

BỘ GIÁO DỤC VÀ ĐÀO TẠO  
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HỒ CHÍ MINH  
KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO CUỐI KỲ

MÔN HỌC: KHAI PHÁ DỮ LIỆU

ĐỀ TÀI: KHAI PHÁ DỮ LIỆU VỀ MARKETING

Giảng viên hướng dẫn: **Nguyễn Văn Thành**

Mã môn học: **DAMI330484**

SV thực hiện:	Lê Hoàng Khang	MSSV: 20133050
	Nguyễn Duy Thái	MSSV: 20133020
	Nguyễn Thanh Hùng	MSSV: 20133045
	Hoàng Uyên	MSSV: 20133008

Tp. Hồ Chí Minh, tháng 5 năm 2023

## LỜI CẢM ƠN

Lời mở đầu, nhóm xin gửi lời cảm ơn đến thầy Nguyễn Văn Thành (Giảng viên hướng dẫn môn Khai phá dữ liệu). Thầy đã cung cấp kiến thức, chỉ bảo và đóng góp những ý kiến quý báu giúp nhóm hoàn thành được đồ án môn học của mình. Trong thời gian một học kỳ thực hiện đề tài, nhóm chúng em đã vận dụng những kiến thức nền tảng đã tích lũy đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới vận dụng tối đa những gì đã thu thập được để hoàn thành đề tài đồ án tốt nhất. Tuy nhiên, trong quá trình thực hiện, nhóm chúng em không tránh khỏi những thiếu sót. Nhóm rất mong nhận sự góp ý từ phía thầy nhằm rút ra những kinh nghiệm quý báu và hoàn thiện vốn kiến thức để nhóm có thể tiếp tục hoàn thành những đồ án khác trong tương lai.

Xin chân thành cảm ơn thầy !

## MỤC LỤC

LỜI CẢM ƠN.....	1
MỤC LỤC.....	3
I. TỔNG QUAN ĐỀ TÀI .....	4
1. Lý do chọn đề tài.....	4
2. Dữ liệu sử dụng .....	5
3. Công cụ và thuật toán sử dụng.....	7
3.1. Công cụ sử dụng .....	7
3.2. Thuật toán sử dụng .....	7
II. XỬ LÝ DỮ LIỆU.....	9
III. QUÁ TRÌNH KHAI PHÁ DỮ LIỆU SỬ DỤNG SSAS .....	12
1. Import dữ liệu từ csv vào database.....	12
2. Thuật toán Microsoft Clustering .....	14
2.1. Thực hiện phân cụm dựa trên các thuộc tính “Education”, “Children”, “Age”, “Income”, “Spent”.....	14
2.2. Thực hiện phân cụm dựa trên các thuộc tính “Marital_Status”, “Is_Parent”, “Family_Size”, “NumWebPurchases”, “NumstorePurchaes” .....	30
3. Thuật toán Decision Tree.....	40
4. Thuật toán Association Rule .....	46
5. Đánh giá thực nghiệm và trực quan hóa dữ liệu: .....	53
IV. KẾT LUẬN.....	70
1. Kết quả đạt được .....	70
2. Hạn chế.....	70
3. Bảng phân công nhiệm vụ .....	70
4. Tài liệu tham khảo.....	73

## I. TỔNG QUAN ĐỀ TÀI

### 1. Lý do chọn đề tài

Nhận thấy Marketing là một lĩnh vực quan trọng trong kinh doanh và được áp dụng rộng rãi trong nhiều ngành công nghiệp. Dữ liệu về marketing cung cấp thông tin về xu hướng tiêu dùng, phản hồi khách hàng, chiến lược tiếp thị và quảng cáo, hiệu quả các chiến dịch tiếp thị, và nhiều yếu tố khác liên quan đến việc xây dựng và quản lý thương hiệu.

Sự bùng nổ của Internet và công nghệ đã tạo ra một môi trường kinh doanh mới, mở ra nhiều cơ hội và thách thức cho các doanh nghiệp. Ngành marketing đã phải thích nghi với việc sử dụng các kênh trực tuyến, mạng xã hội và công nghệ mới để tiếp cận và tương tác với khách hàng. Điều này đã làm tăng sự cần thiết của những chuyên gia marketing có kiến thức về các công nghệ mới và cách sử dụng chúng để tạo ra giá trị cho doanh nghiệp.

Hơn thế, khách hàng ngày càng thông minh và tự tin trong quá trình mua sắm. Họ có khả năng tìm hiểu, so sánh và đánh giá sản phẩm và dịch vụ trước khi quyết định mua hàng. Do đó, doanh nghiệp cần phải đưa ra các chiến lược tiếp thị thông minh và tận dụng những kênh tiếp cận khác nhau để giao tiếp và tương tác với khách hàng. Marketing đóng vai trò quan trọng trong việc tạo ra các chiến dịch tiếp thị nhắm vào nhóm khách hàng cụ thể và xây dựng một môi trường tin cậy và hấp dẫn để thu hút và duy trì khách hàng.

Sử dụng khai phá dữ liệu trong marketing có nhiều lợi ích, bao gồm:

- Hiểu rõ hơn về khách hàng: Khai phá dữ liệu giúp phân tích và hiểu rõ hơn về thông tin khách hàng, từ đó có thể tạo ra chiến lược marketing phù hợp và tăng cường sự tương tác với khách hàng.
- Tối ưu hóa chiến dịch quảng cáo: Bằng cách sử dụng khai phá dữ liệu, các nhà quảng cáo có thể tối ưu hóa chiến dịch quảng cáo của mình, từ việc chọn đối tượng khách hàng phù hợp cho đến tối ưu hóa chiến lược quảng cáo.

- Dự đoán xu hướng thị trường: Khai phá dữ liệu cũng giúp dự đoán và đánh giá các xu hướng thị trường, từ đó giúp các nhà quản lý marketing thích nghi và đưa ra các chiến lược phù hợp.
- Tăng hiệu quả doanh số: Bằng cách sử dụng khai phá dữ liệu, các công ty có thể tăng hiệu quả doanh số của mình bằng cách tối ưu hóa chiến lược giá cả, tăng cường sự tương tác với khách hàng, hoặc tối ưu hóa chiến dịch quảng cáo.

## 2. Dữ liệu sử dụng

Tập dữ liệu mà nhóm chúng em chọn có tên là "Marketing Data" và được lưu trữ trên trang Kaggle. Nguồn dữ liệu thu thập:

<https://www.kaggle.com/datasets/jackdaoud/marketing-data>

Tập dữ liệu này bao gồm thông tin về các chiến dịch tiếp thị của một công ty trong vòng 3 năm (2012-2014). Tập dữ liệu này bao gồm các biến sau:

- ID: Mã số khách hàng.
- Year\_Birth: Năm sinh khách hàng.
- Education: Trình độ học vấn của khách hàng.
- Marital\_Status: Tình trạng hôn nhân của khách hàng.
- Income: Thu nhập hàng năm của khách hàng.
- Kidhome: Số lượng trẻ em trong gia đình của khách hàng dưới 18 tuổi.
- Teenhome: Số lượng trẻ em trong gia đình của khách hàng từ 18 đến 25 tuổi.
- Dt\_Customer: Ngày đăng ký thành viên của khách hàng.
- Recency: Số ngày kể từ khi khách hàng mua sản phẩm của công ty lần cuối cùng.
- MntWines: Số tiền khách hàng đã chi tiêu cho rượu vang trong 2 năm qua.
- MntFruits: Số tiền khách hàng đã chi tiêu cho các loại trái cây trong 2 năm qua.
- MntMeatProducts: Số tiền khách hàng đã chi tiêu cho các sản phẩm từ thịt trong 2 năm qua.

- MntFishProducts: Số tiền khách hàng đã chi tiêu cho các sản phẩm từ hải sản trong 2 năm qua.
- MntSweetProducts: Số tiền khách hàng đã chi tiêu cho các sản phẩm từ kẹo và đồ ngọt trong 2 năm qua.
- MntGoldProds: Số tiền khách hàng đã chi tiêu cho các sản phẩm từ vàng, bạc và kim cương trong 2 năm qua.
- NumDealsPurchases: Số lượng giao dịch mà khách hàng đã tham gia với giá khuyến mãi trong 2 năm qua.
- NumWebPurchases: Số lượng sản phẩm mà khách hàng đã mua trên trang web của công ty trong 2 năm qua.
- NumCatalogPurchases: Số lượng sản phẩm mà khách hàng đã mua thông qua các catalog trong 2 năm qua.
- NumStorePurchases: Số lượng sản phẩm mà khách hàng đã mua trực tiếp tại cửa hàng của công ty trong 2 năm qua.
- NumWebVisitsMonth: Số lượng truy cập trung bình của khách hàng trên trang web của công ty trong một tháng.
- AcceptedCmp3: người đó có chấp nhận tham gia chiến dịch tiếp thị số 3 hay không
- AcceptedCmp4: người đó có chấp nhận tham gia chiến dịch tiếp thị số 4 hay không
- AcceptedCmp5: người đó có chấp nhận tham gia chiến dịch tiếp thị số 5 hay không
- AcceptedCmp1: người đó có chấp nhận tham gia chiến dịch tiếp thị số 1 hay không
- AcceptedCmp2: người đó có chấp nhận tham gia chiến dịch tiếp thị số 2 hay không
- Complain: Khách hàng đã phản đối hoặc khiếu nại về sản phẩm hoặc dịch vụ của công ty hay không.

- Z\_CostContact: Chi phí liên lạc với khách hàng.
- Z\_Revenue: Doanh thu từ khách hàng.
- Response: Khách hàng đã phản hồi với chiến dịch tiếp thị hay không.
- Country: Quốc gia của khách hàng.

Các biến này được sử dụng để phân tích hành vi tiêu dùng của khách hàng và thiết kế các chiến dịch tiếp thị hiệu quả.

### **3. Công cụ và thuật toán sử dụng**

#### *3.1. Công cụ sử dụng*

SSAS (SQL Server Analysis Services) là một công cụ phân tích dữ liệu của Microsoft SQL Server. Nó cho phép người dùng tạo các mô hình dữ liệu đa chiều (multidimensional) và mô hình dữ liệu phẳng (tabular) để phân tích dữ liệu từ các nguồn khác nhau.

SSAS cung cấp cho người dùng các tính năng chính sau:

- Khai thác dữ liệu: SSAS cho phép người dùng khai thác dữ liệu từ các nguồn khác nhau và tạo các mô hình dữ liệu đa chiều hoặc phẳng.
- Tính toán và phân tích dữ liệu: SSAS cho phép người dùng tính toán và phân tích dữ liệu bằng cách sử dụng các tính năng như các công thức tính toán, các bộ lọc dữ liệu và các tính năng tổng hợp dữ liệu.
- Tạo báo cáo: SSAS cho phép người dùng tạo các báo cáo dựa trên các mô hình dữ liệu đã tạo.
- Quản lý dữ liệu: SSAS cho phép người dùng quản lý dữ liệu bằng cách sử dụng các tính năng như xử lý dữ liệu, bảo trì dữ liệu và sao lưu dữ liệu.
- Tích hợp với các công cụ khác: SSAS tích hợp tốt với các công cụ khác của SQL Server, chẳng hạn như SQL Server Integration Services (SSIS) và SQL Server Reporting Services (SSRS).

#### *3.2. Thuật toán sử dụng*

##### *3.2.1. Thuật toán Microsoft Clustering*

Microsoft Clustering là một phần của Microsoft SQL Server Analysis Services (SSAS) và được sử dụng để phân tích dữ liệu và phát hiện các mẫu trong dữ liệu. Microsoft Clustering là một thuật toán phân cụm (clustering algorithm) và có thể được sử dụng để phân loại các đối tượng dữ liệu vào các nhóm dựa trên các đặc tính chung của chúng.

### *3.2.2. Thuật toán Microsoft Decision Tree*

Thuật toán Decision Tree là một thuật toán học máy (machine learning) được sử dụng để phân loại và dự đoán giá trị của các đối tượng dữ liệu dựa trên các đặc tính của chúng. Thuật toán này tạo ra một cây quyết định (decision tree) dựa trên các quyết định được đưa ra dựa trên các đặc tính của dữ liệu.

### *3.2.3. Thuật toán Microsoft Association Rules*

Thuật toán Microsoft Association Rules là một thuật toán khai thác dữ liệu được tích hợp trong Microsoft SQL Server Analysis Services (SSAS). Thuật toán này được sử dụng để tìm kiếm các quy tắc kết hợp (association rules) giữa các mục (items) trong tập dữ liệu.

## II. XỬ LÍ DỮ LIỆU

Thực hiện in vài dòng đầu tiên trong tập dữ liệu:

```
1 df.head()
```

Python

ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	...	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	Accep
5524	1957	Graduation	Single	58138.0	0	0	2012-09-04	58	...	0	0	0	0	0
2174	1954	Graduation	Single	46344.0	1	1	2014-03-08	38	...	0	0	0	0	0
4141	1965	Graduation	Together	71613.0	0	0	2013-08-21	26	...	0	0	0	0	0
6182	1984	Graduation	Together	26646.0	1	0	2014-02-10	26	...	0	0	0	0	0
5324	1981	PhD	Married	58293.0	1	0	2014-01-19	94	...	0	0	0	0	0

Thực hiện đổi tên biến cho phù hợp

```
1 # Đổi tên các cột
2 df = df.rename(columns={'MntWines': 'Wines',
3                         'MntFruits': 'Fruits',
4                         'MntMeatProducts': 'Meat',
5                         'MntFishProducts': 'Fish',
6                         'MntSweetProducts': 'Sweets',
7                         'MntGoldProds': 'Gold'})
```

Làm sạch dữ liệu theo các bước sau:

- Xử lý cột Income có chứa giá trị null

```
1 # Tính trung bình thu nhập dựa theo nhóm education
2 group_means = df.groupby('Education')[['Income']].mean()
3
4 # Tính các giá trị thiếu trong cột income
5 df['Income'] = df['Income'].fillna(df['Education'].map(group_means))
6
```

- Trích xuất "Age" của khách hàng bằng cách sử dụng thuộc tính "Year\_Birth" của họ.

```
df['Age'] = 2022 - df['Year_Birth']
```

- Tạo một thuộc tính mới có tên là "Spent" để cho biết tổng số tiền mà một khách hàng đã chi tiêu cho nhiều danh mục khác nhau trong khoảng thời gian 2 năm.

```
# Tổng chi tiêu cho các danh mục khác nhau  
df['Spent'] = df['Wines'] + df['Fruits'] + \  
|   df['Meat'] + df['Fish'] + df['Sweets'] + df['Gold']
```

- Nhóm thuộc tính "Marital\_Status" thành hai loại: "alone" và "partner".

```
# Nhóm tình trạng hôn nhân thành hai tình trạng duy nhất  
df['Marital_Status'] = df['Marital_Status'].apply(  
|   lambda x: "Partner" if x in {"Married", "Together"} else "Alone")
```

- Tạo một thuộc tính mới gọi là "children" để cho biết tổng số trẻ em trong một hộ gia đình, bao gồm cả trẻ em và thanh thiếu niên.

```
# Tính tổng số trẻ em sống trong hộ gia đình  
df['Children'] = df['Kidhome'] + df['Teenhome']
```

- Tạo một thuộc tính mới gọi là "Family\_Size" để làm rõ hơn quy mô của một hộ gia đình.

```
# Tính tổng số thành viên trong hộ gia đình  
df['Family_Size'] = df['Marital_Status'].replace(  
|   {"Alone": 1, "Partner": 2}) + df['Children']
```

- Tạo một thuộc tính mới gọi là "Is\_Parent" để cho biết khách hàng có phải là cha mẹ hay không.

```
# Kiểm tra phải là cha mẹ  
df['Is_Parent'] = np.where(df.Children > 0, 1, 0)
```

- Đơn giản hóa thuộc tính "Education" bằng cách nhóm các giá trị của nó thành ba loại.

```
# Phân chia trình độ học vấn theo ba nhóm
df['Education'] = df['Education'].replace({'Basic': 'Undergrade',
                                         '2n Cycle': 'Undergraduate',
                                         'Graduation': 'Graduate',
                                         'Master': 'Postgraduate',
                                         'PhD': 'Postgraduate'})
```

- Loại bỏ một số biến dư thừa không cần thiết cho phân tích của nhóm.

```
to_drop = ['Dt_Customer', 'Z_CostContact',
           'Z_Revenue', 'Year_Birth', 'Unnamed: 0', 'Kidhome', 'Teenhome']
df = df.drop(to_drop, axis=1)
```

Kết quả cuối cùng sau khi tiền xử lí:

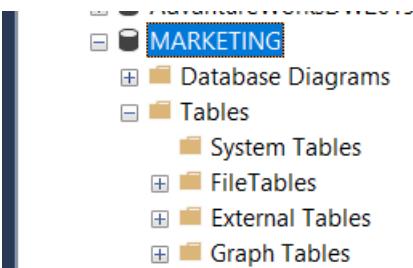
ID	Education	Marital_Status	Income	Recency	Wines	Fruits	Meat	Fish	Sweets	Gold	NumDealsP	NumWebP	NumCatalog	NumStoreP	NumWebVi	AcceptedCn	AcceptedCn	AcceptedCn	AcceptedCn
5524	Graduate	Alone	58138	58	635	88	546	172	88	88	3	8	10	4	7	0	0	0	0
2174	Graduate	Alone	46344	38	11	1	6	2	1	6	2	1	1	2	5	0	0	0	0
4141	Graduate	Partner	71613	26	426	49	127	111	21	42	1	8	2	10	4	0	0	0	0
6182	Graduate	Partner	26646	26	11	4	20	10	3	5	2	2	0	4	6	0	0	0	0
5324	Postgraduat	Partner	58293	94	173	43	118	46	27	15	5	5	3	6	5	0	0	0	0
7446	Postgraduat	Partner	62513	16	520	42	98	0	42	14	2	6	4	10	6	0	0	0	0
965	Graduate	Alone	55635	34	235	65	164	50	49	27	4	7	3	7	6	0	0	0	0
6177	Postgraduat	Partner	33454	32	76	10	56	3	1	23	2	4	0	4	8	0	0	0	0
4855	Postgraduat	Partner	30351	19	14	0	24	3	3	2	1	3	0	2	9	0	0	0	0
5899	Postgraduat	Partner	5648	68	28	0	6	1	13	1	1	0	0	0	20	1	0	0	0
1994	Graduate	Partner	52720374	11	5	5	6	0	2	1	1	1	0	2	7	0	0	0	0
387	Undergraduate	Partner	7500	59	6	16	11	11	1	16	1	2	0	3	8	0	0	0	0
2125	Graduate	Alone	63033	82	194	61	480	225	112	30	1	3	4	8	2	0	0	0	0
8180	Postgraduat	Alone	59354	53	233	2	53	3	5	14	3	6	1	5	6	0	0	0	0
2569	Graduate	Partner	17323	38	3	14	17	6	1	5	1	1	0	3	8	0	0	0	0
2114	Postgraduat	Alone	82800	23	1006	22	115	59	68	45	1	7	6	12	3	0	0	0	1
9736	Graduate	Partner	41850	51	53	5	19	2	13	4	3	3	0	3	8	0	0	0	0
4939	Graduate	Partner	37760	20	84	5	38	150	12	28	2	4	1	6	7	0	0	0	0
6565	Postgraduat	Partner	76995	91	1012	80	498	0	16	176	2	11	4	9	5	0	0	0	1
2278	Undergrade	Alone	33812	86	4	17	19	30	24	39	2	2	1	3	6	0	0	0	0
9360	Graduate	Partner	37040	41	86	2	73	69	38	48	1	4	2	5	8	0	0	0	0
5376	Graduate	Partner	2447	42	1	1	1725	1	1	1	15	0	28	0	1	0	0	0	0
1993	Postgraduat	Partner	58607	63	867	0	86	0	0	19	3	2	3	9	8	0	0	1	0
4047	Postgraduat	Partner	65324	0	384	0	102	21	32	5	3	6	2	9	4	0	0	0	0
1409	Graduate	Partner	40689	69	270	3	27	39	6	99	7	7	1	5	8	0	0	0	0
7892	Graduate	Alone	18589	89	6	4	25	15	12	13	2	2	1	3	7	0	0	0	0
2404	Graduate	Partner	53359	4	173	4	30	3	6	41	4	5	1	4	7	0	0	0	0
5255	Graduate	Alone	52720374	19	5	1	3	3	263	362	0	27	0	0	0	1	0	0	0
9477	Graduate	Partner	38360	26	36	7	42	20	21	10	7	7	1	4	3	0	0	0	0

J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	Sweets	Gold	NumDealsP	NumWebP	NumCatalog	NumStoreP	NumWebVi	AcceptedCn	AcceptedCn	AcceptedCn	AcceptedCn	Complain	Response	Country	Age	Children	Family_Size	Is_Parent	Spent
2	88	88	3	8	10	4	7	0	0	0	0	0	1 US	65	0	1	0	1617	
3	1	6	2	1	1	2	5	0	0	0	0	0	0 US	68	2	3	1	27	
4	21	42	1	8	2	10	4	0	0	0	0	0	0 US	57	0	2	0	776	
5	3	5	2	0	4	6	0	0	0	0	0	0	0 US	38	1	3	1	53	
6	27	15	5	5	3	6	5	0	0	0	0	0	0 US	41	1	3	1	422	
7	42	14	2	6	4	10	6	0	0	0	0	0	0 US	55	1	3	1	716	
8	49	27	4	7	3	7	6	0	0	0	0	0	0 US	51	1	2	1	590	
9	1	23	2	4	0	4	8	0	0	0	0	0	0 US	37	1	3	1	169	
10	3	2	1	3	0	2	9	0	0	0	0	0	0 US	48	1	3	1	46	
11	1	13	1	1	0	0	20	1	0	0	0	0	0 US	72	2	4	1	49	
12	2	1	1	1	0	2	7	0	0	0	0	0	0 US	39	1	3	1	19	
13	1	16	1	2	0	3	8	0	0	0	0	0	0 US	46	0	2	0	61	
14	112	30	1	3	4	8	2	0	0	0	0	0	0 US	63	0	1	0	1102	
15	5	14	3	6	1	5	6	0	0	0	0	0	0 US	70	2	3	1	310	
16	1	5	1	1	0	3	8	0	0	0	0	0	0 US	35	0	2	0	46	
17	68	45	1	7	6	12	3	0	0	1	1	0	0 US	76	0	1	0	1315	
18	13	4	3	3	0	3	8	0	0	0	0	0	0 US	42	2	4	1	96	
19	12	28	2	4	1	6	7	0	0	0	0	0	0 US	76	0	2	0	317	
20	16	176	2	11	4	9	5	0	0	0	1	0	0 US	73	1	3	1	1782	
21	24	39	2	2	1	3	6	0	0	0	0	0	0 US	37	1	2	1	133	
22	38	48	1	4	2	5	8	0	0	0	0	0	0 US	40	0	2	0	316	
23	1	1	15	0	28	0	1	0	0	0	0	0	0 US	43	1	3	1	1730	
24	0	19	3	2	3	9	8	0	1	0	0	0	0 US	73	1	3	1	972	
25	32	5	3	6	2	9	4	0	0	0	0	0	0 US	68	1	3	1	544	
26	6	99	7	7	1	5	8	0	0	0	0	0	0 US	71	1	3	1	444	
27	12	13	2	2	1	3	7	0	0	0	0	0	0 US	53	0	1	0	75	
28	6	41	4	5	1	4	7	0	0	0	0	0	0 US	46	2	4	1	257	
29	263	362	0	27	0	0	1	0	0	0	0	0	0 AUS	36	1	2	1	637	
30	71	10	7	7	1	4	7	0	0	0	0	0	0 AUS	22	1	3	1	111	

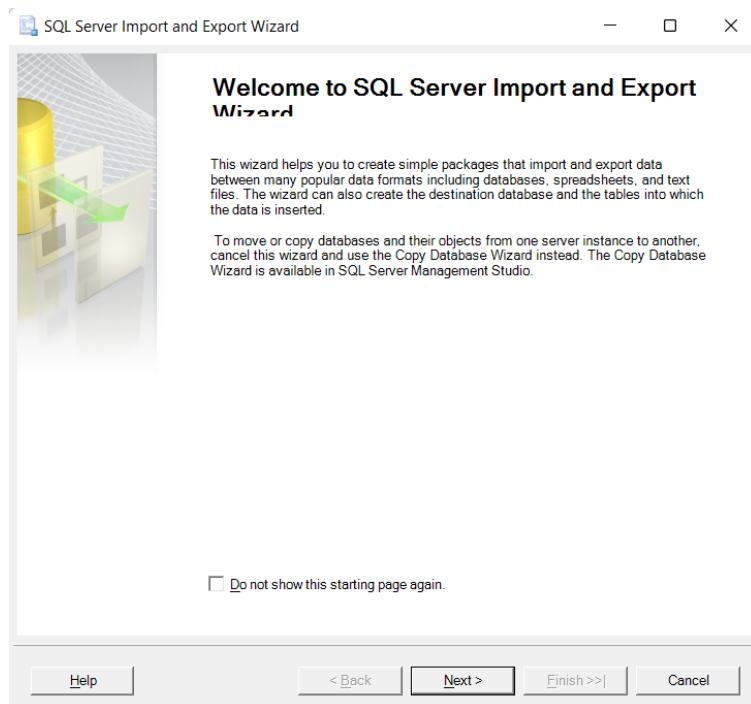
### III. QUÁ TRÌNH KHAI PHÁ DỮ LIỆU SỬ DỤNG SSAS

#### 1. Import dữ liệu từ csv vào database

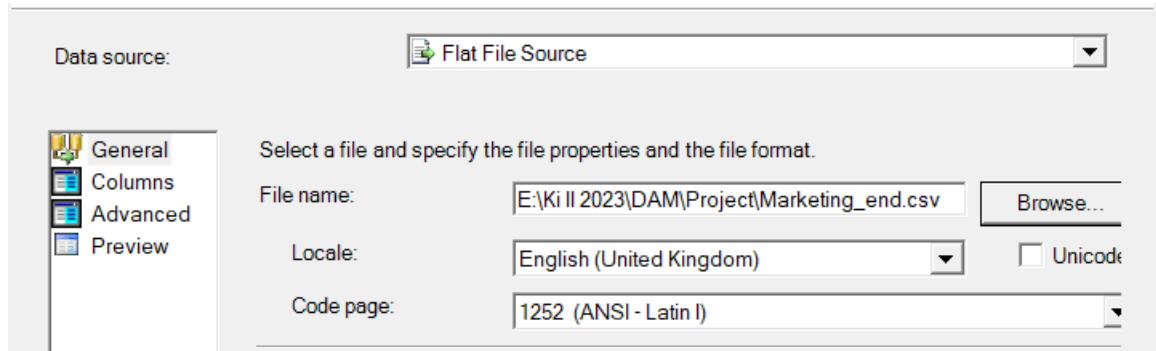
Tạo database tên là MARKETING trong CSDL



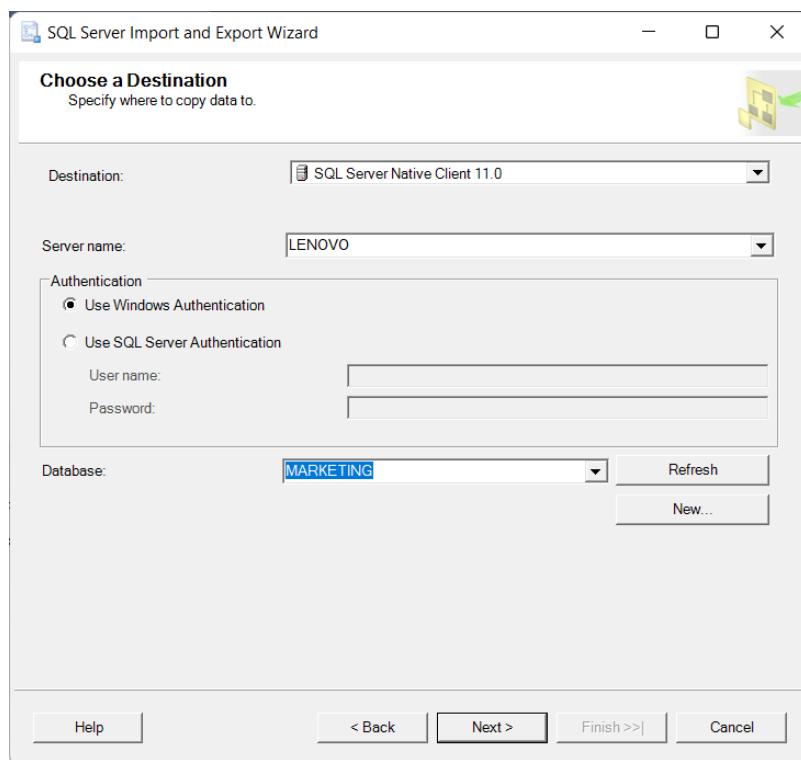
Chọn Tasks -> Import Data



Click next, chọn Flat File Source và đường dẫn chứa file



Click next, chọn server name và database



Kết quả khi import vào database

ID	Education	Marital_Status	Income	Recency	Wines	Fruits	Meat	Fish	Sweets	Gold	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	AcceptedCmp3	Ac
949	Graduate	Partner	39771	92	6	2	18	2	8	14	1	2	0	3	4	0	0
950	2882	Undergraduate	67893	31	711	28	142	49	18	47	2	3	2	9	8	0	0
951	7574	Postgraduate	27922	80	11	0	13	2	4	11	1	2	0	3	4	0	0
952	6387	Postgraduate	52190	39	42	0	17	0	0	18	3	2	1	3	5	1	0
953	5320	Postgraduate	44051	20	79	7	58	6	3	18	4	3	1	4	6	0	0
954	5048	Postgraduate	42767	53	20	6	43	19	5	38	1	3	1	2	8	0	0
955	8146	Postgraduate	46106	84	30	0	8	2	0	14	1	1	1	2	6	0	0
956	100...	Postgraduate	16927	50	20	2	23	3	1	4	5	3	0	4	8	0	0
957	5748	Graduate	59754	96	115	27	44	4	146	139	3	5	2	6	5	0	0
958	2134	Graduate	53700	94	263	5	233	69	41	83	4	5	5	8	5	0	0
959	1523	Graduate	59041	25	69	2	15	2	2	6	2	2	0	4	5	0	0
960	9665	Postgraduate	54237	48	267	3	30	4	0	57	4	5	2	5	6	0	0
961	4640	Graduate	70647	65	561	85	171	25	123	114	2	4	7	13	2	0	0
962	3635	Postgraduate	52597	69	492	0	37	7	0	42	3	6	3	8	5	0	0
963	3547	Postgraduate	41021	12	14	7	9	6	16	12	2	2	0	3	6	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

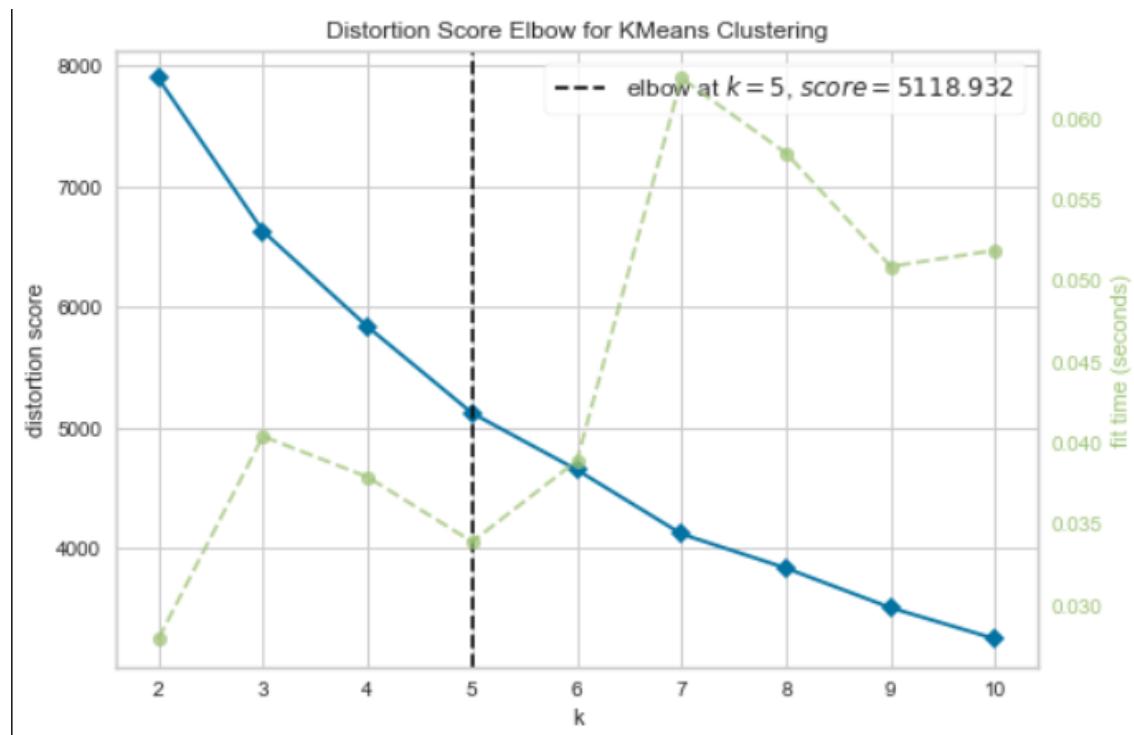
## 2. Thuật toán Microsoft Clustering

Thực hiện 2 lần thuật toán để phân cụm dựa trên 2 tập thuộc tính:

- Tập 1: "Education", "Children", "Age", "Income", "Spent"
- Tập 2: "Marital\_Status", "Is\_Parent", "Family\_Size", "NumWebPurchases", "NumStorePurchases"

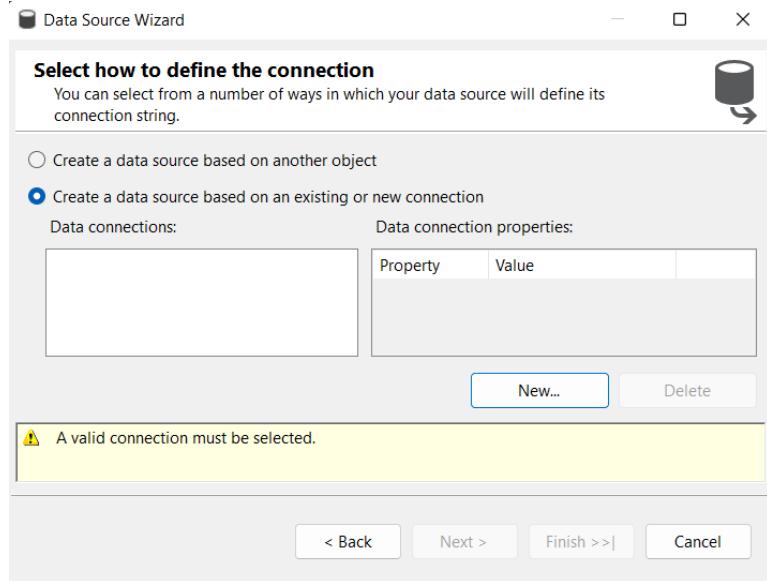
### 2.1. Thực hiện phân cụm dựa trên các thuộc tính “Education”, “Children”, “Age”, “Income”, “Spent”

Trước khi thực hiện phân cụm, ta cần lựa chọn các thuộc tính đó sử dụng phương pháp elbow để xác định tối ưu mà ta cần phân chia, ở hình dưới ta thấy  $k=5$  là tối ưu.

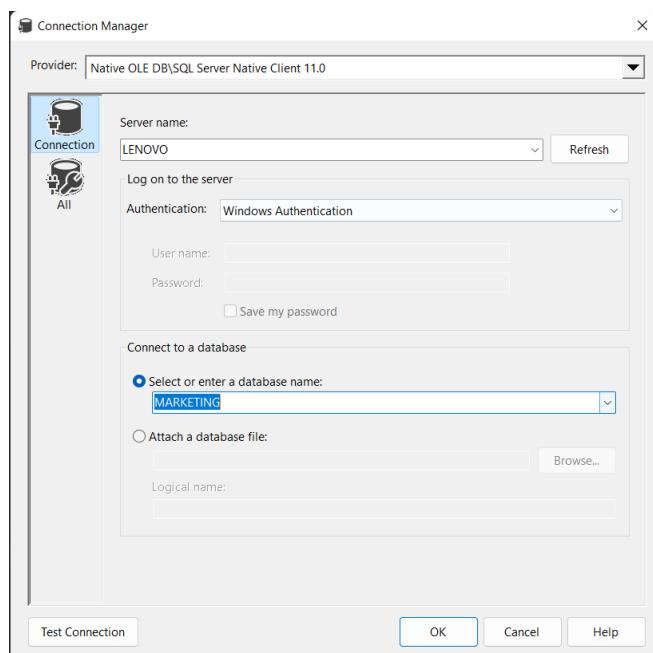


➤ Các bước thực hiện:

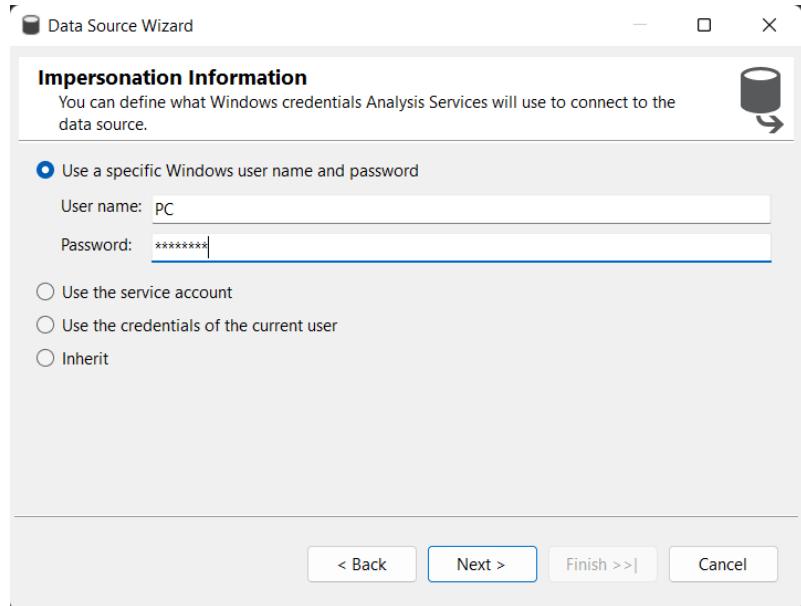
Click phải chuột New Data Source..



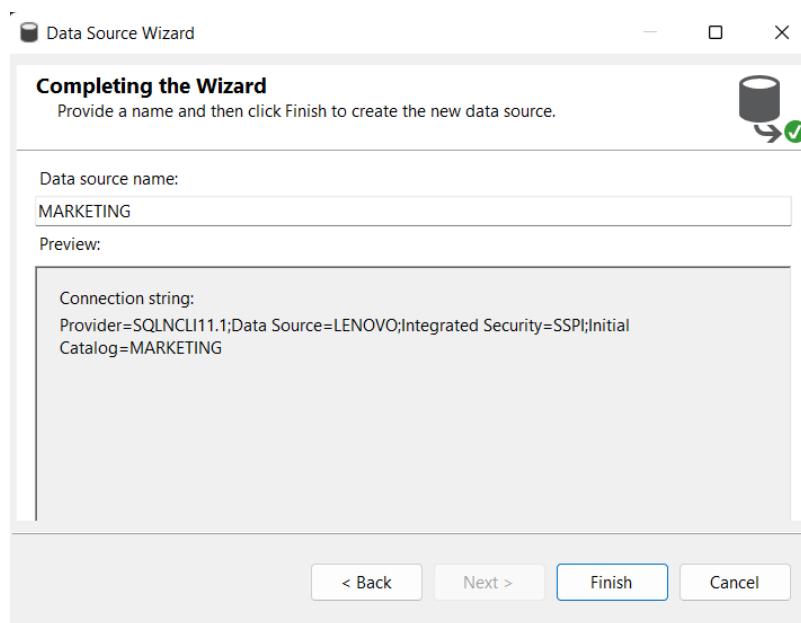
### Chọn new và cấu hình Connection Manager



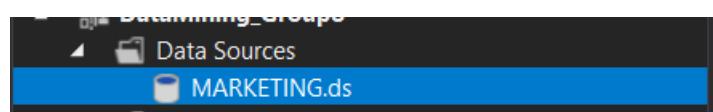
Nhập username và password



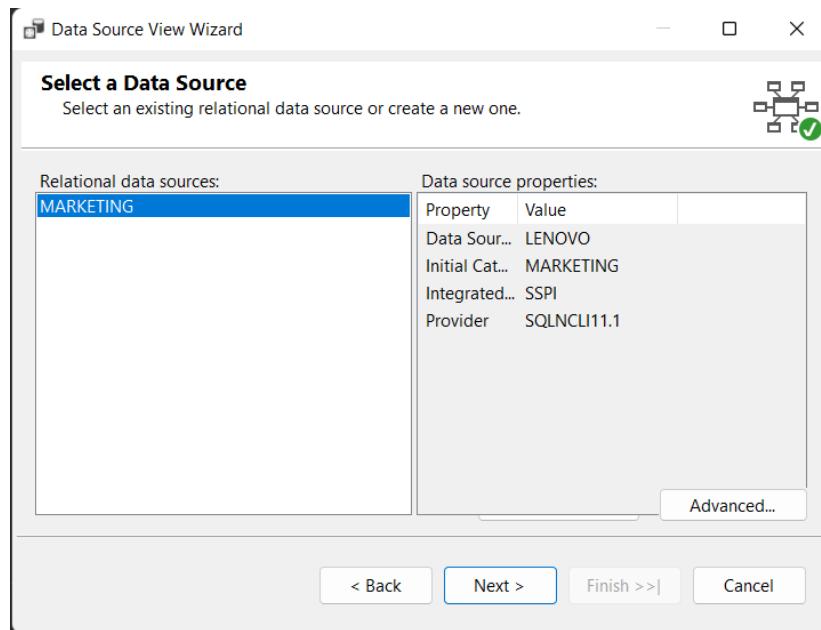
Đặt tên cho Data source name và click finish



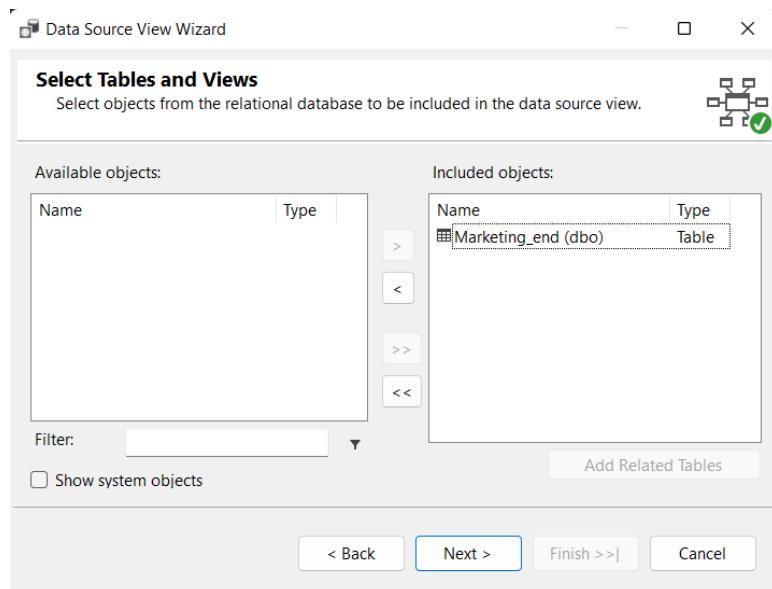
Kết quả:



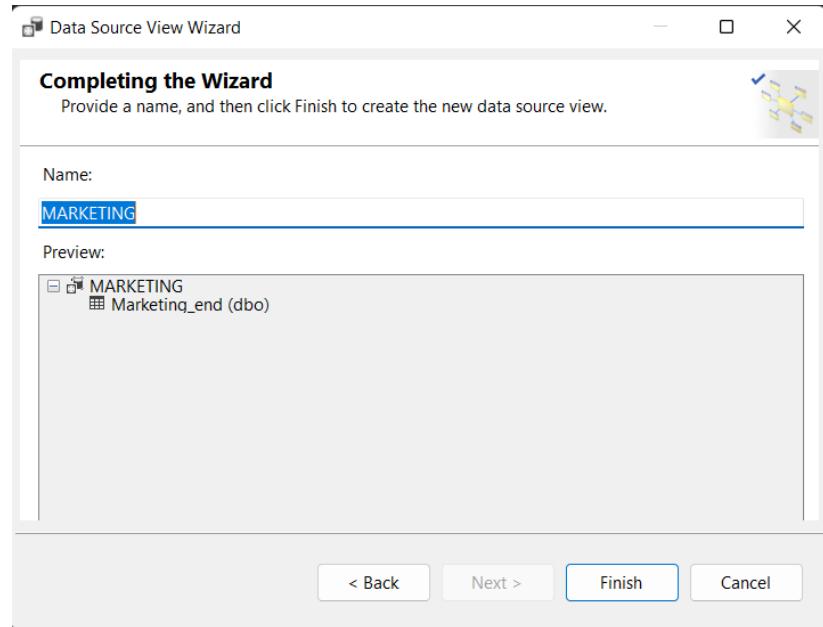
Click phải chuột data source views chọn new data source..



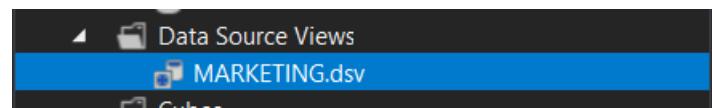
Click next, chọn bảng Marketing\_end



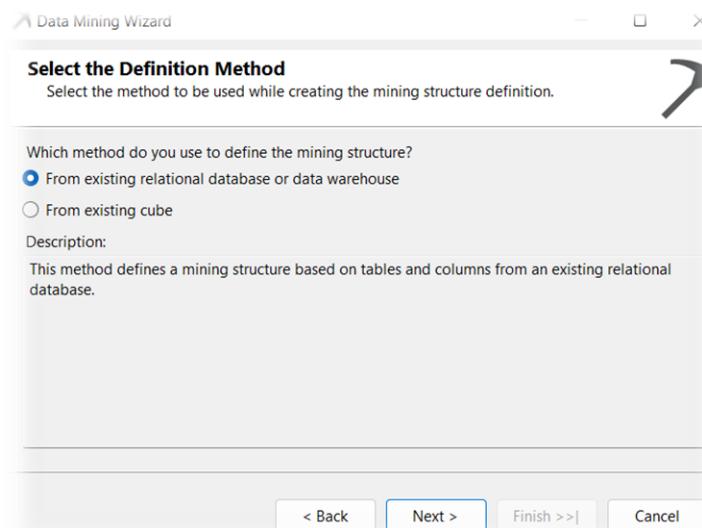
Click next, đặt tên và click finish



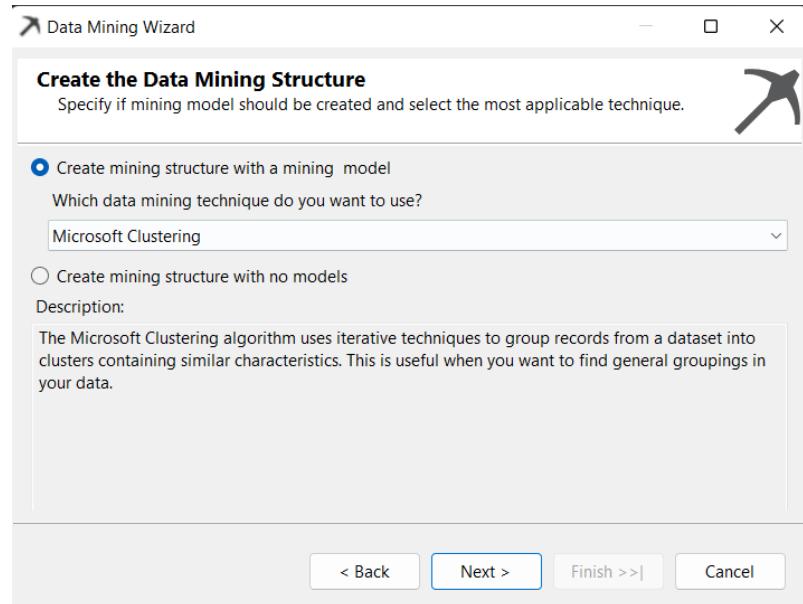
Kết quả



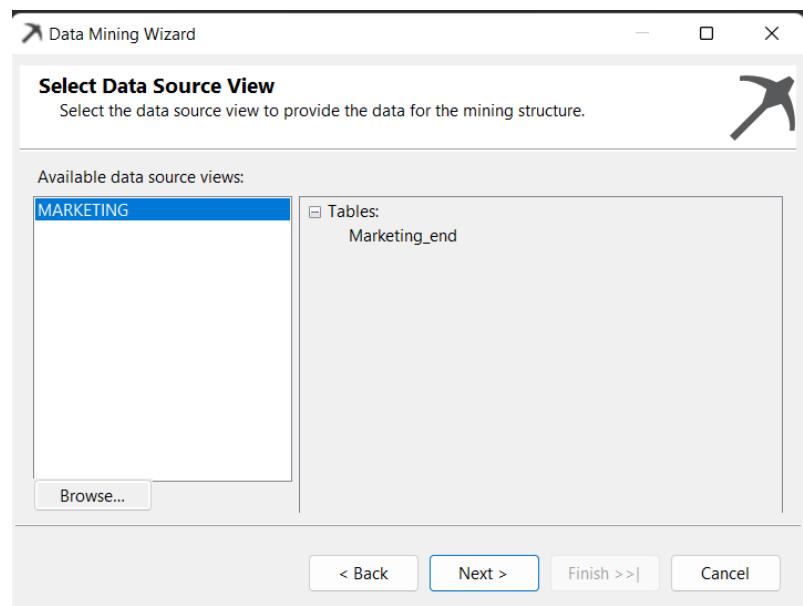
Click phải chuột chọn new mining structure



## Chọn thuật toán Microsoft Clustering



Click next



Chọn các thuộc tính và key phù hợp cho tập đầu tiên gồm: "Age", "Children", "Education", "Income", "Spent"

Click next, định dạng kiểu dữ liệu cho phù hợp

	Content Type	Data Type
Age	Continuous	Double
Children	Continuous	Double
Education	Discrete	Text
ID	Key	Text
Income	Continuous	Double
Spent	Continuous	Double

Click next, vì ta không cần chia tập train test nên để là 0

Đặt tên cho Mining strucer name và click finish

↗ Data Mining Wizard

### Completing the Wizard

Completing the Data Mining Wizard by providing a name for the mining structure.

---

Mining structure name:

Marketing\_Clustering1

Mining model name:

Marketing End

Preview:

Marketing\_Clustering1

- Columns
  - Age
  - Children
  - Education
  - ID
  - Income
  - Spent

Kết quả:

Marketing\_Clustering1.dmm [Design] ✎ ×

Mining Structure Mining Models Mining Model Viewer Mining Accuracy Ch... Mining Model Prediction

Marketing\_Clustering1

Columns

- Age
- Children
- Education
- ID
- Income
- Spent

Data Source View

Marketing\_end

- ID
- Education
- Marital\_Status
- Income
- Recency
- Wines
- Fruits
- Meat
- Fish
- Sweets
- Gold
- NumDealsPurchases
- NumWebPurchases

## Di chuyển qua tab Mining Model

The screenshot shows the Microsoft Data Mining Wizard interface. The top navigation bar has tabs: 'Mining Structure', 'Mining Models' (which is selected and highlighted in blue), 'Mining Model Viewer', 'Mining Accuracy Ch...', and 'Mining Model Prediction'. Below the tabs is a toolbar with icons for refresh, search, and other functions. The main area is divided into two panes. The left pane, titled 'Structure', displays a tree structure for a mining structure named 'Marketing End'. The root node is 'Microsoft\_Clustering'. Under it are six input nodes: 'Age', 'Children', 'Education', 'ID', 'Income', and 'Spent'. The right pane, titled 'Mining Model', shows the details of the 'Microsoft\_Clustering' model.

Cấu hình các tham số cho thuật toán K-Means, ở cột cluster\_count ta chọn là 5  
theo phương pháp elbow đã vẽ trước đó

The screenshot shows the 'Algorithm Parameters' dialog box for the K-Means algorithm. The title bar says 'Algorithm Parameters'. The main area is titled 'Parameters:' and contains a table:

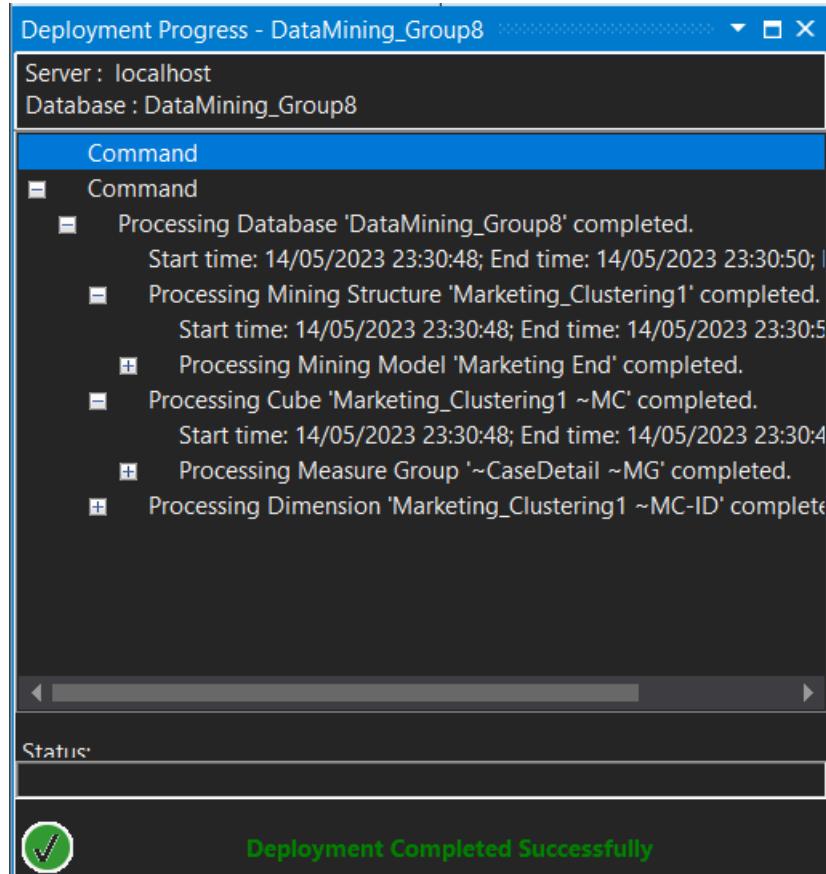
Parameter	Value	Default	Range
CLUSTER_COUNT	5	10	[0,...)
CLUSTER_SEED	0	0	[0,...)
CLUSTERING_METHOD	1	1,2,3,4	
MAXIMUM_INPUT_ATTRIBUTES	255	255	[0,65535]
MAXIMUM_STATES	100	0,[2,6553...	
MINIMUM_SUPPORT	1	(0,...)	
MODELLING_CARDINALITY	10	[1,50]	
SAMPLE_SIZE	50000	0,[100,...)	
STOPPING_TOLERANCE	10	(0,...)	

Below the table is a 'Description:' section with the following text:

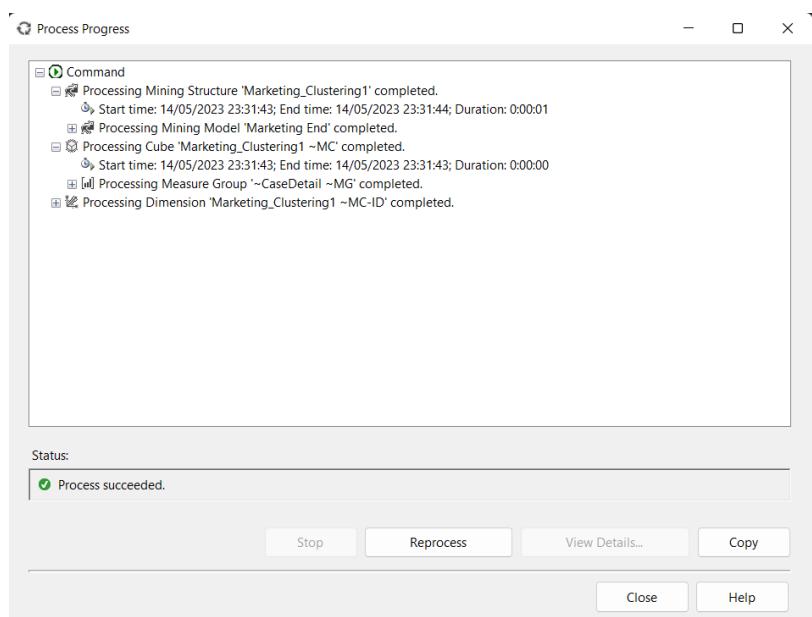
Specifies the approximate number of clusters to be built by the algorithm. If the approximate number of clusters cannot be built from the data, the algorithm builds as many clusters as possible. Setting the CLUSTER\_COUNT parameter to 0 causes the algorithm to use heuristics to best determine the number of clusters to build. The default value is 10.

At the bottom are buttons: 'Add', 'Remove', 'OK' (highlighted in blue), 'Cancel', and 'Help'.

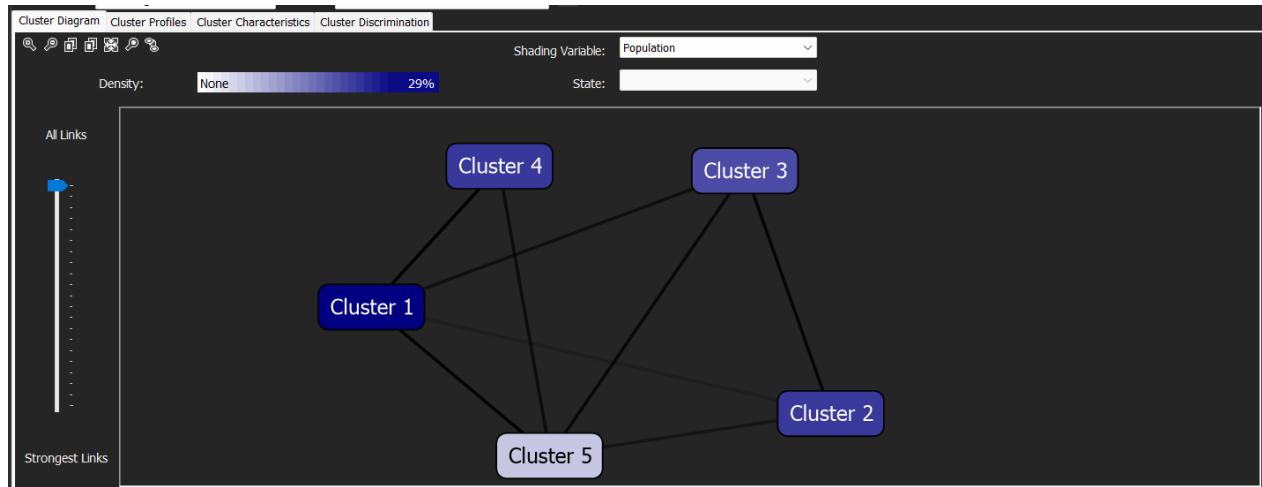
## Deploy project



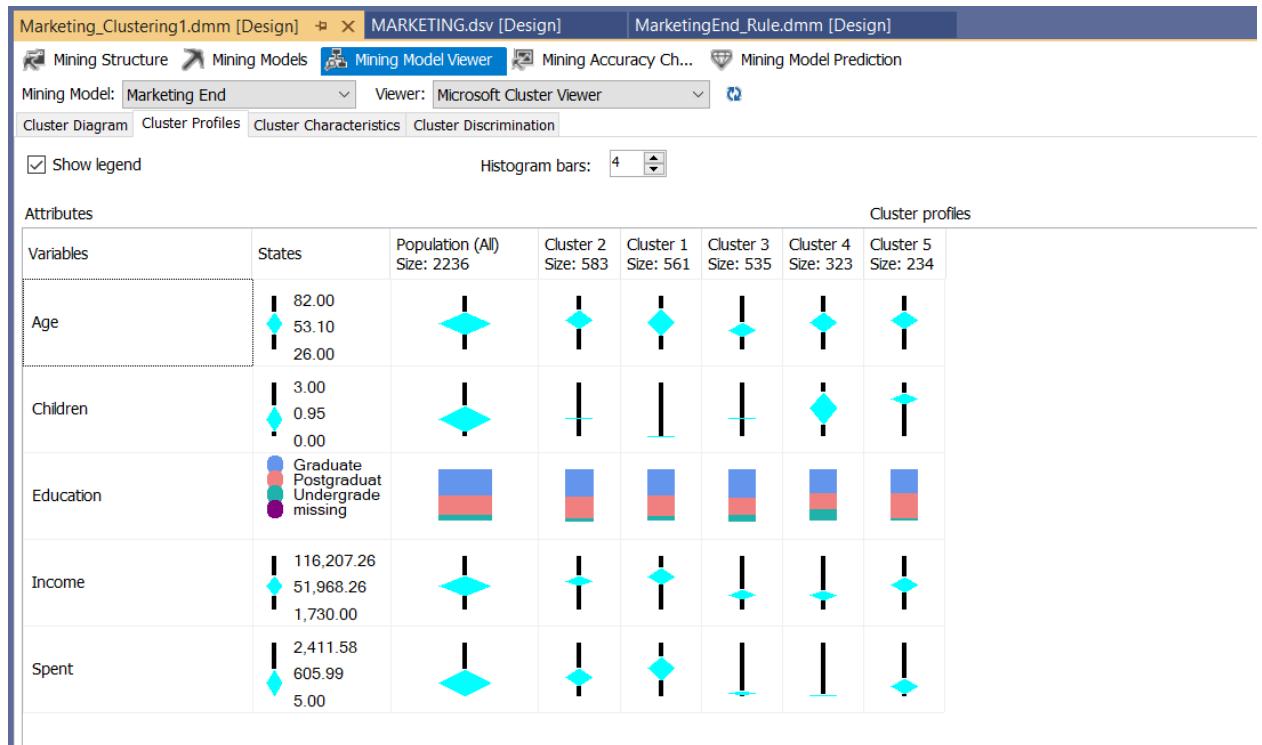
Tiến hành run model, kết quả như hình dưới



Chuyển qua tab Mining Model Viewer, đây là cluster diagram



### 2.1.1. Cluster profiles



➤ Nhận xét chung:

Hình này thể hiện kết quả phân cụm cho nhóm 5 biến bao gồm Age, Children, Education, Income, Spent.

Cột states là một cột được tạo ra bởi thuật toán clustering để đánh giá độ phân bố của các điểm dữ liệu trong các nhóm.

Chúng ta có thể thấy ở tập dữ liệu của chúng ta, độ tuổi trung bình phân bố ở 53,10 tuổi, người có độ tuổi cao nhất là 82, thấp nhất là 26 tuổi.

Tương tự như vậy ta có thấy, sự phân bố thu nhập ở đây cao nhất là 116.207 đô la, thấp nhất là 1.730 đô, thu nhập trung bình của khách hàng là 51.968 đô la.

Các khách hàng có số trẻ con trong gia đình cao là 3 người con, trung bình là 1 con, và không có đứa con nào trong gia đình.

Ở biển spent, cho ta thấy mức chi tiêu trung bình của khách hàng là 605 đô la, cao nhất là 2400 đô, thấp nhất là 5 đô la.

Ở hàng biển education, cho thấy tập dữ liệu chứa cả 3 trình độ học vấn.

Cột population, cho biết số lượng có tất cả 2236 quan sát:

- Spent: giá trị phân bố chủ yếu ở 605,80+-/602,25.
- Age: 53,10+-/11,70.
- Income: 51.968+-/21.413.
- Children : 0.95+-/0.75.
- Education : 1126 Graduate, 885 Postgraduat, 225 Undergrade.

➤ Nhận xét từng cụm:

Ở đây thuật toán đã phân chia ra được 5 cụm dữ liệu, sau đây chúng em sẽ đánh giá kết quả của từng cụm:

- Clustering 2: Là nhóm có số lượng đông nhất với 583 khách hàng, có mức chi tiêu dao động từ 851.18+-412.01 đô la, thu nhập từ 61.799+-/10.658 đô la, thường là có 1 con, độ tuổi dao động 56,87+-/9,87, có đến 51,6% là người có trình độ Graduate, 41,2% là Postgraduation, 7,72% là Undergrade.

- Clustering 1: Là nhóm có số lượng nhiều thứ 2 với 561 khách hàng, có mức chi tiêu dao động từ 1245+-550 đô la, thu nhập từ 71.413+-18.213 đô la, thường không có con , độ tuổi giao động 53,9+-14,33,có đến 50,5% là người có trình độ Graduate, 40,3% là Postgraduation, 9,2 % Undergarde.
- Clutesring 3: Là nhóm có số lượng đông thứ 3 với 535 khách hàng, có mức chi tiêu dao động từ 109+-102.39 đô la, thu nhập từ 33.257+-10928 đô la, thường có 1 con , độ tuổi giao động 46+-8.22, có đến 53.5% là người có trình độ Graduate, 33,6% là Postgraduation, 12,9% là Undergrade.
- Clutesring 4: Là nhóm có số lượng với 323 khách hàng, có mức chi tiêu dao động tu 50+-26 đô la, thu nhập từ 31.518+-11.347 đô la,thường có 2-3 con ,độ tuổi giao động 53.95+-11.21,có đến 45,9% là người có trình độ Graduate,30,9% là Postgraduation, có 23,2% là undergrade.
- Clutesring 5: Là nhóm có số lượng ít nhất với 234 khách hàng, có mức chi tiêu dao động từ 449,70+-366,32 đô la, thu nhập từ 54.863+-16.868 đô la, thường có 2 con , độ tuổi giao động 56,7+-9,26, có đến 46,5% là người có trình độ Graduate, 47,6% là Postgraduation, 5,9% là Undergrade.

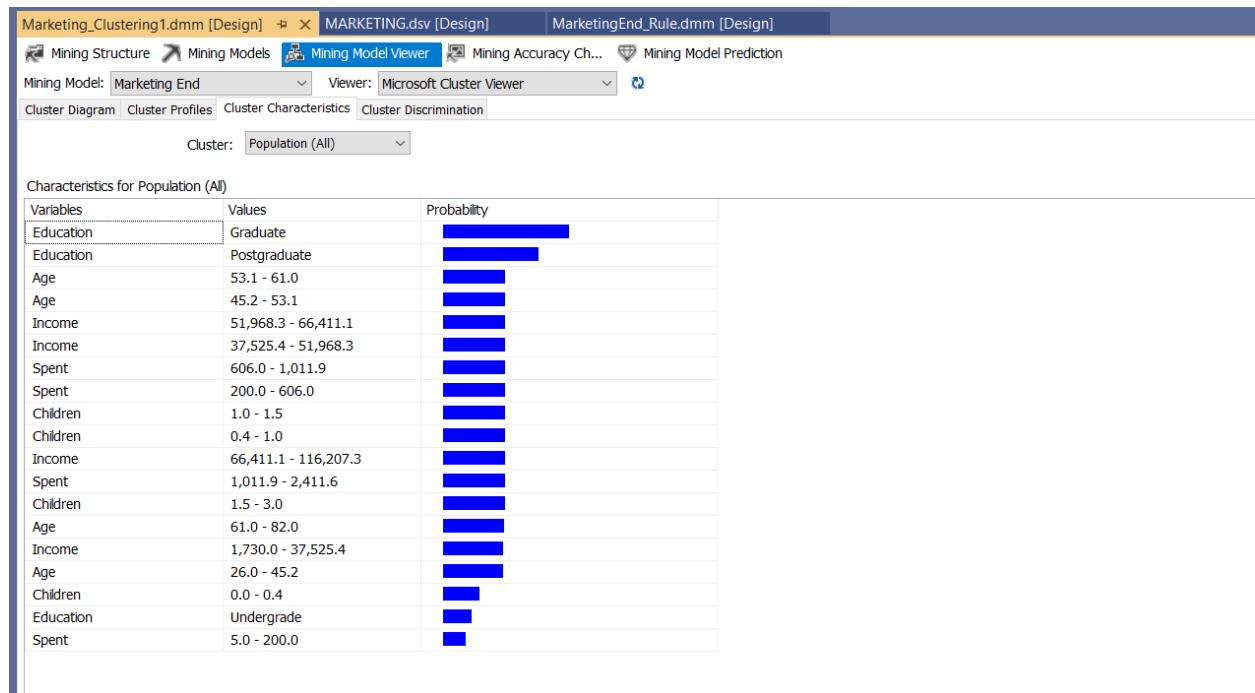
➤ Gán nhãn cụm khách hàng:

- Cluster 2: Khách hàng tốt. Đây là nhóm khách hàng có chi tiêu, thu nhập, tần suất mua hàng và sự hài lòng cao thứ hai. Họ là những khách hàng tiềm năng và có thể trở thành khách hàng ưu tú nếu được chăm sóc tốt.
- Cluster 1: Khách hàng ưu tú. Đây là nhóm khách hàng có chi tiêu, thu nhập, tần suất mua hàng và sự hài lòng cao nhất. Họ là những khách hàng trung thành và quan trọng nhất của bạn.
- Cluster 4: Khách hàng kém. Đây là nhóm khách hàng có chi tiêu, lợi nhuận, tần suất mua hàng và sự hài lòng thấp nhất. Họ là những khách hàng không quan tâm hoặc không phù hợp với sản phẩm hoặc dịch vụ của bạn. Bạn cần xem xét lại chiến lược phục vụ hoặc từ bỏ họ để tập trung vào các nhóm khách hàng khác.

- Cluster 3: Khách hàng bình thường. Đây là nhóm khách hàng có doanh số, lợi nhuận, tần suất mua hàng và sự hài lòng ở mức trung bình. Họ là những khách hàng ổn định và cần được duy trì mối quan hệ.
- Cluster 5: Khách hàng có tiềm năng. Đây là nhóm khách hàng có doanh số, lợi nhuận, tần suất mua hàng thấp nhưng có sự hài lòng cao. Họ là những khách hàng có nhu cầu và mong muốn mua hàng của bạn nhưng chưa được kích hoạt hoặc thuyết phục đủ. Bạn cần tăng cường các chiến dịch marketing và bán hàng để chuyển đổi họ thành khách hàng tốt hoặc ưu tú.

### 2.1.2. Cluster Characteristics

Chuyển qua tab cluster characteristics, ta có thể xem chi tiết các cụm và cho biết xem với 1 biến có giá trị đó thì xác suất nó nằm ở cụm nào là cao nhất.



#### ➤ Nhận xét chung

Bảng này cho thấy các biến số quan trọng nhất để phân biệt các nhóm khách hàng trong tập dữ liệu. Các biến số này được sắp xếp theo thứ tự giảm dần của xác suất xuất hiện trong các nhóm khách hàng.

Biến số Education có giá trị Graduate có xác suất cao nhất (50.358%), cho thấy đây là trình độ học vấn phổ biến nhất trong tập dữ liệu. Điều này cũng có nghĩa là hầu hết khách hàng đều có trình độ học vấn cao và có thể có nhu cầu và khả năng chi tiêu cao hơn.

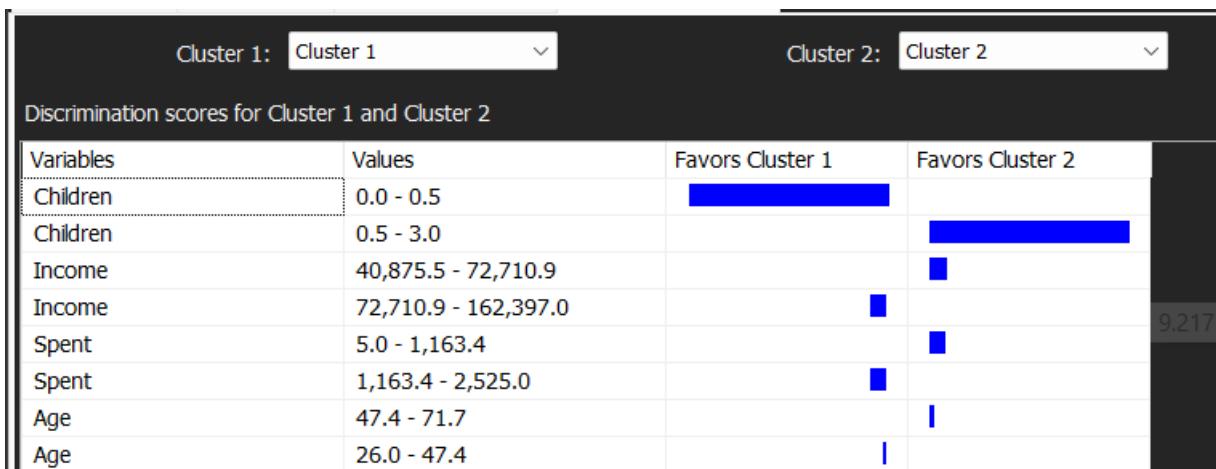
Các biến số Income, Spent, Children và Age có nhiều giá trị khác nhau với xác suất tương đối bằng nhau (24.980%), cho thấy đây là các biến số có sự phân bố đồng đều trong tập dữ liệu. Điều này cũng có nghĩa là các nhóm khách hàng của bạn có thể có sự khác biệt lớn về thu nhập, chi tiêu, số con và độ tuổi.

Biến số Education có giá trị Undergrade có xác suất thấp nhất (11.404%), cho thấy đây là trình độ học vấn ít gặp nhất trong tập dữ liệu. Điều này cũng có nghĩa là khách hàng có trình độ học vấn thấp có thể không phải là mục tiêu chính của doanh nghiệp.

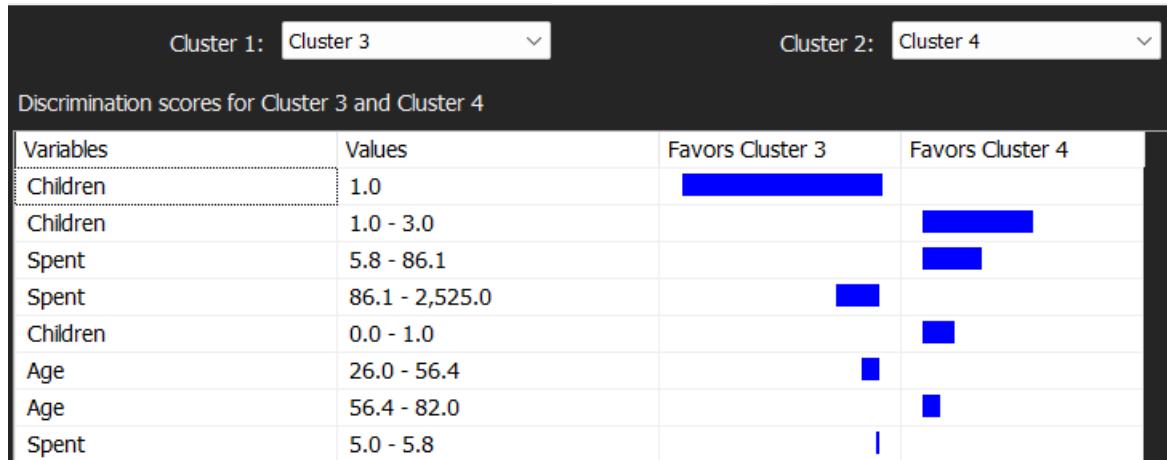
### 2.1.3. Cluster Discrimination

Chuyển qua tab cluster Discrimination, ta có thể so sánh giữa 2 cụm:

- Nếu bạn muốn tìm hiểu về nhóm khách hàng có chi tiêu cao nhất, bạn có thể so sánh cụm 2 và cụm 1, vì hai cụm này có giá trị Spent cao nhất.



- Nếu bạn muốn tìm hiểu về nhóm khách hàng có thu nhập thấp nhất, bạn có thể so sánh cụm 4 và cụm 3, vì hai cụm này có giá trị Income thấp nhất.



#### 2.1.4. Prediction

Chuyển qua tab Mining Model Prediction, ta có thể dự đoán 1 người có các đặc tính như thế thì sẽ thuộc cụm nào. Ở dưới ta cần phân cụm 1 người có tuổi là 33, có số con là 2 , trình độ Education là Postgraduate, thu nhập đạt 22000 và chi tiêu khoảng 3000.

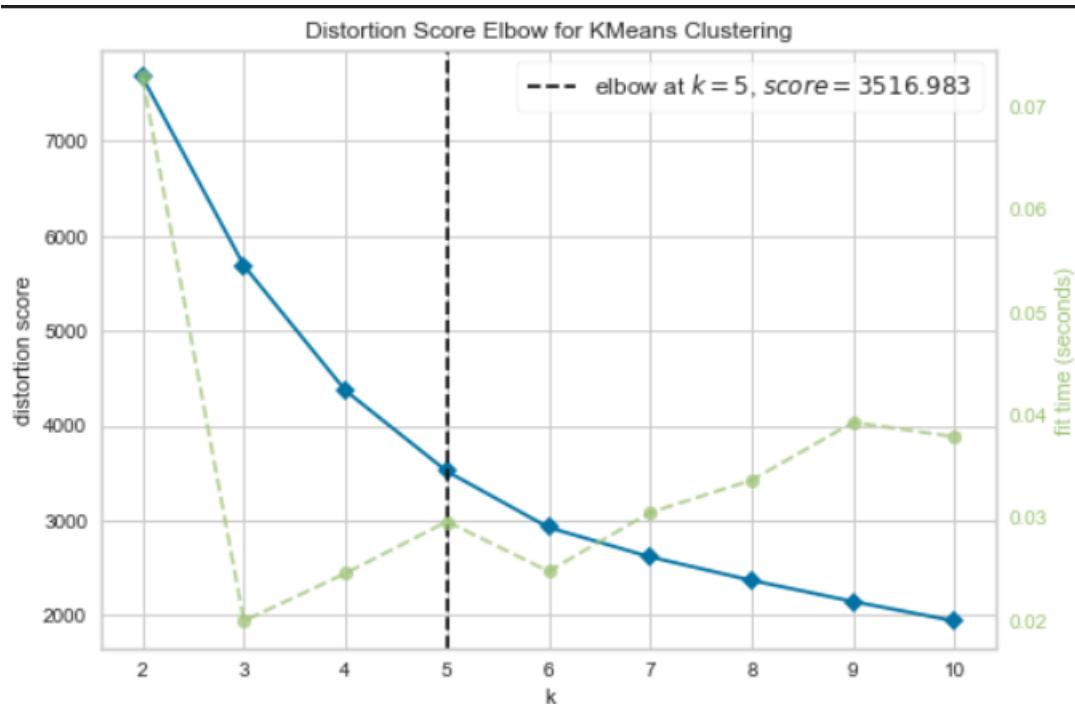
Source	Field	Alias	Show	Group	And/Or	Criteria/Argument
Prediction Function...	Cluster	Predict Cluster	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Sau đó, click vào result. Kết quả người đó thuộc cụm 5.



## 2.2. Thực hiện phân cụm dựa trên các thuộc tính “Marital\_Status”, “Is\_Parent”, “Family\_Size”, “NumWebPurchases”, “NumstorePurchaes”

Trước khi thực hiện phân cụm, ta cần sử dụng thuật toán elbow trên các biến đã được chọn để xác định số cụm mà ta cần phân chia. Theo hình dưới, ta có thể thấy k=5 là số cụm tối ưu nhất.

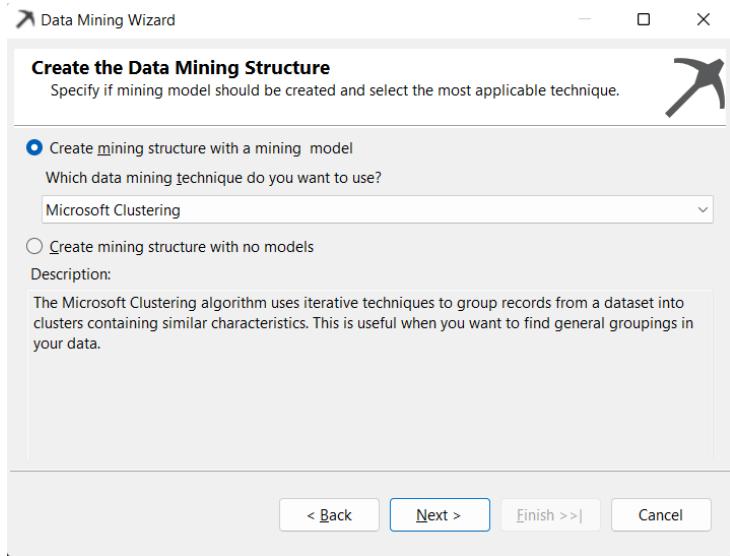


### ➤ Các bước thực hiện

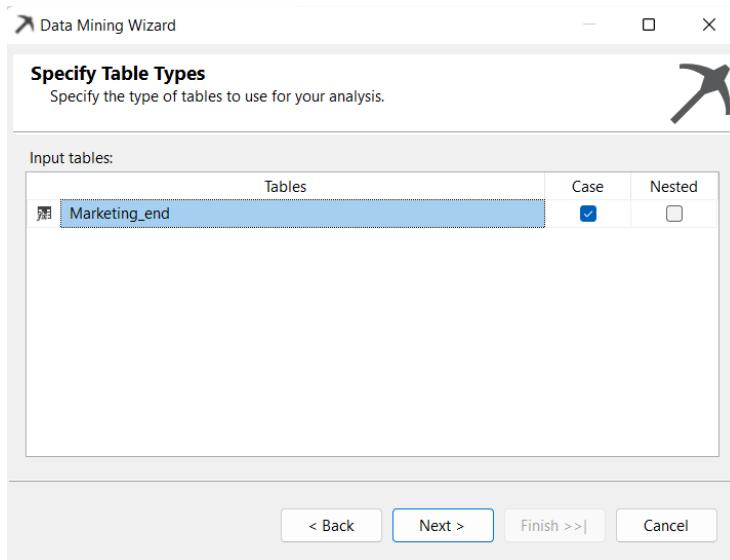
Đã tạo data source ở phía trước

Đã tạo data source view ở phía trước

Click phải chuột chọn New mining structure, chọn thuật toán Microsoft Clustering



Click next

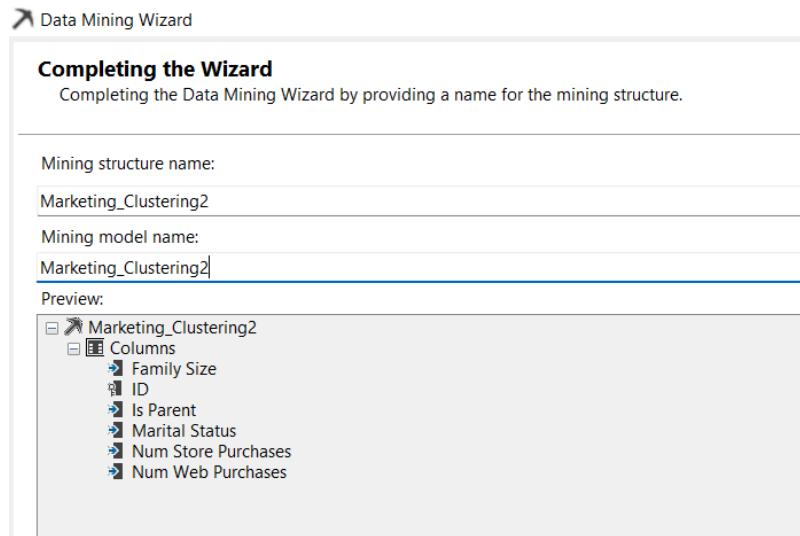


Chọn các input là các thuộc tính cần phân cụm: "Marital\_Status", "Is\_Parent", "Family\_Size", "NumWebPurchases", "NumStorePurchases"

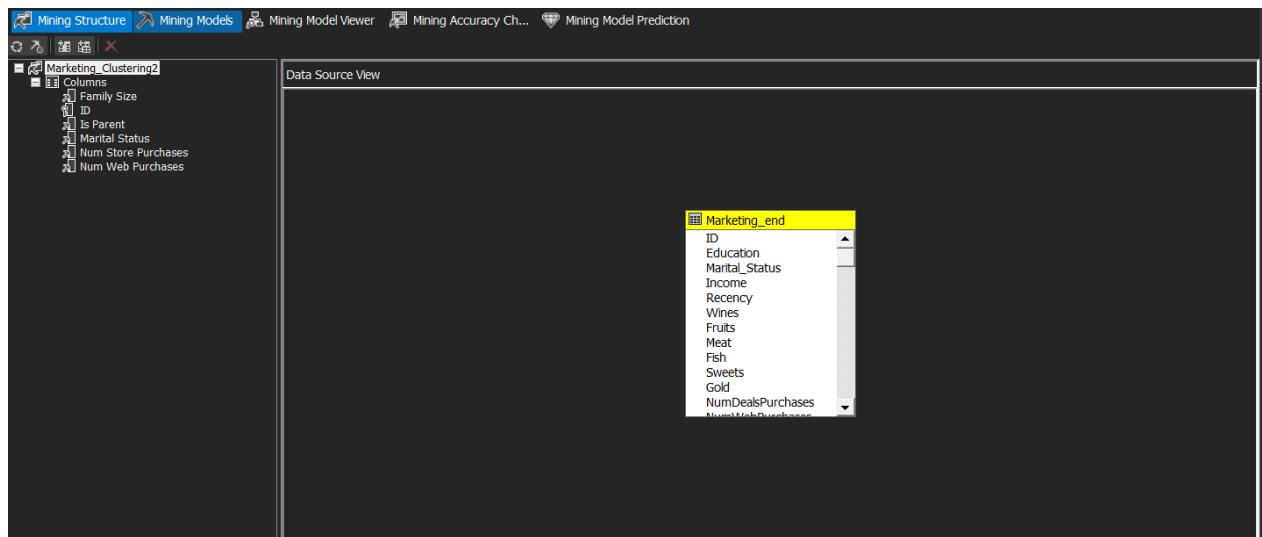
Click next, chỉnh lại các kiểu dữ liệu cần thiết

Không cần chia tập train và test cho thuật toán này

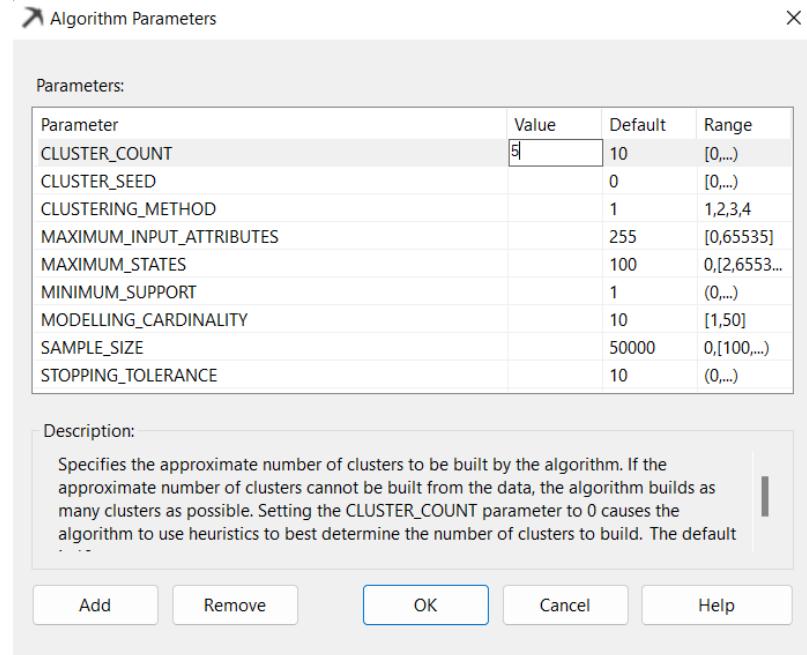
Click next , đặt tên cho mining strucer name và finish



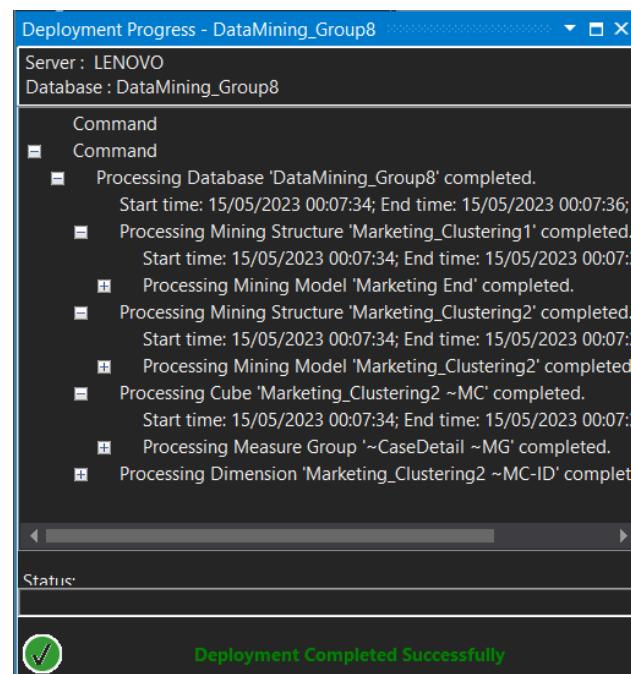
Kết quả:



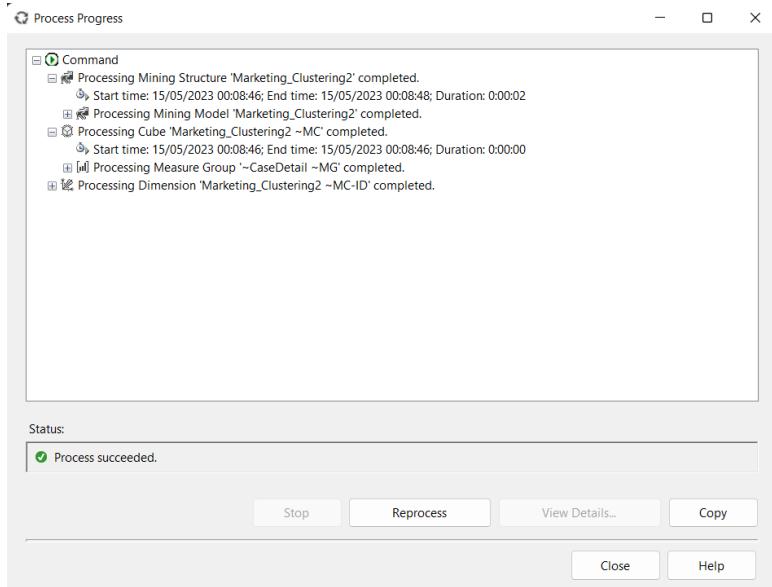
Tiến hành cài đặt các tham số cho mô hình, ở đây số cụm cần phân chia là 5 theo thuật toán elbow



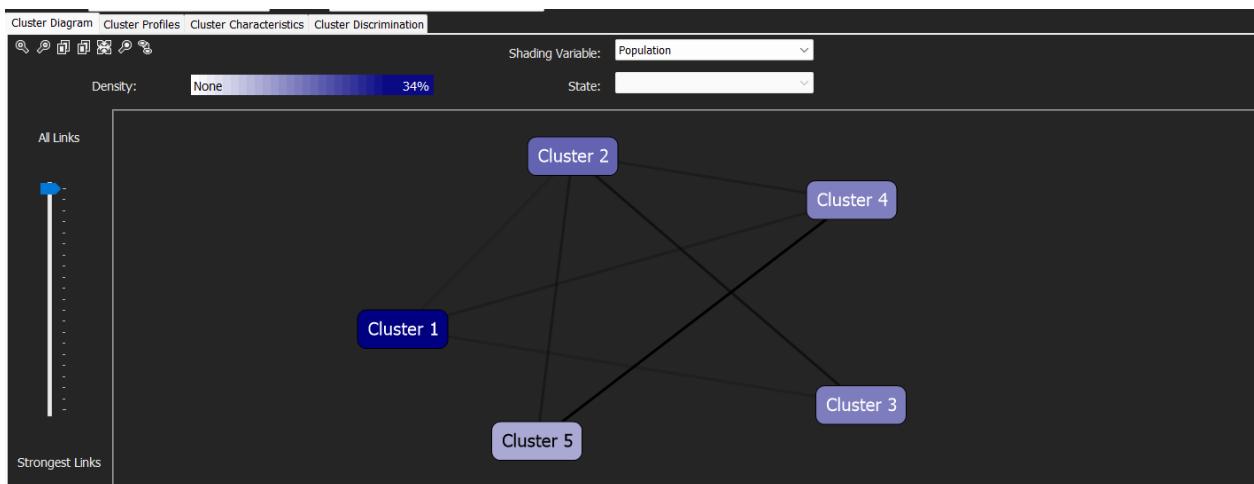
Tiến hành deploy project



## Process model

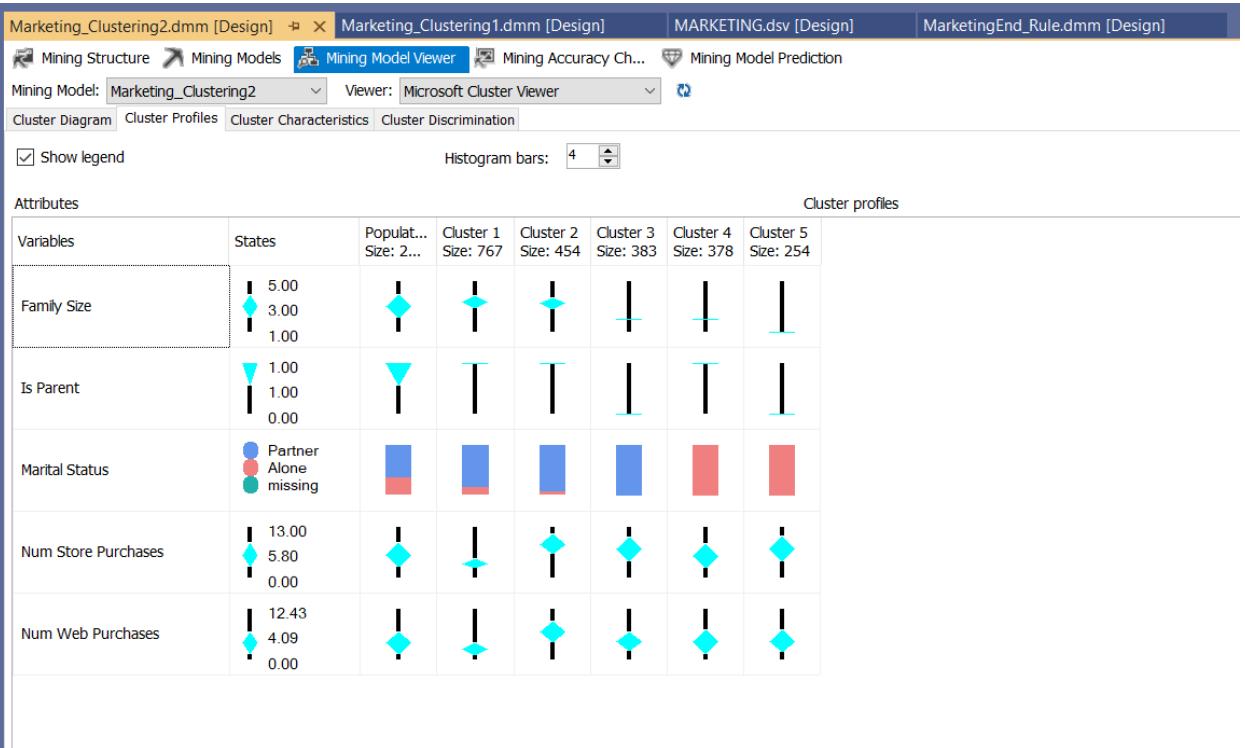


Chuyển qua tab Mining Model Viewer , đây là Cluster Diagram sau khi phân cụm



### 2.2.1. Cluster Profiles

Chuyển qua tab Cluster Profile, đây là hình ảnh mô tả chi tiết các thuộc tính của từng cụm



➤ Nhận xét chung:

Hình cluster diagram, hình này thể hiện kết quả phân cụm cho nhóm 5 biến bao gồm: Family\_Size, Is Parent, Marital Status, Num Store Purchases, Num Web Purchases.

Cột states là một cột được tạo ra bởi thuật toán clustering để đánh giá độ phân bố của các điểm dữ liệu trong các nhóm.

Có thể thấy, kích thước gia đình phân bố đa số ở 3 thành viên có nghĩa là đa số là 1 cặp vợ chồng và có 1 con, số thành viên tối đa là 3, ít nhất là 1(trường hợp độc thân).

Tương tự như vậy ta có thấy, đa số các khách hàng đa số đều là cha hoặc mẹ.

Khách hàng có tới 1442 là những người có cặp/đôi, còn lại là 794 người là độc thân.

Đa số khách hàng mua hàng tại cửa hàng là 5 lần, cao nhất là 13 lần.

Tương tự số lần khách hàng mua qua web là 4 lần, cao nhất là 12 lần.

➤ Nhận xét từng cụm:

Ở đây thuật toán đã phân chia ra được 5 cụm dữ liệu, sau đây chúng em sẽ đánh giá kết quả của từng cụm:

- Clustering 1: Là nhóm có số lượng đông nhất với 767 khách hàng, có số lượng thành viên gia đình  $3,33+0,53$ , hoàn toàn là cha hoặc là mẹ, tình trạng hôn nhân bao gồm 84% là cặp/đôi, 15,6% là độc thân, số lần mua hàng tại cửa hàng là  $3,47+1,24$ , mua hàng qua trang web là  $2,48+1,61$ .
- Clustering 2: Là nhóm có số lượng nhiều thứ hai với 454 khách hàng, có số lượng thành viên gia đình  $3,24+0,47$ , hoàn toàn là cha hoặc là mẹ, tình trạng hôn nhân bao gồm 91,5% là cặp/đôi, 8,5% là độc thân, số lần mua hàng tại cửa hàng là  $8,57+2,66$ , mua hàng qua trang web là  $6,73+2,64$ .
- Clustersring 3: Là nhóm có số lượng đông thứ 3 với 383 khách hàng, có số lượng thành viên gia đình 2 thành viên, hoàn toàn không phải là cha hoặc mẹ, tình trạng hôn nhân bao gồm 100% là cặp/đôi, số lần mua hàng tại cửa hàng là  $7,20+3,26$ , mua hàng qua trang web là  $4,37+2,29$ .
- Clustersring 4: Là nhóm có số lượng với 378 khách hàng, có số lượng thành viên gia đình tất cả là 2 thành viên, hoàn toàn là cha hoặc là mẹ, tình trạng hôn nhân bao gồm 100% là độc thân, số lần mua hàng tại cửa hàng là  $5,48+3,15$ , mua hàng qua trang web là  $4,31+3,04$ .
- Clustersring 5: Là nhóm có số lượng ít nhất với 254 khách hàng, có số lượng thành viên gia đình 1 thành viên, hoàn toàn không phải cha hoặc là mẹ, tình trạng hôn nhân bao gồm 100% là độc thân, số lần mua hàng tại cửa hàng là  $7,37+3,34$ , mua hàng qua trang web là  $4,43+2,76$ .

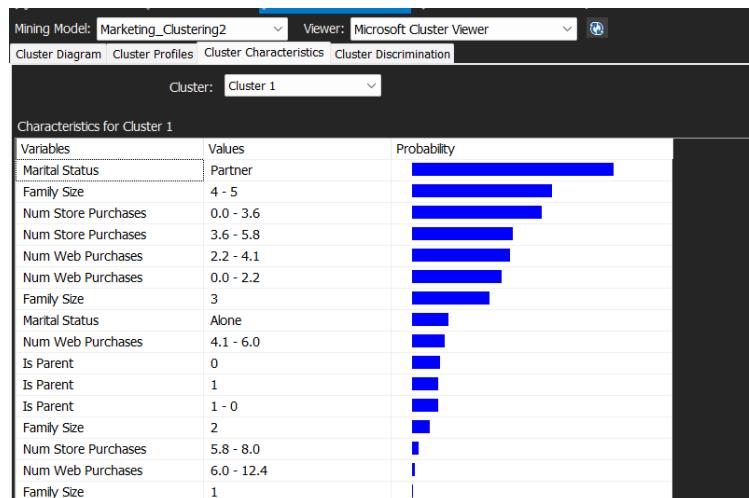
➤ Gán nhãn cho từng cụm khách hàng:

- Cụm 3: Ta có thể gán nhãn cho cụm này là “Khách hàng tiềm năng”, vì họ có số lần mua hàng tại cửa hàng và qua trang web cao, và họ là những cặp đôi vì có số thành viên là 2.

- Cụm 2: Ta có thể gán nhãn cho cụm này là “Khách hàng ưu tú”, vì họ có số lần mua hàng tại cửa hàng và qua trang web ở mức cao, và họ là cha hoặc mẹ và có số lượng thành viên gia đình cáo từ 3-5 thành viên.
- Cụm 1: Ta có thể gán nhãn cho cụm này là “Khách hàng kém”, vì họ có số lần mua hàng tại cửa hàng và qua trang web thấp nhất, và họ là cha hoặc mẹ, và họ có đến 3-4 thành viên gia đình.
- Cụm 4: Ta có thể gán nhãn cho cụm này là “Khách hàng trung bình”, vì họ có số lần mua hàng tại cửa hàng và qua trang web ở mức trung bình, và họ là cha hoặc mẹ, và họ cũng là độc thân.
- Cụm 5: Ta có thể gán nhãn cho cụm này là “Khách hàng tốt”, vì họ có số lần mua hàng tại cửa hàng cao nhưng qua trang web trung bình, và họ cũng không phải là cha hoặc mẹ, và họ cũng là độc thân.

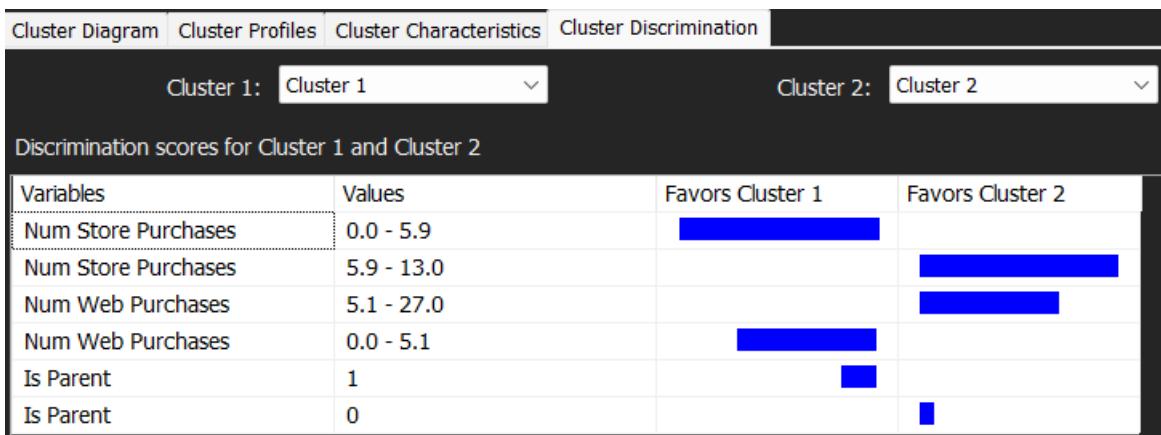
### 2.2.2. Cluster Characteristics

Xem chi tiết từng cụm trong tab Cluster Characteristics



### 2.2.3. Cluster Discrimination

So sánh 2 cụm, cụ thể so sánh giữa hai cụm 1 và 2



#### 2.2.4. Predict

Tiến hành phân cụm một 1 người dựa trên các thuộc tính. Ví dụ 1 người là cha/ mẹ, số thành viên trong gia đình là 3, tình trạng kết hôn là Partner, số lượng sản phẩm mà khách hàng đã mua tại cửa hàng của công ty trong 2 năm qua là 3 và số lần truy cập trang web trong 1 tháng là 5.

The screenshot shows the Microsoft Data Mining Modeler interface. On the left, there is a tree view of the mining model "Marketing\_Clustering2" with nodes for Family Size, ID, Is Parent, Marital Status, Num Store Purchases, and Num Web Purchases. Below the tree is a button "Select Model...". To the right, there is a "Singleton Query Input" window containing a table with columns "Mining Model Column" and "Value". The table has rows for Family Size (3), Is Parent (1), Marital Status (Partner), Num Store Purchases (3), and Num Web Purchases (5). At the bottom, there is a table with columns "Source", "Field", "Alias", "Show", "Group", "And/Or", and "Criteria/Argument". The "Source" column has "Prediction Fun..." and "Cluster". The "Field" column has "Cluster". The "Alias" column has "Predict Cluster". The "Show" column has a checked checkbox. The "Group" and "And/Or" columns have dropdown menus. The "Criteria/Argument" column has an empty checkbox.

Source	Field	Alias	Show	Group	And/Or	Criteria/Argument
Prediction Fun...	Cluster	Predict Cluster	<input checked="" type="checkbox"/>			<input type="checkbox"/>

Ấn vào result, kết quả người đó thuộc cụm 1.

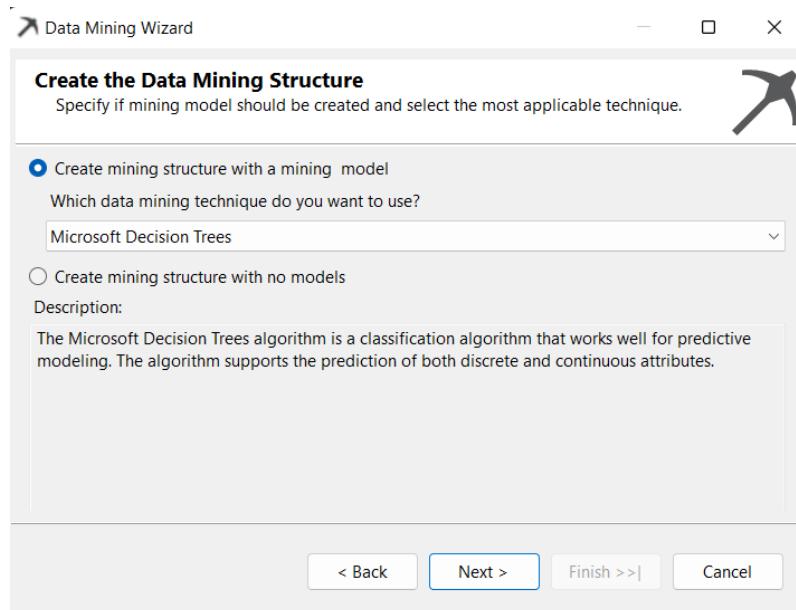


### 3. Thuật toán Decision Tree

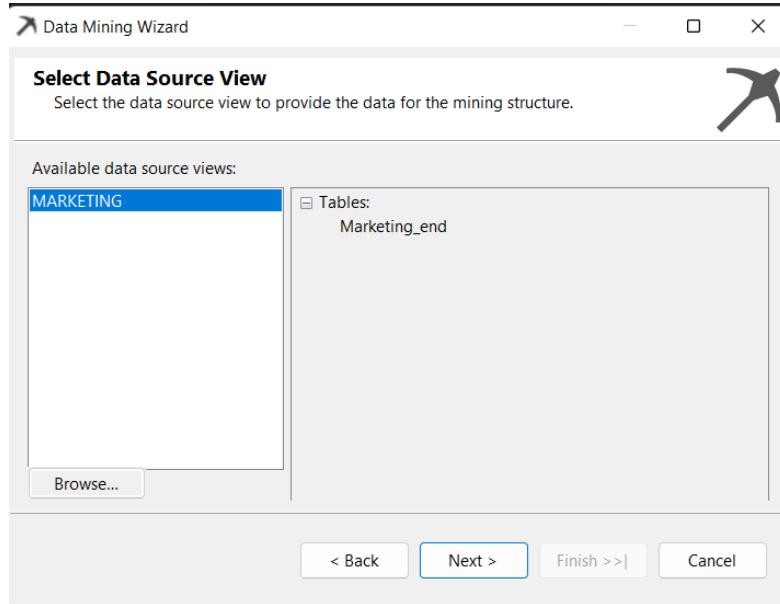
Sử dụng thuật toán decision tree để xây dựng các luật IF - THEN để dự đoán khả năng người đó có phản hồi lại chiến dịch tiếp thị mới nhất hay không.

- Các bước thực hiện:

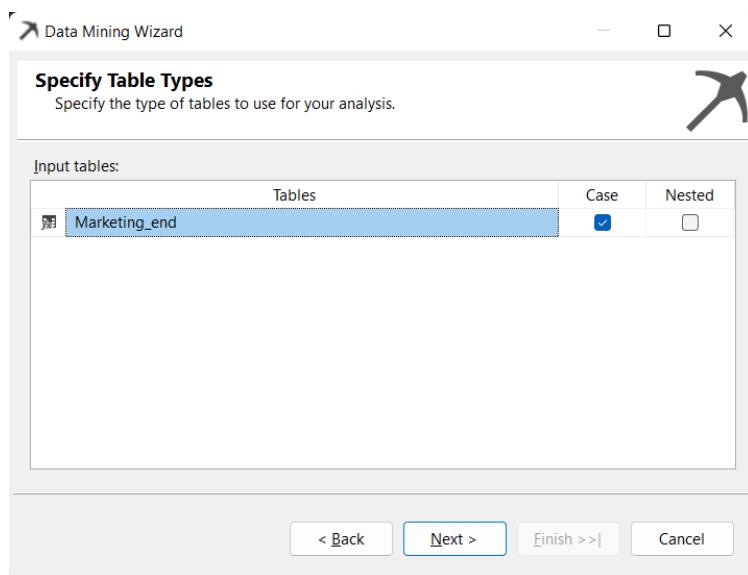
Tạo mới một New Mining Structure, sử dụng thuật toán decision tree



Chọn data source, click next



Click next



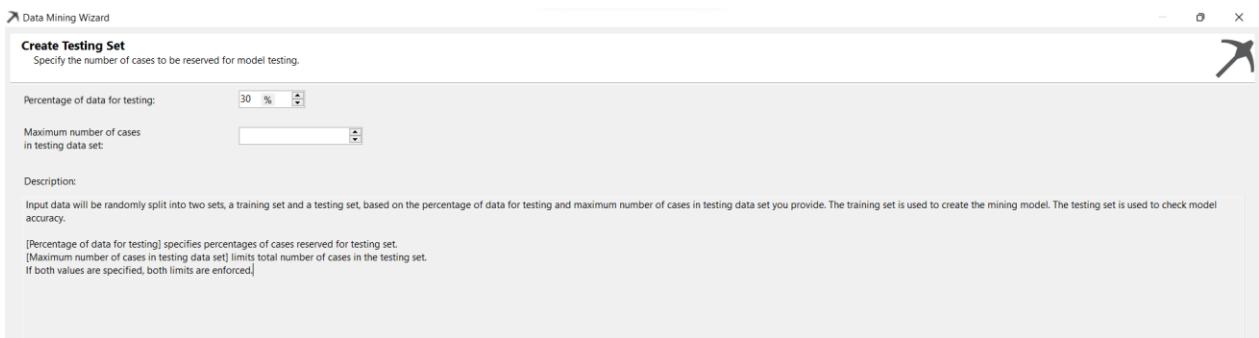
Chọn key, input và biến dự đoán là response

	Tables/Columns	Key	Input	Predict...
-	Marketing_end			
<input checked="" type="checkbox"/>	AcceptedCmp1		<input checked="" type="checkbox"/>	
<input checked="" type="checkbox"/>	AcceptedCmp2		<input checked="" type="checkbox"/>	
<input checked="" type="checkbox"/>	AcceptedCmp3		<input checked="" type="checkbox"/>	
<input checked="" type="checkbox"/>	AcceptedCmp4		<input checked="" type="checkbox"/>	
<input checked="" type="checkbox"/>	AcceptedCmp5		<input checked="" type="checkbox"/>	
<input type="checkbox"/>	Age			
<input type="checkbox"/>	Children			
<input type="checkbox"/>	Column 0			
<input type="checkbox"/>	Complain			
<input type="checkbox"/>	Country			
<input type="checkbox"/>	Education			
<input type="checkbox"/>	Family_Size			
<input checked="" type="checkbox"/>	Fish			
<input checked="" type="checkbox"/>	Fruits			
<input checked="" type="checkbox"/>	Gold			
<input checked="" type="checkbox"/>	ID			
<input checked="" type="checkbox"/>	Income			
<input type="checkbox"/>	IncomeFilter			
<input type="checkbox"/>	Is_Parent			
<input type="checkbox"/>	Marital_Status			
<input checked="" type="checkbox"/>	Meat			
<input type="checkbox"/>	NumCatalogPurchases			
<input type="checkbox"/>	NumDealsPurchases			
<input type="checkbox"/>	NumStorePurchases			

Click next, thay đổi các kiểu thuộc tính

	Columns	Content Type	Data Type
<input checked="" type="checkbox"/>	Accepted Cmp1	Discrete	Text
<input checked="" type="checkbox"/>	Accepted Cmp2	Discrete	Text
<input checked="" type="checkbox"/>	Accepted Cmp3	Discrete	Text
<input checked="" type="checkbox"/>	Accepted Cmp4	Discrete	Text
<input checked="" type="checkbox"/>	Accepted Cmp5	Discrete	Text
<input checked="" type="checkbox"/>	Fish	Continuous	Double
<input checked="" type="checkbox"/>	Fruits	Continuous	Double
<input checked="" type="checkbox"/>	Gold	Continuous	Double
<input checked="" type="checkbox"/>	ID	Key	Text
<input checked="" type="checkbox"/>	Income	Continuous	Double
<input type="checkbox"/>	IncomeFilter	Continuous	Double
<input type="checkbox"/>	Is_Parent	Discrete	Text
<input type="checkbox"/>	Marital_Status	Continuous	Double
<input checked="" type="checkbox"/>	Meat	Continuous	Double
<input type="checkbox"/>	NumCatalogPurchases	Continuous	Double
<input type="checkbox"/>	NumDealsPurchases	Continuous	Double
<input type="checkbox"/>	NumStorePurchases	Continuous	Double
			Double

Chia tập dữ liệu train và test



Tiến hành đổi tên và finish

↗ Data Mining Wizard

### Completing the Wizard

Completing the Data Mining Wizard by providing a name for the mining structure.

---

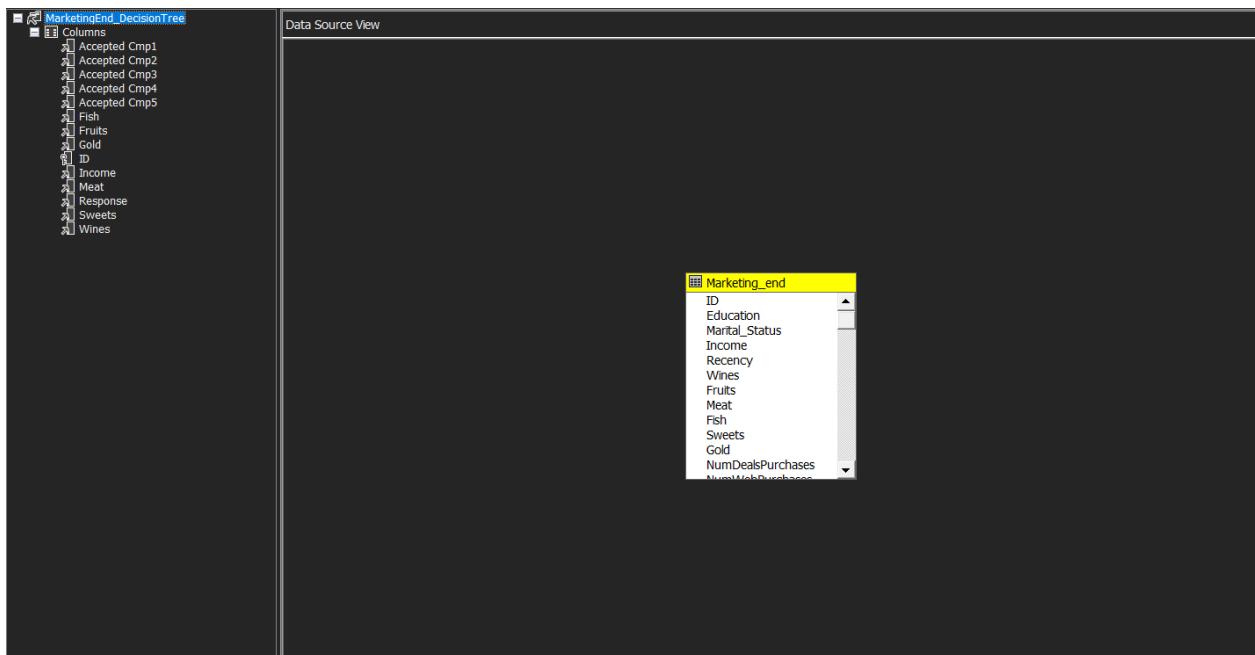
Mining structure name:

MarketingEnd\_DecisionTree

Mining model name:

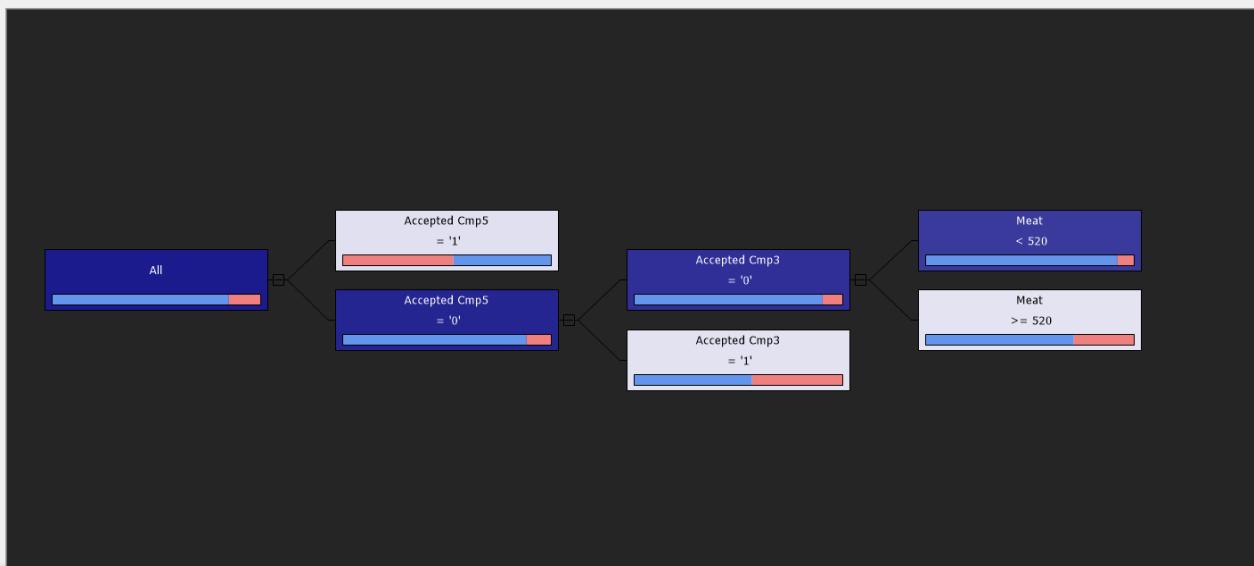
MarketingEnd\_DecisionTree

Kết quả:



The screenshot shows the 'Data Source View' window for the 'MarketingEnd\_DecisionTree' mining structure. On the left, the 'Columns' pane lists various columns: Accepted Cmp1, Accepted Cmp2, Accepted Cmp3, Accepted Cmp4, Accepted Cmp5, Fish, Fruits, Gold, ID, Income, Meat, Response, Sweets, and Wines. On the right, the 'Marketing\_end' data source is expanded, showing its columns: ID, Education, Marital\_Status, Income, Recency, Wines, Fruits, Meat, Fish, Sweets, Gold, NumDealsPurchases, and NumMalsPurchases. The 'Marketing\_end' data source is highlighted with a yellow background.

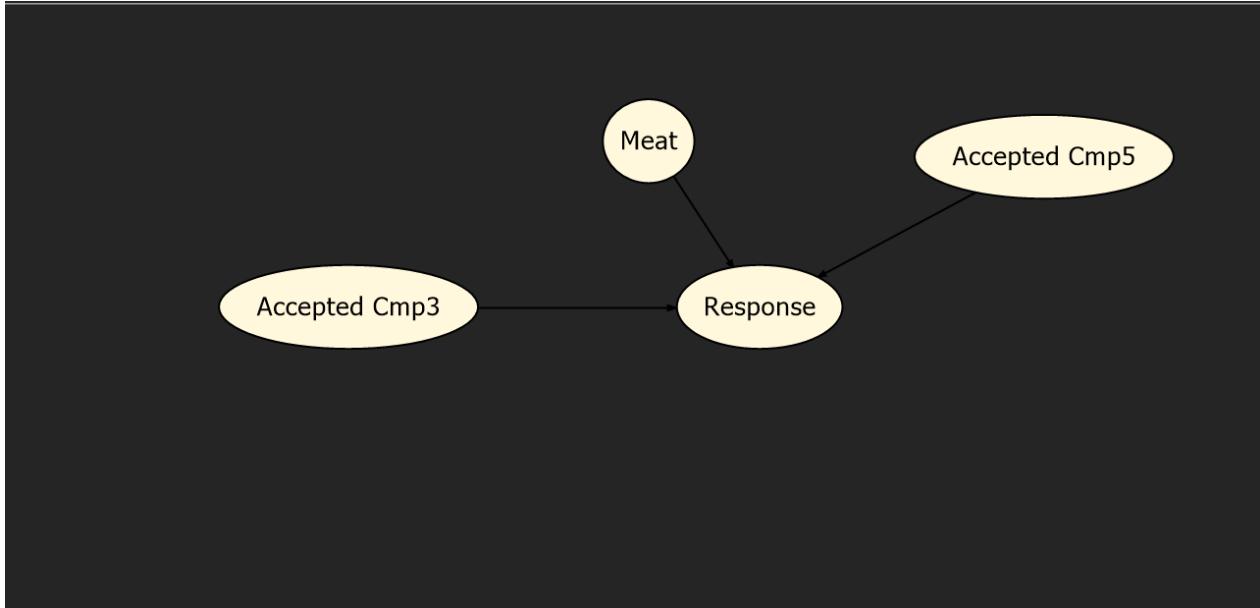
## Kết quả sau khi process model



➤ Từ kết quả trên ta có thể gom lại được 1 số luật như sau:

- Đối với khách hàng phản hồi với chiến dịch tiếp thị (response = 1)
- Nếu khách hàng không chấp nhận tham gia chiến dịch tiếp thị 5 nhưng tham gia vào chiến dịch tiếp thị 3 thì khả năng người đó phản hồi lại chiến dịch tiếp thị mới nhất đạt gần 44%.
  - Đối với khách hàng không phản hồi với chiến dịch tiếp thị (response = 0)
- Đối với khách hàng không chấp nhận tham gia chiến dịch tiếp thị 5 và chiến dịch tiếp thị 3 và số tiền bỏ ra mua thịt nhỏ hơn 520\$ trong 2 năm thì khả năng người đó không phản hồi lại chiến dịch mới đạt khoảng 91,5%.
- Cũng tương tự khách hàng không chấp nhận tham gia chiến dịch tiếp thị 5 và chiến dịch tiếp thị 3 nhưng số tiền chi ra mua thịt lớn hơn 520\$ trong 2 năm thì khả năng người đó không phản hồi lại chiến dịch đạt khoảng 70,5%.

Trong số các biến dự đoán, ta thấy 3 biến Accepted Cmp3, Accepted Cmp5 và Meat quan trọng và ảnh hưởng nhất đến biến kết quả.



Qua đó ta có thể thấy chiến dịch tiếp thị số 3, chiến dịch tiếp thị số 5 và số tiền chi ra để mua thịt ảnh hưởng rất nhiều đến việc người đó có phản hồi lại chiến dịch tiếp thị mới nhất hay không. Việc tập trung vào 2 chiến dịch này có thể thu hút thêm được nhiều khách hàng hơn.

Thực hiện đánh giá mô hình dựa vào confusion matrix

Predicted	1 (Actual)	0 (Actual)
1	28	13
0	67	562

Độ chính xác của mô hình đạt 88% .

Thực hiện dự đoán xem người đó có phản hồi lại chiến dịch tiếp thị dựa trên các input đầu vào. Các thông số được nhập vào như hình dưới.

Mining Model

- MarketingEnd\_DcisionTree
  - Accepted Cmp1
  - Accepted Cmp2
  - Accepted Cmp3
  - Accepted Cmp4
  - Accepted Cmp5
  - Fish
  - Fruits
  - Gold
  - ID
  - Income
  - Meat
  - Response
  - Sweets

Select Model...

Singleton Query Input

Mining Model Column	Value
Accepted Cmp1	1
Accepted Cmp2	0
Accepted Cmp3	1
Accepted Cmp4	0
Accepted Cmp5	1
Fish	100
Fruits	8
Gold	250
Income	60000
Meat	600
Response	
Sweets	100
Wines	400

Source Field Alias Show Group And/Or Criteria/Argument

MarketingEnd_DcisionTree	Response	Predict Value	<input checked="" type="checkbox"/>		[MarketingEnd_DcisionTree].[Response]
Prediction Function	PredictProbability	Predict Probabi...	<input checked="" type="checkbox"/>	<input type="checkbox"/>	

Predict Value Predict Prob...

1	0.5277044...
---	--------------

Kết quả xác suất 52% người đó phản hồi lại chiến dịch tiếp thị mới nhất. Kết quả này đúng với dự đoán của nhóm. Vì ta thấy người này có chấp nhận tham gia cả 3 chiến dịch là 1,3,5 và số tiền chi ra mua thịt > 600\$ nên khả năng cao lớn hơn 50% người đó phản hồi lại chiến dịch tiếp thị mới nhất.

#### 4. Thuật toán Association Rule

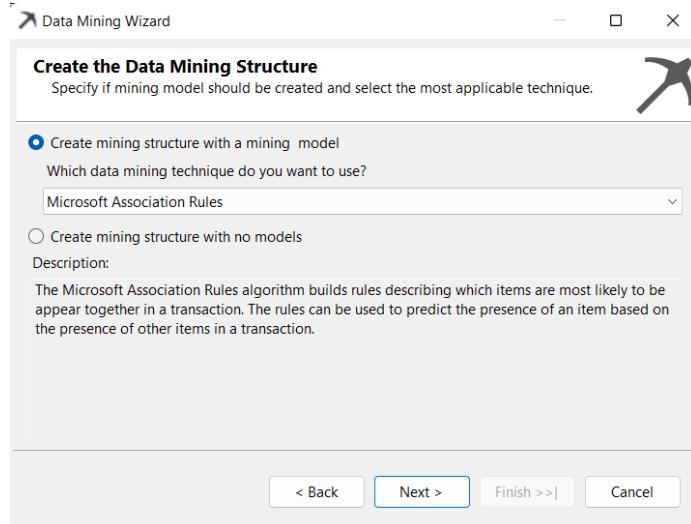
Thực hiện tìm ra các mối quan hệ tương quan giữa các đối tượng trong tập dữ liệu. Một luật kết hợp X kéo theo Y sẽ gồm 2 thành phần:

X : antecedent (if) và Y : consequent (then).

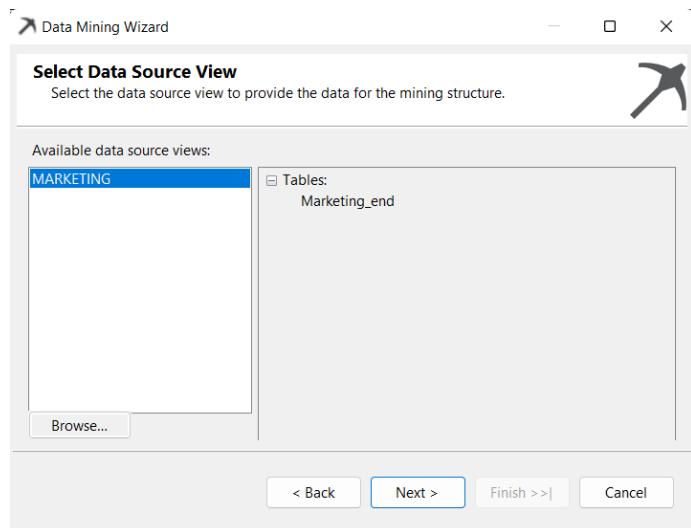
Nhóm sử dụng các biến AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, Fish, Fruits, Gold, Meat, Sweets, Wines cho mệnh đề if và biến response cho mệnh đề Y.

➤ Các bước thực hiện

Tạo mới một mining structure:



Click next



Click next

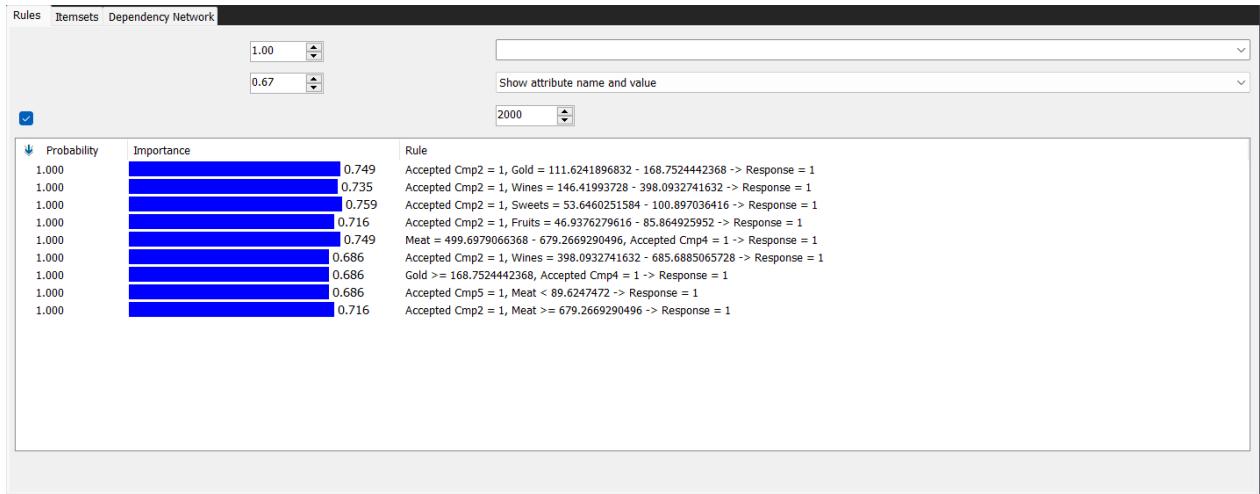
	Tables/Columns	Key	Input	Predict...
<input type="checkbox"/>	Column 0			
<input type="checkbox"/>	Complain			
<input type="checkbox"/>	Country			
<input type="checkbox"/>	Education			
<input type="checkbox"/>	Family_Size			
<input checked="" type="checkbox"/>	Fish			
<input checked="" type="checkbox"/>	Fruits			
<input checked="" type="checkbox"/>	Gold			
<input checked="" type="checkbox"/>	ID			
<input type="checkbox"/>	Income			
<input type="checkbox"/>	Incomefilter			
<input type="checkbox"/>	Is_Parent			
<input type="checkbox"/>	Marital_Status			
<input checked="" type="checkbox"/>	Meat			
<input type="checkbox"/>	NumCatalogPurchases			
<input type="checkbox"/>	NumDealsPurchases			
<input type="checkbox"/>	NumStorePurchases			
<input type="checkbox"/>	NumWebPurchases			
<input type="checkbox"/>	NumWebVisitsMonth			
<input type="checkbox"/>	Recency			
<input checked="" type="checkbox"/>	Response			
<input type="checkbox"/>	Spent			
<input checked="" type="checkbox"/>	Sweets			
<input checked="" type="checkbox"/>	Wines			

Click next, chuyển đổi các kiểu dữ liệu sang dạng phù hợp

	Columns	Content Type	Data Type
<input checked="" type="checkbox"/>	Accepted Cmp1	Discrete	Text
<input checked="" type="checkbox"/>	Accepted Cmp2	Discrete	Text
<input checked="" type="checkbox"/>	Accepted Cmp3	Discrete	Text
<input checked="" type="checkbox"/>	Accepted Cmp4	Discrete	Text
<input checked="" type="checkbox"/>	Accepted Cmp5	Discrete	Text
<input checked="" type="checkbox"/>	Fish	Discretized	Double
<input checked="" type="checkbox"/>	Fruits	Discretized	Double
<input checked="" type="checkbox"/>	Gold	Discretized	Double
<input checked="" type="checkbox"/>	ID	Key	Text
<input checked="" type="checkbox"/>	Meat	Discretized	Double
<input checked="" type="checkbox"/>	Response	Discrete	Text
<input checked="" type="checkbox"/>	Sweets	Discretized	Double
<input checked="" type="checkbox"/>	Wines	Discretized	Double

Đổi tên và click finish

➤ Chuyển qua tab Rules trong phần Mining Model Viewer



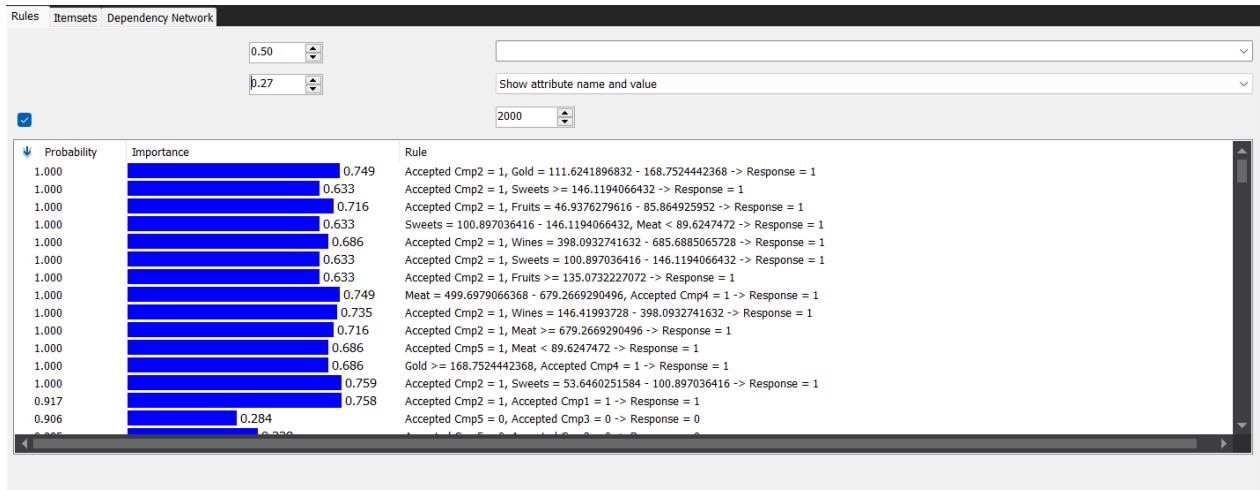
Giá trị Probability là một số từ 0-1 thể hiện mức độ có thể xảy ra của 1 sự kiện. Probability thể hiện xác suất của một itemset (tập hợp các item) xuất hiện trong tập dữ liệu.

Giá trị Importance (tầm quan trọng) là một đánh giá về sự quan trọng của một association rule. Có nhiều phương pháp để tính toán độ quan trọng của một rule, nhưng phổ biến nhất là dựa trên các thông số như support (tần suất xuất hiện của itemset), confidence (độ tin cậy của rule) và lift (tính tương quan giữa các item). Các rule có giá trị importance cao thường được coi là những rule có tính khả thi và ý nghĩa cao hơn trong việc khai thác tri thức từ tập dữ liệu.

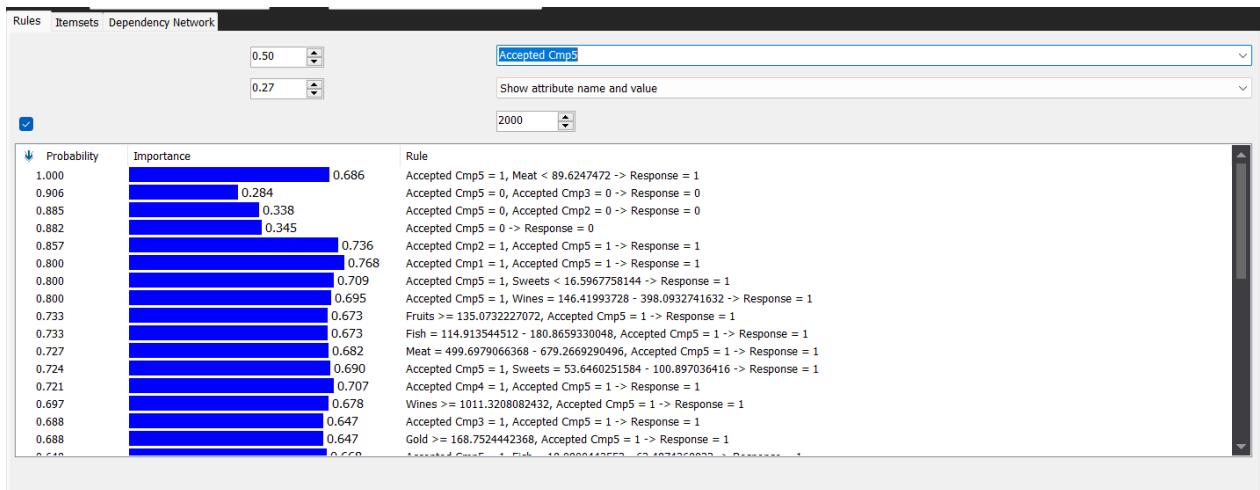
Từ các luật trên, ta có rút ra được 1 số luật quan trọng và có thể diễn giải như sau:

- 100% những người tham gia vào chiến dịch số 2 và số tiền chi để mua vàng nằm trong khoảng 111\$ - 168\$ thì người đó sẽ phản hồi lại chiến dịch mới nhất với giá trị Importance đạt 0,749.
- 100% những người tham gia vào chiến dịch số 2 và số tiền chi để mua vàng nằm trong khoảng 146\$ - 398\$ thì người đó sẽ phản hồi lại chiến dịch mới nhất với giá trị Importance đạt 0,735.
- Tương tự với các luật còn lại.

Ta cũng có thể thay đổi các giá trị Minimum probability và Minimum importance



Ta cũng có thể điền vào ô filter để lọc ra các rule liên quan đến điều kiện đó, ở đây nhóm lọc theo các rule có chứa biến Accepted Cmp5



### ➤ Chuyển qua tab Itemsets

Ở đây ta có thể điều chỉnh các giá trị Minimum support, Minimum itemset size, Maximum rows theo các giá trị mình mong muốn

Rules   Itemsets   Dependency Network

2   ▲   ▼

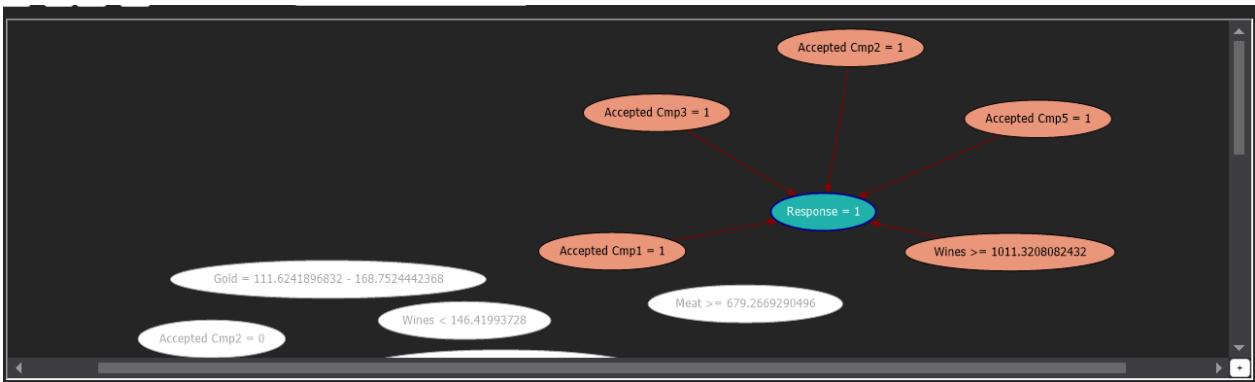
3   ▲   ▼   Show attribute name and value

2000   ▲   ▼

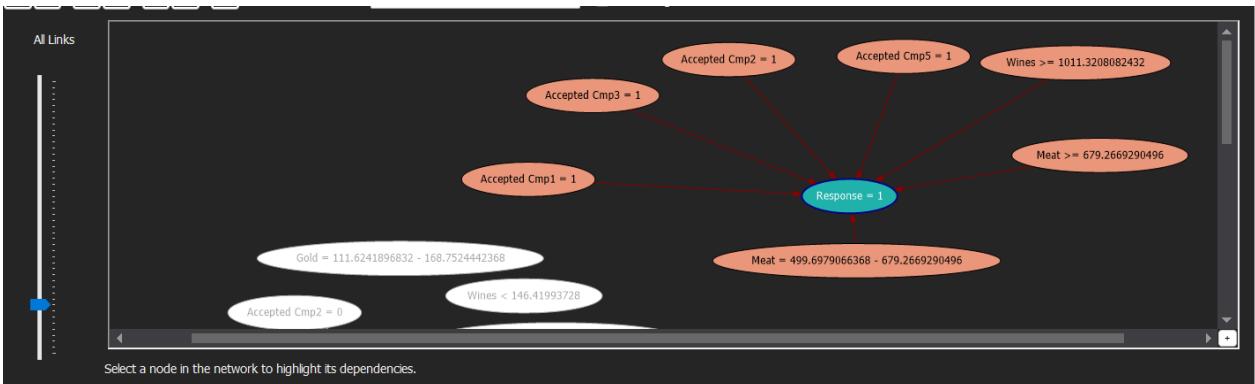
	Support	Size	Itemset
1387	3	Accepted Cmp5 = 0, Accepted Cmp1 = 0, Accepted Cmp2 = 0	
1374	3	Accepted Cmp4 = 0, Accepted Cmp1 = 0, Accepted Cmp2 = 0	
1366	3	Accepted Cmp5 = 0, Accepted Cmp4 = 0, Accepted Cmp2 = 0	
1359	3	Accepted Cmp3 = 0, Accepted Cmp1 = 0, Accepted Cmp2 = 0	
1338	3	Accepted Cmp5 = 0, Accepted Cmp3 = 0, Accepted Cmp2 = 0	
1331	3	Accepted Cmp4 = 0, Accepted Cmp3 = 0, Accepted Cmp2 = 0	
1326	3	Accepted Cmp5 = 0, Accepted Cmp4 = 0, Accepted Cmp1 = 0	
1304	3	Accepted Cmp5 = 0, Accepted Cmp3 = 0, Accepted Cmp1 = 0	
1279	3	Accepted Cmp4 = 0, Accepted Cmp3 = 0, Accepted Cmp1 = 0	
1273	3	Response = 0, Accepted Cmp1 = 0, Accepted Cmp2 = 0	
1270	3	Accepted Cmp5 = 0, Accepted Cmp4 = 0, Accepted Cmp3 = 0	
1270	3	Response = 0, Accepted Cmp5 = 0, Accepted Cmp2 = 0	
1257	3	Response = 0, Accepted Cmp3 = 0, Accepted Cmp2 = 0	
1245	3	Response = 0, Accepted Cmp4 = 0, Accepted Cmp2 = 0	
1241	3	Response = 0, Accepted Cmp5 = 0, Accepted Cmp1 = 0	

### ➤ Chuyển qua tab Dependency Network

Mô tả mức độ phụ thuộc của các item trong 1 luật kết hợp và mô tả theo độ mạnh. Càng tăng mức thanh kéo thì độ phụ thuộc giảm dần.



Nhìn vào kết quả trên, ta thấy các biến Accepted Cmp5, Accepted Cmp2, Accepted Cmp1, Accepted Cmp3 và Wines lần lượt ảnh hưởng đến biến response.



Dựa vào cột All Links mà ta có thể biết được các luật phụ thuộc với nhau như thế nào và độ mạnh của nó.

### ➤ Chuyển sang Mining Model Prediction

Tiến hành dự đoán một người nào đó có phản hồi lại chiến dịch tiếp thị mới nhất không dựa vào các input đầu vào.

The screenshot shows the Microsoft Data Mining tool interface. On the left, a tree view titled "Mining Model" shows nodes like "MarketingEnd\_AssociationRule", "Accepted Cmp1" through "Accepted Cmp5", "Fish", "Fruits", "Gold", "ID", "Meat", "Response", "Sweets", and "Wines". Below the tree is a button "Select Model...". To the right, a window titled "Singleton Query Input" displays a table of mining model columns and their values:

Mining Model Column	Value
Accepted Cmp1	1
Accepted Cmp2	0
Accepted Cmp3	0
Accepted Cmp4	0
Accepted Cmp5	0
Fish	18.0900443552 - 63.4874368832
Fruits	46.9376279616 - 85.864925952
Gold	111.6241896832 - 168.7524442368
Meat	499.6979066368 - 679.2669290496
Response	146.1194066432 - 198
Sweets	0 - 146.41993728
Wines	0 - 146.41993728

At the bottom, there is a table titled "Source" with columns "Source", "Field", "Alias", "Show", "Group", "And/Or", and "Criteria/Argument". It contains two rows:

Source	Field	Alias	Show	Group	And/Or	Criteria/Argument
MarketingEnd_AssociationRule	Response		<input checked="" type="checkbox"/>			[MarketingEnd_AssociationRule].[Response]
Prediction Function	PredictProbability	Predict Probabi...	<input checked="" type="checkbox"/>			

Trước khi nhấn result, nhóm dự đoán rằng người này sẽ không phản hồi bởi vì người này chỉ có tham gia một chiến dịch tiếp thị số 1 và số tiền chi ra mua rượu nhỏ hơn 146\$. Sau khi click result:

Response	Predict Probability
0	0.844827586206897

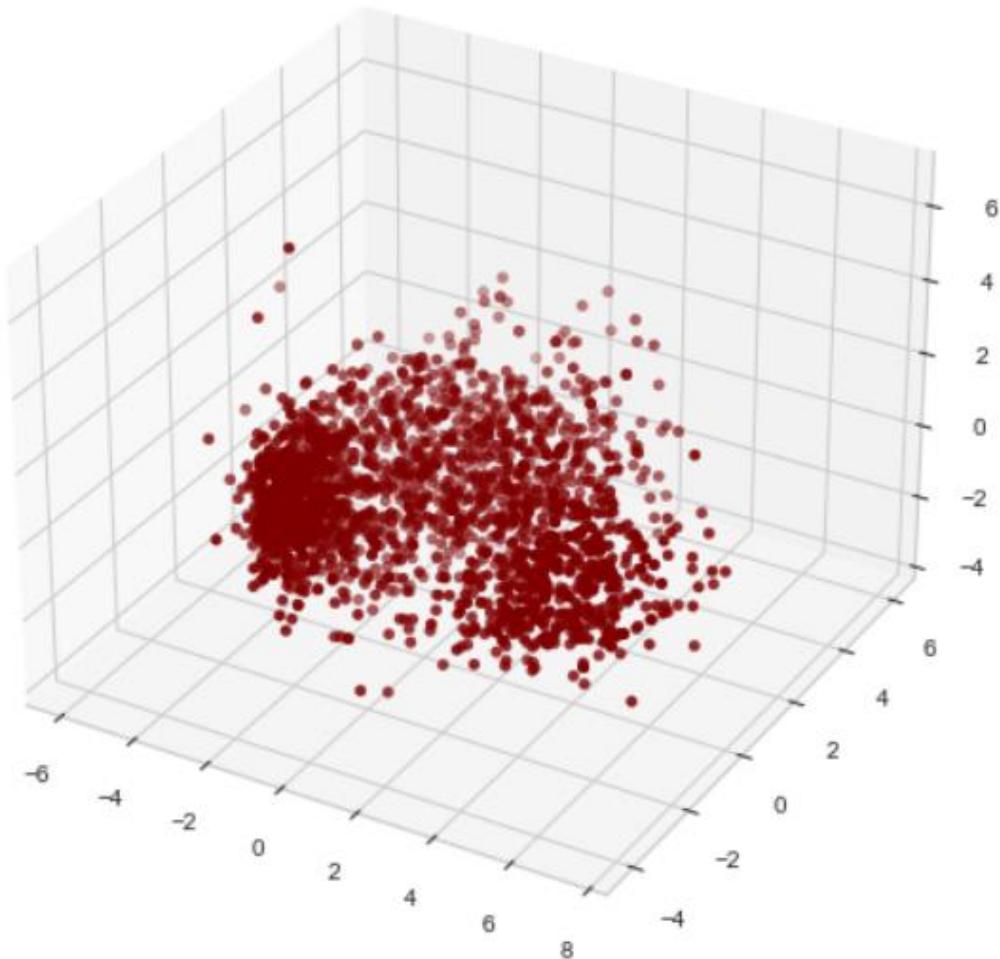
Kết quả đúng như nhóm dự đoán là 0, và xác suất cao đạt 84%.

## 5. Đánh giá thực nghiệm và trực quan hóa dữ liệu:

Sau khi khai phá trên các thuật toán trên SSAS, nhóm sẽ tiến hành phân tích thực nghiệm lại trên toàn bộ tập dữ liệu trên 2236 quan sát.

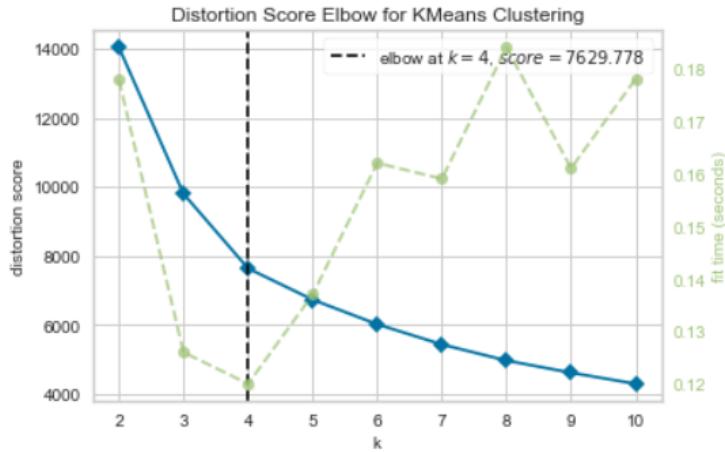
- Trực quan toàn bộ các quan sát:

A 3D Projection Of Data In The Reduced Dimension



- Tiến hành phân cụm lại trên toàn bộ cụm dữ liệu:

Elbow Method to determine the number of clusters to be formed:

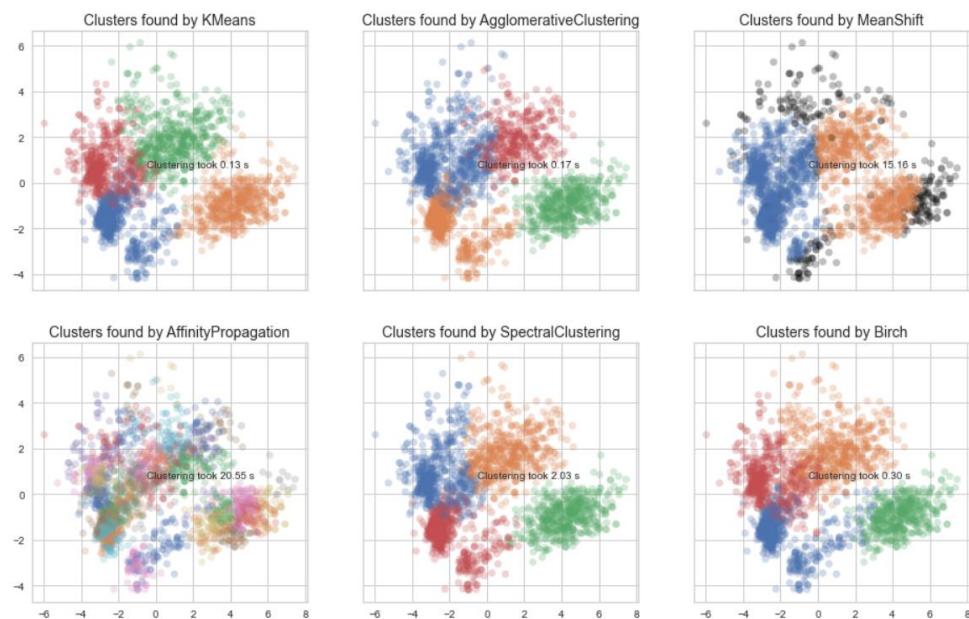


Sau khi chạy phương pháp elbow lại, số cụm phù hợp cho ra là k=4.

- Chạy phân cụm trên 6 thuật toán

Tiến hành chạy phân cụm trên nhiều thuật toán, để đánh giá kết quả phân cụm và lựa chọn thuật toán phù hợp nhất.

Dánh giá kết quả theo điểm Silhouette.

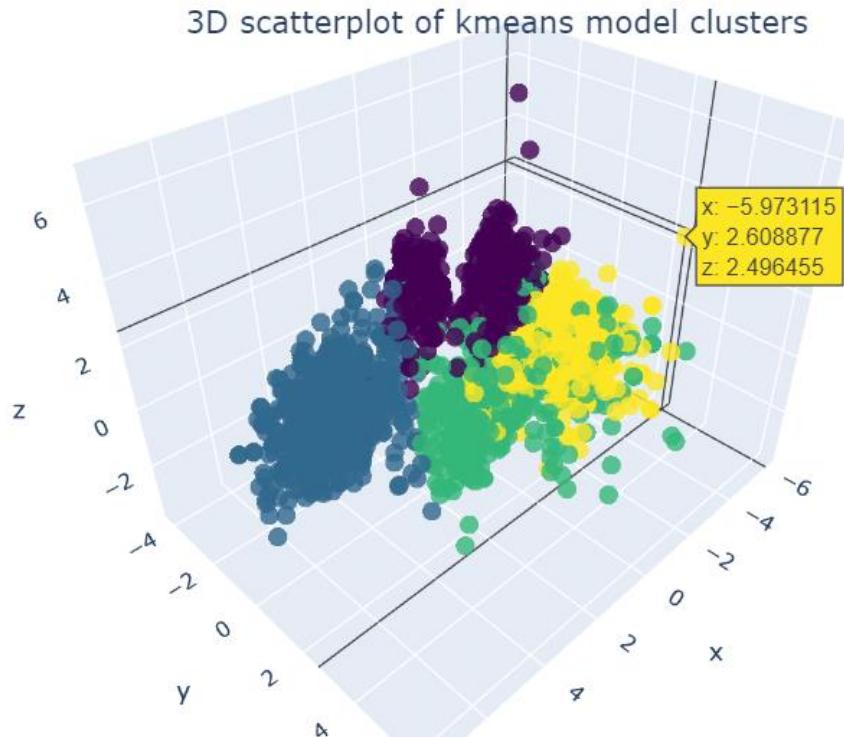


```
Silhouette score for KMeans: 0.374
Silhouette score for AgglomerativeClustering: 0.333
Silhouette score for MeanShift: 0.338
Silhouette score for AffinityPropagation: 0.270
Silhouette score for SpectralClustering: 0.367
Silhouette score for Birch: 0.347
```

Kết quả cho thấy thuật toán k-means là phân cụm nhanh và tốt nhất. Vì vậy nhóm sẽ chạy thuật toán K-means cho bài này.

#### ➤ Trực quan kết quả phân cụm

Phần này trực quan để có cái nhìn trực quan và hình dung về các nhóm khách hàng phân bố.



### ➤ Đánh giá mô hình

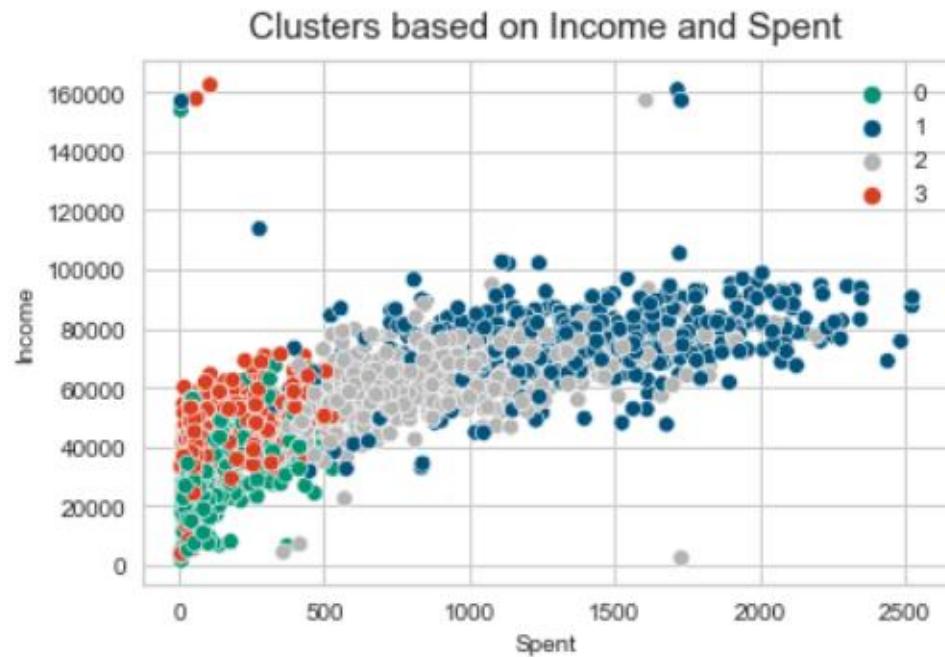
Vì đây là quá trình phân cụm không giám sát, nên không có đặc trưng được gán nhãn để sử dụng cho việc đánh giá hoặc tính điểm mô hình của chúng ta. Thay vào đó, mục tiêu của phần này là khám phá các mẫu trong các cụm được hình thành và hiểu bản chất của chúng. Để làm được điều này, chúng ta sẽ tiến hành phân tích dữ liệu khám phá để xem xét dữ liệu liên quan đến các cụm và rút ra kết luận từ những quan sát của chúng ta. Để bắt đầu, chúng ta sẽ xem xét phân bố của các nhóm trong các cụm.

### ➤ Sự phân bố của các cụm



Các cụm này cho thấy một phân bố tương đối đều trên tập dữ liệu.

➤ Đánh giá kết quả phân cụm dựa trên Spent và Income:

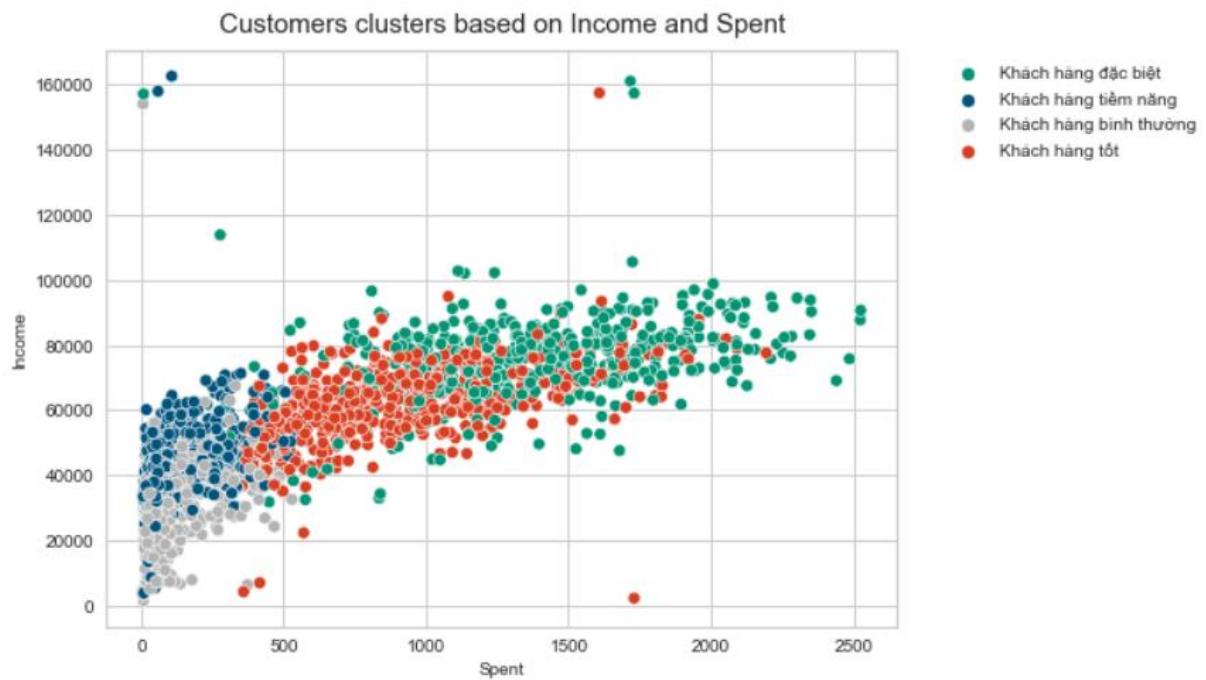


Như đã trình bày ở trên, Thu nhập và Chi tiêu có mối quan hệ mạnh mẽ. Các cụm cho thấy những thông tin chi tiết hơn về các mẫu của chúng:

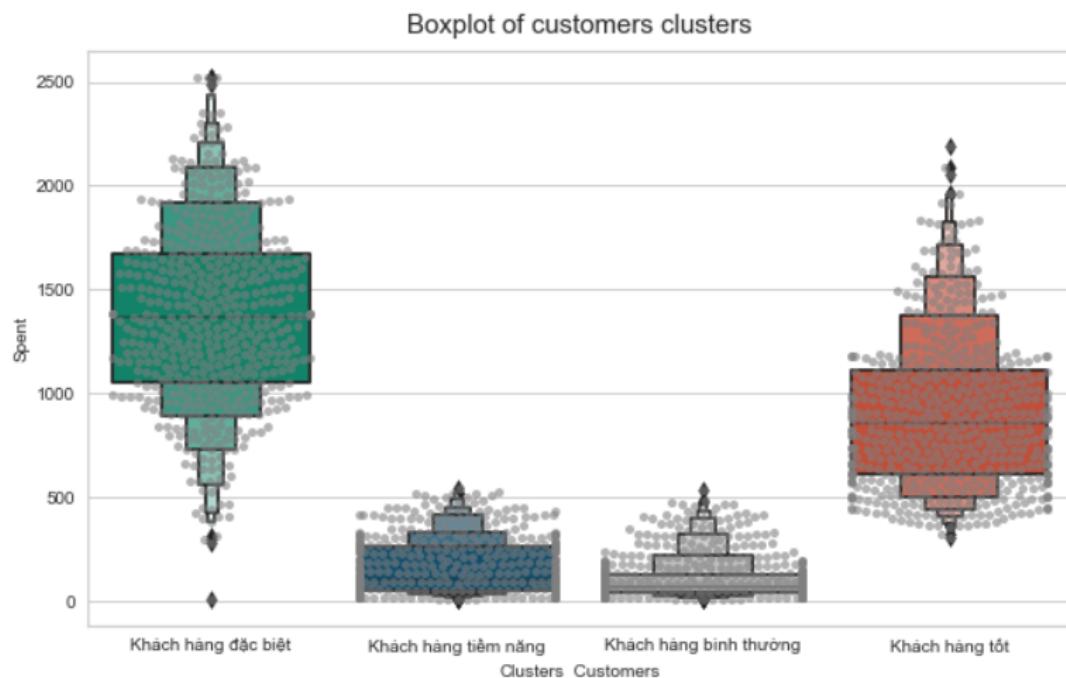
- \* Nhóm 0: chi tiêu thấp và thu nhập thấp
- \* Nhóm 3: chi tiêu thấp và thu nhập trung bình
- \* Nhóm 2: chi tiêu cao và thu nhập trung bình
- \* Nhóm 1: chi tiêu cao và thu nhập cao

Để dễ hiểu hơn, chúng ta sẽ phân loại nhóm 0 là khách hàng bình thường, nhóm 1 là khách hàng đặc biệt, nhóm 2 là khách hàng tốt, và nhóm 3 là khách hàng có tiềm năng.

Kết quả sau khi gán nhãn lại:



➤ Đánh giá phân cụm dựa trên mức chi tiêu:



Biểu đồ hộp cho thấy Khách hàng Đặc biệt và Khách hàng Tốt chi tiêu đáng kể hơn mỗi khách hàng, trung bình là 1400 bảng và 800 bảng tương ứng, so với Khách hàng Bình thường và Khách hàng Có tiềm năng có mức chi tiêu chỉ khoảng 500 bảng. Do đó, các chiến dịch tiếp thị khác nhau sẽ được phát triển cho mỗi nhóm. Để bắt đầu quá trình này, chúng ta sẽ xem xét các chiến dịch tiếp thị trước đây để xem mỗi nhóm đã chi bao nhiêu cho các chiến lược tiếp thị được nhắm mục tiêu.

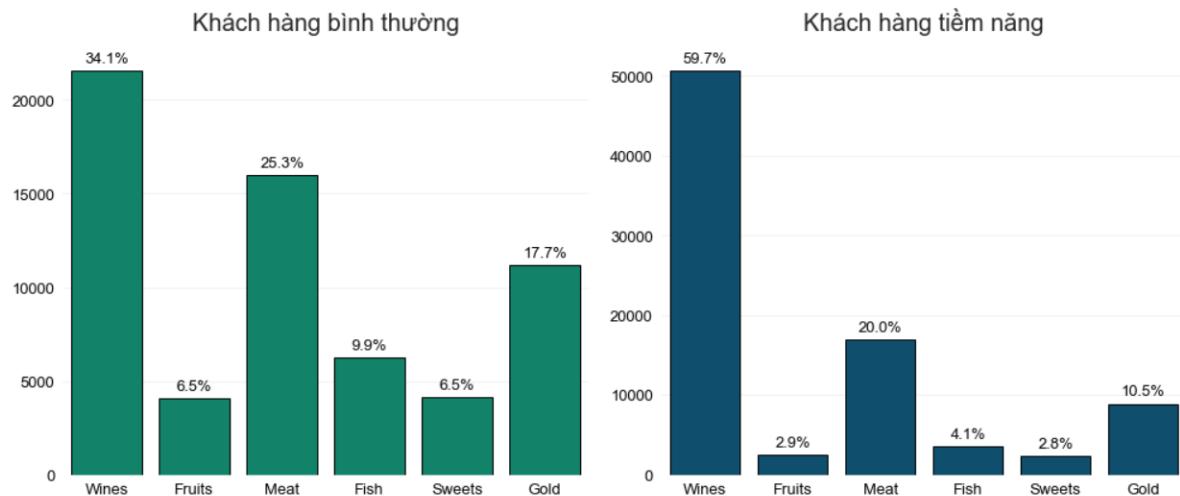
➤ Đánh giá phân cụm khách hàng dựa trên tổng chi tiêu cho từng mặt hàng

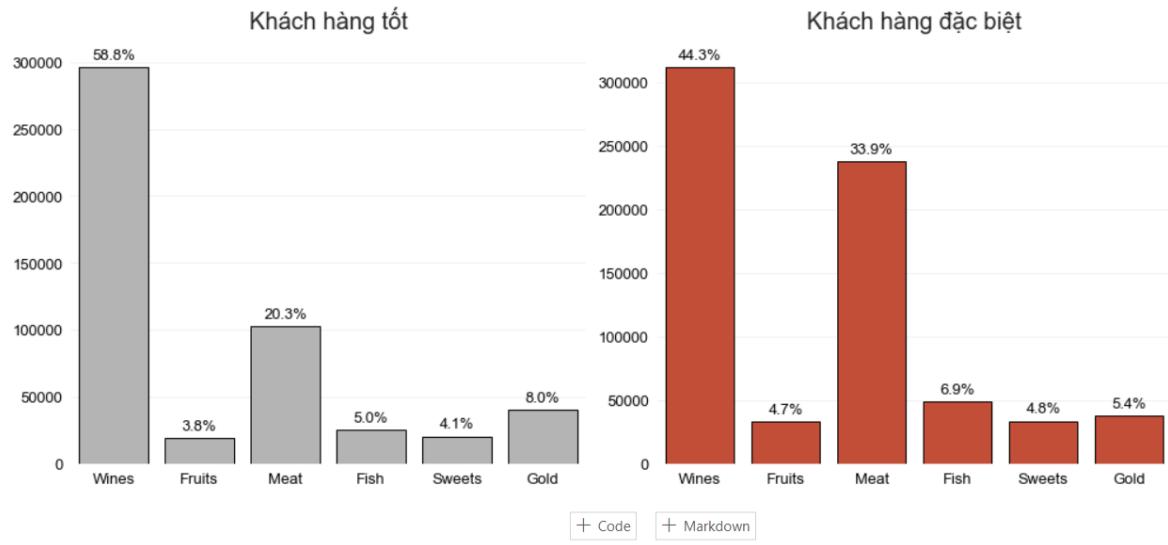
Bảng thể hiện số liệu mức chi tiêu của từng cụm khách hàng trên từng mặt hàng

Clusters_Customers	Category	Khách hàng bình thường	Khách hàng tiềm năng	Khách hàng tốt	Khách hàng đặc biệt
0	Wines	21566	50664	296082	311717
1	Fruits	4078	2487	19127	33061
2	Meat	15992	16961	102359	238063
3	Fish	6229	3487	25407	48808
4	Sweets	4131	2376	20469	33576
5	Gold	11186	8866	40503	37791

Biểu đồ thể hiện phần trăm mức chi tiêu của từng mặt hàng:

Spending of different customer groups





Tất cả bốn nhóm đều chi số tiền cao nhất cho Rượu, theo sau là Thịt, trong sáu loại sản phẩm khác nhau.

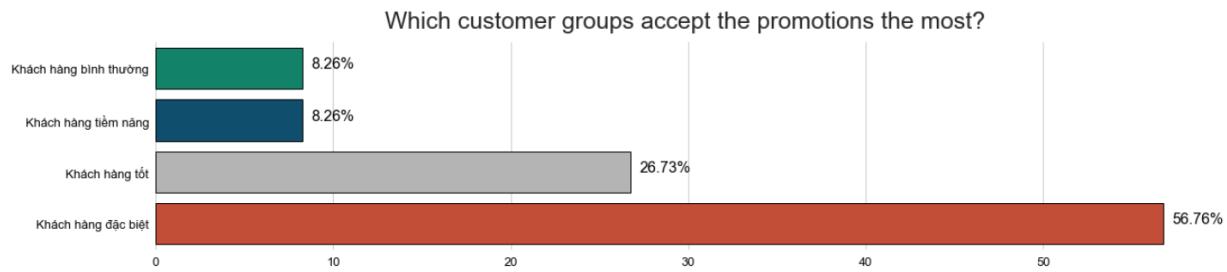
Về việc mua Vàng, Khách hàng Bình thường và Khách hàng Có tiềm năng có xu hướng làm vậy hơn, với tỷ lệ lần lượt là 17,7% và 10,5%, trong khi hai nhóm còn lại chỉ có tỷ lệ 8% và 5,4%.

Khách hàng Đặc biệt, ngược lại, có xu hướng mua Thịt nhiều hơn, chiếm khoảng 240.000 bảng hoặc 33,9% tổng chi tiêu của họ, cao hơn so với chi tiêu của các nhóm khác dưới 26%.

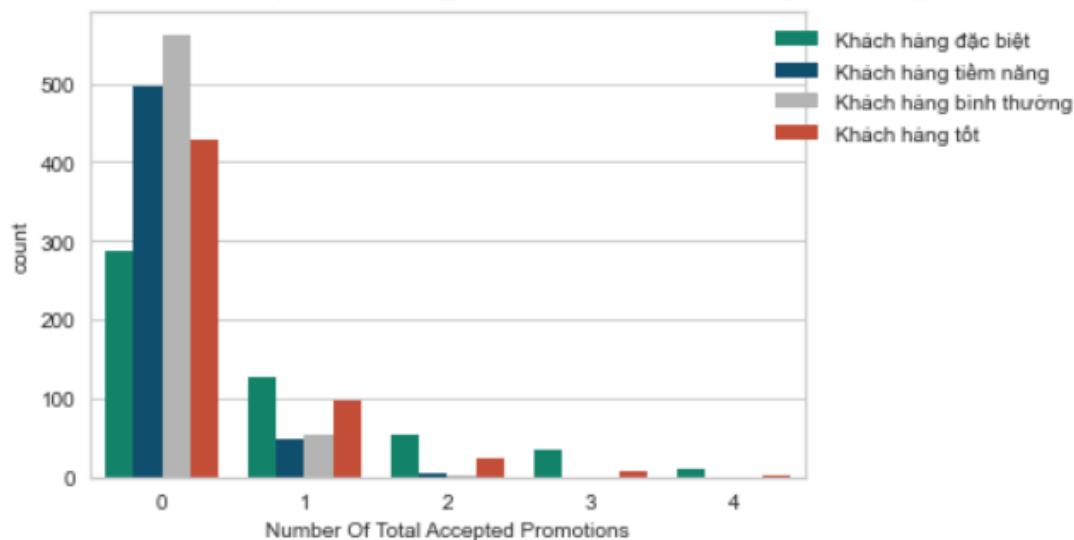
➤ Đánh giá phân cụm khách hàng dựa trên chiến dịch quảng cáo:

Bảng số liệu số lần chấp nhận và phần trăm trên từng cụm khách hàng.

	<b>Clusters_Customers</b>	<b>Total_Promos</b>	<b>Percentage_Promos</b>
0	Elite customer	378	56.76
1	Good customer	178	26.73
2	Ordinary customer	55	8.26
3	Potential good customer	55	8.26



Promotions accepted by customer groups in each marketing campaign

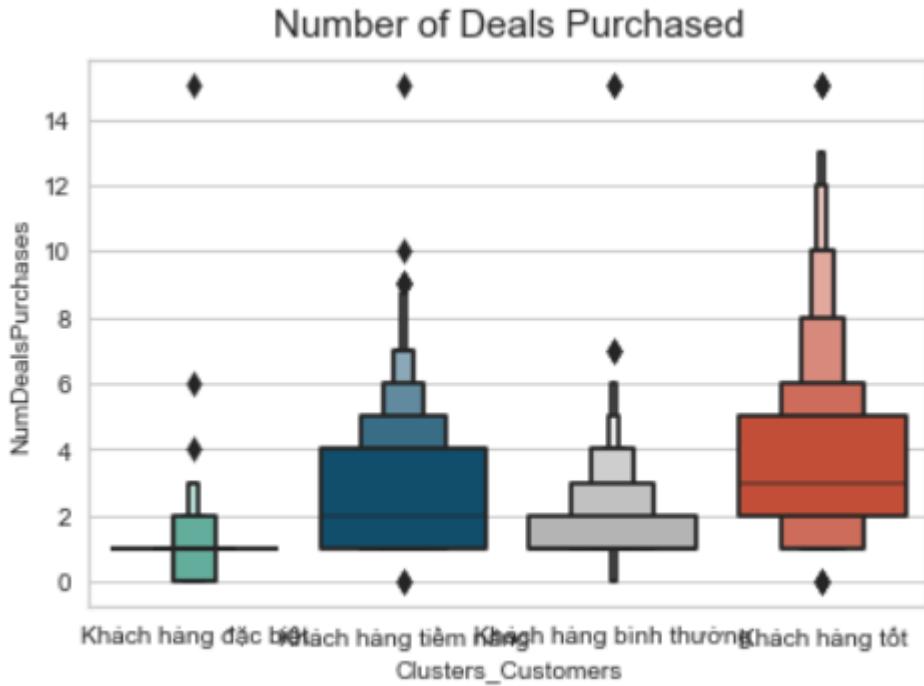


Nhận xét:

Trong chiến dịch tiếp thị ban đầu, nhóm Khách hàng Bình thường và Khách hàng Có tiềm năng có khoảng 500 lượt chấp nhận mỗi nhóm, trong khi nhóm Khách hàng Tốt và Khách hàng Đặc biệt có 400 và 300 lượt chấp nhận tương ứng.

Trong các chiến dịch tiếp theo, số lượng khuyến mãi được chấp nhận giảm đáng kể. Khách hàng Bình thường và Khách hàng Có tiềm năng có xu hướng không chấp nhận ưu đãi, trong khi chỉ có một phần nhỏ của Khách hàng Tốt và Khách hàng Đặc biệt chấp nhận nó.

- Đánh giá kết quả phân cụm dựa trên lượt mua hàng qua giảm giá:



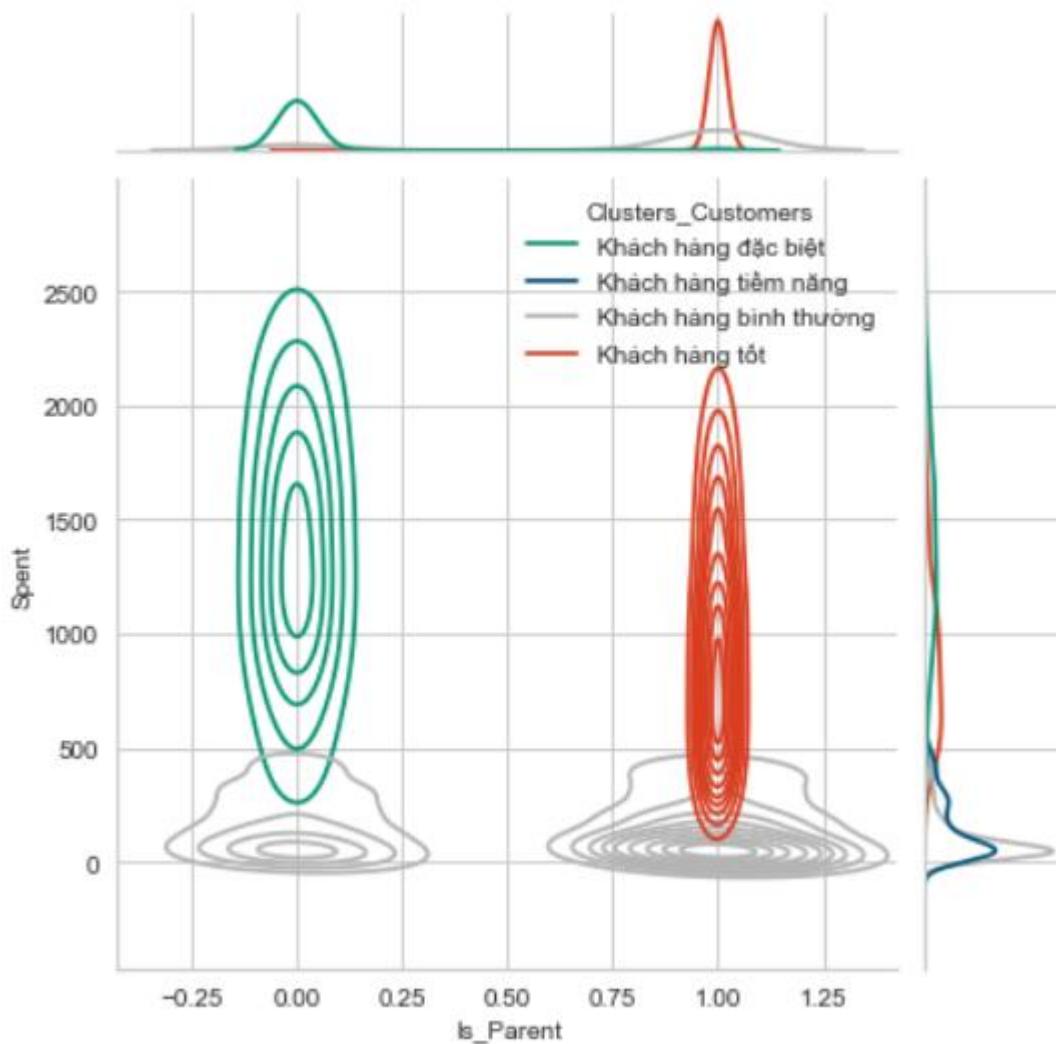
Các giao dịch dường như thành công trong các nhóm khách hàng tốt và khách hàng tốt tiềm năng. Tuy nhiên, nhóm khách hàng ưu tú không thể hiện sự quan tâm đến các giao dịch được cung cấp.

Đánh giá tổng quan:

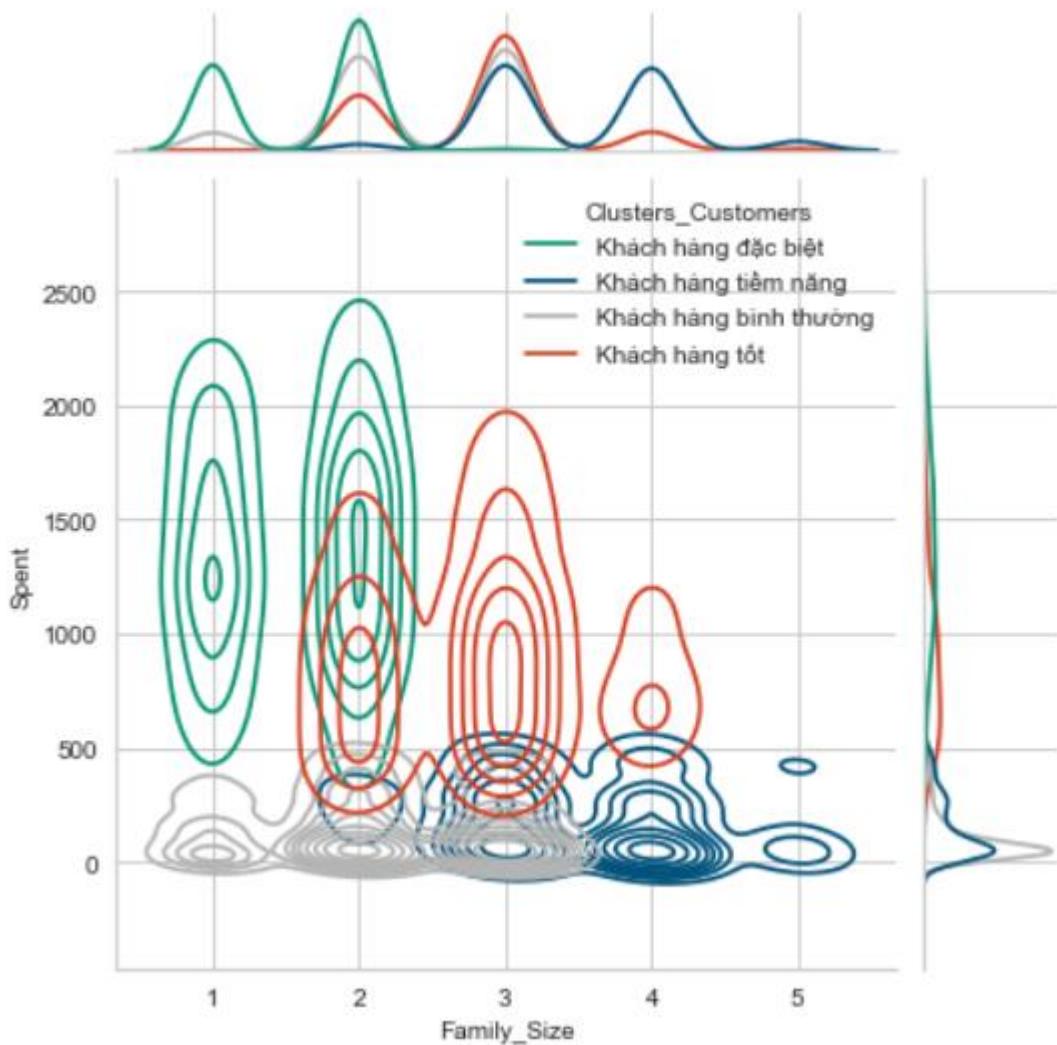
Bây giờ ta đã hình thành các cụm và xem xét thói quen mua hàng của họ. Vì vậy, ta sẽ lập hồ sơ các cụm được hình thành và đi đến kết luận về ai là khách hàng ngôi sao của mình và ai cần sự chú ý nhiều hơn từ nhóm tiếp thị của cửa hàng bán lẻ.

Phần sau đây, ta sẽ vẽ các nhóm khách hàng trên thuộc tính spent và các thuộc tính cá nhân của họ:

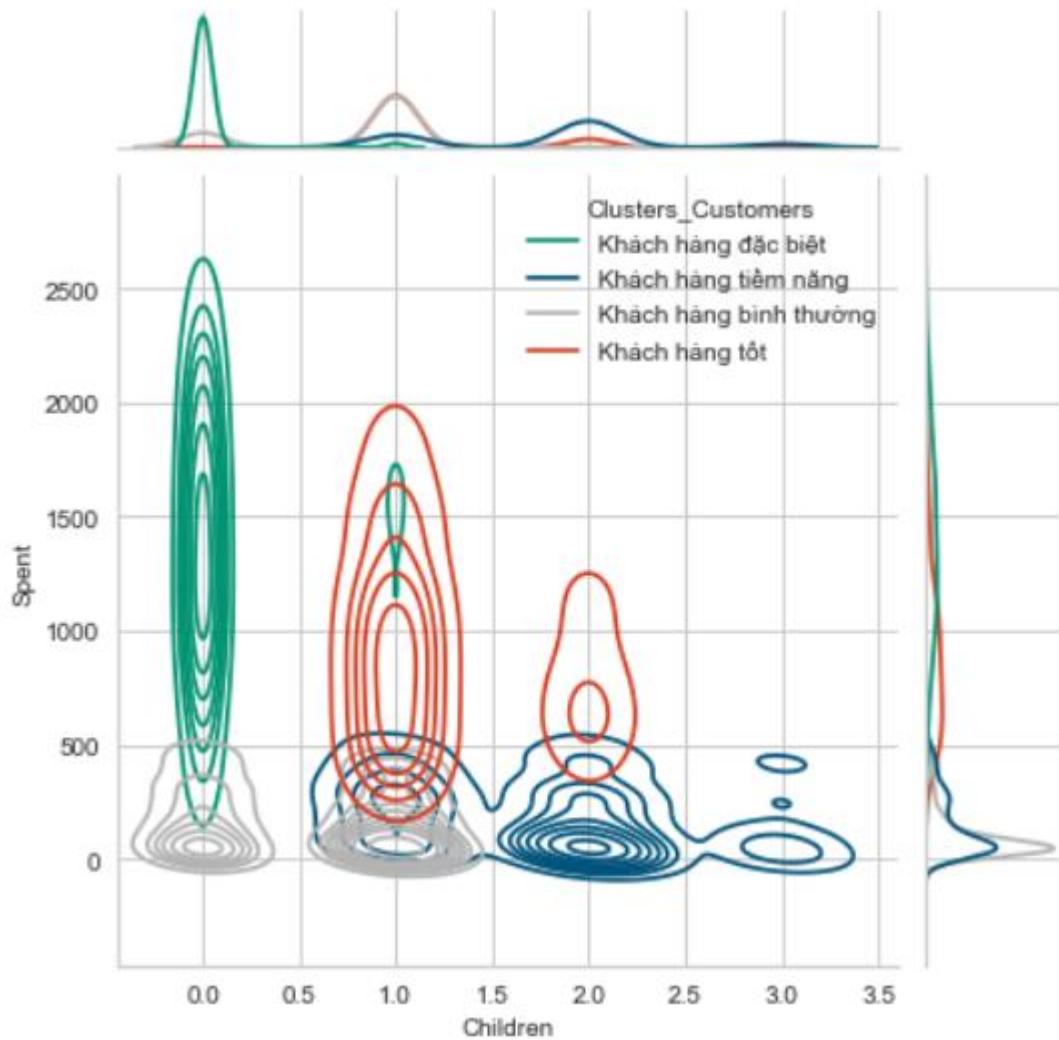
Hình ảnh trực quan của Spent và Is\_Parent(là cha mẹ hay không).



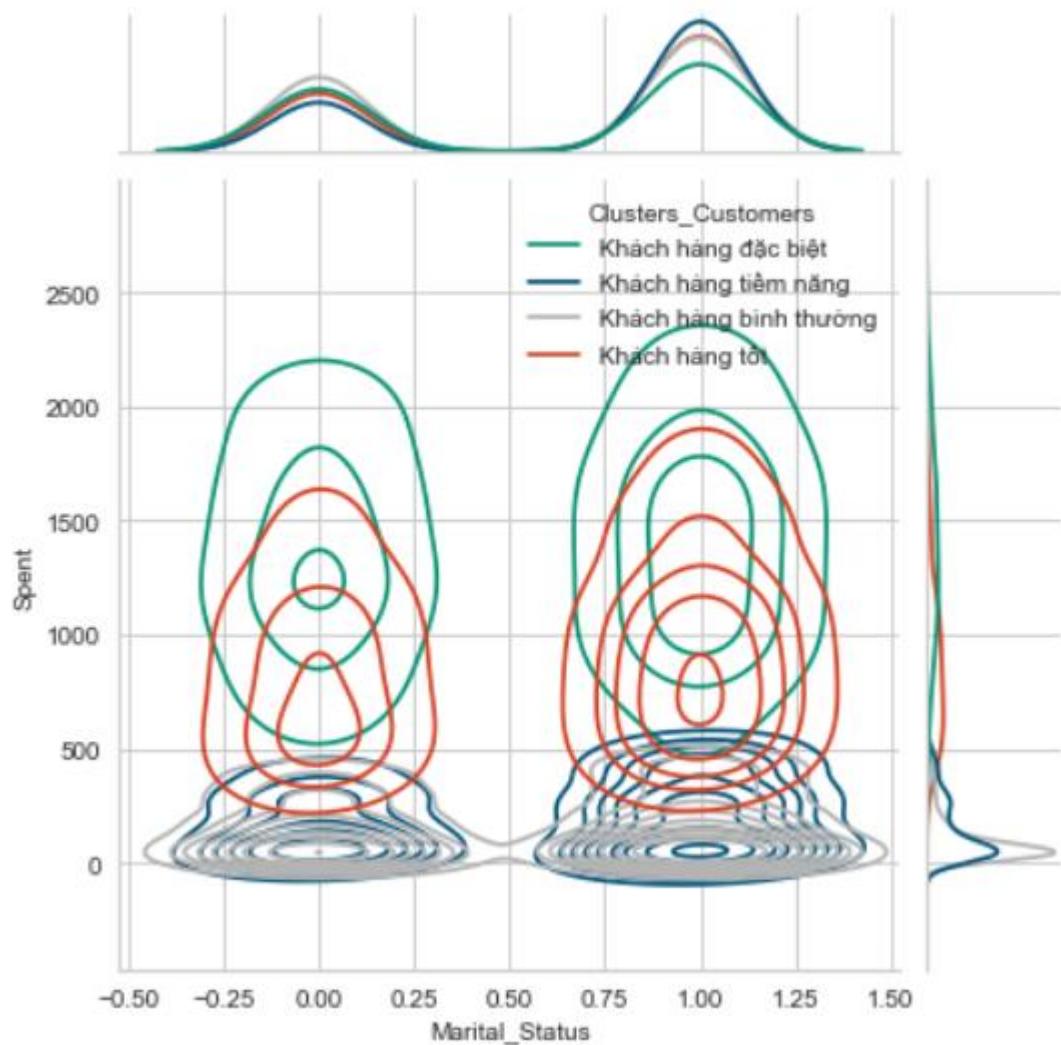
Hình ảnh trực quan của Spent và Family\_size(kích cỡ gia đình).



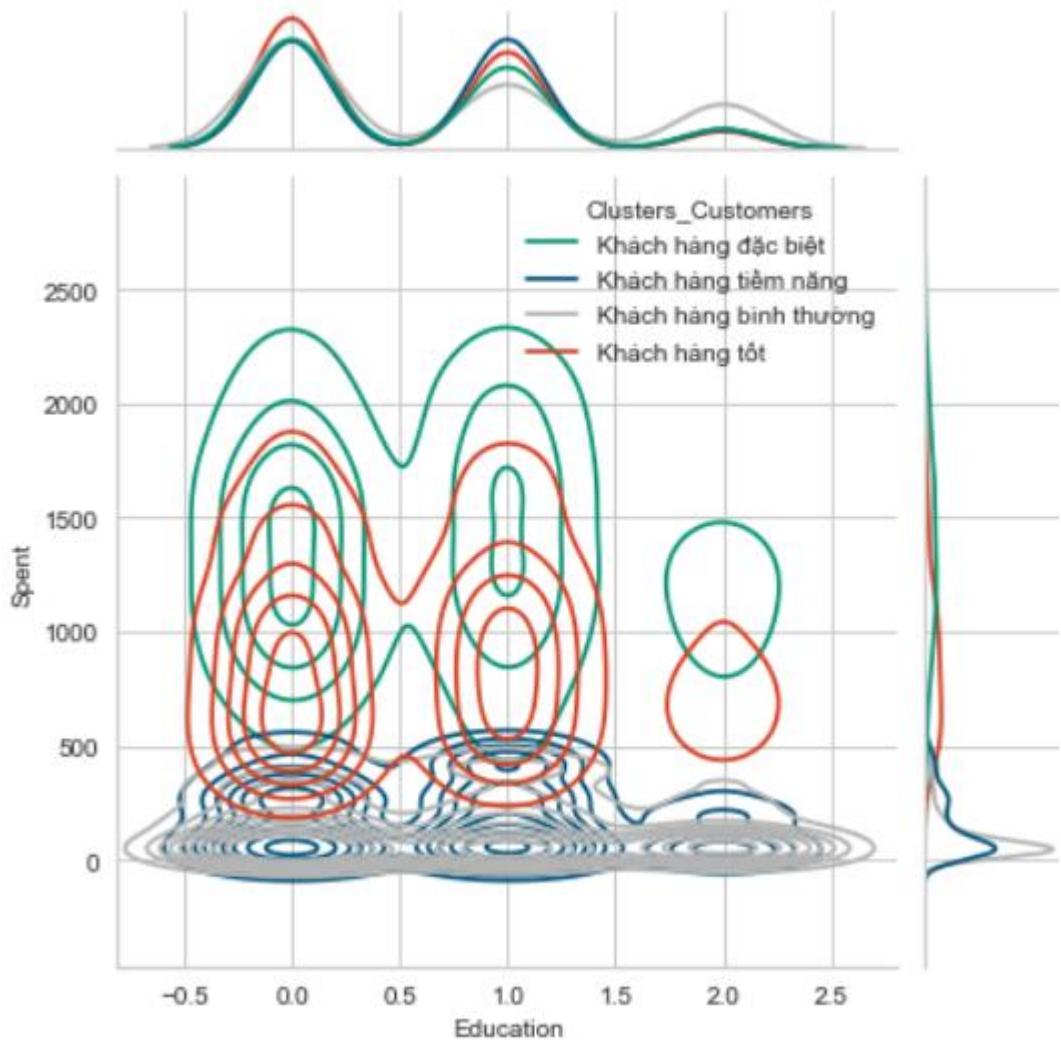
Hình ảnh trực quan của Spent và Children(có bao nhiêu con).



Hình ảnh trực quan của Spent và Marital\_Status(Tình trạng hôn nhân).



Hình ảnh trực quan của Spent và Education(trình độ giao dục).



## Nhận xét:

### 1. Nhóm Đặc biệt

- Chi tiêu cao và thu nhập cao
- Họ chắc chắn không phải là cha mẹ
- Số lượng thành viên trong gia đình tối đa là hai người, không có con
- Họ có hơi nhiều cặp đôi hơn người độc thân

### 2. Nhóm Tốt

- Chi tiêu cao và thu nhập trung bình
- Rất có thể là cha mẹ
- Kích thước gia đình của họ có thể từ hai đến bốn thành viên
- Chủ yếu có một con (hầu hết là thanh thiếu niên)

### 3. Nhóm Có tiềm năng

- Chi tiêu thấp và thu nhập trung bình
- Tất cả đều là cha mẹ
- Kích thước gia đình của họ dao động từ hai đến năm thành viên
- Độ tuổi của họ từ 35 đến khoảng 80 tuổi

### 4. Nhóm Bình thường

- Chi tiêu thấp và thu nhập thấp
- Phần lớn họ là cha mẹ
- Kích thước gia đình tối đa là ba người
- Họ thường có một con (hầu hết là trẻ em)

Kết luận:

Dự án này về phân khúc khách hàng từ một tập dữ liệu không giám sát. Phân tích thành phần chính được sử dụng để giảm xuống chỉ còn 3 chiều. Sáu phương pháp phân cụm được vẽ và tính điểm silhouette để so sánh và tìm ra phương pháp phù hợp nhất là KMeans. 4 nhóm khách hàng được phân cụm và phân tích dựa trên thu nhập và chi tiêu cũng như thông tin cá nhân của họ. Bằng cách đó, các mẫu được xác định có thể hữu ích trong việc phát triển các chiến lược tiếp thị hiệu quả hơn.

## IV. KẾT LUẬN

### 1. Kết quả đạt được

Trong quá trình học và quá trình làm đồ án, nhóm đã tìm hiểu và vận dụng kiến thức về cách sử dụng công cụ SSAS, Python và các thuật toán để tiến hành khai phá. Những kết quả mà nhóm nhận thấy được:

- Nắm rõ các kiến thức về các thuật toán mà nhóm sử dụng để thực hiện đề tài.
- Trang bị kiến thức về công cụ SSAS, cách triển khai thuật toán trong Python, và cách trực quan hóa dữ liệu để có cái nhìn tổng quan.
- Khai phá được nhiều những tri thức mới trong tập dữ liệu.

### 2. Hạn chế

Do thời gian hạn ngắn cộng với khối lượng công việc nhiều nên trong quá trình thực hiện đồ án nhóm còn gặp phải một số vấn đề :

- Khả năng hiểu hết các thuật toán đã học.
- Quá trình SSAS gặp một số vấn đề khó khăn.

### 3. Bảng phân công nhiệm vụ

Công việc	Lê Hoàng Khang	Nguyễn Thanh Hùng	Nguyễn Duy Thái	Hoàng Uyên
Xác định đề tài và chọn tập dữ liệu	x	x	x	x
Chọn thuật toán cho đề tài	x	x	x	x
Tiền xử lý dữ liệu: missing data, đổi tên cột, thêm cột mới.				x

Xác định các biến cần phân cụm cho thuật toán K-means				x
Sử dụng elbow xác định số cụm cho thuật toán K-means				x
Thực hiện thuật toán Microsoft Clustering trên SSAS	x	x	x	x
Đọc kết quả và nhận xét thuật toán Microsoft Clustering	x	x	x	
Vẽ biểu đồ tương quan và xác định số biến quan trọng sử dụng cho Decision Tree và Association Rule	x		x	
Thực hiện thuật toán Decision Tree sử dụng SSAS	x	x	x	x
Đọc kết quả và nhận xét cho thuật toán Decision Tree	x			

Thực hiện thuật toán Association Rule sử dụng SSAS	X	X	X	X
Đọc kết quả và nhận xét cho thuật toán Association Rule			X	
Đánh giá thực nghiệm trên toàn bộ tập dữ liệu sử dụng Python	X	X	X	
Nhận xét đưa ra kết luận cho phần thực nghiệm		X		
Viết báo cáo	X		X	
Làm slide thuyết trình		X		X

#### **4. Tài liệu tham khảo**

- Slide bài giảng của thầy Nguyễn Văn Thành
- Youtube:
  - Applied Association Rule Mining in Banking with SSAS

Link: <https://www.youtube.com/watch?v=FaPRd6xIX4A>

- Building Clustering model with SSAS

Link: <https://www.youtube.com/watch?v=gBxCexfKDFo>

- Building Classification model with Decision Tree Using SSAS

Link: <https://www.youtube.com/watch?v=o8tEEBy5zjQ>