|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC THỦY LỢI



HỌ VÀ TÊN: TRẦN VĂN SINH

**PHÂN CỤM KHÁCH HÀNG SỬ DỤNG THUẬT TOÁN K-MEANS**

ĐỒ ÁN TỐT NGHIỆP

HÀ NỘI, NĂM 2023

|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC THỦY LỢI

HỌ VÀ TÊN: TRẦN VĂN SINH

**PHÂN CỤM KHÁCH HÀNG SỬ DỤNG THUẬT TOÁN K-MEANS**

|  |  |
| --- | --- |
| Ngành : | Hệ thống thông tin |
| Mã số : | 7480104 |

|  |  |
| --- | --- |
| NGƯỜI HƯỚNG DẪN | 1. TS. NGUYỄN TU TRUNG |

HÀ NỘI, NĂM 2023

**GÁY BÌA ĐỒ ÁN TỐT NGHIỆP, KHÓA LUẬN TỐT NGHIỆP**

**HỌ VÀ TÊN: TRẦN VĂN SINH ĐỒ ÁN/KL TỐT NGHIỆP HÀ NỘI, NĂM: 2023 ……**

|  |  |
| --- | --- |
|  | CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM  **Độc lập - Tự do - Hạnh phúc**  --------------------  **NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP** |

|  |  |
| --- | --- |
| Họ tên sinh viên: **Trần Văn Sinh** | Hệ đào tạo: **Đại học chính quy** |
| Lớp: **61HT** | Ngành: **Hệ thống thông tin** |
| Khoa: **Công nghệ thông tin** |  |

1. TÊN ĐỀ TÀI:

PHÂN CỤM KHÁCH HÀNG SỬ DỤNG THUẬT TOÁN K-MEANS

1. CÁC TÀI LIỆU CƠ BẢN:
2. T. M. Tuan, Bài giảng khai phá dữ liệu
3. D. T. Hien và N. D. Huy, Data Mining
4. https://trungkhithe.com/, [Trực tuyến]. Available: https://trungkhithe.com/weka-la-gi/.
5. L. D. Nhuong, 2014. [Trực tuyến]. Available:

https://123docz.net/document/1096864-khai-pha-du-lieu-dung-thuat-toan-k-mean-va-naive-bayes-tren-wave.htm

1. https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python
2. NỘI DUNG CÁC PHẦN THUYẾT MINH VÀ TÍNH TOÁN:

Chương 1: Tổng quan

Chương 2: Thuật toán K-Means và công nghệ sử dụng

Chương 3: Xây dựng chương trình phân cụm khách hàng

Chương 4: Thực nghiệm và đánh giá

1. GIÁO VIÊN HƯỚNG DẪN TỪNG PHẦN

|  |  |
| --- | --- |
| **Phần** | **Họ tên giáo viên hướng dẫn** |
| Chương 2. Thuật toán K-Means và công nghệ sử dụng | TS. Nguyễn Tu Trung |
| Chương 3. Xây dựng chương trình phân cụm khách hàng | TS. Nguyễn Tu Trung |

1. NGÀY GIAO NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP

Ngày .... tháng …. năm 2023.

|  |  |  |
| --- | --- | --- |
| **Trưởng Bộ môn**  *(Ký và ghi rõ Họ tên)* |  | **Giáo viên hướng dẫn chính**  *(Ký và ghi rõ Họ tên)* |

Nhiệm vụ Đồ án tốt nghiệp đã được Hội đồng thi tốt nghiệp của Khoa thông qua.

|  |
| --- |
| Ngày …. tháng …. năm 2023.  **Chủ tịch Hội đồng**  *(Ký và ghi rõ Họ tên)* |

Sinh viên đã hoàn thành và nộp bản Đồ án tốt nghiệp cho Hội đồng thi.

|  |
| --- |
| Ngày .... tháng …. năm 2023.  **Sinh viên làm Đồ án tốt nghiệp**  *(Ký và ghi rõ Họ tên)* |

|  |  |
| --- | --- |
| logo | TRƯỜNG ĐẠI HỌC THUỶ LỢI  **KHOA CÔNG NGHỆ THÔNG TIN**  BẢN TÓM TẮT ĐỀ CƯƠNG ĐỒ ÁN TỐT NGHIỆP |

TÊN ĐỀ TÀI: **PHÂN CỤM KHÁCH HÀNG SỬ DỤNG THUẬT TOÁN K-MEANS**

*Sinh viên thực hiện*: Trần Văn Sinh

*Lớp*: 61HT

*Mã sinh viên :* 1951060979

*Giáo viên hướng dẫn*: Nguyễn Tu Trung

**TÓM TẮT ĐỀ TÀI**

Hiện nay, công nghệ thông tin đóng một vai trò rất quan trọng trong đời sống của chúng ta. Sự bùng nổ của công nghệ thông tin cũng đã kéo theo sự phát triển mạnh mẽ của lĩnh vực Trí tuệ nhân tạo, đặc biệt trong vấn đề kinh doanh trên internet.

Do đó nên đề tài này em sử dụng thuật toán K-Means để phân chia khách hàng thành nhiều cụm/nhóm có chung sự tương đồng theo những tiêu chí như giới tính, tuổi tác, sở thích, thu nhập và thói quen chi tiêu, hành vi mua sắm,… để doanh nghiệp có phương thức tiếp thị hiệu quả. Khi thực hiện được phân cụm khách hàng giúp đơn vị giải quyết đúng các yêu cầu của từng khách hàng, giúp tăng lợi nhuận, giữ chân các khách hàng quan trọng, cũng như thực hiện các chiến dịch, chiến lược marketing hiệu quả hơn.

**CÔNG NGHỆ SỬ DỤNG**

Phần mềm : Visual Studio

Ngôn ngữ lập trình : C#

**CÁC MỤC TIÊU CHÍNH**

Lý thuyết

* Tổng quan về phương pháp phân cụm K-Means.
* Tìm hiểu chi tiết về phân cụm K-Means.
* Ứng dụng thuật toán K-Means vào bài toán phân cụm khách hàng.

Kỹ năng

* Sinh viên tìm hiểu kiến thức về phương pháp phân cụm K-Means
* Sinh viên ứng dụng kiến thức nắm được và sử dụng dữ liệu để phân cụm khách hàng. Từ đó đánh giá kết quả đầu ra giúp đơn vị giải quyết đúng các yêu cầu của từng khách hàng, giúp tăng lợi nhuận, giữ chân các khách hàng quan trọng, cũng như thực hiện các chiến dịch, chiến lược marketing hiệu quả hơn.

**KẾT QUẢ DỰ KIẾN**

Kiến thức

* Sinh viên hiểu được về những kiến thức tìm hiểu trong báo cáo.
* Sinh viên nắm rõ và sử dụng thuật toán phân cụm K-Means.

Kỹ năng

* Báo cáo tổng hợp về đề tài nghiên cứu
* Thu thập dữ liệu, tiền xử lí dữ liệu.
* Sinh viên ứng dụng cài đặt thành công chương trình trên ngôn ngữ lập trình C# với bài toán phân cụm khách hàng.
* Đánh giá kết quả đầu vào, đánh giá kết quả đầu ra của bài toán, đề xuất quan điểm và giải pháp nâng cao hiệu quả kinh doanh cho doanh nghiệp.

**LỜI CAM ĐOAN**

Tác giả xin cam đoan đây là Đồ án tốt nghiệp của bản thân tác giả. Các kết quả trong đồ án tốt nghiệp này là trung thực, và không sao chép từ bất kỳ một nguồn nào và dưới bất kỳ hình thức nào. Việc tham khảo các nguồn tài liệu (nếu có) đã được thực hiện trích dẫn và ghi nguồn tài liệu tham khảo đúng quy định.

|  |
| --- |
| **Tác giả ĐATN**  **Trần Văn Sinh** |

**LỜI CẢM ƠN**

Sau gần 4 năm học tập, rèn luyện tại trường Đại học Thủy Lợi nói chung và khoa Công nghệ thông tin nói riêng, em đã nhận được sự hướng dẫn và giúp đỡ của các thầy cô cùng với các bạn trong ngành học và cuộc sống.

Lời đầu tiên, em xin chân thành cảm ơn các giảng viên Trường đại học Thủy Lợi và đặc biệt là các thầy cô bộ môn khoa Công nghệ thông tin đã truyền dạy kiến thức cần thiết để giúp em thực hiện đồ án. Đặc biệt, trong thời gian làm đồ án, em đã được sự hướng dẫn tận tình của TS. Nguyễn Tu Trung . Em xin chân thành cảm ơn thầy đã giúp đỡ, bổ sung cho em những kiến thức và cho em những lời khuyên, gợi ý để em có thể hoàn thành đồ án một cách nhanh chóng và hiệu quả nhất.

Trong quá trình học tập và thực hiện đồ án em đã may mắn được sự chỉ bảo, hướng dẫn tận tình của các thầy cô giáo và được gia đình, bạn bè quan tâm, động viên, luôn ở bên và tạo mọi điều kiện thuận lợi để hoàn thành tốt đồ án này. Trong suốt quá trình làm đồ án với đề tài “**PHÂN CỤM KHÁCH HÀNG SỬ DỤNG THUẬT TOÁN K-MEANS**”, em đã nỗ lực hết sức để xây dựng và hoàn thiện đồ án một cách tốt nhất, nhưng do kiến thức còn hạn chế, thời gian hoàn thành đồ án có hạn và thiếu kinh nghiệm thực tế nên không thể tránh những sai sót. Một lần nữa, em xin chân thành cảm ơn thầy cô giáo, bạn bè và gia đình, những người đã giúp đỡ, ủng hộ em trong thời gian vừa qua.

**EM XIN CHÂN THÀNH CẢM ƠN**

*Sinh viên*

*Trần Văn Sinh*

MỤC LỤC

[LỜI MỞ ĐẦU 1](#_Toc138590445)

[DANH SÁCH CÁC HÌNH 2](#_Toc138590446)

[Chương I. Tổng quan 3](#_Toc138590447)

[1. Giới thiệu đề tài 3](#_Toc138590448)

[2. Tổng quan vể khai phá dữ liệu 4](#_Toc138590449)

[2.1. Khái niệm 4](#_Toc138590450)

[2.2. Quy trình khám phá tri thức trong CSDL 5](#_Toc138590451)

[2.3. Quy trình khai phá dữ liệu 5](#_Toc138590452)

[2.4. Ứng dụng của khai phá dữ liệu trong các lĩnh vực 6](#_Toc138590453)

[Chương II. Thuật toán phân cụm K-Means và công nghệ sử dụng 8](#_Toc138590454)

[1. Giới thiệu về kỹ thuật phân cụm trong khai phá dữ liệu 8](#_Toc138590455)

[2. Thuật toán K-Means 9](#_Toc138590456)

[2.1. Giới thiệu về K-Means 9](#_Toc138590457)

[2.2. Ưu điểm 9](#_Toc138590458)

[2.3. Hạn chế và thách thức: 9](#_Toc138590459)

[2.4. Ứng dụng của K-Means trong các lĩnh vực: 10](#_Toc138590460)

[3. Các chỉ số đo lường dữ liệu 10](#_Toc138590461)

[3.1. Các độ đo về xu hướng chính của dữ liệu 10](#_Toc138590462)

[3.2. Các độ đo về sự phân tán của dữ liệu 11](#_Toc138590463)

[3.3. Ý nghĩa 11](#_Toc138590464)

[4. Công nghệ và ngôn ngữ lập trình được sử dụng 11](#_Toc138590465)

[4.1. Về công nghệ : 11](#_Toc138590466)

[4.2. Về ngôn ngữ lập trình 13](#_Toc138590467)

[Chương III. Xây dựng chương trình phân cụm khách hàng 14](#_Toc138590468)

[1. Mô tả bài toán phân cụm khách hàng 14](#_Toc138590469)

[2. Chuẩn bị dữ liệu 15](#_Toc138590470)

[2.1. Thu thập dữ liệu 15](#_Toc138590471)

[2.2. Tiền xử lý dữ liệu bằng weka 15](#_Toc138590472)

[3. Xây dựng chương trình 21](#_Toc138590473)

[3.1. Xây dựng giao diện bằng Winform 21](#_Toc138590474)

[3.2. Code các module chương trình 21](#_Toc138590475)

[Chương IV. Thực nghiệm và đánh giá 29](#_Toc138590476)

[1. Phân cụm theo các thuộc tính của dữ liệu. 29](#_Toc138590477)

[2. Đánh giá kết quả phân cụm. 39](#_Toc138590478)

[KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 42](#_Toc138590479)

[TÀI LIỆU THAM KHẢO 43](#_Toc138590480)

# LỜI MỞ ĐẦU

Trong thời đại số hóa phát triển mạnh mẽ như hiện nay, việc hiểu dõ và phân loại khách hàng là một yếu tố quan trọng để đạt được thành công trong kinh doanh. Phân cụm khách hàng là quá trình nhóm các khách hàng có đặc điểm chung lại với nhau để nhận biết các nhóm khách hàng mục tiêu, nắm bắt nhu cầu của từng nhóm và phục vụ họ một cách tốt nhất.

Qua việc phân cụm khách hàng thì doanh nghiệp có thể nắm bắt được sự đa dạng của khách hàng và cung cấp các giải pháp phù hợp cho từng nhóm để từ đó tăng cường mối quan hệ, độ hài lòng và tin tưởng của khách hàng. Bên cạnh đó việc phân cụm khách hàng còn giúp cho doanh nghiệp tối ưu hóa chiến dịch tiếp thị, quản lý tài nguyên và định hình chiến lược kinh doanh.

Trong nghiên cứu này, luận văn đã tập trung vào việc phân cụm khách hàng và áp dụng thuật toán phân cụm K-Means để xác định nhóm khách hàng có đặc điểm tương tự nhau bằng cách tạo ra các nhóm khách hàng với file dữ liệu khách hàng có sẵn, em hy vọng rằng nghiên cứu này sẽ đóng góp vào việc nâng cao hiệu quả quản lý khách hàng và tăng cường khả năng tiếp cận và tương tác với khách hàng.

# DANH SÁCH CÁC HÌNH

Hình 1.1. Quy trình khai phá dữ liệu

Hình 1.2. Dữ liệu khách hàng thu thập trên Kaggle.com

Hình 1.3. Mở ứng dụng Weka

Hình 1.4. Mở file dữ liệu trên Weka

Hình 1.5. Xử lý dữ liệu bị thiếu

Hình 1.6. Chuyển dữ liệu sang kiểu số

Hình 1.7. File dữ liệu sau khi tiền xử lý với Weka

Hình 1.8. Giao diện chương trình phân cụm khách hàng

Hình 1.9. Kết quả phân cụm thuộc tính Spending và Income với k=2

Hình 2.1. Kết quả cụm 0 thuộc tính Spending và Income với k=2

Hình 2.2. Kết quả cụm 1 thuộc tính Spending và Income với k=2

Hình 2.3. Kết quả phân cụm thuộc tính Spending và Income với k=4

Hình 2.4. Kết quả cụm 0 thuộc tính Spending và Income với k=4

Hình 2.5. Kết quả cụm 1 thuộc tính Spending và Income với k=4

Hình 2.6. Kết quả cụm 2 thuộc tính Spending và Income với k=4

Hình 2.7. Kết quả cụm 3 thuộc tính Spending và Income với k=4

Hình 2.8. Kết quả phân cụm thuộc tính Age với k=2

Hình 2.9. Kết quả cụm 0 thuộc tính Age với k=2

Hình 3.1. Kết quả cụm 1 thuộc tính Age với k=2

Hình 3.2. Kết quả phân cụm thuộc tính Age với k=3

Hình 3.3. Kết quả cụm 0 thuộc tính Age với k=3

Hình 3.4. Kết quả cụm 1 thuộc tính Age với k=3

Hình 3.5. Kết quả cụm 2 thuộc tính Age với k=3

Hình 3.6. Các chỉ số đo lường dữ liệu

# Chương I. Tổng quan

## 1. Giới thiệu đề tài

Trong đề tài này, luận văn tập trung vào việc áp dụng thuật toán K-means để phân cụm khách hàng. K-means là một trong những phương pháp phân cụm phổ biến và mạnh mẽ, được sử dụng rộng rãi trong lĩnh vực khai phá dữ liệu và học máy.

Mục tiêu chính của nghiên cứu này là xác định các nhóm khách hàng có đặc điểm tương tự nhau bằng cách sử dụng dữ liệu có sẵn. Thuật toán K-means sẽ được áp dụng để tạo ra các cụm dữ liệu dựa trên sự tương đồng giữa các khách hàng. Quá trình này sẽ giúp chúng ta nhận biết được các nhóm khách hàng có nhu cầu, sở thích hoặc hành vi tiêu dùng tương tự.

Trong quá trình nghiên cứu, em sẽ thu thập và phân tích các thuộc tính quan trọng liên quan đến khách hàng như độ tuổi, giới tính, thu nhập, chi tiêu, … để xây dựng mô hình phân cụm. Sau đó áp dụng thuật toán K-means để phân cụm khách hàng thành các nhóm có đặc điểm tương đồng.

Kết quả từ nghiên cứu này sẽ mang lại những thông tin quan trọng về khách hàng, giúp cho doanh nghiệp hiểu dõ hơn về thị trường tiêu dùng, tạo ra các chiến lược tiếp thị phù hợp để tiếp cận đến khách hàng tốt hơn.

Hy vọng rằng nghiên cứu này sẽ đóng góp vào việc nâng cao hiệu quả quản lý khách hàng, tối ưu hóa chiến dịch tiếp thị và tăng cường sự cạnh tranh của các doanh nghiệp trên thị trường.

## 2. Tổng quan vể khai phá dữ liệu

### 2.1. Khái niệm

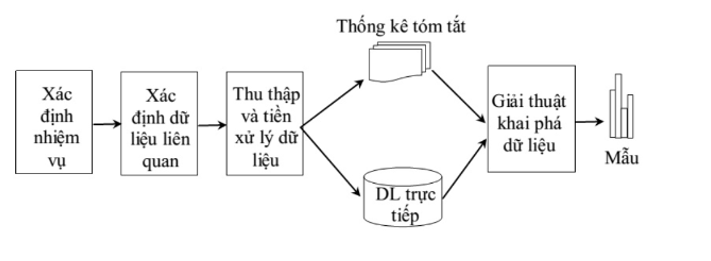
Khai phá dữ liệu là tiến trình khái quát các sự kiện rời rạc trong dữ liệu thành các tri thức mang tính khái quát, tính quy luật hỗ trợ tích cực cho các tiến trình ra quyết định. Khai phá dữ liệu là việc trích rút tri thức một cách tự động và hiệu quả từ một khối dữ liệu rất lớn.Tri thức đó thường ở dạng các mẫu tin có tính chất không tầm thường, không tường minh, chưa được biết đến và có tiềm năng mang lại lợi ích. [1]

Quá trình khai phá dữ liệu sẽ tiến hành qua các giai đoạn sau:

Bước 1: Lọc dữ liệu được thực hiện trong quá trình tiền xử lý. Công việc đầu tiên là tích hợp và chỉnh sửa dữ liệu. Khi dữ liệu được thu thập từ nhiều nguồn khác nhau nên có thể có những sai sót, dư thừa và trùng lặp. Lọc dữ liệu là cắt bỏ những dư thừa để dữ liệu được định dạng thống nhất. Dữ liệu sau khi lọc và chỉnh sửa sẽ nhỏ hơn, xử lý nhanh chóng hơn. [1]

Bước 2: Khai phá dữ liệu, là công việc chính sử dụng các thuật toán khác nhau để khai phá các kiến thức tiềm ẩn trong dữ liệu. [1]

Bước 3: Sau xử lý, là quá trình ước lượng kết quả khai phá theo yêu cầu của người dùng. Nhiều kỹ thuật khai phá được ứng dụng cho một nguồn dữ liệu, các kỹ thuật cho các kết quả có thể khác nhau. Các kết quả được ước lượng bởi những quy tắc nào đó, nếu cuối cùng kết quả không thỏa mãn yêu cầu, chúng ta phải làm lại với kỹ thuật khác cho đến khi có kết quả mong muốn. [1]



Hình 1.1. Quy trình khai phá dữ liệu

Phát hiện tri thức (Knowledge Discovery) trong các cơ sở dữ liệu là một quy trình nhận biết các mẫu hoặc các mô hình trong dữ liệu với các tính năng: hợp thức, mới, khả ích, và có thể hiểu được. [1]

Khai phá dữ liệu (Data mining) được định nghĩa như sau: “Data mining là một quá trình tìm kiếm, phát hiện các tri thức mới, tiềm ẩn, hữu dụng trong CSDL lớn”. [1]

Khai phá dữ liệu có thể được sử dụng cho các lĩnh vực y tế, phân tích thị trường, xây dựng ... có thể được xem như là kết quả của sự tiến triển tự nhiên của công nghệ thông tin. [1]

### 2.2. Quy trình khám phá tri thức trong CSDL

Quá trình khám phá tri thức bao gồm các bước:

+ Trích chọn dữ liệu: trích chọn những tập dữ liệu cần khai phá từ các tập dữ liệu khác nhau theo một tiêu chí nhất định. [2]

+ Tiền xử lý dữ liệu: Làm sạch dữ liệu, rút gọn dữ liệu, rời rạc hoá dữ liệu . [2]

+ Biến đổi dữ liệu: là bước chuẩn hoá và làm mịn dữ liệu để đưa dữ liệu về dạng thuận lợi phục vụ cho các kỹ thuật khai phá ở bước sau. [2]

+Chuyển đổi dữ liệu: Dữ liệu sau khi được chọn lọc sẽ được chuyển đổi hay hợp nhất về dạng thích hợp cho việc khai phá. [2]

+ Khai phá dữ liệu: áp dụng các kỹ thuật phân tích (thường là các kỹ thuật của học máy) nhằm: Khai thác dữ liệu, trích chọn mẫu thông tin, xây dựng tri thức. [2]

+ Đánh giá và biểu diễn tri thức: Những mẫu thông tin và mã liên hệ trong dữ liệu đã được khám phá ở bước trên được chuyển về biểu diễn ở một dạng gần với thế giới thực của người sử dụng như: đồ thị, cây, bảng biểu, luật,… Đánh giá những tri thức khám phá được theo những tiêu chí nhất định. [2]

### 2.3. Quy trình khai phá dữ liệu

Nghiên cứu lĩnh vực:

Ta cần nghiên cứu lĩnh vực cần sử dụng Data mining để xác định được những tri thức ta cần chất lọc, từ đó định hướng để tránh tốn thời gian cho những tri thức không cần thiết. [3]

Tạo tập tin dữ liễu đầu vào:

Ta xây dựng tập tin để lưu trữ các dữ liệu đầu vào để máy tính có thể lưu trữ và xử lý. [3]

Tiền xử lý, làm sạch, mã hóa:

Ở bước này ta tiến hành bỏ bớt những dữ liệu rườm rà, không cần thiết, tinh chỉnh lại cấu trúc của dữ liệu và mã hóa chúng để tiện cho quá trình xử lý . [3]

Rút gọn chiều:

Thông thường một tập dữ liệu có chiều khá lớn sẽ sinh ra một lượng dự liệu khổng lồ,ví dụ với n chiều ta sẽ có 2^n nguyên tổ hợp .Do đó, đây là một bước quan trọng giúp giảm đáng kể hao tổn hề tài nguyên trong quá trình xử lý tri thức.Thông thường ta sẽ dùng Rough set (http://en.wikipedia.org/wiki/Rough\_set) để giảm số chiều. [3]

Chọn tác vụ khai thác dữ liệu:

Để đạt được mục đích ta cần, ta chọn được tác vụ khai thác dữ liệu sao cho phù hợp.Thông thường có các tác vụ sau:

• Đặc trưng(feature)

• Phân biệt(discrimination)

• Kết hợp(association)

• Phân lớp(classification)

• Gom cụm(cluster city)

• Xu thế(trend analysis)

• Phân tích độ lệch

• Phân tích độ hiếm [3]

Khai thác dữ liệu: Tìm kiếm tri thức

Sau khi tiến hành các bước trên thì đây là bước chính của cả quá trình , ta sẽ tiến hành khai thác và tìm kiếm tri thức. [3]

Đánh giá mẫu tìm được:

Ta cần đánh giá lại trong các tri thức tìm được, ta sẽ sử dụng được những tri thức nào, những tri thức nào dư thừa, không cần biết. [3]

Biểu diễn tri thức:

Ta biểu diễn tri thức vừa thu nhập được dưới dạng ngôn ngữ tự nhiên và hình thức sao cho người dùng có thể hiểu được những tri thức đó. [4]

### 2.4. Ứng dụng của khai phá dữ liệu trong các lĩnh vực

Khai phá dữ liệu được ứng dụng trong nhiều lĩnh vực của đời sống, là công nghệ căn bản trong triển khai các giải pháp chuyển đổi số, hỗ trợ ra quyết định. Có 7 lĩnh vực được ứng dụng khai phá dữ liệu nhiều nhất bao gồm:  kinh doanh; viễn thông; ngân hàng; thương mại điện tử và bán lẻ; tài chính; y tế và chăm sóc sức khỏe; an ninh, bảo mật mạng. Trong giới hạn bài này giới thiệu chi tiết việc ứng dụng hai lĩnh vực là chăm sóc sức khỏe và thương mại điện tử để có thông tin chi tiết hơn. [4]

Ứng dụng khai thác dữ liệu trong lĩnh vực chăm sóc sức khỏe:

Ngành chăm sóc sức khỏe ngày nay tạo ra một lượng lớn dữ liệu phức tạp về bệnh nhân, tài nguyên bệnh viện, chẩn đoán bệnh, hồ sơ bệnh nhân điện tử và các loại thiết bị y tế khác nhau. Lượng dữ liệu lớn hơn là nguồn tài nguyên quan trọng cần được xử lý và phân tích để sản sinh ra kiến ​​thức hoặc khai thác thông tin cho phép hỗ trợ nghiệp vụ mang lại lợi ích tiết kiệm chi phí và ra quyết định. Một số lợi ích có thể kể đến là: [4]

Nâng cao hiệu quả điều trị: Các ứng dụng khai thác dữ liệu có thể phát triển để đánh giá hiệu quả của các phương pháp điều trị y tế trong bệnh viện. Khai thác dữ liệu có thể đưa ra phân tích về quá trình hành động nào đó chứng tỏ hiệu quả bằng cách so sánh và đối chiếu các nguyên nhân, triệu chứng và các phương pháp điều trị khác nhau. [4]

Quản lý chăm sóc sức khỏe: Các ứng dụng khai thác dữ liệu có thể được phát triển để tìm ra và theo dõi tốt hơn các trạng thái bệnh mãn tính và những bệnh nhân có nguy cơ cao, thiết kế các biện pháp can thiệp phù hợp và giảm thiểu số lần nhập viện và yêu cầu hỗ trợ quản lý chăm sóc sức khỏe. [4]

Quản trị quan hệ khách hàng: Quản lý quan hệ khách hàng là một quy trình cốt lõi để duy trì sự tương tác giữa các tổ chức và khách hàng, hoạt động này có mặt ở nhiều ngành nghề như ngân hàng, bán lẻ nhưng quan trọng hơn trong quản lý chăm sóc sức khỏe. Tương tác của khách hàng có thực hiện thông qua trung tâm cuộc gọi chăm sóc bệnh nhân, tại các văn phòng bác sĩ, bộ phận thanh toán, cơ sở nội trú và cơ sở chăm sóc sức khỏe. [4]

Xử lý lạm dụng quỹ bảo hiểm: Lạm dụng, trục lợi quỹ bảo hiểm thường diễn ra khi quản lý khám chữa bệnh bảo hiểm y tế dựa trên thẻ giấy. Các vụ việc nhận diện được như một người khám nhiều lần trong khoảng thời gian ngắn hoặc cùng một người khám tại nhiều bệnh viện tại cùng một thời điểm... Khai phá dữ liệu sẽ giúp giảm, hạn chế tình trạng này. Ví dụ cụ thể Bảo hiểm xã hội Việt Nam triển khai hệ thống giám định trên cơ sở khai phá dữ liệu khám chữa bệnh đã cải thiện rõ rệt tình trạng lạm dụng quỹ bảo hiểm y tế và giúp tiết kiệm ngân sách hàng trăm tỉ đồng mỗi năm. [4]

Ứng dụng khai phá dữ liệu trong ngành bán lẻ trực tuyến

Ngành bán lẻ trong những năm qua đã có sự tăng trưởng vượt bậc. So với mua sắm truyền thống tại cửa hàng, mua sắm trực tuyến có một số đặc điểm riêng: quá trình mua sắm của khách hàng có thể theo dõi ngay lập tức, đơn hàng gắn liền với địa chỉ giao và thanh toán. Mỗi khách hàng có phương thức thanh toán cụ thể. Các thông tin này cho phép người bán hàng phân biệt các khách hàng khác nhau và tối ưu hóa, cá thể hóa sự phục vụ cho từng khách hàng. Để làm được điều này, các nền tảng bán hàng trực tuyến phải dựa trên trí tuệ nhân tạo và khai phá dữ liệu. [4]

Khai phá dữ liệu bán lẻ có thể giúp xác định hành vi mua hàng của người dùng, tìm ra các xu hướng và xu hướng mua sắm của người dùng, nâng cao chất lượng dịch vụ người dùng, đạt được sự hài lòng và duy trì người dùng tốt hơn, tăng tỷ lệ tiêu thụ hàng hóa, thiết kế các chính sách vận chuyển và phân phối hàng hóa hiệu quả hơn, đồng thời giảm chi phí việc kinh doanh. [4]

Một số ứng dụng về khai thác dữ liệu trong ngành bán lẻ:

Thiết kế và xây dựng kho hàng dựa trên lợi ích của việc khai thác dữ liệu – Hàng hóa phân phối bán lẻ có phạm vi rộng. Do đó dữ liệu bán hàng sẽ cung cấp các thông tin cụ thể về khách hàng, nhân viên xử lý, quá trình vận chuyển, phân phối. Dữ liệu này có thể được khai phá để cung cấp thông tin tổ chức các kho lưu trữ hàng tối ưu về mặt địa lý giúp việc lưu thông phân phối hiệu quả hơn. [4]

Phân tích đa chiều về doanh số, khách hàng, sản phẩm, thời gian và khu vực - Thị trường bán lẻ cần dữ liệu kịp thời về yêu cầu của khách hàng, doanh số bán sản phẩm, xu hướng và thời trang cũng như chất lượng, chi phí, lợi nhuận và dịch vụ của hàng hóa. Điều cần thiết là phải cung cấp các công cụ phân tích và trực quan đa chiều động, chẳng hạn như việc xây dựng các khối dữ liệu phức tạp theo yêu cầu của phân tích dữ liệu. [4]

Phân tích hiệu quả của các chiến dịch tiếp thị bán hàng: Thị trường bán lẻ thực hiện các chiến dịch tiếp thị bán hàng bằng cách sử dụng quảng cáo, phiếu giảm giá và một số hình thức giảm giá, thưởng để quảng bá sản phẩm và thu hút người dùng. Phân tích dữ liệu đánh giá hiệu quả của các chiến dịch bán hàng có thể hỗ trợ cải thiện lợi nhuận của doanh nghiệp. [4]

Phân tích đa chiều có thể được sử dụng cho các mục tiêu này bằng cách so sánh số lượng bán hàng và nhiều giao dịch bao gồm các mặt hàng bán hàng trong thời gian bán hàng với các giao dịch bao gồm các mặt hàng giống nhau trước hoặc sau chiến dịch bán hàng. Phân tích liên kết có thể tiết lộ những mặt hàng nào có khả năng được mua cùng với các mặt hàng đang giảm giá, cụ thể là so với doanh số bán hàng trước hoặc sau chiến dịch. [4]

# Chương II. Thuật toán phân cụm K-Means và công nghệ sử dụng

## 1. Giới thiệu về kỹ thuật phân cụm trong khai phá dữ liệu

Phân cụm là kỹ thuật rất quan trọng trong khai phá dữ liệu, nó thuộc lớp các phương pháp ***Unsupervised Learning*** trong Machine Learning. Có rất nhiều định nghĩa khác nhau về kỹ thuật này, nhưng về bản chất ta có thể hiểu phân cụm là các qui trình tìm cách nhóm các đối tượng đã cho vào các cụm (clusters), sao cho các đối tượng trong cùng 1 cụm tương tự (similar) nhau và các đối tượng khác cụm thì không tương tự (Dissimilar) nhau. [5]

Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu. Các thuật toán phân cụm đều sinh ra các cụm. Tuy nhiên, không có tiêu chí nào là được xem là tốt nhất để đánh hiệu của của phân tích phân cụm. [5]

Kỹ thuật phân cụm có thể áp dụng trong rất nhiều lĩnh vực như:

-Marketing: Xác định các nhóm khách hàng (khách hàng tiềm năng, khách hàng giá trị, phân loại và dự đoán hành vi khách hàng,…) sử dụng sản phẩm hay dịch vụ của công ty để giúp công ty có chiến lược kinh doanh hiệu quả hơn. [6]

- Sinh vật học: Phân nhóm động vật và thực vật dựa vào các thuộc tính của chúng. [6]

- Thư viện: Theo dõi độc giả, sách, dự đoán nhu cầu của độc giả,… [6]

- Bảo hiểm, tài chính: Phân nhóm các đối tượng sử dụng bảo hiểm và các dịch vụ tài chính, dự đoán xu hướng (trend) của khách hàng. [6]

## 2. Thuật toán K-Means

### 2.1. Giới thiệu về K-Means

K-Means là thuật toán được sử dụng phổ biến trong kỹ thuật phân cụm. Tư tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng đã cho vào K cụm sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm là nhỏ nhất. [6]

Thuật toán K-Means được thực hiện qua các bước sau:

-Bước 1: Chọn ngẫu nhiên k tâm cho k cụm

-Bước 2: Tính toán khoảng cách giữa các đối tượng đến tâm (thường dùng khoảng cách Euclidean) và nhóm các đối tượng vào tâm gần nhất.

+Công thức độ đo khoảng cách Euclidean:

d(p,q) = (1) [7]

-Bước 3: Tính toán lại tâm mới cho các cụm

-Bước 4: Lặp lại bước 2 và 3 cho đến khi không có sự thay đổi về tâm cụm cũng như các đối tượng dữ liệu

### 2.2. Ưu điểm

K-means có một số ưu điểm, bao gồm tính đơn giản, tốc độ thực hiện nhanh và khả năng mở rộng cho dữ liệu lớn, phù hợp với các cụm có dạng hình cầu. [6]

### 2.3. Hạn chế và thách thức:

Kết quả đầu ra phụ thuộc vào điểm khởi tạo ban đầu: Kết quả của K-means có thể khác nhau tùy thuộc vào điểm khởi tạo ban đầu. Nếu chọn các điểm trung tâm ban đầu không tốt, thuật toán có thể dẫn đến kết quả phân cụm không tốt. [6]

Khả năng xử lý dữ liệu lớn: K-means có thể gặp khó khăn trong việc xử lý dữ liệu lớn vì nó yêu cầu tính toán khoảng cách giữa tất cả các cặp điểm dữ liệu và các điểm trung tâm. Với dữ liệu lớn, quá trình này trở nên phức tạp và tốn nhiều thời gian và tài nguyên tính toán. [6]

Phụ thuộc vào số cụm được xác định trước: K-means yêu cầu người dùng xác định số lượng cụm trước. Tuy nhiên, trong một số trường hợp, việc xác định số cụm tối ưu có thể là một thách thức. Nếu số cụm không được chọn đúng, kết quả phân cụm có thể không chính xác hoặc không phản ánh đúng cấu trúc của dữ liệu. [6]

Phân cụm không trực quan: K-means có thể cho kết quả phân cụm không trực quan khi các cụm có kích thước và hình dạng khác nhau. Điều này có thể gây khó khăn trong việc diễn giải và sử dụng kết quả phân cụm trong các bài toán thực tế. [6]

### 2.4. Ứng dụng của K-Means trong các lĩnh vực:

Một ứng dụng quan trọng của K-means là trong phân cụm hình ảnh. Thuật toán này có thể được sử dụng để phân loại các hình ảnh thành các nhóm tương tự nhau. Ví dụ, trong phân loại ảnh chó và mèo, K-means có thể được sử dụng để tạo ra các cụm dữ liệu chứa các hình ảnh chó và mèo riêng biệt. [6]

K-means được sử dụng trong phân loại văn bản. Với dữ liệu văn bản, chúng ta có thể biểu diễn các văn bản thành các vectơ đặc trưng dựa trên tần số xuất hiện của các từ. K-means sau đó có thể được áp dụng để phân loại các văn bản vào các nhóm dựa trên sự tương đồng giữa các vectơ đặc trưng. [6]

Trong lĩnh vực tiếp thị, K-means có thể được sử dụng để phân nhóm khách hàng. Bằng cách sử dụng các thuộc tính khách hàng như độ tuổi, giới tính, thu nhập và hành vi mua hàng, K-means có thể tạo ra các nhóm khách hàng tương tự nhau. Điều này giúp các doanh nghiệp hiểu rõ hơn về khách hàng của mình và tạo ra các chiến lược tiếp thị phù hợp. [6]

Ngoài ra, K-means cũng có thể được sử dụng trong phân tích dữ liệu vệ tinh. Với dữ liệu vệ tinh, K-means có thể phân loại các vùng đất thành các cụm dựa trên các đặc trưng như màu sắc và cấu trúc. Điều này hữu ích trong việc nghiên cứu các hiện tượng tự nhiên, giám sát sự thay đổi môi trường và quản lý tài nguyên. [6]

## 3. Các chỉ số đo lường dữ liệu

### 3.1. Các độ đo về xu hướng chính của dữ liệu

Mean:

Là giá trị trung bình của tất cả các giá trị trong tập dữ liệu. Nó cho biết giá trị trung bình của dữ liệu.

Công thức:

. (2) [7]

Median(Trung vị):

Là giá trị ở vị trí giữa của dữ liệu khi được sắp xếp theo thứ tự tăng dần hoặc giảm dần. Nó cho biết giá trị trung tâm của dữ liệu.

Công thức:

. (3) [7]

Mode(Yếu vị):

Là giá trị xuất hiện thường xuyên nhất trong tập dữ liệu

Midrange: Giá trị trung bình của các giá trị lớn nhất và nhỏ nhất trong tập dữ liệu

### 3.2. Các độ đo về sự phân tán của dữ liệu

Quartile:

+The first quartile (Q1)

+The second quartile (Q2)

+The third quartile (Q3)

Interquartile Range(IQR) = Q3-Q1

+Outliers : giá trị nằm cách trên Q3 hay dưới Q1 một khoảng 1.5xIQR [7]

Variance

[]. (4) [7]

### 3.3. Ý nghĩa

Xác định các thuộc tính tiêu biểu của dữ liệu về xu hướng chính và sự phân tán của dữ liệu làm nổi bật các giá trị dữ liệu nên được xem như nhiễu hoặc phần tử biên (outliers), giúp chúng ta đo lường mức độ biến đổi và đồng đều của dữ liệu cung cấp cái nhìn tổng quan về dữ liệu từ đó đóng vai trò quan trọng trong việc phân tích và hiểu hơn về dữ liệu. [7]

## 4. Công nghệ và ngôn ngữ lập trình được sử dụng

### 4.1. Về công nghệ :

+Visual studio

Visual studio là một phần mềm hỗ trợ đắc lực hỗ trợ công việc lập trình website. Công cụ này được tạo lên và thuộc quyền sở hữu của ông lớn công nghệ Microsoft. Năm 1997, phần mềm lập trình nay có tên mã Project Boston. Nhưng sau đó, Microsoft đã kết hợp các công cụ phát triển, đóng gói thành sản phẩm duy nhất. [8]

Visual Studio là hệ thống tập hợp tất cả những gì liên quan tới phát triển ứng dụng, bao gồm trình chỉnh sửa mã, trình thiết kế, gỡ lỗi. Tức là, bạn có thể viết code, sửa lỗi, chỉnh sửa thiết kế ứng dụng dễ dàng chỉ với 1 phần mềm Visual Studio mà thôi. Không dừng lại ở đó, người dùng còn có thể thiết kế giao diện, trải nghiệm trong Visual Studio như khi phát triển ứng dụng Xamarin, UWP bằng XAML hay Blend vậy. [8]

Visual Studio hỗ trợ nhiều ngôn ngữ lập trình phổ biến như C++, C#, Python,… Với giao diện người dùng thân thiện và linh hoạt, nó giúp lập trình viên tạo và quản lý các dự án một cách dễ dàng. [8]

Những tính năng cần thiết của Visual Studio

+Hỗ trợ nhiều ngôn ngữ lập trình: Tính năng này giúp phát hiện bất kỳ lỗi hoặc tham chiếu ngôn ngữ chéo (cross-language reference) nào một cách dễ dàng. [9]

+Intelli-Sense: Là một tính năng giúp phát hiện có bất kỳ đoạn code nào bị bỏ sót hay không, tự động thực thi cú pháp biến (variable syntaxes) và khai báo biến (variable declarations). Ví dụ: Nếu một biến nào đó đang được sử dụng trong chương trình và người dùng quên khai báo, intellisense sẽ khai báo biến đó cho người dùng. [9]

+Hỗ trợ đa nền tảng: Visual Studio hoạt động trên cả 3 nền tảng Windows, Linux, Mac. [9]

+Tiện ích mở rộng và Hỗ trợ: Các extension hay tiện ích mở rộng của Visual Studio giúp tăng tốc quá trình phát triển ứng dụng mà không ảnh hưởng đến hiệu suất của editor. [9]

+Repository: Visual Studio được kết nối với Git hoặc có thể được kết nối với bất kỳ Repository nào khác. [9]

+Code editor: Visual Studio có class Code editor tốt nhất hỗ trợ nhiều chức năng. Nó cho phép bookmark trong code để kết hợp Quick Navigation. Visual Studio cũng có chức năng Incremental Search, Regex Search, Multi-item Clipboard và Task-list. [9]

+Web-Support: Các ứng dụng web có thể được xây dựng và hỗ trợ trong Visual Studio. [9]

+Hỗ trợ Terminal: Visual Studio hỗ trợ Terminal hoặc Console tích hợp giúp người dùng không cần chuyển đổi giữa hai màn hình. [9]

+Hỗ trợ Git: Tài nguyên có thể được lấy từ Github Repo trực tuyến và ngược lại giúp tiết kiệm thời gian và công sức. [9]

+Debugger: Đây là một tính năng hữu ích cho phép nhà phát triển kiểm tra trạng thái của chương trình và phát hiện bug ở đâu. Bạn cũng có thể xem source code chương trình của mình bằng cách sử dụng các công cụ gỡ lỗi của debugger. [9]

+Thiết kế đa dạng: Visual Studio cung cấp một số visual designer để trợ giúp trong việc phát triển các ứng dụng [9]

Với Visual Studio, lập trình viên có thể phát triển các ứng dụng trên nhiều nền tảng khác nhau. Nó là một công cụ hữu ích cho các nhà phát triển phần mềm và đội ngũ phát triển ứng dụng, giúp tăng cường hiệu suất và chất lượng phát triển phần mềm. [8]

+Weka :

WEKA có tên đầy đủ Waikato Environment for Knowledge Analysis. Đây là bộ phần mềm mã nguồn mở miễn phí khai thác dữ liệu thuộc dự án nghiên cứu của đại học Waikato, New Zealand. Weka được xây dựng bằng ngôn ngữ lập trình Java, theo kiến trúc hướng đối tượng, được tổ chức thành thư viện phục vụ cho lĩnh vực học máy và khai phá dữ liệu. Phần mềm được tạo ra với mục tiêu xây dựng một công cụ hiện đại nhằm mục đích phát triển các kỹ thuật máy học và áp dụng chúng vào bài toán khai thác dữ liệu trong điều kiện thực tế. [10]

Với Weka người dùng có thể thực hiện nhiều tác vụ khác nhau như tiền xử lý dữ liệu, phân loại, gom cụm, hồi quy,… Weka cung cấp một giao diện đồ họa dễ sử dụng, cho phép người dùng tạo, thiết lập và thực hiện các thí nghiệm khai phá dữ liệu một cách trực quan. [9]

Weka là một công cụ hữu ích cho các nhà nghiên cứu, kỹ sư và chuyên gia trong lĩnh vực khai phá dữ liệu và học máy. Nó cho phép người dùng thực hiện các tác vụ phân tích dữ liệu một cách hiệu quả và nhanh chóng, giúp khám phá thông tin quan trọng từ dữ liệu và đưa ra quyết định thông minh trong nhiều lĩnh vực ứng dụng [9]

### 4.2. Về ngôn ngữ lập trình

+C# :

C# (hay C sharp) là một ngôn ngữ lập trình đơn giản, được phát triển bởi đội ngũ kỹ sư của Microsoft vào năm 2000. C# là ngôn ngữ lập trình hiện đại, hướng đối tượng và được xây dựng trên nền tảng của hai ngôn ngữ mạnh nhất là C++ và Java. [11]

C# với sự hỗ trợ mạnh mẽ của .NET Framework giúp cho việc tạo một ứng dụng Windows Forms hay WPF (Windows Presentation Foundation), phát triển game, ứng dụng Web, ứng dụng Mobile trở nên rất dễ dàng. [11]

Đặc trưng của ngôn ngữ lập trình C#:

+C# là ngôn ngữ đơn giản

Ngôn ngữ C# đơn giản vì nó dựa trên nền tảng C và C++. Nếu chúng ta thân thiện với C và C++ hoặc thậm chí là Java, chúng ta sẽ thấy C# khá giống về diện mạo, cú pháp, biểu thức, toán tử và những chức năng khác được lấy trực tiếp từ ngôn ngữ C và C++, nhưng nó đã được cải tiến để làm cho ngôn ngữ đơn giản hơn. [11]

+C# là ngôn ngữ hiện đại

Điều gì làm cho một ngôn ngữ hiện đại? Những đặc tính như là xử lý ngoại lệ, thu gom bộ nhớ tự động, những kiểu dữ liệu mở rộng, và bảo mật mã nguồn là những đặc tính được mong đợi trong một ngôn ngữ hiện đại. C# chứa tất cả những đặc tính trên. Nếu là người mới học lập trình có thể chúng ta sẽ cảm thấy những đặc tính trên phức tạp và khó hiểu. Tuy nhiên, cũng đừng lo lắng chúng ta sẽ dần dần được tìm hiểu những đặc tính qua các nội dung khoá học này. [11]

+C# là một ngôn ngữ lập trình thuần hướng đối tượng Lập trình hướng đối tượng (OOP: Object-oriented programming) là một phương pháp lập trình có 4 tính chất. Đó là tính trừu tượng (abstraction), tính đóng gói (encapsulation), tính đa hình (polymorphism) và tính kế thừa (inheritance). C# hỗ trợ cho chúng ta tất cả những đặc tính trên. [11]

+C# là một ngôn ngữ ít từ khóa C# là ngôn ngữ sử dụng giới hạn những từ khóa. Phần lớn các từ khóa được sử dụng để mô tả thông tin. Chúng ta có thể nghĩ rằng một ngôn ngữ có nhiều từ khóa thì sẽ mạnh hơn. Điều này không phải sự thật, ít nhất là trong trường hợp ngôn ngữ C#, chúng ta có thể tìm thấy rằng ngôn ngữ này có thể được sử dụng để làm bất cứ nhiệm vụ nào. [11]

Với sự phổ biến và tính linh hoạt của C#, lập trình viên có thể tham gia vào nhiều lĩnh vực công nghệ thông tin khác nhau, từ phát triển phần mềm, ứng dụng di động, trò chơi điện tử đến các dự án web. C# là một ngôn ngữ mạnh mẽ và linh hoạt, giúp lập trình viên tạo ra các ứng dụng đáng tin cậy và hiệu quả trên nhiều nền tảng khác nhau. [10]

# Chương III. Xây dựng chương trình phân cụm khách hàng

## 1. Mô tả bài toán phân cụm khách hàng

-Phân cụm khách hàng có ý nghĩa quan trọng trong việc hiểu và phục vụ khách hàng một cách tốt nhất. Dưới đây là những ý nghĩa chính của việc phân cụm khách hàng: +Hiểu rõ hơn về khách hàng: Phân cụm khách hàng giúp chúng ta nhận biết và hiểu rõ hơn về các đặc điểm và nhu cầu của từng nhóm khách hàng. Điều này giúp chúng ta tạo ra các chiến lược và sản phẩm/dịch vụ phù hợp với mỗi nhóm, từ đó nâng cao hiệu quả kinh doanh. [12]

+Tăng khả năng tương tác: Khi chúng ta biết rõ đối tượng khách hàng mình đang hướng đến, chúng ta có thể tạo ra các chiến lược tương tác phù hợp. Việc giao tiếp, tiếp cận và quảng bá thông tin sẽ dễ dàng hơn khi chúng ta đã phân cụm khách hàng. +Cải thiện trải nghiệm khách hàng: Bằng cách hiểu rõ nhu cầu và mong muốn của từng nhóm khách hàng, chúng ta có thể cung cấp trải nghiệm tốt hơn cho khách hàng. Điều này góp phần tạo ra sự hài lòng, tăng tính trung thành và khả năng tạo ra khách hàng tiềm năng mới. [12]

+Tối ưu hóa chiến lược kinh doanh: Phân cụm khách hàng giúp chúng ta tập trung vào nhóm khách hàng có tiềm năng cao nhất và tạo ra các chiến lược tối ưu hóa để tăng cường hiệu quả kinh doanh. Chúng ta có thể tập trung vào việc phát triển sản phẩm/dịch vụ phù hợp nhất và cải thiện quy trình bán hàng để đáp ứng nhu cầu của từng nhóm khách hàng. [12]

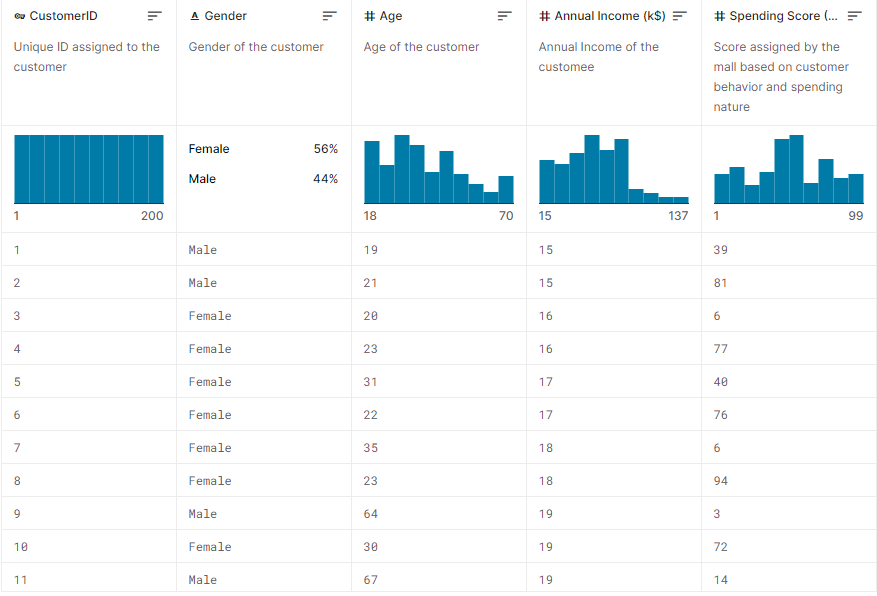
Tóm lại, phân cụm khách hàng giúp chúng ta hiểu rõ hơn về khách hàng và tạo ra các chiến lược kinh doanh phù hợp, từ đó nâng cao hiệu quả kinh doanh và tạo ra sự hài lòng cho khách hàng. [12]

## 2. Chuẩn bị dữ liệu

### 2.1. Thu thập dữ liệu

Dữ liệu được lấy từ : <https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python>

Dữ liệu bao gồm 5 thuộc tính và 200 dòng dữ liệu.



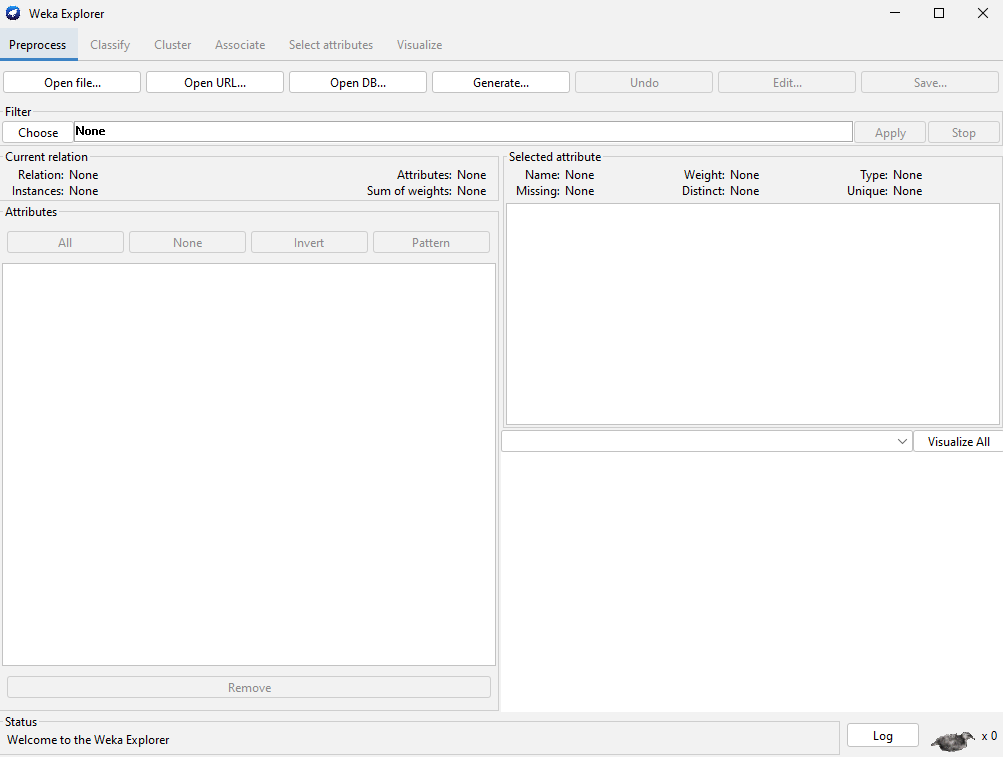
Hình 1.2. Dữ liệu khách hàng thu thập trên Kaggle.com

### 2.2. Tiền xử lý dữ liệu bằng weka

-Là một bước vô cùng quan trọng trước khi đưa dữ liệu vào phân cụm.

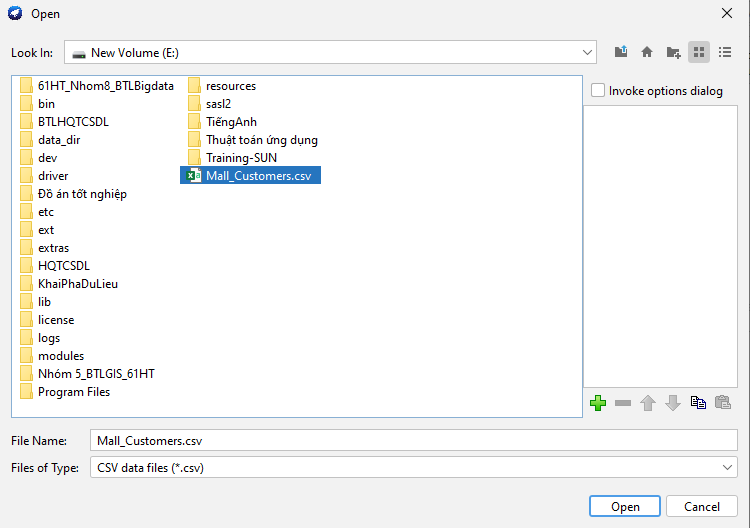
-Trong luận văn đã xử lý các dữ liệu bị thiếu đồng thời chuyển đổi thuộc tính của Gender về dạng số để phù hợp trong quá trình phân cụm và sau đây là quá trình thực hiện:

+Bước 1: Mở phần mềm Weka



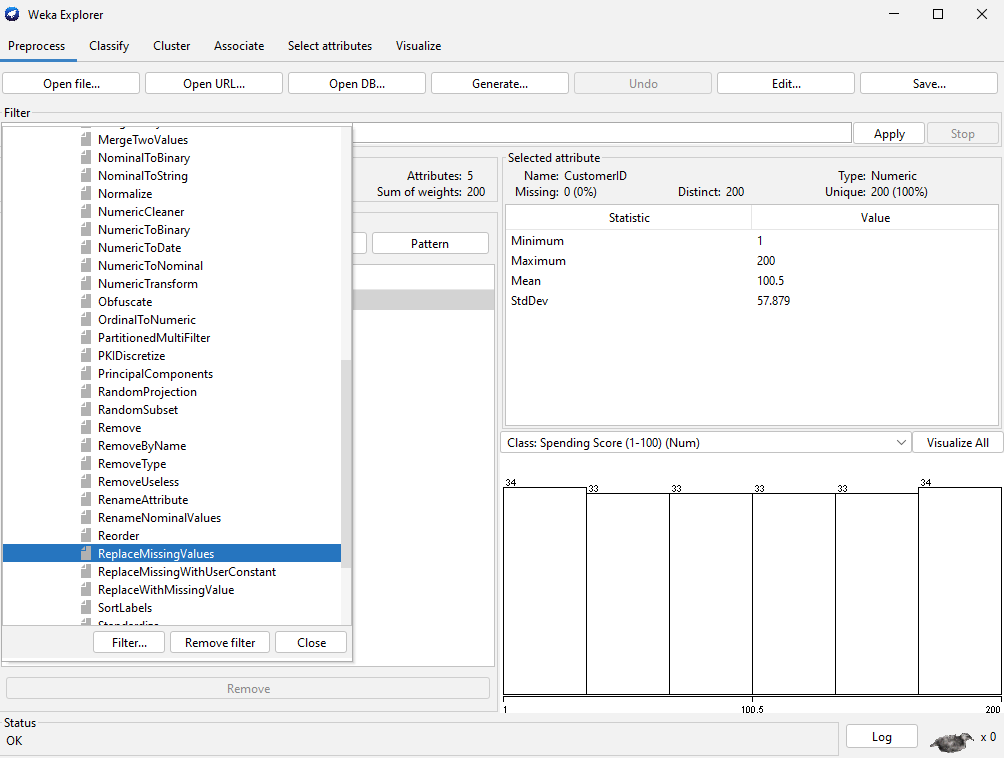
Hình 1.3. Mở ứng dụng Weka

+Bước 2: Mở file dữ liệu



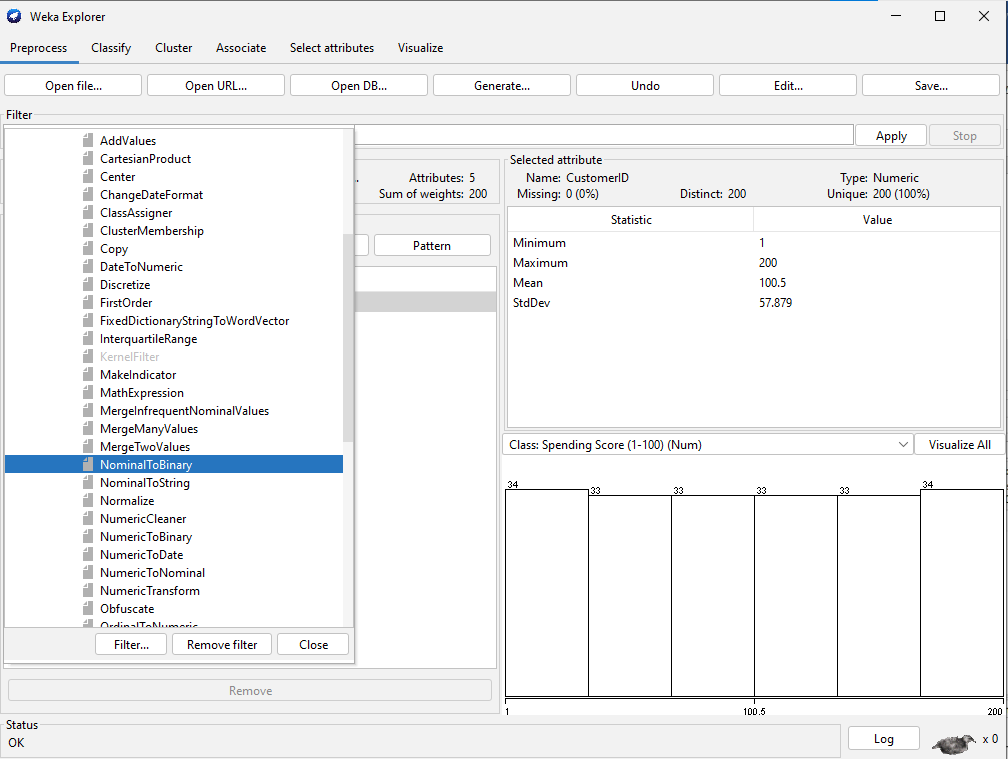
Hình 1.4. Mở file dữ liệu trên Weka

+Bước 3: Xử lý dữ liệu bị thiếu



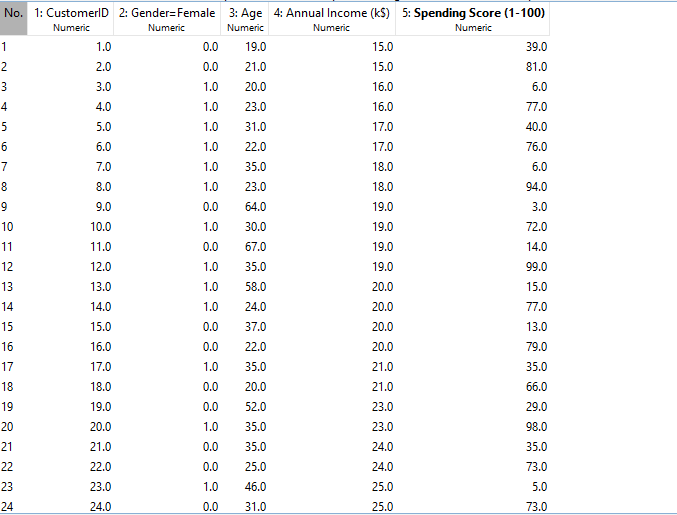
Hình 1.5. Xử lý dữ liệu bị thiếu

+Bước 4: Chuyển dữ liệu sang kiểu số



Hình 1.6. Chuyển dữ liệu sang kiểu số

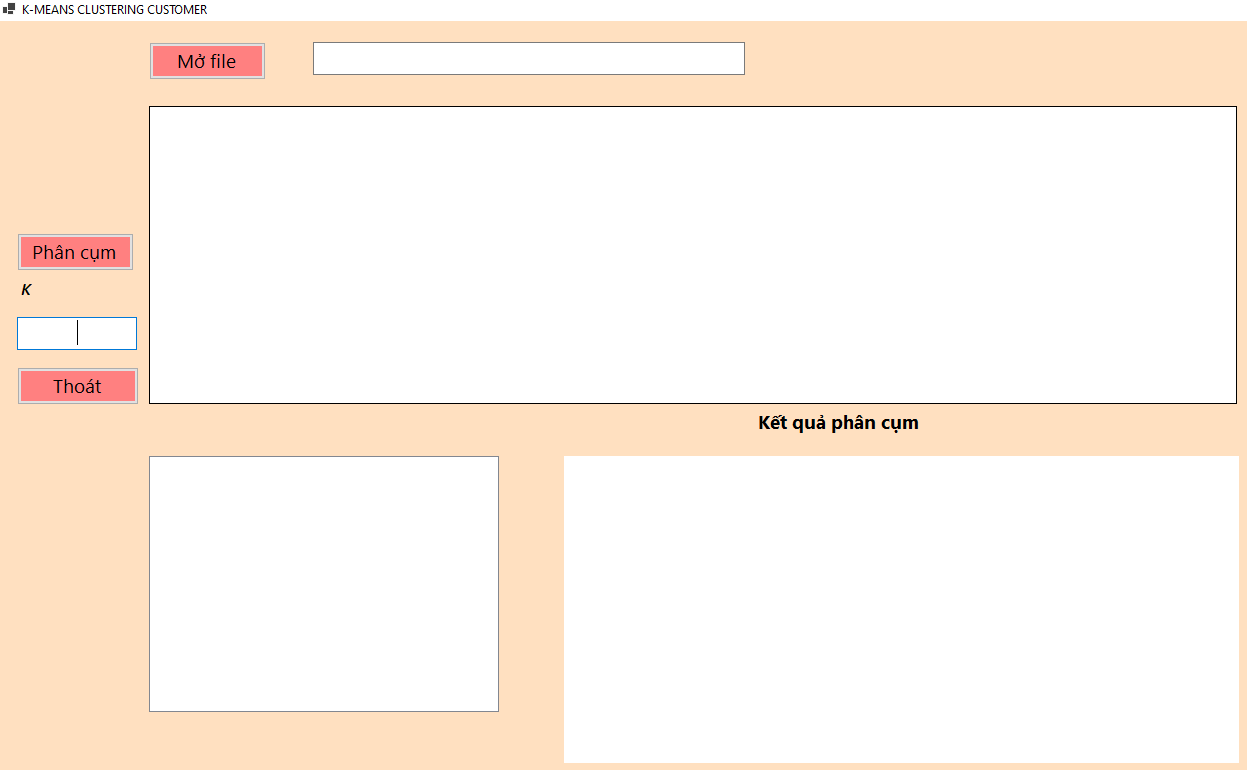
File dữ liệu sau khi tiền xử lý:



Hình 1.7. File dữ liệu sau khi tiền xử lý với Weka

## 3. Xây dựng chương trình

### 3.1. Xây dựng giao diện bằng Winform



Hình 1.8. Giao diện chương trình phân cụm khách hàng

Giao diện bao gồm:

+ 1 button Mở file để thực hiện mở dữ liệu và 1 textBox để hiển thị đường dẫn đến file dữ liệu 1 datagridview để hiển thị file dữ liệu.

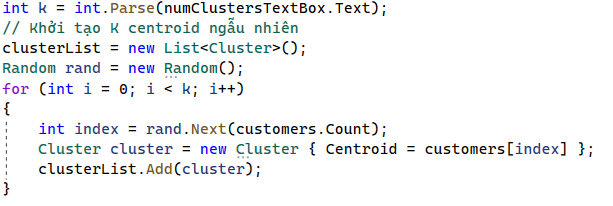
+ 1 button Phân cụm để thực hiện phân cụm dữ liệu 1 textBox để nhập số cụm 1 checklistBox để chọn trường dữ liệu muốn phân cụm và 1 plotView để hiển thị kết quả phân cụm.

+ 1 button Thoát để kết thúc chương trình.

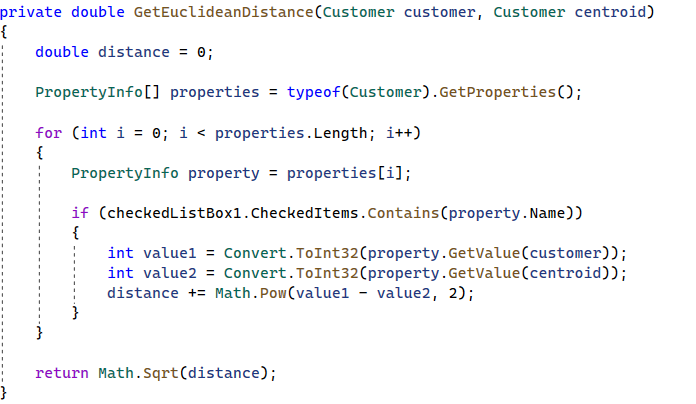
### 3.2. Code các module chương trình

Lựa chọn số cụm

Khởi tạo các tâm cụm ban đầu

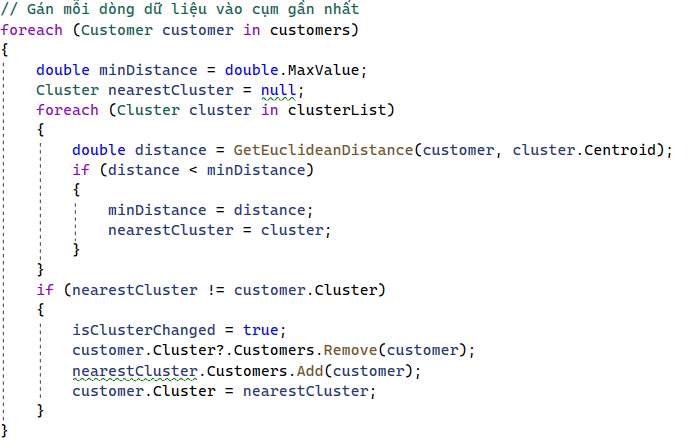


Xây dựng hàm tính khoảng cách

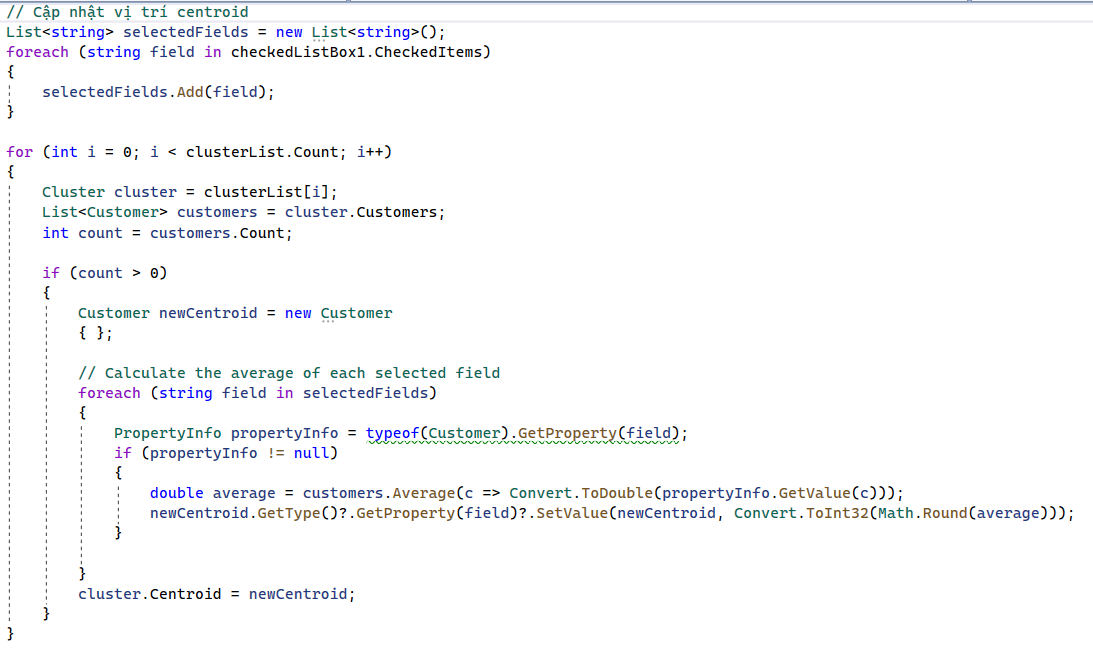


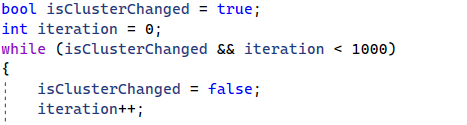
Lặp qua các bước sau cho đến khi hội tụ:

Gán mỗi điểm dữ liệu vào cụm gần nhất dựa trên khoảng cách Euclidean giữa các điểm và trung tâm cụm.

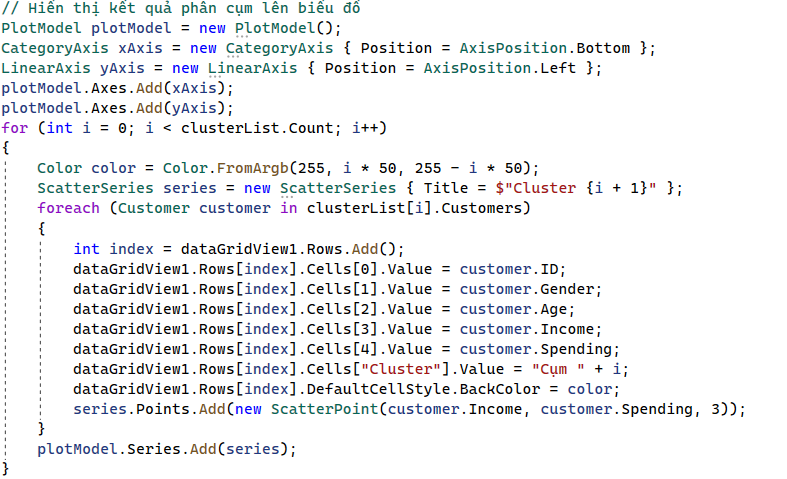


Tính toán lại vị trí trung tâm của các cụm dựa trên các điểm dữ liệu đã được gán vào từng cụm.



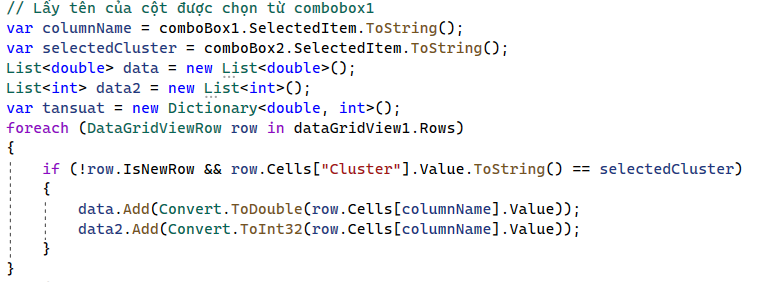
Lặp lại quá trình gán và cập nhật trung tâm cho đến khi không có sự thay đổi đáng kể trong việc gán các điểm dữ liệu vào cụm.

Hiển thị kết quả

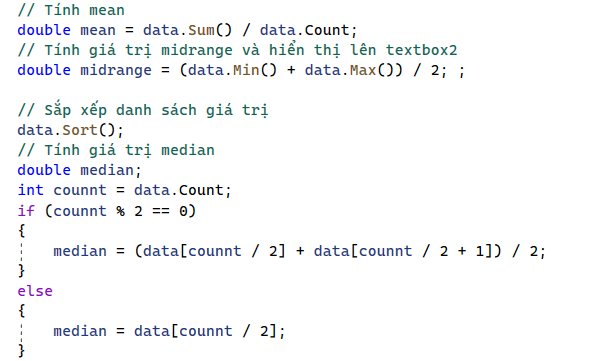


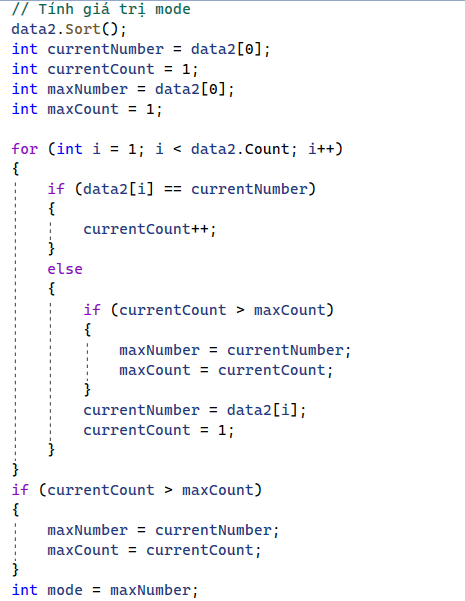
Xây dựng code để tính toán các chỉ số đo lường dữ liệu theo các thuộc tính trong mỗi cụm:

Lấy dữ liệu theo từng cụm của từng thuộc tính lưu vào 1 list để thực hiện tính toán:

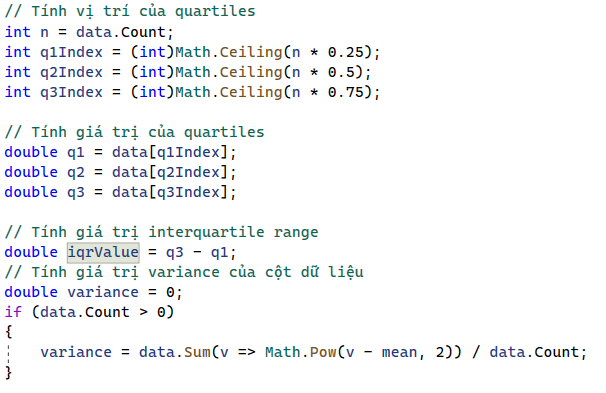


Tính giá trị các độ đo về xu hướng chính của dữ liệu:

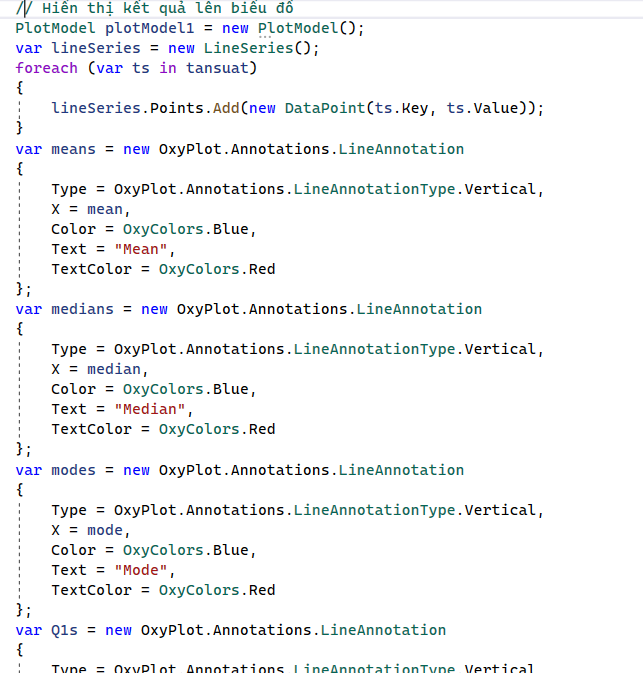




Tính giá trị các độ đo về sự phân tán của dữ liệu:



Hiển thị kết quả lên biểu đồ:

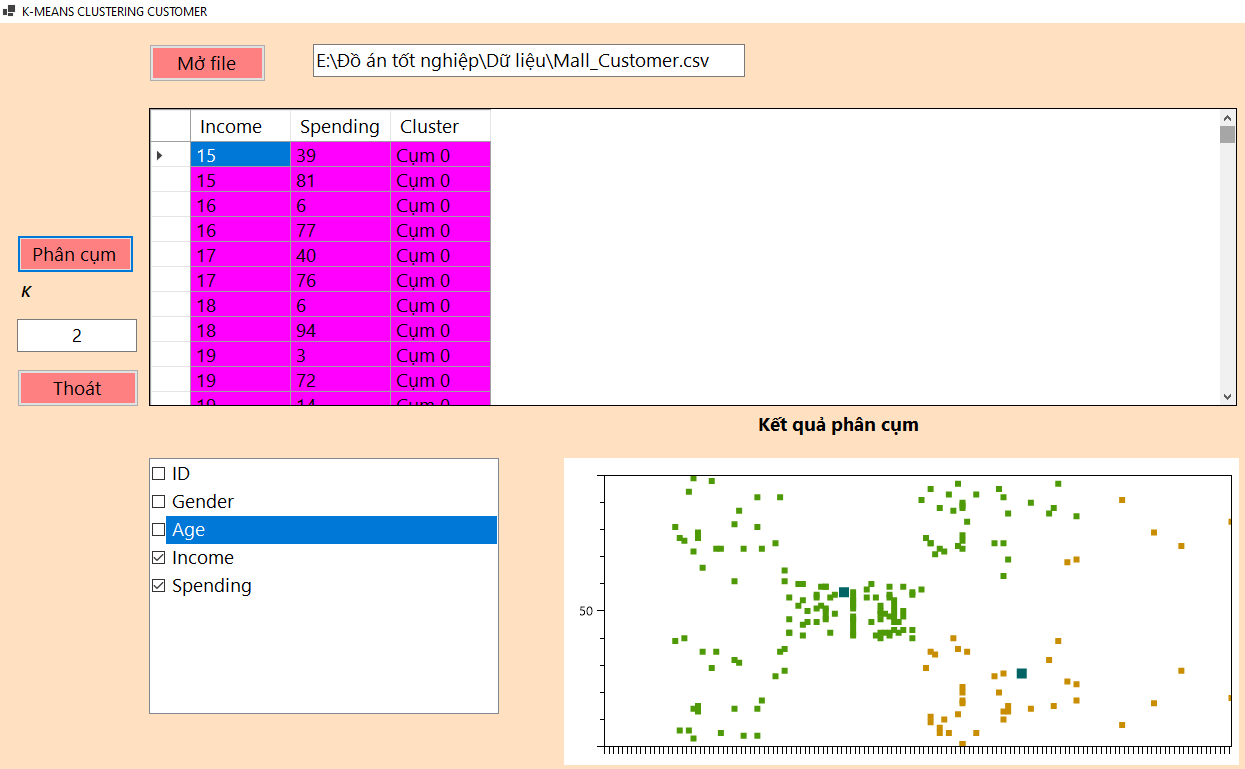


# Chương IV. Thực nghiệm và đánh giá

## 1. Phân cụm theo các thuộc tính của dữ liệu.

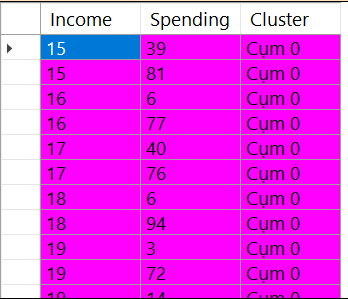
-Phân cụm theo 2 thuộc tính Income và Spending

K=2



Hình 1.9. Kết quả phân cụm thuộc tính Spending và Income với k=2

Cụm 0



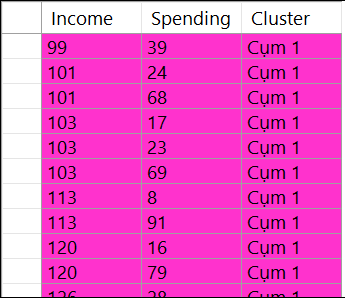
Hình 2.1. Kết quả cụm 0

Income có giá trị từ 15-69

Spending có giá trị từ 3-94

=>Cụm này có giá trị Income thấp và Spending có cả mức thấp và mức cao.

Cụm 1



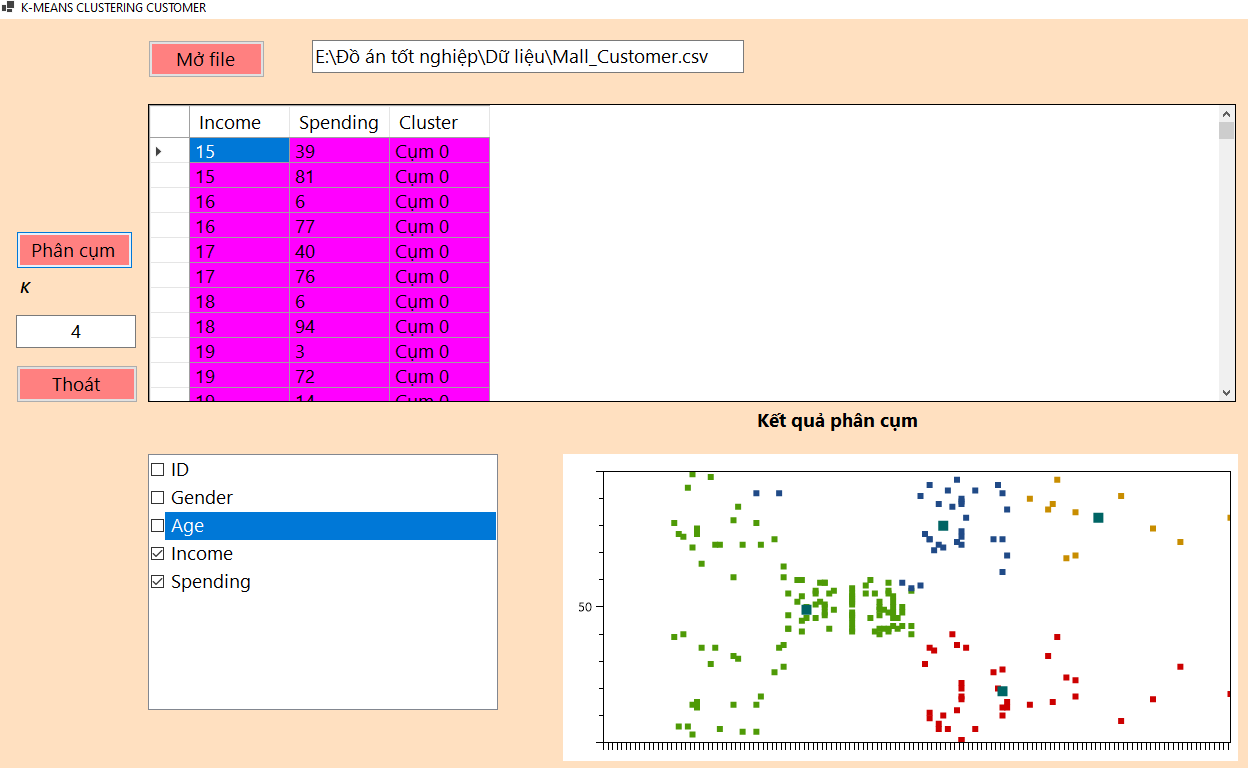
Hình 2.2. Kết quả cụm 1

Income có giá trị từ 70-126

Spending có giá trị từ 8-91

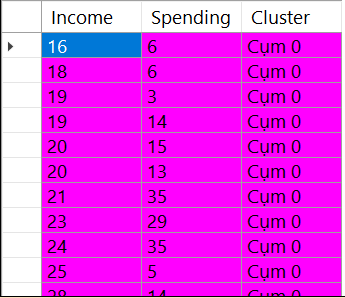
=>Cụm này có giá trị Income cao và Spending có cả mức thấp và mức cao.

K=4



Hình 2.3. Kết quả phân cụm thuộc tính Spending và Income với k=4

Cụm 0



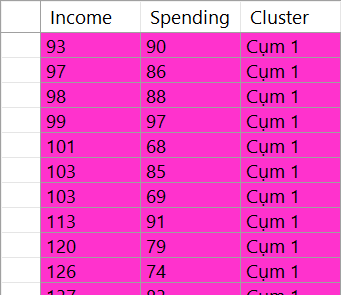
Hình 2.4. Kết quả cụm 0

Income có giá trị từ 15-69

Spending có giá trị từ 6-35

=>Cụm này có giá trị Income thấp và Spending thấp.

Cụm 1



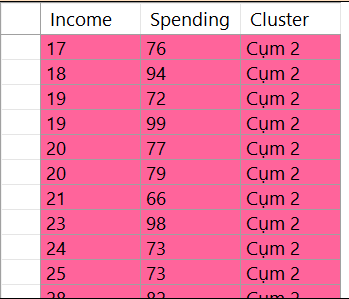
Hình 2.5. Kết quả cụm 1

Income có giá trị từ 93-127

Spending có giá trị từ 74-97

=>Cụm này có giá trị Income cao và Spending cao.

Cụm 2



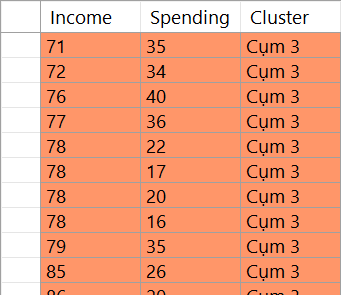
Hình 2.6. Kết quả cụm 2

Income có giá trị từ 17-69

Spending có giá trị từ 73-98

=>Cụm này có giá trị Income thấp và Spending cao.

Cụm 3



Hình 2.7. Kết quả cụm 3

Income có giá trị từ 71-125

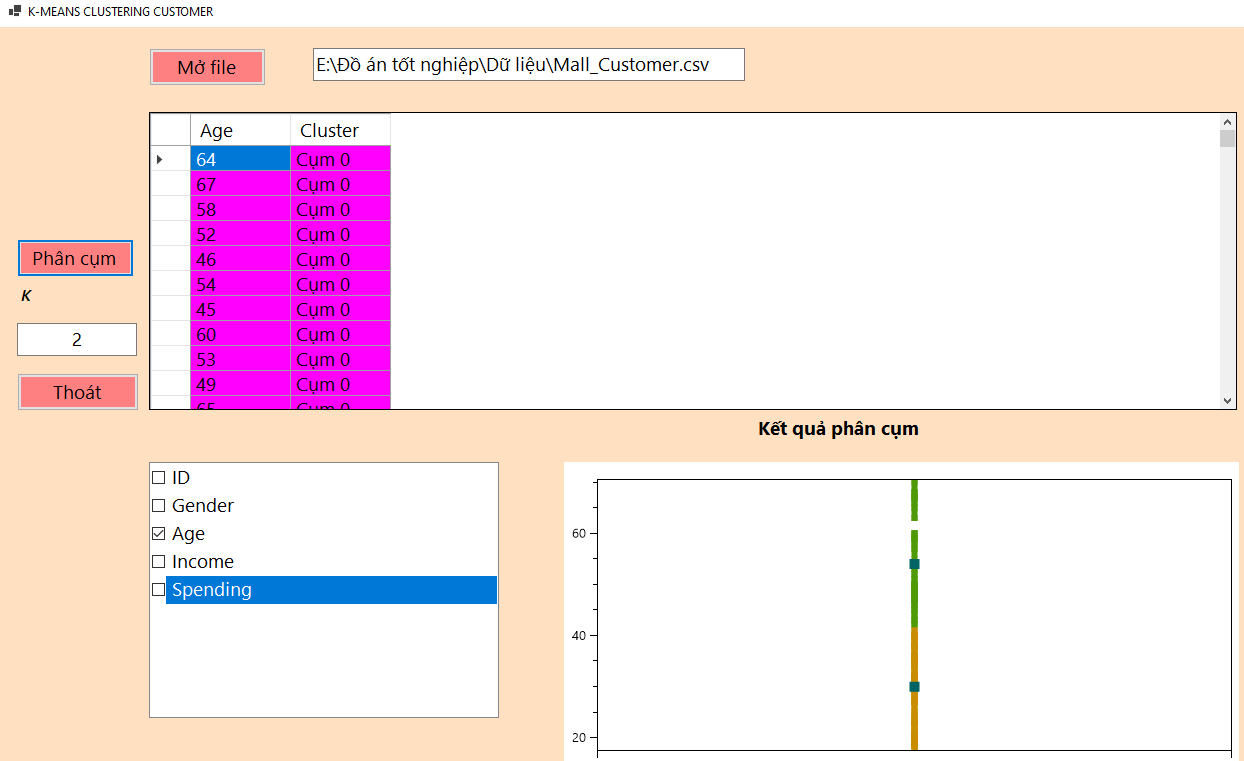
Spending có giá trị từ 16-35

=>Cụm này có giá trị Income cao và Spending thấp

Kết luận: Như vậy ta có thể kết luận rằng phân cụm khách hàng theo 2 thuộc tính Income và Spending thì K=4 sẽ hợp lý hơn vì từ kết quả phân cụm k=4 ta có thể đánh giá khách hàng 1 cách tổng quan về sự khác nhau giữa các cụm.

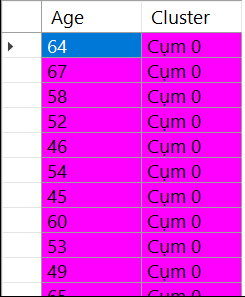
-Phân cụm theo thuộc tính Age

K=2



Hình 2.8. Kết quả phân cụm thuộc tính Age với k=2

Cụm 0

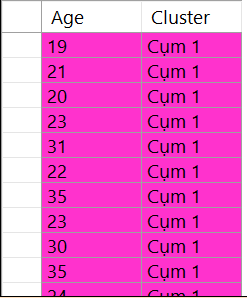


Hình 2.9. Kết quả cụm 0

Kết quả cụm 0 là cụm các khách hàng có độ tuổi từ 18-41

=>Nhóm khách hàng từ người trẻ tuổi đến trung niên.

Cụm 1

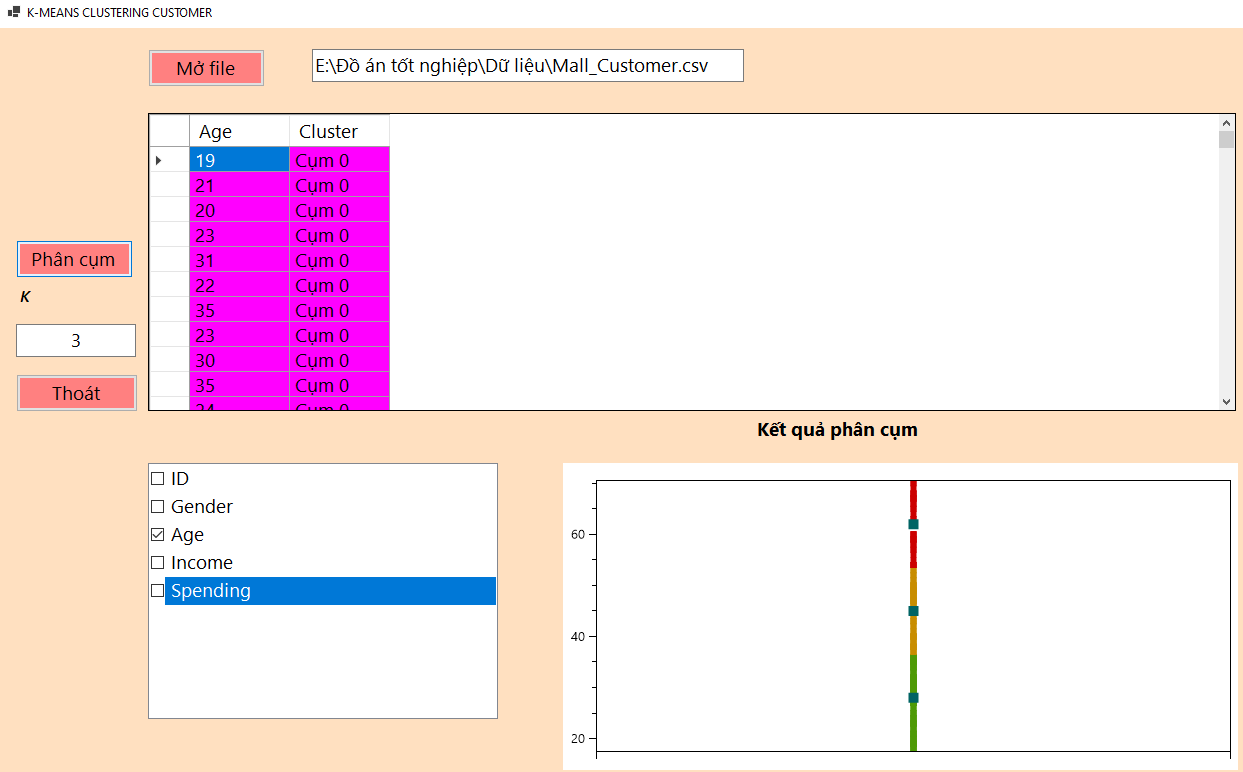


Hình 3.1. Kết quả cụm 1

Kết quả cụm 1 là cụm có độ tuổi từ 42-70

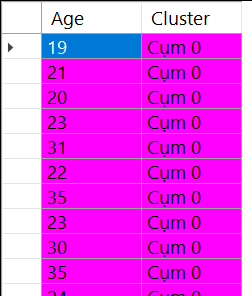
=>Nhóm khách hàng từ người trung niên đến cao tuổi

K=3



Hình 3.2. Kết quả phân cụm thuộc tính Age với k=3

Cụm 0

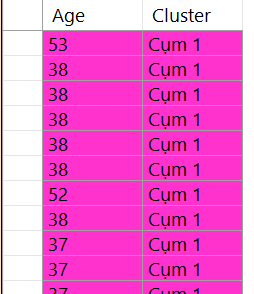


Hình 3.3. Kết quả cụm 0

Kết quả cụm 0 là cụm có độ tuổi từ 18-36

=>Nhóm khách hàng trẻ tuổi.

Cụm 1

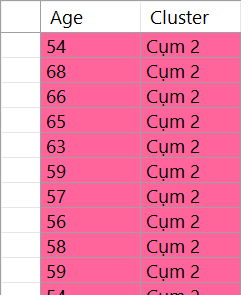


Hình 3.4. Kết quả cụm 1

Kết quả cụm 1 là cụm có độ tuổi từ 37-53

=>Nhóm khách hàng trung tuổi.

Cụm 2



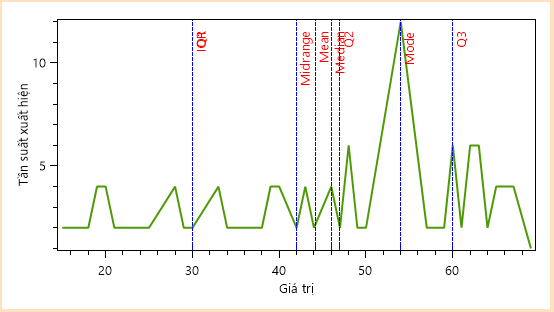
Hình 3.5. Kết quả cụm 2

Kết quả cụm 2 là cụm có độ tuổi từ 54-70

=>Nhóm khách hàng cao tuổi

Kết luận: Đối với phân cụm theo độ tuổi chúng ta có thể nhận thấy k=3 là hợp lý nhất bởi vì k=3 cho ta thấy được sự khách nhau về độ tuổi giữa các cụm.

Các chỉ số đo lường dữ liệu



Hình 3.6. Các chỉ số đo lường dữ liệu

## 2. Đánh giá kết quả phân cụm.

Với mô hình phân cụm khách hàng sau khi được xây dựng em đã chạy mô hình với file dữ liệu khách hàng được lấy trên <https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python>

Với 5 thuộc tính và 200 bản ghi

-Phân cụm khách hàng theo 2 thuộc tính Income và Spending

Với K=2

+Ta được cụm 0

Income có giá trị từ 15-69

Spending có giá trị từ 3-94

Có giá trị Income thấp và Spending có cả mức thấp và mức cao.

+Ta được cụm 1

Income có giá trị từ 70-126

Spending có giá trị từ 8-91

Có giá trị Income cao và Spending có cả mức thấp và mức cao.

Với K=4

+Ta được cụm 0

Income có giá trị từ 15-69

Spending có giá trị từ 6-35

Có giá trị Income thấp và Spending thấp.

+Ta được cụm 1

Income có giá trị từ 93-127

Spending có giá trị từ 74-97

Có giá trị Income cao và Spending cao.

+Ta được cụm 2

Income có giá trị từ 17-69

Spending có giá trị từ 73-98

Có giá trị Income thấp và Spending cao.

+Ta được cụm 3

Income có giá trị từ 71-125

Spending có giá trị từ 16-35

Có giá trị Income cao và Spending thấp

-Phân cụm khách hàng theo thuộc tính Age

Với K=2

+Ta được cụm 0

Kết quả cụm 0 là cụm các khách hàng có độ tuổi từ 18-41

Nhóm khách hàng từ người trẻ tuổi đến trung niên.

+Ta được cụm 1

Kết quả cụm 1 là cụm có độ tuổi từ 42-70

Nhóm khách hàng từ người trung niên đến cao tuổi

Với K=3

+Ta được cụm 0

Kết quả cụm 0 là cụm có độ tuổi từ 18-36

Nhóm khách hàng trẻ tuổi.

+Ta được cụm 1

Kết quả cụm 1 là cụm có độ tuổi từ 37-53

Nhóm khách hàng trung tuổi.

+Ta được cụm 2

Kết quả cụm 2 là cụm có độ tuổi từ 54-70

Nhóm khách hàng cao tuổi

Kết luận: Qua những kết quả phân cụm trên ta dễ dàng nhận thấy sự khách biệt giữa các cụm khách hàng với việc phân cụm theo 2 thuộc tính income và spending ta có thể phân theo số cụm k=4, phân cụm theo thuộc tính age ta có thể phân theo số cụm k=3. Vì nó cho ta thấy dõ sự khác nhau dõ dệt giữa các cụm sau khi thực hiện thuật toán.

**Đánh giá:**

Có 4 kết quả phân cụm khách hàng dựa trên thu nhập và chi tiêu :

1.Nhóm income cao và spending cao: Nhóm này bao gồm khách hàng có thu nhập cao và đồng thời chi tiêu cao.

=> Điều này cho thấy họ có khả năng mua sắm và tiêu dùng sản phẩm/dịch vụ cao cấp.

Đối với doanh nghiệp, nhóm này có tiềm năng tiêu thụ cao và có thể trở thành nhóm khách hàng quan trọng trong việc tạo ra doanh thu và lợi nhuận.

Chiến lược tiếp thị và quảng cáo có thể tập trung vào việc tạo ra giá trị và đảm bảo trải nghiệm mua hàng cao cấp cho nhóm này.

2.Nhóm income cao và spending thấp: Đây là nhóm khách hàng có thu nhập cao nhưng chi tiêu thấp.

=> Có thể rằng họ có xu hướng tiết kiệm hoặc đầu tư tiền của họ vào những mục tiêu khác.

Đối với doanh nghiệp, nhóm này có thể đòi hỏi một chiến lược tiếp thị khác biệt, tập trung vào việc tạo ra giá trị và khuyến khích việc tiêu dùng thông minh hoặc đề xuất các gói sản phẩm/dịch vụ có giá trị cao hơn.

3.Nhóm income thấp và spending cao: Đây là nhóm khách hàng có thu nhập thấp nhưng chi tiêu cao.

=> Họ có thể dùng hết thu nhập của mình để đáp ứng nhu cầu cơ bản hoặc đầu tư vào việc tiêu dùng.

Đối với doanh nghiệp, việc hiểu và tạo ra các sản phẩm/dịch vụ phù hợp với túi tiền của nhóm này có thể giúp thu hút và duy trì khách hàng.

Chiến lược tiếp thị có thể tập trung vào việc tạo ra giá trị tốt nhất với giá trị thấp hơn hoặc đề xuất các phương thức thanh toán linh hoạt.

4.Nhóm income thấp và spending thấp: Nhóm này bao gồm khách hàng có thu nhập thấp và chi tiêu thấp.

=> Họ có thể có hạn chế về khả năng mua sắm và tiêu dùng.

Đối với doanh nghiệp, việc tìm cách tạo ra giá trị với giá cả phù hợp hoặc tìm kiếm cách tiếp cận và hỗ trợ nhóm khách hàng này trong việc tiếp cận sản phẩm/dịch vụ của bạn có thể là một chiến lược hữu ích. Điều quan trọng là đảm bảo rằng giá cả và giá trị sản phẩm/dịch vụ phù hợp với tài chính của nhóm này.

Ý nghĩa của phân cụm khách hàng dựa trên thu nhập và chi tiêu là giúp doanh nghiệp hiểu rõ hơn về nhu cầu và hành vi tiêu dùng của từng nhóm khách hàng. Điều này giúp tối ưu hóa chiến lược tiếp thị, phục vụ khách hàng một cách hiệu quả hơn và tạo ra lợi ích kinh doanh lâu dài.

Có 3 kết quả phân cụm khách hàng dựa trên độ tuổi:

Nhóm tuổi từ 18 đến 36: Đây là nhóm khách hàng trẻ tuổi. Họ có thể là sinh viên, người mới ra trường hoặc đang bắt đầu sự nghiệp. Nhóm này có thể có nhu cầu mua sắm thời trang, sản phẩm công nghệ, giải trí, và đồ gia dụng tiện ích.

Nhóm tuổi từ 37 đến 53: Đây là nhóm khách hàng trung niên. Họ có thể là những người đã có gia đình, có thu nhập ổn định và có trách nhiệm với gia đình và công việc. Nhóm này có thể quan tâm đến các sản phẩm và dịch vụ liên quan đến gia đình, sức khỏe, du lịch, và các sản phẩm lưu trữ thông tin.

Nhóm tuổi từ 54 đến 70: Đây là nhóm khách hàng lớn tuổi. Họ có thể là người đã về hưu hoặc đang tiếp tục làm việc nhưng với tư cách thấp hơn. Nhóm này có thể có nhu cầu về sản phẩm và dịch vụ chăm sóc sức khỏe, du lịch, sản phẩm đa phương tiện, và các sản phẩm hỗ trợ trong cuộc sống hàng ngày.

Qua việc phân cụm khách hàng theo nhóm tuổi, bạn có thể nắm bắt được các đặc điểm và nhu cầu khác nhau của từng nhóm để phát triển chiến lược kinh doanh hiệu quả.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

**Kết luận**: Sau khi thực hiện xong đồ án “**PHÂN CỤM KHÁCH HÀNG SỬ DỤNG THUẬT TOÁN K-MEANS**”, em đã nỗ lực hết sức để xây dựng và hoàn thiện đồ án một cách tốt nhất, nhưng do kiến thức còn hạn chế, thời gian hoàn thành đồ án có hạn và thiếu kinh nghiệm thực tế nên không thể tránh những sai sót.

**Hướng phát triển**: Sau này em sẽ tiếp tục cố gắng để có thể nâng cấp và xây dựng mô hình có tốc độ và độ chính xác cao hơn để có thể áp dụng thật tốt trong các doanh nghiệp khiến cho mức độ bán hàng và đáp ứng được nhu cầu trong cuộc sống.

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | L. D. Nhuong, 2014. [Trực tuyến]. Available: https://123docz.net/document/1096864-khai-pha-du-lieu-dung-thuat-toan-k-mean-va-naive-bayes-tren-wave.htm. |
| [2] | T. M. Tuan, *Bài giảng khai phá dữ liệu.* |
| [3] | “Cuc chuyen doi so quoc gia,” 29 09 2022. [Trực tuyến]. Available: https://aita.gov.vn/khai-pha-du-lieu-%E2%80%93-tong-quan-ung-dung-va-cac-nen-tang-thong-dung-de-khai-pha-du-lieu. |
| [4] | “https://bienuit.wordpress.com/,” 7 9 2013. [Trực tuyến]. Available: https://bienuit.wordpress.com/2013/09/07/quy-trinh-khai-pha-du-lieu-process-of-data-mining/. |
| [5] | [Trực tuyến]. Available: https://tigosoftware.com/vi/thuat-toan-k-means-voi-bai-toan-phan-cum-du-lieu. |
| [6] | “https://www.studocu.com/,” [Trực tuyến]. Available: https://www.studocu.com/vn/document/truong-dai-hoc-phenikaa/kinh-te-vi-mo/tim-hieu-ve-thuat-toan-k-means/31354487. |
| [7] | D. T. Hien và N. D. Huy, *Data Mining.* |
| [8] | “https://visualcpp.net/,” [Trực tuyến]. Available: https://visualcpp.net/visual-studio-la-gi-gioi-thieu-phan-mem-visual-studio/. |
| [9] | “https://bizflycloud.vn/,” [Trực tuyến]. Available: https://bizflycloud.vn/tin-tuc/visual-studio-la-gi-nhung-tinh-nang-can-thiet-cua-visual-studio-20210324175835691.htm. |
| [10] | “https://trungkhithe.com/,” [Trực tuyến]. Available: https://trungkhithe.com/weka-la-gi/. |
| [11] | “https://codegym.vn/,” [Trực tuyến]. Available: https://codegym.vn/blog/2020/06/29/c-la-gi-tim-hieu-ve-ngon-ngu-lap-trinh-c/. |
| [12] | “https://vjol.info.vn/,” [Trực tuyến]. Available: https://vjol.info.vn/index.php/tcidhmodiachat/article/view/55494. |