**MỤC LỤC**

[CHƯƠNG 1: TỔNG QUAN 1](#_Toc149502416)

[1. Giới thiệu đề tài: 1](#_Toc149502417)

[2. Mục tiêu đề tài: 1](#_Toc149502418)

[3. Giải pháp thực hiện: 1](#_Toc149502419)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 2](#_Toc149502420)

[1. Bài toán phân cụm khách hàng 2](#_Toc149502421)

[1.1. Định nghĩa: 2](#_Toc149502422)

[1.2. Ứng dụng: 2](#_Toc149502423)

[2. Áp dụng K-means vào bài toán phân cụm khách hàng: 3](#_Toc149502424)

[3. Các phương pháp sử dụng: 5](#_Toc149502425)

[3.1. Xử lý dữ liệu bằng KNNImputer: 5](#_Toc149502426)

[3.2. Chuẩn hóa dữ liệu bằng StandardScaler: 7](#_Toc149502427)

[3.3. Tìm số cụm K tốt nhất bằng Elbow: 8](#_Toc149502428)

[CHƯƠNG 3: BỘ DỮ LIỆU 10](#_Toc149502429)

[1. Thu thập bộ dữ liệu: 10](#_Toc149502430)

[2. Mô tả bộ dữ liệu: 10](#_Toc149502431)

[CHƯƠNG 4: THỰC NGHIỆM 12](#_Toc149502432)

[1. Tiền xử lý dữ liệu: 12](#_Toc149502433)

[2. Phân tích thăm dò (EDA) 13](#_Toc149502434)

[2.1. Phân tích hạn mức và số dư dựa trên kỳ hạn 13](#_Toc149502435)

[2.2. Phân tích số lượng mua so với tổng giao dịch 13](#_Toc149502436)

[2.3. Phân tích sự tương quan của các thuộc tính: 14](#_Toc149502437)

[3. Chuẩn hóa dữ liệu: 15](#_Toc149502438)

[4. Xây dựng mô hình: 16](#_Toc149502439)

[4.1. Lựa chọn số cụm 16](#_Toc149502440)

[4.2. Xây dựng mô hình K-mean: 17](#_Toc149502441)

[5. Kết quả và đánh giá: 18](#_Toc149502442)

[CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 23](#_Toc149502443)

[TÀI LIỆU THAM KHẢO 24](#_Toc149502444)

MỤC LỤC HÌNH ẢNH

[Hình 1. Ý tưởng bài toán phân cụm khách hàng 3](#_Toc149503012)

[Hình 2. Xác định tâm cụm. 4](#_Toc149503012)

[Hình 3. Total Within-Cluster Variation. 5](#_Toc149503013)

[Hình 4. Xử lý dữ liệu bằng KNNImputer. 6](#_Toc149503014)

[Hình 5. Phương pháp khuỷu tay Elbow. 9](#_Toc149503015)

[Hình 6. Thống kê dữ liệu. 12](#_Toc149503016)

[Hình 7. Biểu đồ Scatter phân tích hạn mức và số dư dựa trên kỳ hạn. 13](#_Toc149503017)

[Hình 8. Biểu đồ dumbbell phân tích số lượng mua so với tổng giao dịch. 14](#_Toc149503018)

[Hình 9. Ma trận tương quan giữa các thuộc tính. 15](#_Toc149503019)

[Hình 10. Biểu đồ xác định số cụm bằng phương pháp Elbow. 17](#_Toc149503020)

[Hình 11. Một số biểu đồ trực quan kết quả của mô hình. 18](#_Toc149503021)

[Hình 12. Insight của kết quả phân cụm. 19](#_Toc149503022)

[Hình 13. Phân tích kết quả của mô hình. 21](#_Toc149503023)

# CHƯƠNG 1: TỔNG QUAN

## Giới thiệu đề tài:

Đề tài nghiên cứu này tập trung vào việc phân cụm khách hàng và dự đoán nhóm khách hàng tiềm năng cho chiến dịch tiếp thị. Trong môi trường kinh doanh ngày nay, việc hiểu rõ hành vi của khách hàng và xác định những khách hàng có tiềm năng là một phần quan trọng của chiến lược tiếp thị hiệu quả. Đề tài này sẽ sử dụng phân cụm khách hàng để nhận biết các nhóm khách hàng có đặc điểm chung và sau đó dự đoán những khách hàng có tiềm năng nhất để tiếp cận trong chiến dịch tiếp thị.

## Mục tiêu đề tài:

* Phân cụm khách hàng dựa trên các biến số quan trọng về hành vi sử dụng thẻ tín dụng, như số lần sử dụng thẻ, hạn mức tín dụng, số lần rút tiền mặt trước, v.v.
* Dự đoán nhóm khách hàng tiềm năng cho chiến dịch tiếp thị bằng việc sử dụng mô hình học máy hoặc các phương pháp dự đoán tương tự. Điều này sẽ giúp doanh nghiệp xác định đối tượng tiềm năng nhất để tập trung tiếp thị và tối ưu hóa chiến dịch tiếp thị của họ.

## Giải pháp thực hiện:

Để đạt được mục tiêu của đề tài, chúng tôi sẽ sử dụng các phương pháp phân cụm dữ liệu để chia khách hàng thành các nhóm dựa trên đặc điểm chung về hành vi sử dụng thẻ tín dụng. Sau đó, chúng tôi sẽ sử dụng mô hình học máy hoặc các kỹ thuật dự đoán để dự đoán nhóm khách hàng có tiềm năng cho chiến dịch tiếp thị. Điều này sẽ giúp doanh nghiệp tập trung tiếp thị một cách hiệu quả và tối ưu hóa việc tiếp cận đối tượng mục tiêu.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## Bài toán phân cụm khách hàng

### Định nghĩa:

Bài toán phân cụm khách hàng là một phần quan trọng của lĩnh vực quản lý khách hàng và tiếp thị, có sự ứng dụng rộng rãi trong nhiều ngành như bán lẻ, dịch vụ tài chính, thương mại điện tử, và nhiều lĩnh vực khác. Chia khách hàng thành các nhóm có đặc điểm tương tự giúp doanh nghiệp xác định mô hình hành vi của khách hàng và hiểu rõ hơn về họ.

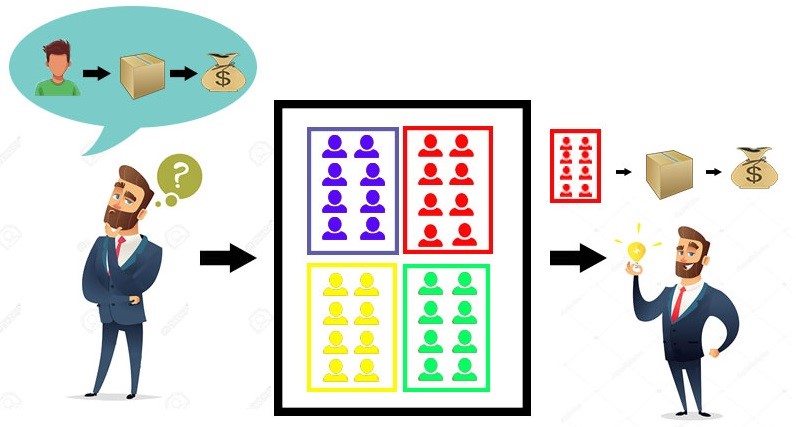
Một điểm quan trọng cần lưu ý là không có một cách tiếp cận phân cụm khách hàng duy nhất hoặc "kích cỡ giày vừa" cho mọi tình huống. Phương pháp phân cụm cần được tùy chỉnh cho từng dự án cụ thể dựa trên mục tiêu cụ thể của doanh nghiệp và dữ liệu khách hàng cụ thể.

### Ứng dụng:

Ứng dụng của bài toán phân cụm khách hàng:

* **Phân loại Khách Hàng:** Bài toán phân cụm khách hàng có thể giúp doanh nghiệp phân loại khách hàng thành các nhóm như "khách hàng trung thành," "khách hàng tiềm năng," và "khách hàng chưa quen biết." Dựa vào việc phân loại này, doanh nghiệp có thể xây dựng chiến lược tương ứng.
* **Tối Ưu Hóa Chiến Dịch Tiếp Thị:** Bằng cách hiểu rõ nhóm khách hàng cụ thể, doanh nghiệp có thể tối ưu hóa chiến dịch tiếp thị để đáp ứng nhu cầu của từng nhóm. Điều này bao gồm cả việc tạo nội dung tiếp thị, lựa chọn kênh tiếp thị, và xác định thời điểm thích hợp.
* **Xác Định Lợi Nhuận và Rủi Ro:** Bài toán phân cụm khách hàng có thể giúp xác định những khách hàng có tiềm năng lớn và giá trị cao cho doanh nghiệp. Nó cũng có thể giúp xác định những khách hàng có rủi ro cao, chẳng hạn như khách hàng có nguy cơ chuyển đổi sang đối thủ cạnh tranh.
* **Tạo Sản Phẩm và Dịch Vụ Tùy Chỉnh:** Hiểu rõ nhu cầu cụ thể của từng nhóm khách hàng giúp doanh nghiệp tạo ra sản phẩm và dịch vụ tùy chỉnh để đáp ứng nhu cầu của họ. Điều này giúp tạo sự hài lòng và trung thực từ phía khách hàng.
* **Quản Lý Khách Hàng:** Phân cụm khách hàng cũng có thể giúp trong việc quản lý khách hàng. Điều này bao gồm việc xác định cách tương tác với từng nhóm khách hàng và đảm bảo rằng họ nhận được sự chăm sóc phù hợp.

Trong quá trình giải quyết bài toán phân cụm khách hàng, việc áp dụng các phương pháp xử lý dữ liệu và lựa chọn số lượng cụm tối ưu đóng vai trò quan trọng để đảm bảo tính hiệu quả của quy trình phân cụm.



Hình 1. Ý tưởng bài toán phân cụm khách hàng.

## Áp dụng K-means vào bài toán phân cụm khách hàng:

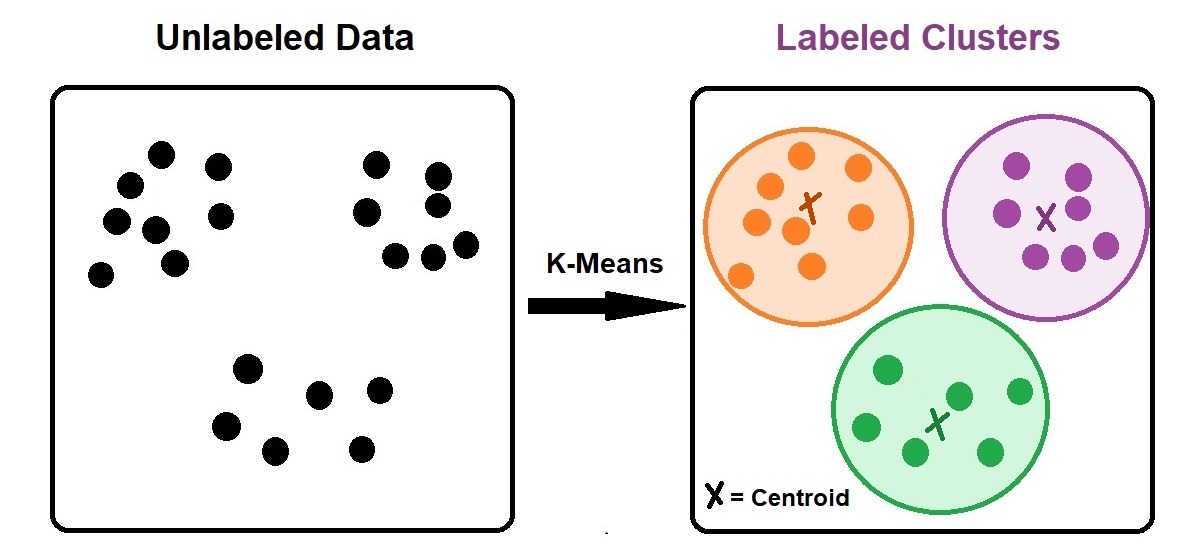
Đây là một phương pháp phổ biến để thực hiện việc phân loại và tạo ra các nhóm khách hàng dựa trên đặc điểm và hành vi của họ. K-means là một thuật toán phân cụm dựa trên tâm cụm, nó hoạt động theo các bước cụ thể để tạo ra các nhóm cụm. Dưới đây là một số khái niệm và công thức quan trọng liên quan đến K-means:

**Tâm Cụm (Centroid):**

Trong K-means, mỗi cụm được đại diện bởi một điểm tâm cụm, là trung bình của tất cả các điểm dữ liệu trong cụm đó. Tâm cụm thường được ký hiệu bằng và được tính bằng công thức:

Trong đó:

* là tâm cụm.
* là số lượng điểm dữ liệu trong cụm.
* là điểm dữ liệu trong cụm.



Hình 2. Xác định tâm cụm.

**Khoảng cách Euclidean:**

Để đo khoảng cách giữa hai điểm dữ liệu và , K-means thường sử dụng khoảng cách Euclidean, được tính bằng công thức:

Trong đó:

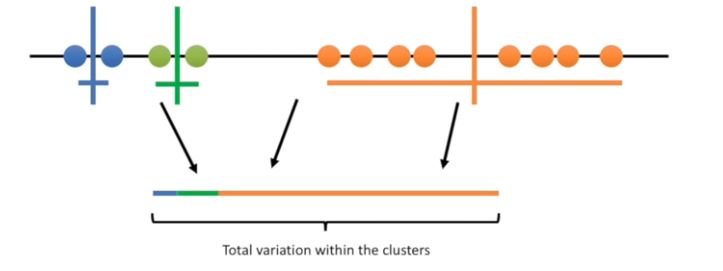
* là khoảng cách giữa điểm và
* là số chiều của dữ liệu (số lượng đặc trưng).

**Sự Khác Biệt Tổng Cộng (Total Within-Cluster Variation):**

K-means cố gắng tối thiểu hóa sự khác biệt tổng cộng, tức là tổng khoảng cách từ mỗi điểm dữ liệu đến tâm cụm của nó. Sự khác biệt tổng cộng được tính bằng công thức:

Trong đó:

* là số lượng cụm.
* là tập hợp các điểm dữ liệu trong cụm thứ .
* là tâm cụm của cụm thứ .



Hình 3. Total Within-Cluster Variation.

**Lựa Chọn Số Lượng Cụm (K) Tối Ưu:**

Một trong những thách thức quan trọng của K-means là lựa chọn số lượng cụm tối ưu (K). Công thức Elbow là một phương pháp thường được sử dụng để xác định K tối ưu. Nó đo lường sự giảm đột ngột trong sự khác biệt tổng cộng khi K tăng lên.

## Các phương pháp sử dụng:

### Xử lý dữ liệu bằng KNNImputer:

KNNImputer là một phương pháp xử lý dữ liệu được sử dụng để điền các giá trị bị thiếu hoặc còn gọi là giá trị NaN trong tập dữ liệu. Thuật toán này hoạt động bằng cách tìm kiếm các điểm dữ liệu tương tự nhất với điểm cần điền và sử dụng giá trị của các điểm tương tự để điền vào giá trị cần thiết. Trong bài toán phân cụm khách hàng, KNNImputer có thể được sử dụng để xử lý các giá trị bị thiếu trong các biến quan trọng như hạn mức tín dụng hoặc số lần rút tiền mặt trước.

Dữ liệu bị thiếu xảy ra khi một hoặc nhiều giá trị trong hàng dữ liệu bị trống hoặc không có giá trị. Nguyên nhân có thể là do lỗi ghi nhận, không thu thập thông tin, hoặc nhiều lý do khác. Xử lý dữ liệu bị thiếu là một phần quan trọng trong quá trình tiền xử lý dữ liệu, bởi vì nó có thể ảnh hưởng đáng kể đến kết quả phân tích và mô hình hóa.

A diagram of a imputation

Description automatically generated

Hình 4. Xử lý dữ liệu bằng KNNImputer.

KNNImputer là một phương pháp dựa trên nguyên tắc "KNN." Ý tưởng chính là sử dụng thông tin từ các hàng (hoặc điểm dữ liệu) gần nhất để ước tính giá trị bị thiếu. Để thực hiện điều này, các bước cơ bản bao gồm:

* **Xác định khoảng cách giữa các hàng dữ liệu:** Thường sử dụng khoảng cách Euclidean để đo sự tương tự giữa các hàng.
* **Chọn giá trị K:** Đây là số lượng hàng gần nhất mà chúng ta muốn xem xét. Lựa chọn giá trị K đòi hỏi sự cân nhắc, vì giá trị K quá nhỏ có thể dẫn đến dự đoán không ổn định, trong khi giá trị K quá lớn có thể làm mất đi tính cục bộ của các ước tính.
* **Tìm các hàng gần nhất:** Sử dụng khoảng cách đã xác định, tìm ra K hàng gần nhất mà không có giá trị bị thiếu ở biến cần ước tính.
* **Ước Tính Giá Trị:** ƨử dụng thông tin từ các hàng gần nhất, ước tính giá trị bị thiếu bằng một trung bình có trọng số của các hàng gần nhất.

Công thức cơ bản cho KNNImputer để ước tính giá trị bị thiếu là:

Trong đó:

* là giá trị bị thiếu cần ước tính.
* là số lượng hàng gần nhất (lựa chọn bằng).
* là giá trị của biến tương ứng trong hàng gần nhất thứ .

Một trong những thách thức quan trọng của việc sử dụng KNNImputer là lựa chọn giá trị K tối ưu. Lựa chọn K phụ thuộc vào từng tình huống cụ thể và đôi khi đòi hỏi sự thử nghiệm. Giá trị K tùy thuộc vào tính chất của dữ liệu và mục tiêu của mô hình.

### Chuẩn hóa dữ liệu bằng StandardScaler:

Chuẩn hóa dữ liệu là một bước quan trọng trong tiền xử lý dữ liệu trước khi áp dụng các thuật toán máy học, bao gồm cả K-means clustering. Trong ngữ cảnh phân cụm, việc chuẩn hóa dữ liệu giúp đảm bảo rằng tất cả các biến số có cùng miền giá trị, loại bỏ các sai biệt tỷ lệ và đảm bảo rằng tất cả các biến số đều được xem xét với cùng một trọng số.

StandardScaler là một công cụ phổ biến được sử dụng để chuẩn hóa dữ liệu. Cách hoạt động của nó đơn giản: nó biến đổi dữ liệu sao cho giá trị trung bình của mỗi biến là 0 và độ lệch chuẩn là 1. Công thức chuẩn hóa dữ liệu bằng StandardScaler cho một biến là:

Trong đó:

* là giá trị đã được chuẩn hóa.
* là giá trị gốc của biến.
* là giá trị trung bình (mean) của biến .
* là độ lệch chuẩn (standard deviation) của biến .

Mặc dù chuẩn hóa dữ liệu có nhiều ưu điểm, nhưng cũng có nhược điểm cần xem xét. Một số điểm quan trọng:

* **Ảnh Hưởng Đến Biến Sao Lệch:** Chuẩn hóa dữ liệu có thể làm thay đổi phân phối của biến số. Điều này có thể ảnh hưởng đến tính phân phối ban đầu của biến số, đặc biệt đối với các biến số không tuân theo phân phối chuẩn ban đầu.
* **Không Phù Hợp Cho Dữ Liệu Rời Rạc:** Chuẩn hóa dữ liệu làm mất đi tính rời rạc của biến số, điều này không phù hợp khi làm việc với dữ liệu rời rạc (categorical data).

Khi áp dụng K-means clustering, việc chuẩn hóa dữ liệu bằng StandardScaler thường được thực hiện để đảm bảo rằng khoảng cách giữa các điểm dữ liệu là đồng nhất và không bị ảnh hưởng bởi sai biệt tỷ lệ của biến số. Điều này cải thiện tính hiệu quả của phân cụm và giúp đưa ra kết quả phân chia tốt hơn dựa trên cơ sở khoảng cách.

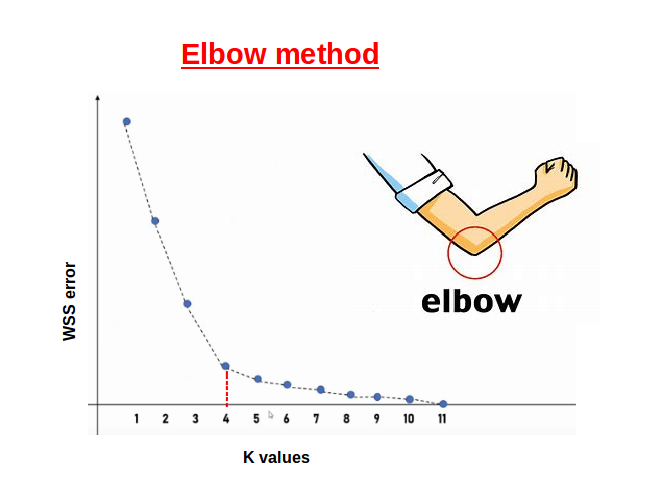
### Tìm số cụm K tốt nhất bằng Elbow:

Tìm số lượng phân cụm (clusters) tốt nhất cho một bài toán phân cụm là một phần quan trọng của quá trình thực hiện K-means clustering. Một trong những cách phổ biến để xác định số K tốt nhất là sử dụng phương pháp gọi là "phương pháp Elbow" (Elbow method). Phương pháp này giúp đánh giá hiệu suất của việc phân cụm dựa trên sự thay đổi của tổng bình phương khoảng cách giữa các điểm dữ liệu và trung tâm cụm khi thay đổi số lượng phân cụm.

Phương pháp Elbow hoạt động như sau:

* Thực hiện K-means clustering với một loạt giá trị K khác nhau, tức là thử nghiệm từ 1 đến một số lớn tùy chọn.
* Đối với mỗi giá trị K, tính tổng bình phương khoảng cách (Sum of Squared Distances, SSD) giữa các điểm dữ liệu và trung tâm cụm. SSD là một phép đo cho biết mức độ phân tán của các điểm dữ liệu trong một phân cụm. SSD thường là một giá trị dương.
* Trình bày SSD theo K. Thông thường, khi K tăng lên, SSD sẽ giảm vì mỗi cụm sẽ có ít điểm dữ liệu hơn và khoảng cách giữa các điểm dữ liệu và trung tâm cụm sẽ giảm đi. Tuy nhiên, tồn tại một điểm trong biểu đồ SSD (thường gọi là "điểm Elbow") khi mà sự giảm đi đột ngột của SSD trở nên đáng kể chậm lại.
* Điểm Elbow là giá trị K mà chúng ta chọn làm số lượng phân cụm tốt nhất cho bài toán phân cụm.

Tên "Elbow method" xuất phát từ hình dạng của biểu đồ SSD khi vẽ theo giá trị K. Biểu đồ thường có hình dạng giống cánh cổng của một khuôn cầu, với điểm Elbow nằm ở chỗ cong, giống hình dáng của "khuỷu tay" (elbow).



Hình 5. Phương pháp khuỷu tay Elbow.

Tuy nhiên, phương pháp Elbow cũng có hạn chế. Đôi khi biểu đồ SSD không có một điểm rõ ràng mà có thể coi là điểm Elbow, và việc lựa chọn K có thể trở nên mơ hồ. Ngoài ra, Elbow method không phụ thuộc hoàn toàn vào số lượng phân cụm tốt mà còn phụ thuộc vào cấu trúc dữ liệu và mục tiêu của việc phân cụm. Do đó, nó thường cần được sử dụng kết hợp với kiến thức về lĩnh vực cụ thể và kiểm tra kết quả phân cụm.

# CHƯƠNG 3: BỘ DỮ LIỆU

## Thu thập bộ dữ liệu:

Bộ dữ liệu được lấy từ KAGGLE với tên bộ dữ liệu là **CreditCardData.**

**Link:** https://www.kaggle.com/datasets/vikashsingh999/creditcarddata

## Mô tả bộ dữ liệu:

Bộ dữ liệu mẫu tổng hợp hành vi sử dụng của khoảng 9000 chủ thẻ tín dụng hoạt động trong vòng 6 tháng qua. Bộ dữ liệu gồm 18 thuộc tính, trong đó, bao gồm 7 thuộc tính phân loại và 11 thuộc tính số.

|  |  |  |
| --- | --- | --- |
| Kiểu | Tên thuộc tính | Mô tả |
| PHÂN LOẠI | CUST\_ID | Mã số nhận dạng của chủ thẻ tín dụng |
| BALANCE\_FREQUENCY | Tần suất cập nhật số dư |
| PURCHASES\_FREQUENCY | Tần suất mua hàng |
| ONEOFFPURCHASESFREQUENCY | Tần suất mua hàng lớn trong một lần |
| PURCHASESINSTALLMENTSFREQUENCY | Tần suất mua hàng trả góp |
| CASHADVANCEFREQUENCY | Tần suất thanh toán trước bằng tiền mặt |
| TENURE | Kì hạn thẻ tín dụng của người dùng |
| Số | BALANCE | Số dư còn lại trong tài khoản |
| PURCHASES | Số tiền đã mua hàng từ tài khoản |
| ONEOFF\_PURCHASES | Số tiền mua hàng lớn nhất trong một lần |
| INSTALLMENTS\_PURCHASES | Số tiền mua hàng theo hình thức trả góp |
| CASH\_ADVANCE | Số tiền trả trước được cấp bởi người dùng |
| CASHADVANCETRX | Số giao dịch được thực hiện với "Thanh toán trước bằng tiền mặt" |
| PURCHASES\_TRX | Số giao dịch mua hàng đã thực hiện |
| CREDIT\_LIMIT | Hạn mức thẻ tín dụng cho người dùng |
| PAYMENTS | Số tiền thanh toán được thực hiện bởi người dùng |
| MINIMUM\_PAYMENTS | Số tiền tối thiểu thanh toán |
| PRCFULLPAYMENT | Tỷ lệ thanh toán đầy đủ của người dùng |

# CHƯƠNG 4: THỰC NGHIỆM

## Tiền xử lý dữ liệu:

Dưới đây là bảng thống kê và mô tả kiểu dữ liệu của từng thuộc tính:

A screenshot of a computer

Description automatically generated

Hình 6. Thống kê dữ liệu.

Dựa vào bảng trên, ta có thể thấy có 2 thuộc tính chứa giá trị NULL đó là:

CREDIT\_LIMIT(1 điểm dữ liệu), MINIMUM\_PAYMENTS(313 điểm dữ liệu).

Có tổng cộng 14 thuộc tính thuộc kiểu float64, 3 thuộc tính thuộc kiểu int64 và 1 thuộc tính thuộc kiểu object.

Nhóm tiến hành thực hiện xử lý 2 vấn đề là:

* Xóa thuộc tính CUST\_ID.
* Xử lý giá trị khuyết bằng KNN Imputer



## Phân tích thăm dò (EDA)

### 2.1. Phân tích hạn mức và số dư dựa trên kỳ hạn

Ở phần này, nhóm dùng thư viện quen thuộc để vẽ biểu đồ là Matplotlib để vẽ biểu đồ Scatter nhằm thể hiện rõ sự phân bố của hạn mức và số dư của tài khoản thông qua kỳ hạn tín dụng của khách hàng.

A graph of purple and yellow dots

Description automatically generated

Hình 7. Biểu đồ Scatter phân tích hạn mức và số dư dựa trên kỳ hạn.

**Nhận xét**:

Biểu đồ scatter cho thấy rằng khi các giá trị Credit-limit tăng lên, phân phối/variance balance cũng tăng lên cho đến khi nó hình thành một hình dạng hình nón (heteroscedasticity). Ngoài ra, như đã đề cập trước đó, phần lớn khách hàng sử dụng thẻ tín dụng ưa chuộng thời gian sử dụng là 12 tháng.

### 2.2. Phân tích số lượng mua so với tổng giao dịch

Ở phần này, nhóm dùng các thư viện quen thuộc để vẽ biểu đồ là Matplotlib và Seaborn để vẽ biểu đồ dumbbell.

A graph with purple dots and white text

Description automatically generated

Hình 8. Biểu đồ dumbbell phân tích số lượng mua so với tổng giao dịch.

**Nhận xét**: Từ biểu đồ dumbbell ở trên, có thể kết luận rằng khách hàng sử dụng thẻ tín dụng với 12 tháng kỳ hạn thường sẵn sàng tiêu nhiều hơn và có nhiều giao dịch mua sắm hơn so với khách hàng có các kỳ hạn khác. Điều này liên quan đến đề cập trước đó, nơi khách hàng thường sẵn sàng trả nợ trong một khoảng thời gian dài để tăng lãi suất. Hơn nữa, một số khách hàng cố ý không thực hiện bất kỳ giao dịch nào (0 giao dịch và mua sắm) để có được hạn mức tín dụng cao, điều này ảnh hưởng đến việc tăng điểm tín dụng và tỷ lệ sử dụng tín dụng.

Biểu đồ trên cũng cho thấy rằng khách hàng có mười hai tháng kỳ hạn có số tiền mua sắm cao hơn so với khách hàng có mười một tháng kỳ hạn. Tuy nhiên, điều này lại nghịch đảo với tổng số giao dịch, trong đó khách hàng có mười tháng kỳ hạn có ít giao dịch hơn so với khách hàng có mười một tháng kỳ hạn. Một ví dụ khác là khách hàng có bảy tháng kỳ hạn có số tiền mua sắm thấp hơn so với khách hàng có tám tháng kỳ hạn. Tuy nhiên, tổng số giao dịch của khách hàng có bảy tháng kỳ hạn lại cao hơn so với khách hàng có tám tháng kỳ hạn.

### 2.3. Phân tích sự tương quan của các thuộc tính:

Ở phần này, nhóm dùng các thư viện quen thuộc để vẽ biểu đồ là Matplotlib và Seaborn để vẽ biểu đồ heatmap.

Biểu đồ correlation matrix sẽ được đính kèm các giá trị tương quan giữa các thuộc tính với nhau, các giá trị này được xác định là tốt hay không tốt bằng:

* Nếu -0.3 < p-value < 0.3 thì độ tương quan của hai thuộc tính không tốt.
* Nếu -0.7 < p-value < -0.3 hoặc 0.7 > p-value > 0.3 thì độ tương quan của hai thuộc tính tương đối tốt.
* Nếu -1 < p-value < -0.7 hoặc 1 > p-value > 0.7 thì độ tương quan của hai thuộc tính rất tốt.

A screenshot of a computer

Description automatically generated

Hình 9. Ma trận tương quan giữa các thuộc tính.

## Chuẩn hóa dữ liệu:

StandardScaler là một công cụ trong thư viện scikit-learn của Python được sử dụng để chuẩn hóa dữ liệu. Chuẩn hóa dữ liệu là quá trình biến đổi các biến số trong tập dữ liệu của bạn sao cho chúng có giá trị trung bình xấp xỉ 0 và phương sai xấp xỉ 1. Điều này giúp trong việc loại bỏ các ảnh hưởng của tỷ lệ giữa các biến số, giúp thuật toán học máy hoạt động tốt hơn.

Sau khi có dữ liệu đã được chuẩn hóa, nhóm tiến hành giảm chiều dữ liệu bằng PCA.

Principal Component Analysis (PCA) là một phương pháp sử dụng để giảm số chiều của dữ liệu bằng cách chuyển dữ liệu gốc sang một không gian mới, sao cho các thành phần chính (principal components) giữ lại phần lớn thông tin của dữ liệu. PCA thường được áp dụng sau khi đã chuẩn hóa dữ liệu bằng StandardScaler.

## Xây dựng mô hình:

### 4.1. Lựa chọn số cụm

Việc lựa chọn số cụm là rất cần thiết cho một mô hình K-mean tối ưu.

Phương pháp "Elbow" là một cách phổ biến để xác định số lượng thành phần chính (hoặc số chiều) tối ưu trong PCA hoặc số cụm tối ưu trong các thuật toán gom cụm như K-means. Phương pháp này giúp chúng ta tìm ra một giá trị "elbow" trên biểu đồ, nơi sự giảm giảm của phương sai hoặc lỗi bắt đầu chậm lại sau một số thành phần chính (hoặc số cụm). Giá trị ở "elbow" thường được coi là số lượng thành phần chính hoặc cụm tối ưu. Dưới đây là cách thực hiện phương pháp "Elbow":

* Vẽ biểu đồ với số lượng thành phần chính trên trục x và tỷ lệ phương sai đã giải thích bởi từng thành phần chính trên trục y.
* Trong biểu đồ, tìm điểm "elbow" hoặc điểm mà sự giảm giảm của tỷ lệ phương sai bắt đầu chậm lại. Điểm này thường là điểm mà bạn cân nhắc giữa việc giảm số lượng thành phần chính và giữ lại một lượng đáng kể của thông tin.
* Số lượng thành phần chính tối ưu là giá trị ứng với điểm "elbow" trên biểu đồ.

A graph with lines and numbers

Description automatically generated

Hình 10. Biểu đồ xác định số cụm bằng phương pháp Elbow.

Sau khi chạy phương pháp Elbow, chúng tôi thấy rằng khi K=4, mô hình sẽ được tối ưu nhất.

### 4.2. Xây dựng mô hình K-mean:

Sau khi có số cụm tối ưu,chúng tôi tiến hành xây dựng mô hình K-mean với K=4. Mô hình được xây dựng dựa trên thư viện sklearn. Dưới đây là một số biểu đồ trực quan kết quả của mô hình K-mean với k=4.

Hình 11A close-up of a graph

Description automatically generated. Một số biểu đồ trực quan kết quả của mô hình.

Silhouette plots là biểu đồ trực quan được sử dụng để đánh giá chất lượng của các phương pháp phân cụm (clustering). Silhouette plots biểu diễn đồ họa cho mỗi điểm dữ liệu trong các cụm và giá trị silhouette score của chúng, thể hiện độ tách biệt giữa các cụm. Giá trị silhouette score trung bình càng gần 1 mô hình phân cụm tốt, còn giá trị gần -1 thì mô hình phân cụm không tốt.

Clusters Distribution là biểu đồ phân bố của các điểm dữ liệu trong các cụm (clusters) sau khi phân cụm bằng K-means (clustering) được áp dụng lên dữ liệu.

Waffle chart là biểu đồ thể hiện phần trăm và tần suất xuất hiện của khách hàng của mỗi cụm trong tập dữ liệu.

## Kết quả và đánh giá:

Độ đo Davis-Bouldin (Davis-Bouldin) là một độ đo được sử dụng để đánh giá hiệu quả của phương pháp phân cụm. Độ đo này tính toán sự khác biệt giữa các cụm (cluster) dựa trên độ tương đồng giữa các điểm dữ liệu bên trong mỗi cụm và khoảng cách giữa các trung tâm cụm. Độ đo Davis-Bouldin càng nhỏ thì phương pháp phân cụm càng tốt.

Độ đo silhouette (Silhouette Coefficient/Score) là một độ đo được sử dụng để đánh giá chất lượng của phương pháp phân cụm. Độ đo này tính toán độ tương đồng giữa một điểm dữ liệu với cụm mà nó thuộc về so với các cụm khác. Độ đo silhouette có giá trị từ -1 đến 1, với giá trị càng gần 1 thì cụm càng tốt, còn giá trị gần -1 thì cụm không tốt.

Độ đo Calinski-Harabasz (Calinski-Harabasz Index) là một độ đo được sử dụng để đánh giá hiệu quả của phương pháp phân cụm. Độ đo này tính toán sự khác biệt giữa các cụm dựa trên tổng trọng số giữa các cụm và tổng trọng số giữa các điểm dữ liệu và trung tâm của chúng. Độ đo Calinski-Harabasz càng lớn thì phương pháp phân cụm càng tốt.

|  |  |
| --- | --- |
| Độ đo | Kết quả |
| Davies-Bouldin | 0.801 |
| Silhouette Score | 0.407 |
| Calinski Harabasz | 5817.434 |

Dựa vào các độ đo, chúng ta có thể thấy rằng mô hình phân cụm khách hàng mà chúng ta đã xây dựng có độ chính xác và độ phân cụm rất là tốt.

A screenshot of a computer

Description automatically generated

Hình 12. Insight của kết quả phân cụm.

Nhận xét về Insight:

Cụm CÓ BALANCE=1009 (Người mới sử dụng/sinh viên): khách hàng hiếm khi/gần như không bao giờ sử dụng thẻ tín dụng cho các giao dịch và trả góp trong. Điều này là do khách hàng có số dư trong tài khoản tương đối nhỏ, tần suất của số dư hiếm khi thay đổi và các khoản trả góp rất thấp. . Ngoài ra, số dư thấp cho phép khách hàng trong cụm này là sinh viên hoặc người dùng mới sử dụng thẻ tín dụng tại ngân hàng này.

Cụm CÓ BALANCE=3407 (Khách hàng lâu năm, trả tiền đầy đủ): tần số của số dư thường thay đổi và số dư cao hơn so với các cụm khác. Khách hàng trong cụm này cũng tích cực sử dụng thẻ tín dụng để giao dịch và trả góp. Các khoản tạm ứng tiền mặt, giao dịch và trả góp trong cụm này cũng xảy ra thường xuyên hơn. Thời gian sử dụng lâu cũng cho thấy điểm tín dụng trong cụm này là rất tốt.

Cụm CÓ BALANCE=4507 (Khách hàng dùng thẻ để ứng trước\rút tiền mặt): Khách hàng trong cụm này có số dư cao, tần suất số dư luôn thay đổi, tần suất tiền mặt ứng trước cao. Ngoài ra, khách hàng trong cụm này có lãi suất thấp nhất so với các cụm khác và có hạn mức tín dụng, thanh toán cao thứ hai trong số bốn cụm. Tuy nhiên, người dùng thẻ tín dụng trong cụm này hiếm khi trả góp hoặc mua hàng một lần và có nhiệm kỳ cao thứ ba trong bốn cụm. Như vậy, có thể kết luận khách hàng trong cụm này chỉ sử dụng thẻ tín dụng cho nhu cầu rút tiền hoặc ứng tiền mặt.

Cụm CÓ BALANCE=826 (Người dùng trả góp): mức độ giao dịch trả góp trong cụm này tương đối cao. Hơn nữa, khách hàng trong cụm này thường thực hiện các giao dịch với số tiền rất lớn cho mỗi giao dịch và tần suất và giao dịch tiền mặt trước là rất nhỏ. Khách hàng trong cụm này rất hiếm khi thực hiện thanh toán bằng tiền mặt trước. Có thể kết luận rằng khách hàng trong cụm này rất phù hợp với thẻ tín dụng dành riêng cho nhu cầu trả góp.

A chart of a credit limit

Description automatically generated

Hình 13. Phân tích kết quả của mô hình.

Dựa trên kết quả phân tích trên, dưới đây là một số đề xuất cho chiến lược tiếp thị:

Nhóm 1: Khách hàng trong nhóm này có thể trở thành mục tiêu chính cho tiếp thị thẻ tín dụng. Điều này là do khách hàng trong nhóm này rất tích cực sử dụng thẻ tín dụng, và họ có thời hạn và giới hạn tín dụng cao nhất so với các nhóm khác. Bằng việc tập trung tiếp thị vào nhóm này, ngân hàng có thể tăng lợi nhuận bằng cách sử dụng thẻ tín dụng thường xuyên hơn và tối ưu hóa chi phí tiếp thị phát sinh. Ngân hàng có thể cung cấp các ưu đãi hoặc phần thưởng để kích thích khách hàng sử dụng thẻ tín dụng thường xuyên hơn.

Dành riêng cho thẻ trả góp: Ngân hàng có thể tập trung tiếp thị thẻ tín dụng trả góp cho khách hàng thuộc nhóm 3. Điều này là do khách hàng trong nhóm 3 có khả năng cao sử dụng thẻ tín dụng cho mục đích trả góp. Ngân hàng có thể cung cấp các chương trình trả góp có lãi suất thấp hoặc 0% và có thể sử dụng cho các nhu cầu trả góp khác nhau để thu hút khách hàng sử dụng thẻ tín dụng. Các yêu cầu trả góp có thể liên quan đến thanh toán du lịch, thiết bị gia đình, thiết bị điện tử, điện thoại thông minh hoặc các thương hiệu cụ thể đang được công chúng yêu thích.

Nhóm 2: Đối với nhóm 2, ngân hàng có thể cung cấp thẻ tín dụng đặc biệt cho người mới hoặc sinh viên (thẻ tín dụng cấp đầu) có thể thiếu một hồ sơ tín dụng chi tiết. Thẻ tín dụng này có thể được sử dụng để xây dựng hồ sơ tín dụng của họ và học cách sử dụng thẻ tín dụng một cách có trách nhiệm. Thẻ này có thể bao gồm một số tính năng, chẳng hạn như miễn phí phí, phần thưởng cho việc thanh toán đúng hẹn, rào cản thấp để trở thành chủ thẻ và lãi suất linh hoạt. Hơn nữa, ngân hàng có thể cung cấp cơ hội nâng cấp lên các sản phẩm mới và điều kiện tốt hơn nếu khách hàng thanh toán hóa đơn đúng hẹn. Ngoài ra, ngân hàng có thể cung cấp phần thưởng đăng ký để thu hút khách hàng chưa sử dụng thẻ tín dụng của ngân hàng.

Nhóm 4: Vì khách hàng trong nhóm 4 thường tiến hành rút tiền mặt trước, ngân hàng có thể cung cấp các thẻ tín dụng đặc biệt với nhiều lợi ích. Những lợi ích này có thể liên quan đến mức phí rút tiền mặt thấp hoặc không mức phí quản lý thấp, lãi suất thấp, thời hạn tương đối cao, v.v. Ngoài ra, ngân hàng cũng có thể cung cấp các chương trình ngân hàng khác ngoài thẻ tín dụng, chẳng hạn như các chương trình trả sau với sự hợp tác của bên thứ ba hoặc các khoản vay cá nhân do ngân hàng cung cấp.

# CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Phân cụm khách hàng là một phương pháp quan trọng để hiểu hành vi của khách hàng và tạo ra chiến dịch tiếp thị hiệu quả. Dựa trên việc phân cụm, chúng ta có thể xác định những đặc điểm chung giữa các nhóm khách hàng và dự đoán những nhóm tiềm năng cho chiến dịch tiếp thị. Quá trình này giúp tối ưu hóa chiến dịch tiếp thị, tập trung vào đúng đối tượng và tăng cơ hội thành công.

Hướng phát triển:

* Tối ưu hóa Dữ liệu: Để cải thiện kết quả phân cụm, chúng ta cần tối ưu hóa quá trình xử lý dữ liệu và làm sạch dữ liệu ban đầu. Điều này bao gồm loại bỏ dữ liệu nhiễu, xử lý dữ liệu bị thiếu và chuẩn hóa dữ liệu.
* Sử dụng Thuật toán Phân cụm Tốt hơn: Ngoài việc sử dụng K-Means, có nhiều thuật toán phân cụm khách hàng khác có thể được áp dụng. Việc nghiên cứu và thử nghiệm các thuật toán khác nhau có thể cải thiện khả năng phân cụm.
* Kết hợp Dữ liệu Bổ sung: Có thể kết hợp dữ liệu khách hàng từ nhiều nguồn khác nhau, chẳng hạn như dữ liệu mạng xã hội, để cung cấp thông tin thêm về hành vi và sở thích của khách hàng.
* Phân tích Chi tiết Hơn: Khám phá các biến số quan trọng hơn trong quá trình phân cụm và xác định những yếu tố quyết định đằng sau hành vi của khách hàng.

# TÀI LIỆU THAM KHẢO

[1] A. Sagar, “Customer Segmentation Using K Means Clustering - Towards Data Science,” *Medium*, Aug. 24, 2019. https://towardsdatascience.com/customer-segmentation-using-k-means-clustering-d33964f238c3 (accessed Oct. 29, 2023).

‌ [2] P. Majumder, “K-Means clustering with Mall Customer Segmentation Data | Full Detailed Code and Explanation,” *Analytics Vidhya*, May 25, 2021. https://www.analyticsvidhya.com/blog/2021/05/k-means-clustering-with-mall-customer-segmentation-data-full-detailed-code-and-explanation/ (accessed Oct. 29, 2023).

‌ [3] Phạm Kiên Trung, Nguyễn Đức Thắng, Lê Văn Chiến, and Nguyễn Văn Thưởng, “Ứng dụng thuật toán K-Means trong phân cụm khách hàng mục tiêu,” *Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất*, vol. 61, no. 5(KTQTKD), pp. 145–150, 2020, Accessed: Oct. 29, 2023. [Online]. Available: https://vjol.info.vn/index.php/tcidhmodiachat/article/view/55494

‌ [4] T. Vu, *Tiep Vu’s blog*, 2018. https://machinelearningcoban.com/ (accessed Oct. 29, 2023).

‌ [5] Qualtrics, “Customer Segmentation Analysis: Definition & Methods,” *Qualtrics*, Apr. 12, 2021. https://www.qualtrics.com/au/experience-management/brand/customer-segmentation/ (accessed Oct. 29, 2023).

‌ [6] A. S. Gillis, “customer segmentation,” *Customer Experience*, 2023. https://www.techtarget.com/searchcustomerexperience/definition/customer-segmentation (accessed Oct. 29, 2023).

‌