

# PHÂN CỤM PHIM VÀ PHÂN LOẠI RATING PHIM DỰA TRÊN THÔNG TIN BỘ PHIM ĐÓ

## NHÓM 5

Thành viên:  
Nguyễn Lê Việt Hoàng  
Võ Đắc Bảo Ân  
Hà Ngọc Hưng



Tên	Nhiệm vụ	Đánh giá
Nguyễn Lê Việt Hoàng	<ul style="list-style-type: none"><li>• Thu thập dữ liệu</li><li>• Trích xuất đặc trưng</li><li>• Làm bài toán phân loại rating</li></ul>	<ul style="list-style-type: none"><li>• Đã hoàn thành</li><li>• Đã hoàn thành</li><li>• Đã hoàn thành</li></ul>
Võ Đắc Bảo Ân	<ul style="list-style-type: none"><li>• Làm sạch dữ liệu</li><li>• Trích xuất đặc trưng</li><li>• Làm bài toán phân cụm phim</li></ul>	<ul style="list-style-type: none"><li>• Đã hoàn thành</li><li>• Đã hoàn thành</li><li>• Đã hoàn thành</li></ul>
Hà Ngọc Hưng	<ul style="list-style-type: none"><li>• Trực quan hóa dữ liệu</li></ul>	<ul style="list-style-type: none"><li>• Đã hoàn thành</li></ul>

# Mục lục

01

Giới thiệu

02

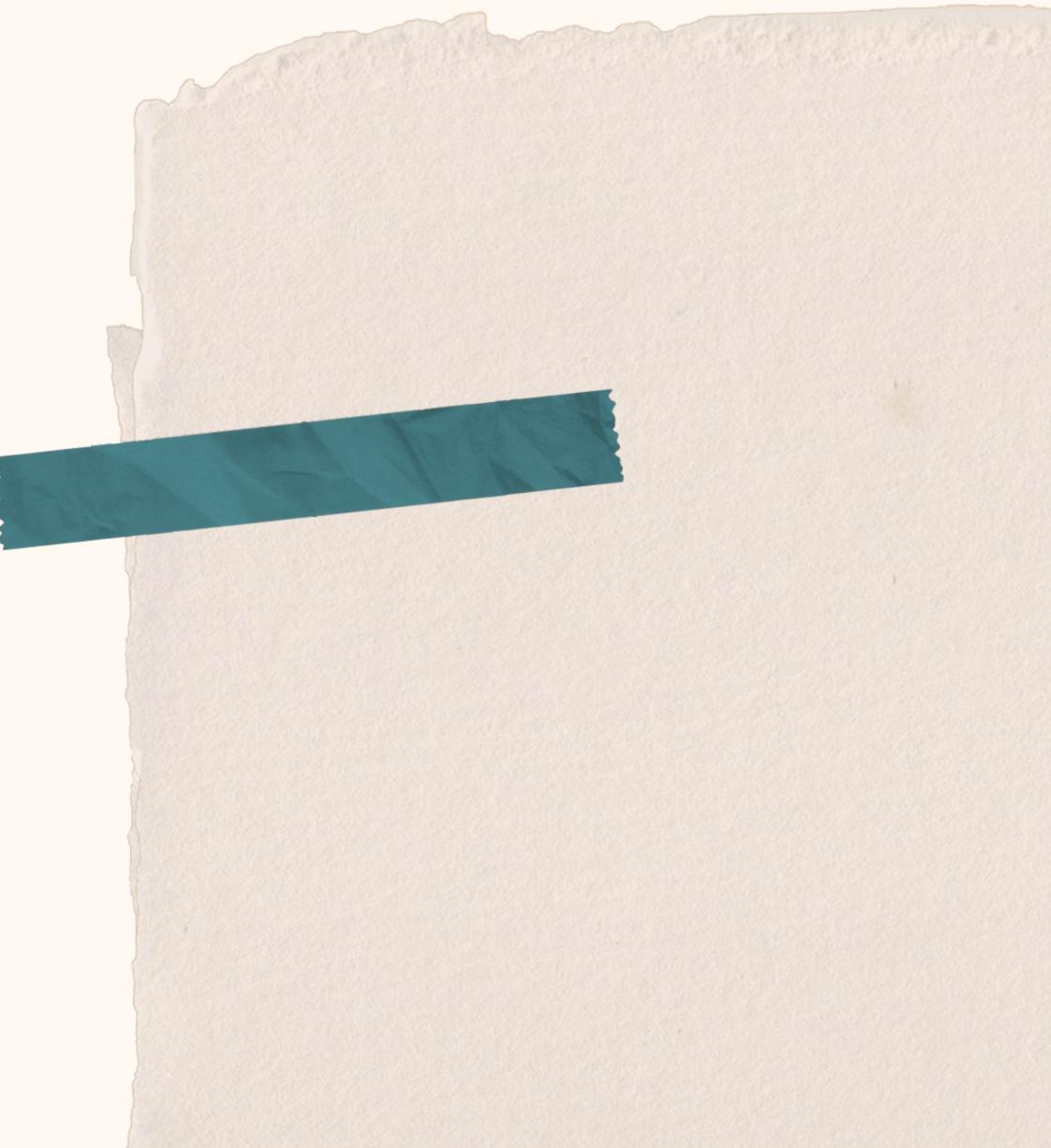
Thu thập và  
mô tả dữ liệu

03

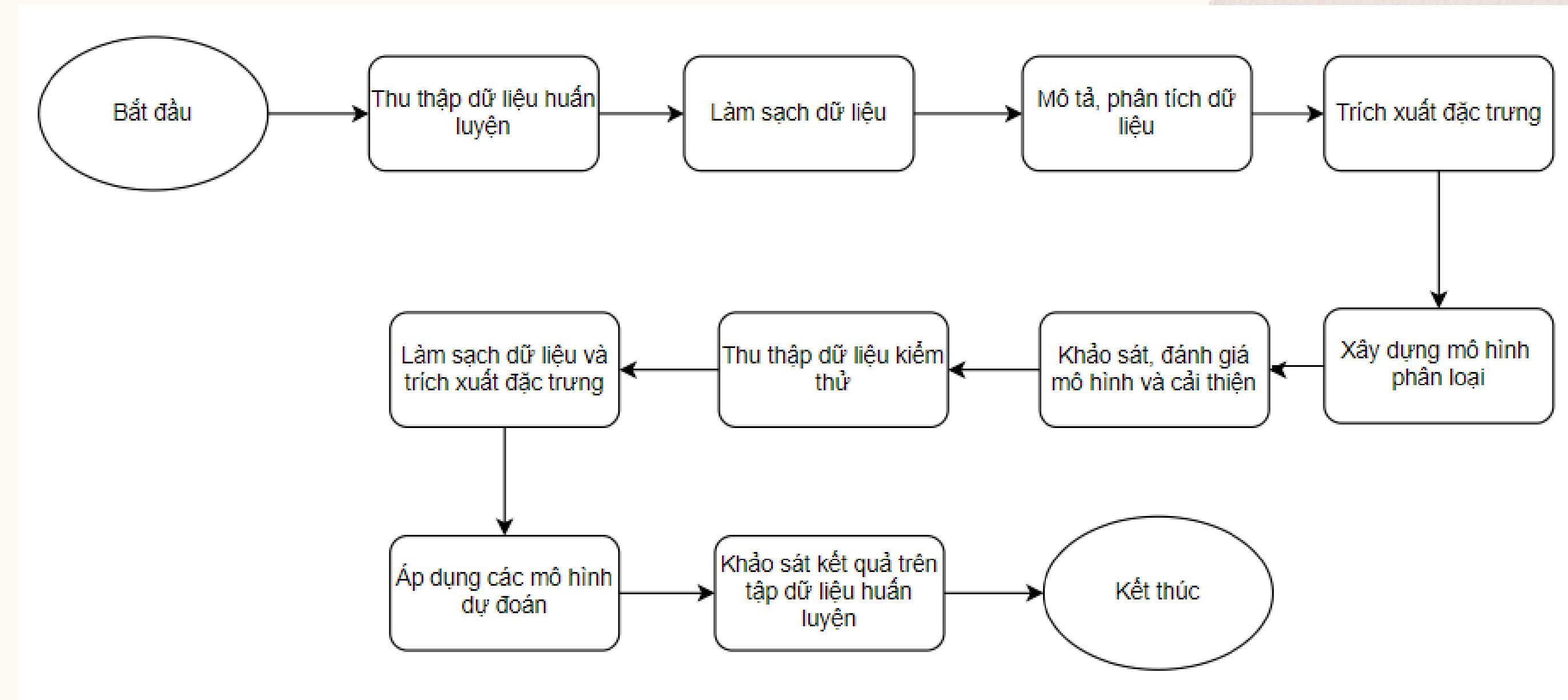
Trích xuất  
đặc trưng

04

Mô hình hóa  
dữ liệu và  
kết luận

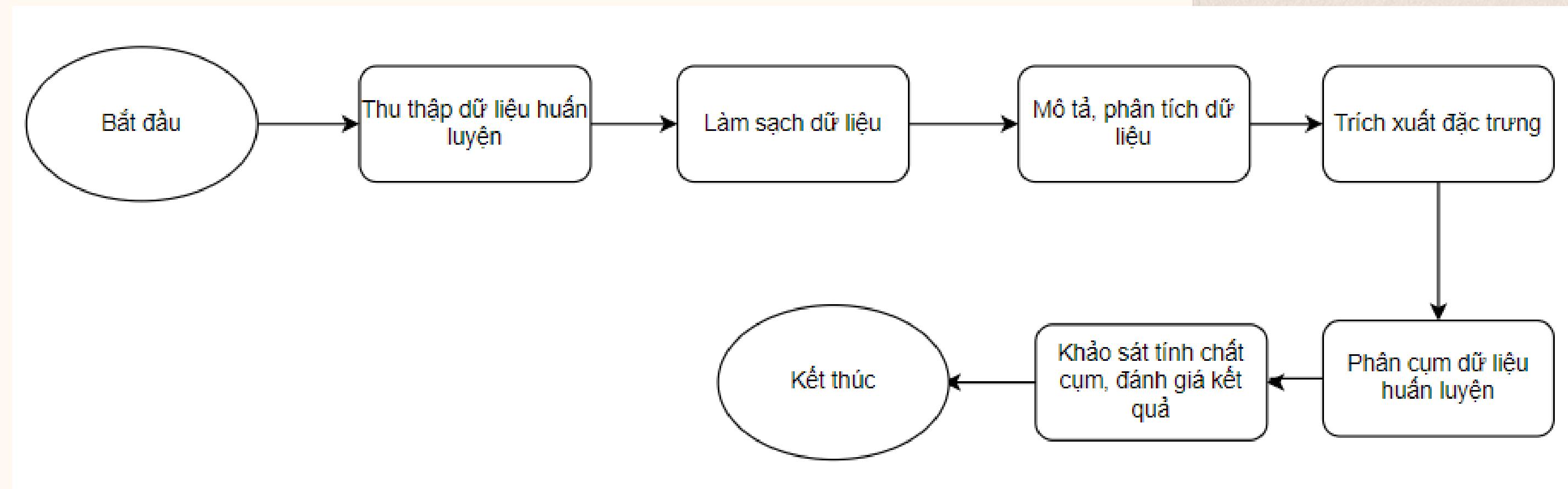


# 1. Giới thiệu



Giải pháp tổng quan về phân loại

# 1. Giới thiệu



Giải pháp tổng quan về phân cụm

## 2. Thu thập và mô tả dữ liệu

- Nguồn dữ liệu: <https://www.imdb.com/>
- Công cụ thu thập: ngôn ngữ python, thư viện requests và BeautifulSoup
- Chương trình python thu thập dữ liệu sẽ duyệt qua lần lượt từng URL với định dạng “<https://www.imdb.com/title/tt>”+id\_film.

 <https://www.imdb.com/title/tt1200000>

Over the course of several years, two convicts redemption through basic compassion.

Director [Frank Darabont](#)

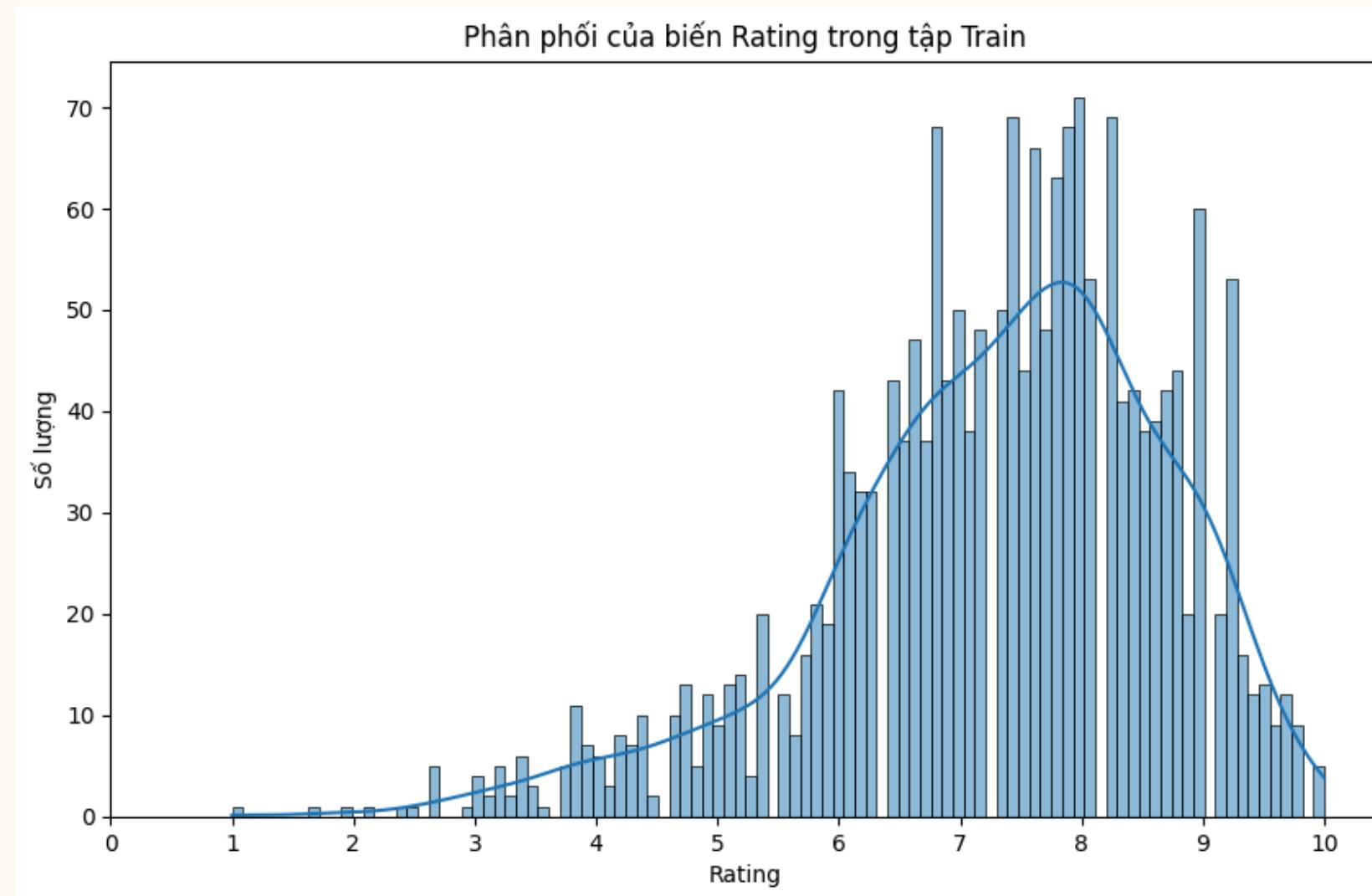
Writers [Stephen King](#) · [Frank Darabont](#)

Stars [Tim Robbins](#) · [Morgan Freeman](#) · [Bob](#)

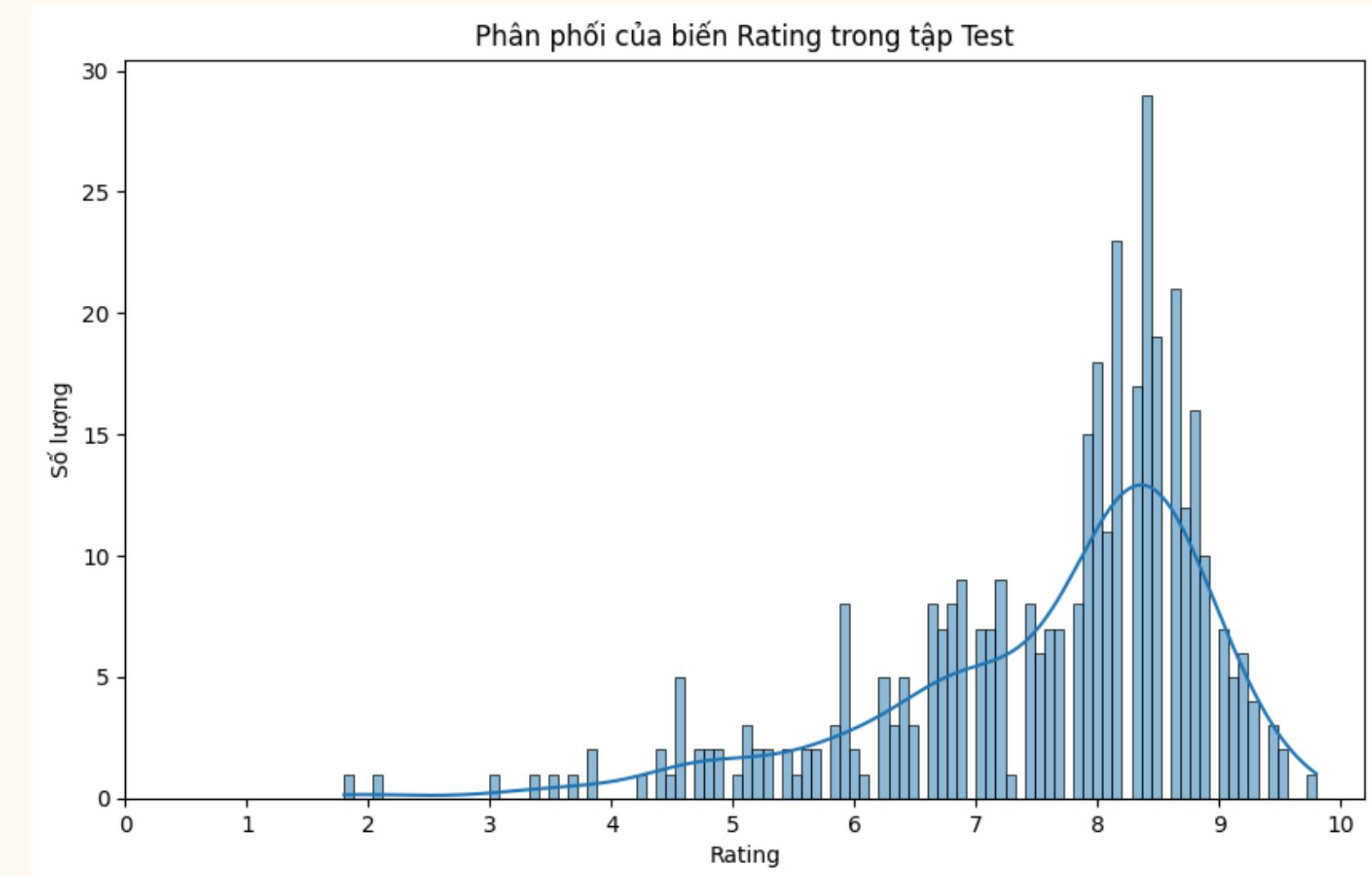
IMDbPro See production info at [IMDbPro](#) 

## 2. Thu thập và mô tả dữ liệu

### 2.2 Mô tả và trực quan hóa dữ liệu

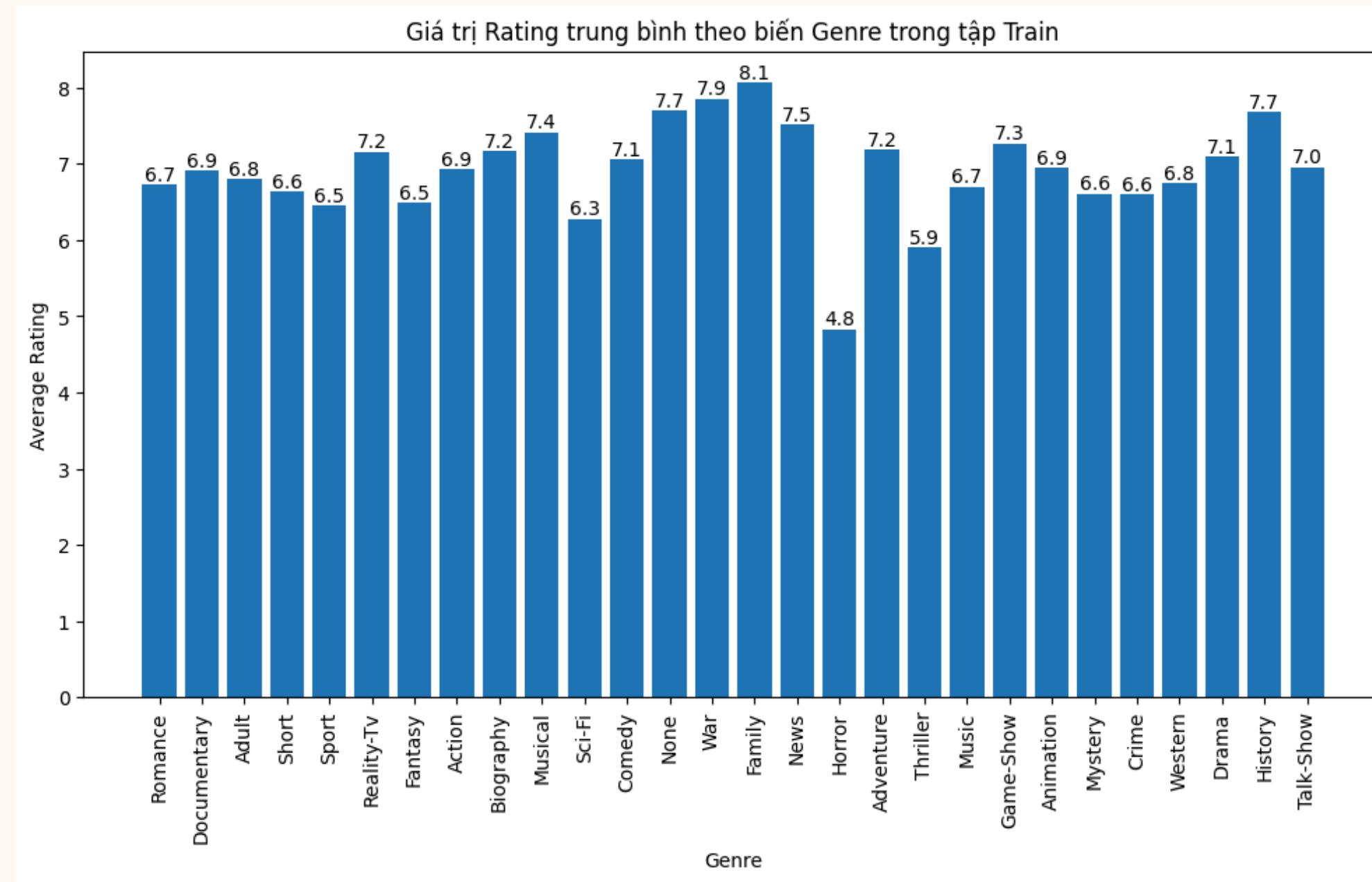


Số mẫu có Rating trong khoảng 7 đến 9 chiếm nhiều phần nhất trong tập Train, có rất ít mẫu có điểm Rating từ 0-4  
Đường cong của biểu đồ cho thấy phân phối chuẩn của dữ liệu đạt đỉnh ở khoảng giá trị rating từ 7 đến 8



Ở tập Test, số mẫu có điểm Rating trong khoảng 8 đến 9 chiếm nhiều nhất trong tập, có rất ít mẫu nằm trong khoảng Rating từ 0-4  
Đường cong phân phối chuẩn của biến Rating đạt đỉnh ở khoảng giá trị từ 8 đến 9

# Trong tập Train

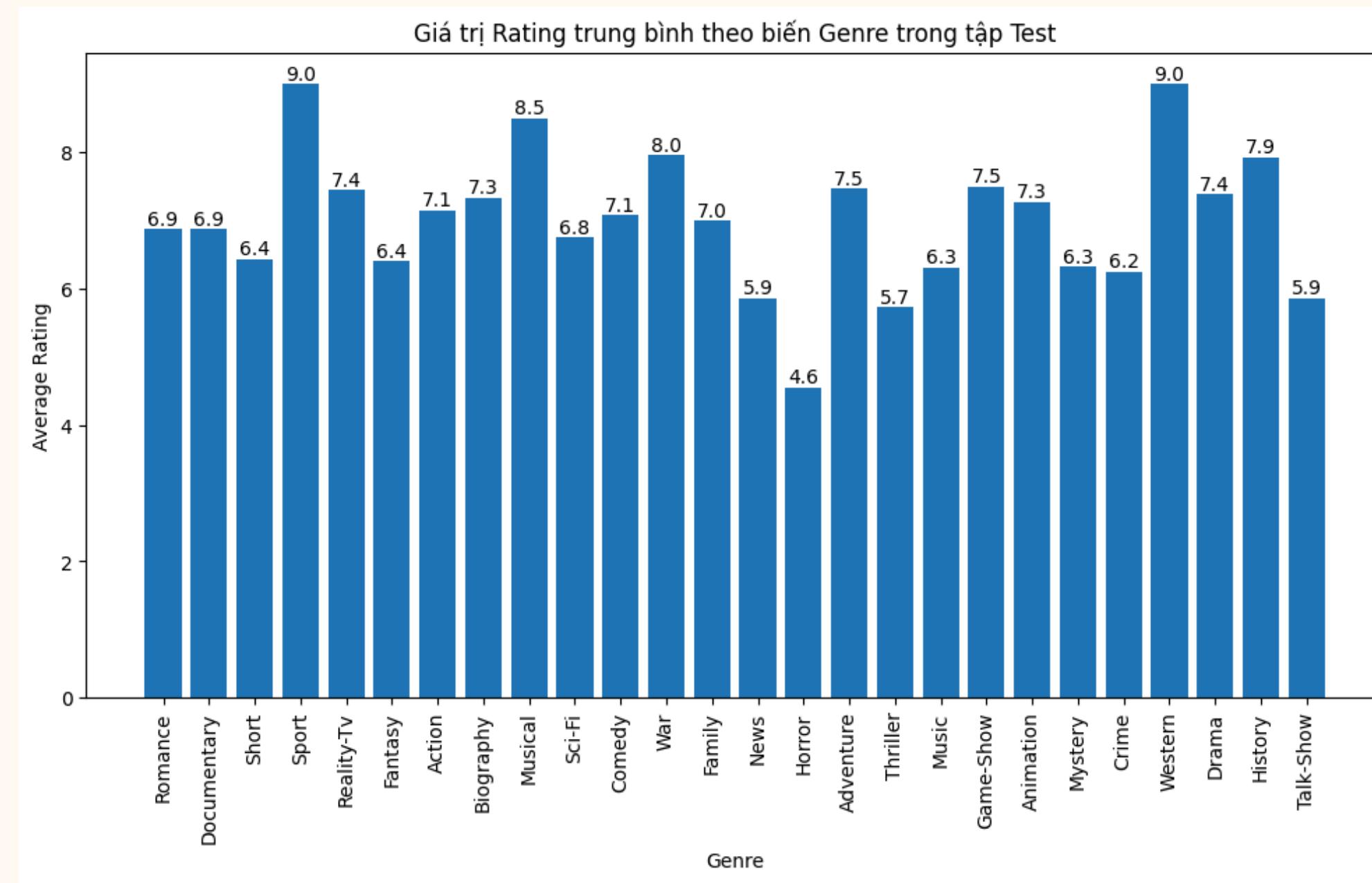


Family là thể loại phim (Genre) có điểm Rating cao nhất 8.1

Horror là thể loại phim có điểm Rating thấp nhất 4.9

Bên cạnh 2 thể loại phim có mức Rating cao nhất và thấp nhất, nhìn chung các thể loại phim khác có sự chênh lệch Rating với nhau không quá cao

# Trong tập Test

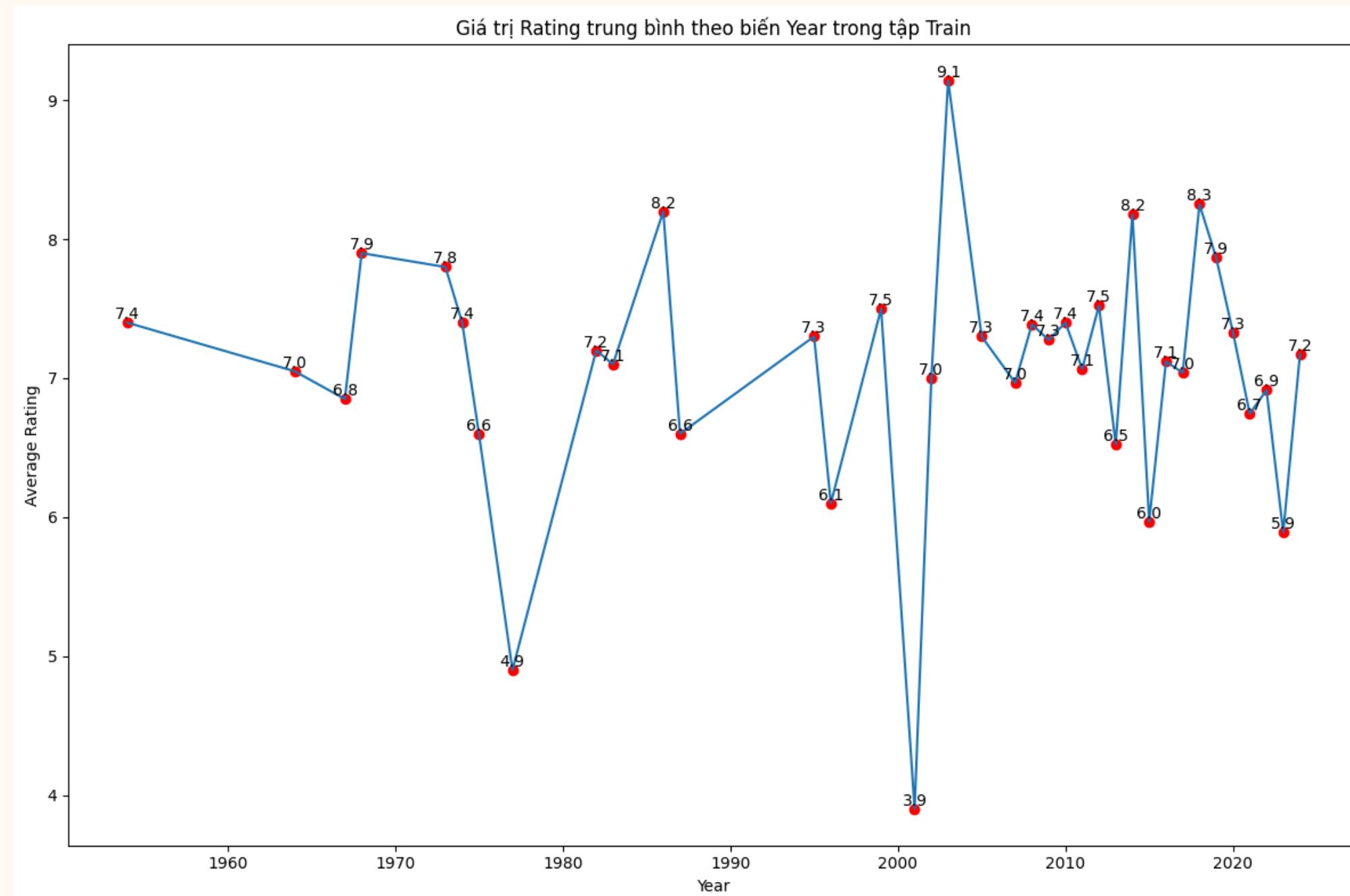


Sport và Western là thể loại phim (Genre) có điểm Rating cao nhất 9.0

Horror là thể loại phim có điểm Rating thấp nhất 4.6

Bên cạnh 2 thể loại phim có mức Rating cao nhất và thấp nhất, nhìn chung các thể loại phim khác có sự chênh lệch Rating với nhau không quá đáng kể

# Trong tập Train

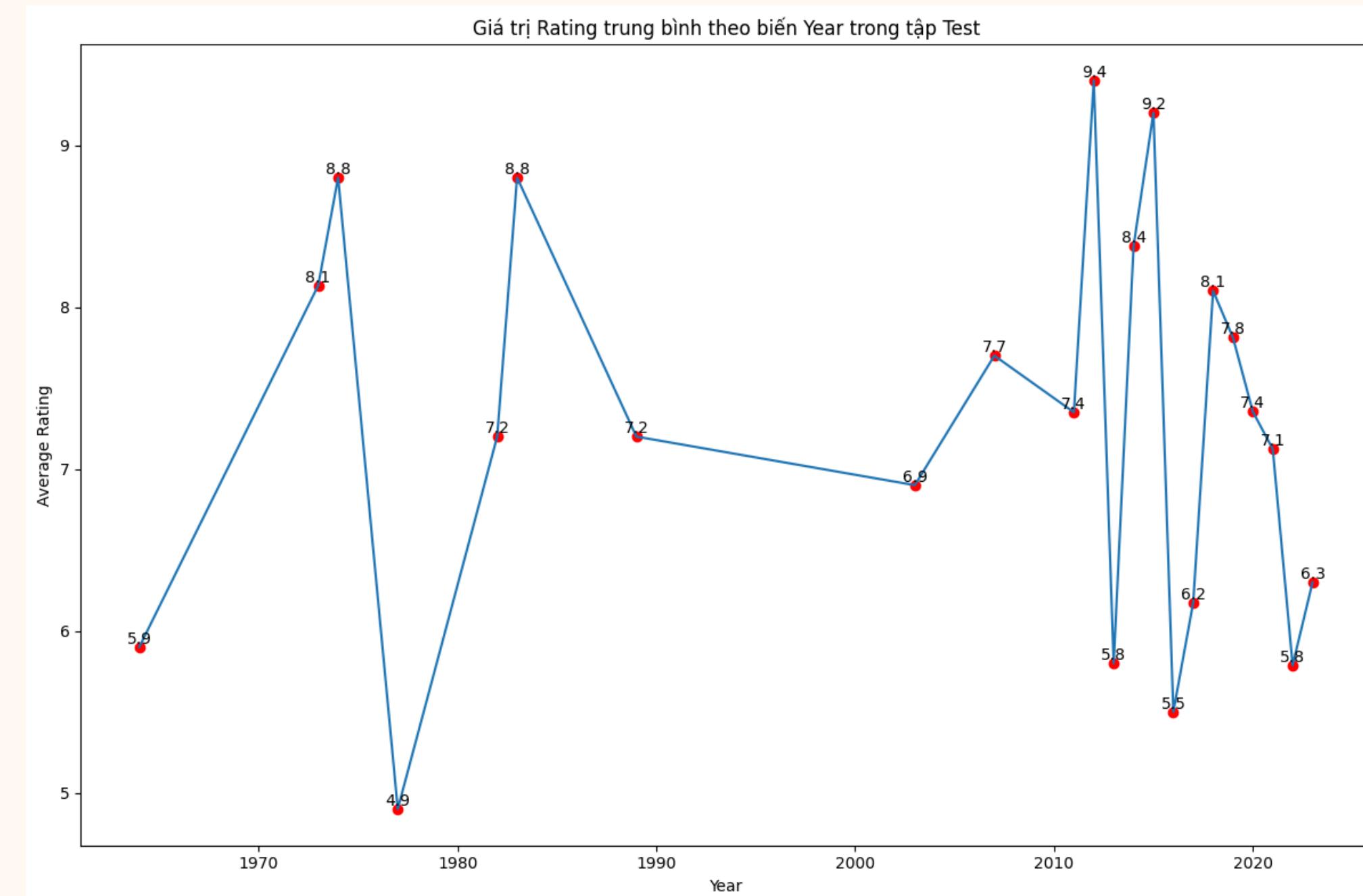


Điểm rating trung bình năm cao nhất là 9.1 và nằm trong khoảng thời gian từ 2000-2010

Điểm rating trung bình năm thấp nhất là 3.9 và nằm trong khoảng thời gian từ 2000-2010

Điểm rating trung bình giữa các khoảng thời gian khác có sự chênh lệch lớn

# Trong tập Test

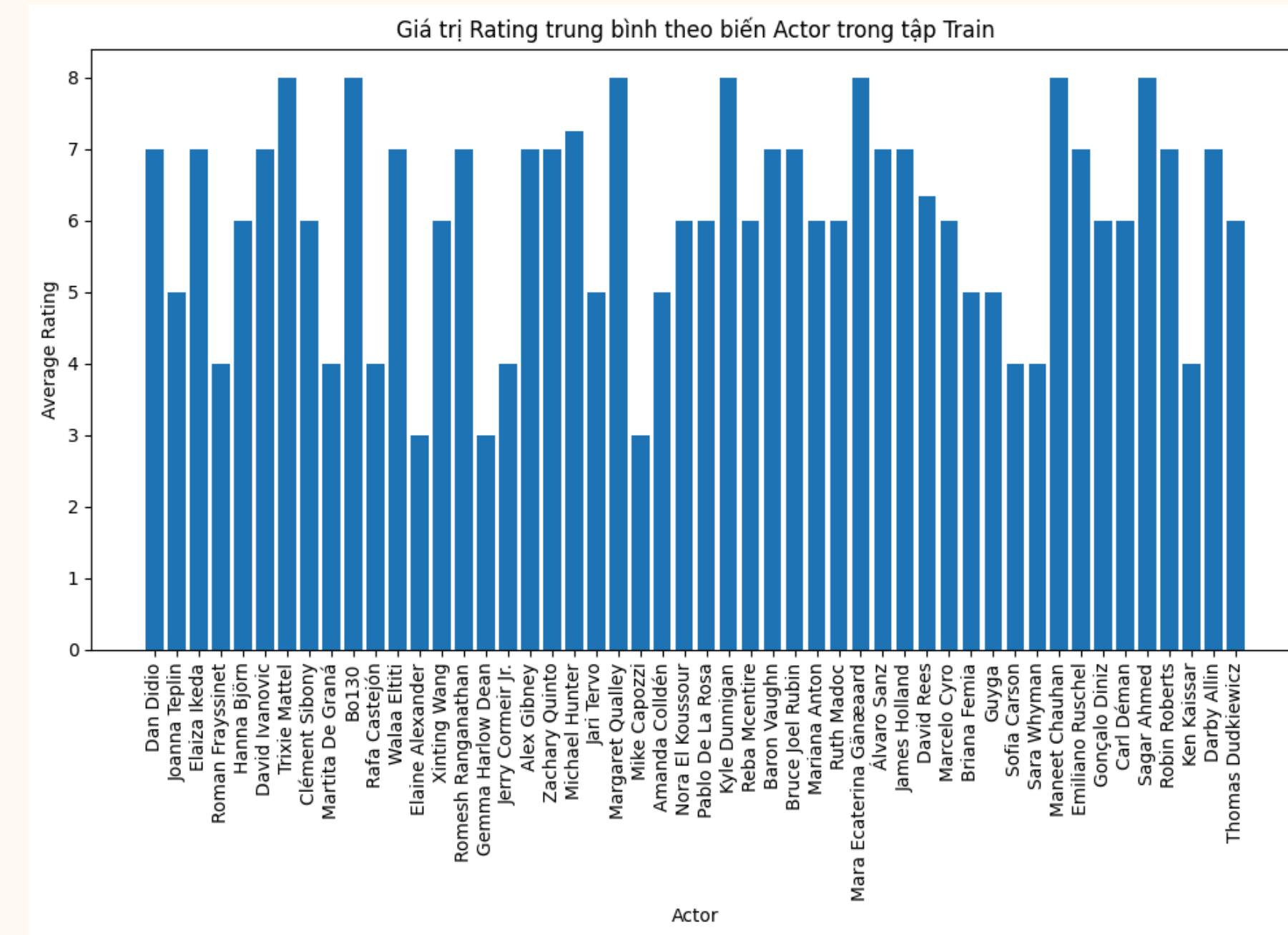


Điểm rating trung bình năm cao nhất là 9.4 và nằm trong khoảng thời gian từ 2000-2010

Điểm rating trung bình năm thấp nhất là 4.9 và nằm trong khoảng thời gian từ 1970-1980

Điểm rating trung bình giữa các khoảng thời gian khác có sự chênh lệch lớn

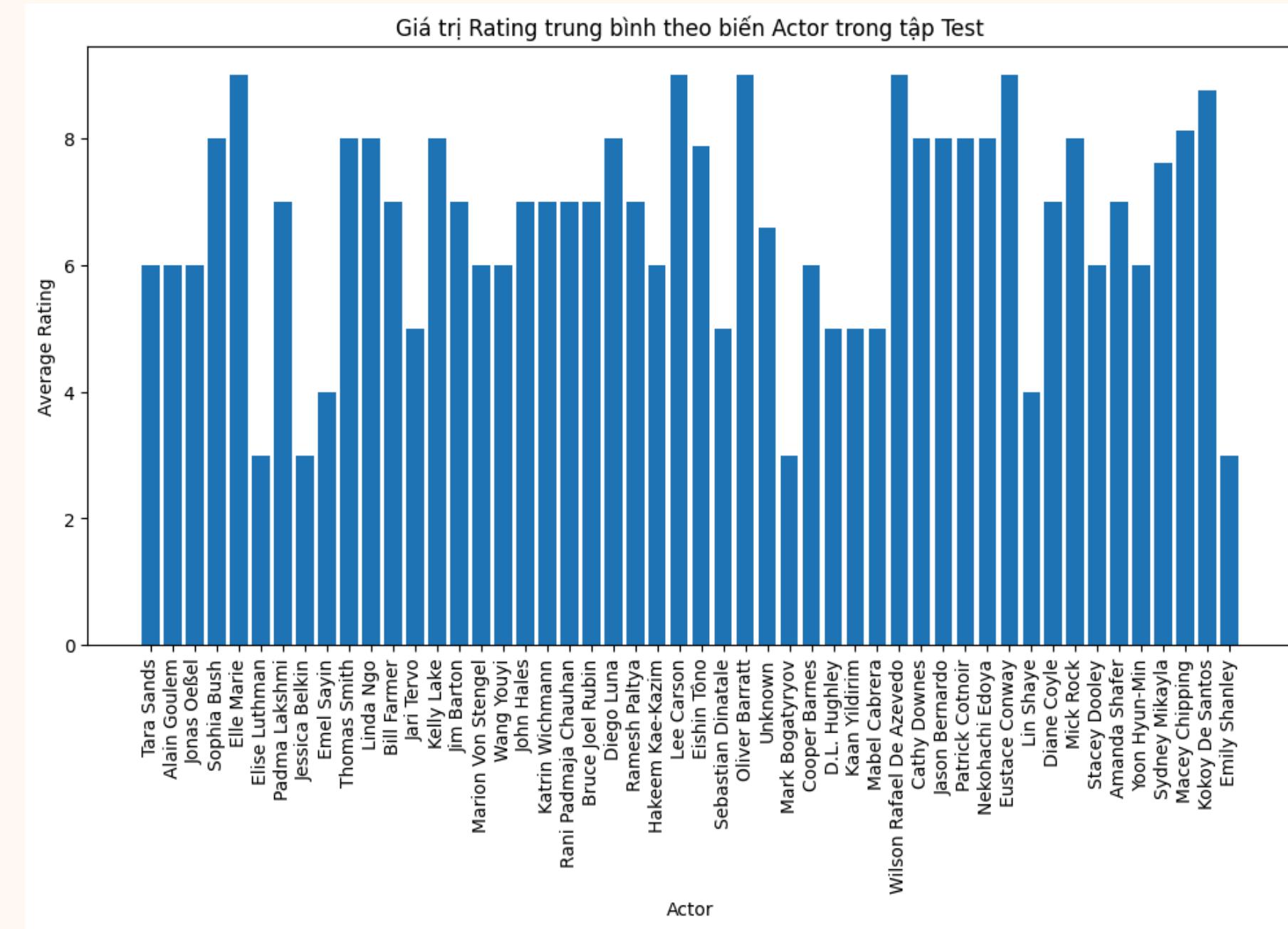
# Trong tập Train



Có khoảng cách lớn giữa Actor có điểm Rating trung bình cao nhất và thấp nhất (Trixie Mattel - Mike Capozzi)

Giữa những Actor khác cũng có sự chênh lệch điểm Rating lớn

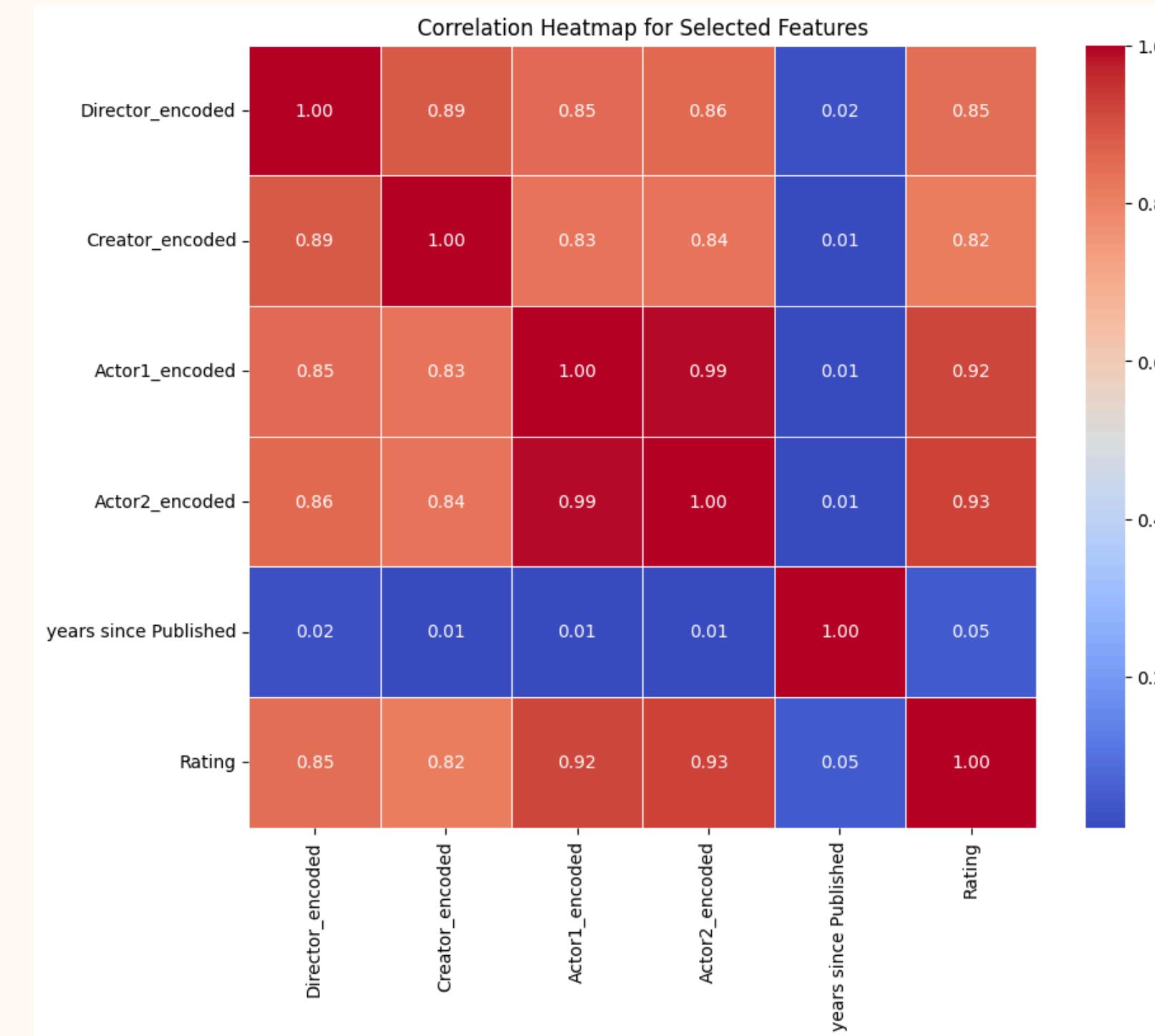
# Trong tập Test



Có khoảng cách lớn giữa Actor có điểm Rating trung bình cao nhất và thấp nhất (Ellie Marie - Mark Bogatyryov)

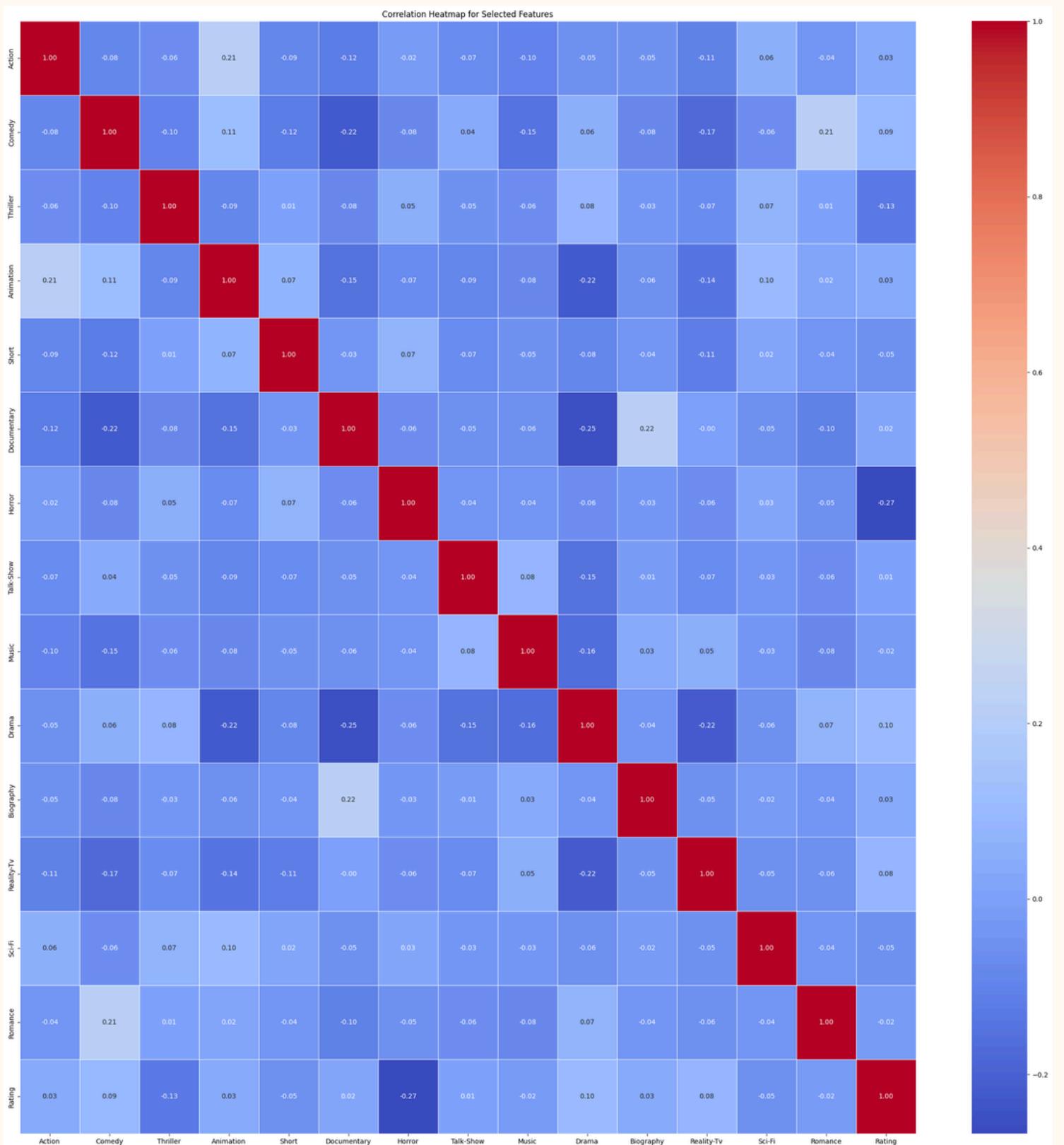
Giữa những Actor khác cũng có sự chênh lệch điểm Rating lớn

# Trong tập Train



Biểu đồ heatmap biểu diễn sự tương quan giữa các biến với biến Rating

# Trong tập Train

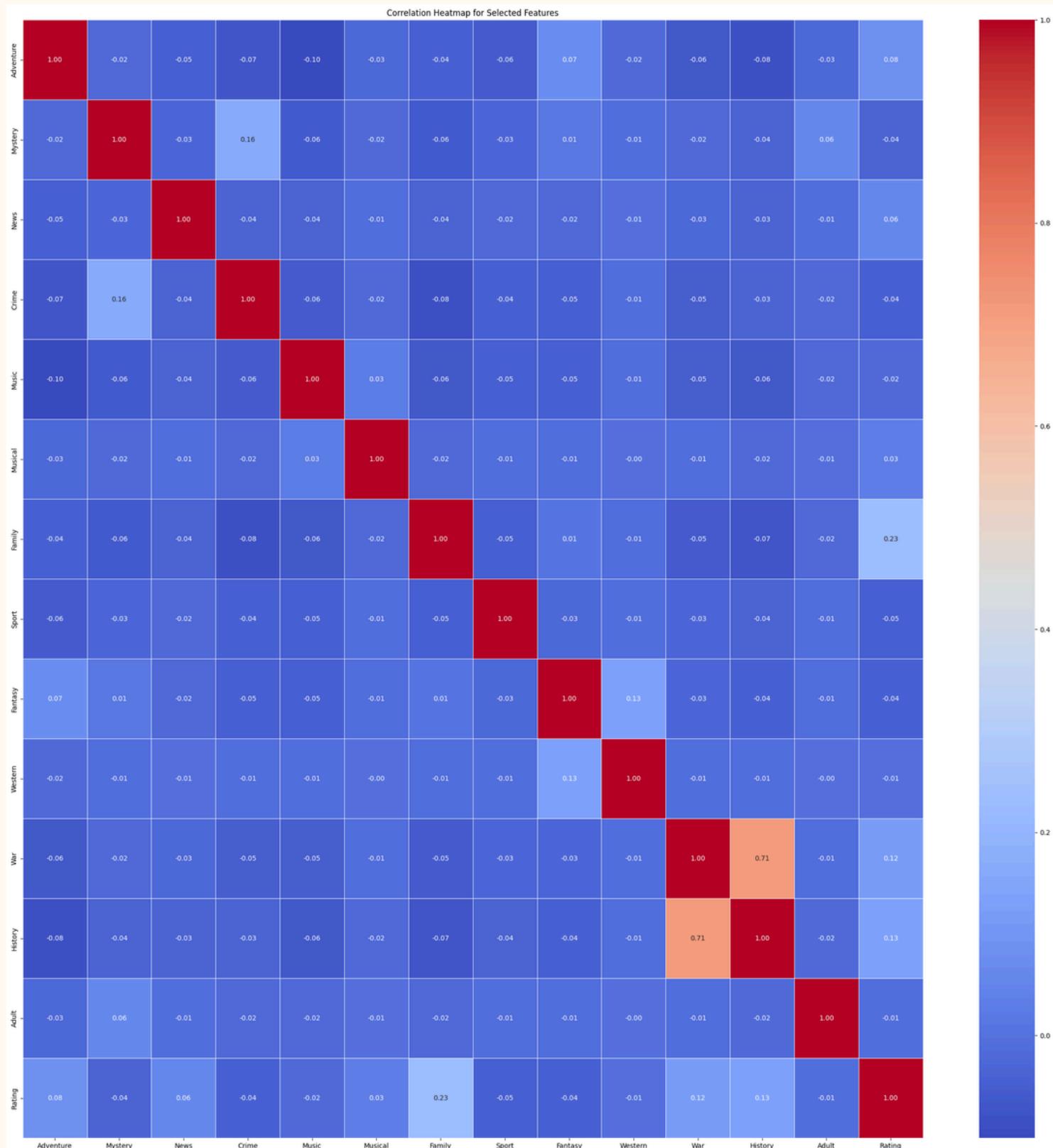


Biểu đồ heatmap biểu diễn sự tương quan giữa các biến Genre với biến Rating (1)

Mỗi biến Genre đại diện cho 1 thể loại phim và heatmap này có 15 biến Genre

- Action
- Comedy
- Thriller
- Animation
- Short
- Document
- Horror
- Talk-Show
- Music
- Drama
- Biography
- Reality-Tv
- Sci-Fi
- Romacne

# Trong tập Train

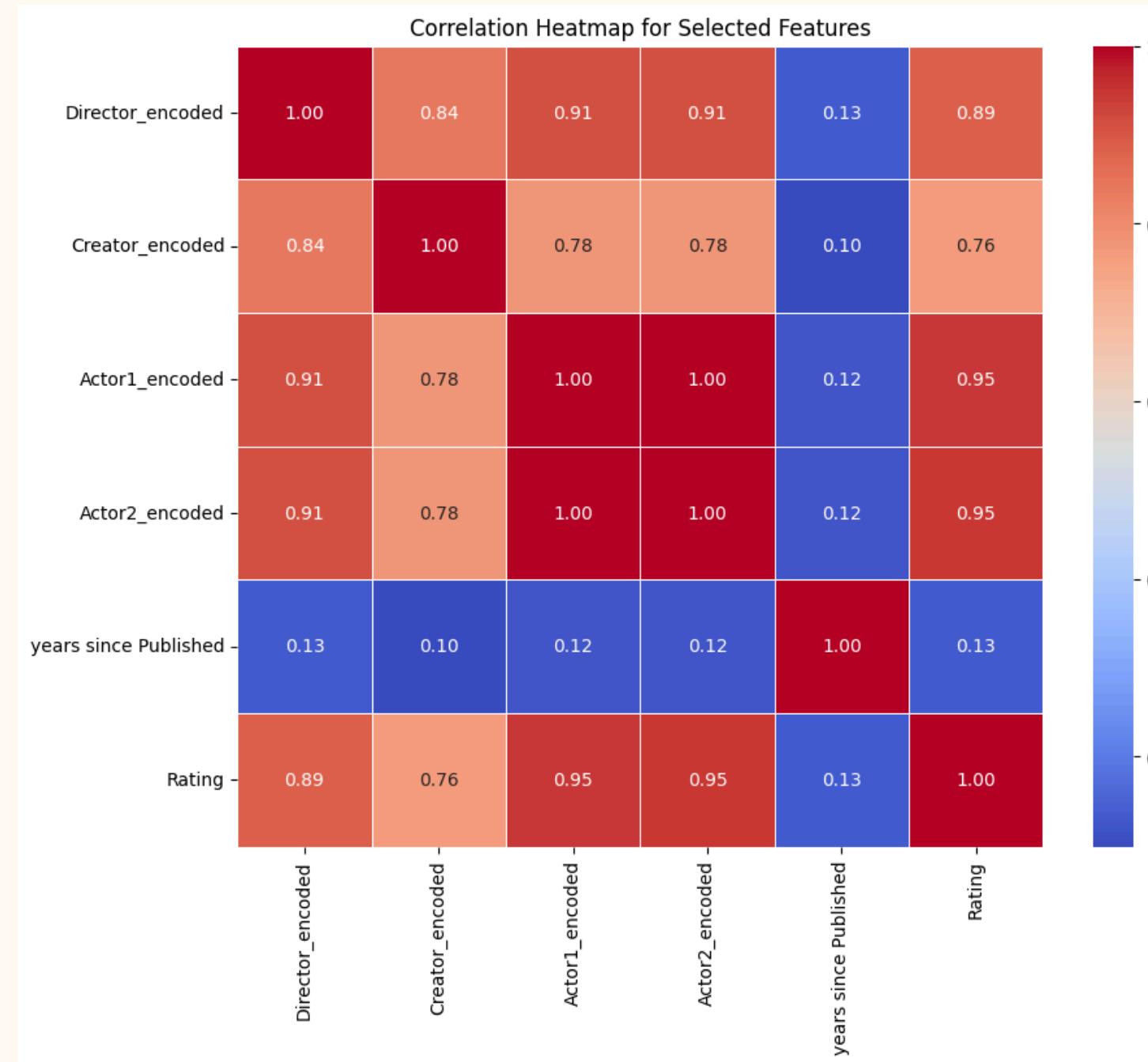


Biểu đồ heatmap biểu diễn sự tương quan giữa các biến Genre với biến Rating (2)

Mỗi biến Genre đại diện cho 1 thể loại phim và heatmap này có 14 biến Genre:

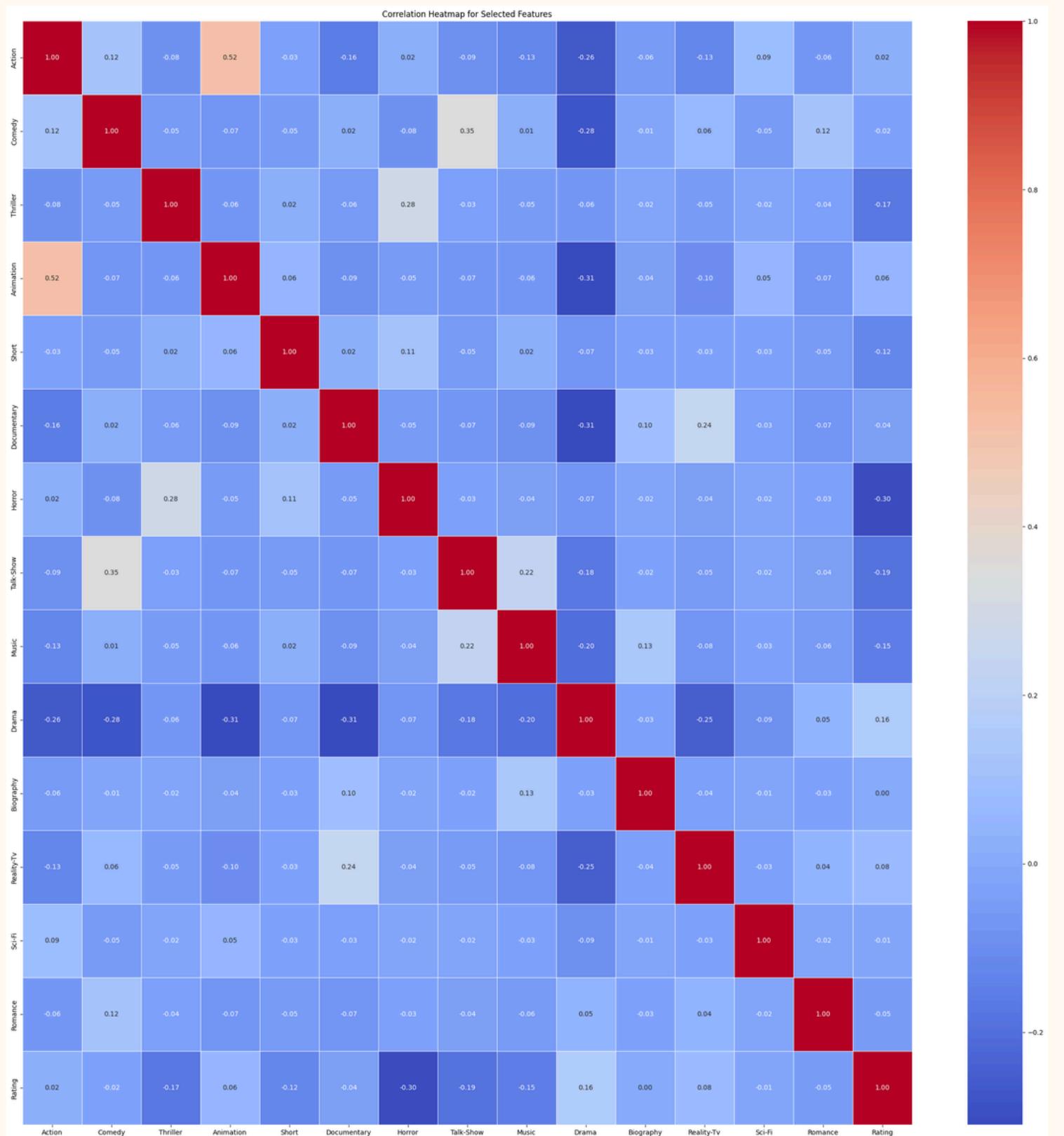
- Adventure
- Mystery
- News
- Crime
- Music
- Family
- Sport
- Fantasy
- Western
- War
- History
- Adult
- Rating

# Trong tập Test



Biểu đồ heatmap biểu diễn sự tương quan giữa các biến với biến Rating

# Trong tập Test

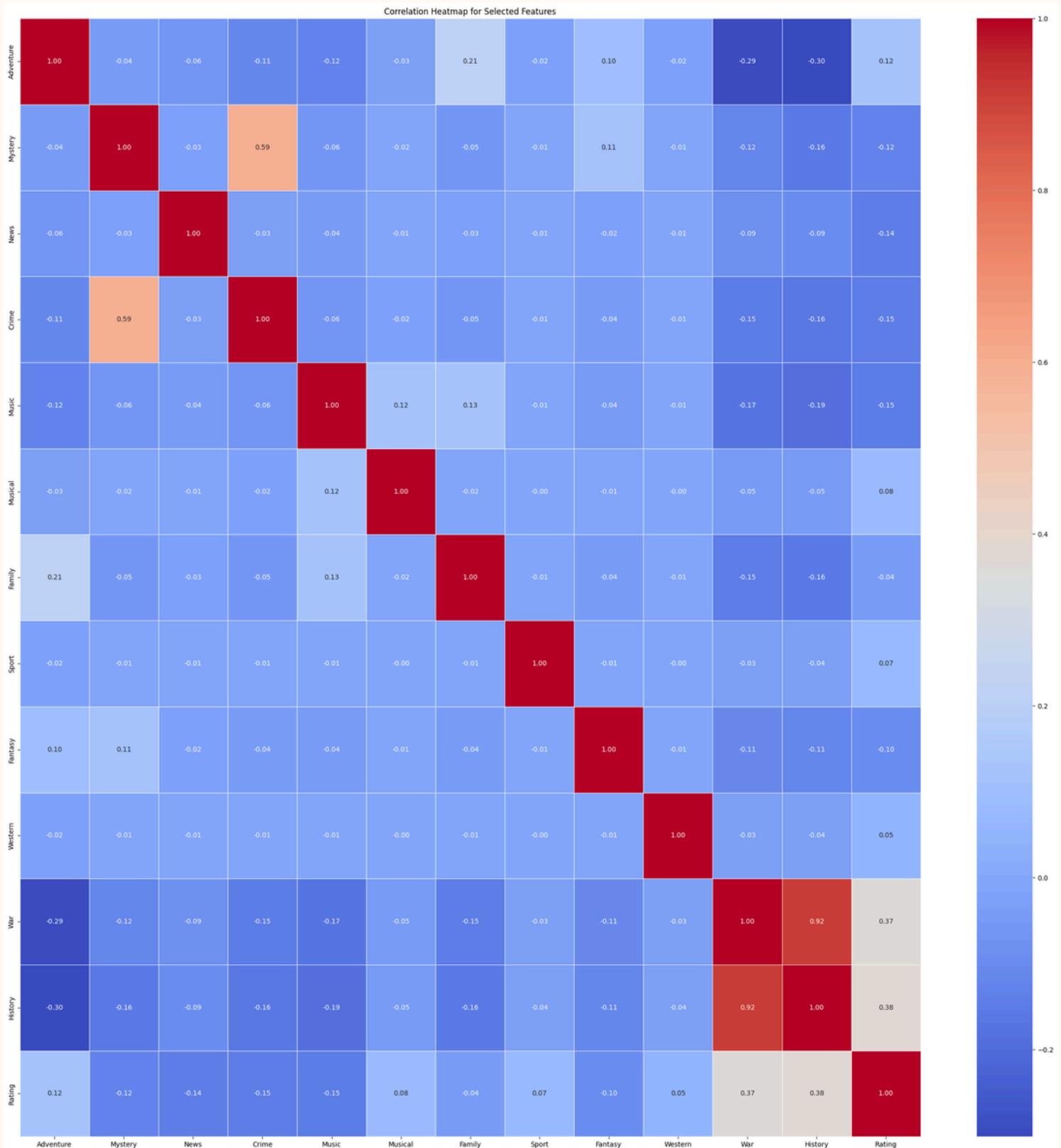


Biểu đồ heatmap biểu diễn sự tương quan giữa các biến Genre với biến Rating (1)

Mỗi biến Genre đại diện cho 1 thể loại phim và heatmap này có 15 biến Genre

- Action
- Comedy
- Thriller
- Animation
- Short
- Document
- Horror
- Talk-Show
- Music
- Drama
- Biography
- Reality-Tv
- Sci-Fi
- Romacne

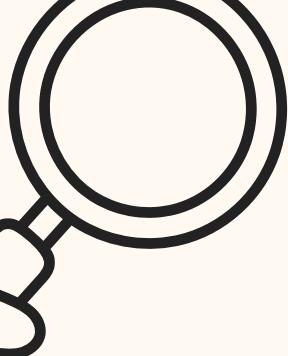
# Trong tập Test



Biểu đồ heatmap biểu diễn sự tương quan giữa các biến Genre với biến Rating (2)

Mỗi biến Genre đại diện cho 1 thể loại phim và heatmap này có 13 biến Genre:

- Adventure
- Mystery
- News
- Crime
- Music
- Musical
- Family
- Sport
- Fantasy
- Western
- War
- History
- Rating



### 3. Trích xuất đặc trưng

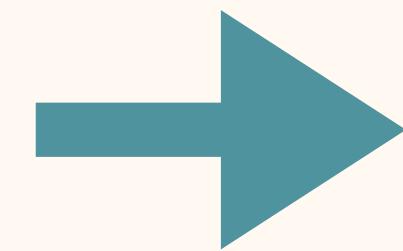
Dành cho bài toán phân loại



# Dành cho bài toán phân loại

- Cột Genre : Chứa danh sách các thể loại phim  
-> Thực hiện Multi-hot encoding

Genre
['Action', 'Comedy', 'Thriller']
['Comedy', 'Drama']
['Animation', 'Short']
['Documentary']
['Documentary', 'Short']
['Short', 'Horror']
['Comedy']
['Short', 'Drama']
['Comedy', 'Talk-Show']
['Music']

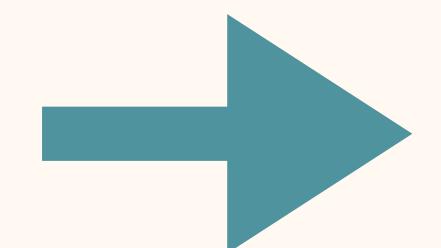


Action	Adult	...	None	Reality-Tv	Romance	Sci-Fi	Short	Sport	Talk-Show	Thriller	War	Western
1	0	...	0	0	0	0	0	0	0	1	0	0
0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	...	0	0	0	0	1	0	0	0	0	0
0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	...	0	0	0	0	1	0	0	0	0	0

# Dành cho bài toán phân loại

Cột DatePublished: Chuyển giá trị từ ngày/tháng/năm thành số năm kể từ ngày sản xuất đến thời điểm hiện tại. Chuẩn hóa lại dữ liệu bằng kĩ thuật chuẩn hóa MinMax.

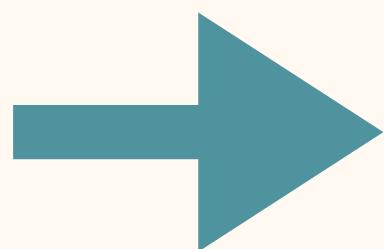
DatePublished	years since Published
2020-06-11	0.023339
2021-06-11	0.053357
2019-11-01	0.050018
2020-06-12	0.418962
2020-06-12	0.072149
2020-06-11	
2020-06-10	
2020-06-08	



# Dành cho bài toán phân loại

Các cột Director, Creator, Actor 1, Actor 2: Thực hiện target encoding và chuẩn hóa lại dữ liệu bằng kĩ thuật chuẩn hóa MinMax.

Director	Creator	Actor 1	Actor 2
Unknown	Unknown	Raleigh Avery	Eustace Conway
Macarena Astorga	Sandra García Nieto	Javier Rey	Paz Vega
Unknown	Unknown	Unknown	Unknown
Unknown	Unknown	Blake Douglas	Albie Robles



Director_encoded	Creator_encoded	Actor1_encoded	Actor2_encoded
0.700000	0.640000	0.700000	0.700000
0.777778	0.700000	0.750000	0.750000
0.744444	0.693333	0.678679	0.678679
0.695305	0.643068	0.700000	0.700000
0.600000	0.643068	0.600000	0.600000

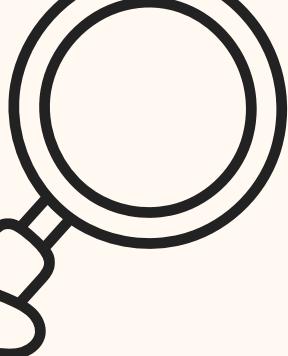
# Dành cho bài toán phân loại

Lựa chọn đặc trưng:

Chỉ lựa chọn những đặc trưng có độ tương quan với biến mục tiêu (Rating) lớn hơn 0.2.

Các đặc trưng được lựa chọn:

- Family: bộ phim đó có thuộc thể loại Family hay không ?
- Horror: bộ phim đó có thuộc thể loại Horror hay không ?
- Director\_encoded: đạo diễn
- Creator\_encoded: biên tập
- Actor1\_encoded: diễn viên 1
- Actor2\_encoded: diễn viên 2



### 3. Trích xuất đặc trưng

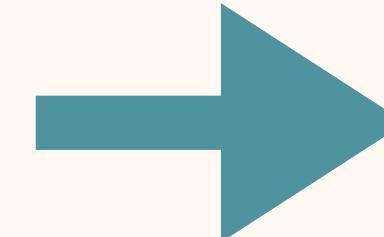
Dành cho bài toán  
phân cụm



# Dành cho bài toán phân cụm

Cột Genre : Chứa danh sách các thể loại phim  
-> Thực hiện Multi-hot encoding

Genre	
['Action', 'Comedy', 'Thriller']	
['Comedy', 'Drama']	
['Animation', 'Short']	
['Documentary']	
['Documentary', 'Short']	
['Short', 'Horror']	
['Comedy']	
['Short', 'Drama']	
['Comedy', 'Talk-Show']	
['Music']	



Action	Adult	...	None	Reality-Tv	Romance	Sci-Fi	Short	Sport	Talk-Show	Thriller	War	Western
1	0	...	0	0	0	0	0	0	0	1	0	0
0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	...	0	0	0	0	1	0	0	0	0	0
0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	...	0	0	0	0	1	0	0	0	0	0

# Dành cho bài toán phân cụm

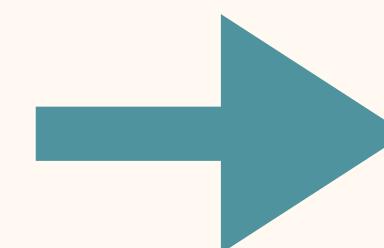
Cột DatePublished: Chuyển giá trị từ ngày/tháng/năm thành số năm kể từ ngày sản xuất đến thời điểm hiện tại. Chuẩn hóa lại dữ liệu bằng z-score.

DatePublished	years since Published
2022-08-05	1.812457
2020-07-02	3.904175
2020-09-25	3.671458
1995-01-10	29.379877
NaN	5.213569
2020-08-01	3.822040
2021-09-08	2.718686
2020-09-10	3.712526

# Dành cho bài toán phân cụm

Cột Genre : Chứa danh sách các thể loại phim  
-> Thực hiện Multi-hot encoding

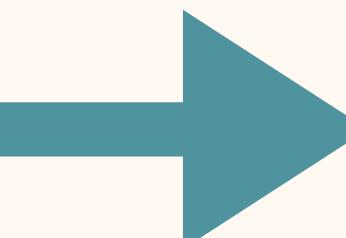
Genre
['Action', 'Comedy', 'Thriller']
['Comedy', 'Drama']
['Animation', 'Short']
['Documentary']
['Documentary', 'Short']
['Short', 'Horror']
['Comedy']
['Short', 'Drama']
['Comedy', 'Talk-Show']
['Music']



Action	Adult	...	None	Reality-Tv	Romance	Sci-Fi	Short	Sport	Talk-Show	Thriller	War	Western
1	0	...	0	0	0	0	0	0	0	1	0	0
0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	...	0	0	0	0	1	0	0	0	0	0
0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	...	0	0	0	0	1	0	0	0	0	0

# Dành cho bài toán phân cụm

Các cột Director, Creator, Actor 1, Actor 2: Thực hiện frequency encoding và chuẩn hóa dữ liệu theo z-score.

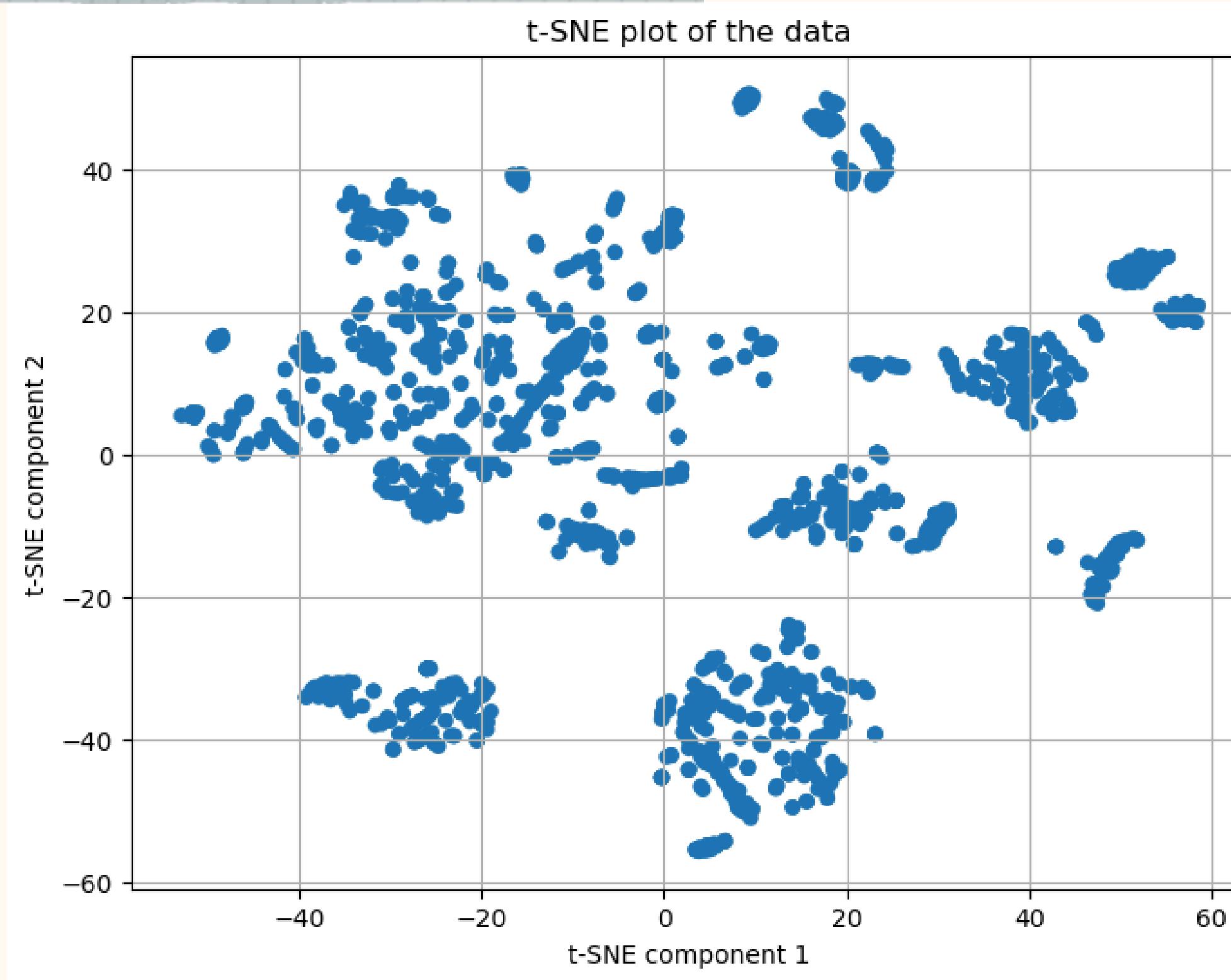


Director	Creator	Actor 1	Actor 2	Director_freq	Creator_freq	Actor1_freq	Actor2_freq
David Leitch	Zak Olkewicz	Brad Pitt	Joey King	-0.745861	-0.798474	-0.535346	-0.498119
Erin Ehrlich	Sarah Watson	Katie Stevens	Aisha Dee	-0.745861	-0.795489	-0.521322	-0.483967
Thiago Martins De Melo	Thiago Martins De Melo	Unknown	Unknown	-0.745861	-0.798474	2.564013	2.629480
Unknown	Unknown	Christopher Chacon	Lars Svedberg	1.429641	1.300233	-0.535346	-0.498119

# Dành cho bài toán phân cụm

- Sử dụng t-SNE để giảm chiều dữ liệu xuống còn 2 chiều trước khi thực hiện phân cụm.
- Các đặc trưng được lựa chọn để thực hiện t-SNE:
  - Rating: đánh giá của bộ phim
  - years since Published : số năm kể từ ngày phát hành.
  - Biến 1-0 cho từng thể loại (có 28 thể loại)
  - Director\_freq: đạo diễn
  - Creator\_freq: biên tập
  - Actor1\_freq: diễn viên 1
  - Actor2\_freq: diễn viên 2

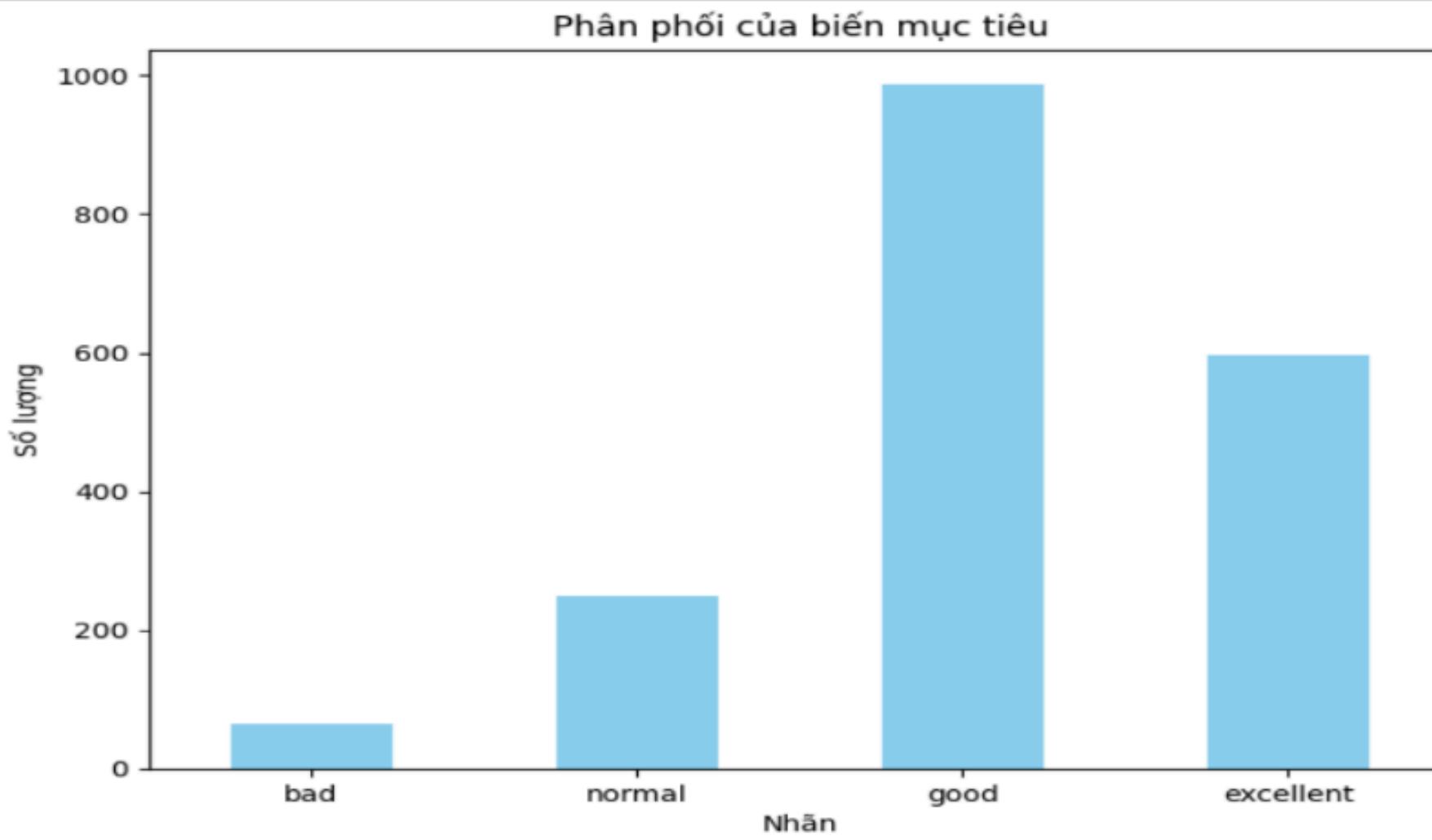
# Sử dụng T-SNE để giảm chiều dữ liệu



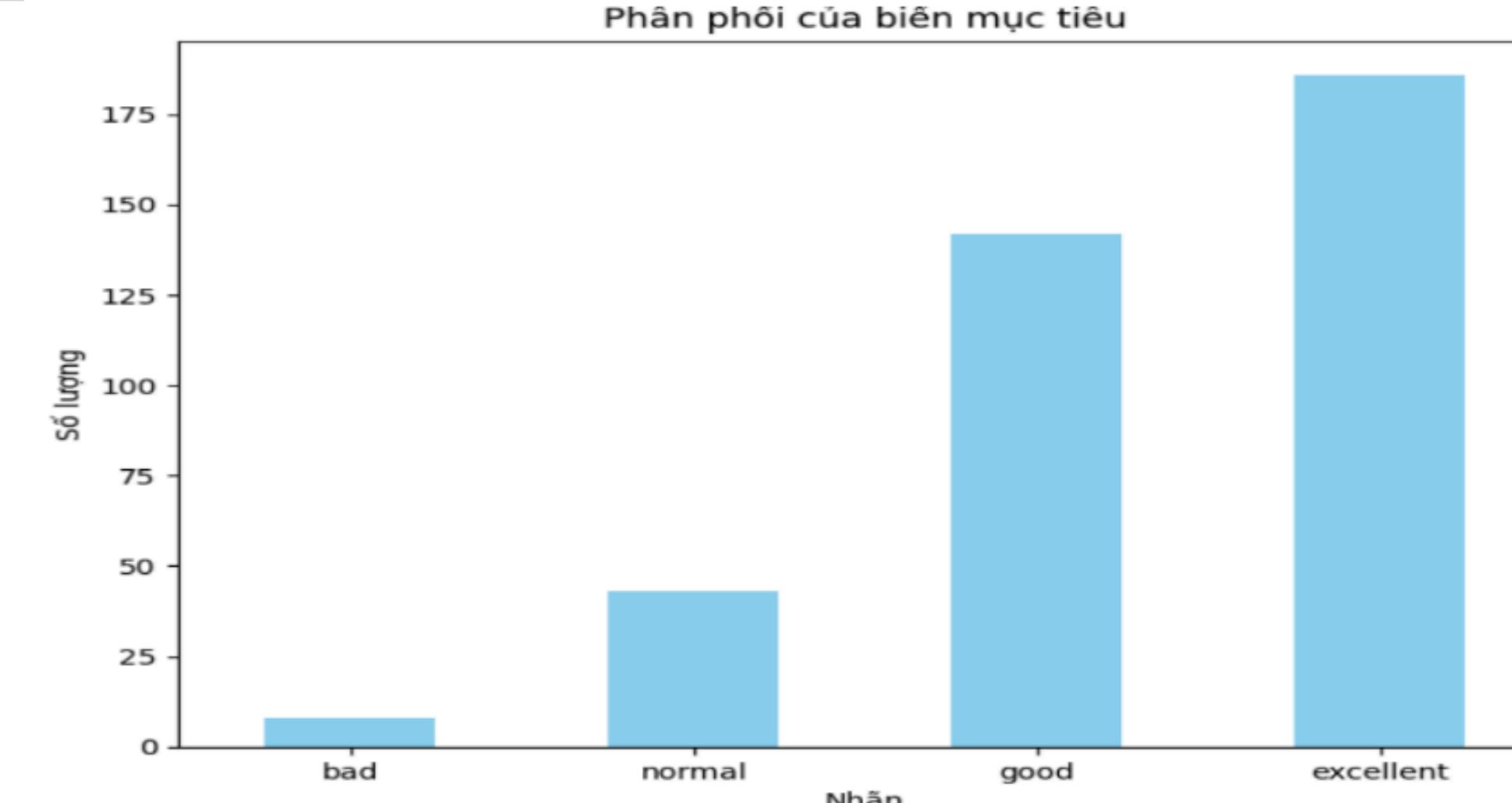
Phân bố điểm dữ liệu ở dạng 2 chiều sau khi thực hiện t-SNE

# MÔ HÌNH PHÂN LOAI RATING PHIM

# TỔNG QUAN VỀ PHÂN LOẠI RATING PHIM



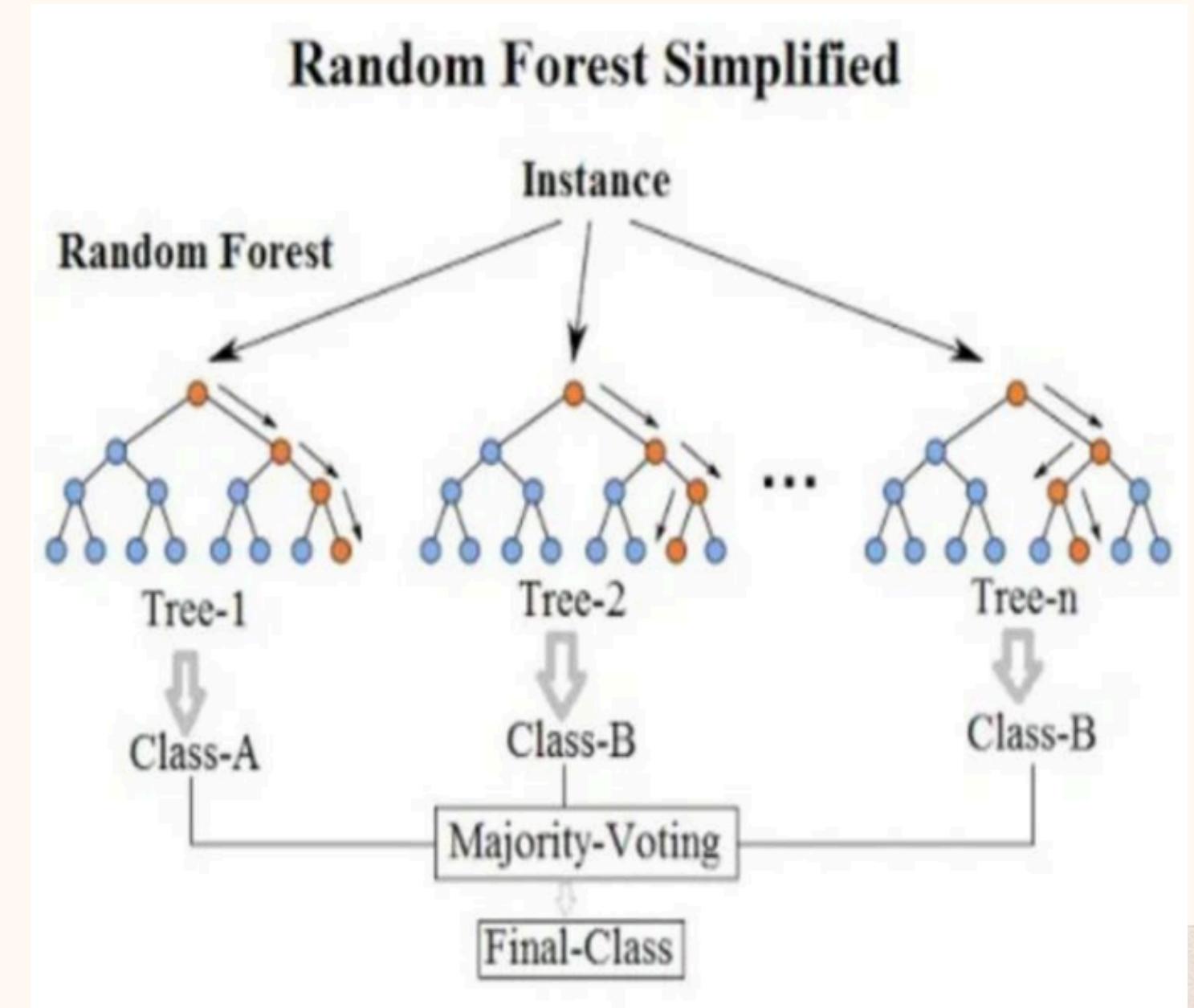
Biến rating trên tập huấn luyện



Biến rating trên tập kiểm thử

# Random forest

- Random Forest là một kỹ thuật học máy thuộc nhóm các phương pháp ensemble learning
- Nó kết hợp nhiều cây quyết định (decision trees) để tạo ra một mô hình mạnh mẽ hơn.
- Nguyên lý của Random Forest là kết hợp nhiều cây quyết định yếu (weak learners) để giảm thiểu vấn đề overfitting và tăng độ chính xác của dự đoán.
- Mỗi cây quyết định trong Random Forest được xây dựng từ một mẫu dữ liệu khác nhau, lấy ngẫu nhiên từ tập huấn luyện (bagging)



# Random forest

Sử dụng RandomSearchSV để tìm các siêu tham số cho mô hình:

- n\_estimators: 80
- max\_features: log2
- max\_depth: 7
- min\_samples\_split: 2
- min\_samples\_leaf: 4
- criterion: gini
- bootstrap: True

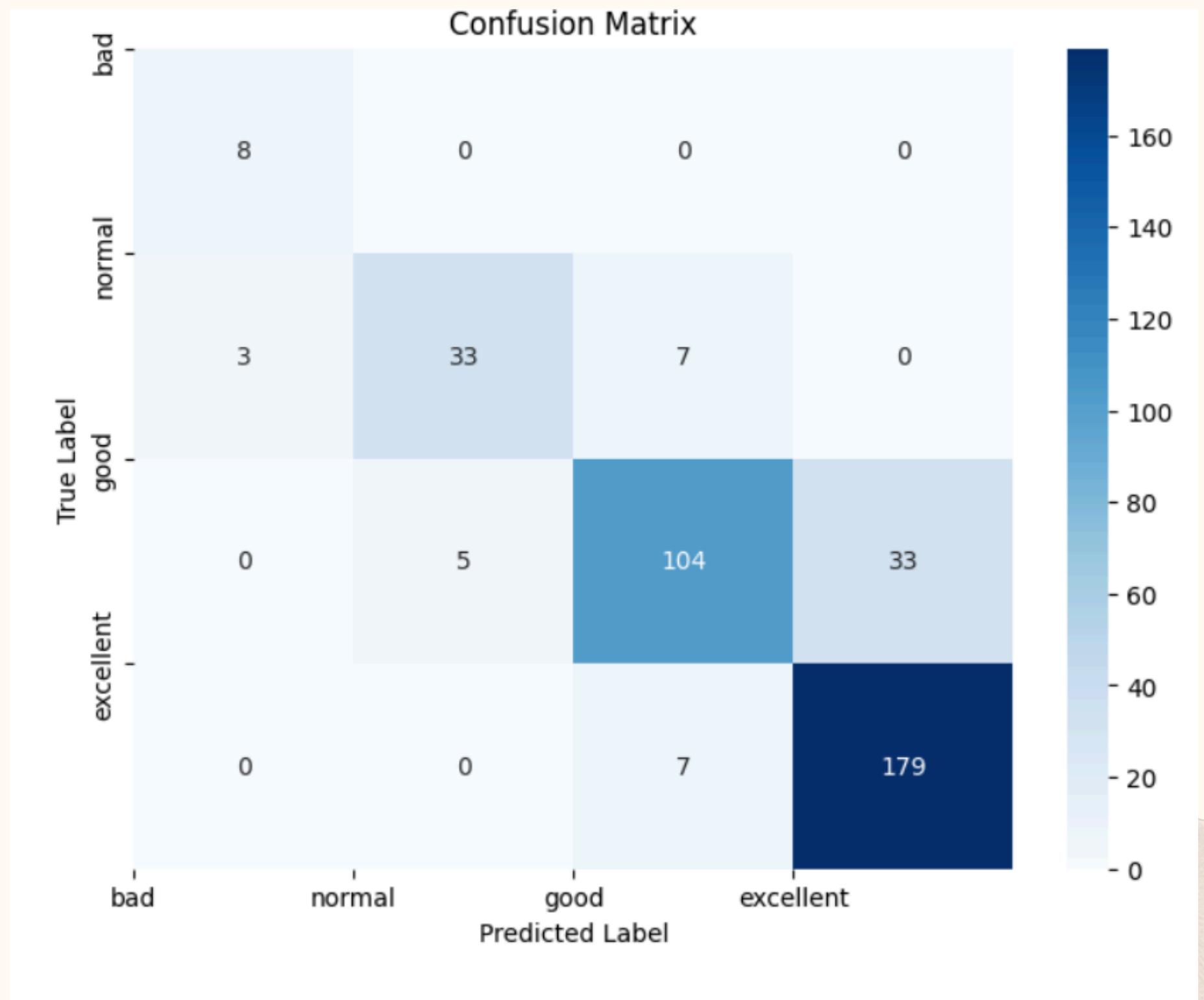
```
Random forest accuracy on training set: 91.36453526697429
```

```
Random forest accuracy on validation set: 88.94736842105263
```

```
Random forest accuracy on test set: 85.4881266490765
```

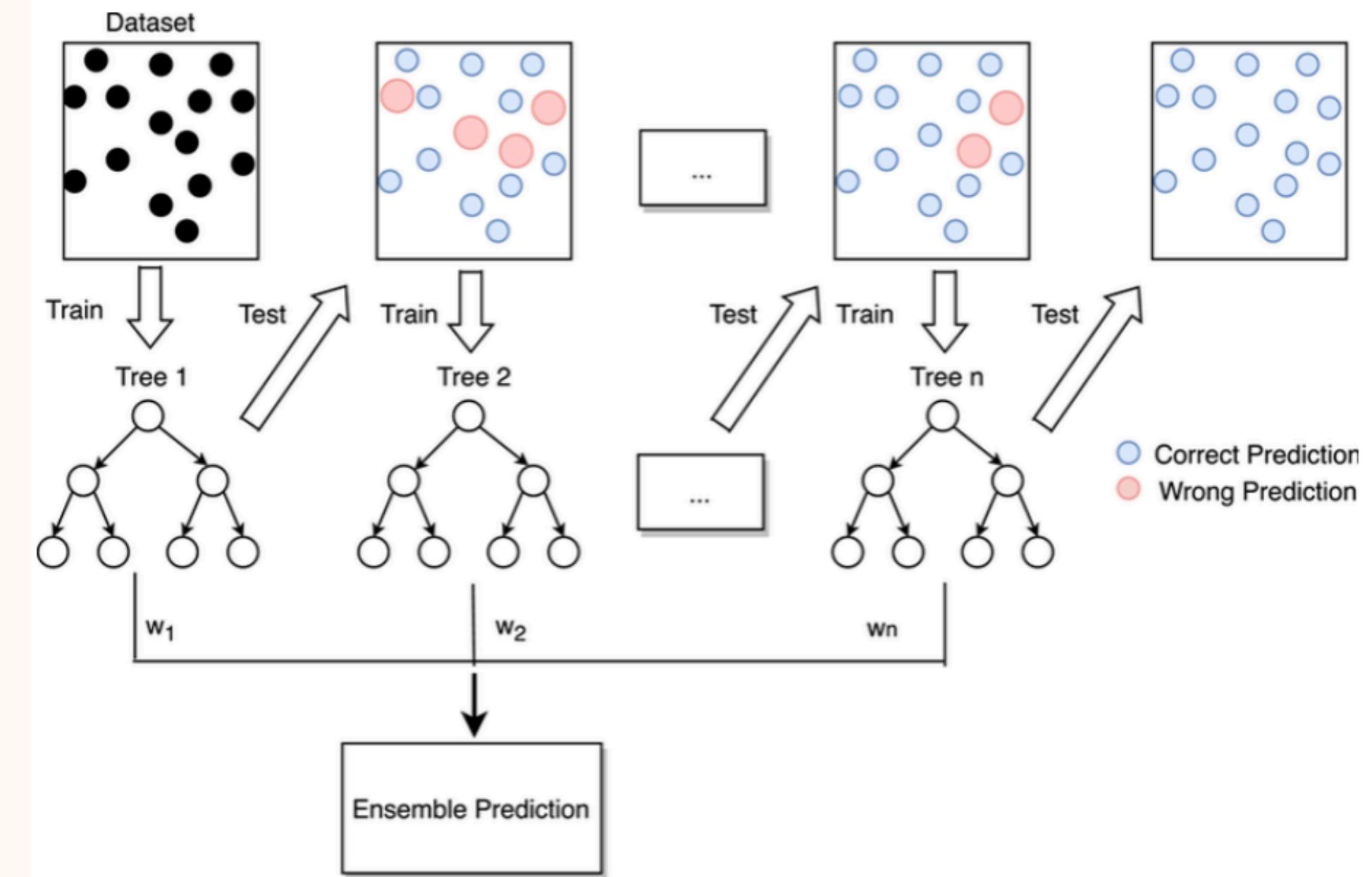
# Random forest

Ma trận nhầm lẫn trên tập kiểm thử



# Gradient Boosting

- Gradient Boosting là một kỹ thuật học máy khác thuộc nhóm các phương pháp ensemble learning.
- Trong đó các mô hình yếu được xây dựng theo tuần tự và mỗi mô hình mới cố gắng sửa chữa lỗi của mô hình trước đó.
- Kỹ thuật này thường sử dụng cây quyết định nhỏ (shallow trees) làm mô hình cơ sở (base learners).



# Gradient Boosting

Sử dụng RandomSearchSV để tìm các siêu tham số cho mô hình:

- n\_estimators: 10
- learning\_rate: 0.1
- max\_depth: 4
- min\_samples\_split: 10
- min\_samples\_leaf: 4
- subsample: 0.8

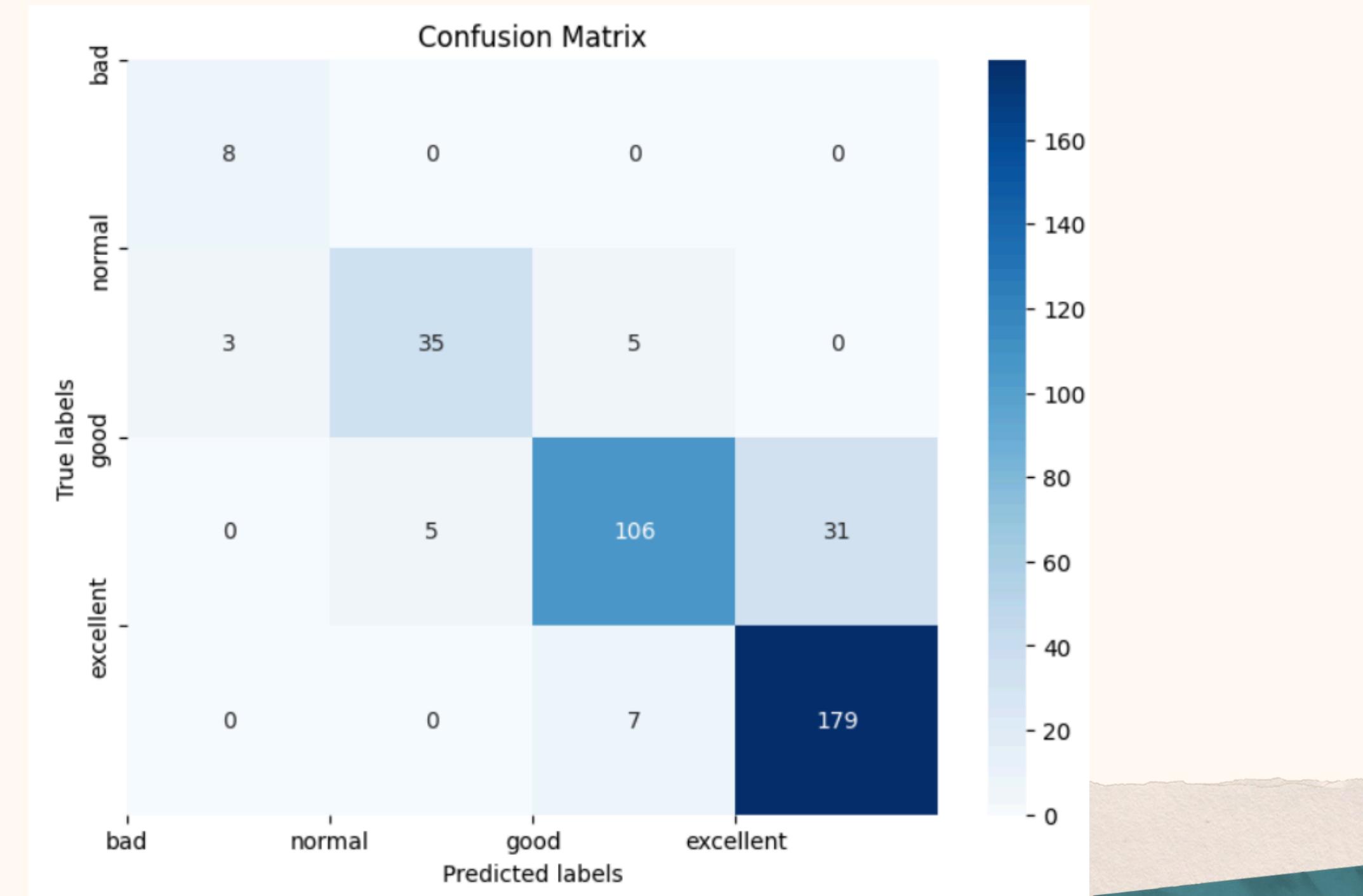
Gradient Boosting accuracy on training set: 91.36453526697429

Gradient Boosting accuracy on validation set: 88.68421052631578

Gradient Boosting accuracy on test set: 86.54353562005277

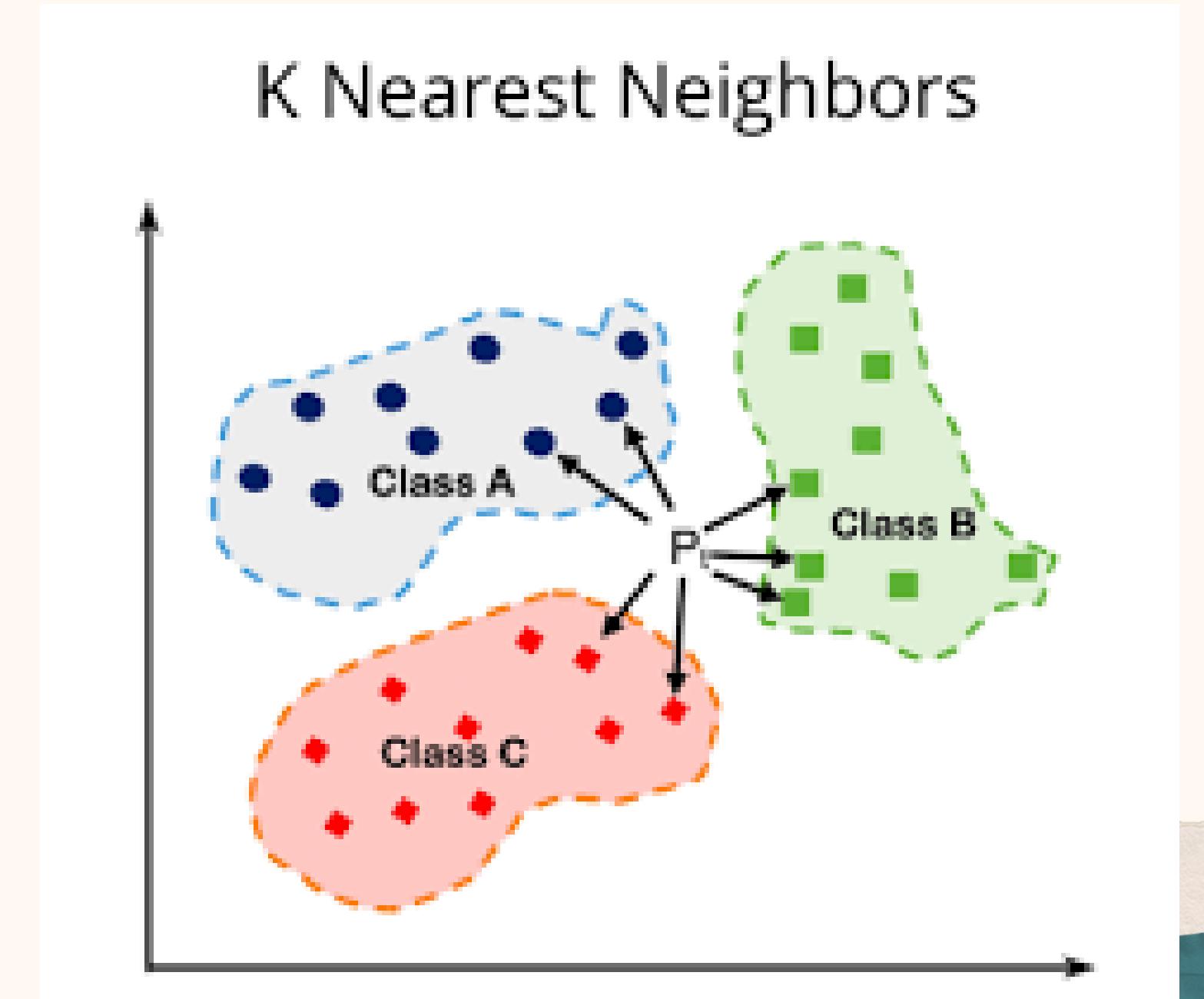
# Gradient Boosting

Ma trận nhầm lẫn trên tập  
kiểm thử



# KNN

- K-Nearest Neighbors (KNN) là một thuật toán đơn giản nhưng mạnh mẽ, hoạt động dựa trên nguyên tắc tìm các điểm dữ liệu gần nhất trong không gian đặc trưng để đưa ra dự đoán.
- Thuật toán này không yêu cầu mô hình huấn luyện phức tạp và thường được gọi là thuật toán "lazy learning" vì nó không học một mô hình tường minh mà chỉ lưu trữ toàn bộ dữ liệu huấn luyện.
- Khi cần phân loại một mẫu mới, KNN tính khoảng cách từ mẫu này đến tất cả các mẫu trong tập huấn luyện.



# KNN

Sử dụng RandomSearchSV để tìm các siêu tham số cho mô hình:

- weights: distance
- n\_neighbors: 9

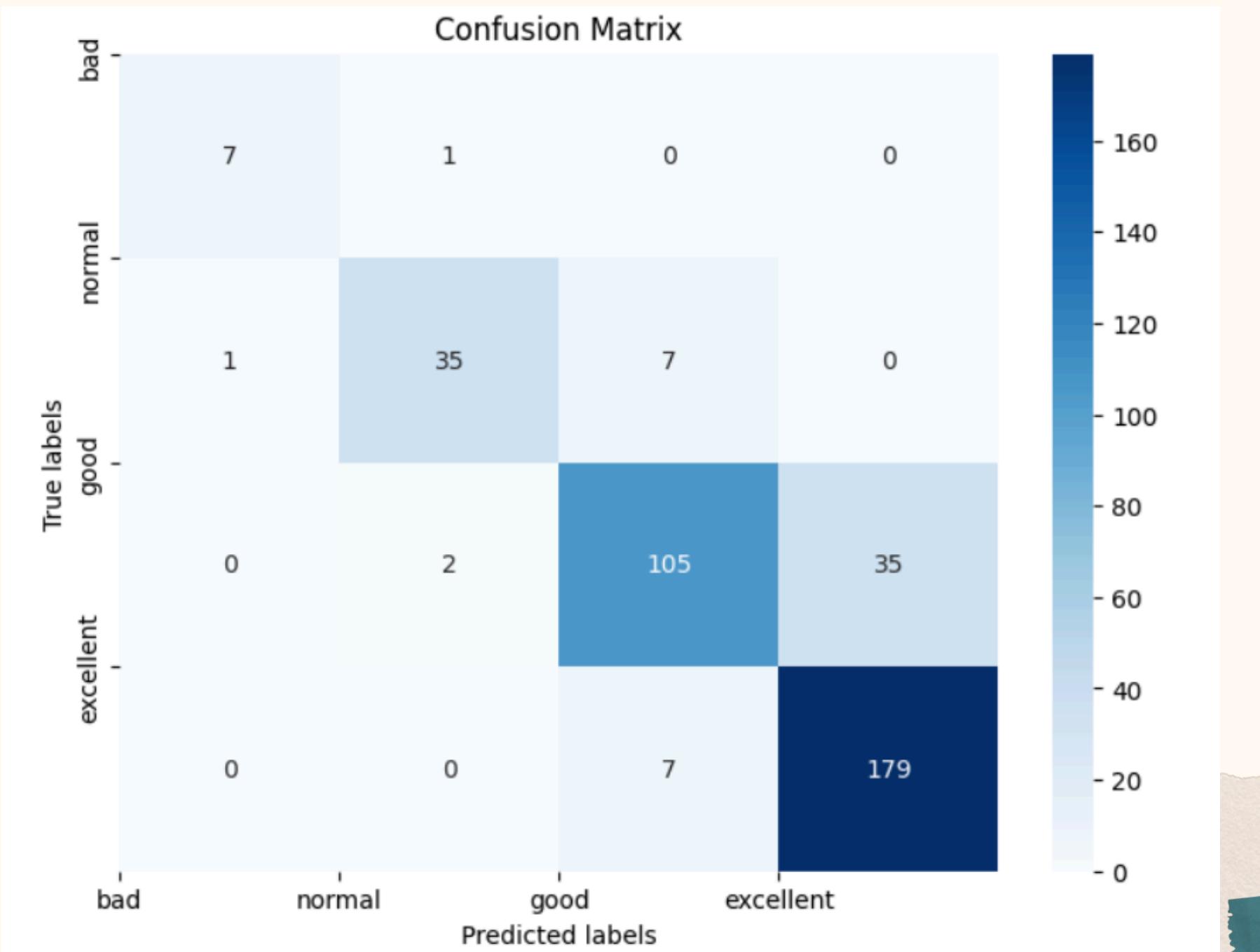
KNN accuracy on training set: 92.089650626236

KNN accuracy on validation set: 86.57894736842105

KNN accuracy on test set: 86.01583113456465

# KNN

Ma trận nhầm lẫn trên tập  
kiểm thử



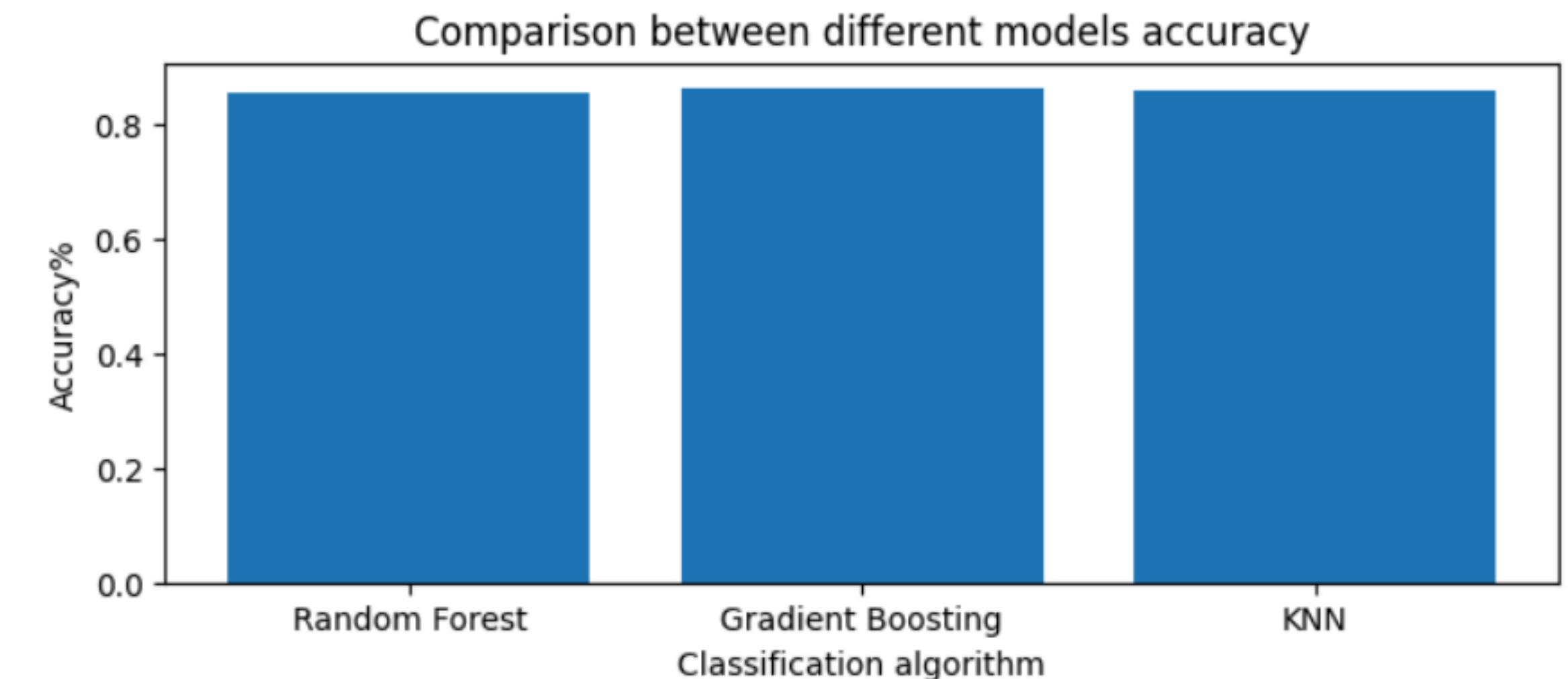
# SO SÁNH 3 MODEL

Gradient Boosting là mô hình tốt nhất trong ba mô hình, với độ chính xác cao nhất.

Tuy nhiên, KNN và Random Forest cũng đạt kết quả rất cao, cho thấy chúng đều là những mô hình mạnh mẽ và đáng tin cậy. Sự chênh lệch nhỏ giữa các mô hình cho thấy tất cả đều đủ tốt để sử dụng.

=> Chọn Gradient Boosting làm mô hình chính

[0.855, 0.865, 0.86]

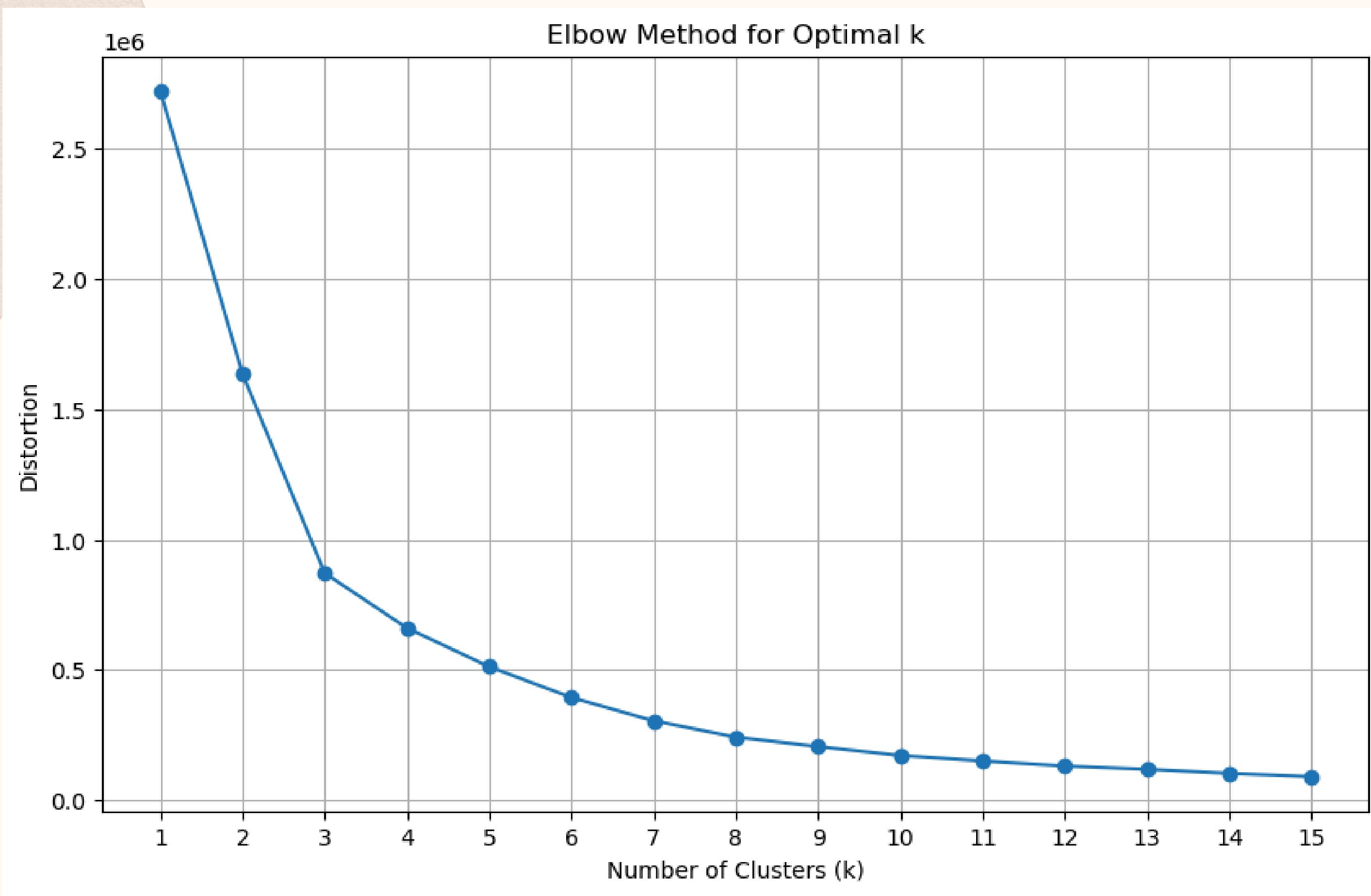


# MÔ HÌNH PHẦN CỤM PHIM

# K-means clustering

- Thuật toán phân cụm phi giám sát phổ biến được sử dụng để nhóm các điểm dữ liệu có cùng đặc điểm.
- Tối ưu hóa tổng bình phương khoảng cách từ mỗi điểm dữ liệu đến tâm của cụm mà nó thuộc về
  - Cần xác định số cụm ban đầu  
=> Sử dụng phương pháp Elbow để tìm số cụm K tối ưu:

### Elbow Method for Optimal k

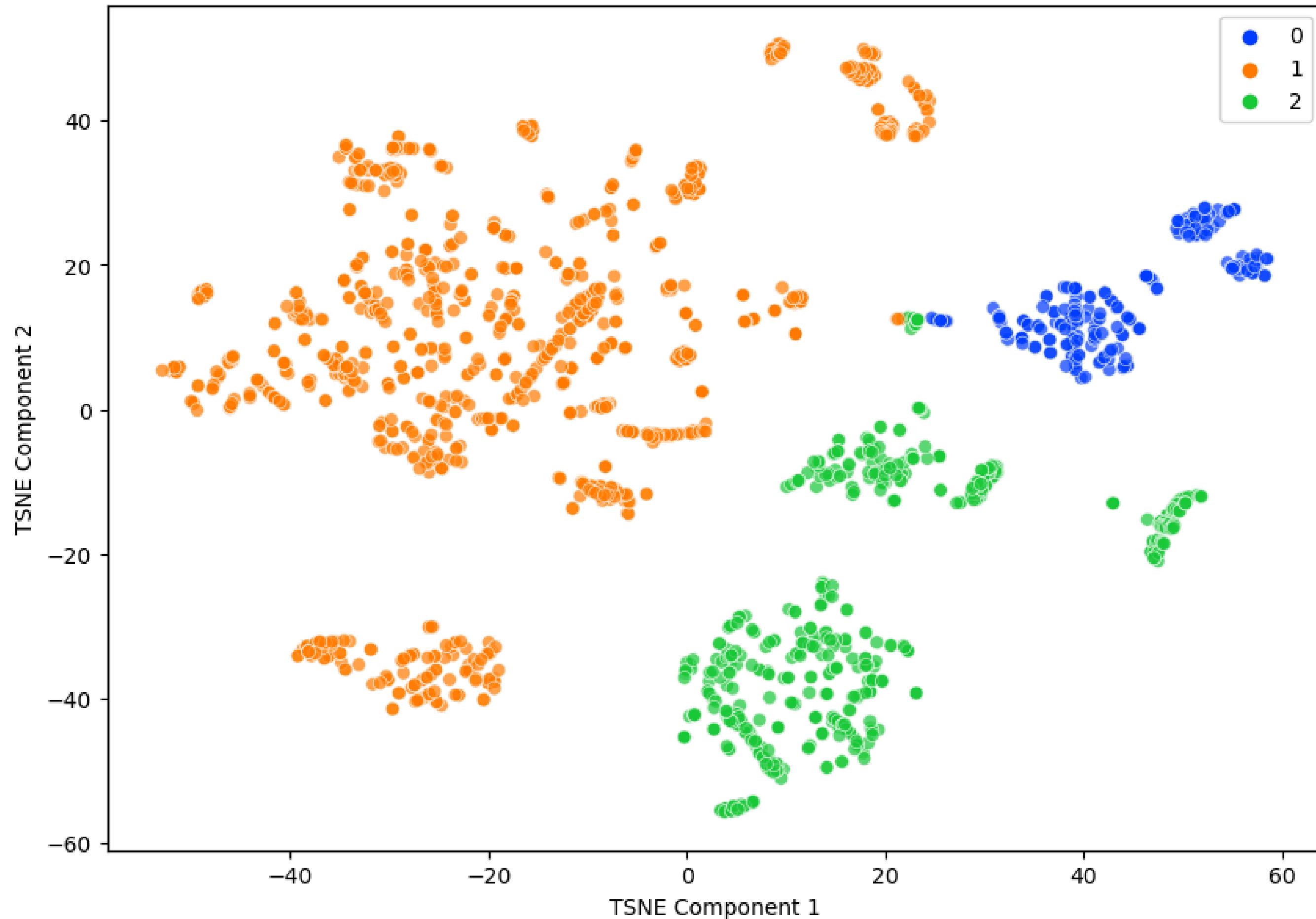


Đồ thị hiệu quả phân cụm theo số cụm k.

# K-means clustering

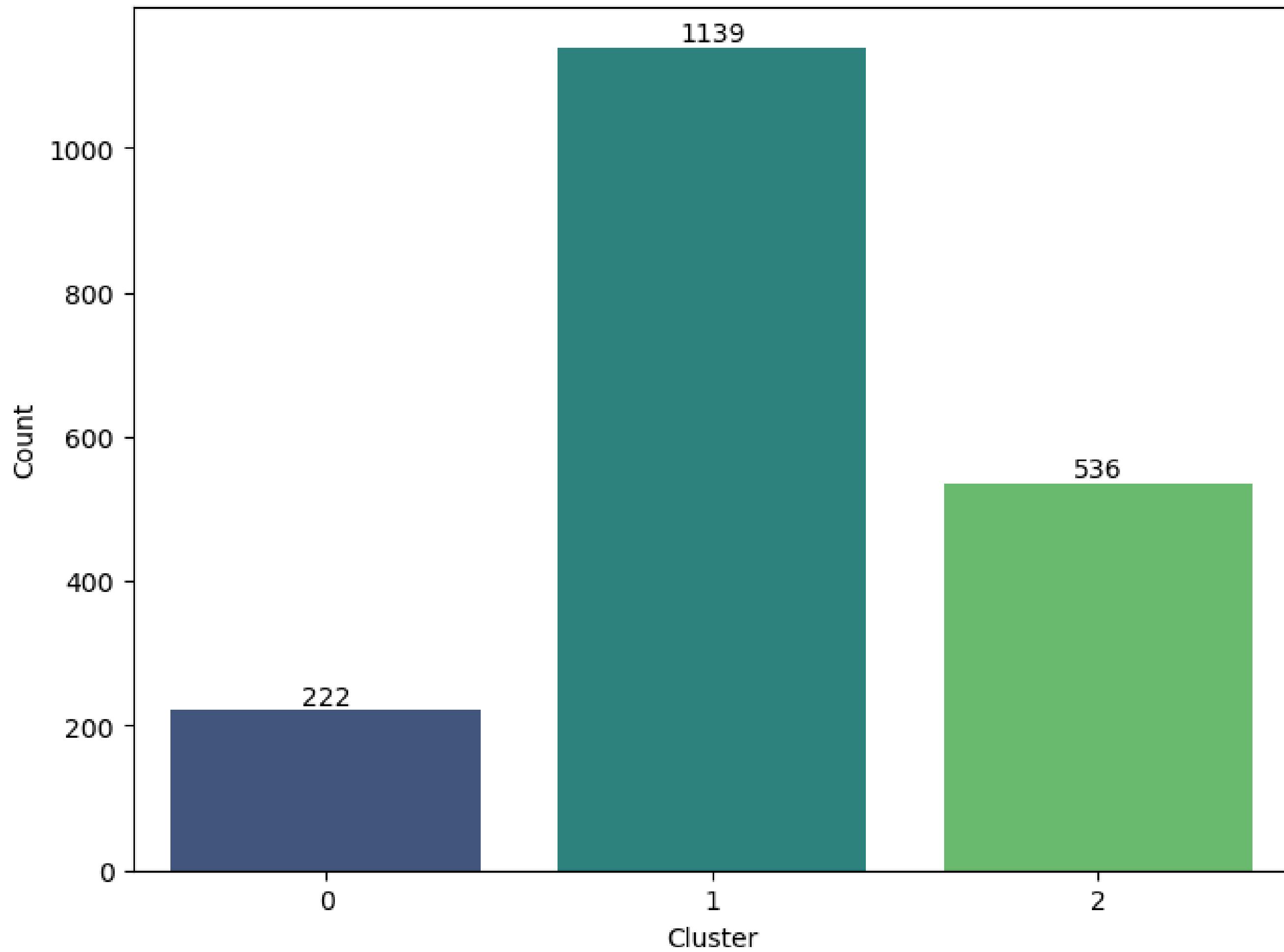
- n\_clusters : 3
- init : 'k-means++'
- n\_init : 10
- max\_iter : 300
- tol : 1e-4
- algorithm : 'auto'
- random\_state : 42

## t-SNE Visualization of Clustering Results



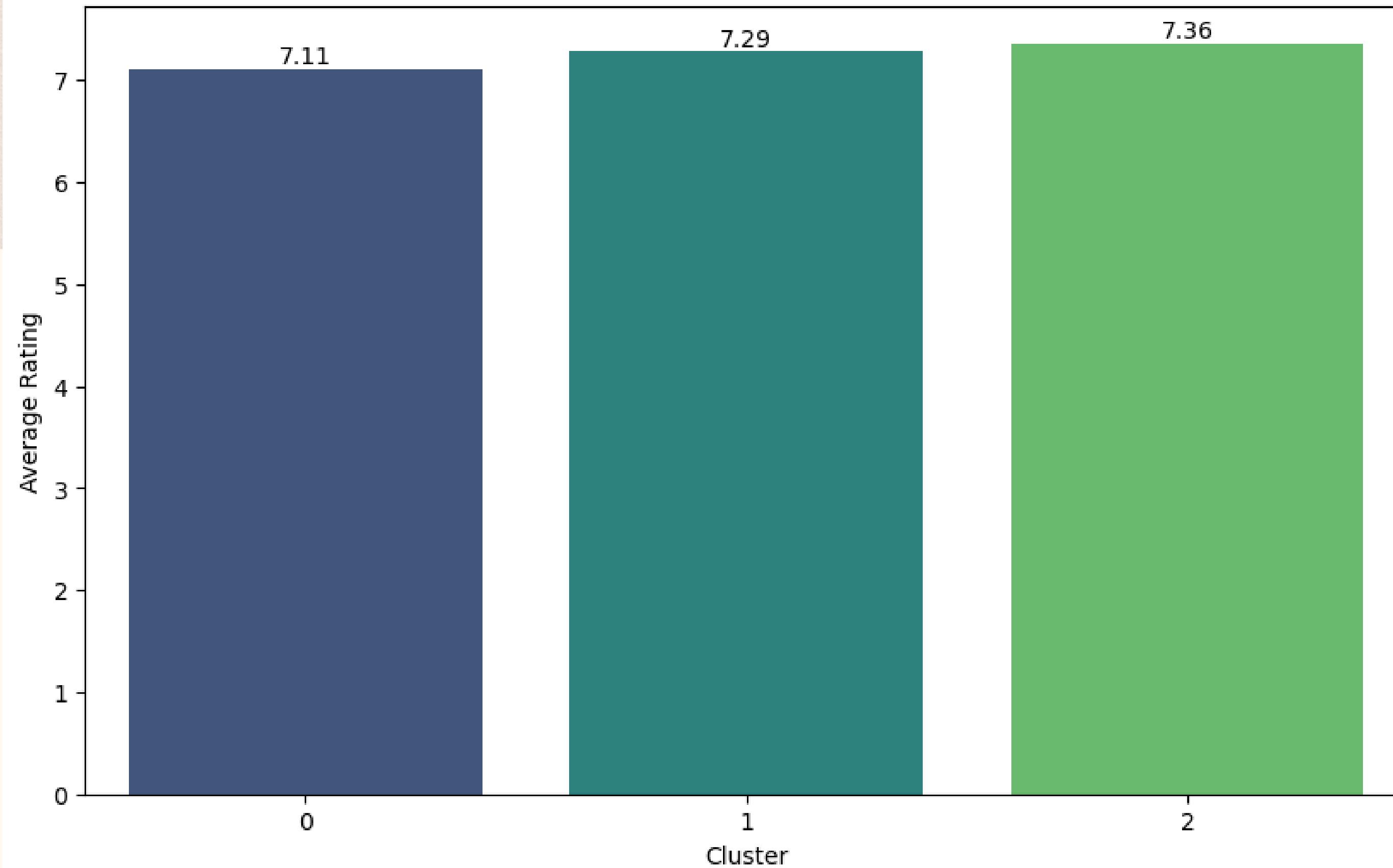
Kết quả phân cụm theo K-means Clustering.

Number of Movies in Each Cluster



Số phim ở mỗi cụm theo K-means Clustering

Average Rating in Each Cluster



Rating trung bình của các phim ở mỗi cụm theo K-means Clustering

# K-means clustering

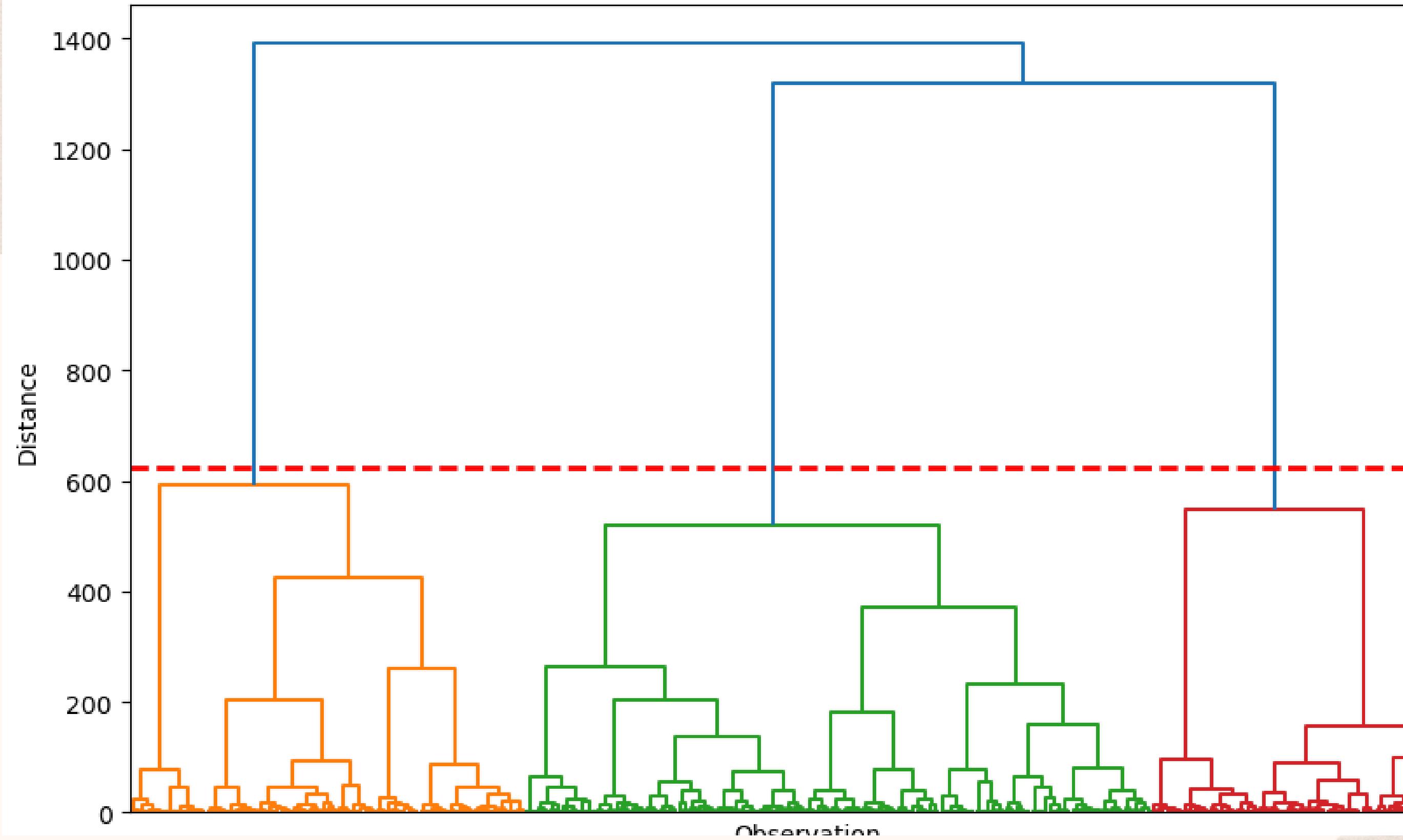
Nhận xét:

- Mô hình phân cụm rõ ràng
- Phân bổ dữ liệu vào các cụm không đồng đều
- Có sự khác biệt nhẹ ở rating trung bình trong từng cụm

# Hierarchical Clustering

- Phương pháp phân cụm dữ liệu dựa trên việc xây dựng một hệ thống phân cấp của các cụm.
- Ban đầu mỗi điểm là một cụm riêng biệt. Thuật toán phân cụm phân cấp sẽ tạo ra các cụm lớn hơn bằng cách sáp nhập các cụm nhỏ hơn gần nhau nhất tại mỗi vòng lặp.
- Dùng biểu đồ Dendrogram để xác định số lượng cụm tối ưu

### Hierarchical Clustering Dendrogram

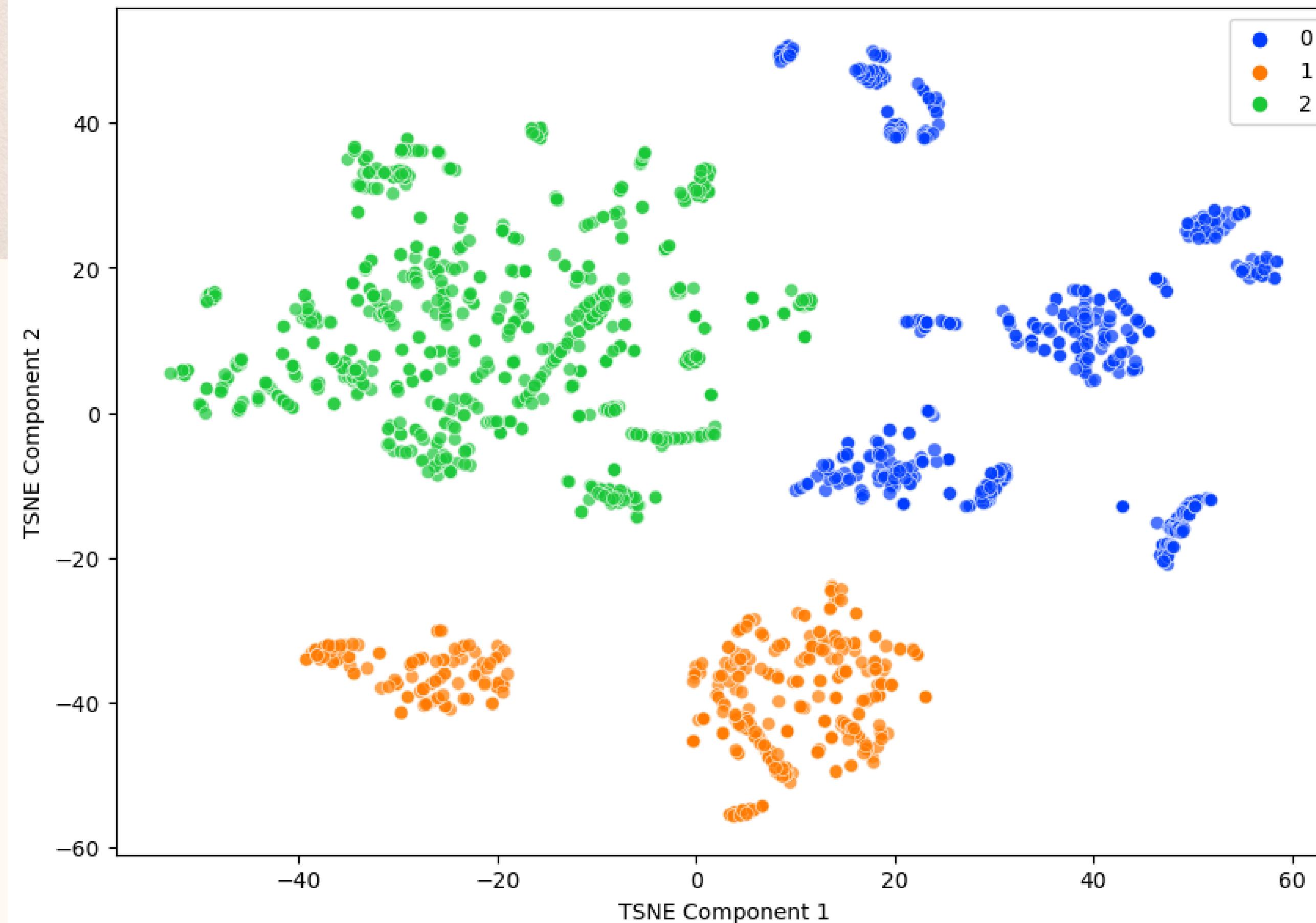


Biểu đồ Dendrogram theo Agglomerative hierarchical clustering.

# Hierarchical Clustering

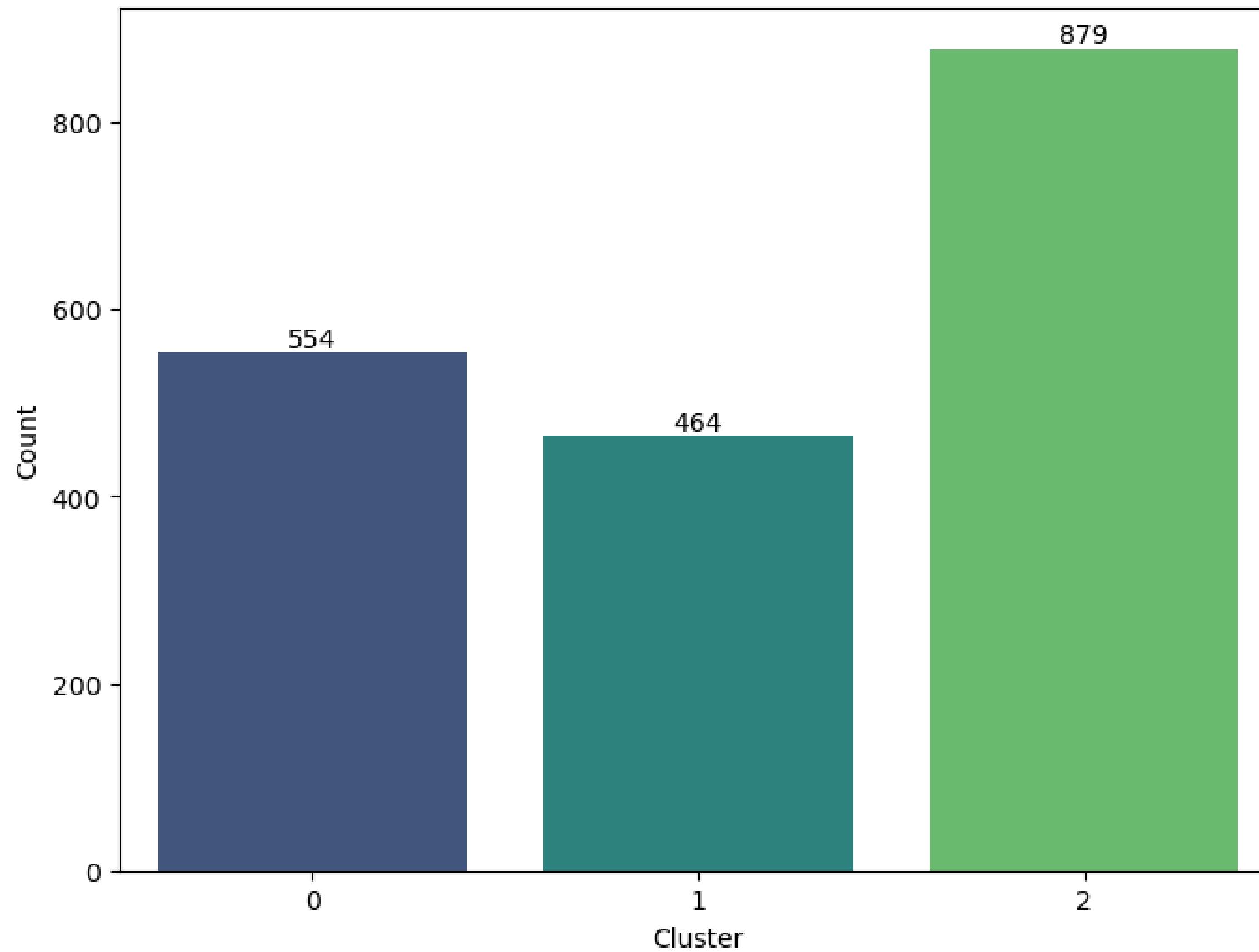
- linkage : 'ward'.
- n\_clusters : 3.
- affinity : 'euclidean'.
- distance\_threshold : None.
- memory : None.
- connectivity : None.

t-SNE Visualization of Clustering Results



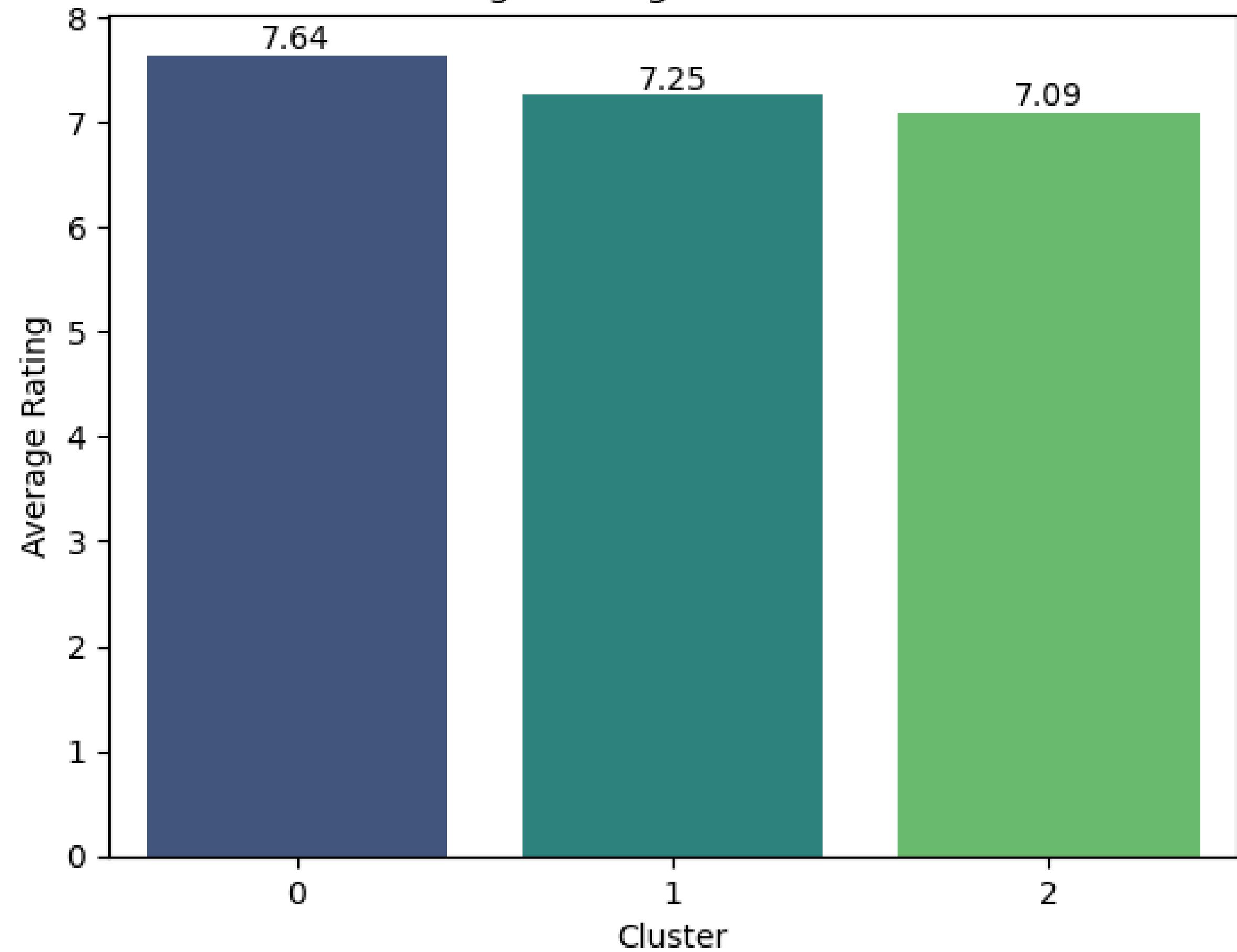
Kết quả phân cụm theo Hierarchical Clustering

Number of Movies in Each Cluster



Số phim ở mỗi cụm theo Hierarchical Clustering

### Average Rating in Each Cluster



Rating trung bình của các phim ở mỗi cụm theo Agglomerative hierarchical clustering

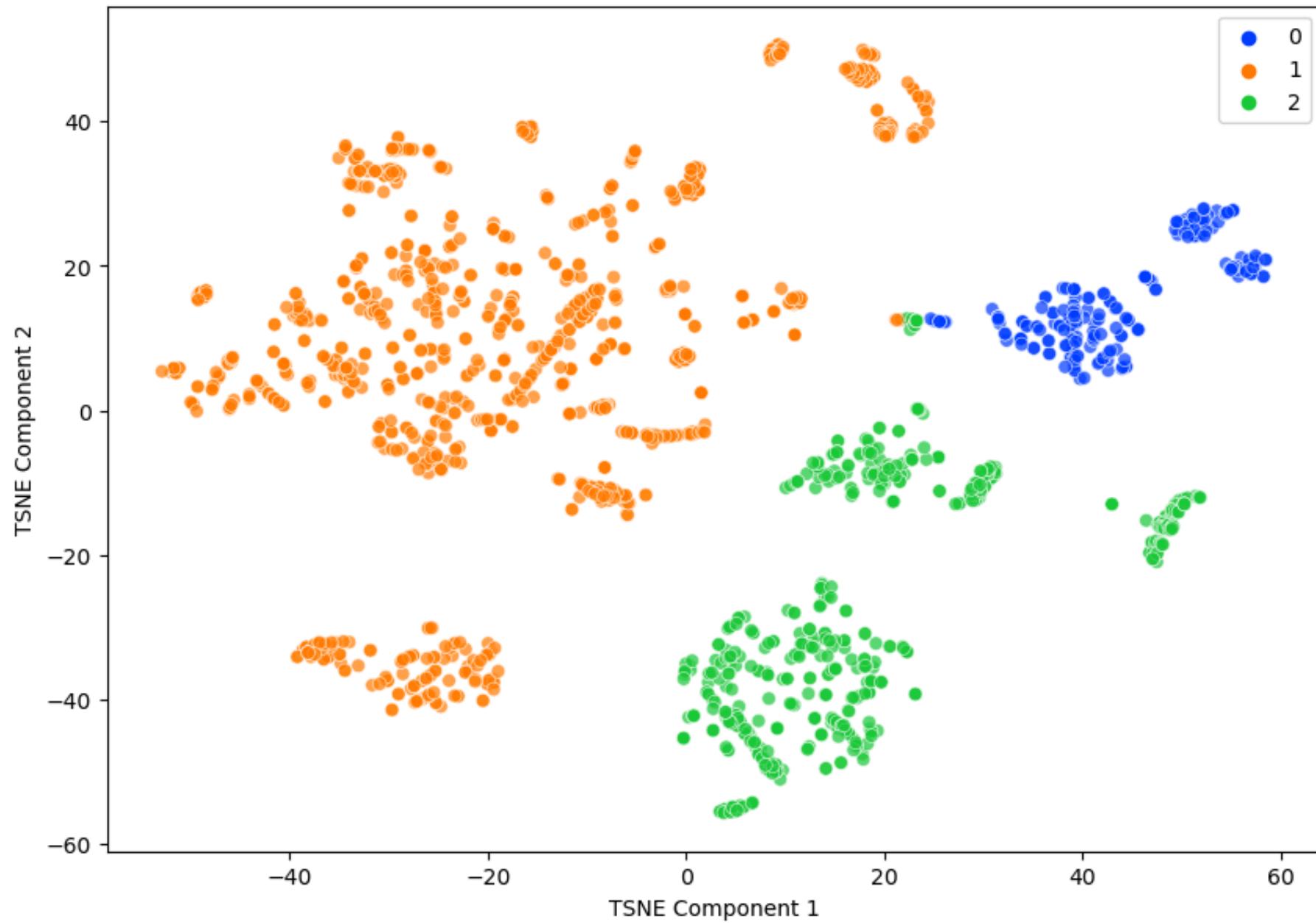
# Hierarchical Clustering

Nhận xét:

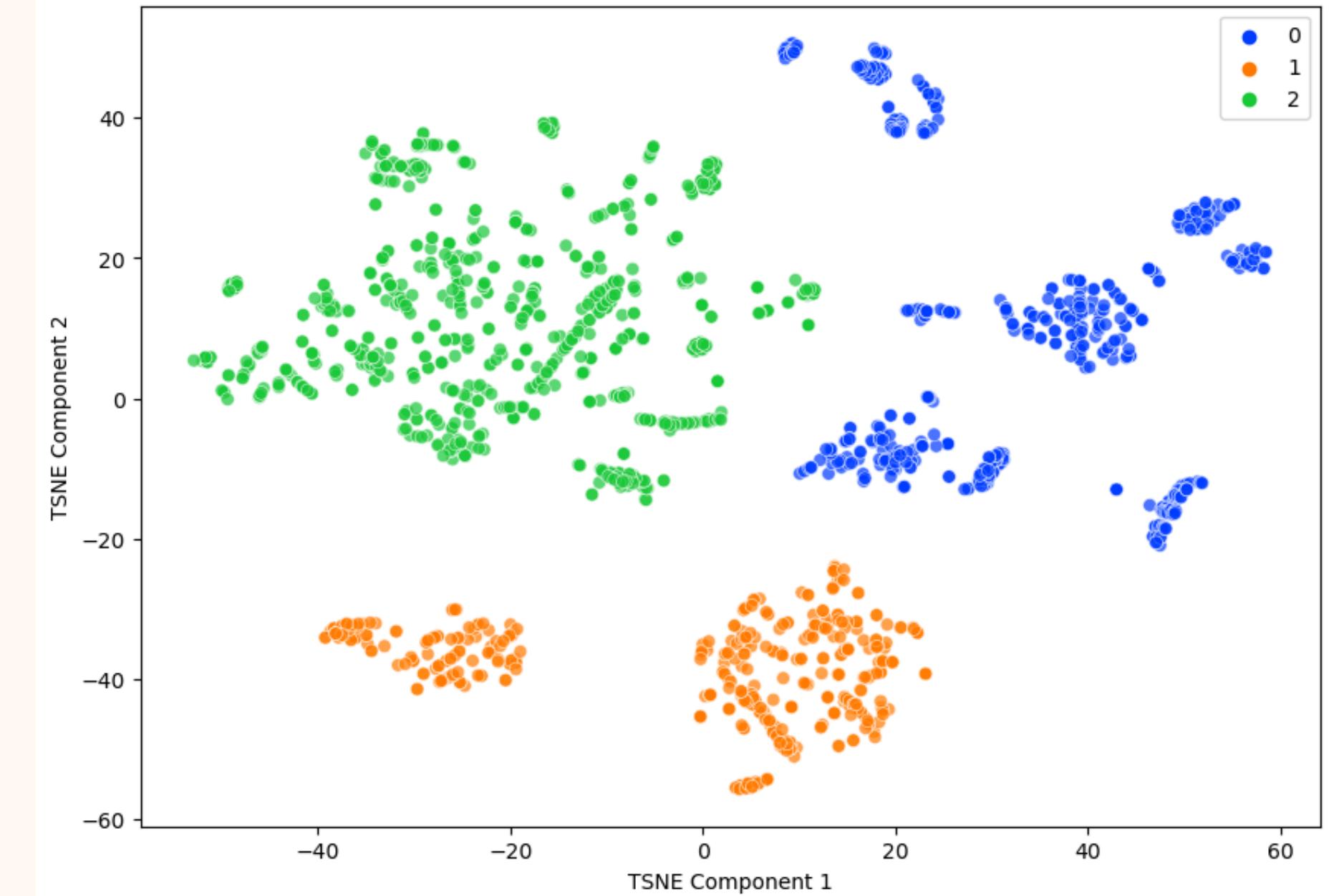
- Mô hình phân cụm rõ ràng
- Phân bổ dữ liệu vào các cụm không đồng đều
- Có sự khác biệt rõ ở rating trung bình trong từng cụm

# SO SÁNH 2 MODEL

t-SNE Visualization of Clustering Results



t-SNE Visualization of Clustering Results

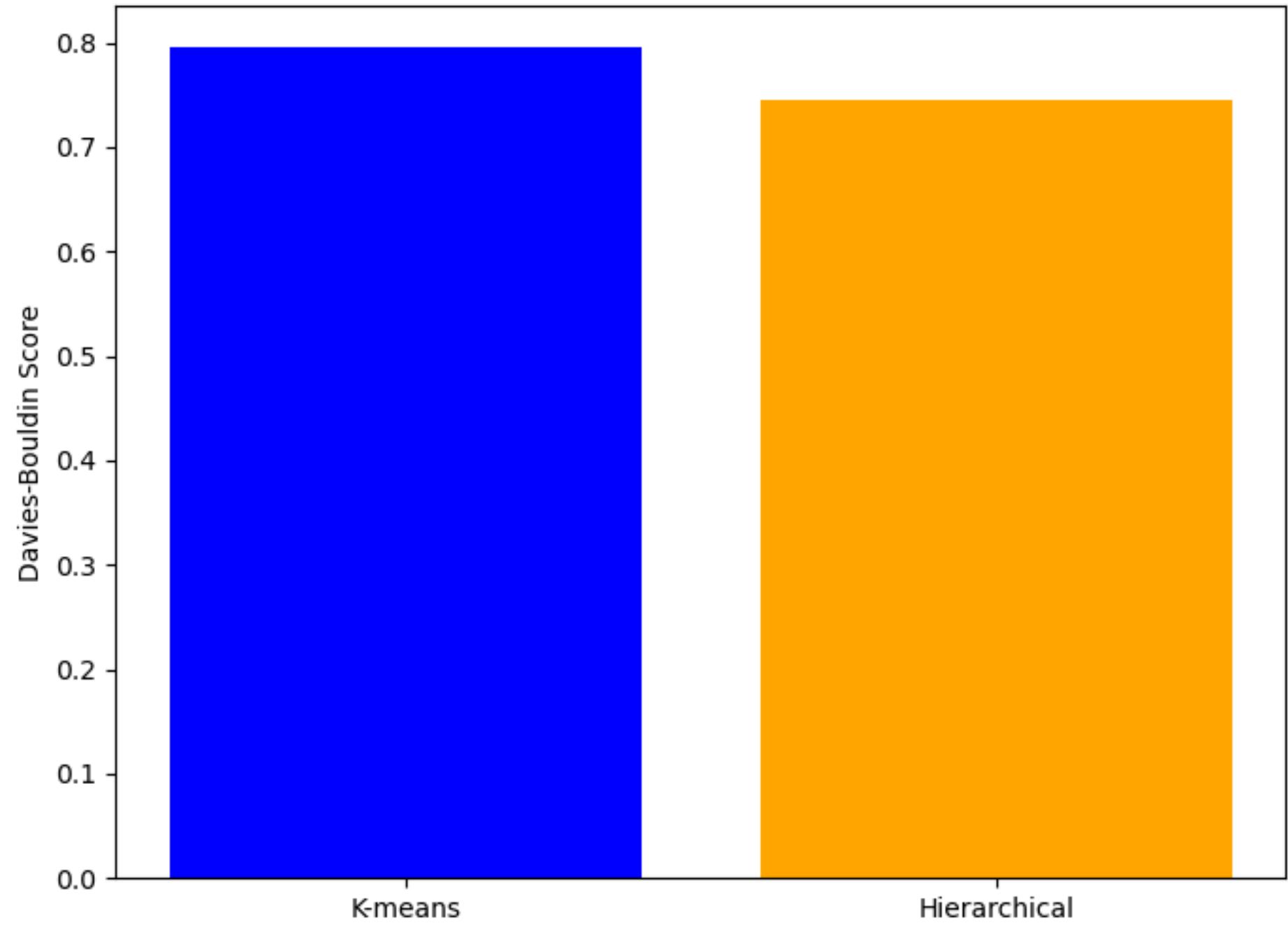


K-means Clustering.

Hierarchical Clustering

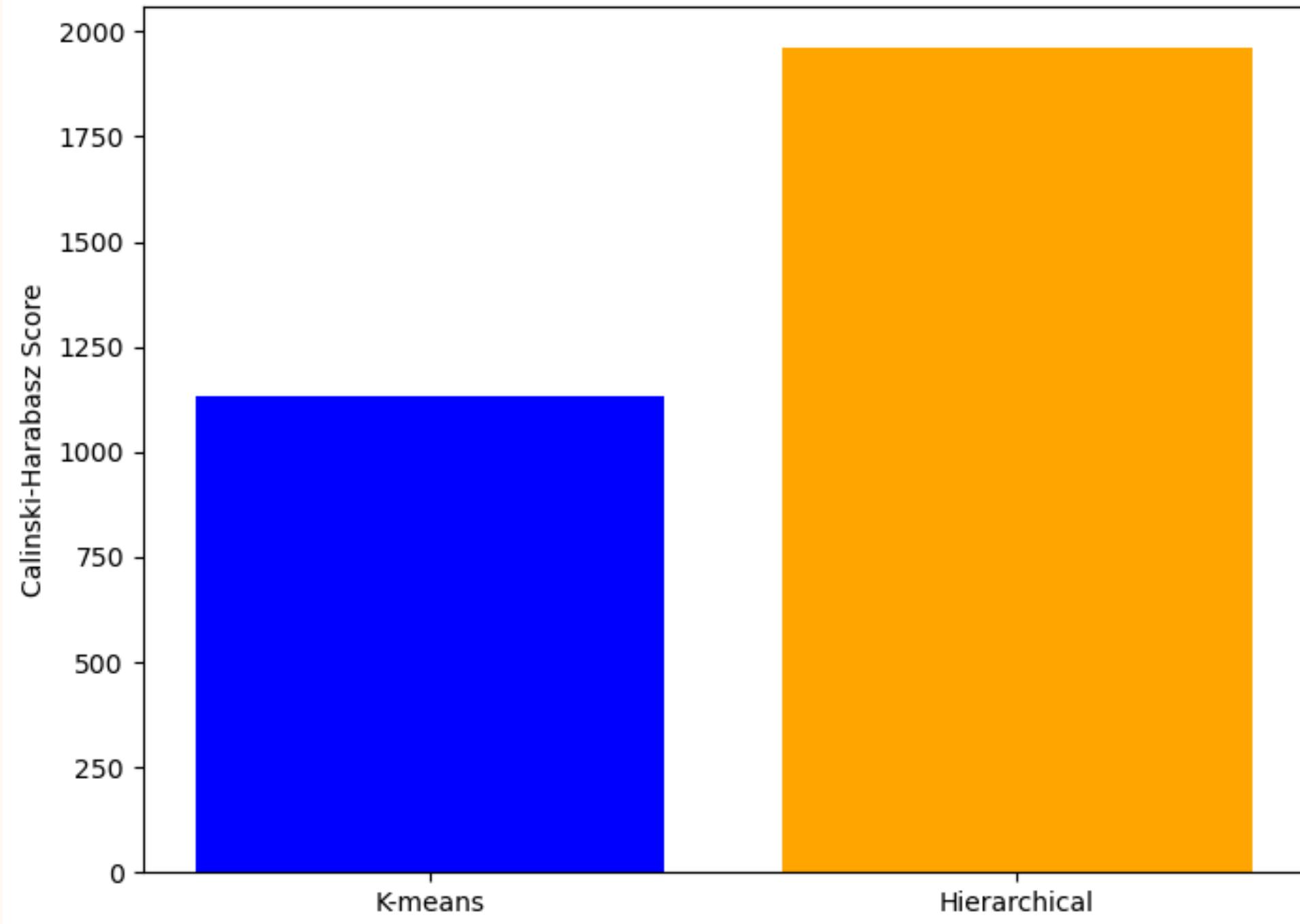
# SO SÁNH 2 MODEL

Davies-Bouldin Score Comparison ( lower is better )



chỉ số Davies-Bouldin của 2 mô hình

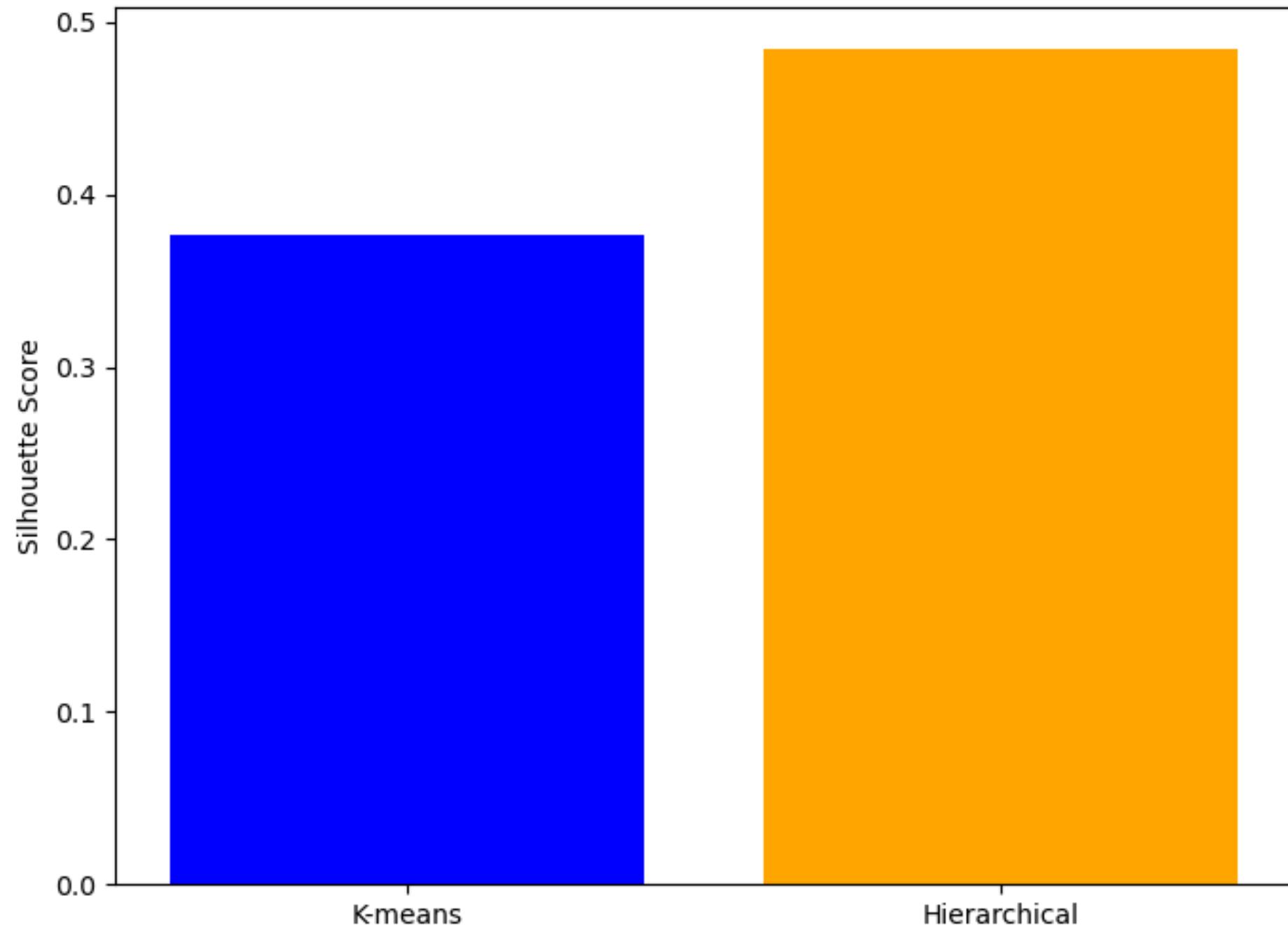
Calinski-Harabasz Score Comparison ( higher is better )



chỉ số Calinski-Harabasz của 2 mô hình

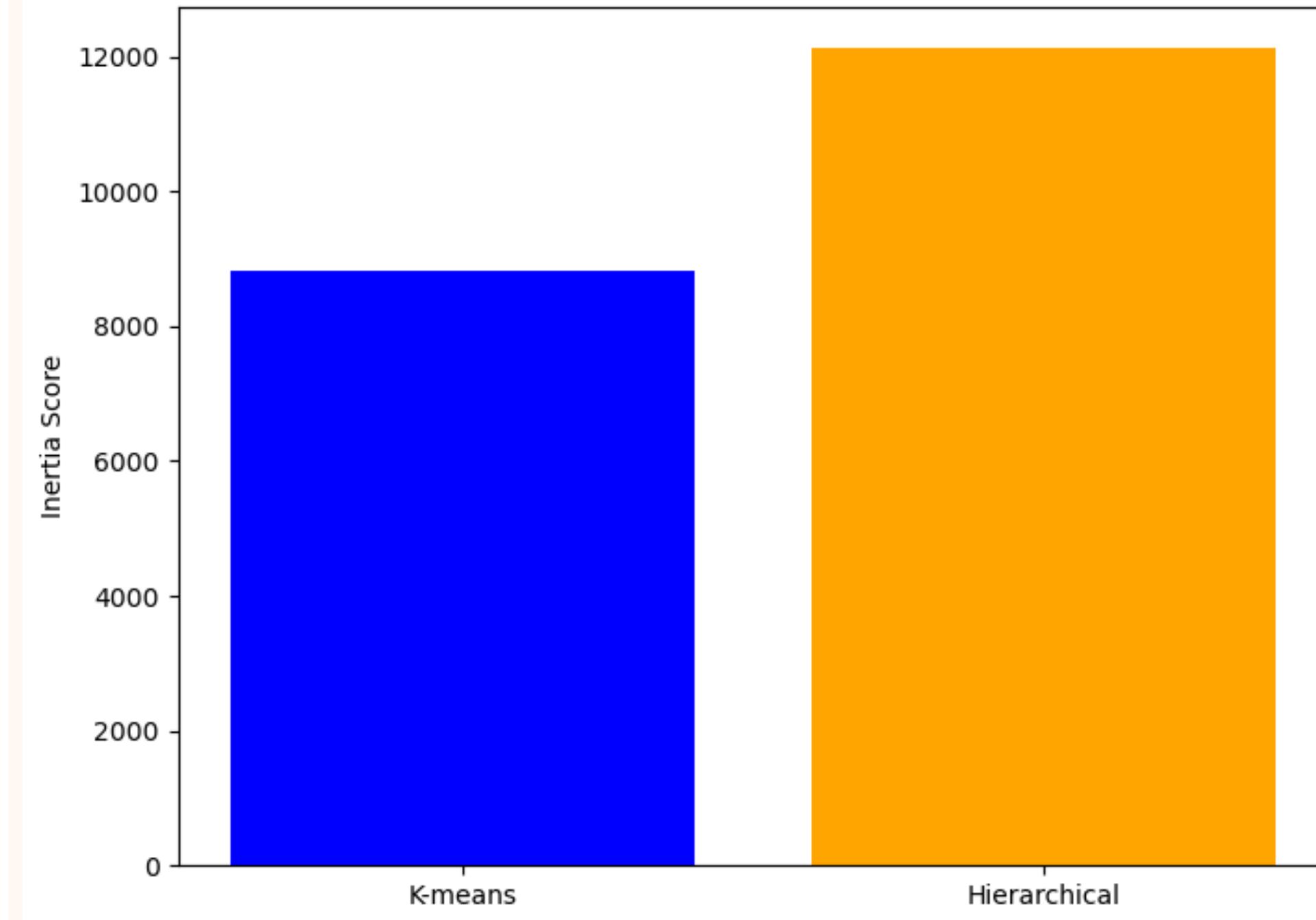
# SO SÁNH 2 MODEL

Silhouette Score Comparison ( higher is better )



chỉ số Silhouette của 2 mô hình

Inertia Comparison ( lower is better )



chỉ số Inertia của 2 mô hình

# SO SÁNH 2 MODEL

- Cả hai mô hình đều cho ra kết quả phân cụm rõ ràng
- Hierarchical Clustering cho ra kết quả phân cụm tốt hơn K-means ở 3 chỉ số đánh giá Davies-Bouldin Score, Calinski-Harabasz Score, Silhouette Score nhưng tệ hơn ở chỉ số Inertia
- K-means Clustering chọn cách phân cụm tối ưu chỉ số Intertia nên các chỉ số khác không được tối ưu.  
=> Chọn mô hình Hierarchical Clustering để phân cụm dữ liệu.

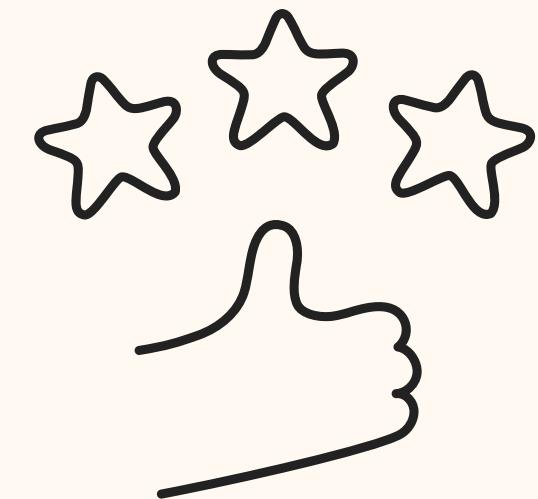
# KẾT LUẬN

## Phân loại

- Gradient Boosting là mô hình tốt nhất trong ba mô hình, với độ chính xác cao nhất.
- Tuy nhiên, KNN và Random Forest cũng đạt kết quả cao, cho thấy chúng đều là những mô hình mạnh mẽ và đáng tin cậy. Sự chênh lệch nhỏ giữa các mô hình cho thấy tất cả đều đủ tốt để sử dụng.  
Chọn Gradient Boosting làm mô hình chính để phân loại dữ liệu

## Phân cụm

- Cả hai mô hình đều cho ra kết quả phân cụm rõ ràng
- Agglomerative hierarchical clustering cho ra kết quả phân cụm tốt hơn K-means ở 3 chỉ số đánh giá Davies-Bouldin Score, Calinski-Harabasz Score, Silhouette Score nhưng tệ hơn ở chỉ số Inertia
- K-means Clustering chọn cách phân cụm tối ưu chỉ số Intertia nên các chỉ số khác không được tối ưu.  
Chọn mô hình Agglomerative hierarchical clustering để phân cụm dữ liệu.



# THANK YOU

Contact us if you are interested in our project.

[www.reallygreatsite.com](http://www.reallygreatsite.com)