**🟨 Pemilihan Model & Estimator**

**Q: Kenapa memilih Random Forest sebagai model utama?**  
✅ *Karena Random Forest mampu menangani data dengan banyak fitur kategorikal dan numerik sekaligus. Model ini juga tangguh terhadap outlier dan tidak mudah overfitting, apalagi setelah dituning.*

**Q: Kenapa tidak pakai XGBoost, SVR, atau model lain?**  
✅ *Fokus proyek ini adalah interpretabilitas dan efisiensi waktu. Random Forest sudah memberi hasil sangat baik (R² > 0.98), sehingga menambah model lain tidak memberi peningkatan signifikan.*

**Q: Kenapa menggunakan n\_estimators = 50 saat awal modelling?**  
✅ *Sebagai default atau baseline awal, 50 estimator cukup untuk model ringan dan evaluasi cepat. Setelah itu, dilakukan tuning sampai 200 estimator untuk akurasi optimal.*

**🟦 Tuning dan Evaluasi**

**Q: Kenapa perlu GridSearchCV?**  
✅ *Agar model tidak hanya mengandalkan parameter default. GridSearch mencari kombinasi hyperparameter terbaik yang menghasilkan performa optimal.*

**Q: Apa indikator bahwa model tidak overfitting?**  
✅ *Nilai R² dan error pada data train dan test sangat mirip (misal: 0.998 vs 0.985), artinya model tidak hanya hafal data train, tapi bisa generalisasi ke data baru.*

**Q: Kenapa tidak pakai scoring ‘recall’ saat GridSearchCV?**  
✅ *Karena ini adalah masalah* ***regresi****, bukan klasifikasi. Jadi digunakan metrik seperti r2, neg\_mean\_absolute\_error, dll.*

**🟥 EDA dan Feature Importance**

**Q: Kenapa perlu visualisasi distribusi harga dan days left?**  
✅ *Agar bisa mengidentifikasi pola penting seperti distribusi harga bimodal (dua puncak) dan hubungan waktu pemesanan dengan harga.*

**Q: Apa fitur paling berpengaruh pada prediksi harga?**  
✅ *Berdasarkan feature importance dari Random Forest: Kelas Penerbangan, Days Left, Airline, dan Durasi adalah yang paling berpengaruh.*

**🟩 Strategi Pemisahan Model**

**Q: Kenapa dilakukan modelling terpisah untuk kelas Economy dan Business?**  
✅ *Karena distribusi harga antara dua kelas sangat berbeda (terlihat di histogram), dan memisahkan model dapat membantu menangkap karakteristik masing-masing dengan lebih akurat.*

**Q: Kenapa akhirnya memilih model gabungan sebagai model utama?**  
✅ *Meskipun model per kelas menarik, model gabungan memberikan hasil yang sangat baik dan stabil tanpa perlu memisahkan data. Ini lebih efisien dan fleksibel untuk implementasi.*

**🟧 Implementasi dan Bisnis**

**Q: Bagaimana model ini bisa digunakan oleh bisnis OTA atau maskapai?**  
✅ *Model ini bisa digunakan untuk prediksi harga otomatis, sistem notifikasi (alert harga), rekomendasi waktu beli terbaik, atau bahkan dinamisasi harga oleh maskapai.*

**Q: Apa rencana pengembangan ke depan?**  
✅ *Model bisa di-deploy ke web (misal Streamlit), dikembangkan ke dashboard dinamis, atau digabungkan dengan data eksternal seperti musim libur atau tren permintaan.*

**🔹 Tentang Dataset dan Business Understanding**

**Q: Kenapa memilih prediksi harga tiket sebagai proyek akhir?**  
✅ *Karena harga tiket sangat dinamis dan berdampak langsung pada bisnis OTA maupun konsumen. Prediksi harga dapat membantu pengambilan keputusan lebih baik.*

**Q: Apakah data yang digunakan real atau simulasi?**  
✅ *Data berasal dari dataset publik, merepresentasikan skenario dunia nyata dengan fitur seperti maskapai, waktu keberangkatan, kota asal, kelas, dan harga.*

**Q: Apa tujuan bisnis utama dari model ini?**  
✅ *Membantu konsumen membeli tiket pada waktu terbaik, serta membantu maskapai/OTA dalam menentukan strategi harga berdasarkan pola historis.*

**🔹 Tentang Data Preparation**

**Q: Bagaimana menghandle missing value atau outlier?**  
✅ *Dataset ini tidak memiliki missing value signifikan. Untuk outlier, model tree-based seperti Random Forest relatif robust, tapi tetap dilakukan validasi visual.*

**Q: Apakah dilakukan feature engineering?**  
✅ *Ya. Beberapa fitur seperti ‘days\_left’ dihitung dari tanggal keberangkatan. Selain itu, dilakukan encoding pada fitur kategorikal seperti maskapai dan kota.*

**🔹 Tentang Visualisasi dan EDA**

**Q: Kenapa distribusi harga ada dua puncak?**  
✅ *Karena perbedaan besar antara harga tiket Economy dan Business. Inilah alasan kami mempertimbangkan pemodelan terpisah.*

**Q: Apa insight terkuat dari EDA?**  
✅ *Harga naik tajam mendekati tanggal keberangkatan, harga Business sangat jauh di atas Economy, dan maskapai tertentu cenderung lebih mahal.*

**🔹 Tentang Model Evaluation & Validasi**

**Q: Kenapa tidak pakai cross-validation di evaluasi akhir?**  
✅ *GridSearchCV sudah melakukan cross-validation. Setelah itu, model terbaik di-evaluasi lagi di data test yang dipisahkan sejak awal.*

**Q: Bagaimana mengukur bahwa model ini siap digunakan?**  
✅ *Model diuji pada data test dengan performa konsisten, lalu disimpan ke .pkl, diuji pada Streamlit untuk memastikan prediksi dapat dilakukan secara live.*

**🔹 Tentang Deployment dan Pengembangan Lanjut**

**Q: Apakah model ini sudah siap di-deploy?**  
✅ *Ya. Model sudah disimpan dalam .pkl dan tersedia fungsi prediksi berbasis Streamlit untuk demo interaktif.*

**Q: Apa yang bisa dikembangkan lagi dari proyek ini?**  
✅ *Menambahkan variabel eksternal seperti cuaca, musim liburan, atau permintaan pasar. Juga bisa dikembangkan ke sistem rekomendasi harga terbaik.*

**Q: Apakah model ini bisa digunakan real-time?**  
✅ *Dengan pengolahan fitur yang sesuai dan integrasi ke sistem backend, model ini bisa digunakan sebagai engine prediksi real-time.*

**Tentang Proses**

**Q: Apa tantangan terbesar selama mengerjakan proyek ini?**  
✅ *Tantangan utama adalah mengelola distribusi harga yang bimodal, sehingga perlu eksplorasi model terpisah per kelas dan evaluasi train/test yang teliti.*

**Q: Apa pertimbangan dalam membagi train dan test?**  
✅ *Menggunakan stratified split tidak relevan di regresi, jadi pembagian dilakukan random dengan seed agar reproducible.*

**🔸 Tentang Validasi & Umum**

**Q: Apakah model overfitting? Bagaimana kamu tahu?**  
✅ *Tidak. Nilai R² dan error pada train dan test sangat dekat (0.998 vs 0.985), dan evaluasi dilakukan secara eksplisit untuk membandingkan performa.*

**Q: Apakah kamu menggunakan cross-validation saat evaluasi?**  
✅ *Ya, saat GridSearchCV dilakukan 3-fold CV. Itu membantu mengurangi risiko model terlalu cocok ke satu subset data.*

**Q: Apa baseline model kamu, dan kenapa penting?**  
✅ *Baseline adalah model mean (rata-rata harga di training set). Ini penting agar kita tahu apakah model machine learning benar-benar memberikan peningkatan.*

**🔸 Tentang Implementasi & Praktis**

**Q: Bagaimana cara menggunakan model ini di dunia nyata?**  
✅ *Model bisa di-deploy ke aplikasi seperti Streamlit untuk demo interaktif. Juga bisa diintegrasikan ke backend sebagai API untuk prediksi harga.*

**Q: Apa fitur minimum yang dibutuhkan agar model bisa memprediksi?**  
✅ *Fitur seperti: airline, source, destination, class, duration, dan days\_left. Semua ini sudah digunakan saat training dan tersimpan dalam file model\_features.pkl.*

**Q: Apakah model ini bisa digunakan untuk memprediksi harga di masa depan?**  
✅ *Ya, selama pola data masa depan masih serupa dan input fitur tersedia. Namun performa tetap perlu dimonitor secara berkala.*

**🔸 Tentang Kelemahan & Pengembangan**

**Q: Apa kelemahan dari model ini?**  
✅ *Model belum mempertimbangkan faktor musiman atau event besar (libur nasional, lebaran, dll). Juga belum menangani data streaming.*

**Q: Apa rencana pengembangan ke depan?**  
✅ *Integrasi ke dashboard, menambahkan data eksternal (cuaca, musim, dll), serta penggunaan model untuk alert sistem atau strategi pricing dinamis.*

### 🔸 ****Model & Algoritma****

**Q: Kenapa Random Forest bisa meng-handle fitur kategorikal setelah encoding?**  
✅ Karena Random Forest tidak membuat asumsi distribusi data. Setelah kategorikal diubah ke numerik (via one-hot encoding atau ordinal), model bisa memprosesnya secara langsung sebagai input split decision.

**Q: Kenapa tidak pakai Linear Regression sebagai model utama?**  
✅ Karena Linear Regression mengasumsikan hubungan linier antara fitur dan target, padahal di data ini terdapat hubungan non-linear (misalnya: harga naik tajam saat days\_left < 5). Random Forest lebih fleksibel.

**Q: Apakah Random Forest bisa overfitting? Bagaimana dicegah?**  
✅ Bisa, terutama jika *max\_depth* terlalu tinggi atau *min\_samples\_split* terlalu rendah. Makanya dilakukan GridSearchCV untuk men-tune parameter tersebut agar menjaga generalisasi model.

### 🔸 ****GridSearchCV dan Tuning****

**Q: Apa yang dimaksud neg\_mean\_absolute\_error dalam GridSearchCV?**  
✅ Karena *GridSearchCV* mengharapkan skor yang semakin tinggi semakin baik, maka sklearn menggunakan nilai negatif dari MAE untuk metrik evaluasi regresi agar tetap konsisten (makin kecil error, makin besar nilai negasinya).

**Q: Apakah kamu melakukan stratified sampling saat train-test split?**  
✅ Tidak, karena ini regresi. Stratified sampling hanya relevan untuk klasifikasi. Tapi kami memastikan data di-randomize dan reproducible via seed.

**Q: Kenapa memilih parameter seperti n\_estimators = [100, 200] dan max\_depth = [10, 20, None]?**  
✅ Dipilih berdasarkan pengalaman umum dengan Random Forest. Estimator lebih banyak umumnya meningkatkan akurasi, dan max\_depth digunakan untuk mengontrol kompleksitas model.

### 🔸 ****Evaluasi dan Error****

**Q: Kenapa pakai R², MAE, dan RMSE bersama-sama?**  
✅ \*Karena ketiganya saling melengkapi:

* **R²** → seberapa besar variasi dijelaskan model
* **MAE** → error rata-rata
* **RMSE** → penalti lebih besar untuk error besar\*

**Q: Apa arti R² negatif? Apakah mungkin?**  
✅ Ya, R² negatif artinya model lebih buruk dari baseline (mean prediksi). Ini bisa terjadi jika model gagal memahami pola data.

**Q: Residual plot kamu simetris, kenapa itu penting?**  
✅ Karena menunjukkan error tersebar merata → model tidak bias ke over atau under prediction. Jika miring, bisa jadi model punya bias sistematis.

### 🔸 ****Feature Importance dan Interpretasi****

**Q: Bagaimana Random Forest menentukan feature importance?**  
✅ Dihitung dari seberapa besar penurunan impurity (MSE) saat fitur digunakan dalam split. Semakin sering dan semakin besar gain-nya, semakin penting fitur tersebut.

**Q: Apakah feature importance bisa bias terhadap fitur kategorikal dengan banyak nilai?**  
✅ Ya. Fitur hasil one-hot encoding (misal 'airline\_A', 'airline\_B', dst) bisa dianggap lebih penting karena terdiri dari banyak kolom. Ini bisa dimitigasi dengan memeriksa grouping importance atau model lain seperti permutation importance.

### 🔸 ****Deployment & Produksi****

**Q: Bagaimana kamu pastikan input prediksi sesuai dengan data saat training?**  
✅ Disimpan daftar fitur dalam *model\_features.pkl* agar setiap input prediksi bisa dicek kesesuaiannya (urutan, jumlah, nama kolom).

**Q: Kenapa .pkl kamu besar? Bisa diperkecil?**  
✅ Karena Random Forest menyimpan semua pohon dan struktur split-nya. Bisa dikompresi dengan *joblib*, *gzip*, atau kurangi *n\_estimators*.