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: DƯƠNG HOÀNG VŨ

  Lớp: K57KMT

Giáo viên giảng dạy: TS. NGUYỄN VĂN HUY

Link GitHub: https://github.com/hoangvu168/PhanLoaiKhachHang\_Mall

**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:Dương Hoàng Vũ*

*Lớp*: *K57KMT*……………………. *Ngành: Kỹ thuật máy tính*

*Giáo viên hướng dẫn:Nguyễn Văn Huy*

*Ngày giao đề:20-5-2025*  *Ngày hoàn thành:30-5-2025*

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

*Yêu cầu :* Dữ liệu từ: 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** 7](#_Toc199371133)

[**1.1. Mô tả bài toán** 7](#_Toc199371134)

[**1.2. Tính năng của chương trình** 8](#_Toc199371135)

[**1.3. Thách thức trong quá trình thực hiện** 8](#_Toc199371136)

[**1.4. Kiến thức vận dụng để làm bài** 9](#_Toc199371137)

[**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT** 11](#_Toc199371138)

[**2.1 Các thư viện và công cụ sử dụng trong chương trình** 11](#_Toc199371139)

[**2.1.1. Thư viện Pandas** 11](#_Toc199371140)

[**2.1.2. Thư viện NumPy** 11](#_Toc199371141)

[**2.1.3. Thư viện Scikit-learn** 11](#_Toc199371142)

[**2.1.4. Thư viện Seaborn và Matplotlib** 12](#_Toc199371143)

[**2.1.5. Thư viện Streamlit** 12](#_Toc199371144)

[**2.2. Cơ sở lý thuyết về thuật toán KMeans** 12](#_Toc199371145)

[**2.3. Xử lý dữ liệu trước khi phân cụm (Data Preprocessing)** 13](#_Toc199371146)

[**2.4. Trực quan hóa dữ liệu** 13](#_Toc199371147)

[**2.5. Phát triển giao diện bằng Streamlit** 14](#_Toc199371148)

[**CHƯƠNG 3: THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH** 15](#_Toc199371149)

[**3.1. Sơ đồ khối hệ thống** 15](#_Toc199371150)

[**3.1.1 Các module chính của hệ thống:** 15](#_Toc199371151)

[**3.1.2. Mô tả các thuật toán chính:** 16](#_Toc199371152)

[**3.3. Cấu trúc dữ liệu** 17](#_Toc199371153)

[**3.4. Chương trình** 17](#_Toc199371154)

[**3.4.1. Các hàm chính trong chương trình:** 17](#_Toc199371155)

[**3.4.2. Một số ví dụ gọi hàm:** 18](#_Toc199371156)

[**CHƯƠNG 4: THỰC NGHIỆM VÀ KẾT LUẬN** 19](#_Toc199371157)

[**4.1 Thực nghiệm*4.1.1 Môi trường thực nghiệm*** 19](#_Toc199371158)

[***4.1.2 Các bước thực nghiệm*** 19](#_Toc199371159)

[***4.1.3 Hình ảnh minh họa kết quả:*** 20](#_Toc199371160)

[**4.2. Kết luận** 25](#_Toc199371161)

[**4.2.1. Sản phẩm đã thực hiện được:** 25](#_Toc199371162)

[**4.2.2. Kiến thức và kỹ năng đạt được:** 25](#_Toc199371163)

[**4.2.3. Hướng cải tiến trong tương lai:** 25](#_Toc199371164)

**LỜI CẢM ƠN**

Em xin chân thành cảm ơn **Thầy Nguyễn Văn Huy** – giảng viên môn *Khoa học dữ liệu* – người đã tận tình giảng dạy, hướng dẫn và hỗ trợ em trong suốt quá trình học tập và thực hiện bài tập này. Những kiến thức mà thầy truyền đạt không chỉ giúp em hoàn thành tốt bài tập mà còn là hành trang quý báu cho chặng đường phát triển sau này.

# LỜI NÓI ĐẦU

Trong thời đại bùng nổ dữ liệu hiện nay, việc thu thập và khai thác thông tin từ dữ liệu không còn là lợi thế mà đã trở thành nhu cầu tất yếu đối với các tổ chức và doanh nghiệp. Môn học *Khoa học dữ liệu* đã cung cấp cho sinh viên những kiến thức nền tảng và kỹ năng thực tiễn để phân tích, xử lý và rút trích giá trị từ dữ liệu.

Bài tập 10 với đề tài **“Phân loại khách hàng tiềm năng”** là một ứng dụng thiết thực của Khoa học dữ liệu trong lĩnh vực marketing và quản trị khách hàng. Thông qua việc sử dụng bộ dữ liệu *Mall Customer Segmentation Data*, nhóm đã xây dựng một **ứng dụng web** hỗ trợ phân tích, phân loại khách hàng dựa trên thông tin giao dịch, từ đó đưa ra những biểu đồ trực quan giúp doanh nghiệp nhận diện được các nhóm khách hàng mục tiêu.

Bài tập không chỉ giúp em củng cố kiến thức về tiền xử lý dữ liệu, thuật toán phân cụm (clustering), mà còn rèn luyện kỹ năng triển khai mô hình học máy trên nền tảng web, đáp ứng nhu cầu thực tiễn trong các hệ thống hỗ trợ ra quyết định.

CHƯƠNG I: GIỚI THIỆU ĐẦU BÀI

**1.1. Mô tả bài toán**

Bài toán đặt ra là xây dựng một **ứng dụng web** có khả năng **phân loại khách hàng tiềm năng** dựa trên dữ liệu giao dịch mua sắm tại một trung tâm thương mại (Mall).  
 Dữ liệu sử dụng là **“Mall Customer Segmentation Data”**, trong đó mỗi khách hàng có các thuộc tính: Mã khách hàng, Giới tính, Tuổi, Thu nhập hàng năm (nghìn đô), và Điểm chi tiêu (Spending Score từ 1 đến 100)

**Mục tiêu chính**:

* Phân nhóm khách hàng thành các cụm có hành vi tiêu dùng tương đồng.
* Hiển thị kết quả phân cụm thông qua bảng dữ liệu và các biểu đồ trực quan.
* Cho phép người dùng lọc, xem và tải xuống dữ liệu đã phân cụm.

**Các bước xử lý gồm:**

1. Đọc và xử lý dữ liệu đầu vào từ file CSV.

2. Tiền xử lý dữ liệu: Đổi tên cột, chọn các cột đặc trưng cần phân cụm.

3. Áp dụng thuật toán **KMeans** để phân loại khách hàng thành 5 cụm.

4. Lưu kết quả phân cụm vào file CSV.

5. Xây dựng giao diện ứng dụng web với **Streamlit**, hỗ trợ lọc dữ liệu và vẽ các biểu đồ phân tích:

* Biểu đồ phân nhóm khách hàng theo thu nhập và điểm chi tiêu
* Số lượng khách hàng theo cụm
* Biểu đồ phân cụm khách hàng theo Thu nhập và Điểm chi tiêu
* Biểu đồ phân bố số lượng khách theo từng cụm
* Cho phép tải xuống dữ liệu phân cụm sau lọc.

**1.2. Tính năng của chương trình**

Ứng dụng web được xây dựng nhằm phân loại khách hàng từ dữ liệu giao dịch tại trung tâm thương mại với nhiều tính năng nổi bật, cụ thể như sau:

**Phân loại khách hàng tự động bằng thuật toán KMeans**: Dựa trên các đặc trưng như độ tuổi, thu nhập và điểm chi tiêu, hệ thống sử dụng thuật toán học máy KMeans để tự động phân chia khách hàng thành các cụm có hành vi tiêu dùng tương đồng. Việc phân cụm này giúp doanh nghiệp nhận diện và xây dựng chiến lược tiếp cận phù hợp với từng nhóm khách hàng.

**Hiển thị dữ liệu phân cụm bằng biểu đồ trực quan**: Sau khi phân loại, hệ thống cung cấp nhiều dạng biểu đồ sinh động như scatter plot (biểu đồ phân tán 2 chiều), count plot (biểu đồ số lượng theo cụm), hist plot (biểu đồ phân phối thu nhập), và boxplot (biểu đồ hộp so sánh điểm chi tiêu). Những biểu đồ này giúp người dùng dễ dàng quan sát và phân tích dữ liệu đã phân cụm.

**Bộ lọc dữ liệu linh hoạt**: Giao diện web hỗ trợ bộ lọc thông minh, cho phép người dùng chọn lọc dữ liệu theo nhiều tiêu chí khác nhau như giới tính, cụm phân loại, độ tuổi và thu nhập hàng năm. Việc này giúp phân tích sâu hơn từng phân khúc khách hàng theo yêu cầu.

**Lưu và tải dữ liệu đã phân cụm**: Người dùng có thể dễ dàng tải dữ liệu đã phân cụm dưới định dạng tệp CSV. Dữ liệu tải về có thể được lọc trước đó, rất tiện lợi cho việc báo cáo hoặc phân tích chuyên sâu bằng các công cụ khác.

**Giao diện thân thiện và dễ sử dụng trên nền tảng web**: Ứng dụng được phát triển bằng thư viện Streamlit nên có thể triển khai nhanh chóng dưới dạng web app. Người dùng chỉ cần trình duyệt để truy cập và không cần cài đặt phần mềm phức tạp. Giao diện đơn giản, dễ sử dụng phù hợp với cả người không chuyên về lập trình.

**1.3. Thách thức trong quá trình thực hiện**

Trong quá trình thực hiện bài toán, một số khó khăn và thách thức được ghi nhận như:

* **Lựa chọn đặc trưng phù hợp** để đưa vào thuật toán KMeans nhằm đảm bảo phân cụm có ý nghĩa thực tiễn.
* **Xác định số lượng cụm (K)** hợp lý mà không có gợi ý từ trước. Việc lựa chọn K = 5 là một quyết định mang tính thực nghiệm dựa trên quan sát biểu đồ.
* **Làm việc với dữ liệu thực tế**, cần đổi tên cột, chuẩn hóa dữ liệu đầu vào và xử lý ngôn ngữ hiển thị (từ tiếng Anh sang tiếng Việt).
* **Thiết kế giao diện trực quan, dễ sử dụng** bằng thư viện Streamlit trong khi vẫn đảm bảo đầy đủ các chức năng phân tích.
* **Lưu file và tải dữ liệu lọc theo yêu cầu người dùng** – xử lý encoding tiếng Việt để tải file chính xác.

**1.4. Kiến thức vận dụng để làm bài**

Để xây dựng và hoàn thiện ứng dụng phân loại khách hàng này, nhóm đã vận dụng tổng hợp nhiều kiến thức lý thuyết và công nghệ thực hành từ các môn học và tài liệu tham khảo. Cụ thể:

* **Kiến thức về Học máy (Machine Learning)**: Bài toán áp dụng thuật toán KMeans, một phương pháp phân cụm không giám sát phổ biến, để chia khách hàng thành các nhóm dựa trên hành vi tiêu dùng. Kiến thức về vector đặc trưng, khoảng cách Euclidean và quá trình lặp trong KMeans đã được áp dụng hiệu quả.
* **Xử lý và trực quan hóa dữ liệu**: Sử dụng thư viện Pandas để đọc và xử lý dữ liệu, cùng với Seaborn và Matplotlib để hiển thị biểu đồ. Việc lựa chọn loại biểu đồ phù hợp cho từng loại thông tin (phân cụm, phân phối, so sánh cụm) thể hiện rõ hiểu biết về trực quan hóa dữ liệu trong thực tế.
* **Phát triển ứng dụng web với Streamlit**: Streamlit là công cụ mạnh mẽ để nhanh chóng triển khai ứng dụng học máy dưới dạng giao diện web. Nhóm đã sử dụng các thành phần như **st.selectbox, st.slider, st.dataframe, st.download\_button** để xây dựng hệ thống lọc dữ liệu, hiển thị bảng và cho phép người dùng tải kết quả.
* **Làm việc với tệp và encoding**: Việc lưu và tải dữ liệu tiếng Việt yêu cầu xử lý đúng encoding UTF-8, tránh lỗi font khi mở tệp trong Excel hoặc trình soạn thảo khác. Nhóm đã áp dụng kiến thức xử lý tệp văn bản để đảm bảo tính toàn vẹn dữ liệu sau tải về.

*Tóm tắt chương:*

*Chương này đã trình bày tổng quan về bài toán phân loại khách hàng từ dữ liệu Mall. Ứng dụng web được xây dựng có khả năng phân nhóm khách hàng theo hành vi chi tiêu bằng thuật toán KMeans, đồng thời cung cấp giao diện thân thiện với đầy đủ biểu đồ phân tích và khả năng tải kết quả. Việc hoàn thiện ứng dụng giúp người học nắm vững kỹ thuật xử lý dữ liệu, phân cụm khách hàng và trình bày kết quả một cách khoa học, trực quan.*

**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT**

**2.1 Các thư viện và công cụ sử dụng trong chương trình**

Trong quá trình xây dựng ứng dụng phân loại khách hàng sử dụng thuật toán KMeans, nhóm đã vận dụng nhiều thư viện lập trình và công cụ hỗ trợ trong ngôn ngữ Python, cụ thể như sau:

**2.1.1. Thư viện Pandas**

Pandas là một thư viện mã nguồn mở trong Python được sử dụng để thao tác và phân tích dữ liệu. Pandas cung cấp các cấu trúc dữ liệu linh hoạt như **DataFrame** và **Series** giúp dễ dàng thao tác dữ liệu dưới dạng bảng tính (tương tự Excel). Trong chương trình, Pandas được dùng để:

* Đọc và xử lý dữ liệu từ file CSV.
* Thực hiện các thao tác lọc, nhóm, thay đổi giá trị cột, đổi tên cột, chuẩn hóa dữ liệu.
* Truy xuất các thống kê cơ bản như trung bình, min, max, số lượng theo nhóm, v.v.

**2.1.2. Thư viện NumPy**

NumPy là thư viện hỗ trợ xử lý số học và mảng đa chiều. Trong chương trình, NumPy được sử dụng hỗ trợ cho các phép tính toán ma trận và xử lý dữ liệu trước khi đưa vào mô hình học máy.

**2.1.3. Thư viện Scikit-learn**

Scikit-learn là một trong những thư viện mạnh mẽ nhất trong Python dành cho học máy (machine learning). Trong bài toán này, nhóm sử dụng Scikit-learn để triển khai thuật toán **KMeans Clustering**. Các chức năng chính được sử dụng từ thư viện này bao gồm:

* **KMeans()**: thực hiện phân cụm không giám sát trên tập dữ liệu.
* **fit()** và **predict()**: huấn luyện mô hình và dự đoán nhãn cụm cho mỗi dữ liệu.
* **StandardScaler**: Chuẩn hóa dữ liệu trước khi đưa vào thuật toán để tăng độ chính xác.

**2.1.4. Thư viện Seaborn và Matplotlib**

Đây là hai thư viện hỗ trợ trực quan hóa dữ liệu (data visualization). Chúng được dùng để tạo biểu đồ thể hiện kết quả phân cụm và phân tích hành vi khách hàng theo từng đặc trưng:

* **Biểu đồ phân tán (scatterplot)** giúp trực quan hoá sự phân bố các cụm khách hàng.
* **Biểu đồ boxplot** để so sánh các cụm theo điểm chi tiêu.
* **Histogram** hiển thị phân phối thu nhập.
* **Countplot** thể hiện số lượng khách hàng ở mỗi cụm.

**2.1.5. Thư viện Streamlit**

Streamlit là một thư viện mã nguồn mở dùng để xây dựng ứng dụng web một cách nhanh chóng. Thư viện này giúp lập trình viên có thể dễ dàng triển khai các ứng dụng dữ liệu mà không cần kiến thức về phát triển web truyền thống. Trong bài toán này, Streamlit được sử dụng để:

* Xây dựng giao diện web thân thiện với người dùng.
* Tạo các thành phần tương tác như bảng chọn, nút tải dữ liệu, khung nhập liệu.
* Hiển thị bảng dữ liệu, biểu đồ và kết quả phân cụm trực tiếp trên web.

**2.2. Cơ sở lý thuyết về thuật toán KMeans**

Thuật toán **KMeans** là một thuật toán học máy không giám sát (unsupervised learning), được sử dụng để giải bài toán phân cụm (clustering). Mục tiêu của KMeans là chia tập dữ liệu thành **K cụm (clusters)** sao cho các điểm dữ liệu trong cùng một cụm có độ tương đồng cao nhất, trong khi các cụm khác biệt nhau nhiều nhất.

**Nguyên lý hoạt động của KMeans:**

1. Chọn số cụm **K** cần phân chia.
2. Khởi tạo **K điểm tâm cụm (centroids)** một cách ngẫu nhiên.
3. Phân phối từng điểm dữ liệu đến cụm có tâm gần nhất (dựa vào khoảng cách Euclidean).
4. Tính lại tâm cụm mới bằng cách lấy trung bình các điểm trong cụm.
5. Lặp lại bước 3 và 4 cho đến khi các tâm cụm hội tụ (không thay đổi hoặc thay đổi rất nhỏ).

**Lợi ích khi sử dụng KMeans trong bài toán:**

* Dễ cài đặt và thực hiện.
* Phù hợp với bài toán phân nhóm khách hàng dựa trên đặc trưng tiêu dùng như thu nhập và điểm chi tiêu.
* Có thể mở rộng và cải tiến bằng các biến thể khác như MiniBatch KMeans, KMeans++.

**2.3. Xử lý dữ liệu trước khi phân cụm (Data Preprocessing)**

Để đảm bảo hiệu quả khi áp dụng thuật toán, dữ liệu cần được xử lý trước, bao gồm:

* **Làm sạch dữ liệu (Data Cleaning)**: Kiểm tra và xử lý các giá trị thiếu hoặc không hợp lệ.
* **Chọn lọc đặc trưng**: Lựa chọn những cột đặc trưng quan trọng (tuổi, thu nhập, điểm chi tiêu) để đưa vào mô hình.
* **Chuẩn hóa dữ liệu (Standardization)**: Sử dụng StandardScaler để chuẩn hóa các đặc trưng về cùng một thang đo (mean = 0, std = 1), giúp thuật toán hoạt động hiệu quả hơn.
* **Gán nhãn cụm sau phân loại**: Sau khi mô hình huấn luyện, mỗi điểm dữ liệu được gán nhãn cụm tương ứng.

**2.4. Trực quan hóa dữ liệu**

Việc trực quan hóa dữ liệu sau khi phân cụm giúp dễ dàng hiểu được hành vi khách hàng trong từng cụm. Một số dạng biểu đồ được sử dụng bao gồm:

* **Biểu đồ phân tán 2D** thể hiện sự phân bố khách hàng theo cụm.
* **Boxplot** so sánh đặc trưng (ví dụ: điểm chi tiêu) giữa các cụm.
* **Histogram** thể hiện phân phối thu nhập.
* **Biểu đồ cột (countplot)** hiển thị số lượng khách hàng ở mỗi cụm.

Trực quan hóa giúp người dùng hoặc nhà quản lý nhanh chóng xác định được nhóm khách hàng tiềm năng, nhóm chi tiêu nhiều, nhóm có thu nhập cao, v.v.

**2.5. Phát triển giao diện bằng Streamlit**

Streamlit là thư viện được sử dụng để hiển thị toàn bộ dữ liệu, biểu đồ và kết quả mô hình lên giao diện web. Một số thành phần tương tác trên giao diện:

* **Selectbox / Multiselect**: Chọn lọc dữ liệu theo cụm, giới tính, độ tuổi.
* **Slider**: Lọc dữ liệu theo khoảng thu nhập.
* **DataFrame hiển thị dữ liệu phân cụm**.
* **Nút tải dữ liệu** giúp người dùng lưu kết quả dưới dạng CSV.

Streamlit giúp việc chia sẻ và vận hành ứng dụng trở nên đơn giản mà không cần xây dựng backend phức tạp.

*Tóm tắt chương:*

*Trong chương này, đã trình bày các kiến thức lý thuyết và công cụ đã được áp dụng trong chương trình. Bao gồm: các thư viện Pandas, Numpy, Scikit-learn, Seaborn, Matplotlib và Streamlit; cùng với lý thuyết về thuật toán KMeans – một phương pháp phân cụm phổ biến trong học máy. Ngoài ra, chương cũng đề cập đến các kỹ thuật xử lý dữ liệu, chuẩn hóa, và trực quan hóa nhằm giúp người dùng dễ dàng nhận diện các phân khúc khách hàng dựa trên hành vi tiêu dùng. Những cơ sở lý thuyết này đóng vai trò nền tảng vững chắc cho quá trình xây dựng ứng dụng thực tế trong các chương sau*.

**CHƯƠNG 3: THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH**

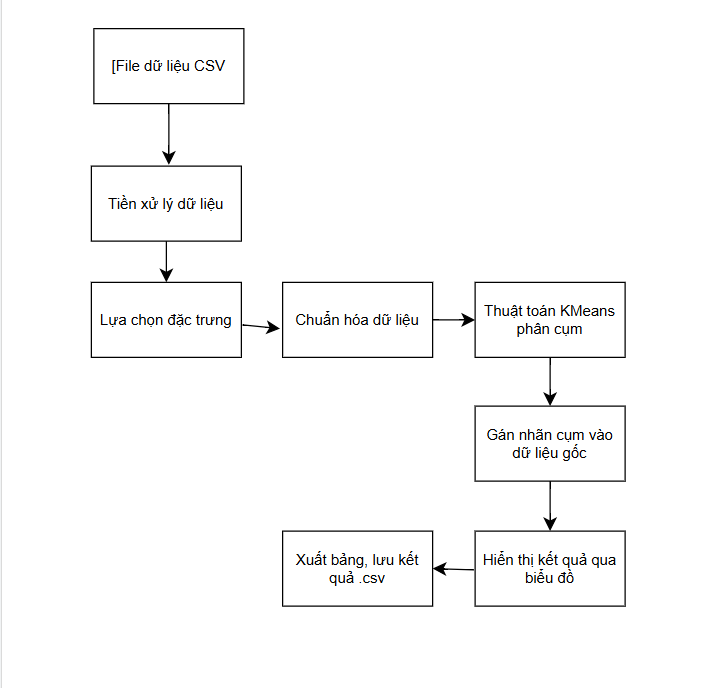
**3.1. Sơ đồ khối hệ thống**

Chương trình được chia thành nhiều module chức năng, mỗi module đảm nhận một nhiệm vụ cụ thể trong quá trình thu thập, xử lý, phân tích và hiển thị dữ liệu khách hàng. Các module chính gồm:

**3.1.1 Các module chính của hệ thống:**

1. **Module tải dữ liệu (Data Loading):**
   * Chức năng: Đọc dữ liệu từ file CSV chứa thông tin khách hàng.
   * Công cụ: pandas.read\_csv().
2. **Module xử lý dữ liệu (Data Preprocessing):**
   * Làm sạch dữ liệu, chuẩn hóa, chọn lọc đặc trưng đầu vào.
3. **Module phân cụm (KMeans Clustering):**
   * Áp dụng thuật toán KMeans để phân nhóm khách hàng.
4. **Module phân tích và trực quan hóa (Visualization):**
   * Hiển thị các cụm và mối quan hệ giữa các đặc trưng.
5. **Module giao diện (Streamlit UI):**
   * Tạo giao diện người dùng để nhập, chọn và xem kết quả.

Biểu đồ phân cấp chức năng



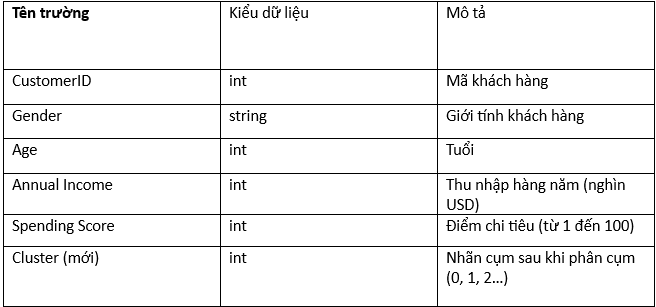
**3.1.2. Mô tả các thuật toán chính:**

* **KMeans Clustering:**
  + **Đầu vào:** Tập dữ liệu đã chuẩn hóa với các đặc trưng như Thu nhập, Điểm chi tiêu.
  + **Xử lý:** Phân nhóm khách hàng thành K cụm dựa trên khoảng cách Euclidean.
  + **Đầu ra:** Nhãn cụm cho mỗi khách hàng, biểu thị họ thuộc nhóm tiêu dùng nào.
* **Chuẩn hóa dữ liệu:**
  + Sử dụng StandardScaler() của Scikit-learn để đảm bảo mỗi cột đặc trưng có trung bình 0, phương sai 1.
* **Trực quan hóa:**
  + Dùng biểu đồ 2D để thể hiện cụm.
  + Dùng biểu đồ cột, boxplot để phân tích sâu từng đặc điểm cụ thể theo từng cụm.

**3.3. Cấu trúc dữ liệu**

Chương trình sử dụng tập dữ liệu đầu vào dưới dạng bảng CSV, trong đó mỗi hàng đại diện cho một khách hàng, các cột là thông tin đặc trưng.

**a. Bảng dữ liệu khách hàng (Mall\_Customers.csv):**



**b. Dữ liệu sau khi xử lý:**

Sau khi xử lý, dữ liệu có thể có thêm các cột như:

* Cluster: nhãn cụm phân nhóm khách hàng.
* Scaled Income, Scaled Score: các đặc trưng đã chuẩn hóa.

**3.4. Chương trình**

**3.4.1. Các hàm chính trong chương trình:**

1. **load\_data()**
   * Đọc dữ liệu từ file CSV.
   * Trả về DataFrame gốc.
2. **preprocess\_data(df)**
   * Làm sạch, chuẩn hóa, chọn đặc trưng.
   * Trả về dữ liệu đã sẵn sàng cho mô hình.
3. **run\_kmeans(df, n\_clusters)**
   * Áp dụng thuật toán KMeans.
   * Trả về nhãn cụm, mô hình đã huấn luyện.
4. **visualize\_clusters(df)**
   * Vẽ biểu đồ phân cụm.
   * Hiển thị biểu đồ phân tích các đặc trưng theo từng cụm.
5. **main()**
   * Hàm chính chạy Streamlit UI.
   * Gọi tất cả các hàm trên để tải dữ liệu, xử lý, phân cụm và hiển thị.

**3.4.2. Một số ví dụ gọi hàm:**

python

Sao chép mã

df = load\_data()

df\_processed = preprocess\_data(df)

labels, model = run\_kmeans(df\_processed, n\_clusters=5)

df["Cluster"] = labels

visualize\_clusters(df)

*Tóm tắt chương:*

*Chương 3 đã trình bày chi tiết về kiến trúc và thiết kế chương trình, từ cấu trúc module, biểu đồ phân cấp chức năng, đến sơ đồ khối thuật toán chính được sử dụng. Các thành phần dữ liệu và hàm xử lý cũng được mô tả cụ thể để đảm bảo tính hệ thống và dễ bảo trì. Phần thiết kế chặt chẽ này là nền tảng giúp chương trình hoạt động chính xác và hiệu quả trong việc phân cụm khách hàng và trực quan hóa kết quả trên giao diện web.*

**CHƯƠNG 4: THỰC NGHIỆM VÀ KẾT LUẬN**

**4.1 Thực nghiệm**

***4.1.1 Môi trường thực nghiệm***

**Ngôn ngữ lập trình:** Python 3.10

**Thư viện sử dụng:** Pandas, Scikit-learn, Matplotlib, Seaborn

**Dữ liệu đầu vào:** File CSV chứa thông tin về [nêu tên bộ dữ liệu, ví dụ: bệnh nhân, điểm thi, đặc điểm sản phẩm,..

**4.1.2 Các bước thực nghiệm**

**Đọc và tiền xử lý dữ liệu:**

* Kiểm tra và xử lý giá trị khuyết (NaN)
* Chuẩn hóa dữ liệu bằng MinMaxScaler / StandardScaler
* Chọn các đặc trưng quan trọng để phân cụm

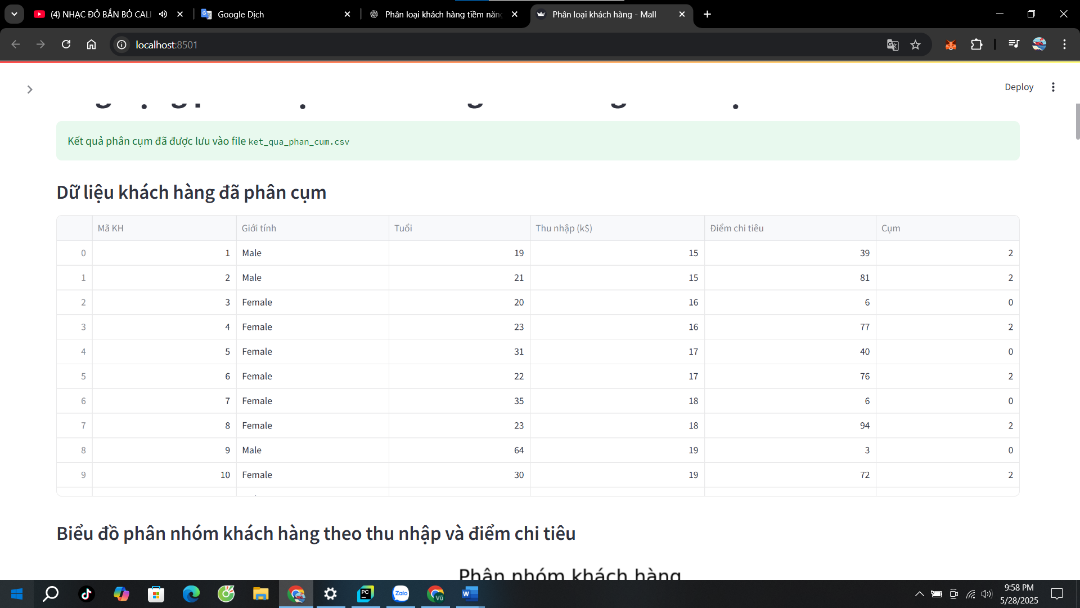
**Áp dụng thuật toán KMeans:**

* Chạy KMeans với số cụm k=nk = nk=n (ví dụ: k = 3)
* Gán nhãn cụm vào dữ liệu gốc
* Trực quan hóa kết quả bằng biểu đồ Scatter, Elbow Method

**Ghi nhận kết quả:**

* Kết quả phân cụm thể hiện rõ sự khác biệt giữa các nhóm
* Biểu đồ minh họa kết quả trực quan, dễ hiểu
* Xuất kết quả ra file .csv gồm dữ liệu gốc + nhãn cụm mới

**4.1.3 Hình ảnh minh họa kết quả:**

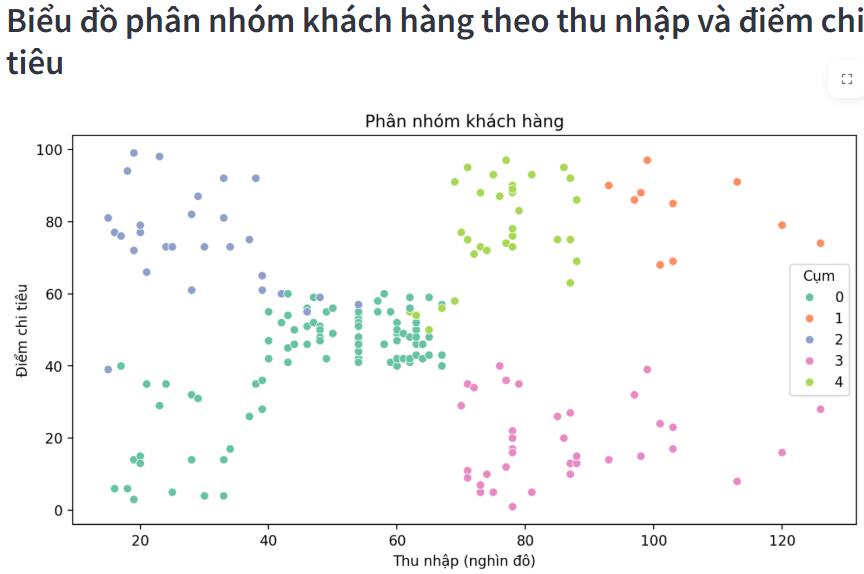


Ứng dụng có khả năng **hiển thị bảng dữ liệu khách hàng** sau khi đã được phân cụm bằng thuật toán **KMeans**. Kết quả phân cụm sẽ được lưu vào file ket\_qua\_phan\_cum.csv và thể hiện ngay trên giao diện chính của ứng dụng.

**Tính năng lưu kết quả:**

Sau khi phân cụm, ứng dụng tự động lưu kết quả ra file .csv giúp người dùng:

* Tiện lợi trong việc theo dõi hoặc tải về để phân tích sâu hơn.
* Dễ dàng chia sẻ với bộ phận khác như marketing hoặc quản lý bán hàng

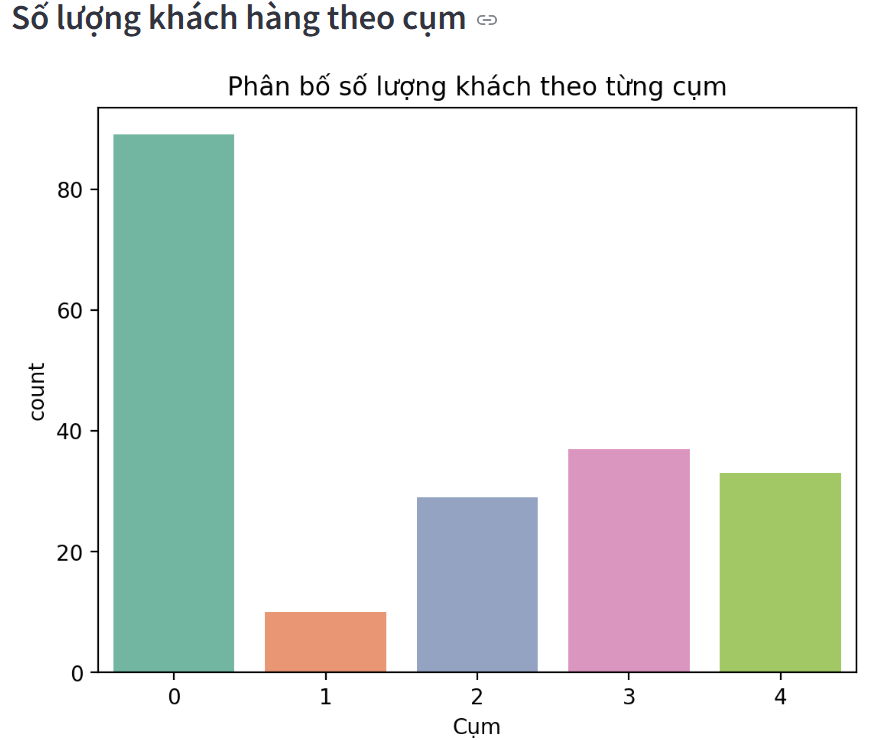


Hình 4.2 – Biểu đồ phân cụm khách hàng theo Thu nhập và Điểm chi tiêu

**Tính năng:**

* Giúp trực quan hóa việc **phân nhóm khách hàng** dựa trên **thu nhập** và **điểm chi tiêu**.
* Sử dụng thuật toán **KMeans** để gom các khách hàng có hành vi tiêu dùng tương đồng vào cùng một cụm.
* Hỗ trợ nhà quản lý nhận diện các nhóm như:
  + Khách hàng thu nhập cao – chi tiêu thấp
  + Khách hàng thu nhập thấp – chi tiêu cao
  + Hoặc các nhóm tiềm năng khá

**Tính năng chính** là giúp doanh nghiệp dễ dàng xây dựng **chiến lược tiếp cận riêng cho từng nhóm khách hàng**.

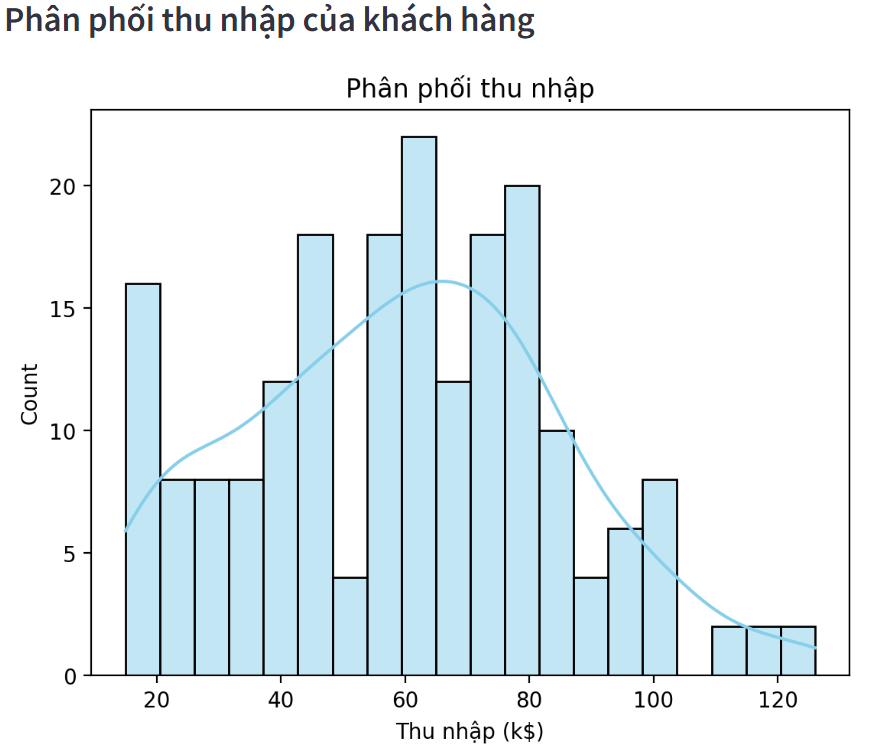
****

**Hình 4.3 – Biểu đồ phân bố số lượng khách theo từng cụm**

**Tính năng:**

* Hiển thị **số lượng khách hàng** thuộc từng cụm sau khi phân nhóm.
* Giúp đánh giá **kích thước từng phân khúc khách hàng**.
* Hỗ trợ đưa ra quyết định như:
  + Tập trung nguồn lực cho cụm đông khách
  + Tìm hiểu lý do cụm ít khách

**Tính năng chính** là cung cấp **cái nhìn tổng quan về tỉ lệ khách hàng trong từng nhóm**, từ đó tối ưu hoạt động marketing và chăm sóc khách hàng.

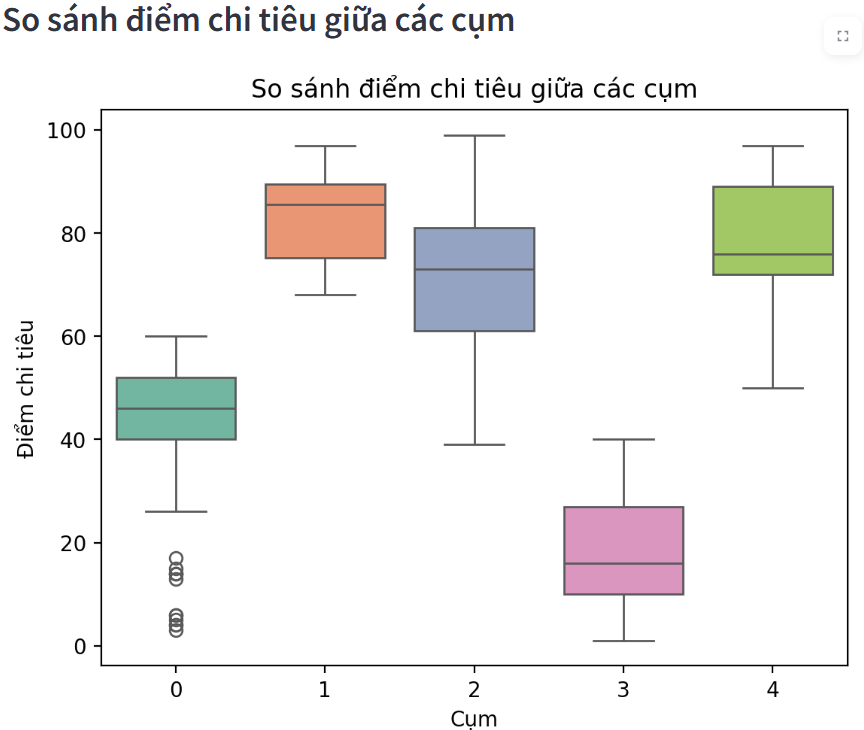


Hình 4.4 Biểu đồ Phân phối thu nhập của khách hàng

**Tính năng:**

* Thể hiện **mức độ phân bố thu nhập** của toàn bộ khách hàng trong tập dữ liệu.
* Giúp nhận diện:
  + Nhóm khách có thu nhập thấp, trung bình, cao
  + Mức độ tập trung thu nhập (thu nhập khách hàng phân bố đều hay bị lệch về một phía)
* Hỗ trợ doanh nghiệp đánh giá được **đối tượng khách hàng mục tiêu** phù hợp với sản phẩm/dịch vụ của mình.

**Tính năng chính:** Phân tích mức thu nhập chung của khách hàng để xây dựng chiến lược giá phù hợp.



Hình 4.5 Biểu đồ So sánh điểm chi tiêu giữa các cụm

**Tính năng:**

* So sánh **mức độ chi tiêu** trung bình giữa các cụm khách hàng đã được phân nhóm.
* Cho thấy:
  + Cụm nào có hành vi chi tiêu cao => nhóm khách tiềm năng
  + Cụm nào chi tiêu thấp => cần chiến lược kích cầu
* Kết hợp với dữ liệu thu nhập để đánh giá mức độ **chi tiêu hợp lý** theo thu nhập từng nhóm.

**Tính năng chính:** Hỗ trợ xác định **cụm khách hàng giá trị cao**, từ đó đề xuất ưu đãi và chăm sóc đặc biệt cho nhóm này.

**Ý nghĩa thực tiễn**:

* Giúp **doanh nghiệp** xác định cụm khách hàng nào là nhóm **tiêu dùng cao**, tiềm năng để chăm sóc, ưu đãi hoặc marketing.
* Phát hiện các cụm có **điểm chi tiêu thấp** để nghiên cứu thêm hoặc điều chỉnh chính sách kinh doanh
* **Dễ hiểu cho người dùng**:
  + Biểu đồ cột (hoặc biểu đồ hộp nếu dùng Boxplot) dễ nhìn, dễ so sánh trực quan.
  + Mỗi cụm được đánh số hoặc đặt tên rõ ràng, giúp người dùng không chuyên về kỹ thuật vẫn hiểu được nội dung.

**Hỗ trợ ra quyết định**:

* Nhờ biểu đồ này, doanh nghiệp có thể **tập trung nguồn lực** vào những nhóm khách hàng mang lại giá trị cao.

Biểu đồ so sánh điểm chi tiêu giữa các cụm là một công cụ trực quan giúp làm rõ sự khác biệt trong hành vi tiêu dùng giữa các nhóm khách hàng. Từ biểu đồ, ta có thể dễ dàng xác định được cụm khách hàng có xu hướng chi tiêu cao, từ đó định hướng các chiến lược chăm sóc và tiếp thị phù hợp. Đây là tính năng quan trọng hỗ trợ doanh nghiệp trong quá trình phân loại và tối ưu hóa giá trị khách hàng

**4.2. Kết luận**

**4.2.1. Sản phẩm đã thực hiện được:**

* Đọc và xử lý dữ liệu từ file CSV
* Lựa chọn và chuẩn hóa đặc trưng
* Áp dụng thuật toán KMeans để phân cụm dữ liệu
* Gán nhãn cụm và trực quan hóa kết quả bằng biểu đồ
* Xuất kết quả ra file .csv để lưu trữ hoặc tiếp tục phân tích

**4.2.2. Kiến thức và kỹ năng đạt được:**

* Hiểu rõ hơn về quy trình xử lý dữ liệu và thuật toán KMeans
* Biết cách sử dụng thư viện Python để phân tích và trực quan hóa dữ liệu
* Kỹ năng làm việc với file dữ liệu, thao tác tiền xử lý và đánh giá kết quả mô hình

**4.2.3. Hướng cải tiến trong tương lai:**

* Cho phép người dùng nhập số cụm (k) linh hoạt
* Thêm chức năng tự động chọn k tối ưu (dùng Elbow, Silhouette)
* Phát triển giao diện người dùng thân thiện (GUI)
* Hỗ trợ nhiều định dạng dữ liệu (JSON, Excel...)

TÀI LIỆU THAM KHẢO

<https://www.kaggle.com/datasets/shwetabh123/mall-customers>