DỰ ĐOÁN SỐ ĐIỂM IMDB CỦA PHIM

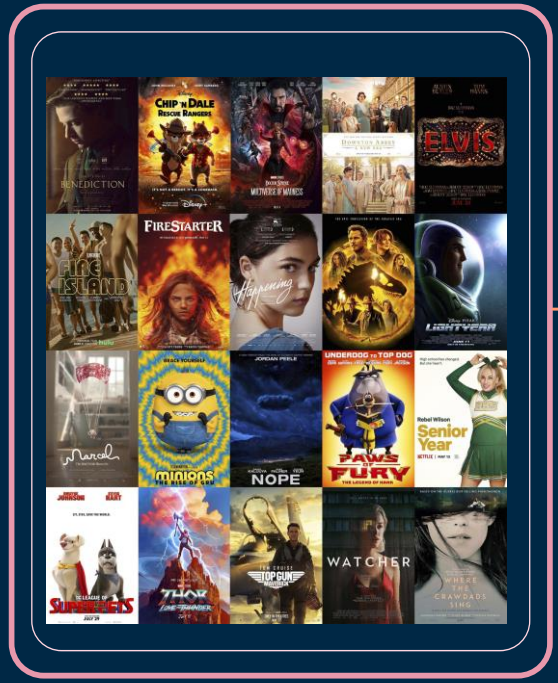
Thành viên:
Nguyễn Công Anh
Trần Xuân Nguyên
Võ Hữu Nam Trường

Bảng phân công nhiệm vụ:

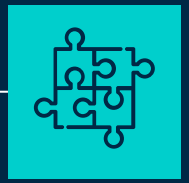
Sinh viên thực hiện	Các nhiệm vụ	Tự đánh giá
Trần Xuân Nguyên	Thu thập dữ liệu	Đã hoàn thành
Nguyễn Công Anh	Trích xuất đặc trưng	Đã hoàn thành
Võ Hữu Nam Trường	Mô hình hóa dữ liệu	Đã hoàn thành

Đặt vấn đề

Trong thời gian gần đây, các phim mới được ra mắt ngày càng nhiều bên cạnh đó cũng có được rất nhiều sự quan tâm chú ý của khán giả. Ngoài nội dung của phim thì các thông tin khác về diễn viên, đạo diễn, nhà làm phim,... rất được nhiều người chú ý đến khi xem một bộ phim. Do đó, dự báo số điểm IMDB của phim là một hình thức để thông qua cách xem điểm số đánh giá về bộ phim, người xem có thể cân nhắc không xem những phim có đánh giá thấp để đỡ mất thời gian



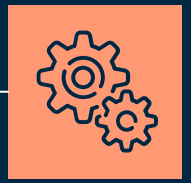
Các bước thực hiện đề tài



01

Thu thập dữ liệu

Sử dụng công cụ để thu thập dữ liệu



02

Xử lí dữ liệu & Trực quan hóa dữ liệu

Xử lí và trực quan hóa dữ liệu

03

Xây dựng mô hình

Xây dựng mô hình dự đoán phù hợp với tập dữ liệu và nhận xét

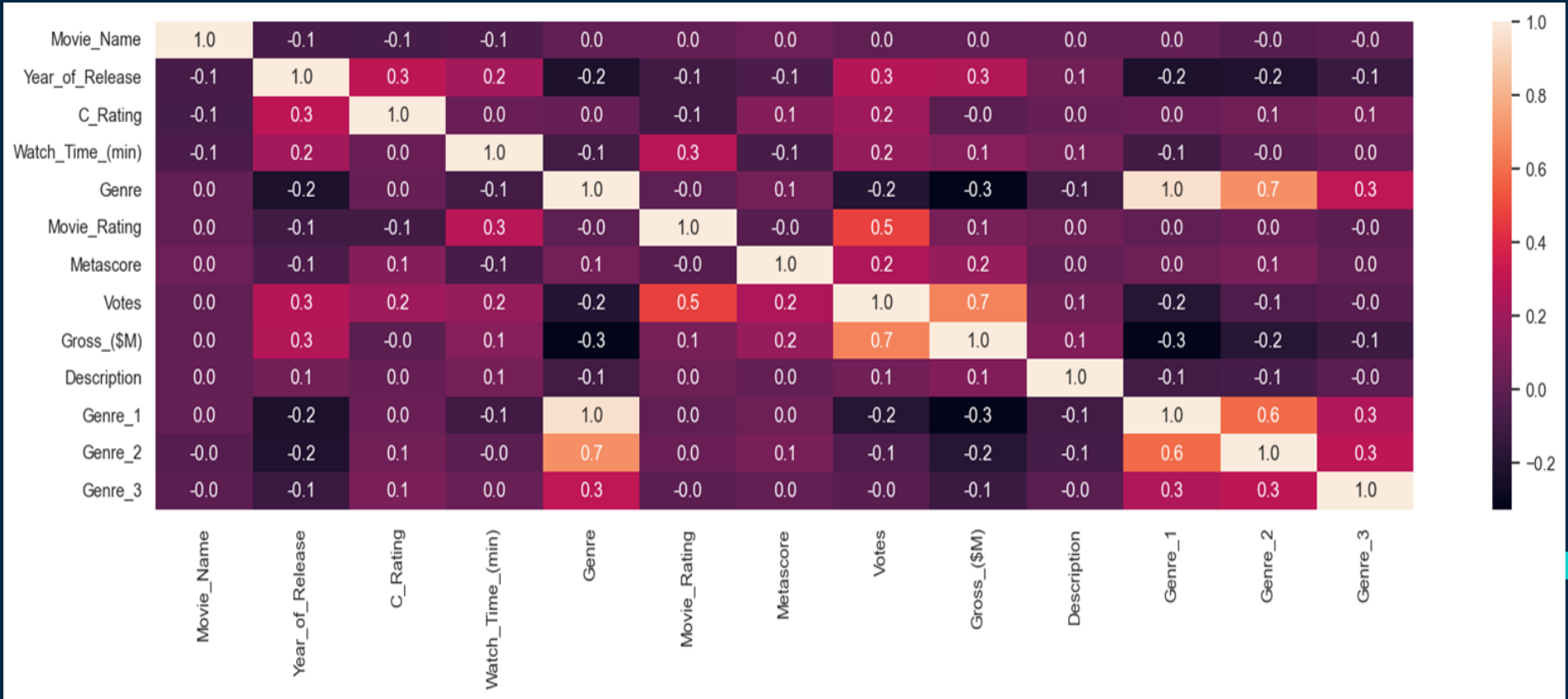
1. Thu thập dữ liệu

Nguồn thu thập dữ liệu:
https://www.imdb.com

Dữ liệu được thu thập bằng BeautifulSoup.

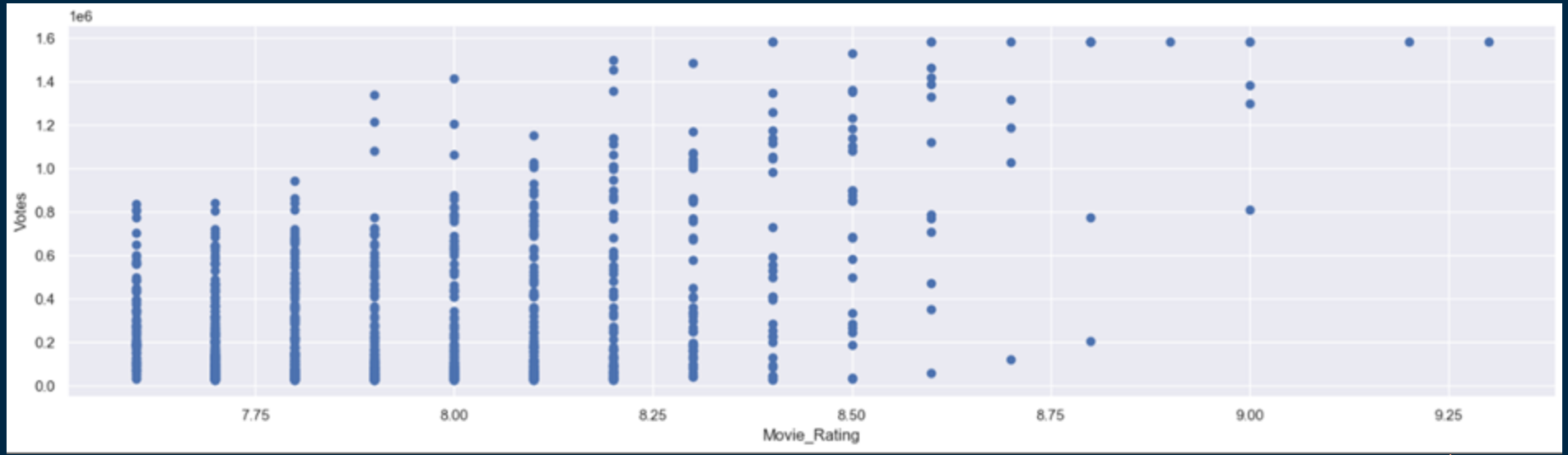
Những đặc trưng đã thu thập được

Tên đặc trưng	Kiểu dữ liệu	Số mẫu dữ liệu trống
Movie_name	string	0
Year_of_Release	int	0
C_Rating	string	16
Watch_Time_(min)	int	0
Genre	string	0
Movie_Rating	float	0
Metascore	float	156
Votes	int	0
Gross_($M)	float	186
Description	string	0



Sự tương quan giữa các đặc trưng với nhau

Ví dụ: Sự tương quan giữa movie_rating là lượng votes



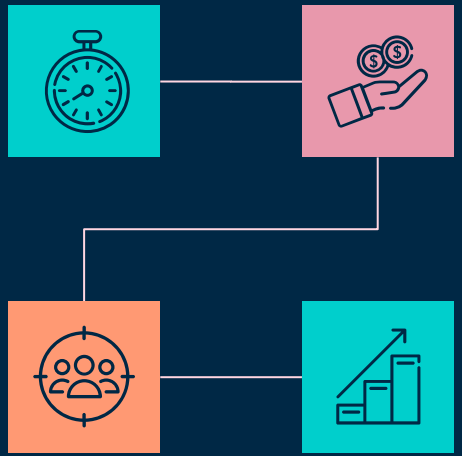
2. Xử lí dữ liệu

Làm sạch dữ liệu

Sử dụng các kỹ thuật xử lí dữ liệu trống và tạo đặc trưng mới

Lựa chọn đặc trưng

Lựa chọn đặc trưng liên quan cho mô hình dự đoán



Xử lí ngoại lệ

Xử lí ngoại lệ cho dữ liệu

Scalling

Thay đổi phạm vi của dữ liệu mà không ảnh hưởng đến dữ liệu

2. Xử lí dữ liệu

2.1. Làm sạch dữ liệu:

2.1.1. Sử dụng các kĩ thuật sau đây để làm sạch dữ liệu:

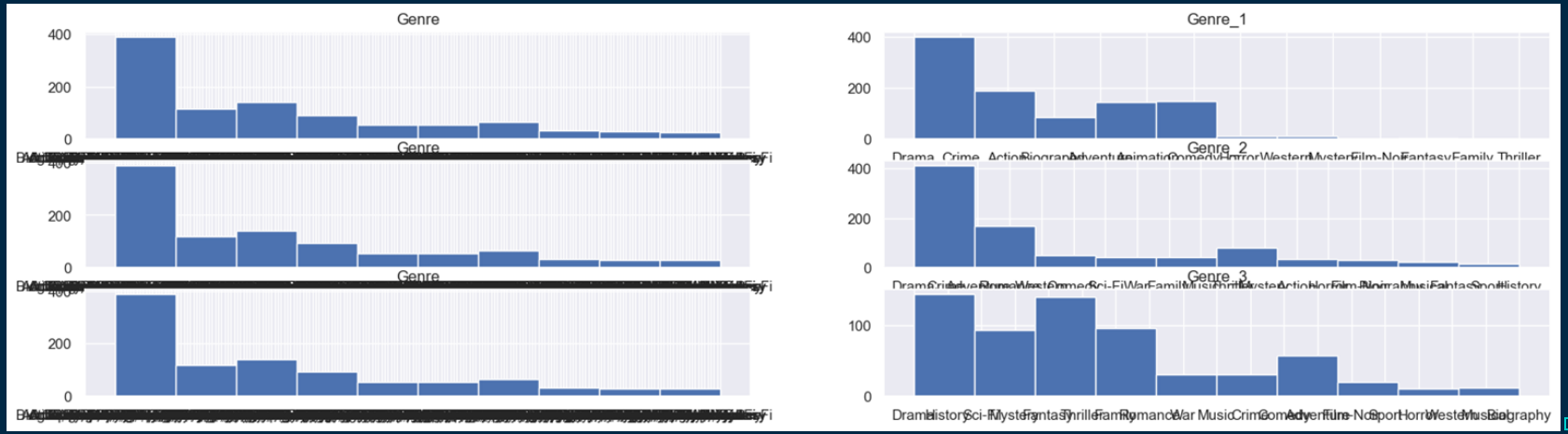
- Xóa cột thứ tự
- Xóa những bộ phim trùng lặp
- Xóa bộ phim có C_Rating chứa dữ liệu NA (Not a number)

2. Xử lí dữ liệu

2.1. Làm sạch dữ liệu:

2.1.2. Tạo đặc trưng mới

Tạo 3 đặc trưng genre 1, 2, 3 là các đặc trưng mới. Ta có hình minh họa như sau



2. Xử lí dữ liệu

2.1. Làm sạch dữ liệu:

2.1.3. Xử lí dữ liệu trống

Các đặc trưng chứa dữ liệu trống bao gồm: C_Rating (16 DL trống), Metascore (156 DL trống) và Gross($) với 186 DL trống

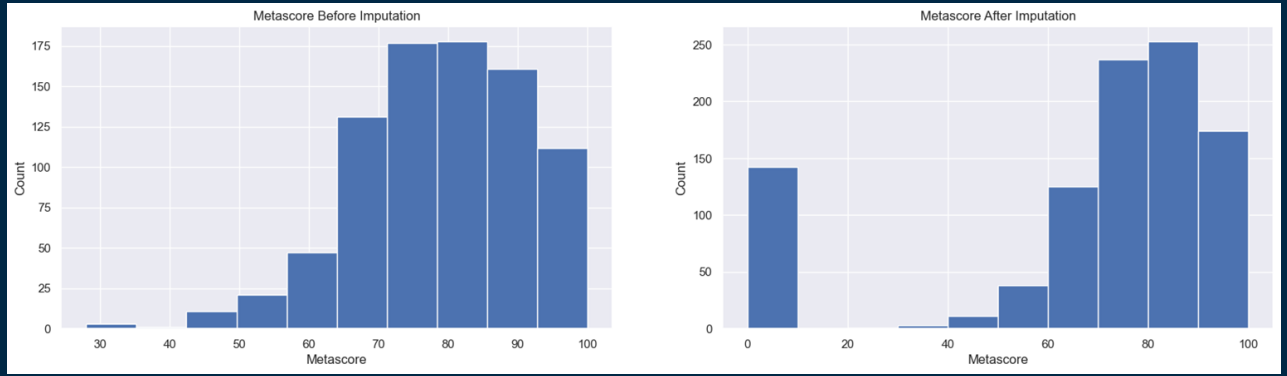
C_Rating	16
Watch_Time_(min)	0
Genre	0
Movie_Rating	0
Metascore	156
Votes	0
Gross_($M)	186

2. Xử lí dữ liệu

2.1. Làm sạch dữ liệu:

2.1.3. Xử lí dữ liệu trống

- Điền vào dữ liệu trống một giá trị bất kì (giá trị 0):



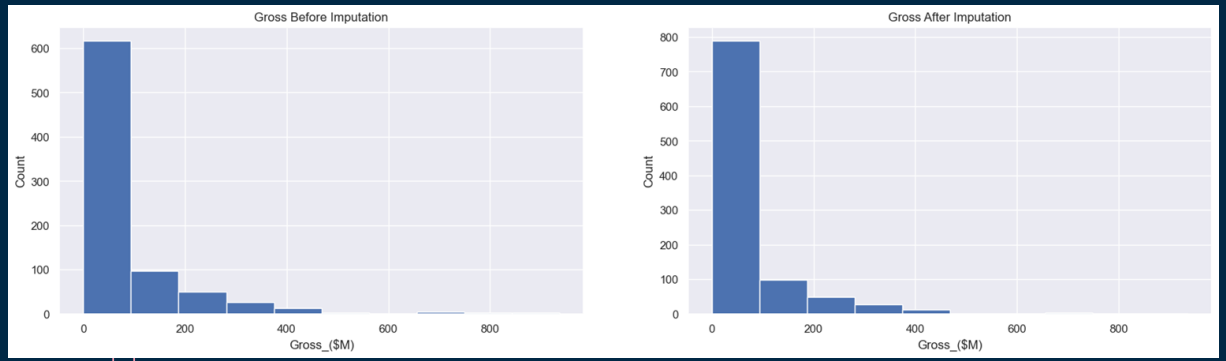
Histogram của Metascore khi điền vào dữ liệu trống giá trị 0

2. Xử lí dữ liệu

2.1. Làm sạch dữ liệu:

2.1.3. Xử lí dữ liệu trống

- Lấp dữ liệu trống bằng giá trị median:



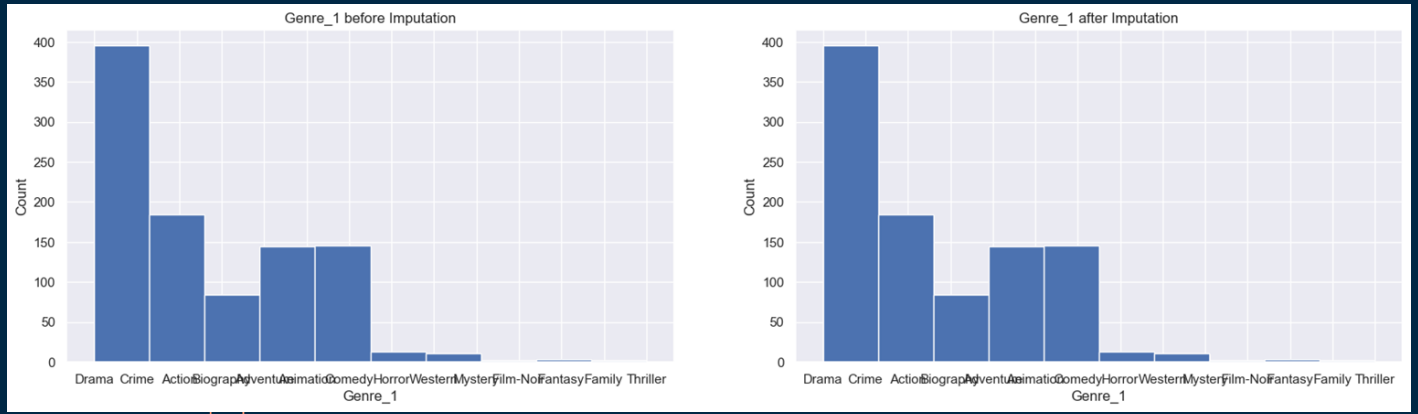
Histogram của Gross trước và sau khi điền vào median

2. Xử lí dữ liệu

2.1. Làm sạch dữ liệu:

2.1.3. Xử lí dữ liệu trống

- Lấy dữ liệu trống bằng giá trị random:

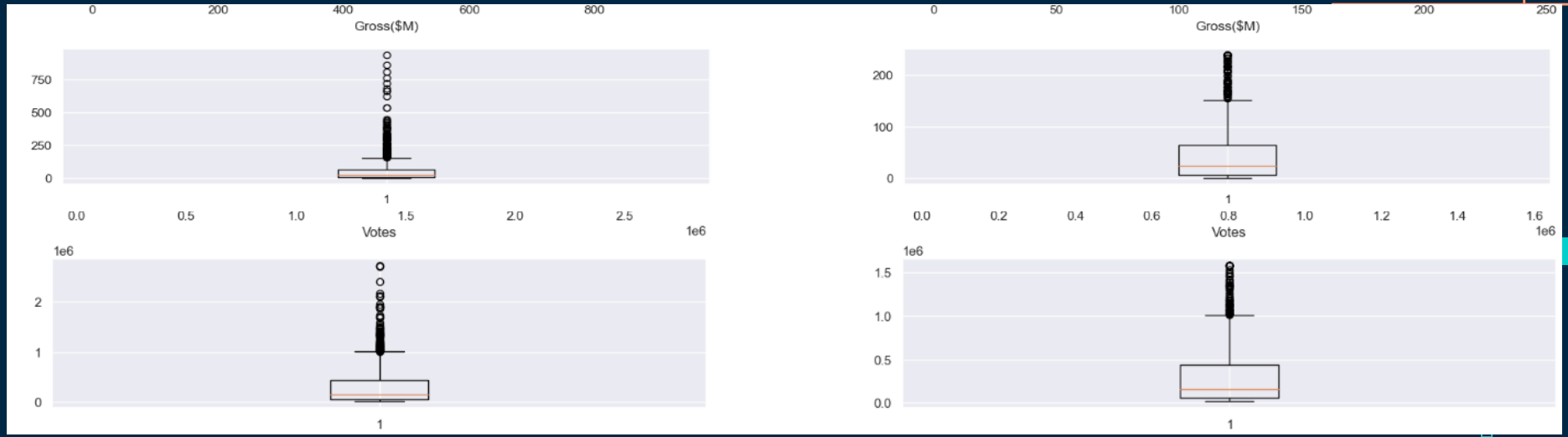


Histogram của Genre trước và sau khi điền vào random

2. Xử lí dữ liệu

2.2. Outliers:

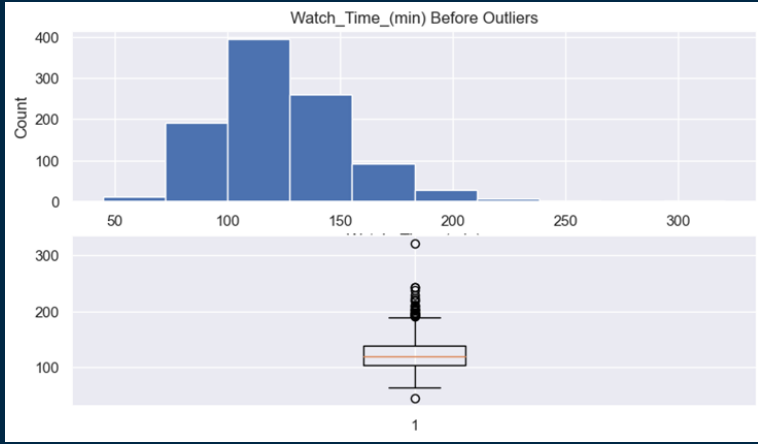
- Đối với Vote và Gross: xử lí ngoại lệ bằng phương pháp tính IQR = phân vị 75 trừ cho phân vị 25 (vì dữ liệu skewed)
- Kết quả sau khi xử lí ngoại lệ như sau

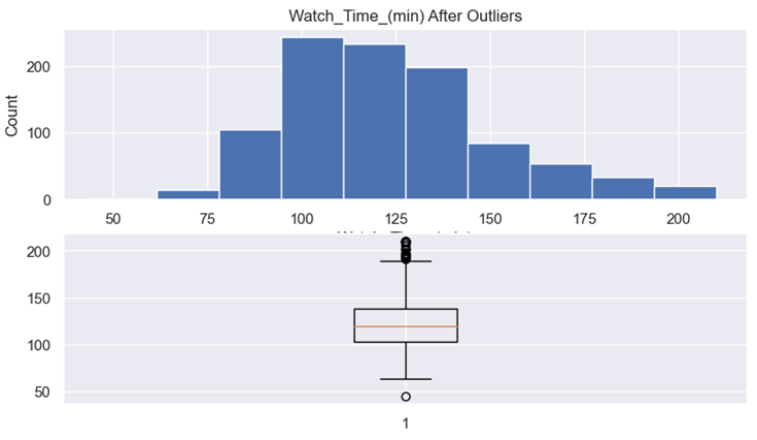


2. Xử lí dữ liệu

2.2. Outliers:

- Đối với Watch_time: Xử lí bằng cách tính upper boundary và lower boundary (Vì dữ liệu distributed)
- Kết quả sau khi xử lí ngoại lệ như sau

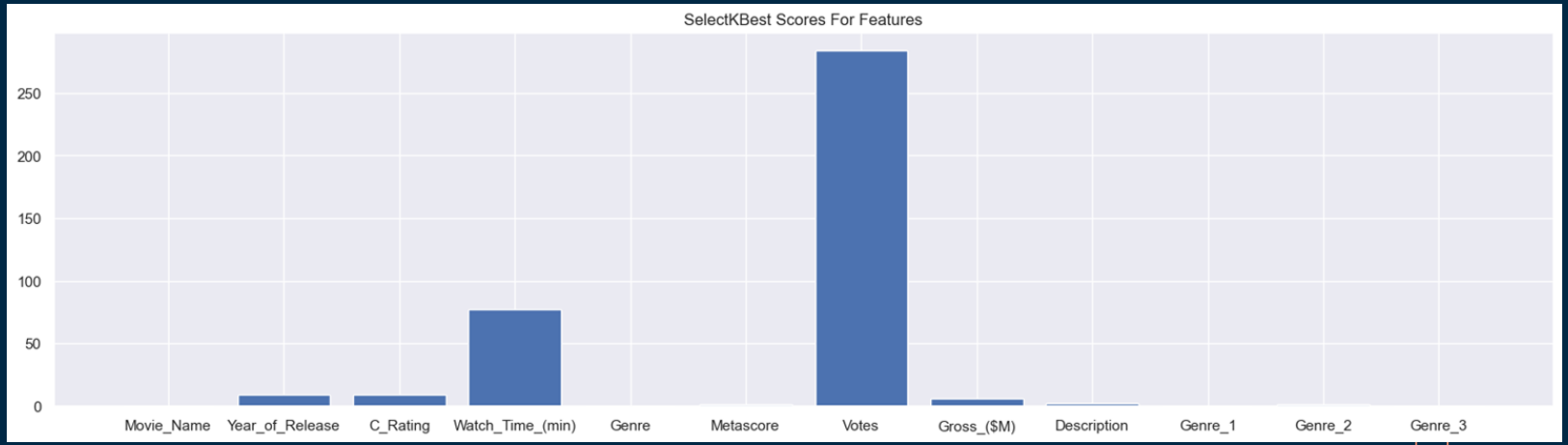




2. Xử lí dữ liệu

2.3. Lựa chọn đặc trưng:

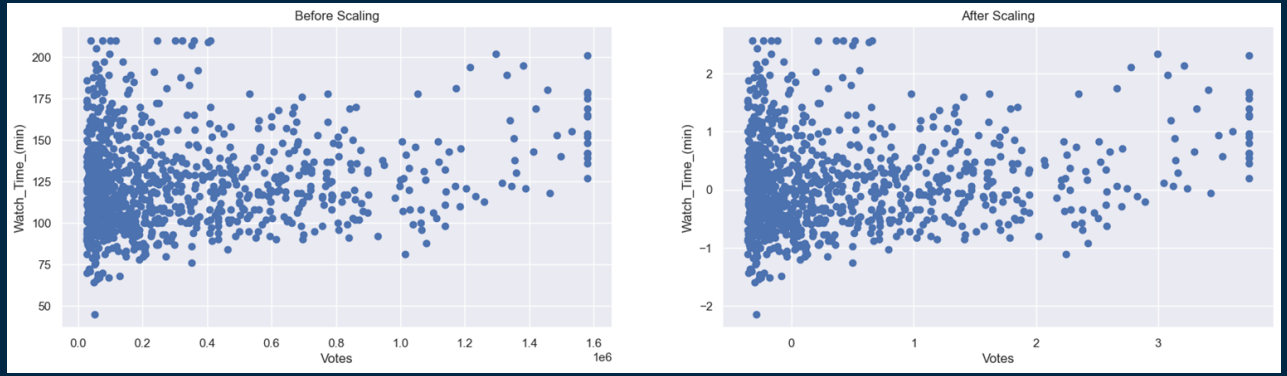
- Sử dụng SelectKBest để có tiêu chuẩn đánh giá mức độ quan trọng, từ đó chọn lọc các đặc trưng cần thiết



2. Xử lí dữ liệu

2.4. Scalling

- Sử dụng phương pháp RobustScaling để giảm độ ảnh hưởng bởi các giá trị ngoại lệ
- Áp dụng vào các đặc trưng đã chọn lọc, ta có kết quả



Dữ liệu về votes và watch time trước và sau khi scale

Xây dựng mô hình

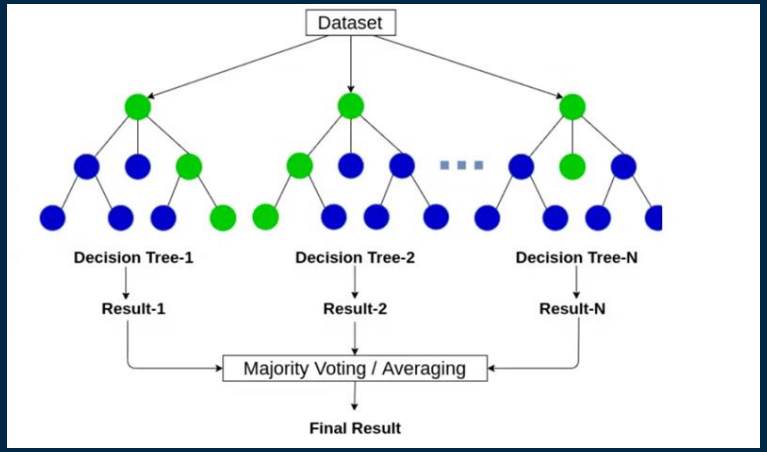
Các mô hình được sử dụng:

- Random Forest Regression
- K-Nearest Neighbors Regression

Xây dựng mô hình

Random Forest

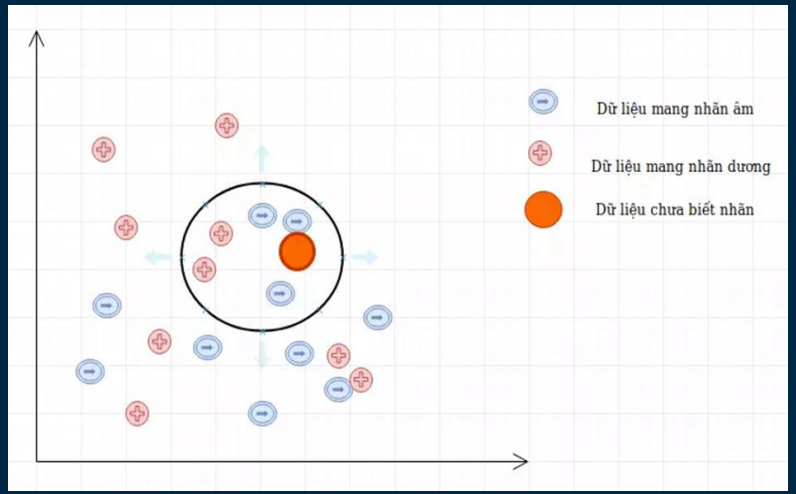
Là một phương pháp thống kê mô hình hóa bằng máy (machine learning statistic) dùng để phục vụ các mục đích phân loại, tính hồi quy và các nhiệm vụ khác bằng cách xây dựng nhiều cây quyết định (Decision tree)



Xây dựng mô hình

K-Nearest Neighbors

Bằng cách tính khoảng cách giữa điểm cần phân lớp và các hàng xóm có sẵn sau đó chọn K các láng giềng gần nhất. Đếm số điểm dữ liệu của mỗi danh mục. Gán các điểm dữ liệu mới cho danh mục đó mà số lượng hàng xóm nhiều nhất.

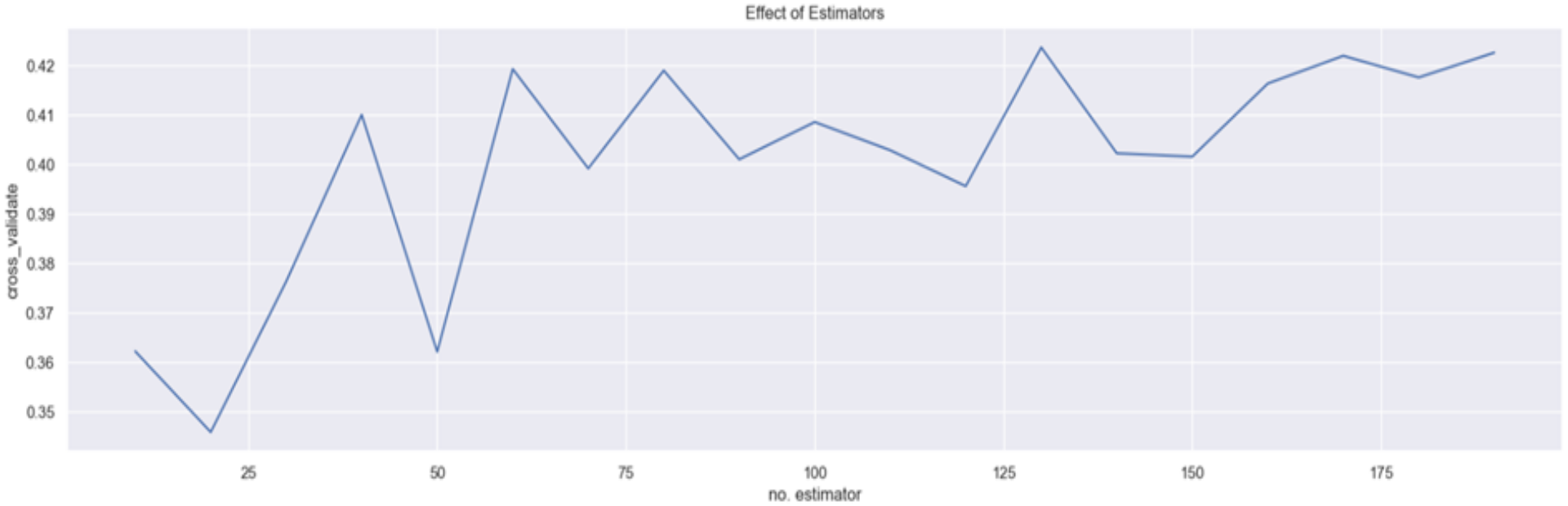


Huấn luyện mô hình

Kết quả huấn luyện mô hình với Small Dataset:

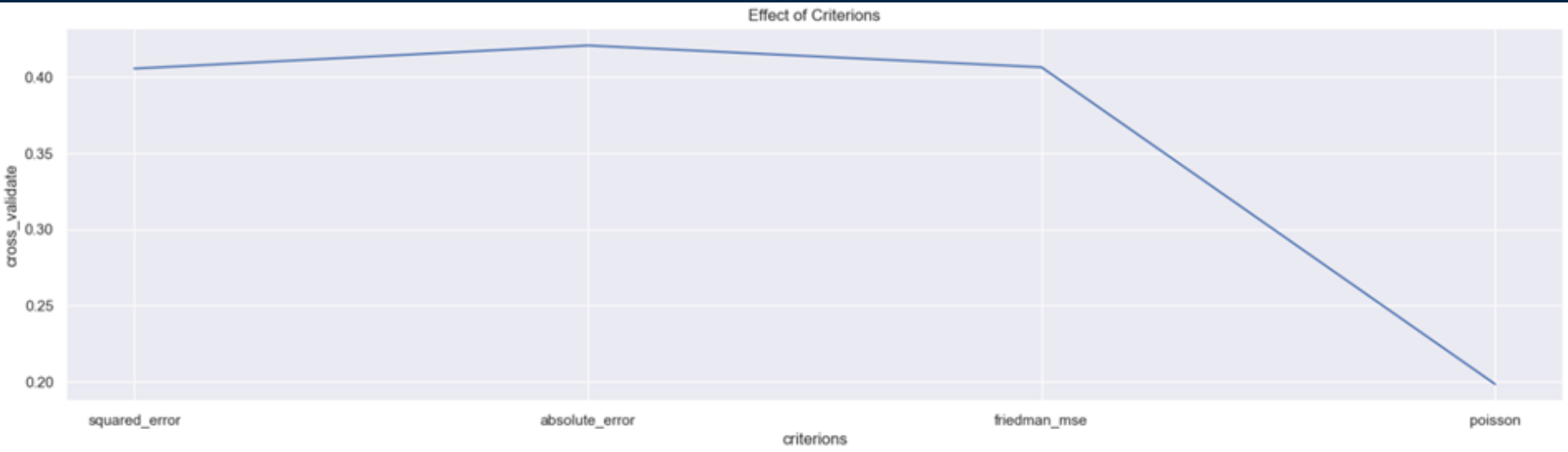
- Random Forest Regression:
  - n_estimators: 80-160 (thay đổi liên tục)
  - criterion: squared_error/absolute_error/friedman_mse (thay đổi liên tục)
- K-Nearest Neighbors Regression:
  - n_neighbors: 8
  - weights: distance

Huấn luyện mô hình



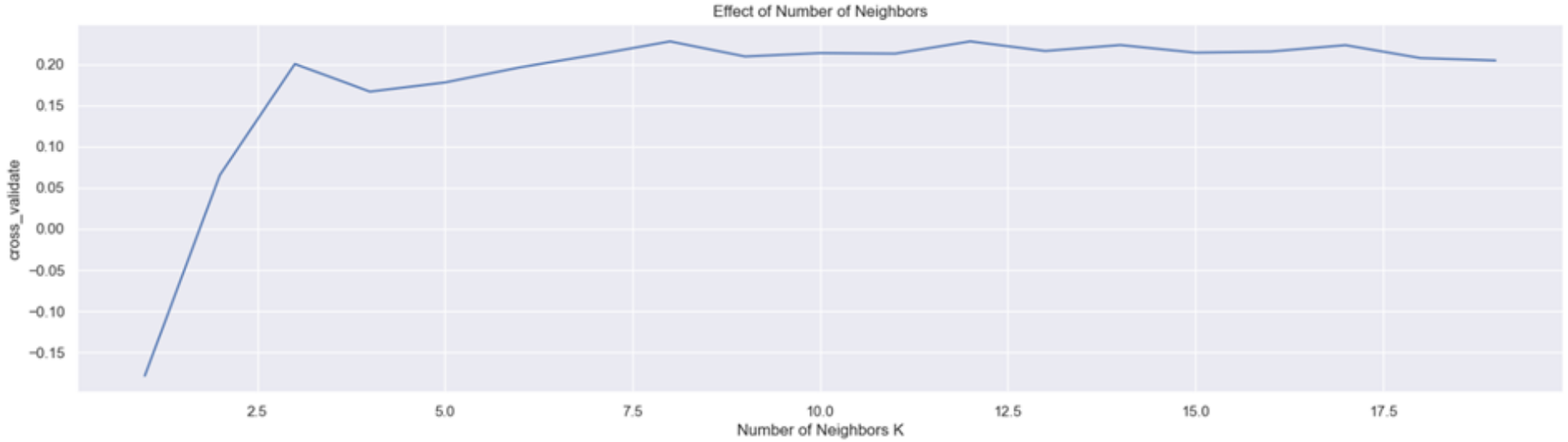
Đồ thị Cross Validation theo n_estimators

Huấn luyện mô hình



Đồ thị Cross Validation theo criterion

Huấn luyện mô hình



Đồ thị Cross Validation theo n_neighbors

Huấn luyện mô hình

Đồ thị Cross Validation theo weights

Huấn luyện mô hình

Kết quả huấn luyện mô hình với Big Dataset:

- Random Forest Regression:
  - n_estimators: 100-190 (thay đổi liên tục)
  - criterion: squared_error/absolute_error/friedman_mse (thay đổi liên tục)
- K-Nearest Neighbors Regression:
  - n_neighbors: 13
  - weights: distance

Đánh giá mô hình

Small Dataset:

	MAE	RMSE	R2 score
Random Forest	~0,1370	~0,1735	~62,02%
K-Nearest Neighbors	~0,1531	~0,1895	~54,66%

Big Dataset:

	MAE	RMSE	R2 score
Random Forest	~0,3939	~0,5133	~63,77%
K-Nearest Neighbors	~0,4291	~0,5558	~57,52%

Kết luận

Những việc đã làm:

- Crawl dữ liệu với số lượng mẫu nhỏ (1000 mẫu) và số lượng mẫu lớn (10000 mẫu)
- Sử dụng các kĩ thuật feature engineering (create new features, data cleaning, missing value imputation, label encoding, outliers, feature selection, feature scaling)
- Xây dựng mô hình (sử dụng mô hình random forest và k nearest neighbor và so sánh 2 mô hình đó).

Kết quả đạt được:

- Xây dựng thành công mô hình dự đoán điểm của phim.
- Thử nghiệm các kỹ thuật nâng cao độ chính xác của mô hình

- Hạn chế:

- Độ chính xác của mô hình dự đoán còn thấp.
- Chưa khai phá hoàn toàn dữ liệu được thu thập.