**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT HƯNG YÊN**

****

**BÀI TẬP LỚN**

**PHÂN TÍCH XỬ LÝ DỮ LIỆU**

**SÀN THƯƠNG MẠI ĐIỆN TỬ MERCADO LIVRE**

NGÀNH: KHOA HỌC MÁY TÍNH

CHUYÊN NGÀNH: TRÍ TUỆ NHÂN TẠO VÀ KHOA HỌC DỮ LIỆU

SINH VIÊN: **VŨ QUANG PHÚC**

MÃ SINH VIÊN: **12522079**

MÃ LỚP: **124221**

HƯỚNG DẪN: **TS. HOÀNG QUỐC VIỆT**

**HƯNG YÊN – 2024**

**NHẬN XÉT**

**Nhận xét của giáo viên hướng dẫn**

**GIẢNG VIÊN HƯỚNG DẪN**

*(Ký và ghi rõ họ tên)*

**LỜI CAM ĐOAN**

Em xin cam đoan bài tập lớn môn Lập trình Python nâng cao có tên là “Phân tích xử lý dữ liệu sàn thương mại điện tử Mercado Livre” là kết quả thực hiện của bản thân em dưới sự hướng dẫn của thầy Hoàng Quốc Việt

Những phần sử dụng tài liệu tham khảo trong bài tập lớn đã được nêu rõ trong phần tài liệu tham khảo. Các kết quả trình bày trong bài tập lớn và chương trình xây dựng được hoàn toàn là kết quả do bản thân em thực hiện.

Nếu vi phạm lời cam đoan này, em xin chịu hoàn toàn trách nhiệm trước khoa và nhà trường.

*Hưng Yên, ngày … tháng*…*năm*…

Sinh viên

…………………………………..

**LỜI CẢM ƠN**

Để có thể hoàn thành bài tập lớn này, lời đầu tiên em xin phép gửi lời cảm ơn tới bộ môn Khoa học máy tính, Khoa Công nghệ thông tin – Trường Đại học Sư phạm Kỹ thuật Hưng Yên đã tạo điều kiện thuận lợi cho em thực hiện bài tập lớn môn học này.

Đặc biệt em xin chân thành cảm ơn thầy Hoàng Quốc Việt đã rất tận tình hướng dẫn, chỉ bảo em trong suốt thời gian thực hiện bài tập lớn vừa qua.

Em cũng xin chân thành cảm ơn tất cả các Thầy, các Cô trong Trường đã tận tình giảng dạy, trang bị cho em những kiến thức cần thiết, quý báu để giúp em thực hiện được bài tập lớn này.

Mặc dù em đã có cố gắng, nhưng với trình độ còn hạn chế, trong quá trình thực hiện đề tài không tránh khỏi những thiếu sót. Em hy vọng sẽ nhận được những ý kiến nhận xét, góp ý của các Thầy cô về những kết quả triển khai trong bài tập lớn.

Em xin trân trọng cảm ơn!

**MỤC LỤC**

[CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI 7](#_Toc165580055)

[1.1 Lý do chọn đề tài 7](#_Toc165580056)

[1.2 Mục tiêu của đề tài 7](#_Toc165580057)

[1.2.1 Mục tiêu tổng quát 7](#_Toc165580058)

[1.2.2 Mục tiêu cụ thể 7](#_Toc165580059)

[1.3 Giới hạn và phạm vi của đề tài 8](#_Toc165580060)

[1.3.1 Đối tượng nghiên cứu 8](#_Toc165580061)

[1.3.2 Phạm vi 8](#_Toc165580062)

[1.4 Nội dung thực hiện 8](#_Toc165580063)

[1.5 Phương pháp tiếp cận 8](#_Toc165580064)

[CHƯƠNG 2: KHẢO SÁT VÀ PHÂN TÍCH THIẾT KẾ HỆ THỐNG 9](#_Toc165580065)

[2.1 Phát biểu yêu cầu 9](#_Toc165580066)

[2.1.1 Danh sách các yêu cầu 10](#_Toc165580067)

[2.1.2 Biểu đồ ca sử dụng 14](#_Toc165580068)

[2.1.3 Phân rã usecase mua hàng 14](#_Toc165580069)

[2.1.4 Phân rã usecase chức năng nhận hàng 16](#_Toc165580070)

[2.1.5 Phân rã usecase Thêm vào giỏ hàng 18](#_Toc165580071)

[2.1.6 Phân rã usecase chức năng tìm kiếm sản phẩm 19](#_Toc165580072)

[2.1.7 Phân rã usecase đánh giá sản phẩm 21](#_Toc165580073)

[2.1.8 Phân rã usecase Đăng kí tài khoản 22](#_Toc165580074)

[2.1.9 Phân rã chức năng cập nhật thông tin khách hàng 24](#_Toc165580075)

[2.1.10 Phân rã usecase đăng kí cửa hàng 26](#_Toc165580076)

[2.1.11 Phân rã usecase Tạo Dashboard 28](#_Toc165580077)

[2.1.12 Phân rã usecase thêm sản phẩm 29](#_Toc165580078)

[2.1.13 Phân rã usecase xóa sản phẩm 31](#_Toc165580079)

[2.1.14 Phân rã usecase cập nhật sản phẩm 33](#_Toc165580080)

[2.1.15 Phân rã usecase xác nhận đơn hàng 34](#_Toc165580081)

[2.1.16 Phân rã usecase kiểm duyệt sản phẩm 36](#_Toc165580082)

[2.1.17 Đặc tả ca usecase thêm nhân viên 38](#_Toc165580083)

[2.2 Biểu đồ lớp thực thể 40](#_Toc165580084)

[2.2.1 Danh sách các lớp đối tượng 40](#_Toc165580085)

[2.2.2 Chi tiết hóa các lớp đối tượng 41](#_Toc165580086)

[2.2.3 Mô hình hóa các lớp đối tượng 51](#_Toc165580087)

[2.3 Thiết kế CSDL 52](#_Toc165580088)

[2.3.1 Lược đồ CSDL 52](#_Toc165580089)

[2.3.2 Mô tả chi tiết từng bảng trong CSDL 52](#_Toc165580090)

[CHƯƠNG 3: XÂY DỰNG ỨNG DỤNG WINDOWS FORMS 62](#_Toc165580091)

[3.1 Triển khai các chức năng nghiệp vụ 62](#_Toc165580092)

[3.1.1 Giao diện các sản phẩm mới nhất 62](#_Toc165580093)

[3.1.2 Giao diện chi tiết sản phẩm 63](#_Toc165580094)

[3.1.3 Chức năng tìm kiếm sàn phẩm 64](#_Toc165580095)

[3.1.4 Giao diện Giỏ hàng 65](#_Toc165580096)

[3.1.5 Giao diện dặt hàng 66](#_Toc165580097)

[3.1.6 Giao diện thông tin khách hàng 67](#_Toc165580098)

[3.1.7 Giao diện các đơn hàng đã đặt 67](#_Toc165580099)

[3.1.8 Giao diện các sản phẩm của shop 68](#_Toc165580100)

[3.1.9 Giao diện cập nhật sản phẩm 68](#_Toc165580101)

[3.1.10 Chức năng quản lý sản phẩm 69](#_Toc165580102)

[3.2 Triển khai các chức năng thống kê, báo cáo 69](#_Toc165580103)

[3.2.1 Giao diện Dashboard 69](#_Toc165580104)

[3.3 Kiểm thử và triển khai ứng dụng 70](#_Toc165580105)

[3.3.1 Kiểm thử 70](#_Toc165580106)

[3.3.2 Đóng gói ứng dụng 70](#_Toc165580107)

[3.3.3 Triển khai ứng dụng 70](#_Toc165580108)

[KẾT LUẬN 71](#_Toc165580109)

[TÀI LIỆU THAM KHẢO 73](#_Toc165580110)

**DANH MỤC HÌNH ẢNH**

No table of figures entries found.

**DANH MỤC BẢNG BIỂU**

[Bảng 1.1 Danh sách các features trong dataset 7](#_Toc168582787)

# GIỚI THIỆU BÀI TOÁN

## Bài toán

Hiện nay các sàn thương mại điện tử đang là xu thế của toàn cầu về quá trình mua bán giữa người tiêu dùng và doanh nghiệp. Nó không những công nghệ hóa các giai đoạn, thao tác của quá trình vận chuyển hàng hóa đến tay người tiêu dùng. Là 1 trong những mắt xích không thể thiếu trong thời đại công nghệ phát triển như hiện nay, việc mua bán đã trở lên dễ dàng hơn bao giờ hết chỉ với 1 chiếc smartphone có kết nối internet là chúng ta đã có thể mua được hàng hóa, dịch vụ ở khắp mọi nơi trên thế giới. Và thị trường Brazil là 1 trong số các thị trường đã và đang ứng dụng các công nghệ của sàn thương mại điện tử rất tốt để nâng cao trải nghiệm của người dùng trong thời đại cách mạng AI hiện nay và Mercado Livre là 1 trong những sàn thương mại điện tử lớn nhất ở Brazil thời điểm hiện tại.

Bằng cách phân tích dữ liệu của Mercado Livre, chúng ta có thể có được 1 cái nhìn tổng quan về xu hướng mua sắm, các danh mục sản phẩm được ưa chuộng của người tiêu dùng Brazil. Sẽ có rất nhiều các danh mục đi theo đó là các sàn phẩm kèm theo. Dữ liệu này được tổng hợp từ rất nhiều nguồn khác nhau nên sẽ không đảm bảo được phần uy tín cũng như độ xác thực 100% so với thực tế. Nếu dữ liệu không phải cho sàn cung cấp thì gần như không có độ đáng tin.

Kể cả như vậy nhưng điều đó không ảnh hưởng quá lớn đến quá trình phân tích bộ dữ liệu này. Chúng ta vẫn sẽ cần phải hiểu được bộ dữ liệu này và từ đó đưa ra được các sự thiếu sót, không đồng nhất, mất mát dữ liệu .. bên trong nó. Bộ dữ liệu sẽ giúp chúng ta hiểu thêm về xu hướng, mức độ gia tăng về giá cả lẫn khác hàng, sản phẩm qua các năm,.. Từ đó đưa ra được những đánh giá khách quan và giải pháp cũng như vấn đề mà doanh nghiệp đang gặp phải(nếu có). Lược bỏ các thông số, nội dung không có ý nghĩa nhiều, tập trung vào các nội dung trọng tâm, mang tích chiến lược từ đó có thể đưa ra giải pháp cho doanh nghiệp hoặc vấn đề mà chúng ta đang băn khoăn.

Để có thể hiểu và làm tốt quá trình phân tích này chúng ta cần các kiến thức liên quan đến nghiệp vụ của 1 sàn thương mại điện tử, các công cụ phục vụ cho việc phân tích dữ liệu như IPython, jupyter notebook or jupyter lab, pandas, seaborn,...

## Trình bày dữ liệu của bài toán

Dữ liệu của bài toán đang ở link sau: [E-commerce\_dataset(full).csv | Github](https://github.com/vuquangphucliar/Ecommerce-analysis/blob/main/ecom_data(full).csv)

Bộ dữ liệu này thuộc sàn thương mại Mercado Lirve trong khoảng thời gian từ 10/2016 - 9/2018 được tổng hợp lại từ các giao dịch của người tiêu dùng trên phạm vi toàn lãnh thổ Brazil.

Mỗi 1 bản ghi đại diện do 1 giao dịch trên sàn lưu thông tin về người bán, người mua, sản phẩm, giá cả và số lượng. Do tất cả các bản ghi trong dataset đều mang mã đơn hàng khác nhau nên có thể coi mỗi 1 bản ghi như là 1 chi tiết của 1 đơn hàng.

Phần mô tả dữ liệu bên dưới sẽ tập trung ngắn gọn về các features trong dataset, Chi tiết mô tả về bộ dữ liệu có thể được tìm thấy tại đây: [E-commerce-analysis(full)](https://github.com/vuquangphucliar/Ecommerce-analysis.git).

Các features có trong dataset:

Bảng 1.1 Danh sách các features trong dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Feature** | **Not Null** | **Type** | **Desctiption** |
| 1 | Order\_id | 97254 | object | Id của đơn hàng |
| 2 | Customer\_id | 97254 | object | Id của khách hàng |
| 3 | Quantity | 97254 | int | Số lượng sản phẩm đã mua |
| 4 | Price\_MRP | 97254 | float(64) | Giá bán lẻ |
| 5 | Payment | 97254 | float(64 | Giá thanh toán |
| 6 | Timestamp | 97254 | object | Thời gian mua |
| 7 | Rating | 97254 | int | Đánh giá |
| 8 | Category | 97254 | object | Danh mục sản phẩm |
| 9 | Product\_id | 97254 | object | Id sản phẩm |
| 10 | Payment\_type | 97254 | object | Phương thức thanh toán |
| 11 | Order\_status | 97254 | object | Trạng thái đơn hàng |
| 12 | Weight | 97254 | int | Trọng lượng sàn phẩm |
| 13 | Length | 97254 | int | Độ dài sàn phẩm |
| 14 | Width | 97254 | int | Chiều rộng sàn phẩm |
| 15 | Height | 97254 | int | Chiều cao sàn phẩm |
| 16 | Customer\_state | 97254 | object | Bang của khách hàng |
| 17 | Seller\_id | 97254 | object | Id người bán |
| 18 | seller\_state | 97254 | object | Bang của người bán |

Bộ dữ liệu trên có tổng cộng là 97254 rows x 19 columns ( column ở cuối cùng Payment\_installments là thời hạn trả góp hiện chưa cần dùng tới nên sẽ không đề cập trong báo cáo này).

Thống kê cơ bản các biến định lượng trong dataset:

Mô tả cơ bản các biến định tính trong dataset:

* rating: chứa các số từ 1-5 đại diện cho số sao khách hàng đánh giá
* category: chứa tên của 71 danh mục trên sàn được giao dịch
* payment\_type: gồm 5 phương thức thanh toán, credit card, debit card,..
* order\_status: gồm 7 trạng thái của đơn hàng invoiced, delivered, processing,..
* customer\_state: gồm các tên viết tắt của 27 bang tại Brazil
* seller\_state: tương tự như customer\_state( tên đầy đủ được gắn tại link trên)
* payment\_installments: gồm các số từ 1-24 đại diện cho 12 tháng.

## Tiền xử lý dữ liệu

## Thống kê dữ liệu

## Trực quan hóa dữ liệu

# CƠ SỞ LÝ THUYẾT

## Numpy

NumPy (Numeric Python) là một thư viện Python mã nguồn mở được sử dụng trong hầu hết các lĩnh vực khoa học và kỹ thuật. Đây là tiêu chuẩn chung để làm việc với dữ liệu số bằng Python và là tiêu chuẩn cốt lõi của hệ sinh thái Python và PyData. Numpy hỗ trợ mạnh mẽ việc tính toán với ma trận, vector và các hàm đại số tuyến tính cơ bản. Điều này làm cho nó trở thành một công cụ quan trọng trong việc hiện thực các thuật toán Machine Learning.

## Pandas

Pandas là một thư viện Python cung cấp các cấu trúc dữ liệu nhanh, mạnh mẽ, linh hoạt và mang hàm ý. Tên thư viện được bắt nguồn từ “panel data” (bảng dữ liệu). Pandas được thiết kế để làm việc dễ dàng và trực quan với dữ liệu có cấu trúc (dạng bảng, đa chiều, có tiềm năng không đồng nhất) và dữ liệu chuỗi thời gian1. Thư viện này thường được sử dụng để thao tác, phân tích và dọn dẹp dữ liệu. Pandas cung cấp nhiều cấu trúc dữ liệu và phép tính hỗ trợ thao tác dữ liệu số và dữ liệu thời gian2. Nó là một công cụ toàn diện cho xử lý dữ liệu trong Python, phù hợp cho nhiều loại tác vụ phân tích và quản lý dữ liệu

## Matplotlib

Matplotlib là một thư viện vẽ đồ thị cho ngôn ngữ lập trình Python và được mở rộng từ thư viện toán học số học NumPy. Nó cung cấp một API hướng đối tượng để nhúng các biểu đồ vào ứng dụng sử dụng các toolkit GUI đa năng như Tkinter, wxPython, Qt hoặc GTK. Matplotlib cho phép bạn tạo và hiển thị các biểu đồ, hình ảnh và các hình vẽ khác, rất hữu ích cho việc trực quan hóa dữ liệu trong Python. Một phần quan trọng của Matplotlib là Pyplot, một module cung cấp các hàm đơn giản để thêm các thành phần plot như đường, hình ảnh, văn bản và nhiều hơn nữa vào các khung vẽ (axes).

# GIẢI PHÁP

## Mã nguồn mô tả dữ liệu

## Mã nguồn tiền sử lý dữ liệu

## Mã nguồn thống kê dữ liệu

## Mã nguồn trực quan hóa dữ liệu

# KẾT LUẬN

**Kết quả đạt được**

Sau khi hoàn thành xong bài tập lớn phân tích và xử lý dữ liệu sàn thương mại điện tử Mercado Livre với Python đã giúp em học thêm nhiều kiến thức hơn về Python, và các thông tin cơ bản về tình hình kinh tế, xu hướng tiêu dùng của Brazil. Quan trọng hơn hết vẫn là các kiến thức liên quan đến quá trình phân tích, xử lý 1 bộ dữ liệu khi lần đầu chạm vào. Về kiến thức python em đã đạt được:

* Cách sử dụng IPython, Jupyter notebook
* Các thư viện phục vụ cho việc phân tích và trực quan dữ liệu: numpy, pandas,...

Về quá trình các bước phân tích 1 bộ dữ liệu mới như sau:

* Tìm hiểu về nghiệp vụ của dataset( dữ liệu lấy từ đâu, dữ liệu mô tả về lĩnh vực gì, vấn đề mà doanh nghiệp đang phải đối mặt( hoặc vấn đề chung của lĩnh vực này), bộ dữ liệu đã cho có khả năng đưa ra các insight có ích giải quyết được vấn đề không, nếu không thì có có thể cho biết điều gì,..).
* Khai thác, phân tích bộ dữ liệu, (chia bộ dữ liệu thành 2 loại biến, định tính và định lượng, phân tích độ phân phối, tương quan, xu hướng tập trung của dữ liệu, cố gắng hiểu được bộ dữ liệu, xem nó đang mô tả điều gì) .
* Lấy insight trong dataset( đặt câu hỏi và kết hợp với trực quan cho dữ liệu, tập trung vào những câu hỏi mang giá trị cao, trực quan mang tính so sánh cao).
* Biến insight thành actionable insight( chọn lọc các câu hỏi để đi đến giải pháp, các đề xuất, giải pháp phải tính hiện thực cao và có khả năng là solution cho vấn đề, bài toán).

Các điều cần chú ý về kĩ thuật khi phân tích 1 bộ dữ liệu (Pov):

* Phân tích luôn luôn đi với trực quan, không phân tích mọi thứ lan man.
* Dữ liệu hashed, id không có khả năng trực quan đơn lẻ, nên dùng để đếm.
* Dữ liệu nào có thể số hóa thì nên số hóa (phải kèm theo quy tắc số hóa)
* Nếu dữ liệu có số lượng unique value cao, nên dùng top để so sánh.
* Dữ liệu trực quan phải có sự so sánh, tương phản, tương cao.
* Chọn biểu đồ thích hợp dự trên số lượng feature, tính chất feature,..
* Dữ liệu ngày tháng thường thống kê theo tháng, năm, mùa,.
* Chú thích sao cho người xem không cần phải đối chiếu qua trục để lấy value.

**Hạn chế của đề tài**

Tuy đề tài này đã được hoàn thành 1 cách chu đáo nhưng đâu đó vẫn có những điểm bất thường hoặc cũng có thể là sai sót. Một phần do đây là đề tài đầu tiên nên không thể chuẩn bị kiến thức 1 cách hoàn thiện được, cộng thêm thời gian tìm tòi nghiên cứu khá lâu để có thể ra được kết quả như hiện tại. Các hạn chế của đề tài:

* Thời gian chuẩn bị lâu
* Dữ liệu thiếu chính xác
* Đội ngũ thiếu nghiệp vụ chuyên môn, chưa có trải nghiệm thực tế

**Hướng phát triển**

Đề tài này đã trở thành 1 bước đệm vững chắc cho em để có thể tiếp tục khám phá, phân tích các bộ dữ liệu mới sau này 1 cách dễ dàng hơn, giúp em trang bị các kiến thức cần thiết để có thể học được các kiến thức nâng cao về sau như Machine learning.

# TÀI LIỆU THAM KHẢO

|  |  |  |
| --- | --- | --- |
| [1] |  | K. CNTT, Phân tích thiết kế phần mềm, Hưng Yên: Đại học SPKT H. |
| [2] |  | K. CNTT, Cơ sở dữ liệu, Hưng Yên: Đại học SPKT Hưng Yên, 2012. |
| [3] |  | K. CNTT, Lập trình ứng dụng windows forms, Hưng Yên: Đại học SPKT Hưng Yên, 2019. |
| [4] |  | W3schools, "W3schools," [Online]. Available: https://www.w3schools.com/. |
| [5] |  | K. CNTT, Kiểm thử phần mềm, Hưng Yên: Đại học SPKT Hưng Yên, 2016. |