TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI
KHOA: CÔNG NGHỆ THÔNG TIN

BÁO CÁO THỰC NGHIỆM
HỌC PHẦN: PHÂN TÍCH DỮ LIỆU LỚN

ĐỀ TÀI: PHÂN TÍCH DỰ ĐOÁN GIÁ MÁY TÍNH XÁCH TAY BẰNG MÔ HÌNH HỒI QUY TUYẾN TÍNH

Giảng viên hướng dẫn: TS. Nguyễn Mạnh Cường

Sinh viên thực hiện: Nguyễn Trung Hiếu - 2022600419

Vương Trí Tín - 2021603785

Trần Văn Trường - 2022607222

Nhóm: 17

Lớp: 20241IT6077002

Hà Nội, năm 2024

PHIẾU HỌC TẬP CÁ NHÂN/NHÓM

I. Thông tin chung

1. Tên lớp: 2022DHCNTT01. Khóa: 17.

2. Họ và tên sinh viên: **Nguyễn Trung Hiếu.** Mã sinh viên: 2022600419.

2. Tên nhóm: Nhóm 17.

II. Nội dung học tập

1. Tên chủ đề: Phân tích dự đoán giá máy tính xách tay bằng mô hình hồi quy tuyến tính.

2. Hoạt động của sinh viên:

- Hoạt động/Nội dung 1: *Thành lập nhóm học tập, lập kế hoạch làm bài tập lớn, thực hiện nghiên cứu phát biểu bài toán.*

Mục tiêu/chuẩn đầu ra: L1, L2, L4.

- Hoạt động/Nội dung 2: *Tìm hiểu các kỹ thuật phổ biến giải quyết bài toán. Tìm hiểu kỹ thuật chính sẽ sử dụng để thực nghiệm trong bài tập lớn.*

Mục tiêu/chuẩn đầu ra: L1, L2, L4.

- Hoạt động/Nội dung 3: *Tiến hành các bước thu thập, tiền xử lý dữ liệu; sử dụng các công cụ phù hợp để thực nghiệm; tổng hợp, so sánh, đánh giá kết quả. Xây dựng chương trình demo (nếu có), viết báo cáo bài tập lớn .*

3. Sản phẩm nghiên cứu: *Báo cáo thí nghiệm/ Thực nghiệm + Chương trình demo (nếu có).*

III. Nhiệm vụ học tập

1. Tổng hợp kiến thức đã học trong học phần, ứng dụng kiến thức, kỹ năng học được để giải quyết một bài toán thực tế liên quan tới phân tích dữ liệu.
2. Hoàn thành bài tập lớn theo đúng thời gian quy định (từ ngày 09/09/2024, đến ngày 22/12/2024).
3. Nộp bài và báo cáo sản phẩm theo chủ đề được giao trước giảng viên và những sinh viên khác.

IV. Học liệu thực hiện Tiểu luận, Bài tập lớn, Bài tập lớn/Dự án

1. Tài liệu học tập: Các tài liệu hướng dẫn thực hiện bài tập lớn do giảng viên cung cấp, các tài liệu, code mẫu tham khảo trong bài giảng và trên mạng internet.

2. Phương tiện, nguyên liệu thực hiện bài tập lớn: sử dụng các công cụ phù hợp: Excel, Weka, PyCharm, Anaconda, Jupiter Notebook, Google Collab, R, Apache Hadoop, Apache Spark…

PHIẾU HỌC TẬP CÁ NHÂN/NHÓM

I. Thông tin chung

1. Tên lớp: 2021DHKHMT01. Khóa: 16.

2. Họ và tên sinh viên: **Vương Trí Tín.** Mã sinh viên: 2021603785.

2. Tên nhóm: Nhóm 17.

II. Nội dung học tập

1. Tên chủ đề: Phân tích dự đoán giá máy tính xách tay bằng mô hình hồi quy tuyến tính.

2. Hoạt động của sinh viên:

- Hoạt động/Nội dung 1: *Thành lập nhóm học tập, lập kế hoạch làm bài tập lớn, thực hiện nghiên cứu phát biểu bài toán.*

Mục tiêu/chuẩn đầu ra: L1, L2, L4.

- Hoạt động/Nội dung 2: *Tìm hiểu các kỹ thuật phổ biến giải quyết bài toán. Tìm hiểu kỹ thuật chính sẽ sử dụng để thực nghiệm trong bài tập lớn.*

Mục tiêu/chuẩn đầu ra: L1, L2, L4.

- Hoạt động/Nội dung 3: *Tiến hành các bước thu thập, tiền xử lý dữ liệu; sử dụng các công cụ phù hợp để thực nghiệm; tổng hợp, so sánh, đánh giá kết quả. Xây dựng chương trình demo (nếu có), viết báo cáo bài tập lớn .*

3. Sản phẩm nghiên cứu: *Báo cáo thí nghiệm/ Thực nghiệm + Chương trình demo (nếu có).*

III. Nhiệm vụ học tập

1. Tổng hợp kiến thức đã học trong học phần, ứng dụng kiến thức, kỹ năng học được để giải quyết một bài toán thực tế liên quan tới phân tích dữ liệu.
2. Hoàn thành bài tập lớn theo đúng thời gian quy định (từ ngày 09/09/2024, đến ngày 22/12/2024).
3. Nộp bài và báo cáo sản phẩm theo chủ đề được giao trước giảng viên và những sinh viên khác.

IV. Học liệu thực hiện Tiểu luận, Bài tập lớn, Bài tập lớn/Dự án

1. Tài liệu học tập: Các tài liệu hướng dẫn thực hiện bài tập lớn do giảng viên cung cấp, các tài liệu, code mẫu tham khảo trong bài giảng và trên mạng internet.

2. Phương tiện, nguyên liệu thực hiện bài tập lớn: sử dụng các công cụ phù hợp: Excel, Weka, PyCharm, Anaconda, Jupiter Notebook, Google Collab, R, Apache Hadoop, Apache Spark…

PHIẾU HỌC TẬP CÁ NHÂN/NHÓM

I. Thông tin chung

1. Tên lớp: 2022DHKHMT01. Khóa : 17.

2. Họ và tên sinh viên: **Trần Văn Trường.** Mã sinh viên: 2022607222.

2. Tên nhóm: Nhóm 17.

II. Nội dung học tập

1. Tên chủ đề: Phân tích dự đoán giá máy tính xách tay bằng mô hình hồi quy tuyến tính.

2. Hoạt động của sinh viên:

- Hoạt động/Nội dung 1: *Thành lập nhóm học tập, lập kế hoạch làm bài tập lớn, thực hiện nghiên cứu phát biểu bài toán.*

Mục tiêu/chuẩn đầu ra: L1, L2, L4.

- Hoạt động/Nội dung 2: *Tìm hiểu các kỹ thuật phổ biến giải quyết bài toán. Tìm hiểu kỹ thuật chính sẽ sử dụng để thực nghiệm trong bài tập lớn.*

Mục tiêu/chuẩn đầu ra: L1, L2, L4.

- Hoạt động/Nội dung 3: *Tiến hành các bước thu thập, tiền xử lý dữ liệu; sử dụng các công cụ phù hợp để thực nghiệm; tổng hợp, so sánh, đánh giá kết quả. Xây dựng chương trình demo (nếu có), viết báo cáo bài tập lớn .*

3. Sản phẩm nghiên cứu: *Báo cáo thí nghiệm/ Thực nghiệm + Chương trình demo (nếu có).*

III. Nhiệm vụ học tập

1. Tổng hợp kiến thức đã học trong học phần, ứng dụng kiến thức, kỹ năng học được để giải quyết một bài toán thực tế liên quan tới phân tích dữ liệu.
2. Hoàn thành bài tập lớn theo đúng thời gian quy định (từ ngày 09/09/2024, đến ngày 22/12/2024).
3. Nộp bài và báo cáo sản phẩm theo chủ đề được giao trước giảng viên và những sinh viên khác.

IV. Học liệu thực hiện Tiểu luận, Bài tập lớn, Bài tập lớn/Dự án

1. Tài liệu học tập: Các tài liệu hướng dẫn thực hiện bài tập lớn do giảng viên cung cấp, các tài liệu, code mẫu tham khảo trong bài giảng và trên mạng internet.

2. Phương tiện, nguyên liệu thực hiện bài tập lớn: sử dụng các công cụ phù hợp: Excel, Weka, PyCharm, Anaconda, Jupiter Notebook, Google Collab, R, Apache Hadoop, Apache Spark…

KẾ HOẠCH THỰC HIỆN TIỂU LUẬN, BÀI TẬP LỚN, ĐỒ ÁN

Tên lớp: 20241IT6077002 Khóa: K16-K17

Họ tên thành viên của nhóm: 1) Nguyễn Trung Hiếu; 2) Vương Trí Tín; 3) Trần Văn Trường

Tên nhóm: Nhóm 17

Tên chủ đề: Phân tích dự đoán giá máy tính xách tay bằng mô hình hồi quy tuyến tính.

Tuần	Người thực hiện	Nội dung công việc	Kết quả đạt được	Phương pháp thực hiện
1 - 3	Nguyễn Trung Hiếu	Chọn đề tài, tìm kiếm dữ liệu	Đề tài: Phân tích dự đoán giá máy tính xách tay bằng mô hình hồi quy tuyến tính Bộ dữ liệu máy tính xách tay thu thập trên sàn Amazon	Sưu tầm tài liệu và dữ liệu
	Vương Trí Tín			
	Trần Văn Trường			
4	Nguyễn Trung Hiếu	Giới thiệu tổng quan về đề tài, các chương, kỳ vọng và kết quả	Thực hiện được mở đầu báo cáo	Sưu tầm tài liệu Nghiên cứu tài liệu
	Vương Trí Tín	Tổng quan về bài toán dự báo và tình hình nghiên cứu	Tổng quan bài toán dự báo, tình hình nghiên cứu quốc tế và ở Việt Nam	Sưu tầm tài liệu Nghiên cứu tài liệu
	Trần Văn Trường	Giới thiệu về bài toán dự đoán giá máy tính xách tay	Giới thiệu được bài toán, mô tả đầu vào và đầu ra cùng khó khăn, miền ứng dụng bài toán	Sưu tầm tài liệu Nghiên cứu tài liệu
5	Nguyễn Trung Hiếu	Tìm hiểu các phương pháp giải quyết bài toán và lựa chọn phương pháp	Các phương pháp phân tích hồi quy được đưa ra và hồi quy tuyến tính	Sưu tầm tài liệu Nghiên cứu tài liệu

			được chọn để giải quyết	Tổng hợp/Đánh giá
	Vương Trí Tín	Tìm hiểu tổng quan về mô hình hồi quy tuyến tính	Tổng quan, ưu và nhược điểm của mô hình hồi quy tuyến tính với bài toán	Sưu tầm tài liệu Nghiên cứu tài liệu Tổng hợp/Đánh giá
	Trần Văn Trường	Lựa chọn công cụ hỗ trợ giải quyết bài toán	Công cụ python cùng các thư viện liên quan được lựa chọn	Sưu tầm tài liệu Nghiên cứu tài liệu Tổng hợp/Đánh giá
6	Nguyễn Trung Hiếu	Mô tả bộ dữ liệu và làm sạch	Các cột đầu vào kiểu chữ (mang bản chất số) được chuyển về dạng số	Thực nghiệm Tổng hợp/Đánh giá
	Vương Trí Tín	Thực hiện tóm lược dữ liệu	Thực hiện thống kê mô tả các cột dữ liệu sau khi được làm sạch	Thực nghiệm Tổng hợp/Đánh giá
	Trần Văn Trường	Thực hiện chuyển đổi dữ liệu	Ánh xạ các cột thích hợp từ kiểu phi số thành kiểu số	Thực nghiệm Tổng hợp/Đánh giá

7	Nguyễn Trung Hiếu	Phân tích mô tả bộ dữ liệu sau tiền xử lý	Thấy được mức tương quan các thuộc tính so với thuộc tính đích	Thực nghiệm Tổng hợp/Đánh giá
	Vương Trí Tín	Chỉnh sửa và nhận xét các về mức độ ảnh hưởng của các thuộc tính so với thuộc tính giá máy tính	Nhận xét mức độ tương quan và giải thích	Thực nghiệm Tổng hợp/Đánh giá
	Trần Văn Trường			
8	Nguyễn Trung Hiếu	Tạo và huấn luyện mô hình hồi quy	Dữ liệu được huấn luyện và đưa ra các thông số của mô hình	Thực nghiệm
	Vương Trí Tín	Phân tích các thông số của mô hình, ưu và nhược điểm từ đó đưa ra đánh giá nhận xét	Bảng kết quả các lần chạy được lập, từ đó giải thích được hiệu suất của mô hình	Tổng hợp/Đánh giá
	Trần Văn Trường			Thực nghiệm
9 - 10	Nguyễn Trung Hiếu	Kết luận sau quá trình thực nghiệm, chỉnh sửa và bổ sung báo cáo	Quyển báo cáo thực nghiệm hoàn thiện cùng đoạn mã của quá trình thực nghiệm trên bộ dữ liệu	Tổng hợp/Đánh giá
	Vương Trí Tín			
	Trần Văn Trường			

Ngày 15 tháng 9 năm 2024

XÁC NHẬN CỦA GIẢNG VIÊN

(Kí, ghi rõ họ tên)

Nguyễn Mạnh Cường

BÁO CÁO HỌC TẬP CÁ NHÂN/NHÓM

Tên lớp: 20241IT6077002 Khóa: K16-K17

Họ tên thành viên của nhóm: 1) Nguyễn Trung Hiếu; 2) Vương Trí Tín; 3) Trần Văn Trường

Tên nhóm: Nhóm 17

Tên chủ đề: Phân tích dự đoán giá máy tính xách tay bằng mô hình hồi quy tuyến tính.

Tuần	**Người thực hiện**	**Nội dung công việc**	**Kết quả đạt được**	**Kiến nghị với giảng viên hướng dẫn**
1 - 3	Nguyễn Trung Hiếu Vương Trí Tín Trần Văn Trường	Chọn đề tài, tìm kiếm dữ liệu	Đề tài: Phân tích dự đoán giá máy tính xách tay bằng mô hình hồi quy tuyến tính Bộ dữ liệu máy tính xách tay thu thập trên sàn Amazon	Góp ý cấu trúc nội dung và tính logic của phần mở đầu.
4	Nguyễn Trung Hiếu	Giới thiệu tổng quan về đề tài, các chương, kỳ vọng và kết quả	Thực hiện được mở đầu báo cáo	Kiểm tra sự đầy đủ và hợp lý của phần mở đầu.
	Vương Trí Tín	Tổng quan về bài toán dự báo và tình hình nghiên cứu	Tổng quan bài toán dự báo, tình hình nghiên cứu quốc tế và ở Việt Nam	Đánh giá phần tổng quan và đề xuất bổ sung nếu cần.
	Trần Văn Trường	Giới thiệu về bài toán dự đoán giá máy tính xách tay	Giới thiệu được bài toán, mô tả đầu vào và đầu ra cùng khó khăn, miền ứng dụng bài toán	Kiểm tra tính chính xác và bổ sung khía cạnh thực tiễn nếu thiếu.
5	Nguyễn Trung Hiếu	Tìm hiểu các phương pháp giải quyết bài toán và lựa chọn phương pháp	Các phương pháp phân tích hồi quy được đưa ra và hồi	Đánh giá lý do chọn hồi quy tuyến tính và gợi ý bổ sung

			quy tuyến tính được chọn để giải quyết	các phương pháp khác.
	Vương Trí Tín	Tìm hiểu tổng quan về mô hình hồi quy tuyến tính	Tổng quan, ưu và nhược điểm của mô hình hồi quy tuyến tính với bài toán	Góp ý đánh giá ưu, nhược điểm và bổ sung khía cạnh thực tiễn nếu cần.
	Trần Văn Trường	Lựa chọn công cụ hỗ trợ giải quyết bài toán	Công cụ python cùng các thư viện liên quan được lựa chọn	Xác nhận sự phù hợp của công cụ và thư viện đã chọn.
6	Nguyễn Trung Hiếu	Mô tả bộ dữ liệu và làm sạch	Các cột đầu vào kiểu chữ (mang bản chất số) được chuyển về dạng số	Đánh giá tính hợp lý và khoa học của các bước làm sạch và chuyển đổi dữ liệu.
	Vương Trí Tín	Thực hiện tóm lược dữ liệu	Thực hiện thống kê mô tả các cột dữ liệu sau khi được làm sạch	Góp ý về sự hợp lý của các thống kê mô tả, bổ sung thông tin nếu cần.
	Trần Văn Trường	Thực hiện chuyển đổi dữ liệu	Ánh xạ các cột thích hợp từ kiểu phi số thành kiểu số	Xem xét sự hợp lý và logic của ánh xạ các dữ liệu phi số.
7	Nguyễn Trung Hiếu	Phân tích mô tả bộ dữ liệu sau tiền xử lý	Thấy được mức tương quan các thuộc tính so với thuộc tính đích	Đánh giá mức độ tương quan và tính khả thi của dữ liệu để dự báo.
	Vương Trí Tín			

	Trần Văn Trường	Chỉnh sửa và nhận xét các về mức độ ảnh hưởng của các thuộc tính so với thuộc tính giá máy tính	Nhận xét mức độ tương quan và giải thích	Bổ sung ý kiến về các nhận xét và giải thích liên quan đến thuộc tính.
8	Nguyễn Trung Hiếu	Tạo và huấn luyện mô hình hồi quy	Dữ liệu được huấn luyện và đưa ra các thông số của mô hình	Xem xét kết quả huấn luyện và đánh giá tính hợp lý của các thông số.
	Vương Trí Tín	Phân tích các thông số của mô hình, ưu và nhược điểm từ đó đưa ra đánh giá nhận xét	Bảng kết quả các lần chạy được lập, từ đó giải thích được hiệu suất của mô hình	Đưa ra nhận xét, đề xuất cải thiện các yếu tố ảnh hưởng đến hiệu suất mô hình.
	Trần Văn Trường			
9 - 10	Nguyễn Trung Hiếu	Kết luận sau quá trình thực nghiệm, chỉnh sửa và bổ sung báo cáo	Quyển báo cáo thực nghiệm hoàn thiện cùng đoạn mã của quá trình thực nghiệm trên bộ dữ liệu	Kiểm tra báo cáo tổng thể và đánh giá sự logic, chặt chẽ của nội dung báo cáo.
	Vương Trí Tín			
	Trần Văn Trường			

Ngày 25 tháng 12 năm 2024

XÁC NHẬN CỦA GIẢNG VIÊN

(Kí, ghi rõ họ tên)

Nguyễn Mạnh Cường

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

DANH MỤC KÝ HIỆU, THUẬT NGỮ, TỪ VIẾT TẮT

STT	Ký hiệu, từ tắt	Tên tiếng anh	Tên tiếng việt
1	AI	Artificial Intelligence	Trí tuệ nhân tạo
2	RAM	Random Access Memory	Bộ nhớ truy cập ngẫu nhiên
3	CPU	Central Processing Unit	Bộ xử lý trung tâm
4	HDD	Hard Disk Drive	Ổ đĩa cứng
5	SSD	Solid State Drive	Ổ đĩa bán dẫn
6	GPU	Graphics Processing Unit	Bộ xử lý đồ họa
7	USD	United States Dollar	Đô la Mỹ
8	VNĐ	Việt Nam Đồng	
9	IDE	Integrated Development Environment	Môi trường phát triển tích hợp
10	MSE	Mean Squared Error	Sai số toàn phương trung bình
11	R^2	R-squared	Hệ số xác định

LỜI CẢM ƠN

Đầu tiên, chúng em xin gửi lời cảm ơn đến quý thầy cô, nhà trường và bạn bè đã quan tâm và động viên chúng em trong suốt thời gian qua. Sự quan tâm và sự hỗ trợ của quý thầy cô không chỉ là nguồn động lực mà còn là một phần quan trọng trong việc xác định hướng đi và hoàn thiện đề tài của chúng em.

Chúng em cũng muốn bày tỏ lòng biết ơn đặc biệt đến thầy **TS. Nguyễn Mạnh Cường** vì sự tận tâm và sự chỉ bảo tận tình trong quá trình hướng dẫn chúng em. Những kiến thức mà thầy đã truyền đạt và những gợi ý quý báu đã giúp chúng em hiểu sâu hơn về mô hình hồi quy tuyến tính và áp dụng nó vào đề tài: "Phân tích dự đoán giá máy tính xách tay bằng mô hình hồi quy tuyến tính". Sự kiên nhẫn và sự đồng hành của thầy đã giúp chúng em vượt qua những khó khăn và hoàn thành đề tài một cách tốt nhất.

Chúng em cũng xin gửi lời cảm ơn đến tất cả những người đã giúp đỡ và hỗ trợ chúng tôi trong quá trình nghiên cứu này. Những đóng góp, ý kiến và sự hỗ trợ của các bạn đã đóng vai trò quan trọng trong việc nâng cao chất lượng của báo cáo này.

Cuối cùng, chúng em xin kết thúc lời cảm ơn này bằng việc cam kết sẽ tiếp tục nỗ lực, phấn đấu và trưởng thành hơn trong con đường nghiên cứu và học tập. Chúng em sẽ luôn đặt lòng biết ơn và trân trọng những sự giúp đỡ mà chúng em đã nhận được và sẽ cố gắng chia sẻ kiến thức và kinh nghiệm của mình để đóng góp vào sự phát triển của cộng đồng.

Nhóm sinh viên thực hiện

Nguyễn Trung Hiếu

Trần Văn Trường

Vương Trí Tín

LỜI NÓI ĐẦU

Trong thời đại số hóa và bùng nổ công nghệ như hiện nay, máy tính xách tay đã trở thành thiết bị không thể thiếu trong học tập, công việc, và giải trí của con người. Sự phát triển mạnh mẽ của lĩnh vực công nghệ thông tin, trí tuệ nhân tạo, và khoa học dữ liệu đã tạo ra một khối lượng dữ liệu khổng lồ, mở ra nhiều cơ hội mới cho việc phân tích và dự báo. Trong bối cảnh đó, việc ứng dụng các phương pháp phân tích dữ liệu để dự đoán giá cả sản phẩm, đặc biệt là sản phẩm công nghệ cao như máy tính xách tay, ngày càng thu hút sự quan tâm và chú ý từ các nhà nghiên cứu và doanh nghiệp.

Đề tài "Phân tích dự đoán giá máy tính xách tay bằng mô hình hồi quy tuyến tính" hướng đến việc xây dựng một mô hình dự báo giá cho các loại máy tính xách tay dựa trên các đặc điểm cơ bản của chúng. Bằng cách áp dụng phương pháp hồi quy tuyến tính – một phương pháp phổ biến trong dự báo và phân tích dữ liệu – đề tài không chỉ giúp tìm hiểu mối quan hệ giữa các đặc tính của sản phẩm và giá trị thị trường mà còn có thể cung cấp cái nhìn trực quan về xu hướng giá cả. Đề tài hứa hẹn mang lại những giá trị ứng dụng thực tiễn, hỗ trợ người tiêu dùng trong việc lựa chọn sản phẩm phù hợp và giúp các nhà sản xuất, nhà phân phối định giá sản phẩm một cách hợp lý.

Báo cáo này được cấu trúc thành ba chương chính như sau:

- **Chương 1:** Tổng quan về đề tài, trình bày các khái niệm cơ bản và tổng quan tình hình nghiên cứu liên quan đến việc dự báo giá sản phẩm công nghệ.
- **Chương 2:** Phương pháp luận và công cụ sử dụng, giới thiệu về mô hình áp dụng để dự báo và các công cụ, thư viện hỗ trợ trong quá trình phân tích dữ liệu.
- **Chương 3:** Thực nghiệm và kết quả, trong đó thực hiện phân tích, huấn luyện mô hình dự đoán và đánh giá kết quả dự báo.

Với đề tài này, chúng em kỳ vọng rằng mô hình dự báo giá máy tính xách tay sẽ đạt được độ chính xác cao, đóng góp vào việc tối ưu hóa quy trình định giá sản phẩm công nghệ.

CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI

1.1. Tổng quan về phân tích dữ liệu

1.1.1. Phân tích dữ liệu là gì

Phân tích dữ liệu là quá trình thu thập, làm sạch, xử lý và mô hình hoá dữ liệu để rút ra các thông tin có giá trị và hỗ trợ cho quá trình ra quyết định. Quá trình này là một phần không thể thiếu trong các lĩnh vực như kinh doanh, y tế, tài chính, giáo dục và công nghệ thông tin, nơi các quyết định chiến lược thường dựa trên các phân tích định lượng hoặc định tính từ dữ liệu thu thập được.

Hình 1.1. Hình ảnh minh họa phân tích dữ liệu (Nguồn: base.vn)

Phân tích dữ liệu giúp hiểu sâu hơn về các hiện tượng thông qua việc tìm kiếm các mẫu, xu hướng và mối quan hệ trong dữ liệu. Tuỳ thuộc vào mục tiêu, phương pháp và phạm vi phân tích, phân tích dữ liệu thường được chia thành bốn loại chính:

- **Phân tích mô tả (Descriptive Analytics):** Đây là dạng phân tích cơ bản nhất, tập trung vào việc tóm tắt và trình bày dữ liệu dưới dạng bảng biểu, biểu đồ hoặc các số liệu thống kê. Phân tích mô tả giúp trả lời các câu hỏi như: *"Điều gì đã xảy ra?"*. Ví dụ, trong kinh doanh,

phân tích mô tả có thể cho biết doanh thu hàng tháng hoặc tỷ lệ khách hàng quay lại.

- **Phân tích chẩn đoán (Diagnostic Analytics):** Phân tích chẩn đoán giải thích lý do xảy ra các hiện tượng đã được phát hiện trong phân tích mô tả. Nó sử dụng các phương pháp như phân tích hồi quy, phân tích mối tương quan và khai phá dữ liệu để xác định các yếu tố ảnh hưởng. Câu hỏi chính mà loại phân tích này giải quyết là: *"Tại sao điều đó xảy ra?"*.
- **Phân tích dự đoán (Predictive Analytics):** Phân tích dự đoán sử dụng các mô hình thống kê và thuật toán học máy để dự đoán các sự kiện trong tương lai. Loại phân tích này thường dựa trên các tập dữ liệu lịch sử để tạo ra các dự đoán về xu hướng, hành vi hoặc kết quả tiềm năng. Ví dụ, một công ty có thể dự đoán nhu cầu sản phẩm dựa trên dữ liệu bán hàng trước đây.
- **Phân tích đề xuất (Prescriptive Analytics):** Đây là dạng phân tích tiên tiến nhất, không chỉ dự đoán kết quả mà còn đề xuất các hành động tối ưu để đạt được mục tiêu mong muốn. Phân tích đề xuất thường sử dụng mô phỏng, tối ưu hóa và các công cụ AI để hỗ trợ ra quyết định.

Trong nghiên cứu này, chúng tôi sử dụng phân tích dữ liệu với mục tiêu dự đoán giá máy tính xách tay dựa trên các đặc trưng kỹ thuật và thông số sản phẩm. Phương pháp phân tích dự đoán sẽ được áp dụng để xây dựng một mô hình học máy, giúp ước lượng giá trị của máy tính xách tay dựa trên các thông tin như bộ xử lý, RAM, dung lượng lưu trữ, và thương hiệu.

Việc ứng dụng phân tích dữ liệu không chỉ giúp hiểu rõ hơn về cách các yếu tố kỹ thuật ảnh hưởng đến giá cả mà còn hỗ trợ người tiêu dùng và nhà sản xuất trong việc đưa ra các quyết định mua bán hoặc thiết kế sản phẩm một cách hiệu quả hơn.

1.1.2. Quy trình phân tích dữ liệu

Hình 1.2: Quy trình phân tích dữ liệu [1]

Quy trình phân tích dữ liệu bao gồm các bước cơ bản:

– **Xác định mục tiêu:**

Bước đầu tiên trong phân tích dữ liệu là xác định rõ mục tiêu của phân tích, giúp định hướng các yêu cầu và phương pháp thực hiện. Trong bối cảnh nghiên cứu dự đoán giá máy tính xách tay, mục tiêu là xây dựng một mô hình có khả năng dự báo giá bán dựa trên các đặc trưng kỹ thuật của sản phẩm. Sau khi xác định được mục tiêu, việc thu thập dữ liệu được tiến hành, đảm bảo rằng dữ liệu chứa đủ các thông tin cần thiết để đạt được mục tiêu đó.

– **Thu thập dữ liệu:**

Dữ liệu có thể được thu thập từ nhiều nguồn khác nhau như cơ sở dữ liệu nội bộ, các trang thương mại điện tử hoặc các nền tảng cung cấp dữ liệu mở. Trong nghiên cứu này, dữ liệu được thu thập từ Amazon, bao gồm các đặc trưng của máy tính xách tay như thương hiệu, bộ vi xử lý, RAM, dung lượng ổ cứng và kích thước màn hình, ….

– **Tiền xử lý dữ liệu:**

Dữ liệu thô sau khi thu thập thường chứa các giá trị bị thiếu, lỗi hoặc không nhất quán, ảnh hưởng đến độ chính xác của mô hình phân tích. Do đó, dữ liệu cần được tiền xử lý để loại bỏ các yếu tố gây nhiễu, đảm bảo dữ liệu sẵn sàng cho quá trình phân tích.

– **Phân tích dữ liệu:**

Sau khi tiền xử lý, dữ liệu đã sẵn sàng để phân tích. Phân tích dữ liệu có thể sử dụng các phương pháp: phân tích mô tả, phân tích hồi quy v.v.

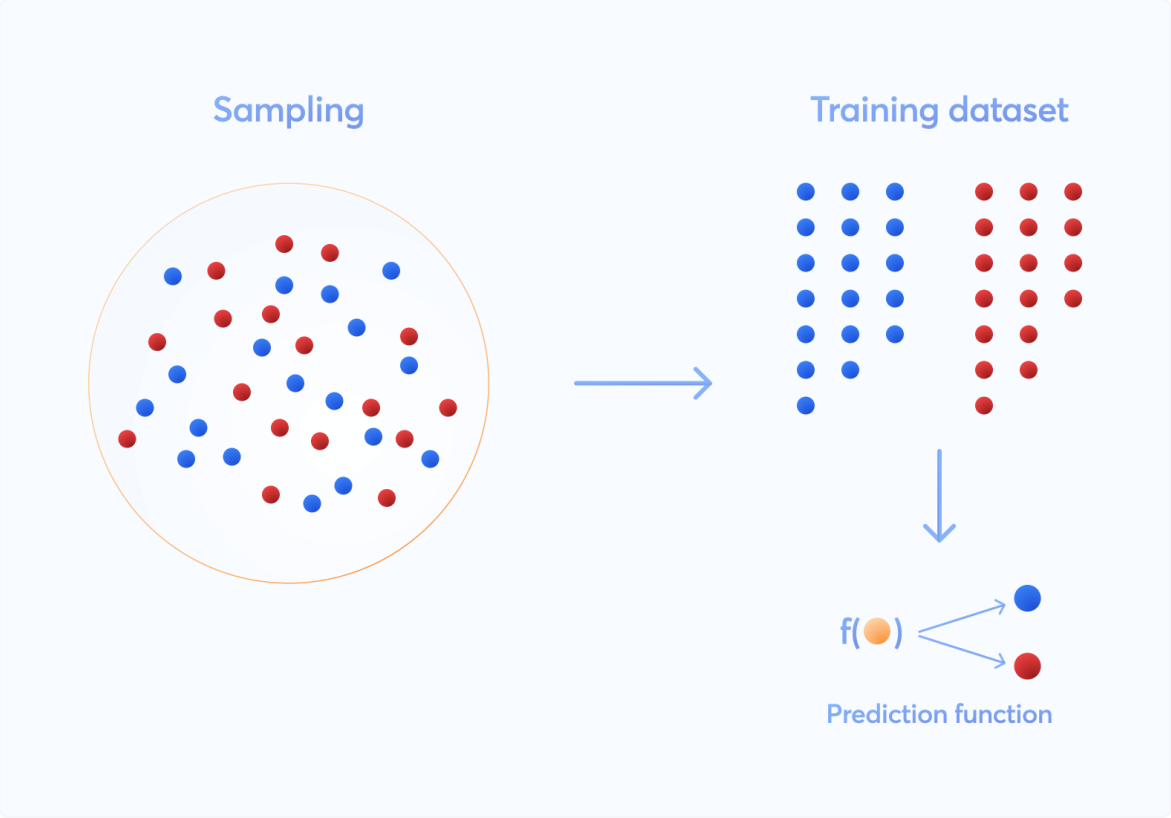
– **Kết quả và dự đoán:**

Bước cuối cùng trong quy trình phân tích dữ liệu là đưa ra các dự đoán và đánh giá kết quả. Với mô hình đã xây dựng, dữ liệu kiểm tra được sử dụng để đánh giá độ chính xác của mô hình.

1.2. Tổng quan về bài toán dự báo

1.2.1. Bài toán dự báo

Bài toán dự đoán là một trong những lĩnh vực cốt lõi trong khoa học dữ liệu và phân tích, giúp các tổ chức và cá nhân đưa ra quyết định dựa trên dữ liệu hiện tại và dữ liệu lịch sử. Dự đoán bao gồm các phương pháp và kỹ thuật sử dụng dữ liệu đầu vào để xây dựng mô hình dự báo cho các kết quả hoặc xu hướng tương lai.



Hình 1.3. Hình minh họa bài toán dự báo (Nguồn: censius.ai)

Sự phát triển nhanh chóng của công nghệ thông tin và dữ liệu lớn đã mở ra nhiều cơ hội cho phân tích dự đoán, không chỉ để hiểu rõ hơn về dữ liệu mà còn để tạo ra giá trị thực tiễn trong nhiều lĩnh vực như tài chính, y tế, sản xuất, và thương mại điện tử, mua bán phù hợp và tối ưu.

Quá trình dự đoán thường được tổ chức theo các bước cơ bản sau:

1) **Xác định vấn đề cần dự đoán:** Đây là bước đầu tiên và quan trọng nhất, đảm bảo rằng mục tiêu của bài toán được xác định rõ ràng. Các câu hỏi như "Chúng ta cần dự đoán gì?" và "Dự đoán này sẽ giúp ích như thế nào trong việc ra quyết định?" cần được trả lời cụ thể.
2) **Thu thập và chuẩn bị dữ liệu:** Dữ liệu là yếu tố cốt lõi cho bất kỳ phân tích dự đoán nào. Quy trình này bao gồm thu thập dữ liệu từ các nguồn khác nhau (hệ thống quản lý, cảm biến, cơ sở dữ liệu trực tuyến) và chuẩn bị dữ liệu để loại bỏ nhiễu, xử lý các giá trị thiếu hoặc sai lệch.
3) **Chọn mô hình và thuật toán:** Các mô hình dự đoán, từ đơn giản như hồi quy tuyến tính đến phức tạp như mạng nơ-ron, đều có ưu và nhược điểm riêng. Việc lựa chọn đúng mô hình phụ thuộc vào bản chất của dữ liệu và mục tiêu dự đoán.
4) **Huấn luyện và kiểm tra mô hình:** Dữ liệu sẽ được chia thành hai tập: tập huấn luyện và tập kiểm tra. Mô hình được huấn luyện trên tập huấn luyện và kiểm tra hiệu suất dựa trên tập kiểm tra để đánh giá khả năng dự đoán của nó.
5) **Triển khai và cải tiến:** Sau khi đạt được hiệu suất mong muốn, mô hình dự đoán sẽ được triển khai trong môi trường thực tế. Quá trình này bao gồm cả việc giám sát và cải tiến mô hình để duy trì độ chính xác trong dài hạn.

Ứng dụng của dự đoán

Dự đoán không chỉ là công cụ hỗ trợ trong quản lý mà còn mở ra những hướng đi mới trong đổi mới và sáng tạo:

- Tài chính: Dự báo giá cổ phiếu, rủi ro tín dụng, và nhu cầu thị trường.
- Chăm sóc sức khỏe: Hỗ trợ bác sĩ trong chẩn đoán bệnh, dự báo dịch bệnh và cải thiện hiệu quả điều trị.
- Logistics: Tối ưu hóa chuỗi cung ứng, dự đoán nhu cầu hàng hóa.
- Giáo dục: Cá nhân hóa quá trình học tập, dự báo kết quả học tập dựa trên dữ liệu hành vi của học sinh.

1.2.2. Bài toán dự báo ở quốc tế

Trên thế giới, bài toán dự báo được ứng dụng rộng rãi trong nhiều lĩnh vực như kinh tế, tài chính, y tế, và thời tiết. Ví dụ, trong lĩnh vực kinh tế, các nhà nghiên cứu sử dụng các mô hình dự báo để dự đoán tăng trưởng GDP, lạm phát, và tỷ lệ thất nghiệp. Các phương pháp dự báo phổ biến bao gồm phân tích hồi quy, mô hình chuỗi thời gian, và mô hình học máy. Các mô hình này giúp các nhà hoạch định chính sách và doanh nghiệp đưa ra các quyết định chiến lược dựa trên dự đoán về xu hướng tương lai. Ngoài ra, trong lĩnh vực y tế, dự báo dịch bệnh là một công cụ quan trọng giúp các cơ quan y tế chuẩn bị và ứng phó kịp thời với các đợt bùng phát dịch bệnh.

1.2.3. Bài toán dự báo ở Việt Nam

Ở Việt Nam, bài toán dự báo cũng được áp dụng rộng rãi trong nhiều lĩnh vực. Một ví dụ cụ thể là dự báo dân số, nơi các nhà nghiên cứu sử dụng các mô hình thống kê như hồi quy, chuỗi thời gian và chuỗi thời gian mờ để dự báo dân số trong tương lai. Trong lĩnh vực năng lượng, dự báo phụ tải hệ thống điện là một công cụ quan trọng giúp các nhà quản lý dự đoán nhu cầu điện năng trong ngắn hạn và dài hạn, từ đó đảm bảo cung cấp điện ổn định cho toàn quốc. Ngoài ra, các tổ chức như Ngân hàng Phát triển châu Á (ADB) thường công bố các báo cáo dự báo tăng trưởng kinh tế của Việt Nam, dựa trên các yếu tố như hoạt động thương mại, sản xuất chế biến, và các biện pháp hỗ trợ tài chính. Những

dự báo này giúp giảm bớt rủi ro và hỗ trợ các nhà hoạch định chính sách trong việc phát triển kinh tế và xã hội.

1.3. Giới thiệu bài toán dự đoán giá máy tính xách tay

1.3.1. Giới thiệu bài toán

Dự đoán giá máy tính xách tay dựa trên thông số kỹ thuật là một bài toán quan trọng trong thương mại điện tử và quản lý chuỗi cung ứng. Với sự phát triển nhanh chóng của thị trường máy tính xách tay, việc hiểu và dự đoán giá sản phẩm không chỉ giúp nhà bán lẻ đưa ra quyết định định giá phù hợp mà còn hỗ trợ khách hàng trong việc lựa chọn sản phẩm phù hợp với nhu cầu và ngân sách.

Hình 1.4. Bài toán dự báo giá máy tính xách tay (Nguồn: medium.com)

1.3.2. Mô tả chi tiết đầu vào và đầu ra

Dữ liệu đầu vào bao gồm các thuộc tính cụ thể của máy tính xách tay, chẳng hạn như bộ xử lý (CPU), bộ nhớ RAM, dung lượng và loại ổ cứng (HDD hoặc SSD), kích thước màn hình, và GPU. Ngoài ra, các yếu tố như thương hiệu, hệ điều hành, trọng lượng sản phẩm, ngày phát hành và các tính năng đặc

biệt (ví dụ: màn hình cảm ứng, độ phân giải cao) cũng được đưa vào phân tích. Dữ liệu này thường được thu thập từ các nguồn đáng tin cậy như các nền tảng thương mại điện tử lớn, điển hình là Shopee ở thị trường Việt Nam hay Amazon ở thị trường quốc tế, nhằm đảm bảo độ chính xác và độ tin cậy.

Đầu ra của bài toán là mức giá bán dự đoán cho mỗi máy tính xách tay, được biểu diễn dưới dạng số cụ thể, chẳng hạn bằng USD hoặc VNĐ. Mô hình dự đoán phải đảm bảo khả năng cung cấp kết quả với sai số nhỏ nhất, phù hợp với thực tế thị trường. Độ chính xác của mô hình phụ thuộc lớn vào chất lượng dữ liệu và phương pháp phân tích được áp dụng, bao gồm các kỹ thuật như hồi quy tuyến tính hoặc các mô hình phức tạp hơn để xử lý mối quan hệ phi tuyến giữa các thuộc tính và giá sản phẩm.

Việc dự đoán giá không chỉ có giá trị trong định giá sản phẩm tại các nền tảng thương mại mà còn hỗ trợ người tiêu dùng trong việc đưa ra quyết định mua hàng. Đồng thời, nó cũng giúp các doanh nghiệp tối ưu hóa chiến lược kinh doanh, quản lý hàng tồn kho và phân tích xu hướng thị trường một cách hiệu quả.

1.3.3. Các khó khăn và thách thức của bài toán

Bài toán dự đoán giá máy tính xách tay phải đối mặt với nhiều thách thức liên quan đến tính đa dạng và phức tạp của dữ liệu. Một trong những vấn đề lớn nhất là sự đa dạng về các thuộc tính của máy tính xách tay, từ các thông số phần cứng như bộ vi xử lý, bộ nhớ, ổ cứng cho đến các yếu tố thương hiệu và hệ điều hành. Điều này tạo ra một lượng lớn dữ liệu không đồng nhất, yêu cầu kỹ thuật xử lý và chuẩn hóa dữ liệu mạnh mẽ để đảm bảo rằng các mô hình dự đoán có thể tiếp nhận và xử lý một cách hiệu quả. Các dữ liệu không đồng nhất này có thể đến từ nhiều nguồn khác nhau như các trang thương mại điện tử và có sự khác biệt lớn về độ tin cậy.

Một yếu tố khó khăn khác là ảnh hưởng của các yếu tố phi kỹ thuật, đặc biệt là thương hiệu. Thương hiệu, mặc dù không phải là thông số kỹ thuật nhưng lại có tác động rất lớn đến giá bán của máy tính xách tay. Các thương hiệu như

Apple, Dell, và HP có thể có giá bán cao hơn so với các thương hiệu ít nổi tiếng hơn, mặc dù cấu hình phần cứng có thể tương đương. Việc xác định và lượng hóa yếu tố này để đưa vào mô hình dự đoán là một thách thức lớn, vì giá trị cảm nhận của thương hiệu có thể thay đổi tùy thuộc vào người tiêu dùng và thị trường.

Bên cạnh đó, bài toán này cũng gặp phải sự phức tạp trong mối quan hệ phi tuyến giữa các yếu tố đầu vào và giá sản phẩm. Các thông số như bộ vi xử lý, RAM, và kích thước màn hình có thể không tác động theo một cách tuyến tính đến giá máy tính xách tay. Một sự thay đổi nhỏ trong một thông số có thể dẫn đến sự thay đổi lớn về giá, điều này yêu cầu các mô hình học máy phức tạp hơn, chẳng hạn như hồi quy phi tuyến hoặc các phương pháp học sâu, để mô hình có thể bắt kịp được các mối quan hệ phức tạp này.

Cuối cùng, một vấn đề không thể bỏ qua là sự thiếu vắng của dữ liệu hoặc dữ liệu nhiễu. Các máy tính xách tay mới ra mắt có thể thiếu thông tin về cấu hình, và đôi khi các mô tả sản phẩm trên các trang thương mại điện tử không chính xác hoặc không đầy đủ. Việc xử lý các giá trị thiếu và dữ liệu không chính xác là một trong những công việc quan trọng trong tiền xử lý dữ liệu, giúp cải thiện độ chính xác của mô hình dự đoán. Những vấn đề này đòi hỏi phải có sự kết hợp của nhiều kỹ thuật làm sạch và xử lý dữ liệu trước khi tiến hành phân tích.

1.3.4. Các miền ứng dụng của bài toán

- **Định giá sản phẩm trong thương mại điện tử:** Hỗ trợ nhà bán lẻ đưa ra mức giá cạnh tranh, phù hợp với xu hướng thị trường.
- **Tư vấn mua hàng:** Giúp người tiêu dùng tìm được sản phẩm tốt nhất theo mức giá mong muốn, đồng thời nâng cao trải nghiệm mua sắm.
- **Dự báo và quản lý hàng tồn kho:** Dự đoán xu hướng giá cả để tối ưu hóa lượng hàng nhập kho, giảm thiểu rủi ro tồn kho sản phẩm giá trị cao.

- **Phân tích thị trường:** Hỗ trợ các doanh nghiệp nghiên cứu thị trường, phát hiện xu hướng và đưa ra chiến lược kinh doanh phù hợp.

1.5. Kết luận chương 1

Chương 1 đã cung cấp cái nhìn tổng quan về bài toán dự đoán giá máy tính xách tay, từ việc định nghĩa vấn đề đến việc phân tích các yếu tố ảnh hưởng đến giá sản phẩm. Chúng ta đã khám phá các yếu tố kỹ thuật và phi kỹ thuật cần thiết cho việc dự đoán giá, cũng như những khó khăn và thách thức trong quá trình xử lý và phân tích dữ liệu. Các yếu tố như sự đa dạng về thông số kỹ thuật, ảnh hưởng của thương hiệu, mối quan hệ phi tuyến giữa các đặc tính của máy tính xách tay, và vấn đề thiếu dữ liệu là những yếu tố cần được giải quyết để đạt được kết quả dự đoán chính xác. Chương này đã khái quát các yếu tố quan trọng, đồng thời chuẩn bị nền tảng cho việc áp dụng các phương pháp và công cụ trong các chương sau của báo cáo.

CHƯƠNG 2: PHƯƠNG PHÁP VÀ CÔNG CỤ

2.1. Phương pháp phân tích hồi quy

Phương pháp phân tích hồi quy là một kỹ thuật thống kê mạnh mẽ được sử dụng để mô hình hóa mối quan hệ giữa một biến phụ thuộc và một hoặc nhiều biến độc lập. Mục tiêu của phương pháp này là dự đoán giá trị của biến phụ thuộc dựa trên các giá trị của các biến độc lập, qua đó giúp nhận diện và phân tích các yếu tố ảnh hưởng đến biến cần dự đoán. Hồi quy có thể được áp dụng trong nhiều lĩnh vực khác nhau, từ tài chính, marketing cho đến y tế, và trong nghiên cứu dữ liệu lớn, hồi quy là công cụ quan trọng để hiểu và dự đoán các xu hướng từ dữ liệu.

2.1.1. Các phương pháp hồi quy giải quyết bài toán

Có nhiều phương pháp hồi quy được phát triển nhằm giải quyết các vấn đề phức tạp trong phân tích dự báo. Các phương pháp phổ biến bao gồm:

- **Hồi quy Ridge (Ridge Regression)**: Là một phương pháp hồi quy tuyến tính mở rộng nhằm xử lý vấn đề đa cộng tuyến trong dữ liệu. Ridge Regression thêm một hệ số điều chuẩn (regularization term) vào hàm mục tiêu để làm giảm hệ số hồi quy, từ đó giảm thiểu sự ảnh hưởng của các biến độc lập có mối tương quan cao.
- **Hồi quy phi tuyến (Nonlinear Regression)**: Khi mối quan hệ giữa biến phụ thuộc và biến độc lập không tuyến tính, hồi quy phi tuyến có thể mô hình hóa các quan hệ này. Mặc dù linh hoạt, hồi quy phi tuyến phức tạp và đòi hỏi nhiều tính toán hơn, thích hợp với các dữ liệu mà các phương pháp hồi quy tuyến tính không đáp ứng được.
- **Hồi quy logistic (Logistic Regression)**: Dù không phải là hồi quy tuyến tính, hồi quy logistic cũng là một phương pháp phổ biến trong dự đoán biến phân loại. Hồi quy logistic dựa trên hàm sigmoid để mô hình hóa xác suất thuộc các lớp phân loại khác nhau, thay vì dự đoán giá trị thực.

- **Hồi quy tuyến tính (Linear Regression)**: Hồi quy tuyến tính dự đoán giá trị mục tiêu dựa trên biến độc lập bằng cách tìm đường thẳng "tốt nhất" vượt qua dữ liệu. Phương pháp này đơn giản và phù hợp với dữ liệu có mối quan hệ tuyến tính. Tuy nhiên, nó có thể không xử lý được dữ liệu phi tuyến và ảnh hưởng bởi nhiễu dữ liệu.

2.1.2. Lựa chọn phương pháp

Với bài toán dự báo giá máy tính xách tay, **phương pháp hồi quy tuyến tính** được lựa chọn do tính phù hợp về mặt dữ liệu với mong muốn đầu vào và đầu ra. Trước hết, hồi quy tuyến tính là mô hình đơn giản và dễ hiểu, cho phép chúng ta nhanh chóng xác định và phân tích mối quan hệ giữa các yếu tố đầu vào và giá trị đầu ra. Điều này đặc biệt quan trọng trong các bài toán dự đoán giá trị liên tục, nơi mối quan hệ giữa các đặc tính kỹ thuật và giá trị dự đoán có thể được giả định là tuyến tính hoặc gần tuyến tính.

Hơn nữa, phương pháp hồi quy tuyến tính giúp chúng ta dễ dàng giải thích được các yếu tố ảnh hưởng đến giá bán của máy tính xách tay, như sự thay đổi trong giá trị của từng yếu tố (chẳng hạn như RAM, bộ vi xử lý, hoặc kích thước màn hình) sẽ ảnh hưởng như thế nào đến mức giá. Các nghiên cứu trước đây đã chỉ ra rằng hồi quy tuyến tính rất hữu ích trong việc xử lý các bài toán phân tích và dự đoán giá trị của sản phẩm khi các yếu tố ảnh hưởng tương đối rõ ràng và có thể mô hình hóa bằng các quan hệ tuyến tính cơ bản [3].

Mặc dù có các phương pháp phức tạp hơn như hồi quy phi tuyến tính hoặc các mô hình học sâu, nhưng hồi quy tuyến tính vẫn là lựa chọn hợp lý trong trường hợp này do tính dễ triển khai, khả năng giải thích mô hình rõ ràng và hiệu quả trong các bài toán có dữ liệu không quá phức tạp. Vì vậy, hồi quy tuyến tính là công cụ lý tưởng để bắt đầu, và nếu cần, có thể mở rộng sang các mô hình phức tạp hơn sau khi đánh giá kết quả sơ bộ từ mô hình tuyến tính.

2.1.3. Tổng quan về hồi quy tuyến tính

Hồi quy tuyến tính là một phương pháp phân tích thống kê được sử dụng rộng rãi trong việc mô hình hóa mối quan hệ giữa một biến phụ thuộc (còn gọi

là biến mục tiêu) và một hoặc nhiều biến độc lập (hay biến giải thích). Phương pháp này được ứng dụng phổ biến trong nhiều lĩnh vực như tài chính, kinh tế, khoa học dữ liệu và kỹ thuật, nhờ khả năng dự đoán và diễn giải đơn giản. Hồi quy tuyến tính giả định mối quan hệ giữa các biến là tuyến tính, tức là có thể được biểu diễn dưới dạng một đường thẳng trên đồ thị.

Biểu thức của mô hình hồi quy tuyến tính một biến độc lập được viết như sau:

$$y = ax + b$$

Trong đó:

- y là biến phụ thuộc mà ta muốn dự đoán,
- x là biến độc lập,
- b là hệ số chặn,
- a là hệ số hồi quy, đại diện cho sự thay đổi của y khi x tăng thêm một đơn vị.

Phương pháp bình phương tối thiểu là kỹ thuật phổ biến để ước lượng các hệ số này, nhằm giảm thiểu tổng bình phương sai số giữa giá trị thực tế và giá trị dự báo.

Tổng bình phương độ lệch (với Ŷ là Y dự báo)

$$J(a,b) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Bài toán trở thành: Tìm a, b để J đạt min

$$\min J(a,b) = \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\min J(a,b) = \frac{1}{2n} \sum_{i=1}^{n} (Y_i - aX_i - b)^2$$

Trong trường hợp có nhiều biến độc lập, hồi quy tuyến tính trở thành hồi quy tuyến tính đa biến, có công thức tổng quát:

$$y = \beta_0 + \beta_1 x_1 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

Ở đây, $x_1, x_2, \dots, x_m$ là các biến độc lập, và $\beta_1, \beta_2, \dots, \beta_m$ là các hệ số hồi quy tương ứng. Mục tiêu của hồi quy tuyến tính là tìm các hệ số $\beta_0, \beta_1, \dots, \beta_m$ sao cho sai số giữa giá trị dự báo và giá trị thực là nhỏ nhất.

Ưu điểm của hồi quy tuyến tính:

- Mô hình dễ hiểu và dễ triển khai, phù hợp để giải thích mối quan hệ giữa các biến.
- Hiệu quả tính toán cao, phù hợp với các bộ dữ liệu có kích thước vừa và lớn.
- Cung cấp thông tin chi tiết về mức độ ảnh hưởng của từng biến độc lập lên biến phụ thuộc, giúp hiểu rõ hơn về dữ liệu.

Nhược điểm của hồi quy tuyến tính:

- Chỉ mô hình hóa tốt các mối quan hệ tuyến tính, trong khi không phù hợp cho các quan hệ phi tuyến.
- Dễ bị ảnh hưởng bởi các giá trị ngoại lai và hiện tượng đa cộng tuyến giữa các biến độc lập.

Hồi quy tuyến tính là nền tảng cho nhiều ứng dụng thực tế trong phân tích dữ liệu và dự báo. Trong đề tài này, hồi quy tuyến tính được sử dụng để dự đoán giá của máy tính xách tay dựa trên các yếu tố như cấu hình và thương hiệu, giúp người dùng có thể đưa ra các lựa chọn mua hàng phù hợp hơn.

2.2 Công cụ phân tích

2.2.1. Giới thiệu Python và các thư viện liên quan

Để thực hiện phân tích và xây dựng mô hình dự đoán giá máy tính xách tay, ngôn ngữ lập trình Python được chọn làm công cụ chính. Python là một ngôn ngữ linh hoạt và mạnh mẽ, rất phổ biến trong các lĩnh vực khoa học dữ liệu và học máy nhờ vào khả năng thao tác dữ liệu dễ dàng và tích hợp với nhiều thư viện hỗ trợ. Trong quá trình thực hiện, mã nguồn được phát triển và chạy trong môi trường **PyCharm**, một IDE mạnh mẽ dành cho Python, giúp hỗ trợ các tác vụ như kiểm tra mã, gợi ý mã, và dễ dàng quản lý các dự án.

Các thư viện quan trọng trong Python mà nghiên cứu này sử dụng gồm:

- **numpy**: Thư viện cung cấp các công cụ để xử lý các phép toán mảng và ma trận, giúp thực hiện các phép toán số học cơ bản một cách nhanh chóng và hiệu quả.
- **pandas**: Đây là thư viện giúp thao tác với dữ liệu dạng bảng (DataFrame). Pandas hỗ trợ nhập và xuất dữ liệu, tiền xử lý dữ liệu, cũng như các thao tác cơ bản như xử lý giá trị thiếu và chuyển đổi định dạng dữ liệu.
- **scikit-learn**: Là thư viện chính cho các thuật toán học máy trong Python, bao gồm các mô hình hồi quy và phân loại. Scikit-learn cung cấp các công cụ hỗ trợ xây dựng mô hình hồi quy tuyến tính, lựa chọn đặc trưng, và đánh giá mô hình.
- **matplotlib** và **seaborn**: Các thư viện này hỗ trợ trực quan hóa dữ liệu và kết quả phân tích, giúp tạo ra các biểu đồ dễ hiểu và sinh động, phục vụ cho việc trình bày dữ liệu và mối quan hệ giữa các yếu tố.
- Ngoài ra còn một số thư viện khác như **math, re** được sử dụng để hỗ trợ tính toán và **time, platform, psutil** giúp ta có thông tin sâu hơn về quá trình phân tích dữ liệu.

PyCharm cung cấp một môi trường làm việc tích hợp với các tính năng hỗ trợ quản lý mã nguồn, kiểm tra và gỡ lỗi, giúp quá trình phát triển và triển khai mô hình học máy trở nên mượt mà và hiệu quả hơn. Cộng với sự mạnh mẽ của Python, PyCharm là công cụ lý tưởng cho việc triển khai và kiểm thử các mô hình phân tích dữ liệu phức tạp.

2.2.2. Lựa chọn công cụ

Dựa trên yêu cầu của bài toán và đặc điểm của tập dữ liệu, chúng tôi lựa chọn Python và các thư viện Pandas, NumPy, Matplotlib, Seaborn và Scikit-Learn làm các công cụ chính cho quy trình phân tích và xây dựng mô hình dự đoán. Các công cụ này được lựa chọn dựa trên các yếu tố sau:

- **Khả năng xử lý dữ liệu mạnh mẽ**: Pandas và NumPy giúp tối ưu hóa các bước xử lý và chuyển đổi dữ liệu, hỗ trợ cho các bước tiền xử lý nhanh chóng và hiệu quả.
- **Đa dạng phương pháp trực quan hóa**: Matplotlib và Seaborn cung cấp các công cụ tạo biểu đồ và đồ thị, giúp trực quan hóa các mối quan hệ giữa các biến, từ đó rút ra các kết luận về dữ liệu.
- **Hỗ trợ toàn diện cho học máy**: Scikit-Learn cung cấp các thuật toán học máy phổ biến, bao gồm hồi quy tuyến tính, cũng như các công cụ để chuẩn hóa, mã hóa và đánh giá mô hình, đáp ứng đầy đủ các yêu cầu của quy trình phân tích và dự báo giá máy tính xách tay.

Với sự kết hợp của các công cụ này, Python cung cấp một môi trường phát triển mạnh mẽ và linh hoạt, phù hợp để triển khai các bước phân tích và xây dựng mô hình dự đoán một cách hiệu quả.

2.3. Kết luận chương 2

Chương 2 đã giới thiệu chi tiết về các phương pháp phân tích và công cụ được sử dụng trong nghiên cứu này. Phương pháp phân tích hồi quy, đặc biệt là hồi quy tuyến tính, đã được chọn vì tính đơn giản, khả năng ứng dụng rộng rãi, và hiệu quả trong việc dự đoán các giá trị liên tục như giá của máy tính xách tay. Ngoài ra, các công cụ và thư viện Python mạnh mẽ như numpy, pandas, scikit-learn, matplotlib, và seaborn đã được lựa chọn để xử lý dữ liệu, xây dựng mô hình, và trực quan hóa kết quả phân tích.

Công cụ Python, kết hợp với PyCharm, cung cấp một môi trường phát triển mạnh mẽ, giúp tối ưu hóa các tác vụ lập trình và phân tích. Việc sử dụng những công cụ này đảm bảo rằng quá trình phân tích dữ liệu sẽ được thực hiện một cách hiệu quả, chính xác và dễ dàng mở rộng trong tương lai.

Tóm lại, với phương pháp hồi quy tuyến tính và các công cụ Python, nghiên cứu này sẽ tiếp cận bài toán dự đoán giá máy tính xách tay một cách khoa học và hệ thống, đảm bảo tính chính xác và khả năng mở rộng của mô hình phân tích.

CHƯƠNG 3: THỰC NGHIỆM VÀ PHÂN TÍCH KẾT QUẢ

3.1. Dữ liệu thực nghiệm

Bộ dữ liệu được sử dụng trong nghiên cứu này là dữ liệu về máy tính xách tay được bán trên Amazon. Dữ liệu bao gồm **3,988** bản ghi với **64** cột, cung cấp các đặc điểm kỹ thuật như kích thước màn hình, độ phân giải, bộ vi xử lý, dung lượng RAM, loại ổ cứng, bộ xử lý đồ họa, và giá bán. Trong đó, giá bán (Price) là biến mục tiêu được sử dụng để xây dựng mô hình hồi quy tuyến tính.

Cụ thể thông tin như sau:

- **Tên bộ dữ liệu**: Laptop Dataset
- **Nguồn**: https://github.com/MiladNooraei/Amazon-Laptop-Analysis.
- **Số lượng bản ghi**: 3,988 bản ghi.
- **Số lượng cột**: 64 cột thông tin, bao gồm các đặc điểm kỹ thuật của laptop như kích thước màn hình, độ phân giải, bộ vi xử lý, RAM, giá bán, v.v.

Bảng 3.1 dưới đây là 20 dòng đầu tiên của bộ dữ liệu.

Bảng 3.1: 20 dòng đầu tiên của bộ dữ liệu

Index	Standing screen display size	Screen Resolution	Max Screen Resolutio	Processor	RAM	Memory Spee	Hard Driv	Graphics Coprocessor	Chipset Brar	Card Descriptio	Graphic	Wireless Type	Numb
1													
2													
3	15.6 Inches	1920 x 1080 pixels	1920x1080	4.1 GHz ryzen_3	8 GB LPDDR5	3200 MHz	128 GB SSD	AMD Radeon Graphics	AMD	Integrated	8 GB	Bluetooth, 802.11ax	3
4													
6													
7	14 Inches	1366 x 768 pixels	1366 x 768 Pixels	1.1 GHz celeron_n	8 GB DDR4	2400 MHz	64 GB SSD	Intel UHD Graphics 600	Intel	Integrated	8 GB	Bluetooth, 802.11a/g	3
8													
9	16 Inches	1366 x 768 pixels	1366 x 768 pixels	4.9 GHz core_i7	16 GB DDR5		SSD	NVIDIA GeForce RTX 4060	NVIDIA	Dedicated			2
10	15.6 Inches	1366 x 768 pixels	1366 x 768 Pixels	3 GHz core_i3_family	32 GB DDR4		1 TB SSD	Intel UHD Graphics	Intel	Integrated	32 GB	Bluetooth	1
11													
12	14 Inches	1366 x 768 pixels	1366 x 768 Pixels	2.6 GHz celeron	4 GB DDR4	2.6 GHz	64 GB SSD	Intel	Intel	Integrated	4 GB	Bluetooth, 802.11a	1
13	14	1366 x 768 pixels	1366x768 Pixels	2.6 GHz celeron_n	4 GB DDR4	2400 MHz	64 GB Emm	Intel UHD Graphics 600	Intel	Integrated	4 GB	802.11a/b/g/n/ac	3
14	15.6 Inches	1927 x 1080 pixels	1920x1080	2.4 GHz apple_ci7	16 DDR5		512 GB SSD	NVIDIA GeForce RTX 4050	NVIDIA	Dedicated	6 GB	Bluetooth, 802.11ax	4
15	15.6 Inches	1920 x 1080 pixels	1920x1080	2.1 GHz apple_ci5	8 DDR5		512 GB SSD	NVIDIA GeForce RTX 4050	NVIDIA	Dedicated	6 GB	Bluetooth, 802.11ax	4
16	15.6 Inches	1920 x 1080 pixels	1920x1080 Pixels	2.1 GHz amd_ryzen_5_5500	8 GB DDR4		512 GB SSD	Intel HD Graphics 400	AMD	Integrated		802.11bgn	1
18	13.3 Inches	1366 x 768 pixels	1440 x 900 Pixels	1.8 GHz core_i5	8 GB LPDDR3	1.8 GHz	128 GB	Intel HD Graphics 6000	Intel	Integrated		Bluetooth, 802.11a	
19	15.6 Inches	1366 x 768 pixels	1366 x 768 Pixels	4.1 GHz apple_ci3	16 GB DDR4	2400 MHz	1 TB SSD	Intel UHD Graphics	Intel	Integrated	1 GB	Bluetooth, 802.11ab	1
20	15.59	1920 x 1080 pixels		1.3 celeron	16 GB DDR4		1 TB SSD		Intel	Integrated		Bluetooth	2

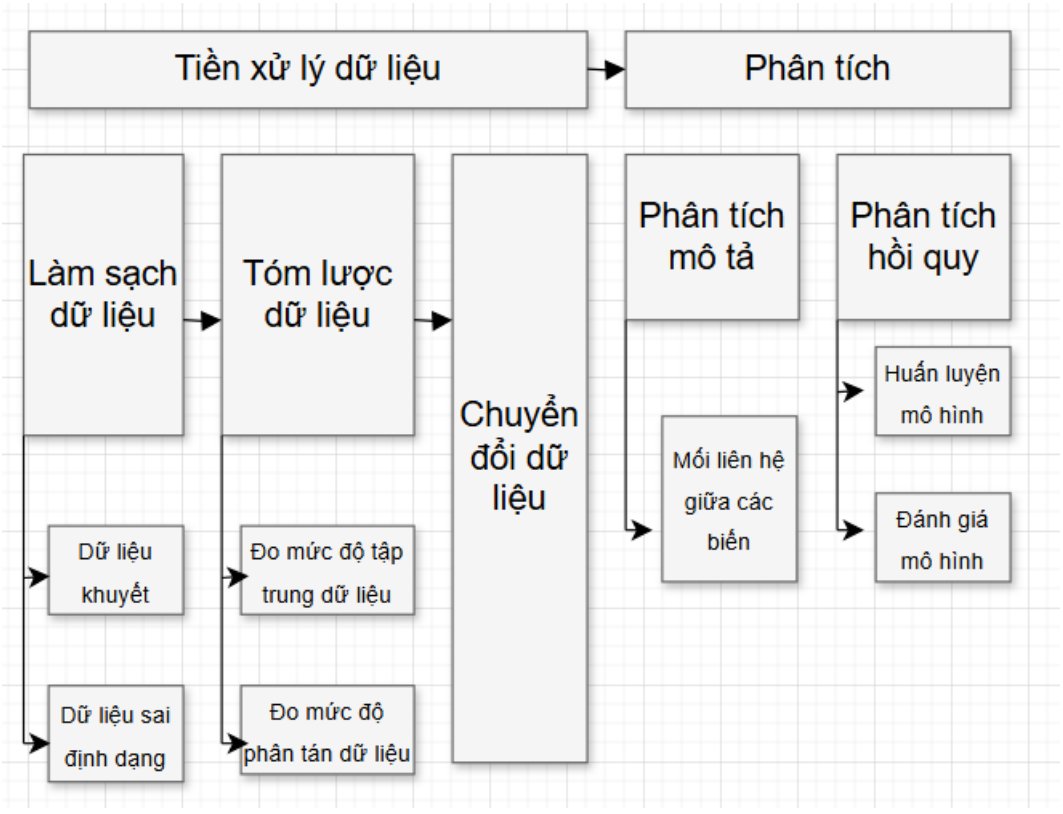
Bộ dữ liệu bao gồm các cột chính như:

1) **"Standing screen display size"**: Kích thước màn hình laptop (ví dụ: 15.6 Inches).
2) **"Screen Resolution"**: Độ phân giải màn hình (ví dụ: 1920 x 1080 pixels).
3) **"Processor"**: Loại bộ vi xử lý (ví dụ: 4.1 GHz ryzen_3).
4) **"RAM"**: Dung lượng RAM (ví dụ: 8 GB).

5) **"Memory Speed"**: Tốc độ bộ nhớ (ví dụ: 3200 MHz).
6) **"Hard Drive"**: Loại và dung lượng ổ cứng (ví dụ: 128 GB SSD).
7) **"Graphics Coprocessor"**: Bộ xử lý đồ họa (ví dụ: AMD Radeon Graphics).
8) **"Price($)"**: Giá bán laptop (biến mục tiêu trong mô hình).

Điểm nổi bật của bộ dữ liệu này là sự đa dạng và phong phú trong thông tin cung cấp. Tuy nhiên, khi xem xét chi tiết, nhiều cột có tỷ lệ giá trị bị thiếu rất cao, khiến chúng không có nhiều giá trị trong phân tích. Vậy nên, trước khi dùng để huấn luyện mô hình ta cần loại bỏ các cột có tỷ lệ giá trị bị thiếu quá cao, đồng thời xử lý các giá trị bị thiếu trong các cột quan trọng để đảm bảo tính đầy đủ và nhất quán.

3.2. Quy trình thực nghiệm



Hình 3.1: Quy trình thực nghiệm bài toán

Các bước thực nghiệm tới sẽ được tiến hành trên hệ thống máy tính cá nhân có thông số như sau:

- **Hệ điều hành**: Windows
- **Kiến trúc máy**: AMD64
- **Bộ xử lý**: Intel64 Family 6 Model 158 Stepping 10, GenuineIntel
- **Số lõi CPU**: 6
- **Số lõi CPU logic**: 12
- **Bộ nhớ RAM**: 15.85 GB
- **Tình trạng sử dụng CPU**: 3.0%
- **Tình trạng sử dụng bộ nhớ**: 82.9%
- **Thời gian thực thi (Trung bình)**: 3.02248 giây

Để thực hiện quy trình phân tích và huấn luyện mô hình hồi quy tuyến tính, tôi sử dụng môi trường Python và công cụ PyCharm. Tất cả các bước từ xử lý dữ liệu, huấn luyện mô hình đến đánh giá kết quả đều được thực hiện trên máy tính với các đặc điểm phần cứng như trên.

3.2.1 Tiền xử lý dữ liệu

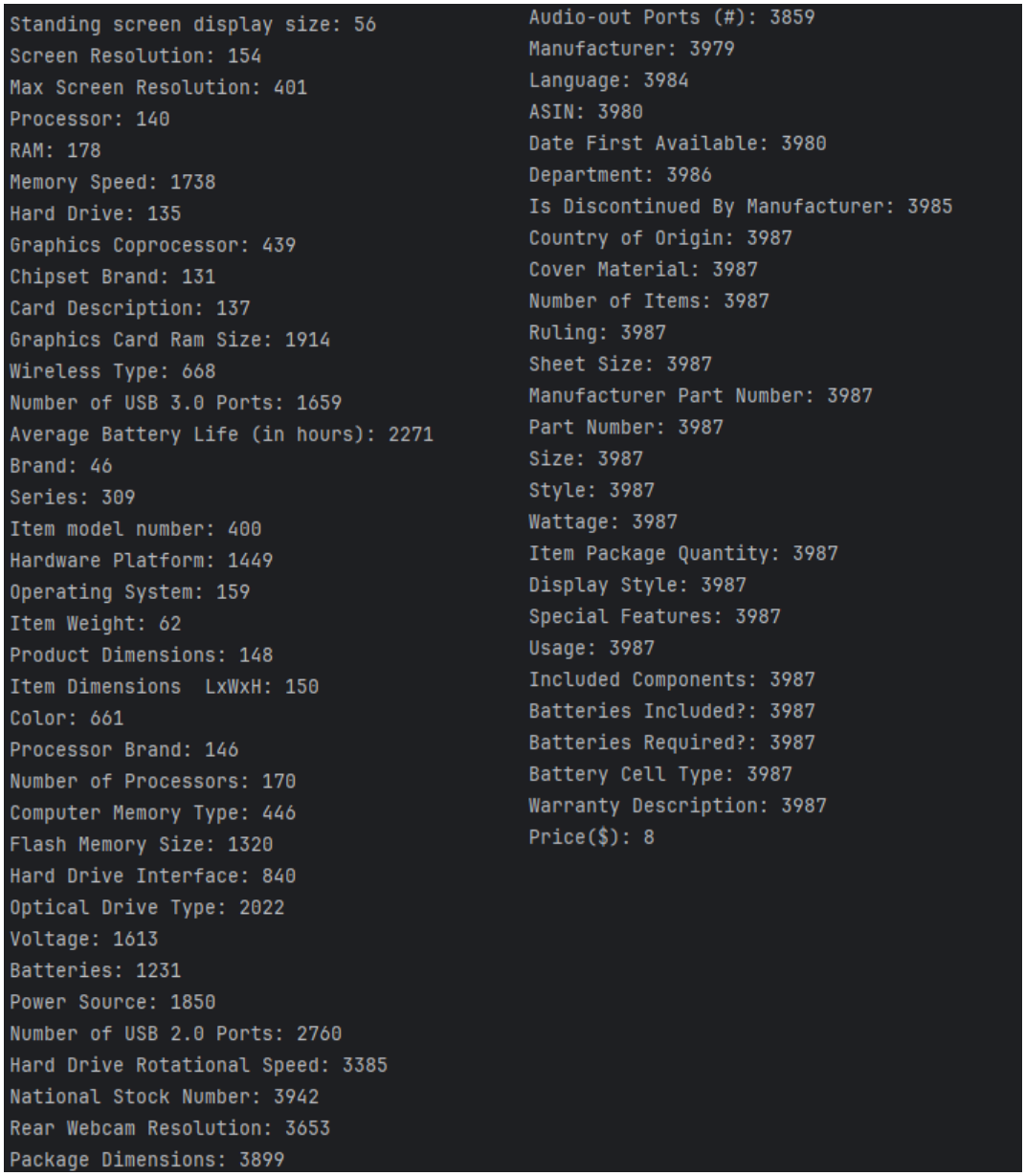
3.2.1.1. Làm sạch dữ liệu

Kiểm tra dữ liệu bị thiếu

Trong quá trình chuẩn bị dữ liệu cho mô hình hồi quy tuyến tính dự đoán giá máy tính xách tay, cần tiến hành kiểm tra số lượng dữ liệu bị khuyết trong từng cột của tập dữ liệu. Từ đó, chọn ra các cột có ít dữ liệu bị khuyết để đảm bảo tính chính xác và đầy đủ của dữ liệu đầu vào cho mô hình.

Các bước thực hiện:

1) **Tính toán số lượng dữ liệu bị thiếu trong từng cột**:
   - Đầu tiên cần đếm số lượng giá trị bị khuyết trong mỗi cột. Việc này giúp xác định các cột có mức độ dữ liệu bị thiếu thấp, có thể dễ dàng xử lý mà không làm mất quá nhiều thông tin.
   - Kết quả kiểm tra giúp đưa ra quyết định về những cột sẽ giữ lại cho mô hình, dựa trên ngưỡng dữ liệu khuyết có thể chấp nhận được.



Hình 3.2: Số lượng giá trị khuyết theo cột

2) **Lọc ra các cột có ít dữ liệu bị thiếu**:

- Dựa trên kết quả kiểm tra, chọn ra các cột có tỷ lệ dữ liệu bị khuyết thấp, đảm bảo đủ thông tin quan trọng và hạn chế ảnh hưởng tiêu cực đến mô hình. Các cột được lựa chọn gồm:
  - "Brand"
  - "Standing screen display size"
  - "Processor"
  - "RAM"
  - "Memory Speed"

- "Hard Drive"
- "Graphics Coprocessor"
- "Chipset Brand"
- "Card Description"
- "Processor Brand"
- "Operating System"
- "Item Weight"
- "Price($)"

3) **Xóa các dòng có giá trị bị khuyết trong các cột đã chọn:**

- Sau khi xác định các cột quan trọng và ít dữ liệu bị khuyết, tiến hành xóa tất cả các dòng có giá trị bị khuyết trong bất kỳ cột nào trong số này.
- Xóa các dòng thiếu dữ liệu nhằm đảm bảo dữ liệu đầu vào cho mô hình là đầy đủ và phù hợp cho phân tích hồi quy tuyến tính.

Kiểm tra dữ liệu sai định dạng

Trong quá trình tiền xử lý, một số biến trong bộ dữ liệu ban đầu tồn tại dưới dạng chuỗi văn bản, ví dụ như “Memory Speed” có các giá trị dạng như “4.1 GHz” hoặc “3200 MHz.” Các biến này cần được chuyển đổi sang định dạng số để sử dụng trong mô hình hồi quy. Cụ thể:

1) Chuyển đổi tốc độ bộ nhớ sang Hz:
   - Cột Memory Speed chứa các giá trị tần số kèm theo đơn vị như "MHz" hoặc "GHz." Để chuẩn hóa, chuyển đổi tất cả các giá trị này sang đơn vị Hz, đơn vị đo lường tần số trong hệ SI. Ví dụ, "3200 MHz" sẽ được chuyển đổi thành "3,200,000,000 Hz," và "4.1 GHz" sẽ thành "4,100,000,000 Hz."

	Memory Speed
2	3200 MHz
5	2400 MHz
10	2.6 GHz
11	2400 MHz
15	1.8 GHz
16	2400 MHz
19	3200 MHz
20	3200 MHz
27	2.8 GHz
30	5200 MHz

	Memory Speed(Hz)
2	3.200000e+09
5	2.400000e+09
10	2.600000e+09
11	2.400000e+09
15	1.800000e+09
16	2.400000e+09
19	3.200000e+09
20	3.200000e+09
27	2.800000e+09
30	5.200000e+09

Hình 3.3: Dữ liệu cột Memory Speed trước và sau chuyển đổi

2) Chuyển đổi khối lượng sang đơn vị Pound:
   - Cột Weight có thể bao gồm các đơn vị như "ounces" hoặc "pounds". Để đồng nhất, tất cả khối lượng được chuyển sang pound. Một khối lượng "1 ounce" sẽ được chuyển đổi thành "0.0625 pounds".

	Item Weight
2	3.92 pounds
5	3.24 pounds
10	4.45 pounds
11	0.634 ounces
15	3 pounds
16	3.74 pounds
19	3.52 pounds
20	3.3 pounds
27	2.65 pounds
30	4.54 pounds

	Item Weight(pounds)
2	3.920000
5	3.240000
10	4.450000
11	0.039625
15	3.000000
16	3.740000
19	3.520000
20	3.300000
27	2.650000
30	4.540000

Hình 3.4: Dữ liệu cột Weight trước và sau chuyển đổi

3) Loại bỏ đơn vị "Inches" khỏi cột kích thước màn hình:
   - Cột Standing screen display size chứa kích thước màn hình với đơn vị "Inches". Đơn vị này được loại bỏ và chỉ giữ lại giá trị

số để dễ dàng sử dụng trong phân tích. Ví dụ, giá trị "15.6 Inches" sẽ thành "15.6".

	Standing screen display size
2	15.6 Inches
5	14 Inches
10	14 Inches
11	14
15	13.3 Inches
16	15.6 Inches
19	15.6 Inches
20	14 Inches
27	11.6 Inches
30	15 Inches

	Standing screen display size(Inches)
2	15.6
5	14.0
10	14.0
11	14.0
15	13.3
16	15.6
19	15.6
20	14.0
27	11.6
30	15.0

Hình 3.5: Dữ liệu cột Standing screen display size trước và sau chuyển đổi

4) Làm sạch cột RAM:
   - Cột RAM có chứa dữ liệu kích thước RAM kèm theo đơn vị, chẳng hạn như "8 GB". Chúng tôi loại bỏ đơn vị và chuyển đổi giá trị thành số nguyên hoặc số thực để thuận tiện cho việc phân tích. Ví dụ, "8 GB" sẽ được chuyển thành "8".

	RAM
2	8 GB LPDDR5
5	8 GB DDR4
10	4 GB DDR4
11	4 GB DDR4
15	8 GB LPDDR3
16	16 GB DDR4
19	16 GB DDR4
20	16 GB DDR4
27	4 GB LPDDR4
30	16 GB DDR5

	RAM(GB)
2	8
5	8
10	4
11	4
15	8
16	16
19	16
20	16
27	4
30	16

Hình 3.6: Dữ liệu cột RAM trước và sau chuyển đổi

5) Làm sạch cột Processor:
   - Cột Processor chứa thông tin về loại và tốc độ của bộ xử lý, ví dụ "4.1 GHz ryzen_3". Chúng tôi tách riêng các thông tin này thành hai phần: tốc độ (sau khi chuyển sang Hz như ở bước 1)

và loại bỏ bộ xử lý, để sử dụng phù hợp trong phân tích hồi quy.

	Processor		Processor(GHz)
2	4.1 GHz ryzen_3	2	4.1
5	1.1 GHz celeron_n	5	1.1
10	2.6 GHz celeron	10	2.6
11	2.6 GHz celeron_n	11	2.6
15	1.8 GHz core_i5	15	1.8
16	4.1 GHz apple_ci3	16	4.1
19	3.7 GHz pentium	19	3.7
20	2.6 GHz celeron	20	2.6
27	2.8 GHz celeron	27	2.8
30	5.2 GHz core_i7	30	5.2

Hình 3.7: Dữ liệu cột Processor trước và sau chuyển đổi

6) Làm sạch cột Hard Drive:
   - Cột Hard Drive chứa thông tin về dung lượng ổ cứng, ví dụ "128 GB SSD". Chúng tôi sẽ chỉ lấy dung lượng là "128" để phân tích.

	Hard Drive		Hard Drive(GB)
2	128 GB SSD	0	128.0
5	64 GB SSD	1	64.0
10	64 GB SSD	2	64.0
11	64 GB Emmc	3	64.0
15	128 GB	4	128.0
16	1 TB SSD	5	1000.0
19	128 GB SSD	6	128.0
20	64 GB SSD	7	64.0
27	32 GB Embedded MultiMediaCard	8	32.0
30	1 TB SSD	9	1000.0

Hình 3.8: Dữ liệu cột Hard Drive trước và sau chuyển đổi

3.2.1.2. Tóm lược dữ liệu

Sau khi tiến hành xóa các giá trị khuyết cũng như chuyển các cột dữ liệu thích hợp về kiểu số để tiện cho tính toán, ta có thể thấy kiểu dữ liệu của các cột được thể hiện ở hình bên dưới.

```
Data columns (total 13 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Brand                                 1589 non-null   object
 1   Standing screen display size(Inches)  1589 non-null   float64
 2   Processor(GHz)                        1589 non-null   float64
 3   RAM(GB)                               1589 non-null   int64
 4   Memory Speed(Hz)                      1589 non-null   float64
 5   Hard Drive(GB)                        1589 non-null   float64
 6   Graphics Coprocessor                  1589 non-null   object
 7   Chipset Brand                         1589 non-null   object
 8   Card Description                      1589 non-null   object
 9   Processor Brand                       1589 non-null   object
 10  Operating System                      1589 non-null   object
 11  Item Weight(pounds)                   1589 non-null   float64
 12  Price($)                              1589 non-null   float64
dtypes: float64(6), int64(1), object(6)
memory usage: 161.5+ KB
```

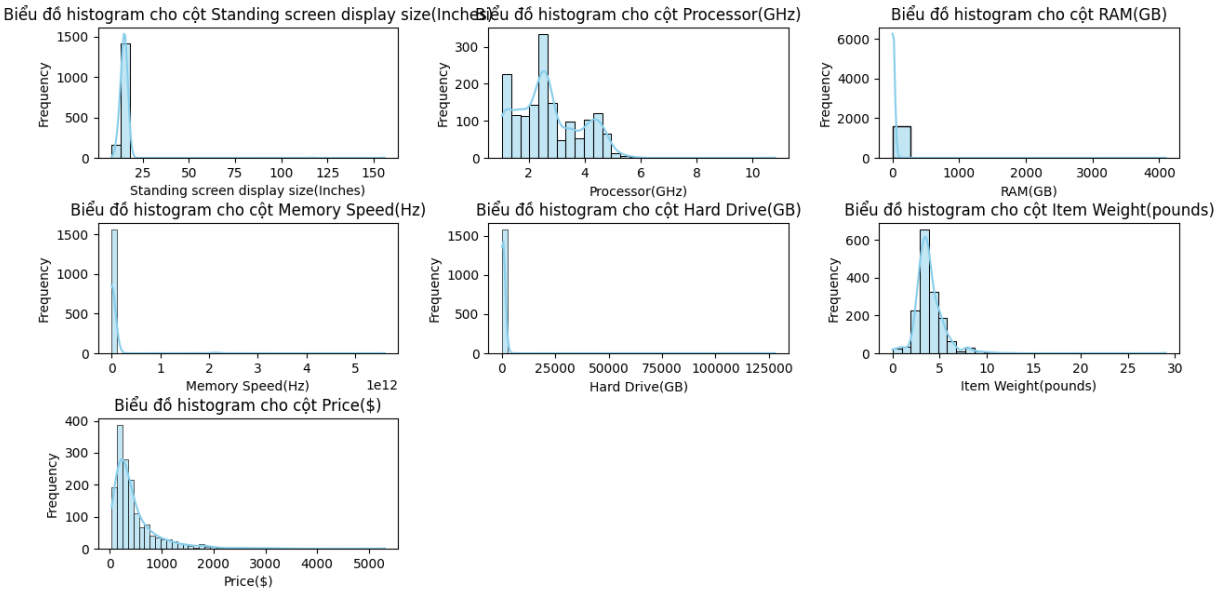
Hình 3.9: Kiểu dữ liệu tất cả các cột sau khi được làm sạch

Để có cái nhìn tổng quát về dữ liệu này, bảng thống kê mô tả được lập ra để hiểu rõ hơn về đặc trưng của các cột dữ liệu dạng số. Bảng này thể hiện số các giá trị (*count*), giá trị trung bình (*mean*), độ lệch chuẩn (*std*), các giá trị cực trị (*min* và *max*), cũng như tứ phân vị của dữ liệu (*25%*, *50%*, *75%*).

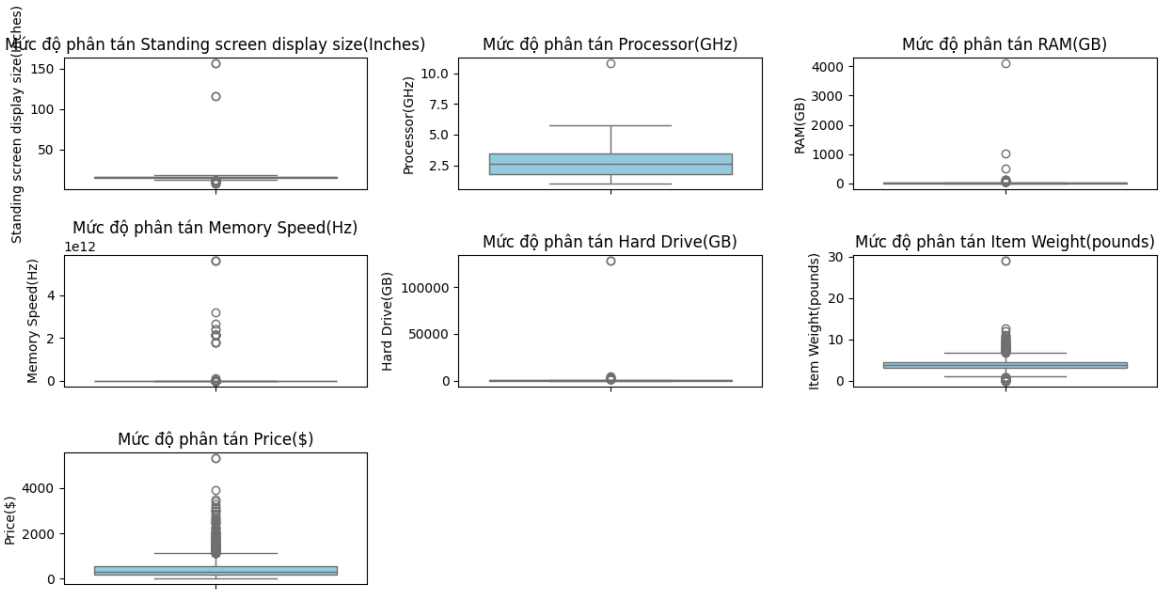
	Standing screen display size(Inches)	Processor(GHz)	RAM(GB)	Memory Speed(Hz)	Hard Drive(GB)	Item Weight(pounds)	Price($)
count	1589.000000	1589.000000	1589.000000	1.589000e+03	1589.000000	1589.000000	1589.000000
mean	14.972026	2.701246	15.700441	4.480794e+10	615.515772	3.924764	485.625758
std	6.335969	1.136559	106.865229	3.557609e+11	4545.218547	1.715258	509.537368
min	8.000000	1.040000	2.000000	2.100000e+00	0.062500	0.000625	33.880000
25%	14.000000	1.800000	4.000000	2.400000e+09	128.000000	3.100000	190.000000
50%	15.600000	2.600000	8.000000	2.666000e+09	256.000000	3.700000	325.000000
75%	15.600000	3.500000	16.000000	3.200000e+09	512.000000	4.520000	568.990000
max	156.000000	10.800000	4096.000000	5.600000e+12	128000.000000	29.000000	5299.000000

Hình 3.10: Bảng thống kê mô tả các cột dữ liệu kiểu số

Sau khi tóm tắt các thông tin thống kê mô tả, chúng ta sẽ vẽ biểu đồ histogram và boxplot (whisker plot) để kiểm tra mức độ tập trung, phân tán của các cột dữ liệu. Hai loại biểu đồ này giúp phát hiện các giá trị ngoại lai, điểm trung vị, khoảng phân phối của dữ liệu, v.v.



Hình 3.11: Biểu đồ histogram cho các cột dữ liệu kiểu số



Hình 3.12: Biểu đồ boxplot cho các cột dữ liệu kiểu số

Từ bảng và các biểu đồ trên, ta thấy:

1) **Standing screen display size(Inches):** Có độ trải giữa hẹp, cho thấy các mẫu laptop hiện nay tập trung chủ yếu từ khoảng 14 – 16 inches tuy nhiên có giá trị ngoại lai đặc biệt là 156 inches khả năng cao là sai số khi nhập dữ liệu.
2) **Processor(GHz):** Dữ liệu phân bố từ khoảng 1.8 – 3.5 GHz nhưng có dữ liệu lớn nhất là 10 GHz. Giá trị ngoại lai này sẽ được loại bỏ

trong các quy trình tiếp theo do tốc độ xung nhịp CPU lớn nhất hiện nay cho thấy rơi vào khoảng 5 GHz [4].

3) **RAM(GB):** Bộ nhớ RAM được biểu thị cho thấy dữ liệu thực tế với các mẫu máy tính xách tay hiện nay, từ 4 – 16 GB. Cá biệt có dữ liệu lớn nhất lên tới 4096 GB sẽ được loại bỏ trong quá trình tính toán.
4) **Memory Speed(Hz):** Tốc độ RAM có mức độ tập trung lớn ở khoảng từ 2 – 3 MHz và có những giá trị trên 5 MHz xuất hiện ở các mẫu laptop cao cấp.
5) **Hard Drive(GB):** Bộ nhớ máy tính tập trung ở các giá trị 128, 256 và 512 GB như thực tế hiện nay cho thấy dữ liệu khá chính xác tuy nhiên lại xuất hiện giá trị cực tiểu 0.0625 và 128000 có thể là sai số trong quá trình lấy dữ liệu.
6) **Item Weight(pounds):** Cân nặng của các máy tính có độ trải giữa nhỏ ở khoảng 3 – 4.5 pounds (khoảng 1.35 – 2 Kg) nhưng lại xuất hiện giá trị không hợp lí là 29 pounds (khoảng 13 Kg) hay 0.000625 pounds (khoảng 0,0002835 Kg, nhẹ hơn một tờ giấy A4)
7) **Price($):** Giá của các mẫu máy tính xách tay có độ trải giữa lớn và cho thấy dữ liệu lệch trái (dương) với nhiều giá trị nằm trong khoảng từ 200 – 550 USD (Khoảng 5,028,400 – 13,828,100 VND theo tỉ giá tờ đô la hiện tại[11]). Cá biệt có mẫu có giá lên tới 5,000 USD (Khoảng 125,710,000 VND)

3.2.1.3. Chuyển đổi dữ liệu

Tóm lược dữ liệu cho chúng ta cái nhìn tổng quan về bộ dữ liệu tuy nhiên còn các cột phi số còn chưa được đề cập. Các thuộc tính như thương hiệu hay hệ điều hành ảnh hưởng khá nhiều đến quyết định mua hàng nên không thể loại bỏ các thuộc tính này mà cần chuyển về dạng phù hợp để phân tích. Trong báo cáo này, chúng tôi áp dụng kĩ thuật **Label Encoding** để ánh xạ các cột kiểu

chuỗi sang kiểu số. Bên dưới là danh sách một số ánh xạ cho các cột được chuyển đổi.

```
Mapping dữ liệu của cột 'Brand':
  acer -> 51
  HP -> 23
  Apple -> 9
Mapping dữ liệu của cột 'Graphics Coprocessor':
  AMD Radeon Graphics -> 16
  Intel UHD Graphics 600 -> 103
  Intel -> 52
Mapping dữ liệu của cột 'Chipset Brand':
  AMD -> 0
  Intel -> 10
  NVIDIA -> 14
Mapping dữ liệu của cột 'Card Description':
  Integrated -> 9
  RTX 4070 -> 45
  RTX 4060 -> 44
Mapping dữ liệu của cột 'Processor Brand':
  AMD -> 0
  Intel -> 3
  Qualcomm -> 7
Mapping dữ liệu của cột 'Operating System':
  Windows 11 S -> 24
  Windows 11 Home -> 21
  Windows 11 -> 19
```

Hình 3.13: Một số ánh xạ dữ liệu kiểu chuỗi sang kiểu số của các cột phi số

3.2.2. Phân tích mô tả

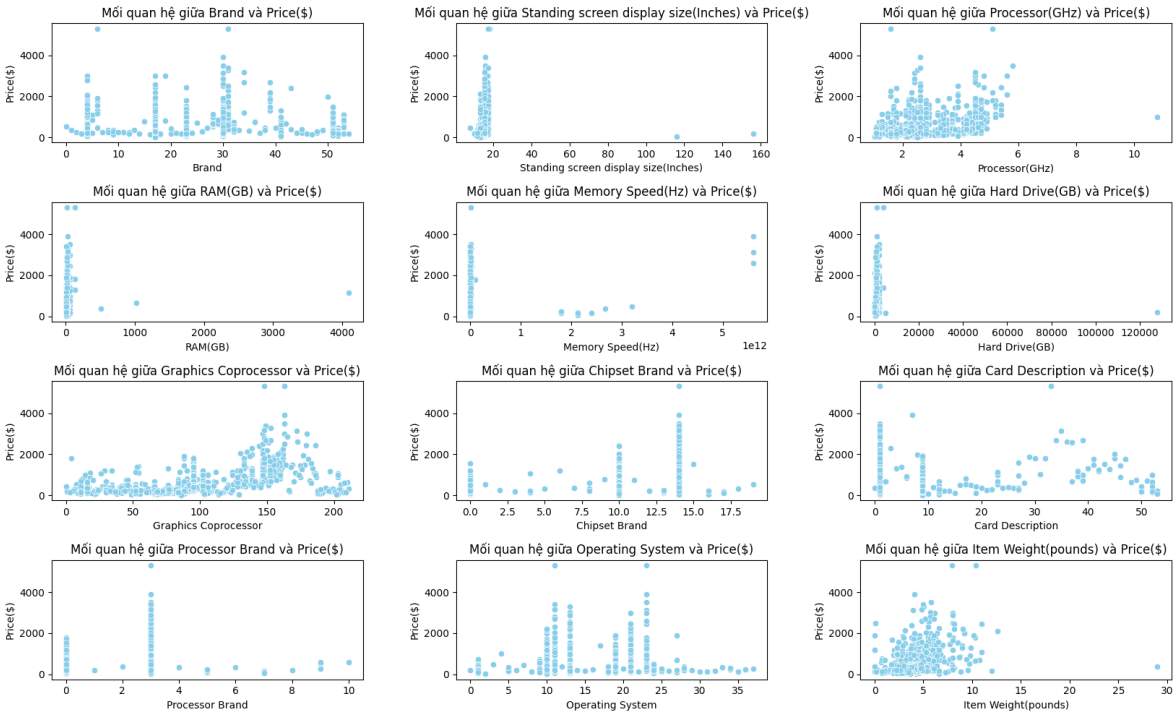
Sau khi tiến hành tiền xử lý dữ liệu bao gồm loại bỏ các giá trị thiếu, làm sạch dữ liệu về dạng phù hợp và chuyển đổi các dữ liệu định lượng về dạng số, chúng tôi thu được một tập dữ liệu đầy đủ, sẵn sàng cho phân tích. Mục tiêu của phần này là khám phá các đặc trưng quan trọng trong dữ liệu thông qua các biểu đồ phân bố nhằm hiểu rõ các đặc điểm chính, mối tương quan giữa các biến số, và đánh giá sự sẵn sàng của dữ liệu cho bước dự báo giá laptop bằng mô hình hồi quy tuyến tính ở các phần tiếp theo.

Để biết được sự ảnh hưởng của các thuộc tính với giá của sản phẩm, ta có hình phía dưới.

```
Tương quan của các thuộc tính với giá (Price($)):
Price($)                                1.000000
Graphics Coprocessor                    0.366602
Item Weight(pounds)                     0.358916
Processor(GHz)                          0.350643
Chipset Brand                           0.282447
Operating System                        0.222398
Memory Speed(Hz)                        0.104182
RAM(GB)                                 0.090495
Standing screen display size(Inches)    0.056859
Hard Drive(GB)                          0.033494
Processor Brand                        -0.008149
Card Description                       -0.029556
Brand                                  -0.050585
Name: Price($), dtype: float64
```

Hình 3.14: Tương quan của các thuộc tính với thuộc tính giá

Từ đó ta có biểu đồ Scatter thể hiện tương quan các cột thuộc tính với thuộc tính Price($):



Hình 3.15: Biểu đồ Scatter của các cột dữ liệu với thuộc tính giá (Price($))

Nhận xét:

- Các thuộc tính như Standing screen display size(Inches), Processor(GHz), RAM(GB), Memory Speed(Hz) và Hard Drive(GB) cho thấy mối quan hệ tăng dần với giá trị Price($). Những laptop có RAM lớn, bộ xử lý mạnh mẽ hơn, màn hình rộng hơn, và dung lượng ổ cứng cao hơn có xu hướng có giá cao hơn.
- Thuộc tính Item Weight(pounds) có mối quan hệ không rõ ràng với Price($). Thuộc tính này có sự phân tán rộng và không cho thấy mối quan hệ tuyến tính mạnh mẽ với giá.
- Một số thuộc tính như Graphics Coprocessor, ChipsetBrand, và Operating System có sự phân bố điểm dữ liệu rất phân tán và không cho thấy mối quan hệ chặt chẽ với giá laptop. Điều này có thể do những yếu tố này ảnh hưởng ít đến giá trị thực tế của laptop so với những yếu tố như bộ xử lý và dung lượng RAM.

Như vậy, phân tích mô tả cho thấy rằng dữ liệu có sự phân bố đa dạng và chứa nhiều thông tin giá trị. Những nhận định này tạo cơ sở cho các bước tiếp theo, cụ thể là xây dựng mô hình hồi quy tuyến tính để dự đoán giá laptop dựa trên các thuộc tính đầu vào.

3.2.3. Tạo và huấn luyện mô hình hồi quy

Trong bước này, chúng tôi xây dựng mô hình hồi quy tuyến tính để dự đoán giá máy tính xách tay dựa trên các đặc trưng đã được làm sạch và chuyển đổi từ dữ liệu ban đầu. Quy trình tạo và huấn luyện mô hình bao gồm các bước chính như chuẩn bị dữ liệu đầu vào và mục tiêu, xử lý giá trị ngoại lai, chia tập dữ liệu thành tập huấn luyện và kiểm tra, huấn luyện mô hình và đánh giá hiệu suất của mô hình dự đoán.

Chuẩn bị dữ liệu

Để mô hình hoạt động hiệu quả, chúng tôi tách dữ liệu thành:

- **Biến đầu vào (X)**: chứa các cột đặc trưng như thương hiệu, kích thước màn hình, bộ vi xử lý, RAM, tốc độ bộ nhớ, card đồ họa, hệ điều hành... đã được chuẩn hóa và chuyển đổi.
- **Biến mục tiêu (y)**: cột giá sản phẩm *Price($)* được sử dụng làm biến mục tiêu, là giá trị mà mô hình sẽ dự đoán.

Xử lí các giá trị ngoại lai

Để giảm thiểu tác động của các giá trị ngoại lệ, áp dụng kỹ thuật *clipping*. Dữ liệu của mỗi cột trong **X** được giới hạn trong khoảng từ **1% đến 99%**. Điều này đảm bảo rằng các giá trị bất thường không làm ảnh hưởng đáng kể đến hiệu quả của mô hình hồi quy.

Chia tập dữ liệu và chuẩn hóa

Sau khi chuẩn bị dữ liệu, chia chúng ngẫu nhiên thành 2 phần: tập huấn luyện và tập kiểm tra với tỷ lệ 80:20. Tập huấn luyện sẽ được sử dụng để huấn luyện mô hình, trong khi tập kiểm tra sẽ dùng để đánh giá mô hình đã huấn luyện. Để đảm bảo rằng tất cả các đặc trưng có cùng tỷ lệ ảnh hưởng đến mô hình, dữ liệu đầu vào được chuẩn hóa bằng phương pháp *StandardScaler* (Chuẩn hóa z-scores) sao cho trung bình của mỗi đặc trưng là 0 và độ lệch chuẩn là 1.

Huấn luyện mô hình

Mô hình hồi quy tuyến tính được khởi tạo và huấn luyện bằng cách sử dụng tập huấn luyện. Hồi quy tuyến tính là phương pháp dự báo tuyến tính đơn giản và dễ giải thích, giúp xác định mức độ ảnh hưởng của từng đặc trưng lên giá trị dự đoán của biến mục tiêu.

Đánh giá mô hình và các chỉ số

Sau khi tạo và huấn luyện mô hình hồi quy tuyến tính, chúng tôi đánh giá hiệu quả của mô hình dựa trên tập dữ liệu kiểm tra bằng cách sử dụng các chỉ số đánh giá phổ biến, bao gồm Mean Squared Error (MSE) và R-squared (R^2). Các chỉ số này giúp đo lường độ chính xác của mô hình dự báo giá máy tính

xách tay, từ đó đánh giá xem mô hình có đạt yêu cầu để áp dụng vào thực tế hay không.

```
Mean Squared Error (MSE): 82877.3567357474
R-squared (R2): 0.5095376461409411
```

Hình 3.16: MSE và R-squared

3.3. Phân tích kết quả dự đoán

3.3.1. Kết quả đạt được

Để tăng tính tin cậy của thông số, chúng tôi cho thực hiện chạy 5 lần rồi tính trung bình và được bảng kết quả bên dưới:

Bảng 3.2: Kết quả các lần chạy

Thông số	**Lần 1**	**Lần 2**	**Lần 3**	**Lần 4**	**Lần 5**	**TB**
R^2	0.51	0.55	0.54	0.48	0.50	0.516
MSE	82877	128539	118546	126983	148530	121095
Thời gian thực hiện (giây)	2.9760	3.1164	2.9184	3.2298	2.8718	3.02248

3.3.2. Phân tích chi tiết

Sau khi huấn luyện và kiểm tra mô hình hồi quy tuyến tính trên tập dữ liệu đã qua tiền xử lý, các kết quả thực nghiệm được ghi nhận như sau:

1) R^2 (Hệ số xác định): Giá trị trung bình $R^2 = 0.516$ cho thấy mô hình hồi quy tuyến tính giải thích được khoảng **51,6% sự biến thiên của giá laptop** trong dữ liệu. Đây là một kết quả khả quan, nhưng còn nhiều dư địa để cải thiện độ chính xác của mô hình. Giá trị R^2 dao động từ **0.48** đến **0.55**, phản ánh sự ổn định tương đối trong các lần chạy thử.

2) MSE (Mean Squared Error): Sai số bình phương trung bình (MSE) trung bình đạt 121095, cho thấy độ chênh lệch giữa giá trị dự đoán và giá trị thực tế. Tuy nhiên, sự chênh lệch lớn giữa các lần chạy, ví dụ từ 82877 (Lần 1) đến 148530 (Lần 5), chỉ ra rằng mô hình có thể bị ảnh hưởng bởi

các yếu tố như biến động trong dữ liệu kiểm tra hoặc chưa tối ưu trong việc loại bỏ nhiễu.

3) **Thời gian thực hiện**: Thời gian chạy trung bình của mỗi lần thực nghiệm là khoảng 3.02 giây, cho thấy mô hình hoạt động khá hiệu quả trên hệ thống. Chênh lệch thời gian nhỏ giữa các lần chạy phản ánh sự ổn định trong hiệu năng của thuật toán hồi quy tuyến tính.

3.3.3. Đánh giá và nhận xét

Ưu điểm:

- Mô hình hồi quy tuyến tính đơn giản nhưng có khả năng dự đoán mức giá laptop khá tốt, đặc biệt khi xem xét khối lượng dữ liệu và số lượng đặc trưng.
- Thời gian chạy nhanh, phù hợp cho việc xử lý và dự đoán trên các bộ dữ liệu lớn.

Nhược điểm:

- Giá trị $R^2 = 0.516$ cho thấy mô hình chỉ đạt mức dự đoán trung bình, chưa đủ cao để ứng dụng cho các hệ thống dự báo yêu cầu độ chính xác cao.
- MSE lớn và có sự biến động giữa các lần chạy, cho thấy mô hình có thể bị ảnh hưởng bởi các nhiễu trong dữ liệu hoặc chưa khai thác tốt các mối quan hệ phi tuyến giữa các đặc trưng.

Hướng cải thiện:

- **Tăng cường xử lý dữ liệu**: Bổ sung các đặc trưng liên quan hoặc áp dụng các kỹ thuật chọn lọc đặc trưng.
- **Thử nghiệm mô hình nâng cao**: Áp dụng các thuật toán phi tuyến như Random Forest, Gradient Boosting hoặc Neural Networks để khai thác các mối quan hệ phức tạp hơn trong dữ liệu.
- **Xử lý nhiễu và ngoại lệ**: Nghiên cứu thêm các phương pháp giảm ảnh hưởng của các giá trị ngoại lệ.

3.3.4. Kết luận

Mô hình hồi quy tuyến tính đã cung cấp một cái nhìn tổng quan và tạo cơ sở ban đầu để dự đoán giá laptop. Tuy nhiên, để đạt được độ chính xác cao hơn, cần phải cải thiện thêm ở các bước tiền xử lý và lựa chọn mô hình.

3.4. Kết luận chương 3

Chương 3 này nhóm đã xây dựng được cơ sở vững chắc cho việc dự đoán giá laptop thông qua mô hình hồi quy tuyến tính. Tuy nhiên, các phân tích cũng chỉ ra một số hạn chế của mô hình, đặc biệt trong việc xử lý các mối quan hệ phi tuyến giữa các đặc trưng. Do đó, các chương sau sẽ tập trung vào việc cải thiện mô hình bằng cách thử nghiệm các thuật toán nâng cao hơn hoặc tối ưu hóa các bước xử lý dữ liệu.

KẾT LUẬN

Trong bài tiểu luận, chúng em tập trung vào phân tích và dự báo giá máy tính xách tay bằng cách sử dụng mô hình hồi quy tuyến tính. Bằng việc thu thập và xử lý các dữ liệu liên quan đến các yếu tố như thương hiệu, tốc độ bộ nhớ, RAM và các yếu tố khác, chúng ta sẽ xây dựng một mô hình hồi quy tuyến tính nhằm dự đoán giá máy tính trong thị trường công nghệ. Mục tiêu của nghiên cứu này là cung cấp một công cụ hữu ích cho người dùng và chuyên gia trong lĩnh vực công nghệ để đưa ra quyết định dựa trên dự báo giá sản phẩm chính xác.

Kết quả phân tích và dự báo sử dụng mô hình hồi quy tuyến tính cho thấy mô hình có khả năng dự báo giá sản phẩm với độ chính xác tương đối. Tuy nhiên, cần lưu ý rằng mô hình hồi quy tuyến tính có một số giới hạn và nhược điểm như đòi hỏi mối quan hệ tuyến tính giữa biến đầu vào và giá sản phẩm. Điều này có thể không phù hợp trong trường hợp dữ liệu có tính chất phi tuyến.

Dựa trên kết quả phân tích và dự báo, chúng ta có thể đưa ra một số kiến nghị nhằm cải thiện quá trình dự báo giá sản phẩm trong lĩnh vực thương mại:

- **Tăng cường thu thập dữ liệu:** Cần tập trung vào việc thu thập dữ liệu chi tiết hơn về các yếu tố quan trọng ảnh hưởng đến giá sản phẩm như tốc độ sử ký bộ nhớ, thương hiệu sản phẩm, hiệu suất sản phẩm và tuổi thọ pin. Điều này giúp cung cấp thông tin đầy đủ và đáng tin cậy cho mô hình dự báo.
- **Sử dụng mô hình phức tạp hơn:** Trong tương lai, cần xem xét sử dụng các phương pháp hồi quy phi tuyến hoặc mô hình học máy phức tạp hơn như mạng nơ-ron, hoặc máy vector hỗ trợ (SVM) để đối phó tốt hơn với sự phức tạp và phi tuyến của dữ liệu. Những mô hình này có thể xử lý tốt hơn các yếu tố phi tuyến và tương tác phức tạp giữa các biến đầu vào.

- **Mở rộng dữ liệu và phạm vi:** Để cải thiện độ chính xác của dự báo, cần thu thập dữ liệu từ nhiều ngành công nghiệp khác nhau và mở rộng phạm vi đối tượng nghiên cứu. Điều này giúp chúng ta hiểu rõ hơn về tác động của các yếu tố khác nhau đến giá và có cái nhìn tổng quan về thị trường thương mại.
- **Khám phá phương pháp phân tích sâu hơn:** Ngoài mô hình hồi quy tuyến tính, cần khám phá các phương pháp phân tích sâu hơn như phân tích chuỗi thời gian để tìm ra sự biến đổi của giá máy tính theo thời gian và các yếu tố ảnh hưởng. Phân tích chuỗi thời gian có thể giúp chúng ta phát hiện xu hướng, mùa vụ và các yếu tố biến đổi khác có thể ảnh hưởng đến giá trị sản phẩm.

Tổng kết lại, nghiên cứu về phân tích và dự báo giá máy tính sử dụng mô hình hồi quy tuyến tính đã mang lại những kết quả khá khả quan. Tuy nhiên, để cải thiện dự báo và đáp ứng tốt hơn với sự phức tạp của thị trường thương mại, cần tiếp tục mở rộng nghiên cứu.

TÀI LIỆU THAM KHẢO

[1] Slide bài giảng học phần Phân tích dữ liệu lớn: https://docs.google.com/presentation/d/1goDrQmjT0oLHFLfi4NOgpT75gci67_hJ/, 4/12/2024.

[2] Douglas C. Montgomery, Introduction to Linear Regression Analysis. John Wiley & Sons, Inc, 2018.

[3] Douglas C. Montgomery, Introduction to Time Series Analysis and Forecasting. Wiley & Sons, Inc, 2015.

[4] Hỏi đáp Quora: https://www.quora.com/What-is-the-fastest-processor-currently-available-for-laptops-and-desktops, 4/12/2024.

[5] Chuyển đổi ngoại tệ Vietcombank: https://www.vietcombank.com.vn/vi-VN/KHCN/Cong-cu-Tien-ich/Ty-gia, 4/12/2024.

CODE

```
import pandas as pd
import numpy as np
import math
import re
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler, LabelEncoder,
PolynomialFeatures
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt
import seaborn as sns
import time
import platform
import psutil
# Bắt đầu đo thời gian
start_time = time.time()

laptops_df = pd.read_csv("laptops_dataset.csv")
print(laptops_df.head(10).to_string())

# Làm sạch dữ liệu
nan_counts = laptops_df.isna().sum()
for column, count in nan_counts.items():
    print(f"{column}: {count}")

selected_columns = [
    "Brand", "Standing screen display size", "Processor", "RAM",
```

```
    "Memory Speed", "Hard Drive", "Graphics Coprocessor",
    "Chipset Brand", "Card Description", "Processor Brand",
    "Operating System", "Item Weight", "Price($)"
]

laptops_df_cleaned = laptops_df[selected_columns].dropna()



# Chuyển Memory speed sang Hz
def convert_to_hz(speed):
    speed = "".join(filter(str.isprintable, speed))
    speed = speed.strip()
    if "GHz" in speed:
        return float(speed.replace("GHz", "").strip()) * 1e9
    elif "MHz" in speed:
        return float(speed.replace("MHz", "").strip()) * 1e6
    elif "KHz" in speed:
        return float(speed.replace("KHz", "").strip()) * 1e3
    else:
        return float(speed)



print(laptops_df_cleaned[['Memory Speed']].head(10).to_string(index=True,
header=True))
laptops_df_cleaned["Memory Speed"] = laptops_df_cleaned["Memory
Speed"].apply(convert_to_hz)
laptops_df_cleaned = laptops_df_cleaned.rename(columns={"Memory
Speed": "Memory Speed(Hz)"})
print(laptops_df_cleaned[['Memory
```

```
Speed(Hz)']].head(10).to_string(index=True, header=True))


# Chuyển trọng lượng sang pounds
def clean_and_convert_weight(weight_str):
    cleaned = re.sub(r"[^\d.]+", "", weight_str).lower()

    if "ounce" in weight_str:
        cleaned = float(cleaned) * 0.0625

    return float(cleaned)


print(laptops_df_cleaned[['Item Weight']].head(10).to_string(index=True,
header=True))
laptops_df_cleaned["Item Weight"] = laptops_df_cleaned["Item
Weight"].apply(clean_and_convert_weight)
laptops_df_cleaned = laptops_df_cleaned.rename(columns={"Item Weight":
"Item Weight(pounds)"})
print(laptops_df_cleaned[['Item
Weight(pounds)']].head(10).to_string(index=True, header=True))

# Loại bỏ chuỗi "Inches" trong Standing screen display size
print(laptops_df_cleaned[['Standing screen display
size']].head(10).to_string(index=True, header=True))

laptops_df_cleaned["Standing screen display size"] =
laptops_df_cleaned["Standing screen display size"].str.replace(
    " Inches", "")
```

```
laptops_df_cleaned["Standing screen display size"] =
laptops_df_cleaned["Standing screen display size"].str.replace(
    r"[^\d\.]", "", regex=True)
laptops_df_cleaned["Standing screen display size"] =
laptops_df_cleaned["Standing screen display size"].astype(
    float).round(2)
laptops_df_cleaned = laptops_df_cleaned.rename(
    columns={"Standing screen display size": "Standing screen display
size(Inches)"})
print(laptops_df_cleaned[['Standing screen display
size(Inches)']].head(10).to_string(index=True, header=True))


# Làm sạch Ram
def clean_ram(value):
    match = re.search(r"\d+", value)
    if match:
        ram_value = int(match.group())
        if "1 TB" in value:
            return 1024
        elif ram_value % 2 != 0:
            return int(math.pow(2, math.ceil(math.log2(ram_value))))
        else:
            return ram_value


print(laptops_df_cleaned[['RAM']].head(10).to_string(index=True,
header=True))
laptops_df_cleaned["RAM"] =
```

```
laptops_df_cleaned["RAM"].apply(clean_ram)
laptops_df_cleaned.dropna(subset=["RAM"], inplace=True)
laptops_df_cleaned = laptops_df_cleaned.rename(columns={"RAM":
"RAM(GB)"})
print(laptops_df_cleaned[['RAM(GB)']].head(10).to_string(index=True,
header=True))

# Làm sạch Processor
print(laptops_df_cleaned[['Processor']].head(10).to_string(index=True,
header=True))
laptops_df_cleaned["Processor"] = laptops_df_cleaned["Processor"].apply(
    lambda x: re.search(r"(\d+\.\d+)\s*GHz", x).group(1) if
re.search(r"(\d+\.\d+)\s*GHz", x) else None)
laptops_df_cleaned.dropna(subset=["Processor"], inplace=True)
laptops_df_cleaned["Processor"] =
laptops_df_cleaned["Processor"].astype(float)
laptops_df_cleaned = laptops_df_cleaned.rename(columns={"Processor":
"Processor(GHz)"})
print(laptops_df_cleaned[['Processor(GHz)']].head(10).to_string(index=True,
header=True))


# Làm sạch Hard Drive
def convert_hard_drive_size(size_str):
    match = re.search(r"(\d+\.?\d*)", size_str)
    if match:
        value = float(match.group(1))
        if ("TB" in size_str) and (not "GB" in size_str):
            value *= 1000
```

```
        elif "MB" in size_str:
            value /= 1024
        elif "GB" not in size_str:
            if 10 < value < 500:
                return value
            elif value < 10:
                value *= 1000
            else:
                value /= 1000000000
        return value
    return None


print(laptops_df_cleaned[['Hard Drive']].head(10).to_string(index=True,
header=True))
laptops_df_cleaned["Hard Drive"] = laptops_df_cleaned["Hard
Drive"].apply(convert_hard_drive_size)
laptops_df_cleaned = laptops_df_cleaned.dropna().reset_index(drop=True)
laptops_df_cleaned = laptops_df_cleaned.rename(columns={"Hard Drive":
"Hard Drive(GB)"})
print(laptops_df_cleaned[['Hard Drive(GB)']].head(10).to_string(index=True,
header=True))

print("\nDữ liệu sau khi làm sạch: ")
print(laptops_df_cleaned.head(10).to_string())

# Tóm lược dữ liệu
print("\n")
print(laptops_df_cleaned.describe().to_string())
```

```
df_numeric = laptops_df_cleaned.select_dtypes(include=['number'])
# Vẽ boxplot cho tất cả các thuộc tính kiểu số
plt.figure(figsize=(15, 10))
for i, col in enumerate(df_numeric, 1):
    plt.subplot(4, 3, i)
    sns.boxplot(y=df_numeric[col], color='skyblue')
    plt.title(f'Mức độ phân tán {col}')
    plt.ylabel(col)

plt.subplots_adjust(wspace=0.3, hspace=0.5)
plt.show()

# Vẽ histogram cho các thuộc tính kiểu số
plt.figure(figsize=(15, 10))
for i, col in enumerate(df_numeric, 1):
    plt.subplot(4, 3, i)
    # Điều chỉnh bins dựa trên đặc điểm của từng cột
    if df_numeric[col].nunique() < 20:
        bins = df_numeric[col].nunique()
    elif df_numeric[col].max() - df_numeric[col].min() > 1000:
        bins = 50
    else:
        bins = 30
    sns.histplot(df_numeric[col], kde=True, bins=bins, color='skyblue')
    plt.title(f'Biểu đồ histogram cho cột {col}')
    plt.xlabel(col)
    plt.ylabel('Frequency')
```

```
plt.subplots_adjust(wspace=0.3, hspace=0.5)
plt.show()

# Chuyển đổi dữ liệu
# Chuyển đổi các cột phân loại sang dạng số bằng Label Encoding
label_encoder = LabelEncoder()

categorical_columns = ["Brand", "Graphics Coprocessor", "Chipset Brand",
                       "Card Description", "Processor Brand", "Operating System"]

# for col in categorical_columns:
#     laptops_df_cleaned[col] =
label_encoder.fit_transform(laptops_df_cleaned[col])

mapping_dict = {}
for col in categorical_columns:
    # Lấy danh sách giá trị gốc và giá trị đã mã hóa
    original_values = laptops_df_cleaned[col].unique()
    encoded_values = label_encoder.fit_transform(original_values)
    mapping_dict[col] = dict(zip(original_values, encoded_values))
    laptops_df_cleaned[col] =
label_encoder.fit_transform(laptops_df_cleaned[col])

for col, mapping in mapping_dict.items():
    print(f"Mapping dữ liệu của cột '{col}':")
    for original, encoded in list(mapping.items())[:3]:
        print(f"  {original} -> {encoded}")
```

```
# Tính hệ số tương quan giữa các thuộc tính và giá trị mục tiêu Price($)
correlation_matrix = laptops_df_cleaned.corr()
price_correlation =
correlation_matrix["Price($)"].sort_values(ascending=False)
print("\nTương quan của các thuộc tính với giá (Price($)):")
print(price_correlation)

# Vẽ biểu đồ scatter cho tất cả các thuộc tính với Price($)
plt.figure(figsize=(20, 15))
for i, col in enumerate(laptops_df_cleaned.columns[:-1], 1):
    plt.subplot(5, 3, i)
    sns.scatterplot(x=laptops_df_cleaned[col],
y=laptops_df_cleaned["Price($)"], color='skyblue')
    plt.title(f'Mối quan hệ giữa {col} và Price($)')
    plt.xlabel(col)
    plt.ylabel('Price($)')

plt.subplots_adjust(wspace=0.3, hspace=0.5)
plt.show()

# Tạo và huấn luyện mô hình hồi quy
# Tạo các biến đầu vào X và biến mục tiêu y
X = laptops_df_cleaned.drop("Price($)", axis=1)
y = laptops_df_cleaned["Price($)"]



# Handle outliers by clipping the data to the 1st and 99th percentiles
def clip_outliers(df, columns):
    for col in columns:
```

```
        lower = df[col].quantile(0.01)
        upper = df[col].quantile(0.99)
        df[col] = np.clip(df[col], lower, upper)
    return df


X = clip_outliers(X, X.columns)


# Create a pipeline to include polynomial features and standard scaling
pipeline = Pipeline([
    ("poly_features", PolynomialFeatures(degree=2, include_bias=False)),
    ("scaler", StandardScaler()),
    ("linear_regression", LinearRegression())
])


# Chia dữ liệu thành tập huấn luyện và kiểm tra (train/test = 8/2)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

# Khởi tạo và huấn luyện mô hình hồi quy tuyến tính
model = LinearRegression()
model.fit(X_train, y_train)

# Dự đoán trên tập kiểm tra
y_pred = model.predict(X_test)

# Đánh giá mô hình
mse = mean_squared_error(y_test, y_pred)
```

```
r2 = r2_score(y_test, y_pred)

print("Mean Squared Error (MSE):", mse)
print("R-squared (R2):", r2)



# Kết thúc đo thời gian
end_time = time.time()
# Tính toán thời gian thực thi
execution_time = end_time - start_time

# Lấy thông tin cơ bản về hệ thống
system_info = {
    "Hệ điều hành": platform.system(),
    "Kiến trúc máy": platform.machine(),
    "Bộ xử lý": platform.processor(),
    "Số lõi CPU": psutil.cpu_count(logical=False),  # Số lõi vật lý
    "Số lõi CPU logic": psutil.cpu_count(logical=True),  # Số lõi logic (bao
gồm cả hyperthreading)
    "Bộ nhớ": round(psutil.virtual_memory().total / (1024 ** 3), 2),  # Bộ nhớ
tính theo GB
}

# In ra thông số hệ thống
print("\nThông số hệ thống:")
for key, value in system_info.items():
    print(f"{key}: {value}")

# Lấy thông tin về tình trạng sử dụng CPU và bộ nhớ
```

```
cpu_usage = psutil.cpu_percent(interval=1)  # Sử dụng CPU theo phần trăm
memory_usage = psutil.virtual_memory().percent  # Sử dụng bộ nhớ theo
phần trăm

print(f"\nTình trạng sử dụng CPU: {cpu_usage}%")
print(f"Tình trạng sử dụng bộ nhớ: {memory_usage}%")
print(f"Thời gian thực thi: {execution_time:.4f} giây")
```