**TRƯỜNG ĐẠI HỌC THỦY LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

**----------------------------------------**  
  


**KHAI PHÁ DỮ LIỆU**

**BÁO CÁO BÀI TẬP LỚN**

**Đề tài: Phân loại khách hàng sử dụng**

**phương pháp phân cụm K-Means**

Nhóm sinh viên thực hiện: **Nhóm 15**

*Nguyễn Đức Trình – 2051063932*

*Phạm Quang Trung - 2051063933*

**Giảng viên hướng dẫn: TS.Trần Mạnh Tuấn**

**Hà Nội, Tháng 12 Năm 2023.**

# Lời nói đầu

*Trong thời đại ngày nay, sự phát triển không ngừng của khoa học và công nghệ đã mở ra một thế giới tri thức vô cùng phong phú và đa dạng. Kho dữ liệu to lớn và nguồn tri thức ngày càng mở rộ đã tạo ra một thách thức đầy thú vị và quan trọng: làm thế nào chúng ta có thể tận dụng tri thức đó để tạo ra giá trị thực sự cho xã hội và cuộc sống?*

*Trong bối cảnh hiện nay, việc tìm kiếm, xử lý và khai thác thông tin đã trở thành một nhiệm vụ cấp thiết, đặc biệt là trong lĩnh vực quản lý và hoạt động kinh doanh. Để đáp ứng yêu cầu này, nhiều công cụ tìm kiếm và xử lý thông tin đã được phát triển, nhưng vẫn còn nhiều thách thức phải đối mặt. Các phương pháp truyền thống khai thác cơ sở dữ liệu thường không đủ mạnh mẽ để hiểu biết sâu rộng về tri thức ẩn sau dữ liệu.*

*Nhận thức về những thách thức này, nhóm em đã đặt mục tiêu nghiên cứu vào một hướng mới mẻ: "****Phân loại khách hàng sử dụng phương pháp phân cụm K-Means****". Mục đích của báo cáo này là tìm hiểu và áp dụng kỹ thuật khai phá dữ liệu để phân loại khách hàng dựa vào dữ liệu của mua sắm. Nhóm em hy vọng rằng việc áp dụng phương pháp phân cụm K-Means sẽ giúp chúng ta nhận biết các nhóm khách hàng có đặc điểm tương tự, từ đó đưa ra thông tin về các loại khách hàng.*

*Trong phạm vi báo cáo này, nhóm em sẽ trình bày quy trình thực hiện nghiên cứu, từ việc thu thập dữ liệu, tiền xử lý, áp dụng phương pháp K-Means, đến việc phân tích và trình bày kết quả. Hy vọng rằng thông qua nỗ lực nghiên cứu này, nhóm em sẽ đóng góp một phần nhỏ vào việc tận dụng tri thức từ dữ liệu để nâng cao chất lượng quản lý khách hàng của các của hàng online.*

*1*

**Mục Lục**

# 

# 

# 

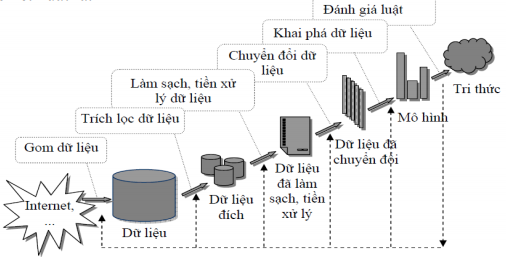
# Chương 1: Tổng quan về khai phá dữ liệu

## Phát hiện tri thức và khai phá dữ liệu

**Phát hiện tri thức (Knowledge Discovery)** trong các cơ sở dữ liệu là một quá trình nhận ra các mẫu hoặc mô hình ẩn chứa trong dữ liệu, mang những đặc điểm: hợp thức, mới mẻ, hữu ích và có thể hiểu được.

**Khai phá dữ liệu (Data Mining)** được định nghĩa: “Data Mining là một quá trình tìm kiếm, phát hiện các tri thức mới, tiềm ẩn, hữu dụng trong CSDL lớn”. Việc khai thác dữ liệu có thể áp dụng vào nhiều lĩnh vực như y tế, phân tích thị trường, xây dựng và có thể được xem là thành quả của sự phát triển tự nhiên trong lĩnh vực công nghệ thông tin.

## Quy trình khám phá tri thức trong CSDL



*Hình 1.1 Quá trình khai phá tri thức từ CSDL*

Phát hiện tri thức trong CSDL gồm những bước quan trọng sau:

* **Làm sạch dữ liệu**: Loại bỏ các dữ liệu nhiễu và dữ liệu không nhất quán.
* **Tích hợp dữ liệu**: Tổng hợp dữ liệu từ nhiều nguồn khác nhau để tạo ra một tập dữ liệu toàn diện hơn.
* **Lựa chọn dữ liệu**: Trích xuất các dữ liệu phù hợp với mục tiêu phân tích từ CSDL.
* **Chuyển đổi dữ liệu**: Chuyển đổi và tạo một cấu trúc dữ liệu thích hợp để dễ dàng tiến hành khai phá.
* **Khai phá dữ liệu**: Bước quan trọng, sử dụng các kỹ thuật thông minh để tìm ra những mẫu và thông tin quan trọng từ dữ liệu.
* **Đánh giá mẫu**: Các chuyên gia dữ liệu sẽ áp dụng các chỉ số và phương pháp để định giá giá trị thực sự và tầm quan trọng của các mẫu tri thức.
* **Biểu diễn tri thức**: Ở giai đoạn này, các kỹ thuật biểu diễn và hiển thị dữ liệu được sử dụng để trình bày tri thức đã khám phá cho người dùng.

## Một số kỹ thuật khai phá dữ liệu

* **Kỹ thuật khai phá luật kết hợp (Association Rule Mining)**: Kỹ thuật này tập trung vào việc tìm hiểu mối quan hệ giữa các mục trong dữ liệu. Nó phát hiện ra các luật kết hợp giữa các mục xuất hiện cùng nhau trong tập dữ liệu. Ví dụ, từ dữ liệu bán hàng, kỹ thuật này có thể phát hiện ra rằng nếu khách hàng mua sản phẩm A thì họ cũng có xu hướng mua sản phẩm B.
* **Kỹ thuật phân lớp (Classification)**: Kỹ thuật này dự đoán và gán các mẫu dữ liệu vào một số lớp hay nhãn đã xác định trước dựa trên các đặc điểm của chúng. Ví dụ, trong phân tích email, kỹ thuật này có thể phân loại email là "spam" hoặc "không phải spam" dựa trên nội dung và thuộc tính khác.
* **Kỹ thuật phân cụm (Clustering)**: Kỹ thuật này nhóm các mẫu dữ liệu vào các cụm dựa trên sự tương đồng hoặc thuộc tính chung. Nó giúp phát hiện cấu trúc ẩn trong dữ liệu mà không cần biết trước các nhãn. Ví dụ, trong phân tích dữ liệu khách hàng, kỹ thuật này có thể nhóm khách hàng có sở thích tương tự vào các nhóm.

Những kỹ thuật này cùng với nhiều kỹ thuật khác trong khai phá dữ liệu giúp chúng ta hiểu rõ hơn về dữ liệu, phát hiện ra mẫu ẩn và thông tin hữu ích, và từ đó hỗ trợ trong việc ra quyết định và đưa ra dự đoán.

**Chương 2: Mô tả bài toán và sử lý sữ liệu thô**

**2.1: Mô tả bài toán**

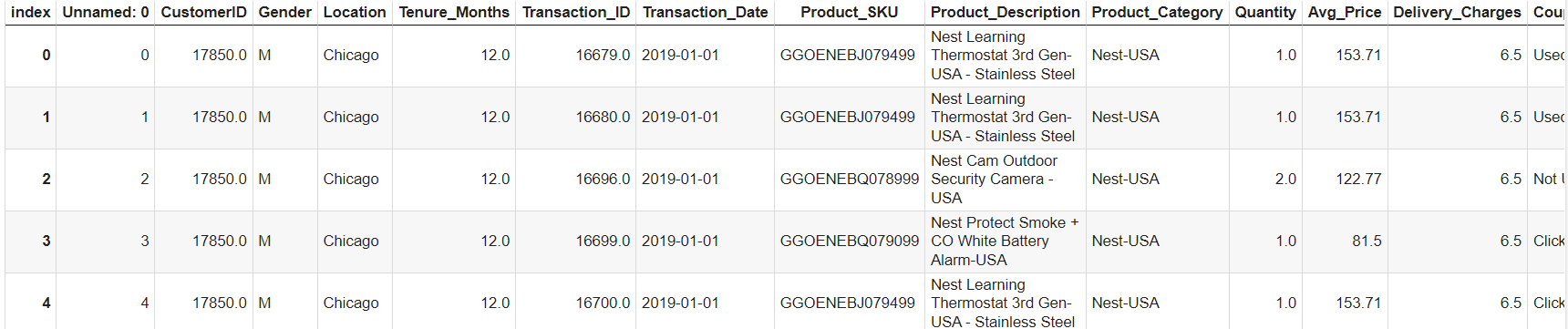
Bài toán đề xuất tập trung vào khai phá dữ liệu khách hàng thông qua việc áp dụng phương pháp phân cụm K-means. Bằng cách sử dụng bộ dữ liệu chứa thông tin về các thong tin mua bán của khách hàng, mục tiêu của bài toán là thực hiện việc phân cụm để xác định loại khách hàng thuộc nhóm nhóm khách nào dựa trên đặc trưng liên quan đến hành vi mua sắm của họ. Từ đó xây dựng ứng dụng dự đoán nhóm khách hàng của người mua.

* Đầu vào: Tập dữ liệu chứa các thuộc tính về thông tin mua bán online
* Đầu ra: Dữ liệu về các tâm cụm, thể hiện cho đặc trưng về khách hàng ở mỗi cụm

## 2.2. Phân tích dữ liệu thô

Nguồn dữ liệu thô:[🛒 Online Shopping Dataset 📊📉📈](https://www.kaggle.com/datasets/jacksondivakarr/online-shopping-dataset), bộ dữ liệu thô về mua sắm online dùng cho việc phân cụm

* Số lượng thuộc tính: 20
* Số lượng bản ghi: 52954



Hình 2.1: Dữ liệu ban đầu

| Số thứ tự | Thuộc tính | Ý nghĩa thuộc tính |
| --- | --- | --- |
| 1 | CustomerID | ID khách hàng |
| 2 | Gender | Giới tính |
| 3 | Location | Địa chỉ của khách hàng |
| 4 | Tenure\_Months | Thời gian khách hàng đã làm việc hoặc sử dụng dịch vụ trong tháng. |
| 5 | Transaction\_ID | Định danh giao dịch |
| 6 | Transaction\_Date | Ngày thực hiện giao dịch |
| 7 | Product\_SKU | Định danh sản phẩm theo mã SKU |
| 8 | Product\_Description | Mô tả sản phẩm |
| 9 | Product\_Category | Nhóm sản phẩm thuộc loại nào |
| 10 | Quantity | Số lượng sản phẩm được mua |
| 11 | Avg\_Price | Giá trung bình của sản phẩm |
| 12 | Delivery\_Charges | Chi phí vận chuyển |
| 13 | Coupon\_Status | Trạng thái của mã giảm giá |
| 14 | GST | Thuế giá trị gia tăng |
| 15 | Date | Ngày thực hiện giao dịch hoặc sự kiện khác |
| 16 | Offline\_Spend | Chi phí chi tiêu ngoại tuyến |
| 17 | Online\_Spend | Chi phí chi tiêu trực tuyến |
| 18 | Month | Tháng của giao dịch hoặc sự kiện |
| 19 | Coupon\_Code | Mã giảm giá được sử dụng |
| 20 | Discount\_pct | Tỷ lệ giảm giá áp dụng |

2.2: Liệt kê các thuộc tính của dữ liệu

Chương 3: Tiền xử lý dữ liệu

## **3.1. Làm sạch dữ liệu**

## Làm sạch dữ liệu là quá trình nhận dạng dữ liệu đã có để tiến hành xử lý các dữ liệu bị thiết (missing data), xử lý dữ liệu bị nhiễu (noisy data) và dữ liệu không nhất quán (inconsistent data).

Đọc file dữ liệu và thực hiện làm sạch dữ liệu bằng Python:

* Xóa các thuộc tính không cần thiết như: 'CustomerID','Transaction\_ID','Product\_Description','Product\_SKU','Coupon\_Code'



* Xử lý dữ liệu bị thiếu: Xóa các dữ liệu bị thiếu



* Xử lý dữ liệu bị thiếu: chèn giá trị của biến df.columns vào chuỗi cần in ra.



## **3.2. Tích hợp dữ liệu**

Tích hợp dữ liệu là quá trình kết hợp dữ liệu từ các nguồn khác nhau để tạo thành một tập dữ liệu hoàn chỉnh hơn hoặc để tạo ra các kết hợp mới của dữ liệu.

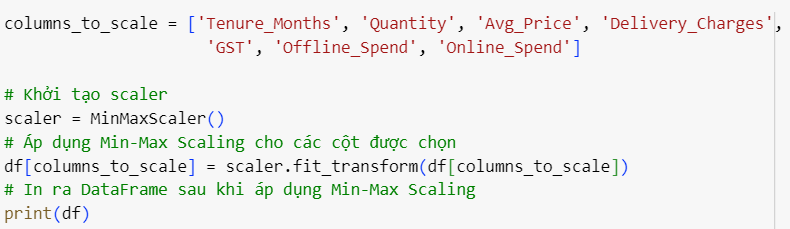
=> Dữ liệu lấy từ một nguồn nên không thực hiện quá trình này

## **3.3. Biến đổi dữ liệu**

Biến đổi dữ liệu là quá trình thay đổi cấu trúc hoặc nội dung của dữ liệu để tạo ra các biến mới hoặc thay đổi dạng của dữ liệu hiện có. Mục tiêu của bước này là cải thiện tính phù hợp của dữ liệu với mục tiêu phân tích hoặc mô hình hóa.

Đọc file dữ liệu và thực hiện biến đổi dữ liệu bằng Python:

* Biến đổi dữ liệu bằng phương pháp Min Max Scaling: đưa các giá trị dữ liệu về khoảng giá trị cụ thể, thường là trong khoảng từ 0 đến 1.



-Công thức:

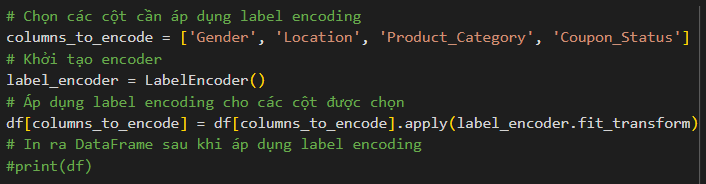
v'=(v-minA)/(maxA-minA )

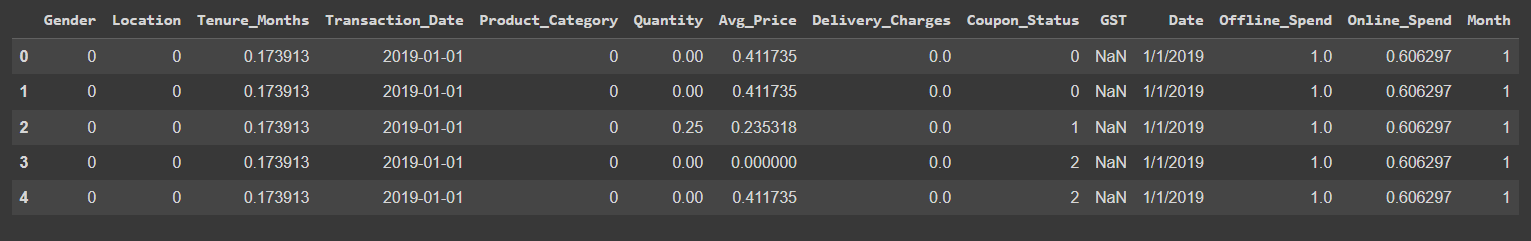
Trong đó:

v = [minA, maxA] là giá trị cũ

v’ = [0,1] là giá trị mới

* Biến đổi các giá trị chữ thành số:





Hình 31: Bảng giá trị sau khi chuẩn hóa

**Chương 4: Khai phá dữ liệu bằng thuật toán K-means**

Giới thiệu về kỹ thuật phân cụm

* Khái niệm:

Phân cụm thuộc loại học không giám sát (Unsupervised learning) là một dữ liệu là bài toán gom nhóm các đối tượng dữ liệu vào thánh từng cụm (cluster) sao cho các đối tượng trong cùng một cụm có sự tương đồng theo một tiêu chí nào đó. Ví dụ: phân nhóm khách hàng dựa trên hành vi mua hàng.

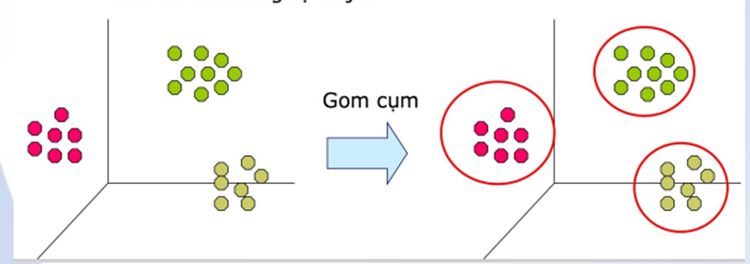
Là một kỹ thuật trong data mining nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên tiềm ẩn, quan tâm trong tập dữ liệu lớn, từ đó cung cấp thông tin, tri thức hữu ích cho ra quyết định

Các hướng tiếp cận trong phân cụm:

* Tìm phần tử ngoại lai (Outlier)

Xác định một số nhóm nhỏ các đối tượng dữ liệu “Khác thường” so với các dữ liệu trong để tránh sự ảnh hưởng của chúng tới quá trình kết quả của Phân cụm dữ liệu

* Làm sạch dữ liệu
* Các bước xây dựng:
* Xây dựng hàm tính độ tương tự
* Xây dựng các tiêu chuẩn phân cụm
* Xây dựng mô hình cho cấu trúc cụm dữ liệu
* Xây dựng thuật toán phân cụm và xác lập các điều kiện khởi tạo
* Xây dựng các thủ tục biểu diễn và đánh giá kết quả phân cụm
* Kỹ thuật gom cụm:
* Là quá trình gom / nhóm cụm dữ liệu / đối tượng vào các lớp / cụm
* Các đối tượng cùng 1 cụm tương tự với nhau hơn là đối tượng ở các cụm



Hình 4.1 Mô tả về kỹ thuật gom cụm

4.2. Thuật toán K-means

Thuật toán K-means là một trong các thuật toán phân cụm đơn giản và điển hình nhất, do MacQueen đề xuất trong lĩnh vực thống kê năm 1967. Mục đích của thuật toán là sinh ra k cụm dữ liệu từ một tập dữ liệu ban đầu gồm n đối tượng trong không gian p chiều.

* Phát biểu bài toán phân cụm:
  + Input: n đối tượng và số các cụm k
  + Output: các cụm Ci (i = 1…k) sao cho hàm tiêu chuẩn E đạt giá trị tối thiểu

Các bước của thuật toán K-mean: [3]

**Bước 1: Khởi tạo tâm cụm**

Chọn k đối tượng mj (j=1…k) là trọng tâm ban đầu của k cụm từ tập dữ liệu

**Bước 2: Tính toán khoảng cách và gán cụm**

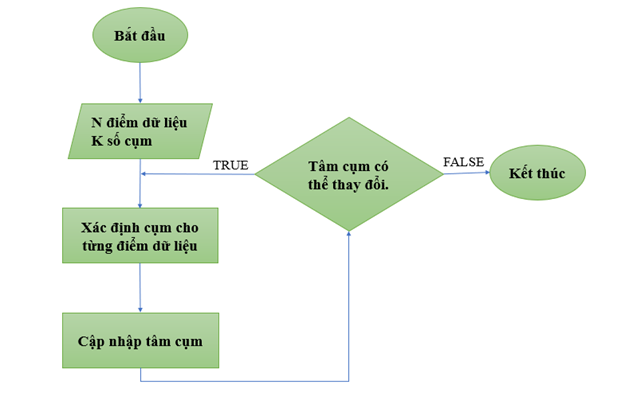
* Với mỗi điểm dữ liệu, tính toán khoảng cách từ nó tới mỗi tâm cụm đã được khởi tạo, sau đó tìm tâm cụm gần nhất đối với mỗi đối tượng
* Công thức tính khoảng cách Euclidean giữa hai điểm A (a1, a2, ..., an) và B (b1, b2, ..., bn) trong không gian n chiều là:
* Distance(A,B)=√((a1-b1)^2+ (a2-b2)^2+⋯+ (an-bn)^2 )

**Bước 3: Cập nhật lại trọng tâm**

Với mỗi j=1,…,k , cập nhật trọng tâm cụm mj bằng cách xác định trung bình cộng của các vector đối tượng dữ liệu

**Bước 4: Kiểm tra điều kiện dừng**

Lặp lại các bước 2 và 3 cho đế khi các tâm cụm không thay đổi



Hình 4.2 Lưu đồ thuật toán K-means

**4.3.Thực hiện Phân cụm dữ liệu bệnh nhân bằng thuật toán K-means**

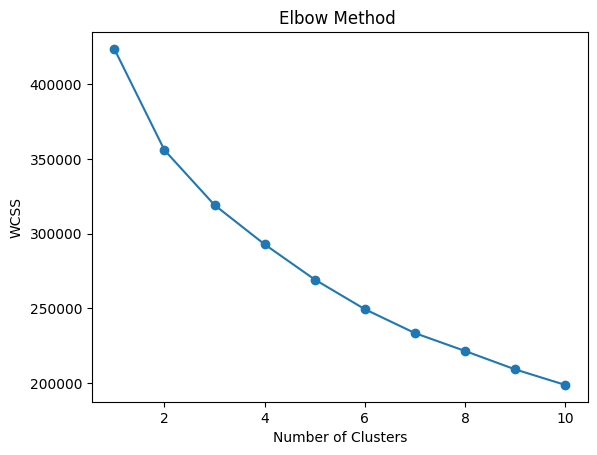
4.3.1. Xác định số tâm cụm tối ưu:

Để xác định số lượng cụm K, cần sử dụng một biểu đồ được tạo bằng Elbow Method (phương pháp khuỷu tay) – là một phương pháp thường được sử dụng để chọn số lượng tốt nhất của các cụm trong thuật toán phân cụm. Phương pháp này giúp xác định số lượng cụm tối ưu dựa trên sự thay đổi của tổng bình phương khoảng cách (Within-Cluster Sum of Squares - WCSS) giữa các điểm dữ liệu và trung tâm của cụm.

Cách thức thực hoạt động của Elbow Method như sau:

* Thực hiện thuật toán KMeans với một loạt số lượng cụm khác nhau (thường từ 1 đến một giới hạn tối đa đã xác định từ trước).
* Đối với mỗi số lượng cụm, tính giá trị WCSS (tổng bình phương khoảng cách giữa các điểm dữ liệu và trung tâm cụm tương ứng).
* Vẽ biểu đồ Elbow Method với trục x là số lượng cụm và trục y là giá trị WCSS. Đường cong trên biểu đồ thường có dạng "cổ điển" như cùi chỏ, giống như đoạn của một cánh tay và một bàn tay đang giơ lên.
* Xác định điểm "bẻ" của đường cong (nơi mà đường cong thay đổi hướng từ việc giảm dần sang giảm chậm hơn). Điểm này thường gọi là "đỉnh của cùi chỏ" (elbow point).
* Số lượng cụm tốt nhất sẽ là số lượng cụm tương ứng với điểm "bẻ" trên biểu đồ Elbow Method.

Ý nghĩa của Elbow Method là khi số lượng cụm tăng, WCSS thường giảm vì các điểm dữ liệu được gom vào cụm một cách gần hơn với trung tâm cụm. Tuy nhiên, tại một điểm nào đó, việc tăng số lượng cụm không còn gây ra sự giảm đáng kể trong WCSS, và đây chính là điểm mà đường cong bẻ. Điều này xảy ra vì thêm một cụm mới không còn đáng giá trong việc giảm tổng bình phương khoảng cách nữa, và sự thay đổi giữa số lượng cụm đã đạt một mức ổn định.



Hình 4.3: Biểu đồ Elbow

Từ biểu đồ trên, nhận thấy rằng tại số tượng cụm bằng 3 thì biểu đồ thay đổi hướng từ việc giảm dần sang giảm chậm hơn. Kết luận rằng số lượng cụm tối ưu cho dữ liệu là 3 cụm

### 4.3.3. Cài đặt thuật toán Kmeans bằng Python

Các bước cài đặt thuật toán Kmeans:

* Bước 1: Thực hiện kết nối đến Driver, lấy dữ liệu của các khách hàng để thực hiện phân cụm

=> Bước này không cần thực hiện nên bỏ qua

* Bước 2: Cài đặt hàm thực hiện tính toán khoảng cách từ một điểm dữ liệu đến các tâm cụm