RESEARCH REQUIREMENT

Ringkasan Proyek

1. **Nama Proyek**

   Telco Product Recommendation (Prototype / Capstone)

2. **Tujuan Utama**

   Mengembangkan sistem rekomendasi produk telekomunikasi yang mampu memprediksi kategori penawaran terbaik (target_offer) untuk setiap pelanggan berdasarkan perilaku mereka.

3. **Output yang Diharapkan:**
   - Model klasifikasi multi-kelas untuk prediksi kategori penawaran produk.
   - API yang dapat digunakan untuk melakukan inference menggunakan model tersebut.
   - Prototipe aplikasi web dengan antarmuka pengguna (UI) yang dapat dimanfaatkan untuk keperluan marketing dan demonstrasi.

Objective (Tujuan Riset)

1. Mengevaluasi kelayakan data perilaku pelanggan (customer behavior) dalam memprediksi target_offer dengan kategori multi-kelas: General Offer, Data Booster, Device Upgrade Offer, dan Top-up Promo.
2. Mendesain pipeline data yang reproducible mencakup tahap pembersihan data (cleaning), rekayasa fitur (feature engineering), pemodelan, dan inference untuk deployment prototipe.
3. Menganalisis metrik performa model
4. Mendokumentasikan batasan yang ada antara simulasi dan produksi serta kebutuhan integrasi jika teknologi ini diadopsi dalam lingkungan operasional perusahaan telekomunikasi sebenarnya.

Ruang Lingkup (Scope)

Scope yang termasuk (MVP):

1. Data ingestion dari file CSV (data_capstone.csv) dan/atau upload manual.
2. Exploratory Data Analysis (EDA), pembersihan data, dan feature engineering dengan pipeline yang reproducible.

3. Training beberapa model baseline seperti Logistic Regression, Random Forest, dan XGBoost/LightGBM.
4. Evaluasi model menggunakan teknik cross-validation, confusion matrix, dan metrik per kelas.
5. API inference yang dikembangkan dengan Flask atau FastAPI untuk prediksi single maupun batch.
6. Prototype frontend berbasis React atau Vue yang dapat digunakan untuk demo, upload file CSV, menampilkan rekomendasi, dan simulasi proses checkout.

Scope yang dikecualikan (eksplisit):

1. Integrasi real-time dengan sistem billing, network monitoring, atau CRM telco.
2. Payment gateway dan provisioning produk nyata.
3. Production hardening seperti autoscaling dan advanced monitoring; fokus pada proof-of-concept.

Dataset & Data Dictionary (Awal)

1. Lokasi Data Sample: data/raw/data_capstone.csv
2. Target: target_offer
3. Contoh Fitur Utama
    - customer_id (ID pelanggan)
    - plan_type (Prepaid / Postpaid)
    - device_brand
    - avg_data_usage_gb
    - pct_video_usage
    - avg_call_duration
    - sms_freq
    - monthly_spend
    - topup_freq
    - travel_score
    - complaint_count

Research Questions (Pertanyaan Riset)

1. Seberapa baik model machine learning (multi-class) dapat memprediksi target_offer berdasarkan fitur perilaku pelanggan yang tersedia?
2. Fitur mana yang paling signifikan berkontribusi terhadap keputusan rekomendasi produk (melalui analisis feature importance atau metode interpretabilitas seperti SHAP)?
3. Bagaimana pengaruh ketidakseimbangan kelas (class imbalance) terhadap performa setiap kelas, dan strategi apa yang paling efektif untuk mengatasinya (misalnya penyesuaian class weight, oversampling seperti SMOTE, atau undersampling)?
4. Apakah pipeline feature engineering yang diusulkan dapat memberikan peningkatan performa dibandingkan baseline sederhana?

Metodologi

1. Exploratory Data Analysis (EDA)

   Inspeksi nilai hilang, duplikasi, distribusi variabel kategorikal dan numerikal, korelasi antar fitur, serta identifikasi outlier menggunakan metode IQR.
2. Data Cleaning

   Menangani missing values dengan imputasi atau penghapusan data, menghapus duplikasi, dan memperbaiki nilai negatif (contoh: avg_call_duration yang negatif diimputasi atau di-set ke 0).
3. Feature Engineering

   Membuat fitur domain-driven seperti data_per_1k_spend, comm_activity_score, customer_value_score, serta pembuatan kategori binned seperti data_usage_category, spend_category.
4. Encoding & Scaling

   Melakukan one-hot encoding atau target encoding untuk fitur kategorikal dan label encoding untuk target variabel, serta scaling (StandardScaler) untuk algoritma yang membutuhkannya.
5. Splitting & Sampling

   Melakukan stratified train-test split dengan perbandingan 80/20, dan jika terjadi ketidakseimbangan kelas lebih dari 2:1, diterapkan teknik SMOTE, ADASYN, atau penyesuaian class weights.

6. Model Training

   Melatih berbagai model baseline dan kompleks: Logistic Regression multinomial, Decision Tree, Random Forest, XGBoost atau LightGBM; opsional KNN dan multilayer perceptron sederhana jika dataset besar.

7. Hyperparameter Tuning

   Menggunakan GridSearchCV, RandomizedSearchCV, atau Optuna untuk optimasi hyperparameter.

8. Model Evaluation

   Menggunakan metrik evaluasi seperti Accuracy, Precision, Recall, F1-score per kelas, Confusion Matrix, macro dan micro F1, juga ROC-AUC per kelas jika relevan; menggunakan cross-validation Stratified K-Fold (k=5).

9. Interpretation & Business Validation

   Menggunakan analisis feature importance dan SHAP values untuk interpretasi model dan validation bisnis melalui studi kasus rekomendasi spesifik pada pelanggan.

10. Packaging & Inference

    Melakukan serialisasi model dan pipeline preprocessing (dengan joblib, pickle, atau ONNX) serta expose endpoint API untuk inference.

Risiko & Mitigasi

1. Risiko: Ketidakseimbangan data (data imbalance) yang besar

   Mitigasi: Menggunakan teknik sampling (oversampling/undersampling), penyesuaian class weight, dan upaya pengumpulan data sample lebih banyak

2. Risiko: Banyaknya fitur kategorikal dengan cardinality tinggi (misal: device_brand)

   Mitigasi: Mengelompokkan kategori ke dalam top-k brands dan menggunakan teknik encoding seperti target encoding atau frequency encoding.

3. Risiko: Overfitting pada model yang kompleks

   Mitigasi: Melakukan validasi silang (cross-validation), menerapkan regularisasi, dan early stopping pada model boosting.