

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH



BÀI TẬP LỚN MÔN KHAI PHÁ DỮ LIỆU
(Mã môn học: CO3029)

HỌC KỲ: HK251 _ NĂM HỌC: 2025

**PHÂN TÍCH GIỎ HÀNG ĐỂ ĐƯA RA KHUYẾN
NGHỊ SẢN PHẨM**

Giảng viên hướng dẫn: Thầy ĐỖ THANH THÁI

STT	Họ và tên	MSSV
1	Bùi Ngọc Diễm Quỳnh	2212878
2	Đặng Công Sơn	2212932
3	Nguyễn Gia Nguyên	2212303

GitHub: [CO3029_DATA_MINING_L01](#)

Thuyết trình: [Youtube](#)

TP. Hồ Chí Minh, Tháng 11/2025



Mục lục

1 Giới thiệu	5
2 Cơ sở lý thuyết	5
2.1 Khai phá dữ liệu (Data Mining)	5
2.2 Phân tích giỏ hàng (Market Basket Analysis)	5
2.3 Khai phá luật kết hợp (Association Rule Mining)	6
2.3.1 Khái niệm	6
2.3.2 Các chỉ số đánh giá luật	6
2.4 Thuật toán Apriori	6
2.4.1 Tổng quan	6
2.4.2 Quy trình thuật toán	7
2.4.3 Lựa chọn ngưỡng hỗ trợ	7
2.5 Ma trận giao dịch (Transaction Matrix)	7
2.6 Trực quan hóa dữ liệu	8
2.7 Hệ thống gợi ý sản phẩm (Recommendation System)	8
3 Quy trình xử lý và phân tích dữ liệu	8
3.1 Thu thập dữ liệu	8
3.1.1 Mô tả tổng quan	8
3.1.2 Các thuộc tính chính	9
3.1.3 Lý do lựa chọn	9
3.2 Khám phá dữ liệu ban đầu	9
3.2.1 Thống kê mô tả	9
3.2.2 Các vấn đề chính	9
3.3 Làm sạch dữ liệu	9
3.4 Biến đổi dữ liệu	10
3.4.1 Tạo bảng giao dịch	10
3.4.2 Mã hóa nhị phân	10
3.4.3 Kích thước ma trận cuối cùng	10
3.5 Phân tích khám phá dữ liệu (EDA)	10
3.5.1 Top sản phẩm bán chạy	10
3.5.2 Phân bố số lượng mặt hàng mỗi hóa đơn	10
3.5.3 Heatmap đồng xuất hiện	11
3.5.4 Quan hệ giữa Support – Confidence – Lift	11
4 Chuẩn bị dữ liệu cho mô hình dự đoán	11
4.1 Tải dữ liệu	11
4.2 Làm sạch dữ liệu	11
4.2.1 Loại bỏ giá trị thiếu	12
4.2.2 Loại bỏ hóa đơn bị hủy	12
4.2.3 Giữ lại giao dịch tại Vương quốc Anh	12
4.2.4 Chuẩn hóa văn bản mô tả sản phẩm	12
4.3 Tạo bảng giao dịch (Transaction Table)	12
4.4 Phân tích thống kê sơ bộ	13
5 Xây dựng mô hình dữ đoán	13
5.1 Sinh luật kết hợp	14
5.2 Trực quan hóa và Phân tích kết quả	16
5.2.1 Phân bố kích thước giỏ hàng	16
5.2.2 Tương quan Support, Confidence và Lift	16
5.2.3 Heatmap đồng xuất hiện (Top 20 sản phẩm)	17
5.2.4 Đồ thị mạng lưới	19
5.3 Phân tích kết quả mô hình	21
5.4 Ứng dụng mô hình khuyến nghị	21



6 Phân tích xu hướng mua sắm chủ đạo	22
6.1 Xu hướng "Sưu Tầm Trọn Bộ" (Product Set Pattern)	22
6.2 Sức Hút Của Mua Sắm Theo Chủ Đề và Dịp Lễ	22
6.3 Tác Động Của Nguyên Tắc Pareto (80/20)	22
6.4 Sức Mạnh Kết Hợp: Sản Phẩm Bổ Sung và "Combo Ngách"	23
7 Kết luận và phương hướng phát triển	23
7.1 Kết luận	23
7.2 Phương hướng phát triển	23



Bảng phân công nhiệm vụ

STT	MSSV	Họ và tên	Đóng góp	Nhiệm vụ
1	2212932	Đặng Công Sơn	100%	Tìm kiếm và tiền xử lý dữ liệu, phân tích dữ liệu, làm slide, chuẩn bị dữ liệu cho mô hình dự đoán.
2	2212303	Nguyễn Gia Nguyên	100%	Xây dựng mô hình dự đoán, khai phá dữ liệu, viết báo cáo.
3	2212878	Bùi Ngọc Diễm Quỳnh	100%	Khai phá thông tin từ dữ liệu, phân tích xu hướng mua sắm và đưa ra phương hướng phát triển, viết báo cáo.



Lời nói đầu

Trong bối cảnh thương mại điện tử phát triển mạnh mẽ, việc thấu hiểu hành vi mua sắm của khách hàng trở thành yếu tố then chốt giúp doanh nghiệp nâng cao trải nghiệm người dùng và tối ưu doanh thu. Một trong những phương pháp phổ biến và hiệu quả nhất hiện nay là phân tích giỏ hàng nhằm khám phá mối liên hệ giữa các sản phẩm mà khách hàng thường mua cùng nhau. Thông qua đó, hệ thống có thể đưa ra những gợi ý sản phẩm phù hợp, hỗ trợ khách hàng lựa chọn dễ dàng hơn, đồng thời giúp doanh nghiệp xây dựng chiến lược bán hàng thông minh.

Đề tài “Phân tích giỏ hàng để đưa ra khuyến nghị sản phẩm” được thực hiện với mục tiêu tìm hiểu, áp dụng các thuật toán khai phá dữ liệu, đặc biệt là kỹ thuật phân tích luật kết hợp (Association Rules), để xây dựng mô hình gợi ý sản phẩm dựa trên dữ liệu thực tế. Bằng cách phân tích thói quen mua hàng, đề tài hướng đến giải pháp gợi ý tự động, góp phần tăng mức độ hài lòng của khách hàng và hiệu quả kinh doanh.

Trong quá trình thực hiện, nhóm nhận được sự hướng dẫn của thầy Đỗ Thanh Thái, người đã hỗ trợ định hướng và cung cấp những kiến thức quý báu giúp nhóm hoàn thành đề tài. Nhóm xin chân thành cảm ơn thầy vì sự tận tâm và đồng hành trong suốt thời gian qua. Bài tập này đã giúp các thành viên trong nhóm học hỏi được rất nhiều kiến thức bổ ích từ tiền xử lý dữ liệu, phân tích, khai phá thông tin tiềm ẩn cho đến trực quan hóa kết quả

Mặc dù đã cố gắng hoàn thiện nội dung một cách đầy đủ tuy nhiên đề tài khó tránh khỏi những hạn chế nhất định. Nhóm rất mong nhận được sự góp ý của thầy để đề tài được hoàn thiện hơn.

Một lần nữa, nhóm xin chân thành cảm ơn thầy Đỗ Thanh Thái đã giao đề tài này, giúp nhóm không chỉ rèn luyện kỹ năng phân tích dữ liệu mà còn có cơ hội gắn kết và làm việc nghiêm túc!



1 Giới thiệu

Trong bối cảnh thương mại điện tử và ngành bán lẻ hiện đại phát triển mạnh mẽ, việc thấu hiểu hành vi mua sắm của khách hàng đóng vai trò vô cùng quan trọng trong chiến lược kinh doanh của các doanh nghiệp. Không chỉ cần biết sản phẩm nào được ưa chuộng, doanh nghiệp còn phải nắm bắt được mối quan hệ giữa các sản phẩm — chẳng hạn như những mặt hàng thường được mua cùng nhau — để từ đó đưa ra chiến lược bán chéo (cross-selling) hoặc khuyến nghị sản phẩm phù hợp.

Một trong những hướng tiếp cận phổ biến và hiệu quả trong lĩnh vực này là phân tích giỏ hàng (Market Basket Analysis). Phương pháp này giúp phát hiện các mẫu kết hợp (association patterns) ẩn trong dữ liệu giao dịch, từ đó hỗ trợ doanh nghiệp hiểu rõ xu hướng tiêu dùng và tối ưu hóa danh mục sản phẩm.

Trong nghiên cứu này, chúng tôi sử dụng thuật toán Apriori, một trong những thuật toán kinh điển trong khai phá luật kết hợp (association rule mining), để phân tích dữ liệu giỏ hàng của khách hàng. Bằng cách phát hiện các tập sản phẩm thường xuyên xuất hiện cùng nhau (frequent itemsets) và các luật kết hợp có độ tin cậy cao, hệ thống có thể đưa ra các khuyến nghị sản phẩm tự động cho người dùng hoặc đề xuất các nhóm sản phẩm (combo) phù hợp cho doanh nghiệp.

Nghiên cứu hướng đến các mục tiêu:

- Xác định những sản phẩm có mức độ mua chung cao.
- Khai thác các luật kết hợp nhằm đưa ra gợi ý sản phẩm phù hợp cho khách hàng.
- Minh họa tính ứng dụng thực tế của thuật toán Apriori trong phân tích dữ liệu bán hàng.

Qua kết quả phân tích, nhóm mong muốn chứng minh tính hiệu quả của kỹ thuật khai phá luật kết hợp trong việc nâng cao trải nghiệm người dùng, tăng doanh thu và tối ưu hóa chiến lược marketing trong các hệ thống thương mại điện tử.

2 Cơ sở lý thuyết

2.1 Khai phá dữ liệu (Data Mining)

Khai phá dữ liệu là quá trình khám phá và trích xuất những thông tin có giá trị, những quy luật ẩn giấu trong khối dữ liệu lớn. Đây là một phần quan trọng của Khoa học dữ liệu và các Hệ thống hỗ trợ ra quyết định.

Quá trình này trải qua nhiều giai đoạn: từ thu thập dữ liệu, làm sạch, tiền xử lý, chuyển đổi, cho đến xây dựng mô hình và đánh giá kết quả. Mục đích cuối cùng là tìm ra những tri thức tiềm ẩn mà con người khó có thể phát hiện chỉ bằng quan sát thông thường.

Trong bài toán phân tích giỏ hàng, khai phá dữ liệu được sử dụng để phát hiện những mối liên hệ về việc các sản phẩm thường xuất hiện cùng nhau.

2.2 Phân tích giỏ hàng (Market Basket Analysis)

Phân tích giỏ hàng là kỹ thuật tìm hiểu các mối quan hệ giữa những sản phẩm thường được mua chung trong cùng một lần giao dịch. Ví dụ điển hình: khách hàng mua trà thường có xu hướng mua thêm bánh quy.

Kỹ thuật này có nhiều ứng dụng thực tế như:

- Xây dựng hệ thống gợi ý sản phẩm.
- Chiến lược bán chéo (cross-selling).
- Tối ưu hóa cách sắp xếp sản phẩm trong cửa hàng.
- Thiết kế các combo sản phẩm hấp dẫn.
- Phân tích và hiểu rõ hành vi mua sắm của khách hàng.

Nền tảng của phân tích giỏ hàng chính là khai phá luật kết hợp (Association Rule Mining).



2.3 Khai phá luật kết hợp (Association Rule Mining)

2.3.1 Khái niệm

Luật kết hợp là dạng tri thức mô tả mối quan hệ:

$$A \Rightarrow B$$

Trong đó:

- **A:** là sản phẩm có trong giỏ hàng.
- **B:** là tập sản phẩm có khả năng sẽ được mua tiếp theo.

Luật này được hiểu là: "Nếu khách hàng mua A thì có khả năng sẽ mua thêm B".

2.3.2 Các chỉ số đánh giá luật

1. Độ hỗ trợ (Support):

Support đo lường mức độ thường xuyên mà một tập sản phẩm xuất hiện trong tổng số giao dịch.

$$\text{Support}(A) = \frac{\text{Số giao dịch chứa } A}{\text{Tổng số giao dịch}}$$

Chỉ số này giúp:

- Loại bỏ những sản phẩm xuất hiện quá ít
- Tăng tốc độ tính toán

2. Độ tin cậy (Confidence):

Confidence đo xác suất khách hàng sẽ mua B khi đã mua A:

$$\text{Confidence}(A \Rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$

Confidence càng cao nghĩa là luật càng đáng tin cậy.

3. Độ nâng (Lift):

Lift đo mức độ A và B xuất hiện cùng nhau có cao hơn (hoặc thấp hơn) so với sự ngẫu nhiên không.

$$\text{Lift}(A \Rightarrow B) = \frac{\text{Confidence}(A \Rightarrow B)}{\text{Support}(B)}$$

Với:

- Lift > 1: A và B liên quan tích cực
- Lift = 1: không có mối quan hệ
- Lift < 1: quan hệ âm (hiếm dùng để gợi ý sản phẩm)

Trong phân tích giỏ hàng, ba chỉ số support, confidence và lift được kết hợp để lọc ra những luật có giá trị.

2.4 Thuật toán Apriori

2.4.1 Tổng quan

Apriori là thuật toán kinh điển và được sử dụng rộng rãi nhất để tìm kiếm:

- Các tập mục thường xuyên (Frequent Itemsets)
- Các luật kết hợp (Association Rules)

Ý tưởng cốt lõi của Apriori (còn gọi là tính chất Apriori):

- Nếu một tập sản phẩm là phổ biến, thì mọi tập con của nó cũng đều phải phổ biến.

2.4.2 Quy trình thuật toán

Thuật toán hoạt động theo các bước:

1. Bắt đầu bằng việc tìm tất cả các tập 1 sản phẩm đạt ngưỡng hỗ trợ tối thiểu
2. Từ đó sinh ra các tập 2 sản phẩm, rồi 3 sản phẩm...
3. Loại bỏ những tập không đạt ngưỡng support
4. Lặp lại cho đến khi không còn tập mới nào được tạo ra

Apriori sử dụng chiến lược "phát sinh – kiểm tra" (generate-and-test), nên chi phí tính toán tăng nhanh khi dữ liệu có quy mô lớn.

2.4.3 Lựa chọn ngưỡng hỗ trợ

Việc chọn ngưỡng min_support có ảnh hưởng lớn đến:

- Số lượng tập mục được sinh ra
- Số luật kết hợp tìm thấy
- Thời gian xử lý

Ví dụ thực nghiệm trong đề tài:

Min support	Số itemsets	Số luật	Thời gian
2.0%	236	30	14.74s
1.5%	441	67	21.76s
1.0%	971	267	30.72s
0.5%	4074	3730	69.17s

Bảng 1: Kết quả thử nghiệm các mức min_support khác nhau

Nhận xét: Nguồn support càng thấp thì kết quả càng phong phú, nhưng đổi lại là thời gian xử lý tăng đáng kể.

2.5 Ma trận giao dịch (Transaction Matrix)

Để thuật toán Apriori có thể hoạt động, dữ liệu cần được biểu diễn dưới dạng ma trận:

- **Hàng (dòng):** Mô hình hóa đơn (InvoiceNo)
- **Cột:** Mỗi sản phẩm
- **Giá trị:**
 - 1 → sản phẩm có trong hóa đơn
 - 0 → sản phẩm không có

Ví dụ minh họa:

InvoiceNo	item A	item B	item C
536365	1	0	1
536366	0	1	1

Ma trận này mang lại các lợi ích:

- Giúp thuật toán Apriori xử lý hiệu quả
- Tính support nhanh chóng
- Dễ dàng tính toán sự đồng xuất hiện giữa các sản phẩm

Trong đề tài này, ma trận giao dịch được xây dựng từ dữ liệu thực tế với quy mô:

- Hơn 16.649 hóa đơn
- 3.833 sản phẩm



2.6 Trực quan hóa dữ liệu

Để hiểu sâu hơn về đặc điểm mua sắm, nhóm nghiên cứu áp dụng các phương pháp trực quan hóa:

- **Biểu đồ tần suất mua hàng:** Xác định những sản phẩm được ưa chuộng nhất
- **Phân bố kích thước giỏ hàng:** Thể hiện mỗi hóa đơn trung bình có bao nhiêu sản phẩm
- **Biểu đồ scatter: support – confidence – lift:** Đánh giá chất lượng của các luật
- **Heatmap đồng xuất hiện:** Xác định những nhóm sản phẩm thường được mua cùng nhau
- **Đồ thị mạng luật kết hợp (Network Graph):** Thể hiện một cách trực quan mối liên hệ giữa các sản phẩm

2.7 Hệ thống gợi ý sản phẩm (Recommendation System)

Gợi ý sản phẩm là ứng dụng chính và quan trọng nhất của phân tích giỏ hàng. Hệ thống hoạt động dựa trên các luật có dạng $A \Rightarrow B$. Khi khách hàng mua sản phẩm A (hoặc một danh sách sản phẩm), hệ thống sẽ tìm những luật có phần điều kiện (antecedent) trùng với giỏ hàng và đề xuất phần kết quả (consequent). Trong đề tài, quy trình gợi ý được thực hiện như sau:

- Sử dụng tập luật kết hợp đã được lọc từ thuật toán Apriori
- Áp dụng confidence làm điểm số để xếp hạng mức độ gợi ý
- Tổng hợp từ nhiều luật để chọn ra những đề xuất tốt nhất

3 Quy trình xử lý và phân tích dữ liệu

Quy trình xử lý và phân tích dữ liệu được thực hiện theo phương pháp luận CRISP-DM (Cross Industry Standard Process for Data Mining) – mô hình chuẩn phổ biến trong khai phá dữ liệu. Mô hình gồm sáu giai đoạn chính: (1) Hiểu bài toán kinh doanh, (2) Hiểu dữ liệu, (3) Chuẩn bị dữ liệu, (4) Mô hình hóa, (5) Đánh giá và (6) Triển khai. Trong nghiên cứu này, nhóm tập trung vào bốn giai đoạn đầu tiên, đặc biệt là các bước thu thập, khám phá, làm sạch, biến đổi và phân tích khám phá dữ liệu (EDA), nhằm xây dựng tập dữ liệu chất lượng cao trước khi áp dụng thuật toán Apriori.

3.1 Thu thập dữ liệu

Bộ dữ liệu được sử dụng là *Online Retail Dataset* do UCI Machine Learning Repository cung cấp. Đây là một trong những bộ dữ liệu kinh điển trong lĩnh vực phân tích giỏ hàng và khai phá luật kết hợp, được sử dụng rộng rãi trong nghiên cứu học thuật và giảng dạy tại nhiều trường đại học.

3.1.1 Mô tả tổng quan

- Số quan sát: 541.909 dòng
- Số thuộc tính: 8 cột
- Thời gian thu thập: từ 01/12/2010 đến 09/12/2011
- Nguồn: dữ liệu giao dịch thực tế của một cửa hàng bán lẻ trực tuyến tại châu Âu
- Định dạng: mỗi dòng tương ứng với một mặt hàng trong một hóa đơn

3.1.2 Các thuộc tính chính

Thuộc tính	Ý nghĩa
InvoiceNo	Mã hóa đơn (có thể trùng nếu hóa đơn chứa nhiều sản phẩm)
StockCode	Mã sản phẩm
Description	Tên sản phẩm
Quantity	Số lượng mua
InvoiceDate	Ngày giờ tạo hóa đơn
UnitPrice	Giá đơn vị
CustomerID	Mã khách hàng
Country	Quốc gia của khách hàng

Bảng 2: Các thuộc tính chính của bộ dữ liệu

3.1.3 Lý do lựa chọn

- Dữ liệu thực tế, phản ánh đúng hành vi mua sắm của khách hàng
- Quy mô lớn, đa dạng sản phẩm (gần 4.000 mặt hàng khác nhau)
- Phù hợp với mục tiêu nghiên cứu: khai thác luật kết hợp để xây dựng hệ thống gợi ý
- Được cộng đồng học thuật công nhận và sử dụng rộng rãi, đảm bảo tính khách quan

3.2 Khám phá dữ liệu ban đầu

Giai đoạn này nhằm hiểu rõ cấu trúc, phân bố và các vấn đề tiềm ẩn của dữ liệu.

3.2.1 Thông kê mô tả

Một số nhận xét quan trọng:

- Tỷ lệ giá trị thiếu cao ở cột CustomerID (khoảng 24,9%)
- Tồn tại các hóa đơn hủy (InvoiceNo bắt đầu bằng ký tự “C”)
- Tên sản phẩm chưa được chuẩn hóa (viết hoa/thường khác nhau, ký tự thừa)
- 90% giao dịch đến từ United Kingdom, các quốc gia khác chiếm tỷ lệ rất nhỏ

3.2.2 Các vấn đề chính

- Giá trị thiếu:** ảnh hưởng đến việc phân nhóm khách hàng
- Ngoại lệ:** giao dịch có số lượng âm hoặc cực lớn
- Nhiều văn bản:** cùng một sản phẩm nhưng có nhiều cách mô tả khác nhau
- Độ thưa của ma trận giao dịch:** gây khó khăn cho thuật toán Apriori

Những vấn đề này nếu không được xử lý sẽ làm giảm đáng kể độ chính xác của các chỉ số Support, Confidence và Lift.

3.3 Làm sạch dữ liệu

Để đảm bảo chất lượng đầu vào cho mô hình Apriori, nhóm đã thực hiện các bước làm sạch sau:

- Lọc bỏ quan sát thiếu thông tin quan trọng:** xóa toàn bộ các dòng không có Description hoặc CustomerID
- Lọc bỏ hóa đơn hủy:** loại bỏ các hóa đơn có InvoiceNo bắt đầu bằng “C” (tránh giá trị Quantity âm)

3. **Giữ lại chỉ số dữ liệu từ United Kingdom:** giảm nhiễu thông kê do các quốc gia khác chiếm tỷ lệ quá nhỏ (dưới 10%)

4. **Chuẩn hóa văn bản tên sản phẩm:**

- Chuyển toàn bộ về chữ thường (lowercase)
- Loại bỏ khoảng trắng thừa (strip)
- Xóa các ký tự đặc biệt không cần thiết

Kết quả: số lượng sản phẩm duy nhất giảm đáng kể, giúp ma trận giao dịch gọn hơn.

3.4 Biến đổi dữ liệu

Để áp dụng thuật toán Apriori, dữ liệu giao dịch cần được chuyển thành ma trận nhị phân (basket matrix).

3.4.1 Tạo bảng giao dịch

- Gom nhóm theo cặp (InvoiceNo, Description)
- Tính tổng số lượng sản phẩm trong mỗi hóa đơn
- Chuyển sang dạng bảng pivot

3.4.2 Mã hóa nhị phân

Áp dụng hàm:

```
encode_units(x) = 1 nếu x > 0 else 0
```

Kết quả: mỗi ô trong ma trận chỉ nhận giá trị 0 hoặc 1 (có/không xuất hiện trong giỏ hàng).

3.4.3 Kích thước ma trận cuối cùng

- Số hóa đơn duy nhất: 16.649
- Số sản phẩm duy nhất: ≈ 3.500
- Tổng số phần tử: ≈ 58 triệu ô

Do ma trận rất thưa, việc lựa chọn ngưỡng min_support hợp lý là yếu tố then chốt để tránh tràn bộ nhớ.

3.5 Phân tích khám phá dữ liệu (EDA)

3.5.1 Top sản phẩm bán chạy

Các mặt hàng phổ biến nhất thuộc nhóm đồ trang trí, túi xách phong cách retro, bộ âm trà Regency – phù hợp với định vị cửa hàng chuyên bán quà tặng và đồ lưu niệm.

3.5.2 Phân bố số lượng mặt hàng mỗi hóa đơn

Phân bố lệch phai mạnh:

- Trung bình: 20,68 sản phẩm/hóa đơn
- Trung vị: 15 sản phẩm/hóa đơn
- Một số hóa đơn có hàng trăm sản phẩm (mua sỉ)



3.5.3 Heatmap đồng xuất hiện

Quan sát thấy các nhóm sản phẩm thường được mua cùng nhau:

- Bộ Lunch Bag (pink, blue, red)
- Bộ đĩa/trà Regency
- Đồ trang trí treo dạng “hanging heart”
- Đồ chơi Poppy’s Playhouse

3.5.4 Quan hệ giữa Support – Confidence – Lift

Biểu đồ phân tán cho thấy:

- Các luật có Lift cao thường có Support thấp (hành vi hiếm nhưng rất mạnh)
- Các luật có Support cao lại có Lift thấp (hành vi phổ biến nhưng ít giá trị gợi ý)

Đây là đặc trưng điển hình của dữ liệu bán lẻ quy mô lớn: những mẫu hình hiếm mới là nguồn insight giá trị nhất cho chiến lược cross-selling và up-selling.

4 Chuẩn bị dữ liệu cho mô hình dự đoán

Trong bất kỳ quy trình khai phá dữ liệu nào, bước chuẩn bị dữ liệu (Data Preparation) luôn đóng vai trò đặc biệt quan trọng, quyết định trực tiếp đến chất lượng mô hình và độ tin cậy của kết quả phân tích. Trong đề tài này, nhóm thực hiện chuẩn bị dữ liệu từ bộ dữ liệu Online Retail theo các bước: tải dữ liệu, làm sạch dữ liệu, xây dựng bảng giao dịch và phân tích thống kê sơ bộ. Các bước này được mô tả chi tiết như sau.

4.1 Tải dữ liệu

Bộ dữ liệu được sử dụng là *Online Retail Dataset*, chứa thông tin giao dịch của một công ty bán lẻ trực tuyến đặt tại Vương quốc Anh. Bộ dữ liệu bao gồm hơn 540.000 giao dịch diễn ra trong giai đoạn từ tháng 12/2010 đến tháng 12/2011 và được công bố trên nền tảng Kaggle.

Mỗi dòng trong tập dữ liệu tương ứng với một sản phẩm thuộc một hóa đơn, bao gồm các thuộc tính:

- **InvoiceNo:** Mã hóa đơn.
- **StockCode:** Mã sản phẩm.
- **Description:** Tên sản phẩm.
- **Quantity:** Số lượng sản phẩm mua trong hóa đơn.
- **InvoiceDate:** Thời gian lập hóa đơn.
- **UnitPrice:** Đơn giá sản phẩm.
- **CustomerID:** Mã khách hàng.
- **Country:** Quốc gia của khách hàng.

Tập dữ liệu được đọc bằng ngôn ngữ Python thông qua thư viện *pandas*. Sau khi tải lên, nhóm kiểm tra kích thước ban đầu và nhận thấy dữ liệu gồm 8 cột thông tin và 541.910 dòng.

4.2 Làm sạch dữ liệu

Trước khi áp dụng bất kỳ thuật toán khai phá nào, việc làm sạch dữ liệu là yêu cầu thiết yếu nhằm loại bỏ dữ liệu thiếu, dữ liệu nhiễu hoặc không hợp lệ. Nhóm tiến hành các bước xử lý sau:



4.2.1 Loại bỏ giá trị thiếu

Một số dòng dữ liệu không có thông tin ở cột *Description* hoặc *CustomerID*. Những dòng này không hữu ích cho phân tích giỏ hàng nên được loại bỏ:

```
df.dropna(subset=['Description', 'CustomerID'], inplace=True)
```

Kết quả: số dòng giảm từ 541.909 xuống còn khoảng 406.829 dòng hợp lệ.

4.2.2 Loại bỏ hóa đơn bị hủy

Các hóa đơn bị hủy được nhận diện bằng mã *InvoiceNo* bắt đầu bằng ký tự "C" (Credit Note). Đây là các giao dịch không phản ánh hành vi thực tế của khách hàng nên nhóm loại bỏ hoàn toàn:

```
df = df[~df['InvoiceNo'].astype(str).str.startswith('C')]
```

4.2.3 Giữ lại giao dịch tại Vương quốc Anh

Bộ dữ liệu chứa giao dịch từ 37 quốc gia khác nhau. Tuy nhiên, hơn 90% giao dịch đến từ United Kingdom. Để đảm bảo tính ổn định của mô hình và tránh nhiễu, nhóm chỉ giữ lại dữ liệu từ quốc gia này:

```
df = df[df['Country'] == 'United Kingdom']
```

4.2.4 Chuẩn hóa văn bản mô tả sản phẩm

Để giảm trùng lặp do khác biệt trong cách viết, nhóm chuẩn hóa cột *Description* bằng cách:

- Chuyển toàn bộ về chữ thường.
- Loại bỏ khoảng trắng dư thừa.

```
df['Description'] = df['Description'].str.strip().str.lower()
```

Việc chuẩn hóa giúp gom nhóm các sản phẩm về dạng thống nhất, phục vụ trực tiếp cho bước tạo bảng giao dịch.

4.3 Tạo bảng giao dịch (Transaction Table)

Sau khi làm sạch dữ liệu, nhóm tiến hành chuyển đổi dữ liệu từ dạng giao dịch từng dòng thành dạng bảng giao dịch (*transaction matrix*), cần thiết để áp dụng thuật toán Apriori. Quá trình gồm các bước:

- Gom nhóm dữ liệu theo cặp (*InvoiceNo*, *Description*) và tính tổng số lượng sản phẩm.
- Mỗi sản phẩm trở thành một cột trong bảng.
- Mỗi hóa đơn tương ứng với một dòng.
- Các ô chứa số lượng được mã hóa thành dạng nhị phân:
 - 1 nếu sản phẩm xuất hiện trong hóa đơn.
 - 0 nếu không xuất hiện.

```
basket = (df.groupby(['InvoiceNo', 'Description'])['Quantity']
           .sum().unstack().reset_index().fillna(0)
           .set_index('InvoiceNo'))

def encode_units(x):
    return 1 if x > 0 else 0

basket_sets = basket.applymap(encode_units)
```

Sau khi chuyển đổi, nhóm thu được một ma trận nhị phân có dạng sau:

InvoiceNo	milk	bread	apple	butter
10001	1	1	0	0
10002	1	0	1	0
10003	1	1	1	1

Kết quả: ma trận giỏ hàng thu được 4.338 giao dịch và hơn 1.200 sản phẩm khác nhau.

4.4 Phân tích thống kê sơ bộ

Truước khi tiến hành khai phá luật kết hợp, nhóm thực hiện phân tích thống kê để hiểu rõ hơn về đặc tính của tập dữ liệu:

- Tính tần suất xuất hiện của từng sản phẩm.
- Vẽ biểu đồ top 10 sản phẩm phổ biến nhất bằng *Matplotlib* và *Seaborn*.

Kết quả phân tích cho thấy một số sản phẩm xuất hiện với tần suất vượt trội, tiêu biểu như:

- white hanging heart t-light holder
- jumbo bag red retrospot
- assorted colour bird ornament

Điều này phản ánh quy luật Pareto trong bán lẻ: 20% sản phẩm mang lại phần lớn số lượng giao dịch. Việc hiểu rõ đặc điểm này hỗ trợ việc lựa chọn *min_support* hợp lý cho mô hình Apriori ở bước tiếp theo.

5 Xây dựng mô hình dữ đoán

Sau khi dữ liệu được làm sạch và chuyển đổi sang dạng nhị phân, bước quan trọng đầu tiên là xác định ngưỡng hỗ trợ tối thiểu (*min_support*). Ngưỡng này quyết định một sản phẩm (hoặc cụm sản phẩm) cần xuất hiện trong bao nhiêu phần trăm tổng số giao dịch để được coi là "phổ biến". Để tìm ra *min_support* phù hợp, nhóm đã thử nghiệm các mức *min_support* khác nhau trên tổng số 16649 giao dịch thu được kết quả sau:

Bảng 3: Kết quả thử nghiệm các mức Min Support

Min Support (%)	Số giao dịch tối thiểu	Số tập phổ biến (Itemsets)	Số luật (Rules)	Thời gian (giây)
2.0%	332	236	30	14.74s
1.5%	249	441	67	21.76s
1.0%	166	971	267	30.72s
0.5%	83	4074	3730	69.17s

Trong đó:

- Min Support (%) (Độ hỗ trợ tối thiểu):** Đây là tham số đầu vào quan trọng nhất mà bạn thiết lập. Nó là một tỷ lệ phần trăm, quy định mức độ "phổ biến" tối thiểu mà một nhóm sản phẩm (itemset) phải đạt được thì mới được coi là đáng quan tâm. Ví dụ, 1.0% có nghĩa là "Tôi chỉ quan tâm đến các nhóm sản phẩm mà xuất hiện trong ít nhất 1% tổng số các giao dịch".
- Số giao dịch tối thiểu:** Đây là con số tuyệt đối được tính từ Min Support (%). Nó bằng Min Support (%) nhân với tổng số giao dịch mà bạn đang phân tích. Đây chính là "ngưỡng hỗ trợ tuyệt đối".
- Số tập phổ biến (Itemsets):** Đây là kết quả của thuật toán. Nó đếm số lượng các nhóm sản phẩm (ví dụ: {Bánh mì, Sữa}, {Bia, Tã, Sữa}) mà có số lần xuất hiện lớn hơn hoặc bằng Số giao dịch tối thiểu (đã giải thích ở trên).
- Số luật (Rules):** Đây là số lượng các "quy tắc kết hợp" (ví dụ: Nếu mua {Bia, Tã} thì sẽ mua {Sữa}) được sinh ra từ các tập phổ biến đó. Thông thường, các luật này cũng đã được lọc qua một ngưỡng *min_confidence* (ở đây nhóm chọn *min_confidence* = 50%) nên mới có ý nghĩa.



- **Thời gian (giây):** Đây là thời gian (tính bằng giây) mà máy tính cần để hoàn thành toàn bộ quá trình: từ việc tìm tất cả các tập phổ biến cho đến việc sinh ra các luật, ứng với mức Min Support đã chọn.

Ta thấy:

- Nếu min_support quá cao (ví dụ: 2%, 1,5%), thuật toán sẽ chạy nhanh nhưng chỉ phát hiện được rất ít luật kết hợp (30 luật), bỏ lỡ nhiều mối quan hệ tiềm năng.
- Nếu min_support quá thấp (ví dụ: 0.5%), thuật toán sẽ tạo ra hàng ngàn luật (3730 luật), gây nhiều và tốn nhiều thời gian xử lý (69.17s).

Vì vậy, nhóm quyết định chọn min_support là 1%. Mức này cung cấp sự cân bằng tốt nhất, tạo ra 971 tập mục phổ biến và 267 luật kết hợp có ý nghĩa, với thời gian xử lý hợp lý (30.72s).

Sau khi dữ liệu được làm sạch và chuyển đổi sang dạng nhị phân (mỗi sản phẩm biểu diễn bằng giá trị 0 hoặc 1), nhóm tiến hành áp dụng thuật toán Apriori nhằm phát hiện các tập mục phổ biến (frequent itemsets) trong giỏ hàng.

Cụ thể, nhóm sử dụng thư viện mlxtend của Python, với parameter hỗ trợ tối thiểu (min_support) đặt là 0.01, nghĩa là chỉ xem xét những sản phẩm hoặc tập hợp sản phẩm xuất hiện trong ít nhất 1% tổng số giao dịch.

```
1 basket_sets = pd.read_csv("cleaned_basket.csv", index_col=0)
2 frequent_itemsets = apriori(basket_sets, min_support=0.01, use_colnames=True,
3 low_memory=True)
print(frequent_itemsets.sort_values(by='support', ascending=False).head(10))
```

Kết quả cho thấy các sản phẩm xuất hiện thường xuyên nhất bao gồm:

Bảng 4: Các sản phẩm xuất hiện thường xuyên nhất

Tên sản phẩm	Support
white hanging heart t-light holder	0.113
jumbo bag red retrospot	0.087
regency cakestand 3 tier	0.085
assorted colour bird ornament	0.078
party bunting	0.078
lunch bag red retrospot	0.067
set of 3 cake tins pantry design	0.060
lunch bag black skull	0.0598
paper chain kit 50's christmas	0.057
natural slate heart chalkboard	0.056

Nhận xét: Những sản phẩm có support cao thường là đồ trang trí, quà tặng hoặc sản phẩm gia dụng nhỏ, được khách hàng mua với tần suất lớn trong các dịp lễ và sự kiện.

5.1 Sinh luật kết hợp

Dựa trên các tập phổ biến vừa thu được, nhóm sử dụng hàm association_rules() để sinh ra các luật kết hợp (association rules). Mục tiêu là xác định mối quan hệ giữa các sản phẩm được mua cùng nhau.

```
1 rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1.0)
2 rules = rules[(rules['confidence'] >= 0.5) & (rules['lift'] >= 1.0)]
3 rules = rules.sort_values(by='lift', ascending=False)
4 print(rules[['antecedents', 'consequents', 'support', 'confidence', 'lift']].head(10))
```

Một số luật kết hợp mạnh được phát hiện như sau:



Bảng 5: Một số luật kết hợp mạnh

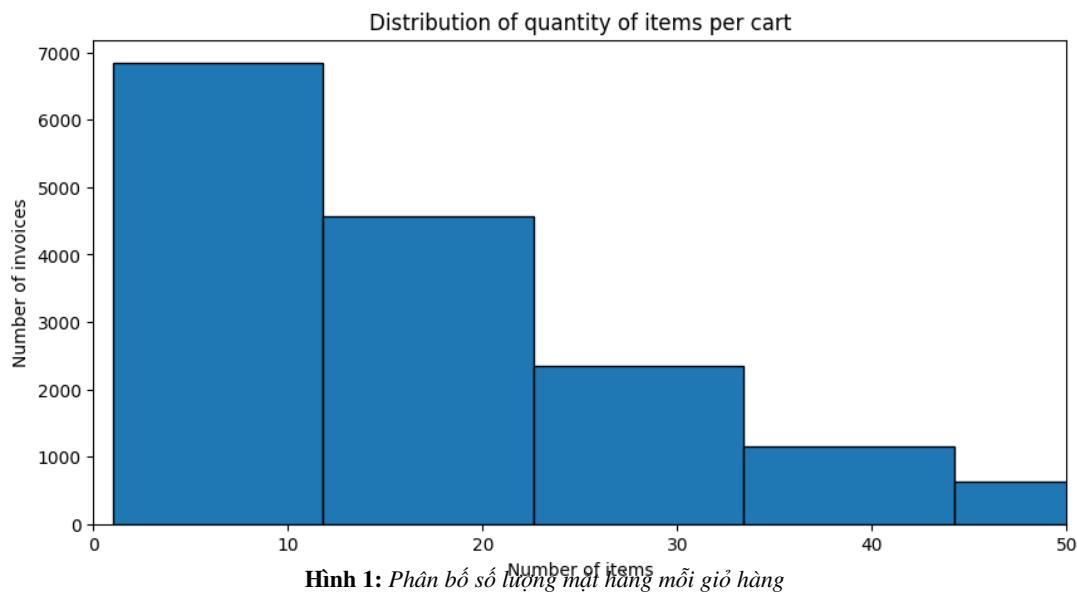
Luật kết hợp	Support	Confidence	Lift
{herb marker thyme} → {herb marker rosemary}	0.01	0.94	86.84
{herb marker rosemary} → {herb marker thyme}	0.01	0.93	86.84
{regency tea plate green} → {regency tea plate roses}	0.012	0.85	52.94
{regency tea plate roses} → {regency tea plate green}	0.012	0.72	52.94
{poppy's playhouse bedroom} → {poppy's playhouse livingroom}	0.01	0.65	51.78
{poppy's playhouse livingroom} → {poppy's playhouse bedroom}	0.01	0.81	51.78
{set of 3 wooden tree decorations} → {set of 3 wooden stocking decoration}	0.01	0.75	50.22
{set of 3 wooden stocking decoration} → {set of 3 wooden tree decorations}	0.01	0.69	50.22
{poppy's playhouse kitchen} → {poppy's playhouse livingroom}	0.011	0.62	49.23
{poppy's playhouse livingroom} → {poppy's playhouse kitchen}	0.011	0.85	49.23

Nhận xét:

- Các sản phẩm thuộc cùng dòng “herb marker”, “regency tea plate”, “poppy's playhouse”,... thường xuyên được mua cùng nhau.
- Support (Độ hỗ trợ):** Các giá trị Support đều ở mức thấp (khoảng 0.010 - 0.012, tức là 1% - 1.2%).
 - Ý nghĩa:** Điều này có nghĩa là các cặp sản phẩm này (ví dụ: {herb marker thyme} VÀ {herb marker rosemary}) chỉ xuất hiện cùng nhau trong khoảng 1% tổng số các giao dịch. Đây là điều bình thường trong các bộ dữ liệu bán lẻ lớn, nơi có hàng nghìn sản phẩm khác nhau. Mặc dù chúng hiếm khi xuất hiện cùng nhau, nhưng khi đã xuất hiện thì mối liên hệ lại rất mạnh.
- Confidence (Độ tin cậy):** Các giá trị Confidence nhìn chung rất cao, nhiều luật vượt trên 0.80 (80%) và thậm chí 0.90 (90%).
 - Ý nghĩa:** Chỉ số này cho thấy khả năng dự đoán cao.
 - Ví dụ:** Luật {herb marker thyme} → {herb marker rosemary} có Confidence là 0.94 (94%). Điều này có nghĩa là: "Trong số những khách hàng đã mua {herb marker thyme}, có đến 94% cũng đã mua {herb marker rosemary}". Đây là một tỷ lệ rất đáng tin cậy.
- Lift (Độ nâng):** Đây là chỉ số vô cùng quan trọng trong bảng.
 - Ý nghĩa:** Lift đo lường mức độ mà sự xuất hiện của vế trái (antecedent) làm tăng khả năng xuất hiện của vế phải (consequent), so với khả năng xuất hiện thông thường của vế phải.
 - Ví dụ:** Luật {herb marker thyme} → {herb marker rosemary} có Lift là 86.84. Điều này có nghĩa là một khách hàng mua {herb marker thyme} có khả năng mua {herb marker rosemary} cao gấp gần 87 lần so với một khách hàng ngẫu nhiên bước vào cửa hàng.

5.2 Trực quan hóa và Phân tích kết quả

5.2.1 Phân bố kích thước giỏ hàng



Biểu đồ này cho thấy số lượng mặt hàng trong mỗi hóa đơn.

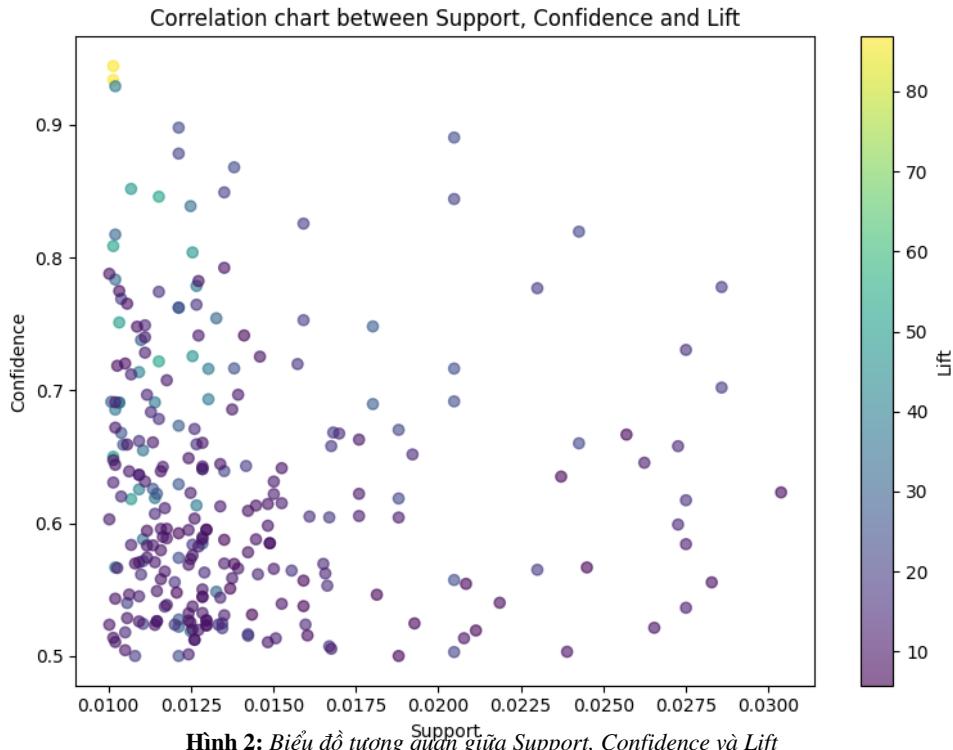
Nhận xét:

- Hầu hết các giỏ hàng có dưới 50 mặt hàng.
- Giỏ hàng trung bình có: 20.68 mặt hàng
- Giỏ hàng phổ biến nhất (median) có: 15.0 mặt hàng
- Điều này cho thấy có một số giỏ hàng rất lớn (mua sỉ hoặc cho sự kiện) kéo giá trị trung bình lên.

5.2.2 Tương quan Support, Confidence và Lift

Để trực quan hóa mối quan hệ giữa các luật, nhóm sử dụng biểu đồ phân tán (scatter plot) biểu diễn mối tương quan giữa Support, Confidence và Lift:

```
1 plt.figure(figsize=(8,6))
2 plt.scatter(rules['support'], rules['confidence'], alpha=0.6, c=rules['lift'],
3             cmap='viridis')
4 plt.title('Correlation chart between Support, Confidence and Lift')
5 plt.xlabel('Support')
6 plt.ylabel('Confidence')
7 plt.colorbar(label='Lift')
8 plt.tight_layout()
9 plt.show()
```



Hình 2: Biểu đồ tương quan giữa Support, Confidence và Lift

Trong đó:

- Trục hoành (X-axis) - Support (Độ hỗ trợ):** Cho biết mức độ phổ biến của một luật (tỷ lệ phần trăm các giao dịch có chứa tổ hợp sản phẩm đó). Giá trị càng sang phải, luật càng phổ biến.
- Trục tung (Y-axis) - Confidence (Độ tin cậy):** Cho biết độ tin cậy của luật (ví dụ: 90% khách hàng mua A cũng sẽ mua B). Giá trị càng lên cao, luật càng đáng tin cậy.
- Thanh màu (Color bar) - Lift (Độ nâng):**
 - Lift > 1: Có mối liên hệ — mua A làm tăng khả năng mua B.
 - Lift cao (màu vàng): Mối liên hệ mạnh, bất ngờ → luật tốt.
 - Lift thấp (màu tím): Mối liên hệ yếu.
- Mỗi chấm tròn** trên biểu đồ đại diện cho **một luật kết hợp** (ví dụ: Nếu mua {Sữa} thì mua {Bánh mì}).

Biểu đồ cho thấy một sự đánh đổi rất rõ ràng:

- Luật giá trị nhất nằm ở góc bên trái:** Các luật có Confidence (độ tin cậy) cao nhất (trên 0.9) và Lift (độ thu vị) cao nhất (> 70) đều có Support (độ phổ biến) rất thấp, tập trung quanh mốc 0.01.
- Luật phổ biến thì yếu:** Ngược lại, các luật có Support cao (phía bên phải biểu đồ) thường có Confidence và Lift thấp (nằm ở nửa dưới, màu tím sẫm).

Kết luận: Những mối liên hệ mạnh nhất và đáng tin cậy nhất lại là những mối liên hệ ít phổ biến nhất. Ngược lại các mối quan hệ yếu hơn và ít tin cậy hơn thì rất phổ biến.

5.2.3 Heatmap đồng xuất hiện (Top 20 sản phẩm)

Ta vẽ heatmap để trực quan hóa các sản phẩm bán chạy nhất xuất hiện cùng nhau:

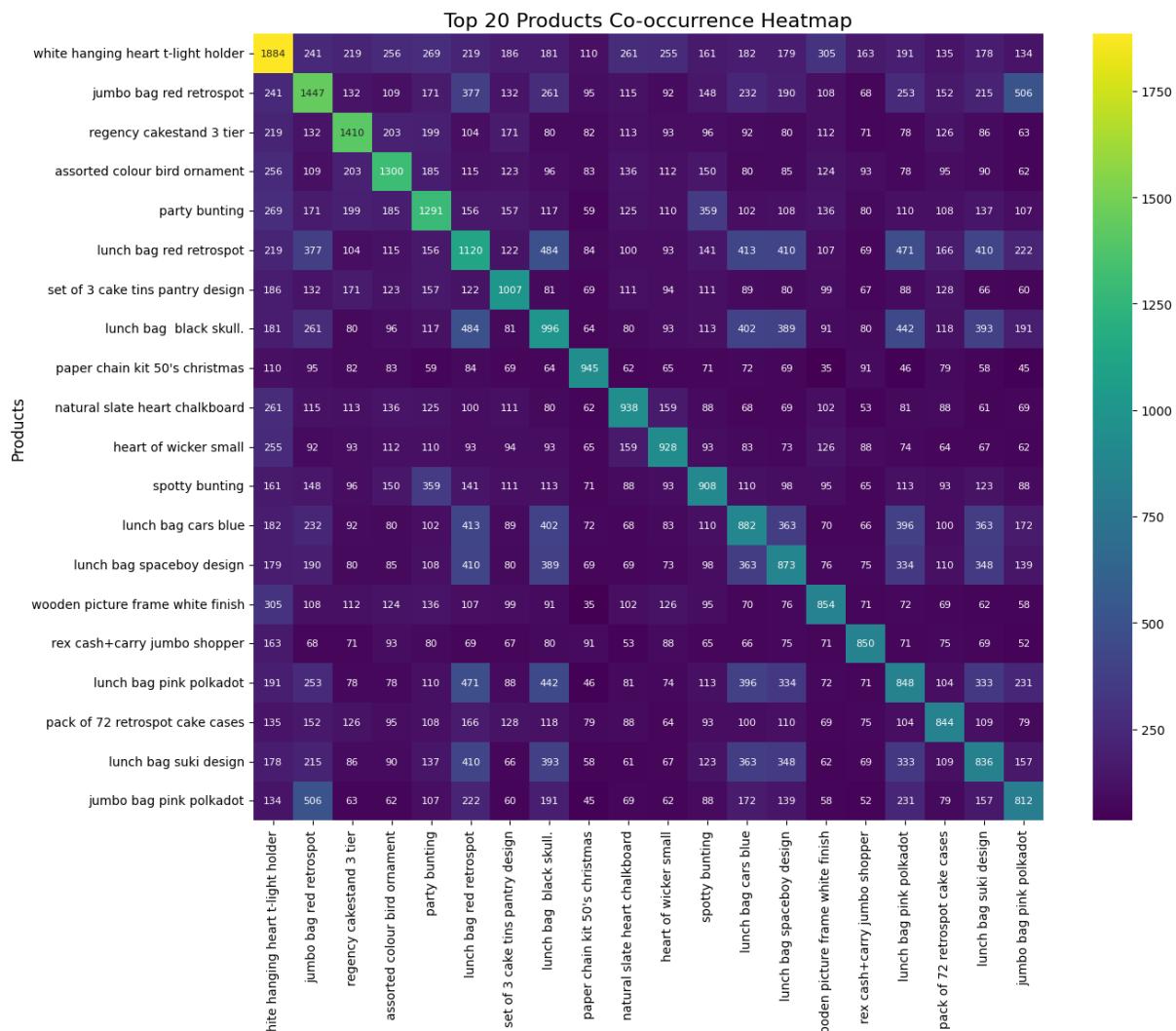
```

1 basket_sets_int = basket_sets.astype(int)
2 item_support = basket_sets_int.sum(axis=0) / basket_sets_int.shape[0]
3 topM = 20
4 top_products = item_support.sort_values(ascending=False).head(topM).index.tolist()
    
```

```

5 sub_basket = basket_sets_int[top_products]
6 cooc_matrix = np.dot(sub_basket.T, sub_basket)
7 cooc_df = pd.DataFrame(cooc_matrix, index=top_products, columns=top_products)
8 plt.figure(figsize=(14, 12))
9 sns.heatmap(cooc_df, annot=True, fmt='d', cmap='viridis', annot_kws={"size": 8})
10 plt.title(f"Top {len(top_products)} Products Co-occurrence Heatmap", size=16)
11 plt.xlabel("Products", size=12)
12 plt.ylabel("Products", size=12)
13 plt.xticks(rotation=90)
14 plt.yticks(rotation=0)
15 plt.tight_layout()

```



Hình 3: Heatmap đồng xuất hiện của Top 20 sản phẩm

Trong đó:

- Các ô màu sáng (vàng, xanh lá) biểu thị các cặp có tần suất đồng xuất hiện cao
- Trục X và Trục Y:** Cả hai trục đều liệt kê 20 sản phẩm bán chạy nhất (Top 20 sản phẩm).
- Ô giao nhau (Cell):** Mỗi ô vuông trên biểu đồ hiển thị số lần mà sản phẩm ở hàng Y và sản phẩm ở cột X xuất hiện chung trong một giỏ hàng.
- Màu sắc (Thanh màu bên phải):**



- Màu càng sáng (Vàng/Xanh lá nhạt): Chỉ ra số lần đồng xuất hiện càng cao. Đây là những cặp sản phẩm rất thường được mua cùng nhau.
 - Màu càng tối (Tím): Chỉ ra số lần đồng xuất hiện càng thấp.
- **Đường chéo (Từ trên trái xuống dưới phải):** Các ô màu vàng sáng nhất nằm trên đường chéo chính. Chúng biểu thị tổng số lần sản phẩm đó được bán (ví dụ: "white hanging heart t-light holder" giao với chính nó là 1884 lần). Con số này chính là Support (Hỗ trợ) của sản phẩm đó.

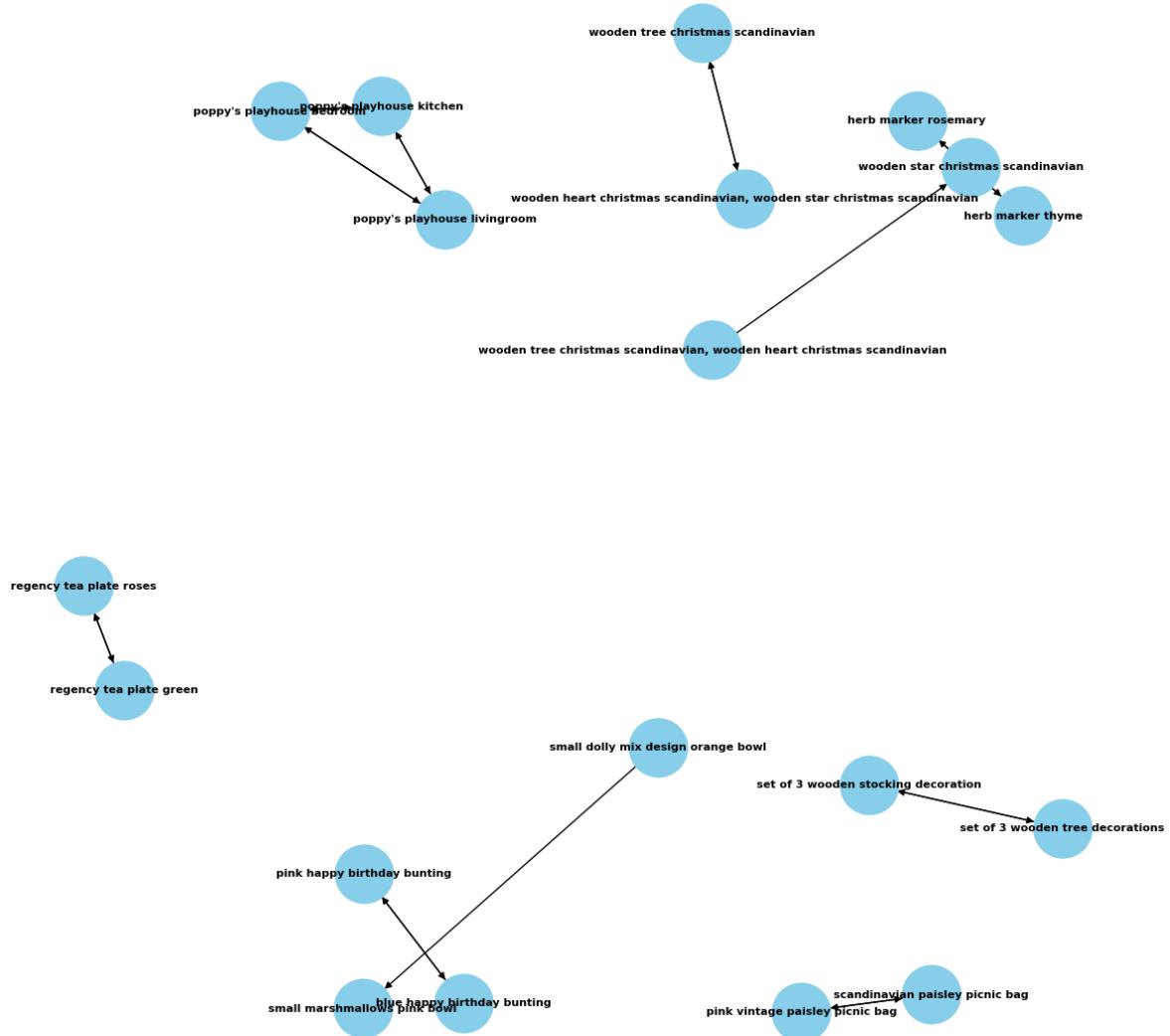
Nhận xét:

- **Cặp đôi mạnh nhất:** Ô giao giữa "regency cakestand 3 tier" (khay bánh 3 tầng) và "white hanging heart t-light holder" (đèn nến trái tim treo) có màu xanh lá cây sáng (giá trị 219). Đây là hai trong số các sản phẩm được mua cùng nhau nhiều nhất trong top 20.
- Dựa theo heatmap ta thấy: "white hanging heart t-light holder" (sản phẩm bán chạy nhất) có tần suất xuất hiện cùng "jumbo bag red retrospot" (241 lần) và "regency cakestand 3 tier" (219 lần) khá cao.
- Các sản phẩm trong cùng bộ "Lunch Bag" (ví dụ: red retrospot, black skull, spaceboy design, suki design) cũng thường được mua cùng nhau, tạo thành một cụm màu sáng rõ rệt.
- **Cụm sản phẩm (Cluster):** Có thể thấy một cụm màu sáng (xanh lá) xung quanh các sản phẩm "Lunch Bag". Ví dụ: "lunch bag black skull." (túi ăn trưa đầu lâu) thường được mua cùng "lunch bag red retrospot" (484 lần) và "lunch bag spaceboy design" (413 lần). Điều này cho thấy khách hàng khi mua túi ăn trưa thường có xu hướng mua nhiều kiểu dáng khác nhau cùng một lúc.

5.2.4 Đồ thị mạng lưới

Đồ thị : Top 20 luật kết hợp mạnh nhất

Network graph of the Top 20 rules (according to Lift)



Hình 4: Đồ thị mạng lưới của Top 20 luật (theo Lift)

Trong đó:

- **Nút (Node):** Mỗi vòng tròn màu xanh da trời là một sản phẩm đơn lẻ.
- **Mũi tên (Edge):** Một mũi tên đi từ sản phẩm A đến sản phẩm B biểu thị một luật kết hợp $\{A\} \rightarrow \{B\}$. Mũi tên chỉ ra rằng khách hàng mua A cũng có xu hướng mua B.
- **Cụm (Cluster):** Các nút được nối với nhau dày đặc (ví dụ: bộ "Herb Marker") được gọi là một cụm.
- Thuật toán spring_layout tự động kéo các nút có liên kết mạnh lại gần nhau và đẩy các nút không liên quan ra xa.

Nhận xét:

- Đồ thị này cho thấy rất rõ xu hướng mua hàng theo bộ sưu tập.
- Thay vì có một mạng lưới phức tạp nối tất cả sản phẩm với nhau, chúng ta thấy các "cụm" hay "hòn đảo" riêng biệt:
 - **Cụm 1 (Góc trên trái):** Bộ đồ chơi "Poppy's Playhouse". Các món như poppy's playhouse kitchen (nhà bếp), bedroom (phòng ngủ), và livingroom (phòng khách) được liên kết chặt chẽ với nhau.



- * **Điễn giải:** Khách hàng khi mua một phòng trong bộ đồ chơi này có xu hướng rất cao sẽ mua các phòng còn lại để hoàn thiện bộ sưu tập.
- **Cụm 2 (Góc trên phải): Bộ thẻ thảo mộc "Herb Marker".** Các thẻ herb marker rosemary (hương thảo), thyme (húng tây), parsley (ngò tây), và mint (bạc hà) tạo thành một cụm dày đặc.
 - * **Điễn giải:** Đây là luật mạnh nhất trong bộ dữ liệu (Lift > 80). Nếu khách hàng mua bất kỳ thẻ nào trong số này, họ gần như chắc chắn sẽ mua các thẻ còn lại.
- **Cụm 3 (Giữa bên phải): Bộ trang trí Giáng sinh.** Các món wooden tree christmas decorations, wooden heart christmas decoration, và wooden star christmas decoration được liên kết với nhau.
 - * **Điễn giải:** Khách hàng mua đồ trang trí Giáng sinh bằng gỗ cũng có xu hướng mua theo bộ chủ đề (cây thông, trái tim, ngôi sao).
- **Cụm 4 (Góc dưới trái): Bộ đĩa trà "Regency Tea Plate".** Các đĩa màu roses (hoa hồng), green (xanh lá), và pink (hồng) được nối với nhau, cho thấy hành vi mua sắm tương tự.

Kết luận: Đồ thị này trực quan hóa một cách hiệu quả rằng hành vi mua hàng mạnh mẽ nhất trong bộ dữ liệu này không phải là mua ngẫu nhiên, mà là mua có chủ đích để hoàn thành một bộ sưu tập cụ thể.

5.3 Phân tích kết quả mô hình

Kết quả cho thấy khách hàng có xu hướng:

- Mua theo bộ sản phẩm cùng chủ đề, ví dụ như bộ tách trà Regency với các màu khác nhau.
- Mua nhóm sản phẩm trang trí cùng loại, chẳng hạn “hanging heart t-light holder” thường đi kèm với “retrosport bag”.
- Các sản phẩm có tính bổ sung hoặc tính thẩm mỹ tương đồng thường tạo ra lift cao.

Bảng 6: Phân tích diễn giải các luật mạnh

Luật kết hợp	Điễn giải
{herb marker thyme} → {herb marker rosemary}	Nếu một khách hàng mua {Thẻ tên cây húng tây} (herb marker thyme) thì họ cũng có 94% khả năng họ sẽ mua {Thẻ tên cây hương thảo} (herb marker rosemary) trong cùng một giao dịch.
{regency tea plate green} → {regency tea plate roses}, {regency tea plate roses} → {regency tea plate green}	Khách hàng có xu hướng mua sắm trọn bộ đĩa trà Regency.
{set of 3 wooden tree decorations} → {set of 3 wooden stocking decoration}, {set of 3 wooden stocking decoration} → {set of 3 wooden tree decorations}	Khách hàng mua sản phẩm trang trí liên quan đến nhau (như đồ trang trí cây mùa cùng đồ trang trí tất cho giáng sinh).

5.4 Ứng dụng mô hình khuyến nghị

Từ các luật kết hợp thu được, nhóm đề xuất cách triển khai mô hình khuyến nghị như sau:

- Khi khách hàng thêm “herb marker thyme” vào giỏ hàng → hệ thống gợi ý thêm “green regency teacup”.
- Khi mua “set of 3 wooden tree decorations” → gợi ý “set of 3 wooden stocking decoration” để trang trí cho ngày lễ.
- Khi mua một sản phẩm thuộc bộ “poppy’s playhouse” → gợi ý thêm các sản phẩm còn lại để hoàn thiện bộ sản phẩm.

Các gợi ý này giúp:

- Tăng tỷ lệ mua thêm (upselling & cross-selling)
- Cải thiện trải nghiệm người dùng khi mua hàng trực tuyến
- Tối ưu doanh thu theo hành vi mua sắm thực tế



6 Phân tích xu hướng mua sắm chủ đạo

Việc phân tích hành vi mua sắm của khách hàng có tầm quan trọng chiến lược vô cùng lớn. Thay vì chỉ tập trung vào việc "khách hàng mua gì", chúng ta cần đi sâu vào việc tìm hiểu "tại sao" và "bằng cách nào" họ đưa ra quyết định mua hàng. Việc nắm bắt được những động lực sâu xa này là nền tảng để xây dựng các chiến lược kinh doanh hiệu quả, tạo ra lợi thế cạnh tranh bền vững và xây dựng lòng trung thành của khách hàng.

6.1 Xu hướng "Sưu Tầm Trọn Bộ" (Product Set Pattern)

Một trong những xu hướng rõ nét và có giá trị thương mại cao nhất là thói quen mua sản phẩm theo bộ sưu tập hoàn chỉnh, phản ánh tâm lý "sưu tầm" của người tiêu dùng.

- **Minh chứng điển hình:** Bộ sản phẩm tách trà Regency với các biến thể màu sắc (hồng, xanh, hoa hồng) cho thấy một mối liên kết đặc biệt mạnh mẽ.
- **Các chỉ số chiến lược:**
 - **Lift > 20:** Chỉ số này cho thấy khả năng khách hàng mua một màu tách trà Regency sẽ mua thêm một màu khác cao hơn 20 lần so với mức trung bình. Nói cách khác, hành vi mua một chiếc tách trà Regency là một chỉ báo dự đoán cực kỳ mạnh mẽ cho việc mua một chiếc khác trong cùng bộ, biến đây thành một hành vi có thể dự đoán và thúc đẩy một cách có chủ đích.
 - **Confidence từ 84-89%:** Khi một khách hàng đã mua một sản phẩm trong bộ, có tới 84-89% khả năng họ sẽ mua thêm một sản phẩm khác. Đây là một tỷ lệ dự báo hành vi cao đến mức có thể được xem là một cam kết mua hàng ngầm.

Ý nghĩa kinh doanh: Xu hướng này đặc biệt phổ biến với các mặt hàng trang trí và quà tặng. Khách hàng không chỉ mua một sản phẩm, họ đang đầu tư vào một bộ sưu tập hoàn chỉnh, và chúng ta phải đáp ứng vẹn hành trình đó.

6.2 Sức Hút Của Mua Sắm Theo Chủ Đề và Dip Lễ

Dữ liệu khẳng định rằng các sản phẩm trang trí và quà tặng là động lực doanh thu cốt lõi, đặc biệt chịu ảnh hưởng mạnh mẽ bởi tính mùa vụ. Các sản phẩm phổ biến nhất trong danh mục bao gồm:

- White hanging heart t-light holder (11.3%)
- Jumbo bag red retrospot (8.7%)
- Assorted colour bird ornament (7.8%)
- Party bunting (7.8%)

Dữ liệu trong giai đoạn 12/2010 - 12/2011 cho thấy nhu cầu đối với nhóm hàng này tăng vọt trong các dịp lễ lớn như Giáng sinh và năm mới. Sự thống trị của nhóm hàng này trong danh sách bán chạy nhất cũng chính là một biểu hiện của Nguyên tắc Pareto, cho thấy động lực doanh thu chính của chúng ta mang tính mùa vụ sâu sắc.

6.3 Tác Động Của Nguyên Tắc Pareto (80/20)

Phân tích dữ liệu đã xác nhận một cách thuyết phục sự hiện diện của nguyên tắc Pareto: khoảng 20% sản phẩm đang tạo ra 80% tổng doanh số. Tần suất mua của các sản phẩm "hot" này vượt trội hoàn toàn so với phần còn lại của danh mục. **Hàm ý chiến lược:** Phát hiện này bắt buộc chúng ta phải thay đổi tư duy phân bổ nguồn lực. Thay vì dàn trải, chúng ta phải tập trung một cách quyết liệt vào việc bảo vệ và khuếch đại doanh số từ nhóm sản phẩm cốt lõi này. Mọi quyết định về marketing, tồn kho và chuỗi cung ứng phải bắt đầu từ đây.

6.4 Sức Mạnh Kết Hợp: Sản Phẩm Bổ Sung và "Combo Ngách"

Dữ liệu còn tiết lộ sức mạnh của việc kết hợp các sản phẩm có tính bổ trợ, thúc đẩy khách hàng mua nhiều hơn trong một lần giao dịch.

- Sản phẩm bổ sung: Các sản phẩm có tính thẩm mỹ tương đồng hoặc chức năng bổ trợ thường được mua cùng nhau, ví dụ như Hanging heart t-light holder đi kèm Jumbo bag retrospot (để đóng gói quà), hoặc Jam making set kết hợp với Jam jar (lọ đựng mứt).
- "Combo Ngách" (Niche Combos): Phân tích biểu đồ scatter plot cho thấy các luật kết hợp có Lift > 15 thường có Confidence > 0.7. Đáng chú ý, ngay cả những cặp sản phẩm có độ phổ biến thấp (Support thấp) vẫn có thể sở hữu chỉ số Lift rất cao. Đây chính là các "combo ngách" — những cơ hội vàng cho các chiến dịch marketing siêu nhỏ, nhắm đến các phân khúc khách hàng chuyên biệt mà đối thủ cạnh tranh có thể bỏ qua. Việc khai thác chúng cho phép chúng ta chiếm lĩnh thị trường ngách với chi phí thấp và tỷ lệ chuyển đổi cao.

7 Kết luận và phương hướng phát triển

7.1 Kết luận

Trong đề tài “Phân tích giỏ hàng để đưa ra khuyến nghị sản phẩm”, nhóm đã hoàn thành toàn bộ quy trình phân tích dữ liệu theo hướng tiếp cận CRISP-DM, từ làm sạch dữ liệu, chuyển đổi sang dạng nhị phân, phân tích khám phá cho đến khai phá luật kết hợp bằng thuật toán Apriori. Việc thử nghiệm nhiều nguồn hỗ trợ khác nhau cho thấy sự đánh đổi rõ rệt giữa độ chi tiết của mô hình và thời gian xử lý, qua đó nhóm đã lựa chọn nguồn min_support = 1% như một mức tối ưu.

Kết quả phân tích cho thấy trong bộ dữ liệu Online Retail, khách hàng có xu hướng mua hàng theo bộ hoặc theo chủ đề. Nhiều luật kết hợp thu được có confidence rất cao (trên 0.8 – 0.9) và lift cực lớn (từ 20 đến 80), phản ánh mối quan hệ mua sắm rất mạnh, dù các luật này có độ phổ biến (support) thấp. Điều này đặc biệt nổi bật ở các nhóm sản phẩm như Herb Marker, Regency Tea Plate, Poppy's Playhouse, hoặc các sản phẩm trang trí Giáng sinh. Các biểu đồ scatter plot, heatmap và đồ thị mang lưới đã trực quan hóa rõ ràng những cụm sản phẩm thường đồng xuất hiện và hành vi mua theo bộ của khách hàng.

Từ các luật kết hợp, nhóm đã xây dựng được một mô hình gợi ý đơn giản dựa trên confidence, có thể áp dụng cho các hệ thống thương mại điện tử nhằm hỗ trợ người dùng, tăng tỉ lệ mua thêm (cross-selling) và tối ưu doanh thu. Phân tích tổng thể cũng cho thấy tác động mạnh của tính mùa vụ, sự thống trị của các sản phẩm thuộc nhóm trang trí/quà tặng và xác nhận sự hiện diện của nguyên tắc Pareto 80/20 trong dữ liệu bán lẻ.

Nhìn chung, đề tài đã minh chứng tính hữu ích của phân tích giỏ hàng trong việc hiểu hành vi mua sắm và hỗ trợ ra quyết định kinh doanh. Mặc dù còn một số hạn chế về phạm vi dữ liệu và phương pháp, đây vẫn là cơ sở quan trọng để phát triển các mô hình khuyến nghị phức tạp hơn trong tương lai.

7.2 Phương hướng phát triển

Dựa trên kết quả hiện tại, để mở rộng và nâng cao hiệu quả mô hình đề tài có thể được nghiên cứu và phát triển theo các hướng tiếp theo:

1. Kết hợp thêm dữ liệu người dùng (User profiling)

Hiện tại mô hình chỉ dựa trên dữ liệu giao dịch. Trong tương lai, có thể bổ sung:

- Thông tin khách hàng (CustomerID, độ tuổi, phân khúc)
- Lịch sử mua sắm theo thời gian
- Địa điểm, mùa vụ và hoá đơn theo dịp lễ.

2. Kết hợp nhiều phương pháp gợi ý khác nhau

Có thể tích hợp thêm:



- Collaborative Filtering (gợi ý dựa vào khách hàng tương tự),
- Content-based filtering (gợi ý dựa trên thông tin sản phẩm),
- Hybrid Recommendation System (kết hợp nhiều phương pháp).

3. Đánh giá mô hình trên môi trường thực tế

Hiện mô hình chưa được triển khai trong hệ thống thương mại điện tử để đánh giá qua:

- Tỉ lệ click (CTR)
- Tỉ lệ chuyển đổi
- Doanh thu tăng thêm từ gợi ý
- A/B Testing giữa các mô hình khuyến nghị

4. Phân tích nâng cao theo thời gian và mùa vụ

Dữ liệu Online Retail chịu ảnh hưởng mạnh bởi dịp lễ (Giáng sinh, năm mới). Tương lai có thể:

- Áp dụng time-series analysis
- Phân tích theo tháng/quý
- Khai phá luật kết hợp theo từng mùa vụ
- So sánh hành vi khách hàng giữa các giai đoạn



Phụ lục

1. Nguồn dữ liệu

- **Bộ dữ liệu:** Online Retail Dataset (2010–2011)
- **Nguồn:** UCI Machine Learning Repository
- **Liên kết:** <https://archive.ics.uci.edu/ml/datasets/Online+Retail>
- **Mô tả:** Dữ liệu giao dịch bán lẻ tại Anh, được sử dụng để phân tích giỏ hàng và khai phá luật kết hợp.

2. Nguồn công cụ và thư viện

- Python 3.10, Jupyter Notebook
- Các thư viện: *pandas, numpy, mlxtend, matplotlib, seaborn, networkx*
- Công cụ trực quan hóa: Excel / Tableau (nếu có)
- Phần mềm soạn thảo: LaTeX / Microsoft Word

Tài liệu tham khảo

1. Agrawal, R., Imieliński, T., & Swami, A. (1993). *Mining Association Rules between Sets of Items in Large Databases*.
2. Agrawal, R., & Srikant, R. (1994). *Fast Algorithms for Mining Association Rules*.
3. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*.
4. Bài giảng môn *Khai phá dữ liệu* – Khoa Khoa học & Kỹ thuật Máy tính, ĐHBK TP.HCM.
5. GeeksforGeeks. *Apriori Algorithm – Implementation & Examples*. <https://www.geeksforgeeks.org/apriori-algorithm/>
6. Towards Data Science. *Market Basket Analysis & Association Rules*. <https://towardsdatascience.com/>
7. Mlxtend Documentation – Apriori and Association Rules. <http://rasbt.github.io/mlxtend/>
8. pandas Documentation – DataFrame Processing. <https://pandas.pydata.org/docs/>

Mã nguồn

Link GitHub: [CO3029_DATA_MINING_L01](https://github.com/CO3029_DATA_MINING_L01)