RAINING PREDICTION

OVERVIEW

- Purpose
- Data Exploring
- Data Preparation
- Model Implementation
- Conclusion

PURPOSE

- ทำนายว่าวันพรุ่งนี้ฝนจะตกหรือไม่
- เปรียบเทียบความแม่นยำของโมเดลที่ใช้ในการจำแนกประเภทของระดับความน่าเชื่อถือ

DATA EXPLORING

DATA DICTATION

Column meanings

To save space on the screen, most of the columns have abbreviated headings:

Heading		Meaning	Units
Date		Day of the month	
Day		Day of the week	first two letters
Temps	**Min**	Minimum temperature in the 24 hours to 9am. Sometimes only known to the nearest whole degree.	degrees Celsius
	Max	Maximum temperature in the 24 hours from 9am. Sometimes only known to the nearest whole degree.	degrees Celsius
Rain		Precipitation (rainfall) in the 24 hours to 9am. Sometimes only known to the nearest whole millimetre.	millimetres
Evap		"Class A" pan evaporation in the 24 hours to 9am	millimetres
Sun		Bright sunshine in the 24 hours to midnight	hours
Max wind gust	**Dirn**	Direction of strongest gust in the 24 hours to midnight	16 compass points
	Spd	Speed of strongest wind gust in the 24 hours to midnight	kilometres per hour
	Time	Time of strongest wind gust	local time hh:mm
9 am	**Temp**	Temperature at 9 am	degrees Celsius
	RH	Relative humidity at 9 am	percent
	Cld	Fraction of sky obscured by cloud at 9 am	eighths
	Dirn	Wind direction averaged over 10 minutes prior to 9 am	compass points
	Spd	Wind speed averaged over 10 minutes prior to 9 am	kilometres per hour
	MSLP	Atmospheric pressure reduced to mean sea level at 9 am	hectopascals
3 pm	**Temp**	Temperature at 3 pm	degrees Celsius
	RH	Relative humidity at 3 pm	percent
	Cld	Fraction of sky obscured by cloud at 3 pm	eighths
	Dirn	Wind direction averaged over 10 minutes prior to 3 pm	compass points
	Spd	Wind speed averaged over 10 minutes prior to 3 pm	kilometres per hour
	MSLP	Atmospheric pressure reduced to mean sea level at 3 pm	hectopascals

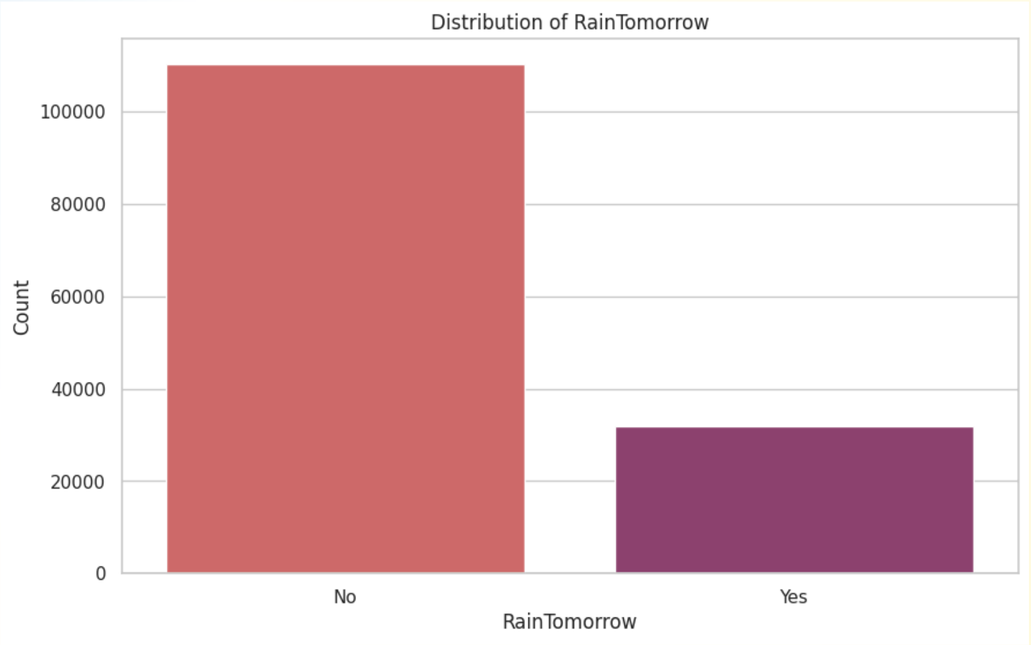
DATA DESCRIPTION

ขนาดข้อมูล

- 145460 Samples
- 23 parameters

Target Classes

- ฝนตก (Yes) , 31877 samples
- ฝนไม่ตก (0) , 110316 samples



จำนวนข้อมูลแต่ละประเภท ของค่าเป้าหมาย หรือ RainTomorrow

DATA DESCRIPTION

Data Info

#	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

Missing Data

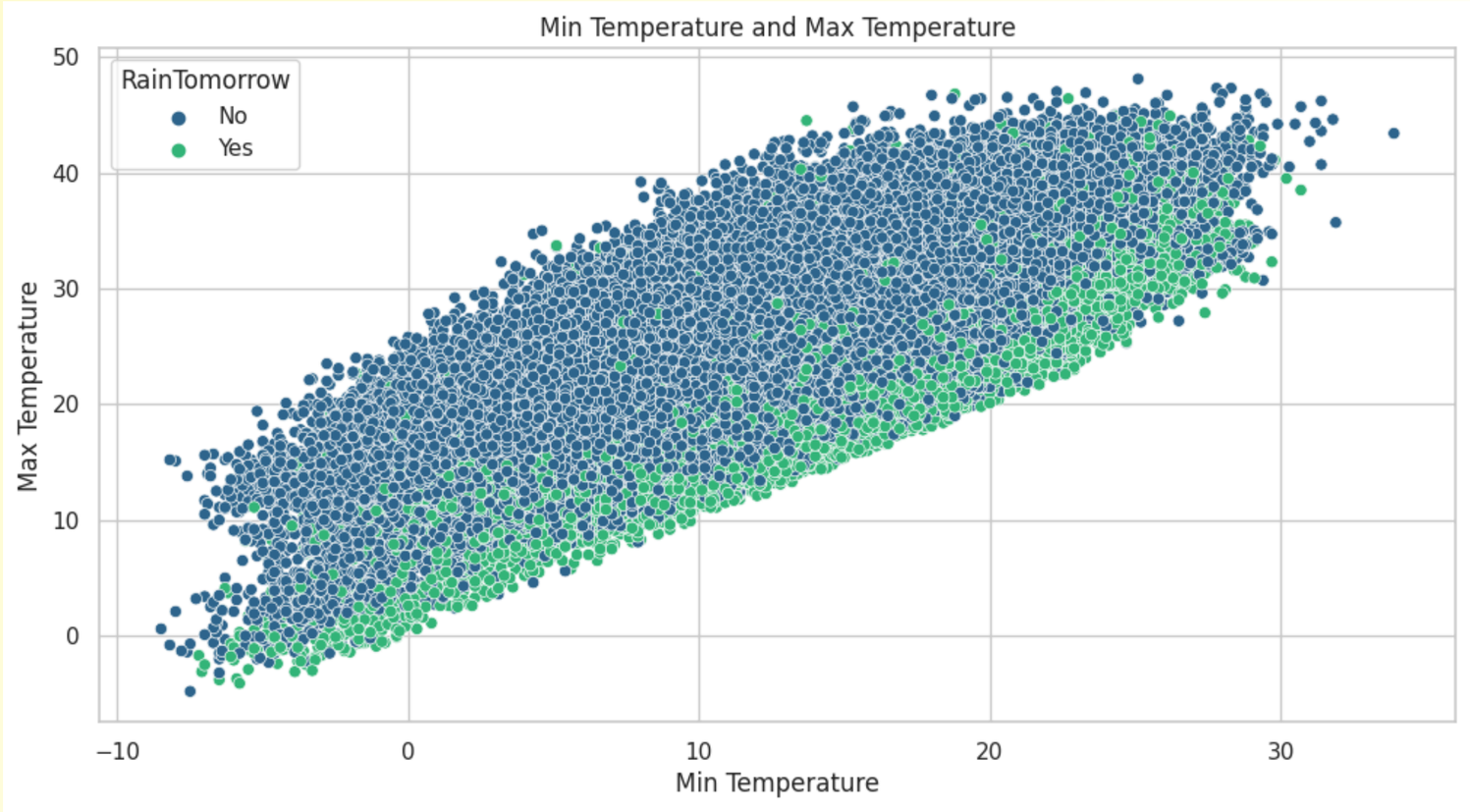
Date	0
Location	0
MinTemp	1485
MaxTemp	1261
Rainfall	3261
Evaporation	62790
Sunshine	69835
WindGustDir	10326
WindGustSpeed	10263
WindDir9am	10566
WindDir3pm	4228
WindSpeed9am	1767
WindSpeed3pm	3062
Humidity9am	2654
Humidity3pm	4507
Pressure9am	15065
Pressure3pm	15028
Cloud9am	55888
Cloud3pm	59358
Temp9am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

EXAMPLE

Min Temperature

And

Max Temperature

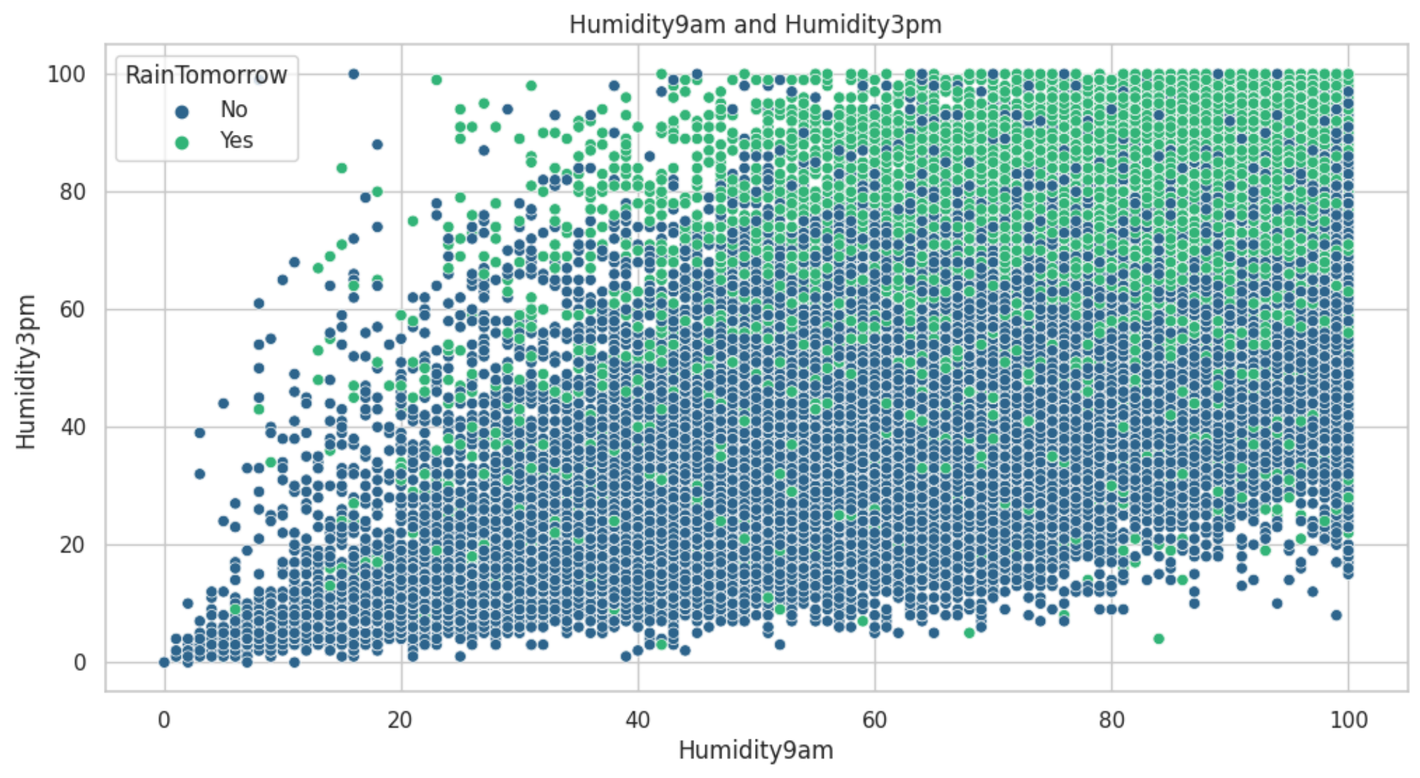


EXAMPLE

Humidity9am

And

Humidity3pm

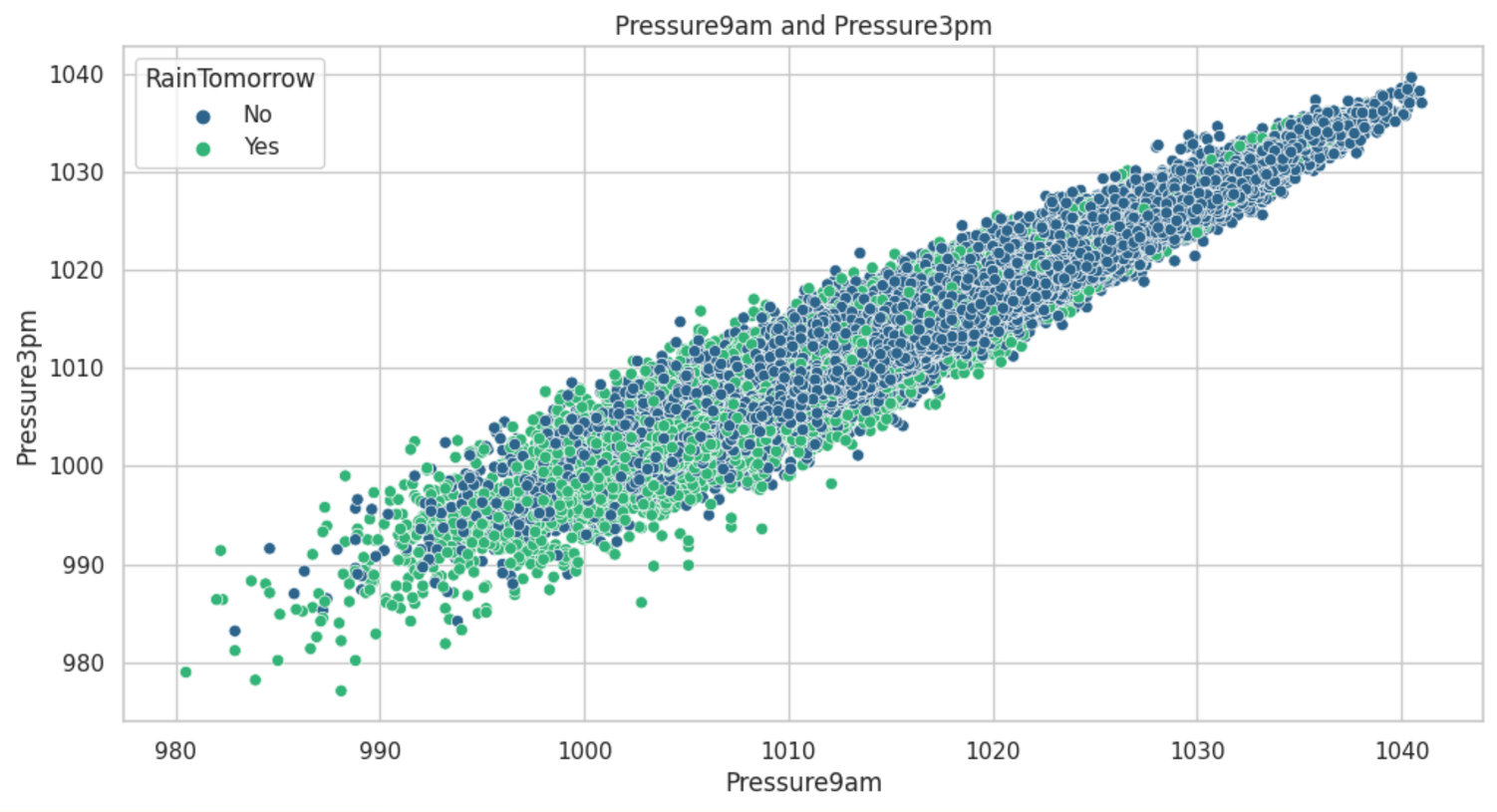


EXAMPLE

Pressure9am

And

Pressure3pm



DATA PREPARETION

DATA PREPARATION

Handle Missing Data

จัดการกับข้อมูลที่หายไป โดยการลบข้อมูลในส่วนที่ไม่ใช่ตัวเลขออก

Feature Engineering

สร้างฟีเจอร์ที่เหมาะสม โดยการตั้งเงื่อนไขกับตัวฟีเจอร์เพื่อหาความสัมพันธ์

Data imputation

ต่อเติมข้อมูล โดยใช้ค่าทางสถิติเข้ามาแทนในส่วนของข้อมูลที่หายไป

Label Encoding

แปลงข้อมูลที่ไม่ได้เป็นตัวเลขให้อยู่ในรูปของตัวเลข

Feature Scaling

ปรับสเกลของข้อมูล เพื่อลดการเหวี่ยงของตัวแปรที่มีค่ามาก

Feature Selection

เลือกฟีเจอร์ที่เหมาะสม โดยใช้ Pearson Correlation

DATA PREPARATION

Handle Missing Data

Before

```
df.shape
```

```
(145460, 23)
```

ก่อนตัดข้อมูลออก ในส่วนของข้อมูลตัวอักษร

After

```
df_drop.shape
```

```
(123710, 23)
```

หลังตัดข้อมูลออก

DATA PREPARATION

Feature Enginerring

Create Feature

- สร้างโดยใช้วิธีดูลักษณะข้อมูลที่เป็นตัวเลข ที่เกิดการหายไป (missing value)

New Features

- Is_MinTemp
- Is_MaxTemp
- Is_Rainfal
- Is_Evaporation
- Is_Sunshine
- Is_WindGustSpeed
- Is_WindSpeed9am
- Is_WindSpeed3pm
- Is_Humidity9am
- Is_Humidity3pm
- Is_Pressure9am
- Is_pressure3pm
- Is_Cloud9am
- Is_Cloud3pm
- Is_Temp9am
- Is_Temp3pm

16 features

DATA PREPARATION

Data Imputation

Filling Missing Data

- แทนที่ค่าที่หายไปด้วยค่าเฉลี่ย (mean)

Imputed Data

- ไม่มีข้อมูลที่หายไป พร้อมนำข้อมูลที่เป็นตัวเลข ไปดำเนินการต่อ

DATA PREPARATION

Label Encoding

String

- ตัวอักษร
- ข้อความ

Integer

- 0,1,2...,n

DATA PREPARATION

Feature Scaling

Stardard Scalar

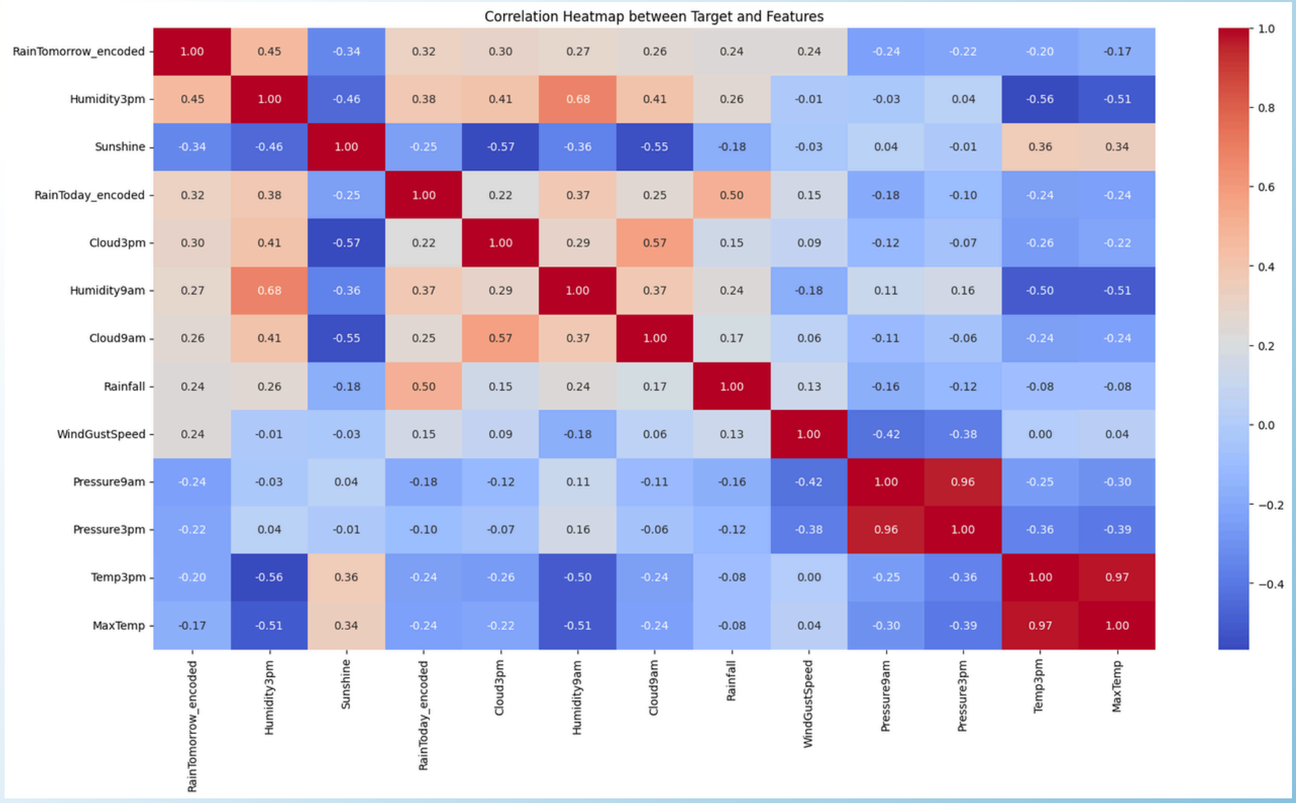
Scaled Data

DATA PREPARATION

Feature Selection

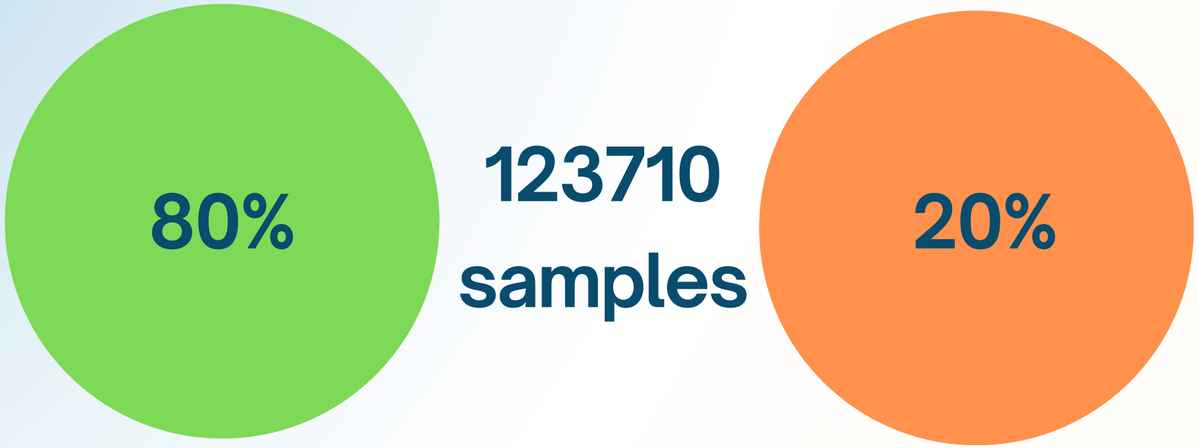
Pearson Correlation

- เลือกฟีเจอร์ที่ค่า
  มากกว่า 0.16



MODEL IMPREMENTATION

MODEL



Train Set

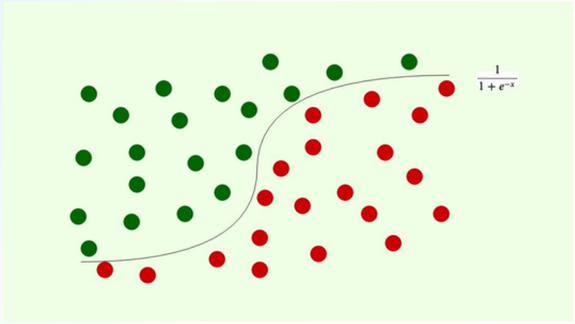
98968 sample

Test Set

24742 samples

MODEL

Decision Tree



Logistic Regression

Random Forest

MODEL

Decision Tree

- Accuracy = 0.8329

Hyperparameters	Values
criterion	entropy
max_depth	8
min_samples_split	10
min_samples_leaf	4

Confusion Matrix

```
Classification Report:
              precision    recall  f1-score   support

           0     0.8562    0.9406    0.8964     15850
           1     0.7064    0.4748    0.5679      4768

    accuracy                         0.8329     20618
   macro avg     0.7813    0.7077    0.7322     20618
weighted avg     0.8216    0.8329    0.8205     20618
```

MODEL

Logistic Regression

- Accuracy = 0.8416

Hyperparameters	Values
penalty	L1
C	1
solver	liblinear
max_iter	1000

Confusion Metrix

```
Classification Report:
              precision    recall  f1-score   support

           0     0.8636    0.9457    0.9028     19230
           1     0.7163    0.4788    0.5739      5512

    accuracy                         0.8416     24742
   macro avg     0.7900    0.7122    0.7383     24742
weighted avg     0.8308    0.8416    0.8295     24742
```

MODEL

Random Forest

- Accuracy = 0.8393

Hyperparameters	Values
n_estimators	200
criterion	gini
max_depth	8
max_features	auto

Confusion Metrix

```
Classification Report:
              precision    recall  f1-score   support

           0     0.8530    0.9557    0.9014     15850
           1     0.7545    0.4524    0.5656      4768

    accuracy                         0.8393     20618
   macro avg     0.8037    0.7041    0.7335     20618
weighted avg     0.8302    0.8393    0.8238     20618
```

CONCLUSION

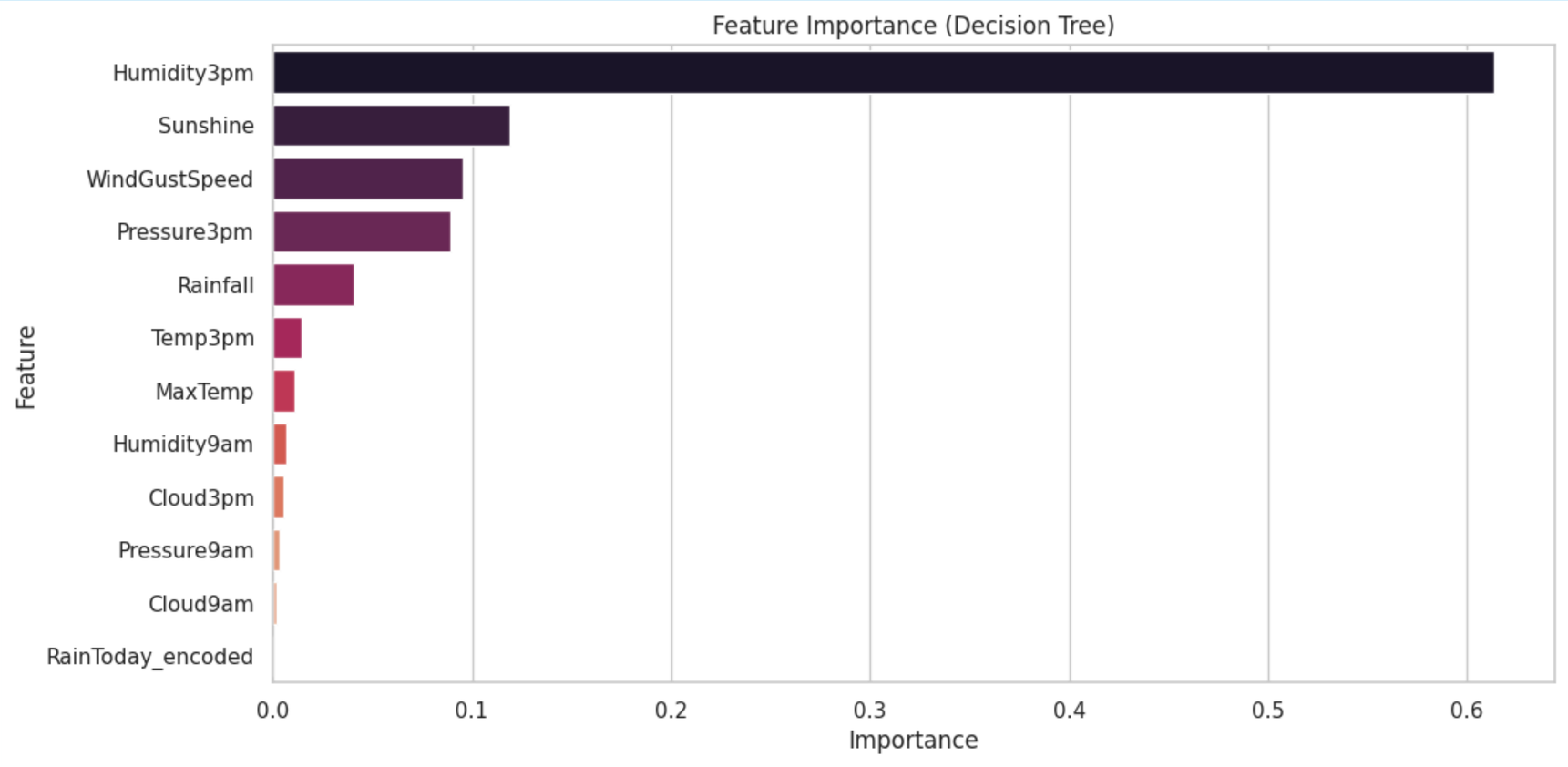
RESULTS

Model	Test score	Validation score	Precision	Recall	F1-Score
Decision Tree	0.8329	0.8398	0.8562	0.9406	0.8964
Logistic Regression	0.8416	0.8439	0.8636	0.9457	0.9028
Random Forest	0.8393	0.8472	0.8530	0.9557	0.9014

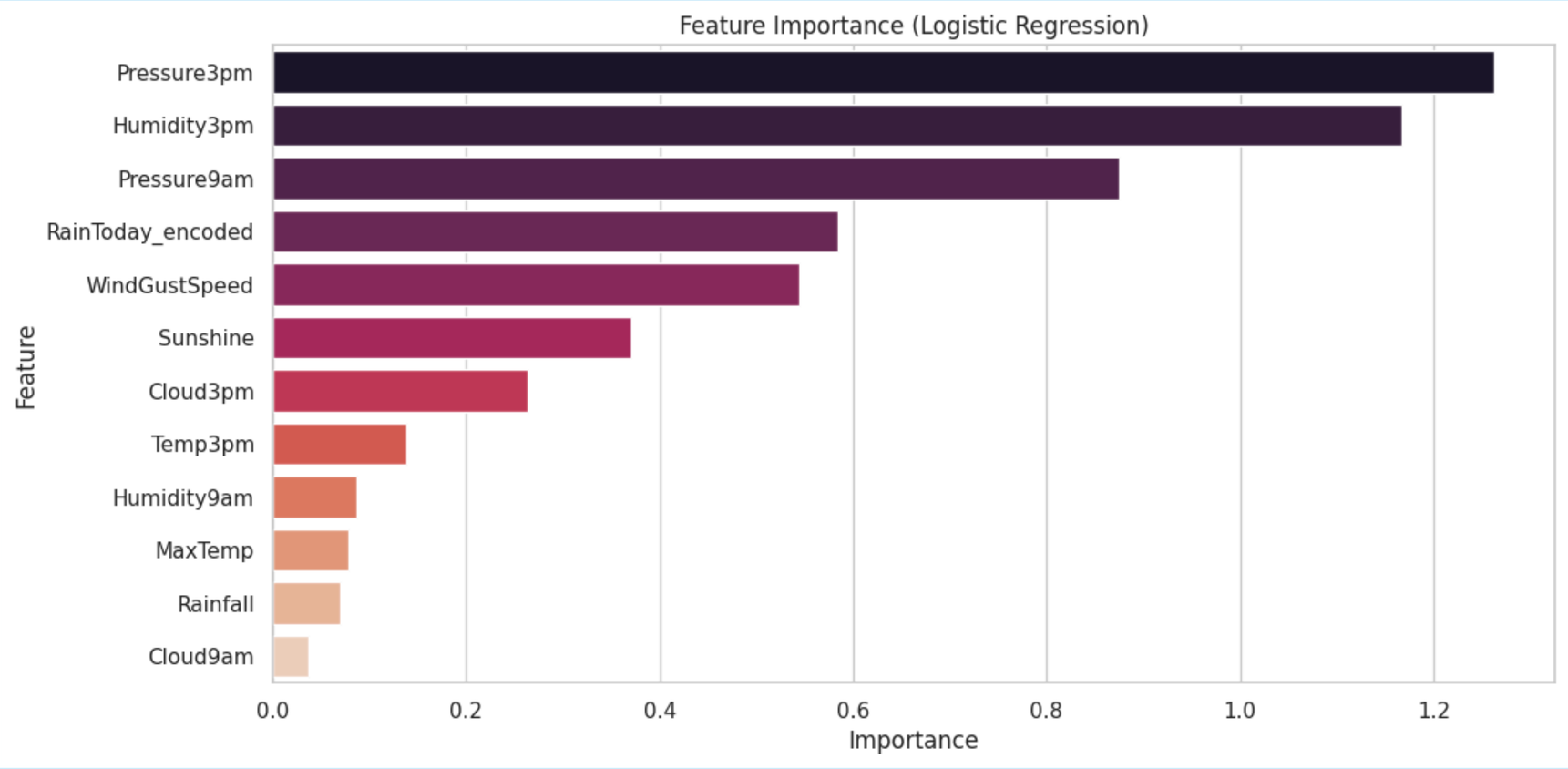
Highest Score !!

FEATURE IMPORTANT

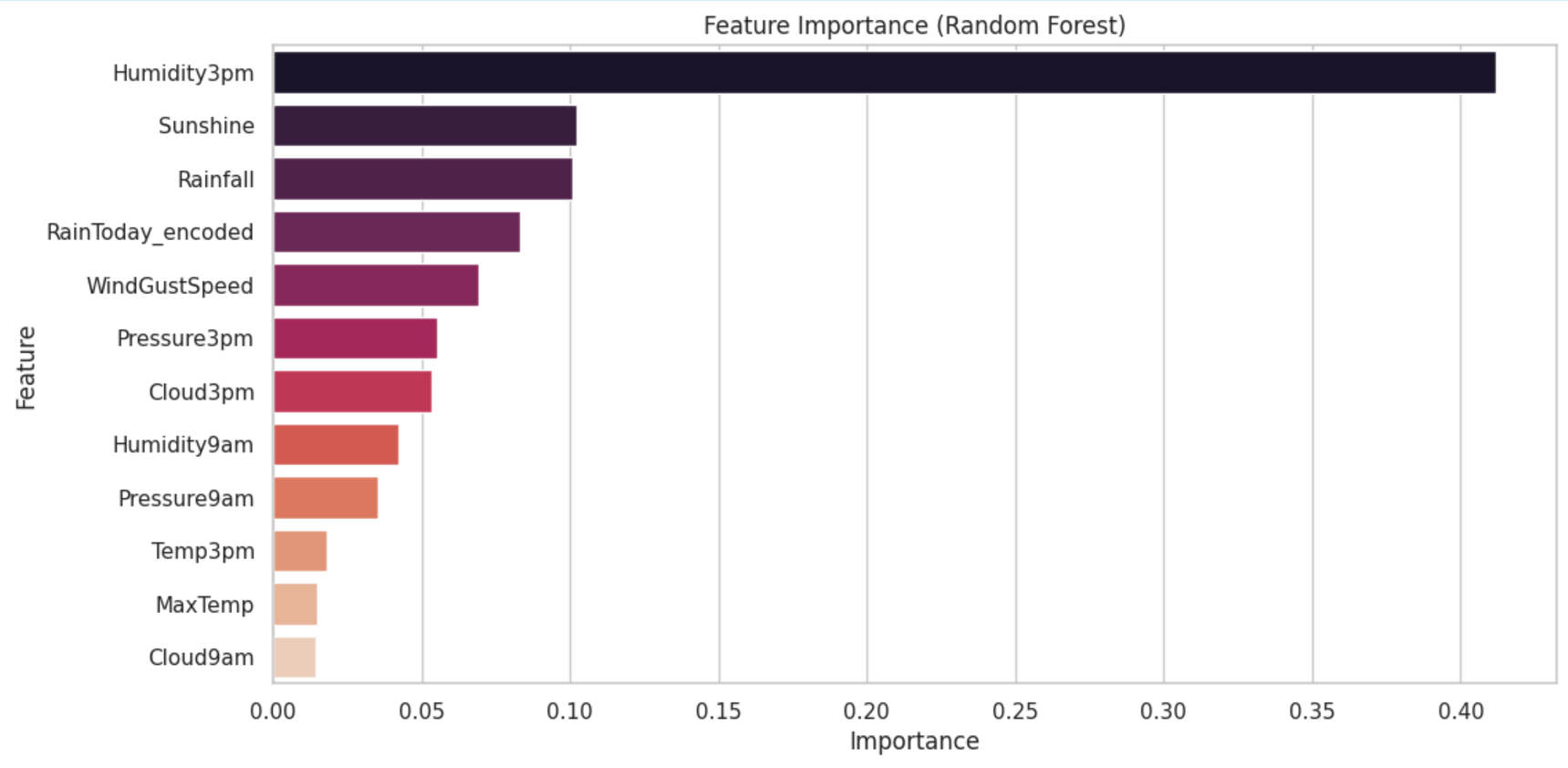
Decision Tree



Logistic Regression



Random Forest



Q&A

Thank you for your Attention