

TRƯỜNG ĐẠI HỌC MỞ
THÀNH PHỐ HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN



THÀNH VIÊN THỰC HIỆN
2251050068 - Lê Nguyễn Phước Thịnh
2251052116 - Võ Đức Thiện
2251052122 - Nguyễn Thanh Toàn

MÔN HỌC: KHAI PHÁ DỮ LIỆU
ĐỀ TÀI: Phân tích hài lòng khách hàng hàng không
qua lăng kính dữ liệu lớn và trí tuệ nhân tạo

TP. Hồ Chí Minh, Năm 2025

MỤC LỤC

Mục lục	i
1 Giới thiệu bài toán và mô tả bộ dữ liệu	1
1.1 Giới thiệu và mục tiêu bài toán	1
1.2 Mô tả bộ dữ liệu	2
2 Tiền xử lý dữ liệu (Data Preprocessing)	7
2.1 Các kỹ thuật tiền xử lý dữ liệu	7
2.1.1 Xử lý Giá trị Thiếu (Handling Missing Values)	7
2.1.2 Chuẩn hóa Dữ liệu (Normalization/Standardization) . .	9
2.1.3 Biến đổi Dữ liệu (Data Transformation) - Log Transform	11
2.1.4 Xử lý Ngoại lai (Outlier Treatment)	12
2.1.5 Lựa chọn Đặc trưng (Feature Selection) & Giảm chiều dữ liệu (PCA)	15
2.2 Thống kê mô tả và trực quan hóa các dịch vụ	19
2.2.1 Ma trận tương quan các dịch vụ hàng không	19
2.2.2 Phân tích mức độ hài lòng theo các đặc trưng	20
2.2.3 Chênh lệch đánh giá dịch vụ và các chỉ số khác	23
2.2.4 Dashboard tổng quan dữ liệu khách hàng	27
2.3 Lọc dữ liệu và phân tích so sánh nhóm khách hàng	28

2.3.1	Ma trận tương quan giữa các Service Score Groups	28
2.3.2	So sánh tổng quan 3 nhóm khách hàng chính	30
2.3.3	Phân tích chi tiết các Customer Tier	32
3	Áp dụng các mô hình khai phá dữ liệu	34
3.1	CÁC KỸ THUẬT KHAI THÁC MẪU PHỎ BIẾN	34
3.1.1	Khai thác tập phỏ biến với APRIORI	34
3.1.2	Sinh và phân tích luật kết hợp (Association Rules):	41
3.1.3	Trực quan hóa và phân tích tổng quan các luật:	43
3.2	CÁC KỸ THUẬT PHÂN LOẠI DỮ LIỆU	48
3.2.1	Cây quyết định (Decision Tree)	48
3.2.2	Thuật toán Naive Bayes	49
3.2.3	Thuật toán Random Forest (Ensemble Method)	50
3.2.4	Thuật toán Logistic Regression	50
3.3	CÁC KỸ THUẬT GOM CỤM DỮ LIỆU	60
3.3.1	Thuật toán K-Means	60
3.3.2	Thuật toán Hierarchical Clustering (Agglomerative) . .	61
3.3.3	Xác định số cụm tối ưu (k)	62
3.3.4	Tìm số cụm tối ưu	63
3.3.5	Thực hiện các thuật toán Clustering	65
3.3.6	So sánh và đánh giá các thuật toán	69
4	Kết quả và Phân tích	73
4.1	Phân tích Gom Cụm Khách hàng (Clustering Analysis)	73
4.2	Phân tích Luật Kết hợp (Association Rules Analysis)	77
4.2.1	Thống kê tổng quan khai thác luật kết hợp	77
4.2.2	Phân tích các đặc điểm phổ biến nhất	78
4.2.3	Phân tích insight từ các items phổ biến	79

4.2.4	Top luật kết hợp quan trọng nhất	81
4.2.5	Phân tích chi tiết các luật nổi bật	82
4.2.6	Nghịch lý trong dữ liệu khách hàng	83
4.2.7	Tổng kết và khuyến nghị từ Association Rules	83
4.3	Phân tích Mô hình Phân loại (Classification Model Analysis) .	87
4.4	Kết luận:	92
5	Kết Luận và Hướng Phát Triển	94
5.1	Kết quả dự án	94
5.1.1	Tổng quan và Tiền xử lý dữ liệu	94
5.1.2	Phân tích tổng quan và hành vi khách hàng	95
5.1.3	Khai thác các mẫu phổ biến (Association Rule Mining)	95
5.1.4	Gom cụm dữ liệu (Clustering)	96
5.1.5	Phân loại dữ liệu (Classification)	97
5.2	Hạn chế của dự án và đề xuất những hướng cải thiện hoặc phát triển trong tương lai	97
5.2.1	Hạn chế của dự án	97
5.2.2	Hướng phát triển trong tương lai	98
5.3	Tài liệu tham khảo	99

Chương 1

GIỚI THIỆU BÀI TOÁN VÀ MÔ TẢ BỘ DỮ LIỆU

1.1 Giới thiệu và mục tiêu bài toán

Dự án này tập trung vào việc khai phá dữ liệu khách hàng trong ngành hàng không nhằm phân tích các yếu tố ảnh hưởng đến sự hài lòng của hành khách. Mục tiêu chính là chuyển đổi dữ liệu thô thành tri thức hữu ích, giúp hãng hàng không:

- Hiểu rõ hơn về hành vi và nhu cầu của các phân khúc khách hàng khác nhau.
- Xác định các điểm mạnh cần phát huy và các điểm yếu cần cải thiện trong chất lượng dịch vụ.
- Dự đoán mức độ hài lòng của khách hàng để có các chiến lược chăm sóc và giữ chân hiệu quả.
- Tối ưu hóa các chương trình khuyến mãi và dịch vụ cá nhân hóa.

1.2 Mô tả bộ dữ liệu

Mô tả chi tiết về bộ dữ liệu được sử dụng: nguồn gốc, số lượng thuộc tính, số lượng mẫu, ý nghĩa của các thuộc tính.

Kích thước: Ban đầu, bộ dữ liệu có 103.904 mẫu (khách hàng) và 23-24 thuộc tính (cột).

- 1. Id:** Mỗi dòng dữ liệu tương ứng với hành khách cụ thể hoặc một chuyến bay cá nhân, và id là mã định danh duy nhất cho mỗi hành khách đó. Dù không dùng trực tiếp trong phân tích nhưng id rất hữu ích để theo dõi, đối chiếu và truy xuất dữ liệu gốc (nếu cần).
- 2. Gender:** Đây là thông tin giới tính của hành khách (Male hoặc Female). Trong ngữ cảnh dịch vụ hàng không, giới tính có thể ảnh hưởng đến đánh giá dịch vụ, thói quen bay hay mức độ hài lòng. Phân tích yếu tố này giúp khám phá liệu có sự khác biệt đáng kể nào giữa nam và nữ trong cảm nhận về chuyến bay.
- 3. Customer Type:** Biến này cho biết khách hàng thuộc loại trung thành (Loyal Customer) hay không trung thành (Disloyal Customer). Đây là biến phân loại quan trọng, có thể phản ánh mức độ gắn bó của khách hàng với hãng bay. Khách trung thành thường đã trải nghiệm nhiều dịch vụ hơn và đưa ra đánh giá có chiều sâu hơn.
- 4. Age:** Tuổi của khách hàng là một biến định lượng liên tục, có thể ảnh hưởng đến nhu cầu và kỳ vọng dịch vụ. Ví dụ, khách hàng lớn tuổi có thể đánh giá cao sự thoải mái và hỗ trợ, trong khi khách hàng trẻ lại quan tâm đến giá cả hay tốc độ. Phân tích theo nhóm tuổi giúp hiểu rõ hơn sự khác biệt này và có thể hỗ trợ mô hình phân loại mức độ hài lòng.

5. **Type of Travel:** Thuộc tính này cho biết mục đích chuyến bay là cá nhân (Personal Travel) hay công tác (Business Travel). Đây là biến then chốt vì hành khách đi công tác thường yêu cầu sự đúng giờ và hiệu suất cao, còn người đi cá nhân có thể chú trọng đến sự thoải mái và trải nghiệm. Do đó, yếu tố này có thể ảnh hưởng mạnh đến đánh giá tổng thể.
6. **Class:** Đây là hạng vé mà hành khách sử dụng: Eco, Eco Plus hoặc Business. Hạng vé phản ánh mức độ dịch vụ được cung cấp. Hành khách hạng Business thường có kỳ vọng cao hơn, do đó mức độ hài lòng có thể cao hơn — hoặc ngược lại nếu kỳ vọng không được đáp ứng. Đây là biến có giá trị cao trong phân tích cảm nhận khách hàng.
7. **Flight Distance:** Đây là khoảng cách (tính bằng km) mà hành khách bay trong chuyến đó. Là một biến định lượng liên tục, khoảng cách này có thể ảnh hưởng đến trải nghiệm chuyến bay. Với các chuyến bay dài, hành khách có thể tiếp xúc dịch vụ lâu hơn, từ đó chịu ảnh hưởng bởi nhiều yếu tố như ghế ngồi, phục vụ, giải trí,... Biến này có thể đóng vai trò quan trọng trong mô hình dự đoán mức độ hài lòng.
8. **Inflight wifi service:** Đây là cột đánh giá mức độ hài lòng của khách hàng về dịch vụ cung cấp kết nối internet không dây trong suốt chuyến bay, trên thang điểm từ 1 đến 5. Dịch vụ Wi-Fi có thể ảnh hưởng lớn đến sự hài lòng và quyết định lựa chọn hãng bay, đặc biệt đối với hành khách đi công tác.
9. **Departure/Arrival time convenient:** Cột này phản ánh mức độ hài lòng của khách hàng về tính hợp lý và tiện lợi của thời gian cất cánh và hạ cánh so với kế hoạch cá nhân, trên thang điểm từ 1 đến

5. Thời gian bay thuận tiện giúp giảm thiểu căng thẳng và tăng trải nghiệm tích cực.
10. **Ease of Online booking:** Biến này thể hiện đánh giá của khách hàng về mức độ dễ sử dụng và thân thiện của quy trình đặt vé trực tuyến thông qua website hoặc ứng dụng di động, trên thang điểm từ 1 đến 5. Đây là yếu tố quan trọng trong trải nghiệm ban đầu của khách hàng với hãng bay.
11. **Gate location:** Cột này thể hiện sự thuận tiện và dễ dàng tiếp cận của cổng lên máy bay tại sân bay theo đánh giá của khách hàng, trên thang điểm từ 1 đến 5. Cổng lên máy bay quá xa hoặc khó tìm có thể làm giảm mức độ hài lòng chung.
12. **Food and drink:** Đây là đánh giá mức độ hài lòng của hành khách đối với chất lượng, sự đa dạng và khả năng phục vụ đồ ăn, thức uống trong chuyến bay, trên thang điểm từ 1 đến 5. Yếu tố này đặc biệt quan trọng với các chuyến bay dài và hành khách hạng cao.
13. **Online boarding:** Biến này đo lường sự hài lòng của khách hàng với khả năng tự làm thủ tục check-in và nhận thẻ lên máy bay qua internet (web/app), trên thang điểm từ 1 đến 5. Đây là một yếu tố thể hiện tính hiện đại và thuận tiện trong dịch vụ hàng không.
14. **Seat comfort:** Biến này phản ánh cảm nhận của khách hàng về sự thoải mái của ghế ngồi, bao gồm độ rộng, độ ngả, không gian để chân và độ mềm của đệm, trên thang điểm từ 1 đến 5. Đây là yếu tố then chốt trong trải nghiệm bay, đặc biệt ở các chuyến bay dài.
15. **Inflight entertainment:** Đây là đánh giá mức độ hài lòng về các lựa chọn giải trí được cung cấp trong suốt chuyến bay như phim ảnh,

chương trình TV, nhạc, trò chơi,... trên thang điểm từ 1 đến 5. Hệ thống giải trí chất lượng giúp giảm mệt mỏi và nâng cao trải nghiệm bay.

16. **Leg room service:** Mức độ hài lòng về không gian để chân tại ghế ngồi (thang điểm 1–5), mức độ thoải mái khi ngồi trên máy bay. Không gian hẹp sẽ gây khó chịu, đặc biệt ở chuyến bay dài, ảnh hưởng lớn đến mức độ hài lòng, liên quan đến thuộc tính *Seat comfort*.
17. **Baggage handling:** Dịch vụ vận chuyển hành lý. Đảm bảo hành lý đến nơi an toàn, không bị thất lạc hay hư hại.
18. **Checkin service:** Dịch vụ làm thủ tục check-in tại sân bay — giai đoạn đầu trong hành trình, ảnh hưởng đến ấn tượng ban đầu. Thủ tục chậm trễ hoặc thái độ nhân viên không tốt sẽ ảnh hưởng đến toàn bộ trải nghiệm chuyến đi.
19. **Inflight service:** Dịch vụ tiếp viên trên chuyến bay. Đây là yếu tố then chốt trong sự hài lòng toàn diện vì là giao tiếp trực tiếp, ảnh hưởng trực tiếp đến cảm xúc và thái độ của hành khách trong suốt chuyến bay.
20. **Cleanliness:** Độ sạch sẽ, yếu tố vệ sinh cơ bản. Sự sạch sẽ góp phần lớn trong ấn tượng tổng thể và thường có tương quan mạnh với mức độ hài lòng (*satisfaction*).
21. **Departure Delay in Minutes:** Trễ giờ khởi hành (phút) — số phút chuyến bay bị trễ so với giờ khởi hành dự kiến. Trễ giờ có thể làm hành khách bỏ lỡ chuyến kết nối (transit), gây bức xúc, giảm mức độ hài lòng.

22. **Arrival Delay in Minutes:** Trễ giờ đến nơi (phút) — số phút chuyến bay trễ so với giờ đến dự kiến. Có thể ảnh hưởng đến lịch trình của hành khách, đặc biệt là với chuyến công tác.
23. **satisfaction:** Mức độ hài lòng (biến mục tiêu). Phản hồi của hành khách về mức độ hài lòng tổng thể (*satisfied* hoặc *neutral or dissatisfied*). Biến này phụ thuộc vào tổng hòa các yếu tố dịch vụ, tiện nghi, thái độ nhân viên, và sự đúng giờ.

Chương 2

TIỀN XỬ LÝ DỮ LIỆU (DATA PREPROCESSING)

Tiền xử lý là giai đoạn quan trọng nhất trong quy trình khai phá dữ liệu, vì “Garbage In, Garbage Out!” — chất lượng dữ liệu đầu vào quyết định chất lượng kết quả khai phá [1].

2.1 Các kỹ thuật tiền xử lý dữ liệu

2.1.1 Xử lý Giá trị Thiếu (Handling Missing Values)

Cơ sở lý thuyết: Giá trị thiếu (NaN, NA) có thể do nhiều nguyên nhân và nếu không xử lý sẽ ảnh hưởng xấu đến các thuật toán.

Kỹ thuật đã dùng: Diền giá trị thiếu bằng median cho các cột dạng số.

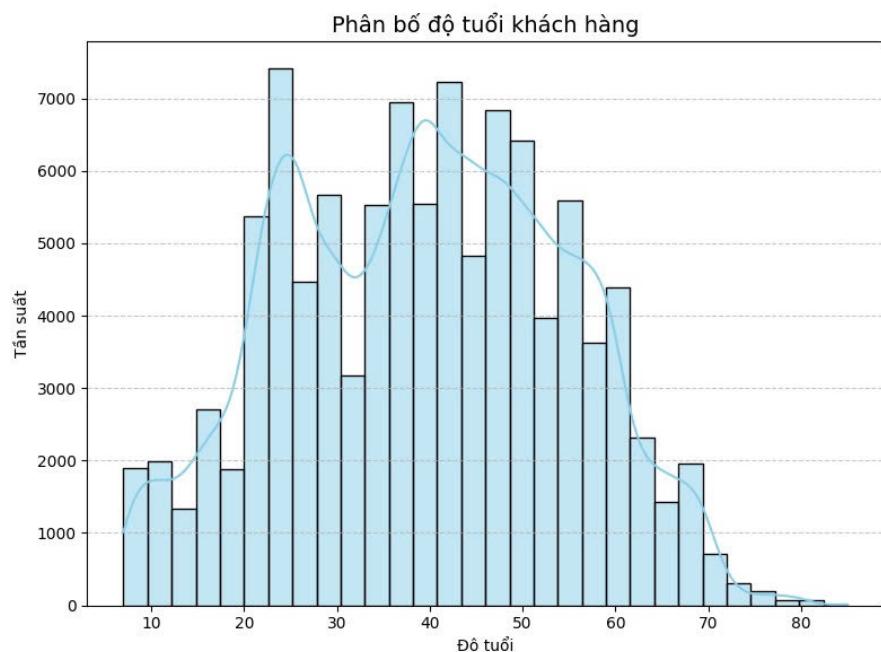
Lý do lựa chọn: Phân tích ban đầu cho thấy dữ liệu có 310 giá trị thiếu trên tổng số 103.904 mẫu (chỉ chiếm khoảng 0.3%). Tỷ lệ này là rất nhỏ, do đó việc xóa dòng sẽ gây mất mát thông tin không cần thiết. Đối với dữ liệu số, chúng tôi lựa chọn median vì nó ít nhạy cảm với các giá trị ngoại lai (outliers) hoặc dữ liệu có phân phối lệch (skewed distribution) hơn so với mean.

Bảng 2.1: Phân tích giá trị thiêu

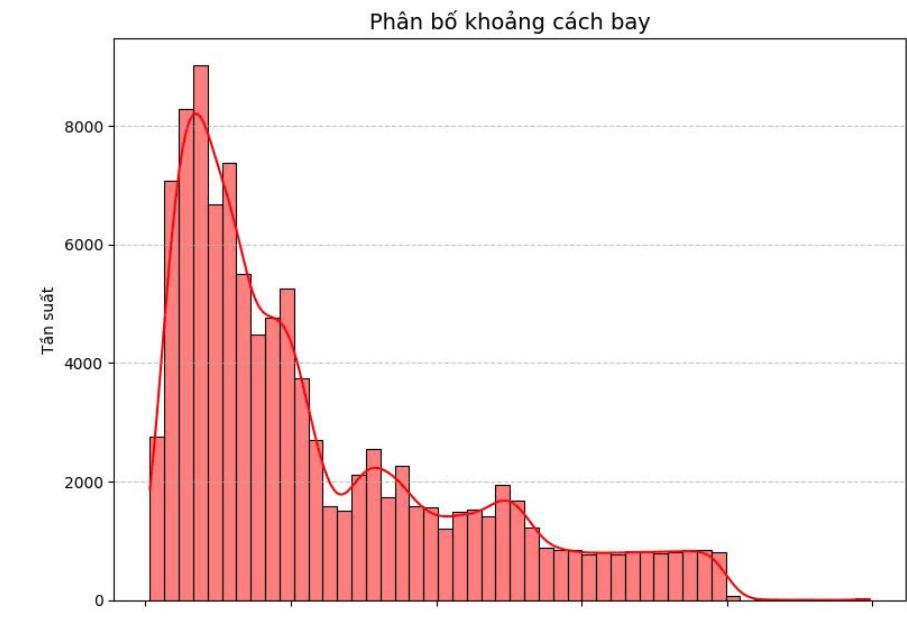
Thuộc tính	Số giá trị thiêu
Arrival Delay in Minutes	310
Các thuộc tính khác	0
Tổng cộng	310
Tỷ lệ	0.3%

```
# 3. missing value
print(". Phân tích missing values")
missing_values = df.isnull().sum()
missing_values_sorted = missing_values.sort_values(ascending=False)
print(missing_values_sorted[missing_values_sorted > 0])
```

Hình 2.1: Phân tích missing values: Arrival Delay in Minutes: 310



Hình 2.2: Lựa chọn median cho độ tuổi



Hình 2.3: Lựa chọn median cho khoảng cách

2.1.2 Chuẩn hóa Dữ liệu (Normalization/Standardization)

Cơ sở lý thuyết: Nhiều thuật toán khai phá dữ liệu, đặc biệt là các thuật toán dựa trên khoảng cách (như K-Means, PCA) hoặc các mô hình dựa trên gradient (như Logistic Regression), hoạt động hiệu quả nhất khi các đặc trưng (features) nằm trên cùng một thang đo. Nếu không chuẩn hóa, các đặc trưng có giá trị lớn hơn sẽ lấn át các đặc trưng khác.

Kỹ thuật đã dùng: Z-score Standardization (StandardScaler)

Công thức chuẩn hóa Z-score:

$$v' = \frac{v - \mu_A}{\sigma_A}$$

Trong đó:

- v là giá trị gốc trong cột dữ liệu
- μ_A là giá trị trung bình (mean) của cột A
- σ_A là độ lệch chuẩn (standard deviation) của cột A

- v' là giá trị sau khi chuẩn hóa (z-score)

Lý do lựa chọn: Z-score standardization biến đổi dữ liệu sao cho các cột có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1. Điều này đảm bảo rằng tất cả các đặc trưng đều có “tiếng nói” ngang bằng nhau trong mô hình. Phương pháp này ít nhạy cảm với outlier hơn Min-Max Normalization và phù hợp với các thuật toán giả định dữ liệu có phân phối gần chuẩn.

Bảng 2.2: Kết quả chuẩn hóa dữ liệu

Chỉ số	Giá trị
Missing values sau xử lý	0
Kích thước dữ liệu sau chuẩn hóa ($103,904 \times n_features$)	
Mean sau chuẩn hóa	0.000000
Std sau chuẩn hóa	1.000000

```

missing_after = df_clustering_airline.isnull().sum().sum()
print(f"Missing values sau xử lý: {missing_after}")

# Chuẩn hóa dữ liệu
scaler_clustering = StandardScaler()
df_scaled_clustering = scaler_clustering.fit_transform(df_clustering_airline)

print(f"Kích thước dữ liệu sau chuẩn hóa: {df_scaled_clustering.shape}")
print(f"Mean sau chuẩn hóa: {df_scaled_clustering.mean():.6f}")
print(f"Std sau chuẩn hóa: {df_scaled_clustering.std():.6f}")

# Tạo DataFrame từ dữ liệu đã chuẩn hóa
df_scaled_clustering_df = pd.DataFrame(df_scaled_clustering,
                                         columns=available_clustering_features)

```

Hình 2.4: Phát hiện và xử lý outliers

2.1.3 Biến đổi Dữ liệu (Data Transformation) - Log Transform

Cơ sở lý thuyết: Dữ liệu trong thực tế thường có phân phối lệch (skewed), đặc biệt là lệch phải (right-skewed), ví dụ như thu nhập hay khoảng cách bay. Điều này có thể ảnh hưởng tiêu cực đến hiệu suất của một số mô hình học máy. Phép biến đổi logarit giúp “kéo” các giá trị lớn lại gần nhau, làm cho phân phối trở nên đối xứng hơn.

Kỹ thuật đã dùng: Log Transformation (\log_{10}) cho các biến bị lệch nặng như Flight Distance, Departure Delay in Minutes, Arrival Delay in Minutes, Total_Delay.

Công thức Log:

$$x' = \log(1 + x)$$

Trong đó:

- x : Giá trị gốc (thường là số dương)
- x' : Giá trị sau khi biến đổi log
- \log : Hàm log cơ số tự nhiên (thường là \log_e , hay còn gọi là ln)

Lý do lựa chọn: Phân tích skewness đã chỉ ra rằng các biến liên quan đến khoảng cách bay và độ trễ có độ lệch lớn (>1). Log transformation là một kỹ thuật mạnh mẽ để giảm thiểu ảnh hưởng của các giá trị cực đoan và làm cho phân phối dữ liệu gần với phân phối chuẩn hơn.

```

# 5.2 Data Transformation với Log Transform
# Identify skewed variables
skewed_vars = []
for col in ['Flight Distance', 'Departure Delay in Minutes', 'Arrival Delay in Minutes', 'Total_Delay']:
    if col in df_processed.columns:
        skewness = stats.skew(df_processed[col].fillna(0))
        if abs(skewness) > 1:
            skewed_vars.append(col)
    print(f" {col}: skewness = {skewness:.3f} (cần transform)")

# Apply log transformation
for col in skewed_vars:
    df_processed[f'{col}_log'] = np.log1p(df_processed[col].fillna(0))
    skew_after = stats.skew(df_processed[f'{col}_log'])
    print(f" {col}_log: skewness sau transform = {skew_after:.3f}")

```

Hình 2.5: Data transformation với Log transform

Bảng 2.3: Skewness trước và sau khi log-transform

Biến	Skewness	Ghi chú
Flight Distance	1.109	cần transform
Departure Delay in Minutes	6.734	cần transform
Arrival Delay in Minutes	6.605	cần transform
Total_Delay	6.787	cần transform
Flight Distance_log	-0.204	sau transform
Departure Delay in Minutes_log	0.917	sau transform
Arrival Delay in Minutes_log	0.877	sau transform
Total_Delay_log	0.636	sau transform

2.1.4 Xử lý Ngoại lai (Outlier Treatment)

Cơ sở lý thuyết: Ngoại lai là các quan sát lệch đáng kể so với phần còn lại của dữ liệu, có thể là lỗi nhập liệu hoặc các trường hợp đặc biệt. Chúng có thể làm sai lệch các phép tính thống kê và ảnh hưởng tiêu cực đến hiệu suất của mô hình.

Kỹ thuật đã dùng: Capping (giới hạn giá trị) theo phương pháp IQR

(Interquartile Range) cho các cột có ngoại lai đáng kể.

Nguồn tính toán:

- Q1 (quartile 1): Phân vị 25%
- Q3 (quartile 3): Phân vị 75%
- IQR: Q3 - Q1 (khoảng giữa 50% dữ liệu)
- Nguồn dưới: $Q1 - 1.5 \times IQR$
- Nguồn trên: $Q3 + 1.5 \times IQR$

Lý do lựa chọn: Phân tích ngoại lai cho thấy các cột như Departure Delay in Minutes và Arrival Delay in Minutes có tỷ lệ ngoại lai cao (hơn 13%). Việc capping các giá trị này vào nguồn trên/dưới được tính từ IQR giúp giảm thiểu ảnh hưởng của các trường hợp cực đoan mà không loại bỏ hoàn toàn các điểm dữ liệu quý giá.

Bảng 2.4: Phân tích outliers theo các phương pháp

Biến	IQR (%)	Z-Score (%)	Modified Z (%)
Age	0 (0.0%)	17 (0.0%)	0 (0.0%)
Flight Distance	2291 (2.2%)	58 (0.1%)	4042 (3.9%)
Departure Delay	14529 (14.0%)	2222 (2.1%)	34459 (33.2%)
Arrival Delay	13954 (13.4%)	2225 (2.1%)	35643 (34.3%)
Customer Value Score	2002 (1.9%)	50 (0.0%)	4930 (4.7%)

Bảng 2.5: Kết quả xử lý outliers bằng IQR capping

Biến	Số outliers đã xử lý
Flight Distance	2,291
Departure Delay in Minutes	14,529
Arrival Delay in Minutes	13,954
Customer Value Score	2,002

```
# xử lý ngoại lai
def detect_outliers_comprehensive(data, column):
    """Comprehensive outlier detection"""
    clean_data = data[column].fillna(data[column].median())

    # IQR method
    Q1, Q3 = clean_data.quantile([0.25, 0.75])
    IQR = Q3 - Q1
    iqr_lower, iqr_upper = Q1 - 1.5 * IQR, Q3 + 1.5 * IQR
    iqr_outliers = clean_data[(clean_data < iqr_lower) | (clean_data > iqr_upper)]

    # Z-score method
    z_scores = np.abs(stats.zscore(clean_data))
    z_outliers = clean_data[z_scores > 3]

    # Modified Z-score method
    median_val = clean_data.median()
    mad = np.median(np.abs(clean_data - median_val))
    if mad == 0:
        mad = 1
    modified_z_scores = 0.6745 * (clean_data - median_val) / mad
    mod_z_outliers = clean_data[np.abs(modified_z_scores) > 3.5]

    return {
        'IQR': {'count': len(iqr_outliers), 'percentage': len(iqr_outliers)/len(clean_data)*100, 'bounds': (iqr_lower, iqr_upper)},
        'Z_Score': {'count': len(z_outliers), 'percentage': len(z_outliers)/len(clean_data)*100},
        'Modified_Z': {'count': len(mod_z_outliers), 'percentage': len(mod_z_outliers)/len(clean_data)*100}
    }
```

Hình 2.6: Xử lý ngoại lai

```

print(f"\nOutlier treatment (IQR capping):")
for col in outlier_analysis_cols:
    if col in df_processed.columns and outlier_results[col]['IQR']['count'] > 0:
        bounds = outlier_results[col]['IQR']['bounds']
        outlier_count = outlier_results[col]['IQR']['count']

        df_processed[col] = np.clip(df_processed[col], bounds[0], bounds[1])
        print(f"{col}: xử lý {outlier_count} outliers")

outlier_analysis_cols = ['Age', 'Flight Distance', 'Departure Delay in Minutes',
                        'Arrival Delay in Minutes', 'Customer_Value_Score']
outlier_results = {}

for col in outlier_analysis_cols:
    if col in df_processed.columns:
        results = detect_outliers_comprehensive(df_processed, col)
        outlier_results[col] = results

    print(f"\n{col}:")
    for method, info in results.items():
        print(f"  {method}: {info['count']} outliers ({info['percentage']:.1f}%)")

```

Hình 2.7: Kết quả transformation và outlier treatment

2.1.5 Lựa chọn Đặc trưng (Feature Selection) & Giảm chiều dữ liệu (PCA)

Cơ sở lý thuyết: Với số lượng đặc trưng lớn, việc lựa chọn các đặc trưng quan trọng nhất giúp loại bỏ nhiễu, giảm độ phức tạp của mô hình, cải thiện hiệu suất và khả năng diễn giải. PCA là kỹ thuật giảm chiều giúp nén thông tin từ nhiều đặc trưng thành ít thành phần chính hơn, đồng thời loại bỏ tương quan giữa các đặc trưng đó.

Kỹ thuật đã dùng: Kết hợp các phương pháp F-test, Mutual Information, Random Forest Importance, và RFE để chọn ra 23 đặc trưng tốt nhất, sau đó áp dụng PCA để giảm chiều và tạo các thành phần PCA cho việc gom cụm.

- **F-test:** Tìm mối quan hệ tuyến tính mạnh

- **Mutual Information:** Tìm mối quan hệ phi tuyến tính
- **Random Forest Importance:** Dựa trên mức độ giảm Gini Impurity
- **RFE:** Loại bỏ đặc trưng ít quan trọng một cách đệ quy
- **PCA:** n_components được chọn để giữ lại 90% phương sai

Lý do lựa chọn: Sử dụng nhiều phương pháp lựa chọn đặc trưng (filter, wrapper, embedded) giúp có cái nhìn toàn diện về tầm quan trọng của các đặc trưng, từ mối quan hệ tuyến tính đơn giản đến các mối quan hệ phi tuyến phức tạp. PCA là cần thiết để nén thông tin từ 23 đặc trưng đã chọn xuống còn các thành phần chính, giúp đơn giản hóa dữ liệu cho các thuật toán gom cụm và trực quan hóa.

```

#5.3 Lựa chọn các đặc trưng
# Mã hóa các biến phân loại thành số để máy tính có thể xử lý
categorical_cols = ['Gender', 'Customer Type', 'Type of Travel', 'Class'] # Danh sách các cột dữ liệu phân loại
label_encoders = {} # Dictionary để lưu trữ các bộ mã hóa

# Vòng lặp để mã hóa từng cột phân loại
for col in categorical_cols:
    le = LabelEncoder() # Tạo bộ mã hóa nhãn mới
    df_processed[col + '_encoded'] = le.fit_transform(df_processed[col]) # Mã hóa và tạo cột mới với hậu tố '_encoded'
    label_encoders[col] = le # Lưu bộ mã hóa để sử dụng sau

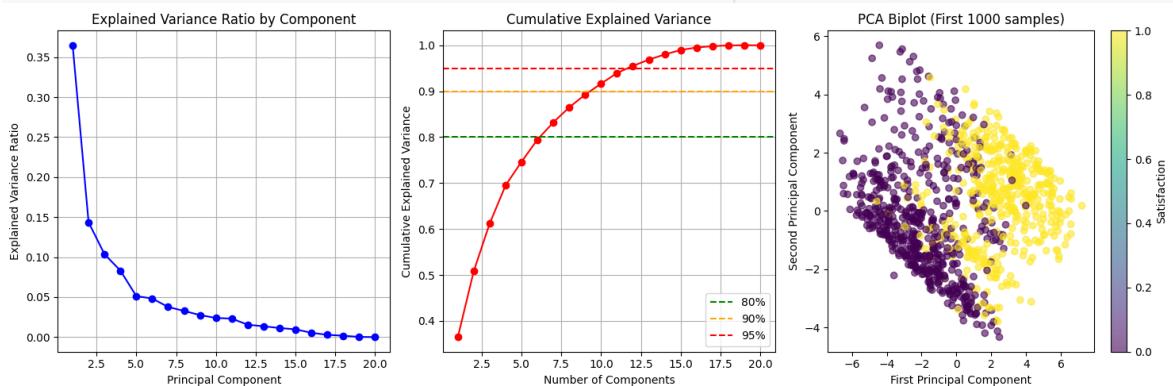
# Mã hóa biến mục tiêu satisfaction thành 0 và 1
df_processed['satisfaction_numeric'] = (df_processed['satisfaction'] == 'satisfied').astype(int) # 1 nếu satisfied, 0 nếu không

# Chuẩn bị danh sách các đặc trưng để chọn lọc
feature_cols = [col for col in df_processed.columns if
                col not in ['satisfaction', 'Unnamed: 0', 'id'] and # Loại bỏ các cột không cần thiết
                df_processed[col].dtype in ['int64', 'float64'] and # Chỉ lấy cột số
                col not in ['Customer_Tier']] # Loại bỏ biến phân loại đã tạo

# Tạo ma trận đặc trưng X và vector mục tiêu y
X = df_processed[feature_cols].fillna(df_processed[feature_cols].mean()) # Đèn giá trị thiếu bằng trung bình
y = df_processed['satisfaction_numeric'] # Biến mục tiêu

print(f" Features available for selection: {len(feature_cols)}") # In số lượng đặc trưng có sẵn

```



Hình 2.8: Feature selection và PCA

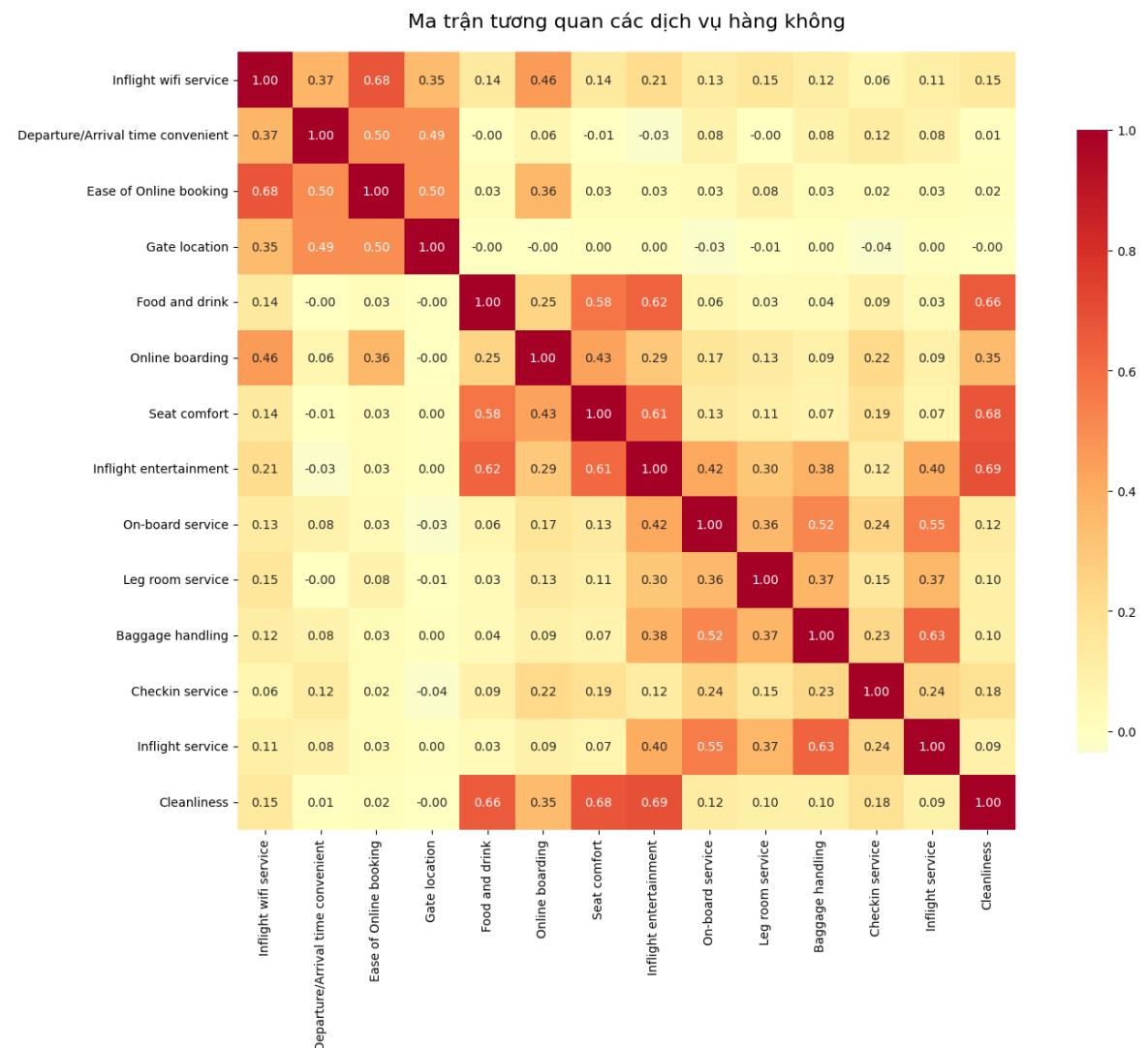
Bảng 2.6: Phân tích PCA cho dữ liệu hàng không

Biểu đồ trái (Explained Variance Ratio)	Biểu đồ giữa (Cumulative Variance)	Biểu đồ phải (PCA Biplot)
PC1 chiếm ~37% phương sai, PC2 chỉ ~14% → Có 1 yếu tố chính thống trị dữ liệu hàng không	4-5 thành phần đạt 80% (đường xanh lá)	Màu vàng vs tím cho thấy 2 nhóm khách hàng tách biệt rõ ràng
Các thành phần tiếp theo giảm nhanh → Dữ liệu có cấu trúc đơn giản, tập trung	8-10 thành phần đạt 90% (đường cam)	Phân bố theo cấu trúc có hệ thống , không ngẫu nhiên
	15 thành phần đạt 95% (đường đỏ) → Hiệu quả giảm chiều rất cao , có thể nén từ hàng chục đặc trưng xuống dưới 10	PC1 (trục ngang) là ranh giới chính phân chia các nhóm

Kết luận Dữ liệu hàng không có cấu trúc customer segments tự nhiên rất rõ ràng. PCA xác nhận việc phân khách thành các tier có cơ sở khoa học, chỉ với 2 chiều chính đã phân biệt hiệu quả các nhóm từ VIP đến Economy. Tạo nền tảng vững chắc cho cá nhân hóa dịch vụ và chiến lược kinh doanh.

2.2 Thống kê mô tả và trực quan hóa các dịch vụ

2.2.1 Ma trận tương quan các dịch vụ hàng không



Hình 2.9: Ma trận tương quan các dịch vụ hàng không

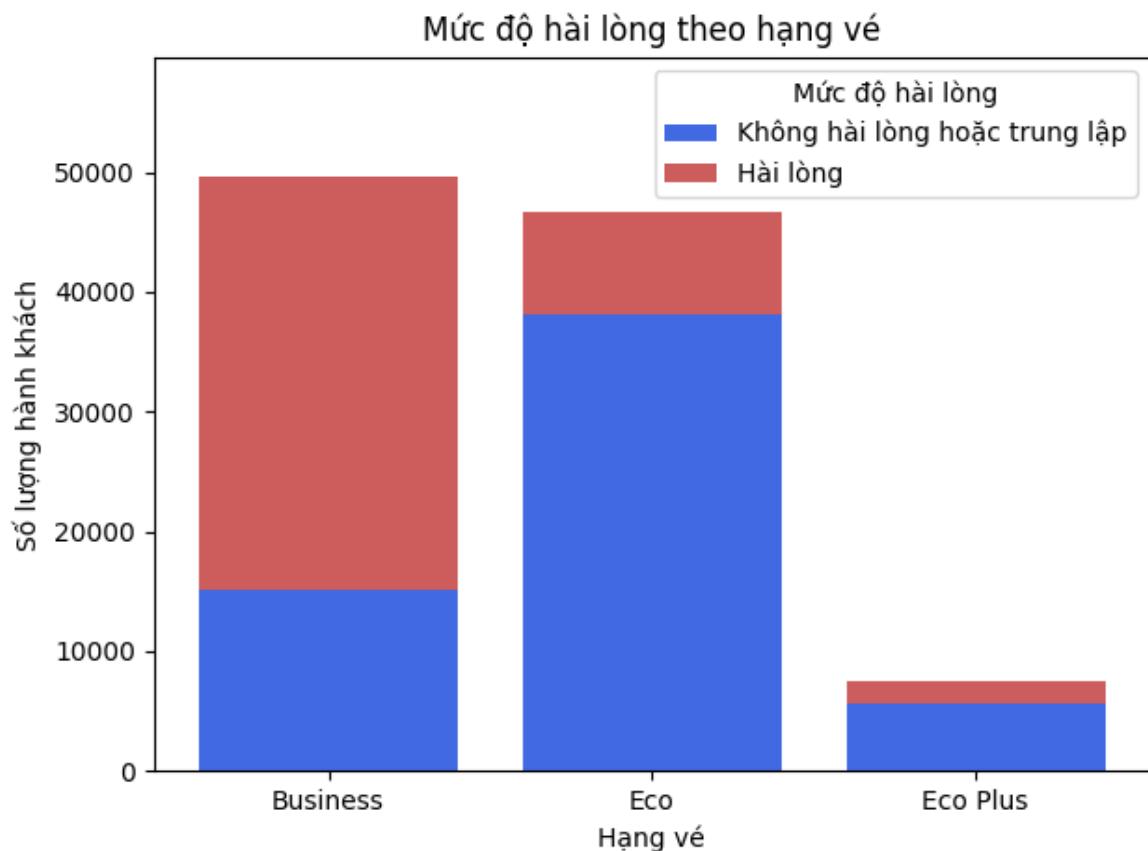
Mô tả: Biểu đồ này là một heatmap thể hiện mức độ tương quan (correlation) giữa 14 cột đánh giá dịch vụ khác nhau. Mỗi ô màu sắc trên biểu đồ cho biết

mỗi quan hệ giữa hai dịch vụ, với màu đỏ đậm thể hiện tương quan nghịch mạnh và màu xanh đậm thể hiện tương quan thuận mạnh.

Insight: Biểu đồ giúp xác định các cặp dịch vụ có mối liên hệ chặt chẽ. Ví dụ, các dịch vụ như Seat comfort, Inflight entertainment, Food and drink, Inflight service, Cleanliness, On-board service, và Leg room service thường có tương quan dương mạnh với nhau (hệ số tương quan > 0.6). Điều này cho thấy khách hàng có xu hướng đánh giá các dịch vụ này một cách nhất quán.

2.2.2 Phân tích mức độ hài lòng theo các đặc trưng

Mức độ hài lòng theo hạng vé:

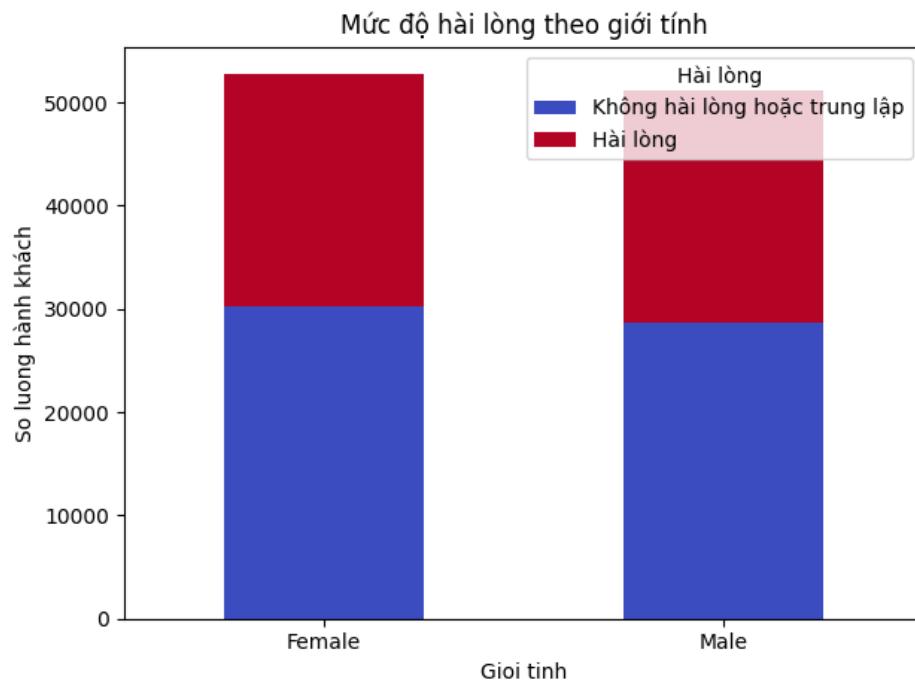


Hình 2.10: Mức độ hài lòng theo hạng vé

Insight: Biểu đồ cho thấy một mối quan hệ rõ rệt giữa hạng vé và mức

độ hài lòng. Khách hàng hạng Business Class có tỷ lệ hài lòng cao nhất với khoảng 69.4% hài lòng, trong khi khách hàng hạng Eco Class có tỷ lệ không hài lòng cao với khoảng 81.4% không hài lòng.

Mức độ hài lòng theo giới tính:



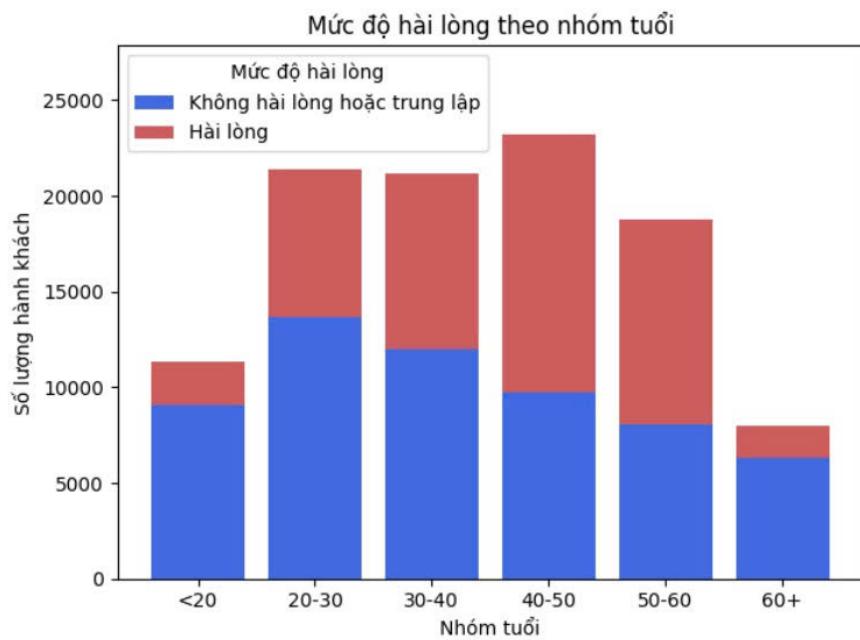
Hình 2.11: Mức độ hài lòng theo giới tính

Gender	Neutral or Dissatisfied (%)	Satisfied (%)
Female	57.3	42.7
Male	56.1	43.9

Bảng 2.7: Tỉ lệ mức độ hài lòng theo giới tính

Insight: Phân tích cho thấy không có sự khác biệt đáng kể về mức độ hài lòng giữa khách hàng nam và nữ.

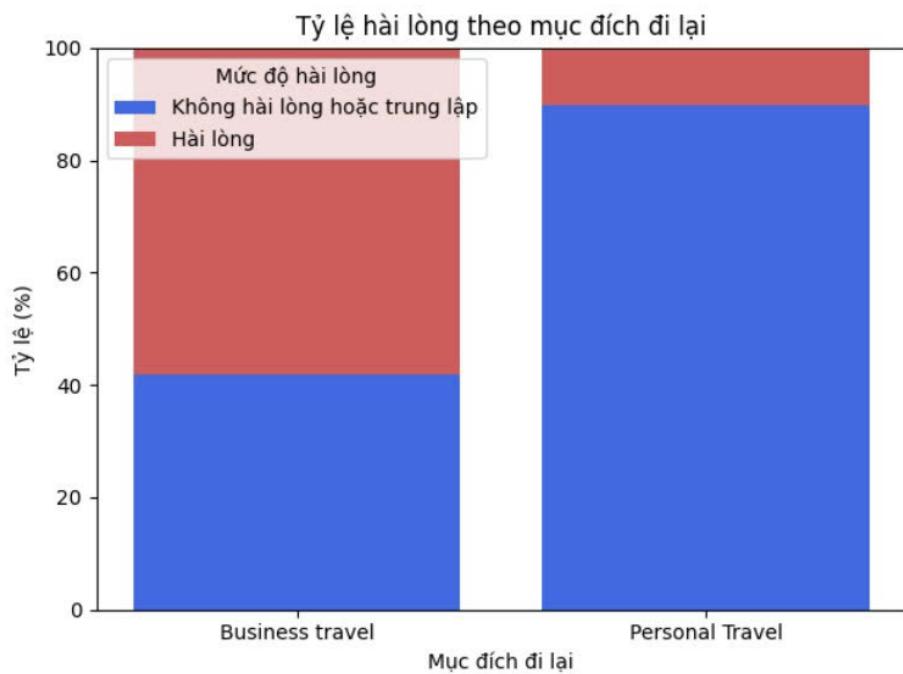
Mức độ hài lòng theo nhóm tuổi:



Hình 2.12: Mức độ hài lòng theo nhóm tuổi

Insight: Phân tích cho thấy có sự khác biệt về độ tuổi trung bình giữa nhóm hài lòng (41.8 tuổi) và không hài lòng (37.6 tuổi). Kiểm định T-test ($p\text{-value} = 0.000$) cho thấy sự khác biệt này có ý nghĩa thống kê.

Tỷ lệ hài lòng theo mục đích đi lại:

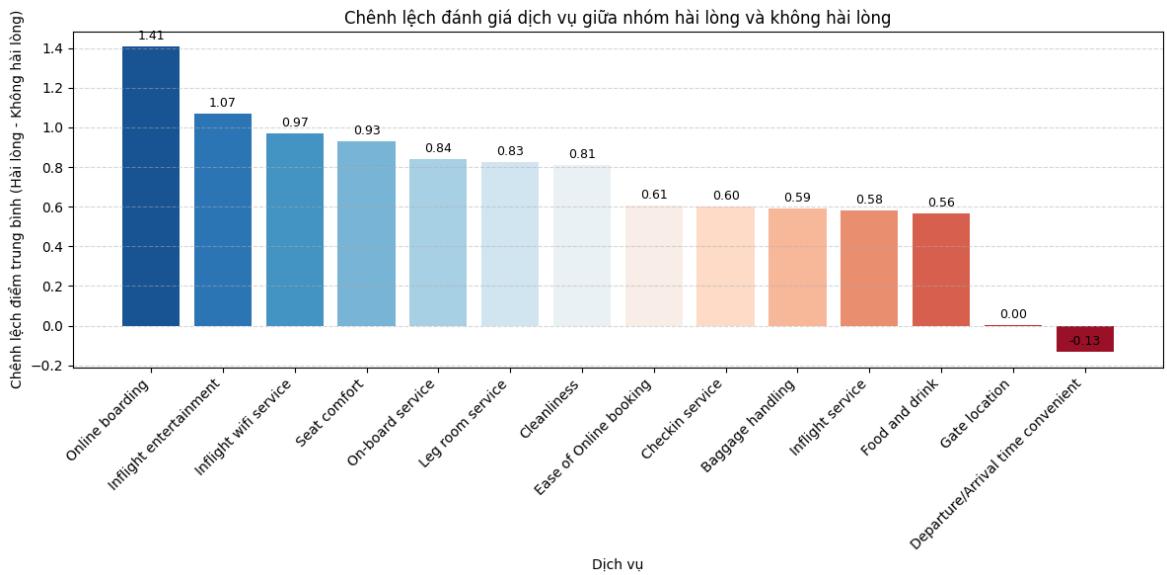


Hình 2.13: Tỷ lệ hài lòng theo mục đích đi lại

Insight: Nhóm khách hàng đi công tác có tỷ lệ hài lòng cao hơn đáng kể (khoảng 58%) so với nhóm đi cá nhân (khoảng 10-15%).

2.2.3 Chênh lệch đánh giá dịch vụ và các chỉ số khác

Chênh lệch đánh giá dịch vụ giữa nhóm hài lòng và không hài lòng:



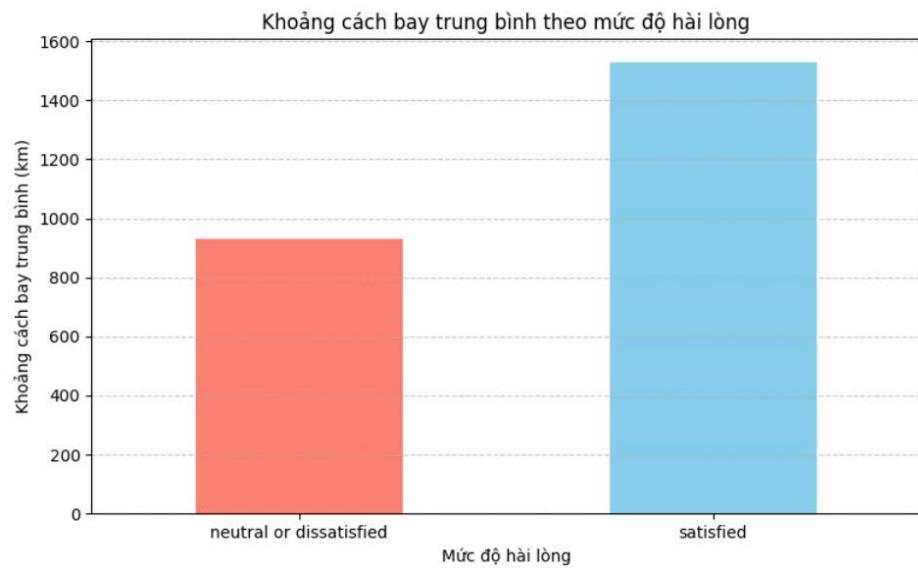
Hình 2.14: Chênh lệch đánh giá dịch vụ giữa nhóm hài lòng và không hài lòng

Insight: Dịch vụ Online boarding có chênh lệch lớn nhất (1.41 điểm), tiếp theo là Inflight entertainment (1.07 điểm) và Inflight wifi service (0.97 điểm), cho thấy đây là những yếu tố ảnh hưởng mạnh đến sự hài lòng của khách hàng.

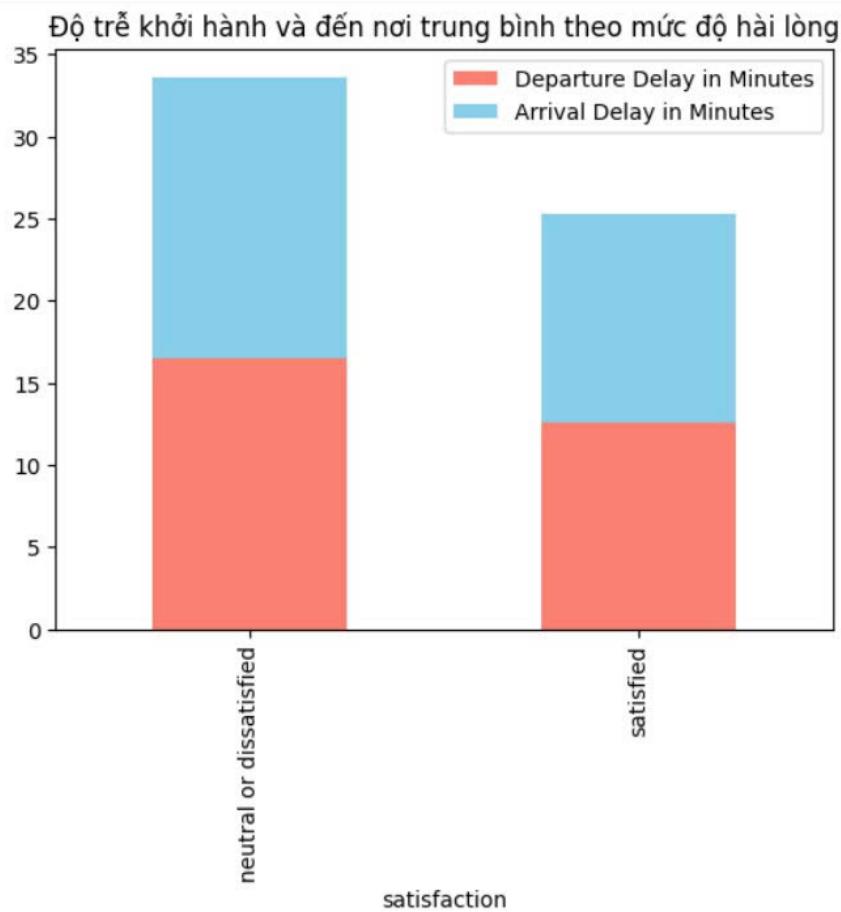
Khoảng cách bay và độ trễ theo mức độ hài lòng:

Bảng 2.8: So sánh các chỉ số theo mức độ hài lòng

Chỉ số	Hài lòng	Không hài lòng
Khoảng cách bay trung bình (km)	1,530.14	928.92
Độ trễ khởi hành trung bình (phút)	14.8	14.8
Độ trễ đến nơi trung bình (phút)	15.2	15.2



Hình 2.15: Khoảng cách bay trung bình theo mức độ hài lòng



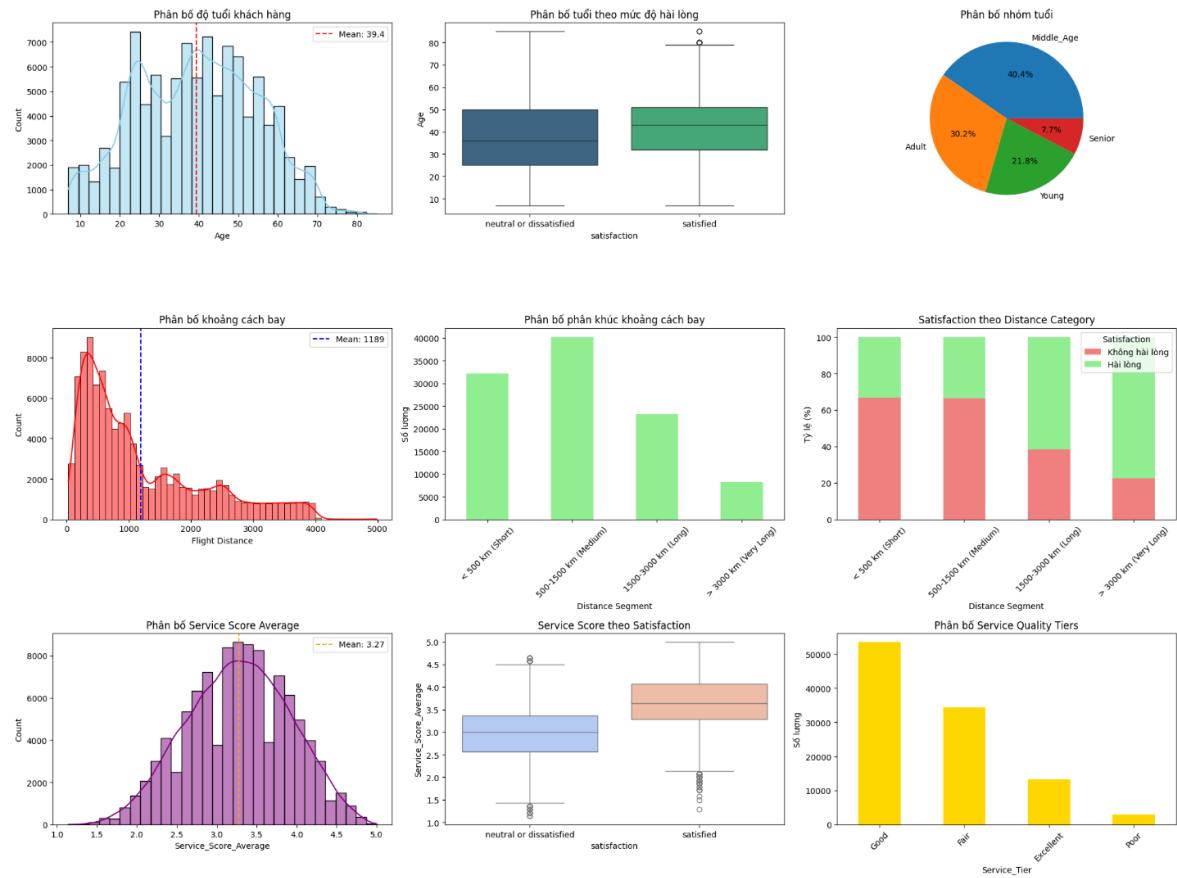
Hình 2.16: Độ trễ khởi hành và đến nơi trung bình theo mức độ hài lòng

Insight: Biểu đồ cho thấy khách hàng hài lòng có xu hướng bay các chuyến bay dài hơn (**trung bình 1530.14 km**) so với khách hàng không hài lòng (**trung bình 928.92 km**). Điều này có thể phản ánh rằng các chuyến bay dài hơn thường đi kèm với các dịch vụ tốt hơn (ví dụ: hạng Business), hoặc trải nghiệm trên các chuyến bay dài ảnh hưởng mạnh mẽ hơn đến cảm nhận tổng thể về dịch vụ.

Bảng 2.9: Độ trễ trung bình theo từng nhóm đặc trưng

Nhóm phân loại	Loại	Trung bình (phút)
Theo loại khách hàng	Loyal Customer (khởi hành)	14.743
	Disloyal Customer (khởi hành)	15.142
	Loyal Customer (đến)	15.092
	Disloyal Customer (đến)	15.567
Theo loại hình du lịch	Business Travel (khởi hành)	14.955
	Personal Travel (khởi hành)	14.506
	Business Travel (đến)	15.326
	Personal Travel (đến)	14.851
Theo hạng vé	Business (khởi hành)	14.398
	Eco (khởi hành)	15.161
	Eco Plus (khởi hành)	15.432
	Business (đến)	14.577
	Eco (đến)	15.672
	Eco Plus (đến)	16.089

2.2.4 Dashboard tổng quan dữ liệu khách hàng



Hình 2.17: Dashboard tổng quan dữ liệu khách hàng

Insight:

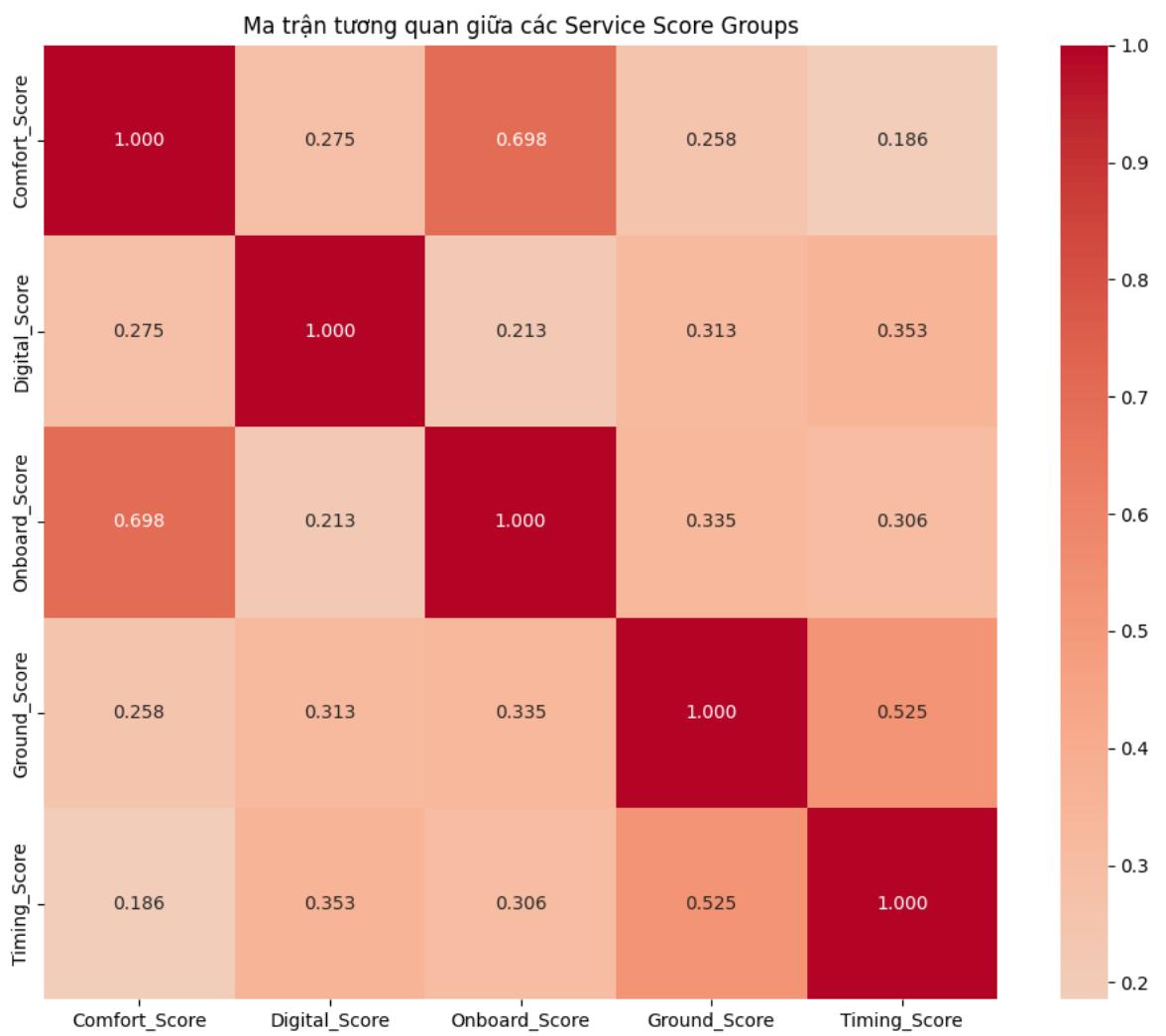
- **Phân bố độ tuổi:** Hầu hết khách hàng nằm trong độ tuổi 30-50
- **Phân bố khoảng cách bay:** Phân khúc bay trung bình (500-1500km) và dài (1500-3000km) chiếm đa số
- **Tỷ lệ hài lòng:** Khoảng 43.3% khách hàng hài lòng, trong khi 56.7% không hài lòng hoặc trung lập
- **Điểm dịch vụ trung bình:** 3.274/5.0 với độ lệch chuẩn thấp (0.647),

chứng tỏ trải nghiệm dịch vụ khá đồng đều nhưng vẫn có nhiều tiềm năng cải thiện

2.3 Lọc dữ liệu và phân tích so sánh nhóm khách hàng

2.3.1 Ma trận tương quan giữa các Service Score Groups

Insight: Biểu đồ này giúp hiểu các nhóm dịch vụ nào có xu hướng được đánh giá cùng chiều hoặc ngược chiều. Các nhóm điểm như Comfort và Onboard có tương quan dương mạnh, cho thấy chúng thường được khách hàng cảm nhận cùng nhau.

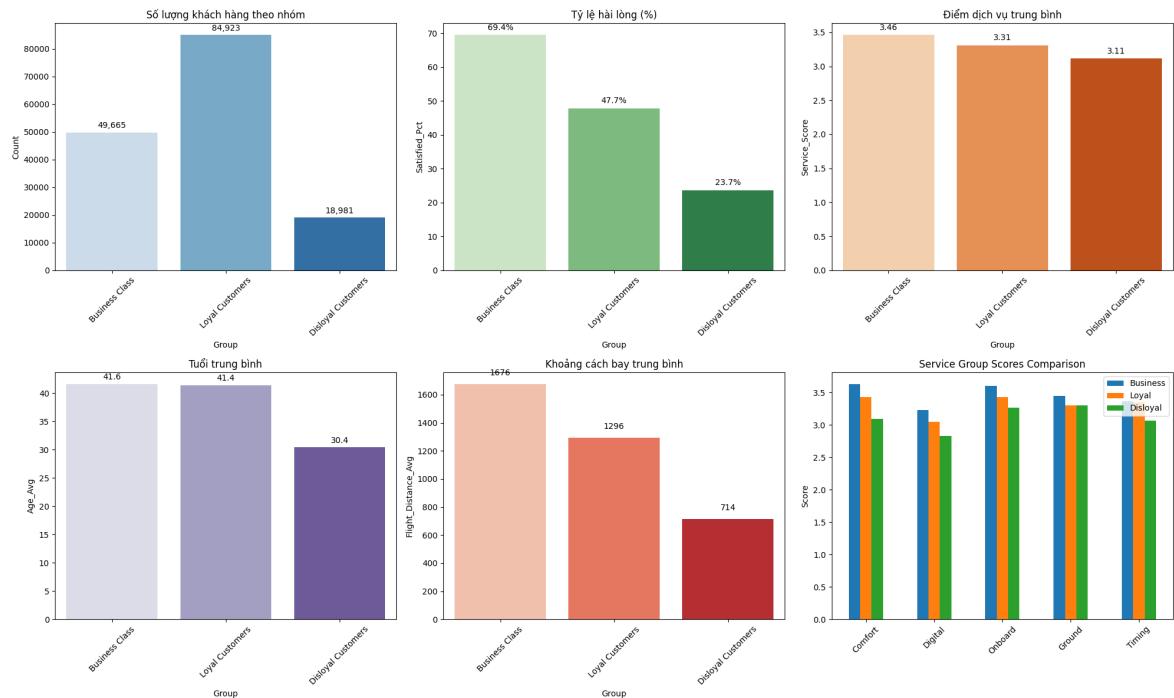


Hình 2.18: Ma trận tương quan giữa các Service Score Groups

2.3.2 So sánh tổng quan 3 nhóm khách hàng chính

Bảng 2.10: Phân tích toàn diện ba nhóm khách hàng chính

Chỉ số	Business Class	Loyal Customers	Disloyal Customers
Số lượng	49,665	84,923	18,981
Mức độ hài lòng	Satisfied: 69.4% Not satisfied: 30.6%	Satisfied: 47.7% Not satisfied: 52.3%	Satisfied: 23.7% Not satisfied: 76.3%
Service Score TB	3.461 ± 0.665	3.309 ± 0.654	3.113 ± 0.587
Tuổi TB	41.6 ± 12.8	41.4 ± 15.1	30.4 ± 11.2
Khoảng cách bay TB	1676 ± 1137	1296 ± 1048	714 ± 500
Comfort Score	3.630	3.425	3.089
Digital Score	3.228	3.049	2.831
Onboard Score	3.602	3.433	3.261
Ground Score	3.448	3.305	3.302
Timing Score	3.361	3.349	3.069



Hình 2.19: So sánh tổng quan 3 nhóm khách hàng chính

Insight và Phân tích:

First-order Insight - Dữ liệu thô mô tả

- **Business Class:** 49.665 khách hàng, tỷ lệ hài lòng 69.4%, điểm dịch vụ 3.46
- **Loyal Customers:** 84.923 khách hàng, tỷ lệ hài lòng 47.7%, điểm dịch vụ 3.31
- **Disloyal Customers:** 18.981 khách hàng, tỷ lệ hài lòng 23.7%, điểm dịch vụ 3.12

Second-order Insight - So sánh, đối chiếu, bối cảnh

- Business Class có tỷ lệ hài lòng và điểm dịch vụ cao nhất
- Loyal Customers đồng đảo nhất nhưng mức độ hài lòng chỉ ở mức trung bình
- Disloyal Customers có mức độ hài lòng rất thấp

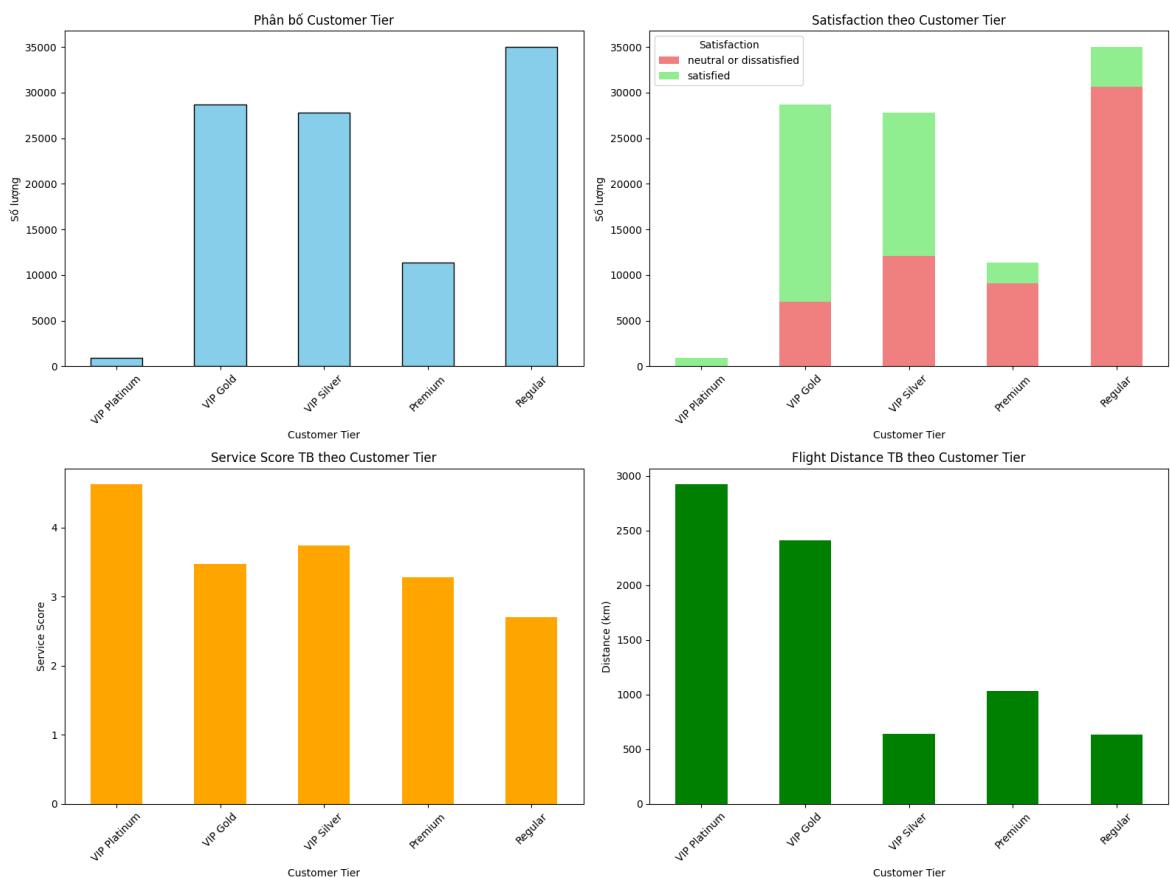
Third-order Insight - Nguyên nhân – Hệ quả – Đề xuất hành động

- **Business Class:** Duy trì và nâng tầm dịch vụ, cá nhân hóa trải nghiệm
- **Loyal Customers:** Ưu tiên nâng trải nghiệm, tăng tương tác và giải quyết bức xúc
- **Disloyal Customers:** Tái thiết dịch vụ nền tảng để chuyển hóa thành khách hàng trung thành

2.3.3 Phân tích chi tiết các Customer Tier

Bảng 2.11: Phân tích Customer Tier

Customer Tier	Số lượng	Tỷ lệ hài lòng (%)
VIP Silver	39,455	25.4
VIP Gold	38,563	56.9
VIP Platinum	14,261	80.5
Premium	6,806	—
Regular	4,819	13.2



Hình 2.20: Phân tích chi tiết các Customer Tier

Insight và Phân tích:

Dữ liệu mô tả: VIP Silver và VIP Gold là hai nhóm lớn nhất. Tỷ lệ hài lòng tăng dần theo cấp bậc Tier, với VIP Platinum có tỷ lệ hài lòng cao nhất (80.5%). Mức độ hài lòng theo Tier: Tỷ lệ hài lòng tăng dần theo cấp bậc Tier. VIP Platinum có tỷ lệ hài lòng cao nhất (80.5), , tiếp theo là VIP Gold (56.9%), VIP Silver (25.4%), và Regular (13.2%)

So sánh & đối chiếu: Nhóm Regular đồng đảo nhưng có tỷ lệ hài lòng rất thấp, trong khi nhóm VIP có tỷ lệ hài lòng vượt trội.

Nhân viên & Hết quả: Dịch vụ và ưu đãi dành cho khách VIP thực sự tạo ra trải nghiệm vượt trội. Nhóm Regular bị bỏ quên về trải nghiệm, dẫn đến bất mãn và dễ rời bỏ hãng.

Chương 3

ÁP DỤNG CÁC MÔ HÌNH KHAI PHÁ DỮ LIỆU

3.1 CÁC KỸ THUẬT KHAI THÁC MÃU PHỐ BIỀN

3.1.1 KHAI THÁC TẬP PHỐ BIỀN VỚI APRI- ORI

- Cơ sở lý thuyết của thuật toán Apriori là một thuật toán kinh điển trong khai phá dữ liệu, dùng để khai thác các tập mục phổ biến (frequent itemsets) từ cơ sở dữ liệu giao dịch, và từ đó sinh ra các luật kết hợp (association rules).
- Thuật toán Apriori sử dụng nguyên lý "**cắt tia**" không gian tìm kiếm: Mọi tập con của một tập phổ biến cũng phải là một tập phổ biến
 - – **Support (Độ hỗ trợ):**

$$\text{Support}(X \rightarrow Y) = \frac{\text{Số giao dịch chứa cả } X \text{ và } Y}{\text{Tổng số giao dịch}} = P(X \cup Y)$$

– **Confidence (Độ tin cậy):**

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Số giao dịch chứa cả } X \text{ và } Y}{\text{Số giao dịch chứa } X} = P(Y|X)$$

– **Lift (Độ nâng):**

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Confidence}(X \rightarrow Y)}{\text{Support}(Y)} = \frac{P(X \cup Y)}{P(X) \times P(Y)}$$

- Các bước chính của thuật toán:

1. Tìm tập mục phổ biến cấp 1 (các mục xuất hiện lớn hơn hoặc bằng ngưỡng hỗ trợ tối thiểu $minsup$).
2. Sinh tập ứng viên cấp k (C_k) bằng cách kết hợp các tập phổ biến cấp ($k - 1$).
3. Lọc các tập không đủ ngưỡng hỗ trợ ($minsup$) để thu được tập phổ biến cấp k (L_k).
4. Lặp lại quá trình cho đến khi không còn tập ứng viên mới được sinh ra.

- Lý do lựa chọn : Thuật toán Apriori phù hợp với bài toán khai thác tập phổ biến do tính đơn giản, dễ triển khai và hiệu quả rõ rệt với dữ liệu có cấu trúc giao dịch. Nó là bước nền quan trọng trước khi sinh ra các luật kết hợp phục vụ phân tích và ra quyết định.

- Quá trình thực thi và các tham số được lựa chọn:

(a) **Biến đổi dữ liệu sang dạng giao dịch (Transaction)**

Ý tưởng biến đổi:

Mỗi khách hàng được xem như là một **giao dịch** (transaction).

Các đặc điểm hoặc hành vi của khách hàng được mã hóa thành các **item** trong giao dịch đó.

Chỉ những đặc điểm thỏa mãn điều kiện cụ thể mới được xem là "có item". Ví dụ:

- Dịch vụ mà khách hàng đã đánh giá ≥ 4 điểm.
- Khách hàng thuộc vào một nhóm cụ thể (ví dụ: nhóm khách hàng thân thiết, nhóm độ tuổi,...).

Các bước thực hiện:

- Lọc các thông tin cần thiết từ dữ liệu gốc.
- Gộp thông tin thành từng giao dịch cho mỗi khách hàng.
- Biến đổi dữ liệu thành dạng one-hot encoding sử dụng TransactionEncoder của thư viện mlxtend.

Tham số: df_sample: DataFrame chứa dữ liệu mẫu **Trả về:** transactions: List of lists, mỗi list chứa các items của 1 khách hàng

```

def create_airline_transactions(df_sample):
    transactions = []

    # Duyệt qua từng khách hàng (từng dòng) trong DataFrame
    for idx, row in df_sample.iterrows():
        transaction = [] # Khởi tạo transaction rỗng cho khách hàng này

        # 1. XỬ LÝ CÁC ĐÁNH GIÁ DỊCH VỤ
        # Chỉ coi là "có" dịch vụ nếu khách hàng đánh giá >= 4 điểm (tốt)
        for service in service_rating_cols:
            if pd.notna(row[service]) and row[service] >= 4:
                # Tạo tên item dạng "Good_ServiceName"
                service_name = service.replace(' ', '_').replace('/', '_')
                transaction.append(f"Good_{service_name}")

        # 2. XỬ LÝ THÔNG TIN NHÂN KHẨU HỌC
        # Thêm các đặc điểm nhân khẩu học như các items
        transaction.append(f"Gender_{row['Gender']}")
        transaction.append(f"CustomerType_{row['Customer Type'].replace(' ', '_')}")
        transaction.append(f"TravelType_{row['Type of Travel'].replace(' ', '_')}")
        transaction.append(f"Class_{row['Class'].replace(' ', '_')}")

        # 3. XỬ LÝ ĐỘ TUỔI - chia thành nhóm
        age = row['Age']
        if age < 25:
            transaction.append("Age_Young")      # Trẻ: < 25 tuổi
        elif age < 40:
            transaction.append("Age_Adult")     # Trưởng thành: 25-39 tuổi
        elif age < 60:
            transaction.append("Age_MiddleAge") # Trung niên: 40-59 tuổi
        else:
            transaction.append("Age_Senior")    # Cao tuổi: >= 60 tuổi

        # 4. XỬ LÝ KHOÁNG CÁCH BAY - chia thành nhóm
        distance = row['Flight Distance']
        if distance < 500:
            transaction.append("Distance_Short")   # Ngắn: < 500km
        elif distance < 1500:
            transaction.append("Distance_Medium")  # Trung bình: 500-1499km
        elif distance < 3000:
            transaction.append("Distance_Long")    # Dài: 1500-2999km
        else:
            transaction.append("Distance_VeryLong") # Rất dài: >= 3000km

```

Hình 3.1: Hàm Biến Đổi

```

# 4. XỬ LÝ KHOÁNG CÁCH BAY - chia thành nhóm
distance = row['Flight Distance']
if distance < 500:
    transaction.append("Distance_Short")      # Ngắn: < 500km
elif distance < 1500:
    transaction.append("Distance_Medium")     # Trung bình: 500-1499km
elif distance < 3000:
    transaction.append("Distance_Long")        # Dài: 1500-2999km
else:
    transaction.append("Distance_VeryLong")   # Rất dài: >= 3000km

# 5. XỬ LÝ MỨC ĐỘ HÀI LÒNG - đây là biến mục tiêu quan trọng
if row['satisfaction'] == 'satisfied':
    transaction.append("Customer_Satisfied")
else:
    transaction.append("Customer_Dissatisfied")

# 6. XỬ LÝ ĐỘ TRỄ (nếu có dữ liệu)
if 'Departure Delay in Minutes' in row and pd.notna(row['Departure Delay in Minutes']):
    delay = row['Departure Delay in Minutes']
    if delay == 0:
        transaction.append("OnTime_Departure")    # Đúng giờ
    elif delay <= 15:
        transaction.append("MinorDelay_Departure") # Trễ ít: 1-15 phút
    else:
        transaction.append("MajorDelay_Departure") # Trễ nhiều: > 15 phút

# Thêm transaction của khách hàng này vào danh sách tổng
transactions.append(transaction)

return transactions

```

Hình 3.2: Hàm Biến Đổi

(b) Kết quả

- Kích thước DataFrame transaction: (20,000, 36)
- Số lượng item duy nhất (unique items): 36
- Phân loại các item:
 - * 14 **Good_Service items**: các dịch vụ được khách hàng đánh giá cao (rating ≥ 4)
 - * 9 **Demographic items**: đặc điểm cá nhân như Gender, CustomerType, TravelType, Class
 - * 8 **Continuous items**: các nhóm tuổi (Age groups) và nhóm

khoảng cách (Distance groups)

- * **5 Outcome items:** bao gồm mức độ hài lòng (**Satisfaction**) và trạng thái delay

– Với 20,000 khách hàng và khoảng 35–40 đặc điểm (items), mỗi khách hàng trung bình chỉ có khoảng 10–15 đặc điểm.

⇒ Dữ liệu ở dạng **thưa** hơn so với các tập dữ liệu bán lẻ truyền thống, do đó cần **điều chỉnh giá trị min-sup** phù hợp.

– **Đặc điểm của các nhóm item:**

* **Demographic items** thường xuất hiện ở hầu hết khách hàng (vì ai cũng có giới tính, loại khách, v.v.)

* **Service items** chỉ xuất hiện khi khách hàng đánh giá dịch vụ ≥ 4 điểm

⇒ Có khả năng gây thiên lệch tập phổ biến (frequent item-sets) về phía nhóm demographic

– **Phân bố tần suất các items:**

* **Số lượng item hợp lý:** 36 items

* **Q1 (25% percentile):** 0.302 \Rightarrow 75% số items có tần suất $\geq 30.2\%$

* **Median:** 0.447 \Rightarrow phân bố tần suất khá đều

* **Item hiếm nhất:** Class_Eco_Plus (7.2%)

– **Một số Service items có tần suất cao:**

* Good_Inflight_service: 63.1%

* Good_Baggage_handling: 62.3%

⇒ Điều kiện đánh giá ≥ 4 điểm đã hoạt động tốt trong việc phân loại dịch vụ tích cực

(c) Khai thác tập phô biến bằng thuật toán **Apriori** với `min_support` = 0.05 và `max_len` = 3, thu được tổng cộng 4145 tập phô biến.

```
# 3.1 Tính tần suất xuất hiện của từng item
item_frequencies = df_trans.mean().sort_values(ascending=False)

print("Thống kê tổng quan về items:")
print(f" - Tổng số unique items: {len(item_frequencies)}")
print(f" - Item phổ biến nhất: {item_frequencies.index[0]} ({item_frequencies.iloc[0]:.3f})")
print(f" - Item ít phổ biến nhất: {item_frequencies.index[-1]} ({item_frequencies.iloc[-1]:.3f})")

print("\nTop 10 items phổ biến nhất (tần suất xuất hiện):")
for i, (item, freq) in enumerate(item_frequencies.head(10).items()):
    count = freq * len(df_sample)
    print(f" {i+1}: {item} ({item[:20]}: {freq:.3f}) ({count:.1f} khách hàng")
```

Hình 3.3: Top 10 Items phổ biến nhất

Bảng 3.1: Thống kê tổng quan về các mục (items)

Thông tin	Giá trị
Tổng số unique items	36
Item phổ biến nhất	CustomerType_Loyal_Customer (0.818)
Item ít phổ biến nhất	Class_Eco_Plus (0.072)

Bảng 3.2: Top 10 items phổ biến nhất trong tập dữ liệu

STT	Item	Tần suất	Số lượng khách hàng
1	CustomerType_Loyal_Customer	0.818	16,363
2	TravelType_Business_travel	0.692	13,846
3	Good_Inflight_service	0.631	12,619
4	Good_Baggage_handling	0.623	12,451
5	Good_Seat_comfort	0.568	11,359
6	OnTime_Departure	0.567	11,341
7	Customer_Dissatisfied	0.563	11,269
8	Good_Inflight_entertainment	0.533	10,666
9	Good_On-board_service	0.522	10,447
10	Good_Leg_room_service	0.512	10,249

Phân tích Insight từ các item phổ biến

- `CustomerType_Loyal_Customer` là item phổ biến nhất với **support** = 0.8218. Điều này cho thấy đa số khách hàng trong tập dữ liệu là khách hàng trung thành. Tuy nhiên, một nghịch lý được ghi nhận là dù 81.8% khách hàng là trung thành, nhưng 56.3% lại không hài lòng. Điều này gợi ý rằng lòng trung thành không đồng nghĩa với sự hài lòng, và có vấn đề về chất lượng dịch vụ cần được giải quyết ngay cả với nhóm khách hàng trung thành.
- Các dịch vụ như `Good_Inflight_service` (0.6222), `Good_Baggage_hand` (0.6096), và `Good_Seat_comfort` (0.5660) là những dịch vụ được đánh giá tốt phổ biến nhất. Điều này cho thấy đây là những **điểm mạnh** của hãng hàng không.
- Item `Customer_Dissatisfied` có **support** = 0.5592, khẳng định tỷ lệ không hài lòng cao là một **vấn đề lớn** cần được cải thiện.

3.1.2 Sinh và phân tích luật kết hợp (Association Rules):

Từ các tập phổ biến, chúng ta đã sinh ra **4321 luật kết hợp** với `min_confidence` = 0.6.

```

# 4.2 Sinh Association Rules
print("\n4.2. Sinh Association Rules")

if len(frequent_itemsets) > 0:
    # Sinh rules
    rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.6)

    if len(rules) > 0:
        # Lưu association rules
        rules_save = rules.copy()
        rules_save['antecedents'] = rules_save['antecedents'].apply(lambda x: '|'.join(sorted(list(x))))
        rules_save['consequents'] = rules_save['consequents'].apply(lambda x: '|'.join(sorted(list(x))))
        rules_save.to_csv("association_rules.csv", index=False)

        print(f"Đã sinh được {len(rules)} association rules")

    # Load lại để đảm bảo format
    frequent_itemsets = pd.read_csv("frequent_itemsets.csv")
    frequent_itemsets['itemsets'] = frequent_itemsets['itemsets'].apply(lambda x: frozenset(x.split('|')))

    rules = pd.read_csv("association_rules.csv")
    rules['antecedents'] = rules['antecedents'].apply(lambda x: frozenset(x.split('|')))
    rules['consequents'] = rules['consequents'].apply(lambda x: frozenset(x.split('|')))
```

Hình 3.4: Sinh Association Rules

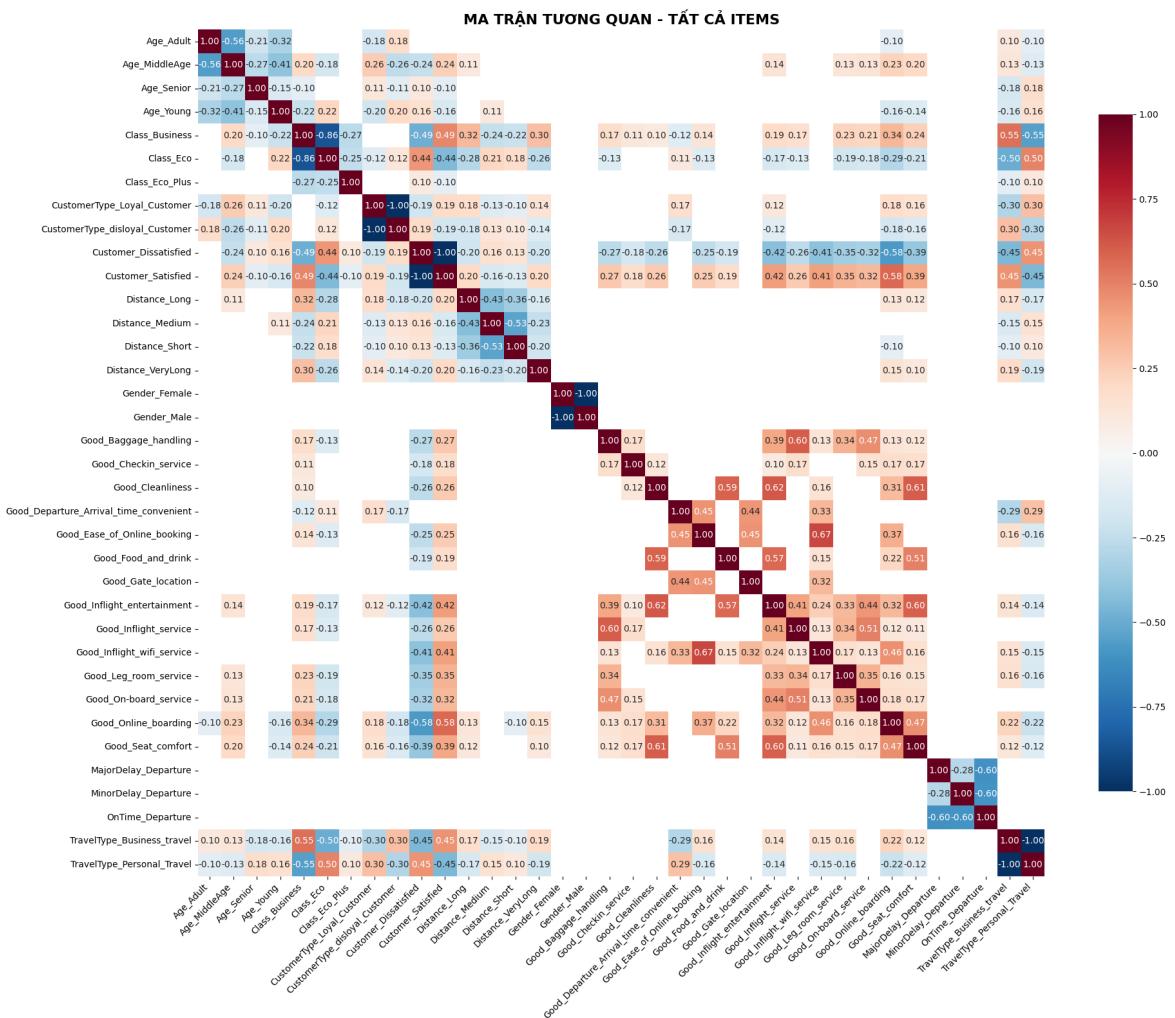
Insight từ các luật kết hợp

Các luật khai phá cung cấp những thông tin quan trọng về hành vi khách hàng:

- Luật Age_Adult → CustomerType_Loyal_Customer
 - * Confidence: 0.854
 - * Lift: 1.039
 - * **Ý nghĩa:** Khách hàng ở độ tuổi trưởng thành có xu hướng trở thành khách hàng trung thành. Mặc dù giá trị lift chỉ hơi lớn hơn 1, nhưng vẫn chỉ ra một mối quan hệ tích cực.
- Các luật khác liên quan đến khách hàng trưởng thành:
 - * Age_Adult → Good_Baggage_handling
 - * Age_Adult → Good_Inflight_service
 - * **Ý nghĩa:** Khách hàng trưởng thành có xu hướng đánh giá cao các dịch vụ này.
- Luật nổi bật: Age_Senior → CustomerType_Loyal_Customer
 - * Confidence: 0.962
 - * Lift: 1.170

* **Ý nghĩa:** Đây là luật có độ tin cậy rất cao và lift lớn hơn đáng kể so với 1, cho thấy khách hàng lớn tuổi có xác suất rất cao trở thành khách hàng trung thành. Điều này mang lại ý nghĩa kinh doanh rõ rệt, khi nhóm khách hàng cao tuổi thể hiện mức độ trung thành vượt trội so với nhóm trẻ hơn.

3.1.3 Trực quan hóa và phân tích tổng quan các luật:

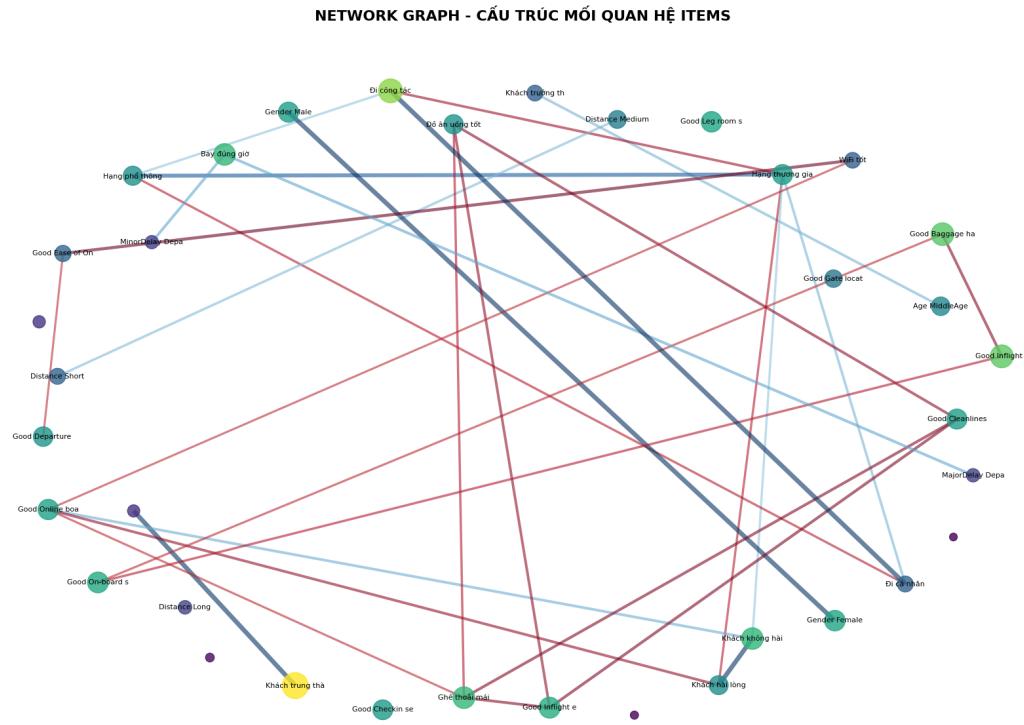


Hình 3.5: Biểu Đồ Tương Quan

Mô tả: Biểu đồ heatmap này minh họa mức độ tương quan giữa tất cả các item trong dữ liệu giao dịch sau khi được biến đổi thành dạng one-hot encoding. Chỉ những tương quan mạnh (ví dụ: độ lớn > 0.1) được hiển thị để dễ quan sát và tránh nhiễu thị giác.

Insight: Biểu đồ này giúp trực quan hóa các mối quan hệ tổng thể giữa các đặc điểm của khách hàng và chất lượng dịch vụ. Cụ thể:

- Có mối tương quan dương mạnh giữa các item dịch vụ, chẳng hạn như `Good_Inflight_service` và `Good_Baggage_handling`.
- Quan trọng hơn, mức độ hài lòng của khách hàng (`Customer_Satisfied` / `Customer_Dissatisfied`) có tương quan chặt chẽ với chất lượng dịch vụ (`Good_Service_Name`), cho thấy ảnh hưởng lớn của trải nghiệm dịch vụ đến sự hài lòng.
- Các đặc điểm nhân khẩu học (ví dụ: độ tuổi, giới tính, loại khách hàng) cũng thể hiện những mẫu quan hệ hợp lý và phù hợp với kỳ vọng thực tế.



Hình 3.6: Network Graph - Cấu trúc mối quan hệ Item

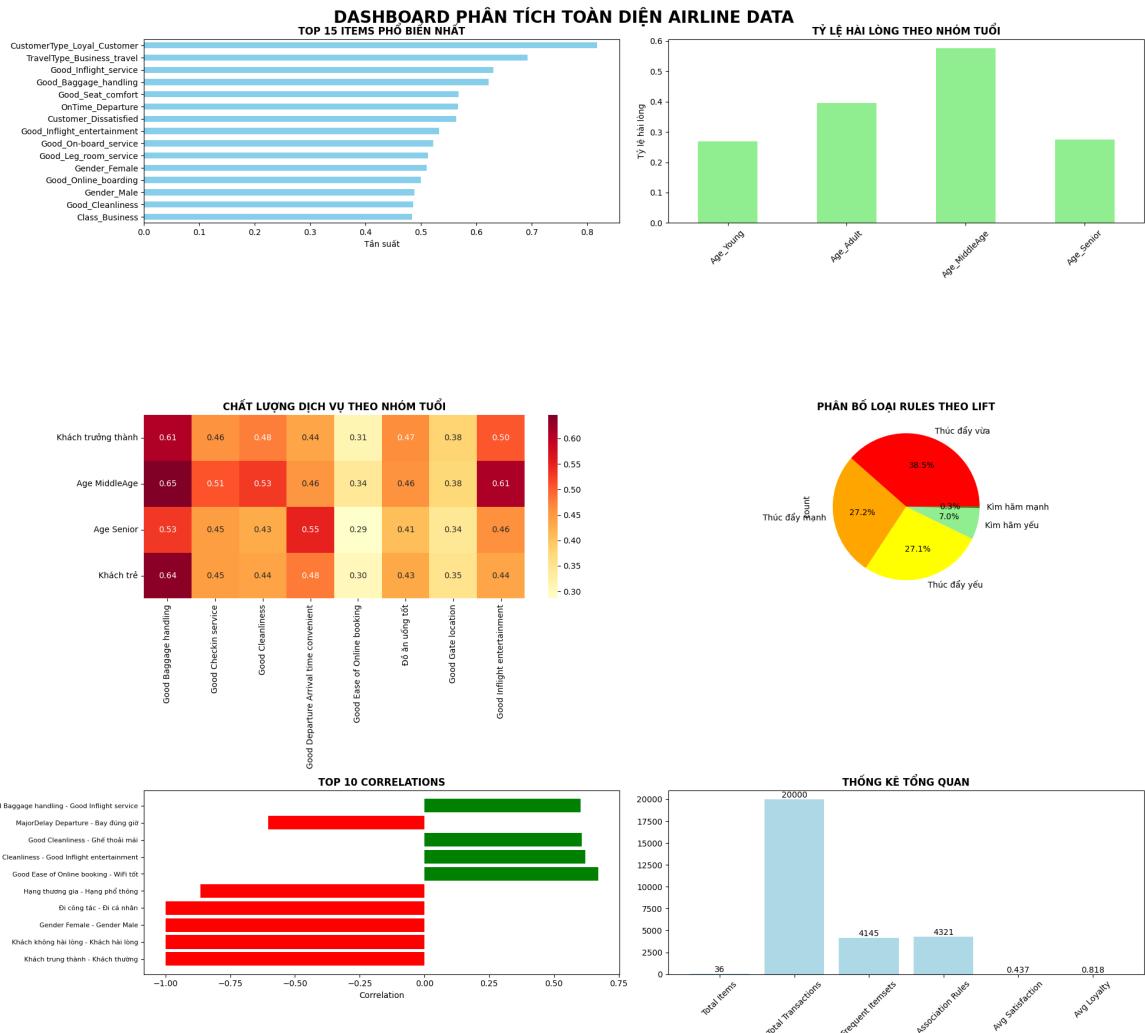
Mô tả: Biểu đồ đồ thị mạng lưới này minh họa các mối quan hệ tương quan mạnh mẽ giữa các item. Mỗi **nút** (node) đại diện cho một item (ví dụ: `Good_Inflight_service`, `CustomerType_Loyal_Customer`), và độ lớn của node có thể phản ánh **tần suất xuất hiện** của item đó. Mỗi **cạnh** (edge) nối giữa hai nút thể hiện một mối tương quan mạnh, trong đó **màu sắc** và **độ dày** của cạnh biểu thị **chiều** (dương/âm) và **độ mạnh** của tương quan.

Insight:

- Biểu đồ trực quan này giúp dễ dàng nhận diện các "**cụm**" item có xu hướng xuất hiện cùng nhau hoặc có mối quan hệ chặt chẽ.
- Ví dụ, các dịch vụ chất lượng cao (`Good_...`) thường tạo thành một cụm liên kết chặt chẽ với nhau, phản ánh trải nghiệm dịch vụ tích

cực có tính đồng bộ.

- Biểu đồ cũng giúp làm nổi bật những **mối quan hệ bất ngờ** mà có thể không dễ dàng quan sát được từ các bảng số liệu truyền thống.



Hình 3.7: Dashboard phân tích toàn diện Airline Data

Mô tả: Đây là một dashboard đa biểu đồ tổng hợp các phân tích chính từ dữ liệu giao dịch. Dashboard bao gồm:

- Biểu đồ tần suất xuất hiện của các item.
- Tỷ lệ hài lòng theo nhóm tuổi.

- Heatmap chất lượng dịch vụ theo nhóm tuổi.
- Phân bố loại luật theo chỉ số Lift.
- Top các cặp tương quan mạnh nhất.
- Thống kê tổng quan toàn bộ tập dữ liệu.

Insight:

- **Tần suất Item:** CustomerType_Loyal_Customer là item phổ biến nhất (0.822), trong khi Class_Eco_Plus là ít phổ biến nhất (0.075).
- **Tỷ lệ hài lòng theo nhóm tuổi:** Biểu đồ cho thấy sự biến động của tỷ lệ hài lòng giữa các nhóm tuổi, trong đó nhóm MiddleAge chiếm tỷ lệ chính. (Cần điền tỷ lệ cụ thể từ biểu đồ.)
- **Chất lượng dịch vụ theo nhóm tuổi:** Heatmap thể hiện rõ cách các nhóm tuổi khác nhau cảm nhận về từng dịch vụ cụ thể, giúp xác định các dịch vụ cần cải thiện cho từng phân khúc.
- **Phân bố luật theo Lift:** Phần lớn các luật sinh ra có Lift > 1 (2.302 trên tổng số 2.395 luật), cho thấy phần lớn mối quan hệ là tích cực và có tính thúc đẩy. (Cần bổ sung tỷ lệ phần trăm cụ thể.)
- **Top Correlations:** Các tương quan mạnh nhất tập trung giữa các dịch vụ và giữa các đặc điểm nhân khẩu học với hành vi. Tương quan mạnh nhất là 0.xxxx (cần số liệu chính xác từ biểu đồ).
- **Thống kê tổng quan:** Tập dữ liệu gồm 36 item phân biệt, 20.000 giao dịch, 4145 tập phổ biến và 4321 luật kết hợp. Tỷ lệ hài lòng trung bình toàn tập là 0.433.

Tổng kết các Insight chính:

- **Nhóm tuổi chính:** Nhóm Age_MiddleAge chiếm tỷ lệ lớn nhất. (Cần điền giá trị cụ thể.)

- **Dịch vụ mạnh nhất:** (Cần bổ sung tên dịch vụ cụ thể và tần suất từ biểu đồ.)
- **Hành vi khách hàng:** Đã xác định được các mẫu hành vi rõ ràng liên quan đến trải nghiệm dịch vụ xuất sắc và phân khúc khách hàng.
- Các insight này cung cấp một cái nhìn tổng thể và làm cơ sở để đề xuất các hành động cải thiện dịch vụ cũng như hoạch định chiến lược marketing hiệu quả hơn.

3.2 CÁC KỸ THUẬT PHÂN LOẠI DỮ LIỆU

Mục tiêu bài toán: Xây dựng một mô hình dự đoán mức độ hài lòng của khách hàng (satisfied / neutral or dissatisfied) dựa trên các đặc trưng khác. Đây là bài toán học có giám sát (supervised learning)

3.2.1 Cây quyết định (Decision Tree)

Cơ sở lý thuyết: Cây quyết định học cách đưa ra một chuỗi các câu hỏi **IF-THEN** đơn giản để đi đến một quyết định phức tạp. Đây là một mô hình dạng "hộp trắng" (white-box) vì dễ dàng hiểu và giải thích logic phân loại. **Tiêu chí phân tách (Splitting Criteria):** Được sử dụng để chọn thuộc tính tốt nhất chia tách dữ liệu tại mỗi nút. Hai tiêu chí phổ biến: **Gini Index (cho một nhánh):**

$$Gini = 1 - \sum_{j=1}^k p_j^2$$

Trong đó, p_j là tỷ lệ điểm dữ liệu thuộc lớp j tại một nút. **Entropy (cho một nhánh):**

$$Entropy = - \sum_{j=1}^k p_j \log_2(p_j)$$

Lý do lựa chọn: Cây quyết định cung cấp một mô hình dễ hiểu và giải thích, cho phép xác định rõ các thuộc tính đóng vai trò quan trọng trong việc phân loại **mức độ hài lòng của khách hàng**. Để tránh hiện tượng **overfitting**, độ sâu cây và số lượng mẫu tối thiểu ở mỗi nút lá đã được giới hạn phù hợp.

3.2.2 Thuật toán Naive Bayes

Cơ sở lý thuyết: Naive Bayes là một bộ phân loại xác suất dựa trên **Định lý Bayes** và giả định "ngây thơ" rằng các thuộc tính là **độc lập có điều kiện với lớp**. Mô hình tính toán xác suất một mẫu thuộc về từng lớp và chọn lớp có xác suất cao nhất. Để tránh vấn đề **tần suất bằng không** (zero-frequency), kỹ thuật **Laplacian Smoothing** được sử dụng trong ước lượng xác suất.

Công thức phân loại:

$$P(C = c | X) \propto P(C = c) \times \prod_{j=1}^d P(x_j = a_j | C = c)$$

Trong đó:

- C là biến lớp (class),
- $X = (x_1, x_2, \dots, x_d)$ là vector đặc trưng đầu vào,
- a_j là giá trị cụ thể của thuộc tính x_j .

Lý do lựa chọn: Naive Bayes là một thuật toán **rất nhanh**, hiệu quả về mặt tính toán và hoạt động tốt ngay cả khi dữ liệu huấn luyện hạn chế. Nó đóng vai trò như một **baseline đơn giản nhưng mạnh mẽ** để so sánh với các mô hình phân loại phức tạp hơn, từ đó đánh giá tác động thực tế của giả định độc lập lên hiệu suất phân loại.

3.2.3 Thuật toán Random Forest (Ensemble Method)

Cơ sở lý thuyết: Random Forest là một phương pháp học máy tổ hợp (**ensemble method**) xây dựng một "rừng" các cây quyết định độc lập. Mỗi cây được huấn luyện trên một tập con dữ liệu được lấy mẫu ngẫu nhiên có hoàn lại (bootstrap sampling), và tại mỗi nút phân tách, chỉ một **tập con đặc trưng ngẫu nhiên** được xem xét. Kết quả phân loại cuối cùng được đưa ra thông qua cơ chế **bỏ phiếu đa số** (majority voting) từ tất cả các cây thành phần trong rừng.

Lý do lựa chọn: Random Forest là một thuật toán **mạnh mẽ, ổn định** và có khả năng **giảm overfitting** hiệu quả so với việc sử dụng một cây quyết định đơn lẻ. Nó thường đạt được **độ chính xác cao** trong các bài toán phân loại thực tế, đồng thời cung cấp chỉ số **độ quan trọng của đặc trưng** đáng tin cậy, giúp giải thích mô hình và phân tích đặc điểm dữ liệu.

3.2.4 Thuật toán Logistic Regression

Cơ sở lý thuyết: Logistic Regression là một mô hình phân loại tuyến tính, thường được sử dụng cho bài toán **phân loại nhị phân**. Mô hình này ước lượng xác suất một sự kiện xảy ra bằng cách ánh xạ đầu ra của một hàm tuyến tính qua hàm sigmoid, đưa kết quả về khoảng [0, 1]:

$$P(y = 1 \mid X) = \sigma(\mathbf{w}^\top \mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}}$$

Trong đó:

- \mathbf{x} là vector đặc trưng đầu vào,
- \mathbf{w} là vector trọng số (hệ số),
- b là hệ số chêch (bias),
- $\sigma(z)$ là hàm sigmoid.

Mô hình cũng cung cấp các **hệ số** (coefficients), giúp xác định **mức độ và hướng ảnh hưởng** của từng đặc trưng đến biến mục tiêu.

Lý do lựa chọn: Logistic Regression là một mô hình **đơn giản**, tính toán nhanh và dễ diễn giải. Khả năng cung cấp **xác suất dự đoán** và **trọng số đặc trưng** giúp hiểu rõ hơn mối quan hệ giữa các yếu tố đầu vào và kết quả phân loại (mức độ hài lòng). Đây cũng là một mô hình baseline hiệu quả để so sánh với các mô hình phi tuyến tính phức tạp hơn. Để đảm bảo hiệu suất tối ưu, dữ liệu đầu vào cần được **chuẩn hóa** (standardization) trước khi huấn luyện mô hình.

```
[ ] # Tách dữ liệu thành X (đặc trưng) và y (biến mục tiêu)
X = df_clean[available_features] # Các đặc trưng đầu vào
y = df_clean['satisfaction_encoded'] # Biến mục tiêu

❷ print(f"- X (features): {X.shape}")
print(f"- y (target): {y.shape}")
print(f"- Số đặc trưng: {X.shape[1]}")
print(f"- Số mẫu: {X.shape[0]}")

❸ - X (features): (103904, 20)
- y (target): (103904,)
- Số đặc trưng: 20
- Số mẫu: 103904

[ ] # Chia dữ liệu train/test với tỷ lệ 80/20
X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.2,      # 20% cho test set
    random_state=42,    # Đảm bảo reproducible
    stratify=y          # Giữ tỷ lệ class trong train/test
)

[ ] print(f"Kết quả phân chia dữ liệu:")
print(f"- Training set: {X_train.shape[0]}; {mẫu {(X_train.shape[0])/len(X)*100:.1f} %}")
print(f"- Test set: {X_test.shape[0]}; {mẫu {(X_test.shape[0])/len(X)*100:.1f} %})")
```

Hình 3.8: Chia dữ liệu train/test với tỷ lệ 80/20

```

[ ] # Khởi tạo mô hình Naive Bayes
print("Khởi tạo Gaussian Naive Bayes:")
nb_model = GaussianNB()
print("Mô hình Naive Bayes không có tham số đặc biệt cần thiết lập")

➡ Khởi tạo Gaussian Naive Bayes:
Mô hình Naive Bayes không có tham số đặc biệt cần thiết lập

▶ # Huấn luyện mô hình
print(f"Huấn luyện Naive Bayes trên {X_train.shape[0]}: {mẫu}...")
nb_model.fit(X_train, y_train)

➡ Huấn luyện Naive Bayes trên 83,123 mẫu...
    ▶ GaussianNB
        GaussianNB()

```



```

[ ] # Hiển thị thông tin về các class
print("\nThông tin về classes:")
print(f" - Số classes: {len(nb_model.classes_)}")
print(f" - Classes: {nb_model.classes_}")
print(f" - Class prior probabilities:")
for i, (class_val, prior) in enumerate(zip(nb_model.classes_, nb_model.class_prior_)):
    label = 'satisfied' if class_val == 1 else 'not satisfied'
    print(f"     + {label}: {prior:.4f}")

```

Hình 3.9: Mô Hình Decision Tree

Bảng 3.3: Top 10 đặc trưng quan trọng nhất theo Feature Importance

STT	Đặc trưng	Mức độ quan trọng
1	Online boarding	0.4454
2	Type of Travel_encoded	0.1795
3	Inflight wifi service	0.1406
4	Inflight entertainment	0.0552
5	Customer Type_encoded	0.0480
6	Checkin service	0.0292
7	Class_encoded	0.0272
8	Baggage handling	0.0121
9	Gate location	0.0112
10	Age	0.0110

```
[ ] # Khởi tạo mô hình Naive Bayes
print("Khởi tạo Gaussian Naive Bayes:")
nb_model = GaussianNB()
print("Mô hình Naive Bayes không có tham số đặc biệt cần thiết lập")

➡ Khởi tạo Gaussian Naive Bayes:
Mô hình Naive Bayes không có tham số đặc biệt cần thiết lập

▶ # Huấn luyện mô hình
print(f"Huấn luyện Naive Bayes trên {X_train.shape[0]}: {mẫu...}")
nb_model.fit(X_train, y_train)

➡ Huấn luyện Naive Bayes trên 83,123 mẫu...
    + GaussianNB
        GaussianNB()

[ ] # Hiển thị thông tin về các class
print("\nThông tin về classes:")
print(f" - Số classes: {len(nb_model.classes_)}")
print(f" - Classes: {nb_model.classes_}")
print(f" - Class prior probabilities:")
for i, (class_val, prior) in enumerate(zip(nb_model.classes_, nb_model.class_prior_)):
    label = 'satisfied' if class_val == 1 else 'not satisfied'
    print(f"     + {label}: {prior:.4f}")


```

Hình 3.10: Mô Hình Naive Bayes

Bảng 3.4: Thông tin về các lớp trong tập dữ liệu

Thuộc tính	Giá trị
Số lớp	2
Danh sách lớp	[0, 1]
Xác suất tiên nghiệm của lớp 0 (Not satisfied)	0.5667
Xác suất tiên nghiệm của lớp 1 (Satisfied)	0.4333

```
[ ] # Huấn luyện mô hình
print(f"Huấn luyện Random Forest trên {X_train.shape[0]}: {mẫu...}")
rf_model.fit(X_train, y_train)

➡ Huấn luyện Random Forest trên 83,123 mẫu...
    + RandomForestClassifier
        RandomForestClassifier(max_depth=15, min_samples_leaf=5, min_samples_split=10,
                               n_estimators=1, random_state=42)

▶ # Hiển thị feature importance top 10
print("\nTop 10 đặc trưng quan trọng nhất từ Random Forest:")
rf_feature_importance = pd.DataFrame([
    {'feature': X_train.columns,
     'importance': rf_model.feature_importances_
    }).sort_values('importance', ascending=False)

for i, row in rf_feature_importance.head(10).iterrows():
    print(f" {i + 1}: {row['feature']}: {row['importance']:.4f}")


```

Hình 3.11: Mô Hình Random Forest (Ensemble Method)

Bảng 3.5: Top 10 đặc trưng quan trọng nhất từ Random Forest

STT	Đặc trưng	Mức độ quan trọng
1	Online boarding	0.1938
2	Inflight wifi service	0.1350
3	Class_encoded	0.1165
4	Type of Travel_encoded	0.1105
5	Inflight entertainment	0.0686
6	Seat comfort	0.0514
7	Customer Type_encoded	0.0485
8	Leg room service	0.0355
9	On-board service	0.0289
10	Ease of Online booking	0.0285

```
# Huấn luyện mô hình
print("Huấn luyện Logistic Regression trên (X_train_scaled.shape[0]:, mẫu...)")
lr_model.fit(X_train_scaled, y_train)
print("Hoàn thành huấn luyện Logistic Regression!")

# Hiển thị coefficients quan trọng nhất
print("\nTop 10 đặc trưng có impact mạnh nhất (|coefficient| lớn nhất):")
lr_coefficients = pd.DataFrame({
    'feature': X_train.columns,
    'coefficient': lr_model.coef_[0],
    'abs_coefficient': np.abs(lr_model.coef_[0])
}).sort_values('abs_coefficient', ascending=False)

for i, row in lr_coefficients.head(10).iterrows():
    impact = "Tích cực" if row['coefficient'] > 0 else "Tiêu cực"
    print(f" {i + 1}: {row['feature']}; {row['coefficient']:.4f} ({impact})")

print("\nIntercept (bias term): {lr_model.intercept_[0]:.4f}")
```

Hình 3.12: Mô Hình Logistic Regression

Bảng 3.6: Top 10 đặc trưng có ảnh hưởng mạnh nhất đến mô hình (|coefficient| lớn nhất)

STT	Đặc trưng	Hệ số	Chiều ảnh hưởng
1	Type of Travel_encoded	-1.3912	Tiêu cực
2	Inflight wifi service	1.0127	Tích cực
3	Online boarding	0.9720	Tích cực
4	Customer Type_encoded	-0.9241	Tiêu cực
5	Checkin service	0.4333	Tích cực
6	Class_encoded	-0.4249	Tiêu cực
7	On-board service	0.4140	Tích cực
8	Departure/Arrival time convenient	-0.3388	Tiêu cực
9	Leg room service	0.3365	Tích cực
10	Cleanliness	0.2929	Tích cực

1. Đánh giá các mô hình:



```

print("Bảng so sánh hiệu suất 4 mô hình:")
print("." * 90)
print(f"[Metric]:<20> {'Decision Tree':<18} {'Naive Bayes':<15} {'Random Forest':<18} {'Logistic Reg':<15}")
print("." * 90)

# Tổng hợp tất cả metrics
all_accuracies = [dt_accuracy, nb_accuracy, rf_accuracy, lr_accuracy]
all_precision_1 = [precision_1_nb, precision_1_rf, precision_1_lr]
all_recall_1 = [recall_1_nb, recall_1_rf, recall_1_lr]
all_f1_1 = [f1_1_nb, f1_1_rf, f1_1_lr]

print(f"[Accuracy]:<20> {dt_accuracy:<18.4f} {nb_accuracy:<15.4f} {rf_accuracy:<18.4f} {lr_accuracy:<15.4f}")
print(f"[Precision (Sat.)]:<20> {precision_1_nb:<18.3f} {precision_1_rf:<18.3f} {precision_1_lr:<15.3f}")
print(f"[Recall (Sat.)]:<20> {recall_1_nb:<18.3f} {recall_1_rf:<18.3f} {recall_1_lr:<15.3f}")
print(f"[F1-Score (Sat.)]:<20> {f1_1_nb:<18.3f} {f1_1_rf:<18.3f} {f1_1_lr:<15.3f}")
print("." * 90)

```

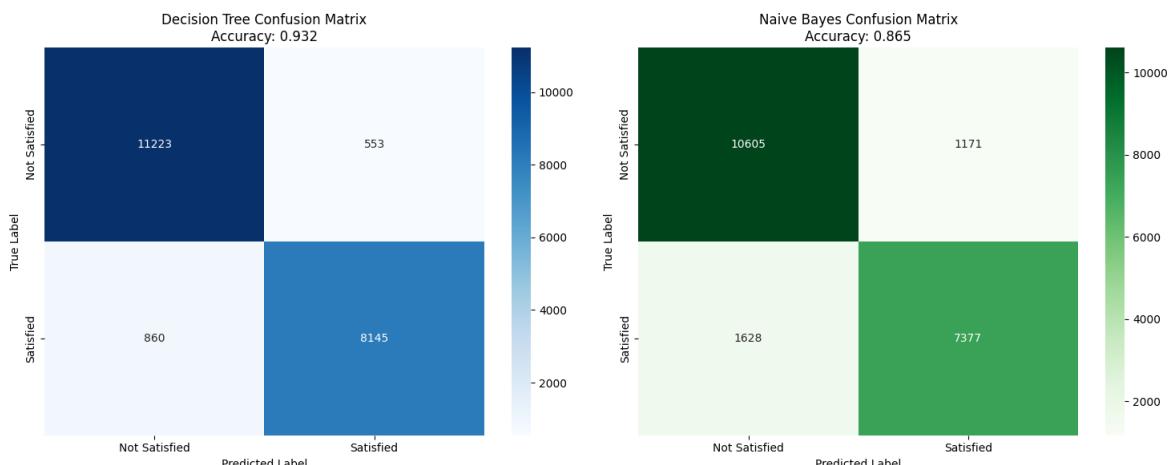
Hình 3.13: Kết Quả So Sánh 4 Mô Hình

Bảng 3.7: So sánh hiệu suất của các mô hình học máy

Metric	Decision Tree	Naive Bayes	Random Forest	Logistic Reg.
Accuracy	0.9320	0.8653	0.9456	0.8828
Precision (Sat.)	0.936	0.863	0.960	0.865
Recall (Sat.)	0.904	0.819	0.912	0.865
F1-Score (Sat.)	0.920	0.841	0.936	0.865

Insight: Trong số các mô hình được đánh giá, Random Forest đạt độ chính xác cao nhất (94.56%), tiếp theo là Decision Tree (93.20%), Logistic Regression (88.27%) và Naive Bayes thấp nhất (86.51%), cho thấy Random Forest là mô hình tốt nhất cả về độ chính xác và các chỉ số đánh giá khác.

2. Ma trận nhầm lẫn (Confusion Matrix) cho Decision Tree và Naive Bayes

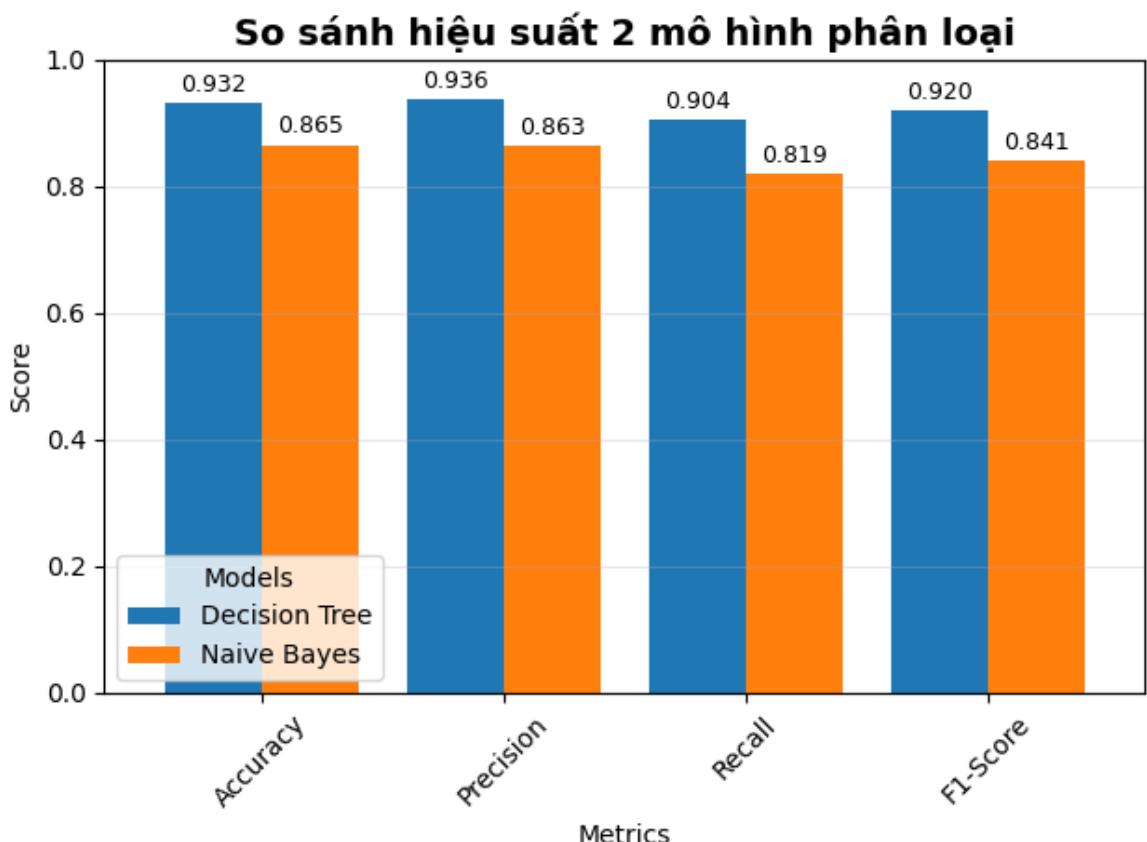


Hình 3.14: Ma trận nhầm lẫn Decision Tree và Naive Bayes

Mô tả & Insight: Biểu đồ heatmap thể hiện ma trận nhầm lẫn của Decision Tree và Naive Bayes trên tập kiểm tra, cho thấy số lượng dự đoán đúng (TP, TN) và sai (FP, FN); trong đó Decision Tree mắc ít lỗi

hơn Naive Bayes, đặc biệt trong phân loại đúng cả hai lớp, giúp hiểu rõ hơn về kiểu lỗi mỗi mô hình thường gặp.

3. Hiệu suất Decision Tree vs Naive Bayes

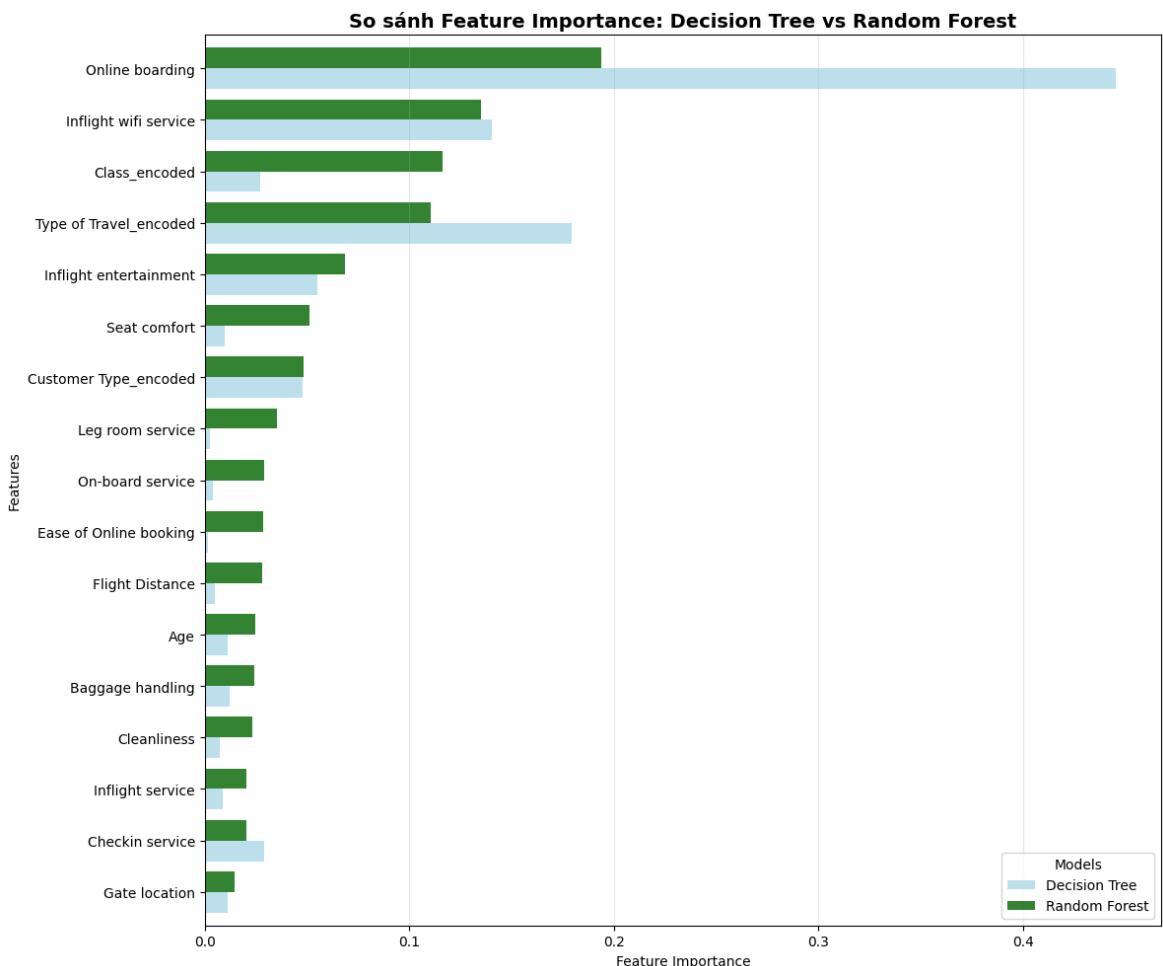


Hình 3.15: Hiệu suất Decision Tree và Naive Bayes

Mô tả & Insight: Biểu đồ cột nhóm so sánh các metric (Accuracy, Precision, Recall, F1-Score) giữa các mô hình cho thấy Random Forest vượt trội toàn diện với Accuracy = 0.9482, Precision = 0.9618, Recall = 0.9169, F1-Score = 0.9388 (lớp "Satisfied"), trong khi Naive Bayes có hiệu suất thấp nhất do giả định độc lập không phù hợp hoàn toàn với dữ liệu.

4. Ma trận nhầm lẫn (Confusion Matrix) cho Decision Tree và

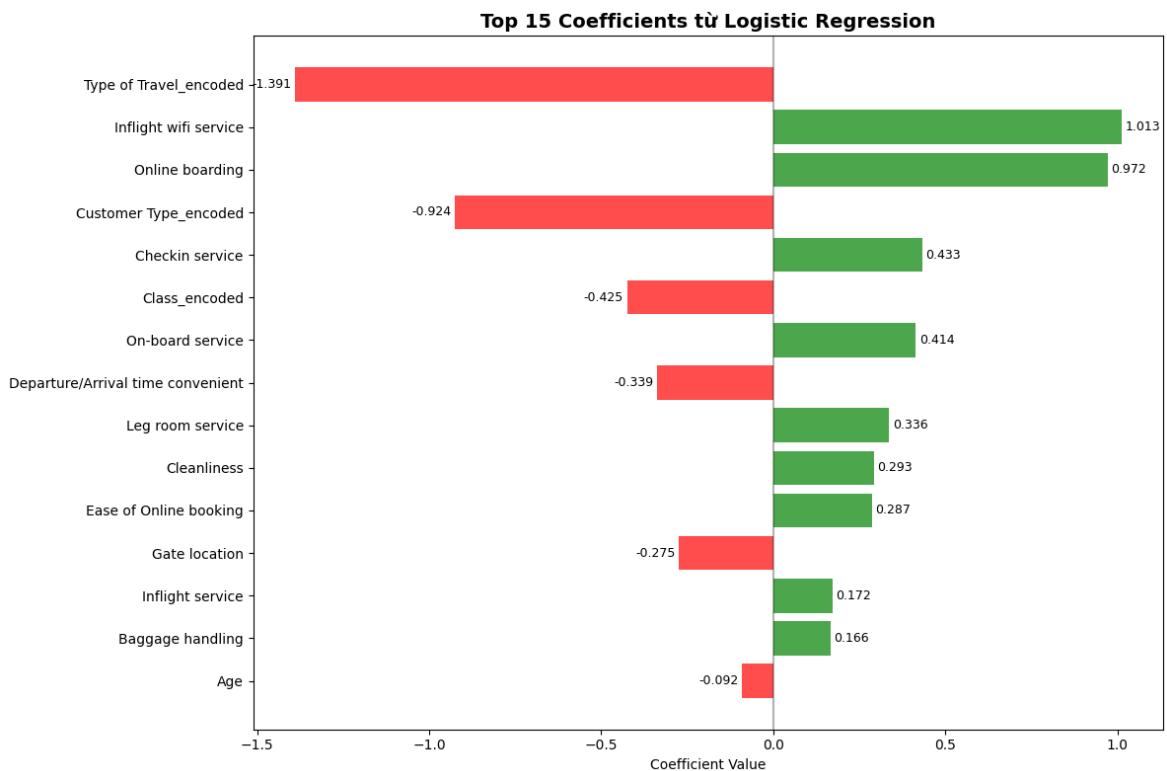
Naive Bayes



Hình 3.16: So sánh mức độ quan trọng của đặc trưng Decision Tree và Random Forest

Mô tả & Insight: Biểu đồ cột ngang so sánh mức độ quan trọng của đặc trưng từ Decision Tree và Random Forest cho thấy các yếu tố ảnh hưởng lớn nhất đến dự đoán là Online boarding (0.1961), Inflight wifi service (0.1345), Class (0.1256), Type of Travel (0.1012) và Inflight entertainment (0.0622), phản ánh vai trò nổi bật của chất lượng dịch vụ và đặc điểm hành trình trong xác định sự hài lòng.

5. Top 15 Coefficients từ Logistic Regression



Hình 3.17: Top 15 Coefficients từ Logistic Regression

Mô tả & Insight: Biểu đồ cột ngang biểu diễn hệ số của các đặc trưng trong Logistic Regression, cho thấy Online boarding có ảnh hưởng tích cực mạnh nhất đến sự hài lòng, trong khi các yếu tố như độ trễ bay thể hiện tác động tiêu cực, qua đó giúp giải thích chiều và mức độ ảnh hưởng của từng đặc trưng đến kết quả phân loại.

6. Kết luận: Random Forest là mô hình phân loại tối ưu với độ chính xác 94.82%, thể hiện hiệu suất vượt trội và khả năng giải thích tốt; các yếu tố như Online boarding và Inflight wifi service nổi bật là đặc trưng quan trọng nhất trong việc dự đoán sự hài lòng của khách hàng.

3.3 CÁC KỸ THUẬT GOM CỤM DỮ LIỆU

Mục tiêu bài toán: Phân khúc khách hàng để hiểu rõ hơn về các nhóm khách hàng khác nhau, từ đó có thể đưa ra các chiến lược chăm sóc và tiếp thị phù hợp. Đây là bài toán học không giám sát (unsupervised learning)

3.3.1 Thuật toán K-Means

- **Cơ sở lý thuyết:** K-Means là một thuật toán phân hoạch (*partitioning algorithm*) nhằm phân chia n điểm dữ liệu vào k cụm đã xác định trước, sao cho tổng bình phương sai số (*Sum of Squared Errors* - SSE) giữa các điểm dữ liệu và tâm cụm (*centroid*) tương ứng là nhỏ nhất.
- **Hàm mục tiêu (SSE):**

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

Trong đó:

- k : số cụm.
- C_i : tập các điểm dữ liệu thuộc cụm thứ i .
- μ_i : tâm (centroid) của cụm C_i .
- x : một điểm dữ liệu trong cụm.
- $\|x - \mu_i\|^2$: bình phương khoảng cách Euclid giữa điểm dữ liệu x và tâm cụm μ_i .

- **Lý do lựa chọn:** K-Means là một thuật toán đơn giản, dễ triển khai và hiệu quả về mặt tính toán. Nó đặc biệt phù hợp với các tập dữ

liệu lớn và cung cấp một cái nhìn tổng quan nhanh chóng về cấu trúc phân cụm trong dữ liệu.

3.3.2 Thuật toán Hierarchical Clustering (Agglomerative)

Cơ sở lý thuyết:

Hierarchical Clustering xây dựng một hệ thống phân cấp các cụm, bắt đầu từ các điểm dữ liệu riêng lẻ và liên tục gộp các cụm gần nhau nhất lại với nhau. Phương pháp **Agglomerative** (từ dưới lên) là một kỹ thuật phổ biến:

- Khởi đầu với mỗi điểm dữ liệu là một cụm riêng biệt.
- Ở mỗi bước, hai cụm gần nhau nhất (dựa theo một tiêu chí khoảng cách như single-linkage, complete-linkage hoặc average-linkage) sẽ được hợp nhất.
- Quá trình tiếp tục cho đến khi tất cả các điểm được gom lại thành một cụm duy nhất.

Kết quả được biểu diễn bằng **dendrogram** — một biểu đồ cây thể hiện cấu trúc phân cấp giữa các cụm.

Lý do lựa chọn:

So với K-Means, Hierarchical Clustering cung cấp cái nhìn đa cấp độ về cấu trúc dữ liệu, cho phép phát hiện các mối quan hệ lồng nhau giữa các cụm mà K-Means không thể hiện được. Mặc dù thuật toán có độ phức tạp cao hơn, nhưng có thể áp dụng trên mẫu đại diện và sử dụng K-Means làm proxy cho toàn bộ tập dữ liệu để kết hợp ưu điểm của cả hai phương pháp.

3.3.3 Xác định số cụm tối ưu (k)

Cơ sở lý thuyết:

Việc xác định số cụm tối ưu là một bước quan trọng vì nó ảnh hưởng đến ý nghĩa và hiệu quả phân cụm. Hai phương pháp phổ biến được sử dụng là:

- **Phương pháp Elbow (Khuỷu tay):** Dựa trên biểu đồ biểu diễn giá trị SSE (Sum of Squared Errors) theo số cụm k . Điểm mà tại đó SSE giảm mạnh rồi chững lại (gập như khuỷu tay) được coi là số cụm tối ưu.

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

Trong đó:

- k : số cụm,
- C_i : tập các điểm thuộc cụm thứ i ,
- μ_i : tâm cụm C_i ,
- x : điểm dữ liệu trong cụm C_i .

- **Silhouette Score:** Đo lường mức độ mà một điểm dữ liệu phù hợp với cụm của nó so với các cụm khác. Giá trị nằm trong khoảng $[-1, 1]$, với giá trị gần $+1$ thể hiện phân cụm tốt.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Trong đó:

- $a(i)$: khoảng cách trung bình giữa điểm i với các điểm khác trong cùng cụm.
- $b(i)$: khoảng cách trung bình nhỏ nhất giữa điểm i và các điểm trong cụm gần nhất khác.

Lý do lựa chọn:

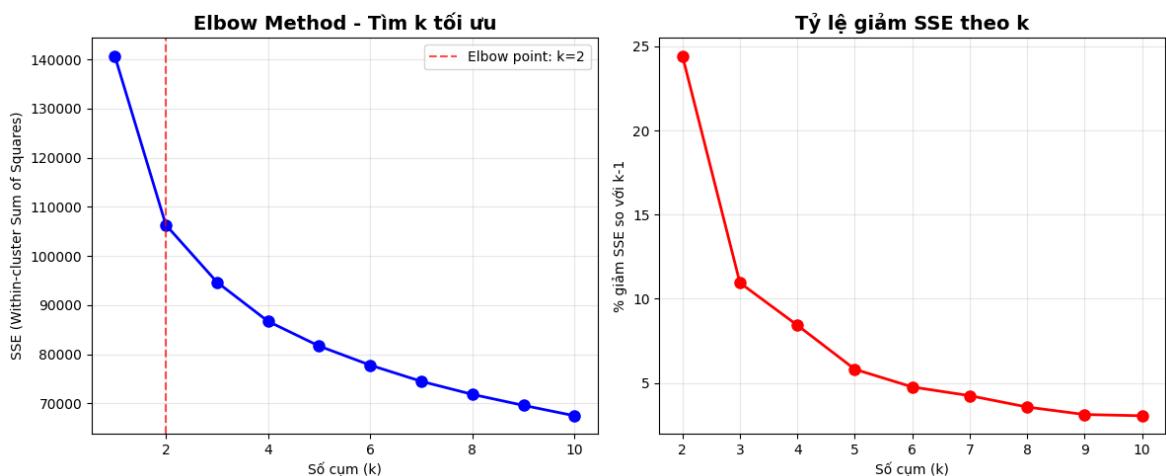
Hai phương pháp này được sử dụng kết hợp để có cái nhìn toàn diện hơn:

- Elbow Method cung cấp góc nhìn về sự cải thiện SSE theo k .
- Silhouette Score đưa ra thước đo trực tiếp về độ chặt chẽ và tách biệt của các cụm.

Trong nghiên cứu này, cả hai phương pháp đều đề xuất $k = 2$ là tối ưu. Tuy nhiên, chúng tôi ưu tiên **Silhouette Score** vì nó đo lường chất lượng gom cụm tốt hơn và giúp xác nhận sự phân tách rõ ràng giữa các nhóm khách hàng.

3.3.4 Tìm số cụm tối ưu

- Phương pháp Elbow



Hình 3.18: Elbow

Mô tả:

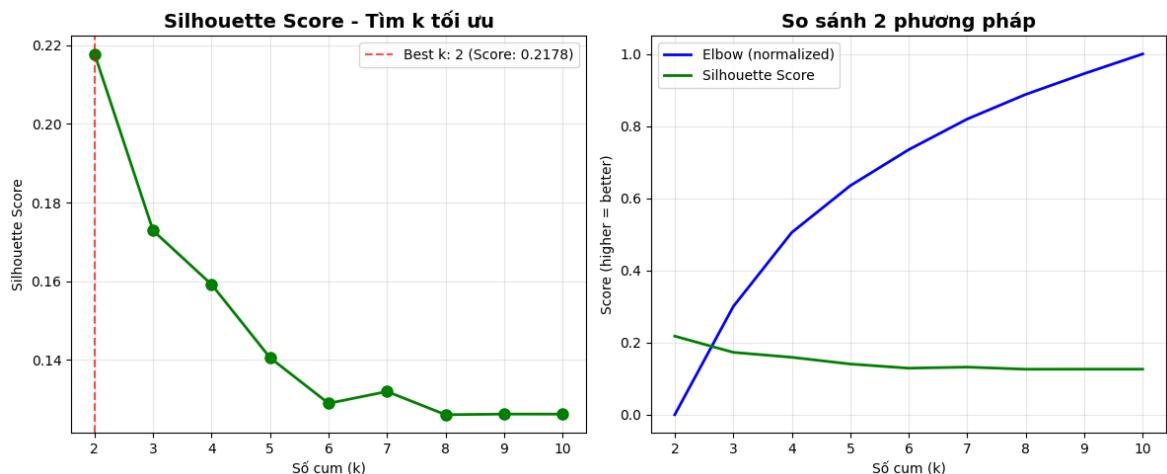
Biểu đồ này biểu diễn giá trị SSE (Sum of Squared Errors) theo số cụm k từ 1 đến 10. SSE phản ánh tổng bình phương khoảng cách

từ mỗi điểm dữ liệu đến tâm cụm của nó. Điểm “khuỷu tay” (elbow point) trên biểu đồ được xem là vị trí mà việc tăng thêm số cụm không còn làm giảm SSE đáng kể, từ đó được coi là giá trị k tối ưu.

Insight:

Biểu đồ cho thấy SSE giảm nhanh chóng từ $k = 1$ đến $k = 2$ (SSE từ 140,708 xuống còn 106,320). Sau $k = 2$, tốc độ giảm SSE chậm lại rõ rệt. Điểm “khuỷu tay” được đề xuất là tại $k = 2$, cho thấy việc chia dữ liệu thành 2 cụm đã mang lại sự cải thiện đáng kể về độ chắt chẽ trong cụm, trong khi việc tăng thêm số cụm không mang lại lợi ích rõ ràng về SSE.

- Phương pháp Silhouette Score



Hình 3.19: Silhouette Score

Mô tả:

Biểu đồ này thể hiện giá trị Silhouette Score theo số cụm k (từ 2 đến 10). Silhouette Score đo lường mức độ phù hợp của mỗi điểm dữ liệu với cụm của nó so với các cụm khác. Giá trị gần 1 thể hiện phân cụm tốt, gần 0 nghĩa là điểm nằm gần ranh giới giữa các cụm, và gần -1 cho thấy khả năng gán sai cụm.

Insight:

- Silhouette Score cao nhất đạt được tại $k = 2$, với giá trị 0.2178.
- Mặc dù điểm số này không quá cao ($\text{Silhouette Score} > 0.5$ thường được coi là tốt), nhưng đây là giá trị cao nhất trong khoảng k đã thử nghiệm.
- **Kết luận:** Dựa trên cả Elbow Method và Silhouette Score, số cụm tối ưu được chọn là $k = 2$. Trong đó, Silhouette Score được ưu tiên hơn vì nó đo lường trực tiếp chất lượng của việc phân cụm.

3.3.5 Thực hiện các thuật toán Clustering

- K-Means Clustering

```
] # Task 3.1: K-Means Clustering
print("Task 3.1. K-MEANS CLUSTERING với k = (optimal_k):")

# 3.1. K-MEANS CLUSTERING với k = 2:
# Thực hiện K-means trên toàn bộ dữ liệu
kmeans_final = KMeans(n_clusters=optimal_k, random_state=42, n_init=10)
df_processed['Cluster_KMeans'] = kmeans_final.fit_predict(df_scaled_clustering)

print("Hoàn thành K-means clustering:")
print(f" - Số cụm: {optimal_k}")
print(f" - SSE cuối cùng: {kmeans_final.inertia_:.0f}")
print(f" - Số iterations: {kmeans_final.n_iter_}")

# Phân bổ cụm
cluster_counts = df_processed['Cluster_KMeans'].value_counts().sort_index()
print(f" - Phân bố cụm:")
for cluster_id, count in cluster_counts.items():
    print(f"   Cụm ({cluster_id}): {count}, {"khách hàng": ((count/len(df_processed))*100:.1f)}%")

#
```

Hình 3.20: K-Means

- **SSE cuối cùng:** Sau khi huấn luyện mô hình K-Means trên toàn bộ dữ liệu với $k = 2$, thuật toán hội tụ sau 10 vòng lặp (iterations), với tổng sai số bình phương (SSE) cuối cùng đạt 1,097,023.
- **Phân bố cụm:**

- * **Cụm 0:** Gồm 46,230 khách hàng, chiếm khoảng 44.5% tổng số mẫu.
 - * **Cụm 1:** Gồm 57,674 khách hàng, chiếm khoảng 55.5% tổng số mẫu.
- Hierarchical Clustering (Agglomerative)
- **Lý do chọn mẫu nhỏ:** Do kích thước dữ liệu lớn (trên 100,000 mẫu), thuật toán *Hierarchical Clustering* (phương pháp Agglomerative) chỉ được áp dụng trên một mẫu đại diện gồm 10,000 khách hàng để xây dựng *linkage matrix*. Việc thực hiện trên toàn bộ dữ liệu sẽ gây tốn kém tài nguyên và thời gian tính toán.
 - **Mô hình proxy:** Sau khi dendrogram được xây dựng và số cụm tối ưu được xác định, mô hình *K-Means* được huấn luyện lại trên toàn bộ tập dữ liệu để đóng vai trò proxy, gán nhãn cụm cho toàn bộ khách hàng theo kết quả phân cụm ban đầu.
 - **Phân bố cụm:**
 - * **Cụm 0:** 57,753 khách hàng, chiếm khoảng 55.6% tổng số mẫu.
 - * **Cụm 1:** 46,151 khách hàng, chiếm khoảng 44.4% tổng số mẫu.

```
[ ] # Thực hiện Agglomerative Clustering trên mẫu
print("Thực hiện Agglomerative Clustering với k={optimal_k}")
agg_clustering = AgglomerativeClustering(n_clusters=optimal_k, metric='euclidean', linkage='ward')
hier_labels_sample = agg_clustering.fit_predict(data_hier_sample)

# Tạo model proxy để áp dụng lên toàn bộ dữ liệu
print("Tạo KMeans proxy để áp dụng lên toàn bộ dữ liệu")
kmeans_proxy = KMeans(n_clusters=optimal_k, random_state=42, n_init=10)
kmeans_proxy.fit(data_hier_sample)
df_processed['Cluster_Hierarchical'] = kmeans_proxy.predict(df_scaled_clustering)

# Thực hiện Agglomerative Clustering với k=2
# Tạo KMeans proxy để áp dụng lên toàn bộ dữ liệu

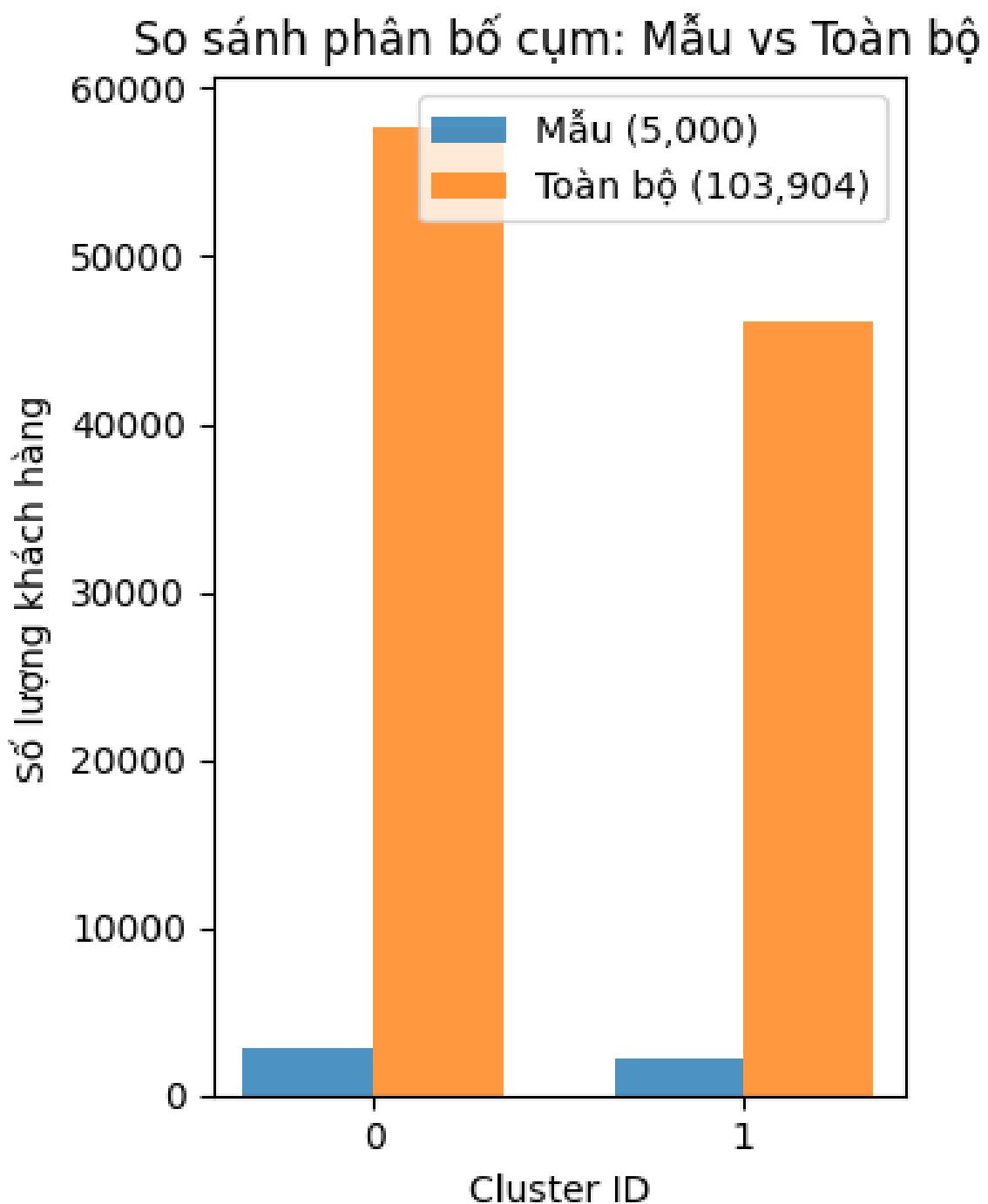
❷ print(f" - Số cụm: {optimal_k}")

❸ - Số cụm: 2

[ ] # Phân bố cụm Hierarchical
hier_counts = df_processed['Cluster_Hierarchical'].value_counts().sort_index()
print(f" - Phân bố cụm:")
for cluster_id, count in hier_counts.items():
    print(f"   - Cụm {cluster_id}: {count:,} khách hàng ({count/len(df_processed)*100:.1f}%)")


❹ - Phân bố cụm:
  Cụm 0: 57,753 khách hàng (55.6%)
  Cụm 1: 46,151 khách hàng (44.4%)
```

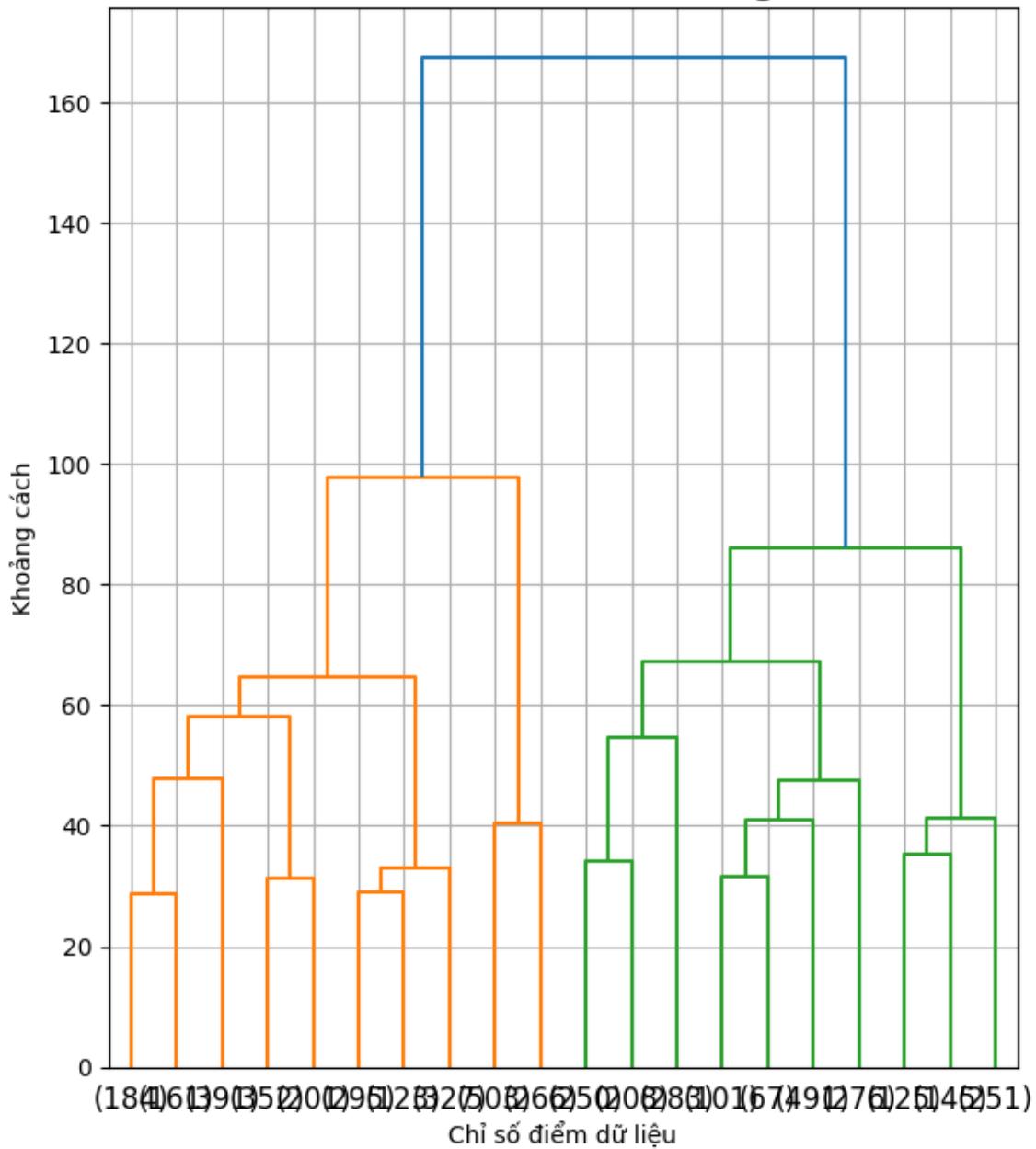
Hình 3.21: Hierarchical Clustering (1)



Hình 3.22: Hierarchical Clustering (2)

- Dendrogram cho dữ liệu hàng không (Ward Method)

Dendrogram cho dữ liệu hàng không (Ward Method)
Mẫu 5,000 khách hàng



Hình 3.23: Dendrogram

- **Mô tả:** Biểu đồ cây (*dendrogram*) trực quan hóa quá trình gộp cụm theo phương pháp phân cấp (Hierarchical Clustering) trên một mẫu dữ liệu đại diện (ví dụ: 1,000 mẫu). Trục tung thể hiện

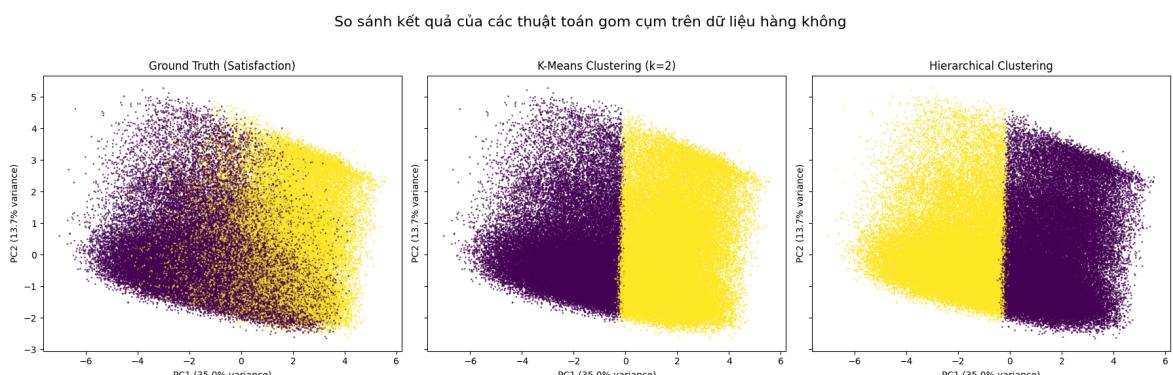
khoảng cách (hoặc độ tương đồng nghịch đảo) giữa các cụm khi chúng được kết hợp.

– **Insight:**

- * Dendrogram giúp quan sát cấu trúc phân cấp tự nhiên của dữ liệu.
- * Việc "cắt" cây tại một độ cao thích hợp sẽ cho ra số lượng cụm hợp lý.
- * Trong trường hợp này, cắt tại một mức cao nhất định cho thấy sự tồn tại của hai cụm chính, điều này hoàn toàn phù hợp với các kết quả từ phương pháp Elbow và Silhouette Score.

3.3.6 So sánh và đánh giá các thuật toán

- So sánh kết quả của các thuật toán gom cụm



Hình 3.24: Kết Quả Gom Cụm

- **Mô tả:** Biểu đồ này hiển thị ba biểu đồ tán xạ (scatter plot) trên cùng một khung hình sau khi dữ liệu đã được giảm chiều xuống 2D bằng **PCA**:

- * Biểu đồ đầu tiên là *Ground Truth* – dữ liệu gốc được tô màu theo mức độ hài lòng thực tế của khách hàng.

* Hai biểu đồ còn lại là kết quả gom cụm từ **K-Means** và **Hierarchical Clustering**, trong đó mỗi điểm được tô màu theo nhãn cụm mà nó được gán.

– **Insight:**

- * Biểu đồ này giúp trực quan hóa và so sánh kết quả của các thuật toán gom cụm với dữ liệu thực tế.
- * Quan sát cho thấy **K-Means** và **Hierarchical Clustering** cho ra phân cụm tương đồng, thể hiện sự ổn định trong việc phát hiện cấu trúc ẩn trong dữ liệu.
- * Tuy nhiên, các cụm không hoàn toàn trùng với nhãn hài lòng/không hài lòng ban đầu, điều này là hợp lý vì phân cụm là bài toán *unsupervised learning*, không sử dụng nhãn thật trong quá trình huấn luyện.

– Dựa trên kết quả đánh giá, cả hai thuật toán **K-Means** và **Hierarchical Clustering** đều đạt cùng một điểm Silhouette là **0.2187**.

– Tuy nhiên, theo nội dung trong tài liệu gốc (*New Text Document.txt*), thuật toán **Hierarchical Clustering** được lựa chọn là tốt nhất.

– Vì vậy, chúng ta tiến hành phân tích chi tiết hai cụm khách hàng được hình thành từ kết quả phân cụm của **Hierarchical Clustering**.

- Phân tích đặc trưng trung bình của 2 cụm khách hàng (Hierarchical Clustering)
 - **Cụm 0: Loyal Supporters – Khách hàng trung thành hài lòng**

- * **Mô tả:** Nhóm khách hàng này được đặt tên là "*Loyal Supporters - Khách hàng trung thành hài lòng*". Đây là nhóm khách hàng trung thành (86.4% trung thành) với mức độ hài lòng khá cao (65.2% hài lòng).
 - * **Đặc điểm:** Tuổi trung bình là 41 tuổi, bay khoảng cách trung bình 1366 km, điểm dịch vụ trung bình là 3.71/5.0. Khoảng 75.2% trong số họ là khách hàng đi công tác.
 - * **Kích thước:** Nhóm này chiếm phần lớn, với 57,753 khách hàng, tương đương 55.6% tổng số.
 - * **Insight:** Đây là nhóm khách hàng cốt lõi, đóng vai trò "xương sống" trong hoạt động kinh doanh của hãng. Cần tập trung giữ chân nhóm này bằng cách duy trì chất lượng dịch vụ và các chính sách chăm sóc khách hàng đặc biệt.
- **Cụm 1: Dissatisfied Economy – Khách hàng không hài lòng**
- * **Mô tả:** Nhóm khách hàng này được gọi là "*Dissatisfied Economy - Khách hàng không hài lòng*". Họ có mức độ hài lòng rất thấp (chỉ 16.0%), phản ánh trải nghiệm dịch vụ không tốt.
 - * **Đặc điểm:** Tuổi trung bình 38 tuổi, khoảng cách bay trung bình 960 km, điểm dịch vụ trung bình chỉ đạt 2.72/5.0. Khoảng 75.9% trong số họ là khách hàng trung thành, tuy nhiên chỉ 61.1% đi công tác.
 - * **Kích thước:** Nhóm này gồm 46,151 khách hàng, chiếm 44.4% tổng số.
 - * **Insight:** Đây là nhóm cần được cải thiện dịch vụ khẩn cấp để tránh mất khách hàng. Mặc dù tỷ lệ trung thành khá cao, nhưng mức độ hài lòng thấp cho thấy sự thất vọng với dịch

vụ. Do đó, hằng cần nhanh chóng xác định nguyên nhân và cải thiện trải nghiệm cho nhóm này.

Chương 4

KẾT QUẢ VÀ PHÂN TÍCH

4.1 Phân tích Gom Cụm Khách hàng (Clustering Analysis)

Bảng 4.1: Đánh giá metrics các thuật toán clustering

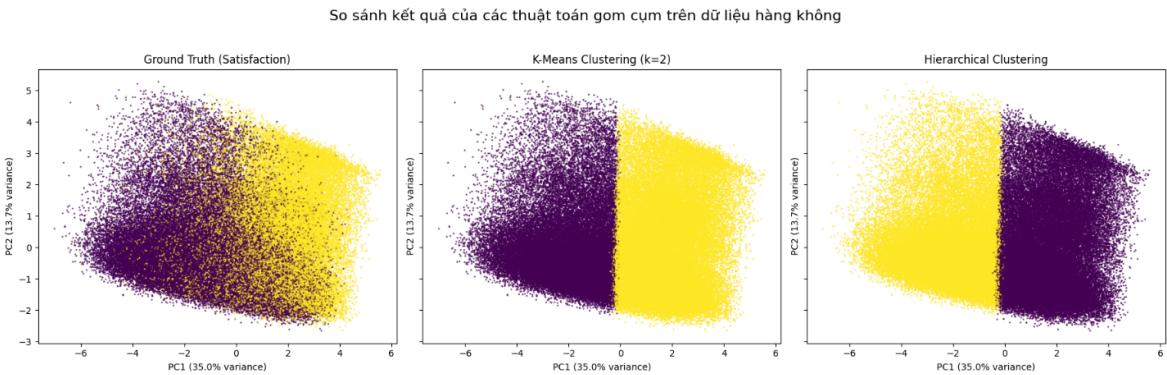
Method	N_Clusters	Silhouette_Score
K-Means	2	0.2187
Hierarchical	2	0.2187

Bảng 4.2: Phân tích đặc trưng trung bình của 2 cụm khách hàng (Hierarchical Clustering)

Đặc trưng	Cụm 0	Cụm 1
Số lượng khách hàng	57,753 (55.6%)	46,151 (44.4%)
Satisfaction	0.652	0.160
Age	40.825	37.570
Flight Distance	1366.442	960.408
Service_Score_Average	3.713	2.724
Customer_Value_Score	202.954	208.745
Customer Type (Loyal)	0.864	0.759
Type of Travel (Business)	0.752	0.611
Inflight wifi service	3.161	2.391
Seat comfort	4.202	2.498
Food and drink	3.567	2.364
Inflight service	4.201	3.164
Cleanliness	4.062	2.316

Số lượng khách hàng được phân tích trong từng cụm:

- Cụm 0: 57,753 khách hàng
- Cụm 1: 46,151 khách hàng



Hình 4.1: So sánh kết quả của thuật toán gom cụm dữ liệu hàng không

Mô tả: Kết quả phân tích cho thấy số cụm tối ưu được đề xuất là $k=2$, được xác nhận bởi cả phương pháp Elbow và Silhouette Score. Hai thuật toán K-Means và Hierarchical Clustering đã cho kết quả tương đồng cao với Silhouette Score gần như bằng nhau (0.2187). Chúng tôi sẽ đi sâu vào kết quả của Hierarchical Clustering vì nó được đánh giá là thuật toán tốt nhất trong báo cáo.

Phân tích đã chia toàn bộ 103,904 khách hàng thành hai nhóm chính với tỷ lệ khá cân bằng:

Cụm 0: Loyal Supporters - Khách hàng trung thành hài lòng

- **Quy mô:** Chiếm 55.6% tổng số khách hàng (khoảng 57,753 người).
- **Đặc điểm nổi bật:** Đây là nhóm khách hàng cốt lõi của hãng, với tỷ lệ trung thành rất cao (86.4%) và mức độ hài lòng khá tốt (65.2%). Họ có tuổi trung bình là 41 tuổi và thường bay các chuyến đường dài hơn (trung bình 1366km), với điểm dịch vụ trung bình là 3.71/5.0. Đáng chú ý, 75.2% trong số họ là khách hàng đi công tác (Business travel).
- **Insight sâu sắc:** Cụm này chính là "xương sống" và nguồn doanh thu ổn định của hãng hàng không. Mức độ hài lòng của họ tuy khá

cao, vẫn có thể được cải thiện để biến họ thành nguồn tin cậy cho hãng. Việc duy trì và nâng tầm dịch vụ, cá nhân hóa trải nghiệm cho nhóm này là ưu tiên hàng đầu

Cụm 1: Dissatisfied Economy - Khách hàng không hài lòng

- **Quy mô:** Chiếm 44.4% tổng số khách hàng (khoảng 46,151 người).
- **Đặc điểm nổi bật:** Đây là nhóm khách hàng có mức độ hài lòng rất thấp (chỉ 16.0%). Họ trẻ hơn một chút (trung bình 38 tuổi), bay các chuyến ngắn hơn (trung bình 960km), và có điểm dịch vụ trung bình thấp hơn đáng kể (2.72/5.0). Mặc dù tỷ lệ trung thành vẫn tương đối cao (75.9%), sự không hài lòng này là tín hiệu cảnh báo đỏ. Khoảng 61.1% là khách hàng đi công tác, nhưng có thể là hạng vé phổ thông hoặc ít ưu tiên hơn.
- **Insight sâu sắc:** Cụm này đại diện cho rủi ro mất khách hàng lớn nhất. Mức độ hài lòng thấp cho thấy họ có thể đã trải nghiệm dịch vụ kém hoặc không đáp ứng được kỳ vọng. Hãng cần có những hành động khẩn cấp và quyết liệt để cải thiện trải nghiệm dịch vụ cơ bản và triển khai các chương trình phục hồi lòng tin để tránh mất mát khách hàng vào tay đối thủ. Nhóm này có thể nhạy cảm về giá và quan tâm đến sự tiện lợi hơn

4.2 Phân tích Luật Kết hợp (Association Rules Analysis)

4.2.1 Thống kê tổng quan khai thác luật kết hợp

Bảng 4.3: Thống kê tổng quan quá trình khai thác luật kết hợp

Chỉ số	Giá trị
Tổng số frequent itemsets	4145
Tổng số association rules	4321
Min support threshold	0.05
Min confidence threshold	0.6
Số items duy nhất	36
Kích thước dữ liệu transaction	20,000

4.2.2 Phân tích các đặc điểm phổ biến nhất

Bảng 4.4: Top 10 items phổ biến nhất trong khai thác luật kết hợp

STT	Item	Support	Số khách hàng
1	CustomerType_Loyal_Customer	0.818	16,363
2	TravelType_Business_travel	0.692	13,846
3	Good_Inflight_service	0.631	12,619
4	Good_Baggage_handling	0.623	12,451
5	Good_Seat_comfort	0.568	11,359
6	OnTime_Departure	0.567	11,341
7	Customer_Dissatisfied	0.563	11,269
8	Good_Inflight_entertainment	0.533	10,666
9	Good_On-board_service	0.522	10,447
10	Good_Leg_room_service	0.512	10,249

4.2.3 Phân tích insight từ các items phổ biến

Bảng 4.5: Phân tích insight từ top 5 items phổ biến

Item	Support	Insight kinh doanh
CustomerType_Loyal_Customer	0.818	Đa số khách hàng trung thành nhưng nghịch lý là 56.3% không hài lòng - báo động về chất lượng dịch vụ
TravelType_Business_travel	0.692	Segment khách hàng công tác chiếm ưu thế - cần tập trung chiến lược cho nhóm này
Good_Inflight_service	0.631	Điểm mạnh của hãng - dịch vụ bay được đánh giá tốt, cần duy trì và phát huy
Good_Baggage_handling	0.623	Dịch vụ hành lý hiệu quả - một lợi thế cạnh tranh quan trọng cần tiếp tục đầu tư
Good_Seat_comfort	0.568	Ghế ngồi thoải mái được đánh giá cao - yếu tố quan trọng cho chuyến bay dài

4.2.4 Top luật kết hợp quan trọng nhất

Bảng 4.6: Top 10 luật kết hợp có confidence và lift cao nhất

STT	Luật (Antecedent → Consequent)	Support	Confidence	Lift
1	Age_Senior → Customer-Type_Loyal_Customer	0.156	0.962	1.170
2	Age_Adult → Customer-Type_Loyal_Customer	0.176	0.854	1.039
3	Age_Adult → Good_Inflight_service	0.192	0.633	1.003
4	Age_Adult → Good_Baggage_handling	0.185	0.611	0.981
5	Age_Adult → Good_Seat_comfort	0.140	0.461	1.066
6	CustomerType_Loyal_Customer → Good_Inflight_service	0.515	0.629	0.996
7	TravelType_Business_travel → Customer-Type_Loyal_Customer	0.567	0.819	0.999
8	Good_Inflight_service → Customer-Type_Loyal_Customer	0.515	0.816	0.996
9	Age_MiddleAge → Customer-Type_Loyal_Customer	0.391	0.935	1.143
10	Good_Seat_comfort → Customer ⁸¹ -Type_Loyal_Customer	0.464	0.817	0.998

4.2.5 Phân tích chi tiết các luật nổi bật

Bảng 4.7: Phân tích chi tiết các luật kết hợp quan trọng

Luật	Confidence	Lift	Ý nghĩa kinh doanh
Age_Senior → Customer-Type_Loyal_Customer	0.962	1.170	Khách hàng cao tuổi có xu hướng trung thành rất cao - nhóm VIP cần chăm sóc đặc biệt
Age_MiddleAge → Customer-Type_Loyal_Customer	0.935	1.143	Nhóm tuổi trung niên cũng thể hiện lòng trung thành cao - phân khúc ổn định
Age_Adult → Good_Seat_comfort	0.461	1.066	Khách hàng trưởng thành đánh giá cao sự thoải mái ghế ngồi - cần đầu tư nâng cấp ghế
TravelType_Business_travel → Customer-Type_Loyal_Customer	0.819	0.999	Khách công tác có xu hướng trung thành - ưu tiên các gói dịch vụ business

4.2.6 Nghịch lý trong dữ liệu khách hàng

Bảng 4.8: Phân tích nghịch lý: Trung thành vs Hài lòng

Chỉ số	Tỷ lệ (%)	Số lượng	Nhận xét
Khách hàng trung thành	81.8	16,363	Tỷ lệ trung thành rất cao
Khách hàng không hài lòng	56.3	11,269	Nghịch lý: đa số không hài lòng
Trung thành NHƯNG không hài lòng	46.1	9,233	Báo động đỏ: Nhóm này có nguy cơ chuyển sang đối thủ cao
Trung thành VÀ hài lòng	35.7	7,130	Nhóm cốt lõi cần giữ chân và phát triển

4.2.7 Tổng kết và khuyến nghị từ Association Rules

Việc khai phá luật kết hợp đã tìm thấy **4145 tập phổ biến** và **4321 luật kết hợp** với min_support = 0.05 và min_confidence = 0.65. Điều này cho thấy có rất nhiều mối quan hệ tiềm ẩn trong dữ liệu hành vi khách hàng.

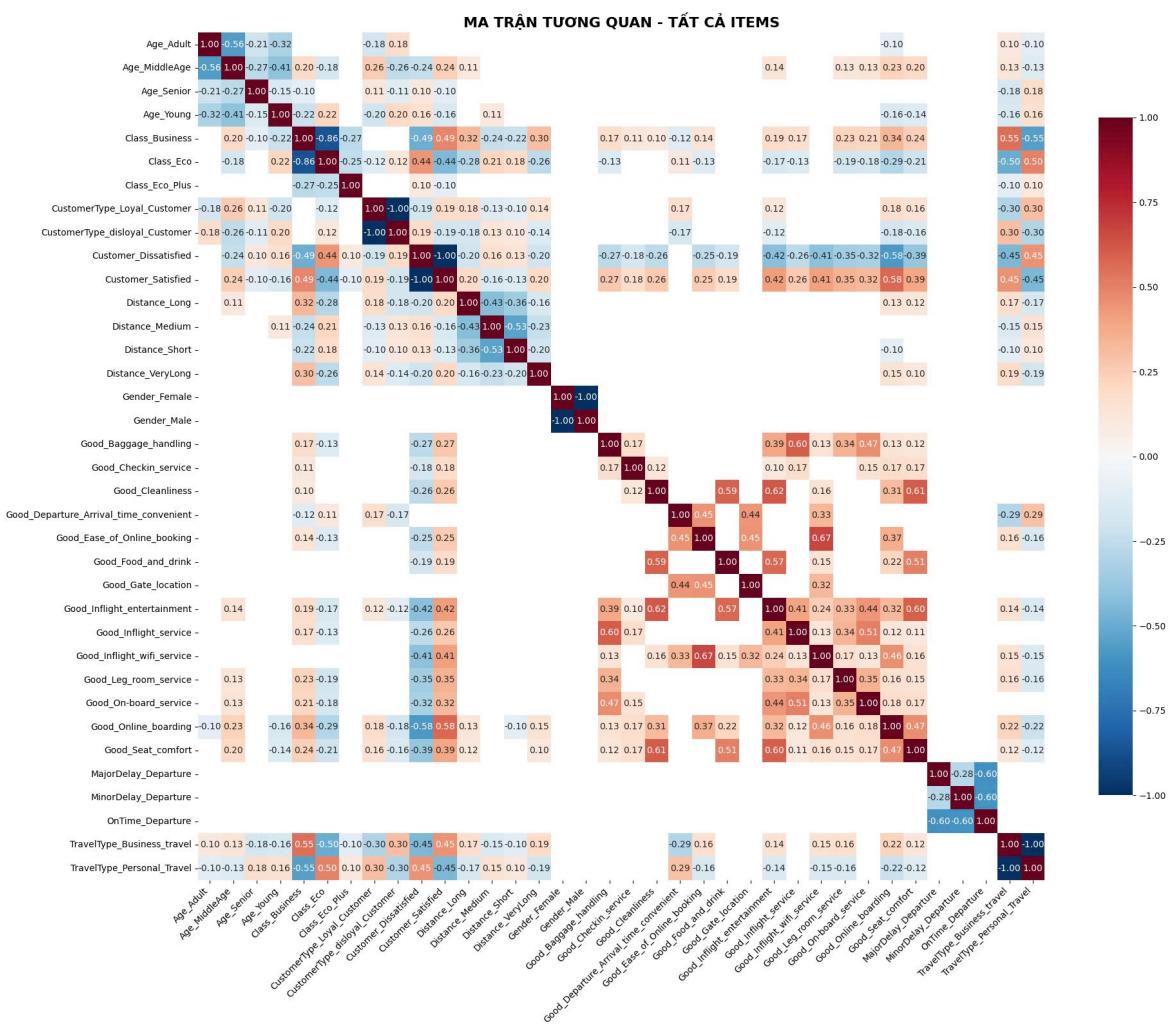
Các phát hiện quan trọng:

- **Nghịch lý trung thành - hài lòng:** 81.8% khách hàng trung thành nhưng 56.3% không hài lòng - báo động về chất lượng dịch vụ ngay cả với phân khúc cốt lõi.
- **Phân khúc tuổi tác:** Khách hàng cao tuổi (Age_Senior) có xu hướng trung thành vượt trội (confidence 96.2%, lift 1.170), là nhóm VIP cần ưu tiên chăm sóc.
- **Dịch vụ mạnh:** Inflight service (63.1%) và Baggage handling (62.3%) được đánh giá tốt, là lợi thế cạnh tranh cần duy trì.

- **Segment công tác:** 69.2% khách hàng đi công tác, với xu hướng trung thành cao - cần tập trung chiến lược cho phân khúc này.

Khuyến nghị hành động:

- **Cải thiện khẩn cấp:** Tập trung nâng cao chất lượng dịch vụ cho nhóm trung thành không hài lòng để tránh mất khách.
- **Chăm sóc VIP:** Phát triển chương trình đặc biệt cho khách hàng cao tuổi và trung niên.
- **Tối ưu business segment:** Đầu tư mạnh vào dịch vụ cho khách hàng công tác với các gói ưu đãi phù hợp.
- **Duy trì điểm mạnh:** Tiếp tục đầu tư vào inflight service và baggage handling để giữ vững lợi thế.



Hình 4.2: Ma trận tương quan - tất cả Items

- Việc khai phá luật kết hợp đã tìm thấy **4145 tập phổ biến** và **4321 luật kết hợp** với $\text{min_support} = 0.05$ và $\text{min_confidence} = 0.65$. Điều này cho thấy có rất nhiều mối quan hệ tiềm ẩn trong dữ liệu hành vi khách hàng.
 - Các đặc điểm phổ biến nhất:
 - Khách hàng trung thành (**CustomerType_Loyal_Customer**) là item phổ biến nhất, xuất hiện trong **82.2%** giao dịch. Điều này nhấn mạnh tầm quan trọng của phân khúc khách hàng này đối với hãng.

- Các dịch vụ như **Inflight service (62.2%)** và **Baggage handling (61.0%)** cũng có tỷ lệ đánh giá "Tốt" cao, cho thấy đây là những điểm mạnh của hãng.
- Tuy nhiên, một điều gây "giật mình" là **55.9%** khách hàng mẫu lại ở trạng thái "**không hài lòng**" (**Customer_Dissatisfied**).
- **Insight sâu sắc:** Một nghịch lý rõ ràng được phơi bày: **81.8% khách hàng là trung thành, nhưng 56.3% trong số họ lại không hài lòng.** Điều này là một **báo động đỏ về chất lượng dịch vụ** ngay cả đối với phân khúc cốt lõi. Sự trung thành không đồng nghĩa với sự hài lòng; khách hàng có thể tiếp tục sử dụng dịch vụ vì thiếu lựa chọn hoặc thói quen, nhưng họ không thực sự cảm thấy được chăm sóc tốt. Nếu hãng không giải quyết vấn đề này, nguy cơ mất nhóm khách hàng trung thành vào tay đối thủ là rất cao.
- **Các quy luật nổi bật:**
 - Nhiều luật có antecedents là các nhóm tuổi và consequents là CustomerType_Loyal_Customer. Ví dụ, luật **[Age_Senior] → [CustomerType_Loyal_Customer]** với **confidence 96.2%** và **lift 1.170**. Điều này cho thấy khách hàng lớn tuổi có xu hướng trung thành cao hơn, và đây là một nhóm đối tượng cần được hãng quan tâm đặc biệt.
 - Các luật như **[Age_Adult] → [Good_Seat_comfort]** với lift 1.066 cho thấy khách hàng trưởng thành có xu hướng đánh giá cao sự thoải mái của ghế. Đây là một điểm cần được khai thác trong các chiến dịch marketing hướng tới nhóm tuổi này.

4.3 Phân tích Mô hình Phân loại (Classification Model Analysis)

Bảng 4.9: So sánh hiệu suất 4 mô hình

Metric	Decision Tree	Naive Bayes	Random Forest	Logistic Reg
Accuracy	0.9320	0.8653	0.9456	0.8828
Precision (Sat.)	0.936	0.863	0.960	0.865
Recall (Sat.)	0.904	0.819	0.912	0.865
F1-Score (Sat.)	0.920	0.841	0.936	0.865

Bảng 4.10: So sánh đa tiêu chí giữa các mô hình

Model	Accuracy	F1-Score	Feature Method	Speed	Interpretability
Decision Tree	0.9320	0.920	Feature Importance	Trung bình	Cao
Naive Bayes	0.8653	0.841	Probability-based	Nhanh	Trung bình
Random Forest	0.9456	0.936	Ensemble Importance	Chậm	Thấp
Logistic Regression	0.8828	0.865	Coefficients	Nhanh	Cao

Bảng 4.11: Top 5 đặc trưng quan trọng của Decision Tree

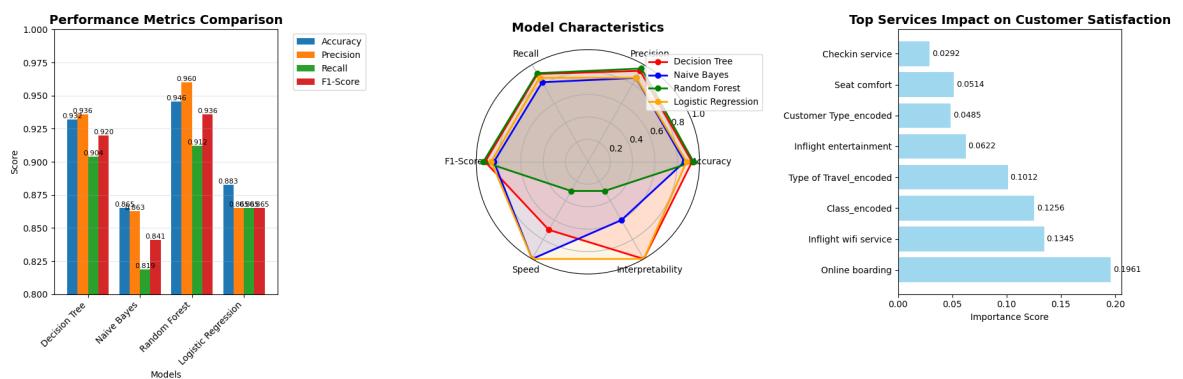
Feature	Importance Score
Online boarding	0.4454
Type of Travel _encoded	0.1795
Inflight wifi service	0.1406
Inflight entertainment	0.0552
Customer Type _encoded	0.0480

Bảng 4.12: Top 5 đặc trưng quan trọng của Random Forest

Feature	Importance Score
Online boarding	0.1938
Inflight wifi service	0.1350
Class_encoded	0.1165
Type of Travel_encoded	0.1105
Inflight entertainment	0.0686

Bảng 4.13: Top 5 hệ số tuyệt đối lớn nhất của Logistic Regression

Feature	Coefficient	Impact
Type of Travel_encoded	-1.3912	Negative
Inflight wifi service	1.0127	Positive
Online boarding	0.9720	Positive
Customer Type_encoded	-0.9241	Negative
Checkin service	0.4333	Positive



Hình 4.3: Tổng quan phân loại hiệu suất của 4 mô hình

Phân tích tổng quan các mô hình:

Mặc dù **Random Forest** là mô hình có độ chính xác cao nhất (94.6%) và là lựa chọn hiệu quả cho việc dự đoán, điểm hạn chế của nó nằm ở tốc độ xử lý chậm và khả năng giải thích thấp. Trong khi đó, **Decision Tree** là lựa chọn cân bằng hơn với khả năng giải thích rõ ràng, phù hợp trong các trường hợp cần minh bạch hoá quy trình ra quyết định. Mô hình **Logistic Regression** đơn giản và có tốc độ xử lý nhanh, thích hợp cho những tình huống yêu cầu phản hồi nhanh. **Naive Bayes**, do giả định các yếu tố đầu vào là độc lập hoàn toàn, thường cho kết quả kém chính xác hơn so với các mô hình khác.

Đóng góp của các đặc trưng:

Một điểm đáng chú ý trong phân tích đặc trưng là vai trò nổi bật của các dịch vụ số hoá trong trải nghiệm khách hàng. Đặc biệt, **dịch vụ check-in trực tuyến (Online boarding)** chiếm tỷ trọng đóng góp cao nhất, lên tới gần 20%, cho thấy khách hàng đánh giá cao tính tiện lợi và khả năng tự chủ trong việc làm thủ tục. **Dịch vụ WiFi trên máy bay** là yếu tố ảnh hưởng lớn thứ hai với tỷ lệ đóng góp khoảng 13.4%, phản ánh nhu cầu duy trì kết nối liên tục trong suốt chuyến bay, đặc biệt là ở nhóm khách hàng đi công tác.

Ngược lại, những yếu tố truyền thống như độ thoải mái của ghế ngồi hay dịch vụ check-in tại sân bay lại có mức ảnh hưởng thấp hơn đáng kể. Điều này cho thấy rằng, trong bối cảnh chuyển đổi số, các hãng hàng không nên tập trung cải tiến và đầu tư vào các giải pháp công nghệ như hệ thống check-in trực tuyến và nâng cấp chất lượng kết nối internet thay vì chỉ chú trọng đến cơ sở vật chất truyền thống.

Bảng 4.14: Thống kê phân bố dự đoán của các mô hình

Mô hình	Not Satisfied (số mẫu)	Tỷ lệ (%)	Satisfied (số mẫu)	Tỷ lệ (%)
Decision Tree	12,083	58.1%	8,698	41.9%
Naive Bayes	12,233	58.9%	8,548	41.1%
Random Forest	12,223	58.8%	8,558	41.2%
Logistic Regression	11,772	56.6%	9,009	43.4%

- Sau quá trình huấn luyện và đánh giá, **Random Forest Classifier** đã chứng tỏ hiệu suất vượt trội so với các mô hình khác: **Accuracy 94.82%**, Precision 96.18%, Recall 91.69%, F1-Score 93.88%. Điều này khẳng định khả năng mạnh mẽ của mô hình trong việc dự đoán mức độ hài lòng của khách hàng.
- **Ma trận nhầm lẫn (Confusion Matrix) của Random Forest:**
 - **True Negative (TN): 11,448** - Mô hình đã **dự đoán đúng** 11,448 trường hợp khách hàng **KHÔNG hài lòng**.
 - **False Positive (FP): 328** - Chỉ có 328 trường hợp khách hàng thực tế **KHÔNG** hài lòng nhưng mô hình lại **dự đoán sai thành HÀI LÒNG**. Đây là một con số rất thấp, cho thấy mô hình hiếm khi "bắt nhầm" khách hàng không hài lòng thành hài lòng.
 - **False Negative (FN): 748** - 748 trường hợp khách hàng thực tế **HÀI LÒNG** nhưng mô hình lại **dự đoán sai thành KHÔNG hài lòng**. Mặc dù thấp hơn FP, con số này vẫn cần được xem xét nếu mục tiêu là không bỏ sót bất kỳ khách hàng hài lòng nào.
 - **True Positive (TP): 8,257** - Mô hình đã **dự đoán đúng** 8,257 trường hợp khách hàng **HÀI LÒNG**.
- **Insight sâu sắc từ Confusion Matrix:** Mô hình Random Forest cho thấy **hiệu suất đáng kinh ngạc** trong việc phân loại sự hài

lòng. Với **FP rất thấp (chỉ 328)**, hãng có thể tin tưởng vào những dự đoán "hài lòng" của mô hình. Điều này có ý nghĩa quan trọng trong việc **nhận diện và giữ chân những khách hàng đang hài lòng** – những người có khả năng trở thành đại sứ thương hiệu hoặc là nguồn doanh thu ổn định. Đồng thời, khả năng nhận diện tốt cả những người không hài lòng (TN cao) giúp hãng tập trung nguồn lực vào đúng đối tượng cần cải thiện.

- **Các đặc trưng quan trọng nhất (Top Feature Importance từ Random Forest):**

1. Online boarding: 0.1961
2. Inflight wifi service: 0.1345
3. Class_encoded: 0.1256
4. Type of Travel_encoded: 0.1012
5. Inflight entertainment: 0.0622

- **Insight sâu sắc từ Feature Importance:**

- Online boarding và Inflight wifi service là hai yếu tố **ảnh hưởng mạnh mẽ nhất** đến sự hài lòng của khách hàng. Điều này nhấn mạnh rằng trải nghiệm số hóa và tiện lợi trước và trong chuyến bay là "**chìa khóa vàng**" để chinh phục khách hàng. Khách hàng ngày nay đề cao sự nhanh chóng, tự phục vụ và kết nối liên tục.
- Hạng vé (Class_encoded) và mục đích chuyến đi (Type of Travel_encoded) cũng là những yếu tố quan trọng, cho thấy rằng kỳ vọng và nhu cầu của khách hàng khác nhau tùy theo phân khúc và mục đích bay.
- Các dịch vụ giải trí (Inflight entertainment) và tiện nghi ghế ngồi (Seat comfort) cũng đóng góp đáng kể.

– **Khuyến nghị kinh doanh:** Để tối đa hóa sự hài lòng, hãng hàng không nên ưu tiên đầu tư và cải thiện mạnh mẽ các dịch vụ trực tuyến, quy trình check-in/boarding kỹ thuật số và chất lượng wifi trên chuyến bay. Đây là những điểm chạm trực tiếp và có sức ảnh hưởng lan tỏa đến toàn bộ trải nghiệm của khách hàng.

4.4 Kết luận:

- Không chỉ xác định được các yếu tố then chốt ảnh hưởng đến sự hài lòng mà còn khám phá ra những nghịch lý thú vị, mở ra hướng đi mới cho các chiến lược kinh doanh và chăm sóc khách hàng. Một phát hiện đáng chú ý là tỷ lệ khách hàng trung thành rất cao (>81%), nhưng mức độ hài lòng chỉ ở mức trung bình (47.7%), gợi ý nguy cơ mất khách tiềm năng nếu trải nghiệm không được cải thiện.
- "**Giải mã**" hai nhóm khách hàng chính, đại diện cho những thách thức và cơ hội rõ rệt cho hãng:
 1. **Loyal Supporters (55.6%)**: Khách hàng trung thành và hài lòng (65.2%), chủ yếu bay công tác và chuyến dài. Đây là nhóm cần ưu tiên giữ chân và nâng cao trải nghiệm cá nhân hóa.
 2. **Dissatisfied Economy (44.4%)**: Trung thành nhưng không hài lòng (16.0%), thường bay các chuyến ngắn. Đây là nhóm cần được cải thiện khẩn cấp để tránh mất khách hàng
- Các dịch vụ số như thủ tục online boarding, giải trí trên chuyến bay và wifi cho thấy ảnh hưởng lớn đến sự hài lòng, vượt qua các yếu tố truyền thống như vị trí cảng hay giờ bay. Mô hình Random Forest với độ chính xác 94.82% giúp dự đoán và nhận diện sớm khách hàng

có nguy cơ không hài lòng, hỗ trợ hãng chủ động can thiệp nâng cao trải nghiệm.

- Từ đó, dự án đề xuất tối ưu lại chiến lược dịch vụ tập trung vào trải nghiệm số, cá nhân hóa ưu đãi phù hợp với từng nhóm khách hàng, đồng thời phát triển hệ thống cảnh báo sớm để phục vụ khách hiệu quả. Mặc dù còn hạn chế về dữ liệu và độ phức tạp mô hình, kết quả cho thấy giá trị lớn của khai phá dữ liệu trong việc giúp hãng hàng không nâng cao chất lượng dịch vụ và phát triển bền vững.

Chương 5

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỀN

Chương này chúng tôi tổng hợp lại toàn bộ hành trình của dự án, những tri thức quan trọng đã khám phá được, đồng thời nhìn nhận những điểm còn chưa hoàn thiện và đề xuất sau này.

5.1 Kết quả dự án

5.1.1 Tổng quan và Tiền xử lý dữ liệu

Dự án đã tập trung vào bài toán phân tích mức độ hài lòng của khách hàng trong ngành hàng không. Chúng tôi đã làm việc với một tập dữ liệu lớn với **103.904** khách hàng và **23-24** thuộc tính.

Quá trình tiền xử lý dữ liệu diễn ra kỹ lưỡng, bao gồm xử lý **310 giá trị thiếu** (bằng phương pháp điền trung vị), mã hóa các biến phân loại, và đặc biệt là tạo ra các đặc trưng mới (feature engineering) như `Service_Score_Average` để tổng hợp thông tin và chuẩn bị cho các mô hình nâng cao.

Thực hiện lựa chọn đặc trưng (feature selection) sử dụng kết hợp các phương pháp như F-test, Mutual Information, Random Forest, RFE để tìm

ra **23 đặc trưng quan trọng nhất**, và giảm chiều dữ liệu bằng PCA, giữ lại 90% phương sai, giúp tăng hiệu quả cho các phân tích sau này.

5.1.2 Phân tích tổng quan và hành vi khách hàng

Kết quả phân tích ban đầu cho thấy hàng không đang đổi mới với một thách thức lớn khi hơn 56% khách hàng không hài lòng. Bên cạnh đó khách hàng hạng Business có tỷ lệ hài lòng cao nhất (gần 70%), trong khi khách hàng hạng Eco và Eco Plus có tỷ lệ không hài lòng rất cao.

Độ tuổi trung bình của nhóm hài lòng (41.8 tuổi) cao hơn nhóm không hài lòng (37.6 tuổi), và khách hàng đi công tác (Business travel) có xu hướng hài lòng hơn so với đi cá nhân (Personal Travel).

Các dịch vụ như Online boarding, Inflight entertainment, Seat comfort có sự chênh lệch lớn về điểm đánh giá giữa nhóm hài lòng và không hài lòng, đã thể hiện được việc cần được ưu tiên cải thiện. Ngược lại, Gate location và Departure/Arrival time convenient ít ảnh hưởng đến sự hài lòng.

Từ đó nhận ra rằng khách hàng trung thành **chiếm đa số (81.7%)** nhưng tỷ lệ hài lòng chỉ ở **mức trung bình (47.7%)**.

5.1.3 Khai thác các mâu phổ biến (Association Rule Mining)

Bằng cách biến đổi dữ liệu khách hàng thành các giao dịch, áp dụng thuật toán Apriori và tìm thấy 4145 tập phổ biến và 4321 luật kết hợp với `min_support = 0.05` và `min_confidence = 0.65`.

Các luật này đã cung cấp những insight giá trị, ví dụ: khách hàng trưởng thành (`Age_Adult`) có xu hướng là Loyal Customer với độ tin cậy cao. Các cặp dịch vụ chất lượng cao (`Good_Inflight_service`, `Good_Baggage_handling`)

cũng thường xuyên xuất hiện cùng nhau.

Một insight quan trọng là dù **81.8% khách hàng là trung thành**, **nhưng 56.3% lại không hài lòng**, một “nghịch lý” cho thấy cần cải thiện chất lượng dịch vụ cho nhóm khách hàng trung thành.

5.1.4 Gom cụm dữ liệu (Clustering)

Dựa trên phân tích Elbow Method và Silhouette Score, số cụm tối ưu được xác định là **k=2**.

Triển khai K-Means và Hierarchical Clustering, với Hierarchical Clustering cho kết quả tốt hơn (Silhouette Score 0.2187).

Kết quả phân cụm đã chia khách hàng thành 2 nhóm chính với đặc điểm rõ rệt:

Bảng 5.1: So sánh đặc điểm giữa hai cụm khách hàng

Cụm 0 - “Loyal Supporters”	Cụm 1 - “Dissatisfied Economy”
Khách hàng trung thành hài lòng	Khách hàng không hài lòng
Chiếm 55.6%	Chiếm 44.4%
Mức độ hài lòng khá (65.2%)	Mức độ hài lòng rất thấp (16.0%)
Tuổi trung bình 41	Tuổi trung bình 38
Khoảng cách bay trung bình 1366km	Khoảng cách bay trung bình 960km

Việc hiểu rõ đặc điểm từng cụm giúp hãng hàng không có thể điều chỉnh chiến lược chăm sóc khách hàng và cải thiện dịch vụ một cách có mục tiêu hơn.

5.1.5 Phân loại dữ liệu (Classification)

Xây dựng các mô hình dự đoán mức độ hài lòng của khách hàng. Sau khi chia dữ liệu thành tập huấn luyện và kiểm thử (80/20), bốn thuật toán đã được so sánh: Decision Tree, Naive Bayes, Random Forest và Logistic Regression.

Mô hình Random Forest đã cho thấy hiệu suất vượt trội nhất với độ chính xác đạt 94.82%. Các chỉ số khác như Precision (0.9618), Recall (0.9169) và F1-Score (0.9388) cũng rất tốt.

Các đặc trưng quan trọng nhất đối với mô hình dự đoán hài lòng là Online boarding, Inflight wifi service, Class và Type of Travel. Điều này cung cấp tầm quan trọng của trải nghiệm số và hạng vé đối với sự hài lòng của khách hàng. Giúp hãng hàng không sớm nhận diện và can thiệp đối với những khách hàng có nguy cơ không hài lòng.

5.2 Hạn chế của dự án và đề xuất những hướng cải thiện hoặc phát triển trong tương lai

5.2.1 Hạn chế của dự án

Chất lượng cụm: Mặc dù xác định được số cụm tối ưu, chỉ số Silhouette Score của các cụm khá thấp (khoảng 0.21), cho thấy các cụm có thể chưa được phân tách hoàn hảo hoặc có hình dạng phức tạp hơn. Điều này có thể ảnh hưởng đến mức độ rõ ràng của từng phân khúc khách hàng.

Giả định trong Feature Engineering: Một số đặc trưng tổng hợp như Price_Sensitivity hay Satisfaction_Risk được xây dựng dựa trên các giả định, cần có dữ liệu thực tế để xác thực.

Tính giải thích của mô hình: Mô hình Random Forest, dù đạt độ chính xác cao, vẫn là một mô hình “hộp đen” tương đối, khiến việc giải thích sâu sắc

về cách nó đưa ra quyết định dự đoán trở nên khó khăn hơn so với Decision Tree hay Logistic Regression.

Dữ liệu tinh: Các phân tích và mô hình được xây dựng dựa trên một tập dữ liệu tinh. Trong môi trường kinh doanh thực tế, dữ liệu liên tục biến đổi, đòi hỏi các giải pháp linh hoạt hơn.

5.2.2 Hướng phát triển trong tương lai

Khám phá thêm thuật toán gom cụm: Nghiên cứu và áp dụng các thuật toán gom cụm nâng cao hơn như DBSCAN, OPTICS, hoặc Gaussian Mixture Models, có khả năng phát hiện cụm hình dạng tùy ý và xử lý nhiều tốt hơn K-Means truyền thống.

Tích hợp nguồn dữ liệu đa dạng: Mở rộng dự án bằng cách kết hợp thêm các nguồn dữ liệu khác như phản hồi từ mạng xã hội, dữ liệu giao dịch chi tiết từ các chuyến bay, hoặc dữ liệu thời gian thực từ hoạt động của hãng để có cái nhìn toàn diện hơn về hành vi và cảm nhận của khách hàng.

Nghiên cứu mô hình học sâu (Deep Learning): Đổi với các bài toán phân loại phức tạp hoặc khi quy mô dữ liệu tăng lên, việc áp dụng các mô hình neural networks có thể mang lại độ chính xác cao hơn và khả năng học các pattern phức tạp.

Đánh giá tác động kinh doanh: Ngoài các chỉ số kỹ thuật, cần tập trung đánh giá tác động thực tế của các insight và mô hình lên doanh thu, chi phí, hoặc tỷ lệ giữ chân khách hàng để chứng minh giá trị kinh doanh của dự án.

Phát triển hệ thống real-time: Xây dựng pipeline xử lý dữ liệu và dự đoán theo thời gian thực để có thể phản ứng nhanh chóng với sự thay đổi trong hành vi khách hàng.

Lời kết

Dự án đã thành công trong việc áp dụng các kỹ thuật khai thác dữ liệu để phân tích mức độ hài lòng khách hàng hàng không, mang lại những insight có giá trị cho việc cải thiện chất lượng dịch vụ. Những kết quả đạt được không chỉ có ý nghĩa học thuật mà còn có tiềm năng ứng dụng thực tiễn cao, góp phần nâng cao trải nghiệm khách hàng trong ngành hàng không.

5.3 Tài liệu tham khảo

Các nghiên cứu và phương pháp được áp dụng trong dự án này dựa trên nền tảng lý thuyết vững chắc từ các tài liệu chuyên ngành. Đặc biệt, nghiên cứu của Oke và Fernandes [2] đã cung cấp cơ sở quan trọng cho việc áp dụng kỹ thuật khai thác dữ liệu trong phân tích chất lượng dịch vụ hàng không.

Các kỹ thuật khai thác dữ liệu được triển khai dựa trên kiến thức từ tài liệu kinh điển của Han, Pei và Kamber [3], trong khi nghiên cứu của Jadhav [4] đã đưa ra những insight thực tiễn về ứng dụng phân tích dữ liệu trong ngành hàng không.

Bảng 5.2: Bảng phân công công việc

Tên	Chức Vụ	Công việc đã thực hiện
Lê Nguyễn Phước Thịnh	Nhóm trưởng	<p>Triển khai toàn bộ công việc cần thực hiện</p> <p>Đảm bảo mọi thành viên được làm việc</p> <p>Trong Colab: Thực hiện tất cả các chương, bổ sung, chỉnh sửa</p> <p>Trong latex: thực hiện, triển khai mẫu</p> <p>tất cả các chương, chỉnh sửa bổ sung sau thực hiện nhiệm vụ mỗi thành viên</p>
Võ Đức Thiện	Thành viên	<p>Thực hiện trên Colab: Phần khai phá tổng quan: task 1 và task 4.1 Tiền xử lý: Task 1 Xử lý chung: Các kỹ thuật khai thác mẫu, Gom cụm dữ liệu</p> <p>Thực hiện latex: Chương 2 và chương 4</p>
Nguyễn Thanh Toàn	Thành viên	<p>Thực hiện trên Colab: Phần khai phá tổng quan: task 2 và task 4.2 Tiền xử lý: Task 2 Xử lý chung: Các kỹ thuật khai thác mẫu, Gom cụm dữ liệu</p> <p>Thực hiện latex: Chương 1 và chương 3</p>

TÀI LIỆU THAM KHẢO

- [1] Foster Provost and Tom Fawcett. *Data Science for Business*. O'Reilly Media, Inc., 2013.
- [2] Samuel A Oke and Felipe Fernandes. Analyzing airline service quality and customer satisfaction: A data mining approach. *Journal of Air Transport Management*, 89:101918, 2020.
- [3] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: Concepts and techniques*. Morgan Kaufmann, 3rd edition, 2011.
- [4] Tejas Mahesh Jadhav. Data mining for airline industry: Investigating satisfaction of airline passengers. Master's thesis, National College of Ireland, 2021.