháp – EDA 2 biến



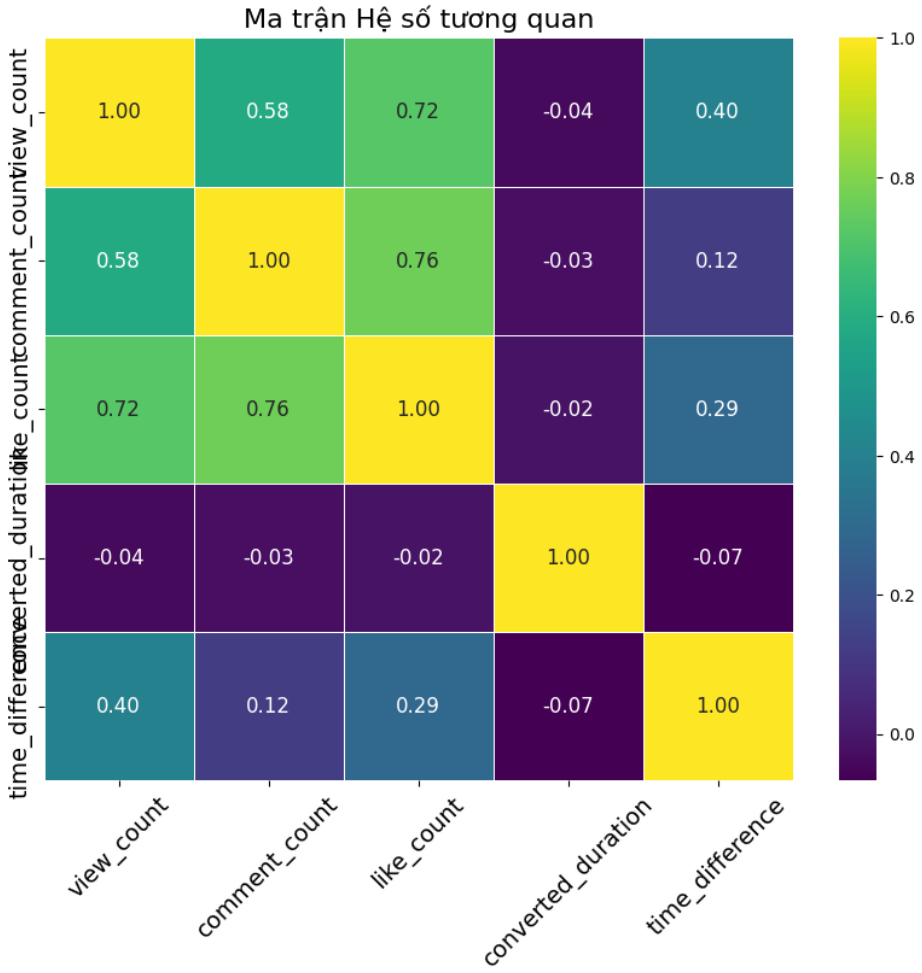


- **Thời lượng**: Không có tương quan rõ ràng với lượt xem.
- **Thời gian đăng tải**: Video mới thường có view cao ban đầu, nhưng video cũ vẫn có thể viral.

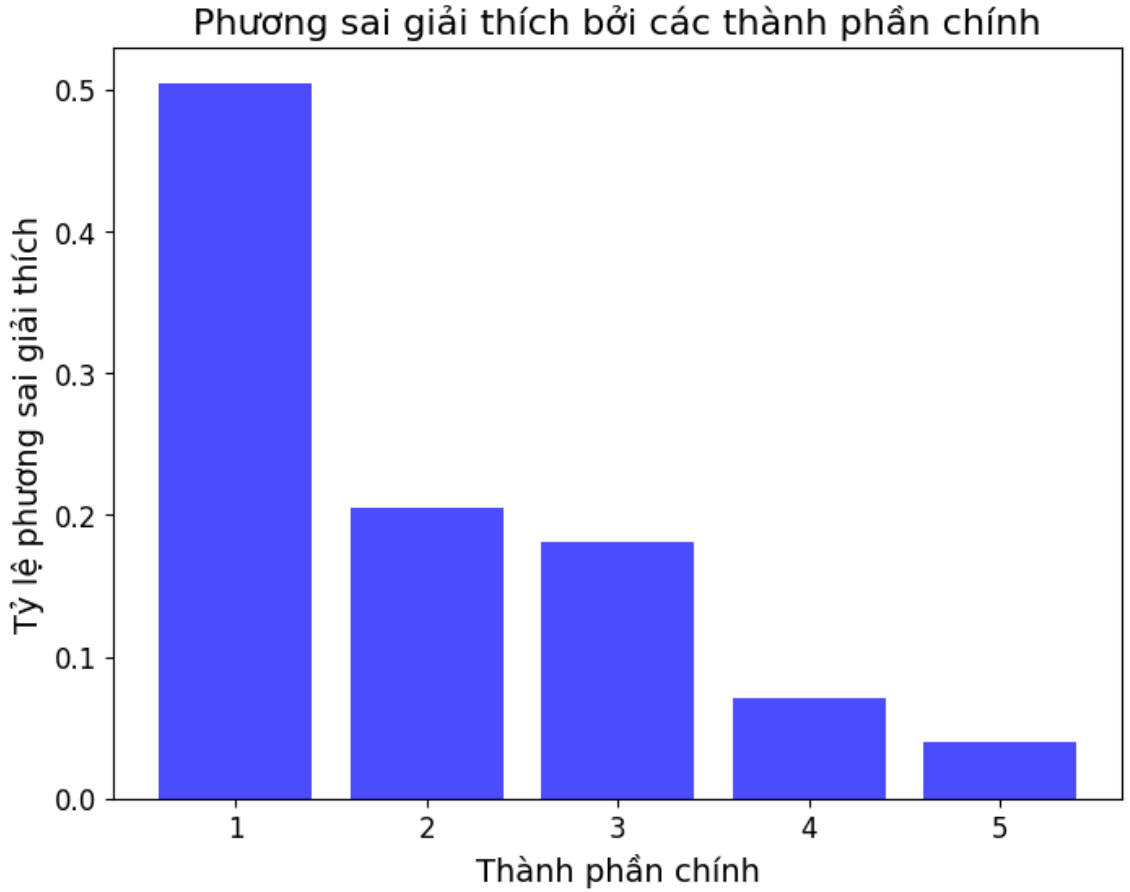
Phương pháp – EDA 2 biến

- **Kết luận**:
  - Lượt thích có ảnh hưởng tích cực nhất đến lượt xem.
  - Thời lượng và thời gian đăng tải ít có mối quan hệ trực tiếp đến lượt xem.
  - Bình luận có thể đóng góp nhưng không phải yếu tố chính ảnh hưởng.

Phương pháp – EDA đa biến



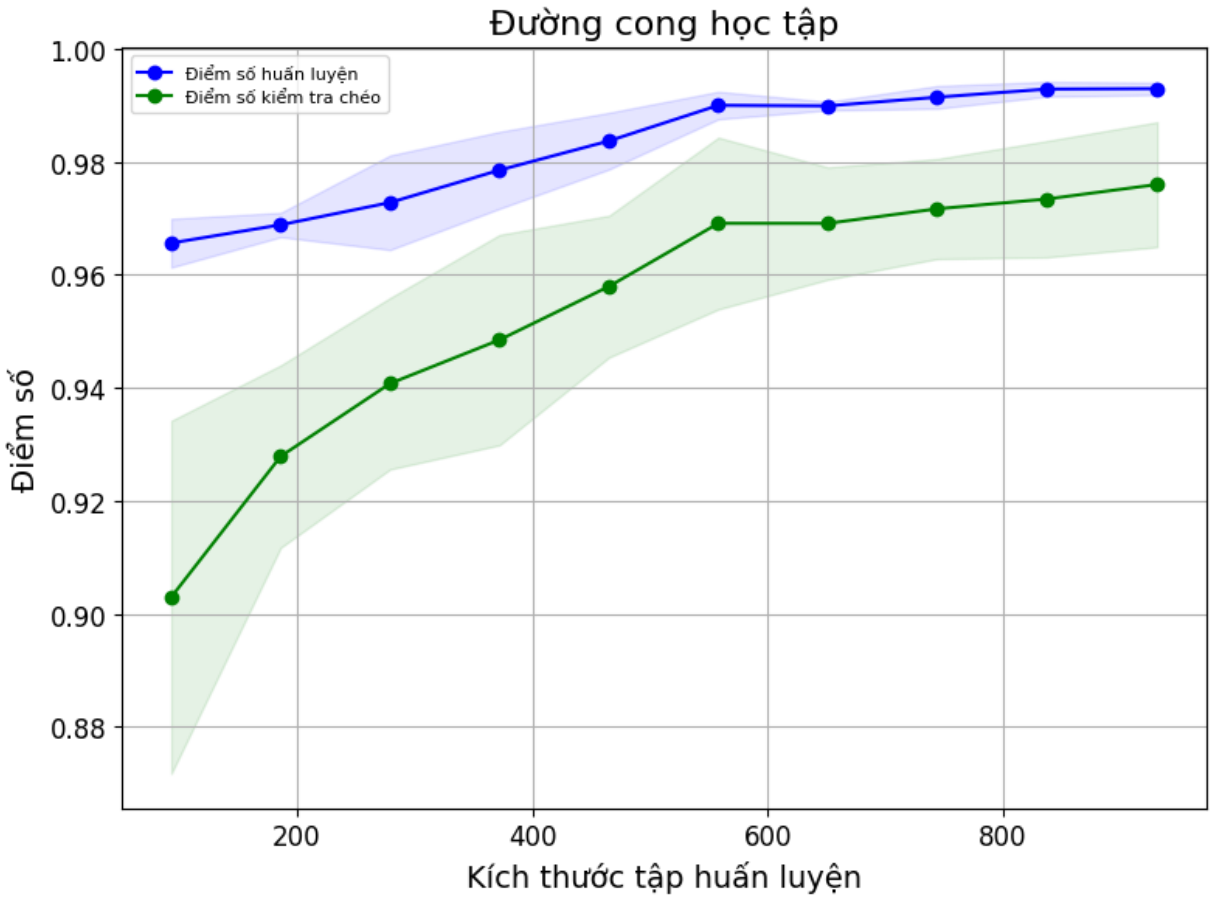
Phương pháp – EDA đa biến



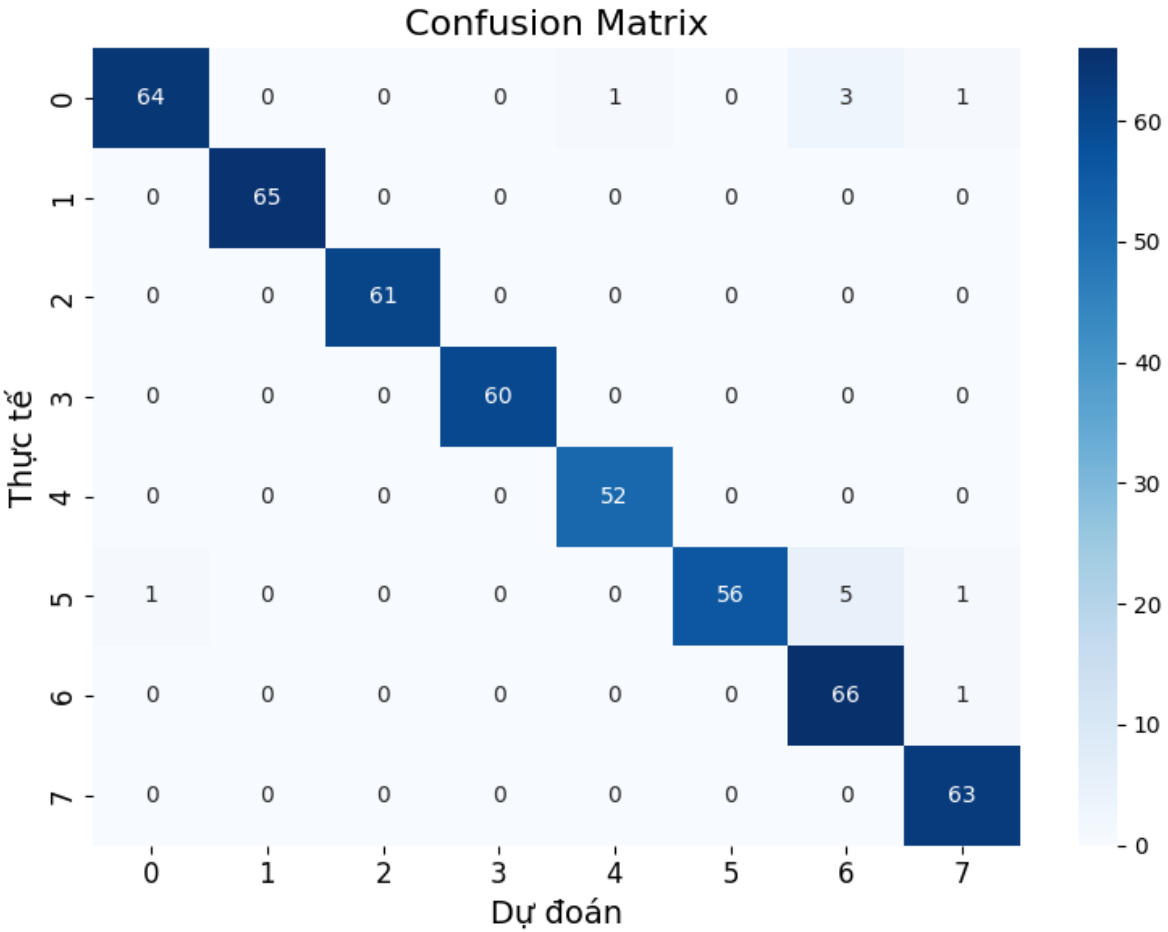
Phương pháp – EDA đa biến

- **Tương tác**: Lượt xem, thích, bình luận có xu hướng đi cùng nhau. Lượt thích tương quan mạnh nhất với lượt xem.
- **Thời lượng & Thời gian**: Ít tương quan trực tiếp với tương tác.
- **PCA**: 2-3 thành phần chính giữ lại phần lớn thông tin.
- **Ý nghĩa**: Tương tác là yếu tố then chốt. Thời lượng và thời gian đăng tải ít ảnh hưởng trực tiếp đến mức độ phổ biến.

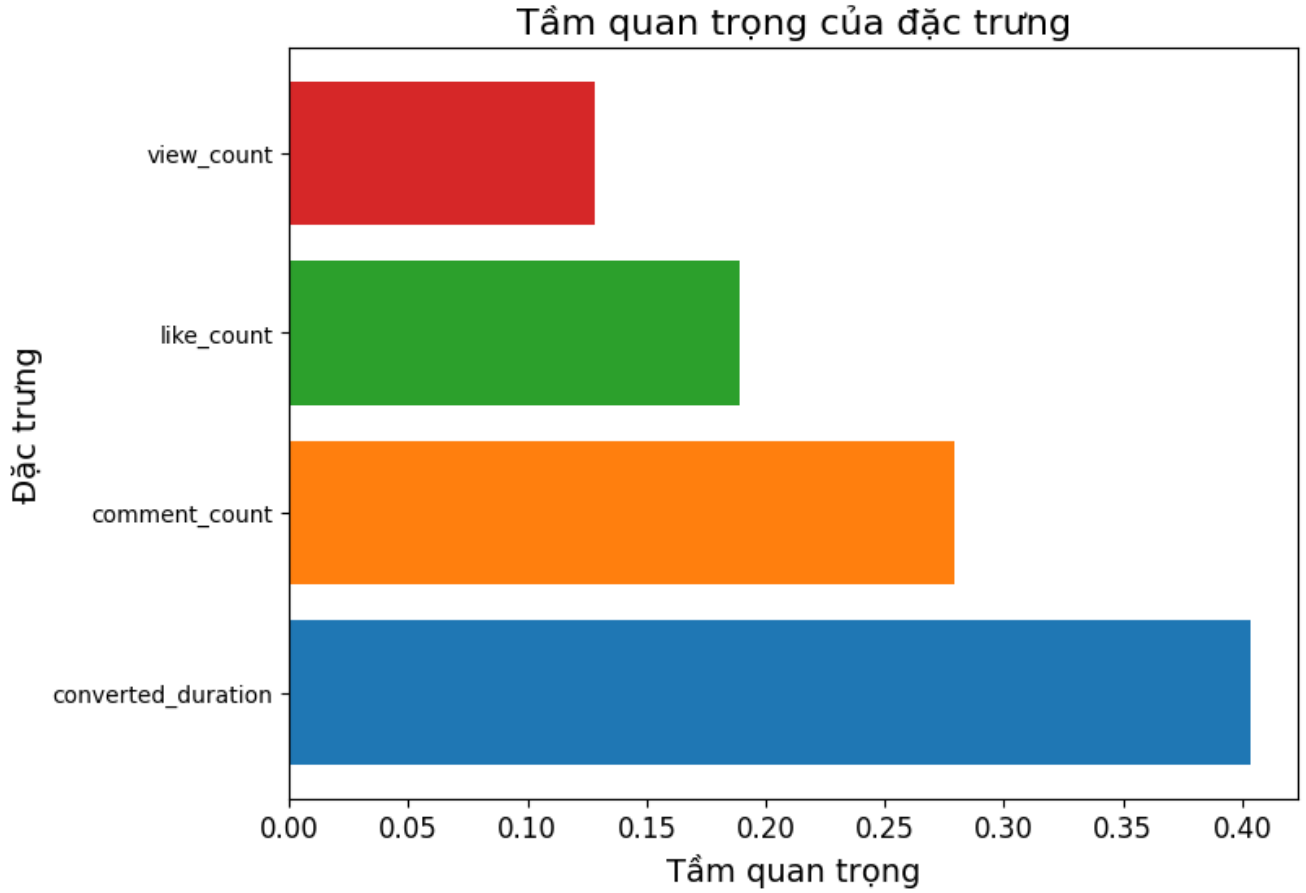
Phương pháp – Mô hình RF



Phương pháp – Mô hình RF



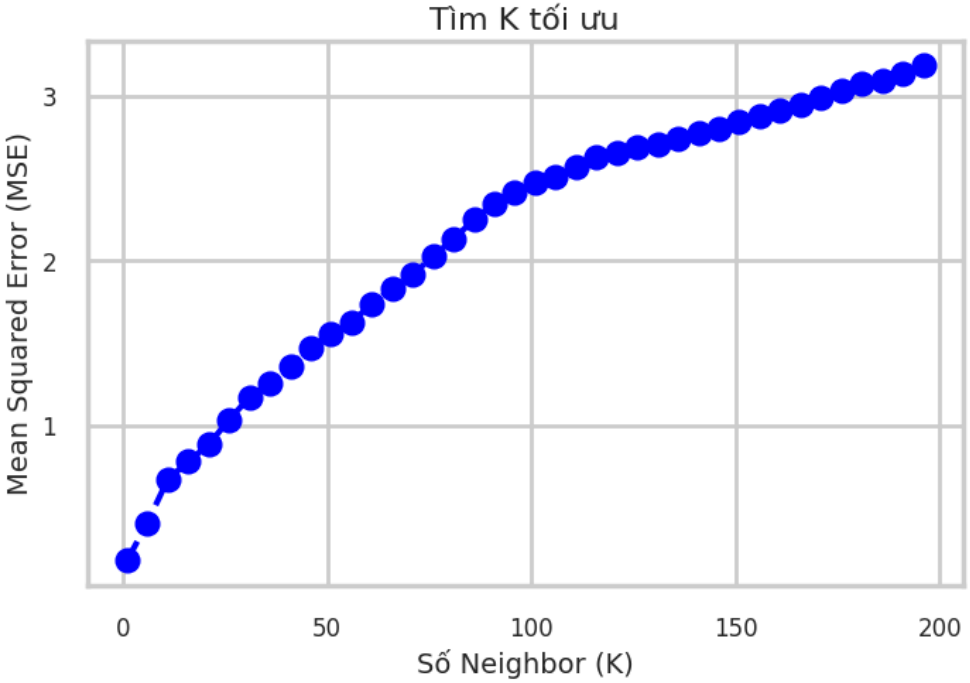
Phương pháp – Mô hình RF

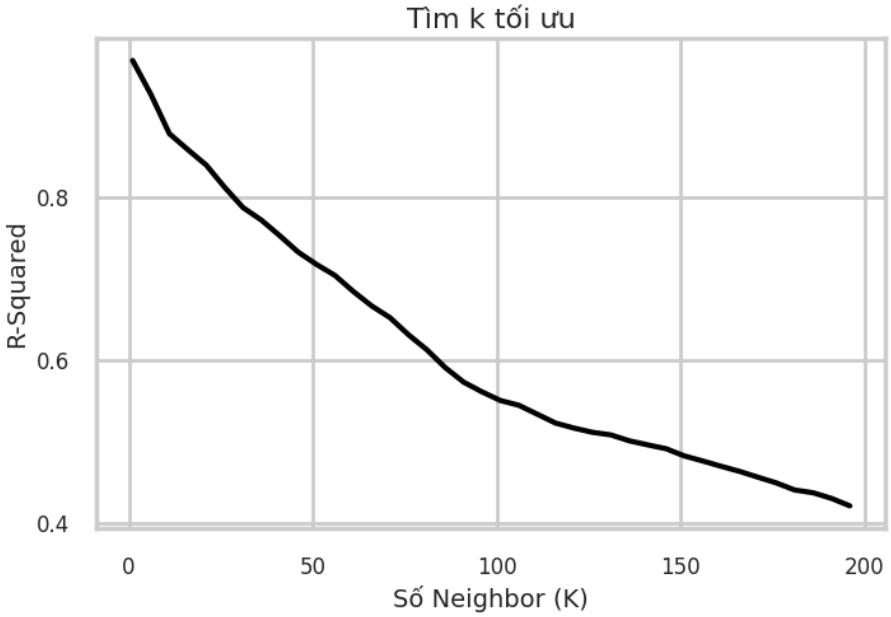


Phương pháp – Mô hình RF

- **Độ chính xác**: 98% (tuning siêu tham số).
- **Độ chính xác Cross-validation**: Trên 96%.
- **Độ chính xác từng cụm**: Hầu hết trên 90%, 3 cụm đạt 100%.
- **Độ quan trọng đặc trưng**:
  - Thời lượng video: Cao nhất.
  - Lượt xem: Thấp nhất.
- **Kết luận**: Mô hình dự đoán cụm video rất tốt. Thời lượng video và số lượng bình luận là hai yếu tố quan trọng nhất để phân loại.

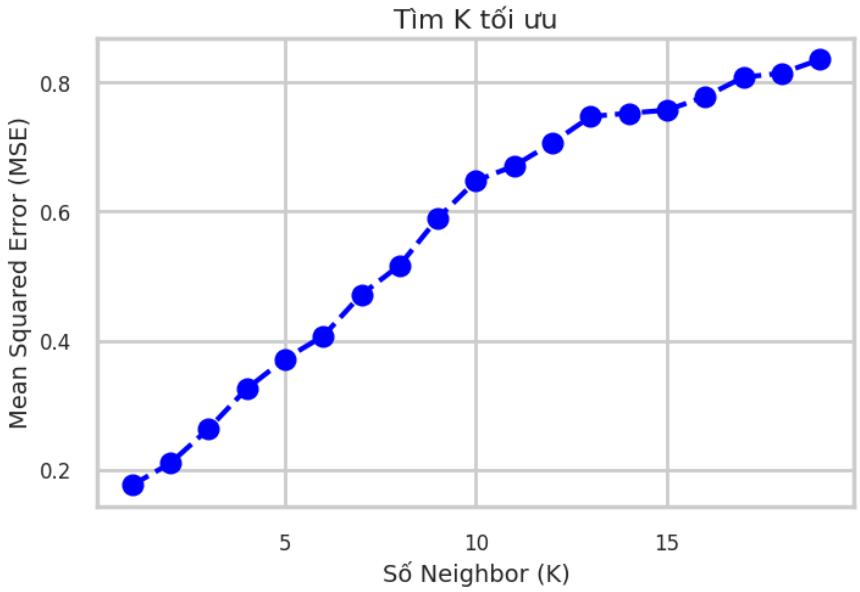
Phương pháp – Mô hình KNN

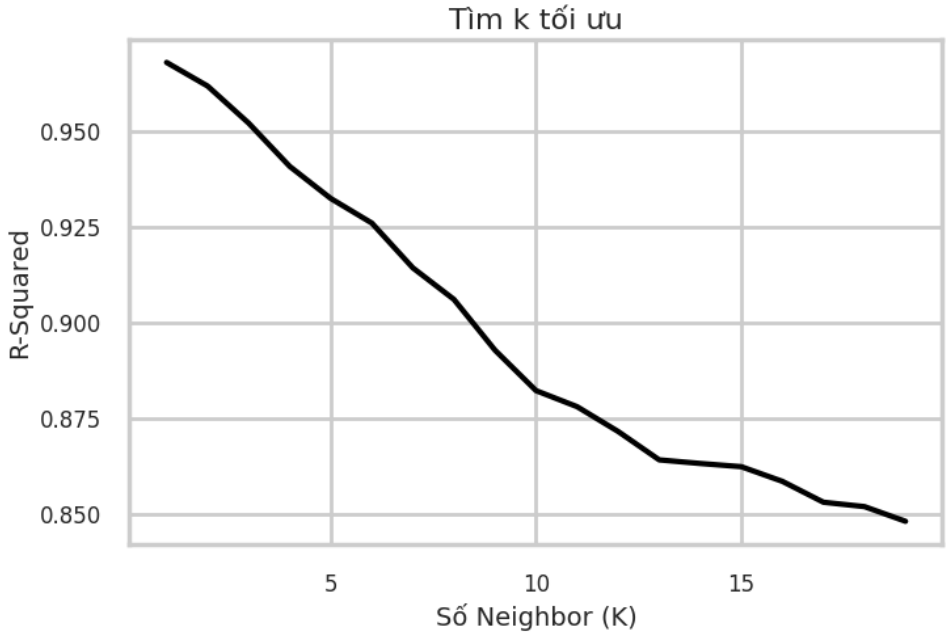




- **Mô hình:** KNN với giá trị k từ 1 đến 200.

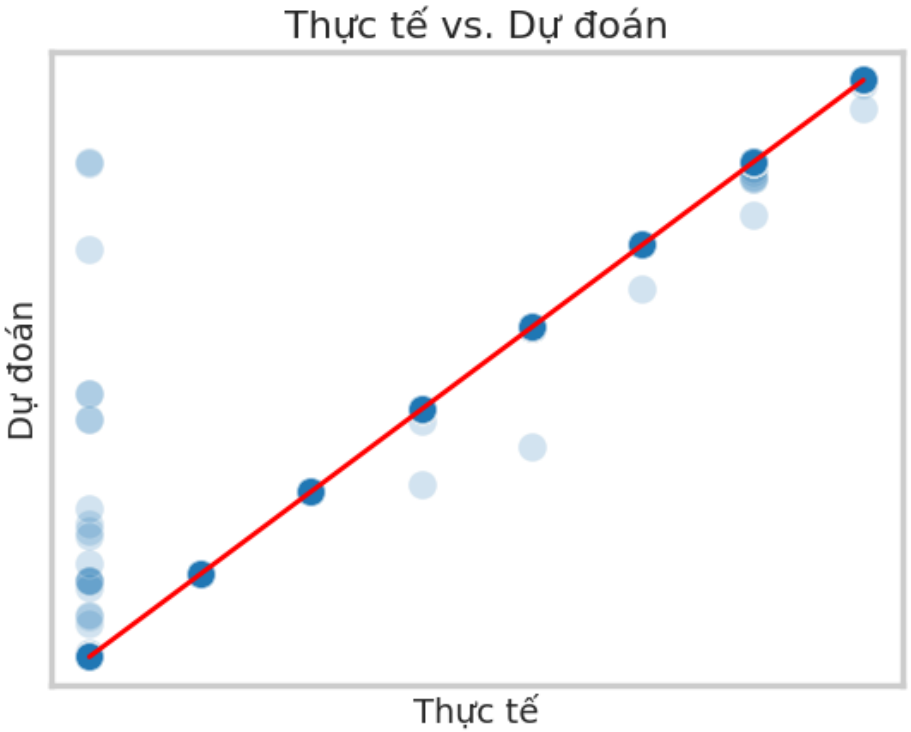
Phương pháp – Mô hình KNN

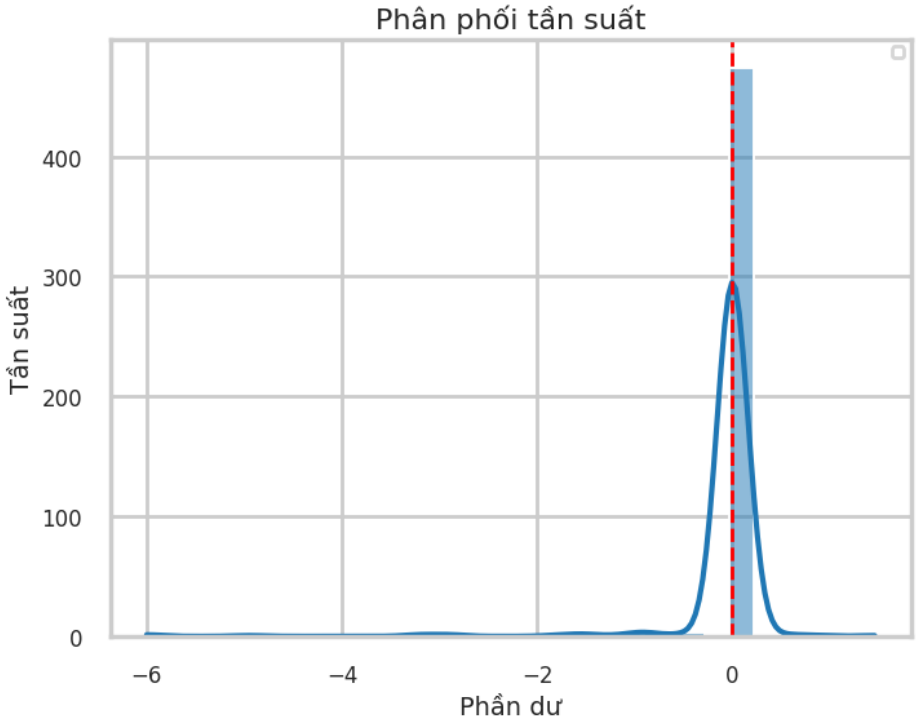




- **Mô hình:** KNN với giá trị k từ 1 đến 19.

Phương pháp – Mô hình KNN

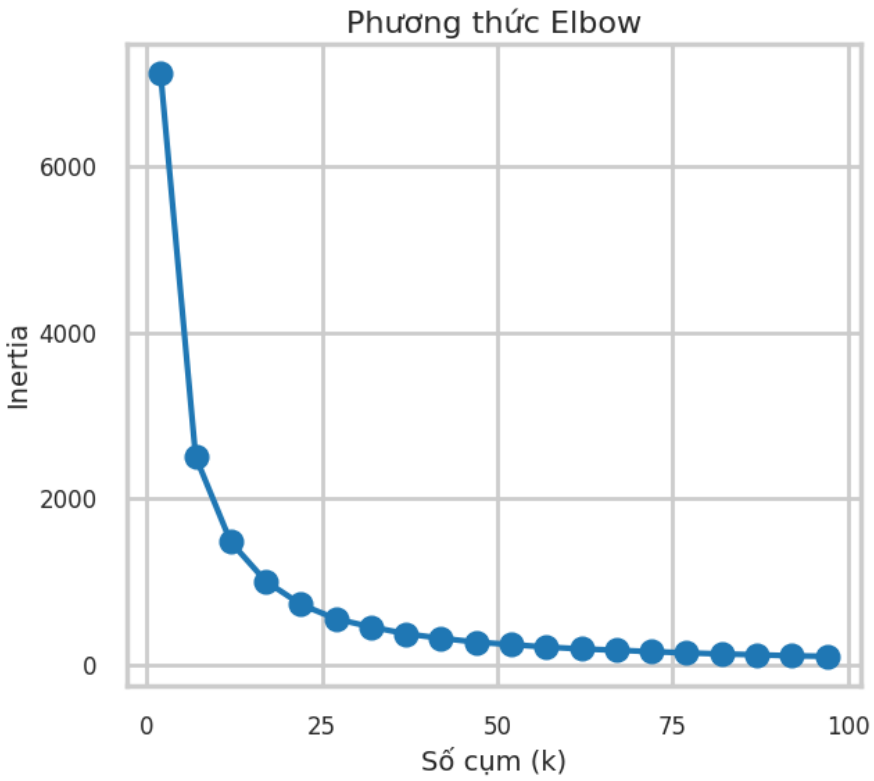


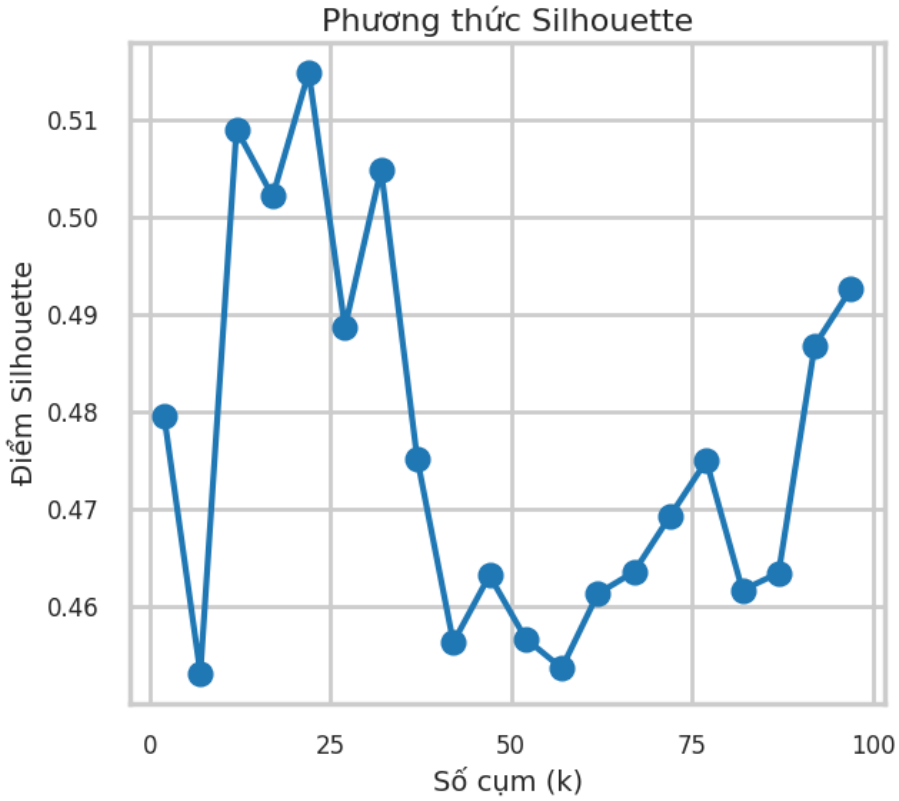


Phương pháp – Mô hình KNN

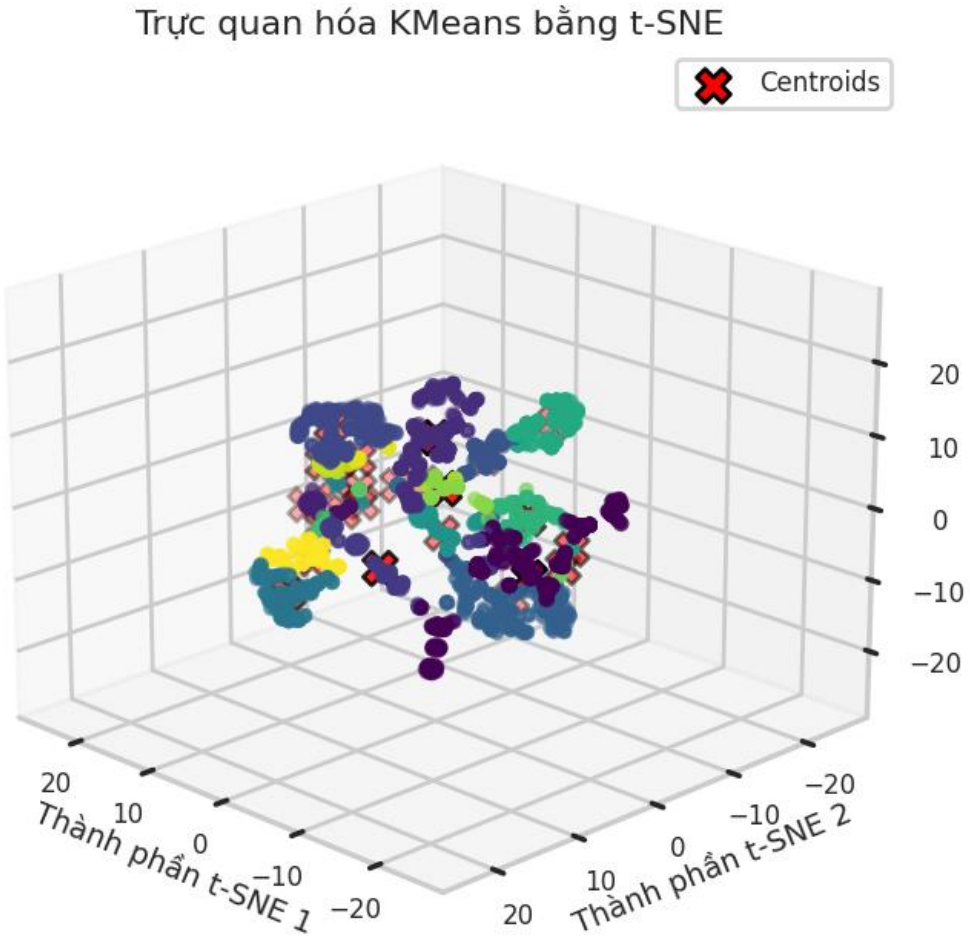
- **Mô hình:** KNN (k=15).
- **R-squared: 0.944** (độ chính xác cao).
- **Lỗi:** MAE=0.095, MSE=0.31, RMSE=0.55 (tương đối thấp).
- **Đặc trưng quan trọng**:
  - Thời lượng video (converted_duration)
  - Lượt thích (like_count)
  - Lượt bình luận (comment_count)
  - Thời gian đăng tải (time_difference)
- **Kết luận:** KNN dự đoán lượt xem tốt, các yếu tố đầu vào quan trọng cho đề xuất YouTube.

Phương pháp – Mô hình KMeans





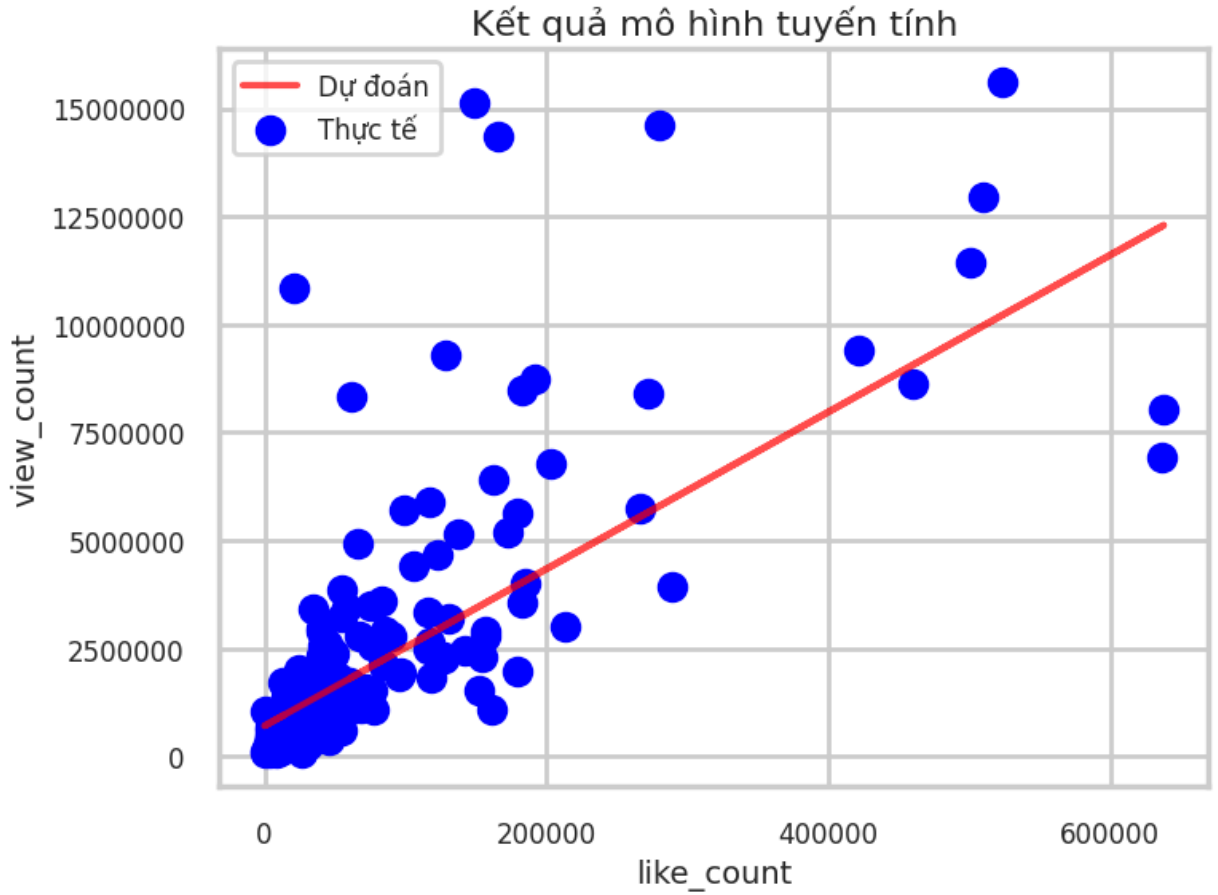
Phương pháp – Mô hình KMeans



Phương pháp – Mô hình KMeans

- **Tìm k tối ưu**:
  - Elbow: Gợi ý k=20-25.
- **Đánh giá** (k=20-24):
  - Silhouette Score: 0.49 - 0.52 (tương đối tốt).
  - Davies-Bouldin Index: 0.68 - 0.76 (tương đối tốt, k=22 & 24 tốt hơn).
- **Số cụm lựa chọn**: k=22 hoặc k=24.
- **Kết luận**: K-Means phân cụm dữ liệu thịnh hành tương đối tốt, với 22 hoặc 24 cụm là lựa chọn phù hợp dựa trên các chỉ số đánh giá.

Phương pháp – Hồi quy tuyến tính



Phương pháp – Hồi quy tuyến tính

- **Đa cộng tuyến:** Không nghiêm trọng (VIF < 5).
- **Like Count và View Count:** Tương quan dương.
- **R^2 = 0.55:** Mô hình giải thích ~55% biến động lượt xem.
- **MAE ~ 1 triệu:** Sai số dự đoán lớn.
- **Kết luận:** Like Count ảnh hưởng đến View Count nhưng không phải yếu tố duy nhất. Cần xem xét thêm biến. Nên kết hợp hồi quy đa biến.

Kết quả – Thảo luận

- **Hiệu suất mô hình:**
  - Random Forest là mô hình tốt nhất.
  - KNN có tiềm năng tốt.
  - K-Means hiệu quả.
- **Đóng góp:**
  - Dữ liệu đã xử lý, chuẩn bị kỹ lưỡng.
  - Kết quả có ý nghĩa thống kê, tiềm năng ứng dụng thực tế.
- **Hạn chế:**
  - Lượng dữ liệu hạn chế về quy mô và độ đa dạng.
  - Xử lý ngoại lai và phân tích cảm xúc còn đơn giản.
  - Chưa triệt để mối quan hệ phi tuyến.

Kết luận – Đề nghị

- **Kết luận:**
  - Xác định đặc trưng video thịnh hành.
  - Ứng dụng học máy và NLP.
  - Nền tảng cho nghiên cứu tiếp theo.
- **Đề nghị:**
  - Mở rộng dữ liệu.
  - Cải thiện mô hình.
  - Giảm bất định.
  - Xây dựng dashboard.
  - Thử nghiệm thực tế.