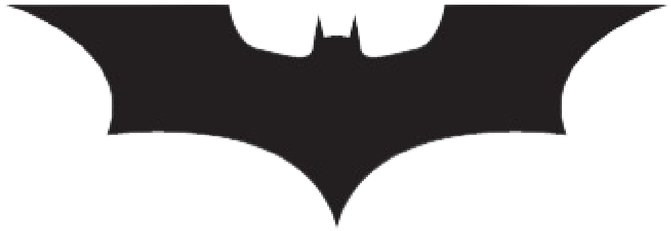
DigitalSkola
Uncover The World of Digital Skills with Us

Project Team Batman



Rain in Australia

Introduction

Team 3 : BATMAN

Farhan Adyatma

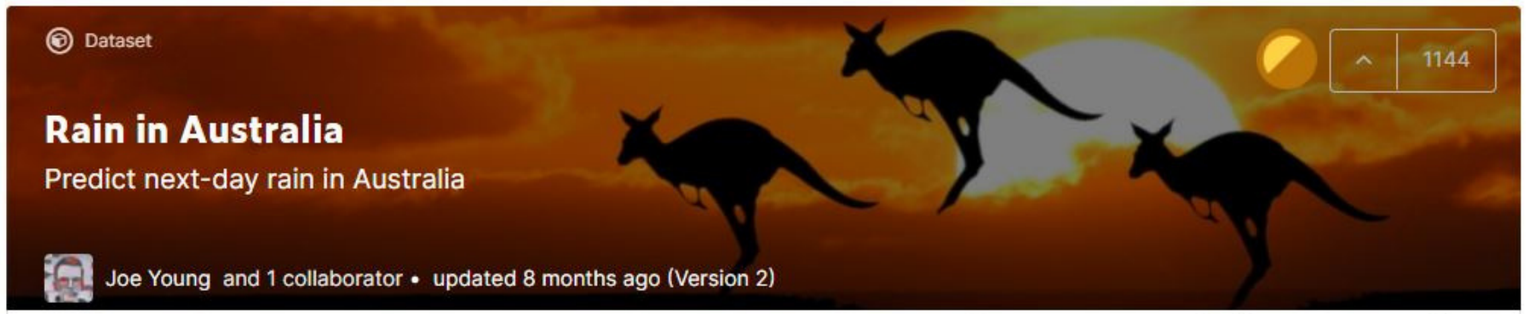
Saddam Annias

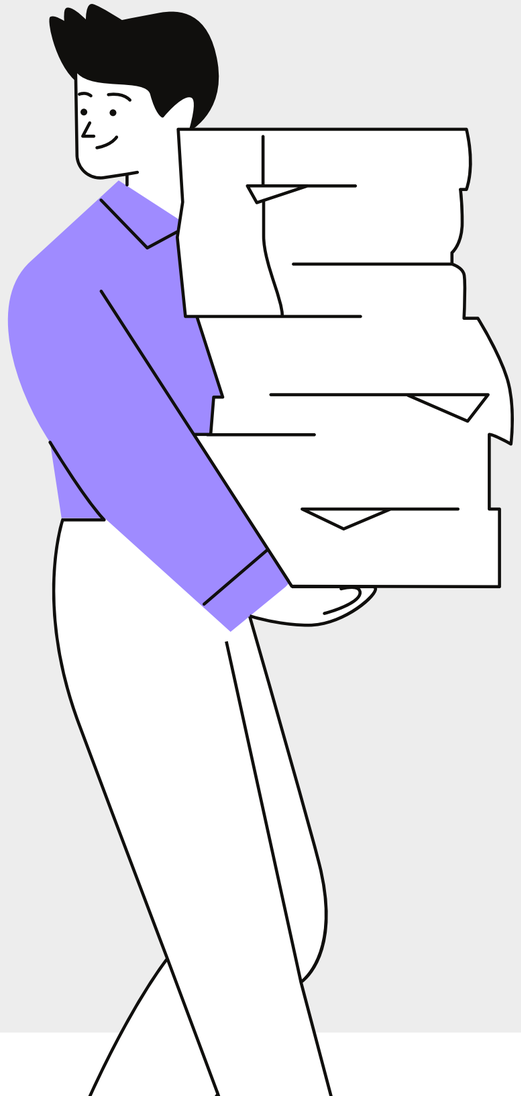
Wahyu Hidayat

Main Topics

https://www.kaggle.com/jsphyg/weather-dataset-rattle-package

Rain In Australia





Today's Agenda

1. DATA UNDERSTANDING
2. EXPLORATORY DATA ANALYSIS (EDA)
3. DATA PRE-PROCESSING
4. MACHINE LEARNING DEVELOPMENT

Data Understanding

Dataset ini kita peroleh dari kaggle, Data ini diperoleh dari BMKG nya australia, dataset ini merupakan hasil observasi selama beberapa tahun dan terdiri dari beberapa wilayah sana.

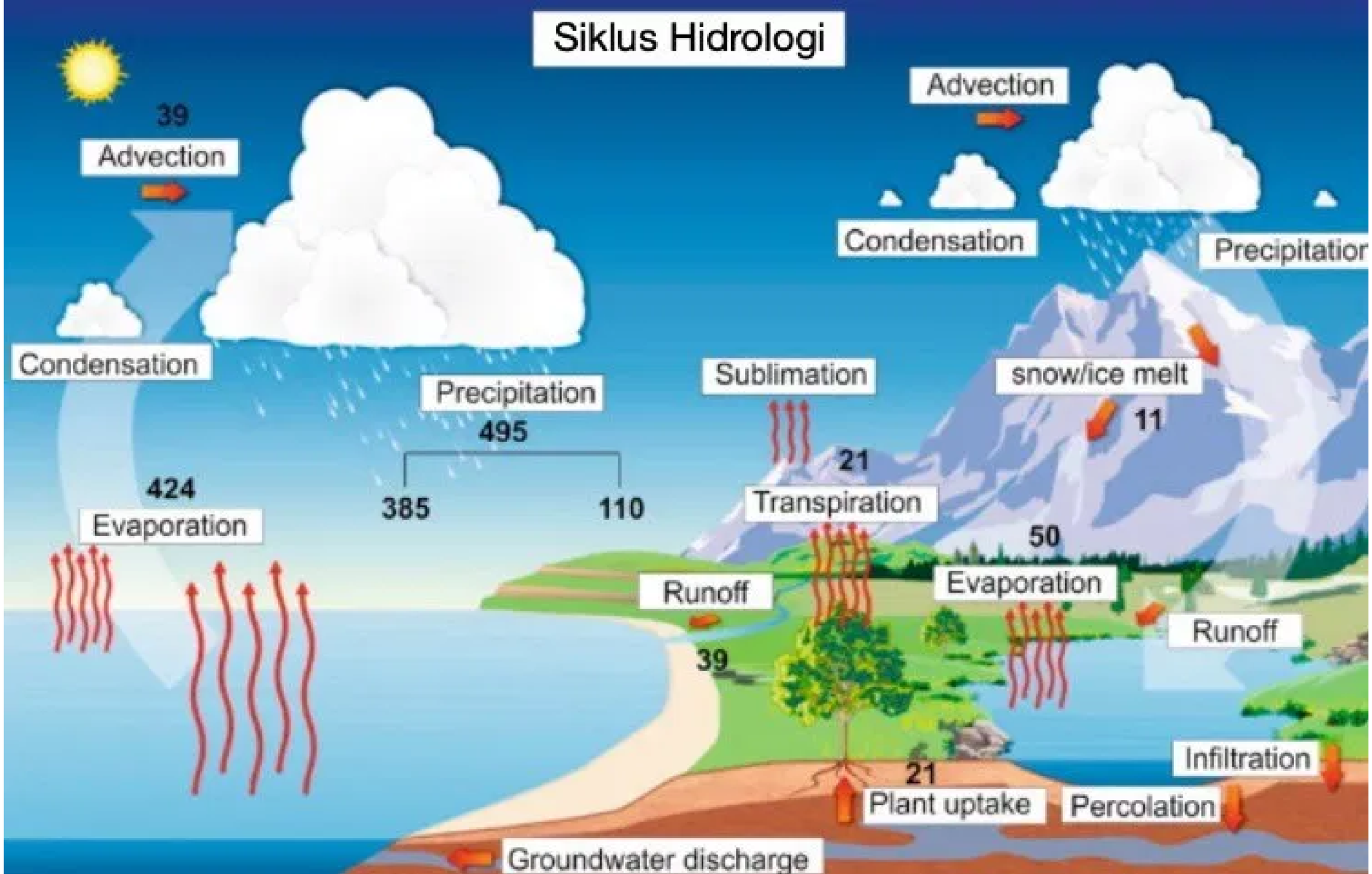
DigitalSkola
Uncover The World of Digital Skills with Us

Data Understanding

Apa itu hujan?
Hujan adalah peristiwa turunnya butir-butir air dari langit ke permukaan bumi akibat terjadinya kondensasi.

Apa saja yang dapat mempengruhi hujan?

- **Suhu**
- **Angin**
- **Kelembapan Udara**
- **Tekanan Udara**
- **Radiasi Matahari**


Siklus Hidrologi
39
Advection
Condensation
Precipitation
495
385
110
424
Evaporation
Sublimation
21
Transpiration
Runoff
39
Advection
Condensation
Precipitation
snow/ice melt
11
50
Evaporation
Runoff
Infiltration
21
Plant uptake
Percolation
Groundwater discharge

Siklus hidrologi adalah tahapan yang terjadi di lingkungan perairan (seperti laut, danau, sungai, tanah, dan atmosfer). Siklus ini akan terus berkelanjutan hingga ketersediaan air di bumi tidak pernah habis.

Urutan siklus hidrologi dalam prosesnya, dimana air dari atmosfer turun ke bumi dalam bentuk hujan dan juga salju akan kembali ke atmosfer secara berulang dan terus menerus.

Deskripsi Tiap Kolom

1. **Date, Tanggal pengambilan data**
2. **Location, Lokasi tempat pengambilan data.**
3. **MinTemp, Temperature terendah dalam derajat Celcius.**
4. **MaxTemp, Temperature tertinggi dalam derajat Celsius.**
5. **Rainfall, Curah hujan dalam satuan mm.**
6. **Evaporation, Tingkat evaporasi dalam satuan mm.**
7. **Sunshine, Lama waktu sinar matahari cerah dalam satuan jam.**
8. **WindGustDir, Arah angin terkuat.**
9. **WindGustSpeed, Kecepatan (km/h) angin terkuat.**
10. **WindDir9am, Arah angin ketika jam 9 pagi.**

Deskripsi Tiap Kolom

11. **WindDir3pm, Arah angin ketika jam 3 sore.**
12. **WindSpeed9am, Rata-rata kecepatan angin (km/hr) dalam waktu 10 mnt menuju jam 9 pagi.**
13. **WindSpeed3pm, Rata-rata kecepatan angin (km/hr) dalam waktu 10 mnt menuju jam 3 sore.**
14. **Humidity9am, Kelembapan (persen) jam 9 pagi**
15. **Humidity3pm, Kelembapan (persen) jam 3 sore**
16. **Pressure9am, Tekanan atmosfer (hpa) dikurangi menjadi rata-rata pada permukaan laut pukul 9 pagi**
17. **Pressure3pm, Tekanan atmosfer (hpa) dikurangi menjadi rata-rata pada permukaan laut pukul 3 sore**

Deskripsi Tiap Kolom

18 **Cloud9am, Bagian langit yang tertutup awan pada jam 9 pagi. Diukur pada satuan ”oktas”. Angka 0 menandakan langit tanpa awan sedangkan angka 8 menandakan langit tertutup semua oleh awan.**

19 **Cloud3pm, Bagian langit yang tertutup awan pada jam 3 sore.**

20 **Temp9am, Temperatur (derajat C) pukul 9 pagi.**

21 **Temp3pm, Temperatur (derajat C) pukul 3 sore.**

22 **RainToday, Boolean: 1 jika hujan, 0 jika tidak hujan**

23 **RainTomorrow, 1 jika hujan, 0 jika tidak hujan pada hari berikutnya**

Menampilkan 5 baris data

Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed
2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0
2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0
2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0
2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0
2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0

WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am
W	WNW	20.0	24.0	71.0	22.0	1007.7
NNW	WSW	4.0	22.0	44.0	25.0	1010.6
W	WSW	19.0	26.0	38.0	30.0	1007.6
SE	E	11.0	9.0	45.0	16.0	1017.6
ENE	NW	7.0	20.0	82.0	33.0	1010.8

Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
1007.1	8.0	NaN	16.9	21.8	No	No
1007.8	NaN	NaN	17.2	24.3	No	No
1008.7	NaN	2.0	21.0	23.2	No	No
1012.8	NaN	NaN	18.1	26.5	No	No
1006.0	7.0	8.0	17.8	29.7	No	No

Tipe Variable Masing-Masing Kolom

Numerical

- Continuous :

MinTemp, MaxTemp, Rainfall, Evaporation, Sunshine, WindGustSpeed, WindSpeed9am, WindSpeed3pm, Humidity9am Humidity3pm, Pressure9am, Pressure3pm, Temp9am, Temp3pm, Cloud9am & Cloud3pm

- Discrete : None

Catgeorical

- Ordinal : Date
- Nominal : Location, WindGustDir, WindDir9am, WindDir3pm, RainToday, RainTomorrow

Missing Value

The number of missing value per column :

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

Percentage of the missing value per column :

Date	0.000000
Location	0.000000
MinTemp	1.020899
MaxTemp	0.866905
Rainfall	2.241853
Evaporation	43.166506
Sunshine	48.009762
WindGustDir	7.098859
WindGustSpeed	7.055548
WindDir9am	7.263853
WindDir3pm	2.906641
WindSpeed9am	1.214767
WindSpeed3pm	2.105046
Humidity9am	1.824557
Humidity3pm	3.098446
Pressure9am	10.356799
Pressure3pm	10.331363
Cloud9am	38.421559
Cloud3pm	40.807095
Temp9am	1.214767
Temp3pm	2.481094
RainToday	2.241853
RainTomorrow	2.245978

Statistik Keseluruhan Setiap Variabel

	count	unique	top	freq	mean	min	25%	50%	75%	max	std
Date	145460	NaN	NaN	NaN	2013-04-04 21:08:51.907053568	2007-11-01 00:00:00	2011-01-11 00:00:00	2013-06-02 00:00:00	2015-06-14 00:00:00	2017-06-25 00:00:00	NaN
Location	145460	49	Canberra	3436	NaN	NaN	NaN	NaN	NaN	NaN	NaN
MinTemp	143975.0	NaN	NaN	NaN	12.194034	-8.5	7.6	12.0	16.9	33.9	6.398495
MaxTemp	144199.0	NaN	NaN	NaN	23.221348	-4.8	17.9	22.6	28.2	48.1	7.119049
Rainfall	142199.0	NaN	NaN	NaN	2.360918	0.0	0.0	0.0	0.8	371.0	8.47806
Evaporation	82670.0	NaN	NaN	NaN	5.468232	0.0	2.6	4.8	7.4	145.0	4.193704
Sunshine	75625.0	NaN	NaN	NaN	7.611178	0.0	4.8	8.4	10.6	14.5	3.785483
WindGustDir	135134	16	W	9915	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WindGustSpeed	135197.0	NaN	NaN	NaN	40.03523	6.0	31.0	39.0	48.0	135.0	13.607062
WindDir9am	134894	16	N	11758	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WindDir3pm	141232	16	SE	10838	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WindSpeed9am	143693.0	NaN	NaN	NaN	14.043426	0.0	7.0	13.0	19.0	130.0	8.915375
WindSpeed3pm	142398.0	NaN	NaN	NaN	18.662657	0.0	13.0	19.0	24.0	87.0	8.8098
Humidity9am	142806.0	NaN	NaN	NaN	68.880831	0.0	57.0	70.0	83.0	100.0	19.029164
Humidity3pm	140953.0	NaN	NaN	NaN	51.539116	0.0	37.0	52.0	66.0	100.0	20.795902

	count	unique	top	freq	mean	min	25%	50%	75%	max	std
Pressure9am	130395.0	NaN	NaN	NaN	1017.64994	980.5	1012.9	1017.6	1022.4	1041.0	7.10653
Pressure3pm	130432.0	NaN	NaN	NaN	1015.255889	977.1	1010.4	1015.2	1020.0	1039.6	7.037414
Cloud9am	89572.0	NaN	NaN	NaN	4.447461	0.0	1.0	5.0	7.0	9.0	2.887159
Cloud3pm	86102.0	NaN	NaN	NaN	4.50993	0.0	2.0	5.0	7.0	9.0	2.720357
Temp9am	143693.0	NaN	NaN	NaN	16.990631	-7.2	12.3	16.7	21.6	40.2	6.488753
Temp3pm	141851.0	NaN	NaN	NaN	21.68339	-5.4	16.6	21.1	26.4	46.7	6.93665
RainToday	142199	2	No	110319	NaN	NaN	NaN	NaN	NaN	NaN	NaN
RainTomorrow	142193	2	No	110316	NaN	NaN	NaN	NaN	NaN	NaN	NaN

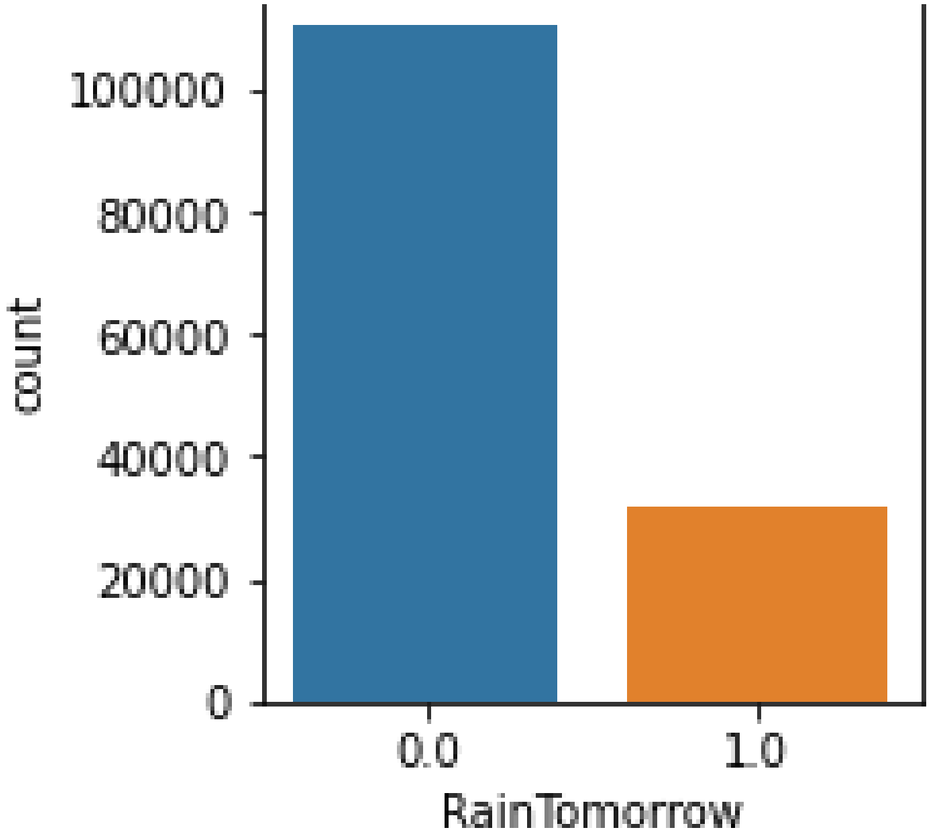
Exploratory Data Analysis

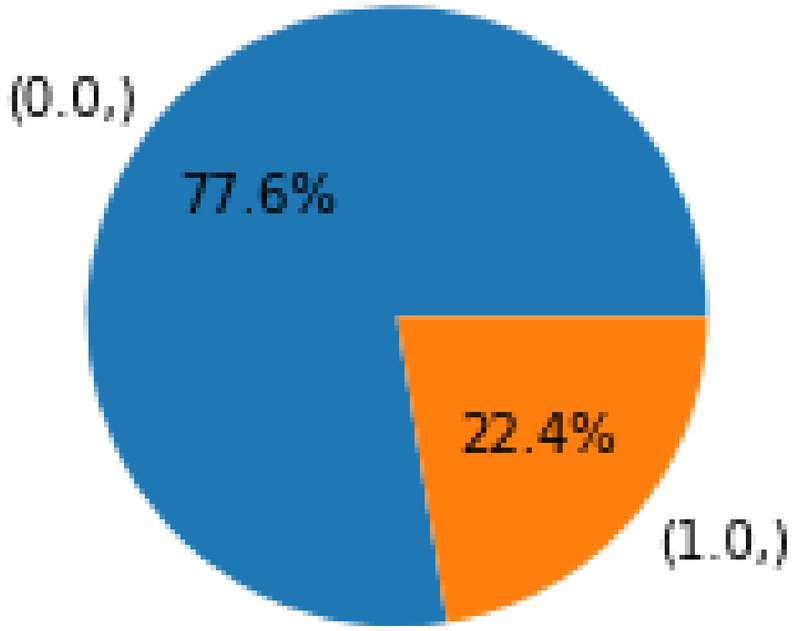
Dalam Tahap ini Kami akan mengeksplorasi data lebih dalam ke dataset untuk menemukan hubungan antara fitur dan hasil yang diinginkan.

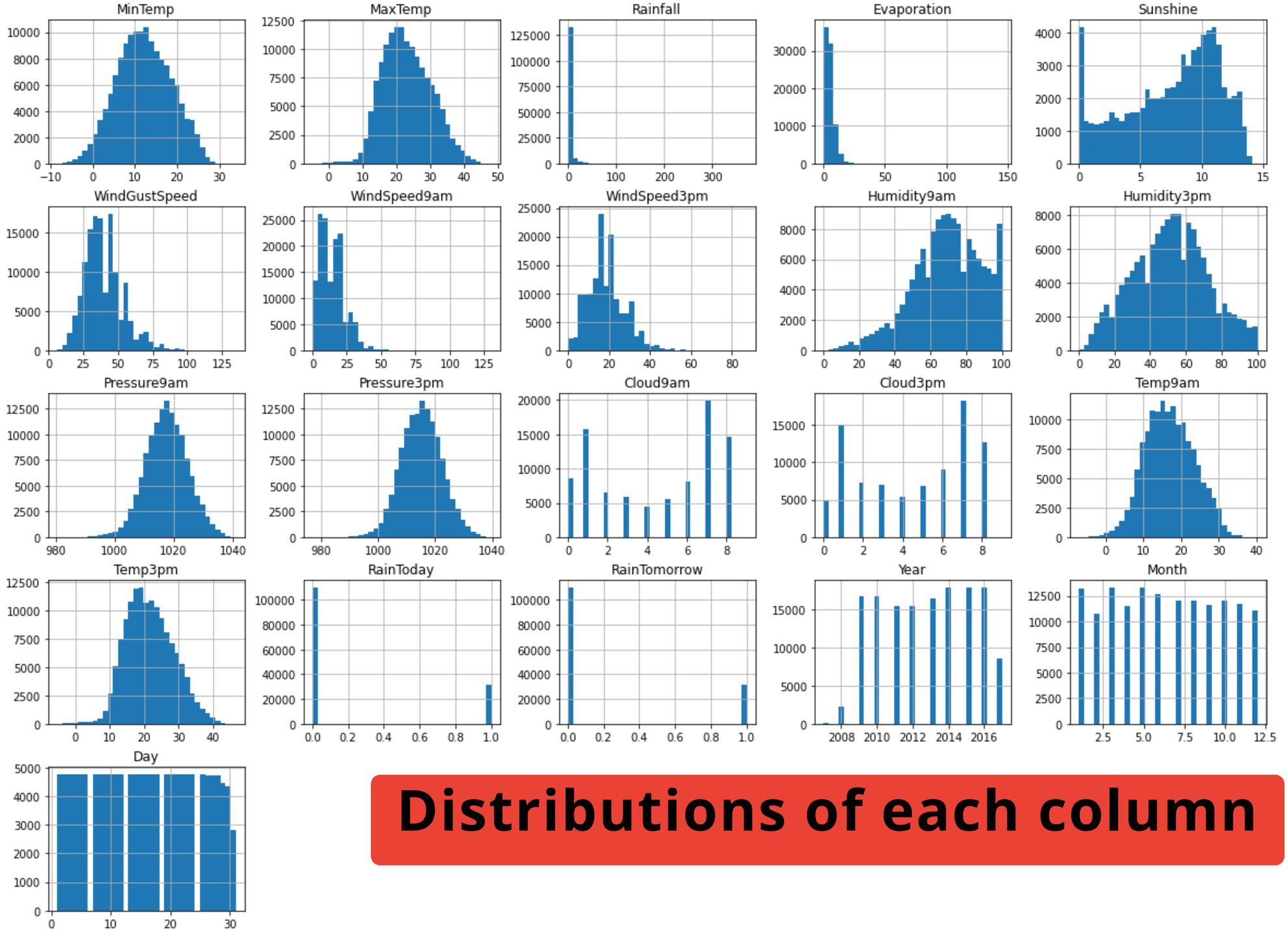
```
#Changing the RainToday and RainTomorrow Column into boolean in order to make analysis easier
df.loc[df['RainToday']=='No', 'RainToday']=0
df.loc[df['RainToday']=='Yes', 'RainToday']=1
df.loc[df['RainTomorrow']=='No', 'RainTomorrow']=0
df.loc[df['RainTomorrow']=='Yes', 'RainTomorrow']=1

df.RainToday=df.RainToday.astype(float)
df.RainTomorrow=df.RainTomorrow.astype(float)
```

Univariate Analysis

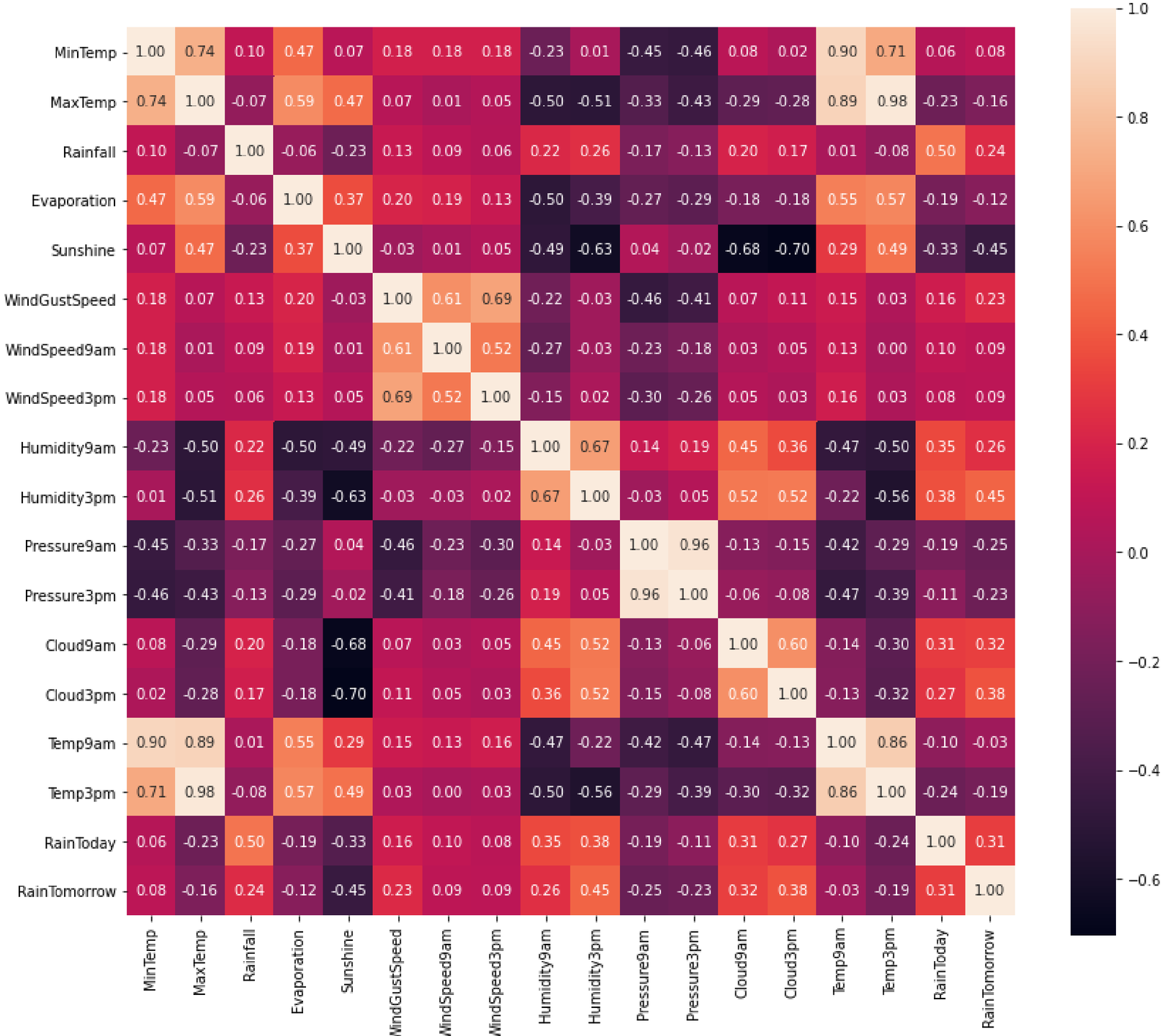
```
The frequency of variable in target columns is :
 RainTomorrow
0.0              110316
1.0               31877
```






MinTemp
MaxTemp
Rainfall
Evaporation
Sunshine
WindGustSpeed
WindSpeed9am
WindSpeed3pm
Humidity9am
Humidity3pm
Pressure9am
Pressure3pm
Cloud9am
Cloud3pm
Temp9am
Temp3pm
RainToday
RainTomorrow
Year
Month
Day
Distributions of each column

Bivariate Analysis



-

- Rainfall memiliki korelasi yang rendah terhadap RainTomorrow, yaitu 0.24. Hal ini terjadi karena ketika terjadi hujan dengan curah hujan yang tinggi pada suatu hari, keesokan harinya juga memiliki probabilitas untuk hujan juga.

- Cloud memiliki korelasi terhadap RainTomorrow. Cloud3pm memiliki korelasi 0.38 sedangkan Cloud9am memiliki korelasi 0.32. Hal ini terajadi karena salah satu komponen untuk terjadinya hujan adalah awan. Ketika awan menutupi langit maka ada kemungkinan terjadinya hujan.

- Sunshine memiliki korelasi negatif yang sedang terhadap RainTomorrow yaitu -0.45. Hal ini terjadi karena sebelum terjadi hujan, awan gelap biasanya menyelimuti langit sehingga sinar matahari terhalang.

- Humidity memiliki korelasi yang cukup tinggi terhadap RainTomorrow. Khususnya Humidity3pm memiliki korelasi 0.45, sedangkan Humidity9am hanya berkorelasi 0.26. Hal ini terjadi karena sebelum terjadi hujan,awan gelap biasanya menyelimuti langit sehingga suhu udara menjadi lebih rendah. Penurunan suhu udara tersebut mengakibatkan kenaikan Humidity.

- Kolom yang mengukur ukuran yang sama seperti min temp dengan max temp, windgust9am dengan windgust3pm, dll cenderung memiliki korelasi yang cukup tinggi.

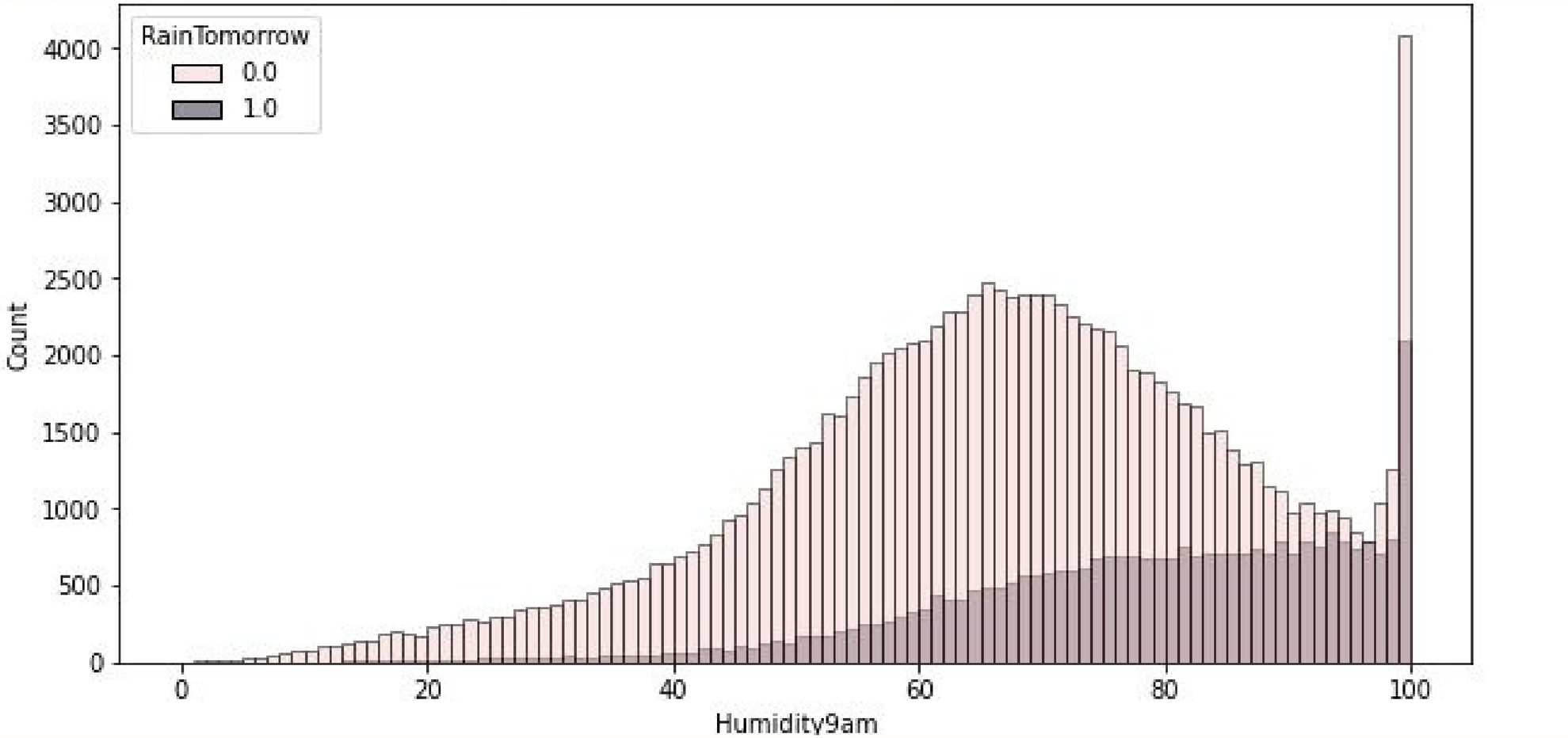
- Humidity dan temperature memiliki nilai korelasi negatif yang sedang yaitu sampai -0.50. Hal ini terjadi karena ketika suhu udara meningkat, udara dapat menahan lebih banyak molekul air, dan kelembaban relatifnya menurun. Ketika suhu turun, kelembaban relatif meningkat.

- Temperature dan tekanan memiliki nilai korelasi negatif yang sedang yaitu sampai -0.40. Hal ini terjadi karena ketika suhu udara meningkat, kerapatan antar udara akan menurun sehingga tekanan udaranya menurun juga.

- Cloud dan sunshine memiliki nilai korelasi negatif yang tinggi yaitu sampai -0.70. Hal ini terjadi karena ketika semakin tinggi nilai Cloud maka semakin rendah pula cahaya matahari yang menyentuh permukaan. Begitu pula sebaliknya.

- Tekanan udara dan Kecepatan angin memiliki nilai korelasi negatif yang sedang yaitu sampai -0.40. Hal ini terjadi karena angin merupakan suatu fluida. Udara akan memiliki tekanan yang lebih kecil ketika udara bergerak dengan cepat. Begitu pula sebaliknya.

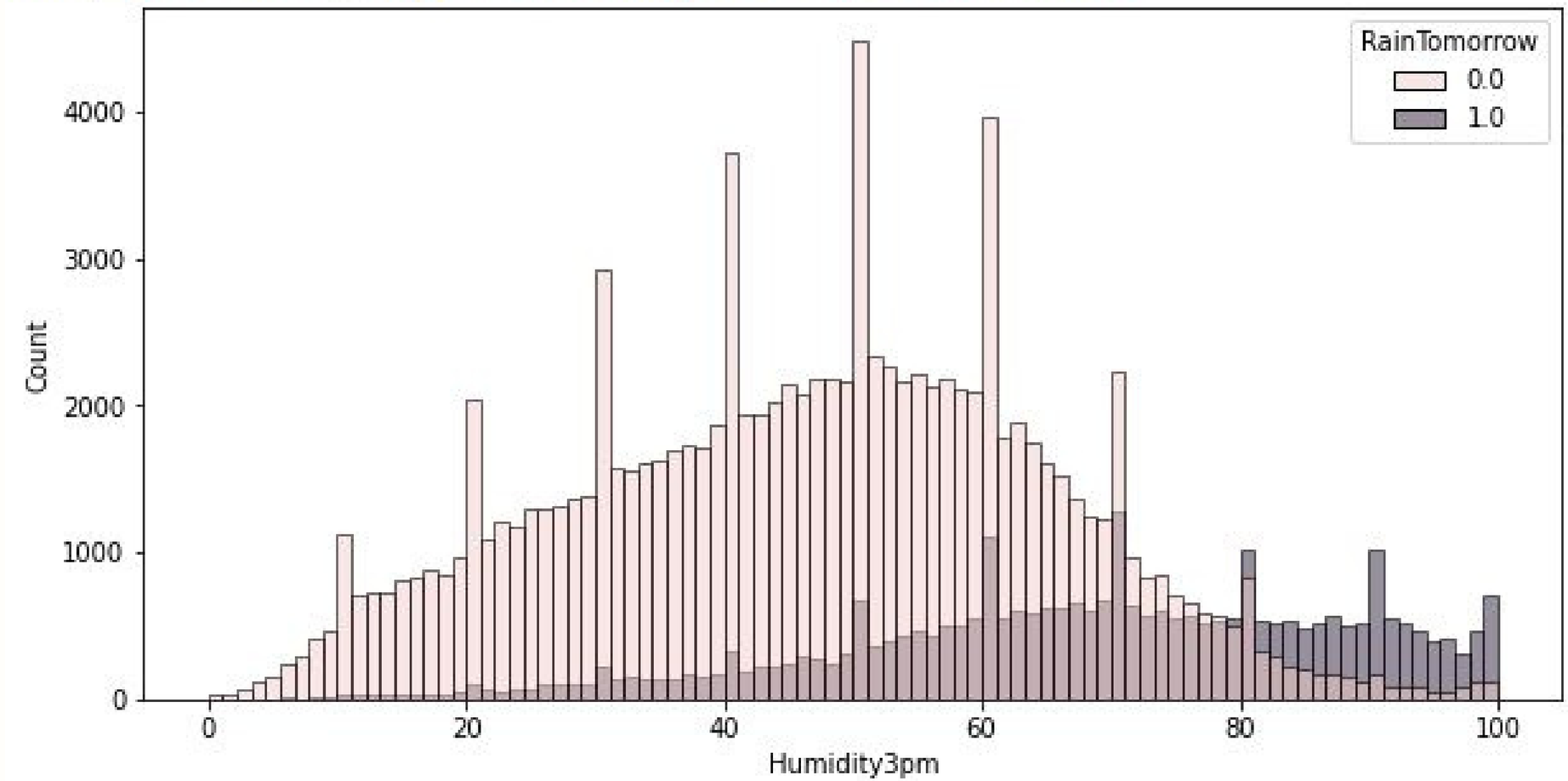
Distribusi dari Humidity9am dengan hue RainTomorrow



Jika Humidity9am bernilai lebih dari 85%, terdapat 38.94% kemungkinan RainTomorrow.

Jika Humidity9am bernilai kurang dari 85%, terdapat 16.95% kemungkinan RainTomorrow.

Distribusi dari Humidity3pm dengan hue RainTomorrow



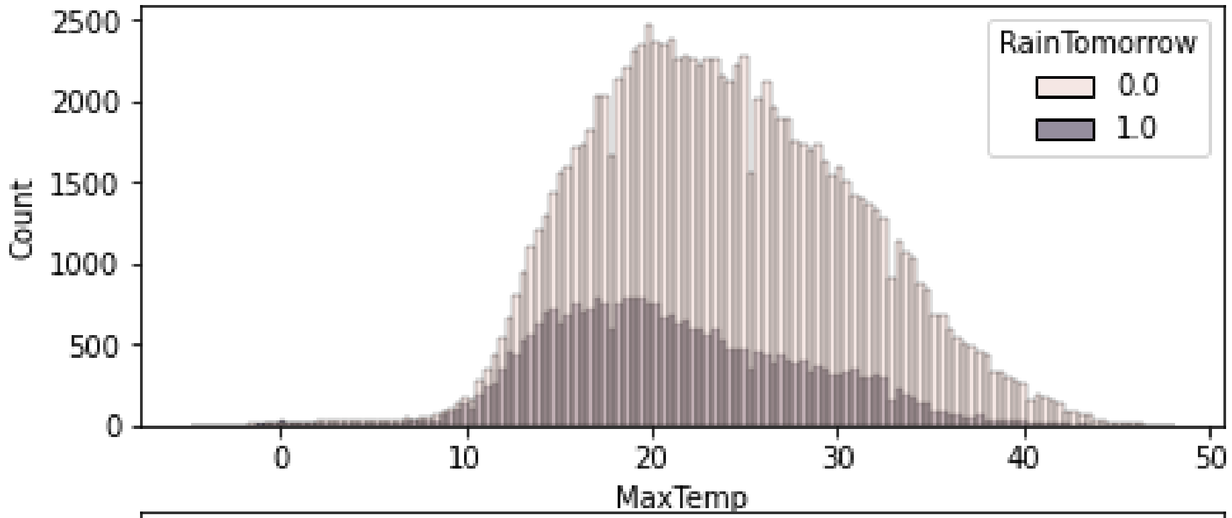
Jika Humidity3pm bernilai lebih dari 80%, terdapat 73.70% kemungkinan RainTomorrow.
Jika Humidity3pm bernilai kurang dari 80%, terdapat 16.40% kemungkinan RainTomorrow.

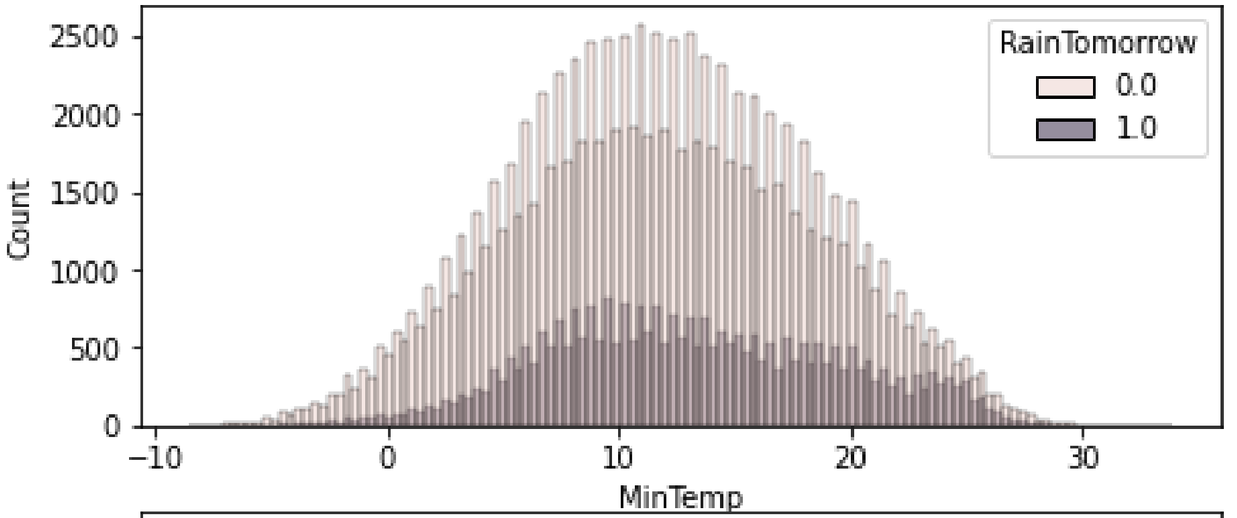
Kesimpulan dari Humidity dan RainTomorrow

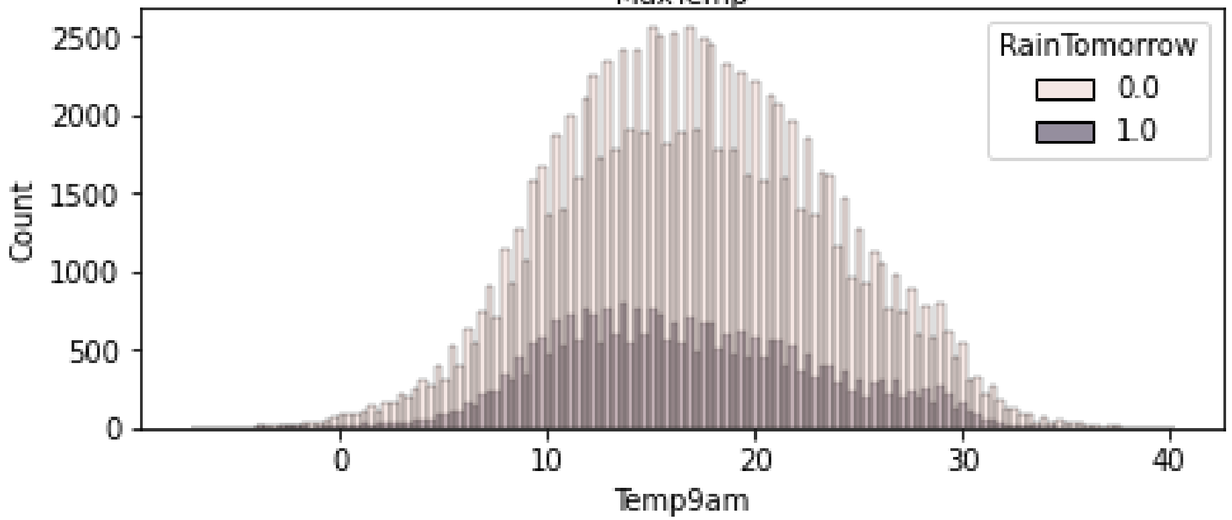
Nilai Humidity3pm sangat berpengaruh terhadap terjadinya hujan pada keesokan harinya karena memiliki probabilitas yang lumayan tinggi jika valuenya lebih dr 80.

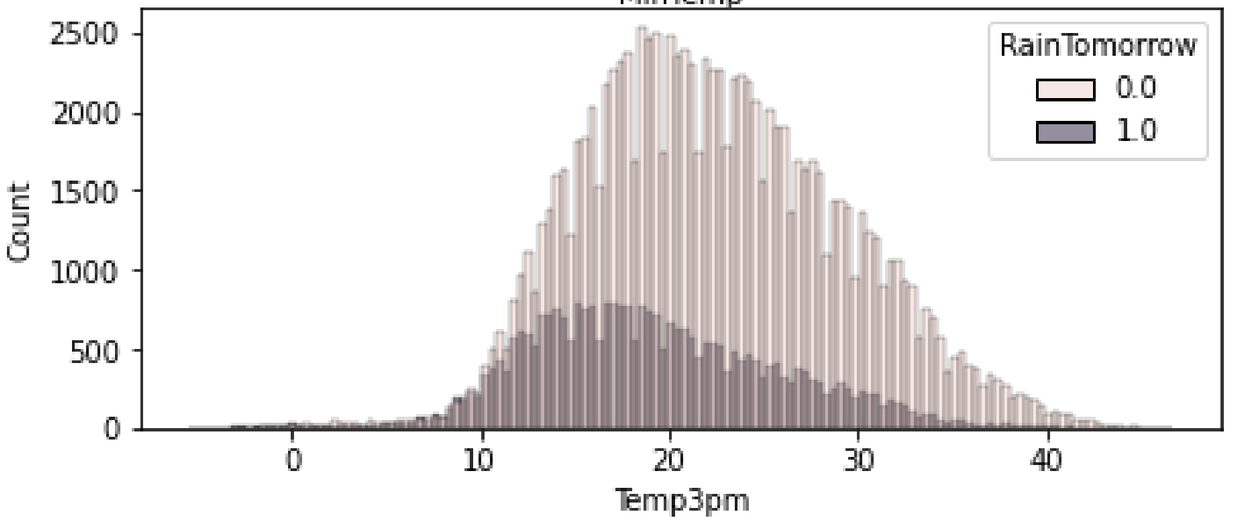
Kami akan membuat fitur baru bernama "HighHumidity" dengan nilai boolean dimana nilai 1 jika Humidity3pm di atas 80 dan nilai 0 jika Humidity3pm di bawah 80.

Distribusi dari Temperature dengan hue RainTomorrow









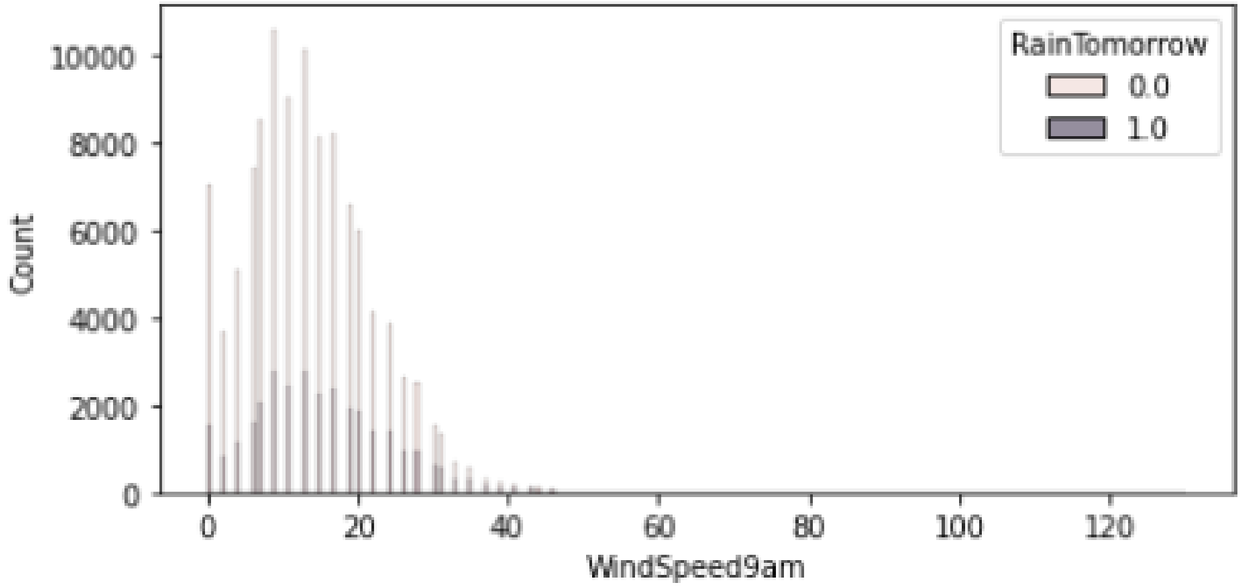
Kemungkinan RainTomorrow jika suhu di bawah 0°C sepanjang hari adalah 58.27 %

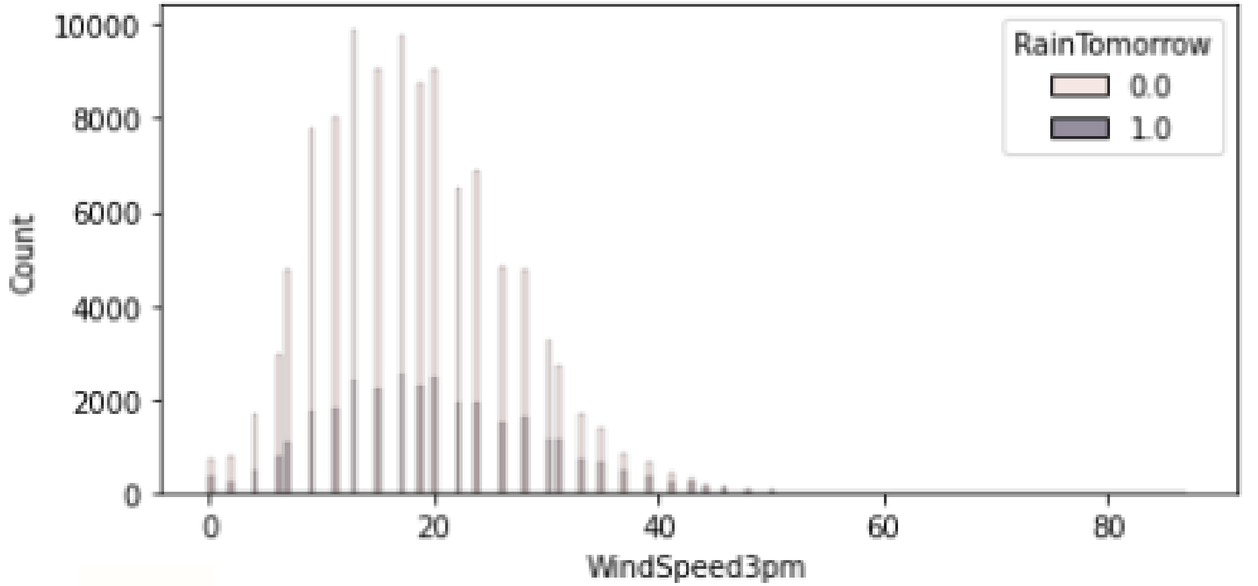
Kesimpulan dari Temperature dan RainTomorrow

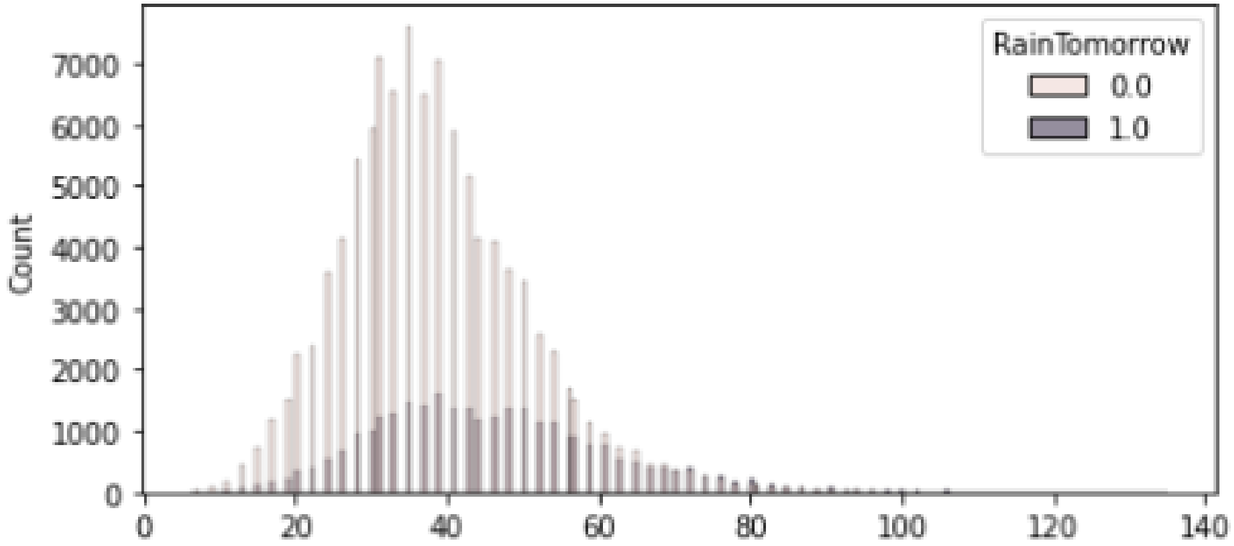
Nilai MaxTemp berpengaruh terhadap terjadinya hujan pada keesokan harinya karena memiliki probabilitas yang moderate jika valuenya kurang dr 0.

Kami akan membuat fitur baru bernama "Freezing" dengan nilai boolean dimana nilai 1 jika MaxTemp di bawah 0 dan nilai 0 jika MaxTemp di atas 0.

Distribusi dari WindSpeed dengan hue RainTomorrow







Kemungkinan RainTomorrow jika kecepatan angin lumayan tinggi adalah 56.20 %

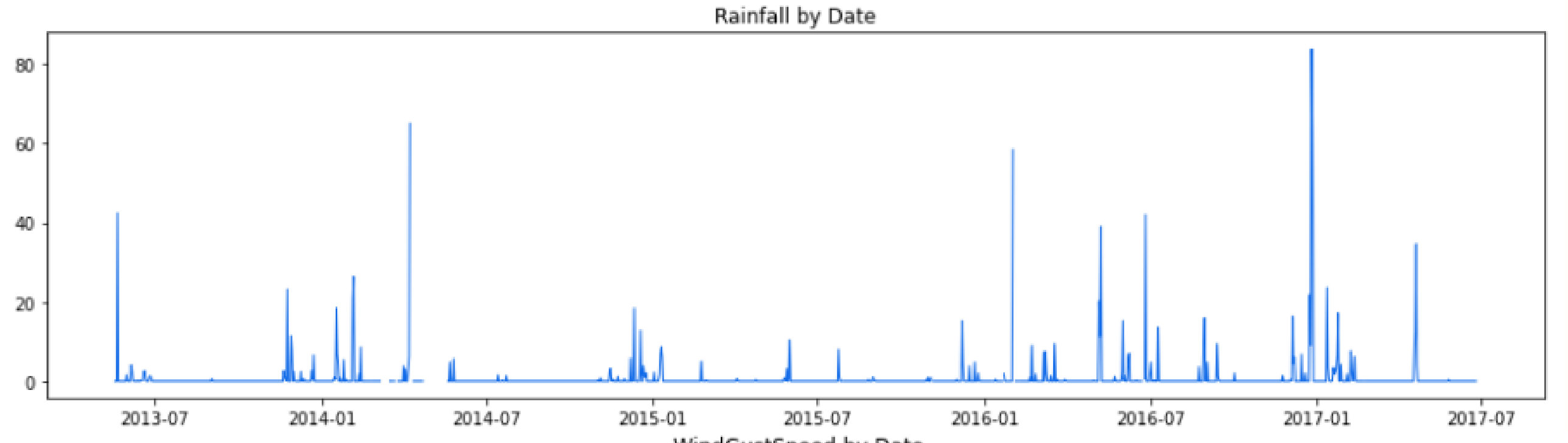
Kesimpulan dari WindSpeed dan RainTomorrow

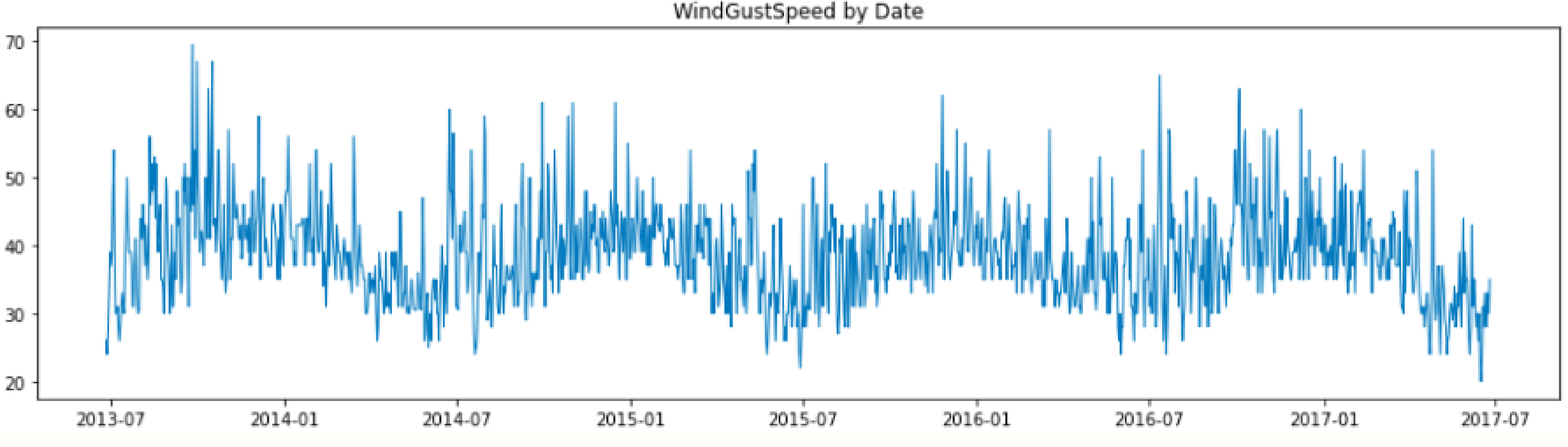
Nilai WindSpeed berpengaruh terhadap terjadinya hujan pada keesokan harinya karena memiliki probabilitas yang moderate jika value dari WindGustSpeed lebih besar dr 80 atau WindSpeed9am lebih dr 60 atau WindSpeed3pm lebih dr 60.

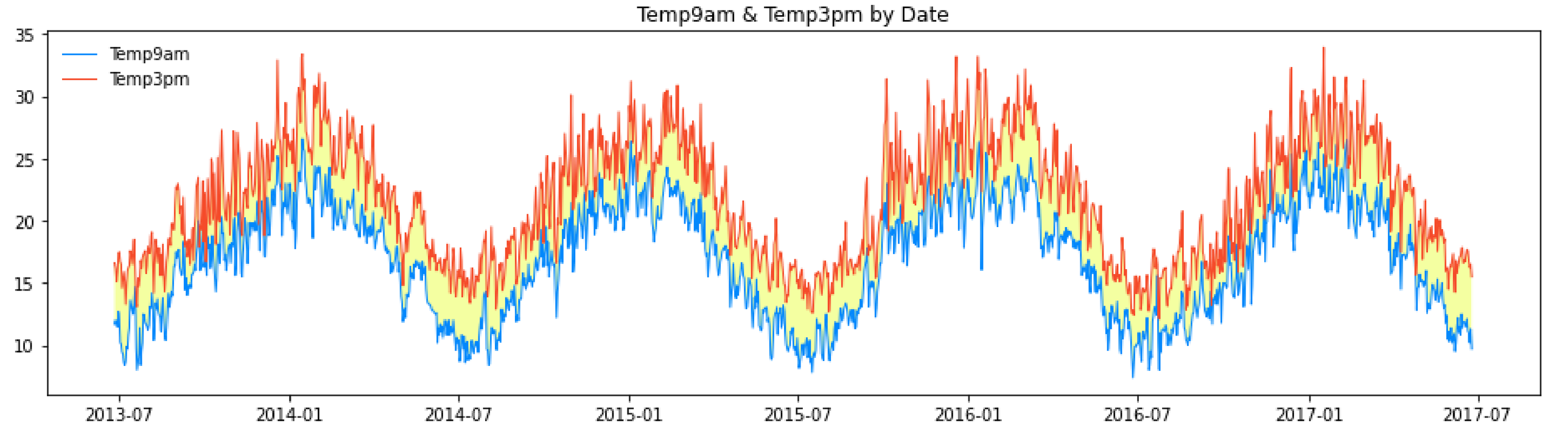
Kami akan membuat fitur baru bernama "HighWind" dengan nilai boolean.

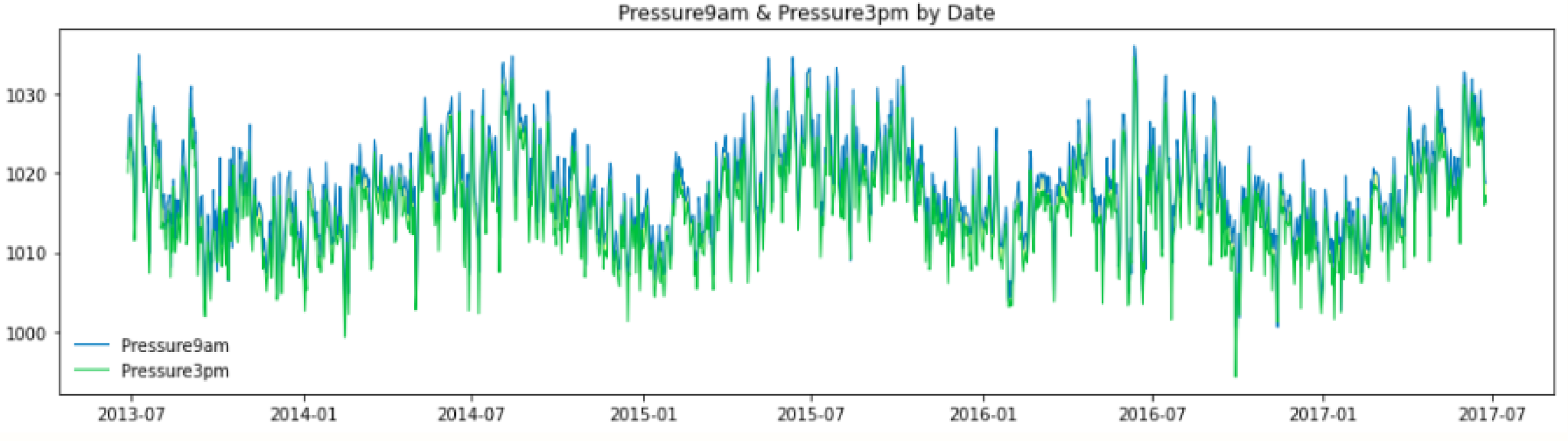
- Bernilai 1 jika value dari WindGustSpeed lebih besar dr 80 atau WindSpeed9am lebih dr 60 atau WindSpeed3pm lebih dr 60.
- Sisanya bernilai 0

Korelasi antara kolom temperature, pressure, rainfall dan wind speed dengan Date.

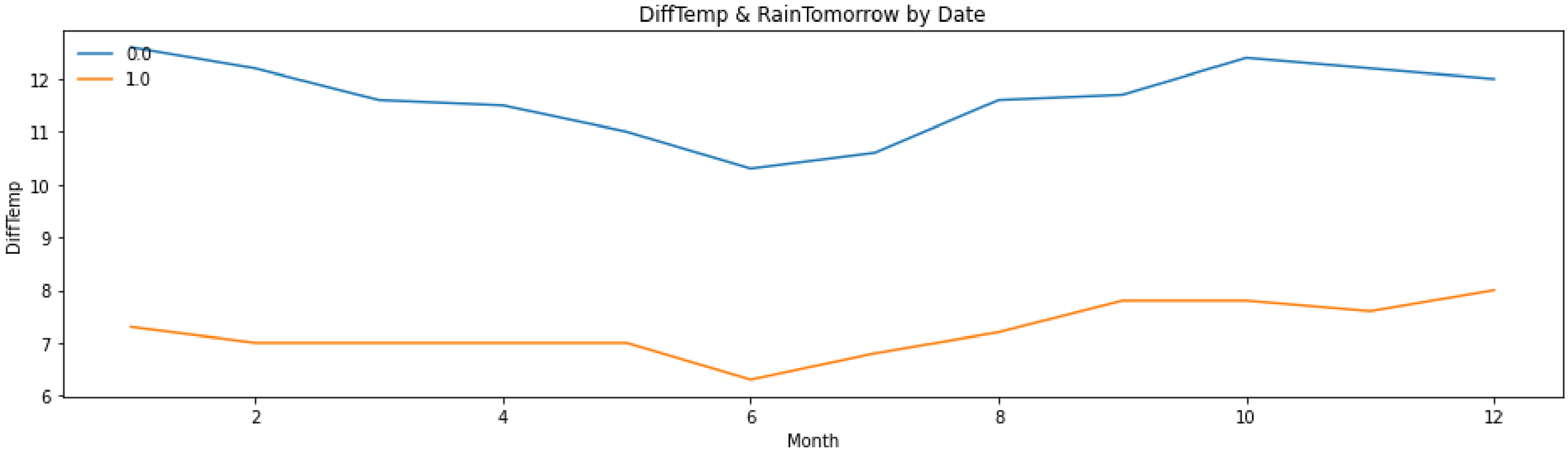

Rainfall by Date
80
60
40
20
0
2013-07
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07


WindGustSpeed by Date
70
60
50
40
30
20
2013-07
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07

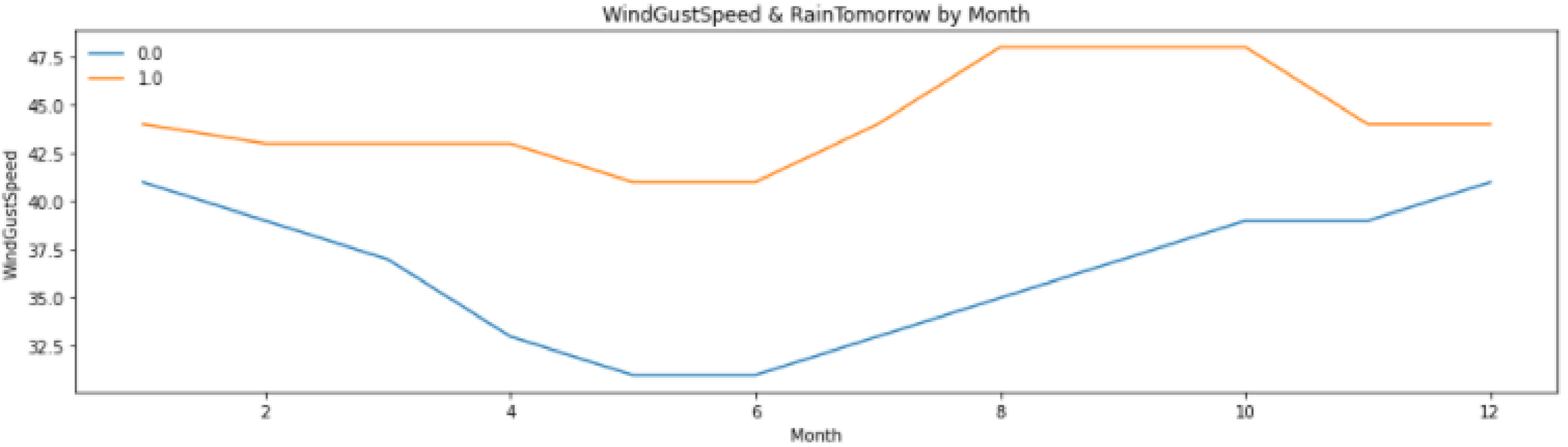

Temp9am & Temp3pm by Date
Temp9am
Temp3pm
35
30
25
20
15
10
2013-07
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07


Pressure9am & Pressure3pm by Date
1030
1020
1010
1000
Pressure9am
Pressure3pm
2013-07
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07

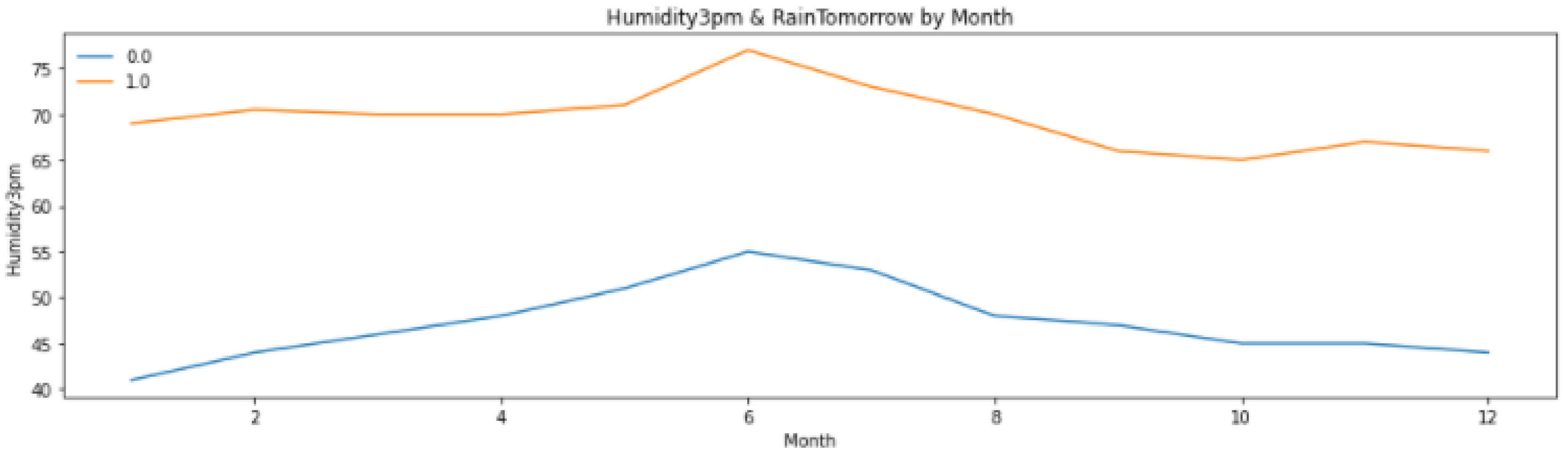
Multivariate Analysis



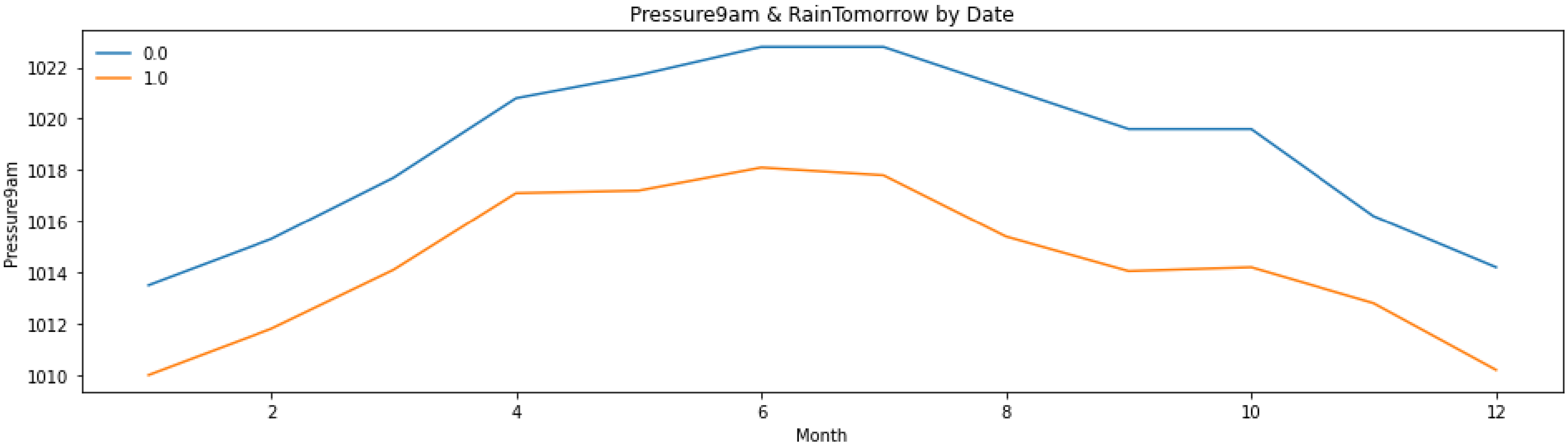
jika Difftemp rendah cenderung hujan dan jika Defftemp tinggi cenderung tidak hujan.



jika WindGustSpeed tinggi cenderung hujan dan jika WindGustSpeed rendah cenderung tidak hujan.



jika Humidity3PM tinggi cenderung hujan dan jika Humidity3PM rendah cenderung tidak hujan.



jika Pressure9am rendah cenderung hujan dan jika Pressire9am tinggi cenderung tidak hujan.

Middle Divider Dataframe

	Month	middle_DiffTemp	middle_WindGustSpeed	middle_Humidity3pm	middle_Pressure9am
0	1	9.95	42.5	55.00	1011.75
1	2	9.60	41.0	57.25	1013.55
2	3	9.30	40.0	58.00	1015.90
3	4	9.25	38.0	59.00	1018.95
4	5	9.00	36.0	61.00	1019.45

Pre-Processing

1 Dropping Duplicate

```
Shape dari dataset sebelum drop duplicates : (145460, 24)
Shape dari dataset setelah drop duplicates : (145460, 24)
```

Tidak ada perubahan shape dari dataset sehingga dapat disimpulkan bahwa tidak ada data yang duplikat

2 Dropping High Missing Values Column

```
Showing the percentage of the NaN value :
Date                0.000000
Location            0.000000
MinTemp             1.020899
MaxTemp             0.866905
Rainfall            2.241853
Evaporation        43.166506
Sunshine           48.009762
WindGustDir         7.098859
WindGustSpeed       7.055548
WindDir9am          7.263853
WindDir3pm          2.906641
WindSpeed9am        1.214767
WindSpeed3pm        2.105046
Humidity9am         1.824557
Humidity3pm         3.098446
Pressure9am        10.356799
Pressure3pm        10.331363
Cloud9am           38.421559
Cloud3pm           40.807095
Temp9am             1.214767
Temp3pm             2.481094
RainToday           2.241853
RainTomorrow        2.245978
dtype: float64
```

```
df.drop(['Evaporation',
         'Sunshine',
         'Cloud9am',
         'Cloud3pm'], axis=1, inplace=True)
```

Kita mendrop kolom yang lebih dari 38% datanya hilang

3 Filling the Missing Values of Each Column

MinTemp, MaxTemp, WindGustSpeed, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Temp9am, and Temp3pm memiliki std yang relatif kecil dibandingkan dengan rata-ratanya. Karena itu, mereka akan diisi dengan rata-rata setiap kolom antara 7 hari sebelum dan 7 hari setelah dari tanggal nilai hilang NaN dengan masing-masing lokasinya.

WindSpeed9am, WindSpeed3pm and Rainfall memiliki std yang relatif besar dibandingkan dengan rata-ratanya. Oleh karena itu, mereka akan diisi dengan median masing-masing kolom antara 7 hari sebelum dan 7 hari setelah dari tanggal nilai hilang NaN dengan masing-masing lokasinya.

Nilai kosong dari kolom WindGustDir, WindDir9am dan WindDir3pm akan diisi dengan moudus setiap kolom antara 8 hari sebelum dan 7 hari setelah dari tanggal nilai hilang NaN dengan masing-masing lokasinya.

```
Showing the NaN value of the 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
dtype: int64
```

```
Showing the percentage of the NaN value :
Date             0.000000
Location         0.000000
MinTemp          1.020899
MaxTemp          0.866905
Rainfall         2.241853
Evaporation     43.166506
Sunshine        48.009762
WindGustDir      7.098859
WindGustSpeed    7.055548
WindDir9am       7.263853
WindDir3pm       2.906641
WindSpeed9am     1.214767
WindSpeed3pm     2.105046
Humidity9am      1.824557
Humidity3pm      3.098446
Pressure9am     10.356799
Pressure3pm     10.331363
Cloud9am        38.421559
Cloud3pm        40.807095
Temp9am          1.214767
Temp3pm          2.481094
RainToday        2.241853
RainTomorrow     2.245978
dtype: float64
```

```
Showing the NaN value of the data :
Date                0
Location            0
MinTemp             0
MaxTemp             0
Rainfall          650
WindGustDir      8813
WindGustSpeed       0
WindDir9am       6574
WindDir3pm       2965
WindSpeed9am      573
WindSpeed3pm     1944
Humidity9am         0
Humidity3pm         0
Pressure9am         0
Pressure3pm         0
Temp9am             0
Temp3pm             0
RainToday        3261
RainTomorrow     3267
Year                0
Month               0
Day                 0
Year_Month          0
DiffTemp         1881
dtype: int64
```

```
Showing the percentage of the NaN value :
Date             0.000000
Location         0.000000
MinTemp          0.000000
MaxTemp          0.000000
Rainfall         0.446858
WindGustDir      6.058710
WindGustSpeed    0.000000
WindDir9am       4.519456
WindDir3pm       2.038361
WindSpeed9am     0.393923
WindSpeed3pm     1.336450
Humidity9am      0.000000
Humidity3pm      0.000000
Pressure9am      0.000000
Pressure3pm      0.000000
Temp9am          0.000000
Temp3pm          0.000000
RainToday        2.241853
RainTomorrow     2.245978
Year             0.000000
Month            0.000000
Day              0.000000
Year_Month       0.000000
DiffTemp         1.293139
dtype: float64
```

```
Showing the NaN value of the data :
Date                0
Location            0
MinTemp             0
MaxTemp             0
Rainfall            0
WindGustDir         0
WindGustSpeed       0
WindDir9am          0
WindDir3pm          0
WindSpeed9am        0
WindSpeed3pm        0
Humidity9am         0
Humidity3pm         0
Pressure9am         0
Pressure3pm         0
Temp9am             0
Temp3pm             0
RainToday        3261
RainTomorrow     3267
Year                0
Month               0
Day                 0
Year_Month          0
DiffTemp         1881
dtype: int64
```

Setelah menggunakan cara pengisian data yang kosong berdasarkan waktu dan lokasi, ternyata masih ada data yang kosong. Data kosong tersebut tidak terisi karena terdapat data yang semuanya kosong dalam jangka waktu yang ditentukan. Oleh karena itu, kami akan mengisi data yang kosong tersebut menggunakan mean atau median atau modus tiap kolom secara keseluruhan.

Sekarang sudah tidak ada data yang kosong kecuali pada kolom target dan kolom feature engineering. Data yang kosong tersebut akan dibenahi pada tahap selanjutnya

4 Creating New Feature

```
#HighWind
df.loc[((df.WindGustSpeed>=80) | (df.WindSpeed9am>=60) | (df.WindSpeed3pm>=60))
      & (df.RainTomorrow==1), 'HighWind']=1

#HighHumidity
df.loc[(df.Humidity3pm>=80), 'HighHumidity']=1

#Freezing
df.loc[(df.MaxTemp<=0), 'Freezing']=1

#Making a new column called "TempDiff"
df['DiffTemp']=df.MaxTemp-df.MinTemp
```

```
# Middle Divider
df=df.merge(middle_divider, on='Month')

df.loc[df.DiffTemp <= df.middle_DiffTemp, 'Binary_DiffTemp']=1
df.loc[df.WindGustSpeed >= df.middle_WindGustSpeed, 'Binary_WindGustSpeed']=1
df.loc[df.DiffTemp >= df.middle_Humidity3pm, 'Binary_Humidity3pm']=1
df.loc[df.DiffTemp <= df.middle_Pressure9am, 'Binary_Pressure9am']=1
```

5 Dropping Unimportant Columns

```
Date                       0
Location                   0
MinTemp                    0
MaxTemp                    0
Rainfall                   0
WindGustDir                0
WindGustSpeed              0
WindDir9am                 0
WindDir3pm                 0
WindSpeed9am               0
WindSpeed3pm               0
Humidity9am                0
Humidity3pm                0
Pressure9am                0
Pressure3pm                0
Temp9am                    0
Temp3pm                    0
RainToday               3261
RainTomorrow            3267
Year                       0
Month                      0
Day                        0
Year_Month                 0
DiffTemp                   0
HighWind                   0
HighHumidity               0
Freezing                   0
middle_DiffTemp            0
middle_WindGustSpeed       0
middle_Humidity3pm         0
middle_Pressure9am         0
Binary_DiffTemp            0
Binary_WindGustSpeed       0
Binary_Humidity3pm         0
Binary_Pressure9am         0
dtype: int64
```

```
Location                   0
MinTemp                    0
MaxTemp                    0
Rainfall                   0
WindGustDir                0
WindGustSpeed              0
WindDir9am                 0
WindDir3pm                 0
WindSpeed9am               0
WindSpeed3pm               0
Humidity9am                0
Humidity3pm                0
Pressure9am                0
Pressure3pm                0
Temp9am                    0
Temp3pm                    0
RainToday               3261
RainTomorrow            3267
Month                      0
DiffTemp                   0
HighWind                   0
HighHumidity               0
Freezing                   0
Binary_DiffTemp            0
Binary_WindGustSpeed       0
Binary_Humidity3pm         0
Binary_Pressure9am         0
dtype: int64
```

6 Dropping Missing Values target

Location	0
MinTemp	0
MaxTemp	0
Rainfall	0
WindGustDir	0
WindGustSpeed	0
WindDir9am	0
WindDir3pm	0
WindSpeed9am	0
WindSpeed3pm	0
Humidity9am	0
Humidity3pm	0
Pressure9am	0
Pressure3pm	0
Temp9am	0
Temp3pm	0
RainToday	3261
RainTomorrow	3267
Month	0
DiffTemp	0
HighWind	0
HighHumidity	0
Freezing	0
Binary_DiffTemp	0
Binary_WindGustSpeed	0
Binary_Humidity3pm	0
Binary_Pressure9am	0

dtype: int64

Location	0
MinTemp	0
MaxTemp	0
Rainfall	0
WindGustDir	0
WindGustSpeed	0
WindDir9am	0
WindDir3pm	0
WindSpeed9am	0
WindSpeed3pm	0
Humidity9am	0
Humidity3pm	0
Pressure9am	0
Pressure3pm	0
Temp9am	0
Temp3pm	0
RainToday	0
RainTomorrow	0
Month	0
DiffTemp	0
HighWind	0
HighHumidity	0
Freezing	0
Binary_DiffTemp	0
Binary_WindGustSpeed	0
Binary_Humidity3pm	0
Binary_Pressure9am	0

dtype: int64

7 Label Encoding Categorical Values

```
from sklearn.preprocessing import LabelEncoder

labelencoder = LabelEncoder()
df.loc[:,'Location']=labelencoder.fit_transform(df['Location'])
df.loc[:,'WindGustDir']=labelencoder.fit_transform(df['WindGustDir'])
df.loc[:,'WindDir9am']=labelencoder.fit_transform(df['WindDir9am'])
df.loc[:,'WindDir3pm']=labelencoder.fit_transform(df['WindDir3pm'])
```

8 **Feature Scaling (Standardization)**

9 **Splitting into train and test dataset**

10 **Upsampling minority target on training dataset**

Processed Dataset

	Location	MinTemp	MaxTemp	Rainfall	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Temp9am
0	-1.525999	0.189378	-0.047299	-0.206727	1.063165	0.304544	1.305710	1.365499	0.674848	0.611094	0.112775	-1.429106	-1.473894	-1.220076	-0.010230
1	-1.525999	-0.747978	0.261996	-0.277606	1.276185	0.304544	-0.243158	1.584742	-1.127268	0.382348	-1.306625	-1.283605	-1.044462	-1.115388	0.035872
2	-1.525999	0.111265	0.346349	-0.277606	1.489206	0.456556	1.305710	1.584742	0.562216	0.839840	-1.622047	-1.041104	-1.488702	-0.980790	0.619831
3	-1.525999	-0.466771	0.669703	-0.277606	-0.854023	-1.215580	0.420642	-1.703906	-0.338842	-1.104499	-1.254055	-1.720109	-0.007901	-0.367621	0.174178
4	-1.525999	0.829905	1.274233	-0.159474	1.063165	0.076525	-1.349493	-0.169203	-0.789371	0.153602	0.691049	-0.895603	-1.014846	-1.384584	0.128076

Temp3pm	RainToday	Month	DiffTemp	HighWind	HighHumidity	Freezing	Binary_DiffTemp	Binary_WindGustSpeed	Binary_Humidity3pm	Binary_Pressure9am	RainTomorrow
0.017678	-0.536378	1.633271	-0.312712	-0.082816	-0.321879	-0.02809	1.167421	1.138936	0.0	0.0	0.0
0.380037	-0.536378	1.633271	1.343086	-0.082816	-0.321879	-0.02809	-0.856589	1.138936	0.0	0.0	0.0
0.220599	-0.536378	1.633271	0.353646	-0.082816	-0.321879	-0.02809	-0.856589	1.138936	0.0	0.0	0.0
0.698913	-0.536378	1.633271	1.565206	-0.082816	-0.321879	-0.02809	-0.856589	-0.878012	0.0	0.0	0.0
1.162733	-0.536378	1.633271	0.757499	-0.082816	-0.321879	-0.02809	-0.856589	-0.878012	0.0	0.0	0.0

Machine Learning

1. Splitting into x_train, y_train, x_test, y_test
2. Making Function
3. Training Model (Decision Tree, Random Forest, XGBoost, LGBM Classifier)
4. Model Selection
5. Feature Selection

1 Splitting into x_train, y_train, x_test, y_test

train/test split membagi dataset menjadi train set dan test set, atau dengan kata lain, data yang digunakan untuk proses training dan testing merupakan kumpulan data yang berbeda.

train/test split ini akan memberikan hasil prediksi yang lebih akurat untuk new data atau data yang belum pernah di-train.

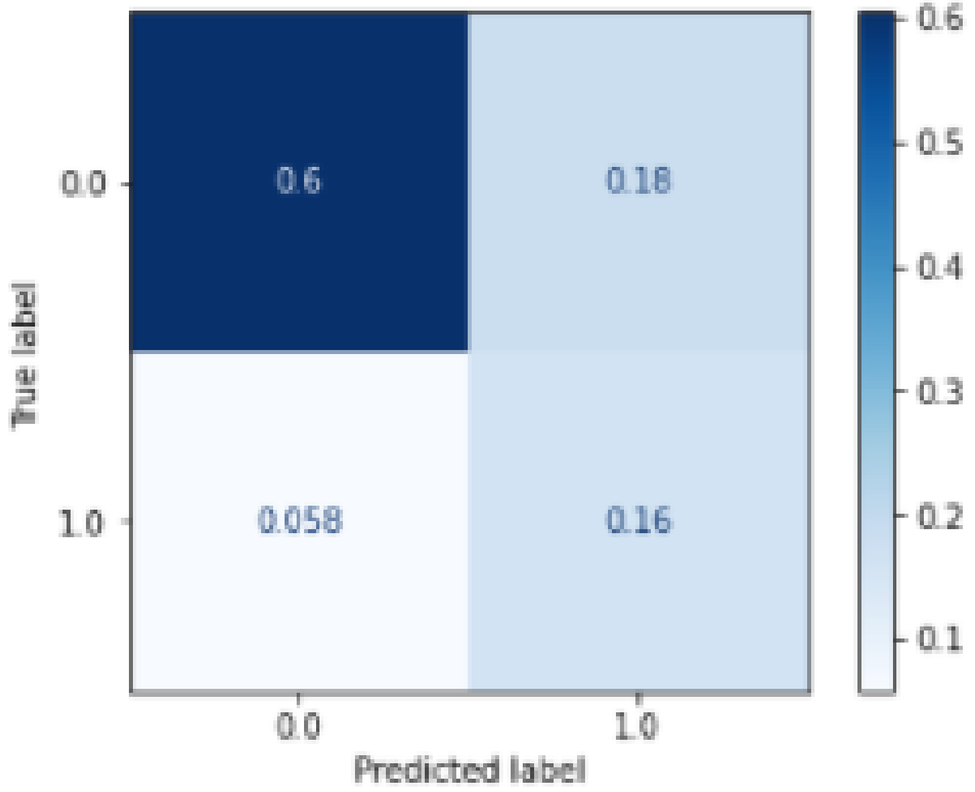
2 Making Function

suatu proses untuk mempersingkat agar tidak membuat ulang metriks.

3 Model Traning

a. Decision Tree

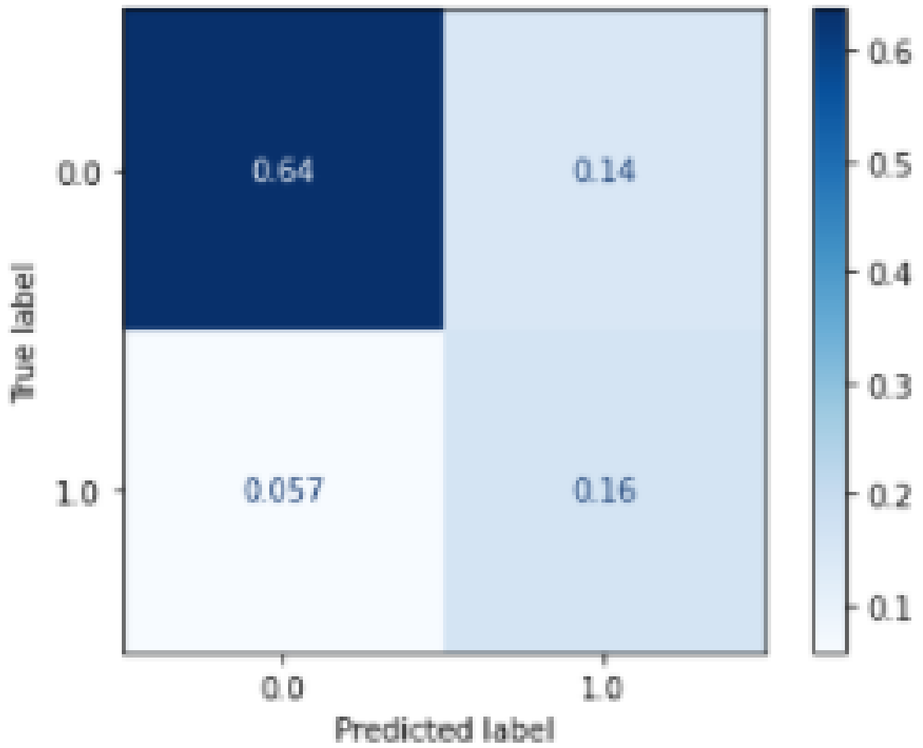
```
Accuracy test = 0.7639036863413595
Accuracy train = 0.7791919330341338
ROC AUC test = 0.7532951761270075
ROC AUC train = 0.7791919330341338
Cohen's Kappa test = 0.4225081997503618
Time taken = 0.6345860958099365
              precision    recall  f1-score   support

         0.0    0.91251   0.77211   0.83646     22019
         1.0    0.47329   0.73448   0.57564      6139

    accuracy                        0.76390     28158
   macro avg    0.69290   0.75330   0.70605     28158
weighted avg    0.81675   0.76390   0.77959     28158
```



b. Random Forest

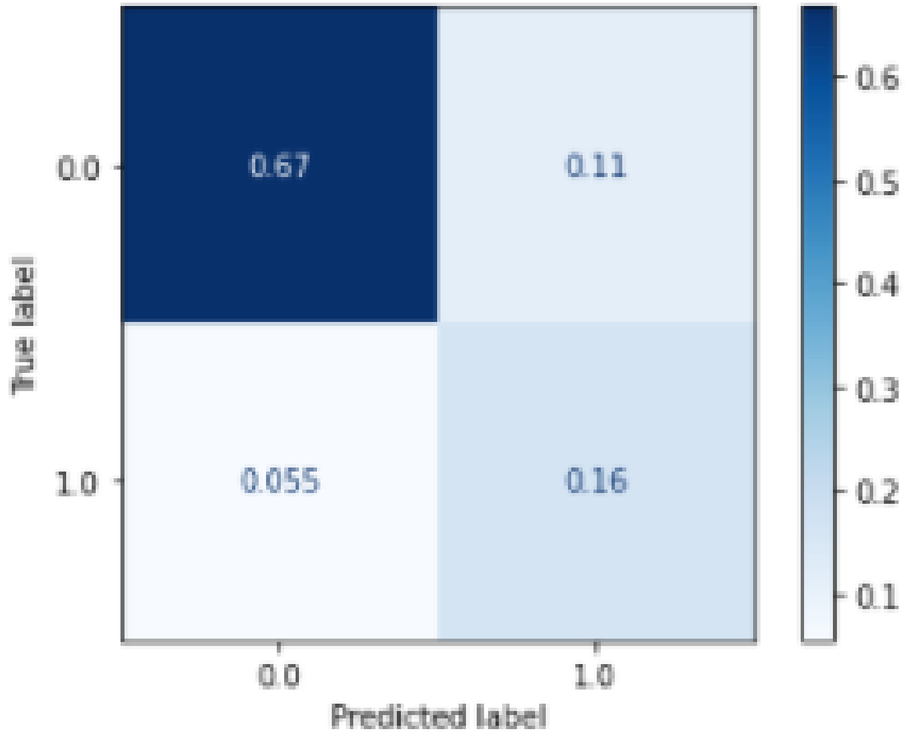
```
Accuracy test = 0.7981035584913702
Accuracy train = 0.7901606769673507
ROC AUC test = 0.7765723860883048
ROC AUC train = 0.7901606769673507
Cohen's Kappa test = 0.4829857240535238
Time taken = 24.85660076141357
```

	precision	recall	f1-score	support
0.0	0.91783	0.81475	0.86323	22019
1.0	0.52636	0.73839	0.61460	6139
accuracy			0.79810	28158
macro avg	0.72210	0.77657	0.73891	28158
weighted avg	0.83249	0.79810	0.80902	28158



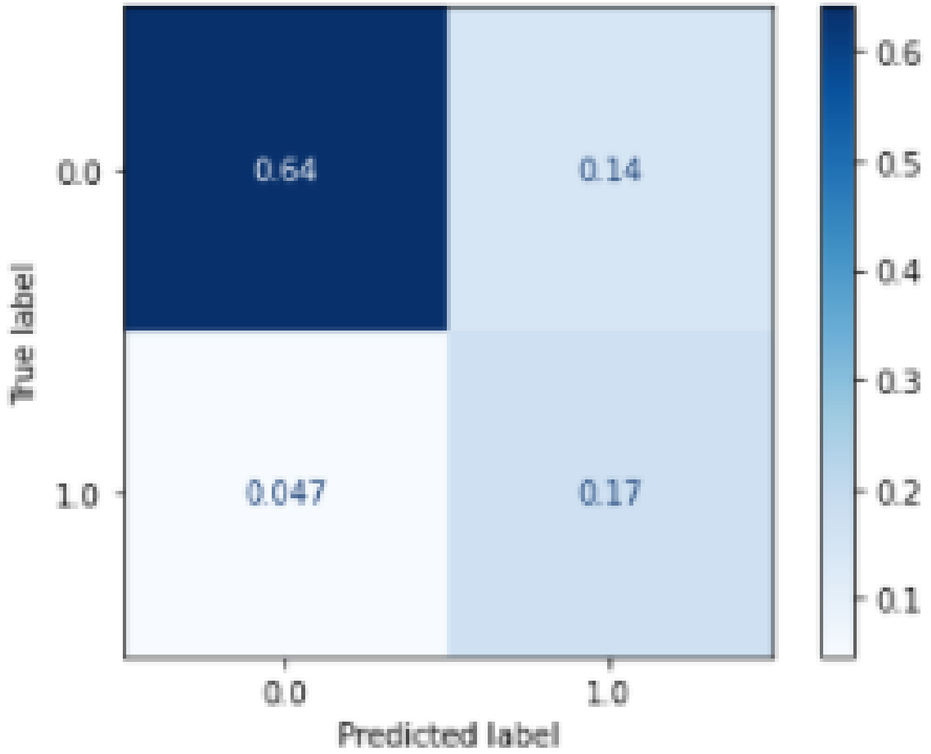
c. XGBoost

```
Accuracy test = 0.830314653029345
Accuracy train = 0.8933388148503432
ROC AUC test = 0.800516348328973 9
ROC AUC train = 0.8933388148503432
Cohen's Kappa test = 0.5470492073530036
Time taken = 97.88796186447144
              precision    recall  f1-score   support

         0.0    0.92384   0.85335   0.88720     22019
         1.0    0.58703   0.74768   0.65769      6139

    accuracy                        0.83031     28158
   macro avg    0.75544   0.80052   0.77244     28158
weighted avg    0.85041   0.83031   0.83716     28158
```



d. LGBM Classifier

```
Accuracy test = 0.8117053768023297
Accuracy train = 0.8191556179839438
ROC AUC test = 0.8015988146770813
ROC AUC train = 0.8191556179839437
Cohen's Kappa test = 0.5220578620401934
Time taken = 6.613329172134399
              precision    recall  f1-score   support

         0.0    0.93145   0.81952   0.87191     22019
         1.0    0.54764   0.78368   0.64473      6139

    accuracy                        0.81171     28158
   macro avg    0.73954   0.80160   0.75832     28158
weighted avg    0.84777   0.81171   0.82238     28158
```



4 Model Selection

Sebelum hyperparameter tuning

Model Name	Accuracy(test)	Accuracy(train)	ROC AUC(test)	ROC AUC(train)	Precision	Recall	f1	False Positive	False Negative
Decision Tree	0.7870	1.0000	0.6875	1.0000	0.5115	0.5110	0.5112	0.1114	0.1066
Random Forest	0.8539	1.0000	0.7489	1.0000	0.7075	0.5626	0.6268	0.1227	0.0954
XGBoost	0.7945	0.7894	0.7827	0.7894	0.5196	0.7619	0.6178	0.1661	0.0519
LGBM	0.8113	0.8190	0.8006	0.8190	0.5470	0.7816	0.6436	0.1704	0.0476

Setelah hyperparameter tuning

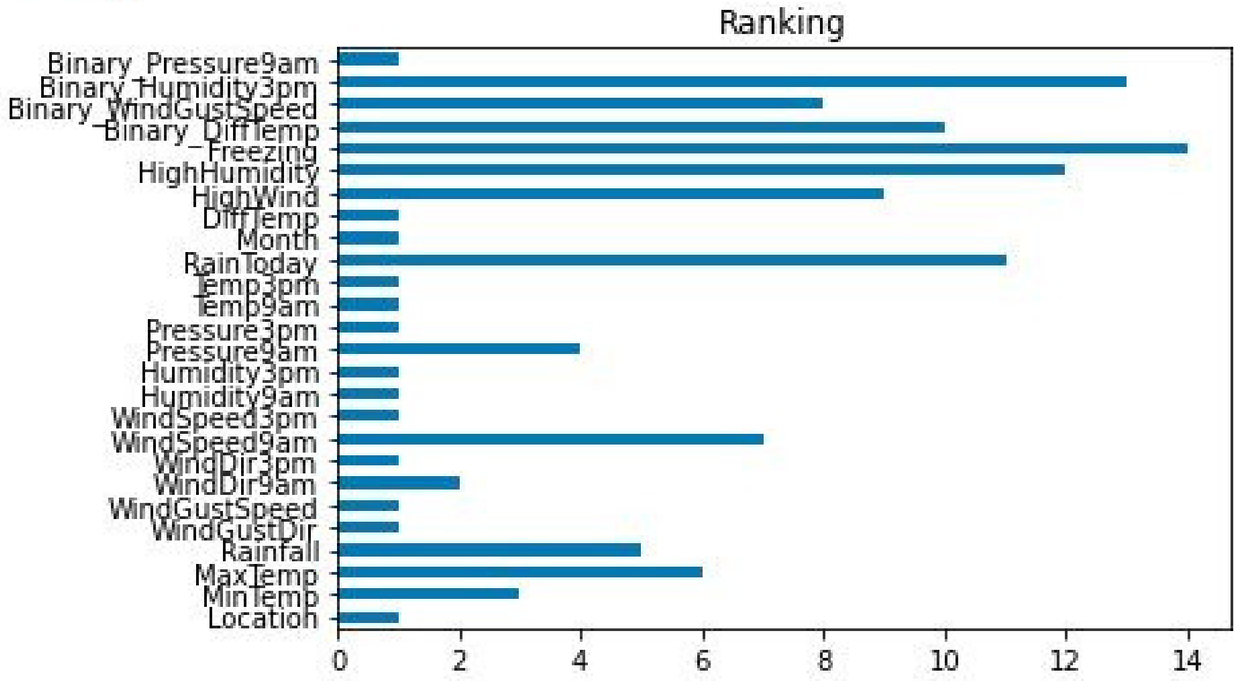
Model Name	Accuracy(test)	Accuracy(train)	ROC AUC(test)	ROC AUC(train)	Precision	Recall	f1	False Positive	False Negative
Decision Tree	0.7805	0.7883	0.7545	0.7883	0.4976	0.7084	0.5846	0.1544	0.0636
Random Forest	0.7981	0.7902	0.7766	0.7902	0.5264	0.7384	0.6146	0.1610	0.0570
XGBoost	0.8303	0.8933	0.8005	0.8933	0.5870	0.7477	0.6577	0.1630	0.0550
LGBM	0.8117	0.8192	0.8016	0.8192	0.5476	0.7837	0.6447	0.1709	0.0472

Dari hasil di atas dapat terlihat bahwa model XGBoost dan LGBM 2 model yang paling unggul serta memiliki score yang hampir sama. Terdapat overfit pada model XGBoost serta XGBoost memiliki nilai recall serta ROC AUC yang lebih rendah daripada LGBM. Selain itu, LGBM memiliki false negative yang lebih rendah dan false positive yang lebih tinggi. Oleh karena itu, kami akan memilih model LGBM.

5 Feature Selection

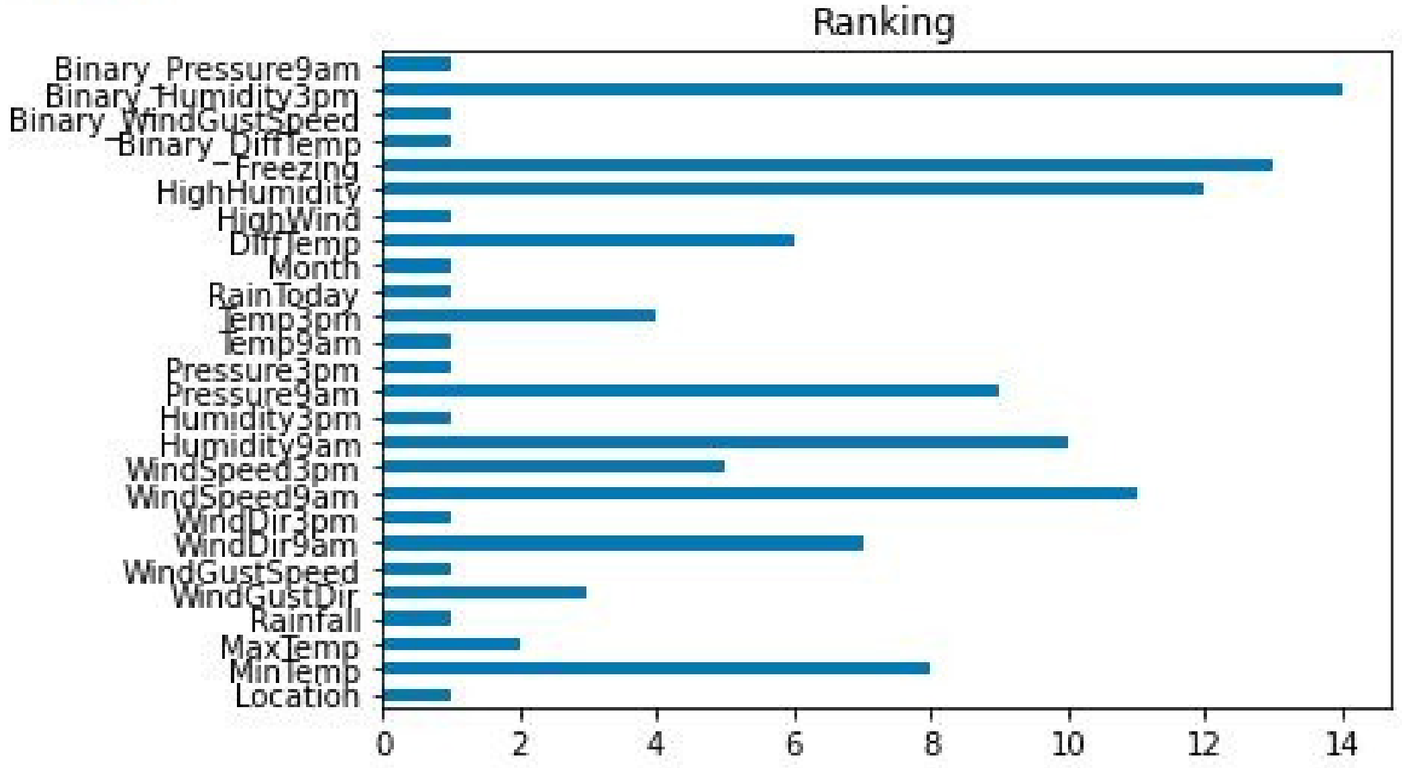
a. LGBM Classifier

```
Sebelum feature selection  (175134, 26)
Setelah feature selection  (175134, 13)
Score of features  [ 1  3  6  5  1  1  2  1  7  1  1  1  4  1  1  1 11  1  1  9 12 14 10  8
 13  1]
```



b. XGBoost

```
Sebelum feature selection  (175134, 26)
Setelah feature selection  (175134, 13)
Score of features  [ 1  8  2  1  3  1  7  1 11  5 10  1  9  1  1  4  1  1  6  1 12 13  1  1
 14  1]
```



Result

	Model Name	Accuracy(test)	Accuracy(train)	ROC AUC(test)	ROC AUC(train)	Precision	Recall	f1	False Positive	False Negative
0	XGBoost	0.8320	0.8987	0.7987	0.8987	0.5917	0.7397	0.6575	0.1613	0.0568
0	LGBM	0.8117	0.8192	0.8016	0.8192	0.5476	0.7837	0.6447	0.1709	0.0472
0	LGBM Optimized	0.8035	0.8150	0.7941	0.8150	0.5340	0.7775	0.6331	0.1695	0.0485
0	XGBoost Optimized	0.8214	0.8735	0.7916	0.8735	0.5697	0.7387	0.6433	0.1611	0.0570

Thank You