BAB I
PENDAHULUAN

1.1 Latar Belakang

Asma bronkial tetap menjadi beban kesehatan global yang kompleks, dengan manifestasi klinis yang dipengaruhi oleh interaksi multivariat yang mencakup usia, indeks massa tubuh, dan paparan terhadap polutan udara. Etiologi penyakit ini tidak hanya bergantung pada riwayat keluarga atau status merokok, tetapi juga melibatkan biomarker seperti tingkat FeNO dan laju aliran ekspirasi puncak, yang mencerminkan peradangan saluran napas. Dalam konteks big data kesehatan, mengeksplorasi faktor-faktor heterogen ini memerlukan teknik penambangan data khusus untuk mengungkap korelasi tersembunyi di antara variabel demografis (jenis kelamin, pekerjaan), klinis (penyakit penyerta, kepatuhan terhadap obat), dan lingkungan (polusi udara, alergen) yang sulit diidentifikasi secara manual.

Peran inti penambangan data beralih dari eksplorasi umum ke pembangunan sistem klasifikasi prediktif, mengubah 15 variabel heterogen (demografis, klinis, lingkungan) menjadi matriks fitur terstruktur untuk mengidentifikasi pola risiko tersembunyi. Teknik seperti seleksi fitur mengoptimalkan dimensi masukan dengan menyaring noise (misalnya, korelasi rendah antara jenis pekerjaan dan kepatuhan terhadap obat), sementara resampling mengatasi ketidakseimbangan kelas antara pasien asma dan non-asma. Untuk menjembatani kesenjangan metodologis ini, studi ini menerapkan evaluasi komparatif yang ketat terhadap tiga paradigma algoritmik yang berbeda: Logistic Regeresion sebagai model linier dasar, Random Forest yang menggunakan pembelajaran ensambel, dan Support Vector Machine (SVM) yang memanfaatkan fungsi kernel untuk pemetaan ruang fitur non-linier[1], [2]. Perbandingan multidimensi ini dirancang untuk mengungkap keunggulan relatif masing-masing arsitektur dalam menangani karakteristik unik dataset asma, termasuk interaksi variabel tingkat tinggi, jenis data campuran (kategorikal-numerik), dan ketidakseimbangan kelas. Ketiga model pembelajaran mesin ini dipilih karena kekuatan komplementernya: Logistic Regeresion (LR) menawarkan interpretabilitas koefisien melalui fungsi logit untuk analisis pentingnya fitur. Random Forest (RF) mengelola interaksi kompleks melalui bagging dan acak fitur, mengoptimalkan kinerja pada data non-IID , SVM dengan kernel RBF mentransformasi fitur menjadi ruang pemisahan optimal. Random Forest (RF) mengelola interaksi kompleks melalui bagging dan acak fitur, mengoptimalkan kinerja pada data non-IID [3]–[5]. SVM dengan kernel RBF

mengubah fitur menjadi hiperplane pemisah optimal (margin maksimum), terbukti efektif untuk data berisik tinggi. Pemilihan ini menyeimbangkan trade-off antara keterjelaskan model dan kapasitas generalisasi [6], [7].

Dalam analisis biomedis, mengubah data klinis menjadi kecerdasan pengambilan keputusan meningkatkan ketepatan diagnosis neurologis. Misalnya, multiple sclerosis (MS) menunjukkan perkembangan gejala yang tidak stabil, menghasilkan hiperplane kompleks di ruang fitur akibat variabilitas biomarker temporal[8]. Studi perbandingan sebelumnya menerapkan SVM dan LR pada kohort pasien MS, di mana SVM menunjukkan margin pemisahan yang lebih unggul melalui transformasi kernel RBF mencapai akurasi 88,33%. Kinerja ini mencerminkan kemampuan optimasi margin maksimum untuk mengkodekan pola lesi spasial-waktu, menegaskan relevansinya sebagai mesin klasifikasi untuk gangguan neurodegeneratif berdimensi tinggi[9], [10].

Salma Rihadatul Ais (2025) [11] menunjukkan bahwa kanker tiroid, meskipun prevalensinya relatif rendah, memerlukan pendekatan diagnostik yang akurat untuk mengurangi risiko kekambuhan. Dengan menerapkan Logistic Regeresion, random forest, dan XGBoost untuk memprediksi kekambuhan menggunakan 14 variabel klinis dari 2.000 pasien Rumah Sakit Ken Saras, Logistic Regeresion mencapai akurasi tertinggi (83%), mengungguli model lainnya. Keunggulannya meliputi interpretabilitas tinggi untuk mengidentifikasi faktor risiko dominan (misalnya usia, ukuran tumor), efisiensi komputasi untuk integrasi klinis real-time, dan output risiko probabilistik yang memungkinkan stratifikasi pasien. Kombinasi akurasi, interpretabilitas, dan efisiensi ini menjadikan LR sebagai alat pendukung keputusan klinis yang ideal, menunggu validasi klinis holistik.

Support Vector Machine (SVM) [12]membangun hiperplane optimal di ruang fitur berdimensi tinggi untuk memisahkan kelas risiko asma. Algoritma ini memaksimalkan margin terlebar antara titik data terdekat dari kedua kelas (vektor dukungan) untuk meningkatkan generalisasi model. Untuk data yang tidak dapat dipisahkan secara linier (misalnya, interaksi kompleks antara gen dan polusi), SVM menggunakan trik kernel untuk memetakan fitur ke ruang baru (misalnya, kernel RBF) tanpa perhitungan nonlinier eksplisit. Parameter regularisasi C mengontrol toleransi kesalahan klasifikasi, sementara penimbangan kelas mengatasi ketidakseimbangan data. Ketahanannya terhadap noise dan efisiensinya dalam ruang dimensi tinggi menjadikannya kandidat kuat untuk mengklasifikasikan risiko dari biomarker klinis yang saling bergantung dan variabel lingkungan.

Hevi Alvina Damayanti (2025) [13] menyoroti keunggulan teknis SVM dalam membangun hiperplane optimal melalui mekanisme maximisasi margin sebuah mekanisme yang secara inheren meningkatkan generalisasi dan mengurangi overfitting dengan memperluas batas pemisahan kelas. Evaluasi empiris mengonfirmasi stabilitas diskriminatifnya, dibuktikan dengan AUC-ROC yang konsisten di atas 0,93, mencerminkan presisi pemisahan fitur berdimensi tinggi. Algoritma ini juga menunjukkan ketahanan terhadap ketidakseimbangan kelas, mempertahankan presisi ≥0,87 dan recall ≥0,96 untuk kelas mayoritas, yang disebabkan oleh transformasi fitur yang diinduksi kernel yang mempertahankan minimisasi risiko struktural meskipun ada ketidakseimbangan sampel.

Random Forest (RF) membangun ensambel pohon keputusan melalui bootstrap aggregating (bagging), di mana setiap pohon dilatih pada subset data acak menggunakan seleksi fitur acak. Hal ini mengurangi varians model dan mencegah overfitting yang umum terjadi pada pohon Tunggal [14]–[16]. Untuk klasifikasi risiko asma, RF secara implisit menangani interaksi non-aditif di antara prediktor heterogen (misalnya, korelasi antara status sosial-ekonomi dan paparan alergen) serta nilai yang hilang. Pentingnya fitur diukur melalui pengurangan ketidakpastian Gini atau kesalahan out-of-bag, memberikan wawasan tentang kontribusi variabel demografis, klinis, dan lingkungan. Model ini secara inheren mendukung data yang tidak seimbang melalui penyesuaian bobot kelas atau bootstrapping terstratifikasi sambil mempertahankan akurasi prediksi yang tinggi pada dataset yang berisik.

R.M. Aldani Adi Bhirawa (2025) [17] menyoroti pengurangan overfitting RF melalui bagging dan pemilihan fitur acak, serta penanganan yang andal terhadap data numerik/kategorikal dan ketahanan terhadap outlier. Estimasi pentingnya fitur yang kritis membantu analisis klinis. Gradient boosting (terutama XGBoost) beroperasi sebagai ensembel boosting berurutan, memperbaiki residu dari model sebelumnya melalui regularisasi, optimisasi komputasi, dan fitur yang meningkatkan kecepatan/akurasi prediksi.

Penelitian ini mengkaji kemampuan komparatif tiga algoritma klasifikasi, yaitu Logistic Regeresion (LR), Random Forest (RF), dan Support Vector Machine (SVM), dalam membangun prediktor risiko asma dari 15 variabel demografis, klinis, dan lingkungan yang heterogen. Melalui optimasi seleksi fitur untuk mengurangi redundansi dan prediktor dengan varians rendah (misalnya, korelasi antara pekerjaan dan kepatuhan pengobatan), tujuan utamanya adalah: Memetakan perbedaan kinerja algoritma dalam menangani interaksi tingkat tinggi (misalnya, sinergi polutan-gen) dan kompleksitas

data campuran. Mengidentifikasi arsitektur optimal untuk mengubah data klinis-lingkungan mentah menjadi matriks fitur siap keputusan.

Secara metodologis, studi ini mengatasi kesenjangan pemilihan model dengan mengevaluasi trade-off antara interpretabilitas dan kinerja. LR dievaluasi melalui pelacakan koefisien logit untuk analisis penjelasan. RF diuji dalam menangkap hubungan non-aditif menggunakan pentingnya fitur berbasis Gini. SVM diverifikasi pada manifold dimensi tinggi melalui transformasi yang diinduksi kernel.

Hasil ini memungkinkan sistem dukungan keputusan klinis untuk mengekstrak pola risiko tersembunyi dari biomarker (FeNO, aliran ekspirasi) dan variabel sosio-lingkungan. Implementasi praktis meliputi mesin stratifikasi risiko real-time untuk intervensi pencegahan yang ditargetkan, sementara kerangka kerja rekayasa fitur dapat disesuaikan untuk penyakit multifaktorial lainnya. Secara komputasi, temuan ini menetapkan preseden untuk penerapan model yang efisien sumber daya dalam infrastruktur kesehatan yang terbatas.

1.2 Rumusan Masalah

Penelitian ini berangkat dari beberapa kesenjangan dalam membangun prediktor risiko asma dari data yang heterogen. Pertama, terdapat kebutuhan untuk mengatasi redundansi fitur dan variabel prediktor dengan varians rendah (seperti korelasi antara pekerjaan dan kepatuhan pengobatan) yang dapat mengganggu kinerja model. Kedua, belum jelas bagaimana perbandingan kinerja algoritma klasifikasi seperti Logistic Regeresion (LR), Random Forest (RF), dan Support Vector Machine (SVM) dalam menangani kompleksitas data campuran serta interaksi non-linier tingkat tinggi (misalnya, sinergi antara polutan dan gen). Ketiga, terdapat trade-off atau pertukaran antara interpretabilitas model dan kinerja prediktif yang belum terpecahkan dalam pemilihan model untuk konteks ini. Terakhir, diperlukan suatu arsitektur atau kerangka kerja optimal untuk mengubah data klinis dan lingkungan mentah menjadi matriks fitur yang siap digunakan untuk pengambilan keputusan klinis.

1.3 Batasan Masalah

1. Algoritma yang Dibandingkan: Penelitian ini hanya membandingkan tiga algoritma klasifikasi, yaitu Logistic Regeresion (LR), Random Forest (RF), dan Support Vector Machine (SVM).

2. Cakupan Variabel: Variabel prediktor dibatasi pada 15 variabel yang bersifat demografis, klinis, dan lingkungan. Proses optimasi akan berfokus pada seleksi fitur dari kumpulan variabel ini.
3. Fokus Evaluasi Model: Evaluasi kinerja dan interpretabilitas model akan berpusat pada metode yang telah disebutkan, seperti pelacakan koefisien logit untuk LR, pentingnya fitur berbasis Gini untuk RF, dan transformasi kernel untuk SVM.
4. Konteks Aplikasi: Meskipun kerangka kerja yang dihasilkan dapat disesuaikan untuk penyakit lain, implementasi dan validasi langsung dalam penelitian ini difokuskan pada prediksi risiko asma.

1.4 Tujuan dan Manfaat

1.4.1 Tujuan Penelitian

Adapun tujuan penelitian ini, diantaranya :

1) Menganalisis dan membandingkan kinerja tiga algoritma klasifikasi — *Logistic Regression* (LR), *Random Forest* (RF), dan *Support Vector Machine* (SVM) — dalam memprediksi risiko asma berdasarkan data klinis, demografis, dan lingkungan yang heterogen.
2) Mengoptimalkan proses seleksi fitur untuk mengurangi redundansi serta meningkatkan relevansi prediktor yang berpengaruh terhadap risiko asma.
3) Mengevaluasi trade-off antara interpretabilitas model (seperti koefisien logit pada LR) dan kemampuan generalisasi model (seperti margin maksimum pada SVM dan pembelajaran ensambel pada RF).
4) Mengembangkan kerangka kerja analitik yang dapat mengubah data mentah klinis-lingkungan menjadi matriks fitur terstruktur untuk sistem pendukung keputusan klinis.

1.4.2 Manfaat Penelitian

Beberapa manfaat dari penelitian ini :

1) Memberikan pemahaman komparatif mengenai efektivitas tiga algoritma pembelajaran mesin dalam menangani data campuran dan ketidakseimbangan kelas pada kasus risiko asma.
2) Menyediakan pendekatan metodologis yang dapat digunakan oleh praktisi kesehatan dalam mengidentifikasi faktor risiko asma secara lebih akurat dan efisien.

3) Mendorong pengembangan sistem pendukung keputusan klinis berbasis data (*data-driven clinical decision support*) yang mampu mengekstrak pola risiko tersembunyi dari biomarker dan variabel lingkungan.
4) Menjadi referensi bagi penelitian lanjutan dalam penerapan *machine learning* pada penyakit multifaktorial lain, dengan menekankan keseimbangan antara akurasi prediksi dan interpretabilitas model.

1.5 Sistematika Penulisan

1.5.1 BAB I PENDAHULUAN

- 1.1 Latar Belakang

  Uraian mengenai tantangan dalam mendiagnosis dan memprediksi risiko asma akibat kompleksitas faktor genetik, klinis, dan lingkungan, serta peluang pemanfaatan algoritma machine learning untuk analisis data heterogen.
- 1.2 Rumusan Masalah

  Pertanyaan utama yang mencakup perbandingan kinerja algoritma, penanganan interaksi non-linier, dan optimasi rekayasa fitur.
- 1.3 Batasan Masalah

  Penjelasan mengenai ruang lingkup algoritma, variabel, dan fokus penelitian.
- 1.4 Tujuan dan Manfaat Penelitian

  Penjabaran tujuan spesifik serta manfaat teoritis, praktis, dan adaptif dari penelitian.
- 1.5 Sistematika Penulisan

  Penjelasan singkat mengenai isi dan alur setiap bab dalam laporan.

1.5.2 BAB II TINJAUAN PUSTAKA

- 2.1 Asma sebagai Penyakit Multifaktorial

  Tinjauan mengenai patofisiologi asma dan faktor risiko demografis, klinis (seperti FeNO, aliran ekspirasi), dan lingkungan.
- 2.2 Konsep Machine Learning dalam Kesehatan

  Penjelasan tentang peran klasifikasi dan prediksi dalam membangun sistem pendukung keputusan klinis.
- 2.3 Algoritma Klasifikasi yang Digunakan

  Dasar teoretis dari Logistic Regeresion (LR), Random Forest (RF), dan Support Vector Machine (SVM), termasuk kekuatan dan kelemahannya.

- 2.4 Teknik Seleksi dan Rekayasa Fitur

  Metode untuk menangani redundansi, varians rendah, dan mengoptimalkan matriks fitur dari data heterogen.
- 2.5 Evaluasi Model dan Interpretabilitas

  Metrik evaluasi kinerja model (seperti Akurasi, Presisi, Recall, AUC-ROC) dan teknik interpretasi model (koefisien logit, importance Gini).
- 2.6 Penelitian Terkait

  Studi literatur dari penelitian sebelumnya yang menerapkan algoritma serupa untuk prediksi penyakit.
- 2.7 Kerangka Pemikiran

  Diagram alur yang menghubungkan variabel input, proses pemodelan, evaluasi, dan output yang diharapkan.

1.5.3 BAB III METODOLOGI PENELITIAN

- 3.1 Sumber dan Pengumpulan Data

  Penjelasan mengenai sumber dataset yang digunakan, yang terdiri dari 15 variabel demografis, klinis, dan lingkungan.
- 3.2 Pra-pemrosesan Data

  Tahapan handling missing data, encoding variabel kategorikal, dan normalisasi atau standardisasi data numerik.
- 3.3 Rekayasa dan Seleksi Fitur

  Proses identifikasi dan mengurangi redundansi fitur (seperti korelasi pekerjaan dan kepatuhan) serta memilih subset fitur yang paling informatif untuk pemodelan.
- 3.4 Pembangunan Model

  Implementasi dan optimasi (hyperparameter tuning) dari tiga algoritma: Logistic Regeresion (LR),Random Forest (RF), dan Support Vector Machine (SVM).
- 3.5 Evaluasi Model

  Protokol validasi (misalnya, Cross-Validation) dan pengukuran kinerja menggunakan metrik yang telah ditentukan untuk membandingkan ketiga algoritma secara objektif.
- 3.6 Analisis Interpretabilitas

Metode untuk menganalisis model yang dibangun, termasuk pelacakan koefisien LR, pentingnya fitur Gini pada RF, dan analisis dukungan vektor pada SVM.

1.5.4 BAB IV HASIL DAN PEMBAHASAN

- 4.1 Karakteristik Data

  Deskripsi statistik dari dataset setelah melalui tahap pra-pemrosesan dan seleksi fitur.
- 4.2 Hasil Evaluasi Kinerja Model

  Penyajian performa ketiga algoritma (LR, RF, SVM) berdasarkan metrik evaluasi yang dipilih dalam bentuk tabel dan grafik (misalnya, ROC Curves).
- 4.3 Analisis Interpretabilitas Model

  Pembahasan mengenai insight yang diperoleh dari model, seperti variabel prediktor paling penting dari RF dan makna klinis dari koefisien Logistic Regeresion.
- 4.4 Pembahasan Komparatif

  Analisis mendalam mengenai trade-off antara kinerja dan interpretabilitas dari setiap algoritma, serta keefektifannya dalam menangani interaksi kompleks dan data campuran. Pembahasan juga mencakup kelebihan dan kekurangan setiap pendekatan dalam konteks prediksi risiko asma.

1.5.5 BAB V PENUTUP

- 5.1 Kesimpulan

  Ringkasan temuan utama dan pencapaian tujuan penelitian.
- 5.2 Ucapan Terima Kasih

1.5.6 DAFTAR PUSTAKA

Referensi buku, jurnal, dan sumber online yang digunakan dalam penelitian.

1.5.7 LAMPIRAN

- Gambar Code
- Source Code