Đại học Quốc gia Thành phố Hồ Chí Minh

Trường Đại học Khoa học Tự nhiên

ĐỒ ÁN
BÀI TOÁN KHÍ HẬU

Nhóm sinh viên thực hiện:

20120014 - Vương Gia Huy

20120021 - Hồ Văn Sơn

20120304 - Phan Trần Khanh

Giảng viên hướng dẫn: PGS. TS. Nguyễn Đình Thúc

Học phần Toán ứng dụng & thống kê

Lớp 20TN

04 - 2022

Mục Lục

1. Đặt vấn đề

- Xác định và hình thức hóa mục tiêu bài toán: Xác định **lượng mưa** của một tháng bất kì trong tương lai tại một thành phố nhất định.
- Dạng bài toán:
  - Kiểm định: Sự tương quan giữa các đối tượng như nhiệt độ, gió, ánh sáng mặt trời đối với lượng mưa.
  - Dự đoán: Từ dữ liệu hiện có trong hiện tại và quá khứ để đưa ra dự đoán dữ liệu trong tương lai chưa biết.
- Đối tượng được chọn cho bài toán:
  - Nhiệt độ
  - Gió
  - Lượng mưa
  - Ánh sáng mặt trời
- Phạm vi, mức độ, quy mô của bài toán:
  - Theo không gian, dữ liệu được thu thập trong một thành phố của một tiểu bang của một quốc gia: Thành phố Seattle, tiểu bang Washington, Mỹ
  - Khi tải dữ liệu về và quan sát chúng tôi quyết định lấy theo thời gian trong 39 năm từ tháng 01 - 1984 đến tháng 02 - 2022, vì dữ liệu được ghi nhận liên tục theo từng tháng.

2. Thu thập dữ liệu

Các bước nhóm đã thực hiện để thu thập dữ liệu cho bài toán dự đoán:

Bước 1. Truy cập vào *nguồn dữ liệu khí hậu của NOAA*. [1]

Bước 2. Sử dụng *Search Tool* để tìm kiếm dữ liệu.

Bước 3. Chọn các mục tương ứng *Global Summary of the Month* → *1763-01-01 to 2022-03-01* → *Stations*.

Bước 4. Tìm kiếm các trạm tại thành phố Seattle bằng cách nhập từ khóa "Seattle" ở mục **Enter a Search Term** và chọn *SEARCH* để kho dữ liệu thực hiện việc tìm kiếm.

Theo đánh giá của nhóm thực hiện thì trạm tại **SEATTLE TACOMA AIRPORT, WA US** có dữ liệu đủ lớn và được cập nhật liên tục cho đến thời điểm hiện tại, vì khoảng thời gian đã được ghi lại là từ 01/01/1948 đến 01/03/2022, hơn 7 thập kỉ.

Bước 5. Chọn *ADD TO CART* đối với trạm **SEATTLE TACOMA AIRPORT, WA US**.

Bước 6. Vào *CART(Free Data)* để hoàn tất thủ tục nhận dữ liệu từ hệ thống lưu trữ.

Bước 7. Nhóm thực hiện đề tài trên tập tin dữ liệu dạng ***.csv** nên đã chọn *Custom Global Summary of The Month CSV* → *CONTINUE*.

Bước 8. Chọn các mục tương ứng *Station Name* và **Units** dạng *Standard* → *Select All* ở mục **Select data types for custom output** → *CONTINUE*.

Theo đánh giá của nhóm thực hiện thì dữ liệu có đơn vị thuộc Metric System (Hệ thống đơn vị đo lường quốc tế SI) chúng có giá trị quá nhỏ dẫn đến bị lỗi số thực và không thể thực hiện huấn luyện mô hình hồi quy tuyến tính sau quá trình tiền xử lý dữ liệu.

Dưới đây là hình ảnh so sánh dữ liệu thu thập được từ NOAA theo Standard System và Metric System.

NAME	DATE	AWND	CDSD	CLDD	DP01	DP10	DSND	DSNW	DT00	DT32	DX32	DX70	DX90	EMNT
SEATTLE TACOMA AIRPORT, WA US	2002-01	4	0	0	19	13			0	5	0	0	0	-3.9
SEATTLE TACOMA AIRPORT, WA US	2002-02	3.6	0	0	18	8			0	5	0	0	0	-2.2
SEATTLE TACOMA AIRPORT, WA US	2002-03	4.4	0	0	13	8			0	9	0	0	0	-2.2
SEATTLE TACOMA AIRPORT, WA US	2002-04	3.6	0	0	14	9			0	0	0	0	0	1.1
SEATTLE TACOMA AIRPORT, WA US	2002-05	3.3	0	0	14	5			0	0	0	1	0	1.7
SEATTLE TACOMA AIRPORT, WA US	2002-06	3.3	17.7	17.7	10	4			0	0	0	11	1	5.6
SEATTLE TACOMA AIRPORT, WA US	2002-07	3.2	50.4	32.7	4	1			0	0	0	22	0	9.4
SEATTLE TACOMA AIRPORT, WA US	2002-08	2.7	81.8	31.4	3	0			0	0	0	24	1	10
SEATTLE TACOMA AIRPORT, WA US	2002-09	3	85.1	3.3	5	2			0	0	0	16	0	7.2
SEATTLE TACOMA AIRPORT, WA US	2002-10	2.5	85.1	0	8	1			0	0	0	1	0	1.1
SEATTLE TACOMA AIRPORT, WA US	2002-11	2.7	85.1	0	13	10			0	4	0	0	0	-1.1
SEATTLE TACOMA AIRPORT, WA US	2002-12	3.8	85.1	0	19	15			0	1	0	0	0	-2.2

Hình 1: Dữ liệu vào năm 2002 tại trạm Seattle Tacoma Airport, WA US có đơn vị theo Metric System

NAME	DATE	AWND	CDSD	CLDD	DP01	DP10	DSND	DSNW	DT00	DT32	DX32	DX70	DX90	EMNT
SEATTLE TACOMA AIRPORT, WA US	2002-01	8.9	0	0	19	13			0	5	0	0	0	25
SEATTLE TACOMA AIRPORT, WA US	2002-02	8.1	0	0	18	8			0	5	0	0	0	28
SEATTLE TACOMA AIRPORT, WA US	2002-03	9.8	0	0	13	8			0	9	0	0	0	28
SEATTLE TACOMA AIRPORT, WA US	2002-04	8.1	0	0	14	9			0	0	0	0	0	34
SEATTLE TACOMA AIRPORT, WA US	2002-05	7.4	0	0	14	5			0	0	0	1	0	35
SEATTLE TACOMA AIRPORT, WA US	2002-06	7.4	32	32	10	4			0	0	0	11	1	42
SEATTLE TACOMA AIRPORT, WA US	2002-07	7.2	91	59	4	1			0	0	0	22	0	49
SEATTLE TACOMA AIRPORT, WA US	2002-08	6	147	56	3	0			0	0	0	24	1	50
SEATTLE TACOMA AIRPORT, WA US	2002-09	6.7	153	6	5	2			0	0	0	16	0	45
SEATTLE TACOMA AIRPORT, WA US	2002-10	5.6	153	0	8	1			0	0	0	1	0	34
SEATTLE TACOMA AIRPORT, WA US	2002-11	6	153	0	13	10			0	4	0	0	0	30
SEATTLE TACOMA AIRPORT, WA US	2002-12	8.5	153	0	19	15			0	1	0	0	0	28

Hình 2: Dữ liệu vào năm 2002 tại trạm Seattle Tacoma Airport, WA US có đơn vị theo Standard System (US Standard Units)

Để phù hợp, nhóm chúng tôi sẽ chọn đơn vị theo **Standard System (US Standard Units)**

Bước 9. Thực hiện nhập Email và chọn *SUBMIT ORDER*.

Bước 10. Kiểm tra Email thường xuyên để nhận dữ liệu từ NOAA và thực hiện việc tải xuống dữ liệu từ đường dẫn của hòm thư NOAA.

3. Xử lý dữ liệu

Các bước nhóm đã thực hiện trích xuất dữ liệu cần thiết và tiền xử lý dữ liệu cho bài toán dự đoán

Bước 1. Loại bỏ cột dữ liệu **STATION**, **NAME** và **DATE**.

Bước 2. Loại bỏ các hàng dữ liệu có giá trị tại tất cả thuộc tính là rỗng.

Bước 3. Loại bỏ các cột thuộc tính có chứa giá trị rỗng.

Sau khi loại bỏ các cột thuộc tính có giá trị NaN bộ dữ liệu còn lại 17 thuộc tính thuộc 2 đối tượng là nhiệt độ và lượng mưa.

- Nhiệt độ: CLDD, HTDD, DX32, DX70, DX90, DT00, DT32, CDSD, EMNT, EMXT, TAVG, TMAX, TMIN.
- Lượng mưa: DP01, DP10, EMXP, PRCP.

Bước 4. Sử dụng các giá trị của thuộc tính **PRCP** làm giá trị được dự đoán và thực hiện chia bộ dữ liệu thành hai bộ *train* và *test* sử dụng *sklearn.model_selection.train_test_split* với tỉ lệ là 20% cho bộ *test* và 80% cho bộ *train*.

4. Phân tích, đánh giá

Về những mô hình phù hợp cho bài toán, nhóm đã tìm hiểu được ba loại mô hình thường được sử dụng để dự đoán lượng mưa như sau:

- Cây quyết định (Decision trees): Gini Index, Classification And Regression Tree (CART).
- Hồi quy (Regression): Linear Regression, Logistic Regression.
- Mạng nơ-ron nhân tạo (Neural networks): Multilayered Feedforward Neural Network (MLFNN), Radial Basis Function Neural Network (RBFNN), Focused Time Delay Neural Network (FTDNN), Nonlinear Autoregressive Exogenous Neural Network (NARXNN), Artificial Neural Network (ANN), ...

Nhóm đã quyết định chọn mô hình hồi quy tuyến tính (Linear Regression). Nguyên nhân:

- Đây là một mô hình đơn giản và quen thuộc khi đã học qua học phần Xác suất Thống kê, cũng như là Toán ứng dụng và Thống kê.
- Mô hình dễ thực thi, dễ huấn luyện, thông dịch nhanh.
- Có thể sử dụng các kĩ thuật đi kèm (cross-validation, regularization, dimensionality reduction) để xử lý hiện tượng overfitting trên tập dữ liệu.
- Đối với tập dữ liệu của NOAA, có thể dễ dàng kiểm định được quan hệ tương quan giữa các thuộc tính, vậy nên ta có thể chọn lựa được các thuộc tính phù hợp để mô hình hồi quy tuyến tính đạt hiệu quả hơn.

4.1. Kiểm định sự tương quan giữa các thuộc tính đầu vào với thuộc tính đầu ra (lượng mưa)

Nhóm sử dụng kiểm định T để kiểm định sự tương quan của từng thuộc tính với thuộc tính đầu ra (lượng mưa).

Mô hình hồi quy tổng thể:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + e$$

Mô hình hồi quy mẫu:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p$$

Trong đó: b_0 là ước lượng cho β_0, b_1 là ước lượng cho β_1, ...

Các bước kiểm định sự tương quan của thuộc tính CLDD với thuộc tính PRCP:

- **Bước 1.** Giả sử β là hệ số trong mô hình hồi quy tuyến tính tổng thể tương ứng với thuộc tính CLDD, b là hệ số được ước tính cho thuộc tính CLDD trong mô hình hồi quy tuyến tính mẫu. Đặt giả thuyết rằng hệ số β có đóng góp không đáng kể trong việc dự đoán giá trị PRCP. Giả thuyết được viết như sau:

$$H_0: \beta = 0$$
$$H_1: \beta \neq 0$$

- **Bước 2.** Đặt độ tin cậy là 95%.
- **Bước 3.** Tính giá trị p-value
  - Tính giá trị thống kê kiểm định: $t_{stat} = \dfrac{b}{Se}$

    Trong đó: b là hệ số được ước tính tương ứng với thuộc tính CLDD trong mô hình hồi quy tuyến tính mẫu, Se là sai số chuẩn của thuộc tính CLDD.
  - Tính p-value: $p_{value} = 2P(|t_{stat}| > t_{n-k+1})$. Với n là số bộ dữ liệu, k là số thuộc tính.
- **Bước 4.** Bác bỏ H_0 nếu p-value ≤ 0.05 (vì độ tin cậy ta chọn ở bước 2 là 95%, tức có sự tương quan giữa thuộc tính CLDD và thuộc tính PRCP. Ngược lại không đủ cơ sở để bác bỏ H_0, tức không có sự tương quan giữa CLDD và PRCP.

Trên đây là bước kiểm định sự tương quan giữa hai thuộc tính CLDD và PRCP. Ta thực hiện tương tự đối với các thuộc tính đầu vào khác còn lại. Ta được kết quả như hình sau:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.925
Model:	OLS	Adj. R-squared:	0.923
Method:	Least Squares	F-statistic:	671.6
Date:	Wed, 06 Apr 2022	Prob (F-statistic):	0.00
Time:	08:22:08	Log-Likelihood:	-943.15
No. Observations:	890	AIC:	1920.
Df Residuals:	873	BIC:	2002.
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
CONST	-0.3786	2.950	-0.128	0.898	-6.168	5.411
CDSD	0.0009	0.000	3.060	0.002	0.000	0.001
CLDD	-0.0004	0.002	-0.181	0.857	-0.005	0.004
DP01	0.0350	0.011	3.195	0.001	0.013	0.056
DP10	0.3078	0.012	25.286	0.000	0.284	0.332
DT00	2.3345	0.886	2.636	0.009	0.596	4.073
DT32	-0.0146	0.011	-1.310	0.191	-0.036	0.007
DX32	-0.0336	0.034	-0.977	0.329	-0.101	0.034
DX70	0.0350	0.009	3.708	0.000	0.016	0.053
DX90	0.0630	0.047	1.338	0.181	-0.029	0.156
EMNT	-0.0173	0.009	-1.951	0.051	-0.035	0.000
EMXP	1.7180	0.052	32.746	0.000	1.615	1.821
EMXT	-0.0092	0.006	-1.420	0.156	-0.022	0.004
HTDD	0.0002	0.001	0.103	0.918	-0.003	0.003
TAVG	0.8664	0.667	1.299	0.194	-0.443	2.176
TMAX	-0.4503	0.335	-1.345	0.179	-1.107	0.207
TMIN	-0.4039	0.335	-1.206	0.228	-1.061	0.253

Omnibus:	193.199	Durbin-Watson:	2.043
Prob(Omnibus):	0.000	Jarque-Bera (JB):	962.030
Skew:	0.901	Prob(JB):	1.25e-209
Kurtosis:	7.764	Cond. No.	6.21e+04

Chú ý ở cột $P > |t|$ ta nhận thấy rằng các thuộc tính có sự tương quan đối với **PRCP** là: **CDSD**, **DP01**, **DP10**, **DT00**, **DX70**, **EMXP**. Do đó ta sẽ chọn các thuộc tính này để huấn luyện mô hình hồi quy tuyến tính nhằm dự đoán lượng mưa trong tương lai.

Có thể tìm hiểu thêm tại đây [2]

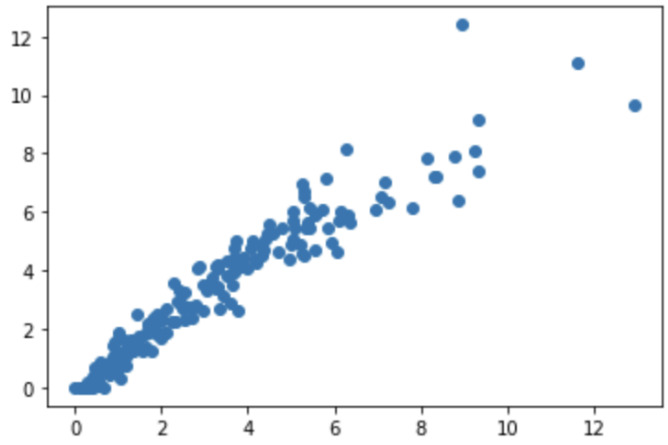
4.2. Huấn luyện mô hình

Mô hình đạt độ chính xác là **0.9116303786**. Sau đây là bảng kết quả của một số dự đoán từ mô hình tuyến tính.

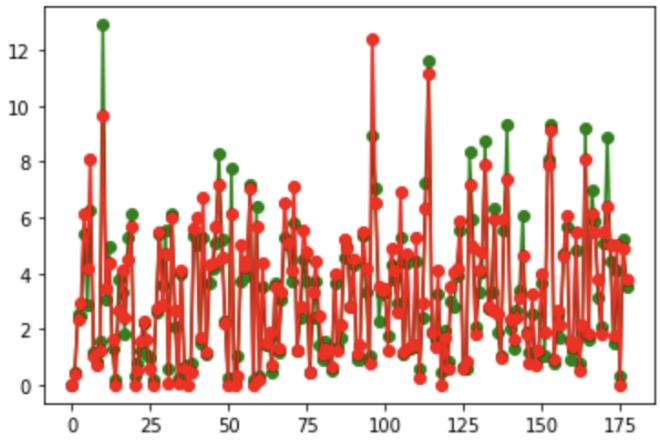
Predict	Actual	Predict	Actual
0	0	0.379321	0.42
2.346860	2.55	2.908851	2.45
6.126028	5.43	4.158078734	2.87
...	...	...	...
5.049497676	4.42	1.760970341	1.48
5.003678974	4.11	0.02118464785	0.31
4.911260927	5.22	3.795602661	3.51

Thông tin chi tiết: Bảng kết quả dự đoán lượng mưa tại Seattle

```
Accuracy: 0.9116303786
```



Hình 3. Biểu đồ phân tán



Hình 4. Biểu đồ lượng mưa dự đoán và lượng mưa thực tế

Để tìm hiểu thêm có thể tham khảo tại đây [3].

Bảng kết quả R - Squared (R^2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE).

R - Squared (R^2)	0.9133964148085189
Mean Squared Error (MSE)	0.5288315711922605
Root Mean Squared Error (RMSE)	0.7272080659565463
Mean Absolute Error (MAE)	0.509451016353164

Trước tiên, chúng tôi xin trình bày về các khái niệm như sau:

- R - Squared (R^2) là một thước đo thống kê về sự phù hợp cho biết mức độ biến thiên của một biến phụ thuộc được giải thích bởi (các) biến độc lập trong mô hình hồi quy. Được tính bởi công thức như sau:

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\Sigma_i(y_i - \hat{y})^2}{\Sigma_i(y_i - \bar{y_i})^2}$$

- Trong khi R - Squared (R^2) là thước đo tương đối để đánh giá mức độ phù hợp của mô hình với các biến phụ thuộc, thì Mean Squared Error (MSE) là thước đo tuyệt đối về mức độ phù hợp của mô hình. Được tính bởi công thức sau:

$$MSE = \frac{1}{N}\sum_{i}^{N}(y_i - \hat{y_i})^2$$

- Root Mean Square Error (RMSE) là căn bậc 2 của MSE, nó được sử dụng phổ biến hơn Mean Squared Error bởi vì đôi lúc, giá trị của MSE khá lớn để so sánh. Ngoài ra, nó được tính bằng bình phương sai số, và do đó căn bậc hai đưa nó trở lại cùng mức sai số dự đoán và giúp việc giải thích dễ dàng hơn.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i}^{N}(y_i - \hat{y_i})^2}$$

- Mean Absolute Error (MAE) tương tự như Mean Squared Error (MSE). Tuy nhiên, thay vì tính tổng bình phương sai số, MAE tính tổng trị tuyệt đối của sai số như sau:

$$MAE = \frac{1}{N}\sum_{i}^{N}|y_i - \hat{y_i}|$$

Để tìm hiểu thêm bạn có thể nhấp vào đây [4].

Dựa vào các thông số ở bảng kết quả, ta có thể đưa ra kết luận về mô hình này là **tốt** vì:

- $R^2 = 0.9133964148085189$ một giá trị tiệm cận giá trị 1, có ý nghĩa là mô hình này phù hợp với tập dữ liệu ở mức 91.34%. Nói cách khác, 91.34% biến thiên của biến phụ thuộc được giải thích bởi các biến độc lập (còn 8.66% còn lại ở đâu, dĩ nhiên là do sai số đo lường, do cách thu thập dữ liệu, do có thể có biến độc lập khác giải thích cho biến phụ thuộc mà chưa được được vào mô hình nghiên cứu... vv). Dựa vào R^2 cũng cho thấy mô hình phù hợp mạnh mẽ.

- Ta để ý rằng, RMSE và MAE có giá trị rất gần nhau. Hơn nữa, giá trị chúng bé đã chỉ ra rằng mô hình có lỗi khá thấp (MAE và RMSE càng thấp thì các ít lỗi).

- Khi ta tính độ chính xác của tập dữ liệu test thì cho ra con số chính xác (accuracy) = 0.9116303786, con số khá cao cũng cho ta thấy độ chính xác của mô hình, hơn nữa dựa vào biểu đồ phân tán cũng có thể thấy các điểm dữ liệu gần như phân bố tuyến tính.

- Ngoài ra, để ý rằng nhóm chúng tôi còn có sử dụng kĩ thuật kiểm định sự tương quan để đưa ra những thuộc tính thực sự liên quan mật thiết đến việc dự đoán lượng mưa. Điều này đã góp phần giúp cho mô hình có một kết quả có thể xem là tích cực.

5. Kết luận

Như chúng ta được biết, tình hình biến đổi khí hậu ngày càng trở nên phức tạp, khó lường, nếu như con người dự đoán được những đợt mưa bất thường thì sẽ có tác động rất lớn đến công tác phòng bị cho thiên tai. Ngoài ra, con người cũng rất cần thông tin về lượng mưa để đưa ra quyết định sử dụng tài nguyên nước một cách hiệu quả, nhằm nâng cao năng suất cây trồng và có sẵn hoạch định trước việc tưới tiêu. Như vậy, việc có những mô hình dự báo thời tiết là rất cần thiết. Mô hình **multiple linear regression** được chúng tôi trình bày cho một kết quả tương đối khả quan, độ lỗi khá thấp và độ chính xác khá cao. Tất nhiên, mô hình còn có thể cải thiện được do lượng dữ liệu chúng tôi sử dụng là còn nhỏ, nếu có thêm dữ liệu tin cậy thì chúng tôi tin là mô hình này sẽ còn tốt hơn nữa. Bên cạnh mô hình trên, chúng tôi sẽ tiếp tục nghiên cứu và thực nghiệm với một số khác cao cấp hơn chẳng hạn mạng nơ ron nhân tạo (ANN), mạng nơ ron hồi quy (RNN)... để có thể cho ra một kết luận tốt nhất về mô hình nên sử dụng.

6. Mô tả thuộc tính

Thuộc tính	Mô tả chi tiết	Đơn vị (US Standard)
AWND	Tốc độ gió trung bình hàng tháng	Miles/hour
CLDD	Nhiệt độ trung bình mỗi ngày - 65 độ F	Độ F
CDSD	CLDD tính theo mùa từ tháng 7 ở Bắc Bán cầu	Độ F
DP01	Số ngày có lượng mưa >= 0.1 inch trong tháng	Ngày
DP10	Số ngày có lượng mưa >= 1 inch trong tháng	Ngày
DT00	Số ngày có nhiệt độ thấp nhất <= 0 độ F trong tháng	Ngày
DT32	Số ngày có nhiệt độ thấp nhất <= 32 độ F trong tháng	Ngày
DX32	Số ngày có nhiệt độ cao nhất <= 32 độ F trong tháng	Ngày
DX70	Số ngày có nhiệt độ cao nhất <= 70 độ F trong tháng	Ngày
DX90	Số ngày có nhiệt độ cao nhất <= 90 độ F trong tháng	Ngày
EMNT	Nhiệt độ thấp nhất trong tháng	Độ F
EMXP	Tổng lượng mưa cao nhất trong tháng	Inches
EMXT	Nhiệt độ cao nhất trong tháng	Độ F
HTDD	65 độ F - nhiệt độ trung bình mỗi ngày	độ F
HDSD	HTDD tính theo mùa từ tháng 7 ở Bắc Bán cầu	Độ F
PRCP	Tổng lượng mưa hàng tháng	Inches
TAVG	Nhiệt độ trung binh hàng tháng	Độ F
TMAX	Trung bình nhiệt độ cao nhất mỗi ngày trong tháng	Độ F
TMIN	Trung bình nhiệt độ thấp nhất mỗi ngày trong tháng	Độ F

Tài liệu tham khảo

[1] URL: https://www.ncdc.noaa.gov/cdo-web/.

[2] URL: https://medium.com/nerd-for-tech/hypothesis-testing-on-linear-regression-c2a1799ba964.

[3] URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.

[4] URL: https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b.