TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP

**KHOA ĐIỆN TỬ**

**Bộ môn: Công nghệ Thông tin.**

**BÀI TẬP KẾT THÚC MÔN HỌC**

MÔN HỌC

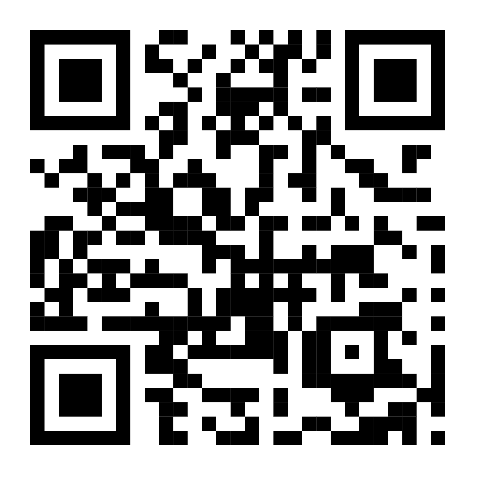
**KHOA HỌC DỮ LIỆU**

Sinh viên: Phạm Quang Trường – K214580106129

Lớp: K57KMT

Giáo viên hướng dẫn: TS. Nguyễn Văn Huy

Link Github: github.com/BeoTrang/BTL\_KHDL



**Thái Nguyên – 2025**

|  |  |
| --- | --- |
| **TRƯỜNG ĐHKTCN** | **CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM** |
| **KHOA ĐIỆN TỬ** | ***Độc lập - Tự do - Hạnh phúc*** |

**BÀI TẬP KẾT THÚC MÔN HỌC**

MÔN HỌC: KHOA HỌC DỮ LIỆU

BỘ MÔN: CÔNG NGHỆ THÔNG TIN

*Sinh viên:* Phạm Quang Trường  *MSSV:* K215480106129

*Lớp*: 57KMT *Ngành:* Kỹ thuật máy tính

*Giáo viên hướng dẫn:* TS.Nguyễn Tuấn Linh

*Ngày giao đề*………………………. *Ngày hoàn thành*………………………

*Tên đề tài*: Phân loại khách hàng tiềm năng.

Yêu cầu:

**Đầu bài:**

* [Mall Customer Segmentation Data](https://www.kaggle.com/datasets/shwetabh123/mall-customers)

**Đầu ra:**

* Phân loại khách hàng và biểu đồ phân nhóm khách hàng.

|  |
| --- |
| **GIÁO VIÊN HƯỚNG DẪN** |
| *(Ký và ghi rõ họ tên)* |
|  |

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

Thái Nguyên, ngày….tháng…..năm 20....

**GIÁO VIÊN HƯỚNG DẪN**

*(Ký ghi rõ họ tên)*

**MỤC LỤC**

[CHƯƠNG I. GIỚI THIỆU ĐẦU BÀI 5](#_Toc199287453)

[1.1. Mô tả đề tài. 5](#_Toc199287454)

[1.2. Mục tiêu. 5](#_Toc199287455)

[1.3. Tính năng của chương trình. 5](#_Toc199287456)

[1.4. Thách thức. 5](#_Toc199287457)

[1.5. Kiến thức áp dụng. 5](#_Toc199287458)

[CHƯƠNG II. CƠ SỞ LÝ THUYẾT 6](#_Toc199287459)

[2.1. Danh sách và xử lý dữ liệu. 6](#_Toc199287460)

[2.2. Mô hình KMeans 6](#_Toc199287461)

[2.3. Đánh giá mô hình phân cụm 6](#_Toc199287462)

[2.4. Thư viện Streamlit 6](#_Toc199287463)

[2.5. Trực quan hóa dữ liệu 6](#_Toc199287464)

[CHƯƠNG III. THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH 7](#_Toc199287465)

[3.1. Sơ đồ khối hệ thống 7](#_Toc199287466)

[3.2. Sơ đồ khối thuật toán 7](#_Toc199287467)

[3.3. Cấu trúc dữ liệu 7](#_Toc199287468)

[a. Bảng dữ liệu đầu vào – Mall\_Customers.csv 7](#_Toc199287469)

[b. Bảng dữ liệu sau khi xử lý – df sau phân cụm 8](#_Toc199287470)

[c. Bảng thống kê cụm – Kết quả phân tích theo cụm 8](#_Toc199287471)

[3.4. Các hàm trong chương trình. 8](#_Toc199287472)

[CHƯƠNG IV. THỰC NGHIỆM VÀ KẾT LUẬN 9](#_Toc199287473)

[4.1. Thực nghiệm 9](#_Toc199287474)

[4.2. Kết luận 10](#_Toc199287475)

# CHƯƠNG I. GIỚI THIỆU ĐẦU BÀI

* 1. **Mô tả đề tài.**

Đề tài tập trung xây dựng một ứng dụng web cho phép phân loại khách hàng tiềm năng dựa trên dữ liệu giao dịch từ tệp CSV. Người dùng có thể tải dữ liệu lên, lựa chọn các thuộc tính đầu vào, điều chỉnh số cụm (k) và xem kết quả phân cụm dưới dạng biểu đồ trực quan.

* 1. **Mục tiêu.**

Xây dựng một hệ thống giúp:

* Phân nhóm khách hàng dựa trên các yếu tố như độ tuổi, thu nhập, điểm chi tiêu...
* Tạo điều kiện cho doanh nghiệp cá nhân hóa chiến lược marketing.
* Hỗ trợ trực quan kết quả phân cụm qua biểu đồ, giúp dễ hiểu và dễ áp dụng vào thực tế.
  1. **Tính năng của chương trình.**

Những tính năng của chương trình :

* Tải dữ liệu khách hàng (dạng CSV)
* Lựa chọn thuộc tính để phân cụm
* Lựa chọn số cụm (k)
* Phân cụm bằng thuật toán KMeans
* Hiển thị kết quả bằng biểu đồ: scatter, pairplot, biểu đồ tròn
* Cho phép tải về dữ liệu đã phân cụm
* Tính toán và hiển thị Silhouette Score và biểu đồ Elbow.
  1. **Thách thức.**

Thách thức của bài toán là :

* Không thể trực quan hóa dữ liệu > 2 chiều bằng biểu đồ scatter.
* Việc lựa chọn k phù hợp cần dựa trên phương pháp đánh giá.
* Hiệu quả phân cụm phụ thuộc chất lượng dữ liệu đầu vào.
  1. **Kiến thức áp dụng.**

Các kiến thức để xử lý bài toán :

* Ngôn ngữ Python và các thư viện như: pandas, matplotlib, seaborn, scikit-learn, streamlit.
* Kiến thức về học máy không giám sát: mô hình KMeans.
* Kiến thức về giao diện người dùng (UI/UX) và thiết kế hệ thống.

# CHƯƠNG II. CƠ SỞ LÝ THUYẾT

## 2.1. Danh sách và xử lý dữ liệu.

Dữ liệu đầu vào ở dạng bảng (CSV), mỗi dòng là thông tin một khách hàng.

A white sheet with numbers and text

AI-generated content may be incorrect.

Dùng pandas để xử lý và trích xuất dữ liệu theo cột: lọc, gom nhóm, chuẩn hóa,...

## 2.2. Mô hình KMeans

* KMeans là một thuật toán phân cụm không giám sát phổ biến.
* Thuật toán sẽ chia dữ liệu thành k cụm sao cho khoảng cách nội cụm nhỏ nhất.
* Mỗi khách hàng sẽ được gán vào cụm gần nhất theo khoảng cách Euclidean.

## 2.3. Đánh giá mô hình phân cụm

* Elbow method: kiểm tra độ giảm WCSS theo k → xác định “góc khuỷu” để chọn k tối ưu.
* Silhouette Score: đánh giá chất lượng phân tách cụm (giá trị từ -1 đến 1).

## 2.4. Thư viện Streamlit

* Cho phép tạo giao diện web bằng Python dễ dàng.
* Hỗ trợ tải file, chọn tùy chọn, hiển thị biểu đồ trực tiếp.

## 2.5. Trực quan hóa dữ liệu

* Dùng scatter plot và pairplot để thể hiện kết quả phân cụm.
* Dùng biểu đồ cột/tròn để thể hiện phân bố khách hàng theo cụm.

# CHƯƠNG III. THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH

## 3.1. Sơ đồ khối hệ thống

Các thành phần chính:

* Giao diện người dùng (UI): Nhận tệp và đầu vào từ người dùng.
* Xử lý dữ liệu: Tiền xử lý dữ liệu, chuẩn hóa.
* Phân cụm: Gọi mô hình KMeans và dự đoán nhãn cụm.
* Hiển thị kết quả: Biểu đồ, bảng, báo cáo thống kê cụm.

[User Upload File] → [Select Features + k] → [KMeans Clustering]

→ [Evaluate & Visualize] → [Download Result]

Ứng dụng phân loại khách hàng

├── Nhập dữ liệu

├── Cấu hình phân cụm (chọn k, chọn thuộc tính)

├── Phân tích và phân cụm

│ ├── KMeans

│ ├── Đánh giá (Elbow, Silhouette)

├── Trực quan hóa dữ liệu

│ ├── Scatter plot

│ ├── Pair plot

│ └── Biểu đồ cụm

└── Tải kết quả xuống

## 3.2. Sơ đồ khối thuật toán

**Input:** DataFrame gồm các cột người dùng chọn (vd: Tuổi, Thu nhập, Điểm chi tiêu).

**Thuật toán chính:**

* Chuẩn hóa dữ liệu đầu vào.
* Áp dụng KMeans(n\_clusters=k) để phân cụm.
* Gán nhãn cụm vào cột Cluster.
* Tính WCSS và Silhouette Score cho đánh giá.

## 3.3. Cấu trúc dữ liệu

### a. Bảng dữ liệu đầu vào – Mall\_Customers.csv

Đây là tập dữ liệu gồm thông tin của 200 khách hàng tại một trung tâm thương mại. Dữ liệu được lưu trữ dưới dạng bảng với các trường sau:

|  |  |  |
| --- | --- | --- |
| **Tên trường** | **Kiểu dữ liệu** | **Mô tả** |
| CustomerID | Số nguyên (int) | Mã định danh duy nhất của khách hàng |
| Gender | Chuỗi (string) | Giới tính của khách hàng (Male hoặc Female) |
| Age | Số nguyên (int) | Tuổi của khách hàng |
| Annual Income (k$) | Số nguyên (int) | Thu nhập hàng năm của khách hàng (đơn vị: ngàn USD) |
| Spending Score (1-100) | Số nguyên (int) | Điểm thể hiện mức độ chi tiêu của khách hàng (1 = thấp, 100 = cao) |

### b. Bảng dữ liệu sau khi xử lý – df sau phân cụm

Sau khi chạy thuật toán KMeans, bảng dữ liệu sẽ được thêm một cột mới:

|  |  |  |
| --- | --- | --- |
| **Tên trường** | **Kiểu dữ liệu** | **Mô tả** |
| Cluster | Số nguyên (int) | Mã cụm mà khách hàng được phân vào (từ 0 đến k-1) |

### c. Bảng thống kê cụm – Kết quả phân tích theo cụm

Sau khi phân cụm, chương trình tạo một bảng tổng hợp thống kê như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **Age (mean)** | **Income (mean)** | **Spending Score (mean)** |
| 0 | 26.4 | 35.2 | 79.1 |
| 1 | 45.1 | 75.5 | 20.2 |
| ... | ... | ... | ... |

## 3.4. Các hàm trong chương trình.

|  |  |
| --- | --- |
| **Tên hàm** | **Chức năng** |
| load\_data(file) | Đọc file CSV và tạo DataFrame |
| run\_kmeans(X, k) | Áp dụng KMeans và trả về nhãn cụm |
| plot\_elbow(X) | Tính và vẽ biểu đồ Elbow |
| plot\_silhouette(X, labels) | Tính và in Silhouette Score |
| plot\_clusters(df, features) | Vẽ biểu đồ phân cụm (scatter/pairplot) |

# CHƯƠNG IV. THỰC NGHIỆM VÀ KẾT LUẬN

## 4.1. Thực nghiệm

Dữ liệu đầu vào:

* Dùng file Mall\_Customers.csv từ Kaggle.
* Gồm 200 khách hàng với 5 cột: ID, Gender, Age, Income, Spending Score.

Các bước test:

1. Tải lên file CSV.
2. Chọn 2 hoặc 3 thuộc tính (Tuổi, Thu nhập, Điểm chi tiêu).
3. Thử nhiều giá trị k (từ 2 đến 8).
4. Quan sát kết quả phân cụm + điểm đánh giá.

Kết quả:

* Elbow cho thấy k=5 là hợp lý.
* Silhouette Score = 0.554 (phân cụm tốt).
* Biểu đồ scatter cho thấy cụm phân bố rõ ràng.

A screenshot of a computer

AI-generated content may be incorrect.

A chart with many colored dots

AI-generated content may be incorrect.

A colorful pie chart with numbers

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

## 4.2. Kết luận

Những gì đã đạt được:

* Ứng dụng web hoàn chỉnh, dễ sử dụng.
* Phân loại khách hàng hiệu quả với mô hình KMeans.
* Có hỗ trợ đánh giá chất lượng phân cụm.

Những gì học được:

* Áp dụng học máy vào bài toán thực tế.
* Thiết kế giao diện người dùng đơn giản với Streamlit.
* Kỹ thuật trực quan hóa và đánh giá mô hình phân cụm.

Hướng cải tiến trong tương lai:

* Thêm mô hình phân cụm khác như DBSCAN, GMM.
* Cho phép gợi ý k tối ưu tự động.
* Nâng cấp UI để có dashboard quản lý.