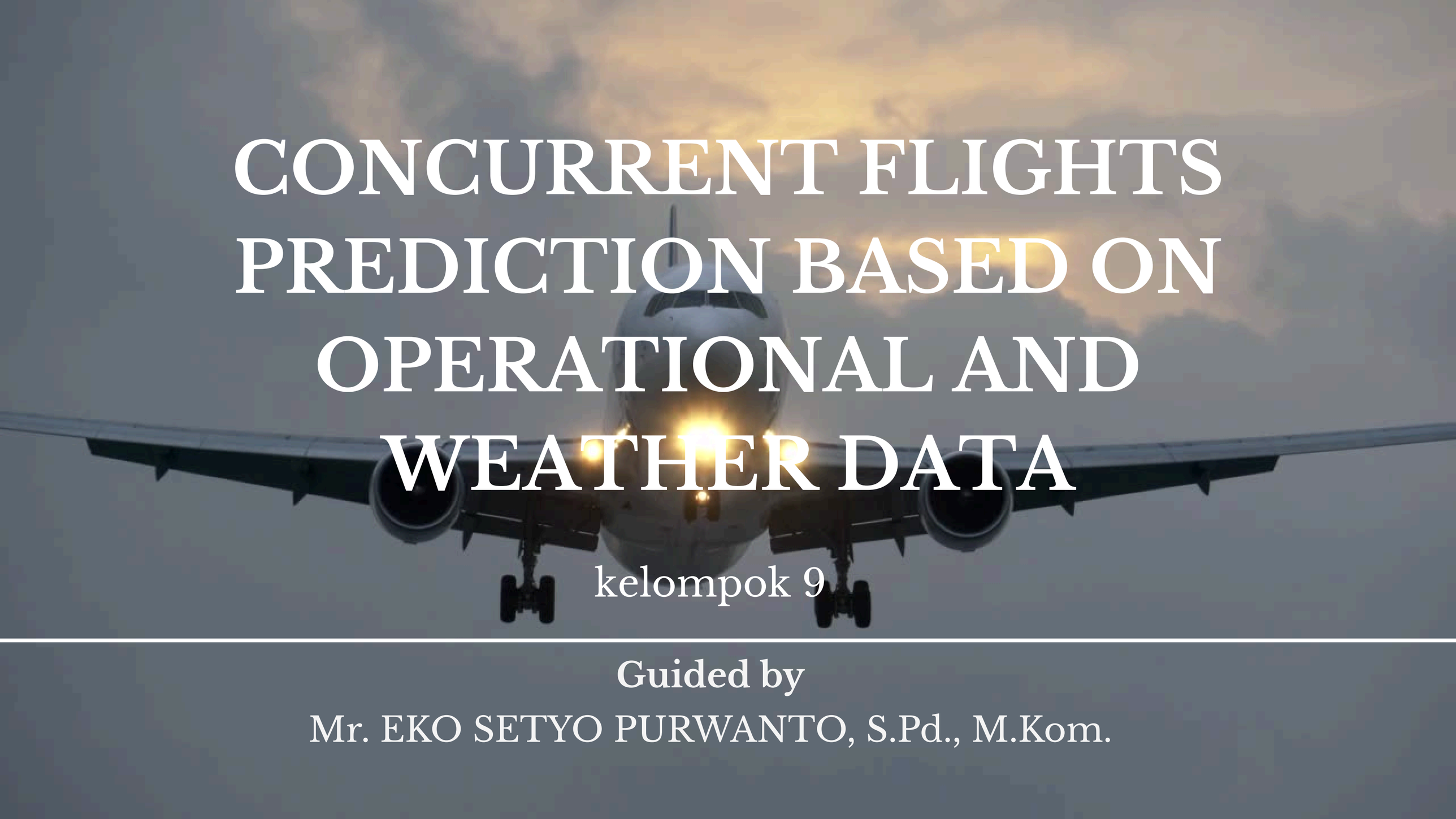


CONCURRENT FLIGHTS PREDICTION BASED ON OPERATIONAL AND WEATHER DATA

kelompok 9

Guided by

Mr. EKO SETYO PURWANTO, S.Pd., M.Kom.

NAMA ANGGOTA

NICHOLAS SINCLAIR ALFIANTO
2702208581

NATASHA DIAN MAHARDITA
2702320210

PAZELLA MUTIA REFLIN
2702343523

VANESSA NAYLA PUTRI
2702235084

PUTRI MAHARANI SETIAWAN
2702375866

LAVINIA NATANIELA NOVYANDI
2702331763

SYAUQINA ZHAFIRAH NUR FAKHRANA
2702360750

DAFTAR ISI

LATAR BELAKANG

Manajemen lalu lintas udara yang efektif membutuhkan prediksi yang akurat terhadap jumlah penerbangan yang terjadi bersamaan (concurrent flights) di setiap waktu tertentu.

Kelebihan kapasitas tanpa perencanaan dapat menyebabkan keterlambatan, pemborosan biaya, dan penurunan layanan.

Dengan memanfaatkan 15 atribut utama dari 26 atribut seperti waktu penerbangan, jarak tempuh, karakteristik pesawat, hingga kondisi cuaca, kami mengembangkan model prediksi menggunakan XGBoost dan Random Forest untuk mendukung optimalisasi operasional bandara dan maskapai.

Analisis data berbasis machine learning ini diharapkan meningkatkan efisiensi, keselamatan, dan pengalaman pelanggan secara keseluruhan.

DATA

# MONTH	# DAY_OF_WEEK	# DEP_DEL15	A DEP_TIME_BLK	# DISTANCE_GROUP	# SEGMENT_NUMBER	# CONCURRENT_FLIG...
Month	Day of Week	TARGET Binary of a departure delay over 15 minutes (1 is yes)	Distance group to be flown by departing aircraft	Distance group to be flown by departing aircraft	The segment that this tail number is on for the day	Concurrent flights leaving from the airport in the same departure block

# NUMBER_OF_SEATS	A CARRIER_NAME	# AIRPORT_FLIGHTS_...	# AIRLINE_FLIGHTS_M...	# AIRLINE_AIRPORT_F...	# AVG_MONTHLY_PA...	# AVG_MONTHLY_PA...
Number of seats on the aircraft	Carrier	Avg Airport Flights per Month	Avg Airline Flights per Month	Avg Flights per month for Airline AND Airport	Avg Passengers for the departing airport for the month	Avg Passengers for airline for month

# FLT_ATTENDANTS_...	# GROUND_SERV_PER...	# PLANE_AGE	A DEPARTING_AIRPORT	A LATITUDE	A LONGITUDE	A PREVIOUS_AIRPORT
Flight attendants per passenger for airline	Ground service employees (service desk) per passenger for airline	Age of departing aircraft	Departing Airport	Latitude of departing airport	Longitude of departing airport	Previous airport that aircraft departed from

# PRCP	# SNOW	# SNWD	# TMAX	# AWND
Inches of precipitation for day	Inches of snowfall for day	Inches of snow on ground for day	Max temperature for day	Max wind speed for day

Sumber Data : https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations

PEMILIHAN ATRIBUT

Atribut dipilih untuk merepresentasikan faktor waktu, operasional penerbangan, serta karakteristik maskapai dan pesawat yang berpengaruh terhadap jumlah penerbangan bersamaan (concurrent flights).

- **Waktu (MONTH, DAY_OF_WEEK, DEP_TIME_BLK): Menangkap pola musiman dan harian penerbangan.**
- **Kinerja Penerbangan (DEP_DEL15, DISTANCE_GROUP, SEGMENT_NUMBER): Menggambarkan potensi keterlambatan dan kompleksitas rute.**
- **Kapasitas (NUMBER_OF_SEATS): Mempengaruhi trafik udara.**
- **Karakteristik Maskapai (CARRIER_NAME, AIRLINE_FLIGHTS_MONTH, AIRLINE_AIRPORT_FLIGHTS_MONTH): Menunjukkan perbedaan frekuensi antar maskapai.**
- **Aktivitas Bandara (AIRPORT_FLIGHTS_MONTH): Mewakili tingkat kepadatan di bandara.**
- **Rasio Layanan (FLT_ATTENDANTS_PER_PASS, GROUND_SERV_PER_PASS): Mempengaruhi kecepatan operasional.**
- **Usia Pesawat (PLANE_AGE): Berkaitan dengan ketepatan operasional.**
- **CONCURRENT_FLIGHTS: Target variabel untuk prediksi.**

DATA CLEANING

1. Memuat Data

Kita mulai dengan memuat dataset yang berisi data tentang penerbangan, bandara, maskapai, dan cuaca.

2. Menghapus Kolom yang Tidak Penting

Beberapa kolom yang tidak terlalu penting untuk prediksi kita dihapus, seperti:

- Nama bandara, maskapai, dan data geografis (koordinat lokasi bandara).
- Informasi cuaca yang bisa mengganggu prediksi.
- Data jumlah penumpang di bandara yang sudah ada di kolom lain.

3. Mengatasi Data yang Hilang

Data yang memiliki nilai kosong (missing) kita hapus supaya model tidak bingung saat dilatih.

DATA PREPROCESSING

4. Menambah Kategori Waktu Keberangkatan

Kolom DEP_TIME_BLK yang berisi jam penerbangan kita bagi menjadi kategori berdasarkan waktu:

- Subuh, Pagi, Siang, Sore, Malam

Lalu, kategori ini diubah menjadi format yang bisa diproses komputer (one-hot encoding).

5. Menentukan Fitur dan Target

- Fitur (X): Kolom-kolom yang digunakan untuk memprediksi (seperti jarak penerbangan, jumlah kursi, dan lainnya).
- Target (y): Kolom yang ingin kita prediksi, yaitu jumlah penerbangan yang bersamaan (CONCURRENT_FLIGHTS).

6. Membagi Data untuk Training dan Testing

Data dibagi menjadi dua bagian:

- 70% untuk melatih model (training).
- 30% untuk menguji model setelah dilatih (testing).

DATA CLEANING & PREPROCESSING

	MONTH	DAY_OF_WEEK	DEP_DEL15	DISTANCE_GROUP	SEGMENT_NUMBER	CONCURRENT_FLIGHTS	NUMBER_OF_SEATS	AIRPORT_FLIGHTS_MONTH
0	1	7	0	2	1	25	143	13056
1	1	7	0	7	1	29	191	13056
2	1	7	0	7	1	27	199	13056
3	1	7	0	9	1	27	180	13056
4	1	7	0	7	1	10	182	13056
...	...	...	...	...	...	...	...	...
6489057	12	7	0	1	11	3	123	1318
6489058	12	7	0	1	11	2	123	1318
6489059	12	7	0	1	11	2	123	1318
6489060	12	7	0	1	12	3	123	1318
6489061	12	7	1	1	12	3	123	1318

AIRLINE_FLIGHTS_MONTH	AIRLINE_AIRPORT_FLIGHTS_MONTH	AVG_MONTHLY_PASS_AIRLINE	FLT_ATTENDANTS_PER_PASS	GROUND_SERV_PER_PASS
107363	5873	13382999	0.000062	0.000099
73508	1174	12460183	0.000144	0.000149
73508	1174	12460183	0.000144	0.000149
73508	1174	12460183	0.000144	0.000149
15023	1257	2688839	0.000009	0.000125
...	...	...	...	...
7268	757	905990	0.000120	0.000198
7268	757	905990	0.000120	0.000198
7268	757	905990	0.000120	0.000198
7268	757	905990	0.000120	0.000198
7268	757	905990	0.000120	0.000198

DATA CLEANING & PREPROCESSING

PLANE_AGE	TMAX	DEP_TIME_Malam	DEP_TIME_Pagi	DEP_TIME_Siang	DEP_TIME_Sore	DEP_TIME_Subuh
8	65.0	0	1	0	0	0
3	65.0	0	1	0	0	0
18	65.0	0	1	0	0	0
2	65.0	0	1	0	0	0
1	65.0	0	0	0	0	1
...	...	...	...	...	...	...
18	84.0	1	0	0	0	0
16	84.0	0	0	0	1	0
18	84.0	0	0	0	1	0
18	84.0	0	0	0	1	0
15	84.0	0	0	0	1	0

XGBOOST

VS

RANDOM FOREST

MODEL YANG DIGUNAKAN

1.XGBoost

XGBoost (Extreme Gradient Boosting) adalah algoritma machine learning decision tree yang menggunakan teknik gradient boosting untuk membuat prediksi yang lebih akurat

- Cocok untuk dataset sedang ke besar (6 juta baris) dan fitur numerik (hasil one-hot).
- Cepat melatih model dengan paralel processing.
- Kuat menangani data kompleks dan outlier.
- Punya regularisasi untuk menghindari overfitting saat prediksi concurrent flights.

TRAINING DAN TESTING

Training model awal menggunakan XGBRegressor

```python
# Define features and target
X, y = df.drop('CONCURRENT_FLIGHTS', axis=1), df.CONCURRENT_FLIGHTS

# Split data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Train XGBoost Model
xgb_reg = xgb.XGBRegressor(
    n_estimators=100,
    learning_rate=0.1,
    max_depth=6,
    subsample=0.8,
    colsample_bytree=0.8,
    random_state=42,
    n_jobs=-1
)

xgb_reg.fit(X_train, y_train)

# Predict
y_pred_xgb = xgb_reg.predict(X_test)
```

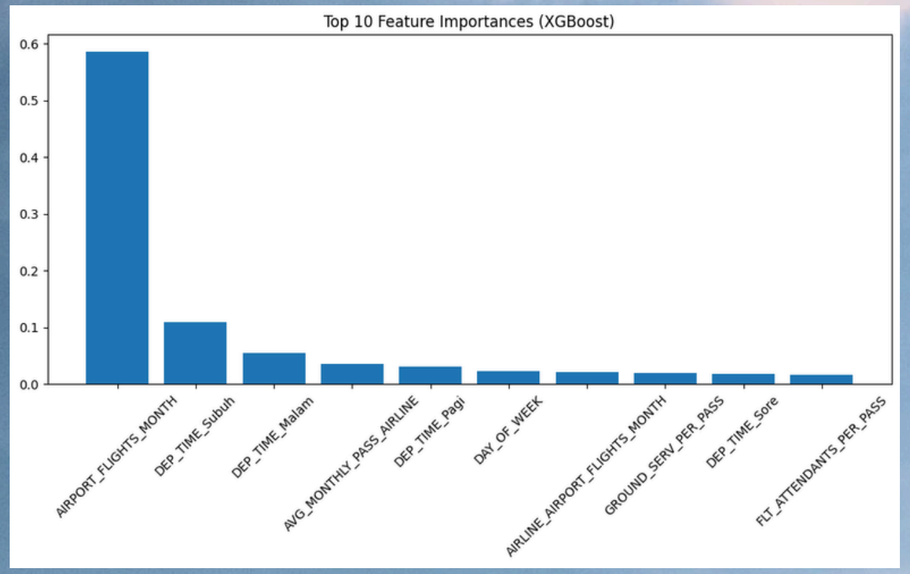
```python
# Evaluate Performance
r2_train_xgb = xgb_reg.score(X_train, y_train)
r2_test_xgb = xgb_reg.score(X_test, y_test)

print(f"Akurasi model pada data training: {r2_train_xgb * 100:.2f}%")
print(f"Akurasi model pada data testing: {r2_test_xgb * 100:.2f}%")

mae_xgb = mean_absolute_error(y_test, y_pred_xgb)
rmse_xgb = math.sqrt(mean_squared_error(y_test, y_pred_xgb))

print(f"Rata-rata kesalahan prediksi (MAE): {mae_xgb:.2f}")
print(f"Rata-rata kesalahan terbesar (RMSE): {rmse_xgb:.2f}")
```

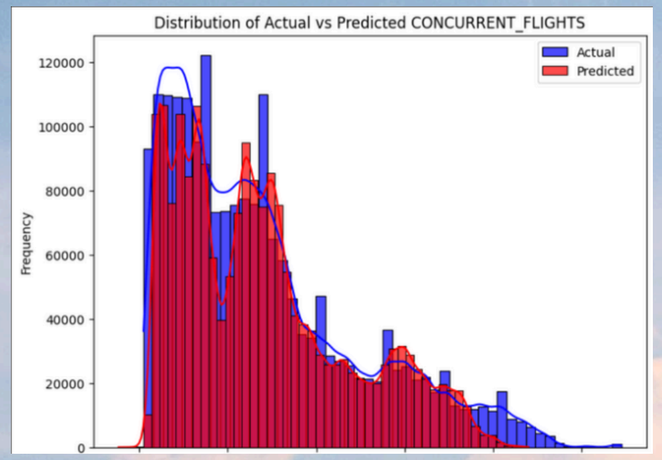
✓ 2.2s

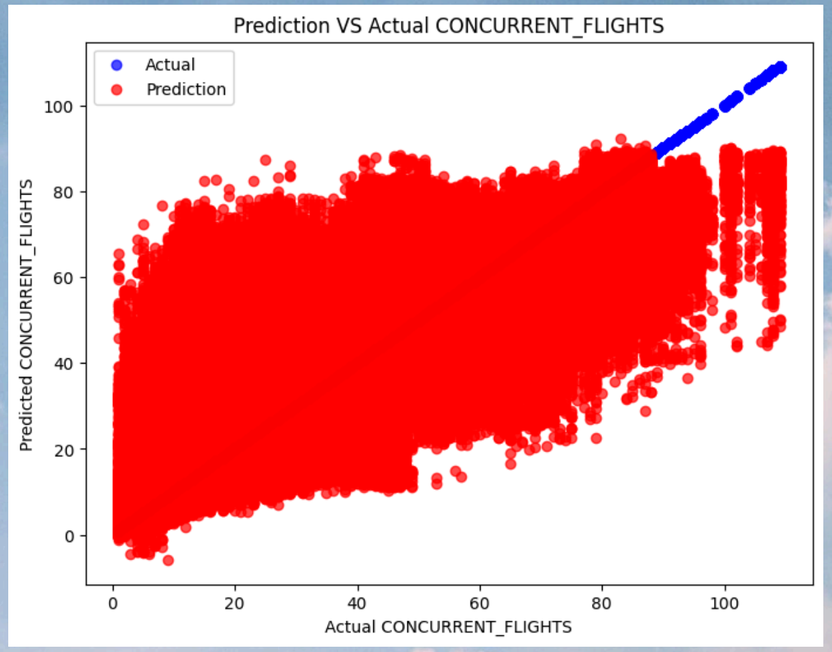
```
Akurasi model pada data training: 81.84%
Akurasi model pada data testing: 81.86%
Rata-rata kesalahan prediksi (MAE): 6.17
Rata-rata kesalahan terbesar (RMSE): 9.16
```

TRAINING DAN TESTING

Actual dan predict concurrent flight







HYPERPARAMETER TUNING

Model di tuning menggunakan XGBRegressor

```
# Sample a smaller dataset for tuning
X_sample, _, y_sample, _ = train_test_split(X, y, train_size=2000000, random_state=42)
```

Memakai sample data agar proses tuning berjalan lebih cepat

```
Step 1 Best Params: {'max_depth': 11, 'min_child_weight': 1}
Step 2 Best Params: {'colsample_bytree': 1.0, 'subsample': 0.9}
Step 3 Best Params: {'learning_rate': 0.05, 'n_estimators': 700}
Step 4 Best Params: {'gamma': 0}
```

Didapatkan hasil parameter sebagai berikut

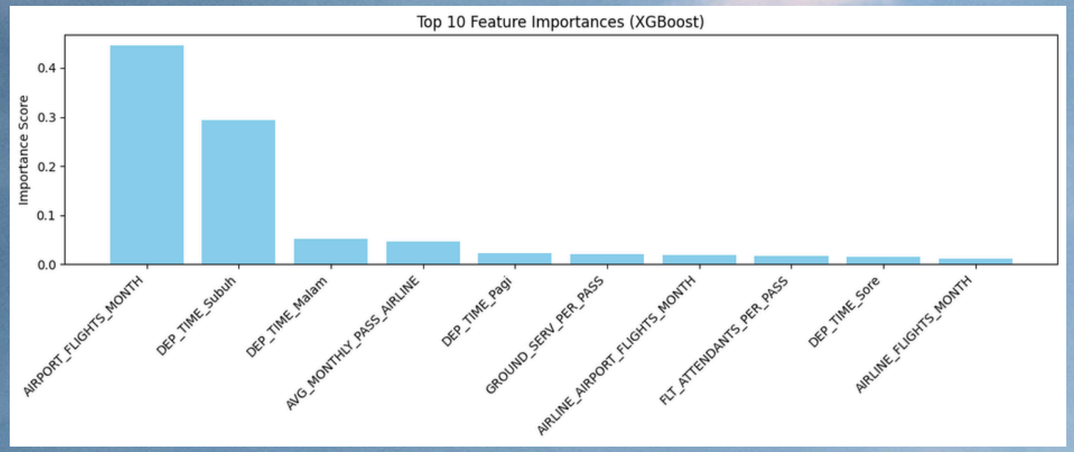
```
FINAL MODEL PERFORMANCE
Akurasi model pada data training: 90.68%
Akurasi model pada data testing: 86.73%
Rata-rata kesalahan prediksi (MAE): 5.02
Rata-rata kesalahan terbesar (RMSE): 7.84
```

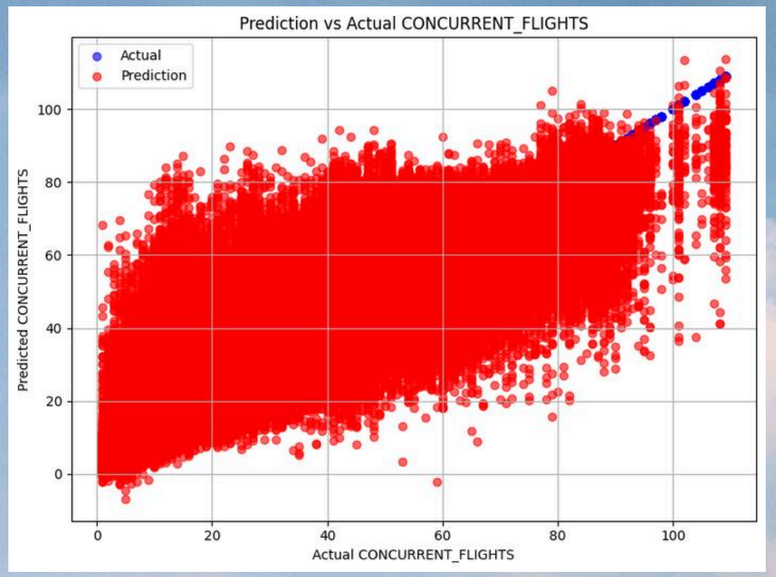
```
# Parameter space untuk RandomizedSearchCV
xgb_param = {
    'n_estimators': [100, 300, 500],
    'max_depth': [3, 5, 7, 9],
    'learning_rate': [0.01, 0.05, 0.1],
    'subsample': [0.6, 0.8, 1.0],
    'colsample_bytree': [0.6, 0.8, 1.0],
    'gamma': [0, 0.1, 0.3]
}
```

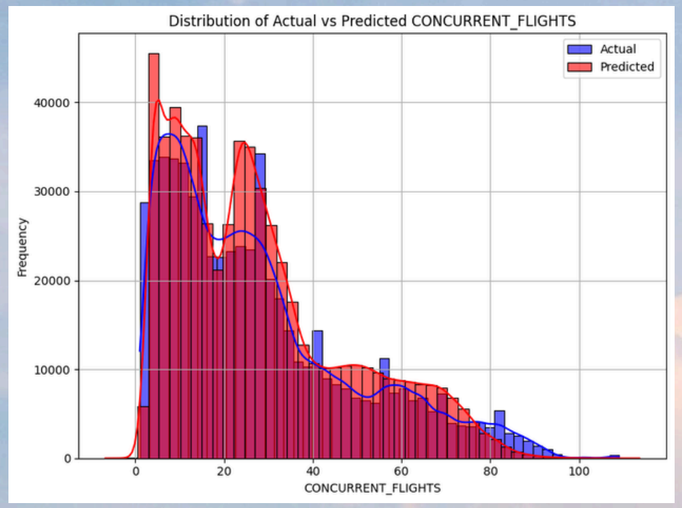
Model di tuning menggunakan XGBRegressor untuk meningkatkan akurasi model agar tidak terjadi overfitting (model machine learning terlalu kompleks sehingga "menghafal" pola dan noise pada data latih (training data), tetapi gagal melakukan generalisasi dengan baik pada data baru atau data uji (testing data).

TRAINING DAN TESTING

Actual dan predict concurrent flight setelah dilakukan tuning



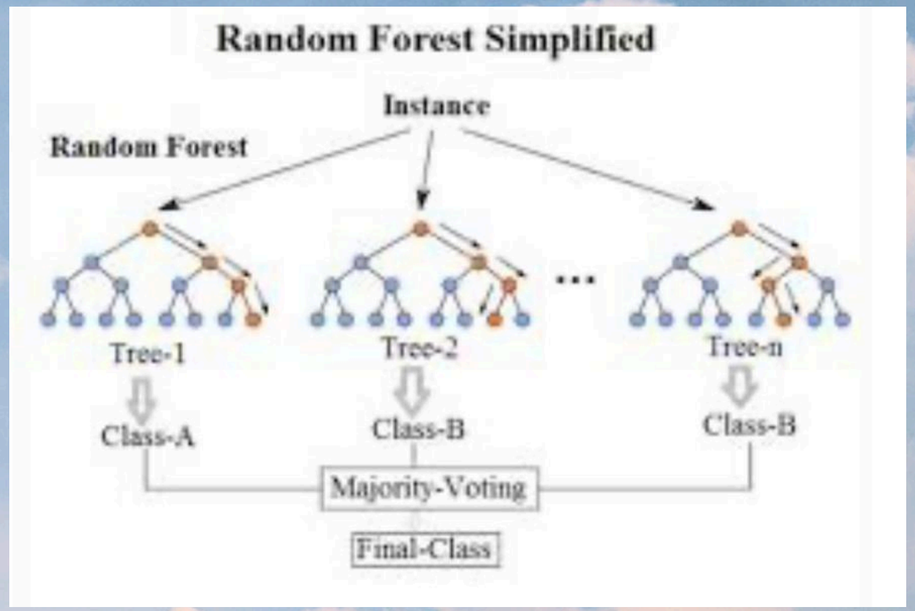




MODEL YANG DIGUNAKAN

2.Random forest

Random Forest adalah ensemble learning method yang menggabungkan banyak pohon keputusan (decision trees) untuk meningkatkan akurasi prediksi dan mengurangi overfitting.



bootstrap aggregating atau bagging. → Decision tree [splitting] → Voting decisions

TRAINING & TESTING

Training dan Testing awal Random Forest

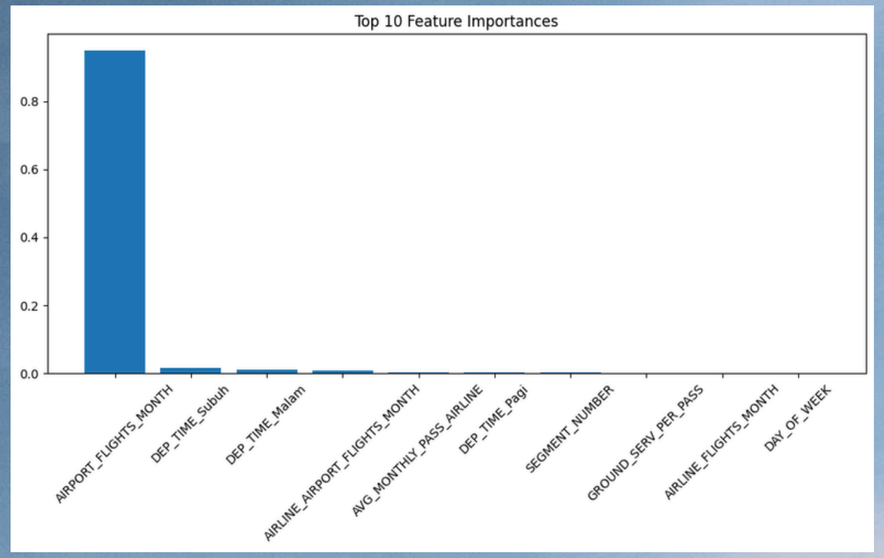
```python
# Train Regression Model with Optimized Hyperparameters
reg = RandomForestRegressor(
    n_estimators=100,
    max_depth=6,
    min_samples_split=10,
    min_samples_leaf=5,
    n_jobs=-1,
    random_state=42
)
reg.fit(X_train, y_train)

# Predict
y_pred = reg.predict(X_test)
```

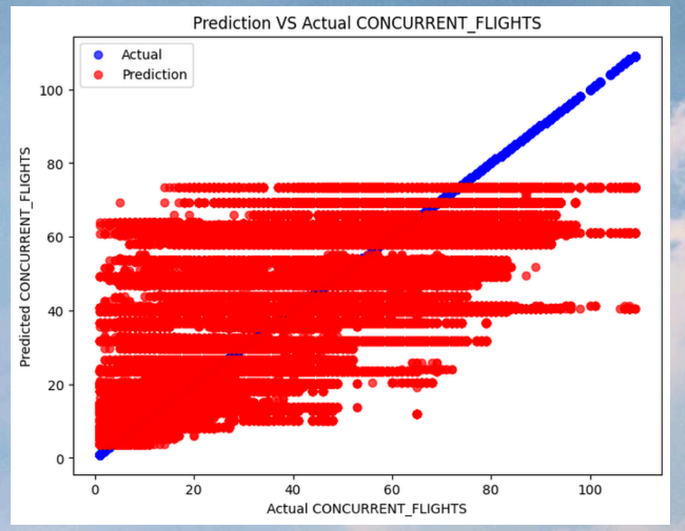
```python
# Evaluate Performance in Simple Terms
r2_train = reg.score(X_train, y_train)
r2_test = reg.score(X_test, y_test)

print(f"Akurasi model pada data training: {r2_train * 100:.2f}%")
print(f"Akurasi model pada data testing: {r2_test * 100:.2f}%")

mae = mean_absolute_error(y_test, y_pred)
rmse = math.sqrt(mean_squared_error(y_test, y_pred))

print(f"Rata-rata kesalahan prediksi (MAE): {mae:.2f}")
print(f"Rata-rata kesalahan terbesar (RMSE): {rmse:.2f}")
```

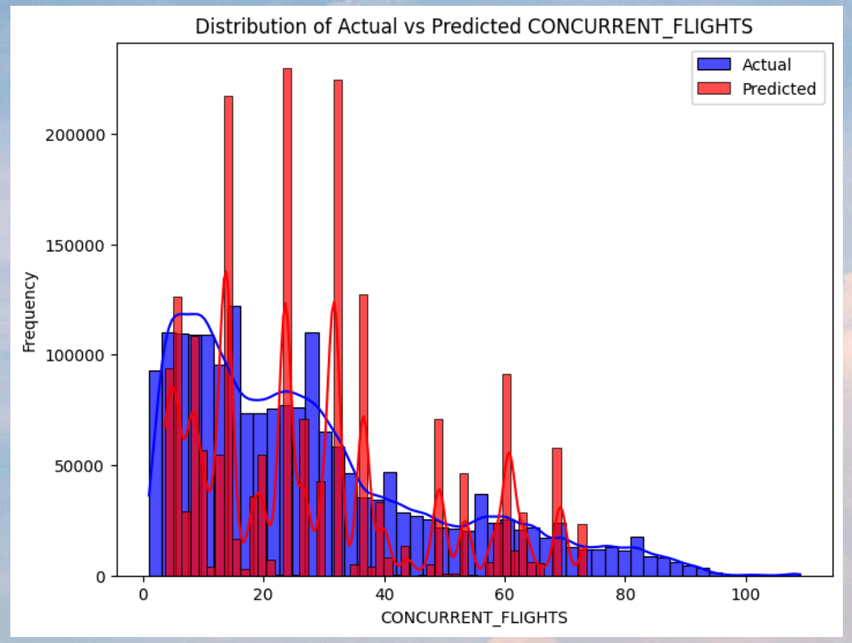
```
Akurasi model pada data training: 76.72%
Akurasi model pada data testing: 76.74%
Rata-rata kesalahan prediksi (MAE): 7.09
Rata-rata kesalahan terbesar (RMSE): 10.38
```

TRAINING & TESTING

Training dan Testing awal Random Forest







HYPERPARAMETER TUNING

Model di tuning menggunakan Random forest

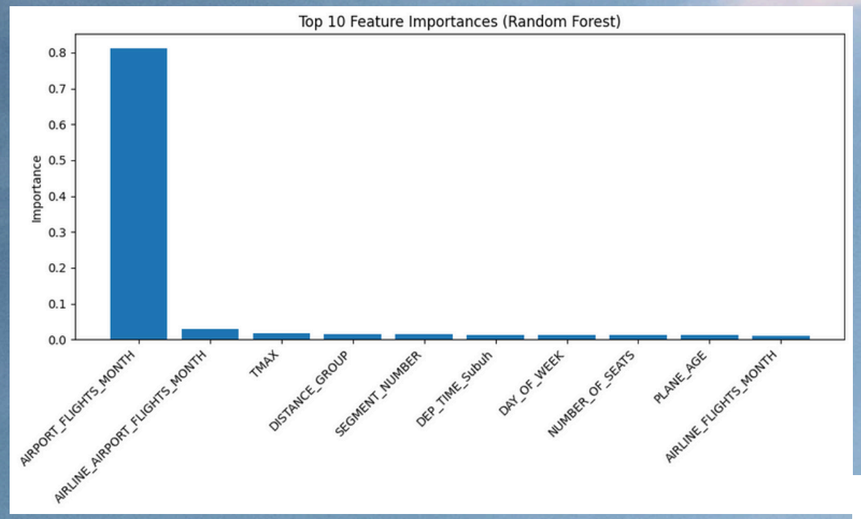
```
param_dist = {
    'n_estimators': [100, 300, 500],
    'max_depth': [None, 20, 30, 40],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4, 8],
    'max_features': ['sqrt', 'log2'],
    'bootstrap': [True, False]
}
```

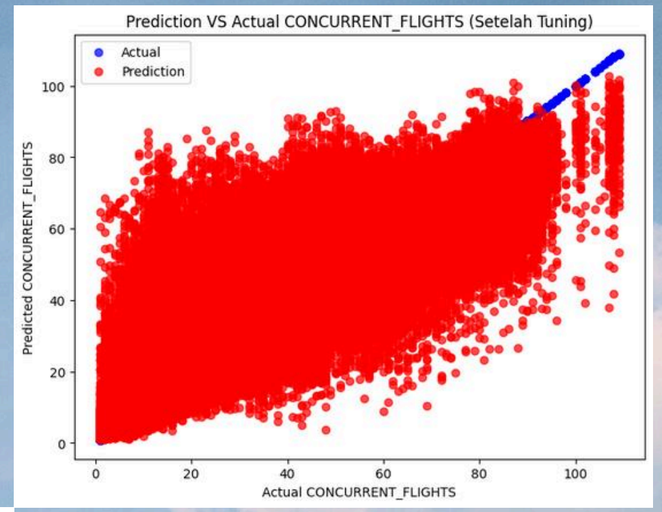
```
r2_train = best_model.score(X_train, y_train)
r2_test = best_model.score(X_test, y_test)

print(f"Akurasi model pada data training: {r2_train * 100:.2f}%")
print(f"Akurasi model pada data testing: {r2_test * 100:.2f}%")

mae = mean_absolute_error(y_test, y_pred)
rmse = math.sqrt(mean_squared_error(y_test, y_pred))

print(f"Rata-rata kesalahan prediksi (MAE): {mae:.2f}")
print(f"Rata-rata kesalahan terbesar (RMSE): {rmse:.2f}")
```

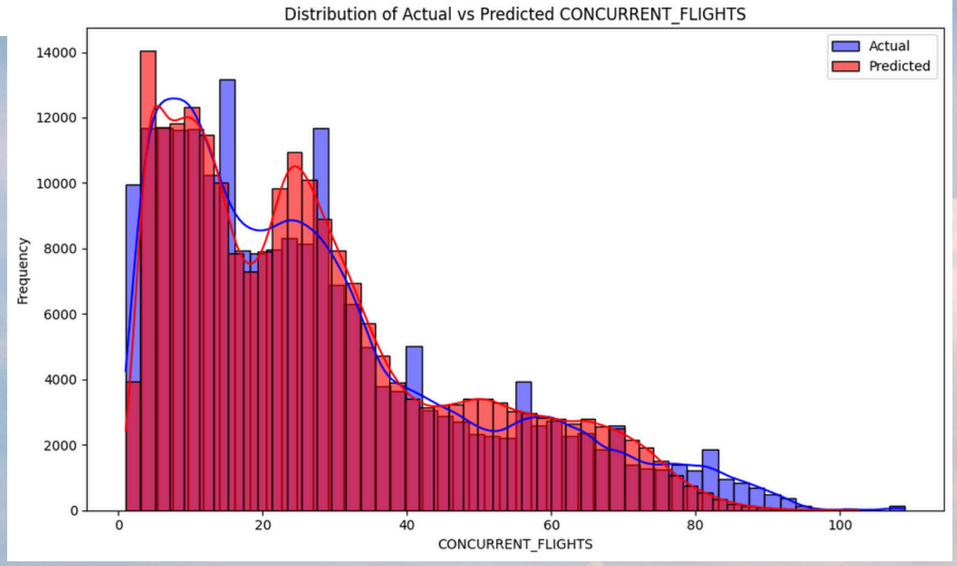
```
Akurasi model pada data training: 91.24%
Akurasi model pada data testing: 86.28%
Rata-rata kesalahan prediksi (MAE): 4.91
Rata-rata kesalahan terbesar (RMSE): 7.98
```

TRAINING DAN TESTING

Model di tuning menggunakan Random forest







DASHBORD INTERAKTIF INSIGHT BISNIS - POWERBI

Dashboard Prediksi & Analisis Data Penerbangan

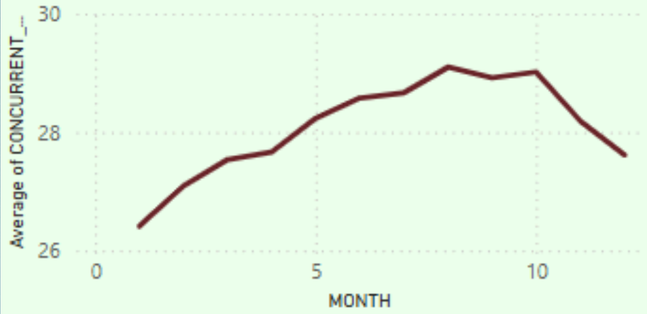
DEPARTING_AIRP... All | PREVIOUS_AIRPO... All | DEP_TIME_B... All | CARRIER_NAME Multiple selections | DAY_OF_... All

DISTANCE_G... All | MONTH All

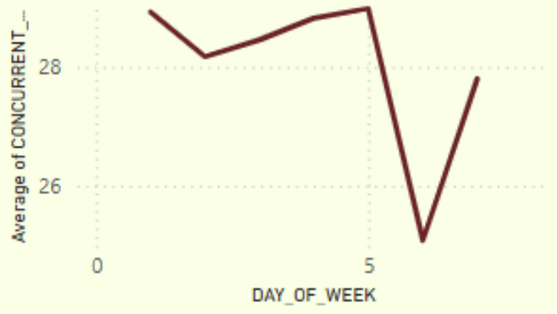
AVG Airplane Airport Flights Month	AVG Airline Flights Month	Total Airport Flights Month	Total Concurent Flights
3.50K	64.53K	80bn	176M

Delayed Percentage	Total Seats Booked	AVG Max Temperature	AVG Plane Age
19.01	829M	716.29	11.66

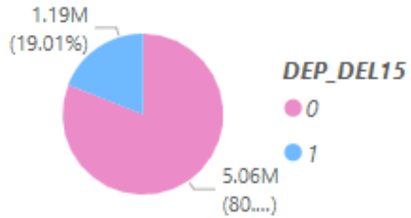
Trend of Average CONCURRENT_FLIGHTS by MONTH



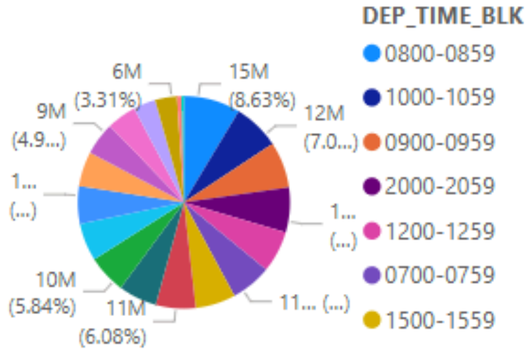
Trend of Average CONCURRENT_FLIGHTS by WEEK



Flight Delay Distribution

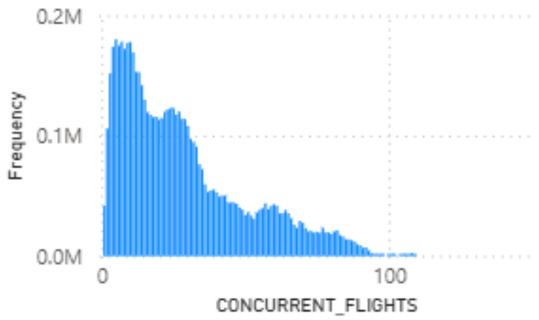


Flight Percentage per Dep Time

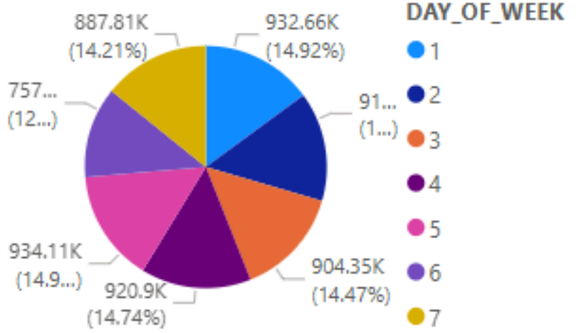


DEPARTING_AIRPORT	Total Concurrent Flights	Seats Sold	Airport Flights Month
Puerto Rico International	630	13573	192612
Sanford NAS	10831	295797	2628985
Portland International Jetport	14037	359924	4358465
Pensacola Regional	12180	395185	4268926
Spokane International	10242	398392	3406904
Palm Springs International	19089	468702	6175898
Total	**175746730**	**829280092**	**79863280023**

Distribution of CONCURRENT_FLIGHTS



Percentage of Flight by Day of the Week



KESIMPULAN

Analisis data ini berhasil membangun model prediksi jumlah penerbangan bersamaan (concurrent flights) menggunakan dataset besar (6 juta row) yang telah dibersihkan dan diproses, termasuk menghapus kolom tidak relevan, penanganan missing values, dan kategorisasi waktu keberangkatan. Dua algoritma yang digunakan adalah Random Forest dan XGBoost, dengan hasil akhir yang seimbang Random Forest mencapai akurasi testing 86.3% dan XGBoost 86.7% setelah hyperparameter tuning. Kedua model menunjukkan performa yang baik, namun XGBoost lebih unggul karena fleksibilitas dan kemampuan regulasinya menjadikannya pilihan yang tepat untuk kasus prediksi berskala besar seperti ini. Model ini dapat dikembangkan lebih lanjut untuk mendukung pengambilan keputusan dalam pengelolaan lalu lintas udara.

DOKUMENTASI

Link Google drive Dokumentasi:

https://drive.google.com/drive/folders/1aasLtL91DVTQG4ZR6IDApnOjpN63xhyc?usp=sharing


Q&A


TERIMA
KASIH