TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN**

**NHẬP MÔN HỌC MÁY**

**ỨNG DỤNG THUẬT TOÁN LINEAR REGRESSION ĐỂ**

**DỰ ĐOÁN GIÁ CẢ CỦA LAPTOP**

|  |  |
| --- | --- |
| **Sinh viên thực hiện : PHẠM DUY PHÚC NGUYÊN**  **NGUYỄN CHÍ CƯỜNG**  **Giảng viên hướng dẫn : VŨ VĂN ĐỊNH**  **Ngành : CÔNG NGHỆ THÔNG TIN**  **Chuyên Ngành : CÔNG NGHỆ PHẦN MỀM**  **Lớp : D16CNPM5**  **Khóa : 2021 - 2026** |  |
|  |  |

***Hà Nội, tháng 04 năm 2024***

**PHIẾU CHẤM ĐIỂM**

Sinh viên thực hiện:

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **Chữ ký** | **Nội dung thực hiện** |
| Phạm Duy Phúc Nguyên  21810310118 |  |  |
| Nguyễn Chí Cường  21810310494 |  |  |

Giảng viên chấm:

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **Chữ ký** | **Ghi chú** |
| Giảng viên chấm 1: |  |  |
| Giảng viên chấm 2: |  |  |

**MỤC LỤC**

[DANH MỤC HÌNH ẢNH 4](#_Toc164220719)

[LỜI CẢM ƠN 6](#_Toc164220720)

[LỜI NÓI ĐẦU 7](#_Toc164220721)

[CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI 8](#_Toc164220722)

[1.1 Đặt vấn đề 8](#_Toc164220723)

[1.2 Cơ sở hình thành đề tài 8](#_Toc164220724)

[1.3 Mục tiêu đề tài 9](#_Toc164220725)

[CHƯƠNG 2: GIỚI THIỆU VỀ HỌC MÁY VÀ THUẬT TOÁN HỒI QUY TUYẾN TÍNH TRONG HỌC MÁY 10](#_Toc164220726)

[2.1 Giới thiệu về học máy 10](#_Toc164220727)

[2.1.1 Khái niệm về học máy 10](#_Toc164220728)

[2.1.2 Ứng dụng của học máy 10](#_Toc164220729)

[2.2 Thuật toán hồi quy tuyến tính (Linear Regression) trong học máy 11](#_Toc164220730)

[2.2.1 Khái niệm 11](#_Toc164220731)

[2.2.2 Dạng của Linear Regression 11](#_Toc164220732)

[2.2.3 Hàm mất mát 12](#_Toc164220733)

[2.2.4 Tìm nghiệm của mô hình hồi quy tuyến tính 12](#_Toc164220734)

[2.2.5 Mức độ lỗi của mô hình hồi quy tuyến tính 13](#_Toc164220735)

[CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ 15](#_Toc164220736)

[3.1 Bài toán dự đoán điều kiện nhập học với mô hình Regression 15](#_Toc164220737)

[3.1.1 Phát biểu bài toán 15](#_Toc164220738)

[3.1.2 Chuẩn bị dữ liệu 15](#_Toc164220739)

[3.2 Cài đặt code 16](#_Toc164220740)

[3.3 Kết quả 23](#_Toc164220741)

[3.3.1 Biểu đồ kết quả sản lượng sản phẩm của mỗi hãng 23](#_Toc164220742)

[3.3.2 Biểu đồ kết quả về giá laptop theo độ rộng của màn hình 24](#_Toc164220743)

[3.3.3 Biểu đồ kết quả số lượng CPU được sử dụng 25](#_Toc164220744)

[3.3.4 Biểu đồ kết quả về số lượng Ram được sử dụng 26](#_Toc164220745)

[3.3.5 Bản đồ nhiệt kiểm tra độ tương quan giữa các cột 27](#_Toc164220746)

[3.3.6 Biểu đồ phân tán so sánh giữa giá thực tế so với giá từ dữ liệu 28](#_Toc164220747)

[KẾT LUẬN 29](#_Toc164220748)

[DANH MỤC TÀI LIỆU THAM KHẢO 30](#_Toc164220749)

# DANH MỤC HÌNH ẢNH

Hình 3.1.2 Dữ liệu

Hình 3.2.1 ảnh import thư viện

Hình 3.2.2 đọc file

Hình 3.2.3 Hiển thị thông tin

Hình 3.2.4 Ảnh import thư viện seaborn và vẽ biểu đồ phân phối của cột 'Price'

Hình 3.2.5 Vẽ biểu đồ của số lượng dữ liệu trong cột ‘Company’

Hình 3.2.6 Vẽ biểu đồ cột số lượng dữ liệu trong cột ‘Typename’

Hình 3.2.7 vẽ biểu đồ phân tán của dữ liệu trong cột 'Inches' trên trục x và cột 'Price' trên trục y

Hình 3.2.8 Tách chuỗi dữ liệu về kích thước chiều rộng và chiều cao của màn hình

Hình 3.2.9 Tách chuỗi và lưu trữ cpu

Hình 3.2.10 Phân loại CPU

Hình 3.2.11 Vẽ biểu đồ cột của CPU

Hình 3.2.12 Vẽ biểu đồ cột từ dữ liệu của Ram

Hình 3.2.13 Xem số lượng bộ nhớ cho mỗi sản phẩm

Hình 3.2.14 Biểu đồ cho GPU

Hình 3.2.15 Biểu đồ cho OpSys

Hình 3.2.16 Biểu đồ phân phối xác suất của các giá trị trong cột “Weight”

Hình 3.2.17 Biểu đồ phân tán hiển thị mối quan hệ giữa trọng lượng và giá của các sản phẩm

Hình 3.2.18 Bản đồ nhiệt để kiểm tra sự tương quan

Hình 3.2.19 Import một số class và module

Hình 3.2.20 Tạo một pipeline để xử lý dữ liệu

Hình 3.2.21 Tính toán và in ra màn hình kết quả của hai metric là R2 score và MAE

Hình 3.2.22 Biểu đồ phân tán hiển thị mối quan hệ

Hình 3.3.1 Biểu đồ số lượng sản phẩm của mỗi hãng laptop

Hình 3.3.2 Biểu đồ phân tán giá laptop theo độ rộng của màn hình

Hình 3.3.3 số lượng CPU được dùng

Hình 3.3.4 số lượng Ram được sử dụng

Hình 3.3.5 Bản đồ nhiệt kiểm tra độ tương quan giữa các cột

Hình 3.3.6 Biểu đồ phân tán so sánh giữa giá thực tế so với giá từ dữ liệu

**LỜI CẢM ƠN**

Trên thực tế, không có sự thành công nào mà không gắn liền với những sự hỗ trợ, sự giúp đỡ dù ít hay nhiều, dù là trực tiếp hay gián tiếp của người khác. Trong suốt thời gian từ khi bắt đầu học tập ở giảng đường Đại học đã đến nay, em đã nhận được rất nhiều sự quan tâm, giúp đỡ của thầy cô, gia đình và bạn bè.

Với lòng biết ơn sâu sắc nhất, em xin gửi đến thầy cô ở Khoa Công Nghệ Thông Tin- trường Đại Học Điện Lực đã cùng với tri thức và tâm huyết của mình để truyền đạt vốn kiến thức quý báu cho em trong suốt thời gian học tập tại trường. Và đặc biệt, trong kỳ này, em được tiếp cận với môn học rất hữu ích đối với sinh viên ngành Công Nghệ Thông Tin. Đó là môn: *“Nhập môn học máy”.*

Em xin chân thành cảm ơn thầy Vũ Văn Định đã tận tâm hướng dẫn chúng em qua từng buổi học trên lớp cũng như những buổi nói chuyện, thảo luận về môn học. Trong thời gian được học tập và thực hành dưới sự hướng dân của thầy, em không những thu được rất nhiều kiến thức bổ ích, mà còn được truyền sự say mê và thích thú đối với bộ môn học phần. Nếu không có những lời hướng dẫn, dạy bảo của thầy thì em nghĩ báo cáo này rất khó có thể hoàn thành được.

Mặc dù đã rất cố gắng hoàn thiện báo cáo với tất cả sự nỗ lực, tuy nhiên, do bước đầu đi vào thực tế, tìm hiểu và xây dựng báo cáo trong thời gian có hạn, và kiến thức còn hạn chế, nhiều bỡ ngỡ, nên báo cáo “ *Ứng dụng thuật toán Linear regression dự đoán giá cả của laptop thông qua các thông số kỹ thuật của chúng*” chắc chắn sẽ không thể tránh khỏi những thiếu sót. Em rất mong nhận được sự quan tâm, thông cảm và những đóng góp quý báu của các thầy cô và các bạn để báo cáo này được hoàn thiện hơn.

Một lần nữa, em xin chân thành cảm ơn và luôn mong nhận được sự đóng góp của mọi người.

Trân trọng.

**LỜI NÓI ĐẦU**

Cùng với sự phát triển không ngừng về kỹ thuật máy tính và mạng điện tử, công nghệ thông tin cũng có được những công nghệ có đẳng cấp và lần lượt chinh phục hết đỉnh cao này đến đỉnh cao khác. Mạng internet là một trong những sản phẩm có giá trị hết sức lớn lao và ngày càng trở thành một công cụ không thể thiếu, là nền tảng chính cho sự truyển tải, trao đổi thông tin trên toàn cầu.

Giờ đây, mọi việc liên quan đến thông tin trở nên thật dễ dàng cho người sử dụng, chỉ cần có một máy tính kết nối Internet và một dòng dữ liệu truy tìm thì gần như lập tức cả thế giới về vấn đề mà bạn đang quan tâm sẽ hiện ra, có đầy đủ thông tin, hình ảnh và thậm chí đôi lúc có cả những âm thanh nếu bạn cần.

Bằng Internet, chúng ta đã thực hiện được nhiều công việc với tốc độ nhanh hơn và chi phí thấp hơn nhiều so với cách thức truyền thống. Chính điều này, đã thúc đẩy sự khai sinh và khám phá ra các dữ liệu của tất cả mọi thứ ở trên thế giới nâng cao chất lượng cuộc sống con người.

Vì vậy nhóm em chọn đề tài : “*Ứng dụng thuật toán Linear regression dự đoán giá cả của laptop thông qua các thông thông số kỹ thuật của chúng*”. Trong quá trình tìm hiểu và hoàn thiện, đề tài sẽ không thể tránh khỏi những sai sót, khuyết điểm. Vì vậy, nhóm thực hiện chúng em hy vọng nhận được sự đánh giá và đóng góp nhiệt tình từ phía thầy và các bạn để báo cáo của nhóm em được hoàn thiện hơn.

Chúng em xin chân thành cảm ơn!

# CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

* 1. **Đặt vấn đề**

Đề tài này nhìn chung là về việc áp dụng thuật toán Linear Regression để dự đoán giá cả của laptop dựa trên các thông số kỹ thuật của chúng. Đây là một ứng dụng cụ thể của học máy, trong đó mục tiêu là tạo ra một mô hình có thể dự đoán giá của một laptop dựa trên thông số kỹ thuật như CPU, RAM, dung lượng ổ cứng, độ phân giải màn hình, và các yếu tố khác.

Việc áp dụng Linear Regression trong trường hợp này có thể được thực hiện bằng cách thu thập một tập dữ liệu các laptop đã được bán và thông tin kỹ thuật của chúng cùng với giá bán. Sau đó, thuật toán Linear Regression được sử dụng để tạo ra một mô hình có thể dự đoán giá của laptop mới dựa trên thông số kỹ thuật mà nó cung cấp.

Kết quả của đề tài sẽ là một mô hình Linear Regression có thể được sử dụng để dự đoán giá của các laptop mới dựa trên thông số kỹ thuật mà chúng cung cấp.

* 1. **Cơ sở hình thành đề tài**

Ngày nay có rất nhiều loại laptop với các hãng khác nhau được bán trên thị trường. Điều này gây khó khăn cho khách hàng khi lựa chọn. Việc dự đoán giá laptop sẽ giúp khách hàng dễ dàng tìm kiếm hơn.

Phân loại laptop cần xét đến nhiều thông tin như tên thương hiệu, chủng loại, chất liệu, màu sắc, tuổi thọ, độ bền, giá cả,... Dữ liệu liên quan đến vấn đề này thường có dạng bảng với nhiều cột và hàng. Do đó, thuật toán Linear Regression là một phương pháp phù hợp để dự đoán thông qua các thông số. Bài toán đặt ra là áp dụng thuật toán Linear Regression trên dữ liệu về thông số của laptop và đưa ra được các nhóm (cluster) laptop để phục vụ khách hàng dễ dàng lựa chọn hơn.

Tóm lại, sự kết hợp giữa nhu cầu thị trường, tiến bộ trong lĩnh vực học máy và dữ liệu, cùng với tiềm năng ứng dụng thực tiễn đã tạo ra cơ sở lý tưởng để nghiên cứu và phát triển đề tài về dự đoán giá cả của laptop thông qua các thông số kỹ thuật của chúng.

## 1.3 Mục tiêu đề tài

Mục tiêu chính của đề tài này là xây dựng một mô hình dự đoán giá cả của laptop dựa trên các thông số kỹ thuật của chúng, sử dụng thuật toán Linear Regression. Cụ thể, mục tiêu có thể được phân thành các phần như sau:

* Thu thập dữ liệu: Thu thập một tập dữ liệu đủ lớn và đa dạng về các laptop đã được bán, bao gồm thông số kỹ thuật và giá bán.
* Tiền xử lý dữ liệu: Tiền xử lý dữ liệu để loại bỏ nhiễu, xử lý dữ liệu thiếu và chuẩn hóa các biến nếu cần.
* Phân chia dữ liệu: Phân chia tập dữ liệu thành tập huấn luyện và tập kiểm tra để đào tạo và đánh giá mô hình.
* Xây dựng mô hình: Sử dụng thuật toán Linear Regression để xây dựng một mô hình có khả năng dự đoán giá của laptop dựa trên thông số kỹ thuật mà nó cung cấp.
* Đánh giá mô hình: Đánh giá hiệu suất của mô hình bằng cách sử dụng các độ đo như sai số trung bình (mean squared error) hoặc hệ số xác định (coefficient of determination) trên tập kiểm tra.
* Tinh chỉnh mô hình (nếu cần): Nếu cần thiết, thực hiện các phương pháp tinh chỉnh mô hình như tối ưu hóa siêu tham số hoặc thử nghiệm các biến đổi dữ liệu khác nhau để cải thiện hiệu suất của mô hình.
* Triển khai và ứng dụng: Cuối cùng, triển khai mô hình đã xây dựng để có thể được sử dụng trong các ứng dụng thực tế, chẳng hạn như cung cấp dự đoán giá cả cho người dùng hoặc tích hợp vào các hệ thống bán hàng trực tuyến.

Tóm lại, mục tiêu của đề tài là tạo ra một mô hình dự đoán giá cả của laptop dựa trên thông số kỹ thuật, cung cấp một công cụ hữu ích cho cả người mua và người bán trong thị trường laptop.

# CHƯƠNG 2: GIỚI THIỆU VỀ HỌC MÁY VÀ THUẬT TOÁN HỒI QUY TUYẾN TÍNH TRONG HỌC MÁY

## 2.1 Giới thiệu về học máy

### 2.1.1 Khái niệm về học máy

Học máy (Machine learning) là một lĩnh vực con của Trí tuệ nhân tạo (Artificial Intelligence) sử dụng các thuật toán cho phép máy tính có thể học từ dữ liệu để thực hiện các công việc thay vì được lập trình một cách rõ ràng, cung cấp cho hệ thống khả năng tự động học hỏi và cải thiện hiệu suất, độ chính xác dựa trên những kinh nghiệm từ dữ liệu đầu vào. Học máy tập trung vào việc phát triển các phần mềm, chương trình máy tính có thể truy cập vào dữ liệu và tận dụng nguồn dữ liệu đó để tự học.

Học máy vẫn đòi hỏi sự đánh giá của con người trong việc tìm hiểu dữ liệu cơ sở và lựa chọn các kĩ thuật phù hợp để phân tích dữ liệu. Đồng thời, trước khi sử dụng, dữ liệu phải sạch, không có sai lệch và không có dữ liệu giả.

Các mô hình học máy yêu cầu lượng dữ liệu đủ lớn để "huấn luyện" và đánh giá mô hình. Trước đây, các thuật toán học máy thiếu quyền truy cập vào một lượng lớn dữ liệu cần thiết để mô hình hóa các mối quan hệ giữa các dữ liệu. Sự tăng trưởng trong dữ liệu lớn (big data) đã cung cấp các thuật toán học máy với đủ dữ liệu để cải thiện độ chính xác của mô hình và dự đoán.

### 2.1.2 Ứng dụng của học máy

Nhiều hoạt động hàng ngày của chúng ta được trợ giúp bởi các thuật toán machine learning, bao gồm:

* Trong y tế: xác định bệnh lý của người bệnh mới dựa trên dữ liệu lịch sử của các bệnh nhân có cùng bệnh lý có cùng các đặc điểm đã được chữa khỏi trước đây, hay xác định loại thuốc phù hợp
* Trong lĩnh vực ngân hàng: xác định khả năng khách hàng chậm trả các khoản vay hoặc rủi ro tín dụng do nợ xấu dựa trên phân tích Credit score; xác định xem liệu các giao dịch có hành vi phạm tội, lừa đảo hay không.
* Trong giáo dục: phân loại các học sinh theo hoàn cảnh, học lực để xem xem cần hỗ trợ gì cho những học sinh ví dụ như hoàn cảnh sống khó khăn nhưng học lực lại tốt.
* Trong thương mại điện tử: phân loại khách hàng theo sở thích cụ thể để hỗ trợ personalized marketing hay xây dựng hệ thống khuyến nghị, dựa trên dữ liệu từ website, social media.

## 2.2 Thuật toán hồi quy tuyến tính (Linear Regression) trong học máy

### 2.2.1 Khái niệm

Hồi quy tuyến tính la một phương pháp thống kê để hồi quy dữ liệu với biến phụ thuộc có giá trị liên tục trong khi các biến độc lập có thể có một trong hai giá trị liên tục hoặc là giá trị phân loại. Hồi quy tuyến tính là một trong hai dạng lớn của học có giám sát (supervised learning) dựa trên tập dữ liệu mẫu.

Nói cách khác "Hồi quy tuyến tính" là một phương pháp để dự đoán biến phụ thuộc (Y) dựa trên giá trị của biến độc lập (X). Nó có thể được sử dụng cho các trường hợp chúng ta muốn dự đoán một số lượng liên tục. Ví dụ, dự đoán giao thông ở một cửa hàng bán lẻ, dự đoán thời gian người dùng dừng lại một trang nào đó hoặc số trang đã truy cập vào một website nào đó v.v...

### 2.2.2 Dạng của Linear Regression

Hồi quy tuyến tính có phương trình đạng :

F(x) = w0 + w1x1 + w2x2 + … + wnxn (1)

Trong đó, w1, w2, wn, w0 là các hằng số, w0 còn được gọi là bias hay sai số.

Mối quan hệ giữa y f(x) bên trên là một mối quan hệ tuyến tính (linear). Bài toán chúng ta đang làm là một bài toán thuộc loại regression. Bài toán đi tìm các

hệ số tối ưu { w1, w2, wn, w0 } chính vì vậy được gọi là bài toán Linear Regression (Hồi quy tuyến tính).

Trong phương trình (1) nếu chúng ta đặt **w = [**w0, w1, w2, wn**]**T là một vecter (cột) hệ số cần phải tối ưu và = [1, x1, x2, xn] (đọc là x bar trong tiếng Anh) là vector (hàng) dữ liệu đầu vào mở rộng. Số 1 ở đầu được thêm vào để phép tính đơn giản hơn và thuận tiện cho việc tính toán. Khi đó, phương trình (1) có thể được viết lại dưới dạng: y ( trong đó là một vecter hàng).

### 2.2.3 Hàm mất mát

Máy học từ giá trị trung bình của một hàm mất mát. Đây là một phương pháp đánh giá độ hiệu quả của một thuật toán nào đó trên bộ dữ liệu cho trước. Nếu kết quả dự đoán chênh lệch quá nhiều so với kết quả thực tế, hàm mất mát sẽ là một số rất lớn. Điều tương tự xảy ra với tất cả các cặp (xi, yi), i = 1, 2, 3, …, N với N là số lượng dữ liệu quan sát được. Để hàm mất mát nhỏ nhất khi đó tổng sai số là nhỏ nhất tương đương với việc tìm **w** để hàm số sau đạt giá trị nhỏ nhất:

J(w) = (2)

Hàm số J(w) được gọi là hàm mất mát (loss function) của bài toán Linear Regression. Chúng ta luôn mong muốn rằng sự mất mát (sai số) là nhỏ nhất, điều đó đồng nghĩa với việc tìm vector hệ số **w** sao cho giá trị của hàm mất mát này càng nhỏ càng tốt.

Trước khi đi tìm lời giải, chúng ta đơn giản hóa phép toán trong phương trình hàm mất mát (2). Đặt là một vector cột chứa tất cả các output của training data; = là ma trận dữ liệu đầu vào (mở rộng) mà mỗi hàng của nó là một điểm dữ liệu.

### 2.2.4 Tìm nghiệm của mô hình hồi quy tuyến tính

Để tìm nghiệm cho một bài toán tối ưu chúng ta thường giải phương trình đạo hàm J(w) = bằng 0.

Đạo hàm theo **w** của hàm mất mát là:



Trên thực tế A có thể không khả nghịch nên ta sẽ dùng ma trận giả nghịch đảo nên ta có W = hay W = Đây chính là nghiệm tổng quát của hồi quy tuyến tính.

Để tìm nghiệm của mô hình hồi quy tuyến tính, ta cần thực hiện các bước sau:

* Xây dựng ma trận đặc trưng X: Ma trận này sẽ bao gồm các thông số kỹ thuật của các laptop, mỗi hàng là một mẫu dữ liệu và mỗi cột là một đặc trưng (ví dụ: CPU, RAM, dung lượng ổ cứng, độ phân giải màn hình, ...). Đồng thời, ta cần xây dựng một vector cột y chứa các giá cả tương ứng với từng mẫu dữ liệu.
* Thêm cột 1 vào ma trận X (nếu cần): Để tính toán được hệ số điều chỉnh tự do (intercept), ta cần thêm một cột toàn giá trị 1 vào ma trận X.
* Tìm nghiệm w của mô hình: Sử dụng phương pháp tối thiểu hóa hàm mất mát, chẳng hạn như phương pháp Least Squares, để tìm ra vector tham số w của mô hình.

Cụ thể, nghiệm của mô hình được tính theo công thức:

Trong đó:

* *X* là ma trận đặc trưng.
* *y* là vector cột chứa giá cả của các mẫu dữ liệu.
* là ma trận chuyển vị của *X*.
* là ma trận nghịch đảo của tích của với *X*.
* Sử dụng nghiệm để dự đoán: Sau khi tìm được vector tham số w, ta có thể sử dụng nó để dự đoán giá cả của các laptop mới dựa trên các thông số kỹ thuật của chúng.
  + 1. **Mức độ lỗi của mô hình hồi quy tuyến tính**

Mức độ lỗi của một mô hình hồi quy tuyến tính thường được đo bằng các độ đo sai số, phổ biến nhất là sai số trung bình (mean squared error - MSE) và hệ số xác định (coefficient of determination - ). Dưới đây là cách tính và ý nghĩa của từng độ đo:

* Sai số trung bình (MSE): MSE là trung bình của bình phương của sai số giữa giá trị dự đoán và giá trị thực tế. Công thức tính MSE như sau:

Trong đó:

* *n* là số lượng mẫu dữ liệu.
* ​ là giá trị thực tế của mẫu dữ liệu thứ *i*.
* là giá trị dự đoán của mô hình cho mẫu dữ liệu thứ *i*.

MSE càng nhỏ, mô hình càng tốt vì nó cho biết mức độ chính xác của các dự đoán so với giá trị thực tế.

* Hệ số xác định (): đo lường mức độ biến thiên của mục tiêu mà mô hình có thể giải thích được. Nó là tỷ lệ giữa phương sai của dự đoán và phương sai của giá trị thực tế. Công thức tính như sau:

Trong đó:

* ​ là tổng bình phương của các sai số dự đoán.
* là tổng bình phương của sự biến thiên của giá trị thực tế.

càng gần 1, mô hình càng tốt vì nó cho biết phần lớn sự biến thiên của mục tiêu được mô hình hóa bởi các đặc trưng.

Việc kiểm tra và so sánh MSE và giữa mô hình trên tập huấn luyện và tập kiểm tra giúp đánh giá hiệu suất của mô hình và đảm bảo rằng nó không bị overfitting (quá mức).

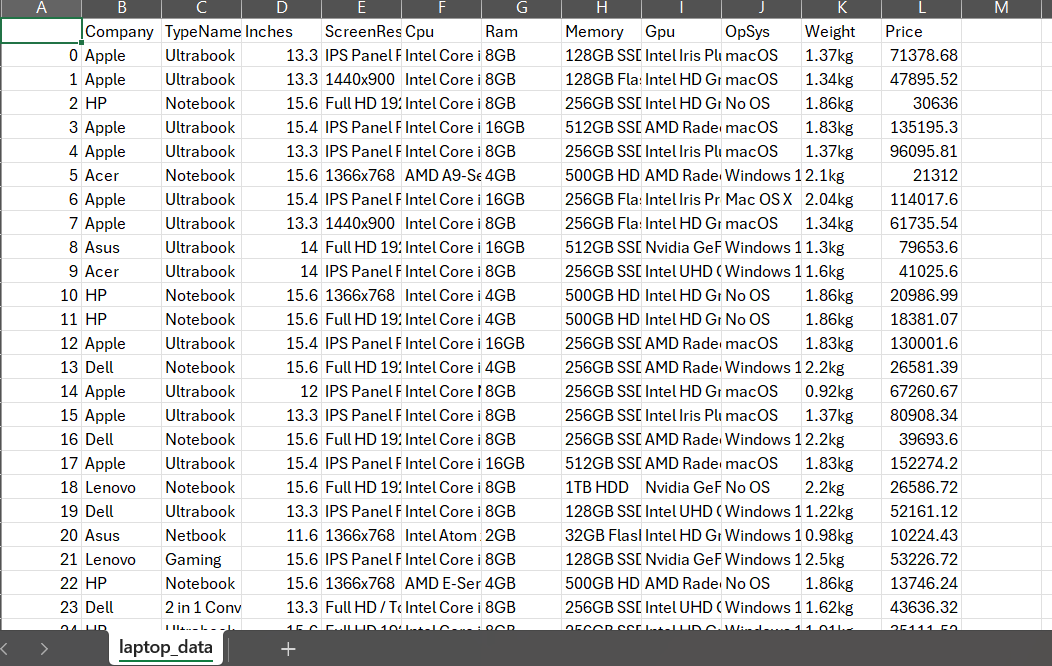
# CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

## 3.1 Bài toán dự đoán điều kiện nhập học với mô hình Regression

### 3.1.1 Phát biểu bài toán

* Bài toán đưa ra tất cả thông tin cần thiết để đưa ra giá hợp lý thông qua các thông số của các laptop để từ đó căn cứ dự đoán giá ở thời điểm hiện tại đến thời điểm sau đó.
* Bài toán sẽ lấy dữ liệu trên Kaggle để phân tích, huấn luyện để dự đoán điều kiện mua hàng với giá cả hợp lý.

### 3.1.2 Chuẩn bị dữ liệu



*Hình 3.1.2 Dữ liệu*

## 3.2 Cài đặt code

Import thư viện:

**A screen shot of a computer program

Description automatically generated**

*Hình 3.2.1 ảnh import thư viện*

Đọc file và load dữ liệu từ file laptop\_data.csv:

*A screenshot of a computer

Description automatically generated*

*Hình 3.2.2 đọc file*

Hiển thị thông tin mô tả về DataFrame, bao gồm: tên, số lượng dòng, cột,…

*A screenshot of a computer program

Description automatically generated*

*Hình 3.2.3 Hiển thị thông tin*

Import thư viện seaborn và vẽ biểu đồ phân phối của cột 'Price' :

**A screenshot of a computer

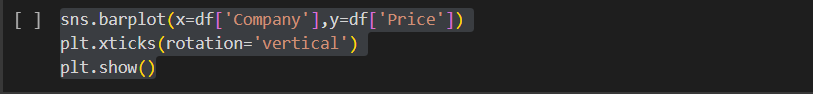
Description automatically generated**

*Hình 3.2.4 Ảnh import thư viện seaborn và vẽ biểu đồ phân phối của cột 'Price'*

Vẽ biểu đồ cột của số lượng dữ liệu trong cột 'Company':

**A black rectangular sign with white text

Description automatically generated**

******

*Hình 3.2.5 Vẽ biểu đồ của số lượng dữ liệu trong cột ‘Company’*

Vẽ biểu đồ cột của số lượng dữ liệu trong cột 'Typename':

*A black and white text

Description automatically generated*

*A black background with colorful text

Description automatically generated*

*Hình 3.2.6 Vẽ biểu đồ cột số lượng dữ liệu trong cột ‘Typename’*

Vẽ biểu đồ phân tán dữ liệu trong cột 'Inches' trên trục x và cột 'Price' trên trục y:



*Hình 3.2.7 vẽ biểu đồ phân tán của dữ liệu trong cột 'Inches' trên trục x và cột 'Price' trên trục y*

Tách chuỗi dư liệu về kích thước chiều rộng và chiều cao của màn hình:

A screenshot of a computer

Description automatically generated

*Hình 3.2.8 Tách chuỗi dữ liệu về kích thước chiều rộng và chiều cao của màn hình*

Tách chuỗi và lưu trữ cpu:

A screenshot of a computer

Description automatically generated *Hình 3.2.9 Tách chuỗi và lưu trữ cpu*

Phân loại cpu:

A screenshot of a computer

Description automatically generated

*Hình 3.2.10 Phân loại CPU*

Vẽ biểu đồ cột từ dữ liệu của CPU:



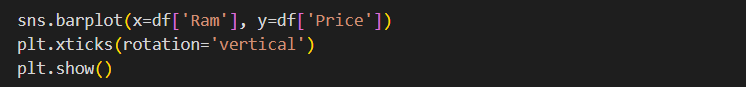
A computer screen shot of a black background

Description automatically generated

*Hình 3.2.11 Vẽ biểu đồ cột của CPU*

Vẽ biểu đồ cột từ dữ liệu của Ram:



*Hình 3.2.12 Vẽ biểu đồ cột từ dữ liệu của Ram*

Xem số lượng bộ nhớ cho mỗi sản phẩm:

A black background with white text

Description automatically generated

A computer screen with colorful lights

Description automatically generated

*Hình 3.2.13 Xem số lượng bộ nhớ cho mỗi sản phẩm*

Vẽ biểu đồ cho GPU:

A black background with white text

Description automatically generated

A screenshot of a computer

Description automatically generated

*Hình 3.2.14 Biểu đồ cho GPU*

Vẽ biểu đồ cho OpSys:

A black rectangular object with a black border

Description automatically generated

*Hình 3.2.15 Biểu đồ cho OpSys*

Biểu đồ này sẽ hiển thị phân phối xác suất của các giá trị trong cột 'Weight':



*Hình 3.2.16 Biểu đồ phân phối xác suất của các giá trị trong cột “Weight”*

Biểu đồ phân tán hiển thị mối quan hệ giữa trọng lượng và giá của các sản phẩm:



*Hình 3.2.17 Biểu đồ phân tán hiển thị mối quan hệ giữa trọng lượng và giá của các sản phẩm*

Chúng ta sẽ sử dụng bản đồ nhiệt để kiểm tra độ tương quan giữa các cột:

A screen shot of a computer code

Description automatically generated

*Hình 3.2.18 Bản đồ nhiệt để kiểm tra sự tương quan*

Import một số class và module từ các thư viện sklearn và XGBoost để thực hiện các công việc liên quan đến huấn luyện và đánh giá mô hình hồi quy:

A computer screen with white text

Description automatically generated

*Hình 3.2.19 Import một số class và module*

Tạo một pipeline để xử lý dữ liệu và huấn luyện một mô hình hồi quy tuyến tính

A screen shot of a computer

Description automatically generated

*Hình 3.2.20 Tạo một pipeline để xử lý dữ liệu*

Tính toán và in ra màn hình kết quả của hai metric là R2 score và MAE:

A screen shot of a computer

Description automatically generated

*Hình 3.2.21 Tính toán và in ra màn hình kết quả của hai metric là R2 score và MAE*

Biểu đồ phân tán hiển thị mối quan hệ giữa giá trị thực tế (trục x) và giá trị dự đoán (trục y) của mô hình trên tập dữ liệu kiểm tra:

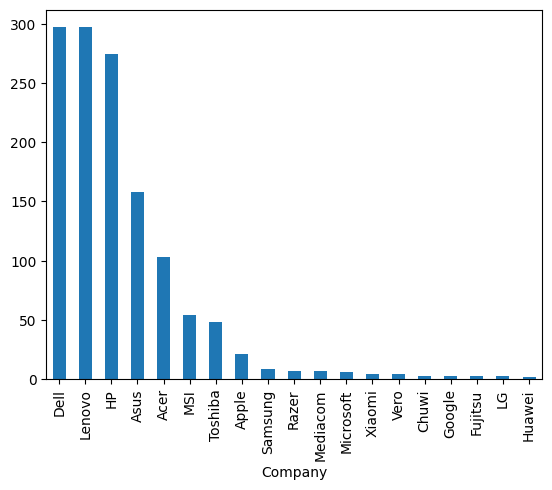
A black screen with a black background

Description automatically generated

*Hình 3.2.22 Biểu đồ phân tán hiển thị mối quan hệ*

## 3.3 Kết quả

### 3.3.1 Biểu đồ kết quả sản lượng sản phẩm của mỗi hãng



*Hình 3.3.1 Biểu đồ số lượng sản phẩm của mỗi hãng laptop*

* Số lượng hãng được khách dùng nhiều nhất lần lượt giảm dần là: Dell, Lenovo, HP, Asus, Acer, Msi, Toshiba, Apple, SamSung, Razer.
* Hãng dung ít nhất như là LG, Huawei.

### 3.3.2 Biểu đồ kết quả về giá laptop theo độ rộng của màn hình

A graph with blue dots

Description automatically generated

*Hình 3.3.2 Biểu đồ phân tán giá laptop theo độ rộng của màn hình*

Theo hình trên, ta thấy được:

* Giá cao nhất thuộc về màn hình rộng 18 inches.
* Giá thấp nhất thuộc về màn hình 10 inches.

### 3.3.3 Biểu đồ kết quả số lượng CPU được sử dụng

A graph of different types of computer components

Description automatically generated

*Hình 3.3.3 số lượng CPU được dùng*

* CPU được dùng nhiều nhất là inter Core i7.
* CPU được dùng ít nhất là AMD processor.

### 3.3.4 Biểu đồ kết quả về số lượng Ram được sử dụng

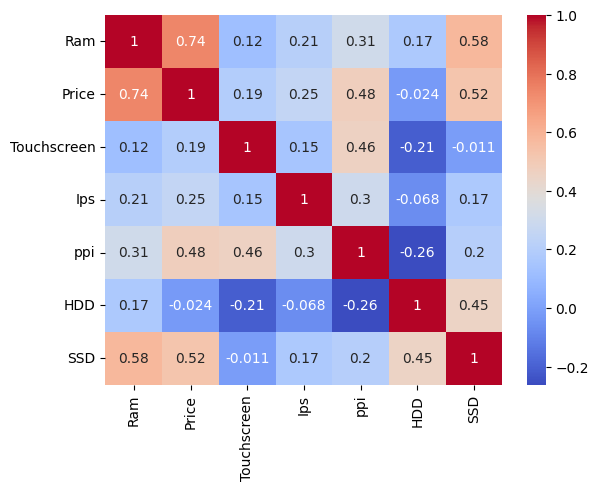
A graph with blue bars

Description automatically generated

*Hình 3.3.4 số lượng Ram được sử dụng*

* Ram được sử dụng nhiều nhất là 8GB
* Ram được sử dụng ít nhất là 24GB và 64GB

### 3.3.5 Bản đồ nhiệt kiểm tra độ tương quan giữa các cột



*Hình 3.3.5 Bản đồ nhiệt kiểm tra độ tương quan giữa các cột*

Qua đó, ta phân tích được các cột có giá trị tương quan như như thế nào với nhau. Về cơ bản, cột giá (price) có chút tương quan với các cột còn lại nhiều nhất, chứng tỏ các yếu tố đó có tác động ít nhiều lên giá laptop.

### 3.3.6 Biểu đồ phân tán so sánh giữa giá thực tế so với giá từ dữ liệu

A graph with blue dots

Description automatically generated

*Hình 3.3.6 Biểu đồ phân tán so sánh giữa giá thực tế so với giá từ dữ liệu*

# KẾT LUẬN

Qua việc thực hiện nghiên cứu đề tài “*Ứng dụng thuật toán Linear regression dự đoán giá cả của laptop thông qua các thông số kỹ thuật của chúng*”, nhóm chúng em đã được biết thêm rất nhiều kiến thức vềthuật toán cũng như bước đầu nắm bắt được ứng dụng của nhập môn học máy.

Kết quả đạt được: chúng em đã cài đặt được thuật toán và sử dụng dụng thư viện scikit-learn trong quá trình học tập. Nhưng bên cạnh đó thuật toán vẫn còn những ưu nhược điểm như:

* Ưu điểm: Nhanh chóng để mô hình hóa và đặc biệt hữu ích khi mối quan hệ được mô hình hóa không quá phức tạp và nếu bạn không có nhiều dữ liệu. Hồi quy tuyến tính là đơn giản để hiểu, nó rất có giá trị cho các quyết định kinh doanh.
* Nhược điểm: Đối với dữ liệu phi tuyến tính, hồi quy đa thức có thể khá khó khăn để thiết kế, vì người ta phải có một số thông tin về cấu trúc của dữ liệu và mối quan hệ giữa các biến tính năng.

# DANH MỤC TÀI LIỆU THAM KHẢO

[1] George A. F. Seber và Alan J. Lee, "Linear Regression Analysis", Wiley, 2012.

[2] Sanford Weisberg, "Applied Linear Regression", Wiley, 2014.

[3] Kevin P. Murphy, "Machine Learning: A Probabilistic Perspective", MIT Press, 2012.

[4] Christopher M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.

[5] Ethem Alpaydin, "Introduction to Machine Learning", MIT Press, 2009.

[6] Aurélien Géron, "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow", O'Reilly Media, 2019.

[7] Tom Mitchell, "Machine Learning"*,* McGraw-Hill, 1997.

[8] <https://tailieuhust.com/nhap-mon-hoc-may-va-khai-pha-du-lieu/>