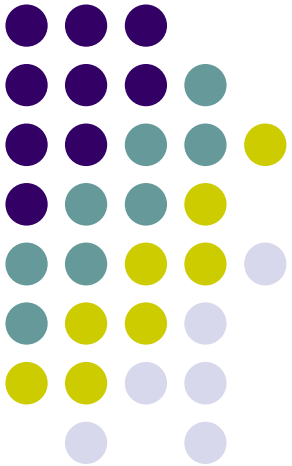
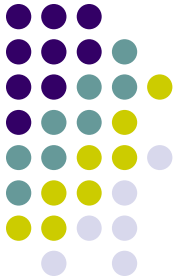
Nhóm 4

DỰ ĐOÁN SỰ HÀI LÒNG CỦA KHÁCH HÀNG

Ứng dụng phân tích mô tả, phân cụm, phân loại, khai phá luật



- GVHD: Đỗ Như Tài
- Khai phá dữ liệu

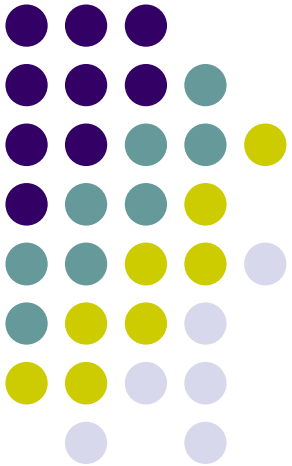


STT	Họ Tên	MSSV	Mức độ đóng góp
1	Hồ Gia Bảo	3123580003	33.33%
2	Nguyễn Gia Huy	3123580016	33.33%
3	Trần Nguyễn Minh Tiến	3123580052	33.33%

CHƯƠNG 1: GIỚI THIỆU

- Trong bối cảnh thương mại điện tử phát triển mạnh, đặc biệt tại các quốc gia có tốc độ số hóa cao như Brazil, việc phân tích dữ liệu đóng vai trò then chốt trong việc hiểu hành vi khách hàng và tối ưu hoạt động kinh doanh.

- Bộ dữ liệu *Brazilian E-Commerce Public Dataset* phản ánh toàn bộ quy trình mua sắm từ đặt hàng, giao hàng đến đánh giá, là nguồn dữ liệu thực tế và quy mô lớn, phù hợp để áp dụng các kỹ thuật khai phá dữ liệu như tiền xử lý, phân tích mô tả, phân cụm, phân loại và luật kết hợp.



TỔNG QUAN DỮ LIỆU

THU THẬP DỮ LIỆU

Bộ dữ liệu gồm khoảng 100.000 đơn hàng tại Brazil trong giai đoạn 2016–2018, thu thập từ nhiều sàn thương mại điện tử. Dữ liệu phản ánh toàn diện quá trình mua sắm, bao gồm trạng thái đơn hàng, giá cả, thanh toán, vận chuyển, thông tin khách hàng, sản phẩm và đánh giá. Ngoài ra, bộ dữ liệu còn tích hợp thông tin địa lý liên kết mã bưu chính với tọa độ vĩ độ/kinh độ. Toàn bộ dữ liệu đã được ẩn danh nhằm đảm bảo quyền riêng tư.

LINK NGUỒN DỮ LIỆU

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

MÔ TẢ DỮ LIỆU

Phạm vi: Nghiên cứu giới hạn trong dữ liệu thương mại điện tử tại Brazil giai đoạn 2016–2018, tập trung phân tích đơn hàng, vận chuyển, thanh toán và sự hài lòng của khách hàng.

Kích thước: 40 biến 115609 quan sát

Các thuộc tính chính: float64(10), int64(6), object(24)

```
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)
```

VÍ DỤ MẪU

```
df.describe()
```

Python

	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	

MỤC TIÊU NGHIÊN CỨU

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.

CÂU HỎI NGHIÊN CỨU

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?.

CHƯƠNG 2: PHÂN TÍCH DỮ LIỆU

Tiền xử lý dữ liệu và xử lý mất cân bằng

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.

Các bước tiền xử lý

Excel

- Làm sạch và chọn lọc dữ liệu
- Loại bỏ cột không cần thiết.

Python

- Chuẩn hóa định dạng và xử lý missing values
- Xóa dòng trùng lặp
- Chuyển định dạng:

→ Số thực cho biến định lượng

→ Chuỗi cho biến định tính

- Xóa ký tự thừa

Mô tả các biến

customers: chứa thông tin cơ bản về khách hàng.

geolocation: là bảng có kích thước lớn nhất, lưu trữ thông tin vị trí địa lý.

order_items: mô tả chi tiết các sản phẩm trong từng đơn hàng.

order_payments: chứa thông tin về hình thức và giá trị thanh toán.

order_reviews: ghi nhận đánh giá và phản hồi của khách hàng.

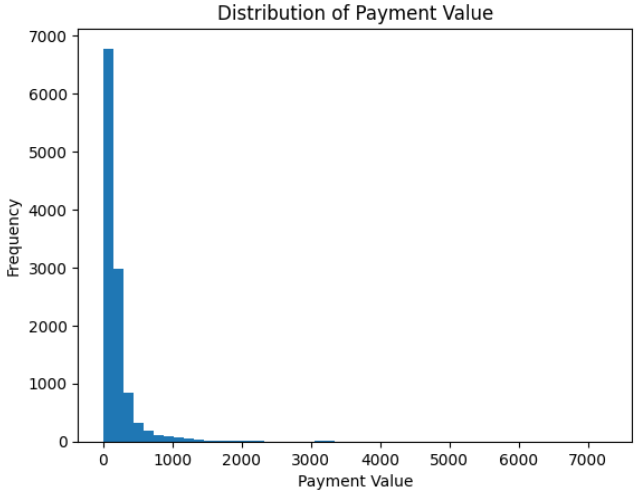
orders: là bảng trung tâm liên kết hầu hết các bảng còn lại.

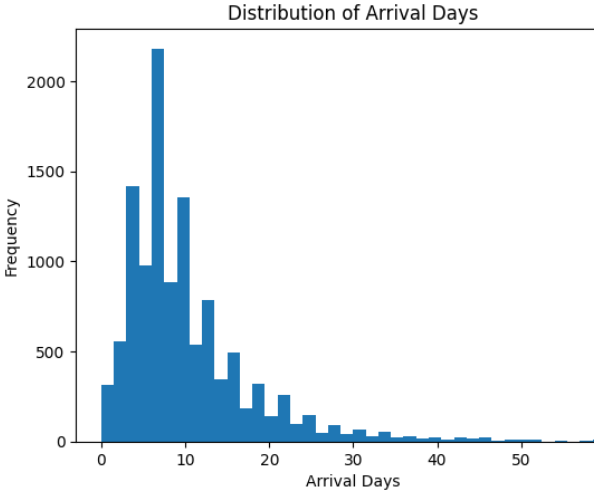
products: cung cấp thông tin chi tiết về sản phẩm.

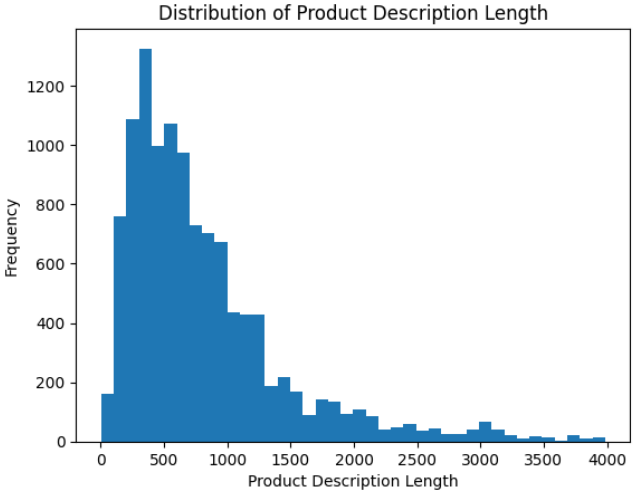
sellers: lưu trữ dữ liệu về người bán.

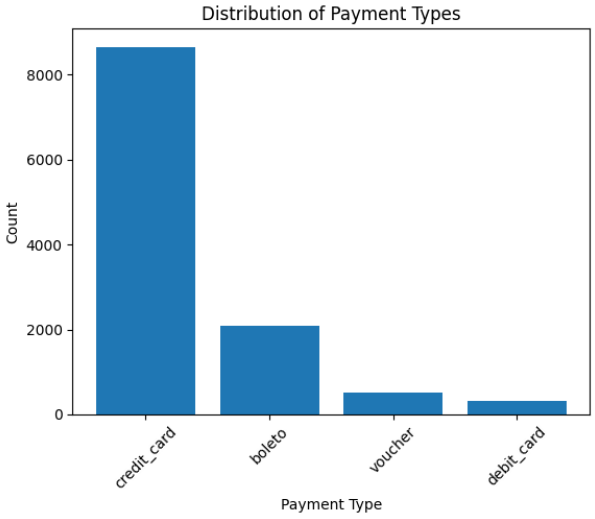
category_translation: dùng để dịch tên danh mục sản phẩm sang tiếng Anh.

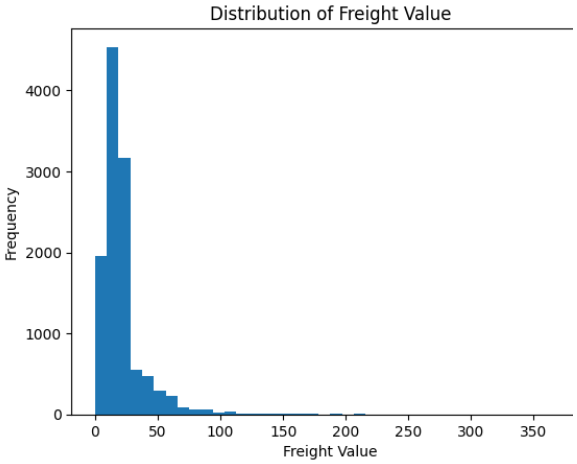
Phân phối các biến bằng Histogram

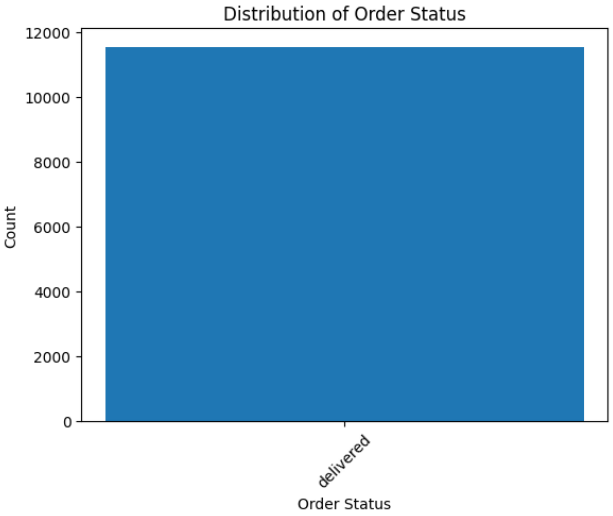






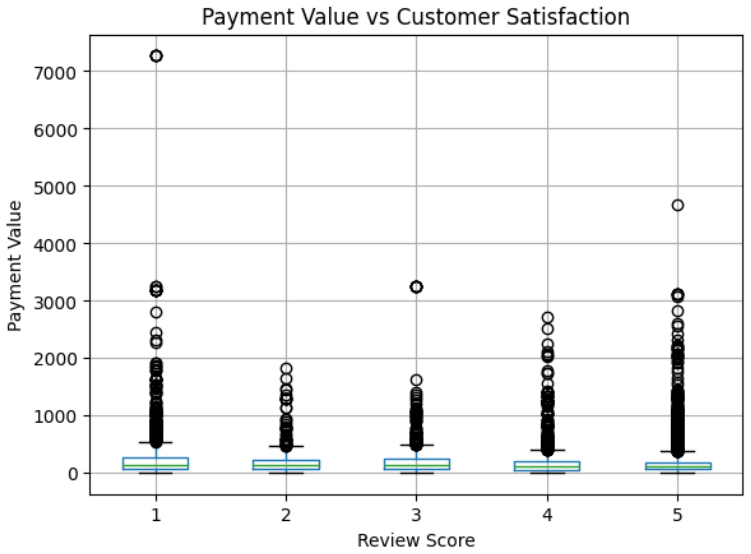






Mối quan hệ giữa payment_value và mức độ hài lòng

Kết quả so sánh giá trị thanh toán giữa hai nhóm khách hàng hài lòng và không hài lòng cho phép đánh giá vai trò của *payment_value* trong việc ảnh hưởng đến mức độ hài lòng và xu hướng đánh giá của khách hàng.

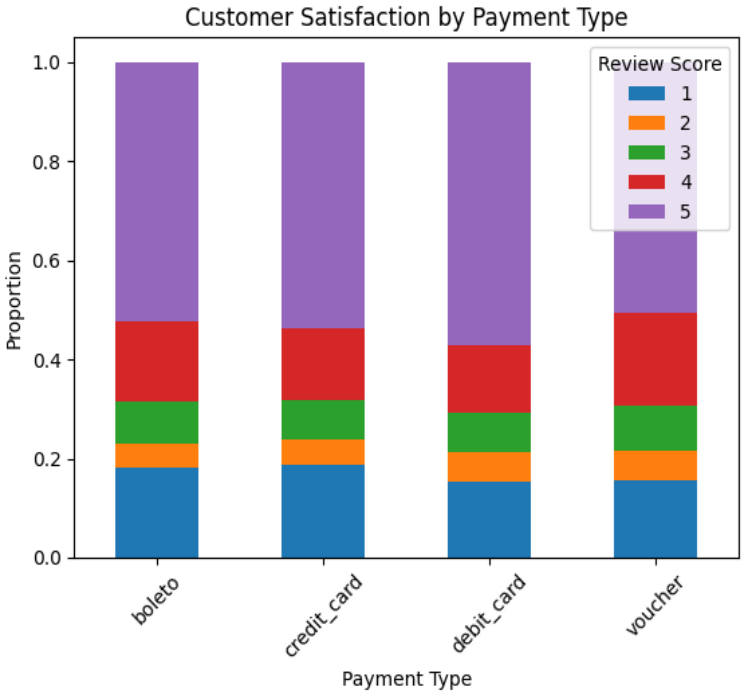


Mối quan hệ giữa freight_value và mức độ hài lòng

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.

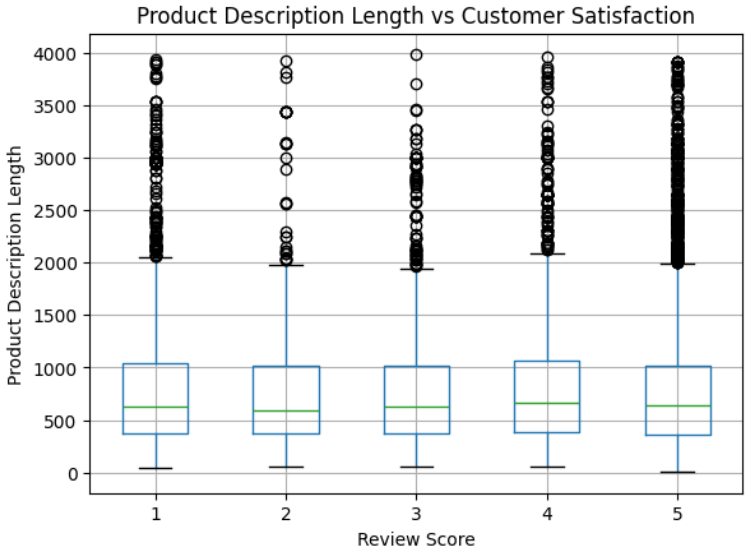
So sánh trực quan mức độ hài lòng theo hình thức thanh toán

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.



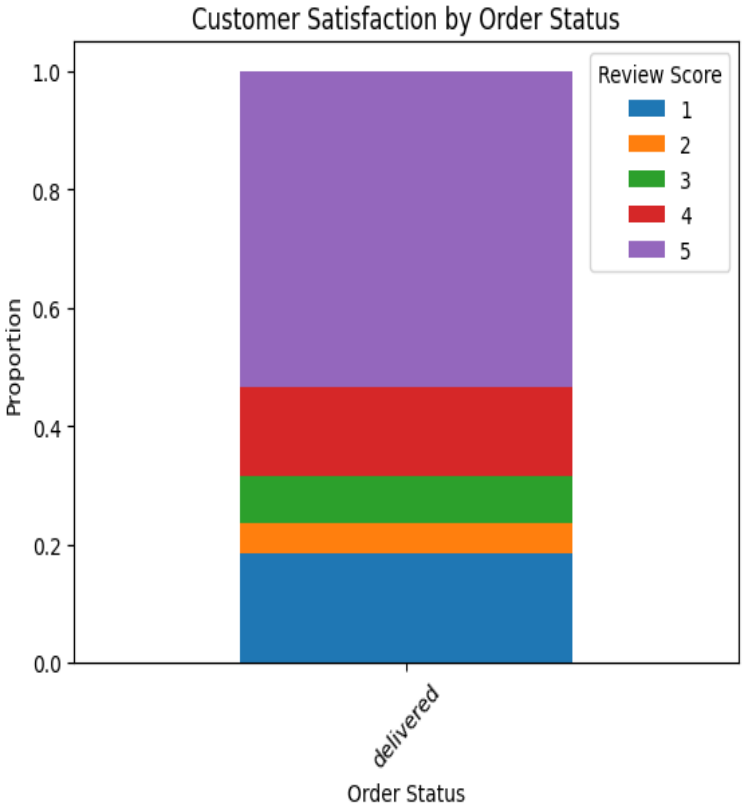
Mối quan hệ giữa độ dài mô tả và mức độ hài lòng

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.



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.



XÂY DỰNG MÔ HÌNH PHÂN NHÓM

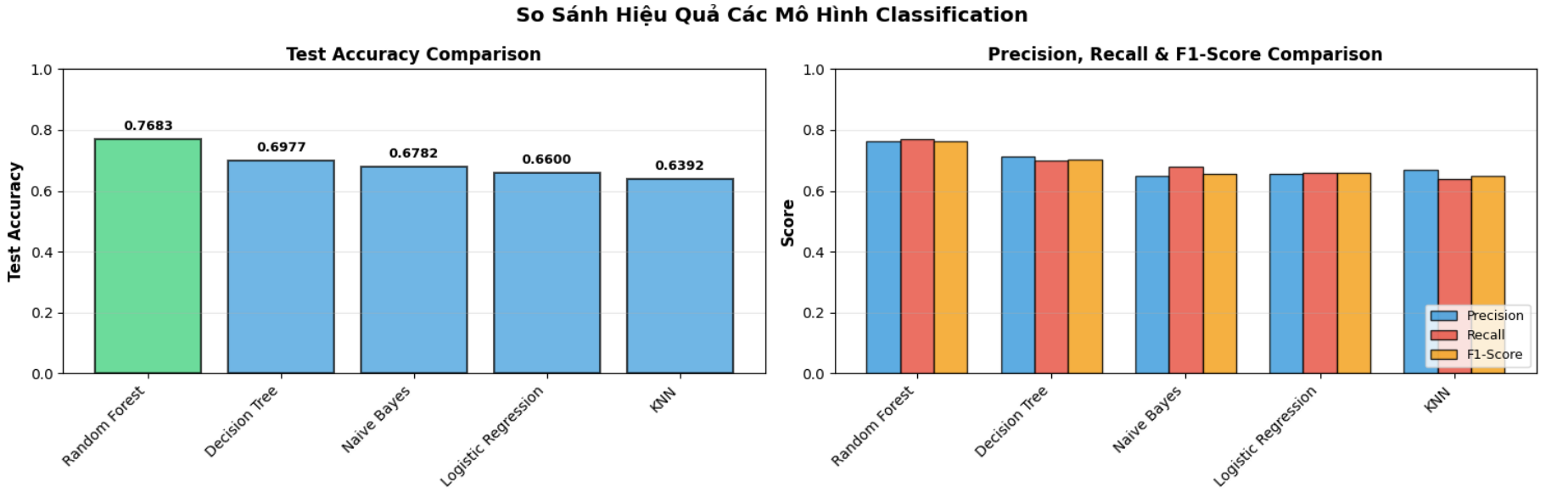
Mô hình sử dụng

- Logistic Regression
- Decision Tree
- Random Forest
- **K-Nearest Neighbors**
- **Naïve Bayes**

Lý do thử nhiều mô hình

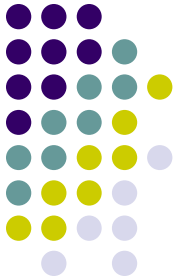
- Đánh giá độ phù hợp với dữ liệu hài lòng
- Chọn mô hình có khả năng dự đoán cao & ổn định nhất

SO SÁNH HIỆU SUẤT 5 MÔ HÌNH



Random Forest được chọn làm mô hình chính nhờ hiệu suất phân loại cao nhất và khả năng phân tích mức độ quan trọng của các đặc trưng.

Mô hình Random Forest



Tạo nhiều cây quyết định từ các mẫu dữ liệu con (Bootstrap Sampling).
Mỗi cây chỉ chọn ngẫu nhiên một tập đặc trưng khi chia nhánh.
Kết quả dự đoán cuối cùng lấy theo bỏ phiếu đa số giữa các cây.
→ Giúp mô hình ổn định và chính xác hơn trên dữ liệu thực tế.

Kết quả mô hình Random Forest

Evaluation on Training

	precision	recall	f1-score	support
Not Satisfied	1.00	1.00	1.00	6320
Satisfied	1.00	1.00	1.00	6320
accuracy			1.00	12640
macro avg	1.00	1.00	1.00	12640
weighted avg	1.00	1.00	1.00	12640

Evaluation on Testing

	precision	recall	f1-score	support
Not Satisfied	0.67	0.57	0.61	728
Satisfied	0.81	0.87	0.84	1581
accuracy			0.77	2309
macro avg	0.74	0.72	0.73	2309
weighted avg	0.77	0.77	0.77	2309

Mô hình Random Forest đạt kết quả hoàn hảo trên tập huấn luyện (Accuracy = 100%), cho thấy khả năng ghi nhớ dữ liệu rất cao. Tuy nhiên, trên tập kiểm tra, độ chính xác giảm xuống 77%, phản ánh hiện tượng quá khớp (overfitting). Mô hình dự đoán tốt nhóm *Satisfied* (F1-score = 0.84) nhưng kém hơn với nhóm *Not Satisfied* (F1-score = 0.61), cho thấy khả năng nhận diện khách hàng không hài lòng còn hạn chế.

DỰ ÁN KHOA HỌC DỮ LIỆU

Mô hình phân cụm

Phân khúc khách hàng thương mại điện tử

Từ vấn đề kinh doanh đến các đề xuất có thể thực hiện được

DỮ LIỆU
11.542 khách hàng
Biến số

THUẬT TOÁN
K-Means & Phân cấp

KHÁCH QUAN
Phân đoạn Có thể hành động

NGÀY
Tháng 12 2025

DỰ ÁN KHOA HỌC DỮ LIỆU

Mô hình phân cụm

Phân khúc khách hàng thương mại điện tử

Từ vấn đề kinh doanh đến các đề xuất có thể thực hiện được

DỮ LIỆU
11.542 khách hàng
Biến số

THUẬT TOÁN
K-Means & Phân cấp

KHÁCH QUAN
Phân đoạn Có thể hành động

NGÀY
Tháng 12 2025

Bản chất của vấn đề

Hiểu rõ những thách thức của việc phân khúc khách hàng trong thương mại điện tử

Bối cảnh thương mại điện tử

Ngành thương mại điện tử đang đối mặt với **hành vi mua sắm không đồng nhất . Với lượng khách hàng lớn và đa dạng, phương pháp "một kích cỡ phù hợp cho tất cả" không còn khả thi. Việc cá nhân hóa trên quy mô lớn** là vô cùng cần thiết để duy trì khả năng cạnh tranh.

Có vấn đề

Làm thế nào để chuyển đổi dữ liệu thô thành thông tin chi tiết? Thách thức chính là **xác định các phân khúc nhất quán và riêng biệt** trong hàng nghìn giao dịch để kích hoạt các chiến lược tiếp thị và bán hàng mục tiêu, thay vì chỉ dựa vào trực giác.

Mục tiêu chiến lược

- Hiểu rõ **giá trị vòng đời khách hàng** (LTV) và sở thích thanh toán.
- Phân tích tác động của **chi phí hậu cần** đến việc mua hàng.
- Cải thiện **tỷ lệ giữ chân khách hàng** và tối ưu hóa ROI chiến dịch thông qua việc nhắm mục tiêu chính xác.

Hạn chế kỹ thuật

Bộ dữ liệu này đặt ra những thách thức cụ thể: **Dữ liệu hỗn hợp** (số và phân loại), thang đo rất khác nhau (giá cả so với điểm số), **khối lượng lớn** (hơn 11.000 dòng) và nhu cầu tuyệt đối về **khả năng diễn giải** cho các nhóm kinh doanh.

Câu hỏi & Biến số

Khung phân tích và cấu trúc dữ liệu

6 câu hỏi quan trọng

Câu 1 Liệu có thể phân khúc khách hàng một cách hiệu quả dựa trên **mức chi tiêu** , phương thức **thanh toán** và **mức độ hài lòng** của họ hay không?

Câu 2 **Những đặc điểm hành vi** nào là yếu tố then chốt để hình thành các nhóm này?

Câu 3 Liệu thuật toán **K-Means** và thuật toán **phân cấp** có tạo ra các cụm ổn định không?

Quý 4 **Số lượng phân đoạn tối ưu** cho phân tích kinh doanh là bao nhiêu ?

Câu 5 Thuật toán nào mang lại hiệu suất kỹ thuật tốt nhất (**Silhouette, Inertia**)?

Câu 6 Liệu các nhóm khách hàng đã xác định **có thể được sử dụng hiệu quả** trong hoạt động tiếp thị và hậu cần không?

Biến số phân tích

BIẾN SỐ

`giá trị thanh toán` `giá_hàng_hàng` `trả góp` `số lượng ảnh sản phẩm` `độ dài mô tả sản phẩm` `điểm đánh giá`

BIẾN PHÂN LOẠI (ONE-HOT)

`thẻ tín dụng` `thẻ ghi nợ` `chứng từ` `trạng thái đơn hàng`

TIỀN XỬ LÝ

- Giá trị thiếu : Điền giá trị bằng 0.
- Chuẩn hóa : Tiêu chuẩn hóa (ngoại trừ review_score)

TỔNG SỐ MẪU

11.542

Khách hàng độc đáo

Phương pháp luận và quy trình làm việc

Một phương pháp tiếp cận có cấu trúc gồm 5 bước để chuyển đổi dữ liệu thô thành các phân khúc khách hàng có thể hành động được.

1

Sự chuẩn bị

- Điền các giá trị bị thiếu (0)
- Mã hóa one-hot (Thanh toán/Trạng thái)
- Tiêu chuẩn hóa & Vệ sinh

2

Mô hình hóa

- K-Means (K được khám phá từ 2 đến 10)
- Phân cụm phân cấp (mẫu 5k)
- So sánh các thuật toán

3

Đánh giá

- Điểm số hình bóng
- Phương pháp quán tính và khuỷu tay
- Độ ổn định (ARI) & Khoảng cách

4

Giải thích

- Thống kê mô tả theo cụm
- Tầm quan trọng của các biến số
- Phân tích hành vi

5

Lựa chọn

- Chất lượng so với tay nghề: Sự thỏa hiệp
- Định nghĩa nhân vật
- Chiến lược hành động

Phân bố cụm

Phân tích K-Means trên 11.542 khách hàng

Tổng cộng: 11.542

Cụm 0

00

Tiêu chuẩn & Tiền mặt

Giỏ trung bình	172,9 đô la
Thanh toán hàng tháng	1.00
Chi phí vận chuyển	21,5 đô la

Dân số — 311 khách hàng (2,7%)

Đa số cụm 1

01

Giá trị cao & Thanh toán hàng tháng

Giỏ trung bình	202,3 đô la
Thanh toán hàng tháng	3,75
Chi phí vận chuyển	22,1 đô la

Dân số — **8.647 khách hàng (74,9%)**

Cụm 2

02

Giá trị cao & Tiền mặt

Giỏ trung bình	213,1 đô la
Thanh toán hàng tháng	1.00
Chi phí vận chuyển	21,1 đô la

Cụm 3

03

Tiết kiệm & Tiền mặt

Giỏ trung bình	72,8 đô la
Thanh toán hàng tháng	1.00
Chi phí vận chuyển	23,4 đô la

Điểm mấu chốt: Phân khúc này chủ yếu (75%) bao gồm các giao dịch mua hàng giá trị cao được thanh toán trả góp.

Các yếu tố quyết định

Phân tích tầm quan trọng của tính năng

Phân tích K-Means

Sự chi phối của phương thức thanh toán

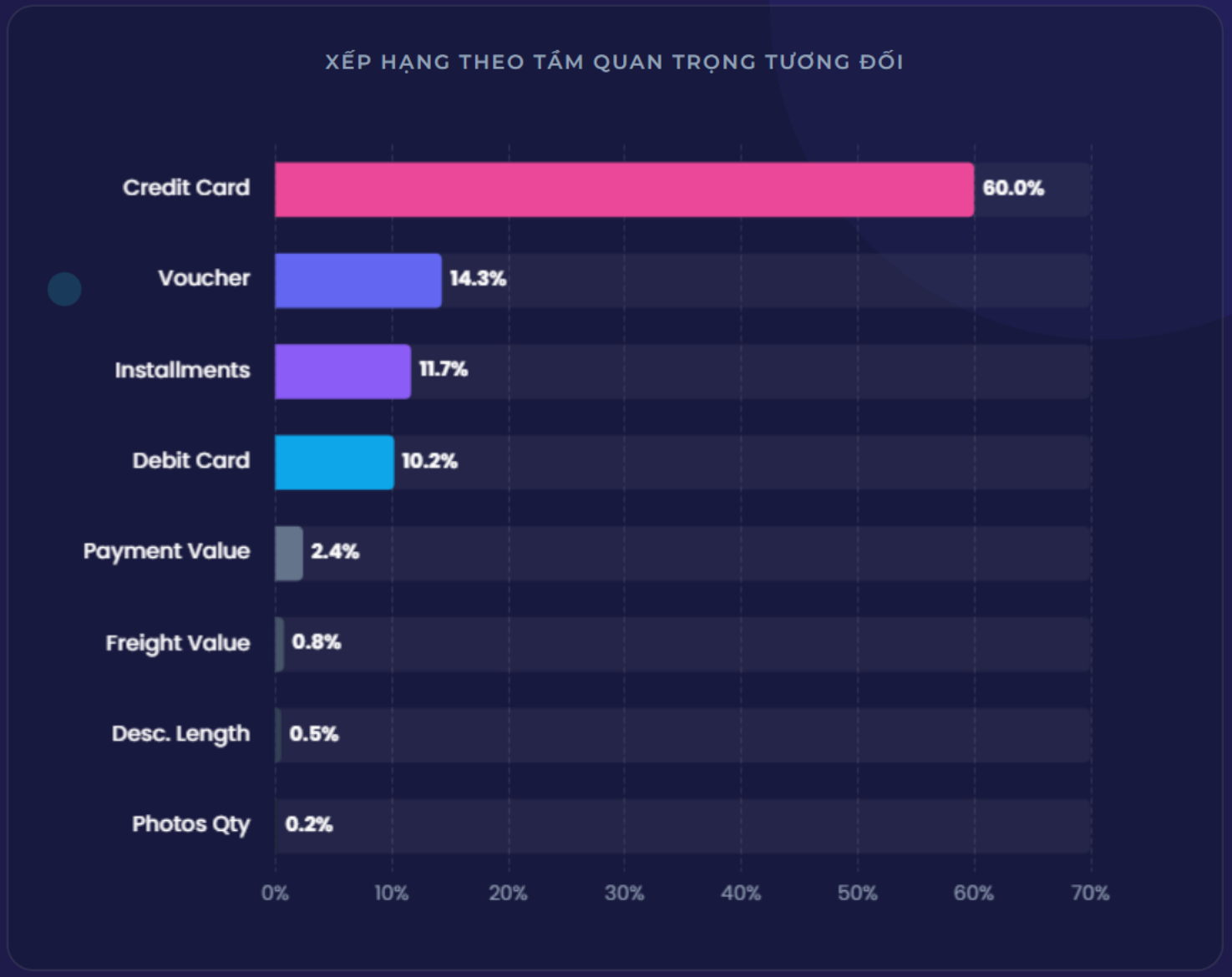
Thanh toán **bằng thẻ tín dụng** là yếu tố quan trọng nhất (60%). Điều này cho thấy sự khác biệt rõ rệt giữa người mua bằng thẻ tín dụng và những người mua khác.

Vai trò của các khoản thanh toán hàng tháng

Khả năng thanh toán **trả góp** (Trả góp: 12%) và việc sử dụng Phiếu mua hàng (14%) tạo nên cấu trúc nhóm rõ rệt hơn nhiều so với chỉ riêng số tiền.

Giá trị & Hậu cần

Điều đáng ngạc nhiên là **Giá trị thanh toán** (2,4%) và **Giá trị vận chuyển** (0,8%) đóng vai trò thứ yếu so với các cơ chế tài chính.



Hiệu suất thuật toán

So sánh kỹ thuật: K-Means so với Phân cụm phân cấp

CHỈ SỐ CHÍNH

Điểm số hình bóng

K-Means

✓ ĐÃ CHỌN

Điểm số hình bóng Chất lượng tách biệt	0.3284
Độ ổn định (ARI) Tính nhất quán (5 lần chạy)	0.3131 ± 0,3435 SD
Khoảng cách tâm Khoảng cách trung bình	5,61
Quán tính Tính nhỏ gọn	55.571 bộ dữ liệu đầy đủ

TẠI SAO LẠI CHỌN CÁCH NÀY?

Điểm Silhouette tốt nhất và độ ổn định đã được chứng minh. Cho phép xử lý hiệu quả tất cả dữ liệu (hơn 11.000) với khả năng diễn giải kinh doanh tốt.

VS

Phân cấp

Điểm số hình bóng Chất lượng tách biệt	0.3097
Sự ổn định Độ nhạy tiếng ồn	Yếu đuối Chi phí tính toán cao
Khoảng cách tâm Khoảng cách trung bình	5,81 +0,20 (Tốt hơn một chút)
Quán tính Tính nhỏ gọn	23.267 Mẫu (5k)

⚠ Bị hạn chế bởi độ phức tạp tính toán. Yêu cầu lấy mẫu (5k điểm) làm giảm độ chính xác tổng thể.

Giải pháp — Số lượng phân đoạn tối ưu

Cân bằng giữa độ chính xác thống kê và khả năng vận hành kinh doanh

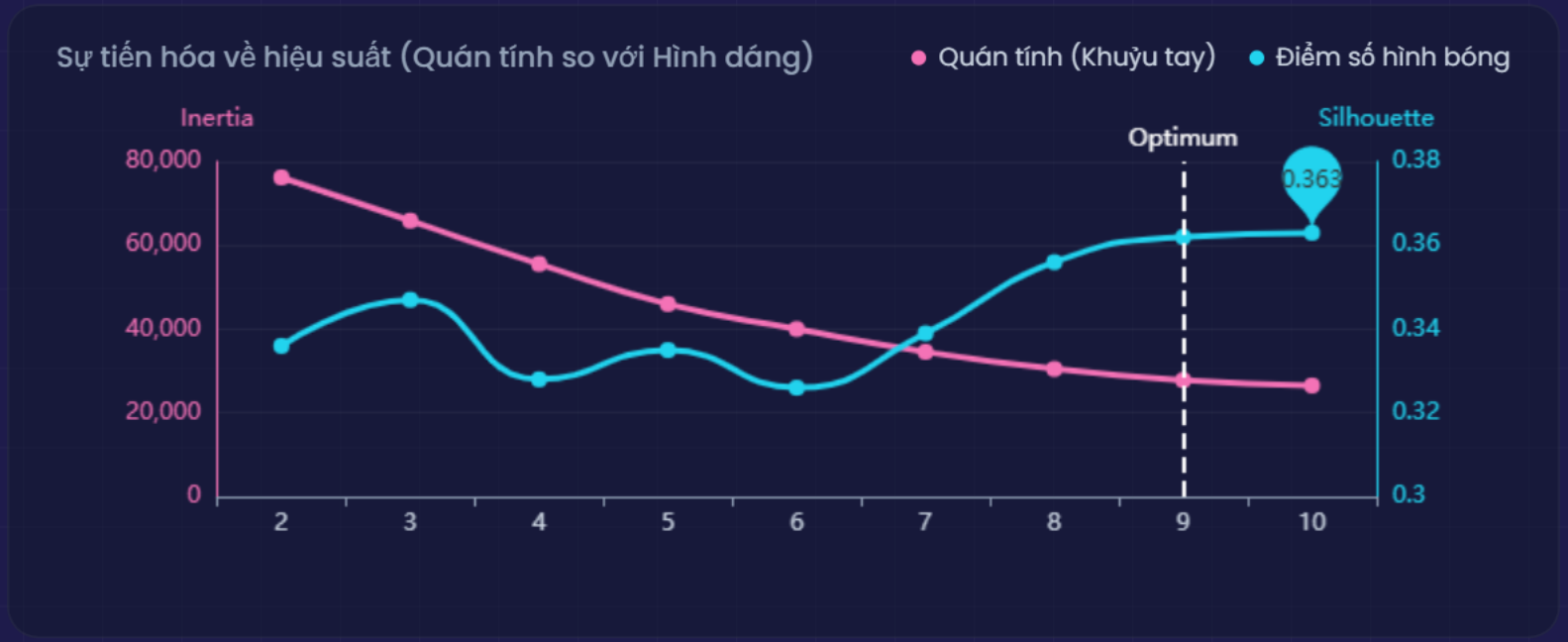
SỰ GIỚI THIỆU

9

Các cụm

Sự thỏa hiệp lý tưởng

Tối đa hóa sự khác biệt về hành vi (Hình bóng) trong khi vẫn duy trì cấu trúc mạch lạc (Quán tính).



Lý giải kỹ thuật

Phương pháp khuỷu tay	**K = 8**
Hình bóng (Max)	**K = 10 (0,363)**
Hình bóng (K=9)	**0,362 (Rất gần)**

K=9 nắm bắt được những sắc thái tinh tế mà không cần phân đoạn quá mức.

Chiến lược triển khai

Giai đoạn 1: Giao tiếp	**K = 4**
Giai đoạn 2: Kích hoạt	**K = 9**
Khách quan	**Cá nhân hóa cao cấp**

Sử dụng K=4 cho báo cáo dành cho cấp quản lý, K=9 cho CRM.

Ứng dụng thực tiễn

Chuyển đổi các phân khúc (cụm) thành đòn bẩy hoạt động cho công ty.

Tiếp thị & Giữ chân khách hàng

SỰ MUA LẠI

- **Chương trình Khách hàng thân thiết VIP (Nhóm 1 & 2):** Các chiến dịch email độc quyền dành cho khách hàng có giá trị cao.
- **Bán chéo sản phẩm có mục tiêu:** Đề xuất dựa trên lịch sử mua hàng để tăng giá trị đơn hàng trung bình.
- **Quy trình tiếp nhận khách hàng mới (Cụm 3):** Chuỗi chào mừng mang tính giáo dục nhằm chuyển đổi khách hàng nhỏ mới thành khách hàng tiềm năng.

Chiến lược định giá

THU NHẬP

- **Ưu đãi cao cấp:** Tập trung vào giá trị gia tăng thay vì giảm giá cho các phân khúc khách hàng không nhạy cảm với giá cả.
- **Các chương trình khuyến mãi mang tính chiến thuật:** Phiếu giảm giá nhắm mục tiêu để kích hoạt lại các khách hàng "không hoạt động" nhạy cảm với giá cả.
- **Phương thức thanh toán:** Thanh toán trả góp (Installments) được làm nổi bật trên trang sản phẩm có giá cao.

Sản phẩm chào bán

DANH MỤC

- **Điều chỉnh danh mục sản phẩm:** Làm nổi bật các sản phẩm "Bán chạy nhất" tương ứng với từng đối tượng khách hàng.
- **Nâng cao nội dung:** Cải thiện hình ảnh và mô tả cho các sản phẩm có mức độ tương tác cao (Nhóm 2).

Dịch vụ khách hàng & Hậu cần

KINH NGHIỆM

- **Dịch vụ ưu tiên:** Hàng chờ riêng dành cho khách hàng VIP để đảm bảo sự hài lòng tối đa.
- **Tối ưu hóa hậu cần:** Các tùy chọn giao hàng tiết kiệm là mặc định cho các đơn hàng nhỏ.

BÁO CÁO PHÂN TÍCH DỮ LIỆU | Khai phá dữ liệu & Học máy

Kết quả thực thi thuật toán Apriori & FP-Growth

Phân tích toàn diện về tập phổ biến, luật kết hợp và so sánh hiệu năng trên tập dữ liệu giao dịch.

KẾT QUẢ KHAI PHÁ
111 Tập phổ biến
237 Luật kết hợp

LĨNH VỰC TRỌNG TÂM
Giao hàng & Sản phẩm
Phân tích hành vi đánh giá

KẾT QUẢ THỰC THI THUẬT TOÁN

Tổng quan kết quả Apriori

Số lượng tập phổ biến và các luật kết hợp được phát hiện

111

TẬP PHỔ BIẾN

(Frequent Itemsets)

237

LUẬT KẾT HỢP

(Association Rules)

Thỏa mãn ngưỡng thiết lập

Tất cả các luật được sinh ra đều thỏa mãn ngưỡng **Support** (độ hỗ trợ) và **Confidence** (độ tin cậy) đã được thiết lập từ trước, đảm bảo tính chặt chẽ của kết quả.

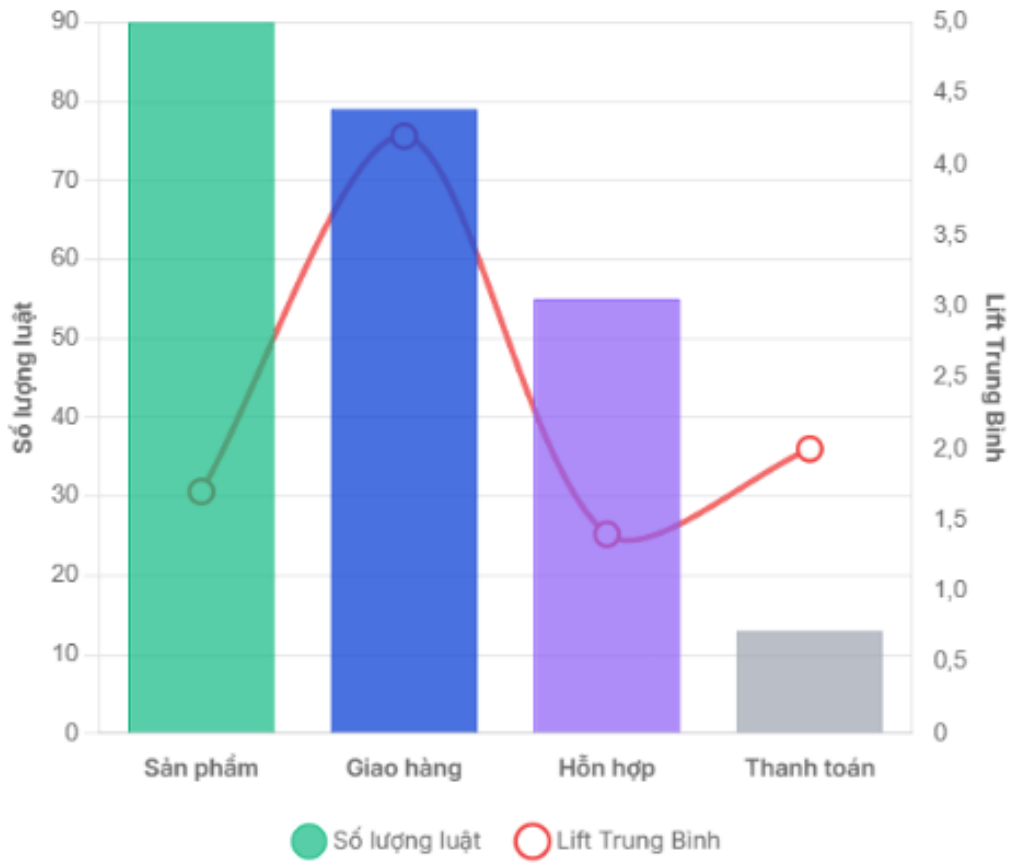
Dữ liệu đủ độ dày

Số lượng luật và tập phổ biến thu được cho thấy dữ liệu đầu vào có độ dày đủ lớn để khai phá các mối quan hệ có ý nghĩa (meaningful relationships), không bị phân tán hay nhiễu quá mức.

PHÂN TÍCH CHI TIẾT

Phân bố luật theo nhóm vấn đề

Phân tích 237 luật kết hợp dựa trên nội dung và độ đo Lift



Giao hàng

Chiếm tỷ trọng lớn nhất

79 LUẬT

Lift TB: ~4.2 (Cao nhất) *Mối liên hệ rất mạnh với đánh giá thấp*

Sản phẩm

90 LUẬT

Lift TB: ~1.7 Số lượng nhiều nhưng tác động kém hơn giao hàng.

Hỗn hợp (Đa nguyên nhân)

55 LUẬT

Lift TB: ~1.4 Phản ánh tính phức tạp, đa yếu tố.

Thanh toán

Lift TB: ~2.0

13 LUẬT

PHÂN TÍCH CHUYÊN SÂU

Top 10 Luật có Lift cao nhất

Các quy luật mạnh nhất chi phối sự không hài lòng của khách hàng

Đặc điểm chung (Lift > 6.0)

Xếp hạng theo Lift giảm dần

1 LUẬT ĐIỂN HÌNH NHẤT

Giao hàng rất trễ → Giao hàng trễ + Đánh giá rất thấp + Vận chuyển lâu

LIFT 6.8

Confidence: **0.97** Rất phổ biến

2 Vận chuyển > 7 ngày → Đánh giá 1 sao — **Lift: 6.5**

Khách hàng mất kiên nhẫn dẫn đến đánh giá tiêu cực ngay lập tức.

3 Giao hàng trễ + Hư hỏng nhẹ → Đánh giá 2 sao — **Lift: 6.2**

Sự kết hợp giữa thời gian và chất lượng tạo ra phản ứng tiêu cực mạnh.

CHỈ SỐ THỐNG KÊ NHÓM TOP 10

0.88 - 0.97

Confidence Range

Khi giao hàng trễ xảy ra, gần như chắc chắn (>88%) sẽ nhận đánh giá thấp.

> 6.0

Lift Score

Mối liên kết mạnh gấp 6 lần so với ngẫu nhiên.

Kết luận quan trọng

Điều này khẳng định **Giao hàng (Logistics)** là yếu tố kích hoạt chính (primary trigger) của sự không hài lòng.

- Các vấn đề thường không đứng riêng lẻ mà cộng hưởng (ví dụ: Trễ + Hàng hỏng).
- Ngưỡng chịu đựng của khách hàng rất thấp đối với việc "Giao hàng rất trễ".

ĐÁNH GIÁ TỔNG THỂ

So sánh Apriori & FP-Growth

Đối chiếu hiệu năng, độ chính xác và khả năng ứng dụng

TIÊU CHÍ SO SÁNH	APRIORI	FP-GROWTH	ĐÁNH GIÁ CHUNG
Độ chính xác	**100%** Trùng khớp hoàn toàn	**100%** Trùng khớp hoàn toàn	✓ Ngang nhau
Số luật phát hiện	**237 luật**	**237 luật**	✓ Ngang nhau
Thời gian (Tập này)	**~0.1195s** ▲ Nhanh hơn	**~0.1338s** Chậm hơn không đáng kể	Chênh lệch nhỏ với dữ liệu vừa & nhỏ.
Khả năng mở rộng	Trung bình Quét CSDL nhiều lần (k lần)	ƯU VIỆT Chỉ quét CSDL 2 lần (FP-Tree)	FP-Growth vượt trội khi dữ liệu lớn.

Khi nào dùng Apriori?

Phù hợp cho các tập dữ liệu **vừa và nhỏ**, hoặc mục đích nghiên cứu, giảng dạy do tính đơn giản và dễ cài đặt.

Khi nào dùng FP-Growth?

Lựa chọn tối ưu cho hệ thống **Big Data**, dữ liệu giao dịch lớn cần tốc độ xử lý cao và bộ nhớ hiệu quả.

ĐÁNH GIÁ TỔNG THỂ

So sánh Apriori & FP-Growth

Đối chiếu hiệu năng, độ chính xác và khả năng ứng dụng

TIÊU CHÍ SO SÁNH	APRIORI	FP-GROWTH	ĐÁNH GIÁ CHUNG
Độ chính xác	**100%** Trùng khớp hoàn toàn	**100%** Trùng khớp hoàn toàn	✓ Ngang nhau
Số luật phát hiện	**237 luật**	**237 luật**	✓ Ngang nhau
Thời gian (Tập này)	**~0.1195s** ▲ Nhanh hơn	**~0.1338s** Chậm hơn không đáng kể	Chênh lệch nhỏ với dữ liệu vừa & nhỏ.
Khả năng mở rộng	Trung bình Quét CSDL nhiều lần (k lần)	ƯU VIỆT Chỉ quét CSDL 2 lần (FP-Tree)	FP-Growth vượt trội khi dữ liệu lớn.

Khi nào dùng Apriori?

Phù hợp cho các tập dữ liệu **vừa và nhỏ**, hoặc mục đích nghiên cứu, giảng dạy do tính đơn giản và dễ cài đặt.

Khi nào dùng FP-Growth?

Lựa chọn tối ưu cho hệ thống **Big Data**, dữ liệu giao dịch lớn cần tốc độ xử lý cao và bộ nhớ hiệu quả.

Kết luận chính

Tổng hợp kết quả thực thi & giá trị nghiên cứu

Thuật toán ổn định & Hiệu quả

Cả **Apriori** và **FP-Growth** đều cho kết quả trùng khớp 100% (111 tập phổ biến, 237 luật). Dữ liệu nhất quán và đáng tin cậy.

Giá trị thực tiễn cao

Các luật kết hợp không ngẫu nhiên mà phản ánh chính xác hành vi thực tế, tạo cơ sở vững chắc cho các quyết định quản trị.

Nguyên nhân cốt lõi: Giao hàng

Khẳng định **Vận chuyển/Giao hàng** là yếu tố chính gây ra đánh giá thấp. Các vấn đề thường xuất hiện đồng thời (Combo lỗi).

Cải tiến dựa trên dữ liệu

Kết quả cung cấp lộ trình rõ ràng để tối ưu quy trình logistics và nâng cao trải nghiệm khách hàng một cách khoa học.

Cảm ơn đã theo dõi!