**ĐẠI HỌC KINH TẾ QUỐC DÂN**

**TRƯỜNG CÔNG NGHỆ**

KHOA HỆ THỐNG THÔNG TIN QUẢN LÝ

-----□□🕮□□-----



**BỘ MÔN: ỨNG DỤNG TRÍ TUỆ NHÂN TẠO TRONG KINH DOANH VÀ QUẢN LÝ**

**ĐỀ TÀI:**

**PHÂN LOẠI PHÂN KHÚC KHÁCH HÀNG TRONG LĨNH VỰC BÁN LẺ TRỰC TUYẾN DỰA TRÊN DỮ LIỆU UCI**

**Nhóm thực hiện : Nhóm 03**

**Giảng viên : Ths. Phạm Thảo**

**Lớp học phần : TIHT1123(224)\_02**

*Hà Nội, tháng 5 năm 2025*

# THÀNH VIÊN NHÓM 03

|  |  |
| --- | --- |
| **Họ và tên** | **Mã sinh viên** |
| Nguyễn Tuấn Anh | 11220494 |
| Nguyễn Thanh Chúc | 11221121 |
| Nguyễn Cẩm Tú | 11226662 |
| Nguyễn Sơn Tùng | 11226745 |

# MỤC LỤC

[**THÀNH VIÊN NHÓM 03 1**](#_p4bcx4sc9hjx)

[**MỤC LỤC 2**](#_fx181xt9678m)

[**LỜI MỞ ĐẦU 4**](#_k3rfjmudu5ie)

[**I. MÔ TẢ CHUNG 5**](#_wmo4ts2nd8mq)

[1. Bài toán thực tế 5](#_9uwtlgng8sec)

[2. Quy trình truyền thống 6](#_frze49ckva2m)

[3. Giải pháp ứng dụng AI cho doanh nghiệp 7](#_1wgb1tx7ixd8)

[**II. XÁC ĐỊNH BÀI TOÁN 7**](#_63y8yp1d3di0)

[1. Tên đề tài 7](#_1al1wtb3q2ax)

[2. Mục tiêu của đề tài 8](#_rn05f9rp505b)

[**III. XÁC ĐỊNH, CHUẨN HÓA DỮ LIỆU 8**](#_axrptpcxwt2b)

[1. Xác định nguồn dữ liệu và phương pháp thu thập 8](#_xq2nmd6ww14i)

[2. Tiền xử lý dữ liệu 10](#_4sjs5jb9a7kl)

[2.1. Làm sạch dữ liệu 10](#_whmx4tl9wv4i)

[2.2. Biến đổi dữ liệu 11](#_2tvn8st6ts0i)

[3. Chuẩn hóa dữ liệu 11](#_vv6ny4s14g4w)

[3.1. Phương pháp chuẩn hóa 11](#_dig8dr7ttu6q)

[3.2. Các bước thực hiện: 11](#_c4qy9yndhecf)

[**IV. THIẾT KẾ VÀ XÂY DỰNG MÔ HÌNH 12**](#_fp7m23pbytuh)

[1. Thiết kế 12](#_w6gv3aqcs8w3)

[1.1. Khái niệm thuật toán Kmeans 12](#_d16a3uxfnkr8)

[1.2. Mô hình toán học của thuật toán K-Means 13](#_hx5g0ubyhyp4)

[1.3. Ưu điểm và hạn chế 13](#_7agrchmlb2td)

[1.4. Ứng dụng trong phân cụm khách hàng theo RFM 14](#_4r8tvwsjrk72)

[2. Xây dựng mô hình 14](#_63f791r7dqd)

[1.1. Đầu vào 14](#_te0p3go1ipaq)

[1.2. Đầu ra 15](#_sgg8od3n41lg)

[1.3. Xuất ra mô hình 17](#_ttofu7udeuqa)

[**V. THỬ NGHIỆM VÀ ĐÁNH GIÁ MÔ HÌNH 17**](#_eftskfmch4bs)

[1. Làm thế nào để biết mô hình tốt hay chưa tốt? 17](#_2zulj8h1xwwd)

[2. Thử và so sánh 20](#_7w93j5csxftz)

[3. Phương pháp đánh giá 29](#_wtzn1f6qiekw)

[3.1. Đánh giá hình học 30](#_byec7yprzvwj)

[3.2. Đánh giá ý nghĩa và khả năng giải thích cụm 31](#_mwjdj4qkrq3l)

[4. Kết luận 32](#_bhfw9m8gg28l)

[4.1. Trường hợp tốt nhất & ứng dụng 35](#_o5upbf79bo1i)

[4.2. Lý giải hiệu quả của Cluster 1 35](#_cvk5jhm50edw)

[4.3. Ví dụ minh họa với 3 features RFM 36](#_7xwwmr7emqnt)

[5. Xuất mô hình tốt nhất 37](#_mvu83w2m6nx7)

[6. Viết một đoạn code để có thể thực thi các trường hợp thử này 38](#_mrtmxixw47ah)

[**VI. XÂY DỰNG GIAO DIỆN NGƯỜI SỬ DỤNG 38**](#_4et34o1jdkan)

[1. Xác định đầu vào 38](#_6biuzlceytzp)

[1.1. Đầu vào tương ứng với phương thức nhập liệu thủ công 39](#_m3gk4q9pytyi)

[1.2. Đầu vào tương ứng với phương thức tải lên tệp dữ liệu (CSV) 40](#_rcgk9gr1rr0d)

[2. Xác định đầu ra 41](#_skp0zkrz8yod)

[2.1. Đầu ra tương ứng với phương thức nhập liệu thủ công 42](#_zgb808fd2p8a)

[2.2. Đầu ra tương ứng với phương thức tải lên tệp dữ liệu (CSV) 42](#_tkf5nrpho60g)

[**LỜI KẾT 45**](#_1skx4z70kna9)

# LỜI MỞ ĐẦU

Trong bối cảnh chuyển đổi số đang diễn ra mạnh mẽ, lĩnh vực bán lẻ trực tuyến không ngừng mở rộng và cạnh tranh ngày càng khốc liệt. Việc thấu hiểu hành vi và nhu cầu của khách hàng trở thành yếu tố then chốt giúp doanh nghiệp nâng cao trải nghiệm người dùng, tối ưu hóa chiến lược marketing và tăng trưởng bền vững. Với sự hỗ trợ của trí tuệ nhân tạo (AI), đặc biệt là các phương pháp học máy, việc phân loại và phân khúc khách hàng từ những tập dữ liệu lớn đã trở nên khả thi, chính xác và hiệu quả hơn bao giờ hết.

Xuất phát từ nhu cầu thực tiễn đó, nhóm 3 thực hiện đề tài “Phân loại phân khúc khách hàng trong lĩnh vực bán lẻ trực tuyến dựa trên dữ liệu UCI” nhằm nghiên cứu và ứng dụng các thuật toán AI trong việc khai thác, xử lý và phân tích dữ liệu khách hàng. Dữ liệu được sử dụng trong đề tài là tập dữ liệu khách hàng từ kho dữ liệu UCI - một nguồn thông tin có độ tin cậy cao và được sử dụng rộng rãi trong nghiên cứu khoa học.

Đề tài được thực hiện dưới sự hướng dẫn tận tình của thầy Phạm Thảo. Nhóm 3 xin chân thành cảm ơn thầy vì những góp ý chuyên môn quý báu giúp chúng em định hướng rõ ràng và hoàn thiện nội dung nghiên cứu.

Hy vọng rằng, với những kiến thức đã tiếp thu được cùng quá trình tìm hiểu nghiêm túc, đề tài sẽ mang lại cái nhìn thực tiễn về việc ứng dụng AI trong phân khúc khách hàng và đóng góp phần nào vào việc nâng cao hiệu quả trong hoạt động kinh doanh trực tuyến!

## I. MÔ TẢ CHUNG

### 1. Bài toán thực tế

- Trong bối cảnh cạnh tranh ngày càng gay gắt của ngành bán lẻ, các doanh nghiệp đang đối mặt với thách thức trong việc hiểu và đáp ứng nhu cầu đa dạng của khách hàng. Một nhà bán lẻ trực tuyến tại Anh cũng không ngoại lệ khi gặp khó khăn trong việc tiếp cận đúng đối tượng khách hàng với các chiến dịch marketing và dịch vụ cá nhân hóa. Nguyên nhân chính xuất phát từ việc thiếu một hệ thống phân tích dữ liệu khách hàng chuyên sâu và khoa học, dẫn đến nhiều hệ lụy trong hoạt động kinh doanh:

* Thứ nhất, doanh nghiệp đang lãng phí đáng kể ngân sách marketing khi triển khai các chiến dịch quảng cáo chung chung, không có sự phân tầng rõ ràng. Cùng một thông điệp tiếp thị được gửi đến tất cả khách hàng mà không tính đến sự khác biệt về nhu cầu, hành vi mua sắm hay giá trị mang lại cho doanh nghiệp.
* Thứ hai, cơ hội nâng cao trải nghiệm khách hàng thông qua dịch vụ cá nhân hóa bị bỏ ngỏ. Do không phân tích được dữ liệu hành vi, doanh nghiệp không thể cung cấp các ưu đãi, sản phẩm hay dịch vụ phù hợp với từng phân khúc khách hàng cụ thể.
* Thứ ba, việc xác định và chăm sóc nhóm khách hàng có giá trị cao (high-value customers) trở nên khó khăn. Doanh nghiệp đang đánh giá khách hàng theo cảm tính hoặc chỉ dựa trên một vài tiêu chí đơn giản, dẫn đến bỏ sót nhiều cơ hội kinh doanh tiềm năng.

- Giải pháp ứng dụng AI mà nhóm đề xuất hướng đến việc giải quyết hiệu quả các vấn đề trên. Hệ thống phân loại khách hàng tự động dựa trên phân tích RFM (Recency, Frequency, Monetary) kết hợp với thuật toán K-means sẽ giúp phân chia khách hàng thành các nhóm có đặc điểm rõ ràng. Quy trình triển khai bao gồm các bước quan trọng từ tiền xử lý dữ liệu, xây dựng mô hình đến ứng dụng thực tế. Mỗi nhóm khách hàng sau khi được xác định sẽ nhận được các chiến dịch marketing phù hợp, từ voucher mặt hàng cao cấp cho nhóm tiềm năng đến chương trình ưu đãi đặc biệt nhằm kéo nhóm khách đang dần rời bỏ doanh nghiệp quay trở lại.

- Kết quả kỳ vọng của đề tài là một mô hình hoàn chỉnh có thể tích hợp vào hệ thống quản lý khách hàng hiện có của doanh nghiệp. Hệ thống này sẽ thay thế phương pháp phân loại thủ công truyền thống bằng giải pháp tự động, nhanh chóng và chính xác hơn. Về lâu dài, việc ứng dụng mô hình này sẽ góp phần nâng cao tỷ lệ quay lại mua hàng của khách hàng, đồng thời tối ưu hóa chi phí marketing và gia tăng hiệu quả kinh doanh tổng thể cho doanh nghiệp.

### 2. Quy trình truyền thống

- Trước khi ứng dụng AI, các doanh nghiệp thường phân loại khách hàng theo phương pháp thủ công với nhiều hạn chế. Quy trình này gồm các bước:

* Thu thập dữ liệu: Nhân viên tổng hợp từ hệ thống POS (mã vạch, tổng chi tiêu) vào Excel.
* Phân nhóm đơn giản:
  + Nhóm "VIP" nếu tổng chi tiêu > 20 triệu/năm.
  + Nhóm "Thường xuyên" nếu mua > 10 lần/năm.
* Marketing thủ công:
  + Gửi SMS khuyến mãi giống nhau cho từng nhóm.
  + Phiếu giảm giá in tại quầy thu ngân không cá nhân hóa.
* Đánh giá: Dựa trên doanh thu chung, không đo lường được hiệu quả từng phân khúc.

### 3. Giải pháp ứng dụng AI cho doanh nghiệp

- Nhóm đề xuất hệ thống phân loại khách hàng tự động bằng AI, tập trung vào:

* Phân tích RFM (Recency, Frequency, Monetary):
  + **Recency**: Ngày mua gần nhất (khách mua gần đây dễ phản hồi khuyến mãi).
  + **Frequency**: Tần suất mua (khách thường xuyên cần chính sách duy trì).
  + **Monetary**: Giá trị chi tiêu (ưu tiên nhóm có đóng góp doanh thu cao).
* Kết hợp thuật toán K-means để chia cụm khách hàng thành 3 nhóm:
  + **Khách hàng phổ thông**: vẫn còn mua hàng trong khoảng gần đây nhưng tần suất và chi tiêu không quá cao.
  + **Khách hàng VIP**: mua hàng gần đây, thường xuyên, và chi tiêu cao nhất. Đây là nhóm khách hàng tiềm năng nhất, cần ưu tiên giữ chân.
  + **Khách hàng đã rời đi hoặc ít tương tác**: đã lâu không quay lại, mua ít, chi tiêu thấp.

## II. XÁC ĐỊNH BÀI TOÁN

### 1. Tên đề tài

Phân loại phân khúc khách hàng trong lĩnh vực bán lẻ trực tuyến dựa trên dữ liệu UCI

### 2. Mục tiêu của đề tài

- Đề tài này nhằm xây dựng một mô hình phân loại khách hàng dựa trên các tiêu chí như tần suất mua hàng, giá trị đơn hàng và thời gian gần nhất mua hàng. Mục tiêu của nhóm là xây dựng và triển khai mô hình bằng cách sử dụng các thuật toán phân cụm có sẵn từ thư viện Scikit-learn, nhằm hỗ trợ doanh nghiệp trong việc phân nhóm khách hàng một cách hiệu quả.

- Mô hình cho phép người dùng là nhân viên marketing hoặc quản lý kinh doanh có khả năng thực hiện việc xác định các nhóm khách hàng tiềm năng, từ đó đưa ra các chiến lược tiếp thị và chăm sóc phù hợp với từng nhóm.

- Hơn nữa, hệ thống còn cho phép người dùng là chuyên viên phân tích dữ liệu có khả năng khám phá các đặc điểm hành vi tiêu dùng theo từng nhóm, hỗ trợ việc cá nhân hóa dịch vụ, đề xuất sản phẩm và cải thiện trải nghiệm khách hàng.

## III. XÁC ĐỊNH, CHUẨN HÓA DỮ LIỆU

### 1. Xác định nguồn dữ liệu và phương pháp thu thập

- Nguồn dữ liệu:

* Dataset: Online Retail từ kho lưu trữ UCI, lưu trữ dưới định dạng CSV.
* Đặc điểm dữ liệu: Gồm 8 cột:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tên Biến** | **Vai Trò** | **Kiểu Dữ Liệu** | **Mô Tả** | **Đơn Vị** | **Missing Values** |
| InvoiceNo | ID | Categorical | Một số nguyên 6 chữ số được gán duy nhất cho mỗi giao dịch. Nếu bắt đầu bằng chữ 'C', tức là giao dịch đã bị huỷ | – | Không |
| StockCode | ID | Categorical | Một số nguyên 5 chữ số được gán duy nhất cho mỗi sản phẩm riêng biệt | – | Không |
| Description | Feature | Categorical | Tên sản phẩm | – | Không |
| Quantity | Feature | Integer | Số lượng sản phẩm trong mỗi giao dịch | – | Không |
| InvoiceDate | Feature | Date & Time | Ngày và thời gian giao dịch được tạo | – | Không |
| UnitPrice | Feature | Continuous | Giá bán của mỗi đơn vị sản phẩm | Bảng Anh | Không |
| CustomerID | Feature | Categorical | Một số nguyên 5 chữ số được gán duy nhất cho mỗi khách hàng | – | Không |
| Country | Feature | Categorical | Tên quốc gia nơi mỗi khách hàng cư trú | \_ | Không |

* Dữ liệu giao dịch từ 38 quốc gia, chủ yếu ở Vương quốc Anh (91.4%), bao gồm tất cả các giao dịch diễn ra trong khoảng thời gian từ ngày 01/12/2010 đến ngày 09/12/2011 của một doanh nghiệp bán lẻ trực tuyến không có cửa hàng, có trụ sở và đăng ký tại Vương quốc Anh. Công ty chủ yếu kinh doanh các mặt hàng quà tặng độc đáo dùng cho nhiều dịp khác nhau. Nhiều khách hàng của công ty là các nhà bán sỉ.

- Lý do chọn bộ dữ liệu: Phù hợp để xây dựng mô hình RFM (Recency, Frequency, Monetary) phân nhóm khách hàng.

- Phương pháp thu thập: Sử dụng thư viện pandas trong Python để đọc file CSV từ Google Drive.

### 2. Tiền xử lý dữ liệu

#### 2.1. Làm sạch dữ liệu

**- Xử lý missing values:**

* **CustomerID:** Loại bỏ các bản ghi thiếu CustomerID (do không thể xác định khách hàng)
* **Description**: Điền giá trị thiếu bằng chuỗi rỗng (nếu không ảnh hưởng đến phân tích)

**- Loại bỏ dữ liệu không hợp lệ:**

* **Giao dịch hủy**: Xóa các giao dịch có Quantity âm và InvoiceNo bắt đầu bằng "C".
* **Mã sản phẩm bất thường**: Loại bỏ các mã StockCode không đúng định dạng (ví dụ: “POST”, “BANK CHARGES”)
* **Giá sản phẩm bất thường:** loại bỏ các sản phẩm có giá bằng 0
* Sử dụng thuật toán Isolation Forest để phát hiện điểm bất thường (outliers / anomalies) trong dữ liệu.

#### 2.2. Biến đổi dữ liệu

- Tạo đặc trưng RFM:

* **Recency (R)**: Số ngày từ lần mua cuối đến ngày hiện tại.
* **Frequency (F)**: Tổng số giao dịch của khách hàng.
* **Monetary (M)**: Tổng giá trị đơn hàng (Quantity \* UnitPrice).

### 3. Chuẩn hóa dữ liệu

#### 3.1. Phương pháp chuẩn hóa

- Sử dụng StandardScaler từ thư viện scikit-learn để chuẩn hóa, biến đổi dữ liệu về dạng có mean=0 và variance=1, giúp các thuật toán phân cụm (như Kmeans) hoạt động hiệu quả.

#### 3.2. Các bước thực hiện:

**- Bước 1:** Chuẩn bị dữ liệu:

* Loại trừ các cột không cần chuẩn hóa (CustomerID, Is\_UK) bằng cách tạo danh sách columns\_to\_exclude.
* Xác định các cột cần chuẩn hóa (columns\_to\_scale) bằng phép lấy hiệu giữa tất cả cột và columns\_to\_exclude.

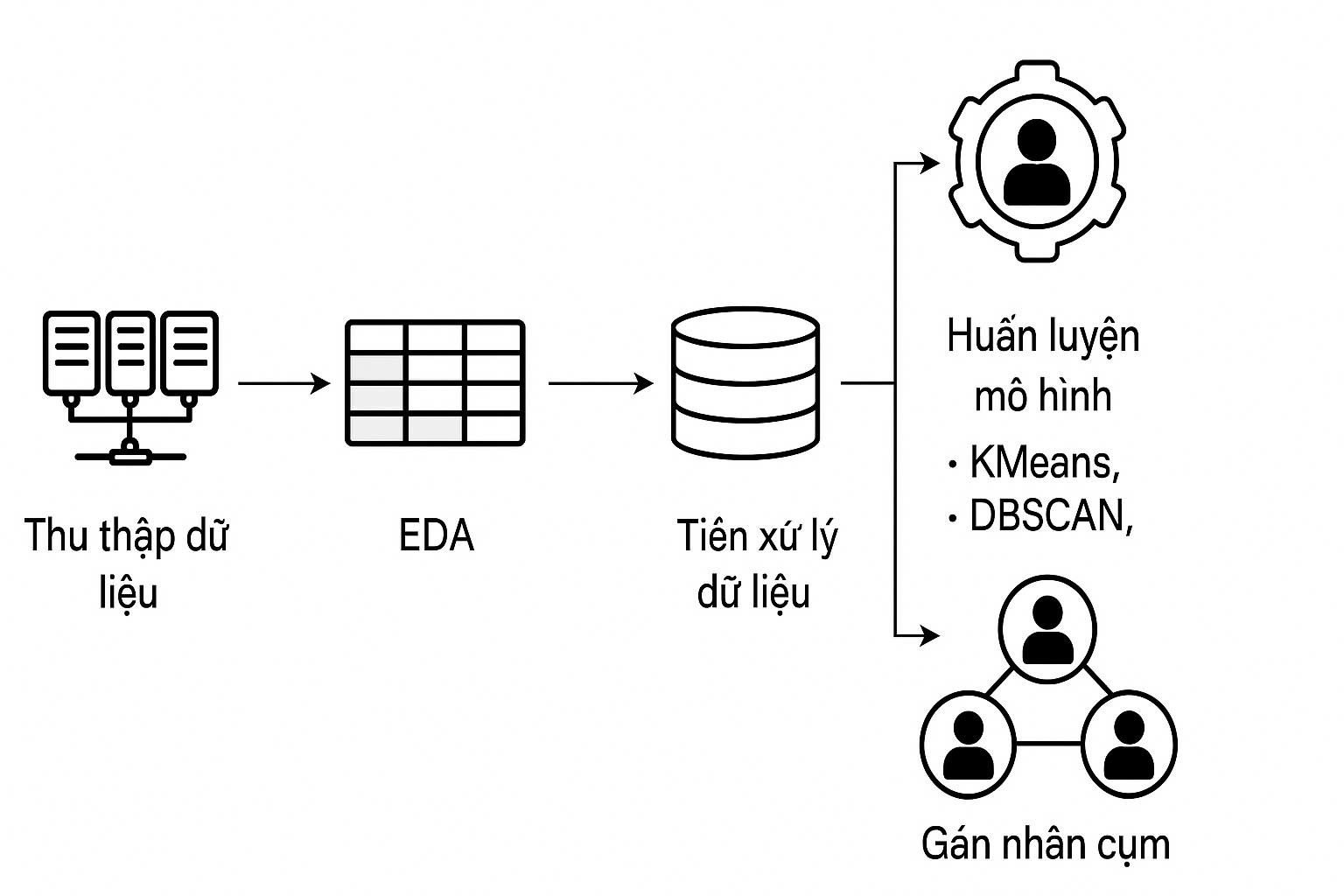
**- Bước 2:** Tạo bản sao dữ liệu:

* Tạo bản copy của DataFrame gốc (customer\_data\_cleaned) để tránh thay đổi dữ liệu gốc.

**- Bước 3:** Áp dụng StandardScaler

* Chuẩn hóa các cột đã chọn (columns\_to\_scale) bằng fit\_transform.

## IV. THIẾT KẾ VÀ XÂY DỰNG MÔ HÌNH



### 1. Thiết kế

#### 1.1. Khái niệm thuật toán Kmeans

- K-Means là một thuật toán phân cụm phổ biến trong học máy không giám sát (unsupervised learning), có nhiệm vụ chia dữ liệu thành k nhóm sao cho khoảng cách giữa các điểm dữ liệu với tâm cụm (centroid) gần nhất là nhỏ nhất. Đây là phương pháp thường được sử dụng trong phân tích hành vi khách hàng, phát hiện bất thường, và khai phá dữ liệu.

#### 1.2. Mô hình toán học của thuật toán K-Means

- Giả sử tập dữ liệu gồm n điểm dữ liệu X = {x₁, x₂, ..., xₙ}, trong đó mỗi điểm xᵢ ∈ ℝᵈ là một vector d chiều. Mục tiêu của thuật toán là chia tập dữ liệu thành k cụm C = {C₁, C₂, ..., Cₖ} sao cho hàm mục tiêu sau được tối ưu:

minimize ∑\_{j=1}^{k} ∑\_{xᵢ ∈ Cⱼ} ‖xᵢ − μⱼ‖²

- Trong đó:

* μⱼ là trung tâm (centroid) của cụm Cⱼ, được tính bằng trung bình các điểm dữ liệu thuộc cụm đó:

μⱼ = (1 / |Cⱼ|) × ∑\_{xᵢ ∈ Cⱼ} xᵢ

* ‖xᵢ − μⱼ‖² là khoảng cách Euclid giữa điểm dữ liệu xᵢ và trung tâm cụm μⱼ.

- Thuật toán hoạt động theo quy trình lặp sau:

* B1. Khởi tạo: Chọn ngẫu nhiên k điểm làm trung tâm cụm ban đầu.
* B2. Gán cụm: Gán mỗi điểm dữ liệu xᵢ vào cụm có trung tâm gần nhất theo công thức:

Cⱼ = {xᵢ : ‖xᵢ − μⱼ‖² ≤ ‖xᵢ − μₗ‖², ∀ l = 1,...,k}

* B3. Cập nhật: Tính lại các trung tâm cụm μⱼ theo công thức trung bình.
* B4. Lặp lại bước 2 và 3 cho đến khi hội tụ (tức là các cụm không thay đổi đáng kể).

#### 1.3. Ưu điểm và hạn chế

- Ưu điểm:

* Đơn giản, dễ hiểu và dễ triển khai.
* Tính toán nhanh, hiệu quả với tập dữ liệu lớn.

- Hạn chế:

* Phải xác định trước số lượng cụm k.
* Nhạy cảm với các giá trị ngoại lai (outlier) và khởi tạo ban đầu.
* Không đảm bảo hội tụ tới nghiệm tối ưu toàn cục.

#### 1.4. Ứng dụng trong phân cụm khách hàng theo RFM

- Trong bài toán thực tế, mỗi khách hàng được biểu diễn bằng một vector xᵢ = (Rᵢ, Fᵢ, Mᵢ) tương ứng với các chỉ số:

* Rᵢ: Khoảng thời gian kể từ lần mua hàng gần nhất (Recency).
* Fᵢ: Tần suất mua hàng (Frequency).
* Mᵢ: Tổng giá trị giao dịch (Monetary).

- Trước khi áp dụng K-Means, dữ liệu RFM được chuẩn hóa (scaling) để đảm bảo các đặc trưng có cùng độ lớn, tránh làm sai lệch kết quả phân cụm.

- Sau khi thực hiện phân cụm, mỗi nhóm khách hàng sẽ có đặc điểm hành vi mua sắm tương tự nhau, giúp doanh nghiệp đưa ra chiến lược tiếp thị và chăm sóc khách hàng phù hợp như: tập trung vào nhóm khách hàng trung thành, tái kích hoạt nhóm khách hàng ít mua, hoặc tìm kiếm cơ hội từ nhóm có tiềm năng cao.

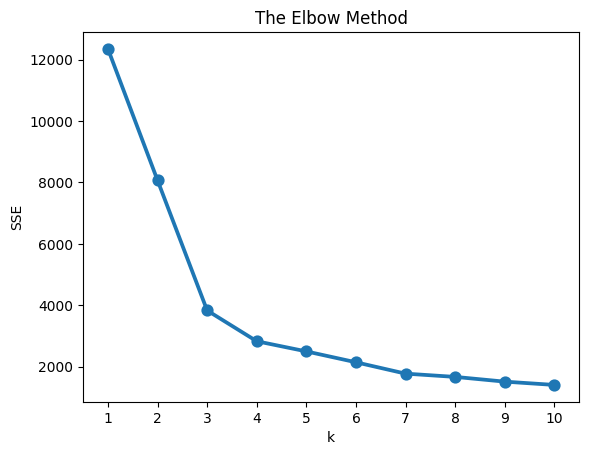
### 2. Xây dựng mô hình

#### 1.1. Đầu vào

- Có 3 feature đầu vào mô hình:

* Recency: số ngày kể từ lần mua hàng gần nhất của khách hàng.
* Frequency: tổng số giao dịch mà khách hàng thực hiện.
* Monetary: tổng số tiền mà khách hàng chi tiêu cho sản phẩm.

- Xử lý: Sử dụng StandardScaler để đưa dữ liệu RFM về cùng phân phối chuẩn.

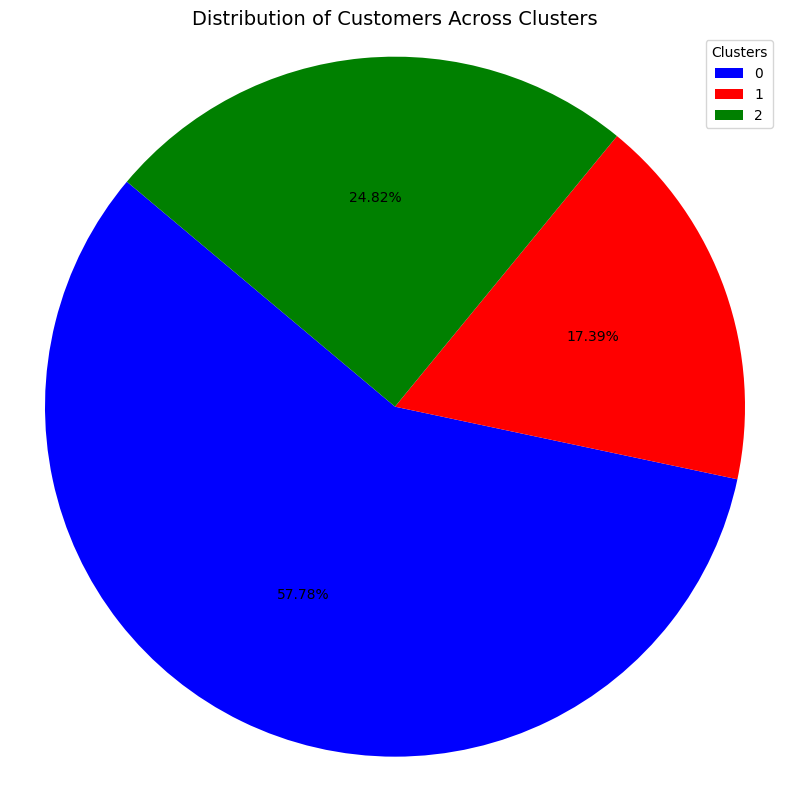


- Dùng Elbow Method (cho K-Means) Chạy K-Means với nhiều giá trị K, vẽ biểu đồ "Within-cluster SSE (Sum of Squared Errors)". Tìm “khuỷu tay” (elbow) – điểm mà SSE không giảm nhiều nữa ⇒ chọn K = 3

#### 1.2. Đầu ra

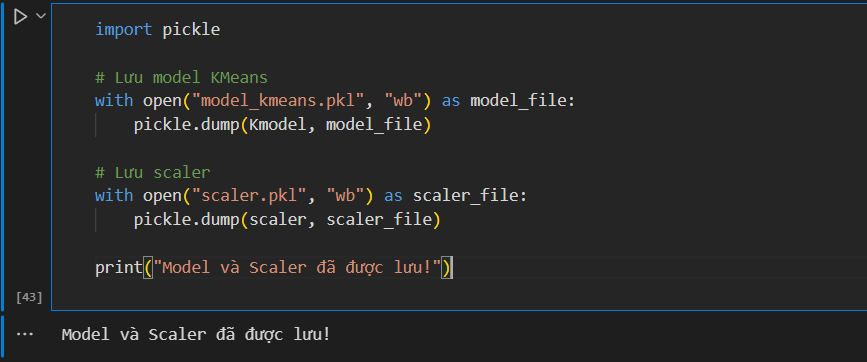
- Bảng phân cụm khách hàng: mỗi khách hàng được gán vào một cụm cụ thể dựa trên RFM.

- Biểu đồ thể hiện phân bố cụm khách hàng:



* **Cụm 0 (màu xanh dương)**: chiếm 57.78% khách hàng - Khách hàng phổ thông: vẫn còn mua hàng trong khoảng gần đây nhưng tần suất và chi tiêu không quá cao.
* **Cụm 1 (màu đỏ)**: chiếm 17.39% khách hàng - Khách hàng VIP: mua hàng gần đây, thường xuyên, và chi tiêu cao nhất. Đây là nhóm khách hàng tiềm năng nhất, cần ưu tiên giữ chân.
* **Cụm 2 (màu xanh lá)**: chiếm 24.82% khách hàng - Khách hàng đã rời đi hoặc ít tương tác: đã lâu không quay lại, mua ít, chi tiêu thấp.

#### 1.3. Xuất ra mô hình



## V. THỬ NGHIỆM VÀ ĐÁNH GIÁ MÔ HÌNH

### 1. Làm thế nào để biết mô hình tốt hay chưa tốt?

- Các chỉ số đánh giá nội tại mô hình:

* Silhouette Score có giá trị nằm trong khoảng [-1, 1]:
  + > 0.5: mô hình rất tốt.
  + 0.2 - 0.5: chấp nhận được (phân cụm có giá trị).
  + < 0.2: cụm chồng lấn, cần xem lại.
* Calinski-Harabasz Score có giá trị càng cao càng tốt và không có ngưỡng cố định, nhưng so sánh giữa các mô hình với số lượng cụm khác nhau sẽ hữu ích.
* Davies-Bouldin Score có giá trị càng thấp càng tốt và < 1.0 là tốt.

- Khi không có nhãn dữ liệu thực tế, ngoài việc sử dụng các chỉ số định lượng như Silhouette, Calinski-Harabasz và Davies-Bouldin, chúng ta cần đánh giá định tính - kiểm tra xem các cụm mà mô hình phân cụm tạo ra có thực sự “có ý nghĩa” và có thể áp dụng được trong thực tiễn hay không. Dưới đây là 3 tiêu chí chính để bạn đánh giá định tính mô hình:

- Các cụm có phản ánh hành vi tiêu dùng khác biệt không?

* Sau khi phân cụm, mô hình của bạn tạo ra 3 nhóm:
  + Cụm 1 gồm các khách hàng có Recency thấp, Frequency và Monetary cao, cho thấy đây là nhóm khách hàng có giá trị lớn và đang tích cực mua sắm – thường được nhận định là nhóm khách hàng VIP nên ưu tiên chăm sóc.
  + Cụm 2 có Recency rất cao cùng Frequency và Monetary thấp, đại diện cho nhóm khách hàng đã lâu không mua hàng và ít tương tác – là nhóm có dấu hiệu rời bỏ cần tái kích hoạt.
  + Cụm 0 bao gồm các khách hàng có Recency trung bình, Frequency vừa phải và chi tiêu ở mức trung bình, được xem là nhóm khách hàng phổ thông và tiềm năng để nuôi dưỡng lên nhóm VIP.

=> Như vậy, sự khác biệt giữa các cụm không chỉ thể hiện qua số liệu định lượng mà còn phản ánh rõ nét hành vi tiêu dùng thực tiễn.

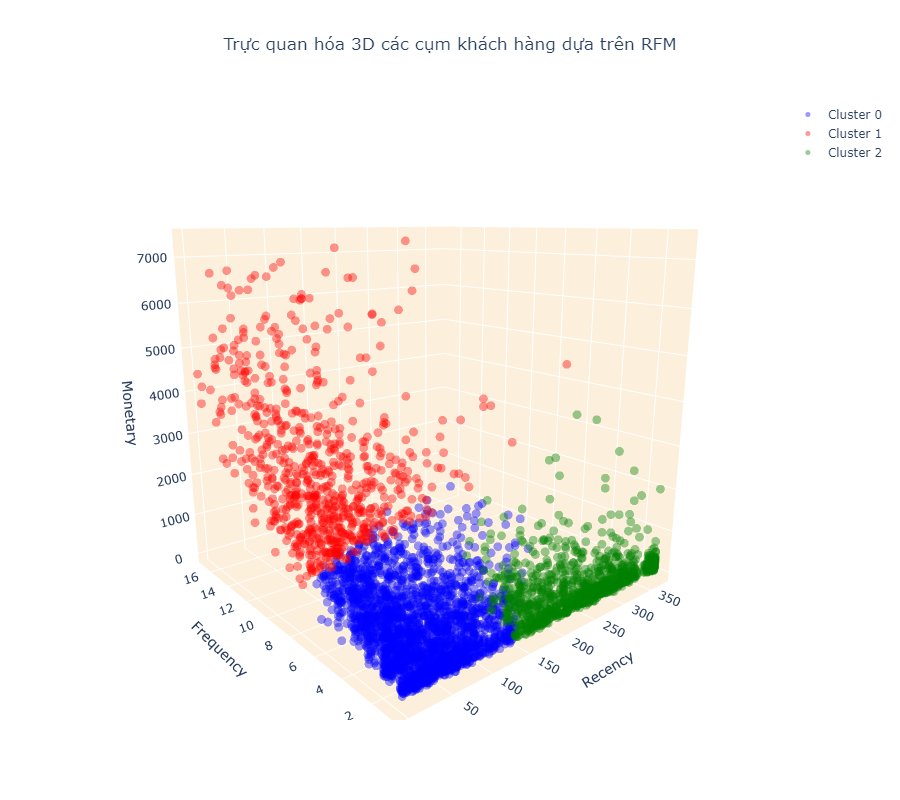
- Phân bố tỷ lệ giữa các cụm có hợp lý không?

* Phân bố này phù hợp với thực tiễn kinh doanh, nơi mà:
  + Phần lớn khách hàng là nhóm trung bình (Cluster 0)
  + Chỉ một phần nhỏ là khách chi tiêu cao (Cluster 1)
  + Một số khách có dấu hiệu rời bỏ (Cluster 2)

=> Phân bố này được đánh giá là hợp lý và phù hợp với thực tế kinh doanh trong ngành bán lẻ, nơi phần lớn khách hàng thường nằm trong nhóm phổ thông, trong khi nhóm khách hàng có giá trị cao (VIP) chỉ chiếm tỷ lệ nhỏ. Sự phân bố đồng đều và không quá chênh lệch cũng góp phần củng cố tính ổn định của mô hình.

- Mức độ ứng dụng thực tế của các cụm: Doanh nghiệp có thể làm gì từ kết quả này?

* Kết quả phân cụm hoàn toàn có thể hỗ trợ doanh nghiệp triển khai các chiến lược marketing và chăm sóc khách hàng khác nhau:
  + Nhóm VIP có thể được ưu tiên nhận các chương trình tri ân, khuyến mãi độc quyền hoặc nâng hạng thành viên.
  + Nhóm khách hàng có dấu hiệu rời bỏ có thể là đối tượng của các chiến dịch tái kích hoạt.
  + Nhóm phổ thông có thể được nuôi dưỡng nhằm nâng cao giá trị vòng đời khách hàng.



### 2. Thử và so sánh

**KMeans**

- Không xóa phần giao dịch bị hủy:

* 6 features

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **k** | **init** | **max\_iter** | **n\_init** | **Silhouette Score** | **Davies-Bouldin Score** | **Calinski-Harabasz Score** |
| 3 | k-means++ | 100 | 10 | 0.365406852 | 1.146351217 | 1942.700514 |
| 100 | 20 | 0.365406852 | 1.146351217 | 1942.700514 |
| 200 | 10 | 0.365406852 | 1.146351217 | 1942.700514 |
| 200 | 20 | 0.365406852 | 1.146351217 | 1942.700514 |
| 300 | 10 | 0.365406852 | 1.146351217 | 1942.700514 |
| 300 | 20 | 0.365406852 | 1.146351217 | 1942.700514 |
| random | 100 | 10 | 0.365406852 | 1.146351217 | 1942.700514 |
| 100 | 20 | 0.365406852 | 1.146351217 | 1942.700514 |
| 200 | 10 | 0.365406852 | 1.146351217 | 1942.700514 |
| 200 | 20 | 0.365406852 | 1.146351217 | 1942.700514 |
| 300 | 10 | 0.365406852 | 1.146351217 | 1942.700514 |
| 300 | 20 | 0.365406852 | 1.146351217 | 1942.700514 |
| 4 | k-means++ | 100 | 10 | 0.369804752 | 0.98719756 | 2079.59754 |
| 100 | 20 | 0.369804752 | 0.98719756 | 2079.59754 |
| 200 | 10 | 0.369804752 | 0.98719756 | 2079.59754 |
| 200 | 20 | 0.369804752 | 0.98719756 | 2079.59754 |
| 300 | 10 | 0.369804752 | 0.98719756 | 2079.59754 |
| 300 | 20 | 0.369804752 | 0.98719756 | 2079.59754 |
| random | 100 | 10 | 0.370862794 | 0.990075593 | 2079.33278 |
| 100 | 20 | 0.370711401 | 0.988505277 | 2079.362889 |
| 200 | 10 | 0.370862794 | 0.990075593 | 2079.33278 |
| 200 | 20 | 0.370711401 | 0.988505277 | 2079.362889 |
| 300 | 10 | 0.370862794 | 0.990075593 | 2079.33278 |
| 300 | 20 | 0.370711401 | 0.988505277 | 2079.362889 |
| 5 | k-means++ | 100 | 10 | 0.375883502 | 1.078538887 | 2078.577155 |
| 100 | 20 | 0.375883502 | 1.078538887 | 2078.577155 |
| 200 | 10 | 0.375883502 | 1.078538887 | 2078.577155 |
| 200 | 20 | 0.375883502 | 1.078538887 | 2078.577155 |
| 300 | 10 | 0.375883502 | 1.078538887 | 2078.577155 |
| 300 | 20 | 0.375883502 | 1.078538887 | 2078.577155 |
| random | 100 | 10 | 0.375883502 | 1.078538887 | 2078.577155 |
| 100 | 20 | 0.375883502 | 1.078538887 | 2078.577155 |
| 200 | 10 | 0.375883502 | 1.078538887 | 2078.577155 |
| 200 | 20 | 0.375883502 | 1.078538887 | 2078.577155 |
| 300 | 10 | 0.375883502 | 1.078538887 | 2078.577155 |
| 300 | 20 | 0.375883502 | 1.078538887 | 2078.577155 |

* 10 features

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **k** | **init** | **max\_iter** | **n\_init** | **Silhouette Score** | **Davies-Bouldin Score** | **Calinski-Harabasz Score** |
| 3 | k-means++ | 100 | 10 | 0.267291543 | 1.197049164 | 1550.938594 |
| 100 | 20 | 0.267294804 | 1.196586174 | 1550.938912 |
| 200 | 10 | 0.267291543 | 1.197049164 | 1550.938594 |
| 200 | 20 | 0.267294804 | 1.196586174 | 1550.938912 |
| 300 | 10 | 0.267291543 | 1.197049164 | 1550.938594 |
| 300 | 20 | 0.267294804 | 1.196586174 | 1550.938912 |
| random | 100 | 10 | 0.267294804 | 1.196586174 | 1550.938912 |
| 100 | 20 | 0.267294804 | 1.196586174 | 1550.938912 |
| 200 | 10 | 0.267294804 | 1.196586174 | 1550.938912 |
| 200 | 20 | 0.267294804 | 1.196586174 | 1550.938912 |
| 300 | 10 | 0.267294804 | 1.196586174 | 1550.938912 |
| 300 | 20 | 0.267294804 | 1.196586174 | 1550.938912 |
| 4 | k-means++ | 100 | 10 | 0.295033279 | 1.06001582 | 1449.622142 |
| 100 | 20 | 0.295033279 | 1.06001582 | 1449.622142 |
| 200 | 10 | 0.295033279 | 1.06001582 | 1449.622142 |
| 200 | 20 | 0.295033279 | 1.06001582 | 1449.622142 |
| 300 | 10 | 0.295033279 | 1.06001582 | 1449.622142 |
| 300 | 20 | 0.295033279 | 1.06001582 | 1449.622142 |
| random | 100 | 10 | 0.295121878 | 1.059851674 | 1449.614228 |
| 100 | 20 | 0.295033279 | 1.06001582 | 1449.622142 |
| 200 | 10 | 0.295121878 | 1.059851674 | 1449.614228 |
| 200 | 20 | 0.295033279 | 1.06001582 | 1449.622142 |
| 300 | 10 | 0.295121878 | 1.059851674 | 1449.614228 |
| 300 | 20 | 0.295033279 | 1.06001582 | 1449.622142 |
| 5 | k-means++ | 100 | 10 | 0.298244318 | 1.134733875 | 1480.510752 |
| 100 | 20 | 0.298244318 | 1.134733875 | 1480.510752 |
| 200 | 10 | 0.298244318 | 1.134733875 | 1480.510752 |
| 200 | 20 | 0.298244318 | 1.134733875 | 1480.510752 |
| 300 | 10 | 0.298244318 | 1.134733875 | 1480.510752 |
| 300 | 20 | 0.298244318 | 1.134733875 | 1480.510752 |
| random | 100 | 10 | 0.297881627 | 1.139591948 | 1480.299242 |
| 100 | 20 | 0.297881627 | 1.139591948 | 1480.299242 |
| 200 | 10 | 0.297881627 | 1.139591948 | 1480.299242 |
| 200 | 20 | 0.297881627 | 1.139591948 | 1480.299242 |
| 300 | 10 | 0.297881627 | 1.139591948 | 1480.299242 |
| 300 | 20 | 0.297881627 | 1.139591948 | 1480.299242 |

- Xóa phần giao dịch bị hủy

* 3 features

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **k** | **init** | **max\_iter** | **n\_init** | **Silhouette Score** | **Davies-Bouldin Score** | **Calinski-Harabasz Score** |
| 3 | k-means++ | 100 | 10 | 0.484495937 | 0.728612548 | 4576.935005 |
| 100 | 20 | 0.484495937 | 0.728612548 | 4576.935005 |
| 300 | 10 | 0.484495937 | 0.728612548 | 4576.935005 |
| 300 | 20 | 0.484495937 | 0.728612548 | 4576.935005 |
| random | 100 | 10 | 0.48442864 | 0.728707714 | 4576.932592 |
| 100 | 20 | 0.48442864 | 0.728707714 | 4576.932592 |
| 300 | 10 | 0.48442864 | 0.728707714 | 4576.932592 |
| 300 | 20 | 0.48442864 | 0.728707714 | 4576.932592 |
| 4 | k-means++ | 100 | 10 | 0.456494041 | 0.817335565 | 4618.161553 |
| 100 | 20 | 0.456494041 | 0.817335565 | 4618.161553 |
| 300 | 10 | 0.456494041 | 0.817335565 | 4618.161553 |
| 300 | 20 | 0.456494041 | 0.817335565 | 4618.161553 |
| random | 100 | 10 | 0.456494041 | 0.817335565 | 4618.161553 |
| 100 | 20 | 0.456509198 | 0.817343857 | 4618.160282 |
| 300 | 10 | 0.456494041 | 0.817335565 | 4618.161553 |
| 300 | 20 | 0.456509198 | 0.817343857 | 4618.160282 |
| 5 | k-means++ | 100 | 10 | 0.397839555 | 0.87429291 | 4186.636628 |
| 100 | 20 | 0.397831468 | 0.873626826 | 4186.664252 |
| 300 | 10 | 0.397839555 | 0.87429291 | 4186.636628 |
| 300 | 20 | 0.397831468 | 0.873626826 | 4186.664252 |
| random | 100 | 10 | 0.391051231 | 0.947972486 | 4186.831123 |
| 100 | 20 | 0.391051231 | 0.947972486 | 4186.831123 |
| 300 | 10 | 0.391051231 | 0.947972486 | 4186.831123 |
| 300 | 20 | 0.391051231 | 0.947972486 | 4186.831123 |

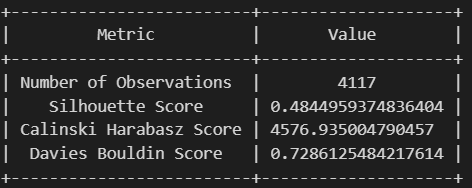
* 8 features

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **k** | **init** | **max\_iter** | **n\_init** | **Silhouette Score** | **Davies-Bouldin Score** | **Calinski-Harabasz Score** |
| 3 | k-means++ | 100 | 10 | 0.267291543 | 1.197049164 | 1550.938594 |
| 100 | 20 | 0.267294804 | 1.196586174 | 1550.938912 |
| 200 | 10 | 0.267291543 | 1.197049164 | 1550.938594 |
| 200 | 20 | 0.267294804 | 1.196586174 | 1550.938912 |
| 300 | 10 | 0.267291543 | 1.197049164 | 1550.938594 |
| 300 | 20 | 0.267294804 | 1.196586174 | 1550.938912 |
| random | 100 | 10 | 0.267294804 | 1.196586174 | 1550.938912 |
| 100 | 20 | 0.267294804 | 1.196586174 | 1550.938912 |
| 200 | 10 | 0.267294804 | 1.196586174 | 1550.938912 |
| 200 | 20 | 0.267294804 | 1.196586174 | 1550.938912 |
| 300 | 10 | 0.267294804 | 1.196586174 | 1550.938912 |
| 300 | 20 | 0.267294804 | 1.196586174 | 1550.938912 |
| 4 | k-means++ | 100 | 10 | 0.295033279 | 1.06001582 | 1449.622142 |
| 100 | 20 | 0.295033279 | 1.06001582 | 1449.622142 |
| 200 | 10 | 0.295033279 | 1.06001582 | 1449.622142 |
| 200 | 20 | 0.295033279 | 1.06001582 | 1449.622142 |
| 300 | 10 | 0.295033279 | 1.06001582 | 1449.622142 |
| 300 | 20 | 0.295033279 | 1.06001582 | 1449.622142 |
| random | 100 | 10 | 0.295121878 | 1.059851674 | 1449.614228 |
| 100 | 20 | 0.295033279 | 1.06001582 | 1449.622142 |
| 200 | 10 | 0.295121878 | 1.059851674 | 1449.614228 |
| 200 | 20 | 0.295033279 | 1.06001582 | 1449.622142 |
| 300 | 10 | 0.295121878 | 1.059851674 | 1449.614228 |
| 300 | 20 | 0.295033279 | 1.06001582 | 1449.622142 |
| 5 | k-means++ | 100 | 10 | 0.298244318 | 1.134733875 | 1480.510752 |
| 100 | 20 | 0.298244318 | 1.134733875 | 1480.510752 |
| 200 | 10 | 0.298244318 | 1.134733875 | 1480.510752 |
| 200 | 20 | 0.298244318 | 1.134733875 | 1480.510752 |
| 300 | 10 | 0.298244318 | 1.134733875 | 1480.510752 |
| 300 | 20 | 0.298244318 | 1.134733875 | 1480.510752 |
| random | 100 | 10 | 0.297881627 | 1.139591948 | 1480.299242 |
| 100 | 20 | 0.297881627 | 1.139591948 | 1480.299242 |
| 200 | 10 | 0.297881627 | 1.139591948 | 1480.299242 |
| 200 | 20 | 0.297881627 | 1.139591948 | 1480.299242 |
| 300 | 10 | 0.297881627 | 1.139591948 | 1480.299242 |
| 300 | 20 | 0.297881627 | 1.139591948 | 1480.299242 |

### 3. Phương pháp đánh giá

- Để đánh giá hiệu quả của mô hình phân cụm khách hàng dựa trên chỉ số RFM (Recency, Frequency, Monetary), nghiên cứu sử dụng hai nhóm tiêu chí chính: các chỉ số đánh giá hình học của phân cụm và mức độ diễn giải và tính thực tiễn của các cụm được tạo ra.

#### 3.1. Đánh giá hình học

****

- Các chỉ số hình học được sử dụng nhằm đo lường mức độ phân tách và đồng nhất của các cụm bao gồm:

* Silhouette Score (Chỉ số đường viền): Đo lường mức độ gần nhau giữa các điểm dữ liệu trong cùng một cụm và mức độ khác biệt với các cụm còn lại. Giá trị dao động từ -1 đến 1, trong đó giá trị gần 1 thể hiện phân cụm tốt. Trong nghiên cứu này, Silhouette Score đạt 0.484, cho thấy các cụm có sự phân tách tương đối rõ ràng.
* Calinski-Harabasz Score (Chỉ số phân tán giữa cụm): Là tỷ lệ giữa độ phân tán giữa các cụm và độ phân tán bên trong cụm. Giá trị càng cao cho thấy mô hình phân cụm càng hiệu quả. Mô hình thu được Calinski-Harabasz Score là 4576.92 – một giá trị cao, cho thấy cụm được hình thành rõ ràng.
* Davies-Bouldin Score (Chỉ số tương đồng cụm): Đo lường mức độ tương đồng giữa các cụm, trong đó giá trị càng thấp càng tốt. Với Davies-Bouldin Score đạt 0.728, mô hình cho thấy các cụm tương đối khác biệt và có độ đồng nhất bên trong cụm tốt.

- Những chỉ số trên giúp xác nhận rằng mô hình phân cụm đã đạt được hiệu quả về mặt toán học trong việc phân chia khách hàng thành các nhóm có đặc điểm tương đồng.

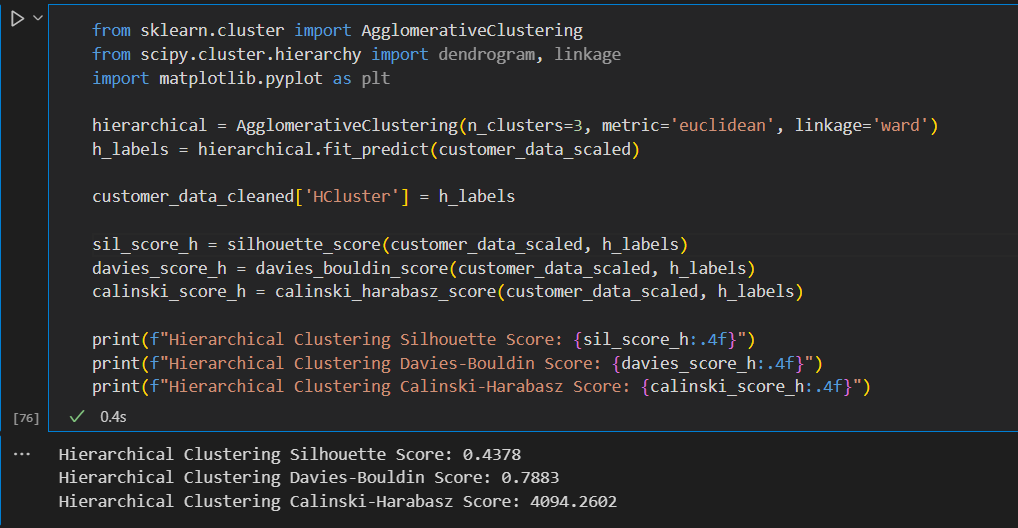
#### 3.2. Đánh giá ý nghĩa và khả năng giải thích cụm

Bên cạnh đánh giá toán học, một yếu tố quan trọng trong bài toán phân cụm là mức độ giải thích được và khả năng ứng dụng thực tiễn của các cụm. Dựa trên giá trị trung bình của các chỉ số RFM trong từng cụm, ta có:

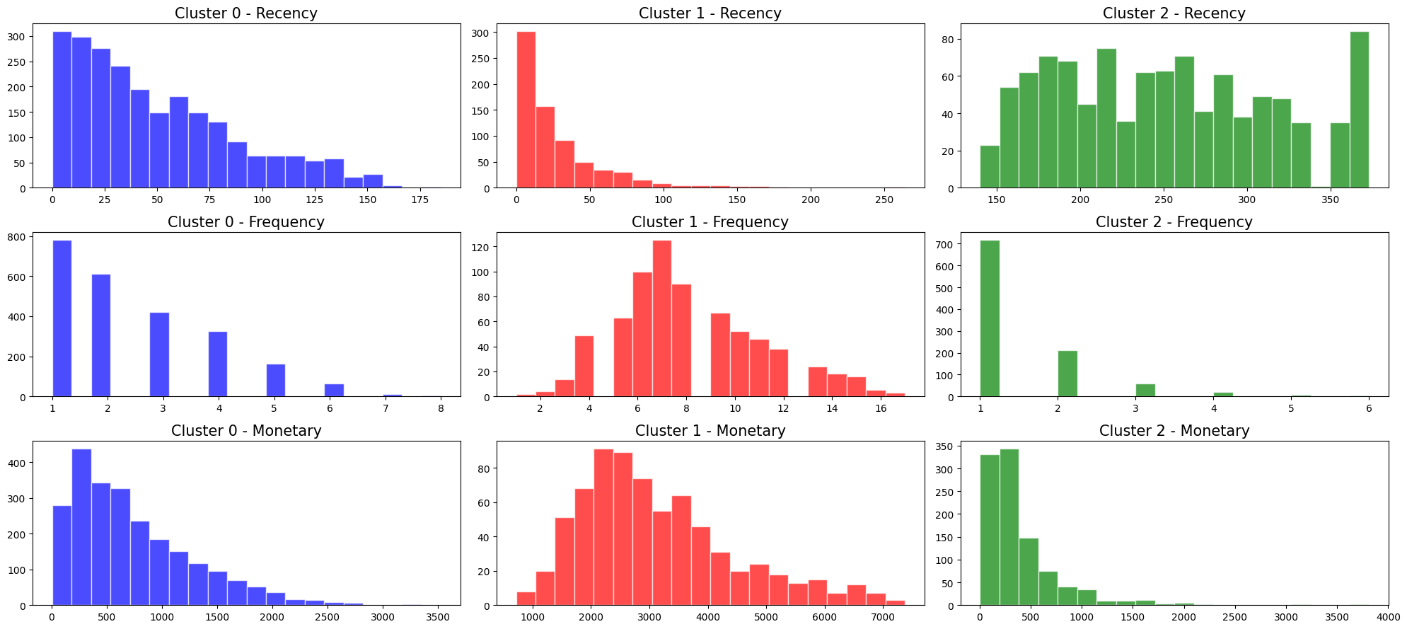
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cụm** | **Recency** | **Frequency** | **Monetary** | **Diễn giải** |
| 0 | 49.57 | 2.47 | 738.32 | Khách mua gần đây, mua ít, chi tiêu trung bình – nhóm khách hàng phổ thông |
| 1 | 28.50 | 8.07 | 3133.57 | Khách hàng mua thường xuyên, gần đây và có giá trị cao – **khách hàng VIP** cần được ưu đãi và giữ chân |
| 2 | 251.49 | 1.44 | 410.44 | Khách hàng lâu không quay lại, ít mua, chi tiêu thấp – **khách hàng không còn hoạt động**, cần chiến lược tái kích hoạt |

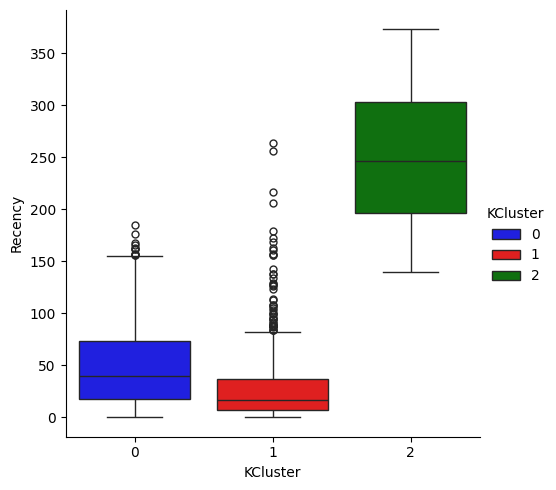
- Việc phân cụm như vậy không chỉ phản ánh sự khác biệt rõ ràng giữa các nhóm khách hàng trên các chỉ số RFM, mà còn cung cấp các gợi ý thiết thực cho chiến lược chăm sóc khách hàng, marketing cá nhân hóa và tối ưu doanh thu.

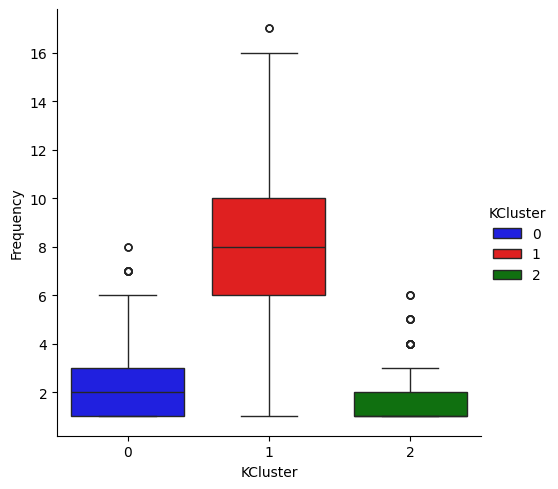
- Hierarchical Clustering

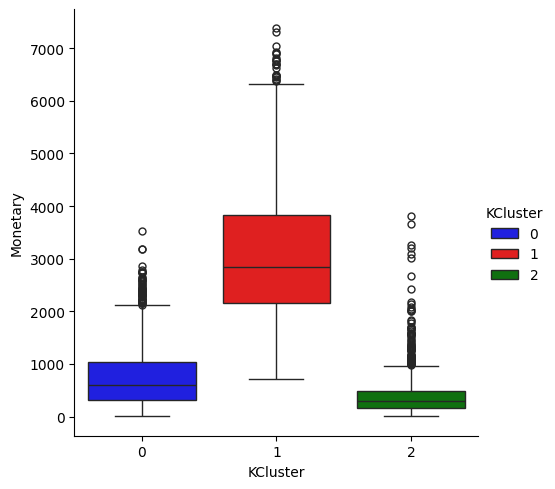


### 4. Kết luận









#### 4.1. Trường hợp tốt nhất & ứng dụng

- Cluster 1 (Khách hàng VIP) là nhóm tốt nhất để tập trung nguồn lực.

- Sử dụng trong trường hợp:

* Triển khai chiến dịch giữ chân khách hàng trung thành.
* Phát triển sản phẩm/dịch vụ cao cấp nhắm đến nhóm có giá trị chi tiêu cao.
* Ví dụ: Ưu đãi "Mua 5 lần được 1 lần miễn phí" hoặc gói membership VIP.

#### 4.2. Lý giải hiệu quả của Cluster 1

- Cluster 1 đạt hiệu quả cao do kết hợp tối ưu 3 chỉ số RFM:

* **Recency thấp** (28.5 ngày): Khách hàng mua hàng gần đây → Có xu hướng quay lại nhanh.
* **Frequency cao** (8.07 lần): Tương tác thường xuyên → Độ trung thành và gắn bó cao.
* **Monetary cao** (3,133.57): Đóng góp 80-90% doanh thu (giả định) → Nhóm sinh lời chính.

- Cơ chế khoa học:

* **Phương pháp Elbow**: Chọn K=3 giúp phân nhóm rõ ràng, tránh overlap dữ liệu.
* **Tối ưu hóa mục tiêu kinh doanh**: Tập trung vào nhóm mang lại lợi nhuận cao nhất (Cluster 1) thay vì dàn trải nguồn lực.

#### 4.3. Ví dụ minh họa với 3 features RFM

- Scenario: Một khách hàng thuộc Cluster 1 có:

* Recency = 28 ngày: Lần cuối mua hàng cách đây 28 ngày → Gần đây.
* Frequency = 8 lần: Mua hàng 8 lần trong 3 tháng → Thường xuyên.
* Monetary = 3,133.57: Tổng chi tiêu 3,133.57 đơn vị tiền tệ → Giá trị cao.

- Lý do hiệu quả:

* Recency thấp → Khả năng phản hồi với ưu đãi nhanh (ví dụ: email giảm 15% có tỷ lệ mở 40%).
* Frequency cao → Dễ dàng dự đoán hành vi mua tiếp theo → Tối ưu thời gian chạy chiến dịch.
* Monetary cao → Tăng ROI khi đầu tư vào nhóm này (ví dụ: 1 đồng marketing thu về 5 đồng doanh thu).

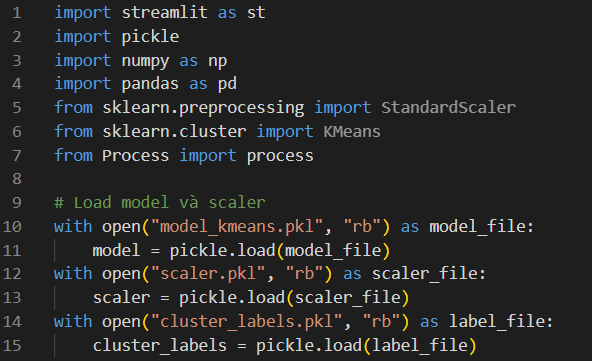
### 5. Xuất mô hình tốt nhất

- Trong số các trường hợp ở trên, trường hợp 3 feature cho kết quả tốt nhất phù hợp với ngữ cảnh sẽ được xuất ra và load vào hệ thống cho người sử dụng có thể thao tác được.

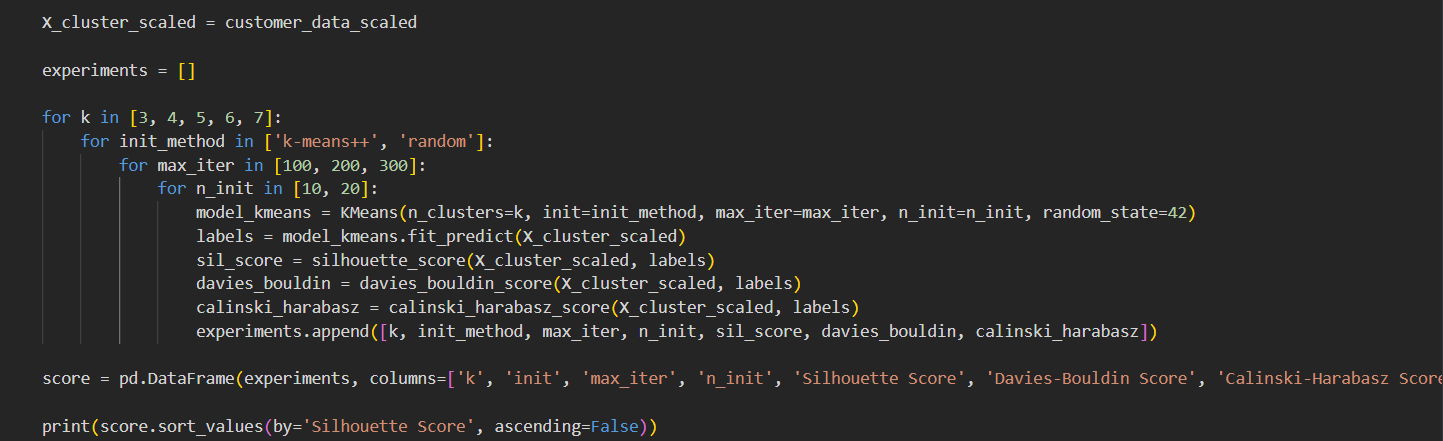
- Xuất mô hình:



- Load mô hình:



### 6. Viết một đoạn code để có thể thực thi các trường hợp thử này



## VI. XÂY DỰNG GIAO DIỆN NGƯỜI SỬ DỤNG

### 1. Xác định đầu vào

- Hệ thống được xây dựng hướng đến phục vụ các đối tượng có liên quan trực tiếp đến việc quản lý và khai thác dữ liệu khách hàng trong doanh nghiệp, cụ thể:

* Nhân viên Marketing: là người trực tiếp sử dụng kết quả phân nhóm để triển khai các chiến dịch khuyến mãi, chăm sóc khách hàng, hoặc cá nhân hóa trải nghiệm mua sắm.
* Quản lý kinh doanh: sử dụng hệ thống để giám sát hoạt động phân khúc khách hàng, từ đó đưa ra các quyết định chiến lược phù hợp với từng nhóm.
* Chuyên viên phân tích dữ liệu (Data Analyst): có khả năng sử dụng hệ thống để khám phá hành vi tiêu dùng, khai thác đặc điểm của từng nhóm khách hàng phục vụ mục tiêu tư vấn hoặc tối ưu hóa hiệu quả kinh doanh.

#### 1.1. Đầu vào tương ứng với phương thức nhập liệu thủ công

- Trong trường hợp người dùng muốn kiểm tra riêng lẻ từng khách hàng, hệ thống cho phép nhập trực tiếp vào giao diện các trường thông tin sau:

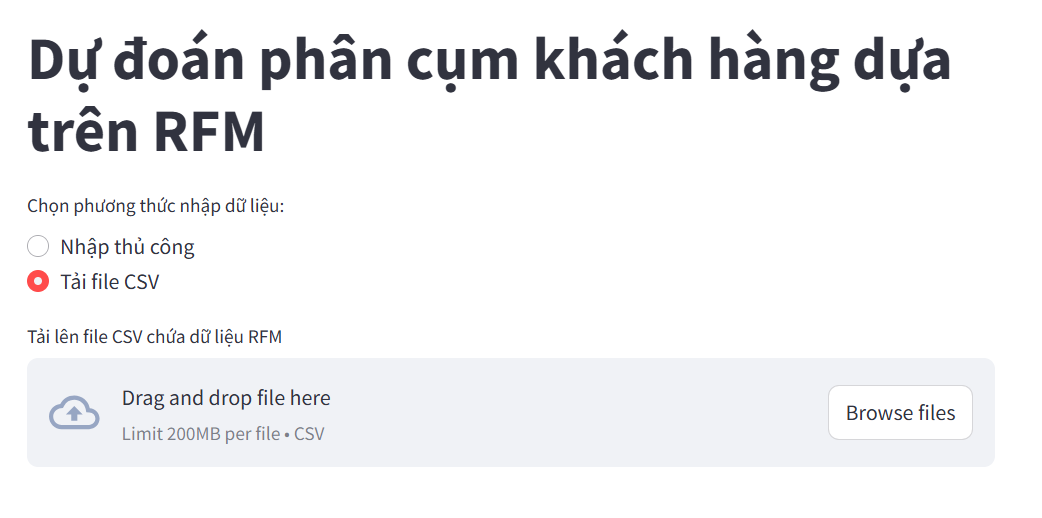
* Ngày mua hàng gần nhất: Ngày gần nhất giao dịch tại cửa hàng.
* Frequency: tổng số lần mua hàng.
* Monetary: tổng giá trị chi tiêu.

- Việc nhập tay từng bản ghi phù hợp với các tình huống như:

* Kiểm tra nhanh trạng thái một khách hàng cụ thể.
* Dùng thử hoặc demo hệ thống trong các cuộc họp nhỏ.



#### 1.2. Đầu vào tương ứng với phương thức tải lên tệp dữ liệu (CSV)



- Đối với người dùng muốn phân tích nhiều khách hàng cùng lúc, hệ thống cho phép tải lên file CSV, trong đó mỗi dòng chứa thông tin một khách hàng. Cấu trúc file gồm 8 cột:

* InvoiceNo
* StockCode
* Description
* Quantity
* InvoiceDate
* UnitPrice
* CustomerID
* Country

- Đây là phương thức nhập dữ liệu phù hợp với:

* Các chiến dịch marketing quy mô lớn.
* Phân tích dữ liệu định kỳ hàng tuần, hàng tháng.
* Tự động hóa hệ thống CRM.

### 2. Xác định đầu ra

- Trong hệ thống Phân loại khách hàng dựa trên mô hình RFM này, đầu ra được xác định là tập hợp các kết luận về nhóm khách hàng, được biểu diễn thông qua giao diện người dùng trên nền tảng Web. Hệ thống hỗ trợ hai phương thức nhập dữ liệu khác nhau, tương ứng với đó là hai dạng đầu ra chính:

#### 2.1. Đầu ra tương ứng với phương thức nhập liệu thủ công



- Khi người dùng lựa chọn hình thức nhập liệu thủ công, hệ thống cho phép nhập trực tiếp các giá trị đặc trưng cho từng khách hàng. Sau khi dữ liệu được nhập và người dùng nhấn nút “Dự đoán”, hệ thống sẽ tiến hành xử lý và hiển thị một kết luận ngắn gọn, ví dụ: “Khách hàng thuộc nhóm: Vip”.

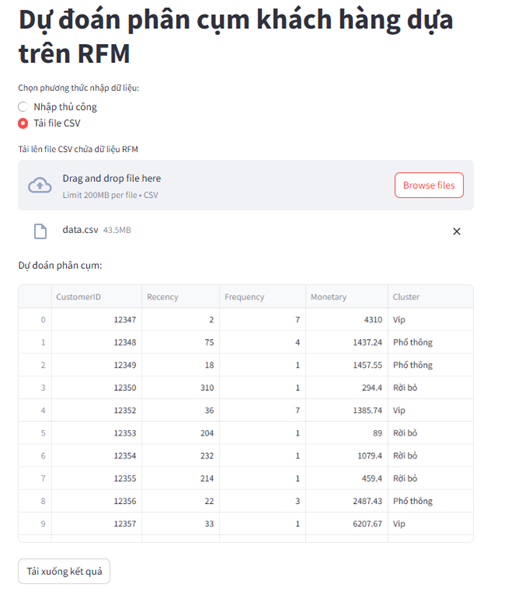
- Đây là dạng đầu ra phù hợp cho các trường hợp cần kiểm tra riêng lẻ hoặc đánh giá nhanh một khách hàng cụ thể.

#### 2.2. Đầu ra tương ứng với phương thức tải lên tệp dữ liệu (CSV)

- Trong trường hợp người dùng lựa chọn tải lên một tệp dữ liệu định dạng CSV chứa thông tin của nhiều khách hàng, hệ thống sẽ thực hiện phân cụm cho toàn bộ tập dữ liệu và xuất ra một bảng kết quả tổng hợp. Mỗi hàng trong bảng thể hiện một khách hàng với các thông tin bao gồm:

* Mã khách hàng (CustomerID)
* Recency (số ngày kể từ lần mua gần nhất)
* Frequency (số lần mua hàng)
* Monetary (tổng giá trị chi tiêu)
* Nhóm phân cụm (Cluster)

- Đây là dạng đầu ra bảng biểu tổng hợp, hỗ trợ người dùng trong việc phân tích tập trung và đánh giá toàn diện tập khách hàng dựa trên các đặc trưng hành vi mua sắm.



# LỜI KẾT

Sau quá trình tìm hiểu và thực hiện, nhóm 3 đã vận dụng thành công một số thuật toán trí tuệ nhân tạo vào việc phân tích và phân khúc khách hàng từ tập dữ liệu bán lẻ trực tuyến của UCI. Thông qua việc tiền xử lý dữ liệu, lựa chọn mô hình phù hợp và đánh giá kết quả, đề tài đã bước đầu chứng minh được hiệu quả của AI trong việc hỗ trợ doanh nghiệp hiểu rõ hơn về khách hàng mục tiêu.

Mặc dù còn một số hạn chế nhất định về phạm vi và độ sâu phân tích, song đề tài là bước đầu quan trọng giúp nhóm tiếp cận thực tế ứng dụng AI trong kinh doanh đồng thời củng cố kiến thức đã học trong môn học.

Nhóm 3 xin trân trọng cảm ơn thầy Phạm Thảo đã tạo điều kiện cho chúng em có cơ hội nghiên cứu, học tập và tiếp cận những kiến thức thực tiễn giá trị. Rất mong nhận được sự đóng góp ý kiến từ thầy để đề tài được hoàn thiện hơn!