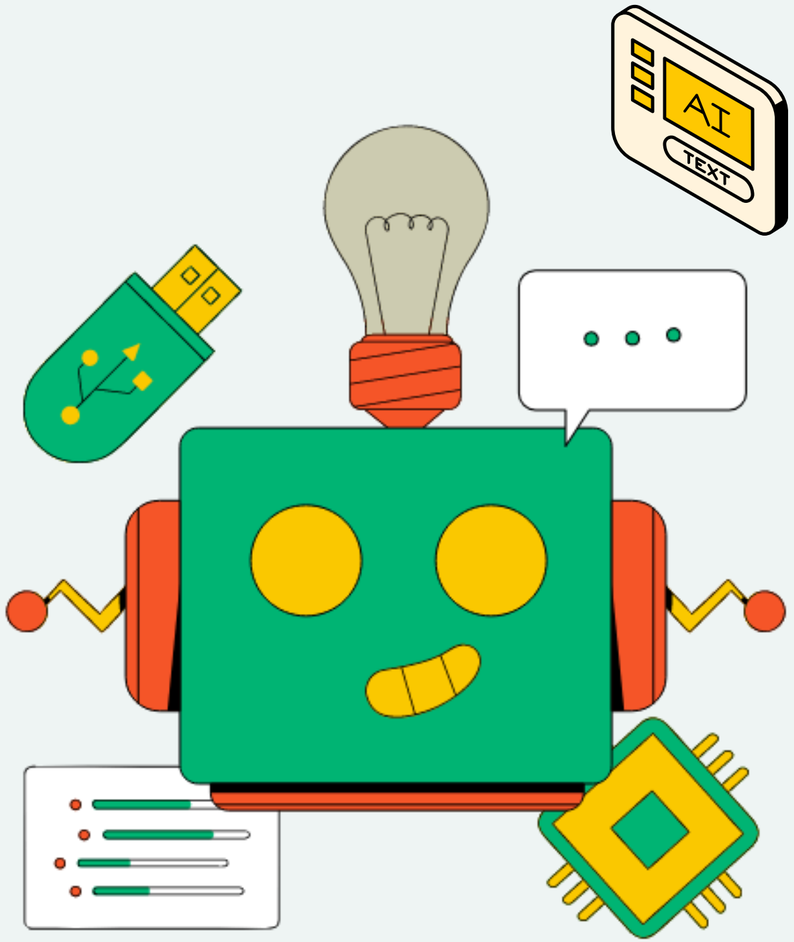
FINAL PROJECT
DATA SCIENTIST



FLIGHT PRICE PREDICTION (STUDY CASE : INDIAN DOMESTIC FLIGHT)

PRESENTATION

PRESENTED BY:

INDAH RESTUMI

ABOUT ME

Hi, I am Indah. An aspiring Data Scientist eager to apply the skills I have developed during my bootcamp journey. Passionate about exploring data, uncovering insights, and creating meaningful visualizations, I aim to support smarter business decisions. I am motivated to keep learning, grow professionally, and contribute to impactful, data-driven solutions as part of a professional team.

DIBIMBING ID - FULL STACK DATA SCIENCE BOOTCAMP

(MARCH 2025-PRESENT)

UNIVERSITAS GADJAH MADA - GEODETIC ENGINEERING

(AUG 2015-AUG 2019)

WORKING EXPERIENCE

DIREKTORAT JENDERAL BINAMARGA

GIS EXPERT ASSISTANT
(FEB 2023–NOV 2024)

- Ensuring roads data quality and integrity
- Maintained and updated LRS spatial data/

WIKA SERANG PANIMBANG TOLL ROAD

COMMERCIAL STAFF & WEBGIS COORDINATOR
(MAR 2021–NOV 2023)

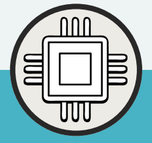
- Supported infrastructure project planning and execution through budget management, compliance review, and stakeholder reporting.

KEMENTERIAN AGRARIA DAN TATA RUANG
BADAN PERTANAHAN NASIONAL

GIS ANALYST
(JUNE 2020–JAN 2024)

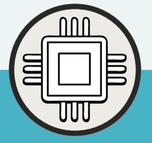
- Executed spatial data integration and land parcel validation to support national land administration improvements.

PROJECT OVERVIEW



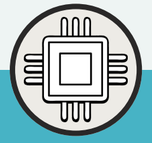
RETAIL

Customer Segmentation



PROPERTY

Boston Housing Price Predictions



TELCO

Telco Customer Churn

THE MAIN PROJECT

FLIGHT PRICE PREDICTION (INDIAN DOMESTIC FLIGHT)

PRESENTATION OUTLINE

- Business Understanding
- Data Understanding
- The Stage of Machine Learning
- Data Prepocessing-1
- Explanatory Data Analysis
- Data Preprocessing-2
- Model Building
- Model Evaluation
- Hyperparameter Tuning
- Conclusions & Recommendation

BUSINESS UNDERSTANDING

ISSUES

- Harga tiket pesawat sangat dinamis, dipengaruhi oleh jenis maskapai, kelas penerbangan, durasi, dan waktu pemesanan sehingga sulit untuk menetapkan harga yang tepat.

BUSINESS IMPACT

- Prediksi harga yang tidak tepat dan akurat dapat menyebabkan underpricing atau overpricing, sehingga memengaruhi tingkat keterisian kursi (load factor).
- Kondisi ini akan menurunkan pendapatan, mengurangi konversi transaksi, dan menurunkan kepuasan pelanggan.

OBJECTIVES

BUSINESS OBJECTIVES

- Mendukung penetapan harga dinamis (dynamic pricing) agar lebih kompetitif.
- Meningkatkan customer engagement dengan transparansi harga.
- Mengurangi risiko kehilangan pelanggan akibat harga yang terlalu tinggi.

PROJECT OBJECTIVES

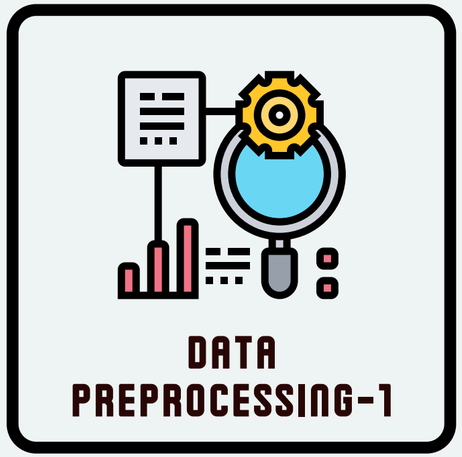
- Mengembangkan model prediksi harga tiket dengan akurasi tinggi.
- Mengidentifikasi faktor utama yang memengaruhi harga tiket pesawat.
- Menyediakan insight berbasis data untuk mendukung strategi dynamic pricing dan optimasi load factor.

DATA UNDERSTANDING

- Data dapat di download di Kaggle
- Dataset ini berisi informasi mengenai opsi pemesanan penerbangan dari sebuah Online Travel Agency di India.
- Dataset ini terdiri dari 300.261 baris dan 11 kolom.
- Dataset merepresentasikan karakteristik penerbangan, seperti maskapai, jumlah transit, durasi, hingga harga tiket

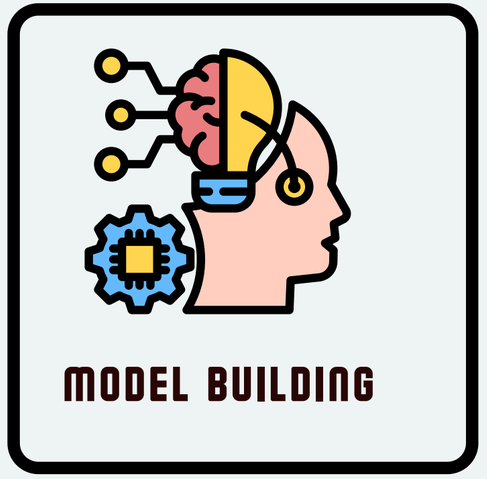
THE STAGE OF MACHINE LEARNING

DATA COLLECTION

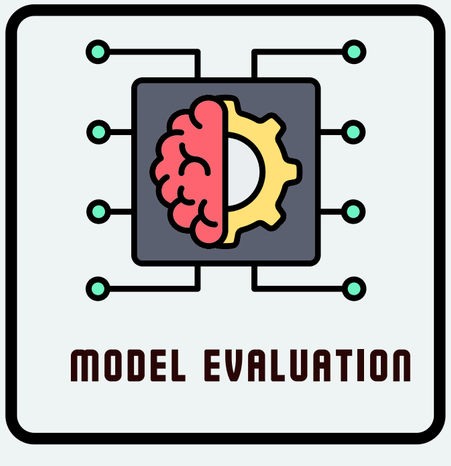


DATA PREPROCESSING-1

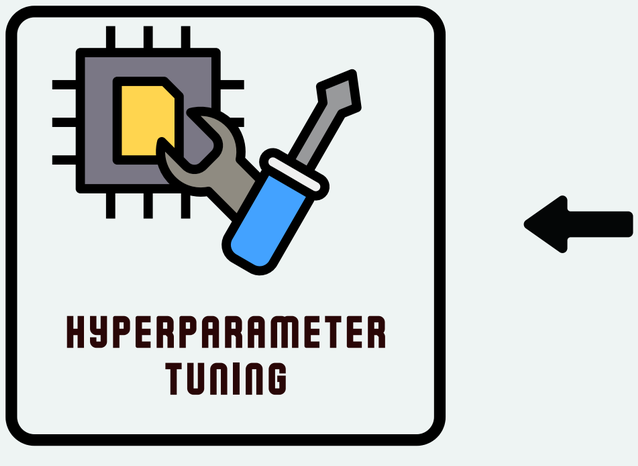
DATA PREPROCESSING-2



MODEL BUILDING



MODEL EVALUATION



HYPERPARAMETER TUNING

DATA PREPROCESSING - 1

DATA TYPE INSPECTION

CHECK MISSING VALUES

CHECK DUPLICATED VALUES

FEATURE ENGINEERING

0%

MISSING VALUES

0%

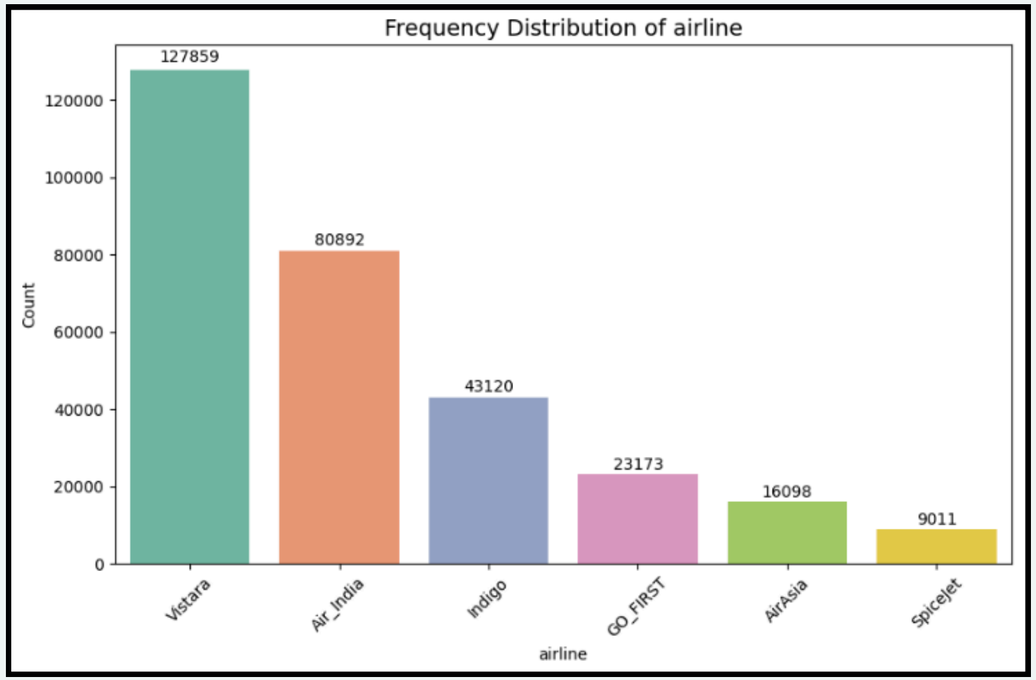
DUPLICATED ROW

0%

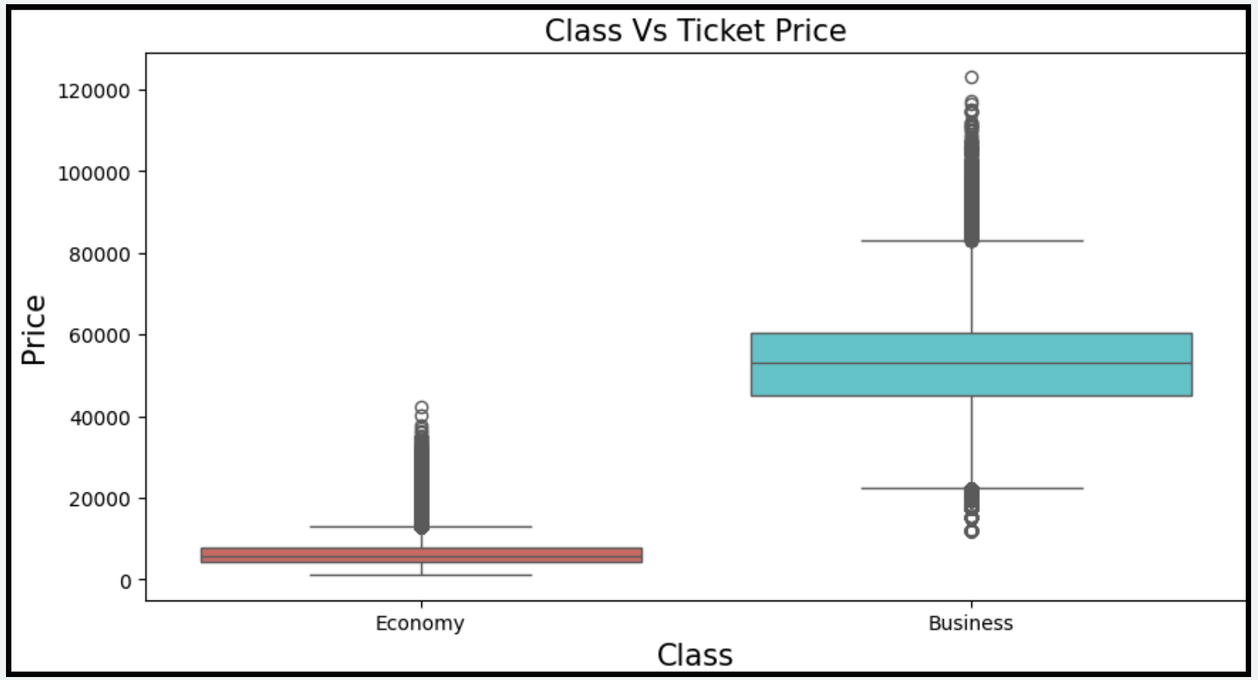
COLUMN ADDED

EXPLORATORY DATA ANALYSIS

AIRLINES DISTRIBUTION



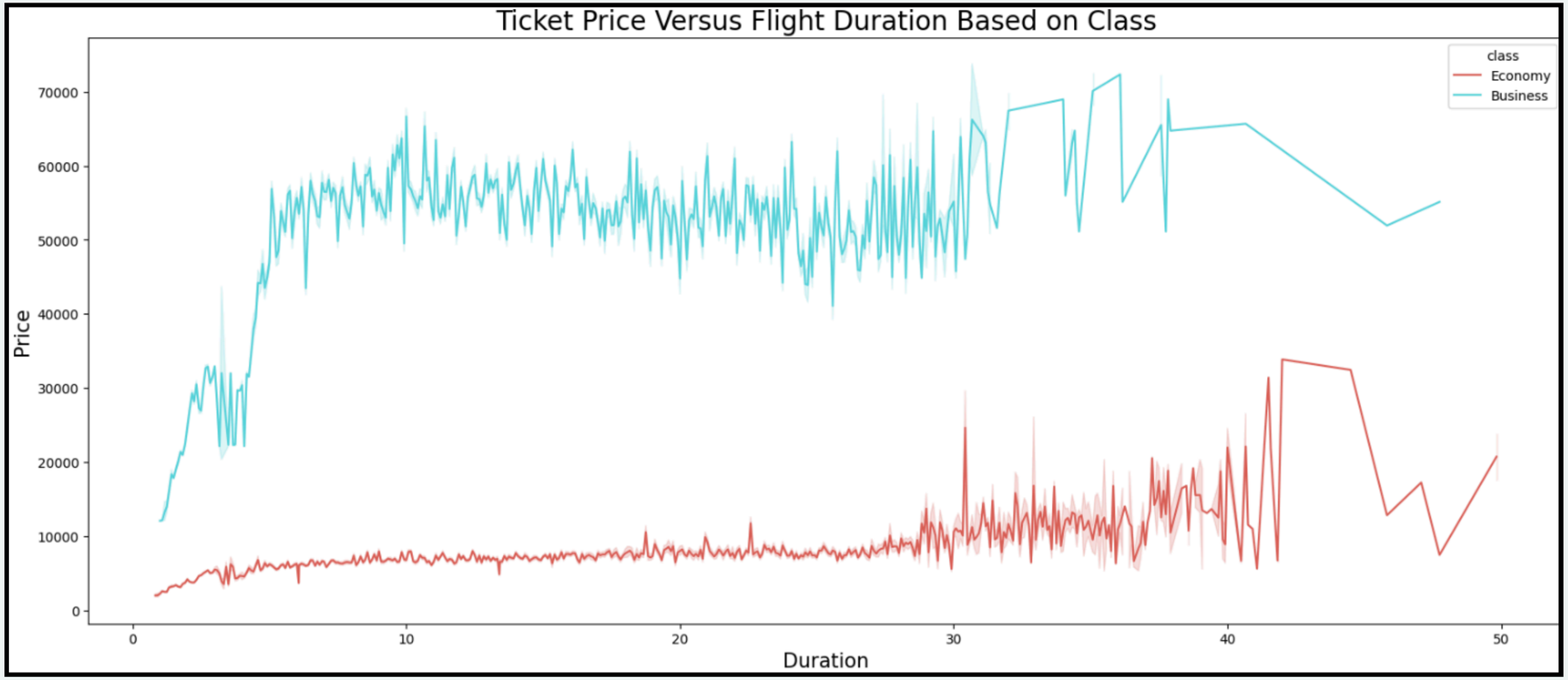
CLASS VS TICKET PRICE



- Terdapat 6 airlines dengan dua kategori yaitu **Premium Airlines (Vistara & Air India**) dan **On-Budget Airlines (Indigo, Go First, AirAsia, Spicejet).**
- Terdapat **dua kelas** penerbangan (**Ekonomi & Business**).
- Outlier disebabkan oleh **kategori & kelas penerbangan, waktu penerbangan (duration > 10 jam), pemesanan mendekati hari penerbangan.**

EXPLORATORY DATA ANALYSIS

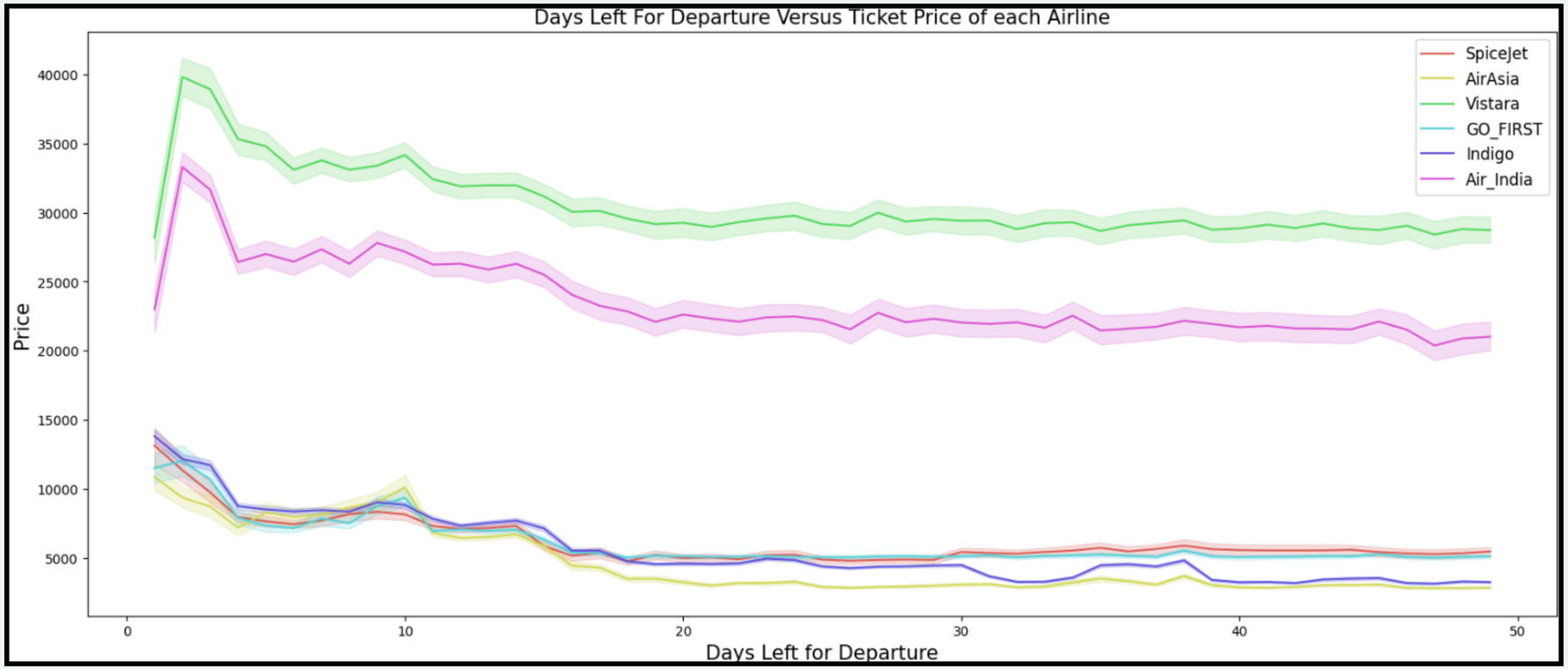
DURATION VS PRICE



- **Semakin lama durasinya** maka **harganya semakin tinggi** baik di kelas bisnis ataupun ekonomi

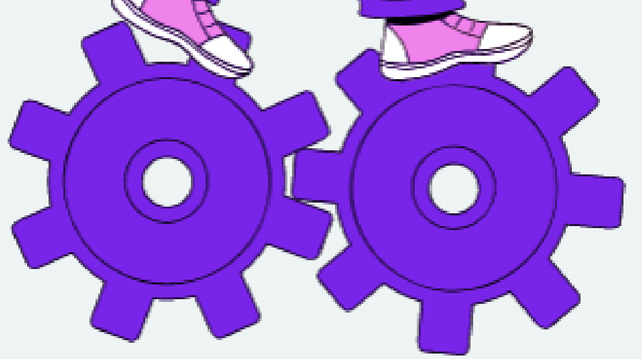
EXPLORATORY DATA ANALYSIS

DAYS LEFT VS PRICE



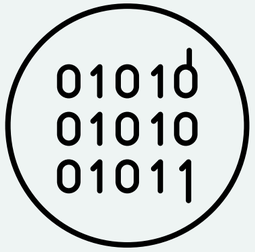
- Harga tiket **cenderung sangat tinggi saat mendekati tanggal keberangkatan.**
- Harga **semakin murah jika dipesan lebih awal.**
- Pola ini wajar karena maskapai menggunakan **dynamic pricing**: mendekati hari keberangkatan, kursi tersisa lebih sedikit → harga melonjak.

DATA PREPROCESSING - 2



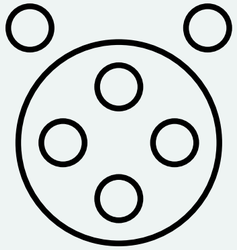
DATA SPLIT

- **Train data : 80%**
- **Test Data : 20%**
- Drop column yang tidak relevan



FEATURE ENCODING

- **Manual mapping** untuk class karena ini fitur ordinal. Economy : 0, Business : 1
- **One Hot Encoding** untuk **fitur kategorikal** seperti airline, kota, atau durasi tidak memiliki urutan alami



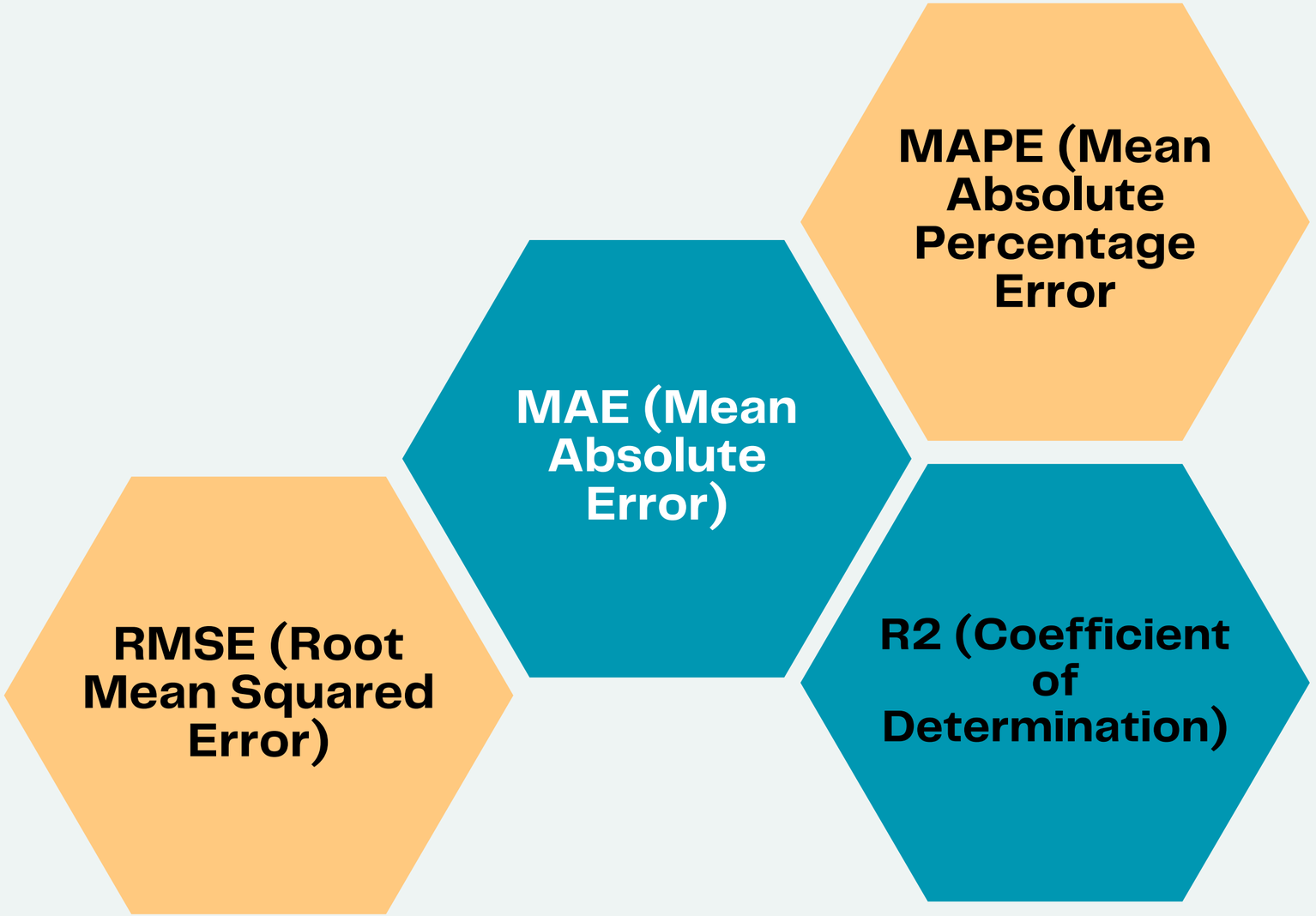
OUTLIER HANDLING

- Tidak dilakukan outlier handling pada project ini karena harga yang outlier masih masuk akal yang disebabkan oleh beberapa faktor.

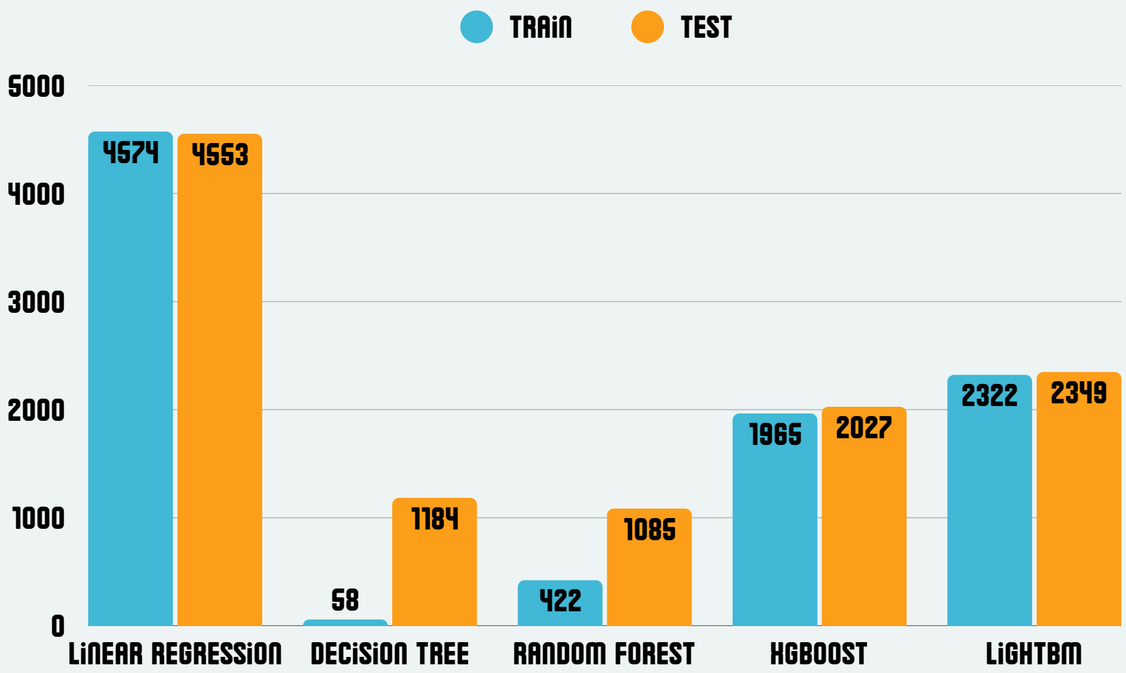
MODEL BUILDING

- Baseline Model
- Decision Tree
- Random Forest
- XGBoost
- LightBM

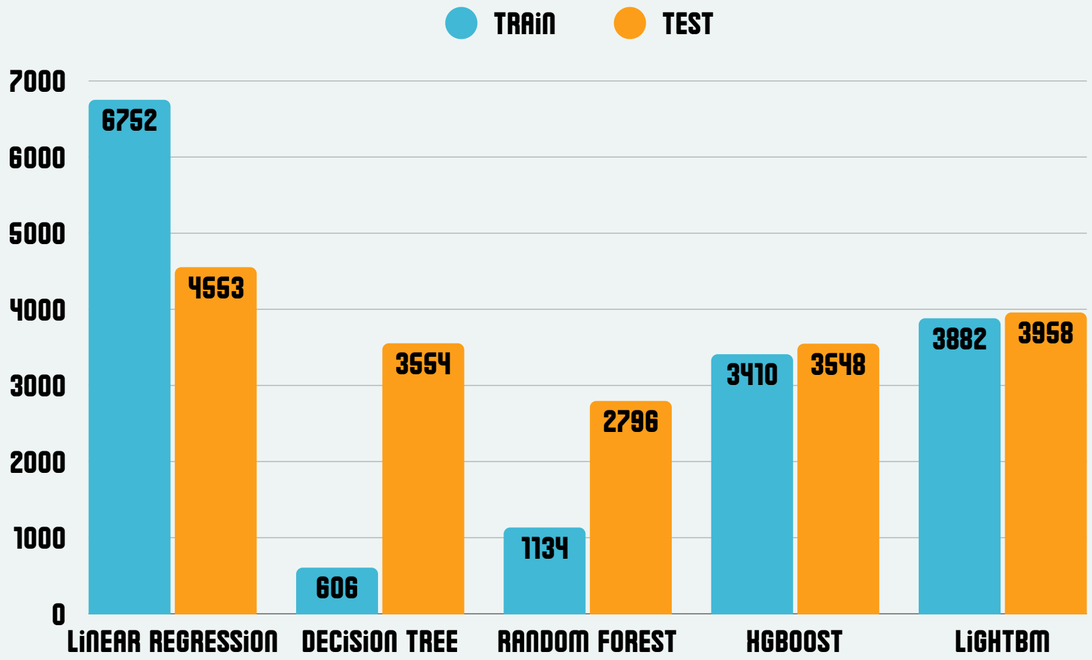
MODEL EVALUATION



MAE COMPARATION

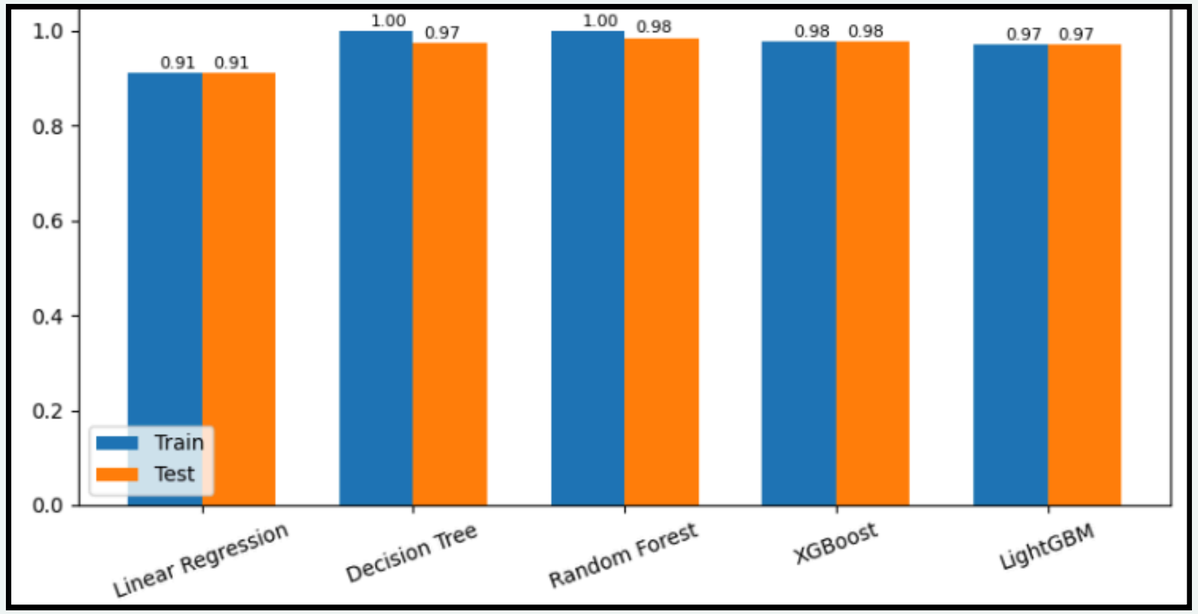


RMSE COMPARATION

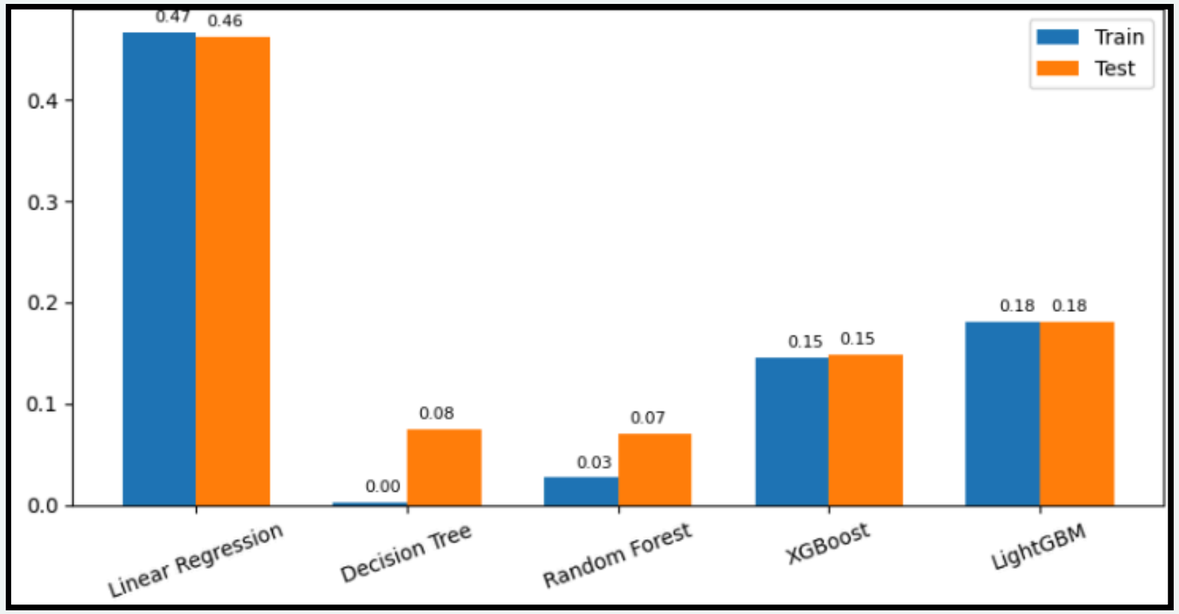


- Decision Tree memberikan **prediksi yang akurat di train namun error melonjak di test** sehingga bisa mengindikasikan **overfitting.**
- Linear Regression, XGBoost, LightGBM, MAE dan RMSE lebih konsisten, namun **error lebih tinggi dibanding Random Forest.**
- Random Forest dipilih sebagai model paling optimal, karena memberikan **prediksi dengan error rata-rata terkecil (MAE rendah) sekaligus menjaga kestabilan terhadap error besar (RMSE rendah)**.

R2 COMPARATION



MAPE COMPARATION



- Decision Tree menunjukkan **R^2 sangat tinggi di train** namun sedikit menurun di test, sementara **MAPE meningkat.**
- Linear Regression, XGBoost, dan LightGBM konsisten, namun **R^2 lebih rendah atau MAPE lebih tinggi dibanding Random Forest.**
- Random Forest dipilih sebagai model paling optimal, karena mampu **menjaga R^2 tetap tinggi sekaligus mempertahankan MAPE yang rendah dan stabil pada train maupun test.**

HYPERPARAMETER TUNING

BEFORE TUNING

Metric	Train	Test
MAE	421.72	1084.96
RMSE	1133.67	2796.19
R^2	0.9975	0.9848
MAPE	0.0272	0.0707

AFTER TUNING

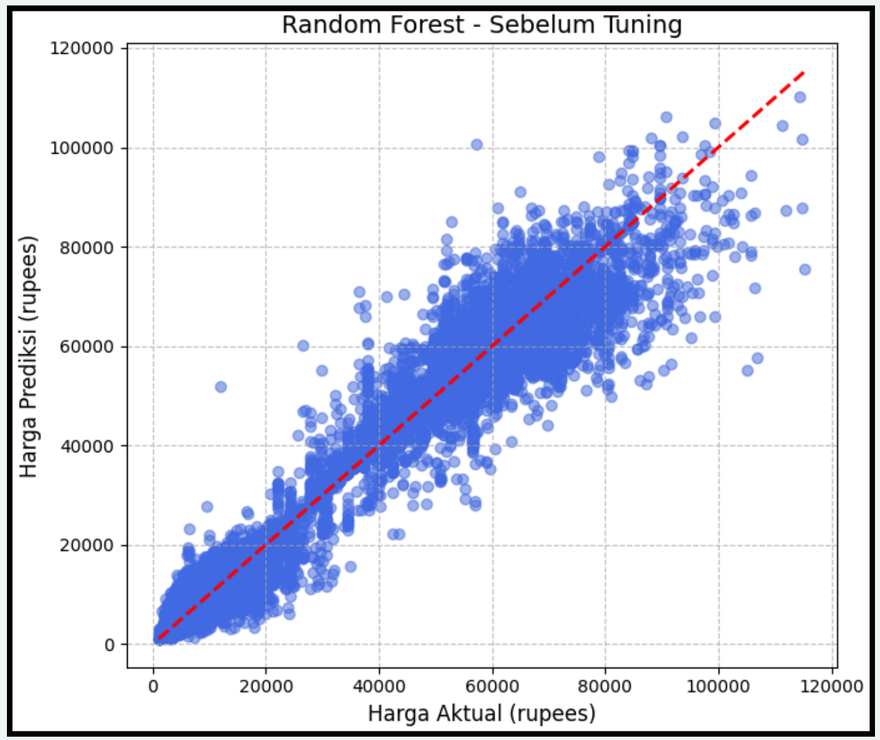
Metric	Train	Test
MAE	1099.5	1377.91
RMSE	2269.97	2837.71
R^2	0.9900	0.9844
MAPE	0.0791	0.0985

INDICATOR

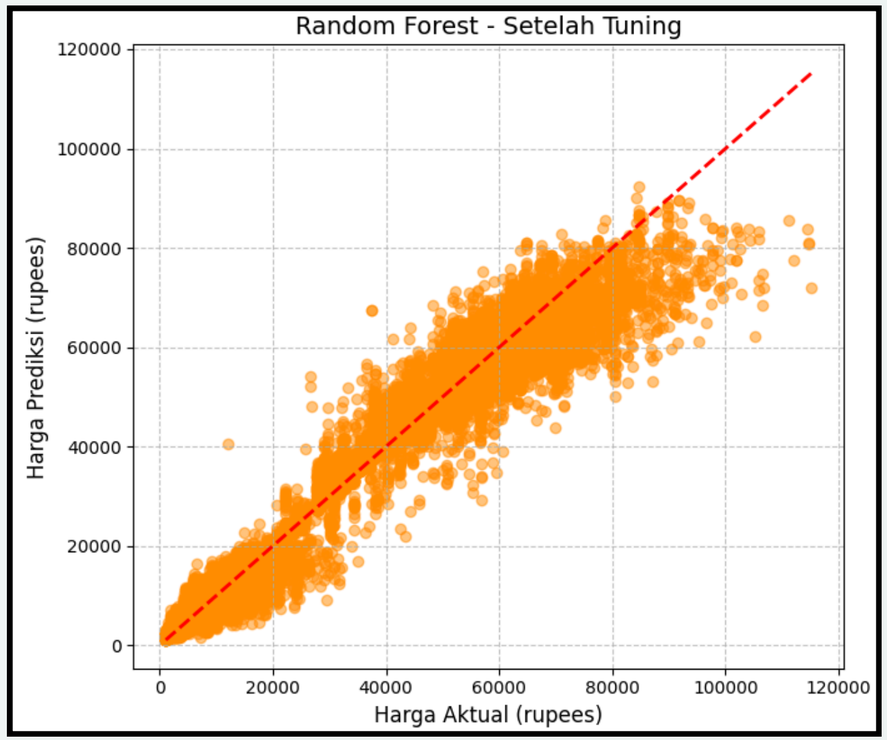
- Ruang parameter yang diuji sempit.
- Scoring hanya fokus ke MAE.
- Iterasi search terlalu sedikit.

HYPERPARAMETER TUNING

BEFORE TUNING

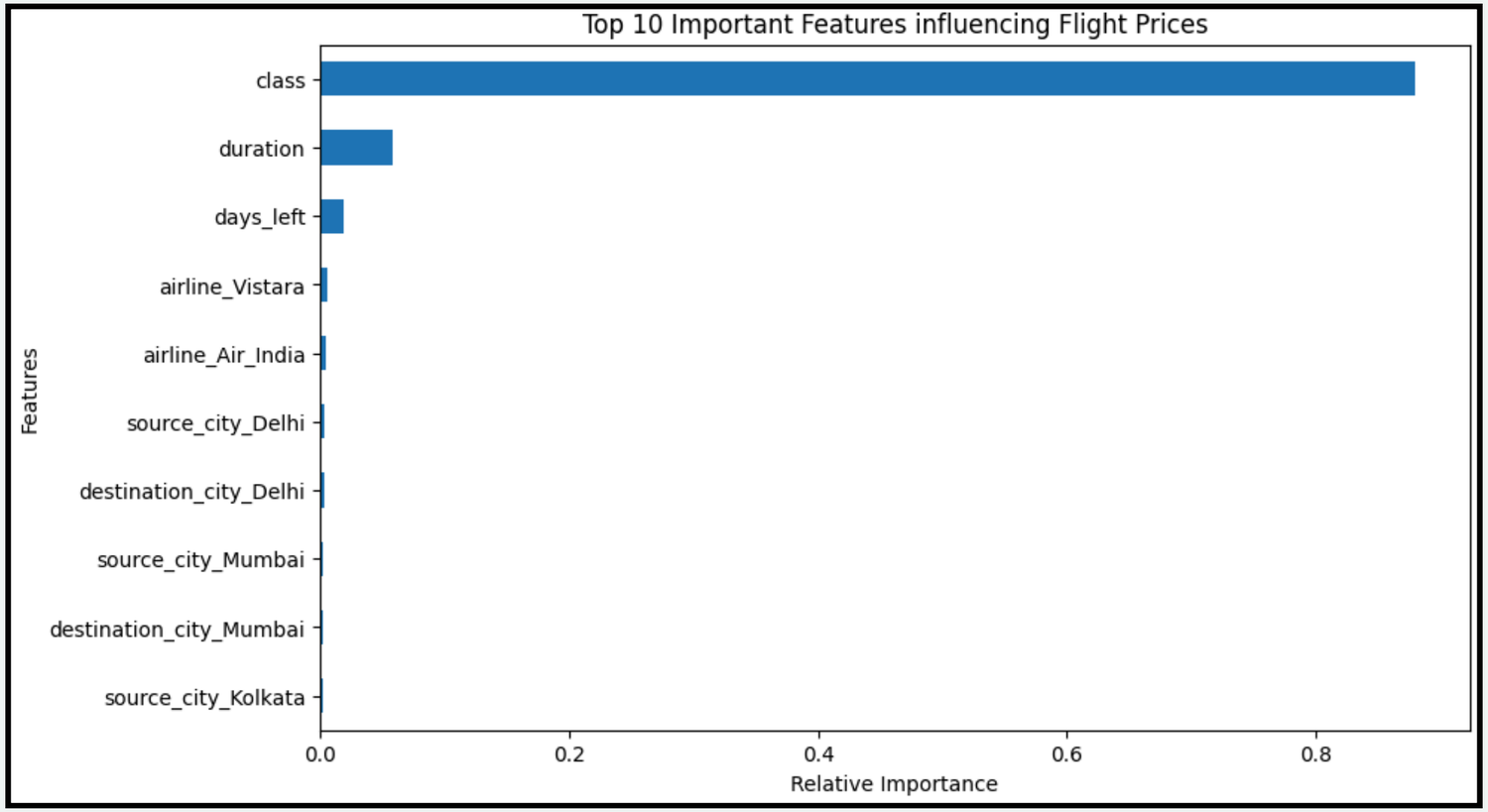


AFTER TUNING



- Mayoritas titik biru rapat di sekitar garis merah artinya **model cukup akurat dalam memprediksi harga tiket.**
- Model **sedikit kesulitan memprediksi tiket yang sangat mahal**, sehingga error lebih besar (wajar, karena RMSE lebih sensitif ke outlier tiket mahal).

FEATURE IMPORTANCE



- Faktor paling dominan memengaruhi harga tiket adalah **kelas penerbangan (class),** diikuti oleh **durasi perjalanan (duration)** dan **jarak waktu pembelian terhadap keberangkatan (days_left).**
- Faktor lain seperti maskapai (airline) dan kota asal/tujuan juga berpengaruh, tapi relatif kecil.

CONCLUSIONS

- Hasil analisis menunjukkan bahwa **Random Forest merupakan model terbaik** karena mampu memberikan prediksi harga tiket yang paling akurat, stabil, dan dengan tingkat kesalahan yang rendah.
- Faktor utama yang memengaruhi harga adalah **kelas penerbangan, durasi, dan waktu pembelian.**

RECOMMENDATIONS

- **Revenue Growth** : Menerapkan **dynamic pricing** berbasis **kelas, durasi, dan timing pembelian** untuk memaksimalkan pendapatan.

- **Operational Efficiency** : Mengoptimalkan **load factor** agar kursi kosong berkurang dan profitabilitas meningkat.

- **Customer Loyalty** : Menggunakan prediksi harga untuk personalisasi **promosi** sehingga **meningkatkan retensi dan value pelanggan.**

THANK YOU!!

Indah Restumi

Portofolio

Github

indahrestumi97@gmail.com