**TRƯỜNG ĐH THỦ DẦU MỘT** **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**KHOA CÔNG NGHỆ THÔNG TIN**

**\*\*\*\*\*\*\*\*\*\*\***

****

**BÁO CÁO THỰC NGHIỆM HỆ HỖ TRỢ**

**QUYẾT ĐỊNH**

**ĐỀ TÀI**

**Nghiên cứu về Linear Regression - Hồi quy tuyến tính trong Machine Learning**

|  |  |  |  |
| --- | --- | --- | --- |
| Cán bộ hướng dẫn: | | Ths. Trần Thanh Hùng | |
| Nhóm: | | | 16 |
| Thành viên: | 1. Vũ Văn Hiếu - 2019606157 | | |
|  | 1. Đỗ Đình Hòa - 2019605868 | | |
|  | 1. Phạm Văn Mạnh - 2019605915 | | |
|  | 1. Nguyễn Minh Vũ - 2019603212 | | |
|  |  | | |

**Hà Nội, 2022**

**Hà Nội – Năm 2021**

**VIỆN KT-CN Độc lập - Tự do - Hạnh phú**c

**PHIẾU NHẬN XÉT BÁO CÁO TỐT NGHIỆP**

*(Dành cho giảng viên hướng dẫn)*

**…………………………………………………………………………………**

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

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

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

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

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

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

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

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

*Bình Dương, ngày...... tháng...... năm 2020*

***Giảng viên hướng dẫn***

***(Ký tên và ghi rõ họ tên)***

|  |
| --- |
|  |
| TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI  KHOA CÔNG NGHỆ THÔNG TIN  --------------------------------------- |
|  |
| BÁO CÁO THỰC NGIỆM  MÔN HỌC: HỆ HỖ TRỢ QUYẾT ĐỊNH |
|  |
| **Nghiên cứu về Linear Regression - Hồi quy tuyến tính trong Machine Learning** |
|  |
|  |
| GVHD : ThS.Trần Thanh Hùng  Nhóm : 16 |
| Sinh viên :  Vũ Văn Hiếu - 2019606157  Đỗ Đình Hòa - 2019605868  Phạm Văn Mạnh - 2019605915  Nguyễn Minh Vũ - 2019603212 |
| Lớp: 20221IT6058001 Khóa 14 |
|  |
|  |
|  |
| Hà Nội – Năm 2022 |

**Mục Lục**

[**LỜI NÓI ĐẦU** 4](#_Toc123252881)

[**CHƯƠNG 1. CƠ SỞ LÝ THUYẾT** 5](#_Toc123252882)

[**I.Lý thuyết ra quyết định** 5](#_Toc123252883)

[**1.** **Quyết định là gì?** 5](#_Toc123252884)

[**2.** **Ra quyết định là gì** 5](#_Toc123252885)

[**3.** **Tại sao phải hỗ trợ quyết định** 5](#_Toc123252886)

[**4.** **Bản chất của hỗ trợ ra quyết định** 6](#_Toc123252887)

[5. **Các yếu tố ảnh hưởng đến ra quyết định** 6](#_Toc123252888)

[**6.** **Người ra quyết định** 6](#_Toc123252889)

[**7.** **Hệ hỗ trợ quyết định** 6](#_Toc123252890)

[. 11](#_Toc123252891)

[**CHƯƠNG 2: ĐỀ TÀI** 12](#_Toc123252892)

[**I. Tính cấp thiết của việc xây dựng đề tài.** 12](#_Toc123252893)

[**1. Sơ lược về Machine Learning.** 12](#_Toc123252894)

[**2. Hồi quy tuyến tính.** 13](#_Toc123252895)

[**3. Tại sao hồi quy tuyến tính lại quan trọng?** 13](#_Toc123252896)

[**4. Hồi quy tuyến tính hoạt động như thế nào?** 13](#_Toc123252897)

[**5. Các bước trong hồi quy tuyến tính** 13](#_Toc123252898)

[**6.** **Hồi quy tuyến tính trong máy học là gì?** 14](#_Toc123252899)

[**7.** **Mối quan hệ tuyến tính** 14](#_Toc123252900)

[**8.** **Phần dư độc lập** 14](#_Toc123252901)

[**9.** **Tính chuẩn** 14](#_Toc123252902)

[**10.** **Phương sai không đổi** 14](#_Toc123252903)

[**11.** **Có những loại hồi quy tuyến tính nào?** 15](#_Toc123252904)

[**II. Mục đích nghiên cứu** 15](#_Toc123252905)

[**Chương 3: Xây dựng chương trình** 16](#_Toc123252906)

[**1.** **Mô tả bài toán** 16](#_Toc123252907)

[**2.** **Thực hiện bài toán** 16](#_Toc123252908)

[**TỔNG KẾT** 23](#_Toc123252909)

[**Tài liệu tham khảo** 25](#_Toc123252910)

# **LỜI NÓI ĐẦU**

Trong những năm gần đây, cùng với chức năng khai thác có tính chất tác nghiệp, việc khai thác các cơ sở dữ liệu phục vụ các yêu cầu trợ giúp quyết định ngày càng có ý nghĩa quan trọng và là nhu cầu to lớn trong mọi lĩnh vực hoạt động kinh doanh, quản lý kinh tế cũng như nghiên cứu khoa học, thống kê tình hình phát triển xã hội, dân số, .... Dữ liệu được lưu trữ và thu thập ngày càng nhiều nhưng người ra quyết định trong quản lý, kinh doanh lại cần những thông tin dưới dạng “tri thức” rút ra từ những nguồn dữ liệu đó hơn là chính những dữ liệu đó cho việc ra quyết định.Quá trình ra quyết định cần có nhiều phương pháp hỗ trợ, trong số những phương pháp đó có phương pháp Learner Regressionđược sử dụng rất phổ biến.

Em xin chân thành cảm ơn thầy **Trần Thanh Hùng** đã cung cấp kiến thức và tài liệu, hướng dẫn để có thể hoàn thành tiểu luận này.

# **CHƯƠNG 1. CƠ SỞ LÝ THUYẾT**

## **I.Lý thuyết ra quyết định**

Việc đưa ra quyết định đối với một vấn đề xuất hiện trong khắp các lĩnh vực, hoạt động sản xuất, đời sống con người, nghiên cứu, thống kê...mà đôi khi chúng ta không nhận ra. Từ những việc đơn giản như chọn một đôi dép phù hợp với bộ quần áo cho đến những việc lớn như phân bổ ngân sách vào các chương trình quốc gia đều là công việc đưa ra quyết định.

### **Quyết định là gì?**

Đó là một lựa chọn về “đường lối hành động” (Simon,1960; Costello & Zalkind,1963; Churchman,1968), hay “chiến lược hành động” (Fishburn,1964) dẫnđến “một mục tiêu mong muốn” (Churchman,1968).

### **Ra quyết định là gì**

Một quá trình lựa chọn có ý thức giữa hai hay nhiều phương án để chọn ra một phương án tạo ra được kết quả mong muốn trong các điều kiện ràng buộc đã biết.

Quyết định có thể là nhận thức ở dạng sự kiện,

“Chi $10,000 cho quảng cáo trong quý 2”

Quyết định có thể là nhận thức ở dạng quá trình,

“Trước tiên thực hiện A, sau đó B hai lần và nếu có đáp ứng tốt hãy thực hiện C”

Quyết định có thể là một hoạt động giàu kiến thức,

Quyết định có kết luậnnào thìhợp lý/hợp lệ trong hoàn cảnh nào?

Quyết định có thể là những thay đổi trạng thái kiến thức

Quyết định có chấp nhận một kiến thức mới không?

### **Tại sao phải hỗ trợ quyết định**

* Nhu cầu hỗ trợ ra quyết định

+ Ra quyết định luôn cần xử lý kiến thức

+ Kiến thức là nguyên liệu và thành phẩm của ra quyết định, cần được sở hữu hoặc tích lũy bởi người ra quyết định.

* Giới hạn về nhận thức: trí nhớ con người là có hạn trong khi con người có vô vàn các mối quan hệ cần phải nhớ phải ra quyết định.
* Giới hạn về kinh tế: Do vấn đề kinh phí cho dự án luôn có hạn nên muốn có một dự án thành công thì cần phải có kế hoạch sử dụng kinh phí hợp lý.
* Giới hạn về thời gian: Một dự án không thể kéo dài phải có kế hoạch thực hiện trong một khoảng thời gian nhất định, như vậy cần có kế hoạch phân công công việc phù hợp để kịp tiến độ, đảm bảo chất lượng.
* Áp lực cạnh tranh: kế hoạch và chiến lược thực hiện dự án hợp lý, chính xác luôn tạo nên thế mạnh cho doanh nghiệp trong cuộc cạnh tranh

### **Bản chất của hỗ trợ ra quyết định**

* Cung cấp thông tin, tri thức
* Có thể thể hiện qua tương tác người - máy, qua mô phỏng

### **Các yếu tố ảnh hưởng đến ra quyết định**

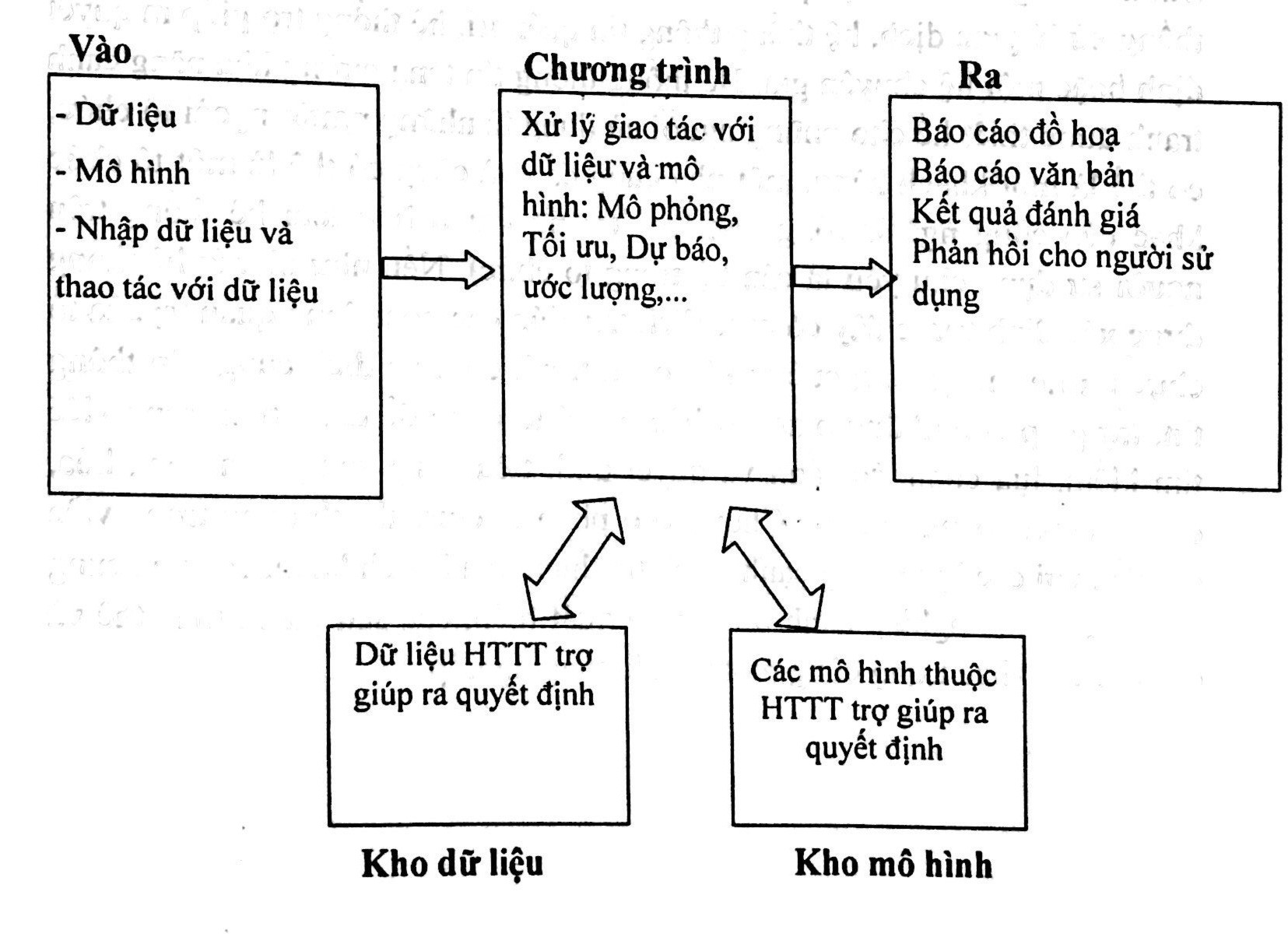
* Công nghệ - thông tin – máy tính.
* Tính cạnh tranh – sự phức tạp về cấu trúc.
* Các thay đổi biến động

### **Người ra quyết định**

* Ở cấp quản lý thấp hay tổ chức quy mô nhỏ: chính cá nhân là người ra quyết định. Đối với một cá nhân cũng có thể có nhiều mục tiêu xung đột.
* Tổ chức vừa và lớn: thường là nhóm ra quyết định, như vậy thường hay có nhiều mục tiêu xung đột.
* Đồng thuận là khó khăn nên quá trình nhóm ra quyết định rất phức tạp, thường cần máy tính hỗ trợ.

### **Hệ hỗ trợ quyết định**

* Khái niệm



Trong thập niên 1970, Scott Morton đưa ra khái niệm đầu tiên về Hệ hỗ trợ ra quyết định (Decision Support Systems -DSS). Ông định nghĩa DSS như là những hệ thống máy tính tương tác nhằm giúp những người ra quyết định sử dụng dữ liệu và mô hình để giải quyết các vấn đề không có cấu trúc.

Hệ hỗ trợ quyết định –HHTQĐ kết hợp trí lực của con người với năng lực của máy tính để cải tiến chất lượng của quyết định. Đây là các hệ dựa vào máy tính hỗ trợ cho người ra quyết định giải các bài toán nửa cấu trúc (Keen and Scott Morton, 1978).

HHTQĐ là tập các thủ tục dựa trên mô hình nhằm xử lý dữ liệu và phán đoán của con người để giúp nhà quản lý ra quyết định (Little, 1970).

Ưu thế của người ra quyết định:kinh nghiệm, khả năng trực giác,cóóc phán đoán, có trí thức. Ưu thế của máy tính:tốc độ, thông tin, khả năng xử lý.

Kết hợp cả ưu thế của người ra quyết định và máy tính, ta có ưu thế của Hệ hỗ trợ ra quyết định:tăng hiệu quả, tăngsự hiểu biết, tăng tốc độ, tăngtính linh hoạt, giảm sự phức tạp, giảm chi phí.

Hiện tại chưa có một định nghĩa thống nhất nào về DSS. Tuy nhiên tất cả đều đồng ý mục đích cơ bản nhấtcủa DSS là để hỗ trợ và cải tiến việc ra quyết định.



* Sử dụng hệ hỗ trợ quyết định

DSS có thể được sử dụng bởi quản trị vận hành và các phòng kế hoạch khác trong một tổ chức để biên soạn thông tin và dữ liệu, sau đó tổng hợp nó thành tin tức. Trên thực tế, hệ thống này chủ yếu được sử dụng bởi quản lí cấp trung đến cấp cao hơn.

* Đặc điểm của hệ hỗ trợ quyết định

Mục đích chính của việc sử dụng DSS là trình bày thông tin cho khách hàng một cách dễ hiểu. Một DSS rất hữu ích vì nó có thể được lập trình để tạo nhiều loại báo cáo, tất cả dựa trên thông số kĩ thuật của người dùng.

Khi công nghệ tiếp tục phát triển, phân tích dữ liệu không còn giới hạn trong những chiếc máy tính lớn, cồng kềnh. Vì DSS về cơ bản là một ứng dụng, nó có thể được tải trên hầu hết các hệ thống máy tính, cho dù trên máy tính bàn hay máy tính xách tay. Một số ứng dụng DSS cũng có sẵn trên các thiết bị di động.

Tính linh hoạt của hệ hỗ trợ quyết định cực kì có lợi cho những người dùng thường xuyên phải di chuyển. Điều này mang lại cho họ cơ hội nhận được thông tin chất lượng mọi lúc mọi nơi, cung cấp khả năng đưa ra quyết định tốt nhất cho công ty và khách hàng của họ khi đang di chuyển hoặc thậm chí ngay tại chỗ. (Theo Investopedia)

* Lý do dùng hệ hỗ trợ quyết định
* Nhu cầu về hệ hỗ trợ quyết định

Vào các năm 1980, 1990 điều tra các công ty lớn cho thấy:

* Kinh tế thiếu ổn định.
* Khó theo dõi vận hành của doanh nghiệp.
* Cạnh tranh gay gắt.
* Xuất hiện thương mại điện tử (e-commerce).
* Bộ phận IT quá bận, không giải quyết được các yêu cầu quản lý.
* Cần phân tích lợi nhuận, hiệu quả và thông tin chính xác, mới, kịp thời.
* Giảm giá phí hoạt động.
* Cải thiện tốc độ tính toán.
* Cải tiến kỹ thuật trong việclưu trữ, tìm kiếm, trao đổi dữ liệu trong và ngoài tổ chức theo hướng nhanh và kinh tế.
* Nâng cao chất lượng của các quyết định đưa ra.
* Tăng cường năng lực cạnh tranh của tổ chức.
* Khắc phục khả năng hạn chế của người trong việc xử lý và lưu chứa thông tin.
* Thuận lợi của hệ hỗ trợ quyết định
* Tăng số phương án xem xét, so sánh, phân tích độ nhanh nhạy, hiệu quả.
* Hiểu rõ các quan hệ nghiệp vụ trong toàn hệ thống tốt hơn.
* Đáp ứng nhanh trước các tình hướng không mong đợi, dễ điều chỉnh và thay đổi khi cần thiết.
* Có thể thực hiện các phân tích phi chính qui.
* Học tập và hiểu biết thêm các nguồn tài nguyên chưa được tận dụng.
* Cải thiện những cách thực hiện truyền thống.
* Kiểm soát kế hoạch, tiêu chuẩn hóa các thủ tục tính toán.
* Tiết kiệm chi phí cho các thủ tục hànhchính.
* Quyết định tốt hơn.
* Tiết kiệm thời gian.
* Dùng các nguồn dữ liệu tốt, có chọn lọc.

**Ưu điểm**

* Nó tiết kiệm thời gian bằng cách tăng tốc quá trình ra quyết định.
* Cải thiện giao tiếp giữa mọi người thông qua các cuộc họp, các buổi động não.
* Các báo cáo do hệ thống hỗ trợ huyết định tạo ra có thể được sử dụng làm bằng chứng.
* Nó giúp tự động hóa các quy trình.
* Giảm chi phí

**Nhược điểm**

* Giảm trạng thái
* Hiệu ứng không lường trước
* Chi phí bằng tiền

Quá nhiều phụ thuộc DSS

* Người dùng

Thành phần chính của Hệ thống Hỗ trợ Quyết định là người dùng. Vì DSS sử dụng thông thường là các nhà quản lý, hoạch định chính sách, có thể không phải là chuyên gia máy tính đủ trình độ, do đó DSS nên cung cấp các giao diện dễ sử dụng và một số hướng dẫn để sử dụng DSS cũng như tương tác với mô hình, chẳng hạn như nhận các khuyến nghị từ nó.

Mục tiêu chính của DSS là đảm bảo rằng người dùng đang sử dụng và được hưởng lợi từ DSS.

* Hệ thống quản lý cơ sở dữ liệu DBMS

DBMS hoạt động như một ngân hàng dữ liệu cho DSS. Nó lưu trữ số lượng lớn dữ liệu được thu thập từ các nguồn khác nhau. Nó cung cấp các cấu trúc dữ liệu logic để người dùng tương tác. Đầu vào và đầu ra được lưu trữ trong Cơ sở dữ liệu. Tất cả quá trình xử lý được thực hiện trong cơ sở dữ liệu.

* Hệ thống quản lý dựa trên mô hình (MBMS)

Chức năng của hệ thống này là tìm nạp dữ liệu và từ DBMS và chuyển đổi dữ liệu đó thành thông tin giúp đưa ra quyết định phù hợp. Nó cũng phải cung cấp hỗ trợ thích hợp cho người dùng để phát triển mô hình.

* Sử dụng hệ thống ra quyết định - DSS

Bản chất của DSS là gì không còn phải giới thiệu vì bạn đọc đã nắm được chi tiết thông tin của nó. Tuy nhiên, DSS chủ yếu được sử dụng trong kinh doanh và quản lý, giúp xác định các xu hướng tiêu cực và phân bổ kinh doanh tốt hơn.

Ngoài ra, nó có thể sử dụng để hỗ trợ quyết định lâm sàng trong chẩn đoán y tế. Nhân viên cho vay ngân hàng có thể sử dụng DSS để xác minh tín dụng của người xin vay hoặc một công ty kỹ thuật đang thực hiện các dự án khác nhau và muốn đảm bảo liệu họ có thể cạnh tranh với các khoản chi của mình hay không.

DSS không giới hạn trong máy tính, cồng kềnh khi nó cơ bản là một phần mềm và bạn có thể sử dụng thông qua điện thoại, máy tính. Điều này cho phép bạn luôn được cung cấp thông tin đầy đủ, cho phép bạn đưa ra quyết định tốt nhất mọi lúc cho doanh nghiệp và khách hàng của mình. Các hệ thống lý tưởng có thể sử dụng và phân tích thông tin để đưa ra lựa chọn của người dùng. DSS bao gồm dữ liệu tập trung vào mục tiêu hoặc dự báo doanh thu và bán hàng, số liệu bán hàng trong khoảng thời gian và hàng tồn kho khác.

# .

# **CHƯƠNG 2: ĐỀ TÀI**

## **I. Tính cấp thiết của việc xây dựng đề tài.**

### **1. Sơ lược về Machine Learning.**

Machine Learning là một tập hợp con của [trí tuệ nhân tạo (AI)](https://itgtechnology.vn/tri-tue-nhan-tao-la-gi-ai-thay-doi-tuong-lai-cua-loai-nguoi-ra-sao/)  và đang làm thay đổi các hoạt động nền sản xuất hiện đại. Ứng dụng phương pháp này, các nhà máy có thể dễ dàng rút ngắn thời gian sản xuất thông qua giảm thao tác. Vậy, Machine Learning là gì?

**Machine Learning là gì?**

Machine Learning có nghĩa là học máy. Đây là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống “học” tự động từ dữ liệu để giải quyết những vấn đề cụ thể. Các thuật toán của phương pháp học máy là các chương trình máy tính có khả năng học hỏi về cách hoàn thành các nhiệm vụ và cách cải thiện hiệu suất theo thời gian.

Máy có khả năng thích nghi với các điều kiện môi trường xung quanh để rút trích ra các nguyên lý từ tri thức thu nhận được phục vụ cho việc ra quyết định. Có thể kể đến một số sản phẩm của phương pháp học máy như: Cảnh báo giao thông trên ứng dụng Google Maps, Deepface của mạng xã hội Facebook, các máy có thể “học” cách phân loại thư điện tử xem có phải thư rác (spam) hay không và tự động xếp chúng vào thư mục tương ứng.

**Machine learning hoạt động như thế nào?**

Thuật toán Machine Learning được hướng dẫn để sử dụng một bộ dữ liệu đào tạo, từ đó tạo ra một mô hình nguyên mẫu. Khi thuật toán này tiếp nhận dữ liệu mới, nó sẽ đưa ra những dự đoán phân tích dựa trên nguyên mẫu căn bản.

Những phân tích nói trên sẽ được đánh giá về độ chính xác. Nếu độ chính xác này được công nhận, thuật toán Machine Learning sẽ tiến hành triển khai. Ngược lại, nếu độ chính xác không được công nhận, thuật toán sẽ được hướng dẫn lại nhiều lần với một bộ dữ liệu hướng dẫn tăng dần.

**Phương pháp học máy**

* Học máy có giám sát (Supervised Learning)
* Học máy không giám sát (Unsupervised Learning)
* Học tập bán giám sát (Semi-Supervised Learning)

Trong học phần này, chúng tôi sẽ giới thiệu một trong những thuật toán cơ bản nhất của Machine Learning. Đây là thuật toán Linear Regression (Hồi Quy Tuyến Tính) thuộc nhóm Supervised learning ( Học có giám sát )

### **2. Hồi quy tuyến tính.**

"Hồi quy tuyến tính" là 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. 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...

### **3. Tại sao hồi quy tuyến tính lại quan trọng?**

Các mô hình hồi quy tuyến tính tương đối đơn giản và cung cấp một công thức toán học dễ giải thích để đưa ra các dự đoán. Hồi quy tuyến tính là một kỹ thuật thống kê được sử dụng từ lâu và áp dụng dễ dàng cho phần mềm và tính toán. Các doanh nghiệp sử dụng nó để chuyển đổi dữ liệu thô một cách đáng tin cậy và có thể dự đoán được thành nghiệp vụ thông minh và thông tin chuyên sâu hữu ích. Các nhà khoa học trong nhiều lĩnh vực, bao gồm sinh học và các ngành khoa học hành vi, môi trường, và xã hội, sử dụng hồi quy tuyến tính để tiến hành phân tích dữ liệu sơ bộ và dự đoán các xu hướng tương lai. Nhiều phương pháp khoa học dữ liệu, chẳng hạn như máy học và trí tuệ nhân tạo, sử dụng hồi quy tuyến tính để giải quyết các bài toán phức tạp.

### **4. Hồi quy tuyến tính hoạt động như thế nào?**

Về bản chất, một kỹ thuật hồi quy tuyến tính đơn giản cố gắng vẽ một đồ thị đường giữa hai biến dữ liệu, x và y. Là biến độc lập, x được vẽ dọc theo trục hoành. Các biến độc lập còn được gọi là biến giải thích hoặc biến dự báo. Biến phụ thuộc, y, được vẽ trên trục tung. Bạn cũng có thể tham chiếu các giá trị y như các biến phản hồi hoặc các biến dự báo.

### **5. Các bước trong hồi quy tuyến tính**

Để có cái nhìn tổng quan, hãy xem xét dạng đơn giản nhất của phương trình đồ thị đường giữa y và x; y=c\*x+m, trong đó c và m là hằng số cho tất cả các giá trị có thể có của x và y. Vì vậy, chẳng hạn giả sử rằng tập dữ liệu đầu vào cho (x,y) là (1,5), (2,8), và (3,11). Để xác định phương pháp hồi quy tuyến tính, bạn sẽ thực hiện các bước sau:

Vẽ một đường thẳng và đo lường mối tương quan giữa 1 và 5.

Tiếp tục thay đổi hướng của đường thẳng cho các giá trị mới (2,8) và (3,11) cho đến khi tất cả các giá trị đều phù hợp.

Xác định phương trình hồi quy tuyến tính là y=3\*x+2.

Ngoại suy hoặc dự đoán y là 14 khi x là

### **Hồi quy tuyến tính trong máy học là gì?**

Trong máy học, các chương trình máy tính gọi là thuật toán phân tích các tập dữ liệu lớn và làm việc ngược từ dữ liệu đó để tính toán phương trình hồi quy tuyến tính. Các nhà khoa học dữ liệu đầu tiên sẽ đào tạo thuật toán trên các tập dữ liệu đã biết hoặc được dán nhãn và sau đó sử dụng thuật toán để dự đoán các giá trị chưa biết. Dữ liệu thực tế phức tạp hơn so với ví dụ trước. Đó là lý do tại sao phân tích hồi quy tuyến tính phải sửa đổi hoặc biến đổi các giá trị dữ liệu về mặt toán học để đáp ứng bốn giả định sau đây.

### **Mối quan hệ tuyến tính**

Mối quan hệ tuyến tính phải tồn tại giữa các biến độc lập và phụ thuộc. Để xác định mối quan hệ này, các nhà khoa học dữ liệu tạo ra một biểu đồ phân tán – một tập hợp ngẫu nhiên các giá trị x và y – để xem liệu chúng có nằm trên một đường thẳng hay không. Nếu không, bạn có thể áp dụng các hàm phi tuyến tính như căn bậc hai hoặc log để tạo ra mối quan hệ tuyến tính giữa hai biến về mặt toán học.

### **Phần dư độc lập**

Các nhà khoa học dữ liệu sử dụng phần dư để đo lường độ chính xác dự đoán. Phần dư là chênh lệch giữa dữ liệu quan sát được và giá trị dự đoán. Phần dư không được có một mô hình nhận dạng được giữa chúng. Ví dụ: bạn không muốn phần dư tăng lên theo thời gian. Bạn có thể sử dụng các bài kiểm tra toán học khác nhau, như kiểm tra Durbin-Watson, để xác định phần dư độc lập. Bạn có thể sử dụng dữ liệu giả để thay thế bất kỳ dữ liệu biến thiên nào, chẳng hạn như dữ liệu theo mùa.

### **Tính chuẩn**

Các kỹ thuật vẽ đồ thị như các biểu đồ Q-Q xác định xem các phần dư có phân bố bình thường hay không. Các phần dư sẽ nằm dọc theo một đường chéo ở trung tâm của đồ thị. Nếu phần dư không được chuẩn hóa, bạn có thể kiểm tra dữ liệu cho các điểm ngoại lai ngẫu nhiên hoặc các giá trị không phải là điển hình. Loại bỏ các điểm ngoại lai hoặc thực hiện phép biến đổi phi tuyến tính có thể khắc phục sự cố.

### **Phương sai không đổi**

Phương sai không đổi giả định rằng phần dư có phương sai cố định hoặc độ lệch chuẩn so với trung bình cho mọi giá trị của x. Nếu không, kết quả phân tích có thể không chính xác. Nếu giả định này chưa được đáp ứng, bạn có thể phải thay đổi biến phụ thuộc. Bởi vì phương sai xuất hiện tự nhiên trong các tập dữ liệu lớn, giải thích được lý do phải thay đổi quy mô của biến phụ thuộc. Ví dụ: thay vì sử dụng quy mô dân số để dự đoán số trạm cứu hỏa trong một thành phố, có thể sử dụng kích thước dân số để dự đoán số trạm cứu hỏa trên mỗi người.

### **Có những loại hồi quy tuyến tính nào?**

Một số loại phân tích hồi quy phù hợp để xử lý các tập dữ liệu phức tạp hơn các loại khác. Sau đây là một số ví dụ.

## **II. Mục đích nghiên cứu**

- Nghiên cứu, tìm hiểu Linear Regression trong Machine Learning

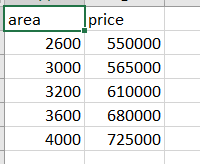
- Đánh giá hiệu quả của thuật toán

-Nhận xét, ưu, nhược điểm của thuật toán

# **Chương 3: Xây dựng chương trình**

1. **Mô tả bài toán**

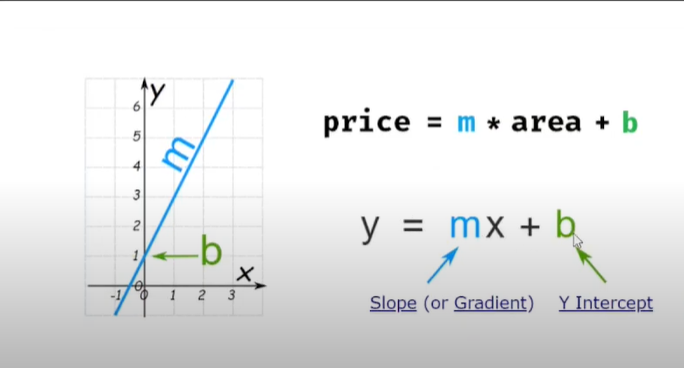
* Bài toán dự đoán giá nhà ở Monroe, Mỹ
* Bảng dưới đây thể hiện giá nhà ở thị trấn monroe dựa trên diện tích(m2)



Vấn đề: Với dữ liệu trên, hãy xây dựng mô hình học máy có thể dự đoán giá nhà dựa trên diện tích mét vuông

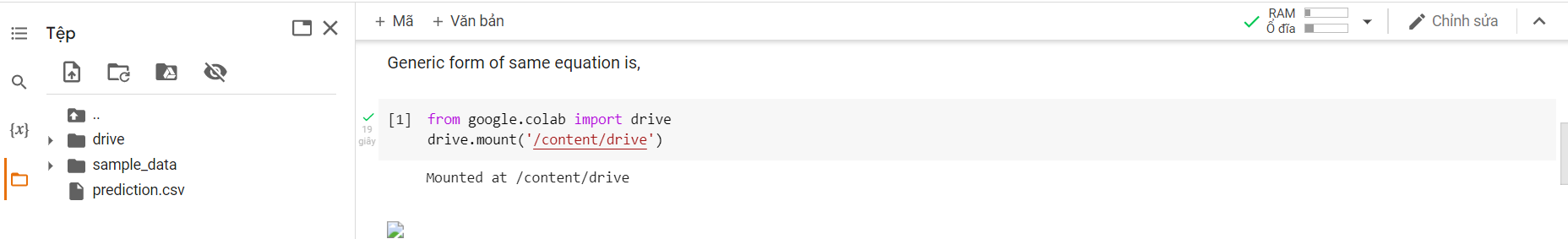
Có thể biểu thị các giá trị trong bảng dưới dạng biểu đồ phân tán (các giá trị được hiển thị bằng các điểm dấu màu đỏ). Sau đó, người ta có thể vẽ một đường thẳng phù hợp nhất với các giá trị trên biểu đồ

* Lưu ý: phương trình tuyến tính từ lớp toán trung cấp. giá nhà có thể được trình bày: price = m\*area +b
* Dạng chung của cùng một phương trình:



1. **Thực hiện bài toán**

* Sử dụng Colab. Tạo kết nối với Drive



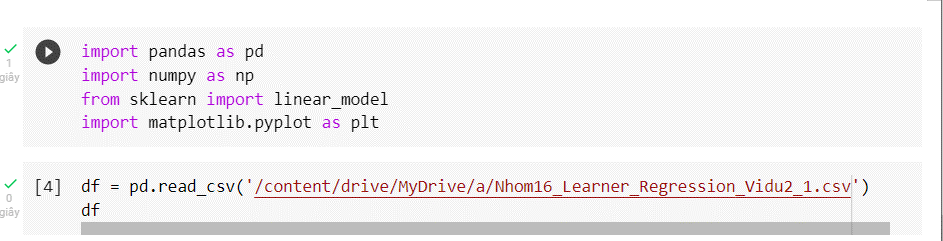
* Sử dụng các thư viện cần thiết

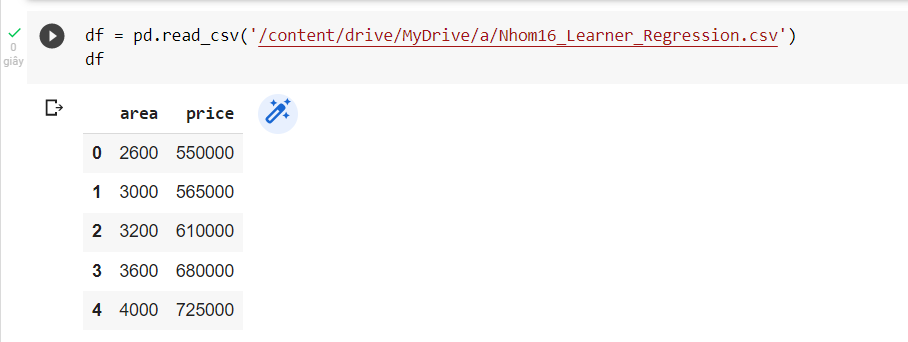
Pandas: Phân tích và xử lí dữ liệu

Numpy: Gói thư viện xử lý dãy số và ma trận nhiều chiều

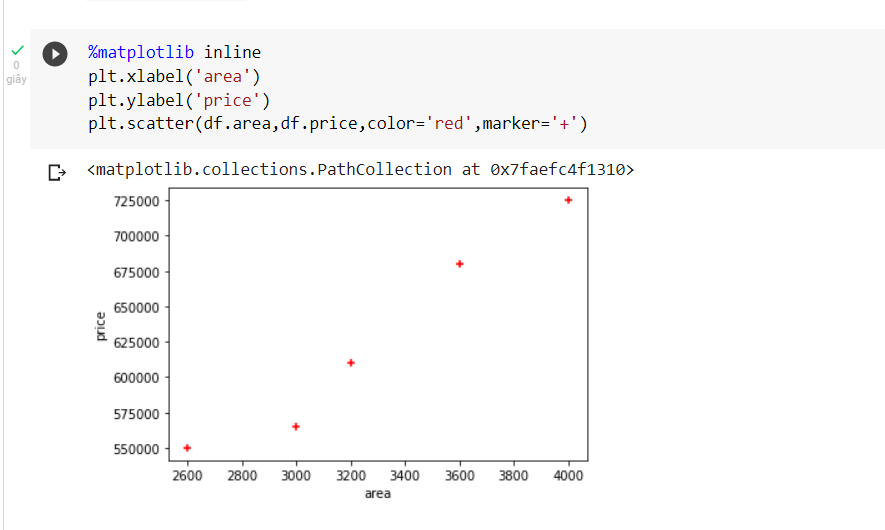
Sklearn: ác thuật toán học máy được viết trên ngôn ngữ Python

Matplotlib: biểu diễn dữ liệu dưới dạng đồ thị 2 chiều

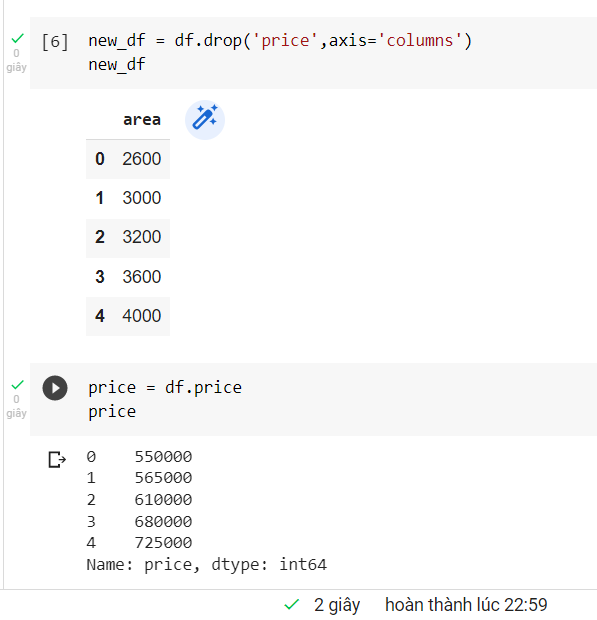




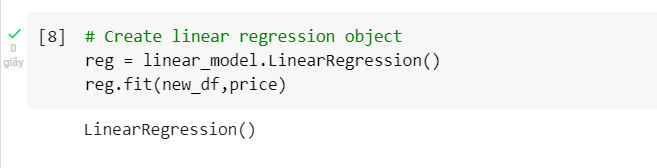
* Vẽ biểu thị của giá trị Price và Area với trục x là Price và y là Area.



* Tạo biến độc lập new\_df để thực hiện dự đoán cho biến price (X\_train) và gán giá trị cột price cho biến price (Y\_train)

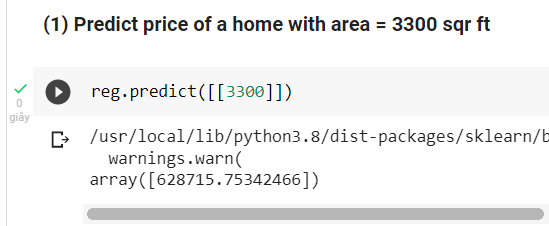


* Sử dụng mô hình hồi quy tuyến tính



Sử dụng mô hình gọi đến phương thức predict([Giá\_trị]) để lấy giá trị price

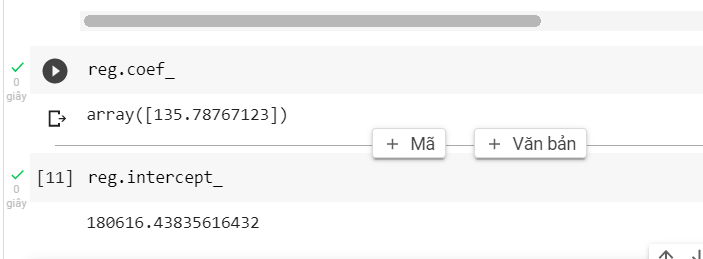
Ví dụ. tính giá của ngôi nhà với diện tích là 3300 m2



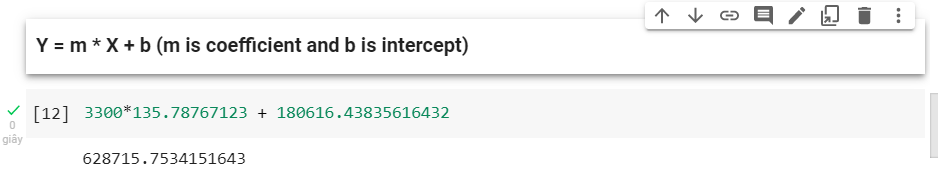
* Giá = 628715,75342…

\*Giải thích mô hình

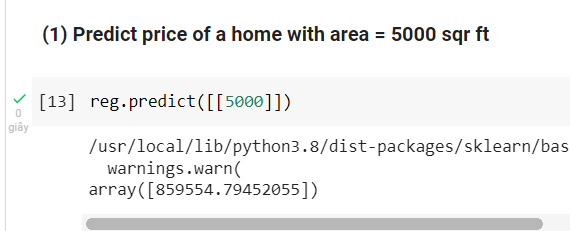
ứng dụng phương trình y = m\*X +b (m là hệ số và b là hằng số chặn)



ở đây X = giá trị trong predict, m = reg.coef\_, b = reg.intercept

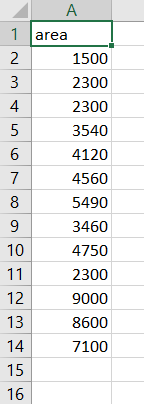


* Được giá trị Y. ta thấy Y = m\*X + b = predict([3300]) của mô hình hồi quy tuyến tính.
* Tiếp tục, với diện tích 5000

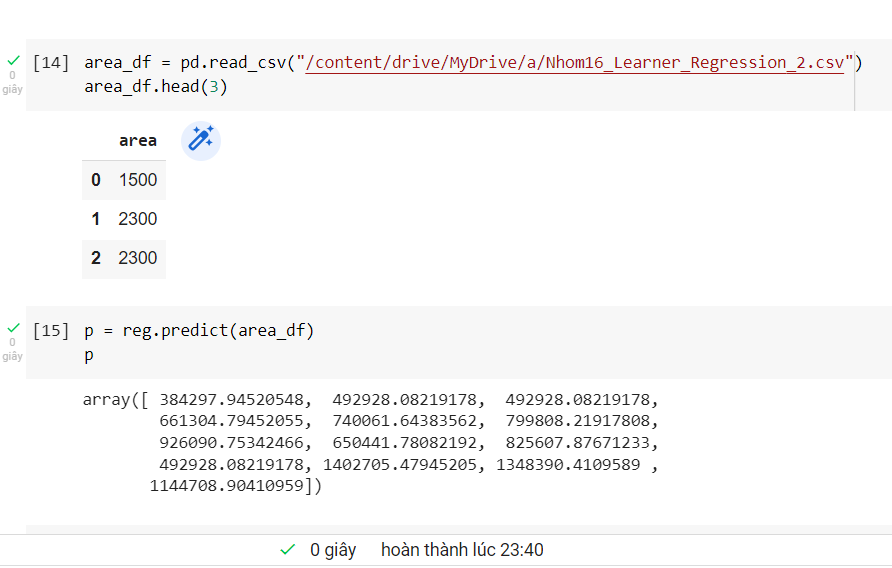


* Tương tự, lọc dữ liệu price để lấy dữ liệu area

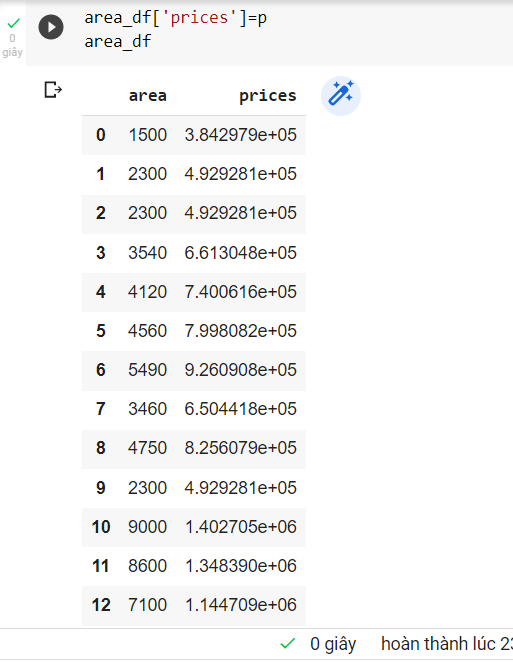
Input:



Ta tạo 1 file area.csv chứa dữ liệu area



Output:



Đây là dự đoán mô hình 1 biến.

# **TỔNG KẾT**

**Những kết quả đạt được:**

- Sự hiểu biết về thuật toán Learner Regression cơ bản tương đối tốt. (thuộc nhóm Supervised learning)

- Ứng dụng được vào các bài toán, lĩnh vực thực tế trong đời sống

- Từ những gì đã làm được, từ đó hiểu biết thêm về AI, ứng dụng vào đời sống công nghệ hiện đại.

- Làm quen ngôn ngữ lập trình Python.

**Những hạn chế:**

- Linear Regression là một thuật toán đơn giản và dễ cài đặt cho một bài toán với quan hệ tuyến tính giữa input và output. Tuy vậy Linear Regression có một nhược điểm lớn là nhạy cảm với nhiễu.

- Độ chính xác của thuật toán chưa tuyệt đối.

# **Tài liệu tham khảo**

[1] https://aicurious.io/posts/linear-regression/

[2] https://viblo.asia/p/linear-regression-hoi-quy-tuyen-tinh-trong-machine-learning-4P856akRlY3

[3] https://dominhhai.github.io/vi/2017/12/ml-linear-regression/