Project Big Data Processing

- Gandi Napoleon Putra – 2702236250
- Benny Linardi Efendy – 2702277105
- Matthew Rafael Suleman - 2702232082
- Ferdy Yusuf Tuharea - 2702277124
- Nicholas Tantama - 2702212414
- Nathaniel Kevin Heriawan - 2702254165

Prediksi Harga Tiket Pesawat Menggunakan Random Forest

DAFTAR ISI

Latar Belakang

Transportasi udara merupakan moda penting dalam perjalanan jarak jauh, dan harga tiket menjadi faktor utama dalam pengambilan keputusan penumpang. Sayangnya, harga tiket sangat dinamis dan dipengaruhi oleh berbagai faktor seperti musim, permintaan, jumlah kursi, dan strategi maskapai. Sistem dynamic pricing yang digunakan maskapai membuat harga dapat berubah dalam hitungan menit, menyulitkan penumpang menentukan waktu terbaik untuk membeli tiket.

Seiring berkembangnya teknologi dan ketersediaan big data, pendekatan berbasis Machine Learning kini memungkinkan untuk memprediksi harga tiket secara lebih akurat. Salah satu algoritma yang digunakan adalah Random Forest Regressor, yang unggul karena mampu menangani data besar, mengurangi overfitting, dan memberikan prediksi stabil.

Latar Belakang

Proyek ini memanfaatkan dataset Flight Price Prediction dari Kaggle dengan lebih dari 300.000 data penerbangan domestik dan internasional di India. Tujuannya adalah untuk:

- Membangun model prediksi harga tiket berbasis machine learning.
- Mengevaluasi performa model dengan metrik regresi.
- Menunjukkan potensi machine learning dalam industri penerbangan.

Dengan pendekatan ini, diharapkan tercipta solusi cerdas dalam membantu konsumen dan maskapai membuat keputusan yang lebih efisien dan strategis.

Jika perlu, saya bisa bantu mendesain slide-nya langsung juga.

Borcelle

Dataset

Flight Price Prediction

by : Shubham Bathwal

Link Dataset :

https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction

Methodology dan Alur Pekerjaan

1. **DATA COLLECTION**
   Mengambil dataset dari Kaggle berisi >300.000 data penerbangan.

2. **DATA PREPARATION**
   Membersihkan data, menghapus kolom tidak relevan, dan ubah durasi ke format numerik.

3. **DATA PREPROCESSING**
   Encoding data kategorikal, normalisasi, dan split data menjadi training dan testing.

4. **MODEL TRAINING (RANDOM FOREST)**
   Melatih model Random Forest untuk prediksi harga tiket secara akurat.

5. **MODEL EVALUATION**
   Evaluasi performa model menggunakan MAE, RMSE, dan R^2 Score.

6. **VISUALIZATION & INTERPRETATION**
   Visualisasi hasil prediksi dan analisis fitur yang paling memengaruhi harga.

Data Preparation

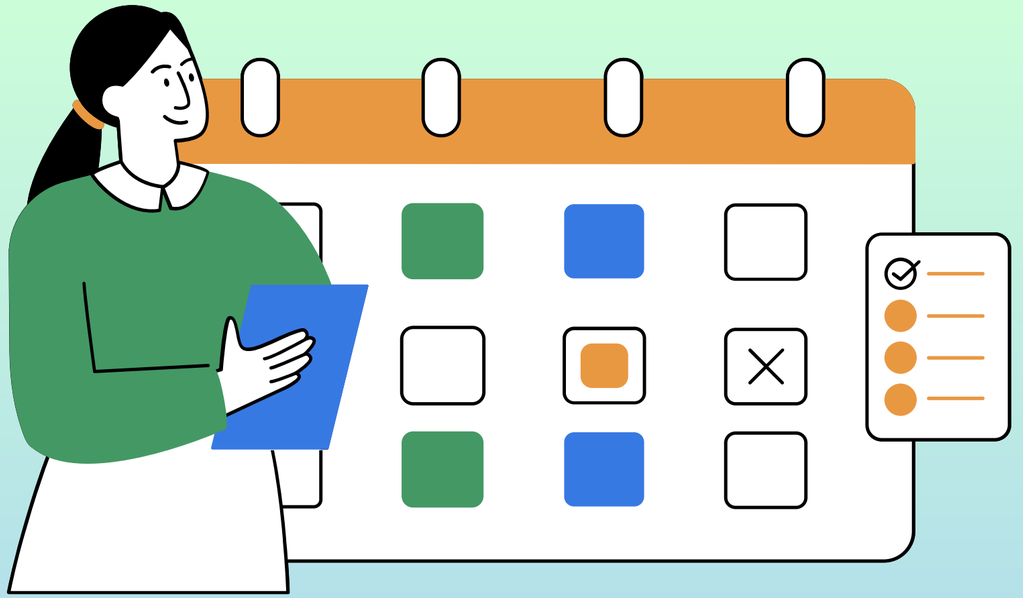
PROSES DIAWALI DENGAN MEMBACA DATASET DALAM FORMAT CSV KE DALAM SPARK DATAFRAME. DATASET TERDIRI DARI 300.153 BARIS DATA. KOLOM _C0 YANG MERUPAKAN INDEKS DIHAPUS, DAN DILAKUKAN PEMERIKSAAN NILAI KOSONG UNTUK MEMASTIKAN TIDAK ADA DATA YANG PERLU DI-DROP.

```
from google.colab import drive
drive.mount('/content/drive')
```

SALAH SATU PREPROCESSING PENTING ADALAH KONVERSI KOLOM DURATION, YANG AWALNYA BERUPA STRING WAKTU (CONTOHNYA "2H 30M") MENJADI NILAI NUMERIK (DALAM SATUAN JAM DESIMAL) AGAR DAPAT DIPROSES DALAM MODEL REGRESI.

Borcelle

Data Preparation



Tahap selanjutnya adalah inisialisasi Spark session dengan nama “Flight Price Prediction”. Setelah session berhasil dibuat, dataset dimuat ke dalam DataFrame PySpark dan ditampilkan beberapa baris pertama untuk validasi isi dan struktur datanya.

Skema data juga diperiksa untuk memastikan semua tipe kolom sesuai dengan yang diharapkan (misalnya kolom price sebagai integer dan duration sebagai double).

```
import pandas as pd
import numpy as np
import time
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

file_path = '/content/drive/MyDrive/GoogleCollab/Clean_Dataset.csv'
df = pd.read_csv(file_path)

print("Jumlah data:", len(df))
print("\nContoh data:")
print(df.head())
```

Data Preprocessing

```
label_cols = ['airline', 'flight', 'source_city', 'departure_time', 'stops',
              'arrival_time', 'destination_city', 'class']

le = LabelEncoder()
for col in label_cols:
    df[col] = le.fit_transform(df[col])

X = df.drop('price', axis=1)
y = df['price']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

TAHAP INI MELIBATKAN BEBERAPA LANGKAH TRANSFORMASI UNTUK MENYIAPKAN DATA MENJADI INPUT YANG SESUAI UNTUK MODEL MACHINE LEARNING:

- IDENTIFIKASI KOLOM KATEGORIKAL SEPERTI AIRLINE, SOURCE_CITY, DEPARTURE_TIME, STOPS, ARRIVAL_TIME, DESTINATION_CITY, DAN CLASS.
- KONVERSI DATA KATEGORIKAL KE BENTUK NUMERIK DENGAN STRINGINDEXER.
- TRANSFORMASI HASIL INDEKS KE VEKTOR NUMERIK MENGGUNAKAN ONEHOTENCODER.
- PENGGABUNGAN SEMUA FITUR NUMERIK DAN ENCODED KE DALAM SATU VEKTOR MENGGUNAKAN VECTORASSEMBLER.
- NORMALISASI FITUR MENGGUNAKAN STANDARDSCALER UNTUK MENCEGAH FITUR DENGAN SKALA BESAR MENDOMINASI.
- DATASET DIBAGI MENJADI DUA BAGIAN: 80% UNTUK TRAINING DAN 20% UNTUK TESTING MENGGUNAKAN FUNGSI RANDOMSPLIT().

Model Training (Random Forest)

```
start_time = time.time()
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
training_time = time.time() - start_time

y_pred = model.predict(X_test)
```

Model yang digunakan dalam eksperimen ini adalah Random Forest Regressor, yang merupakan algoritma ensemble berbasis Decision Tree.
Langkah - Langkah:

Inisialisasi Random Forest Regressor dengan kolom input (featuresCol) dan target (labelCol = price).

Training model dilakukan pada data training dengan metode .fit().

Model yang telah dilatih kemudian digunakan untuk memprediksi harga tiket pada data testing dengan metode .transform().

Model Evaluation

Evaluasi model dilakukan dengan membandingkan hasil prediksi terhadap nilai aktual pada data testing menggunakan tiga metrik regresi:

```
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, y_pred)

print("\n=== HASIL EVALUASI MODEL ===")
print(f"Training time: {training_time:.2f} detik")
print(f"MAE (Mean Absolute Error): {mae:.2f}")
print(f"MSE (Mean Squared Error): {mse:.2f}")
print(f"RMSE (Root MSE): {rmse:.2f}")
print(f"R² Score: {r2:.4f}")
```

1. Mean Absolute Error (MAE) – menunjukkan rata-rata kesalahan absolut antara prediksi dan nilai aktual.
2. Root Mean Squared Error (RMSE) – mengukur besar kesalahan dengan penalti lebih tinggi untuk prediksi yang jauh dari nilai sebenarnya.
3. R-squared (R^2) – mengukur seberapa besar variasi dalam data dapat dijelaskan oleh model.

Evaluasi dilakukan dengan menggunakan RegressionEvaluator dari PySpark MLlib.

Visualization & Interpretation

```python
import matplotlib.pyplot as plt

plt.scatter(y_test, y_pred, alpha=0.5)
plt.xlabel('Harga Asli')
plt.ylabel('Harga Prediksi')
plt.title('Prediksi vs Harga Asli')
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], 'r--')
plt.show()

residuals = y_test - y_pred
plt.scatter(y_pred, residuals, alpha=0.5)
plt.hlines(y=0, xmin=min(y_pred), xmax=max(y_pred), colors='r')
plt.xlabel('Prediksi')
plt.ylabel('Residual (Error)')
plt.title('Plot Residual')
plt.show()

importances = model.feature_importances_
features = X.columns
imp_df = pd.DataFrame({'Feature': features, 'Importance': importances})
imp_df = imp_df.sort_values(by='Importance', ascending=False)
print(imp_df)
```

Tahap akhir dari proses adalah visualisasi hasil dan pemahaman data:

1. Menampilkan tabel hasil prediksi sampel: harga aktual vs hasil prediksi.
2. Visualisasi distribusi harga dan hubungan antara prediksi dengan nilai aktual melalui scatter plot.
3. Membuat grafik perbandingan MAE, RMSE, dan R^2 secara visual untuk interpretasi performa model.
4. Analisis tambahan dilakukan terhadap fitur-fitur penting yang paling berpengaruh dalam prediksi berdasarkan featureImportances dari model Random Forest.

Evaluasi : Alur Pemrosesan Data

I . Label Encoding

Beberapa kolom dengan data kategorik dikonversi ke bentuk numerik menggunakan teknik Label Encoding dari pustaka Scikit-Learn. Kolom-kolom tersebut meliputi:

- airline
- flight
- source_city
- departure_time

II. Pemilihan Target

Fitur target atau variabel dependen yang diprediksi adalah price, yaitu harga tiket pesawat.

III. Pembagian Data (Train-Test Split)

Setelah tahap encoding selesai, data dibagi menjadi dua bagian menggunakan train_test_split dari Scikit-Learn, dengan proporsi:

- 80% untuk data latih (training set)
- 20% untuk data uji (testing set)

Tujuan pembagian ini adalah untuk memastikan model dapat diuji pada data yang tidak pernah dilihat sebelumnya, sehingga hasil evaluasi mencerminkan kemampuan generalisasi model terhadap data baru.

Evaluasi : Pelatihan dan Performa

No	Metrics	Score
1	MAE (Mean Absolute Error)	771.64
2	MSE (Mean Squared Error)	4571152.26
3	RMSE (Root MSE):	2138.03
4	R^2 Score	0.991

MAE (Mean Absolute Error): 771.64
→ Rata-rata selisih absolut antara harga asli dan prediksi adalah ±771.64 rupiah.
Contoh: Jika harga asli tiket adalah 50.000, maka model memprediksi rata-rata di kisaran ±49.200 hingga 50.800.
MSE (Mean Squared Error): 4.571.152,26
→ Rata-rata kuadrat dari error. Karena dikuadratkan, outlier akan memberi dampak besar.
Digunakan untuk menghitung RMSE.
RMSE (Root Mean Squared Error): 2138.03
→ Akar dari MSE, menunjukkan kesalahan prediksi dalam satuan yang sama dengan target (harga).
→ Artinya, prediksi harga tiket rata-rata meleset sebesar ±2.138 rupiah.
R^2 Score (Koefisien Determinasi)

- R^2 Score: 0.9911
- → Artinya model dapat menjelaskan 99,11% variasi harga tiket dari fitur-fitur yang diberikan.
- → Nilai ini sangat tinggi, menandakan bahwa model memiliki performa prediksi yang sangat baik.

Keunggulan Model Random Forest

I. Mengatasi Overfitting

Random Forest menggabungkan banyak Decision Tree dan melakukan averaging pada hasilnya, sehingga secara signifikan mampu mengurangi risiko overfitting yang umum terjadi pada pohon keputusan tunggal.

II. Dapat Menangani Fitur Kategorikal

Setelah proses encoding, Random Forest dapat dengan mudah memproses fitur kategorikal dan menangkap hubungan kompleks antar fitur tanpa perlu eksplisit menyatakan interaksi tersebut.

III. Skalabilitas terhadap Data Besar dan Berdimensi Tinggi

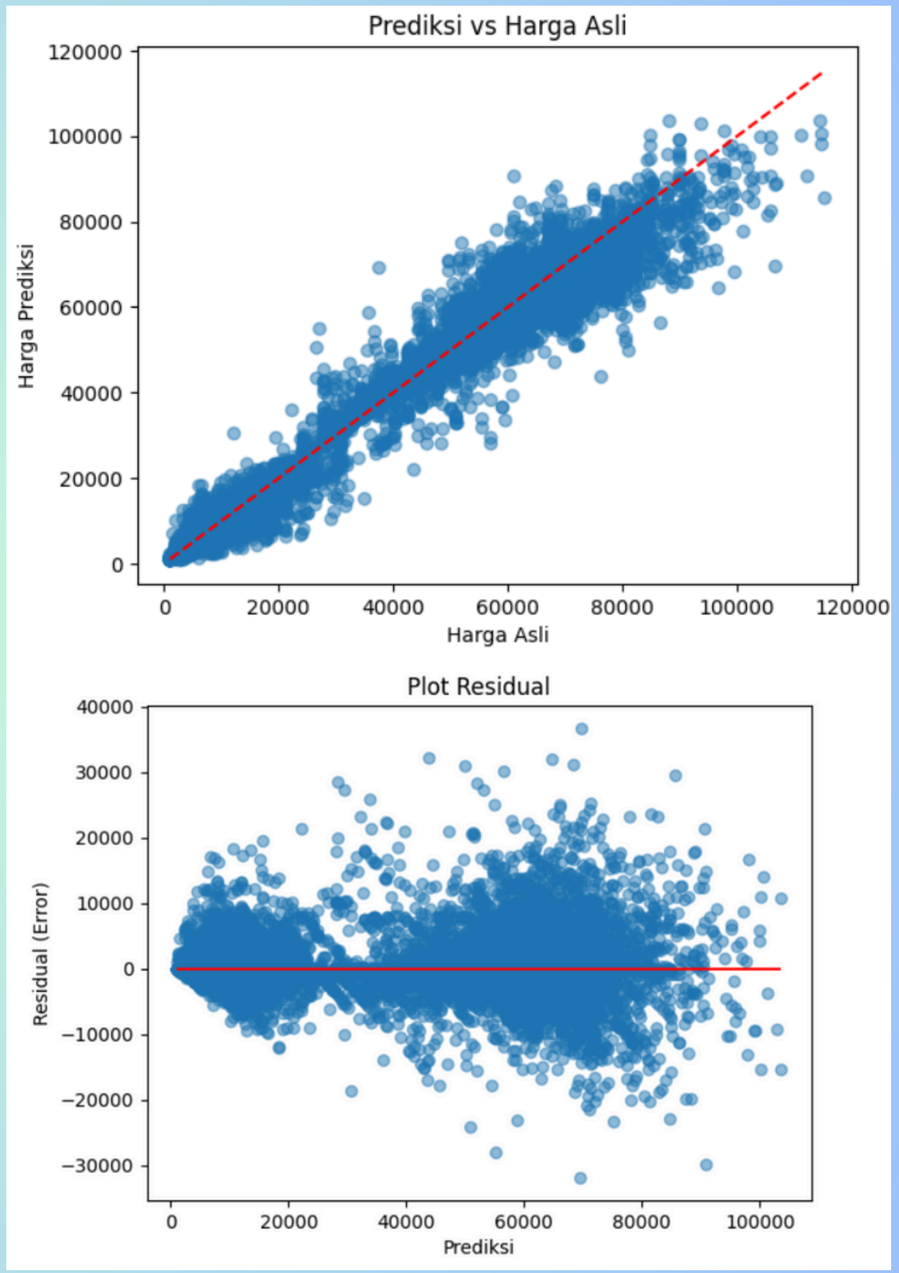
Model ini terbukti mampu menangani dataset besar dengan cepat dan efisien, baik dari segi pelatihan maupun prediksi.

IV. Ketahanan terhadap Outliers dan Noise

Berbeda dengan model regresi linier yang sangat sensitif terhadap data ekstrem, Random Forest cenderung menstabilkan pengaruh noise dengan melakukan averaging pada banyak model dasar.

Scatter Chart

1. PREDIKSI SANGAT MENDEKATI NILAI AKTUAL
GRAFIK PREDIKSI VS HARGA ASLI MENUNJUKKAN BAHWA SEBAGIAN BESAR TITIK DATA TERSEBAR DEKAT DENGAN GARIS DIAGONAL (GARIS MERAH PUTUS-PUTUS).
INI MENANDAKAN BAHWA MODEL BERHASIL MEMPREDIKSI HARGA TIKET DENGAN TINGKAT AKURASI YANG TINGGI.
2. DISTRIBUSI ERROR CUKUP SIMETRIS
PLOT RESIDUAL (ERROR) MEMPERLIHATKAN SEBARAN ERROR YANG RELATIF SIMETRIS DI SEKITAR GARIS NOL (GARIS MERAH), TANPA POLA SISTEMATIS.
ARTINYA, MODEL TIDAK MEMILIKI BIAS SIGNIFIKAN, DAN KESALAHAN PREDIKSI TERSEBAR MERATA.
3. FITUR PALING BERPENGARUH
TIGA FITUR PALING PENTING DALAM PREDIKSI HARGA TIKET:
UNNAMED: 0 (SERIAL NUMBER) → 0.83
CLASS (KELAS TIKET: EKONOMI/BISNIS) → 0.86
DURATION (DURASI PENERBANGAN) → 0.64
INI MENUNJUKKAN BAHWA KELAS PENERBANGAN DAN DURASI MEMILIKI PENGARUH BESAR TERHADAP HARGA.



Borcelle

Kesimpulan

Model Random Forest Regressor Efektif untuk Prediksi Harga Tiket

Model mampu memberikan prediksi harga yang akurat berdasarkan fitur-fitur penting seperti maskapai, kota asal/tujuan, waktu keberangkatan, dan jumlah hari sebelum keberangkatan.

Machine Learning Berpotensi Mendukung Perencanaan Perjalanan

Dengan prediksi harga yang lebih tepat, calon penumpang dapat menentukan waktu terbaik untuk membeli tiket, sehingga menghemat biaya perjalanan.

Pemanfaatan Data Besar Mendukung Inovasi di Industri Penerbangan

Dataset besar seperti Flight Price Prediction memungkinkan pengembangan solusi cerdas untuk strategi harga dan manajemen penumpang di masa depan.

Referensi

[1] SIMPLE FLYING. (2023). 5 KEY FACTORS THAT INFLUENCE AIRLINE TICKET PRICES. HTTPS://SIMPLEFLYING.COM/AIRLINE-TICKET-PRICES-INFLUENTIAL-FACTORS-LIST/

[2] AVIATIONFILE. (2022). *FACTORS AFFECTING FLIGHT TICKET PRICES*. HTTPS://WWW.AVIATIONFILE.COM/FACTORS-AFFECTING-FLIGHT-TICKET-PRICES/

[3] OAG. (2022). *THE STORY OF AIRLINE PRICING STRATEGIES*. HTTPS://WWW.OAG.COM/BLOG/THE-STORY-OF-AIRLINE-PRICING-STRATEGIES

[4] ANALYTICS VIDHYA. (2022). *FLIGHT PRICE PREDICTION USING MACHINE LEARNING*. HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2022/01/FLIGHT-FARE-PREDICTION-USING-MACHINE-LEARNING/

[5] SCIKIT-LEARN. (2024). RANDOMFORESTREGRESSOR — SKLEARN.ENSEMBLE.RANDOMFORESTREGRESSOR. HTTPS://SCIKIT-LEARN.ORG/STABLE/MODULES/GENERATED/SKLEARN.ENSEMBLE.RANDOMFORESTREGRESSOR.HTML

[6] KAGGLE. (N.D.). *FLIGHT PRICE PREDICTION DATASET*. HTTPS://WWW.KAGGLE.COM/DATASETS/SHUBHAMBATHWAL/FLIGHT-PRICE-PREDICTION

Thank You