

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Tuberkulosis (TBC) merupakan salah satu penyakit menular mematikan yang tetap menjadi tantangan kesehatan global yang utama. Berdasarkan laporan *World Health Organization* (WHO) dalam *Global TB Report 2023*, TBC merupakan penyebab kematian kedua di dunia dari agen infeksi tunggal. Indonesia sendiri menempati peringkat kedua dengan beban kasus TBC tertinggi di dunia setelah India, dengan estimasi jumlah kasus mencapai 1.060.000 kasus pertahun dan angka kematian yang masih sangat tinggi (WHO, 2023).

Tuberkulosis paru dapat berkembang menjadi kondisi kronis yang fatal jika tidak terdeteksi sedini mungkin atau jika pengobatan dilakukan secara tidak konsisten (Fahdhienie & Darwis, 2023). Tuberkulosis paru merupakan penyakit menular yang disebabkan oleh *Mycobacterium tuberculosis* dan masih menjadi salah satu masalah kesehatan utama di Indonesia. Keterlambatan dalam mengenali gejala penyakit sering menyebabkan penanganan pasien tidak dilakukan secara optimal, sehingga berpotensi meningkatkan risiko penularan kepada masyarakat (Takwa, 2025). Oleh karena itu, diperlukan sebuah pendekatan berbasis data yang mampu mengidentifikasi pola gejala klinis secara cepat dan akurat sebagai pendukung skrining awal di fasilitas pelayanan kesehatan.

Dalam upaya melakukan diagnosis dini yang lebih efektif, peran teknologi informasi melalui pemanfaatan *Machine Learning* menjadi sangat krusial. Selama ini, tantangan terbesar bagi tenaga medis di lapangan adalah menganalisis sekumpulan gejala klinis yang kompleks dan sering kali tumpang tindih dengan penyakit pernapasan lainnya. Penerapan algoritma klasifikasi memungkinkan sistem untuk mempelajari pola dari data historis pasien TBC, sehingga mampu memberikan prediksi status kesehatan dengan tingkat presisi yang lebih tinggi dibandingkan penilaian subjektif (Mutiarra, 2020). Dengan penerapan sistem klasifikasi berbasis data, proses identifikasi awal pasien yang diduga menderita tuberkulosis paru dapat dilakukan secara lebih cepat dan objektif. Hasil klasifikasi

tersebut diharapkan dapat membantu tenaga medis dalam proses skrining awal sehingga penanganan dan pemeriksaan lanjutan dapat dilakukan secara lebih tepat waktu.

Penerapan *Machine Learning* dalam memprediksi status penyakit telah menunjukkan performa yang jauh lebih unggul dibandingkan model statistik tradisional karena kemampuannya dalam menangani pola data kesehatan yang kompleks dan non-linear (Mthethwa & Melesse, 2025). Dalam kasus tuberkulosis paru, data klinis sering kali melibatkan hubungan antar gejala dan faktor risiko yang tidak sederhana sehingga memerlukan model yang mampu menangkap interaksi variabel secara efektif. Salah satu algoritma yang banyak digunakan dalam klasifikasi medis adalah *Classification and Regression Trees* (CART), karena mampu membentuk aturan keputusan dalam struktur pohon yang mudah dipahami dan dapat membantu tenaga medis dalam menginterpretasikan hasil prediksi secara transparan (Hendri, 2025).

Meskipun demikian, CART sebagai model pohon tunggal memiliki keterbatasan, terutama rentan terhadap *overfitting* serta kurang optimal ketika dihadapkan pada dataset dengan variasi tinggi. Kondisi ini menjadi tantangan dalam klasifikasi penyakit tuberkulosis paru, mengingat data klinis sering kali bersifat beragam dan tidak sederhana. Sebagai alternatif, algoritma *ensemble learning* seperti *XGBOOST* (*Extreme Gradient Boosting*) muncul sebagai metode yang sangat efisien dalam menangani data yang hilang (*missing values*) serta kelas yang tidak seimbang (*imbalanced data*), yang merupakan karakteristik umum dalam dataset rekam medis (Abdillah et al., 2025). Penelitian oleh Zakaria et al. (2025) menunjukkan bahwa model berbasis ensemble seperti Random Forest dan *XGBOOST* memiliki kemampuan generalisasi yang lebih baik dibandingkan model pohon tunggal dalam memprediksi status penyakit tuberkulosis paru berdasarkan data klinis. Dengan demikian, jika *Random Forest* terbukti stabil, maka *XGBOOST* sebagai pengembangan dari pendekatan boosting diharapkan mampu memberikan akurasi dan performa yang lebih optimal dalam klasifikasi penyakit tuberkulosis paru dibandingkan model CART tunggal.

Dalam implementasinya, algoritma CART bekerja dengan membangun pohon keputusan melalui proses pemilihan atribut terbaik menggunakan nilai Gini Index. Setiap atribut akan dievaluasi untuk menentukan pembagian data (*splitting*) yang menghasilkan tingkat kemurnian kelas tertinggi. Proses tersebut dilakukan secara berulang hingga terbentuk node akhir (*leaf node*) yang merepresentasikan hasil klasifikasi. Sementara itu, algoritma *XGBOOST* membangun model secara bertahap melalui serangkaian *decision tree* yang saling melengkapi. Setiap pohon baru dibentuk berdasarkan kesalahan prediksi dari pohon sebelumnya sehingga model mampu memperbaiki performa secara iteratif. Pada penelitian ini, kedua algoritma akan diterapkan pada data pasien tuberkulosis paru yang telah melalui tahap preprocessing dan pembagian *data training* serta *testing*. Selanjutnya, hasil klasifikasi dari masing-masing metode akan dievaluasi dan dibandingkan berdasarkan nilai *Accuracy*, *Precision*, *Recall*, dan *F1-Score* untuk menentukan algoritma yang memiliki performa terbaik dalam mengklasifikasikan status penyakit tuberkulosis paru.

Perbandingan antara algoritma CART dan *XGBOOST* dalam penelitian ini dilakukan untuk mengetahui algoritma yang memberikan performa terbaik dalam klasifikasi penyakit tuberkulosis paru. Algoritma CART dipilih karena menghasilkan model yang sederhana dan mudah diinterpretasikan, sedangkan algoritma *XGBOOST* dipilih karena mampu membangun model prediksi melalui teknik *iterative boosting*. Kedua algoritma kemudian dievaluasi menggunakan *Accuracy*, *Precision*, *Recall*, dan *F1-Score* sehingga diperoleh informasi mengenai algoritma yang paling sering sesuai untuk diterapkan pada sistem klasifikasi penyakit tuberkulosis paru.

Penelitian ini menggunakan data rekam medis pasien Tuberkulosis (TB) Paru yang diperoleh dari RS Khusus Paru Sumatera Utara dengan izin penggunaan data untuk kepentingan penelitian. Dataset terdiri dari 3.732 data pasien yang dikumpulkan pada periode tahun 2024. Data yang digunakan memiliki atribut, yaitu umur, jenis kelamin, demam, batuk lebih dari 2 minggu, sesak napas, berat badan turun, keringat malam, nafsu makan turun, riwayat kontak TB, diabetes mellitus,

dan diagnosis TB Paru. Variabel diagnosis terdiri atas dua kelas, yaitu TB Paru positif dan TB negatif.

Penelitian ini bertujuan untuk mengembangkan model klasifikasi penyakit tuberkulosis paru menggunakan algoritma CART dan *XGBOOST* dengan membandingkan performa keduanya. Hasil penelitian diharapkan dapat memberikan kontribusi dalam pengembangan sistem pendukung keputusan klinis untuk deteksi dini tuberkulosis paru.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan diatas, maka rumusan permasalahannya adalah :

1. Bagaimana hasil klasifikasi penyakit tuberkulosis paru menggunakan algoritma CART berdasarkan nilai *Accuracy*, *Precision*, *Recall*, dan *F1-Score*?
2. Bagaimana hasil klasifikasi penyakit tuberkulosis paru menggunakan algoritma *XGBOOST* berdasarkan nilai *Accuracy*, *Precision*, *Recall*, dan *F1-Score*
3. Bagaimana hasil perbandingan kinerja algoritma CART dan *XGBOOST* dalam mengklasifikasikan penyakit tuberkulosis paru berdasarkan nilai *Accuracy*, *Precision*, *Recall*, dan *F1-score*?

1.3 Tujuan Penelitian

Adapun tujuan adanya penelitian ini adalah :

1. Membangun model klasifikasi untuk mendiagnosis penyakit Tuberkulosis Paru menggunakan metode CART dan *XGBOOST*.
2. Menganalisis perbandingan performa antara algoritma CART dan *XGBOOST* untuk menentukan model terbaik dalam klasifikasi TBC Paru.
3. Membantu mempercepat proses skrining awal penyakit Tuberkulosis Paru berdasarkan gejala yang dialami pasien.

1.4 Manfaat Penelitian

Adapun manfaat penelitian ini yaitu :

1. Dapat memberikan informasi kepada Rumah Sakit dalam mendiagnosis penyakit tuberkulosis paru.
2. Dapat membantu pasien yang terkena penyakit tuberkulosis paru dilihat dari gejala yang ditimbulkan dan segera mendapatkan pengobatan yang tepat.
3. Dapat digunakan sebagai referensi dalam melakukan penelitian selanjutnya menggunakan metode klasifikasi CART (*Classification and Regression Tree*) dan *XGBOOST*.

1.5 Ruang Lingkup dan Batasan Penelitian

Adapun batasan masalahnya adalah sebagai berikut :

1. Data yang digunakan meliputi data gejala klinis dan rekam medis pasien terkait Tuberkulosis Paru yang bersumber dari RS Khusus Paru Sumatera Utara.
2. Dataset yang digunakan dalam penelitian terdiri dari 3.732 data pasien yang digunakan sebagai data penelitian.
3. Penelitian hanya membahas klasifikasi penyakit Tuberkulosis Paru dan tidak mencakup klasifikasi jenis penyakit paru lainnya.
4. Algoritma *Machine Learning* yang digunakan dalam penelitian dibatasi pada dua algoritma, yaitu *Classification and Regression Tree* (CART) dan *Extreme Gradient Boosting* (XGBOOST).
5. Variabel target dalam penelitian ini adalah hasil klasifikasi penyakit TB Paru yang terdiri dari TB Paru Positif dan TB Paru Negatif.
6. Perbandingan kinerja CART dan XGBOOST dilakukan berdasarkan hasil nilai *Accuracy*, *Precision*, *Recall*, dan *F1-Score* yang diperoleh pada data pengujian.
7. Sistem hanya menggunakan atribut yang telah ditentukan dalam dataset penelitian sebagai dasar proses prediksi dan tidak menambahkan atribut baru di luar ruang lingkup penelitian.