**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

****



**đỒ ÁN KẾT THÚC HỌC PHẦN**

**Môn: PHÂN TÍCH DỮ LIỆU**

Mã môn học: 841432

Khoa: Công nghệ thông tin

Chuyên ngành: Hệ thống thông tin

Nhóm nghiên cứu:

Trần Nguyên Lộc – 3120410297

Võ Đăng Quang – 3120410429

**Giảng viên phụ trách:**

**PHAN THÀNH HUẤN**

**Thành phố Hồ Chí Minh, tháng 12 năm 2023**

# **Lời cảm ơn**

Trước hết em xin gửi đến lời cảm ơn chân thành và sâu sắc nhất đến thầy TS. Phan Thành Huấn người trực tiếp hướng dẫn và tận tình chỉ bảo cho nhóm chúng em cho tới khi em hoàn thành đồ án của mình.

Tiếp đến em xin giành lời cảm ơn đến quý thầy cô Trường Đại học Sài Gòn – khoa Công nghệ thông tin đã truyền đạt cho em những kiến thức vô cùng quý báu và bổ ích trong suốt quá trình nghiên cứu và học tập tại trường.

Xin chân thành cảm ơn tới những người bạn đã luôn sát cánh cùng em, những lời động viên, những lần hỗ trợ những lúc cần thiết đã phần nào giúp em hoàn thành đồ án này.

Cuối cùng, em xin cảm ơn đến ba mẹ và người thân trong gia đình đã hỗ trợ và tạo điều kiện thuận lợi cho em trong suốt thời gian học tập và nghiên cứu tại Đại học Sài Gòn.

# **TÓM TẮT ĐỒ ÁN**

Đồ án này trình bày quá trình tìm kiếm, khám phá và phân tích thông tin từ tập dữ liệu tìm được. Trong đó có sử dụng các các thuật toán và phương pháp khai thác dữ liệu. Sau khi áp dụng các phương pháp, cần đánh giá hiệu suất của mô hình và lựa chọn mô hình tốt nhất dựa trên các tiêu chí như độ chính xác, độ tin cậy, độ phủ,… Cuối cùng, quá trình khai thác dữ liệu dẫn dến việc trích xuất tri thức từ dữ liệu và đưa ra quyết định thông minh có lợi cho việc marketing.

# **BẢNG ĐÁNH GIÁ MỨC ĐỘ THAM GIA**

|  |  |  |
| --- | --- | --- |
| Danh sách thành viên | MSSV | Mức độ tham gia |
| Trần Nguyên Lộc | 3120410297 | 100% |
| Võ Đăng Quang | 3120410429 | 100% |

# **Mục lục**

[**Lời cảm ơn** i](#_Toc150235405)

[**TÓM TẮT ĐỒ ÁN** ii](#_Toc150235406)

[**BẢNG ĐÁNH GIÁ MỨC ĐỘ THAM GIA** ii](#_Toc150235407)

[**Mục lục** iii](#_Toc150235408)

[**Danh mục hình ảnh** iv](#_Toc150235409)

[**Danh mục bảng biểu** v](#_Toc150235410)

[**Chương 1: KHÁI QUÁT ĐỒ ÁN** 1](#_Toc150235411)

[**1.1.** **Hành vi tiêu dùng của khách hàng** 1](#_Toc150235412)

[**1.2.** **Lí do chọn đề tài** 1](#_Toc150235413)

[**1.3.** **Mục tiêu đề tài** 1](#_Toc150235414)

[**1.4.** **Mô tả dữ liệu và cấu trúc dữ liệu** 2](#_Toc150235415)

[**Chương 2: QUY TRÌNH THỰC HIỆN VÀ KẾT QUẢ** 4](#_Toc150235416)

[**Tài liệu tham khảo** 5](#_Toc150235417)

# **Danh mục hình ảnh**

[Hình 1. 1. Kết quả trả về từ câu lệnh info() đối với bộ dữ liệu orders.csv 4](#_Toc150285481)

[Hình 1. 2. Kết quả trả về từ câu lệnh isnal().sum() đối với bộ dữ liệu orders.csv 4](#_Toc150285482)

[Hình 1. 3. Kết quả trả về từ câu lệnh isnal().sum() đối với bộ dữ liệu orders.csv 5](#_Toc150285483)

[Hình 1. 4. Kết quả trả về từ câu lệnh info() đối với bộ dữ liệu product-supplier.csv 5](#_Toc150285484)

[Hình 1. 5. Kết quả trả về từ câu lệnh isna().sum() đối với bộ dữ liệu product-supplier.csv 5](#_Toc150285485)

[Hình 1. 6. Kết quả trả về từ câu lệnh isnull().sum() đối với bộ dữ liệu product-supplier.csv 6](#_Toc150285486)

[Hình 1. 7. Kiểu dữ liệu ngày và tháng ban đầu 6](#_Toc150285487)

[Hình 1. 8. Kiểu dữ liệu ngày tháng sau khi được biến đổi để phù hợp cho việc phân tích 6](#_Toc150285488)

[Hình 1. 9. Lỗi không thống nhất dữ liệu trên cột Customer Status 7](#_Toc150285489)

[Hình 1. 10. Thông nhất kiểu dữ liệu trên cột Customer Status 7](#_Toc150285490)

[Hình 1. 11. Thêm một cột dữ liệu mới tên là Item Retail Value 7](#_Toc150285491)

[Hình 1. 12. Tạo dataframe mới nhằm thống kê số lượng sản phẩm được bán cũng như giá trị vốn - lãi trung bình của sản phẩm 8](#_Toc150285492)

[Hình 1. 13. Thông tin của bảng dữ liệu được merge lại với nhau 8](#_Toc150285493)

[Hình 1. 14. Dữ liệu của bảng dữ liệu được merge lại với nhau 9](#_Toc150285494)

[Hình 1. 15. Bộ dữ liệu cuối cùng trong quá trình tiền xử lý 9](#_Toc150285495)

# **Danh mục bảng biểu**

Bảng 1.4a

Bảng 1.4b

# **Chương 1: KHÁI QUÁT ĐỒ ÁN**

## **Hành vi tiêu dùng của khách hàng**

Ngày nay, khi cuộc sống chúng ta ngày một phát triển thì việc mua bán, trao đổi hàng hoá ngày càng được chú trọng và quan tâm. Khách hàng ngày càng có nhiều sự lựa chọn, nhiều cơ hội để tìm cho mình một loại sản phẩm phù hợp nhất và giá cả lại phải chăng. Sự hiểu biết về hành vi tiêu dùng là chìa khóa cho một chiến lược marketing thành công cả trong nước và quốc tế. Những hoạt động marketing chủ yếu này sẽ hiệu quả hơn khi được đặt trên cơ sở một sự hiểu biết về hành vi tiêu dùng.

## **Lí do chọn đề tài**

Xuất phát từ nhận thức trên, nhận thấy được tầm quan trọng của sự thay đổi hành vi của người tiêu dùng của trong thời điểm hiện tại, nên nhóm chúng em quyết định lựa chọn đề tài: ”Phân tích các mẫu mua hàng trực tuyến để tìm hiểu các luật kết hợp giữa các mặt hàng” để nhận diện ra mặt hàng nào là tiềm năng, phát hiện những yếu tố quan trọng tác động đến hành vi mua sắm từ đó có những chiến lược Marketing phù hợp.

## **Mục tiêu đề tài**

Mục tiêu của đề tài này sẽ phân tích:

* Dữ liệu về lịch sử mua hàng: sử dụng để xác định các mặt hàng thường được mua cùng nhau, các mặt hàng thường được mua trước hoặc sau một mặt hàng khác, và các mặt hàng thường được mua bởi các khách hàng có chung sở thích.
* Dữ liệu về dữ liệu vị trí: sử dụng để xác định các mặt hàng thường được mua ở cùng một địa điểm, các mặt hàng thường được mua bởi các khách hàng ở cùng một khu vực, và các mặt hàng thường được mua vào cùng một thời điểm.
* Dữ liệu về dữ liệu thanh toán: sử dụng để xác định các mặt hàng thường được mua bởi cùng một khách hàng, các mặt hàng thường được mua cùng với một phương thức thanh toán cụ thể, và các mặt hàng thường được mua vào cùng một thời điểm trong năm.

## **Mô tả dữ liệu và cấu trúc dữ liệu**

Dữ liệu được tổng hợp từ các nguồn sau Wholesale & Retail Orders Dataset (kaggle.com):

<https://www.kaggle.com/datasets/gabrielsantello/wholesale-and-retail-orders-dataset?select=orders.csv>

<https://www.kaggle.com/datasets/gabrielsantello/wholesale-and-retail-orders-dataset?select=product-supplier.csv>

Đây là những nguồn được đánh giá là có những bộ dữ liệu đáng tin cậy để phục vụ cho việc nghiên cứu hành vi tiêu dùng.

*Bảng 1.4a. Bảng mô tả cấu trúc của bộ dữ liệu orders.csv*

|  |  |
| --- | --- |
| Thuộc tính | Ý nghĩa |
| Customer ID | Mã khách hàng |
| Customer Status | Trạng thái khách hàng thân thiết (Ví dụ: Gold, Slive, Platinum) |
| Date Order was placed | Ngày đặt hàng |
| Delivery Date | Ngày giao hàng |
| Order ID | Mã đặt hàng |
| Product ID | Mã sản phẩm |
| Quantity Ordered | Số lượng sản phẩm đặt |
| Total Retail Price for This Order | Tổng giá bán lẻ cho đơn hàng này |
| Cost Price Per Unit | Giá vốn của mỗi món hàng |

*Bảng 1.4b. Bảng mô tả cấu trúc của bộ dữ liệu product-supplier.csv*

|  |  |
| --- | --- |
| Thuộc tính | Ý nghĩa |
| Product ID | Mã sản phẩm |
| Product Line | Dòng sản phẩm |
| Product Category | Danh mục sản phẩm |
| Product Group | Nhóm sản phẩm |
| Product Name | Tên sản phẩm |
| Supplier Country | Nước phân phối |
| Supplier Name | Tên nhà phân phối |
| Supplier ID | Mã nhà phân phối |

# **Chương 2: QUY TRÌNH THỰC HIỆN VÀ KẾT QUẢ**

## **Phân tích bộ dữ liệu**

Trong đồ án môn học này, chúng em sử dụng bộ dữ liệu trong hoạt động mua bán sỉ lẻ, bộ dữ liệu phân tích gồm 2 file csv tương ứng:

* *Bộ dữ liệu orders.csv* *(gồm 185004 bản ghi và 9 trường thuộc tính):*
* Là bộ dữ liệu chứa thông tin về lịch sử giao dịch về mua bán sản phẩm đối với khách hàng và người bán.
* Trong đó, mỗi bản ghi (record) trong bộ dữ liệu tượng trưng cho việc thông tin sản phẩm trong quá trình mua bán như mã sản phẩm, mã đặt hàng, mã khách hàng, mức giá, ngày đặt, ngày giao, số lượng giao dịch, tổng vốn thu được, giá trị mỗi sản phẩm.

Các thuộc tính trong bộ dữ liệu được mô tả bao như sau:

* + **Customer ID:** Mã khách hàng
  + **Customer Status:** Trạng thái khách hàng thân thiết
  + **Date Order was placed:** Ngày đặt hàng
  + **Delivery Date:** Ngày giao hàng
  + **Order ID:** Mã đặt hàng
  + **Product ID:** Mã sản phẩm
  + **Quantity Ordered:** Số lượng sản phẩm được đặt hàng
  + **Total Retail Price for This Order:** Tổng giá bán lẻ cho cả đơn hàng
  + **Cost Price Per Unit:** Giá vốn thu được của mỗi món hàng
  + Qua việc đánh giá bộ dữ liệu trên, nhóm em có một vài nhận định như sau: Đối với mỗi bản ghi trong bộ dữ liệu chỉ đánh giá cụ thể đối với một sản phẩm thông qua mã sản phẩm và bộ dữ liệu này cũng là một bộ dữ liệu đầy đủ không thiếu hay NULL dòng/trường nào.
* *Bộ dữ liệu product-supplier.csv (gồm 5505 bản ghi và 8 trường thuộc tính):*
* Là bộ dữ liệu mô tả thông tin chi tiết về các sản phẩm được mua bán.
* Trong đó, mỗi bản ghi (record) thể hiện thông tin chi tiết về từng thuộc tính của sản phẩm như mã sản phẩm, tên sản phẩm, nhóm sản phẩm, phân loại, dòng sản phẩm, nước phân phối, tên nhà phân phối, mã nhà phân phối.

Các thuộc tính trong bộ dữ liệu được mô tả như sau:

* + **Product ID:** Mã sản phẩm
  + **Product Line:** Dòng sản phẩm
  + **Product Category:** Danh mục sản phẩm
  + **Product Group:** Nhóm sản phẩm
  + **Product Name:** Tên sản phẩm
  + **Supplier Country:** Nước phân phối
  + **Supplier Name:** Tên nhà phân phối
  + **Supplier ID:** Mã nhà phân phối

Qua việc đánh giá bộ dữ liệu trên, nhóm em có một vài nhận định như sau:

* Bộ dữ liệu trên mô tả rất đầy đủ thông tin chi tiết của từng đối tượng sản phẩm.
* Do đó có thể thấy rõ tầm quan trọng không thể thiếu của nó đối trong việc phân tích dữ liệu bởi vì ta có thể kết hợp với bộ dữ liệu *order.csv* ở trên thì chúng ta có thể tìm ra các thông tin và các mối liên hệ giữa các sản phẩm với nhau. Ví dụ: sản phẩm bán chạy, lợi nhuận sản phẩm nào nhiều nhất ….

Đánh giá tổng thể cho thấy đây là một bộ dữ liệu khá đầy đủ cho việc phân tích dữ liệu. Ngoài ra, hướng đi chính cho quá trình phân trình phân tích dữ liệu trong đồ án môn học này sẽ là tìm ra các bộ luật nhằm xác định các mối liên hệ giữa sản phẩm và người dùng.

## **Tiền xử lí dữ liệu**

## **Kiểm tra bộ dữ liệu**

Trước khi tiền xử lý dữ liệu, nhóm chúng em tiến hành kiểm tra bộ dữ liệu một cách tổng quan nhất. Bắt đầu với bộ dữ liệu *orders.csv* trước tiên ta đọc bộ dữ liệu và gán vào biến dataset bằng lệnh ***info()***. Kết quả trả về thu được sẽ là:



Hình 1. 1. Kết quả trả về từ câu lệnh info() đối với bộ dữ liệu orders.csv

Điều đó cho thấy kiểu dữ liệu tổng thể của toàn bộ bộ dữ liệu trên. Tiếp sau đó là kiểm tra xem bộ dữ liệu trên có chứa tập Null và NaN hay không bằng cách đếm số lượng của chúng chứa trong bộ dữ liệu bằng lệnh *isna().sum()* và *isnull().sum()*, sau đó kết quả trả về sẽ là:



Hình 1. 2. Kết quả trả về từ câu lệnh isnal().sum() đối với bộ dữ liệu orders.csv



Hình 1. 3. Kết quả trả về từ câu lệnh isnal().sum() đối với bộ dữ liệu orders.csv

Tiếp sau đó ta vẫn tiến hành kiểm tra bộ dữ liệu như trên đối với bộ dữ liệu *product-supplier.csv* và kết quả thu được như sau:



Hình 1. 4. Kết quả trả về từ câu lệnh info() đối với bộ dữ liệu product-supplier.csv



Hình 1. 5. Kết quả trả về từ câu lệnh isna().sum() đối với bộ dữ liệu product-supplier.csv



Hình 1. 6. Kết quả trả về từ câu lệnh isnull().sum() đối với bộ dữ liệu product-supplier.csv

Kết quả trả về từ việc kiểm tra 2 bộ dữ liệu trên cho thấy không hề có một bản ghi nào bị trống dữ liệu.

## **Biến đổi và thêm mới dữ liệu**

Để thuận tiện hơn trong quá trình phân tích dữ liệu, chúng em tiến hành biến đổi dữ liệu bằng cách thay đổi một số trường dữ liệu, thay đổi kiểu dữ liệu, tạo dữ liệu mới và kết hợp dữ liệu với nhau tạo ra một dataframe phù hợp cho quá trình phân tích.

Đầu tiên, đối với bộ dữ liệu *orders.csv*, ta chuyển kiểu dữ liệu của cột *Date Order was placed* và *Delivery Date* về lại dạng ngày tháng bằng số hoàn chỉnh, thay vì kiểu dữ liệu viết tắt tháng như ban đầu.



Hình 1. 7. Kiểu dữ liệu ngày và tháng ban đầu



Hình 1. 8. Kiểu dữ liệu ngày tháng sau khi được biến đổi để phù hợp cho việc phân tích

Trong quá trình kiểm tra lại bộ dữ liệu thì nhóm chúng em có nhận thấy một thiếu sót trong quá trình tổng hợp dữ liệu của bên cung cấp bộ dữ liệu, bằng cách sử dụng câu lệnh *unique()* đối với mỗi cột trong bộ dữ liệu, chúng em nhận thấy đối với cột *Customer Status* bị lỗi lặp dữ liệu trùng trong một số bản ghi. Ví dụ: Gold và GOLD, Silver và SILVER, một số lỗi viết hoa và viết thường ko thông nhất dẫn đến lỗi không khớp dữ liệu với nhau:



Hình 1. 9. Lỗi không thống nhất dữ liệu trên cột Customer Status

Để xử lý vấn đề này, chúng em xử dụng câu lệnh *str.lower()* và *str.capitalize()* để thống nhất lại kiểu dữ liệu chung trong cột này. Ngoài ra chúng em còn chuyển kiểu dữ liệu của cột từ *object* sang *string* để thuận tiện cho việc đọc – ghi dữ liệu:



Hình 1. 10. Thông nhất kiểu dữ liệu trên cột Customer Status

Ngoài ra, để thuận tiện hơn cho việc phân tích nhóm chúng em quyết định thêm mới dữ liệu bằng cách tạo một cột mới tên là *Item Retail Value* (giá bán lẻ trên từng sản phẩm)dựa trên dữ liệu đến từ cột *Total Retail Price for This Order* (Tổng giá bán lẻ trên order này)và *Quanlity Order* (Số lượng sản phẩm) theo công thức *Item Retail Value = Total Retail Price for This Order / Quanlity Order*.



Hình 1. 11. Thêm một cột dữ liệu mới tên là Item Retail Value

Bên cạnh đó, để có thể biết được số lượng được bán ra của từng sản phẩm cũng như là giá trị vốn lẫn lãi trung bình thu được của sản phẩm đó, nhóm chúng em tiến hành tạo mới một dataframe dựa trên bộ dữ liệu *orders.csv* đã được tiền xử lý trước như sau: Đầu tiên nhóm (group) các mã sản phẩm với nhau (nhằm tạo ra một bản ghi về sản phẩm là duy nhất trong dataframe), song đếm số lượng sản phẩm được group lại, bên cạnh đó tính toán lại giá trị trung bình trên từng sản phẩm của các cột như *Cost Price Per Unit* (Giá vốn trên từng sản phẩm) và *Item Retail Value* (giá bán lẻ trên từng sản phẩm) sao cho phù hợp với lại từng sản phẩm đã được nhóm (group) lại với nhau.



Hình 1. 12. Tạo dataframe mới nhằm thống kê số lượng sản phẩm được bán cũng như giá trị vốn - lãi trung bình của sản phẩm

## **Bộ dữ liệu tổng thể**

Sau quá trình tiền xử lý ở phần trên, dữ liệu chúng ta thu được ở bước này là bộ dữ liệu dataframe về số lượng, giá vốn – lãi trung bình trên từng mặt hàng, do đó cái mà nhóm chúng em thiếu để có thể phân tích đặc tính của mặt hàng đó chính là thông tin tổng thể của từng mặt hàng. Thông tin tổng thể của từng mặt hàng được chứa trong bộ dữ liệu *product-supplier.csv*, cho nên nhóm chúng em quyết định trộn (merge) dữ liệu từ 2 dataframe đó lại với nhau nhằm mục đích tạo ra một bộ dữ liệu hoàn chỉnh hơn.

Kết hợp dữ liệu tư 2 bộ dữ liệu với nhau khác giống với việc join 2 table lại với nhau trong SQL, bằng cách sử dụng câu lệnh *merge()*với tham số điều kiện kết hợp bảng là *Product ID* thì ta có thể trộn dữ liệu từ 2 bảng lại với nhau. Dữ liệu được kết hợp lại sẽ có mô tả như sau:



Hình 1. 13. Thông tin của bảng dữ liệu được merge lại với nhau

Và bộ dữ liệu sẽ có dạng như sau:



Hình 1. 14. Dữ liệu của bảng dữ liệu được merge lại với nhau

Tuy vậy, dữ liệu bên trong khá là nhiều và có thể gây tốn kém tài nguyên trong quá trình phân tích. Nhóm chúng em quyết định lược bỏ đi những cột không cần thiết như *Product Line, Product Group, Supplier Country, Supplier Name* và *Supplier ID*. Bên cạnh đó, đổi tên các cột sao cho dễ hiễu và phù hợp hơn như *N Rows* đổi thành *Total Sold* (Tổng sản phẩm bán được), *Cost Price Per Unit* đổi thành *Wholesale Price* (Giá sỉ), *Item Retail Value* đổi thành *Retail Price* (Giá lẻ).



Hình 1. 15. Bộ dữ liệu cuối cùng trong quá trình tiền xử lý

Tóm lại, qua một chuỗi tiền xử lý trên, từ 2 bộ dữ liệu về lịch sử mua bán và thông tin mặt hàng, nhóm chúng em đã tóm gọn và xử lý thành một bộ dữ liệu hoàn chỉnh và đầy đủ về thông tin được bán ra như số lượng, giá sỉ-lẻ của từng mặt hàng. Bộ liệu thu được qua quá trình tiền xử lý trên sẽ chính là tiền đề để nhóm chúng em phân tích sâu hơn trong những phần sau.

# **Tài liệu tham khảo**