

BAB I

PENDAHULUAN

1.1 LATAR BELAKANG MASALAH

Kanker paru-paru merupakan salah satu tantangan kesehatan terbesar di dunia dan dikenal sebagai penyebab utama kematian akibat kanker secara global. Berdasarkan data dari Organisasi Kesehatan Dunia (WHO), penyakit ini menyebabkan sekitar 1,8 juta kematian pada tahun 2020 [1], dan jumlah tersebut meningkat menjadi 2.480.675 jiwa, atau sekitar 12,4% dari total kematian akibat kanker, pada tahun 2022 [2]. Angka ini diperkirakan akan terus meningkat seiring dengan pertumbuhan populasi global dan perubahan gaya hidup. Deteksi dini kanker paru-paru sangat penting untuk meningkatkan tingkat kelangsungan hidup pasien. *American Cancer Society* melaporkan bahwa tingkat kelangsungan hidup 5 tahun untuk kanker paru-paru meningkat dari 18% menjadi 55% jika terdeteksi pada tahap awal [3].

Dalam upaya meningkatkan deteksi dini dan prediksi risiko kanker paru-paru, para peneliti telah mengidentifikasi berbagai faktor risiko dan gejala yang berkorelasi dengan penyakit ini. Atribut seperti jenis kelamin, usia, riwayat merokok, gejala fisik (seperti jari kuning, batuk, dan sesak napas), serta faktor psikologis dan lingkungan (seperti kecemasan dan tekanan dari teman sebaya) telah diidentifikasi sebagai indikator potensial dalam penilaian risiko kanker paru-paru [4]. Namun, kompleksitas

interaksi antara faktor-faktor ini membuat prediksi manual menjadi tantangan besar bagi para praktisi medis.

Dalam ranah prediksi risiko kanker paru-paru, tantangan utama terletak pada kemampuan mengidentifikasi pola-pola kompleks dari berbagai faktor klinis dan demografis. Serangkaian studi terdahulu telah mengeksplorasi potensi algoritma *machine learning* untuk mengatasi kompleksitas ini. Penelitian Putra, T. D., Utami, E., & Kurniawan, M. P. [5]. Menggunakan Algoritma *Artificial Neural Network* (ANN) untuk melakukan klasifikasi pasien kanker paru-paru dengan atribut dataset yang serupa, mencapai akurasi *training* 92.79% dengan presisi 86.98% dan akurasi *testing* sebesar 95.12% dengan presisi 90.23%. Sementara itu, Wulandari, E. [6]. melakukan klasifikasi menggunakan *Naive Bayes* dengan atribut dataset yang sama menghasilkan akurasi data test dengan tingkat persentasi *recall* 98,77%, persentasi *precision* 95,24% dan tingkat persentasi akurasinya sebesar 94,62%. Selain itu, Desiani dkk. [7] membandingkan kinerja SVM dan *K-Nearest Neighbor* (K-NN) dalam klasifikasi kanker paru-paru dan menemukan bahwa SVM dengan metode *percentage split* memberikan hasil terbaik, dengan akurasi 95,16%, presisi 88%, dan *recall* 82,5%. Namun, Penelitian yang secara spesifik berfokus membandingkan metode *random forest* dan *extra trees* belum ada.

Meskipun berbagai metode *machine learning* telah menunjukkan hasil yang menjanjikan, masih terdapat beberapa gap yang perlu diperhatikan. Pertama, dari jurnal yang paparkan atas bahwa studi berfokus pada satu metode *machine learning* tertentu, sehingga perbandingan langsung antara berbagai metode masih terbatas. Kedua,

interpretabilitas model, yang sangat penting dalam konteks medis, seringkali kurang diperhatikan dalam pengembangan model prediksi.

Untuk mengatasi gap-gap tersebut, penelitian ini mengusulkan pengembangan model komparatif yang membandingkan secara langsung metode *Random Forest* dan *Extra Trees* dalam konteks prediksi kanker paru-paru. Kedua metode ini dipilih karena kemampuannya dalam menangani dataset kompleks dengan berbagai jenis variable [8], seperti yang terdapat dalam atribut yang digunakan dalam penelitian ini (*gender, age, smoking, yellow_fingers, anxiety, peer_pressure, chronic disease, fatigue, allergy, wheezing, alcohol consuming, coughing, shortness of breath, swallowing difficulty, chest pain*).

Pemilihan *Random Forest* dan *Extra Trees* didasarkan pada beberapa pertimbangan. Pertama, kedua metode ini mampu menangani interaksi kompleks antar variabel, yang penting dalam konteks faktor risiko kanker paru-paru [9]. Kedua, keduanya menawarkan interpretabilitas yang lebih baik dibandingkan model "*black box*", memungkinkan identifikasi faktor-faktor kunci dalam prediksi [10]. Ketiga, *Random Forest* dan *Extra Trees* memiliki resistensi yang baik terhadap *overfitting*, *crucial* untuk generalisasi model pada data pasien baru [11]. Terakhir, perbedaan dalam proses pemilihan fitur dan pembagian node antara kedua metode dapat menghasilkan *insights* yang berharga dalam konteks prediksi kanker paru-paru [12].

Berdasarkan uraian permasalahan dan penelitian penelitian sebelumnya, maka penulis tertarik untuk melakukan penelitian guna memberi solusi terhadap masalah

yang terjadi dengan mengangkat judul “**Perbandingan Metode Klasifikasi *Random Forest* dan *Extra Trees* Dalam Memprediksi Kanker Paru.**”

Melalui tinjauan terhadap penelitian sebelumnya, dapat disimpulkan bahwa klasifikasi menggunakan *Machine Learning* sangat tepat dan sesuai untuk penelitian ini, Oleh karena itu pada penelitian ini akan di ujikan perbandingan klasifikasi algoritma dan *Naïve bayes* dengan parameter terdapat pada dataset.

1.2 RUMUSAN MASALAH

Berdasarkan uraian yang telah dijelaskan pada latar belakang, maka rumusan masalah pada penelitian ini adalah:

1. Bagaimana cara mengimplementasikan metode Random Forest dan Extra Tree dalam klasifikasi dataset kanker paru?
2. Bagaimana perbandingan hasil evaluasi antara implementasi metode Random Forest dan Extra Tree dalam klasifikasi dataset kanker paru?

1.3 BATASAN MASALAH

Agar dapat lebih fokus dan pembahasan tidak menyimpang dari permasalahan yang ada, maka penulis membatasi masalah sebagai berikut:

1. Penelitian ini hanya membandingkan metode *Random Forest* dan *Extra Trees* dalam konteks klasifikasi.
2. Data yang digunakan merupakan dataset publik yang diambil dari situs website yang bernama KAGGLE.
3. Tools yang digunakan dalam penelitian ini adalah Google Colab.

1.4 TUJUAN DAN MANFAAT PENELITIAN

Adapun tujuan dan manfaat dari penelitian ini yang dilakukan akan dijelaskan sebagai berikut:

1.4.1 TUJUAN PENELITIAN

Berdasarkan rumusan masalah yang didapat, maka tujuan penelitian ini adalah sebagai berikut :

1. Mengimplementasikan metode *Random Forest* dan *Extra Tree* dalam proses klasifikasi dataset kanker paru-paru.
2. Membandingkan hasil evaluasi kerja dari metode *Random Forest* dan *Extra Tree* dalam klasifikasi dataset kanker paru-paru untuk menentukan metode yang lebih efektif.

1.4.2 MANFAAT PENELITIAN

Adapun penelitian ini memiliki manfaat sebagai berikut :

1. Penelitian memberikan pemahaman yang lebih mendalam tentang performa dan keandalan dua metode klasifikasi berbasis *ensemble learning*, yakni *Random Forest* dan *Extra Trees*.
2. Penelitian ini dapat dijadikan referensi bagi studi-studi selanjutnya yang ingin mengembangkan atau membandingkan metode lain dalam *machine learning*.
3. Bagi penulis, dapat mengembangkan pengetahuan dan wawasan yang baru. melalui penelitian ini, mahasiswa akan belajar menangani data medis yang sensitif dan kompleks, serta memahami variabel-variabel yang relevan dalam prediksi penyakit.

4. Hasil perbandingan ini dapat menjadi referensi untuk penelitian lebih lanjut, terutama bagi peneliti yang tertarik menggunakan model klasifikasi dalam diagnosis penyakit.

Penelitian ini menunjukkan bagaimana teknik *machine learning* dapat diterapkan dalam dunia medis, khususnya dalam mendiagnosis penyakit serius seperti kanker, yang membuka peluang penggunaan teknologi yang lebih luas dalam bidang kesehatan.

1.4.3 SISTEMATIKA PENULISAN

Untuk memudahkan pemahaman terhadap isi laporan penelitian ini, maka disajikan struktur penulisan penelitian ini sebagai berikut:

BAB I :PENDAHULUAN

Bab ini menguraikan tentang latar belakang masalah, rumusan masalah,

Batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika

BAB II : LANDASAN TEORI

Landasan teori dalam penelitian ini mencakup beberapa konsep utama. Pertama, dibahas teori mengenai kanker paru-paru yang meliputi definisi, karakteristik klinis, dan faktor risiko penyakit. Selanjutnya, dijelaskan konsep *machine learning* dengan fokus pada algoritma klasifikasi dan penerapannya dalam analisis medis. Pembahasan dilanjutkan dengan eksplorasi metode *ensemble learning*, khususnya *Random Forest* dan *Extra Trees*, yang

mencakup struktur algoritma, mekanisme kerja, serta keunggulan masing-masing pendekatan. Terakhir, dijelaskan metode evaluasi model menggunakan metrik kinerja seperti akurasi, presisi, dan *recall* untuk memungkinkan analisis komparatif yang objektif antara algoritma klasifikasi dalam prediksi risiko kanker paru-paru.

BAB III : METODOLOGI PENELITIAN

Dalam bab ini berisi deskripsi bagaimana penelitian dilakukan, mencakup metode yang akan digunakan dalam penelitian.

BAB IV : HASIL DAN PEMBAHASAN

Pada bab ini berisi tentang gambaran umum tentang data yang akan di analisis dan bab ini membahas hasil analisis dari pengujian data yang dilakukan dan hasil yang dicapai dari pengujian data yang telah ditentukan.

BAB V : PENUTUP

Bab ini disajikan simpulan dari hasil penelitian dan juga saran-saran yang ditujukan kepada semua pihak yang bersangkutan.