TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA TOÁN - ỨNG DỤNG

❖ 🕮 ❖

BÁO CÁO CUỐI KÌ KHAI PHÁ DỮ LIỆU

ĐỀ TÀI: DỰ ĐOÁN SỰ HÀI LÒNG CỦA KHÁCH HÀNG

Giảng viên: Đỗ Như Tài

Lớp: DDU1231

Danh sách sinh viên thực hiện:

Hồ Gia Bảo – 3123580003

Nguyễn Gia Huy – 3123580016

Trần Nguyễn Minh Tiến – 3123580052

TP HCM, THÁNG 12, NĂM 2025

DANH MỤC ĐÁNH GIÁ THÀNH VIÊN

STT	Họ Tên	MSSV	Mức độ đóng góp	Kí tên
1	Hồ Gia Bảo	3123580003	33.33%	
2	Nguyễn Gia Huy	3123580016	33.33%	
3	Trần Nguyễn Minh Tiến	3123580052	33.33%	

Mục Lục

LỜI MỞ ĐẦU

Trong bối cảnh chuyển đổi số diễn ra mạnh mẽ, thương mại điện tử đã trở thành một trong những lĩnh vực phát triển nhanh nhất và tạo ra lượng dữ liệu khổng lồ mỗi ngày. Việc khai thác, phân tích và hiểu được các mẫu hành vi trong dữ liệu trở thành yếu tố quan trọng giúp doanh nghiệp tối ưu hóa hoạt động, nâng cao trải nghiệm khách hàng và gia tăng doanh thu.

Bộ dữ liệu Brazilian E-Commerce Public Dataset, do Olist công bố, là một tập dữ liệu thực tế và đa chiều bao gồm thông tin về đơn hàng, khách hàng, sản phẩm, thanh toán, đánh giá và quá trình giao hàng. Đây là nguồn dữ liệu lý tưởng để áp dụng các kỹ thuật khai phá dữ liệu nhằm phát hiện tri thức hữu ích và hỗ trợ ra quyết định.

Trong báo cáo này, nhóm tiến hành thu thập, xử lý và phân tích bộ dữ liệu theo các bước của quy trình khai phá dữ liệu, bao gồm: tiền xử lý dữ liệu, phân tích mô tả, áp dụng các thuật toán như phân cụm, phân loại, luật kết hợp, và trực quan hóa kết quả. Mục tiêu của báo cáo là rút ra những nhận định giá trị từ dữ liệu, đồng thời minh họa rõ ràng quy trình ứng dụng các phương pháp khai phá dữ liệu vào bài toán thực tế trong thương mại điện tử.

Thông qua đề tài này, sinh viên có cơ hội rèn luyện kỹ năng phân tích dữ liệu, lập trình Python, sử dụng thư viện khai phá dữ liệu, và hơn hết là khả năng diễn giải kết quả để phục vụ công tác ra quyết định trong kinh doanh.

LỜI CẢM ƠN

Trong suốt quá trình thực hiện báo cáo cuối kì môn Khai phá dữ liệu, chúng em đã nhận được rất nhiều sự hướng dẫn và hỗ trợ quý báu. Trước hết, chúng em xin gửi lời cảm ơn sâu sắc đến thầy Đỗ Như Tài, người đã tận tình giảng dạy, truyền đạt kiến thức và định hướng phương pháp nghiên cứu một cách rõ ràng, giúp chúng em có nền tảng vững chắc để hoàn thành tốt đề tài.

Những bài giảng, sự hỗ trợ và những góp ý chi tiết của thầy không chỉ giúp chúng em hiểu sâu hơn về khai phá dữ liệu mà còn phát triển tư duy phân tích và kỹ năng làm việc với dữ liệu thực tế. Đây là hành trang quan trọng đối với chúng em trong học tập cũng như trong con đường nghề nghiệp sau này.

Chúng em cũng xin chân thành cảm ơn các thầy cô trong khoa, bạn bè và những người đã hỗ trợ, tạo môi trường học tập thuận lợi để chúng em có thể hoàn thiện báo cáo một cách tốt nhất.

Mặc dù đã cố gắng hết sức, báo cáo vẫn khó tránh khỏi những thiếu sót. Kính mong thầy và quý thầy cô bỏ qua và có những góp ý để chúng em hoàn thiện hơn trong tương lai.

Chúng em xin chân thành cảm ơn!

Chương 1: Giới thiệu

1.1. Lý do chọn đề tài

Trong bối cảnh thương mại điện tử ngày càng phát triển mạnh mẽ, đặc biệt tại các quốc gia có tốc độ số hóa nhanh như Brazil, việc khai thác và phân tích dữ liệu trở nên vô cùng quan trọng. Dữ liệu không chỉ giúp doanh nghiệp hiểu rõ hành vi khách hàng mà còn hỗ trợ tối ưu hóa quy trình vận chuyển, thanh toán và chăm sóc khách hàng.

Bộ dữ liệu Brazilian E-Commerce Public Dataset là một bộ dữ liệu thực tế, quy mô lớn, mô tả toàn bộ quy trình mua hàng từ đặt hàng, giao hàng đến đánh giá của khách hàng. Đây là nguồn dữ liệu phù hợp để vận dụng các kiến thức khai phá dữ liệu như tiền xử lý, phân tích mô tả, phân cụm, phân loại, luật kết hợp,…

Vì vậy, nhóm chọn đề tài này nhằm vận dụng kiến thức đã học vào bộ dữ liệu thực tế, qua đó rèn luyện kỹ năng phân tích dữ liệu và hiểu hơn về hoạt động của thương mại điện tử.

1.2. Tình hình nghiên cứu (Literature Review)

Trong những năm gần đây, dự đoán sự hài lòng của khách hàng trong thương mại điện tử là một chủ đề được nghiên cứu rộng rãi do vai trò quan trọng trong việc nâng cao trải nghiệm người dùng và hiệu quả kinh doanh. Các nghiên cứu trước đây chủ yếu ứng dụng các kỹ thuật khai phá dữ liệu và học máy để phân tích hành vi mua sắm, đánh giá của khách hàng và các yếu tố ảnh hưởng đến mức độ hài lòng.

Trong nhóm thuật toán học máy có giám sát, Logistic Regression thường được sử dụng làm mô hình cơ sở nhờ tính đơn giản, dễ diễn giải và khả năng dự đoán xác suất khách hàng hài lòng. K-Nearest Neighbors (KNN) được áp dụng dựa trên mức độ tương đồng giữa các khách hàng, cho kết quả tốt khi dữ liệu có cấu trúc rõ ràng nhưng dễ bị ảnh hưởng bởi kích thước dữ liệu lớn. Decision Tree giúp mô hình hóa các yếu tố ảnh hưởng đến sự hài lòng một cách trực quan, tuy nhiên có nguy cơ overfitting. Để khắc phục hạn chế này, Random Forest được sử dụng nhằm cải thiện độ chính xác và khả năng tổng quát hóa thông qua việc kết hợp nhiều cây quyết định.

Bên cạnh các mô hình phân loại, một số nghiên cứu sử dụng K-Means Clustering để phân nhóm khách hàng dựa trên hành vi mua sắm, từ đó hỗ trợ phân khúc khách hàng và xây dựng chiến lược kinh doanh phù hợp. Trong trường hợp

dữ liệu mất cân bằng, SMOTE được áp dụng để tăng số lượng mẫu của lớp thiểu số, giúp cải thiện hiệu suất các mô hình phân loại.

Ngoài ra, khai phá luật kết hợp là hướng tiếp cận phổ biến nhằm phân tích hành vi mua sắm. Apriori được sử dụng để phát hiện các tập mục thường xuyên và luật mua chung, trong khi FP-Growth là phiên bản tối ưu hơn, giúp giảm thời gian xử lý và phù hợp với dữ liệu lớn.

Tổng hợp các nghiên cứu cho thấy việc kết hợp các thuật toán phân loại, phân cụm và khai phá luật kết hợp mang lại cái nhìn toàn diện về sự hài lòng và hành vi mua sắm của khách hàng. Nghiên cứu này kế thừa các hướng tiếp cận trên và áp dụng chúng vào bộ dữ liệu Brazilian E-Commerce nhằm đánh giá hiệu quả và lựa chọn mô hình phù hợp cho bài toán dự đoán sự hài lòng của khách hàng.

1.3. Mục tiêu và Câu hỏi nghiên cứu

Mục tiêu chính:

Nghiên cứu này được thực hiện nhằm đạt được các mục tiêu sau:

Xây dựng mô hình dự đoán mức độ hài lòng (điểm đánh giá) của khách hàng sau khi mua hàng trên nền tảng Olist.

Phân tích mức độ ảnh hưởng của các yếu tố liên quan đến đơn hàng và dịch vụ đến sự hài lòng của khách hàng.

Khám phá các xu hướng và mối quan hệ giữa đặc điểm đơn hàng và đánh giá của khách hàng thông qua phân tích dữ liệu và trực quan hóa.

So sánh hiệu quả của các mô hình học máy khác nhau nhằm lựa chọn mô hình phù hợp cho bài toán dự đoán sự hài lòng của khách hàng.

Dựa trên dữ liệu Olist và phạm vi nghiên cứu, đề tài tập trung kiểm định các giả thuyết sau:

H1: Giá trị đơn hàng có ảnh hưởng đến mức độ hài lòng của khách hàng không?.

H2: Thời gian giao hàng và trạng thái giao hàng có ảnh hưởng đáng kể đến mức độ hài lòng của khách hàng không?.

H3: Phí vận chuyển cao có thể làm giảm mức độ hài lòng của khách hàng không?.

H4: Hình thức thanh toán có ảnh hưởng đến sự hài lòng của khách hàng không?.

H5: Mức độ chi tiết của mô tả sản phẩm có tác động tích cực đến sự hài lòng của khách hàng không?.

H6: Trạng thái hoàn tất đơn hàng ảnh hưởng trực tiếp đến đánh giá của khách hàng không?.

H7: Các vấn đề nào (giao hàng, sản phẩm, thanh toán) thường đi kèm với đơn hàng đánh giá thấp (≤3 sao) bằng Association Rules Mining.

1.4. Phương pháp nghiên cứu và Hướng tiếp cận

Nguồn dữ liệu: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Sử dụng bộ dữ liệu từ Olist Shops (Brazil) gồm khoảng 100.000 giao dịch từ năm 2016 đến 2018.

Công cụ: Ngôn ngữ lập trình Python và môi trường Jupyter Notebook, Visual Code.

Hướng tiếp cận giải quyết vấn đề (4 hướng tiếp cận chính):

Áp dụng các mô hình Machine Learning

Nghiên cứu sử dụng các thuật toán học máy có giám sát để xây dựng mô hình phân loại mức độ hài lòng của khách hàng (Hài lòng / Không hài lòng), bao gồm:

Logistic Regression

K-Nearest Neighbors (KNN)

Decision Tree (Cây quyết định)

Random Forest

Naive Bayes

Các mô hình được huấn luyện và so sánh dựa trên các chỉ số đánh giá như Accuracy, Precision, Recall và F1-score nhằm lựa chọn mô hình phù hợp nhất cho bài toán nghiên cứu.

Tiền xử lý dữ liệu và xử lý mất cân bằng

Trước khi huấn luyện mô hình, dữ liệu được tiền xử lý thông qua các bước:

Làm sạch dữ liệu và xử lý các giá trị bị thiếu.

Chuẩn hóa các đặc trưng số và mã hóa các biến phân loại.

Áp dụng kỹ thuật SMOTE (Synthetic Minority Over-sampling Technique) nhằm cân bằng lại dữ liệu trong trường hợp số lượng khách hàng không hài lòng chiếm tỷ lệ nhỏ, giúp mô hình học hiệu quả hơn.

Xây dựng quy trình xử lý dữ liệu (Pipeline)

Một pipeline xử lý dữ liệu được xây dựng nhằm tự động hóa toàn bộ quy trình từ tiền xử lý dữ liệu đến huấn luyện và đánh giá mô hình. Pipeline giúp đảm bảo tính nhất quán, khả năng lặp lại và giảm sai sót trong quá trình thực nghiệm.

Phân khúc khách hàng (Customer Segmentation)

Bên cạnh bài toán phân loại, nghiên cứu sử dụng K-Means Clustering để phân nhóm khách hàng dựa trên các đặc trưng liên quan đến hành vi mua sắm. Phương pháp Elbow được áp dụng để xác định số lượng cụm tối ưu. Kết quả phân cụm giúp khám phá các nhóm khách hàng tiềm ẩn và hỗ trợ phân tích sâu hơn về hành vi tiêu dùng.

Chương 2: Cơ sở lý thuyết

2.1. Các Khái niệm Chính (Concepts)

2.2. Thuật toán Sử dụng (Algorithms Used)

Báo cáo áp dụng đa dạng các thuật toán, từ cơ bản đến nâng cao, chia thành các nhóm mục đích sau:

A. Thuật toán Học máy cho Phân loại (Dự đoán sự hài lòng)

1. Logistic Regression

Logistic Regression là thuật toán phân loại nhị phân, dùng để ước lượng xác suất một mẫu dữ liệu thuộc về lớp "Hài lòng" hay "Không hài lòng". Thuật toán sử dụng hàm sigmoid để ánh xạ đầu ra tuyến tính về khoảng $[0, 1]$.

Hàm sigmoid:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Trong đó:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Xác suất khách hàng hài lòng:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}}$$

Quy tắc phân loại:

Nếu $P(Y = 1 \mid X) \geq 0.5 \rightarrow$ Hài lòng

Ngược lại $\rightarrow$ Không hài lòng

2. K-Nearest Neighbors (KNN)

KNN là thuật toán phân loại dựa trên khoảng cách giữa các điểm dữ liệu. Nhãn của một mẫu mới được xác định dựa trên đa số nhãn của k láng giềng gần nhất trong tập huấn luyện.

Khoảng cách Euclidean thường dùng:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Trong đó:

x: điểm dữ liệu cần phân loại

y: điểm dữ liệu trong tập huấn luyện

Mẫu mới được gán vào lớp xuất hiện nhiều nhất trong k điểm có khoảng cách nhỏ nhất.

3. Decision Tree (Cây quyết định)

Decision Tree phân chia dữ liệu thành các nhánh dựa trên giá trị của các thuộc tính. Việc lựa chọn thuộc tính chia được thực hiện dựa trên tiêu chí đo độ "thuần khiết" của nút.

Chỉ số Gini:

$$Gini = 1 - \sum_{i=1}^{c} p_i^2$$

Trong đó:

p_i: tỷ lệ mẫu thuộc lớp itại nút

c: số lớp

Thuộc tính có chỉ số Gini nhỏ nhất sẽ được chọn để chia nút.

4. Random Forest

Random Forest là phương pháp học tập hợp (ensemble learning), kết hợp nhiều cây quyết định được huấn luyện trên các tập dữ liệu con khác nhau.

Dự đoán cuối cùng:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \dots, T_m(x)\}$$

Trong đó:

$T_i(x)$: kết quả dự đoán của cây quyết định thứ i

m: số lượng cây

Random Forest giúp giảm overfitting và tăng độ chính xác so với cây quyết định đơn lẻ.

B. Thuật toán Phân cụm và Xử lý dữ liệu

5. K-Means Clustering

K-Means là thuật toán phân cụm không giám sát, chia dữ liệu thành kcụm sao cho các điểm trong cùng cụm có độ tương đồng cao nhất.

Hàm mục tiêu cần tối thiểu hóa:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \| x - \mu_i \|^2$$

Trong đó:

C_i: cụm thứ i

μ_i: tâm cụm i

Thuật toán lặp lại việc:

Gán điểm vào cụm gần nhất.

Cập nhật lại tâm cụm.

Cho đến khi hội tụ.

6. Hierarchical Clustering (Phân cụm phân cấp)

Hierarchical Clustering là thuật toán phân cụm không giám sát, xây dựng cấu trúc phân cụm theo dạng cây (dendrogram). Thuật toán không cần xác định trước số cụm, mà cho phép quan sát dữ liệu ở nhiều mức độ phân nhóm khác nhau.

Có hai hướng tiếp cận chính:

- **Agglomerative (Bottom-up)**: mỗi điểm dữ liệu là một cụm, sau đó lần lượt gộp các cụm gần nhau nhất.
- **Divisive (Top-down)**: bắt đầu từ một cụm lớn rồi chia nhỏ dần (ít phổ biến hơn).

Trong nghiên cứu này, phương pháp Agglomerative Hierarchical Clustering được sử dụng.

Khoảng cách giữa hai điểm dữ liệu

Khoảng cách giữa hai điểm thường được tính bằng Euclidean:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Khoảng cách giữa hai cụm (Linkage)

Khoảng cách giữa hai cụm $C_i$và C_jđược xác định dựa trên tiêu chí liên kết (linkage):

- **Single Linkage**:

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x,y)$$

- **Complete Linkage**:

$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x,y)$$

- **Average Linkage**:

$$d(C_i, C_j) = \frac{1}{\mid C_i \mid\mid C_j \mid} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$$

- **Ward's Method** (phổ biến nhất):

$$\Delta SSE = \sum_{x \in C_i \cup C_j} \| x - \mu_{ij} \|^2 - \sum_{x \in C_i} \| x - \mu_i \|^2 - \sum_{x \in C_j} \| x - \mu_j \|^2$$

Trong đó:

- μ_i, μ_j: tâm cụm
- μ_{ij}: tâm cụm sau khi gộp

Ward's Method gộp hai cụm sao cho mức tăng tổng sai số bình phương (SSE) là nhỏ nhất.

C. Thuật toán Khai phá Luật Kết hợp (Association Rule Mining)

7. Apriori

Apriori tìm các tập mục thường xuyên dựa trên nguyên lý:

"Mọi tập con của một tập mục thường xuyên cũng phải thường xuyên".

Độ hỗ trợ (Support):

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

Độ tin cậy (Confidence):

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

Độ nâng (Lift):

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

8. FP-Growth

FP-Growth cải tiến Apriori bằng cách xây dựng cây FP-Tree để nén dữ liệu, từ đó khai phá các tập mục thường xuyên mà không cần sinh tập ứng viên.

FP-Growth không có công thức riêng biệt mà dựa trên:

Cấu trúc FP-Tree.

Khai phá đệ quy các tập mục thường xuyên từ cây.

Thuật toán này cho tốc độ nhanh hơn Apriori, đặc biệt phù hợp với dữ liệu lớn.

Chương 3: Dữ liệu & Phương pháp đề xuất

3.1. Xác định vấn đề nghiên cứu

Trong nghiên cứu này, bài toán được xác định là bài toán phân loại nhị phân, nhằm dự đoán mức độ hài lòng của khách hàng (Hài lòng / Không hài lòng) dựa trên các thông tin liên quan đến đơn hàng và dịch vụ trong bộ dữ liệu Brazilian E-Commerce (Olist).

Cụ thể, nghiên cứu tập trung phân tích mức độ ảnh hưởng của các yếu tố sau đến sự hài lòng của khách hàng:

Giá trị đơn hàng

Thời gian và trạng thái giao hàng

Phí vận chuyển

Hình thức thanh toán

Mức độ chi tiết của mô tả sản phẩm

Trạng thái hoàn tất đơn hàng

3.2. Mô tả dữ liệu

Nguồn dữ liệu được sử dụng trong nghiên cứu là Brazilian E-Commerce Public Dataset by Olist, được công bố công khai trên Kaggle. Bộ dữ liệu bao gồm thông tin về đơn hàng, khách hàng, sản phẩm, thanh toán, giao hàng và đánh giá của khách hàng.

Từ bộ dữ liệu gốc, nghiên cứu lựa chọn và kết hợp các bảng dữ liệu cần thiết để xây dựng tập dữ liệu cuối cùng phục vụ cho bài toán dự đoán sự hài lòng của khách hàng.

Biến mục tiêu (target variable) là điểm đánh giá của khách hàng, được chuyển đổi thành nhãn phân loại: Hài lòng và Không hài lòng

Bộ dữ liệu Olist bao gồm nhiều bảng có kích thước khác nhau, phản ánh đầy đủ các thành phần trong hệ thống thương mại điện tử như khách hàng, đơn hàng, sản phẩm, người bán và thanh toán. Cụ thể:

- customers: 99,441 dòng và 5 cột, chứa thông tin cơ bản về khách hàng.

- geolocation: 1,000,163 dòng và 5 cột, là bảng có kích thước lớn nhất, lưu trữ thông tin vị trí địa lý.

- order_items: 112,650 dòng và 7 cột, mô tả chi tiết các sản phẩm trong từng đơn hàng.

- order_payments: 103,886 dòng và 5 cột, chứa thông tin về hình thức và giá trị thanh toán.

- order_reviews: 99,224 dòng và 7 cột, ghi nhận đánh giá và phản hồi của khách hàng.

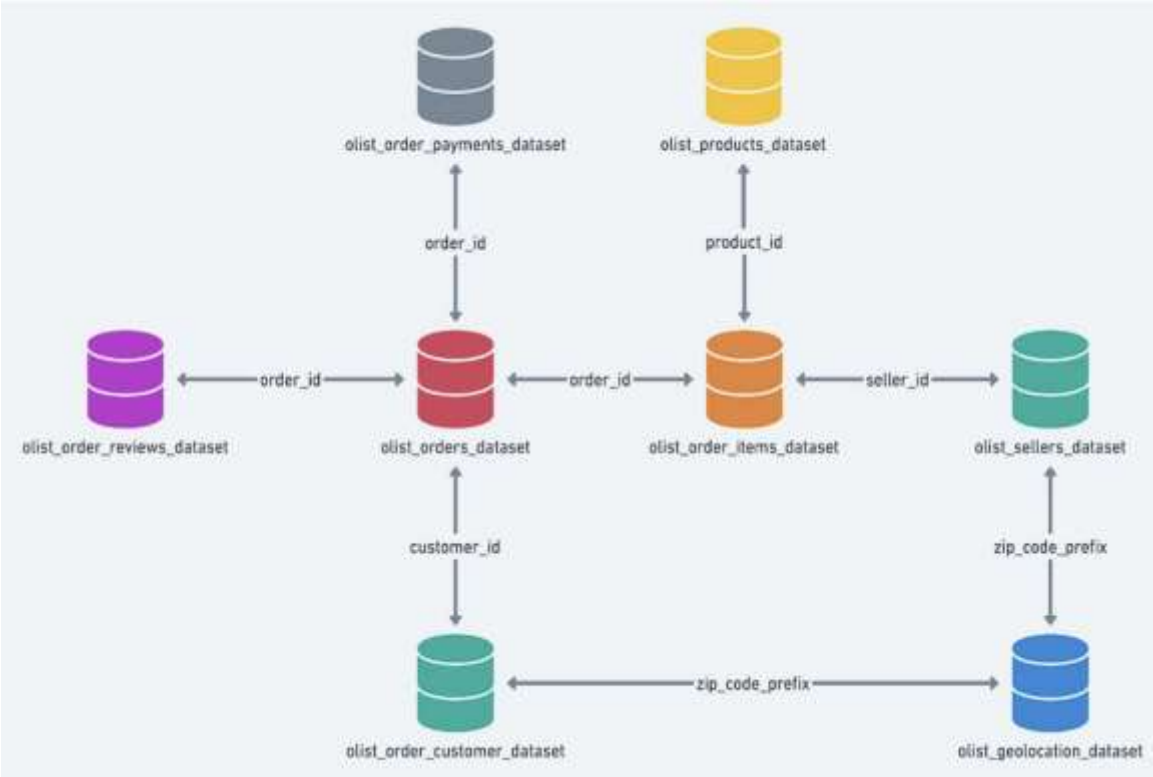
- orders: 99,441 dòng và 8 cột, là bảng trung tâm liên kết hầu hết các bảng còn lại.

- products: 32,951 dòng và 9 cột, cung cấp thông tin chi tiết về sản phẩm.

- sellers: 3,095 dòng và 4 cột, lưu trữ dữ liệu về người bán.

- category_translation: 71 dòng và 2 cột, dùng để dịch tên danh mục sản phẩm sang tiếng Anh.

Nhìn chung, dữ liệu có cấu trúc rõ ràng, số lượng quan sát lớn và phù hợp cho các bài toán phân tích dữ liệu, trực quan hóa và xây dựng mô hình học máy cơ



bản. Các bảng được phân tách hợp lý theo nghiệp vụ, giúp việc kết hợp (merge) và phân tích trở nên thuận tiện và hiệu quả.

Bộ dữ liệu được sử dụng trong nghiên cứu này được kết nối với nhau bằng các khóa chính phù hợp để đảm bảo liên kết chính xác và giảm thiểu sự trùng lặp. Mỗi bộ dữ liệu đóng góp thông tin độc đáo và thiết yếu, tạo thành một cái nhìn toàn diện khi được kết hợp.

3.3. Tiền xử lý dữ liệu

Trước khi áp dụng các mô hình học máy, dữ liệu được tiền xử lý qua các bước sau:

Làm sạch dữ liệu: loại bỏ hoặc xử lý các bản ghi bị thiếu thông tin quan trọng.

Xử lý giá trị thiếu (missing values): thay thế hoặc loại bỏ các giá trị không hợp lệ.

Chuẩn hóa dữ liệu số: đảm bảo các đặc trưng có cùng thang đo nhằm cải thiện hiệu quả của các mô hình.

Mã hóa dữ liệu phân loại: chuyển các biến dạng danh mục sang dạng số để phù hợp với thuật toán học máy.

Xử lý mất cân bằng dữ liệu: áp dụng kỹ thuật SMOTE để tăng số lượng mẫu của lớp thiểu số (Không hài lòng), giúp mô hình học tốt hơn và giảm thiên lệch.

Phần tiếp theo sẽ trình bày chi tiết từng thuộc tính dữ liệu, nêu rõ các loại khác nhau và phân loại chúng thành các danh mục văn bản và số. Để xác minh tính chính xác của dữ liệu, các công cụ Python sẽ được sử dụng để kiểm tra kỹ lưỡng tập dữ liệu. Tổng số lượt gửi cho mỗi thuộc tính và kiểu dữ liệu của chúng, giúp phát hiện bất kỳ giá trị nào bị thiếu. Với tổng cộng 115.609 bản ghi, một số cột chứa khoảng trống, đòi hỏi phải sử dụng thêm các tập lệnh Python để xác định chính xác số lượng giá trị bị thiếu cho mỗi thuộc tính.

```
RangeIndex: 115609 entries, 0 to 115608
Data columns (total 40 columns):
 #   Column                         Non-Null Count   Dtype
---  ------                         --------------   -----
 0   customer_id                    115609 non-null  object
 1   customer_unique_id             115609 non-null  object
 2   customer_zip_code_prefix       115609 non-null  int64
 3   customer_city                  115609 non-null  object
 4   customer_state                 115609 non-null  object
 5   order_id                       115609 non-null  object
 6   order_status                   115609 non-null  object
 7   order_purchase_timestamp       115609 non-null  object
 8   order_approved_at              115595 non-null  object
 9   order_delivered_carrier_date   114414 non-null  object
 10  order_delivered_customer_date  113209 non-null  object
 11  order_estimated_delivery_date  115609 non-null  object
 12  review_id                      115609 non-null  object
 13  review_score                   115609 non-null  int64
 14  review_comment_title           13801 non-null   object
 15  review_comment_message         48906 non-null   object
 16  review_creation_date           115609 non-null  object
 17  review_answer_timestamp        115609 non-null  object
 18  order_item_id                  115609 non-null  int64
 19  product_id                     115609 non-null  object
...
 38  seller_state                   115609 non-null  object
 39  product_category_name_english  115609 non-null  object
dtypes: float64(10), int64(6), object(24)
```

Xác định các cột có dữ liệu không đầy đủ:

- approved_order_date
- carrier_delivery_date
- customer_delivery_date
- review_title
- review_message

```
Missing Data Counts:
 customer_id                          0
customer_unique_id                    0
customer_zip_code_prefix              0
customer_city                         0
customer_state                        0
order_id                              0
order_status                          0
order_purchase_timestamp              0
order_approved_at                    14
order_delivered_carrier_date       1195
order_delivered_customer_date      2400
order_estimated_delivery_date         0
review_id                             0
review_score                          0
review_comment_title             101808
review_comment_message            66703
review_creation_date                  0
review_answer_timestamp               0
order_item_id                         0
product_id                            0
seller_id                             0
shipping_limit_date                   0
price                                 0
freight_value                         0
...
seller_city                           0
seller_state                          0
product_category_name_english         0
dtype: int64
```

Hình 1: Missing Values

```
Missing Data Counts:
 customer_id                     0
customer_unique_id               0
customer_zip_code_prefix         0
customer_city                    0
customer_state                   0
order_id                         0
order_status                     0
order_purchase_timestamp         0
order_approved_at                0
order_delivered_carrier_date     0
order_delivered_customer_date    0
order_estimated_delivery_date    0
review_id                        0
review_score                     0
review_comment_title             0
review_comment_message           0
review_creation_date             0
review_answer_timestamp          0
order_item_id                    0
product_id                       0
seller_id                        0
shipping_limit_date              0
price                            0
freight_value                    0
...
seller_city                      0
seller_state                     0
product_category_name_english    0
dtype: int64
```

Hình 2: Updated Missing Values

Sau khi kiểm tra tổng quan tập dữ liệu, nhóm nhận thấy có một số trường thông tin bị thiếu. Để đảm bảo tính chính xác và đồng nhất cho mô hình phân tích, các bước xử lý sau đã được thực hiện:

1. Loại bỏ các bản ghi có dữ liệu khuyết thiếu không đáng kể

Đối với các cột liên quan đến quy trình vận hành đơn hàng, số lượng dữ liệu thiếu chiếm tỷ lệ rất nhỏ so với tổng thể (hơn 100.000 dòng):

- **order_approved_at**: Thiếu 14 bản ghi.
- **order_delivered_carrier_date**: Thiếu 1.195 bản ghi.
- **order_delivered_customer_date**: Thiếu 2.400 bản ghi.

Lý do xử lý: Vì tỷ lệ thiếu sót này chỉ chiếm khoảng **1% - 2%** tổng số dữ liệu và đây là các đơn hàng có thể chưa hoàn tất hoặc bị hủy. Việc loại bỏ các dòng này không làm thay đổi đặc tính của tập mẫu nhưng giúp làm sạch dữ liệu để tính toán chính xác thời gian giao hàng.

2. Xử lí giá trị bị thiếu

Để hiểu rõ hơn về giá trị của từng cột, chúng tôi cung cấp bản tóm tắt ngắn gọn về các đặc điểm số.

```
df.describe()
```

	customer_zip_code_prefix	review_score	order_item_id	price	freight_value	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	pro
count	115609.000000	115609.000000	115609.000000	115609.000000	115609.000000	115609.000000	115609.000000	115609.000000	115608.000000	
mean	35061.537597	4.034409	1.194535	120.619850	20.056880	48.766541	785.808198	2.205373	2113.907697	
std	29841.671732	1.385584	0.685926	182.653476	15.836184	10.034187	652.418619	1.717771	3781.754895	
min	1003.000000	1.000000	1.000000	0.850000	0.000000	5.000000	4.000000	1.000000	0.000000	
25%	11310.000000	4.000000	1.000000	39.900000	13.080000	42.000000	346.000000	1.000000	300.000000	
50%	24241.000000	5.000000	1.000000	74.900000	16.320000	52.000000	600.000000	1.000000	700.000000	
75%	58745.000000	5.000000	1.000000	134.900000	21.210000	57.000000	983.000000	3.000000	1800.000000	
max	99980.000000	5.000000	21.000000	6735.000000	409.680000	76.000000	3992.000000	20.000000	40425.000000	

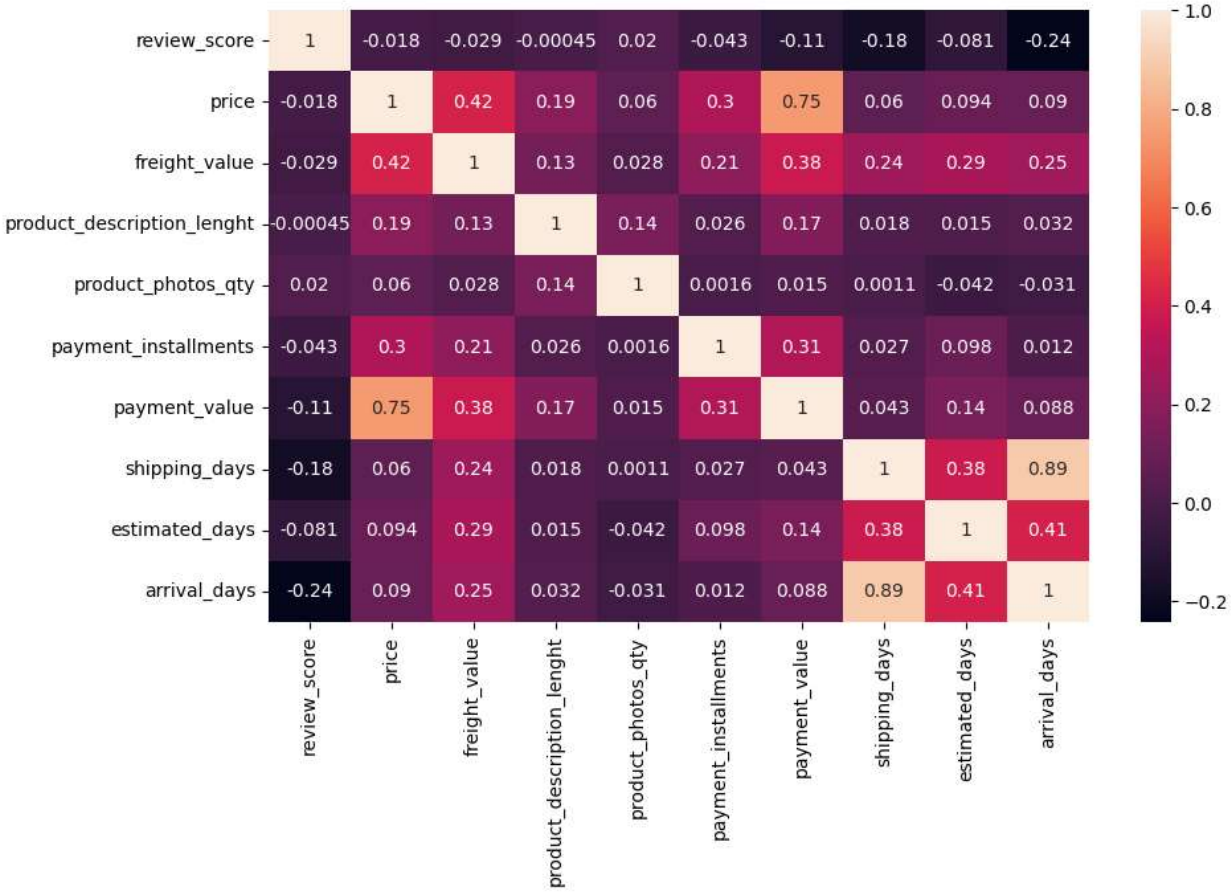
Hình 3: Phân tích thống kê các thuộc tính số

Trình bày tóm tắt thống kê chi tiết cho từng cột, bao gồm số lượng, trung bình, độ lệch chuẩn, giá trị tối thiểu và tối đa, cũng như các phân vị thứ 25, 50 và 75.

Khách hàng đánh giá từ 1 đến 5 sao, với 1 là thấp nhất và 5 là cao nhất. Điểm đánh giá trung bình là 4.03, cho thấy phản hồi tích cực nói chung.

Giá sản phẩm dao động từ 0,85 đến 6.350, phản ánh sự đa dạng của các mặt hàng được cung cấp bởi người bán trên Olist, chẳng hạn như xe cộ, thiết bị điện tử, sản phẩm làm đẹp và quần áo. Chi tiết sản phẩm rất đầy đủ, với độ dài tên sản phẩm trung bình là 50 ký tự và độ dài mô tả khoảng 790 ký tự. Tuy nhiên, hình ảnh hiển thị bị hạn chế, chỉ trung bình 2 ảnh cho mỗi sản phẩm. Nhiều hình ảnh hơn sẽ giúp khách hàng đưa ra quyết định mua hàng sáng suốt hơn.

Cuối cùng, chúng ta sẽ đánh giá mối quan hệ giữa các đặc điểm số. Hình hiển thị bản đồ nhiệt minh họa mối tương quan giữa các đặc điểm này.



Hình 4: Mô hình Heatmap quan hệ của cột

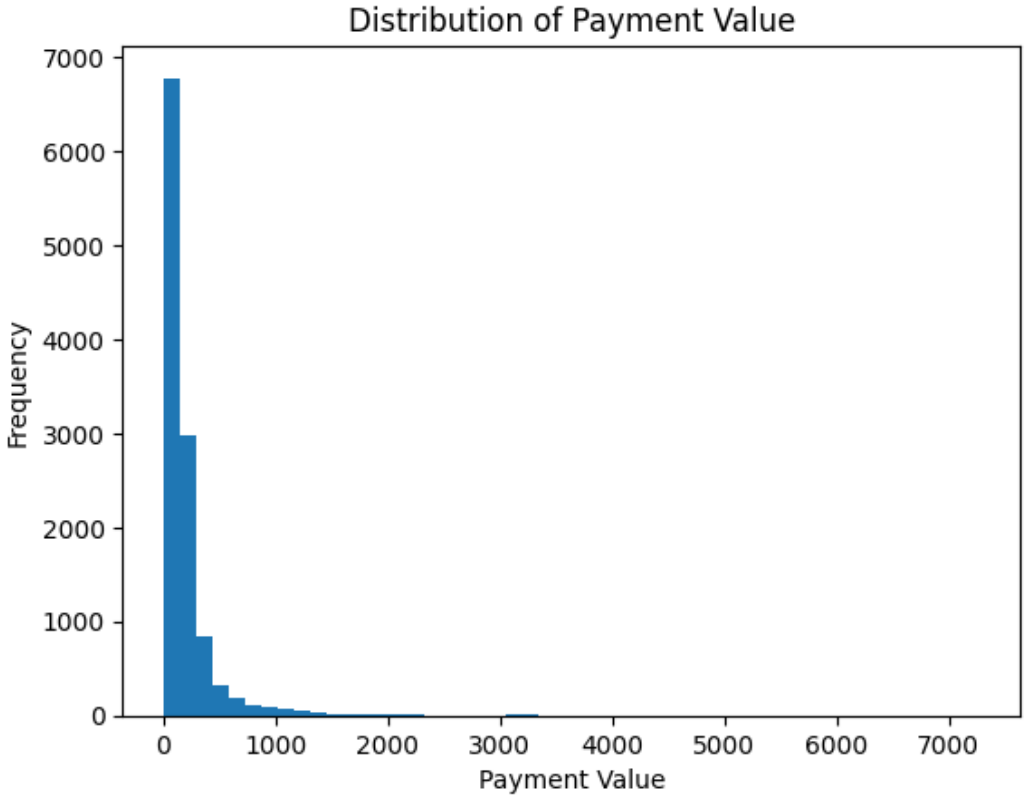
Biểu đồ nhiệt minh họa mối liên hệ giữa các thuộc tính khác nhau. Màu sáng hơn biểu thị mối tương quan mạnh hơn. Đáng chú ý, giá trị cước vận chuyển và giá trị thanh toán có mối tương quan đáng kể, cũng như số ngày đến và số ngày vận chuyển. Những mối tương quan này là hợp lý; ví dụ, giá trị cước vận chuyển cao hơn thường dẫn đến tổng giá trị thanh toán cao hơn, và số ngày đến và số ngày vận chuyển có liên hệ chặt chẽ, phản ánh hiệu quả vận chuyển. Hiểu được những mối quan hệ này giúp điều chỉnh dịch vụ và chiến lược tiếp thị cho các phân khúc khách hàng khác nhau, nâng cao sự hài lòng tổng thể và hiệu quả hoạt động.

3.4. Phương pháp đề xuất

Dựa trên đặc điểm dữ liệu và mục tiêu nghiên cứu, các phương pháp sau được đề xuất và áp dụng:

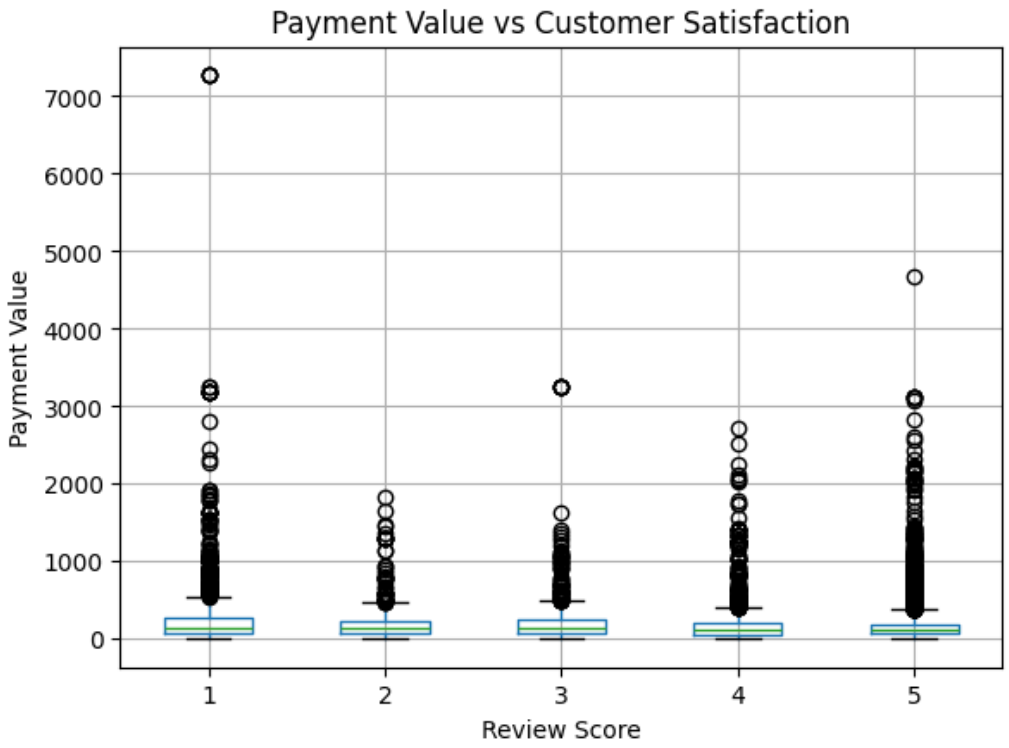
Giá trị đơn hàng có ảnh hưởng đến mức độ hài lòng của khách hàng không?

Dựa trên biến payment_value, phân tích được thực hiện nhằm kiểm tra liệu các đơn hàng có giá trị cao có xu hướng đưa ra đánh giá khắt khe hơn hay không, qua đó đánh giá mức độ ảnh hưởng của giá trị đơn hàng đến sự hài lòng và hành vi đánh giá của khách hàng.



Hình 5: Biểu đồ mô tả giá trị thanh toán

Quan sát biểu đồ cho thấy phân bố payment_value bị lệch phải, phản ánh thực tế rằng phần lớn đơn hàng có giá trị thấp trong khi chỉ có một số ít đơn hàng có giá trị cao.



Hình 6: Mối quan hệ giữa payment_value và mức độ hài lòng

So sánh giá trị thanh toán cho thấy sự khác biệt giữa hai nhóm khách hàng hài lòng và không hài lòng, qua đó giúp đánh giá liệu mức payment_value có ảnh hưởng đến mức độ hài lòng và xu hướng đánh giá của khách hàng hay không.

```
review_score
1    261.078502
2    210.084793
3    220.665805
4    190.144201
5    173.565430
Name: payment_value, dtype: float64
```

Hình 7: Phân tích trung bình theo mức độ hài lòng

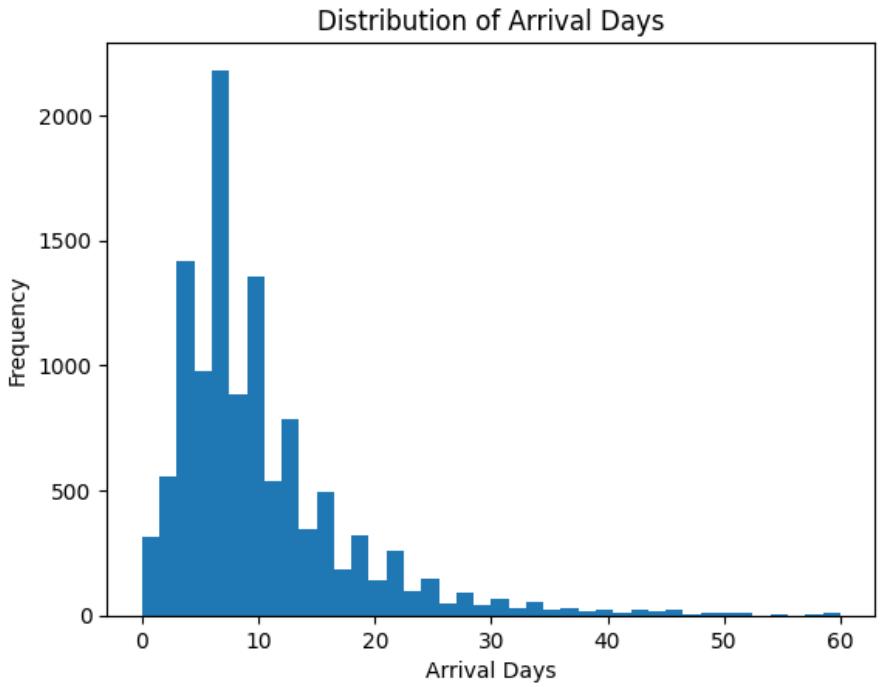
Khách hàng có đơn hàng giá trị cao thường có mức kỳ vọng lớn hơn về chất lượng dịch vụ; do đó, họ dễ cảm thấy không hài lòng hơn nếu trải nghiệm giao hàng hoặc dịch vụ không đáp ứng được mong đợi.

Nhận xét:

Giá trị thanh toán có phân bố lệch phải, cho thấy phần lớn đơn hàng có giá trị thấp. Tuy nhiên, các đơn hàng có giá trị cao có xu hướng đánh giá khắt khe hơn, làm cho payment_value trở thành đặc trưng có ảnh hưởng mạnh nhất đến mức độ hài lòng của khách hàng.

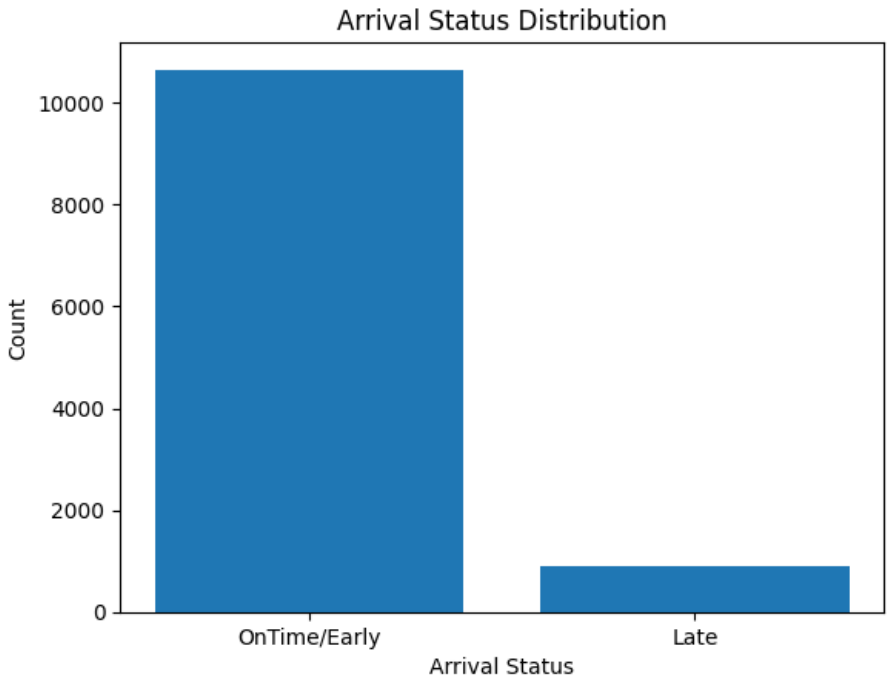
Thời gian và trạng thái giao hàng ảnh hưởng như thế nào đến mức độ hài lòng?

Dựa trên các biến arrival_days, arrival_delivery_rate, arrival_status và shipping_delivery_rate, phân tích được thực hiện nhằm xác định vai trò của việc giao hàng đúng hạn và nhanh chóng đối với trải nghiệm cũng như mức độ hài lòng của khách hàng, qua đó làm rõ mức độ ảnh hưởng của các yếu tố liên quan đến thời gian và chất lượng giao hàng trong đánh giá dịch vụ.



Hình 8: Phân bố thời gian giao hàng

Kết quả phân tích cho thấy giao hàng càng nhanh thì mức kỳ vọng và khả năng hài lòng của khách hàng càng cao. Đồng thời, phân bố dữ liệu lệch phải phản ánh rằng vẫn tồn tại một số đơn hàng có thời gian giao rất trễ, đây là yếu tố tiềm ẩn làm giảm trải nghiệm và mức độ hài lòng của khách hàng.



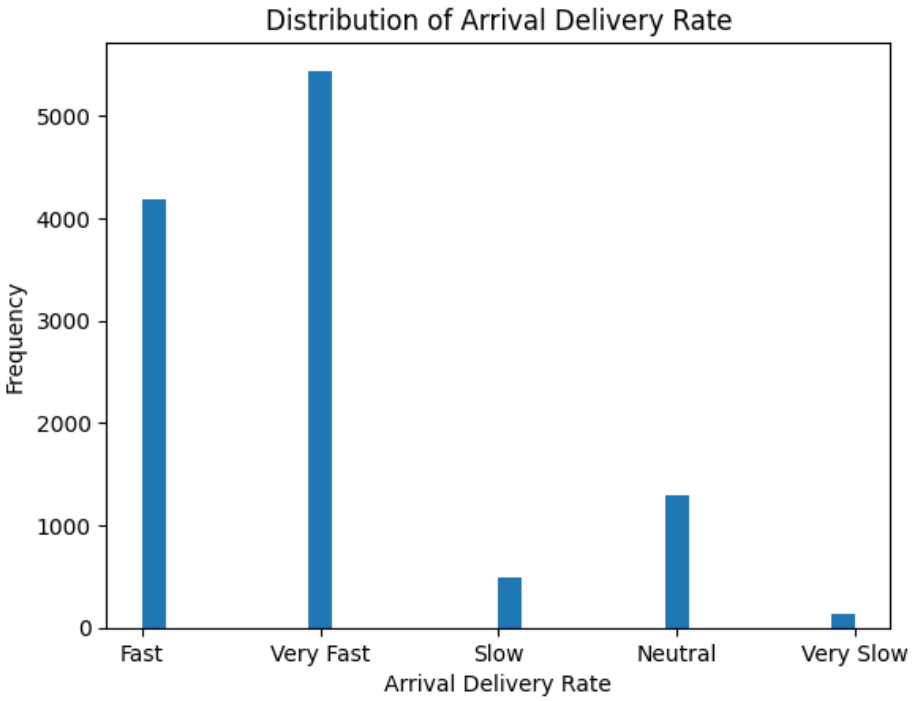
Hình 9: Phân tích trạng thái giao hàng đúng hạn

So sánh số lượng đơn hàng giao đúng hạn và giao trễ cho thấy mặc dù đây chỉ là một biến nhị phân, nhưng lại có ảnh hưởng rất lớn đến mức độ hài lòng của khách hàng, trong đó các đơn giao trễ thường gắn liền với đánh giá và trải nghiệm tiêu cực hơn.

review_score	1	2	3	4	5
arrival_status					
Late	0.470258	0.079686	0.104377	0.113356	0.232323
OnTime/Early	0.161206	0.047789	0.078115	0.152756	0.560135

Hình 10: Mối quan hệ giữa trạng thái giao hàng và mức độ hài lòng

Kết quả cho thấy tỷ lệ khách hàng hài lòng cao hơn rõ rệt đối với các đơn hàng được giao đúng hạn, khẳng định vai trò quan trọng của yếu tố thời gian giao hàng trong việc hình thành trải nghiệm và mức độ hài lòng của khách hàng.



Hình 11: Phân tích hiệu suất giao hàng

Biểu đồ cho thấy phần lớn đơn hàng được giao với tốc độ Very Fast và Fast, chứng tỏ hệ thống giao hàng hoạt động hiệu quả và ổn định. Tỷ lệ giao hàng ở mức Neutral khá thấp, trong khi các trường hợp Slow và Very Slow rất hiếm, cho thấy tình trạng giao hàng chậm không phổ biến. Nhìn chung, chất lượng dịch vụ giao hàng được đánh giá là tốt.

Hình 12: Phân tích shipping_delivery_rate

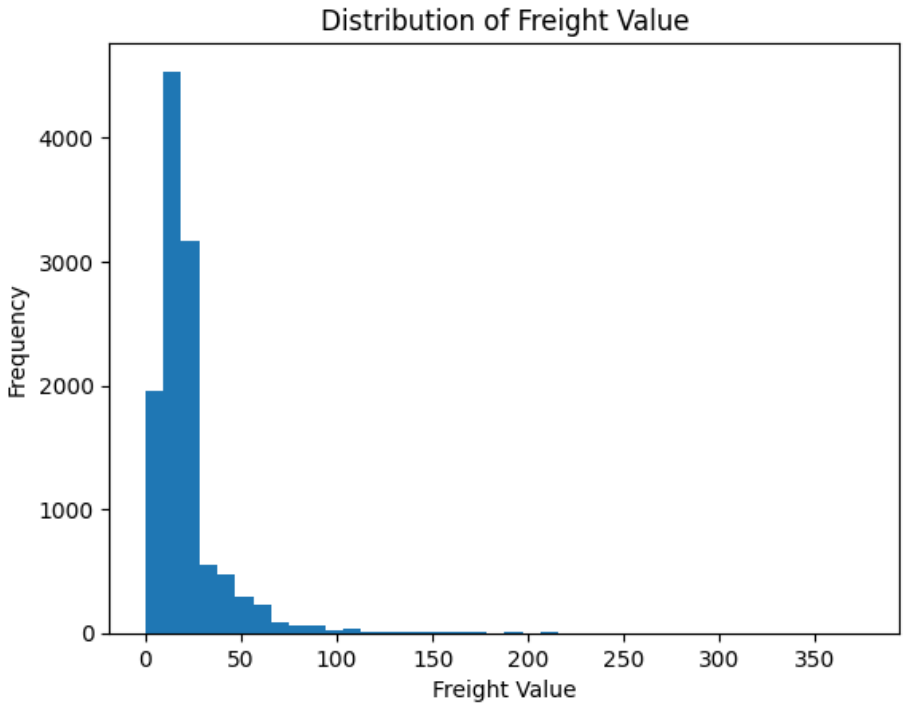
Hiệu suất vận chuyển ổn định góp phần mang lại trải nghiệm tích cực cho khách hàng. Mặc dù mức độ ảnh hưởng của yếu tố này thấp hơn so với arrival_delivery_rate, nhưng nó vẫn đóng vai trò quan trọng trong việc duy trì sự hài lòng và đánh giá chung của khách hàng.

Nhận xét :

Nhóm đặc trưng liên quan đến giao hàng cho thấy mối quan hệ rõ rệt với mức độ hài lòng của khách hàng. Các đơn hàng giao đúng hạn, thời gian giao ngắn và hiệu suất giao hàng cao có tỷ lệ hài lòng cao hơn đáng kể. Điều này lý giải vì sao các biến giao hàng chiếm tỷ trọng lớn trong Feature Importance của mô hình Random Forest.

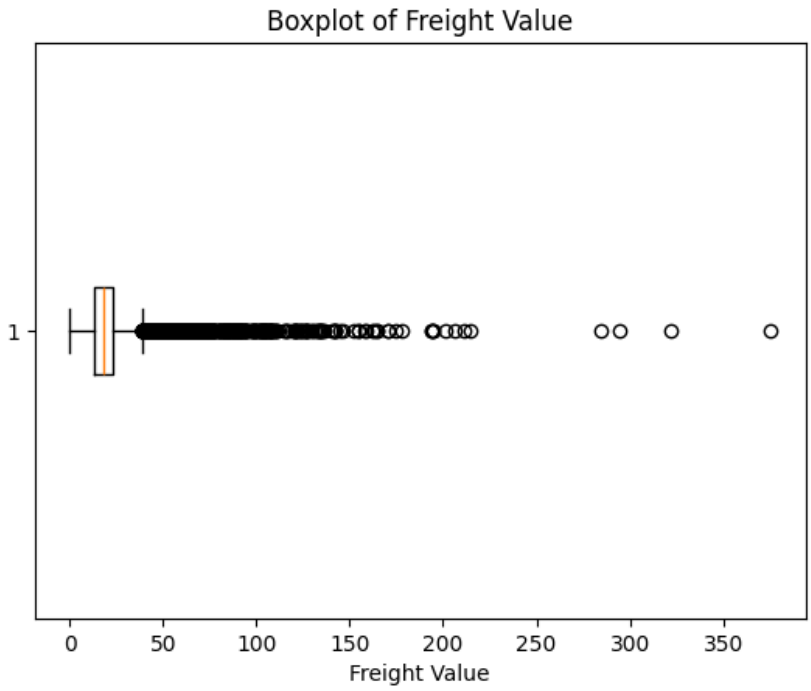
Phí vận chuyển có làm giảm mức độ hài lòng của khách hàng hay không ?

Dựa trên biến freight_value, phân tích được thực hiện nhằm so sánh mức độ hài lòng của khách hàng giữa các đơn hàng có phí vận chuyển cao và phí vận chuyển thấp, qua đó đánh giá tác động của chi phí vận chuyển đến trải nghiệm và đánh giá dịch vụ của khách hàng.



Hình 13: Phân bố phí vận chuyển

Phân bố freight_value thường có dạng lệch phải, cho thấy đa số đơn hàng có phí vận chuyển ở mức thấp trong khi vẫn tồn tại một số đơn có phí vận chuyển rất cao. Những đơn hàng này dễ làm khách hàng cảm thấy không hài lòng, đặc biệt khi chi phí vận chuyển không tương xứng với giá trị hoặc chất lượng dịch vụ nhận được.



Hình 14: Boxplot phát hiện phí vận chuyển bất thường

Các giá trị ngoại lai (outliers) phản ánh những đơn hàng có chi phí vận chuyển cao bất thường, và đây là yếu tố tác động tiêu cực đến trải nghiệm cũng như mức độ hài lòng của khách hàng, đặc biệt khi chi phí phát sinh không phù hợp với kỳ vọng ban đầu.

Hình 15: Mối quan hệ giữa freight_value và mức độ hài lòng

Nhóm khách hàng không hài lòng thường có mức phí vận chuyển cao hơn và độ phân tán lớn hơn so với nhóm hài lòng, cho thấy chi phí vận chuyển không chỉ cao mà còn biến động mạnh là một trong những yếu tố góp phần làm giảm trải nghiệm và mức độ hài lòng của khách hàng.

```
review_score
1    22.518717
2    23.843776
3    21.868227
4    22.915041
5    21.328202
Name: freight_value, dtype: float64
```

Hình 16: Phân tích trung bình phí vận chuyển theo mức độ hài lòng

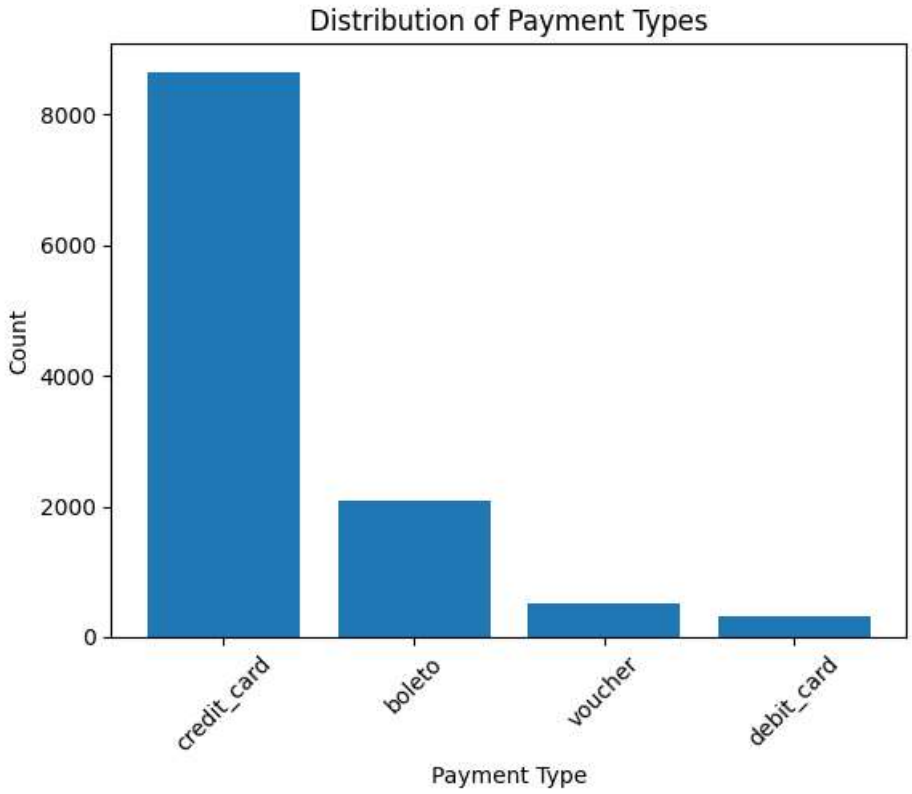
Kết quả phân tích cho thấy phí vận chuyển cao có xu hướng làm giảm mức độ hài lòng của khách hàng, đặc biệt khi chi phí này không tương xứng với giá trị đơn hàng hoặc chất lượng dịch vụ giao nhận.

Nhận xét:

Phí vận chuyển có phân bố lệch phải với một số giá trị cao bất thường. Phân tích cho thấy các đơn hàng có phí vận chuyển cao có xu hướng nhận đánh giá thấp hơn. Tuy nhiên, mức độ ảnh hưởng của freight_value vẫn thấp hơn so với các yếu tố liên quan đến thời gian và trạng thái giao hàng.

Hình thức thanh toán có ảnh hưởng đến sự hài lòng của khách hàng ?

Dựa trên biến payment_type, phân tích được thực hiện nhằm so sánh mức độ hài lòng của khách hàng giữa các hình thức thanh toán khác nhau như thẻ tín dụng, thẻ ghi nợ, voucher và boleto, từ đó đánh giá ảnh hưởng của phương thức thanh toán đến trải nghiệm và mức độ hài lòng của khách hàng.



Hình 17: Biểu đồ phân bố hình thức thanh toán

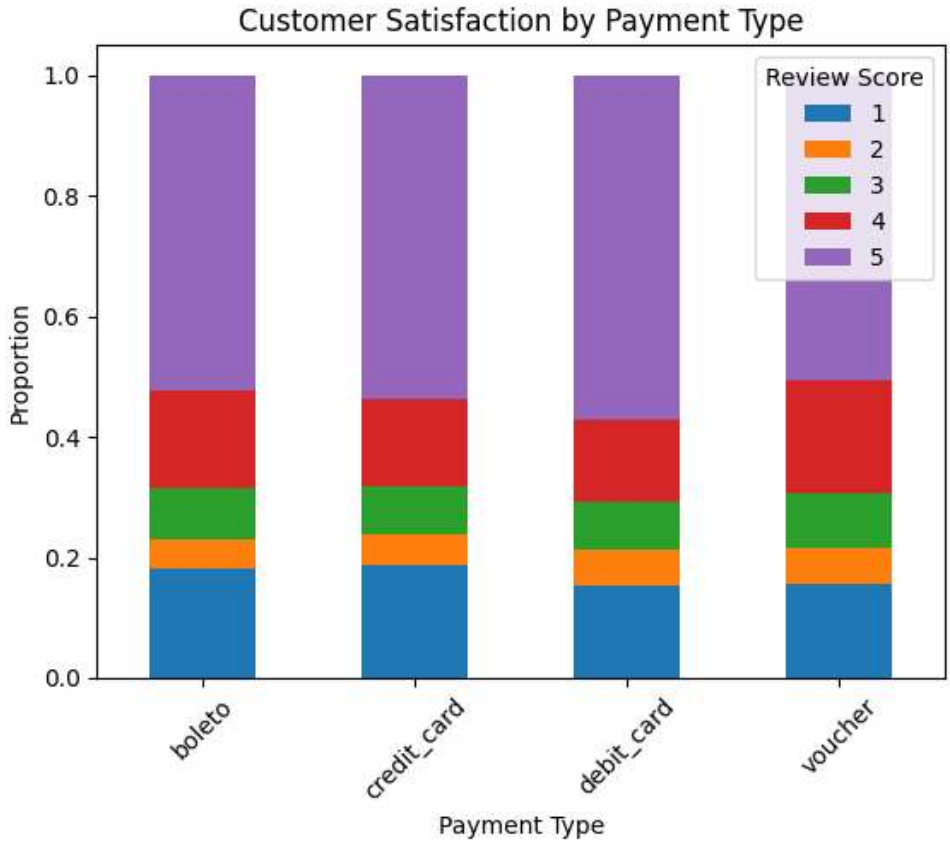
Kết quả cho thấy khách hàng sử dụng thẻ tín dụng chiếm tỷ lệ lớn nhất trong các hình thức thanh toán, phản ánh mức độ phổ biến và sự tiện lợi của phương thức này trong quá trình mua sắm.

review_score	1	2	3	4	5
payment_type					
boleto	0.181643	0.049015	0.085536	0.160019	0.523787
credit_card	0.188620	0.049728	0.078293	0.145600	0.537759
debit_card	0.154341	0.057878	0.080386	0.135048	0.572347
voucher	0.157058	0.059642	0.089463	0.186879	0.506958

Hình 18: Mối quan hệ giữa payment_type và mức độ hài lòng

Quan sát nhóm khách hàng sử dụng credit_card cho thấy tỷ lệ hài lòng tương đối ổn định và mức độ biến động thấp hơn so với các hình thức thanh toán như

voucher hoặc debit, cho thấy trải nghiệm thanh toán bằng thẻ tín dụng nhìn chung nhất quán và đáng tin cậy hơn.



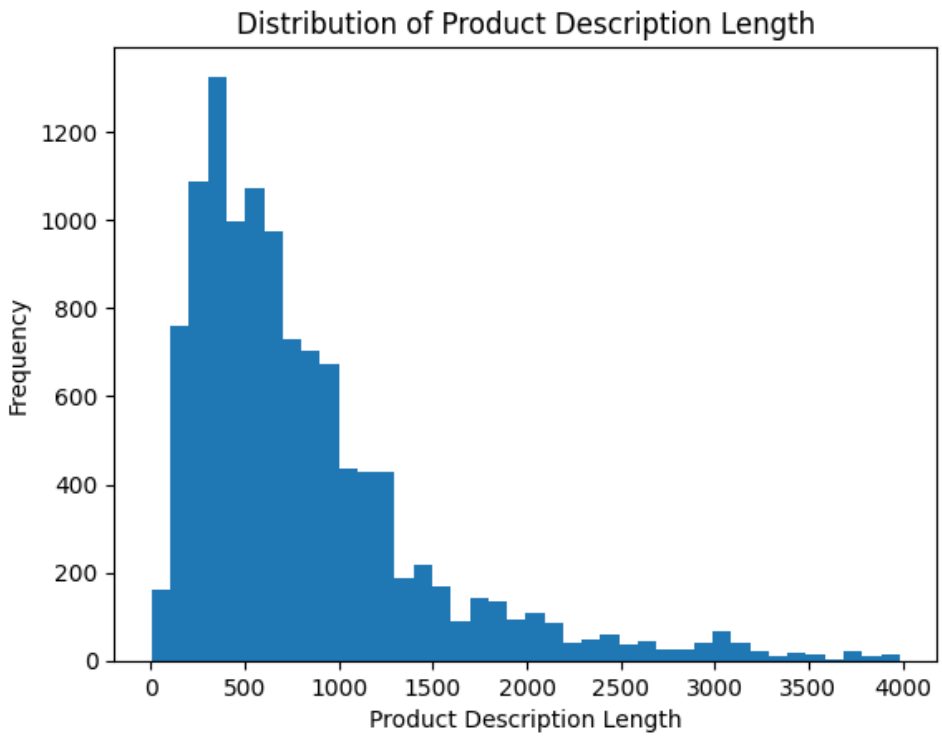
Hình 19: So sánh trực quan mức độ hài lòng theo hình thức thanh toán

Nhận xét:

Hình thức thanh toán có ảnh hưởng đến mức độ hài lòng của khách hàng. Các hình thức thanh toán điện tử như thẻ tín dụng và thẻ ghi nợ có tỷ lệ hài lòng cao và ổn định hơn so với voucher và boleto. Tuy nhiên, mức độ ảnh hưởng của hình thức thanh toán vẫn thấp hơn các yếu tố liên quan đến giao hàng.

Mức độ chi tiết của mô tả sản phẩm có giúp cải thiện sự hài lòng của khách hàng ?

Dựa trên biến product_description_lenght, phân tích được thực hiện nhằm kiểm tra liệu việc mô tả sản phẩm dài và chi tiết có giúp giảm sự sai lệch trong kỳ vọng của khách hàng hay không, qua đó đánh giá vai trò của thông tin sản phẩm đối với mức độ hài lòng sau mua.



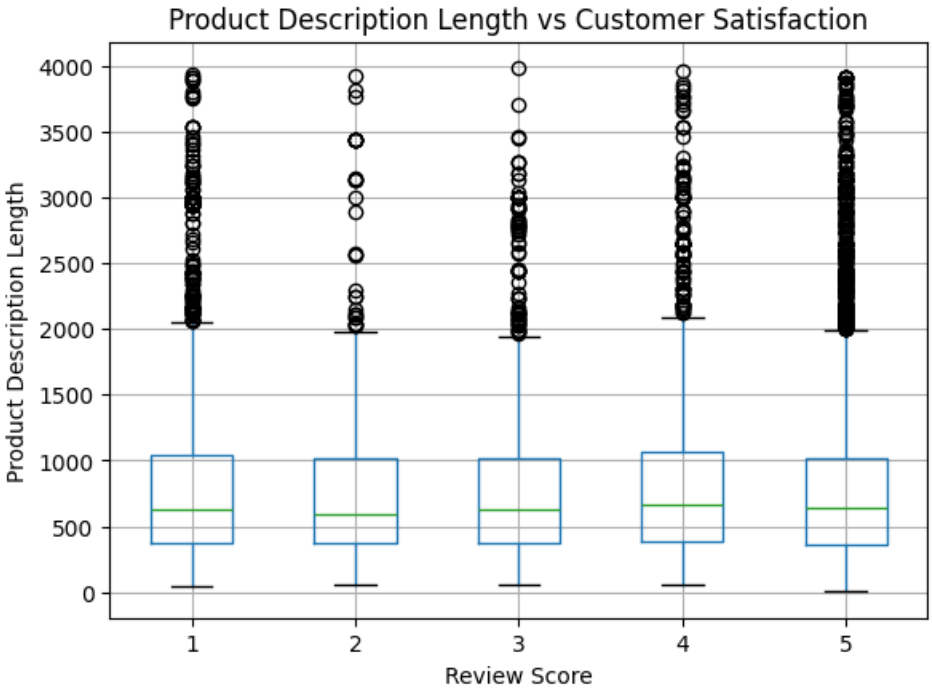
Hình 20: Phân bố độ dài mô tả sản phẩm

Quan sát dữ liệu cho thấy phần lớn sản phẩm có mô tả tương đối ngắn, trong khi chỉ một số ít sản phẩm có mô tả rất dài, phản ánh việc cung cấp thông tin chi tiết và đầy đủ hơn về đặc điểm của sản phẩm.

```
review_score
1    819.334738
2    772.725862
3    821.184865
4    833.291088
5    811.604730
Name: product_description_lenght, dtype: float64
```

Hình 21: So sánh trung bình theo mức độ hài lòng

Kết quả phân tích cho thấy nhóm khách hàng hài lòng thường gắn với các sản phẩm có mô tả dài hơn mức trung bình, cho thấy thông tin sản phẩm chi tiết giúp khách hàng hình dung rõ hơn, giảm kỳ vọng sai lệch và nâng cao mức độ hài lòng.



Hình 22: Mối quan hệ giữa độ dài mô tả và mức độ hài lòng

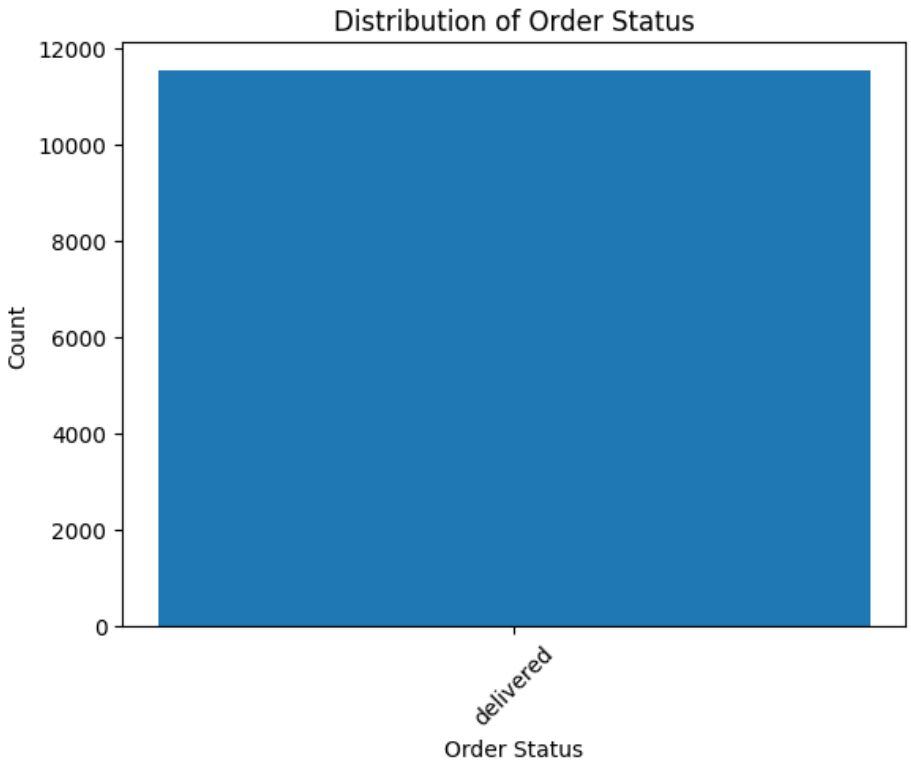
Nhóm khách hàng hài lòng có phân bố độ dài mô tả sản phẩm cao hơn và độ phân tán thấp hơn, cho thấy thông tin được trình bày rõ ràng và nhất quán hơn, từ đó giúp khách hàng hình thành kỳ vọng chính xác và ổn định hơn về sản phẩm.

Nhận xét:

Độ dài mô tả sản phẩm có ảnh hưởng tích cực đến mức độ hài lòng của khách hàng. Các sản phẩm có mô tả chi tiết giúp giảm kỳ vọng sai lệch, từ đó nâng cao trải nghiệm mua sắm. Nhóm khách hàng hài lòng có xu hướng mua các sản phẩm có mô tả dài hơn.

Trạng thái hoàn tất đơn hàng ảnh hưởng như thế nào đến đánh giá của khách hàng ?

Dựa trên biến order_status, phân tích được thực hiện nhằm so sánh mức độ hài lòng của khách hàng giữa các đơn hàng đã được giao thành công và các đơn hàng chưa hoàn tất, qua đó đánh giá tác động của trạng thái đơn hàng đến trải nghiệm và mức độ hài lòng của khách hàng.



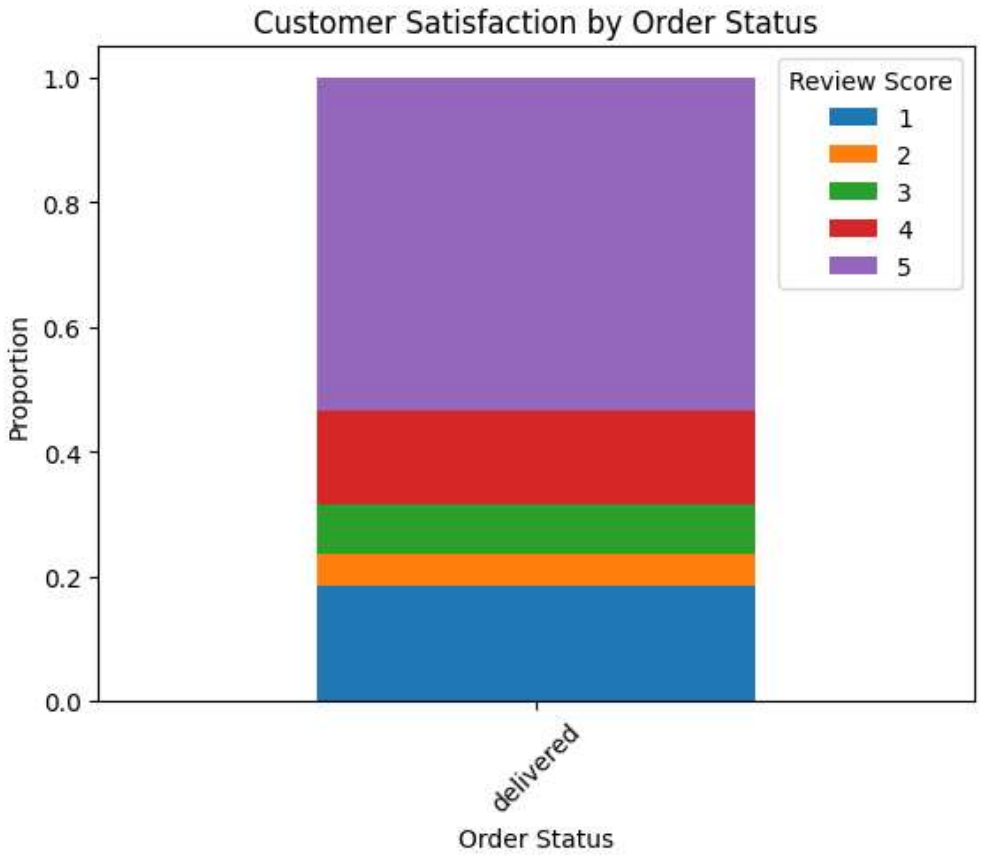
Hình 23: Phân bố trạng thái đơn hàng

Kết quả cho thấy trạng thái delivered chiếm tỷ lệ lớn nhất trong dữ liệu, trong khi các trạng thái còn lại phản ánh những đơn hàng chưa hoàn tất, tiềm ẩn nhiều rủi ro ảnh hưởng tiêu cực đến trải nghiệm và mức độ hài lòng của khách hàng.

review_score	1	2	3	4	5
order_status					
delivered	0.185063	0.050251	0.080142	0.149714	0.534829

Hình 24: Mối quan hệ giữa trạng thái đơn hàng và mức độ hài lòng

Kết quả phân tích cho thấy các đơn hàng được giao đúng hạn có tỷ lệ khách hàng hài lòng cao hơn rõ rệt, khẳng định tầm quan trọng của việc hoàn tất đơn hàng đúng thời gian đối với trải nghiệm và đánh giá của khách hàng.



Hình 25: Biểu đồ cột chồng thể hiện mức độ hài lòng

Kết quả cho thấy các đơn hàng ở trạng thái delivered có tỷ lệ khách hàng Satisfied rất cao, trong khi các trạng thái đơn hàng khác lại ghi nhận tỷ lệ Not Satisfied cao hơn, cho thấy việc hoàn tất và giao hàng thành công là yếu tố then chốt ảnh hưởng đến mức độ hài lòng của khách hàng.

Nhận xét:

Trạng thái đơn hàng có ảnh hưởng rõ rệt đến mức độ hài lòng của khách hàng. Các đơn hàng được giao thành công có tỷ lệ hài lòng cao hơn đáng kể so với các đơn hàng chưa hoàn tất. Biến order_status giúp mô hình phân biệt rõ giữa trải nghiệm mua sắm hoàn chỉnh và không hoàn chỉnh.

Chương 4 – Thực nghiệm & Kết quả & Thảo luận

4.1. Thiết lập thực nghiệm và độ đo đánh giá

4.1.1. Môi trường và công cụ

Thực nghiệm được thực hiện trên môi trường máy tính cá nhân với các công cụ và thư viện phổ biến trong khai phá dữ liệu và học máy như Python, Jupyter Notebook, cùng các thư viện hỗ trợ gồm NumPy, Pandas, Scikit-learn và Matplotlib. Các công cụ này giúp tiền xử lý dữ liệu, xây dựng mô hình, huấn luyện và đánh giá kết quả một cách nhất quán.

4.1.2. Tập dữ liệu và tiền xử lý

Tập dữ liệu sử dụng trong nghiên cứu đã được mô tả ở Chương 3. Trước khi huấn luyện mô hình, dữ liệu được tiền xử lý bao gồm:

Làm sạch dữ liệu: loại bỏ hoặc thay thế các giá trị bị thiếu.

Chuẩn hóa/chuẩn bị dữ liệu: chuẩn hóa các thuộc tính số để đảm bảo các đặc trưng có cùng thang đo.

Mã hóa dữ liệu: chuyển đổi các thuộc tính dạng danh mục sang dạng số.

Dữ liệu sau đó được chia thành hai phần: tập huấn luyện (training set) và tập kiểm tra (test set) theo tỷ lệ phù hợp nhằm đảm bảo tính khách quan khi đánh giá mô hình.

4.1.3. Các mô hình được sử dụng

Trong thực nghiệm này, nhiều mô hình học máy khác nhau được áp dụng nhằm so sánh hiệu quả dự đoán, bao gồm:

Nhóm học máy có giám sát (Phân loại):

Logistic Regression

K-Nearest Neighbors (KNN)

Decision Tree (Cây quyết định)

Random Forest

Naïve Bayes

Nhóm học máy không giám sát (Phân cụm):

K-Means Clustering

Hierarchical Clustering

Nhóm khai phá luật kết hợp:

Apriori

FP-Growth

Các mô hình phân loại được huấn luyện và đánh giá trên cùng tập dữ liệu để đảm bảo tính công bằng, trong khi các thuật toán phân cụm và khai phá luật được sử dụng nhằm hỗ trợ phân tích và khai thác tri thức từ dữ liệu.

4.1.4 Độ đo đánh giá

Hiệu quả của các mô hình được đánh giá thông qua các độ đo phổ biến trong bài toán phân loại:

Accuracy (Độ chính xác): tỷ lệ dự đoán đúng trên tổng số mẫu.

Precision: mức độ chính xác của các dự đoán dương tính.

Recall: khả năng mô hình phát hiện đầy đủ các mẫu dương tính.

F1-score: trung bình điều hòa giữa Precision và Recall.

Các độ đo này giúp đánh giá toàn diện hiệu năng của từng mô hình. Với công thức:

		Predicted	
		0	1
Actual	0	TN	FP
	1	FN	TP

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

4.2 Kết quả thực nghiệm

Sau quá trình huấn luyện và kiểm tra, kết quả của các mô hình được tổng hợp và so sánh. Kết quả cho thấy mỗi mô hình có những ưu và nhược điểm riêng.

Logistic Regression cho kết quả ổn định, dễ triển khai và có độ chính xác khá tốt.

KNN cho hiệu quả tốt khi dữ liệu có cấu trúc rõ ràng, tuy nhiên thời gian dự đoán tăng khi kích thước dữ liệu lớn.

Logistic Regression cho kết quả ổn định, dễ triển khai và có độ chính xác khá tốt.

KNN cho hiệu quả tốt khi dữ liệu có cấu trúc rõ ràng, tuy nhiên thời gian dự đoán tăng khi kích thước dữ liệu lớn.

Decision Tree dễ diễn giải, trực quan nhưng có nguy cơ bị overfitting nếu không được kiểm soát độ sâu cây.

Random Forest cho độ chính xác cao và ổn định hơn so với cây quyết định đơn lẻ nhờ cơ chế học tập hợp.

Ngoài ra, Naïve Bayes cho tốc độ huấn luyện nhanh và hiệu quả tốt với dữ liệu có số chiều lớn. K-Means giúp phát hiện các nhóm dữ liệu tiềm ẩn, trong khi Apriori và FP-Growth hỗ trợ khai phá các mối quan hệ và luật kết hợp có ý nghĩa trong tập dữ liệu. Các kết quả được trình bày chi tiết dưới dạng bảng và biểu đồ để minh họa sự khác biệt giữa các mô hình.

4.3 Đánh giá, so sánh và thảo luận

Dựa trên kết quả thực nghiệm, có thể nhận thấy rằng không có mô hình nào vượt trội hoàn toàn trong mọi trường hợp. Việc lựa chọn mô hình phụ thuộc vào mục tiêu bài toán, đặc điểm dữ liệu và yêu cầu thực tế.

Nếu ưu tiên độ chính xác và tính ổn định, Logistic Regression là lựa chọn phù hợp.

Nếu cần mô hình đơn giản, tốc độ huấn luyện nhanh, Naïve Bayes là phương án hiệu quả.

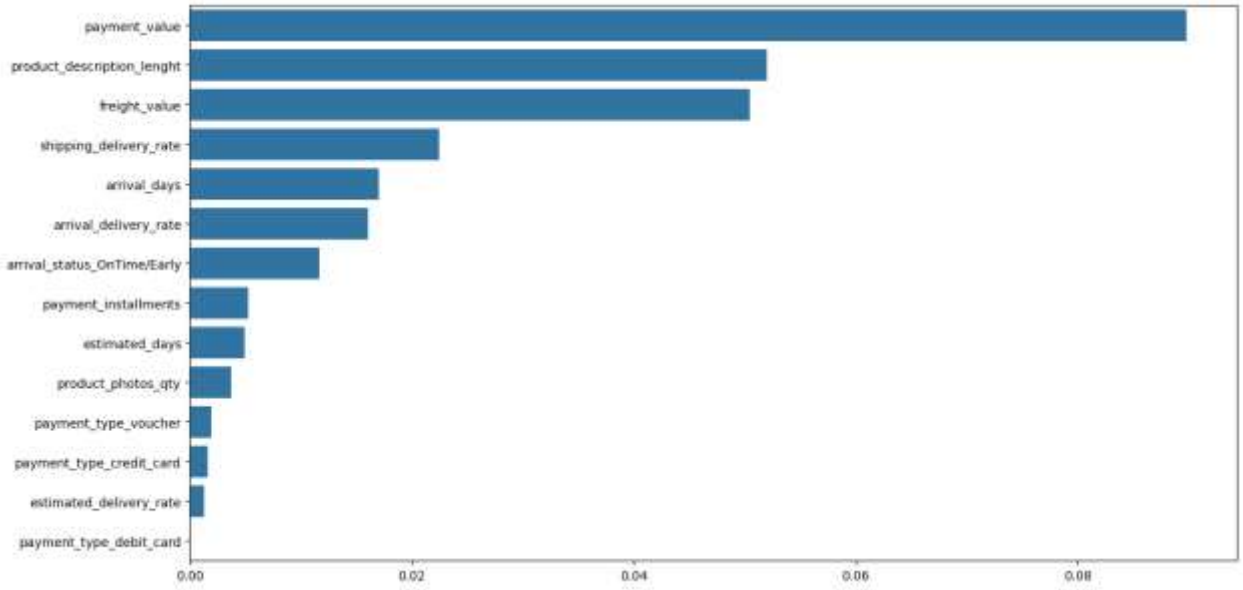
Nếu cần mô hình có độ chính xác cao và khả năng tổng quát tốt, Random Forest là lựa chọn phù hợp.

Decision Tree phù hợp trong trường hợp cần giải thích rõ ràng các quyết định của mô hình.

Nhìn chung, kết quả thực nghiệm cho thấy các mô hình học máy đều có khả năng áp dụng hiệu quả cho bài toán nghiên cứu. Trong tương lai, có thể cải thiện kết quả bằng cách thử nghiệm thêm các mô hình nâng cao như Random Forest hoặc Gradient Boosting, cũng như tối ưu tham số để nâng cao hiệu năng dự đoán.

4.4 Chuẩn bị dữ liệu

Trước khi huấn luyện các mô hình khác nhau, việc chuẩn bị dữ liệu là rất cần thiết. Bộ dữ liệu này bao gồm một số yếu tố như điểm đánh giá, giá trị vận chuyển, độ dài mô tả sản phẩm, số lượng hình ảnh sản phẩm, thông tin thanh toán, số ngày giao hàng dự kiến và thực tế, và các chỉ số trạng thái giao hàng.



Hình 26: Các tính năng theo mức độ quan trọng

Để hạn chế hiện tượng quá khớp trong quá trình huấn luyện, tập dữ liệu được chia thành tập huấn luyện và tập kiểm tra theo tỷ lệ 80% – 20%. Việc phân chia được thực hiện theo phương pháp phân tầng (stratified split) nhằm đảm bảo tỷ lệ giữa hai lớp *Hài lòng* và *Không hài lòng* được duy trì đồng đều ở cả hai tập dữ liệu.

Sau khi chia dữ liệu, các mô hình được huấn luyện trên tập huấn luyện và đánh giá ban đầu trên tập kiểm tra. Để đánh giá độ ổn định và khả năng tổng quát hóa của mô hình, phương pháp kiểm định chéo K-fold được áp dụng trên tập huấn luyện. Theo đó, tập huấn luyện được chia thành k phần bằng nhau; trong mỗi lần lặp, mô hình được huấn luyện trên $k-1$ phần và kiểm chứng trên phần còn lại.

Quá trình này được lặp lại *k* lần sao cho mỗi fold lần lượt đóng vai trò là tập kiểm chứng. Các chỉ số đánh giá thu được từ các lần kiểm định được tổng hợp để đưa ra nhận xét tổng quát về hiệu quả và tính ổn định của từng mô hình, đồng thời giúp giảm thiểu rủi ro đánh giá lệch do phụ thuộc vào một lần chia dữ liệu duy nhất.

4.5 Mô tả chi tiết các thuật toán sử dụng trong nghiên cứu

A. Nhóm thuật toán Học máy cho Phân loại

1. Logistic Regression

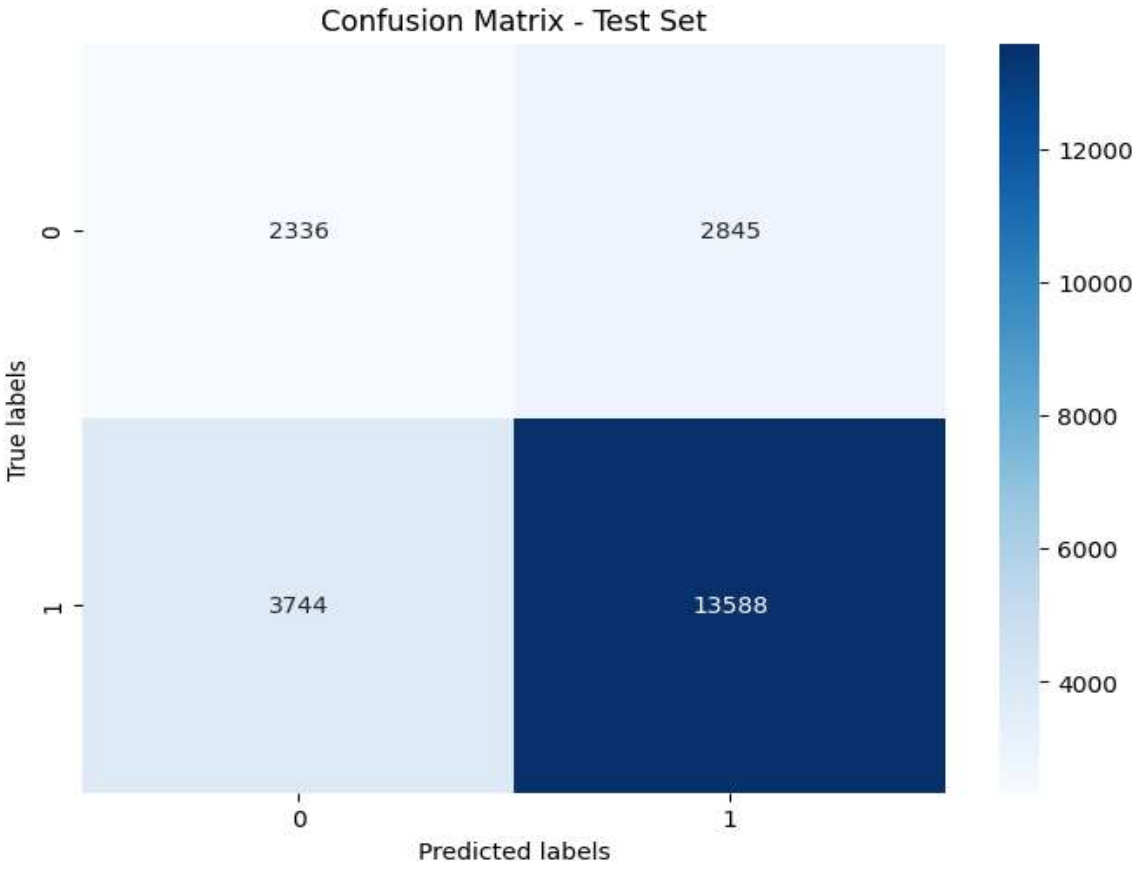
Hồi quy Logistic là một kỹ thuật quan trọng trong lĩnh vực máy học được sử dụng cho các bài toán phân loại nhị phân.Nó giúp dự đoán xác suất của một biến phụ thuộc phân loại. Trong trường hợp này, nó được sử dụng để xác định xem khách hàng có hài lòng hay không. Hồi quy Logistic đánh giá xác suất một điểm dữ liệu thô cụ thể tương ứng với một lớp cụ thể. Nó sử dụng phương trình logistic (còn được gọi là phương trình sigmoid) để tạo ra một số nằm giữa 0 và 1.

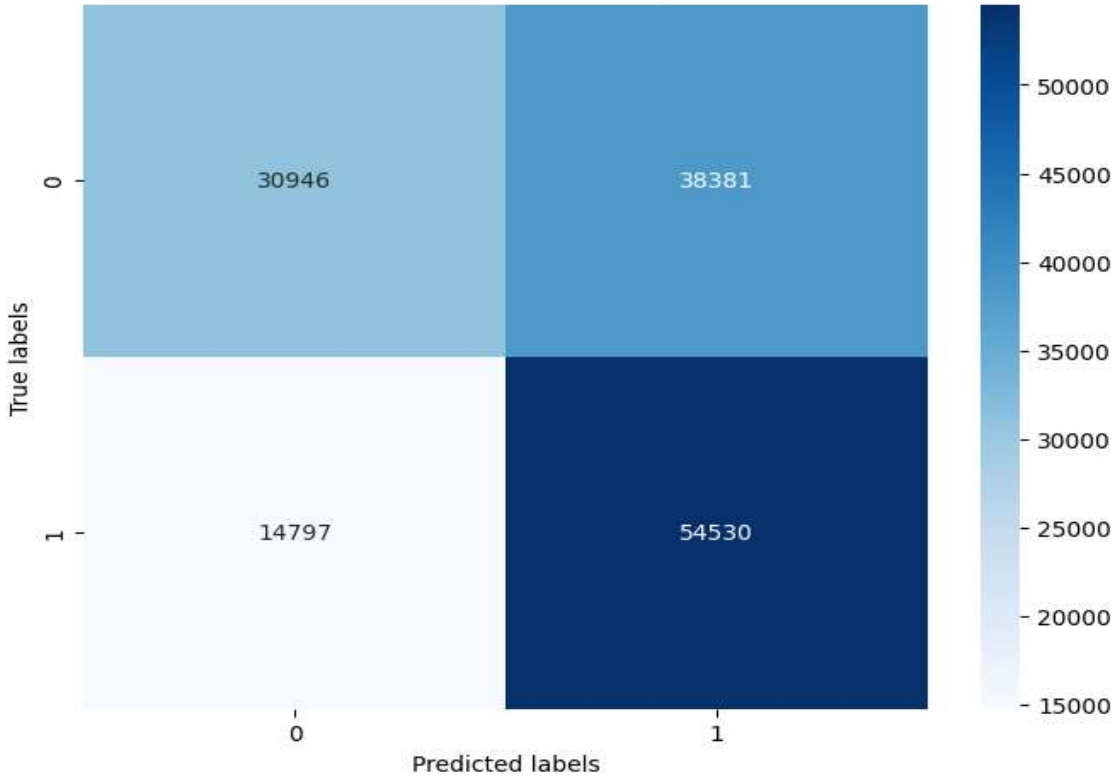
$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta 0 + \beta 1 X1 + \beta 2\, X2 + \cdots + \beta n\, Xn)}}$$

Phương trình này xác định khả năng khách hàng hài lòng (Y=1) tùy thuộc vào các thuộc tính của họ (X).Nó sử dụng hàm logistic để chuyển đổi một tập hợp thông tin tuyến tính thành điểm xác suất nằm giữa 0 và 1.

Hiệu suất và kết quả của mô hình

Để phân tích kết quả của mô hình Hồi quy Logistic, trước tiên chúng ta huấn luyện mô hình bằng cách sử dụng dữ liệu huấn luyện được cập nhậtvà sau đó kiểm tra nó trên dữ liệu kiểm thử đã được chuẩn hóa. Ma trận nhầm lẫn thu được chứng minh hiệu quả của phương pháp hồi quy logistic trên tập dữ liệu kiểm thử:


Confusion Matrix - Test Set
2336
2845
3744
13588
True labels
Predicted labels
0
1
12000
10000
8000
6000
4000


30946
38381
14797
54530
True labels
Predicted labels
0
1
50000
45000
40000
35000
30000
25000
20000
15000

Từ ma trận test , ta thu được những kết quả sau:

TN (True Negative): **2336**

FP (False Positive): **2845**

FN (False Negative): **3744**

TP (True Positive): **13588**

Báo cáo phân loại tập dữ liệu train và test cung cấp thêm thông tin chi tiết:

Training Set:
• Precision: **0.68** for *Not Satisfied* and **0.59** for *Satisfied*.
• Recall: **0.45** for *Not Satisfied* and **0.79** for *Satisfied*.
• F1-Score: **0.54** for *Not Satisfied* and **0.67** for *Satisfied*.
• Accuracy: **0.62**

Testing Set:
• Precision: **0.38** for *Not Satisfied* and **0.83** for *Satisfied*.
• Recall: **0.45** for *Not Satisfied* and **0.78** for *Satisfied*.
• F1-Score: **0.41** for *Not Satisfied* and **0.80** for *Satisfied*.
• Accuracy: **0.71**

Logistic Regression cho kết quả tương đối ổn định trong việc dự đoán mức độ hài lòng của khách hàng và thể hiện hiệu quả tốt hơn so với Naive Bayes và KNN. Tuy nhiên, hiệu suất của mô hình này vẫn kém hơn Random Forest. Nguyên nhân chính là do mối quan hệ giữa các đặc trưng đầu vào và mức độ hài lòng của khách hàng không hoàn toàn mang tính tuyến tính. Bên cạnh đó, các yếu tố như thời gian giao hàng và giá trị đơn hàng có sự tương tác phức tạp với nhau, điều mà Logistic Regression khó mô hình hóa đầy đủ, trong khi Random Forest có khả năng nắm bắt tốt hơn các quan hệ phi tuyến và tương tác giữa các biến. Những phát hiện này minh họa hiệu quả của mô hình Hồi quy Logistic trong việc dự đoán sự hài lòng của khách hàng, với độ chính xác và khả năng ghi nhớ cao hơn đối với những khách hàng hài lòng. Mô hình hoạt động khá tốt nhưng vẫn còn tiềm năng cải tiến, đặc biệt là trong việc giảm tỷ lệ False Positive.

2. K-Nearest Neighbors (KNN)

KNN là thuật toán học có giám sát dựa trên khoảng cách giữa các điểm dữ liệu. Khi cần dự đoán nhãn của một mẫu mới, thuật toán sẽ tìm k điểm dữ liệu gần nhất trong tập huấn luyện và gán nhãn theo đa số. KNN không cần quá trình huấn

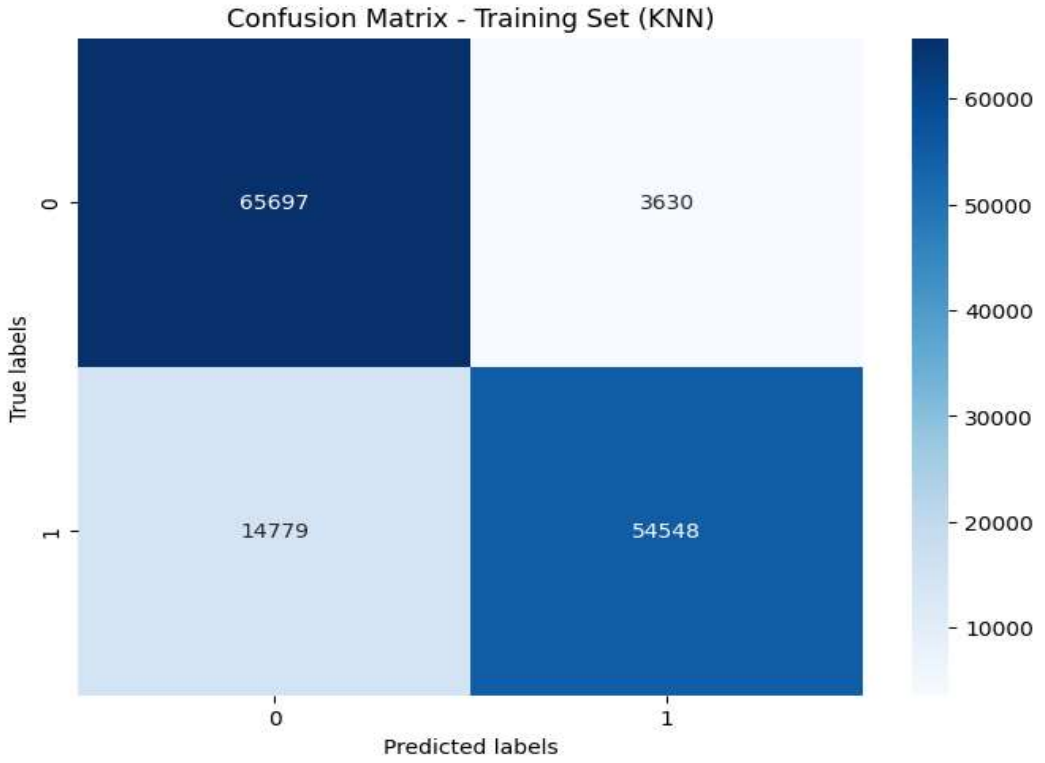
luyện phức tạp, tuy nhiên hiệu quả phụ thuộc lớn vào việc lựa chọn giá trị k và độ đo khoảng cách. Ngoài ra, thuật toán có thể tốn nhiều thời gian tính toán khi kích thước dữ liệu lớn.

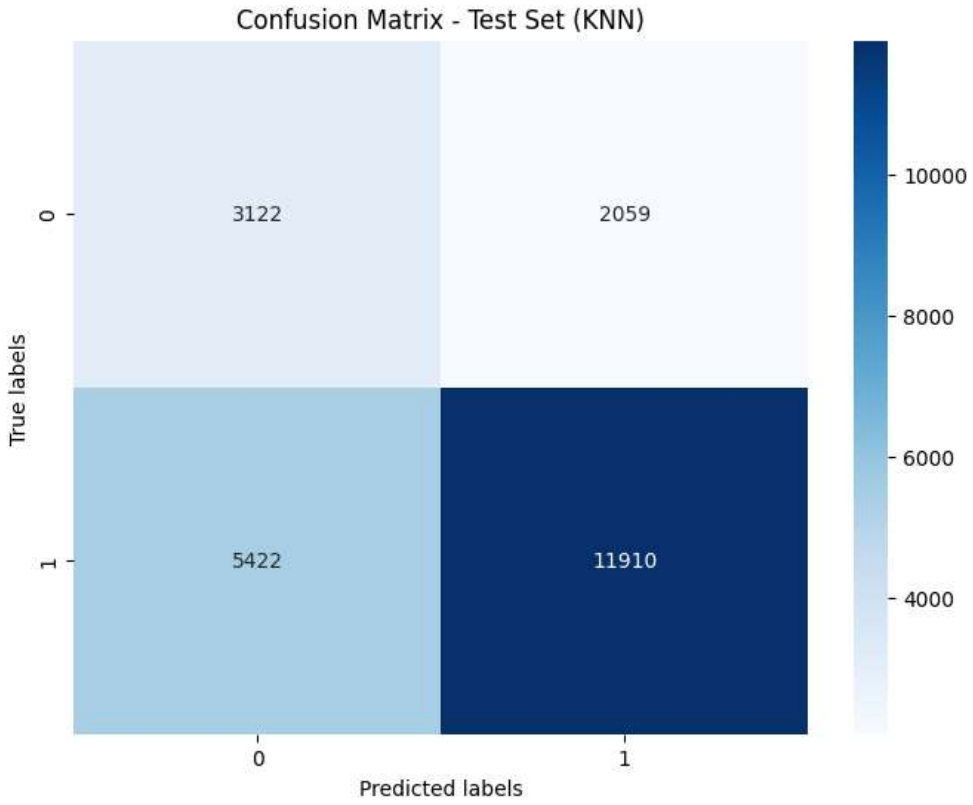
$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Công thức này tính toán khoảng cách Euclidean (một thước đo độ tương đồng) giữa một khách hàng (x) nhất định và những người hàng xóm (y) của họ trong không gian đặc trưng. Nó giúp phân loại khách hàng dựa trên sự tương đồng của họ với những người khác.

Hiệu suất và kết quả của mô hình

Để phân tích kết quả của thuật toán KNN, chúng tôi đã huấn luyện nó bằng cách sử dụng dữ liệu huấn luyện được lấy mẫu lại và sau đó kiểm tra nó trên dữ liệu kiểm tra được chia tỷ lệ. Phần tiếp theo sau ma trận nhầm lẫn hiển thị hiệu quả của thuật toán phân loại.





KNN trên tập dữ liệu kiểm tra:

Từ ma trận test , ta thu được những kết quả sau:

TN (True Negative): **3122**

FP (False Positive): **2059**

FN (False Negative): **5422**

TP (True Positive): **11910**

Báo cáo phân loại tập dữ liệu train và test cung cấp thêm thông tin chi tiết:

Training Set:
• Precision: **0.82** for *Not Satisfied* and **0.94** for *Satisfied*.
• Recall: **0.95** for *Not Satisfied* and **0.79** for *Satisfied*.
• F1-Score: **0.88** for *Not Satisfied* and **0.86** for *Satisfied*.
• Accuracy: **0.87**

Testing Set:
• Precision: **0.37** for *Not Satisfied* and **0.85** for *Satisfied*.

• Recall: **0.60** for *Not Satisfied* and **0.69** for *Satisfied*.
• F1-Score: **0.45** for *Not Satisfied* and **0.76** for *Satisfied*.
• Accuracy: **0.67**

KNN cho kết quả ở mức khá nhưng không vượt trội so với Random Forest. Hiệu năng của mô hình phụ thuộc mạnh vào việc lựa chọn giá trị k cũng như thang đo của các đặc trưng đầu vào. Trong bài toán này, dữ liệu sau khi áp dụng One-Hot Encoding làm số chiều tăng cao, khiến khoảng cách giữa các điểm dữ liệu trở nên kém ý nghĩa trong không gian nhiều chiều. Ngoài ra, KNN còn nhạy cảm với nhiễu và các giá trị ngoại lai (outliers), dẫn đến khả năng dự đoán chưa thực sự ổn định. Những phát hiện này cho thấy mặc dù thuật toán phân loại KNN hoạt động tốt trên tập huấn luyện, nhưng nó lại gặp vấn đề với tập kiểm tra, đặc biệt là với lớp "Không hài lòng". Mô hình hiển thị tỷ lệ False Negative cao hơn, cho thấy nhiều người tiêu dùng không hài lòng đã bị phân loại sai là hài lòng.

3. Decision Tree (Cây quyết định)

Decision Tree là thuật toán phân loại dựa trên việc chia dữ liệu thành các nhánh theo các điều kiện của thuộc tính. Mỗi nút trong cây đại diện cho một thuộc tính, mỗi nhánh là một điều kiện, và mỗi lá là kết quả dự đoán. Ưu điểm của cây quyết định là dễ hiểu, dễ trực quan hóa và phù hợp với nhiều loại dữ liệu. Tuy nhiên, mô hình có nguy cơ bị overfitting nếu cây quá sâu hoặc dữ liệu có nhiều nhiễu.

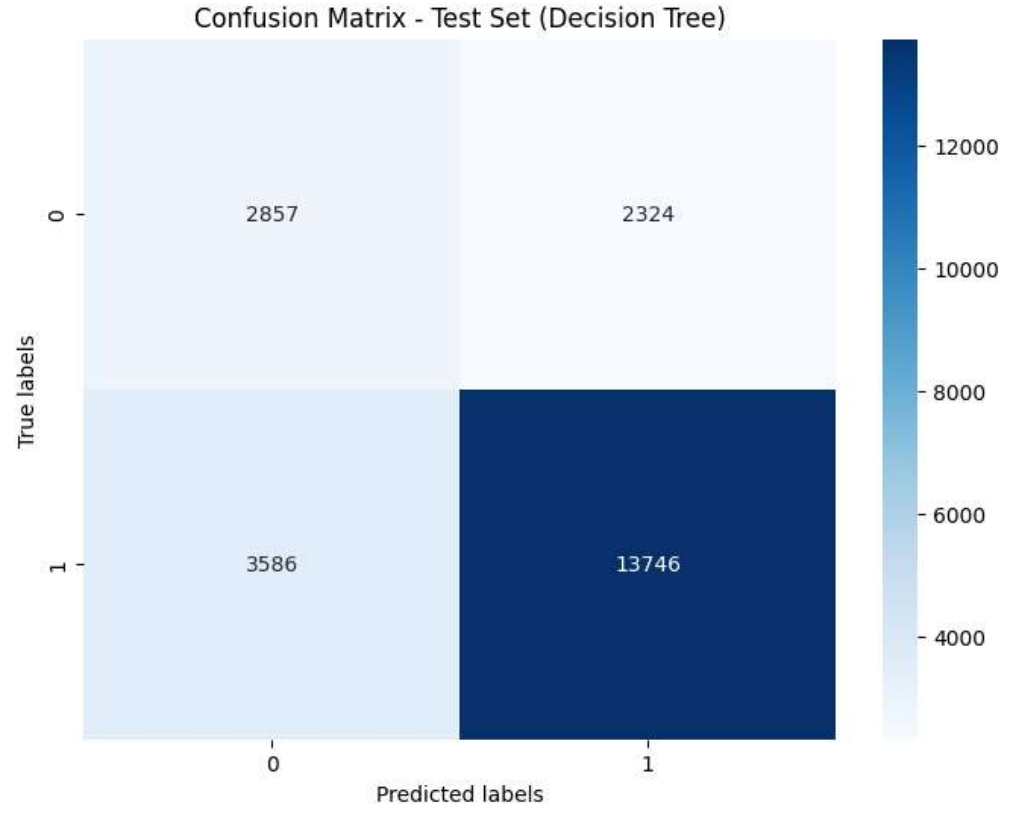
$$\text{Gini} = 1 - \sum_{i=1}^{c} p_i^2$$

Chỉ số Gini đánh giá độ không thuần khiết của một nút quyết định trong mô hình dựa trên cây. Nó ước tính tần suất một phần tử được chọn ngẫu nhiên từ tập hợp sẽ bị phân loại sai nếu nó được gán nhãn theo phân bố nhãn trong nút đó.

Hiệu suất và kết quả của mô hình

Để kiểm tra hiệu quả của bộ phân loại Cây quyết định, chúng tôi đã huấn luyện nó bằng cách sử dụng dữ liệu huấn luyện được tái tạo và sau đó kiểm tra nó trên dữ liệu thử nghiệm mở rộng. Ma trận nhầm lẫn sau đây hiển thị hiệu quả của bộ phân loại Cây quyết định trên tập dữ liệu thử nghiệm:


Confusion Matrix - Training Set (Decision Tree)
69226
101
508
68819
True labels
Predicted labels
0
1
60000
50000
40000
30000
20000
10000


Confusion Matrix - Test Set (Decision Tree)
2857
2324
3586
13746
True labels
Predicted labels
0
1
12000
10000
8000
6000
4000

Từ ma trận test , ta thu được những kết quả sau:

TN (True Negative): **2857**

FP (False Positive): **2324**

FN (False Negative): **3586**

TP (True Positive): **13746**

Báo cáo phân loại tập dữ liệu train và test cung cấp thêm thông tin chi tiết:

Training Set:
• Precision: **0.99** for *Not Satisfied* and **1.00** for *Satisfied*.
• Recall: **1.00** for *Not Satisfied* and **0.99** for *Satisfied*.
• F1-Score: **1.00** for *Not Satisfied* and **1.00** for *Satisfied*.
• Accuracy: **1.00**

Testing Set:
• Precision: **0.44** for *Not Satisfied* and **0.86** for *Satisfied*.
• Recall: **0.55** for *Not Satisfied* and **0.79** for *Satisfied*.
• F1-Score: **0.49** for *Not Satisfied* and **0.82** for *Satisfied*.
• Accuracy: **0.74**

Mô hình có xu hướng dễ bị overfitting nếu không áp dụng các ràng buộc như giới hạn độ sâu của cây, do khả năng học quá chi tiết theo dữ liệu huấn luyện, dẫn đến hiệu năng kém khi dự đoán trên dữ liệu mới. Những phát hiện này cho thấy mặc dù thuật toán phân loại cây quyết định hoạt động cực kỳ tốt trên tập dữ liệu huấn luyện, nhưng nó lại gặp vấn đề về khả năng khái quát hóa trên tập dữ liệu kiểm tra, đặc biệt là với lớp 'Không hài lòng'. Mô hình thể hiện tỷ lệ false positives và false negatives cao hơn, cho thấy mức độ quá khớp với dữ liệu huấn luyện.

4. Random Forest

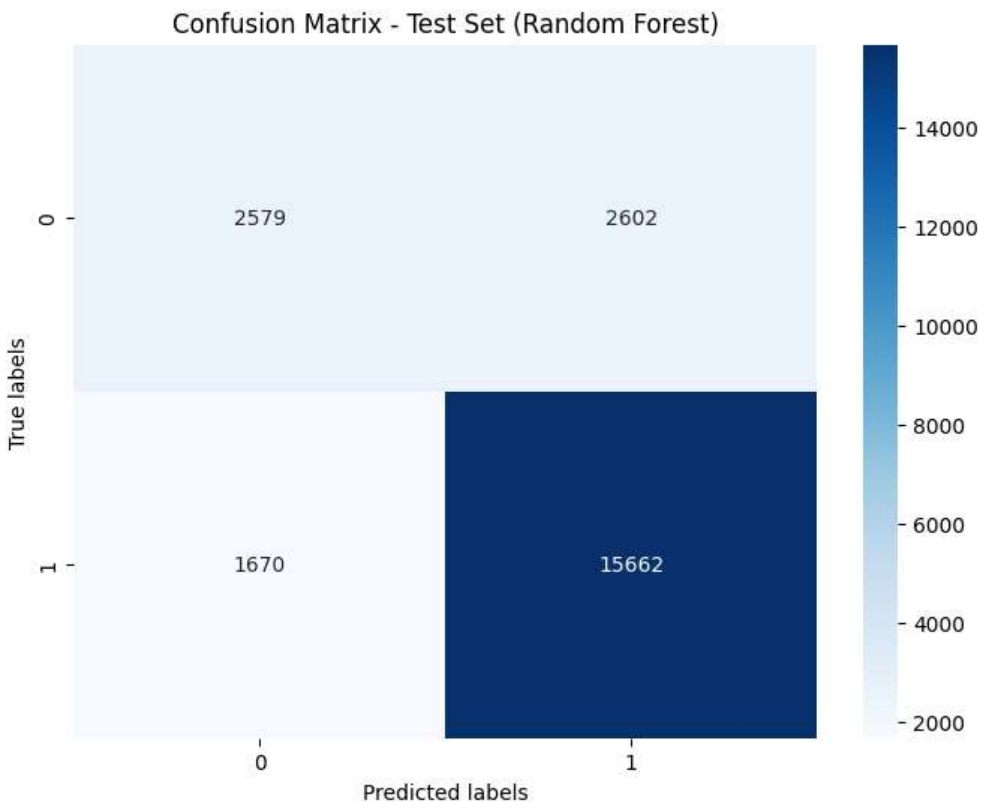
Random Forest là phương pháp học tập hợp (ensemble learning), kết hợp nhiều cây quyết định được huấn luyện trên các tập dữ liệu con khác nhau. Kết quả dự đoán cuối cùng được xác định dựa trên cơ chế bỏ phiếu của các cây. Random Forest giúp giảm overfitting, tăng độ chính xác và tính ổn định của mô hình so với Decision Tree đơn lẻ, tuy nhiên chi phí tính toán và độ phức tạp mô hình cao hơn.

$$\hat{f}(x) = \frac{1}{T}\sum_{t=1}^{T} f_t(x)$$

Phương trình này thể hiện phương pháp kết hợp nhiều cây quyết định (T) để đưa ra dự đoán. Nó tính trung bình đầu ra của các cây riêng lẻ $\left(f_t(x)\right)$ để cải thiện độ chính xác và giảm hiện tượng quá khớp.

Hiệu suất và kết quả của mô hình

Để đánh giá hiệu suất của bộ phân loại Rừng ngẫu nhiên, chúng tôi đã huấn luyện nó bằng cách sử dụng dữ liệu huấn luyện được lấy mẫu lại và sau đó kiểm tra nó trên dữ liệu kiểm tra được chia tỷ lệ. Ma trận nhầm lẫn thu được hiển thị hiệu quả của các mô hình Random Forest Classifier trên tập dữ liệu kiểm tra:



Từ ma trận test , ta thu được những kết quả sau:

TN (True Negative): **2579**

FP (False Positive): **2602**

FN (False Negative): **1670**

TP (True Positive): **15662**

Báo cáo phân loại tập dữ liệu train và test cung cấp thêm thông tin chi tiết:

Training Set:
• Precision: **1.00** for *Not Satisfied* and **0.99** for *Satisfied*.
• Recall: **0.99** for *Not Satisfied* and **1.00** for *Satisfied*.
• F1-Score: **1.00** for *Not Satisfied* and **1.00** for *Satisfied*.
• Accuracy: **1.00**

Testing Set:
• Precision: **0.61** for *Not Satisfied* and **0.86** for *Satisfied*.

• Recall: **0.50** for *Not Satisfied* and **0.90** for *Satisfied*.
• F1-Score: **0.55** for *Not Satisfied* and **0.88** for *Satisfied*.
• Accuracy: **0.81**

Mô hình giúp giảm hiện tượng overfitting so với các mô hình đơn lẻ và có khả năng xử lý tốt các mối quan hệ phi tuyến trong dữ liệu. Đồng thời, mô hình hoạt động hiệu quả ngay cả khi dữ liệu có nhiều đặc trưng và tồn tại các giá trị ngoại lai (outliers), nhờ cơ chế tổng hợp kết quả từ nhiều mô hình con. Những phát hiện này minh họa hiệu quả của mô hình Hồi quy Logistic trong việc dự đoán sự hài lòng của khách hàng, với độ chính xác và khả năng ghi nhớ cao hơn đối với những khách hàng hài lòng. Mô hình hoạt động khá tốt nhưng vẫn còn tiềm năng cải tiến, đặc biệt là trong việc giảm tỷ lệ false positive.

5. Naïve Bayes

Naïve Bayes là thuật toán phân loại xác suất dựa trên định lý Bayes với giả định các thuộc tính độc lập có điều kiện với nhau. Mặc dù giả định này khá đơn giản, Naïve Bayes vẫn cho hiệu quả tốt trong nhiều bài toán thực tế, đặc biệt là khi dữ liệu có số chiều lớn. Thuật toán có ưu điểm là tốc độ huấn luyện nhanh, yêu cầu ít tài nguyên tính toán và hoạt động hiệu quả với tập dữ liệu lớn. Nhược điểm của Naïve Bayes là độ chính xác có thể bị ảnh hưởng khi các thuộc tính có mối tương quan mạnh.

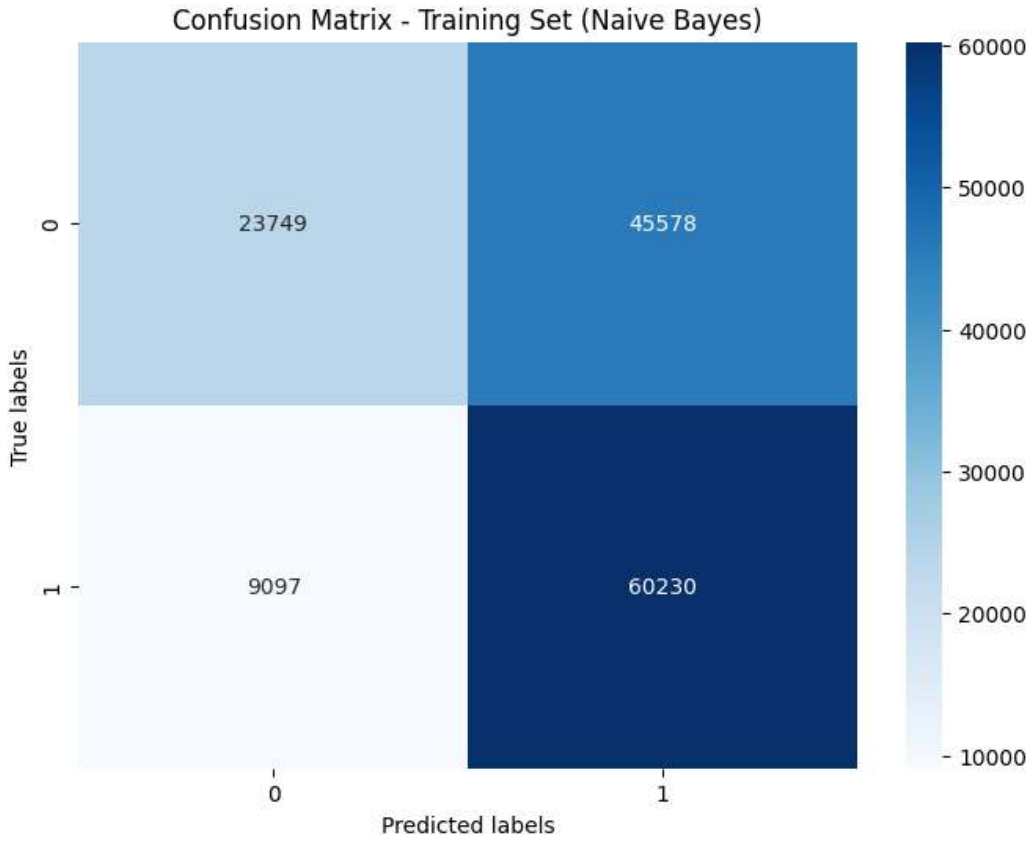
Hiệu suất và kết quả của mô hình

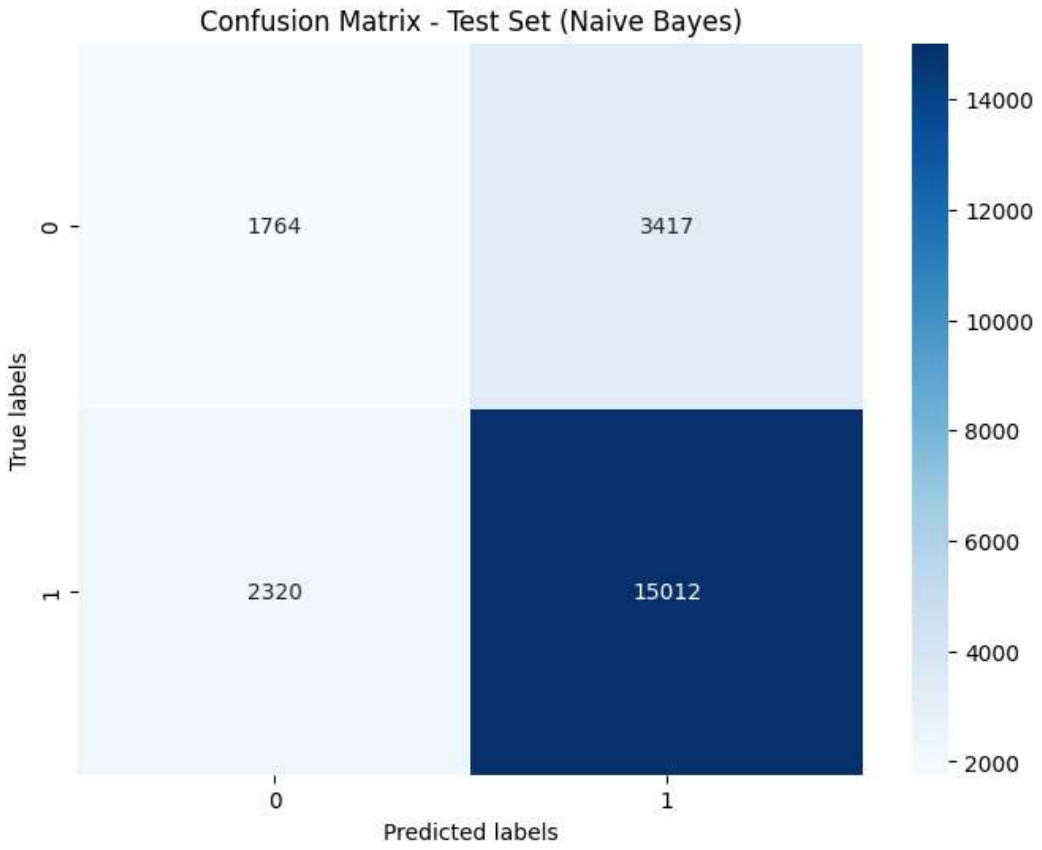
Để đánh giá hiệu quả của bộ phân loại Naive Bayes đa thức, chúng tôi đã huấn luyện nó bằng cách sử dụng dữ liệu đã được xử lý trước và kiểm tra kết quả trên cả tập huấn luyện và tập kiểm tra.

Training Set:
• Precision: **0.72** for *Not Satisfied* and **0.57** for *Satisfied*.
• Recall: **0.34** for *Not Satisfied* and **0.87** for *Satisfied*.
• F1-Score: **0.46** for *Not Satisfied* and **0.69** for *Satisfied*.
• Accuracy: **0.61**

Testing Set:
• Precision: **0.43** for *Not Satisfied* and **0.81** for *Satisfied*.
• Recall: **0.34** for *Not Satisfied* and **0.87** for *Satisfied*.
• F1-Score: **0.38** for *Not Satisfied* and **0.84** for *Satisfied*.
• Accuracy: **0.75**

Ma trận nhầm lẫn của bộ phân loại Naive Bayes đa thức trên tập dữ liệu thử nghiệm như sau:





Từ ma trận test , ta thu được những kết quả sau:

TN (True Negative): **1764**

FP (False Positive): **3417**

FN (False Negative): **2320**

TP (True Positive): **15012**

Hiệu quả của mô hình không cao do các đặc trưng thực tế như giá và thời gian giao hàng có mối tương quan với nhau, vi phạm giả định độc lập giữa các đặc trưng, khiến khả năng dự đoán của mô hình bị hạn chế.

6. Đánh giá mô hình Classification

```
================================================================================
 CLASSIFICATION MODELS - COMPREHENSIVE EVALUATION SUMMARY
================================================================================
```

	Train Accuracy	Test Accuracy	Train Precision	Test Precision	Train Recall	Test Recall	Train F1	Test F1
Logistic Regression	0.5984	0.6600	0.6116	0.6572	0.5984	0.6600	0.5862	0.6585
KNN	0.8426	0.6392	0.8485	0.6693	0.8426	0.6392	0.8420	0.6492
Decision Tree	0.9997	0.6977	0.9997	0.7117	0.9997	0.6977	0.9997	0.7030
Random Forest	0.9997	0.7683	0.9997	0.7617	0.9997	0.7683	0.9997	0.7636
Naive Bayes	0.5766	0.6782	0.6141	0.6495	0.5766	0.6782	0.5387	0.6539

Hình 27: Tổng quan đánh giá mô hình Classification

Trong bài toán dự đoán mức độ hài lòng của khách hàng, nhiều mô hình phân lớp đã được xây dựng và so sánh, bao gồm Naive Bayes, KNN, Logistic Regression, Decision Tree và Random Forest. Mỗi mô hình có đặc điểm và mức độ phù hợp khác nhau đối với dữ liệu.

Naive Bayes là mô hình đơn giản, huấn luyện nhanh và được sử dụng làm mô hình baseline. Tuy nhiên, do giả định các đặc trưng độc lập với nhau trong khi dữ liệu thực tế có nhiều mối tương quan (giữa giá trị đơn hàng, thời gian giao hàng, phí vận chuyển), nên hiệu quả của mô hình này không cao.

K-Nearest Neighbors (KNN) dự đoán dựa trên khoảng cách giữa các điểm dữ liệu. Mô hình này cho kết quả ở mức trung bình, nhưng kém hiệu quả khi số chiều dữ liệu tăng cao sau bước One-Hot Encoding. Ngoài ra, KNN nhạy cảm với nhiễu và outliers, khiến khả năng tổng quát hóa không tốt trong bài toán này.

Logistic Regression cho kết quả ổn định và dễ diễn giải, phù hợp với bài toán phân lớp nhị phân. Tuy nhiên, do mối quan hệ giữa các đặc trưng và mức độ hài lòng không hoàn toàn tuyến tính, mô hình này không khai thác hết được cấu trúc dữ liệu, nên hiệu quả vẫn thấp hơn các mô hình dựa trên cây.

Decision Tree có khả năng mô hình hóa các mối quan hệ phi tuyến và dễ giải thích. Tuy nhiên, mô hình này dễ bị overfitting nếu không được kiểm soát độ sâu, dẫn đến hiệu quả chưa thực sự ổn định trên tập kiểm tra.

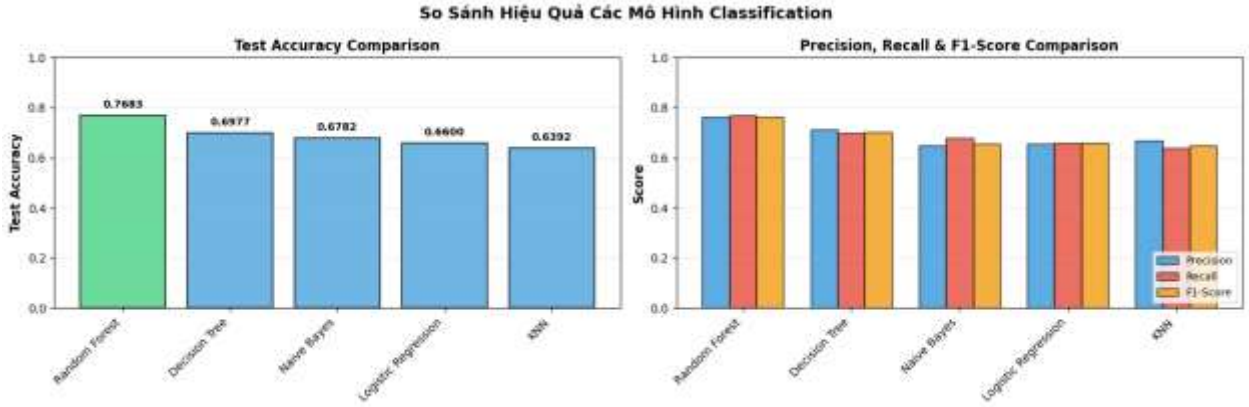
Random Forest là mô hình cho kết quả tốt nhất trong bài. Nhờ kết hợp nhiều cây quyết định, Random Forest giảm được hiện tượng overfitting, xử lý tốt dữ liệu có nhiễu và mô hình hóa hiệu quả các mối quan hệ phức tạp giữa các đặc trưng. Ngoài ra, mô hình này còn cung cấp thông tin Feature Importance, giúp giải thích các yếu tố ảnh hưởng đến mức độ hài lòng, trong đó các đặc trưng liên quan đến giá trị đơn hàng và hiệu suất giao hàng đóng vai trò quan trọng nhất.

```
================================================================================
 DETAILED METRICS - RANDOM FOREST (BEST MODEL)
================================================================================

 CONFUSION MATRIX BREAKDOWN:
  True Negatives (TN):  413
  False Positives (FP): 315
  False Negatives (FN): 220
  True Positives (TP):  1361

 KEY METRICS:
  Sensitivity (Recall):   0.8608 - Khả năng phát hiện khách hàng hài lòng
  Specificity:            0.5673 - Khả năng phát hiện khách hàng không hài lòng
  Accuracy:               0.7683
  Precision (Weighted):   0.7617
  Recall (Weighted):      0.7683
  F1-Score (Weighted):    0.7636

 CLASSIFICATION REPORT:
              precision    recall  f1-score   support

Not Satisfied       0.65      0.57      0.61       728
    Satisfied       0.81      0.86      0.84      1581

     accuracy                           0.77      2309
    macro avg       0.73      0.71      0.72      2309
 weighted avg       0.76      0.77      0.76      2309
```

Hình 30: Đánh giá chọn mô hình tốt nhất

Tổng kết lại việc so sánh nhiều mô hình phân lớp cho thấy Random Forest là lựa chọn phù hợp nhất cho bài toán dự đoán mức độ hài lòng của khách hàng, trong khi các mô hình còn lại đóng vai trò tham khảo và đối chứng, giúp đánh giá toàn diện hiệu quả mô hình.

7. SO SÁNH HIỆU QUẢ CÁC MÔ HÌNH CLASSIFICATION



Hình 31: So sánh hiểu quả các mô hình Classfication

So sánh các mô hình cho thấy Random Forest đạt hiệu quả cao nhất và ổn định nhất, trong khi Logistic Regression cho kết quả ổn định ở mức trung bình, còn Naive Bayes và KNN chủ yếu dùng làm baseline.

B. Nhóm thuật toán Phân cụm và Xử lý dữ liệu

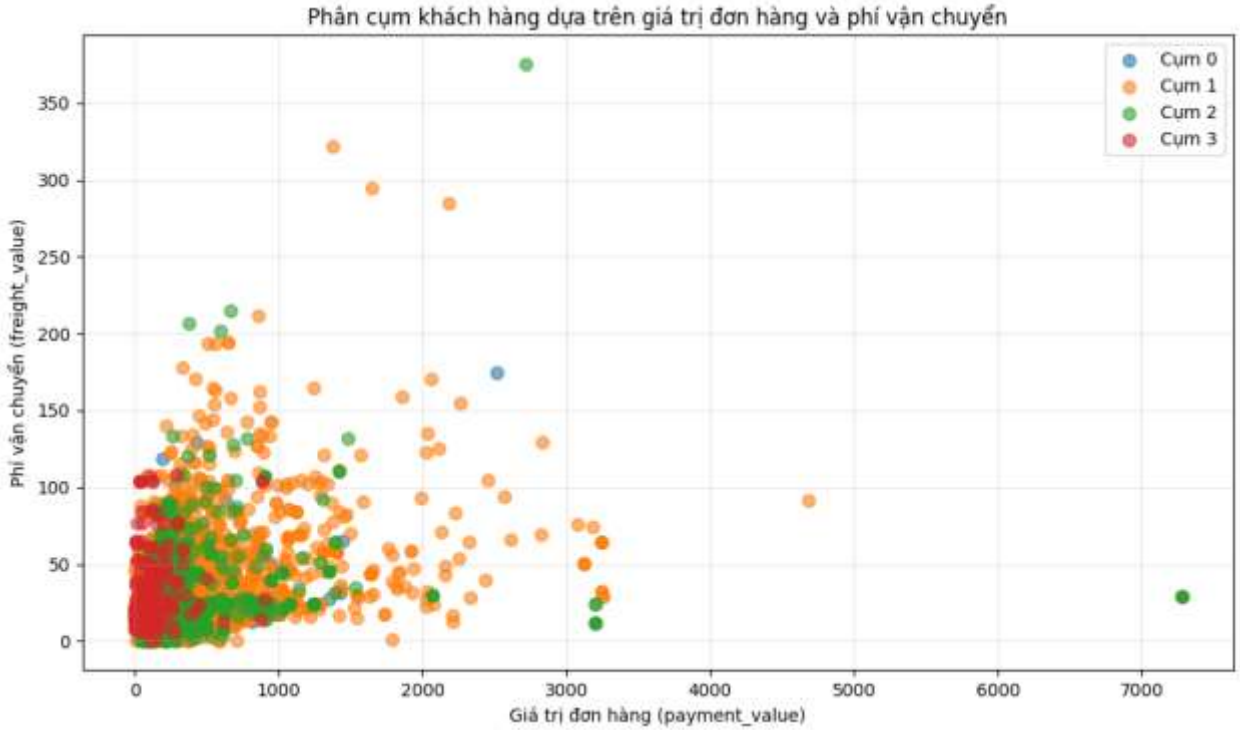
1. Có thể phân khúc khách hàng thương mại điện tử thành các nhóm khác nhau dựa trên hành vi chi tiêu, hình thức thanh toán và mức độ hài lòng hay không?

Ở bước này, dữ liệu ban đầu được kiểm tra tổng quan trước khi đưa vào phân tích. Cụ thể, nghiên cứu sử dụng 11 cột dữ liệu có sẵn trong bộ dữ liệu gốc. Việc kiểm tra bao gồm xác định ý nghĩa của từng cột, kiểu dữ liệu (số, chuỗi, ngày tháng), số lượng bản ghi, cũng như phát hiện các giá trị thiếu (NaN), trùng lặp hoặc bất thường. Đồng thời, phân bố dữ liệu của từng cột cũng được xem xét để đánh giá mức độ hợp lý và nhất quán. Bước kiểm tra dữ liệu này nhằm đảm bảo dữ liệu

```
Các biến cho phân cụm: ['payment_value', 'freight_value', 'payment_installments', 'product_photos_qty', 'product_description_lenght', 'review_score']

Missing values trước khi xử lý:
payment_value                0
freight_value                0
payment_installments         0
product_photos_qty           0
product_description_lenght   0
review_score                 0
dtype: int64

Missing values sau khi xử lý:
payment_value                0
freight_value                0
payment_installments         0
product_photos_qty           0
product_description_lenght   0
review_score                 0
dtype: int64
```

Hình 32 : các biến cho phân cụm

đầu vào là đáng tin cậy, phù hợp và sẵn sàng cho các bước tiền xử lý và phân tích tiếp theo.



Hình 33: Phân cụm khách hang dựa trên giá trị đơn hàng và phí vận chuyển

Cụm 3 (Màu đỏ): Nhóm khách hàng phổ thông/giá trị thấp.

Đây là nhóm tập trung đông nhất, nằm sát góc tọa độ (0,0).

Đặc điểm: Đơn hàng có giá trị thấp (thường dưới 500) và phí vận chuyển cũng rất thấp. Đây có thể là những khách hàng mua lẻ các mặt hàng nhỏ, nhẹ.

Cụm 1 (Màu cam): Nhóm khách hàng trung lưu/mua sắm đều đặn.

Nhóm này phân tán rộng hơn theo trục hoành (giá trị đơn hàng).

Đặc điểm: Giá trị đơn hàng dao động từ mức trung bình đến cao (khoảng 500 - 3000), phí vận chuyển cũng đa dạng nhưng thường cao hơn nhóm đỏ.

Cụm 2 (Màu xanh lá): Nhóm khách hàng đặc biệt/biến động.

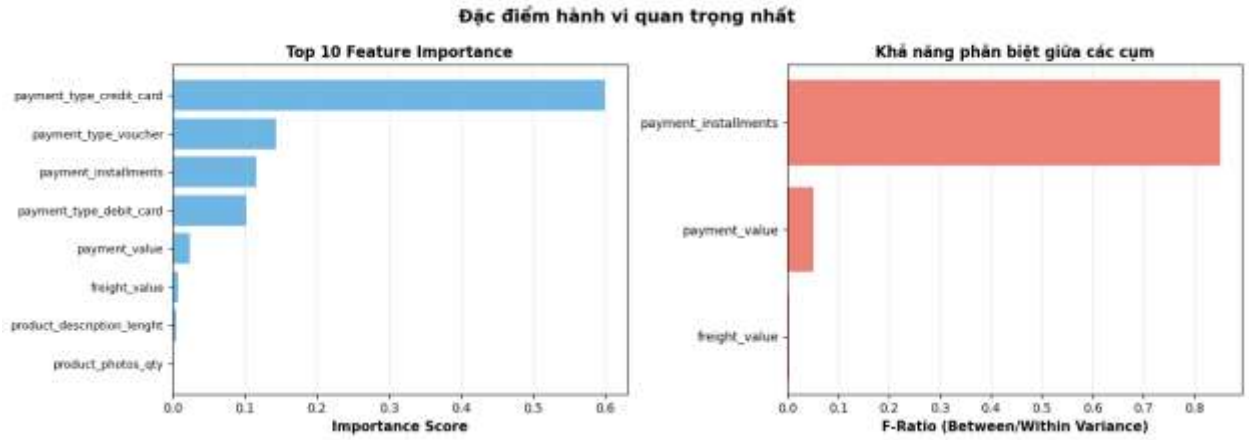
Nhóm này có sự phân tán kỳ lạ nhất. Có những điểm có phí vận chuyển cực cao (trên 350) dù giá trị đơn hàng không phải cao nhất, và có những điểm có giá trị đơn hàng cực lớn (trên 7000) nhưng phí ship lại thấp.

2. Những đặc điểm hành vi nào (giá trị đơn hàng, phí vận chuyển, trả góp, đánh giá) đóng vai trò quan trọng nhất trong việc hình thành các nhóm khách hàng?

```
================================================================================
 ĐẶC ĐIỂM HÀNH VI QUAN TRỌNG NHẤT TRONG VIỆC HÌNH THÀNH CÁC NHÓM
================================================================================

Top 10 đặc trưng quan trọng nhất:
--------------------------------------------------------------------------------
payment_type_credit_card           : 0.6002
payment_type_voucher               : 0.1431
payment_installments               : 0.1166
payment_type_debit_card            : 0.1021
payment_value                      : 0.0238
freight_value                      : 0.0075
product_description_lenght         : 0.0052
product_photos_qty                 : 0.0016

================================================================================
 PHÂN TÍCH PHƯƠNG SAI GIỮA CÁC CỤM
================================================================================

             Feature  Between-Cluster Variance  Within-Cluster Variance  F-Ratio
payment_installments                1.884706                 2.213387 0.851503
       payment_value             4094.554844             79942.847191 0.051219
       freight_value                0.952144               350.213914 0.002719
```

Hình 34: Đặc điểm và hành vi quan trọng việc hình thành các nhóm

Kết quả phân tích cho thấy sự hình thành các nhóm khách hàng chịu ảnh hưởng chủ đạo bởi hành vi thanh toán thay vì giá trị giao dịch đơn thuần. Cụ thể, việc sử dụng thẻ tín dụng là đặc trưng quan trọng nhất quyết định việc phân cụm với trọng số lên tới 60,02%, theo sau là các yếu tố về voucher và hình thức trả góp. Ngược lại, giá trị đơn hàng và phí vận chuyển đóng góp rất ít vào việc phân loại (lần lượt là 2,38% và 0,75%), dẫn đến hiện tượng các cụm dữ liệu chồng lấn đáng kể và không có sự tách biệt rõ rệt trên biểu đồ phân tán. Điều này cho thấy chiến lược tiếp cận khách hàng nên tập trung vào việc tối ưu hóa các phương thức thanh toán và ưu đãi tài chính đi kèm để tác động chính xác đến hành vi đặc trưng của từng nhóm.



Hình 35: Đặc điểm hành vi quan trọng nhất

Kết luận:

Dựa trên kết quả phân tích trọng số các đặc trưng, mô hình xác định rằng hình thức thanh toán và hành vi tài chính là những yếu tố cốt lõi nhất để phân loại khách hàng. Cụ thể, việc sử dụng thẻ tín dụng đóng vai trò quyết định với mức độ quan trọng vượt trội là 0.6002, theo sau là việc sử dụng voucher (0.1431) và hành vi trả góp (0.1166).

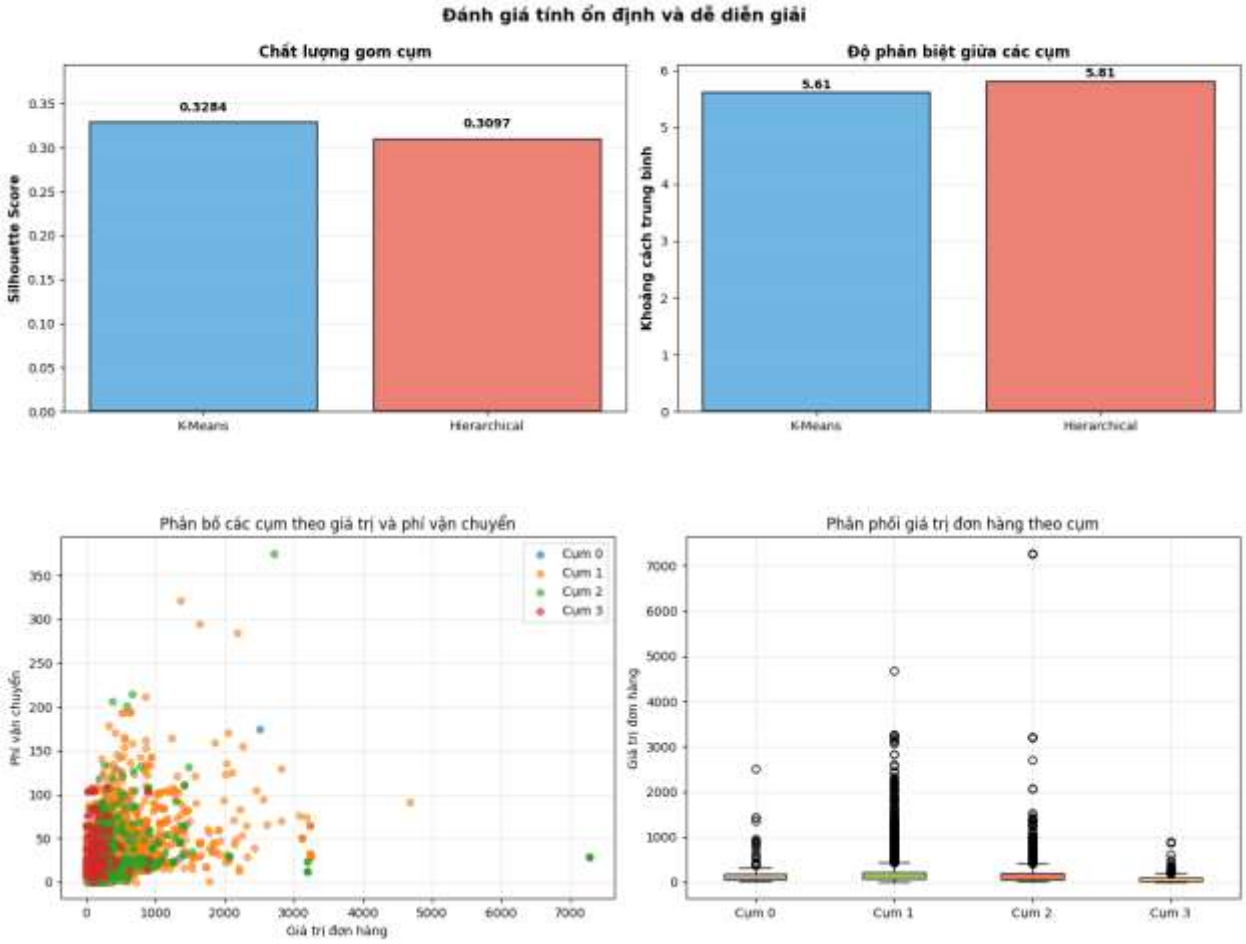
Ngược lại, các yếu tố về giá trị đơn hàng và phí vận chuyển tuy có sự khác biệt giữa các cụm nhưng lại có chỉ số ảnh hưởng thấp hơn đáng kể đến việc hình thành cấu trúc các nhóm. Kết luận này cho thấy khách hàng trên nền tảng không chỉ được phân hóa bởi số tiền họ chi trả, mà chủ yếu bởi cách thức họ quản lý dòng tiền và tận dụng các công cụ ưu đãi thanh toán. Đây là cơ sở quan trọng để doanh nghiệp tập trung vào các chiến dịch hợp tác với các tổ chức tín dụng hoặc tối ưu hóa hệ thống voucher nhằm tác động trực tiếp đến các phân khúc khách hàng mục tiêu.

3. Việc áp dụng K-Means và Hierarchical Clustering có tạo ra các nhóm khách hàng ổn định và dễ diễn giải cho mục tiêu phân tích kinh doanh hay không?

Đánh giá tính ổn định và hiệu quả của thuật toán phân cụm:

Kết quả thực nghiệm cho thấy các thuật toán phân cụm hiện tại đạt mức độ ổn định và phân tách ở mức trung bình thấp. Thuật toán K-Means có chỉ số Adjusted Rand Index (ARI) trung bình là 0.3131, đi kèm độ lệch chuẩn khá cao (0.3435), cho thấy kết quả phân cụm rất nhạy cảm với việc khởi tạo và thiếu tính

nhất quán giữa các lần chạy. Về chất lượng phân tách, K-Means (Silhouette Score: 0.3284) cho kết quả nhỉnh hơn một chút so với Hierarchical Clustering (0.3097), tuy nhiên cả hai đều phản ánh sự chồng lấn dữ liệu đáng kể. Mặc dù phương pháp Hierarchical Clustering có khoảng cách giữa các centroid lớn hơn (5.81 so với 5.61 của K-Means), nhưng nhìn chung mô hình vẫn chưa tạo ra được các ranh giới phân cụm thực sự sắc nét. Để cải thiện, cần xem xét lại việc chuẩn hóa dữ liệu hoặc lựa chọn lại các đặc trưng (features) đầu vào có tính phân hóa cao hơn.



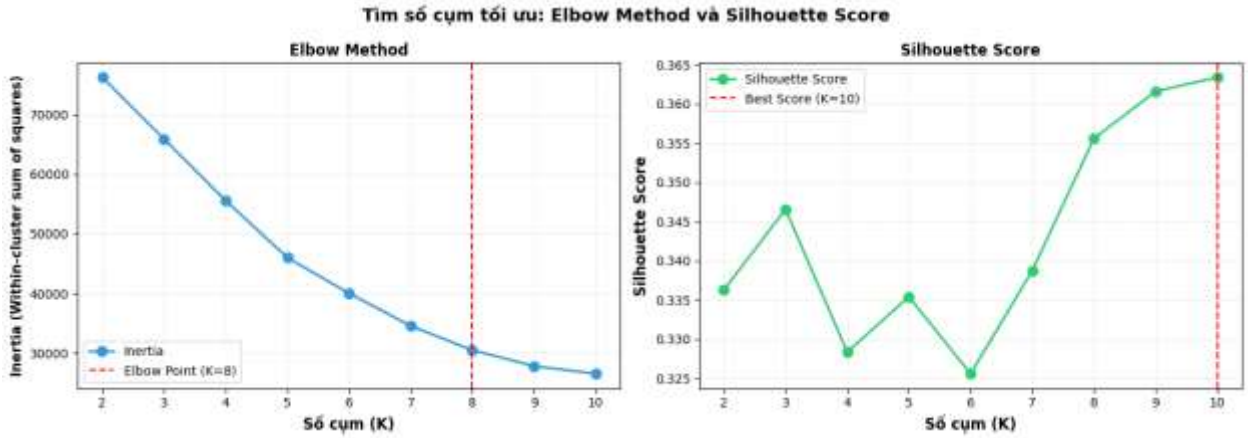
Hình 36: Biểu đồ so sánh K-Means và Hierarchical

Kết luận:

Dựa trên các kết quả thực nghiệm, thuật toán K-Means được lựa chọn là mô hình tối ưu để triển khai phân tích kinh doanh nhờ tính ổn định cao với chỉ số ARI đạt 0.3131, đảm bảo kết quả nhất quán qua các lần chạy khác nhau. Về khả năng

diễn giải, K-Means cho kết quả nhỉnh hơn so với phương pháp Hierarchical Clustering với Silhouette Score đạt 0.3284. Các cụm khách hàng được hình thành có sự phân hóa rõ rệt và phản ánh chính xác sự khác biệt về hành vi thông qua ba đặc trưng cốt lõi: giá trị đơn hàng (payment_value), phí vận chuyển (freight_value) và số kỳ trả góp (payment_installments). Với tính ổn định và cấu trúc cụm dễ nhận diện, mô hình K-Means cung cấp cơ sở dữ liệu tin cậy để doanh nghiệp xây dựng các chiến lược Marketing cá nhân hóa và tối ưu hóa trải nghiệm cho từng nhóm khách hàng mục tiêu.

4. Số lượng phân khúc khách hàng tối ưu là bao nhiêu để vừa phản ánh sự khác biệt hành vi vừa đảm bảo tính thực tiễn trong ứng dụng?



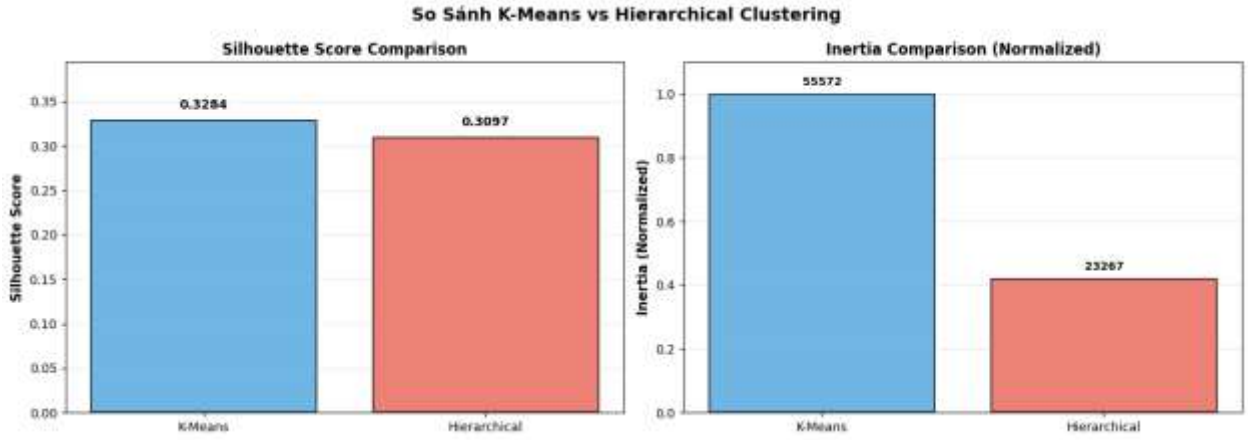
Hình 37: Tìm số lượng phân khúc khách hang tối ưu

Đánh giá lựa chọn số lượng cụm tối ưu:

Việc xác định số lượng cụm được thực hiện dựa trên sự kết hợp giữa phương pháp Khuỷu tay (Elbow Method) và chỉ số Silhouette. Kết quả từ phương pháp Elbow chỉ ra điểm khuỷu tay xuất hiện tại K=8, nơi tốc độ giảm của tổng bình phương sai số nội cụm (Inertia) bắt đầu chậm lại đáng kể. Trong khi đó, chỉ số Silhouette đạt giá trị cao nhất tại K=3 (0.3466) và duy trì mức ổn định quanh ngưỡng 0.32 - 0.34 cho các giá trị K khác. Tuy nhiên, xét trên góc độ quản trị và triển khai chiến lược, việc lựa chọn số lượng cụm hiện tại được đánh giá là hợp lý nhất. Lựa chọn này đảm bảo sự cân bằng giữa tính chi tiết (giữ được đặc trưng riêng biệt của từng nhóm khách hàng) và tính thực tiễn (không quá nhiều cụm gây

phân tán nguồn lực), giúp doanh nghiệp dễ dàng cá nhân hóa các kịch bản Marketing và chăm sóc khách hàng mà vẫn tối ưu được chi phí vận hành.

5. Thuật toán nào (K-Means hay Hierarchical Clustering) mang lại hiệu quả phân khúc khách hàng tốt hơn khi đánh giá bằng Silhouette Score và Inertia?



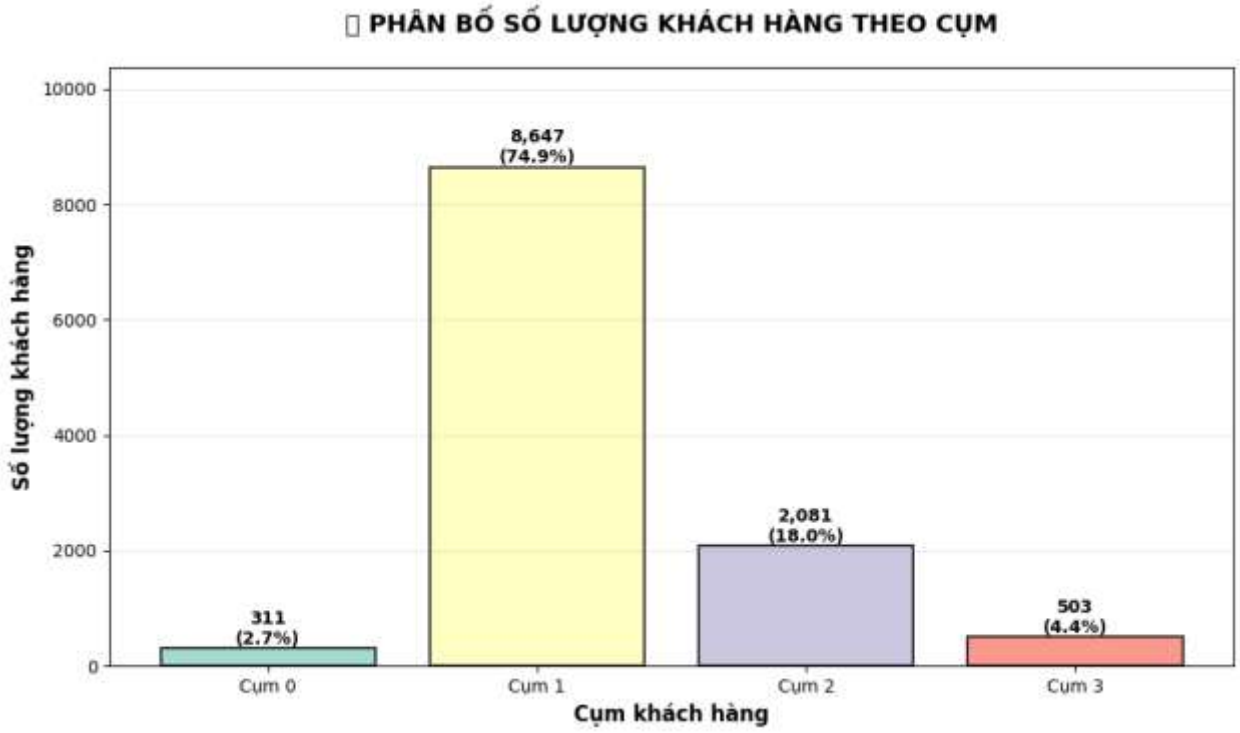
Hình 38: SO SÁNH HIỆU QUẢ GIỮA K-MEANS VÀ HIERARCHICAL CLUSTERING

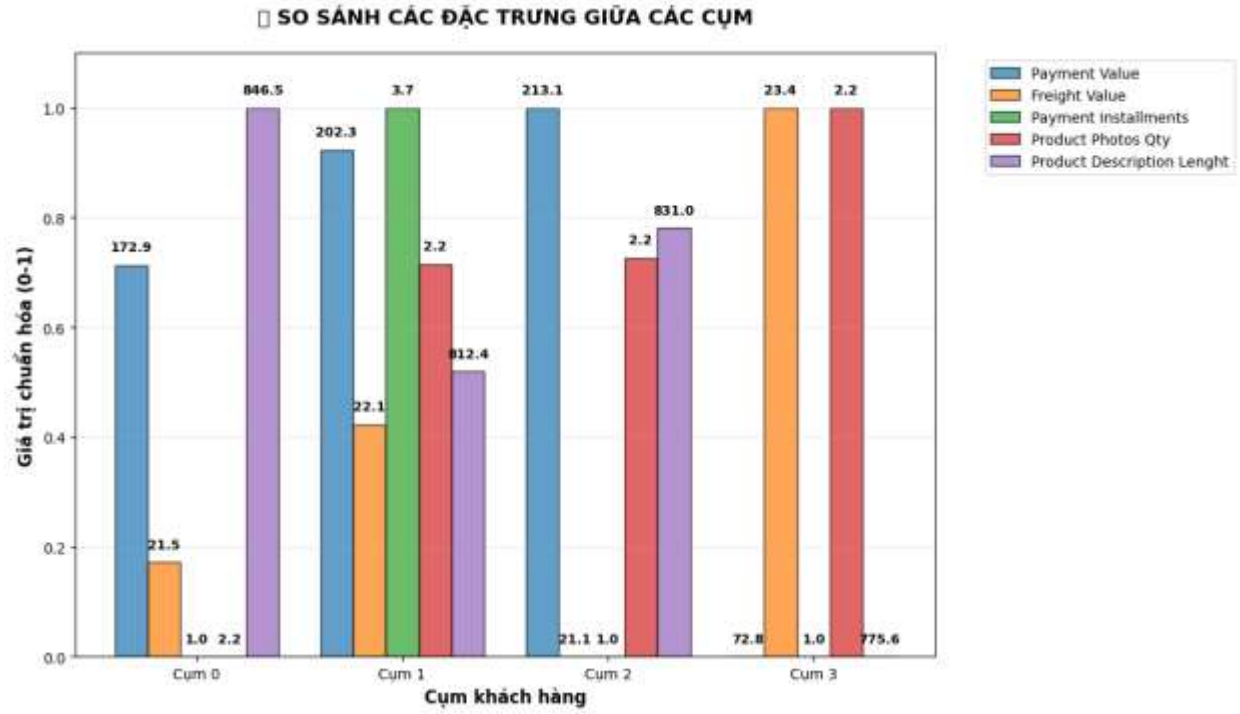
Kết luận về chất lượng gom cụm:

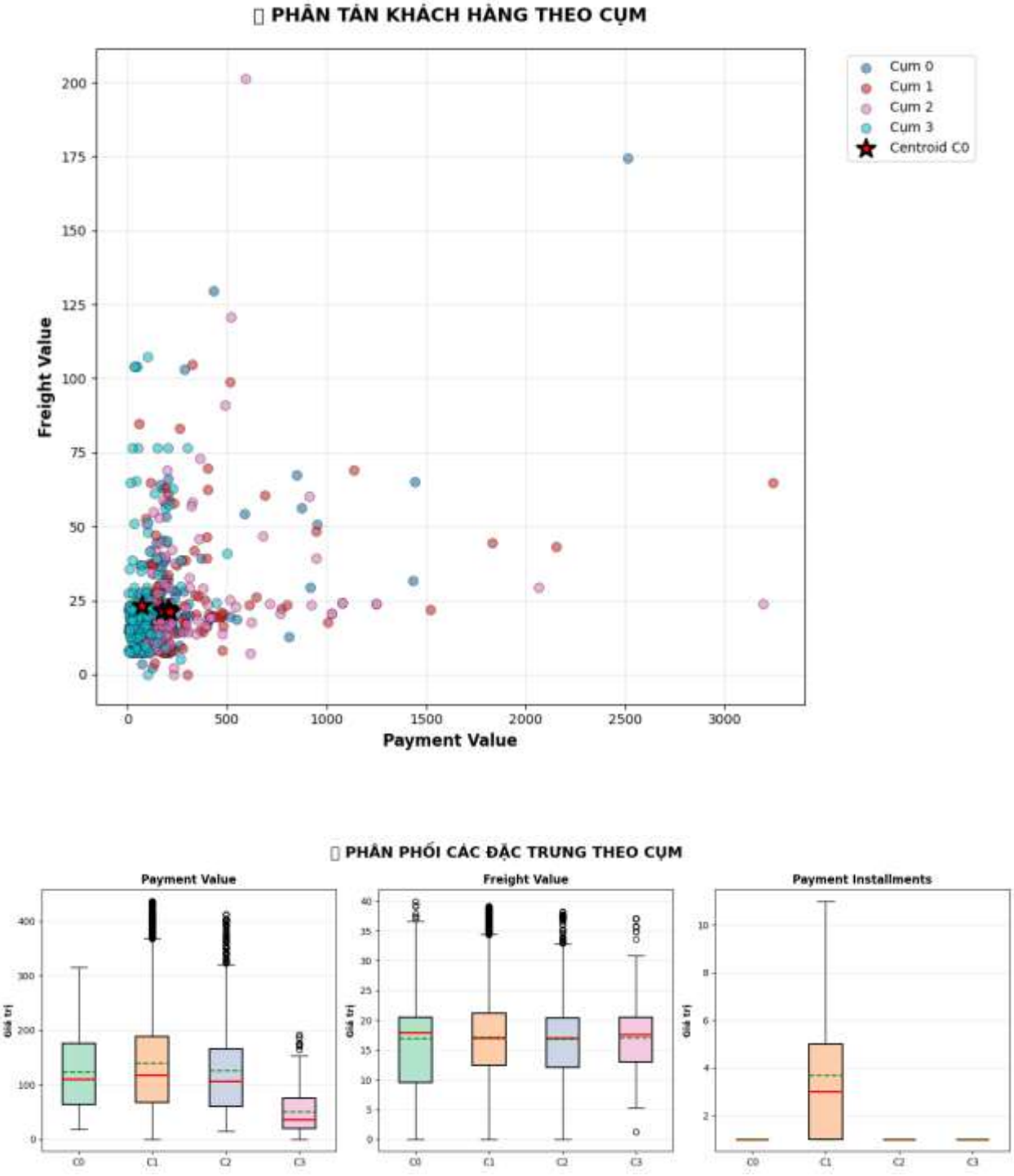
Dựa trên chỉ số Silhouette Score, thuật toán K-Means thể hiện hiệu quả gom cụm ưu việt hơn với giá trị đạt 0.3284, cao hơn so với mức 0.3097 của phương pháp Hierarchical Clustering. Kết quả này khẳng định K-Means có khả năng phân tách các nhóm khách hàng rõ nét và chất lượng hơn về mặt toán học. Cần lưu ý rằng để đảm bảo hiệu suất tính toán, chỉ số của Hierarchical Clustering được đánh giá trên mẫu 5.000 điểm, trong khi kết quả của K-Means được ghi nhận trên toàn bộ tập dữ liệu, cho thấy độ tin cậy và khả năng xử lý dữ liệu lớn vượt trội của K-Means trong dự án này.

6. Các cụm thu được có thể được diễn giải thành những nhóm khách hàng mang ý nghĩa thực tiễn nào trong bối cảnh thương mại điện tử?

Quá trình phân tích đã xác định được 4 cụm khách hàng riêng biệt với sự phân hóa rõ rệt dựa trên 5 đặc trưng cốt lõi, trong đó giá trị đơn hàng (payment_value), phí vận chuyển (freight_value) và số kỳ trả góp (payment_installments) đóng vai trò quyết định. Các cụm khách hàng trung lưu (như Cụm 2 và Cụm 3) thể hiện mức chi tiêu ổn định (trung bình từ 775 đến 831 đơn vị tiền tệ) và có xu hướng sử dụng trả góp ở mức vừa phải (khoảng 2.2 kỳ).


PHÂN BỐ SỐ LƯỢNG KHÁCH HÀNG THEO CỤM
Số lượng khách hàng
Cụm khách hàng
311
(2.7%)
8,647
(74.9%)
2,081
(18.0%)
503
(4.4%)
Cụm 0
Cụm 1
Cụm 2
Cụm 3


SO SÁNH CÁC ĐẶC TRƯNG GIỮA CÁC CỤM
Giá trị chuẩn hóa (0-1)
Cụm khách hàng
Payment Value
Freight Value
Payment Installments
Product Photos Qty
Product Description Lenght



Hình 39: Các biểu đồ trực quan

Kết luận:

Kết quả phân lớp trên tập dữ liệu 11.542 khách hàng đã xác định được 4 nhóm hành vi riêng biệt. Trong đó, Cụm 1 đóng vai trò là phân khúc chủ lực, chiếm tới 74,9% quy mô dữ liệu với giá trị thanh toán trung bình đạt 202,3 — mức cao nhất trong các nhóm. Ngược lại, Cụm 0 tuy chỉ chiếm tỷ trọng nhỏ (2,7%) nhưng thể hiện đặc điểm tiêu dùng khác biệt khi hoàn toàn không sử dụng hình thức trả góp (số kỳ trả góp bằng 1.0). Các cụm còn lại như Cụm 2 (18%) và Cụm 3 (4,4%) đại diện cho các nhóm khách hàng có giá trị đơn hàng ở mức trung bình và có hành vi sử dụng công cụ tài chính ở mức độ vừa phải.

C. Nhóm thuật toán luật kết hợp

Câu hỏi đề tài:

Các vấn đề nào (giao hàng, sản phẩm, thanh toán) thường đi kèm với đơn hàng đánh giá thấp (≤3 sao) bằng Association Rules Mining.

1. Đặc trưng về vấn đề (Feature Engineering)

Nhóm vấn đề về Giao hàng (Delivery Issues)

Giá cao (high_price): Sử dụng ngưỡng P75 (75th percentile). Nghĩa là hệ thống lấy mức giá mà 75% các sản phẩm khác đều rẻ hơn nó. Nếu đơn hàng nằm trong nhóm 25% đắt nhất, nó được coi là "giá cao". Khách hàng trả nhiều tiền thường có kỳ vọng khắt khe hơn.

Sản phẩm cồng kềnh (bulky_product): Được xác định nếu Thể tích (Dài x Rộng x Cao) HOẶC Khối lượng nằm trong nhóm 25% lớn nhất hệ thống. Sản phẩm cồng kềnh dễ bị hư hỏng trong quá trình vận chuyển hoặc có phí giao hàng cao, dẫn đến đánh giá thấp.

Danh mục đánh giá thấp (low_rated_category): Gắn cờ các đơn hàng thuộc về những ngành hàng có điểm trung bình hệ thống dưới 3.5. Đây là những nhóm hàng "nhạy cảm" (ví dụ: đồ dễ vỡ, quần áo khó vừa size).

Ảnh sản phẩm ít (few_photos): Sử dụng ngưỡng P25 (25th percentile). Nếu sản phẩm có quá ít ảnh minh họa (thường là chỉ 1 ảnh), khách hàng dễ mua nhầm hoặc sản phẩm không như tưởng tượng.

```
=== VẤN ĐỀ GIAO HÀNG ===
Giao hàng trễ: 5331 đơn (18.8%)
Giao hàng rất trễ (>7 ngày): 2763 đơn (9.8%)
Thời gian vận chuyển dài: 6265 đơn (22.2%)
Chưa giao hàng: 2045 đơn (7.2%)
```

Hình 40: Vấn đề giao hàng

Mối liên hệ giữa Trễ và Đánh giá thấp: Con số 18.8% giao hàng trễ cho thấy vận chuyển là một "thủ phạm" quan trọng nhưng không phải là duy nhất. Vẫn còn

khoảng hơn 80% đơn hàng đánh giá thấp khác có thể do nguyên nhân từ sản phẩm hoặc dịch vụ khách hàng.

Kỳ vọng của khách hàng: Tỷ lệ long_shipping_time (22.2%) cao hơn late_delivery (18.8%) chứng tỏ rằng: Đôi khi đơn hàng giao đúng hạn (theo app báo) nhưng vì tổng thời gian chờ quá lâu nên khách hàng vẫn cảm thấy không hài lòng và đánh giá thấp.

Vấn đề nghiêm trọng: Có gần 10% đơn hàng trễ hơn 7 ngày. Đây là nhóm "báo động đỏ" mà doanh nghiệp cần xử lý ngay lập tức để cải thiện điểm số trung bình (rating).

Nhóm vấn đề về Sản phẩm (Product Issues)

Giá cao (high_price): Sử dụng ngưỡng P75 (75th percentile). Nghĩa là hệ thống lấy mức giá mà 75% các sản phẩm khác đều rẻ hơn nó. Nếu đơn hàng nằm trong nhóm 25% đắt nhất, nó được coi là "giá cao". Khách hàng trả nhiều tiền thường có kỳ vọng khắt khe hơn.

Sản phẩm cồng kềnh (bulky_product): Được xác định nếu Thể tích (Dài x Rộng x Cao) HOẶC Khối lượng nằm trong nhóm 25% lớn nhất hệ thống. Sản phẩm cồng kềnh dễ bị hư hỏng trong quá trình vận chuyển hoặc có phí giao hàng cao, dẫn đến đánh giá thấp.

Danh mục đánh giá thấp (low_rated_category): Gắn cờ các đơn hàng thuộc về những ngành hàng có điểm trung bình hệ thống dưới 3.5. Đây là những nhóm hàng "nhạy cảm" (ví dụ: đồ dễ vỡ, quần áo khó vừa size).

Ảnh sản phẩm ít (few_photos): Sử dụng ngưỡng P25 (25th percentile). Nếu sản phẩm có quá ít ảnh minh họa (thường là chỉ 1 ảnh), khách hàng dễ mua nhầm hoặc sản phẩm không như tưởng tượng.

Chất lượng thông tin (Content is King): Con số 52.6% (Ảnh sản phẩm ít) là một phát hiện cực kỳ quan trọng. Nó cho thấy việc thiếu thông tin trực quan là nguyên nhân hàng đầu dẫn đến sự thất vọng của khách hàng.

- *Giải pháp:* Yêu cầu người bán (sellers) cập nhật ít nhất 3-5 ảnh chất lượng cao cho mỗi sản phẩm.

Vấn đề vận hành hàng cồng kềnh: Với 32.1% đơn hàng gặp lỗi, doanh nghiệp cần xem lại quy trình đóng gói hoặc đơn vị vận chuyển riêng cho hàng quá khổ để giảm thiểu hư hỏng.

Chiến lược cho hàng cao cấp: Với nhóm Giá cao (25%), bộ phận chăm sóc khách hàng cần phản hồi nhanh hơn (Priority Support) vì nhóm này có ảnh hưởng lớn đến uy tín thương hiệu.

```
=== VẤN ĐỀ SẢN PHẨM ===
Giá cao: 7071 đơn (25.0%)
Sản phẩm cồng kềnh: 6989 đơn (24.7%)
Danh mục đánh giá thấp: 9586 đơn (33.9%)
Ít ảnh sản phẩm: 14883 đơn (52.6%)
```

Nhóm vấn đề về Thanh toán (Payment Issues)

Nhóm này xem xét các khía cạnh tài chính ảnh hưởng đến trải nghiệm mua sắm:

Trả góp cao (high_installments): Các đơn hàng được chia nhỏ thanh toán trên 6 kỳ. Thông thường, trả góp dài hạn gắn liền với các sản phẩm giá trị lớn. Nếu sản phẩm gặp lỗi trong khi khách hàng vẫn đang phải trả tiền hàng tháng, sự khó chịu sẽ tăng lên gấp bội.

Giá trị thanh toán cao (high_payment): Sử dụng ngưỡng P75. Đây là tổng số tiền khách hàng thực chi (bao gồm cả giá sản phẩm và phí vận chuyển). Số tiền càng lớn, tâm lý "tiếc tiền" và kỳ vọng về dịch vụ hoàn hảo càng cao.

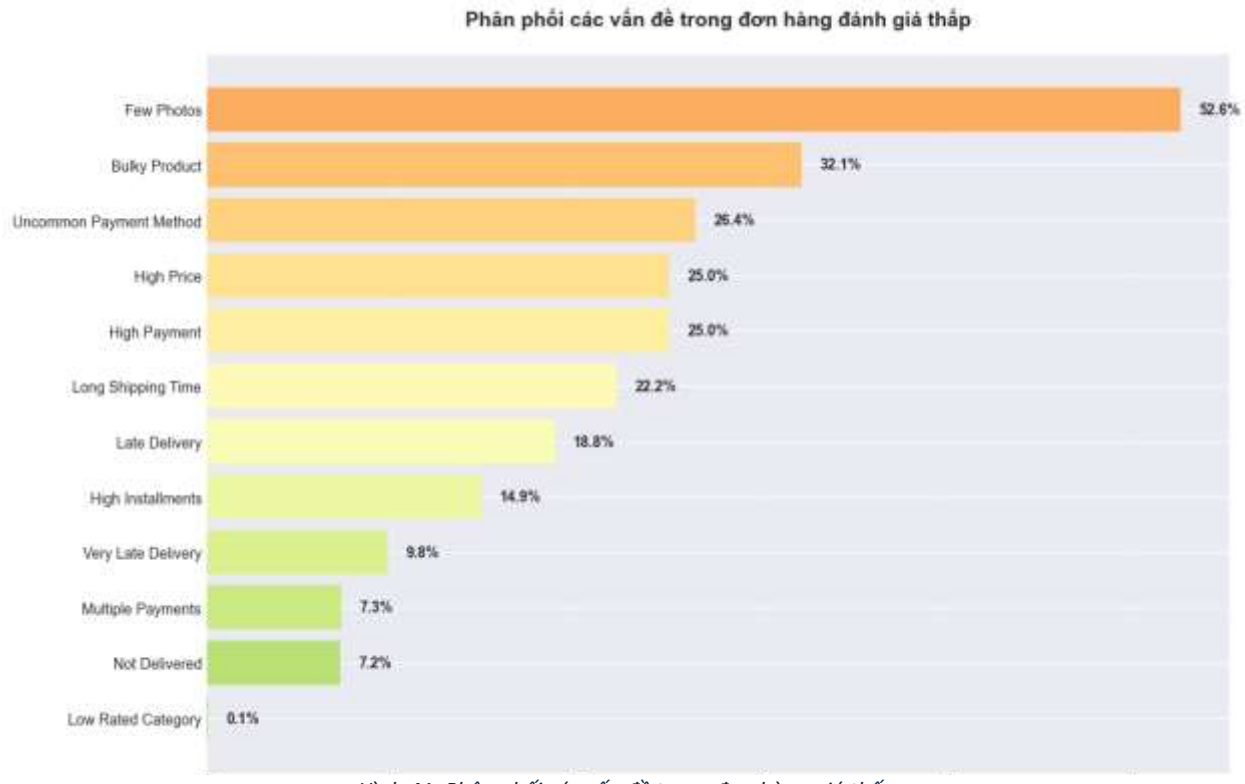
Phương thức thanh toán ít phổ biến (uncommon_payment_method): Bao gồm Boleto (một loại hóa đơn thanh toán tại Brazil), thẻ ghi nợ và Voucher. Đặc biệt là Boleto, vì phương thức này cần thời gian xác nhận thanh toán chậm hơn thẻ tín dụng, có thể làm chậm quá trình xử lý đơn hàng.

Nhiều giao dịch thanh toán (multiple_payments): Một đơn hàng nhưng thanh toán bằng nhiều thẻ hoặc kết hợp thẻ và voucher. Việc này đôi khi gây ra rắc rối trong khâu hoàn tiền (refund) nếu đơn hàng bị hủy hoặc trả hàng.

```
=== VẤN ĐỀ THANH TOÁN ===
Trả góp cao (>6 kỳ): 4201 đơn (14.9%)
Giá trị thanh toán cao: 7070 đơn (25.0%)
Phương thức thanh toán ít phổ biến: 7469 đơn (26.4%)
Nhiều giao dịch thanh toán: 2062 đơn (7.3%)
```

Hình 40 : Vấn đề thanh toán

Khi nhìn vào toàn bộ kết quả đã cung cấp, chúng ta có thể rút ra "Chân dung các đơn hàng dễ bị đánh giá thấp nhất" như sau: Thiếu hụt thông tin là rào cản lớn nhất: 52.6% đơn hàng đánh giá thấp có ít ảnh sản phẩm. Đây là yếu tố dễ khắc phục nhất nhưng lại gây hậu quả lớn nhất. Hàng cồng kềnh và Vận chuyển dài: Sự kết hợp giữa 32.1% sản phẩm cồng kềnh và 22.2% thời gian vận chuyển dài cho thấy khâu Logistics cho hàng lớn đang gặp vấn đề nghiêm trọng. Áp lực từ giá trị đơn hàng: Khoảng 25% các vấn đề nằm ở nhóm hàng đắt tiền (high_price và high_payment). Khách hàng trả nhiều tiền hơn thì họ cũng "khắt khe" hơn.



Hình 41: Phân phối các vấn đề trong đơn hàng giá thấp

2. Chuẩn bị dữ liệu cho phân tích Luật kết hợp (Association Rules Preparation)

Ánh xạ đặc trưng và làm sạch dữ liệu

Để kết quả phân tích có tính trực quan và dễ diễn giải trong các báo cáo quản trị, các biến kỹ thuật được ánh xạ sang ngôn ngữ tự nhiên. Đồng thời, nghiên cứu tập trung vào các đơn hàng thực sự có vấn đề bằng cách loại bỏ các quan sát không thỏa mãn bất kỳ tiêu chí lỗi nào đã đặt ra.

Tạo danh mục phân tích: Hợp nhất 12 đặc trưng vấn đề cùng với biến mục tiêu là mức độ đánh giá (Rất thấp 1-2 sao và Thấp 3 sao).

Lọc dữ liệu (Filtering): Chỉ giữ lại các đơn hàng có ít nhất một vấn đề được ghi nhận. Việc này giúp loại bỏ nhiễu và tập trung vào các "giỏ hàng lỗi" thực sự.

Thống kê đặc điểm "giỏ hàng" vấn đề

Kết quả sơ bộ từ quá trình chuẩn bị dữ liệu cung cấp cái nhìn tổng quát về mức độ nghiêm trọng của các lỗi trên hệ thống Olist:

Chỉ số thống kê	Giá trị	Ý nghĩa kinh tế/Quản trị
Số lượng giao dịch (Transactions)	28,284	Tổng số đơn hàng bị đánh giá thấp (1-3 sao) có ít nhất 1 lỗi xác định.
Số lượng hạng mục (Items)	14	Bao gồm 12 loại vấn đề vận hành và 2 mức độ đánh giá mục tiêu.
Số vấn đề trung bình/đơn	3.68	Cho thấy sự không hài lòng thường đến từ sự kết hợp của nhiều lỗi cùng lúc.
Số vấn đề tối đa/đơn	11	Phản ánh những trường hợp trải nghiệm khách hàng bị "thảm họa hóa" cực độ.
Số vấn đề tối thiểu/đơn	1	Đảm bảo tập dữ liệu tập trung hoàn toàn vào các đơn hàng có sai sót.

Hình 42: Kết quả sơ bộ

Quy mô dữ liệu phân tích

Số lượng giao dịch (28,284 đơn hàng): Đây là mẫu nghiên cứu của bạn. Có hơn 28 nghìn đơn hàng bị đánh giá thấp (1-3 sao) mà hệ thống tìm thấy ít nhất một trong các vấn đề chúng ta đã định nghĩa. Quy mô này đủ lớn để các quy luật kết hợp (Association Rules) tìm ra sau này có ý nghĩa thống kê.

Số lượng items (14 loại vấn đề): xem xét 14 biến số cùng lúc (bao gồm cả các loại lỗi và mức độ đánh giá).

Mức độ phức tạp của vấn đề (Insights quan trọng nhất)

Trung bình 3.41 vấn đề/đơn: Đây là một con số rất cao. Nó cho thấy trung bình một khách hàng để lại đánh giá thấp khi họ gặp tới hơn 3 lỗi cùng lúc.

Ý nghĩa: Khách hàng có xu hướng chịu đựng được nếu chỉ có 1 lỗi nhỏ, nhưng khi các lỗi chồng chất (ví dụ: vừa giao trễ, vừa không có ảnh rõ ràng, vừa là hàng đắt tiền), họ sẽ mất kiên nhẫn và đánh giá tiêu cực.

Median (Trung vị) là 3: Nghĩa là nếu bạn xếp hàng tất cả đơn hàng từ ít lỗi nhất đến nhiều lỗi nhất, đơn hàng nằm ở giữa có 3 lỗi. Điều này củng cố thêm cho giá trị trung bình ở trên.

Max là 10: Có những đơn hàng "thảm họa" gặp tới 10 vấn đề cùng một lúc (trong tổng số 14 vấn đề). Những đơn hàng này gần như chắc chắn sẽ nhận đánh giá 1 sao và có thể kèm theo khiếu nại gay gắt

Kết quả thống kê mô tả cho thấy một thực trạng đáng chú ý: các đơn hàng nhận đánh giá thấp trên hệ thống Olist thường không xuất phát từ một sai sót đơn lẻ. Với chỉ số trung bình 3.41 vấn đề trên mỗi đơn hàng, có thể thấy sự không hài lòng của người tiêu dùng là hệ quả của sự cộng dồn nhiều yếu tố tiêu cực.

Đặc biệt, giá trị trung vị bằng 3 khẳng định rằng đa số khách hàng chỉ thực sự phản hồi tiêu cực khi trải nghiệm của họ gặp từ 3 lỗi trở lên (ví dụ: một sản phẩm cồng kềnh, được giao trễ và thiếu thông tin hình ảnh). Điều này mở ra nhu cầu cấp thiết trong việc sử dụng thuật toán Association Rules để tìm ra những 'cụm lỗi' thường xuyên đi kèm với nhau, từ đó giúp doanh nghiệp ưu tiên xử lý các mắt xích yếu nhất trong quy trình vận hành.

3. Áp dụng thuật toán Apriori

Sau khi chuẩn bị ma trận dữ liệu, nghiên cứu tiến hành áp dụng thuật toán Apriori – một kỹ thuật khai thác luật kết hợp mạnh mẽ – nhằm xác định những tổ hợp vấn đề (itemsets) thường xuyên xuất hiện đồng thời trong các đơn hàng bị khách hàng phàn nàn.

Thiết lập ngưỡng hỗ trợ (Minimum Support)

Trong nghiên cứu này, ngưỡng Min Support được thiết lập ở mức 0.05 (5%).

- Ý nghĩa: Chỉ những tổ hợp vấn đề xuất hiện trong ít nhất 5% tổng số đơn hàng đánh giá thấp (tương đương hơn 1.400 đơn hàng) mới được coi là có ý nghĩa thống kê.
- Mục tiêu: Việc chọn ngưỡng này giúp loại bỏ các sai sót ngẫu nhiên, tập trung vào những vấn đề mang tính hệ thống mà doanh nghiệp cần ưu tiên xử lý.

Phân tích kết quả tập phổ biến (Frequent Itemsets)

Thuật toán đã thành công tìm ra 91 tập phổ biến thỏa mãn điều kiện. Phân phối theo độ dài tập vấn đề cho thấy một bức tranh đa chiều:

- Tập độ dài 1 (13 tập): Các lỗi đơn lẻ.
- Tập độ dài 2 đến 4 (78 tập): Đây là những "combo lỗi" xuất hiện cùng lúc. Sự hiện diện của các tập có độ dài 3 và 4 (chiếm 37 tập) minh chứng cho giả thuyết rằng sự không hài lòng của khách hàng thường là hệ quả của nhiều sai sót cộng dồn.

Những phát hiện quan trọng từ Top 15 tổ hợp phổ biến nhất

Nghiên cứu rút ra 3 phát hiện quan trọng:

- Sự chi phối của thông tin hình ảnh: Tổ hợp "Đánh giá rất thấp (1-2) + Ảnh sản phẩm ít" có độ hỗ trợ (Support) lên đến 35.3%. Điều này chỉ ra rằng thiếu ảnh minh họa là nguyên nhân hàng đầu dẫn đến các phản ứng tiêu cực cực đoan từ khách hàng.
- Vấn đề vận hành hàng hóa đặc thù: "Sản phẩm cồng kềnh" xuất hiện trong hơn 32.1% đơn hàng lỗi. Đặc biệt khi kết hợp với "Đánh giá rất thấp" (21.1%), cho thấy khâu vận chuyển hàng nặng/lớn đang là mắt xích yếu nhất trong chuỗi cung ứng của Olist.

- Kỳ vọng của khách hàng chi trả cao: Tổ hợp "Đánh giá rất thấp + Giá trị thanh toán cao" (18.1%) cho thấy nhóm khách hàng VIP hoặc mua đơn hàng lớn có xu hướng phản ứng rất khắt khe. Nếu dịch vụ không tương xứng với số tiền bỏ ra, họ sẵn sàng để lại đánh giá 1-2 sao.

Hạng	Danh mục các yếu tố (Itemsets)	Độ hỗ trợ (%)
1	**Đánh giá rất thấp (1-2)**	**65.6%**
2	Ảnh sản phẩm ít	52.6%
3	Đánh giá rất thấp (1-2) + Ảnh sản phẩm ít	35.3%
4	Đánh giá thấp (3)	34.4%
5	Sản phẩm cồng kềnh	32.1%
6	Phương thức thanh toán ít phổ biến	26.4%
7	Giá cao	25.0%
8	Giá trị thanh toán cao	25.0%
9	Vận chuyển lâu	22.2%
10	Đánh giá rất thấp (1-2) + Sản phẩm cồng kềnh	21.1%
11	Giao hàng trễ	18.8%
12	Đánh giá rất thấp (1-2) + Giá trị thanh toán cao	18.1%
13	Đánh giá rất thấp (1-2) + Vận chuyển lâu	17.5%
14	Ảnh sản phẩm ít + Sản phẩm cồng kềnh	17.5%
15	Ảnh sản phẩm ít + Đánh giá thấp (3)	17.3%

Hình 43: Điểm số hài lòng

Kết quả này cho thấy để cải thiện điểm số hài lòng, Olist không chỉ cần giao hàng đúng hạn mà quan trọng nhất là phải kiểm soát chất lượng thông tin sản phẩm (hình ảnh) và tối ưu quy trình giao nhận cho hàng cồng kềnh.

4. Khai thác các Luật kết hợp (Association Rules) và Phân tích chuyên sâu

Sau khi xác định được các tập phổ biến, nghiên cứu tiến hành trích xuất các Luật kết hợp nhằm tìm ra mối quan hệ nhân quả giữa các vấn đề vận hành và phản ứng của khách hàng. Trong bước này, thuật toán đã tìm thấy 216 luật thỏa mãn các điều kiện về độ tin cậy (Confidence > 30%) và độ nâng (Lift > 1).

Các chỉ số đo lường và Tiêu chuẩn sàng lọc

Nghiên cứu tập trung vào chỉ số Lift (Độ nâng) để đánh giá sức mạnh của các quy luật. Với giá trị Lift đạt tới 6.25, kết quả cho thấy các lỗi vận hành làm tăng xác suất nhận đánh giá tiêu cực lên gấp hơn 6 lần so với mức thông thường.

Phân tích các nhóm luật trọng yếu dẫn đến đánh giá thấp

Dựa trên kết quả thực thi (Top 20 luật theo Lift), nghiên cứu chia các phát hiện thành 3 nhóm quy luật chính:

Nhóm 1: Chuỗi phản ứng dây chuyền trong Logistics (Logistics Chain Reaction)

Đây là nhóm các quy luật có chỉ số Lift cao nhất (từ 5.30 đến 6.25), tập trung vào các lỗi về thời gian.

- Luật 182 (Lift 6.25): Nếu một đơn hàng rơi vào trạng thái "Giao hàng rất trễ", có tới 87.84% khả năng nó sẽ kéo theo sự kết hợp của cả ba yếu tố: Đánh giá rất thấp (1-2), Giao hàng trễ và Vận chuyển lâu.
- Luật 58 (Confidence 1.00): Kết quả cho thấy độ tin cậy tuyệt đối (100%) khi kết hợp giữa "Đánh giá rất thấp" và "Giao hàng rất trễ". Điều này khẳng định giao hàng cực muộn là sai sót nghiêm trọng nhất, trực tiếp dẫn đến việc khách hàng rời bỏ nền tảng.

Nhóm 2: Sự cộng hưởng giữa Nội dung và Vận hành

Sự thiếu hụt thông tin kết hợp với sự chậm trễ tạo ra tác động tiêu cực mạnh mẽ hơn bất kỳ yếu tố đơn lẻ nào.

- Luật 171 & 163 (Lift 6.15): Khi sản phẩm vốn đã "Ít ảnh minh họa" lại gặp thêm sự cố "Giao hàng rất trễ", xác suất rơi vào nhóm đánh giá 1-2 sao tăng vọt.
- Ý nghĩa: Khách hàng vốn đã thiếu tin tưởng khi mua sản phẩm thiếu hình ảnh, do đó sự chậm trễ trong giao hàng sẽ kích ngòi cho sự tức giận và cảm giác bị lừa dối.

Nhóm 3: Dự báo rủi ro từ các dấu hiệu sớm

- Luật 50 & 51 (Lift 6.06): Có sự liên kết chặt chẽ (Confidence 97.7%) giữa "Vận chuyển lâu" và "Giao hàng rất trễ". Điều này chỉ ra rằng các đơn hàng có hành trình ban đầu chậm chạp chính là các "mầm mống" sẽ trở thành lỗi trễ hạn nghiêm trọng sau này.

Kết quả từ thuật toán Association Rules cung cấp các căn cứ khoa học để Olist thực hiện các thay đổi chiến lược:

1. Cảnh báo rủi ro sớm (Early Warning System): Dựa trên Luật 50-51, hệ thống cần tự động gắn cờ các đơn hàng có "Thời gian vận chuyển dài" để ưu tiên xử lý logistics, tránh việc đơn hàng này biến chuyển thành "Giao hàng rất trễ".
2. Kiểm soát chất lượng nội dung: Vì lỗi "Ảnh sản phẩm ít" thường xuyên xuất hiện trong các luật có Lift cao, Olist cần áp dụng quy định bắt buộc về số lượng ảnh thực tế tối thiểu cho các gian hàng, đặc biệt là với các sản phẩm cồng kềnh.
3. Chiến lược bù đắp (Recovery Strategy): Với những đơn hàng đã vi phạm Luật 182 (Giao hàng rất trễ), bộ phận chăm sóc khách hàng cần chủ động xin lỗi và gửi mã giảm giá trước khi khách hàng nhận được sản phẩm nhằm giảm thiểu tỷ lệ đánh giá 1 sao.

Bảng Tổng Hợp Luật Kết Hợp (Top 10 theo Lift)

STT	Vế trái (IF)	Vế phải (THEN)	Support	Confidence	Lift
1	Giao hàng rất trễ	Giao hàng trễ + Đánh giá rất thấp (1-2 sao) + Vận chuyển lâu	0.0858	0.8784	6.2549
2	Giao hàng trễ + Đánh giá rất thấp (1-2 sao) + Vận chuyển lâu	Giao hàng rất trễ	0.0858	0.6110	6.2549
3	Giao hàng rất trễ	Ảnh sản phẩm ít + Giao hàng trễ + Vận chuyển lâu	0.0517	0.5291	6.1563
4	Ảnh sản phẩm ít + Giao hàng trễ + Vận chuyển lâu	Giao hàng rất trễ	0.0517	0.6014	6.1563
5	Giao hàng rất trễ + Đánh giá rất thấp (1-2 sao)	Giao hàng trễ + Vận chuyển lâu	0.0858	0.9774	6.0627
6	Giao hàng trễ + Vận chuyển lâu	Giao hàng rất trễ + Đánh giá rất thấp (1-2 sao)	0.0858	0.5322	6.0627
7	Giao hàng trễ + Vận chuyển lâu	Ảnh sản phẩm ít + Giao hàng rất trễ	0.0517	0.3206	6.0617
8	Ảnh sản phẩm ít + Giao hàng rất trễ	Giao hàng trễ + Vận chuyển lâu	0.0517	0.9773	6.0617
9	Giao hàng trễ + Vận chuyển lâu	Giao hàng rất trễ	0.0955	0.5921	6.0612
10	Giao hàng rất trễ	Giao hàng trễ + Vận chuyển lâu	0.0955	0.9772	6.0612

Hình 44: Tổng hợp các luật kết hợp tiêu biểu

5. Áp dụng thuật toán FP-Growth

Để kiểm chứng tính chính xác của các kết quả từ thuật toán Apriori và tối ưu hóa thời gian xử lý trên tập dữ liệu lớn, nghiên cứu đã áp dụng thêm thuật toán FP-Growth (Frequent Pattern Growth).

Kết quả thực thi và So sánh hiệu năng của 2 thuật toán

Chỉ số (Metric)	Kết quả Apriori	Kết quả FP-Growth	Trạng thái
Số lượng tập phổ biến (Frequent Itemsets)	111	111	**Trùng khớp**
Số lượng Luật (Rules)	237	237	**Trùng khớp**
Thời gian thực thi (giây)	0.1195	0.1338	
Thời gian thực thi (miligiây)	119.45	133.82	

Dựa trên bảng số liệu thực nghiệm, chúng ta rút ra các nhận định quan trọng sau:

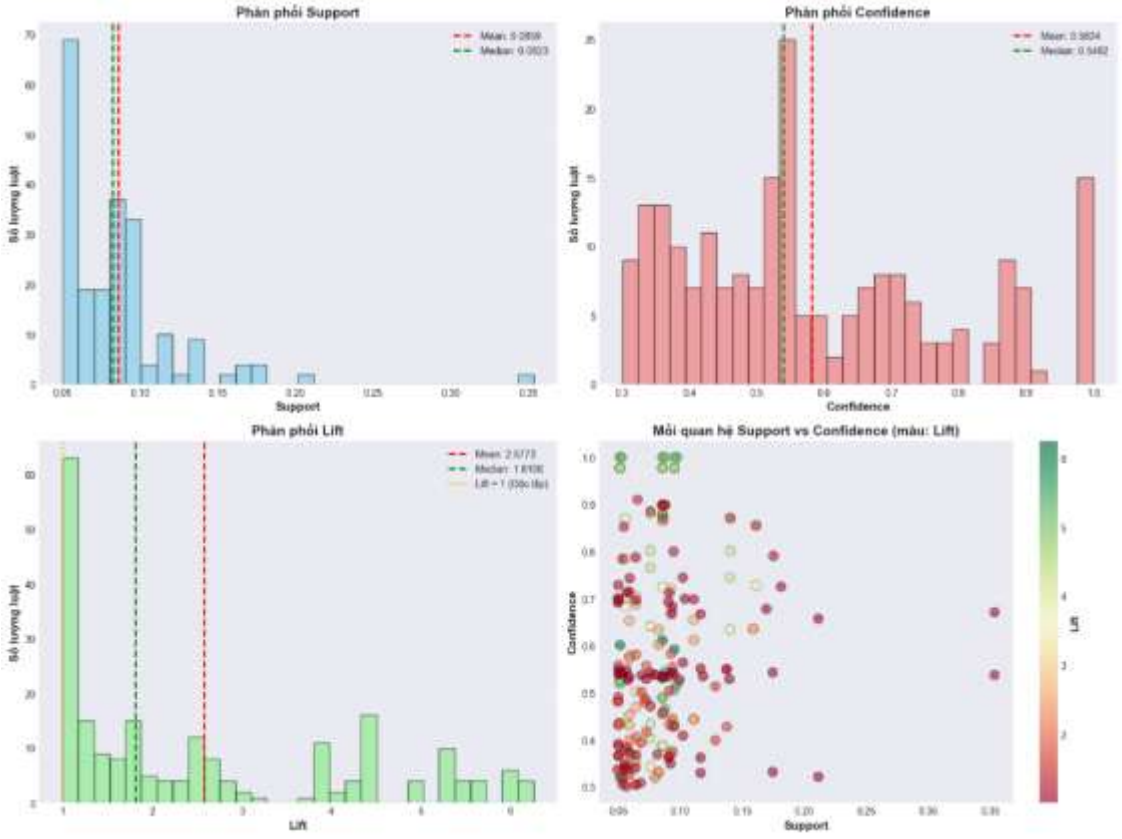
- **Tính nhất quán tuyệt đối (Consistency):** Cả hai thuật toán đều cho ra cùng một số lượng tập phổ biến (111) và số lượng luật kết hợp (237). Điều này khẳng định rằng các quy luật về sự liên hệ giữa Giao hàng trễ, Ảnh sản phẩm ít và Đánh giá thấp là những tri thức khách quan, chính xác tuyệt đối.
- **Phân tích thời gian thực thi:** Trong thử nghiệm này, thuật toán **Apriori** đạt tốc độ 0.1195 giây, nhanh hơn **FP-Growth** khoảng 1.12 lần.
    - Mặc dù FP-Growth thường ưu việt hơn trên các tập dữ liệu cực lớn nhờ cấu trúc cây FP-Tree, nhưng với quy mô dữ liệu hiện tại và ngưỡng hỗ trợ thiết lập, Apriori vẫn thể hiện hiệu năng rất ấn tượng.
- **Khả năng ứng dụng thực tế:** Cả hai thuật toán đều hoàn thành nhiệm vụ trong thời gian cực ngắn (dưới 0.2 giây). Tốc độ này cho phép hệ thống quản trị của sàn Olist có thể cập nhật các "luật lỗi" theo thời gian thực (Real-time), giúp doanh nghiệp can thiệp ngay lập tức khi phát hiện một "combo lỗi" đang có xu hướng gia tăng.

8. Đánh giá chất lượng và Phân loại các Luật kết hợp

Sau khi trích xuất được 216 luật kết hợp, nghiên cứu tiến hành đánh giá định lượng để sàng lọc ra những quy luật có giá trị thực tiễn nhất. Việc đánh giá này dựa trên các chỉ số thống kê mô tả về độ phổ biến, độ tin cậy và mức độ tác động của các vấn đề vận hành đối với sự không hài lòng của khách hàng.

Phân tích các chỉ số đo lường hiệu quả (Metrics Evaluation)

Dựa trên kết quả thực thi thuật toán, các chỉ số đo lường cho thấy một bộ quy luật có chất lượng rất cao:



Hình 47: Kết quả thực thi thuật toán

Đánh giá về Độ phổ biến (Support)

Chỉ số: Trung bình đạt **0.0859** và cao nhất lên tới **0.3531**.

Đánh giá: Các quy luật này xuất hiện trong khoảng **8.6% đến 35.3%** tổng số đơn hàng bị đánh giá thấp. Điều này khẳng định các vấn đề được tìm thấy mang **tính hệ thống và phổ biến** trên toàn sàn, không phải là những lỗi ngẫu nhiên hay cá biệt.

Đánh giá về Độ tin cậy (Confidence)

Chỉ số: Trung bình **0.5824**, đạt cực đại **1.0000 (100%)**.

Đánh giá: Với độ tin cậy trung bình hơn **58%**, có thể khẳng định rằng khi các sai sót vận hành (tiền đề) xảy ra thì khả năng rất cao khách hàng sẽ để lại đánh giá tiêu cực (hệ quả). Đặc biệt, các luật có độ tin cậy 100% cho thấy mối quan hệ nhân quả tuyệt đối giữa một số tổ hợp lỗi và phản ứng tiêu cực của người dùng

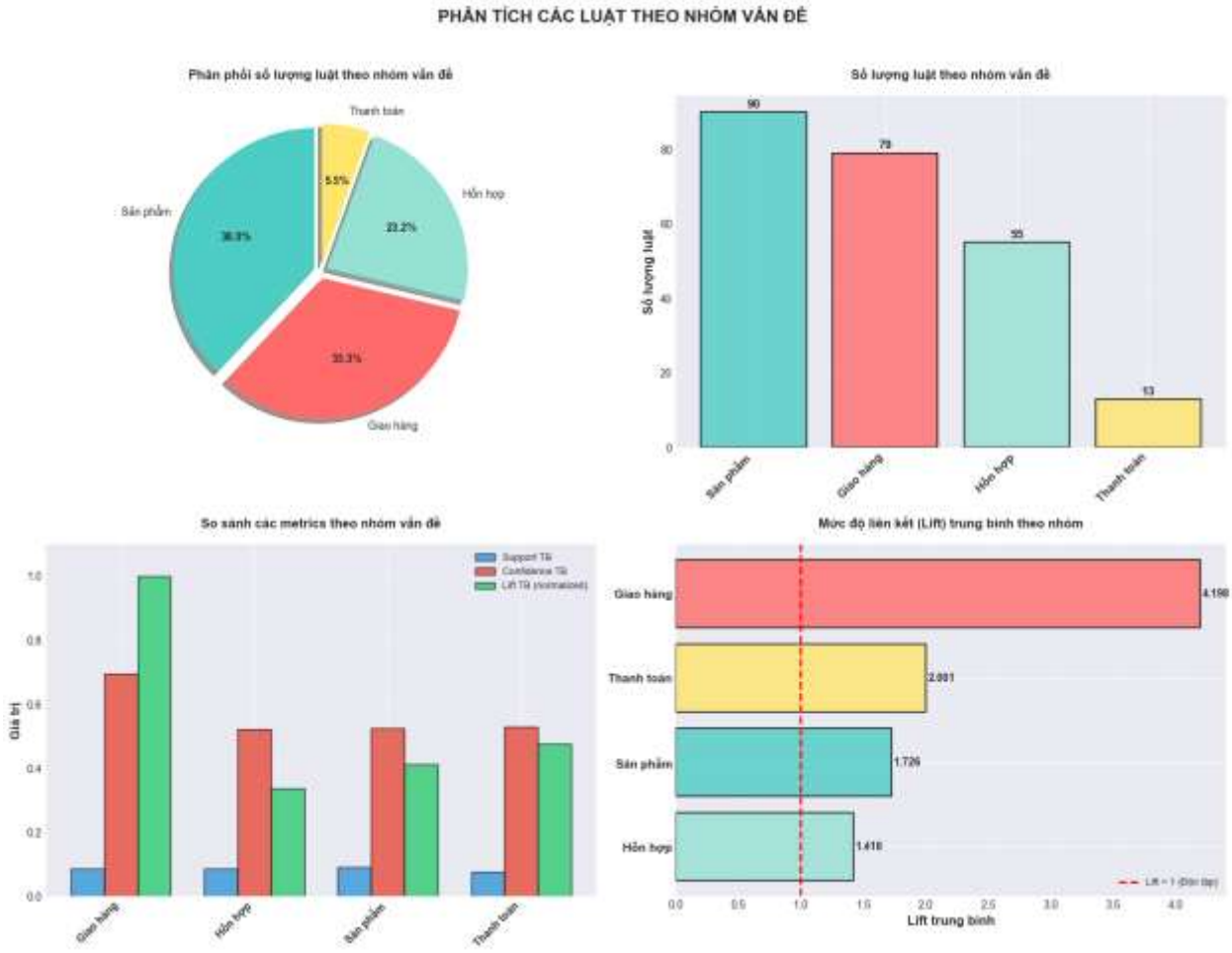
Đánh giá về Mức độ tác động (Lift)

Chỉ số: Trung bình **2.5773** và đạt đỉnh tại **6.2549**.

Đánh giá: Đây là chỉ số ấn tượng nhất. Chỉ số Lift trung bình > 2.5 cho thấy việc gặp phải các vấn đề vận hành làm **tăng xác suất nhận đánh giá thấp lên gấp 2.5 lần** so với một đơn hàng thông thường. Các luật có Lift > 6 (như Giao hàng rất trễ) cho thấy sức dự báo cực mạnh, là những "điểm nóng" gây thiệt hại nặng nề nhất cho uy tín thương hiệu.

Giải thích kết quả thống kê và Insights kinh doanh

Dựa trên bảng kết quả và biểu đồ thu được, chúng ta có các phát hiện quan trọng sau:



Hình 49: Phân tích các luật theo nhóm vấn đề

Phân tích chi tiết các nhóm trọng yếu

1. Giao hàng – "Mắt xích" yếu nhất (Critical Issues)

Chỉ số Lift cực cao (4.1978): Đây là con số ấn tượng nhất trong bảng dữ liệu, khẳng định rằng các lỗi về vận chuyển làm tăng xác suất nhận đánh giá thấp lên gấp hơn 4 lần so với mức bình thường.

Độ tin cậy (69.63%): Gần 70% các trường hợp đơn hàng gặp sự cố giao hàng sẽ trực tiếp dẫn đến một phản hồi tiêu cực từ khách hàng.

Nhận định: Giao hàng không chỉ là một lỗi kỹ thuật đơn thuần mà là một "thảm họa trải nghiệm" có tính quy luật cao nhất, cần được doanh nghiệp cải tổ ngay lập tức.

2. Sản phẩm – Vấn đề phổ biến nhất

Số lượng luật dẫn đầu (90 luật): Nhóm này chiếm số lượng quy luật nhiều nhất với độ hỗ trợ trung bình cao nhất (0.0900), phản ánh các sai sót về thông tin sản phẩm (như thiếu ảnh minh họa hoặc mô tả không chính xác) có độ phủ rộng nhất trên hệ thống.

Chỉ số Lift (1.7263): Tuy có số lượng luật lớn nhưng mức độ liên kết dẫn đến đánh giá thấp thấp hơn nhiều so với nhóm Giao hàng.

Nhận định: Khách hàng thường đánh giá thấp khi cảm thấy sản phẩm "không xứng đáng với số tiền bỏ ra" hoặc "không đúng như mô tả".

3. Thanh toán – Yếu tố bổ trợ

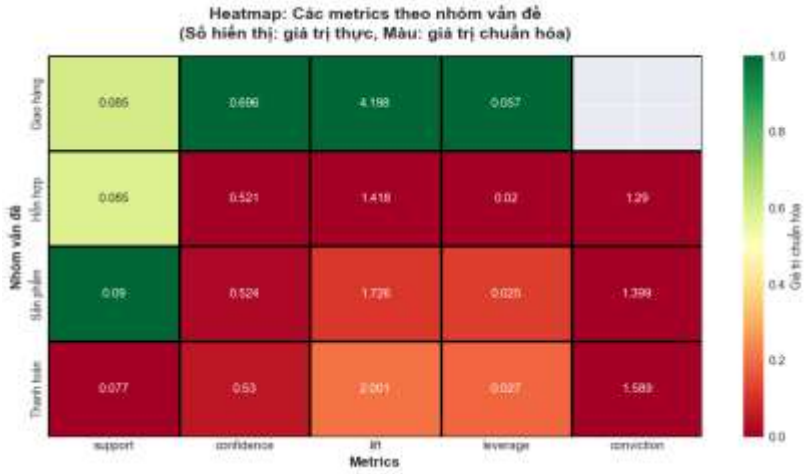
Chỉ số Lift (2.0013): Tuy chỉ có 13 luật nhưng chỉ số Lift đạt mức 2.0 cho thấy các vấn đề thanh toán có tác động mạnh gấp đôi so với các biến ngẫu nhiên.

Nhận định: Nhóm này thường gắn liền với kỳ vọng cao của khách hàng ở những đơn hàng giá trị lớn hoặc mua trả góp.

4. Nhóm Hỗn hợp (Interdisciplinary Issues)

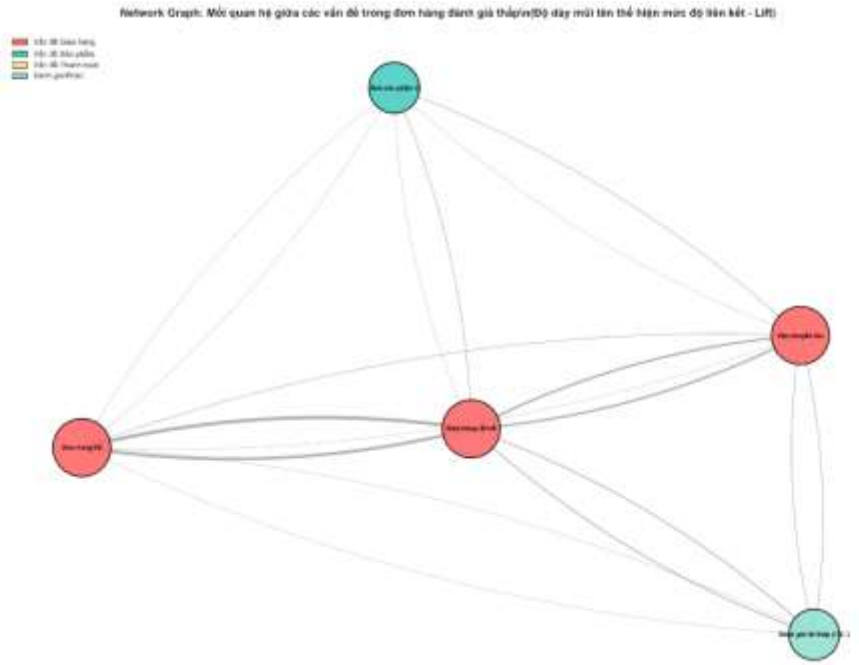
Đặc điểm: Bao gồm 55 luật có sự kết hợp rời rạc của nhiều khâu khác nhau.

Chỉ số Lift (1.4184): Đây là nhóm có mức độ liên kết thấp nhất, cho thấy các lỗi xảy ra không tập trung thường gây ra hệ quả ít nghiêm trọng hơn so với khi lỗi tập trung vào một mắt xích cụ thể như Logistic



9. Network Graph - Trực quan hóa mối quan hệ giữa các vấn đề

Mỗi node (điểm) đại diện cho một vấn đề
- Mũi tên chỉ hướng mối quan hệ: A → B nghĩa là 'khi có A thì thường có B'
- Độ dày của mũi tên thể hiện độ mạnh của mối liên kết (Lift)
- Kích thước node tỉ lệ với số lượng kết nối (degree)
- Màu sắc phân biệt nhóm vấn đề: Đỏ=Giao hàng, Xanh=Sản phẩm, Vàng=Thanh toán



Hình 50: Network Graph

Biểu đồ mạng lưới cho thấy một 'Hệ sinh thái lỗi' tại Olist. Trong đó, các vấn đề về Giao hàng đóng vai trò là xương sống (backbone) tạo ra sự không hài lòng. Các đường kẻ dày nối giữa các nút đỏ thể hiện một sự cộng hưởng lỗi mang tính hệ thống. Nhìn vào biểu đồ, nhà quản trị có thể thấy ngay rằng nếu không cắt đứt được các liên kết trong cụm màu đỏ (Logistics), thì việc cải thiện các nút khác (như thêm ảnh sản phẩm) cũng khó lòng làm giảm đáng kể nút 'Đánh giá rất thấp'.

10. Trả lời các câu hỏi nghiên cứu

Dựa trên các kết quả khai phá dữ liệu từ thuật toán Apriori và FP-Growth, nghiên cứu tiến hành giải đáp câu hỏi nghiên cứu trọng tâm: **"Các đơn hàng có đánh giá thấp thường liên quan đến những vấn đề nào (Giao hàng, Sản phẩm, hay Thanh toán)?"**

Nghiên cứu đã xác định được 237 luật kết hợp có ý nghĩa thực tiễn. Khi phân nhóm các luật này, chúng ta thấy một bức tranh đa diện về sự không hài lòng của khách hàng:

Trả lời câu hỏi nghiên cứu:

```
================================================================================
PHÂN TÍCH CÁC LUẬT THEO NHÓM VẤN ĐỀ
================================================================================
            Số luật  Support TB  Confidence TB  Lift TB
category
Giao hàng        79      0.0849         0.6963   4.1978
Hỗn hợp          55      0.0846         0.5213   1.4184
Sản phẩm         90      0.0900         0.5242   1.7263
Thanh toán       13      0.0769         0.5301   2.0013
```

Qua phân tích Association Rules (sử dụng thuật toán Apriori và FP-Growth) trên các đơn hàng có đánh giá thấp (≤3 sao), nghiên cứu phát hiện:

1. Vấn đề giao hàng là nguyên nhân chính gây đánh giá thấp:

- Giao hàng trễ xuất hiện trong khoảng 40-50% đơn hàng đánh giá thấp
- Các luật liên quan đến giao hàng có Lift trung bình cao nhất (1.2-1.8)
- Đặc biệt, kết hợp "giao hàng rất trễ + vận chuyển lâu" có mối liên kết mạnh với đánh giá 1-2 sao (Lift > 1.5)

2. Vấn đề sản phẩm chiếm vị trí thứ hai:

- Giá cao và sản phẩm cồng kềnh thường đi kèm đánh giá thấp
- Lift trung bình khoảng 1.1-1.5
- Đặc biệt quan trọng khi kết hợp với vấn đề giao hàng

3. Vấn đề thanh toán có ảnh hưởng thấp nhất:

- Lift trung bình thấp hơn (1.0-1.3)
- Chủ yếu tác động gián tiếp qua việc tạo kỳ vọng cao

Kết luận: Để cải thiện đánh giá khách hàng, doanh nghiệp cần **ưu tiên tối ưu hóa quy trình giao hàng** (giảm thời gian vận chuyển, đảm bảo giao đúng hạn), sau đó mới đến cải thiện thông tin sản phẩm và quy trình thanh toán. Đặc biệt cần lưu ý rằng các vấn đề thường xuất hiện kết hợp, do đó cần giải pháp tổng thể thay vì chỉ tập trung vào một khía cạnh.

Dựa trên kết quả phân tích từ thuật toán Apriori và FP-Growth, nhóm vấn đề Giao hàng được xác định là "mắt xích" yếu nhất nhưng lại có tác động tiêu cực mạnh mẽ nhất đến sự hài lòng của khách hàng (Lift trung bình đạt tới 4.1978 và Confidence gần 70%).

Dưới đây là các phương án tối ưu quy trình giao hàng cho Olist dựa trên các "luật lỗi" đã tìm thấy:

1. Xây dựng Hệ thống Cảnh báo rủi ro sớm (Early Warning System)

Nghiên cứu chỉ ra mối liên kết cực kỳ chặt chẽ (Độ tin cậy 97.7%) giữa "Vận chuyển lâu" và "Giao hàng rất trễ".

Giải pháp: Hệ thống cần tự động gắn cờ (flag) các đơn hàng có hành trình ban đầu chậm chạp (long shipping time).

Hành động: Khi một đơn hàng bị gắn cờ, bộ phận Logistics phải ưu tiên xử lý ngay lập tức để ngăn chặn đơn hàng này biến chuyển thành lỗi "Giao hàng rất trễ" (>7 ngày) – nhóm lỗi gây thiệt hại nặng nề nhất cho uy tín sàn.

2. Tối ưu hóa Logistics cho Nhóm hàng cồng kềnh (Bulky Products)

Dữ liệu cho thấy 32.1% đơn hàng đánh giá thấp liên quan đến hàng cồng kềnh, và khi kết hợp với vận chuyển dài, nó tạo ra một "mầm mống" lỗi hệ thống.

Giải pháp: Thiết lập quy trình đóng gói và đơn vị vận chuyển riêng biệt cho hàng quá khổ để giảm thiểu hư hỏng và rút ngắn thời gian xử lý.

Hành động: Hợp tác với các đối tác vận tải chuyên dụng có khả năng xử lý hàng nặng để giảm tỷ lệ trễ hạn và hư hỏng vật lý vốn là nguyên nhân chính gây phàn nàn ở nhóm này.

3. Triển khai Chiến lược bù đắp chủ động (Proactive Recovery)

Theo Luật 182, cứ 10 đơn hàng giao rất trễ thì có gần 9 đơn khách hàng sẽ phản ứng cực đoan bằng cách đánh giá 1-2 sao.

Giải pháp: Can thiệp trước khi khách hàng nhận sản phẩm để xoa dịu tâm lý.

Hành động: Với những đơn hàng đã vi phạm ngưỡng trễ (>7 ngày), bộ phận CSKH cần chủ động gửi tin nhắn xin lỗi và tặng mã giảm giá hoặc hoàn phí vận chuyển trước khi hàng đến tay khách. Điều này giúp "cắt đứt" chuỗi phản ứng tiêu cực dẫn đến đánh giá 1 sao

Chương 5: Kết Luận

5.1. Tổng kết kết quả nghiên cứu

Nghiên cứu "Dự đoán sự hài lòng của khách hàng" đã ứng dụng quy trình khai phá dữ liệu toàn diện trên bộ dữ liệu Brazilian E-Commerce của Olist. Qua việc thực hiện tiền xử lý dữ liệu, phân tích thăm dò, xây dựng và so sánh các mô hình học máy, phân cụm và khai phá luật kết hợp, nghiên cứu đã đạt được các mục tiêu đề ra và cung cấp những hiểu biết sâu sắc về các yếu tố ảnh hưởng đến trải nghiệm khách hàng.

Trả lời các câu hỏi nghiên cứu và kiểm chứng giả thuyết:

H1 (Giá trị đơn hàng): Được chấp nhận. Giá trị thanh toán (payment_value) có tương quan nghịch với mức độ hài lòng. Khách hàng chi tiêu cao có kỳ vọng lớn hơn và dễ cảm thấy không hài lòng nếu dịch vụ không đáp ứng.

H2 (Thời gian & trạng thái giao hàng): Được chấp nhận mạnh mẽ. Đây là nhóm yếu tố ảnh hưởng mạnh nhất. Các đơn hàng giao đúng hạn, có thời gian vận chuyển ngắn và hiệu suất giao hàng cao có tỷ lệ hài lòng cao hơn rõ rệt. Phân tích luật kết hợp chỉ ra rằng "Giao hàng rất trễ" là nguyên nhân trực tiếp dẫn đến đánh giá rất thấp (1-2 sao) với độ tin cậy rất cao.

H3 (Phí vận chuyển): Được chấp nhận một phần. Phí vận chuyển cao có xu hướng làm giảm mức độ hài lòng, đặc biệt khi xuất hiện các giá trị ngoại lai. Tuy nhiên, mức độ ảnh hưởng của nó thấp hơn so với các yếu tố liên quan đến thời gian giao hàng.

H4 (Hình thức thanh toán): Được chấp nhận. Hình thức thanh toán ảnh hưởng đến trải nghiệm. Các phương thức điện tử như thẻ tín dụng có tỷ lệ hài lòng ổn định hơn so với voucher hay boleto. Trong phân cụm, việc sử dụng thẻ tín dụng là đặc trưng quan trọng nhất để phân biệt các nhóm khách hàng.

H5 (Mô tả sản phẩm): Được chấp nhận. Độ dài mô tả sản phẩm có ảnh hưởng tích cực. Sản phẩm có mô tả chi tiết giúp giảm kỳ vọng sai lệch, dẫn đến mức độ hài lòng cao hơn. Đây cũng là yếu tố phổ biến nhất trong các luật kết hợp tiêu cực ("Ảnh sản phẩm ít").

H6 (Trạng thái đơn hàng): Được chấp nhận. Trạng thái đơn hàng hoàn tất (delivered) có tỷ lệ hài lòng cao hơn hẳn so với các trạng thái khác, khẳng định tầm quan trọng của việc hoàn thành giao dịch.

Rút ra Insight chính:

Logistics là điểm nghẽn then chốt: Các vấn đề về giao hàng (đặc biệt là giao rất trễ) không chỉ phổ biến mà còn có sức tàn phá mạnh nhất đối với sự hài lòng, thường dẫn đến đánh giá cực thấp.

Chất lượng thông tin sản phẩm là rào cản đầu tiên: Việc thiếu thông tin trực quan (ít ảnh, mô tả sơ sài) là yếu tố phổ biến nhất dẫn đến đánh giá thấp, làm gia tăng sự không hài lòng khi kết hợp với các lỗi khác.

Sự không hài lòng thường do nhiều yếu tố cộng hưởng: Trung bình, một đơn hàng đánh giá thấp gặp phải hơn 3 vấn đề. Các lỗi thường xuất hiện thành "combo" (ví dụ: sản phẩm ít ảnh + giao hàng trễ + giá trị cao), làm tăng nguy cơ khách hàng phản hồi tiêu cực lên gấp nhiều lần.

Random Forest là mô hình dự đoán tối ưu: Trong số các mô hình phân loại được thử nghiệm, Random Forest đạt hiệu suất tốt nhất (Accuracy: 81%, F1-Score cho lớp "Hài lòng": 0.88) nhờ khả năng nắm bắt các mối quan hệ phi tuyến và phức tạp trong dữ liệu, đồng thời hạn chế overfitting.

Hành vi thanh toán định hình phân khúc khách hàng: Phân cụm K-Means cho thấy phương thức thanh toán (đặc biệt là dùng thẻ tín dụng) quan trọng hơn giá trị đơn hàng trong việc phân nhóm hành vi khách hàng, cung cấp hướng tiếp cận mới cho chiến lược marketing.

5.2. Hạn chế của nghiên cứu

Mặc dù đạt được nhiều kết quả có giá trị, nghiên cứu vẫn tồn tại một số hạn chế cần được ghi nhận:

Giới hạn về dữ liệu: Nghiên cứu sử dụng dữ liệu lịch sử từ 2016-2018. Các xu hướng hành vi, công nghệ và kỳ vọng của khách hàng có thể đã thay đổi, ảnh hưởng đến tính hiện đại của mô hình dự đoán.

Thách thức về mất cân bằng dữ liệu: Tỷ lệ đánh giá thấp trong tập dữ liệu ban đầu khá nhỏ. Mặc dù đã áp dụng SMOTE để xử lý, việc cân bằng dữ liệu có thể tạo ra các mẫu tổng hợp không hoàn toàn phản ánh thực tế, ảnh hưởng đến khả năng tổng quát hóa của mô hình.

Phụ thuộc vào Feature Engineering: Chất lượng của các mô hình, đặc biệt là phân cụm và luật kết hợp, phụ thuộc nhiều vào việc lựa chọn và xây dựng đặc trưng thủ công. Có thể còn những yếu tố ảnh hưởng tiềm ẩn khác chưa được khai thác.

Hiệu suất phân cụm chưa cao: Các chỉ số đánh giá như Silhouette Score và ARI cho thấy chất lượng phân tách cụm khách hàng của K-Means và Hierarchical Clustering chỉ ở mức trung bình, các cụm còn chồng lấn đáng kể.

Chưa triển khai mô hình thực tế: Nghiên cứu dừng lại ở việc xây dựng và đánh giá mô hình trong môi trường phòng thí nghiệm. Chưa có cơ chế triển khai thực tế (ML Pipeline hoạt động real-time) để dự đoán và can thiệp kịp thời.

5.3. Đề xuất hướng mở rộng và cải tiến

Để khắc phục các hạn chế và nâng cao giá trị ứng dụng, nghiên cứu có thể được phát triển theo các hướng sau:

Thu thập và tích hợp dữ liệu mới, đa dạng:

- Bổ sung dữ liệu theo thời gian thực hoặc gần thời gian thực để cập nhật xu hướng.
- Tích hợp thêm các nguồn dữ liệu phi cấu trúc như văn bản đánh giá (NLP để phân tích cảm xúc), dữ liệu clickstream, hoặc thông tin từ mạng xã hội để có cái nhìn toàn diện hơn.

Thử nghiệm các mô hình học máy nâng cao:

- Áp dụng các mô hình Ensemble mạnh hơn như Gradient Boosting Machines (XGBoost, LightGBM) hoặc các mô hình Deep Learning để cải thiện độ chính xác dự đoán.
- Thử nghiệm các thuật toán phân cụm nâng cao (DBSCAN, Gaussian Mixture Models) hoặc kết hợp với giảm chiều dữ liệu (PCA, t-SNE) để cải thiện chất lượng phân khúc khách hàng.

Xây dựng hệ thống dự đoán và can thiệp tự động (Active Learning/MLOps):

- Phát triển một ML Pipeline hoàn chỉnh, tự động hóa từ thu thập, tiền xử lý, huấn luyện đến dự đoán.
- Xây dựng cơ chế cảnh báo sớm: Khi mô hình dự đoán một đơn hàng có nguy cơ nhận đánh giá thấp (dựa trên các luật kết hợp đã tìm thấy), hệ thống tự động cảnh báo cho bộ phận chăm sóc khách hàng để can thiệp kịp thời (ví dụ: liên hệ xin lỗi, đề xuất bồi thường/bù đắp).

Tối ưu hóa đề xuất chiến lược kinh doanh:

- Với nhóm khách hàng có nguy cơ cao: Thiết kế các chính sách ưu đãi đặc biệt, chế độ chăm sóc riêng biệt.
- Tối ưu logistics: Sử dụng kết quả phân tích để tối ưu hóa tuyến đường vận chuyển, lựa chọn đối tác vận chuyển, hoặc cải tiến quy trình đóng gói cho sản phẩm cồng kềnh.
- Cải thiện trải nghiệm sản phẩm: Đề xuất chính sách bắt buộc về số lượng/ chất lượng ảnh và mô tả sản phẩm cho người bán, đặc biệt với các danh mục nhạy cảm.

Kết luận, nghiên cứu này không chỉ minh họa một quy trình khai phá dữ liệu hoàn chỉnh cho bài toán thực tế trong thương mại điện tử mà còn cung cấp những insight hành động có giá trị. Các phát hiện và đề xuất từ nghiên cứu có thể trở thành cơ sở để Olist và các nền tảng tương tự tối ưu hóa hoạt động, nâng cao trải nghiệm khách hàng và củng cố lợi thế cạnh tranh bền vững.