CUSTOMER SHOPPING DATA MINING

Phân Tích Hành Vi Mua Sắm Khách Hàng Tại Istanbul

Huỳnh Nhật Thành, Lê Thành Danh, Nguyễn Quốc Thuận, Nguyễn Trí Sự

GIỚI THIỆU

Đặt vấn đề: Dữ liệu mua sắm chứa đựng nhiều thông tin ẩn về hành vi khách hàng. Dự án áp dụng các kỹ thuật Data Mining để phân tích **99,457 giao dịch**.

Nguồn dữ liệu: https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset/data

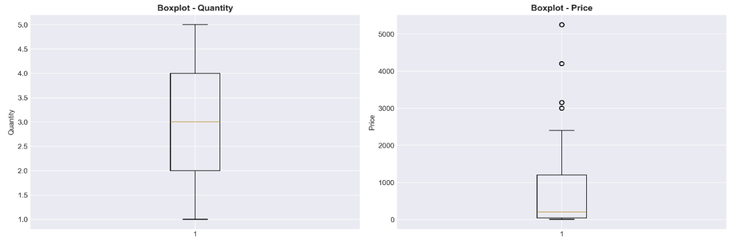
Mục tiêu chính:

- Phân nhóm khách hàng (Segmentation).
- Tìm luật kết hợp mua sắm (Market Basket).
- Dự báo hành vi tiêu dùng (Prediction).

TIỀN XỬ LÝ DỮ LIỆU

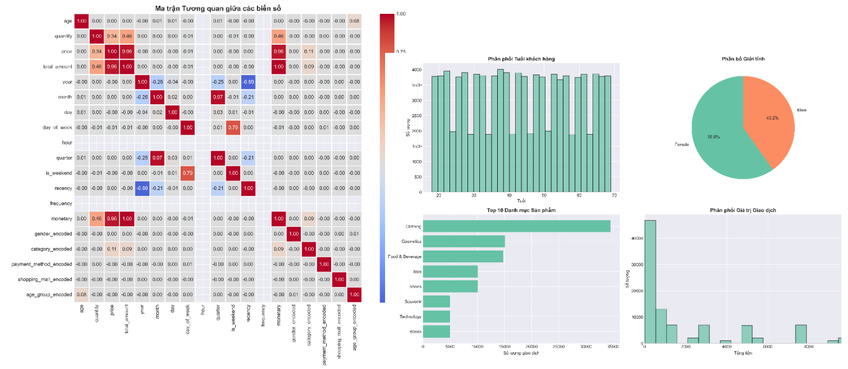
Quy trình xử lý dữ liệu:

1. **Cleaning:** Xử lý Missing Values, loại bỏ trùng lặp.
2. **Outliers**: Phát hiện ngoại lai bằng IQR (Quantity, Price). Quyết định giữ lại vì là giao dịch mua sỉ.
3. **Feature Engineering:**
   - RFM (Recency, Frequency, Monetary).
   - Nhóm tuổi (Teen, Adult, Senior).



PHÂN TÍCH KHÁM PHÁ DỮ LIỆU (EDA)

Hai biểu đồ cho thấy đặc điểm phân bố dữ liệu khách hàng và mối quan hệ giữa các biến, trong đó *monetary* có tương quan mạnh với *price* và *quantity* còn các biến khác tương quan thấp.

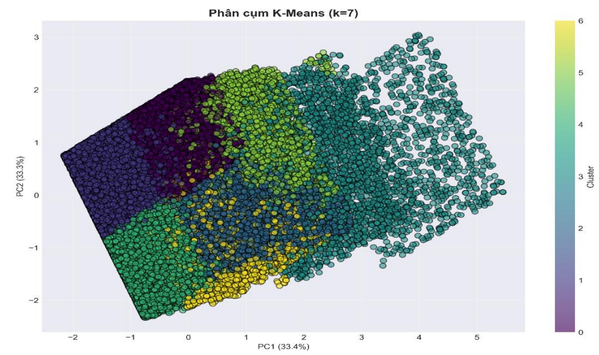


PHƯƠNG PHÁP 1: PHÂN CỤM

Sử dụng bộ chỉ số **RAM (Recency - Age - Monetary)** để phân nhóm khách hàng theo hành vi mua sắm và độ tuổi, nhằm tối ưu hóa các chiến dịch Marketing cá nhân hóa. So sánh hiệu quả phân cụm giữa thuật toán **K-Means** và **Hierarchical Clustering**.

Thuật toán	Số cụm (k)	Silhouette Score
K-Means	7	0.3325
Hierarchical	3	0.2665

K-Means (k=7 gộp thành 4 nhóm) tối ưu cho quy mô lớn, trong khi Hierarchical (k=3) phù hợp cho quy mô nhỏ.

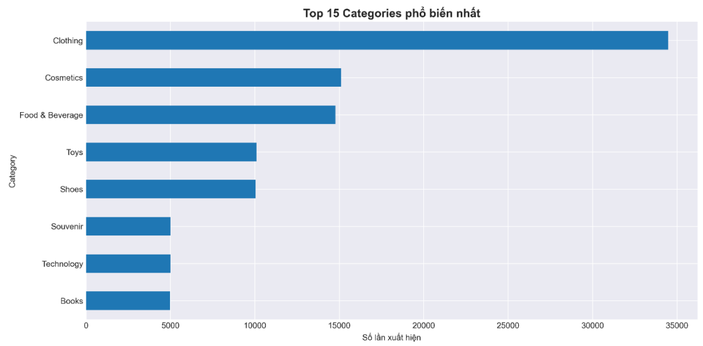


PHƯƠNG PHÁP 2: LUẬT KẾT HỢP

Sử dụng **Apriori** và **FP-Growth.**

Tần suất xuất hiện (Support):

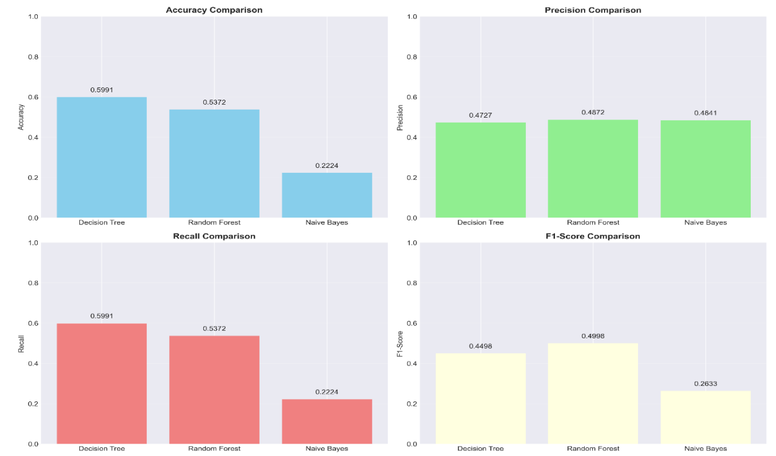
- Clothing: 34.7% (Cao nhất)
- Cosmetics: 15.2%
- Food & Beverage: 14.9%



Không tìm thấy luật kết hợp mạnh (Lift < 2). Khách hàng có xu hướng mua lẻ từng món.

PHƯƠNG PHÁP 3: PHÂN LOẠI

Dự đoán danh mục sản phẩm (Category) dựa trên thông tin nhân khẩu học (Age, Gender, Mall...).



Model	Accuracy	F1-Score
Decision Tree	59.91%	0.450
Random Forest	53.72%	0.500
Naive Bayes	22.24%	0.263

Kết quả: Decision Tree cho độ chính xác cao nhất (~60%), phù hợp để giải thích các quyết định mua hàng đơn giản.

KẾT LUẬN

Tổng kết dự án:

Quy trình toàn diện: Hoàn thiện luồng khai phá dữ liệu từ Tiền xử lý, EDA đến Mô hình hóa.

Phân cụm linh hoạt: K-Means (k=7) gộp thành 4 nhóm) tối ưu cho quy mô lớn, trong khi Hierarchical (k=3) phù hợp cho quy mô nhỏ.

Insight ngành hàng: Thời trang (Clothing) là ngành hàng chủ lực, chiếm ưu thế tuyệt đối về doanh thu và tần suất.

Khuyến nghị chiến lược:

Tập trung nguồn lực: Ưu tiên ngân sách Marketing cho ngành hàng Clothing.

Chiến lược phân hóa: Thay vì dàn trải, hãy cá nhân hóa hành động cho 4 nhóm khách hàng mục tiêu (VIP, Tiềm năng, Phổ thông, Nguy cơ).