

TRƯỜNG ĐẠI HỌC KINH TẾ
KHOA THÔNG KÊ – TIN HỌC



BÁO CÁO TỐT NGHIỆP

**NGÀNH HỆ THỐNG THÔNG TIN QUẢN LÝ
CHUYÊN NGÀNH QUẢN TRỊ HỆ THỐNG THÔNG TIN**

**PHÂN TÍCH DỮ LIỆU LAPTOP CỦA THẾ GIỚI DI
ĐỘNG VÀ XÂY DỰNG MÔ HÌNH ĐỂ XUẤT LAPTOP
CHO KHÁCH HÀNG**

Sinh viên thực hiện	: Nguyễn Thị Duyên Nguyễn Thị Mùi
Lớp	: 47K21.1
Đơn vị thực tập	: Trung tâm VNPT IT – KV3
Cán bộ hướng dẫn	: Đặng Thái Bình
Giảng viên hướng dẫn	: ThS. Nguyễn Văn Chức

Đà Nẵng, 06/2025

NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP

NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP

Họ và tên sinh viên: Nguyễn Thị Duyên
Lớp: 4TK21_1 Khoa: Thông tin và Truyền thông

Thực tập từ ngày: 10.1.2025 đến ngày: 10.5.2025

Tại: Đà Nẵng, VNPT - IT khu vực 3
Địa chỉ: Số 44, đường 2/9, Hòa Cường Bắc, Hải Châu, TP. Đà Nẵng

Sau quá trình thực tập tại đơn vị của sinh viên, chúng tôi có một số nhận xét, đánh giá như sau:

STT	Mục đánh giá	Rất tệ	Tệ	Bình thường	Tốt	Rất tốt
1	Về thái độ, ý thức, đạo đức, kỷ luật và văn hóa công ty	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
2	Kiến thức chuyên môn	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
3	Khả năng hòa nhập, thích nghi và tác phong nghề nghiệp	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
4	Trách nhiệm, sáng tạo trong công việc	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Các nhận xét khác (Khoa chúng tôi mong muốn nhận thêm những ý kiến khác từ quý doanh nghiệp nhằm nâng cao chất lượng đào tạo)

Chị Duyên có ý thức làm việc cẩn thận, khéo léo, linh hoạt để đạt kết quả cao. Cần tiếp tục rèn luyện nâng cao kỹ năng mềm.

Điểm:

8.5/10

Hạnh

Đặng Thị Bình

Đà Nẵng ngày 10 tháng 5 năm 2025

Xác nhận của đơn vị thực tập



Phan Văn Thảo

NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP

Họ và tên sinh viên: Nguyễn Thị Mùi.....
 Lớp: K21.1..... Khoa: Thống tin - Tin học Trường: Đại học Kinh tế - ĐHQGHN
 Thực tập từ ngày: 10.1.2025 đến ngày: 10.5.2025
 Tại: Trung tâm VNPT - Khu vực 3.....
 Địa chỉ: 344 Nguyễn 2/9, Hòa Lạc, Huyện Hòa Lạc, Thành phố Hồ Chí Minh
 Sau quá trình thực tập tại đơn vị của sinh viên, chúng tôi có một số nhận xét, đánh giá như sau:

STT	Mục đánh giá	Rất tệ	Tệ	Bình thường	Tốt	Rất tốt
1	Về thái độ, ý thức, đạo đức, kỷ luật và văn hóa công ty	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
2	Kiến thức chuyên môn	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
3	Khả năng hòa nhập, thích nghi và tác phong nghề nghiệp	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
4	Trách nhiệm, sáng tạo trong công việc	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Các nhận xét khác (Khoa chúng tôi mong muốn nhận thêm những ý kiến khác từ quý doanh nghiệp nhằm nâng cao chất lượng đào tạo)

Sinh viên Mùi có ý chí làm việc cao, nỗ lực, năng động, linh hoạt và phụ trách tài liệu nghiên cứu...
 Cố gắng, nhiệt huyết, vui vẻ, hòa đồng.

Điểm:

Q5/10

Đà Nẵng, ngày 10 tháng 5 năm 2025

Xác nhận của đơn vị thực tập



Phan Văn Thảo

Võ Anh
 Đồng Thái Bình

LỜI CẢM ƠN

Đầu tiên, chúng em muốn gửi lời cảm ơn chân thành các thầy cô trong Khoa Thống Kê – Tin Học và trường Đại học Kinh tế - Đại học Đà Nẵng đã truyền đạt nhiều kiến thức bổ ích trong suốt thời gian vừa qua. Những kiến thức và kỹ năng được học tập đã giúp chúng em có nền tảng vững chắc để thực hiện tốt đê tài và các công việc tại đơn vị thực tập.

Đặc biệt, chúng em xin gửi lời cảm ơn sâu sắc đến thầy Nguyễn Văn Chúc - người thầy đã tận tình hướng dẫn theo sát quá trình chúng em thực hiện báo cáo, thầy đã hỗ trợ và đưa ra các lời khuyên cần thiết để chúng em hoàn thành bài báo cáo một cách hoàn chỉnh.

Chúng em cũng xin gửi lời cảm ơn đến quý công ty và anh Đặng Thái Bình tại đơn vị thực tập đã tạo điều kiện thuận lợi để chúng em có cơ hội học hỏi và tiếp thu nhiều kiến thức thực tế tại công ty trong quá trình thực tập.

Vì kiến thức bản thân còn nhiều hạn chế, trong quá trình hoàn thiện đề án không tránh khỏi sai sót. Chúng em rất mong nhận được sự thông cảm và những ý kiến đóng góp, chỉ bảo từ quý thầy cô, cũng như quý công ty để em có thể khắc phục và hoàn thiện bản thân mình hơn trong quá trình làm việc sau này. Cuối cùng, em kính chúc quý thầy cô luôn dồi dào sức khỏe và thành công trong sự nghiệp giảng dạy cao quý.

Chúng em xin chân thành cảm ơn!

LỜI CAM ĐOAN

Chúng em xin cam đoan rằng đề tài “Phân tích dữ liệu laptop của Thế giới di động và xây dựng mô hình đề xuất laptop cho khách hàng” là kết quả do nhóm chúng em nghiên cứu, tổng hợp và thực hiện dưới sự hướng dẫn tận tâm của thầy Nguyễn Văn Chức và mentor tại đơn vị thực tập. Chúng em xin đảm bảo rằng những nội dung được trình bày trong báo cáo được tham khảo và tổng hợp từ các nguồn tài liệu khác nhau đã được trích dẫn rõ ràng, trung thực và không hề tồn tại sự gian lận. Nếu có điều gì sai phạm, chúng em xin chịu mọi hình thức kỷ luật theo quy định.

MỤC LỤC

NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP	1
LỜI CẢM ƠN.....	i
LỜI CAM ĐOAN.....	ii
MỤC LỤC.....	iii
DANH MỤC HÌNH ẢNH.....	vi
DANH MỤC BẢNG BIỂU	ix
DANH MỤC CÁC TỪ VIẾT TẮT	x
LỜI MỞ ĐẦU	1
CHƯƠNG 1. TỔNG QUAN VỀ ĐƠN VỊ THỰC TẬP VÀ VỊ TRÍ DATA ANALYST	4
1.1. Tổng quan về công ty VNPT – IT3	4
1.1.1. Thông tin chung.....	4
1.1.2. Tâm nhìn và sứ mệnh	5
1.1.3. Giá trị cốt lõi.....	5
1.1.4. Cơ cấu tổ chức.....	5
1.2. Giới thiệu vị trí Data Analyst.....	6
1.2.1. Data Analyst là gì?	6
1.2.2. Các công việc của Data Analyst:.....	6
1.2.3. Các kỹ năng cần thiết của Data Analyst:.....	8
CHƯƠNG 2. CƠ SỞ LÝ THUYẾT	9
2.1. Machine learning.....	9
2.1.1. Thuật toán Logistic Regression	9
2.1.2. Thuật toán Random Forest.....	10
2.1.3. Thuật toán K-Means.....	11
2.1.4. Hệ thống đề xuất.....	12

2.1.5. Các chỉ số đánh giá mô hình.....	14
2.2. Các công cụ sử dụng.....	15
2.2.1. Jupyter Notebook.....	15
2.2.2. Python.....	16
2.2.3. Power BI.....	17
CHƯƠNG 3. PHÂN TÍCH KHÁM PHÁ VÀ XỬ LÝ DỮ LIỆU.....	19
3.1. Giới thiệu dữ liệu.....	19
3.2. Tiền xử lý dữ liệu.....	22
3.3. Thống kê mô tả	27
3.4. Trực quan hóa dữ liệu.....	28
CHƯƠNG 4. PHÂN CỤM DỮ LIỆU, XÂY DỰNG MÔ HÌNH DỰ ĐOÁN SỰ HÀI LÒNG VÀ ĐỀ XUẤT SẢN PHẨM CHO KHÁCH HÀNG	36
4.1. Dự đoán mức độ hài lòng của khách hàng.....	36
4.1.1. Chuẩn bị dữ liệu	36
4.1.2. Mô hình Hồi quy Logistic	40
4.1.3. Mô hình Random Forest.....	43
4.1.4. So sánh các chỉ số của 2 mô hình.....	48
4.1.5. Kết luận	49
4.2. K-Means phân cụm sản phẩm.....	49
4.2.1. Thực hiện phân cụm	49
4.2.2. Kết quả phân cụm	51
4.3. Xây dựng mô hình đề xuất.....	55
4.3.1. Chuẩn bị dữ liệu	55
4.3.2. Ma trận đặc trưng	57
4.3.3. Đưa ra gợi ý	59
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	62

TÀI LIỆU THAM KHẢO.....	64
CHECK LIST CỦA BÁO CÁO.....	66
PHỤ LỤC.....	1

DANH MỤC HÌNH ẢNH

Hình 1.1 Logo công ty VNPT IT	4
Hình 2.1 Mô hình Logistic Regression	9
Hình 2.2 Minh họa thuật toán Random Forest.....	10
Hình 2.3 Minh họa thuật toán K-Means.....	11
Hình 2.4 Minh họa lọc cộng tác	13
Hình 2.5 Minh họa dựa trên nội dung	13
Hình 2.6 Confusion matrix.....	14
Hình 2.7 Logo Jupyter Notebook.....	16
Hình 2.8 Logo Python.....	17
Hình 2.9 Logo Power BI.....	18
Hình 3.1 Bộ dữ liệu đầu vào.....	19
Hình 3.2 Import thư viện và đọc dữ liệu vào	22
Hình 3.3 Kiểm tra thông tin các cột	22
Hình 3.4 Kiểm tra giá trị duy nhất.....	23
Hình 3.5 Kiểm tra giá trị null	24
Hình 3.6 Kiểm tra sự trùng lặp.....	25
Hình 3.7 Xóa cột thiếu hoặc trống dữ liệu	25
Hình 3.8 Kết quả dữ liệu mới	26
Hình 3.9 Các cột sau xử lý	26
Hình 3.10 Các chỉ số thống kê mô tả của dữ liệu.....	27
Hình 3.11 Biểu đồ tổng quan thông tin Laptop tại Thế Giới Di Động.....	28
Hình 3.12 Biểu đồ top sản phẩm theo Rating và Hài lòng.....	29
Hình 3.13 Biểu đồ số lượng bán ra theo các chỉ số liên quan đến Ram và lưu trữ	30
Hình 3.14 Biểu đồ số lượng bán ra theo các chỉ số liên quan đến màn hình....	31
Hình 3.15 Biểu đồ số lượng bán ra theo các chỉ số liên quan đến CPU.....	32

Hình 3.16 Biểu đồ số lượng bán ra theo các chỉ số liên quan đến Công nghệ màn hình và công kết nối	33
Hình 3.17 Biểu đồ số lượng bán ra theo các chỉ số liên quan đến đặc điểm thiết kế của sản phẩm	34
Hình 3.18 Biểu đồ số lượng bán ra theo các chỉ số liên quan đến hệ điều hành và kích thước sản phẩm	35
Hình 4.1 Đọc file dữ liệu	36
Hình 4.2 Chuyển đổi dữ liệu về dạng số	37
Hình 4.3 Biểu đồ tương quan của dữ liệu	38
Hình 4.4 Dataframe mới được chọn từ ma trận tương quan	39
Hình 4.5 Cân bằng dữ liệu	39
Hình 4.6 Chuẩn hóa Min-max	40
Hình 4.7 Chia tập dữ liệu	40
Hình 4.8 Hàm dự đoán	41
Hình 4.9 Confusion matrix mô hình Logistic Regression	41
Hình 4.10 Biểu đồ ROC mô hình Logistic Regression	42
Hình 4.11 Biểu đồ Learning Curve của mô hình Random Forest	44
Hình 4.12 Cây hoàn chỉnh trong Random forest	45
Hình 4.13 Confusion matrix mô hình Random Forest	46
Hình 4.14 Biểu đồ Roc của mô hình Random Forest	47
Hình 4.15 Đọc dữ liệu	49
Hình 4.16 Biểu đồ tìm số cụm k	50
Hình 4.17 Kết quả phân cụm	51
Hình 4.18 Chỉ số của K-means	54
Hình 4.19 Đọc dữ liệu	55
Hình 4.20 Tạo hàm xử lý	56
Hình 4.21 Giá trị cột Tag	57
Hình 4.22 Tạo dataframe mới	57

Hình 4.23 Đọc file stopwords.....	58
Hình 4.24 Dữ liệu sau khi vector hóa.....	58
Hình 4.25 Độ tương đồng dữ liệu.....	59
Hình 4.26 Hàm xử lý đầu vào	60
Hình 4.27 Hàm gợi ý sản phẩm.....	60
Hình 4.28 Kết quả gợi ý.....	61

DANH MỤC BẢNG BIỂU

Bảng 3.1 Mô tả trường dữ liệu	20
Bảng 4.1 So sánh các chỉ số của 2 mô hình	48
Bảng 4.2 Đặc trưng của từng cụm.....	52

DANH MỤC CÁC TỪ VIẾT TẮT

DA	: Data Analyst
AUC	: Area Under The Curve
ROC	: Receiver Operating Characteristics.
SMOTE	: Synthetic Minority Over-sampling
WSS	: Within Sum of Squares
BSS	: Between Sum of Squares
TF – IDF	: Term Frequency – Inverse Document Frequency

LỜI MỞ ĐẦU

1. Lý do chọn đề tài

Trong bối cảnh công nghệ số ngày càng phát triển, đặc biệt là khi làm việc từ xa và học trực tuyến trở thành xu hướng phổ biến đã dẫn tới sự gia tăng về nhu cầu sử dụng laptop. Khách hàng không chỉ quan tâm đến giá cả mà còn chú trọng vào hiệu suất, thiết kế và các tính năng của sản phẩm. Điều này khiến người tiêu dùng dễ bị choáng ngợp và khó đưa ra quyết định mua hàng hơn. Việc cung cấp thông tin chi tiết cùng với các phân tích rõ ràng không chỉ mang lại giá trị thiết thực cho người tiêu dùng mà còn đưa ra các mẫu sản phẩm phù hợp với nhu cầu thực tế của từng cá nhân. Điều này không chỉ tạo ra trải nghiệm tốt hơn cho khách hàng mà còn góp phần tăng doanh số bán hàng trong việc tìm kiếm thông tin, tối ưu hóa trải nghiệm mua sắm.

Thế Giới Di Động nổi bật là một trong những chuỗi bán lẻ lớn nhất tại Việt Nam, có lượng dữ liệu lớn và đa dạng về các sản phẩm laptop. Tuy nhiên, với sự gia tăng cạnh tranh và đa dạng sản phẩm, doanh nghiệp cần nắm bắt nhanh chóng nhu cầu và sở thích của khách hàng để duy trì vị thế dẫn đầu. Việc phân tích dữ liệu sẽ cung cấp thông tin cần thiết giúp doanh nghiệp có cái nhìn sâu sắc hơn về thị trường, khách hàng để điều chỉnh chiến lược marketing và cải thiện dịch vụ bán hàng để phục vụ khách hàng tốt hơn.

Đề tài này không chỉ mang lại lợi ích cho Thế Giới Di Động mà còn góp phần vào sự phát triển bền vững của ngành công nghiệp công nghệ tại Việt Nam. Khi tập trung vào việc cung cấp sản phẩm phù hợp với nhu cầu thực tế, doanh nghiệp không chỉ nâng cao giá trị thương hiệu mà còn tạo ra một thị trường tiêu dùng thông minh hơn.

2. Mục tiêu của đề tài

- Phân tích khám phá, trực quan hóa dữ liệu tìm ra xu hướng đặc điểm của dữ liệu.
- Phân cụm sản phẩm theo các đặc trưng nhất định, áp dụng thuật toán phân cụm K-means để nhóm các sản phẩm laptop có sự tương đồng.

- Xây dựng mô hình để dự đoán sự hài lòng của khách hàng dựa trên các đặc trưng sản phẩm và biến mục tiêu hài lòng. Lựa chọn thuật toán phù hợp với bộ dữ liệu.
- Xây dựng mô hình để xuất các mẫu laptop phù hợp với nhu cầu sở thích của khách hàng dựa trên các đặc trưng về cấu hình và dữ liệu về hành vi khách hàng.
- Đánh giá hiệu suất các mô hình thông qua các chỉ số cụ thể.

3. Phương pháp nghiên cứu

- Nghiên cứu được nhóm thực hiện bằng cách thu thập thông tin từ nguồn dữ liệu trên trang web thegioididong.com. Từ những thông tin, dữ liệu thu thập được, tiến hành xử lý và phân tích dữ liệu, sử dụng các thuật toán phân cụm, đưa ra dự đoán sự hài lòng và phát triển hệ thống để xuất. Ứng dụng các công cụ phân tích dữ liệu như BI, Jupyter Notebook hoặc Python.

4. Đối tượng và phạm vi nghiên cứu

- **Đối tượng nghiên cứu:** Dữ liệu khách hàng tiêu dùng cá nhân mua sắm laptop tại Thế Giới Di Động, dữ liệu bao gồm các thông tin sản phẩm và phản hồi của khách hàng.
- **Phạm vi nghiên cứu:** Tập trung vào các mẫu laptop đang được bán tại Thế Giới Di Động trong khoảng thời gian từ năm 2020 đến 2024. Nghiên cứu sẽ phân tích đặc điểm kỹ thuật của sản phẩm và phản hồi của khách hàng để thực hiện phân cụm sản phẩm, dự đoán mức độ hài lòng và xây dựng hệ thống gợi ý laptop phù hợp.

5. Nội dung của đề tài

Đề tài được tổ chức gồm phần mở đầu, 4 chương nội dung và phần kết luận...

- Mở đầu
- **Chương 1:** Tổng quan về đơn vị thực tập và vị trí Data Analyst
- **Chương 2:** Cơ sở lý thuyết
- **Chương 3:** Phân tích khám phá và xử lý dữ liệu

- **Chương 4:** Phân cụm dữ liệu, xây dựng mô hình dự đoán sự hài lòng và đề xuất sản phẩm cho khách hàng
- Kết luận và hướng phát triển

CHƯƠNG 1. TỔNG QUAN VỀ ĐƠN VỊ THỰC TẬP VÀ VỊ TRÍ DATA ANALYST

1.1. Tổng quan về công ty VNPT – IT3

1.1.1. Thông tin chung

Công ty Công nghệ thông tin VNPT (viết tắt: VNPT-IT) được thành lập ngày 01/03/2018 theo Quyết định số 39/QĐ-VNPT-HĐTV-NL của Chủ tịch Tập đoàn Bưu chính Viễn thông Việt Nam. Công ty được tái tổ chức từ các nhiệm vụ và nguồn lực công nghệ thông tin thuộc Tập đoàn. VNPT-IT tập trung vào nghiên cứu, phát triển và tích hợp các sản phẩm, dịch vụ công nghệ thông tin, không chỉ phục vụ nội bộ Tập đoàn VNPT mà còn mở rộng đến khách hàng bên ngoài, bao gồm cả thị trường quốc tế.



Hình 1.1 Logo công ty VNPT IT

Công ty đặt mục tiêu xây dựng hệ sinh thái tích hợp trọn gói các sản phẩm và dịch vụ công nghệ thông tin và Internet lớn nhất Việt Nam, từ đó mở rộng ra thị trường quốc tế. Để đạt được điều này, VNPT-IT định hướng phát triển dựa trên bốn giá trị cốt lõi: Con người là chìa khóa, khách hàng là trung tâm, sáng tạo không ngừng, và đối tác đáng tin cậy.

Về lĩnh vực kinh doanh VNPT-IT chuyên nghiên cứu, phát triển, sản xuất và kinh doanh các sản phẩm, dịch vụ công nghệ thông tin để cung cấp cho cả nội bộ VNPT và khách hàng bên ngoài. Công ty đầu tư, phát triển và quản lý các hệ thống, nền tảng công nghệ thông tin; vận hành, khai thác hệ thống điều hành sản

xuất kinh doanh, đồng thời đảm bảo an toàn và bảo mật thông tin cho các sản phẩm và dịch vụ công nghệ thông tin mà VNPT cung cấp cho khách hàng. [1]

1.1.2. *Tầm nhìn và sứ mệnh*

Tầm nhìn của VNPT-IT thể hiện khát vọng mạnh mẽ và định hướng rõ ràng trong việc trở thành đơn vị tiên phong trong lĩnh vực công nghệ thông tin. Với khát vọng tiên phong trong chiến lược đầu tư và cung cấp giải pháp, sản phẩm, dịch vụ công nghệ thông tin, VNPT-IT, công ty hàng đầu trong lĩnh vực này tại Việt Nam, cam kết phát triển theo hướng nghiên cứu, gia công và thử nghiệm. Công ty hướng tới cung cấp các sản phẩm đa dạng, uy tín và chất lượng cao, đáp ứng nhu cầu của xã hội và góp phần nâng cao chất lượng cuộc sống của người Việt, đồng thời nâng tầm vị thế của Việt Nam trên thị trường quốc tế. [1]

1.1.3. *Giá trị cốt lõi*

Là công ty công nghệ trọng điểm quốc gia, Công ty Công nghệ thông tin VNPT-IT, thành viên của Tập đoàn Bưu chính Viễn thông Việt Nam, sẽ đóng vai trò chủ lực trong việc triển khai VNPT 4.0 và tham gia mạnh mẽ vào quá trình chuyển đổi nền kinh tế số của Việt Nam. VNPT-IT không chỉ cung cấp giải pháp xây dựng chính quyền số và nền kinh tế số mà còn hỗ trợ doanh nghiệp Việt Nam chuyển đổi thành doanh nghiệp số, cung cấp các nền tảng phát triển công nghệ số trong tương lai. VNPT-IT chính là hoài bão và khát vọng của VNPT trong hành trình khai phá lĩnh vực CNTT.

Với sứ mệnh và trọng trách cao cả, công ty cam kết cung cấp các giải pháp, dịch vụ và sản phẩm thông minh cho xã hội trong mọi lĩnh vực, từ giáo dục, quản lý, chăm sóc sức khỏe, đến ứng dụng và chuyển giao công nghệ, tất cả đều đạt chất lượng cao với thương hiệu và danh tiếng đẳng cấp khu vực và quốc tế. VNPT-IT góp phần quan trọng vào sự nghiệp công nghiệp hóa, hiện đại hóa đất nước trong bối cảnh hội nhập kinh tế toàn cầu. [1]

1.1.4. *Cơ cấu tổ chức*

Về Cơ cấu của VNPT-IT bao gồm Ban Tổng giám đốc, Văn phòng các Ban chức năng, các Trung tâm trực thuộc và các Trung tâm tại Hà Nội, Hồ Chí

Minh, Đà Nẵng, Hải Phòng và Tiền Giang. Các Trung tâm này là đơn vị hạch toán phụ thuộc của Công ty.

Là một trong những đơn vị chủ chốt của Tập đoàn Bưu chính Viễn thông Việt Nam, VNPT-IT không ngừng phấn đấu nâng cao chất lượng sản phẩm và dịch vụ về mọi mặt để trở thành thương hiệu uy tín trong lĩnh vực công nghệ thông tin, góp phần đưa VNPT đạt mục tiêu trở thành Tập đoàn Viễn thông - CNTT hàng đầu quốc gia và giữ vai trò chủ đạo trong lĩnh vực Viễn thông và CNTT tại Việt Nam. [1]

1.2. Giới thiệu vị trí Data Analyst

1.2.1. Data Analyst là gì?

Data Analyst (DA) – chuyên viên phân tích dữ liệu, là công việc được hoạt động gắn liền với khoa học dữ liệu, nhiệm vụ chính của họ là thu thập, chọn lọc, xử lý và phân tích chuyên sâu dữ liệu thu thập được để phát hiện ra vấn đề, hiện trạng cần giải quyết. Tùy theo mỗi ngành nghề khác nhau, mà dữ liệu Data Analyst xử lý sẽ khác nhau (có thể là số, hình ảnh, hay các dạng data khác) và kết luận được đưa ra sẽ còn dựa trên đặc thù của ngành nghề. [2]

1.2.2. Các công việc của Data Analyst:

Quá trình phân tích dữ liệu của Data Analyst thường bao gồm các bước cơ bản sau:

Xác định mục tiêu: Trước khi bắt đầu phân tích, Data Analyst cần hiểu rõ mục tiêu và câu hỏi cụ thể mà phân tích dữ liệu phải trả lời. Điều này bao gồm việc xác định vấn đề cần giải quyết hoặc các mục tiêu kinh doanh cần đạt được.

Thu thập dữ liệu: Dữ liệu có thể được thu thập từ nhiều nguồn khác nhau, bao gồm cơ sở dữ liệu nội bộ, bảng khảo sát, hệ thống giao dịch, hoặc các nguồn dữ liệu bên ngoài. Data Analyst cần phải đảm bảo rằng dữ liệu thu thập được là đầy đủ và phù hợp với mục tiêu phân tích.

Làm sạch dữ liệu: Dữ liệu thô thường chứa lỗi, thiếu sót, hoặc thông tin không nhất quán. Bước này bao gồm việc xử lý các giá trị thiếu, loại bỏ dữ liệu trùng lặp, và chuẩn hóa dữ liệu để đảm bảo tính chính xác và đồng nhất.

Khám phá dữ liệu: Trong bước này, Data Analyst sẽ thực hiện các phân tích mô tả để hiểu rõ hơn về dữ liệu, bao gồm việc kiểm tra các mô hình phân phối, phát hiện các giá trị ngoại lệ, và đánh giá các mối quan hệ giữa các biến. Các công cụ trực quan hóa dữ liệu như biểu đồ, đồ thị và bảng tổng hợp thường được sử dụng trong bước này.

Phân tích dữ liệu: Dựa trên mục tiêu đã xác định, Data Analyst áp dụng các phương pháp phân tích thống kê và kỹ thuật phân tích nâng cao như hồi quy, phân tích cụm (clustering), hoặc phân tích chuỗi thời gian. Bước này nhằm phát hiện các xu hướng, mẫu và mối quan hệ trong dữ liệu.

Trực quan hóa dữ liệu: Dữ liệu phân tích được chuyển thành các biểu đồ, đồ thị, và bảng điều khiển dễ hiểu để giúp các bên liên quan nhanh chóng nắm bắt thông tin. Trực quan hóa dữ liệu giúp trình bày kết quả phân tích một cách rõ ràng và trực quan.

Điển giải kết quả: Data Analyst diễn giải các kết quả phân tích để đưa ra các thông tin có giá trị và khuyến nghị cụ thể. Điều này bao gồm việc kết nối các phát hiện với mục tiêu kinh doanh và giải thích ý nghĩa của các xu hướng hoặc mẫu phát hiện được.

Trình bày và báo cáo: Data Analyst chuẩn bị báo cáo và trình bày kết quả phân tích cho các bên liên quan. Báo cáo thường bao gồm các phát hiện chính, khuyến nghị, và các hành động được đề xuất dựa trên phân tích.

Đánh giá và cải thiện: Sau khi báo cáo được trình bày, Data Analyst có thể thu thập phản hồi từ các bên liên quan và đánh giá hiệu quả của các khuyến nghị. Dựa trên phản hồi, họ có thể thực hiện các cải thiện hoặc phân tích thêm để đạt được kết quả tốt hơn. [2]

1.2.3. Các kỹ năng cần thiết của Data Analyst:

Làm sạch và phân tích dữ liệu: Đây là một trong những kỹ năng cần thiết trong danh sách các kỹ năng của Data Analyst. Về cơ bản, phân tích dữ liệu bao gồm việc biến đổi một câu hỏi hoặc nhu cầu kinh doanh thành một câu hỏi dữ liệu cụ thể. Sau đó, bạn sẽ phải thực hiện quá trình biến đổi và phân tích dữ liệu để tìm ra câu trả lời cho câu hỏi đó.

Am hiểu kiến thức thống kê: Nếu bạn muốn trở thành một Data Analyst, thì việc nắm vững các khái niệm và công cụ sử dụng trong phân tích dữ liệu là điều vô cùng quan trọng và không thể bỏ qua. Cần nắm vững các mô hình phân tích dữ liệu như Logistic Regression (Hồi quy Logistic), Linear Regression (Hồi quy tuyến tính), và nhiều mô hình khác thuộc loại Regression models. Đây là những công cụ quan trọng để dự đoán và hiểu mối quan hệ giữa các biến.

Sử dụng các công cụ phân tích dữ liệu: Cần phải thành thạo trong việc sử dụng các công cụ phân tích dữ liệu chính như Python, R và SQL. Python và R là hai ngôn ngữ lập trình phổ biến để xử lý dữ liệu, thực hiện phân tích thống kê và xây dựng mô hình dự đoán. SQL (Structured Query Language) là ngôn ngữ quản lý cơ sở dữ liệu, giúp trích xuất, cập nhật và tương tác với cơ sở dữ liệu một cách hiệu quả. Hiểu biết sâu rộng về các công cụ này là một trong những nền tảng quan trọng để trở thành một Data Analyst giỏi. [3]

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

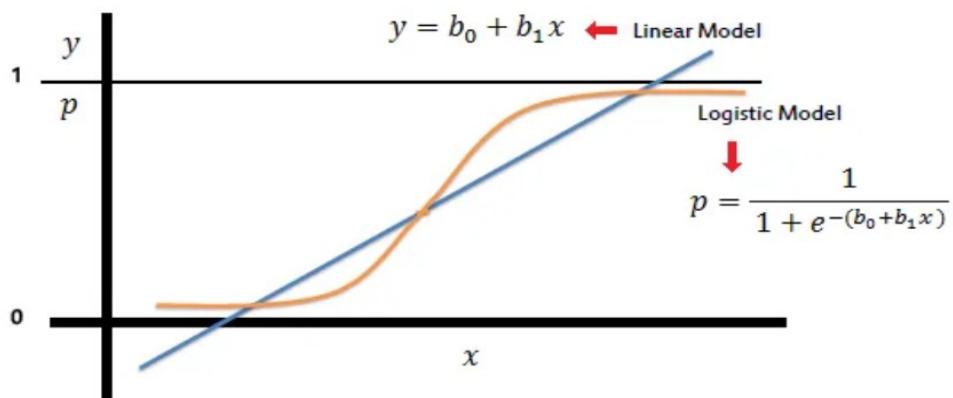
2.1. Machine learning

2.1.1. Thuật toán Logistic Regression

2.1.1.1. Khái niệm

Hồi quy Logistic là một mô hình thống kê được sử dụng để phân loại nhị phân, tức dự đoán một đối tượng thuộc vào một trong hai nhóm. Hồi quy Logistic làm việc dựa trên nguyên tắc của hàm sigmoid – một hàm phi tuyến tự chuyển đầu vào của nó thành xác suất thuộc về một trong hai lớp nhị phân. [4]

2.1.1.2. Mô hình Logistic Regression



Hình 2.1 Mô hình Logistic Regression

Hàm Sigmoid nhận đầu vào là một giá trị z bất kỳ, và trả về đầu ra là một giá trị xác suất nằm trong khoảng [0,1]. Khi áp dụng vào mô hình Hồi quy Logistic với đầu vào là ma trận dữ liệu X và trọng số w, ta có $z=Xw$.

Việc huấn luyện của mô hình là tìm ra bộ trọng số w sao cho đầu ra dự đoán của hàm Sigmoid gần với kết quả thực tế nhất. Để làm được điều này, ta sử dụng hàm mất mát (Loss Function) để đánh giá hiệu năng của mô hình. Mô hình càng tốt khi hàm mất mát càng nhỏ.

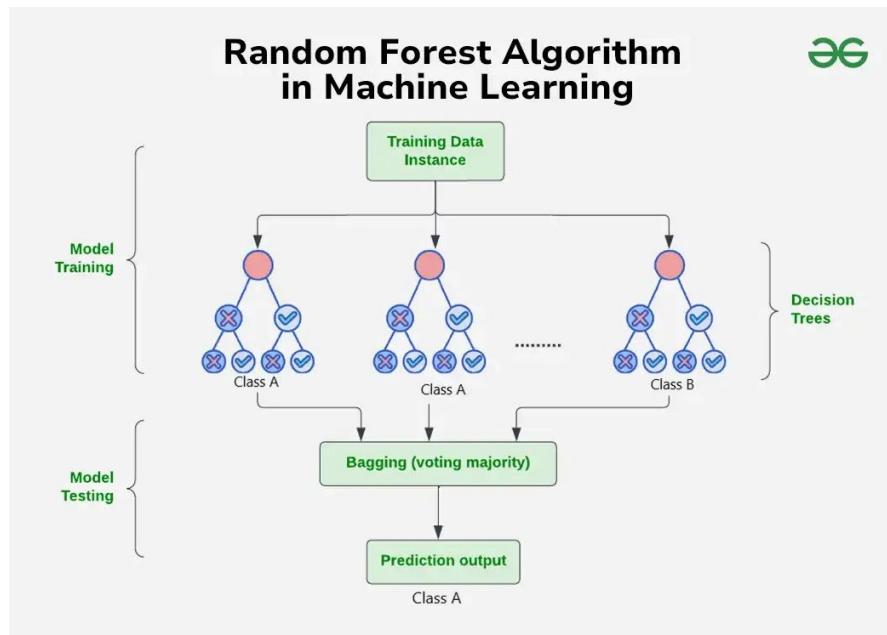
Trong quá trình huấn luyện, chúng ta tìm cách cập nhật bộ trọng số w sao cho giá trị hàm mất mát Cross-Entropy đạt giá trị nhỏ nhất, dẫn đến một mô hình dự đoán tốt nhất.

Để tìm giá trị tối ưu cho bộ trọng số w , chúng ta có thể sử dụng kỹ thuật Gradient Descent. Tại mỗi bước lặp, chúng ta cập nhật w theo phương trình ứng với đạo hàm của hàm mất mát $L(w)$ theo w . [4]

2.1.2. Thuật toán Random Forest

2.1.2.1. Khái niệm

Random Forest là một thuật toán học máy thuộc nhóm học có giám sát (supervised learning), dựa trên kỹ thuật ensemble learning (học tổ hợp), kết hợp nhiều mô hình con để cải thiện độ chính xác và độ ổn định. Cụ thể, Random Forest xây dựng nhiều cây quyết định (decision trees) trên các tập con ngẫu nhiên của dữ liệu huấn luyện và tổng hợp kết quả dự đoán từ các cây đó thông qua biểu quyết đa số (với bài toán phân loại) hoặc trung bình (với bài toán hồi quy). [5]



Hình 2.2 Minh họa thuật toán Random Forest

2.1.2.2. Cách thức hoạt động

Random Forest hoạt động dựa trên sự kết hợp giữa hai kỹ thuật chính:

Bootstrap Aggregating (Bagging): Kỹ thuật lấy mẫu ngẫu nhiên với hoàn lại từ tập huấn luyện ban đầu để tạo ra nhiều tập con khác nhau. Mỗi cây quyết định sẽ được huấn luyện độc lập trên một trong các tập con đó.

Lựa chọn đặc trưng ngẫu nhiên: Tại mỗi nút của cây, thay vì xem xét toàn bộ các đặc trưng để tìm điểm chia tối ưu, chỉ một tập con ngẫu nhiên các đặc trưng được chọn để quyết định chia. Điều này giúp các cây trở nên khác biệt và giảm sự tương quan giữa chúng.

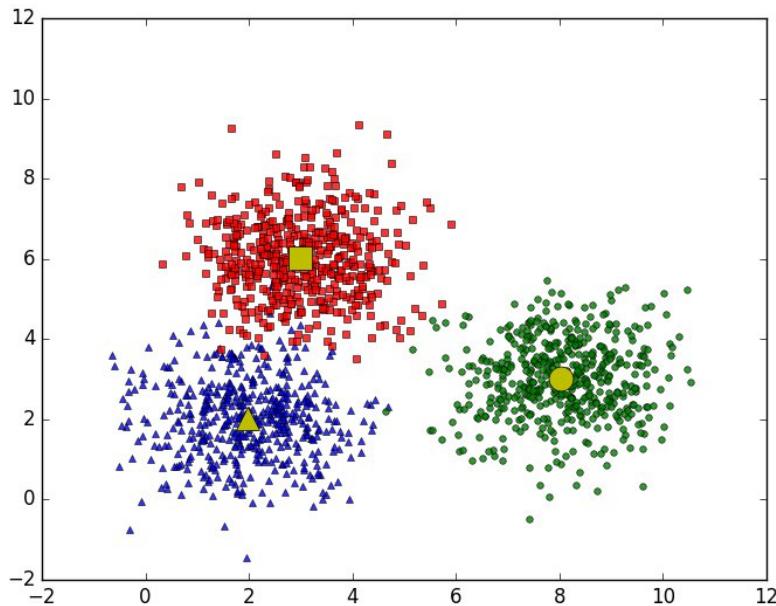
Quy trình xây dựng mô hình Random Forest bao gồm:

- Bước 1. Tạo n tập con ngẫu nhiên (có hoàn lại) từ tập dữ liệu gốc.
- Bước 2. Huấn luyện một cây quyết định trên mỗi tập con, với quá trình lựa chọn đặc trưng ngẫu nhiên tại mỗi điểm chia.
- Bước 3. Tổng hợp dự đoán từ tất cả các cây: Sử dụng biểu quyết đa số với phân loại. Tính trung bình giá trị dự đoán với hồi quy. [5]

2.1.3. Thuật toán K-Means

2.1.3.1. Khái niệm

K-means Clustering là một thuật toán phân cụm cỗ điển và phổ biến, được sử dụng rộng rãi trong nhiều lĩnh vực của khoa học dữ liệu như phân loại khách hàng, xử lý ảnh, giảm chiều dữ liệu, và phát hiện outlier. [6]



Hình 2.3 Minh họa thuật toán K-Means

2.1.3.2. Cách thức hoạt động

Bước 1: Khởi tạo

Thuật toán K-means Clustering bắt đầu bằng việc chọn k số điểm dữ liệu ngẫu nhiên (cụm) trong tập dữ liệu. K là số cụm cần phân loại, được lựa chọn trước khi thiết lập thuật toán.

Bước 2: Gán nhãn cho từng điểm dữ liệu

Sau khi có k cụm ban đầu, chúng ta sẽ tính toán khoảng cách giữa từng điểm dữ liệu với k cụm này và gán điểm dữ liệu đó vào cụm gần nó nhất. Khoảng cách giữa hai điểm dữ liệu thường được tính bằng cách Euclidean, công

$$\text{thức như sau: Euclidean} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Bước 3: Cập nhật tâm của cụm

Sau khi đã gán nhãn cho tất cả các điểm dữ liệu, chúng ta cần xác định lại tâm của các cụm để cải thiện hiệu quả của thuật toán. Tâm mới của cụm sẽ được xác định bằng cách tính trung bình vị trí của tất cả các điểm dữ liệu thuộc cụm đó.

Bước 4: Kiểm tra điều kiện dừng

Quá trình gán nhãn và cập nhật tâm cụm sẽ được lặp lại cho đến khi tâm cụm không thay đổi sau mỗi vòng lặp (hay chênh lệch đủ nhỏ) hoặc đạt số lần lặp tối đa. [6]

2.1.4. Hệ thống đề xuất

2.1.4.1. Khái niệm

Hệ thống đề xuất (recommender system) là một dạng công cụ lọc thông tin (information filtering) cho phép suy diễn, dự đoán các sản phẩm, dịch vụ, nội dung mà người dùng có thể quan tâm dựa trên những thông tin thu thập được về người dùng, về các sản phẩm, dịch vụ, về các hoạt động, tương tác cũng như đánh giá của người dùng đối với các sản phẩm, dịch vụ trong quá khứ. [7]

2.1.4.2. Các cách tiếp cận

Collaborative filtering - CF (lọc cộng tác)

Lọc cộng tác thực hiện tư vấn (gợi ý) các sản phẩm, dịch vụ, nội dung cho một người dùng nào đó dựa trên mối quan tâm, sở thích (preferences) của những người dùng tương tự đối với các sản phẩm, dịch vụ, nội dung đó.

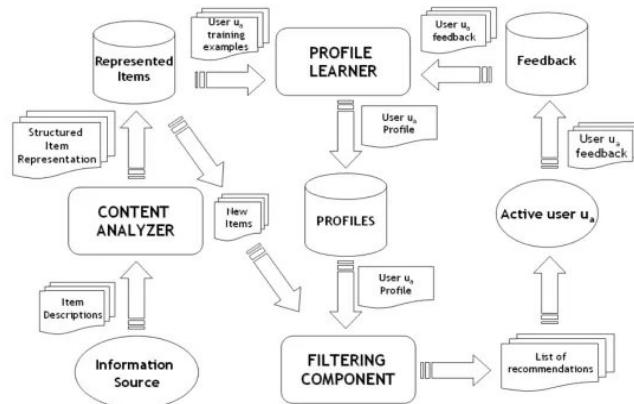


Hình 2.4 Minh họa lọc cộng tác

Content-based (dựa trên nội dung)

Các phương pháp lọc cộng tác chỉ dựa trên tương tác của người dùng, thì các phương pháp dựa trên nội dung sử dụng thông tin bổ sung (features) về người dùng và sản phẩm.

Sau đó, ý tưởng của phương pháp dựa trên nội dung là cố gắng xây dựng một mô hình dựa trên các thông tin bổ sung.



Hình 2.5 Minh họa dựa trên nội dung

Hybrid approach (phương pháp kết hợp)

Phương pháp này là sự kết hợp các phương pháp tiếp cận dựa trên nội dung và lọc cộng tác, chúng cho kết quả tốt trong nhiều trường hợp và do đó,

được sử dụng trong nhiều hệ thống đề xuất quy mô lớn hiện nay. Sự kết hợp được thực hiện trong các phương pháp lai có thể chủ yếu có hai dạng:

Huấn luyện hai mô hình một cách độc lập (một mô hình lọc cộng tác và một mô hình dựa trên nội dung) và kết hợp các đề xuất của chúng.

Trực tiếp xây dựng một mô hình để thống nhất cả hai cách tiếp cận bằng cách sử dụng làm thông tin trước khi nhập (về người dùng và / hoặc vật phẩm) cũng như thông tin tương tác của cộng tác trực tuyến. [7]

2.1.5. Các chỉ số đánh giá mô hình

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Hình 2.6 Confusion matrix

Để đánh giá hiệu suất mô hình một cách toàn diện hơn, Confusion matrix sử dụng các chỉ số đánh giá sau:

Accuracy (Độ chính xác): Tỷ lệ dự đoán đúng trên tổng số mẫu dữ liệu.

$$\text{Accuracy} = \frac{\text{Số lượng dự đoán đúng}}{\text{Tổng số mẫu dữ liệu}}$$

Precision (Độ chính xác của lớp dương): Tỷ lệ khách hàng được dự đoán là sẽ rời bỏ dịch vụ thực sự rời bỏ dịch vụ.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall (Độ phủ của lớp dương): Tỷ lệ khách hàng thực sự rời bỏ dịch vụ được dự đoán là sẽ rời bỏ dịch vụ.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1-score: Trung bình điều hòa giữa Precision và Recall, là một chỉ số tổng hợp giữa hai chỉ số trên.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

AUC (Area Under the ROC Curve): Diện tích dưới đường cong ROC. Một chỉ số tổng hợp cho biết mức độ mà mô hình phân loại phân biệt được giữa các lớp. [8]

2.2. Các công cụ sử dụng

2.2.1. Jupyter Notebook

Jupyter Notebook là một nền tảng tính toán khoa học mã nguồn mở, bạn có thể sử dụng để tạo và chia sẻ các tài liệu có chứa code trực tiếp, chương trình, trực quan hóa dữ liệu và văn bản tường thuật.

Jupyter Notebook được coi là môi trường điện toán tương tác đa ngôn ngữ, hỗ trợ hơn 40 ngôn ngữ lập trình cho người dùng. Với Jupyter Notebook, người dùng có thể đưa dữ liệu, code, hình ảnh, công thức, video,... vào trong cùng một file, giúp cho việc trình bày trở nên dễ dàng hơn.



Hình 2.7 Logo Jupyter Notebook

Jupyter cho phép người dùng xem kết quả của code in-line (mã inline) mà không cần phụ thuộc vào các phần khác của code.

Từng ô tự duy trì trạng thái hoạt động sẽ hơi khó, nhưng với Jupyter, công việc này sẽ được thực hiện tự động. Vì Jupyter lưu trữ kết quả hoạt động của mọi ô đang chạy, cho dù là code đang đào tạo mô hình machine learning hay code đang tải xuống gigabyte dữ liệu từ một máy chủ từ xa.

Jupyter Notebook ở định dạng JSON, vì thế nó được biết đến là một nền tảng độc lập cũng như độc lập về ngôn ngữ.

hỗ trợ trực quan hóa dữ liệu và hiển thị thêm một số đồ họa và biểu đồ. Ngoài ra, Jupyter còn cho phép người dùng cùng chia sẻ code và bộ dữ liệu hoặc thay đổi tương tác với nhau.

Cung cấp cho người dùng giao diện chuẩn nhằm khám phá sự tương tác trực tiếp với code và với dữ liệu. Người dùng có thể chỉnh sửa và chạy code, làm cho code của Jupyter non-static.

Jupyter giúp người dùng dễ dàng giải thích từng dòng code của họ với các phản hồi được đính kèm. Dù trong code đã có đầy đủ các chức năng nhưng người dùng vẫn có thể tăng thêm sự tương tác bằng các lời giải thích. [9]

2.2.2. Python

Python là một ngôn ngữ lập trình thông dịch, dễ đọc và dễ hiểu. Nền tảng nổi tiếng với cú pháp đơn giản và được sử dụng rộng rãi trong nhiều lĩnh vực

khác nhau. Điện hình như phát triển web, phân tích dữ liệu, trí tuệ nhân tạo và nhiều ứng dụng khác.



Hình 2.8 Logo Python

Python có cú pháp linh hoạt và cấu trúc dữ liệu mạnh mẽ, công nghệ được hỗ trợ bởi một cộng đồng lớn. Điều này đã mang đến các thư viện và framework phong phú mà người dùng có thể sử dụng để xây dựng các ứng dụng phức tạp. Python cũng là một trong những ngôn ngữ phổ biến cho người mới học lập trình nhờ vào tính linh hoạt của nó. [10]

Thư viện python là một tập hợp các mã thường xuyên được sử dụng mà các nhà phát triển có thể bao gồm trong những chương trình Python của họ để không phải lập trình từ đầu. Theo mặc định, Python đi kèm với Thư viện chuẩn, chứa rất nhiều các hàm có thể tái sử dụng. Ngoài ra, hơn 137.000 thư viện Python có sẵn cho các ứng dụng khác nhau, bao gồm phát triển web, khoa học dữ liệu và máy học (ML). [11]

2.2.3. **Power BI**

Power BI là một công cụ phân tích dữ liệu do Microsoft phát triển, giúp người dùng kết nối với nhiều nguồn dữ liệu khác nhau, truy vấn, lọc, chuyển đổi và trực quan hóa dữ liệu để tạo báo cáo và bảng điều khiển thông tin. Power BI cung cấp nhiều tính năng mạnh mẽ, dễ sử dụng, phù hợp cho người dùng ở mọi cấp độ kỹ thuật, từ người mới bắt đầu đến chuyên gia phân tích dữ liệu. [12]



Hình 2.9 Logo Power BI

Kết nối với nguồn dữ liệu đa dạng, từ những tệp dữ liệu phẳng (flat file), file excel, đến các cơ sở dữ liệu như SQL Server, MySQL, Access,... hay dữ liệu trên đám mây (cloud). Ngoài ra, Power BI cũng có thể kết nối được với các nguồn dữ liệu đặc biệt như SharePoint, dữ liệu Web, PDF,... hay kết nối trực tiếp đến các hệ thống ERP, SAP,...

Làm sạch, biến đổi và chuẩn hóa dữ liệu đầu vào với Power Query Trước khi đưa vào báo cáo phân tích. Power BI Query có khả năng thực hiện nhiều thao tác phức tạp như lọc dữ liệu, xóa thông tin trùng lặp,...và trả về dữ liệu với độ chính xác cao, đầy đủ và phù hợp để phân tích.

Xây dựng các mô hình dữ liệu làm gọn nhẹ dữ liệu hơn so với việc phải kết hợp các dữ liệu thành một file duy nhất. Giống với các lược đồ dữ liệu trong các cơ sở dữ liệu, Power BI có khả năng kết nối các bảng để tạo thành một mô hình có tính liên kết giúp việc truy xuất dữ liệu được hiệu quả.

Trực quan hóa dữ liệu bằng các biểu đồ, dashboard mang tính tương tác một cách linh hoạt, từ đó có những phân tích theo nhiều chiều, từ tổng quan đến chi tiết để có các phát hiện, kết luận và đưa ra quyết định kinh doanh.

Hỗ trợ các hàm biểu thức tính toán phân tích dữ liệu DAX (data analysis expression) cực kì mạnh mẽ để xây dựng các công thức phân tích và hiển thị dữ liệu. [13]

CHƯƠNG 3. PHÂN TÍCH KHÁM PHÁ VÀ XỬ LÝ DỮ LIỆU

3.1. Giới thiệu dữ liệu

Bộ dữ liệu được lấy từ trang web của Thế Giới Di Động

Link	Thương hiệu	Tên sản phẩm	Thông tin	Giá	Tên khái niệm	Rating	Hài lòng	Đánh giá	RAM	Loại RAM	Tốc độ	Bu	Hỗ trợ RAM	Ổ cứng	Công nghệ	Số nhân	
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	12490000	Phạm Đức	1	1	Máy để bàn	16 GB	LPDDR4X (4266 MHz)	Không hỗ	512 GB	SSD	AMD Ryzen	8		
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	12490000	Lê Văn Đồ	1	1	Mua sắm	16 GB	LPDDR4X (4266 MHz)	Không hỗ	512 GB	SSD	AMD Ryzen	8		
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	12490000	Long Hoar	5	1	trong tầm	16 GB	LPDDR4X (4266 MHz)	Không hỗ	512 GB	SSD	AMD Ryzen	8		
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	12490000	Hải Đăng	3	1	không có	16 GB	LPDDR4X (4266 MHz)	Không hỗ	512 GB	SSD	AMD Ryzen	8		
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	12490000	Nguyễn H	4	1	Mới mua	16 GB	LPDDR4X (4266 MHz)	Không hỗ	512 GB	SSD	AMD Ryzen	8		
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	12490000	Nguyễn Th	2	0	Tôi mới m	8 GB	DDR4 2 kh	Từ 2400 M	32 GB		512 GB	SSD	Intel Core	10
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	12490000	Tuyết	5	0	máy oki	8 GB	DDR4 2 kh	Từ 2400 M	32 GB		512 GB	SSD	Intel Core	10
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	12490000	Lương Thế	3	0	Máy tạm	8 GB	DDR4 2 kh	Từ 2400 M	32 GB		512 GB	SSD	Intel Core	10
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	12490000	Công Ty Tr	5	1	Sản phẩm	8 GB	DDR4 2 kh	Từ 2400 M	32 GB		512 GB	SSD	Intel Core	10
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	12490000	Thương	5	1	Tốt	8 GB	DDR4 2 kh	Từ 2400 M	32 GB		512 GB	SSD	Intel Core	10
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	12490000	Anh Long	5	0	Rất tốt	8 GB	DDR4 2 kh	Từ 2400 M	32 GB		512 GB	SSD	Intel Core	10
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	12490000	VO PHAN	4	0	Tốt	8 GB	DDR4 2 kh	Từ 2400 M	32 GB		512 GB	SSD	Intel Core	10
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	12490000	Ngô Mỹ	4	0	Sé giới thi	8 GB	DDR4 2 kh	Từ 2400 M	32 GB		512 GB	SSD	Intel Core	10
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	16990000	Đinh Võ Q	3	1	Laptop m	16 GB	DDR4 2 kh	3200 MHz	32 GB		512 GB	SSD	Intel Core	8
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	16990000	Nguyễn Th	3	1	Máy sài	16 GB	DDR4 2 kh	3200 MHz	32 GB		512 GB	SSD	Intel Core	8
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	16990000	Huỳnh Vũ	3	1	Máy dẹp	16 GB	DDR4 2 kh	3200 MHz	32 GB		512 GB	SSD	Intel Core	8
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	16990000	Nguyễn Th	3	1	mình v	16 GB	DDR4 2 kh	3200 MHz	32 GB		512 GB	SSD	Intel Core	8
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	16990000	Hoàng Văn	3	0	Máy dùng	16 GB	DDR4 2 kh	3200 MHz	32 GB		512 GB	SSD	Intel Core	8
https://www.Acer.com	Acer	Laptop Acer	Bộ sưu tập	16990000	Lê Văn Ph	4	0	máy rất	16 GB	DDR4 2 kh	3200 MHz	32 GB		512 GB	SSD	Intel Core	8

Số lượng	Tốc độ CPU	Tốc độ tối đa NPU	Hiệu năng	Màn hình	Độ phân giải	Tần số quay	Độ phủ m	Công nghệ	Màn hình	Độ sáng	S Card	màn	Công nghệ	DTS	Công giao
16	1.8GHz	Turbo Boost 4.3 GHz	14"	WUXGA	60Hz	45% NTSC	Tấm nền IPS				Card tích	I Acer Purified Voice	Jack tai ng		
16	1.8GHz	Turbo Boost 4.3 GHz	14"	WUXGA	60Hz	45% NTSC	Tấm nền IPS				Card tích	I Acer Purified Voice	Jack tai ng		
16	1.8GHz	Turbo Boost 4.3 GHz	14"	WUXGA	60Hz	45% NTSC	Tấm nền IPS				Card tích	I Acer Purified Voice	Jack tai ng		
16	1.8GHz	Turbo Boost 4.3 GHz	14"	WUXGA	60Hz	45% NTSC	Tấm nền IPS				Card tích	I Acer Purified Voice	Jack tai ng		
16	1.8GHz	Turbo Boost 4.3 GHz	14"	WUXGA	60Hz	45% NTSC	Tấm nền IPS				Card tích	I Acer Purified Voice	Jack tai ng		
12	1.3GHz	Turbo Boost 4.4 GHz	15.6"	Full HD (160Hz)				TFT			Card tích	I Stereo speakers	LAN (RJ45)		
12	1.3GHz	Turbo Boost 4.4 GHz	15.6"	Full HD (160Hz)				TFT			Card tích	I Stereo speakers	LAN (RJ45)		
12	1.3GHz	Turbo Boost 4.4 GHz	15.6"	Full HD (160Hz)				TFT			Card tích	I Stereo speakers	LAN (RJ45)		
12	1.3GHz	Turbo Boost 4.4 GHz	15.6"	Full HD (160Hz)				TFT			Card tích	I Stereo speakers	LAN (RJ45)		
12	1.3GHz	Turbo Boost 4.4 GHz	15.6"	Full HD (160Hz)				TFT			Card tích	I Stereo speakers	LAN (RJ45)		
12	1.3GHz	Turbo Boost 4.4 GHz	15.6"	Full HD (160Hz)				TFT			Card tích	I Stereo speakers	LAN (RJ45)		
12	2.1GHz	Turbo Boost 4.6 GHz	15.6"	Full HD (160Hz)				Tấm nền IPS			Card rời	- Acer Purified Voice	LAN (RJ45)		
12	2.1GHz	Turbo Boost 4.6 GHz	15.6"	Full HD (160Hz)				Tấm nền IPS			Card rời	- Acer Purified Voice	LAN (RJ45)		
12	2.1GHz	Turbo Boost 4.6 GHz	15.6"	Full HD (160Hz)				Tấm nền IPS			Card rời	- Acer Purified Voice	LAN (RJ45)		
12	2.1GHz	Turbo Boost 4.6 GHz	15.6"	Full HD (160Hz)				Tấm nền IPS			Card rời	- Acer Purified Voice	LAN (RJ45)		
12	2.1GHz	Turbo Boost 4.6 GHz	15.6"	Full HD (160Hz)				Tấm nền IPS			Card rời	- Acer Purified Voice	LAN (RJ45)		
12	2.1GHz	Turbo Boost 4.6 GHz	15.6"	Full HD (160Hz)				Tấm nền IPS			Card rời	- Acer Purified Voice	LAN (RJ45)		

Kết nối kh	Đèn bàn	p	Tính năng	Tản nhiệt	Khe đọc t	Kích thước	Chất liệu	Thông tin	Hệ điều h	Thời di	đêm	Dài(mm)	Rộng(mm)	Dày(mm)	Trọng Lượ
Wi-Fi 6 (802.11ax)	Không có đèn				Dài 318.2	Vỏ nhựa	3-cell, 50V	Windows	2024	318.2	225.5	17.8	1.4		
Wi-Fi 6 (802.11ax)	Không có đèn				Dài 318.2	Vỏ nhựa	3-cell, 50V	Windows	2024	318.2	225.5	17.8	1.4		
Wi-Fi 6 (802.11ax)	Không có đèn				Dài 318.2	Vỏ nhựa	3-cell, 50V	Windows	2024	318.2	225.5	17.8	1.4		
Wi-Fi 6 (802.11ax)	Không có đèn				Dài 318.2	Vỏ nhựa	3-cell, 50V	Windows	2024	318.2	225.5	17.8	1.4		
Bluetooth 5.2	Không có đèn				Dài 362.9	Vỏ nhựa	3-cell, 40V	Windows	2022	362.9	241.26	19.9	1.7		
Bluetooth 5.2	Không có đèn				Dài 362.9	Vỏ nhựa	3-cell, 40V	Windows	2022	362.9	241.26	19.9	1.7		
Bluetooth 5.2	Không có đèn				Dài 362.9	Vỏ nhựa	3-cell, 40V	Windows	2022	362.9	241.26	19.9	1.7		
Bluetooth 5.2	Không có đèn				Dài 362.9	Vỏ nhựa	3-cell, 40V	Windows	2022	362.9	241.26	19.9	1.7		
Bluetooth 5.2	Không có đèn				Dài 362.9	Vỏ nhựa	3-cell, 40V	Windows	2022	362.9	241.26	19.9	1.7		
Wi-Fi 6E (6 Full HD)	Đơn sắc - Bảo mật vân tay				Dài 361 m	Vỏ nhựa	3-cell, 50V	Windows	2023	361	237	17.9	1.7		
Wi-Fi 6E (6 Full HD)	Đơn sắc - Bảo mật vân tay				Dài 361 m	Vỏ nhựa	3-cell, 50V	Windows	2023	361	237	17.9	1.7		
Wi-Fi 6E (6 Full HD)	Đơn sắc - Bảo mật vân tay				Dài 361 m	Vỏ nhựa	3-cell, 50V	Windows	2023	361	237	17.9	1.7		
Wi-Fi 6E (6 Full HD)	Đơn sắc - Bảo mật vân tay				Dài 361 m	Vỏ nhựa	3-cell, 50V	Windows	2023	361	237	17.9	1.7		
Wi-Fi 6E (6 Full HD)	Đơn sắc - Bảo mật vân tay				Dài 361 m	Vỏ nhựa	3-cell, 50V	Windows	2023	361	237	17.9	1.7		

Hình 3.1 Bộ dữ liệu đầu vào

Tập dữ liệu gồm 47 trường và 1027 bản ghi

Bảng 3.1 Mô tả trường dữ liệu

Tên trường	Mô tả	Kiểu dữ liệu
Link	Đường dẫn đến trang sản phẩm	object
Thương hiệu	Tên thương hiệu của laptop	object
Tên sp	Tên đầy đủ của laptop	object
Thông tin	Mô tả tổng quát về laptop	object
Giá	Giá của laptop	float64
Tên khách hàng	Tên hoặc username của người đánh giá	object
Rating	Điểm đánh giá	float64
Hài lòng	Mức độ hài lòng của khách hàng	float64
Đánh giá	Nhận xét chi tiết của khách hàng	object
RAM	Dung lượng RAM	object
Loại RAM	Kiểu RAM	object
Tốc độ Bus RAM	Tốc độ RAM	object
Hỗ trợ RAM tối đa	Dung lượng RAM tối đa mà laptop hỗ trợ	object
Ổ cứng	Loại và dung lượng ổ cứng	object
Công nghệ CPU	Dòng CPU	object
Số nhân	Số nhân của CPU	int64
Số luồng	Số luồng xử lý của CPU	object
Tốc độ CPU	Xung nhịp cơ bản	object
Tốc độ tối đa	Xung nhịp tối đa khi Turbo Boost	object
NPU	Bộ xử lý AI	object
Hiệu năng xử lý AI (TOPS)	Công suất xử lý AI	object
Màn hình	Kích thước màn hình	object

Độ phân giải	Độ phân giải màn hình	object
Tần số quét	Tốc độ làm tươi màn hình	object
Độ phủ màu	Độ bao phủ màu sắc	object
Công nghệ màn hình	Loại màn hình	object
Màn hình cảm ứng	Laptop có hỗ trợ cảm ứng không	object
Độ sáng SDR	Độ sáng màn hình khi không bật HDR	object
Card màn hình	Bộ xử lý đồ họa	object
Công nghệ âm thanh	Công nghệ âm thanh tích hợp	object
DTS	Hỗ trợ âm thanh DTS hay không	object
Cổng giao tiếp	Các cổng kết nối	object
Kết nối không dây	Chuẩn kết nối Wi-Fi, Bluetooth	object
Webcam	Độ phân giải camera trước	object
Đèn bàn phím	Bàn phím có đèn nền không	object
Tính năng khác	Các tính năng bổ sung	object
Tản nhiệt	Công nghệ làm mát	object
Khe đọc thẻ nhớ	Khe đọc thẻ SD	object
Kích thước	Kích thước tổng thể	object
Chất liệu	Vật liệu vỏ máy	object
Thông tin Pin	Dung lượng pin	object
Hệ điều hành	Hệ điều hành cài sẵn	object
Thời điểm ra mắt	Năm hoặc tháng phát hành laptop	object
Dài(mm)	Chiều dài máy	float64
Rộng(mm)	Chiều rộng máy	float64
Dày(mm)	Độ dày máy	float64
Trọng Lượng(kg)	Trọng lượng máy tính	float64

3.2. Tiết xử lý dữ liệu

- Import thư viện và dữ liệu

```
import numpy as np
import pandas as pd
data = pd.read_csv('data_laptop.csv')
```

Hình 3.2 Import thư viện và đọc dữ liệu vào

Ta thực hiện việc đọc tập dữ liệu vào DataFrame có tên là “data” và dùng lệnh “data.head” hiển thị các dòng đầu tiên của dữ liệu để hiểu rõ hơn về thông tin các dòng các cột của bộ dữ liệu.

- Kiểm tra dữ liệu

```
RangeIndex: 1027 entries, 0 to 1026
Data columns (total 47 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Link             1027 non-null   object  
 1   Thương hiệu       1027 non-null   object  
 2   Tên sp            1027 non-null   object  
 3   Thông tin          1005 non-null   object  
 4   Giá              1017 non-null   float64 
 5   Tên khách hàng    816 non-null   object  
 6   Rating            816 non-null   float64 
 7   Hải lòng          816 non-null   float64 
 8   Đánh giá          816 non-null   object  
 9   RAM               1027 non-null   object  
 10  Loại RAM          1027 non-null   object  
 11  Tốc độ Bus RAM    1027 non-null   object  
 12  Hỗ trợ RAM tối đa  1027 non-null   object  
 13  Ổ cứng            1027 non-null   object  
 14  Công nghệ CPU      1027 non-null   object  
 15  Số nhân            1027 non-null   int64  
 16  Số luồng          1027 non-null   object  
 17  Tốc độ CPU          1027 non-null   object  
 18  Tốc độ tối đa        1027 non-null   object  
 19  NPU                24 non-null    object  
 20  Hiệu năng xử lý AI (TOPS) 129 non-null   object  
 21  Màn hình            1027 non-null   object  
 22  Độ phân giải        1027 non-null   object  
 23  Tần số quét         1027 non-null   object  
 24  Độ phủ màu          557 non-null   object  
 25  Công nghệ màn hình    1026 non-null   object  
 26  Màn hình cảm ứng      40 non-null    object  
 27  Độ sáng SDR          11 non-null    object  
 28  Card màn hình        827 non-null   object  
 29  Công nghệ âm thanh    806 non-null   object  
 30  DTS                 59 non-null    object  
 31  Cổng giao tiếp        827 non-null   object  
 32  Kết nối không dây    827 non-null   object  
 33  Webcam              821 non-null   object  
 34  Đèn bàn phím          827 non-null   object  
 35  Tính năng khác        503 non-null   object  
 36  Tản nhiệt            171 non-null   object  
 37  Khe đục thẻ nhớ      224 non-null   object  
 38  Kích thước            827 non-null   object  
 39  Chất liệu             827 non-null   object  
 40  Thông tin Pin          827 non-null   object  
 41  Hẹ điều hành          827 non-null   object  
 42  Thời điểm ra mắt      824 non-null   float64 
 43  Dài(mm)              827 non-null   float64 
 44  Rộng(mm)              827 non-null   float64 
 45  Dày(mm)              827 non-null   float64 
 46  Trọng Lượng(kg)        827 non-null   float64 
dtypes: float64(8), int64(1), object(38)
```

Hình 3.3 Kiểm tra thông tin các cột

Bộ dữ liệu data_laptop gồm có 47 cột và 1027 dòng tương ứng. Kết quả sau khi sử dụng câu lệnh “data.info” hiển thị có 8 cột với kiểu dữ liệu là float64, 1 cột có kiểu dữ liệu là int64 và 38 cột có kiểu dữ liệu là object.

- **Kiểm tra các cột có giá trị duy nhất**

Link	340
Thương hiệu	11
Tên sp	341
Thông tin	155
Giá	171
Tên khách hàng	746
Rating	5
Hài lòng	2
Đánh giá	736
RAM	11
Loại RAM	28
Tốc độ Bus RAM	17
Hỗ trợ RAM tối đa	9
Ổ cứng	28
Công nghệ CPU	83
Số nhân	10
Số luồng	13
Tốc độ CPU	33
Tốc độ tối đa	20
NPU	2
Hiệu năng xử lý AI (TOPS)	6
Màn hình	14
Độ phân giải	27
Tần số quét	10
Độ phủ màu	5
Công nghệ màn hình	28
Màn hình cảm ứng	1
Độ sáng SDR	2
Card màn hình	26
Công nghệ âm thanh	24
DTS	1
Cổng giao tiếp	8
Kết nối không dây	10
Webcam	8
Đèn bàn phím	12
Tính năng khác	12
Tản nhiệt	10
Khe đọc thẻ nhớ	3
Kích thước	83
Chất liệu	11
Thông tin Pin	42
Hệ điều hành	8
Thời điểm ra mắt	6
Dài(mm)	50
Rộng(mm)	60
Dày(mm)	40
Trọng Lượng(kg)	55

Hình 3.4 Kiểm tra giá trị duy nhất

Để kiểm tra các giá trị duy nhất ta sử dụng câu lệnh “data.unique”. Kết quả cho thấy rằng không có sự biến đổi giá trị nào trong 2 cột là cột Màn hình cảm ứng và DTS. Do không giúp ích trong việc phân loại hoặc dự đoán, nên xóa 2 cột này giúp giảm kích thước cho dữ liệu.

- **Kiểm tra giá trị null của mỗi cột**

Sử dụng câu lệnh `data.isnull().sum()` để xác định các giá trị null thấy được:

Link	0
Thương hiệu	0
Tên sp	0
Thông tin	22
Giá	10
Tên khách hàng	211
Rating	211
Hài lòng	211
Đánh giá	211
RAM	0
Loại RAM	0
Tốc độ Bus RAM	0
Hỗ trợ RAM tối đa	0
Ổ cứng	0
Công nghệ CPU	0
Số nhân	0
Số luồng	0
Tốc độ CPU	0
Tốc độ tối đa	0
NPU	1003
Hiệu năng xử lý AI (TOPS)	898
Màn hình	0
Độ phân giải	0
Tần số quét	0
Độ phủ màu	470
Công nghệ màn hình	1
Màn hình cảm ứng	987
Độ sáng SDR	1016
Card màn hình	200
Công nghệ âm thanh	221
DTS	968
Cổng giao tiếp	200
Kết nối không dây	200
Webcam	206
Đèn bàn phím	200
Tính năng khác	524
Tản nhiệt	856
Khe đọc thẻ nhớ	803
Kích thước	200
Chất liệu	200
Thông tin Pin	200
Hệ điều hành	200
Thời điểm ra mắt	203
Dài(mm)	200
Rộng(mm)	200
Dày(mm)	200
Trọng Lượng(kg)	200

Hình 3.5 Kiểm tra giá trị null

Các cột không có giá trị null và một số cột khác có giá trị null thấp như cột ‘Giá’ thiếu 10 giá trị, cột ‘Thông tin’ thiếu 22 giá trị, cột ‘Công nghệ màn hình’ chỉ thiếu 1 giá trị. Tiến hành xóa các dòng có dữ liệu null với số lượng ít này vì không ảnh hưởng nhiều đến phân tích.

Các cột NPU, Hiệu năng xử lý AI (TOPS), Độ phủ màu, Màn hình cảm ứng, DTS, Độ sáng SDR, Tính năng khác, Tản nhiệt và Khe đọc thẻ nhớ, link có số lượng giá trị null rất lớn, điều này có thể ảnh hưởng đến các phân tích và mô hình hóa dữ liệu. Do đó đối với những cột này ta sẽ tiến hành loại bỏ hẳn.

Các cột như kích thước, chất liệu, webcam, đèn bàn phím, trọng lượng (kg),.. đều thiếu khoảng 200 giá trị null, con số khá cao. Tuy nhiên cũng không thể xử lý các giá trị thiếu bằng cách điền giá trị trung bình hoặc median được vì sẽ làm thay đổi rất lớn đến kết quả tổng thể. Do đó đối với những cột này ta sẽ chỉ tiến hành loại bỏ các dòng có dữ liệu null.

- **Kiểm tra các giá trị trùng lặp**

```
#Kiểm tra sự trùng lặp của dữ liệu  
data.duplicated().sum()
```

0

Hình 3.6 Kiểm tra sự trùng lặp

Thực hiện việc kiểm tra các dữ liệu trùng lặp bằng lệnh data.duplicated().sum(). Kết quả trả về là 0 cho thấy không có hàng nào trong DataFrame “data” bị trùng lặp. Đó là một kết quả tích cực trong việc xử lý dữ liệu và giúp đảm bảo tính toàn vẹn cho dữ liệu, tránh sai lệch và tiết kiệm thời gian phân tích dữ liệu.

- **Xử lý các cột có giá trị thiếu nhiều hoặc trống dữ liệu**

```
#Loại bỏ các cột thiếu nhiều dữ liệu và trống dữ liệu  
df_clean = data.drop(['Link', 'Thông tin', 'NPU', 'Hiệu năng xử lý AI (TOPS)', 'Tính năng khác',  
'Màn hình cảm ứng', 'Độ sáng SDR', 'DTS', 'Tản nhiệt', 'Khe đọc thẻ nhớ', 'Độ phủ màu',  
'Kích thước', 'Đày(mm)'], axis=1)  
  
df = df_clean.dropna()
```

Hình 3.7 Xóa cột thiếu hoặc trống dữ liệu

Sau khi xóa bỏ 11 cột ở trên, thực hiện kết quả kiểm tra lại dữ liệu, ta thu được 34 cột dữ liệu và 793 bản ghi

```
df.shape
```

(793, 34)

Hình 3.8 Kết quả dữ liệu mới

- **Kiểm tra lại các giá trị sau khi đã xử lý**

Thương hiệu	0
Tên sp	0
Giá	0
Tên khách hàng	0
Rating	0
Hài lòng	0
Đánh giá	0
RAM	0
Loại RAM	0
Tốc độ Bus RAM	0
Hỗ trợ RAM tối đa	0
Ổ cứng	0
Công nghệ CPU	0
Số nhân	0
Số luồng	0
Tốc độ CPU	0
Tốc độ tối đa	0
Màn hình	0
Độ phân giải	0
Tần số quét	0
Công nghệ màn hình	0
Card màn hình	0
Công nghệ âm thanh	0
Cổng giao tiếp	0
Kết nối không dây	0
Webcam	0
Đèn bàn phím	0
Chất liệu	0
Thông tin Pin	0
Hệ điều hành	0
Thời điểm ra mắt	0
Dài(mm)	0
Rộng(mm)	0
Trọng Lượng(kg)	0

Hình 3.9 Các cột sau xử lý

Sử dụng lệnh `data.isnull().sum()` để kiểm tra lại số lượng giá trị null (NaN) cho từng cột trong DataFrame “`data`” sau khi đã xử lý. Kết quả cho biết tình trạng dữ liệu của từng cột sau khi đã loại bỏ các hàng chứa giá trị null trước đó là 0. Điều này cho thấy đã có một DataFrame sạch với dữ liệu đầy đủ cho tất cả các cột.

3.3. Thống kê mô tả

	Giá	Rating	Hài lòng	Số nhân	Thời điểm ra mắt	Dài(mm)	Rộng(mm)	Trọng Lượng(kg)
count	7.930000e+02	793.000000	793.000000	793.000000	793.000000	793.000000	793.000000	793.000000
mean	1.749903e+07	3.958386	0.593947	9.414880	2022.766709	349.240391	236.692333	1.74169
std	5.563504e+06	1.275065	0.491405	2.617083	0.884662	18.473169	14.373650	0.30767
min	8.490000e+06	1.000000	0.000000	2.000000	2020.000000	304.100000	210.000000	1.20000
25%	1.449000e+07	3.000000	0.000000	8.000000	2022.000000	354.000000	227.600000	1.60000
50%	1.669000e+07	4.000000	1.000000	10.000000	2023.000000	358.600000	235.560000	1.70000
75%	1.919000e+07	5.000000	1.000000	12.000000	2023.000000	359.860000	249.100000	1.86000
max	6.949000e+07	5.000000	1.000000	20.000000	2024.000000	369.000000	279.900000	2.60000

Hình 3.10 Các chỉ số thống kê mô tả của dữ liệu

Giải thích các thông số có trong dữ liệu:

Giá: Giá laptop trải rộng từ 8.49 triệu đến 69.49 triệu VND, nhưng phần lớn nằm trong khoảng 14-19 triệu VND tức là nằm ở phân khúc tầm trung, phù hợp cho sinh viên, dân văn phòng và người dùng phổ thông. Độ lệch chuẩn cao (5.56 triệu VND) cho thấy thị trường laptop có rất nhiều sự lựa chọn, từ laptop giá rẻ dưới 10 triệu cho đến những mẫu cao cấp gần 70 triệu.

Rating (Điểm đánh giá): Đánh giá trung bình 3.96/5 là một tín hiệu tích cực, cho thấy phần lớn người dùng hài lòng với sản phẩm. Có khoảng 25% laptop nhận rating ≤ 3 sao, điều này có thể do chất lượng sản phẩm kém, hiệu suất không như mong đợi hoặc trải nghiệm người dùng chưa tốt.

Hài lòng (Mức độ hài lòng của khách hàng): Tỷ lệ hài lòng 59% không quá cao, tức là vẫn có tới 41% người dùng không hài lòng về sản phẩm. Điều này cho thấy một số laptop vẫn chưa đáp ứng tốt nhu cầu thực tế. So với rating trung bình gần 4 sao, có vẻ như một số laptop dù có điểm rating ổn nhưng vẫn khiến người dùng không hài lòng hoàn toàn.

Số nhân CPU: Trung bình 9.41 nhân, thấp nhất là 2 nhân và cao nhất là 20 nhân. Laptop hiện nay chủ yếu có từ 8-12 nhân, phù hợp cho hầu hết nhu cầu từ văn phòng đến gaming. Có một số mẫu chỉ có 2 nhân, điều này có thể ảnh hưởng đến hiệu suất đáng kể, nhất là với các tác vụ đa nhiệm nặng.

Thời điểm ra mắt: Trung bình các sản phẩm được ra mắt vào khoảng năm 2022 – 2023. Đây có thể là giai đoạn máy được người dùng ưa thích nhất.

Kích thước vật lý: Dài 349.24 mm (~13.75 inch), cao độ 304.1 mm - 369 mm. Rộng 236.69 mm (~9.32 inch), cao độ 210 mm - 279.9 mm. Phần lớn laptop có kích thước tiêu chuẩn 13-15 inch.

Trọng lượng laptop: Phần lớn laptop nặng từ 1.5 - 2.5 kg, phù hợp để di chuyển.

Kết luận: Phần lớn laptop trong bộ dữ liệu thuộc phân khúc tầm trung, có hiệu năng tốt và được đánh giá cao, nhưng vẫn có một số laptop khiến khách hàng chưa thực sự hài lòng.

3.4. Trực quan hóa dữ liệu



Hình 3.11 Biểu đồ tổng quan thông tin Laptop tại Thế Giới Di Động

Tổng doanh thu đạt 14 tỷ đồng. Có 124 dòng sản phẩm laptop khác nhau. Tổng cộng có 793 đơn hàng được bán ra.

Giá bán trung bình đạt 17,5 triệu đồng. Mức rating trung bình là 3,96/5 và mức độ hài lòng trung bình là 0,59.

Asus dẫn đầu về cả doanh thu (4 tỷ – 32,16%) và số lượng sản phẩm (263 sản phẩm), cho thấy đây là thương hiệu chủ lực trong danh mục laptop.

Năm 2023 là giai đoạn bán hàng mạnh nhất với 350 sản phẩm (chiếm 44,14%), tiếp theo là năm 2022 và 2024.

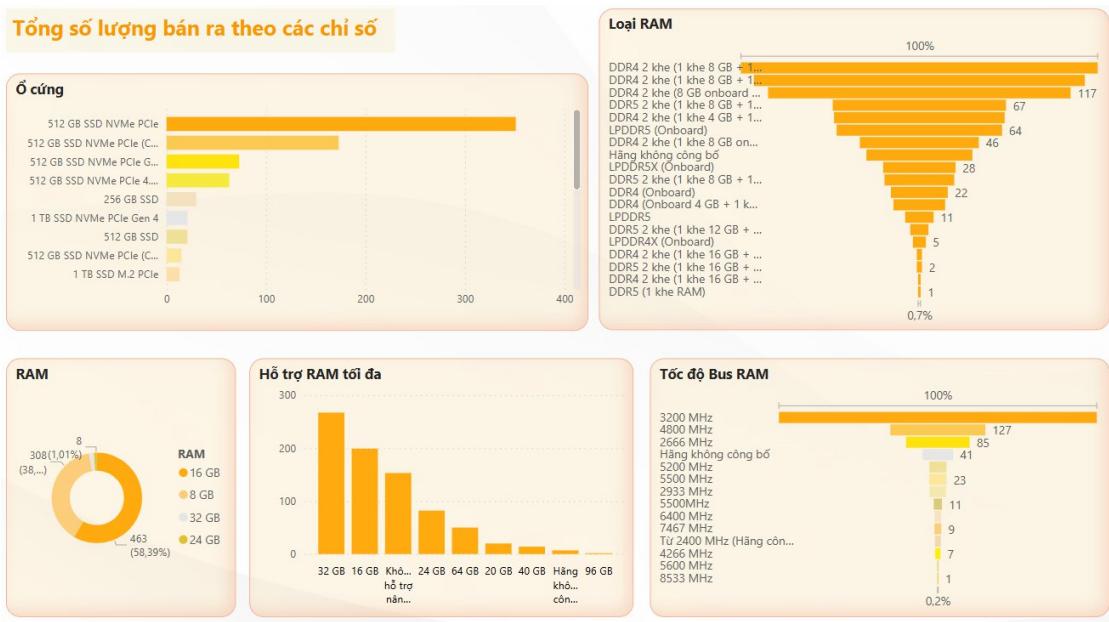
Rating 5 sao chiếm gần 50%, trong khi các mức 1–2 sao khá thấp. Tuy nhiên, vẫn có khoảng 40% khách hàng không hài lòng, là điểm cần chú ý để cải thiện chất lượng sản phẩm hoặc dịch vụ.

TOP SP ĐƯỢC ƯA CHUỘNG NHẤT DỰA TRÊN CHỈ SỐ KẾT HỢP RATING & MỨC ĐỘ HÀI LÒNG

Tên sp	Rating TB	Hài lòng TB
Laptop MSI Gaming Thin 15 B12UCX i5 12450H/16GB/512GB/4GB RTX2050/144Hz/Balo/Win11 (2046VN)	5,00	1,00
Laptop MSI Gaming Sword 16 HX B14VFKG i7 14700HX/16GB/1TB/8GB RTX4060/240Hz/Balo/Win11 (045VN)	5,00	1,00
Laptop MSI Gaming GF63 Thin 12UC i7 12650H/8GB/512GB/4GB RTX3050/144Hz/Win11 (887VN)	5,00	1,00
Laptop Lenovo Gaming LOQ 15ARP9 R7 7435HS/24GB/512GB/6GB RTX4050/144Hz/Win11 (83JC0040VN)	5,00	1,00
Laptop HP Pavilion 15 eg3095TU i5 1335U/8GB/512GB/Win11 (8C5L6PA)	5,00	1,00
Laptop HP Gaming VICTUS 16 r0128TX i5 13450HX/16GB/512GB/144Hz/6GB RTX4050/Win11 (8C5N3PA)	5,00	1,00
Laptop HP 240 G9 i3 1215U/8GB/512GB/Win11 (AG2J5AT)	5,00	1,00
Laptop HP 15 fd0303TU i3 1315U/8GB/512GB/Win11 (A2NL4PA)	5,00	1,00
Laptop Dell Vostro 15 3520 i5 1235U/16GB/512GB/120Hz/OfficeHS/KYHD/Win11 (i5U165W11GRU)	5,00	1,00
Laptop Dell Inspiron 15 3530 i7 1355U/16GB/512GB/120Hz/OfficeHS/Win11 (71043888)	5,00	1,00
Laptop Dell Inspiron 15 3530 i7 1355U/16GB/1TB/120Hz/OfficeHS/Win11 (P16WD)	5,00	1,00
Laptop Dell Inspiron 15 3520 i5 1235U/16GB/512GB/120Hz/OfficeHS/Win11 (N5I5052W1)	5,00	1,00
Laptop Asus Zenbook S 14 UX5406SA Ultra 7 258V/32GB/1TB/120Hz/Túi/OfficeHS/Win11 (PV140WS)	5,00	1,00
Laptop Asus Vivobook 15 X1504ZA i5 1235U/16GB/1TB/Win11 (NJ1528W)	5,00	1,00
Laptop Asus TUF Gaming A15 FA507NV R7 7735HS/16GB/512GB/8GB RTX4060/144Hz/Win11 (LP031W)	5,00	1,00
Laptop Asus TUF Gaming A15 FA506NC R5 7535HS/16GB/512GB/4GB RTX3050/144Hz/Win11 (HN017W)	5,00	1,00
Laptop Asus Gaming TUF A15 FA507NUR R7 7435HS/16GB/512GB/6GB RTX4050/144Hz/Win11 (LP057W)	5,00	1,00
Laptop Apple MacBook Pro 16 inch M4 Pro 24GB/1TB	5,00	1,00
Laptop Acer Gaming Nitro AN515 58 773Y i7 12700H/16GB/512GB/4GB RTX3050Ti/144Hz/Win11 (NH.QFKSV.001.16G)	5,00	1,00
Laptop Acer Aspire Lite 14 51M 36PN i3 1215U/8GB/512GB/Win11 (NX.KTVWSV.001)	5,00	1,00
Laptop Acer Aspire Lite 14 51M 36MH i3 1215U/8GB/256GB/Win11 (NX.KTVSV.001)	5,00	1,00
Tổng	5,00	1,00

Hình 3.12 Biểu đồ top sản phẩm theo Rating và Hài lòng

Bảng thống kê thể hiện danh sách Top 23 sản phẩm laptop đạt điểm tối đa về cả Rating trung bình (5.0) và mức độ hài lòng trung bình (1.0) – cho thấy đây là những sản phẩm được khách hàng đánh giá rất cao và hài lòng tuyệt đối.



Hình 3.13 Biểu đồ số lượng bán ra theo các chỉ số liên quan đến Ram và lưu trữ

Ổ cứng 512 GB SSD NVMe PCIe là loại ổ cứng được lựa chọn lưu trữ phổ biến, với số lượng bán lớn hơn đáng kể so với các loại khác.

Loại RAM DDR4 2 khe là loại RAM có người dùng phổ biến nhất, đặc biệt là cấu hình DDR4 2 khe (1 khe 8 GB + 1 khe 8 GB). Các loại RAM khác vẫn còn chiếm tỷ lệ nhỏ hơn.

RAM 16 GB chiếm tỷ lệ lớn nhất (56.39% với 463 chiếc). Tiếp theo là RAM 8 GB. Các dung lượng lớn hơn như 32 GB và 24 GB chiếm tỷ lệ nhỏ hơn đáng kể.

Các dòng laptop có hỗ trợ tốc độ Bus RAM 3200 MHz chiếm đa số trong tổng số sản phẩm bán ra. Các tốc độ khác phân bố rải rác với số lượng bán ít hơn đáng kể.

Tổng số lượng bán ra theo các chỉ số



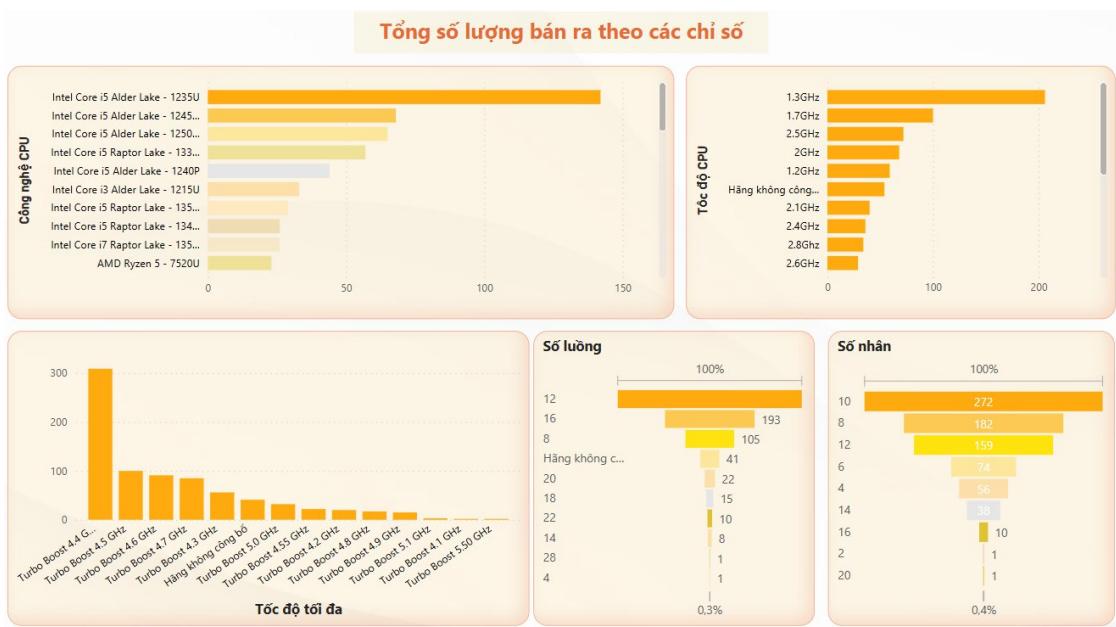
Hình 3.14 Biểu đồ số lượng bán ra theo các chỉ số liên quan đến màn hình

Nhờ khả năng tái tạo màu sắc chính xác và sống động, cùng với góc nhìn rộng. Tấm nền IPS trở thành công nghệ màn hình chiếm ưu thế rõ rệt trong tất cả các loại.

Kích thước màn hình 15.6 inch chiếm phần nhiều so với các loại khác. Vì kích thước này đủ lớn để làm việc đa nhiệm thoải mái, xem nội dung giải trí hấp dẫn, nhưng vẫn đủ gọn gàng để có thể mang theo khi cần thiết.

Độ phân giải Full HD (1920 x 1080) đã trở thành một tiêu chuẩn vàng trong nhiều năm qua khi luôn dẫn đầu về số lượng bán ra với 598 chiếc.

Mặc dù các tần số quét cao hơn mang lại trải nghiệm mượt mà hơn đáng kể nhưng chúng thường đi kèm với chi phí sản xuất cao hơn và có thể tiêu thụ nhiều năng lượng hơn. Do đó, 60Hz vẫn là lựa chọn phổ biến nhất của phần lớn khách hàng khi mua laptop.



Hình 3.15 Biểu đồ số lượng bán ra theo các chỉ số liên quan đến CPU

Các dòng CPU Intel Core i5 thuộc thế hệ Alder Lake (đầu 12xx) phân phối với số lượng đáng kể, đặc biệt là Core i5-1235U, phản ánh xu hướng tiêu dùng nghiêng mạnh về dòng chip Alder Lake của Intel. Trong khi đó, sự hiện diện của CPU AMD là khá khiêm tốn.

Các CPU có tốc độ xung nhịp cơ bản khoảng 1.3GHz và tốc độ Turbo Boost tối đa khoảng 4.4 - 4.7 GHz là phổ biến.

Phần lớn laptop bán ra sử dụng CPU có 8, 10 hoặc 12 nhân và sở hữu 12 luồng xử lý – một cấu hình phù hợp cho đa số người dùng hiện nay, cân bằng giữa khả năng xử lý và mức giá.



Hình 3.16 Biểu đồ số lượng bán ra theo các chỉ số liên quan đến Công nghệ màn hình và công kết nối

Card đồ họa tích hợp Intel Iris Xe Graphics là lựa chọn phổ biến nhất. Trong khi đó, những dòng sử dụng Intel UHD Graphics cũng có sức tiêu thụ đáng kể.

SonicMaster audio xuất hiện ở phần lớn các sản phẩm được bán. Sự xuất hiện của các công nghệ cao cấp như Audio by B&O, Hi-Res Audio,... tuy ít sản phẩm hơn nhưng vẫn cho thấy sự sự hiện diện của phân khúc người dùng chú trọng đến trải nghiệm âm thanh chất lượng cao.

USB Type-C và Jack tai nghe 3.5mm có mặt ở nhiều sản phẩm đã được tiêu thụ, điều này cho thấy người dùng ưu tiên sự linh hoạt và khả năng tương thích cao với nhiều thiết bị ngoại vi. Ngoài ra, cổng mạng LAN (RJ45) cũng giữ vai trò quan trọng với nhiều người dùng.

Wi-Fi 6E và 6 (802.11ax) vẫn là chuẩn Wi-Fi hiện diện phổ biến nhất. Bluetooth 5.2 cũng khá thịnh hành, phù hợp với nhu cầu sử dụng phụ kiện không dây ngày càng cao như tai nghe, bàn phím, chuột...



Hình 3.17 Biểu đồ số lượng bán ra theo các chỉ số liên quan đến đặc điểm thiết kế của sản phẩm

Số lượng sản phẩm được đánh giá là Hài lòng chiếm đa số với 471 sản phẩm (59.39%). Số lượng sản phẩm bị đánh giá là Không hài lòng là 322 sản phẩm (40.61%). Điều này cho thấy đa phần khách hàng khá hài lòng với sản phẩm.

Mức đánh giá Rating 5 sao chiếm tỷ lệ cao nhất với 376 sản phẩm (47.41%). Các mức đánh giá thấp hơn có số lượng giảm dần tuy nhiên số lượng vẫn khá cao.

Webcam HD có số lượng bán ra cao nhất, vượt trội so với các loại webcam khác. Các loại webcam có độ phân giải cao hơn như Full HD, 1080p, 720p có số lượng bán ra thấp hơn đáng kể so với webcam HD.

Đa số sản phẩm bán ra là các sản phẩm không có đèn bàn phím. Trong số các sản phẩm có đèn, đèn màu trắng và đèn chuyển màu RGB 4 vùng là phổ biến hơn so với các loại đèn bàn phím khác.

Vỏ nhựa là chất liệu phổ biến nhất với 471 sản phẩm (59.39%). Các loại chất liệu khác bằng kim loại hoặc chứa một phần kim loại có số lượng ít hơn. Điều này có thể do khách hàng có sự ưu tiên về giá cả hoặc tính phổ biến của chất liệu nhựa.



Hình 3.18 Biểu đồ số lượng bán ra theo các chỉ số liên quan đến hệ điều hành và kích thước sản phẩm

Windows 11 Home Single Language vượt trội hơn so với các phiên bản Windows 11 Home khác (bao gồm cả Office Home & Student) và hệ điều hành macOS.

Các sản phẩm sử dụng pin 3 cell với dung lượng 41Wh, 42Wh và 50Wh phổ biến hơn hẳn so với các cấu hình pin khác về số lượng sản phẩm bán ra.

Chiều dài và chiều rộng của sản phẩm có sự phân bố khá phân tán, cho thấy sự đa dạng về kích thước của các mẫu sản phẩm được bán ra.

Trọng lượng sản phẩm tập trung nhiều nhất ở mức khoảng 1.70 kg, cho thấy đây có thể là một trọng lượng phổ biến cho các sản phẩm được người dùng lựa chọn.

CHƯƠNG 4. PHÂN CỤM DỮ LIỆU, XÂY DỰNG MÔ HÌNH DỰ ĐOÁN SỰ HÀI LÒNG VÀ ĐỀ XUẤT SẢN PHẨM CHO KHÁCH HÀNG

4.1. Dự đoán mức độ hài lòng của khách hàng

4.1.1. Chuẩn bị dữ liệu

- Import dữ liệu

Sử dụng bộ dữ liệu đã được xử lý ở trên để tiến hành xây dựng mô hình. Thực hiện import các thư viện cần thiết và đọc dữ liệu.

Thương hiệu	Tên sp	Giá khách hàng	Rating	Hài lòng	Đánh giá	RAM	Loại RAM	Tốc độ Bus RAM	Kết nối không dây	Webcam	Đèn bàn phím	Chất liệu	Thông tin Pin	Hệ điều hành	Thời diểm ra mắt	Dài(mm)	Rộng(mm)	Trọng Lượng(kg)	
0 Acer	Laptop Acer Aspire 3 A314-42P R3B3 R7 5700U/16...	12490000	Phạm Dung	1	1	Máy để chế độ sleep khoảng 12 tiếng sẽ bị nóng...	16 GB	LPDDR4X (Onboard)	4266 MHz	Wi-Fi 6 (802.11ax)	HD webcam	Không có đèn	Vỏ nhựa	3-cell, 50Wh	Windows 11 Home SL	2024	318.2	225.5	1.4
1 Acer	Laptop Acer Aspire 3 A314-42P R3B3 R7 5700U/16...	12490000	Lê Văn Đối	1	1	Mua sản phẩm vào tháng 09/2024 đến tháng 02/20...	16 GB	LPDDR4X (Onboard)	4266 MHz	Wi-Fi 6 (802.11ax)	HD webcam	Không có đèn	Vỏ nhựa	3-cell, 50Wh	Windows 11 Home SL	2024	318.2	225.5	1.4

Hình 4.1 Đọc file dữ liệu

- Chuyển đổi dữ liệu

Để đảm bảo biến mục tiêu (Hài lòng) có kiểu dữ liệu phù hợp cho các thuật toán học máy, ta cần chuyển cột Hài lòng sang kiểu số nguyên (int64). Điều này giúp các mô hình phân loại dễ dàng tiếp nhận và xử lý biến mục tiêu trong quá trình huấn luyện.

- Mã hóa các biến dạng chuỗi

Cột Tên khách hàng chỉ mang tính chất định danh, không đóng góp giá trị cho việc dự đoán sự hài lòng. Do đó, cần loại bỏ cột này để tránh gây nhiễu dữ liệu và giúp mô hình tập trung vào các đặc trưng thực sự liên quan.

Các biến có kiểu dữ liệu object (chuỗi) cần được mã hóa thành số trước khi đưa vào mô hình học máy. Sử dụng LabelEncoder để mã hóa từng cột dạng chuỗi, đảm bảo rằng tất cả dữ liệu đầu vào đều ở dạng số nguyên.

Thương hiệu	Tên sp	Giá	Rating	Hàng	Đánh giá	RAM	Loại RAM	Tốc độ Bus RAM	Hỗ trợ RAM tối đa	...	Kết nối không dây	Webcam	Đèn bàn phím	Chất liệu	Thông tin Pin	Né hành	Thời điểm ra mắt	Dài(mm)	Rộng(mm)	Trọng Lượng(kg)	
0	0	1	12490000	1	1	315	0	15	3	8	...	9	5	1	6	10	2	2024	318.2	225.5	1.4
1	0	1	12490000	1	1	163	0	15	3	8	...	9	5	1	6	10	2	2024	318.2	225.5	1.4
2	0	1	12490000	5	1	674	0	15	3	8	...	9	5	1	6	10	2	2024	318.2	225.5	1.4
3	0	1	12490000	3	1	588	0	15	3	8	...	9	5	1	6	10	2	2024	318.2	225.5	1.4
4	0	1	12490000	4	1	365	0	15	3	8	...	9	5	1	6	10	2	2024	318.2	225.5	1.4
...
788	1	26	16990000	5	1	206	0	1	2	2	...	11	5	6	8	10	2	2024	358.6	249.1	1.8
789	1	26	16990000	5	1	207	0	1	2	2	...	11	5	6	8	10	2	2024	358.6	249.1	1.8
790	1	26	16990000	1	0	463	0	1	2	2	...	11	5	6	8	10	2	2024	358.6	249.1	1.8
791	1	26	16990000	5	0	137	0	1	2	2	...	11	5	6	8	10	2	2024	358.6	249.1	1.8
792	1	26	16990000	5	0	477	0	1	2	2	...	11	5	6	8	10	2	2024	358.6	249.1	1.8

793 rows x 33 columns

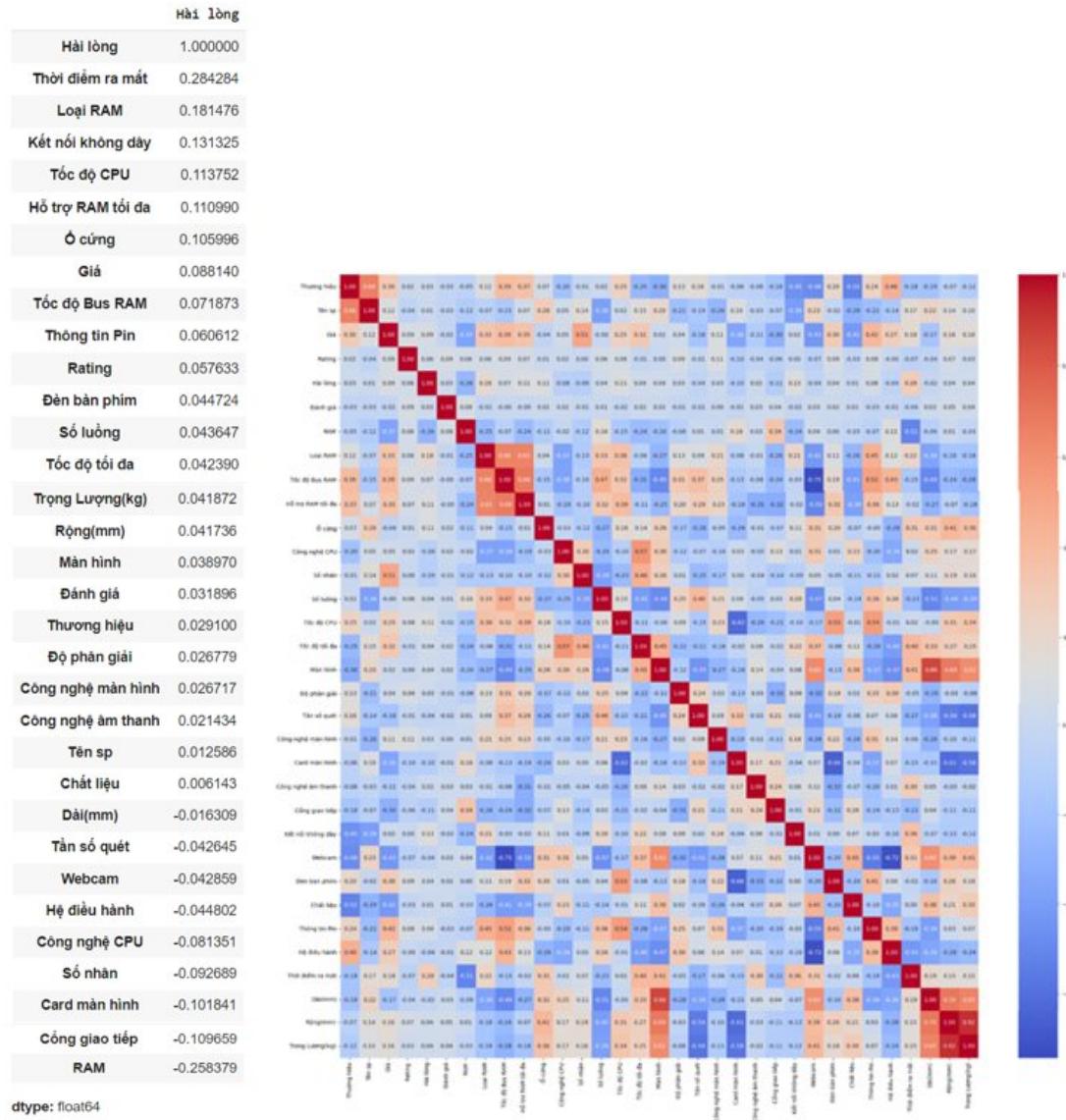
Hình 4.2 Chuyển đổi dữ liệu về dạng số

• Giảm chiều dữ liệu

Đôi khi dữ liệu có quá nhiều đặc trưng (features) không cần thiết, có thể dẫn đến hiện tượng overfitting và tăng chi phí tính toán. Vậy nên chúng ta cần giảm bớt các chiều không cần thiết và giữ lại các chiều quan trọng để xây dựng mô hình nhằm bớt chi phí và thời gian xây dựng mô hình.

Để xác định được các features quan trọng mình cần chọn được ngưỡng xác định phù hợp với bộ dữ liệu. Nếu chọn ngưỡng xác định quá thấp, bộ dữ liệu vẫn còn lại rất nhiều feature và những feature không quan trọng có thể chưa được loại bỏ. Nếu chọn ngưỡng xác định quá cao, bộ dữ liệu sẽ có thể mất đi nhiều feature quan trọng và còn quá ít feature để xây dựng mô hình.

Sử dụng ma trận tương quan (correlation matrix) và lựa chọn ngưỡng (threshold) để loại bỏ các biến có mức độ tương quan thấp.



Hình 4.3 Biểu đồ tương quan của dữ liệu

Với sự so sánh các hiệu suất của mô hình ở các ngưỡng xác định khác nhau cho ra hiệu suất mô hình khác nhau. Trong đó ngưỡng xác định > 0.05 cho ra hiệu suất sau khi xây dựng mô hình là cao nhất.

Tạo một dataframe mới chứa biến phụ thuộc Hài lòng và biến độc lập được chọn từ ma trận tương quan với ngưỡng xác định > 0.05 gồm Giá, Rating, RAM, Loại RAM, Tốc độ Bus RAM, Hỗ trợ RAM tối đa, Ổ cứng, Công nghệ CPU, Số nhân, Tốc độ CPU, Card màn hình, Cổng giao tiếp, Kết nối không dây, Thông tin Pin, Thời điểm ra mắt.

Số biến được chọn: 16																
	Giá Rating	Hài lòng	RAM	Loại RAM	Tốc độ Bus RAM	Hỗ trợ RAM tối đa	Ổ cứng	Công nghệ CPU	Số nhân	Tốc độ CPU	Card màn hình	Công giao tiếp	Kết nối không dây	Thông tin Pin	Thời điểm ra mắt	
0	12490000	1	1	0	15	3	8	7	3	8	4	14	1	9	10	2024
1	12490000	1	1	0	15	3	8	7	3	8	4	14	1	9	10	2024
2	12490000	5	1	0	15	3	8	7	3	8	4	14	1	9	10	2024
3	12490000	3	1	0	15	3	8	7	3	8	4	14	1	9	10	2024
4	12490000	4	1	0	15	3	8	7	3	8	4	14	1	9	10	2024

Hình 4.4 Dataframe mới được chọn từ ma trận tương quan

• Cân bằng dữ liệu

Chia dữ liệu thành 2 tập biến độc lập x và biến phụ thuộc y trước khi cân bằng dữ liệu.

Trong tập dữ liệu biến phụ thuộc y, biến Hài lòng phân chia khách hàng thành 471 khách hàng hài lòng và 322 khách hàng không hài lòng. Tuy chênh lệch chưa cực kỳ lớn, nhưng vẫn có sự mất cân bằng giữa hai nhóm, điều này có thể làm mô hình thiên lệch dự đoán về phía nhóm đông hơn (khách hàng hài lòng) đồng thời mô hình sẽ khó học được đặc trưng của nhóm khách hàng không hài lòng - nhóm mà doanh nghiệp cần chú ý cải thiện.

Để giải quyết vấn đề mất cân bằng dữ liệu sẽ sử dụng kỹ thuật SMOTE một phương pháp phổ biến dùng để cân bằng dữ liệu sẽ giúp cải thiện hiệu suất của mô hình dự đoán bằng cách giảm đi dữ liệu lớp đa số hoặc tạo ra các dữ liệu mới cho lớp thiểu số (Không hài lòng). Ở dữ liệu này, chúng ta sẽ dùng SMOTE để tạo thêm dữ liệu mới cho lớp 0.

Hài lòng	count		Dữ liệu trước khi cân bằng: Hài lòng
	1	0	1
1	471		471
0		322	322
			Name: count, dtype: int64
dtype: int64	Dữ liệu sau khi cân bằng: Hài lòng		
	0	1	471
			Name: count, dtype: int64

Hình 4.5 Cân bằng dữ liệu

• Chuẩn hóa dữ liệu

Chuẩn hóa dữ liệu là quá trình chuẩn hóa dữ liệu về đơn vị đo lường phù hợp nhằm chuẩn bị dữ liệu đầu vào phù hợp giúp tối ưu hóa và cải thiện hiệu suất của các mô hình học máy, đồng thời giúp các thuật toán học máy làm việc hiệu quả hơn với dữ liệu thực tế.

Để đảm bảo các biến có thang đo khác nhau sẽ có phạm vi giá trị tương đồng, tránh một vài biến có ảnh hưởng quá lớn so với các biến khác, chúng ta sẽ sử dụng phương pháp chuẩn hóa Min-max để chuẩn hóa các biến về khoảng 0-1.

	Giá	Rating	RAM	Loại RAM	Tốc độ Bus RAM	Hỗ trợ RAM tối đa	Ó cứng	Công nghệ CPU	Số nhân	Tốc độ CPU	Card màn hình	Công giao tiếp	Kết nối không dây	Thông tin Pin	Thời điểm ra mắt
5	0.085574	0.25	1.0	0.333333	1.000000	0.375	0.368421	0.461538	0.444444	0.047619	0.904762	0.285714	0.000000	0.138889	0.50
361	0.188525	0.50	0.0	0.444444	0.153846	0.250	0.315789	0.564103	0.555556	0.476190	0.095238	0.142857	0.846154	0.277778	1.00
52	0.040984	1.00	1.0	0.333333	0.153846	0.375	0.578947	0.076923	0.333333	0.190476	0.666667	0.285714	0.692308	0.277778	0.75
147	0.339344	0.75	0.0	0.777778	0.923077	1.000	0.263158	0.205128	0.333333	0.238095	0.428571	0.428571	0.692308	1.000000	0.50
186	0.229608	0.75	0.0	0.277778	0.153846	0.375	0.052632	0.564103	0.555556	0.476190	0.095238	0.285714	0.153846	0.555556	0.75

Hình 4.6 Chuẩn hóa Min-max

4.1.2. Mô hình Hồi quy Logistic

4.1.2.1. Huấn luyện mô hình

- Chia tập dữ liệu

Mô hình được chia thành 2 phần là x_train và y_train chiếm 80% bộ dữ liệu dùng để training mô hình, x_test và y_test chiếm 20% bộ dữ liệu dùng để test mô hình và đánh giá hiệu quả mô hình.

Số mẫu trong bộ train: 753
Số mẫu trong bộ test: 189

Hình 4.7 Chia tập dữ liệu

- Khởi tạo mô hình

Trong mô hình Logistic Regression các thông số xây dựng mô hình gồm có:

- penalty = L2: dùng để giảm thiểu overfitting và cải thiện tổng quát hóa của mô hình.
- max_iter = 100: xác định số lần lặp tối đa mà thuật toán sẽ thực hiện trong quá trình huấn luyện.
- tol = 1e-4 (0.0001): tol là ngưỡng dừng, ví dụ 1e-4 tức là dừng khi hàm mất mát thấp hơn 1e-4.

- Huấn luyện mô hình

Quá trình huấn luyện mô hình sẽ sử dụng bộ dữ liệu x_train và y_train để xây dựng ra mô hình dự đoán với các thông số chúng ta khởi tạo.

- **Dự báo nhãn của bộ dữ liệu test**

Sau khi đã huấn luyện mô hình trên tập dữ liệu huấn luyện, việc dự báo nhãn trên tập dữ liệu test là bước quan trọng trong quá trình đánh giá và kiểm tra hiệu suất của mô hình.

- **Hàm dự đoán**

Hàm dự đoán có dạng: $y_{\text{pred}} = \sigma(b + W \times X)$

```
Hàm dự đoán:  
y_pred = sigmoid([-1.39956951] + [[ 0.04456671  0.38176152 -0.6194309 -0.03241243  0.63179232  0.12414764  
  0.30679516 -0.28307281 -1.11243086  0.3102352  0.06881999  0.08565513  
 -0.15913699  0.22942948  2.14058529]] * x)
```

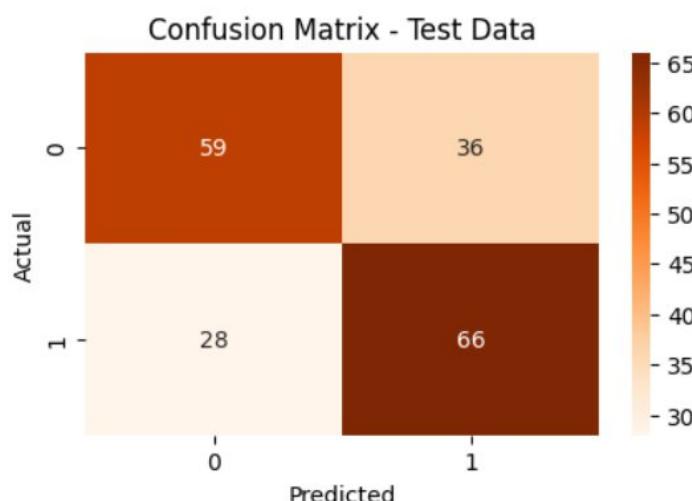
Hình 4.8 Hàm dự đoán

4.1.2.2. Đánh giá mô hình

- **Đánh giá các chỉ số**

Dựa vào kết quả dự báo nhãn của bộ dữ liệu test (y_{prep}) và nhãn của bộ dữ liệu test (y_{test}) chúng ta sẽ đánh giá mô hình qua các chỉ số sau:

Độ chính xác của tập test: 66.14%				
Classification Report:				
	precision	recall	f1-score	support
0	0.68	0.62	0.65	95
1	0.65	0.70	0.67	94
accuracy			0.66	189
macro avg	0.66	0.66	0.66	189
weighted avg	0.66	0.66	0.66	189



Hình 4.9 Confusion matrix mô hình Logistic Regression

Kết quả đánh giá của mô hình Logistic

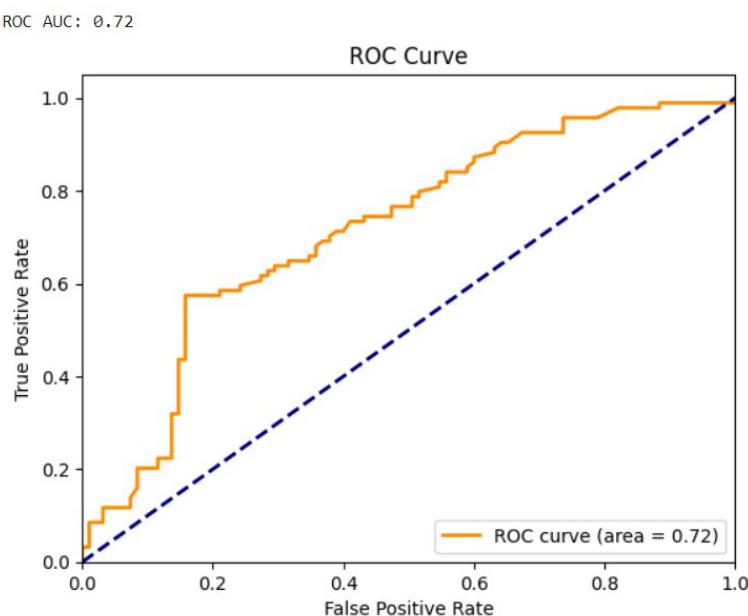
Độ chính xác (Accuracy) là tỷ lệ các dự đoán đúng so với tổng số dự đoán. Trong mô hình, độ chính xác là 66.14% điều này nghĩa là mô hình dự đoán đúng khoảng 66% nhãn của toàn bộ khách hàng. Đây là mức độ chính xác ở mức khá thấp, cho thấy mô hình còn khả năng phân loại nhầm khá nhiều.

Một số lỗi mà mô hình nhầm lẫn có thể được giải thích như sau:

- 59: Số khách hàng thực tế hài lòng được mô hình dự đoán đúng là hài lòng.
- 28: Số khách hàng thực tế hài lòng nhưng bị mô hình dự đoán sai là không hài lòng.
- 36: Số khách hàng thực tế không hài lòng nhưng bị mô hình dự đoán sai là hài lòng.
- 66: Số khách hàng thực tế không hài lòng được mô hình dự đoán đúng là không hài lòng.

Mô hình bị nhầm lẫn khá nhiều giữa hai lớp khi tỷ lệ dự đoán sai (sai nhãn giữa hai nhóm) vẫn còn khá cao (28 và 36 trường hợp cho mỗi nhóm). Điều này cho thấy mô hình vẫn còn khó khăn trong việc phân biệt chính xác giữa khách hàng hài lòng và không hài lòng.

• Đánh giá đồ thị AUC-ROC



Hình 4.10 Biểu đồ ROC mô hình Logistic Regression

Độ cong của đường ROC: Đường cong ROC có xu hướng cong lên từ góc dưới bên trái về phía góc trên bên trái, tuy nhiên không quá rõ rệt. Điều này cho thấy mô hình có khả năng phân biệt giữa hai nhóm khách hàng hài lòng và không hài lòng chỉ ở mức khá, chưa thực sự mạnh

Diện tích dưới đường cong (AUC): Giá trị AUC đạt 0.72, cho thấy mô hình có độ chính xác tổng thể ở mức khá ổn trong việc phân loại cảm nhận của khách hàng. Giá trị này cao hơn mức ngẫu nhiên (0.5) nhưng chưa đạt mức xuất sắc (trên 0.8).

Nhìn chung, với ROC AUC = 0.72, mô hình có hiệu quả ở mức chấp nhận được trong việc dự đoán cảm nhận khách hàng. Doanh nghiệp có thể sử dụng mô hình như một công cụ tham khảo trong phân tích hành vi khách hàng, đồng thời nên kết hợp với các phương pháp tối ưu hóa khác để nâng cao hiệu suất dự đoán.

4.1.3. Mô hình Random Forest

4.1.3.1. Huấn luyện mô hình

- Chia tập dữ liệu**

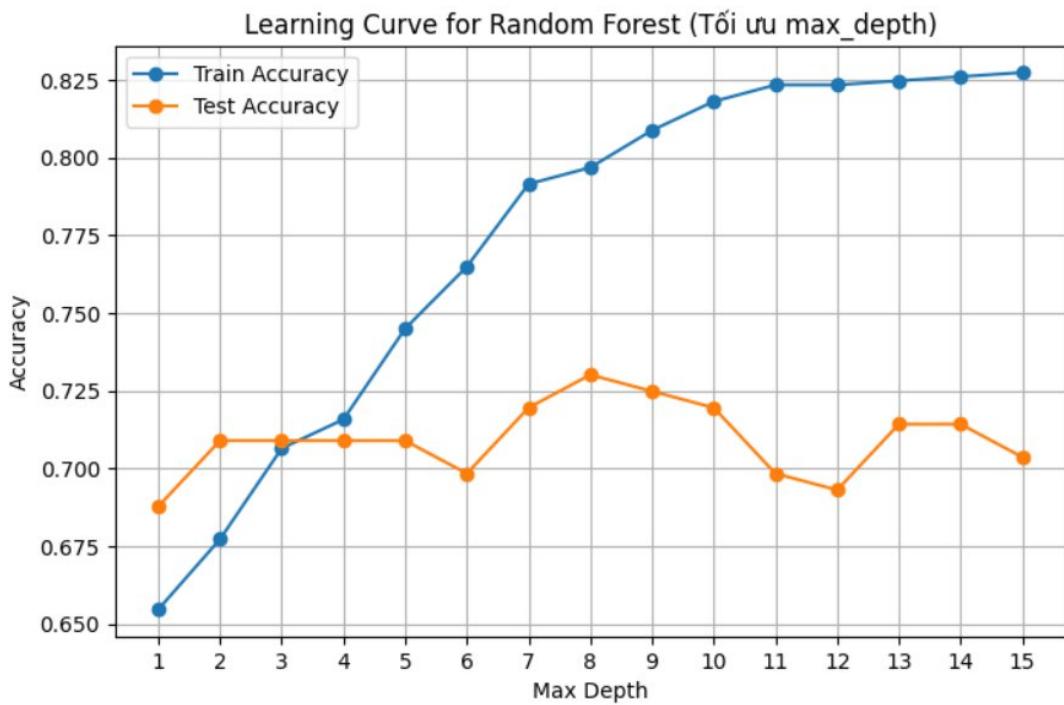
Mô hình được chia thành 2 phần là x_train và y_train chiếm 80% bộ dữ liệu dùng để training mô hình, x_test và y_test chiếm 20% bộ dữ liệu dùng để test mô hình và đánh giá hiệu quả mô hình.

- Chọn giá trị max_depth phù hợp với bộ dữ liệu**

Max_depth: Là tham số quy định độ sâu tối đa của cây quyết định. Nó xác định số lượng các node từ gốc (root) đến lá (leaf) mà cây được phép phát triển.

Biểu đồ Learning Curve:

- Train Accuracy: độ chính xác trên tập huấn luyện
- Test Accuracy: độ chính xác trên tập testing



Hình 4.11 Biểu đồ Learning Curve của mô hình Random Forest

Quan sát biểu đồ Learning Curve:

- Train Accuracy tăng dần đều đặn theo từng giá trị max_depth và dần tiến tới trạng thái bão hòa, cho thấy mô hình càng học sâu thì càng khớp dữ liệu huấn luyện tốt hơn.
- Test Accuracy dao động nhẹ quanh mức 0.70–0.73 và không có xu hướng tăng rõ rệt khi max_depth tăng, thậm chí từ $\text{max_depth} > 9$, độ chính xác trên tập test có xu hướng giảm nhẹ hoặc dao động thất thường, là dấu hiệu của overfitting.

Vì vậy, giá trị max_depth hợp lý cho mô hình nên nằm trong khoảng 6 đến 9, vì tại đây mô hình đạt độ chính xác tốt trên tập kiểm tra trong khi vẫn chưa bị overfitting mạnh

- **Khởi tạo mô hình**

Khởi tạo mô hình là quá trình tạo ra một cấu trúc mô hình từ dữ liệu huấn luyện và đặt các thông số cần thiết cho mô hình trước khi bắt đầu quá trình huấn luyện.

Trong mô hình Random Forest các thông số xây dựng mô hình gồm có:

- criterion = “entropy” : Chỉ số đo mức độ hỗn loạn của lá.
- max_depth = 8 : Độ sâu của cây.
- n_estimators = 99 : Số cây xây dựng của rừng

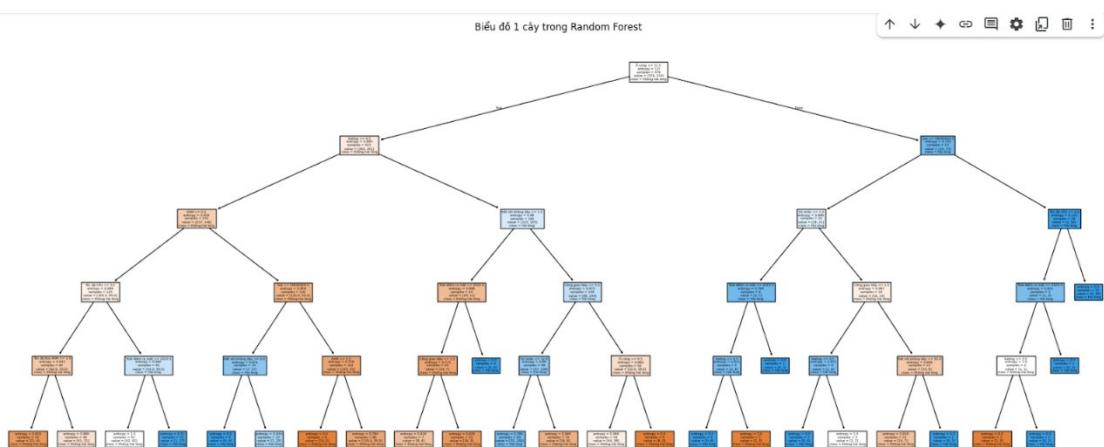
- **Huấn luyện mô hình**

Quá trình huấn luyện mô hình sẽ sử dụng bộ dữ liệu x_train và y_train để xây dựng ra mô hình dự đoán với các thông số chúng ta khởi tạo.

- **Dự báo nhãn của bộ dữ liệu test**

Sau khi đã huấn luyện mô hình trên tập dữ liệu huấn luyện, việc dự báo nhãn trên tập dữ liệu test là bước quan trọng trong quá trình đánh giá và kiểm tra hiệu suất của mô hình.

- **Cây của mô hình Random Forest**



Hình 4.12 Cây hoàn chỉnh trong Random forest

Các chỉ số trong 1 nút bao gồm:

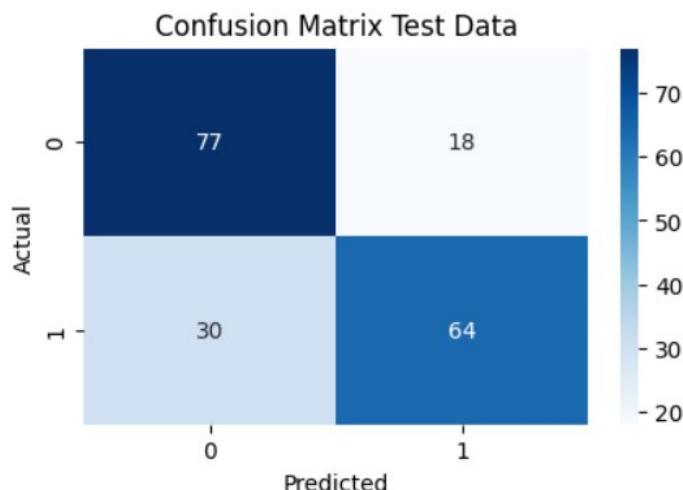
- Tên feature và giá trị
- Entropy: là một độ đo của sự không chắc chắn, entropy đo lường mức độ hỗn loạn trong việc phân loại các mẫu tại node đó. Giá trị entropy càng cao thể hiện sự không chắc chắn càng lớn trong việc phân loại các mẫu.
- Samples: Đây là số lượng mẫu tại node đó.
- Value: Giá trị này biểu thị số lượng của lớp “Hài lòng” và “Không hài lòng” tại node đó so với bộ dữ liệu training.
- Class: Đây là lớp được dự đoán cho node đó dựa trên quy tắc phân loại của cây quyết định.

4.1.3.2. Đánh giá mô hình

- **Đánh giá các chỉ số**

Dựa vào kết quả dự báo nhãn của bộ dữ liệu test (y_prep) và nhãn của bộ dữ liệu test (y_test) chúng ta sẽ đánh giá mô hình qua các chỉ số sau:

Độ chính xác trên tập test: 74.60%				
Classification Report:				
	precision	recall	f1-score	support
0	0.72	0.81	0.76	95
1	0.78	0.68	0.73	94
accuracy			0.75	189
macro avg	0.75	0.75	0.74	189
weighted avg	0.75	0.75	0.74	189



Hình 4.13 Confusion matrix mô hình Random Forest

Độ chính xác (Accuracy) là tỷ lệ các dự đoán đúng so với tổng số dự đoán. Trong mô hình, độ chính xác là 74.60%, nghĩa là mô hình đã dự đoán đúng 75% nhãn của toàn bộ khách hàng. Đây là một tỷ lệ khá cao, cho thấy mô hình có khả năng dự đoán khá chính xác.

Ma trận nhầm lẫn có thể được giải thích như sau:

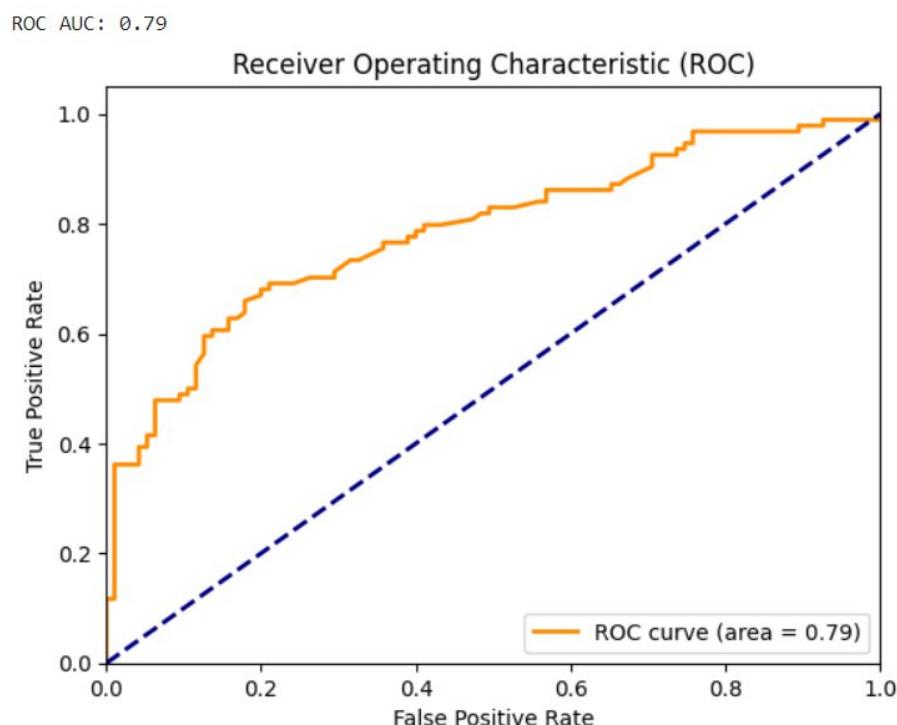
- 77: Số khách hàng thực tế không hài lòng được dự đoán đúng là “Không hài lòng” (True Negative).
- 64: Số khách hàng thực tế hài lòng được dự đoán đúng là “Hài lòng” (True Positive).

- 18: Số khách hàng thực tế không hài lòng nhưng bị mô hình dự đoán nhầm là “Hài lòng” (False Positive).
 - 30: Số khách hàng thực tế hài lòng nhưng bị mô hình dự đoán nhầm là “Không hài lòng” (False Negative).
- **Đánh giá đồ thị AUC-ROC**

Chỉ dựa vào độ chính xác không đủ để đánh giá toàn diện hiệu quả của mô hình này. Do đó, cần bổ sung thêm các chỉ số khác để có cái nhìn tổng quan hơn.

Vì vậy, ta có thể sử dụng điểm AUC-ROC làm thước đo hiệu quả mô hình, AUC-ROC (Area Under the Receiver Operating Characteristic Curve) là một thước đo hiệu quả phổ biến cho các mô hình phân loại, đánh giá khả năng phân biệt giữa các lớp (ví dụ: 1 và 0, khách hàng hài lòng và không hài lòng).

Điểm AUC càng cao, mô hình càng có khả năng phân biệt chính xác giữa các lớp.



Hình 4.14 Biểu đồ Roc của mô hình Random Forest

Dựa vào biểu đồ ROC, ta có thể thấy:

- Độ cong của đường ROC: Đường cong ROC có xu hướng cong lên từ góc dưới bên trái về phía góc trên bên trái. Điều này cho thấy mô hình có khả

năng phân biệt khá tốt giữa hai nhóm khách hàng: hài lòng và không hài lòng.

- Diện tích dưới đường cong (AUC): Giá trị AUC đạt 0.79, cho thấy mô hình Random Forest có độ chính xác tổng thể ở mức khá tốt trong việc phân loại cảm nhận của khách hàng. Mô hình có khả năng phân biệt được khách hàng hài lòng và không hài lòng tốt hơn so với mô hình ngẫu nhiên (AUC = 0.5).

Nhìn chung, với ROC AUC = 0.79, mô hình Random Forest có hiệu quả khá trong việc dự đoán cảm nhận khách hàng. Do đó, doanh nghiệp có thể sử dụng mô hình này để hỗ trợ phân tích tâm lý khách hàng và xây dựng các chiến lược nâng cao mức độ hài lòng.

4.1.4. So sánh các chỉ số của 2 mô hình

Bảng 4.1 So sánh các chỉ số của 2 mô hình

Mô hình/Chỉ số	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	74.60%	78%	68%	73%	0.79
Regression Logistic	66.14%	65%	70%	67%	0.72

Accuracy: Mô hình Random Forest đạt độ chính xác cao hơn đáng kể (74.60%) so với Logistic Regression (66.14%), cho thấy Random Forest có khả năng phân loại chính xác tổng thể tốt hơn trong việc đánh giá mức độ hài lòng của khách hàng.

Precision: Precision của Random Forest khá cao (78%), tức là trong số các trường hợp mô hình dự đoán khách hàng hài lòng, phần lớn khách hàng thực sự hài lòng. Trong khi đó, Logistic Regression có Precision thấp hơn (65%), cho thấy khả năng dự đoán đúng các trường hợp hài lòng chưa cao.

Recall: Logistic Regression có recall cao hơn một chút (70%) so với Random Forest (68%), nghĩa là Logistic Regression phát hiện được nhiều hơn các khách hàng thực sự hài lòng, mặc dù đánh đổi bằng precision thấp hơn.

F1-score: Random Forest đạt F1-score (73%), cao hơn Logistic Regression (62%), cho thấy sự cân bằng tốt hơn giữa khả năng phát hiện và dự đoán chính xác khách hàng hài lòng.

AUC: Với AUC = 0.79, Random Forest thể hiện khả năng phân biệt tốt hơn giữa khách hàng hài lòng và không hài lòng so với Logistic Regression (AUC = 0.72) hiệu quả thấp hơn.

4.1.5. Kết luận

Mô hình Random Forest vượt trội hơn Logistic Regression trên hầu hết các chỉ số đánh giá, đặc biệt là độ chính xác tổng thể (Accuracy), khả năng phân biệt (AUC), và cân bằng giữa precision – recall (F1-score). Logistic Regression tuy có recall cao hơn, nhưng lại kém về precision và hiệu suất tổng thể. Do đó, Random Forest là mô hình phù hợp hơn để ứng dụng vào dự đoán sự hài lòng của khách hàng trong bộ dữ liệu này.

4.2. K-Means phân cụm sản phẩm

4.2.1. Thực hiện phân cụm

- Import dữ liệu

Sử dụng bộ dữ liệu đã thông qua xử lý về dạng số, chọn ra các cột phù hợp như bài toán dự đoán mức độ hài lòng của khách hàng

	Giá	Rating	Hài lòng	RAM	Loại RAM	Tốc độ Bus RAM	Hỗ trợ RAM tối đa	Ổ cứng	Công nghệ CPU	Số nhân	Tốc độ CPU	Card màn hình	Cổng giao tiếp	Kết nối không dây	Thông tin Pin	Thời điểm ra mắt
0	12490000	1.0	1	0	15	3	8	7	3	8	4	14	1	9	10	4
1	12490000	1.0	1	0	15	3	8	7	3	8	4	14	1	9	10	4
2	12490000	5.0	1	0	15	3	8	7	3	8	4	14	1	9	10	4
3	12490000	3.0	1	0	15	3	8	7	3	8	4	14	1	9	10	4
4	12490000	4.0	1	0	15	3	8	7	3	8	4	14	1	9	10	4

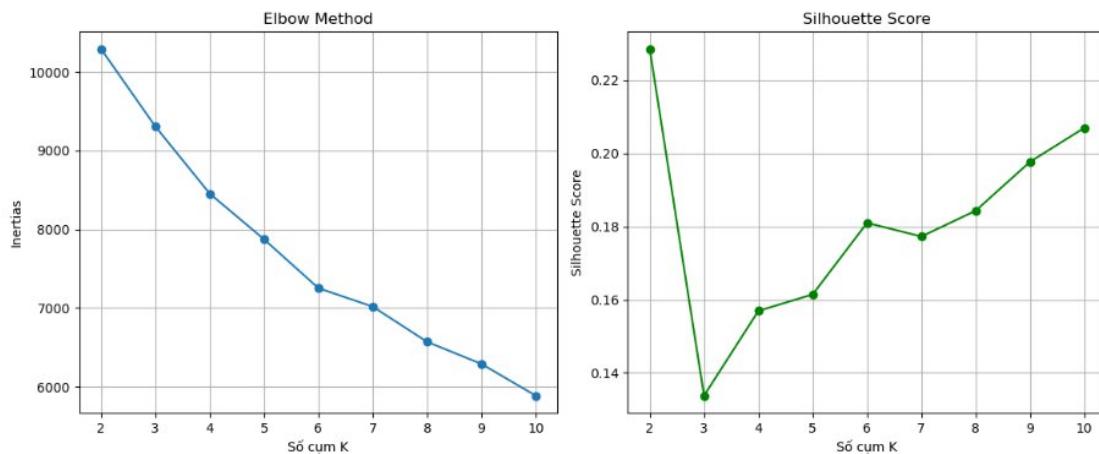
Hình 4.15 Đọc dữ liệu

- Chuẩn hóa dữ liệu

Các biến có các khoảng giá trị hoặc đơn vị khác nhau sẽ ảnh hưởng tới kết quả phân cụm nên trước khi thực hiện phân cụm tiến hành chuẩn hóa dữ liệu để đảm bảo sự đồng đều giữa các biến, bằng cách sử dụng lớp StandardScaler từ thư viện scikit-learn chuẩn hóa theo phân phối chuẩn z-score.

- **Tìm số cụm k tối ưu**

Sau khi chuẩn hóa dữ liệu sẽ thực hiện tìm ra số cụm k tối ưu bằng phương pháp Elbow và Silhouette score. Elbow sẽ tính tổng bình phương khoảng cách giữa các điểm đến tâm cụm, đưa ra điểm gãy trong biểu đồ hay còn gọi là điểm khuỷu tay nơi mà Inertia giảm chậm.



Hình 4.16 Biểu đồ tìm số cụm k

Như trên biểu đồ Elbow có thể thấy đường cong đang giảm chậm ở mức $k=6$ hoặc $k=7$, điểm khuỷu tay xuất hiện ở hai giá trị k này. Biểu đồ Silhouette thấy được điểm k có chỉ số Silhouette tốt là ở $k=8$, $k=9$, $k=10$. Nhìn vào biểu đồ Silhouette $k=2$ là vị trí có chỉ số Silhouette cao nhất nhưng mô hình phân cụm với 2 cụm thì quá đơn giản, không phản ánh được sự đa dạng của dữ liệu. Ta có thể lấy $k = 6$, giá trị k này có thể trung hòa cho cả hai dạng biểu đồ với chỉ số Silhouette khoảng 0.16 ở mức chấp nhận được.

- **Tiến hành phân cụm**

Tiến hành phân cụm với $k = 6$, mô hình sẽ tìm ra 6 tâm cụm và gán nhãn cụm cho từng dòng dữ liệu. Nhãn cụm mô hình dự đoán thể hiện ở cột ‘Cluster’ trong bộ dữ liệu ban đầu.

4.2.2. Kết quả phân cụm

Trước khi thực hiện phân cụm, đã thực hiện việc chuyển đổi các cột có kiểu dữ liệu là object sang dạng số. Để đưa ra các đặc điểm chung của các cụm, ta cần biết được các giá trị đã chuyển đổi thành giá trị nào. Bảng dưới thể hiện giá trị chuyển đổi của các cột. (Mục 2 Phụ lục)

Đặc điểm các cụm:						
	Giá	Rating	Hải lòng	RAM	Loại RAM	Tốc độ Bus RAM
0	1.721696e+07	3.637255	0.627451	-1.110223e-15	6.147059	1.553922
1	1.310168e+07	3.948598	0.425234	3.000000e+00	5.313084	1.985981
2	1.948205e+07	4.173516	0.666667	9.726027e-01	8.178082	2.867580
3	3.119455e+07	4.159091	0.613636	6.136364e-01	16.772727	6.431818
4	1.369417e+07	3.944444	0.763889	2.500000e-01	16.708333	4.875000
5	2.339000e+07	4.275000	0.600000	1.500000e+00	14.000000	11.000000
Hỗ trợ RAM tối đa						
	Hỗ trợ RAM tối đa	Ổ cứng	Công nghệ CPU	Số nhân	Tốc độ CPU	\
0	1.931373	6.475490	23.348039	10.617647	3.651961	
1	1.317757	6.626168	20.121495	8.747664	2.191589	
2	3.150685	8.687306	20.100457	9.378995	12.242809	
3	7.000000	6.704545	16.181818	14.000000	6.454545	
4	8.000000	7.097222	16.777778	6.083333	10.291667	
5	8.000000	3.500000	7.500000	8.000000	12.500000	
Card màn hình						
	Card màn hình	Cổng giao tiếp	Kết nối không dây	Thông tin Pin	Tốc độ CPU	\
0	16.833333	3.019608	3.960784	6.632353		
1	17.897196	4.663551	3.495327	7.869159		
2	3.365297	2.447489	3.876712	16.785388		
3	14.295455	0.727273	5.068182	20.545455		
4	15.791667	2.388889	4.902778	14.069444		
5	10.750000	2.000000	2.500000	35.500000		
Thời điểm ra mắt						
	Thời điểm ra mắt					
0	2022.975490					
1	2022.285047					
2	2023.091324					
3	2023.568182					
4	2023.111111					
5	2021.000000					
Số điểm trong mỗi cụm:						
	Số điểm trong mỗi cụm:					
0	284					
1	214					
2	219					
3	44					
4	72					
5	40					

Hình 4.17 Kết quả phân cụm

Dựa trên kết quả phân cụm ta có thể rút ra được đặc điểm chung của các cụm như sau:

- Giá: Các cụm có sự phân hóa rõ rệt giữa các nhóm laptop giá thấp (cụm 1, 4 với khoảng 13 triệu) và nhóm giá cao (cụm 3 với khoảng 31 triệu). Thể hiện được sự phân khúc thị trường.

- Rating: Các cụm có khoảng rating laptop khá cao, đồng đều nhau, phô biến là 3 và 4 sao, đa số được đánh giá ở mức trung bình khá.
- Hài lòng: Người dùng khá hài lòng về sản phẩm trong các cụm (gần 1).
- RAM: Dao động từ 8 GB đến 32 GB, khá đồng đều, phô biến ở mức 24 GB (cụm 2 và cụm 3)
- Loại RAM: Đa phần các cụm sản phẩm đều sử dụng loại DDR4, một số ít sử dụng DDR5
- Tốc độ Bus RAM: Có sự phân hóa rõ rệt giữa các cụm, có cụm cao đến hơn 5200 MHz và cụm thấp chỉ khoảng 2933–4800 MHz.
- Hỗ trợ RAM tối đa: Đa số các cụm có hỗ trợ khoảng từ 20 GB đến 32 GB, một số ít sản phẩm không có hỗ trợ RAM
- Ổ cứng: Các cụm thể hiện sự khác biệt về loại và dung lượng ổ cứng, đa số các cụm có ổ cứng 512 GB SSD
- Công nghệ CPU: Có sự phân biệt giữa các cụm, có sự hiện diện của các dòng CPU thấp đến cao(core i3, i5 và apple)
- Số nhân: Dao động từ 8 đến 14 nhân.
- Tốc độ CPU: Các cụm có tốc độ CPU dao động từ 1.4GHz đến 2.8GHz, có sự khác biệt giữa các cụm
- Card màn hình: Có sự khác biệt, đa số các cụm có sản phẩm sử dụng Card tích hợp, chỉ 1 cụm sử dụng Card rời
- Cổng giao tiếp: Các cụm chủ yếu sử dụng cổng giao tiếp là LAN
- Kết nối không dây: Các cụm đều sử dụng Wifi 6
- Thông tin Pin: Các sản phẩm trong cụm chủ yếu sử dụng pin 3-cell, một số sản phẩm sử dụng pin 4-cell
- Thời điểm ra mắt: Các cụm có sản phẩm được ra mắt trong khoảng thời gian 2022-2023

Bảng 4.2 Đặc trưng của từng cụm

Cụm	Số lượng	Đặc trưng	Nhóm khách hàng
0	204	- Giá tầm trung (17 triệu)	Phù hợp với sinh viên, nhân viên văn

		<ul style="list-style-type: none"> - RAM 16 GB và ổ cứng 512 GB SSD NVMe PCIe - Card tích hợp - Intel Iris Xe Graphics - Pin trung bình 	phòng, khách hàng không có nhu cầu về đồ họa hay chơi game dung lượng nặng.
1	214	<ul style="list-style-type: none"> - Giá rẻ nhất (13 triệu) - Ram 8 GB và ổ cứng 512 GB SSD NVMe PCIe (có thẻ nâng cấp) - Card tích hợp - Intel Iris Xe Graphics 	Giống với cụm 0 phù hợp với sinh viên, người không có nhu cầu đồ họa, nhưng có ngân sách thấp
2	219	<ul style="list-style-type: none"> - Giá trung bình cao (19.5 triệu) - Rating cao (4.17) - Hiệu năng cao với RAM 24 GB và CPU Alder Lake mạnh - Card rời - NVIDIA GeForce RTX 3050, 6 GB - Tốc độ Bus RAM cao và ổ cứng có thẻ nâng cấp. 	Phù hợp với những khách hàng làm việc liên quan đến đồ họa, lập trình viên, nhu cầu chơi game, làm việc đa nhiệm cần hiệu năng tốt.
3	44	<ul style="list-style-type: none"> - Laptop cao cấp với giá cao (31.2 triệu) - RAM mạnh 24 GB, loại RAM LPDDR5 tốc độ cao - Chip Intel Core i3 thế hệ 13 tiết kiệm điện 	Khách hàng thường xuyên làm việc di động cần mỏng nhẹ.

		- Card rời AMD	
4	72	<ul style="list-style-type: none"> - Giá trung bình thấp (13.7 triệu) - RAM 16 GB LPDDR5 tốc độ cao. - Ổ cứng 512 GB SSD NVMe PCIe(có thể nâng cấp) - Thời lượng pin tốt 	Sinh viên, người dùng mong muốn máy mượt, nhẹ, pin khỏe
5	40	<ul style="list-style-type: none"> - Giá trung bình cao (23.4 triệu) - RAM 32 GB hiệu năng cao - Card tích hợp Apple GPU - Công nghệ CPU Apple M2 hiệu năng mạnh, tiết kiệm điện 	<p>Khách hàng yêu thích dòng sản phẩm Apple, giao diện macOS</p> <p>Khách hàng làm việc trong lĩnh vực sáng tạo nội dung</p>

WSS: 7253.457904290316

BSS: 71.18460165424494

Silhouette Score: 0.18100574718686305

Hình 4.18 Chỉ số của K-means

Để đánh giá mức độ phân cụm, sẽ sử dụng các chỉ số như WSS, BSS và Silhouette Score

WSS sẽ tính bình phương khoảng cách từ mỗi điểm đến tâm cụm của nó, WSS càng thấp thì các điểm nằm càng gần tâm cụm. Ở đây có chỉ số WSS khá cao cho thấy các cụm chưa được chặt chẽ.

BSS tính bình phương giữa các tâm cụm và trung bình tổng thể của dữ liệu, BSS càng cao thì các cụm càng tách biệt rõ. Chỉ số BSS ở đây lại khá thấp thể hiện các cụm chưa tách biệt rõ ràng khó phân biệt giữa các cụm.

Silhouette Score đo mức độ tách biệt và liên kết giữa các cụm, chỉ số này càng gần 1 thì càng tốt. Tuy nhiên chỉ số silhouette của bài toán đang nằm ở mức trung bình cần cải thiện

4.3. Xây dựng mô hình đề xuất

Bộ dữ liệu chứa chủ yếu các cột mô tả về thông số kỹ thuật của sản phẩm, không có chứa thông tin về lịch sử mua hàng của khách hàng như ai đã mua cái gì. Để xây dựng mô hình đề xuất cho bộ dữ liệu này sẽ lựa chọn sử dụng thuật toán Content - Based Filtering, thuật toán này xây dựng mô hình dựa trên sự phân tích các mô tả về sản phẩm, sau đó đưa ra đề xuất các sản phẩm có điểm tương đồng nhất với yêu cầu của khách hàng.

4.3.1. Chuẩn bị dữ liệu

Import vào thư viện cần thiết và bộ dữ liệu đã được xử lý

Tuy nhiên sẽ không sử dụng tất cả các cột trong bộ dữ liệu mà sẽ chọn ra những cột chứa các thông tin quan trọng trong việc ra quyết định mua sản phẩm của khách hàng.

	Thương hiệu	Tên sp	Giá	Đánh giá	RAM	Loại RAM	Ổ cứng	Công nghệ CPU	Tốc độ CPU	Màn hình	Độ phân giải	Card màn hình	Chất liệu	Thông tin Pin	Hệ điều hành
0	Acer	Laptop Acer Aspire 3 A314 42P R3B3	12490000	Máy để chế độ sleep khoảng 12 tiếng sẽ bị nóng...	16 GB	LPDDR4X (Onboard)	512 GB SSD NVMe PCIe (Có thể tháo ra, lắp thanh...)	AMD Ryzen 7 - 5700U	1.8GHz	14"	WUXGA	Card tích hợp - AMD Radeon Graphics	Vỏ nhưa	3-cell, 50Wh	Windows 11 Home SL
1	Acer	Laptop Acer Aspire 3 A314 42P R3B3	12490000	Mua sản phẩm vào tháng 09/2024 đến tháng 02/20...	16 GB	LPDDR4X (Onboard)	512 GB SSD NVMe PCIe (Có thể tháo ra, lắp thanh...)	AMD Ryzen 7 - 5700U	1.8GHz	14"	WUXGA	Card tích hợp - AMD Radeon Graphics	Vỏ nhưa	3-cell, 50Wh	Windows 11 Home SL

Hình 4.19 Đọc dữ liệu

Các chọn được chọn bao gồm 'Thương hiệu', 'Tên sp', 'Giá', 'Đánh giá', 'RAM', 'Loại RAM', 'Ổ cứng', 'Công nghệ CPU', 'Tốc độ CPU', 'Màn hình', 'Độ phân giải', 'Card màn hình', 'Chất liệu', 'Thông tin Pin', 'Hệ điều hành'. Đây là các thông tin khi khách hàng mua sản phẩm sẽ quan tâm đến.

Đưa các cột đã chọn vào dataframe tên là 'data'.

- **Tạo hàm xử lý văn bản**

```
def clean(text):
    if pd.isnull(text):
        return ''
    text = text.lower().strip()
    text = re.sub(r'^\w\s.', ' ', text, flags=re.UNICODE)
    for unit in ['gb', 'tb', 'mb', 'ghz', 'mhz', 'hz', 'wh', 'inch', 'cell']:
        text = re.sub(rf'(\d+(?:\.\d+)?)\s*{unit}', rf'\1{unit}', text)
    text = re.sub(r'\s+', ' ', text).strip()
    return text
```

Hình 4.20 Tạo hàm xử lý

Bộ dữ liệu đã thông qua bước tiền xử lý để làm sạch dữ liệu, tuy nhiên bộ dữ liệu vẫn sẽ còn chứa những dữ liệu bị lỗi chính tả, chứa nhiều ký tự hay khoảng trắng... để thuật toán hoạt động tốt và chính xác hơn sẽ tiến hành làm sạch lại để loại bỏ những lỗi đó.

Tạo hàm clean() để làm sạch và chuẩn hóa dữ liệu bao gồm các công việc:

- Nếu dữ liệu bị thiếu sẽ trả về chuỗi rỗng
- Sử dụng hàm lower() chuyển chữ hoa thành chữ thường để tránh hiểu sai giữa '8GB' và '8gb'. Hàm strip() để xóa đi khoảng trắng ở đầu và cuối của chuỗi.
- Để tránh các ký tự gây nhiễu khi xử lý văn bản, sử dụng hàm sub() của thư viện 're' để xóa. Dòng `[\w\s.]` ở đây sẽ xóa tất cả mọi ký tự không phải là chữ cái, chữ số, dấu gạch dưới, khoảng trắng, dấu chấm. Các ký tự không thuộc nhóm trên sẽ bị xóa như là !@#\$%^&.... và sẽ được thay thế bằng dấu cách.
- Để tránh trường hợp khi thực hiện vector hóa bị nhận diện sai, ví dụ như dữ liệu có giá trị là '16 GB' khi vector hóa sẽ bị tách thành hai từ là '16' và 'GB', để tránh việc này sẽ tiến hành gộp các số có đơn vị theo sau nhưng bị tách bởi dấu cách. Đầu tiên sẽ liệt kê ra các đơn vị có thể xuất hiện liên quan tới bộ dữ liệu 'gb', 'tb', 'mb', 'ghz', 'mhz', 'hz', 'wh', 'inch', 'cell', tìm trong bộ dữ liệu các số và đơn vị để ghép chúng lại với nhau.
- Giữa những từ trong bộ dữ liệu có thể bị tách với nhau bởi 2 hoặc nhiều dấu cách, để gọn gàng hơn sẽ xóa đi những khoảng trống thừa đó.

Áp dụng hàm clean() cho các cột 'Thương hiệu', 'Đánh giá', 'RAM', 'Loại RAM', 'Ổ cứng', 'Công nghệ CPU', 'Tốc độ CPU', 'Màn hình', 'Độ phân giải', 'Card màn hình', 'Chất liệu', 'Thông tin Pin', 'Hệ điều hành' nhưng không có cột 'Tên sp' và 'Giá' vì 2 cột này sẽ chứa tên nhận diện của sản phẩm, đặc điểm phân biệt và hàm này áp dụng cho các cột có giá trị dạng văn bản nên không cần thiết để áp dụng cho cột số.

- **Cột tổng hợp đặc trưng**

Sau khi thực hiện tất cả việc làm sạch sẽ tạo một cột mới có tên là 'Tag'. Cột này được tạo nên bởi việc gộp tất các giá trị trong dòng của các cột đã chọn thành một chuỗi văn bản tổng hợp tất cả các đặc điểm của sản phẩm. Giai đoạn này rất cần thiết cho các bước vector hóa, nếu trước khi thực hiện vector hóa không gộp các cột cần thiết lại, áp dụng vectorizer lên từng cột riêng lẻ sẽ khó hiểu được tất cả thông tin cũng như sẽ không thể hiện được mối liên hệ giữa các đặc trưng. Vì vậy việc gộp các cột là cần thiết cho các bước về sau.

```
data.iloc[10]['Tag']
```

```
'acer rất tốt 8gb ddr4 2 khe 1 khe 8gb 1 khe rời 512gb ssd nvme pcie có thể tháo ra lắp thanh khác
tối đa 1tb intel core i5 alder lake 1235u 1.3ghz 15.6 full hd 1920 x 1080 card tích hợp intel uhd
graphics iris xe graphics chỉ hoạt động với ram kênh đôi vỏ nhựa 3cell 40wh windows 11 home sl'
```

Hình 4.21 Giá trị cột Tag

Đây là kết quả của công đoạn trên, như đã thấy giá trị đã được làm sạch tất cả chữ hoa, khoảng trắng, ký tự.. và được nối lại với nhau thành một đoạn văn bản.

4.3.2. Ma trận đặc trưng

- **Dataframe mới**

	Tên sp	Giá	Tag
0	Laptop Acer Aspire 3 A314 42P R3B3 R7 5700U/16...	12490000	acer máy để chế độ sleep khoảng 12 tiếng sẽ bị...
1	Laptop Acer Aspire 3 A314 42P R3B3 R7 5700U/16...	12490000	acer mua sản phẩm vào tháng 09 2024 đến tháng ...

Hình 4.22 Tạo dataframe mới

Tạo dataframe mới tên là ‘data_new’ với 3 cột ‘Tên sp’, ‘Giá’, ‘Tag’. Đây là các thông tin cần thiết để hiển thị ra khi đề xuất, sẽ không giữ tất cả các cột để gọn hơn.

- **Đọc file stopwords**

```
with open('vietnamese-stopwords.txt', 'r', encoding = 'utf-8') as f:  
    vn_stopwords = [line.strip() for line in f if line.strip()]  
    vn_stopwords = [word.replace(' ', '_') for word in vn_stopwords]
```

Hình 4.23 Đọc file stopwords

Trước khi thực hiện Vectorizer, sẽ thực hiện loại bỏ từ dừng. Dữ liệu vẫn đang còn chứa những từ không mang nhiều ý nghĩa như ‘và’, ‘là’, ‘có’,... những từ này có thể gây nhiễu dẫn đến ma trận TF-IDF sẽ thưa thớt, tăng kích thước ma trận, tốn thời gian xử lý và khi tính toán độ tương đồng sẽ không chính xác.

Tiến hành mở file từ điển, trong từ điển sẽ chứa các từ dừng như ‘và’, ‘trong’, ‘có’, ‘sẽ’,.... Duyệt từng dòng bỏ qua các dòng rỗng và đưa vào danh sách tên là ‘vn_stopwords’ vì TfidfVectorizer không chấp nhận đường dẫn file.

- **Vector hóa dữ liệu**

```
array([[0.          , 0.          , 0.          , ..., 0.04483891, 0.          ,  
       0.          ],  
       [0.          , 0.          , 0.1284892 , ..., 0.          , 0.          ,  
       0.          ],  
       [0.          , 0.          , 0.          , ..., 0.          , 0.          ,  
       0.          ],  
       ...,  
       [0.          , 0.          , 0.          , ..., 0.15292327, 0.          ,  
       0.          ],  
       [0.          , 0.          , 0.          , ..., 0.          , 0.          ,  
       0.          ],  
       [0.          , 0.          , 0.          , ..., 0.          , 0.          ,  
       0.          ]])
```

Hình 4.24 Dữ liệu sau khi vector hóa

Sau khi đọc vào file stopword và loại bỏ các từ dừng, tiến hành biến đổi chuỗi văn bản thành vector số. TF-IDF sẽ tách chuỗi văn bản thành những token riêng biệt có ý nghĩa. Sau đó sẽ tính TF – tần suất thuật ngữ, tức là sẽ tính tần số xuất hiện của từng token trong các dòng dữ liệu ở cột ‘Tag’, chỉ số TF càng cao chứng tỏ từ đó xuất hiện nhiều lần trong dữ liệu. Tiếp theo sẽ tính IDF – ước

lượng tầm quan trọng của thuật ngữ, IDF càng cao thì từ đó sẽ chỉ xuất hiện ở một vài dòng, những từ này sẽ mang các đặc trưng để phân biệt dữ liệu, ngược lại IDF càng thấp thì đây là các từ xuất hiện ở nhiều trong dữ liệu thường không mang nhiều ý nghĩa. Cuối cùng sẽ tính TF – IDF, nếu cả hai đều cao thì TF – IDF sẽ cao chứng tỏ từ đó có tầm quan trọng lớn, phân biệt dữ liệu và không phải là những từ phổ biến.

Khi thực hiện Vectorize xong sẽ cho ra được ma trận chứa vector TF – IDF và mỗi cột tương ứng với một từ trong từ điển. Ma trận này sẽ dùng cho việc tính toán độ tương đồng giữa các vector.

- **Tính độ tương đồng**

```
array([[1.          , 0.14747765, 0.04711437, ... , 0.03878998, 0.03477464,
       0.07024941],
      [0.14747765, 1.          , 0.04964731, ... , 0.09442679, 0.03598281,
       0.12043372],
      [0.04711437, 0.04964731, 1.          , ... , 0.01945393, 0.03488233,
       0.02481191],
      ... ,
      [0.03878998, 0.09442679, 0.01945393, ... , 1.          , 0.29729169,
       0.54143819],
      [0.03477464, 0.03598281, 0.03488233, ... , 0.29729169, 1.          ,
       0.37917147],
      [0.07024941, 0.12043372, 0.02481191, ... , 0.54143819, 0.37917147,
       1.          ]])
```

Hình 4.25 Độ tương đồng dữ liệu

Sử dụng ma trận Cosine Similarity để tính độ tương đồng từng vector một với nhau, độ tương đồng giữa các sản phẩm với nhau. Ma trận sẽ tính tích vô hướng giữa hai vector chia cho tích độ dài hai vector.

$$\text{Similarity}(A, B) = \frac{\vec{A} \times \vec{B}}{|\vec{A}| \times |\vec{B}|}$$

Nếu Cosine similarity càng gần 1 thì hai càng giống nhau, tức là có thể nói hai sản phẩm tương đồng nhau, càng gần 0 thì hai sản phẩm sẽ khác nhau.

4.3.3. *Đưa ra gợi ý*

- **Tạo hàm xử lý đầu vào**

```

def query_tag(des):
    t = []
    for col in columns:
        feature = des.get(col, '')
        feature = clean(feature)
        t.append(feature)
    return ' '.join(t)

```

Hình 4.26 Hàm xử lý đầu vào

Tạo một hàm mới có tên là query_tag(des) để xử lý các mô tả mới mà người dùng đưa vào. Các mô tả sẽ được đưa vào chuỗi rỗng ‘t’, duyệt các cột qua các cột đã chọn lúc ban đầu, nếu mô tả có cột nào thì lấy giá trị cột đó không có thì sẽ trả về rỗng. Sau đó cũng sử dụng hàm clean() để làm sạch mô tả nhập vào đảm bảo định dạng, thêm vào danh sách ‘t’.

- **Tạo hàm gợi ý**

```

def recommend(des):
    query = query_tag(des)
    query_vec = tfidf.transform([query]).toarray()
    similarity_scores = cosine_similarity(query_vec, vector)[0]
    product = similarity_scores.argsort()[:-1][1:10]
    print('Các sản phẩm tương tự: ')
    for i in product:
        name = data_new.iloc[i]['Tên sp'].title()
        price = data_new.iloc[i]['Giá']
        print(f'{name} - Giá: {price}')

```

Hình 4.27 Hàm gợi ý sản phẩm

Sau khi tạo hàm để xử lý cho các mô tả mới, tạo một hàm để xuất là recommend(des). ‘query’ là một chuỗi văn bản được gộp lại từ mô tả của sản phẩm, giống với kiểu của cột ‘Tag’.

Giống như vector hóa cho cột ‘Tag’, ‘query’ cũng sử dụng TF – IDF để vector hóa về số học với tên mới là ‘query_vec’.

Sau khi thực hiện Vectorizer cho cột ‘query’ sẽ đi tính độ tương đồng giữa sản phẩm mới với các sản phẩm trong bộ dữ liệu. Kết quả của ‘similarity_scores’ sẽ là một mảng chứa độ tương đồng đã tính được.

Để in ra được các sản phẩm có độ tương đồng cao thì sẽ thực hiện việc sắp xếp lại độ tương đồng. Sử dụng hàm argsort() hàm này sẽ trả về index của các

phần tử trong mảng đã được sắp xếp từ thấp đến cao theo độ tương đồng. Vì argsort() chỉ sắp xếp tăng dần, nên phải thực hiện đảo ngược mảng [::-1], kết quả trả về là độ tương đồng giảm dần được lưu trong ‘product’.

Sẽ cho in ra 10 đề xuất có độ tương đồng gần nhất với mô tả sản phẩm mới, cho hiển thị ra cột thông tin như ‘Tên sp’, ‘Giá’, có thể nói đây là hai thông tin khách hàng quan tâm nhất khi quyết định mua sản phẩm.

Các sản phẩm tương tự:

Laptop Hp 15S Fq5162Tu I5 1235U/8Gb/512Gb/Win11 (7C134Pa) - Giá: 12990000
Laptop Hp 240 G9 I5 1235U/8Gb/512Gb/Win11 (6L1Y2Pa) - Giá: 12990000
Laptop Hp Pavilion 15 Eg2081Tu I5 1240P/16Gb/512Gb/Win11 (7C0Q4Pa) - Giá: 16890000
Laptop Hp Pavilion 15 Eg2081Tu I5 1240P/16Gb/512Gb/Win11 (7C0Q4Pa) - Giá: 16890000
Laptop Hp Pavilion 15 Eg2081Tu I5 1240P/16Gb/512Gb/Win11 (7C0Q4Pa) - Giá: 16890000
Laptop Hp Pavilion 15 Eg3091Tu I7 1355U/16Gb/512Gb/Win11 (8C5L2Pa) - Giá: 20990000
Laptop Hp Pavilion 15 Eg2081Tu I5 1240P/16Gb/512Gb/Win11 (7C0Q4Pa) - Giá: 16890000
Laptop Hp Gaming Victus 15 Fa1139Tx I5 12450H/16Gb/512Gb/4Gb Rtx2050/144Hz/Win11 (8Y6W3Pa) - Giá: 17690000
Laptop Hp Pavilion 15 Eg2081Tu I5 1240P/16Gb/512Gb/Win11 (7C0Q4Pa) - Giá: 16890000

Hình 4.28 Kết quả gợi ý

Đây là 10 sản phẩm có độ tương đồng gần nhất với sản phẩm mới có mô tả 'Thương hiệu': 'hp', 'Giá': '20000000', 'RAM': '16gb'.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

- **Đạt được:**

- Đề tài đã tìm hiểu các kỹ thuật học máy như Logistic Regression, Random Forests từ đó ứng dụng kỹ thuật này vào xây dựng mô hình khai phá dữ liệu nhằm dự đoán sự hài lòng của khách hàng.
- Đánh giá được hiệu quả của mô hình dự đoán thông qua các chỉ số như độ chính xác, độ nhạy, độ đặc trưng, v.v. Các mô hình học máy đều cho kết quả dự đoán ở mức khá về mức độ hài lòng của khách hàng, với độ chính xác 74,60% đối với mô hình Random Forests. Kết quả phân tích này cũng góp phần hỗ trợ doanh nghiệp chủ động nắm bắt và đáp ứng nhu cầu của khách hàng, từ đó giảm tỷ lệ khách hàng không hài lòng.
- Thực hiện được việc phân cụm sản phẩm thành các cụm, mỗi cụm sẽ phù hợp với một nhóm khách hàng. Hỗ trợ doanh nghiệp biết được mỗi cụm sẽ gồm những sản phẩm có đặc điểm ra sao, từ đó đưa ra các thông tin marketing cho từng nhóm sản phẩm. Giúp khách hàng có thể dễ dàng lựa chọn sản phẩm phù hợp với nhu cầu của bản thân.
- Xây dựng được mô hình gợi ý sản phẩm cho khách hàng, gợi ý các sản phẩm có đặc điểm tương đồng với mong muốn về đặc điểm kỹ thuật của sản phẩm giúp khách hàng có thể tìm kiếm được sản phẩm ưng ý.

- **Hạn chế:**

- Mô hình dự đoán vẫn chưa đạt được độ chính xác 100%, vẫn còn một số trường hợp dự đoán không chính xác, do đó cần tiếp tục cải thiện các mô hình.
- Chưa thể phân tích sâu các yếu tố ảnh hưởng đến sự hài lòng của khách hàng.
- Các điểm trong cụm vẫn chưa được chặt chẽ, các cụm chưa tách biệt rõ ràng với nhau.

- **Hướng phát triển:**

- Tiếp tục nghiên cứu và cải tiến mô hình học máy để nâng cao độ chính xác và hiệu quả dự đoán.
- Ứng dụng nhiều kỹ thuật học máy khác nhau để khai thác tốt hơn về dữ liệu về vấn đề hài lòng của khách hàng giúp có cái nhìn chi tiết, đa chiều hơn.
- Mở rộng phạm vi nghiên cứu, thu thập thêm dữ liệu về hành vi và nhu cầu của khách hàng để phân tích sâu hơn các yếu tố ảnh hưởng, phân biệt rõ ràng từng cụm với nhau.
- Xây dựng một trang web để đưa ra các đề xuất cho khách hàng với giao diện dễ thao tác hơn.

TÀI LIỆU THAM KHẢO

- [1] V. -. IT, "VNPT - IT," [Online]. Available: <https://vnptit.vn/>.
- [2] TOPDev, "Data Analyst là gì? Khám phá công việc của Data Analyst," [Online]. Available: <https://topdev.vn/blog/data-analyst-la-gi/>.
- [3] IZISOLUTION, "Các kỹ năng cần thiết của một Data Analyst," 2023. [Online]. Available: <https://izisolution.vn/cac-ky-nang-can-thiet-cua-mot-data-analyst/>.
- [4] T. t. n. tạo, "Bài 6: Logistic Regression (Hồi quy Logistic)," [Online]. Available: <https://trituenhantao.io/machine-learning-co-ban/bai-6-logistic-regression-hoi-quy-logistic/>.
- [5] ScholarHub, "Random forest là gì? Các công bố khoa học về Random forest," [Online]. Available: <https://scholarhub.vn/topic/random%20forest>.
- [6] T. t. n. tạo, "Bài 4: K-means Clustering," [Online]. Available: <https://trituenhantao.io/machine-learning-co-ban/bai-4-k-means-clustering/>.
- [7] N. T. Hop, "Introduction to Recommender Systems," 2020. [Online]. Available: <https://viblo.asia/p/introduction-to-recommender-systems-aWj53LQ8K6m>.
- [8] P.C.A.Huy, "Trí tuệ nhân tạo: Các phương pháp đánh giá một mô hình phân loại," 2021. [Online]. Available: <https://tapit.vn/cac-phuong-phap-danh-gia-mot-mo-hinh-phan-loai/>.
- [9] Pum, "Jupyter Notebook là gì? Hướng dẫn cài đặt và sử dụng

Jupyter Notebook," 2022. [Online]. Available:
<https://200lab.io/blog/jupyter-notebook-la-gi>.

- [10] N. Liên, "Python là gì? Tổng hợp tất tần tật kiến thức về ngôn ngữ Python có thể bạn chưa biết," 2024. [Online]. Available: <https://fptshop.com.vn/tin-tuc/danh-gia/python-la-gi-168825>.
- [11] aws, "Python là gì?", [Online]. Available: <https://aws.amazon.com/vi/what-is/python/>.
- [12] L. H. Hạnh, "Power BI là gì? Tìm hiểu cách sử dụng Power BI cho doanh nghiệp," 2024. [Online]. Available: <https://base.vn/blog/power-bi-la-gi/>.
- [13] T. Nguen, "Power BI là gì? Tại sao các doanh nghiệp nên sử dụng Power BI," 2023. [Online]. Available: <https://fptshop.com.vn/tin-tuc/danh-gia/power-bi-la-gi-tai-sao-cac-doanh-nghiep-nen-su-dung-power-bi-146601>.

CHECK LIST CỦA BÁO CÁO

STT	Nội dung công việc	Có	Không	Ghi chú
1	Báo cáo được trình bày (định dạng) đúng với yêu cầu.			
2	Báo cáo có số lượng trang đáp ứng đúng yêu cầu (50-80 trang)			
3	Báo cáo trình bày được đầy đủ phần mở đầu			
4	Báo cáo trình bày được cơ sở lý thuyết phù hợp với nội dung của đề tài và yêu cầu			
5	Nội dung chính của đề tài được trình bày hợp lý như đặt vấn đề rõ ràng, giải quyết vấn đề và kết quả.			
6	Báo cáo có phần kết luận và hướng phát triển của đề tài (Kết luận về kết quả đề tài và kết quả của bản thân thu được qua quá trình thực tập tại Doanh nghiệp)			

PHỤ LỤC

1. Tiền xử lý dữ liệu

```
#import thư viện cần thiết
import numpy as np
import pandas as pd

#Đọc dữ liệu
data = pd.read_csv('data_laptop.csv')
data.head(5)

#Kiểm tra thông tin dữ liệu
data.info()
#Kiểm tra giá trị duy nhất của mỗi cột
data.nunique()
#Kiểm tra giá trị null của mỗi cột
data.isnull().sum()
#Kiểm tra sự trùng lặp của dữ liệu
data.duplicated().sum()

# Xác định kiểu dữ liệu của cột
column_data_types = data.dtypes
# Đếm các cột số và phân loại
numerical_count = 0
categorical_count = 0
for column_name, data_type in column_data_types.items():
    if np.issubdtype(data_type, np.number):
        numerical_count += 1
    else:
        categorical_count += 1
print(f"Trong bộ dữ liệu có :")
print(f"Có {numerical_count} cột biến số")
print(f"Có {categorical_count} cột biến phân loại")

#Loại bỏ các cột thiếu nhiều dữ liệu và trống dữ liệu
df_clean = data.drop(['Link', 'Thông tin', 'NPU', 'Hiệu năng xử lý AI (TOPS)', 'Tính năng khác',
                      'Màn hình cảm ứng', 'Độ sáng SDR', 'DTS', 'Tần nhiệt', 'Khe đọc thẻ nhớ', 'Độ phủ màu',
                      'Kích thước', 'Dày(mm)'], axis=1)

df = df_clean.dropna()

df.shape

(793, 34)

df.isnull().sum()

#Thống kê mô tả
df.describe()

#Xuất file đã xử lý
df.to_csv('D:/TTTN/laptop_dx1.csv', index=False, encoding="utf-8-sig")
```

2. Bảng chuyển đổi giá trị các cột

Bảng chuyển đổi giá trị cột RAM

Giá trị	Giá trị chuyển đổi
0	16 GB
1	24 GB
2	32 GB
3	8 GB

Bảng chuyển đổi giá trị cột Loại RAM

Giá trị	Giá trị chuyển đổi
0	DDR4 (Onboard 4 GB + 1 khe 4 GB)
1	DDR4 (Onboard)
2	DDR4 2 khe (1 khe 16 GB + 1 khe 16 GB)
3	DDR4 2 khe (1 khe 16 GB + 1 khe rời)
4	DDR4 2 khe (1 khe 4 GB + 1 khe 4 GB)
5	DDR4 2 khe (1 khe 8 GB + 1 khe 8 GB)
6	DDR4 2 khe (1 khe 8 GB + 1 khe rời)
7	DDR4 2 khe (1 khe 8 GB onboard + 1 khe trống)
8	DDR4 2 khe (8 GB onboard + 1 khe 8 GB)
9	DDR5 (1 khe RAM)
10	DDR5 2 khe (1 khe 12 GB + 1 khe 12 GB)
11	DDR5 2 khe (1 khe 16 GB + 1 khe rời)
12	DDR5 2 khe (1 khe 8 GB + 1 khe 8 GB)

13	DDR5 2 khe (1 khe 8 GB + 1 khe trống)
14	Hàng không công bố
15	LPDDR4X (Onboard)
16	LPDDR5
17	LPDDR5 (Onboard)
18	LPDDR5X (Onboard)

Bảng chuyển đổi giá trị cột Tốc độ Bus RAM

Giá trị	Giá trị chuyển đổi
0	2666 MHz
1	2933 MHz
2	3200 MHz
3	4266 MHz
4	4800 MHz
5	5200 MHz
6	5500 MHz
7	5600 MHz
8	6400 MHz
9	7467 MHz
10	8522 MHz
11	Hàng không công bố
12	Từ 2400 MHz (Hàng công bố)

Bảng chuyển đổi giá trị cột Hỗ trợ RAM tối đa

Giá trị	Giá trị chuyển đổi
0	16 GB
1	20 GB
2	24 GB
3	32 GB
4	40 GB
5	64 GB
6	96 GB
7	Hỗ trợ không công bố
8	Không hỗ trợ nâng cấp

Bảng chuyển đổi giá trị cột Ổ cứng

Giá trị	Giá trị chuyển đổi
0	1 TB SSD
1	1 TB SSD M.2 PCIe
2	1 TB SSD NVMe PCIe Gen 4
3	256 GB SSD
4	256 GB SSD NVMe PCIe
5	512 GB SSD
6	512 GB SSD NVMe PCIe

7	512 GB SSD NVMe PCIe (Có thẻ tháo ra, lắp thanh khác tối đa 1 TB)
8	512 GB SSD NVMe PCIe (Có thẻ tháo ra, lắp thanh khác tối đa 2 TB (2280) / 1 TB (2230))
9	512 GB SSD NVMe PCIe (Có thẻ tháo ra, lắp thanh khác tối đa 2 TB)
10	512 GB SSD NVMe PCIe 4.0 (Có thẻ tháo ra, lắp thanh khác tối đa 1 TB)
11	512 GB SSD NVMe PCIe Gen 4 (Có thẻ tháo ra, lắp thanh khác tối đa 2 TB)
12	512 GB SSD NVMe PCIe Gen 4.0
13	512 GB SSD NVMe PCIe Gen 4.0 (Có thẻ tháo ra, lắp thanh khác tối đa 1 TB (2280) / 512 GB (2242))
14	512 GB SSD NVMe PCIe SED (Có thẻ tháo ra, lắp thanh khác nâng cấp tối đa 2 TB)

Bảng chuyển đổi giá trị cột Công nghệ CPU

Giá trị	Giá trị chuyển đổi
0	AMD Ryzen 5 - 7430U
1	AMD Ryzen 5 - 7520U
2	AMD Ryzen 5 - 7535HS
3	AMD Ryzen 7 - 5700U
4	AMD Ryzen 7 - 7435HS
5	AMD Ryzen 7 - 7735HS

6	AMD Ryzen 7 - 8845HS
7	Apple M1
8	Apple M2
9	Apple M4 Pro - Hãng không công bố
10	Intel Core Ultra 5 Meteor Lake - 125H
11	Intel Core Ultra 7 Lunar Lake - 258V
12	Intel Core Ultra 7 Meteor Lake - 155H
13	Intel Core Ultra 7 Meteor Lake - 155U
14	Intel Core Ultra 9 Meteor Lake - 185H
15	Intel Core i3 Alder Lake - 1215U
16	Intel Core i3 Raptor Lake - 1315U
17	Intel Core i3 Tiger Lake - 1115G4
18	Intel Core i5 Alder Lake - 1235U
19	Intel Core i5 Alder Lake - 1240P
20	Intel Core i5 Alder Lake - 12450H
21	Intel Core i5 Alder Lake - 12450HX
22	Intel Core i5 Alder Lake - 12500H
23	Intel Core i5 Raptor Lake - 1334U
24	Intel Core i5 Raptor Lake - 1335U
25	Intel Core i5 Raptor Lake - 13420H
26	Intel Core i5 Raptor Lake - 13450HX

27	Intel Core i5 Raptor Lake - 13500H
28	Intel Core i5 Raptor Lake - 13500HX
29	Intel Core i5 Tiger Lake - 1135G7
30	Intel Core i5 Tiger Lake - 1155G7
31	Intel Core i7 Alder Lake - 1255U
32	Intel Core i7 Alder Lake - 1260P
33	Intel Core i7 Alder Lake - 12650H
34	Intel Core i7 Alder Lake - 12700H
35	Intel Core i7 Raptor Lake - 1355U
36	Intel Core i7 Raptor Lake - 13620H
37	Intel Core i7 Raptor Lake - 13700H
38	Intel Core i7 Raptor Lake - 14700HX
39	Ryzen 5

Bảng chuyển đổi giá trị cột Tốc độ CPU

Giá trị	Giá trị chuyển đổi
0	1.2GHz
1	1.3GHz
2	1.4GHz
3	1.7GHz
4	1.8GHz
5	100GB/ s

6	2.1GHz
7	2.2GHz
8	2.3GHz
9	2.4GHz
10	2.5GHz
11	2.6GHz
12	2.8GHz
13	273 GB/s memory bandwidth
14	2GHz
15	3.1GHz
16	3.2GHz
17	3.3GHz
18	3.8GHz
19	3GHz
20	Hãng không công bố

Bảng chuyển đổi giá trị cột Card màn hình

Giá trị	Giá trị chuyển đổi
0	Card rời - AMD Radeon RX 6550M, 4 GB
1	Card rời - NVIDIA GeForce RTX 2050, 4 GB
2	Card rời - NVIDIA GeForce RTX 3050, 4 GB
3	Card rời - NVIDIA GeForce RTX 3050, 6 GB

4	Card rời - NVIDIA GeForce RTX 3050Ti, 4 GB
5	Card rời - NVIDIA GeForce RTX 4050, 6 GB
6	Card rời - NVIDIA GeForce RTX 4060
7	Card rời - NVIDIA GeForce RTX 4060, 8 GB
8	Card rời - NVIDIA GeForce MX550 2 GB
9	Card tích hợp - 10 nhân GPU
10	Card tích hợp - 20 nhân GPU
11	Card tích hợp - 7 nhân GPU
12	Card tích hợp - 8 nhân GPU
13	Card tích hợp - AMD Radeon 610M Graphics
14	Card tích hợp - AMD Radeon Graphics
15	Card tích hợp - Intel Arc Graphics
16	Card tích hợp - Intel Graphics
17	Card tích hợp - Intel Iris Xe Graphics
18	Card tích hợp - Intel UHD Graphics
19	Card tích hợp - Intel UHD Graphics (Iris Xe Graphics chỉ hoạt động với RAM kênh đôi)
20	Card tích hợp - Intel Graphics
21	Card tích hợp - Intel Iris Xe Graphics

Bảng chuyển đổi giá trị cột Cổng giao tiếp

Giá trị	Giá trị chuyển đổi
0	HDMI
1	Jack tai nghe 3.5 mm
2	LAN (RJ45)
3	MagSafe 3
4	Thunderbolt 4 USB-C
5	USB 2.0
6	USB Type-C
7	USB Type-C (hỗ trợ truyền dữ liệu, Power Delivery 3.0 và DisplayPort 1.2)

Bảng chuyển đổi giá trị cột Kết nối không dây

Giá trị	Giá trị chuyển đổi
0	Bluetooth
1	Bluetooth 5.0
2	Bluetooth 5.2
3	Bluetooth 5.3
4	Wi-Fi 6 (802.11ax)
5	Wi-Fi 6E (802.11ac)
6	Wi-Fi 6E (802.11ax)
7	Wi-Fi 7 (802.11be)

8	Wi-Fi 802.11 a/b/g/n/ac
---	-------------------------

Bảng chuyển đổi giá trị cột Thông tin Pin

Giá trị	Giá trị chuyển đổi
0	3 cell, 41.04 Wh
1	3-Cell, 39.3 Wh
2	3-cell Li-ion, 52.4 Wh
3	3-cell Li-ion, 58 Wh
4	3-cell Li-ion, 59 Wh
5	3-cell, 40Wh
6	3-cell, 41Wh
7	3-cell, 42Wh
8	3-cell, 43Wh
9	3-cell, 48Wh
10	3-cell, 50Wh
11	3-cell, 51Wh
12	3-cell, 52.5Wh
13	3-cell, 53.5Wh
14	3-cell, 59 Wh
15	3-cell, 63Wh
16	36Wh
17	38Wh

18	4-cell Li-ion, 75 Wh
19	4-cell, 54Wh
20	4-cell, 56Wh
21	4-cell, 57.5Wh
22	4-cell, 65Wh
23	4-cell, 70Wh
24	4-cell, 72Wh
25	4-cell, 76Wh
26	4-cell, 90Wh
27	42 Wh
28	47 Wh
29	48Wh
30	56.6Wh
31	57Wh
32	60 Wh
33	71Wh
34	Khoảng 10 tiếng
35	Li-Po, 100 Wh
36	Lên đến 18 giờ

3. Chuẩn bị dữ liệu dự đoán

```

# Khai báo các thư viện cần thiết
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Đọc dữ liệu
from google.colab import files
uploaded = files.upload()
df = pd.read_csv('laptop_dxl.csv')
df

# Chuyển đổi cột dữ liệu của cột hài lòng
df['Hài lòng'] = df['Hài lòng'].astype('int64')

# Loại bỏ cột không cần thiết & sử dụng LabelEncoder để mã hóa các
cột có kiểu object
from sklearn.preprocessing import LabelEncoder
label_encoders = {}
df = df.drop(columns=['Tên khách hàng'], errors='ignore')
for col in df.select_dtypes(include='object').columns:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    label_encoders[col] = le

# Tạo bản sao DataFrame df lưu vào biến log
log=df.copy()

```

```

# Tính toán ma trận tương quan giữa các cột trong DataFrame log
log.corr()['Hài lòng'].sort_values(ascending = False)
plt.figure(figsize=(25, 25))
sns.heatmap(log.corr(), annot=True, fmt=".2f", cmap="coolwarm",
annot_kws={"size":15})

# Lọc các biến độc lập có mối tương quan đủ mạnh với biến Hài lòng
corr = log.corr()
cols = corr['Hài lòng'].abs()[corr['Hài lòng'].abs() > 0.05].index
data = log[cols]
print(f"Số biến được chọn: {len(cols)}")
data.head()

# Tách dữ liệu & Đếm số lượng các giá trị trong biến phụ thuộc
x = data.drop(['Hài lòng'], axis=1)
y = data[['Hài lòng']]
y.value_counts()

```

```

# Sử dụng SMOTE để cân bằng dữ liệu
from imblearn.over_sampling import SMOTE
smote = SMOTE()
x_smote, y_smote = smote.fit_resample(x,y)
print('Dữ liệu trước khi cân bằng:',y.value_counts())
print('Dữ liệu sau khi cân bằng:',y_smote.value_counts())

# Sử dụng MinMaxScaler để chuẩn hóa dữ liệu
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
cols = x_smote.select_dtypes(include=['float64', 'int64']).columns
x_scaled = x_smote.copy()
x_scaled[cols] = scaler.fit_transform(x_smote[cols])
x_scaled.sample(5)

# Tách dữ liệu chuẩn bị cho quá trình huấn luyện mô hình
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x_scaled,
y_smote, test_size = 0.2, random_state = 42, stratify = y_smote)
print(f"Số mẫu trong bộ train: {x_train.shape[0]}")
print(f"Số mẫu trong bộ test: {x_test.shape[0]}")

```

4. Huấn luyện mô hình Hồi quy Logistic

```

from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix,
classification_report
from sklearn.metrics import roc_curve, auc
model = LogisticRegression(max_iter=100,tol=1e-4)
model.fit(x_train, y_train)

# Độ chính xác trên tập dữ liệu kiểm
y_pred = model.predict(x_test)
test_accuracy = accuracy_score(y_pred, y_test)
print(f'Dộ chính xác của tập test: {test_accuracy*100:.2f}%')

# In ra classification report
print("Classification Report:\n", classification_report(y_test,
y_pred))

# Confusion Matrix - Test
cm_test = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(5,3))
sns.heatmap(cm_test, annot=True, fmt='g', cmap='Oranges')
plt.title('Confusion Matrix - Test Data')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()

```

```

# Vẽ đường cong ROC
y_pred_prob = model.predict_proba(x_test)[:, 1]
fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob)
roc_auc = auc(fpr, tpr)
print(f"ROC AUC: {roc_auc:.2f}")

plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=2, label='ROC curve
(area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.legend(loc="lower right")
plt.show()

# Giải thích hệ số mô hình
weights = model.coef_
intercept = model.intercept_
print("Hàm dự đoán:")
print(f"y_pred = sigmoid({intercept} + {weights} * X)")
# Dự đoán trên mẫu dữ liệu mới
X1 = {
    'Giá': 16990000,
    'Rating': 1,
    'RAM': 0,
    'Loại RAM': 1,
    'Tốc độ Bus RAM': 2,
    'Hỗ trợ RAM tối đa': 2,
    'Ổ cứng': 6,
    'Công nghệ CPU': 22,
    'Số nhân': 12,
    'Tốc độ CPU': 10,
    'Card màn hình': 2,
    'Cổng giao tiếp': 1,
    'Kết nối không dây': 11,
    'Thông tin Pin': 10,
    'Thời điểm ra mắt': 2024
}
X1_df = pd.DataFrame(X1, index=[0])
X1_scaled = scaler.transform(X1_df)
Y1 = model.predict(X1_df)
if Y1 == 0:
    print("Dự báo khách hàng X1 bằng mô hình Hồi Quy Logistic: 0
(Không hài lòng)")
else:
    print("Dự báo khách hàng X1 bằng mô hình Hồi Quy Logistic: 1
(Hài lòng)")

```

5. Huấn luyện với mô hình Random Forest

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt

depths = list(range(1, 16))
train_scores = []
test_scores = []

for depth in depths:
    rf_model = RandomForestClassifier(max_depth=depth,
random_state=42, n_estimators=100)
    rf_model.fit(x_train, y_train)

    # Dự đoán và tính độ chính xác trên tập huấn luyện
    train_pred = rf_model.predict(x_train)
    train_acc = accuracy_score(y_train, train_pred)
    train_scores.append(train_acc)

    # Dự đoán và tính độ chính xác trên tập kiểm tra
    test_pred = rf_model.predict(x_test)
    test_acc = accuracy_score(y_test, test_pred)
    test_scores.append(test_acc)

# Vẽ Learning Curve
plt.figure(figsize=(8, 5))
plt.plot(depths, train_scores, marker='o', label='Train Accuracy')
plt.plot(depths, test_scores, marker='o', label='Test Accuracy')
plt.xlabel('Max Depth')
plt.ylabel('Accuracy')
plt.title('Learning Curve for Random Forest (Tối ưu max_depth)')
plt.xticks(depths)
plt.legend()
plt.grid(True)
plt.show()
```

```

# Khởi tạo mô hình Random Forest với các tham số tối ưu
rf_model = RandomForestClassifier(criterion="entropy",
max_depth=8, n_estimators=99, random_state=42)

rf_model.fit(x_train, y_train)
y_pred_rf = rf_model.predict(x_test)
# Độ chính xác trên tập test
rf_test_accuracy = accuracy_score(y_pred_rf, y_test)
print(f'Dộ chính xác trên tập test: {rf_test_accuracy*100:.2f}%')

# In ra Classification report
print("Classification Report:\n", classification_report(y_test,
y_pred_rf))

# Confusion Matrix
cm_rf_test = confusion_matrix(y_test, y_pred_rf)
plt.figure(figsize=(5,3))
sns.heatmap(cm_rf_test, annot=True, fmt='g', cmap='Blues')
plt.title('Confusion Matrix Test Data')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()

# Vẽ đường cong ROC
from sklearn.metrics import roc_curve, roc_auc_score
import matplotlib.pyplot as plt

y_pred_prob_rf = rf_model.predict_proba(x_test)[:, 1]
fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob_rf)
roc_auc = roc_auc_score(y_test, y_pred_prob_rf)
print(f"ROC AUC: {roc_auc:.2f}")

plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=2, label='ROC curve
(area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC)')
plt.legend(loc="lower right")
plt.show()

```

6. K-means phân cụm dữ liệu

```

#Import thư viện cần thiết
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
np.random.seed(0)

#Đọc dữ liệu
data = pd.read_csv('Laptop_dx1.csv')
data

#Chuyển đổi dữ liệu
from sklearn.preprocessing import LabelEncoder

label_encoders = {}
for col in data.select_dtypes(include='object').columns:
    le = LabelEncoder()
    data[col] = le.fit_transform(data[col])
    label_encoders[col] = le

#Lấy cột cho phân cụm
corr = data.corr()
cols = corr['Hài lòng'].abs()[corr['Hài lòng'].abs() > 0.05].index
kmean = data[cols]
kmean.head()

#Chuẩn hóa dữ liệu
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
data_scaled = scaler.fit_transform(kmean)

#Tìm số cụm k
from sklearn.cluster import KMeans
from sklearn import metrics

inertias = []
silhouette_scores=[]
for k in range(2, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(data_scaled)
    inertias.append(kmeans.inertia_)

silhouette_score=metrics.silhouette_score(data_scaled, kmeans.labels_)
silhouette_scores.append(silhouette_score)

```

```

#Vẽ biểu đồ Elbow và Silhouette
plt.figure(figsize=(12, 5))

# Biểu đồ Elbow Method
plt.subplot(1, 2, 1)
plt.plot(range(2, 11), inertias, marker='o')
plt.title('Elbow Method')
plt.xlabel('Số cụm K')
plt.ylabel('Inertias')
plt.grid(True)

# Biểu đồ Silhouette Score
plt.subplot(1, 2, 2)
plt.plot(range(2, 11), silhouette_scores, marker='o', color='green')
plt.title('Silhouette Score')
plt.xlabel('Số cụm K')
plt.ylabel('Silhouette Score')
plt.grid(True)

plt.tight_layout()
plt.show()

```

```

#Phân cụm và đưa ra đặc điểm các cụm
kmeans=KMeans(n_clusters=6, random_state=42)
kmeans.fit(data_scaled)
data['Cluster']=kmeans.labels_

cluster_centers=scaler.inverse_transform(kmeans.cluster_centers_)
cluster_df=pd.DataFrame(cluster_centers, columns=['Giá', 'Rating', 'Hài lòng', 'RAM', 'Loại RAM', 'Tốc độ Bus RAM',
                                                 'Hỗ trợ RAM tối đa', 'Ổ cứng', 'Công nghệ CPU', 'Số nhân', 'Tốc độ CPU',
                                                 'Card màn hình', 'Cổng giao tiếp', 'Kết nối không dây', 'Thông tin Pin',
                                                 'Thời điểm ra mắt'])

print('Đặc điểm các cụm:')
print(cluster_df)

cluster_counts=pd.Series(kmeans.labels_).value_counts().sort_index()
print('Số điểm trong mỗi cụm:')
print(cluster_counts)

```

```

#Ánh xạ lại dữ liệu
print("ÁNH XẠ LABEL ENCODING CHO CÁC CỘT ĐÃ CHỌN\n")
selected_col = ['RAM', 'Loại RAM', 'Tốc độ Bus RAM',
                'Hỗ trợ RAM tối đa', 'Ổ cứng', 'Công nghệ CPU', 'Tốc độ CPU',
                'Card màn hình', 'Cổng giao tiếp', 'Kết nối không dây', 'Thông tin Pin']

for col in selected_col:
    if col in label_encoders:
        le = label_encoders[col]
        print(f"--- Mapping cho cột '{col}':")
        for i, label in enumerate(le.classes_):
            print(f"{i} → {label}")
        print("\n")
    else:
        print(f"Cột '{col}' không có trong label_encoders.\n")

```

```

#Chỉ số của k-means
from sklearn.metrics import silhouette_score

wss=kmeans.inertia_
print('WSS:', wss)

centroids=kmeans.cluster_centers_
mean=data_scaled.mean(axis=0)
bss=((centroids - mean)**2).sum()
print('BSS:', bss)

silhouette_avg = silhouette_score(data_scaled, kmeans.labels_)
print('Silhouette Score:', silhouette_avg)

```

7. Xây dựng mô hình gợi ý

```
#Import thư viện và đọc dữ liệu
import pandas as pd
import numpy as np
import re
from sklearn.preprocessing import StandardScaler
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

laptop=pd.read_csv('laptop_dxl.csv')
laptop.head(5)

#Chọn cột cần thiết
data = laptop[['Thương hiệu', 'Tên sp', 'Giá', 'Đánh giá', 'RAM', 'Loại RAM', 'Ổ cứng', 'Công nghệ CPU', 'Tốc độ CPU',
               'Màn hình', 'Độ phân giải', 'Card màn hình', 'Chất liệu', 'Thông tin Pin', 'Hệ điều hành']]
data.head(2)

#Hàm xử lý văn bản
def clean(text):
    if pd.isnull(text):
        return ''
    text = text.lower().strip()
    text = re.sub(r'^\w\s.', ' ', text, flags=re.UNICODE)
    for unit in ['gb', 'tb', 'mb', 'ghz', 'mhz', 'hz', 'wh', 'inch', 'cell']:
        text = re.sub(rf'(\d+(?:\.\d+)?)\s*{unit}', rf'\1{unit}', text)
    text = re.sub(r'\s+', ' ', text).strip()
    return text

#Áp dụng hàm
columns = ['Thương hiệu', 'Đánh giá', 'RAM', 'Loại RAM', 'Ổ cứng', 'Công nghệ CPU', 'Tốc độ CPU',
           'Màn hình', 'Độ phân giải', 'Card màn hình', 'Chất liệu', 'Thông tin Pin', 'Hệ điều hành']

for col in columns:
    data[col] = data[col].apply(clean)

#Tạo cột Tag
data['Tag'] = data[columns].apply(lambda row: ' '.join(row), axis=1)
data.iloc[10]['Tag']

#Tạo dataframe mới
data_new = data[['Tên sp', 'Giá', 'Tag']]
data_new.head(2)

#Đọc file từ đằng
with open('vietnamese-stopwords.txt', 'r', encoding = 'utf-8') as f:
    vn_stopwords = [line.strip() for line in f if line.strip()]
    vn_stopwords = [word.replace(' ', '_') for word in vn_stopwords]

#Thực hiện TF-IDF
tfidf = TfidfVectorizer(stop_words = vn_stopwords)
vector = tfidf.fit_transform(data_new['Tag']).toarray()
vector

#Tính độ tương đồng
from sklearn.metrics.pairwise import cosine_similarity
similarity_matrix = cosine_similarity(vector)
similarity_matrix
```

```

#Hàm xử lý đầu vào
def query_tag(des):
    t = []
    for col in columns:
        feature = des.get(col, '')
        feature = clean(feature)
        t.append(feature)
    return ' '.join(t)

#Hàm gợi ý
def recommend(des):
    query = query_tag(des)
    query_vec = tfidf.transform([query]).toarray()
    similarity_scores = cosine_similarity(query_vec, vector)[0]
    product = similarity_scores.argsort()[:-1][1:10]
    print('Các sản phẩm tương tự:')
    for i in product:
        name = data_new.iloc[i]['Tên sp'].title()
        price = data_new.iloc[i]['Giá']
        print(f'{name} - Giá: {price}')

```

```

#gợi ý sản phẩm
recommend({'Thương hiệu': 'hp',
           'Giá': '20000000',
           'RAM': '16gb'})

```

```

#gợi ý sản phẩm
recommend({'Thương hiệu': 'hp',
           'Giá': '20000000',
           'RAM': '16gb'})

```