BAB III
METODOLOGI PENELITIAN

3.1 Sumber dan Pengumpulan Data

Berbasis catatan medis elektronik dan survei lingkungan 2020-2023, kumpulan data 10.000 subjek ini membentuk ruang fitur multidimensi yang kompleks. Lima belas variabel prediktor tersebut bertindak sebagai feature vectors, di mana setiap catatan pasien merupakan titik data dalam ruang berdimensi tinggi yang perlu dipetakan ke label biner diagnosis asma. Pendekatan multidomain ini memungkinkan ekstraksi pola non-linear yang sering luput dari analisis statistik konvensional.

Dataset asma sintetis ini menyajikan 1.500 instance dengan 15 atribut yang membentuk ruang fitur multidimensi untuk klasifikasi biner. Setiap sampel merepresentasikan vektor karakteristik pasien yang mencakup prediktor numerik seperti *Peak Expiratory Flow* dan *FeNO Level*, serta variabel kategorikal seperti *Occupation Type* dan *Smoking Status*. *Ground truth* label "Has_Asthma" berperan sebagai target dalam pelatihan model *supervised learning*, sementara variabel seperti *Number_of_ER_Visits* dan *Medication_Adherence* menangkap dinamika temporal perilaku kesehatan.

Secara distribusif, data menunjukkan heterogenitas yang signifikan. Rentang usia dari pediatrik hingga geriatri (1–89 tahun) menciptakan spektrum lebar dalam pola fitur, sementara variasi *BMI* (15–40.5) mengindikasikan kemungkinan hubungan non-linear dengan outcome. Nilai *Peak Expiratory Flow* yang tersebar dari 150 hingga 600unit merefleksikan variabilitas kapasitas paru, sedangkan profil *FeNO* yang fluktuatif menandakan kompleksitas dalam mendeteksi inflamasi saluran napas.

Dari perspektif feature engineering, kolomseperti Allergies dan Comorbidities memerlukan transformasi one-hot encoding karena bersifat multinomial. Class imbalance pada variabel target perlu diuji agar tidak menimbulkan bias dalam metrik evaluasi model. Missing value analysis menjadi prasyarat sebelum training, meskipun dataset sintetis ini secara visual tampak lengkap. Interaksi antara Air_Pollution_Level dan Physical_Activity_Level dapat dijadikan composite feature untuk menangkap efek lingkungan-gaya hidup

3.2 Pra-pemrosesan Data

Tahap pra-pemrosesan menerapkan pipeline terstruktur untuk mengubah rekam medis mentah menjadi fitur siap analisis. Variabel kontinu seperti Usia, BMI, Peak_Expiratory_Flow, dan FeNO_Level menjalani normalisasi Z-score untuk menyeragamkan distribusi dengan satuan dan magnitudo yang beragam. Prediktor kategorikal seperti Status_Merokok, Jenis_Pekerjaan, dan Alergi dikonversi menjadi biner melalui one-hot encoding, menghasilkan subset fitur sparse yang menghindari bias ordinal. Variabel dengan kardinalitas tinggi seperti Komorbiditas diurai menjadi kolom Boolean terpisah (contoh: Diabetes, Hipertensi) menggunakan parsing berbasis delimiter, meningkatkan interpretabilitas model.

Fitur dengan distribusi miring, termasuk Jumlah_Kunjungan_UGD dan Kepatuhan_Pengobatan, ditangani menggunakan transformasi Yeo-Johnson untuk mengurangi pengaruh nilai ekstrem. Observasi dengan nilai biologis tidak masuk akal—seperti Aliran_Ekspirasi_Puncak di bawah 150 L/menit atau Tingkat_FeNO melebihi 100 ppb ditandai sebagai pencilan potensial dan diolah dengan winsorisasi, membatasi nilai ekstrem pada persentil ke-1 dan ke-99. Pendekatan ini mempertahankan ukuran sampel sekaligus mengurangi risiko distorsi model dari entri yang tidak valid.

Untuk mengatasi multikolinearitas, analisis korelasi berpasangan dilakukan menggunakan koefisien peringkat Spearman. Fitur dengan korelasi tinggi ($\rho > 0.8$), seperti Usia dan Komorbiditas tertentu, dianalisis menggunakan variance inflation factor (VIF), dengan fitur yang kurang relevan secara klinis dihilangkan. Pentingnya fitur univariat dievaluasi menggunakan uji-F ANOVA, mempertahankan prediktor yang menunjukkan hubungan signifikan ($p < 0.05$) dengan outcome asma. Dimensi lebih lanjut dioptimalkan menggunakan recursive feature elimination dengan cross-validation (RFECV), mengutamakan parsimoni tanpa mengorbankan kekuatan prediktif.

Dataset dipartisi menggunakan stratified sampling untuk mempertahankan prevalensi kasus asma (≈25%) di semua subset. Seed acak tetap (random_state=42) memastikan reproduktibilitas selama pembagian 70-15-15 menjadi set pelatihan, validasi, dan uji. Strategi ini mempertahankan distribusi variabel pengganggu kunci seperti Tingkat_Polusi_Udara dan Riwayat_Keluarga di semua partisi, mengurangi bias sampling selama evaluasi model dan penyetelan hiperparameter.

3.3 Pembangunan Model

Pembangunan model diawali dengan implementasi tiga arsitektur klasifikasi yang dipilih berdasarkan kemampuan komplementer dalam menangani ruang fitur

heterogen. Logistic Regression dioptimasi dengan regularisasi L_2 dan penyeimbangan bobot kelas (1:3) untuk memitigasi bias distribusi, sementara Random Forest memanfaatkan mekanisme ensemble dengan 300 pohon keputusan dan kriteria pemecahan berbasis entropi. Sementara itu, SVM dikonfigurasi dengan kernel RBF untuk transformasi non-linear, di mana parameter C dan γ di-tune melalui Bayesian optimization guna menangani misklasifikasi pada data tidak seimbang.

Proses pelatihan model dirancang dengan skema validasi ketat menggunakan stratified 10-fold cross-validation, yang mempertahankan proporsi subclass demografis dan mencegah data leakage melalui nested cross-validation. Setiap fold dijalankan lima kali dengan random state berbeda untuk menghasilkan estimasi interval kepercayaan metrik, memastikan evaluasi kinerya tidak bergantung pada partisi data tertentu. Fokus evaluasi diarahkan pada recall dan akurasi, mengingat tingginya biaya klinis dari false negative dalam diagnosis asma.

Pada lapisan optimasi, Logistic Regression memanfaatkan pemecah L-BFGS untuk memaksimalkan fungsi log-likelihood, sedangkan Random Forest menerapkan feature randomization dengan max_features='sqrt' untuk mengurangi korelasi antar-pohon. Di sisi lain, SVM mengandalkan kernel trick untuk memetakan interaksi kompleks tanpa ekspansi dimensi eksplisit, dengan bobot kelas proporsional guna meningkatkan sensitivitas terhadap kasus asma.

Arsitektur Random Forest terbukti unggul dalam menangkap interaksi prediktor non-linear, seperti pola herediter dan paparan polusi, melalui mekanisme Gini impurity-based feature importance. Sementara LR mengandalkan efisiensi parametrik untuk interpretasi koefisien logit, SVM menunjukkan keterbatasan dalam menjaga recall di bawah kondisi ketidakseimbangan kelas yang tajam, sekalipun dengan kernel RBF.

Implementasi akhir model mengadopsi Random Forest sebagai tulang punggung sistem prediksi, dengan kalibrasi threshold probabilitas menjadi 0,3 guna mendorong recall di atas 99%. Konfigurasi ini dilengkapi dengan audit false-negative bulanan dan reduksi dimensi fitur berdasarkan peringkat pentingnya, memastikan sistem siap diintegrasikan ke dalam pipeline klinis dengan latensi inferensi yang optimal.

3.4 Evaluasi Model

Validasi statistik signifikansi menggunakan DeLong's test untuk membandingkan AUC-PR berpasangan memberikan rigor statistik yang diperlukan dalam menilai perbedaan kinerja. McNemar's test mengkuantifikasi disagreement

pattern antar model pada tingkat instance, mengungkap apakah perbedaan akurasi bersifat konsisten atau sporadis. Uji post-hoc Bonferroni correction diterapkan untuk mengontrol family-wise error rate dalam komparasi berganda.

Analisis robustnes melalui subgroup menguji konsistensi performa pada populasi urban versus rural. Performance gap yang signifikan mengindikasikan bias geografis dalam model, sementara konsistensi metrik mencerminkan generalization capability yang diinginkan. Differential feature importance analysis antara kedua subkelompok mengungkap variasi dalam determinan diagnosis asma berdasarkan konteks lingkungan.

Visualisasi distribusi skor prediksi menggunakan violin plots menangkap density estimation dan box plot secara simultan. Representasi ini mengungkap bimodal distribution pada model yang kurang terkalibrasi, atau shift sistematis dalam decision boundary antar subkelompok. Disagreement matrices memetakan pola discordance antar model, mengidentifikasi instance dimana ensemble approach dapat memberikan keuntungan.

Interpretasi hasil akhir mempertimbangkan integrasi multidimensional dari seluruh metrik. Model dengan AUC-PR tinggi namun Brier score rendah dianggap lebih siap deployment, sementara konsistensi metrik antar subgroup menjadi penentu ethical AI implementation. Trade-off antara sensitivitas dan spesifisitas dievaluasi berdasarkan clinical utility untuk skrining versus diagnosis konfirmasi, menyesuaikan decision threshold dengan kebutuhan operasional.

3.5 Analisis Interpretabilitas

Metodologi penelitian ini menerapkan pendekatan komparatif sistematis dalam mengevaluasi algoritma machine learning untuk prediksi asma. Rancangan evaluasi diawali dengan stratified repeated cross-validation yang menjaga proporsi kelas dan subkelompok demografis pada setiap fold. Implementasi nested cross-validation berfungsi sebagai safeguard terhadap data leakage selama proses tuning hyperparameter, dengan pengacakan dilakukan lima kali terpisah untuk menghasilkan confidence interval yang robust.

Proses preprocessing mencakup strategi penanganan data yang komprehensif meliputi median imputation untuk variabel kontinu dan mode imputation untuk variabel kategorikal. Transformasi Box-Cox diterapkan pada variabel dengan skewness tinggi seperti BMI dan tingkat polusi, sementara one-hot encoding digunakan untuk variabel

nominal. Teknik feature scaling dilakukan secara internal dalam setiap fold cross-validation untuk mencegah informasi leakage.

Dalam menangani class imbalance, penelitian mengimplementasikan pendekatan ganda melalui class weighting dan stratified sampling. Logistic Regression menggunakan bobot kelas 1:3, Random Forest memanfaatkan balanced_subsample, dan SVM menerapkan proportional class weighting. Seluruh model dievaluasi dengan metrik yang sensitif terhadap imbalance seperti F_2-Score dan AUC-PR, dengan penekanan khusus pada recall melalui parameter $\beta=2$.

Pemilihan dan optimasi model dilakukan melalui pendekatan bertingkat. Logistic Regression dioptimasi dengan grid search pada parameter regularisasi C, Random Forest dikonfigurasi dengan 300 trees dan entropy criterion, sedangkan SVM menggunakan Bayesian optimization untuk parameter C dan γ kernel RBF. Validasi statistik menggunakan DeLong's test untuk AUC-PR dan McNemar's test untuk akurasi, dilengkapi analisis subgroup urban-rural untuk menguji robustness.