

BAB I

PENDAHULUAN

1.1 Latar Belakang

Tuberkulosis (TBC) merupakan penyakit menular yang masih menjadi masalah kesehatan utama di dunia, terutama di negara berkembang dengan kepadatan penduduk tinggi dan tingkat sosial ekonomi menengah ke bawah. Penyakit ini disebabkan oleh *Mycobacterium tuberculosis*, yang menyerang paru-paru dan dapat menyebar ke organ lain bila tidak segera ditangani. Berdasarkan laporan *Global Tuberculosis Report* tahun 2024, Indonesia menempati posisi kedua dengan jumlah kasus baru tertinggi setelah India, menunjukkan bahwa upaya pengendalian penyakit ini masih menghadapi berbagai tantangan (*World Health Organization.*, 2024a).

Penyakit TBC tidak hanya berdampak pada kesehatan individu, tetapi juga memberikan beban sosial dan ekonomi yang besar bagi masyarakat. Pasien yang menderita TBC sering kehilangan produktivitas akibat lamanya masa pengobatan, sementara keluarga dan lingkungan sekitar berisiko tertular. Di sisi lain, sistem pelayanan kesehatan harus menanggung biaya pemeriksaan dan pengobatan jangka panjang yang cukup tinggi. Oleh karena itu, deteksi dini menjadi salah satu langkah penting dalam upaya pengendalian penyebaran penyakit. Dengan perkembangan teknologi informasi dan ketersediaan data rekam medis digital, pemanfaatan analisis data menjadi salah satu strategi potensial untuk mendukung diagnosis awal TBC (Kementerian Kesehatan Republik Indonesia, 2020).

Salah satu pendekatan modern yang banyak digunakan dalam analisis data kesehatan adalah *machine learning*. Teknologi ini memungkinkan komputer mempelajari pola data dan menghasilkan prediksi secara otomatis tanpa pemrograman eksplisit. Dalam TBC, penerapan algoritma *machine learning* dapat membantu memprediksi status penyakit berdasarkan parameter klinis dan riwayat pasien, seperti lama batuk, berat badan, riwayat pengobatan, dan hasil pemeriksaan laboratorium. Dari berbagai algoritma yang ada, *Gradient Boosting* termasuk yang paling efektif karena kemampuannya dalam menggabungkan sejumlah model sederhana untuk membentuk model akhir yang kuat dan akurat (Hossain *et al.*, 2025).

Gradient Boosting merupakan metode *ensemble learning* yang bekerja secara bertahap dengan memperbaiki kesalahan prediksi pada setiap iterasi. Algoritma ini dipilih karena kemampuannya dalam menangkap hubungan yang tidak linier serta interaksi antarvariabel klinis, di mana pengaruh suatu gejala terhadap status TBC tidak selalu bersifat proporsional dan dapat bergantung pada kombinasi dengan gejala lain, seperti lama batuk yang berkaitan dengan jenis batuk atau penurunan berat badan. Dalam berbagai studi klasifikasi medis, *Gradient Boosting* terbukti menghasilkan tingkat akurasi yang tinggi dan lebih stabil dibanding algoritma lain seperti *Decision Tree* tunggal atau *K-Nearest Neighbor* (Hossain *et al.*, 2025). Performa model ini dipengaruhi oleh kualitas dan distribusi data yang digunakan. Salah satu permasalahan umum dalam data medis adalah ketidakseimbangan kelas (*imbalanced data*), dengan jumlah pasien pada salah satu kelas jauh lebih dominan dibandingkan kelas lainnya.

Ketidakseimbangan data tersebut menyebabkan model cenderung bias terhadap kelas mayoritas dan gagal mengenali pola penting dari kelas minoritas. Sebagai contoh, pada data TBC jumlah pasien positif umumnya lebih sedikit dibanding pasien negatif, sehingga model lebih cenderung menghasilkan prediksi pada kelas negatif. Untuk mengatasi permasalahan ini, digunakan teknik *Synthetic Minority Oversampling Technique* (SMOTE), yaitu metode penyeimbangan data dengan menambah sampel sintetis pada kelas minoritas. Teknik ini menciptakan data baru berdasarkan interpolasi antara titik-titik tetangga terdekat, sehingga menghasilkan distribusi data yang lebih proporsional tanpa menimbulkan duplikasi berlebihan (Elreedy *et al.*, 2024).

SMOTE tidak hanya meningkatkan jumlah data pada kelas minoritas, tetapi juga memperluas representasi distribusi variabel, sehingga proses pembelajaran model menjadi lebih seimbang. Penelitian yang dilakukan menunjukkan bahwa penerapan SMOTE mampu meningkatkan sensitivitas model terhadap data minoritas dan menghasilkan prediksi yang lebih adil antara kedua kelas (Elreedy *et al.*, 2024). Pendekatan serupa digunakan oleh peneliti lain yang menggabungkan metode *K-Means* dan SMOTE untuk menyeimbangkan data sebelum proses klasifikasi. Hasilnya, model mengalami peningkatan akurasi secara signifikan karena data menjadi lebih proporsional antara kelas mayoritas dan minoritas (Maulana *et al.*, 2024).

Interpretabilitas model merupakan salah satu tantangan penting dalam penerapan *machine learning* di bidang medis. Tenaga medis tidak hanya membutuhkan hasil akhir berupa prediksi, tetapi juga penjelasan yang jelas mengenai alasan di balik keputusan model. Dalam layanan kesehatan, setiap hasil

analisis harus dapat dijelaskan secara rasional agar dapat dipercaya dan digunakan sebagai dasar pengambilan keputusan klinis. Oleh karena itu, muncul konsep *Explainable Artificial Intelligence* (XAI) yang bertujuan memberikan pemahaman terhadap proses internal model *machine learning* yang kompleks.

Dua metode yang paling sering digunakan dalam XAI adalah *SHapley Additive exPlanations* (SHAP) dan *Local Interpretable Model-agnostic Explanations* (LIME). SHAP menjelaskan kontribusi masing-masing variabel terhadap hasil prediksi secara global, sedangkan LIME berfokus pada penjelasan lokal untuk kasus tertentu. Kombinasi kedua metode tersebut memberikan pemahaman yang komprehensif terhadap perilaku model, baik secara global maupun pada tingkat individu (Hasan, 2023). Penerapan SHAP dan LIME pada model prediksi dapat memberikan informasi yang transparan, seperti variabel yang paling memengaruhi hasil diagnosis pasien TBC.

Kedua metode tersebut digunakan untuk meningkatkan interpretabilitas model dengan mengidentifikasi faktor-faktor klinis yang berpengaruh terhadap hasil klasifikasi dalam studi deteksi penyakit Alzheimer (Vimbi *et al.*, 2024). Hasil serupa juga diperoleh pada penelitian mengenai prediksi risiko rawat inap ulang pasien diabetes dengan menggunakan pendekatan *Explainable Ensemble Learning* yang mengintegrasikan SMOTE, SHAP, dan LIME. Kombinasi tersebut berhasil meningkatkan performa model sekaligus menghasilkan interpretasi yang jelas bagi tenaga medis (Pinem *et al.*, 2025).

Integrasi *Gradient Boosting*, SMOTE, serta *Explainable AI* (SHAP dan LIME) menjadi solusi potensial dalam pengembangan sistem klasifikasi tuberkulosis berbasis data medis berdasarkan efektivitas metode tersebut. *Gradient*

Boosting digunakan untuk membangun model dengan akurasi tinggi, SMOTE diterapkan untuk menyeimbangkan data agar model tidak bias terhadap kelas mayoritas, sedangkan SHAP dan LIME berfungsi menjelaskan hasil prediksi secara transparan. Pendekatan ini memungkinkan sistem tidak hanya menghasilkan klasifikasi, tetapi juga memberikan alasan yang dapat dipahami oleh tenaga kesehatan.

Penerapan konsep ini relevan di Puskesmas Gayamsari, Semarang, sebagai salah satu fasilitas layanan kesehatan dasar yang menangani banyak kasus TBC setiap tahunnya. Data rekam medis pasien yang tersimpan di puskesmas berisi berbagai parameter klinis seperti gejala batuk kronis, berat badan, dan riwayat pengobatan yang dapat diolah menjadi sumber informasi untuk pelatihan model. Dengan menerapkan sistem klasifikasi otomatis berbasis *Gradient Boosting* dan *Explainable AI*, tenaga medis dapat memperoleh bantuan dalam mendeteksi kemungkinan TBC sejak tahap awal pemeriksaan. Pendekatan tersebut berpotensi mempercepat proses diagnosis serta meningkatkan efektivitas pengobatan.

Penelitian ini memiliki nilai ilmiah yang tinggi dari sisi akademik karena menggabungkan tiga pendekatan penting dalam satu kerangka kerja yang komprehensif. Pertama, *Gradient Boosting* digunakan sebagai algoritma utama untuk membangun model klasifikasi dengan performa tinggi. Kedua, SMOTE berperan sebagai strategi penyeimbangan data yang memastikan distribusi antar kelas tetap proporsional. Ketiga, SHAP dan LIME diterapkan untuk meningkatkan transparansi model sehingga hasil prediksi dapat dipahami secara klinis oleh tenaga kesehatan. Sinergi ketiga metode ini diharapkan dapat mengatasi dua persoalan

utama dalam penerapan *machine learning* di bidang medis, yaitu ketidakseimbangan data dan kurangnya interpretabilitas model.

1.2 Rumusan Masalah

Rumusan masalah yang akan dibahas pada penelitian ini:

1. Bagaimana algoritma *Gradient Boosting* dapat digunakan untuk mengklasifikasikan penyakit tuberkulosis berdasarkan parameter klinis dan riwayat pasien?
2. Bagaimana cara melakukan optimasi *hyperparameter Gradient Boosting* untuk memperoleh performa model terbaik dalam klasifikasi penyakit tuberkulosis?
3. Bagaimana interpretasi kontribusi variabel klinis dan riwayat pasien terhadap prediksi penyakit tuberkulosis menggunakan metode *Explainable AI* (XAI) berbasis SHAP dan LIME?

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini:

1. Penelitian ini menggunakan dataset rekam medis pasien tuberkulosis dari Puskesmas Gayamsari Semarang dengan variabel yang terbatas pada jenis kelamin, umur, demam, penurunan berat badan, hilang nafsu makan, sesak napas, sakit dada, lemas, lama batuk, dan jenis batuk, dengan variabel keluaran berupa status TBC.
2. Proses pemodelan menggunakan algoritma *Gradient Boosting* dengan optimasi *hyperparameter* yang difokuskan pada metode *Grid Search* dan OPTUNA, terbatas pada *hyperparameter* utama seperti *n_estimators*, *max_depth*, dan *learning_rate*.

3. Evaluasi kinerja model dibatasi pada metrik *Accuracy*, *Precision*, *Recall*, *F1-score*, *Specificity*, serta visualisasi menggunakan *ROC-AUC Curve* dan *Confusion Matrix*.
4. Interpretasi model hanya dilakukan dengan metode *Explainable AI* (XAI) berbasis SHAP dan LIME untuk mengetahui kontribusi masing-masing variabel terhadap prediksi TBC.

1.4 Tujuan Penelitian

Tujuan penelitian ini:

1. Mengimplementasikan algoritma *Gradient Boosting* untuk melakukan klasifikasi penyakit tuberkulosis berdasarkan parameter klinis dan riwayat pasien.
2. Melakukan optimasi *hyperparameter* pada model *Gradient Boosting* menggunakan *Grid Search* dan *OPTUNA* untuk memperoleh konfigurasi parameter terbaik dan mengevaluasi performa model *Gradient Boosting*.
3. Melakukan interpretasi hasil prediksi menggunakan *Explainable AI* (XAI) berbasis SHAP dan LIME untuk mengetahui kontribusi variabel klinis dan riwayat pasien terhadap klasifikasi penyakit tuberkulosis.