**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG CƠ SỞ TẠI THÀNH PHỐ HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN II**

-----

|  |
| --- |
|  |
|  |  |

**THIẾT KẾ HỆ THỐNG THÔNG MINH**

**BÁO CÁO ĐỀ TÀI**

**THIẾT KẾ HỆ THỐNG TRI THỨC ĐƯA KIẾN NGHỊ DỰA TRÊN PHONG CÁCH THỜI TRANG CHO KHÁCH HÀNG TRÊN WEBSITE**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **GVHD: Nguyễn Ngọc Duy**   |  |  | | --- | --- | | Giang Mạnh Tuấn | N19DCCN176 | | Nguyễn Minh Thuận | N19DCCN202 | | Nguyễn Xuân Thường | N19DCCN206 | | Nguyễn Xuân Tiến | N19DCCN169 | | Nguyễn Thành Trung | N19DCCN216 | | D19CQCNPM02 | | |

**TPHCM, tháng 12 năm 2022**

**MỤC LỤC**

[**I.** **GIỚI THIỆU ĐỀ TÀI** 4](#_Toc122902773)

[1. Lý do chọn đề tài 4](#_Toc122902774)

[2. Mục Tiêu Nghiên Cứu 5](#_Toc122902775)

[3. Mô Tả Đề Tài 5](#_Toc122902776)

[4. Đối Tượng Và Phạm Vi Nghiên Cứu 5](#_Toc122902777)

[1. Đối tượng nghiên cứu: Dữ liệu về áo. 5](#_Toc122902778)

[2. Phạm vi nghiên cứu: Dữ liệu áo của cửa hàng online. 5](#_Toc122902779)

[**II.** **CƠ SỞ LÝ THUYẾT VÀ CÔNG NGHỆ** 5](#_Toc122902780)

[1. Giới Thiệu Về Học Máy: 5](#_Toc122902781)

[2. Khái Quát Về Phân Lớp: 6](#_Toc122902782)

[3. Thuật Toán Máy Vector Hỗ Trợ (Support Vector Machines – SVM): 6](#_Toc122902783)

[**3.1**  **Phương Pháp SVM** 7](#_Toc122902784)

[**3.2** **Dữ liệu chồng nhau và phương pháp biên mềm** 10](#_Toc122902785)

[**3.3 Dữ liệu phân tách phi tuyến và phương pháp kernel** 12](#_Toc122902786)

[**3.4** **Ưu và Nhược Điểm** 13](#_Toc122902787)

[ Ưu điểm: 13](#_Toc122902788)

[ Nhược Điểm: 14](#_Toc122902789)

[4. Tiền Xử Lý Dữ Liệu 14](#_Toc122902790)

[**III.** **GIỚI THIỆU VÀ ĐẶC TẢ DỮ LIỆU** 15](#_Toc122902791)

[1. Tổng quan về các phong cách thời trang thường thấy 15](#_Toc122902792)

[2. Các yếu tố xác định phong cách của người dùng 16](#_Toc122902793)

[3. Dữ liệu 20](#_Toc122902794)

[**Nghiệp vụ:** 20](#_Toc122902795)

[**Xác định các thuộc tính:** 21](#_Toc122902796)

[4. Chuẩn bị Dữ liệu (DATA preparation) 22](#_Toc122902797)

[5. Tiền xử lý Dữ liệu (DATA PREPROCESSING) 23](#_Toc122902798)

[6. Trích chọn thuộc tính (FEATURE EXTRACTION): 25](#_Toc122902799)

[7. Đánh giá dữ liệu qua công cụ WEKA 26](#_Toc122902800)

[IV. XÂY DỰNG CẤU HÌNH: 30](#_Toc122902801)

[1. Đọc dữ liệu từ file csv: 30](#_Toc122902802)

[2. Cấu hình SVM và xuất file đã được train: 30](#_Toc122902803)

[3. Quy đổi chuyền dữ liệu từ Java sang Python: 32](#_Toc122902804)

[4. Đưa dữ liệu vào huấn luyện: 33](#_Toc122902805)

[**V.** **PHÂN TÍCH THIẾT KẾ HỆ THỐNG** 34](#_Toc122902806)

[1. Thiết kế cơ sở dữ liệu: 34](#_Toc122902807)

[2. Thiết kế giao diện: 41](#_Toc122902808)

[3. Thiết kế giao diện thông minh: 45](#_Toc122902809)

[4. Chức năng dựa trên tri thức: 50](#_Toc122902810)

[**VI. KẾT LUẬN 51**](#_Toc122902813)

[**VII. TÀI LIỆU THAM KHẢO 51**](#_Toc122902814)

1. **GIỚI THIỆU ĐỀ TÀI**
   1. **Lý do chọn đề tài**

Công nghệ thông tin và thương mại điện tử đã được phát triển và ứng dụng rộng rãi vào đời sống xã hội nói chung và doanh nghiệp nói riêng. Đối với doanh nghiệp, thương mại điện tử góp phần hình thành những mô hình kinh doanh mới, tăng doanh thu, giảm chi phí, nâng cao hiệu quả kinh doanh và mở ra một thị trường rộng lớn với mọi đối tượng khách hàng trong và ngoài nước. Giả sử bạn có một gian hàng rất lớn, có thể chứa tới vài ngàn sản phẩm. Theo nguyên lý pareto thì 20% các sản phẩm quan trọng nhất sẽ mang lại 80% doanh thu, trong khi không gian trưng bày là có hạn nên bạn không thể show ra hết 100% các sản phẩm bạn có. Một cách thông thường nhất mà chúng ta thường nghĩ tới đó là:

* Trưng bày những sản phẩm phổ biến hoặc bán chạy nhất tại vị trí dễ tiếp cận nhất.
* Gom các sản phẩm có công dụng tương tự nhau thành các nhóm sản phẩm và trưng bày một ít sản phẩm trong mỗi nhóm tới khách hàng.

Tuy nhiên việc này có nhược điểm là không phải 100% các khách hàng tìm đến đều có nhu cầu mua những sản phẩm phổ biến nhất. Do đó cần có một thuật toán hiểu được với từng nhóm khách hàng có đặc điểm cụ thể sẽ mua gì? cửa hàng cần bày trí các sản phẩm nào? tại đâu? đối với khách hàng nào? thì sẽ tối ưu được doanh thu, tận dụng được nguồn hàng sẵn có và gia tăng khả năng bán hàng.

Đây chính là một trong những yếu tố then chốt tạo nên sự khác biệt giữa các sàn thương mại điện tử vì số lượng hàng hóa của một sàn thương mại điện tử có thể lên tới vài triệu nhưng giao diện chỉ cho phép hiển thị một số lượng ít các sản phẩm. Vấn đề cá nhân hóa (personalization) để đưa ra những hiển thị sản phẩm hợp lý nhất tới từng cá nhân người dùng sẽ là bài toán mà mọi sàn thương mại điện tử luôn tìm cách cải thiện. Có nhiều cách để thực hiện việc cá nhân hóa người dùng và hệ thống khuyến nghị sẽ giúp bạn thực hiện việc đó. Công cụ này được xây dựng dựa trên cơ sở thu thập và phân tích những thông tin về các sản phẩm, khách hàng trong hệ thống để dự đoán mức độ quan tâm của khách hàng tới từng sản phẩm và gợi ý cho khách hàng mua những sản phẩm đó.

Một hệ thống khuyến nghị sản phẩm được xây dựng dựa trên các kỹ thuật như khuyến nghị dựa trên nội dung để khuyến nghị cho khách hàng những sản phẩm tương tự với những sản phẩm mà khách hàng đó đã mua hoặc đã quan tâm, khuyến nghị những sản phẩm mà những khách hàng tương tự khác cũng quan tâm hoặc đã mua, khuyến nghị sản phẩm dựa trên đặc trưng trích xuất từ lựa chọn của khách hàng.

Trong nghiên cứu này, nhóm tập trung nghiên cứu công cụ khuyến nghị sản phẩm được xây dựng trên cơ sở kỹ thuật khuyến nghị dựa trên đặc trưng trích xuất từ lựa chọn khách hàng.

* 1. **Mục Tiêu Nghiên Cứu**

Mô hình thuật toán Vector Machine Support (SVM).

Ứng dụng thuật toán phân lớp Vector Machine Support (SVM) phân loại phong cách thời trang của người dùng đưa ra khuyến nghị sản phẩm phù hợp.

* 1. **Mô Tả Đề Tài**

Từ tập dữ liệu sản phẩm áo đã có của cửa hàng tiến hành xử lý phân loại dán nhãn cho các sản phẩm chia thành các phong cách khác nhau như: elegant (lịch lãm – sang trọng), simple (đơn giản), active (năng động).

Người dùng lựa chọn những đặc trưng phong cách mong muốn về sản phẩm muốn tìm kiếm (chất liệu vải, form, màu sắc) hệ thống sử dụng thuật toán SVM phân loại tìm ra phong cách của khách hàng sau đó đề xuất sản phẩm liên quan.

* 1. **Đối Tượng Và Phạm Vi Nghiên Cứu**
     1. Đối tượng nghiên cứu: Dữ liệu về áo.
     2. Phạm vi nghiên cứu: Dữ liệu áo của cửa hàng online.
  2. **Phương pháp nghiên cứu**

**Phương pháp thực nghiệm khoa học:** Dựa vào tập dữ liệu của cửa hàng, tiến hành tiền xử lý và sử dụng các thuật toán phân lớp svm, đánh giá tính hiệu quả của thuật toán dựa trên tập dữ liệu kiểm thử.

1. **CƠ SỞ LÝ THUYẾT VÀ CÔNG NGHỆ**
   1. **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.

Có 2 loại ML chính bao gồm học có giám sát (supervised learning) và học không giám sát (unsupervised learning):

* Học có giám sát: máy tính học cách mô hình hóa các mối quan hệ dựa trên dữ liệu được gán nhãn (labeled data). Sau khi tìm hiểu cách tốt nhất để mô hình hóa các mối quan hệ cho dữ liệu được gắn nhãn, các thuật toán được huấn luyện được sử dụng cho các bộ dữ liệu mới.
* Học không giám sát: máy tính không được cung cấp dữ liệu được dán nhãn mà thay vào đó chỉ được cung cấp dữ liệu mà thuật toán tìm cách mô tả dữ liệu và cấu trúc của chúng.
  1. **Khái Quát Về Phân Lớp:**

Trong học máy phân lớp (Classification/Categorization) là một khái niệm học có giám sát, về cơ bản phân loại một tập hợp dữ liệu thành các lớp.

Phân lớp thực hiện việc xây dựng (mô tả) các mô hình (hàm) dự báo nhằm mô tả hoặc phát hiện các lớp hoặc khái niệm cho các dự báo tiếp theo. Một số phương pháp điển hình là cây quyết định, luật phân lớp, mạng neuron, SVM. Nội dung của phân lớp chính là học một hàm ánh xạ các dữ liệu vào một trong một số lớp đã biết. Sau khi học được hàm phân lớp, các giải thuật có thể dùng để dự đoán các dữ liệu mới. Tuy nhiên trước khi đem giải thuật vào ứng dụng trong thực tế, các giải thuật phải trải qua bước thứ 2 là bước kiểm tra hiệu năng từ kết quả cho máy học tập trước đó. Để tránh hiện tượng “quá vừa” (overfit), một tập dữ liệu khác được gọi là dữ liệu kiểm thử (testing data) sẽ được sử dụng để đo độ chính xác của giải thuật. Thông thường dữ liệu kiểm thử sẽ không chứa bất kỳ bản ghi nào giống với dữ liệu huấn luyện. Giống như dữ liệu huấn luyện, bộ dữ liệu kiểm thử cũng có nhãn đi kèm đối với từng bản ghi trong nó. Vậy các nhãn này có tác dụng gì? Nó dùng để so sánh với các nhãn có được từ các giải thuật phân lớp. Tỷ lệ dự đoán đúng nhãn của các giải thuật được gọi là độ chính xác (accuracy) của giải thuật. Khi chất lượng phân lớp của các giải thuật là chấp nhận được trong một miền dữ liệu cụ thể nào đó, ta có thể dùng chúng dự đoán lớp của các phần tử dữ liệu mới hoàn toàn chưa biết trước.

* 1. **Thuật Toán Máy Vector Hỗ Trợ (Support Vector Machines – SVM):**

Thuật toán máy vector hỗ trợ (SVM) là một thuộc lớp giải thuật phân lớp thống kê. Ý tưởng của SVM là tìm một siêu phẳng (hyper lane) để phân tách các điểm dữ liệu. Siêu phẳng này sẽ chia không gian n chiều thành các lớp để chúng ta có thể dễ dàng đặc điểm dữ liệu mới vào đúng danh mục trong tương lai. Nó có khả năng xử lý cả dữ liệu tuyến tính và dữ liệu không tuyến tính. Bản chất của giải thuật này là xây dựng một siêu phẳng để phân chia dữ liệu thành 2 nửa. Trong trường hợp nếu dữ liệu là không tuyến tính sử dụng một hàm nhân (kernel function) để chuyển đổi tập dữ liệu ban đầu sang một không gian mới có số chiều lớn hơn để xử lý.

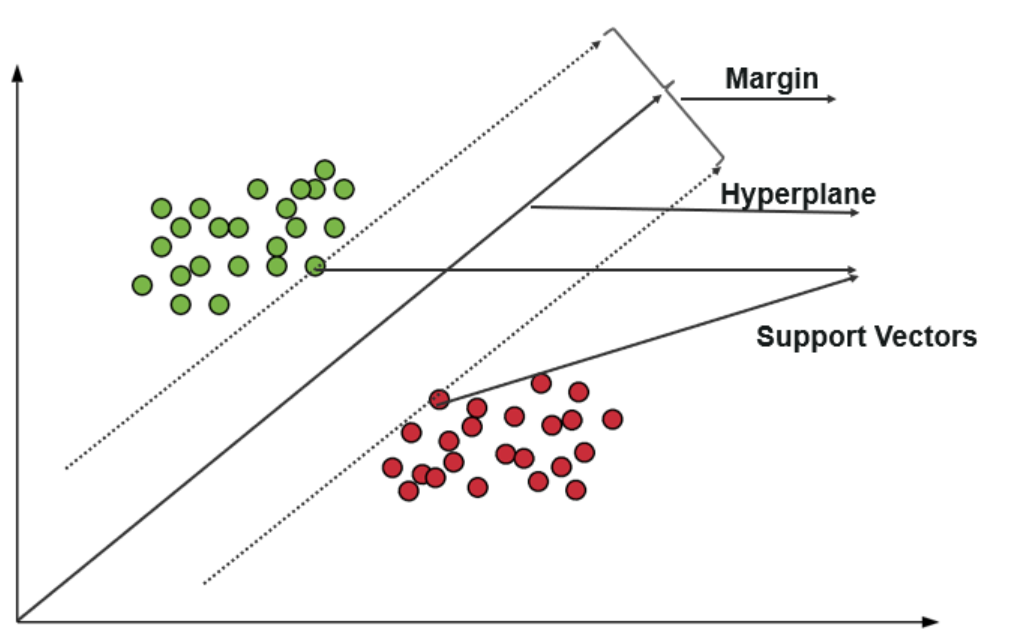


Figure 1: Mô tả về SVM

### **3.1 Phương Pháp SVM**

Như đã biết, với bài toán phân loại nhị phân tuyến tính ta cần vẽ được mặt phân tách (với không gian 2 chiều thì mặt phẳng này là đường phân tách): **wT x**+*b*=0 để phân biệt được dữ liệu.

Khi đó dấu của hàm ước lượng *H* = {**x**↦sgn (**wT x**+*b*); **w**∈RN, *b*∈R} sẽ thể hiện được điểm dữ liệu nằm ở cụm dữ liệu nào.

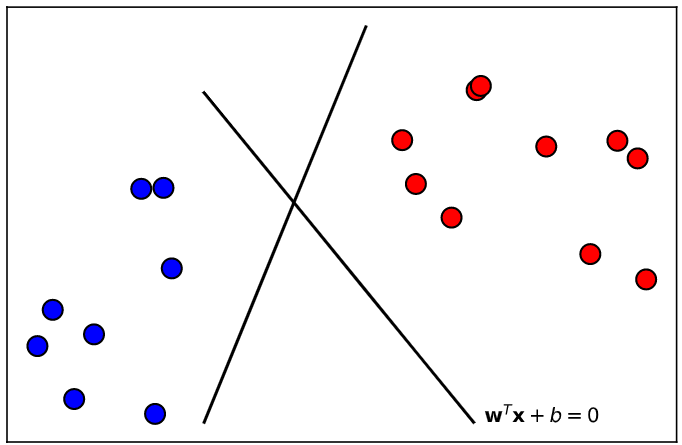


Figure 2: Lựa chọn hyper lane

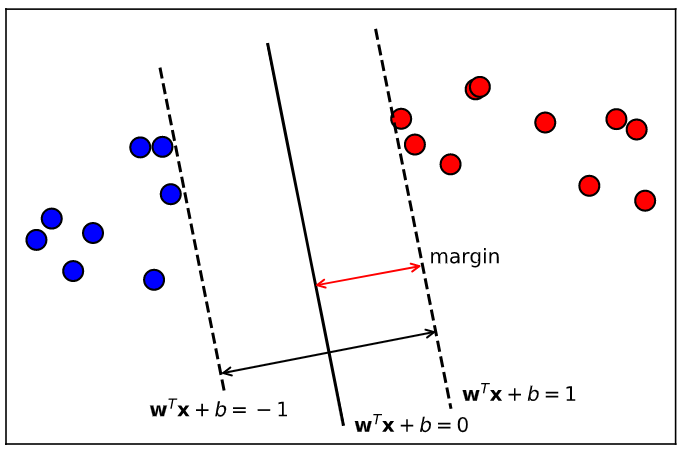
Nếu để ý thì ta có thể có nhiều mặt phân tách thoả mãn được việc này và đương nhiên là nếu chọn được mặt mà phân tách tốt thì kết quả phân loại của ta sẽ tốt hơn. Một lẽ rất tự nhiên là dường như mặt nằm vừa khít giữa 2 cụm dữ liệu sao cho nằm xa các tập dữ liệu nhất là mặt tốt nhất.

Figure 3: Lựa chọn hyper lane tối ưu nhất

**SVM** chính là một biện pháp để thực hiện được phép lấy mặt phẳng như vậy.

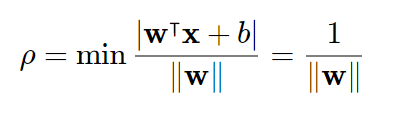
Để xác định mặt phẳng kẹp giữa đó, trước tiên ta cần phải xác định được 2 mặt biên gốc như 2 đường nét đứt ở trên. Các điểm dữ liệu gần với mặt biên gốc này nhất có thể xác định bằng:

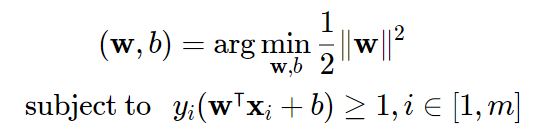
min ∣**wT x**+*b*∣

Để dễ dàng cho việc tính toán thì người ta sẽ chọn **w** và *b* sao cho các điểm gần nhất (mặt biên gốc) thoả mãn: min ∣**wT x**+*b*∣=1 tức là:

min ∣**wT x**+*b*∣=1

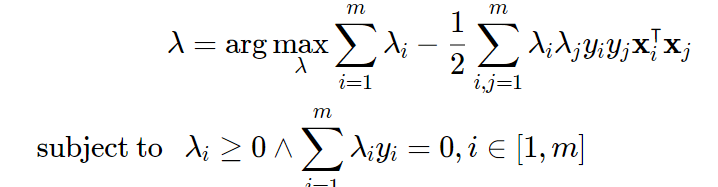
Đương nhiên là có thể tồn tại nhiều cặp đôi mặt biên gốc như vậy và tồn tại nhiều mặt phân đôi kẹp giữa các mặt biên gốc đó. Nên ta phải tìm cách xác định được mặt kẹp giữa tốt nhất bằng cách lấy cặp có khoảng cách xa nhau nhất. Lẽ này là đương nhiên bởi cặp có khoảng cách xa nhất đồng nghĩa với chuyện tập dữ liệu được phân cách xa nhất.

Như vậy, ta có thể thiết lập thông số tính khoảng cách đó bằng phép lấy độ rộng biên từ mặt biên gốc tới mặt phân tách cần tìm.

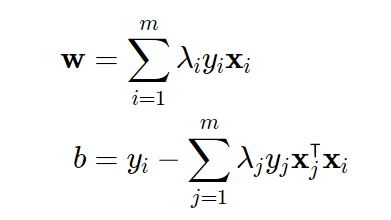
Bài toán của ta bây giờ sẽ là cần xác định **w** và *b* sao cho *ρ* đạt lớn nhất và các điểm dữ liệu *yi*​ (**wT** **x***i*​+*b*) ≥ 1 *ρ* đạt lớn nhất đồng nghĩa với việc ∥**w**∥ đạt nhỏ nhất. Tức là:

Ở đây, *m* là số lượng các điểm dữ liệu (**x***i*​, *yi*​) còn việc lấy bình phương và chia đôi nhằm dễ dàng tính toán và tối ưu lồi.

Bài toán này có thể giải thông qua bài toán đối ngẫu của nó và sử dụng phương pháp nhân tử Lagrance. Lúc này, ta sẽ cần tìm các giá trị *λ* như sau:

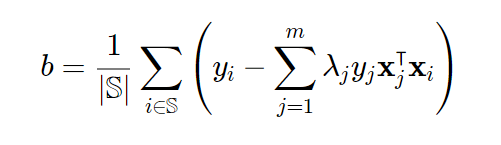


Việc giải *λ* có thể được thực hiện bằng phương pháp quy hoạch động bậc 2 (Quadratic Programing). Với Python ta có thể sử dụng thư viện [CVOPT](http://cvxopt.org/examples/tutorial/qp.html). Sau khi tìm được *λ* thì ta có các tham số:

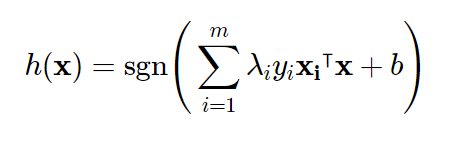


​​

Ở đây, (**x***i*​, *yi*​) là một điểm dữ liệu bất kì nào đó nằm trên đường biên gốc. Điểm dữ liệu này còn được gọi là **Support Vector**. Tên của phương pháp SVM cũng từ đây mà ra. Tuy nhiên, thường người ta tính b*b* bằng phép lấy trung bình tổng của tất cả các *bi.* Giả sử, ta có tập S các Support Vectors thì:



Khi đó, một điểm dữ liệu mới sẽ được phân loại dựa theo:



Như vậy, chỉ cần các điểm Support Vector trên đường biên gốc là ta có thể ước lượng được các tham số tối ưu cho bài toán. Việc này rất có lợi khi tính toán giúp phương pháp này tiết kiệm được tài nguyên thực thi.

### **3.2 Dữ liệu chồng nhau và phương pháp biên mềm**

Trong thực tế tập dữ liệu thường không được sạch như trên mà thường có nhiễu. Nhiễu ở đây là dạng dữ liệu chồng chéo lên nhau như hình bên dưới.

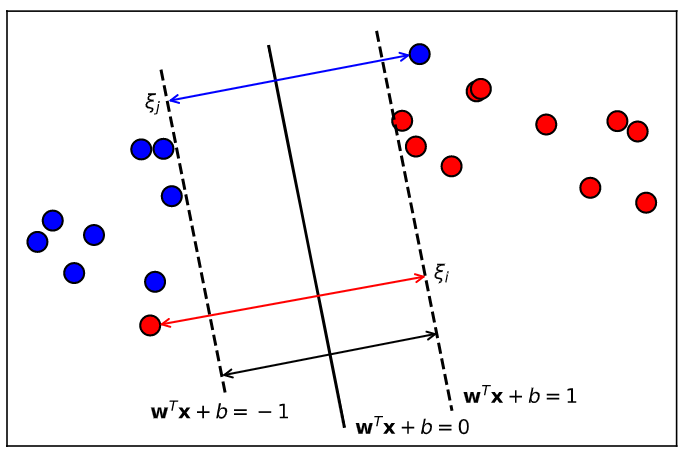


Figure 4: Phương pháp làm mềm biên

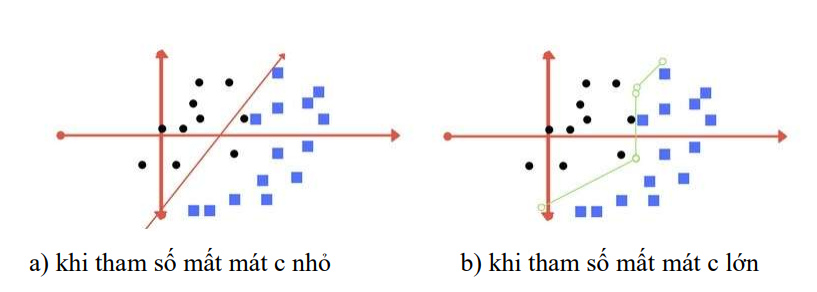
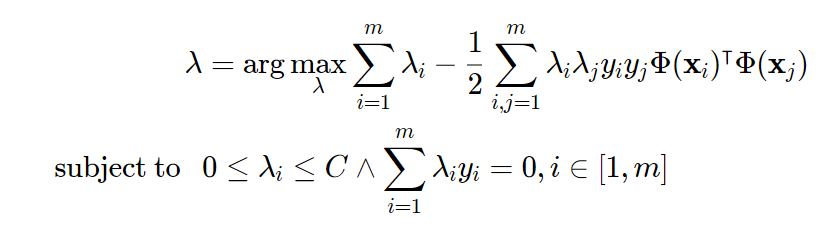
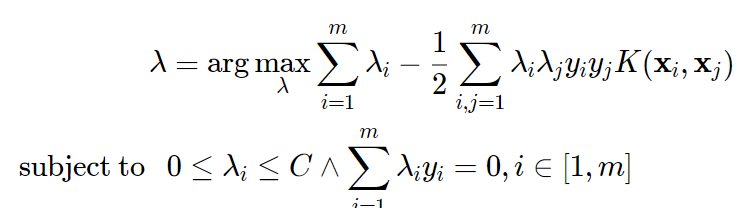
Với dạng dữ liệu như vậy thì mặt phân tách tìm được sẽ khó mà tối ưu được, thậm chí là không tìm được mặt phân tách. chúng ta có thể nới lỏng bằng cách làm rộng lề (margin) hơn một chút, chấp nhận những cá thể bị phân loại sai. Để kiểm soát sai sót của lề mềm trong mức cho phép, mô hình SVM sử dụng đại lượng C biểu thị mức độ phạt khi model dự đoán sai. Khi C càng lớn, SVM model sẽ bị phạt nặng hơn khi dự đoán sai và kết quả sẽ cho ra một lề hẹp. Ở chiều ngược lại khi C càng nhỏ, SVM sẽ càng dễ dãi hơn và kết quả sẽ cho ra một lề rộng. S mở rộng như vậy người ta gọi phương pháp này là phương pháp biên mềm (**Soft-Margin SVM**). Còn phương pháp truyền thống là biên cứng (**Hard-Margin SVM**). Hình ảnh dưới đây cho chúng ta cái nhìn rõ ràng hơn về tính hiệu quả khi sử dụng lề mềm, hình bên phải là siêu phẳng được tạo ra khi sử dụng lề mềm, hình bên trái là siêu phẳng được tạo ra khi sử dụng lề cứng.

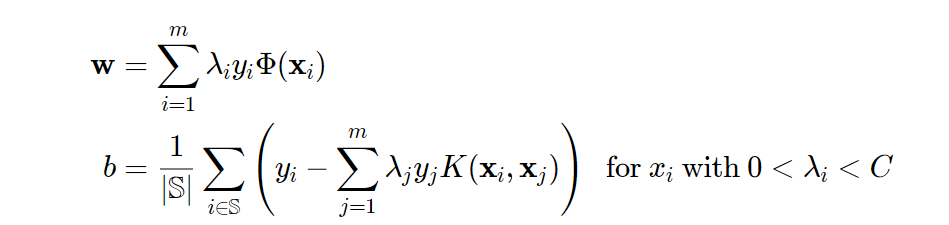
Figure 5: minh họa sự ảnh hưởng của tham số mất mát c

### **3.3 Dữ liệu phân tách phi tuyến và phương pháp kernel**

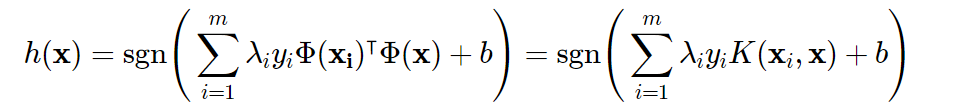
Như các bài trước đã đề cập tới việc sử dụng hàm cơ bản Φ(**x**) để tạo đặc trưng cho tập dữ liệu nhằm nâng được chiều của dữ liệu ban đầu. Bằng các hàm cơ bản này, ta có thể tạo các mặt cong phân tách cho phù hợp với các điểm dữ liệu không phân tách tuyến tính.

Khi đó tối ưu biên mềm được viết dưới dạng:

****Đặt hàm **Kernel** K (**x***i*​, **x***j*​) =Φ(**x***i*​)T Φ(**x***j*​), ta có:

****Khi đó tham số tương ứng sẽ là:

Điểm dữ liệu mới được phân lớp với:

****

Như vậy, chỉ cần hàm Kernel *K* (**x***i*​, **x***j*​) để tính tích vô hướng giữa các điểm dữ liệu trong không gian mới là ta có thể ước lượng được một điểm mới nằm trong phân lớp nào.

Việc sử dụng hàm Kernel ở đây sẽ giúp giảm được công số tính từng hàm Φ và tích vô hướng giữa chúng. Nó có thể tính được cho bất kì không gian nào rất hiệu quả. Kể cả các không gian với số chiều vô hạn. Bởi nó chỉ cần tính tích vô hương giữa các điểm dữ liệu mà thôi. Tất nhiên để làm được điều đó thì Kernel phải thoả mãn điều kiện Mercer.

Khi làm việc người ta thường chọn một hàm Kernel thông dụng sau:

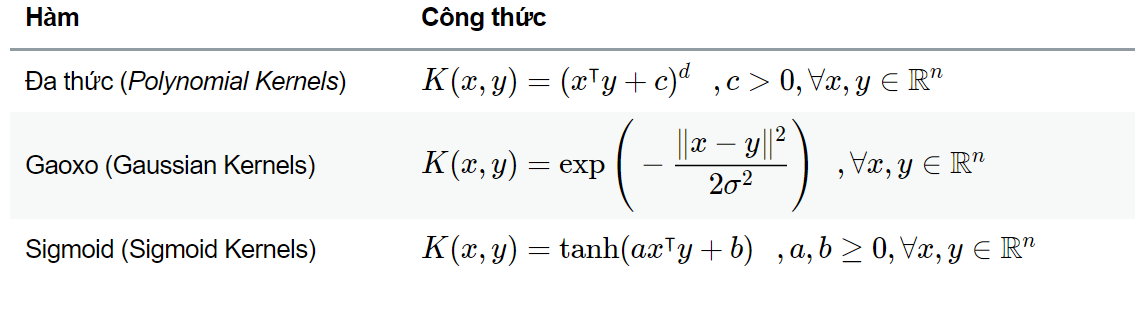
****

Figure 6: Công thức từng hàm

### **3.4 Ưu và Nhược Điểm**

* Ưu điểm:
  + Xử lý trên không gian số chiều cao: SVM là một công cụ tính toán hiệu quả trong không gian chiều cao, trong đó đặc biệt áp dụng cho các bài toán phân loại văn bản và phân tích quan điểm nơi chiều có thể cực kỳ lớn
  + Tiết kiệm bộ nhớ: Do chỉ có một tập hợp con của các điểm được sử dụng trong quá trình huấn luyện và ra quyết định thực tế cho các điểm dữ liệu mới nên chỉ có những điểm cần thiết mới được lưu trữ trong bộ nhớ khi ra quyết định.
  + Tính linh hoạt - phân lớp thường là phi tuyến tính. Khả năng áp dụng Kernel mới cho phép linh động giữa các phương pháp tuyến tính và phi tuyến tính từ đó khiến cho hiệu suất phân loại lớn hơn.
* Nhược Điểm:
  + SVM không có tính chất có thể mở rộng và chúng không hoạt động tốt với các bộ dữ liệu kích thước trung bình hoặc lớn.
  + SVM không phải là thuật toán hiệu quả nhất và việc huấn luyện mô hình có thể khá tốn kém về mặt tính toán. (Khi áp dụng với các kernel và đặc biệt là với các kernel phi tuyến tính).
  1. **Tiền Xử Lý Dữ Liệu**

Có nhiều cách tiền xử lý dữ liệu khác nhau nhằm mục tiêu tăng chất lượng dữ liệu và từ đó có thể làm tăng hiệu quả của các kỹ thuật khai phá dữ liệu. Mỗi một kỹ thuật cụ thể sẽ giúp cải thiện chất lượng dữ liệu theo hướng nhất định và hiệu quả của nó phụ thuộc rất nhiều vào đặc trưng của dữ liệu. Một số kỹ thuật tiền xử lý dữ liệu thường được áp dụng bao gồm:

* **Kỹ thuật làm sạch dữ liệu (data cleaning**) thường được sử dụng để thêm những giá trị bị thiếu, loại bỏ nhiễu, xác định và loại bỏ các giá trị ngoại lai và giải quyết vấn đề không nhất quán của dữ liệu. Hiển nhiên ta sẽ không thể tin tưởng vào kết quả thu được từ bất kỳ thuật toán khai phá dữ liệu nào nếu ta biết chắc rằng dữ liệu còn chưa được làm sạch và có chất lượng tốt. Một số kỹ thuật khai phá dữ liệu đã tích hợp sẵn các modul để loại bỏ nhiễu và xử lý dữ liệu thiếu, tuy nhiên phần lớn chúng hoạt động không thực sự hiệu quả. Vì vậy, thay vì tập trung vào việc làm sạch dữ liệu các thuật toán khai phá dữ liệu có thể tập trung vào việc xây dựng các mô hình hiệu quả hơn. Nhiệm vụ làm sạch dữ liệu sẽ được thực hiện trong quá trình tiền xử lý dữ liệu trước khi sử dụng bất kỳ thuật toán khai phá dữ liệu nào.
* **Kỹ thuật tích hợp dữ liệu (data integration):** cho phép trộn (lắp ghép/ tích hợp) dữ liệu từ nhiều nguồn khác nhau về một kho chứa đồng nhất và có tính gắn kết chặt chẽ phục vụ cho quá trình khai phá dữ liệu tiếp theo. Việc các nguồn dữ liệu khác nhau thì tổ chức và định nghĩa dữ liệu hoàn toàn có thể khác nhau. Quá trình khai phá tri thức sẽ không thể thực hiện, thực hiện chậm hoặc thực hiện không chính xác khi dữ liệu có càng nhiều dữ liệu dư thừa. Hiển nhiên ta thấy trong khi tích hợp dữ liệu các kỹ thuật làm sạch dữ liệu phải được áp dụng nhằm tránh sự dư thừa dữ liệu. Không những thế các kỹ thuật làm sạch còn được áp dụng để phát hiện và loại bỏ các dữ liệu dư thừa sau khi tích hợp dữ liệu từ nhiều nguồn khác nhau.
* **Thu gọn (làm giảm) dữ liệu (data reduction) nhằm** giảm kích cỡ của dữ liệu nhiều nhất có thể mà không làm ảnh hưởng (hoặc ảnh hưởng ở mức chấp nhận được) tới kết quả phân tích. Việc thu gọn dữ liệu thường xảy ra trong trường hợp dữ liệu quá lớn tới mức làm giảm hiệu năng của các kỹ thuật khai phá dữ liệu như thời gian chạy quá lâu hoặc không đủ bộ nhớ để thực hiện… Có hai chiến lược thu gọn dữ liệu là giảm chiều dữ liệu (dimensionality reduction) và giảm số lượng dữ liệu (numeosity reduction).
* **Kỹ thuật chuyển dạng dữ liệu (data transformation)** có thể ứng dụng với dữ liệu có phân bổ không phù hợp với các thuật toán phân tích dữ liệu dựa trên khoảng cách như mạng nơron, phân lớp K-NN,… Với kỹ thuật khai phá dữ liệu này, thông thường dữ liệu cần được chuẩn hóa về cùng một miền dữ liệu thì các độ đo khoảng cách mới được áp dụng một cách hiệu quả. Phương pháp rời rạc hóa (discretization) và tạo cây phân cấp khái niệm (concept hierarchy generation) dữ liệu cũng là những kỹ thuật rất hiệu quả trong việc chuyển dạng dữ liệu.

1. **GIỚI THIỆU VÀ ĐẶC TẢ DỮ LIỆU**
   1. **Tổng quan về các phong cách thời trang thường thấy**

***Phong cách đơn giản (minimalism):*** Đơn giản, nhẹ nhàng nhưng cực kỳ hiện đại và thanh lịch biểu hiện qua những gam màu đơn sắc, các thiết kế thường là sự tổng hòa của những đường nét cơ bản, không rườm rà, phức tạp.

***Phong cách đơn giản (minimalism):*** Đơn giản, nhẹ nhàng nhưng cực kỳ hiện đại và thanh lịch biểu hiện qua những gam màu đơn sắc, các thiết kế thường là sự tổng hòa của những đường nét cơ bản, không rườm rà, phức tạp.

Phong cách minimalism, nói cách khác, là “back to basic” – giản lược mọi thứ, giữ mọi thứ ở mức đơn giản nhất có thể.

Tuy nhiên tối giản không có nghĩa là đơn điệu, minimalism kiến tạo biểu hiện hình dáng cơ thể con người thông qua những đường cắt may tinh xảo, mang đến cho người mặc một sự nổi bật đầy thoải mái, toát lên sự sang trọng và tinh tế khi mặc đem lại một sự thời thượng vô cùng êm ái.

***Phong cách thanh lịch (elegant):***  **Phong cách thời trang elegant**được hiểu là cách ăn mặc tinh tế tế và thanh lịch, phong cách này luôn lọt top các style thời trang sành điệu và đẳng cấp suốt nhiều thập kỷ qua.

Phong cách thanh lịch không chú trọng trang phục đắt tiền, phụ kiện lấp lánh hào nhoáng. Phong cách thanh lịch đáng quý ở chỗ tinh giản nhưng khí chất, chú trọng lối ăn mặc đơn thuần, không lòe loẹt, bắt nguồn từ sự tinh giản và thoải mái, không cần có quá nhiều lớp áo, phụ kiện. Thay vào đó, chú ý tới vào chất liệu, kiểu váy áo và kết hợp phụ kiện vừa đủ.

***Phong cách năng động (*Active Casual*):*** Active Casual style được ra đời vào những năm đầu của thế kỷ 20 và bắt nguồn từ nước Mỹ. Trước khi trở thành một trong những phong cách thịnh hành nhất hiện nay, thời trang Active Casual bắt đầu gói gọn trong các trang phục thể thao, những chiếc áo khoác blazer vải tuýt hay đôi giày oxford. Theo thời gian, phong cách này ngày càng lan tỏa hơn nữa và tiến tới qua bộ trang phục công nhân và cả những bộ trang phục của quân đội*.*

Với tính ứng dụng cao và sự tiện dụng, phong cách Active Casual đã nhanh chóng lan tỏa khắp thế giới từ châu âu đến châu á.Nếu bạn là người yêu thích sự năng động, thoải mái và sôi nổi, thì Active Casual sẽ là phong cách phù hợp với bạn.

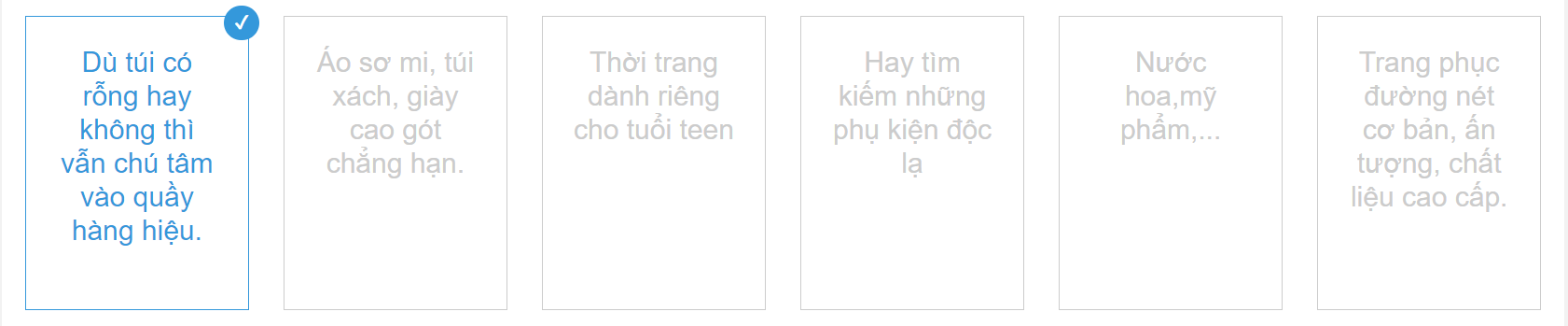
* 1. **Các yếu tố xác định phong cách của người dùng**

Xác định phong cách thời trang của một người là điều không dễ bởi phong cách thời trang có tính chất tương đối. Nhưng có thể sử dụng những đặc điểm sau để xác định được phong cách thời trang của một người:

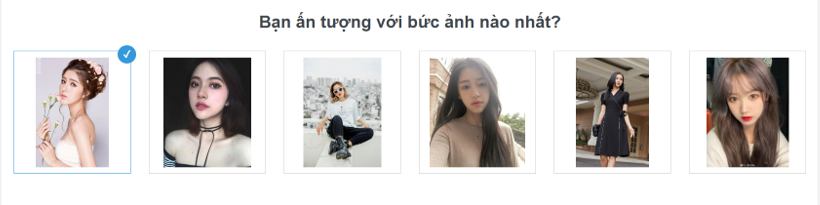
1. Nghĩ về những bộ trang phục làm bạn vui. Món đồ ưa thích trong bộ trang phục của bạn?
   * Phong cách đơn giản (minimalism): bộ trang phục cơ bản, ấn tượng, chất liệu cao cấp. Món đồ ưa thích như áo trắng đen, blazer trung tính, váy liền trơn đơn sắc, …
   * Phong cách năng động (active): Với những items khá quen thuộc như quần jeans, áo thun, sweater… kết hợp cùng với giày sneakers, túi đeo. → Trang phục đường nét cơ bản,ấn tượng,chất liệu cao cấp,thời trang dành cho tuổi teen.
   * Phong cách thanh lịch (elegant): áo sơ mi trắng oversize, túi sách, giày cao gót, blaze,…
2. Xác định nguồn cảm hứng thời trang. Tìm kiếm cảm hứng thông qua hình ảnh, chọn ra hình ảnh bạn cảm thấy nổi bật, có thể có nhiều hình ảnh nổi bật đối với bạn nhưng chắc chắn chúng có điểm chung, điểm chung đó là một phần phong cách của bạn.
3. Xác định dấu ấn đặc trưng. Có thể quan sát học hỏi từ nhiều nguồn cảm hứng nhưng phong cách của một người luôn thể hiện dấu ấn riêng của người đó thông qua gam màu yêu thích, món phụ kiện hay họa tiết nói lên cá tính của bạn, đôi khi kiểu tóc hay phong cách trang điểm cũng là đặc điểm xác định phong cách thời trang của bạn.
   * Phong cách đơn giản (minimalism): thường giảm thiểu phụ kiện có thể là những trang sức nhỏ nhắn, túi xách hay giày cao gót đơn giản, gam màu ưu thích là đen, trắng, tím, xanh dương đậm.
   * Phong cách năng động (active): Thường có ít phụ kiện,đi kèm một đôi giày màu sắc trẻ trung,kiểu dáng thoải mái.
   * Phong cách thanh lịch (elegant): những màu ưa thích như pastel, trắng, nude, xám, hồng, kiểu tóc thường thấy là tóc đuôi ngựa bóng mượt hoặc tóc xoăn bồng bềnh, phụ kiện thường là giày cao gót mũi trơn, đồng hồ hiệu hoặc những phụ kiện nhỏ nhắn.
4. Nhiều thí nghiệm đã được các nhà khoa học tiến hành chứng minh rằng, thời trang là tấm gương phản chiếu tính cách và tâm lý chủ nhân. bởi nó cho thấy cái nhìn tổng quan về những gì một con người suy nghĩ cũng như tính cách "thu nhỏ” của họ. Tính cách là đại diện của mỗi cá nhân. Quần áo thể hiện những suy nghĩ, và tâm hồn bên trong đó. Nhưng vấn đề là chính người mặc sẽ không phát hiện ra mối liên hệ đó nếu họ không được chỉ ra tính cách của mình.
   * Phong cách đơn giản (minimalism): không thích sự bừa bộn nhà của họ thường sạch sẽ và ngăn nắp, chỉ trưng bày những món đồ yêu thích hoặc thiết yếu nhất. Họ ưu tiên sự đơn giản và đa dụng trong mọi lĩnh vực của cuộc sống. Người theo chủ nghĩa tối giản thường là những cá nhân rất độc lập, quyết đoán, hành động theo nhịp trống của riêng họ.
   * Phong cách thanh lịch (elegant): họ thường là những người đĩnh đạc và thanh lịch mang trong mình vẻ duyên dáng có sự tinh tế và gu thẩm mĩ tốt, người thể hiện cách cư xử tốt, hiểu biết sâu rộng và phong thái tự tin.
   * Phong cách năng động (active):Phong cách phối đồ này sẽ rất thích hợp cho những người vui vẻ hòa đồng,tích cực.Thích hợp cho những buổi đi chơi, hẹn hò mà thích sự gọn nhẹ, khỏe khoắn và lịch sự.

Dựa vào những đặc điểm trên nhóm xây dựng bộ test xác định phong cách thời trang của người dùng bao gồm 9 câu hỏi mỗi câu hỏi gồm 6 đáp án:

1. Phong cách mua sắm của bạn là j?



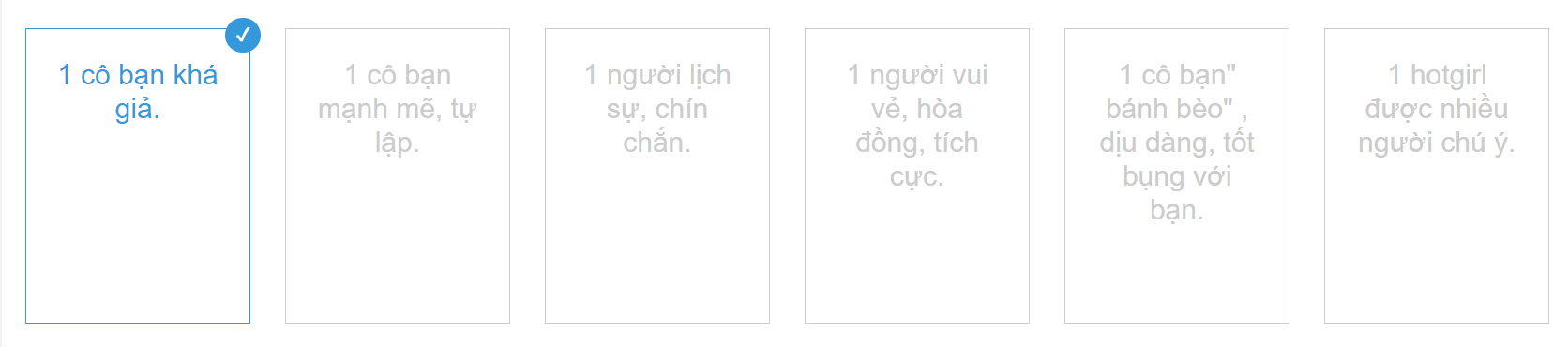
1. Bức ảnh gây ấn tượng với bạn?



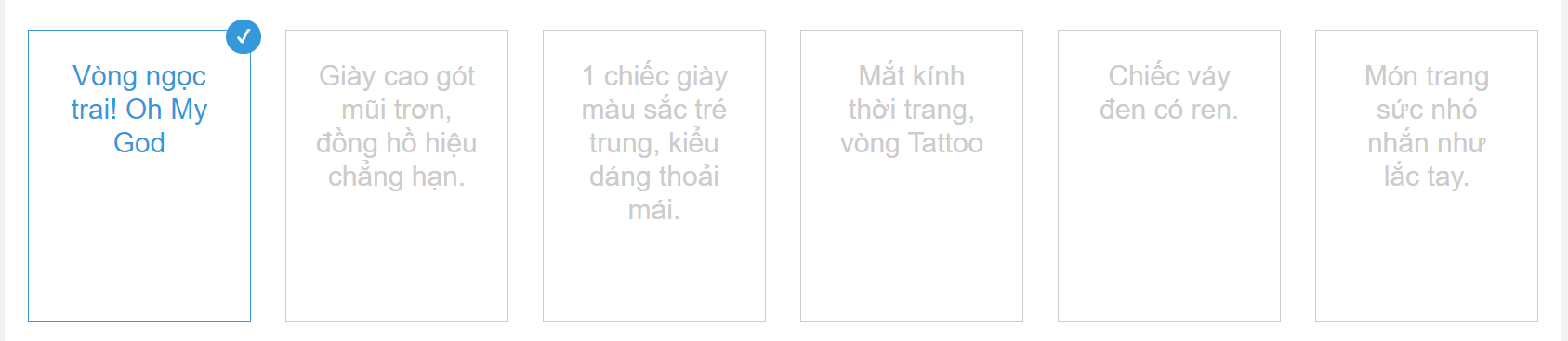
1. Màu sắc mà bạn ưa chuộng?



1. Bạn nghĩ người tâm đầu ý hợp với mình sẽ là?



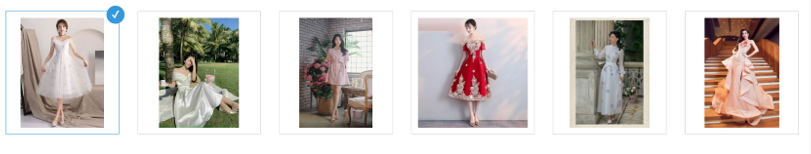
1. Khi bạn được tặng một món quà, bạn mong đó sẽ là?



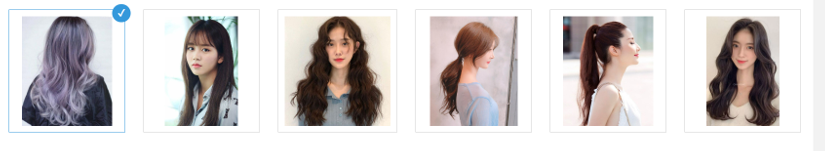
1. Trả lời thật nhé! Đáp án nào không đúng về con người bạn?



1. Đứa bạn thân mời bạn đến dạ hội khi đó bạn sẽ mặc chiếc váy như thế nào?



1. Mái tóc bạn ưa thích?



1. 2 từ khóa nói lên tính cách của bạn.



* 1. **Dữ liệu**

**Nghiệp vụ:**

Website hệ thông bán hàng của nhóm có hai chế độ giao diện phía quản lý và giao diện phía khách hàng. Áp dụng hệ thống tri thức dựa trên kiến nghị được nhóm làm trên giao diện phía khách hàng. Còn giao diện phía quản lý dùng để cập nhật, thêm sửa sản phẩm của hệ thống bán hàng.

Khi quản lý muốn thêm một sản phẩm mới chưa tồn tại trong database thì người quản lý sẽ tiến hành nhập lưu thông tin sản phẩm: tên sản phẩm, danh mục và các thông số khác, … *Dựa vào các thông số đó người quản lý sẽ phân tích và nhận định sản phẩm đó thuộc phong cách nào* (elegent: Thanh lịch, sang trọng – active: năng động – minimalist: tối giản, giản đơn).

\* Dựa vào đâu người quản lý quyết định sản phẩm đó thuộc phong cách nào?

Phong cách thời trang là đa dạng để quyết định gán nhãn như vậy thì nhóm đã dựa trên những bài viết khoa học, chọn lọc thông tin trên Internet và qua những nhận xét của chuyên gia. Dưới đây là link đính kèm nhóm phần tích tìm hiểu phong cách thời trang cho website. [[THU NHẬP THÔNG TIN]](https://docs.google.com/document/d/1GGnvaB646n4HO8NfKLCgiw89qIp-xA3AOAukatuQw0o/edit?usp=sharing)

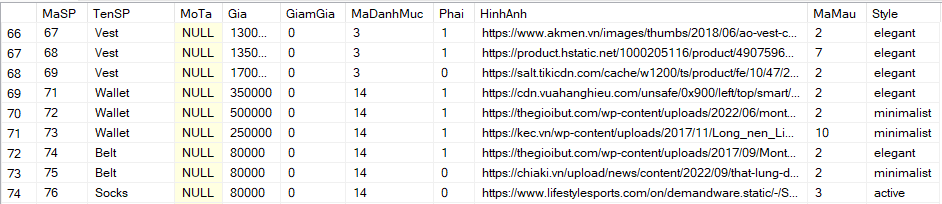
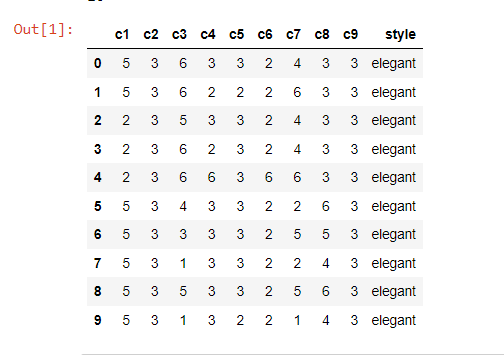
**Xác định các thuộc tính:**

Figure 7: Ảnh minh họa table Sản phẩm trong database

Column c1: chứa đáp án câu hỏi số 1 theo người dùng chọn.

Figure 8: Hình ảnh dataset

Lần lượt c2, c3, … c9 đều là các câu hỏi trong bộ quiz.

Column cuối cùng là có tên là ‘Style’ dùng để gán nhãn.

Style có giá trị:

Style **elegant:** mang phong cáchlịch lãm, sang trọng, thanh lịch, …

* + - * Type: shirt, trouser, jeans, suit (mức độ sang trọng cao nhất)
      * Model: Trouser\_pants (quần tây), trouser\_kaki, jeans\_classic, jeans\_modern, shirt\_long, two (suit 2 mảnh), three (suit 3 mảnh)
      * Color: navy, silver, black, white
      * Material: kaki, cotton, jean, silk (lụa)
      * Form: skinny (ôm), slimfit (vừa vặn)

Style **active:** mang phong cáchnăng động, trẻ trung, vui tươi, …

* + - * Type: shorts (quần short), jeans, jacket, tshirt
      * Model: tshirt-crewneck (thun cổ tròn), tshirt-polo (thun cổ trụ), tshirt-hoodie, jacket-sweater, short-kaki, short-sport
      * Color: red, orange, yellow, green, blue (xanh da trời)
      * Material: cotton\_CM (phù hợp co giãn tốt), cotton\_CVC, vải Tici (mức độ co giản kém), polyester (rẻ nhất)
      * Form: straight (rộng rãi), elastic (co giãn), chỉ có quần jeans rộng

Style **minimalist:** mang phong cáchtối giản, giản đơn, hài hòa, …

* + - * Type: tshirt, shirt, jacket, shorts, trouser
      * Model: shirt\_short (tay ngắn), tshirt-crewneck (thun cổ tròn), tshirt-polo (thun cổ trụ), jacket\_cardigan (áo len)
      * Color: black, white, navy, maroon (đỏ đô), beige
      * Material: coffee (sợi cà phê) =>’chỉ cho áo sơ mi’, cotton, kate (bông + nhân tạo), kaki
      * Form: slimfit, straight (rộng rãi)
  1. **Chuẩn bị Dữ liệu (DATA preparation)**

**Bước 1:** **Thu thập dữ liệu (Collecting Dataset)**

Trước tiên chúng ta cần phải hiểu và nắm được bài toán cần giải quyết là gì? Business value của nó để có thể tìm kiếm chính xác dữ liệu training cho bài toán! Với các bài toán classification có thể dựa vào tên các classes để tạo thành các keywords và sử dung các công cụ crawling data từ Internet cho việc tìm ảnh. Hoặc có thể tìm ảnh, videos từ các trang mạng xã hội, ảnh vệ tinh trên Google, free collected data từ camera công cộng hay xe hơi (Waymo, Tesla), thậm chí có thể mua dữ liệu từ bên thứ 3 (lưu ý tính chính xác của dữ liệu). Nhưng theo đề tài của nhóm, nhóm xây dựng bộ dataset dự trên bộ câu hỏi Quiz.

**Bước 2: Đánh nhãn dữ liệu (Labeling data)**

Đây là bước khá quan trọng vì sẽ đánh giá mô hình chúng ta làm việc tốt hay không! Đánh nhãn sai dữ liệu sẽ làm cho model dự đoán và đánh giá sai -> tốn nhiều thời gian và công sức bỏ ra cho quá trình training. Có hai vấn đề cần lưu ý là:

**Vấn đề 1 - Làm thế nào để đánh nhãn dữ liệu?**

Sau khi đã tìm được dataset cho bài toán, cần xác định xem bài toán thuộc dạng nào? Ví dụ classification, object detection, segmentation… Từ đó có thể tiến hành process data để đánh nhãn cho phù hợp! Trường hợp classification thì các labels chính là các keywords dùng trong quá trình tìm và crawling data từ Internet. Trường hợp instance segmentation cần có nhãn cho từng pixel của ảnh. Lúc này chúng ta cần phải sử dụng tools để tiến hành image annotation (tức là set label và metadata cho ảnh). Các tools phổ biến có thể kể đến là Comma Coloring, Annotorious, LabelMe… Các tool này sẽ hỗ trợ GUI cho việc đánh label từng segment của ảnh.

**Vấn đề 2 - Ai sẽ đánh nhãn dữ liệu?**

**Vấn đề chọn ai đánh giá dữ liệu là một trong những vấn đề quan trọng, thông thường người đánh nhãn dữ liệu phải là người có kinh nghiệm, có kiến thức đủ sâu vì vậy người đánh dãn dữ liệu không ai khác là một chuyên gia. Nhưng do quy mô đề tài chúng ta tạm có thể tự mình đánh giá trên tinh thần một cách khách quan và quan trọng tìm hiểu kĩ vấn đề một cách tường tận nhất.**

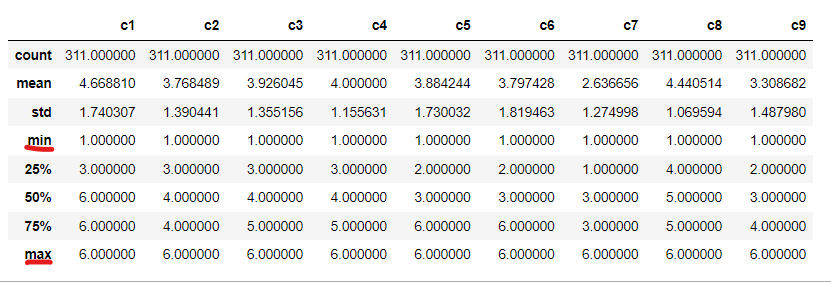
* 1. **Tiền xử lý Dữ liệu (DATA PREPROCESSING)**

**Bước 1: Tìm và xử lý giá trị bị thiếu - NAN (MISSING DATA)**

Do quá trình thu nhập của nhóm có sự giám sát và chọn lọc nên dataset này không có những giá trị NAN cần xử lý.

**Bước 2: Mã hóa giá trị danh mục (Encode Categorical Data)**

Trong bộ dataset nhóm xây dựng không có các đặc trưng ở dạng danh mục (Categorical feature), chính vì vậy bước này nhóm sẽ không thực hiện.

**Bước 3: Đưa Dữ liệu dạng thô về dạng 0 – 1 (DATA SCALING)**

Từ bảng mô tả dữ liệu trên, ta thấy giá trị của từng column trong dataset dao động từ 1 đến 6 (hay nói cách khác là min = 1, max = 6). Để đưa dữ liệu dạng thô về dạng 0 – 1, ta chỉ sử dụng phương pháp Min – Max Scaling được hỗ trợ trong thư viện sklearn. Mục tiêu của phương pháp là đưa các giá trị về gần hơn giá trị mean của features. Phương pháp này đưa các giá trị về 1 khoảng đặc biệt, thường là [0, 1] [0,1] hoặc [-1, 1] [−1,1]. Một trong những hạn chế của phương pháp này là khi áp dụng với một khoảng giá trị nhỏ, ta sẽ thu được độ lệch chuẩn nhỏ hơn, điều này làm giảm weight của các outliers trong dữ liệu.

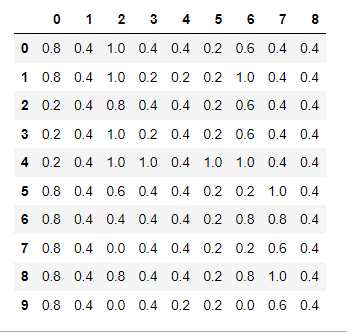
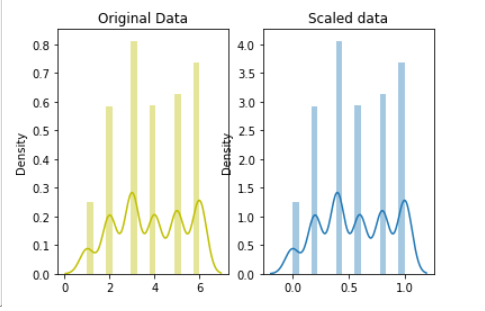


Với Xold: là giá trị ban đầu

Xmax: giá trị lớn nhất

Xmin: giá trị nhỏ nhất

Sau khi scaling:



Qua ví dụ, khẳng định scaling chỉ làm thay đổi khoảng giá trị chứ không làm thay đổi hình dạng phân phối dữ liệu là chính xác.

* 1. **Trích chọn thuộc tính (FEATURE EXTRACTION):**

\* *Tại sao không trích chọn thuộc tính?*

Vì hiện tại các giá trị trong dataset không có độ chênh lệch các giá trị với nhau quá lớn.

* 1. **Đánh giá dữ liệu qua công cụ WEKA**

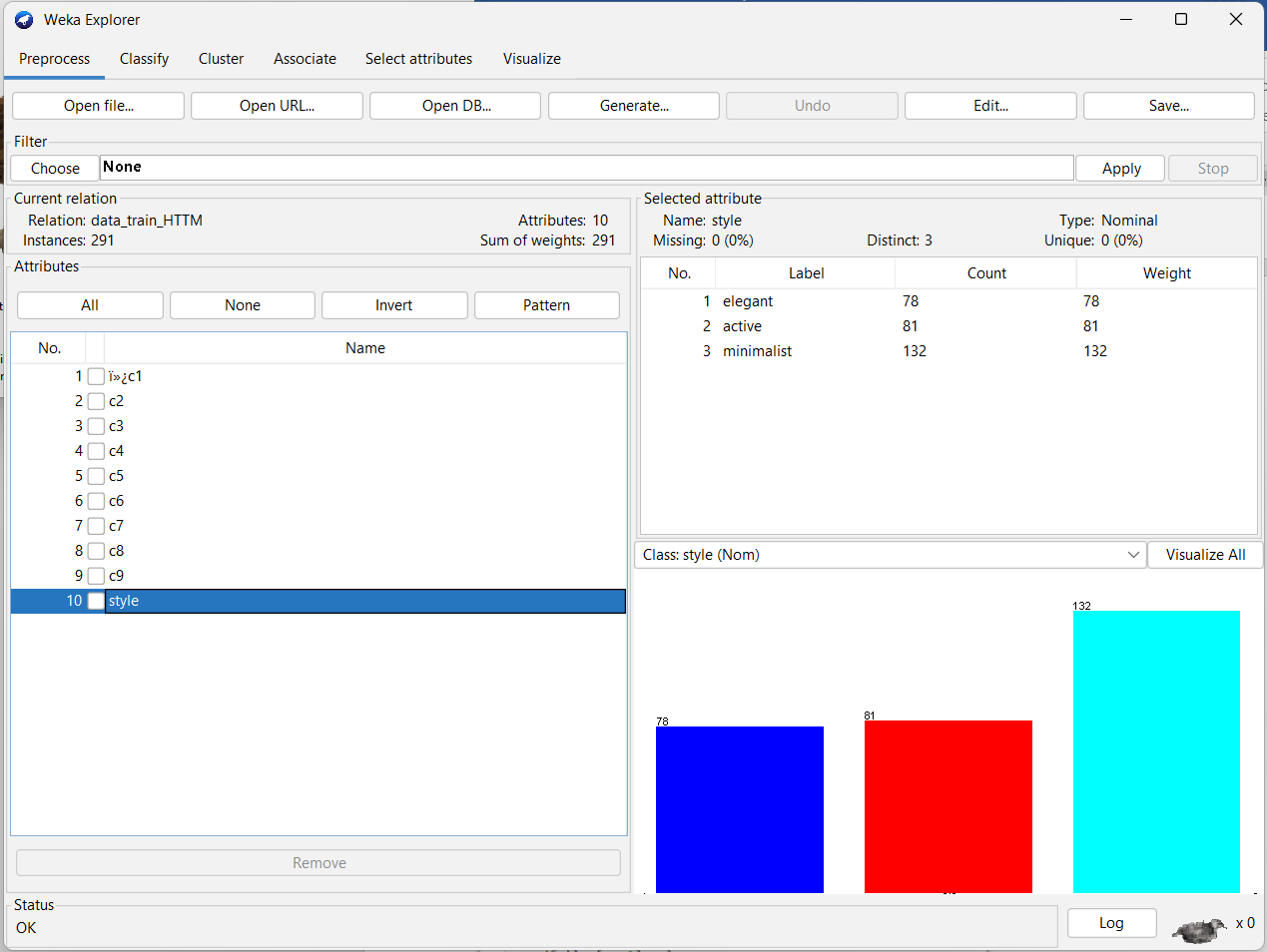
Trong bức ảnh ta thấy được, với nhãn Elegant có số lượng là 78 hàng, nhãn Active có 81 hàng, nhãn Minimalist có 132 hàng.

Figure 9: phân tích Dataset với công cụ Weka

Sử dụng Tab Classify để tiến hành phân lớp bộ dữ liệu chuẩn với thuật toán J48.

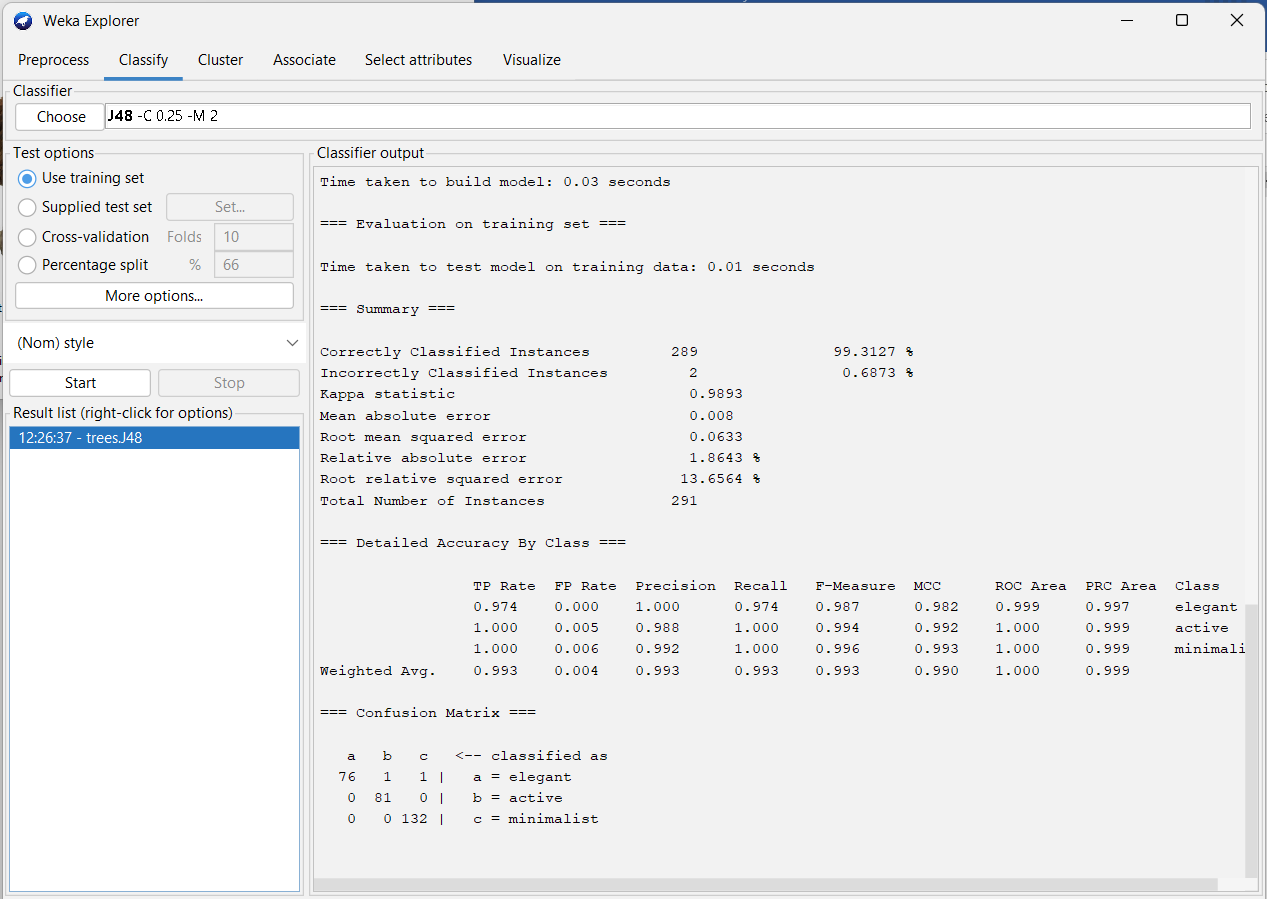
* Use Training set: Cho ra kết quả cao ~ 99% đáng để tin cây

Figure 10: weka với use training set

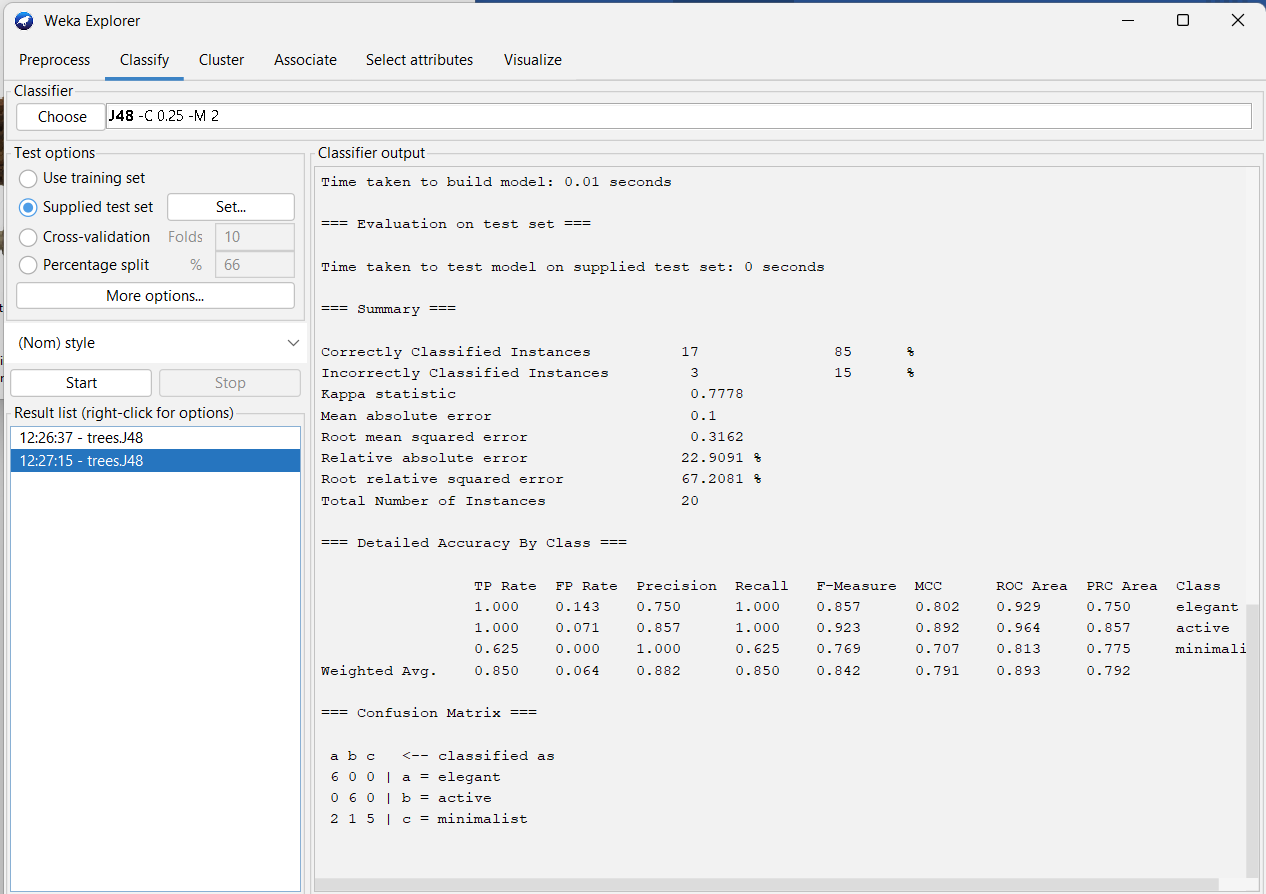
* Supplied test set: Cho ra kết quả 85% khá ổn đáng để tin cậy.

Figure 11: weka với use Supplied test set và kết quả.

* Cross – Validation: Cho ra kết quả lớn hơn 92% khá cao, đáng để tin cậy.

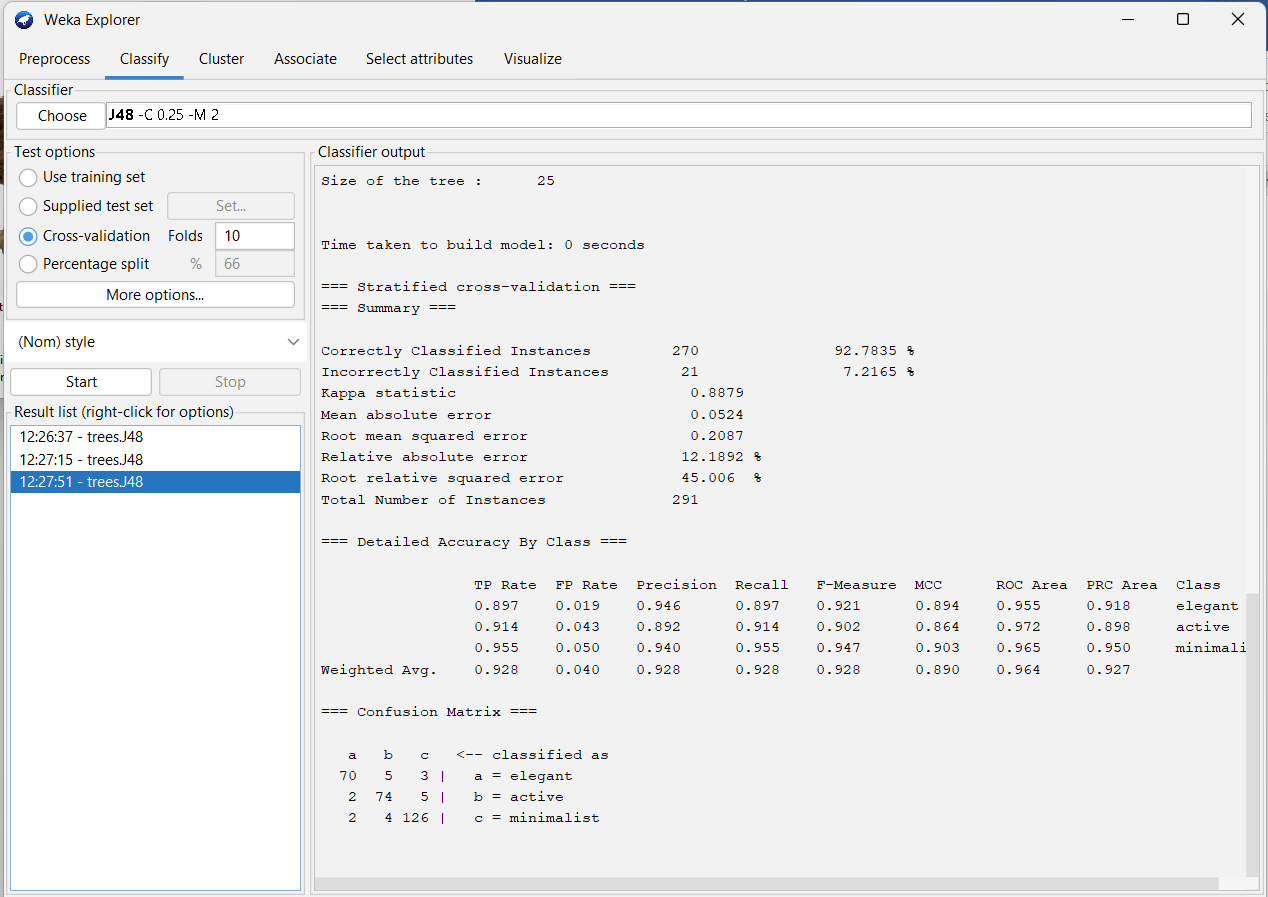


Figure 12: Weka với Cross - Validation và kết quả.

1. **XÂY DỰNG CẤU HÌNH:**
   1. **Đọc dữ liệu từ file csv:**

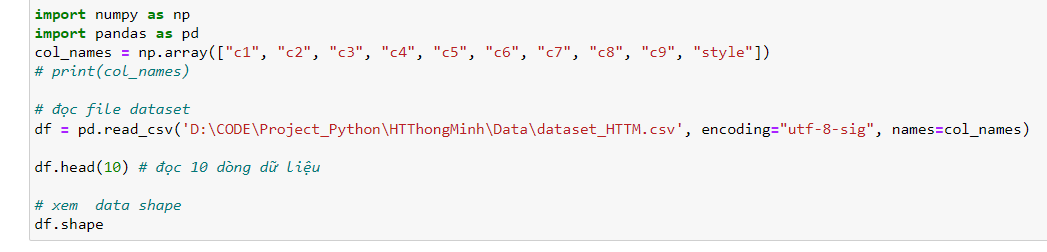
****

Figure 13: Code python đọc file csv với thư viện Pandas.

* 1. **Cấu hình SVM và xuất file đã được train:**

Xác định nhãn và thuộc tính của dữ liệu đầu vào, đưa thuộc tính và nhãn lần lượt vào x\_train, y\_train.

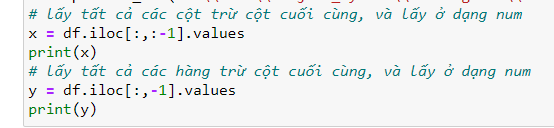
****

Figure 14: Code python xác định x và y

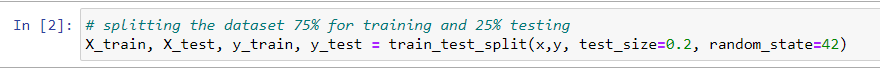
****

Figure 15: Cắt dataset thành 2 bộ train và test

Cấu hình mô hình SVM với các thông số:

Kernel: kernel = ’poly’.

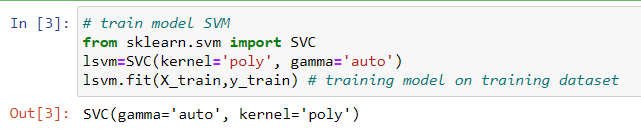
****Gamma: gamma = ‘auto’.

Figure 16: Cấu hình mô hình SVM với thư viện sklearn.

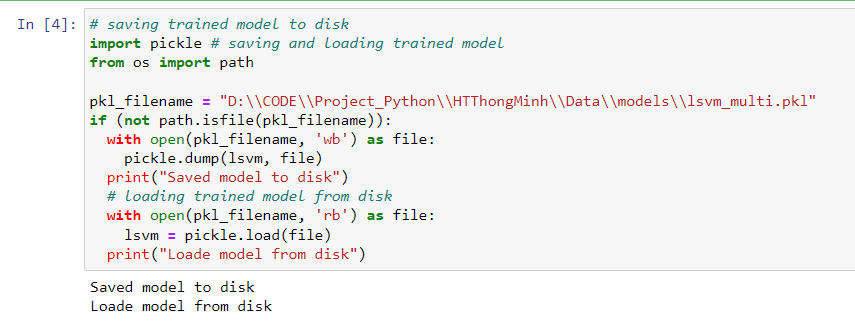
****

Figure 17: Code python lưu lại model đã train với thư viện pickle.

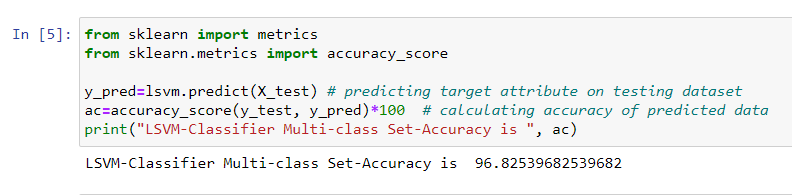
****

Figure 18: Code python dự đoán accuracy\_scorce với thư viện sklearn.

* 1. **Quy đổi chuyền dữ liệu từ Java sang Python:**

ProcessBuilder

Bufferedreader

InputStreamReader

****

Figure 19: Code java dùng để quy đổi dữ liệu từ từ Java sang Python

* 1. **Đưa dữ liệu vào huấn luyện:**

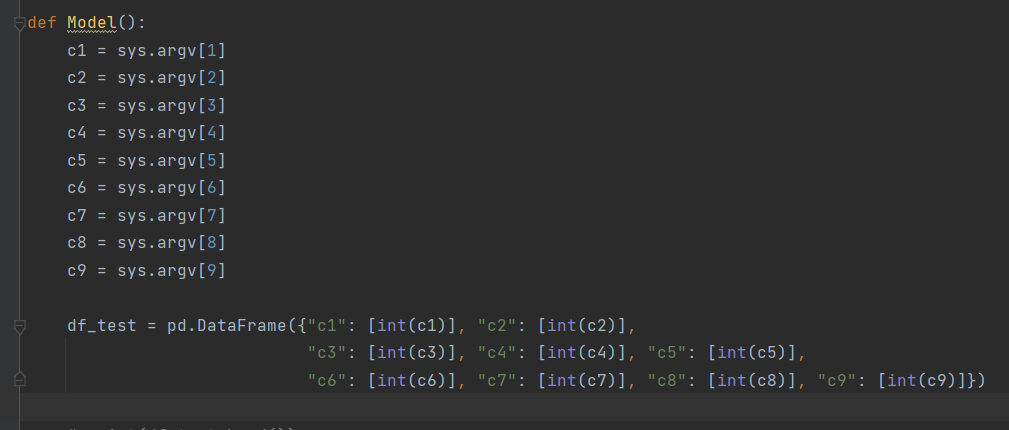


Figure 20: Code python đưa dữa liệu đầu vào.

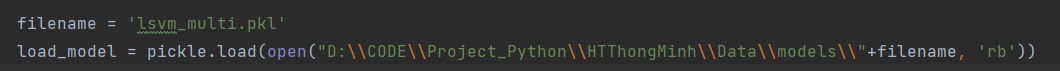


Figure 21: Code đọc file để load model đã train.

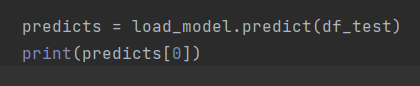


Figure 22: Code Xuất ra kết quả đã dự đoán

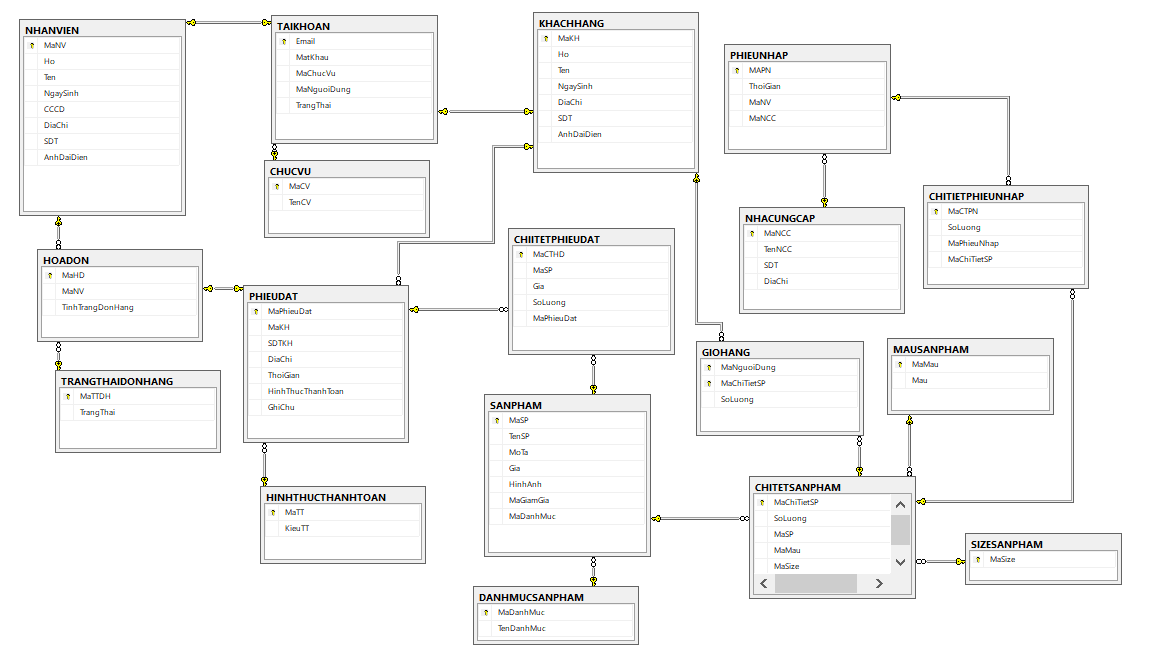
1. **PHÂN TÍCH THIẾT KẾ HỆ THỐNG**
   1. **Thiết kế cơ sở dữ liệu:**

Figure 23: Tổng quan về Diagram của toàn bộ hệ thống

Danh sách thuộc tính từng thành phần

* TAIKHOAN

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | Email | Primary key | Varchar (50) | Not null |  |
| 2 | Matkhau |  | Varchar (50) | Not null |  |
| 3 | Mavaitro | Foreign key | Varchar (10) | Not null |  |
| 4 | Trangthai | Foreign key | Varchar (10) | Not null |  |
| 5 | Manguoidung | Foreign key | Varchar (10) | Not null |  |

* KHACHHANG

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | Makh | Primary key | Varchar (10) | Not null |  |
| 2 | Ho |  | Nvarchar (50) | Not null |  |
| 3 | Ten |  | Nvarchar (30) | Not null |  |
| 4 | Ngaysinh |  | Date | Allow null |  |
| 5 | Diachi |  | Ntext | Not null |  |
| 6 | Sdt |  | Varchar (11) | Not null |  |
| 7 | AnhDaiDien |  | Image | Allow null |  |

* NHANVIEN

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | MaNV | Primary key | Varchar (10) | Not null |  |
| 2 | Ho |  | Nvarchar (50) | Not null |  |
| 3 | Ten |  | Nvarchar (30) | Not null |  |
| 4 | Ngaysinh |  | Date | Not null |  |
| 5 | CCCD |  | Varchar (15) | Not null |  |
| 6 | Diachi |  | Ntext | Not null |  |
| 7 | SDT |  | Varchar (11) | Not null |  |
| 8 | AnhDaiDien |  | Image | Allow null |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | MaChucVu | Primary key | Varchar (10) | Not null |  |
| 2 | TenChucVu |  | Varchar (50) | Not null |  |

* CHUCVU
* SANPHAM

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | MaSP | Primary key | Varchar (10) | Not null |  |
| 2 | TenSP |  | Nvarchar (50) | Not null |  |
| 3 | MoTa |  | Ntext | Allow null |  |
| 4 | Gia |  | Nvarchar (30) | Not null |  |
| 5 | HinhAnh |  | Image | Not null |  |
| 6 | MaGiamGia |  | Varchar (20) | Allow null |  |
| 7 | MaDanhMuc |  | Varchar (10) | Not null |  |

* CHITIETSANPHAM

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | MaChiTietSP | Primary key | Varchar (10) | Not null |  |
| 2 | SoLuong |  | Int | Not null |  |
| 3 | MaSP |  | Varchar (10) | Not null |  |
| 4 | MaMau |  | Varchar (10) | Not null |  |
| 5 | MaSize |  | Varchar (10) | Not null |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | MaMau | Primary key | Varchar (10) | Not null |  |
| 2 | Mau |  | Nvarchar (50) | Not null |  |

* MAUSANPHAM
* SIZESANPHAM

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | MaSIZE | Primary key | Varchar (10) | Not null |  |

* DANHMUCSANPHAM

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | MaDanhMuc | Primary key | Varchar (10) | Not null |  |
| 2 | TenDanhMuc |  | Nvarchar (50) | Not null |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | MaNCC | Primary key | Varchar (10) | Not null |  |
| 2 | TenNCC |  | Nvarchar (50) | Not null |  |
| 3 | SDT |  | Varchar (11) | Not null |  |
| 4 | DiaChi |  | Ntext | Not null |  |

* NHACUNGCAP

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | MaHD | Primary key | Varchar (10) | Not null |  |
| 2 | MaNV |  | Varchar (10) | Not null |  |
| 3 | TinhTrangHoaDon |  | Varchar (10) | Not null |  |

* HOADON
* PHIEUDAT

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | MaPhieuDat | Primary key | Varchar (10) | Not null |  |
| 2 | MaKH |  | Varchar (10) | Not null |  |
| 3 | ThoiGian |  | Date | Not null |  |
| 4 | SDTKH |  | Varchar (50) | Not null |  |
| 5 | DiaChi |  | Ntext | Not null |  |
| 6 | GhiChu |  | Ntext | Allow null |  |
| 7 | HinhThucThanhToan |  | Varchar (10) | Not null |  |

* CHITIETPHIEUDAT

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | MaCTHD | Primary key | Varchar (10) | Not null |  |
| 2 | MaSP |  | Varchar (10) | Not null |  |
| 3 | Gia |  | Nvarchar (30) | Not null |  |
| 4 | SoLuong |  | Int | Not null |  |
| 5 | MaPhieuDat |  | Varchar (10) | Not null |  |

* PHIEUNHAP

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | MaPN | Primary key | Varchar (10) | Not null |  |
| 2 | ThoiGian |  | Date | Not null |  |
| 3 | MaNV |  | Varchar (10) | Not null |  |
| 4 | MaNCC |  | Varchar (10) | Not null |  |

* CHITIETPHIEUNHAP

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | MaCTPN | Primary key | Varchar (10) | Not null |  |
| 2 | SoLuong |  | Int | Not null |  |
| 3 | MaPhieuNhap |  | Varchar (10) | Not null |  |
| 4 | MaChiTietSP |  | Varchar (10) | Not null |  |

* GIOHANG

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | MaNguoiDung | Primary key | Varchar (10) | Not null |  |
| 2 | MaChiTietSP |  | Varchar (10) | Not null |  |
| 3 | SoLuong |  | int | Not null |  |

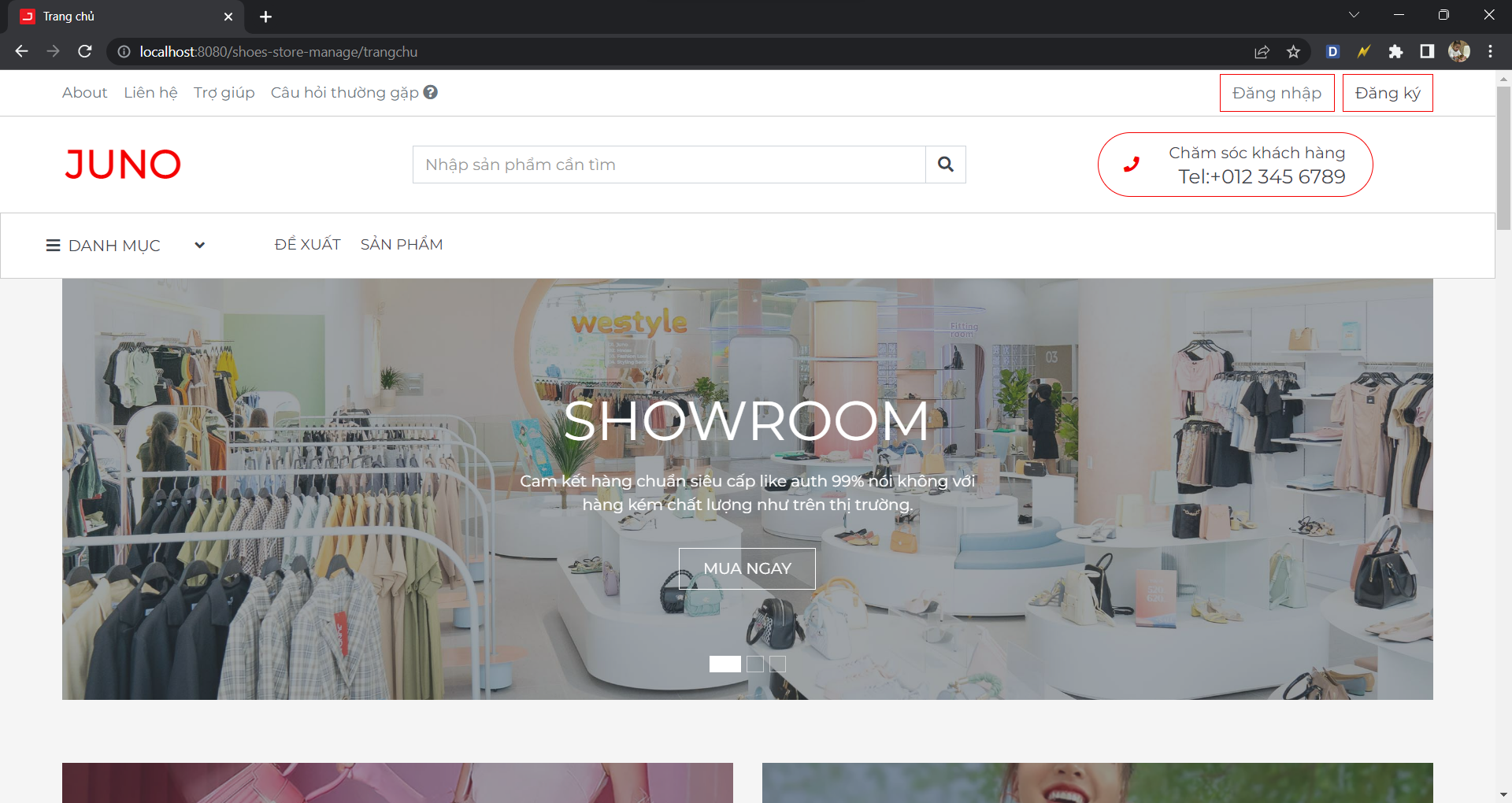
* TINHTRANGDONHANG

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | MaTT | Primary key | Varchar (10) | Not null |  |
| 2 | TenTT |  | Nvarchar (50) | Not null |  |

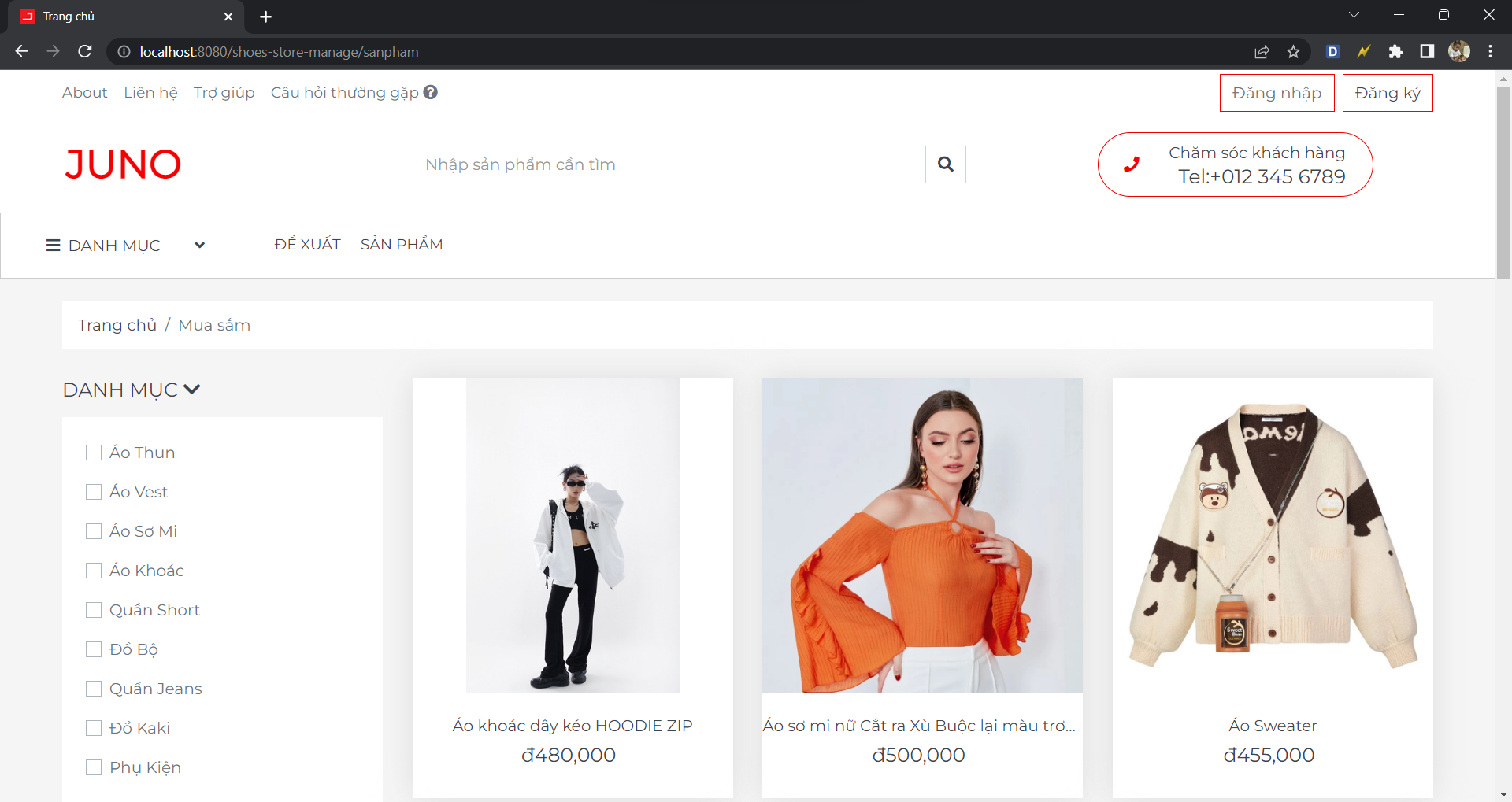
* HINHTHUCTHANHTOAN

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **TÊN** | **LOẠI** | **KIỂU** | **MIỀN GIÁ TRỊ** | **Ý NGHĨA** |
| 1 | MaHTTT | Primary key | Varchar (10) | Not null |  |
| 2 | TenHTTT |  | Nvarchar (50) | Not null |  |

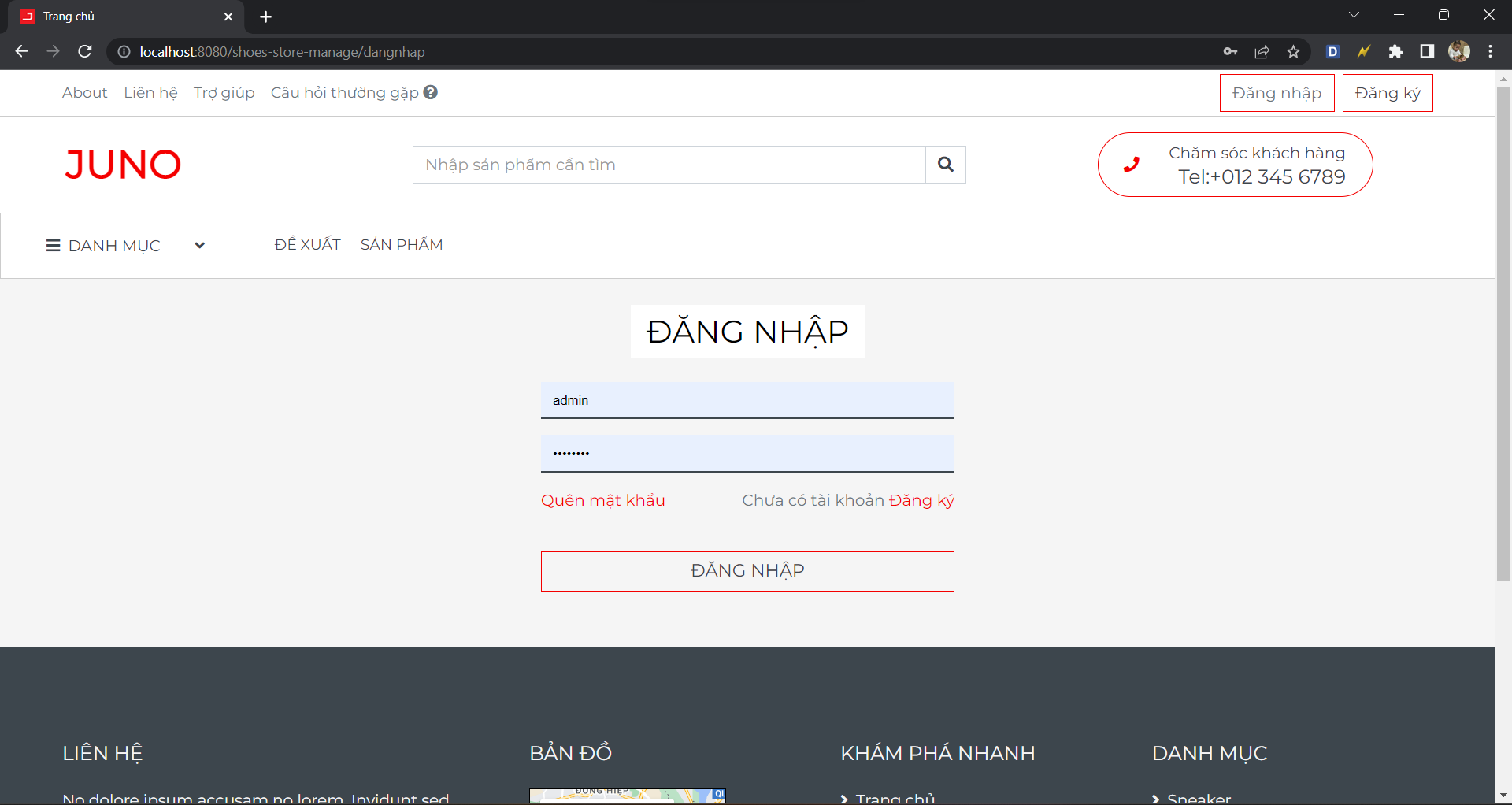
* 1. **Thiết kế giao diện:**



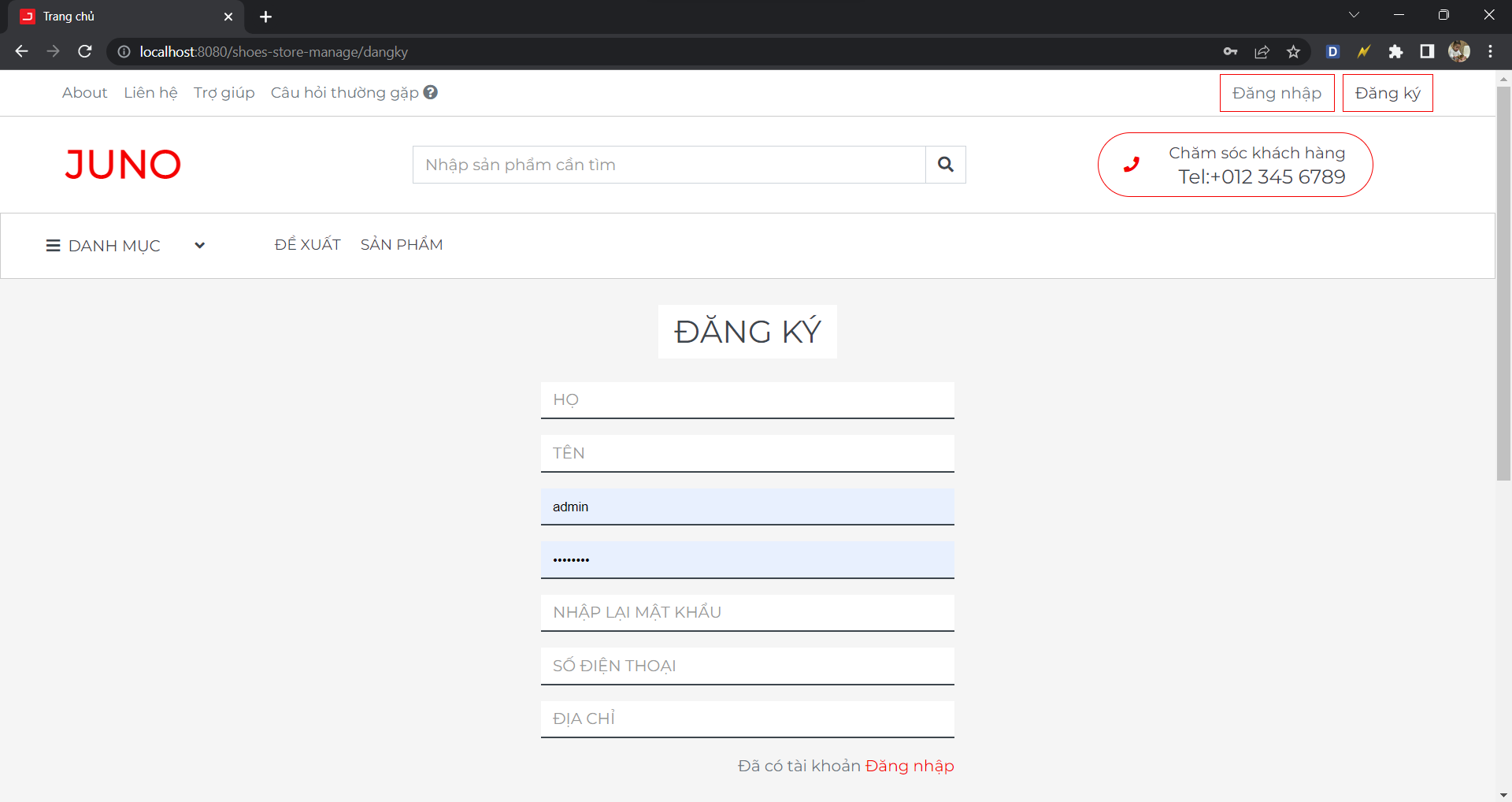
*Hình 2.1:Giao diện trang chủ*



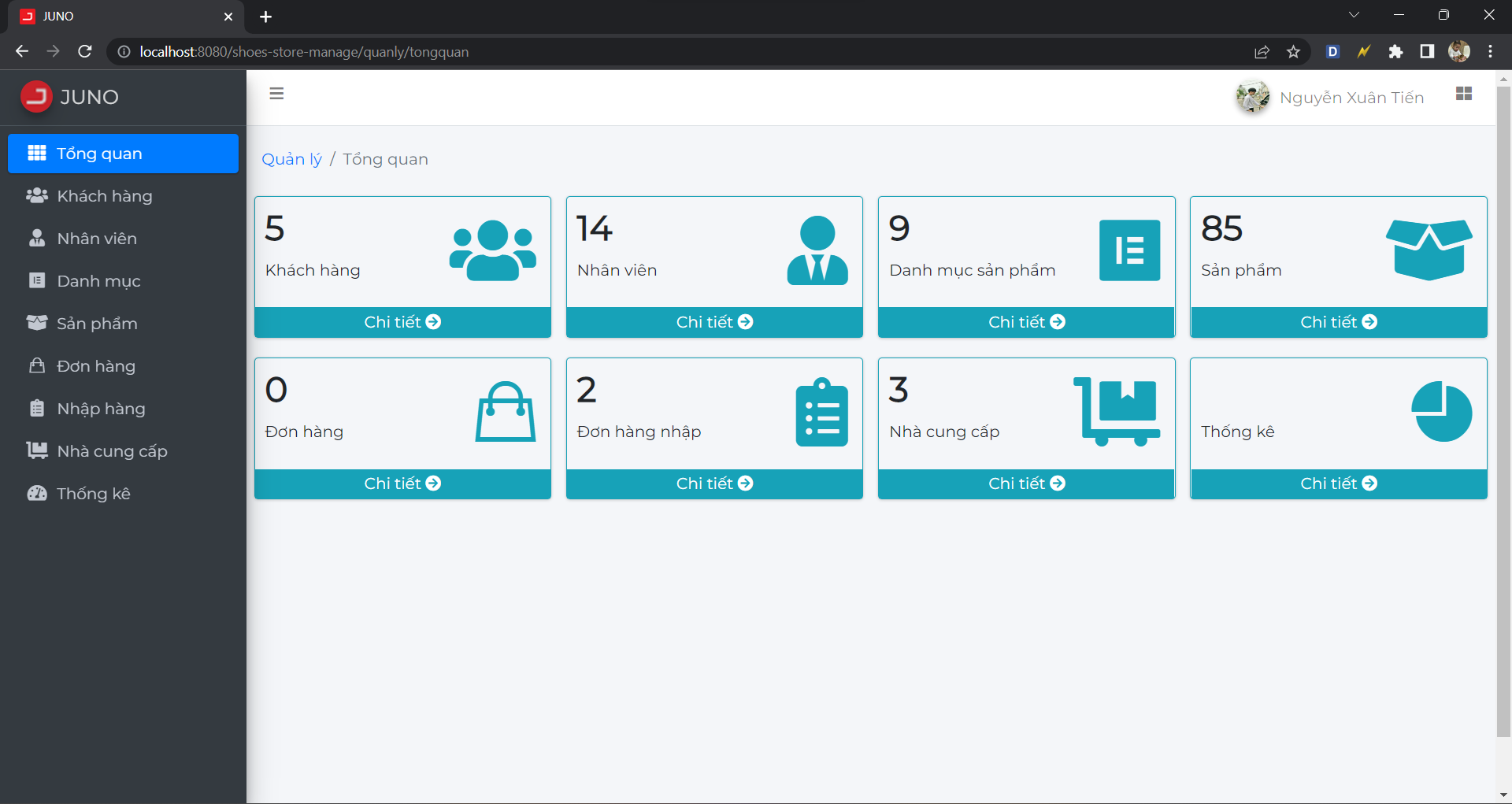
*Hình 2.2:Giao diện trang tất cả sản phẩm*



*Hình 2.3:Giao diện đăng nhập*



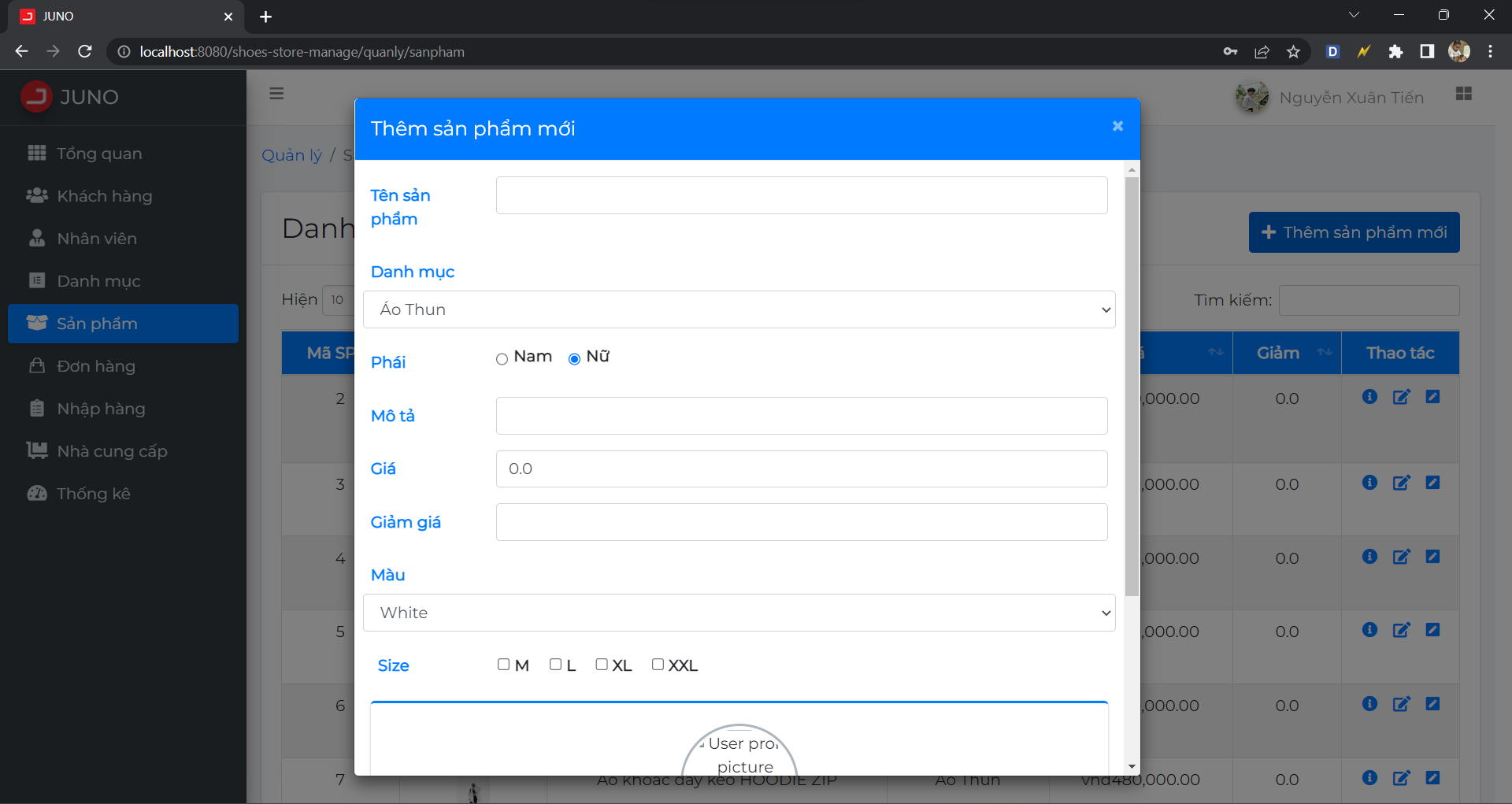
*Hình 2.4:Giao diện đăng ký*



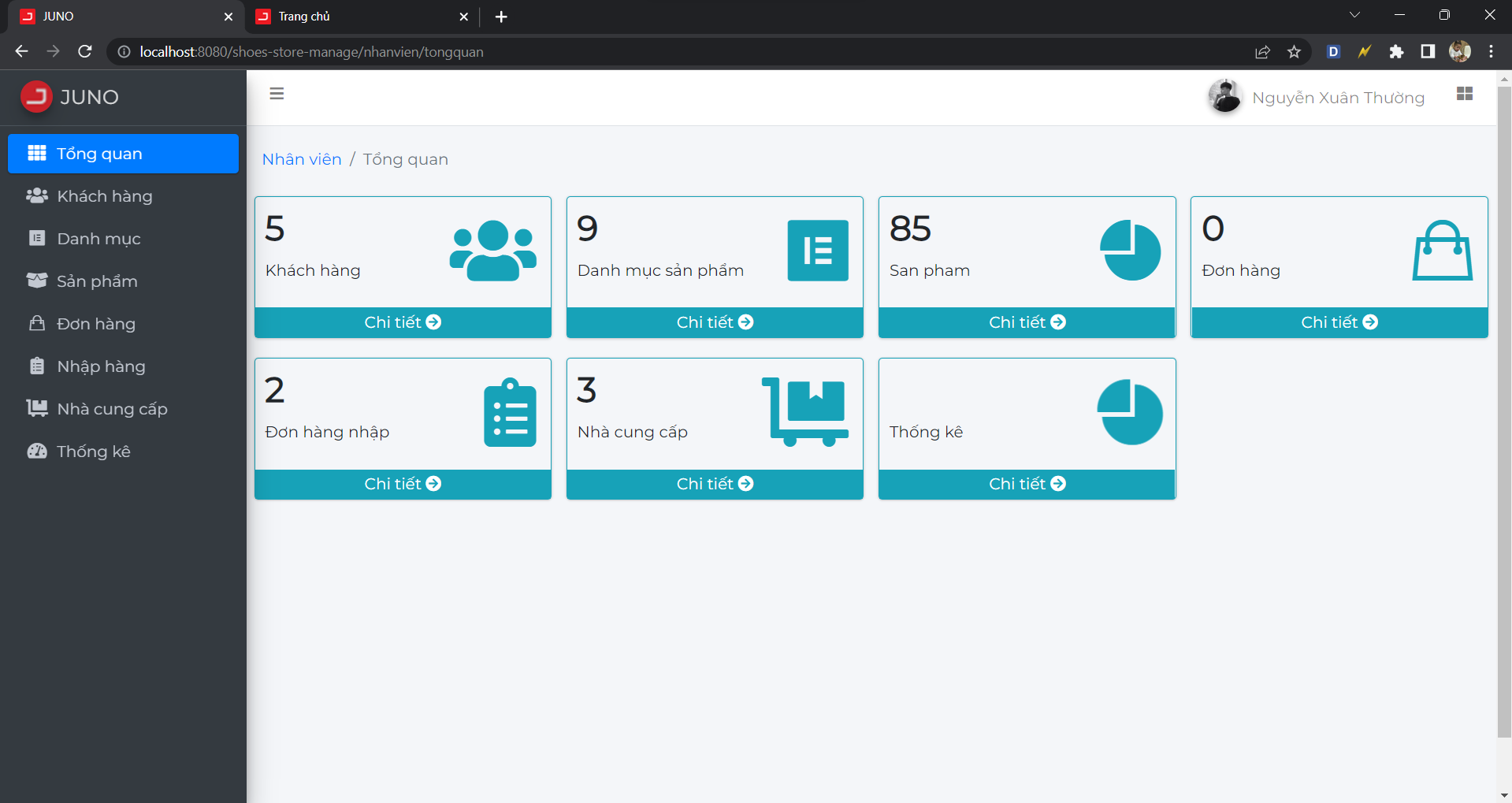
*Hình 2.5:Giao diện quản lý*



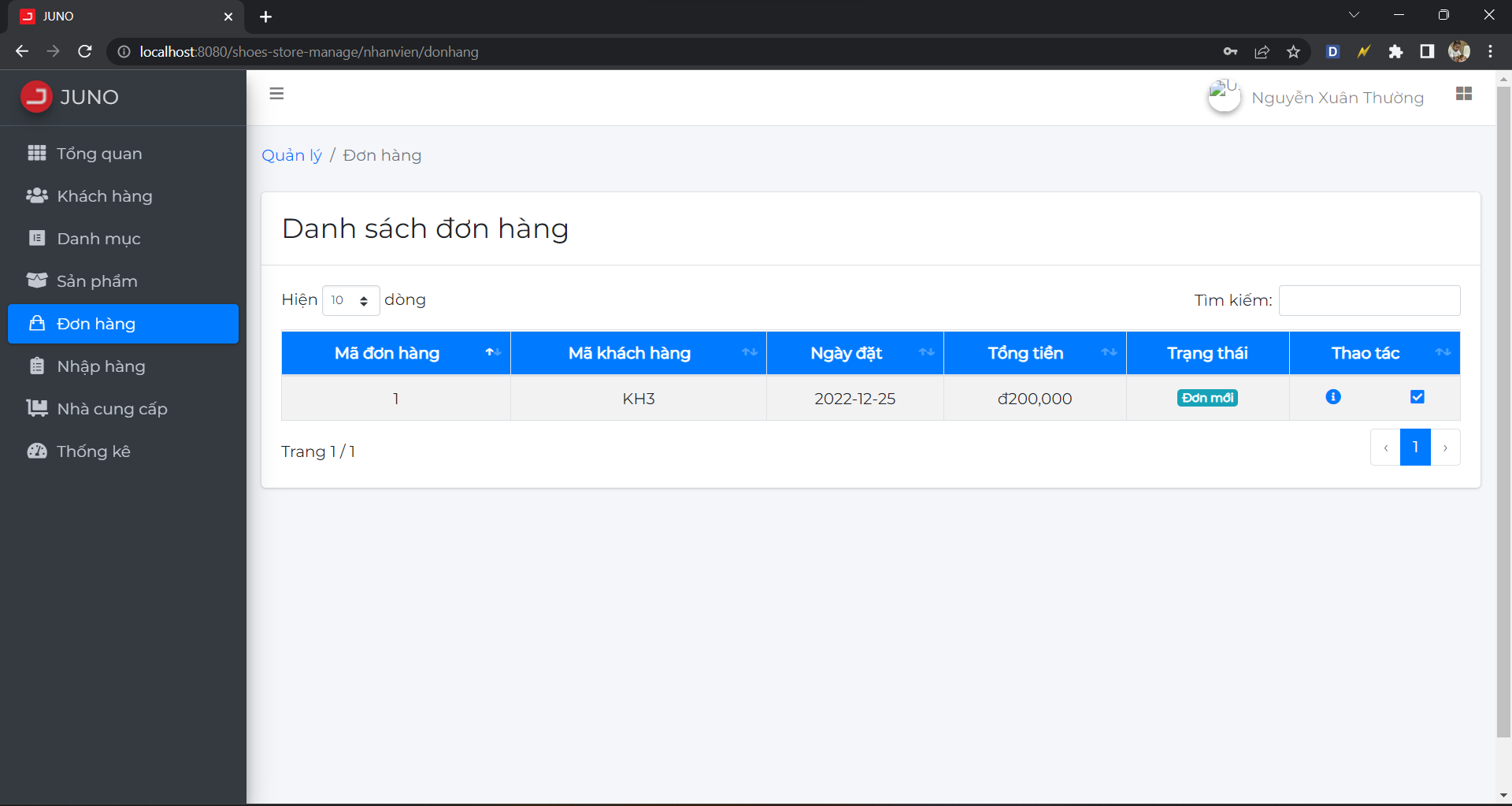
*Hình 2.6:Giao diện quản lý sản phẩm*



*Hình 2.7:Giao diện thêm sản phẩm*

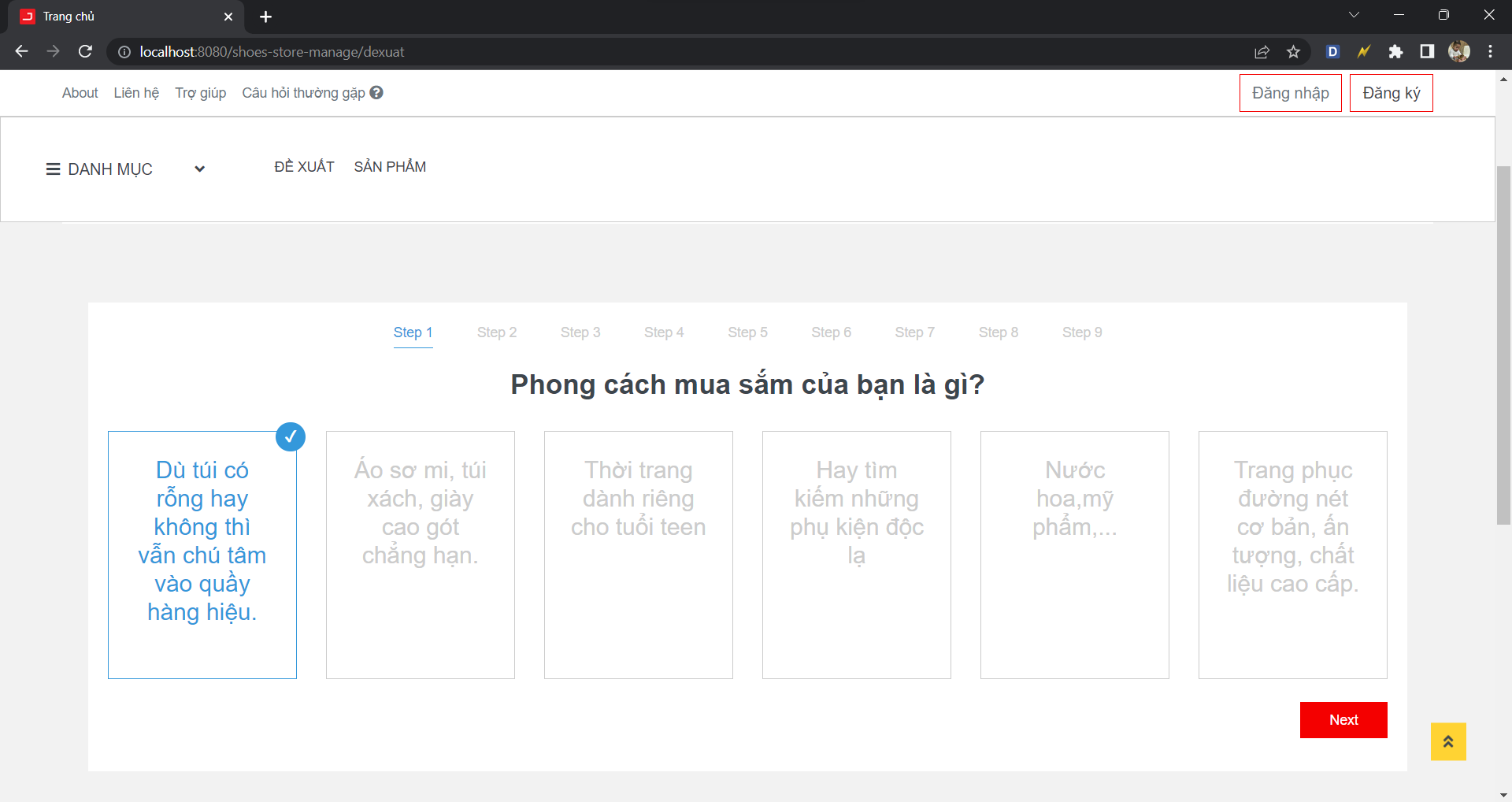


*Hình 2.8: Giao diện nhân viên*

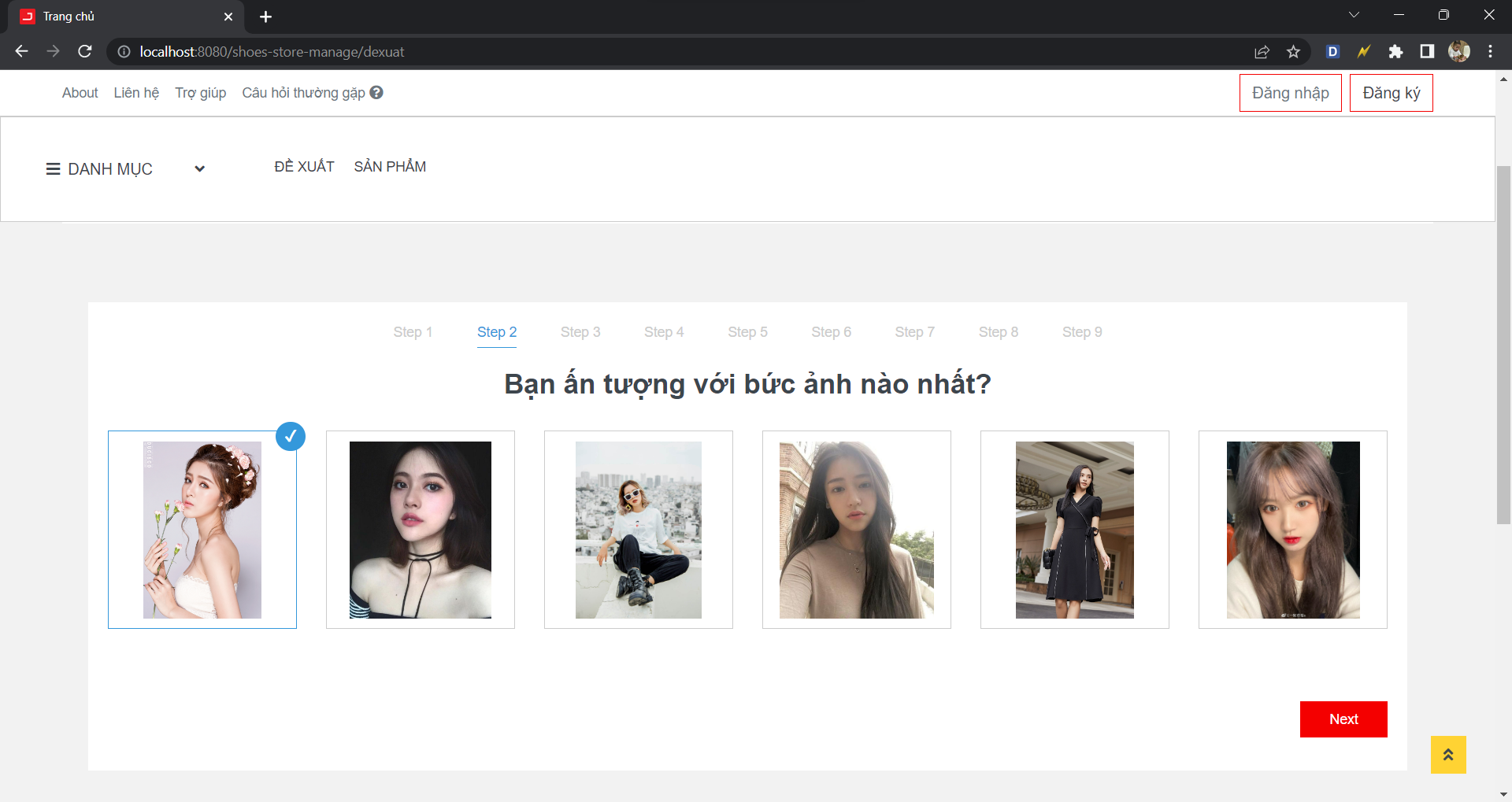


*Hình 2.9: Quản lý đơn hàng*

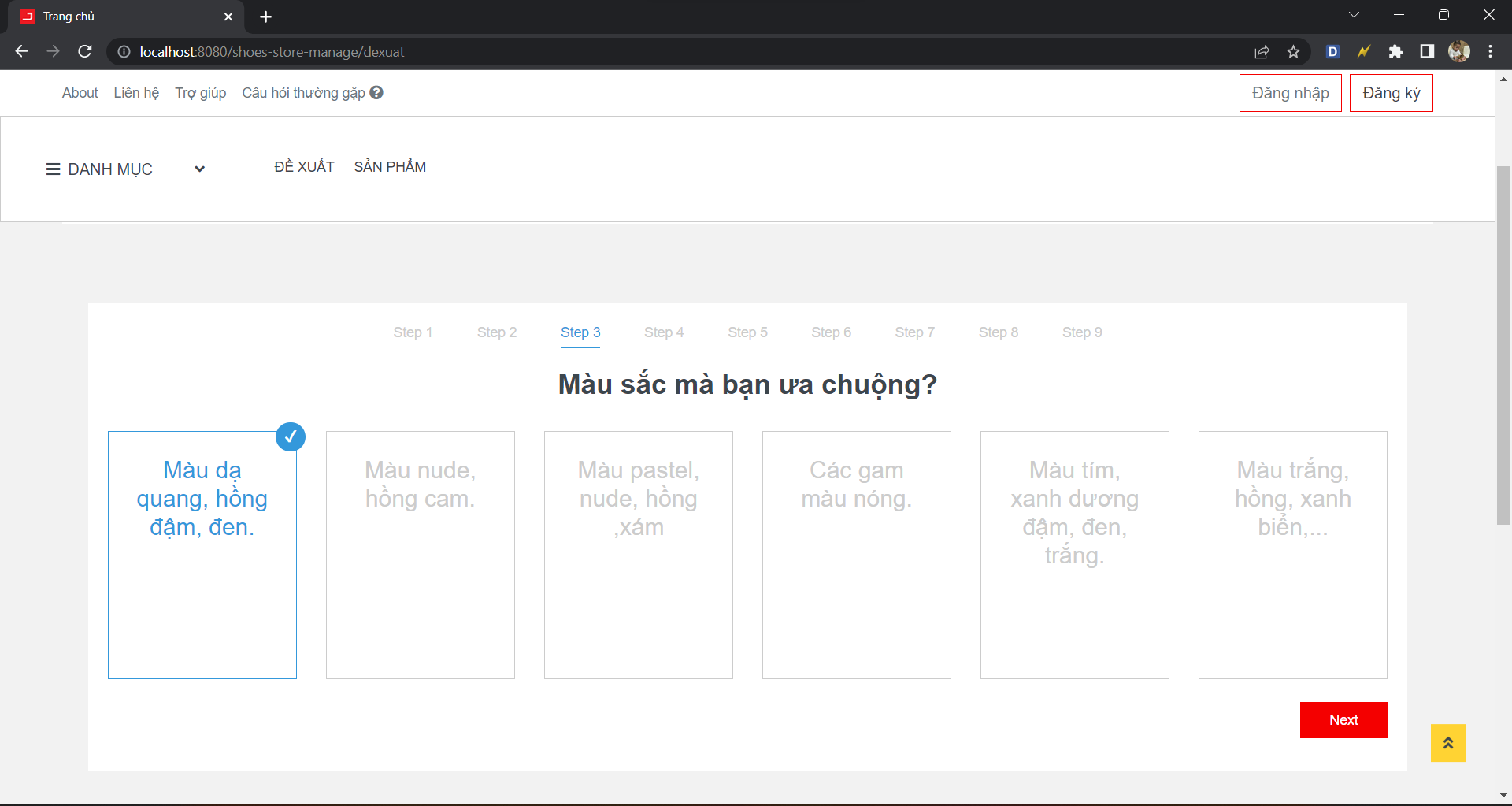
* 1. **Thiết kế giao diện thông minh:**



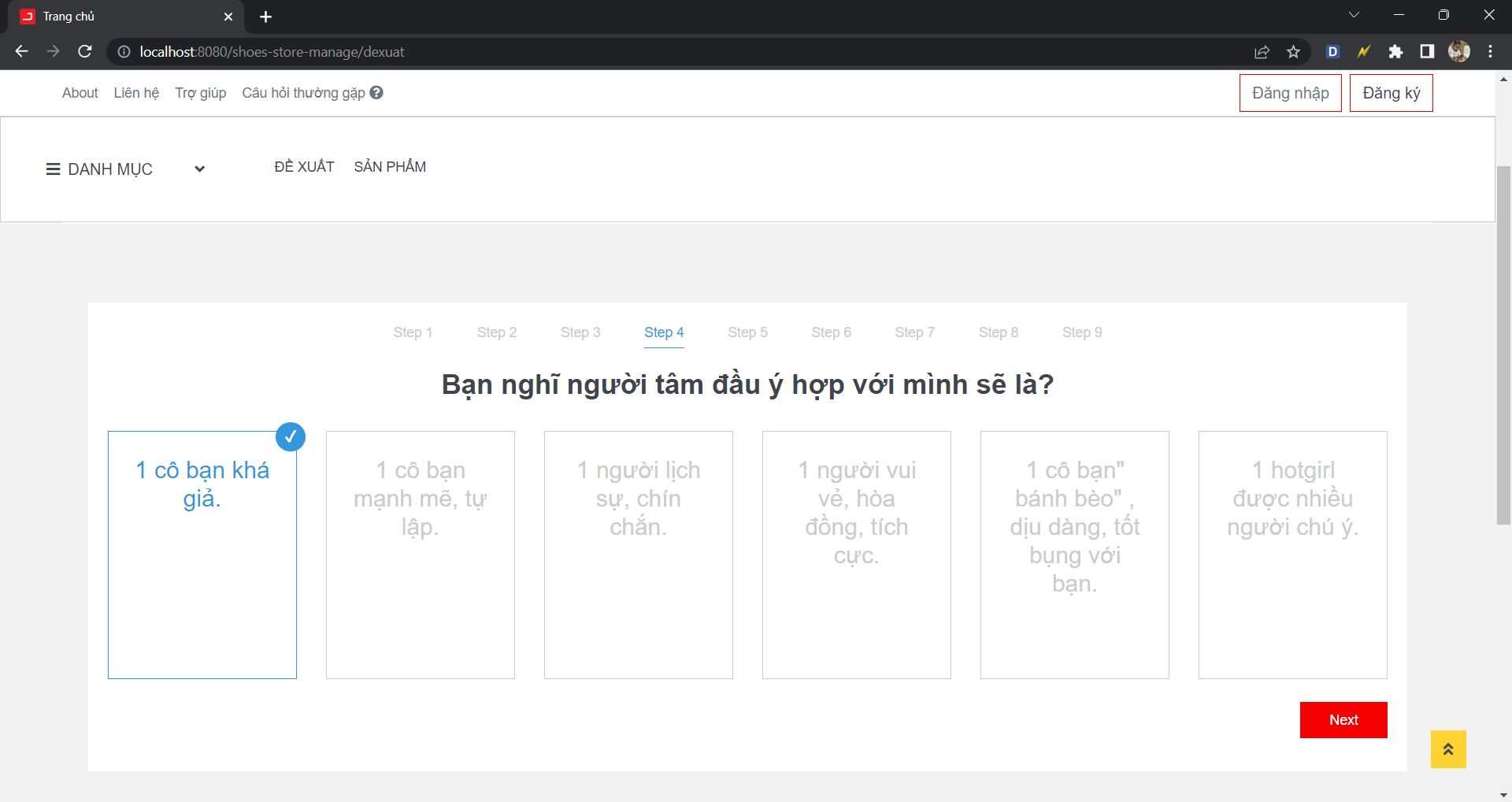
*Hình 3.1:Giao diện trang câu hỏi đề xuất 1*

**

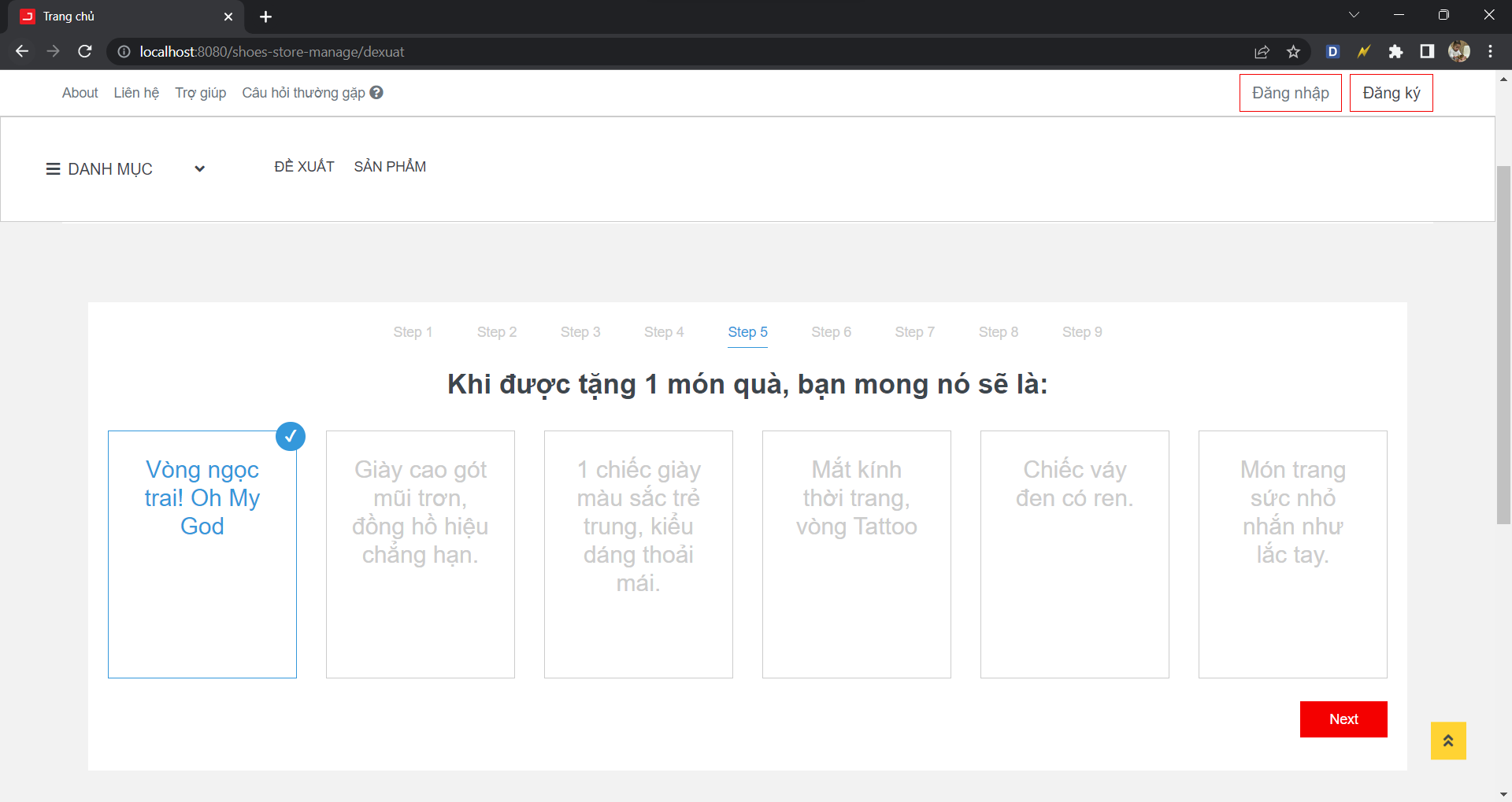
*Hình 3.2:Giao diện trang câu hỏi đề xuất 2*

**

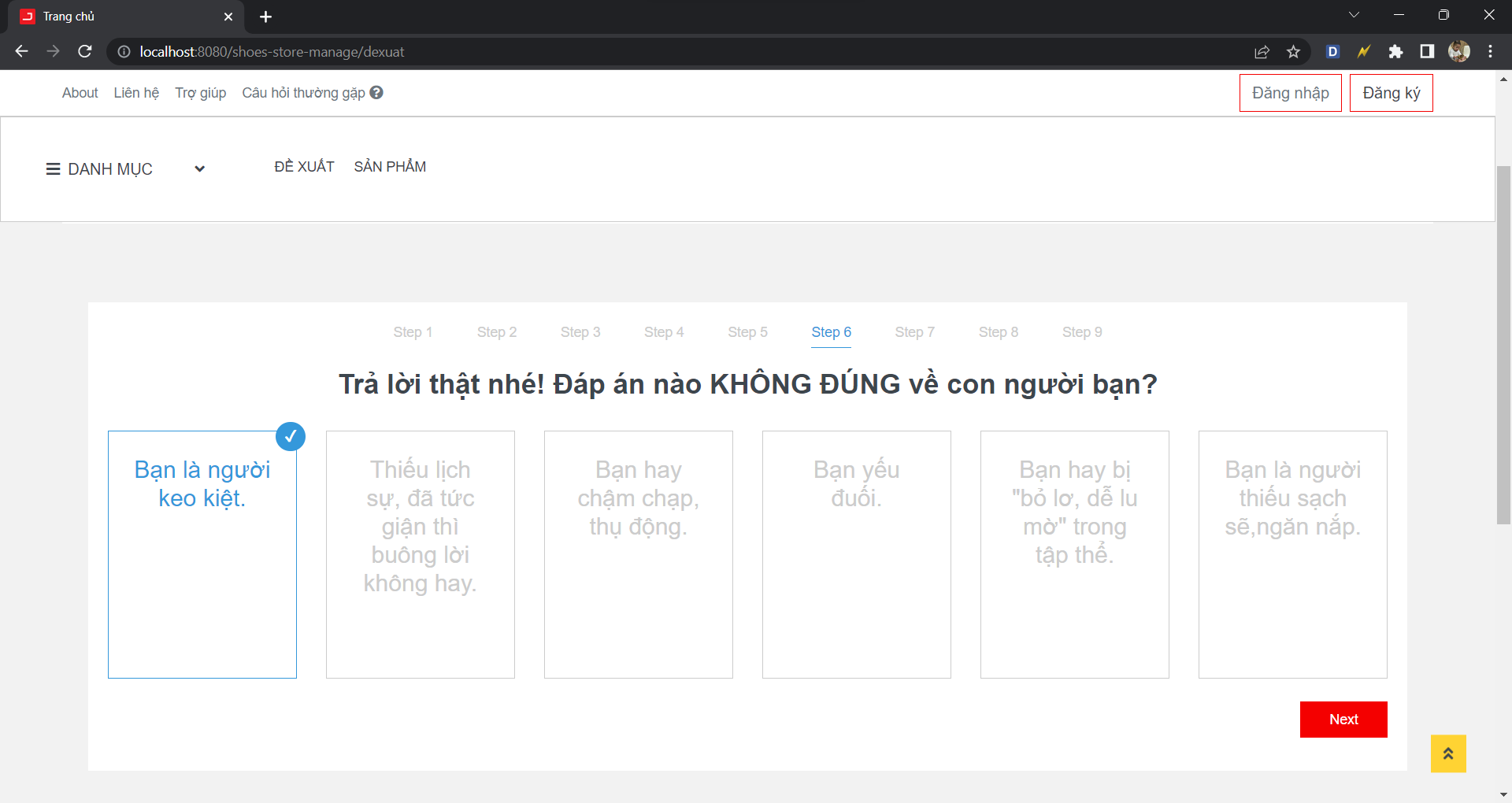
*Hình 3.3:Giao diện trang câu hỏi đề xuất 3*

**

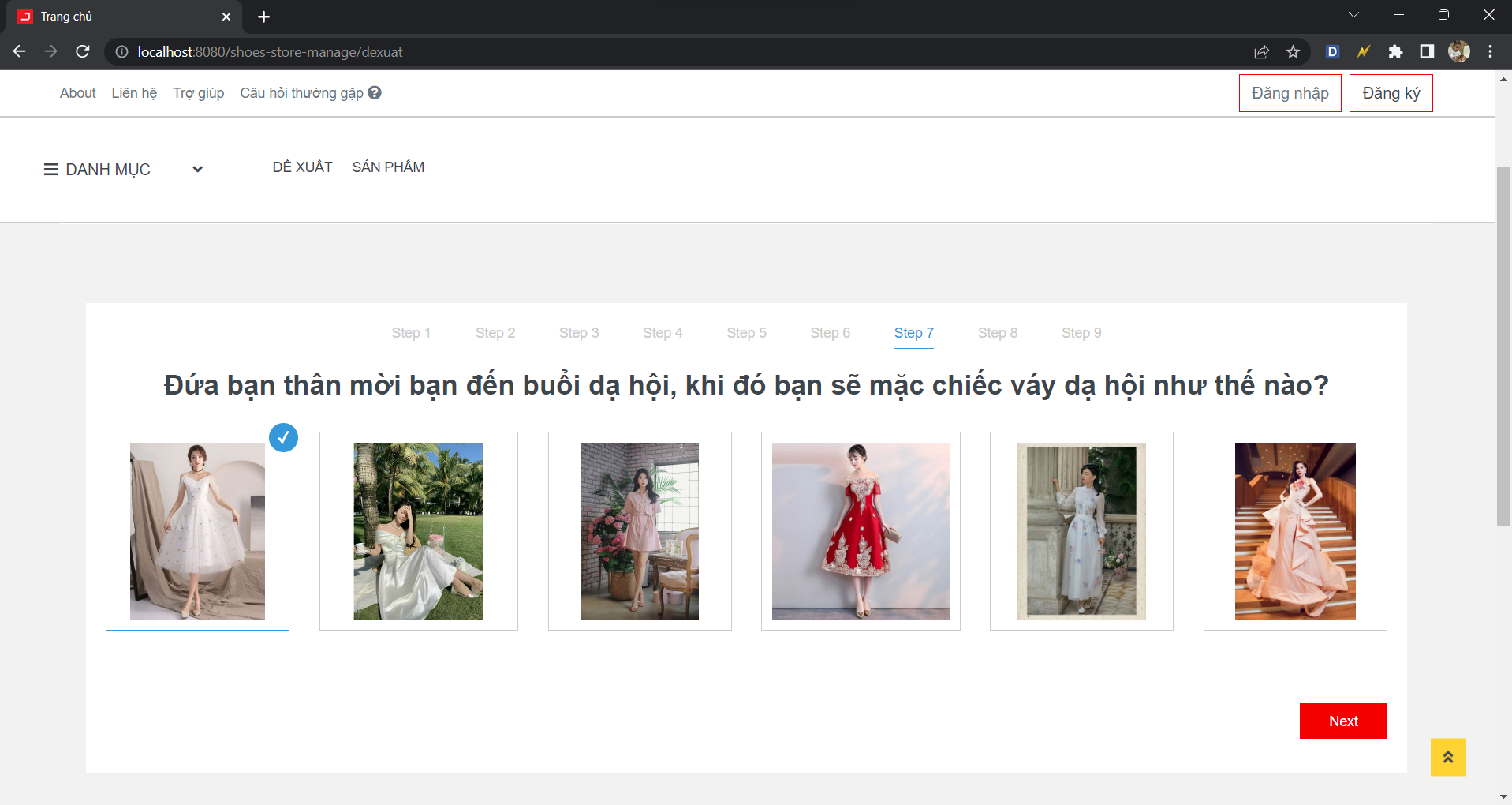
*Hình 3.4:Giao diện trang câu hỏi đề xuất 4*

**

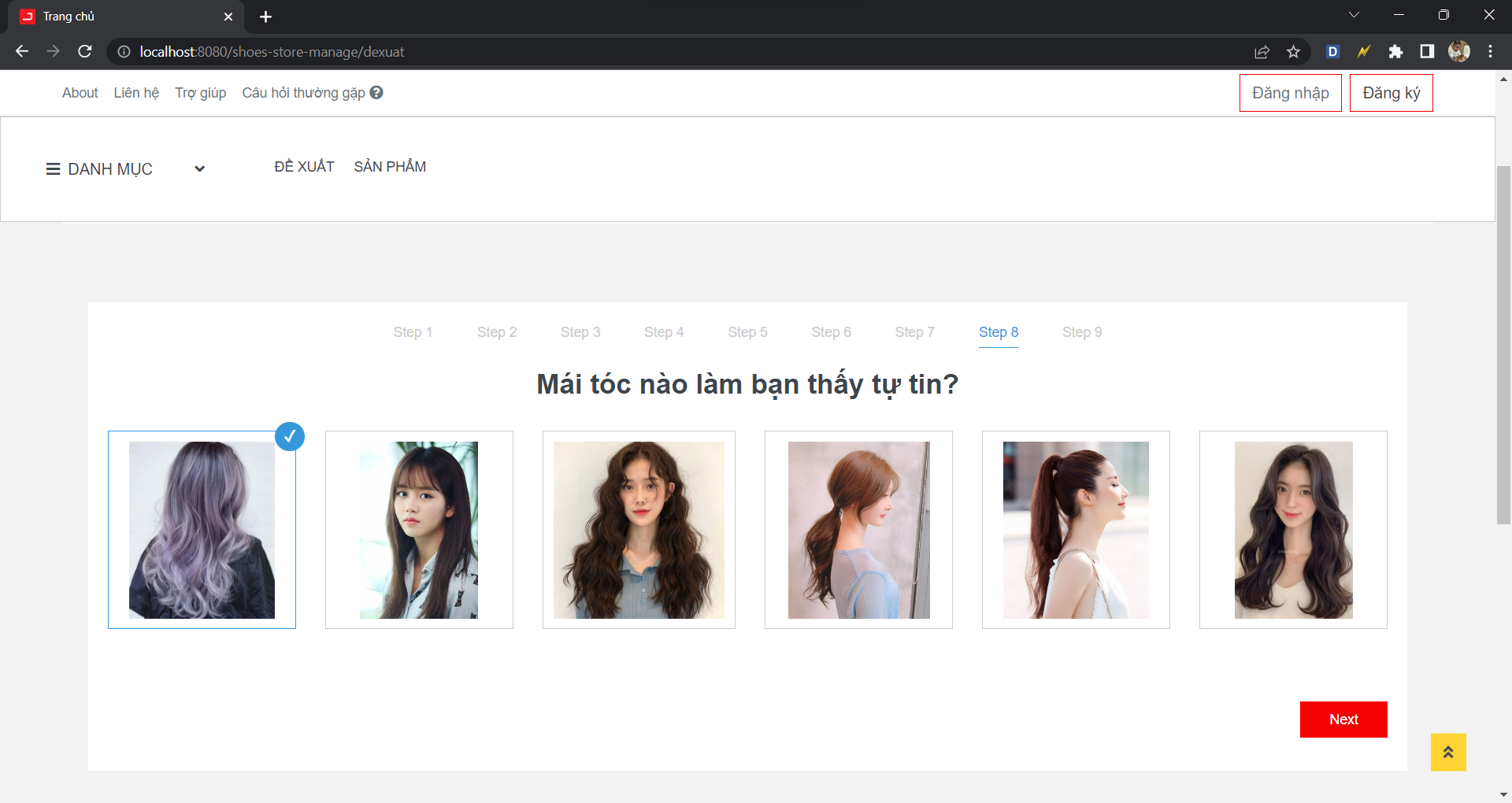
*Hình 3.5:Giao diện trang câu hỏi đề xuất 5*

**

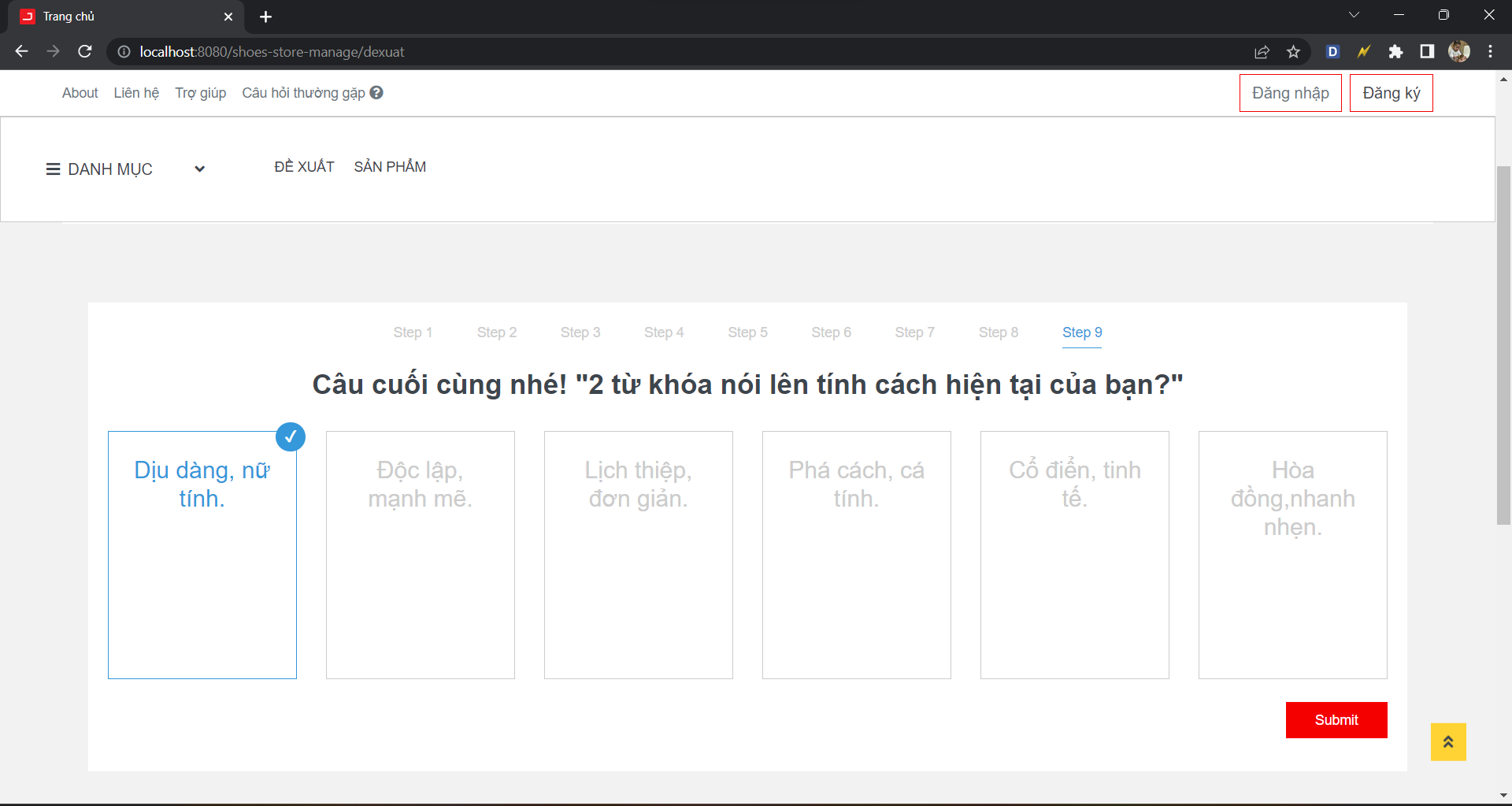
*Hình 3.6:Giao diện trang câu hỏi đề xuất 6*

**

*Hình 3.7:Giao diện trang câu hỏi đề xuất 7*

**

*Hình 3.8:Giao diện trang câu hỏi đề xuất 8*

**

*Hình 3.9:Giao diện trang câu hỏi đề xuất 9*

* 1. **Chức năng dựa trên tri thức:**



*Hình 4.1: Kết quả đề xuất style cho khách hàng*



*Hình 1.3:Giao diện trang đề xuất sản phẩm cho khách hàng theo style*

1. KẾT LUẬN

Đề tài:”Thiết kế hệ thống tri thức đưa kiến nghị dựa trên phong cách thời trang cho khách hàng trên website” đã tổng kết các kiến thức cơ bản của xây dựng hệ thống thông minh, lý thuyết khai phá dữ liệu. Đồng thời, từ việc tìm hiểu phần mềm khai phá dữ liệu Weka và ứng dụng nó trong đề tài đã giúp nhóm đưa ra được những kết quả sau:

* Đã xây dựng được bộ data thô bao gồm 100 dòng dữ liệu các trang phục đã được phân loại gu dựa theo những đặc điểm như kiểu dáng màu sắc, mẫu mã, chất liệu, form dạng hỗ trợ cho việc xây dựng dataset cho website.
* Xây dựng được bộ câu hỏi dạng quiz bao gồm 9 câu hỏi giúp xác định được phong cách thời trang của người dùng được tổng hợp từ những đặc trưng chung cơ bản của những người thuộc ba phong cách phổ biến với độ tin cậy cao, có thể tin tưởng sử dụng trong phần mềm dự đoán phong cách thời trang.
* Xây dựng được bộ dữ liệu bao gồm 300 dòng bao gồm kết quả của 300 bài test phong cách đã xây dựng ở trên phục vụ làm dữ liệu training cho mô hình huấn luyện.

Một khía cạnh khác, mặc dù dữ liệu đầu vào được lấy từ thực tế 100% tuy nhiên việc xác định phong cách thời trang của một người rất tương đối phụ thuộc vào nhiều yếu tố khác do đó bộ data không thể đúng 100%. Ngoài ra bộ dữ liệu chỉ mang tính chất đề tài môn học vì vậy để áp dụng vào thực tiễn thì cần thu thập thêm nhiều thuộc tính và nhiều dữ liệu thực tế khác để bộ dữ liệu đạt được độ chính xác và tin cậy cao hơn.

Bài báo cáo đã xây dựng được website có ứng dụng giải thuật học máy về thời trang và thực hiện phân loại được phong cách của từng khách hàng và đề xuất những sản phẩm thích hợp.

Về mặt lý thuyết bài viết có giới thiệu về học máy, bài toán phân lớp, thuật toán Support Vector Machine và cấu hình thuật toán.

Về mặt thử nghiệm, bài báo cáo giới thiệu kỹ thuật khai phá dữ liệu theo thuật toán Support Vector Machine, sử dụng ngôn ngữ python và java xây dựng phần mềm áp dụng phân loại sản phẩm đầu vào. Phần mềm này ứng dụng vào việc

hỗ trợ phân loại, ứng dụng có thể hoạt động tốt phụ thuộc vào độ tin cậy của bộ dữ

liệu huấn luyện.

Về mặt dữ liệu, kết quả dự trên bộ dữ liệu chuẩn cho ra kết quả chính xác.

Ngoài ra, kết quả huấn luyện trên bộ dữ liệu số lớn cũng cho ra kết quả như mong

muốn. Tuy nhiên, để nâng cao khả năng dự đoán của thuật toán, ta nên tiếp tục thu thập nhiều dữ liệu về bộ test hơn để 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 bán quần áo online hoàn chỉnh cả phần quản trị và khách hàng.

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.

1. TÀI LIỆU THAM KHẢO

Báo cáo đề tài PTUDDTTT

Tên đề tài: “Xây dựng hệ thống định dạng phân khúc sản phẩm điện thoại”  
 GVHD : Nguyễn Ngọc Duy

1. PHÂN CHIA CÔNG VIỆC

|  |  |  |
| --- | --- | --- |
| **Tên** | **Nội dung công việc** | **Ghi chú** |
| Tuấn | - Lên ý tưởng.  - Tìm hiểu đọc tài liệu.  - Viết toàn bộ báo cáo giữa kỳ.  - Thu thập data ‘elegant’.  - Triển khai hệ thống tri thức.  - Viết báo cáo cuối kỳ. |  |
| Thuận | - Tìm hiểu đọc tài liệu.  - Xây dựng và tổng hợp dataset.  - Đặc tả dữ liệu.  - Tiền xử lý dữ liệu.  - Xây dựng model.  - Viết báo cáo cuối kỳ. |  |
| Tiến | - Thu nhập data active.  - Xây dựng giao diện website (Front end).  - Kiểm tra và sửa lỗi cho website.  - Viết báo cáo cuối kỳ. |  |
| Thường | - Thiết kế cơ sở dữ liệu, database.  - Xây dựng các hàm, thủ tục phía Back end của website.  - Xây dựng hệ thống Website (Back end).  - Kiểm tra và sửa lỗi cho website.  - Viết báo cáo cuối kỳ. |  |
| Trung | - Thu nhập dât minimalist.  - Góp ý sửa đổi báo cáo.  - Viết phần cơ sở lý thuyết (báo cáo giữa kỳ).  - Viết báo cáo cuối kỳ. |  |