# Phần 1: Tổng quan

## 1.1 Giới thiệu về học máy

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 (1 - động cơ hơi nước, 2 - năng lượng điện, 3 - công nghệ thông tin). Trí Tuệ Nhân Tạo 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, …

**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) Trong đó:

**Học tập có giám sát** là nơi bạn có các biến đầu vào (X) và biến đầu ra (Y) và bạn sử dụng thuật toán để tìm hiểu hàm ánh xạ từ đầu vào đến đầu ra. Mục đích là để xây dựng hàm ánh xạ một cách tốt nhất có thể để khi bạn có dữ liệu đầu vào mới (X) và bạn có thể dự đoán các biến đầu ra (Y) cho dữ liệu 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(ví dụ như phân cụm,…)

Trong kỷ nguyên số hóa, công nghệ là yếu tố thiết yếu giúp doanh nghiệp thực hiện những cải tiến, nâng cao vị thế cạnh tranh trên thị trường. Nhu cầu nguồn nhân lực chất lượng cao thông hiểu công nghệ  nổi bật liên quan đến AI như Big Data, Data Science, Machine Learning ngày càng nâng cao. Thị trường nhân lực ngành Công nghệ đã phát triển mạnh trong nửa đầu năm nay với số lượng đăng tuyển và ứng tuyển đều tăng cao. Các công ty sẽ đối mặt với thách thức giữ chân nhân tài trong lĩnh vực này bằng cách đề xuất mức lương, thưởng chế độ thăng tiến tương xứng.

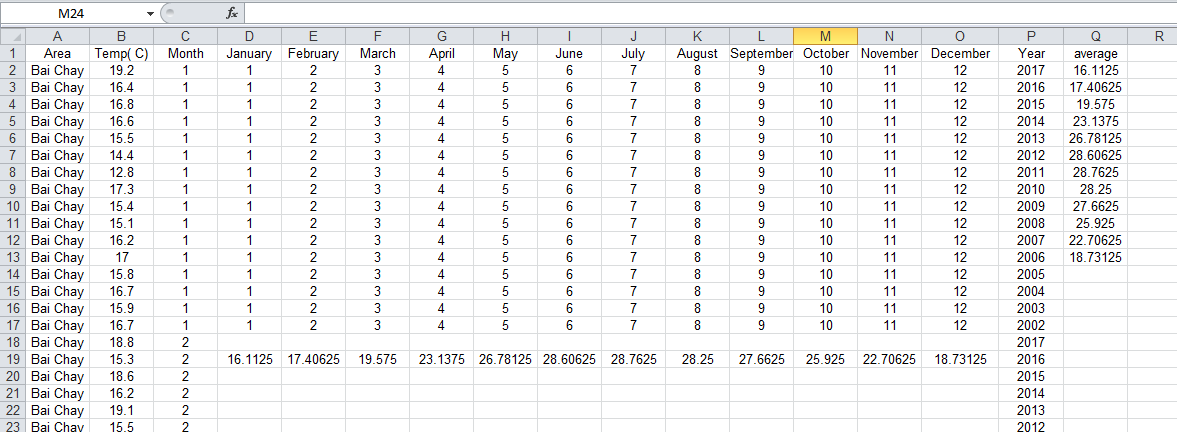
## 1.2 Giới thiệu về thuật toán Linear Regression

Project của nhóm em lần này sẽ sử dụng thuật toán Linear Regression (Hồi Quy Tuyến Tính) thuộc nhóm Supervised learning ( Học có giám sát ). Hồi quy tuyến tính (Linear Regression) là một phương pháp thống kê để thiết lập mối quan hệ giữa một biến phụ thuộc và một nhóm tập hợp các biến độc lập rất đơn giản nhưng đã được chứng minh được tính hữu ích cho một số lượng lớn các tình huống:

Như ta đã biết, chịu tác động của hiệu ứng nhà kính toàn cầu nên nhiệt độ đều có xu hướng biến động và tăng sau mỗi năm. Vì vậy nhóm em sẽ ứng dụng thuật toán Hồi quy tuyến tính nhằm biểu diễn nhiệt độ trung bình sau mỗi năm, ngoài biểu diễn dữ liệu, mô hình còn đưa ra dự đoán nhiệt độ cho thời gian sắp tới nhờ vào việc liên tục cập nhật dữ liệu nhiệt độ, thời tiết qua các năm. Việc này làm cho mô hình của ta càng đáng tin cậy hơn.

## 1.3 Trình bày bài toán

**Input**: Bảng csv data về nhiệt độ trong 15 năm có các thông tin về trung bình nhiệt dộ của các tháng. Bảng được xử lí một số thông số thông qua lệnh Excel( xử lí trung bình nhiệt độ tháng)



Bảng dữ liệu đầu vào có dạng trên.

**Ouput**: Biểu diễn và Dự đoán thời tiết tháng/ của năm sau đó

## 1.4 Trình bày phương pháp học máy được sư dụng trong bài tập lớn

Phương pháp Hồi quy tuyến tính (Linear Regression) thuộc nhóm Supervised learning ( Học có giám sát ).

Trong việc phân tích dữ liệu, thuật ngữ Regression( Hồi quy) để nói lên mối quan hệ giữa 1 biến phụ thuộc và một nhóm tập hợp các biến độc lập. Ở bài toán này ta thiết lập mối quan hệ giữu Nhiệt độ và Năm, Nhiệt độ và các tháng. Biểu diễn đơn giản như sau:

Temp = W0 + W1\*Year  
Và Temp = W0 + W1\*Month

Với việc dự đoán biến phụ thuộc (Temp) dựa trên giá trị biến độc lập(Year/Month). Mô hình còn có thể được sử dụng trong 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...)

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. Vì vậy nó rất nạy cảm với dữ liệu đột biến/ dữ liệu nhiễu.

# Phần 2: Thực nghiệm

## 2.1 Mô tả tập ví dụ huấn luyện và tập test

### a. Tập ví dụ huấn luyện

\* Đối với mô hình Linear Regression:

Tập ví dụ huẩn luyện là tập 192 bản ghi ứng với mỗi 16 năm của 12 tháng tại địa điểm là Bãi Cháy

\*Đối với mô hình Logistic Regression:

Tập ví dụ huấn luyện là tập nhiệt độ của 8 năm 2002: 2009 của 12 tháng

### b. Tập test

\*Đối với mô hình Linear Regression:

Tập test là tập 2 bản ghi dữ liệu nhiệt độ trung bình tháng của 2 năm ngay sau đó là 2018 và 2019 tại địa điểm là Bãi Cháy ta thu nhận được từ việc lấy dữ liệu trên trang nhiệt độ Bãi Cháy.

Với nhiệt độ trung bình của 12 tháng tại Bãi Cháy năm 2018 là:

17,2 16.5 21.2 23.3 28.7 29.4 28.6 28 27.5 25.5 23.4 19.0

Nhiệt độ trung bình của 12 tháng tại Bãi Cháy năm 2019 là:

17.4 16.8 21 23.6 29 29.7 28.5 28.4 27.8 26 23.7 19.7

\*Đối với mô hình Logistic Regression:

Tập test là tập dữ liệu 8 năm sau đó 2010-2017 của 12 tháng

## 2.2 Ý tưởng giải quyết

Bới nhiệt độ trung bình của các tháng được ghi nhận không có sự biến động quá lớn( dữ liệu gây nhiễu) nên lẽ dĩ nhiên ta sẽ tìm ra được mối quan hệ giữa nhiệt độ và các tháng trong năm thông qua 1 đường hồi quy tuyến tính với mức tin tưởng ở một ngưỡng nhất định nào đó.

Và để thẩm định độ chính xác của kết quả dự đoán của mô hình, ta sử dụng đến thuật toán Logistic Regression sau khi đã xử lí một phần dữ liệu:

Trước hết ta sử dụng dữ liệu của 8 năm 2002-2009 để lấy ra đường hồi quy dự đoán nhiệt độ cho 8 năm kế tiếp(1)

Sau đó ta so sánh kết quả của nhiệt độ dự đoán ở (1) với kết quả nhiệt độ thực tế có được trong bảng csv.

Nếu 2 kết quả chênh lệch không quá 1.5 độ C thì sẽ lưu vào một mảng giá trị là 1(đúng)

Nếu ngược lại thì sẽ lưu vào mảng giá trị là 0(sai)

Vậy là ta đã có dữ liệu đầu vào cho mô hình Logistic Regression với

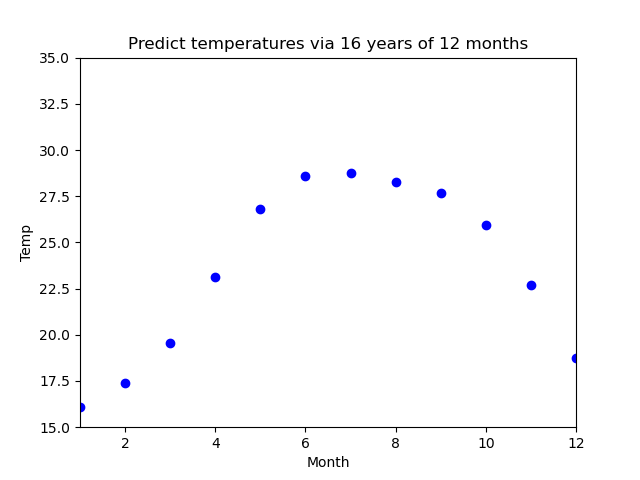
X: dữ liệu nhiệt độ thực tế của 8 năm

L: một mảng 0, 1 thể hiện mức độ chính xác của dữ liệu

## 2.3 Mô tả chương trình demo

Chương trình xử lí, biểu diễn các dữ liệu nhiệt độ thu thập được trong 16 năm từ 2002-2017 và đưa ra được đường mô hình dự đoán nhiệt độ cho các tháng/ năm tiếp theo

* **Biểu diễn nhiệt độ trung bình của cả năm:**



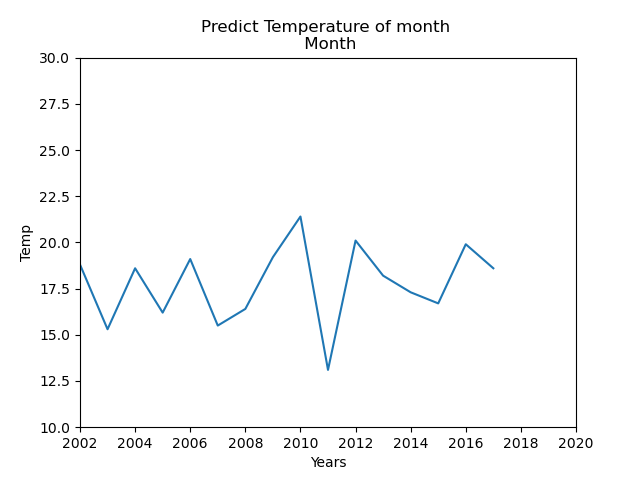
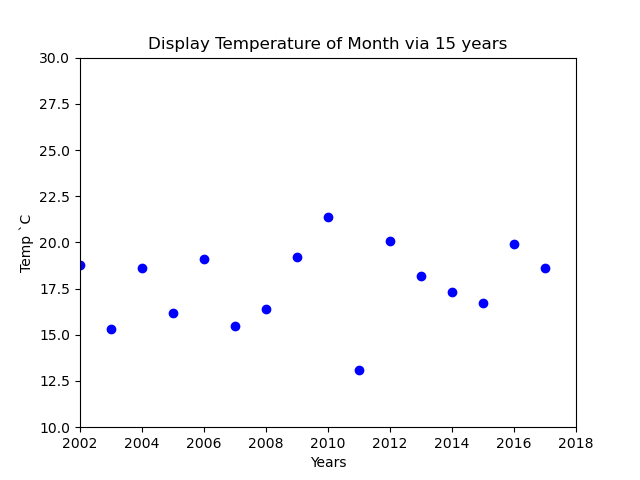
* **Biểu diễn nhiệt độ nhiệt độ của các tháng:**

Tháng 1:



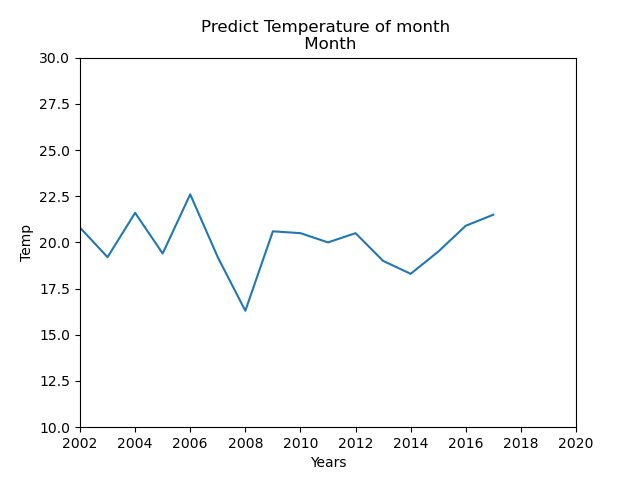
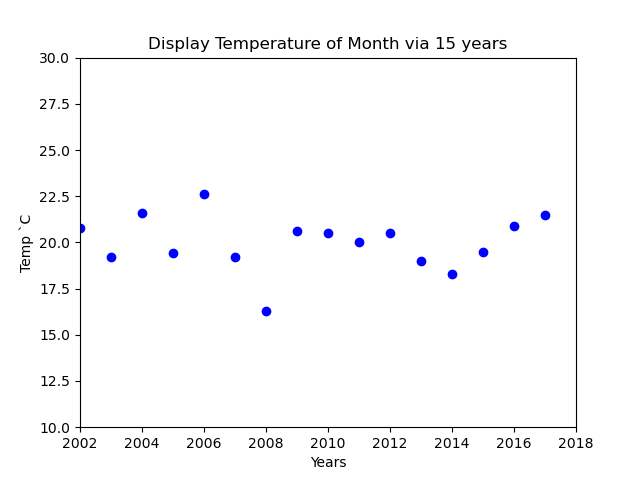
h

Tháng 2:



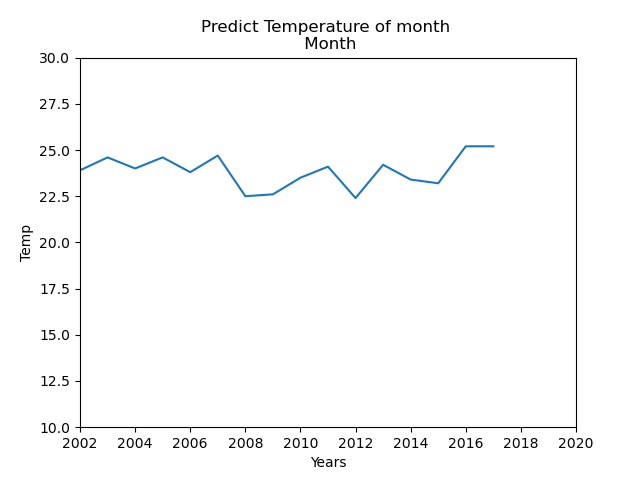
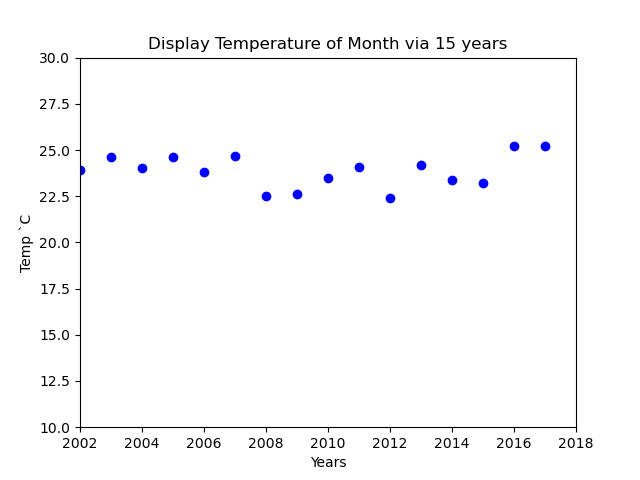
Tháng 3:

Tháng 3:

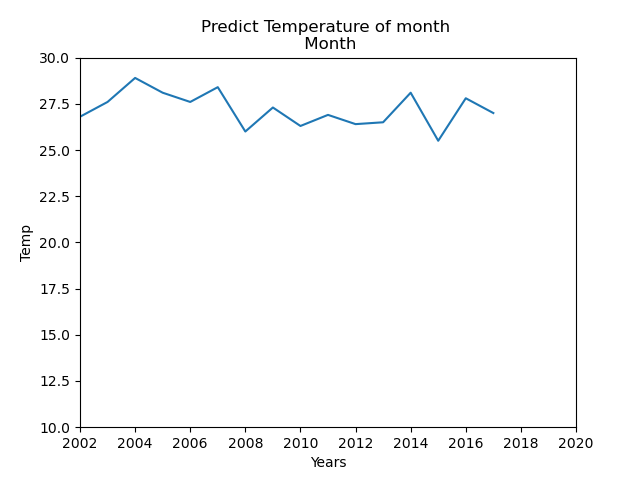
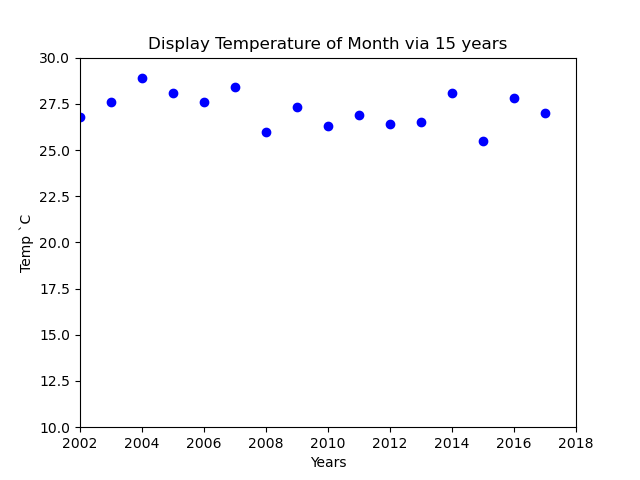


h

Tháng 4

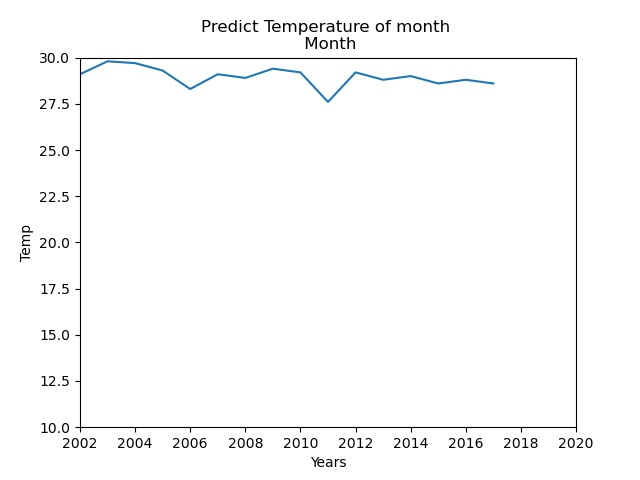
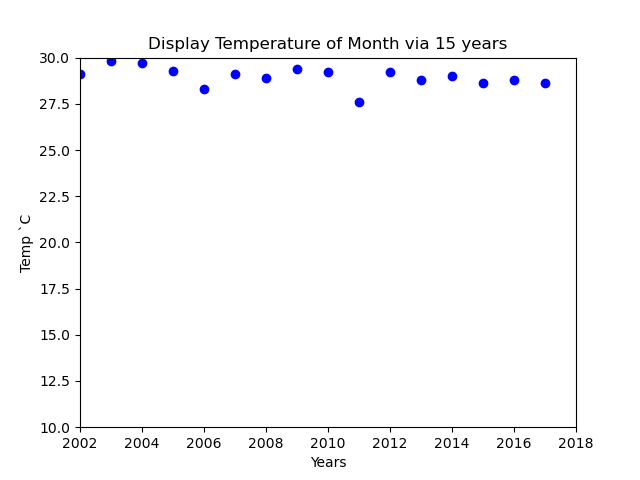


Tháng 5:

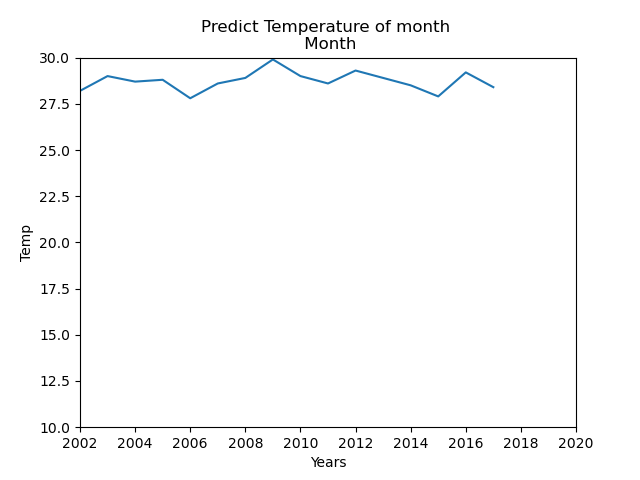
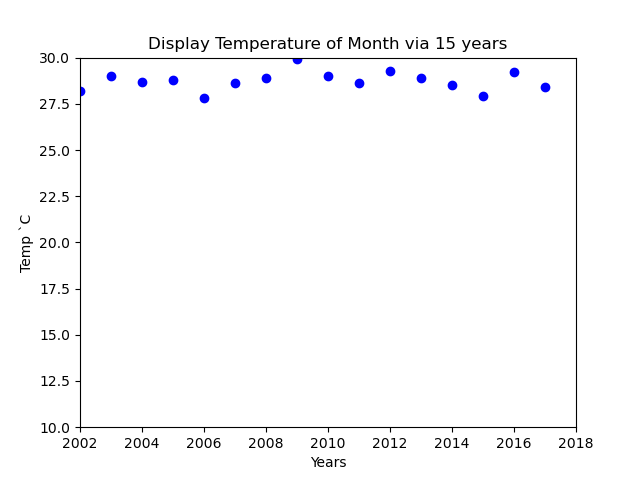


h

Tháng 6:

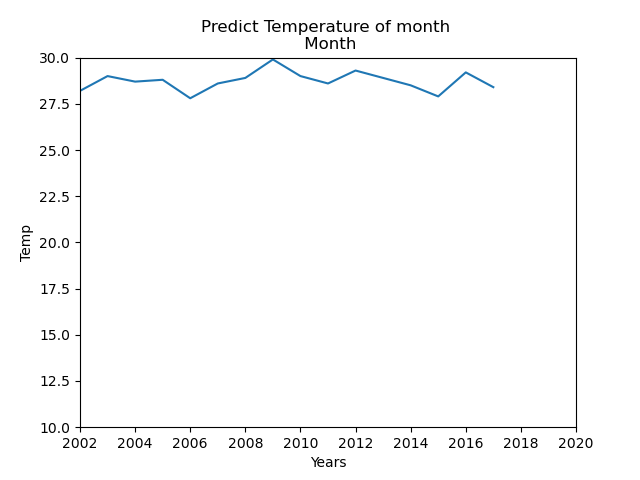
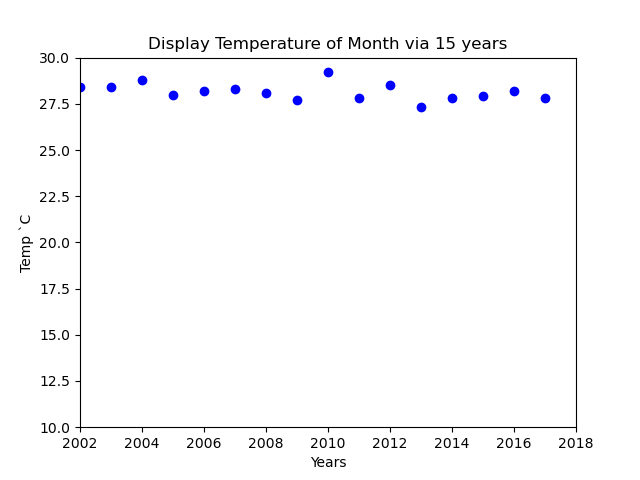


Tháng 7:

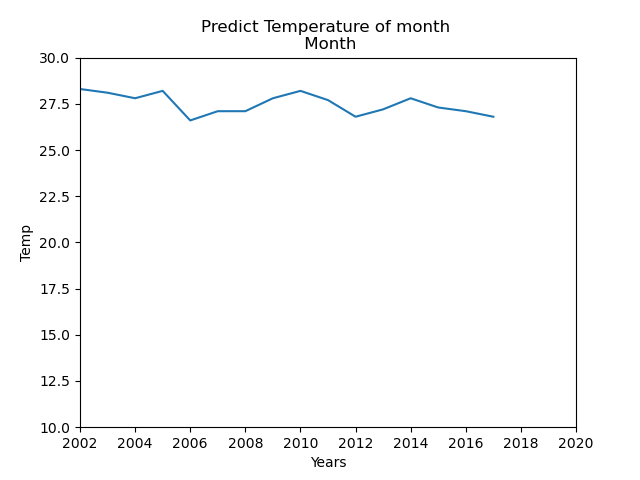
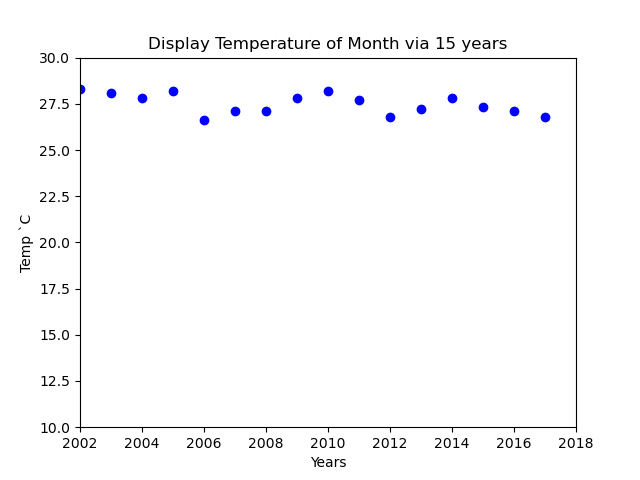


h

Tháng 8:

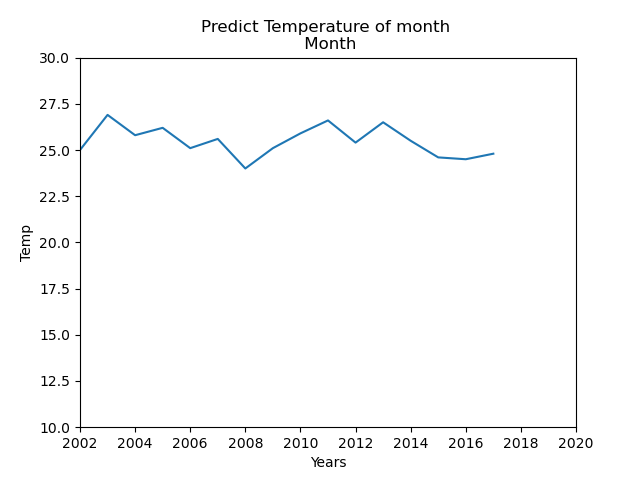
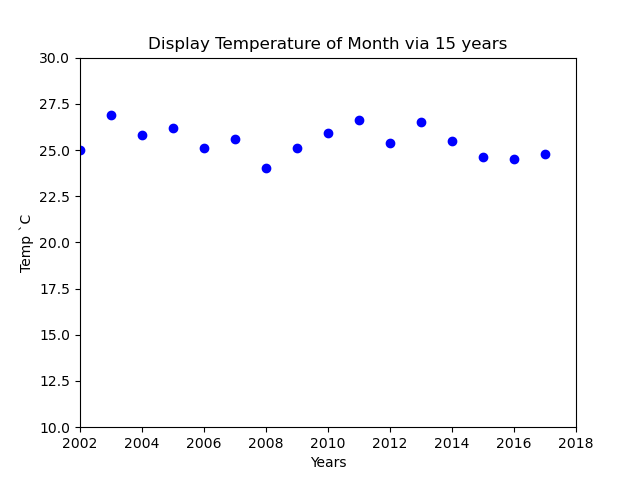


Tháng 9:

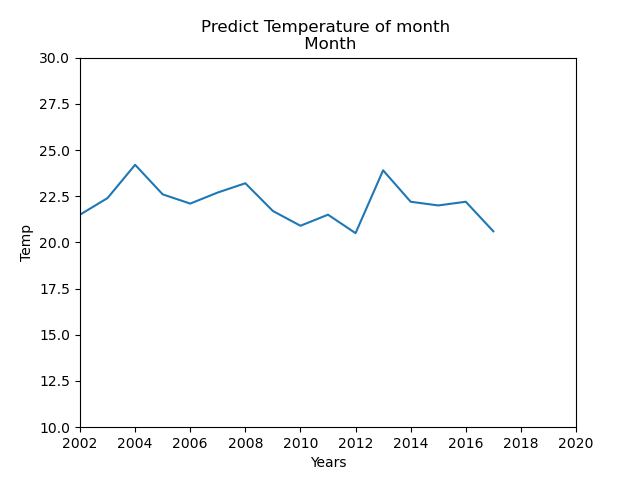
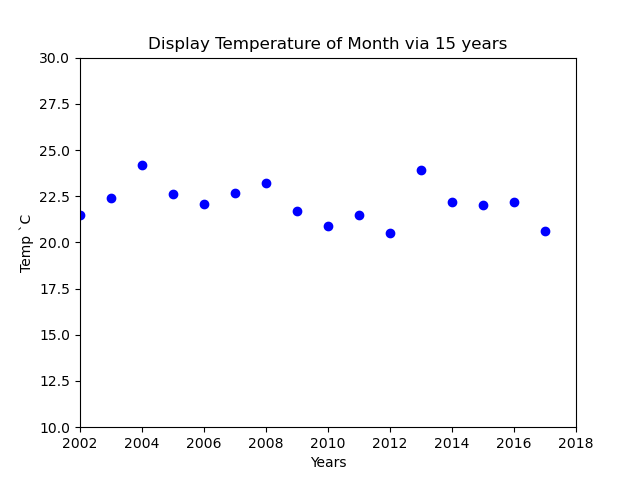


h

Tháng 10:

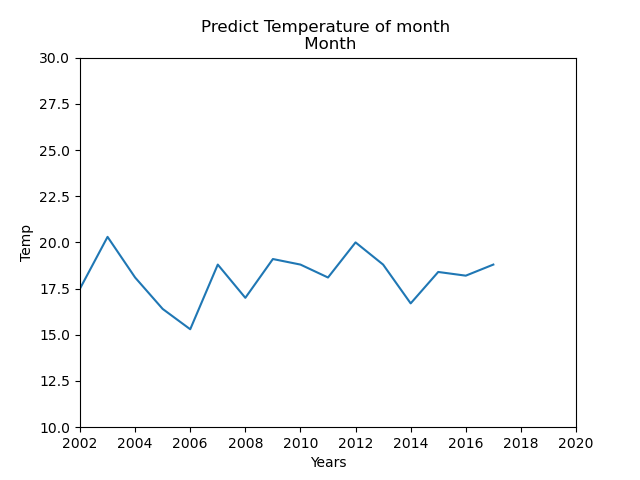
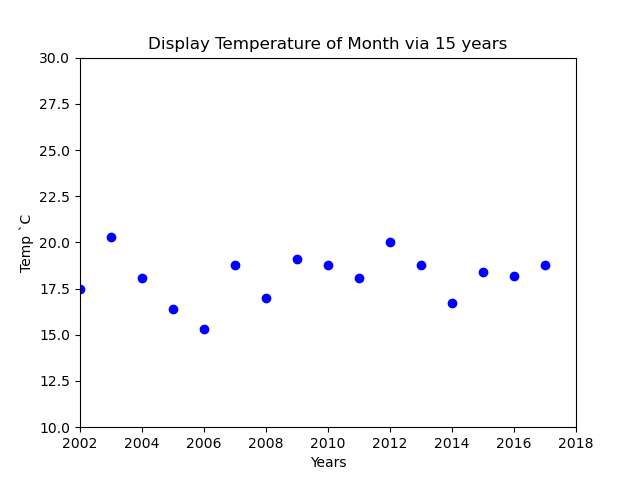


Tháng 11:



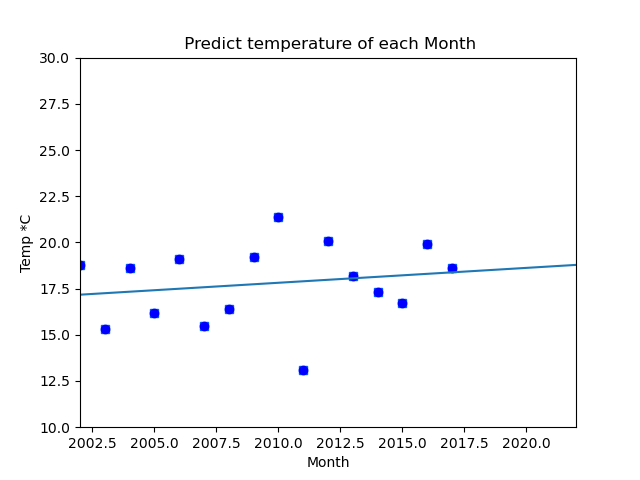
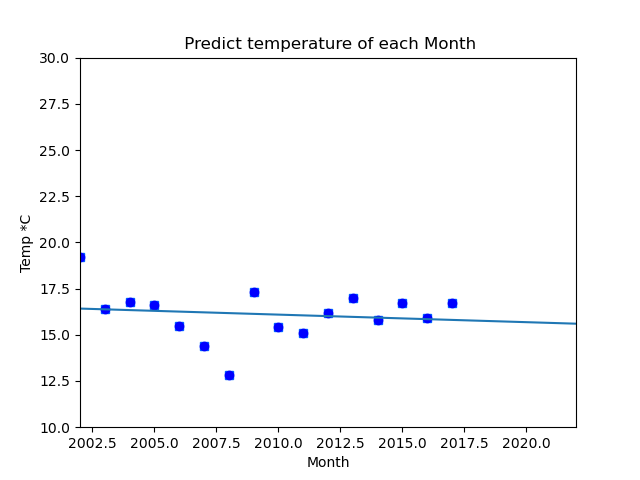
h

**Tháng 12:**



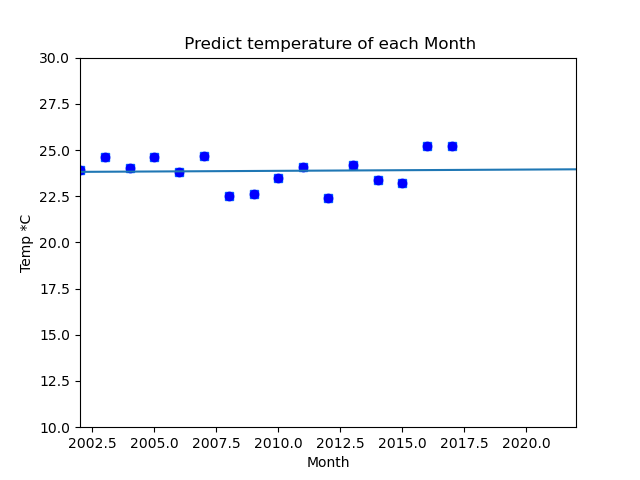
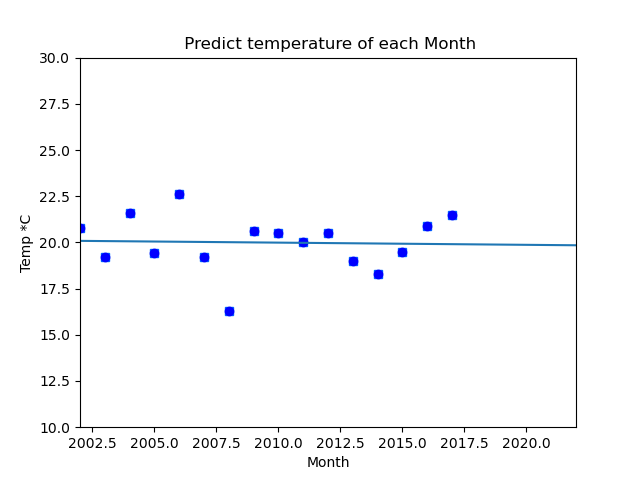
## 2.3. Mô tả kết quả dự đoán: Kết quả với đường Hồi quy tuyến tính:

**Tháng 1+2**

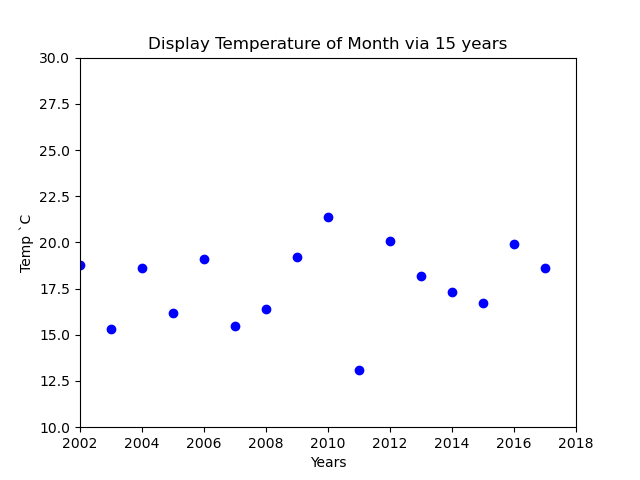


y = -144.1670619966 + x\*0.08058887315453 y = -144.16706199659 + x\*0.04088235405765772

**Tháng 3+4**

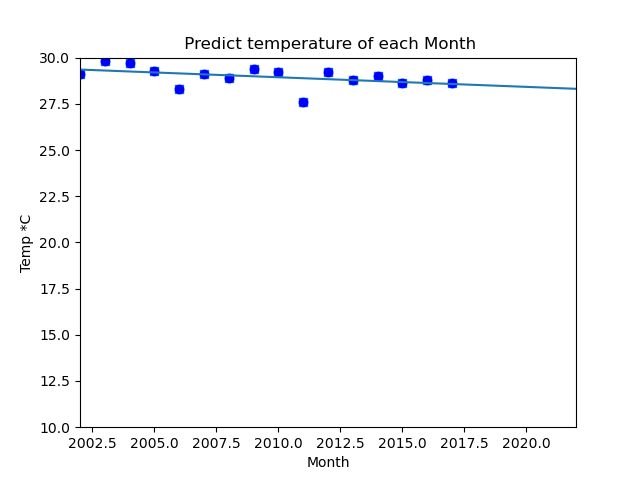
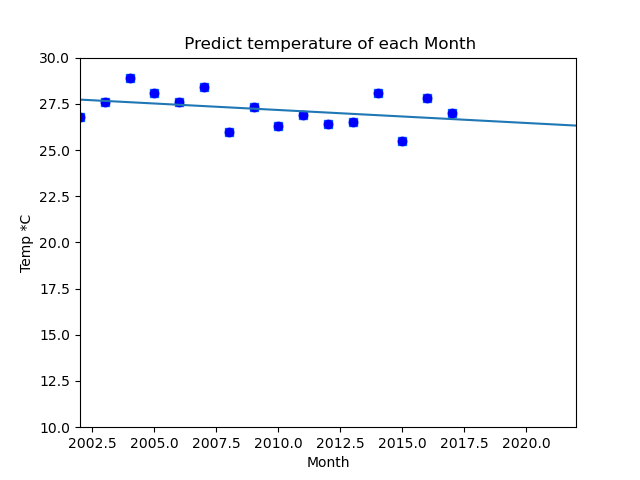


y = -144.1670619966 + x\*0.08058887315453 y = 9.9795591142028 + x\* 0.006911764561209566



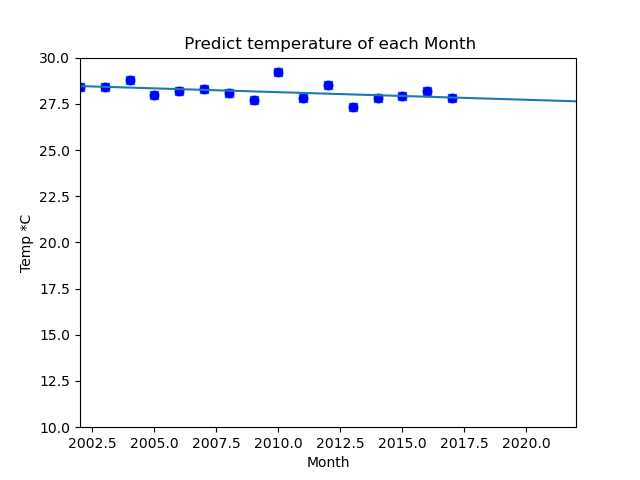
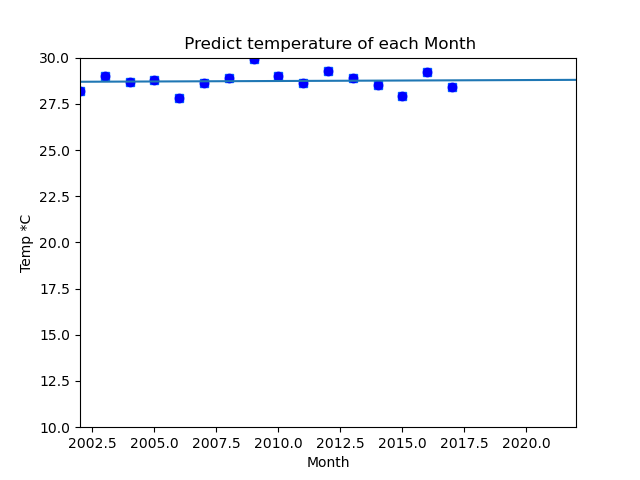
Tháng

**Tháng 5+6**



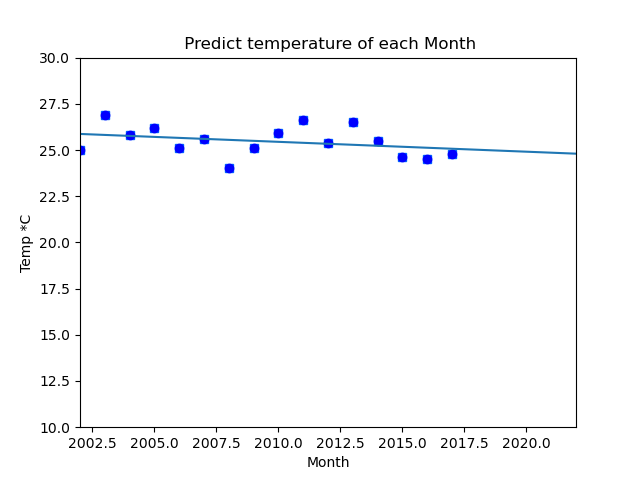
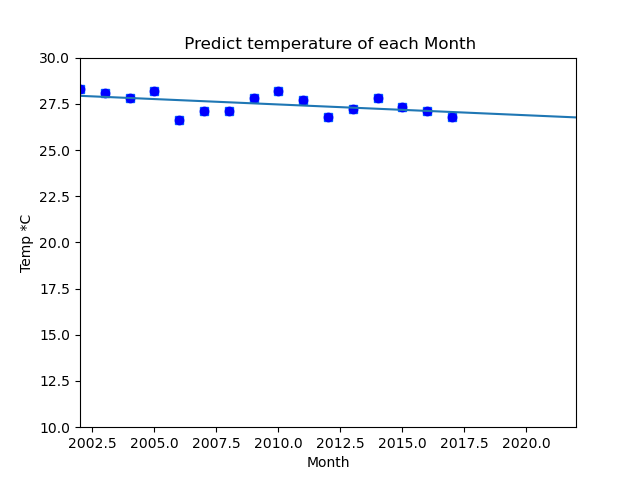
Y = 168.456033256836 – x\*0.070294119560458 Y = 133.574708954 - x\* 0.05205882505788395

**Tháng 7+8**



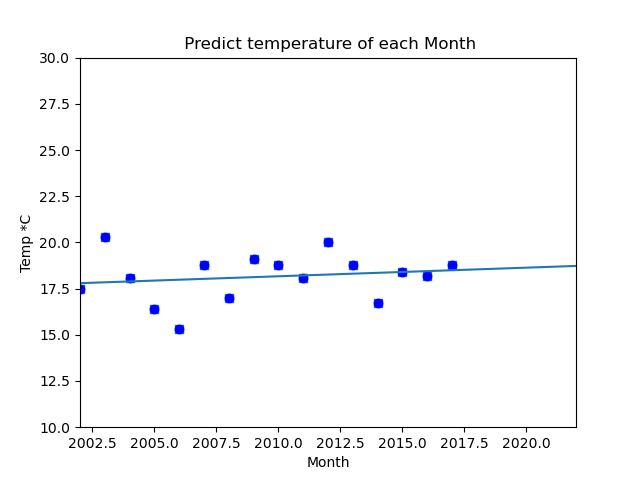
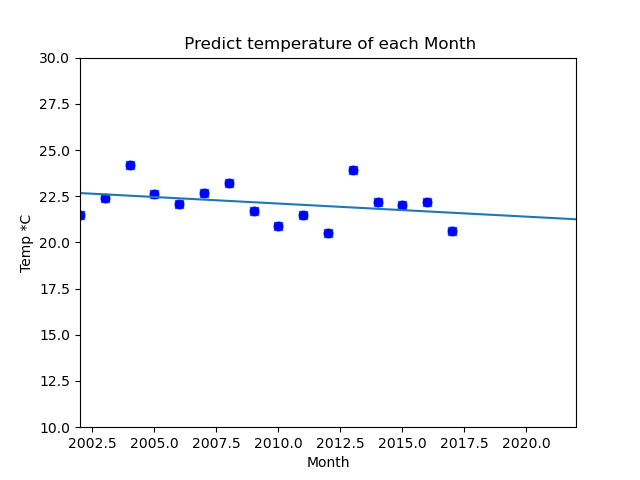
Y = 18.3882357878 + x\* 0.005147058577676944 Y = 133.57470895443 - 0.052058825057883

**Tháng 9+10**



Y = 110.89412020798 - x\* 0.04117647186 Y = 144.8130915537 – x \*0.058382354592595

**Tháng 11+12**



Y = 164.575591978617 – x\*0.070882354804325 Y = -75.53441340150312 +x \* 0.0466176478732

# 3. Kiểm ta với dữ liệu test

* Với nhiệt độ trung bình của 12 tháng tại Bãi Cháy năm 2018 là:

17,2 16.5 21.2 23.3 28.7 29.4 28.6 28 27.5 25.5 23.4 19.0

Áp dụng: với X =2018 cho các phương trình nhiệt độ theo từng tháng ta được

17.27 18.5 21.7 25.1 26.9 27,8 28.5 28.6 26.5 24.1 19.7 18.4

Độ chính xác:

99.6 89.2 97.7 92.8 93.7 94.6 99.6 97.8 96.3 94.5 84.2 96.7

* Nhiệt độ trung bình của 12 tháng tại Bãi Cháy năm 2019 là:

17.4 16.8 21 23.6 29 29.7 28.5 28.4 27.8 26 23.7 19.7

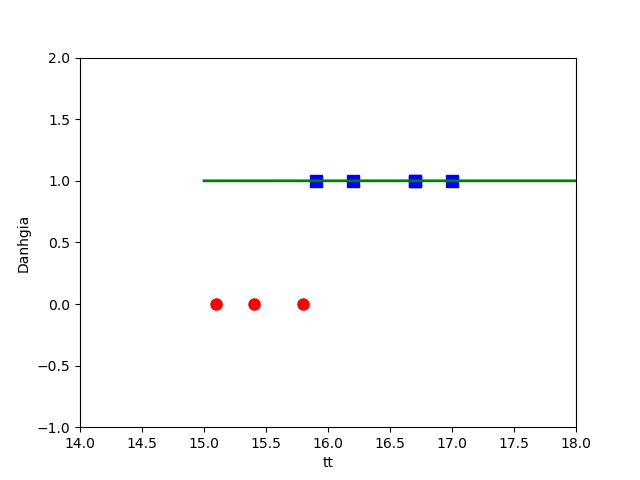
Áp dụng: với X =2019 cho các phương trình nhiệt độ theo từng tháng ta được

15.89 18.5 19.87 23.9 26.7 28 28.7 27.9 26.9 24.9 21.46 17.55

Độ chính xác:

91 90.8 94.6 98.7 92 93.7 94.2 98.2 97.8 96.7 90.5 89.1

Kiểm định độ chính xác với mô hình Logistic Regression( tháng 1):



# 4. Kết luận:

Dự báo thời tiết là một quán trình phức tạp, chưa thể chính xác tuyệt đối. Bên cạnh nguồn dữ liệu lớn, phong phú về( áp suất, tốc gió, độ ẩm,…các yếu tố khác). Sau đó mới cần đến kiến thức về khí quyển của chúng ta và một mô hình học máy phù hợp để đưa ra thông tin cần thiết trong không gian gần.

Bài toán của nhóm em ứng dụng hoàn toán kiến thức đã học được từ các thầy cô cũng như kiến thức tìm hiểu được về Học máy nói chung và Hồi quy tuyến tính nói riêng nhằm đưa ra một cách chính xác, đáng tin cậy nhất về dữ liệu thời tiết của không chỉ một địa điểm là Bãi Cháy như trong bài mà suy rộng ra tất cả các địa điểm khác. Và kết quả nghiệm thu được có độ chính xác khá cao, đủ đạt đến ngưỡng tin tưởng được.

Tuy nhiên do quỹ thời gian và kiến thức hạn chế nên bài toán, ý tưởng của nhóm chưa đi đến được như ý muốn. Nhóm rất mong nhận được những gợi ý và nhận xét của thầy/cô để đi tới 1 mô hình hoàn thiện, có ý nghĩa thực tế.

# 4. Tài liệu tham khảo

**Tham khảo:**

1. Các slide bài giảng của PGS Nguyễn Hữu Quỳnh và kiến thức được giảng dạy qua các buổi thực hành của thầy Cù Mạnh Dũng
2. Lý thuyết về Linear Regression trên trang Machine Learning cơ bản
3. Lý thuyết về Logistic Regression trên trang Machine Learning cơ bản
4. <https://machinelearningcoban.com/2016/12/28/linearregression/>
5. <https://www.geeksforgeeks.org/ml-rainfall-prediction-using-linear-regression/>

**Dữ liệu:**

1. <https://storage.googleapis.com/kaggle-data-sets/296368/607351/bundle/archive.zip?GoogleAccessId=web-data@kaggle-161607.iam.gserviceaccount.com&Expires=1589600596&Signature=lzqcfANFxeq%2BZ9SCNRH8gecPSl0Bm%2Fj0LWwWoy7WzfWFE5QVqHuNvIpjSeV6LWQIVKZBrcR3BJjupUKC9TtA8S6HEW%2BRYW%2BCLcvEtlS6KOuzuvwolbYTRWhmVqMuiVioFXG52AvGRIxuJ5EEN62d0Nw39w6inWgeD2oEJqmW2TeiCjGh1ccBCIXjs%2FZFGzpf5O0jAxHBuDHcn%2BQcs%2BZtFmYbCFsyhSUXXT1tz9zFSRejr7V5nDGW44b%2FAV33IoTBhjv5zAqfkpAZmB4aw%2B06Spz87UqzLVI6IT081Qmsdc72jR9%2FtpfOj87fOhukj0GSXDKjbk0hGyBWN97af0dETQ%3D%3D&response-content-disposition=attachment%3B+filename%3Dvietnam-weather-temperature-2002-2017.zip>
2. <https://www.kaggle.com/muthuj7/weather-dataset?select=weatherHistory.csv>

**Dữ liệu test:**

1. <https://www.gso.gov.vn/SLTK/Selection.aspx?rxid=8f161760-9ba0-4c6d-8898-fdef1a92c072&px_db=01.+%C4%90%C6%A1n+v%E1%BB%8B+h%C3%A0nh+ch%C3%ADnh%2C+%C4%91%E1%BA%A5t+%C4%91ai+v%C3%A0+kh%C3%AD+h%E1%BA%ADu&px_type=PX&px_language=vi&px_tableid=01.+%C4%90%C6%A1n+v%E1%BB%8B+h%C3%A0nh+ch%C3%ADnh%2C+%C4%91%E1%BA%A5t+%C4%91ai+v%C3%A0+kh%C3%AD+h%E1%BA%ADu%5CV01.13.px>

**Thư viện:**

1. <https://numpy.org/>
2. <https://pandas.pydata.org/>
3. <https://matplotlib.org/>
4. <https://pypi.org/project/future/>