BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HCM

KHOA CÔNG NGHỆ THÔNG TIN

MÔN HỌC: PHÂN TÍCH DỮ LIỆU

BÁO CÁO CUỐI KÌ

ĐỀ TÀI: PHÂN TÍCH TẬP DỮ LIỆU LƯỢNG MƯA Ở ÚC

GVHD: Ths.Nguyễn Văn Thành

SVTH:	**MSSV**
Nguyễn Phương Khoa	21133048
Phạm Hữu Dũng	21133022
Hoàng Mạnh Đức	21133027
Trương Quốc Việt	21133092
Lê Lương Trường An	21133001

Thành Phố Hồ Chí Minh, Tháng 05 năm 2024

NHẬN XÉT CỦA GV:

……

……

……

……

……

……

……

……

……

……

……

ĐIỂM **GV ký tên**

Bảng Phân Công Nhiệm Vụ

Họ và Tên	Nhiệm Vụ	Mức độ	Ưu Điểm	Nhược Điểm
Hoàng Mạnh Đức	+Tham gia làm trực quan hóa dữ liệu về: EDA (Thực hiện Tiền Xử lí dữ liệu) +Thực hiện làm mô hình phân tích Random Forest Classifier	100%	+Hoàn thành tốt được nhiệm vụ làm mô hình trên. +Hỗ trợ nhau trong việc thực hiện các mô hình.	Trong quá trình làm vẫn còn nhiều thiếu sót,mô hình chưa được hoàn thiện tốt.
Nguyễn Phương Khoa	+Tham gia làm trực quan hóa dữ liệu về: EDA (Thực hiện kiểm tra dữ liệu) +Thực hiện làm mô hình phân tích Logistic Regression	100%	+Hoàn thành tốt được nhiệm vụ làm mô hình trên.	Trong quá trình làm vẫn còn nhiều thiếu sót,mô hình chưa được hoàn thiện tốt.
Phạm Hữu Dũng	+Thực hiện làm mô hình phân tích KNN +Tham gia làm trực quan hóa dữ liệu về: EDA (Thực hiện Tiền Xử lí dữ liệu)	100%	+Hoàn thành tốt được nhiệm vụ làm mô hình trên. +Hỗ trợ nhau trong việc thực hiện các mô hình.	Trong quá trình làm vẫn còn nhiều thiếu sót,mô hình chưa được hoàn thiện tốt.

Trương Quốc Việt	+Tham gia làm trực quan hóa dữ liệu về: EDA (Thực hiện Phân Tích Sơ bộ) +Thực hiện làm mô hình phân tích DecisionTree Classifier	100%	+Hoàn thành tốt được nhiệm vụ làm mô hình trên.	Trong quá trình làm vẫn còn nhiều thiếu sót,mô hình chưa được hoàn thiện tốt.
Lê Lương Trường An	+Thực hiện làm mô hình phân tích Support Vector Machine +Tham gia làm trực quan hóa dữ liệu về: EDA (Thực hiện Phân Tích Sơ bộ)	100%	+Hoàn thành tốt được nhiệm vụ làm mô hình trên. +Hỗ trợ nhau trong việc thực hiện các mô hình.	Trong quá trình làm vẫn còn nhiều thiếu sót,mô hình chưa được hoàn thiện tốt.

Mục Lục

CHƯƠNG I: TỔNG QUAN VỀ ĐỀ TÀI

1.1. Lý do chọn đề tài

Lượng mưa đóng vai trò nền tảng cho nhiều hoạt động kinh tế - xã hội, đóng góp quan trọng vào nông nghiệp, thủy lợi, giao thông, du lịch,... Phân tích dữ liệu lượng mưa giúp dự báo chính xác, ứng phó hiệu quả với biến đổi khí hậu và hỗ trợ nghiên cứu khoa học. Nhóm chúng em lựa chọn Úc làm đối tượng nghiên cứu bởi đây là quốc gia có nền kinh tế phụ thuộc vào nông nghiệp, sở hữu nguồn dữ liệu phong phú và chịu ảnh hưởng nặng nề bởi biến đổi khí hậu.

Kết quả nghiên cứu có thể ứng dụng vào nhiều lĩnh vực: dự báo lượng mưa, lập kế hoạch tưới tiêu, phát triển hệ thống cảnh báo sớm và nghiên cứu khoa học. Do vậy, phân tích tập dữ liệu lượng mưa ở Úc là đề tài nghiên cứu quan trọng, mang tính chiến lược, giúp nâng cao kiến thức, kỹ năng phân tích dữ liệu và giải quyết các vấn đề thực tế trong cuộc sống.

1.2. Tổng quan về tập dữ liệu

1.2.1. Nguồn dữ liệu

- Nhóm chung em sử dụng Tập dữ liệu "Rain in Australia" được lấy từ trang web Kaggle (https://www.kaggle.com/), một nền tảng cung cấp các tập dữ liệu tin cậy và miễn phí cho mục đích học tập và nghiên cứu.
- Đường dẫn tập dữ liệu: Rain in Australia

1.2.2. Mô tả chi tiết tập dữ liệu

Tập dữ liệu này là một bộ sưu tập với hơn 10 năm quan sát thời tiết hàng ngày từ nhiều địa điểm khác nhau trên toàn bộ lãnh thổ của nước Úc.

Mục tiêu: Dự đoán lượng mưa ngày hôm sau bằng cách huấn luyện các mô hình phân loại trên biến mục tiêu RainTomorrow. RainTomorrow là biến mục tiêu cần dự đoán.

Nó có nghĩa là -- ngày hôm sau trời mưa, Có hay Không? Cột này là Có nếu lượng mưa trong ngày hôm đó từ 1 mm trở lên.

Dữ liệu bao gồm:

- 23 cột.

- 145461 dòng.

Tập dữ liệu bao gồm các thông tin như sau:

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporatio	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
2	12/1/2008	Albury	13.4	22.9	0.6	NA	NA	W	44	W	WNW	20	24	71	22	1007.7	1007.1	8	NA	16.9	21.8	No	No
3	12/2/2008	Albury	7.4	25.1	0	NA	NA	WNW	44	NNW	WSW	4	22	44	25	1010.6	1007.8	NA	NA	17.2	24.3	No	No
4	12/3/2008	Albury	12.9	25.7	0	NA	NA	WSW	46	W	WSW	19	26	38	30	1007.6	1008.7	NA	2	21	23.2	No	No
5	12/4/2008	Albury	9.2	28	0	NA	NA	NE	24	SE	E	11	9	45	16	1017.6	1012.8	NA	NA	18.1	26.5	No	No
6	12/5/2008	Albury	17.5	32.3	1	NA	NA	W	41	ENE	NW	7	20	82	33	1010.8	1006	7	8	17.8	29.7	No	No
7	12/6/2008	Albury	14.6	29.7	0.2	NA	NA	WNW	56	W	W	19	24	55	23	1009.2	1005.4	NA	NA	20.6	28.9	No	No
8	12/7/2008	Albury	14.3	25	0	NA	NA	W	50	SW	W	20	24	49	19	1009.6	1008.2	1	NA	18.1	24.6	No	No
9	12/8/2008	Albury	7.7	26.7	0	NA	NA	W	35	SSE	W	6	17	48	19	1013.4	1010.1	NA	NA	16.3	25.5	No	No
10	12/9/2008	Albury	9.7	31.9	0	NA	NA	NNW	80	SE	NW	7	28	42	9	1008.9	1003.6	NA	NA	18.3	30.2	No	Yes
11	2/10/2008	Albury	13.1	30.1	1.4	NA	NA	W	28	S	SSE	15	11	58	27	1007	1005.7	NA	NA	20.1	28.2	Yes	No
12	2/11/2008	Albury	13.4	30.4	0	NA	NA	N	30	SSE	ESE	17	6	48	22	1011.8	1008.7	NA	NA	20.4	28.8	No	Yes
13	2/12/2008	Albury	15.9	21.7	2.2	NA	NA	NNE	31	NE	ENE	15	13	89	91	1010.5	1004.2	8	8	15.9	17	Yes	Yes
14	2/13/2008	Albury	15.9	18.6	15.6	NA	NA	W	61	NNW	NNW	28	28	76	93	994.3	993	8	8	17.4	15.8	Yes	Yes
15	2/14/2008	Albury	12.6	21	3.6	NA	NA	SW	44	W	SSW	24	20	65	43	1001.2	1001.8	NA	7	15.8	19.8	Yes	No

Ý nghĩa các cột thông tin:

Cột	Ý Nghĩa	Cột	Ý Nghĩa
1. Date	Ngày ghi nhận thông tin thời tiết	13. WindSpeed3pm	Tốc độ gió lúc 3 giờ chiều (đơn vị: km/h)
2. Location	Vị trí địa lý nơi quan sát thời tiết	14. Humidity9am	Độ ẩm lúc 9 giờ sáng (đơn vị: phần trăm)
3. MinTemp	Nhiệt độ tối thiểu trong ngày (đơn vị: độ Celsius)	15. Humidity3pm	Độ ẩm lúc 3 giờ chiều (đơn vị: phần trăm)
4. MaxTemp	Nhiệt độ tối đa trong ngày (đơn vị: độ Celsius)	16. Pressure9am	Áp suất không khí lúc 9 giờ sáng (đơn vị: hPa)
5. Rainfall	Lượng mưa tích lũy trong ngày (đơn vị: mm)	17. Pressure3pm	Áp suất không khí lúc 3 giờ chiều (đơn vị: hPa)
6. Evaporation	Lượng hơi nước bốc hơi trong ngày (đơn vị: mm)	18. Cloud9am	Mây che phủ lúc 9 giờ sáng (đơn vị: oktas)
7. Sunshine	Thời gian nắng trong ngày (đơn vị: giờ)	19. Cloud3pm	Mây che phủ lúc 3 giờ chiều (đơn vị: oktas
8. WindGustDir	Hướng gió mạnh nhất trong ngày	20. Temp9am	Nhiệt độ lúc 9 giờ sáng (đơn vị: độ Celsius)
9. WindGustSpeed	Tốc độ gió mạnh nhất trong ngày (đơn vị: km/h)	21. Temp3pm	Nhiệt độ lúc 3 giờ chiều (đơn vị: độ Celsius)
10. WindDir9am	Hướng gió lúc 9 giờ sáng	22. RainToday	Có mưa trong ngày (Yes: Có, No: Không)
11. WindDir3pm	Hướng gió lúc 3 giờ chiều	23. RainTomorrow	Có mưa vào ngày tiếp theo (Yes: Có, No: Không)
12. WindSpeed9am	Tốc độ gió lúc 9 giờ sáng (đơn vị: km/h)		

CHƯƠNG II: KIỂM TRA VÀ ĐÁNH GIÁ SƠ BỘ VỀ DỮ LIỆU (EDA).

2.1 Kiểm tra dữ liệu

- Kiểm tra kiểu giá trị của dữ liệu:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 145460 entries, 0 to 145459
Data columns (total 23 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   Date           145460 non-null  object
 1   Location       145460 non-null  object
 2   MinTemp        143975 non-null  float64
 3   MaxTemp        144199 non-null  float64
 4   Rainfall       142199 non-null  float64
 5   Evaporation    82670 non-null   float64
 6   Sunshine       75625 non-null   float64
 7   WindGustDir    135134 non-null  object
 8   WindGustSpeed  135197 non-null  float64
 9   WindDir9am     134894 non-null  object
 10  WindDir3pm     141232 non-null  object
 11  WindSpeed9am   143693 non-null  float64
 12  WindSpeed3pm   142398 non-null  float64
 13  Humidity9am    142806 non-null  float64
 14  Humidity3pm    140953 non-null  float64
 15  Pressure9am    130395 non-null  float64
 16  Pressure3pm    130432 non-null  float64
 17  Cloud9am       89572 non-null   float64
 18  Cloud3pm       86102 non-null   float64
 19  Temp9am        143693 non-null  float64
 20  Temp3pm        141851 non-null  float64
 21  RainToday      142199 non-null  object
 22  RainTomorrow   142193 non-null  object
dtypes: float64(16), object(7)
memory usage: 25.5+ MB
```

- Kiểm tra các ký tự đặc biệt cho cột kiểu Object (nếu có):

```
# Kiểm tra các ký tự đặc biệt trong cột kiểu object
object_columns = df.select_dtypes(include=['object']).columns
for col in object_columns:
    special_characters = df[col].str.contains(r'[^\w\s]', na=False)
    if special_characters.any():
        print(f"Column '{col}' contains special characters:")
        print(df[col][special_characters])
```

```
Column 'Date' contains special characters:
0         2008-12-01
1         2008-12-02
2         2008-12-03
3         2008-12-04
4         2008-12-05
             ...
145455    2017-06-21
145456    2017-06-22
145457    2017-06-23
145458    2017-06-24
145459    2017-06-25
Name: Date, Length: 145460, dtype: object
```

- Kiểm tra các dữ liệu Null hoặc bị trùng lặp:

```
df.isnull().sum()
```

```
Date                0
Location            0
MinTemp          1485
MaxTemp          1261
Rainfall         3261
Evaporation     62790
Sunshine        69835
WindGustDir     10326
WindGustSpeed   10263
WindDir9am      10566
WindDir3pm       4228
WindSpeed9am     1767
WindSpeed3pm     3062
Humidity9am      2654
Humidity3pm      4507
Pressure9am     15065
Pressure3pm     15028
Cloud9am        55888
Cloud3pm        59358
Temp9am          1767
Temp3pm          3609
RainToday        3261
RainTomorrow     3267
dtype: int64
```

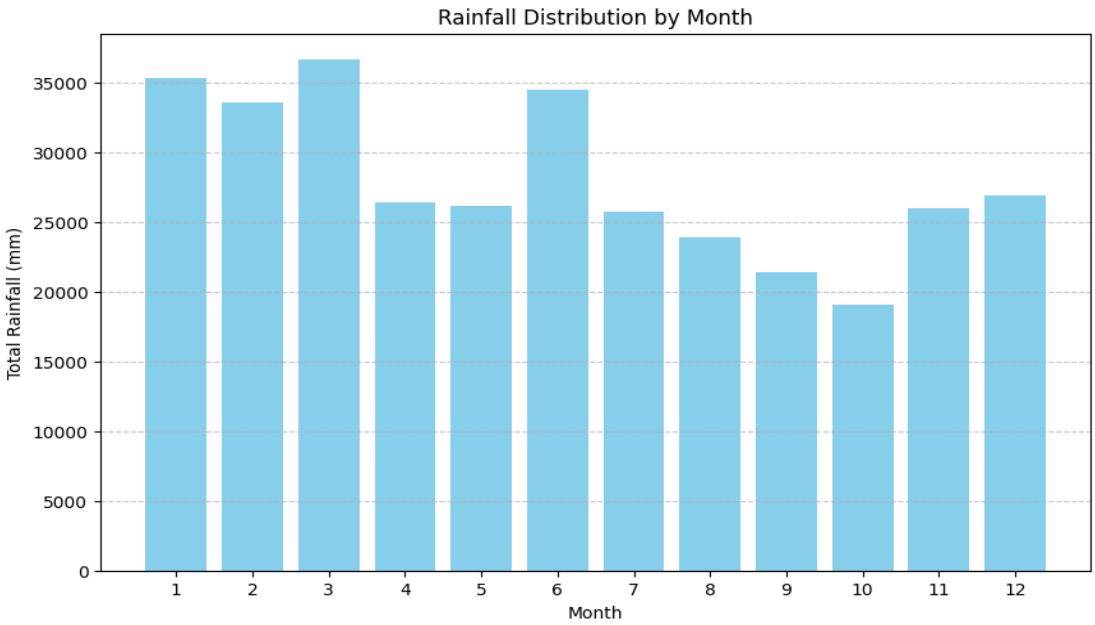
Kiểm tra describe các biến có giá trị là số:

```
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
MinTemp	143975.0	12.194034	6.398495	-8.5	7.6	12.0	16.9	33.9
MaxTemp	144199.0	23.221348	7.119049	-4.8	17.9	22.6	28.2	48.1
Rainfall	142199.0	2.360918	8.478060	0.0	0.0	0.0	0.8	371.0
Evaporation	82670.0	5.468232	4.193704	0.0	2.6	4.8	7.4	145.0
Sunshine	75625.0	7.611178	3.785483	0.0	4.8	8.4	10.6	14.5
WindGustSpeed	135197.0	40.035230	13.607062	6.0	31.0	39.0	48.0	135.0
WindSpeed9am	143693.0	14.043426	8.915375	0.0	7.0	13.0	19.0	130.0
WindSpeed3pm	142398.0	18.662657	8.809800	0.0	13.0	19.0	24.0	87.0
Humidity9am	142806.0	68.880831	19.029164	0.0	57.0	70.0	83.0	100.0
Humidity3pm	140953.0	51.539116	20.795902	0.0	37.0	52.0	66.0	100.0
Pressure9am	130395.0	1017.649940	7.106530	980.5	1012.9	1017.6	1022.4	1041.0
Pressure3pm	130432.0	1015.255889	7.037414	977.1	1010.4	1015.2	1020.0	1039.6
Cloud9am	89572.0	4.447461	2.887159	0.0	1.0	5.0	7.0	9.0
Cloud3pm	86102.0	4.509930	2.720357	0.0	2.0	5.0	7.0	9.0
Temp9am	143693.0	16.990631	6.488753	-7.2	12.3	16.7	21.6	40.2
Temp3pm	141851.0	21.683390	6.936650	-5.4	16.6	21.1	26.4	46.7

2.2 Phân tích sơ bộ

+ Biểu đồ thể hiện lượng mưa theo các tháng ở Úc từ năm 2008 đến 2017



Giải thích:

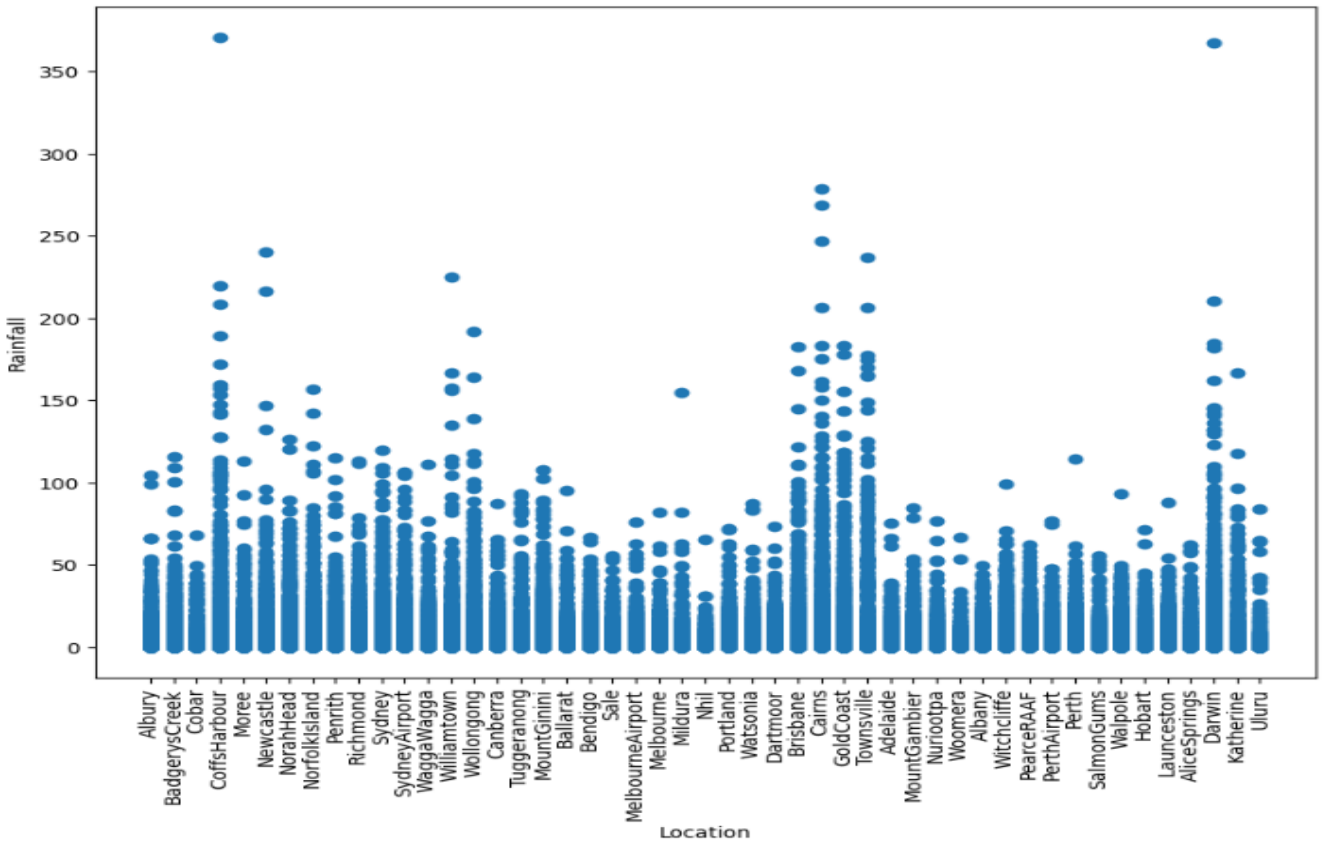
Mùa mưa: Lượng mưa trung bình theo tháng cao nhất vào tháng 3, với lượng mưa trung bình khoảng 36000 mm.

Mùa khô: Lượng mưa trung bình theo tháng thấp nhất vào tháng 10, với lượng mưa trung bình khoảng 19000 mm.

Biên độ mưa: Biên độ mưa giữa các tháng trong năm khá lớn, với lượng mưa trung bình cao nhất gấp khoảng 1,9 lần lượng mưa trung bình thấp nhất.

Phân bố mưa: Lượng mưa phân bố không đều trong năm, tập trung chủ yếu vào tháng 3 .Biểu đồ cho thấy Úc có khí hậu ôn hòa với. Biên độ mưa giữa các tháng trong năm khá lớn, nhưng lượng mưa trung bình theo tháng vẫn ở mức cao, cho thấy Úc được mưa quanh năm.

+ Biểu đồ thể hiện lượng mưa ở các địa điểm trong năm:



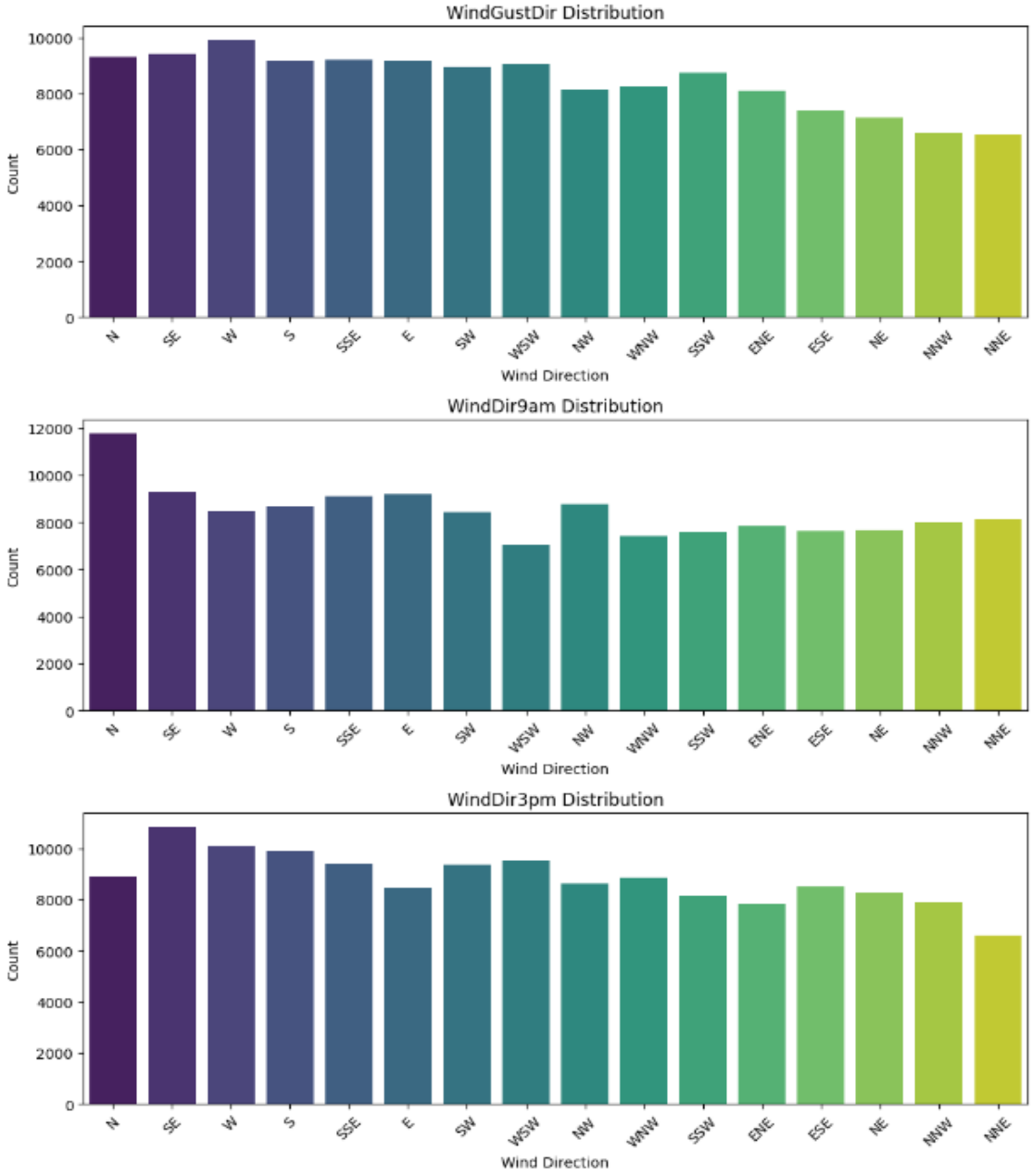
Giải thích:

Đông Úc: Nổi bật với lượng mưa trung bình năm cao nhất, tập trung ở các địa điểm như Cairns, Coffs Harbour, với lượng mưa có thể vượt 350 mm.

Tây Úc và Nam Úc: Lượng mưa trung bình năm thấp hơn đáng kể so với Đông Úc, chỉ dao động từ 50 mm đến 150 mm

Biểu đồ cho thấy sự khác biệt rõ rệt về lượng mưa giữa các khu vực ở Úc và Yếu tố địa ly đóng vai trò quan trọng trong việc phân bố lượng mưa, bao gồm địa hình, hướng gió và dòng hải lưu

+ Biểu đồ thể hiện phân bố hướng gió của tập dữ liệu về lượng mưa ở Úc:



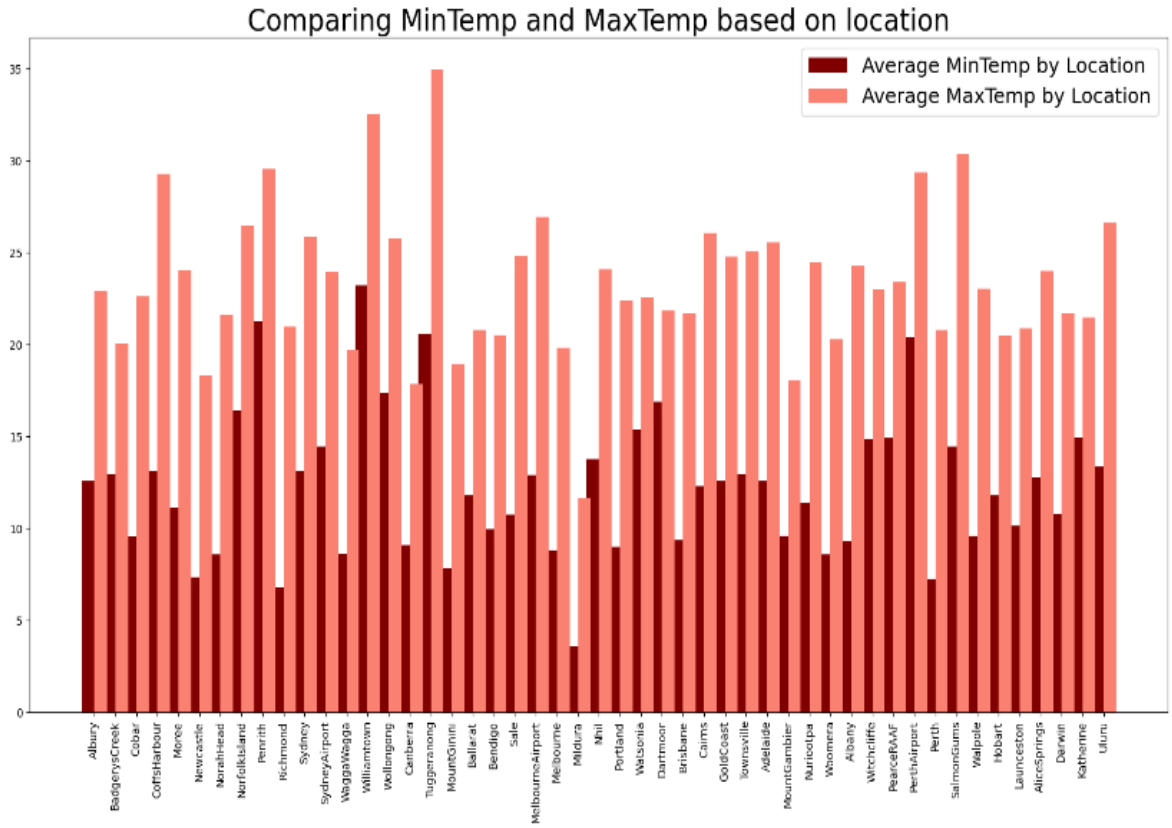
Giải thích:

Hướng gió Tây (W) chiếm tỷ lệ lớn nhất trong ngày với số lần xuất hiện gần 10.000 lần từ năm 2008 đến năm 2017. Đây là hướng có gió mạnh nhất vì có tần suất xuất hiện cao nhất.

Khi xem biểu đồ sự phân phối hướng gió vào lúc 9 giờ sáng, ta thấy hướng Bắc (N) có số lần xuất hiện nhiều nhất, đạt gần 12.000 lần từ năm 2008 đến năm 2017. Vì vậy, gió lúc 9h sáng xuất hiện nhiều nhất ở hướng Bắc.

Với biểu đồ phân bố hướng gió lúc 3 giờ chiều, hướng Đông Nam (SE) chiếm tỷ lệ xuất hiện cao nhất, với số lần xuất hiện dao động trên 10.000 lần trong khoảng thời gian từ năm 2008 đến năm 2017. Kết luận, gió lúc 3h chiều xuất hiện nhiều nhất ở hướng Đông Nam.

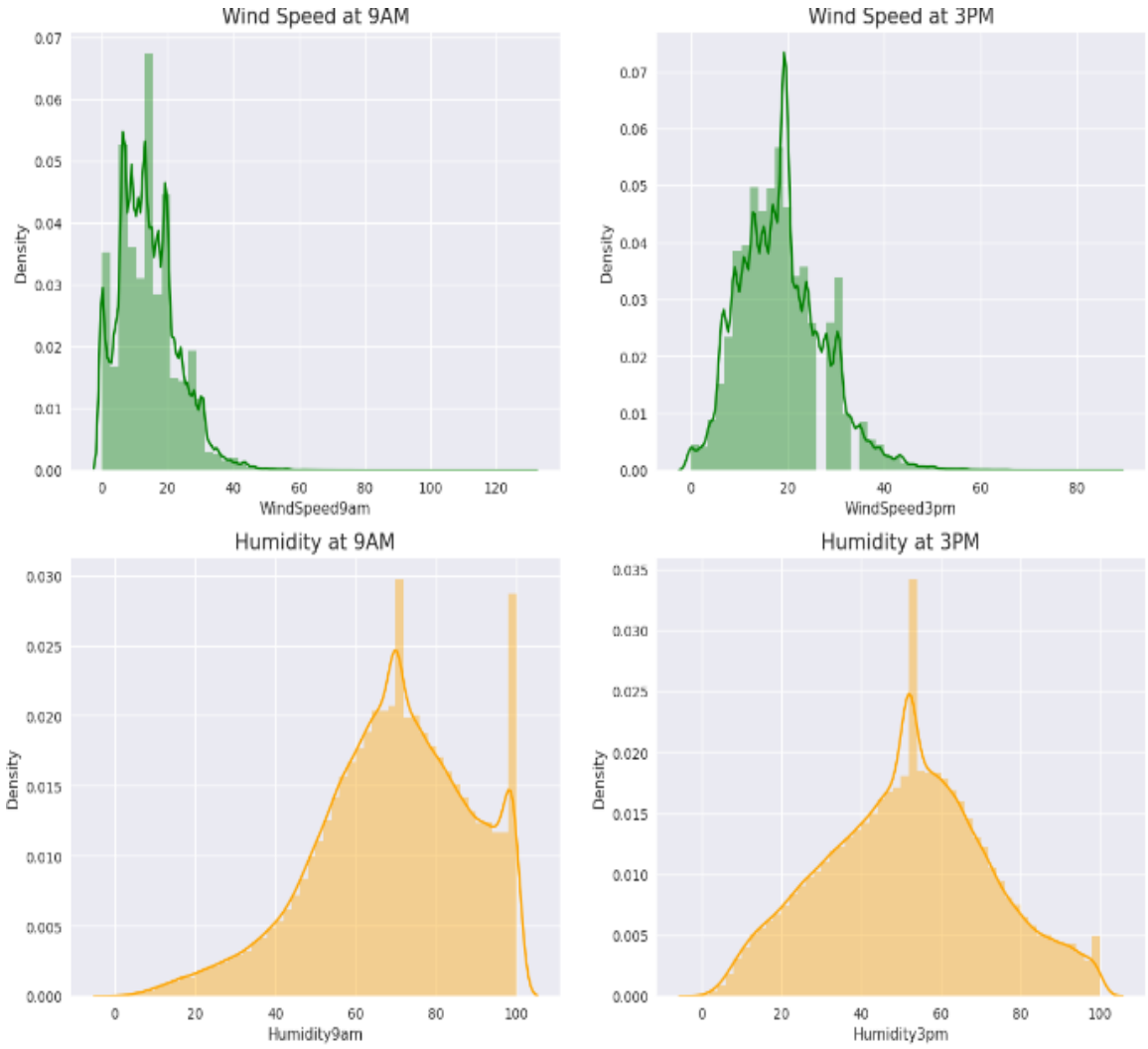
+ Biểu đồ so sánh nhiệt độ thấp nhất và nhiệt độ cao nhất theo khu vực:

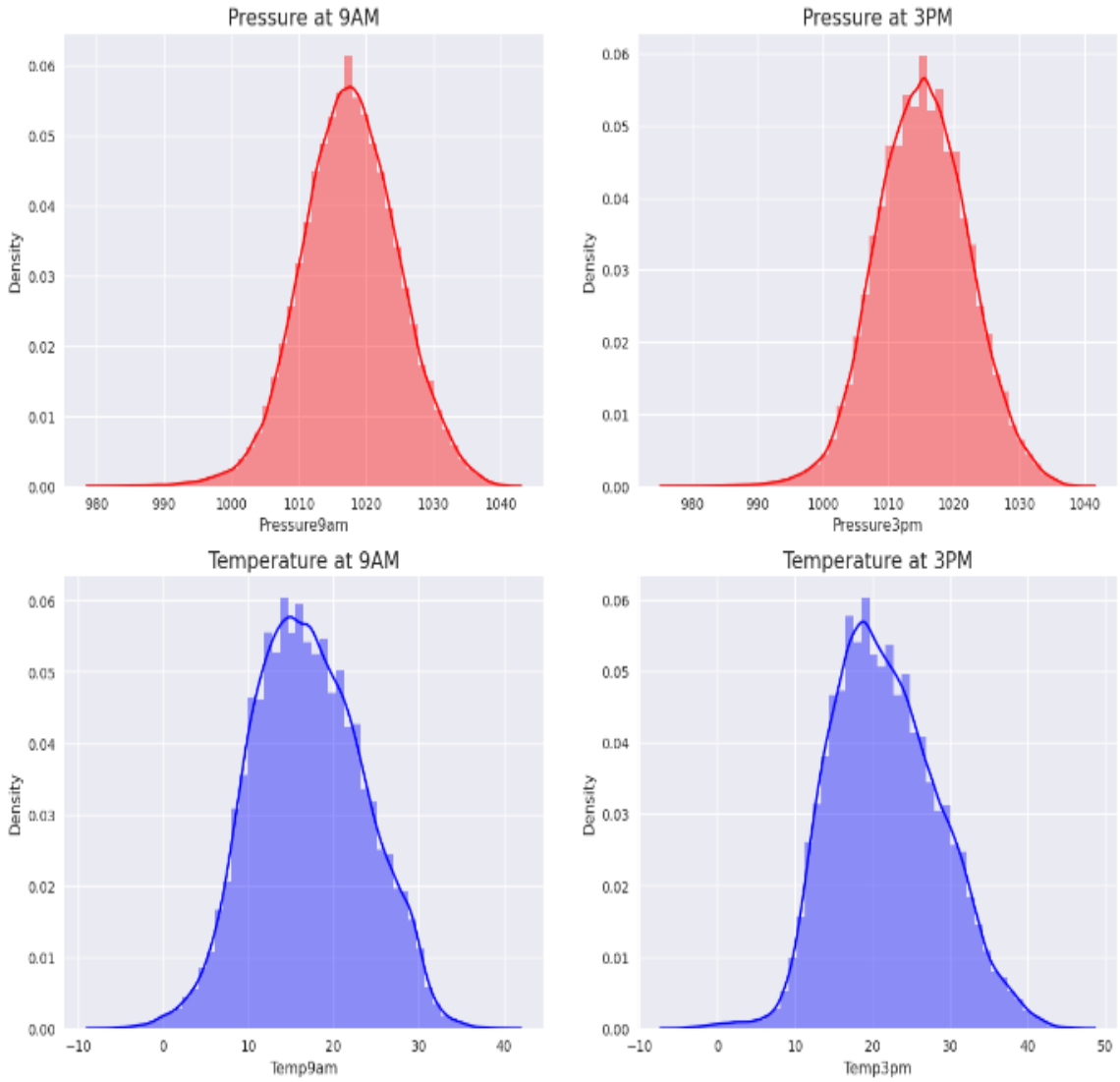


Giải thích:

- Tuggeranong ghi nhận là khu vực có nhiệt độ cao nhất (trong khoảng 34 tới 35 độ)
- Mildura được xác định là khu vực có nhiệt độ thấp nhất (trong khoảng từ 3 tới 4 độ)

+ Biểu đồ thể hiện sự phân bố của các tham số thời tiết (tốc độ gió, độ ẩm, áp suất, nhiệt độ) tại hai thời điểm trong ngày: 9 giờ sáng và 3 giờ chiều. Biểu đồ được chia thành 8 ô con, mỗi ô con dành cho một tham số thời tiết và được hiển thị dưới dạng biểu đồ phân bố (histogram hoặc kernel density plot).





Giải thích:

- Tốc độ gió:

9 giờ sáng: Trung bình dao động từ 0 đến 20 km/h, phân bố đều đặn.

3 giờ chiều: Trung bình dao động từ 0 đến 25 km/h, nghiêng về tốc độ cao hơn, cho thấy gió thường mạnh hơn vào buổi chiều.

- Độ ẩm:

9 giờ sáng: Trung bình dao động từ 50% đến 100%, nghiêng về độ ẩm cao, cho thấy buổi sáng thường ẩm hơn.

3 giờ chiều: Trung bình dao động từ 30% đến 80%, phân bố đều đặn.

- Áp suất:

9 giờ sáng: Trung bình dao động từ 1000 hPa đến 1020 hPa, phân bố đều đặn.

3 giờ chiều: Trung bình dao động từ 990 hPa đến 1010 hPa, phân bố đều đặn.

- Nhiệt độ:

9 giờ sáng: Trung bình dao động từ 15°C đến 25°C, phân bố đều đặn.

3 giờ chiều: Trung bình dao động từ 20°C đến 30°C, nghiêng về nhiệt độ cao hơn, cho thấy nhiệt độ thường cao hơn vào buổi chiều.

2.3. Tiền xử lý dữ liệu

Xóa các cột không sử dụng cho mô hình dự đoán:

```
# Như chúng ta có thể thấy bốn cột đầu tiên có ít hơn 60% dữ liệu, chúng ta có thể bỏ qua bốn cột này
#('Sunshine','Evaporation','Cloud3pm','Cloud9am')
# Chúng ta không cần cột Location và Date vì chúng ta sẽ tìm hiểu xem trời có mưa ở Úc không (không nêu địa điểm cụ thể)
df = df.drop(columns=['Sunshine','Evaporation','Cloud3pm','Cloud9am','Date','Location'],axis=1)
df.shape
```

```
(145460, 20)
```

Xóa giá trị null và các ngoại lệ:

```
#Loại bỏ tất cả các giá trị null trong df
df = df.dropna(how='any')
df.shape
```

```
(112925, 20)
```

```
#Loại bỏ các ngoại lệ trong dữ liệu - sử dụng điểm Z để phát hiện và loại bỏ các ngoại lệ.
from scipy import stats
z = np.abs(stats.zscore(df._get_numeric_data()))
print(z)
df= df[(z < 3).all(axis=1)]
print(df.shape)
```

```
          MinTemp   MaxTemp  Rainfall  WindGustSpeed  WindSpeed9am  \
0        0.117567  0.108221  0.206661       0.241214      0.577742
1        0.841802  0.206845  0.276405       0.241214      1.339742
2        0.037620  0.292772  0.276405       0.391345      0.457900
3        0.553991  0.622159  0.276405       1.260094      0.500842
4        0.773137  1.237969  0.160165       0.016018      0.980214
...           ...       ...       ...            ...           ...
145454   1.465392  0.265754  0.276405       0.734636      0.021471
145455   1.577319  0.036615  0.276405       0.734636      0.261157
145456   1.449403  0.235487  0.276405       1.410225      0.261157
145457   1.161592  0.464626  0.276405       0.284243      0.740528
145458   0.777844  0.478947  0.276405       0.959832      0.261157

        WindSpeed3pm  Humidity9am  Humidity3pm  Pressure9am  Pressure3pm  \
0           0.524408     0.190140     1.380413     1.382962     1.142455
1           0.291310     1.237561     1.235963     0.970598     1.041848
2           0.757507     1.554828     0.995214     1.397181     0.912497
3           1.223831     1.184683     1.669313     0.024764     0.323229
4           0.058211     0.771796     0.850764     0.942159     1.300551
...              ...          ...          ...          ...          ...
145454      0.757634     0.444394     1.139664     1.034344     0.884050
145455      0.990733     0.867416     1.284113     1.020125     0.754699
145456      1.223831     0.603027     1.428563     0.863711     0.582231
145457      1.223831     0.761660     1.284113     0.508225     0.251666
145458      1.456930     0.867416     1.284113     0.280714     0.208549

         Temp9am   Temp3pm   RISK_MM      Year     Month
0       0.088435  0.047870  0.273102  1.878666  1.601848
1       0.041228  0.317768  0.273102  1.878666  1.601848
2       0.556724  0.156887  0.273102  1.878666  1.601848
3       0.100392  0.639531  0.155691  1.878666  1.601848
4       0.053185  1.107548  0.249620  1.878666  1.601848
...          ...       ...       ...       ...       ...
145454  1.268605  0.179500  0.273102  1.678377  0.124126
145455  1.158456  0.039883  0.273102  1.678377  0.124126
145456  1.032571  0.347020  0.273102  1.678377  0.124126
145457  0.780802  0.581028  0.273102  1.678377  0.124126
145458  0.371676  0.566403  0.273102  1.678377  0.124126

[112925 rows x 15 columns]
(106447, 20)
```

Thay đổi yes/no thành 1/0 cho cột RainToday, RainTomorrow và chuyển đổi các cột phân loại:

```
#Thay đổi yes/no thành 1/0 cho RainToday và RainTomorrow

df['RainToday'].replace({'No': 0, 'Yes': 1},inplace = True)
df['RainTomorrow'].replace({'No': 0, 'Yes': 1},inplace = True)
#Xem các giá trị duy nhất và chuyển đổi chúng thành int bằng pd.get_Dummies(

categorical_columns = ['WindGustDir', 'WindDir3pm', 'WindDir9am']
for col in categorical_columns:
    print(np.unique(df[col]))
# chuyển đổi các cột phân loại
df = pd.get_dummies(df, columns=categorical_columns)
df.iloc[4:9]
```

```
['E' 'ENE' 'ESE' 'N' 'NE' 'NNE' 'NNW' 'NW' 'S' 'SE' 'SSE' 'SSW' 'SW' 'W'
 'WNW' 'WSW']
['E' 'ENE' 'ESE' 'N' 'NE' 'NNE' 'NNW' 'NW' 'S' 'SE' 'SSE' 'SSW' 'SW' 'W'
 'WNW' 'WSW']
['E' 'ENE' 'ESE' 'N' 'NE' 'NNE' 'NNW' 'NW' 'S' 'SE' 'SSE' 'SSW' 'SW' 'W'
 'WNW' 'WSW']
```

Chuẩn hóa dữ liệu - sử dụng MinMaxScaler

```
#bước tiếp theo là chuẩn hóa dữ liệu - sử dụng MinMaxScaler

from sklearn import preprocessing
scaler = preprocessing.MinMaxScaler()
scaler.fit(df)
df = pd.DataFrame(scaler.transform(df), index=df.index, columns=df.columns)
df.iloc[4:10]
```

	MinTemp	MaxTemp	Rainfall	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	...	WindDir9am_NN
4	0.628342	0.696296	0.035714	0.465753	0.135135	0.428571	0.797753	0.33	0.342043	0.282974	...	0
5	0.550802	0.632099	0.007143	0.671233	0.459459	0.523810	0.494382	0.23	0.304038	0.268585	...	0
6	0.542781	0.516049	0.000000	0.589041	0.486486	0.523810	0.426966	0.19	0.313539	0.335731	...	0
7	0.366310	0.558025	0.000000	0.383562	0.108108	0.357143	0.415730	0.19	0.403800	0.381295	...	0
8	0.419786	0.686420	0.000000	1.000000	0.135135	0.619048	0.348315	0.09	0.296912	0.225420	...	0
9	0.510695	0.641975	0.050000	0.287671	0.351351	0.214286	0.528090	0.27	0.251781	0.275779	...	0

6 rows × 65 columns

Sử dụng SelectKBest để thấy được những cột quan trọng:

```
#Bây giờ chúng ta đã hoàn tất phần tiền xử lý, hãy cùng xem đâu là những cột quan trọng của RainTomorrow!
#Sử dụng SelectKBest để có được những cột quan trọng!

from sklearn.feature_selection import SelectKBest, chi2
X = df.loc[:,df.columns!='RainTomorrow']
y = df[['RainTomorrow']]
selector = SelectKBest(chi2, k=3)
selector.fit(X, y)
X_new = selector.transform(X)
print(X.columns[selector.get_support(indices=True)]) #top 3 columns
```

```
Index(['Rainfall', 'RainToday', 'RISK_MM'], dtype='object')
```

Nắm bắt các tính năng quan trọng khi gán chúng là X:

```
#Nắm bắt các tính năng quan trọng khi gán chúng là X
df = df[['Humidity3pm','Rainfall','RainToday','RainTomorrow']]
X = df[['Humidity3pm']]
y = df[['RainTomorrow']]
```

CHƯƠNG III: MÔ HÌNH PHÂN TÍCH VÀ KẾT QUẢ

3.1. Logistic Regression:

Mô Hình	Accuracy	Precision		Recall		F1-Score	
		Yes	No	Yes	No	Yes	No
LogisticRegression	84.09%	69.88%	85.46%	31.57%	96.73%	43.49%	43.49%

Độ chính xác (Accuracy):

Độ chính xác của mô hình là 0.8409364196603036 (84.09%), cho thấy mô hình có thể dự đoán đúng khoảng 84.09% các trường hợp trong tập dữ liệu.

Precision và Recall:

Precision cho lớp "Yes" là 69.88%, có nghĩa là trong tất cả các dự đoán là "Yes", có 69.88% là đúng. Precision cho lớp "No" là 85.46%, chỉ ra mô hình có độ chính xác tốt hơn trong dự đoán lớp "No".

Recall cho lớp "Yes" là 31.57%, tức là chỉ khoảng 31.57% các trường hợp thực sự là "Yes" được mô hình dự đoán đúng. Trong khi đó, lớp "No" có Recall là 96.73%, cho thấy mô hình rất tốt trong việc phát hiện các trường hợp thuộc lớp "No".

F1-Score:

F1-Score cho cả hai lớp là 43.49%, cho thấy sự cân bằng giữa Precision và Recall. Mức F1-Score này phản ánh khả năng dự đoán của mô hình với sự cân nhắc giữa độ chính xác và độ nhạy.

Thời gian huấn luyện (Training Time):

Mô hình có thời gian huấn luyện là 0.07479119300842285 giây, tương đối nhanh, chỉ ra rằng LogisticRegression là một mô hình hiệu quả về mặt thời gian.

Kết Luận: Mô hình Logistic Regression cho thấy độ chính xác khá cao, đạt khoảng 84.09%. Tuy nhiên, Precision cho nhóm "Yes" chỉ đạt khoảng 69.88%, còn Recall chỉ khoảng 31.57%. Điều này ngụ ý rằng mô hình có xu hướng dự đoán sai các trường hợp có mưa. Mặc dù Precision cho nhóm "No" đạt 85.46%, nhưng mô hình cũng chỉ có Recall khoảng 31.57%. F1-score đối với cả hai nhóm "Yes" và "No" đều là 43.49%. Thời gian huấn luyện của mô hình khá nhanh, chỉ mất khoảng 0.07 giây. Tóm lại, mặc dù có độ chính xác cao, mô hình cần được cải thiện để nhận diện chính xác các trường hợp có mưa hơn.

3.2. Random Forest Classifier:

Mô Hình	Accuracy	Precision		Recall		F1-Score	
		Yes	No	Yes	No	Yes	No
Random Forest Classifier	83.80%	74.52%	84.55%	27.92%	97.64%	40.62%	40.62%

Độ chính xác (Accuracy):

Độ chính xác của mô hình là 0.8380429881256576 (83.80%), cho thấy rằng mô hình có khả năng dự đoán đúng khoảng 83.80% các trường hợp trong tập dữ liệu.

Precision và Recall:

Precision cho lớp "Yes" là 74.52%, điều này có nghĩa rằng trong số các trường hợp được dự đoán là "Yes", khoảng 74.52% là chính xác. Precision của lớp "No" là 84.55%, cũng khá cao.

Recall cho lớp "Yes" là 27.92%, nghĩa là chỉ khoảng 27.92% các trường hợp thực sự là "Yes" được mô hình nhận diện. Trong khi đó, lớp "No" có recall cao (97.64%), cho thấy mô hình có khả năng nhận diện hầu hết các trường hợp "No".

F1-Score:

F1-Score của cả hai lớp là 40.62%, thể hiện sự cân bằng giữa Precision và Recall. Mức F1-Score này phản ánh khả năng dự đoán tổng thể của mô hình, tuy nhiên cho thấy cần cải thiện khả năng phân loại của lớp "Yes".

Thời gian huấn luyện (Training Time):

Mô hình này có thời gian huấn luyện là 2.631321668624878 giây. So với các mô hình khác như DecisionTreeClassifier, RandomForestClassifier mất nhiều thời gian huấn luyện hơn, nhưng thường có hiệu suất cao hơn nhờ việc kết hợp nhiều cây quyết định

Kết Luận:

Mô hình Random Forest Classifier có độ chính xác cao, khoảng 83.80%, nhưng nhớ lại các trường hợp có mưa không tốt (Recall thấp), chỉ khoảng 29.55%. Độ chính xác trong việc dự đoán các trường hợp không mưa (Precision) là 84.84%. Thời gian huấn luyện của mô hình khá lớn, khoảng 2.55 giây. Tuy nhiên, F1-score chỉ đạt khoảng 41.90% cho cả hai nhóm "Yes" và "No", chỉ ra một sự cân nhắc giữa Precision và Recall. Tóm lại, mặc dù đạt được độ chính xác cao, mô hình cần được cải thiện để nhận diện chính xác hơn các trường hợp có mưa.

3.3. Decision Tree Classifier:

Mô Hình	Accuracy	Precision		Recall		F1-Score	
		Yes	No	Yes	No	Yes	No
DecisionTreeClassifier	83.50%	72.90%	84.53%	29.68%	97.21%	42.18%	42.18%

Độ chính xác (Accuracy):

Độ chính xác của mô hình là 0.8357132120847738 (83.57%), cho thấy rằng mô hình có thể dự đoán đúng khoảng 83.57% các trường hợp trong tập dữ liệu.

Precision và Recall:

Precision cho lớp "Yes" là 72.90%, nghĩa là trong tất cả các dự đoán là "Yes", có 72.90% là đúng. Precision cho lớp "No" là 84.53%, cũng là con số tương đối cao.

Recall cho lớp "Yes" là 29.68%, tức là chỉ khoảng 29.68% các trường hợp thực sự là "Yes" được mô hình dự đoán đúng. Trong khi đó, lớp "No" có recall là 97.21%, thể hiện rằng mô hình này có khả năng nhận diện phần lớn các trường hợp thuộc lớp "No".

F1-Score:

F1-Score cho cả hai lớp là 42.18%, cho thấy sự cân bằng giữa Precision và Recall. Mức F1-Score này phản ánh hiệu suất tổng thể của mô hình trong việc dự đoán cả hai lớp, với trọng tâm rõ ràng hơn về lớp "No".

Thời gian huấn luyện (Training Time):

Mô hình được huấn luyện với thời gian 0.06875991821289062 giây, là thời gian huấn luyện rất nhanh, điều này thể hiện lợi ích về mặt hiệu suất của DecisionTreeClassifier.

Kết Luận:

Mô hình Decision Tree Classifier đạt độ chính xác cao khoảng 83.80%, nhưng chỉ nhận diện đúng khoảng 29.55% các trường hợp có mưa. Precision cho các trường hợp không mưa đạt 84.84%. Thời gian huấn luyện khá lớn, khoảng 2.55 giây. F1-score chỉ đạt khoảng 41.90%, chỉ ra sự cân nhắc giữa Precision và Recall. Mặc dù có độ chính xác cao, mô hình cần cải thiện để nhận diện chính xác hơn các trường hợp có mưa.

3.4. Support Vector Machine:

Mô Hình	Accuracy	Precision		Recall		F1-Score	
		Yes	No	Yes	No	Yes	No
Support Vector Machine	80.50%	0.00%	80.50%	0.00%	100.00%	0.00%	0.00%

Accuracy:

Độ chính xác của mô hình là 0.8049751991582744 (80.50%). Điều này cho thấy mô hình có khả năng dự đoán đúng khoảng 80.50% các trường hợp trong tập dữ liệu.

Precision và Recall:

Precision của các lớp "Yes/No" cho thấy sự khác biệt đáng kể. Lớp "Yes" có độ chính xác là 0.00%, có nghĩa là mọi dự đoán là "Yes" đều không chính xác. Trong khi đó, lớp "No" có độ chính xác là 80.50%, tức là trong các dự đoán là "No", khoảng 80.50% là đúng.

Recall của lớp "Yes" là 0.00%, cho thấy mô hình không bắt được bất kỳ trường hợp nào thuộc lớp này. Ngược lại, lớp "No" có recall là 100.00%, nghĩa là mọi trường hợp thực sự là "No" đều được mô hình dự đoán đúng.

F1-Score:

Vì F1-Score là một trung bình hài hòa giữa Precision và Recall, giá trị F1-Score là 0.00% cho lớp "Yes" là hợp lý, vì cả Precision và Recall đều thấp hoặc bằng 0. Điều này chỉ ra rằng mô hình không có khả năng dự đoán chính xác các trường hợp "Yes".

Training Time:

Thời gian huấn luyện của mô hình là 0.08376646041870117 giây. Đây là thời gian khá ngắn, cho thấy mô hình huấn luyện nhanh.

Kết luận:

Mô hình Support Vector Classifier (SVC) có độ chính xác khoảng 80.50%, nhưng không nhận diện được bất kỳ trường hợp nào thuộc lớp "Yes". Precision của lớp "No" đạt 80.50%, và Recall đạt 100.00%. Tuy nhiên, F1-score đều đạt 0.00% cho cả hai lớp, cho thấy mô hình không đạt được sự cân nhắc giữa Precision và Recall.

3.5. K-Nearest Neighbors:

Mô Hình	Accuracy	Precision		Recall		F1-Score	
		Yes	No	Yes	No	Yes	No
K-Nearest Neighbors	81.26%	55.97%	83.70%	24.84%	95.18%	34.41%	34.41%

Độ chính xác (Accuracy):

Độ chính xác của mô hình là 0.8126033368405231 (81.26%), cho thấy rằng mô hình có khả năng dự đoán đúng khoảng 81.26% trường hợp trong tập dữ liệu.

Độ chính xác và Độ nhớ (Precision and Recall):

Precision cho lớp "Yes" là 55.97%, điều này có nghĩa rằng trong tất cả các trường hợp mà mô hình dự đoán là "Yes", có 55.97% khả năng là đúng. Precision cho lớp "No" là 83.70%, đây cũng là con số khá cao.

Recall cho lớp "Yes" là 24.84%, nghĩa là chỉ khoảng 24.84% các trường hợp thực sự là "Yes" được mô hình dự đoán đúng. Trong khi đó, lớp "No" có recall cao (95.18%), điều này cho thấy mô hình có khả năng nhận diện hầu hết các trường hợp "No".

F1-Score:

F1-Score của cả hai lớp là 34.41%, cho thấy sự cân bằng giữa Precision và Recall. Vì F1-Score là trung bình hài hòa giữa Precision và Recall, con số này phản ánh hiệu suất của mô hình trong việc phân loại đúng cả hai lớp.

Thời gian huấn luyện (Training Time):

Mô hình có thời gian huấn luyện là 0.030718088150024414 giây, cho thấy quá trình huấn luyện diễn ra nhanh chóng.

Kết luận:

Mô hình K-Nearest Neighbors (KNN) có độ chính xác khoảng 81.26%, với Precision của lớp "Yes" đạt 55.97% và của lớp "No" đạt 83.70%. Tuy nhiên, Recall của lớp "Yes" chỉ đạt 24.84%, trong khi của lớp "No" đạt 95.18%. F1-score của cả hai lớp đều khoảng 34.41%. Mặc dù có độ chính xác cao và Precision tốt, nhưng mô hình cần được cải thiện để nhận diện chính xác hơn các trường hợp có mưa và đạt được sự cân nhắc giữa Precision và Recall.

Các công thức Liên Quan đến các mô hình Machine Learning:

1. **Precision (Độ chính xác):**
$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$
2. **Recall (Độ phủ):**
$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$
3. **F1-Score:**
$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Trong các công thức trên:

- True Positives (TP): Số lượng các mẫu dự đoán đúng là positive và thực tế cũng là positive.
- False Positives (FP): Số lượng các mẫu dự đoán là positive nhưng thực tế là negative.
- False Negatives (FN): Số lượng các mẫu dự đoán là negative nhưng thực tế là positive.

Với ma trận confusion, các giá trị này có thể được trích xuất để tính toán precision, recall và F1-score. Trong đoạn code, chúng ta sử dụng hàm calculate_confusion_matrix để tính ma trận confusion, sau đó sử dụng các giá trị trong ma trận này để tính toán các chỉ số.

Precision cao chỉ ra rằng mô hình đưa ra ít dự đoán tích cực sai, tức là ít false positive. Nó là một thước đo quan trọng khi mục tiêu là giảm thiểu false positive.

Recall cao chỉ ra rằng mô hình có khả năng tìm ra nhiều mẫu positive hơn, tức là ít false negative.

F1-score là trung bình điều hòa của precision và recall. Nó cung cấp một đánh giá tổng thể về hiệu suất của mô hình bằng cách cân nhắc cả precision và recall.

CHƯƠNG IV: KẾT LUẬN:

So sánh các kết quả tốt nhất ở các trường hợp cho từng mô hình:

Mô Hình	Accuracy	Precision		Recall		F1-Score	
		Yes	No	Yes	No	Yes	No
Logistic Regression	84.09%	69.88%	85.46%	31.57%	96.73%	43.49%	43.49%
Random Forest Classifier	83.80%	74.52%	84.55%	27.92%	97.64%	40.62%	40.62%
Decision Tree Classifier	83.50%	72.90%	84.53%	29.68%	97.21%	42.18%	42.18%
Support Vector Machine	80.50%	0.00%	80.50%	0.00%	100.00%	0.00%	0.00%
K-Nearest Neighbors	81.26%	55.97%	83.70%	24.84%	95.18%	34.41%	34.41%

Kết Luận:

1. Với tập dữ liệu này thì mô hình sẽ đạt kết quả dự đoán tốt khi sử dụng bốn biến để dự đoán.

2. Logistic Regression là mô hình hoạt động tốt nhất cho việc dự báo thời tiết ở Úc, với độ chính xác cao và F1-Score tương đối ổn định. Tuy nhiên, cần lưu ý rằng Precision và Recall cho lớp "Yes" còn thấp, cho thấy khả năng dự đoán những trường hợp những ngày không có mưa còn hạn chế. Support Vector Machine là mô hình này có độ chính xác thấp hơn đáng kể (80.50%). Các chỉ số Precision, Recall, và F1-Score bằng 0 cho lớp "Yes", điều này cho thấy mô hình không thể dự đoán chính xác cho tập dữ liệu trên.

3. Dựa trên EDA và các mô hình dự đoán, chúng ta có thể dự đoán rằng liệu ngày mai có mưa hay không với độ chính xác cao nhất từ mô hình Logistic Regression. Mô hình này cho thấy hiệu suất tốt nhất trong số các mô hình đã thử nghiệm, nhưng cần cải thiện khả năng dự đoán các trường hợp "Yes" để tăng độ chính xác của dự báo lượng mưa hơn.

CHƯƠNG VII: TÀI LIỆU THAM KHẢO

1. Thanh Nguyen (30/4/2024), Các bài giảng của thầy và *Final-project DA.pptx*.
2. Adamyoung (30/4/2024), Rain in Australia, https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package.
3. "Machine Learning và các khái niệm cơ bản," Trí tuệ nhân tạo, Ngày xuất bản: 08/02/2020, https://trituenhantao.io/kien-thuc/machine-learning-va-cac-khai-niem-co-ban/.
4. "Mô Hình Học Máy", https://www.techtarget.com/searchenterpriseai/tip/What-are-machine-learning-models-Types-and-examples.
5. Tuan Nguyen, "RanDom Forest algorithm", https://machinelearningcoban.com/tabml_book/ch_model/random_forest.html.