RISIKO CUACA PADA EVENT OLAHRAGA

Presented by :
FAHRIAN MUHAMMAD NABIEL

TABLE OF CONTENT

HI, I'M FAHRIAN

Seorang calon Data Scientist yang bersemangat untuk menghubungkan ilmu yang dipelajari dengan penerapan di dunia nyata. Terbiasa mengubah tantangan bisnis menjadi solusi yang bisa digunakan serta menghasilkan dampak nyata. Berkomitmen untuk terus mengembangkan kemampuan teknis dan membangun karier yang kuat di bidang data science.

Saat ini sedang mengikuti program Full Stack Data Science Bootcamp selama 6 bulan di dibimbing.id dengan fokus pada analisis data, machine learning, dan proyek end-to-end. Termotivasi untuk terus belajar, mengasah keterampilan, dan memberikan kontribusi positif dalam bidang data science.

PENDIDIKAN

Universitas Bina Nusantara

Teknik Informatika

2019 - 2023

- Sarjana Teknik Informatika, IPK 3.08/4.00

Dibimbing.id

Full-Stack Data Science Bootcamp

6 Bulan

- Pengetahuan dalam statistik, analisis data, machine learning, dan visualisasi data
- Kemampuan teknis: Python, SQL, Tableau, Google Data Studio, Streamlit
- Teknik: EDA, Hypothesis Testing, Regression (Linear, Ridge, Lasso, Logistic), KNN, Decision Tree, Random Forest, Gradient Boosted Trees, Factor Analysis, K-Means

BACKGROUND PROJECT

- Event Main Bareng (MaBar) olahraga diminati masyarakat → sarana kesehatan & relasi
- Masalah utama: curah hujan tinggi & iklim tidak menentu
- Hujan sering membuat peserta batal hadir mendadak
- Dampak: kerugian penyelenggara karena lapangan/event yang dipesan tidak bisa dipakai & biaya sulit dikembalikan

PROJECT GOALS

Aplikasi XYZ bertujuan untuk:

- Mempermudah penyelenggara event MaBar
- Mengurangi hambatan akibat cuaca
- Menarik lebih banyak peserta
- Menjamin event berjalan lancar

DATA UNDERSTANDING

Dataset Cuaca (2007–2017)

- Total: 145.460 entri, 23 variabel

Kategori	Variabel
Idetitas Data	Date, Location
Suhu	MinTemp, MaxTemp, Temp9am, Temp3pm
Hujan & Kelembapan	Rainfall, Humidity9am, Humidity3pm, RainToday, RainTomo
Angin	WindGustDir, WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm
Tekanan Udara	Pressure9am, Pressure3pm
Awan & Matahari	Cloud9am, Cloud3pm, Sunshine, Evaporation

Kegunaan:

- Analisis pola cuaca
- Prediksi hujan lebih akurat
- Mendukung aplikasi XYZ → event lebih efektif

DATA PREPROCESSING

```
df.isna().sum()
```

	0
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

Dataset memiliki cukup banyak **missing values**. Untuk menjaga **konsistensi** dan **kualitas data**, **entri** dengan **missing values** diputuskan untuk dihapus.

```
[59] df['Date'] = pd.to_datetime(df['Date'])
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 56420 entries, 6049 to 142302
Data columns (total 23 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Date           56420 non-null  datetime64[ns]
 1   Location       56420 non-null  object
 2   MinTemp        56420 non-null  float64
 3   MaxTemp        56420 non-null  float64
 4   Rainfall       56420 non-null  float64
 5   Evaporation    56420 non-null  float64
 6   Sunshine       56420 non-null  float64
 7   WindGustDir    56420 non-null  object
 8   WindGustSpeed  56420 non-null  float64
 9   WindDir9am     56420 non-null  object
 10  WindDir3pm     56420 non-null  object
 11  WindSpeed9am   56420 non-null  float64
 12  WindSpeed3pm   56420 non-null  float64
 13  Humidity9am    56420 non-null  float64
 14  Humidity3pm    56420 non-null  float64
 15  Pressure9am    56420 non-null  float64
 16  Pressure3pm    56420 non-null  float64
 17  Cloud9am       56420 non-null  float64
 18  Cloud3pm       56420 non-null  float64
 19  Temp9am        56420 non-null  float64
 20  Temp3pm        56420 non-null  float64
 21  RainToday      56420 non-null  object
 22  RainTomorrow   56420 non-null  object
dtypes: datetime64[ns](1), float64(16), object(6)
memory usage: 10.3+ MB
```

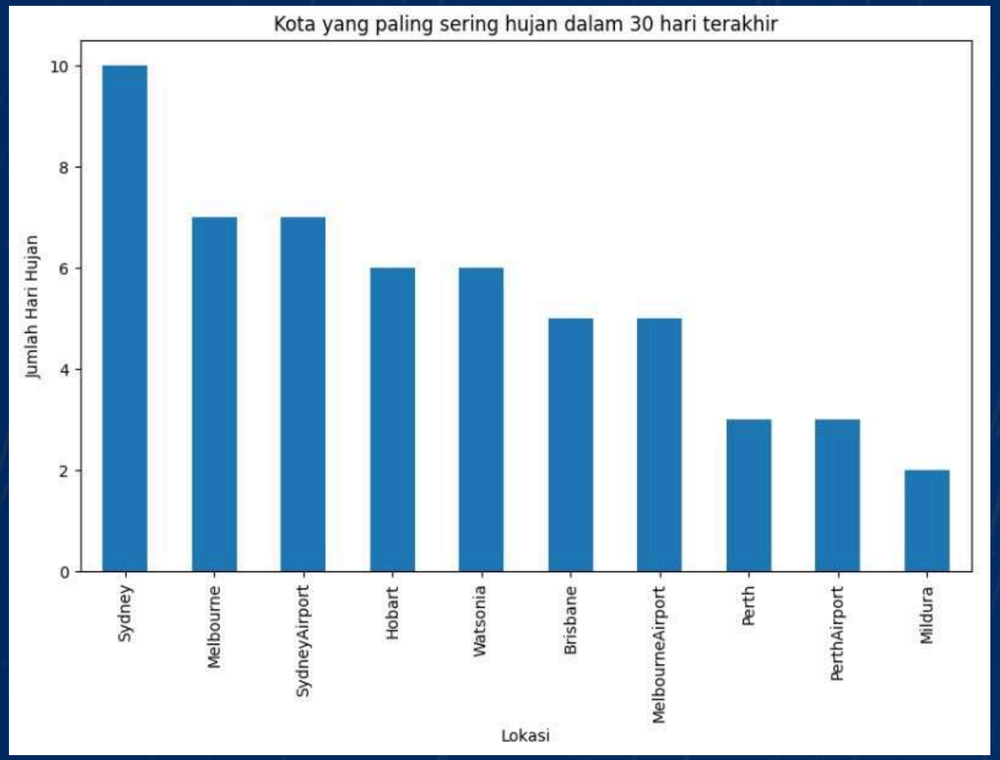
Set Index

```
[61] df = df.set_index('Date')
```

Kolom date diubah ke dalam format **datetime** dan dijadikan sebagai **indeks** agar data lebih terstruktur serta dapat digunakan untuk **keperluan peramalan (forecasting).**

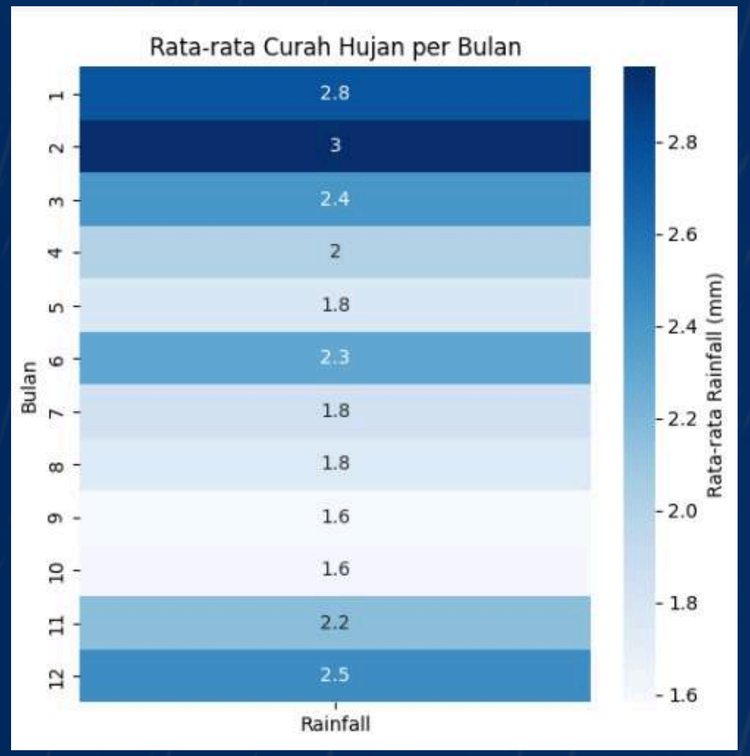
EXPLORATORY DATA ANALYSIS

30 HARI TERAKHIR : KOTA PALING HUJAN



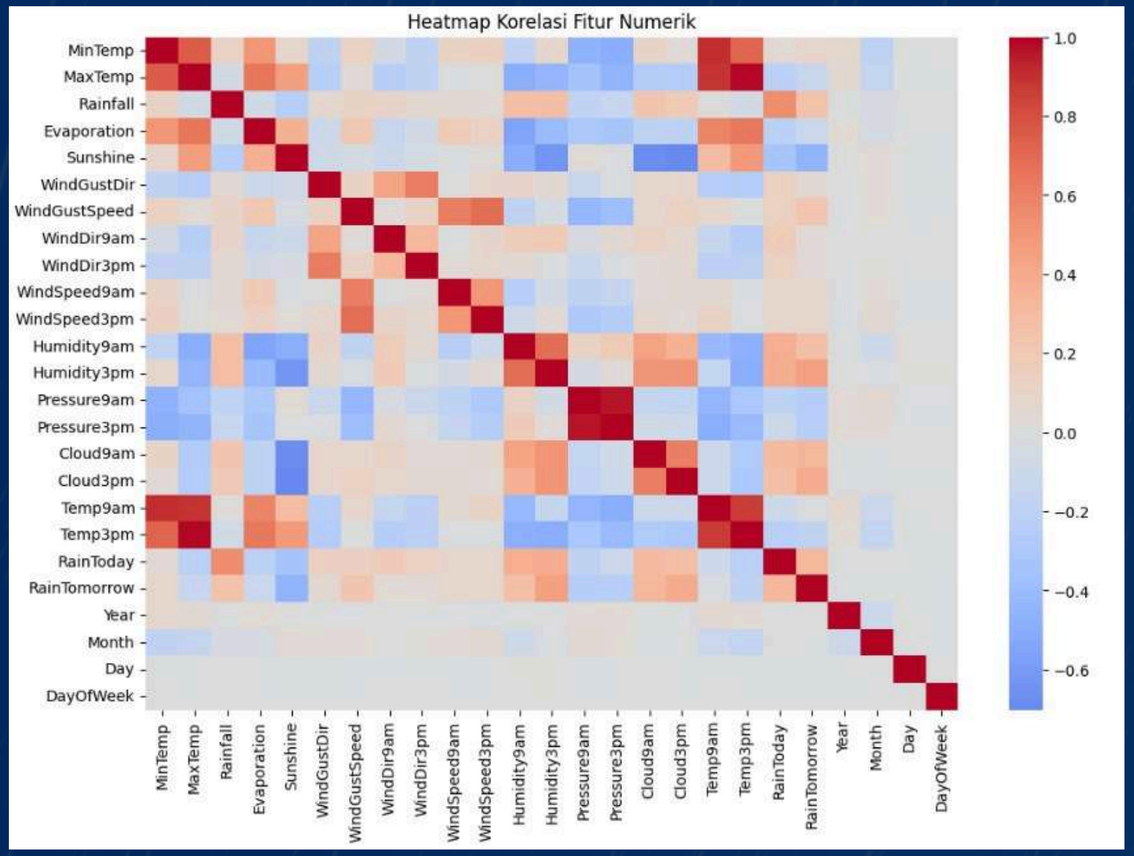
Hujan yang paling banyak selama 30 hari terakhir itu **Sydney** dengan total hujan 10 hari dan disusul oleh **Melbourne** dan **SydneyAirport.**

EXPLORATORY DATA ANALYSIS



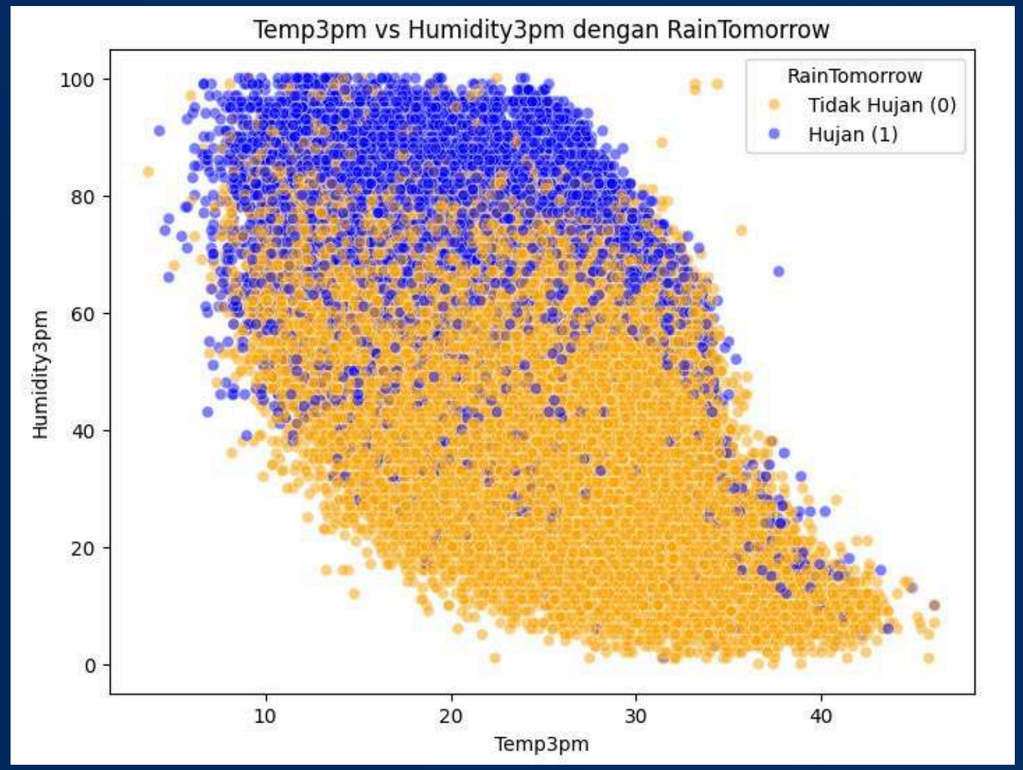
Terlihat bahwa curah hujan **cenderung meningkat** pada akhir tahun dan awal tahun. Hal ini juga diperkuat oleh tabel yang menunjukkan bahwa periode tersebut memang **memiliki intensitas curah hujan yang lebih tinggi dibandingkan bulan lainnya.**

KORELASI ANTAR KOLOM



Terlihat bahwa variabel **Sunshine dan Temp3pm** dengan **RainToday** memiliki warna biru cukup gelap yang menunjukkan adanya korelasi **negatif** dengan R**ainTomorrow**. Sebaliknya, variabel **Humidity3pm** menunjukkan korelasi positif dengan **RainTomorrow**. Hal ini mengindikasikan bahwa **semakin tinggi kelembapan pada sore hari, semakin besar kemungkinan terjadinya hujan keesokan harinya.**

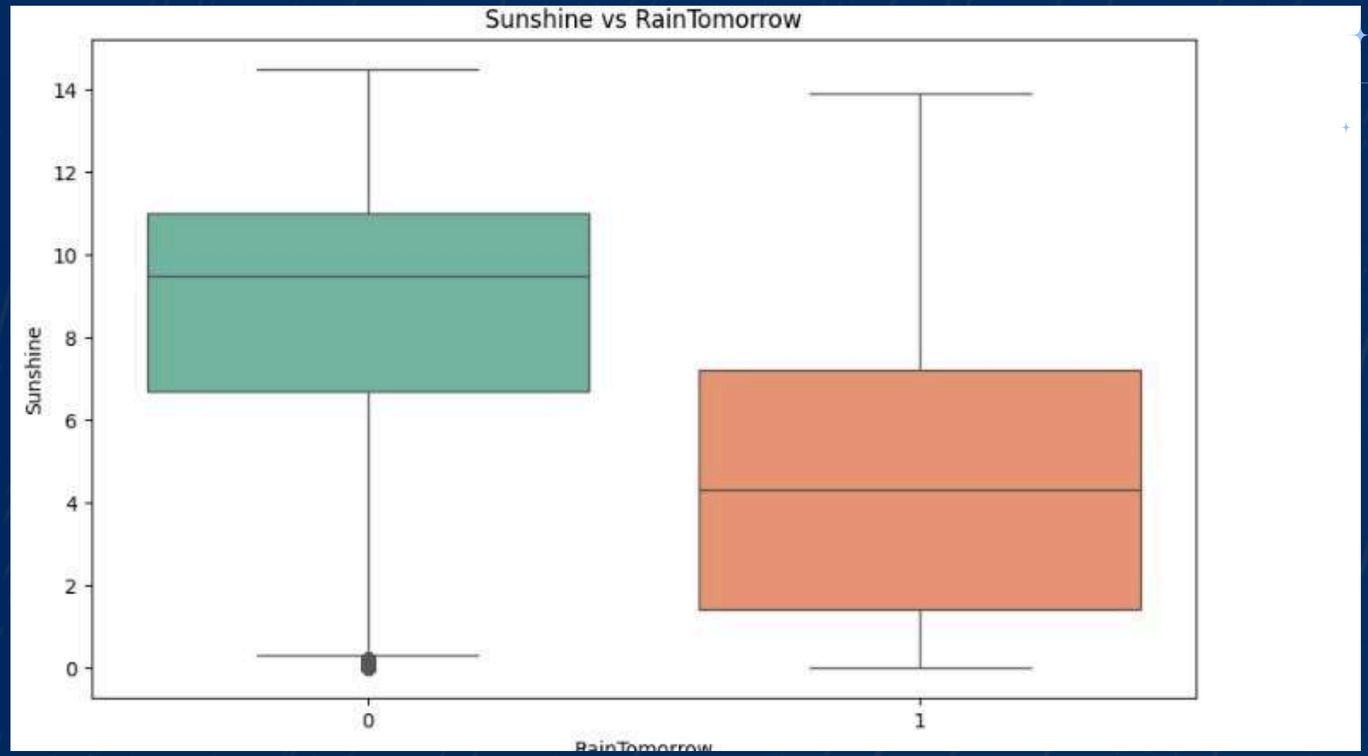
DISTRIBUSI HUJAN BERDASARKAN KELEMBAPAN DAN TEMPERATUR



Persebaran RainTomorrow menunjukkan **bahwa semakin tinggi Humidity3pm dan semakin rendah Temp3pm, maka peluang terjadinya hujan keesokan harinya semakin besar.** Kondisi udara lembap dengan suhu relatif dingin pada sore hari menjadi indikator kuat terbentuknya hujan esok hari.

SUNSHINE VS RAIN TOMORROW

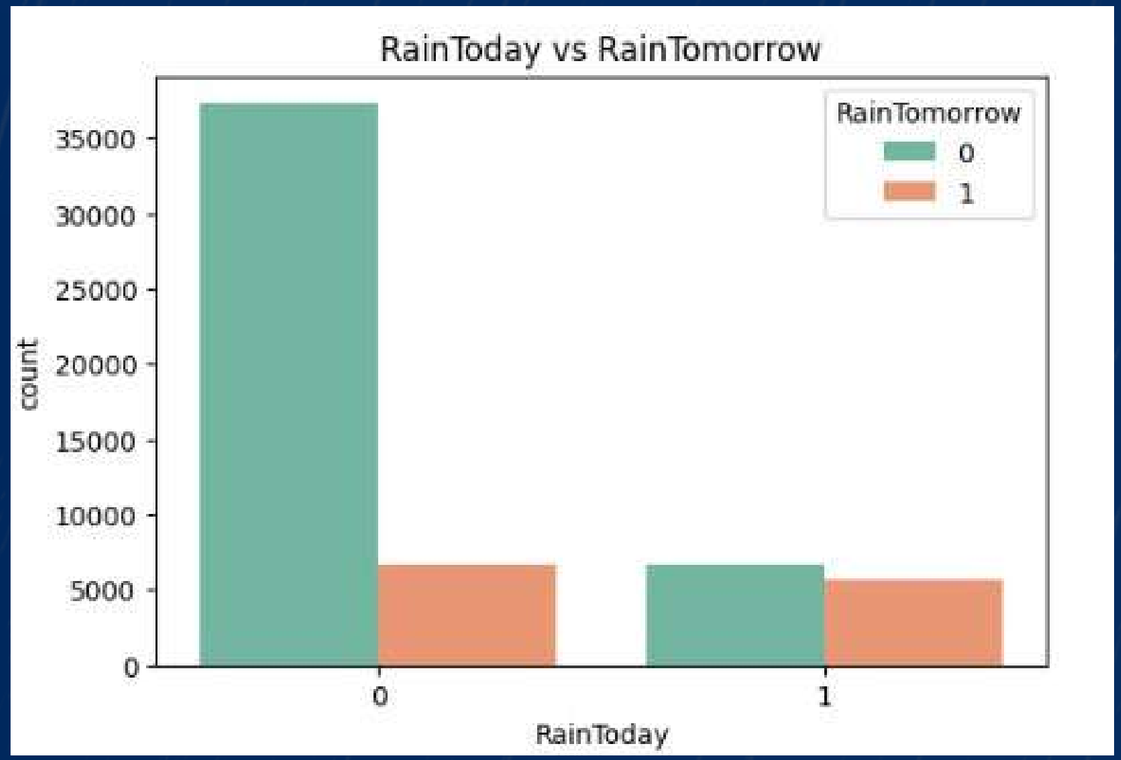
YES : 1
NO : 0



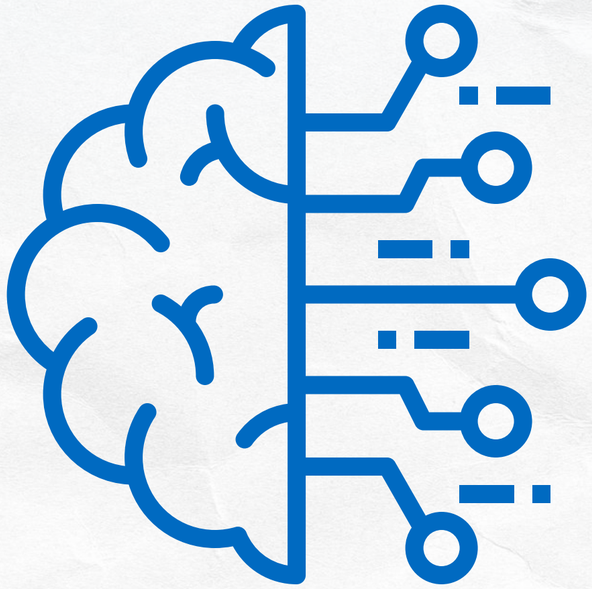
Terlihat bahwa semakin **lama durasi matahari bersinar**, **semakin kecil kemungkinan terjadinya hujan**. Sebaliknya, jika durasi **sinar matahari relatif singkat**, maka **kemungkinan terjadinya hujan menjadi lebih besar.**

RAIN TODAY VS RAIN TOMORROW

YES : 1
NO : 0

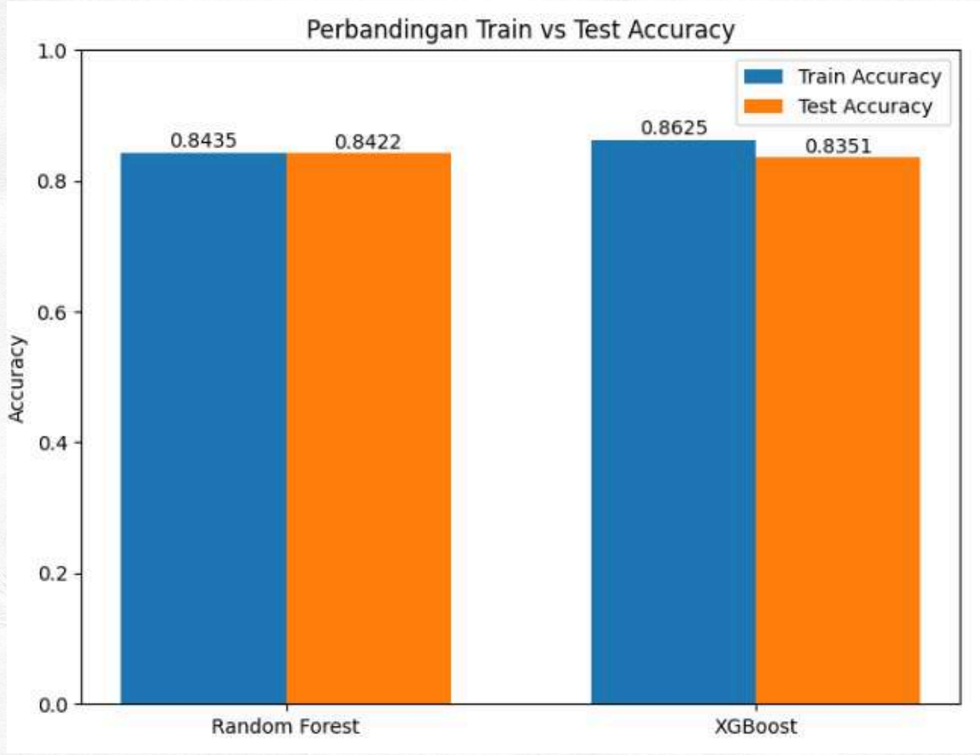


Distribusi data menunjukkan bahwa **jika hari ini terjadi hujan**, **kemungkinan besar besok juga akan hujan**. Namun, ketika **hari ini tidak hujan**, **bukan berarti besok pasti tidak hujan**, karena tabel di atas masih memperlihatkan **adanya kemungkinan hujan** meskipun dengan persentase yang lebih kecil.



MACHINE LEARNING

PERBANDINGAN MODEL RANDOM FOREST DAN XGBOOST



Dari tabel di atas terlihat bahwa **akurasi pada data train dan test untuk model Random Forest relatif seimbang.** Sementara itu, pada model **XGBoost terdapat perbedaan yang cukup besar antara akurasi train dan test**, yang mengindikasikan adanya gejala **overfitting.** Dari hal ini Model yang paling cocok di gunakan adalah **Random Forest.**

RANDOM FOREST

- Precision dipakai jika False Positive penting dan berbahaya.
- Recall dipakai jika False Negative penting dam berbahaya
- Accuracy jika False Positive dan False Negative ga terlalu penting

Test Classification report :

Accuracy : 0.842 **Random forest**

	Precision	Recall	Fi-score	Support
Tidak hujan	0.94	0.86	0.90	9033
Hujan	0.58	0.76	0.66	2251
Accuracy			0.84	11284
Macro avg	0.76	0.81	0.78	11284
Weighted avg	0.86	0.84	0.85	11284

Model Random Forest mencapai akurasi 84,2%, dengan performa sangat baik pada kelas tidak hujan **(precision 0,94; recall 0,86)**. Namun, pada kelas hujan performanya lebih lemah **(precision 0,58; recall 0,76)**. Namun untuk kasus ini angka yang di dapat cukup untuk dapat di pakai dalam aplikasi XYZ karena yang lebih di butuhkan adalah akurasi di banding yang lainnya.

RANDOM FOREST CONFUSION MATRIX

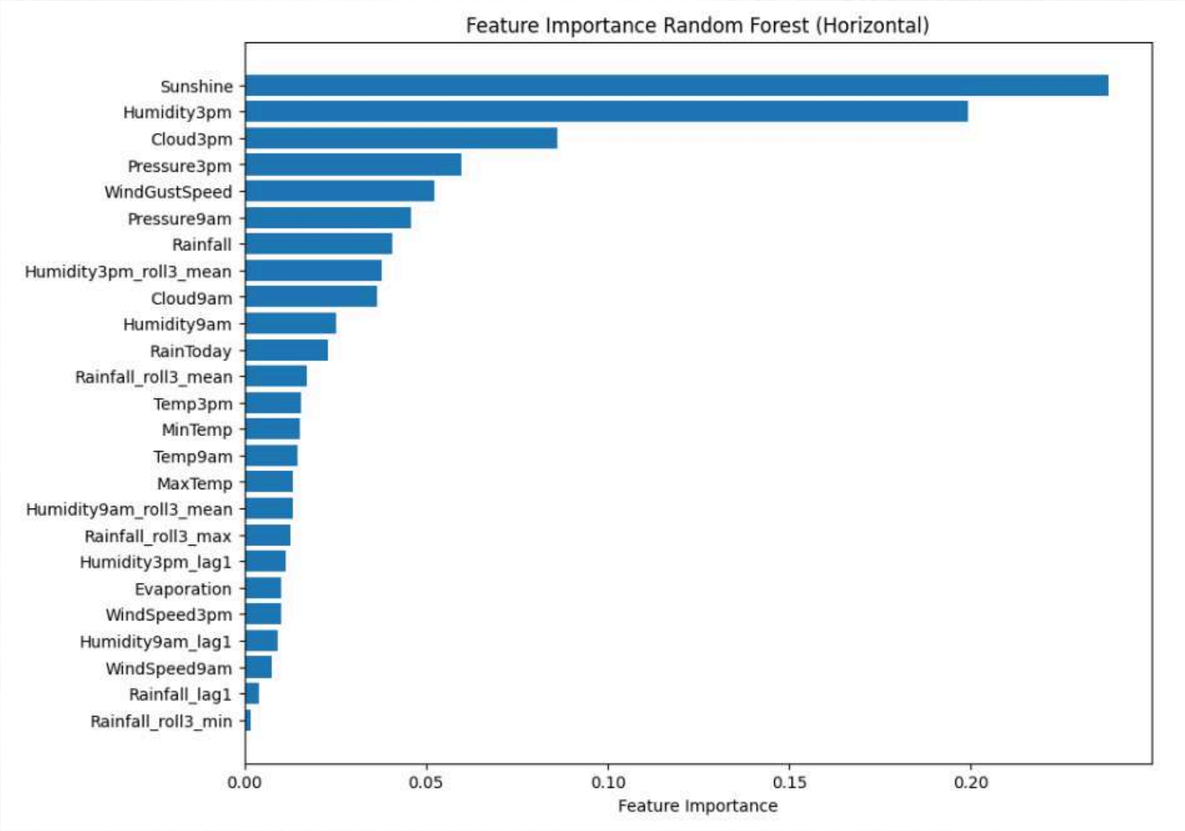
Confusion Matrix :

7792	1241
540	1711

- TP (True Positive): jumlah kasus hujan yang benar diprediksi hujan.
- TN (True Negative): jumlah kasus tidak hujan yang benar diprediksi tidak hujan.
- FP (False Positive): kasus tidak hujan tapi diprediksi hujan (kesalahan "alarm palsu").
- FN (False Negative): kasus hujan tapi diprediksi tidak hujan (kesalahan "missed detection").

Model mampu mengenali hujan sebanyak 1,711 kali dan tidak hujan sebanyak 7,792 kali. tetapi juga ada FP sebanyak 1,241. dan FN sebanyak 540 kali.

FEATURE IMPORTANCE RANDOM FOREST



Berikut adalah **Feature Importance** dari **Random Forest** dimana **Sunshine** dan **Humidity3pm** menjadi salah satu yang terpenting

BUSINESS RECOMMENDATION

BUSINESS RECOMMENDATION

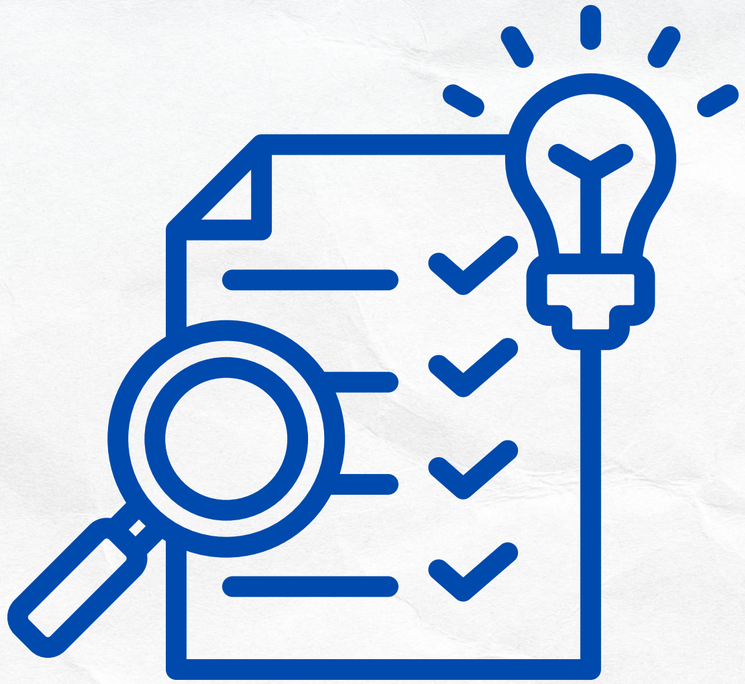
Notifikasi pengingat
Mengirimkan notifikasi jika ada potensi hujan di hari acara, sehingga pengguna bisa antisipasi lebih awal.

Memberikan Rekomendasi Tanggal
Menyarankan tanggal terbaik dengan kemungkinan hujan kecil, agar acara lebih lancar dan partisipasi lebih tinggi..

Menambah predikisi cuaca
Menampilkan prakiraan cuaca untuk membantu pengguna merencanakan event tanpa khawatir hujan.

Menambahkan halaman pembelajaran
Memberikan informasi tanda-tanda alam yang berkaitan dengan hujan, agar pengguna bisa belajar mengenali cuaca.

CONCLUSION



Untuk mengatasi permasalahan ini, Aplikasi XYZ dapat memanfaatkan model **machine learning Random Forest**. **Model ini dapat digunakan untuk mengembangkan fitur prediksi cuaca, khususnya untuk memperkirakan kemungkinan hujan atau tidak pada tanggal tertentu**. Dengan adanya fitur ini, pengguna akan lebih terbantu dalam merencanakan kegiatan, misalnya saat menentukan jadwal bermain bersama.

Link Google Colab

THANK YOU

+62 813 - 8141 - 3963

fahrian1116@gmail.com

FahrianMuhammad