TRƯỜNG CAO ĐẲNG FPT POLYTECHNIC

BÁO CÁO DỰ ÁN TỐT NGHIỆP

Chuyên ngành Xử Lý Dữ Liệu

CINEMA DATA ANALYST

Các thành viên:

- **Huỳnh Nhật Phi – PS38116 (Nhóm trưởng)**
- **Phạm Văn Long – PS39583**
- **Nguyễn Xuân Sơn – PS38373**
- **Phạm Thế Vinh – PS38437**
- **Hà Sơn Bình – PS38037**

GVHD: Thầy Văn Công Khanh

Tp.Hồ Chí Minh, tháng 7 năm 2025

MỤC LỤC

Lời Cảm Ơn

Nhóm chúng em xin gửi lời cảm ơn chân thành đến Thầy Văn Công Khanh – người đã tận tình hướng dẫn, hỗ trợ và đồng hành cùng nhóm trong suốt quá trình thực hiện dự án.

Nhờ sự chỉ dẫn tận tâm, những góp ý thẳng thắn và định hướng rõ ràng từ Thầy, nhóm đã học được rất nhiều kiến thức thực tế về xử lý và phân tích dữ liệu – từ cách tiếp cận bài toán, làm sạch dữ liệu, đến trực quan hóa và rút ra những kết luận có giá trị.

Thầy không chỉ giúp nhóm nhìn vấn đề một cách sâu sắc hơn mà còn tạo động lực để cả nhóm hoàn thành dự án một cách nghiêm túc và có trách nhiệm.

Nhóm cũng xin cảm ơn các anh chị, bạn bè đã hỗ trợ trong quá trình thu thập dữ liệu, chia sẻ kinh nghiệm, và góp ý để nhóm hoàn thiện bài tốt hơn.

Dù đã rất cố gắng nhưng do thời gian hạn chế và kinh nghiệm thực tế chưa nhiều, dự án chắc chắn vẫn còn những thiếu sót. Rất mong nhận được sự thông cảm và góp ý từ Thầy để nhóm có thể rút kinh nghiệm và cải thiện trong những lần sau.

Nhóm chúng em xin chân thành cảm ơn Thầy!

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

Tên đề tài: CINEMA DATA ANALYST

Lớp: DP 19301 **Ngành:** Xử lý dữ liệu – CNTT

1. Tinh thần thái độ làm việc:

.....

.....

2. Mức độ hoàn thành đề tài:

.....

.....

3. Kết quả đạt được:

.....

.....

Nhận xét và đánh giá chung:

.....

.....

.....

Điểm đánh giá (bằng số): ..... **Bằng chữ:** .....

Ngày ..... tháng ..... năm 2025

Giáo viên hướng dẫn
(ký và ghi rõ họ tên)

1 GIỚI THIỆU ĐỀ TÀI

1.1 LÝ DO CHỌN ĐỀ TÀI

Trong bối cảnh ngành công nghiệp giải trí và điện ảnh phát triển mạnh mẽ, dữ liệu về phim ảnh được tạo ra với số lượng lớn và đa dạng. Việc khai thác, phân tích các dữ liệu này mang lại giá trị lớn cho việc dự đoán xu hướng, đánh giá chất lượng nội dung, cũng như hỗ trợ các nền tảng trực tuyến trong việc gợi ý phim phù hợp cho người xem.

Bộ dữ liệu **TMDB 5000 Movie Dataset** cung cấp thông tin phong phú về hàng nghìn bộ phim, bao gồm các chỉ số về doanh thu, đánh giá, thể loại, diễn viên và đội ngũ sản xuất. Đây là nguồn dữ liệu thích hợp để áp dụng các kỹ thuật xử lý dữ liệu, trực quan hóa và phân tích nhằm rút ra những thông tin hữu ích.

Vì vậy, nhóm lựa chọn đề tài này để vừa rèn luyện kỹ năng làm sạch và phân tích dữ liệu, vừa tạo ra sản phẩm có giá trị ứng dụng thực tiễn trong lĩnh vực giải trí.

1.2 MỤC TIÊU PHẠM VI

Mục tiêu:

- Thu thập, tiền xử lý và phân tích bộ dữ liệu TMDB 5000 để tìm ra các yếu tố ảnh hưởng đến thành công của một bộ phim.
- Xây dựng các biểu đồ trực quan giúp người đọc dễ dàng nắm bắt xu hướng và mối quan hệ giữa các đặc trưng dữ liệu.
- Đề xuất các mô hình hoặc tiêu chí dự đoán mức độ thành công của phim dựa trên dữ liệu đã phân tích.

Phạm vi:

- Dữ liệu sử dụng: TMDB 5000 Movie Dataset (bao gồm các tập tin tmdb_5000_movies.csv và tmdb_5000_credits.csv).
- Công cụ xử lý: Python và các thư viện hỗ trợ phân tích dữ liệu (Pandas, NumPy, Matplotlib, Seaborn…).
- Giới hạn: Chỉ phân tích dữ liệu trong bộ TMDB 5000, không thu thập thêm dữ liệu bên ngoài.
- Thời gian thực hiện: Từ [tháng bắt đầu] đến [tháng kết thúc] năm 2025.

1.3 SẢN PHẨM MỤC TIÊU

Bộ dữ liệu đã được xử lý và làm sạch: loại bỏ giá trị thiếu, dữ liệu trùng lặp, định dạng lại các trường thông tin.

Báo cáo phân tích dữ liệu: trình bày kết quả thống kê, các biểu đồ trực quan và nhận xét.

Bảng tổng hợp kết luận & đề xuất: liệt kê các yếu tố ảnh hưởng mạnh đến thành công của phim và gợi ý hướng phát triển nội dung.

Mô hình dự đoán: thử nghiệm mô hình dự đoán doanh thu hoặc điểm đánh giá dựa trên các đặc trưng đầu vào.

1.4 CẤU TRÚC BÁO CÁO

Để viết một báo cáo phân tích dữ liệu hiệu quả, bạn cần thực hiện theo các bước sau:

❖ Xác định mục tiêu của báo cáo

Bạn cần xác định mục tiêu của báo cáo nói về gì. Bạn muốn cung cấp cho người đọc thông tin gì?

❖ Thu thập và phân tích dữ liệu

Bước tiếp theo là thu thập và phân tích dữ liệu. Bạn sẽ sử dụng dữ liệu này để tạo ra các kết luận và đề xuất trong báo cáo của mình.

Khi thu thập dữ liệu, hãy đảm bảo rằng dữ liệu là chính xác, đáng tin cậy và phù hợp với mục tiêu của báo cáo.

Khi phân tích dữ liệu, hãy sử dụng các kỹ thuật phân tích dữ liệu phù hợp để tìm ra các xu hướng, mô hình và mối tương quan trong dữ liệu.

❖ Viết nội dung báo cáo

Nội dung báo cáo của bạn nên viết bao gồm các phần sau:

- Giới thiệu: Giới thiệu mục tiêu của báo cáo, phạm vi của dữ liệu và các phương pháp phân tích dữ liệu đã được sử dụng.
- Phân tích dữ liệu: Trình bày các kết quả của phân tích dữ liệu.

- Kết luận: Tóm tắt các kết luận chính của báo cáo và các đề xuất hành động.

❖ Trình bày báo cáo

Bước cuối cùng là trình bày báo cáo. Bạn có thể trình bày báo cáo bằng cách in ra, trình bày trực tiếp hoặc tải lên trên internet. Khi trình bày bạn nên sử dụng ngôn ngữ rõ ràng,súc tích và dễ hiểu. Không nên sử dụng quá nhiều thật ngữ chuyên môn nếu không cần thiết. Tóm tắt kết luận chính của báo cáo ở đầu và cuối báo cáo.

2 TỔNG QUAN DỮ LIỆU VÀ CÔNG NGHỆ

2.1 MÔ TẢ DỮ LIỆU

Bộ dữ liệu nhóm mình sử dụng là **TMDB 5000 Movie Dataset**, lấy từ nền tảng *The Movie Database (TMDB)* — một trang web lớn chuyên cung cấp thông tin về phim ảnh. Bộ dữ liệu này gồm hai file chính:

1. **tmdb_5000_movies.csv** – Chứa thông tin chi tiết của hơn 4.800 bộ phim như:
    - **title**: tên phim
    - **genres**: thể loại (hành động, hài, tình cảm…)
    - **budget**: kinh phí sản xuất
    - **revenue**: doanh thu
    - **release_date**: ngày phát hành
    - **vote_average** và **vote_count**: điểm đánh giá và số lượt đánh giá
    - Các thông tin khác như quốc gia sản xuất, công ty sản xuất…
2. **tmdb_5000_credits.csv** – Chứa thông tin về **diễn viên** và **đội ngũ sản xuất**:
    - **cast**: danh sách diễn viên
    - **crew**: danh sách thành viên ekip (đạo diễn, biên kịch, quay phim…)
    - **job** và **department**: vai trò và bộ phận đảm nhận

Tổng cộng, dữ liệu có hàng chục cột và nhiều kiểu thông tin khác nhau: số, chuỗi, ngày tháng và cả dạng JSON.

Khi mới tải về, dữ liệu vẫn còn **nhiều giá trị bị thiếu, không đồng nhất**, nên nhóm phải làm bước **làm sạch dữ liệu** trước khi phân tích.

2.2 CÁC CÔNG CỤ, CÔNG NGHỆ, THƯ VIỆN

Để xử lý và phân tích bộ dữ liệu, nhóm sử dụng các công cụ và thư viện sau:

- **Python**: Ngôn ngữ chính dùng để đọc, xử lý và phân tích dữ liệu.
- **Pandas**: Hỗ trợ thao tác với dữ liệu dạng bảng (đọc file CSV, lọc, sắp xếp, tính toán…).
- **NumPy**: Xử lý dữ liệu dạng mảng, hỗ trợ các phép toán nhanh.
- **Matplotlib** và **Seaborn**: Dùng để vẽ biểu đồ, trực quan hóa dữ liệu một cách trực quan và dễ hiểu.
- **Jupyter Notebook**: Môi trường chạy code và hiển thị kết quả ngay trong cùng một trang, thuận tiện khi làm báo cáo.
- **Tableau**: Dùng để trực quan hóa dữ liệu, tạo dashboard và biểu đồ tương tác.
- **GitHub**: Lưu trữ và quản lý mã nguồn, thuận tiện cho việc làm việc nhóm.

2.3 TỔNG QUAN MÔ HÌNH, THUẬT TOÁN LIÊN QUAN

Trong quá trình phân tích và trực quan hóa bộ dữ liệu TMDB 5000, nhóm chủ yếu áp dụng các phương pháp và kỹ thuật thuộc lĩnh vực xử lý dữ liệu và thống kê mô tả, thay vì các mô hình dự đoán phức tạp. Cụ thể:

1. **Thống kê mô tả (Descriptive Statistics)**
   - ❖ Sử dụng các chỉ số như giá trị trung bình (*mean*), trung vị (*median*), giá trị lớn nhất (*max*), nhỏ nhất (*min*) và độ lệch chuẩn (*standard deviation*) để nắm tổng quan về dữ liệu.
   - ❖ Giúp nhận biết xu hướng chung, phân bố dữ liệu và phát hiện các giá trị bất thường.
2. **Tiền xử lý dữ liệu (Data Preprocessing)**
   - ❖ **Xử lý dữ liệu thiếu (Missing Values)**: Điền giá trị thay thế hoặc loại bỏ các bản ghi bị thiếu.
   - ❖ **Chuẩn hóa dữ liệu**: Chuyển đổi định dạng ngày tháng, tách dữ liệu JSON trong cột *genres* và *cast*.

- **Loại bỏ dữ liệu nhiễu**: Xóa các dòng chứa thông tin không hợp lệ (ví dụ: doanh thu = 0 nhưng vẫn có điểm đánh giá cao).

3. **Trực quan hóa dữ liệu (Data Visualization)**
   - **Biểu đồ cột và biểu đồ thanh**: So sánh số lượng phim theo thể loại, quốc gia sản xuất.
   - **Biểu đồ phân tán (Scatter Plot)**: Thể hiện mối quan hệ giữa kinh phí sản xuất và doanh thu.
   - **Biểu đồ tròn (Pie Chart)**: Thể hiện tỉ lệ các thể loại phim.
   - **Biểu đồ hộp (Box Plot)**: Phát hiện ngoại lệ trong doanh thu và điểm đánh giá.
4. **Phân tích tương quan (Correlation Analysis)**
   - Tính hệ số tương quan giữa các biến số như *budget*, *revenue*, *vote_average*, *vote_count* để xem mức độ liên hệ.
   - Giúp nhận biết yếu tố nào tác động mạnh đến doanh thu hoặc điểm đánh giá.

3 QUY TRÌNH XỬ LÍ DỮ LIỆU

3.1 KHÁM PHÁ DỮ LIỆU

Bắt đầu, nhóm mình tải bộ dữ liệu TMDB 5000 về máy và mở bằng Excel để xem qua các cột và nội dung bên trong. Sau đó, dùng Python (Pandas) để coi thử kích thước, tên các cột, và vài dòng đầu tiên để biết dữ liệu gồm những gì.

Ở bước này, nhóm nhận ra dữ liệu khá nhiều cột, gồm cả dạng số, chữ, ngày tháng, và có mấy cột dạng JSON như *genres* hay *cast*. Ngoài ra, cũng phát hiện một số chỗ bị thiếu dữ liệu hoặc để giá trị 0.

3.2 LÀM SẠCH DỮ LIỆU

Sau khi nắm sơ bộ, nhóm bắt đầu xử lý để dữ liệu gọn gàng hơn:

- Mấy cột quan trọng như tên phim, ngân sách, doanh thu mà bị thiếu thì xóa hẳn dòng đó.
- Cột ít quan trọng như *homepage* mà trống thì để nguyên, coi như giá trị rỗng.

- Xóa các dòng trùng lặp.
- Chỉnh lại định dạng ngày tháng về kiểu năm-tháng-ngày cho đồng nhất.
- Loại mấy phim có ngân sách hoặc doanh thu = 0 mà vẫn có điểm cao (vì khả năng thông tin sai).
- Tách dữ liệu trong các cột JSON để dễ xử lý, ví dụ từ *genres* lấy ra danh sách thể loại phim.

3.3 CHUYỂN ĐỔI DỮ LIỆU

Khi dữ liệu đã sạch, nhóm chuyển đổi để dễ phân tích:

- Tạo thêm cột lợi nhuận = doanh thu – ngân sách.
- Lấy năm phát hành từ cột *release_date*.
- Chuyển cột thể loại sang dạng cột riêng (One-Hot) để sau này vẽ biểu đồ và phân tích.
- Đảm bảo các con số về ngân sách và doanh thu đều tính theo USD.
- Gom nhóm theo thể loại, năm hoặc quốc gia để tạo bảng thống kê nhanh.

Nhờ mấy bước này, dữ liệu nhìn gọn, dễ đọc, và quan trọng là sẵn sàng để đem đi phân tích, vẽ biểu đồ bằng Python hoặc Tableau.

4 XÂY DỰNG VÀ PHÁT TRIỂN SẢN PHẨM

4.1 CÁC BƯỚC XÂY DỰNG SẢN PHẨM

Sau khi có dữ liệu sạch, nhóm bắt đầu thực hiện các bước để xây dựng sản phẩm cuối:

1. **Xác định yêu cầu**: Chốt mục tiêu phân tích (ví dụ tìm yếu tố ảnh hưởng doanh thu, xem xu hướng thể loại phim).
2. **Chuẩn bị môi trường**: Cài Python, các thư viện cần thiết (Pandas, NumPy, Matplotlib, Seaborn) và Tableau.

3. **Xử lý dữ liệu**: Làm sạch, chuyển đổi, tạo thêm cột mới như *profit*, *release_year*.
4. **Phân tích**: Dùng Python để tính toán, tạo bảng thống kê, xem mối quan hệ giữa các yếu tố.
5. **Trực quan hóa**: Vẽ biểu đồ bằng Python và thiết kế dashboard trên Tableau để người xem dễ nắm ý chính.
6. **Hoàn thiện báo cáo**: Tổng hợp kết quả, hình ảnh, nhận xét để đưa vào báo cáo cuối.

4.2 MÔ HÌNH DỰ ĐOÁN / PHÂN TÍCH / BÁO CÁO

Do mục tiêu chính là phân tích dữ liệu và trực quan hóa, nhóm không tập trung vào mô hình AI phức tạp, mà chủ yếu dùng:

- **Thống kê mô tả**: Xem giá trị trung bình, cao nhất, thấp nhất của ngân sách, doanh thu, điểm đánh giá.
- **Phân tích tương quan**: Tìm mối liên hệ giữa ngân sách và doanh thu, hoặc giữa số lượt đánh giá và điểm trung bình.
- **Phân loại xu hướng**: So sánh doanh thu trung bình giữa các thể loại hoặc giữa các năm.

Ngoài ra, nhóm thử nghiệm **hồi quy tuyến tính (Linear Regression)** để ước lượng doanh thu dựa trên ngân sách và số lượt đánh giá, nhưng chỉ ở mức tham khảo.

4.3 TRIỂN KHAI DASHBOARD, CÔNG CỤ, GIAO DIỆN

Nhóm sử dụng **Tableau** để tạo dashboard trực quan từ dữ liệu đã xử lý:

- **Biểu đồ cột**: So sánh số lượng phim theo thể loại.
- **Biểu đồ phân tán**: Thể hiện mối quan hệ ngân sách – doanh thu.
- **Biểu đồ đường**: Thể hiện xu hướng doanh thu qua các năm.
- **Bảng tổng hợp**: Hiển thị top 10 phim doanh thu cao nhất.

Dashboard được thiết kế gọn gàng, dễ đọc, có màu sắc phân biệt rõ ràng. Người dùng có thể lọc theo năm hoặc thể loại để xem thông tin chi tiết.

5 KẾT QUẢ ĐÁNH GIÁ

5.1 KẾT QUẢ ĐẠT ĐƯỢC

Sau khi hoàn thành các bước xử lý và phân tích, nhóm thu được:

- Bộ dữ liệu đã sạch, thống nhất về định dạng, không còn giá trị thiếu hoặc trùng lặp ở các cột quan trọng.
- Thêm các trường dữ liệu mới (*profit*, *release_year*) hỗ trợ phân tích tốt hơn.
- Bộ biểu đồ trực quan bằng Tableau giúp nhìn nhanh xu hướng và mối quan hệ trong dữ liệu.
- Một số phát hiện thú vị, ví dụ:
  - ❖ Ngân sách và doanh thu có mối tương quan dương, nhưng không phải lúc nào ngân sách lớn cũng đồng nghĩa doanh thu cao.
  - ❖ Thể loại phim **Action** và **Adventure** thường có doanh thu cao nhất.
  - ❖ Giai đoạn 2010–2015 là thời kỳ nhiều phim đạt doanh thu đột biến.

5.2 ĐÁNH GIÁ HIỆU QUẢ

Về mục tiêu: Hoàn thành đúng yêu cầu ban đầu là phân tích và trực quan hóa dữ liệu phim TMDB.

Về độ trực quan: Dashboard rõ ràng, dễ lọc, phù hợp để người xem không cần kiến thức sâu vẫn hiểu.

Về quy trình: Các bước xử lý dữ liệu khá mạch lạc, dễ lặp lại nếu áp dụng cho bộ dữ liệu khác.

Về nhóm: Mỗi thành viên đều tham gia vào cả phần xử lý dữ liệu và thiết kế dashboard, nên hiểu tổng thể dự án.

5.3 SO SÁNH TRƯỚC/SAU XỬ LÍ, TÍNH CHÍNH XÁC, TỐC ĐỘ, ĐỘ ỔN ĐỊNH

Độ sạch dữ liệu

- Trước xử lý: Nhiều giá trị thiếu, dữ liệu nhiễu, trùng lặp.
- Sau xử lý: 100% các cột quan trọng đầy đủ và đúng định dạng.

Tính chính xác

- Trước xử lý: Nhiều bản ghi sai (ngân sách/doanh thu = 0).
- Sau xử lý: Loại bỏ sai sót, dữ liệu phản ánh đúng thực tế hơn.

Tốc độ phân tích

- Trước xử lý: Chậm do phải lọc thủ công và gặp lỗi.
- Sau xử lý: Nhanh hơn vì dữ liệu đã chuẩn hóa.

Độ ổn định

- Trước xử lý: Kết quả phân tích dễ thay đổi do dữ liệu lỗi.
- Sau xử lý: Ổn định, có thể tái sử dụng cho nhiều phân tích khác.

6 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.2 TỔNG KẾT NỘI DUNG THỰC HIỆN

Dự án đã trải qua đầy đủ các bước từ thu thập và khám phá dữ liệu, làm sạch – chuyển đổi – chuẩn hóa, đến phân tích và trực quan hóa bằng Tableau. Nhóm đã áp dụng các công cụ như GitHub để quản lý mã nguồn và chia sẻ dữ liệu, cùng Tableau để tạo dashboard trực quan. Quy trình được thực hiện tuần tự, đảm bảo tính logic và có thể tái áp dụng cho các bộ dữ liệu khác.

6.3 ĐÁNH GIÁ NHỮNG GÌ ĐẠT ĐƯỢC

- Hoàn thành đúng mục tiêu đề ra: làm sạch dữ liệu, phân tích và xây dựng dashboard.
- Cải thiện đáng kể chất lượng dữ liệu: loại bỏ giá trị thiếu, dữ liệu nhiễu, bản ghi sai.
- Dashboard trực quan, dễ sử dụng, cho phép người dùng lọc và quan sát xu hướng nhanh chóng.
- Phát hiện một số mối quan hệ quan trọng trong dữ liệu, hỗ trợ ra quyết định.
- Quy trình xử lý dữ liệu đã tối ưu, tiết kiệm thời gian và đảm bảo kết quả ổn định.

6.1 ĐỀ XUẤT CẢI TIẾN VÀ MỞ RỘNG ỨNG DỤNG THỰC TẾ

- Mở rộng tập dữ liệu để phân tích nhiều khía cạnh hơn, ví dụ bổ sung thông tin đánh giá của khán giả.
- Ứng dụng thêm các mô hình dự đoán (machine learning) để dự báo xu hướng doanh thu hoặc đánh giá phim.
- Tích hợp dashboard với dữ liệu cập nhật theo thời gian thực.
- Nâng cấp giao diện người dùng, tăng tính tương tác và trải nghiệm trực quan.
- Áp dụng quy trình này cho các lĩnh vực khác như thương mại, tài chính, giáo dục để khai thác giá trị từ dữ liệu.

7 TỔNG KẾT

7.1 TÀI LIỆU THAM KHẢO

7.2 PHỤ LỤC