Business Understanding		
Determine Business Objective	Background	Proporsi pasien IGD dengan LOS ≥ 6 jam masih tinggi (diatas 10%) dan berkontribusi pada penurunan mutu layanan, keterlambatan intervensi, dan risiko keselamatan pasien.
	Business Objective	Mendefinisikan problem operasional IGD dan tujuan prediksi prolonged LOS (≥6 jam) yang dapat ditindaklanjuti.
	Business Success Criteria	Model memberikan sinyal risiko yang dapat digunakan untuk intervensi dini dan manajemen kapasitas IGD.
		Model mencapai recall tinggi terhadap prolonged LOS dengan tingkat false positive yang masih bisa dikelola secara operasional.
		Model menunjukkan stabilitas performa terhadap variasi desain fitur dan dapat memberikan interpretasi fitur yang dapat divalidasi secara klinis dan operasional
Assess Situation	Inventory of Resources	Dataset ED 2025; N=25.126 Identifikasi 12 variabel mentah Cross-validation dataset: ED 2026 (3 bulan pertama) data untuk evaluasi robustness model.
	Requirements, Asumptions, and Constraint	Requirements : Data kunjungan IGD lengkap selama periode 2025
		Asumptions : Vital sign yang tercatat mencerminkan kondisi awal pasien, Recall lebih diprioritaskan dibanding precision
		Constraint : Single-center study Data hanya satu tahun (2025) Variabel klinis terbatas pada yang tersedia saat awal kedatangan
	Risks and Contingencies	Risk: Data Leakage, Penambahan banyak fitur interaksi menyebabkan overfitting, Synthetic samples menyebabkan model tidak realistis pada data asli.
		Contingencies : Pipeline yang terkontrol, cross validation, ablation
	Cost and Benefits	Cost : Koordinasi tim klinis, potensi penyiapan infrastruktur real-time inference.
		Benefits : Penguatan keselamatan pasien, optimalisasi arus IGD, efisiensi biaya layanan, dukungan evidence-based decision making.
Determine Data Mining Goals	Data mining goals	Primary Modeling Goal : Mengembangkan model klasifikasi biner untuk memprediksi risiko prolonged LOS (≥ 6 jam) menggunakan fitur yang tersedia pada fase awal kedatangan IGD. Performance Goal :Mengoptimalkan kemampuan diskriminasi model dengan memaksimalkan ROC-AUC, mengutamakan recall (sensitivity) untuk kelas prolonged LOS Feature Representation Goal : mengevaluasi representasi fitur terstruktur (Model A) sudah cukup informatif, apakah penambahan interaksi eksplisit berbasis klinisi (Model B) memberikan peningkatan performa Robustness Goal : Menilai stabilitas model terhadap penghilangan fitur dominan (ablation study) dan ketergantungan model terhadap variabel administratif
	Data mining success criteria	Model dianggap berhasil secara teknis apabila memenuhi kriteria berikut: Mencapai kemampuan diskriminasi yang baik dalam memprediksi prolonged LOS (≥ 6 jam), ditunjukkan oleh nilai ROC-AUC yang memadai. Mencapai tingkat sensitivitas (recall) tinggi terhadap kasus prolonged LOS untuk mendukung deteksi dini dalam konteks skrining IGD. Menunjukkan performa yang stabil terhadap variasi desain fitur (Model A vs Model B) dan tidak mengalami penurunan signifikan pada studi ablasi. Memberikan hasil interpretasi berbasis SHAP yang konsisten dengan pemahaman klinis dan operasional IGD. Tidak menunjukkan ketergantungan berlebihan pada satu fitur dominan yang berpotensi menyebabkan bias prediksi.
Produce Project Plan	Project Plan	Fase 1: Pemahaman dan Persiapan Data Fase 2: Rekayasa Fitur dan Pengembangan Model Fase 3: Evaluasi dan Interpretabilitas Fase 4: Analisis Robustness dan Validasi Metodologis
	Initial assesment of tools and techniques	Light Gradient Boosting Machine (LightGBM) sebagai model utama berbasis tree untuk menangkap hubungan non-linear. Logistic Regression sebagai baseline linear untuk pembanding. Optuna untuk optimasi hyperparameter berbasis cross-validation. SHAP (SHapley Additive exPlanations) untuk analisis interpretabilitas model. SMOTE (Synthetic Minority Oversampling Technique) untuk menangani ketidakseimbangan kelas pada data latih. Threshold optimization untuk menyesuaikan titik keputusan klasifikasi dengan kebutuhan operasional IGD.
Data Understanding		
Collect Initial Data	Initial Data Collection Report	• Data period 2025; N=25.126 • Identifikasi 12 variabel mentah • Konfirmasi engineered features belum digunakan di tahap ini
Describe Data	Data Description Report	1. Distribusi Data -Distribusi fitur numerik (usia, LOS, jumlah diagnosis/prosedur) -Distribusi fitur kategorikal (arrival shift, day-of-week, payer type, referral source) 2. Outcome & Imbalance Analysis -Proporsi LOS < 6 vs ≥ 6 jam -Identifikasi high-risk minority class 3. Clinical Pattern Analysis -Profil diagnosis dominan & grouped diagnosis (ICD-10 categories) -Relationship diagnosis → prolonged LOS -Severity proxies: complexity count, life-saving interventions -Operational Context Analysis -Beban kunjungan (load_4h) & korelasi dengan LOS -Temporal risk patterns: jam puncak, hari kerja vs akhir pekan -Pengaruh shift terhadap LOS outcome
	Variabel	
	usia_tahun, durasi_kunjungan_igd,load_4h	Numerik
	kd_pasien, jenis_kelamin,kd_customer, nama_dokter, unit_rwi,status, load_category	Kategorikal
	waktu_masuk, waktu_keluar	datetime
	max_los	Target - Kategorikal
Explore Data	Data Exploration Report	*Imbalance Target Data LOS >=6 jam 24%*
Verify Data Quality	Data Quality Report	-Ada missing pada kolom diagnosis, procedure di sebagian kunjungan, -Ditemukan kasus LOS ≤ 0 atau > 72 jam (anomali), -Ada pasien dengan kunjungan berulang (bukan duplikasi)

Data Preparation

Select Data	Rationale for Inclusion/Exclusion	Included: Semua kunjungan pasien IGD tahun 2025 yang memiliki waktu masuk dan keluar valid. Excluded: -Kunjungan tanpa informasi waktu keluar -LOS ≤ 0 jam (data salah input/time logging error) -LOS > 72 jam (dianggap outlier operasional → transfer/rawat inap tercecer) Privacy measure: kd_pasien dipseudonimkan untuk menjaga kerahasiaan data pasien.
Clean Data	Data Cleaning Report	Drop : jika waktu_keluar = 0 dan durasi_jam > 72 jam, jika Timestamp invalid (waktu_keluar < waktu_masuk).
		Variasi penulisan diagnosis (ICD) distandarisasi. Variasi penulisan tindakan (produk_list) diseragamkan. Format vital sign dikonversi menjadi numerik terstruktur. Tekanan darah (tensi) dipecah menjadi SBP dan DBP.
Construct Data	Derived Atrributes Generated Records	age_bin, weekend_flag, shift_flag, overload_flag, risk_dx_flag, tind_lifesaving, SHAP-guided interactions
Integrate Data	Merged Data	waktu_masuk dan waktu_keluar di merge menjadi durasi rawat IGD
Format Data	Reformatted Data	Membuat list diagnosa dan tindakan-Daftar_diagnosa dan daftar_Produk → diparse menjadi: dx_* (kelompok diagnosis ICD-10) pd_* (top principal diagnosis flags) tind_* (indikator prosedur/tindakan) Format diubah dari text dengan multi-hot encoded binary
	Dataset Description	-Periode Jan – Des 2025 -Total kunjungan setelah cleaning 25,126 -Target LOS ≥ 6 jam (high-risk) -Jumlah fitur akhir untuk modeling 93 features utama + clinician

Modelling

Select Modelling Techniques	Modeling Techniques	Logistic Regression Digunakan sebagai baseline linear untuk mengevaluasi kemampuan diskriminatif dari representasi fitur terstruktur. Light Gradient Boosting Machine (LightGBM) Digunakan sebagai model utama berbasis tree untuk menangkap hubungan non-linear dan interaksi kompleks antar fitur. Metode peningkatan performa: Penanganan class imbalance menggunakan SMOTE yang diterapkan secara eksklusif pada data latih Optimasi hyperparameter menggunakan Optuna berbasis cross-validation Threshold optimization dilakukan untuk menyesuaikan trade-off antara sensitivitas dan presisi sesuai kebutuhan skrining operasional IGD. Analisis interpretabilitas menggunakan SHAP untuk mengevaluasi kontribusi fitur dan membandingkan Model A dan Model B.
	Modeling Assumptions	Pola klinis dan operasional yang tersedia pada fase awal kedatangan pasien mengandung sinyal prediktif yang cukup untuk memodelkan risiko prolonged LOS. Perbandingan antara Model A dan Model B dilakukan dengan asumsi bahwa pipeline eksperimen identik, sehingga perbedaan performa mencerminkan kontribusi desain fitur, bukan variasi prosedur pelatihan.
Generate Test Design	Test Design	Train–test split stratified (80/20) untuk mempertahankan distribusi imbalanced. Evaluasi performa menggunakan: -Recall sebagai primary metric (safety-critical) -F1-score dan ROC-AUC sebagai pendukung Threshold tuning pada test-set tanpa oversampling
Build Model	Parameters Settings	Best param (hasil optuna)
	Models Descriptions	Baseline Model (Model A – Structured Features) Model baseline dikembangkan menggunakan representasi fitur terstruktur hasil structured feature engineering (Model A). Interaction-Extended Model (Model B – Clinician-Guided Features) Model B dibangun dengan menambahkan fitur interaksi eksplisit berbasis hipotesis klinis terhadap dataset Model A.
Assess Model	Model Assesment	Analisis SHAP mengindikasikan bahwa prediksi prolonged LOS didorong oleh kombinasi faktor: Faktor klinis (misalnya kondisi trauma, gangguan respirasi, infeksi, abnormalitas vital sign) Faktor operasional (misalnya beban layanan tinggi, kunjungan malam hari, akhir pekan) Hasil ini menunjukkan bahwa risiko prolonged LOS bersifat multidimensional, melibatkan interaksi antara kondisi klinis dan konteks operasional IGD
	Revised Parameter Settings	Threshold probabilitas diturunkan dari 0,50 menjadi sekitar 0.25untuk meningkatkan sensitivitas terhadap prolonged LOS. Set fitur diperluas melalui penambahan interaksi klinis–operasional (Model B). Optimasi hyperparameter dilakukan secara iteratif menggunakan Optuna. Analisis SHAP digunakan untuk mengevaluasi kontribusi fitur dan menginformasikan penyempurnaan desain fitur.

Evaluation

Evaluate Results	Assesment of Data Mining Results	Model berbasis fitur interaksi (Model B) tidak menunjukkan peningkatan sensitivitas dibandingkan baseline terstruktur (Model A), tree-based model sudah menangkap interaksi implisit Optimasi threshold menghasilkan peningkatan recall yang signifikan dengan konsekuensi peningkatan false positive yang masih dalam batas operasional. Analisis SHAP menunjukkan bahwa prediksi didorong oleh kombinasi faktor klinis dan operasional, bukan hanya satu variabel dominan. Studi ablasi menunjukkan bahwa meskipun beberapa fitur memiliki kontribusi tinggi (misalnya variabel administratif), model tetap mempertahankan performa yang stabil ketika fitur tersebut dihilangkan, sehingga mengurangi risiko ketergantungan berlebihan.
	Business Success Criteria & Approved Models	Model LightGBM dengan fitur interaksi (Model A) dan threshold teroptimasi dipilih sebagai model utama. Model baseline (Logistic Regression dan Clinician Guide Model B) digunakan sebagai pembanding

Review Process	Review of Process	Proses penelitian mengikuti tahapan CRISP-DM secara sistematis: -Dimulai dari pemahaman masalah operasional IGD dan definisi target prolonged LOS. -Dilanjutkan dengan analisis kualitas data, pembersihan, dan rekayasa fitur terstruktur. -Dikembangkan dua desain fitur (Model A dan Model B) dengan pipeline pelatihan yang identik. -Dilakukan optimasi hyperparameter, evaluasi performa, serta penyesuaian threshold. -Interpretabilitas dianalisis menggunakan SHAP. -Robustness diuji melalui studi ablasi fitur. Proses ini memastikan bahwa: -Perbandingan model bersifat terkontrol. -Tidak terjadi kebocoran informasi. -Performa model tidak hanya tinggi secara numerik tetapi juga dapat dijelaskan dan diuji stabilitasnya.
Determine Next Steps	List of Possible Actions Decision	Mengintegrasikan model sebagai alat skrining risiko prolonged LOS dalam sistem informasi IGD. Melakukan uji coba terbatas (pilot implementation) dalam workflow triase. Melakukan validasi temporal tambahan menggunakan data tahun berikutnya.
Deployment		
Plan Deployment	Deployment Plan	- Mode Implementasi Awal: Model berjalan di background tanpa mengintervensi keputusan klinis - Integrasi Sistem: Ditempatkan pada modul Dashboard IGD pada SIMRSHEBAT - Input data otomatis dari registrasi IGD & triase: usia, diagnosa awal, jam masuk, load_4h, shift - Output model: Risk Score LOS ≥ 6 jam Risk Flag (High / Low) -Notifikasi Klinis: Jika High-risk, muncul indikator warna (merah/oranye) Terdapat tombol “Eskalasi ke Bed Management” User Role Access: -Dokter IGD, Perawat Triase, Tim Bed Management, ITISI Admin Pilot Deployment: -Durasi: 1 bulan -Shift fokus awal: Sore–Malam (puncak overload berdasarkan analisis)
Plan Monitoring and Maintenance	Monitoring & Maintenance Plan	-Monitoring Harian Validasi prediksi vs real LOS (feedback klinis) Log jumlah alerts & false alerts -Monitoring Mingguan Update visual dashboard: status kasus high-risk, FN ratio Evaluasi acceptance & workflow impact -Model Performance Tracking Recall, precision, dan SHAP stability setiap bulan Early trigger jika recall turun >5% -Data Drift & Model Retraining Drift diperiksa secara otomatis (ks-test / feature distribution shift) Retraining model berkala dilakukan 3–6 bulan -Backup & Fail-Safe Jika model down maka sistem kembali ke decision manual Semua keputusan final oleh tenaga medis
Produce Final Report	Final Report	Berisi ringkasan lengkap: 1. CRISPDM (Latar belakang klinis & operasional IGD) 2. Metodologi dan Experimental Design 3. Rekap performa model (pre/post refinement) 4. Analisis SHAP dan interpretasi klinis 5. Rencana deployment lanjutan dan rekomendasi pengembangan Disiapkan dalam format: Journal Q4
	Final Presentation	Audience: Direksi, Kabid Pelayanan, Kepala IGD, Tim Bed Management, ITISI
Review Project	Experience Documentation	SOP Model Deployment & Monitoring User Feedback Log (Perawat & Dokter Triase) Incident Record jika ada false alert berdampak klinis