TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN**

**NHẬP MÔN HỌC MÁY**

**ĐỀ TÀI:**

**KỸ THUẬT HỒI QUY TUYẾN TÍNH DỰ ĐOÁN ĐIỀU KIỆN NHẬP HỌC VÀ ĐIỂM TRUNG BÌNH CỦA SINH VIÊN**

|  |
| --- |
| **Sinh viên thực hiện** |
| **Giảng viên hướng dẫn** | **: ĐÀO NAM ANH** |
| **Ngành** | **: CÔNG NGHỆ THÔNG TIN** |
| **Chuyên ngành** | **: HTTMDT** |
| **Lớp** | **: D15HTTMDT2** |
| **Khóa** | **: 2018-2023** |

***Hà Nội, tháng 12 năm 2022***

**PHIẾU CHẤM ĐIỂM**

Sinh viên thực hiện:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Họ và tên** | | **Chữ ký** | | **Ghi chú** |
|  |  | |
|  |  | |
|  |  | |

Giảng viên chấm:

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

**MỤC LỤC**

[LỜI CẢM ƠN 1](#_Toc75878152)

[LỜI MỞ ĐẦU 2](#_Toc75878153)

[CHƯƠNG 1. TỔNG QUAN VỀ HỌC MÁY VÀ KỸ THUẬT HỒI QUY TUYẾN TÍNH, ỨNG DỤNG TRONG GIÁO DỤC 4](#_Toc75878154)

[1.1. Khái niệm về học máy: 4](#_Toc75878155)

[1.2. Học giám sát và học không giám sát: 4](#_Toc75878156)

[1.2.1 Học có giám sát (Supervised Learning)-SL 4](#_Toc75878157)

[1.2.2. Học không giám sát (Unsupervised Learning)-UL: 6](#_Toc75878158)

[1.3. Ứng dụng của học máy: 8](#_Toc75878159)

[1.4. Khái niệm phân lớp dự báo: 9](#_Toc75878160)

[1.5. Mô hình Regression: 11](#_Toc75878161)

[1.5.1 Giới thiệu mô hình Regression: 11](#_Toc75878162)

[1.5.2 Mô hình Linear Regression: 14](#_Toc75878163)

[1.5.3 Sai số của dự đoán 15](#_Toc75878164)

[1.5.4. Xác định Basic Function: 16](#_Toc75878165)

[1.5.5 Hạn chế của mô hình Linear Regression: 16](#_Toc75878166)

[1.6. Ứng dụng trong Giáo dục: 17](#_Toc75878167)

[CHƯƠNG 2. THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ 19](#_Toc75878168)

[2.1 Bài toán dự đoán điều kiện nhập học với mô hình Regression 19](#_Toc75878169)

[2.1.1 Phát biểu bài toán: 19](#_Toc75878170)

[2.1.2 Chuẩn bị dữ liệu: 19](#_Toc75878171)

[2.1.3 Xử lý dữ liệu: 20](#_Toc75878172)

[2.1.4.Code và chạy dữ liệu 21](#_Toc75878173)

[2.2 Bài toán trung bình tích lũy với mô hình Regression: 24](#_Toc75878174)

[2.2.1 Phát biểu bài toán: 24](#_Toc75878175)

[2.2.2 Chuẩn bị dữ liệu: 24](#_Toc75878176)

[2.2.3 Xử lý dữ liệu: 24](#_Toc75878177)

[2.2.4.Code và chạy dữ liệu: 25](#_Toc75878178)

[KẾT LUẬN 29](#_Toc75878179)

[TÀI LIỆU THAM KHẢO 30](#_Toc75878180)

DAnh mục hình ảnh

[Hình 1. 1. Mô hình học có giám sát 5](#_Toc16982)

[Hình 1. 2. Mô hình học không giám sát 6](#_Toc22431)

[Hình 1. 3. Sự khác biệt giữa 2 mô hình SL và UL 7](#_Toc32742)

[Hình 1.4.Mô hình Regression Tree dự báo giá của một chiếc xe Toyota 11](#_Toc10840)

[Hình 1.5. Ví dụ về phép toán 13](#_Toc28190)

[Hình 1.6. Quan hệ y=3+4x 13](#_Toc4774)

[Hình 1. 7. Đồ thị thể hiện giá trị hiện tại và giá trị tương lai của dữ liệu. 17](#_Toc27119)

[Hình 2.1.Dữ liệu điều kiện nhập học với mô hình Regression 19](#_Toc6178)

[Hình 2.2.dữ liệu trung bình tích lũy với mô hình Regression 24](#_Toc10578)

# **LỜI CẢM ƠN**

Lời đầu tiên, em xin chân thành gửi lời cảm ơn tới các thầy cô giáo trong Trường Đại học Điện Lực nói chung và các thầy cô giáo trong Khoa Công nghệ thông tin nói riêng đã tận tình giảng dạy, truyền đạt cho chúng em những kiến thức cũng như kinh nghiệm quý báu trong suốt quá trình học.

Đặc biệt, em gửi lời cảm ơn đến Giáo viên hướng dẫn Đào Nam Anh, thầy đã tận tình theo sát giúp đỡ, trực tiếp chỉ bảo, hướng dẫn trong suốt quá trình nghiên cứu và học tập của chúng em. Trong thời gian học tập với thầy, chúng em không những tiếp thu thêm nhiều kiến thức bổ ích mà còn học tập được tinh thần làm việc, thái độ nghiên cứu khoa học nghiêm túc, hiệu quả. Đây là những điều rất cần thiết cho chúng em trong quá trình học tập và công tác sau này. Chúng em muốn gửi lời cảm ơn đặc biệt nhất, sâu sắc nhất, thân thương nhất đến thầy và chúc thầy luôn dồi dào sức khỏe, tiếp tục giảng dạy hết tâm huyết của mình cho những lứa học trò sau này để đất nước ta ngày càng có nhiều nhân tài, những người giỏi trong các doanh nghiệp, xây dựng đất nước phát triển hơn nữa.

Em xin chân thành cảm ơn!

# **LỜI MỞ ĐẦU**

1. Tính cấp thiết của đề tài

Công nghệ ngày càng phổ biến và không ai có thể phủ nhận được tầm quan trọng và những hiệu quả mà nó đem lại cho cuộc sống chúng ta. Bất kỳ trong lĩnh vực nào, sự góp mặt của trí tuệ nhân tạo sẽ giúp con người làm việc và hoàn thành tốt công việc hơn. Và gần đây, một thuật ngữ “machine learning” rất được nhiều người quan tâm.Thay vì phải code phần mềm với cách thức thủ công theo một bộ hướng dẫn cụ thể nhằm hoàn thành một nhiệm vụ đề ra thì máy sẽ tự “học hỏi” bằng cách sử dụng một lượng lớn dữ liệu cùng những thuật toán cho phép nó thực hiện các tác vụ.

Đây là một lĩnh vực khoa học tuy không mới, nhưng cho thấy lĩnh vực trí tuệ nhân tạo đang ngày càng phát triển và có thể tiến xa hơn trong tương lai. Đồng thời, thời điểm này nó được xem là một lĩnh vực “nóng” và dành rất nhiều mối quan tâm để phát triển nó một cách mạnh mẽ, bùng nổ hơn.

Hiện nay, việc quan tâm machine learning càng ngày càng tăng lên là vì nhờ có machine learning giúp gia tăng dung lượng lưu trữ các loại dữ liệu sẵn, việc xử lý tính toán có chi phí thấp và hiệu quả hơn rất nhiều.

Những điều trên được hiểu là nó có thể thực hiện tự động, nhanh chóng để tạo ra những mô hình cho phép phân tích các dữ liệu có quy mô lớn hơn và phức tạp hơn đồng thời đưa ra những kết quả một cách nhanh và chính xác hơn.

Chính sự hiệu quả trong công việc và các lợi ích vượt bậc mà nó đem lại cho chúng ta khiến machine learning ngày càng được chú trọng và quan tâm nhiều hơn. Vì vậy chúng em đã chọn đề tài: ”Kỹ thuật quy hồi tuyến tính và ứng dụng trong Giáo dục”để làm báo cáo.

2. Mục tiêu nghiên cứu

Báo cáo xây dựng được mô hình dự báo ứng dụng trong Giáo dục. Từ đó là cơ sở cho lập trình xây dựng mô hình dự đoán điểm thi đầu vào của các trường Đại hoc; Điểm sinh viên tốt nghiệp;…...

3. Đối tượng và phạm vi nghiên cứu

**Đối tượng nghiên cứu:** Đối tượng nghiên cứu của báo cáo là phân tích và thiết kế theo hướng đối tượng về Giáo dục như điểm thi, sinh viên đạt học bổng,….

Phạm vi nghiên cứu: Đồ án chỉ nghiên cứu trong phạm quy nhu cầu thực tế của Giáo dục Việt Nam là các trường Đại học.

4. Phương pháp nghiên cứu

Báo cáo sử dụng phương pháp phân tích và diễn giải thực trạng như cầu dự đoán điểm đầu vào, số sinh viên đạt học bổng,.. của các trường Đại học. Tài liệu được tham khảo tại các trường Đại học ở Việt Nam. Ngoài ra báo cáo còn sử dụng phương pháp định tính để đưa ra các giải pháp hoàn thiện cho bài toán cần giải quyết.

5. Kết cấu báo cáo:

Báo cáo gồm 2 chương:

+ Chương 1: Tổng quan về học máy và kỹ thuật quy hồi tuyến tính, ứng dụng trong Giáo dục.

+ Chương 2: Thử nghiệm và đánh giá kết quả

# **CHƯƠNG 1. TỔNG QUAN VỀ HỌC MÁY VÀ KỸ THUẬT HỒI QUY TUYẾN TÍNH, ỨNG DỤNG TRONG GIÁO DỤC**

## **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](https://vietnambiz.vn/du-lieu-lon-big-data-la-gi-nhung-kho-khan-khi-su-dung-du-lieu-lon-20190923114156622.htm) (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.

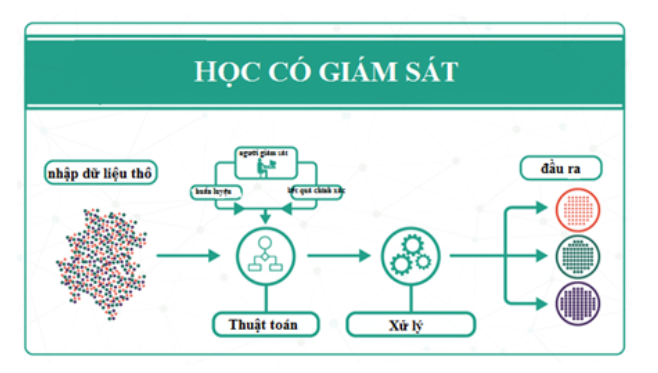
Học máy có 2 loại Học có giám sát (*Supervised Learning) và* Học không có giám sát*(Unsupervised Learning)*

## **1.2. Học giám sát và học không giám sát:**

### **1.2.1 Học có giám sát (*Supervised Learning*)-SL**

- Là phương pháp sử dụng những dữ liệu đã được gán nhãn từ trước để suy luận ra quan hệ giữa đầu vào và đầu ra. Các dữ liệu này được gọi là dữ liệu huấn luyện và chúng là cặp các đầu vào-đầu ra.

Học có giám sát sẽ xem xét các tập huấn luyện này để từ đó có thể đưa ra dự đoán đầu ra cho 1 đầu vào mới chưa gặp bao giờ. Mỗi bộ dữ liệu có cấu trúc theo cặp {x, y} với x được xem là dữ liệu thô (raw data) và y là nhãn của dữ liệu đó. Nhiệm vụ của SL là dự đoán đầu ra mong muốn dựa vào giá trị đầu vào. Dễ nhận ra, học có GIÁM SÁT tức là máy học dựa vào sự trợ giúp của con người, hay nói cách khác con người dạy cho máy học và giá trị đầu ra mong muốn được định trước bởi con người. Tập dữ liệu huấn luyện hoàn toàn được gán nhãn dựa vào con người. Tập càng nhỏ thì máy tính học càng ít. SL cũng được áp dụng cho 2 nhóm bài toán chính là bài toán dự đoán (regression) và bài toán phân lớp (classification), dự đoán giá nhà, phân loại email dựa trên tập dữ liệu mẫu - tập huấn luyện (*training data*).



Hình 1. 1. Mô hình học có giám sát

Học có giám sát là một hướng tiếp cận của Máy học để làm cho máy tính có khả năng "học". Trong hướng tiếp cận này, người ta "huấn luyện" máy tính dựa trên những quan sát có dán nhãn. Ta có thể hình dung những quan sát này như là những câu hỏi, và nhãn của chúng là những câu trả lời.

Ý tưởng của học có giám sát là: bằng việc ghi nhớ và tổng quát hóa một số quy tắc từ một tập câu hỏi có đáp án trước, máy tính sẽ có thể trả lời được những câu hỏi dù chưa từng gặp phải, nhưng có mối liên quan.

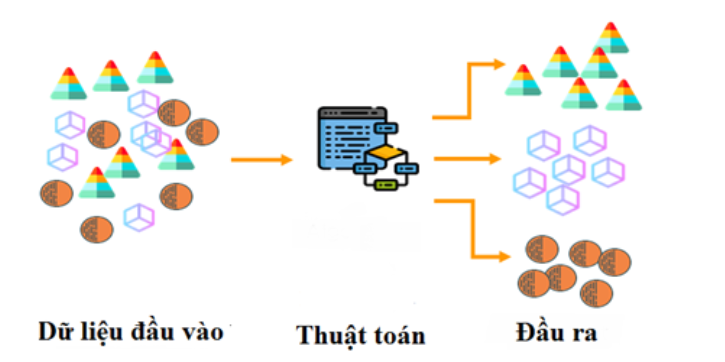
Ví dụ ta dạy máy tính "1 + 1 = 2" và hy vọng nó sẽ học được phép tính cộng x + 1 và trả lời được là "2 + 1 = 3". Học có giám sát mô phỏng việc con người học bằng cách đưa ra dự đoán của mình cho một câu hỏi, sau đó đối chiếu với đáp án. Sau đó con người rút ra phương pháp để trả lời đúng không chỉ câu hỏi đó, mà cho những câu hỏi có dạng tương tự.

Trong học có giám sát, các quan sát bắt buộc phải được dán nhãn trước. Đây chính là một trong những nhược điểm của phương pháp này, bởi vì không phải lúc nào việc dán nhãn chính xác cho quan sát cũng dễ dàng. Ví dụ như trong dịch thuật, từ một câu của ngôn ngữ gốc có thể dịch thành rất nhiều phiên bản khác nhau trong ngôn ngữ cần dịch sang. Tuy nhiên, việc quan sát được dán nhãn cũng lại chính là ưu điểm của học có giám sát bởi vì một khi đã thu thập được một bộ dữ liệu lớn được dán nhãn chuẩn xác, thì việc huấn luyện trở nên dễ dàng hơn rất nhiều so với khi dữ liệu không được dán nhãn.

### **1.2.2. Học không giám sát (*Unsupervised Learning*)-UL:**

Unsupervised Learning là một nhóm thuật toán hay phương pháp kỹ thuật cho phép máy tự học hỏi và tìm ra một mô hình hay cấu trúc nào đó ẩn trong một bộ dữ liệu không được gắn nhãn trước. Điều này đồng nghĩa với việc chúng ta chỉ có bộ dữ liệu đầu vào và hoàn toàn không biết comeout là gì.

Ứng dụng phổ biến nhất của học không giám sát là gom cụm (cluster)

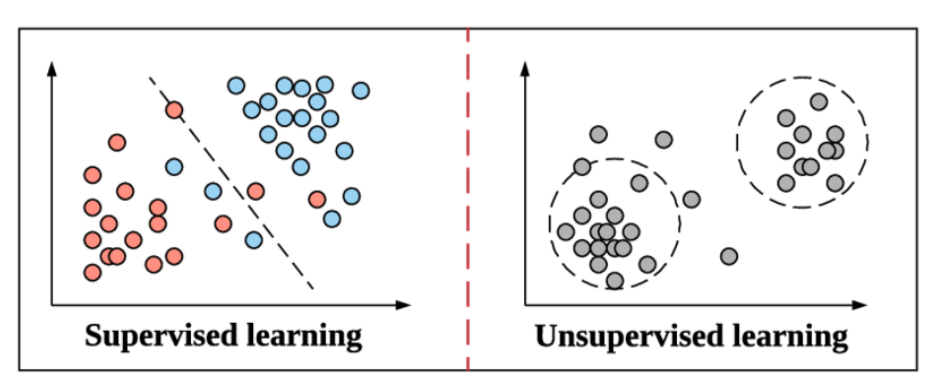


Hình 1. 2. Mô hình học không giám sát

Trong thuật toán này, chúng ta không biết được *dữ liệu đầu ra* hay *nhãn* mà chỉ có dữ liệu đầu vào. Thuật toán Học không giám sát dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ như phân nhóm hoặc giảm số chiều của dữ liệu để thuận tiện trong việc lưu trữ và tính toán.

Một cách toán học, Học không giám sát là khi chúng ta chỉ có dữ liệu vào X mà không biết *nhãn* Y tương ứng.

Sự khác nhau giữa học có giám sát và học không giám sát:



Hình 1. 3. Sự khác biệt giữa 2 mô hình SL và UL

Học có giám sát: Là cách huấn luyện một mô hình trong đó dữ liệu học có đầu vào và đầu ra tương ứng đầu vào đó. Mô hình được huấn luyện bằng cách giảm thiểu sai số lỗi (loss) của các dự đoán tại các vòng lặp huấn luyện. Sau quá trình huấn luyện. mô hình sẽ có khả năng đưa ra dự đoán về đầu ra với một đầu vào mới gặp (không có trong dữ liệu học). Nếu không gian đầu ra được biểu diễn dứới dạng rời rạc, ta gọi đó là bài toán phân loại (classification). Nếu không gian đầu ra được biểu diễn dưới dạng liên tục, ta gọi đó là bài toán hồi quy (regression).

Học không giám sát: Là cách huấn luyện một mô hình trong đó dữ liệu học chỉ bao gồm đầu vào mà không có đầu ra. Mô hình sẽ được huấn luyện cách để tìm cấu trúc hoặc mối quan hệ giữa các đầu vào. Một trong những phương pháp học không giám sát quan trọng nhất là phân cụm (clustering): Tạo các cụm khác nhau với mỗi cụm biểu diễn một đặc trưng nào đó của dữ liệu và phân các đầu vào mới vào các cụm theo các đặc trưng của đầu vào đó. Các phương pháp học không giám sát khác có thể kể đến như: phát hiện điểm bất thường (anomaly detection), Singular-value decomposition, …

Giả sử bạn cần nhận ra xe nào là xe hơi và xe nào là xe máy.

Trong học có giám sát trường hợp học tập, tập dữ liệu đầu vào (đào tạo) của bạn cần được dán nhãn, nghĩa là, đối với mỗi yếu tố đầu vào trong tập dữ liệu đầu vào (đào tạo) của bạn, bạn nên chỉ định nếu nó đại diện cho ô tô hoặc xe máy .

Trong học không giám sát trường hợp học tập, bạn không gắn nhãn đầu vào. Mô hình không giám sát sẽ phân cụm đầu vào thành các cụm dựa trên, ví dụ: trên các tính năng/tính chất tương tự. Vì vậy, trong trường hợp này, không có nhãn như "xe hơi".

## **1.3. Ứng dụng của học máy:**

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

+ Phát hiện gian lận.

+ Các kết quả tìm kiếm trên web.

+ Quảng cáo theo thời gian thực trên các trang web và các thiết bị di động.

+ Điểm số tín dụng và lời mời chào tiếp theo tốt nhất.

+ Dự đoán những hư hỏng của thiết bị.

+ Những mô hình định giá mới.

+ Phát hiện xâm nhập mạng.

+ Nhận dạng qua pattern (mẫu) và hình ảnh.

+ Lọc bỏ các email spam.

+ Trong công nghệ thông tin, cụ thể là lĩnh vực Machine Learning (Học Máy), học không giám sát thường được ứng dụng để chia hay phân loại nhóm và kết hợp, trong đó:

+ *Phân nhóm* hay chia nhóm là khi máy phân tích, tìm hiểu và khám phá các nhóm vốn có bên trong bộ dữ liệu, từ đó giúp bạn phân nhóm các khách hàng của mình theo hành vi mua hàng hay phân loại các bài viết, hình ảnh có cùng nội dung chẳng hạn.

+ *Kết hợp*: được vận dụng khi máy khám phá các quy tắc mô tả dữ liệu, ví dụ như người mua món hàng này sẽ có xu hướng mua một món hàng tương ứng nào đó có liên quan.

+ *Marketing và sales*: Dựa trên hành vi mua hàng trước đây, các trang web sử dụng Machine Learning phân tích lịch sử mua hàng, từ đó giới thiệu những vật dụng mà bạn có thể sẽ quan tâm và yêu thích. Khả năng tiếp nhận dữ liệu, phân tích và sử dụng những dữ liệu đó để cá nhân hóa trải nghiệm mua sắm (hoặc thực hiện chiến dịch Marketing) chính là tương tai của ngành bán lẻ.

+ Ứng dụng trong các mạng xã hội

Học máy đang được sử dụng trong một loạt các ứng dụng ngày nay. Một trong những ví dụ nổi tiếng nhất là Facebook News Feed. Nguồn cấp tin tức sử dụng học máy để cá nhân hóa từng nguồn cấp dữ liệu thành viên. Nếu một thành viên thường xuyên dừng lại để đọc hoặc thích một bài đăng của một người bạn cụ thể, News Feed sẽ bắt đầu hiển thị nhiều hơn về hoạt động của người bạn đó trước đó trong nguồn cấp dữ liệu. Đằng sau hệ thống ấy, phần mềm sử dụng phân tích thống kê và phân tích dự đoán để xác định các mẫu trong dữ liệu người dùng và sử dụng các mẫu đó để điền vào News Feed. Nếu thành viên không còn dừng lại để đọc, thích hoặc bình luận trên các bài đăng của bạn bè, dữ liệu mới đó sẽ được bao gồm trong tập dữ liệu và News Feed sẽ điều chỉnh tương ứng. Không chỉ riêng facebook, ta có thể bắt gặp những tính năng tương tự đó qua các mạng xã hội khác như google, instagram,....

+ Nhận diện hình ảnh

Nhận dạng hình ảnh là một trong những ví dụ về máy học và trí tuệ nhân tạo phổ biến nhất. Về cơ bản, nó là một cách tiếp cận để xác định và phát hiện các đặc trưng của một đối tượng trong hình ảnh kỹ thuật số. Hơn nữa, kỹ thuật này có thể được sử dụng để phân tích sâu hơn, chẳng hạn như nhận dạng mẫu, nhận diện hình khuôn, nhận dạng khuôn mặt, nhận dạng ký tự quang học và nhiều hơn nữa,...

+ Xử lý ngôn ngữ tự nhiên (NLP):

Nếu bạn nghĩ Google Translate tự bản thân nó đã thực sự là một cuốn từ điển hoàn hảo thì bạn nên suy nghĩ lại. Nó thực ra được tạo nên từ một bộ các thuật toán Học máy dựa trên việc cập nhật theo thời gian các đầu vào từ phía người dùng, như từ mới và cú pháp mới. Siri, Alexa, Cortana, và gần đây nhất là Google Assistant, tất cả đều dựa trên việc *Xử lý ngôn ngữ tự nhiên* để Nhận dạng giọng nói, Tổng hợp giọng nói (dựa trên một phần), cho phép chúng hiểu hay phát âm ra những từ chưa từng gặp phải trước đó.

## **1.4. Khái niệm phân lớp dự báo:**

Phân lớp dự báo: là đưa ra số liệu thống kê để dự báo trước cho một đối tượng, giá trị trong tương lai.

Có hai nhóm phương pháp dự báo chính là nhóm định tính và nhóm định lượng

- Các phương pháp dự báo định tính: dựa vào phán đoán chủ quan và trực giác để đưa ra dự báo thay cho vì dựa vào các số liệu quá khứ. Phương pháp dự báo định tính hữu ích cho việc dự báo toàn cục và một số trường hợp mà số liệu quá khứ không hữu ích cho dự báo.

Các kỹ thuật dự báo định lượng: dựa vào việc phân tích số liệu quá khứ để đưa ra dự báo. Giả định của phương pháp này là các nhân tố từng tác động lên biến được dự báo trong quá khứ vẫn tiếp tục ảnh hưởng đến biến này trong tương lai. Vậy dựa vào diễn biến dữ liệu trong quá khứ ta có thể dự báo cho tương lai. Các phương pháp dự báo định lượng lại được chia thành hai nhóm chính: dự báo định lượng mang tính nhân quả và dự báo định lượng mang tính thống kê.

- Các phương pháp dự báo định lượng mang tính nhân quả

Đại diện của nhóm phương pháp này là phân tích hồi quy. Mô hình dự báo có hai nhóm biến số: các biến số được dự báo được gọi là biến độc lập, các biến số dùng để dự báo được gọi là biến phụ thuộc.Các phương pháp dự báo định lượng mang tính thống kê.

Để tiến hành dự báo nhu cầu sản phẩm theo phương pháp định lượng cần thực hiện 8 bước sau:

- Xác định mục tiêu dự báo

- Lựa chọn những sản phẩm cần dự báo

- Xác định độ dài thời gian dự báo

- Chọn mô hình dự báo

- Thu thập các dữ liệu cần thiết

- Phê chuẩn mô hình dự báo

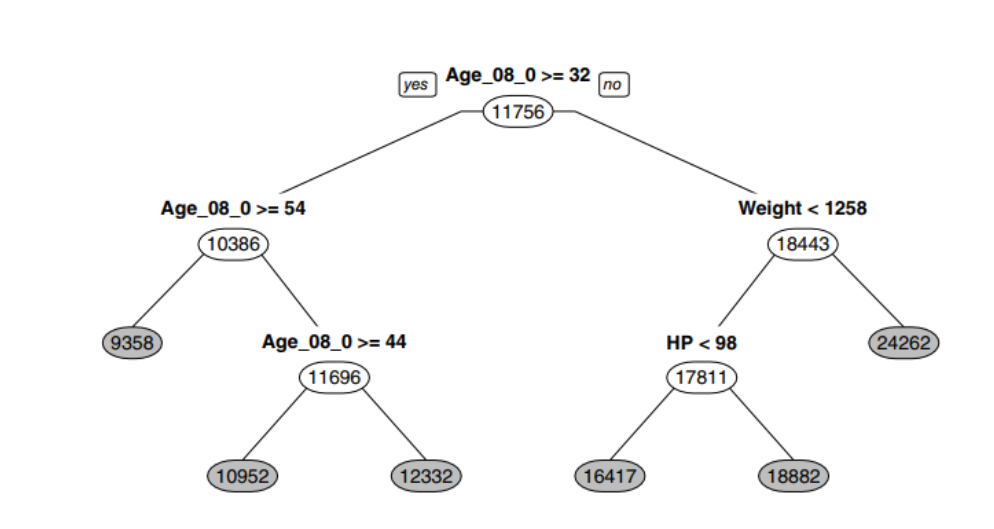
- Tiến hành dự báo

- Áp dụng kết quả dự báo

Nhóm các phương pháp dự báo mang tính thống kê chỉ quan tâm đến quy luật biến thiên của biến cần dự báo trong quá khứ để đưa ra dự báo. Biến thiên của một biến số kinh tế được chia thành các thành phần: xu hướng, chu kỳ, thời vụ và ngẫu nhiên.

Regression Tree, cách sử dụng cây quyết định để dự báo giá trị của biến mục tiêu(biến định lượng) dựa trên thuộc tính, đặc điểm những biến đầu vào còn lại của các đối tượng dữ liệu. Mô hình cây quyết định sử dụng nguyên lý hoạt động của các phương trình hồi quy , tìm ra mối quan hệ giữa các biến độc lập với biến phụ thuộc, giữa biến đầu vào và biến dự báo.

Ví dụ:



Hình 1.4.Mô hình Regression Tree dự báo giá của một chiếc xe Toyota

Trên là mô hình Regression Tree dự báo giá của một chiếc xe Toyota dựa trên 3 biến: Tuổi đời của xe(Age), Trọng lượng (Weight) và mã lực (Horse power-HP). Ví dụ với chiếc xe có độ tuổi là 55, mã lực bằng 100 thì có thể bán với giá 9358$. Ta có thể thấy mối quan hệ giữa độ tuổi với giá trị của chiếc xe, tức độ tuổi cao hay thấp sẽ có tác động tới chiếc xe , giống như mã lực cao hay thấp. Tuy nhiên tác động của độ tuổi và mã lực đến giá xe , tác động nào lớn hơn? Dựa trên mô hình cây này ta khó có thể xác định được. Để diễn giải kết quả phân tích hồi quy hoặc mô tả theo cách định lượng hóa, thì mô hình cây quyết định không được phổ biến và áp dụng. Mà thay vào đó là sử dụng phương trình hay mô hình hồi quy bao gồm công thức định lượng mối quan hệ giữa các biến .

## **1.5. Mô hình Regression:**

### **1.5.1 Giới thiệu mô hình Regression:**

-"Regression" 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 "Regression" 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...

-Mục tiêu của giải thuật hồi quy tuyến tính là dự đoán giá trị của một hoặc nhiều *biến mục tiêu liên tục* (*continuous target variable*) y dựa trên một véc tơ đầu vào X

-Ví dụ: dự đoán giá nhà ở Hà Nội dựa vào thông tin về diện tích, vị trí, năm xây dựng của ngôi nhà thì t ở đây là giá nhà và X=(x1,x2,x3) với x1 là diện tích ,x2 là vị trí và x3 là năm xây dựng

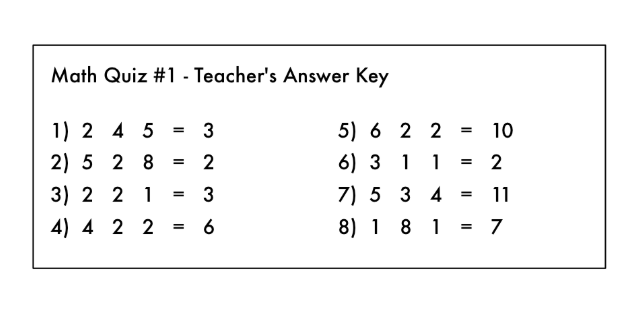
Ví dụ về bài toán dự đoán giá của căn nhà ta có thể thấy rằng :

+Diện tích căn nhà càng lớn=>giá càng cao.

+Số lượng phòng lớn=>Giá càng cao

+Càng xa trung tâm =>Giá càng giảm .

Đây là ví dụ về dự đoán phép toán;



Hình 1.5. Ví dụ về phép toán

Trong vòng khoảng nửa phút ,bạn sẽ ra cách tìm dấu điền vào chỗ trống để có được 1 kết quả đúng .

Và Machine-Learning cũng vậy ,bạn đưa cho máy tính con số và kết quả ,việc của máy tính là tìm ra mối liên hệ giữa các con số để đồng nhất kết quả giữa vế trái và vế phải của phép tính .

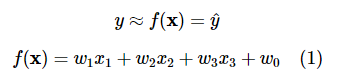
Về cơ bản thì ta sẽ có 1 cặp huấn luyện chứa các cặp (x(i),y(i)) ,tương ứng với nhiệm vụ của ta là phải tìm ra giá trị ứng với 1 đầu vào X mới .

Để làm được điều này ta cần phải tìm được quan hệ giữa x và y để từ đó đưa ra dự đoán .



Hình 1.6. Quan hệ y=3+4x

Một hàm số đơn giản nhất có thể mô tả mối quan hệ giữa giá nhà và 3 đại lượng đầu vào là:



Trong đó ,w1 ,w2 ,w3 ,w0 là các hằng số, w0 còn được gọi là bias.Mối quan hệ 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,w3,,w0} {w1,w2,w3,w0} chính vì vậy được gọi là bài toán Linear Regression.

### **1.5.2 Mô hình Linear Regression:**

Mô hình đơn giản nhất là mô hình kết hợp tuyến tính của các biến đầu vào:



Trong đó  là véc tơ biến đầu vào và  là véctơ trọng số tương ứng. Thường *θ* được gọi là tham số của mô hình. Giá trị của tham số sẽ được ước lượng bằng cách sử dụng các cặp giá trị của tập huấn luyện

Thực ra mô hình tuyến tính là chỉ cần ở mức tuyến tính giữa tham số *θ* và y là đủ. Và mình cho rằng tên gọi tuyến tính là xuất phát giữa *θ* và y

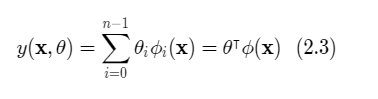
chứ không phải giữa x và y .Nói cách khác, ta có thể kết hợp các x một cách phi tuyến trước khi hợp với *θ* để được y . Một cách đơn giản là sử dụng hàm phi tuyến cho x như sau:



θ được gọi là tham số của mô hình. Giá trị của tham số sẽ được ước lượng bằng cách sử dụng các cặp giá trị (x(i),y(i)) của tập huấn luyện.

*θo*​ được gọi là độ lệch (*bias*) nhằm cắt đi mức độ chênh lệch giữa mô hình và thực tế.

Viết lại công thức trên như sau:



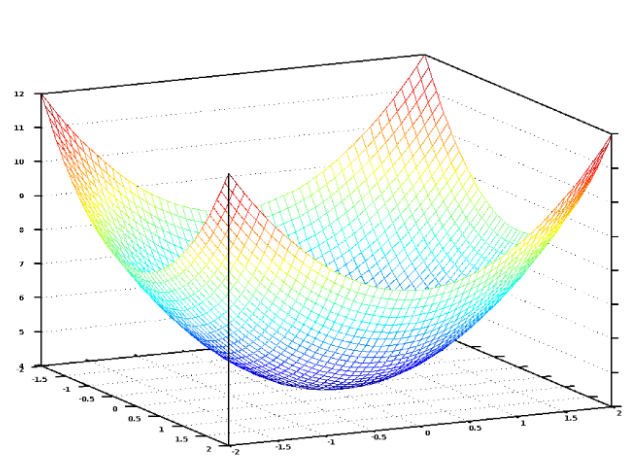
Như quy ước thì tất cả các véc tơ nếu không nói gì thì ta ngầm định với nhau rằng nó là véc tơ cột nên ta có được cách viết nhân ma trận như trên.

### **1.5.3 Sai số của dự đoán**

Chúng ta mong muốn rằng sự sai khác e giữa giá trị thực y và giá trị dự đoán y^ (đọc là *y hat* trong tiếng Anh) là nhỏ nhất. Nói cách khác, chúng ta muốn giá trị sau đây càng nhỏ càng tốt:



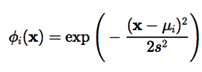
### **1.5.4. Xác định Basic Function:**



- Giữ nguyên đầu vào có ý là không thay đổi giá trị đầu vào ϕi (x)=x

- Chuẩn hoá về đoạn [min, max]

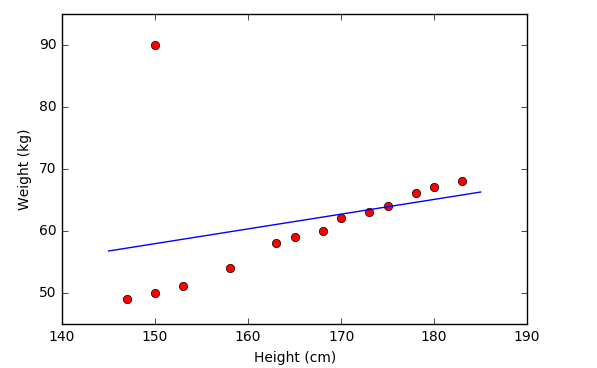
- Sử dụng đa thức bậc cao:

- Sử dụng hàm Gaussian: 

- Sử dụng hàm Sigmod: 

### **1.5.5 Hạn chế của mô hình Linear Regression:**

Hạn chế đầu tiên của Linear Regression là nó rất nhạy cảm với nhiễu (sensitive to noise). Trong ví dụ về mối quan hệ giữa chiều cao và cân nặng bên trên, nếu có chỉ một cặp dữ liệu *nhiễu* (150 cm, 90kg) thì kết quả sẽ sai khác đi rất nhiều. Xem hình dưới đây:



Hình 1. 7. Đồ thị thể hiện giá trị hiện tại và giá trị tương lai của dữ liệu.

Vì vậy, trước khi thực hiện Linear Regression, các nhiễu (*outlier*) cần phải được loại bỏ. Bước này được gọi là tiền xử lý (pre-processing).

Hạn chế thứ hai của Linear Regression là nó không biểu diễn được các mô hình phức tạp. Mặc dù trong phần trên, chúng ta thấy rằng phương pháp này có thể được áp dụng nếu quan hệ giữa *outcome* và *input* không nhất thiết phải là tuyến tính, nhưng mối quan hệ này vẫn đơn giản nhiều so với các mô hình thực tế.

## **1.6. Ứng dụng trong Giáo dục:**

Dự báo là phán đoán những sự kiện sẽ xảy ra trong tương lai trên cơ sở phân tích khoa học các dữ liệu của quá khứ và hiện tại nhờ một số mô hình toán học.

Dự báo Giáo dục là việc đưa ra các dự báo những sự kiện Giáo dục sẽ xảy ra trong tương lai dựa trên cơ sở phân tích khoa học các số liệu kinh tế của quá khứ và hiện tại. Chẳng hạn, nhà quản lý dựa trên cơ sở các số liệu về điểm thi đầu vào của kỳ trước và kỳ này để đưa ra dự báo về điểm tuyển sinh của các trường học trong tương lai.

Do đó, trong hoạt động Giáo dục, dự báo đem lại ý nghĩa rất lớn. Nó là cơ sở để lập các kế hoạch học tập tạo tính hiệu quả và sức cạnh tranh cho các sĩ tử trong tương lai. Dự báo mang tính khoa học và đòi hỏi cả một nghệ thuật dựa trên cơ sở phân tích khoa học các số liệu thu thập được. Bởi lẽ cũng dựa vào các số liệu thời gian nhưng lấy số lượng là bao nhiêu, mức độ ở những thời gian cuối nhiều hay ít sẽ khiến cho mô hình dự đoán phản ánh đầy đủ hay không đầy đủ những thay đổi của các nhân tố mới đối với sự biến động của hiện tượng. Do vậy mà dự báo vừa mang tính chủ quan vừa mang tính khách quan. Dự báo muốn chính xác thì càng cần phải loại trừ tính chủ quan của người dự báo.

# **CHƯƠNG 2. THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ**

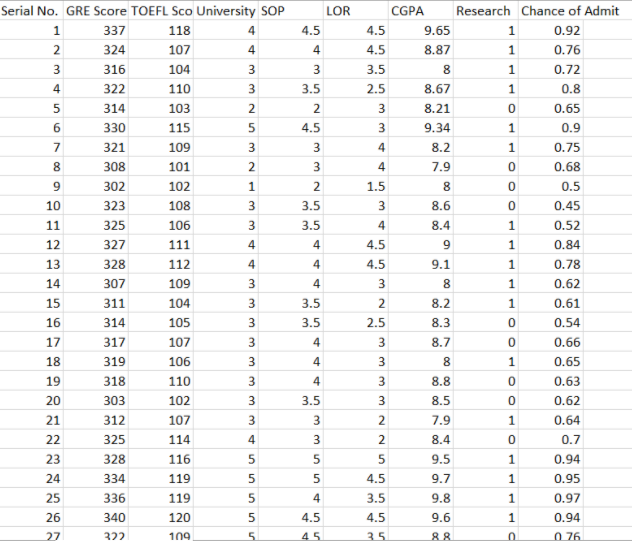
## **2.1 Bài toán dự đoán điều kiện nhập học với mô hình Regression**

### ***2.1.1 Phát biểu bài toán:***

- Bài toán dự nhập học đưa ra tất cả các thông tin cần thiết để tuyển sinh của một trường học để từ đó làm căn cứ dự đoán điểm kiểm tra bài test ở thời điểm hiện tại cho đến thời điểm sau đó .

- Bài toán sẽ lấy dữ liệu trên Kaggle để phân tích ,huấn luyện để dự đoán điều kiện nhập học vào một trường học.

### ***2.1.2 Chuẩn bị dữ liệu:***



Hình 2.1.Dữ liệu điều kiện nhập học với mô hình Regression

*Chú thích*

* GRE Score: Điểm GRE (trên 340)
* TOEFL Score: Điểm TOEFL (trên 120)
* University Rating: Xếp hạng Đại học (trên 5)
* SOP: Tuyên bố Mục đích (trên 5)
* LOR : Thư giới thiệu Điểm mạnh (trên 5)
* CGPA: Điểm trung bình đại học (trên 10)
* Research: Kinh nghiệm nghiên cứu (0 hoặc 1)
* Chance of Admit : Cơ hội nhập học (từ 0 đến 1)

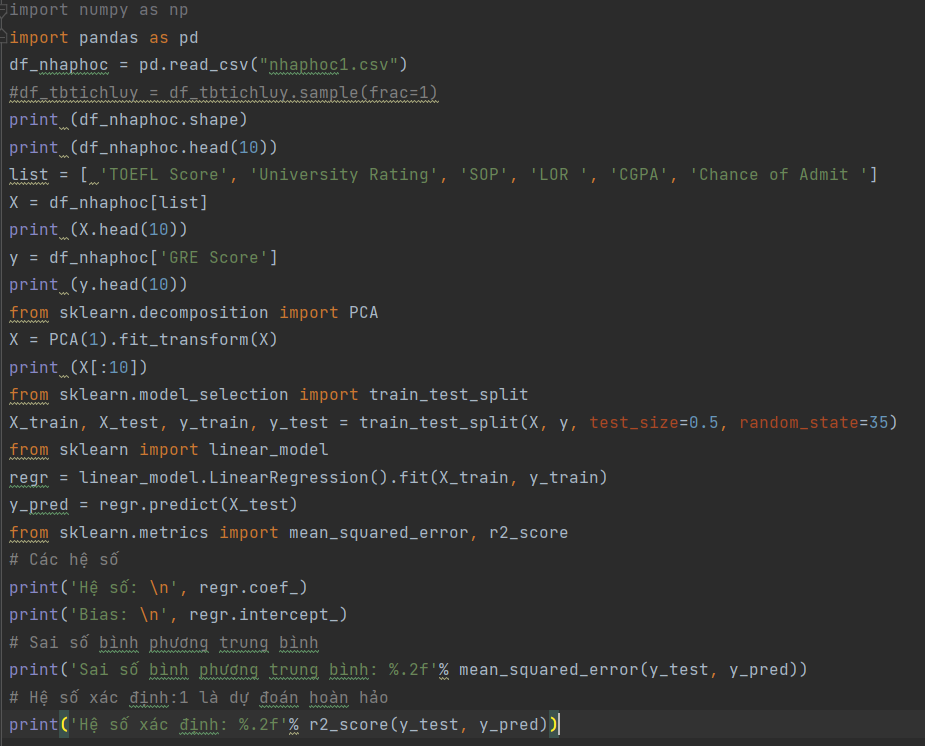
### ***2.1.3 Xử lý dữ liệu:***

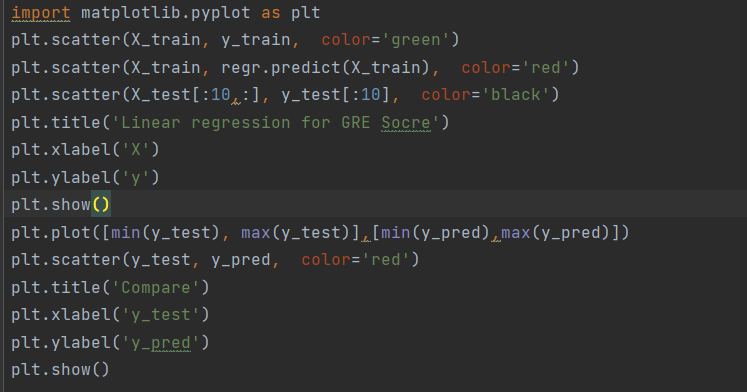
Ở đây chúng em sử dụng Linear regression và dữ liệu lấy từ Excel(file csv) ,subline text,python ,command prompt để hỗ trợ quá trình training .Về cơ bản thì python đã được tích hợp rất nhiều các thuật toán khác nhau, dễ dàng sử dụng, và giúp giảm thời gian xây dựng các hệ thống deep learning. Đồng thời kết hợp với pandas và numpy để phân tích, và xử lý cấu trúc data, và matplotlib dùng để về đồ thị.

Việc vẽ đồ thị rất quan trọng đối với các bài toán thuộc dạng Time Series Analysis như thế này. Vì dĩ nhiên việc đoán trước không thể trả về kết quả chính xác 100% được, Kết quả sẽ là tương đối và có thể có một chút sai số không đáng kể. Vì thế việc vẽ đồ thị sẽ giúp bạn dễ dàng so sánh giữa kết quả dự đoán và thực tế.

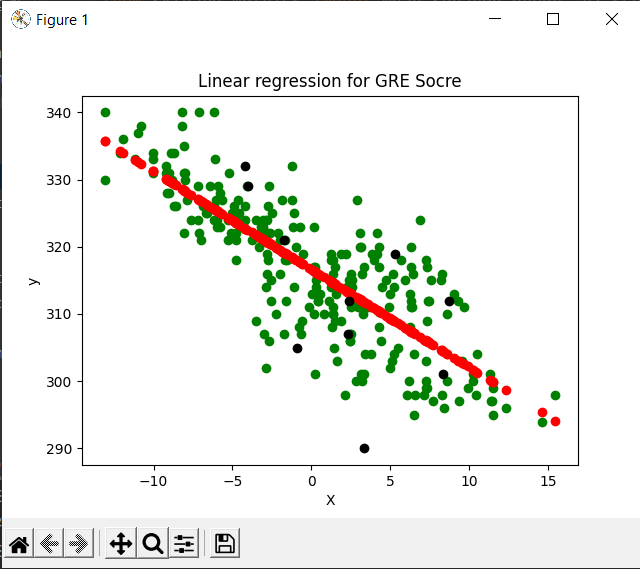
### ***2.1.4.Code và chạy dữ liệu***

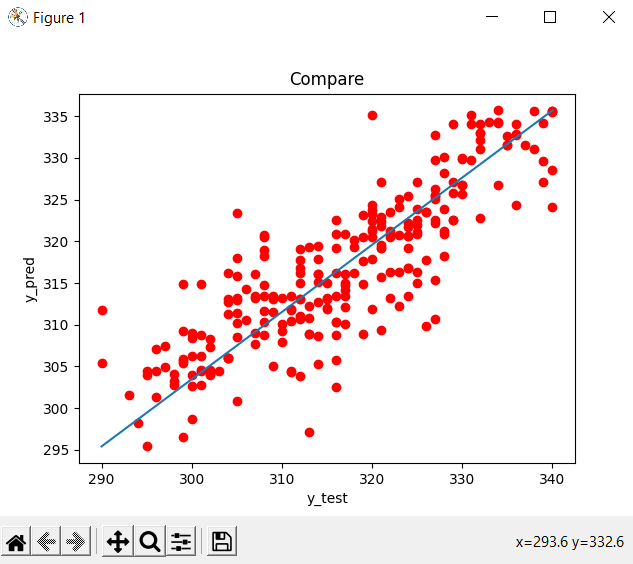
#### *2.1.4.1.Code*

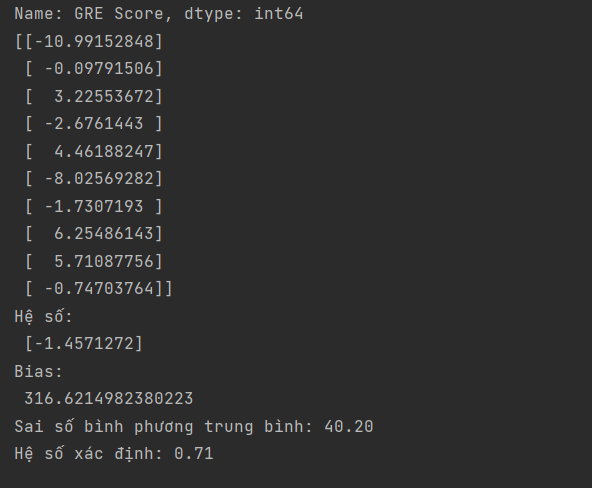




#### *2.1.4.2.Chạy dữ liệu*







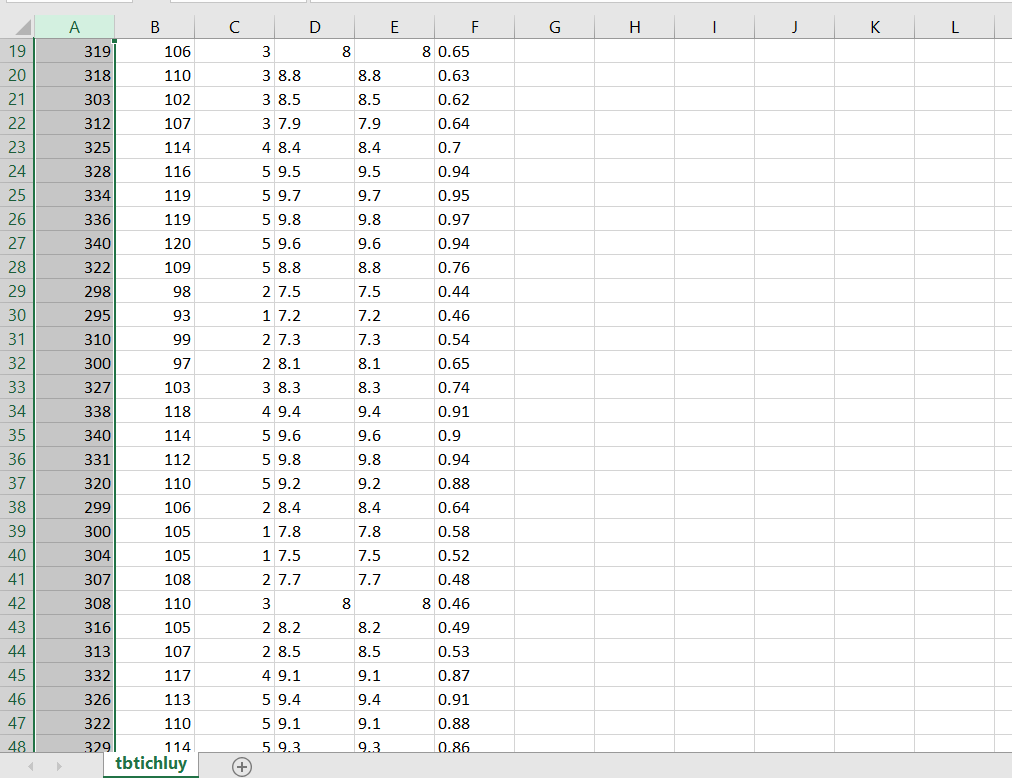
## **2.2 Bài toán trung bình tích lũy với mô hình Regression:**

### ***2.2.1 Phát biểu bài toán:***

- Bài toán dự báo trung bình tích lũy đưa ra tất cả các thông tin học sinh của một trường học để từ đó làm căn cứ dự đoán trung bình tích lũy ở thời điểm hiện tại cho đến thời điểm sau đó .

-.Bài toán sẽ lấy dữ liệu trên Kaggle để phân tích ,huấn luyện để dự đoán trung bình tích lũy của học sinh.

### ***2.2.2 Chuẩn bị dữ liệu:***



Hình 2.2.dữ liệu trung bình tích lũy với mô hình Regression

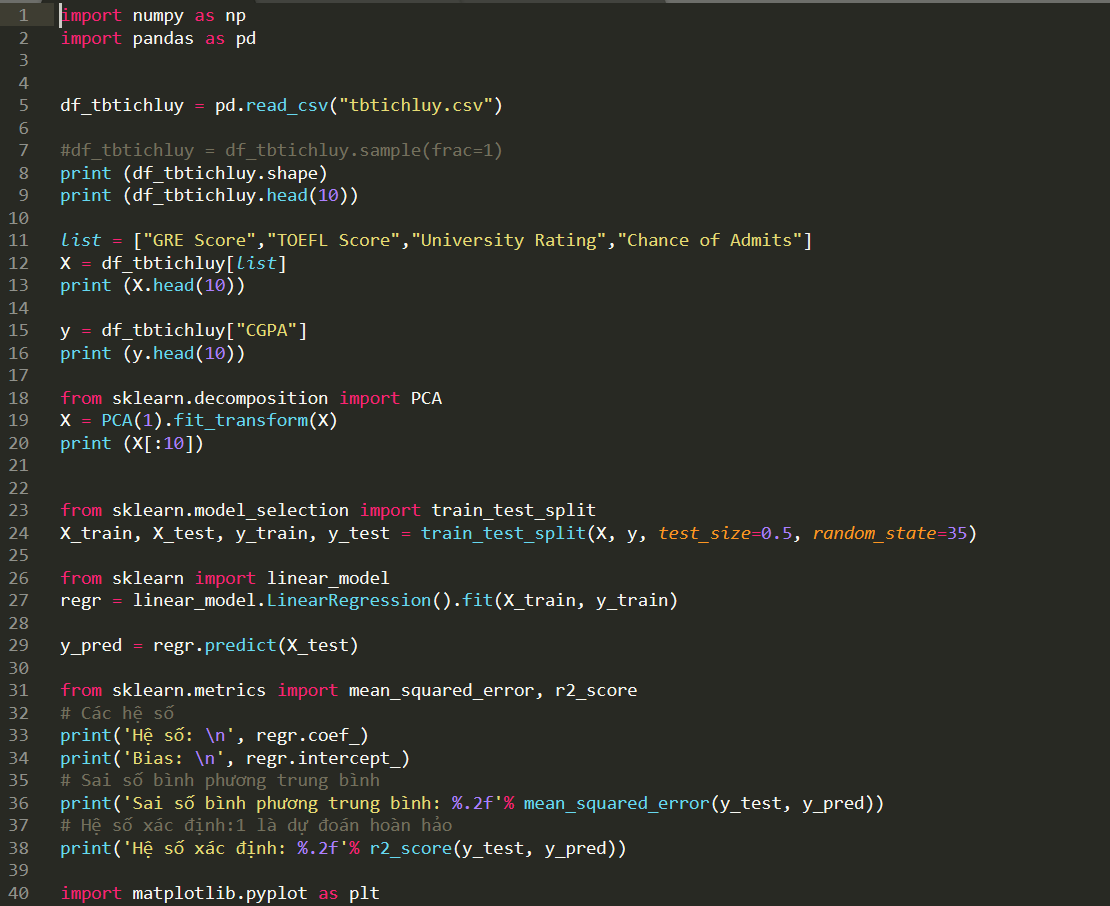
### ***2.2.3 Xử lý dữ liệu:***

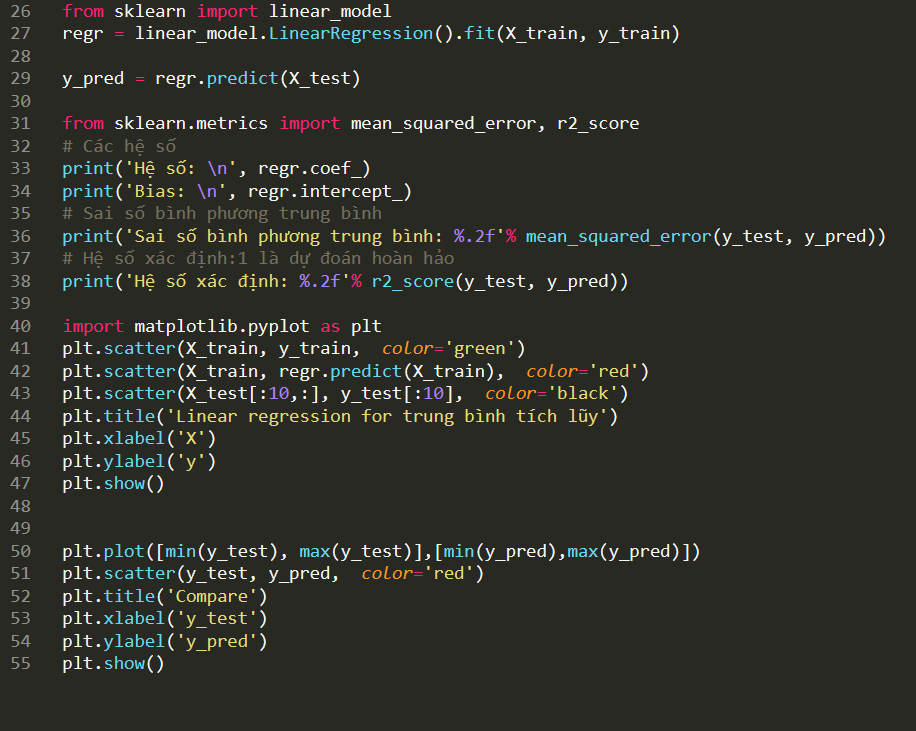
Ở đây chúng em sử dụng Linear regression và dữ liệu lấy từ Excel(file csv) ,subline text,python ,command prompt để hỗ trợ quá trình training .Về cơ bản thì python đã được tích hợp rất nhiều các thuật toán khác nhau, dễ dàng sử dụng, và giúp giảm thời gian xây dựng các hệ thống deep learning. Đồng thời kết hợp với pandas và numpy để phân tích, và xử lý cấu trúc data, và matplotlib dùng để về đồ thị.

Việc vẽ đồ thị rất quan trọng đối với các bài toán thuộc dạng Time Series Analysis như thế này. Vì dĩ nhiên việc đoán trước không thể trả về kết quả chính xác 100% được, Kết quả sẽ là tương đối và có thể có một chút sai số không đáng kể. Vì thế việc vẽ đồ thị sẽ giúp bạn dễ dàng so sánh giữa kết quả dự đoán và thực tế.

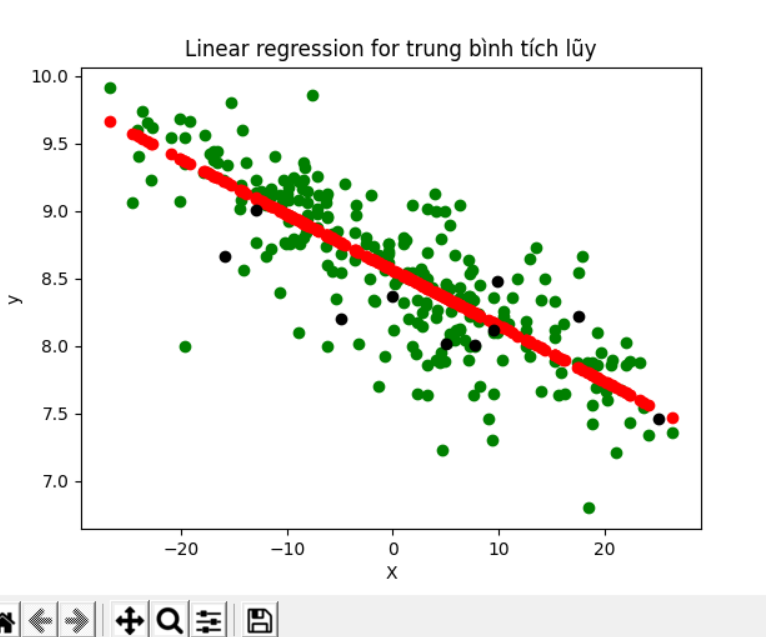
### ***2.2.4.Code và chạy dữ liệu:***

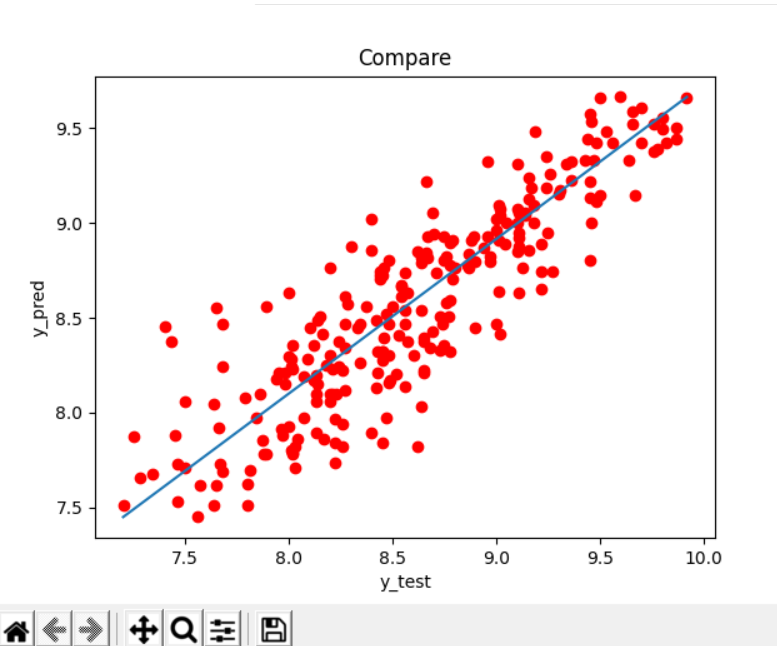
#### *2.2.4.1.Code:*

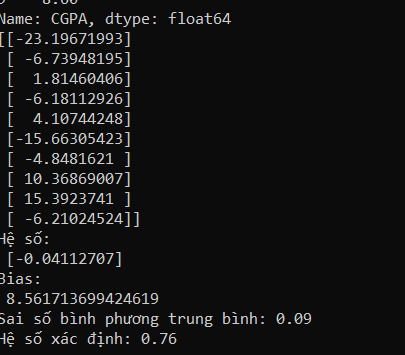




#### *2.2.4.2.Chạy dữ liệu:*







# **KẾT LUẬN**

Qua quá trình tìm hiểu, phân tích nghiên cứu “Kỹ thuật quy hồi tuyến tính và ứng dụng trong Giáo dục” của việc ứng dụng mô hình, chúng em đã xây dựng hoàn thành đề tài: “Kỹ thuật quy hồi tuyến tính và ứng dụng trong Giáo dục” Với việc chính là dự báo một cách chính xác những vấn đề liên quan đến Giáo dục như điểm tuyển sinh các trường Đại học,….

Báo cáo đã thu được một số kết quả như sau:

-Mô hình đã ứng dụng được một cách chính xác các yêu cầu đã đề ra, bám sát thực tế.

-Mô hình đảm bảo lưu trữ chính xác những thông tin cần thiết và chi tiết về thông tin. Và có thể cập nhật thường xuyên và dễ dàng chỉnh sửa thông tin khi cần thiết.

-Xử lý thông tin một cách chính xác, đầy đủ, kiểm tra thông tin khi cần thiết.

-Đảm bảo được tính bảo mật cũng như an toàn về dữ liệu.

-Giao diện thuận tiện dễ sử dụng.

Tuy nhiên mô hình còn một số mặt hạn chế:

-Mô hình chưa có tính chuyên nghiệp cao.

-Chưa giải quyết được trọn vẹn một số vấn đề nảy sinh .

Trên đây là toàn bộ đề tài của chúng em trong quá trình xây dựng đề tài “Kỹ thuật quy hồi tuyến tính và ứng dụng trong Giáo dục”. Tuy đã hoàn thành đề tài nhưng do kinh nghiệm còn thiếu, trong quá trình khảo sát, phân tích thiết kế còn gặp nhiều khó khăn nên mô hình chưa được toàn diện. Ngoài ra, đề tài khi đem áp dụng vào thực tế tuy đã đáp ứng được một số yêu cầu được đặt ra, song còn xuất hiện một số vấn đề nảy sinh, đòi hỏi mô hình cần phải được nâng cấp. Vì vậy nếu có cơ hội, chúng em hi vọng sẽ được sửa chữa và nâng cấp mô hình sao cho phù hợp và đáp ứng đầy đủ với yêu cầu thực tế một cách tốt nhất.

Em xin chân thành cảm ơn!

# 

# **TÀI LIỆU THAM KHẢO**

[1] Trang web: https://en.wikipedia.org/wiki/Linear\_regression

[2] Bài giảng Học máy – thầy Đào Nam Anh

[3] Link data: <https://www.kaggle.com/datasets>

[4] Trang web: https://scikit- learn.org/stable/modules/generated /sklearn.linear\_model.LinearRegression.html