ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO

ĐỒ ÁN: NHẬP MÔN KHOA HỌC DỮ LIỆU

CHỦ ĐỀ: PHÂN TÍCH THỊ HIẾU XEM PHIM CỦA KHÁN GIẢ VÀ DỰ ĐOÁN XU PHƯỚNG PHIM TRONG TƯƠNG LAI

Giáo viên hướng dẫn: Lê Nhựt Nam

Sinh viên thực hiện: 22120117 – Trần Mạnh Hùng
22120339 – Nguyễn Thị Anh Thi
22120389 – Dương Ngọc Kiều Trinh
22120421 – Nguyễn Đoàn Minh Uyên
22120422 – Nguyễn Phạm Tú Uyên

Hồ Chí Minh, tháng 12 năm 2024

MỤC LỤC

THÔNG TIN THÀNH VIÊN NHÓM

STT	MSSV	HỌ VÀ TÊN
1	22120117	Trần Mạnh Hùng
2	22120339	Nguyễn Thị Anh Thi
3	22120389	Dương Ngọc Kiều Trinh
4	22120421	Nguyễn Đoàn Minh Uyên
5	22120422	Nguyễn Phạm Tú Uyên

I. TỔNG QUAN ĐỒ ÁN

1. Giới thiệu về đề tài

a) Lý do chọn đề tài

Trong những năm gần đây ngành công nghiệp điện ảnh đã chứng kiến sự phát triển vượt bậc, không chỉ về số lượng phim được phát hành mà còn về sự đa dạng trong thể loại. Sự thay đổi nhanh chóng trong thị yếu khán giả cùng với những tiến bộ công nghệ đã khiến các nhà làm phim phải tìm cách dự đoán xu hướng để đáp ứng nhu cầu thị trường.

b) Ý nghĩa của đề tài

Nghiên cứu này không chỉ mang lại cái nhìn toàn diện về thị hiếu và hành vi của khán giả mà còn hỗ trợ các nhà làm phim, nhà phát hành, và nhà đầu tư nắm bắt xu hướng để tối ưu hóa chiến lược kinh doanh. Đồng thời, đề tài còn góp phần vào việc phát triển ngành công nghiệp điện ảnh, hướng đến việc tạo ra các sản phẩm giải trí chất lượng, đáp ứng kỳ vọng của khán giả trong tương lai.

2. Mục tiêu của đồ án

Đề tài tập trung vào 2 khía cạnh chính:

- **Phân tích thị hiếu xem phim của khán giả:** Xác định các yếu tố ảnh hưởng đến sở thích phim như thể loại, thời gian phát hành, thời lượng phim,.. trong thời gian 2 năm trở lại đây (2023-2024).
- **Dự đoán điểm đánh giá của phim:** Sử dụng dữ liệu hiện tại để dự đoán đánh giá của phim, phần nào đánh giá độ thành công của các bộ phim, từ đó hỗ trợ các nhà sản xuất phim đưa ra các quyết định chiến lược.

Việc hoàn thành đồ án còn giúp nhóm nâng cao các kỹ năng cần thiết cho ngành Khoa học dữ liệu như kỹ năng đặt câu hỏi, kỹ năng xử lý dữ liệu và áp dụng các mô hình máy học vào dữ liệu thực tiễn.

II. PHÂN CÔNG CÔNG VIỆC

1. Kế hoạch làm việc

Thời gian	Nội dung
25/11 – 08/12	Trả lời subquestion và đưa ra tiểu kết luận
09/12 – 18/12	Xây dựng mô hình học máy
15/12 – 19/12	Hoàn thiện đồ án

2. Chi tiết phân công công việc

STT	Giai đoạn	Công việc	Phụ trách
1	Tìm nguồn dữ liệu	Tìm chủ đề theo sở thích cá nhân, liệt kê các feature và nguồn dữ liệu để thực hiện cào.	Cả nhóm – họp meet, mỗi cá nhân trình bày 1 chủ đề.
2	Hoàn thành đề cương đồ án	Chọn chủ đề, nguồn dữ liệu, định hình những điều cần làm trong đồ án.	Cả nhóm – họp meet và phân công.
3	Thu thập dữ liệu	Cào dữ liệu trên trang web theo các feature mà nhóm đã chọn.	- Kiều Trinh – cào links phim. - Minh Uyên – viết code cào dữ liệu trong OMDb. - Anh Thi – viết code cào dữ liệu trong JustWatch.
4	Khám phá và tiền xử lý dữ liệu	Format và làm sạch dữ liệu.	- Tú Uyên – kiểm tra các kiểu dữ liệu của cột, xử lý dữ liệu kiểu định tính. - Mạnh Hùng – xử lý dữ liệu kiểu định lượng.
5	Đặt câu hỏi	Đặt câu hỏi, giải thích lợi ích khi trả lời câu hỏi và bước đầu định hình cách trả lời câu hỏi.	Mỗi cá nhân đặt ra 2 câu hỏi và lọc ra 6 câu hỏi hợp lý nhất.
6	Viết báo cáo (P1)	Viết báo cáo cho các công việc đã hoàn thiện (tới giai đoạn đặt câu hỏi)	- Anh Thi – Cào và khám phá dữ liệu. - Mạnh Hùng – tiền xử lý và đặt câu hỏi.
7	Trả lời câu hỏi	Tiền xử lý, trực quan và trả lời	- Minh Uyên – Câu 1 - Anh Thi – Câu 2 + Câu 3 (cũ) - Tú Uyên – Câu 4 - Hùng – Câu 5 - Trinh – Câu 6 + gộp Câu 3 và Câu 6 thành Câu 3 mới
8	Tìm hiểu các mô hình phù hợp cho đồ án	Lựa chọn được mô hình phù hợp cho các vấn đề được đặt ra	- Anh Thi – mô hình cho vấn đề 1 - Minh Uyên – mô hình vấn đề 2
9	Xây dựng mô hình	Tiến hành build mô hình	Anh Thi, Minh Uyên – build mô hình
10	Kiểm tra lại bài làm và format lại code	Chỉnh sửa lại code và các file rõ ràng và dễ hiểu	- Trinh - format phần cào dữ liệu - Minh Uyên – format phần trả lời câu hỏi - Hùng – format các phần còn lại
11	Viết báo cáo (P2 – Hoàn thiện)	Viết báo cáo cho các công việc còn lại	- Anh Thi – mô hình học máy 1 - Tú Uyên – các bước và insight câu hỏi, mô hình học máy 2 - Kiều Trinh – các phần khác

III. NỘI DUNG ĐỒ ÁN

1. Thu thập dữ liệu

a) Nguồn dữ liệu

Nhóm lựa chọn thu thập dữ liệu từ 2 trang web là **JustWatch** và **OMDB**. Đây là hai nền tảng phổ biến và đáng tin cậy, cung cấp thông tin phong phú về các bộ phim, đáp ứng được các yêu cầu của đồ án.

- **JustWatch**: JustWatch là một nền tảng chuyên cung cấp thông tin về các dịch vụ phát trực tuyến, bao gồm danh sách phim và chương trình truyền hình, giá cả, và nền tảng phát hành.
- **OMDb:** OMDb (Open Movie Database) là một cơ sở dữ liệu mở, cung cấp API cho phép truy xuất thông tin chi tiết về phim và chương trình truyền hình cũng như thông tin về đánh giá, xếp hạng, doanh thu và giải thưởng của phim.

b) Quy trình thu thập dữ liệu

Bước 1: Thu thập đường dẫn tới các trang phim trong JustWatch.

- Trang JustWatch chứa nhiều bộ lọc như thể loại, năm phát hành, đánh giá,… Vì chủ đề nhóm là phim nên nhóm lựa chọn thu thập dữ liệu thể loại của các bộ phim phát hành trong năm 2023 – 2024 để cào.
- Danh sách phim trong web không chia thành từng trang mà hiển thị dưới dạng cuộn nên nhóm dùng thư viện Selenium mô phỏng thao tác cuộn chuột để lấy đường dẫn của từng phim.

Bước 2: Cào các thông tin như *Id/Title, Release Time, Runtime, Genre, Age Rating* trong JustWatch từ các đường dẫn đã thu thập.

- Ban đầu nhóm chỉ thực hiện cào *Id/Title* trong trang này và dùng *Id/Title* đó để cào các thông tin khác bằng OMDb API. Tuy nhiên, sau khi thực hiện, nhóm phát hiện các thông tin *Release Time, Runtime, Genre, Age Rating* trong OMDb API có tỉ lệ thiếu cao hơn trong JustWatch nên nhóm quyết định lấy thông tin đó trong trang web này để bổ sung vào dữ liệu bị thiếu trong OMDb API.
- Dữ liệu cào được trong trang này được lưu vào các file pre_data_part{1,2,3}.csv. Trong đó, cột *Id/Title* là Id hoặc Title của phim. Lý do của việc lấy cả 2 thông tin này là do trang OMDb API chỉ cần Id/Title để cào thông tin phim, tuy nhiên một số link phim lại không cung cấp Id mà chỉ có Title nên trong trường hợp không cào được Id, nhóm sẽ dùng Title để cào dữ liệu trong OMDb API.

Bước 3: Cào tất cả thông tin trong OMDB và kết hợp với dữ liệu trong trang JustWatch để giảm thiểu dữ liệu bị thiếu.

- Trang OMDb API này cào dữ liệu dựa trên thông tin Id/Title của phim (cần có) và API phiên bản miễn phí.
- Với các dữ liệu bị thiếu khi cào bằng API mà có đủ trong JustWatch, nhóm bổ sung phần thiếu đó bằng dữ liệu trong JustWatch để giảm thiểu giá trị bị thiếu.
- Dữ liệu sau khi cào và kết hợp được lưu vào file csv có tên là movie_data_part{1,2,3} để tiện cho quá trình xử lý dữ liệu.

2. Khám phá và tiền xử lý dữ liệu

a) Sơ bộ các bước khám phá và tiền xử lí dữ liệu

- Tổng hợp dữ liệu
- Kiểm tra số lượng cột, dòng và kiểu dữ liệu của từng đặc trưng
- Ý nghĩa của từng cột dữ liệu
- Phân bố dữ liệu của những cột quan trọng
- Xử lí giá trị trùng lặp, giá trị thiếu

b) Chi tiết

- Hợp nhất các nguồn dữ liệu thành 1 bộ dữ liệu hoàn chỉnh để sử dụng
- Loại bỏ các dòng dữ liệu trùng lặp
- Kiểm tra dữ liệu và chuyển đổi về kiểu dữ liệu mong muốn
- Kiểm tra sự phân bố của dữ liệu, vẽ biểu đồ để thể hiện rõ sự phân bố đó
- Xử lí giá trị thiếu: điền trung bình hoặc trung vị cho các cột số hoặc điền giá trị phổ biến phụ thuộc vào sự phân bố của dữ liệu, xóa đi những cột có tỉ lệ thiếu quá cao
- Không thực hiện xử lý outliers mà để nó ở các bước tiếp theo để quan sát và phân tích sự ảnh hưởng của nó đối với mô hình từ đó quyết định cách xử lý hợp lý.

c) Kết quả

- Tạo được bộ dữ liệu nhất quán cho các tác vụ sau đó
- Dữ liệu đã được làm sạch (xử lí các giá trị thiếu hoặc trùng lặp, loại bỏ những cột có tỉ lệ thiếu cao)

3. Đặt các câu hỏi có ý nghĩa và trả lời

Câu hỏi 1: Với từng thể loại khác nhau thì đánh giá của khán giả khác nhau như thế nào ? Liệu số lượng bình chọn có ảnh hưởng tới đánh giá của từng thể loại hay không?

Lợi ích của việc trả lời câu hỏi này:
Giúp chúng ta tìm hiểu đánh giá của khán giả với từng thể loại phim khác nhau, từ đó phần nào diễn giải sở thích, khẩu vị xem phim của họ.

Thực hiện:

- Sử dụng cột IMDb Rating, IMDb Votes và các cột thể loại phim.
- Trực quan sự phân bố phim và điểm trung bình (dùng công thức Bayesian Average) của từng thể loại.

Kết luận:

- *Số lượng không tỷ lệ thuận với chất lượng*: Drama là thể loại phổ biến nhất nhưng xếp hạng trung bình chỉ ở mức khá. Trong khi đó, History, Biography và Sport mặc dù có số lượng phim ít hơn nhưng xếp hạng trung bình rất cao, cho thấy rằng số lượng lớn không đảm bảo chất lượng.
- *Thể loại ít nhưng chất lượng cao*: History, Biography và Sport là có nội dung sâu sắc, đáp ứng tốt kỳ vọng của đối tượng khán giả cụ thể và phù hợp đầu tư phát triển.
- *Thể loại phổ biến nhưng chất lượng trung bình*: Drama, Comedy và Documentary có số lượng lớn nhưng xếp hạng chỉ ở mức trung bình có thể là do thiếu sự đầu tư trong sản xuất để đáp ứng được khán giả.
- *Thể loại kém hấp dẫn*: Sci-Fi, Horror và Musical vừa có số lượng phim ít vừa có xếp hạng tháp cho thấy cần được khai thác và đầu tư tốt hơn.

Câu hỏi 2: Các phim thuộc thể loại nào có nhiều khả năng được đề cử hoặc giành giải thưởng nhất?

Lợi ích của việc trả lời câu hỏi này:
- Giúp các nhà làm phim định hướng chọn thể loại để tăng khả năng được đề cử và giành giải thưởng.
- Tối ưu hoá chiến lược phát triển nội dung dựa trên xu hướng thành công.
- Nhà đầu tư có thể định hướng tài trợ vào các dự án có khả năng được đánh giá cao.
- Hiểu rõ xu hướng thị hiếu và tiêu chí đánh giá từ các hội đồng giải thưởng.

Thực hiện:
- Sử dụng cột Win, Nomination và các cột thể loại
- Vẽ biểu đồ cột (bar chart) để thể hiện tổng số đề cử và giải thưởng của từng thể loại

Kết luận: Thể loại Drama dẫn đầu cả về số đề cử và thắng giải, khẳng định vị trí quan trọng của nó trong ngành công nghiệp điện ảnh. Các thể loại Comedy, Histogry và Adventure cũng có tiềm năng lớn, trong khi các thể loại khác cần nhiều nỗ lực hơn để thu hút sự chú ý.

Câu hỏi 3: Nhóm tuổi (Age rating) ảnh hưởng thế nào đến sự phổ biến của các thể loại phim qua số lượt bình chọn IMDb?

Lợi ích của việc trả lời câu hỏi này:
- *Hiểu rõ mức độ quan tâm của khán giả*: Tỉ trọng IMDb Votes thể hiện mức độ yêu thích và quan tâm thực tế của khán giả đối với từng thể loại phim trong mỗi nhóm tuổi, giúp đánh giá chính xác thị hiếu.
- *Định hướng chiến lược sản xuất và đầu tư*: Nhà sản xuất và nhà đầu tư có thể dựa vào tỉ trọng bình chọn để ưu tiên phát triển các thể loại phim được khán giả ở từng độ tuổi ưa chuộng nhất, tối ưu hóa lợi nhuận.
- *Phát hiện xu hướng thị trường*: Tỉ trọng IMDb Votes có thể cho thấy sự thay đổi trong sở thích của từng nhóm tuổi, giúp các nhà phát hành phim cập nhật danh mục thể loại để thu hút khán giả.
- *Nâng cao chất lượng nền tảng xem phim*: Các nền tảng phân phối phim trực tuyến có thể gợi ý nội dung phù hợp hơn cho người dùng dựa trên xu hướng đã phân tích, cải thiện trải nghiệm cá nhân hóa.

Thực hiện:
- Sử dụng cột Age rating, IMDb Votes và các cột thể loại.
- Bar chart trực quan số lượt bình chọn trunh bình và heat map cho các nhóm tuổi, thể loại và tỉ trọng IMDb Votes.

Kết luận:
- *Nhóm tuổi PG-13 và R*: là hai nhóm khán giả quan trọng đóng góp lớn vào sự phổ biến và thành công của các thể loại như Action, Drama và Thriller. Các nhà làm phim nên tập trung khai thác đối tượng thanh thiếu niên và trưởng thành để tăng sức hút và mức độ phổ biến.
- *Nhóm tuổi nhỏ*: chủ yếu quan tâm đến Animation và Family nhưng ít đóng góp vào IMDb Votes, cần chiến lược phù hợp để tăng độ nhận diện và thu hút khán giả.

Câu hỏi 4: Thời lượng phim (Runtime) ảnh hưởng như thế nào đến số lượng IMDb Votes và IMDb Rating?

Lợi ích của việc trả lời câu hỏi này:
- Trả lời câu hỏi này giúp chúng ta xác định thời lượng lý tưởng của một bộ phim để thu hút khán giả và đạt được điểm đánh giá cao.

Thực hiện:
- Sử dụng cột Runtime, IMDb Votes và IMDb Rating.
- Xem phân phối của Runtime và biểu đồ bar chart so sánh sự khác biệt trong số lượng IMDb Votes và IMDb Rating theo nhóm thời lượng phim.

Kết luận:
- Phim chất lượng cao: Thời lượng lý tưởng trên 120 phút, phù hợp với các tác phẩm chuyên sâu, mang lại trải nghiệm đa chiều.
- Phim thương mại phổ thông: thời lượng 80-120 phút, cân bằng giữa nội dung hấp dẫn và độ dài vừa phải để thu hút khán giả.

Câu hỏi 5: Thể loại nào có xu hướng phát triển mạnh nhất qua các quý của năm dựa trên số lượng phim phát hành?

Lợi ích của việc trả lời câu hỏi này:
- Trả lời câu hỏi này giúp dự đoán xu hướng phát triển về thể loại của ngành công nghiệp điện ảnh để định hướng đầu tư vào các thể loại tiềm năng trong tương lai.

Thực hiện:
- Sử dụng cột Realease time và các cột thể loại.
- Trực quan hóa xu hướng của mỗi thể loại qua các quý theo năm.

Kết luận:
- Những thể loại ít phổ biến như Short và Western cho thấy thị trường hẹp và ít tiềm năng phát triển.
- Các thể loại như Drama, Documentary và Comedy thống trị thị trường với sự tăng trưởng ổn định, phù hợp với thị hiếu khán giả. Nhà sản xuất có thể tập trung vào những thể loại này để tối ưu hóa lượng phát hành.

4. Xây dựng mô hình học máy

Vấn đề 1: Dự đoán phim có thành công hay không (Ratings > 7.0), phim thành công thường có những yếu tố nào (thời lượng, độ tuổi giới hạn, lượt bình chọn, thể loại).

Mục tiêu mô hình:
- Xây dựng mô hình machine learning để dự đoán phim có thành công hay không dựa trên các đặc trưng như IMDb Rating, số lượt bình chọn, thể loại phim và các thông tin liên quan
- Hiểu rõ các yếu tố ảnh hưởng đến sự thành công của bộ phim.
- Định hướng và cải thiện các dự án phim trong tương lai.

Xử lý dữ liệu:
- Gắn nhãn mục tiêu "success" (thành công) dựa trên điều kiện: IMDb Rating > 7.0

Chia dữ liệu:
- Tập huấn luyện: 70%
- Tập kiểm tra: 30%
- Giải quyết vấn đề mất cân bằng dữ liệu bằng SMOTE (Synthetic Minority Oversampling Technique).

Mô hình Random Forest Class:
- Mô hình mạnh mẽ, hiểu quả với dữ liệu có nhiều đặc trưng
- Không yêu cầu chuẩn hoá đặc trưng

Mô hình Gradient Boosting:
- Xử lý được nhiều loại dữ liệu
- Chống overfitting tốt

Độ đo:
- *ROC Curve*: đồ thị biểu diễn khả năng phân loại của mô hình phân loại nhị phân, giúp đánh giá mô hình phân loại nhị phân ở nhiều ngưỡng phân loại khác nhau. Một mô hình tốt sẽ có một ROC Curve gần với góc trên cùng bên trái của đồ thị và AUC Score.
- *AUC Score*: AUC là diện tích dưới ROC Curve, đo lường khả năng phân biệt giữa các lớp (dương tính và âm tính) của mô hình. AUC có giá trị trong khoảng từ 0 đến 1. Nó giúp cung cấp cái nhìn tổng quan về hiệu suất của mô hình mà không phụ thuộc vào ngưỡng phân loại cụ thể. Một AUC cao cho thấy mô hình có khả năng phân biệt tốt giữa các lớp.
- *Confusion matrix*: bảng hiển thị số lượng dự đoán chính xác và sai của mô hình phân loại, giúp đánh giá hiệu suất của mô hình trong việc phân loại các mẫu. Confusion Matrix cho phép bạn xác định được số lượng dự đoán đúng và sai cho mỗi lớp (class), giúp đánh giá các yếu tố như Precision (độ chính xác), Recall (độ nhạy), và Accuracy (độ chính xác tổng thể).

Đánh giá mô hình:
- *Confusion matrix*: Gradient Boosting ít nhầm lẫn hơn với lớp 0 nhưng số dự đoán đúng cho lớp 1 lại thấp hơn Random Forest.
- *ROC Curve và AUC Score*: Cho thấy khả năng phân biệt giữa hai lớp của hai mô hình là ngang nhau, Gradient Boosting có xu hướng tốt hơn một chút ở giai đoạn đầu. Random Forest có đường cong ổn định nhưng không vượt trội ở bất kỳ khu vực nào.

Đánh giá chung: Gradient Boosting là mô hình tốt hơn, hiệu suất tổng thể cao hơn. Với mục tiêu là xây dựng mô hình dự đoán chính xác các phim thành công, Gradient Boosting là mô hình được khuyến nghị do hiệu suất tổng thể cao hơn và khả năng giảm số lượng dự đoán sai cho lớp 0 (phim không thành công).

Vấn đề 2: Dự đoán Ratings của phim dựa trên các thông tin được cho trước, bao gồm thời lượng, giới hạn độ tuổi, thể loại, số lượt đề cử, số lần thắng giải và số lượt bình chọn.

Kết quả mô hình: Rating phim
- Mô hình đã dự đoán rating phim với sai số khoảng 5-7%. Đây là sai số tạm chấp nhận được với dữ liệu hiện tại.
- Đánh giá mô hình có thể được cân nhắc sử dụng vào bài toán điền dữ liệu rating thiếu nếu các feature input được xử lý tốt và có nhiều dữ liệu train hơn.

Xử lý dữ liệu:
- Xử lý outliers trong cột IMDb Rating và chọn cột này làm nhãn.

Chia dữ liệu:
- Tập huấn luyện: 70%
- Tập kiểm tra: 15%
- Tập kiểm định: 15%

Mô hình cơ sở: Linear Regression
- Được sử dụng làm mô hình cơ sở. Đây là mô hình đơn giản cung cấp một điểm so sánh ban đầu.

Mô hình Random Forest Regressor và Gradient Boosting:
- Phù hợp với nhiều kiểu dữ liệu
- Xử lý tốt dữ liệu phức tạp
- Không yêu cầu chuẩn hóa đặc trưng
- Giảm overfitting bằng cách trung bình hóa nhiều cây
- Hoạt động hiệu quả trên dữ liệu nhiễu

Độ đo:
- MAE (Mean Absolute Error): mức độ sai lệch trung bình giữa giá trị dự đoán và thực tế. MAE càng thấp thì mô hình càng chính xác. Giúp hiểu rõ được mức độ sai lệch tổng thể của các dự đoán, biết được mô hình dự đoán có đúng với thực tế hay không.
- RMSE (Root Mean Squared Error): đo lường sai số bình phương trung bình, nó nhằm vào các lỗi trong mô hình. RMSE càng thấp mô hình càng chính xác. Giúp kiểm tra xem mô hình có đang mắc phải các sai số lớn hay không.

Hai chỉ số này giúp đánh giá mô hình một cách toàn diện.

Đánh giá chung:

- Mô hình Linear Regression có hiệu suất tệ nhất, trong khi đó mô hình Random Forest Regressor có hiệu suất tốt nhất dựa trên so sánh MAE và RMSE.
- Khi áp dụng mô hình Random Forest trên tập test, hiệu suất tương đương giữa tập valid và tập test cho thấy mô hình tương đối đáng tin cậy, có khả năng tổng quát hóa tốt.

IV. ĐÁNH GIÁ MỨC ĐỘ HOÀN THÀNH ĐỒ ÁN

STT	Nội dung	Nhận xét	Đánh giá
1	Thu thập dữ liệu	Thu thập đủ dữ liệu, đạt chuẩn yêu cầu dữ liệu của đề bài.	100%
2	Khám phá và tiền xử lý	Dữ liệu đã được làm sạch.	100%
3	Đặt câu hỏi và trả lời	Câu hỏi hay, tìm được insight phù hợp với chủ đề.	100%
4	Xây dựng mô hình học máy	Lựa chọn mô hình phù hợp, về cơ bản vẫn giải quyết được vấn đề ban đầu nhóm đặt ra.	100%
5	Viết báo cáo	Đầy đủ nội dung, trình bày đẹp.	100%

TÀI LIỆU THAM KHẢO

[1] *API reference*. (n.d.). Scikit-Learn. Retrieved December 19, 2024, from https://scikit-learn.org/1.5/api/index.html

[2] *API reference — pandas 2.2.3 documentation*. (n.d.). Pydata.org. Retrieved December 19, 2024, from https://pandas.pydata.org/docs/reference/

[3] *Matplotlib documentation — Matplotlib 3.9.3 documentation*. (n.d.). Matplotlib.org. Retrieved December 19, 2024, from https://matplotlib.org/stable/

[4] *API reference — seaborn 0.13.2 documentation*. (n.d.). Pydata.org. Retrieved December 19, 2024, from https://seaborn.pydata.org/api.html

[5] *Using the Bayesian average in custom ranking*. (n.d.). Algolia Documentation. Retrieved December 19, 2024, from https://www.algolia.com/doc/guides/managing-results/must-do/custom-ranking/how-to/bayesian-average/