****

**T****RƯỜNG ĐẠI HỌC THỦY LỢI**

KHOA CÔNG NGHỆ THÔNG TIN

-----🙞🙜🕮🙞🙜-----



**BÁO CÁO THỰC HÀNH BÀI TẬP LỚN**

**MÔN HỌC: HỌC MÁY**

**Đề tài:** DỰ ĐOÁN DOANH THU FILM

**Giảng viên hướng dẫn: PGS.TS Nguyễn Hữu Quỳnh**

Nhóm sinh viên thực hiện: Nhóm 2-60Th4

Đinh Tất Hiển -1851061360

Phạm Văn Hà- 1851061760

Phạm Hoàng Minh-1851061718

Hà Nội, 11/ 2021

# MỤC LỤC

[MỤC LỤC 1](#_Toc69409941)

[DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT 1](#_Toc69409942)

[MỞ ĐẦU 2](#_Toc69409943)

[TỔNG QUAN 3](#_Toc69409944)

[PHẦN 1: GIỚI THIỆU VỀ HỌC MÁY 3](#_Toc69409945)

[1.Khái niệm cơ bản 3](#_Toc69409946)

[1.1: Lịch sử của Học Máy 4](#_Toc69409947)

[1.2: Các phân nhóm Học máy 4](#_Toc69409948)

[1.3: Các phương pháp học máy 5](#_Toc69409949)

[PHẦN 2: TRÌNH BÀY BÀI TOÁN 5](#_Toc69409950)

[Bài toán đặt ra 5](#_Toc69409951)

[Input 5](#_Toc69409952)

[Output 6](#_Toc69409953)

[PHẦN 3 : PHƯƠNG PHÁP HỌC MÁY SỬ DỤNG 6](#_Toc69409954)

[THỰC NGHIỆM 7](#_Toc69409955)

[PHẦN 1: MÔ TẢ TẬP VÍ DỤ HUẤN LUYỆN VÀ TẬP TEST 7](#_Toc69409956)

[PHẦN 2: MÔ TẢ CHƯƠNG TRÌNH DEMO 10](#_Toc69409957)

[PHẦN 3: MÔ TẢ KẾT QUẢ DỰ ĐOÁN 10](#_Toc69409958)

[KẾT LUẬN 13](#_Toc69409959)

[TÀI LIỆU THAM KHẢO 14](#_Toc69409960)

# DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

* AI : Artificial Intelligence
* PLA: Perceptron Learning Algorithm
* ID3: Iterative Dichotomiser 3
* PCA: Principal Component Analysis
* SGD: Stochastic Gradient Descent
* SVM: Support Vector Machine

# MỞ ĐẦU

Những năm gần đây, có thể mọi người sẽ nghe thấy rất nhiều cụm từ "cách mạng công nghiệp lần thứ 4" hay "cách mạng công nghiệp 4.0"(1 - động cơ hơi nước, 2 - năng lượng điện, 3 - công nghệ thông tin). Cuộc cách mạng mới này tập trung chủ yếu vào sản xuất thông minh dựa trên sự phát triển đột phá của cách ngành nghề công nghệ thông tin, công nghệ sinh học, công nghệ nano, ... Và đặc biệt, mỗi một cuộc cách mạng công nghệ đều sẽ mang đến một bước ngoặt lớn với cách thức chúng ta sản xuất, lao động, hãy nhìn lại thế giới xung quanh chúng ta đang thay đổi từng ngày như thế nào: chúng ta có các sản phẩm trí tuệ nhân tạo mô phỏng được các hoạt động y hệt con người, thậm chí là giỏi hơn con người, những ứng dụng của trí tuệ nhân tạo còn đang len lỏi vào mọi lĩnh vực trong đời sống mà có thể chúng ta không nhận ra. Xe tự hành của Google và Tesla, hệ thống tự tag khuôn mặt trong ảnh của Facebook, trợ lý ảo Siri của Apple, hệ thống gợi ý sản phẩm của Amazon, hệ thống gợi ý phim của Netflix, máy chơi cờ vây

AlphaGo của Google DeepMind, … và đó cũng chỉ là một vài trong số vô vàn những ứng dụng của AI/Machine Learning. Thực sự, thế giới đang đi những bước dài mỗi ngày, góp một phần không nhỏ trong đó chính là công nghệ thông tin, và cụ thể hơn, một trong các công nghệ góp phần vào bước phát triển của công nghệ thông tin, chính là machine learning.

Machine Learning là một tập con của AI. Theo định nghĩa của Wikipedia, *Machine learning is the subfield of computer science that “gives computers the ability to learn without being explicitly programmed”*. Nói đơn giản, Machine Learning là một lĩnh vực nhỏ của Khoa Học Máy Tính, nó có khả năng tự học hỏi dựa trên dữ liệu đưa vào mà không cần phải được lập trình cụ thể. Machine learning Algorithm được chia làm 2 loại chính là: Supervised Learning (Học có giám sát) và Unsupervised Learning (Học không giám sát). Chính vì vậy nên trong phạm vi nghiên cứu tìm hiểu, nhóm chúng em quyết định sẽ phân tích và làm rõ về thuật toán Linear Regression (Hồi quy tuyến tính) thuộc phương thức Supervised Learning

(Học có giám sát) thông qua việc phân tích và dự đoán chi phí bảo hiểm y tế dựa trên dữ liệu đã được thống kê, tổng hợp.

# TỔNG QUAN

## PHẦN 1: GIỚI THIỆU VỀ HỌC MÁY

### 1.Khái niệm cơ bản

Học máy (Machine learning) là gì ?

**Machine Learning** là một tập con của AI có khả năng tự học hỏi dựa trên dữ liệu đưa vào mà không cần phải được lập trình cụ thể với 2 kỹ thuật chủ yếu là **Học có giám sát** (Supervised Learning) và **Học không giám sát** (Unsupervised Learning)

### 1.1: Lịch sử của Học Máy

Lịch sử của Học máy là một quá trình dài dựa trên các thuật toán từ cả trăm năm trước, cũng như những phát triển về công nghệ máy tính trong 7 thập kỷ qua. Những năm gần đây, AI - Artificial Intelligence (Trí Tuệ Nhân Tạo), và cụ thể hơn là Machine Learning (Học Máy hoặc Máy Học) nổi lên như một bằng chứng của cuộc cách mạng công nghiệp lần thứ tư. Hiện nay, các thuật toán Machine Learning cho phép máy tính có thể giao tiếp với con người, xe hơi tự lái của Google và Tesla, hệ thống tự tag khuôn mặt trong ảnh của Facebook, trợ lý ảo Siri của Apple, hệ thống gợi ý sản phẩm của Amazon, hệ thống gợi ý phim của Netflix, máy chơi cờ vây AlphaGo của Google DeepMind, …, chỉ là một vài trong vô vàn những ứng dụng của AI/Machine Learning , Machine Learning đang tác động một cách sâu sắc đến mọi ngành công nghiệp và các công việc liên quan đến chúng, đó là lý do tại sao mọi người cần phải có ít nhất một số kiến thức về machine learning và nó đã phát triển như thế nào.

### 1.2: Các phân nhóm Học máy

Dựa theo phương thức học ,các thuật toán Machine Learning thường được chia làm 4 nhóm:

* + - * Supervised learning (học có giám sát)
      * Unsupervised learning (học không giám sát)
      * Semi-supervised lerning (học bán giám sát)
      * Reinforcement learning (học củng cố).

Có một số cách phân nhóm không có Semi-supervised learning hoặc Reinforcement learning. Những năm gần đây, khi mà khả năng tính toán của các máy tính được nâng lên một tầm cao mới và lượng dữ liệu khổng lồ được thu thập bởi các hãng công nghệ lớn, Machine Learning đã tiến thêm một bước dài và một lĩnh vực mới được ra đời gọi là Deep Learning (Học Sâu)

Dựa theo chức năng của các thuật toán , ta có một số thuật toán tiêu biểu:

* + Regression Algorithms (hồi quy tuyến tính)
  + Classification Algorithms (phân loại)
  + Clustering Algorithms (phân cụm K-means)
  + Decision tree (cây quyết định)

### 1.3: Các phương pháp học máy

- **Supervised learning** là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (input, outcome) đã biết từ trước. Cặp dữ liệu này còn được gọi là (data, label), tức (dữ liệu, nhãn). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning.

Một cách toán học, Supervised learning là khi chúng ra có một tập hợp biến đầu vào X={x1,x2,…,xN}X={x1,x2,…,xN} và một tập hợp nhãn tương ứng Y={y1,y2,…,yN}Y={y1,y2,…,yN}, trong đó xi,yixi,yi là các vector. Các cặp dữ liệu biết trước (xi,yi)∈X×Y(xi,yi)∈X×Y được gọi là tập training data (dữ liệu huấn luyện). Từ tập training data này, chúng ta cần tạo ra một hàm số ánh xạ mỗi phần tử từ tập XX sang một phần tử (xấp xỉ) tương ứng của tập YY:

yi≈f(xi),  ∀i=1,2,…,Nyi≈f(xi),  ∀i=1,2,…,NMục đích là xấp xỉ hàm số ff thật tốt để khi có một dữ liệu xx mới, chúng ta có thể tính được nhãn tương ứng của nó y=f(x)y=f(x).

#### **Unsupervised Learning (Học không giám sát)**

Trong thuật toán này, chúng ta không biết được outcome hay nhãn mà chỉ có dữ liệu đầu vào. Thuật toán unsupervised learning sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ như phân nhóm (clustering) hoặc giảm số chiều của dữ liệu (dimension reduction) để thuận tiện trong việc lưu trữ và tính toán.

Một cách toán học, Unsupervised learning là khi chúng ta chỉ có dữ liệu vào XX mà không biết nhãn YY tương ứng.

Những thuật toán loại này được gọi là Unsupervised learning vì không giống như Supervised learning, chúng ta không biết câu trả lời chính xác cho mỗi dữ liệu đầu vào. Giống như khi ta học, không có thầy cô giáo nào chỉ cho ta biết đó là chữ A hay chữ B. Cụm không giám sát được đặt tên theo nghĩa này.

#### **Semi-Supervised Learning (Học bán giám sát)**

Các bài toán khi chúng ta có một lượng lớn dữ liệu XX nhưng chỉ một phần trong chúng được gán nhãn được gọi là Semi-Supervised Learning. Những bài toán thuộc nhóm này nằm giữa hai nhóm được nêu bên trên.

Một ví dụ điển hình của nhóm này là chỉ có một phần ảnh hoặc văn bản được gán nhãn (ví dụ bức ảnh về người, động vật hoặc các văn bản khoa học, chính trị) và phần lớn các bức ảnh/văn bản khác chưa được gán nhãn được thu thập từ internet. Thực tế cho thấy rất nhiều các bài toán Machine Learning thuộc vào nhóm này vì việc thu thập dữ liệu có nhãn tốn rất nhiều thời gian và có chi phí cao. Rất nhiều loại dữ liệu thậm chí cần phải có chuyên gia mới gán nhãn được (ảnh y học chẳng hạn). Ngược lại, dữ liệu chưa có nhãn có thể được thu thập với chi phí thấp từ internet.

#### **Reinforcement Learning (Học Củng Cố)**

Reinforcement learning là các bài toán giúp cho một hệ thống tự động xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao nhất (maximizing the performance). Hiện tại, Reinforcement learning chủ yếu được áp dụng vào Lý Thuyết Trò Chơi (Game Theory), các thuật toán cần xác định nước đi tiếp theo để đạt được điểm số cao nhất.

## PHẦN 2: TRÌNH BÀY BÀI TOÁN

### Bài toán đặt ra

Từ dữ liệu đã được thu thập: vốn đầu tư, độ phổ biến, số xuất chiếu, điểm, phân tích và đưa ra dự đoán về doanh thu của một bộ phim.

### Input

* Tập dữ liệu được thu thập có kích thước 898x4
* Gồm 4 cột: budget, popularity, runtime, score

o budget: vốn đầu tư

o popularity: độ phổ biến

* + runtime: số xuất chiếu

o score: điểm

### Output

- Dự đoán doanh thu của một bộ phim chiếu tại có thể đạt được.

## PHẦN 3 : PHƯƠNG PHÁP HỌC MÁY SỬ DỤNG

* Phương pháp học máy : Hồi quy tuyến tính (Linear Regression), Bagging

- Lí do sử dụng :

* + Dữ liệu đầu vào là các tập dữ liệu và tập nhãn đã biết.
  + Mục tiêu là dự đoán giá trị đầu ra từ dữ liệu đã cho và kỳ vọng giá trị dự đoán và giá trị thực (outcome) có sai số chênh lệch thấp.

o Nhãn (label) không được chia thành các nhóm mà là các giá trị cụ thể

* Hồi quy tuyến tính

o Phương trình hồi quy tuyến tính có dạng : y ≈ f(x) = 𝑦̂

f(x) = w1x1 + w2x2 + w3x3 + w0 (1)

y là giá trị thực của outcome (dựa trên số liệu thống kê chúng ta có trong tập training data), trong khi 𝑦̂ là giá trị mà mô hình Linear Regression dự đoán được. Nhìn chung, y và 𝑦̂ là hai giá trị khác nhau do có sai số mô hình, tuy nhiên, chúng ta mong muốn rằng sự khác nhau này rất nhỏ.

* + Nếu đặt w = [w0, w1, w2, w3]𝑇 là vector (cột) hệ số cần phải tối ưu và x̅ = [1, 𝑥1, 𝑥2, 𝑥3]𝑇 là vector (hàng) dữ liệu đầu vào mở rộng thì phương trình hồi quy tuyến tính được viết lại là: y ≈ x̅w = 𝑦̂
  + Hàm số ℒ(w) được gọi là hàm mất mát (loss function) của bài toán Linear Regression.

ℒ(w) =  ‖𝑦 − x̅w‖22

* + - Cách phổ biến nhất để tìm nghiệm cho một bài toán tối ưu là giải phương trình đạo hàm (gradient) bằng 0! Nhưng chỉ trường hợp

Đạo hàm theo w của hàm mất mát là:

𝜕𝐿(𝑤) = 𝑋𝑇(𝑋𝑤 − 𝑦)

𝜕𝑤

Phương trình đạo hàm bằng 0 tương đương:

𝑋𝑇𝑋𝑤 = 𝑋𝑇𝑦 (\*)

Đặt 𝐴 ≜ 𝑋𝑇𝑋 và 𝑏 ≜ 𝑋𝑇𝑦

Nếu ma trận vuông A khả nghịch thì phương trình (\*) có nghiệm duy nhất

𝑤 = 𝐴−1𝑏

* Bagging
  + - Xây dựng một lượng lớn các model (thường là cùng loại) trên những subsamples khác nhau từ tập training dataset (random sample trong 1 dataset để tạo 1 dataset mới).
    - Những model này sẽ được train độc lập và song song với nhau nhưng đầu ra của chúng sẽ được tính là trung bình cộng (hoặc bỏ phiếu) để cho ra kết quả cuối cùng.
* Sử dụng trong bài tập lớn
  + Xác định đây là hồi quy tuyến tính đa biến có phương trình dạng tổng quát.
  + Đặt các hệ số w thành một vector cột và các giá trị x đầu vào là một mà trận vector hàng.
  + Chia tập dữ liệu khảo sát thành 2 phần training và test với tỉ lệ cho tập training là 80% và tập test là 20%.
  + Áp dụng mô hình hồi quy tuyến tính để tìm ra nghiệm hay chính là các giá trị w. Từ đó hoàn thiện phương trình hồi quy tuyến tính.
  + Kiểm tra với tập test sau đó đánh giá sai số bằng phương pháp R-square. Nó giúp hiểu được mức độ biến đổi trong y có thể được giải thích bằng sự phụ thuộc vào x bằng cách sử dụng mô hình hồi quy cụ thể.
  + Nhập dữ liệu bất kỳ để dự đoán doanh thu của một bộ phim.

# THỰC NGHIỆM

## PHẦN 1: MÔ TẢ TẬP VÍ DỤ HUẤN LUYỆN VÀ TẬP TEST

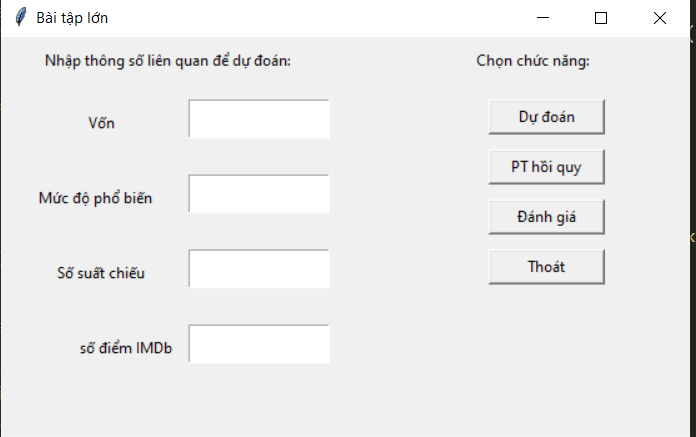
* Tập dữ liệu đầu vào có kích thước 898x4.
* Gồm 4 cột: budget, popularity, runtime, score, revenue
  + Dữ liệu ban đầu
* Tập ví dụ huấn luyện lần lượt là X\_train và Y\_train. X\_train có kích thước 602x4 và Y\_train là 602x1.

* Tập test lần lượt là X\_test và Y\_test. X\_train có kích thước 297x4 và Y\_train là 297x1.

## 

## PHẦN 2: MÔ TẢ CHƯƠNG TRÌNH DEMO

a. Cài đặt các thư viện cần thiết

* Thư viện numpy
* Numpy (Numeric Python): là một thư viện toán học phổ biến và mạnh mẽ của Python. Cho phép làm việc hiệu quả với ma trận và mảng, đặc biệt là dữ liệu ma trận và mảng lớn với tốc độ xử lý nhanh hơn nhiều lần khi chỉ sử dụng “core Python” đơn thuần.
* Numpy đặc biệt hữu ích khi thực hiện các hàm liên quan tới Đại Số Tuyến Tính.
* Cài đặt thư viện: Command Prompt và gõ lệnh: pip install numpy
* Thư viện matplotlib
* Matploblib là một thư viện trực quan hoá dữ liệu phổ biến trong Python. Nó có thể vẽ được nhiều loại đồ thị khác nhau, và rất hữu ích khi làm việc cùng với NumPy.
* Cài đặt thư viện: Command Prompt và gõ lệnh: pip install matplotlib
* Thư viện pandas
* Thư viện pandas trong python là một thư viện mã nguồn mở, hỗ trợ đắc lực trong thao tác dữ liệu.
* Đây cũng là bộ công cụ phân tích và xử lý dữ liệu mạnh mẽ của ngôn ngữ lập trình python.
* Là một công cụ cho phép đọc/ ghi dữ liệu giữa bộ nhớ và nhiều định dạng file: csv, text, excel, sql database,…
* Cài đặt thư viện: Command Prompt và gõ lệnh: pip install pandas
* Giao diện chương trình
* 

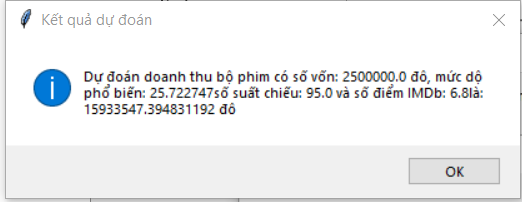
* Chương trình gồm 4 ô nhập dữ liệu đầu vào chính (vốn đầu tư, độ phổ biến, sổ xuất chiếu, điểm) một nút Dự đoán
* Nhập giá trị của các biến để tiến hành dự đoán giá trị của doanh thu 1 bộ phim.
* Giá trị charges dự đoán được xử lý bằng thuật toán Hồi quy tuyến tính và **Bagging** dựa trên những giá trị được nhập vào.
* Thực hiện kiểm tra dự đoán bằng cách nhập thông tin trong tập dữ liệu test sau đó đối chiếu kết quả dự đoán và kết quả chính xác trong tập test.
* Tiến hành nhập thông tin đầu vào ngẫu nhiên và bắt đầu dự đoán.

## PHẦN 3: MÔ TẢ KẾT QUẢ DỰ ĐOÁN

* Kết quả dự đoán dựa trên thông tin đầu vào lấy từ tập dữ liệu test được so sánh với kết quả thực trong tập test.
* Tập dữ liệu dự đoán
* Chọn dữ liệu nhập vào chương trình



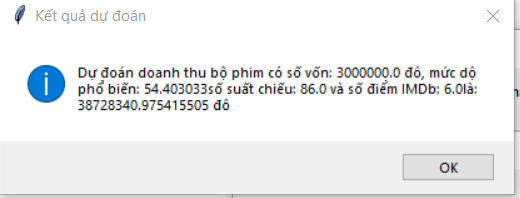
Kết quả



* Chọn dữ liệu nhập vào chương trình



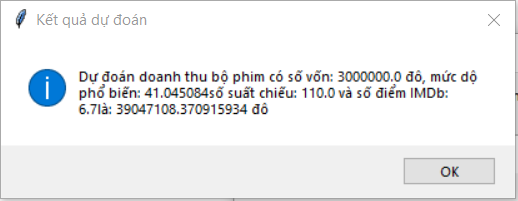
Kết quả:



* Chọn dữ liệu nhập vào chương trình

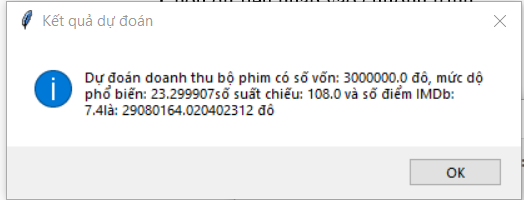


Kết quả



* Chọn dữ liệu nhập vào chương trình

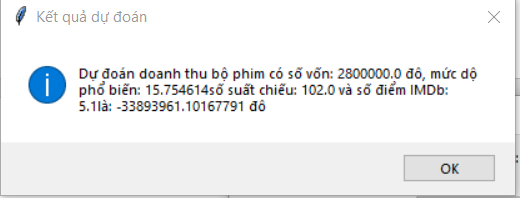


Kết quả

* Chọn dữ liệu nhập vào chương trình



Kết quả



# KẾT LUẬN

Cũng giống như các thuật toán khác, Linear Regression cũng có những hạn chế nhất định:

o Linear Regression rất nhạy cảm với nhiễu (sensitive to noise). o Linear Regression không biểu diễn đươc các mô hình phức tạp.

Sau khi hoàn thành bài tập lớn này, chúng em đã thu được những kiến thức về môn học máy cũng như áp dụng kiến thức đó vào phân tích các bài toán thực tế.

Qua đây chúng em đã:

+ Hiểu được tổng quan về học máy cũng như một số thuật toán học máy cơ bản.

+ Có thể ứng dụng Hồi quy tuyến tính, Bagging vào các bộ dữ liệu khác nhau sau này.

Tuy nhiên:

+ Tất cả các kiến thức nắm được ở mức lý thuyết hoặc thực hành trên máy tính riêng, chưa được thực hành trên thực tế nhiều để hiểu sau rộng về thuật toán cũng như các kỹ thuật khác của học máy.

Qua bài tập này chúng em đã học được:

+ Bài tập đã khái quát được một số vấn đề về bài toán phân lớp bao gồm phương pháp phân lớp dữ liệu và các thuật toán học máy áp dụng vào bài toán phân lớp, trong đó chú trọng nghiên cứu tới phương pháp học có giám sát được sử dụng rất phổ biến hiện nay.

+ Nghiên cứu và trình bày cơ sở của lý thuyết của phương pháp học máy.

Thông qua việc phân tích các dữ liệu đầu vào, ta có thể dự đoán doanh thu của một bộ phim với độ chính xác tương đối.

# TÀI LIỆU THAM KHẢO

1. Các slide bài giảng của PGS Nguyễn Hữu Quỳnh
2. Lý thuyết về [Linear Regression](https://machinelearningcoban.com/2016/12/28/linearregression/) trên trang Machine Learning cơ bản
3. <https://machinelearningcoban.com/2016/12/28/linearregression/>
4. <https://www.geeksforgeeks.org/ml-rainfall-prediction-using-linear-regression/>
5. <https://github.com/bangoc123/learn-machine-learning-in-two-months/tree/master/models/linear-regression>
6. <https://codelearn.io/sharing/ve-bieu-do-voi-thu-vien-matplotlib-p1>
7. https://codelearn.io/sharing/ve-bieu-do-voi-thu-vien-matplotlib-p2
8. https://viblo.asia/p/gioi-thieu-ve-numpy-mot-thu-vien-chu-yeu-phuc-vu-cho-khoa-hoc-may-tinh-cua-python-maGK7kz9Kj2
9. <https://allaravel.com/blog/ve-do-thi-trong-python-voi-thu-vien-matplotlib>
10. https://blog.vietnamlab.vn/tim-hieu-thuat-toan-linear-regression/
11. https://aischool.vn/2019/11/04/ham-mat-mat-trong-machine-learning
12. http://quantri.vn/dict/details/9182-phan-tram-sai-so-tuyet-doi-trung-binh-mape