Trường Đại Học Sài Gòn - SGU

Khoa Toán - Ứng Dụng

BÁO CÁO ĐỒ ÁN CUỐI KÌ KẾT THÚC HỌC PHẦN

Bộ Môn: Machine Learning (858015)

CHỦ ĐỀ: "Phân tích, đánh giá mô hình dữ liệu video Trending trên YouTube"

Giảng viên: **Vũ Ngọc Thanh Sang**

Sinh Viên: **Nguyễn Trương Cao Sơn**

Lớp**: DDU1231**

TP Hồ Chí Minh, ngày 30, tháng 5, năm 2025

MỤC LỤC

LỜI MỞ ĐẦU

Máy học là thuật ngữ chung để chỉ việc máy tính học từ dữ liệu. Đây là một nhánh của Trí tuệ nhân tạo (AI), trong đó các thuật toán được sử dụng để thực hiện một nhiệm vụ cụ thể mà không cần lập trình rõ ràng – thay vào đó, chúng nhận diện các mẫu trong dữ liệu và đưa ra dự đoán dựa trên những gì đã học khi có dữ liệu mới. Máy học (ML) là một cách hiệu quả để tự động hóa các nhiệm vụ phức tạp, vượt xa hơn so với tự động hóa dựa trên quy tắc. Máy học (Machine Learning – ML) đã thu hút được nhiều sự chú ý trong những năm gần đây nhờ việc ứng dụng rộng rãi trong nhiều lĩnh vực. Từ phát hiện gian lận thẻ tín dụng đến quảng cáo nhắm mục tiêu trên mạng xã hội, ML đang được sử dụng cho những nhiệm vụ trước đây do con người thực hiện nhưng giờ có thể tự động hóa thông qua các thuật toán dựa trên cơ sở dữ liệu lớn.

Chương I: GIỚI THIỆU VỀ MÁY HỌC (MACHINE LEARNING)

1.1. Giới thiệu về máy học:

1.1.1. Sơ lượt về máy học:

Học máy hay máy học (machine learning):

Là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể. Các thuật toán học máy xây dựng một mô hình dựa trên dữ liệu mẫu, được gọi là dữ liệu huấn luyện, để đưa ra dự đoán hoặc quyết định mà không cần được lập trình chi tiết về việc đưa ra dự đoán hoặc quyết định này. Ví dụ như các máy có thể "học" cách phân loại thư điện tử xem có phải thư rác (*spam*) hay không và tự động xếp thư vào thư mục tương ứng. Học máy rất gần với suy diễn thống kê (*statistical inference*) tuy có khác nhau về thuật ngữ. Một nhánh của học máy là *học sâu* phát triển rất mạnh mẽ gần đây và có những kết quả vượt trội so với các phương pháp học máy khác. Học máy có liên quan lớn đến thống kê, vì cả hai lĩnh vực đều nghiên cứu việc phân tích dữ liệu, nhưng khác với thống kê, học máy tập trung vào sự phức tạp của các giải thuật trong việc thực thi tính toán. Nhiều bài toán suy luận được xếp vào loại bài toán NP-khó, vì thế một phần của học máy là nghiên cứu sự phát triển các giải thuật suy luận xấp xỉ mà có thể xử lý được.

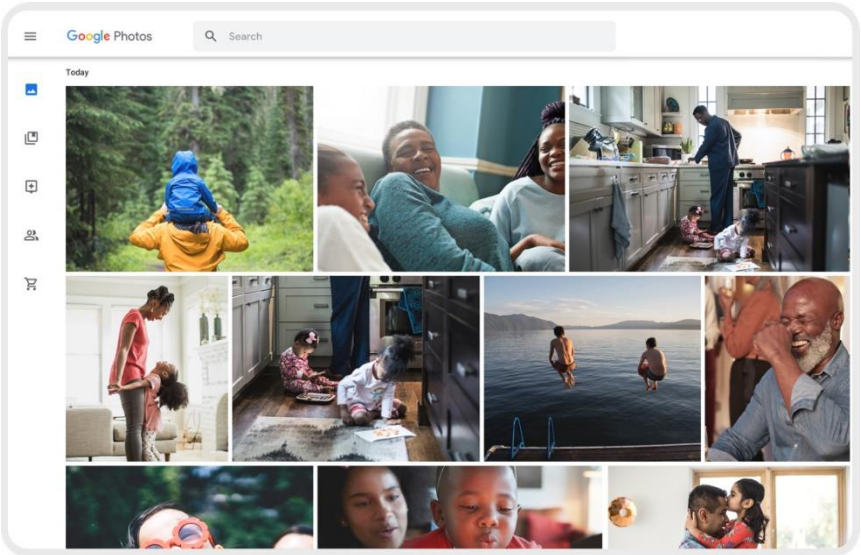
Học máy có hiện nay được áp dụng rộng rãi bao gồm máy truy tìm dữ liệu, chẩn đoán y khoa, phát hiện thẻ tín dụng giả, phân tích thị trường chứng khoán, phân loại các chuỗi DNA, nhận dạng tiếng nói và chữ viết, dịch tự động, chơi trò chơi và cử động rô-bốt (*robot locomotion*).

1.2. Các dạng máy học trong Machine Learning:

Để hiểu Máy học (Machine Learning) được sử dụng như thế nào trong kinh doanh và cách thức hoạt động của nó, điều quan trọng là phải biết các cách khác nhau mà ML có thể hoạt động. Có ba cách phổ biến nhất mà máy móc có thể học: Học có giảm sát (Supervised learning), Học không giám sát (Unsupervised learning) và Học tăng cường (Reinforcement learning)

1.2.1. Học có giám sát là gì? Supervised learning là gì?:

Học có giám sát Supervised learning sử dụng dữ liệu đã được gán nhãn hoặc đánh dấu để huấn luyện các mô hình Máy học (ML). Các thuật toán có thể được đào tạo để **phân loại dữ liệu chính xác** hoặc **dự đoán kết quả**. Do đó, học có giám sát cho phép các doanh nghiệp **giải quyết các vấn đề thực tế ở quy mô lớn**, chẳng hạn như tách thư rác khỏi email.

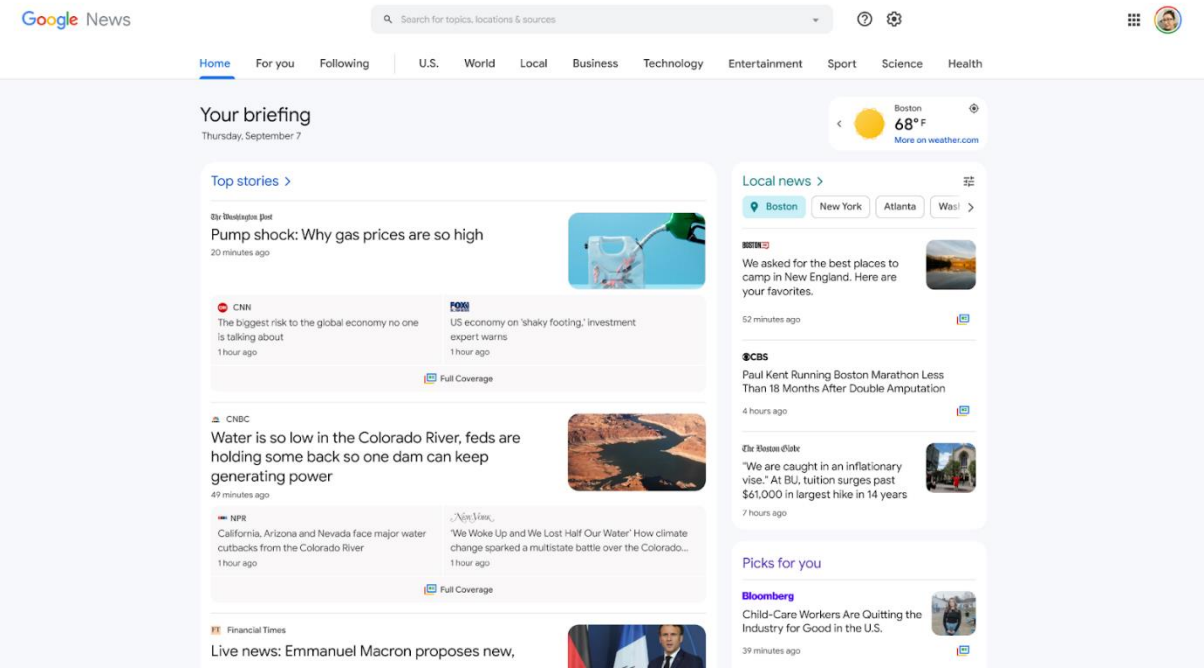


Google Photos sử dụng máy học có giám sát Supervised learning nhận diện và nhóm các hình ảnh theo người hoặc địa điểm bằng cách sử dụng dữ liệu được gán nhãn.

Học có giám sát thu thập dữ liệu từ kinh nghiệm trước đó hoặc tạo ra kết quả dữ liệu từ sự kiện đó. Nó giúp tối ưu hóa các yêu cầu về hiệu suất dựa trên kinh nghiệm trước đó và giải quyết nhiều vấn đề tính toán thực tế khác nhau

1.2.2. Học không giám sát là gì? Unsupervised learning là gì:

Học không giám sát Unsupervised learning đánh giá và phân nhóm dữ liệu không được gán nhãn. Các thuật toán này **tự động phát hiện các mẫu ẩn** hoặc **nhóm dữ liệu.** So với học có giám sát, các thuật toán học không giám sát có thể xử lý các vấn đề phức tạp hơn. Học không giám sát cho phép các công ty **khám phá dữ liệu, giúp họ phát hiện các mẫu nhanh hơn** so với quan sát bằng con người.



Google News phân loại các câu chuyện về cùng một chủ đề từ các nguồn tin tức trực tuyến khác nhau bằng cách sử dụng máy học không giám sát Unsupervised learning

Học không giám sát tìm ra tất cả các mẫu chưa được phát hiện trước đó trong dữ liệu và hỗ trợ trong việc phát hiện các đặc điểm hữu ích cho việc phân loại.

1.2.3. Học tăng cường (Reinforcement learning):

Học tăng cường Reinforcement learning huấn luyện các mô hình máy tính để đưa ra quyết định bằng cách đặt trí tuệ nhân tạo (AI) vào một tình huống giống như trò chơi. Trong học tăng cường, thuật toán **học qua phương pháp thử và sai** (trial and error) bằng cách **sử dụng phản hồi** từ các hành động của nó. Mô hình Học máy (ML) nhận được phản hồi tích cực và tiêu cực về hành vi của nó và **học từ chính kinh nghiệm của nó** để **tối đa hóa độ chính xác của kết quả đầu ra.**

Một hệ thống chơi cờ vua sử dụng máy học tăng cường Reinforcement leanring bằng cách chơi nhiều ván và điều chỉnh chiến lược của nó dựa trên việc chiến thắng hay thất bại.

Chương II: TỔNG QUAN VỀ ĐỀ TÀI

2.1. Giới thiệu về nền tảng giải trí YouTube:

YouTube là một nền tảng chia sẻ video trực tuyến lớn nhất thế giới, thuộc sở hữu của Google. Kể từ khi ra mắt vào năm 2005, YouTube đã nhanh chóng trở thành một kho lưu trữ video phong phú, bao gồm các nội dung giáo dục, giải trí, vlog, nhạc, tin tức, và nhiều lĩnh vực khác. Mỗi ngày, hàng tỷ video được tải lên và hàng tỷ lượt xem được thực hiện.

Danh sách "Trending" trên YouTube là tập hợp các video đang được quan tâm nhiều trong khoảng thời gian ngắn. Việc phân tích xu hướng video trending góp phần hiểu rõ hơn sở thị hành vi người dùng, xu hướng nội dung được yêu thích, và giúp các nhà sáng tạo tối ưu hóa chiến lược phát triển kênh của mình.

2.2. Lý do chọn đề tài:

Tầm ảnh hưởng của YouTube: Là một trong những nền tảng trực tuyến được sử dụng nhiều nhất thế giới, YouTube đang định hình xu hướng nội dung và giáo tiếp số lớn người xem hàng ngày.

Dữ liệu phổ biến và thật: Dữ liệu trending video có các thuộc tính để phân tích như lượt xem (views), lượt like, thời gian đăng, danh mục video, thời lượng video, giúp cho việc khai thác xu hướng trở nên khá đáng tin cậy.

Áp dụng học máy: Việc áp dụng Machine Learning vào việc phân tích dữ liệu trending YouTube mang tính thực tiễn cao, gắn liền với nhu cầu xã hội hiện đại và gợi mở hướng nghiên cứu sau này.

Tác động với nhà sáng tạo và marketer: Các kết quả phân tích giúp đề xuất các chiến lược tối ưu cho người làm sáng tạo nội dung hoặc các nhà quảng cáo.

CHƯƠNG III: PHÂN TÍCH DỮ LIỆU VIDEO TRENDING

3.1. Phương pháp nghiên cứu:

- Tiến hành tiền xử và làm sạch dữ liệu (EDA): Khám phá dữ liệu, thống kê mô tả các trường dữ liệu, trực quan hóa để phát hiện phân bố không đều, giá trị ngoại lệ và mối quan hệ giữa các biến.

	rank	Video	Video views	Likes	Dislikes	Category	published
0	1	20 Tennis shots if they were not filmed, NOBOD...	3,471,237	19,023	859	NaN	2017
1	2	Lil Nas X - Old Town Road (Official Movie) ft....	54,071,677	3,497,955	78,799	Music	2019
2	3	JoJo Siwa - Karma (Official Video)	34,206,747	293,563	NaN	Music	2024
3	4	Wiz Khalifa - See You Again ft. Charlie Puth [...	6,643,904,918	44,861,602	NaN	Music	2015
4	5	伊賀の天然水強炭酸水「家族で、シュワシェア。」篇　15秒	236,085,971	38	NaN	NaN	2021
5	6	JP Saxe - If the World Was Ending (Official Vi...	76,834,495	804,353	21,195	Music	2019
6	7	David Kushner - Daylight (Official Music Video)	18,558,390	680,732	NaN	Music	2023
7	8	Power Star Pawan Kalyan Special Surprise To Se...	96,686	1,007	82	Entertainment	2018
8	9	Kulit Kamu Kulit Kering dan Sensitif? Pakai Av...	9,605,969	6	NaN	NaN	2023
9	10	Totti with a funny penalty	8,353,318	5,613	1,082	Sports	2007

- Hồi quy (Regression): Dự đoán lượt xem (views) dựa vào các yếu tố như số lượt like, số lượt bình luận, độ dài video, thời điểm đăng video, loại nội dung. Mục tiêu là đánh giá khả năng dự đoán và xác định các biến có ảnh hưởng lớn nhất.
- Phân loại (Classification): Gán nhãn các video là "hit" hay không (ví dụ: nếu views > 1 triệu thì là hit). Sử dụng các mô hình học máy để xác định các yếu tố giúp một video có khả năng trở thành hiện tượng.
- Phân cụm (Clustering): Gom nhóm video có đặc điểm tương tự nhau. Nhằm phát hiện các nhóm video phổ biến như video ca nhạc, vlog giải trí, phân tích nội dung, giúp nhận diện nhóm mục tiêu trong marketing.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   rank         1000 non-null   int64
 1   Video        1000 non-null   object
 2   Video views  1000 non-null   object
 3   Likes        973 non-null    object
 4   Dislikes     687 non-null    object
 5   Category     820 non-null    object
 6   published    1000 non-null   int64
dtypes: int64(2), object(5)
memory usage: 54.8+ KB
```

	0
rank	0
Video	0
Video views	0
Likes	27
Dislikes	313
Category	180
published	0

dtype: int64

3.2. Xử lý và làm sạch dữ liệu:

- Loại bỏ các dòng trống, dòng bị lỗi (missing values hoặc dữ liệu không hợp lệ)

- Chuyển đổi các cột thời gian về đúng định dạng datetime để trích xuất ngày, giờ, tháng
- Phân tích các biến dạng text như tiêu đề, mô tả để chuẩn hóa, nếu cần có thể rút trích đặc trưng (feature engineering)
- Mã hóa các biến phân loại như danh mục nội dung (category), quốc gia (region) bằng One-hot encoding hoặc Label encoding
- Chuẩn hóa các biến đầu vào (standardization) để phục vụ các mô hình nhạy cảm với thang đo như KMeans hoặc Logistic Regression

3.3. Triển khai các mô hình Machine Learning:

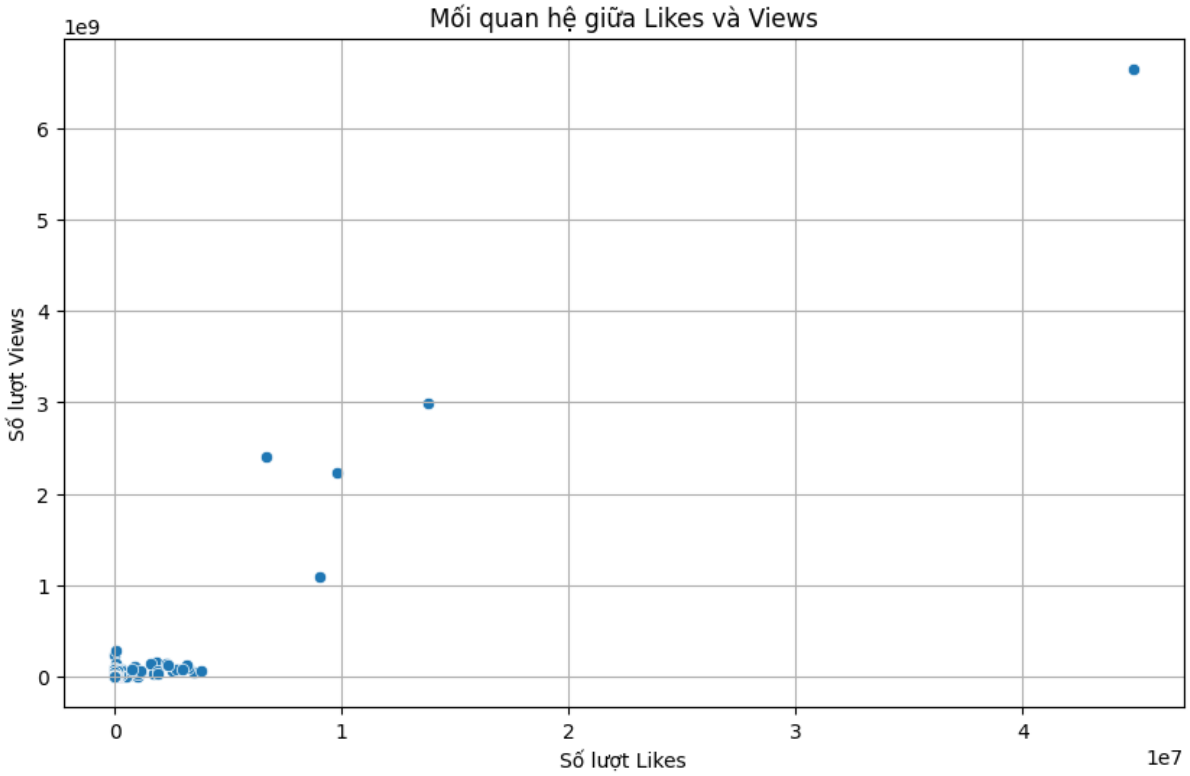
3.3.1. Hồi quy:

- Mô hình sử dụng: Linear Regression, Decision Tree Regressor, Random Forest Regressor
- Biến đầu vào: likes, dislikes, comment count, duration, publish hour
- Biến mục tiêu: views (dạng số thực)
- Kết quả: Random Forest cho kết quả RMSE thấp nhất; số lượt like và thời gian phát hành là yếu tố dự báo mạnh
- Trực quan: biểu đồ scatter giữa views và likes; đồ thị quan hệ giữa views và thời lượng

#INPUT

```
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='Likes', y='Video views')
plt.title('Mối quan hệ giữa Likes và Views')
plt.xlabel('Số lượt Likes')
plt.ylabel('Số lượt Views')
plt.grid(True)
plt.show()
```

#OUTPUT



3.3.2. Phân loại:

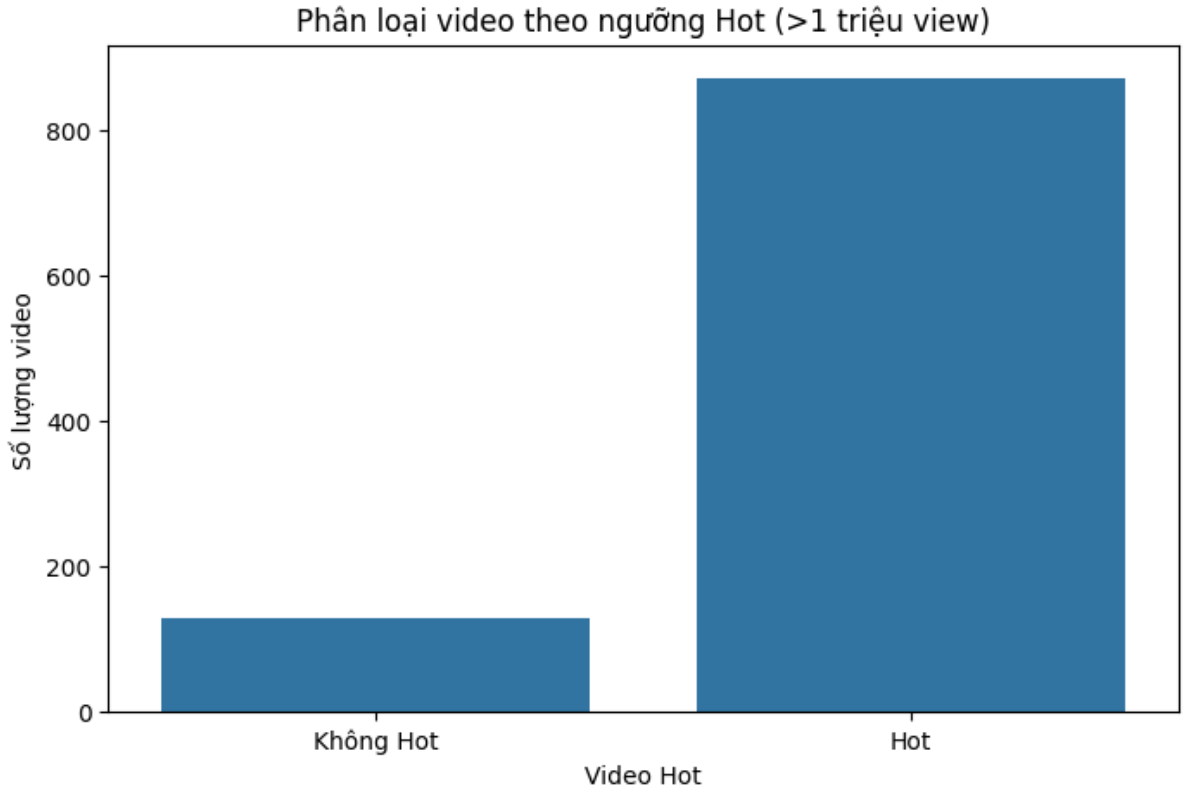
- Tiêu chí gán nhãn: video được xem là "hot" nếu có views > 1,000,000

#INPUT

```
df['is_hot'] = df['Video views'] > 1_000_000

plt.figure(figsize=(8, 5))
sns.countplot(data=df, x='is_hot')
plt.title('Phân loại video theo ngưỡng Hot (>1 triệu view)')
plt.xlabel('Video Hot')
plt.ylabel('Số lượng video')
plt.xticks([0, 1], ['Không Hot', 'Hot'])
plt.show()
```

#OUTPUT



- Mô hình: Logistic Regression, Decision Tree, Random Forest, XGBoost
- Chia tập train/test: 80/20
- Kết quả: Random Forest và XGBoost đạt độ chính xác >85%
- Phân tích độ quan trọng của đặc trưng: likes và publish hour là yếu tố nổi bật
- Đánh giá qua confusion matrix, classification report (precision, recall, F1-score)

3.3.3. Phân cụm:

- Phương pháp: KMeans
- Đầu vào: views, likes, comments, duration
- Chọn số cụm tối ưu bằng Elbow method và Silhouette Score

#INPUT

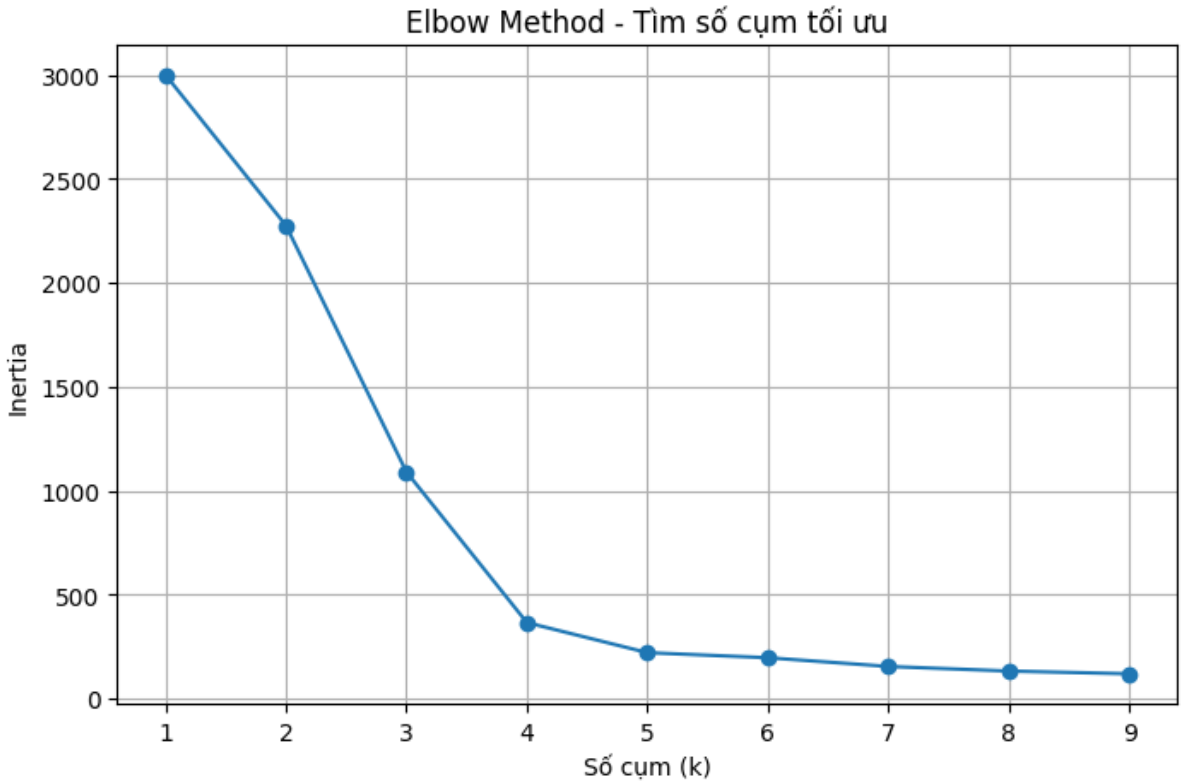
```
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans

# Lấy các biến số để phân cụm
features = df[['Video views', 'Likes', 'Dislikes']].dropna()
scaled = StandardScaler().fit_transform(features)

# Elbow method
inertia = []
K = range(1, 10)
for k in K:
    model = KMeans(n_clusters=k, random_state=42)
    model.fit(scaled)
    inertia.append(model.inertia_)

plt.figure(figsize=(8, 5))
plt.plot(K, inertia, marker='o')
plt.title('Elbow Method - Tìm số cụm tối ưu')
plt.xlabel('Số cụm (k)')
plt.ylabel('Inertia')
plt.grid(True)
plt.show()
```

#OUTPUT



- Kết quả: 3 cụm nổi bật
  - Cụm 1: video giải trí ngắn nhưng tương tác cao (vlog, tiktok cut)
  - Cụm 2: video âm nhạc, MV có lượt xem cực cao
  - Cụm 3: video thông tin chuyên sâu, thời lượng dài, lượt tương tác thấp hơn

#INPUT

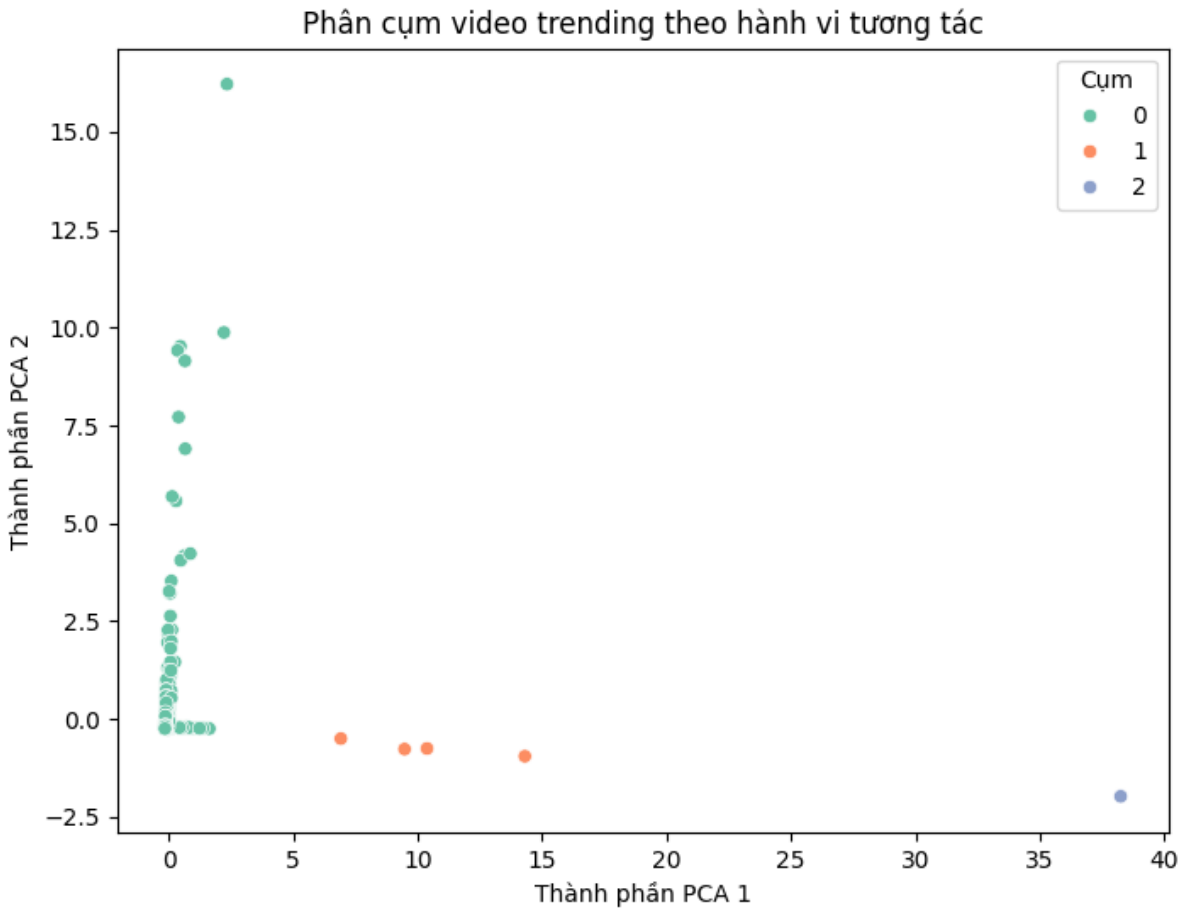
```
from sklearn.decomposition import PCA

# Phân cụm
kmeans = KMeans(n_clusters=3, random_state=42)
labels = kmeans.fit_predict(scaled)

# PCA để giảm xuống 2 chiều
pca = PCA(n_components=2)
components = pca.fit_transform(scaled)

# Vẽ cụm
plt.figure(figsize=(8, 6))
sns.scatterplot(x=components[:, 0], y=components[:, 1], hue=labels, palette='Set2')
plt.title('Phân cụm video trending theo hành vi tương tác')
plt.xlabel('Thành phần PCA 1')
plt.ylabel('Thành phần PCA 2')
plt.legend(title='Cụm')
plt.show()
```

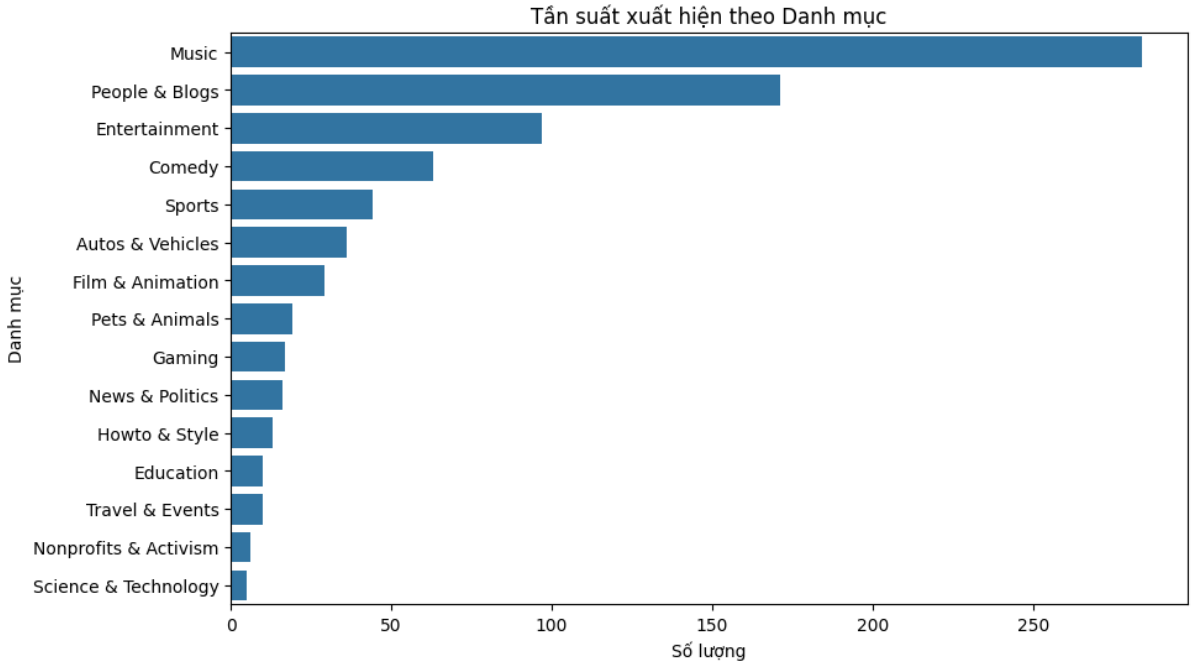
#OUTPUT



- Trực quan hóa bằng biểu đồ scatter 2D và 3D các cụm

3.4. Nhận định và đề xuất:

- Video đăng vào buổi sáng hoặc buổi tối có xu hướng được xem nhiều hơn (có thể do thói quen người dùng)
- Danh mục video như Âm nhạc, Giải trí, Trò chơi có xác suất trending cao



- Like và comment là yếu tố tương quan cao với lượt xem (views) => nên tập trung tối ưu tương tác người dùng
- Các cụm phát hiện giúp nhóm đối tượng cụ thể: người xem ngắn hạn (shorts), người yêu âm nhạc, người thích nội dung học thuật
- Gợi ý cho nhà sáng tạo: kết hợp chủ đề thịnh hành với thời gian phát hành hợp lý; tối ưu tiêu đề, mô tả và thumbnail để tăng tương tác

=> Kết luận: Việc áp dụng ML vào phân tích video trending giúp nhìn rõ xu hướng nội dung và đề xuất hữu ích cho nhiều đối tượng từ marketer đến creator.

KẾT LUẬN

Đề tài "Phân tích, đánh giá mô hình dữ liệu video Trending trên YouTube" đã cho thấy tính khả thi và hữu ích của việc áp dụng các kỹ thuật học máy vào khai phá thông tin từ dữ liệu thực tế. Qua quá trình xử lý, phân tích và áp dụng các mô hình hồi quy, phân loại và phân cụm, nhóm nghiên cứu đã rút ra được những yếu tố chính ảnh hưởng đến độ phổ biến của video trên nền tảng YouTube.

Một số yếu tố như thời gian đăng, danh mục nội dung, lượt thích và bình luận đều đóng vai trò quan trọng trong việc xác định một video có khả năng trở nên trending hay không. Những hiểu biết này mang lại giá trị thực tiễn cho các nhà sáng tạo nội dung, marketer cũng như các nhà nghiên cứu dữ liệu.

Trong tương lai, có thể mở rộng nghiên cứu với tập dữ liệu lớn hơn theo thời gian thực, sử dụng các mô hình học sâu để nâng cao độ chính xác, đồng thời phân tích ảnh hưởng của tiêu đề, mô tả, thẻ hashtag đến khả năng viral của video.

Đây là một bước tiến quan trọng trong việc ứng dụng Machine Learning vào ngành truyền thông số và mở ra nhiều cơ hội khai thác hiệu quả dữ liệu từ các nền tảng mạng xã hội.

Lời cảm ơn

Em xin chân thành cảm ơn *Thầy Vũ Ngọc Thanh Sang* – giảng viên bộ môn Machine Learning – đã tận tình hướng dẫn, góp ý và tạo điều kiện thuận lợi để em hoàn thành đồ án cuối kỳ này.

Trong suốt quá trình thực hiện đề tài, em đã nhận được sự giúp đỡ quý báu từ thầy, bạn bè trong lớp và gia đình. Những lời khuyên chuyên môn, nguồn tài liệu tham khảo và động lực mà mọi người mang lại là một phần không thể thiếu cho sự hoàn thiện của báo cáo.

Qua đề tài này, em không chỉ nâng cao kiến thức chuyên môn về học máy và xử lý dữ liệu thực tế, mà còn rèn luyện kỹ năng tự học, kỹ năng tư duy logic và kỹ năng trình bày học thuật.

Một lần nữa, em xin chân thành cảm ơn tất cả mọi người đã đồng hành và hỗ trợ trong quá trình thực hiện đồ án này.

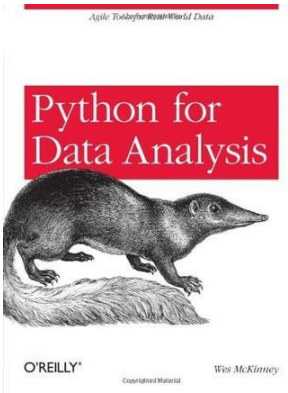
Trân trọng,

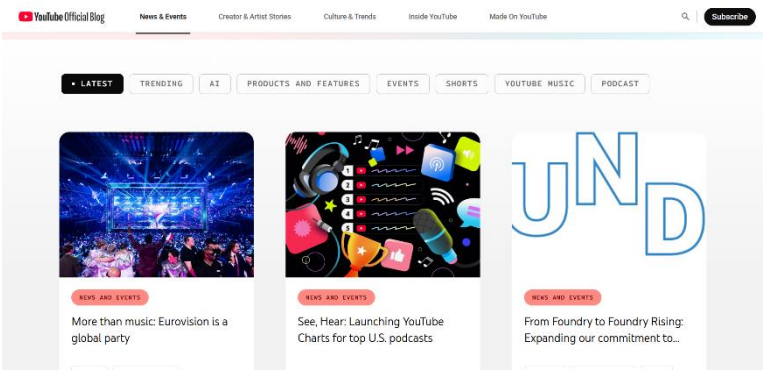
Nguyễn Trương Cao Sơn

30/05/2025

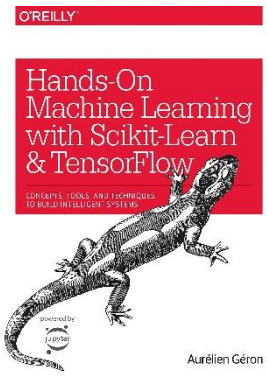
TÀI LIỆU THAM KHẢO

1. Wes McKinney (2018). *Python for Data Analysis*. O'Reilly Media.
2. Aurélien Géron (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.
3. Jake VanderPlas (2016). *Python Data Science Handbook*. O'Reilly Media.
4. YouTube Official Blog: *https://blog.youtube/news-and-events/*





5. Kaggle Datasets - YouTube Trending Videos: *https://www.kaggle.com/datasets*
6. Scikit-learn Documentation: *https://scikit-learn.org/*
7. Matplotlib & Seaborn Documentation: *https://matplotlib.org/* & *https://seaborn.pydata.org/*



8. Statista - YouTube usage statistics: *https://www.statista.com/topics/2019/youtube/*
9. Google Developers - YouTube Data API: *https://developers.google.com/youtube/v3*
10. Medium Articles on Data Science & YouTube Analysis: *https://medium.com/tag/youtube*