ĐẠI HỌC ĐÀ NẴNG
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA CÔNG NGHỆ THÔNG TIN

TIỂU LUẬN CUỐI KỲ
HỌC PHẦN: KHOA HỌC DỮ LIỆU

ĐỀ TÀI:
"Dự đoán lượt xem video trên youtube"

GIẢNG VIÊN HƯỚNG DẪN: TS. Ninh Khánh Duy

Nhóm	4
Họ và tên sinh viên	Lớp HP
Hồ Như Phong	20.N12
Nguyễn Thị Thu Thuyên	
Nguyễn Hoàng Ngọc	

ĐÀ NẴNG, 06/2023

TÓM TẮT

Đề tài của tiểu luận tập trung vào việc dự đoán lượt xem video trên Youtube, một vấn đề quan trọng trong việc tối ưu hóa nội dung và thu hút lượng người xem đông đảo. Để giải quyết vấn đề này, nhóm đã sử dụng phương pháp học máy và xây dựng mô hình dự đoán lượt xem dựa trên các thông tin về video như tiêu đề, mô tả, thời lượng video, lượt đăng kí kênh, chủ đề của video.

Phương pháp giải quyết vấn đề của nhóm bao gồm các bước: thu thập dữ liệu về video trên Youtube, tiền xử lý dữ liệu và chọn lọc các thuộc tính quan trọng, xây dựng các mô hình dự đoán lượt xem bằng phương pháp Machinee Learning (Linear Regression, Random Forest và Support Vector Regressor) sau đó đánh giá độ chính xác của mô hình trên tập dữ liệu kiểm tra bằng các độ đo như MSE (Mean Squared Error), RMSE (Root Mean Squared Error) và R2 Score.

Kết quả đạt được là một mô hình Machine Learning có độ chính xác tương đối xao trong việc dự đoán lượt xem video trên Youtube. Điều này giúp cho các nhà sáng tạo nội dung, quảng cáo trên Youtube hoặc các công ty quảng cáo có thể đưa ra các quyết định hiệu quả hơn trrogn việc tiếp cận khách hàng hoặc tăng lượt xem video trên nền tẳng Youtube. Bên cạnh đó, đề tài này cũng mở ra những hướng nghiên cứu mới trong việc phát triển các mô hình Machine Learning để dự đoán các chỉ số trên các nền tảng mạng xã hội khác nhau.

BẢNG PHÂN CÔNG NHIỆM VỤ

Sinh viên thực hiện	Các nhiệm vụ	Tự đánh giá
Hồ Như Phong	- Crawl dữ liệu - Phân tích dữ liệu - Tiền xử lý dữ liệu - Train model linear regressor - Đánh giá mô hình - Làm báo cáo, slide thuyết trình	Đã hoàn thành
Nguyễn Hoàng Ngọc	- Crawl dữ liệu - Trực quan hóa dữ liệu - Tiền xử lý dữ liệu - Train model SVR - Đánh giá mô hình - Làm báo cáo, slide thuyết trình	Đã hoàn thành
Nguyễn Thị Thu Thuyên	- Phân tích dữ liệu - Tiền xử lý dữ liệu - Train model Random Forest - Đánh giá mô hình - Trực quan hóa kết quả - Làm báo cáo, slide thuyết trình	Đã hoàn thành

MỤC LỤC

1. Giới thiệu

Trong đề tài dự đoán lượt xem trên Youtube, ta cần xây dựng mô hình để dự đoán chính xác số lượt xem của một video dựa trên các thông tin liên quan như: tiêu đề, mô tả, thời lượng video, lượt đăng ký kênh, chủ đề của video và các yếu tố khác.

Để giải quyết vấn đề, nhóm đề xuất sử dụng các mô hình Machine Learning để dự đoán lượt xem của video trên Youtube. Quá trình giải quyết bao gồm các bước:

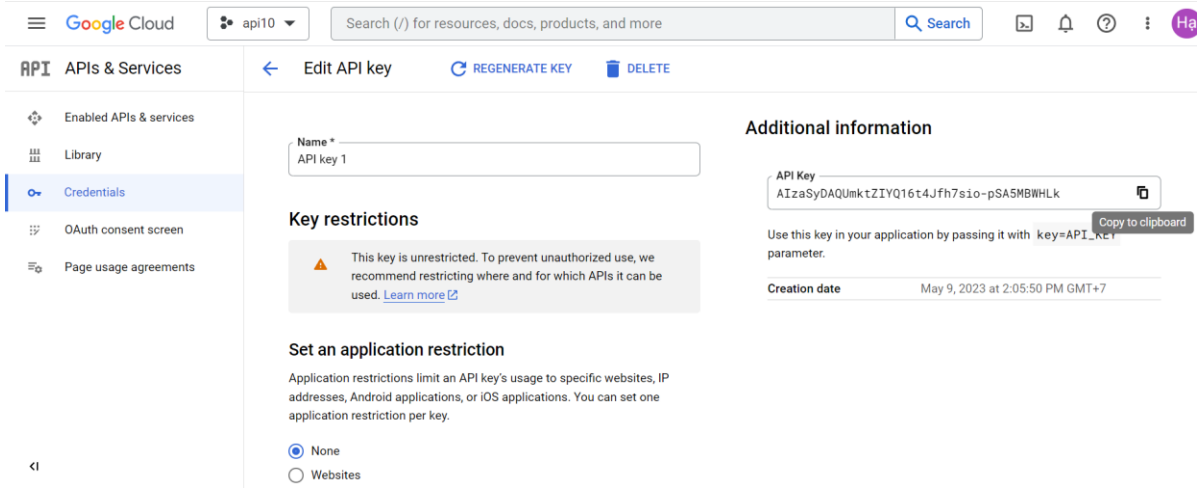
- Thu thập dữ liệu: Thu thập một tập dữ liệu với các thông tin về video.
- Tiền xử lý dữ liệu: Tiến hành xử lý và làm sạch dữ liệu để loại bỏ, xử lý các dữ liệu thiếu và chuyển đổi dữ liệu về dạng phù hợp cho quá trình huấn luyện mô hình.
- Xây dựng mô hình Machine Learning: chọn và xây dựng mô hình phù hợp Linear Regressor, Random Forest và Support Vector Regressor. Đưa các đặc trưng quan trọng vào mô hình và điều chỉnh các siêu tham số để đạt hiệu suất tốt nhất trên tập dữ liệu huấn luyện.
- Huấn luyện và đánh giá mô hình: Huấn luyện mô hình trên tập dữ liệu huấn luyện và đánh giá hiệu suất của mô hình trên tập dữ liệu kiểm tra để đánh giá khả năng dự đoán lượt xem.
- Dự đoán lượt xem: Áp dụng mô hình đã huấn luyện để dự đoán số lượt xem cho các video mới dựa trên các thông tin liên quan

2. Thu thập và mô tả dữ liệu

2.1. Thu thập dữ liệu

Sử dụng dữ liệu từ youtube: https://www.youtube.com/

Để thu thập dữ liệu từ Youtube, có thể sử dụng API Youtube Data v3 để truy xuất thông tin về video, kênh và các bình luận trên nền tảng này. Để sử dụng API Youtube Data v3 cần đăng ký và tạo một API key. API key này là một mã xác thực chho phép truy cập vào API Youtube Data v3 cho phép truy cập dễ dàng vào cấc tài nguyên trên Youtube và lấy các dữ liệu liên quan, tuy nhiên số lượng yêu cầu bị giới hạn.



Công cụ thu thập dữ liệu được xây dựng bằng ngôn ngữ lập trình Python, và các thư viện như googleapiclient và requests. Thư viện googleapiclient cung cấp các phương thức để gửi các yêu cầu tới API và lấy kết quả về. Thư viện requests được sử dụng để gửi các yêu cầu HTTP đến API.

Quá trình thu thập dữ liệu bắt đầu bằng việc xác định đầu vào cho quá trình. Là các từ khóa tìm kiếm liên quan đến đề tài cần nghiên cứu hoặc danh sách các channel trên Youtube. Dựa vào đó có thể sử dụng API để truy cập thông tin về các video liên quan.

Khi đã thu thập dữ liệu, nó được lưu trữ dưới dạng CSV để dễ dàng xử lý và phân tích. Mỗi dòng là một mẫu dữ liệu chứa các thông tin về video .

2.2. Mô tả dữ liệu

- Đối với SmallDS: dữ liệu gồm 1057 dòng, 14 cột
- Đối với BigDS: dữ liệu gồm 9986 dòng, 14 cột
- 2 tập dữ liệu chỉ khác nhau về số lượng mẫu

2.2.1 Dữ liệu trước khi xử lý:

- Dữ liệu gồm 14 cột chứa các thông tin:
  - video_id: id của video
  - channel_id: id của kênh
  - channel_name: tên kênh
  - published_date: ngày đăng video
  - video_title: tên video
  - video_description: mô tả video
  - dislikes: lượt không thích
  - likes: lượt thích
  - views: lượt xem
  - comment_count:số lượt bình luận
  - favorite_count:số lượt yêu thích
  - category: danh mục của video
  - subscribers: số người đăng kí kênh
  - duration: độ dài thời gian video
- Video trùng nhau:

  SmallDS: 34

  Big DS: 229
- Thông tin về dữ liệu:

```
RangeIndex: 1057 entries, 0 to 1056
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   video_id           1057 non-null   object
 1   channel_id         1057 non-null   object
 2   channel_name       1057 non-null   object
 3   published_date     1057 non-null   object
 4   video_title        1057 non-null   object
 5   video_description  899 non-null    object
 6   dislikes           1057 non-null   int64
 7   likes              1057 non-null   int64
 8   views              1057 non-null   int64
 9   comment_count      1057 non-null   int64
 10  favorite_count     1057 non-null   int64
 11  category           1057 non-null   int64
 12  subscribers        1057 non-null   int64
 13  duration           1057 non-null   object
dtypes: int64(7), object(7)
memory usage: 115.7+ KB
```

2.2.2 Dữ liệu sau khi xử lý:

- Sau khi xử lý dữ liệu sẽ có 1023 dòng và 12 cột:

- video_id: id của video
- channel_id: id của kênh
- channel_name: tên kênh
- published_date: ngày đăng video
- video_title: tên video
- video_description: mô tả video
- likes: lượt thích
- views: lượt xem
- comment_count:số lượt bình luận
- category: danh mục của video
- subscribers: số người đăng kí kênh
- duration: độ dài thời gian video

```
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   video_id           1023 non-null   object
 1   channel_id         1023 non-null   object
 2   channel_name       1023 non-null   object
 3   published_date     1023 non-null   datetime64[ns, UTC]
 4   video_title        1023 non-null   object
 5   video_description  869 non-null    object
 6   likes              1023 non-null   int64
 7   views              1023 non-null   int64
 8   comment_count      1023 non-null   int64
 9   category           1023 non-null   int64
 10  subscribers        1023 non-null   int64
 11  duration           1023 non-null   int64
dtypes: datetime64[ns, UTC](1), int64(6), object(5)
memory usage: 136.2+ KB
```

Dữ liệu từ file .csv

	video_id	channel_id	channel_name	published_date	video_title	video_descript	likes	views	comment_	category	subscriber	duration
2	iLBYhZcLPrI	UCXc-GBhPcGrzwkZt67h4suQ	200Lab Education	2021-08-02 13:00:18+00:00	Flutter Tutorial 2021 for Beginners # Trong video		94	9845	15	27	3960	332
3	Duv6hhEMFDw	UCXc-GBhPcGrzwkZt67h4suQ	200Lab Education	2022-08-11 13:00:14+00:00	CÃ¡c bÆ°á»c phÃ¢n tÃch dá»¯ liá»‡u Nhá»¯ng		6	635	1	27	3960	184
4	W1bAvH_PFWQ	UCXc-GBhPcGrzwkZt67h4suQ	200Lab Education	2023-02-22 12:15:02+00:00	Top 4 cÃ¯ng cá»¥ vá»½ dashboard mÃ bá°¡n khÃ´		11	397	1	27	3960	16
5	gqvg7CQkncs	UCXc-GBhPcGrzwkZt67h4suQ	200Lab Education	2021-08-31 13:00:21+00:00	Flutter Tutorial 2021 for Beginners # Náº¿u bá°¡n		9	1395	3	27	3960	542
6	mOU5HFUwS2I	UCXc-GBhPcGrzwkZt67h4suQ	200Lab Education	2023-02-18 04:45:01+00:00	Giá»¯a lÃ n sÃ³ng â€œSA THá°CI" lÃ Ã¡»f nÃ©ng		13	783	0	27	3960	57
7	bf-cGopy-a0	UC4qExVK6hxFaPa2JYfQOCdQ	Glory Education - TS.Trá°§n HoÃ ng Há°£i	2021-04-09 11:25:20+00:00	#2.2 [BÃ€I Tá°¬P] KINH Tá°¾ CHÃN #2.2 [BÃ€I		366	34721	20	27	39600	223
8	oiztBU2NiZo	UC4qExVK6hxFaPa2JYfQOCdQ	Glory Education - TS.Trá°§n HoÃ ng Há°£i	2021-01-01 23:42:20+00:00	KINH Tá°¾ CHÃNH TRá»Š MÃC LÃŠ KINH Tá°¾		852	64550	86	27	39600	533
9	vnu4KKFPDKM	UC4qExVK6hxFaPa2JYfQOCdQ	Glory Education - TS.Trá°§n HoÃ ng Há°£i	2022-11-24 12:00:44+00:00	Tá°¿t HÃ n thá»±c lÃ gÃ¬? #shorts Tá°¿t HÃ n		3	278	0	27	39600	58
10	Qnr2EWcfPD0	UC4qExVK6hxFaPa2JYfQOCdQ	Glory Education - TS.Trá°§n HoÃ ng Há°£i	2022-12-15 13:42:37+00:00	VÃ¬ sao ngÆ°á»i dÃ¢n PhÆ°Æ¡ng TÃ¢y VÃ¬ sao		34	1831	1	27	39600	60
11	0N5k48DKIHM	UCE_dkCIlyCsBu1YH2MM9iZQ	Zila Education Official	2022-10-21 09:00:00+00:00	DU Há»ŒC HÃ€N QUá»C NGÃ€NH Du há»c HÃ n		6	307	2	27	1460	146
12	vm8LKwyuLCo	UCE_dkCIlyCsBu1YH2MM9iZQ	Zila Education Official	2022-11-07 05:00:27+00:00	Buá»•i 5: Ã´n táºp TrÆ°á»ng há»c (P1 Buá»•i 5:		11	380	0	27	1460	2017
13	4zYfww1srgk	UCE_dkCIlyCsBu1YH2MM9iZQ	Zila Education Official	2022-01-18 12:45:00+00:00	Lá»™ trÃ¬nh 30 ngÃ y cá°£i thiá»‡n Giao tiáº¿p		17	370	12	27	1460	460
14	#NAME?	UCE_dkCIlyCsBu1YH2MM9iZQ	Zila Education Official	2022-05-21 12:30:00+00:00	TOPIK lÃ gÃ¬? CÃ¡ch luyá»‡n thi TO TOPIK lÃ		15	1027	0	27	1460	530
15	7D0qdtRr6mg	UCE_dkCIlyCsBu1YH2MM9iZQ	Zila Education Official	2022-10-24 09:00:06+00:00	Ãlá»€U Kiá»†N Sá»¯C KHá»ŽE Ãi DL Má»™t trong		5	203	0	27	1460	136
16	YZpzbyNw_sQ	UC7BBBPge1-mYoh0KL4zI6YQ	NhÃ¢n Viá»‡t Media & Education	2023-02-15 15:55:10+00:00	TÃ©t Reset Tiá»n Vá» 0 Facebook Vì TÃ©t Reset		143	13518	40	22	15800	161
17	TEn5EIPsQUU	UC7BBBPge1-mYoh0KL4zI6YQ	NhÃ¢n Viá»‡t Media & Education	2022-11-18 14:10:06+00:00	Link KhÃ¡ng 273 Má»»i KhÃ¡ng Treo Link KhÃ¡ng		146	15412	55	22	15800	87
18	HofE-rJWfbs	UC7BBBPge1-mYoh0KL4zI6YQ	NhÃ¢n Viá»‡t Media & Education	2022-06-11 12:00:18+00:00	CÃ¡ch ÃÃ¡ Thá°» Ra Khá»i BM Die K CÃ¡ch ÃÃ¡		122	9954	29	22	15800	194
19	pddXmd8r9e0	UC7BBBPge1-mYoh0KL4zI6YQ	NhÃ¢n Viá»‡t Media & Education	2022-06-15 15:43:25+00:00	Paid Náº£ LÃ GÃ¬? - Má»™t ChÃ©t Paid Náº£		255	15152	82	22	15800	782
20	04EOkZTt5Zc	UC7BBBPge1-mYoh0KL4zI6YQ	NhÃ¢n Viá»‡t Media & Education	2022-12-26 07:40:07+00:00	Via Die PTTT Má»»i VÃ Set Camp Q	Via Die PTTT	42	1853	19	22	15800	61
21	cJ4xlKeDQWw	UCSLCRI9oriS5EnS4x0icBHA	DOWNTOWN Education	2023-02-28 20:19:31+00:00	DownTown		225	3358	13	27	19200	61

3. Trích xuất đặc trưng

3.1 Làm sạch dữ liệu

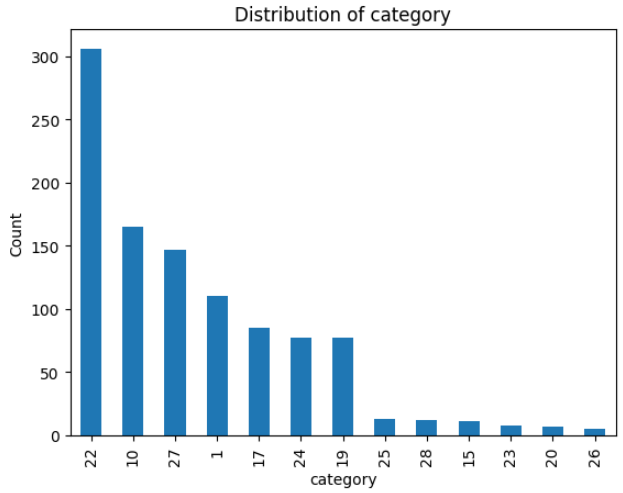
- Loại bỏ các cột không có dữ liệu (dislike, favorite_count)
- Chuyển cột "published_date" thành kiểu dữ liệu datetime.
- Chuyển cột "duration" thành kiểu dữ liệu số (int64)

3.2 Thêm đặc trưng mới

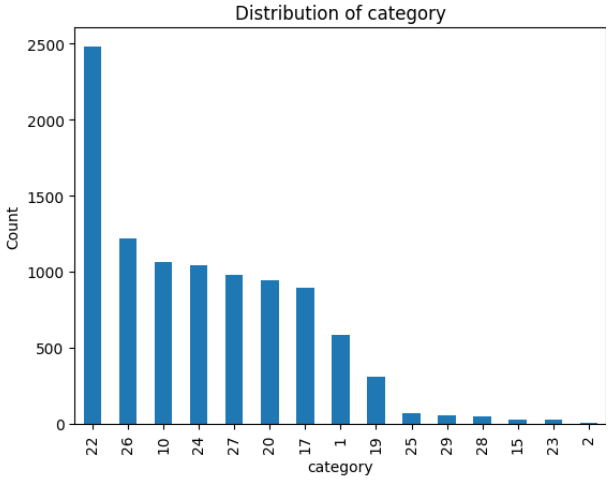
- Tạo đặc trưng số lượng từ trong tiêu đề và mô tả.
- Tạo đặc trưng thời gian từ lúc video được đăng.

3.3 Phân tích dữ liệu

- Số lượng video dựa vào danh mục



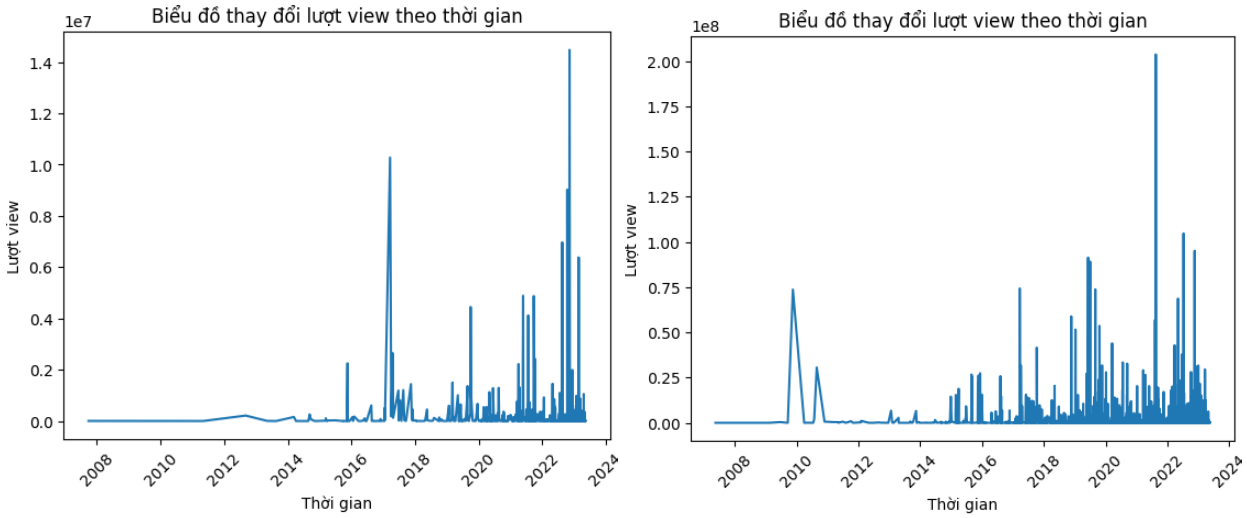
SmallDS



BigDS

Nhận xét: Đối với các danh mục khác nhau thì số lượng video đăng tải cũng khác nhau

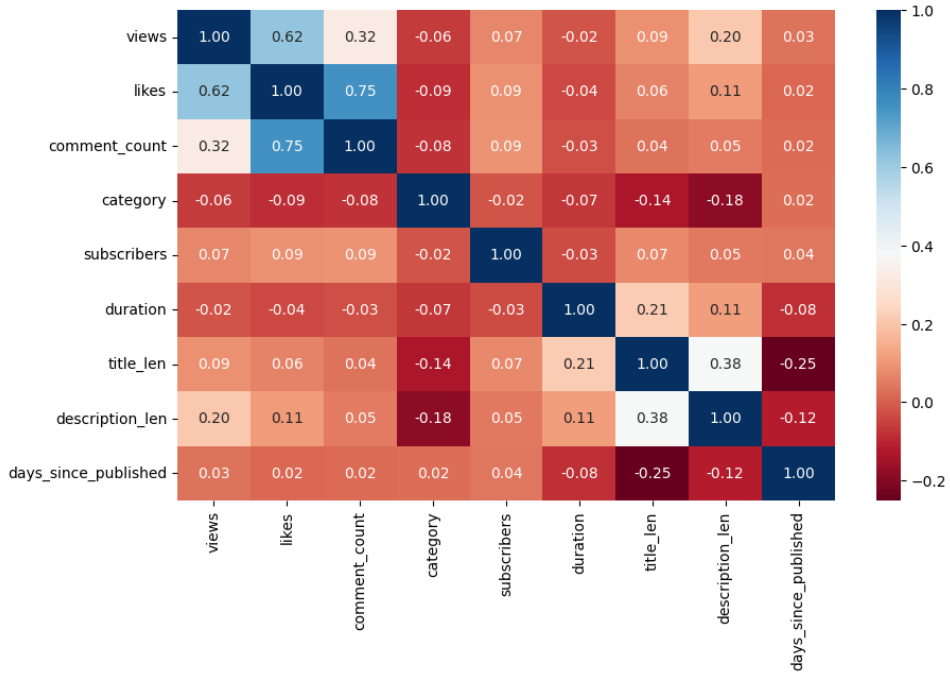
- Biểu đồ thay đổi views theo thời gian



SmallDS BigDS

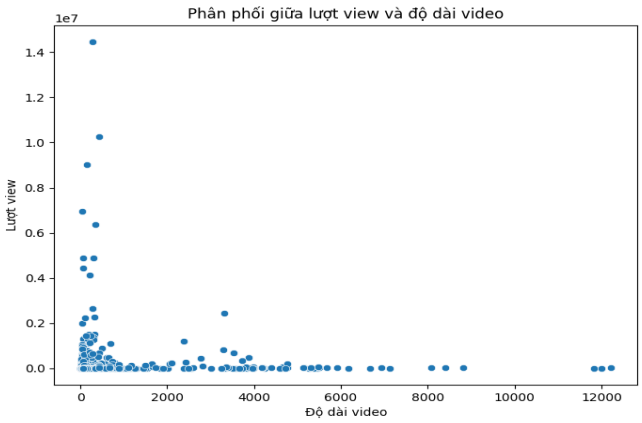
Nhận xét: đối với những video được đăng trong những năm gần đây thì lượt xem cao hơn.

- Độ tương quan giữa các đặc trưng của dữ liệu:

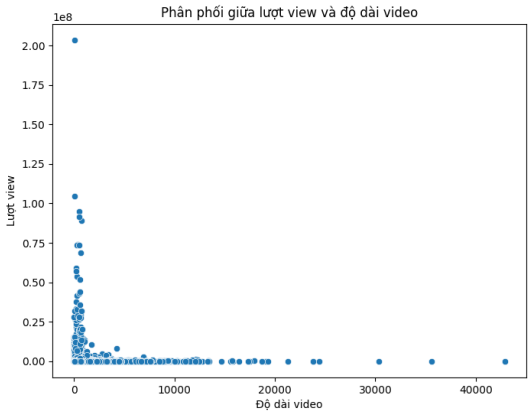


Nhận xét: Dựa vào đồ thị ta thấy độ tương quan của các biến đối với biến "views" là trung bình. Đa số là mối quna hệ tương quan dương (khi biến "views" tăng thì biến "category","subscribers","duration","title_len","description_len","days_since_published " cũng tăng).

- Độ tương quan giữa views và duration



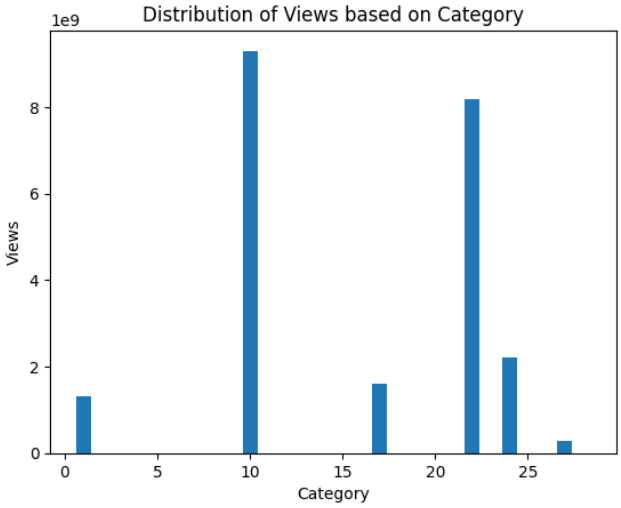
SmallDS



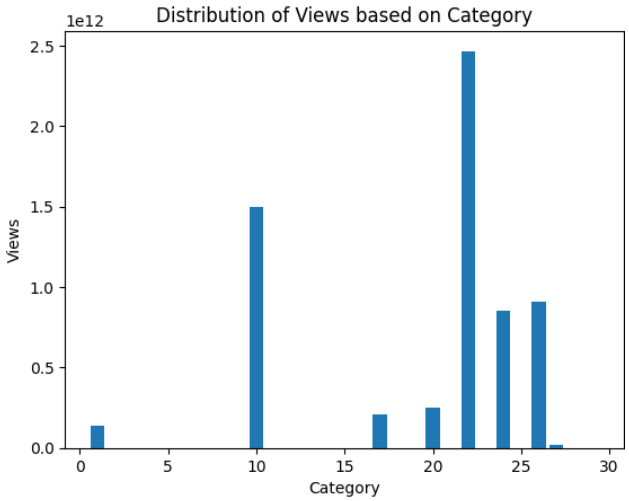
BigDS

Nhận xét: Video có thời gian ngắn thường có nhiều view cao hơn những video có độ dài lớn.

- Độ tương quan giữa views và category:



SmallDS



BigDS

Nhận xét: Danh mục video cũng ảnh hưởng đến lượt xem của video đó.

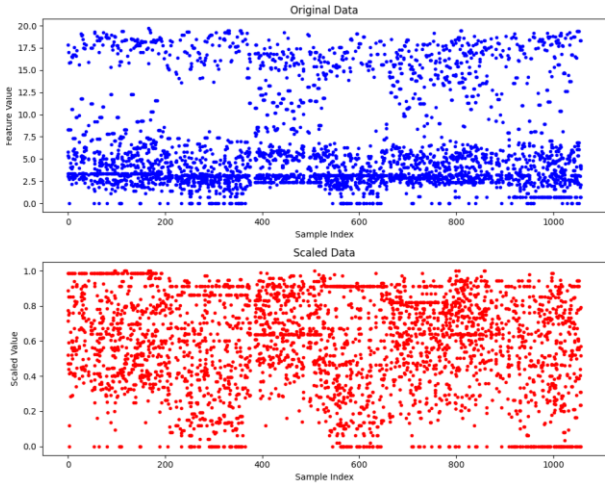
3.4 Lựa chọn đặc trưng

- Sử dụng phương pháp SelectKBest và Random Forest để lấy ra 5 đặc trưng có tương quan mạnh.
- Với SelectKbest thì 5 đặc trưng là: 'category', 'subscribers', 'duration', 'title_len', 'description_len'
- Với Random Forest thì 5 đặc trưng là: 'subscribers', 'days_since_published', 'duration', 'description_len', 'title_len'

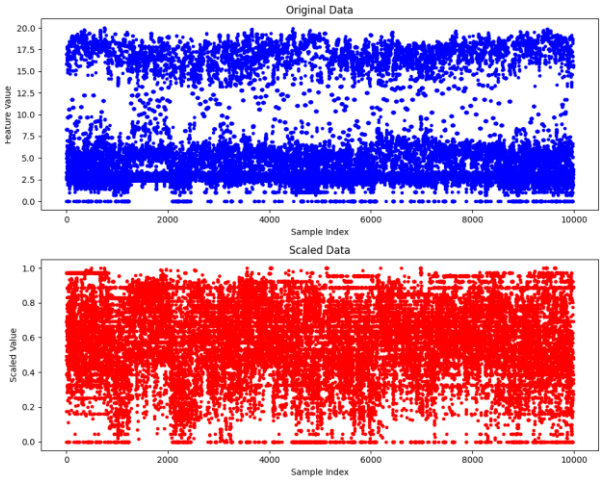
Nhận xét: Do cách tiếp cận và phương pháp tính toán khác nhau, SelectKBest và Random Forest có thể cho kết quả khác nhau khi chọn đặc trưng.

3.5. Xử lý kiểu dữ liệu và chuẩn hóa

- Xử lý kiểu dữ liệu: Sử dụng phương pháp chuyển đổi dữ liệu sang logarithm (log) khi dữ liệu có phân phối lệch. Lý do chính để thực hiện chuyển đổi log là để đạt được phân phối dữ liệu gần với phân phối chuẩn, giúp cải thiện tính chuẩn xác và đồng nhất của dữ liệu. Điều này giúp giảm thiểu sự biến động của dữ liệu, làm cho phân bố dữ liệu đồng nhất hơn và thuận lợi cho việc áp dụng các phương pháp và mô hình học máy tiếp theo.
- Xử lý ngoại lệ: sau khi biến đổi dữ liệu về dạng phân phối chuẩn, áp dụng xử lý ngoại lệ Gaussion cho dữ liệu.
- Chuẩn hóa dữ liệu: sử dụng phương pháp Min-Max Scaler để chuẩn hóa dữ liệu. Phương pháp này hoạt động bằng cách chuyển đổi dữ liệu sao cho nằm trong khoảng [0, 1].
- Lợi ích của việc sử dụng Min-Max Scaler là:
  - Chuẩn hóa dữ liệu về cùng một khoảng giá trị, giúp dễ dàng so sánh và hiểu được quan hệ giữa các đặc trưng.
  - Loại bỏ sự ảnh hưởng của đơn vị đo lường trong dữ liệu ban đầu, giúp mô hình học máy hội tụ nhanh hơn.
  - Giảm thiểu tác động của ngoại lệ (outliers) lên quá trình huấn luyện mô hình, bằng cách đưa các giá trị cận biên vào khoảng giá trị chuẩn hóa.
- Trực quan hóa dữ liệu trước và sau khi chuẩn hóa:



SmallDS



BigDS

Nhận xét:

- Biểu đồ trước khi chuẩn hóa dữ liệu cho thấy các đặc trưng có phạm vi giá trị khác nhau và không đồng nhất

- Sau khi chuẩn hóa, biểu đồ cho thấy các đặc trưng đã được đưa về cùng một phạm vi [0, 1]. Điều này cho thấy các giá trị của các dữ liệu nằm gần nhau và không bị chênh lệch lớn về giá trị.
- Điều này cho thấy việc chuẩn hóa đã giúp đồng nhất dữ liệu và đưa các đặc trưng về cùng một thang đo.

4. Mô hình hóa dữ liệu

4.1 Model Random Forest

4.1.1 Cơ sở lí thuyết

Random Forest là một thuật toán học máy dựa trên việc kết hợp nhiều cây quyết định (decision tree) để tạo ra một mô hình dự đoán mạnh mẽ và ổn định. Mỗi cây quyết định được huấn luyện trên một tập con của dữ liệu và sử dụng một phương pháp chọn đặc trưng ngẫu nhiên để tạo ra sự đa dạng giữa các cây. Kết quả dự đoán cuối cùng của mô hình Random Forest được tính bằng cách lấy trung bình hoặc phương án đa số của dự đoán từ các cây quyết định.

4.1.2 Huấn luyện mô hình

Dữ liệu được chia thành 3 tập (train, validation, test) với tỉ lệ là 70/15/15

Trong quá trình huấn luyện mô hình Random Forest:

- Tập dữ liệu huấn luyện (train): được sử dụng để xây dựng mô hình Random Forest. Trong quá trình huấn luyện, mỗi cây quyết định trong mô hình sẽ được huấn luyện trên một tập con của dữ liệu huấn luyện. Sự đa dạng giữa các cây được đảm bảo thông qua việc sử dụng một số đặc trưng ngẫu nhiên trong quá trình xây dựng mỗi cây. Tập dữ liệu huấn luyện giúp mô hình học các quy tắc và mẫu từ dữ liệu để dự đoán kết quả.
- Tập dữ liệu validation (validate): được sử dụng để đánh giá hiệu suất của mô hình và tinh chỉnh các siêu tham số.

Kết quả huấn luyện

Tập dữ liệu	Tham số	Kỹ thuật SelectKBest	Kỹ thuật Random Forest
SmallDS	MAE	1.51	1.41
	RMSE	1.95	1.82
	R2 score	0.68	0.72
BigDS	MAE	1.10	1.00
	RMSE	1.49	1.37
	R2 score	0.80	0.83

Nhận xét: Dựa trên các kết quả trên, chúng ta có thể kết luận rằng kỹ thuật Random Forest cho kết quả tốt hơn so với kỹ thuật SelectKBest trong việc dự đoán và giải thích biến mục tiêu cho cả 2 tập dữ liệu.

4.1.4 Hiệu chỉnh tham số

Phần hiệu chỉnh tham số được thực hiện bằng cách sử dụng lớp GridSearchCV trong thư viện sklearn.model_selection. Mục đích của việc hiệu chỉnh tham số là tìm ra các giá trị tối ưu cho các siêu tham số của mô hình Random Forest.

Các tham số quan trọng cần xem xét trong hiệu chỉnh là 'n_estimators': số lượng cây quyết định, 'max_depth': độ sâu tối đa, 'min_samples_split': số lượng mẫu tối thiểu cần thiết để tiếp tục phân chia một nút, 'min_samples_leaf': số lượng mẫu tối thiểu cần thiết trong mỗi lá cây và 'max_features': số lượng đặc trưng.

Quá trình tìm kiếm siêu tham số trên mô hình Random Forest có thời gian huấn luyện tương đối lâu, khoảng 4 phút 30 giây.

Kết quả của trên tập val:

Tham số	SmallDS	BigDS
'max_depth'	10	None
'max_features'	'log2'	'sqrt'
'min_samples_leaf'	1	1
'min_samples_split'	2	2
'n_estimators'	300	300
MAE	1.46	1.01
RMSE	1.83	1.38
R2 score	0.72	0.83

Nhận xét: từ bảng trên này cho thấy mô hình có khả năng dự đoán gần đúng giá trị thực tế trên tập validation. Độ chính xác của tập BigDS có khả năng dự đoán cao hơn.

4.1.5 Đánh giá kết quả

Tiến hành dự đoán trên tập test và trực quan hóa kết quả:

Tham số	SmallDS	BigDS
MAE	1.38	1.04
RMSE	1.74	1.44
R2 score	0.75	0.80

Nhận xét:

- Đây là kết quả tốt hơn so với trên tập validation, cho thấy mô hình có khả năng dự đoán chính xác và tốt hơn trên dữ liệu mới
  - MAE (Mean Absolute Error): Giá trị này thấp hơn so với kết quả trước đó, cho thấy mô hình Random Forest đạt được sự chính xác cao hơn trong việc dự đoán giá trị.
  - MSE (Mean Squared Error): Giá trị này cũng thấp hơn so với kết quả trước, cho thấy mô hình có khả năng dự đoán gần giá trị thực tế hơn.
  - R2 score: Giá trị R2 score tăng lên so với kết quả trước, cho thấy mô hình giải thích được khoảng 75% cho tập SmallDS và 80% cho tập BigDS sự biến thiên của dữ liệu, tức là có khả năng dự đoán tốt hơn.
- Tổng thể, mô hình Random Forest đã cho kết quả khá tốt trên cả tập validation và tập test. Điều này cho thấy mô hình có khả năng học và tổng quát hóa tốt trên dữ liệu mới, giúp dự đoán giá trị mục tiêu một cách chính xác và hiệu quả.

4.2 Model Linear Regression

4.2.1 Cơ sở lí thuyết

- Linear Regression là một mô hình học máy thuộc họ mô hình hồi quy. Nó xác định một mối quan hệ tuyến tính giữa các biến đầu vào (đặc trưng) và biến mục tiêu.
- Mô hình Linear Regression giả định rằng mối quan hệ giữa các biến đầu vào và biến mục tiêu có thể được biểu diễn bằng một đường thẳng trong không gian đặc trưng.
- Mục tiêu của Linear Regression là tìm ra các hệ số hợp lý để đường thẳng tạo ra phù hợp nhất với dữ liệu huấn luyện.

4.2.2 Huấn luyện mô hình

Dữ liệu được chia thành 3 tập (train, validation, test) với tỉ lệ là 70/15/15

Trong quá trình huấn luyện mô hình Linear Regression:

- Tập dữ liệu huấn luyện (train) được sử dụng để xây dựng và huấn luyện mô hình Linear Regression. Mô hình sẽ học từ các mẫu dữ liệu trong tập train để tìm ra các tham số tối ưu cho đường thẳng tương ứng với mô hình. Mục tiêu là tìm cách mô hình có thể tìm hiểu các quy luật và mẫu trong tập dữ liệu huấn luyện để dự đoán mục tiêu một cách tốt nhất.
- Tập dữ liệu validation (validate) được sử dụng để đánh giá hiệu suất của mô hình và tinh chỉnh các siêu tham số

4.2.3 Kết quả huấn luyện

Tập dữ liệu	Tham số	Kỹ thuật SelectKBest	Kỹ thuật Random Forest
SmallDS	MAE	1.63	1.60
	RMSE	2.06	1.99

	R2 score	0.64	0.67
BigDS	MAE	1.48	1.46
	RMSE	1.88	1.86
	R2 score	0.69	0.70

Nhận xét: Dựa trên các kết quả trên, chúng ta có thể kết luận rằng kỹ thuật Random Forest cho kết quả tốt hơn so với kỹ thuật SelectKBest trong việc dự đoán và giải thích biến mục tiêu. Từ đó chọn kĩ thuật lựa chọn đặc trưng cho mô hình này.

4.2.4 Hiệu chỉnh tham số

Trong mô hình Linear Regression, tham số fit_intercept quyết định liệu mô hình có tính giá trị chặn (intercept) hay không. Giá trị chặn là giá trị của y khi x = 0. Mặc định, tham số fit_intercept trong scikit-learn được đặt là True, tức là mô hình sẽ tính cả giá trị chặn.

Việc hiệu chỉnh tham số fit_intercept có thể được thực hiện bằng cách truyền giá trị True hoặc False vào tham số này khi khởi tạo mô hình Linear Regression trong scikit-learn

Kết quả của quá trình:

Tham số	SmallDS	BigDS
'fit_intercept'	True	True
MAE	1.60	1.46
RMSE	1.99	1.86
R2 score	0.67	0.70

Nhận xét: từ bảng trên này cho thấy mô hình có khả năng dự đoán gần đúng giá trị thực tế trên tập validation. Độ chính xác của tập BigDS có khả năng dự đoán cao hơn.

4.2.5 Đánh giá kết quả

Tiến hành dự đoán trên tập test và trực quan hóa kết quả:

Tham số	SmallDS	BigDS
MAE	1.47	1.49
RMSE	1.86	1.91
R2 score	0.72	0.66

Nhận xét: Đây là kết quả tốt hơn đối với SmallDS và kém hơn đối với BigDS so với trên tập validation, cho thấy mô hình có khả năng dự đoán chính xác và tốt hơn trên dữ liệu mới. Tuy nhiên, cần tiếp tục đánh giá và kiểm tra hiệu suất của mô hình trên các tập dữ liệu khác để xác nhận tính khái quát và đáng tin cậy của mô hình trên dữ liệu mới.

- MAE (Mean Absolute Error): Giá trị này thấp hơn so với kết quả trước đó, cho thấy mô hình Random Forest đạt được sự chính xác cao hơn trong việc dự đoán giá trị.

- MSE (Mean Squared Error): Giá trị này cũng thấp hơn so với kết quả trước, cho thấy mô hình có khả năng dự đoán gần giá trị thực tế hơn.
- R2 score: giá trị R2 score tăng lên so với kết quả trước , cho thấy mô hình giải thích được sự biến thiên của dữ liệu, tức là có khả năng dự đoán tốt hơn.

Tổng thể, mô hình Random Forest đã cho kết quả khá tốt trên cả tập validation và tập test. Điều này cho thấy mô hình có khả năng học và tổng quát hóa tốt trên dữ liệu mới, giúp dự đoán giá trị mục tiêu một cách chính xác và hiệu quả.

4.3 Model Support Vector Regressor

4.3.1 Cơ sở lí thuyết

- Mô hình SVR (Support Vector Regression) là một phương pháp học máy được áp dụng vào bài toán hồi quy. Nó dựa trên lý thuyết của SVM (Support Vector Machines).
- Mục tiêu của SVR là xây dựng một đường tuyến tính hoặc phi tuyến tính để dự đoán giá trị đầu ra và tối đa hóa khoảng cách giữa các điểm dữ liệu và đường ranh giới.
- Mô hình SVR sử dụng hàm lỗi epsilon-insensitive loss để đánh giá sai số dự đoán và tạo độ linh hoạt cho mô hình.
- Phương pháp tối ưu hóa thường được sử dụng trong SVR là Sequential Minimal Optimization (SMO) hoặc hạ gradient.

4.3.2 Huấn luyện mô hình

Chia dữ liệu thành 3 tập: train (70%), validation (15%), test (15%).

Trong quá trình huấn luyện mô hình SVR:

- Tập dữ liệu huấn luyện (train): được sử dụng để xây dựng mô hình SVR. Trong quá trình huấn luyện, Mô hình SVR được tạo bằng cách tạo một đối tượng SVR từ sklearn.svm và gán cho một biến
- Tập dữ liệu validation (validate): được sử dụng để đánh giá hiệu suất của mô hình và tinh chỉnh các siêu tham số.
- Sau đó, mô hình được huấn luyện bằng cách sử dụng phương thức fit(X_train, y_train) trên đối tượng mô hình. Quá trình này tìm ra các tham số tối ưu để xây dựng siêu mặt phẳng tốt nhất

4.3.3 Đánh giá hiệu suất trên tập dữ liệu validation

Tập dữ liệu	Tham số	Kỹ thuật SelectKBest	Kỹ thuật Random Forest
SmallDS	MAE	1.61	1.57
	RMSE	2.04	1.94
	R2 score	0.65	0.68
BigDS	MAE	1.42	1.36
	RMSE	1.85	1.77

	R2 score	0.70	0.72

Nhận xét: Dựa trên các kết quả trên, chúng ta có thể kết luận rằng kỹ thuật lựa chọn đặc trưng của Random Forest cho kết quả huấn luyện mô hình SVR tốt hơn so với kỹ thuật lựa chọn đặc trưng của SelectKBest trong việc dự đoán và giải thích biến mục tiêu. Từ đó chọn kĩ thuật lựa chọn đặc trưng cho mô hình này.

4.3.4 Hiệu chỉnh tham số

Dựa trên kết quả đánh giá trên tập validation, điều chỉnh các siêu tham số của mô hình SVR để cải thiện hiệu suất

Phần hiệu chỉnh tham số được thực hiện bằng cách sử dụng lớp GridSearchCV trong thư viện sklearn.model_selection. Mục đích của việc hiệu chỉnh tham số là tìm ra các giá trị tối ưu cho các siêu tham số của mô hình Random Forest.

Kết quả hiệu chỉnh:

Tham số	SmallDS	BigDS
'C'	10	10
'epsilon'	0.2	0.3
'gamma'	'auto'	'scale'
'kernel'	'rbf'	'rbf'
MAE	1.57	1.33
RMSE	1.97	1.75
R2 score	0.67	0.73

Nhận xét: từ bảng trên này cho thấy mô hình có khả năng dự đoán gần đúng giá trị thực tế trên tập validation. Độ chính xác của tập BigDS có khả năng dự đoán cao hơn.

4.3.5 Đánh giá kết quả

Tiến hành dự đoán trên tập test và trực quan hóa kết quả:

Tham số	SmallDS	BigDS
MAE	1.45	1.33
RMSE	1.86	1.76
R2 score	0.72	0.71

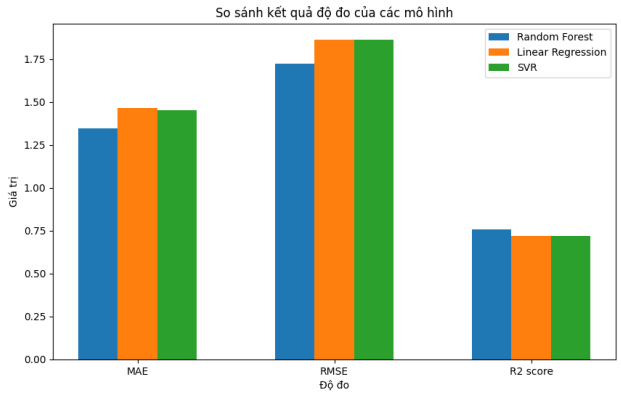
Nhận xét: Đây là kết quả tốt hơn đối với SmallDS và kém hơn đối với BigDS so với trên tập validation, cho thấy mô hình có khả năng dự đoán chính xác và tốt hơn trên dữ liệu mới. Tuy nhiên, cần tiếp tục đánh giá và kiểm tra hiệu suất của mô hình trên các tập dữ liệu khác để xác nhận tính khái quát và đáng tin cậy của mô hình trên dữ liệu mới.

- MAE (Mean Absolute Error): Giá trị này thấp hơn so với kết quả trước đó, cho thấy mô hình Random Forest đạt được sự chính xác cao hơn trong việc dự đoán giá trị.
- MSE (Mean Squared Error): Giá trị này cũng thấp hơn so với kết quả trước, cho thấy mô hình có khả năng dự đoán gần giá trị thực tế hơn.
- R2 score: Giá trị R2 score tăng lên so với kết quả trước, cho thấy mô hình giải thích được sự biến thiên của dữ liệu, tức là có khả năng dự đoán tốt hơn.

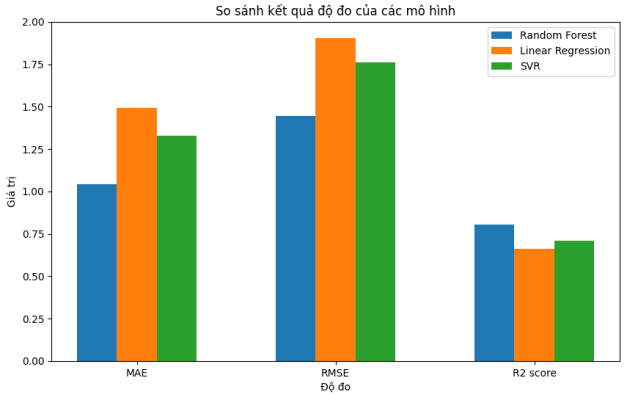
Tổng thể, mô hình Random Forest đã cho kết quả khá tốt trên cả tập validation và tập test. Điều này cho thấy mô hình có khả năng học và tổng quát hóa tốt trên dữ liệu mới, giúp dự đoán giá trị mục tiêu một cách chính xác và hiệu quả.

4.4 Đánh giá hiệu suất các mô hình

Từ các kết quả trên ta vẽ được biểu đồ cột so sánh kết quả của 3 mô hình:



SmallDS



BigDS

Kết luận:

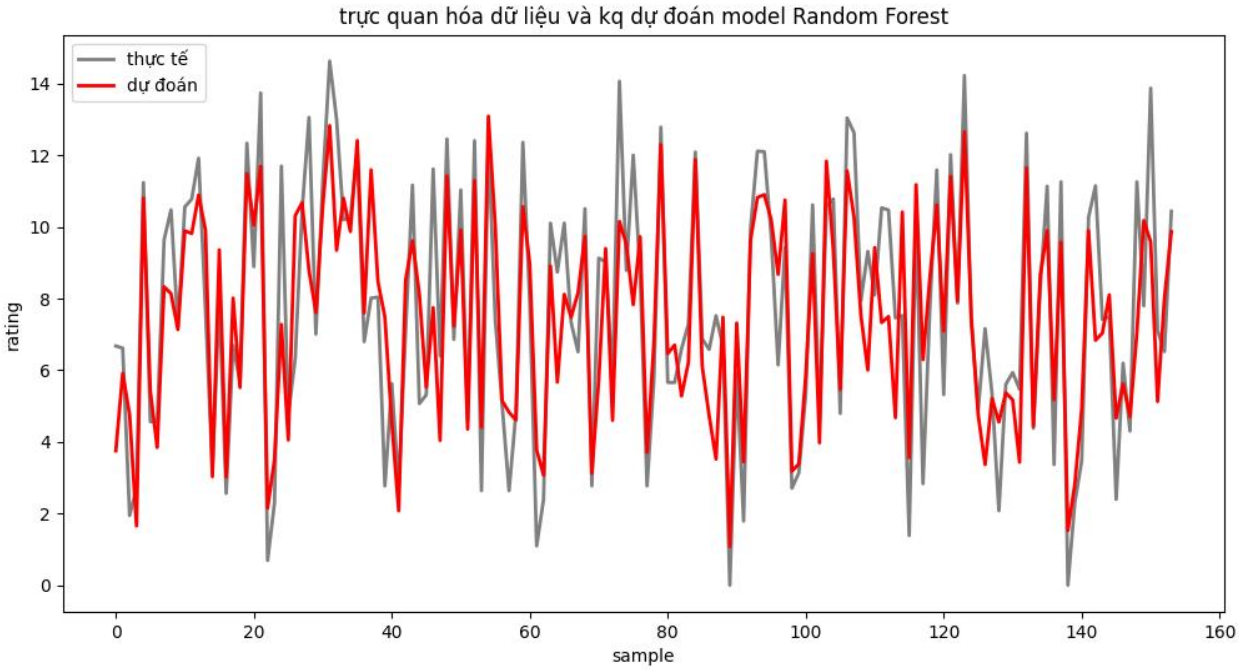
Dựa vào cả 2 biểu đồ và kết quả độ đo, có thể kết luận rằng mô hình Random Forest cho kết quả tốt nhất trong số ba mô hình được so sánh (Random Forest, Linear Regression, SVR).

- Mô hình Random Forest có giá trị MAE thấp hơn so với hai mô hình khác, cho thấy sự chính xác cao hơn trong việc dự đoán giá trị.
- Mô hình Random Forest cũng có giá trị RMSE thấp hơn, cho thấy khả năng dự đoán gần giá trị thực tế hơn.
- Giá trị R2 score của mô hình Random Forest cũng cao hơn, cho thấy mô hình giải thích được sự biến thiên của dữ liệu tốt hơn.

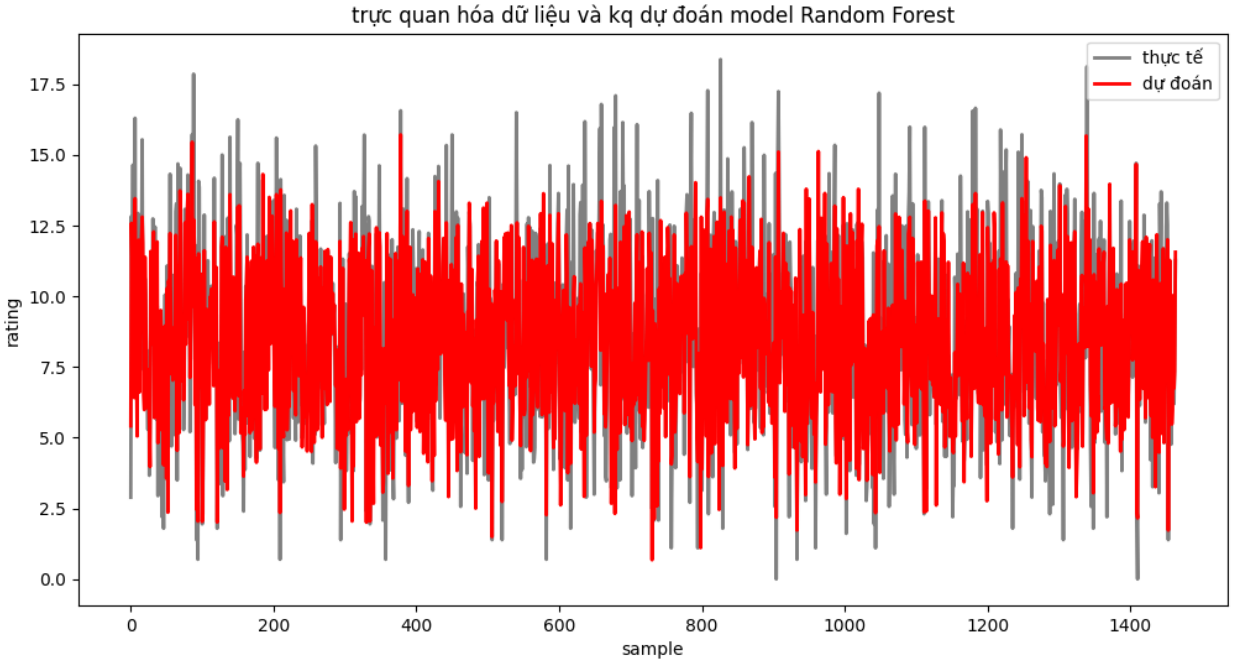
Tổng quan, mô hình Random Forest đạt được hiệu suất tốt hơn so với hai mô hình khác trong việc dự đoán giá trị và giải thích sự biến thiên của dữ liệu.

Tham số	SmallDS	BigDS
MAE	1.38	1.04

RMSE	1.74	1.44
R2 score	0.75	0.80



SmallDS



BigDS

5. Kết luận

5.1 Kết quả đạt được

- Crawl dữ liệu từ web

- Xử lý, phân tích dữ liệu và train model
- Hoàn thành yêu cầu dự đoán với độ chính xác tương đối cao.

5.2 Hướng phát triển

- Thu thập thêm các dữ liệu đa dạng hơn.
- Xây dựng mô hình có độ chính xác cao hơn.
- Áp dụng các kĩ thuật tiền xử lý để tăng độ chính xác của mô hình.

6. Tài liệu tham khảo

https://www.academia.edu/82395781/YouTube_Data_Analysis_and_Prediction_of_Views_and_Categories

https://www.mathworks.com/campaigns/offers/next/machine-learning-vs-deep-learning.html

https://drive.google.com/drive/folders/12KNleapSgtcwCJglSsbpLtkIDVN4pG47