TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP

**KHOA ĐIỆN TỬ**

**Bộ môn: Tin Học Công Nghiệp**

****

**BÀI TẬP LỚN**

**MÔN HỌC**

**KHOA HỌC DỮ LIỆU**

Sinh Viên: Đặng Văn Tuyên

Lớp: K55KMT

Giáo viên hướng dẫn: TS.Nguyễn Văn Huy

**Thái Nguyên-2023**

**TRƯỜNG ĐHKTCN CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**

**KHOA ĐIỆN TỬ *Độc lập –Tự Do-Hạnh phúc***

**BÀI TẬP LỚN**

**MÔN HỌC: KHOA HỌC DỮ LIỆU**

BỘ MÔN:TIN HỌC CÔNG NGHIỆP

*Sinh viên:Đặng Văn Tuyên.*

*Lớp:K55KMT. Ngành:Kỹ Thuật Máy Tính.*

*Giáo viên hướng dẫn: TS.Nguyễn Văn Huy*

*Ngày giao đề: 05/05/2023. Ngày hoàn thành:31/5/2023*

*Tên đề tài:Thiết kế mô hình dự đoán giá laptop.*

*Yêu cầu: ……………………………………………………………………………………………*

*………………………………………………………………………………………………………*

*………………………………………………………………………………………………………*

*………………………………………………………………………………………………………*

*………………………………………………………………………………………………………*

*………………………………………………………………………………………………………*

*………………………………………………………………………………………………………*

**GIÁO VIÊN HƯỚNG DẪN**

*(Ký và ghi rõ họ tên)*

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

*……………………………………………………………………………………………..………………………………………………………………………………………………..………………………………………………………………………………………………..………………………………………………………………………………………………………………………..*

*……………………………………………………………………………………………..………………………………………………………………………………………………..…………………………………………………………………………………………………..……………………………………………………………………………………………………………………..*

*Thái nguyên,ngày….tháng….năm 20….*

**GIÁO VIÊN HƯỚNG DẪN**

**(***Ký và ghi rõ họ tên)*

**MỤC LỤC**

[CHƯƠNG I: GIỚI THIỆU 6](#_Toc136327441)

[**1.1. Phân tích dự đoán là gì?** 6](#_Toc136327442)

[**1.2. Phân tích dự đoán hoạt động như thế nào?** 6](#_Toc136327443)

[**1.3. Ứng dụng của dự đoán giá** 6](#_Toc136327444)

[**1.4. Các công nghệ trong dự đoán giá** 7](#_Toc136327445)

[**1.5. Mô hình hồi quy được áp dụng** 9](#_Toc136327446)

[CHƯƠNG II: CƠ SỞ LÝ THUYẾT 11](#_Toc136327447)

[**2.1. Xử lý dữ liệu** 11](#_Toc136327448)

[*2.1.1. Xử lý dữ liệu là gì?* 11](#_Toc136327449)

[*2.1.2. Dữ liệu được xử lý như thế nào?* 11](#_Toc136327450)

[*2.1.3. Xử lý dữ liệu NAN* 13](#_Toc136327451)

[*2.1.4. Lọc nhiễu* 14](#_Toc136327452)

[**2.1. Model** 15](#_Toc136327453)

[*2.1.1. Model là gì?* 15](#_Toc136327454)

[*2.1.2. Linear Regression (Hồi Quy Tuyến Tính)* 15](#_Toc136327455)

[CHƯƠNG III: XÂY DỰNG CHƯƠNG TRÌNH 17](#_Toc136327456)

[**3.1. Code** 17](#_Toc136327457)

[**3.2. Thuật toán** 20](#_Toc136327458)

[CHƯƠNG IV: KẾT LUẬN 22](#_Toc136327459)

[**4.1. Chương trình đã chạy được** 22](#_Toc136327460)

[**4.2. Mặt tốt và mặt xấu của chương trình** 22](#_Toc136327461)

[*4.2.1. Mặt tốt* 22](#_Toc136327462)

[*4.2.2. Mặt xấu* 23](#_Toc136327463)

[**4.3. Ứng dụng phát triển về sau của chương trình** 24](#_Toc136327464)

# CHƯƠNG I: GIỚI THIỆU

## **1.1. Phân tích dự đoán là gì?**

Phân tích dự đoán là quá trình nghiên cứu dữ liệu trước đây và hiện tại để đưa ra dự đoán trong tương lai. Quá trình này sử dụng một tập hợp bao gồm nhiều kỹ thuật toán học, thống kê và máy học nâng cao để phân tích dữ liệu nhằm xác định và ngoại suy các xu hướng ẩn.

Nhiều doanh nghiệp và tổ chức sử dụng phân tích dự đoán để định hướng các quyết định trong tương lai. Ví dụ: các nhà phân tích tiếp thị sử dụng phân tích dự đoán để xác định doanh số bán hàng trong tương lai cho sản phẩm của họ; các trạm khí tượng sử dụng phân tích dự đoán để dự báo thời tiết và các nhà môi giới chứng khoán sử dụng phương pháp này để tối đa hóa lợi nhuận giao dịch.

## **1.2. Phân tích dự đoán hoạt động như thế nào?**

Ngày nay, phân tích dự đoán chủ yếu dựa trên các kỹ thuật máy học tiên tiến. Các nhà khoa học dữ liệu sử dụng công nghệ học sâu và các thuật toán phức tạp để phân tích nhiều biến số nhằm tạo ra những mô hình dự đoán có khả năng dự báo hành vi có thể xảy ra từ dữ liệu lớn.

Tương tự như các ứng dụng máy học khác, phân tích dự đoán là một hoạt động linh hoạt, liên tục sử dụng dữ liệu mới để cập nhật các dự đoán. Tức là kỹ thuật này sử dụng quy trình công cụ máy cổ điển để làm sạch dữ liệu, đào tạo mô hình, triển khai, phản hồi, đào tạo lại và triển khai lại cùng với khả năng tải nhập dữ liệu gần theo thời gian thực. Các kỹ thuật bao gồm cây quyết định, phân tích hồi quy, phân tích chuỗi thời gian và mạng nơ-ron học sâu.

Các mô hình phân tích dự đoán định hướng quyết định kinh doanh trong tương lai thường phức tạp và liên quan đến nhiều yếu tố. Các mô hình này thường tiêu tốn nhiều thời gian để phát triển cũng như xác thực và cần được cập nhật lại liên tục để thích ứng với những thay đổi trong môi trường kinh doanh và kinh tế.

## **1.3. Ứng dụng của dự đoán giá**

Đoán giá có thể được áp dụng trong nhiều lĩnh vực khác nhau và có nhiều ứng dụng thực tế như:

* Thị trường tài chính: Trong lĩnh vực này, đoán giá được sử dụng để dự đoán giá cổ phiếu, tỷ giá hối đoái, giá vàng và các tài sản tài chính khác. Điều này giúp các nhà đầu tư, ngân hàng và các tổ chức tài chính khác đưa ra quyết định đầu tư và quản lý rủi ro.
* Bất động sản: Đoán giá có thể được sử dụng để ước tính giá trị của các tài sản bất động sản như căn nhà, đất đai hoặc các dự án bất động sản. Điều này hỗ trợ các nhà đầu tư, nhà môi giới và các công ty định giá bất động sản trong việc đưa ra quyết định mua, bán hoặc đầu tư vào bất động sản.
* Tiếp thị và dự báo doanh số: Đoán giá có thể giúp các doanh nghiệp dự đoán doanh số và nhu cầu của sản phẩm hoặc dịch vụ trong tương lai. Điều này hỗ trợ các chiến lược tiếp thị, quản lý sản xuất và quyết định về giá cả.
* Dự báo trong lĩnh vực khoa học và công nghệ: Đoán giá có thể được sử dụng để dự đoán các biến số khoa học và công nghệ, ví dụ như dự báo thời tiết, dự đoán khí hậu, dự báo lưu lượng truy cập trang web hoặc dự đoán sự phát triển của công nghệ trong tương lai.
* Quản lý chuỗi cung ứng: Đoán giá có thể hỗ trợ quản lý chuỗi cung ứng bằng cách dự đoán giá cả và nhu cầu của các nguyên liệu, thành phẩm và dịch vụ trong chuỗi cung ứng. Điều này giúp các doanh nghiệp tối ưu hóa quy trình sản xuất, lưu kho và phân phối.
* Dự báo chi phí: Đoán giá có thể được sử dụng để dự báo chi phí trong các dự án xây dựng, sản xuất và dịch vụ. Việc có dự báo chính xác giúp các doanh nghiệp lập kế hoạch ngân sách và quản lý tài chính một cách hiệu quả.
* Dự báo trong y tế: Đoán giá có thể được áp dụng trong lĩnh vực y tế để dự đoán mức độ lây nhiễm bệnh, nhu cầu dịch vụ y tế và xu hướng dịch tễ học. Các dự báo này có thể hỗ trợ quyết định chính sách y tế, quản lý tài nguyên y tế và phòng ngừa bệnh.

## **1.4. Các công nghệ trong dự đoán giá**

Công nghệ đóng một vai trò quan trọng trong phát triển và áp dụng các phương pháp dự đoán giá hiện đại. Dưới đây là một số công nghệ phổ biến được sử dụng trong dự đoán giá:

* Học máy và Trí tuệ nhân tạo (AI): Học máy và AI đóng vai trò quan trọng trong việc phân tích dữ liệu và tạo ra mô hình dự đoán giá. Các thuật toán học máy như học sâu (deep learning) và mạng nơ-ron nhân tạo (artificial neural networks) có khả năng học từ dữ liệu lịch sử và tìm ra các mẫu và quy luật ẩn trong dữ liệu, từ đó dự đoán giá trong tương lai.
* Mô hình hồi quy (Regression Models): Mô hình hồi quy có thể sử dụng để dự đoán giá dựa trên các biến độc lập (như thông tin lịch sử, chỉ số kinh tế, và yếu tố thị trường). Các phương pháp như hồi quy tuyến tính, hồi quy logistic, và hồi quy phi tuyến (nonlinear regression) có thể được áp dụng để xây dựng mô hình dự đoán giá.
* Mạng nơ-ron nhân tạo (Artificial Neural Networks - ANN): ANN có khả năng học từ dữ liệu và tìm ra các mẫu phức tạp trong dữ liệu giúp dự đoán giá. Các mạng nơ-ron nhân tạo sâu (deep neural networks) như mạng nơ-ron tích chập (convolutional neural networks) và mạng nơ-ron hồi quy (recurrent neural networks) có thể được sử dụng để xây dựng mô hình dự đoán giá mạnh mẽ.
* Học tăng cường (Reinforcement Learning): Học tăng cường có thể được sử dụng để xây dựng các mô hình dự đoán giá trong các môi trường động và không chắc chắn. Các thuật toán học tăng cường như Q-learning và Deep Q-networks (DQN) có thể được áp dụng để tối ưu hóa việc đưa ra quyết định và dự đoán giá.
* Khai phá dữ liệu (Data mining): Công nghệ khai phá dữ liệu giúp phân tích và khám phá thông tin quan trọng từ các nguồn dữ liệu lớn. Các phương pháp như phân tích đa biến, phân cụm và phân tích chuỗi thời gian có thể áp dụng để tìm ra xu hướng và mẫu trong dữ liệu giúp dự đoán giá.
* Tư duy phân tán (Distributed computing): Đối với dự đoán giá phức tạp và việc xử lý các tập dữ liệu lớn, công nghệ tính toán phân tán giúp tăng tốc quá trình tính toán. Sử dụng hệ thống tính toán phân tán, các mô hình dự đoán giá có thể được xây dựng và triển khai một cách hiệu quả.
* Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP): Với dự đoán giá trong lĩnh vực tin tức và thông tin thị trường, công nghệ xử lý ngôn ngữ tự nhiên có thể được sử dụng để phân tích và hiểu nghĩa từ các nguồn tin tức và bài viết liên quan đến giá cả. Điều này giúp cung cấp thông tin quan trọng để dự đoán giá.

Tất cả những công nghệ trên đóng vai trò quan trọng trong việc xây dựng mô hình dự đoán giá và cải thiện độ chính xác của dự đoán.

## **1.5. Mô hình hồi quy được áp dụng**

Bản chất của phân tích hồi quy chính là việc ước lượng mối quan hệ giữa các tham số (variables) trong mô hình mà chúng ta đang xét. Quay về trên quan điểm của xác suất thống kê, mình vẫn nhớ quan điểm của một người thầy rằng **“Không có một sự kiện nào trên đời là ngẫu nhiên, những thứ đang cho là ngẫu nhiên chỉ là những sự kiện ta chưa tìm ra được mô hình để biểu diễn quy luật của chúng”**. Hãy tạm coi như phát biểu của ông thầy kia là đúng và tưởng tượng chúng ta đang là những người đánh xổ số. Giả sử rằng chúng ta có thể tìm ra được mối liên hệ giữa kết quả xổ số của một ngày với một vài tham số **dễ thương** như sau:

* Biến **X1** biểu diễn số trẻ em sinh ra trong ngày hôm nay
* Biến **X2** biểu diễn số người chết đi trong ngày hôm nay
* Biến **X3** biểu diễn lượng mưa trong ngày hôm nay
* Biến **X4** biểu diễn số thằng bị thất tình trong ngày hôm nay...

Và giả sử rằng kết quả xổ số của ngày hôm nay phụ thuộc một cách nào đó vào số người mới sinh, số người chết, lượng mưa và số thằng thất tình… Tức là chúng ta có thể viết dưới dạng toán học như sau:

**Kết quả xổ số = f(X1, X2, X3, X4, …)**

Hàm **f** của chúng ta là gì??? Đó là công việc của hàm hồi quy, việc của chúng ta đó là hãy sưu tập cho máy tính một tập dữ liệu gốc thật chuẩn (**training dataset**). Sau khi tìm ra được các tham số phụ thuộc ví dụ như:

**Kết quả xổ số = 0.6X1 + 3X2 – 9X3 + 0.25X4 +…)**

Chúng ta sẽ sử dụng để so sánh nó trên tập dữ liệu kiểm tra (**testing dataset**). Thay đổi các phương pháp hồi quy cho đến khi sai số trên tập dữ liệu kiểm tra là nhỏ nhất. Như vậy chúng ta đã có một mô hình hồi quy sử dụng để tính toán kết quả xổ sổ rồi. Việc còn lại là tìm các giá trị cho **X1**, **X2**, **X3**…. áp vào mô hình rồi lấy kết quả đi đánh xổ số và tự tin với xác suất thắng lợi giống y như xác suất cao nhất bạn đạt trên tập dữ liệu kiểm tra. Qúa thú vị phải không nào. Hi vọng qua ví dụ hơi **hoang tưởng** trên bạn có thể hình dung ra được việc sử dụng phương pháp hồi quy trong bài toán dự đoán là như thế nào. Tất nhiên, trên thực tế người ta đã sử dụng mô hình hồi quy để ứng dụng vào các bài toán như:

* Dự đoán giá cả của sản phẩm
* Dự đoán biến động chứng khoán
* Dự đoán thời tiết...

# CHƯƠNG II: CƠ SỞ LÝ THUYẾT

## **2.1. Xử lý dữ liệu**

### *2.1.1. Xử lý dữ liệu là gì?*



Xử lý dữ liệu là việc chuyển đổi dữ liệu thành dạng có thể sử dụng được. Trước đây, việc xử lý được thực hiện thủ công, tốn thời gian và dễ sai sót, thì hiện nay hầu hết việc xử lý được thực hiện tự động bằng máy tính sẽ cho kết quả nhanh và chính xác. Trong bài viết này, UniTrain sẽ hướng dẫn bạn cách dữ liệu được xử lý.

### *2.1.2. Dữ liệu được xử lý như thế nào?*

Xử lý dữ liệu bắt đầu với việc thu thập dữ liệu. Dữ liệu được thu thập để chuyển đổi dạng mong muốn phải được xử lý bằng cách xử lý dữ liệu theo từng bước như dữ liệu thu thập được phải được lưu trữ, sắp xếp, xử lý, phân tích và trình bày.

Vì vậy, điều này được chia thành 6 bước cơ bản như sau: thu thập dữ liệu, lưu trữ dữ liệu, sắp xếp dữ liệu, xử lý dữ liệu. phân tích dữ liệu, trình bày dữ liệu và kết luận.

Tìm hiểu chi tiết từng cái một:

* Thu thập dữ liệu

Như chúng ta đã thảo luận về các nguồn thu thập dữ liệu, dữ liệu liên quan về mặt logic được thu thập từ các nguồn khác nhau, định dạng khác nhau, các loại khác nhau như từ XML, tệp CSV, phương tiện truyền thông xã hội, hình ảnh là dữ liệu có cấu trúc hoặc không có cấu trúc,…

* Lưu trữ dữ liệu

Dữ liệu được thu thập bây giờ cần được lưu trữ dưới dạng vật lý như giấy tờ, sổ ghi chép, và tất cả hoặc dưới bất kỳ hình thức vật lý nào khác. Hiện nay do khai thác dữ liệu và dữ liệu lớn, việc thu thập dữ liệu là rất lớn kể cả ở dạng có cấu trúc hay không có cấu trúc. Dữ liệu sẽ được lưu trữ dưới dạng kỹ thuật số để thực hiện phân tích và trình bày có ý nghĩa theo các yêu cầu ứng dụng.

* Sắp xếp dữ liệu

Sau bước lưu trữ, bước ngay lập tức sẽ là phân loại và lọc. Việc phân loại và lọc được yêu cầu để sắp xếp dữ liệu theo một số thứ tự có ý nghĩa và chỉ lọc ra những thông tin được yêu cầu giúp dễ dàng hình dung và phân tích.

* Xử lý dữ liệu

Một loạt quá trình xử lý hoặc sử dụng liên tục và xử lý được thực hiện để xác minh, chuyển đổi, tổ chức, tích hợp và trích xuất dữ liệu ở dạng đầu ra hữu ích để sử dụng xa hơn.

* Phân tích dữ liệu

Phân tích dữ liệu là quá trình áp dụng hoặc đánh giá dữ liệu một cách có hệ thống bằng cách sử dụng lập luận phân tích và logic để minh họa từng thành phần của dữ liệu được cung cấp và để có được kết quả hoặc quyết định được đưa ra.

* Trình bày dữ liệu và kết luận

Khi chúng ta đi đến kết quả phân tích, nó có thể được biểu diễn thành các dạng khác nhau như biểu đồ, tệp văn bản, tệp excel, đồ thị,…

Phần mềm đơn lẻ hoặc kết hợp các phần mềm có thể sử dụng để thực hiện lưu trữ, sắp xếp, lọc và xử lý dữ liệu tùy theo yêu cầu và khả thi. Nó có thể được thực hiện bằng phần mềm cụ thể theo tập hợp các hoạt động được xác định trước theo các yêu cầu ứng dụng.

### *2.1.3. Xử lý dữ liệu NAN*

Xử lý dữ liệu NaN (Not a Number) là một bước quan trọng trong quá trình tiền xử lý và phân tích dữ liệu. Dữ liệu NaN xuất hiện khi có giá trị bị thiếu hoặc không xác định trong tập dữ liệu. Đây có thể là do các giá trị bị lỗi, không hoàn chỉnh hoặc không tồn tại. Dữ liệu NaN có thể gây ảnh hưởng tiêu cực đến kết quả phân tích nếu không được xử lý đúng.

Dưới đây là một số phương pháp phổ biến để xử lý dữ liệu NaN:

* Xóa dữ liệu NaN: Nếu số lượng dữ liệu NaN không quá lớn và không ảnh hưởng đáng kể đến tập dữ liệu, bạn có thể xóa các hàng hoặc cột chứa NaN. Tuy nhiên, cần đánh giá cẩn thận để đảm bảo không mất mát quá nhiều thông tin quan trọng.
* Điền giá trị: Một cách phổ biến để xử lý dữ liệu NaN là điền vào chỗ trống bằng một giá trị nào đó. Có thể sử dụng các phương pháp như điền giá trị trung bình, giá trị trung vị, giá trị mode hoặc giá trị được tính toán từ dữ liệu khác. Việc chọn phương pháp điền giá trị phụ thuộc vào loại dữ liệu và mục tiêu của phân tích.
* Tạo cột mới: Thay vì xóa hoặc điền vào chỗ trống, bạn có thể tạo cột mới để đánh dấu vị trí dữ liệu NaN. Điều này giúp giữ lại thông tin NaN để nắm bắt các xu hướng hoặc mô hình dữ liệu sau này.
* Sử dụng các mô hình học máy đặc biệt: Có các mô hình học máy được phát triển đặc biệt để xử lý dữ liệu NaN, như kĩ thuật học tăng cường, học tập bán giám sát hoặc học tập bằng cách tạo mô hình dự đoán giá trị NaN.
* Thống kê đa biến: Nếu có sự tương quan giữa các biến trong tập dữ liệu, bạn có thể sử dụng các kỹ thuật thống kê đa biến để dự đoán giá trị NaN từ các biến khác.

### *2.1.4. Lọc nhiễu*

Trong Khoa học Dữ liệu, việc lọc nhiễu (noise) là một bước quan trọng để làm sạch dữ liệu và cải thiện chất lượng phân tích. Nhiễu có thể là các giá trị bất thường, dữ liệu không chính xác hoặc không mong muốn trong tập dữ liệu.

Quá trình lọc nhiễu trong khoa học dữ liệu có thể thực hiện qua các bước sau:

* Hiểu dữ liệu: Đầu tiên, bạn cần hiểu rõ về loại dữ liệu mà bạn đang làm việc. Xác định tính chất của dữ liệu, bao gồm loại dữ liệu (số, chuỗi, hình ảnh, âm thanh, v.v.), đặc điểm thống kê, phân phối và nhiễu có thể tồn tại trong dữ liệu.
* Khám phá và phân tích dữ liệu: Tiến hành khám phá và phân tích dữ liệu để xác định loại nhiễu có mặt trong dữ liệu. Các phương pháp như biểu đồ, biểu đồ tần số, đồ thị phân phối, và các công cụ khác có thể được sử dụng để phân tích dữ liệu và xác định nhiễu.
* Xác định phương pháp lọc nhiễu: Dựa trên hiểu biết về dữ liệu và nhiễu, chọn phương pháp lọc nhiễu thích hợp. Có nhiều phương pháp khác nhau như đã đề cập trong câu trả lời trước, bao gồm lọc trung bình, lọc trung vị, lọc Kalman, lọc Savitzky-Golay, lọc Fourier, và nhiều phương pháp khác.
* Thực hiện lọc nhiễu: Áp dụng phương pháp lọc nhiễu đã chọn vào dữ liệu. Các phương pháp này có thể được thực hiện thông qua viết mã hoặc sử dụng các thư viện phân tích dữ liệu phổ biến như Python (ví dụ: NumPy, SciPy) hoặc R.
* Đánh giá kết quả: Đánh giá kết quả của quá trình lọc nhiễu. So sánh dữ liệu trước và sau khi được lọc nhiễu để đảm bảo rằng nhiễu đã được giảm thiểu một cách hiệu quả và thông tin quan trọng không bị mất đi.
* Tinh chỉnh và tối ưu hóa: Nếu kết quả lọc nhiễu chưa đạt yêu cầu, bạn có thể tinh chỉnh và tối ưu hóa phương pháp lọc nhiễu. Điều này có thể bao gồm thay đổi cấu hình cửa sổ lọc, thay đổi tham số hoặc chọn phương pháp lọc khác.

## **2.1. Model**

### *2.1.1. Model là gì?*

Trong ngữ cảnh của dự đoán giá, thuật ngữ "model" ám chỉ một mô hình dự đoán. Một mô hình là một biểu diễn hoặc một hệ thống được xây dựng dựa trên dữ liệu và quy tắc hoặc mẫu có sẵn để dự đoán kết quả.

Trong dự đoán giá, một mô hình có thể là một thuật toán hoặc một tập hợp các quy tắc được sử dụng để dự đoán giá cả. Mô hình có thể được xây dựng từ dữ liệu lịch sử và được huấn luyện để nhận ra các mẫu, quy luật và sự tương quan trong dữ liệu đó. Sau đó, mô hình được sử dụng để dự đoán giá cả trong tương lai dựa trên các thông tin mới.

Một số ví dụ về các mô hình dự đoán giá bao gồm: mô hình hồi quy tuyến tính, mạng nơ-ron nhân tạo, cây quyết định, mô hình ARIMA (AutoRegressive Integrated Moving Average), mô hình học tăng cường, và các mô hình kết hợp như mô hình dự đoán bằng phương pháp bỏ phiếu (ensemble models).

Mô hình dự đoán giá có thể được xây dựng và cải thiện thông qua quá trình huấn luyện, điều chỉnh và đánh giá với mục tiêu đạt được độ chính xác và độ tin cậy tốt nhất trong dự đoán giá cả.

### *2.1.2. Linear Regression (Hồi Quy Tuyến Tính)*

Đây là thuật toán cơ bản nhất của Machine Learning, thuật toán Linear Regression (Hồi Quy Tuyến Tính) thuộc nhóm Supervised learning ( Học có giám sát ).

Trong khi sử dụng hồi quy tuyến tính, mục tiêu của chúng ta là để làm sao một đường thẳng có thể tạo được sự phân bố gần nhất với hầu hết các điểm. Do đó làm giảm khoảng cách (sai số) của các điểm dữ liệu cho đến đường đó.

Quá trình xây dựng mô hình Linear Regression trong khoa học dữ liệu thường bao gồm các bước sau:

1. Chuẩn bị dữ liệu: Tiền xử lý dữ liệu bằng cách loại bỏ các giá trị thiếu, xử lý ngoại lệ, chuẩn hóa dữ liệu nếu cần thiết và phân chia dữ liệu thành tập huấn luyện và tập kiểm tra.
2. Xác định biến độc lập và biến mục tiêu: Xác định biến độc lập (x) là biến mà chúng ta sẽ sử dụng để dự đoán biến mục tiêu (y).
3. Xây dựng mô hình: Sử dụng tập huấn luyện, tìm hiểu các tham số (hệ số và chặn) trong đường thẳng tốt nhất để phù hợp với dữ liệu. Phương pháp phổ biến để tìm hiểu các tham số là phương pháp của bình phương tối thiểu (Least Squares Method) hoặc các phương pháp tối ưu hóa khác như Gradient Descent.
4. Đánh giá mô hình: Đánh giá mô hình bằng cách sử dụng các độ đo như Mean Squared Error (MSE), Root Mean Squared Error (RMSE) hoặc Coefficient of Determination (R-squared) để xem mức độ phù hợp của mô hình với dữ liệu huấn luyện.
5. Dự đoán và đánh giá: Sử dụng mô hình đã được huấn luyện để dự đoán giá trị biến mục tiêu trên tập kiểm tra hoặc dữ liệu mới. Đánh giá mô hình dựa trên các độ đo tương tự như trong bước 4.

# CHƯƠNG III: XÂY DỰNG CHƯƠNG TRÌNH

## **3.1. Code**

**Code**

from sklearn.preprocessing import OrdinalEncoder

oe=OrdinalEncoder()

In [45]:

df[catcols]=oe.fit\_transform(df[catcols])

In [46]:

df[catcols]=df[catcols].astype('int')

In [47]:

from sklearn.linear\_model import LinearRegression

from sklearn.model\_selection import train\_test\_split

In [48]:

li=LinearRegression()

In [49]:

linkcode

X,y=df.drop(columns='Price'),df['Price']

X\_train, X\_test, y\_train, y\_test=train\_test\_split(X,y,random\_state=555)

In [51]:

model=li.fit(X\_train,y\_train)

In [52]:

print("The score of the model with all independent variables is:",model.score(X\_test,y\_test)\*100,"%")

The score of the model with all independent variables is: 62.1813359072383 %

In [53]:

linkcode

*#Process of improving the model, finding the highly correlated columns with price*

plt.figure(figsize=(12,5))

sns.heatmap(df.corr(),annot=True)

plt.show()

train=pd.concat((X\_train,y\_train),axis='columns')

test=pd.concat((X\_test,y\_test),axis='columns')

In [55]:

*#the most correlated independent values to price are 'StorageType', 'ProcessSpeed', 'Ram', 'ScreenResolution' with these values lets find*

*#out the rsquare value of the test dataset*

import statsmodels.formula.api as smf

from sklearn.metrics import r2\_score

In [56]:

model1=smf.ols(data=train,formula='Price~StorageType+ProcessSpeed+Ram+ScreenResolution').fit()

In [57]:

y\_pred1=model1.predict(test)

In [58]:

linkcode

print("The score of the model with 'StorageType', 'ProcessSpeed', 'Ram', 'ScreenResolution' as independent variables:",r2\_score(test['Price'], y\_pred1)\*100,"%")

*#the most correlated independent values to price are 'StorageType', 'Ram', 'ScreenResolution' with these values lets find*

*#out the rsquare value of the test dataset*

In [60]:

model2=smf.ols(data=train,formula='Price~StorageType+Ram+ScreenResolution').fit()

In [61]:

y\_pred2=model2.predict(test)

In [62]:

linkcode

print("The score of the model with 'StorageType', Ram', 'ScreenResolution' independent variables is:",r2\_score(test['Price'], y\_pred2)\*100,"%")

*#the most correlated independent values to price are 'Ram', 'ScreenResolution' with these values lets find*

*#out the rsquare value of the test dataset*

In [64]:

model3=smf.ols(data=train,formula='Price~Ram+ScreenResolution').fit()

In [65]:

y\_pred3=model3.predict(test)

In [66]:

linkcode

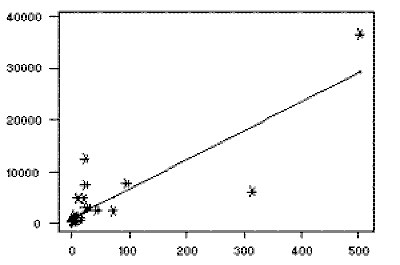
print("The score of the model with 'Ram', 'ScreenResolution' independent variables is:",r2\_score(test['Price'], y\_pred3)\*100,"%")

## **3.2. Thuật toán**

Thuật toán hồi quy tuyến tính là một phương pháp thống kê để tìm mối quan hệ tuyến tính giữa hai biến. Nó được sử dụng để dự đoán giá trị của một biến phụ thuộc (được gọi là biến phụ thuộc) dựa trên giá trị của một hoặc nhiều biến độc lập (được gọi là biến độc lập).

Thuật toán hồi quy tuyến tính giả định rằng mối quan hệ giữa biến phụ thuộc và biến độc lập có thể được biểu diễn bằng một đường thẳng hoặc siêu mặt phẳng trong không gian nhiều chiều. Mục tiêu của thuật toán là tìm ra các hệ số của đường thẳng hoặc siêu mặt phẳng sao cho sai số giữa giá trị dự đoán và giá trị thực tế là nhỏ nhất.

Thuật toán hồi quy tuyến tính thông thường được thực hiện bằng cách sử dụng phương pháp bình phương tối thiểu để tìm ra các hệ số tối ưu. Điều này có nghĩa là thuật toán tìm cách tối thiểu hóa tổng bình phương sai số giữa giá trị dự đoán và giá trị thực tế.



Trong bài toán dự đoán giá này của em thì em sử dụng qua các bước sau:

* Đầu tiên, các cột phân loại sẽ được mã hóa bằng OrdinalEncoder từ sklearn và thay đổi kiểu dữ liệu của nó thành số nguyên
* Chia dữ liệu thành 75% tập dữ liệu đào tạo và 25% tập dữ liệu thử nghiệm và thực hiện hồi quy tuyến tính
* Lắp tập dữ liệu huấn luyện vào mô hình và tính toán sai số bình phương trung bình gốc của tập dữ liệu thử nghiệm.
* Ra được độc chính xác là 62,1%
* Sử dụng mô-đun mô hình thống kê, chúng ta sẽ thực hiện phân tích trọng lượng và kết luận các tính năng quan trọng nhất ảnh hưởng đến giá của laptop.

# CHƯƠNG IV: KẾT LUẬN

## **4.1. Chương trình đã chạy được**

- Đầu tiên chương trình đã chạy được in ra các hàng,các cột trong DATA FRAME

- Dropping các cột không tên, lọc được các dữ liệu không xác định và xóa các giá trị trùng lặp

- In các giá trị duy nhất cho tất cả các cột

- Làm sạch được các cột trong data

- Chạy tìm ra được loại lưu trữ được sử dụng nhiều nhất trong tập dữ liệu

- Chạy tìm ra được công ty bán laptop giá cao nhất

- Chạy tìm ra được hệ điều hành phổ biến nhất

## **4.2. Mặt tốt và mặt xấu của chương trình**

### *4.2.1. Mặt tốt*

Có một số mặt tích cực của dự án dự đoán giá máy tính xách tay, như sau:

- Thuận lợi: Chương trình dự kiến ​​giá máy tính xách tay có thể cung cấp thông tin nhanh chóng và thuận tiện về giá trị dự kiến ​​của một chiếc máy tính xách tay cụ thể. Người dùng không cần phải tìm thông tin tìm kiếm từ nhiều nguồn khác nhau mà chỉ cần nhập thông tin liên quan và chương trình sẽ cung cấp dự đoán giá một cách tức thì.

- Tiết kiệm thời gian: Thay vì phải thăm dò thị trường hoặc tra cứu thông tin từ nhiều nguồn khác nhau, dự án dự đoán giá máy tính xách tay giúp tiết kiệm thời gian tiết kiệm cho người dùng. Việc dự đoán giá một cách nhanh chóng và chính xác có thể giúp người dùng đưa ra quyết định mua hàng nhanh hơn.

- Hỗ trợ ra quyết định: Chương trình dự đoán giá laptop có thể cung cấp cho người dùng một khung giá dự kiến, giúp họ có cái nhìn tổng thể về giá trị thực của laptop mà họ quan tâm. Điều này có thể giúp người dùng đưa ra quyết định mua hàng thông minh tốt hơn và tránh mua máy tính xách tay với giá không phù hợp.

- Thông tin tham khảo: Dự đoán giá máy tính xách tay có thể cung cấp cho người dùng thông tin tham khảo về xu hướng giá cả và các yếu tố ảnh hưởng đến giá trị của máy tính xách tay. Điều này có thể giúp người dùng hiểu rõ hơn về máy tính xách tay thị trường và giúp họ đưa ra quyết định mua sắm thông minh hơn trong tương lai.

- Cập nhật thông tin: Chương trình dự đoán giá laptop có thể được cập nhật thường xuyên để phản ánh các thay đổi trong thị trường. Điều này giúp người dùng có thông tin mới nhất về giá trị laptop và thay đổi xu hướng. Cập nhật thông tin định kỳ giúp chương trình duy trì tính chính xác và tin cậy.

### *4.2.2. Mặt xấu*

Dưới đây là một số mặt hạn chế hoặc xấu của chương trình dự đoán giá laptop:

* Độ chính xác hạn chế: Dù có tính chính xác cao, chương trình dự đoán giá laptop vẫn có thể gặp phải sai sót và đánh giá không chính xác về giá trị thực của một chiếc laptop. Giá cả của sản phẩm có thể bị ảnh hưởng bởi nhiều yếu tố khác nhau và có thể thay đổi theo thời gian. Do đó, dự đoán chỉ mang tính tham khảo và không thể đảm bảo tính chính xác tuyệt đối.
* Thiếu thông tin chi tiết: Chương trình dự đoán giá laptop dựa trên thông tin được cung cấp và các yếu tố được xem xét. Tuy nhiên, nó có thể thiếu thông tin chi tiết hoặc không có quyền truy cập vào dữ liệu mới nhất về giá cả và các yếu tố thị trường. Điều này có thể ảnh hưởng đến tính chính xác của dự đoán và không đưa ra được những thông tin cụ thể về giá trị thực của một sản phẩm.
* Ứng dụng hạn chế: Chương trình dự đoán giá laptop có thể hữu ích trong việc cung cấp thông tin tham khảo, nhưng nó không thể đưa ra các đánh giá sâu hơn về chất lượng, hiệu suất hoặc đặc điểm cụ thể của một sản phẩm. Điều này có nghĩa là người dùng vẫn cần nghiên cứu thêm và tham khảo từ các nguồn tin đáng tin cậy khác để có cái nhìn toàn diện hơn về một chiếc laptop.
* Ảnh hưởng yếu tố không xác định: Chương trình dự đoán giá laptop có thể không xem xét hoặc không đánh giá được một số yếu tố không xác định có thể ảnh hưởng đến giá trị của một laptop. Ví dụ, chương trình có thể không xem xét các yếu tố về thương hiệu, chế độ bảo hành, đánh giá của người dùng và sự khan hiếm của một sản phẩm cụ thể trên thị trường.

Tóm lại, chương trình dự đoán giá laptop có một số hạn chế và mặt hạn chế nhất định. Nó chỉ cung cấp thông tin tham khảo và không thể thay thế quá trình nghiên cứu, đánh giá và tham khảo từ nhiều nguồn thông tin khác nhau để đưa ra quyết định mua hàng thông minh và đáng tin cậy.

## **4.3. Ứng dụng phát triển về sau của chương trình**

Có nhiều ứng dụng phát triển tiềm năng cho dự án dự đoán giá laptop. Dưới đây là một số ví dụ về các phát triển có thể xảy ra trong tương lai:

* Nâng cao tính chính xác: Một mục tiêu phát triển quan trọng là nâng cao tính chính xác của chương trình dự đoán giá laptop. Các phương pháp tiên tiến hơn có thể được áp dụng, bao gồm sử dụng các mô hình học máy phức tạp hơn, xây dựng các mạng nơ-ron sâu (deep neural networks) hoặc sử dụng kỹ thuật học tăng cường (reinforcement learning) để tăng độ chính xác dự đoán.
* Tích hợp dữ liệu thêm: Mở rộng phạm vi dữ liệu đang được sử dụng là một ứng dụng tiềm năng. Bằng cách tích hợp thêm thông tin từ các nguồn dữ liệu khác như đánh giá người dùng, tin tức thị trường, xu hướng công nghệ và dữ liệu xã hội, chương trình có thể cung cấp dự đoán tốt hơn về giá trị thực của laptop.
* Phát triển ứng dụng di động: Xây dựng một ứng dụng di động cho chương trình dự đoán giá laptop sẽ tăng tính tiện lợi và sẵn có cho người dùng. Người dùng có thể truy cập và sử dụng chương trình dễ dàng trên điện thoại di động của họ để tra cứu và so sánh giá trị của các sản phẩm laptop một cách thuận tiện.
* Tích hợp công nghệ trí tuệ nhân tạo (AI) và học máy: Sử dụng công nghệ AI và học máy có thể đưa ra dự đoán giá laptop chính xác hơn và liên tục cải thiện theo thời gian. Các thuật toán học máy có thể được huấn luyện trên dữ liệu ngày càng lớn và đa dạng để cải thiện tính chính xác và khả năng dự đoán của chương trình.
* Phân tích dữ liệu thị trường thời gian thực: Để đảm bảo tính chính xác và đáng tin cậy của dự đoán, chương trình có thể được cập nhật với dữ liệu thị trường thời gian thực. Thông qua việc theo dõi và phân tích các xu hướng, sự biến đổi giá cả và yếu tố thị trường khác, chương trình có thể cung cấp thông tin cập nhật và dự đoán giá trị laptop theo thời gian thực.
* Phân tích dữ liệu người dùng: Chương trình có thể thu thập dữ liệu từ người dùng như phản hồi và đánh giá sau khi mua laptop. Thông qua việc phân tích dữ liệu này, chương trình có thể cải thiện dự đoán giá trị laptop dựa trên các trải nghiệm thực tế của người dùng.