

DỰ ĐOÁN KHÁCH HÀNG TIỀM NĂNG MUA SẮM DỰA TRÊN DỮ LIỆU GIAO DỊCH VÀ HỒ SƠ KHÁCH HÀNG

GVHD: TS. LÊ TRỌNG NGỌC
Thuyết trình bởi nhóm Data Novices



NỘI DUNG

- 1. Bối Cảnh
- 2. Mục tiêu
- 3. Giới thiệu dữ liệu
- 4. Trực quan hóa dữ liệu
- 5. Xử lý dữ liệu
- 6. Xây dựng mô hình
- 7. Hiệu suất mô hình
- 8. Cải thiện mô hình
- 9. Đánh giá tổng quan
- 10. Kết luận

1. BỐI CẢNH

1. BỐI CẢNH

Trong thời đại số hóa và sự phát triển mạnh mẽ của thương mại điện tử, việc thấu hiểu hành vi khách hàng và dự đoán những khách hàng tiềm năng đã trở thành một trong những yếu tố quan trọng giúp doanh nghiệp tối ưu hóa chiến lược kinh doanh. Dự đoán chính xác khách hàng có khả năng mua sắm không chỉ giúp tiết kiệm chi phí tiếp thị, mà còn tạo điều kiện cho doanh nghiệp cung cấp các sản phẩm, dịch vụ phù hợp và kịp thời, từ đó nâng cao trải nghiệm của khách hàng.

2. MỤC TIÊU

2. MỤC TIÊU

1. Xác định các khách hàng tiềm năng

2. Tối ưu hóa chiến lược tiếp thị

3. Nâng cao hiệu quả của các chiến dịch tiếp thị

3. GIỚI THIỆU VỀ DỮ LIỆU

3. GIỚI THIỆU VỀ DỮ LIỆU

Dữ liệu này bao gồm thông tin chi tiết về khách hàng của một doanh nghiệp, cung cấp cái nhìn sâu sắc về nhiều khía cạnh khác nhau như thông tin cá nhân, hành vi mua sắm,.... cũng như mức độ tham gia của khách hàng trong các chiến dịch tiếp thị và quảng cáo.

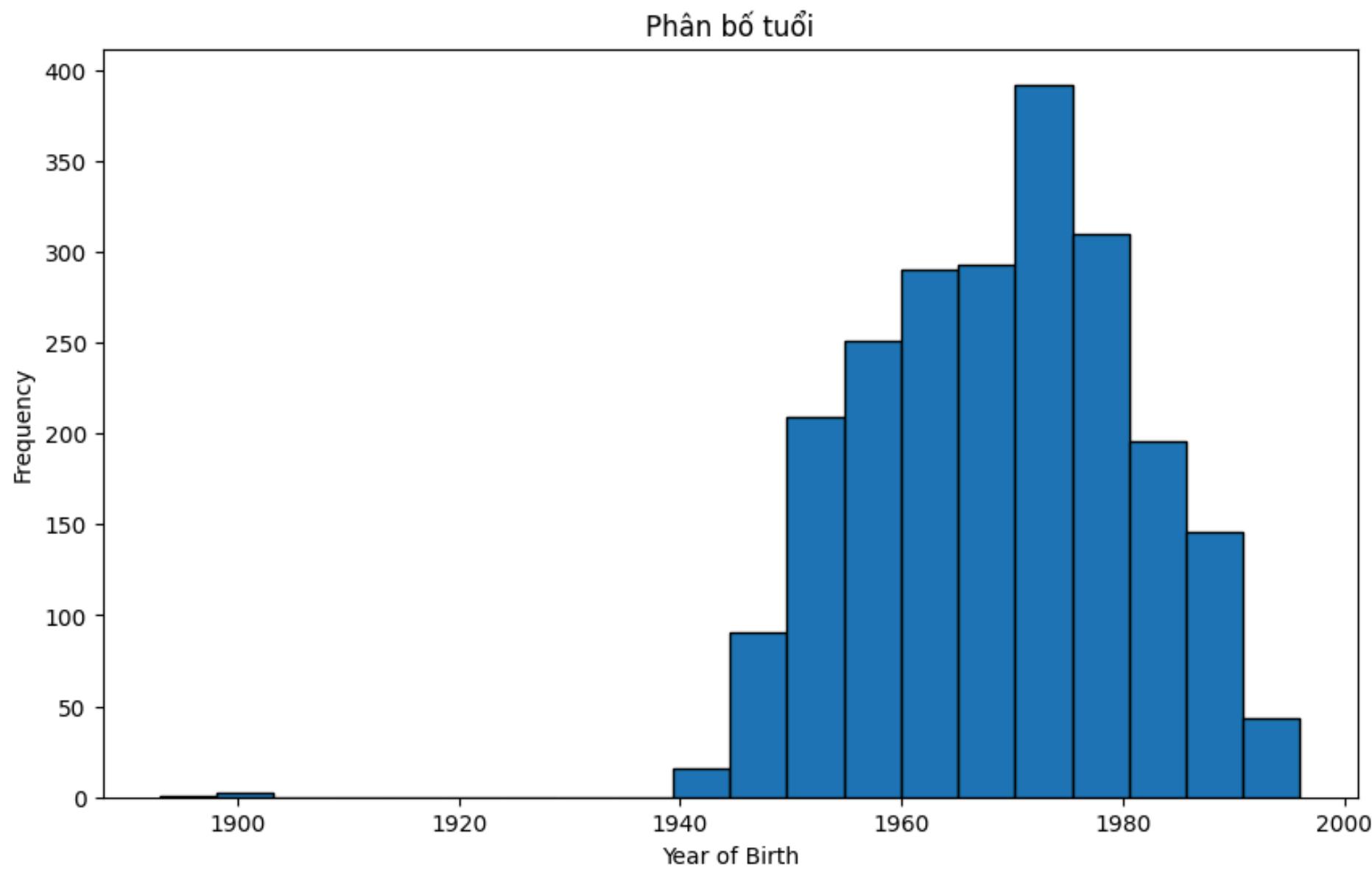
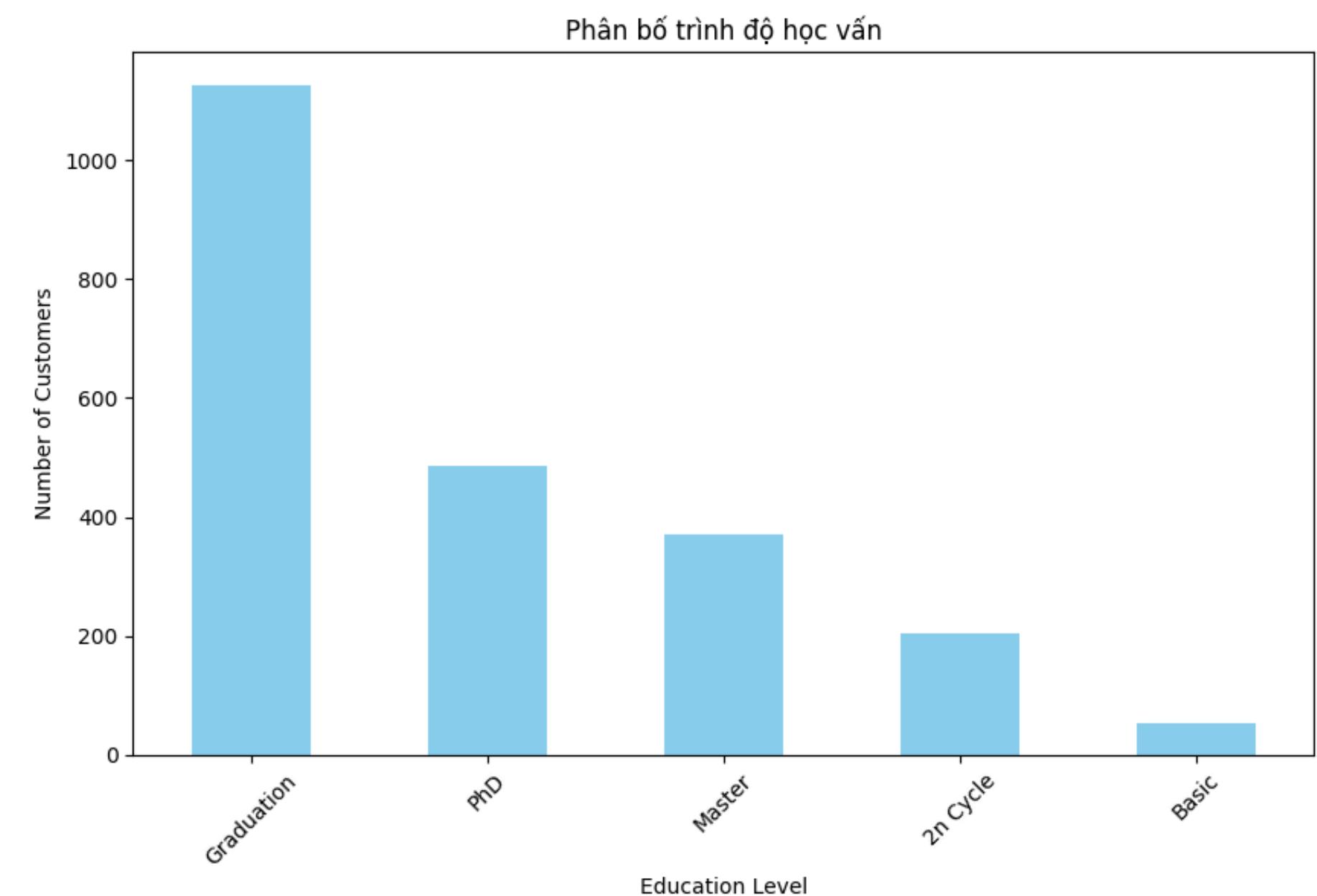
Dữ liệu bao gồm có 29 cột và 2240 dòng.

3. GIỚI THIỆU VỀ DỮ LIỆU

- **ID Khách Hàng:** Mã định danh duy nhất cho mỗi khách hàng.
- **Year_Birth:** Năm sinh của khách hàng, từ đó tính toán tuổi.
- **Education:** Trình độ học vấn của khách hàng.
- **Marital_Status:** Tình trạng hôn nhân của khách hàng.
- **Income:** Thu nhập hàng năm của khách hàng.
- **Kidhome:** Số lượng trẻ em trong hộ gia đình.
- **Teenhome:** Số lượng thanh thiếu niên trong hộ gia đình.
- **Dt_Customer:** Ngày khách hàng được ghi nhận vào cơ sở dữ liệu.
- **Recency:** Số ngày kể từ lần giao dịch cuối cùng của khách hàng.
- **MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds:** Số tiền chi tiêu cho các loại sản phẩm khác nhau.
- **NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases:** Số lượng giao dịch qua các kênh khác nhau.
- **NumWebVisitsMonth:** Số lượt truy cập vào website mỗi tháng.
- **AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5:** Chỉ số cho biết khách hàng có chấp nhận các chiến dịch tiếp thị cụ thể không.
- **Complain:** Chỉ số cho biết khách hàng có khiếu nại hay không.
- **Z_CostContact, Z_Revenue:** Chi phí liên hệ và doanh thu từ các chiến dịch tiếp thị.
- **Response:** Chỉ số cho biết khách hàng có phản hồi lại chiến dịch tiếp thị không.

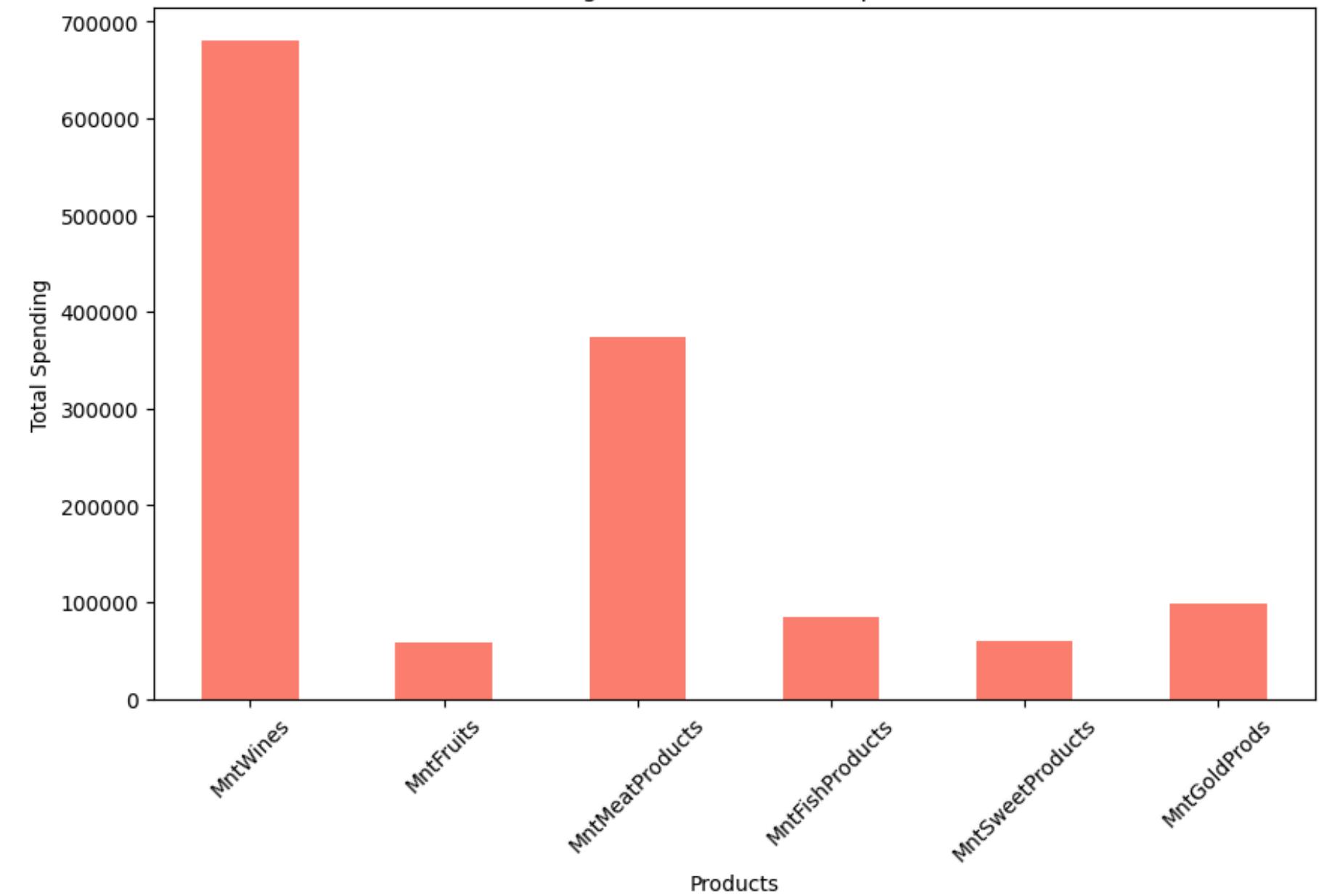
4. TRỰC QUAN HÓA DỮ LIỆU

4. TRỰC QUAN HÓA DỮ LIỆU

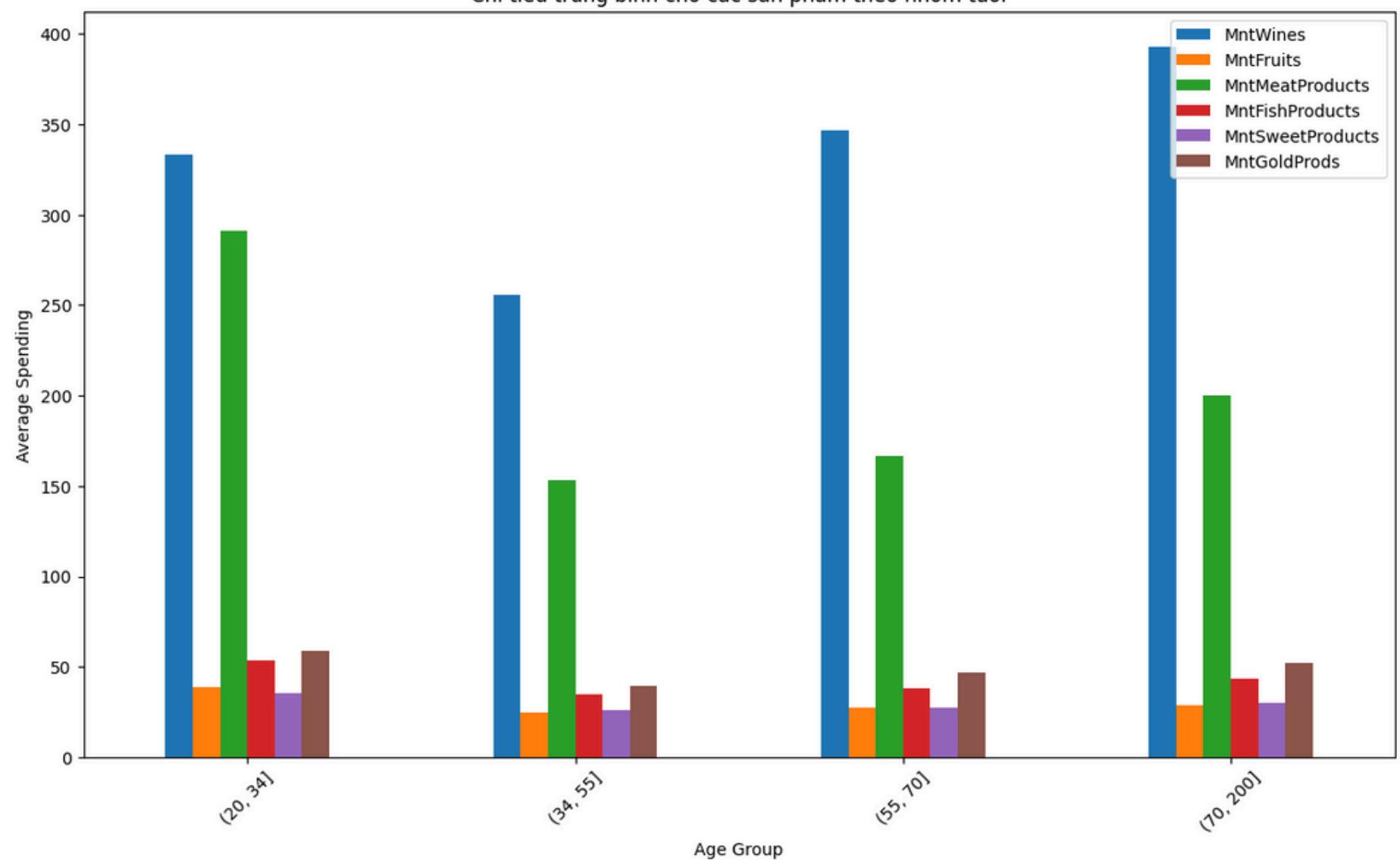


4. TRỰC QUAN HÓA DỮ LIỆU

Tổng chi tiêu cho các sản phẩm

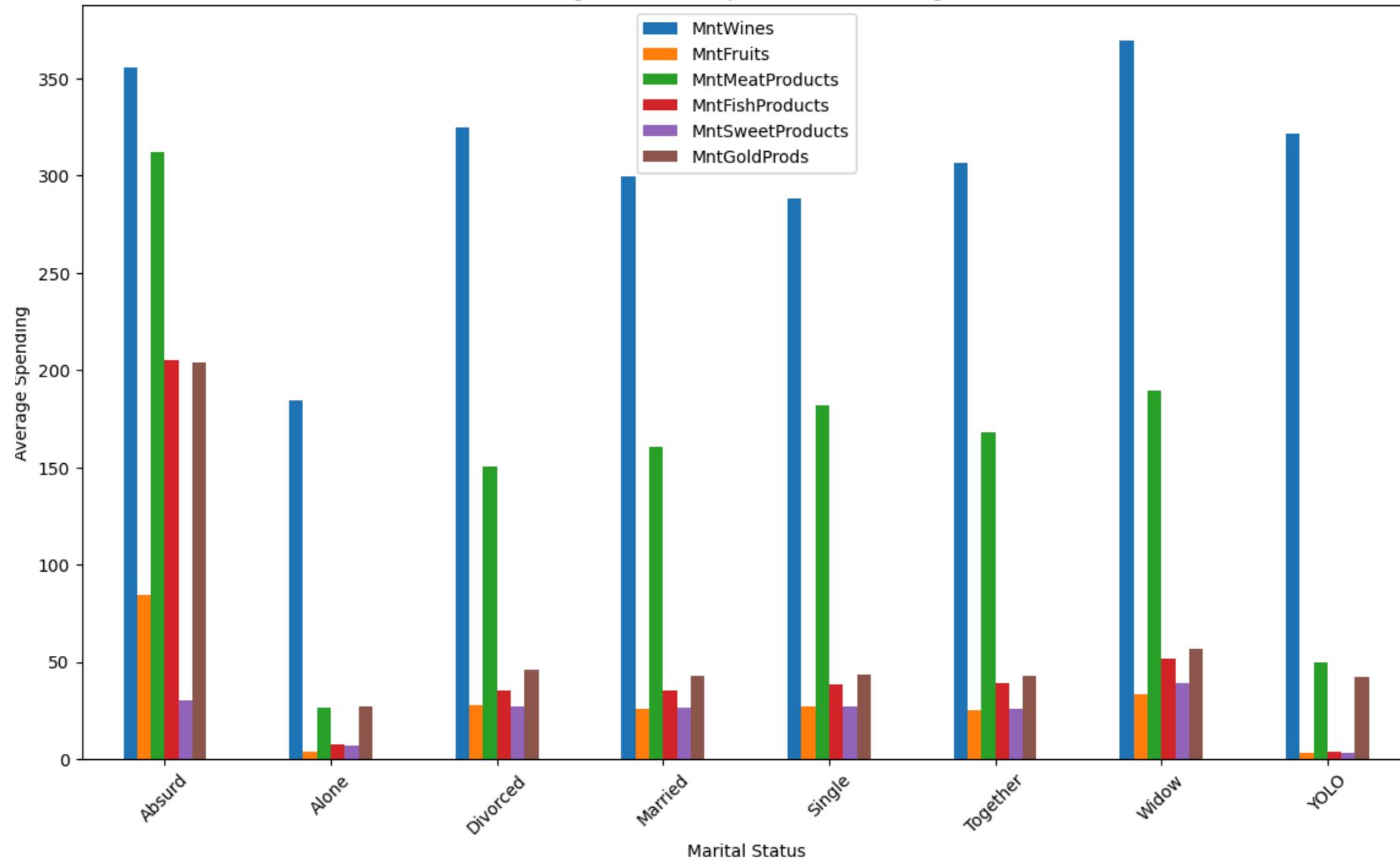


Chi tiêu trung bình cho các sản phẩm theo nhóm tuổi

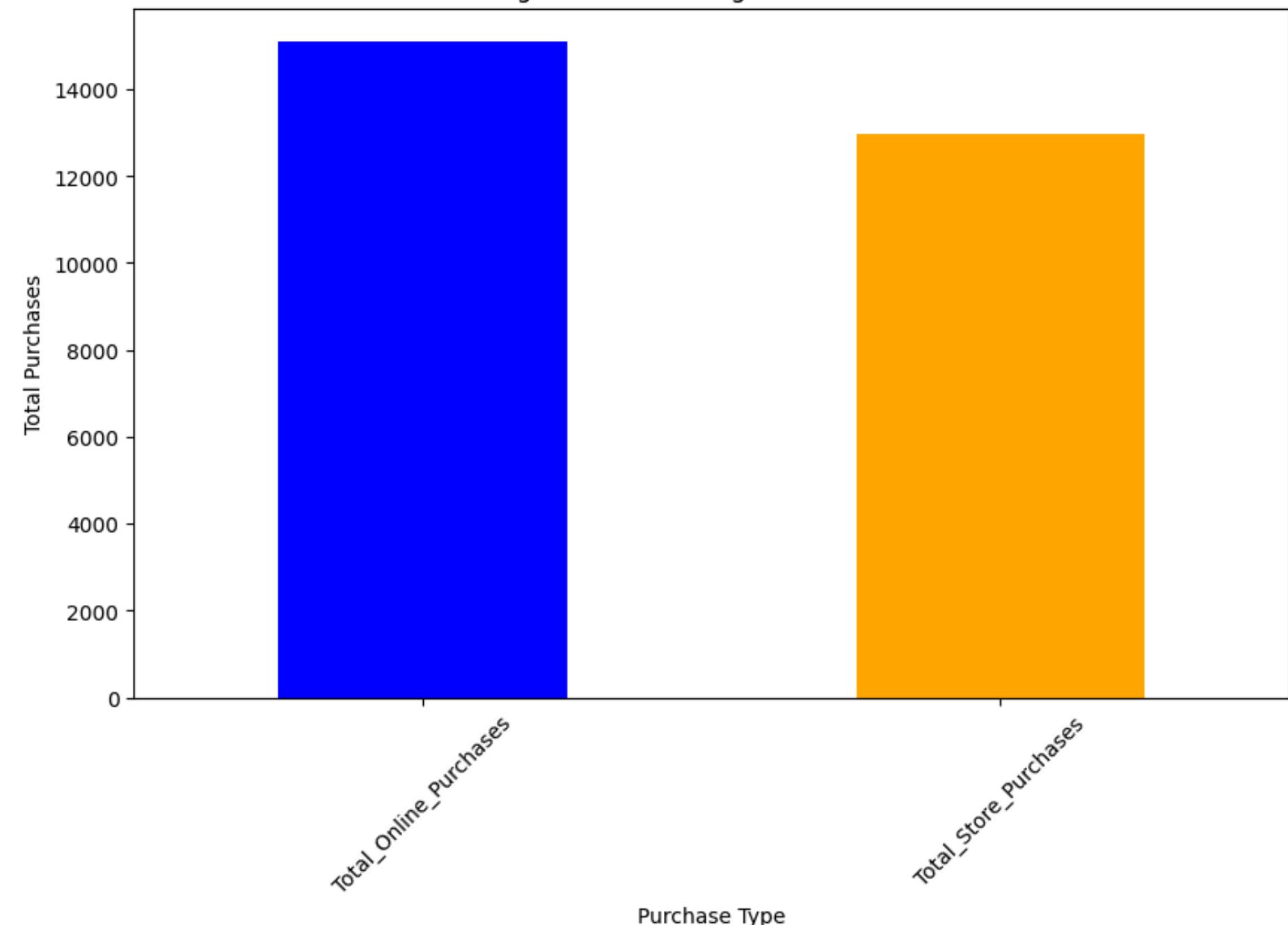


4. TRỰC QUAN HÓA DỮ LIỆU

Chi tiêu trung bình cho sản phẩm theo tình trạng hôn nhân

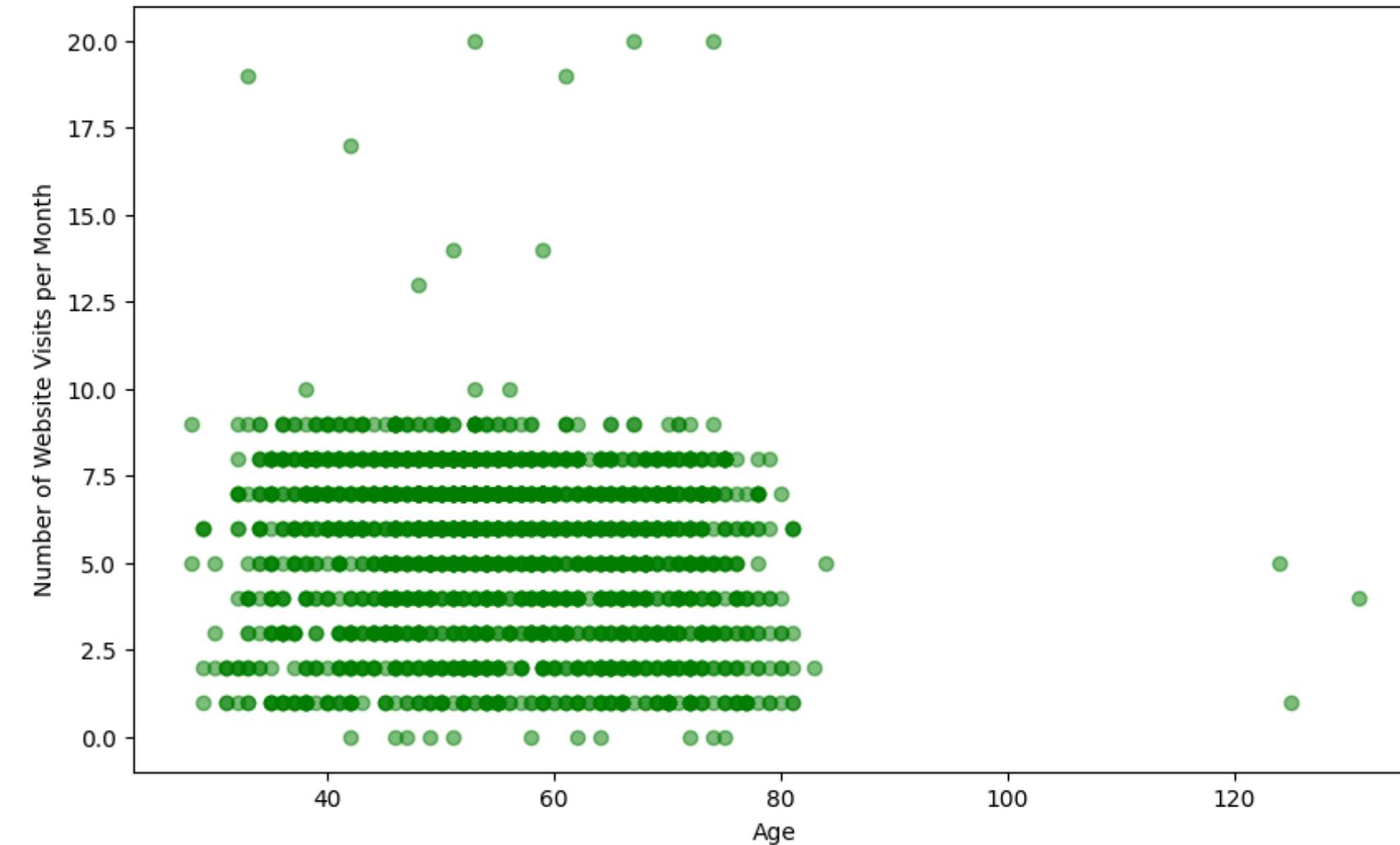


Tổng số lần mua hàng: Online và Store

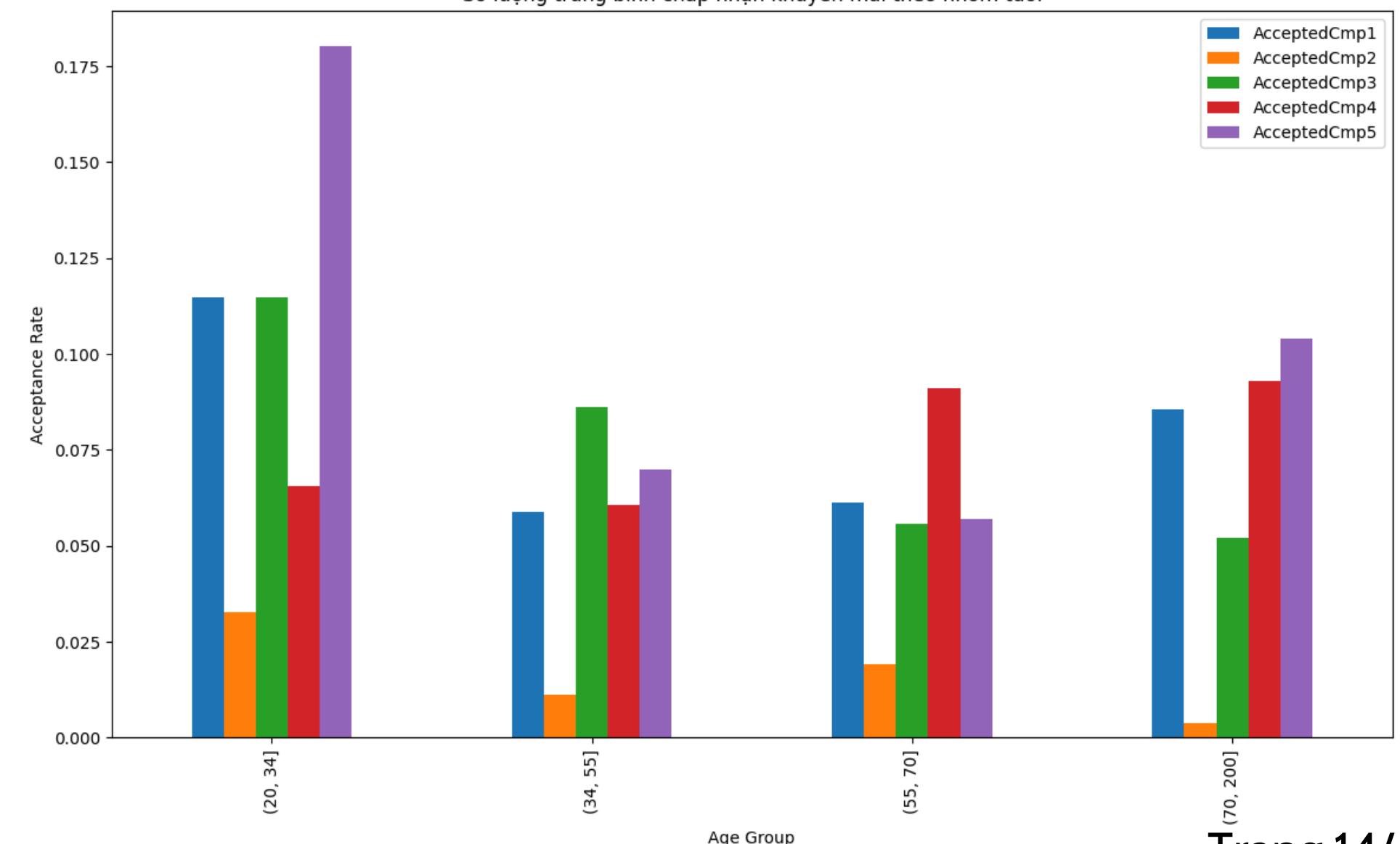


4. TRỰC QUAN HÓA DỮ LIỆU

Mối quan hệ giữa độ tuổi so với số lượt truy cập trang web mỗi tháng



Số lượng trung bình chấp nhận khuyến mãi theo nhóm tuổi



5. XỬ LÝ DỮ LIỆU

5. XỬ LÝ DỮ LIỆU

Loại bỏ các dòng hoặc cột có quá nhiều giá trị thiếu.

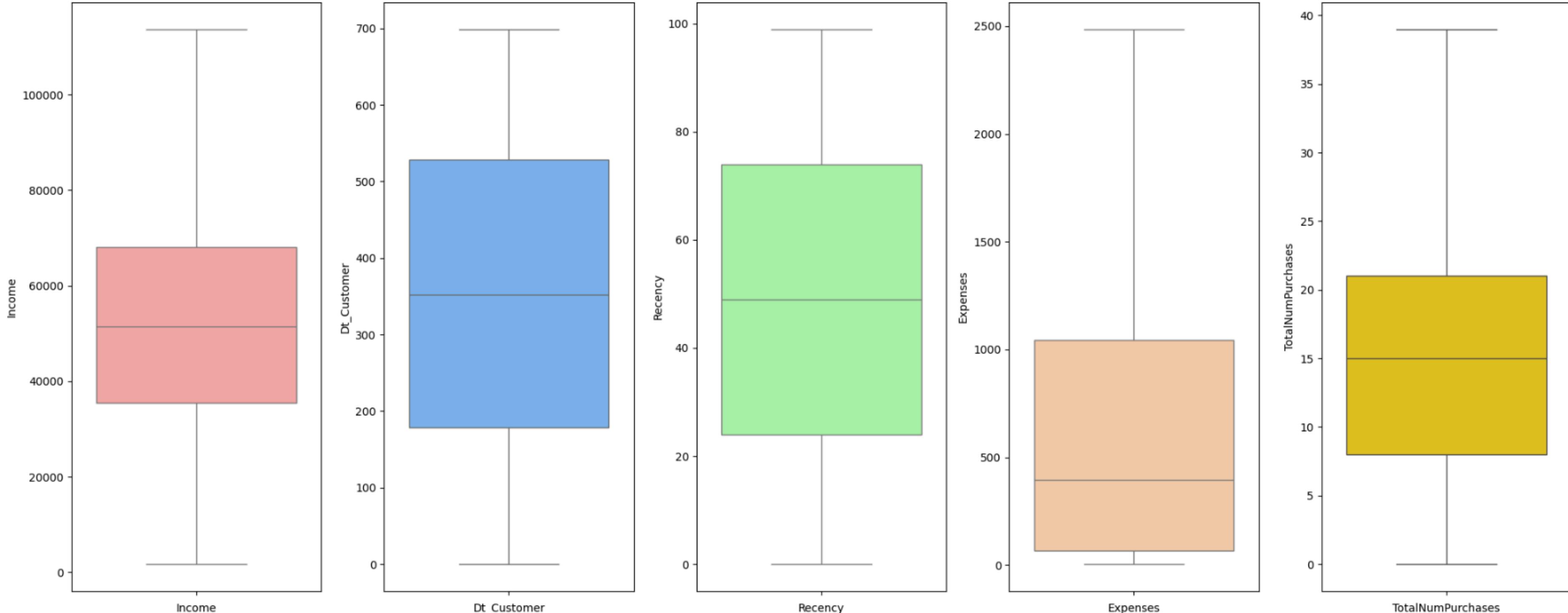
```
df.drop_duplicates(inplace=True) # Loại bỏ các hàng trùng lặp trong DataFrame và cập nhật trực tiếp trên df.  
df.dropna(inplace=True) # Loại bỏ các hàng có giá trị thiếu (NaN) và cập nhật trực tiếp trên df.  
df.shape # Trả về kích thước của DataFrame sau khi loại bỏ.
```

```
df.drop('Year_Birth', axis=1, inplace=True) # Loại bỏ cột 'Year_Birth' khỏi DataFrame và cập nhật trực tiếp lên df.
```

```
data.drop(['ID', 'Z_CostContact', 'Z_Revenue'], axis=1, inplace=True) # Loại bỏ các cột 'ID', 'Z_CostContact', 'Z_Revenue' khỏi DataFrame.
```

5. XỬ LÝ DỮ LIỆU

Sử dụng biểu đồ hộp (boxplot) để nhận diện ngoại lai.



5. XỬ LÝ DỮ LIỆU

Gộp các cột có sự tương đồng lại với nhau để giảm chiều dữ liệu.

- Số trẻ
- Giảm không gian về tình trạng hôn nhân
- Số tiền chi tiêu cho các loại sản phẩm khác nhau
- Số lượng giao dịch qua các kênh khác nhau
- Chỉ số cho biết khách hàng có nhận các chiến dịch

5. XỬ LÝ DỮ LIỆU

Label Encoding: Chuyển các nhãn định tính thành các giá trị số nguyên (0, 1, 2,...).

One-Hot Encoding: Tạo các cột nhị phân cho mỗi giá trị trong biến định tính để tránh ảnh hưởng thứ tự trong mô hình học máy.

6. XÂY DỰNG MÔ HÌNH

6. XÂY DỰNG MÔ HÌNH

K-Means:

- Chọn số cụm (k) bằng Elbow Method hoặc Silhouette Score. K-Means phân cụm khách hàng dựa trên khoảng cách đến tâm cụm, lặp lại cho đến khi ổn định.

Logistic Regression:

- Dự đoán khả năng mua sắm của khách hàng qua hàm sigmoid, tối ưu hóa các tham số để phân loại khách hàng có thể mua hàng.

Random Forest:

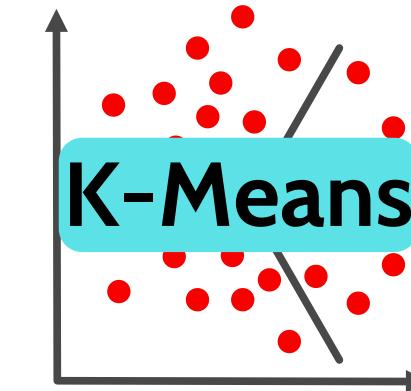
- Tạo nhiều cây quyết định từ mẫu ngẫu nhiên và kết hợp dự đoán, giúp tăng độ chính xác và giảm overfitting trong phân loại khách hàng tiềm năng.

Mô hình phân loại kết hợp phân cụm (LogisticRegression - K-Means):

- Kết hợp cả hai kỹ thuật phân cụm và phân loại để cải thiện khả năng dự đoán và phân tích dữ liệu.

7. HIỆU SUẤT MÔ HÌNH

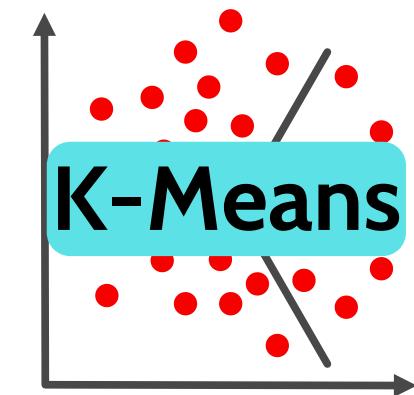
7. Hiệu suất mô hình



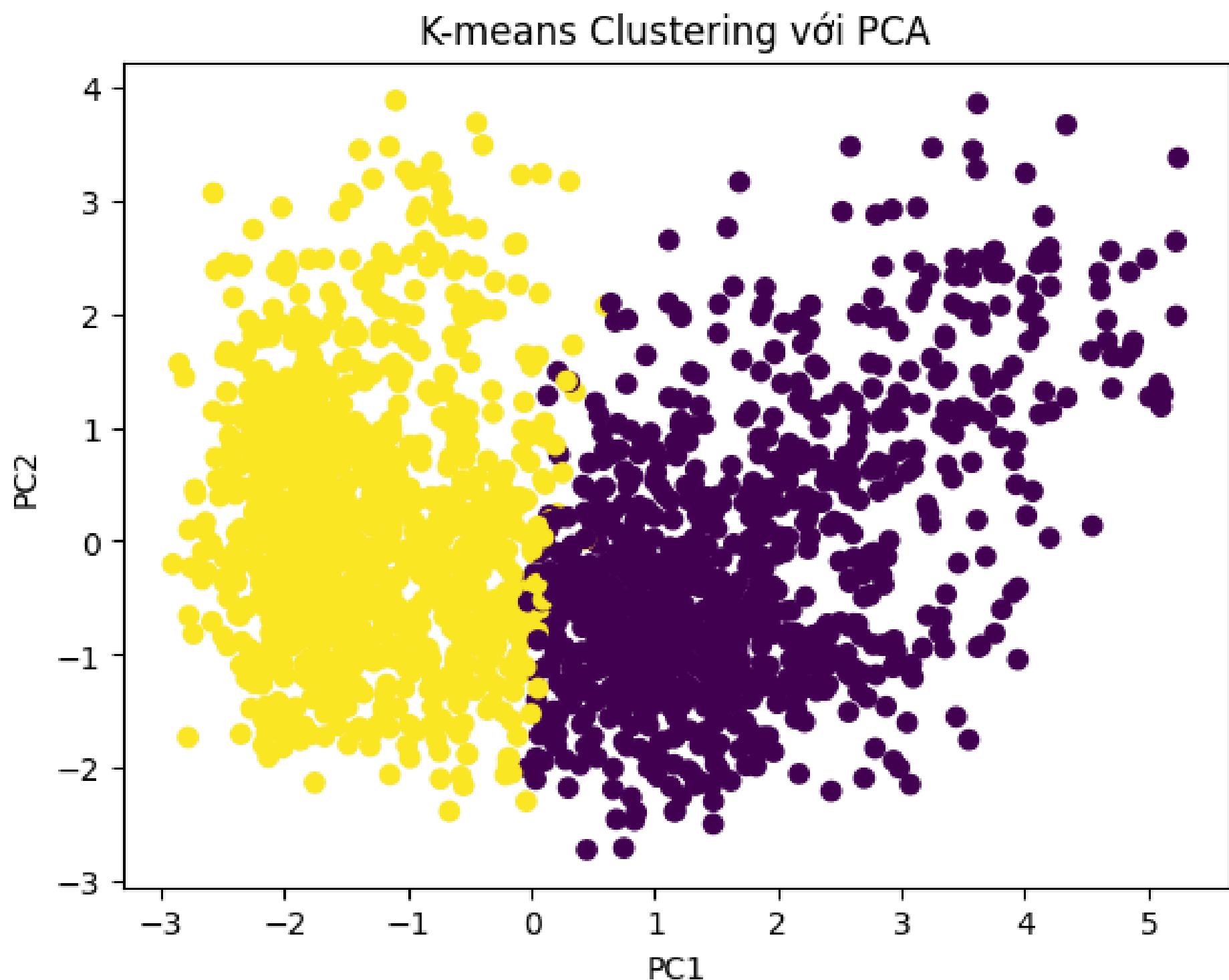
Silhouette Score = 0.1988 khi số cụm là 2 cho thấy rằng các cụm được hình thành không tách biệt rõ ràng. Ta có nhận xét như sau:

- Cụm không đủ chặt chẽ: Giá trị Silhouette Score dưới 0.5 thường cho thấy các đối tượng trong cùng một cụm không giống nhau nhiều, hoặc chúng gần các đối tượng trong cụm khác. Điều này nghĩa là các đối tượng trong cụm có sự phân tán lớn.
- Tách biệt giữa các cụm kém: Vì Silhouette Score khá thấp, các cụm có thể không được phân biệt tốt với nhau. Các đối tượng trong các cụm có thể nằm ở gần ranh giới của các cụm khác, dẫn đến việc phân cụm không rõ ràng.

7. Hiệu suất mô hình



Sử dụng PCA để giảm chiều dữ liệu sau khi phân cụm để trực quan hóa



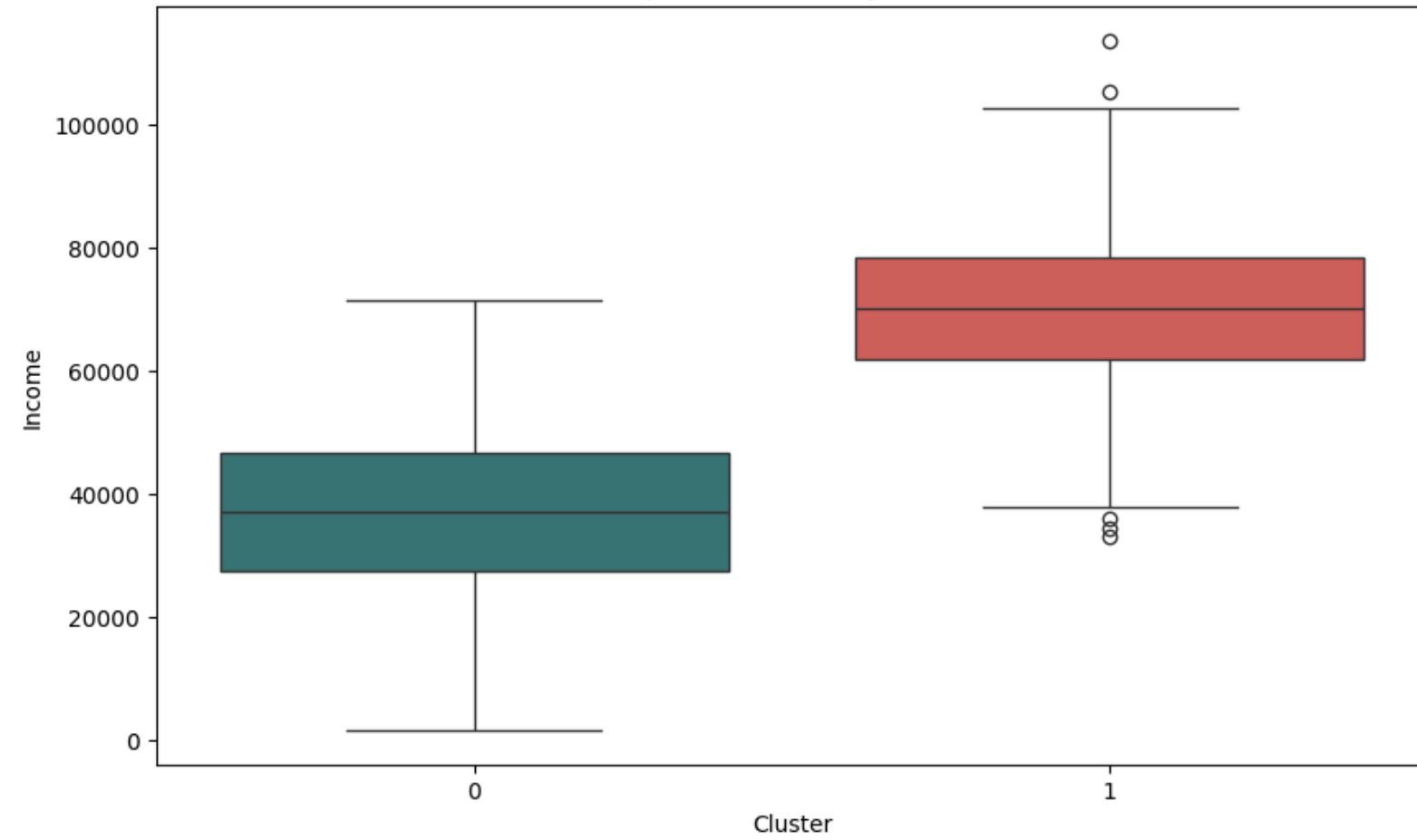
7. Hiệu suất mô hình



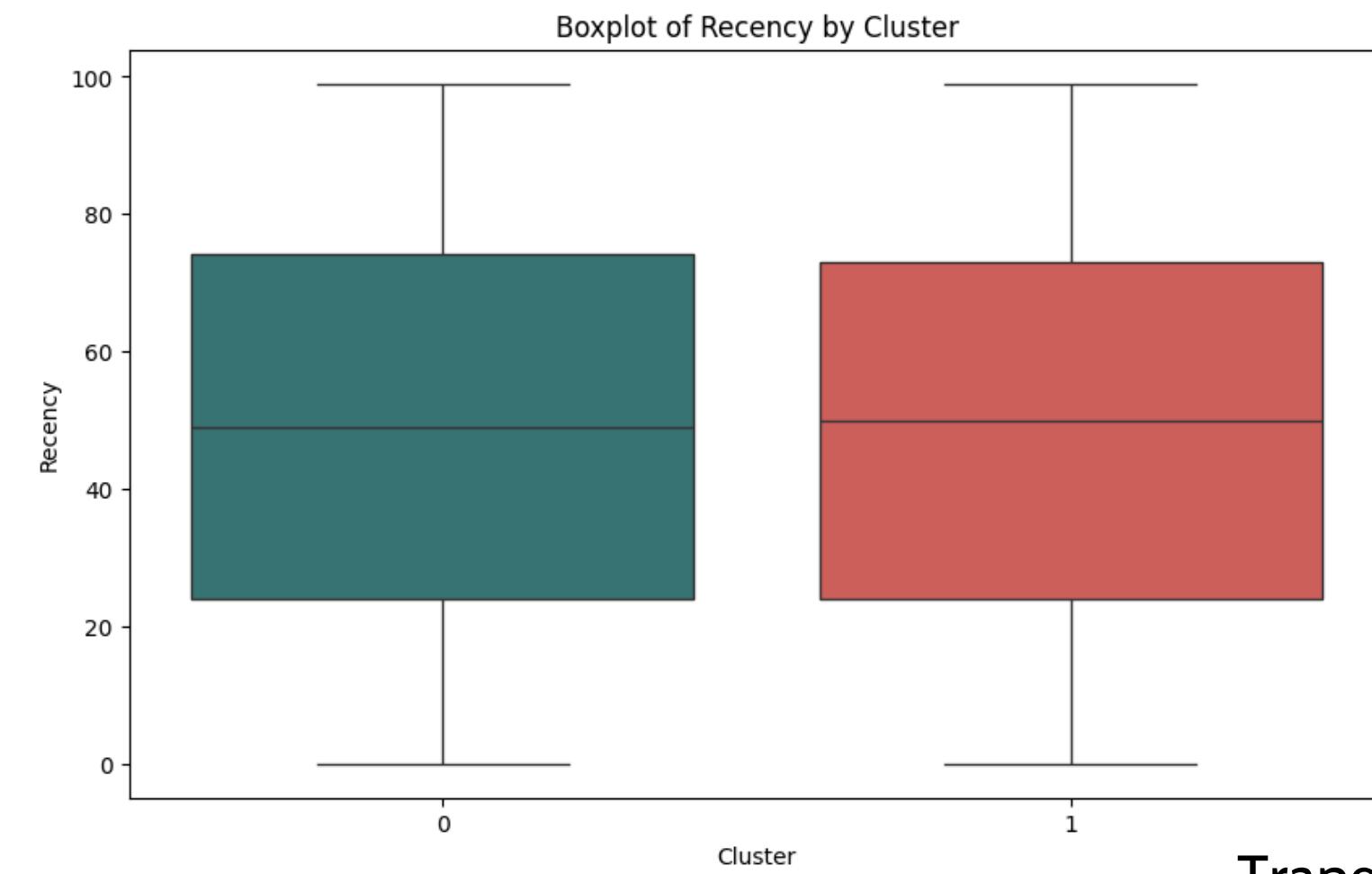
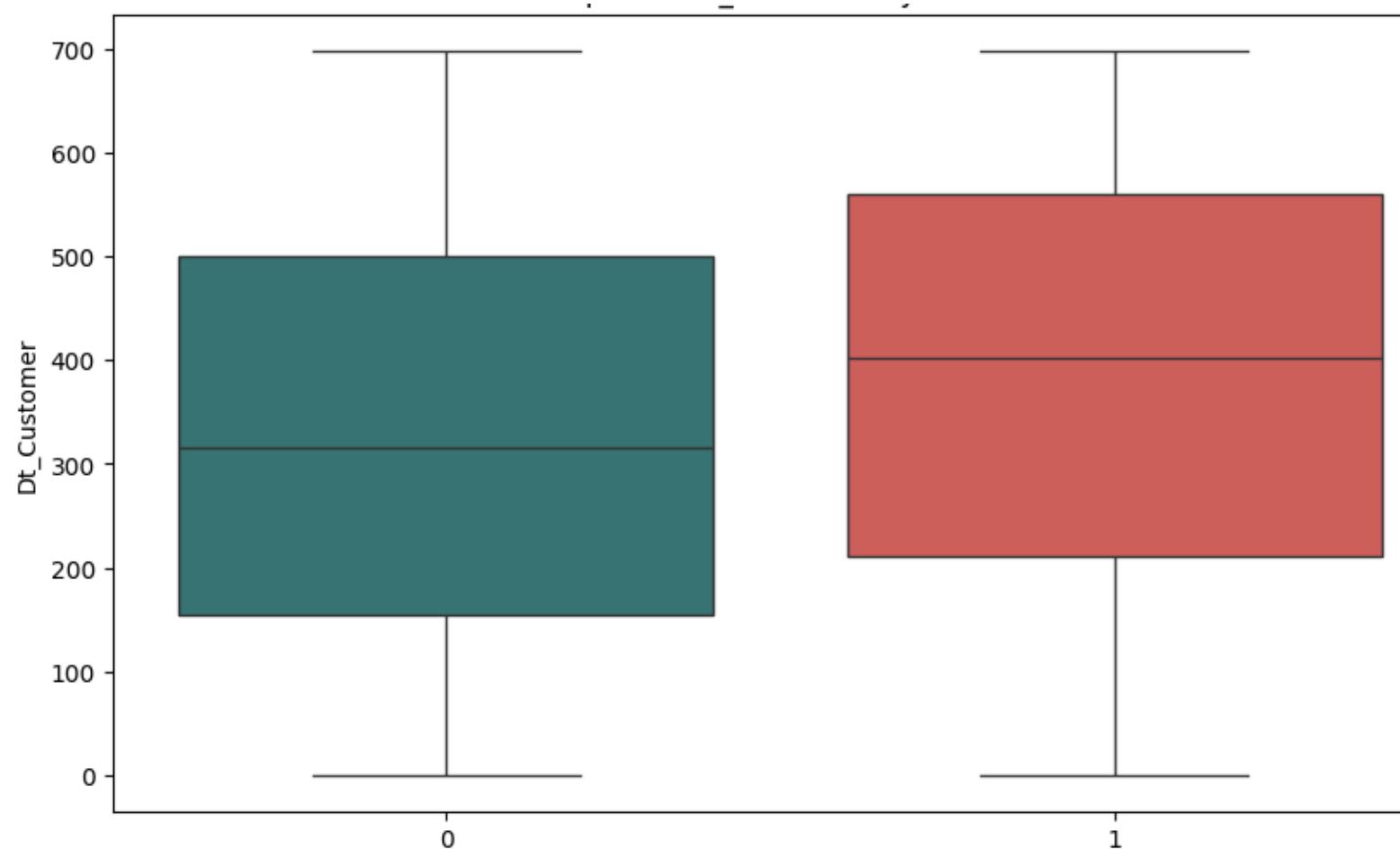
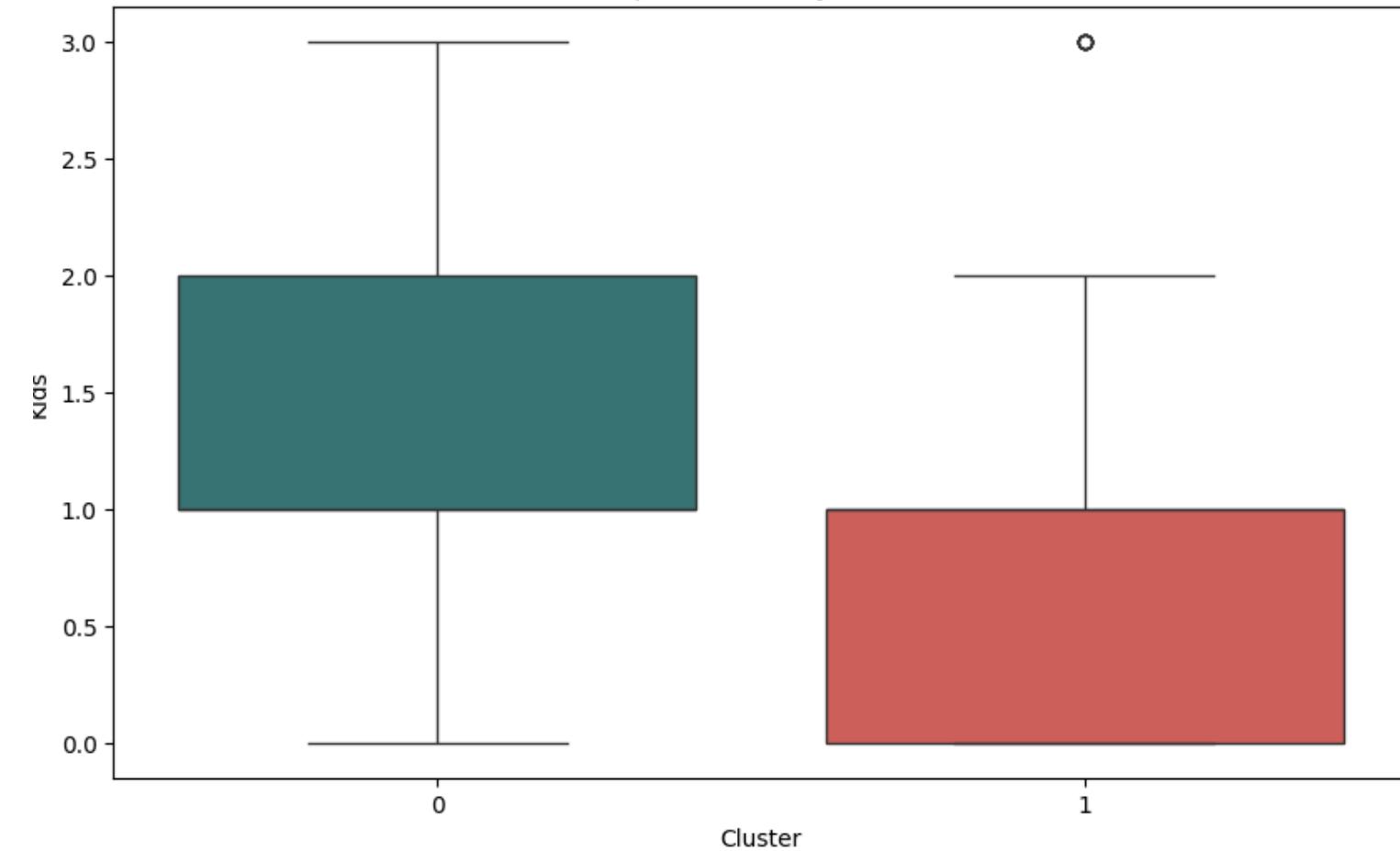
7. Hiệu suất mô hình

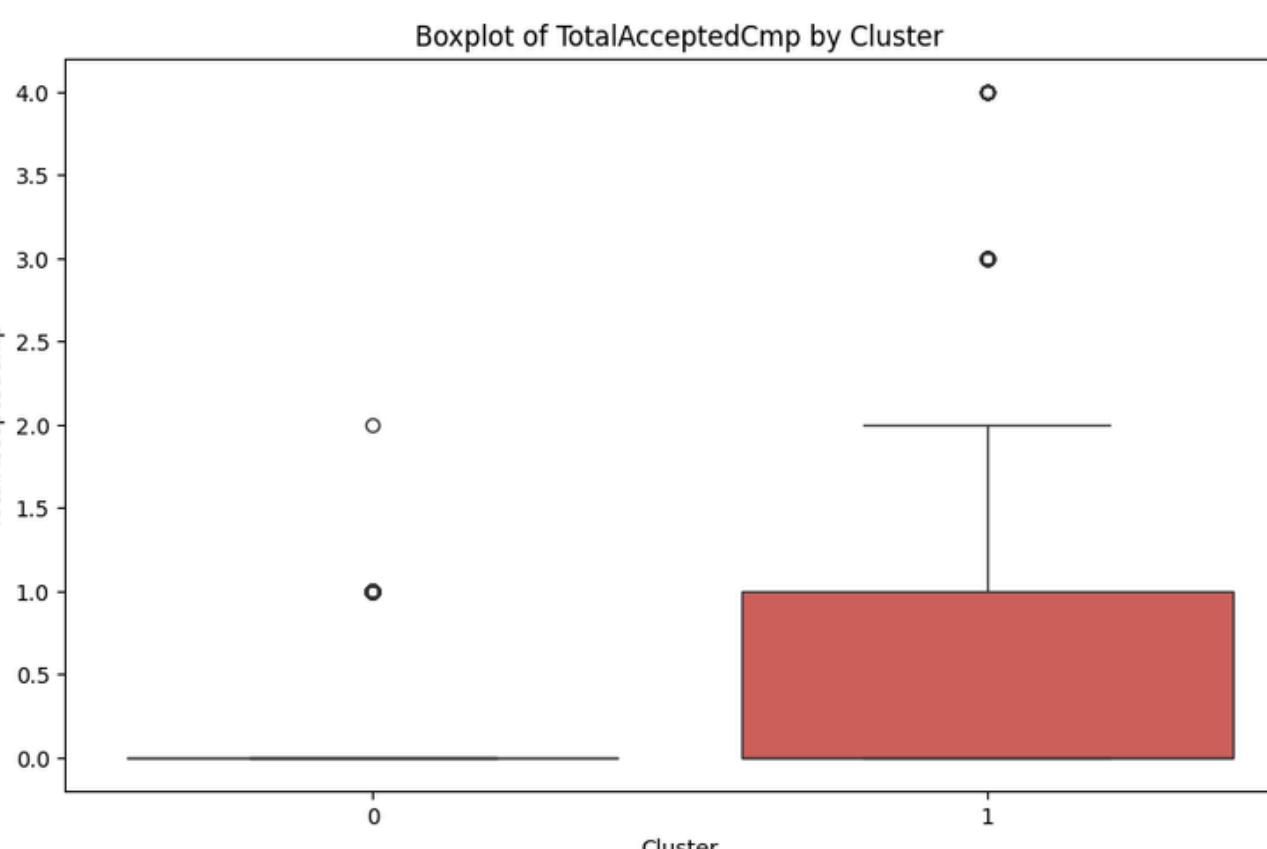
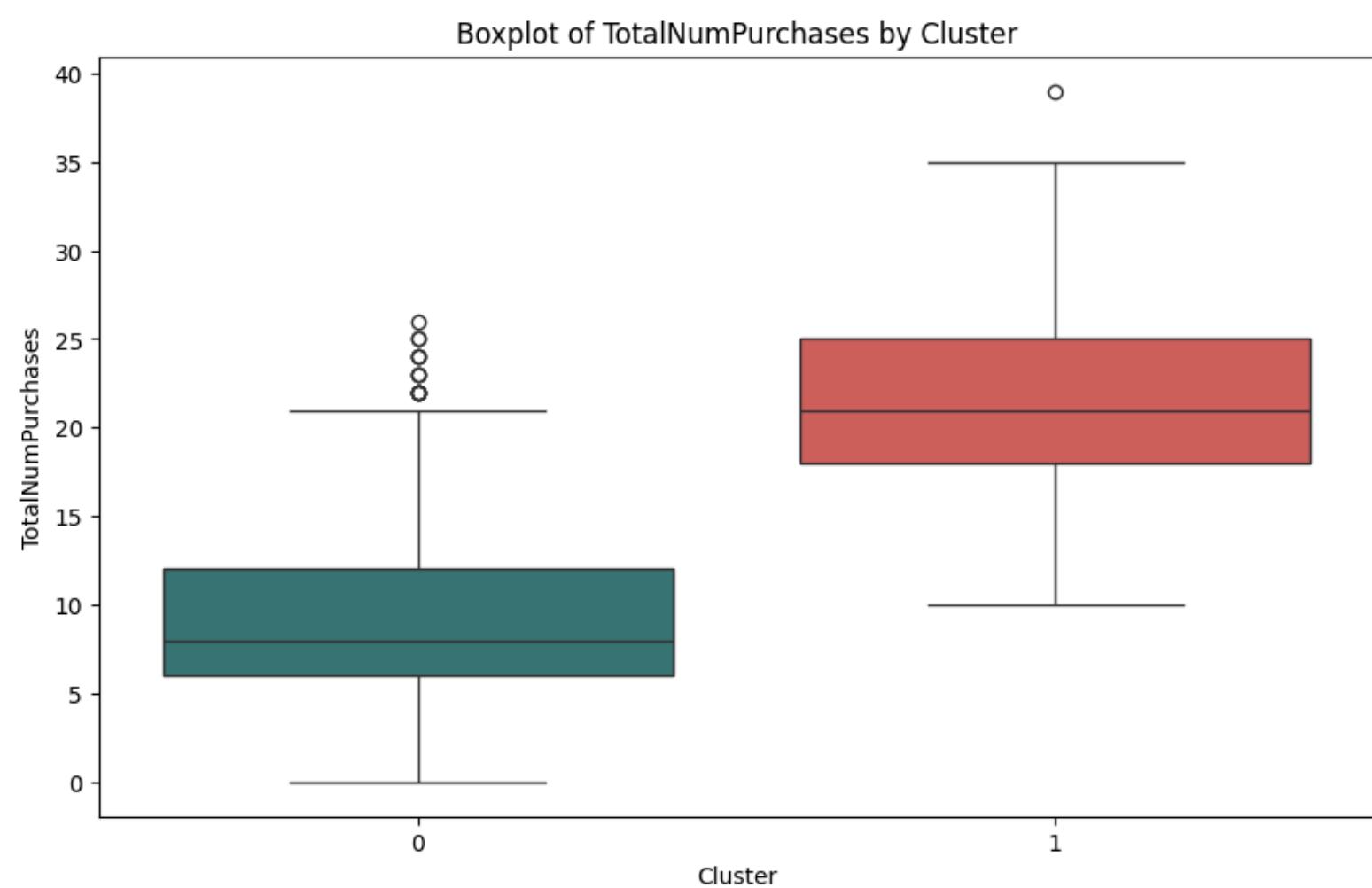
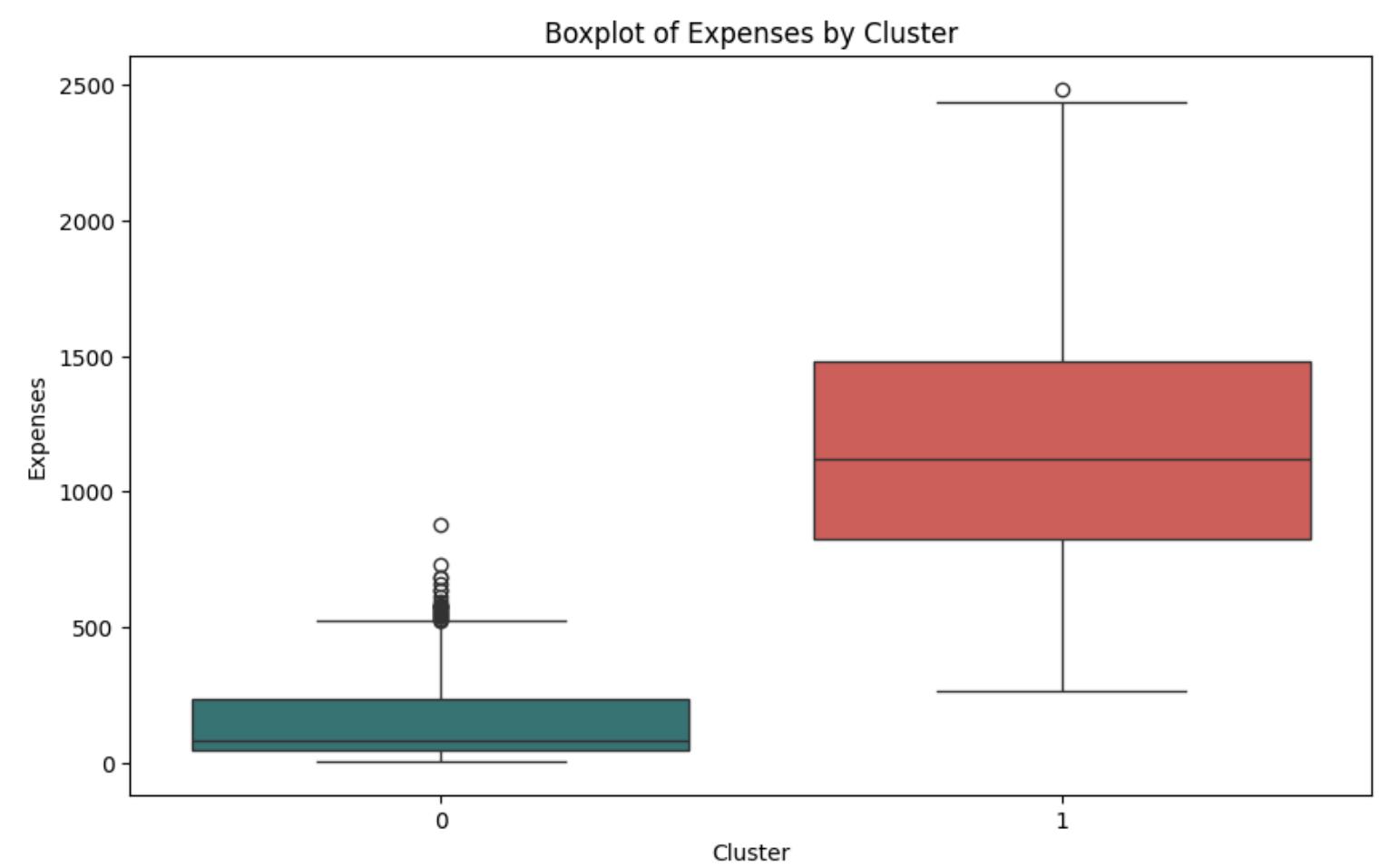


Boxplot of Income by Cluster

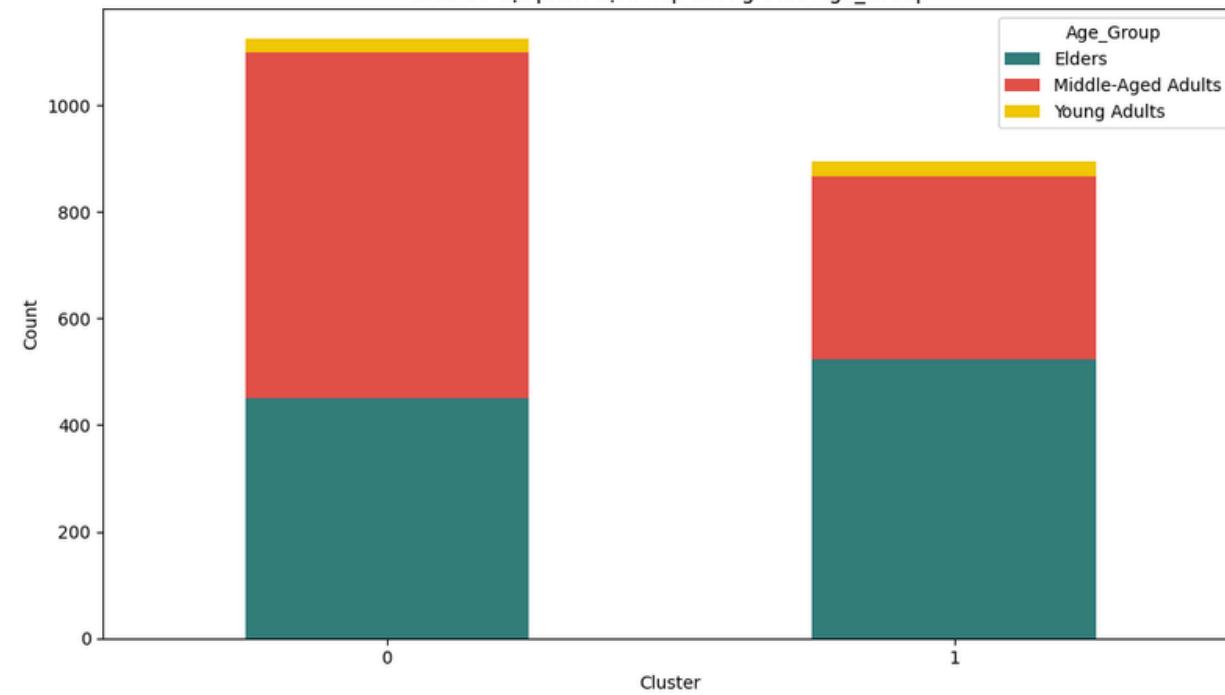


Boxplot of Kids by Cluster

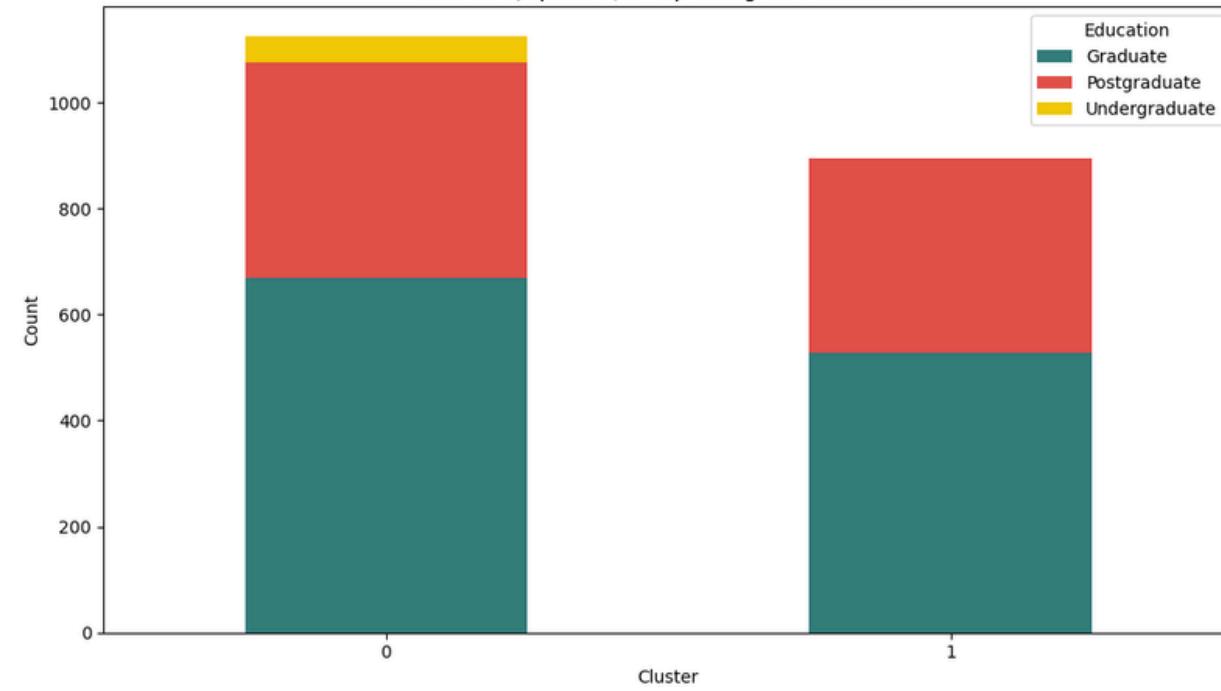




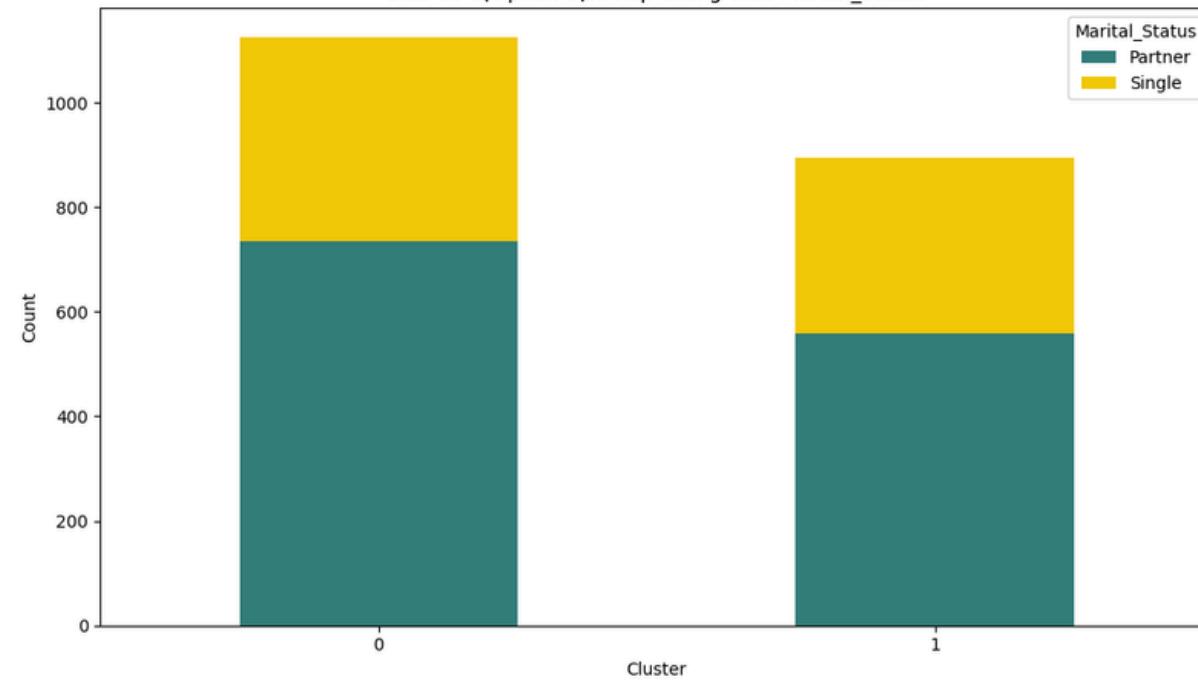
Biểu đồ cột phân cụm xếp chồng theo Age_Group



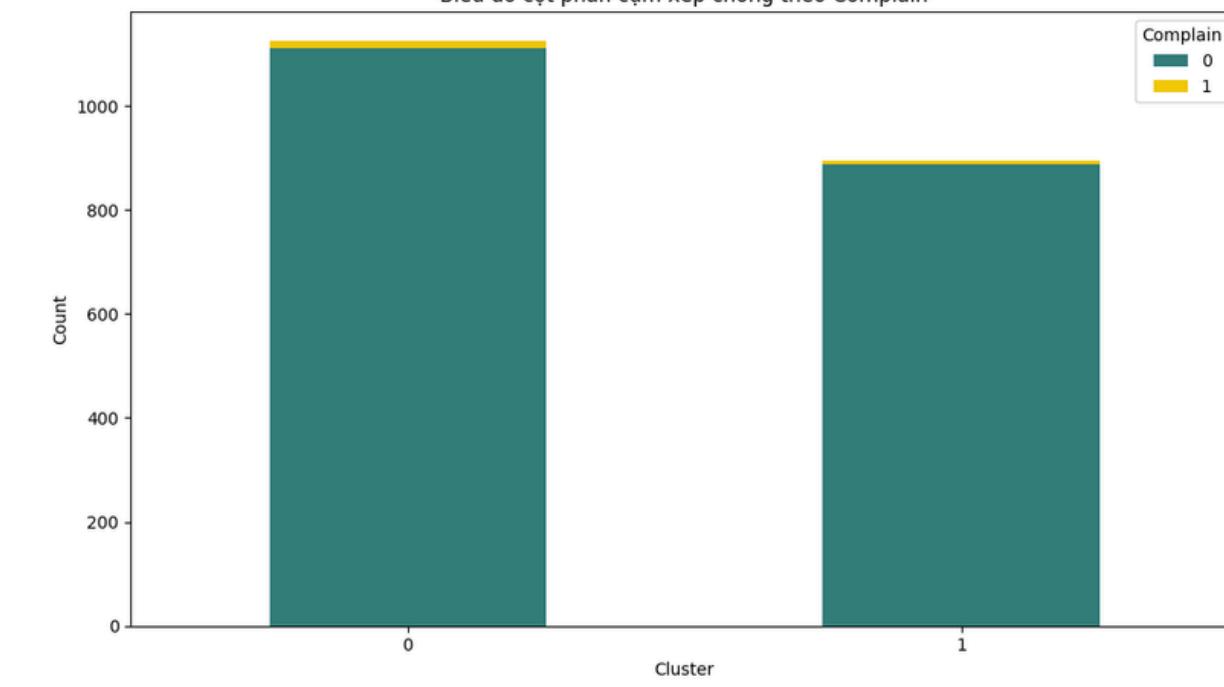
Biểu đồ cột phân cụm xếp chồng theo Education



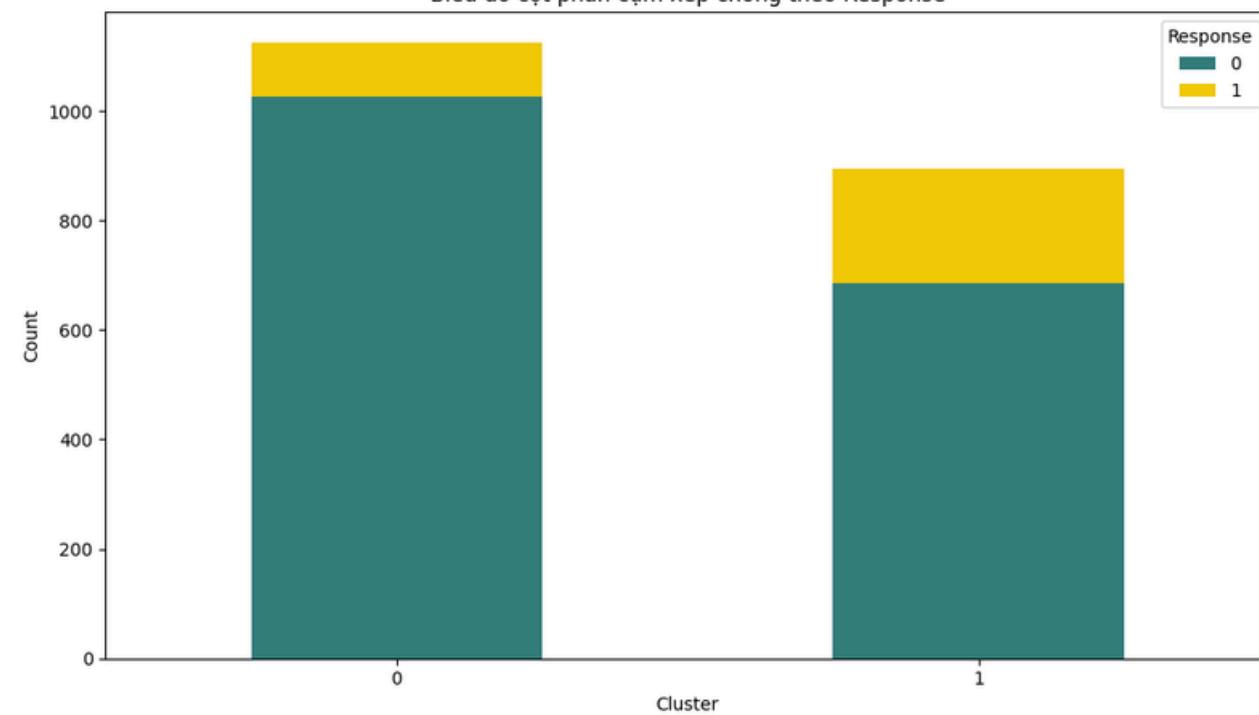
Biểu đồ cột phân cụm xếp chồng theo Marital_Status



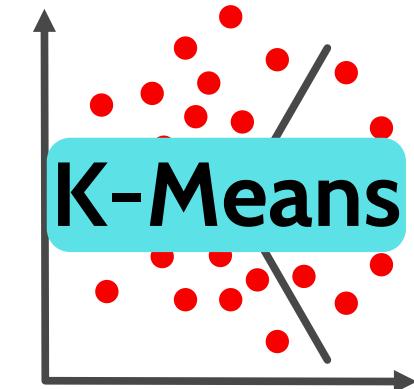
Biểu đồ cột phân cụm xếp chồng theo Complain



Biểu đồ cột phân cụm xếp chồng theo Response



7. Hiệu suất mô hình



💰 Nhóm 0:

- Nhóm thu nhập trung bình và thấp
- Thường có vợ/chồng
- Có 2 con trở lên
- Tốt nghiệp hoặc Sau đại học hoặc chưa học đại học
- Chi phí thấp và số lần mua hàng ít
- Không chấp nhận khuyến mại

\$ 💰 Nhóm 1:

- Nhóm thu nhập cao
- Không có con hoặc 1 con
- Thường có bạn đời
- Tốt nghiệp hoặc sau đại học
- Số lần mua hàng cao
- Chi phí cao

7. HIỆU SUẤT MÔ HÌNH

Logistic Regression

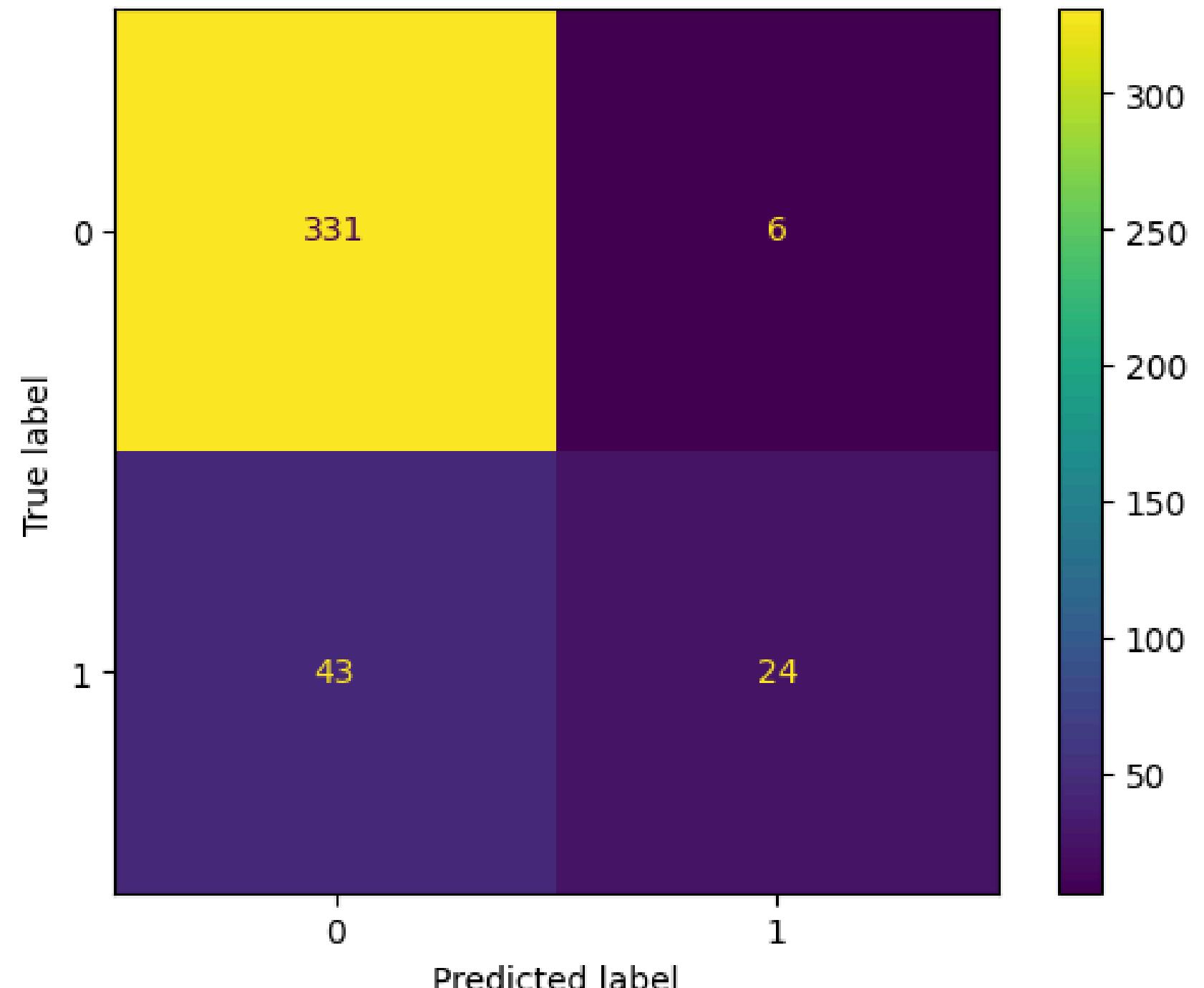
Accuracy: 0.8787

Precision: 0.8

Cross-Validation Accuracy: 0.8798

Mô hình đơn giản và dễ hiểu, phù hợp với các bài toán phân loại cơ bản.

Hiệu suất rất tốt với độ chính xác cao và precision tốt, chứng tỏ mô hình đơn giản này hoạt động rất hiệu quả trên dữ liệu



7. HIỆU SUẤT MÔ HÌNH

Random Forest

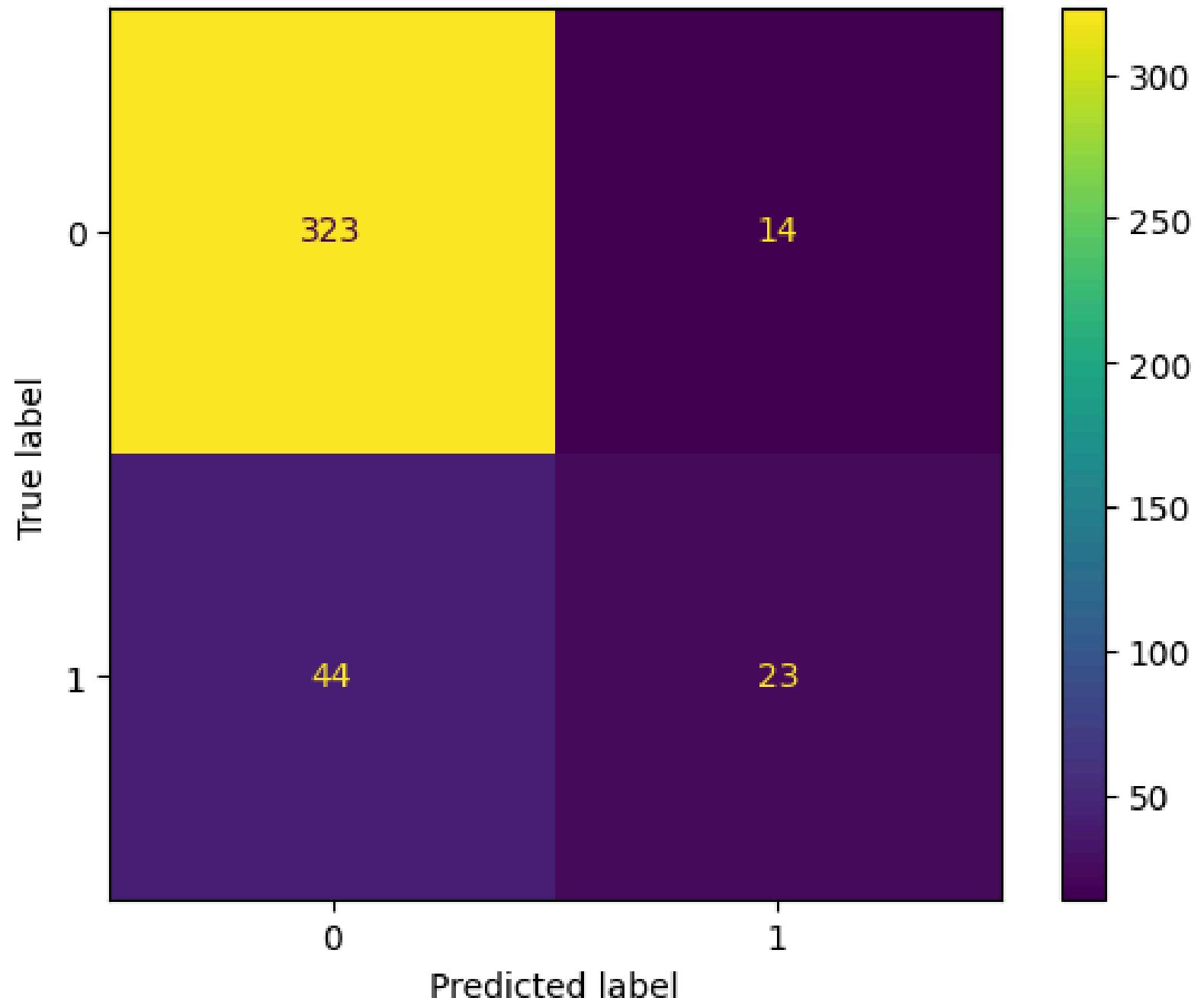
Accuracy: 0.8564

Precision: 0.6216

Cross-Validation Accuracy: 0.8650

Độ chính xác tổng thể khá tốt nhưng độ chính xác của dự đoán positive thấp hơn Logistic Regression

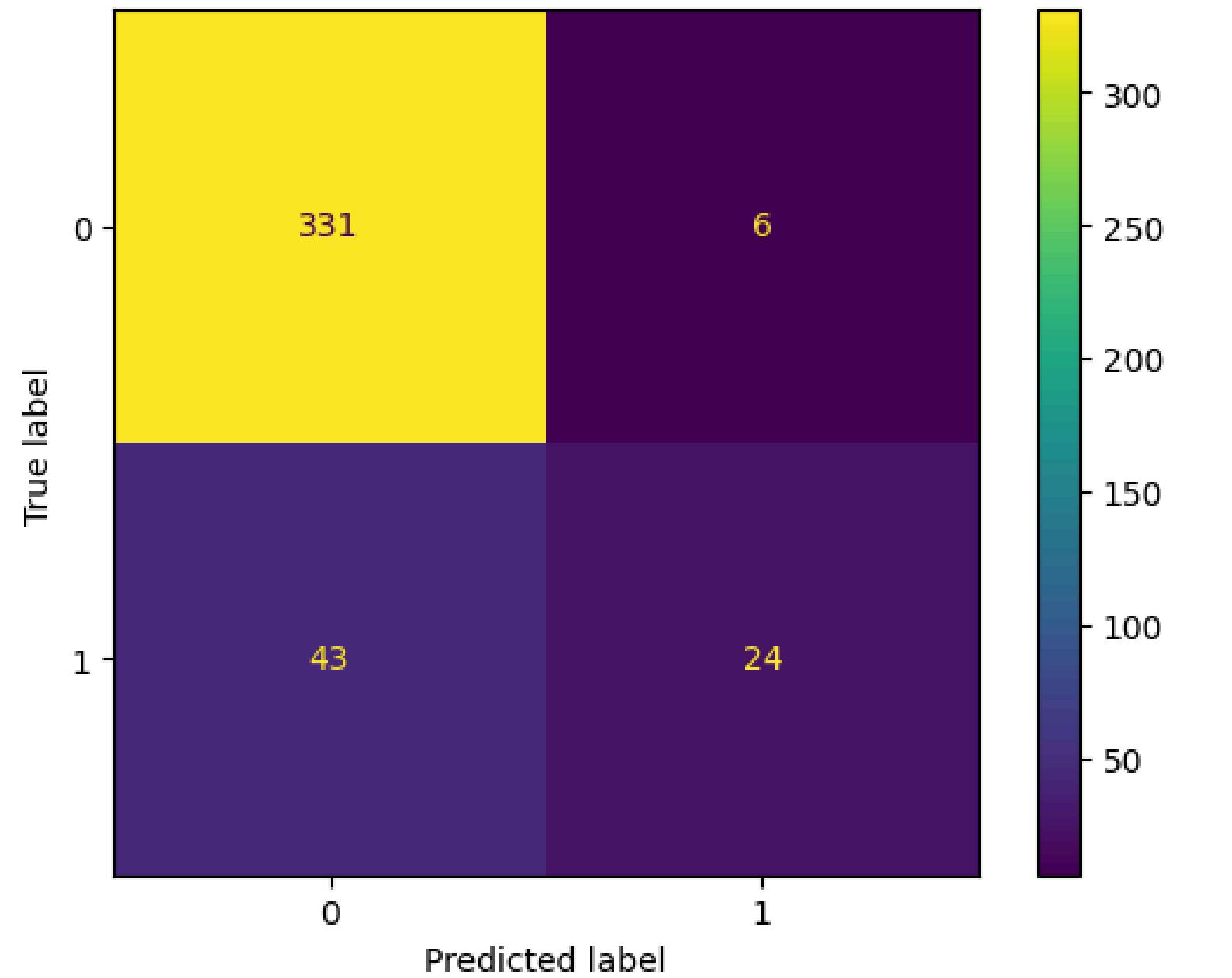
Mô hình này tạo ra nhiều false positives. Cần cải thiện để giảm false positives và tăng precision.



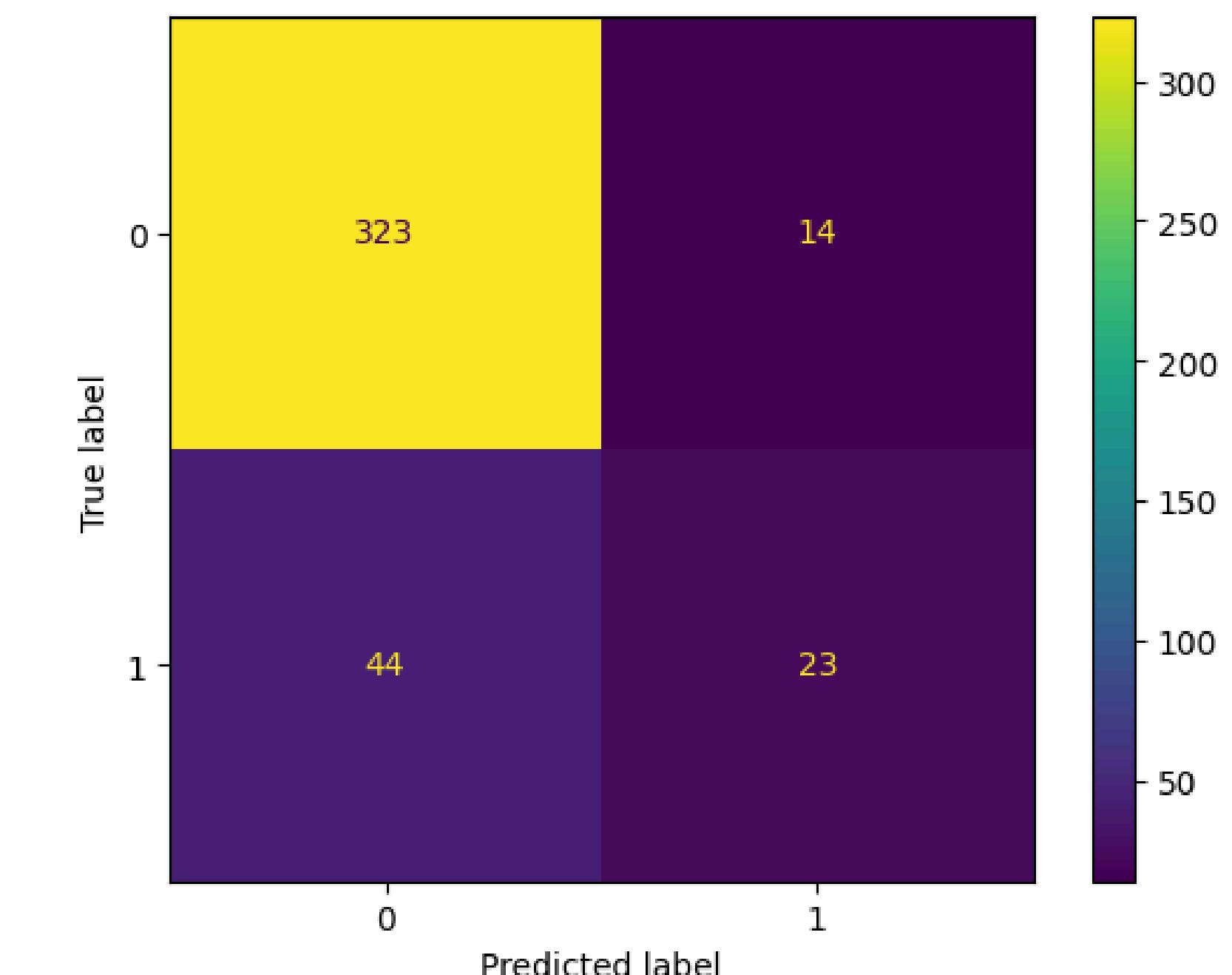
7. HIỆU SUẤT MÔ HÌNH

So sánh

Logistic Regression



Random Forest



7. HIỆU SUẤT MÔ HÌNH

Accuracy: 0.8762

Precision: 0.7741

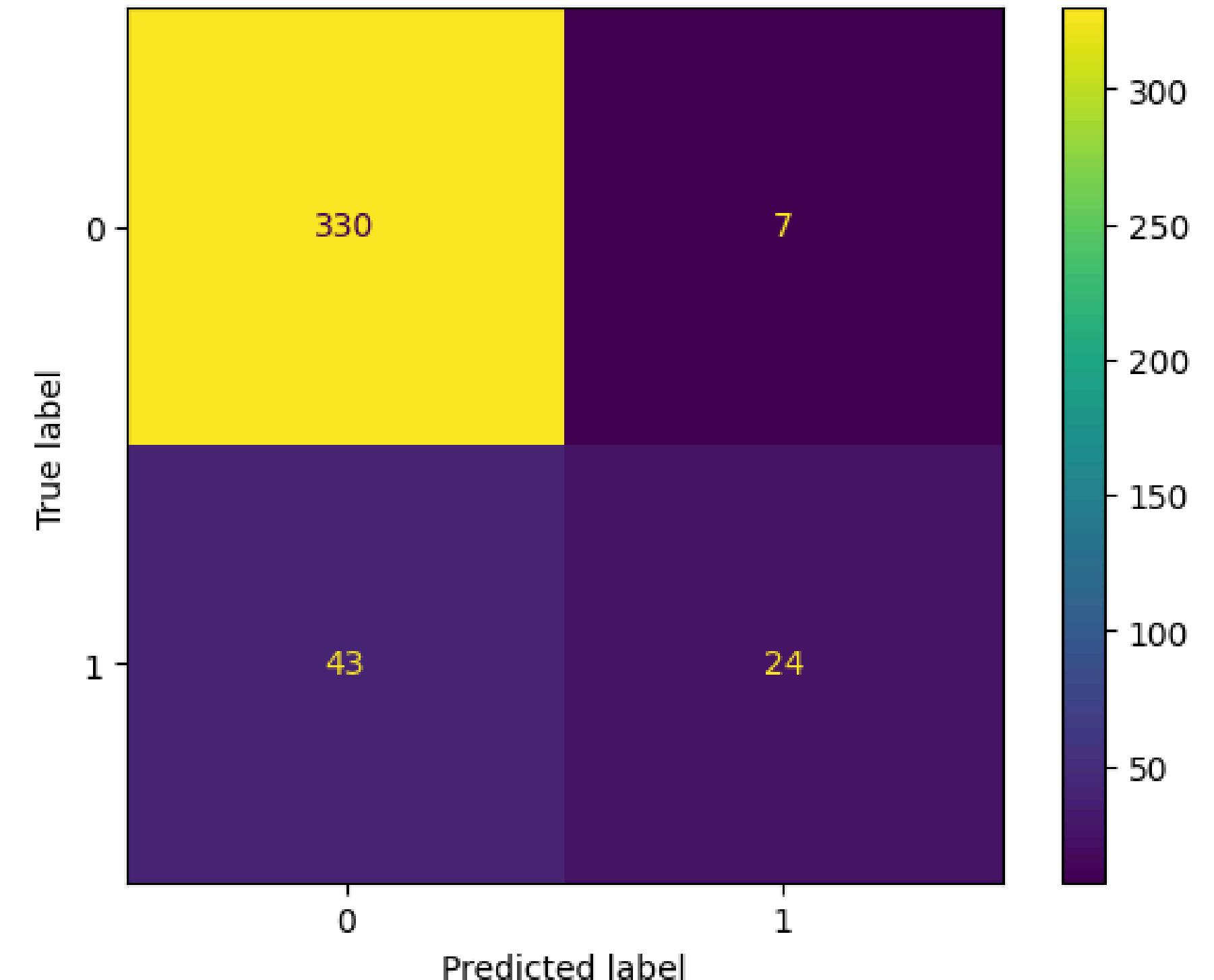
Cross-Validation Accuracy: 0.8792

Sử dụng K-Means giúp Logistic Regression có thêm thông tin từ cấu trúc dữ liệu, cải thiện độ chính xác

Phụ thuộc vào chất lượng phân cụm, việc thêm feature Cluster từ K-Means không mang lại nhiều cải thiện rõ rệt

Quá trình huấn luyện phức tạp hơn, phải chuẩn hóa dữ liệu, thêm Cluster vào Feature,...

**Mô hình kết hợp
(Phân loại - phân cụm)**



8. CẢI THIỆN MÔ HÌNH

8. CẢI THIỆN MÔ HÌNH

Lựa chọn đặc trưng với SelectKBest với f_classif (ANOVA F-test) được sử dụng để đánh giá và lựa chọn các đặc trưng dựa trên ý nghĩa thống kê của chúng.

	Feature	Score
6	Expenses	5203.42
7	TotalNumPurchases	3414.82
2	Income	3349.78
3	Kids	504.28
8	TotalAcceptedCmp	299.08
11	Response	84.87
0	Age_Group	50.79
4	Dt_Customer	33.54
1	Education	14.36
10	Marital_Status_Single	2.01
9	Complain	0.71
5	Recency	0.01

8. CẢI THIỆN MÔ HÌNH

Sinh dữ liệu với CGAN

1. Discriminator (D Loss) ~1.2:

- Discriminator hoạt động ổn định, duy trì khả năng phân biệt dữ liệu thật và giả một cách hợp lý.
- Mức D Loss quanh 1.2 là dấu hiệu cho thấy Discriminator đủ mạnh nhưng không quá áp đảo, đảm bảo sự cân bằng với Generator.

2. Generator (G Loss) ~1.1:

- Generator đã học được cách sinh dữ liệu giả với chất lượng cao và khó phân biệt với dữ liệu thật.
- Mức G Loss > 1.0 cho thấy Generator đang hoạt động tốt, đánh lừa Discriminator hiệu quả.

3. Tổng quan mô hình

- Trạng thái cân bằng:
 - Cả Discriminator và Generator đều đạt sự cân bằng tốt, không thành phần nào áp đảo thành phần còn lại, giúp mô hình học hiệu quả.
- Chất lượng dữ liệu giả sinh ra:
 - Với D Loss và G Loss như hiện tại, dữ liệu giả được kỳ vọng có độ chân thực cao, phân phối sát với dữ liệu thật.
- Hội tụ:
 - Các giá trị D Loss và G Loss ổn định qua các epoch, cho thấy mô hình CGAN đã hội tụ.

```

Epoch 1500/3200, D Loss: 1.3751, G Loss: 0.7003
Epoch 1550/3200, D Loss: 1.3745, G Loss: 0.7228
Epoch 1600/3200, D Loss: 1.3912, G Loss: 0.6852
Epoch 1650/3200, D Loss: 1.3669, G Loss: 0.7285
Epoch 1700/3200, D Loss: 1.3862, G Loss: 0.7242
Epoch 1750/3200, D Loss: 1.3571, G Loss: 0.7111
Epoch 1800/3200, D Loss: 1.3245, G Loss: 0.7488
Epoch 1850/3200, D Loss: 1.3781, G Loss: 0.7284
Epoch 1900/3200, D Loss: 1.3525, G Loss: 0.7123
Epoch 1950/3200, D Loss: 1.3894, G Loss: 0.7337
Epoch 2000/3200, D Loss: 1.3387, G Loss: 0.7666
Epoch 2050/3200, D Loss: 1.3680, G Loss: 0.6444
Epoch 2100/3200, D Loss: 1.3478, G Loss: 0.8267
Epoch 2150/3200, D Loss: 1.3977, G Loss: 0.6593
Epoch 2200/3200, D Loss: 1.2948, G Loss: 0.8193
Epoch 2250/3200, D Loss: 1.2444, G Loss: 0.8219
Epoch 2300/3200, D Loss: 1.2261, G Loss: 0.8465
Epoch 2350/3200, D Loss: 1.2679, G Loss: 0.8529
Epoch 2400/3200, D Loss: 1.2273, G Loss: 0.9068
Epoch 2450/3200, D Loss: 1.1803, G Loss: 0.8875
Epoch 2500/3200, D Loss: 1.2518, G Loss: 0.9373
Epoch 2550/3200, D Loss: 1.1966, G Loss: 0.9099
Epoch 2600/3200, D Loss: 1.1390, G Loss: 0.9753
Epoch 2650/3200, D Loss: 1.1484, G Loss: 0.9468
Epoch 2700/3200, D Loss: 1.2436, G Loss: 1.0424
Epoch 2750/3200, D Loss: 1.0955, G Loss: 1.0459
Epoch 2800/3200, D Loss: 1.1353, G Loss: 1.0517
Epoch 2850/3200, D Loss: 1.1477, G Loss: 1.0201
Epoch 2900/3200, D Loss: 1.1632, G Loss: 1.1070
Epoch 2950/3200, D Loss: 1.1706, G Loss: 1.0180
Epoch 3000/3200, D Loss: 1.1703, G Loss: 1.1252
Epoch 3050/3200, D Loss: 1.1258, G Loss: 1.0970
Epoch 3100/3200, D Loss: 1.1498, G Loss: 1.1079
Epoch 3150/3200, D Loss: 1.0848, G Loss: 1.1235
Epoch 3200/3200, D Loss: 1.2213, G Loss: 1.1059

```

Kết luận: Mô hình CGAN hoạt động hiệu quả, đạt trạng thái cân bằng giữa Discriminator và Generator.

9. ĐÁNH GIÁ TỔNG QUAN

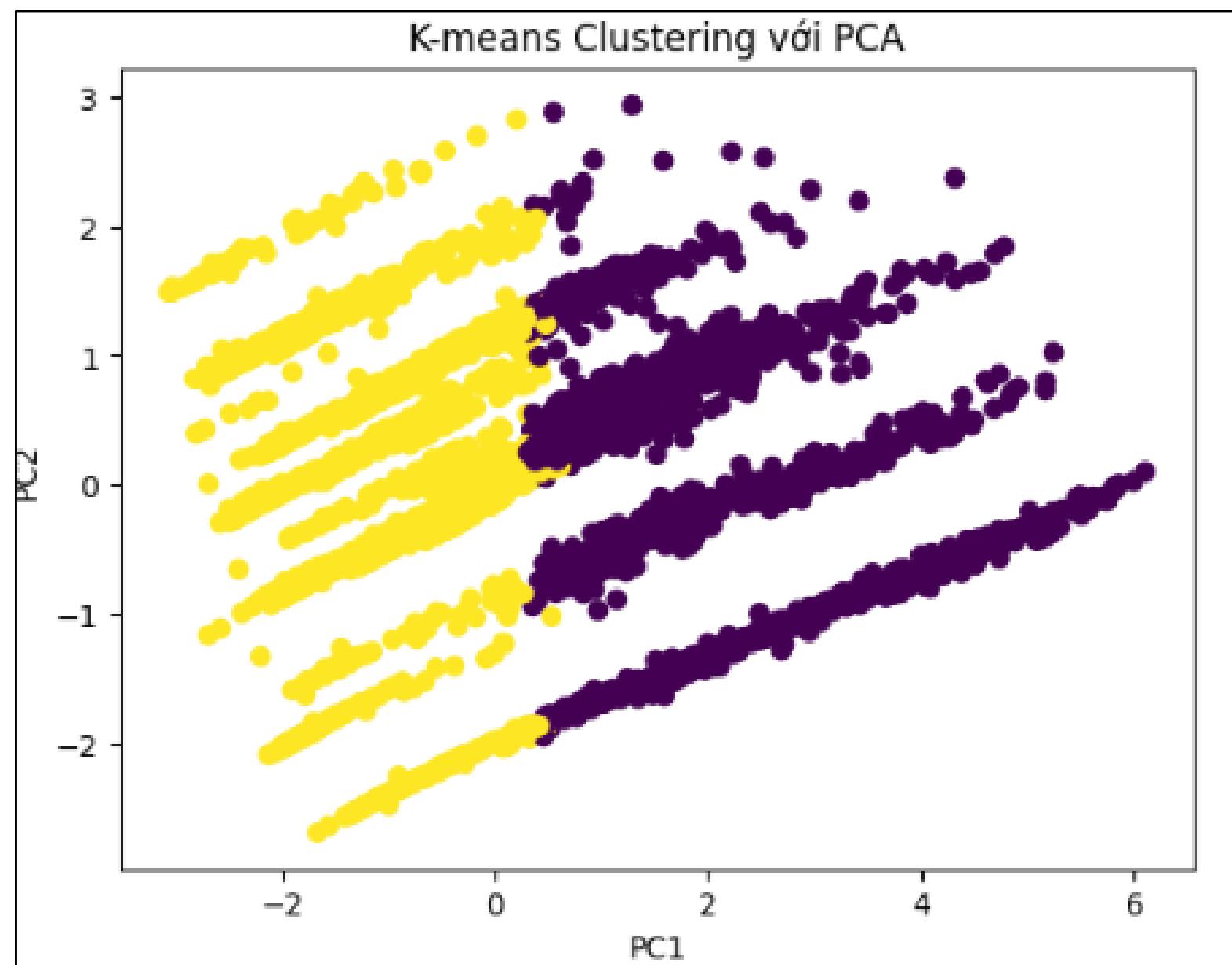
9. ĐÁNH GIÁ TỔNG QUAN



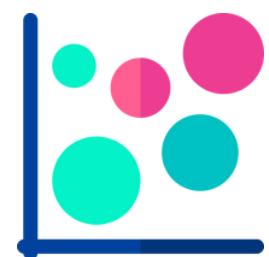
K-Means

Việc cải thiện điểm Silhouette Score từ 0.1988 lên 0.3485 cho thấy rõ ràng rằng:

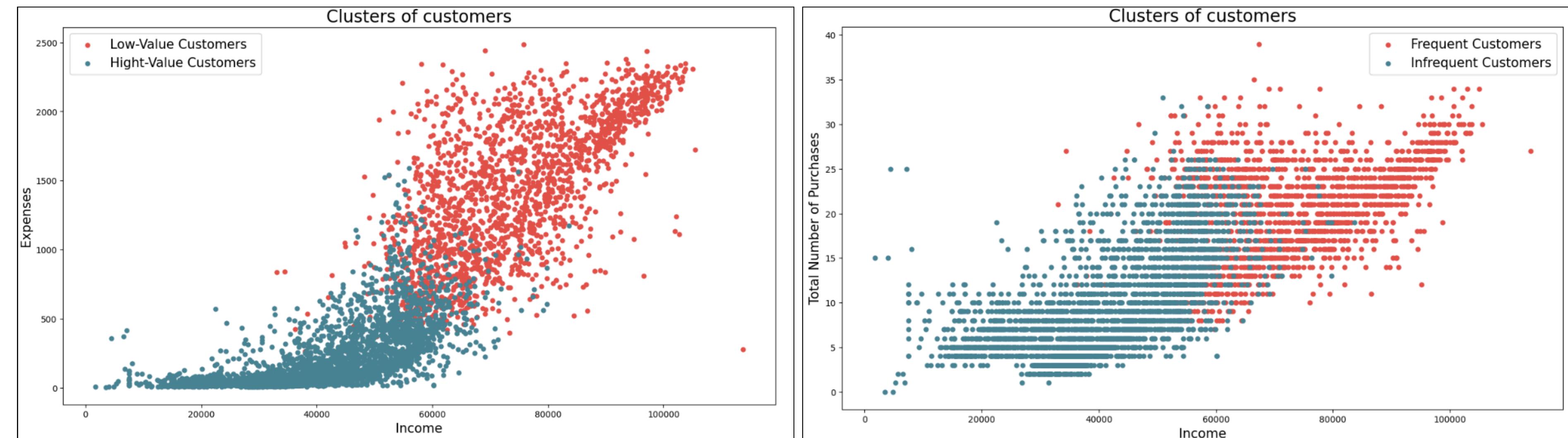
- Chọn đặc trưng phù hợp đã giúp tập trung vào các yếu tố quan trọng, loại bỏ nhiễu, và tăng cường khả năng tách biệt giữa các cụm.
- Sinh dữ liệu mới bằng CGAN đã bổ sung sự đa dạng vào tập dữ liệu, cung cấp thêm thông tin hữu ích, giúp thuật toán K-means xác định ranh giới giữa các cụm tốt hơn.



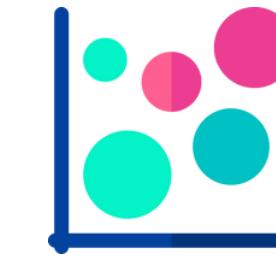
9. ĐÁNH GIÁ TỔNG QUAN



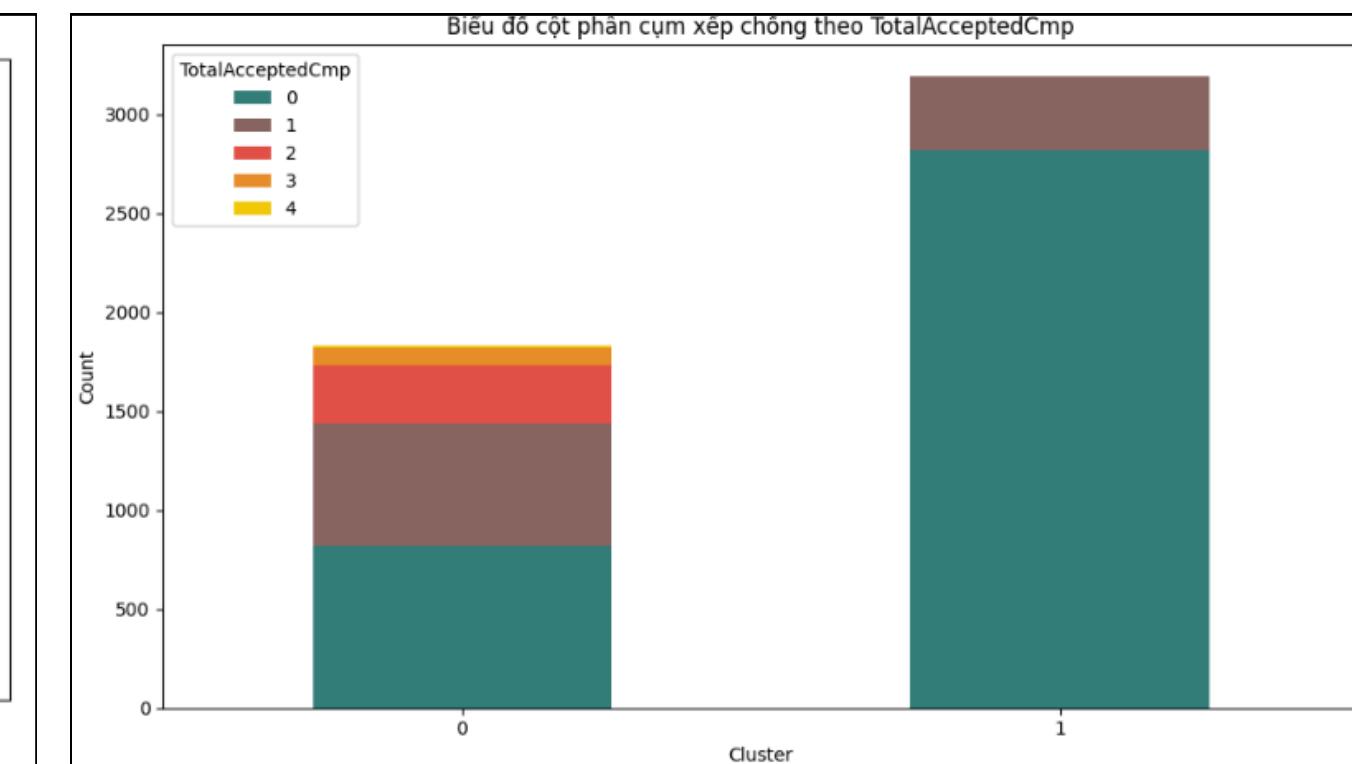
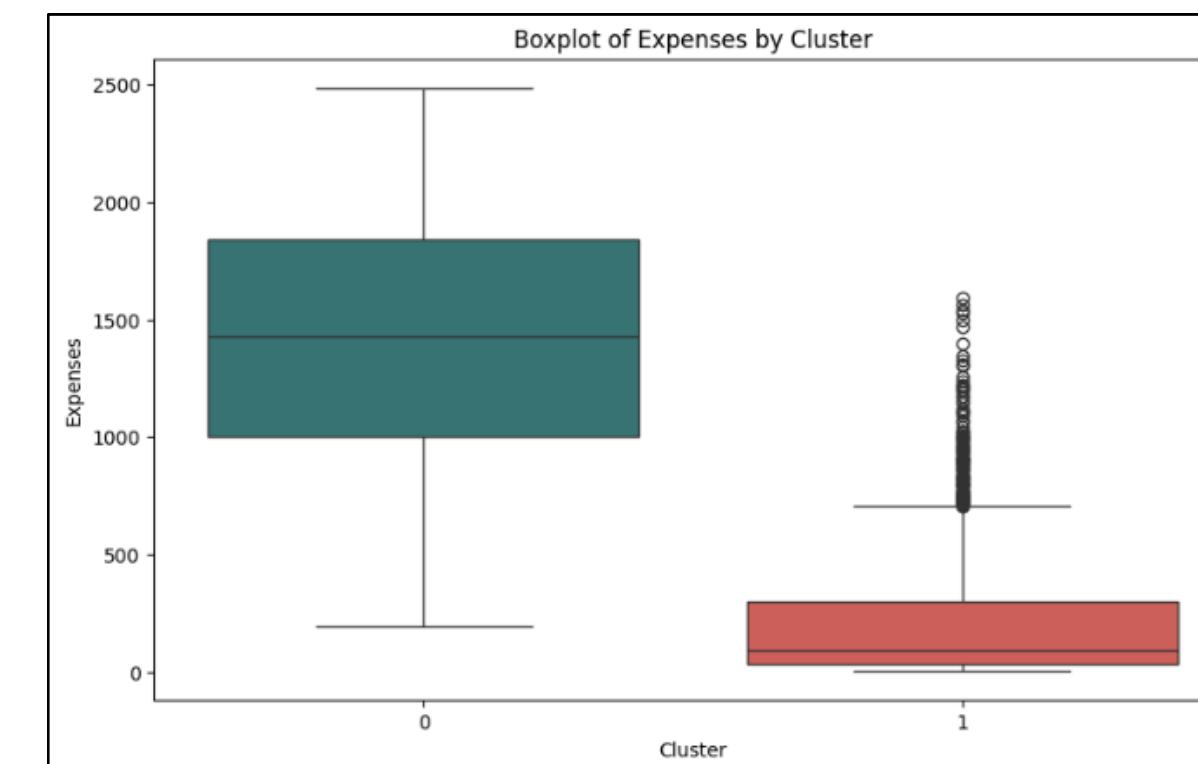
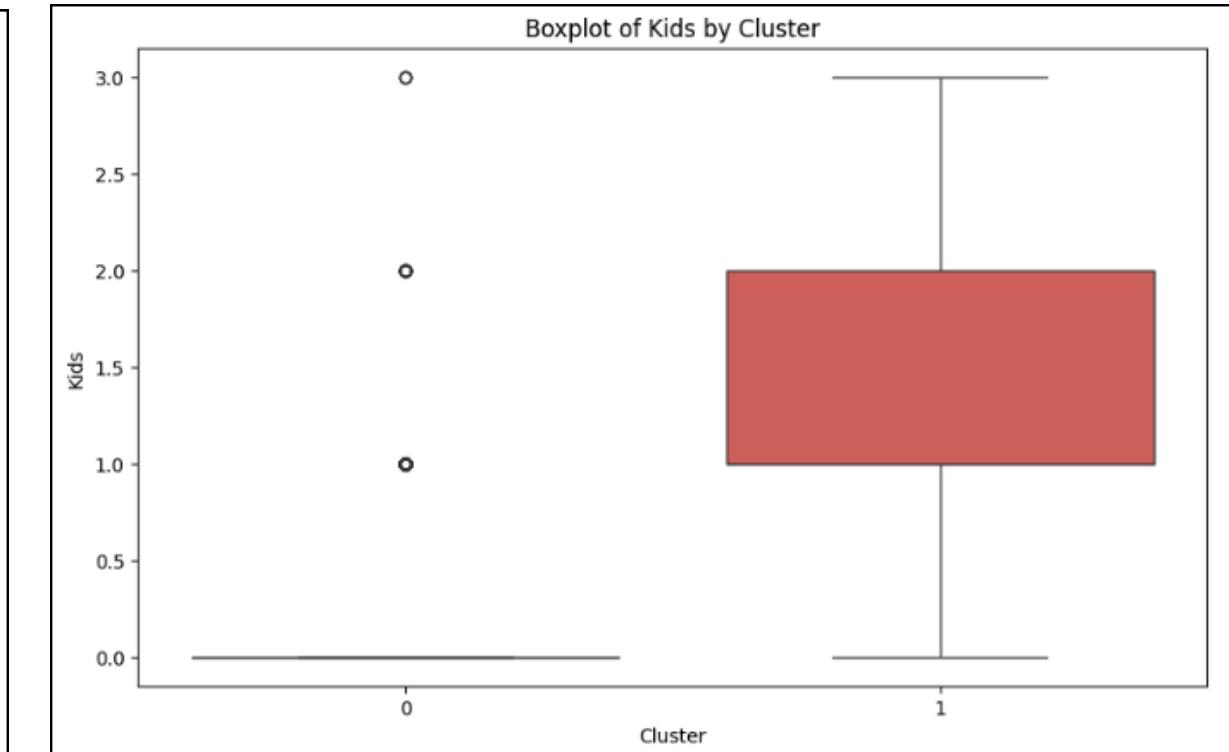
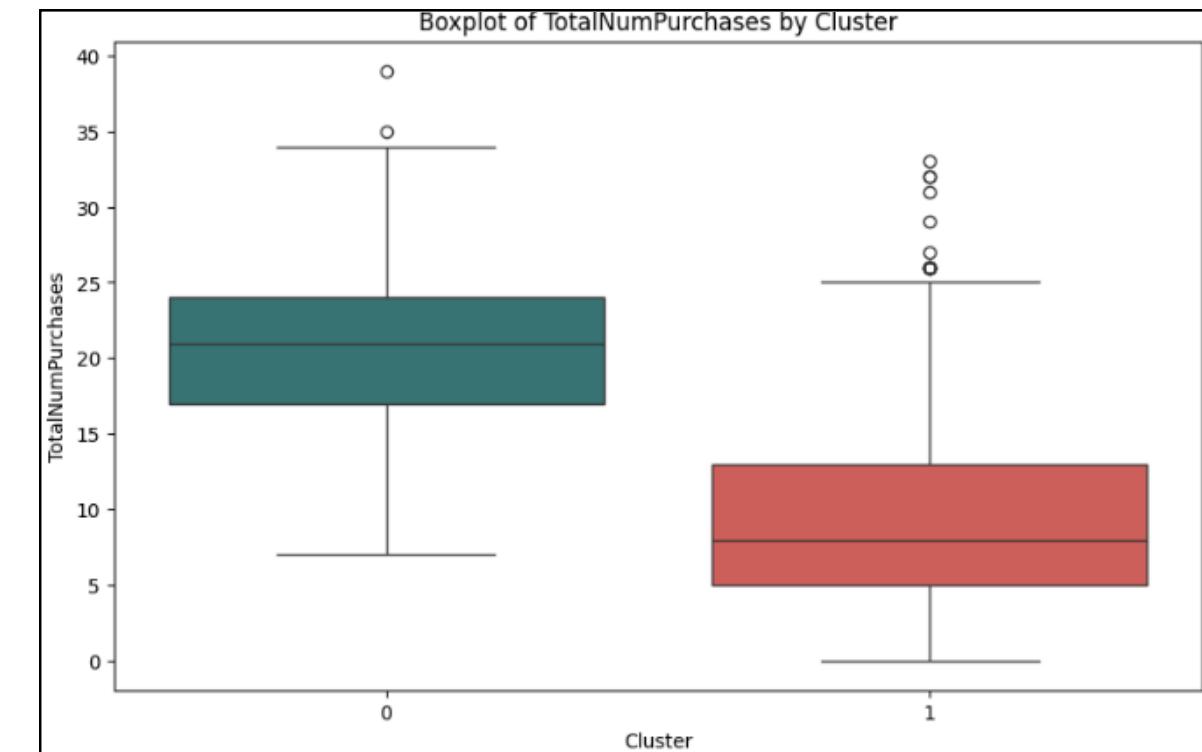
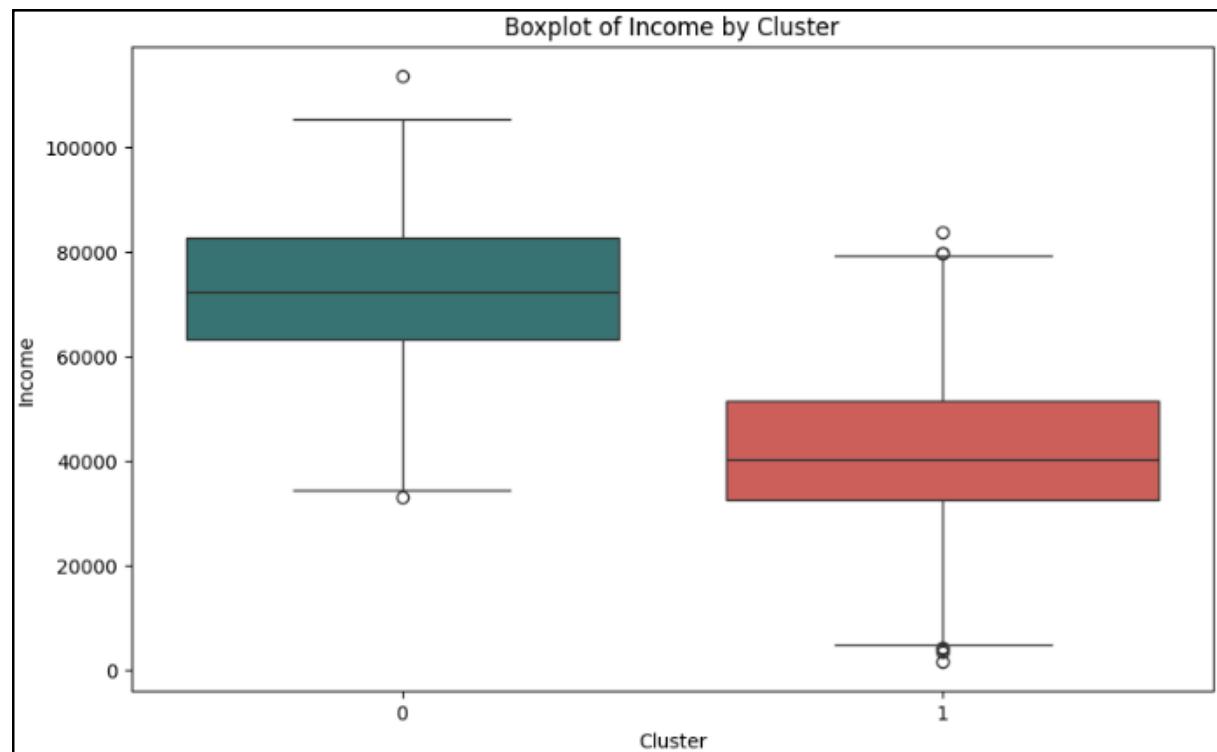
K-Means



9. ĐÁNH GIÁ TỔNG QUAN



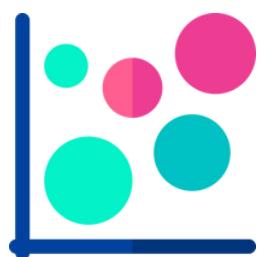
K-Means



9. ĐÁNH GIÁ TỔNG QUAN

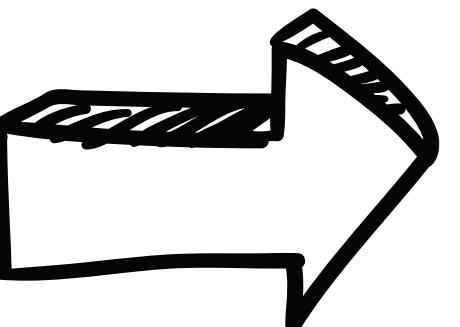
Nhóm 0:

- Nhóm thu nhập cao
- Thường là không có con
- Số lần mua hàng cao
- Chi phí cao
- Chấp nhận khuyến mại



K-Means

K-Means để phân cụm khách hàng thành hai nhóm riêng biệt dựa trên hồ sơ và hành vi giao dịch của họ.



Nhóm 1:

- Nhóm thu nhập trung bình và thấp
- Có nhiều hơn 1 con
- Chi phí thấp và số lần mua hàng ít
- Không chấp nhận khuyến mại

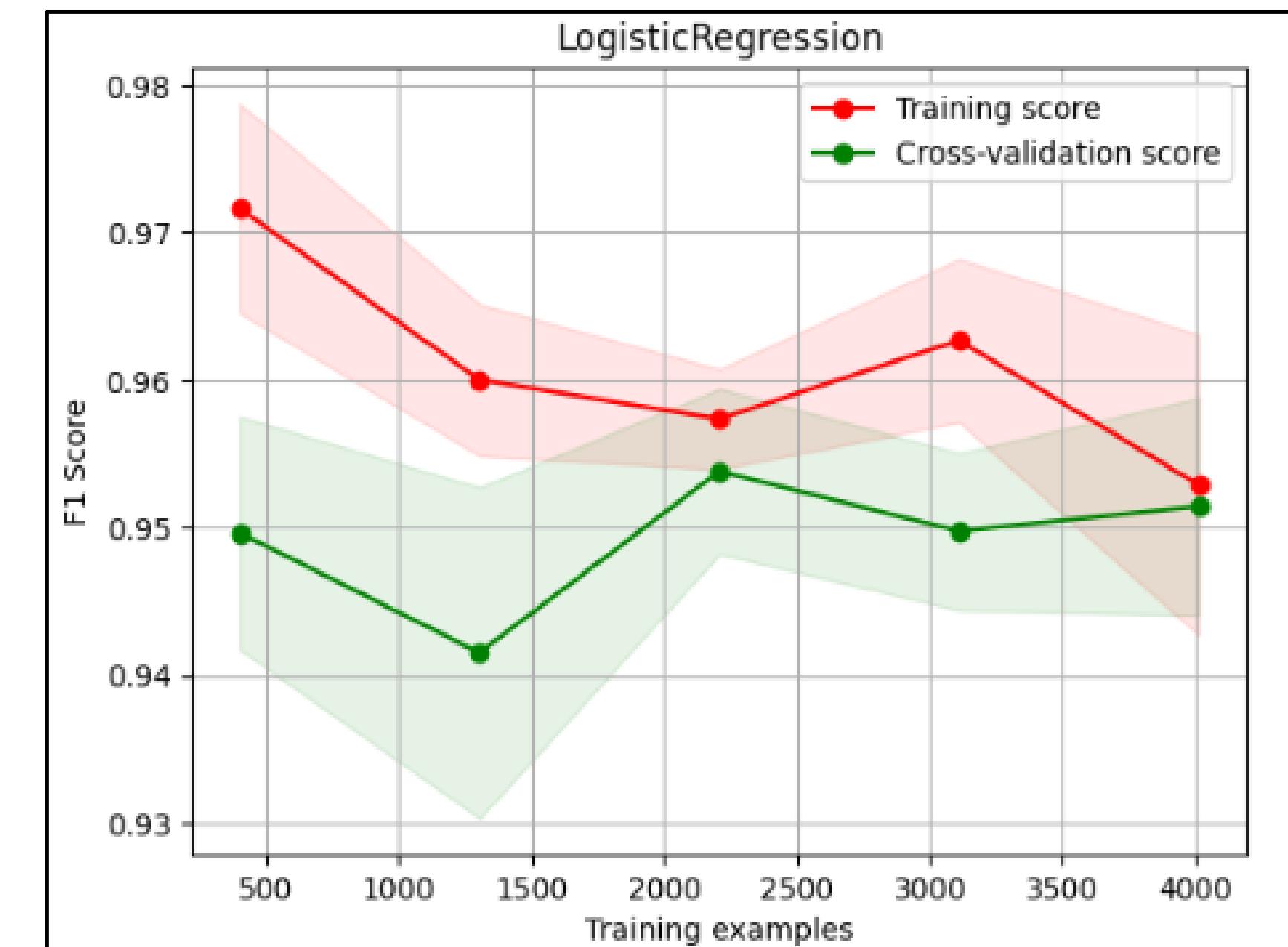
=> Điều này hỗ trợ việc xác định các nhóm khách hàng có những đặc điểm chung, giúp doanh nghiệp có thể đưa ra các chiến lược tiếp thị cụ thể cho từng nhóm.

9. ĐÁNH GIÁ TỔNG QUAN

KẾT QUẢ ĐẠT ĐƯỢC VỚI MÔ HÌNH PHÂN CỤM KẾT HỢP PHÂN LOẠI(Logistic Regression)

So sánh với kết quả khi chưa cải thiện dữ liệu

- Độ chính xác tăng từ 88% lên 99%
- Precision cải thiện từ 78% lên 99%
- Cross-validation score ổn định hơn với độ lệch chuẩn thấp hơn



10. KẾT LUẬN

10. KẾT LUẬN

TỔNG KẾT NGHIÊN CỨU

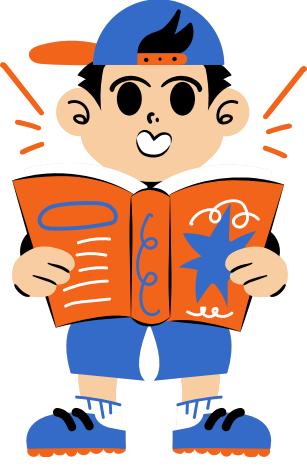
Kết hợp K-means clustering với Logistic Regression giúp cải thiện độ chính xác trong việc phân loại khách hàng tiềm năng bằng cách nhóm khách hàng có hành vi tương tự, từ đó tối ưu hóa chiến lược marketing. Phân cụm khách hàng theo yếu tố như độ tuổi, thu nhập và chi tiêu giúp phân chia khách hàng thành các nhóm hành vi riêng biệt, từ đó doanh nghiệp có thể thiết kế chiến lược marketing phù hợp. Kết quả này giúp xác định chính xác nhóm khách hàng tiềm năng, tối ưu chi phí quảng cáo và gia tăng doanh thu.

HƯỚNG PHÁT TRIỂN

- Thử nghiệm các kỹ thuật sinh dữ liệu khác
- Phát triển mô hình ensemble
- Tích hợp thêm dữ liệu từ nhiều nguồn
- Xây dựng hệ thống dự đoán real-time
- Tối ưu hóa thêm siêu tham số các mô hình



Thành viên nhóm



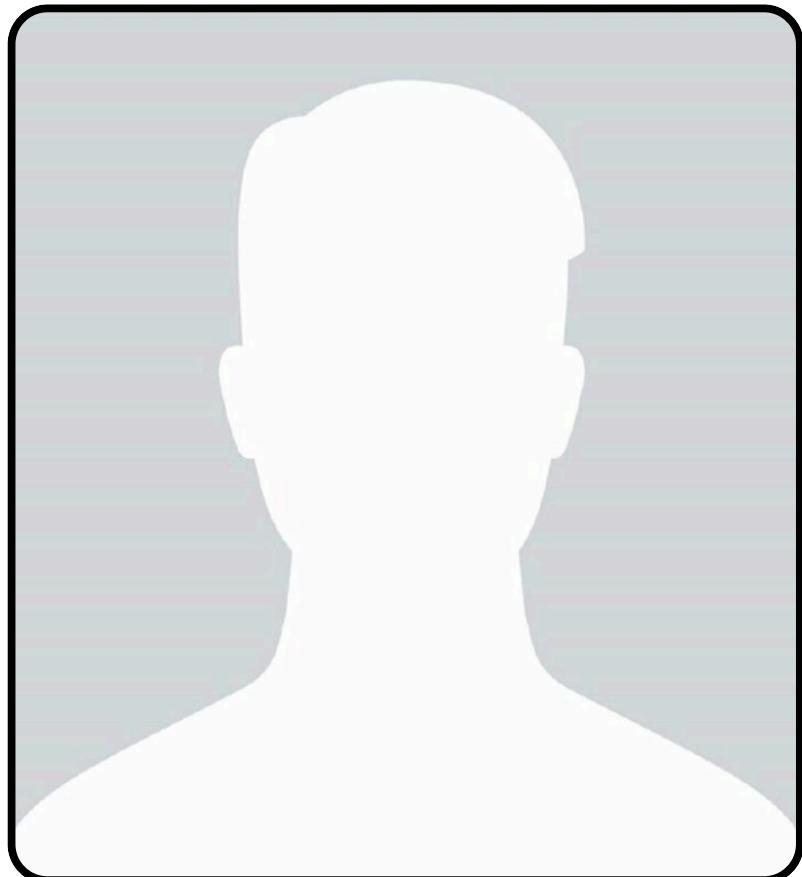
Phạm Gia Khánh



Nguyễn Minh Phúc



Trần Minh Tú



Nguyễn Đức Chiến



Xin cảm ơn!

