BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek penelitian ini berfokus pada proses prioritisasi kunjungan dan rekomendasi channel follow-up yang dijalankan oleh tim sales PT XYZ. Dalam praktiknya, penentuan eksekusi prospek—melalui kunjungan langsung (visit) maupun telepon (phone)—masih sangat dipengaruhi pertimbangan manual dan subjektif sehingga berpotensi menurunkan efisiensi kunjungan serta menimbulkan peluang konversi yang terlewat. Untuk menjawab permasalahan tersebut, penelitian ini mengembangkan pendekatan two-stage prediction yang terdiri atas Stage-1 untuk memprediksi probabilitas booking sebagai dasar penetapan prioritas prospek, dan Stage-2 untuk merekomendasikan channel follow-up optimal (visit/phone) pada prospek yang telah diprioritaskan.

Data penelitian bersumber dari sistem CRM internal yang digunakan tim sales untuk penawaran dan pencatatan interaksi pelanggan, mencakup periode Q2 2025 (April–Juni 2025) dengan cakupan nasional dan ukuran data 579.081 baris serta 31 kolom. Kualitas data secara umum baik; kolom dengan nilai hilang di atas 1% hanya last_type_fu dan last_type_fu_std (±7,28%), sedangkan kolom lainnya memiliki kelengkapan mendekati penuh. Seluruh identitas sensitif diperlakukan secara anonim sesuai kebijakan perusahaan. Target pada Stage-1 adalah status booking (is_book), dengan distribusi kelas NO_BOOKING 82,98% dan BOOKING 17,02% (ketidakseimbangan sekitar 4,88:1). Target pada Stage-2 adalah last_type_fu yang merepresentasikan aksi aktual follow-up dan dinormalisasi pada last_type_fu_std (visit=1, phone=0). Statistik awal menunjukkan perbedaan conversion rate antar kanal, yakni sekitar 14,59% untuk phone dan 16,68% untuk visit, sehingga keputusan kanal yang tepat menjadi krusial setelah prospek dipilih untuk dieksekusi.

Pemodelan memanfaatkan variabel demografi dan sosio-ekonomi (misalnya usia, jenis kelamin, status perkawinan, tingkat pendidikan, tanggungan, status rumah), lokasi (provinsi, kecamatan, kelurahan), karakteristik produk atau objek

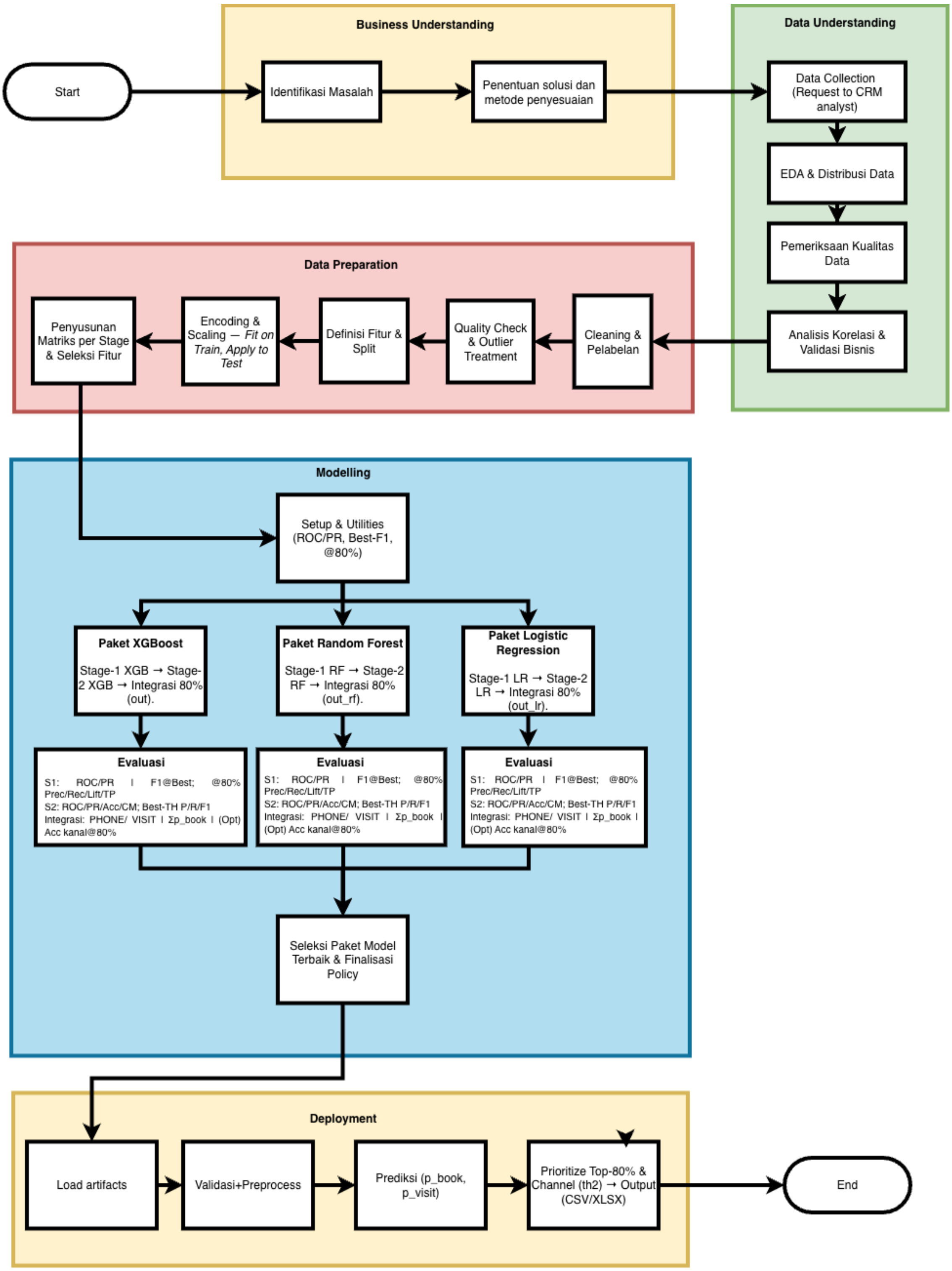
pembiayaan (merek dan tahun objek, plafon, tenor, angsuran bulanan, uang muka kotor), serta konteks dan riwayat CRM (periode distribusi awal, unit bisnis, serta penanda interaksi). Seluruh tahapan transformasi—termasuk encoding dan scaling—dilaksanakan dengan prinsip *fit on train, apply to test* untuk mencegah kebocoran data. Mengingat keterbatasan kapasitas lapangan, kebijakan perusahaan menargetkan eksekusi sekitar 80% prospek setiap siklus; karena itu evaluasi tidak hanya meliputi metrik umum seperti accuracy, precision, recall, F1-score, ROC-AUC, dan PR-AUC, tetapi juga Precision@80, Recall@80, dan Lift@80 guna memastikan kesesuaian kinerja model terhadap kapasitas eksekusi riil. Tiga algoritma—Logistic Regression, Random Forest, dan XGBoost—dibandingkan pada masing-masing stage untuk memperoleh kombinasi model terbaik.

Keluaran operasional sistem berupa daftar prioritas prospek beserta rekomendasi channel yang dihasilkan melalui dashboard berbasis Streamlit. Dashboard tersebut memungkinkan pemrosesan batch dan pengunduhan hasil dalam format CSV/XLSX, sementara distribusi lebih lanjut ke aplikasi yang digunakan tim sales dikelola oleh analis CRM sebagai bagian dari proses operasional internal PT XYZ**.**

3.2 Metode Penelitian

3.2.1 Alur Penelitian

Penelitian ini menggunakan alur kerja yang didasarkan pada kerangka kerja CRISP-DM (Cross-Industry Standard Process for Data Mining). Alur penelitian terdiri dari enam tahapan utama: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment, seperti yang ditunjukkan pada Gambar 3.1. Setiap tahapan memiliki peran penting dalam memastikan model *machine learning* yang dihasilkan dapat memenuhi tujuan penelitian secara optimal.



Gambar 3. 1 Diagram Alur Penelitian

Gambar 3.1 menampilkan diagram alur penelitian yang merangkum proses end-to-end pengembangan sistem two-stage untuk prioritisasi lead dan rekomendasi channel follow-up. Berikut penjelasan setiap tahapan.

Tahap Business Understanding merupakan tahap awal yang terdiri dari dua aktivitas utama yaitu Identifikasi Masalah dan Penentuan Solusi dan Metode Penyesuaian. Identifikasi Masalah bertujuan untuk menganalisis permasalahan efektivitas proses follow-up lead yang masih dilakukan secara manual dan subjektif. Penentuan Solusi dan Metode Penyesuaian bertujuan untuk mendefinisikan sasaran pengembangan sistem prioritas lead berbasis machine learning beserta metrik keberhasilan yang relevan.

Tahap Data Understanding mencakup empat aktivitas yaitu Data Collection, EDA & Distribusi Data, Pemeriksaan Kualitas Data, dan Analisis Korelasi & Validasi Bisnis. Data Collection dilakukan melalui permintaan resmi kepada CRM analyst untuk memperoleh data historis follow-up. EDA & Distribusi Data bertujuan untuk mengeksplorasi karakteristik dan sebaran data. Pemeriksaan Kualitas Data meliputi pengecekan kelengkapan, duplikasi, dan anomali. Analisis Korelasi & Validasi Bisnis dilakukan untuk memastikan variabel yang dipilih relevan dengan konteks bisnis.

Tahap Data Preparation dilakukan secara bertahap yang mencakup lima aktivitas yaitu Cleaning & Pelabelan, Quality Check & Outlier Treatment, Definisi Fitur & Split, Encoding & Scaling, dan Penyusunan Matriks per Stage & Seleksi Fitur. Cleaning & Pelabelan bertujuan untuk membersihkan data dan menentukan label target. Quality Check & Outlier Treatment bertujuan untuk mendeteksi dan menangani nilai ekstrem. Definisi Fitur & Split bertujuan untuk menentukan variabel dan membagi data training-testing. Encoding & Scaling dilakukan dengan prinsip fit on train, apply to test. Penyusunan Matriks per Stage & Seleksi Fitur bertujuan untuk menyiapkan dataset final masing-masing tahap pemodelan.

Tahap Modelling diawali dengan Setup & Utilities yang mencakup penyiapan fungsi evaluasi kurva ROC/PR, penentuan threshold berdasarkan F1 terbaik (Best-

F1), dan pengukuran kinerja pada kapasitas 80% (@80%). Selanjutnya dibangun tiga paket model yaitu Paket XGBoost, Paket Random Forest, dan Paket Logistic Regression. Setiap paket menjalankan Stage-1 (prediksi probabilitas booking) dan Stage-2 (rekomendasi channel) secara berurutan, kemudian diintegrasikan dengan kebijakan 80% untuk menghasilkan output masing-masing.

Setiap paket model kemudian masuk ke tahap Evaluasi dengan komponen sebagai berikut. Komponen S1 merujuk pada evaluasi Stage 1 yang meliputi ROC/PR curve, F1@Best (threshold optimal berdasarkan F1 tertinggi), serta metrik @80% mencakup Precision, Recall, Lift, dan True Positive pada kapasitas 80%. Komponen S2 merujuk pada evaluasi Stage 2 yang mencakup ROC-AUC, PR-AUC, Accuracy, Confusion Matrix, serta optimasi threshold berdasarkan Precision, Recall, dan F1 (Best-TH P/R/F1). Komponen Integrasi menunjukkan evaluasi sistem two-stage secara keseluruhan untuk rekomendasi channel PHONE/VISIT, agregasi probabilitas booking (Σp_book), dan akurasi kanal pada top-80% lead. Hasil evaluasi ketiga paket dibandingkan pada tahap Seleksi Paket Model Terbaik & Finalisasi Policy untuk menentukan model final dan threshold operasional.

Tahap Deployment terdiri dari empat langkah sekuensial yaitu Load Artifacts, Validasi+Preprocess, Prediksi, dan Prioritize Top-80% & Channel. Load Artifacts bertujuan untuk memuat model dan preprocessor yang telah disimpan. Validasi+Preprocess bertujuan untuk memvalidasi format input dan melakukan transformasi data. Prediksi menghasilkan probabilitas booking (p_book) dan probabilitas visit (p_visit) dari kedua stage. Prioritize Top-80% & Channel bertujuan untuk mengurutkan lead berdasarkan p_book, menentukan rekomendasi channel berdasarkan threshold Stage-2, dan mengekspor hasil dalam format CSV/XLSX untuk distribusi ke tim sales.

3.2.2 Metode Data Mining

Penelitian ini menerapkan kerangka CRISP-DM karena bersifat terstruktur dan iteratif, meliputi tahap Business Understanding hingga Deployment. Untuk konteks dua tahap (prediksi booking dan rekomendasi channel), CRISP-DM

dipakai sebagai panduan utama dan dibandingkan secara ringkas dengan KDD dan SEMMA pada Tabel 3.1.

Tabel 3. 1 Perbandingan Framework [62][63]

Aspek	**CRISP-DM**	**KDD**	**SEMMA**
Tahapan Utama	1. *Business Understanding,* 2. *Data Understanding,* 3. *Data Preparation,* 4. *Modeling,* 5. *Evaluation,* 6. *Deployment*	1. *Selection,* 2. *Pre-processing,* 3. *Transformation,* 4. *Data Mining,* 5. *Interpretation*	1. *Sample,* 2. *Explore,* 3. *Modify,* 4. *Model,* 5. *Assess*
Pendekatan	Iteratif, berfokus pada pemahaman bisnis dan siklus pengembangan data mining	Iteratif, berfokus pada pengolahan data mentah menjadi informasi yang berguna	Linear, berfokus pada proses teknis data mining
Kelebihan	1. Cocok untuk berbagai industri 2. Pendekatan yang terstruktur dengan development 3. Iteratif untuk pembaruan	1. Proses yang terdefinisi dengan baik 2. Interaktif untuk kontrol manual dalam proses mining	1. Mudah dipahami 2. Memungkinkan eksplorasi data yang cepat
Kekurangan	1. Membutuhkan penyesuaian untuk menangani volume dan kecepatan data pada proyek big data 2. Tidak secara eksplisit mengatasi pengolahan data streming atau real-time	1. Tidak mencakup fase Deployment 2. Tidak ada definisi teknik spesifik	1. Tidak mencakup pemahaman bisnis 2. Mengabaikan aspek Deployment dan evaluasi pola

Berdasarkan perbandingan tersebut, CRISP-DM dipilih karena mampu mengintegrasikan analisis data dengan implementasi model secara terstruktur sekaligus mendukung proses iteratif untuk penyesuaian model berdasarkan hasil evaluasi. Karakter ini relevan untuk penelitian berbasis machine learning pada penetapan prioritas prospek dan rekomendasi metode follow-up (visit/phone), di mana data bersifat dinamis dan model perlu diperbarui secara berkala.

Berbeda dengan KDD yang lebih menekankan transformasi/penemuan pengetahuan maupun SEMMA yang mengutamakan eksperimen cepat, CRISP-DM menawarkan keseimbangan antara kedalaman analisis dan fleksibilitas proses serta mencakup tahap deployment. Pada penelitian ini, deployment dilakukan melalui

dashboard interaktif berbasis Streamlit yang menyajikan daftar prioritas serta rekomendasi metode follow-up kepada tim sales dan manajemen PT XYZ secara operasional.

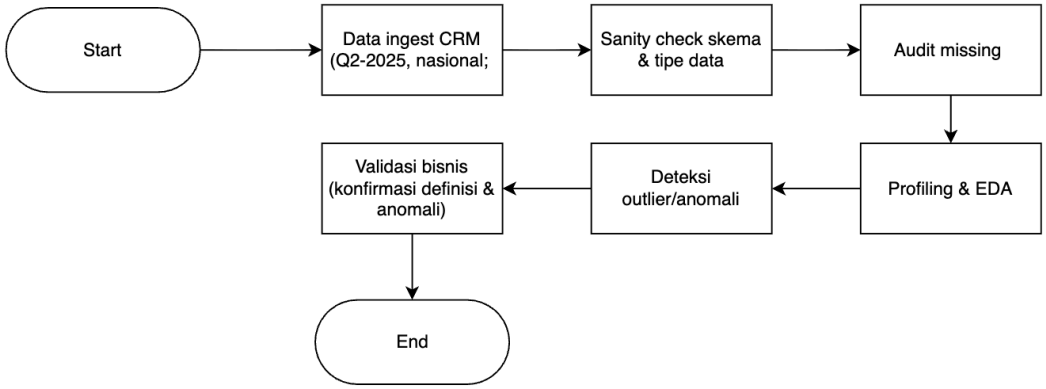
3.2.2.1 Business Understanding

Tahap business understanding bertujuan menetapkan tujuan dan manfaat penelitian dengan merujuk pada kebutuhan operasional PT XYZ. Permasalahan yang dihadapi adalah penetapan prioritas prospek dan pemilihan metode follow-up (visit/phone) yang masih bersifat manual dan subjektif, sehingga alokasi eksekusi sering tidak efisien dan sebagian peluang booking terlewat.

Penelitian ini menargetkan pengembangan pendekatan two-stage: Stage-1 memprediksi probabilitas booking sebagai dasar penyusunan prioritas prospek, sedangkan Stage-2 merekomendasikan metode follow-up yang paling tepat pada prospek terpilih. Rancangan ini diselaraskan dengan keterbatasan kapasitas eksekusi lapangan (sekitar 80% dari total prospek per siklus) agar hasil analitik mudah dioperasionalkan. Manfaat yang diharapkan adalah tersedianya dasar pengambilan keputusan yang objektif bagi tim sales, peningkatan efisiensi penggunaan sumber daya kunjungan/telepon, serta kenaikan peluang konversi melalui penanganan prospek yang lebih terarah.

3.2.2.2 Data Understanding

Tahap Data Understanding dilakukan untuk mengeksplorasi dataset dan memahami pola data yang relevan dengan tujuan penelitian ini. Seperti yang terlihat pada Gambar 3.2, proses ini diawali dengan pengumpulan data dari sistem Customer Relationship Management (CRM) Data Mining PT XYZ , yang memuat data historis dan profil pelanggan.



Gambar 3. 2 Diagram Proses Data Understanding

Gambar 3.2 menunjukkan tahap data understanding yang berisi proses pemahaman data yang akan digunakan dalam penelitian. Langkah pertama pada tahap ini adalah melakukan pengambilan data dari sistem Customer Relationship Management (CRM) internal PT XYZ, yang memuat data historis aktivitas sales dan profil pelanggan. Dataset mencakup periode Q2 2025 (April–Juni 2025) dengan cakupan nasional dan berukuran 579.081 baris × 31 kolom.

Selanjutnya dilakukan pemeriksaan skema dan tipe data untuk memastikan konsistensi antar-atribut serta audit kelengkapan terhadap setiap kolom. Berdasarkan hasil pemeriksaan, tingkat missing value di atas 1% hanya ditemukan pada variabel last_type_fu dan last_type_fu_std (sekitar 7,28%), sedangkan atribut lainnya memiliki kelengkapan hampir 100%.

Tahap berikutnya adalah eksplorasi data (EDA) untuk memahami distribusi dan karakteristik dataset. Analisis ini mencakup identifikasi rasio kelas target (BOOKING 17,02% vs NO_BOOKING 82,98%) serta perbandingan tingkat konversi berdasarkan metode follow-up (Phone 14,59% dan Visit 16,68%). Selain itu, dilakukan deteksi outlier dan anomali pada variabel numerik seperti salary, principal, dan month_inst untuk memastikan bahwa nilai ekstrem tidak mengganggu hasil pemodelan.

Tahap terakhir yaitu validasi bisnis, dilakukan bersama pihak analis untuk mengonfirmasi definisi variabel dan hasil eksplorasi awal, memastikan bahwa pola yang ditemukan sesuai dengan konteks operasional CRM. Hasil dari tahap data understanding ini menjadi dasar untuk proses data preparation selanjutnya, di mana data akan dibersihkan, ditransformasi, dan disiapkan bagi pemodelan machine learning dua tahap untuk prioritisasi prospek dan rekomendasi metode follow-up.

Selain itu, hasil eksplorasi data pada tahap data understanding juga mencakup identifikasi atribut yang akan digunakan dalam pemodelan. Setiap variabel dipilih berdasarkan relevansinya terhadap proses pengambilan keputusan pelanggan dan potensi hubungannya dengan peluang terjadinya booking maupun pemilihan metode follow-up.

Tabel 3.2 berikut menyajikan daftar kolom yang digunakan dalam penelitian ini beserta tipe data dan deskripsinya. Informasi ini menjadi dasar dalam proses data preparation dan feature engineering pada tahap berikutnya.

Tabel 3. 2 Daftar Kolom Dataset

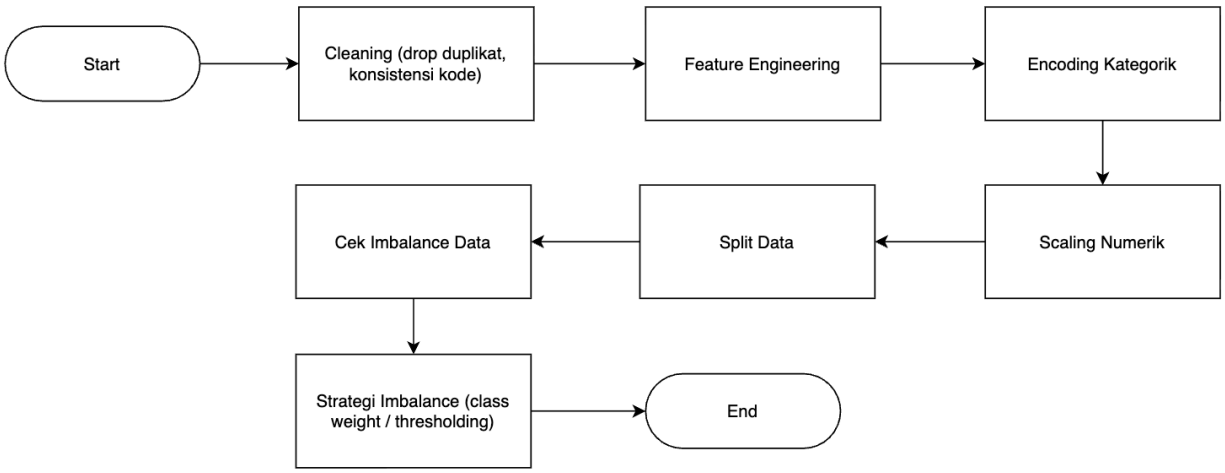
No	Nama Variabel	Tipe Data	Deskripsi
1	usia	Numerical	Umur pelanggan dalam tahun.
2	salary	Numerical	Penghasilan bulanan pelanggan.
3	no_of_depend	Numerical	Jumlah tanggungan keluarga pelanggan.
4	edu_type	Categorical	Tingkat pendidikan terakhir pelanggan (mis. SD, SMA, Sarjana).
5	marital_stat	Categorical	Status pernikahan pelanggan.
6	cust_sex	Categorical	Jenis kelamin pelanggan.
7	house_stat	Categorical	Status kepemilikan rumah (milik, sewa, keluarga, dll).
8	ocpt_code	Categorical	Kode jenis pekerjaan pelanggan.
9	obj_brand	Categorical	Merek objek pembiayaan (mis. Honda).
10	obj_tahun	Numerical	Tahun pembuatan objek pembiayaan.
11	principal	Numerical	Nilai pokok pinjaman (plafon kredit).
12	month_inst	Numerical	Besaran angsuran bulanan.
13	grs_dp	Numerical	Besaran uang muka kotor (gross down payment).
14	top	Numerical	Tenor pembiayaan (jangka waktu dalam bulan).

No	Nama Variabel	Tipe Data	Deskripsi
15	buss_unit	Categorical	Unit bisnis pembiayaan (NMC atau REFI).
16	periode_first_dist	Numerical	Periode distribusi pertama pelanggan.
17	cust_prov	Categorical	Provinsi domisili pelanggan.
18	cust_kec	Categorical	Kecamatan domisili pelanggan.
19	cust_kel	Categorical	Kelurahan domisili pelanggan.
20	appl_no	Numerical	Nomor aplikasi pelanggan.
21	contract_no	Numerical	Nomor kontrak pelanggan (jika ada).
22	cust_no	Numerical	Nomor unik identitas pelanggan.
23	bpkb_same_name	Categorical	Status kesesuaian nama pada dokumen BPKB (Sesuai/Tidak).
24	data_month	Numerical	Bulan periode data (mis. 202504–202506).
25	is_book	Categorical (Target Stage-1)	Status hasil follow-up (0 = No Booking, 1 = Booking).
26	last_type_fu	Categorical (Raw)	Tipe follow-up terakhir pelanggan (Phone/Visit).

Tabel 3.2 di atas menyajikan informasi rinci mengenai atribut-atribut dalam dataset yang digunakan pada penelitian ini, termasuk deskripsi dan tipe data masing-masing kolom. Atribut-atribut tersebut berperan penting dalam mengidentifikasi pola dan karakteristik pelanggan yang memengaruhi probabilitas terjadinya booking serta pemilihan metode follow-up yang optimal. Selain itu, atribut-atribut ini juga membantu menjelaskan hubungan antar fitur dalam proses pengambilan keputusan berbasis model machine learning.

Pemanfaatan dataset ini memungkinkan sistem untuk menganalisis berbagai informasi seperti data demografis pelanggan, kondisi finansial, riwayat kontrak dan pembayaran angsuran, riwayat interaksi atau tindak lanjut sales, serta karakteristik objek pembiayaan. Seluruh informasi tersebut relevan untuk mendukung pengembangan model prediksi dua tahap (two-stage prediction) yang akurat dan efektif, yaitu untuk menentukan prioritas prospek dan merekomendasikan metode follow-up (visit atau phone) secara lebih terarah.

3.2.2.3 Data Preparation



Gambar 3. 3 Diagram Proses Data Preparation

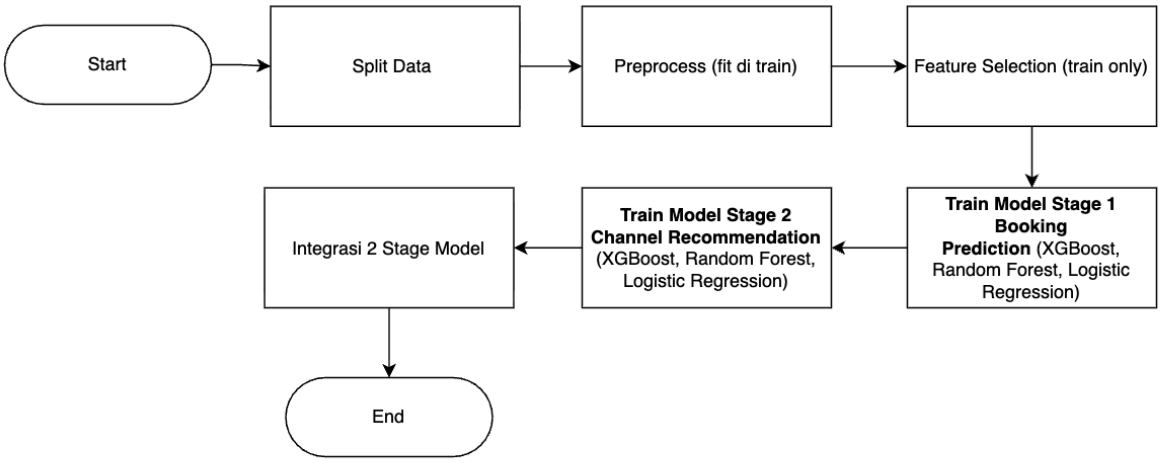
Gambar 3.3 menampilkan diagram proses data preparation yang terdiri dari tujuh langkah sekuensial untuk menyiapkan data agar siap digunakan pada proses pemodelan. Berikut penjelasan setiap langkah.

Proses dimulai dengan Cleaning (drop duplikat, konsistensi kode) untuk membersihkan data dengan menghapus duplikasi berdasarkan customer ID dan application number, menyesuaikan tipe data, serta menghapus kolom tanpa variasi seperti r_n, r_ng, r_nuu, dan r_nuug. Selanjutnya dilakukan Feature Engineering untuk memperkaya informasi model melalui pembuatan fitur baru, seperti rasio plafon terhadap pendapatan (principal/salary) dan beban angsuran bulanan terhadap plafon kredit (month_inst/principal).

Tahap berikutnya adalah transformasi dan pembagian data. Encoding Kategorik mengubah variabel kategorik ke bentuk numerik melalui label encoding dan target encoding. Scaling Numerik menskalakan variabel numerik menggunakan StandardScaler dengan prinsip fit on train, apply to test untuk mencegah kebocoran data. Split Data membagi dataset menjadi data latih, validasi, dan uji secara stratified agar distribusi kelas tetap seimbang.

Tahap terakhir adalah penanganan ketidakseimbangan kelas. Cek Imbalance Data menganalisis ketidakseimbangan pada target variable, di mana target is_book memiliki base rate 17,02%. Strategi Imbalance (class weight/thresholding) menangani ketidakseimbangan tersebut menggunakan kombinasi class weighting serta threshold tuning dengan evaluasi berbasis kapasitas top-80% meliputi Precision@80, Recall@80, dan Lift@80. Hasil dari tahap ini berupa data yang bersih, terstandardisasi, dan siap digunakan untuk tahap modelling.

3.2.2.4 Modelling



Gambar 3. 4 Diagram Proses Modelling

Gambar 3.4 menunjukkan proses pada tahap modelling yang bertujuan untuk membangun dan melatih model klasifikasi berdasarkan data yang telah melalui proses data preparation. Tahap ini diawali dengan pembagian data (split data) menjadi subset pelatihan dan pengujian yang digunakan secara terpisah untuk mencegah kebocoran informasi. Pembagian dilakukan secara stratified agar proporsi kelas target tetap seimbang, baik pada tahap prediksi booking maupun pada tahap rekomendasi metode follow-up.

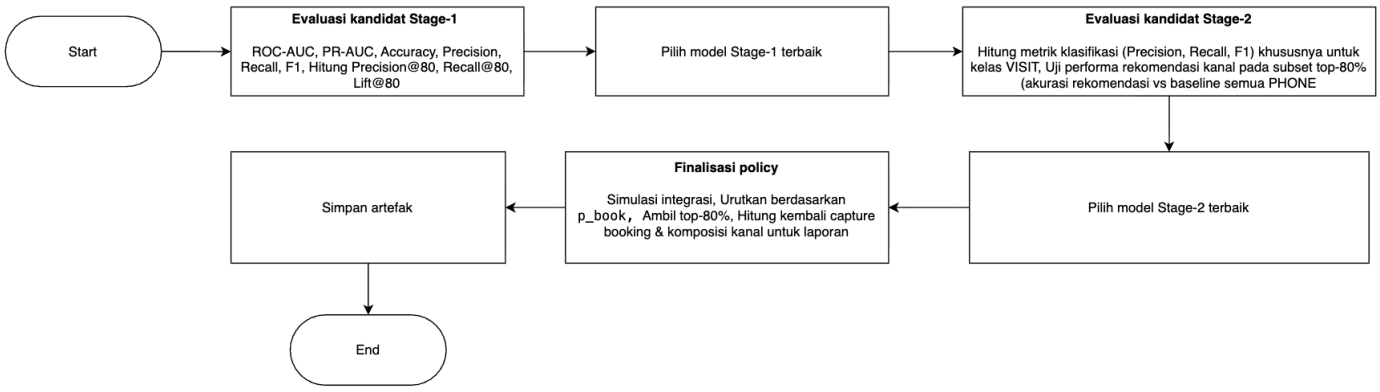
Selanjutnya dilakukan proses praproses dan seleksi fitur pada data latih. Langkah ini mencakup penerapan encoder dan scaler yang telah di-fit pada data pelatihan, kemudian digunakan pada data pengujian dengan prinsip fit on train,

apply to test. Setelah itu, dilakukan seleksi fitur (feature selection) menggunakan model berbasis tree untuk mengidentifikasi atribut dengan kontribusi prediktif tertinggi terhadap variabel target. Fitur terpilih kemudian digunakan dalam proses pelatihan dua tahap (two-stage modelling).

Pada Stage-1, model dilatih untuk memprediksi probabilitas keberhasilan booking (p_book) menggunakan tiga algoritma pembanding, yaitu Logistic Regression, Random Forest, dan XGBoost. Sementara itu, Stage-2 dilatih untuk memprediksi probabilitas kecenderungan kanal tindak lanjut (p_visit) berdasarkan label aktual visit dan phone. Seluruh model dilatih dengan parameter dasar, kemudian dilakukan penyetelan menggunakan metode early stopping pada XGBoost untuk memperoleh kombinasi model terbaik dengan performa stabil.

Tahap terakhir adalah integrasi model dua tahap, di mana hasil prediksi Stage-1 digunakan untuk menentukan prioritas prospek, sedangkan hasil Stage-2 berfungsi memberikan rekomendasi metode follow-up yang paling efektif. Hasil pelatihan kedua model disimpan sebagai artefak untuk digunakan pada tahap evaluation dan deployment berikutnya.

3.2.2.5 Evaluation

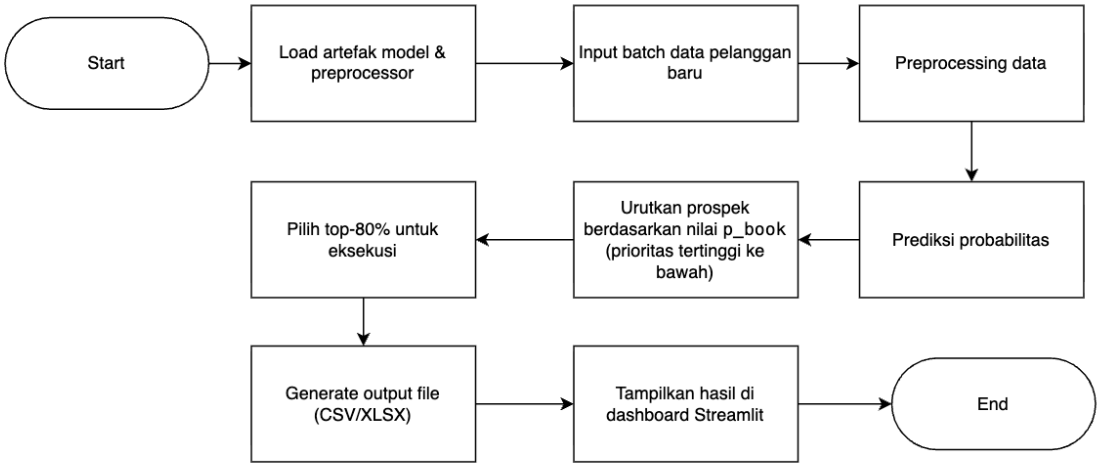


Gambar 3. 5 Diagram Proses Evaluation

Tahap evaluation dilakukan setelah proses pemodelan selesai untuk menilai kinerja model dan menentukan kombinasi model dua tahap yang paling optimal bagi kebutuhan bisnis PT XYZ. Seperti yang ditunjukkan pada Gambar 3.5, proses evaluasi diawali dengan pengujian performa kandidat model pada Stage-1 (prediksi booking) menggunakan data uji. Evaluasi dilakukan berdasarkan metrik ROC-AUC, PR-AUC, accuracy, precision, recall, dan F1-score, serta analisis confusion matrix untuk melihat distribusi prediksi benar dan salah. Selain itu, dilakukan pengukuran tambahan berupa Precision@80, Recall@80, dan Lift@80, dengan cara mengurutkan prospek berdasarkan nilai probabilitas p_book dan mensimulasikan eksekusi pada top-80% prospek untuk mencerminkan kondisi kapasitas eksekusi di lapangan. Model dengan performa terbaik berdasarkan metrik tersebut dipilih sebagai model final untuk Stage-1.

Selanjutnya, pada Stage-2 (rekomendasi metode follow-up), evaluasi difokuskan pada kemampuan model dalam mengklasifikasikan kanal visit dan phone. Pengujian dilakukan dengan menggunakan metrik precision, recall, dan F1-score, khususnya untuk kelas VISIT karena kunjungan lapangan memiliki biaya yang lebih tinggi dibandingkan panggilan telepon. Selain itu, dilakukan simulasi integrasi dua tahap dengan menggunakan hasil prediksi p_book untuk memilih top-80% prospek, kemudian menerapkan hasil prediksi p_visit sebagai dasar rekomendasi kanal. Hasil integrasi ini dibandingkan dengan skenario baseline (semua prospek diasumsikan melalui PHONE) untuk menilai peningkatan efektivitas rekomendasi model. Berdasarkan hasil evaluasi tersebut, ditetapkan model terbaik untuk masing-masing stage beserta ambang keputusan (threshold) yang digunakan, dan seluruh artefak model disimpan untuk digunakan kembali pada tahap deployment di dashboard interaktif..

3.2.2.6 Deployment



Gambar 3. 6 Diagram Proses Deployment

Gambar 3.6 menunjukkan tahapan deployment yang bertujuan untuk mengimplementasikan model ke dalam sistem aplikasi berbasis Streamlit. Proses ini dimulai dengan pemuatan artefak model dan preprocessor hasil tahap evaluation, mencakup model Stage-1 (prediksi probabilitas booking) dan Stage-2 (rekomendasi kanal follow-up).

Data pelanggan baru diunggah melalui dashboard, kemudian diproses menggunakan pipeline preprocessing untuk menghasilkan nilai probabilitas p_book dan p_visit. Hasil prediksi diurutkan berdasarkan nilai p_book, dan top-80% pelanggan dengan peluang tertinggi dipilih untuk dieksekusi. Model Stage-2 kemudian menentukan rekomendasi kanal follow-up (Visit atau Phone) berdasarkan nilai ambang keputusan. Output akhir ditampilkan pada dashboard Streamlit dalam bentuk tabel interaktif serta dapat diekspor ke file CSV/XLSX untuk digunakan oleh tim CRM PT XYZ. Pendekatan ini memastikan model dapat dioperasikan secara praktis dan terintegrasi dalam proses bisnis perusahaan.

3.3 Teknik Pengambilan Data

3.3.1 Populasi dan Sampel

Populasi dalam penelitian ini mencakup seluruh data aktivitas dan profil pelanggan yang tercatat pada sistem Customer Relationship Management (CRM) milik PT XYZ. Data tersebut terdiri atas berbagai informasi seperti karakteristik demografis pelanggan, riwayat transaksi pembiayaan, aktivitas kunjungan sales, serta status hasil penawaran (booking atau no booking).

Sampel penelitian diperoleh dengan menggunakan teknik non-probability sampling melalui pendekatan purposive sampling, yaitu pemilihan data secara sengaja berdasarkan relevansi terhadap tujuan penelitian. Pendekatan ini digunakan untuk memastikan bahwa data yang digunakan memiliki kelengkapan atribut dan konsistensi informasi yang diperlukan untuk analisis prediktif. Dataset yang digunakan merupakan data pelanggan pada periode kuartal II tahun 2025 (April–Juni) dengan cakupan nasional, berjumlah 579.081 baris dan 31 kolom. Data diperoleh melalui proses ekstraksi langsung dari database CRM perusahaan menggunakan TOAD for Oracle, di mana query dijalankan oleh tim analis untuk menarik data sesuai kebutuhan penelitian. Hasil ekstraksi kemudian disimpan dalam format CSV untuk tahap pembersihan dan analisis lanjutan menggunakan Python.

3.3.2 Periode Pengambilan Data

Data yang digunakan dalam penelitian ini diambil dari periode April hingga Juni 2025 (Q2 2025). Pemilihan periode ini dilakukan karena volume data transaksi dan aktivitas pelanggan pada sistem CRM perusahaan berskala nasional sangat besar, sehingga penggunaan data dalam satu kuartal dinilai paling optimal untuk menjaga efisiensi pemrosesan, kestabilan komputasi, serta keterwakilan kondisi operasional sales.

Periode Q2 2025 dipilih karena mencerminkan data terkini yang relevan dengan konteks analisis efektivitas kunjungan sales dan rekomendasi metode follow-up. Selain itu, rentang waktu ini menunjukkan kestabilan performa bisnis

dan tingkat aktivitas follow-up yang konsisten, sehingga hasil analisis dapat merepresentasikan pola kerja aktual dari tim sales perusahaan.

3.4 Variabel Penelitian

Variabel dalam penelitian ini merupakan atribut atau fitur yang diperoleh dari data historis aktivitas sales dan profil pelanggan yang tersimpan pada sistem Customer Relationship Management (CRM) milik PT XYZ. Variabel-variabel ini digunakan sebagai masukan dalam proses pemodelan machine learning untuk memprediksi probabilitas booking pada Stage-1 dan menentukan rekomendasi metode follow-up pada Stage-2. Secara umum, variabel dibagi menjadi beberapa kelompok utama berdasarkan karakteristiknya, yaitu:

1. Data demografis pelanggan, seperti usia (usia), jenis kelamin (cust_sex), status pernikahan (marital_stat), tingkat pendidikan (edu_type), jumlah tanggungan (no_of_depend), dan status kepemilikan rumah (house_stat). Variabel-variabel ini digunakan untuk menggambarkan kondisi sosial ekonomi pelanggan yang dapat memengaruhi peluang pelanggan dalam melakukan pembiayaan.
2. Data geografis, yang meliputi provinsi (cust_prov), kecamatan (cust_kec), dan kelurahan (cust_kel) tempat tinggal pelanggan. Informasi ini membantu mengidentifikasi pola wilayah dan segmentasi lokasi pelanggan terhadap efektivitas penawaran.
3. Data finansial, yang terdiri atas pendapatan bulanan (salary), nilai pokok pembiayaan (principal), tenor atau jangka waktu pembiayaan (top), jumlah angsuran per bulan (month_inst), serta uang muka kotor (grs_dp). Atribut finansial ini merepresentasikan kemampuan ekonomi pelanggan dan risiko kredit yang menjadi indikator penting dalam peluang booking.
4. Data objek pembiayaan, meliputi merek kendaraan (obj_brand) dan tahun kendaraan (obj_tahun). Kedua variabel ini digunakan untuk menilai preferensi produk serta potensi pembiayaan berdasarkan kondisi aset yang dibiayai.
5. Data bisnis dan riwayat CRM, seperti unit bisnis (buss_unit), periode distribusi pertama (periode_first_dist), dan status kesesuaian nama dokumen (bpkb_same_name). Atribut ini membantu model mengenali hubungan historis pelanggan dengan perusahaan dan aktivitas sebelumnya dalam sistem CRM.

Selain variabel utama tersebut, penelitian ini juga menambahkan variabel hasil feature engineering seperti rasio angsuran terhadap pendapatan (installment-to-salary ratio) dan rasio uang muka terhadap nilai pokok pembiayaan (down payment-to-principal ratio). Kedua rasio ini berfungsi untuk memperkuat interpretasi model terhadap kapasitas finansial pelanggan.

Seluruh variabel tersebut digunakan secara bersamaan untuk melatih model klasifikasi pada kedua tahap, dengan penyesuaian ruang lingkup data pada masing-masing stage. Pada Stage-1, variabel-variabel ini digunakan untuk memprediksi status hasil follow-up pelanggan (is_book), yang bersifat biner dan menunjukkan apakah pelanggan berhasil melakukan transaksi (nilai 1 untuk BOOKING dan 0 untuk NO_BOOKING). Pada Stage-2, variabel yang sama digunakan untuk memprediksi jenis metode follow-up terakhir (last_type_fu_std), yang juga bersifat biner dengan nilai 1 untuk VISIT (kunjungan langsung) dan 0 untuk PHONE (panggilan telepon).

Kedua variabel prediksi (is_book dan last_type_fu_std) memiliki hubungan yang saling berurutan, di mana hasil prediksi Stage-1 digunakan untuk menentukan pelanggan dengan peluang booking tertinggi, dan hasil prediksi Stage-2 digunakan untuk menentukan kanal interaksi yang paling efektif bagi kelompok pelanggan tersebut. Dengan demikian, hasil prediksi kedua tahap ini secara bersama-sama mendukung sistem rekomendasi yang mampu meningkatkan efisiensi dan efektivitas kinerja tim sales dalam proses tindak lanjut prospek pelanggan.

3.5 Teknik Analisis Data

3.5.1 Tools Analisis Data

Penelitian ini menggunakan bahasa pemrograman Python dengan environment utama Google Colaboratory sebagai tools analisis data. Google Colaboratory dipilih karena mendukung kebutuhan penelitian berbasis machine learning dengan kemampuan

komputasi yang efisien, mudah digunakan, serta terintegrasi dengan pustaka analisis data populer seperti pandas, NumPy, scikit-learn, dan XGBoost.

Selain itu, Google Colaboratory memiliki dukungan GPU dan TPU yang memungkinkan proses pelatihan model berjalan lebih cepat dibandingkan dengan lingkungan lokal. Tools ini juga berbasis cloud, sehingga tidak memerlukan instalasi perangkat lunak tambahan dan dapat diakses kapan saja melalui browser. Kelebihan lain adalah kemudahan integrasi dengan Google Drive untuk penyimpanan data serta kemudahan kolaborasi antar peneliti. Sebagai perbandingan, Tabel 3.3 berikut menampilkan evaluasi beberapa tools yang umum digunakan untuk penelitian analisis data:

Tabel 3. 3 Perbandingan Tools Analisis Data [64]

Indikator	Google Colaboratory	Jupyter Notebook	R Studio
Bahasa Pemrograman	Python dan mendukung bahasa lain	Julia, Python, R	R
Keunggulan	1. Gratis dengan akses GPU/TPU	1. Fleksibel untuk berbagai bahasa	1. Berfokus pada analisis statistik
	2. Berbasis *cloud*	2. Komputasi cepat	2. Memiliki pustaka bawaan yang lengkap
	3. Terintegrasi dengan Google Drive	3. Dokumentasi dan visualisasi interaktif	
	4. Dapat disimpan ke GitHub dengan mudah		
Kekurangan	1. Membutuhkan koneksi internet	1. Beberapa tugas memerlukan kode tambahan	1. Kurang fleksibel
	2. Batasan waktu *runtime*		2. Hanya optimal untuk bahasa R
Fleksibilitas	Tinggi dan dapat dikonfigurasi	Tinggi	Rendah
Kompleksitas	Rendah	Rendah	Tinggi

Berdasarkan hasil perbandingan tersebut, penelitian ini menggunakan Google Colaboratory karena paling sesuai untuk kebutuhan eksperimen *machine learning* dengan dataset berukuran besar dan proses komputasi intensif. Platform ini menyediakan keseimbangan antara kemudahan penggunaan, kecepatan pemrosesan, serta fleksibilitas dalam integrasi dengan berbagai pustaka Python yang diperlukan dalam pengembangan dan evaluasi model klasifikasi dua tahap.

3.5.1 Metode Validasi Model

Metode validasi model digunakan untuk mengukur performa dan reliabilitas model *machine learning* yang dibangun pada penelitian ini, yaitu model dua tahap (*two-stage prediction model*) yang terdiri atas *Stage-1 (Booking Prediction)* dan *Stage-2 (Channel Recommendation)*. Validasi dilakukan untuk memastikan bahwa model tidak hanya mampu memberikan hasil yang baik pada data pelatihan, tetapi juga memiliki kemampuan generalisasi yang stabil terhadap data baru.

Proses validasi dilakukan dengan menggunakan metode train-test split yang membagi dataset menjadi dua bagian, yaitu 80% sebagai data latih (*training set*) dan 20% sebagai data uji (*testing set*). Pembagian dilakukan secara *stratified* untuk menjaga proporsi kelas target agar tetap representatif, mengingat dataset memiliki ketidakseimbangan antara kelas positif dan negatif.

Selain itu, diterapkan strategi penyesuaian untuk mengatasi permasalahan *class imbalance* dengan pendekatan *class weight* serta penyesuaian ambang batas (*threshold tuning*) berdasarkan nilai *F1-Score* terbaik. Pendekatan ini memastikan model tetap sensitif terhadap kelas minoritas, khususnya pelanggan yang benar-benar melakukan *booking* atau pelanggan yang lebih efektif ditindaklanjuti dengan kanal *visit*.

Evaluasi performa model dilakukan menggunakan beberapa metrik utama yang mencerminkan baik aspek teknis maupun relevansi bisnis, yaitu:

1. ROC-AUC (Receiver Operating Characteristic - Area Under Curve) untuk mengukur kemampuan model dalam membedakan antara kelas positif dan negatif.
2. PR-AUC (Precision-Recall Area Under Curve) untuk mengevaluasi ketepatan model dalam kondisi data tidak seimbang.
3. Accuracy, Precision, Recall, dan F1-Score sebagai metrik dasar dalam menilai keseimbangan antara ketepatan dan kelengkapan prediksi.
4. Precision@80, Recall@80, dan Lift@80, yang menggambarkan kinerja model pada kondisi business constraint aktual, di mana kapasitas tim sales

rata-rata hanya dapat mengeksekusi sekitar 80% dari total prospek pelanggan dalam satu periode kerja.

Tahap akhir validasi dilakukan melalui **simulasi integrasi dua tahap** yang merepresentasikan proses nyata di lapangan. Simulasi ini mencakup pemeringkatan pelanggan berdasarkan probabilitas *booking (p_book)* dari *Stage-1*, kemudian rekomendasi kanal *follow-up* optimal dari *Stage-2* pada subset 80% pelanggan prioritas teratas. Hasil validasi ini menjadi dasar untuk menilai efektivitas sistem rekomendasi dalam meningkatkan efisiensi dan keberhasilan aktivitas penjualan di PT XYZ.