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

**THÀNH PHỐ HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN 2**



**BÁO CÁO**

**PHÁT TRIỂN CÁC HỆ THỐNG THÔNG MINH**

**ĐỀ TÀI: HỆ THỐNG KHUYẾN NGHỊ SẢN PHẨM THỜI TRANG**

**DỰA VÀO SẢN PHẨM KHÁCH HÀNG ĐÃ XEM**

**Giáo viên hướng dẫn : Nguyễn Ngọc Duy**

**Nhóm sinh viên thực hiện : 19**

**Lớp : D19CQCNPM02-N**

**Thành viên nhóm : Hồ Đức Trung : N19DCCN213**

**Lê Thành Trung : N19DCCN214**

**Nguyễn Đăng Bắc : N19DCCN018**

**Trần Quang Ngọc Huỳnh : N19DCCN075**

**Nguyễn Ngọc Đức : N19DCCN039**

**Đinh Trường Sơn : N19DCCN159**

**TP.HCM – Tháng 12, năm 2022**

**LỜI CẢM ƠN**

Trong thời gian thực hiện đề tài với nội dung nghiên cứu “Hệ thống khuyến nghị sản phẩm thời trang thông minh”, nhóm em đã cố gắng vận dụng những kiến thức đã học trên lớp, trong thực tế để hoàn thành tốt đề tài.

Em xin chân thành cảm ơn Thầy Nguyễn Ngọc Duy đã tận tình chỉ bảo và giúp đỡ, giải đáp các vướng mắc để em có thể hoàn thành đề tài cũng như môn học này. Do kiến thức còn hạn chế nên trong quá trình thực hiện đề tài nghiên cứu sẽ không thể tránh khỏi những sai sót, kính mong Thầy chỉ dẫn thêm, bỏ qua và giúp đỡ nhóm em. Nhóm Em rất mong được sự đóng góp của Thầy và các bạn để nội dung đề tài này ngày càng hoàn thiện hơn.

MỤC LỤC

[**CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI 7**](#_heading=h.3znysh7)

[**1.1 Lý do chọn đề tài 7**](#_heading=h.2et92p0)

[**1.2 Mục tiêu nghiên cứu 8**](#_heading=h.tyjcwt)

[**1.3 Đối tượng và phạm vi nghiên cứu 8**](#_heading=h.3dy6vkm)

[**1.4 Phương pháp nghiên cứu 8**](#_heading=h.1t3h5sf)

[**1.5 Một số hạn chế dự kiến và cách khắc phục 9**](#_heading=h.4d34og8)

[**1.6 Nội dung báo cáo 9**](#_heading=h.2s8eyo1)

[**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ CÔNG NGHỆ 5**](#_heading=h.1y810tw)

[**2.1 Giới thiệu về học máy 5**](#_heading=h.17dp8vu)

[**2.2 Kỹ thuật xây dựng hệ thống khuyến nghị dựa trên kỹ thuật không cá nhân hóa 5**](#_heading=h.3rdcrjn)

[**2.3 Machine learning trong hệ thống khuyến nghị 6**](#_heading=h.4i7ojhp)

[**2.4 Thuật toán K-Means. 6**](#_heading=h.lnxbz9)

[**2.4.1 Giới thiệu 6**](#_heading=h.35nkun2)

[**2.4.2 Ý tưởng của bài toán K-Means 7**](#_heading=h.1ksv4uv)

[**2.4.3 Ưu nhược điểm của thuật toán K-Means 9**](#_heading=h.2xcytpi)

[**2.5 Elbow trong việc phân cụm 10**](#_heading=h.41mghml)

[**2.6 Đánh giá mức độ hiệu quả của hệ thống khuyến nghị thời trang thông minh 11**](#_heading=h.2grqrue)

[**2.7 Công nghệ sử dụng 11**](#_heading=h.1v1yuxt)

[**CHƯƠNG 3: XÂY DỰNG CẤU HÌNH 12**](#_heading=h.44sinio)

[**3.1 Tiền xử lý dữ liệu 12**](#_heading=h.28h4qwu)

[**3.2 Cấu hình K-Means 13**](#_heading=h.2lwamvv)

[**3.3 Gọi API Python từ Java (Spring) 14**](#_heading=h.3j2qqm3)

[**CHƯƠNG 4: THIẾT KẾ HỆ THỐNG 14**](#_heading=h.3ygebqi)

[**4.1 Thiết kế cơ sở dữ liệu: 14**](#_heading=h.2dlolyb)

[**4.1.1 Sơ đồ diagram: 14**](#_heading=h.sqyw64)

[**4.1.2 Từ điển dữ liệu: 15**](#_heading=h.3cqmetx)

[**4.2 Thiết kế giao diện: 19**](#_heading=h.1rvwp1q)

[**4.2.1 Giao diện hiển thị sản phẩm cho khách hàng: 19**](#_heading=h.4bvk7pj)

[**4.2.2 Giao diện xem chi tiết sản phẩm. 21**](#_heading=h.2r0uhxc)

[**4.2.3 Giao diện hiển thị danh sách các sản phẩm khuyến nghị thông minh. 21**](#_heading=h.1664s55)

[**4.2.4. Giao diện thêm, sửa, xóa sản phẩm 22**](#_heading=h.3q5sasy)

[**CHƯƠNG 5: KẾT QUẢ THỰC HIỆN THỬ NGHIỆM 23**](#_heading=h.25b2l0r)

[**5.1. Phân cụm dữ liệu bằng K-means 23**](#_heading=h.kgcv8k)

[**5.2. Đánh giá mức độ hiểu quả của hệ thống 23**](#_heading=h.34g0dwd)

[**5.3. Kết luận 23**](#_heading=h.1jlao46)

[**CHƯƠNG 6: KẾT LUẬN 27**](#_heading=h.43ky6rz)

# CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

## Lý do chọn đề tài

Trong những năm gần đây, sự phát triển của thương mại điện tử (E-Commerce) đã đem lại nhiều lợi ích to lớn cho nền kinh tế toàn cầu. Thông qua thương mại điện tử, nhiều loại hình kinh doanh mới được hình thành, trong đó có mua bán hàng trên mạng. Với hình thức này người tiêu dùng có thể tiếp cận với hàng hóa một cách dễ dàng và nhanh chóng hơn rất nhiều so với hình thức mua bán hàng truyền thống.

Hiện nay các hệ thống bán hàng trực tuyến đã tạo nhiều điều kiện thuận lợi để người mua có thể tiếp cận nhiều mặt hàng cùng một lúc. Tuy nhiên, một website thương mại thì luôn luôn mong muốn phát triển số lượng khách hàng, và muốn có nhiều khách hàng thì họ phải đa dạng hóa các loại sản phẩm để đáp ứng được nhu cầu mua sắm của nhiều loại khách hàng, do vậy số lượng sản phẩm và loại sản phẩm được trưng bày trong website ngày càng tăng và sẽ làm hạn chế khả năng giao tiếp chọn sản phẩm của khách hàng, họ phải duyệt qua nhiều liên kết, sàng lọc nhiều thông tin mới có thể tìm được sản phẩm mong muốn. Vậy làm sao hỗ trợ khách hàng trong công việc lựa chọn sản phẩm mua sắm? Cụ thể, những sản phẩm nào nên được đề xuất tiếp theo các sản phẩm đã được khách hàng đánh giá hoặc chọn trong giỏ hàng? Nên đề xuất bao nhiêu sản phẩm là tốt nhất cho khách hàng?

Để khách hàng có thể tìm và mua được một sản phẩm ưng ý thì một lời khuyên, một sự trợ giúp là rất quan trọng. Một người bán trong phương thức mua bán truyền thống là một lợi thế rất lớn. Do đó để hình thức mua bán qua mạng thực sự phát triển thì bên cạnh các lợi thế vốn có của mình việc có thêm một “người trợ giúp” là hết sức cần thiết. Hệ tư vấn được hình thành và phát triển không nằm ngoài mục đích đáp ứng những yêu cầu trên. Một hệ thống tư vấn tốt có thể đóng vai trò như người trung gian hỗ trợ khách hàng đưa ra quyết định chọn hàng. Tiện ích này đóng vai trò như một người bán hàng có khả năng thu thập thông tin về sở thích của khách hàng, sau đó tìm trong kho hàng vô tận của mình những mặt hàng thích hợp nhất với sở thích đó. Thực chất của một hệ thống tư vấn này là quá trình hỗ trợ khách hàng đưa ra quyết định.

Với những lý do trên cùng với sự cho phép của thầy Nguyễn Ngọc Duy, nhóm chúng em quyết định chọn đề tài ***“Xây dựng hệ thống khuyến nghị sản phẩm thời trang thông minh”.***

## Mục tiêu nghiên cứu

Tìm hiểu hệ thống khuyến nghị sản phẩm phù hợp với mô hình kinh doanh thời trang trên website.

Nghiên cứu về thuật toán K-Means và mô hình học máy không giám sát (unsupervised learning).

Thu thập, tìm hiểu, phân tích tài liệu, thông tin liên quan đến đồ án.

Ứng dụng của thuật toán K-Means vào việc phân các cụm các sản phẩm tương đồng phục vụ cho hệ thống khuyến nghị.

Cài đặt thử nghiệm phương pháp và đánh giá kết quả.

## Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: bài toán hệ thống khuyến nghị sản phẩm thời trang trong thương mai điện tử (website), thuật toán K-Means và mô hình học máy không giám sát.

Phạm vi nghiên cứu: Tìm hiểu lý thuyết về hệ thống khuyến nghị sản phẩm trong thương mại điện tử, mô hình học máy không giám sát, K-Means và giải quyết bài toán hệ thống khuyến nghị sản phẩm thời trang thông minh với mô hình học máy không giám sát kết hợp với thuật toán K-Means.

## Phương pháp nghiên cứu

***Phương pháp nghiên cứu lý thuyết***:

- Nghiên cứu, thu thập thông tin về mô hình học máy không giám sát, thuật toán K-Means và hệ thống khuyến nghị sản phẩm thời trang trong thương mại điện tử.

- Nghiên cứu ứng dụng K-Means vào giải quyết các bài toán phân cụm sản phẩm.

- Nghiên cứu ứng dụng deep learning vào giải quyết bài toán hệ thống khuyến nghị sản phẩm trong thương mại điện tử.

- Xin ý kiến tư vấn từ các chuyên gia.

***Phương pháp nghiên cứu thực nghiệm***:

- Xây dựng một website bán thời trang đơn giản áp dụng mô hình thử nghiệm trên bộ dư liệu tự xây dựng và đưa hệ thống khuyến nghị sản phẩm thông minh vào website. Dựa vào kết quả để phân tích, đánh giá hệ thống khuyến nghị thông minh.

## Một số hạn chế dự kiến và cách khắc phục

- Chưa tìm hiểu hết tất cả các cấu hình mô hình thuật toán K-means cũng như những mô hình khác để phục vụ cho hệ thống khuyến nghị một cách tối ưu nhất.

- Ứng dụng hoạt động tốt khi sử dụng bộ dữ liệu được làm sạch, chuẩn bị tốt và cấu hình thuật toán phù hợp với bộ dữ liệu.

## Nội dung báo cáo

Báo cáo được trình bày trong phạm vi chương, bao gồm:

*Chương 1: Giới thiệu về đề tài.*

*Chương 2: Cơ sở lý thuyết*.

*Chương 3: Xây dựng cấu hình.*

*Chương 4: Thiết kế hệ thống.*

*Chương 5: Kết quả thực hiện thử nghiệm.*

*Chương 6: Kết luận.*

Dù trong quá trình làm bài, nhóm chúng em đã cố gắng thu thập, nghiên cứu và tổng hợp song chắc chắn vẫn còn không ít khiếm khuyết. Chúng em mong muốn nhận được những ý kiến đánh giá hữu ích từ thầy để kiến thức về môn học “phát triển hệ thống thông minh” ngày càng trở nên hoàn thiện.

Chúng em xin chân thành cảm ơn sự giúp đỡ từ phía thầy trong thời gian qua.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ CÔNG NGHỆ

## Giới thiệu về học máy

**Học máy** (*machine learning* ) là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể. **Học máy** được chia thành 2 loại chính bao gồm: học có giám sát và học không giám sát:

* + - Học có giám sát là phương pháp sử dụng những dữ liệu được gán nhãn sẵn để suy luận ra quan hệ giữa đầu vào và đầu ra. Sau khi tìm hiểu cách tốt nhất để mô hình hóa các mối quan hệ cho dữ liệu được gán nhãn, thuật toán huấn luyện sẽ được sử dụng cho các bộ dữ liệu mới. Học tập có giám sát có thể được nhóm lại thành các vấn đề về phân loại và hồi quy.
    - Học không giám sát sử dụng những dữ liệu chưa được gán nhãn sẵn để suy luận và tìm cách để mô tả dữ liệu cùng cấu trúc của chúng. Ứng dụng của học không giám sát đó là hỗ trợ phân loại thành các nhóm có đặc điểm tương đồng.

## Kỹ thuật xây dựng hệ thống khuyến nghị dựa trên kỹ thuật không cá nhân hóa

Trong nhóm kỹ thuật này, do chúng khá đơn giản, dễ cài đặt nên nên thường được các website/hệ thống tích hợp vào, gồm cả các website thương mại, website tin tức, hay giải trí. Chẳng hạn như trong các hệ thống bán hàng trực tuyến, người ta thường gợi ý các sản phẩm được xem/mua/bình luận/.. nhiều nhất; gợi ý các sản phẩm mới nhất; gợi ý các sản phẩm cùng loại/ cùng nhà sản xuất/..; gợi ý các sản phẩm được mua/chọn cùng nhau. Một ví dụ khá điển hình là thông qua luật kết hợp (như Apriori), Amazon đã áp dụng khá thành công để tìm ra các sản phẩm hay được mua cùng nhau như minh họa trong Hình 1.

Graphical user interface

Description automatically generated

*Hình 1 Gợi ý sản phẩm thường được mua cùng nhau*

Tuy vậy, bất lợi của các phương pháp này là không cá nhân hóa cho từng người dùng, nghĩa là tất cả các người dùng đều được gợi ý giống nhau khi chọn cùng sản phẩm.

## Machine learning trong hệ thống khuyến nghị

Machine learning là một trong những hướng nghiên cứu rất phát triển của lĩnh vực trí tuệ nhân tạo trong những năm gần đây. Các mô hình Machine learning đã tạo nên những đột phá lớn trong việc giải quyết các bài toán phức tạp. Nó là thuật toán dựa trên một số ý tưởng từ não bộ với việc tiếp thu nhiều tầng biểu diễn, cả cụ thể lẫn trừu tượng, qua đó làm rõ nghĩa của các loại dữ liệu. Với sự phát triển của thương mại điện tử sẽ tạo ra nguồn dữ liệu rất lớn gây khó khăn trong việc phân tích dữ liệu để xây dựng hệ thống khuyến nghị sản phẩm nên việc áp dụng Machine learning vào xây dựng hệ thống khuyến nghị sản phẩm trong thương mại điện tử là một hướng nghiên cứu đúng đắn thời điểm này.

## Thuật toán K-Means.

## Giới thiệu

**K-Means clustering** là một thuật toán phân cụm thuộc nhóm Học không giám sát dùng để hỗ trợ phân loại thành các nhóm sản phẩm khác nhau sao cho các sản phẩm trong cùng một cụm có đặc điểm, tính chất tương đồng.

## Ý tưởng của bài toán K-Means

Ý tưởng đơn giản nhất về cluster (cụm) là tập hợp các điểm *ở gần nhau trong một không gian nào đó* (không gian này có thể có rất nhiều chiều trong trường hợp thông tin về một điểm dữ liệu là rất lớn). Hình bên dưới là một ví dụ về 3 cụm dữ liệu (từ giờ tôi sẽ viết gọn là *cluster*).

Chart, scatter chart

Description automatically generated

Hình 2: Bài toán với 3 clusters

Giả sử mỗi cluster có một điểm đại diện (*center*) hinh tròn. Và những điểm xung quanh mỗi center thuộc vào cùng nhóm với center đó. Một cách đơn giản nhất, xét một điểm bất kỳ, ta xét xem điểm đó gần với center nào nhất thì nó thuộc về cùng nhóm với center đó.

**Tóm tắt thuật toán:**

**Đầu vào:** Dữ liệu X và số lượng cluster cần tìm K.

**Đầu ra:** Các center M và label vector cho từng điểm dữ liệu Y.

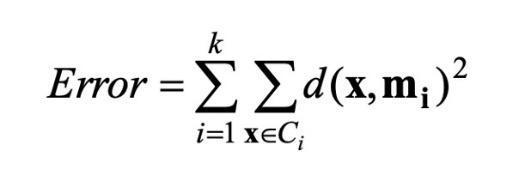
1. Chọn K điểm bất kỳ làm các center ban đầu.
2. Phân mỗi điểm dữ liệu vào cluster có center gần nó nhất.
3. Nếu việc gán dữ liệu vào từng cluster ở bước 2 không thay đổi so với vòng lặp trước nó thì ta dừng thuật toán.
4. Cập nhật center cho từng cluster bằng cách lấy trung bình cộng của tất các các điểm dữ liệu đã được gán vào cluster đó sau bước 2.
5. Quay lại bước 2.

Chúng ta có thể đảm bảo rằng thuật toán sẽ dừng lại sau một số hữu hạn vòng lập. Thật vậy, vì ham mất mát là một số dương và sau mỗi bước 2 hoặc 3, giá trị của hàm mất mát bị giảm đi. Theo kiến thức về dãy số trong phương trình : nếu một dãy số giảm và bị chặn dưới thì nó hội tụ! Hơn nữa, số lượng cách phân nhóm cho toàn bộ dữ liệu là hữu hạn nên đến một lúc nào đó, hàm mất mát sẽ không thể thay đổi, và chúng ta có thể dụng thuật toán tại đây.

* Điều kiện hội tụ (điều kiện dừng thuật toán)

Ta sẽ xác định điều kiện dừng thuật toán theo một số cách như sau:

* + Tại 1 vòng lặp: có ít các điểm dữ liệu được gán sang cluster khác hoặc
  + Điểm trung tâm (centroid) không thay đổi nhiều hoặc
  + Giá trị hàm mất mát không thay đổi nhiều:

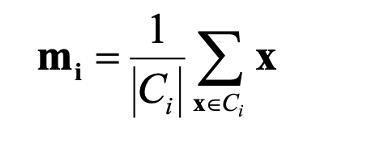


Trong đó Ci là cluster thứ i, mi là điểm trung tâm của cluster Ci tương ứng.

Nhìn chung về điều kiện hội tụ có thể thấy mối liên hệ giữa các điều kiện là gần tương đồng như nhau. Khi có ít điểm dữ liệu được gán sang cluster khác có thể khiến điểm trung tâm không thay đổi nhiều và từ đó hàm mất mát cũng sẽ ít bị ảnh hưởng. Vậy nên chúng ta có thể sử dụng 1 trong 3 cách trên để xác định điều kiện dừng của thuật toán

* Xác định điểm trung tâm của cluster

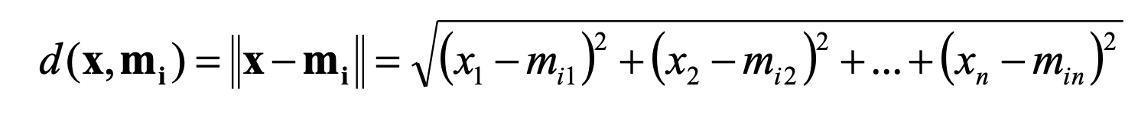
Để xác định điểm trung tâm của cluster ta sử dụng công thức như sau:



Trong đó Ci là cluster thứ i, mi là điểm trung tâm của cluster Ci tương ứng.

* Phép đo khoảng cách

Trong K-means để đánh giá mức độ giống nhau hay khoảng cách giữa 2 điểm dữ liệu ta có thể sử dụng các phép đo khoảng cách khác nhau. Ngoài khoảng cách Euclidean, tuỳ thuộc vào từng bài toán có thể sử dụng phương pháp đo khác (cosine, manhattan…)



Mọi phương pháp tính khoảng cách giữa 2 vectơ đều có thể được sử dụng. Mỗi cách tính khoảng cách thể hiện cách nhìn nhận về dữ liệu

* Có vô số cách tính khoảng cách
* Cách tính khoảng cách nào là tốt? Câu trả lời phụ thuộc vào từng bài toán để đưa ra cách tính khoảng cách phù hợp.
* Ảnh hưởng của outlier

Hiểu đơn giản thì Outliers là một hoặc nhiều cá thể khác hẳn đối với các thành viên còn lại của nhóm. Sự khác biệt này có thể dựa trên nhiều tiêu chí khác nhau như giá trị hay thuộc tính. Ví dụ về outlier có thể như là nhiễu trong các cảm biến hay lỗi trong quá trình nhập liệu của người dùng ảnh hưởng đến chất lượng của dữ liệu.

## Ưu nhược điểm của thuật toán K-Means

## Ưu điểm

## Đơn giản dễ cài đặt.

## Tốc độ xử lý tương đối nhanh, độ phức tạp là O(tkn), trong đó:

## + n: Số điểm trong không gian dữ liệu

## + k: Số cụm

## + t: Số lần lập

## Có hiệu quả trong thực tế.

## Đảm bảo hội tụ trong thời gian đa thức.

## Linh hoạt trong việc lựa chọn phương pháp đo khoảng cách.

## Nhược điểm

## Việc lựa chọn cách tính khoảng cách cho bài toán cụ thể khó.

## Khó xác định số cụm thật sự mà không gian dữ liệu có

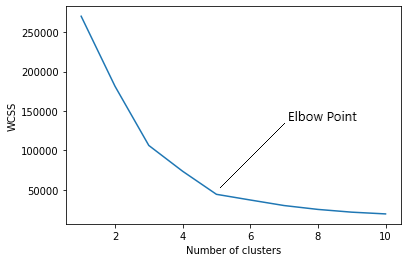
## Nhạy cảm với các điểm dữ liệu outlier.

## Chỉ có thể áp dụng khi tính được trọng tâm.

## Elbow trong việc phân cụm

Trong phương pháp Elbow,thay đổi số lượng cụm ( K ) từ 1 – 10. Đối với mỗi giá trị của K, chúng tôi tính toán WCSS (Trung bình cộng khoảng cách trong cụm).

WCSS là tổng bình phương khoảng cách giữa mỗi điểm và trọng tâm trong một cụm. Sau đó vẽ đồ thị WCSS với giá trị K, đồ thị trông giống như một Khuỷu tay.Khi số cụm tăng lên, giá trị WCSS sẽ bắt đầu giảm .Giá trị WCSS lớn nhất khi K = 1. Khi phân tích đồ thị ta thấy đồ thị sẽ thay đổi nhanh chóng tại một điểm và do đó tạo ra hình dạng khuỷu tay. Từ thời điểm này, đồ thị bắt đầu di chuyển gần như song song với trục X. Giá trị K tương ứng với điểm này là giá trị K tối ưu hoặc số cụm tối ưu.



Hình 3: đồ thị Elbow

## Đánh giá mức độ hiệu quả của hệ thống khuyến nghị thời trang thông minh

## Ý tưởng: mỗi khi người dùng truy cập trang web sẽ lưu lại thời gian và ID của session. Khi người dùng click vào xem những sản phẩm ở danh mục được hệ thống khuyến nghị, lúc này những sản phẩm đó sẽ được lưu lại ở CSDL nhằm phục vụ cho chức năng gợi ý sản phẩm, nâng cấp và phát triển hệ thống sau này.

## Công thức đánh giá:

## Công nghệ sử dụng

## - Front-end: HTML, CSS, JS

## - Back-end: Java (Spring MVC), Python (FastAPI)

## - Model: Python

## - CSDL: MSSQL

# CHƯƠNG 3: XÂY DỰNG CẤU HÌNH

## 

## Tiền xử lý dữ liệu

## - Biến đổi dữ liệu chữ về dạng số nhằm hỗ trợ cho thuật toán K-means đọc dữ liệu số

- Các thuộc tính dùng để phân cụm bao gồm: Nhóm sản phẩm (Category), thương hiệu (Brand), kiểu sản phẩm (Product Type), giới tính (Gender), chất liệu (Material) và năm ra mắt (ReleaseTime). Trong đó:

* Thuộc tính thương hiệu được chuyển đổi từ dạng chuỗi sang dạng số nguyên theo bảng xếp hạng thương hiệu thời trang từ trang Brand Finance ([Apparel 50 2022 | Brand Value Ranking League Table | Brandirectory](https://brandirectory.com/rankings/apparel/table)) năm 2022.
* Các thuộc tính ProductType, Gender, Material được chuyển đổi từ dạng chuỗi sang dạng số nguyên sử dụng phương pháp LabelEncoder.
* Thuộc tính Price được phân lại thành 3 mức

## 

## - Biến đổi miền giá trị

## Text Description automatically generated

## Cấu hình K-Means

## - Sử dụng phương pháp Elbow để tìm ra K cụm

## Text Description automatically generated

## - Huấn luyện model K-means bằng data đã được xử lý và lưu lại model sử dụng module pickle

## A screenshot of a computer screen Description automatically generated with medium confidence

## - Trả về một mảng các giá trị cụm mà sản phẩm đó thuộc về và lưu vào CSDL

## 

## Gọi API Python từ Java (Spring)

Để hệ thống có thể giao tiếp với module thông minh, nhóm dùng cách tiếp cận là viết API cho module thông minh sử dụng Fast API. Sau đó, từ hệ thống được viết dựa trên Spring (Java) chúng ta có thể gọi tới API của module thông minh.

Bên dưới là ví dụ chi tiết gọi API từ hệ thống.

## 

# 

# CHƯƠNG 4: THIẾT KẾ HỆ THỐNG

## 4.1 Thiết kế cơ sở dữ liệu:

### **4.1.1 Sơ đồ diagram:**

**Diagram

Description automatically generated**

*Hình 4: Sơ đồ diagram cơ dữ liệu*

### **4.1.2 Từ điển dữ liệu:**

* + - * **Bảng Product**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ghi chú** |
| 1 | ID | nvarchar | Identity |
| 2 | Name | nvarchar | Tên sản phẩm |
| 3 | IDCategory | int | Khóa ngoại |
| 4 | Price | float | Giá |
| 5 | Image | nvarchar | Url ảnh |
| 6 | Brand | nvarchar | Thương hiệu |
| 7 | Gender | bit | Giới tính |
| 8 | ReleaseTime | smallint | Thời gian ra mắt |
| 9 | ProductType | nvarchar | Khóa ngoại |
| 10 | ProductCluster | smallint | Khóa ngoại |
| 11 | ProductMaterial | nvarchar | Khóa ngoại |

* **Bảng Account**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ghi chú** |
| 1 | ID | Bigint | Identity |
| 2 | Username | nvarchar | Tên đăng nhập |
| 3 | Password | nvarchar | Mật khẩu |
| 4 | Fullname | nvarchar | Họ và tên |
| 5 | Role | int | Khóa ngoại chỉ quyền hạn |
| 6 | Email | nvarchar | Địa chỉ email |
| 7 | Gender | bit | Giới tính |
| 8 | Image | nvarchar | Url ảnh |
| 9 | Birthday | date | Ngày sinh |
| 10 | Phone | Char | Số điện thoại |
| 11 | Address | nvarchar | Địa chỉ |

* **Bảng Cart**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ghi chú** |
| 1 | ID | int | Identity |
| 2 | Username | nvarchar | Tên khách hàng |
| 3 | ProductID | nvarchar | Mật khẩu |
| 4 | Quantity | int | Họ và tên |

* **Bảng Order**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ghi chú** |
| 1 | ID | int | Identity |
| 2 | CusUsername | nvarchar | Tên khách hàng |
| 3 | CusEmail | nvarchar | Địa chỉ email khách hàng |
| 4 | TotaPrice | bigint | Tổng số tiền |
| 5 | Date | date | Ngày sinh |
| 6 | Phone | Char | Số điện thoại |
| 7 | CusAddress | nvarchar | Địa chỉ khách hàng |

* **Bảng OrderDetail**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ghi chú** |
| 1 | ID | int | Identity |
| 2 | IDOrder | nvarchar | ID đơn |
| 3 | IDProduct | nvarchar | ID sản phẩm đặt |
| 4 | Quantity | bigint | Số lượng |
| 5 | Price | date | Giá tiền |

* **Bảng ProductCategory**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ghi chú** |
| 1 | ID | int | Identity |
| 2 | Name | nvarchar | Tên loại sản phẩm |

* **Bảng Rating**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ghi chú** |
| 1 | Username | nvarchar | Tên sản phẩm |
| 2 | Rating | smallint | Tỷ lệ |
| 3 | ProductID | nvarchar | ID sản phẩm |

* **Bảng Role**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ghi chú** |
| 1 | ID | int | Identity |
| 2 | Name | nvarchar | Loại tài khoản |

* **Bảng SizeAndColor**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ghi chú** |
| 1 | ProductID | nvarchar | ID sản phẩm |
| 2 | Size | nvarchar | Kích thước |
| 3 | Color | nvarchar | Màu sắc |
| 4 | Quantity | int | Số lượng |

* **Bảng History**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ghi chú** |
| 1 | Username | nvarchar | Tên đăng nhập |
| 2 | RequestTime | datetime | Thời gian truy cập |
| 3 | SessionID | nvarchar | ID phiên làm việc |
| 4 | ProductID | nvarchar | ID sản phẩm |
| 5 | ID | Int | ID |

* **Bảng SatisfyProduct**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ghi chú** |
| 1 | HistoryID | int | ID lịch sử |
| 2 | ProductID | nvarchar | ID sản phẩm |

## Thiết kế giao diện:

### **4.2.1 Giao diện hiển thị sản phẩm cho khách hàng:**

* **Home**

Graphical user interface, website

Description automatically generated

*Hình 5: trang chủ*

* **Xem danh sách sản phẩm**

*Hình 6: Danh sách sản phẩm*A screenshot of a computer

Description automatically generated with low confidence

### **4.2.2 Giao diện xem chi tiết sản phẩm.**

A computer screen capture

Description automatically generated with medium confidence

*Hình 7: Giao diện chi tiết sản phẩm*

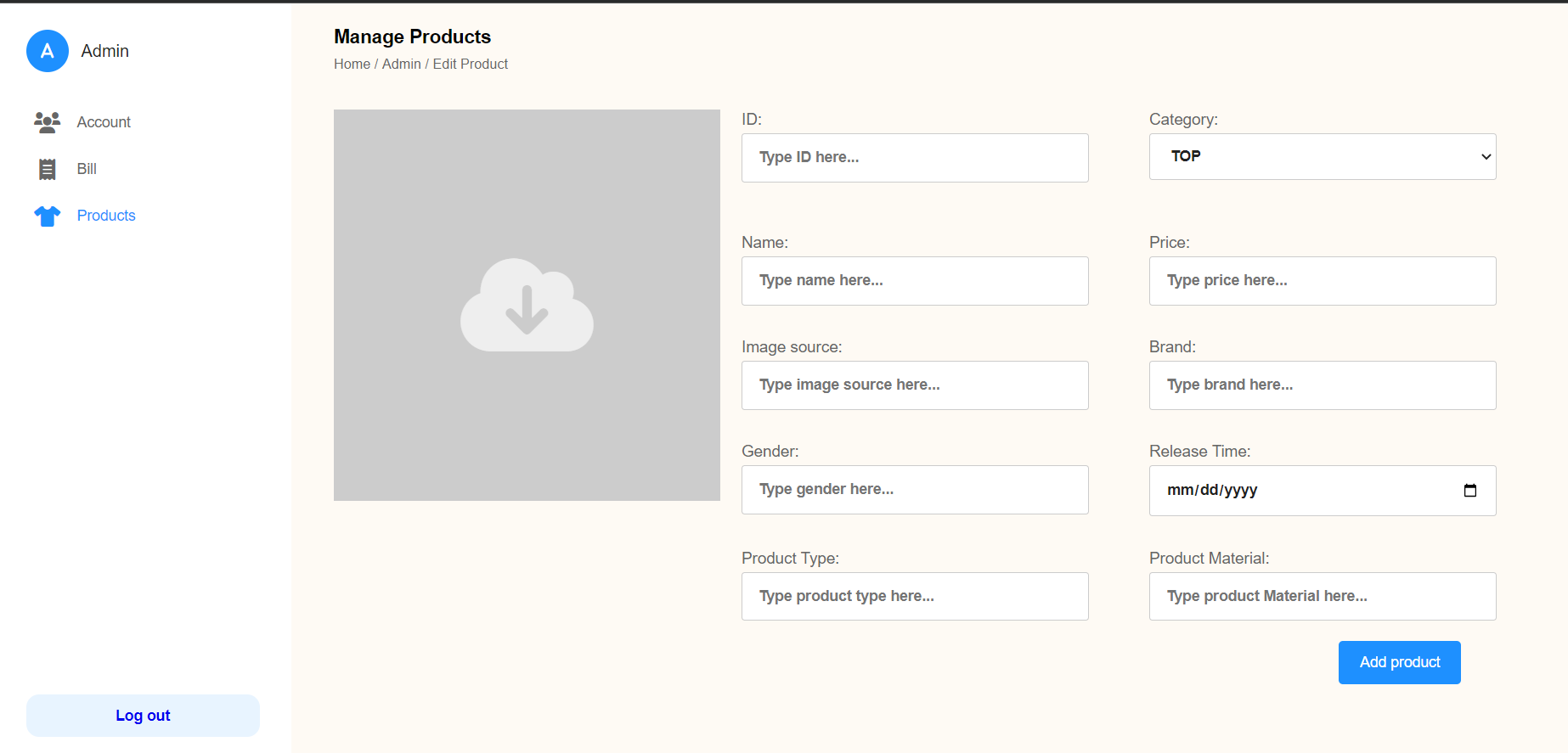
### **4.2.3 Giao diện hiển thị danh sách các sản phẩm khuyến nghị thông minh.**

A screenshot of a computer

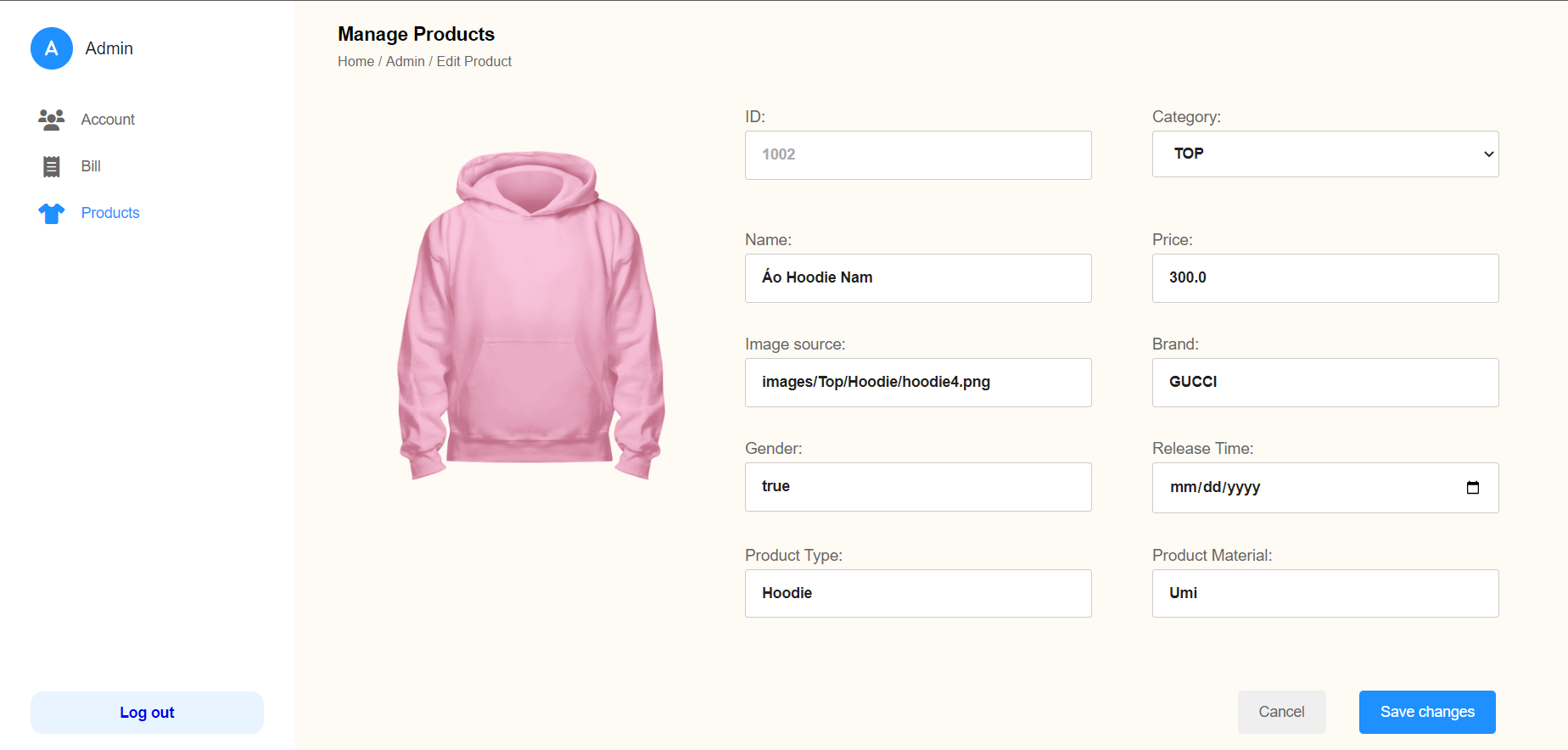
Description automatically generated with low confidence

*Hình 8: Danh sách sản phẩm khuyến nghị sắp xếp theo Rating*

* + 1. **Giao diện thêm, sửa, xóa sản phẩm**



*Hình 8: Giao diện thêm sản phẩm*



*Hình 9: Giao diện sửa sản phẩm*

* 1. **Chức năng dựa trên tri thức**
* **Bước 1:**
* **TH 1:** Dựa vào nguồn dữ liệu trong CSDL, hệ thống sẽ tiền xử lý dữ liệu và đưa vào thuật toán Elbow và K-Means để tim K cụm phù hợp và phân cụm sản phẩm sau đó lưu vào CSDL.
* **TH 2:** Dựa vào lịch sử xem sản phẩm của khách hàng, hệ thống sẽ tính toán giá trị trung bình (mean value) của các sản phẩm khách hàng đã xem trong lịch sử và dùng model K-means để phân cụm dòng dữ liệu trung bình này. Sau khi phân cụm, hệ thống dùng giá trị cụm để lấy ra những sản phẩm thuộc cùng cụm (những sản phẩm tương đồng) và sắp xếp theo điểm đánh giá (rating) từ cao đến thấp để khuyến nghị cho người dùng.
* **Bước 2:** Khi người dùng đăng nhập vào website sẽ lưu lại thời gian cũng như id session. Người dùng tiến hành xem sản phẩm, khi có ý định mua sản phẩm, người dùng sẽ tiến hành xem chi tiết sản phẩm. Khi xem chi tiết sản phẩm, hệ thống sẽ lấy thông tin của sản phẩm đó ra là xem xét coi sản phẩm đó thuộc cụm nào trong số K cụm mà thuật toán đã chia trong CSDL.
* **Bước 3:** Sau khi đã xác định được sản phẩm thuộc cụm nào, hệ thống sẽ tiến hành chọn ra top 7 những sản phẩm có rating cao nhất trong số những sản phẩm thuộc cụm đó và tiến hành đưa ra giao diện danh sách các sản phẩm đề xuất.
* **Bước 4:** Dựa vào số lần khách hàng click vào sản phẩm mà hệ thống đề xuất và số lượng khách hàng truy cập để tính toán xem mức độ hiệu quả của hệ thống đề xuất.
  1. **Kết luận:**

Chương 4 đã phần nào mô tả được giao diện sản phẩm website bán thời trang cũng như giao diện sản phẩm thông minh được khuyến nghị. Đồng thời giải thích cáchsơ lược cách mà hệ thống khuyến nghị sản phẩm thời trang thông minh hoạt động.

# 

# CHƯƠNG 5: KẾT QUẢ THỰC HIỆN THỬ NGHIỆM

## 5.1. Phân cụm dữ liệu bằng K-means

Khi người dùng mới truy cập chưa có lịch sử xem hàng thì thuật toán K-Means được áp dụng với việc lấy dữ liệu từ CSDL bao gồm **1985** sản phẩm thời trang, sau khi tiến hàng áp dụng Elbow, thuật toán đã chia dữ liệu thành **4** cụm sản phẩm thời trang với những thuộc tính tương đồng.

Hệ thống khuyến nghị sản phẩm có thu thập dữ liệu của người dùng bao gồm thời gian, các sản phẩm người dùng đã xem và những sản phẩm khuyến nghị mà người dùng cảm thấy phù hợp để phân tích người dùng. Ngoài ra hệ thống sử dụng lịch sử xem sản phẩm trong một phiên để gợi ý sản phẩm tương đồng bằng cách tính giá trị trung bình của các thuộc tính của những sản phẩm khách hàng đã xem trong phiên đó.

**5.1.1. Đánh giá kết quả phân k cụm bằng K-Means**

Với 4 cụm dữ liệu mà thuật toán K-means đã chia, ta nhận thấy những thuộc tính nào càng có ít giá trị thì càng ảnh hưởng đến đặc trưng của cụm.

Độ ảnh hưởng của thuộc tính: Category, Gender -> Price -> ReleaseTime -> ProductMaterial -> ProductType -> Brand.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cụm | Category | Price | Brand | Gender | Release  Time | Product  Type | Product  Material |
| 0 | Áo | Trải đều qua các mức giá, giá 800 chiếm số lượng nhiều nhất | Trải đều 11 nhãn hàng, nhiều nhất là hãng H&M | Nữ | Trải đều qua 5 mốc thời gian ra mắt, nhiều nhất là 2018 | Trải đều qua 8 kiểu, nhiều nhất là Jacket | Trải đều qua 6 chất liệu, nhiều nhất là Linen |
| 1 | Áo | Trải đều qua các mức giá, giá 150 chiếm số lượng nhiều nhất | Trải đều 11 nhãn hàng, nhiều nhất là hãng Cartier | Nam | Trải đều qua 5 mốc thời gian ra mắt, nhiều nhất là 2018 | Trải đều qua 8 kiểu, nhiều nhất là Hoodie | Trải đều qua 6 chất liệu, nhiều nhất là Cotton |
| 2 | Quần | Trải đều qua các mức giá, giá 900 chiếm số lượng nhiều nhất | Trải đều 11 nhãn hàng, nhiều nhất là hãng Nike | Nữ | Trải đều qua 5 mốc thời gian ra mắt, nhiều nhất là 2018 | Trải đều qua 8 kiểu, nhiều nhất là Legging | Trải đều qua 6 chất liệu, nhiều nhất là Cotton |
| 3 | Quần | Trải đều qua các mức giá, giá 500 chiếm số lượng nhiều nhất | Trải đều 11 nhãn hàng, nhiều nhất là hãng Nike | Nam | Trải đều qua 5 mốc thời gian ra mắt, nhiều nhất là 2019 | Trải đều qua 8 kiểu, nhiều nhất là Baggy | Trải đều qua 6 chất liệu, nhiều nhất là Jean |

Qua bảng thống kê trên, nhóm rút ra được đặc trưng của các cụm như sau:

* Cụm 0: thuộc nhóm sản phẩm là áo, có mức giá vừa và cao chiếm phần lớn, dành cho giới tính nữ.
* Cụm 1: thuộc nhóm sản phẩm là áo, có mức giá thấp và vừa chiếm phần lớn, dành cho giới tính nam.
* Cụm 2: thuộc nhóm sản phẩm là quần, có mức giá vừa và cao chiếm phần lớn, dành cho giới tính nữ.
* Cụm 3: thuộc nhóm sản phẩm là quần, có mức giá vừa chiếm phần lớn, dành cho giới tính nam.

**Đánh giá kết quả phân cụm đối với dữ liệu không đầy đủ**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | **Category** | **Price** | **Brand** | **Gender** | **Release**  **Time** | **Type** | **Material** | **Đánh giá** |
| 1 | BOT | 900 | UNIQLO | NULL | NULL | Legging | Linen | Không tương đồng |
| 2 | BOT | 400 | Nike | NULL | NULL | Jeans | Jean | Tương đồng về nhóm sản phẩm và mức giá |
| 3 | BOT | 150 | Nike | NULL | NULL | Jeans | Jean | Tương đồng về nhóm sản phẩm |
| 4 | TOP | 250 | Adidas | Male |  | Hoodie | Cotton | Tương đồng về giới tính |
| 5 | TOP | 500 | Adidas | Female |  | Hoodie | Cotton | Tương đồng về mức giá và giới tính |
| 6 | BOT | 1000 | H&M | Male |  | Cargo | Kate | Tương đồng về mức giá |
| 7 | BOT | 900 | H&M | Female |  | Jeans | Jean | Tương đồng về mức giá và giới tính |
| 8 | TOP | 150 | ZARA | NULL | 2018 | TankTop | Linen | Tương đồng về nhóm sản phẩm |
| 9 | TOP | 800 | GUCCI | NULL | 2022 | Croptop | Jean | Tương đồng về nhóm sản phẩm và mức giá |
| 10 | BOT | 500 | GUCCI | NULL | 2019 | Joggers | Linen | Tương đồng về nhóm sản phẩm, mức giá và năm ra mắt |
| 11 | BOT | 100 | ZARA | NULL | 2020 | Joggers | Cotton | Tương đồng về mức giá |

## 5.2. Đánh giá mức độ hiểu quả của hệ thống

Để đánh giá mức độ hiệu quả của hệ thống, nhóm sử dụng cách tính như sau:

Với mỗi lần khuyến nghị, nếu khách hàng click vào một hoặc nhiều sản phẩm trong danh sách vừa khuyến nghị thì được coi là 1 lượt đề xuất thành công. Ngược lại sẽ được coi là không thành công.

Độ hiệu quả của hệ thống được tính theo công thức: Số lần khuyến nghị thành công / Tổng số lần khuyến nghị.

## Kết luận

Thông qua những dữ liệu từ sản phẩm và cả từ người dùng, ta thấy thuật toán K-Means hoạt động khá nhanh và hiệu quả, đồng thời việc áp dụng elbow tự tim K cụm đã góp phần giải quyết khó khăn chọn cụm của thuật toán đồng thời còn giúp cho hệ thống linh hoạt trong việc có sự thêm, sửa, xóa dữ liệu sản phẩm mà không ảnh hưởng đến hệ thống hoạt động. Việc này đảm bảo cho việc nâng cấp và phát triển hệ thống trong tương lai.

# CHƯƠNG 6: KẾT LUẬN

Với ý tưởng nghiên cứu ứng dụng của thuật toán phân cụm K-Means vào hệ khuyến nghị sản phẩm, đề tài “Hệ thống khuyến nghị sản phẩm thời trang thông minh” đã đạt được một số mục tiêu sau: Luận văn đã trình bày kiến thức tổng quan về hệ khuyến nghị và các phương pháp tiếp cận cơ bản để xây dựng một hệ thống khuyến nghị cho website thời trang của nhóm. Hệ thống hóa cơ sở lý thuyết về hệ khuyến nghị và phân tích, tổng hợp các nghiên cứu liên quan nhằm đề ra quy trình và lựa chọn các công cụ thích hợp để xây dựng hệ thống khuyến nghị. Bước đầu xây dựng được một mô hình kết hợp học máy vào thương mại điện tử với quy mô gần với thực tế và đánh giá được ưu, nhược điểm của mô hình đề xuất.

Về mặt lý thuyết, bài viết có giới thiệu về học máy, thuật toán K-Means và cấu hình thuật toán. Giới thiệu các hàm đo, các cách tối ưu cụm dữ liệu, các yếu tố ảnh hưởng kết quả thuật toán.

Về mặt thử nghiệm, bài báo cáo giới thiệu thuật toán K-Means, sử dụng ngôn ngữ python và java xây dựng phần mềm khuyến nghị sản phẩm thời trang. Phần mềm này ứng dụng vào việc hỗ trợ người dùng, đóng vai trò phần nào làm người tự vấn cho khách hàng, giúp khách hàng tiết kiệm thời gian khi mua sắm và tìm kiếm sản phẩm.

Về mặt dữ liệu, với nguồn dữ liệu tự random và khai thác từ khách hàng chủ yếu là dạng chuỗi nên ta tiền xử lý dữ liệu để bộ dữ liệu có kết quả tốt hơn.

Về mặt ứng dụng, Giao diện của ứng dụng thân thiện, đơn giản, dễ sử dụng. Chương trình đã xây dựng websites với những chức năng cơ bản và đơn giản nhất:

* + - Thêm, sửa, xóa sản phẩm
    - Sử dụng tri thức, tự động đề xuất, khuyến nghị sản phẩm phù hợp
    - Xem danh sách sản phẩm

Tuy nhiên, ứng dụng này vẫn cần thay đổi và bổ sung thêm chức năng cần thiết của một website. Dựa trên cơ sở kế thừa những gì đã đạt được, tiếp tục sửa chữa và khắc phục hạn chế mà ứng dụng đang hiện có. Hoàn thiện chương trình một cách tối ưu nhất để đem vào sử dụng thực tế và cho người dùng sử dụng một cách thuận tiện nhất.

Trong quá trình thực hiện bài, nhóm đã cố gắng tập trung tìm hiểu và tham khảo các tài liệu liên quan. Tuy nhiên, với thời gian và trình độ có hạn nên không tránh khỏi những hạn chế và thiếu sót. Nhóm rất mong nhận được các nhận xét và góp ý của các thầy để hoàn thiện hơn các kết quả nghiên cứu.