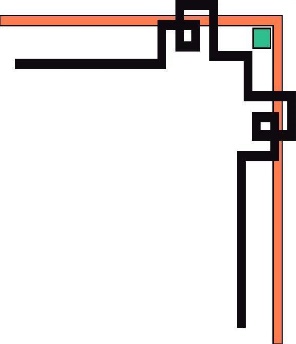
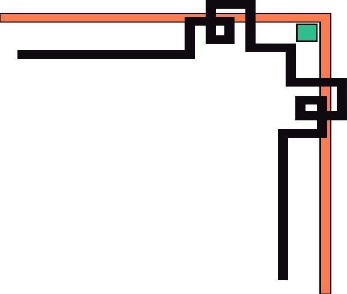
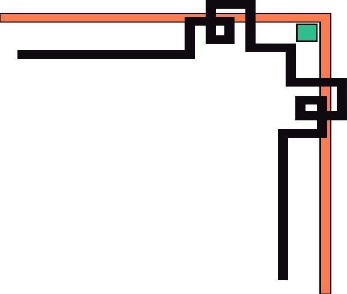
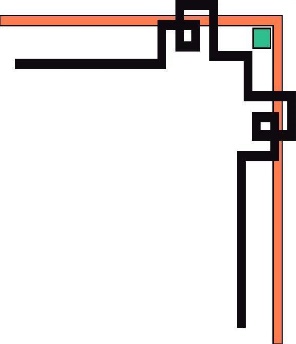
ĐẠI HỌC HUẾ



**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**

A logo with text on it

Description automatically generated

**BÁO CÁO KẾT QUẢ**

**HỌC KÌ I, NĂM HỌC 2022-2023**

**HỌC PHẦN:**

**HỌC MÁY 1**

**Thừa Thiên Huế, ngày 2 tháng 1 năm 2024**

**Giảng viên hướng dẫn**: TS. Hoàng Hữu Trung

**Lớp**: Khoa học dữ liệu và Trí tuệ nhân tạo K3

**Nhóm sinh viên thực hiện**: Hồ Tăng Nhật Hiếu

Phạm Phước Bảo Tín

Nguyễn Văn Minh Khánh

Văn Khiêm Chương

## Phần 1: Giới thiệu dữ liệu

Chúng tôi đã chọn bộ dữ liệu có tên là "**Mall Customer Segmentation”.** Đây là tập dữ liệu về phân khúc khách hàng của trung tâm mua sắm.

Gồm các cột:

CustomerID: Mã khách hàng - mã số khách hàng duy nhất - kiểu số nguyên

Gender: Giới tính

Age: Tuổi – kiểu số nguyên

Annual Income (k$): Thu nhập hằng năm của khách hàng – kiểu số nguyên

Spending Score (1-100): Điểm chi tiêu của khách hàng – kiểu số nguyên

Nguồn: [Salary\_datasets](https://www.kaggle.com/datasets/abhishek14398/salary-dataset-simple-linear-regression)

## Phần 2: Tiền xử lý dữ liệu

Import các thư viện cần thiết.

Đọc file dữ liệu và in ra.

A screenshot of a computer

Description automatically generated

Kiểm tra giá trị thiếu.

A screenshot of a computer

Description automatically generated

Kết quả cho thấy không có cột nào có giá trị thiếu.

Kiểm tra thông tin DataFrame

A screenshot of a computer code

Description automatically generated

Xóa cột “ID” không cần thiết:

A black text with red letters

Description automatically generated

Hiển thị lại DataFrame

A screenshot of a screen shot of a number

Description automatically generated

Kết quả cho thấy đã xóa cột “ID” khỏi DataFrame.

Chuẩn hóa cột Gender thành 0 1.

A close-up of a number

Description automatically generated

Kết quả:

A screenshot of a graph

Description automatically generated

Sử dụng Min-Max Scaler để chuẩn hóa các cột 'Age', 'Annual Income (k$)', và 'Spending Score (1-100)' trong DataFrame.

A screen shot of a computer

Description automatically generated

Kết quả:

A screenshot of a graph

Description automatically generated

Dữ liệu trước khi tiền xử lý:

A screenshot of a table

Description automatically generated

Dữ liệu sau khi tiền xử lý:

A screenshot of a graph

Description automatically generated

## Phần 3: Trực quan hóa dữ liệu

Biểu đồ cột về “Thu nhập trung bình năm theo giới tính”

A close-up of a text

Description automatically generated

Kết quả:

A graph with blue rectangular bars

Description automatically generated with medium confidence

Biểu đồ tròn về “Phân phối giới tính”

A close-up of a computer code

Description automatically generated

Kết quả:

A blue and orange pie chart

Description automatically generated

Biểu đồ Histogram về “Lương thu nhập hằng năm”

A computer code with text

Description automatically generated with medium confidence

Kết quả:

A graph with blue lines

Description automatically generated

Biểu đồ phân tán giữa tuổi và điểm chi tiêu

A screenshot of a computer

Description automatically generated

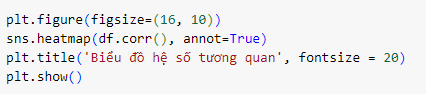
Kết quả:

A graph with blue dots

Description automatically generated

Từ biểu đồ phân tán cho thấy ở thế hệ trẻ có mức chi tiêu cao hơn.

Biểu đồ tương quan



Kết quả:

A screenshot of a computer

Description automatically generated

## PHẦN 4: THUẬT TOÁN KMEANS

Bước đầu tiên trong phân tích phân cụm là nhập K-Means từ thư viện sklearn.



Để phân cụm chỉ có các cột số được sử dụng. Biến nhị phân gender (giới tính) sẽ không được sử dụng.



Để tìm được số cụm thích hợp, phương pháp elbow method sẽ được sử dụng. Trong trường hợp này, quán tính cho một số cụm từ 2 đến 10 sẽ được tính toán. Quy tắc là chọn số cụm mà đường gấp khúc hoặc "elbow method" trong biểu đồ.

A screenshot of a computer code

Description automatically generated

Biểu đồ dưới cho thấy mức giảm điểm biến dạng khi số lượng cụm tăng lên. Tuy nhiên, không có "elbow" rõ ràng. Thuật toán cơ bản gợi ý 5 cụm.

A graph with a blue line and green line

Description automatically generated

Nhập thư viện để tính Kmeans, thời gian và độ chính xác của thuật toán

A close-up of a text

Description automatically generated

Khởi tạo dữ liệu và thuật toán KMeans, tính thời gian và độ chính xác của thuật toán.

A screenshot of a computer program

Description automatically generated

Vẽ biểu đồ trực quan về phân tích cụm (Thu nhập hàng năm so với điểm chi tiêu)

A screenshot of a computer code

Description automatically generated

Kết quả tính thời gian chạy và độ chính xác của thuật toán.



Kết quả vẽ biểu đồ trực quan thuật toán Kmeans.

A diagram of different colored dots

Description automatically generated

Nhận xét biểu đồ:

* Biểu đồ trên thể hiện mối quan hệ giữa thu nhập hàng năm (k$) và điểm chi tiêu (1-100) của một nhóm người. Biểu đồ được chia thành 5 cụm, mỗi cụm đại diện cho một nhóm người có đặc điểm tương đồng về thu nhập và điểm chi tiêu.
* Nhìn chung, có thể thấy rằng thu nhập hàng năm và điểm chi tiêu có mối quan hệ tương quan dương. Điều này có nghĩa là những người có thu nhập cao thường có điểm chi tiêu cao hơn những người có thu nhập thấp.
* Tuy nhiên, biểu đồ cũng cho thấy có một số ngoại lệ. Cụ thể, có một số người có thu nhập cao nhưng điểm chi tiêu thấp (cụm 1) và một số người có thu nhập thấp nhưng điểm chi tiêu cao (cụm 5).
* Các cụm cụ thể có thể được mô tả như sau:
* Cụm 1: Những người có thu nhập cao và điểm chi tiêu thấp. Nhóm người này có thể là những người tiết kiệm, đầu tư hoặc có mục tiêu tài chính cụ thể.
* Cụm 2: Những người có thu nhập trung bình và điểm chi tiêu trung bình. Đây là nhóm người phổ biến nhất trong xã hội.
* Cụm 3: Những người có thu nhập thấp và điểm chi tiêu trung bình. Nhóm người này có thể là những người đang trong giai đoạn khởi nghiệp hoặc gặp khó khăn về tài chính.
* Cụm 4: Những người có thu nhập trung bình và điểm chi tiêu cao. Nhóm người này có thể là những người có sở thích mua sắm hoặc thích tận hưởng cuộc sống.
* Cụm 5: Những người có thu nhập thấp và điểm chi tiêu cao. Nhóm người này có thể là những người có thói quen chi tiêu vượt quá khả năng của mình.