Model

* Thử nghiệm nhiều thuật toán: Gồm các thuật toán phổ biến trong thư viện Surprise như SVD, SVD++, SlopeOne, NMF, các thuật toán KNN, và BaselineOnly.

Đánh giá bằng Cross-Validation:

* Sử dụng cross\_validate để tính các chỉ số RMSE (Root Mean Square Error) và MAE (Mean Absolute Error).
* Sử dụng 3-fold cross-validation (cv=3) để đảm bảo tính tổng quát của đánh giá.

Lợi ích :

Xử lý dữ liệu thiếu

Giảm chiều dữ liệu

Hiệu quả trong xử lý cold-start ( hiện khi một **người dùng mới** tham gia hệ thống và chưa có bất kỳ đánh giá, lịch sử mua sắm hoặc tương tác nào với các sản phẩm. Hệ thống không biết sở thích của người dùng, dẫn đến khó khăn trong việc đưa ra gợi ý phù hợp )

7.1. User-based Filtering (Collaborative Filtering)

Phương pháp này dự đoán điểm đánh giá (rating) của người dùng cho các phim chưa được đánh giá dựa trên sự tương đồng giữa người dùng. Các bước thực hiện gồm:

* Đo lường sự tương đồng giữa người dùng:
* Sử dụng Cosine Similarity hoặc Pearson Correlation để tính toán mức độ tương đồng giữa các vector rating của người dùng.
* Dự đoán rating dựa trên người dùng tương tự:
* Tính trung bình có trọng số các điểm đánh giá từ những người dùng có mức độ tương đồng cao.

7.2. Content-based Filtering

Hệ thống này gợi ý phim dựa trên nội dung như thể loại, diễn viên, hoặc mô tả phim.

* Vector hóa đặc trưng nội dung: Sử dụng TF-IDF, CountVectorizer, hoặc Sentence Transformer để chuyển đổi nội dung phim thành vector số.
* Tính toán mức độ tương đồng: Sử dụng Cosine Similarity để đo lường độ tương đồng giữa các vector nội dung.

Kết luận: Sentence Transformer được đánh giá là phương pháp phù hợp nhất nhờ khả năng biểu diễn ngữ nghĩa tốt.

7.3. Hybrid Approach (Weighted-based Filtering)

Kết hợp phương pháp User-based Filtering và Content-based Filtering để cải thiện hiệu suất gợi ý.

Cách 1: Lồng ghép (Sequential Combination)

* Lấy danh sách phim đã xem của người dùng
* Dùng hàm user-based để lấy phim gợi ý dựa trên sự tương đồng giữa người dùng.
* Với mỗi phim được gợi ý từ user-based, sử dụng hàm content-based để tìm phim tương tự về nội dung.
* Kết hợp điểm số từ hai phương pháp bằng trung bình có trọng số (alpha), loại bỏ phim đã xem và sắp xếp kết quả theo điểm số.

Hạn chế:

* Phụ thuộc vào chất lượng dữ liệu đầu vào.
* Không xử lý tốt trường hợp cold-start hoặc phim có điểm thấp.

Cách 2: Song song (Parallel Combination)

* Tạo từ điển để lưu điểm số gợi ý từ cả user-based và content-based.
* Tính điểm cho từng phim với trọng số alpha (user-based) và 1-alpha (content-based).
* Cộng điểm nếu phim đã có trong từ điển, sắp xếp và chọn top\_n phim.

Hạn chế:

* Không loại bỏ được phim đã xem khỏi danh sách gợi ý.
* Hiệu suất thấp với dữ liệu lớn, tiêu tốn nhiều bộ nhớ.

Cách 3: Kết hợp với chuẩn hóa MinMaxScaler

* Tính điểm user-based: Dự đoán điểm số cho các phim chưa xem dựa trên sự tương đồng giữa người dùng.
* Chuẩn hóa điểm số: Dùng MinMaxScaler để chuẩn hóa về khoảng [0, 1].
* Tính độ tương đồng content-based: Sử dụng cosine similarity để tính mức độ tương đồng giữa các phim.
* Tính điểm số kết hợp (hybrid score): Sử dụng công thức kết hợp điểm từ hai phương pháp.

Gợi ý phim: Loại bỏ phim đã xem, sắp xếp và chọn các phim có điểm số cao nhất.

Ưu điểm:

* Giải quyết các hạn chế của các phương pháp hybrid khác: không gợi ý phim đã xem, gợi ý chủ động. Tuy nhiên nhược điểm là sẽ khá nặng và chạy lâu đối với những thiết bị yếu.

Kết luận chung:

User-based Filtering phù hợp cho người dùng có lịch sử đánh giá phong phú.

Content-based Filtering hữu ích khi có ít dữ liệu người dùng hoặc cần gợi ý dựa trên nội dung.

Hybrid Approach mang lại sự cân bằng, nhưng cần cải thiện hiệu suất và xử lý cold-start.

Cách 3 (chuẩn hóa MinMaxScaler) được đánh giá hiệu quả nhất nhờ khắc phục nhiều hạn chế và đảm bảo tính chính xác của gợi ý.

Kết luận lại hết:  
Trong bài báo cáo này, chúng tôi đã tiến hành phân tích các thuộc tính có ảnh hưởng đến đánh giá phim và triển khai Recommender Systems để dự đoán phim phù hợp cho người dùng. Thông qua việc phân tích dữ liệu, nhóm nhận thấy rằng các thuộc tính riêng biệt có độ tương quan yếu với thuộc tính độ đánh giá phim. Thử nghiệm các phương pháp gợi ý phim, nhóm nhận thấy mỗi phương pháp đều có ưu nhược điểm riêng. Phương pháp Hybrid Approach đã được triển khai với 3 cách tiếp cận khác nhau (lồng user-based với content-based, sử dụng song song cả 2 phương pháp, kết hợp vector dự đoán trên user-based và phim user chọn trên content-based) nhằm kết hợp điểm mạnh của cả hai phương pháp trên. Hybrid Approach ở cách thứ 3 dù chưa tối ưu về hiệu suất, cách tiếp cận này mang lại kết quả tốt hơn, gợi ý chính xác và chủ động hơn, là lựa chọn khả thi để triển khai vào hệ thống gợi ý phim thực tế.