**🎯 Langkah Awal: Kenali Masalahnya**

| **Pertanyaan** | **Jawaban** | **Model yang Cocok** |
| --- | --- | --- |
| **Apakah target (label) datamu numerik?** | Ya | **Regresi** |
|  | Tidak (kategori) | **Klasifikasi** |
| **Apakah kamu ingin memahami hubungan antar fitur?** | Ya | **Model interpretable (misalnya regresi)** |
|  | Tidak masalah interpretasi | **Model kompleks seperti Random Forest atau XGBoost** |
| **Apakah kamu punya banyak data & waktu training tidak masalah?** | Ya | **Model kompleks bisa digunakan** |
|  | Tidak (sedikit data atau waktu terbatas) | **Model ringan (regresi, decision tree)** |

**📦 Jenis-Jenis Model Populer & Kapan Dipakai**

**1. Linear Regression (Regresi Linier)**

* **Gunakan jika:** Target numerik dan kamu ingin model yang sederhana & mudah dijelaskan.
* **Contoh kasus:** Prediksi harga rumah, prediksi pendapatan.
* **Syarat:** Hubungan antar fitur dengan target **linier**, tidak banyak outlier.

**2. Logistic Regression (Regresi Logistik)**

* **Gunakan jika:** Target berupa **kategori biner (0/1)**.
* **Contoh kasus:** Apakah pelanggan akan churn (keluar) atau tidak?
* **Kelebihan:** Mudah dijelaskan (interpretable), ringan, cepat dilatih.

**3. Decision Tree**

* **Gunakan jika:** Ingin model sederhana tapi lebih fleksibel dari regresi.
* **Contoh kasus:** Klasifikasi email spam atau bukan.
* **Kelebihan:** Tidak butuh scaling data, bisa menangani data kategori dan numerik.

**4. Random Forest**

* **Gunakan jika:** Ingin akurasi lebih tinggi dari decision tree.
* **Contoh kasus:** Prediksi penyakit, prediksi kredit macet.
* **Kelebihan:** Kuat terhadap overfitting, akurat.
* **Kekurangan:** Kurang interpretatif.

**5. Gradient Boosting (XGBoost, LightGBM)**

* **Gunakan jika:** Ingin performa terbaik dan rela trade-off dengan waktu training.
* **Contoh kasus:** Kompetisi data science (Kaggle), prediksi click, penipuan, dsb.

**6. KNN (K-Nearest Neighbors)**

* **Gunakan jika:** Data kecil dan sederhana.
* **Kelebihan:** Mudah dipahami.
* **Kekurangan:** Lambat untuk data besar, sensitif terhadap skala data.

**7. SVM (Support Vector Machine)**

* **Gunakan jika:** Data kamu tidak linear dan punya margin pemisahan yang jelas.
* **Cocok untuk:** Klasifikasi dengan data bersih dan tidak terlalu besar.

**🧠 Tips Memilih Model:**

| **Tujuan** | **Model Awal yang Direkomendasikan** |
| --- | --- |
| Ingin baseline cepat | Logistic Regression / Linear Regression |
| Fokus akurasi | Random Forest / XGBoost |
| Fokus interpretasi | Regresi / Decision Tree |
| Banyak fitur | Regularisasi (Lasso, Ridge), Random Forest |
| Data tidak seimbang | Gunakan model + balancing method (misal: SMOTE) |
| Data besar dan kompleks | XGBoost / LightGBM |

**📋 Checklist Sebelum Modeling:**

1. ✅ Pahami tipe masalah (klasifikasi atau regresi)
2. ✅ Lihat distribusi data & missing values
3. ✅ Lakukan EDA (Exploratory Data Analysis)
4. ✅ Normalisasi (jika pakai model sensitif skala, seperti KNN, SVM)
5. ✅ Gunakan **train-test split** atau **cross-validation**
6. ✅ Bandingkan beberapa model dengan metrik yang sesuai

## ****Rangkuman Insight dari EDA yang Relevan untuk Modeling****

| **Insight dari EDA** | **Relevansi Modeling** | **Aksi** |
| --- | --- | --- |
| Distribusi harga bimodal (ekonomi & bisnis) | Data punya 2 segmen | ➤ Lakukan 2 versi modeling: gabungan & per kelas |
| Harga naik saat mendekati hari H (days\_left) | Fitur prediktif kuat | ➤ days\_left wajib digunakan |
| Harga kelas bisnis jauh lebih tinggi | Perlu perlakuan khusus | ➤ Gunakan class sebagai fitur utama atau pecah model |
| Durasi memengaruhi harga | Korelasi moderat (+) | ➤ Gunakan duration sebagai fitur numerik |
| Maskapai dan rute memengaruhi harga | Efek signifikan | ➤ Gunakan one-hot encoding maskapai dan kota |
| Korelasi antar fitur aman (tidak tinggi) | Tidak perlu drop kolom | ➤ Semua fitur numerik tetap digunakan |

**🚀 Langkah Selanjutnya Berdasarkan Evaluasi**

| **Tujuan** | **Aksi** |
| --- | --- |
| Ingin model prediksi praktis & cepat | Gunakan **Linear Regression** dengan fitur utama (days\_left, class, dll) |
| Ingin model prediksi akurat dan kompleks | Gunakan **Random Forest / XGBoost** |
| Ingin segmentasi prediksi ekonomi vs bisnis | ➤ Lanjut ke **modeling per kelas** (Langkah berikutnya!) |
| Ingin tahu fitur paling berpengaruh | ➤ Tambahkan analisis **feature importance** dari RandomForest |

**🧠 Rekomendasi**

* Model Random Forest atau XGBoost dapat digunakan sebagai **model final untuk deployment**
* Tapi, **Linear Regression sudah cukup kuat** dan cepat jika digunakan untuk fitur ringan seperti rekomendasi harga real-time di aplikasi OTA