

TIỂU LUẬN KHOA HỌC DỮ LIỆU

Dự đoán lượt view của video

NHÓM: 4

GVHD: TS. Ninh Khánh Duy

Hồ Như Phong

MSSV: 102200145

Nguyễn Thị Thu Thuyền


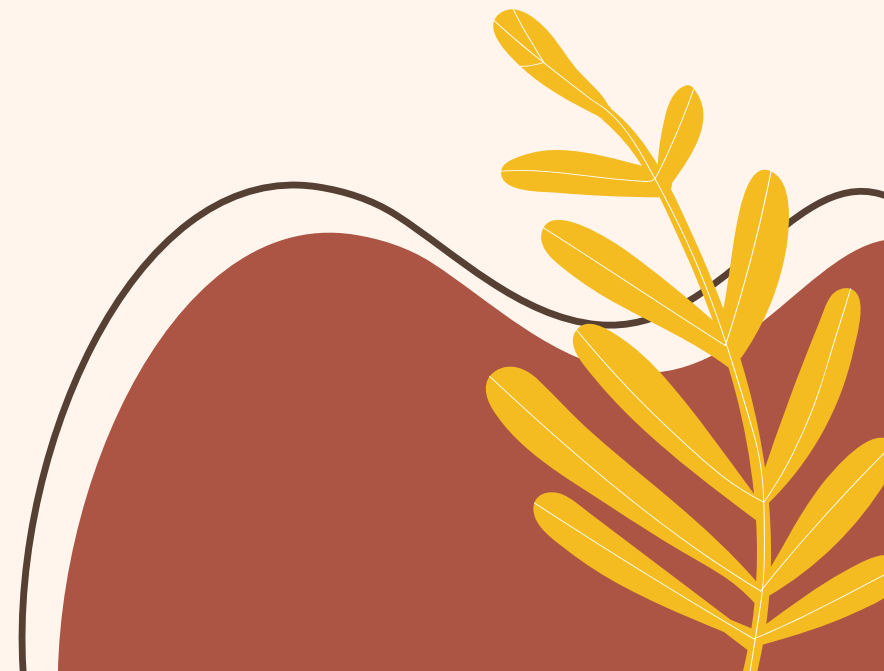
MSSV: 102200156

Nguyễn Hoàng Ngọc

MSSV: 102200141



Nội dung

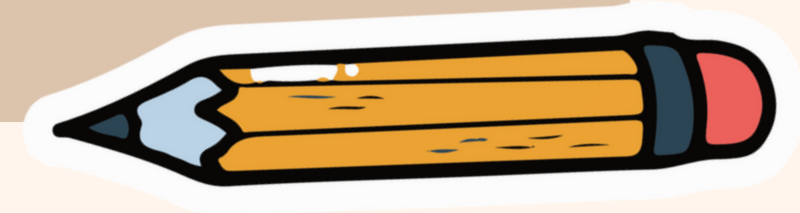
1. Giới thiệu
 2. Thu thập và mô tả dữ liệu
 3. Trích xuất đặc trưng
 4. Mô hình hóa dữ liệu
 5. Đánh giá hiệu suất mô hình
 6. Kết luận
- 
- 

Giới thiệu



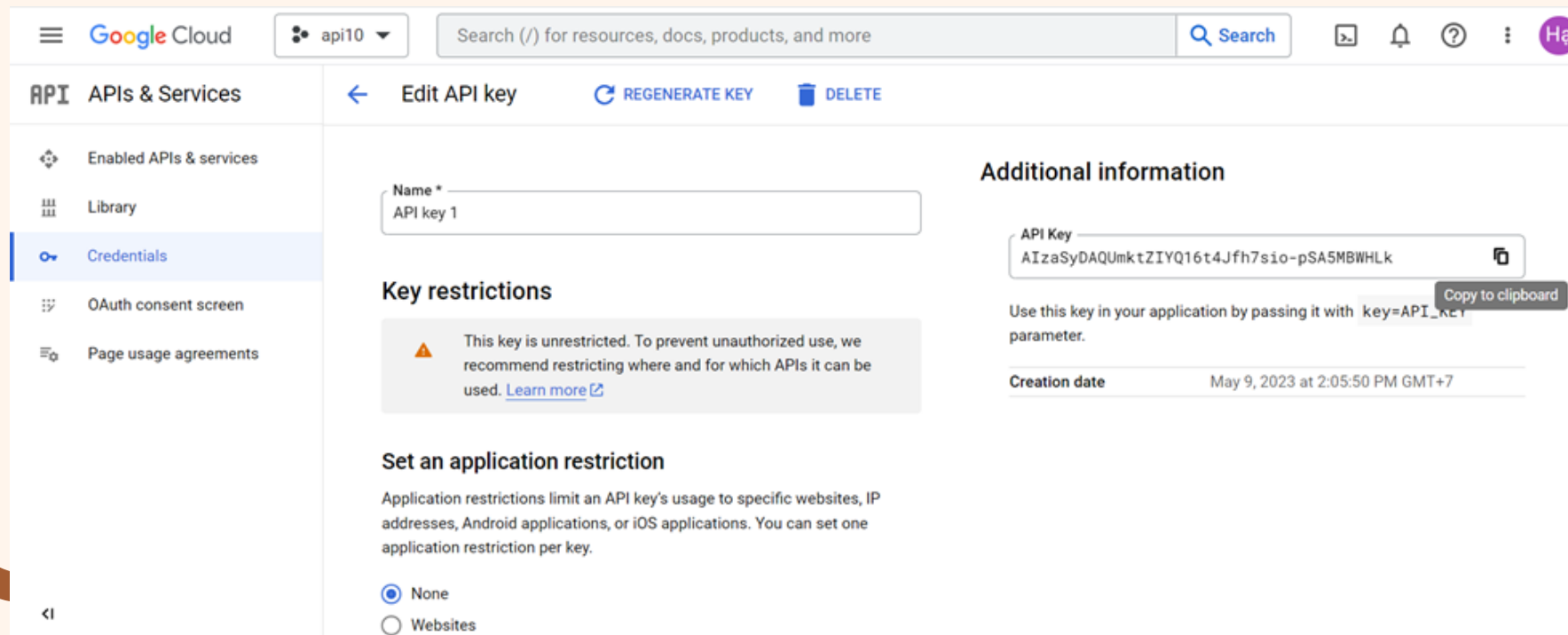
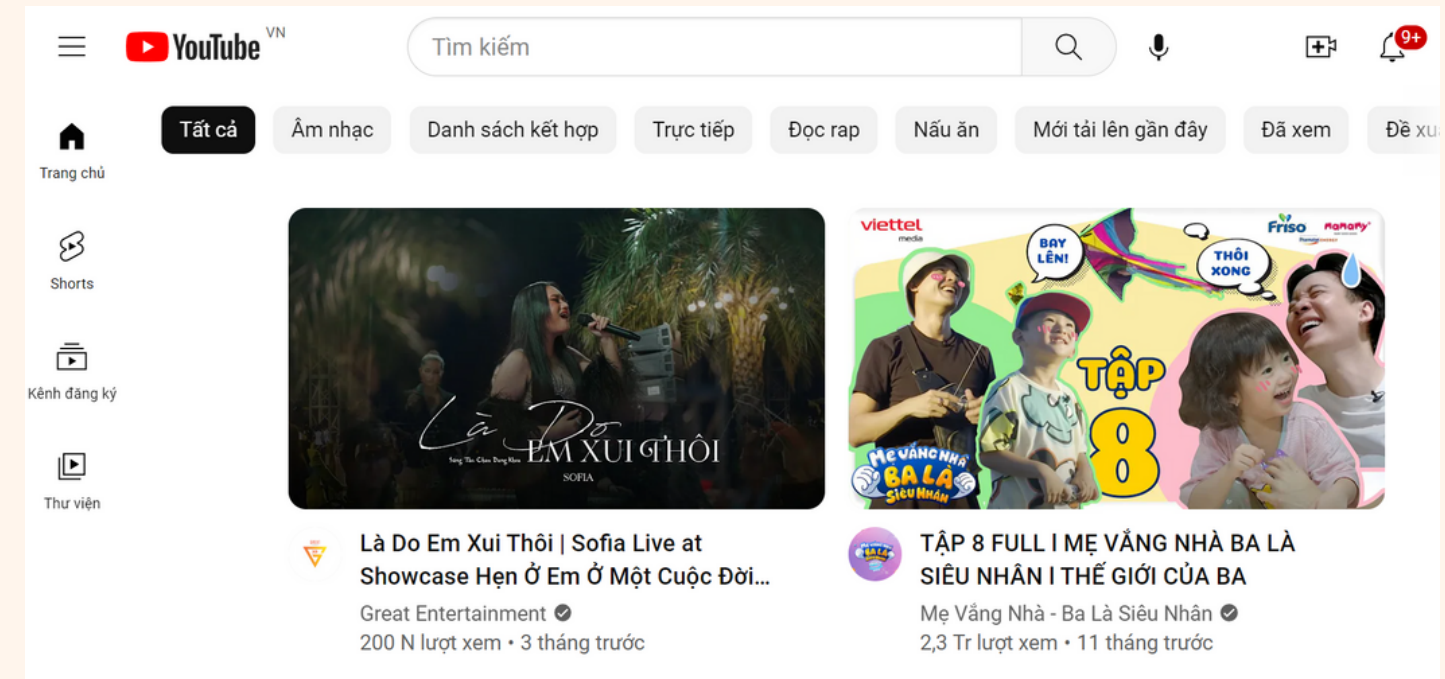
Trong dự án này, chúng ta sẽ tạo một mô hình dự đoán lượt xem của các video trên nền tảng Youtube. Mô hình sẽ sử dụng các thông tin liên quan như tiêu đề, mô tả, thời lượng video, lượt đăng kí kênh, chủ đề của các video và các yếu tố khác để dự đoán chính xác số lượt xem.

Mô hình này có thể hỗ trợ người dùng và các nhà sản xuất, sáng tạo nội dung trong việc đánh giá và tối ưu hóa hiệu quả của các video trước khi công bố.



Thu thập dữ liệu

- Thu thập dữ liệu từ: <https://www.youtube.com/> để lấy các thông tin liên quan về video.
- Có thể sử dụng API Youtube Data v3 để truy xuất thông tin về channel và video trên nền tảng này. Để sử dụng API Youtube Data v3 cần đăng ký và tạo một API key.



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 danh mục video hoặc danh sách các channel trên Youtube. Dựa vào đó có thể sử dụng API để truy cập thông tin về video sau đó lưu trữ dữ liệu dưới dạng .csv

Trích xuất đặc trưng

1. Làm sạch dữ liệu

- Loại bỏ các mẫu có "video_id" trùng nhau
- 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)

2. Thêm đặc trưng mới

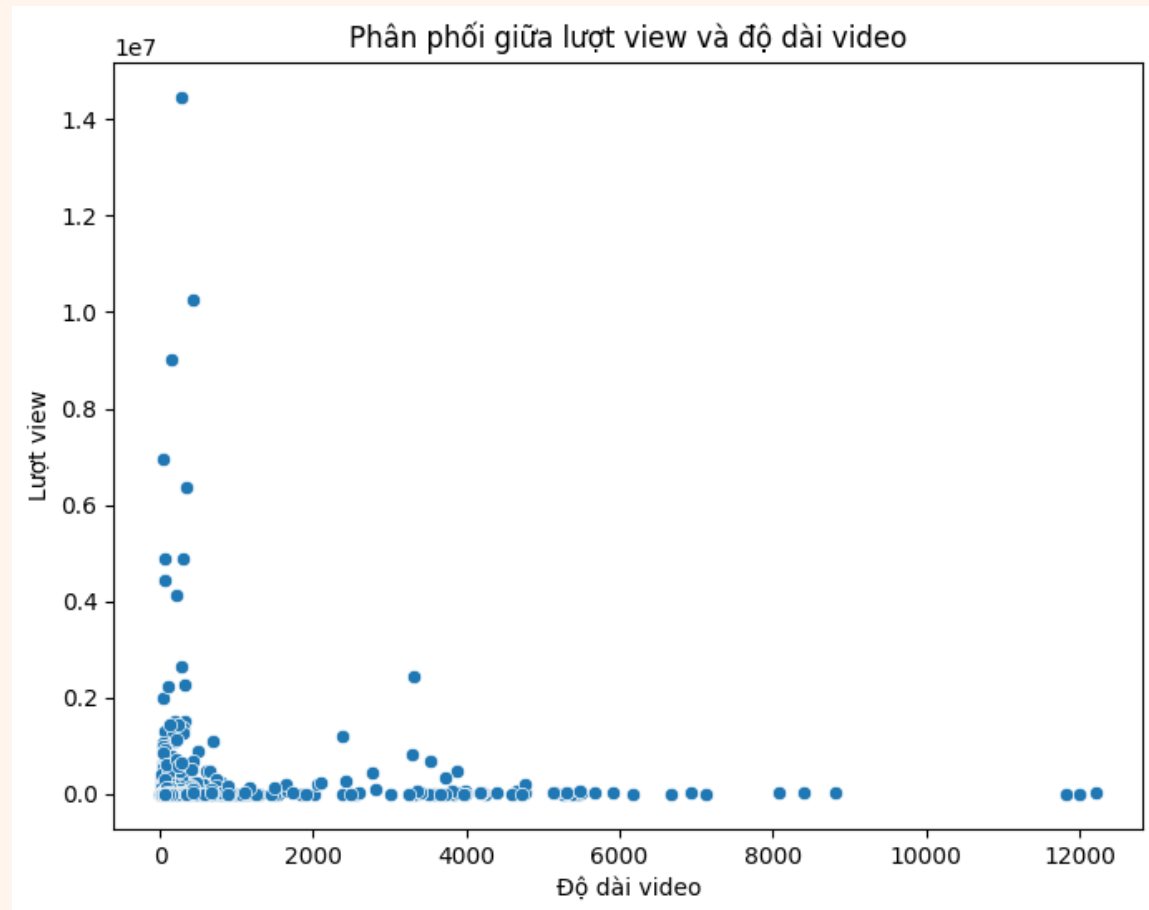
- Tạo đặc trưng số lượng từ trong tiêu đề và mô tả.
- Tạo đặc trưng thời gian từ lúc video được đăng.

Dữ liệu sau khi xử lý sẽ có 1023 dòng và 12 cột

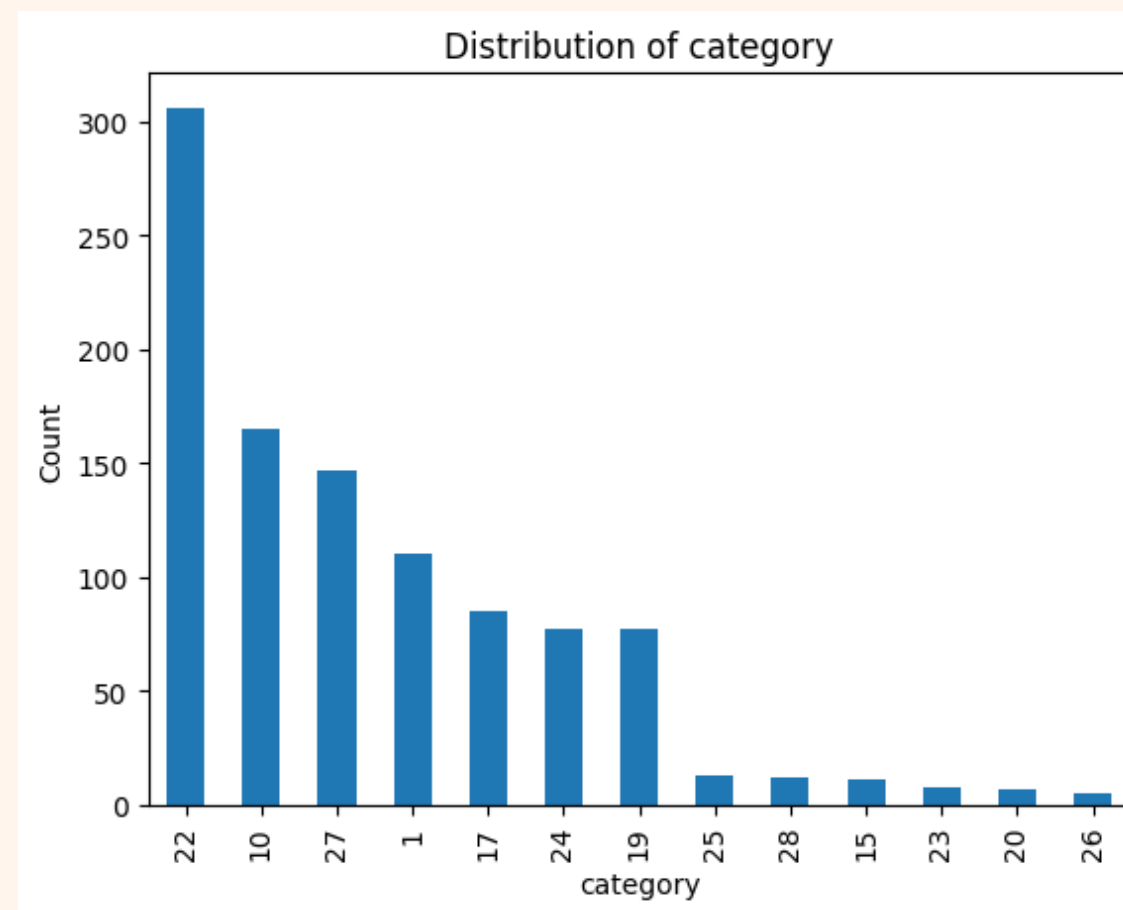
```
Int64Index: 1023 entries, 0 to 1056
Data columns (total 12 columns):
#   Column                Non-Null Count  Dtype
---  -
0   video_id              1023 non-null   object
1   channel_id            1023 non-null   object
2   channel_name          1023 non-null   object
3   published_date        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
```

- 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

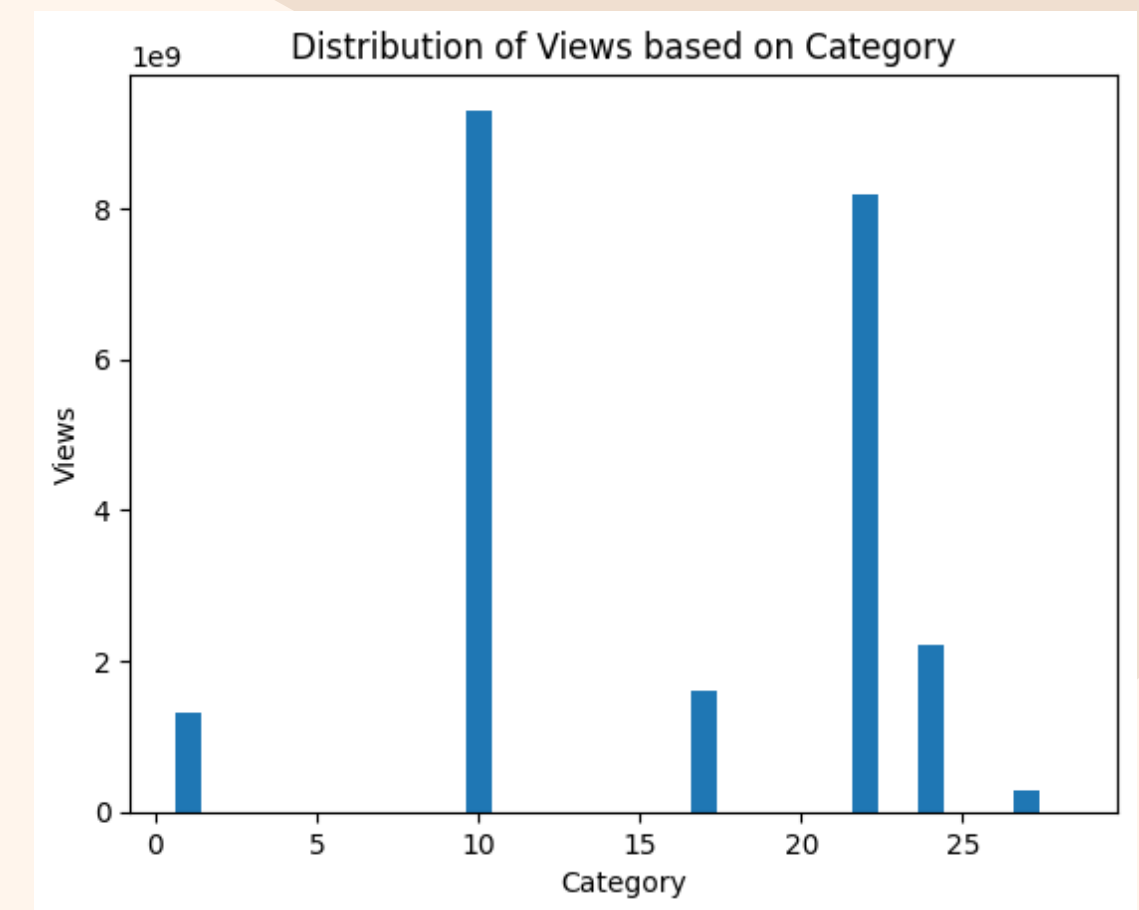
Phân tích dữ liệu



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



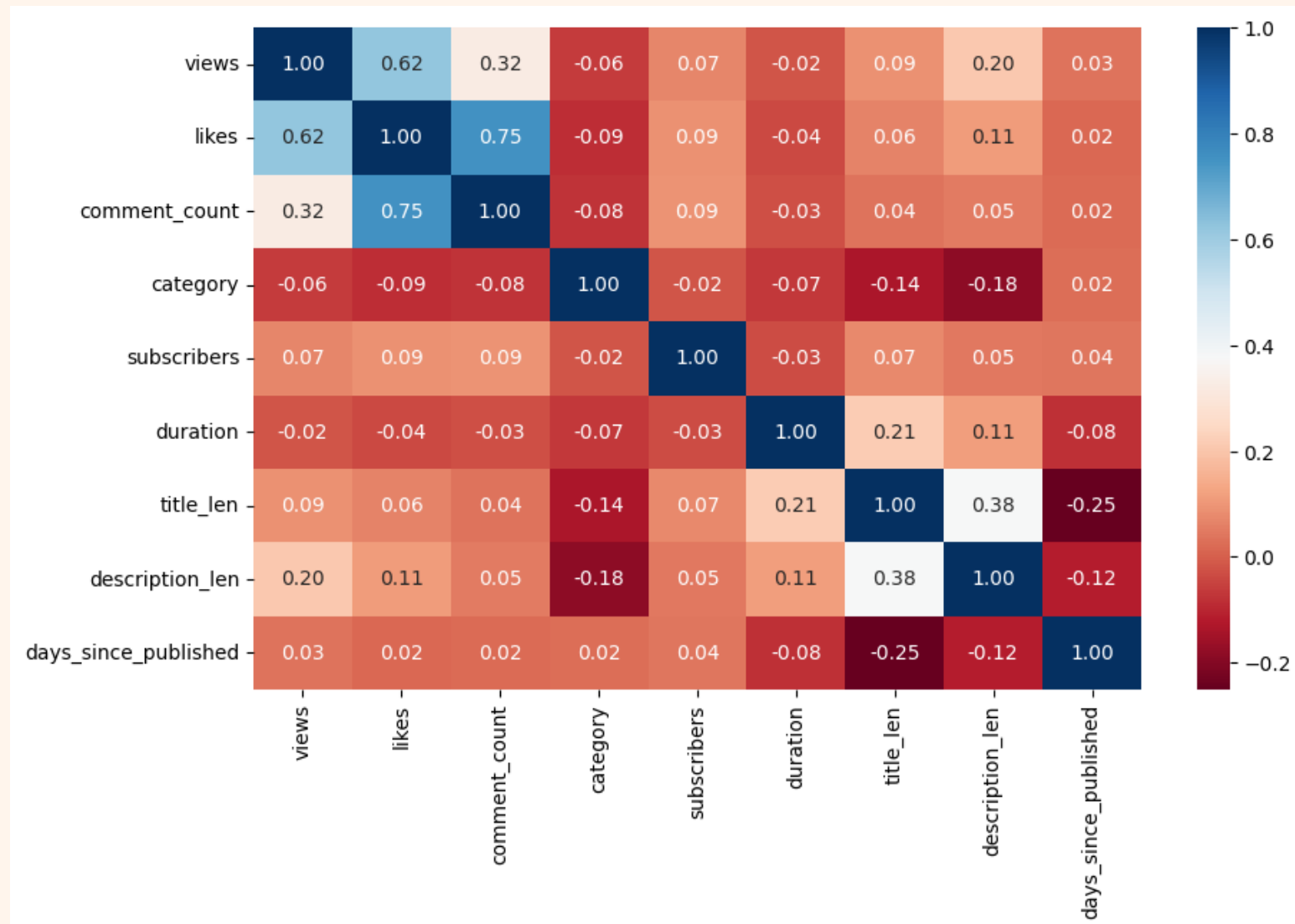
Nhận xét: Đối với các danh mục khác nhau thì số lượng video đăng tải cũng khác nhau



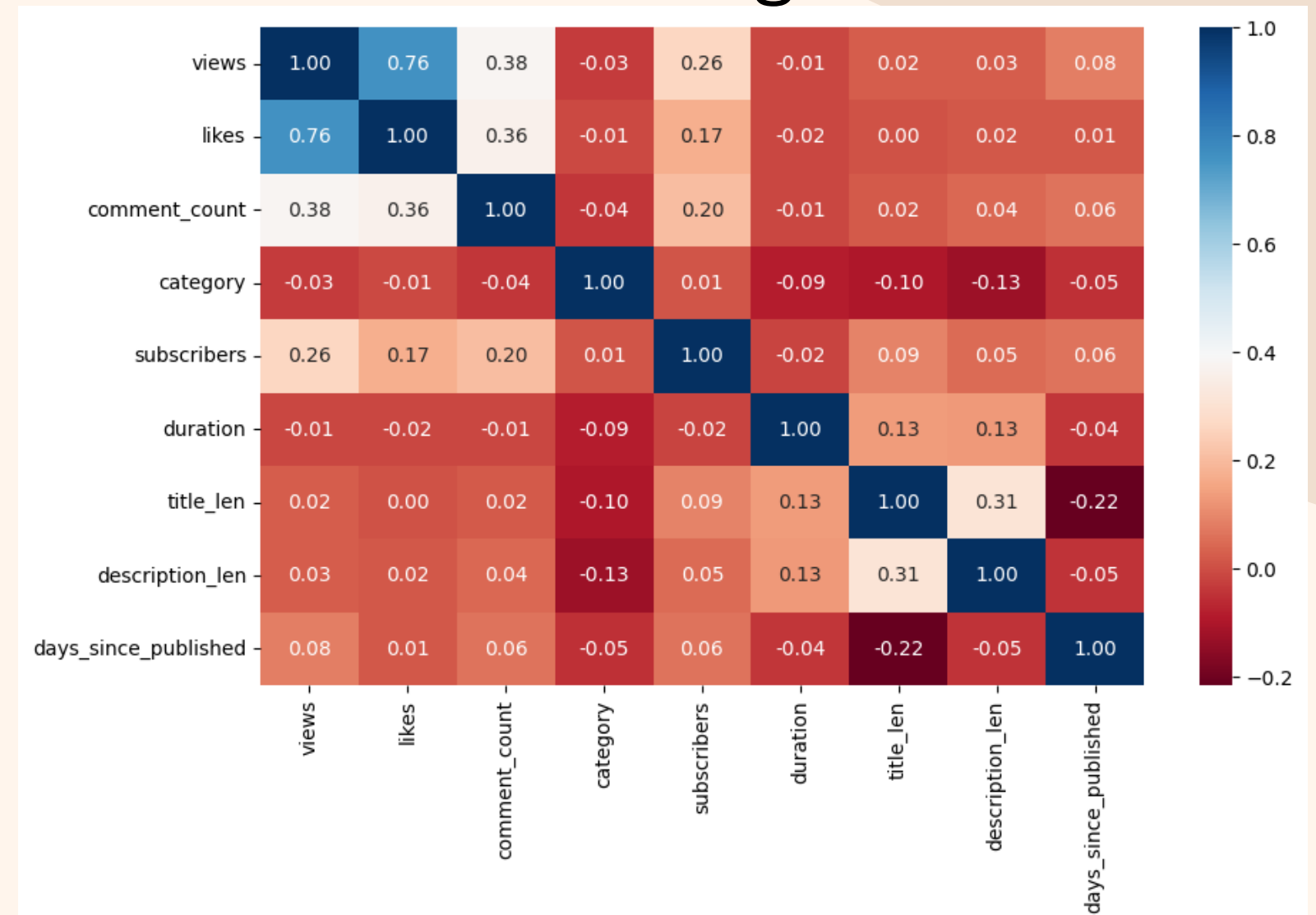
Nhận xét: Danh mục video cũng ảnh hưởng đến lượt xem của video đó. tùy vào từng danh mục và lượt view thay đổi

Phân tích dữ liệu

SmallDS



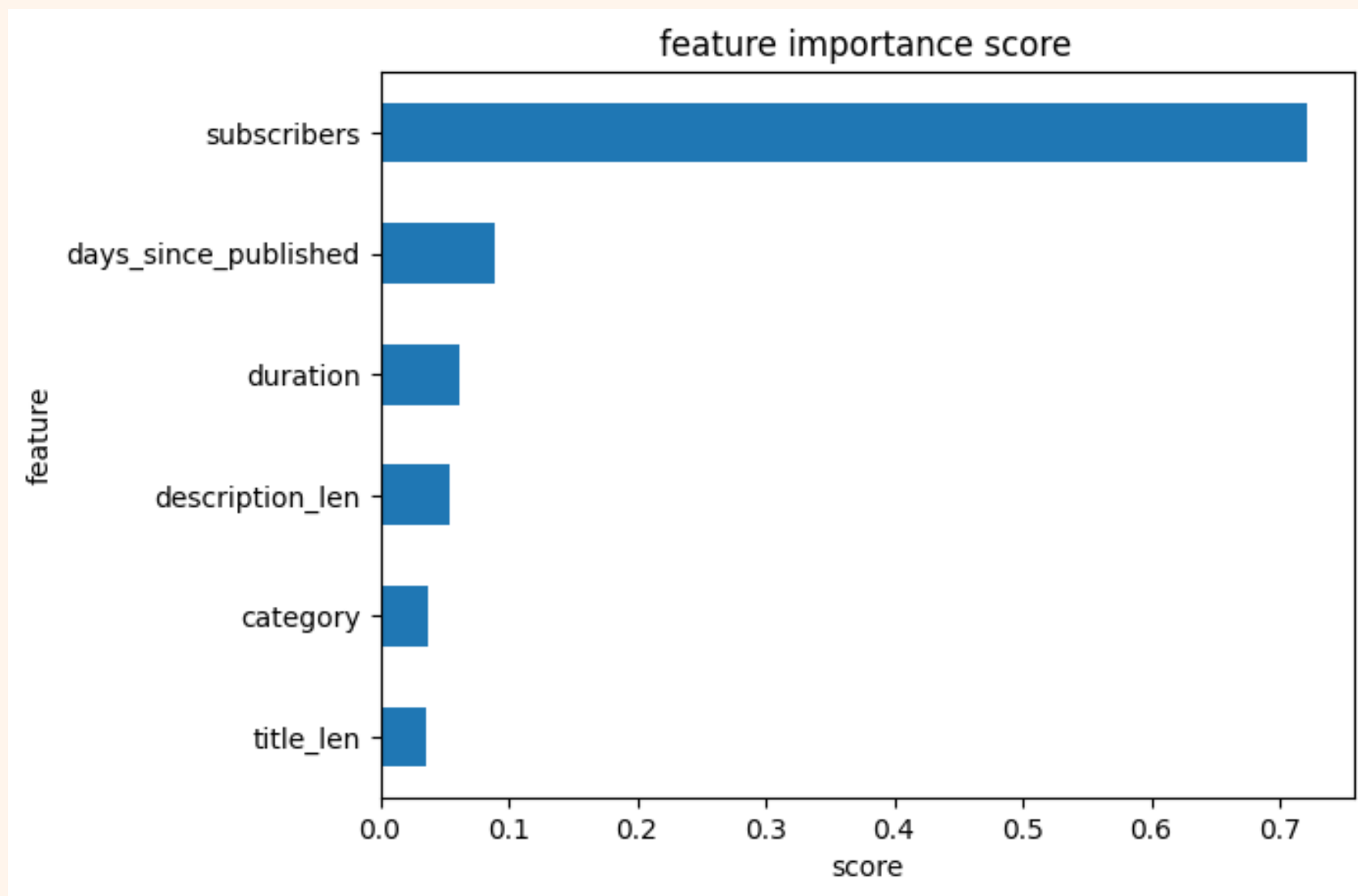
BigDS



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).

Kỹ thuật lựa chọn đặc trưng

Sử dụng kỹ thuật SelectKBest và Random Forest để lấy ra 5 đặc trưng có tương quan mạnh.



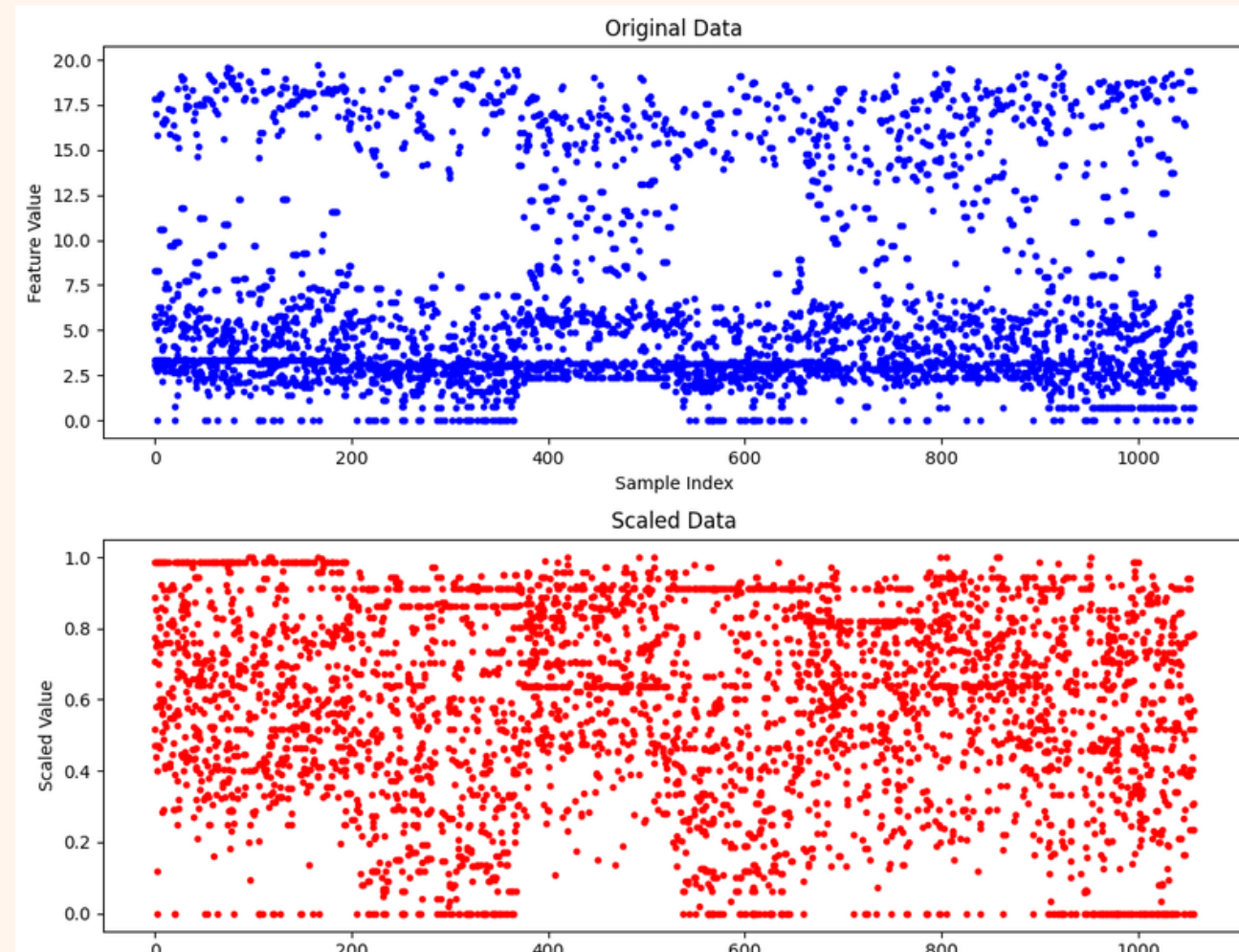
- Với SelectKbest thì 5 đặc trưng là: 'category', 'subscribers', 'duration', 'title_len', 'description_len'
- Với Random Forest thì 5 đặc trưng là: 'subscribers', 'days_since_published', 'duration', 'description_len', 'title_len'

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.

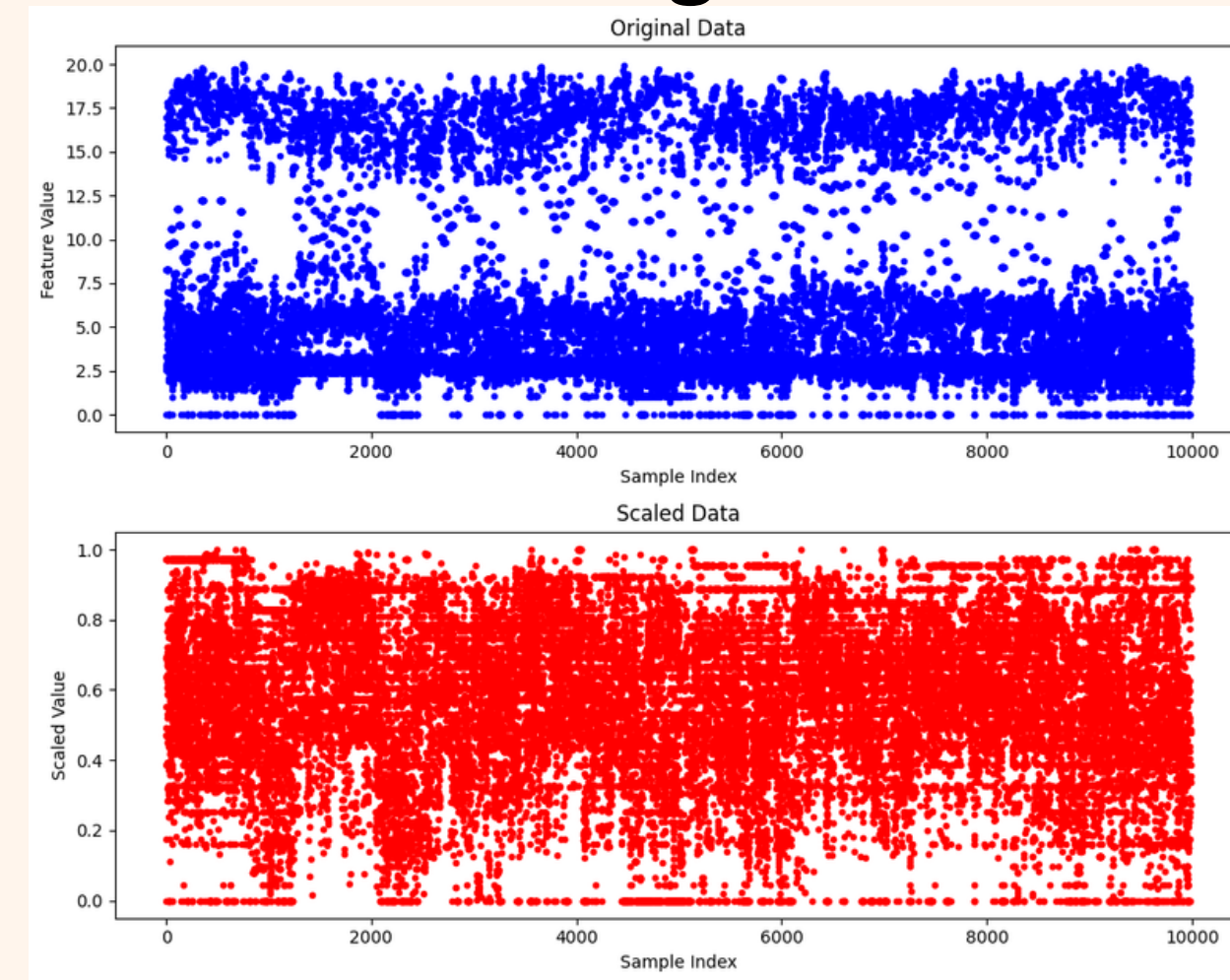
Kỹ thuật Random Forest

Trực quan hóa dữ liệu trước và sau khi chuẩn hóa

SmallDS



BigDS



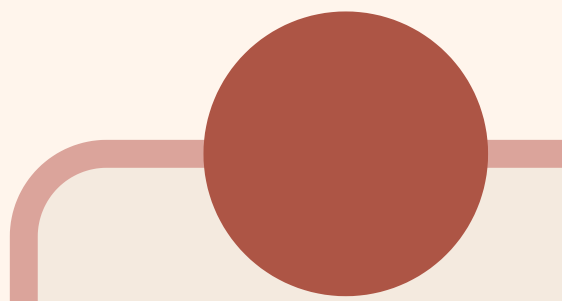
Nhận xét:

- * 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.

Mô hình hóa dữ liệu

- **Model Random Forest**
- **Model Linear Regressor**
- **Model Support Vector Regressor**



Model Random Forest

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.

Kết quả dự đoán trên tập val sau khi được huấn luyện ở tập train:

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.

Model Random Forest

Hiệu chỉnh tham số

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

Đánh giá kết quả
trên tập validation

Kết quả dự
đoán trên tập
test

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

Model Linear Regressor

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.

Kết quả dự đoán trên tập val sau khi được huấn luyện ở tập train:

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 kỹ thuật lựa chọn đặc trưng cho mô hình này.

Model Linear Regressor

Hiệu chỉnh tham số

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

Đánh giá kết quả trên tập validaton

Kết quả dự đoán trên tập test

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

Model Support Vector Regressor

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.

Kết quả dự đoán trên tập val sau khi được huấn luyện ở tập train:

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 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 kỹ thuật lựa chọn đặc trưng cho mô hình này.

Model Support Vector Regressor

Hiệu chỉnh tham số

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

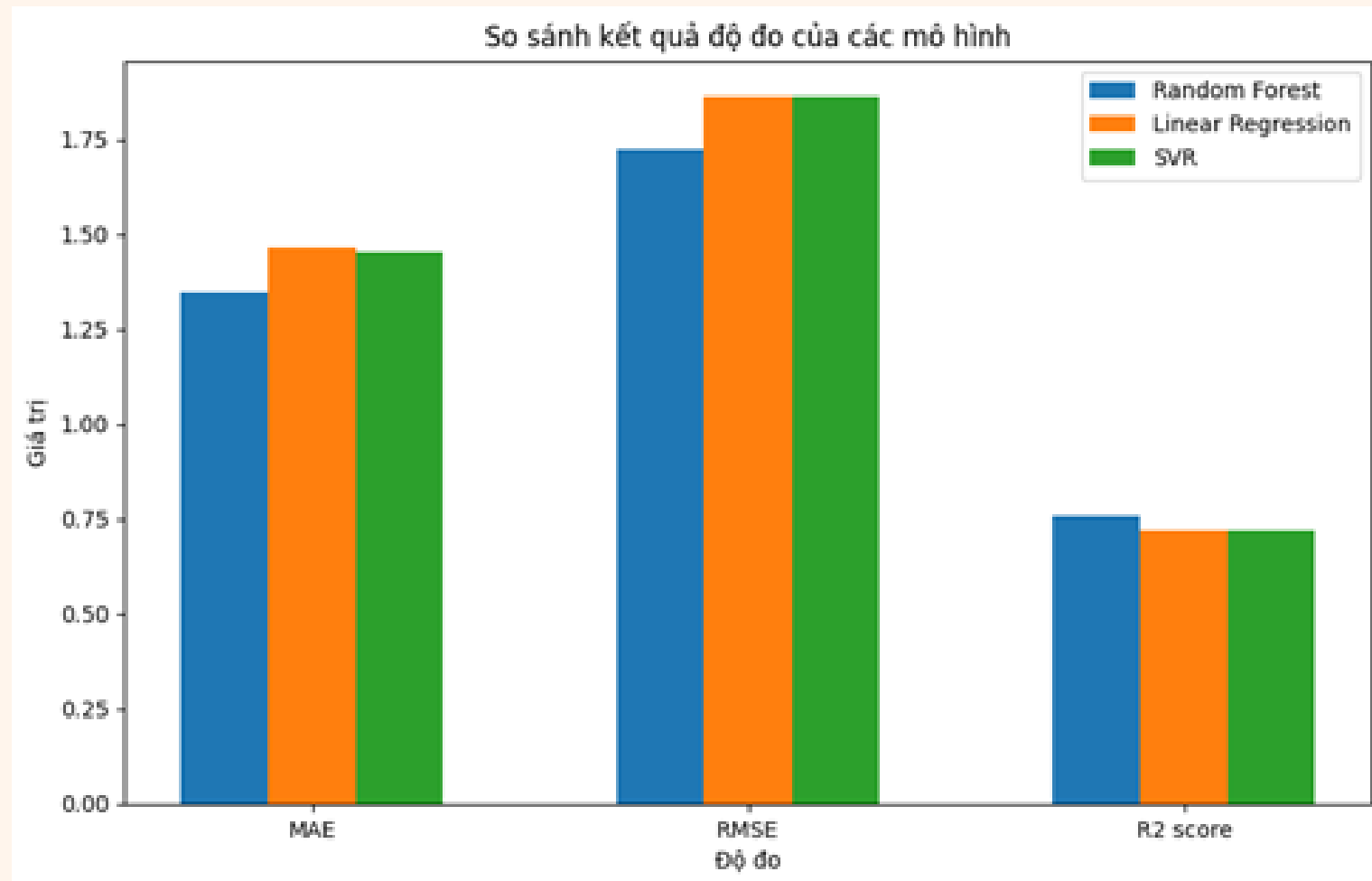
Đánh giá kết quả
trên tập validaton

Kết quả dự
đoán trên tập
test

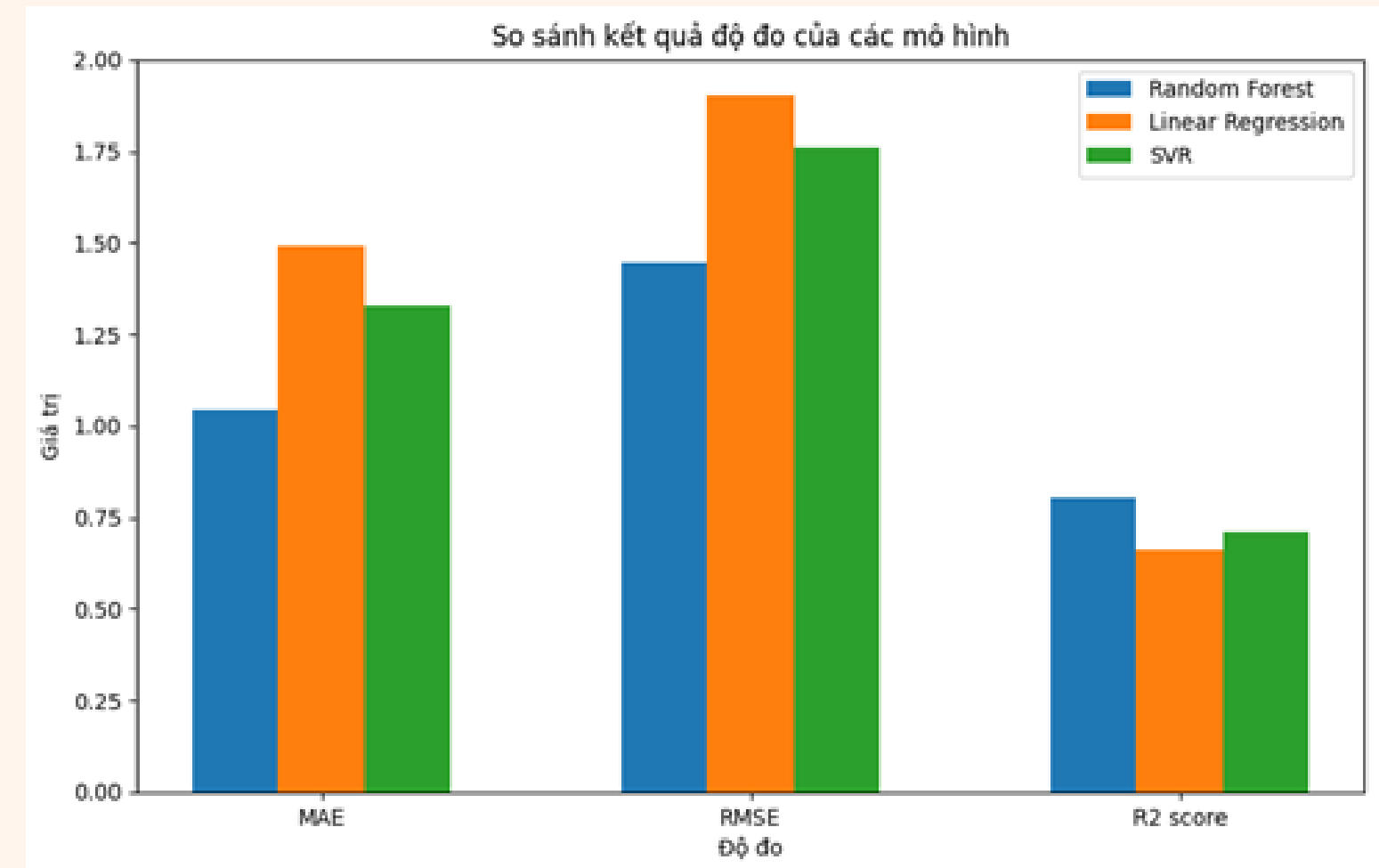
Tham số	SmallDS	BigDS
MAE	1.45	1.33
RMSE	1.86	1.76
R2 score	0.72	0.71

Đánh giá hiệu suất của các mô hình

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



SmallDS



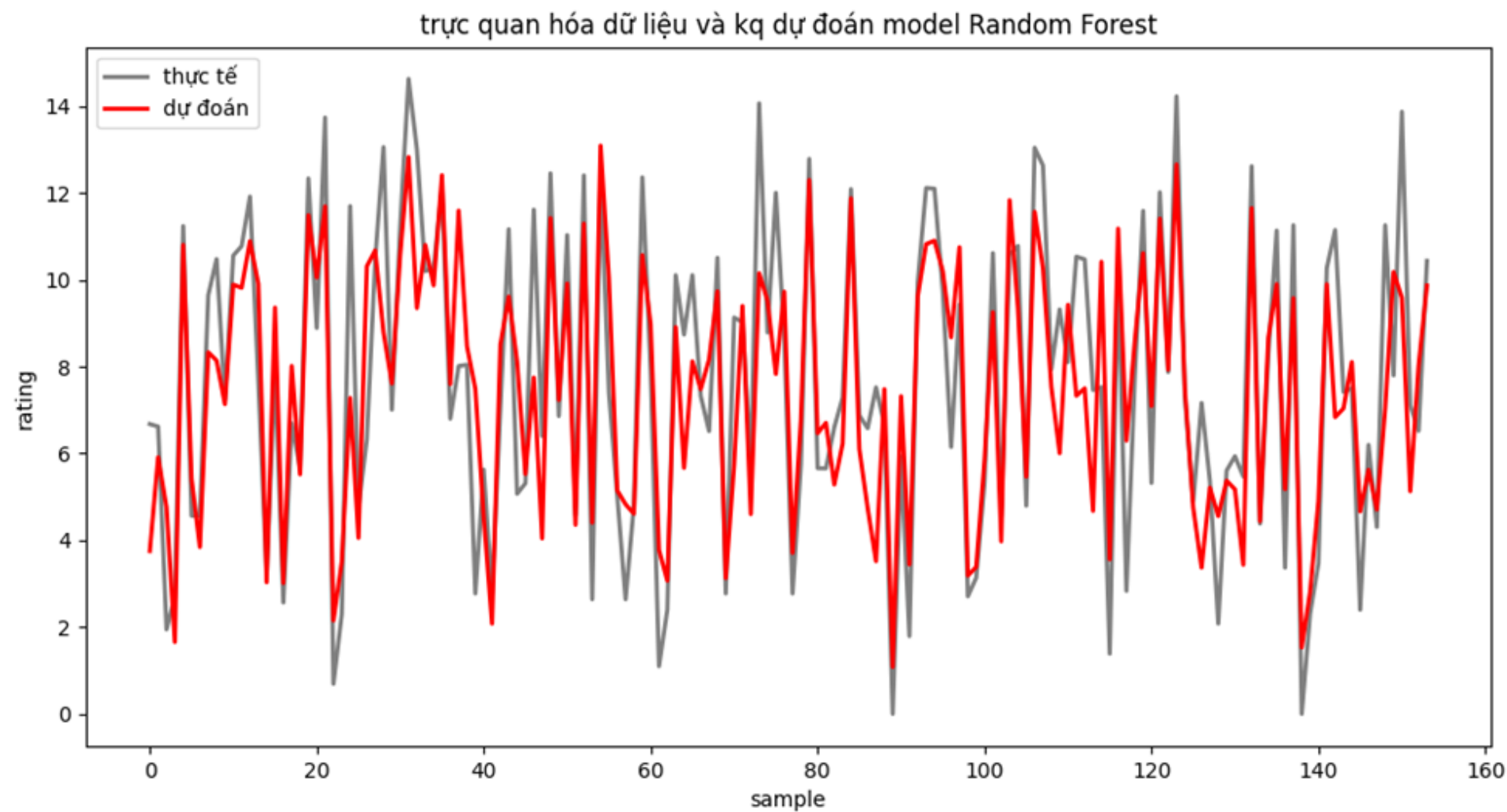
BigDS

Kết luận:

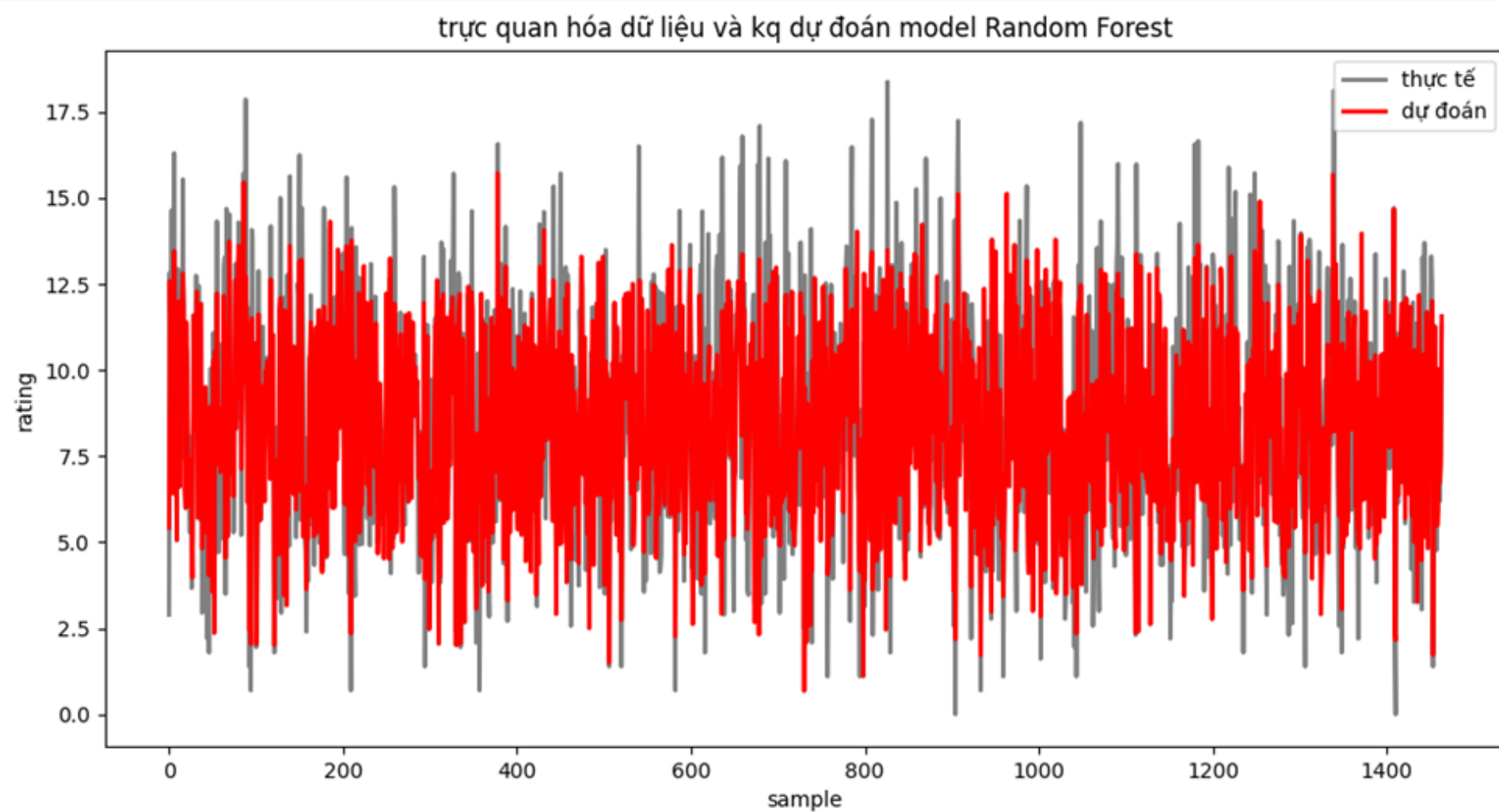
Dựa vào cả 2 biểu đồ và kết quả độ đo, có thể kết luận rằng mô hình Random Forest cho kết quả tốt nhất trong số ba mô hình được so sánh (Random Forest, Linear Regression, SVR).

Đánh giá hiệu suất của các mô hình

Trực quan kết quả dự đoán của mô hình Random Forest

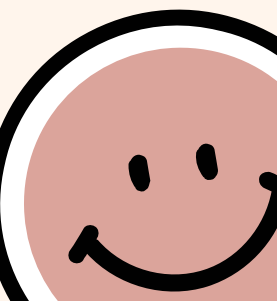


SmallDS

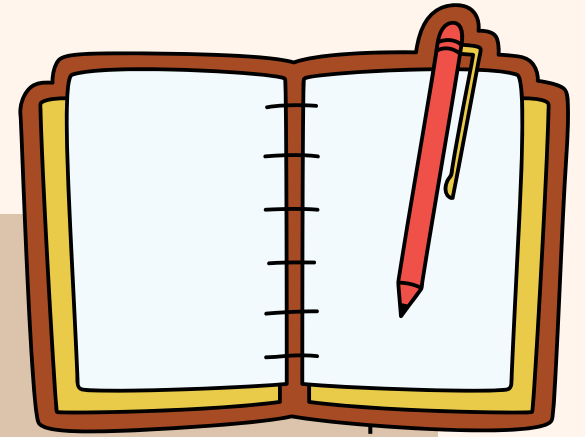


BigDS

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



Kết luận

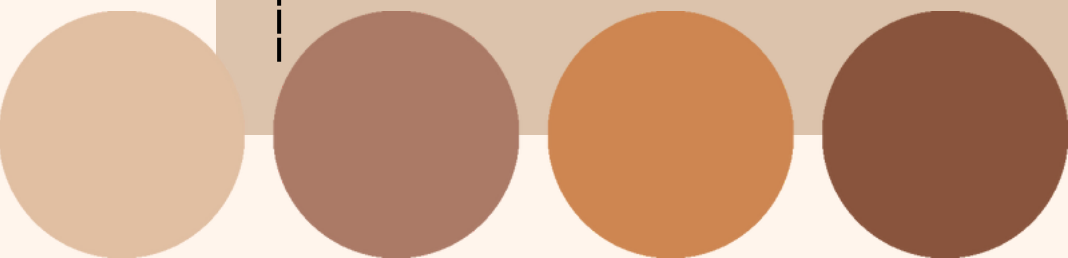


5.1 Kết quả đạt được

- Crawl dữ liệu từ youtube
- 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 tiên xử lý để tăng độ chính xác của mô hình.



**Cảm ơn thầy và các bạn
đã lắng nghe!**

