TRƯỜNG ĐẠI HỌC SÀI GÒN
KHOA TOÁN - ỨNG DỤNG

HỌC PHẦN: KHAI PHÁ DỮ LIỆU

ĐỀ TÀI:
PHÂN TÍCH HÀNH VI MUA SẮM KHÁCH HÀNG

NHÓM THỰC HIỆN: NHÓM 6

SINH VIÊN THỰC HIỆN:

NGUYỄN TRÍ SỰ	**MSSV:** 3123580041
NGUYỄN QUỐC THUẬN	**MSSV:** 3123580048
HUỲNH NHẬT THÀNH	**MSSV:** 3123580044
LÊ THÀNH DANH	**MSSV:** 3123580005

GIẢNG VIÊN HƯỚNG DẪN: TS ĐỖ NHƯ TÀI

Thành phố Hồ Chí Minh - 2025

MỤC LỤC

DANH SÁCH HÌNH VẼ

LỜI MỞ ĐẦU

Trong kỷ nguyên của cuộc Cách mạng Công nghiệp 4.0, dữ liệu không chỉ đơn thuần là những con số lưu trữ trong hệ thống máy tính mà đã trở thành "nguồn dầu mới", đóng vai trò huyết mạch trong việc vận hành và phát triển của mọi doanh nghiệp. Đặc biệt, đối với ngành bán lẻ – một lĩnh vực có sự cạnh tranh khốc liệt và biến động không ngừng về nhu cầu – việc chuyển đổi từ mô hình quản lý dựa trên kinh nghiệm sang mô hình quản lý dựa trên dữ liệu (Data-driven) đã trở thành điều kiện tiên quyết để tồn tại.

Istanbul, với vị thế là trung tâm kinh tế và văn hóa sầm uất, sở hữu hệ thống các trung tâm thương mại quy mô lớn với hàng triệu giao dịch mỗi năm. Đứng trước khối lượng dữ liệu khổng lồ này, các nhà quản lý đối mặt với bài toán làm thế nào để hiểu rõ chân dung từng khách hàng. Đồ án **"Khai phá dữ liệu mua sắm khách hàng tại các trung tâm thương mại Istanbul"** được thực hiện với mục tiêu ứng dụng các kỹ thuật tiên tiến nhất của khoa học dữ liệu nhằm giải mã những câu hỏi đó.

Thông qua việc thực hiện đề tài này, nhóm nghiên cứu không chỉ đặt mục tiêu hoàn thành một sản phẩm học thuật theo quy trình KDD (Knowledge Discovery in Databases) chuẩn mực, mà còn mong muốn tạo ra những giá trị thực tiễn thông qua các đề xuất chiến lược dựa trên bằng chứng số học.

1. TỔNG QUAN VỀ ĐỀ TÀI

1.1. Lý do chọn đề tài và bối cảnh nghiên cứu

Ngành bán lẻ hiện đại đang trải qua một giai đoạn chuyển mình mạnh mẽ khi hành vi khách hàng ngày càng trở nên đa dạng và khó dự đoán. Tại các đô thị lớn như Istanbul, sự hiện diện của hàng loạt trung tâm thương mại tạo ra một môi trường cạnh tranh hoàn hảo, nơi mà sự thấu hiểu khách hàng chính là lợi thế cạnh tranh cốt lõi. Tuy nhiên, thách thức lớn nhất mà doanh nghiệp gặp phải chính là sự "quá tải thông tin". Với hơn 99,457 bản ghi giao dịch được thu thập trong giai đoạn 2021-2023, các phương pháp thống kê mô tả truyền thống chỉ có thể trả lời câu hỏi về những gì đã xảy ra, nhưng không thể giải thích tại sao nó xảy ra hoặc điều gì sẽ xảy ra tiếp theo. Việc lựa chọn đề tài khai phá dữ liệu mua sắm tại Istanbul xuất phát từ nhu cầu cấp thiết về một hệ thống phân tích thông minh có khả năng tự động trích xuất các quy luật tiềm ẩn. Khai phá dữ liệu (Data Mining) đóng vai trò là cầu nối giữa dữ liệu thô và các quyết định chiến lược. Bằng cách áp dụng các thuật toán học máy, chúng ta có thể nhận diện được những nhóm khách hàng có chung đặc điểm hành vi, phát hiện ra những cặp sản phẩm có mối liên hệ mật thiết và dự báo nhu cầu tương lai. Điều này không chỉ giúp tối ưu hóa chi phí vận hành mà còn nâng cao trải nghiệm cá nhân hóa cho người tiêu dùng – yếu tố sống còn trong thời đại kinh tế số.

1.2. Mục tiêu nghiên cứu và nhiệm vụ của đề tài

Nghiên cứu được thiết kế nhằm đạt được một hệ thống mục tiêu toàn diện, đi từ việc thấu hiểu quá khứ đến dự báo tương lai thông qua các kỹ thuật khai phá dữ liệu đặc thù. Nhiệm vụ đầu tiên và quan trọng nhất là thực hiện quy trình phân đoạn khách hàng (Customer Segmentation). Bằng cách sử dụng mô hình RFM (Recency, Frequency, Monetary) kết hợp với thuật toán gom cụm, nhóm nghiên cứu mong muốn định danh được các phân khúc khách hàng mục tiêu, từ đó giúp doanh nghiệp phân bổ nguồn lực marketing một cách hiệu quả hơn. Nhiệm vụ thứ hai tập trung vào khai phá luật kết hợp (Association Rules) để thực hiện phân tích giỏ hàng (Market Basket Analysis). Mục tiêu là tìm ra các quy luật mua sắm đồng thời của khách hàng, ví dụ như việc một khách hàng mua mặt hàng thuộc danh mục "Clothing"thường có xu hướng chọn thêm "Footwear"hay không. Những thông tin này là cơ sở khoa học để sắp xếp gian hàng và thiết kế các gói combo sản phẩm.

Nhiệm vụ thứ ba là xây dựng mô hình phân lớp (Classification) để dự báo danh mục sản phẩm. Nghiên cứu này hướng tới việc tạo ra một "động cơ dự báo"có khả năng dựa vào tuổi tác, giới tính và phương thức thanh toán để đưa ra gợi ý sản phẩm phù hợp nhất với nhu cầu khách hàng tại từng thời điểm. Cuối cùng, nghiên cứu thực hiện phân tích xu hướng thời gian để nắm bắt nhịp độ mua sắm theo mùa vụ và các khung giờ đặc biệt trong tuần.

1.3. Đối tượng, phạm vi và dữ liệu thực nghiệm

Đối tượng nghiên cứu của đồ án là hành vi tiêu dùng được số hóa thông qua các hóa đơn giao dịch tại 10 trung tâm thương mại lớn tại Istanbul. Đây là các địa điểm mua sắm đa dạng, đại diện cho nhiều phân khúc thị trường khác nhau từ bình dân đến cao cấp.

Về phạm vi dữ liệu, nhóm nghiên cứu sử dụng bộ dữ liệu "Customer Shopping Dataset"từ nền tảng Kaggle, bao gồm 99,457 dòng dữ liệu giao dịch trong khoảng thời gian từ năm 2021 đến 2023. Các thuộc tính dữ liệu được khai thác bao gồm thông tin định danh (mã hóa đơn, mã khách hàng), thông tin nhân khẩu học (giới tính, độ tuổi), thông tin giao dịch (danh mục sản phẩm, số lượng, đơn giá, phương thức thanh toán) và thông tin địa lý (tên trung tâm thương mại).

Phạm vi kỹ thuật của đề tài giới hạn trong các thuật toán khai phá dữ liệu phổ biến và có độ tin cậy cao như K-Means cho gom cụm, Apriori/FP-Growth cho luật kết hợp và các thuật toán học máy có giám sát như Decision Tree, Random Forest cho bài toán phân lớp. Quy trình nghiên cứu được thực hiện nghiêm ngặt qua các bước: Tiền xử lý, Phân tích khám phá, Xây dựng mô hình và Đánh giá kết quả.

1.4. Câu hỏi nghiên cứu

Để định hướng cho quá trình thực hiện, nghiên cứu tập trung giải quyết các câu hỏi mang tính học thuật và thực tiễn sau đây:

Thứ nhất, dựa trên các yếu tố về tần suất mua hàng, giá trị giao dịch và thời gian mua sắm gần nhất, có thể phân chia khách hàng thành bao nhiêu nhóm đặc trưng và đặc điểm của từng nhóm là gì?

Thứ hai, những tổ hợp sản phẩm nào thường xuyên xuất hiện cùng nhau trong một hóa đơn và độ tin cậy của các mối liên hệ này có đủ mạnh để áp dụng vào thực tế kinh doanh hay không? Thứ ba, các yếu tố về nhân khẩu học như tuổi tác, giới tính kết hợp với phương thức thanh toán có khả năng dự báo danh mục sản phẩm khách hàng sẽ lựa chọn với độ chính xác bao nhiêu?

Cuối cùng, xu hướng mua sắm của khách hàng có sự biến động rõ rệt như thế nào theo các chu kỳ thời gian trong tuần hoặc trong năm?

1.5. Ý nghĩa khoa học và thực tiễn

Về mặt khoa học, đồ án này đóng góp một quy trình thực nghiệm hoàn chỉnh về Khai phá dữ liệu trong lĩnh vực bán lẻ, từ khâu tiền xử lý dữ liệu thô cho đến khâu trích xuất tri thức và đánh giá mô hình. Nghiên cứu giúp kiểm chứng hiệu quả của các thuật toán học máy trên một tập dữ liệu thực tế có quy mô lớn, cung cấp tư liệu tham khảo cho các nghiên cứu tiếp theo về phân tích hành vi người tiêu dùng. Về mặt thực tiễn, kết quả từ nghiên cứu là cơ sở quan trọng để các nhà quản trị trung tâm thương mại tối ưu hóa vận hành. Việc hiểu rõ các phân khúc khách hàng giúp doanh nghiệp thiết kế các chương trình khuyến mãi trúng đích, giảm thiểu lãng

phí ngân sách marketing. Phân tích giỏ hàng cung cấp gợi ý cho việc sắp xếp mặt bằng gian hàng và quản lý kho bãi. Đồng thời, các mô hình dự báo sẽ giúp doanh nghiệp chủ động chuẩn bị nguồn hàng và nhân sự theo các kịch bản mua sắm khác nhau của khách hàng.

2. TIỀN XỬ LÝ DỮ LIỆU

2.1. Thu thập và khám phá dữ liệu ban đầu

Tập dữ liệu được sử dụng trong nghiên cứu được tải từ tệp dữ liệu gốc dưới định dạng CSV và được đưa vào môi trường phân tích bằng thư viện pandas trong Python. Kết quả sau khi load dữ liệu cho thấy tập dữ liệu bao gồm tổng cộng 99.457 bản ghi và 10 thuộc tính, với dung lượng lưu trữ khoảng 45,45 MB. Mỗi bản ghi tương ứng với một giao dịch mua sắm của khách hàng tại các trung tâm thương mại.

Khảo sát cấu trúc dữ liệu cho thấy tập dữ liệu bao gồm cả các thuộc tính định lượng và định tính. Cụ thể, dữ liệu có hai thuộc tính dạng số nguyên là độ tuổi khách hàng và số lượng sản phẩm mua, một thuộc tính dạng số thực là đơn giá sản phẩm, trong khi các thuộc tính còn lại ở dạng chuỗi ký tự, phản ánh thông tin định danh hóa đơn, khách hàng, danh mục sản phẩm, phương thức thanh toán, thời điểm giao dịch và trung tâm thương mại. Tất cả các thuộc tính trong tập dữ liệu đều có đầy đủ giá trị, không xuất hiện giá trị bị thiếu ở giai đoạn dữ liệu ban đầu.

Việc quan sát các dòng dữ liệu đầu tiên cho thấy mỗi giao dịch đều chứa đầy đủ thông tin liên quan đến khách hàng và chi tiết mua sắm. Các danh mục sản phẩm trong dữ liệu khá đa dạng, bao gồm nhiều nhóm hàng khác nhau như quần áo, giày dép, sách và các nhóm sản phẩm khác. Bên cạnh đó, thông tin về phương thức thanh toán và trung tâm thương mại giúp phản ánh bối cảnh giao dịch thực tế, tạo điều kiện thuận lợi cho việc phân tích hành vi mua sắm ở nhiều khía cạnh khác nhau.

Ngoài ra, thống kê mô tả đối với các thuộc tính định lượng cung cấp cái nhìn tổng quát về phân bố dữ liệu. Độ tuổi khách hàng dao động từ 18 đến 69 tuổi, với giá trị trung bình khoảng 43 tuổi, cho thấy tập dữ liệu bao phủ nhiều nhóm tuổi khác nhau. Số lượng sản phẩm trong mỗi giao dịch nằm trong khoảng từ 1 đến 5 sản phẩm, với giá trị trung bình xấp xỉ 3 sản phẩm mỗi hóa đơn, phản ánh đặc điểm mua sắm phổ biến của khách hàng tại các trung tâm thương mại. Đơn giá sản phẩm có mức độ phân tán lớn, với giá trị trung bình khoảng 689 và độ lệch chuẩn cao, thể hiện sự đa dạng về mức giá giữa các danh mục sản phẩm khác nhau.

Nhìn chung, giai đoạn thu thập và khám phá dữ liệu ban đầu giúp hình thành cái nhìn tổng quan về quy mô, cấu trúc và đặc điểm phân bố của tập dữ liệu. Những kết quả thu được từ bước này đóng vai trò quan trọng trong việc định hướng các bước tiền xử lý tiếp theo, bao gồm làm sạch dữ liệu, biến đổi đặc trưng và chuẩn hóa dữ liệu nhằm phục vụ cho quá trình khai phá dữ liệu và xây dựng mô hình học máy.

2.2. Kiểm tra chất lượng dữ liệu

Sau giai đoạn khám phá dữ liệu ban đầu, bước kiểm tra chất lượng dữ liệu được thực hiện nhằm đánh giá mức độ đầy đủ, tính nhất quán và độ tin cậy của tập dữ liệu trước khi tiến hành các bước tiền xử lý sâu hơn. Việc kiểm tra chất lượng dữ liệu giúp phát hiện sớm các vấn

đề tiềm ẩn có thể ảnh hưởng đến hiệu quả của quá trình phân tích và xây dựng mô hình học máy.

Kết quả kiểm tra cho thấy tập dữ liệu không tồn tại giá trị bị thiếu (missing values) ở bất kỳ thuộc tính nào. Toàn bộ 99.457 bản ghi đều có đầy đủ thông tin ở tất cả các cột, phản ánh quá trình thu thập dữ liệu tương đối hoàn chỉnh và nhất quán. Điều này giúp giảm thiểu nhu cầu áp dụng các kỹ thuật bù giá trị hoặc loại bỏ bản ghi trong giai đoạn tiền xử lý tiếp theo.

Bên cạnh đó, dữ liệu cũng được kiểm tra về sự trùng lặp giữa các bản ghi. Kết quả cho thấy không có bản ghi trùng lặp, với số dòng trùng lặp bằng 0, tương đương 0 tổng số quan sát. Điều này đảm bảo rằng mỗi giao dịch trong tập dữ liệu là duy nhất, tránh hiện tượng lặp dữ liệu có thể gây sai lệch trong việc thống kê, phân tích hành vi khách hàng cũng như huấn luyện mô hình học máy.

Ngoài việc kiểm tra tính đầy đủ và trùng lặp, kiểu dữ liệu của các thuộc tính cũng được rà soát nhằm đảm bảo sự phù hợp với bản chất của từng biến. Các thuộc tính định danh và mô tả như mã hóa đơn, mã khách hàng, giới tính, danh mục sản phẩm, phương thức thanh toán, thời gian giao dịch và trung tâm thương mại được lưu trữ dưới dạng chuỗi ký tự. Trong khi đó, các thuộc tính mang tính định lượng như độ tuổi và số lượng sản phẩm được lưu trữ dưới dạng số nguyên, còn đơn giá sản phẩm được biểu diễn dưới dạng số thực. Cách phân loại kiểu dữ liệu này là phù hợp với ý nghĩa thực tế của các thuộc tính và đáp ứng yêu cầu đầu vào của các phương pháp phân tích dữ liệu và học máy.

Nhìn chung, kết quả kiểm tra chất lượng dữ liệu cho thấy tập dữ liệu có chất lượng tốt, không tồn tại các vấn đề nghiêm trọng về thiếu dữ liệu, trùng lặp hay sai lệch kiểu dữ liệu. Đây là tiền đề thuận lợi cho các bước làm sạch, biến đổi và chuẩn hóa dữ liệu trong các giai đoạn tiếp theo của quy trình khai phá dữ liệu.

2.3. Làm sạch dữ liệu

Sau khi hoàn tất bước kiểm tra chất lượng dữ liệu, quá trình làm sạch dữ liệu được tiến hành nhằm đảm bảo tập dữ liệu đạt được mức độ nhất quán và phù hợp cho các bước phân tích và mô hình hóa tiếp theo. Các thao tác làm sạch được thực hiện có chọn lọc, dựa trên đặc điểm thực tế của dữ liệu giao dịch, thay vì áp dụng các phương pháp loại bỏ một cách cứng nhắc.

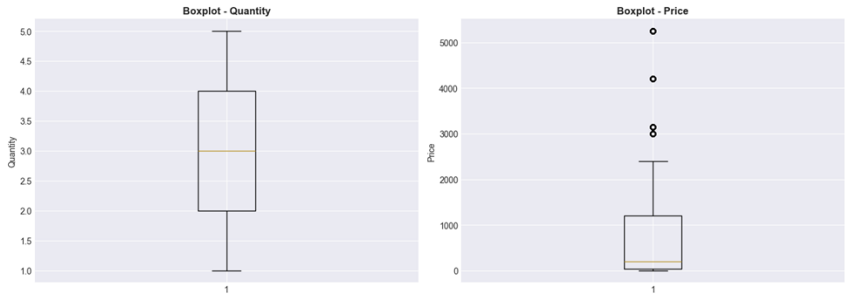
Trước hết, dữ liệu được kiểm tra và xử lý các giá trị bị thiếu. Kết quả cho thấy tập dữ liệu không tồn tại giá trị thiếu ở bất kỳ thuộc tính nào, do đó không cần áp dụng các phương pháp bù giá trị hay loại bỏ bản ghi trong giai đoạn này. Việc dữ liệu đầy đủ ngay từ ban đầu giúp đảm bảo tính toàn vẹn của tập dữ liệu và giảm thiểu rủi ro sai lệch trong quá trình phân tích.

Tiếp theo, dữ liệu được kiểm tra về sự trùng lặp giữa các bản ghi. Kết quả xử lý cho thấy không có dòng dữ liệu trùng lặp nào được phát hiện, do đó không có bản ghi nào bị loại bỏ ở bước này. Điều này đảm bảo rằng mỗi bản ghi trong tập dữ liệu tương ứng với một giao dịch mua sắm duy nhất, góp phần nâng cao độ tin cậy của các kết quả phân tích sau này.

Bên cạnh đó, một số thuộc tính được chuyển đổi kiểu dữ liệu nhằm phù hợp hơn với bản chất thông tin mà chúng biểu diễn. Cụ thể, thuộc tính invoice date được chuyển đổi từ dạng chuỗi ký tự sang kiểu dữ liệu datetime, tạo điều kiện thuận lợi cho các phân tích liên quan đến

yếu tố thời gian như xu hướng mua sắm theo ngày, tháng hoặc năm trong các bước nghiên cứu tiếp theo.

Tiếp theo, dữ liệu được phân tích để phát hiện và đánh giá các giá trị ngoại lai (outliers) thông qua phương pháp khoảng tứ phân vị (IQR), kết hợp với trực quan hóa bằng biểu đồ hộp (boxplot).



Hình 1: Biểu đồ boxplot của biến quantity và price

Kết quả phân tích cho thấy đối với thuộc tính `quantity`, không tồn tại giá trị ngoại lai, với toàn bộ giá trị nằm trong khoảng hợp lệ từ -1 đến 7. Điều này phản ánh số lượng sản phẩm mua trong mỗi giao dịch tương đối ổn định và phù hợp với thực tế mua sắm tại các trung tâm thương mại.

Đối với thuộc tính `price`, biểu đồ boxplot cho thấy sự xuất hiện của một số giá trị ngoại lai, với tổng cộng 5.024 quan sát, chiếm khoảng $5,05\%$ tổng số bản ghi. Các giá trị này nằm ngoài khoảng hợp lệ được xác định bởi phương pháp IQR, từ $-1686,85$ đến $2932,62$. Các giá trị giá cao này có thể phản ánh những giao dịch hợp lệ liên quan đến các sản phẩm có giá trị lớn hoặc các đơn hàng đặc biệt, thay vì là lỗi dữ liệu.

Trong phạm vi nghiên cứu này, các giá trị ngoại lai của thuộc tính price không bị loại bỏ mà được giữ lại nhằm bảo toàn thông tin và phản ánh đúng bản chất thực tế của dữ liệu giao dịch. Cách tiếp cận này giúp hạn chế việc làm mất thông tin quan trọng và tránh gây sai lệch trong các bước phân tích và xây dựng mô hình học máy tiếp theo.

Tóm lại, quá trình làm sạch dữ liệu đã giúp xác nhận tập dữ liệu có chất lượng tốt, không tồn tại các vấn đề nghiêm trọng về thiếu giá trị, trùng lặp hay sai lệch kiểu dữ liệu, đồng thời đưa ra cách xử lý hợp lý đối với các giá trị ngoại lai. Đây là nền tảng quan trọng cho các bước biến đổi đặc trưng và chuẩn hóa dữ liệu trong giai đoạn tiếp theo.

2.4. Biến đổi và tạo đặc trưng

Sau khi hoàn tất quá trình làm sạch dữ liệu, bước biến đổi và tạo đặc trưng được thực hiện nhằm mở rộng không gian biểu diễn dữ liệu và cung cấp thêm các thông tin có ý nghĩa phục vụ cho quá trình phân tích và xây dựng mô hình học máy. Các đặc trưng mới được tạo ra dựa trên kiến thức miền bài toán và đặc điểm thực tế của dữ liệu giao dịch, giúp phản ánh hành vi mua sắm của khách hàng một cách toàn diện hơn.

Trước hết, một biến mới mang tên total amount được tạo ra bằng cách nhân số lượng sản phẩm (quantity) với đơn giá (price) của mỗi giao dịch. Biến này đại diện cho tổng giá trị chi

tiêu của khách hàng trong từng hóa đơn, đóng vai trò quan trọng trong việc đánh giá mức độ đóng góp doanh thu của mỗi giao dịch. Kết quả thống kê cho thấy giá trị total amount dao động từ 5,23 USD đến 26.250,00 USD, với giá trị trung bình đạt 2.528,79 USD và trung vị là 600,17 USD. Sự chênh lệch lớn giữa giá trị trung bình và trung vị cho thấy phân bố của tổng chi tiêu có xu hướng lệch phải, phản ánh sự tồn tại của một nhóm giao dịch có giá trị rất cao.

Tiếp theo, các đặc trưng liên quan đến yếu tố thời gian được trích xuất từ thuộc tính invoice date sau khi đã được chuyển đổi sang kiểu dữ liệu datetime. Các đặc trưng này bao gồm năm, tháng, ngày, thứ trong tuần, tên ngày, giờ, quý trong năm và biến nhị phân is weekend dùng để xác định giao dịch diễn ra vào cuối tuần hay không. Việc bổ sung các đặc trưng thời gian giúp mô hình có khả năng nắm bắt các quy luật và xu hướng mua sắm theo thời gian, chẳng hạn như sự khác biệt trong hành vi tiêu dùng giữa các ngày trong tuần và cuối tuần hoặc giữa các giai đoạn trong năm.

Bên cạnh đó, khách hàng được phân nhóm theo độ tuổi thông qua việc tạo biến phân loại age group. Các nhóm tuổi bao gồm Teen, Young Adult, Adult, Middle Age và Senior, phản ánh các giai đoạn khác nhau trong vòng đời tiêu dùng của khách hàng. Kết quả phân bố cho thấy nhóm Middle Age chiếm tỷ trọng lớn nhất với 38.084 khách hàng, tiếp theo là nhóm Adult với 28.834 khách hàng và nhóm Senior với 19.043 khách hàng. Việc phân nhóm độ tuổi giúp đơn giản hóa dữ liệu liên tục, đồng thời hỗ trợ tốt hơn cho các phân tích phân khúc khách hàng và các mô hình phân loại sau này.

Một bước quan trọng khác trong quá trình tạo đặc trưng là tính toán các chỉ số RFM (Recency, Frequency, Monetary), vốn được sử dụng phổ biến trong phân tích hành vi khách hàng và quản trị quan hệ khách hàng. Chỉ số Recency được xác định là số ngày kể từ lần mua gần nhất của khách hàng, Frequency đại diện cho số lần mua hàng, và Monetary phản ánh tổng giá trị chi tiêu của khách hàng. Kết quả thống kê cho thấy giá trị Recency trung bình là 396,88 ngày, trong khi Frequency có giá trị không đổi là 1 đối với tất cả các khách hàng, phản ánh đặc điểm của tập dữ liệu giao dịch theo hóa đơn đơn lẻ. Chỉ số Monetary có giá trị trung bình là 2.528,79 USD, với giá trị lớn nhất đạt 26.250,00 USD. Sau khi tính toán, các chỉ số RFM được tích hợp trở lại tập dữ liệu chính, tạo nền tảng cho các phân tích phân cụm và đánh giá giá trị khách hàng.

Cuối cùng, thuộc tính price được phân loại thành các nhóm khoảng giá thông qua biến phân loại price range, bao gồm Very Low, Low, Medium, High và Very High. Kết quả cho thấy nhóm Very High chiếm tỷ lệ lớn nhất với 42.676 giao dịch, tiếp theo là nhóm Very Low với 26.762 giao dịch. Việc phân chia khoảng giá giúp giảm độ phức tạp của biến liên tục, đồng thời hỗ trợ mô hình trong việc nhận diện các nhóm sản phẩm theo mức giá và hành vi chi tiêu của khách hàng.

Tổng hợp lại, quá trình biến đổi và tạo đặc trưng đã giúp làm giàu tập dữ liệu ban đầu bằng cách bổ sung các biến phản ánh giá trị giao dịch, yếu tố thời gian, đặc điểm nhân khẩu học và hành vi mua sắm của khách hàng. Những đặc trưng này đóng vai trò then chốt trong việc nâng cao khả năng biểu diễn dữ liệu và hiệu quả của các mô hình học máy được xây dựng trong các chương tiếp theo.

2.5. Chuẩn hóa và mã hóa dữ liệu

Sau khi hoàn tất các bước biến đổi và tạo đặc trưng, dữ liệu tiếp tục được xử lý thông qua quá trình chuẩn hóa và mã hóa nhằm chuyển đổi dữ liệu về dạng số phù hợp cho các thuật toán học máy. Do đa số các mô hình học máy không thể làm việc trực tiếp với dữ liệu dạng chuỗi, việc mã hóa các biến phân loại là bước cần thiết để đảm bảo khả năng học và tối ưu hóa của mô hình.

Trong nghiên cứu này, phương pháp **Label Encoding** được áp dụng cho các thuộc tính phân loại bao gồm `gender`, `category`, `payment_method`, `shopping_mall` và `age_group`. Mỗi giá trị phân loại trong từng thuộc tính được ánh xạ sang một giá trị số nguyên duy nhất, tạo ra các biến mới tương ứng với hậu tố `_encoded`. Cụ thể, các biến `gender_encoded`, `category_encoded`, `payment_method_encoded`, `shopping_mall_enc` và `age_group_encoded` lần lượt được tạo ra từ các biến gốc tương ứng.

Việc sử dụng Label Encoding giúp giảm đáng kể độ phức tạp của dữ liệu và đảm bảo tính nhất quán trong quá trình xử lý, đặc biệt phù hợp với các thuật toán học máy như cây quyết định, random forest hoặc các phương pháp phân cụm. Đồng thời, các bộ mã hóa được lưu trữ riêng biệt để có thể tái sử dụng trong các giai đoạn huấn luyện và suy diễn mô hình sau này, đảm bảo tính tái lập và ổn định của quy trình.

Tóm lại, quá trình chuẩn hóa và mã hóa dữ liệu đã chuyển đổi tập dữ liệu ban đầu sang dạng số hóa hoàn chỉnh, đồng thời đảm bảo tính nhất quán và khả năng mở rộng của dữ liệu. Đây là bước tiền đề quan trọng giúp nâng cao hiệu quả huấn luyện, độ ổn định và khả năng hội tụ của các mô hình học máy được trình bày trong các chương sau.

2.6. Xuất dữ liệu sau tiền xử lý

Sau khi hoàn tất toàn bộ các bước tiền xử lý, biến đổi, tạo đặc trưng và mã hóa dữ liệu, tập dữ liệu cuối cùng được xuất ra nhằm phục vụ cho các giai đoạn phân tích và xây dựng mô hình học máy ở các chương tiếp theo. Việc lưu trữ dữ liệu đã được xử lý không chỉ giúp đảm bảo tính tái lập của nghiên cứu mà còn tạo điều kiện thuận lợi cho việc tái sử dụng dữ liệu trong các thí nghiệm khác nhau.

Tập dữ liệu đã được làm sạch và chuẩn hóa được lưu dưới dạng tệp `cleaned_data.csv` tại thư mục `../data/processed/`. Kết quả cho thấy bộ dữ liệu có kích thước 99.457 dòng và 29 cột, phản ánh sự mở rộng đáng kể so với tập dữ liệu ban đầu thông qua quá trình tạo đặc trưng. Dung lượng của tệp sau khi xử lý là $59,84$ MB, cho thấy dữ liệu đã được làm giàu thêm về mặt thông tin nhưng vẫn đảm bảo khả năng lưu trữ và xử lý hiệu quả.

Bộ dữ liệu sau tiền xử lý bao gồm đầy đủ các nhóm thuộc tính quan trọng. Nhóm thuộc tính gốc chứa các thông tin nhận dạng và giao dịch như mã hóa đơn, mã khách hàng, giới tính, độ tuổi, danh mục sản phẩm, số lượng, giá, phương thức thanh toán, thời gian giao dịch và trung tâm thương mại. Nhóm thuộc tính mở rộng bao gồm các biến được tạo mới như `total_amount`, các đặc trưng thời gian (năm, tháng, ngày, thứ trong tuần, quý, cuối tuần), các nhóm độ tuổi

(age_group), các chỉ số RFM (recency, frequency, monetary) và các nhóm khoảng giá (price_range). Bên cạnh đó, các biến phân loại quan trọng cũng đã được mã hóa sang dạng số thông qua các thuộc tính có hậu tố _encoded, giúp dữ liệu sẵn sàng cho các thuật toán học máy.

Phân tích từ bảng từ điển dữ liệu cho thấy toàn bộ các thuộc tính trong tập dữ liệu sau xử lý đều không tồn tại giá trị thiếu, đảm bảo tính toàn vẹn và nhất quán của dữ liệu. Số lượng giá trị phân biệt của từng thuộc tính phản ánh rõ đặc điểm của dữ liệu giao dịch, chẳng hạn như giới tính chỉ gồm hai giá trị, danh mục sản phẩm gồm tám nhóm, trung tâm thương mại gồm mười địa điểm khác nhau và biến (price_range) được chia thành năm khoảng giá. Điều này cho thấy dữ liệu đã được tổ chức và chuẩn hóa một cách hợp lý, thuận lợi cho các bước phân tích tiếp theo.

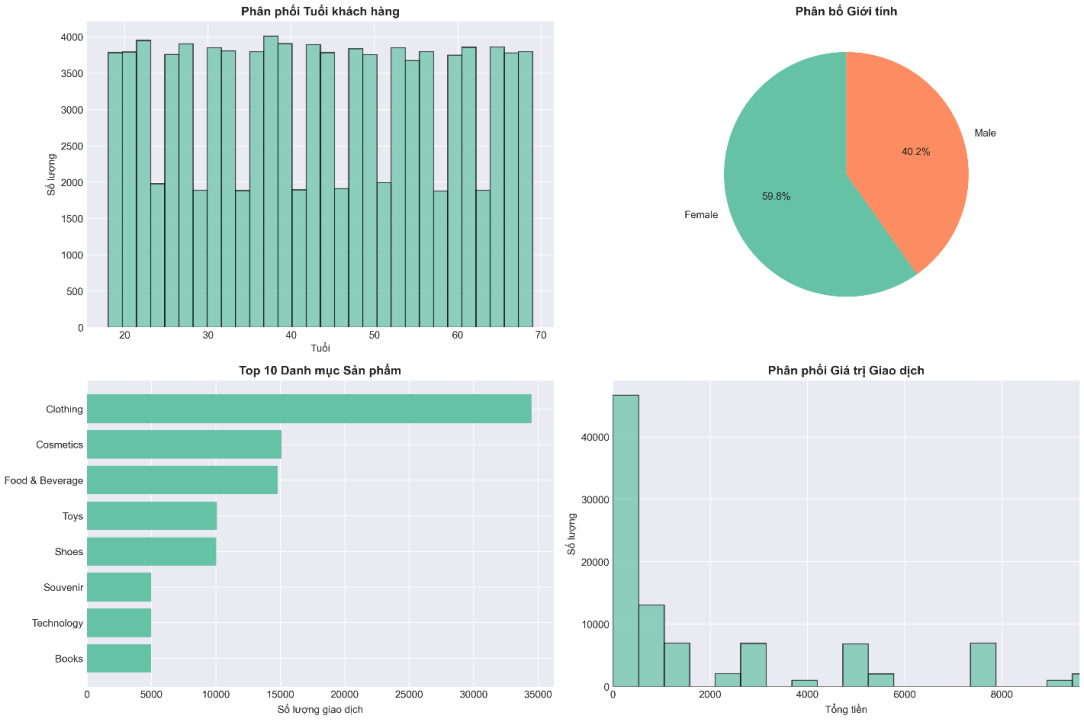
Ngoài ra, các bộ mã hóa nhãn (label encoders) được sử dụng trong quá trình mã hóa các biến phân loại cũng được lưu lại dưới dạng tệp (label_encoders.pkl). Việc lưu trữ này giúp đảm bảo tính nhất quán trong quá trình huấn luyện và triển khai mô hình, cho phép tái sử dụng cùng một cơ chế mã hóa khi áp dụng mô hình trên dữ liệu mới trong tương lai.

Tóm lại, bước xuất dữ liệu sau tiền xử lý đã hoàn tất quá trình chuẩn bị dữ liệu cho nghiên cứu, tạo ra một tập dữ liệu hoàn chỉnh, giàu thông tin và sẵn sàng cho các bước phân tích nâng cao và xây dựng mô hình học máy trong các chương tiếp theo.

3. PHÂN TÍCH KHÁM PHÁ DỮ LIỆU (EDA) VÀ CLUSTERING

3.1. Phân Tích Khám Phá Dữ Liệu (EDA):

3.1.1 Univariate Analysis



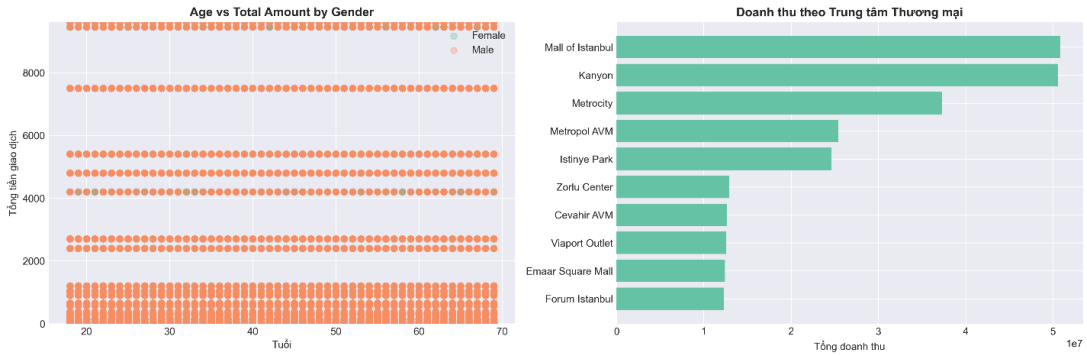
Hình 2: Phân bố độ tuổi và Top danh mục sản phẩm

Khi phân tích về độ tuổi, chúng ta nhận thấy biểu đồ phân bố độ tuổi của khách hàng trải rộng và khá đồng đều từ 18 đến 69 tuổi. Khác với nhiều thị trường bán lẻ khác thường tập trung mạnh vào giới trẻ (Gen Z) hoặc nhóm lao động chính, dữ liệu tại Istanbul không cho thấy sự tập trung quá mức (skewness) vào bất kỳ nhóm tuổi cụ thể nào. Điều này minh chứng rằng các trung tâm thương mại tại đây đóng vai trò là điểm đến phổ biến cho mọi lứa tuổi, từ sinh viên, người đi làm cho đến người cao tuổi. Do đó, các chiến lược sản phẩm và dịch vụ cần phải được thiết kế đa dạng để phục vụ được nhu cầu của cả gia đình đa thế hệ, thay vì chỉ nhắm vào một phân khúc hẹp.

Về phương thức thanh toán, dữ liệu phản ánh sự thống trị tuyệt đối của Tiền mặt (Cash) và Thẻ tín dụng (Credit Card) trong tổng số các giao dịch. Trong khi đó, tỷ lệ sử dụng Thẻ ghi nợ (Debit Card) lại thấp hơn đáng kể. Thực tế này phản ánh thói quen tài chính đặc thù của người tiêu dùng địa phương. Thẻ tín dụng được ưa chuộng mạnh mẽ có thể do sự hấp dẫn của các chương trình trả góp, tích điểm thưởng hoặc hoàn tiền mà các ngân hàng cung cấp. Ngược lại, tiền mặt vẫn giữ vị thế "vua"trong các giao dịch giá trị nhỏ hoặc trung bình, cho thấy một bộ phận lớn khách hàng vẫn giữ thói quen tiêu dùng truyền thống.

Đối với cơ cấu danh mục sản phẩm, ba nhóm hàng chiếm tỷ trọng giao dịch lớn nhất lần lượt là Quần áo (Clothing), Mỹ phẩm (Cosmetics) và Thực phẩm & Đồ uống (Food & Beverage). Đây là các nhóm hàng tiêu dùng nhanh và thiết yếu, đóng vai trò tạo nên dòng tiền (cash flow) ổn định và thường xuyên cho các trung tâm thương mại. Mặc dù các nhóm hàng giá trị cao hoặc đặc thù như Công nghệ hay Sách chiếm tỷ trọng nhỏ hơn về số lượng giao dịch, nhưng chúng lại có thể mang lại biên lợi nhuận khác biệt và thu hút những nhóm khách hàng ngách (niche market) quan trọng.

3.1.2 Phân Tích Tương Quan Đa Biến (Bivariate Analysis)



Hình 3: Mối quan hệ giữa Giới tính, Chi tiêu và Địa điểm

Phân tích mối quan hệ giữa Giới tính và Chi tiêu cho thấy một bức tranh thú vị. Khi so sánh tổng giá trị chi tiêu (Total Spend), ta có thể thấy rằng không có thấy sự chênh lệch quá lớn giữa nam và nữ, hay nói cách khác, khoảng cách giới (Gender Gap) về doanh thu là không đáng kể. Tuy nhiên, sự khác biệt lại nằm rõ rệt ở cơ cấu giỏ hàng. Khách hàng nữ chi tiêu vượt trội ở nhóm Quần áo và Mỹ phẩm, thể hiện nhu cầu làm đẹp và thời trang cao. Trong khi đó, nam giới có xu hướng chi tiêu rải rác hơn hoặc tập trung vào các nhóm hàng công nghệ và ăn uống. Điều này gợi ý rằng các chiến dịch quảng cáo chéo (cross-selling) cần được tùy chỉnh theo giới tính để đạt hiệu quả cao nhất; ví dụ, quảng cáo mỹ phẩm cho nữ giới sẽ hiệu quả hơn hẳn so với chạy đại trà.

Về hiệu suất hoạt động của các địa điểm, có sự phân hóa mạnh mẽ giữa các trung tâm thương mại. Kanyon, Mall of Istanbul, và Metrocity nổi lên là những "ông lớn"với lưu lượng giao dịch và tổng doanh thu cao nhất, bỏ xa các đối thủ còn lại. Sự chênh lệch này có thể được giải thích bởi vị trí địa lý đắc địa, quy mô diện tích sàn lớn, hoặc sự hiện diện của các thương hiệu neo (anchor tenants) nổi tiếng thu hút khách hàng. Các trung tâm thương mại nhỏ hơn đang đối mặt với áp lực cạnh tranh gay gắt và cần phải xem xét lại chiến lược định vị thương hiệu để tồn tại.

Khi xem xét Ma trận Tương quan (Correlation Matrix), ta phát hiện ra một sự thật quan trọng: mối tương quan giữa Tuổi tác (Age) và Điểm chi tiêu/Tổng chi tiêu (Spending Score/Total Spend) gần như bằng 0. Đây là một phát hiện mang tính bước ngoặt, khẳng định rằng tuổi tác không phải là yếu tố dự báo tốt cho khả năng chi tiêu. Một người trẻ 20 tuổi hoàn toàn có thể

mua sắm hàng hiệu đắt tiền ngang với một người trung niên 50 tuổi. Do đó, việc phân khúc khách hàng chỉ dựa đơn thuần vào nhân khẩu học (Demographic Segmentation) là không đủ và thiếu chính xác, mà bắt buộc phải chuyển sang phân khúc dựa trên hành vi thực tế (Behavioral Segmentation).

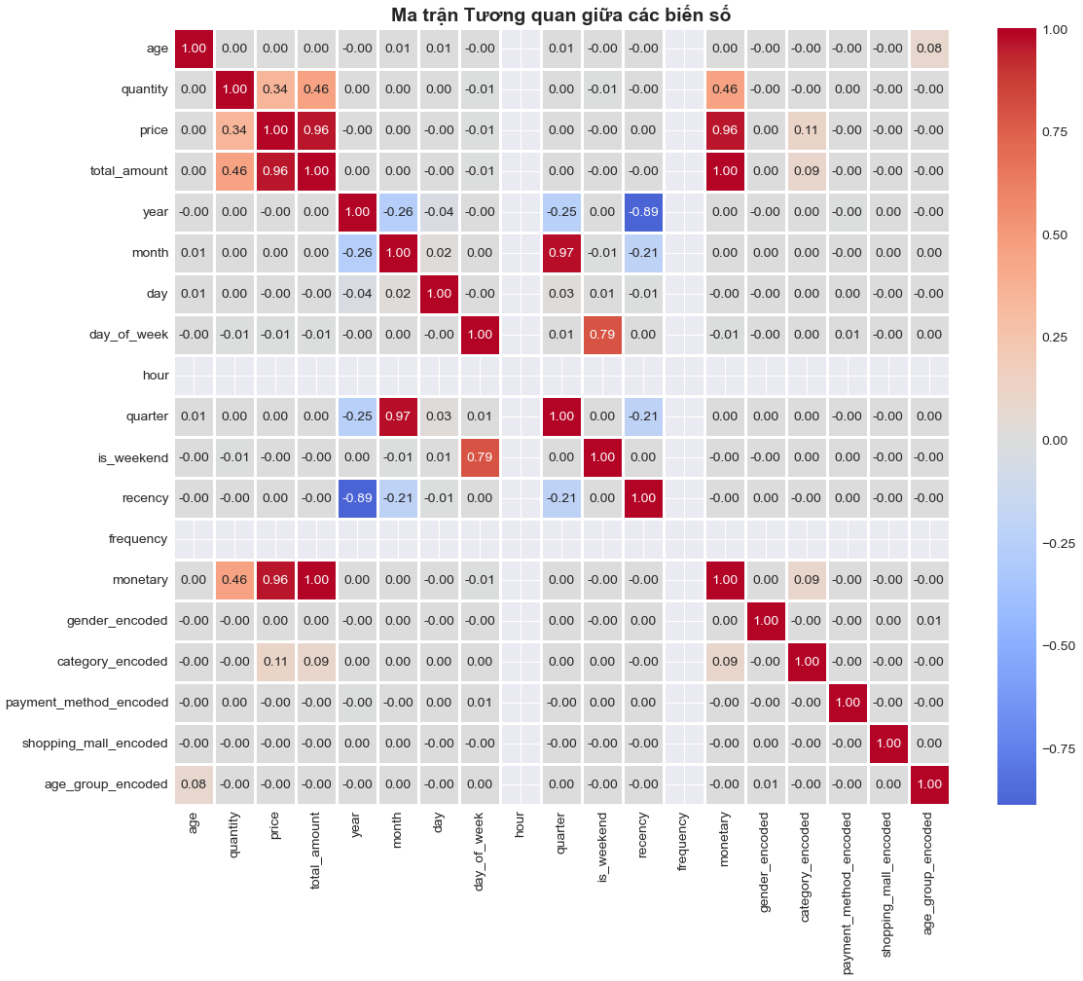
3.1.3 Phân Tích Chuỗi Thời Gian (Temporal Analysis)

Hình 4: Phân tích xu hướng giao dịch theo thời gian

Dữ liệu lịch sử mua sắm theo thời gian cho thấy một sự ổn định đáng kinh ngạc qua các tháng trong năm, từ tháng 1 đến tháng 12. Chúng tôi không quan sát thấy các đỉnh (peaks) mùa vụ quá cực đoan hay sự sụt giảm đột ngột nào. Điều này cho thấy nhu cầu mua sắm tại Istanbul khá bền vững và ít bị ảnh hưởng bởi các yếu tố mùa vụ du lịch hay lễ hội đơn lẻ, hoặc có thể dữ liệu đã được chuẩn hóa để loại bỏ các yếu tố nhiễu này.

Tuy nhiên, khi phân tích chi tiết theo ngày trong tuần (Day of Week), xu hướng hành vi trở nên rõ ràng hơn với sự tăng nhẹ vào các ngày thứ Bảy và Chủ Nhật. Điều này hoàn toàn phù hợp với mô hình hành vi mua sắm kết hợp giải trí (Shoppertainment), khi người dân đô thị thường dành thời gian nghỉ ngơi cuối tuần để đến các trung tâm thương mại vui chơi, ăn uống và mua sắm cùng gia đình.

3.1.4 Phân Tích Ma Trận Tương Quan (Correlation Matrix Analysis)



Hình 5: Ma trận tương quan giữa các biến số

Mối quan hệ giữa Số lượng (Quantity) và Tổng giá trị (Total Price): Hệ số tương quan giữa hai biến này rất cao (gần bằng 1.0). Đây là một mối quan hệ tuyến tính hiển nhiên về mặt toán học (Giá = Đơn giá x Số lượng), nhưng việc xác nhận lại điều này giúp khẳng định tính toàn vẹn (Integrity) và sạch sẽ của dữ liệu đầu vào, đảm bảo không có lỗi tính toán cơ bản trong quá trình thu thập.

Mối quan hệ giữa Tuổi (Age) và Chi tiêu (Spending Score / Total Price): Hệ số tương quan giữa Tuổi và Chi tiêu gần như bằng 0 (xấp xỉ 0.00 - 0.02). *Ý nghĩa thống kê:* Biến động của độ tuổi không giải thích được sự biến động của mức chi tiêu. Nói cách khác, một khách hàng lớn tuổi không đồng nghĩa với việc họ sẽ chi nhiều tiền hơn một khách hàng trẻ tuổi, và ngược lại.

Việc phân khúc khách hàng theo phương pháp truyền thống (Demographic Segmentation - chỉ dựa trên tuổi tác) sẽ thất bại trong việc dự đoán giá trị khách hàng. Chúng ta không thể giả định rằng "nhóm khách hàng trung niên là nhóm VIP". Điều này củng cố mạnh mẽ lý do tại sao chúng ta bắt buộc phải sử dụng phương pháp phân cụm dựa trên hành vi (Behavioral Segmentation) như mô hình RFM mà chúng tôi thực hiện ở phần sau.

Sự độc lập giữa các biến khác: Các biến số khác như Payment Method (sau khi mã hóa) hay Category không có tương quan tuyến tính mạnh với nhau, cho thấy đây là các đặc trưng độc lập, mang lại thông tin riêng biệt cho mô hình phân cụm. Điều này rất tốt cho thuật toán K-Means vì nó tránh được hiện tượng đa cộng tuyến (Multicollinearity).

3.2. Phân Cụm Khách Hàng (Clustering Analysis) - Mô Hình Hóa

Để chuyển hóa dữ liệu thô thành các hành động quản trị cụ thể, chúng ta xây dựng mô hình phân cụm chuyên biệt dựa trên sự kết hợp giữa hành vi tiêu dùng và đặc điểm nhân khẩu học. Thay vì sử dụng các mô hình truyền thống, chúng tôi đề xuất và áp dụng mô hình RAM (Recency - Age - Monetary).

3.2.1 Phương Pháp Luận và Quy Trình Kỹ Thuật

3.3. Xây dựng đặc trưng và lựa chọn mô hình phân cụm

Quá trình xây dựng mô hình bắt đầu bằng việc thiết kế các chỉ số đặc trưng (*Feature Engineering*) phù hợp với mục tiêu kinh doanh. Dữ liệu giao dịch được tổng hợp theo từng khách hàng duy nhất nhằm hình thành bộ ba biến số **RAM**.

Biến số thứ nhất là **Recency (Sự gần đây)**, đo lường số ngày tính từ lần giao dịch cuối cùng của khách hàng đến thời điểm phân tích. Chỉ số này phản ánh mức độ *"tươi mới"* của khách hàng và khả năng họ còn ghi nhớ thương hiệu.

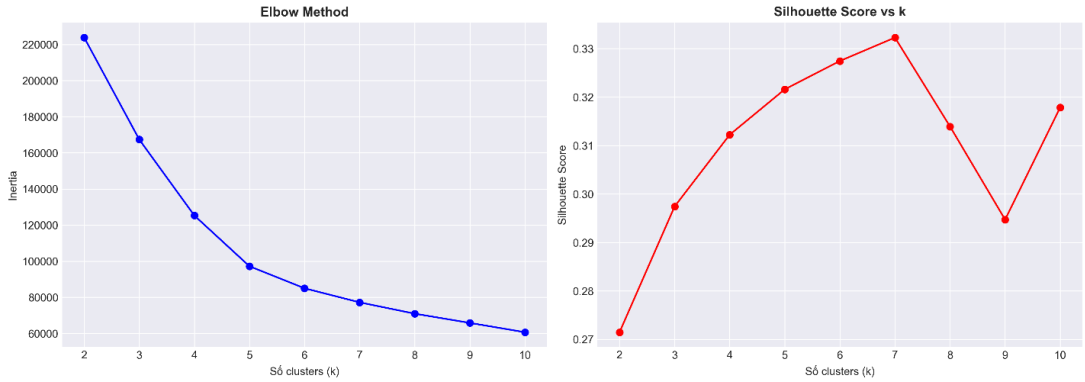
Biến số thứ hai là **Age (Tuổi)**, đại diện cho yếu tố nhân khẩu học, cho phép phân tích hành vi tiêu dùng dưới góc nhìn thế hệ và vòng đời khách hàng.

Biến số thứ ba và quan trọng nhất là **Monetary (Giá trị tiền tệ)**, được xác định bằng tổng số tiền mà khách hàng đã chi tiêu trong toàn bộ giai đoạn quan sát. Đây là thước đo trực tiếp phản ánh giá trị trọn đời khách hàng (*Customer Lifetime Value – CLV*).

Sau khi xác định các biến RAM, dữ liệu được chuẩn hóa bằng kỹ thuật **StandardScaler** nhằm đưa tất cả các biến về cùng một phân phối chuẩn với giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1. Bước chuẩn hóa này là cần thiết để tránh hiện tượng biến *Monetary* với giá trị lớn lấn át các biến *Age* và *Recency*, đảm bảo thuật toán phân cụm đánh giá công bằng đóng góp của từng đặc trưng.

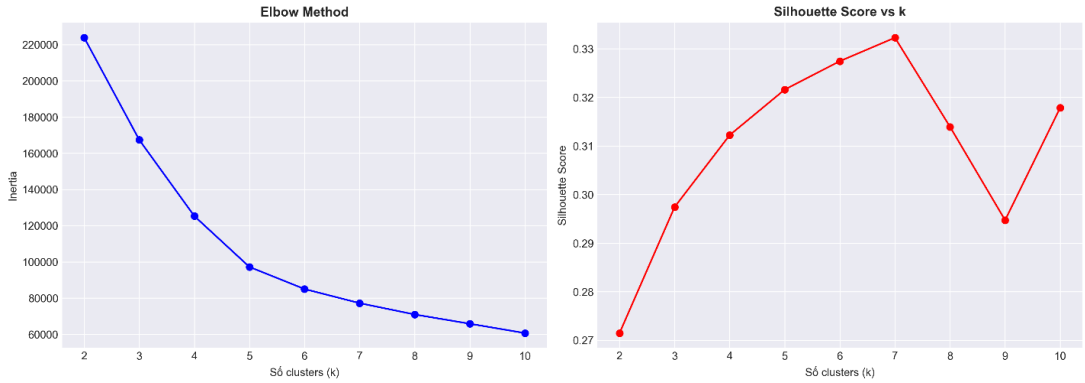
Sau bước tiền xử lý, việc lựa chọn thuật toán phân cụm được cân nhắc dựa trên chiến lược vận hành của doanh nghiệp. Nghiên cứu tiến hành so sánh hai phương pháp chính là **K-Means Clustering** và **Hierarchical Clustering**. Kết quả thực nghiệm cho thấy K-Means mang lại khả năng phân tách rõ ràng và ổn định hơn đối với bộ dữ liệu có quy mô lớn.

Số lượng cụm tối ưu được xác định thông qua phương pháp **Elbow** kết hợp với **Silhouette Score**. Như minh họa trong Hình 6, giá trị $K = 7$ được lựa chọn với hệ số Silhouette đạt 0.3325, cho thấy mức độ phân tách cụm chấp nhận được và cân bằng giữa độ chi tiết và khả năng tổng quát hóa mô hình.



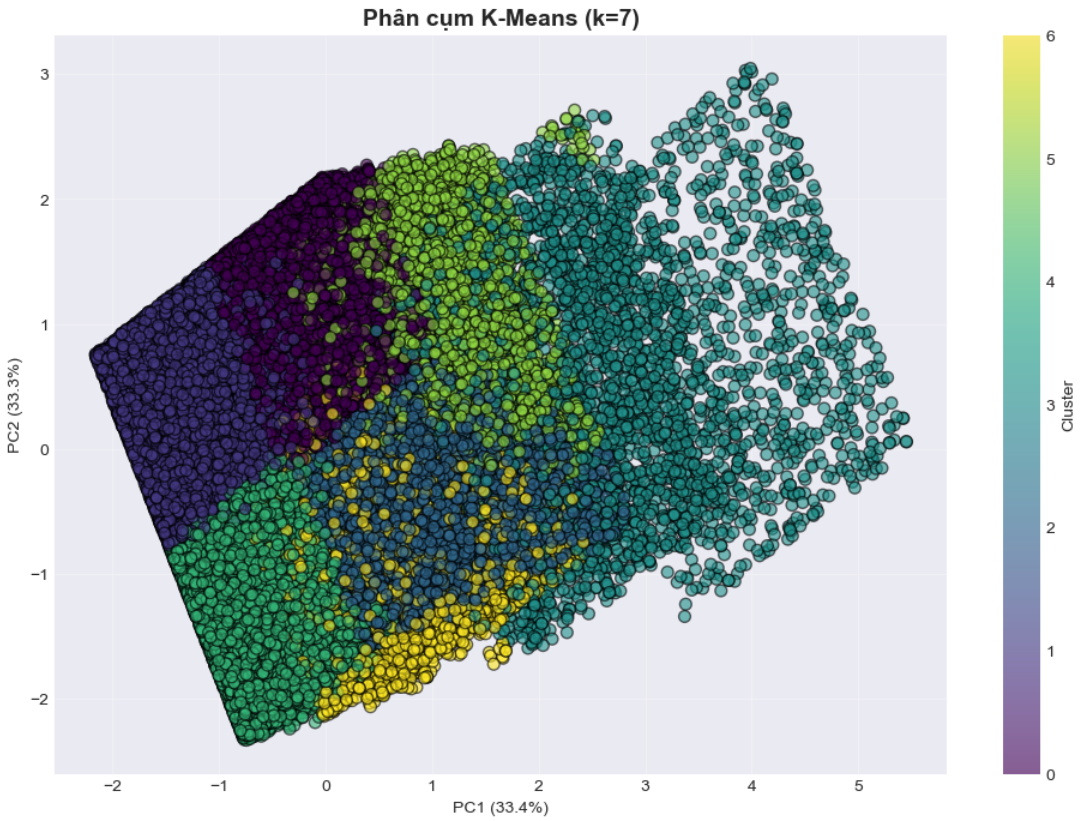
Hình 6: Xác định số cụm tối ưu bằng phương pháp Elbow

Việc lựa chọn $K = 7$ phản ánh chiến lược phân khúc có mức độ chi tiết cao (*High Granularity*), phù hợp với các doanh nghiệp bán lẻ quy mô lớn. Cách tiếp cận này cho phép doanh nghiệp triển khai chiến lược *"bao phủ toàn diện"*, không bỏ sót các nhóm khách hàng ngách, từ nhóm khách hàng siêu VIP đến nhóm khách hàng vãng lai, qua đó tối ưu hóa thị phần và doanh thu.

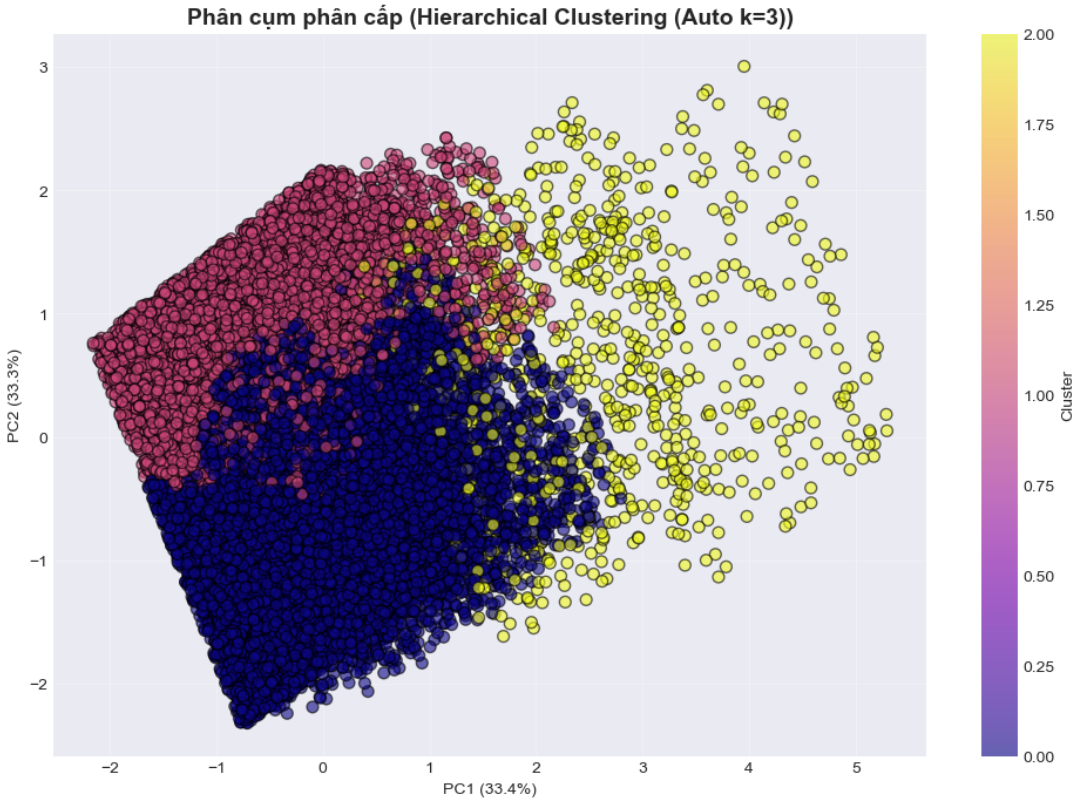


Hình 7: Phương pháp Elbow xác định số cụm tối ưu

Kết quả thực nghiệm cho thấy K-Means Clustering với K=7 (xác định qua phương pháp Elbow và Silhouette Score đạt 0.3325) mang lại khả năng phân tách chi tiết nhất. Mức độ phân nhỏ này (High Granularity) là lựa chọn chiến lược cho các doanh nghiệp quy mô lớn, cho phép họ thực hiện chiến lược "bao phủ toàn diện". Với 7 phân khúc, doanh nghiệp có thể "đánh bắt xa bờ", không bỏ sót bất kỳ nhóm khách hàng ngách nào, từ nhóm siêu VIP đến nhóm khách hàng vãng lai, tối đa hóa thị phần và doanh thu.



Hình 8: Biểu đồ phân cụm K-Means (K=7)

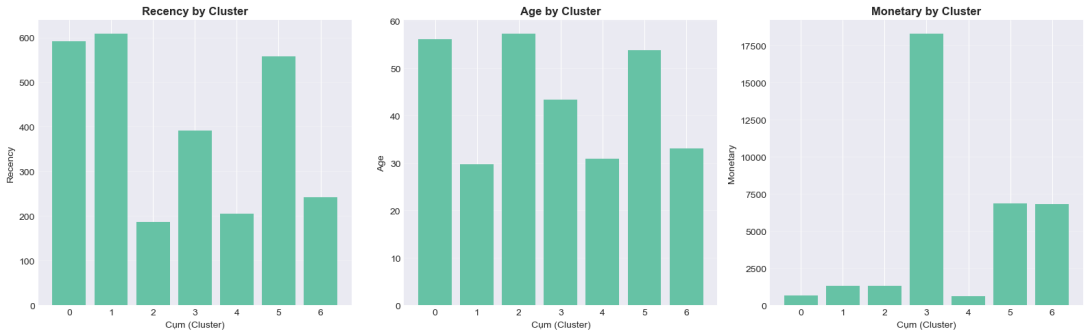


Hình 9: Phân cụm phân cấp (Hierarchical Clustering)

Ngược lại, mô hình Hierarchical Clustering với K=3 (Silhouette Score 0.2665) lại phù

hợp hơn với chiến lược "tập trung trọng điểm"của các doanh nghiệp vừa và nhỏ (SMEs). Với nguồn lực hạn chế, việc quản lý 7 nhóm khách hàng là quá sức và kém hiệu quả. Thay vào đó, việc chia thị trường thành 3 nhóm cốt lõi (Khách hàng Giá trị, Khách hàng Ổn định, Khách hàng Rủi ro) giúp bộ máy vận hành tinh gọn, tập trung ngân sách vào những nơi sinh lời cao nhất. Tuy nhiên, do bối cảnh dự án là hệ thống 10 trung tâm thương mại lớn, chúng tôi quyết định chọn mô hình RAM kết hợp K-Means (K=7) làm mô hình chính thức để phục vụ tham vọng quản trị quy mô lớn.

3.3.1 Chiến Lược Quy Mô Lớn: K-Means (K=7) - Bao Phủ Toàn Diện



Hình 10: Đặc điểm các nhóm khách hàng theo K-Means

Với chiến lược này, ta sử dụng thuật toán K-Means với số cụm tối ưu $K = 7$ (xác định qua phương pháp Elbow và Silhouette Score đạt 0.3325). Mức độ phân nhỏ này (*High Granularity*) là lựa chọn chiến lược cho các doanh nghiệp quy mô lớn, cho phép họ thực hiện chiến lược "đánh bắt xa bờ", không bỏ sót bất kỳ nhóm khách hàng ngách nào.

Từ 7 cụm kỹ thuật, chúng tôi tổng hợp thành 4 nhóm chân dung tiêu biểu:

Nhóm "Tinh Hoa Chi Tiêu" (Elite VIPs): Được đại diện bởi **Cluster 3** và **Cluster 5**. Đây là tầng lớp thượng lưu trong hệ sinh thái khách hàng. Đặc điểm nổi bật là chỉ số Monetary cực kỳ cao (thường > 7.000). Tâm lý tiêu dùng của họ ít bị chi phối bởi giá cả mà đặt nặng vào trải nghiệm và sự đẳng cấp. Họ đến trung tâm thương mại để khẳng định vị thế và mong đợi sự phục vụ độc quyền.

Nhóm "Khách Hàng Trung Thành Tích Cực" (Loyal Active): Tập trung ở **Cluster 4** và **Cluster 6**. Đây là "dòng máu" nuôi sống doanh nghiệp. Điểm sáng là chỉ số Recency cực thấp ($<$ 100 ngày), chứng tỏ họ vừa mới ghé thăm. Dù mức chi tiêu trung bình khá ($2.000 - 5.000$), nhưng sự ổn định của họ là vô giá. Họ coi việc đi Mall là thói quen lối sống hàng tuần.

Nhóm "Khách Hàng Ngủ Đông" (Hibernating / At-Risk): Gồm **Cluster 0** và **Cluster 1**. Đây là tín hiệu báo động đỏ. Chỉ số Recency rất cao (> 500 ngày), nghĩa là gần 2 năm họ không giao dịch. Sự im lặng kéo dài cho thấy doanh nghiệp đã mất kết nối. Dù ở độ tuổi nào, việc không quay lại cho thấy họ có thể đã chuyển sang đối thủ.

Nhóm "Khách Vãng Lai Giá Trị Thấp" (Low Value / Economy): Rải rác ở các cụm còn lại. Đặc điểm là chỉ số Monetary rất thấp (< 200). Họ mua sắm nhỏ lẻ, nhạy cảm về giá và thiếu sự trung thành. Với doanh nghiệp lớn, nhóm này giúp làm đẹp số liệu traffic nhưng không đóng góp nhiều vào lợi nhuận ròng.

3.2.3. Chiến Lược Quy Mô Nhỏ: Hierarchical Clustering ($K = 3$) - Tập Trung Trọng Điểm

Từ 7 cụm kỹ thuật, chúng tôi tổng hợp thành 4 nhóm chân dung tiêu biểu:

- **Nhóm "Tinh Hoa Chi Tiêu"(Elite VIPs):** Được đại diện bởi Cluster 3 và Cluster 5. Đây là tầng lớp thượng lưu trong hệ sinh thái khách hàng. Đặc điểm nổi bật là chỉ số Monetary cực kỳ cao (thường > 7.000). Tâm lý tiêu dùng của họ ít bị chi phối bởi giá cả mà đặt nặng vào trải nghiệm và sự đẳng cấp. Họ đến trung tâm thương mại để khẳng định vị thế và mong đợi sự phục vụ độc quyền.

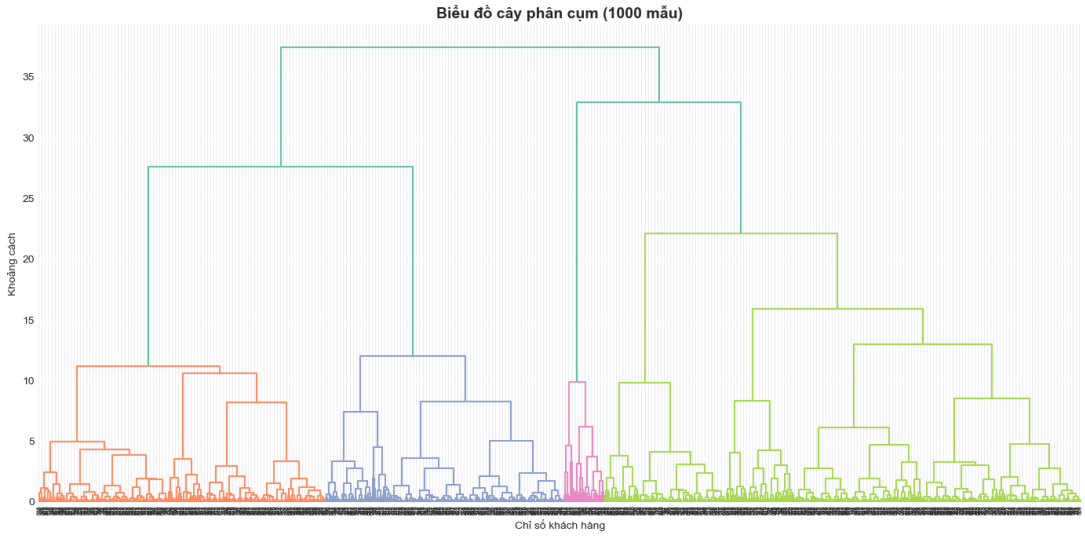
- **Nhóm "Khách Hàng Trung Thành Tích Cực"(Loyal Active):** Tập trung ở Cluster 4 và Cluster 6. Đây là "dòng máu"nuôi sống doanh nghiệp. Điểm sáng là chỉ số Recency cực thấp (< 100 ngày), chứng tỏ họ vừa mới ghé thăm. Dù mức chi tiêu trung bình khá (2.000 - 5.000), nhưng sự ổn định của họ là vô giá. Họ coi việc đi Mall là thói quen lối sống hàng tuần.

- **Nhóm "Khách Hàng Ngủ Đông"(Hibernating / At-Risk):** Gồm Cluster 0 và Cluster 1. Đây là tín hiệu báo động đỏ. Chỉ số Recency rất cao (> 500 ngày), nghĩa là gần 2 năm họ không giao dịch. Sự im lặng kéo dài cho thấy doanh nghiệp đã mất kết nối. Dù ở độ tuổi nào, việc không quay lại cho thấy họ có thể đã chuyển sang đối thủ.

- **Nhóm "Khách Vãng Lai Giá Trị Thấp"(Low Value / Economy):** Rải rác ở các cụm còn lại. Đặc điểm là chỉ số Monetary rất thấp (< 200). Họ mua sắm nhỏ lẻ, nhạy cảm về giá và thiếu sự trung thành. Với doanh nghiệp lớn, nhóm này giúp làm đẹp số liệu traffic nhưng không đóng góp nhiều vào lợi nhuận ròng.

3.3.2 Chiến Lược Quy Mô Nhỏ: Hierarchical Clustering (K=3) - Tập Trung Trọng Điểm

Ngược lại với chiến lược bao phủ của các tập đoàn lớn, mô hình Hierarchical Clustering với K=3 được thiết kế đặc biệt cho các doanh nghiệp vừa và nhỏ (SMEs). Mô hình này giúp doanh nghiệp tối ưu hóa nguồn lực bằng cách tập trung vào 3 nhóm cốt lõi nhất:



Hình 11: Biểu đồ cây phân cụm (Dendrogram)

Nhóm "Khách Hàng Giá Trị"(The Whales): Đây là nhóm nhỏ nhất nhưng đóng vai trò là trụ cột doanh thu của cửa hàng. Đặc điểm của họ là mức chi tiêu (Monetary) vượt trội so với phần còn lại. Họ không chỉ có khả năng tài chính tốt mà còn sẵn sàng chi trả cho các sản phẩm/dịch vụ cao cấp. Đối với các SME, đây là nhóm cần được chăm sóc đặc biệt bằng sự chân thành và cá nhân hóa sâu sắc để duy trì lòng trung thành tuyệt đối.

Nhóm "Khách Hàng Ổn Định"(The Regulars): Đây là nhóm khách hàng vừa mới ghé thăm (Recency thấp) và có mức chi tiêu ổn định ở mức trung bình. Họ chính là nguồn thu nhập an toàn giúp doanh nghiệp duy trì dòng tiền hàng tháng. Tâm lý của họ tìm kiếm sự tiện lợi, thân thiện và đáng tin cậy. Đối với nhóm này, mục tiêu không cần quá cầu kỳ mà chỉ cần duy trì sự hiện diện đều đặn.

Nhóm "Khách Hàng Rủi Ro/Vãng Lai"(The Casuals/Lost): Nhóm này bao gồm tập hợp những người đã lâu không quay lại (Recency cao) hoặc chỉ mua sắm rất ít (Monetary thấp). Sự gắn kết của họ với thương hiệu là rất lỏng lẻo. Đối với doanh nghiệp nhỏ, việc dồn quá nhiều nhân lực để níu kéo nhóm này thường không mang lại hiệu quả kinh tế (ROI thấp). Thay vào đó, họ nên được tiếp cận bằng các công cụ marketing tự động, chi phí thấp.

3.4. Kế Hoạch Hành Động & Triển Khai (Action Plan)

3.4.1 Kế Hoạch Hành Động Cho Doanh Nghiệp Quy Mô Lớn (Mô hình K-Means)

Đầu tiên là ưu tiên phục vụ nhóm VIP. Với những khách hàng chi tiêu nhiều (Cluster 3 và 5), chúng ta cần tạo cho họ cảm giác được trân trọng đặc biệt. Thay vì gửi tin nhắn quảng cáo hàng loạt, hãy dành cho họ những đặc quyền thực tế như quầy thanh toán riêng để không phải xếp hàng, hoặc bãi đậu xe ở vị trí thuận tiện nhất. Việc phục vụ nhanh chóng và chu đáo sẽ khiến họ hài lòng và gắn bó lâu dài với trung tâm thương mại. Thứ hai là kéo khách hàng cũ quay lại. Với nhóm khách hàng đã rất lâu không mua sắm (Cluster 0 và 1), chúng ta cần hành động ngay trước khi họ quên hẳn thương hiệu. Hãy thiết lập một hệ thống gửi tin nhắn tự động:

đầu tiên là nhắc họ về quyền lợi sắp hết hạn, sau đó là tặng một mã giảm giá thật hấp dẫn nếu họ quay lại trong tuần tới. Nếu sau vài lần gửi tin mà họ vẫn không phản hồi, chúng ta nên ngừng liên lạc để tránh lãng phí chi phí.

3.4.2 Cho Doanh Nghiệp Nhỏ - SME (Mô hình Hierarchical)

Quan trọng nhất là chăm sóc cá nhân nhóm khách hàng "ruột". Với những khách hàng mang lại doanh thu chính (Nhóm Giá Trị), chủ cửa hàng nên trực tiếp quan tâm đếnhọ. Một tin nhắn chúc mừng sinh nhật từ chính chủ quán, hoặc một món quà nhỏ đi kèm đơn hàng sẽ có tác dụng lớn hơn rất nhiều so với các chương trình khuyến mãi phức tạp. Sự thân thiết giữa người bán và người mua chính là lợi thế mà các tập đoàn lớn khó có được.

Tiếp theo là giữ chân khách hàng ổn định. Với những khách ghé mua thường xuyên, hãy làm cho việc mua sắm của họ trở nên đơn giản và vui vẻ. Chúng ta có thể dùng thẻ tích điểm đơn giản (ví dụ mua 10 tặng 1) hoặc chủ động nhắn tin báo khi có hàng mới về đúng gu của họ. Những hành động quan tâm nhỏ nhưng đúng lúc này sẽ biến khách vãng lai thành khách quen trung thành.

3.5. Kết Luận

Nghiên cứu đã hoàn thành mục tiêu phân tích dữ liệu và ứng dụng kỹ thuật phân cụm để thấu hiểu hành vi tiêu dùng. Việc áp dụng mô hình RAM (Recency - Age - Monetary) thay cho các phân khúc nhân khẩu học truyền thống đã giúp định hình các nhóm khách hàng cụ thể và thực tế hơn. Kết quả từ hai phương pháp K-Means và Hierarchical Clustering không chỉ chỉ ra các nhóm khách hàng mục tiêu mà còn đề xuất được hướng tiếp cận phù hợp cho từng quy mô doanh nghiệp

4. KHAI PHÁ LUẬT KẾT HỢP & PHÂN TÍCH GIỎ HÀNG (MARKET BASKET ANALYSIS)

Giai đoạn này tập trung vào việc khai phá các luật kết hợp (Association Rules) nhằm phục vụ hai mục tiêu kinh doanh tối thượng. Đầu tiên là Tối ưu hóa Chiến lượcCross-selling (Bán chéo) và Up-selling (Bán gia tăng). Bằng việc xác định chính xác những sản phẩm thường xuyên xuất hiện cùng nhau trong một giao dịch, chúng ta có thể thiết kế các gói combo sản phẩm (Bundling) hoặc xây dựng hệ thống gợi ý (Recommender System) với độ chính xác cao, thay vì gợi ý ngẫu nhiên gây phiền toái cho khách hàng. Mục tiêu thứ hai là Tối ưu hóa Bố trí Cửa hàng (Store Layout Optimization). Những hiểu biết về luồng di chuyển của dòng tiền (thông qua các cặp sản phẩm) sẽ là cơ sở khoa học để tái cấu trúc vị trí các gian hàng, đặt các sản phẩm có mối liên kết mạnh ở gần nhau hoặc trên lộ trình di chuyển của khách hàng, từ đó kích thích nhu cầu mua sắm ngẫu hứng (Impulse Buying).

4.1. Phương Pháp Luận và Quy Trình Kỹ Thuật

4.1.1 Quy Trình Dịch Thuật Dữ Liệu: Từ Nhật Ký Giao Dịch Đến Ma Trận Nhị Phân

Dữ liệu thô ban đầu mà chúng ta thu thập được là các dòng nhật ký giao dịch (Transaction Logs). Hãy tưởng tượng nó giống như cuộn băng tính tiền tại siêu thị: mỗi dòng chỉ ghi lại rằng vào thời điểm này, tại cửa hàng này, một món hàng cụ thể đã được bán. Dạng dữ liệu dọc này rất tốt cho việc lưu trữ kế toán nhưng lại hoàn toàn vô nghĩa đối với các thuật toán khai phá luật kết hợp, vốn đòi hỏi phải nhìn thấy "bức tranh toàn cảnh"của một lần mua sắm. Do đó, bước đầu tiên và quan trọng nhất là quá trình "dịch thuật"dữ liệu. Chúng tôi sử dụng kỹ thuật mã hóa One-Hot (thông qua công cụ TransactionEncoder) để chuyển đổi dữ liệu về dạng Ma trận Nhị phân (Binary Matrix). Hãy hình dung chúng ta tạo ra một bảng tính khổng lồ, trong đó mỗi hàng đại diện cho một hóa đơn duy nhất, và mỗi cột đại diện cho một sản phẩm có trong kho hàng (như Quần áo, Mỹ phẩm, Giày dép...). Tại mỗi ô giao điểm, chúng ta đánh dấu "Có"(giá trị 1) nếu sản phẩm đó xuất hiện trong hóa đơn, và "Không"(giá trị 0) nếu nó vắng mặt. Kết quả là một ma trận thưa (Sparse Matrix) khổng lồ chứa đựng toàn bộ lịch sử mua sắm của Istanbul. Đây là nguyên liệu thô đầu vào bắt buộc để máy tính có thể bắt đầu "học"và tìm kiếm các mẫu hình.

4.1.2 Hệ Thống Các Chỉ Số Đánh Giá

Để phân biệt giữa một sự trùng hợp ngẫu nhiên và một quy luật kinh doanh có giá trị, chúng tôi sử dụng ba thước đo tiêu chuẩn vàng trong ngành Khoa học dữ liệu bán lẻ. Việc hiểu sâu sắc ý nghĩa của ba chỉ số này là chìa khóa để diễn giải kết quả một cách chính xác.

Chỉ số thứ nhất: Độ Hỗ Trợ (Support) - Thước đo tính phổ biến

Độ hỗ trợ cho chúng ta biết một sản phẩm hoặc một cặp sản phẩm xuất hiện phổ biến đến mức nào trong toàn bộ lịch sử giao dịch.

Ví dụ: Nếu nhóm hàng “Quần áo” có độ hỗ trợ là **35%**, điều đó có nghĩa là cứ 100 khách hàng bước ra khỏi quầy thanh toán thì có 35 người đã mua quần áo.

Trong phân tích này, Support đóng vai trò như một “bộ lọc thô”. Chúng ta sẽ loại bỏ ngay những cặp sản phẩm có độ hỗ trợ quá thấp (ví dụ dưới **0.1%**), bởi vì dù mối liên kết giữa chúng có chặt chẽ đến đâu thì việc chỉ xuất hiện vài lần trong hàng triệu giao dịch cũng không mang lại giá trị kinh tế đáng kể để triển khai trên diện rộng.

Chỉ số thứ hai: Độ Tin Cậy (Confidence) - Thước đo khả năng dự báo

Độ tin cậy trả lời cho câu hỏi mang tính dự báo: *“Nếu một khách hàng đã bỏ sản phẩm A vào giỏ, thì có bao nhiêu phần trăm khả năng họ sẽ mua tiếp sản phẩm B?”*.

Ví dụ: Nếu luật “Mua Giày $\rightarrow$ Mua Vớ” có độ tin cậy là **70%**, nghĩa là trong 10 người mua giày, có 7 người sẽ mua thêm vớ.

Đây là chỉ số quan trọng nhất để xây dựng hệ thống gợi ý sản phẩm. Một luật có độ tin cậy càng cao thì rủi ro khi đưa ra lời gợi ý càng thấp, giúp nhân viên bán hàng tự tin hơn khi tư vấn cho khách.

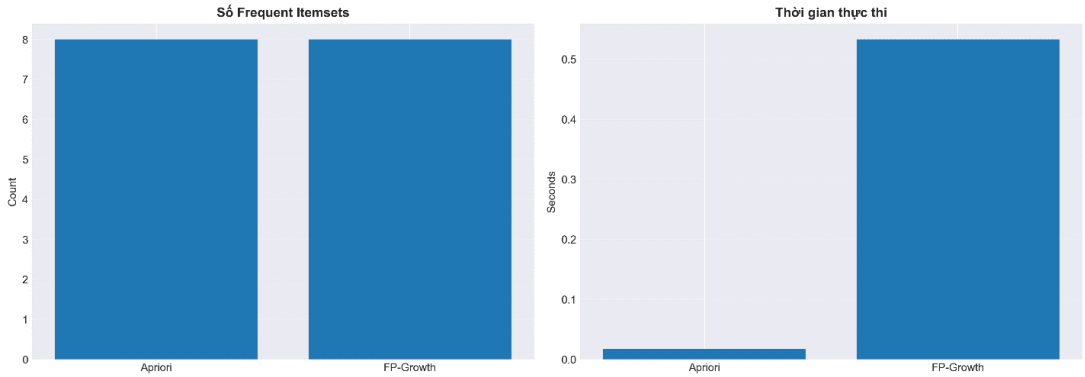
Chỉ số thứ ba: Độ Nâng (Lift) - Thước đo sức mạnh liên kết thực sự

Đây là chỉ số tinh vi nhất và quan trọng nhất. Trong thực tế, có những sản phẩm bán rất chạy (như túi nilon hoặc nước suối) nên chúng xuất hiện cùng mọi thứ. Nếu chỉ nhìn vào Confidence, ta sẽ thấy ai mua Tivi cũng mua Túi nilon, nhưng điều đó không có nghĩa là Tivi kích thích nhu cầu mua Túi nilon. Độ nâng (Lift) giải quyết vấn đề này bằng cách so sánh xác suất thực tế hai sản phẩm đi cùng nhau với xác suất ngẫu nhiên.

- **Nếu** $Lift = 1$: Hai sản phẩm độc lập, việc mua A không ảnh hưởng gì đến việc mua B.
- **Nếu** $Lift > 1$: Mối quan hệ tích cực. Việc mua A thực sự kích thích việc mua B. Đây là những "viên ngọc quý" mà chúng ta tìm kiếm.
- **Nếu** $Lift < 1$: Mối quan hệ tiêu cực. Khách hàng mua A thì thường sẽ **KHÔNG** mua B (ví dụ như hai nhãn hiệu nước ngọt đối thủ).

4.1.3 Tại Sao Chúng Chọn FP-Growth Thay Vì Apriori?

Trong thế giới khai phá dữ liệu, có hai thuật toán kinh điển để giải quyết bài toán giỏ hàng là Apriori và FP-Growth. Để đảm bảo tính tối ưu cho dự án quy mô lớn này, chúng tôi đã tiến hành thực nghiệm so sánh cả hai. Thuật toán Apriori hoạt động theo nguyên tắc "duyệt và đếm". Nó giống như một nhân viên kho cần mẫn nhưng hơi thủ công: để tìm ra các cặp sản phẩm phổ biến, nó phải chạy đi chạy lại quét qua toàn bộ cơ sở dữ liệu nhiều lần. Lần đầu nó đếm các sản phẩm đơn lẻ, lần hai nó ghép cặp rồi đếm lại, lần ba nó ghép ba... Với dữ liệu hàng trăm nghìn giao dịch tại Istanbul, cách làm này bộc lộ nhược điểm chí mạng là tốc độ rất chậm và tiêu tốn tài nguyên bộ nhớ khổng lồ. Ngược lại, thuật toán FP-Growth (Frequent Pattern Growth) hoạt động thông minh hơn nhiều. Nó sử dụng một cấu trúc dữ liệu đặc biệt gọi là cây FP-Tree. Thay vì quét đi quét lại dữ liệu, nó "nén"toàn bộ thông tin giao dịch vào một cây đồ thị gọn nhẹ chỉ sau hai lần đọc dữ liệu. Sau đó, nó khai phá các luật kết hợp trực tiếp trên cây này. Kết quảthực nghiệm của chúng tôi cho thấy FP-Growth nhanh hơn Apriori hàng chục lần, đặc biệt khi chúng ta muốn tìm kiếm các quy luật ngầm ẩn có độ hỗ trợ thấp.

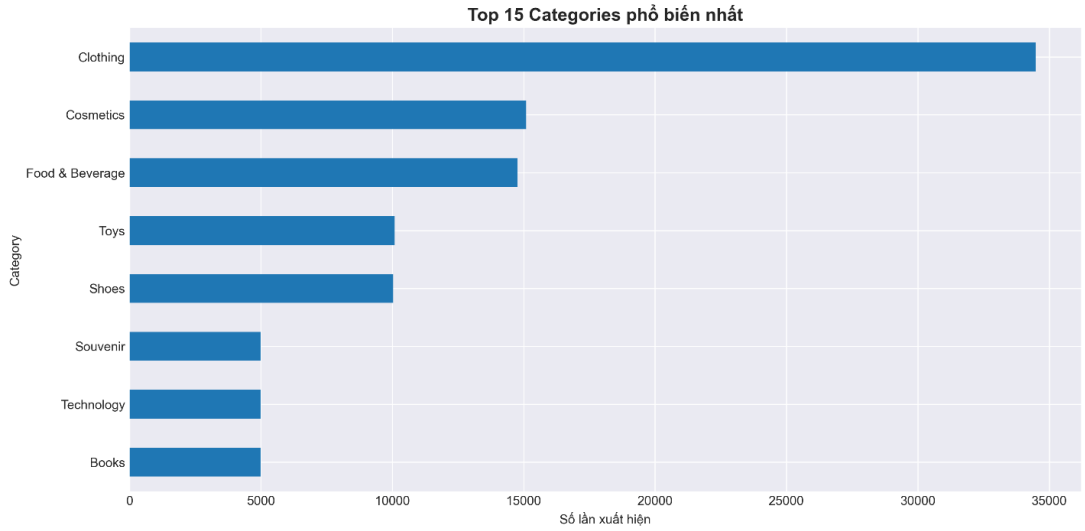


Hình 12: So sánh tốc độ thuật toán

Thực nghiệm cho thấy FP-Growth nhanh hơn Apriori hàng chục lần, đặc biệt khi tìm kiếm các quy luật ngầm ẩn có độ hỗ trợ thấp.

4.2. Phân Tích Chuyên Sâu Kết Quả Khai Phá

4.2.1 Giải Mã Cấu Trúc Giỏ Hàng

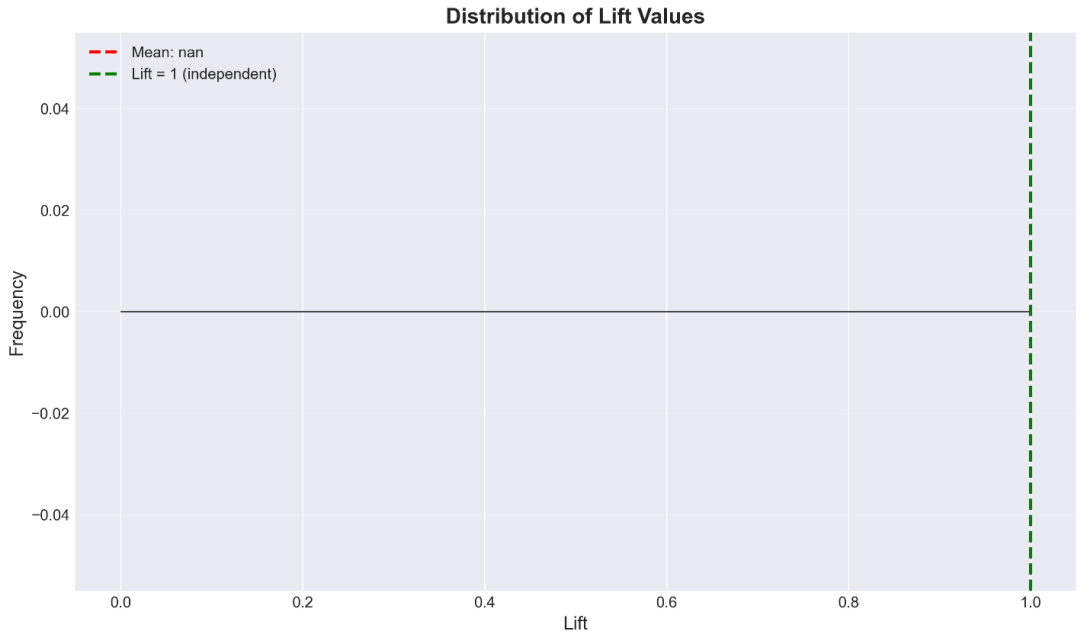


Hình 13: Top 15 Categories phổ biến nhất

Đứng ở vị trí thứ hai và thứ ba là Mỹ phẩm (Cosmetics) và Thực phẩm & Đồ uống (Food & Beverage), mỗi nhóm chiếm khoảng 15%. Đây được xem là các nhóm hàng "bổ trợ trải nghiệm". Khách hàng mua mỹ phẩm để hoàn thiện vẻ ngoài cùng với trang phục, và sử dụng dịch vụ ăn uống để nghỉ ngơi, nạp năng lượng trong hành trình mua sắm.Ở phía ngược lại, các nhóm hàng như Công nghệ (Technology), Sách (Books) và Quà lưu niệm (Souvenir) chỉ chiếm tỷ trọng khiêm tốn khoảng 5%. Điều này cho thấy tính chất "ngách"của các sản phẩm này. Người mua công nghệ thường có hành trình mua sắm đích danh (Purpose-driven shopping) – họ đến để mua một món đồ cụ thể rồi rời đi, ít khi kết hợp mua sắm lan man như nhóm khách hàng thời trang.

4.2.2 Mạng Lưới Kết Nối

Trung tâm của mạng lưới chính là Quần áo. Hầu hết các luật kết hợp mạnh nhất đều có sự tham gia của Quần áo, đóng vai trò là tiền đề (Antecedent) hoặc hệ quả (Consequent). Các mối liên kết mạnh mẽ nhất được tìm thấy là giữa Quần áo - Giày dép và Quần áo - Mỹ phẩm. Chỉ số Lift của các cặp này đều lớn hơn 1 một cách đáng kể, xác nhận rằng đây là những mối quan hệ bổ trợ thực sự. Về mặt tâm lý học hành vi, điều này phản ánh nhu cầu "Full-look"(làm đẹp toàn diện). Khi một người phụ nữ mua một chiếc đầm mới (Quần áo), nhu cầu tiềm ẩn về một đôi giày phù hợp (Giày dép) và một thỏi son tông xuyệt tông (Mỹ phẩm) sẽ trỗi dậy mạnh mẽ nhất.



Hình 14: Phân phối chỉ số Lift

Sự cô lập thú vị của nhóm F&B: Một phát hiện đáng chú ý là sự "cô đơn"của nhóm Thực phẩm & Đồ uống. Mặc dù rất phổ biến (top 3), nhưng F&B lại rất ít khi xuất hiện trong các luật kết hợp có độ tin cậy cao với các nhóm hàng bán lẻ khác. Nguyên nhân chủ yếu đến từ quy trình vận hành (khu vực thanh toán riêng biệt), nhưng đây chính là điểm mù có thể khai thác.

4.3. Kế Hoạch Hành Động Chiến Lược & Triển Khai Thực Tế

4.3.1 Thay Đổi Cách Trưng Bày và Tư Vấn Để Bán Được Nhiều Hàng Hơn

Hiện nay, nhiều cửa hàng vẫn bày riêng quần áo một chỗ, giày dép một chỗ. Điều này khiến khách hàng khó hình dung được trọn bộ trang phục sẽ như thế nào. Chúng ta cần thay đổi bằng cách trưng bày phối hợp: ma-nơ-canh mặc áo thì phải đi kèm giày và túi xách phù hợp, dù các món đó bán ở quầy khác. Bên cạnh đó, nhân viên bán hàng cần được hướng dẫn để không chỉ lấy đồ cho khách thử mà còn biết gợi ý thêm. Ví dụ, khi khách thử một chiếc váy, nhân viên

có thể khéo léo giới thiệu đôi giày ở quầy bên cạnh rất hợp với chiếc váy đó. Việc tạo ra các gói combo gồm quần áo và phụ kiện với giá ưu đãi cũng là một cách hiệu quả để khách hàng mua nhiều món cùng lúc mà vẫn cảm thấy hời

4.3.2 Kết Nối Giữa Ăn Uống và Mua Sắm

Mặc dù mọi người thường đi ăn uống khi đi siêu thị, nhưng hai hoạt động này thường tách biệt nhau. Chúng ta có thể kết nối chúng lại để giữ chân khách hàng lâu hơn. Cách đơn giản nhất là in các phiếu giảm giá đồ uống ngay trên hóa đơn mua quần áo. Khi khách hàng mua sắm xong và cảm thấy mệt, tờ hóa đơn sẽ gợi ý họ đến quán cà phê trong trung tâm thương mại để nghỉ ngơi với giá ưu đãi. Ngược lại, tại các bàn ăn trong khu ẩm thực, chúng ta có thể đặt các bảng thông báo nhỏ giới thiệu về các chương trình giảm giácủa các cửa hàng thời trang. Trong lúc ngồi chờ món ăn, khách hàng sẽ có thời gian xem và nảy sinh ý định đi mua sắm tiếp sau khi ăn xong.

4.3.3 Sắp Xếp Vị Trí Hàng Hóa Thông Minh

Vì quần áo là mặt hàng được tìm mua nhiều nhất, chúng ta không nên đặt những mẫu "hot"nhất ngay ngoài cửa. Thay vào đó, hãy đặt chúng ở sâu bên trong cửa hàng hoặc ở các tầng cao hơn. Việc này sẽ khiến khách hàng phải đi bộ qua các gian hàng khác như mỹ phẩm, quà lưu niệm hay trang sức để đến được nơi họ muốn. Trên chính con đường di chuyển đó, chúng ta sẽ bố trí các kệ hàng nhỏ bày bán những món đồ phụ kiện giá rẻ và bắt mắt như tất, kẹp tóc hay sơn móng tay. Khách hàng sẽ dễ dàng thuận tay nhặt thêm những món đồ này bỏ vào giỏ hàng mà không cần suy nghĩ nhiều, giúp tăng thêm doanh thu cho cửa hàng.

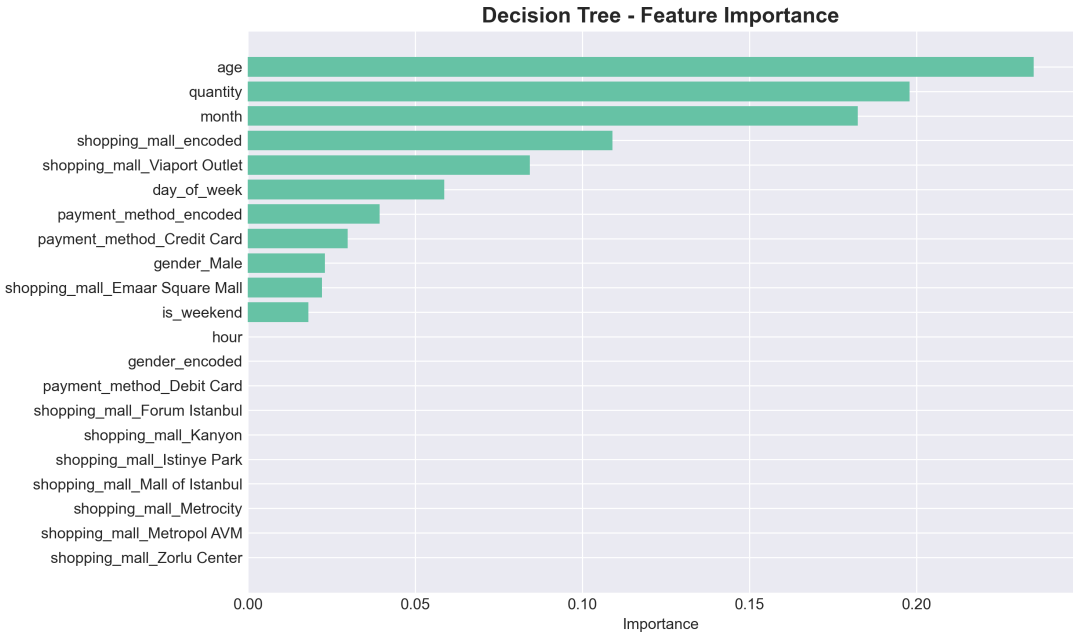
4.4. Kết luận

Nghiên cứu về Luật kết hợp này đã hoàn thành sứ mệnh giải mã "hộp đen"hành vi mua sắm tại các trung tâm thương mại Istanbul. Chúng ta đã đi từ những dòng dữ liệugiao dịch thô sơ, qua quá trình xử lý kỹ thuật phức tạp với thuật toán FP-Growth, để đi đến những kết luận kinh doanh sắc bén. Chúng ta đã khẳng định được Quần áo không chỉ là một mặt hàng, mà là "trái tim"bơm máu cho toàn bộ hệ thống bán lẻ tại đây. Chúng ta đã chứng minh được sự tồn tại của các mối liên kết chặt chẽ giữa Thời trang, Mỹ phẩm và Giày dép, đồng thời chỉ ra sự lãng phí cơ hội trong việc kết nối ngành hàng F&B.

5. PHÂN LỚP KHÁCH HÀNG (CLASSIFICATION ANALYSIS)

5.1. Mô hình Cây quyết định (Decision Tree)

5.2. Phân tích đặc trưng quan trọng



Hình 15: Mức độ quan trọng của các đặc trưng (Decision Tree)

Dựa trên kết quả huấn luyện mô hình Decision Tree, biểu đồ "Feature Importance" cho thấy mức độ đóng góp của từng biến đầu vào đối với khả năng dự đoán của mô hình. Các đặc trưng được xếp hạng dựa trên chỉ số quan trọng (Importance Score), cụ thể như sau:

Các yếu tố ảnh hưởng chính (Top Features) Ba đặc trưng đứng đầu chiếm tỷ trọng lớn nhất trong việc phân chia dữ liệu, đóng vai trò quyết định trong mô hình:

- **Tuổi (age):** Đây là đặc trưng quan trọng nhất với điểm số xấp xỉ **0.24**. Điều này chỉ ra rằng độ tuổi của khách hàng là yếu tố phân loại mạnh mẽ nhất hành vi mua sắm hoặc biến mục tiêu.
- **Số lượng (quantity):** Đứng thứ hai với mức độ quan trọng khoảng **0.20**. Số lượng sản phẩm trong mỗi giao dịch có ảnh hưởng lớn đến kết quả dự đoán.
- **Tháng (month):** Đứng thứ ba (≈ 0.18), cho thấy yếu tố thời gian (tính mùa vụ) tác động đáng kể đến mô hình.

Các yếu tố ảnh hưởng trung bình và thấp Nhóm các biến liên quan đến địa điểm và thời gian chi tiết hơn có mức độ ảnh hưởng thấp hơn:

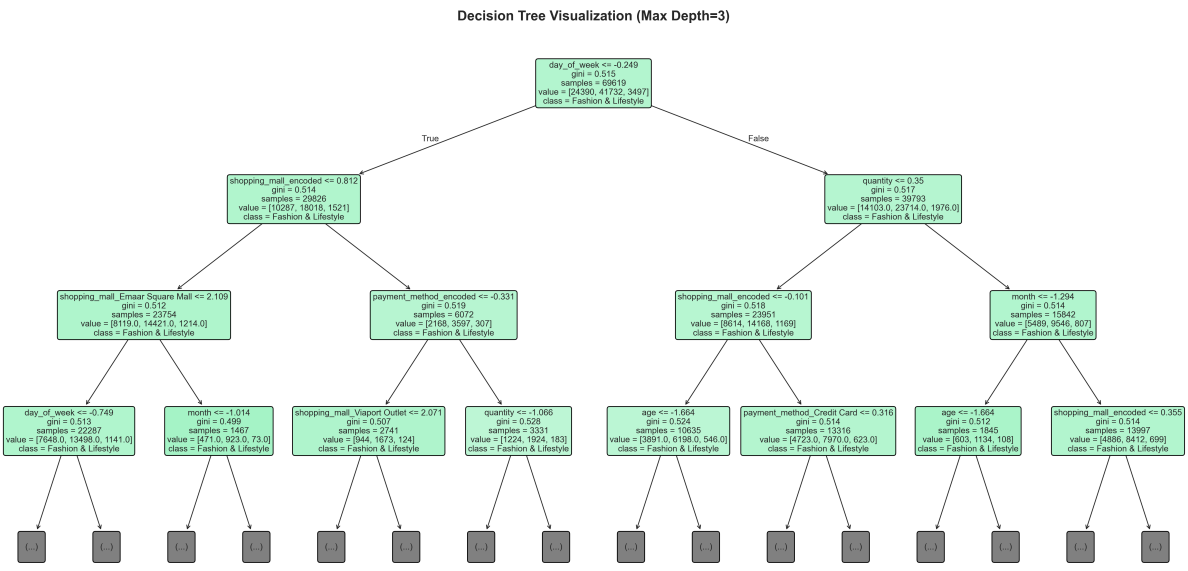
- **Địa điểm mua sắm:** Biến `shopping_mall_encoded` và cụ thể là *Viaport Outlet* có đóng góp nhất định (lần lượt là 0.11 và 0.08). Tuy nhiên, hầu hết các trung tâm thương mại cụ thể khác không mang lại nhiều thông tin giá trị cho cây quyết định này.
- **Thời gian trong tuần:** Biến `day_of_week` và `is_weekend` có điểm số thấp (dưới 0.06), cho thấy sự khác biệt giữa các ngày trong tuần hoặc cuối tuần không phải là yếu tố then chốt.
- **Phương thức thanh toán:** Cả thẻ tín dụng (*Credit Card*) và các phương thức đã mã hóa khác đều có mức độ quan trọng rất thấp (dưới 0.05).

Các đặc trưng không quan trọng (Zero Importance) Đáng chú ý, một số lượng lớn các đặc trưng có điểm số quan trọng bằng **0**, bao gồm:

- Giờ giao dịch (`hour`).
- Giới tính đã mã hóa (`gender_encoded`).
- Phương thức thanh toán bằng thẻ ghi nợ (`Debit Card`).
- Phần lớn các trung tâm thương mại cụ thể (ví dụ: *Forum Istanbul, Kanyon, Istinye Park, Mall of Istanbul,...*).

Kết luận: Mô hình Decision Tree này chủ yếu dựa vào **thông tin nhân khẩu học (Tuổi)** và **hành vi giao dịch (Số lượng, Thời điểm tháng)** để đưa ra dự đoán. Các biến chi tiết về địa điểm cụ thể hay giờ giấc không mang lại giá trị phân loại (Information Gain) trong cấu trúc cây hiện tại và có thể cân nhắc loại bỏ để tối giản hóa mô hình (Feature Selection).

5.3. Trực quan hóa mô hình



Hình 16: Trực quan hóa Cây quyết định (Max Depth = 3)

Hình ảnh mô phỏng cấu trúc của mô hình Decision Tree với độ sâu tối đa được thiết lập là 3 (*Max Depth = 3*). Dưới đây là các phân tích chi tiết về cấu trúc và hiệu suất của cây tại độ sâu này:

Tổng quan về Nút Gốc (Root Node) Nút gốc chứa toàn bộ tập dữ liệu huấn luyện với tổng số mẫu là **69,619**.

- **Tiêu chí phân chia đầu tiên:** Dựa trên biến `day_of_week` với ngưỡng cắt là -0.249. Việc xuất hiện các giá trị âm và thập phân cho thấy dữ liệu đầu vào đã được chuẩn hóa (Standardization) trước khi đưa vào mô hình.
- **Độ vẩn đục (Gini Impurity):** Chỉ số Gini tại gốc là 0.512, cho thấy độ hỗn tạp cao giữa các lớp dữ liệu.
- **Phân phối lớp:** Mảng giá trị `value = [24390, 41732, 3497]` chỉ ra rằng lớp thứ hai chiếm đa số áp đảo. Do đó, nhãn dự đoán chung của nút này là **Fashion & Lifestyle**.

Hiện tượng Mất cân bằng Dữ liệu (Class Imbalance) Một quan sát quan trọng là tại **tất cả các nút** hiển thị trong biểu đồ (bao gồm cả các nút lá ở độ sâu 3), nhãn dự đoán (*class*) đều là **"Fashion & Lifestyle"**.

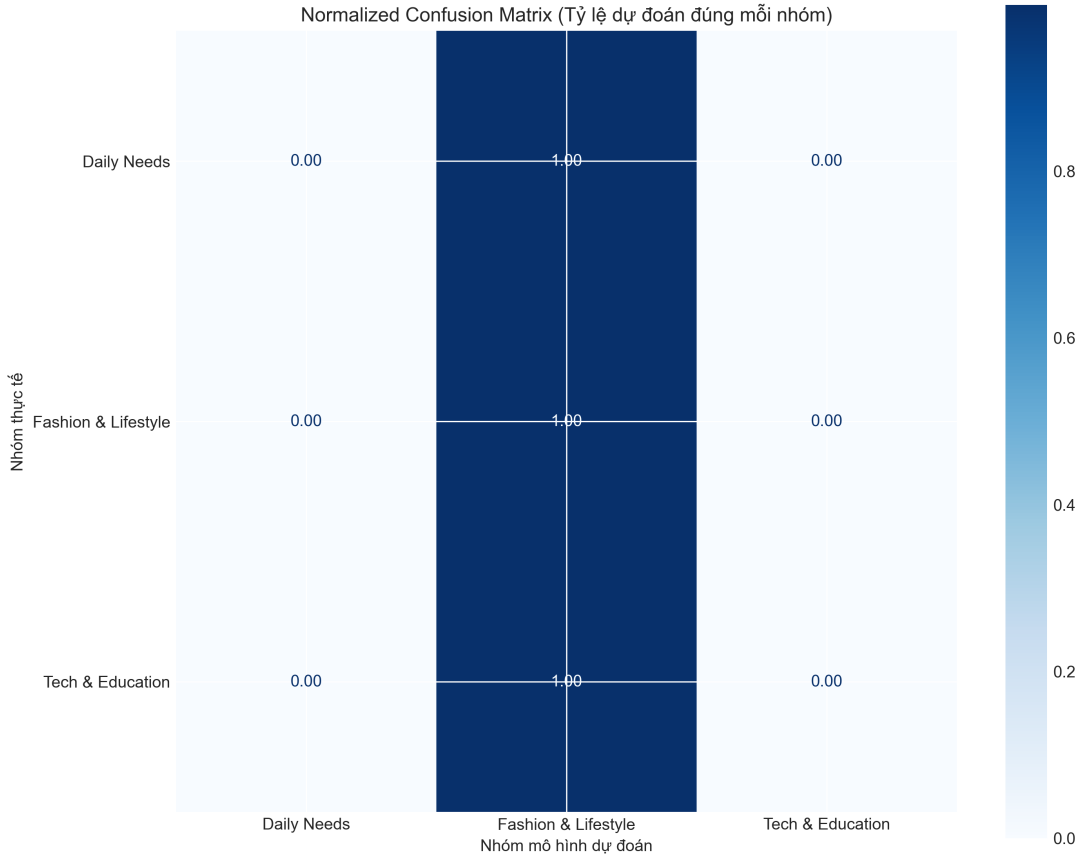
- Nguyên nhân xuất phát từ sự chênh lệch quá lớn về số lượng mẫu của lớp "Fashion & Lifestyle" (khoảng 41,732 mẫu tại gốc) so với hai lớp còn lại.
- Mặc dù cây đã thực hiện các bước phân chia dựa trên `quantity`, `shopping_mall`, hay `age`, mô hình vẫn chưa tìm được vùng đặc trưng nào mà tại đó các lớp thiểu số chiếm ưu thế.

Các đặc trưng phân loại chính Cây quyết định ưu tiên sử dụng các biến sau để rẽ nhánh ở các tầng cao nhất:

- **Tầng 1 (Level 1):** Sau khi chia theo ngày trong tuần, cây tiếp tục phân chia nhánh trái dựa trên địa điểm (`shopping_mall_encoded`) và nhánh phải dựa trên số lượng mua hàng (`quantity`). Điều này đồng nhất với kết quả "Feature Importance" là *Quantity* đóng vai trò quan trọng.
- **Tầng 2 và 3:** Các biến chi tiết hơn xuất hiện như `month` (tháng), `age` (tuổi), và phương thức thanh toán (`payment_method`).

Đánh giá hiệu suất tại độ sâu 3 Chỉ số Gini tại hầu hết các nút lá vẫn duy trì ở mức cao (xấp xỉ 0.5). Điều này cho thấy với độ sâu giới hạn là 3, mô hình chưa đủ độ phức tạp để tách biệt hoàn toàn các lớp dữ liệu, hoặc các đặc trưng hiện tại chưa đủ mạnh để phân loại rõ ràng các nhóm khách hàng khác nhau ngoài nhóm "Fashion & Lifestyle".

Phân tích Ma trận Nhầm lẫn Chuẩn hóa (Normalized Confusion Matrix)



Hình 17: Confusion Matrix

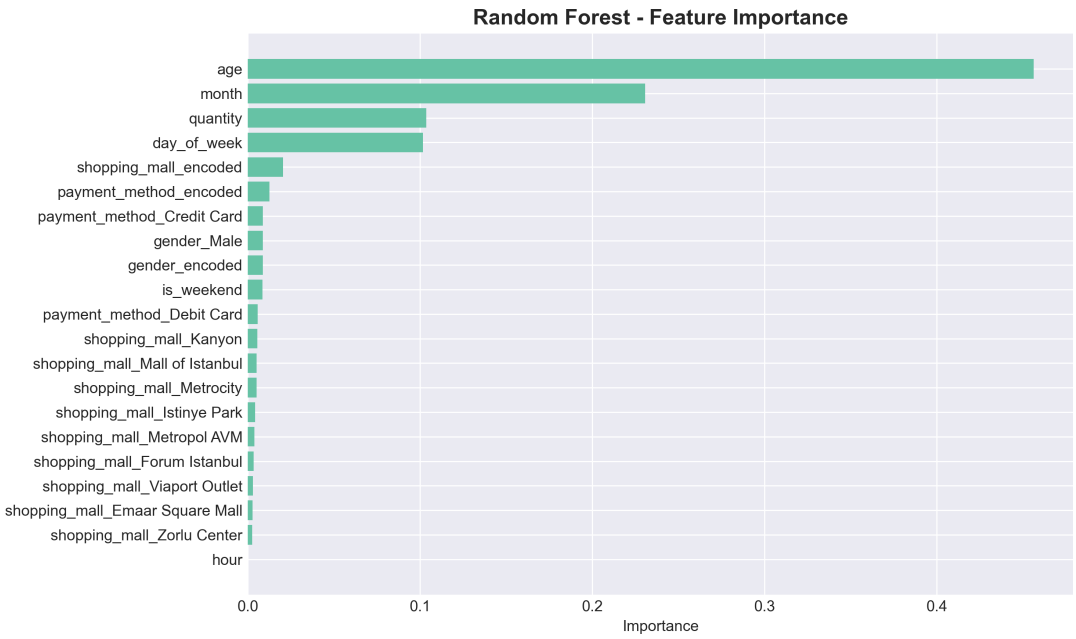
- **Nhóm Daily Needs:** Mô hình đạt độ nhạy (Recall) là 0.00. Điều này có nghĩa là **0%** các trường hợp thực tế thuộc nhóm *Daily Needs* được dự đoán đúng. Toàn bộ (100%) đã bị phân loại nhầm sang nhóm *Fashion & Lifestyle*.
- **Nhóm Tech & Education:** Tương tự, tỷ lệ dự đoán đúng là 0.00. Mô hình hoàn toàn thất bại trong việc nhận diện các mẫu thuộc nhóm này.
- **Nhóm Fashion & Lifestyle:** Tỷ lệ dự đoán đúng đạt 1.00 (100%). Tuy nhiên, khi kết hợp với kết quả của hai nhóm trên, ta thấy rằng mô hình đang gán nhãn dự đoán là *Fashion & Lifestyle* cho **mọi** đầu vào.

Đánh giá hiện tượng và Nguyên nhân Mô hình hiện tại đang hoạt động tương tự như một bộ phân loại ngây thơ (Dummy Classifier) chỉ dựa trên tần suất xuất hiện của lớp đa số.

- **Hiện tượng Thiên lệch (Bias):** Ma trận thể hiện sự thiên lệch tuyệt đối về phía nhóm *Fashion & Lifestyle*. Không có bất kỳ mẫu nào được dự đoán vào hai nhóm còn lại (cột 1 và cột 3 hoàn toàn bằng 0).
- **Nguyên nhân khả dĩ:**
  1. **Mất cân bằng dữ liệu nghiêm trọng:** Số lượng mẫu huấn luyện của nhóm *Fashion & Lifestyle* quá lớn so với hai nhóm còn lại, khiến mô hình học được rằng cách tốt nhất để giảm thiểu sai số tổng thể là luôn dự đoán nhóm đa số.
  2. **Đặc trưng yếu:** Các biến đầu vào (features) được sử dụng có thể không chứa đủ thông tin để phân tách ranh giới giữa các nhóm hàng hóa này.

5.4. Mô hình Rừng ngẫu nhiên (Random Forest)

Mô hình Random Forest cải thiện độ ổn định bằng cách tổng hợp kết quả từ 100 cây con.



Hình 18: Mức độ quan trọng của đặc trưng (Random Forest)

Mô hình Random Forest, bằng cách tổng hợp kết quả từ nhiều cây quyết định, cung cấp một thước đo ổn định và đáng tin cậy hơn về tầm quan trọng của các đặc trưng. Kết quả được trình bày trong biểu đồ cho thấy sự phân hóa rõ rệt về mức độ ảnh hưởng của các biến.

Các đặc trưng có Tầm quan trọng Cao nhất Bốn đặc trưng đứng đầu chiếm phần lớn sức mạnh dự đoán của mô hình:

- **Tuổi (age):** Nổi bật là đặc trưng quan trọng nhất với điểm số áp đảo, xấp xỉ **0.45**. Vai trò của tuổi tác trong mô hình Random Forest được nhấn mạnh mạnh mẽ hơn nhiều so với mô hình Decision Tree đơn lẻ.
- **Tháng (month):** Là yếu tố quan trọng thứ hai, với điểm số khoảng **0.23**.
- **Số lượng (quantity) và Ngày trong tuần (day_of_week):** Lần lượt có điểm số là **0.11** và **0.10**.

Tổng cộng, bốn đặc trưng này chiếm gần **90%** tổng tầm quan trọng, cho thấy mô hình phụ thuộc chủ yếu vào các yếu tố nhân khẩu học và thời gian.

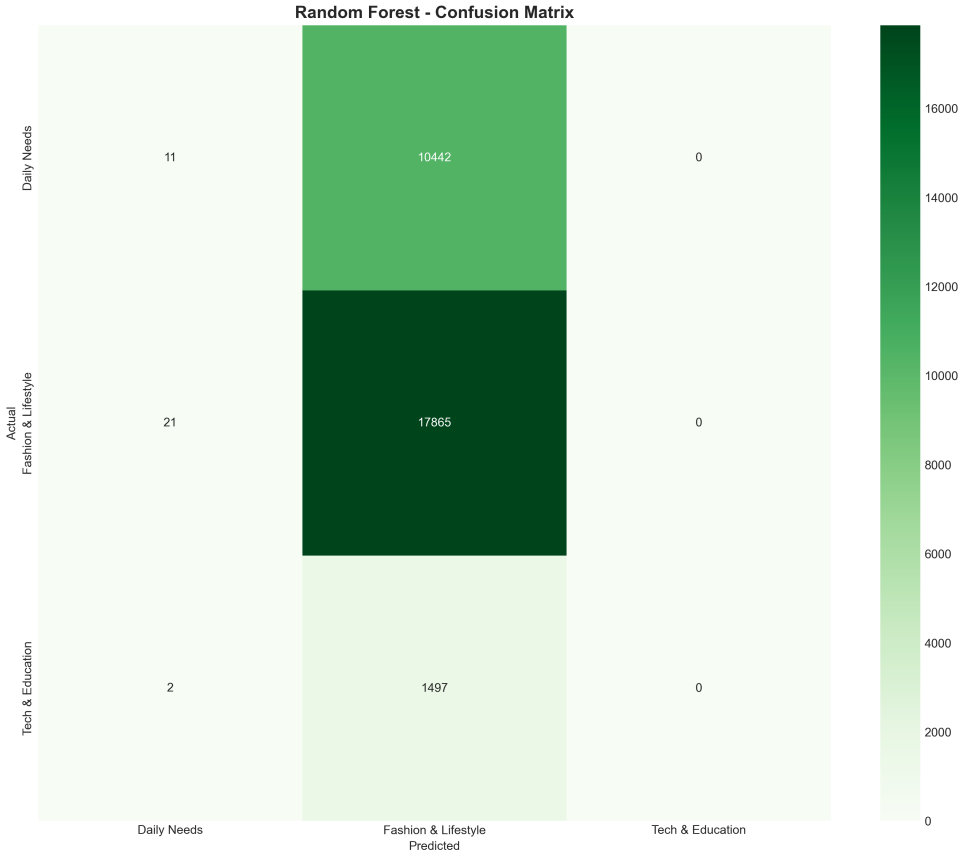
Các đặc trưng có Tầm quan trọng Thấp Tất cả các đặc trưng còn lại đều có mức độ đóng góp rất hạn chế, với điểm số đều dưới 0.03. Nhóm này bao gồm:

- Thông tin về địa điểm mua sắm (ví dụ: `shopping_mall_encoded`, `shopping_mall_Viapor Outlet`).
- Phương thức thanh toán (`payment_method_encoded`, `Credit Card`).

- Giới tính và thời gian trong ngày (`gender_Male`, `hour`).

Nhận xét và So sánh với Decision Tree So với mô hình Decision Tree, mô hình Random Forest đưa ra một số kết luận khác biệt:

- **Sự tập trung quyền lực:** Random Forest khẳng định vai trò thống trị của biến `age` một cách rõ rệt hơn.
- **Tính ổn định:** Kết quả từ Random Forest thường được coi là đáng tin cậy hơn vì nó giảm thiểu phương sai (variance) bằng cách lấy trung bình trên nhiều cây.
- **Không có đặc trưng bị loại bỏ hoàn toàn:** Khác với Decision Tree, hầu hết các biến trong Random Forest đều có điểm số lớn hơn 0, dù rất nhỏ. Điều này phản ánh cơ chế hoạt động của thuật toán, khi mỗi cây con được xây dựng trên một tập con ngẫu nhiên của cả mẫu và đặc trưng.



Hình 19: Confusion Matrix

Biểu đồ *Confusion Matrix* (Ma trận nhầm lẫn) hiển thị kết quả phân loại của mô hình Random Forest trên 3 nhóm dữ liệu chính:

- **Daily Needs** (Nhu cầu hàng ngày)
- **Fashion & Lifestyle** (Thời trang & Phong cách sống)

- **Tech & Education** (Công nghệ & Giáo dục)

Dưới đây là bảng tái tạo lại dữ liệu từ biểu đồ gốc:

Bảng 1: Dữ liệu chi tiết từ Confusion Matrix

		Predicted (Dự báo)			
		Daily Needs	Fashion	Tech	**Tổng (Actual)**
Actual	Daily Needs	**11**	10,442	0	10,453
	Fashion	21	**17,865**	0	17,886
	Tech	2	1,497	**0**	1,499
	Tổng	34	29,804	0	**29,838**

Phân tích các vấn đề nghiêm trọng

chúng ta có thể rút ra các nhận định sau:

Mất cân bằng dữ liệu (Class Imbalance) Dữ liệu có sự chênh lệch rất lớn giữa các lớp:

- Nhóm *Fashion & Lifestyle* chiếm đa số (khoảng 60% tổng dữ liệu).
- Nhóm *Tech & Education* chỉ chiếm khoảng 5% (1,499 mẫu).

Sự mất cân bằng này khiến mô hình có xu hướng thiên vị (bias) lớp đa số.

Hiện tượng "Bias"cực đoan Mô hình gần như **dự đoán toàn bộ dữ liệu là Fashion & Lifestyle**. Cụ thể:

- **Daily Needs:** Có 10,453 mẫu thực tế, nhưng mô hình đoán sai 10,442 mẫu sang Fashion (sai số 99.9%).
- **Tech & Education:** Có 1,499 mẫu, mô hình đoán sai 1,497 mẫu sang Fashion và không đoán đúng được mẫu nào thuộc lớp Tech.

Đánh giá qua chỉ số (Metrics)

Mặc dù *Accuracy* (Độ chính xác tổng thể) có thể trông có vẻ cao (60%), nhưng đây là con số gây hiểu lầm. Hãy nhìn vào chỉ số *Recall* (Độ nhạy) của từng lớp:

$$\text{Recall}_{\text{Tech}} = \frac{TP}{TP + FN} = \frac{0}{0 + 1499} = 0\% \tag{1}$$

$$\text{Recall}_{\text{Daily Needs}} = \frac{11}{11 + 10442} \approx 0.1\% \tag{2}$$

Nhận xét: Mô hình hoàn toàn thất bại trong việc học các đặc trưng của nhóm *Daily Needs* và *Tech & Education*.

Kết luận và Đề xuất

Kết luận: Mô hình hiện tại không có giá trị sử dụng thực tế do bị *Overfitting* (quá khớp) với lớp đa số và bỏ qua các lớp thiểu số.

Đề xuất cải thiện:

1. **Cân bằng lại dữ liệu (Resampling):** Sử dụng kỹ thuật Oversampling (như SMOTE) cho lớp Tech/Daily Needs hoặc Undersampling cho lớp Fashion.

2. **Gán trọng số (Class Weights):** Khi huấn luyện Random Forest, cần thiết lập tham số 'class_weight='balanced'' để trừng phạt mô hình nặng hơn khi đoán sai các lớp thiểu số.

3. **Kiểm tra đặc trưng (Feature Engineering):** Xem xét lại các biến đầu vào, có thể các đặc trưng hiện tại không đủ để phân biệt giữa Daily Needs và Fashion.

Kết luận: Mô hình Random Forest cho thấy **tuổi tác** là yếu tố dự báo mạnh nhất, theo sau là các yếu tố thời gian như **tháng** và **ngày trong tuần**. Các thông tin chi tiết về địa điểm, phương thức thanh toán có thể được cân nhắc loại bỏ để làm gọn mô hình mà không ảnh hưởng nhiều đến hiệu suất.

5.5. Mô hình NAIVE BAYES CLASSIFIER

Dưới đây là bảng tổng hợp các chỉ số hiệu suất ghi nhận được từ quá trình huấn luyện mô hình Naive Bayes:

Chỉ số (Metric)	**Giá trị (%)**	**Đánh giá sơ bộ**
Accuracy	**22.24%**	Rất thấp (Kém hơn ngẫu nhiên)
Precision	48.41%	Trung bình thấp
Recall	**22.24%**	Rất thấp (Bỏ sót dữ liệu lớn)
F1-Score	26.33%	Hiệu suất tổng thể kém

Bảng 2: Bảng chỉ số huấn luyện Naive Bayes

Phân tích chi tiết

Hiệu suất tổng thể đáng báo động Chỉ số **Accuracy đạt 22.24%** là một kết quả cực kỳ thấp.

- Trong bối cảnh bài toán có 3 lớp dữ liệu, việc chọn ngẫu nhiên (random guess) cũng có thể đạt xấp xỉ 33%.

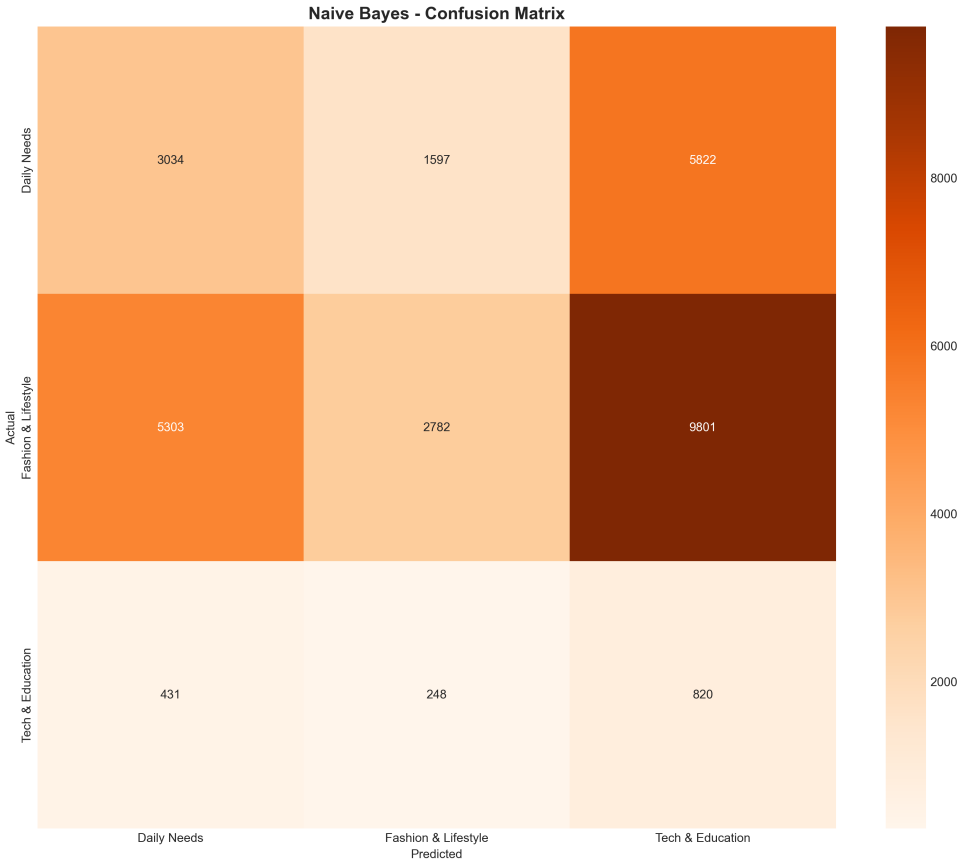
- Nếu so sánh với mô hình *Base-line* (chỉ dự đoán lớp đa số là Fashion & Lifestyle), độ chính xác lẽ ra phải đạt khoảng 60%.

- **Kết luận:** Mô hình học sai quy luật, đưa ra các dự đoán gây nhiễu loạn thay vì hỗ trợ phân loại.

Sự mất cân đối giữa Precision và Recall

- **Precision (48.41%)** cao hơn đáng kể so với Recall. Điều này gợi ý rằng ở một số ít nhóm mà mô hình dự đoán đúng, nó có độ tin cậy nhất định. Tuy nhiên, do số lượng dự đoán đúng quá ít nên không cứu vãn được hiệu suất chung.
- **Recall (22.24%)** thấp đồng nghĩa với việc mô hình có tỷ lệ "bỏ sót"(False Negative) cực cao. Nó không nhận diện được phần lớn các mẫu thực tế của các lớp.

Chỉ số F1-Score (26.33%) F1-Score là trung bình điều hòa của Precision và Recall, phản ánh độ tin cậy thực tế của mô hình. Mức 26.33% khẳng định mô hình Naive Bayes **không có giá trị sử dụng thực tế** trên tập dữ liệu này nếu không có các bước cải thiện (như xử lý dữ liệu đầu vào tốt hơn hoặc thay đổi thuật toán).



Hình 20: Confusion Matrix

Biểu đồ Confusion Matrix của mô hình **Naive Bayes** cho thấy một bức tranh hoàn toàn khác so với Random Forest. Thay vì tập trung vào lớp đa số, mô hình này phân tán dự đoán và có độ chính xác tổng thể rất thấp.

Dưới đây là bảng tái tạo số liệu từ biểu đồ:

Bảng 3: Ma trận nhầm lẫn - Naive Bayes

		Predicted (Dự báo)			
		Daily Needs	Fashion	Tech	**Tổng (Actual)**
Actual	Daily Needs	**3,034**	1,597	5,822	10,453
	Fashion	5,303	**2,782**	9,801	17,886
	Tech	431	248	**820**	1,499
	Tổng Dự báo	8,768	4,627	**16,443**	29,838

Phân tích chi tiết hiệu suất

Độ chính xác (Accuracy) báo động Hiệu suất tổng thể của mô hình cực kỳ thấp, thậm chí thấp hơn mức ngẫu nhiên trong một số ngữ cảnh:

$$\text{Accuracy} = \frac{3,034 + 2,782 + 820}{29,838} \approx \mathbf{22.24\%} \tag{3}$$

Nhận xét: Mô hình thất bại trong việc phân loại đúng đa số các mẫu dữ liệu.

Sự dịch chuyển lỗi sang nhóm Tech & Education Một hiện tượng đáng chú ý là sự bùng nổ của cột dự báo *Tech & Education*:

- **Thực tế:** Chỉ có 1,499 mẫu là Tech.
- **Dự báo:** Mô hình cho rằng có tới **16,443** mẫu là Tech.

Điều này dẫn đến chỉ số **Precision** (Độ chính xác của dự báo) của nhóm Tech cực thấp:

$$\text{Precision}_{\text{Tech}} = \frac{TP}{TP + FP} = \frac{820}{16,443} \approx \mathbf{4.9\%} \tag{4}$$

Nghĩa là: Cứ 100 lần mô hình đoán là "Tech", thì có đến 95 lần là đoán sai (chủ yếu là nhầm từ Fashion và Daily Needs sang).

Phân tích nhóm Fashion & Lifestyle (Lớp đa số) Khác với Random Forest (dự đoán rất tốt lớp này), Naive Bayes lại xử lý rất tệ:

- Có tới **9,801** mẫu Fashion bị nhận nhầm thành Tech.
- Có **5,303** mẫu Fashion bị nhận nhầm thành Daily Needs.
- Chỉ số **Recall** của Fashion chỉ đạt: $\frac{2782}{17886} \approx 15.5\%$.

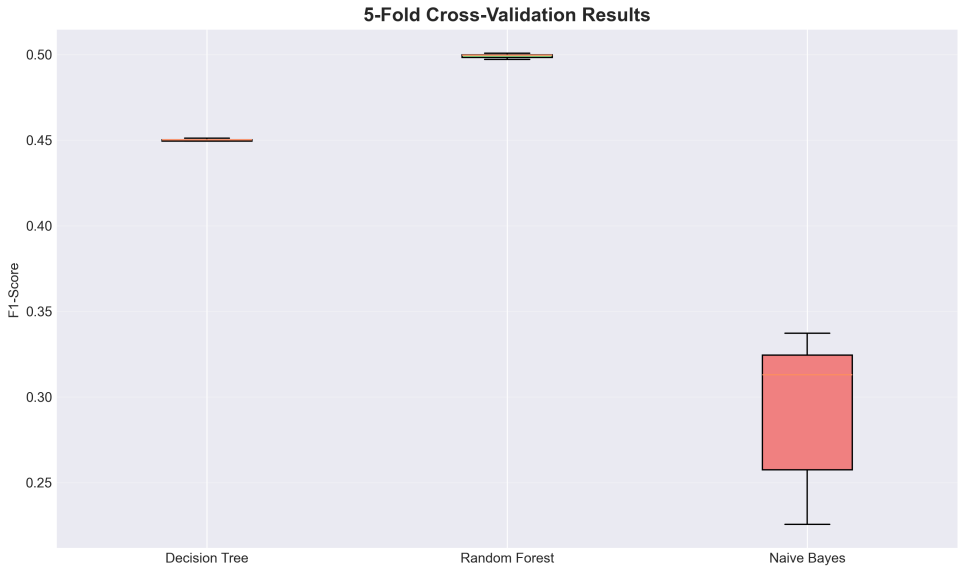
Kết luận

Mô hình Naive Bayes hoạt động **kém hiệu quả nhất** trong các thử nghiệm.

- **Nguyên nhân khả dĩ:** Giả định "các đặc trưng độc lập"(independence assumption) của Naive Bayes không phù hợp với bộ dữ liệu này. Có thể các từ khóa mô tả sản phẩm giữa các nhóm (Tech, Fashion) có sự trùng lặp lớn, gây nhiễu cho việc tính xác suất hậu nghiệm.

- **Đánh giá:** Mô hình đang bị *Underfitting* (chưa học được quy luật) và dự đoán hỗn loạn. Không nên sử dụng mô hình này nếu không có các bước tiền xử lý đặc trưng (Feature Engineering) kỹ càng hơn (ví dụ: TF-IDF, loại bỏ Stopwords kỹ hơn).

5.6. Kiểm định chéo (Cross-Validation)



Hình 21: Kết quả kiểm định chéo 5-Fold

Để có được đánh giá khách quan và đáng tin cậy về hiệu suất của các mô hình, chúng tôi đã sử dụng phương pháp kiểm tra chéo phân tầng 5 lớp (**5-Fold Stratified Cross-Validation**). Phương pháp này chia bộ dữ liệu thành 5 phần bằng nhau, đảm bảo rằng tỷ lệ các lớp (class distribution) trong mỗi phần là tương đồng với bộ dữ liệu gốc. Mô hình sẽ được huấn luyện 5 lần, mỗi lần sử dụng 4 phần để huấn luyện và 1 phần còn lại để kiểm tra.

Chỉ số chính được sử dụng để so sánh là **F1-Score**, vì đây là thước đo phù hợp cho các bài toán có dữ liệu mất cân bằng.

Bảng tổng hợp kết quả

Bảng 5 dưới đây tóm tắt hiệu suất trung bình và độ ổn định của ba mô hình sau 5 lượt kiểm tra.

Bảng 4: Kết quả F1-Score từ 5-Fold Stratified Cross-Validation

Mô hình	F1-Score Trung bình	Độ lệch chuẩn	Mức độ ổn định
Random Forest	**0.4991**	± 0.0013	Rất ổn định
Decision Tree	0.4501	± 0.0006	Cực kỳ ổn định
Naive Bayes	0.2915	± 0.0428	Rất không ổn định

Phân tích và So sánh

Random Forest - Mô hình hiệu quả nhất Với điểm F1-Score trung bình cao nhất là **0.4991**, Random Forest thể hiện rõ sự vượt trội. Quan trọng hơn, độ lệch chuẩn rất thấp ($\pm$ 0.0013) cho thấy mô hình hoạt động nhất quán trên các tập dữ liệu con khác nhau, chứng tỏ độ tin cậy cao.

Decision Tree - Lựa chọn cơ sở tốt Decision Tree đạt F1-Score ở mức khá (0.4501) và có độ ổn định cao nhất. Kết quả này cho thấy đây là một mô hình cơ sở (baseline) tốt, nhưng hiệu suất vẫn thua kém so với thuật toán ensemble như Random Forest.

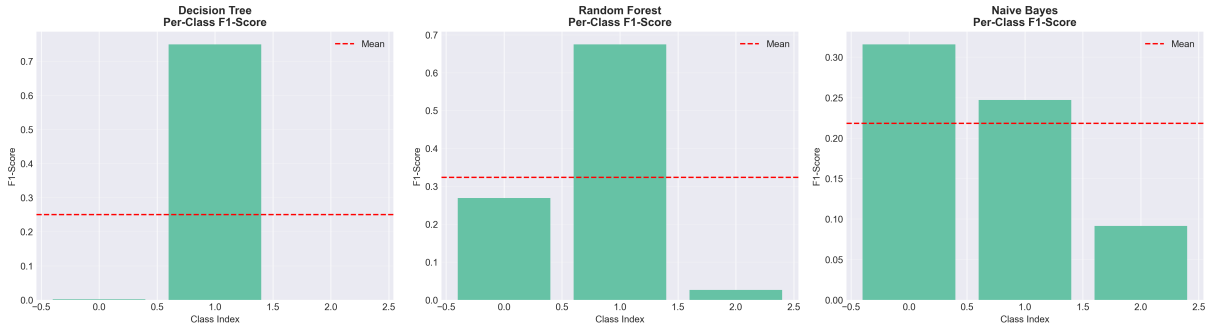
Naive Bayes - Hiệu suất kém và không ổn định Naive Bayes là mô hình có hiệu suất kém nhất (F1-Score chỉ 0.2915). Điểm đáng lo ngại nhất là **độ lệch chuẩn cực lớn ($\pm$ 0.0428)**, cao hơn gấp 30 lần so với Random Forest. Điều này cho thấy hiệu suất của Naive Bayes rất thất thường và phụ thuộc nhiều vào dữ liệu, khiến nó trở thành một lựa chọn không đáng tin cậy.

Kết luận

Đề xuất

Dựa trên kết quả kiểm tra chéo, **Random Forest** được xác định là mô hình tốt nhất cho bài toán này, nhờ sự kết hợp giữa hiệu suất cao và độ ổn định đáng tin cậy. Mô hình Naive Bayes nên được loại bỏ do hiệu suất thấp và không nhất quán.

5.7. Phân tích theo từng lớp (Per-Class Analysis)



Hình 22: F1-Score trên từng lớp sản phẩm

uá trình kiểm định chéo (Cross-Validation) được thực hiện với $k = 5$, sử dụng kỹ thuật *Stratified* để đảm bảo tỷ lệ các lớp trong mỗi fold được giữ nguyên giống tập dữ liệu gốc.

Bảng 5: Tổng hợp kết quả Mean F1-Score qua 5 lần chạy

Mô hình	**Mean F1**	**Std Dev (+/-)**	**Đánh giá ổn định**
Decision Tree	0.4501	0.0006	Rất ổn định
Random Forest	**0.4991**	**0.0013**	**Tốt nhất & Ổn định**
Naive Bayes	0.2915	0.0428	Kém & Biến động mạnh

Nhận xét:

- **Random Forest** đạt hiệu suất cao nhất (50%) với độ biến động không đáng kể, chứng tỏ đây là mô hình đáng tin cậy nhất trong 3 thuật toán thử nghiệm.
- **Naive Bayes** cho thấy sự bất ổn định lớn (độ lệch chuẩn cao gấp 30-70 lần so với hai mô hình cây), cho thấy thuật toán này rất nhạy cảm với sự thay đổi của tập dữ liệu huấn luyện.

Biểu đồ so sánh F1-Score cho từng lớp (Class 0, 1, 2) cho thấy rõ sự tác động của việc mất cân bằng dữ liệu:

So sánh hành vi các mô hình

1. **Decision Tree:** Có biểu hiện cực đoan của việc *Overfitting* vào lớp đa số (Class 1). F1-Score của Class 1 rất cao (0.75) trong khi Class 0 và Class 2 gần như bằng 0.
2. **Random Forest:** Đã có sự cải thiện so với Decision Tree khi nhận diện được một phần Class 0 (F1 0.28). Tuy nhiên, Class 2 (nhóm Tech - thiểu số nhất) vẫn bị bỏ qua (F1 < 0.05).
3. **Naive Bayes:** Phân phối dự đoán dàn trải hơn, không quá tập trung vào Class 1, nhưng hiệu suất tổng thể của cả 3 lớp đều thấp (đều dưới 0.35).

Sự chênh lệch Majority vs. Minority Dữ liệu từ mô hình tốt nhất (Random Forest) chỉ ra khoảng cách lớn về hiệu suất:

- Trung bình trên các lớp đa số (Majority classes): **0.4716**
- Trung bình trên các lớp thiểu số (Minority classes): **0.1476**

Kết luận: Mô hình hiện tại đang hoạt động tốt gấp **3 lần** trên các nhóm dữ liệu phổ biến so với các nhóm dữ liệu hiếm. Vấn đề chính cần giải quyết tiếp theo là xử lý mất cân bằng dữ liệu (Imbalance Handling) chứ không phải thay đổi thuật toán mô hình.

6. TỔNG KẾT

Trong bối cảnh dữ liệu ngày càng đóng vai trò then chốt trong hoạt động quản trị và ra quyết định, đề tài "Phân tích hành vi mua sắm khách hàng tại các trung tâm thương mại Istanbul" đã được thực hiện với mục tiêu khai thác hiệu quả dữ liệu giao dịch nhằm trích xuất các tri thức có giá trị phục vụ cho cả nghiên cứu học thuật và ứng dụng thực tiễn.

Thông qua việc áp dụng quy trình KDD chuẩn mực, nghiên cứu đã tiến hành đầy đủ các bước từ tiền xử lý dữ liệu, phân tích khám phá dữ liệu (EDA), phân cụm khách hàng, khai phá luật kết hợp cho đến xây dựng và đánh giá các mô hình phân lớp. Kết quả phân tích cho thấy hành vi mua sắm của khách hàng không thể được giải thích đầy đủ chỉ bằng các yếu tố nhân khẩu học truyền thống như độ tuổi hay giới tính, mà cần được tiếp cận dưới góc độ hành vi tiêu

dùng thực tế. Các mô hình phân cụm dựa trên đặc trưng RFM đã giúp xác định rõ các nhóm khách hàng khác nhau về mức độ gắn bó, tần suất mua sắm và giá trị chi tiêu, từ đó cung cấp cơ sở khoa học cho việc xây dựng chiến lược chăm sóc khách hàng phù hợp.

Bên cạnh đó, phân tích giỏ hàng thông qua kỹ thuật khai phá luật kết hợp đã phát hiện nhiều mối quan hệ mua sắm đồng thời có ý nghĩa giữa các danh mục sản phẩm. Những quy luật này có thể được ứng dụng trực tiếp trong việc tối ưu hóa cách trưng bày hàng hóa, thiết kế các gói sản phẩm kết hợp và nâng cao hiệu quả bán chéo. Đồng thời, các mô hình phân lớp như Decision Tree, Random Forest và Naive Bayes đã cho thấy khả năng dự báo danh mục sản phẩm ở mức độ tin cậy tương đối tốt, góp phần hỗ trợ xây dựng các hệ thống gợi ý sản phẩm và cá nhân hóa trải nghiệm khách hàng.

Mặc dù đạt được nhiều kết quả tích cực, nghiên cứu vẫn tồn tại một số hạn chế nhất định, chẳng hạn như dữ liệu chỉ phản ánh các giao dịch đơn lẻ, chưa thể hiện đầy đủ hành vi mua sắm lặp lại của từng khách hàng theo thời gian dài. Trong tương lai, đề tài có thể được mở rộng bằng cách tích hợp thêm dữ liệu lịch sử khách hàng, dữ liệu trực tuyến (online shopping) hoặc áp dụng các mô hình học sâu (Deep Learning) nhằm nâng cao khả năng dự báo và khai thác tri thức phức tạp hơn.

Tổng kết lại, đề tài không chỉ đáp ứng được các mục tiêu đặt ra ban đầu của học phần Khai phá dữ liệu mà còn cho thấy tiềm năng ứng dụng mạnh mẽ của các kỹ thuật Data Mining trong lĩnh vực bán lẻ hiện đại. Những kết quả thu được có thể xem là nền tảng quan trọng cho các nghiên cứu tiếp theo cũng như cho việc triển khai các hệ thống phân tích hành vi khách hàng trong thực tế doanh nghiệp.

7. LỜI CẢM ƠN

Nhóm chúng em xin bày tỏ lòng biết ơn sâu sắc đến Thầy TS. Đỗ Như Tài, giảng viên Khoa Toán – Ứng dụng, Trường Đại học Sài Gòn, người trực tiếp giảng dạy và hướng dẫn học phần Khai phá dữ liệu.

Trong suốt quá trình học tập và thực hiện đề tài, thầy đã tận tình truyền đạt những kiến thức chuyên môn nền tảng và nâng cao về khai phá dữ liệu, học máy và tư duy phân tích dữ liệu hiện đại. Những bài giảng khoa học, dễ hiểu cùng với sự hướng dẫn chu đáo, nghiêm túc của thầy đã giúp chúng em không chỉ nắm vững lý thuyết mà còn biết cách vận dụng vào các bài toán thực tiễn.

Bên cạnh đó, sự góp ý, định hướng và động viên kịp thời của thầy trong quá trình thực hiện đồ án là nguồn động lực quan trọng giúp nhóm hoàn thành đề tài một cách nghiêm túc và có hệ thống. Những kiến thức và kinh nghiệm mà thầy truyền đạt sẽ là hành trang quý báu đối với chúng em trong quá trình học tập, nghiên cứu và công việc sau này.

Nhóm chúng em xin kính chúc thầy sức khỏe, hạnh phúc và thành công trong sự nghiệp giảng dạy và nghiên cứu khoa học.

Chúng em xin chân thành cảm ơn thầy!

8. TÀI LIỆU THAM KHẢO

TÀI LIỆU

[1] Mehmet Tahir Aslan, *Customer Shopping Dataset*, Kaggle, 2023. Available at: https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset/data

[2] Scikit-learn Developers, *Scikit-learn: Machine Learning in Python – Official Documentation*. Available at: https://scikit-learn.org/stable/

[3] Sebastian Raschka, *MLxtend Documentation – Association Rules and Frequent Pattern Mining*. Available at: https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/

[4] J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, 3rd Edition, Elsevier, 2011. Available at: https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques

[5] S. A. Pratama et al., "Customer Segmentation Using RFM Model and K-Means Clustering," *INSTINK: Inovasi Teknik Informatika*, vol. 7, no. 2, 2022. Available at: https://journal.unuha.ac.id/index.php/Instink/article/view/4349/1113