ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH VÀ XÂY DỰNG MÔ HÌNH DỰ BÁO THỜI TIẾT CÁC QUẬN HUYỆN THÀNH PHỐ Ở THÀNH PHỐ HỒ CHÍ MINH

Nhóm 7			
Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
1	Hồ Quang Lâm	21521049	HTTT
2	Lê Minh Chánh	21521882	HTTT
3	Trần Thị Thanh Trúc	21522722	MTT&TTDL

TP. HỒ CHÍ MINH – 12/2024

1. GIỚI THIỆU

Trong phạm vi đồ án môn học, nhóm sẽ sử dụng các kĩ thuật xử lí dữ liệu, phân tích, trực quan dữ liệu về các thông số thời tiết tại 22 quận, huyện, thành phố của thành phố Hồ Chí Minh trong vòng 10 năm trở lại đây. Mục đích để tìm ra các yếu tố ảnh hưởng đến khả năng mưa vào ngày tiếp theo và xây dựng mô hình dự báo. Khi có bộ dữ liệu, nhóm đã tiến hành tiền xử lí dữ liệu bằng các công cụ như Pandas, Numpy và sử dụng các thư viện hỗ trợ cho trực quan hoá và phân tích như Matplotlib, Seaborn, Scipy.stats. Cuối cùng sẽ sử dụng các thuật toán RandomForestClassifier, XGBClassifier, LGBMClassifier từ các thư viện có sẵn và tìm ra các thông số của thuật toán sao cho phù hợp với bộ dữ liệu để thu được mô hình có độ chính xác cao. Kết quả thu được là các đánh giá về các thông số thời tiết, các thuộc tính quan trọng ảnh hưởng lớn đến khả năng mưa và mô hình khả thi để xây dựng với bộ dữ liệu là RandomRorest với các thông số đo độ chính xác đều trên 86%.

Điểm mạnh của đề tài là bộ dữ liệu do nhóm tự thu thập, phân tích và thiết kế, không tham khảo đề tài nào khác. Để thu thập được dữ liệu thời tiết, nhóm đã tiến hành lấy thông tin toạ độ của 22 quận huyện thành phố thông qua OpenStreetMap[2], sau đó dùng các thông tin lấy được để thu thập dữ liệu thời tiết từ Open-Meteo Historical Weather API[1], dữ liệu thu thập từ API này khá đầy đủ, góp phần giúp giảm một số bước và thời gian phân tích.

2. MÔ TẢ BỘ DỮ LIỆU

2.1. Giới thiệu tóm tắt bộ dữ liệu

- **Tên bộ dữ liệu**: **HCMCity_weather.csv** (cập nhập lúc 13h00 ngày 26/10/2024).
- **Nguồn bộ dữ liệu**: Bộ dữ liệu được thu thập từ Open-Meteo Historical Weather API [1] dự trên dữ liệu toạ độ của 22 quận, huyện, thành phố của Thành phố Hồ Chí Minh được thu thập từ trang web Nominatim | OpenStreetMap [2].
- **Ý nghĩa bộ dữ liệu**: Bộ dữ liệu chứa các thông tin về đặc điểm thời tiết theo từng ngày tại 22 quận, huyện, thành phố của Thành phố Hồ Chí Minh trong vòng 10 năm gần đây (từ **25/10/2014** đến **25/10/2024**).
- **Kích thước**: 80.388 dòng × 29 thuộc tính.

2.2. Thống kê bộ dữ liệu và mô tả các thuộc tính

Bảng 1: Bảng thống kê bộ dữ liệu

Thống kê bộ dữ liệu		
Số dòng		80.388 dòng dữ liệu
Số cột		29 cột thuộc tính
Số biến phân loại		4
Số biến liên tục		25
Số dữ liệu bị khuyết trên từng thuộc tính	temperature_2m_mean, apparent_temperature_mean, sunshine_duration, precipitation_sum, rain_sum, snowfall_sum, wind_direction_10m_dominant, shortwave_radiation_sum, et0_fao_evapotranspiration	44
	dew_point_2m, pressure_msl, surface_pressure, cloud_cover, weather_code, temperature_2m_max, temperature_2m_min, apparent_temperature_max, apparent_temperature_min, wind_speed_10m_max, wind_gusts_10m_max, relative_humidity_2m	22
Số mẫu có dữ liệu khuyết		44

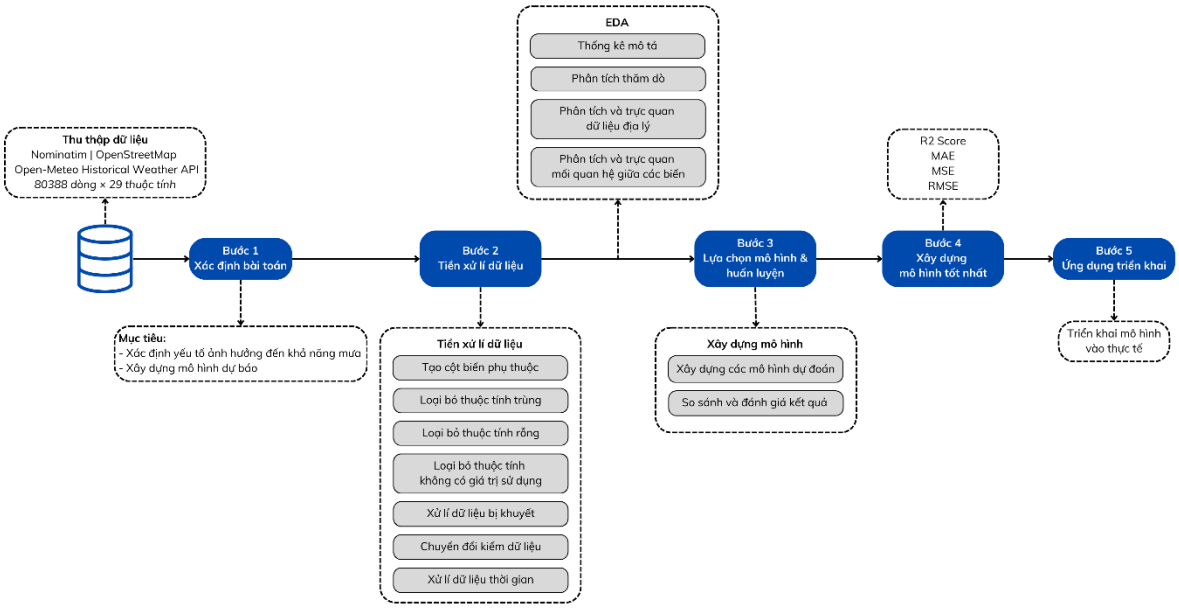
Bảng 2: Bảng mô tả các thuộc tính của bộ dữ liệu

Mô tả thuộc tính			
STT	**Thuộc tính**	**Kiểu**	**Mô tả**
1	district	String	Tên quận, huyện, thành phố trực thuộc
2	latitude	float	Vĩ độ trung tâm đơn vị hành chính

3	longitude	float	Kinh độ trung tâm đơn vị hành chính
4	date	Date	Ngày ghi nhận thông tin
5	relative_humidity_2m	float	Trung bình độ ẩm tương đối ở độ cao 2m (%)
6	dew_point_2m	float	Trung bình nhiệt độ điểm sương ở độ cao 2m (°C)
7	pressure_msl	float	Trung bình áp suất khí quyển tại mực biển (hPa)
8	surface_pressure	float	Trung bình áp suất đo trực tiếp tại mặt đất (hPa)
9	cloud_cover	float	Tỉ lệ độ che phủ mây cao nhất trong ngày(%)
10	weather_code	int	Mã WMO
10	temperature_2m_max	float	Nhiệt độ không khí tối đa ở độ cao 2m (°C)
11	temperature_2m_min	float	Nhiệt độ không khí tối thiểu ở độ cao 2m (°C)
12	temperature_2m_mean	float	Nhiệt độ không khi trung bình ở độ cao 2m (°C)
13	apparent_temperature_max	float	Nhiệt độ tối đa cảm nhận được (°C)
14	apparent_temperature_min	float	Nhiệt độ tối thiểu cảm nhận được (°C)
15	apparent_temperature_mean	float	Nhiệt độ trung bình cảm nhận được (°C)
16	sunrise	Time	Thời gian mặt trời mọc
17	sunset	Time	Thời gian mặt trời lặn
18	daylight_duration	float	Thời lượng ánh sáng (giây), từ bình minh đến hoàng hôn
19	sunshine_duration	float	Thời lượng ánh nắng (giây) khi bức xạ > 120W/m^2 (giây)
20	precipitation_sum	float	Tổng lượng mưa (mưa + mưa rào + tuyết) (mm)
21	rain_sum	float	Tổng lượng mưa (mm)
22	snowfall_sum	float	Tổng lượng tuyết rơi (cm)
24	precipitation_hours	float	Số giờ mưa (giờ)
25	wind_speed_10m_max	float	Tốc độ gió tối đa trung bình độ cao 10m (km/h)
26	wind_gusts_10m_max	float	Tốc độ gió giật tối đa ở độ cao 10 mét (km/h)
27	wind_direction_10m_dominant	float	Hướng gió chủ đạo (0° đến 360°)
28	shortwave_radiation_sum	float	Tổng bức xạ mặt trời (MJ/m^2)
29	et0_fao_evapotranspiration	float	Chỉ số thoát hơi nước chuẩn ET_0 (mm)

3. PHƯƠNG PHÁP PHÂN TÍCH

Sau khi thu thập thành công bộ dữ liệu, nhóm tiến hành phân tích với quy trình các bước theo sơ đồ minh hoạ bên dưới:



Hình 1. Quy trình phân tích dữ liệu

4. TIỀN XỬ LÍ DỮ LIỆU

Nhóm đã tiến hành tiền xử lí dữ liệu thô qua 5 bước như sau:

- **Bước 1:** Tạo cột biến phụ thuộc (RainTomorrow) để làm mục tiêu cho bài toán dự báo khả năng mưa vào ngày tiếp theo mà nhóm đã đặt ra bằng cách kiểm tra lượng mưa thực tế của ngày hôm sau lớn hơn 1mm.
- **Bước 2:** Loại bỏ cột trùng: Loại bỏ cột rain_sum vì toàn bộ giá trị trùng lặp với cột precipitation_sum.
- **Bước 3:** Xoá các cột thuộc tính không có giá trị sử dụng trong bài toán:
    - Cột **sunrise** và **sunset** không chứa dữ liệu thời gian, chỉ chứa số 0.
    - Vì thời tiết TP. Hồ Chí Minh là ở đới khí hậu nhiệt đới nên cột **snowfall_sum** chỉ chứa giá trị 0.
- **Bước 4:** Xử lí dữ liệu bị khuyết: bộ dữ liệu chỉ có 66 dòng dữ liệu bị khuyết
    - Có 19 cột bị khuyết ở 44 dòng vì đây là 2 cột ở 22 địa điểm vào 2 ngày cuối cùng vì hệ thống chưa cập nhập tại thời điểm thu thập.

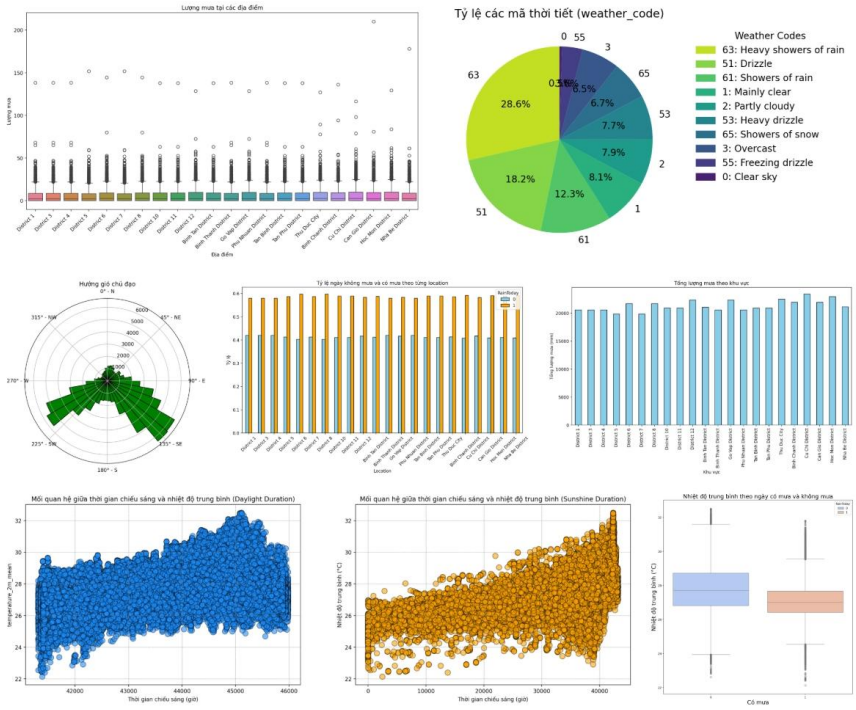
- Có 66 dòng của cột RainTomorrow bị khuyết vì thiếu dữ liệu ở cột precipitation_sum khi tạo.

Vì số dòng dữ liệu khuyết ít nên nhóm sẽ thực hiện xoá dòng khuyết.

- **Bước 5:** Chuyển đổi kiểu dữ liệu:
  - Chuyển đổi kiểu dữ liệu thời gian: dạng dữ liệu dd/mm/yyy.
  - Chuyển đổi kiểu dữ liệu số của 2 cột RainTomorrow và weather_code.

5. PHÂN TÍCH THĂM DÒ

Nội dung của phần này sẽ bao gồm trực quan dữ liệu của tất cả các trường theo thời gian, sự phân bố giá trị và quan hệ tương quan giữa các biến. Từ đó nhóm hiểu được các đặc trưng của bộ dữ liệu.

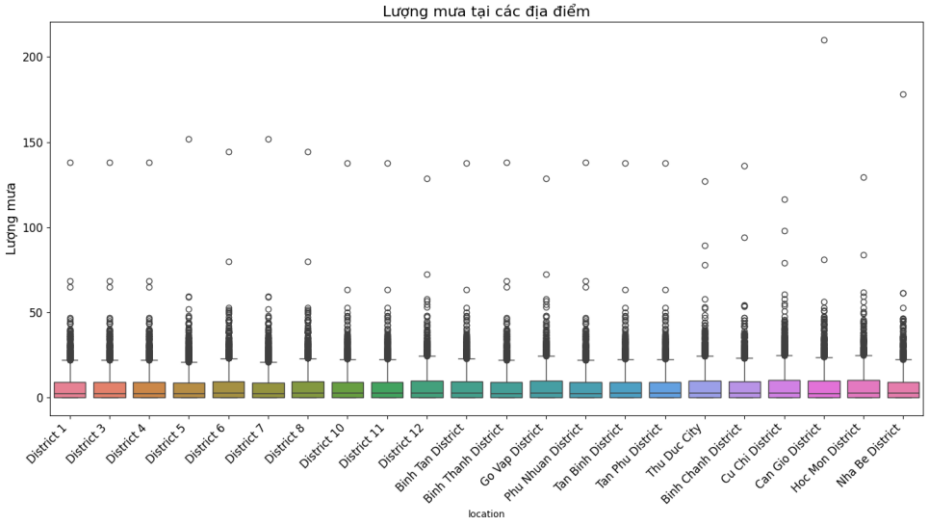


Hình 2. Dashboard các biểu đồ các yếu tố ảnh hưởng đến phân tích

5.1. Thống kê mô tả

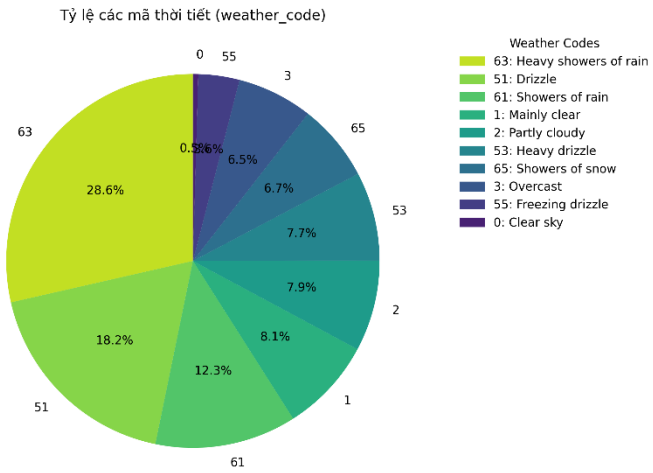
Trong thời gian 10 năm, sự phân bố lượng mưa tại các địa điểm tương đối đồng đều nhau được thể hiện qua các boxplot thấy được các giá trị trung bị, và các thông số khác của boxplot. Nhưng có một vài trường hợp lượng mưa tăng đột biến tại hai khu

vực rìa thành phố là huyện Cần Giờ và huyện Nhà Bè, có thể có trường hợp do ảnh hưởng mạnh của bão tại 2 khu vực gần biển này.



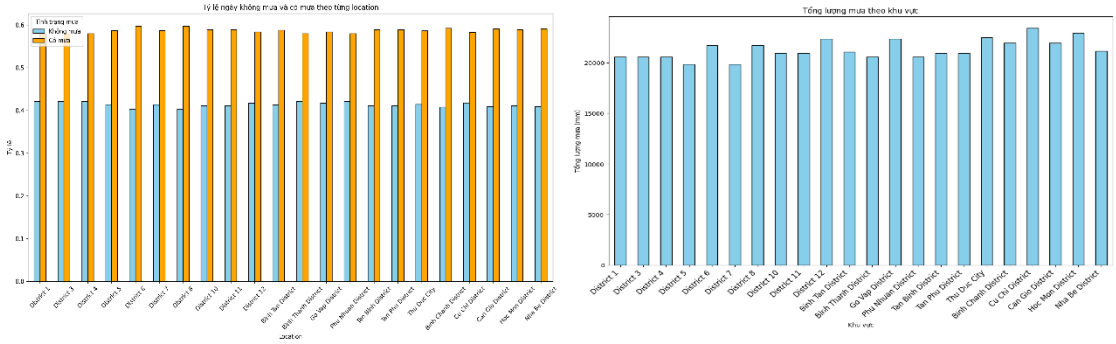
Hình 3. Biểu đồ phân tố lượng mưa của 22 quận huyện thành phố

Trong bộ dữ liệu, dữ liệu thời tiết tại từng địa điểm có số lượng giống nhau, sau khi tổng hợp mã thời tiết đánh giá theo từng ngày cho thấy **mã 63 – mưa lớn** là loại thời tiết phổ biến nhất. Theo WeatherCode có trong bộ dữ liệu, các giá trị từ 0 – 3 là không mưa chiếm số lượng rất thấp, các giá trị lớn hơn 51 là biểu thị cho thời tiết có mưa theo quy định của WMO chiếm hầu hết các trường hợp.



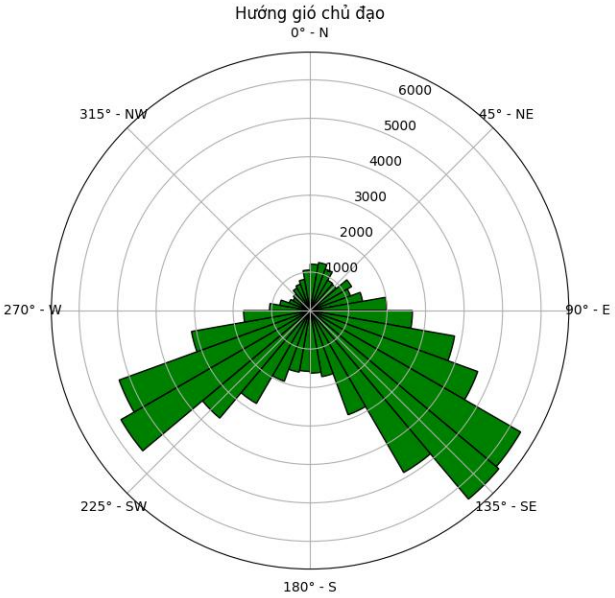
Hình 4. Biểu đồ tỉ lệ các loại thời tiết của thành phố

Diện tích thành phố Hồ Chí Minh là khá nhỏ, nên lượng mưa và số ngày có mưa cũng gần đồng đều nhau giữa từng địa điểm.



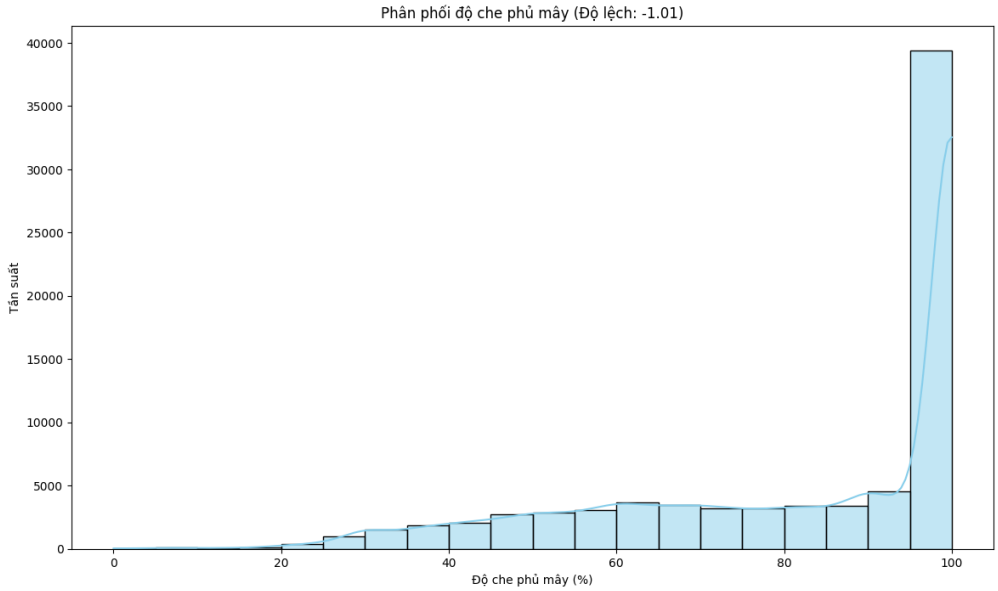
Hình 5. Biểu đồ tỉ lệ số lượng ngày mưa, không mưa và tổng lượng mưa từng khu vực

Về hướng gió chủ đạo ở khu vực toàn thành phố thì vẫn là gió theo mùa với 2 mùa gió chính là Đông Bắc và Tây Nam, và theo biểu đồ dạng Polar ta có thể thấy được là tầng xuất của gió mùa Tây Nam nhiều hơn so với gió mùa Đông Bắc. Như phần trên đã trình bày, thời tiết mưa ở TP. Hồ Chí Minh chiếm nhiều hơn thời tiết không mưa vì sự ảnh hưởng mạnh của gió Tây Nam, mang theo lượng mưa lớn và gây ra mùa mưa cho miền Nam.



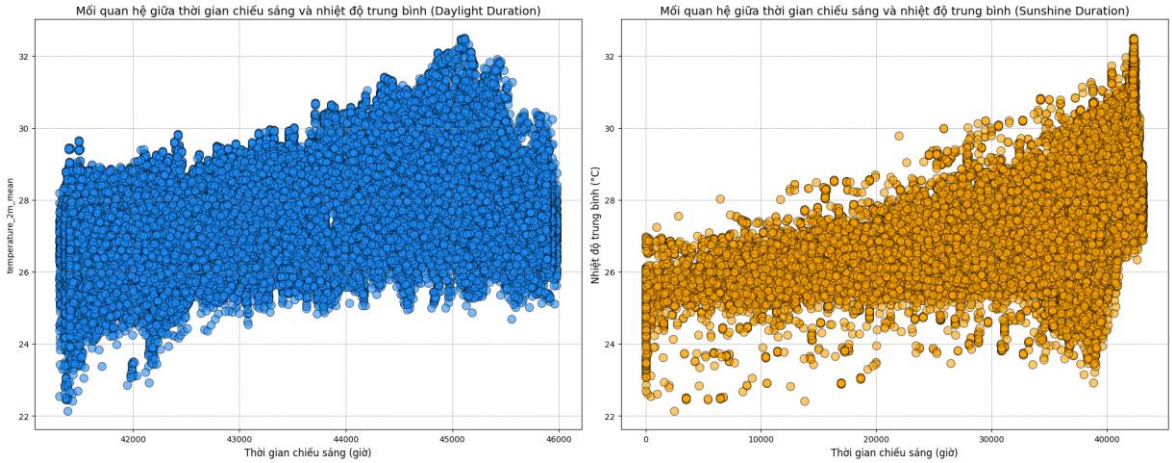
Hình 6. Biểu đồ Polar thể hiện hướng gió chính của thành phố

Thống kê bộ dữ liệu chỉ hơn 80.000 dòng nhưng có khoảng 50% trong số đó có lượng mây che phủ là 100%, biểu đồ lệch trái đáng kể với độ lệch là **-1.01** và các con số còn lại không đánh kể, để đánh giá được ảnh hưởng của mây cần xem xét thêm một số yếu tố khác như ô nhiễm không khí…



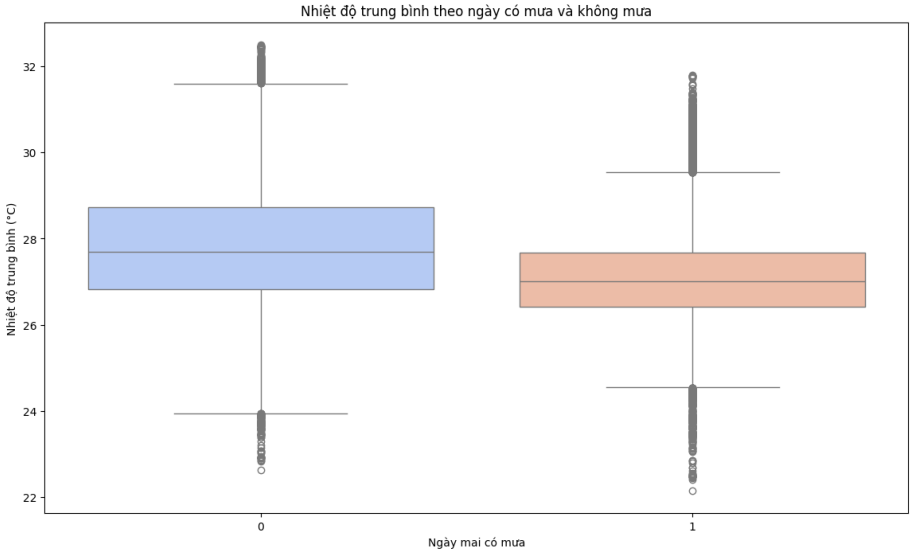
Hình 7. Biểu đồ phân phối độ che phủ mây

Lượng mây che phủ thành phố khá lớn dẫn đến có nhiều khoảng thời gian lượng ánh nắng bức xạ $> 120 W/m^2$ từ mặt trời chiếu có giá trị bằng 0 nhưng nhiệt độ vẫn cao trên 26°C ở nhiều ngày, và ngày thấp nhất ghi nhận được xấp xỉ 22°C.



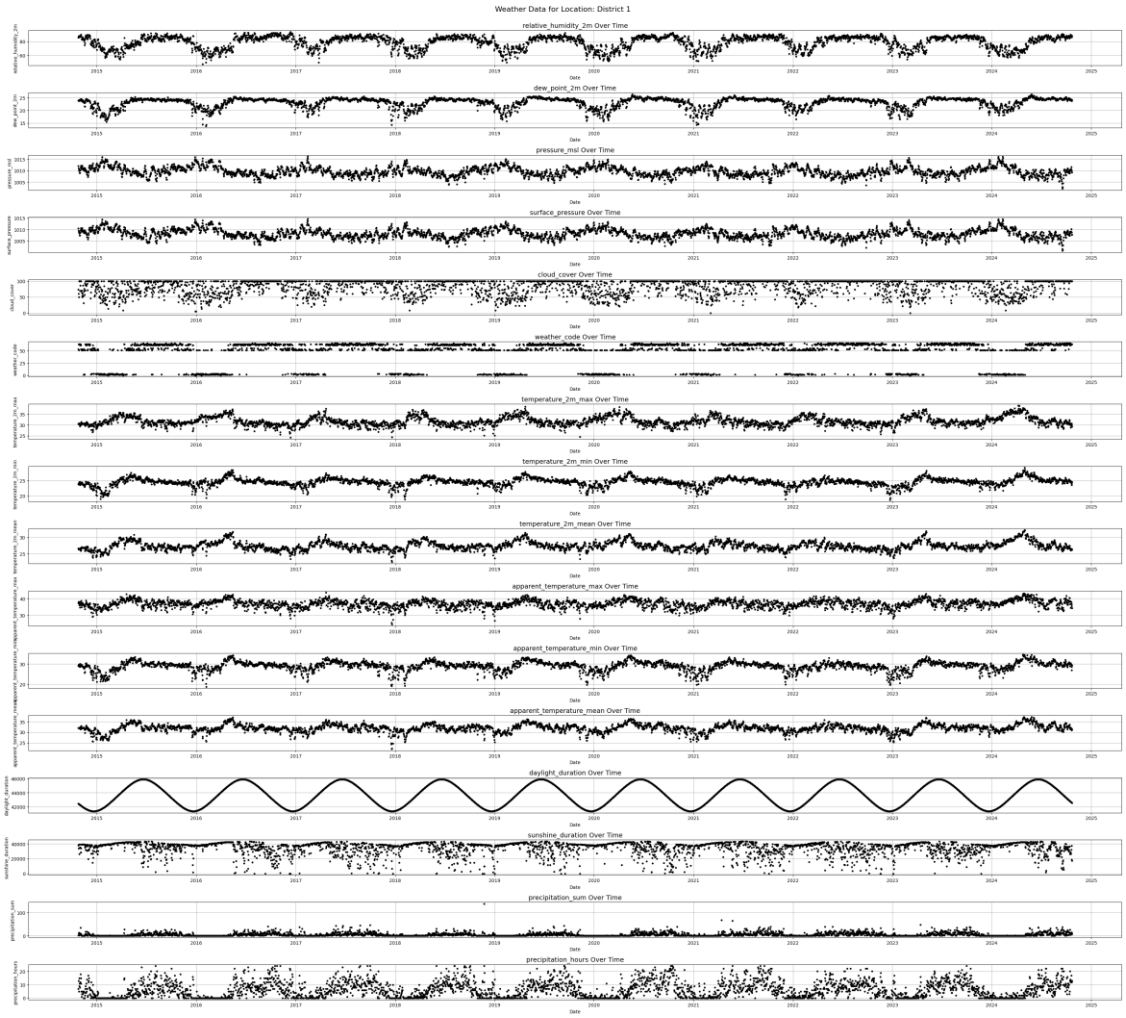
Hình 8. Mối quan hệ giữa nhiệt độ trung bình với thời gian chiếu sáng mặt trời

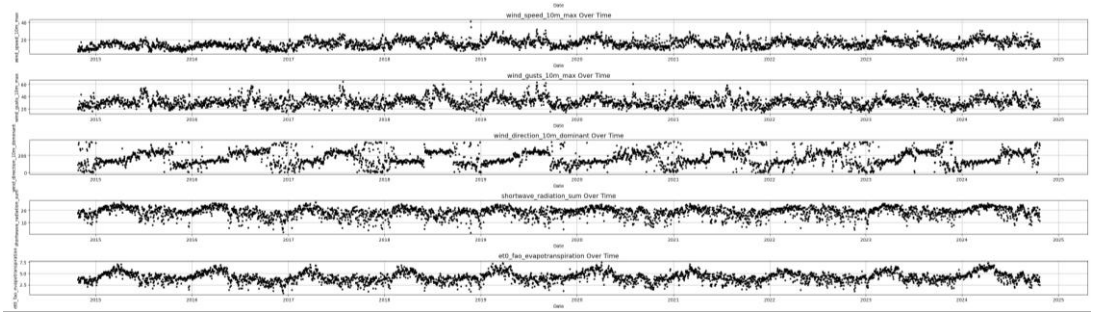
Về phân bố mức nhiệt độ tại những ngày không mưa và có mưa, ngày có mưa ta có thể thấy được mức nhiệt trung bình thấp hơn so với ngày không mưa một chút, nhưng đáng chú ý là vào những ngày có mưa, có thể là lượng mưa không đủ lớn nên mức nhiệt vẫn ở khá cao, có ngày trên 30°C.



Hình 9. Biểu đồ phân phối nhiệt độ trung bình theo ngày có mưa và không mưa

5.2. Phân tích và trực quan dữ liệu thời gian

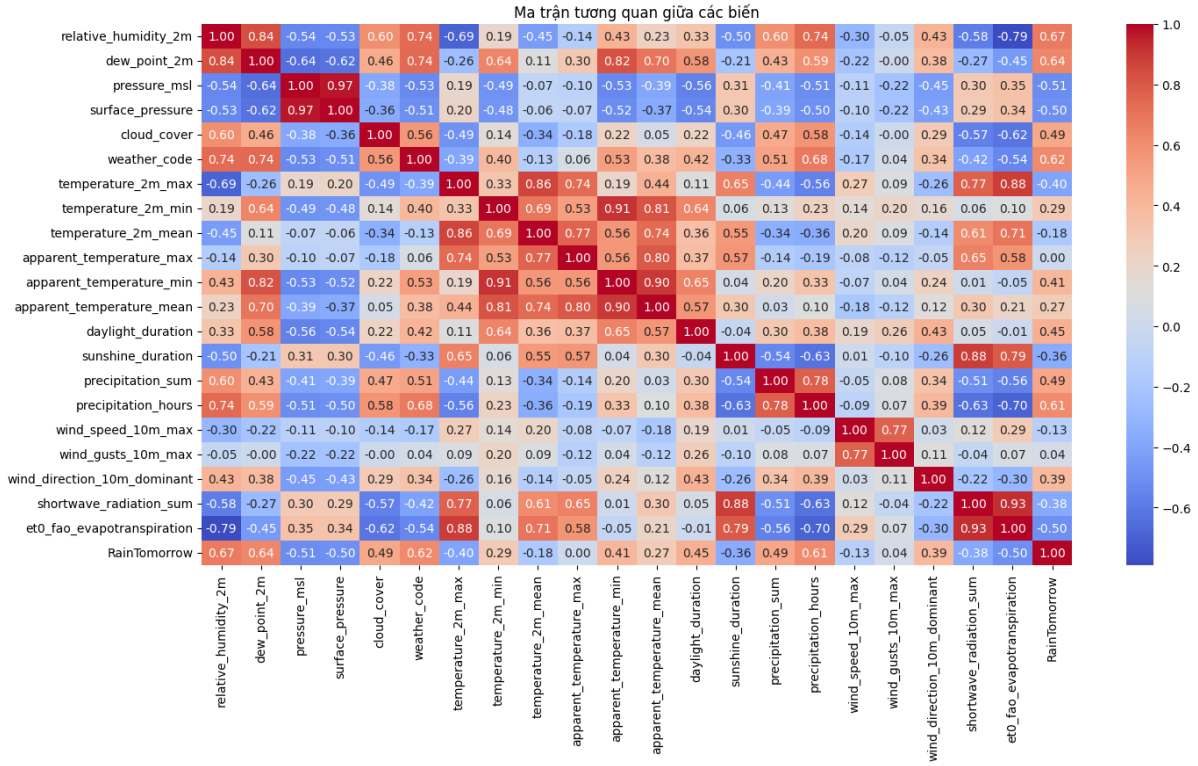




Hình 10. Biểu đồ phân phối các thông số thời tiết theo thời gian

Biểu đồ thể hiện các thông số theo thời gian 10 năm, chúng ta có thể thấy được thời tiết của TP. Hồ Chí Minh tương đối ổn định, ít bị ảnh hưởng bởi các thiên tai như bão lũ hay không khí lạnh từ phía Bắc. Các thông số hầu như là có xu hướng tuần hoàn theo 2 mùa chính là mùa mưa và mùa khô.

5.3. Phân tích và trực quan mối quan hệ tương quan giữa các biến



Hình 11. Ma trận tương quan mối quan hệ giữa các biến

Từ ma trận tương quan, chúng ta có thể rút ra một số nhận xét như sau:

- Một số thuộc tính có mối quan hệ mạnh với nhau, chúng ta chỉ nên giữ lại một trong số chúng để tránh bị trường hợp bị đa cộng tuyến (Multicollinearity) – các biến độc lập sẽ không còn thật sự là độc lập với nhau.

- Giữ lại các biến có mối quan hệ mạnh với biến phụ thuộc (RainTomorrow) và ít có mối quan hệ với các biến độc lập khác.

6. XÂY DỰNG MÔ HÌNH

Bộ dữ liệu dùng để xây dựng mô hình sẽ được phân chia và huấn luyện trên 22 địa điểm riêng lẻ sẽ gồm **15** biến: location, relative_humidity_2m, surface_pressure, cloud_cover, weather_code, temperature_2m_min, temperature_2m_max, daylight_duration, sunshine_duration, precipitation_sum, precipitation_hours, wind_speed_10m_max, wind_direction_10m_dominant, et0_fao_evapotranspiration. Các mô hình được sử dụng từ các thư viện có sẵn và thực hiện chuẩn hoá mô hình để tìm ra các thông số thuật toán sao cho phù hợp mang lại kết quả cao nhất.

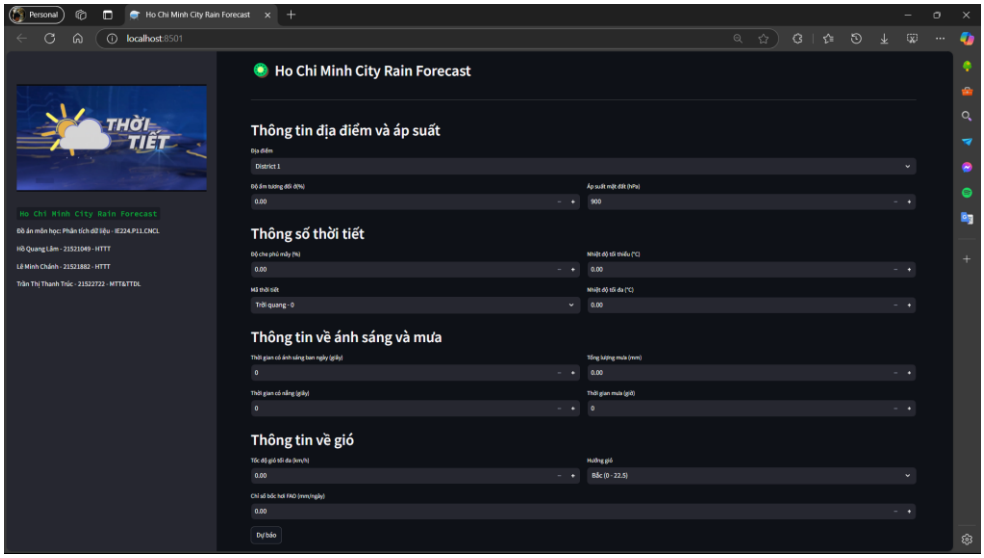
Kiểm tra độ chính xác của 3 thuật toán phổ biến khi cho bài toán phân loại nhị phân: **RandomForestClassifier, XGBClassifier, LGBMClassifier**. Kiểm tra trên các độ đo **"Accuracy", "Precision", "Recall", "F1 Score", "ROC AUC".**

Kết quả thu được tại 22 địa điểm, sau tính trung bình các độ đo và độ lệch chuẩn thì thuật toán **RandomForest** là thuật toán hiệu quả và ổn định nhất với hầu hết các thông số độ đo đều cao nhất.

	Test Accuracy		Test Precision		Test Recall		Test F1-score		Test ROC AUC	
	mean	std	mean	std	mean	std	mean	std	mean	std
Model										
RandomForest	0.86675	0.00515	0.88549	0.00413	0.88822	0.01067	0.88680	0.00430	0.93076	0.00446
XGB	0.86592	0.00751	0.87822	0.00722	0.89626	0.01402	0.88706	0.00659	0.93051	0.00372
LGBM	0.86513	0.00753	0.87673	0.00710	0.89670	0.01223	0.88654	0.00633	0.92835	0.00353

7. ỨNG DỤNG MÔ HÌNH

Nhóm tiến hành xây dựng mô hình với thuật toán RandomForest với dữ liệu huấn luyện là toàn bộ bộ dữ liệu, sau đó ứng dụng vào xây dựng một website bằng thư viện Streamlit của Python để có thể trực quan nhập các thuộc tính của thời tiết làm biến đầu vào và dự báo với một bộ dữ liệu cụ thể và hiển thị kết quả.



Hình 12. Giao diện trang web dự báo mưa áp dụng mô hình

8. KẾT LUẬN

Từ bài toán đã đề ra, sau quá trình xử lí, phân tích, đánh giá và xây dựng mô hình dữ liệu thời tiết của 22 địa phương của TP. Hồ Chí Minh, nhóm đã có cái nhìn tổng quan, hiểu hơn về thời tiết của thành phố, các yếu tố tự nhiên diễn ra hằng ngày.

Bộ dữ liệu có cấu trúc dạng chuỗi thời gian kết hợp với nhiều địa điểm, nhóm đã áp dụng nhiều bước để xử lí dữ liệu, trong quá trình phân tích thăm dò, nhóm đã sử dụng các biểu đồ trực quan khác nhau, chọn lọc các yếu tố tính để xây dựng mô hình. Từ đó nhóm đã nắm bắt được cơ bản kiến thức về thu thập, kĩ thuật tiền xử lí dữ liệu, kĩ thuật phân tích thăm dò, trực quan hoá dữ liệu và xem xét mối quan hệ bên trong. Từ đó ứng dụng xây dựng mô hình một cách hiệu quả để đạt được những đánh giá khách quan và độ chính xác cao.

TÀI LIỆU THAM KHẢO

[1] Historical Weather API. Open-Meteo.com. Link: https://open-meteo.com/en/docs/historical-weather-api (26/10/2024)

[2] Nominatim demo. OpenStreetMap.org. Link: https://nominatim.openstreetmap.org/ui/search.html (26/10/2024)

[3] RandomForestClassifier - Scikit-Learn. Link: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (01/12/2024)

[4] Get Started with XGBoost — xgboost 2.1.3 documentation. Link: https://xgboost.readthedocs.io/en/latest/get_started.html (05/12/2024)

[5] Lightgbm.LGBMClassifier — LightGBM 4.5.0.99 documentation. Link: https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html (05/12/2024)

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Hồ Quang Lâm	Thu thập dữ liệu Phân tích thăm dò Tổng hợp viết báo cáo Chuẩn bị slide thuyết trình
2	Lê Minh Chánh	Thu thập dữ liệu Phân tích thăm dò Xây dựng thuật toán trên model có sẵn Xây dựng website Tổng hợp viết báo cáo
3	Trần Thị Thanh Trúc	Tiền xử lí dữ liệu Xây dựng thuật toán trên mode có sẵn Tổng hợp viết báo cáo Chuẩn bị file thuyết trình