Báo cáo môn Thu thập và tiền xử lý dữ liệu: Bộ dữ liệu các yếu tố thời tiết tại Thành phố Hồ Chí Minh và mô hình dự đoán nhiệt độ

Nguyễn Hoàng Minh[1], Nguyễn Minh Tiến[2], Tạ Nhật Minh[3], Nguyễn Đức Hiển[4]

[1] Trường Đại học Công nghệ thông tin
20521609@gm.uit.edu.vn
[2] Trường Đại học Công nghệ thông tin
20522010@gm.uit.edu.vn
[3] Trường Đại học Công nghệ thông tin
20521614@gm.uit.edu.vn
[4] Trường Đại học Công nghệ thông tin
20521307@gm.uit.edu.vn

Tóm tắt. Trong bài báo cáo này, chúng tôi đề cập đến quá trình thu thập, xử lý và xây dựng bộ dữ liệu về các yếu tố thời tiết tại Thành phố Hồ Chí Minh từ năm 1990 đến năm 2021 được thu thập từ trang POWER Data Access Viewer. Đồng thời thực hiện phân tích và sử dụng bộ dữ liệu trên vào các mô hình máy học. Ở đây, chúng tôi sử dụng các mô hình máy học bao gồm: Linear Regression, Random Forest Regression, Support Vector Regression và đã đạt được các kết quả. Điều này chứng minh rằng bộ dữ liệu có ý nghĩa thống kê, từ đó có thể ứng dụng vào việc xây dựng các mô hình khác.

Từ khóa: Weather elements dataset, Linear Regression (LR), Random Forest Regression (RFR), Support Vector Regression (SVR), Ho Chi Minh city.

Mở đầu

Hiện nay, những tác động của con người đối với môi trường ngày càng nhiều, dẫn đến sự nóng lên toàn cầu và biến đổi khí hậu càng khắc nghiệt hơn. Bài toán được đặt ra là dự đoán nhiệt độ và thời tiết bằng những dữ liệu đã thu thập được trong gần các năm qua, bao gồm: nhiệt độ cao nhất, nhiệt độ thấp nhất, độ ẩm trung bình, v.v. trong trường hợp không thể sử dụng những phương pháp dự báo thời tiết thông thường.

Vì thế, chúng tôi đã thu thập và xây dựng một bộ dữ liệu về những yếu tố thời tiết tại các quận/huyện ở Thành phố Hồ Chí Minh trong hơn 30 năm qua nhằm phục vụ cho việc tạo dựng những mô hình dự đoán thời tiết, khí hậu cũng như những dự án có liên quan. Đồng thời, chúng tôi cũng đã chuẩn bị một số mô hình dự đoán nhiệt độ sử dụng bộ dữ liệu này ở trong báo cáo này.

Bài báo cáo của chúng tôi được sắp xếp như sau: Trong phần 1, chúng tôi sẽ giới thiệu bài toán liên quan đến bộ dữ liệu được nói trên. Phần 2 và 3 sẽ lần lượt nói về quá trình thu thập và xử lý dữ liệu. Tại phần 4, chúng tôi sẽ đưa ra một số ứng dụng và mô hình có sử dụng đến bộ dữ liệu. Cuối cùng, chúng tôi sẽ đưa ra một số kết luận chung và phương hướng trong phần 5.

1 Giới thiệu bài toán

Với yêu cầu của đề tài, mục tiêu đặt ra cho nhóm thực hiện là phải giải quyết được hai bài toán cụ thể:

Bài toán 1. Thu thập và xây dựng bộ dữ liệu các yếu tố về thời tiết tại Thành phố Hồ Chí Minh

Đầu vào: Dữ liệu trên web, API, trung tâm dữ liệu mở, v.v.

Đầu ra: Bộ dữ liệu các yếu tố thời tiết tại Thành phố Hồ Chí Minh dưới dạng file csv.

Yêu cầu: Dữ liệu phải được thu thập từ nguồn uy tín, đảm bảo tính chính xác của dữ liệu.

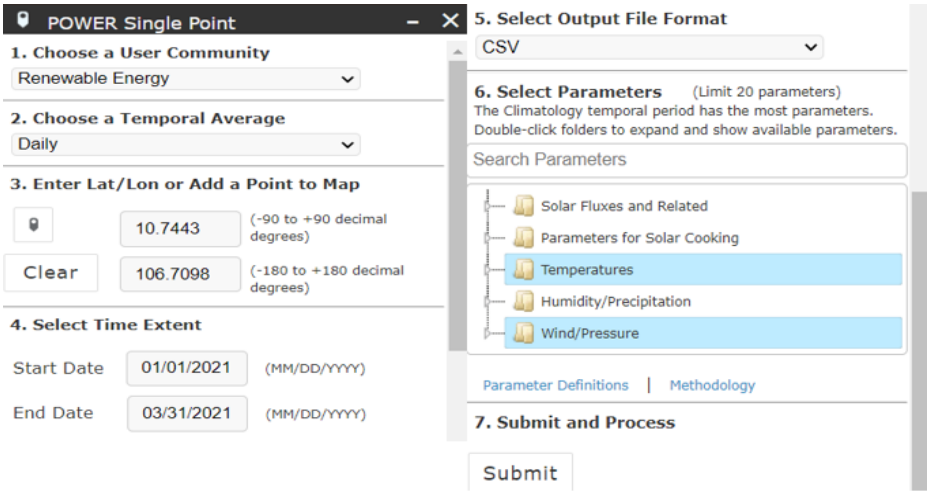
Bài toán 2. Xây dựng các mô hình dự đoán nhiệt độ trên bộ dữ liệu được thu thập

Đầu vào: Các yếu tố đo được từ khí hậu tại Thành phố Hồ Chí Minh

Đầu ra: Giá trị nhiệt độ trung bình máy dự đoán

2 Quá trình thu thập

Trong quá trình nghiên cứu và xây dựng bộ dữ liệu, chúng tôi đã quyết định thu thập dữ liệu thô từ trang POWER | Data Access Viewer [1]. Trang web này có chứa những một lượng lớn dữ liệu về các yếu tố thời tiết trên thế giới qua nhiều năm. Chúng tôi đã chọn Thành phố Hồ Chí Minh làm địa điểm chính để thu thập dữ liệu và xây dựng bộ dữ liệu này.



Hình 1. Hộp thoại tùy chọn thu thập dữ liệu từ trang web

Dữ liệu được thu thập bằng cách điều chỉnh các tùy chọn trong hộp thoại và tọa độ của các quận/huyện thuộc Thành phố Hồ Chí Minh. Sau khi kết thúc quá trình, chúng tôi thu được kết quả là tập dữ liệu thô bao gồm các bản ghi được sắp xếp theo ngày, từ ngày 01/01/1990 đến ngày 31/03/2021 và 24 tệp dữ liệu chứa 14 thuộc tính thời tiết được lưu dưới định dạng .csv

YEAR	MO	DY	WS50M	WS50M_R	QV2M	PRECTOT	ALLSKY_SF	T2M_RAN	ALLSKY_SF	T2M	ALLSKY_K	ALLSKY_SF	WS2M	ALLSKY_SF	ALLSKY_SF	ALLSKY_SFC_PAR_TOT
1990	1	1	4.02	8.31	13	0	-999	13.52	5.35	27.12	0.62	408.8	1.95	-999	-999	110.9
1990	1	2	4.64	7.42	12.82	0	-999	15.11	5.7	26.86	0.66	397.7	2.18	-999	-999	117
1990	1	3	3.78	6.87	12.51	0.03	-999	15.03	5.82	27.34	0.68	395.4	1.76	-999	-999	118.7
1990	1	4	3.75	6.32	12.02	0.04	-999	14.61	5.41	27.55	0.63	415.9	1.7	-999	-999	112.1
1990	1	5	3.55	6.11	11.6	0	-999	13.82	5.65	27.26	0.66	410.8	1.56	-999	-999	117.4
1990	1	6	3.46	7.13	11.41	0	-999	14.66	5.19	27.6	0.6	416.2	1.64	-999	-999	108
1990	1	7	3.71	5.89	10.99	0	-999	15.71	6	27.78	0.7	389.3	1.56	-999	-999	120.4
1990	1	8	3.33	3.98	13.12	0	-999	13.72	5.1	28.4	0.59	428.7	1.51	-999	-999	106.5
1990	1	9	3.82	6.24	14.28	0	-999	11.8	5.24	29.07	0.6	434.1	2	-999	-999	109.3
1990	1	10	4.28	5.53	13.92	0.01	-999	13.41	5.2	29.19	0.6	422.8	2.18	-999	-999	109.9
1990	1	11	4.03	6.04	14.1	0.08	-999	13.57	4.91	29.26	0.56	425.1	1.95	-999	-999	101.5
1990	1	12	5.48	4.17	13.55	0.01	-999	14.02	5.83	30.15	0.67	408.2	2.77	-999	-999	119.6
1990	1	13	5.31	6.73	12.94	0.01	-999	14.7	5.91	29.58	0.67	405	2.74	-999	-999	121.9
1990	1	14	5.39	5.89	12.88	0	-999	15.31	5.71	29.07	0.65	411.9	2.66	-999	-999	118.5
1990	1	15	4.73	9.01	13.18	0	-999	15.44	5.63	29.18	0.64	415.2	2.32	-999	-999	116.7
1990	1	16	4.57	6.23	13	0	-999	15.23	5.19	28.74	0.59	416.4	2.33	-999	-999	108.3
1990	1	17	5.58	6.65	12.51	0.02	-999	14.12	6.02	28.49	0.68	395.9	2.88	-999	-999	122.3
1990	1	18	5	8.34	12.63	0	-999	15.13	5.51	28.31	0.62	406.8	2.59	-999	-999	114.2
1990	1	19	5.56	5.3	12.88	0.03	-999	14.9	5.86	28.34	0.66	415.2	2.95	-999	-999	122.3
1990	1	20	4.43	5.97	12.02	0	-999	14.55	5.03	29.16	0.56	418.5	2.08	-999	-999	105.1
1990	1	21	4.05	6.5	11.05	0	-999	14.88	5.93	28.76	0.66	405.3	1.79	-999	-999	123.8
1990	1	22	3.73	7.46	10.99	0	-999	15.07	5.94	29.3	0.66	408.9	1.66	-999	-999	124.4
1990	1	23	3.23	5.97	10.68	0	-999	14.2	5.98	28.61	0.66	410.7	1.39	-999	-999	123
1990	1	24	4.08	4.79	10.38	0	-999	14.23	6.02	27.59	0.67	403.6	1.8	-999	-999	123.8
1990	1	25	3.11	7.75	10.74	0.01	-999	15.07	5.95	28.08	0.66	404.4	1.56	-999	-999	122.5
1990	1	26	3.68	7.41	10.86	0.05	-999	15.2	6.12	28.26	0.68	404	1.7	-999	-999	125.9
1990	1	27	3.55	6.84	11.05	0.18	-999	13.84	6.24	28.37	0.69	397.4	1.66	-999	-999	128.1
1990	1	28	3.4	8.37	11.05	0.07	-999	14.36	6.18	27.93	0.68	393.9	1.65	-999	-999	125.5
1990	1	29	3.41	6.98	11.6	0	-999	15.47	6.1	27.98	0.67	394.6	1.7	-999	-999	125.1

Hình 2. Một tệp dữ liệu thô bao gồm các thuộc tính

3 Xử lý và tổ chức dữ liệu

3.1 Tổ chức dữ liệu

Sau khi khảo sát trên các file thu thập được, kết quả cho thấy cấu trúc của các file dữ liệu khá sạch và có tính đồng nhất với nhau (Hình 3.1, 3.2). Chính vì vậy, việc đồng nhất dữ liệu sẽ không được xem xét.

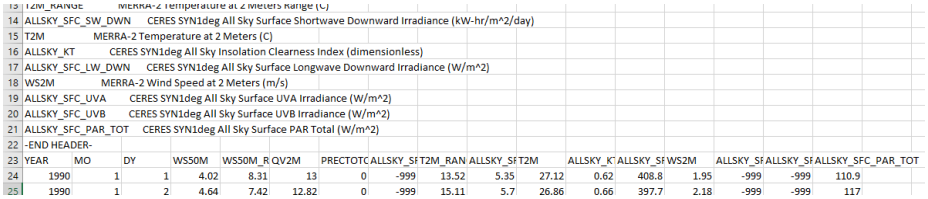
23	YEAR	MO	DY	WS50M	WS50M_R	QV2M	PRECTOTC	ALLSKY_SF	T2M_RAN	ALLSKY_SF	T2M	ALLSKY_K	ALLSKY_SF	WS2M	ALLSKY_SF	ALLSKY_SF	ALLSKY_SFC_PAR_TOT
24	1990	1	1	4.02	8.31	13	0	-999	13.52	5.35	27.12	0.62	408.8	1.95	-999	-999	110.9
25	1990	1	2	4.64	7.42	12.82	0	-999	15.11	5.7	26.86	0.66	397.7	2.18	-999	-999	117
26	1990	1	3	3.78	6.87	12.51	0.03	-999	15.03	5.82	27.34	0.68	395.4	1.76	-999	-999	118.7
27	1990	1	4	3.75	6.32	12.02	0.04	-999	14.61	5.41	27.55	0.63	415.9	1.7	-999	-999	112.1

Hình 3. Hình ảnh một vài dữ liệu ở bảng dữ liệu Quan_1.csv

23	YEAR	MO	DY	WS50M	WS50M_R	QV2M	PRECTOTC	ALLSKY_SF	T2M_RAN	ALLSKY_SF	T2M	ALLSKY_K	ALLSKY_SF	WS2M	ALLSKY_SF	ALLSKY_SF	ALLSKY_SFC_PAR_TO
24	1990	1	1	4.02	8.31	13	0	-999	13.52	5.35	27.12	0.62	408.8	1.95	-999	-999	110.9
25	1990	1	2	4.64	7.42	12.82	0	-999	15.11	5.7	26.86	0.66	397.7	2.18	-999	-999	117
26	1990	1	3	3.78	6.87	12.51	0.03	-999	15.03	5.82	27.34	0.68	395.4	1.76	-999	-999	118.7
27	1990	1	4	3.75	6.32	12.02	0.04	-999	14.61	5.41	27.55	0.63	415.9	1.7	-999	-999	112.1
28	1990	1	5	3.55	6.11	11.6	0	-999	13.82	5.65	27.26	0.66	410.8	1.56	-999	-999	117.4

Hình 4. Hình ảnh một vài dữ liệu ở bảng dữ liệu Quan_Thu_Duc.csv

Ở các dòng đầu của bộ dữ liệu là các chú thích về các thông tin thuộc tính của bộ dữ liệu (xem Hình 5) nên chúng tôi đã loại bỏ các dòng này ra khỏi bảng dữ liệu.



13	T2M_RANGE	MERRA-2 Temperature at 2 Meters Range (C)															
14	ALLSKY_SFC_SW_DWN	CERES SYN1deg All Sky Surface Shortwave Downward Irradiance (kW-hr/m^2/day)															
15	T2M	MERRA-2 Temperature at 2 Meters (C)															
16	ALLSKY_KT	CERES SYN1deg All Sky Insolation Clearness Index (dimensionless)															
17	ALLSKY_SFC_LW_DWN	CERES SYN1deg All Sky Surface Longwave Downward Irradiance (W/m^2)															
18	WS2M	MERRA-2 Wind Speed at 2 Meters (m/s)															
19	ALLSKY_SFC_UVA	CERES SYN1deg All Sky Surface UVA Irradiance (W/m^2)															
20	ALLSKY_SFC_UVB	CERES SYN1deg All Sky Surface UVB Irradiance (W/m^2)															
21	ALLSKY_SFC_PAR_TOT	CERES SYN1deg All Sky Surface PAR Total (W/m^2)															
22	-END HEADER-																
23	YEAR	MO	DY	WS50M	WS50M_R	QV2M	PRECTOTC	ALLSKY_SF	T2M_RAN	ALLSKY_SF	T2M	ALLSKY_K	ALLSKY_SF	WS2M	ALLSKY_SF	ALLSKY_SF	ALLSKY_SFC_PAR_TOT
24	1990	1	1	4.02	8.31	13	0	-999	13.52	5.35	27.12	0.62	408.8	1.95	-999	-999	110.9
25	1990	1	2	4.64	7.42	12.82	0	-999	15.11	5.7	26.86	0.66	397.7	2.18	-999	-999	117

Hình 5. Hình ảnh bảng dữ liệu Quan_Thu_Duc.csv

Sau đó, chúng tôi thêm cột LOC có giá trị là tên các quận/huyện lên từng bộ dữ liệu để ghi lại thông tin dữ liệu thời tiết của quận/huyện trước khi gộp thành một bảng dữ liệu chung của thành phố.

Sau khi dữ liệu đã sẵn sàng, chúng tôi tiến hành gộp 24 bảng dữ liệu của các quận, huyện của thành phố thành một file thống nhất chứa tất cả dữ liệu về thời tiết của thành phố Hồ Chí Minh, kết quả sau khi gộp dữ liệu ta sẽ được một bộ dữ liệu gồm 273912 dòng.

3.2 Làm sạch dữ liệu

Qua kết quả cho thấy các thuộc tính đa phần đều có dạng giống phân phối chuẩn, hơi lệch hoặc hai yếu vị, các đồ thì được xem là không bất thường trong dữ liệu. Ở thuộc tính 'PRECTOTCORR' có khoảng dữ liệu khá dài và cần được xem xét nếu phân tích sâu. Ngoài ra, các giá trị bị thiếu (có giá trị -999) đã làm cho phân phối của các thuộc tính bị chia thành hai cột.

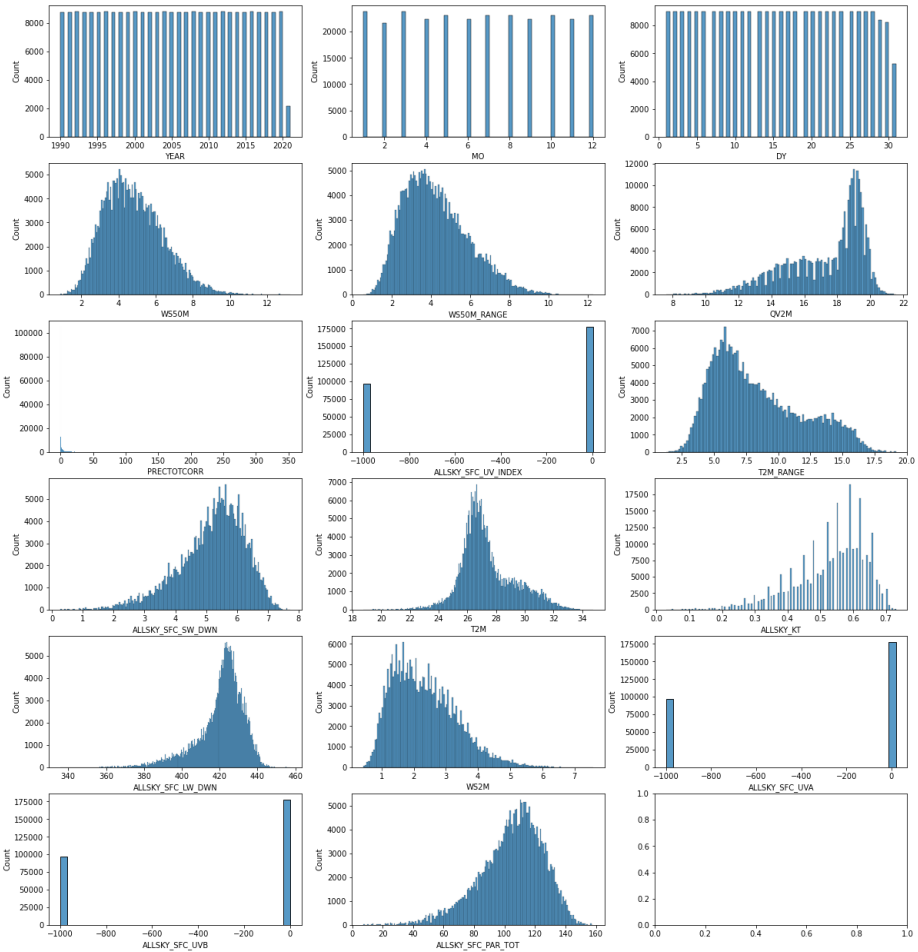
Khảo sát với các thuộc tính có các giá trị bị thiếu, kết quả cho thấy các dữ liệu bị mất nằm ở ba thuộc tính 'ALLSKY_SFC_UV_INDEX', 'ALLSKY_SFC_UVA', 'ALLSKY_SFC_UVB' lần lượt là 96558, 96432, 96432, và các dữ liệu này đa phần đều nằm từ năm 1992 đến năm 1999, lý do có thể là tại thời điểm đó, chưa có máy móc để đo được các thông số này. Với các giá trị bị thiếu, chúng tôi có 3 kịch bản để xử lý các dữ liệu này:

Kịch bản 1: Loại bỏ các dữ liệu bị thiếu (xóa hàng hoặc xóa cột)

Kịch bản 2: Điền thiếu dựa trên các giá trị thuộc tính đã có (điền bằng giá trị mặc định hoặc điền bằng giá trị dự đoán gần nhất)

Kịch bản 3: Điền bằng tay

Với kịch bản 2, bộ dữ liệu sẽ bị giảm đi tính chính xác, chính vì vậy chúng tôi sẽ loại bỏ. Với kịch bản 3, dữ liệu nhiều nên cũng không thực hiện được. Với kịch bản 1, chúng tôi tính số lượng các dữ liệu bị mất khi thực hiện xóa hàng hoặc cột có dữ liệu bị thiếu lần lượt là 1255254 và khi xóa hàng và 532062 khi xóa cột. Như vậy chúng tôi đã quyết định loại bỏ 3 cột này khỏi bộ dữ liệu.



Hình 6. Đồ thị phân phối dữ liệu theo từng thuộc tính

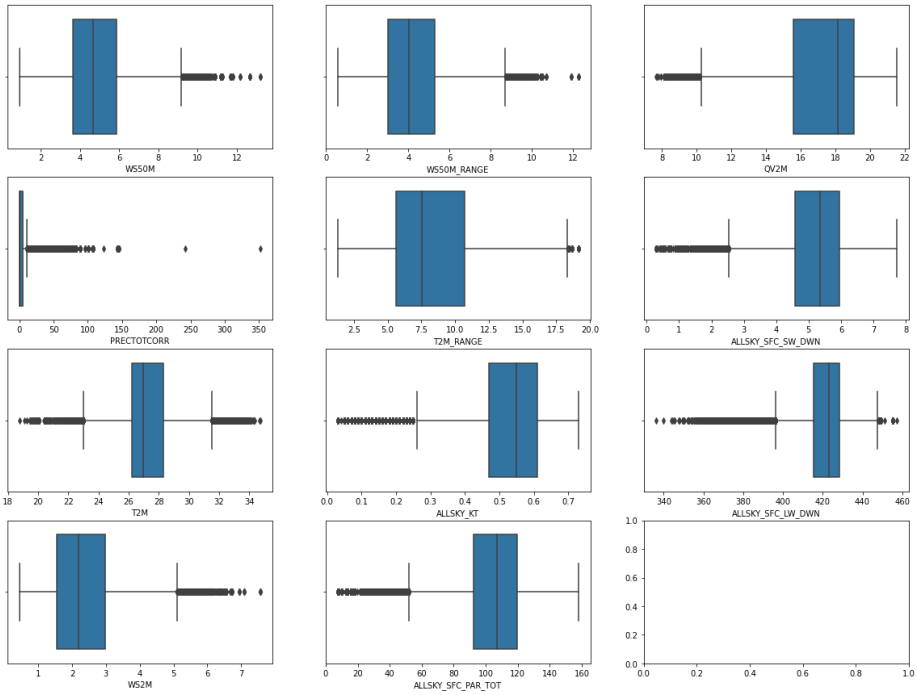
Khảo sát các giá trị bị nhiễu (outlier), chúng tôi dựa vào đồ thị hộp và phương pháp IQR để phát hiện chúng.

IQR có công thức:

$$IQR = Q_3 - Q_1 \quad (1)$$

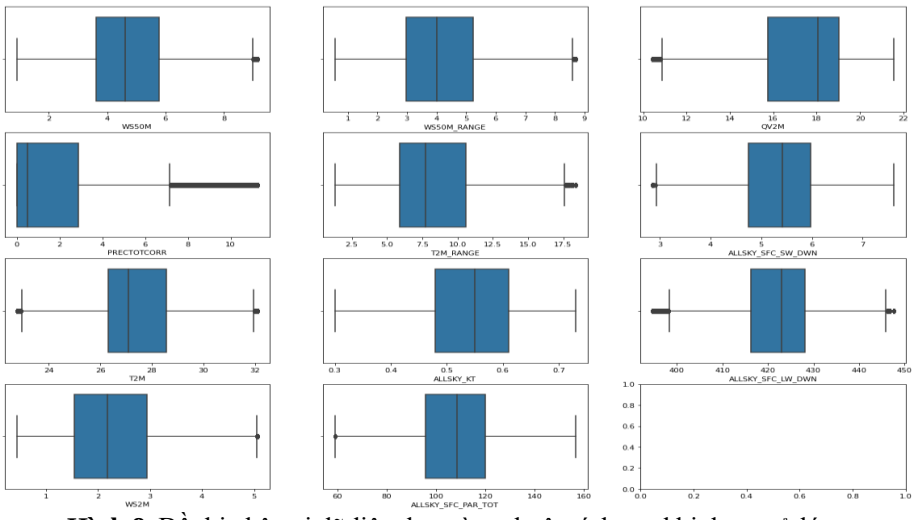
$$Lower\ Bound = Q_1 - 1{,}5 * IQR \quad (2)$$

$$Higher\ Bound = Q_3 + 1{,}5 * IQR \quad (3)$$



Hình 7. Đồ thị tứ phân vị dữ liệu theo từng thuộc tính

Với các giá trị thấp hơn Lower Bound và cao hơn Higher Bound là các giá trị nhiễu (outlier), được xử lý bằng cách xóa các giá trị nằm ngoài khoảng này để đảm bảo tính chính xác của dữ liệu.



Hình 8. Đồ thị phân vị dữ liệu theo từng thuộc tính sau khi được xử lý

3.3 Codebook

Bảng 1. Codebook của bộ dữ liệu

Thông tin	Nội dung
Nguồn thu thập và cách thức thu thập	Nguồn: https://power.larc.nasa.gov/data-access-viewer/ Nhập các thông số phù hợp và tập hợp các thuộc tính cần thiết để thu thập dữ liệu.
Tên bộ dữ liệu	Bộ dữ liệu các yếu tố thời tiết Thành phố Hồ Chí Minh
Kích thước bộ dữ liệu	216575 dòng dữ liệu
Nội dung bộ dữ liệu	Bộ dữ liệu chứa dữ liệu về các yếu tố thời tiết của 24 quận/huyện thuộc Thành phố Hồ Chí Minh được ghi nhận theo từng ngày từ 01/01/1990 đến 31/03/2021
Số thuộc tính	15
Thông tin các thuộc tính	1. **YEAR:** Năm - *Kiểu dữ liệu:* ***Integer***, khoảng giá trị trong bộ dữ liệu [1990,2021] 2. **MO:** Tháng – *Kiểu dữ liệu:* ***Integer***, khoảng giá trị trong bộ dữ liệu [1,12] 3. **DY**: Ngày – *Kiểu dữ liệu:* ***Integer***, khoảng giá trị trong bộ dữ liệu [1,31] 4. **WS50M**: *Wind Speed at 50 Meters (m/s)*, Tốc độ gió trung bình ở độ cao 50 mét so với mặt đất. - *Kiểu dữ liệu:* ***Real***. 5. **WS50M_RANGE**: *Wind Speed at 50 Meters Range (m/s)*, Biên độ tốc độ gió tối thiểu và tối đa mỗi giờ ở độ cao 50 mét so với mặt đất – *Kiểu dữ liệu:* ***Real***. 6. **QV2M**: *Specific Humidity at 2 Meters (g/kg)*, Tỷ số giữa khối lượng hơi nước với tổng khối lượng không khí ở độ cao 2m (g nước / kg tổng khối lượng không khí) – *Kiểu dữ liệu:* ***Real.*** 7. **PRECTOTCORR:** *Precipitation Corrected (mm/day)*, Giá trị trung bình sai lệch của tổng lượng mưa trên bề mặt trái đất trong khối nước (bao gồm cả hàm lượng nước trong tuyết) – *Kiểu dữ liệu:* ***Real***. 8. **T2M_RANGE**: *Temperature at 2 Meters Range (C)*, Biên độ nhiệt không khí (bầu khô) tối thiểu và tối đa mỗi giờ ở độ cao 2 mét so với mặt đất - *Kiểu dữ liệu:* ***Real.*** 9. **ALLSKY_SFC_SW_DWN**: *All Sky Surface Shortwave Downward Irradiance (kW-hr/m^2/day)*, Tổng sự cố bức xạ mặt trời (trực tiếp cộng với khuếch

tán) trên mặt phẳng nằm ngang ở bề mặt trái đất trong mọi điều kiện bầu trời – *Kiểu dữ liệu:* ***Real***.

10**. T2M**: *Temperature at 2 Meters (C),* Nhiệt độ không khí (bầu khô) trung bình ở độ cao 2m so với mặt đất – *Kiểu dữ liệu:* ***Real***.

11. **ALLSKY_KT**: *All Sky Insolation Clearness Index (dimensionless)*, Độ trong của khí quyển, Tỉ số giữa sự cách nhiệt toàn bộ bầu trời truyền qua bầu khí quyển tới bề mặt trái đất với tổng sự cố bức xạ mặt trời trung bình của đỉnh khí quyển – *Kiểu dữ liệu:* ***Real***.

12. **ALLSKY_SFC_LW_DWN**: *All Sky Surface Longwave Downward Irradiance (W/m^2)*, Cường độ bức xạ hồng ngoại ngang từ bầu trời hay bức xạ hồng ngoại nhiệt hướng xuống trong mọi điều kiện bầu trời tới mặt phẳng nằm ngang bề mặt trái đất – *Kiểu dữ liệu:* ***Real***.

13. **WS2M**: *Wind Speed at 2 Meters (m/s),* Tốc độ gió trung bình ở độ cao 2 mét so với mặt đất – *Kiểu dữ liệu:* ***Real***.

14. **ALLSKY_SFC_PAR_TOT**: *All Sky Surface PAR Total (W/m^2)*, Tổng sự cố Bức xạ hoạt động quang hợp (PAR) trên mặt phẳng nằm ngang ở bề mặt trái đất trong mọi điều kiện bầu trời – *Kiểu dữ liệu:* ***Real***

15. **LOC**: *Tên Quận/huyện* thuộc Thành phố Hồ Chí Minh – *Kiểu dữ liệu:* ***String***.

Thông tin tác giả

1. **Nguyễn Hoàng Minh**
   Email: ***20521609@gm.uit.edu.vn***
2. **Nguyễn Minh Tiến**
   Email: ***20522010@gm.uit.edu.vn***
3. **Nguyễn Đức Hiển**
   Email: ***20521307@gm.uit.edu.vn***
4. **Tạ Nhật Minh**
   Email: ***20521614@gm.uit.edu.vn***

4 Ứng dụng

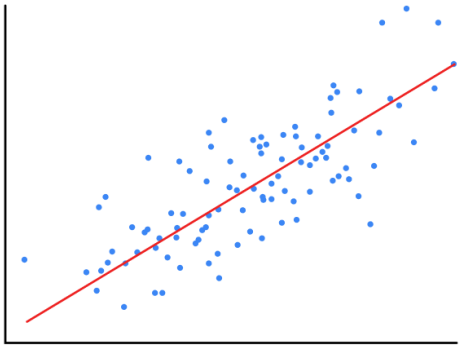
4.1 Phân chia dữ liệu

Bộ dữ liệu của chúng tôi có kích thước khá lớn (216575 dòng dữ liệu) nên chúng tôi quyết định chọn dữ liệu của Thủ Đức để làm tập dữ liệu để huấn luyện và kiểm thử cho các mô hình.

Dữ liệu được chia thành tập huấn huyện và tập kiểm thử theo tỉ lệ 4:1. Chúng tôi sẽ huấn luyện mô hình và sử dụng các phương pháp để tìm mối quan hệ giữa biến độc lập và biến phụ thuộc, mục đích là để tránh overfitting.

4.2 Xây dựng mô hình và đánh giá

Linear Regression. (hồi quy tuyến tính) được sử dụng để tìm mối quan hệ tuyến tính giữa các mục tiêu và một hoặc nhiều yếu tố được dự báo. Có hai loại hồi quy tuyến tính – đơn tuyến tính và đa tuyến tính. Vì hồi quy tuyến tính đơn giản hữu ích cho việc tìm kiếm mối quan hệ chỉ giữa hai biến liên tục trong khi hồi quy nhiều lần là một kỹ thuật thống kê sử dụng một số biến giải thích để dự đoán kết quả của một biến phản hồi. Mục tiêu của hồi quy đa tuyến tính (MLR) là mô hình hóa mối quan hệ tuyến tính giữa các biến giải thích (độc lập) và biến phản hồi (phụ thuộc), tức là Nhiệt độ kỳ vọng. Đó là lý do tại sao chúng tôi đã sử dụng hồi quy nhiều tuyến tính ở đây.



Hình 9. Mô hình Linear Regression

Công thức chung của hồi quy tuyến tính:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ ... + \beta_p x_{ip} + \ \epsilon \qquad (4)$$

y_i : Biến phụ thuộc (Biến mục tiêu)
x_i : Biến độc lập (Biến dự đoán)
β_0 : Điểm cắt của đường thẳng hồi quy và trục
β_p : Hệ số góc.
ε : Sai số.

Support Vector Regression. (SVR) là mô hình cung cấp một sự linh hoạt để xác định mức độ lỗi có thể chấp nhận được trong mô hình và sẽ tìm một đường thích hợp (hoặc là một siêu phẳng ở các kích thước cao hơn) để phù hợp với dữ liệu. SVR sử dụng nguyên tắc tương tự như SVM, nhưng chủ yếu với các vấn đề hồi quy. Có nghĩa là tìm một hàm gần đúng ánh xạ từ miền đầu vào tạo thành các số thực trên cơ sở một mẫu huấn luyện.

Hàm giả thuyết cho SVR:

$$Y = wx + b \quad (6)$$

Khi đó, ta có các phương trình của đường ranh giới quyết định:

$$wx + b = +\epsilon;\ wx + b = -\epsilon \quad (7)$$

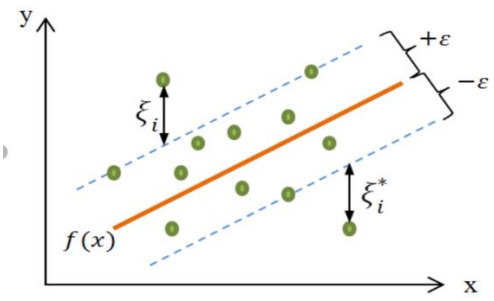
Xét bài toán tối ưu có ràng buộc:

$$((w, b) = \arg \left[\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \quad \xi_i\right] \quad (8)$$

Do đó, bất kỳ siêu phẳng nào của SVR đều phải thoả điều kiện sau:

$$-\epsilon < Y - wx+b < +3\epsilon \quad (9)$$

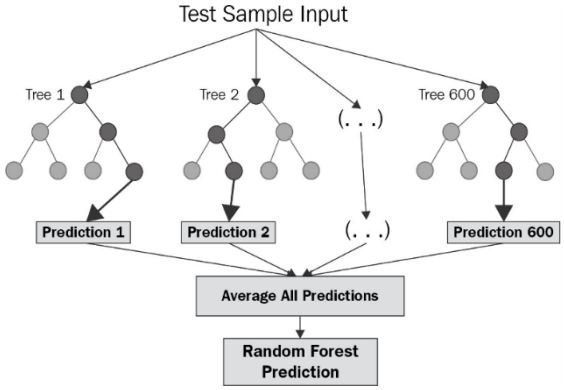
Do đó, ở đây chỉ lấy những điểm nằm trong đường ranh giới quyết định và có tỷ lệ lỗi ít nhất hoặc nằm trong Margin of Tolerance (tạm dịch là phạm vi dung sai). Từ đó sẽ tạo điều kiện thuận lợi để đào tạo một mô hình phù hợp hơn.



Hình 10. Mô hình Support Vector Regression

Random Forest Regression. là một thuật toán học máy có giám sát sử dụng phương pháp học máy theo nhóm để phân loại và hồi quy. Bagging Random Forest không giống như Boosting Random Forest. Nó hoạt động bằng cách xây dựng vô số cây quyết định tại thời điểm huấn luyện và dự đoán trung bình (hồi quy) của các cây riêng lẻ.

Mọi cây quyết định đều có phương sai cao, nhưng khi chúng ta kết hợp tất cả chúng lại với nhau song song thì phương sai của kết quả sẽ thấp vì mỗi cây quyết định được đào tạo hoàn hảo trên dữ liệu mẫu cụ thể và do đó kết quả đầu ra không phụ thuộc vào một cây quyết định mà nhiều cây quyết định. Trong trường hợp có vấn đề phân loại, kết quả cuối cùng được thực hiện bằng cách sử dụng bộ phân loại biểu quyết đa số. Trong trường hợp của một bài toán hồi quy, đầu ra cuối cùng là giá trị trung bình của tất cả các đầu ra. Phần này được gọi là Aggregation.



Hình 11. Mô hình Random Forest Regression

4.3 Thang đo đánh giá mô hình

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \quad (10)$$

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \quad (11)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_i)^2}{n}} \quad (12)$$

$$R^2 = 1 - \frac{RSS}{TSS} \text{ (R2 -score)} \quad (13)$$

MAE, MSE, RMSE là các hàm mất mát do độ sai lệch giữa giá trị dự đoán so với các giá trị thực. R^2 tỉ lệ thuận với mức giải thích của biến độc lập với sự thay đổi của biến phụ thuộc. R^2 cũng thể hiện độ phù hợp của mô hình hồi quy với tập dữ liệu.

4.4 Kết quả và đánh giá

Với mục đích đánh giá hiệu suất các mô hình máy học khác nhau, chúng tôi đã sử dụng ba thuật toán máy học phổ biến gồm: Linear Regression, Support Vector Regression và Random Forest Regression.

Bảng 2. Bảng kết quả độ đo của các mô hình với tham số mặc định

	MAE	MSE	RMSE	R2-score
Linear Regression	0.58	0.58	0.76	0.823
Support Vector Regression	1.19	2.31	1.52	0.293
Random Forest Regression	0.56	0.62	0.79	0.809

Bảng 3. Bảng kết quả độ đo của các mô hình với tham số được hiệu chỉnh

	MAE	MSE	RMSE	R2-score
Linear Regression	0.58	0.58	0.76	0.823

Support Vector Regression	0.50	0.46	0.68	0.858
Random Forest Regression	0.56	0.61	0.78	0.814

Chúng tôi tiếp tục tinh chỉnh tham số của mỗi phương pháp với phương pháp GridSearchCV để tăng hiệu suất mô hình và nhận được kết quả thể hiện ở bảng 3 cùng với bộ tham số thu được như sau:

Thuật toán Support Vector Regression với tham số 'C': 1000.0, 'gamma': 0.001, 'kernel': 'rbf'.

Thuật toán Random Forest Regression với tham số 'bootstrap': False, 'max_depth':15, 'max_feature': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split':10, 'n_estimators': 150.

Nhìn chung, qua quá trình huấn luyện mô hình, chúng tôi thấy Support Vector Regression là thuật toán tốt nhất để dự đoán nhiệt độ với kết quả đạt được là 85.8%. Bên cạnh đó, hai mô hình Linear Regression và Random Forest Regression cũng cho ra kết quả dự đoán tương đối cao (82.3% và 81.4%). Điều này cho thấy rằng bộ dữ liệu của chúng tôi có thể được áp dụng vào các mô hình dự đoán máy học.

5 Kết luận, phương hướng

5.1 Kết quả đạt được

Báo cáo hoàn thành được những mục tiêu đề ra ban đầu là xây dựng được bộ dữ liệu các yếu tố thời tiết tại thành phố Hồ Chí Minh. Thao tác, xây dựng được các phương pháp tiền xử lý dữ liệu. Đồng thời xây dựng được các mô hình dự đoán nhiệt độ từ bộ dữ liệu đã thu thập và xử lý. Cũng như tìm hiểu, sử dụng các phương pháp phân tích, đánh giá các giá trị của bộ dữ liệu.

5.2 Khó khăn

Trong quá trình thu thập dữ liệu, chúng tôi có gặp một số khó khăn về việc tìm kiếm nguồn dữ liệu tin cậy. Ngoài ra, quá trình tiền xử lý cũng đã gây một số rắc rối trong việc huấn luyện mô hình do bài toán đòi hỏi kiến thức nâng cao, tập dữ liệu thô có nhiều giá trị bất thường và kích thước dữ liệu lớn.

5.3 Hướng phát triển

Dự báo thời tiết là một trong những lĩnh vực nghiên cứu quan trọng do khả năng ứng dụng của nó trong các vấn đề thực tế như đo lường, nghiên cứu nông nghiệp, hàng không, sức khỏe,…

Với như cầu ngày càng tăng về thông tin dự báo thời tiết đáng tin cậy và chính xác hơn, việc này giúp xác định phương hướng phát triển của bộ dữ liệu sau này. Chúng tôi sẽ tiếp tục thu thập và phát triển bộ dữ liệu nhằm phân tích sâu hơn về tầm ảnh hưởng của các yếu tố thời tiết. Từ đó xây dựng các mô hình kết hợp để nâng cao dự đoán chính xác với bộ dữ liệu. Kết hợp với các tổ chức hay hỗ trợ Chính phủ về vấn đề dự báo thời

tiết để ứng dụng vào các lĩnh vực như đời sống, nông nghiệp, sức khỏe, du lịch, giao thông vận tải,…

References

1. POWER Data Access Viewer v2.0.0, https://power.larc.nasa.gov/data-access-viewer/