# MINI PROJECT: PHÂN CỤM KHÁCH HÀNG DỰA TRÊN LUẬT KẾT HỢP

ThS. Lê Thị Thùy Trang

2025-12-16

## 1 Phân cụm khách hàng dựa trên luật kết hợp

### 1.1 Giới thiệu dự án

Sử dụng thuật toán K-Means để phân cụm khách hàng thành các nhóm có hành vi mua sắm tương đồng. Điểm khác biệt của Mini Project này là thay vì dùng các đặc trưng RFM hay truyền thống, chúng ta sẽ dựa trên các luật kết hợp. Điều này cho phép phân nhóm khách hàng theo các mẫu sản phẩm thường được mua cùng nhau, từ đó đưa ra những chiến lược marketing phù hợp cho từng nhóm.

Sau khi áp dụng Apriori (Lab 1) và FP-Growth (Lab 2) để khai phá các tập mục phổ biến và luật kết hợp trong hành vi mua sắm, bước kế tiếp là phân cụm để hiểu sâu hơn về nhóm khách hàng, nhóm giỏ hàng, hoặc thậm chí các nhóm sản phẩm thường đi cùng nhau.

Dữ liệu giao dịch không có nhãn "khách hàng tốt/xấu", vì vậy bài toán phù hợp với Unsupervised Learning. K-Means được chọn vì:

- Hoạt động tốt với dữ liệu đa chiều, dễ triển khai và dễ diễn giải theo tâm cụm (centroid).
- Có thể mở rộng cho dữ liệu lớn và dễ kết hợp với các bước chọn K (Elbow/Silhouette).
- Lưu ý quan trọng: K-Means giả định cụm "gần hình cầu" theo khoảng cách Euclidean; vì vậy ta cần chuẩn hóa/scale, và cần kiểm tra tính hợp lý của cụm bằng trực quan + phân tích hồ sơ cụm

Trong Mini Project này, sinh viên sẽ:

1. Trích xuất đặc trưng từ các luật kết hợp (rules).
2. Biến các đặc trưng đó thành dữ liệu đầu vào cho thuật toán phân cụm.
3. Áp dụng các thuật toán phân cụm như KMeans, Agglomerative, hoặc DBSCAN.
4. Diễn giải các cụm tìm được để đưa ra các chiến lược kinh doanh.

### 1.2 Mục tiêu

Sau khi thực hiện xong dự án, sinh viên có thể:

1. Hiểu quy trình kết hợp giữa khai phá luật và phân cụm.
2. Thực hành trích đặc trưng từ dữ liệu luật kết hợp.
3. Áp dụng các thuật toán phân cụm để tìm nhóm hành vi tương đồng.
4. Trực quan hóa và diễn giải các cụm.
5. Đề xuất chiến lược hành động từ từng cụm (segment).

Pipeline thực hiện như sau:

1. Tiền xử lý và khai phá luật (tái sử dụng Lab1/ Lab 2): Tập trung vào các luật kết hợp có giá trị lift cao và support đủ mạnh.
2. Trích xuất đặc trưng từ luật: Xây dựng vector đặc trưng cho từng giỏ hàng, trong đó mỗi chiều đại diện cho một luật mạnh (hoặc một tập sản phẩm phổ biến). Gán nhãn 1 nếu giỏ hàng thỏa mãn luật đó, 0 nếu không.
3. Phân cụm: Sử dụng KMeans (hoặc thuật toán khác) để phân nhóm các giỏ hàng hoặc khách hàng dựa trên đặc trưng. Chọn số cụm bằng phương pháp elbow hoặc silhouette.
4. Diễn giải và trực quan hóa: Phân tích mỗi cụm: nhóm khách hàng nào? đặc điểm giỏ hàng là gì? luật nào chiếm ưu thế? Sau đó, đưa ra chiến lược phù hợp cho từng nhóm.

## 1.3 Danh sách các module trong project

Dự án được tổ chức với các lớp (module) tương tự Lab 1 để đảm bảo tính rõ ràng và tái sử dụng.

Chúng ta tái sử dụng các lớp cũ cho những bước chung và bổ sung lớp mới cho phân cụm. Cụ thể: `DataCleaner` (làm sạch dữ liệu) và `BasketPreparer` (chuẩn bị dữ liệu basket) giữ nguyên như trước.

Thay đổi chính ở giai đoạn tiếp theo nằm trong module phân cụm khách hàng từ luật kết hợp: thay vì chỉ dừng ở việc sinh và trực quan hóa các luật, ta triển khai lớp mới `RuleBasedCustomerClusterer` để biến kết quả luật thành đặc trưng (features) cho từng khách hàng. Cụ thể:

- Lớp này xây dựng ma trận `Customer×Item` dạng boolean (`build_customer_item_matrix`),
- sau đó nạp Top-K luật từ file rules (`load_rules`) và tạo ma trận `Customer×Item` trong đó mỗi cột là một luật, giá trị thể hiện khách hàng có "thỏa" tiền đề của luật hay không (có thể gán trọng số theo lift/confidence/support).
- Tiếp theo, pipeline có thể ghép thêm RFM (Recency–Frequency–Monetary) và chuẩn hóa để tạo vector đặc trưng cuối cùng (`build_final_features`). Trên vector này, notebook sẽ chọn số cụm K bằng silhouette (`choose_k_by_silhouette`), fit K-Means để gán nhãn cụm (`fit_kmeans`), và giảm chiều 2D (PCA/SVD) để trực quan hóa phân bố cụm (`project_2d`). Cuối cùng, kết quả được xuất ra bảng `meta_out` để profiling từng cụm (quy mô cụm, trung bình RFM, hành vi theo luật) và từ đó đề xuất chiến lược marketing phù hợp.

| **Thành phần** | **Ý nghĩa và chức năng** |
|---|---|
| `src/cluster_library.py` | Thư viện trung tâm của dự án, ngoài các class như lab 1, lab 2, chứa lớp (class): `RuleBasedCustomerClusterer` (biến luật kết hợp thành feature vector cho từng khách hàng, sau đó chạy KMeans để phân cụm.). |
| `notebooks/clustering_from_rules.ipynb` | Notebook này lấy kết quả luật kết hợp (rules) để biến thành vector đặc trưng cho từng khách hàng, rồi chạy K-Means để phân cụm khách hàng, cuối cùng vẽ/quan sát cụm và xuất bảng kết quả. |

Table 1: Bảng thành phần Project

## 1.4 Chuẩn bị môi trường thực hành

Kích hoạt môi trường ảo

```
conda activate shopping_env
```

Toàn bộ các thư viện cần thiết cho dự án được liệt kê trong file `requirements.txt`. Vì vậy, sau khi kích hoạt môi trường, cài đặt các thư viện cần thiết bằng câu lệnh:

```
pip install -r requirements.txt
```

# 2 Hướng dẫn thực hiện hoạt động FIT-DNU CONQUER

## 2.1 Mục tiêu của hoạt động

Hoạt động FIT-DNU CONQUER được thiết kế nhằm giúp sinh viên:

- Hiểu sâu bản chất của thuật toán thông qua việc ứng dụng trên dữ liệu thực.
- Tập diễn giải kết quả bằng ngôn ngữ đơn giản, tránh chỉ mô tả code hoặc trích xuất số liệu.
- Rèn luyện khả năng trình bày, giải thích và thuyết phục thông qua hoạt động chia sẻ kết quả.
- Phát triển tư duy phân tích theo góc nhìn kinh doanh và đưa ra đề xuất hành động.

Sinh viên cần xem xét dự án như một nhiệm vụ Data Scientist thực thụ: không chỉ "chạy đúng thuật toán", mà phải chuyển dữ liệu thành tri thức hữu ích.

## 2.2 Yêu cầu

Trong Mini Project này, các nhóm sẽ sử dụng repo cơ sở shop_cluster để xây dựng một pipeline phân khúc khách hàng theo hướng "luật kết hợp → đặc trưng hành vi mua kèm → phân cụm → diễn giải → đề xuất chiến lược marketing".

1. Trước hết, mỗi nhóm cần chạy pipeline để tạo ra hoặc sử dụng lại danh sách luật kết hợp (rules) từ Apriori hoặc FP-Growth. Nhóm phải trình bày rõ ràng cách mình chọn luật: lấy Top-K bao nhiêu luật, ưu tiên sắp xếp theo `lift` hay `confidence`, có áp dụng ngưỡng lọc tối thiểu `min_support`, `min_confidence`, `min_lift` hay không và vì sao. Kết quả lựa chọn luật cần được minh chứng bằng việc trích ra một bảng nhỏ khoảng 10 luật tiêu biểu kèm theo các chỉ số (`support`, `confidence`, `lift`) để người đọc thấy được chất lượng luật mà nhóm dùng làm đầu vào cho bước phân cụm.
2. Tiếp theo, nhóm cần thực hiện bước feature engineering cho phân cụm. Ở đây yêu cầu bắt buộc là nhóm phải xây dựng ít nhất hai biến thể đặc trưng để so sánh. Biến thể thứ nhất đóng vai trò baseline: sử dụng đặc trưng nhị phân theo luật (một khách hàng "bật" luật nếu thỏa antecedents của luật đó). Biến thể thứ hai là biến thể nâng cao, nhóm có thể chọn một trong hai hướng:
    - Đưa trọng số vào đặc trưng luật (ví dụ dùng `lift` hoặc `lift×confidence` để phản ánh độ mạnh của luật);
    - Ghép thêm RFM (Recency–Frequency–Monetary) để bổ trợ thông tin giá trị khách hàng.
    - Với biến thể nâng cao, nhóm phải mô tả rõ các thiết lập quan trọng như cách weighting, có bật RFM hay không, có scale RFM hay không, và có scale phần rule-feature hay không. Nhóm được khuyến khích thử thêm tiêu chí lọc luật theo độ dài antecedent (ví dụ loại các luật antecedent quá ngắn) để quan sát sự thay đổi chất lượng cụm.
3. Sau khi có vector đặc trưng, nhóm cần thực hiện chọn số cụm K và huấn luyện mô hình. Yêu cầu tối thiểu là nhóm phải sử dụng Silhouette score hoặc Eblow để khảo sát `K` trong một khoảng giá trị hợp lý (ví dụ 2 đến 10 hoặc 2 đến 12), sau đó chọn ra `K` tốt nhất theo kết quả và giải thích ngắn gọn lý do lựa chọn. Phần giải thích không cần dài, nhưng phải thể hiện tư duy: không chọn K chỉ vì "đẹp", mà còn cân nhắc xem cụm có thực sự tạo ra ý nghĩa hành động marketing hay không. Sau khi chọn `K`, nhóm huấn luyện K-Means và lưu lại nhãn cụm cho từng khách hàng.
4. Kết quả phân cụm cần được trực quan hóa và đánh giá ở mức tối thiểu. Mỗi nhóm phải thực hiện giảm chiều về 2D bằng PCA hoặc SVD và vẽ scatter plot, tô màu theo cluster để người đọc thấy mức độ tách cụm (tách rõ hay chồng lấn). Nhóm cần nhận xét ngắn về biểu đồ, tránh nhận xét chung chung mà cần bám vào hình ảnh.

5. Thực hiện so sánh có hệ thống giữa các biến thể đặc trưng: rule-only vs rule+RFM, binary vs weighted rules, Top-K nhỏ vs Top-K lớn. Nhóm cần bảng tổng hợp để cho thấy cấu hình nào tốt hơn và vì sao.
6. Quan trọng nhất là phần profiling và diễn giải cụm, đây là nội dung để sinh viên thể hiện năng lực phân tích. Mỗi nhóm phải tạo một bảng thống kê theo cụm, trong đó ít nhất có số lượng khách hàng của cụm. Nếu nhóm có dùng RFM thì bắt buộc báo cáo thêm trung bình hoặc trung vị Recency–Frequency–Monetary theo cụm. Đồng thời, nhóm phải rút ra "dấu hiệu đặc trưng" của cụm dựa trên luật: ví dụ liệt kê Top 10 luật hoặc Top rule-features được kích hoạt nhiều nhất trong cụm. Từ các thông tin này, nhóm phải đặt tên cho từng cụm (một tên tiếng Anh và một tên tiếng Việt dễ nhớ), mô tả persona của cụm trong 1 câu, và đưa ra một chiến lược marketing cụ thể dành cho cụm đó (bundle/cross-sell/upsell, ưu đãi theo nhóm sản phẩm, chăm sóc VIP, chiến dịch kích hoạt khách ngủ đông, v.v.). Chiến lược phải liên hệ trực tiếp đến đặc trưng cụm, không viết chung chung.
7. Xây dựng dashboard Streamlit để đọc các file output và cho phép lọc theo cụm, xem top rules, xem gợi ý bundle/cross-sell theo cụm.

## 2.3 Khuyến khích nâng cấp dự án

Ngoài phần bắt buộc, các nhóm được khuyến khích chọn thêm các hướng mở rộng để nâng chất lượng bài làm (Nhóm nào tham vọng tổng kết 10 thì phải làm thôi):

1. So sánh mô hình phân cụm: ngoài K-Means, thử thêm một thuật toán khác như Agglomerative Clustering hoặc DBSCAN/HDBSCAN, sau đó so sánh kết quả bằng metric (silhouette/DBI/CH) và bằng mức độ "actionable" của cụm.
2. Thử một trong các hướng: phân cụm giỏ hàng (basket clustering), phân cụm sản phẩm, hoặc phân cụm chính các luật (rule clustering), rồi so sánh "góc nhìn nào hữu ích hơn" cho marketing.

## 2.4 Kết quả kỳ vọng

Mỗi nhóm cần hoàn thành:

- Blog/Report (link Notion/GitHub).
- Slide trình bày.

## 2.5 Gợi ý cho phần trình bày tại lớp

1. Giới thiệu bài toán và mục tiêu nhóm.
2. Trình bày kết quả trọng tâm (không kể lể, không giải thích code).
3. Diễn giải các biểu đồ.
4. Kết luận và đề xuất hành động.