**BỘ THÔNG TIN VÀ TRUYỀN THÔNG**

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**CƠ SỞ TẠI THÀNH PHỐ HỒ CHÍ MINH**



**MÔN: KHAI PHÁ DỮ LIỆU ĐA PHƯƠNG TIỆN**

**ĐỀ TÀI: Ứng dụng thuật toán k-means cluster để phân cụm khách hàng theo thu nhập dựa trên dữ liệu phân khúc khách hàng của trung tâm mua sắm**

**Giảng viên hướng dẫn: ThS. NGUYỄN VĂN SÁU**

**Sinh viên thực hiện**

HỌ VÀ TÊN: Phạm Trúc Tâm

MÃ SINH VIÊN: N18DCPT058

LỚP: D18CQPU01-N

**Hồ Chí Minh ngày 24 tháng 05 năm 2022**

NHẬN XÉT CỦA GIẢNG VIÊN

………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

MỤC LỤC

[CHƯƠNG 1. NGUỒN DỮ LIỆU 3](#_Toc104247933)

[1. Nguồn gốc 3](#_Toc104247934)

[2. Lý giải tên đề tài: 3](#_Toc104247935)

[CHƯƠNG 2. ÁP DỤNG THUẬT TOÁN 4](#_Toc104247936)

[2.1 Áp dụng thuật toán KMeans 4](#_Toc104247937)

[2.2 Các thư viện sử dụng: 4](#_Toc104247938)

[2.3 Các tham số 4](#_Toc104247939)

[CHƯƠNG 3. KẾT QUẢ 5](#_Toc104247940)

# CHƯƠNG 1. NGUỒN DỮ LIỆU

## Nguồn gốc

- Lấy dữ liệu từ: <https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python>

- Dữ liệu ảo, thời gian của tập dữ liệu được cập nhật lần cuối là tháng 10 năm 2016.

- Dữ liệu gốc: 200 dòng, 5 cột

Trong đó:

* **CustomerID**: id của khách hàng
* **Gerder**: giới tính của khách hàng
* **Age**: độ tuổi của khách hàng
* **Annual Income (k$)**: thu nhập hằng năm của khách hàng tính theo nghìn đô
* **Spending Score (1-100)**: điểm do trung tâm mua sắm chỉ định dựa trên hành vi của khách hàng và bản chất chi tiêu

Table

Description automatically generated

## Lý giải tên đề tài:

* Trong đề tài này ta sẽ phân tích mức độ mua sắm của khách hàng dựa trên chi tiêu để từ đó lên chiến lượt bán sản phẩm dựa trên nhóm khách hàng.
* Do đó ta sử dụng 2 cột chính để phân tích là cột **Annual Income (k$)** và cột **Spending Score (1-100)**

# CHƯƠNG 2. ÁP DỤNG THUẬT TOÁN

## 2.1 Áp dụng thuật toán KMeans

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

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

## 2.2 Các thư viện sử dụng:

- **KMeans**

- **pandas**: Pandas được thiết kế để làm việc dễ dàng và trực quan với dữ liệu có cấu trúc (dạng bảng, đa chiều, có tiềm năng không đồng nhất) và dữ liệu chuỗi thời gian.

- **matplotlib**: Nó là một thư viện đa nền tảng để tạo các đồ thị 2D từ dữ liệu trong các mảng.

## 2.3 Các tham số

200 dòng và 5 cột

# CHƯƠNG 3. KẾT QUẢ

**Mô tả, diễn giải, ý nghĩa thực tế**

Bạn đang chủ sở hữu một trung tâm mua sắm ở siêu thị và thông qua thẻ thành viên, bạn có một số dữ liệu cơ bản về khách hàng của mình như ID khách hàng, tuổi, giới tính, thu nhập hàng năm và điểm chi tiêu.

Điểm chi tiêu là thứ bạn chỉ định cho khách hàng dựa trên các thông số đã xác định của bạn như hành vi của khách hàng và dữ liệu mua hàng.

Báo cáo vấn đề:

Bạn sở hữu trung tâm thương mại và muốn hiểu khách hàng như những người có thể dễ dàng hội tụ [Khách hàng tiềm năng] để đội ngũ tiếp thị có thể hiểu và hoạch định chiến lược cho phù hợp.

**Kết quả:**

Chart, line chart

Description automatically generated

* Chart, line chart

  Description automatically generated
* Theo biểu đồ Elbow trên thì ta sẽ chọn k = 5 vì ta phải chọn giá trị của k tại vị trí “khủy tay” tức là điểm mà sau đó biến dạng/quán tính bắt đầu giảm theo kiểu tuyến tính. Do đó ta kết luận rằng số cụm tối ưu cho dữ liệu là 5.

Chart, scatter chart

Description automatically generated