**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

# PHÂN TÍCH VÀ DỰ ĐOÁN ĐIỂM SỐ IMDB CỦA CÁC BỘ PHIM



| Sinh viên thực hiện: | | |
|---|---|---|
| STT | Họ tên | MSSV |
| 1 | Lê Nguyễn Bá Duy | 20521232 |
| 2 | | |
| 3 | | |



**TP. HỒ CHÍ MINH – 12/2023**

# 1. GIỚI THIỆU

Thế giới điện ảnh luôn là một lĩnh vực thú vị, nơi mà tính nghệ thuật, tính sáng tạo, công nghệ giao nhau. Đây là một nghành công nghiệp năng động và đang trên đà tiến hoá và thích nghi để có thể phản ánh được sự thanh đổi của xã hội và nhưng tiến bộ về mặt công nghệ. Một yếu tố quan trọng để quyết định sự thành công của một bộ phim đó chính là điểm đánh giá của một bộ phim hay còn gọi là rating, đây được xem như là một thước đo để đánh giá chất lượng của bộ phim và độ hấp dẫn của nó đối với khán giả. Vì vậy nhóm chúng em đã quyết định xây dựng đề tài có liên quan đến chủ đề trên. Mục tiêu của đề tài này bao gồm:

- Phân tích và tìm hiểu những yếu tố sẽ ảnh hưởng đến số điểm đánh giá của các bộ phim chiếu rạp.
- Trực quan một số đặc tính đáng chú ý của bộ dữ liệu
- Xây dựng một mô hình dự đoán rating của các bộ phim.
- Đánh giá hiệu xuất dự đoán của mô hình.

Trong đề tài này bọn em sử dụng ngôn ngữ Python cùng với một số thư viện sau:

- Cào dữ liệu: BeautifulSoup, Selenium
- Trực quan dữ liệu: mathplotlib, seaborn, plotly
- Phân tích và xây dựng mô hình: pandas, numpy, scikit-learn, …

Bộ dữ liệu được sử dụng trong đề tài được chúng em tự thu thập về từ trang web IMDB là một trang web nổi tiếng về cơ sở dữ liệu phim khổng lồ và cung cấp các điểm số đánh giá của người dùng cũng như của các chuyên gia đánh giá về các bộ phim.

Qua đề tài này, chúng em đã có thể phân tích và đưa ra một số thông tin như sự thịnh hành của các thể loại qua từng năm, mối quan hệ giữa các phân loại phim với rating và lợi nhuận, …Và đã có thể xây dựng một mô hình dự đoán được điểm đánh giá với độ chính xác đáng kể.

# 2. MÔ TẢ BỘ DỮ LỆU

Bộ dữ liệu được chúng em thu thập từ trang web IMDB với một số tiêu chí sau:

- Những phim được thu thập là những bộ phim lẻ (phim chiếu rạp)
- Những bộ phim được thu thập đã được có điểm đánh giá rating và số lượng người đánh giá bộ phim là trên 50000 người.

Bộ dữ liệu bao gồm 19 cột và 4073 dòng dữ liệu. Mỗi dòng dữ liệu là một bộ phim khác nhau.

**Các biến được sử dụng trong bộ dữ liệu:**

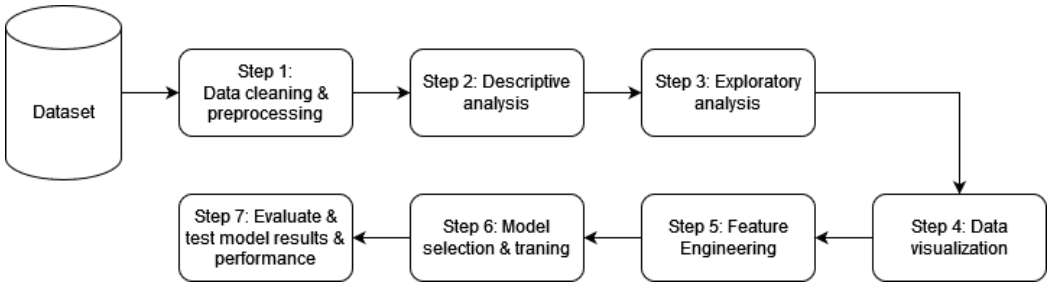
- title: Tiêu đề của bộ phim (Integer)
- num-of-voters: Số lượng người đã cho điểm rating của bộ phim (Integer)
- release-year: mốc thời gian mà bộ phim được ra mắt (DateTime)
- certificate: chứng chỉ độ tuổi của bộ phim (String)
- genres: các thể loại của bộ phim (String)
- length: thời lượng của bộ phim được tính theo phút (Integer)
- director: đạo diễn của bộ phim (String)
- writer: người viết kịch bản cho bộ phim (String)
- cast: diễn viên trong bộ phim (String)
- language: ngôn ngữ được sử dụng trong bộ phim (String)
- company: công ty phụ trách sản xuất ra bộ phim (String)
- origin: nơi bộ phim được sản xuất (String)
- budget: ngân sách được sử dụng cho bộ phim (Float)
- domestic-gross: doanh thu tại nơi bộ phim được sản xuất (Float)
- worldwide-gross: doanh thu toàn cầu của bộ phim (Float)
- rating: số điểm đánh giá IMDB của bộ phim (Float)
- metascore: số điểm đánh giá metascore của bộ phim (Interger)
- metauser: số lượng người xem đã cho điểm metascore (Integer)
- metacritic: số lượng nhà phê bình đã cho điểm metascore (Integer)

**Tiềm năng và hạn chế của bộ dữ liệu:**

- Tiềm năng: bộ dữ liệu có số lượng dữ liệu khá lớn, có thể được sử dụng để phân tích theo nhiều hướng khác nhau như dự đoán rating của bộ phim, phân tích xu hướng thể loại phi, …
- Hạn chế: bộ dữ liệu không bao hàm hết tất cả các bộ phim trên IMDB, có chứa một số bộ phim đã cũ vì vậy một số thông tin sẽ bị lỗi thời hoặc bị khuyết và vì được cào xuống từ web nên có thể sẽ chứa một số lỗi trong đó.

## 3. PHƯƠNG PHÁP PHÂN TÍCH

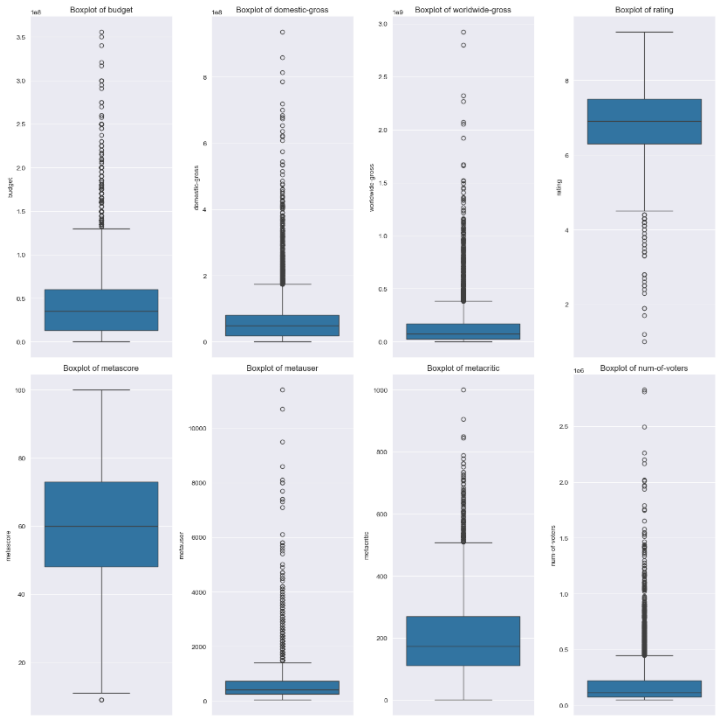
Đối với đề tài này, nhóm em quết định chia ra là 6 bước:




Hình 1: Các bước phân tích dữ liệu

### 3.1. Làm sạch dữ liệu và tiền xử lý

- Biến đổi dữ liệu:
  - Đối với cột release-year, để dễ xử lý hơn thì sẽ chuyển đổi thành kiểu pandas.datetime.
  - Các cột budget, domestic-gross, world-wide vẫn đang là dạng string vì vậy cần phải loại bỏ các ký tự không phải số và biến đổi thành kiểu số. Đồng thời các đơn vị tiền tệ sử dụng trong các cột trên đang khác nhau vì vậy để tăng độ chính xác của bộ dữ liệu thì cần phải chuyển về một đơn vị đồng nhất là Dollar.
  - Cột length đang ở dạng string vì vậy em đã biến đổi và sử dụng thư viện time delta để chuyển về kiểu số với đơn vị là phút.
  - Cột metauser có một số giá trị có kí tự 'K' thể hiện đơn vị là 1000 vì vậy cần biến đổi để tất cả đều có kiểu integer.
- Xử lý giá trị khuyết:
  - Có một số hàng bị thiếu giá trị language và certificate nhưng số lượng không đáng kể nên em đã loại bỏ những hàng đó.
  - Đối với các hàng bị thiếu các giá trị kiểu số thì em đã sử dụng phương pháp  điền khuyết bằng giá trị trung bình.
- Tạo thêm biến đặc trưng:
  - Thông thường với các bộ phim thì điểm đánh giá sẽ tỉ lệ thuận với lợi nhuận mà bộ phim mang đến vì vậy em đã tạo ra biến mới:
    - profit: lợi nhuận mà bộ phim đem lại toàn cầu (Float)
- Xử lý ngoại lệ:
  Để phát hiện các ngoại lệ thì em đã sử dụng boxplot để minh hoạ các biến s




Hình 2: Boxplot của các biến số

Theo như hình thì các hầu hết các biến số đều có một số lượng ngoại lệ nhất định nhưng đây không phải gây ra do lỗi mà là các ngoại lệ hợp lệ

- Với biến rating thì sẽ có những bộ phim không được đánh giá tốt và sẽ xuất hiện ngoại lệ dưới
- Với các biến còn lại thì sẽ xuất hiện những bộ phim thành công hơn nên sẽ các biến còn lại sẽ xuất hiện nhiều ngoại lệ trên.
- Vì là các ngoại lệ là hợp lệ nên thay vì loại bỏ thì em dùng phương pháp winsoration ( thay thế các giá ngoại lệ bằng phần tử nhỏ nhất và phần tử lớn nhất không phải ngoại lệ).

### 3.2. Phân tích mô tả

Ở phần này sẽ phân tích một số đặc tính cơ bản của bộ dữ liệu:

- Kiểm tra kích thước bộ dữ liệu sau khi đã làm sạch và tiền xử lý:

```
Index: 4068 entries, 0 to 4076
Data columns (total 19 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   title            4068 non-null   object
 1   num-of-voters    4068 non-null   int64
 2   certificate      4068 non-null   object
 3   length           4068 non-null   float64
 4   genres           4068 non-null   object
 5   director         4068 non-null   object
 6   writer           4068 non-null   object
 7   cast             4068 non-null   object
 8   language         4068 non-null   object
 9   company          4068 non-null   object
 10  origin           4068 non-null   object
 11  budget           4068 non-null   float64
 12  domestic-gross   4068 non-null   float64
 13  worldwide-gross  4068 non-null   float64
 14  rating           4068 non-null   float64
 15  metascore        4068 non-null   float64
 16  metauser         4068 non-null   int64
 17  metacritic       4068 non-null   int64
 18  profit           4068 non-null   float64
dtypes: float64(7), int64(3), object(9)
```

Hình 3: Thông tin về bộ dữ liệu

- Bảng mô tả thông kê của các biến số:

| | title | certificate | genres | director | writer | cast | language | company | origin |
|---|---|---|---|---|---|---|---|---|---|
| count | 4068 | 4068 | 4068 | 4068 | 4068 | 4068 | 4068 | 4068 | 4068 |
| unique | 3982 | 17 | 330 | 1585 | 3520 | 3820 | 43 | 1183 | 49 |
| top | Halloween | 15 | [Comedy, Drama, Romance] | Steven Spielberg | [Woody Allen] | [Daniel Radcliffe, Emma Watson] | English | Universal Pictures | United States |
| freq | 3 | 1590 | 157 | 30 | 14 | 6 | 3694 | 264 | 2928 |

Hình 4: Mô tả thống kê của biến số

- Bảng mô tả thống kê của các biến phân loại:

| | num-of-voters | length | budget | domestic-gross | worldwide-gross | rating | metascore | metauser | metacritic | profit |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 4068.0000 | 4068.0000 | 4068.0000 | 4068.0000 | 4068.0000 | 4068.0000 | 4068.0000 | 4068.0000 | 4068.0000 | 4068.0000 |
| mean | 168277.7018 | 113.3867 | 42968819.0838 | 59164499.1128 | 118278103.5996 | 6.8517 | 60.3864 | 542.8051 | 202.4594 | 75309284.5157 |
| std | 124280.6181 | 21.6530 | 37465300.4616 | 51046549.6383 | 118209558.1055 | 0.8605 | 17.4271 | 380.5687 | 120.6349 | 96882115.8089 |
| min | 50026.0000 | 42.0000 | 18.0000 | 509.0000 | 67.0000 | 4.5000 | 11.0000 | 49.0000 | 1.0000 | -129821857.0000 |
| 25% | 73812.7500 | 98.0000 | 13000000.0000 | 17592731.0000 | 25243297.5000 | 6.3000 | 48.0000 | 256.0000 | 111.0000 | 3069962.0000 |
| 50% | 115864.0000 | 110.0000 | 35000000.0000 | 48320889.0000 | 76451886.5000 | 6.9000 | 60.0000 | 411.0000 | 174.0000 | 42751256.5000 |
| 75% | 222825.5000 | 124.0000 | 60000000.0000 | 80019981.7500 | 167824819.7500 | 7.5000 | 73.0000 | 730.2500 | 270.0000 | 122400104.5000 |
| max | 446062.0000 | 321.0000 | 130000000.0000 | 173585516.0000 | 381545846.0000 | 9.3000 | 100.0000 | 1400.0000 | 508.0000 | 379093048.0000 |

Hình 5: Bảng mô tả thống kê của các biến phân loại

- Kiểm tra độ đối xứng của các biến số: Các biến số không bị quá mất đối xứng

| | <unnamed> |
|---|---|
| num-of-voters | 1.1788 |
| length | 1.3972 |
| budget | 1.0443 |
| domestic-gross | 0.9323 |
| worldwide-gross | 1.0930 |
| rating | -0.3288 |
| metascore | -0.0591 |
| metauser | 1.0649 |
| metacritic | 0.8552 |
| profit | 1.0267 |

Hình 6: Độ lệch của các biến số

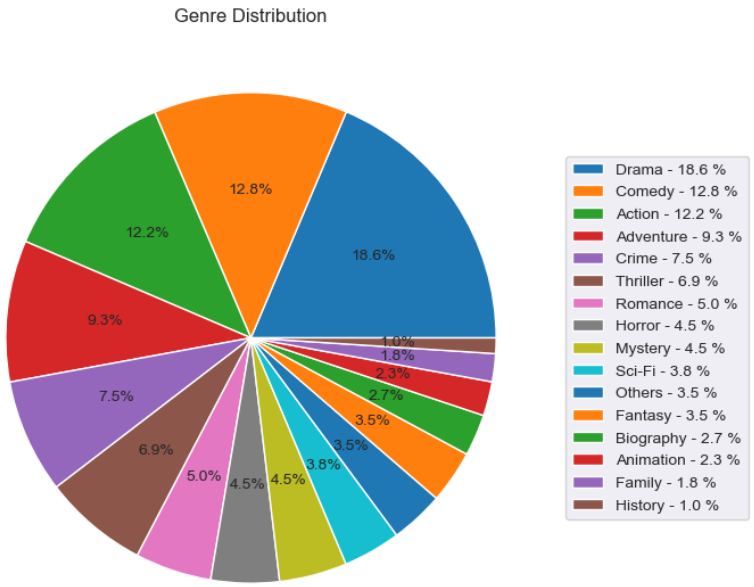
- Kiểm tra độ nhọn của các biến số: phân phối của các biến đều có sự dàn trải nhẹ

| | <unnamed> |
|---|---|
| num-of-voters | 0.1291 |
| length | 4.5998 |
| budget | 0.1785 |
| domestic-gross | -0.0504 |
| worldwide-gross | 0.0411 |
| rating | -0.1477 |
| metascore | -0.4729 |
| metauser | 0.0304 |
| metacritic | 0.0613 |
| profit | 0.1718 |

Hình 7: Độ nhọn của biến số

### 3.3. Khám phá dữ liệu và minh hoạ dữ liệu

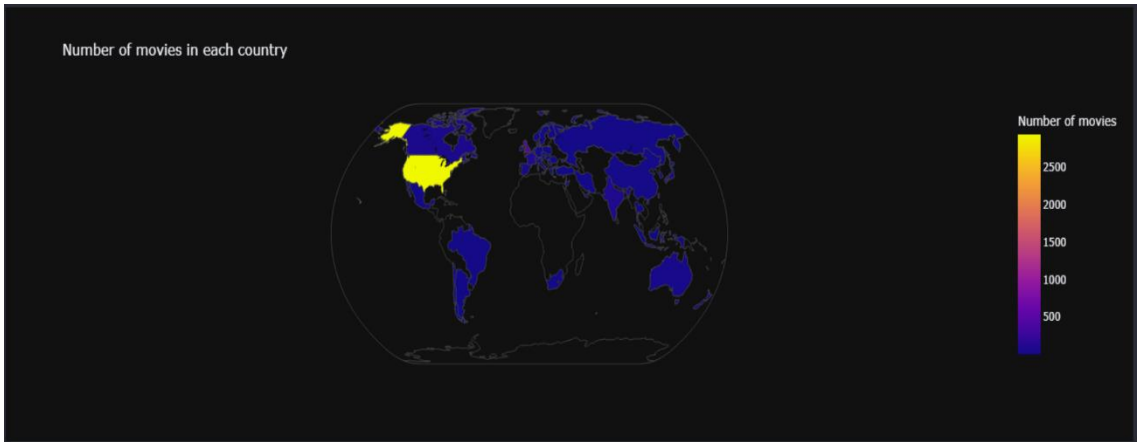
- Sự phân bố của các thể loại phim




Có thể thấy thể loại chính kịch, hài kịch, hành động, tội phạm và giật gân chiếm đến hơn 50% thị trường phim.

Hình 8: Sự phân bố của các thể loại phim

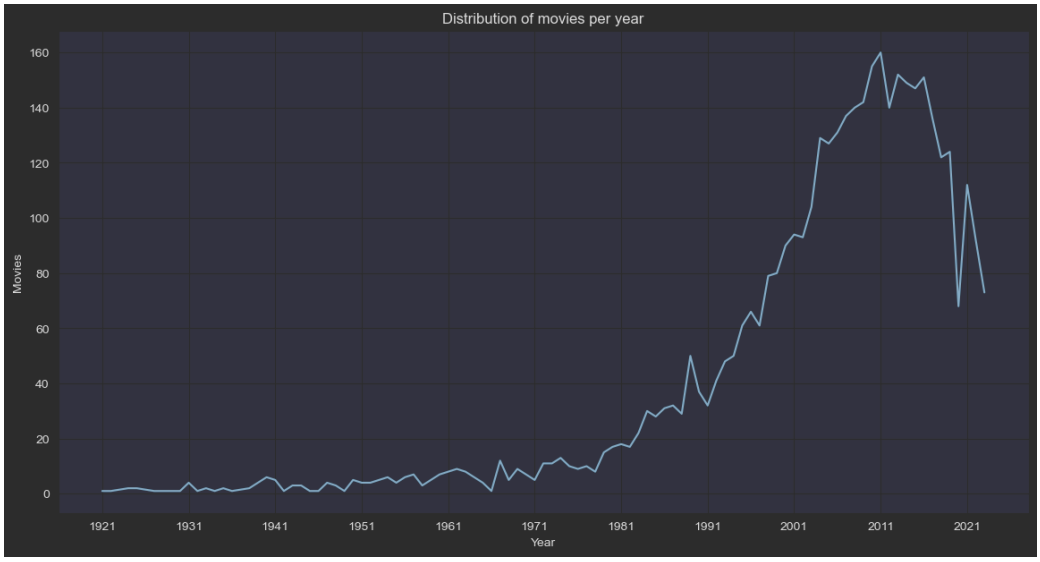
- Mật độ phim đến từ các nước trên thế giới




Hình 9: Mật độ phim đến từ các nước trên thế giới

⇨Những bộ phim có trên 50000 lượt đánh giá được sản xuất tại Mỹ chiếm hơn 50% thị phần phim của bộ dữ liệu.

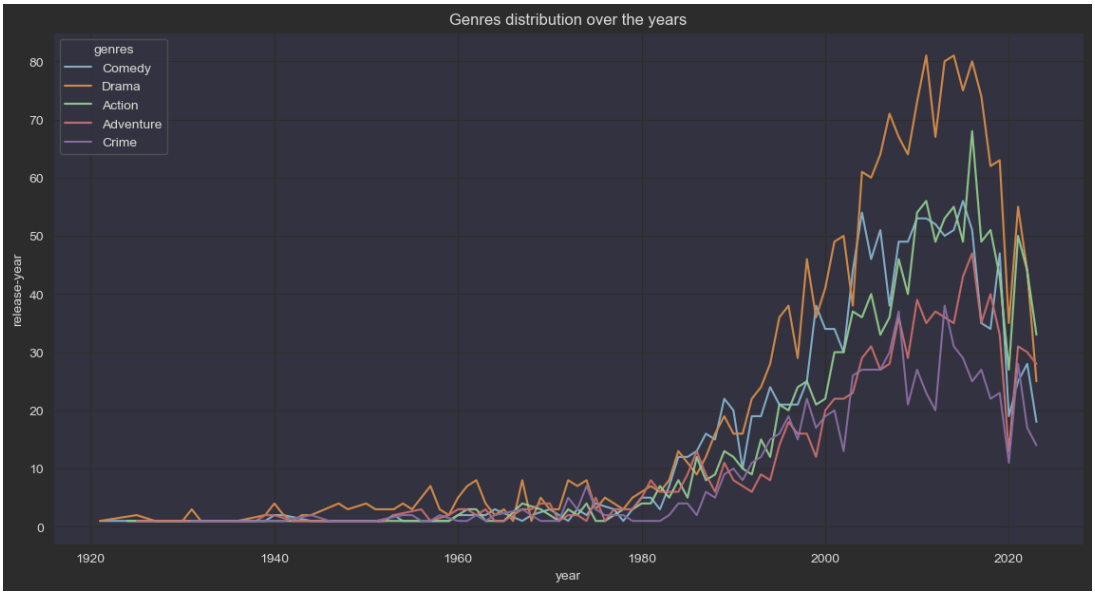
- Số bộ phim được ra mắt theo từng năm




Hình 10: Số bộ phim được ra mắt theo từng năm

⇨Từ năm 1921 đến năm 2011 đã có sự tăng vọt trong số lượng phim ra mắt mỗi năm với đỉnh là nắm 2011 với 160 bộ phim với trên 50000 lượt đánh giá nhưng từ 2011 đến 2021 đã có xu hướng giảm mạnh có thể đặc biệt chỉ còn tầm 70 phim 1 năm.

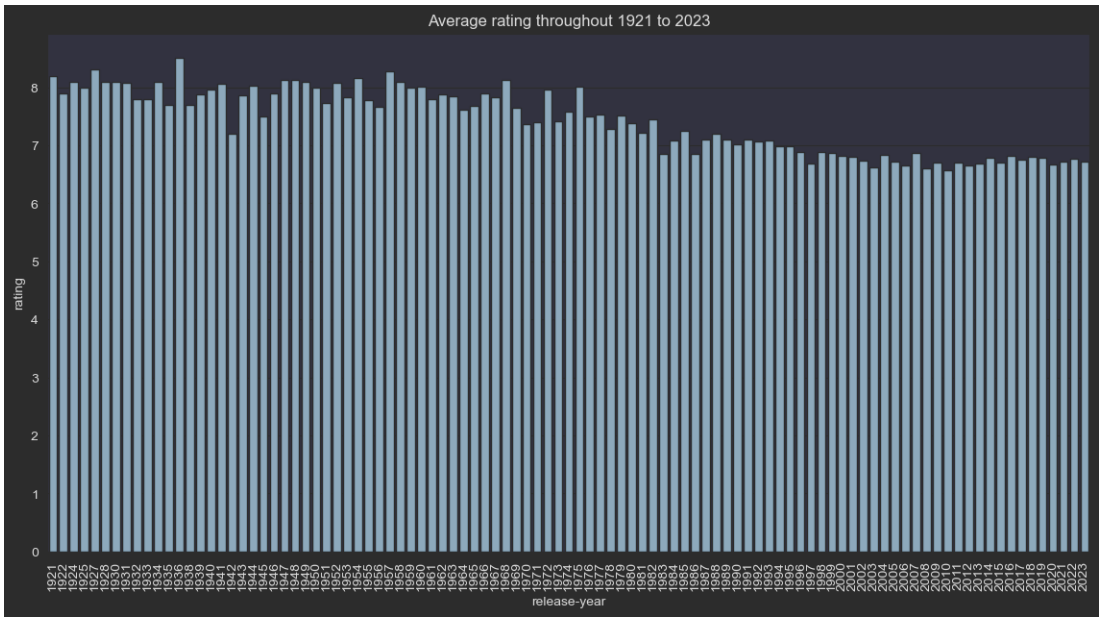
- Sự phân bố của số phim theo thể loại theo từng năm




Hình 11: sự phân bố của số phim theo thể loại theo từng năm

⇨Qua các năm thì chính kịch vẫn luôn là thể loại được lựa chọn nhiều nhất theo sau đó là thể loại hành động và hài kịch thay phiên nhau giữ bị trí thứ 2.

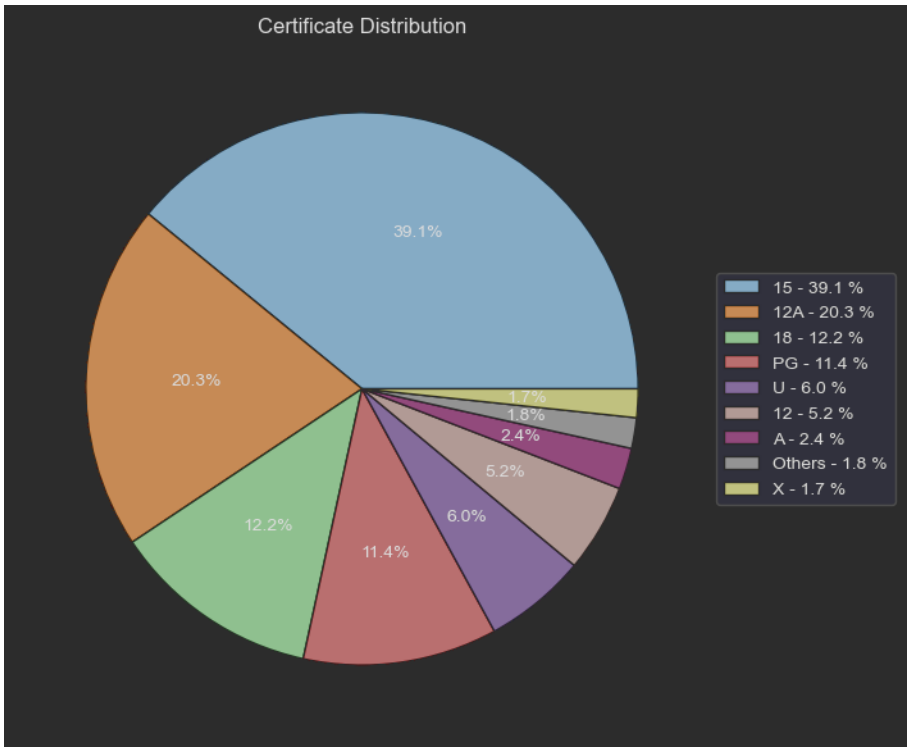
- Rating trung bình của các bộ phim qua các năm




Hình 12: Rating trung bình của các bộ phim qua các năm

➪Điểm số trung bình của các bộ phim khá ổn định và không có quá nhiều thay đổi

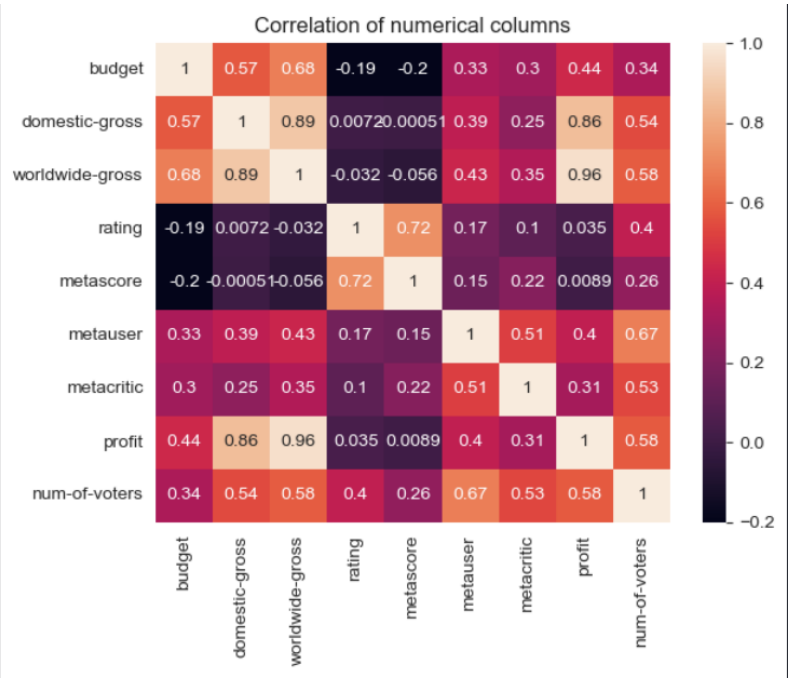
- Sự phân bố của các chứng chỉ




➪Có thể thấy rằng các bộ phim có chứng chỉ 15 chiếm một phần khá lơn điều đó chứng tỏ rằng các nhà làm phim thường nhắm vào các khán giả từ 15 tuổi trở lên.

Hình 13: Sự phân bố của các chứng chỉ

- Bảng chỉ số tương quan giữa các biến số




➪Theo như bảng bên thì biến metascore và biến num-of-voters sẽ có độ tương quan với biến rating cao nhất

Hình 14: Bảng chỉ số tương quan giữa các biến

### 3.4. Trích chọn biến feature

- Đối với các biến có quá nhiều giá trị độc nhất như title, cast, writer, director thì bị sẽ loại bỏ,
- Sau bước phân tích đã cho thấy biến language có hơn 90% giá trị là English vì vậy đối với các biến không phải English em sẽ chuyển nó thành Others.
- Đối với biến origin thì ngoài biến United States và United Kingdom chiếm hơn 82% nên sẽ biến không phải United States và United Kingdom em sẽ chuyển nó thành Others.
- Với biến release-year thì qua từng năm thì điểm đánh giá trung bình của không có sự thay đổi quá nhiều vì vậy ta có thể loại bỏ biến release-year.
- Đối với các biến phân loại genres, language, origin, certificate sẽ tạo các biến dummies của các biến đó và loại bỏ các biến đó

### 3.5. Xây dựng và đánh giá mô hình dự đoán

Để xây dựng mô hình dự đoán đầu tiên thì em sẽ chia ra các biến trong bộ dữ liệu ra làm 2 phần:

- Target (rating): là mục tiêu dự đoán của mô hình.
- Feature (các biến còn lại): là các biến dùng để dự đoán ra biến target

Sau đó sẽ chia target, feature thành 2 bộ là train set và test set với tỉ lệ là 8:2.

Tiếp đó sẽ tạo một pipeline bao gồm các bước sau:

- Chuẩn hoá dữ liệu sử dụng StandardScaler()
- Đưa dữ liệu đã chuẩn hoá vào mô hình và bắt đầu dự đoán.

Ở đây em đã sử dụng tổng cộng 7 thuật toán:

- Random Forest Regressor
- Gradient Forest Regressor
- Linear Regression
- Decision Tree Regressor
- Support Vector Regression
- Kneighbors Regressor

Và cuối cùng để đánh giá mô hình thì em sử dụng 2 giá trị để đánh giá đó là Root mean square error (RMSE), hệ số xác đinh (R2) cùng với phương pháp kiểm chứng chéo (cross validation) bằng cách chia bộ dữ liệu thành 5 phần để đánh giá kết quả của mô hình.

```
Random Forest Regressor:
Cross-validation scores: [0.72650371 0.72847856 0.69325658 0.69172904 0.71414161]
Average score: 0.7108218995673747
Test RMSE: 0.427
Test r2 score: 0.772

Gradient Boosting Regressor:
Cross-validation scores: [0.74420879 0.73488428 0.69719967 0.70486166 0.72730845]
Average score: 0.7216925717203184
Test RMSE: 0.433
Test r2 score: 0.766

Linear Regression
Cross-validation scores: [0.70413265 0.70759546 0.69127005 0.68577171 0.71846168]
Average score: 0.7014463103694868
Test RMSE: 5090423959.627
Test r2 score: -32348009999634862080.000

Decision Tree Regressor:
Cross-validation scores: [0.37760517 0.42139485 0.43478313 0.39725918 0.44179042]
Average score: 0.4145665501256608
Test RMSE: 0.647
Test r2 score: 0.478

Support Vector Regression:
Cross-validation scores: [0.17710004 0.09606182 0.16296448 0.13875673 0.12358977]
Average score: 0.13969456685501902
Test RMSE: 0.437
Test r2 score: 0.761

KNeighbors Regressor:
Cross-validation scores: [0.0974213  0.0981518  0.06948412 0.05161    0.0282128 ]
Average score: 0.06897600429988712
Test RMSE: 0.582
Test r2 score: 0.576
```

Kết quả:

Có thể thấy được 3 mô hinh Random Forest Regressor, Gradient Boostring Regressor, SVR có kết quả dự đoán khá cao khi lần lượt dự đoán đúng được 77%, 76,5% và 76,1% kết quả. Và phương pháp kiểm chứng chéo cũng đã cho ra kết quả tương tự. Mặc dù không quá cao nhưng có thể chấp nhận được vì có một số yếu tố khác ảnh hưởng đến chất lượng của một bộ phim mà không thể được đưa vào bộ dữ liệu.

## 4. KẾT LUẬN

Thông qua đề tài này, nhóm em có cơ hội thực hiện một quá trình phân tích dữ liệu và tạo một mô hình dự đoán rating cho một bộ dữ liệu về phim ảnh.

Quá trình bắt đầu với việc thu thập dữ liệu, làm sạch dữ liệu, tiền xử lý dữ liệu tại đây nhóm em đã thực hiện xử lý giá trị khuyết, ngoại lệ và biến đổi dữ liệu để phù hợp hơn với việc phân tích. Sau đó tại bước khám phá dữ liệu, nhóm chúng em đã có thể trực quan và phân tích dữ liệu từ đó đã tìm ra được một số thông tin hữu ích về những đặc điểm về các bộ phim, mối quan hệ của một số các thuộc tính của bộ phim với điểm đánh giá của nó. Cuối cùng là bước xây dựng mô hình dự đoán điểm số, bằng cách thử nghiệm với nhiều thuật toán khác nhau, bọn em đã có thể tìm ra một số thuật toán có hiệu suất tốt đối với mô hình dự đoán này. Từ đề tài này đã có thể xây dựng ra một mô hình dự đoán điểm số của một bộ phim với độ chính xác đáng kể.

Tuy nhiên, nhóm chúng em đã chưa thể tìm ra cách để đưa được các biến chuỗi như tên diễn viên hay đạo diễn vào để có thể tăng thêm độ chính xác cho mô hình dự đoán.

Nhìn chung, nhóm em đã có thể đạt được mục tiêu là xây dựng được một mô hình dự đoán điểm số. Nhóm em đã có một cái nhìn thấu suốt hơn về những yếu tố đóng góp vào điểm số của một bộ phim bằng một số công cụ phân tích và tương quan dữ liệu. Đề tài này không những tăng thêm sự hiểu biết của nhóm em về nghành công nghiệp phim mà còn cho thấy được sức mạnh của việc phân tích dữ liệu và mô hình dự đoán.

# TÀI LIỆU THAM KHẢO

**[1]** Stephanie (2020) *Winsorize: Definition, examples in easy steps, Statistics How To*. Available at: https://www.statisticshowto.com/winsorize/ (Accessed: 11 December 2023).

**[2]** Saurav Anand (2020) *IMDB score prediction for Movies, Kaggle*. Available at: https://www.kaggle.com/code/saurav9786/imdb-score-prediction-for-movies (Accessed: 11 December 2023).

**[3]** P. Hsu, Yuan-Hong Shen, Xiang-An Xie (2014) *Predicting Movies User Ratings with Imdb Attributes*, published in Rough Sets and Knowledge Technology

**[4]** W. R. Bristi, Z. Zaman and N. Sultana, *Predicting IMDb Rating of Movies by Machine Learning Techniques*, 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-5, doi: 10.1109/ICCCNT45670.2019.8944604.

# PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

| STT | Thành viên | Nhiệm vụ |
|---|---|---|
| 1 | Lê Nguyễn Bá Duy | • Cào dữ liệu (100%)<br>• Làm sạch dữ liệu (50%)<br>• Phân tích dữ liệu (50%)<br>• Tạo mô hình (50%)<br>• Viết báo cáo (100%) |
| 2 | Phạm Tiến | • Phân tích dữ liệu (50%)<br>• Tạo mô hình (50%)<br>• Làm sạch dữ liệu (50%)<br>• Làm slide thuyết trình (100%)<br>• Thuyết trình |