**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

****



**đỒ ÁN KẾT THÚC HỌC PHẦN**

**Môn: PHÂN TÍCH DỮ LIỆU**

Mã môn học: 841432

Khoa: Công nghệ thông tin

Chuyên ngành: Hệ thống thông tin

Nhóm nghiên cứu:

Trịnh Hùng Thái – 3120410471

Phạm Minh Quân – 3120410438

**Giảng viên phụ trách:**

**PHAN THÀNH HUẤN**

**Thành phố Hồ Chí Minh, tháng 12 năm 2023**

# **Lời cảm ơn**

Trước hết em xin gửi đến lời cảm ơn chân thành và sâu sắc nhất đến thầy TS. Phan Thành Huấn người trực tiếp hướng dẫn và tận tình chỉ bảo cho nhóm chúng em cho tới khi em hoàn thành đồ án của mình.

Tiếp đến em xin giành lời cảm ơn đến quý thầy cô Trường Đại học Sài Gòn – khoa Công nghệ thông tin đã truyền đạt cho em những kiến thức vô cùng quý báu và bổ ích trong suốt quá trình nghiên cứu và học tập tại trường.

Xin chân thành cảm ơn tới những người bạn đã luôn sát cánh cùng em, những lời động viên, những lần hỗ trợ những lúc cần thiết đã phần nào giúp em hoàn thành đồ án này.

Cuối cùng, em xin cảm ơn đến ba mẹ và người thân trong gia đình đã hỗ trợ và tạo điều kiện thuận lợi cho em trong suốt thời gian học tập và nghiên cứu tại Đại học Sài Gòn.

# **BẢNG ĐÁNH GIÁ MỨC ĐỘ THAM GIA**

|  |  |  |
| --- | --- | --- |
| Danh sách thành viên | MSSV | Mức độ tham gia |
| Trịnh Hùng Thái | 3120410471 | 100% |
| Phạm Minh Quân | 3120410438 | 100% |

Đồ án này trình bày quá trình tìm kiếm, khám phá và phân tích thông tin từ tập dữ liệu tìm được. Trong đó có sử dụng các các thuật toán và phương pháp khai thác dữ liệu. Cuối cùng, quá trình khai thác dữ liệu dẫn đến việc đưa ra các lời khuyên cũng như hướng giải quyết phù hợp với nhu cầu của khách hàng.

# **Mục lục**

[Lời cảm ơn i](#_Toc152765422)

[BẢNG ĐÁNH GIÁ MỨC ĐỘ THAM GIA ii](#_Toc152765423)

[Mục lục iii](#_Toc152765424)

[Danh mục bảng biểu v](#_Toc152765425)

[Danh mục hình ảnh v](#_Toc152765426)

[Chương 1: KHÁI QUÁT ĐỒ ÁN 1](#_Toc152765427)

[1.1. Lí do chọn đề tài 1](#_Toc152765428)

[1.2. Mô tả dữ liệu và cấu trúc dữ liệu 2](#_Toc152765429)

[Chương 2: KHẢO SÁT VÀ TIỀN XỬ LÝ BỘ DỮ LIỆU 4](#_Toc152765430)

[2.1. Khảo sát bộ dữ liệu 4](#_Toc152765431)

[2.2. Tiền xử lí dữ liệu 5](#_Toc152765432)

[2.2.1. Kiểm tra bộ dữ liệu 5](#_Toc152765433)

[2.2.2. Biến đổi và thêm mới dữ liệu 8](#_Toc152765434)

[2.2.3. Bộ dữ liệu tổng thể 10](#_Toc152765435)

[2.3. Thăm dò dữ liệu sau tiền xử lý 11](#_Toc152765436)

[2.3.1. Khái niệm về EDA 12](#_Toc152765437)

[2.3.2. Mục đích sử dụng EDA 12](#_Toc152765438)

[2.3.3. Thực nghiệm EDA đối với dữ liệu phân tích 13](#_Toc152765439)

[2.3.3.1. Tỷ suất lợi nhuận của mặt hàng sản phẩm 13](#_Toc152765440)

[2.3.3.2. Tổng tiền thu được trên mỗi danh mục mặt hàng 14](#_Toc152765441)

[2.3.3.3. Số lượng đơn hàng được đặt qua từng tháng từ năm 2017 – 2021 17](#_Toc152765442)

[2.4. Kết luận sau quá trình EDA 18](#_Toc152765443)

[Chương 3: BÀI TOÁN VÀ HƯỚNG GIẢI QUYẾT 19](#_Toc152765444)

[3.1. Đặt ra bài toán và hướng xử lí 19](#_Toc152765445)

[3.2. Các thuật toán hồi quy 19](#_Toc152765446)

[3.2.1. Giới thiệu 19](#_Toc152765447)

[3.2.2. Linear Regression 20](#_Toc152765448)

[3.2.3. Support Vector Regression 20](#_Toc152765449)

[3.3. Bài toán 1: Làm thế nào để lấy được giá bán lẻ tốt nhất 22](#_Toc152765450)

[3.3.1. Huấn luyện mô hình 22](#_Toc152765451)

[3.3.1.1. Linear Regression 2 biến 22](#_Toc152765452)

[3.3.1.2. Linear Regression 1 biến 25](#_Toc152765453)

[3.3.1.3. Support Vector Regression 27](#_Toc152765454)

[3.3.2. Thực hiện dự đoán giá bán lẻ tốt nhất của một mặt hàng có giá bán sỉ là x và tổng số lượng bán dự kiếm là y trong vòng 5 năm 29](#_Toc152765455)

[3.3.2.1. Mô tả quá trình dự đoán 29](#_Toc152765456)

[3.3.2.2. Kết quả dự đoán của giá bán sỉ 200 USD và tổng lượng bán 500 31](#_Toc152765457)

[3.3.2.3. Kết quả dự đoán của giá bán sỉ 1000 và tổng lượng bán 500 31](#_Toc152765458)

[3.3.3. So sánh giữa các mô hình huấn luyện 32](#_Toc152765459)

[3.4. Bài toán 2: Có lợi thế về ngày giao hàng khi trở thành khách hàng có thứ hạng Bạch Kim hay không 33](#_Toc152765460)

[KẾT LUẬN 35](#_Toc152765461)

# **Danh mục bảng biểu**

[Bảng 1.1. Bảng mô tả cấu trúc của bộ dữ liệu orders.csv 2](#_Toc152765260)

[Bảng 1.2. Bảng mô tả cấu trúc của bộ dữ liệu product-supplier.csv 3](#_Toc152765261)

# **Danh mục hình ảnh**

[Hình 2.1. Kết quả trả về từ câu lệnh info() đối với bộ dữ liệu orders.csv 6](#_Toc152708275)

[Hình 2.2. Kết quả trả về từ câu lệnh isnal().sum() đối với bộ dữ liệu orders.csv 6](#_Toc152708276)

[Hình 2.3. Kết quả trả về từ câu lệnh isnal().sum() đối với bộ dữ liệu orders.csv 6](#_Toc152708277)

[Hình 2.4. Kết quả trả về từ câu lệnh info() đối với bộ dữ liệu product-supplier.csv 7](#_Toc152708278)

[Hình 2.5. Kết quả trả về từ câu lệnh isna().sum() đối với bộ dữ liệu product-supplier.csv 7](#_Toc152708279)

[Hình 2.6. Kết quả trả về từ câu lệnh isnull().sum() đối với bộ dữ liệu product-supplier.csv 7](#_Toc152708280)

[Hình 2.7. Kiểu dữ liệu ngày và tháng ban đầu 8](#_Toc152708281)

[Hình 2.8. Kiểu dữ liệu ngày tháng sau khi được biến đổi để phù hợp cho việc phân tích 8](#_Toc152708282)

[Hình 2.9. Lỗi không thống nhất dữ liệu trên cột Customer Status 9](#_Toc152708283)

[Hình 2.10. Thông nhất kiểu dữ liệu trên cột Customer Status 9](#_Toc152708284)

[Hình 2.11. Thêm một cột dữ liệu mới tên là Item Retail Value 9](#_Toc152708285)

[Hình 2.12. Tạo dataframe mới nhằm thống kê số lượng sản phẩm được bán cũng như giá trị vốn - lãi trung bình của sản phẩm 10](#_Toc152708286)

[Hình 2.13. Thông tin của bảng dữ liệu được merge lại với nhau 10](#_Toc152708287)

[Hình 2.14. Dữ liệu của bảng dữ liệu được merge lại với nhau 11](#_Toc152708288)

[Hình 2.15. Bộ dữ liệu cuối cùng trong quá trình tiền xử lý 11](#_Toc152708289)

[Hình 2.16. Phân bố sản phẩm dựa trên giá bán lẻ và giá bán sỉ 13](#_Toc152708290)

[Hình 2.17. Mặt hàng có tỷ suất lợi nhuận cao nhất 14](#_Toc152708291)

[Hình 2.18. Biểu đồ trực quan hóa tổng tiền thu được trên các danh mục sản phẩm 14](#_Toc152708292)

[Hình 2.19. Biểu đồ trực quan hóa số tiền thu được đựa trên khoảng mức giá bán sỉ 15](#_Toc152708293)

[Hình 2.20. Mô tả thống kê cơ bản trong cột Wholesale Price 16](#_Toc152708294)

[Hình 2.21. Biểu đồ trực quan hóa số tiền thu được đựa trên khoảng mức giá bán lẻ 16](#_Toc152708295)

[Hình 2.22. Mô tả thống kê cơ bản trong cột Retail Price 17](#_Toc152708296)

[Hình 2.23. Mối tương quan giữa số lượng đặt hàng 17](#_Toc152708297)

[Hình 3.1. Minh hoạ phương pháp SVM 21](#_Toc152765049)

[Hình 3.2. Chọn biến độc lập và biến phụ thuộc LNR-2 22](#_Toc152765050)

[Hình 3.3. Sử dụng hàm train\_test\_split để sinh các x\_train, y\_train, x\_test và y\_test trong LNR-2 22](#_Toc152765051)

[Hình 3.4. Bắt đầu quá trình huấn luyện mô hình LNR-2 23](#_Toc152765052)

[Hình 3.5. Dự đoán kết quả LNR-2 23](#_Toc152765053)

[Hình 3.6. Trực quan hóa kết quả dự đoán LNR-2 23](#_Toc152765054)

[Hình 3.7. Đánh giá chỉ tiêu và điểm số hồi quy LNR-2 24](#_Toc152765055)

[Hình 3.8. Trực quan hóa phân bổ dư lượng LNR-2 24](#_Toc152765056)

[Hình 3.9. Kết quả hệ số của LNR-2 25](#_Toc152765057)

[Hình 3.10. Chọn biến độc lập và biến phụ thuộc LNR-1 25](#_Toc152765058)

[Hình 3.11. Sử dụng hàm train\_test\_split để sinh các x\_train, y\_train, x\_test và y\_test trong LNR-1 26](#_Toc152765059)

[Hình 3.12. Trực quan hóa kết quả dự đoán LNR-1 26](#_Toc152765060)

[Hình 3.13. Trực quan hóa phân bổ dư lượng LNR-1 26](#_Toc152765061)

[Hình 3.14. Đánh giá chỉ tiêu và điểm số hồi quy LNR-1 27](#_Toc152765062)

[Hình 3.15. Kết quả hệ số LNR-1 27](#_Toc152765063)

[Hình 3.16. Biến độc lập, biến phụ thuộc và tham số train\_test\_split của SVR 27](#_Toc152765064)

[Hình 3.17. Xác định các tham số của từng kernel trong SVR 28](#_Toc152765065)

[Hình 3.18. Trực quan hóa kết quả dự đoán của SVR 28](#_Toc152765066)

[Hình 3.19. Điểm số hồi quy của SVR 29](#_Toc152765067)

[Hình 3.20. Dự đoán giá bán lẻ tốt nhất dựa trên giá bán sỉ và số lượng bán 30](#_Toc152765068)

[Hình 3.21. Nạp dữ liệu dự đoán vào các mô hình đã được huấn luyện 30](#_Toc152765069)

[Hình 3.22. Kết quả giá bán lẻ tốt nhất dự đoán được trả về với trường hợp giá bán sỉ là 200 USD và tổng lượng bán là 500 31](#_Toc152765070)

[Hình 3.23. Kết quả giá bán lẻ tốt nhất dự đoán được trả về với trường hợp giá bán sỉ là 1000 USD và tổng lượng bán là 500 31](#_Toc152765071)

[Hình 3.24. Điểm số hồi quy giữa các mô hình huấn luyện 32](#_Toc152765072)

[Hình 3.25. Dataframe mới phù hợp cho việc giải quyết bài toán số 2 33](#_Toc152765073)

[Hình 3.26. Cột Delay for Delivery trước khi chuyển thành số 33](#_Toc152765074)

[Hình 3.27. Cột Delay for Delivery sau khi chuyển thành số 33](#_Toc152765075)

[Hình 3.28. Thời gian trung bình chờ đợi của khách hàng theo từng loại khách hàng 34](#_Toc152765076)

[Hình 3.29. Mô hình hoá dữ liệu của dataframe mới 34](#_Toc152765077)

# **Chương 1: KHÁI QUÁT ĐỒ ÁN**

## **Lí do chọn đề tài**

Phân tích nhu cầu của khách hàng là một hoạt động quan trọng đối với mọi doanh nghiệp. Việc hiểu rõ nhu cầu của khách hàng giúp doanh nghiệp đưa ra các sản phẩm và dịch vụ đáp ứng nhu cầu đó, từ đó thỏa mãn và giữ chân khách hàng.  Tầm quan trọng của phân tích nhu cầu của khách hàng mang lại nhiều lợi ích cho doanh nghiệp, bao gồm:

* Thỏa mãn nhu cầu của khách hàng: Khi doanh nghiệp hiểu rõ nhu cầu của khách hàng, doanh nghiệp có thể đưa ra các sản phẩm và dịch vụ đáp ứng nhu cầu đó, từ đó thỏa mãn khách hàng.
* Giữ chân khách hàng: Khách hàng hài lòng với sản phẩm và dịch vụ của doanh nghiệp sẽ có xu hướng tiếp tục sử dụng sản phẩm và dịch vụ của doanh nghiệp, từ đó giúp doanh nghiệp giữ chân khách hàng.
* Phát triển doanh nghiệp: Việc hiểu rõ nhu cầu của khách hàng giúp doanh nghiệp phát triển các sản phẩm và dịch vụ mới, đáp ứng nhu cầu của thị trường, từ đó phát triển doanh nghiệp.

Xu hướng thay đổi của nhu cầu khách hàng  luôn thay đổi theo thời gian và theo sự phát triển của xã hội. Doanh nghiệp cần thường xuyên cập nhật thực trạng nhu cầu của khách hàng để có thể đáp ứng nhu cầu của khách hàng một cách tốt nhất. Một số xu hướng chính trong nhu cầu của khách hàng hiện nay bao gồm:

* Nhu cầu về sản phẩm và dịch vụ chất lượng cao, đáp ứng nhu cầu cá nhân hóa của khách hàng.
* Nhu cầu về trải nghiệm mua sắm và sử dụng sản phẩm dịch vụ tốt.
* Nhu cầu về giá cả hợp lý.
* Nhu cầu về tính bền vững của sản phẩm và dịch vụ.

Tóm lại, phân tích nhu cầu của khách hàng là một hoạt động quan trọng đối với mọi doanh nghiệp. Việc hiểu rõ nhu cầu của khách hàng giúp doanh nghiệp phát triển bền vững.

## **Mô tả dữ liệu và cấu trúc dữ liệu**

Bộ dữ liệu được tổng hợp từ nguồn Customer Personality Analysis  :

**https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data**

Đây là nguồn dữ liệu về khách hàng được đánh giá là tin cậy nhất trong các bộ dữ liệu của kaggle.

Dữ liệu gồm 3 yếu tố: Con ngườ (People), sản phẩm (Product), khuyến mãi (Promotion) và địa điểm mua hàng (Place).

Về con người:

Bảng 1. 1. Bảng mô tả dữ liệu các thuộc tính của con người

|  |  |  |
| --- | --- | --- |
| Thuộc tính |  | Ý nghĩa |
| ID |  | Mã khách hàng |
| Year\_Birth |  | Năm sinh của khách hàng |
| Education |  | Trình độ học vấn |
| Marital\_Status |  | Tình trạng hôn nhân |
| Income |  | Thu nhập |
| Kidhome |  | Số trẻ em trong nhà |
| Teenhome |  | Số trẻ vị thành niên trong nhà |
| Dt\_Customer |  | Ngày khách hàng đăng kí thẻ |
| Recency |  | Số ngày kể từ lần cuối khách hàng mua hàng |
| Complain |  | 1 nếu khách hàng phàn nàn trong vòng 2 năm đổ lại, ngược lại là 0 |

Bảng 1. 2. Bảng mô tả dữ liệu các thuộc tính của sản phẩm

|  |  |
| --- | --- |
| Thuộc tính | Ý nghĩa |
| MntWines | Số lượng tiêu thụ rượu trong vòng 2 năm gần đây |
| MntFruits | Số lượng tiêu thụ hoa quả trong vòng 2 năm gần đây |
| MntMeatProducts | Số lượng tiêu thụ các sản phẩm từ thịt trong vòng 2 năm gần đây |
| MntFishProduct | Số lượng tiêu thụ các sản phẩm từ cá trong vòng 2 năm gần đây |
| MntSweetProduct | Số lượng tiêu thụ các sản phẩm ăn vặt trong vòng 2 năm gần đây |
| MntGoldProds | Số lượng tiêu thụ rượu trong vòng 2 năm gần đây |

Bảng 1. 3. Bảng mô tả dữ liệu các thuộc tính của các chiến dịch khuyến mãi

|  |  |
| --- | --- |
| Thuộc tính | Ý nghĩa |
| NumDealsPurchases | Số lần khách hàng mua hàng được giảm giá |
| AcceptedCmp1 | 1 nếu khách hàng sử dụng ưu đãi trong chiến dịch đầu tiên, ngược lại là 0 |
| AcceptedCmp2 | 1 nếu khách hàng sử dụng ưu đãi trong chiến dịch đầu tiên, ngược lại là 0 |
| AcceptedCmp3 | 1 nếu khách hàng sử dụng ưu đãi trong chiến dịch đầu tiên, ngược lại là 0 |
| AcceptedCmp4 | 1 nếu khách hàng sử dụng ưu đãi trong chiến dịch đầu tiên, ngược lại là 0 |
| AcceptedCmp5 | 1 nếu khách hàng sử dụng ưu đãi trong chiến dịch đầu tiên, ngược lại là 0 |
| Response | 1 nếu khách hàng sử dụng ưu đãi trong chiến dịch trước, ngược lại là 0 |

Bảng 1. 4. Bảng mô tả dữ liệu các thuộc tính của địa điểm mua bán

|  |  |
| --- | --- |
| Thuộc tính | Ý nghĩa |
| NumWebPurchases | Số lần khách hàng  mua hàng qua website của cửa hàng |
| NumCatalogPurchases | Số lần khách hàng mua hàng được thực hiện bằng danh mục |
| NumStorePurchases | Số lần khách hàng mua hàng tại cửa hàng |
| NumWebVisitMonth | Số lần khách hàng ghé thăm website của cửa hàng vào tháng qua |

# **Chương 2: KHẢO SÁT VÀ TIỀN XỬ LÝ BỘ DỮ LIỆU**

## **Khảo sát bộ dữ liệu**

Trong đồ án môn học này, chúng em sử dụng bộ dữ liệu thông tin cá nhân của khách hàng mua hàng từ năm 2012 đến năm 2014, bộ dữ liệu phân tích chỉ có 1 file csv tương ứng:

Bộ dữ liệu marketing\_campaign.csv (gồm 2240 bản ghi và 29 trường thuộc tính):

Đây là bộ dữ liệu chứa thông tin cá nhân của khách hàng. Trong đó, mỗi bản ghi (record) trong bộ dữ liệu tượng trưng cho thông tin của khách hàng như năm sinh, trình độ học vấn, tình trạng hôn nhân, … .

\*Về con người:

ID: Mã khách hành

Year\_Birth: Năm sinh của khách hàng

Education: Trình độ học vấn

Marital\_Status: Tình trạng hôn nhân

Income: Thu nhập

Kidhome: Số trẻ em trong nhà

Teenhome: Số trẻ vị thành niên trong nhà

Dt\_Customer: Ngày khách hàng đăng kí thẻ

Recency: Số ngày kể từ lần cuối khách hàng mua hàng

Complain: 1 nếu khách hàng phàn nàn trong vòng 2 năm đổ lại, ngược lại là 0

\*Về sản phẩm:

MntWines: Số tiền khách hàng đã chi ra để mua rượu trong vòng 2 năm

MntFruits: Số tiền khách hàng đã chi ra để mua trái cây trong vòng 2 năm

MntMeatProducts: Số tiền khách hàng đã chi ra để mua các sản phẩm từ thịt trong vòng 2 năm

MntFishProducts: Số tiền khách hàng đã chi ra để mua các sản phẩm từ cá trong vòng 2 năm

MntSweetProducts: Số tiền khách hàng đã chi ra để mua các sản phẩm tráng miệng trong vòng 2 năm

MntGoldProds: Số tiền khách hàng đã chi ra để mua vàng trong vòng 2 năm

\*Về các chiến dịch khuyến mãi:

NumDealsPurchases: Số lần khách hàng mua hàng được giảm giá

AcceptedCmp1: 1 nếu khách hàng sử dụng ưu đãi trong chiến dịch đầu tiên, ngược lại là 0

AcceptedCmp2: 1 nếu khách hàng sử dụng ưu đãi trong chiến dịch thứ hai, ngược lại là 0

AcceptedCmp3: 1 nếu khách hàng sử dụng ưu đãi trong chiến dịch thứ ba, ngược lại là 0

AcceptedCmp4: 1 nếu khách hàng sử dụng ưu đãi trong chiến dịch thứ tư, ngược lại là 0

AcceptedCmp5: 1 nếu khách hàng sử dụng ưu đãi trong chiến dịch thứ năm, ngược lại là 0

Response: 1 nếu khách hàng sử dụng ưu đãi trong chiến dịch trước, ngược lại là 0

\*Địa điểm:

NumWebPurchases: Số lần khách hàng mua hàng thông qua trang web của cửa hàng

NumCatalogPurchases: Số lần khách hàng mua hàng thông qua danh mục của cửa hàng

NumStorePurchases: Số lần khách hàng mua hàng trực tiếp tại cửa hàng

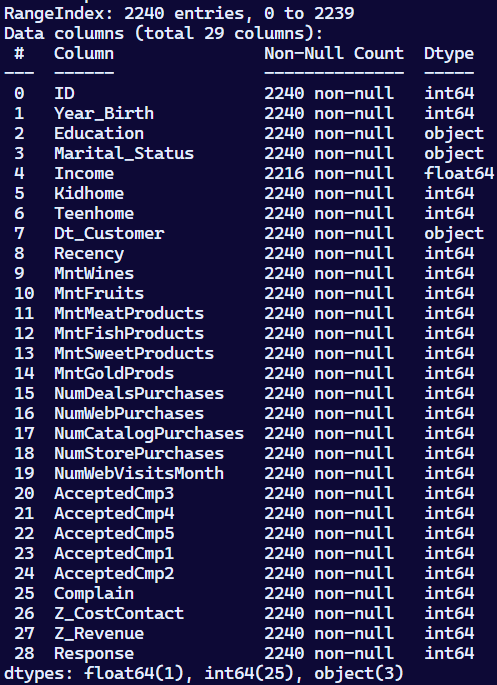
NumWebVisitsMonth: Số lần khách hàng truy cập vào trang web của công ty trong tháng trước

Qua bộ dữ liệu trên, nhóm em có một vài nhận định như sau: Đối với mỗi bản ghi trong bộ dữ liệu, cung cấp đầy đủ thông tin của khách hàng, số lượng tiêu các sản phẩm của khách hàng và bộ dữ liệu này là một bộ dữ liệu không đầy đủ do vẫn còn vài bản ghi có trường để trống.

## **Tiền xử lí dữ liệu**

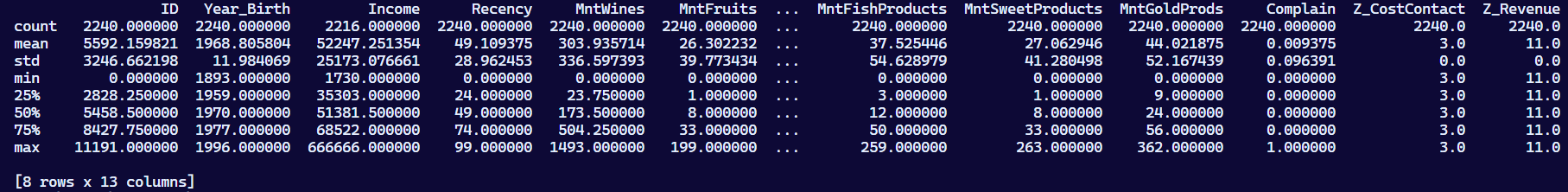
### **Kiểm tra bộ dữ liệu**

Trước khi tiền xử lý dữ liệu, nhóm chúng em tiến hành kiểm tra bộ dữ liệu một cách tổng quan nhất. Trước tiên ta đọc bộ dữ liệu và gán vào biến dataset bằng lệnh ***info()***. Kết quả trả về thu được sẽ là:



Hình 2. 1. Kết quả trả về từ câu lệnh info() đối với bộ dữ liệu

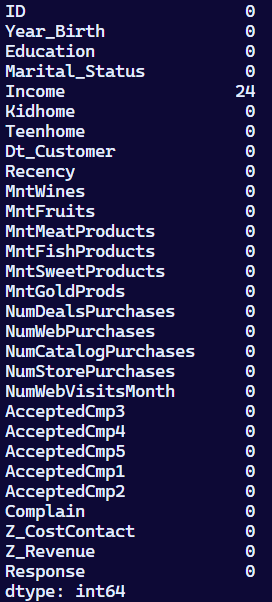
Điều đó cho thấy kiểu dữ liệu tổng thể của toàn bộ bộ dữ liệu trên. Tiếp sau đó là chúng em sử dụng lệnh ***describe()***, lệnh này giúp em đếm tổng số giá trị của các bản ghi có trong bảng dữ liệu. Kết quả trả về sẽ là:



Hình 2. 2. Kết quả trả về từ câu lệnh describe() đối với bộ dữ liệu

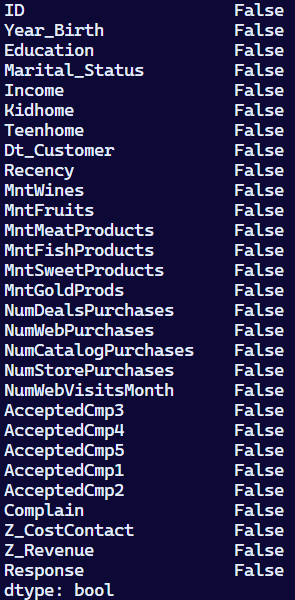
Từ lệnh ***discribe()***, ta có thể thấy được thu nhập tối đa cao hơn rất nhiều so với thu nhập ở tứ phân vị thứ 3 (lớn hơn gần 10 lần). Ngoài ra, số tiền tối đa chi tiêu cho rượu và các sản phẩm từ thịt (Ví dụ: MntWines, MntMeatProduct) lớn hơn đáng kể so với số tiền chi cho các sản phẩm khác

Tiếp theo, em dùng lệnh ***isna().sum()*** để kiểm tra các trường dữ liệu có dữ liệu null không. Kết quả cho ra:



Hình 2. 3. Trường dữ liệu "Income" có 24 bản ghi có giá trị null

Từ bảng trên, ta có thể thấy có trường “income” còn 24 giá trị null. Chúng em sẽ dùng lệnh fillna() để điền vào các giá trị còn trống. Đây là kết quả sau khi điền các giá trị trống:



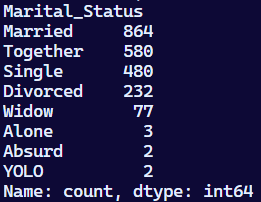
Hình 2. 4. Các trường sau khi được điền các giá trị trống vào

### **Biến đổi và thêm mới dữ liệu**

Để thuận tiện hơn trong quá trình phân tích dữ liệu, chúng em tiến hành biến đổi dữ liệu bằng cách thay đổi một số trường dữ liệu, thay đổi kiểu dữ liệu, tạo dữ liệu mới và kết hợp dữ liệu với nhau tạo ra một dataframe phù hợp cho quá trình phân tích.

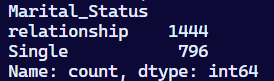
Đầu tiên, từ hình 2.2, chúng ta có thể thấy hai trường “Z\_Contact” và “Z\_Revenue” chưa được dữ liệu mô tả. Em quyết định xoá luôn 2 cột dữ liệu này vì nó không có đóng góp gì trong việc xây dựng mô hình.

Sau đó, chúng em kiểm tra số lượng danh mục duy nhất có trong trường “Marital\_Status”



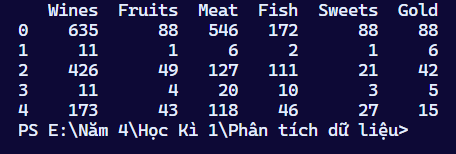
Hình 2. 5. Thông tin từ trường "Marital\_Status"

Nhận thấy quá nhiều giá trị, chúng em dùng lệnh ***replace()*** để thay đổi các giá trị “Maried”, “Together” thành “relationship”, các giá trị còn lại thành “single”



Hình 2.6. Kết quả trả về sau khi thay đổi các giá trị của bảng “Marital\_Status”

Tiếp theo, tạo một bảng DataFrame cho sản phẩm để về sau dễ thực hiện các phương pháp phân tích. Ta được bảng như sau:



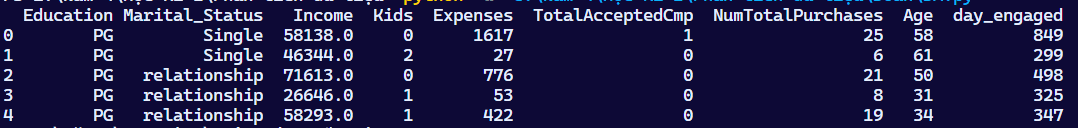
Hình 2. 7. Bảng dữ liệu số lượng sản phẩm tiêu thụ

Sau đó, tiếp tục gọm cụm các trường dữ liệu có nội dung tương tự nhau như trường “Kidhome” và trường “Teenhome” thành một trường “Kids”, gom cụm trường “Expenses” từ các trường “MntWines”, “MntFruits”, “MntMeatProduct”, “MntFishProduct”, “MntSweetsProduct”, “MntGoldProds”. Sau đó, gom các trường tham gia khuyến mãi như “AcceptedCmp1”, “AcceptedCmp2”, … thành 1 trường “TotalAcceptedCmp”. Cuối cùng là các trường “NumWebPurchase”, “NumCatalogPurchases”, … thành trường “NumTotalPurchases”.

Em tạo thêm 1 trường “Age” vì nghĩ nó sẽ rõ ràng hơn trường “Year\_Birth”.

Sau đó, em thay đổi các giá trị trong trường “Education” như “Marital\_Status” để làm giảm độ phức tạp của trường “Education”. Và thay đổi format của Dt\_Customer thành dạng timestamp.

Đồng thời xoá các trường không cần thiết để giảm độ kích thước cũng như độ phức tạp của mô hình. Cuôi cùng ta được bảng dữ liệu như sau:



Hình 2. 8. 5 dòng đầu của bảng dữ liệu sau khi tiền xử lí

## **Thăm dò dữ liệu sau tiền xử lý**

Đối với một nhà phân tích dữ liệu, việc khám phá dữ liệu là một nhiệm vụ quan trọng. Tuy nhiên, việc nhìn vào một cột số cụ thể hay toàn bộ bảng dữ liệu để hiểu rõ về dữ liệu là không dễ dàng. Nếu thực hiện bằng cách thức thủ công, sẽ mất rất nhiều thời gian và mức độ hiệu quả không cao. Vì vậy, EDA – Phân tích khám phá dữ liệu (Exploratory Data Analysis) là một giải pháp hữu ích. EDA sử dụng các kỹ thuật thống kê và trực quan hóa để giúp nhà phân tích dữ liệu hiểu rõ về dữ liệu, trả lời những câu hỏi về dữ liệu và đưa ra những quyết định sáng suốt.

### **Khái niệm về EDA**

Phân tích dữ liệu thăm dò là quá trình mô tả dữ liệu bằng các kỹ thuật thống kê và trực quan hoá nhằm tập trung vào các khía cạnh quan trọng của dữ liệu để tiếp tục phân tích. Điều này bao gồm cả việc kiểm tra tập dữ liệu từ nhiều góc độ, mô tả và tóm tắt nó mà không đưa ra bất kỳ giả định nào khác về nội dung của nó. Trong đề tài của nhóm chúng em sẽ tập trung hiển thị các đặc trưng của dữ liệu trong bộ dữ liệu đã tiền xử lý ở phần trước.

### **Mục đích sử dụng EDA**

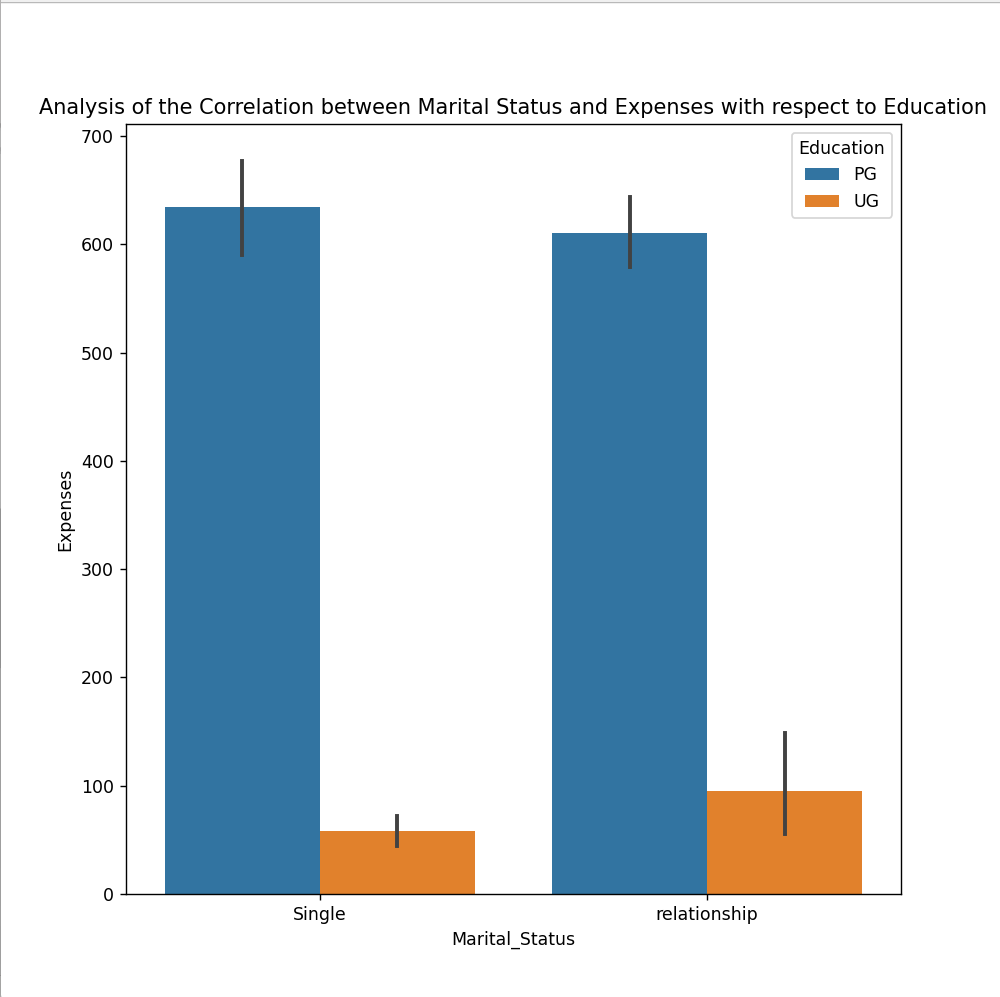
Một số mục đích của việc sử dụng EDA vào các dự án phân tích dữ liệu như:

* Tìm hiểu về cấu trúc dữ liệu: EDA là phương pháp giúp xác định cấu trúc dữ liệu bao gồm số lượng, kiểu dữ liệu, trường dữ liệu, sự liên kết giữa các trường dữ liệu,... Khi xác định được cấu trúc dữ liệu, các nhà phân tích dữ liệu có thể hiểu được mối quan hệ giữa các dữ liệu trong tệp.
* Điều chỉnh và thay đổi: EDA giúp giải quyết các trường hợp thiếu giá trị, dữ liệu lỗi, các ngoại lệ trong dữ liệu. Điều này giúp các nhà phân tích dữ liệu điều chỉnh các phương án khắc phục kịp thời, tránh những ảnh hưởng nghiêm trọng đến dự án.
* Xác định mối tương quan giữa các biến: Các biến đều chứa các giá trị riêng, EDA có khả năng phát hiện các liên hệ tiềm ẩn và sự ảnh hưởng giữa các biến với nhau, tạo sự liên kết giữa các thông tin dữ liệu nhằm xây dựng một quy trình phân tích tổng thể, rõ ràng.
* Xây dựng cơ sở dữ liệu quan hệ: Các đối tượng dữ liệu quan trọng được phát triển mối quan hệ nhằm cấu trúc hóa dữ liệu theo sơ đồ, tiết kiệm thời gian xử lý những thông tin thừa, hạn chế sự sai sót của kết quả phân tích.
* Chuẩn bị cho bước phân tích tiếp theo: Áp dụng EDA giúp loại bỏ các dữ liệu không cần thiết, dữ liệu thiếu giá trị và chuẩn hóa dữ liệu. Đây là yếu tố nền tảng để chuẩn bị cho các bước phân tích bằng thuật toán học máy.

### **Thực nghiệm EDA đối với dữ liệu phân tích**

Để có một cái nhìn bao quát hơn về dữ liệu mà nhóm chúng em đang thực nghiện phân tích, nhóm chúng em tiến hành phân tích và tổng hợp dữ liệu một lần nữa dưới dạng mô hình nhằm đưa ra cái nhìn trực quan hơn về dữ liệu mà chúng em đang làm việc.

#### **Phân tích mối quan hệ giữa “Marital\_Status” và “Expenses” với “Education”**



Hình 2. 9. Số liệu thống kê giữa “Marital\_Status” và “Expenses” với “Education”

Như hình trên (hình 2.9), ta có trực quan hóa toàn bộ khách hàng đang trong tình trạng hôn nhân với mức tiêu thụ dựa trên trình độ học vấn. Do đó, ta có thể thấy các khách hàng đang trong trạng thái độc thân, có trình độ học vấn là PG (Postgraduate) có mức chi tiêu nhiều hơn so với các khách hàng đang trong trạng thái có mối quan hệ, có trình độ học vấn là PG (cụ thể là hơn 30 của cột chi tiêu). Tuy nhiên, trái ngược lại những khách hàng có trình độ học vấn là PG, thì các khách hàng có trình độ học vấn là UG thì những khách đang trong mối quan hệ lại chi tiêu nhiều hơn so với những khách độc thân (cụ thể là hơn 50 của cột chi tiêu). Điều đó cho thấy nhu cầu của mỗi tệp khách hàng khác nhau, nhưng chúng ta nên chú trọng vào những tệp khách có trình độ học vấn là PG, độc thân và các tệp khách có trình độ học vấn là UG, đang trong mối quan hệ.

#### **Phân tích mối quan hệ “Marital\_status” và “Expenses”**

A blue and orange bars

Description automatically generated

Hình 2.10 Số liệu trực quan hóa giữa “Marital\_Status” và “Expenses”

Từ (hình 2.10), ta có trực quan hóa toàn bộ khách hàng đang trong tình trạng hôn nhân dựa trên mức tiêu thụ. Do đó, ta có thể thấy các khách hàng đang trong trạng thái độc thân có mức chi tiêu nhiều hơn so với các khách hàng đang trong trạng thái có mối quan hệ. Điều đó cho thấy nhu cầu của mỗi tệp khách hàng khác nhau nhưng chúng vẫn có mứt tiêu thụ gần bằng nhau.

#### **Phân phối chi phí “Expenses” với “Marital\_status”**

A graph of a number of expenses

Description automatically generated

Hình 2.11 Số liệu phân phối chi phí “Expenses” theo “Marital\_status”

Từ (Hình 2.11) thể hiện sự phân phối của một biến liên tục. Trong trường hợp này, biến liên tục là “Expenses”.Biểu đồ cho thấy rằng những người độc thân có xu hướng chi tiêu ít hơn so với những người đang trong một mối quan hệ và từ đó chỉ ra những người đang trong một mối quan hệ có xu hướng chi tiêu nhiều hơn so với những người độc thân. Điều này có thể do nhiều yếu tố, như nhu cầu tài chính cao hơn do có gia đình, chi phí cho con cái, v.v.

#### **Phân phối chi phí “Expenses” với “Education”**

A graph of a distribution of expenses

Description automatically generated

Hình 2.12 Phân phối chi phí “Expensive” với “Education”

Từ (Hình 2.12) thể hiện sự phân phối của một biến liên tục “Expenses”. Biểu đồ cho thấy rằng những người học sau đại học (PG) có xu hướng chi tiêu nhiều hơn so với những người học đại học (UG). Điều này có thể do nhiều yếu tố, như nhu cầu tài chính cao hơn do học sau đại học.

#### **Phân phối chi phí “NumTotalPurchases” với “Education”**

A graph of a distribution of purchases

Description automatically generated

Hình 2.13 Phân phối chi phí “NumTotalPurchases” với “Education”

Từ (Hình 2.13) thể hiện sự phân phối của một biến liên tục “Expenses”. 'NumTotalPurchases' với 'Education'. Biểu đồ cho thấy rằng những người học sau đại học (PG) có xu hướng mua nhiều hơn so với những người học đại học (UG). Điều này có thể do nhiều yếu tố, như nhu cầu tài chính cao hơn do học sau đại học.

#### **Phân bổ “Age” theo “Marital\_Status”**

A graph of a number of people

Description automatically generated

Hình 2.14 Phân bổ “Age” theo “Marital\_Status”

Từ (Hình 2.14) thể hiện sự phân phối của ‘Age’ (Tuổi) theo ‘Marital\_Status’ (Tình trạng hôn nhân). Biểu đồ cho thấy rằng phân phối tuổi của những người độc thân và những người có mối quan hệ có sự khác biệt. Điều này có thể do nhiều yếu tố, như xu hướng kết hôn ở tuổi nhất định, tuổi trung bình khi bắt đầu một mối quan hệ, v.v.

#### **Phân phối “Income” theo “Marital\_Status”**

A graph of income with numbers and a chart

Description automatically generated with medium confidence

Hình 2.15 Phân phối “Income” theo “Marital\_Status”

Từ (Hình 2.15) thể hiện sự phân phối của ‘Income’ (Tuổi) theo ‘Marital\_Status’ (Tình trạng hôn nhân). Biểu đồ cho thấy rằng phân phối tuổi của những người độc thân và những người có mối quan hệ có sự khác biệt thu nhập khá lớn. Điều này có thể cho thấy nguồn thu nhập chính của sẽ là những người có một mối quan hệ.

#### **Phân tích tỷ lệ khách hàng theo “Marital\_Status”**

A blue and orange pie chart

Description automatically generated

Hình 2.15 Phân tích tỷ lệ khách hàng theo “Marital\_Status”

Từ (Hình 2.15) thể hiện tỉ lệ 35% khách hàng là người độc thân trong khi hơn 64% là khách hàng đang trong mối quan hệ trong biểu đồ hình tròn.

#### **Phân tích tỷ lệ khách hàng theo “Education”**

A blue circle with a orange triangle

Description automatically generated

Hình 2.16 Phân tích tỷ lệ khách hàng theo “Education”

Từ (Hình 2.16) Hơn 97% khách hàng có trong nhóm PG (Postgraduate - Sau đại học). và xấp xỉ. 2% đến từ nhóm UG (Undergraduate - Đại học) được thể hiện trong biểu đồ hình tròn.

#### **Phân tích tổng thu nhập “Income” theo “Education”**

A graph of income based on the education level

Description automatically generated

Hình 2.17 Phân tích tổng thu nhập “Income” theo “Education”

Từ (Hình 2.17) là biểu đồ thể hiện phân tích tổng thu nhập theo trình độ học vấn của khách hàng cho thấy tổng thu nhập khác hàng có trong nhóm PG (Postgraduate - Sau đại học) nhiều gần gấp 3 lần so với những khách hàng có trong nhóm UG (Undergraduate - Đại học).

**Kết luận sau quá trình EDA**

Kết quả thu được sau khi áp dụng phương pháp EDA như sau:

* Mối Quan Hệ Giữa Hôn Nhân, Chi Tiêu và Trình Độ Học Vấn: Những người độc thân có trình độ học vấn sau đại học (PG) có xu hướng chi tiêu nhiều hơn so với những người có mối quan hệ và cũng là PG.Ngược lại, những người có mối quan hệ và trình độ học vấn đại học (UG) chi tiêu nhiều hơn so với những người độc thân UG.
* Mối Quan Hệ Giữa Hôn Nhân và Chi Tiêu: Những người độc thân có chi tiêu cao hơn so với những người có mối quan hệ. Điều này có thể liên quan đến nhu cầu tài chính cao hơn khi có gia đình, trách nhiệm phục vụ cho con cái, v.v.
* Phân Phối Chi Phí và Mối Quan Hệ Hôn Nhân: Biểu đồ phân phối chi phí cho thấy rằng người độc thân có xu hướng chi tiêu ít hơn so với những người có mối quan hệ.
* Phân Phối Chi Phí và Trình Độ Học Vấn: Người học sau đại học (PG) có chi phí chi tiêu cao hơn so với người học đại học (UG). Điều này có thể do nhu cầu tài chính cao hơn của những người học sau đại học.
* Mối Quan Hệ Giữa Thu Nhập và Hôn Nhân: Phân phối thu nhập cho thấy những người có mối quan hệ có thu nhập lớn hơn so với những người độc thân.
* Phân Tích Tổng Thu Nhập Theo Trình Độ Học Vấn: Người học sau đại học (PG) có tổng thu nhập cao gần gấp ba lần so với người học đại học (UG).
* Phân Tích Tỉ Lệ Khách Hàng: Có khoảng 35% khách hàng là người độc thân, trong khi hơn 64% là khách hàng có mối quan hệ.
* Phân Tích Tỉ Lệ Khách Hàng Theo Trình Độ Học Vấn: Hơn 97% khách hàng là người học sau đại học (PG), và xấp xỉ 2% đến từ nhóm học đại học (UG).

Tóm lại, EDA đã giúp chúng ta hiểu rõ hơn về mối quan hệ giữa các biến khác nhau trong dữ liệu, đồng thời cung cấp thông tin quan trọng về xu hướng và đặc điểm của khách hàng. Kết quả này có thể hỗ trợ quá trình đưa ra quyết định và xây dựng mô hình trong các giai đoạn tiếp theo của dự án.

# **Chương 3: BÀI TOÁN VÀ HƯỚNG GIẢI QUYẾT**

## **3.1. Đặt ra bài toán và hướng xử lí**

Đối với dữ liệu mà ta đã tiền xử lý trên, ta đã thu được một bộ dữ liệu chứa các thông tin của các trong trường “Education” như “Marital\_Status” và một bộ dữ liệu chứa các thông tin nhu cầu khách hàng. Với bộ dữ liệu của mình, nhóm chúng em đặt ra 3 bài toán cần được giải quyết như sau:

* Bài toán 1: Cách để tìm các nhóm khách hàng Vip và khách hàng tự do?
* Bài toán 2: Cách tiếp thị những sản phẩm, thu hẹp nhu cầu phù hợp với khách hàng?
* Bài toán 3: Xác định xem khách hàng có mua sản phẩm của công ty hay không?

Đối với bài toán 1, nhóm chúng em có hướng giải quyết là sẽ tìm kiếm và loại các ngoại lệ nhằm xác định hiểu rõ hơn về các giao dịch, hành vi mua sắm của các khách hàng. Tiếp đó chúng em phân cụm (K-means) dựa trên hành vi mua sắm chung, hiểu rõ về đặc điểm và ưu tiên của cụm để tối ưu hóa trải nghiệm mua sắm và tăng cường quản cáo đối với từng nhóm đã được xử lý ở trước.

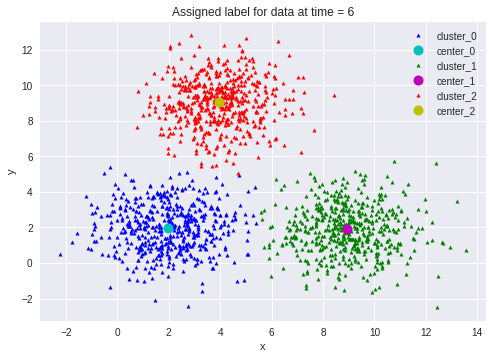
Đối với bài toán 2, nhóm chúng em sẽ lựa chọn phương pháp khai thác quy tắc kết hợp (Thuật toán Apriori) để phát hiện ra sự quan hệ giữa các sản và các yếu tố liên quan như độ tuổi, thu nhập, ngày tham gia, nhằm tối ưu hóa trực tiếp chiến lược đặt hàng và giảm giá.

Đối với bài toán 3, nhóm chúng em sử dụng phân loại (Hồi quy Logistic) để dự đoán cụm khách hàng dựa trên các đặc trưng, giúp tối ưu hóa chiến lược quản cáo và tương tác cá nhân hóa từ đó đánh giá hiệu suất mô hình giúp nhận định độ chính xác của việc dự đoán và hiểu độ hiểu quả của chiến lược tiếp thị đối với từng nhóm.

## **3.2. K-Means**

### **3.2.1. Giới thiệu**

K-means là một thuật toán phân cụm đơn giản thuộc loại học không giám sát(tức là dữ liệu không có nhãn) và được sử dụng để giải quyết bài toán phân cụm. Ý tưởng của thuật toán phân cụm k-means là phân chia 1 bộ dữ liệu thành các cụm khác nhau. Trong đó số lượng cụm được cho trước là k. Công việc phân cụm được xác lập dựa trên nguyên lý: Các điểm dữ liệu trong cùng 1 cụm thì phải có cùng 1 số tính chất nhất định. Tức là giữa các điểm trong cùng 1 cụm phải có sự liên quan lẫn nhau. Đối với máy tính thì các điểm trong 1 cụm đó sẽ là các điểm dữ liệu gần nhau. Thuật toán phân cụm k-means thường được sử dụng trong các ứng dụng cỗ máy tìm kiếm, phân đoạn khách hàng, thống kê dữ liệu, …



Hình 3.1 Kết quả phân cụm của thuật toán kmeans

### **3.2.2. Thuật toán k-means**

Thuật toán phân cụm k-means là một phương pháp được sử dụng trong phân tích tính chất cụm của dữ liệu. Nó đặc biệt được sử dụng nhiều trong khai phá dữ liệu và thống kê. Nó phân vùng dữ liệu thành k cụm khác nhau. Giải thuật này giúp chúng ta xác định được dữ liệu của chúng ta nó thực sử thuộc về nhóm nào.

Ví dụ thực tế như sau: Trong các mô hình kinh doanh, doanh nghiệp sẽ chia nhỏ tệp khách hàng ra thành những nhóm đối tượng khác nhau để có thể áp dụng những chiến lược kinh doanh cụ thể cho từng nhóm đối tượng. Điều này giúp cho khách hàng được tiếp cận với các sản phẩm thật sự phù hợp với bản thân họ. Sự phù hợp đó sẽ kéo doanh số của chúng ta tăng lên. Vấn đề đặt ra là làm sao có thể chia nhỏ tệp khách hàng đó ra khi mà số lượng hóa đơn là rất lớn và chúng ta không thể ngồi để phân tích từng vị khách.

## **3.3. Thuật toán Apriori**

Thuật toán Apriori là một thuật toán quan trọng được sử dụng trong khai thác luật kết hợp (Association Rule Mining) trong dữ liệu. Nó được thiết kế để phân tích mối quan hệ giữa các mục trong tập dữ liệu, đặc biệt là để tìm ra các quy tắc kết hợp (association rules). Dưới đây là một tổng quan về thuật toán Apriori:

* Mục Tiêu Của Apriori: Tìm ra các mối quan hệ liên quan giữa các mục xuất hiện cùng nhau trong tập dữ liệu.
* Cơ Bản Của Thuật Toán:
  + Hỗ Trợ (Support): Là một số liệu thường được sử dụng để đo mức độ phổ biến của một tập hợp các mục trong tập dữ liệu. Hỗ trợ được tính bằng số lần xuất hiện của tập hợp đó chia cho tổng số lượng mẫu.
  + Độ Tin Cậy (Confidence): Đo lường mức độ tin tưởng rằng một quy tắc kết hợp là đúng khi một tập hợp các mục xuất hiện. Độ tin cậy được tính bằng số lần xuất hiện của cả tập hợp mục trong dữ liệu chia cho số lần xuất hiện của tập hợp con.
* Các Bước Chính Của Apriori:
  + Bước 1 (Khởi tạo): Tìm tất cả các mục cá nhân có hỗ trợ lớn hơn một ngưỡng đã đặt trước (min\_support).
  + Bước 2 (Tạo Tập Hợp 2-mục): Tạo tất cả các cặp mục có hỗ trợ lớn hơn min\_support từ các mục cá nhân đã tìm được ở bước 1.
  + Bước 3 (Tạo Tập Hợp k-mục): Lặp lại quá trình để tạo các tập hợp k-mục từ các tập hợp (k-1)-mục đã tạo trước đó, cho đến khi không thể tạo thêm tập hợp mới.
  + Bước 4 (Tìm Quy Tắc): Dựa trên các tập hợp đã tìm được, tạo các quy tắc kết hợp có độ tin cậy lớn hơn một ngưỡng đã đặt trước (min\_confidence).
* Ưu Điểm: Đơn giản và dễ hiểu. Hiệu suất tốt trên dữ liệu với số lượng mẫu lớn.
* Nhược Điểm: Yêu cầu thực hiện nhiều quét qua dữ liệu. Có thể tạo ra quá nhiều quy tắc, đòi hỏi quá trình lọc để chọn ra những quy tắc quan trọng.

**3.4. Hồi quy logistics (Logistiec Regression)**

Hồi quy logistic là một kỹ thuật phân tích dữ liệu sử dụng toán học để tìm ra mối quan hệ giữa hai yếu tố dữ liệu. Sau đó, kỹ thuật này sử dụng mối quan hệ đã tìm được để dự đoán giá trị của những yếu tố đó dựa trên yếu tố còn lại. Dự đoán thường cho ra một số kết quả hữu hạn, như có hoặc không. Ví dụ: giả sử bạn muốn đoán xem khách truy cập trang web của bạn sẽ nhấp vào nút thanh toán trong giỏ hàng của họ hay không. Phân tích hồi quy logistic xem xét hành vi của khách truy cập trước đây, chẳng hạn như thời gian dành cho trang web và số lượng các mặt hàng trong giỏ hàng. Quá trình phân tích này xác định rằng, trước đây, nếu khách truy cập dành hơn năm phút trên trang web và thêm hơn ba mặt hàng vào giỏ hàng, họ sẽ nhấp vào nút thanh toán. Nhờ vào thông tin này, sau đó, hàm hồi quy logistic có thể dự đoán hành vi của một khách mới truy cập trang web.

Hồi quy logistic là một kỹ thuật quan trọng trong lĩnh vực trí tuệ nhân tạo và máy học (AI/ML). Mô hình ML là các chương trình phần mềm có thể được đào tạo để thực hiện các tác vụ xử lý dữ liệu phức tạp mà không cần sự can thiệp của con người. Mô hình ML được xây dựng bằng hồi quy logistic có thể giúp các tổ chức thu được thông tin chuyên sâu hữu ích từ dữ liệu kinh doanh của mình. Họ có thể sử dụng những thông tin chuyên sâu này để phân tích dự đoán nhằm giảm chi phí hoạt động, tăng độ hiệu quả và đổi chỉnh quy mô nhanh hơn. Ví dụ: doanh nghiệp có thể khám phá các mẫu hình cải thiện khả năng giữ chân nhân viên hoặc tạo ra thiết kế sản phẩm mang về nhiều lợi nhuận hơn.

*Tính đơn giản*

Các mô hình hồi quy logistic ít phức tạp về mặt toán học hơn các phương pháp ML khác. Do đó, bạn có thể triển khai chúng ngay cả khi đội ngũ của bạn không ai có chuyên môn sâu về ML.

*Tốc độ*

Các mô hình hồi quy logistic có thể xử lý khối lượng lớn dữ liệu ở tốc độ cao bởi chúng cần ít khả năng điện toán hơn, chẳng hạn như bộ nhớ và sức mạnh xử lý. Điều này khiến các mô hình hồi quy logistic trở nên lý tưởng đối với những tổ chức đang bắt đầu với các dự án ML để đạt được một số thành tựu nhanh chóng.

*Sự linh hoạt*

Bạn có thể sử dụng hồi quy logistic để tìm đáp án cho các câu hỏi có hai hoặc nhiều kết quả hữu hạn. Bạn cũng có thể sử dụng phương pháp này để xử lý trước dữ liệu. Ví dụ: bạn có thể sắp xếp dữ liệu với một phạm vi giá trị lớn, chẳng hạn như giao dịch ngân hàng, thành một phạm vi giá trị hữu hạn, nhỏ hơn nhờ hồi quy logistic. Sau đó, bạn có thể xử lý tập dữ liệu nhỏ hơn này với các kỹ thuật ML khác để phân tích chính xác hơn.

*Khả năng hiển thị*

Phân tích hồi quy logistic cung cấp cho nhà phát triển khả năng nhìn nhận các quy trình phần mềm nội bộ rõ hơn so với các kỹ thuật phân tích dữ liệu khác. Khắc phục sự cố và sửa lỗi cũng trở nên dễ dàng hơn do các phép toán ít phức tạp hơn.

## **3.5. Bài toán 1: Cách để tìm các nhóm khác hàng thân thiết và khác hàng tự do**

### **3.5.1. Tìm các ngoại lệ và loại bỏ**

Thực hiện việc kiểm tra và loại bỏ ngoại lệ từ dữ liệu sử dụng đồ thị Boxplot và phương pháp IQR (Interquartile Range).

*Boxplot (Hình hộp)*

* Hình hộp là một biểu đồ thống kê mô tả được sử dụng để hiển thị phân phối của dữ liệu.
* Boxplot thường bao gồm một hộp, đại diện cho khoảng giữa 25% và 75% của dữ liệu (tương ứng với khoảng IQR).
* Đường giữa hộp là giá trị trung bình (hoặc trung vị) của dữ liệu.
* Ngoại lệ thường được hiển thị dưới dạng các điểm nằm ngoài hai đầu của đường râu (whiskers) hoặc nằm ngoài khoảng IQR.

*Loại bỏ ngoại lệ bằng phương pháp IQR*

* Đầu tiên, đoạn code tính giá trị Q1 (25th percentile) và Q3 (75th percentile) của từng cột trong DataFrame.
* Tính khoảng IQR bằng cách lấy hiệu của Q3 và Q1.
* Xác định giới hạn dưới và giới hạn trên để xác định ngoại lệ.
* Loại bỏ các hàng chứa giá trị nằm ngoài khoảng giới hạn dưới và giới hạn trên.

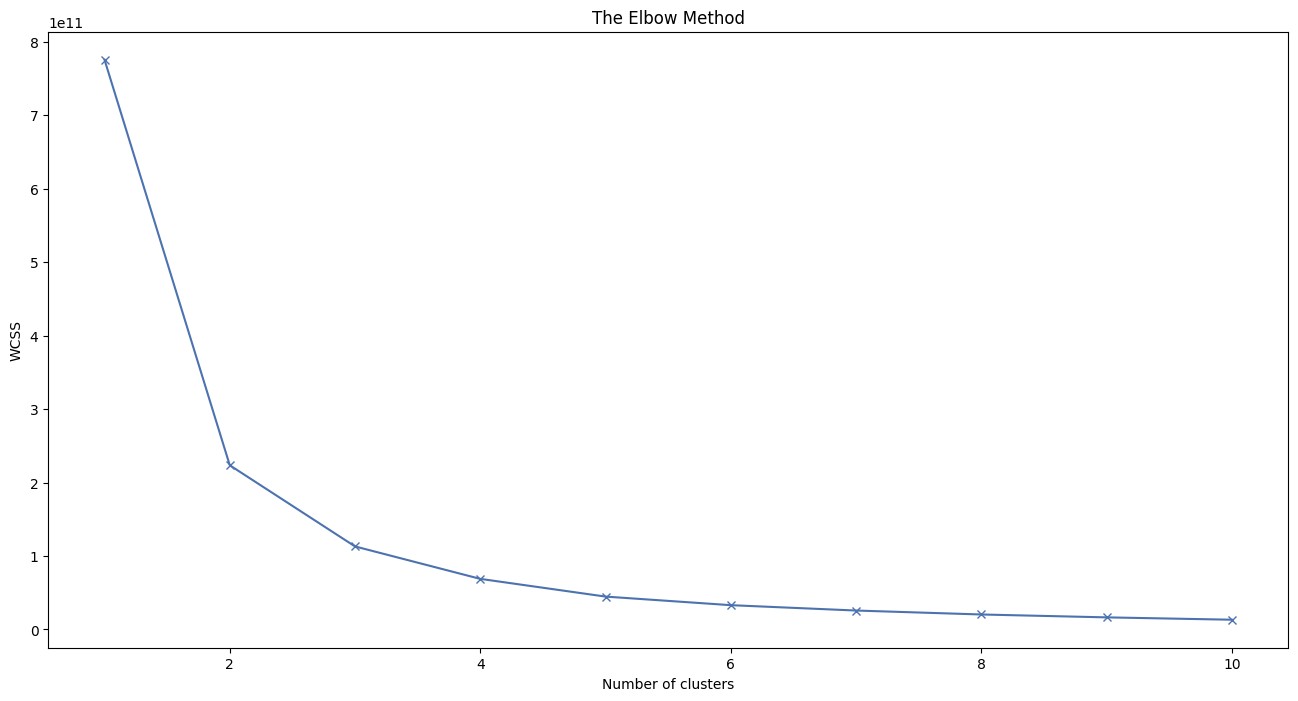
Điều này giúp "lọc" dữ liệu bằng cách loại bỏ những điểm dữ liệu mà theo định nghĩa là ngoại lệ. Việc này có thể cải thiện hiệu suất của các mô hình và phân tích dữ liệu bằng cách loại bỏ ảnh hưởng tiêu cực của các giá trị ngoại lệ đối với các thống kê và phân phối của dữ liệu.

A graph with numbers and a green box

Description automatically generated with medium confidence

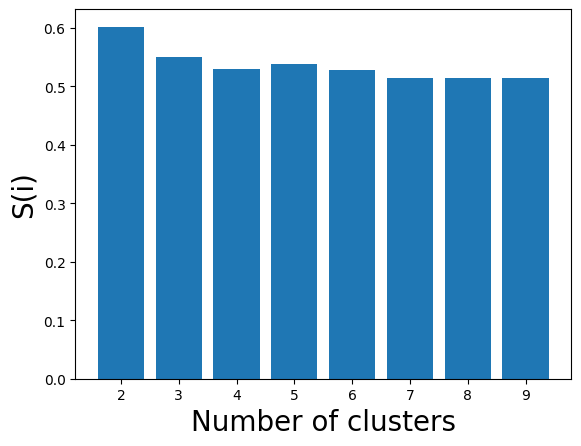
Hình 3.2 Các ngoại lệ trong tập dữ liệu

### **3.5.2. K-means clustering**



Hình 3.3 Mức độ biến động của các điểm dữ liệu

WCSS (Within-Cluster-Sum-of-Squares) là một phép đo cho biết mức độ biến động của các điểm dữ liệu bên trong mỗi cụm. Mục tiêu của phương pháp "The Elbow Method" là tìm giá trị của K (số lượng cụm clusters) sao cho WCSS là nhỏ nhất. Quá trình diễn ra như sau: Thực hiện vòng lặp từ 1 đến 10, tạo mô hình K-Means với mỗi giá trị của K. Mỗi mô hình được khởi tạo với phương pháp k-means++ và seed (random\_state) được đặt là 42 để đảm bảo nhất quán khi chạy lại mô hình. Mỗi mô hình K-Means được huấn luyện trên dữ liệu df1. Đối với mỗi giá trị của K, tính WCSS (Within-Cluster-Sum-of-Squares) và thêm giá trị này vào danh sách wcss. Vẽ đồ thị "The Elbow Method" với trục x là số lượng cụm (clusters) và trục y là giá trị WCSS. Mục tiêu là chọn giá trị K tại điểm " clusters " của đồ thị, nơi mức giảm của WCSS giảm chậm lại. Kết quả là một đồ thị hiển thị giá trị của K mà ở đó có một sự giảm chậm lại đáng kể của WCSS. Giá trị K tương ứng với " clusters " của đồ thị thường được chọn là số lượng cụm tối ưu cho thuật toán K-Means.



Hình 3.4 Sử dụng điểm Silhouette để đo giá trị của K

Sử dụng độ đo Silhouette để đánh giá chất lượng của các cụm tạo bởi thuật toán K-Means với các giá trị khác nhau của K. Điểm Silhouette càng cao, mô hình K-Means càng tốt, với sự phân tách và đồng nhất tốt giữa các cụm.

Tiếp đến ta sử thuật toán K-Means để gán nhãn cụm cho dữ liệu trong DataFrame (df1). Một mô hình K-Means được tạo và huấn luyện với số lượng cụm (number\_of\_clusters) và seed là 42 để đảm bảo kết quả nhất quán. Dự đoán nhãn cụm cho từng điểm dữ liệu trong df1 sử dụng mô hình đã được huấn luyện. Cột "cluster" được thêm vào DataFrame để lưu trữ giá trị của cụm cho mỗi điểm dữ liệu, và giá trị nhãn cụm được bắt đầu từ 1 để thuận tiện cho việc hiểu và phân tích kết quả K-Means.

A black screen with white text

Description automatically generated

Hình 3.5 Kết quả dữ liệu sau khi chạy K-means

## **3.6. Bài toán 2: Cách tiếp thị những sản phẩm, thu hẹp nhu cầu phù hợp với khách hàng**

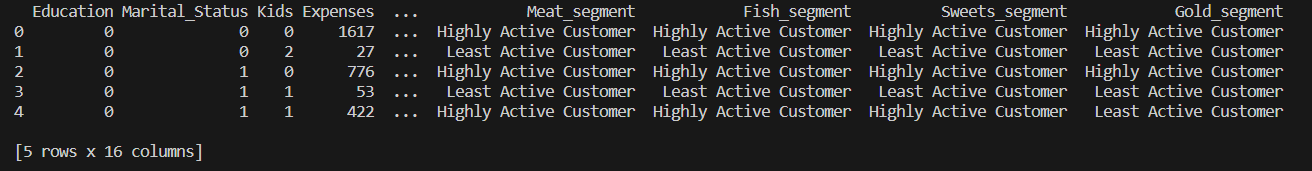
Sử dụng thuật toán này để tìm ra khách hàng nào phù hợp nhất cho một mặt hàng nhất định. Qua đó giúp doanh nghiệp quảng bá đúng đối tượng khách hàng mục tiêu để tăng hiệu quả và tiết kiệm chi phí

*Chuẩn bị dữ liệu*

Xác định các phân khúc mới theo mức chi tiêu của khách hàng cho từng sản phẩm sẽ dựa trên:

Nhóm 1 - Khách hàng năng động cao (Highly Active Customer)

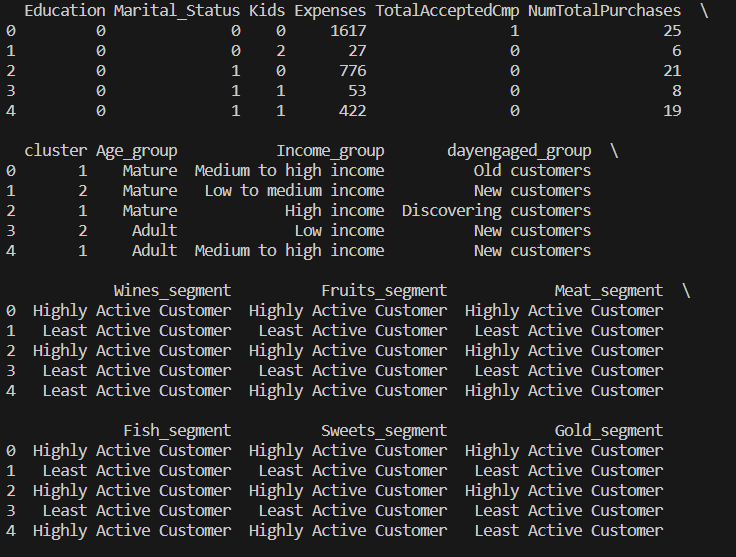
Nhóm 2 - Khách hàng ít hoạt động nhất (Least Active Customer)



Hình 3.7 Dữ liệu sau khi được phân khúc để sử dụng thuật toán apriori

*Áp dụng thuật toán Apriori*

Thực hiện các thiết lập tùy chọn hiển thị của thư viện Pandas để hiển thị toàn bộ nội dung của DataFrame mà không bị cắt bớt. Sau đó, tạo một bản sao của DataFrame và hiển thị năm hàng đầu tiên để kiểm tra kết quả. Điều này giúp làm cho quá trình kiểm tra và phân tích DataFrame trở nên thuận tiện và linh hoạt.



Hình 3.8 Bản sao của DataFrame

Tìm ra khách hàng “Highly Active Customers” khi nói đến “Wine” đầu tiên tìm kiếm và hiển thị luật kết hợp, để tìm kiếm và hiển thị luật kết hợp, trước tiên cần xác định sản phẩm và phân khúc khách hàng mà chúng ta muốn tập trung. Trong trường hợp này, chúng ta muốn tìm hiểu về hành vi mua sắm của các khách hàng thân thiết (VIP) hay mua rượu vang. Sau khi xác định sản phẩm và phân khúc, chúng ta cần tạo một chuỗi mục tiêu. Chuỗi mục tiêu này sẽ được sử dụng để lọc tập luật kết hợp và chỉ hiển thị các luật kết hợp có liên quan đến sản phẩm và phân khúc đã chọn. Cuối cùng, chúng ta có thể lọc và sắp xếp tập luật kết hợp dựa trên độ tin cậy và hiển thị các luật kết hợp có độ tin cậy cao nhất.

A computer screen shot of text

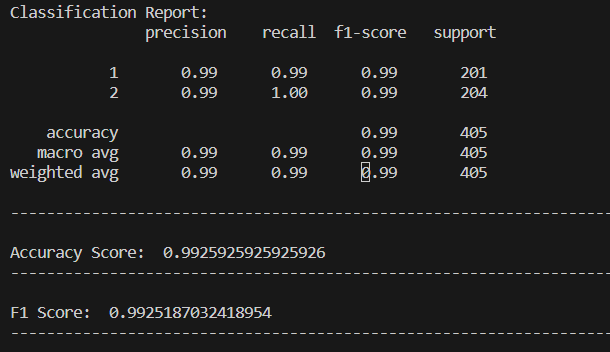
Description automatically generated

Hình 3.9 Kết quả tìm phân khúc khách hàng của “Wine” sử dụng luật kết hợp

## **3.7. Bài toán 3: Xác định xem khách hàng có mua sản phẩm của công ty hay không**

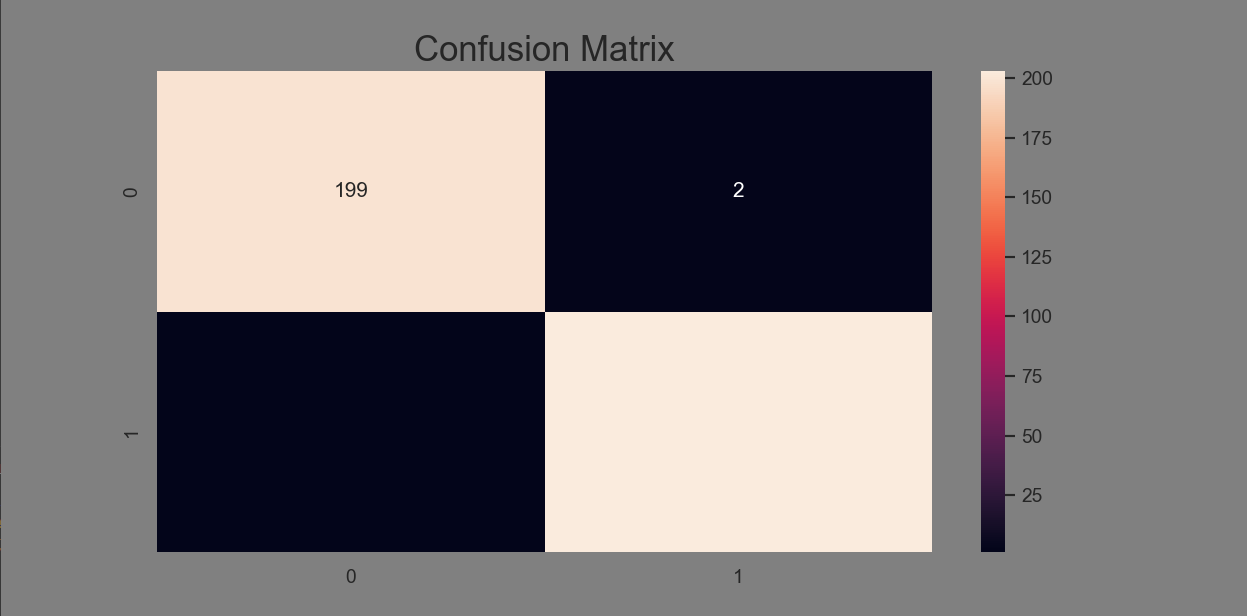
Quá trình huấn luyện một mô hình hồi quy logistic và đánh giá hiệu suất của nó trên tập kiểm tra. Quá trình này bao gồm các bước sau:

* Sao chép DataFrame: Bước đầu tiên là tạo một bản sao DataFrame (df2) từ DataFrame gốc (df1). Điều này cần thiết để đảm bảo rằng dữ liệu gốc vẫn còn nguyên vẹn trong trường hợp có bất kỳ lỗi nào xảy ra trong quá trình huấn luyện mô hình.
* Chuẩn bị dữ liệu cho mô hình: Tiếp theo, chúng ta cần chuẩn bị dữ liệu cho mô hình. Điều này bao gồm việc loại bỏ cột "cluster" khỏi dữ liệu đặc trưng (x) và sử dụng phương pháp train\_test\_split() để chia dữ liệu thành tập huấn luyện và tập kiểm tra.
* Chuẩn hóa dữ liệu: Trước khi huấn luyện mô hình, chúng ta cần chuẩn hóa dữ liệu đặc trưng. Điều này giúp cải thiện hiệu suất của mô hình bằng cách đảm bảo rằng các đặc trưng có cùng quy mô. Chúng ta có thể sử dụng phương pháp StandardScaler() để thực hiện việc này.
* Huấn luyện mô hình: Bước tiếp theo là huấn luyện mô hình trên tập huấn luyện đã được chuẩn hóa. Chúng ta có thể làm điều này bằng cách tạo một đối tượng LogisticRegression() và sử dụng phương thức fit() để huấn luyện mô hình.
* Dự đoán và đánh giá hiệu suất mô hình: Sau khi mô hình được huấn luyện, chúng ta có thể sử dụng nó để dự đoán nhãn trên tập kiểm tra. Chúng ta có thể sử dụng phương thức predict () để thực hiện việc này.
* Cuối cùng, chúng ta có thể đánh giá hiệu suất của mô hình bằng cách sử dụng các chỉ số như điểm chính xác, báo cáo phân loại và ma trận nhầm lẫn. Chúng ta có thể sử dụng các phương thức accuracy\_score(), classification\_report() và confusion\_matrix() để tính toán các chỉ số này. Từ những điều trên cung cấp một ví dụ về cách huấn luyện một mô hình Hồi quy Logistic và đánh giá hiệu suất của nó trên tập kiểm tra. Mã này có thể được sử dụng làm cơ sở để phát triển các ứng dụng phân loại sử dụng Hồi quy Logistic.



Hình 3.10 Kết quả báo cáo phân loại, điểm chính xác, điểm F1 sau khi hồi quy logistic

Cuối cùng thực hiện việc tính toán và hiển thị đồ thị heatmap của ma trận nhầm lẫn (confusion matrix) để đánh giá hiệu suất của mô hình học máy, đặc biệt là trong bài toán phân loại. Heatmap giúp trực quan hóa mức độ chính xác và sai lầm của mô hình đối với từng lớp.



Hình 3.11 Ma trận hỗn loạn

Quan sát: Từ ma trận hỗn loạn ở trên, có thể thấy rằng hồi quy logistic chỉ dự đoán 3 giá trị không chính xác nhưng tất cả các giá trị còn lại đều được dự đoán chính xác, điều đó có nghĩa là trình phân loại hồi quy logistic đã hoạt động khá tốt.

# **KẾT LUẬN**

Trong đồ án này nhóm chúng em đã thực hiện phân tích trên bộ dữ liệu *marketing\_campaign.csv* để giải quyết các mục tiêu đã đề ra.

Bài toán 1: Phân loại khách hàng

Để hiểu rõ hơn về khách hàng của mình, trung tâm thương mại cần phân loại họ thành các nhóm dựa trên hành vi mua sắm. Đồ thị Boxplot được sử dụng để kiểm tra và loại bỏ ngoại lệ từ dữ liệu, giúp cải thiện hiệu suất mô hình phân loại. Thuật toán K-Means sau đó được sử dụng để phân nhóm khách hàng thành các cụm dựa trên các đặc điểm tương đồng. Phương pháp Elbow và Silhouette được sử dụng để xác định số lượng cụm tối ưu.

Bài toán 2: Xác định nhu cầu của khách hàng

Sau khi phân loại khách hàng, trung tâm thương mại có thể sử dụng thuật toán Apriori để tìm hiểu về mối quan hệ mua sắm giữa các sản phẩm và phân khúc khách hàng. Điều này giúp xác định những sản phẩm nào được mua cùng nhau thường xuyên. Các luật kết hợp được lọc và sắp xếp dựa trên độ tin cậy để hiển thị những khách hàng thân thiết mua sản phẩm nhất định.

Bài toán 3: Dự đoán hành vi mua sắm

Mô hình Hồi quy Logistic được sử dụng để dự đoán xác suất mua sản phẩm của khách hàng. Độ chính xác, điểm F1 và ma trận nhầm lẫn được sử dụng để đánh giá hiệu suất của mô hình. Heatmap của ma trận nhầm lẫn giúp trực quan hóa hiệu suất của mô hình.

Từ những quy trình này giúp trung tâm thương mại hiểu rõ hơn về khách hàng của mình, từ việc phân loại họ thành các nhóm, tìm ra mối quan hệ mua sắm, đến việc dự đoán xác suất mua sản phẩm. Điều này có thể hỗ trợ trong việc tối ưu hóa chiến lược tiếp thị và cải thiện trải nghiệm mua sắm của khách hàng.