

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi *Artificial Intelligence* (AI) dan *machine learning* saat ini mengalami kemajuan yang sangat pesat dan telah dimanfaatkan di berbagai bidang, salah satunya di bidang kesehatan. *Machine learning* merupakan bagian dari kecerdasan buatan yang memungkinkan komputer mempelajari pola dari data untuk menghasilkan prediksi atau keputusan tanpa harus diprogram secara eksplisit. Dalam bidang kesehatan, *machine learning* berperan penting dalam membantu tenaga medis, terutama pada proses deteksi dini penyakit, diagnosis, dan pengambilan keputusan medis [1]. Salah satu penerapan *machine learning* yang banyak digunakan di bidang kesehatan adalah metode klasifikasi. Metode klasifikasi bertujuan untuk mengelompokkan data ke dalam kelas tertentu berdasarkan karakteristik yang dimiliki [2]. Metode ini telah banyak diterapkan dalam klasifikasi berbagai penyakit, khususnya penyakit paru-paru seperti *pneumonia*, *tuberkulosis* (TBC), dan kanker paru-paru, dengan tingkat akurasi yang cukup tinggi [3]. Penggunaan metode klasifikasi dapat membantu tenaga medis dalam menganalisis data pasien secara lebih cepat dan konsisten dibandingkan dengan cara manual. *Tuberkulosis* (TBC) merupakan penyakit infeksi menular yang disebabkan oleh bakteri *Mycobacterium tuberculosis*. Penyakit ini terutama menyerang paru-paru, namun juga dapat menyerang organ lain seperti tulang, kelenjar getah bening, dan otak. Hingga saat ini, TBC masih menjadi salah satu masalah kesehatan utama di dunia dan menyebabkan lebih dari satu juta kematian setiap tahunnya [4].

Berdasarkan data Organisasi Kesehatan Dunia (WHO), TBC merupakan penyebab kematian terbesar kedua akibat penyakit menular di dunia. Secara global, penyakit ini menyebabkan sekitar 1,2 juta kematian setiap tahunnya dan tetap menjadi ancaman serius bagi kesehatan masyarakat, terutama di negara-

negara berkembang, termasuk Indonesia. Indonesia tercatat sebagai salah satu negara dengan beban kasus TBC tertinggi di dunia. Pada tahun 2023, Indonesia mencatat sekitar 1,09 juta penderita TBC. Kasus TBC di Indonesia menunjukkan tren peningkatan yang signifikan dalam beberapa tahun terakhir. Pada tahun 2020, tercatat 261.113 kasus TBC yang terkonfirmasi bakteriologis, dengan 157.764 kasus pada laki-laki dan 103.349 kasus pada perempuan. Meskipun terjadi penurunan pada tahun 2021 menjadi 173.221 kasus, angka kasus kembali meningkat drastis pada tahun-tahun berikutnya [5]. Kondisi ini menunjukkan bahwa upaya pengendalian dan deteksi dini TBC masih perlu ditingkatkan.

Permasalahan TBC tidak hanya terjadi di tingkat nasional, tetapi juga di tingkat daerah. Salah satu daerah dengan jumlah kasus TBC yang cukup tinggi adalah Kabupaten Timor Tengah Selatan (TTS), Provinsi Nusa Tenggara Timur (NTT). Berdasarkan data BPS TTS tahun 2021-2023, TBC selalu masuk dalam tiga besar penyakit menular dengan angka kejadian yang terus meningkat setiap tahun. Tercatat 242 kasus pada tahun 2017, meningkat menjadi 497 kasus pada tahun 2018, 468 kasus pada tahun 2019, dan mencapai 640 kasus pada tahun 2023. Dalam kurun waktu tersebut, TBC naik dari peringkat ketiga setelah diare dan malaria menjadi penyakit menular dengan kasus tertinggi kedua di tahun 2023 [6]. Data dari Dinas Kesehatan Kabupaten TTS juga menunjukkan jumlah kasus TBC dalam lima tahun terakhir, yaitu 367 kasus pada tahun 2021, 562 kasus pada tahun 2022, 629 kasus pada tahun 2023, 649 kasus pada tahun 2024, dan 300 kasus pada tahun 2025 (data hingga bulan Oktober). Kondisi ini menyebabkan proses diagnosis TBC sering mengalami keterlambatan karena pasien harus dirujuk ke rumah sakit yang memiliki fasilitas lebih lengkap. Gejala klinis TBC seperti batuk berdahak lebih dari dua minggu, demam berkepanjangan dan penurunan berat badan sering kali mirip dengan penyakit lain [7]. Hal ini menyebabkan diagnosis TBC menjadi sulit apabila hanya berdasarkan pemeriksaan klinis tanpa didukung pemeriksaan laboratorium. Di daerah dengan keterbatasan fasilitas kesehatan, kondisi ini dapat meningkatkan risiko keterlambatan diagnosis dan penularan penyakit.

Berdasarkan permasalahan tersebut, pemanfaatan *machine learning* sebagai alat bantu diagnosis TBC berdasarkan gejala klinis pasien menjadi salah satu solusi yang dapat dipertimbangkan. Dengan memanfaatkan data gejala klinis, algoritma *machine learning* diharapkan mampu membantu mengklasifikasikan apakah seorang pasien berpotensi menderita TBC atau tidak secara lebih cepat dan akurat. Namun, setiap algoritma *machine learning* memiliki karakteristik dan kinerja yang berbeda dalam proses klasifikasi. Dua algoritma yang sering digunakan dalam penelitian klasifikasi adalah *K-Nearest Neighbors* (KNN) dan *Random Forest*. Algoritma KNN bekerja berdasarkan kedekatan jarak antar data, sedangkan *Random Forest* merupakan metode berbasis pohon keputusan yang mampu menangani data dalam jumlah besar dan kompleks [8]. Beberapa penelitian menunjukkan bahwa kedua metode tersebut memiliki kelebihan dan kekurangan masing-masing dalam mengklasifikasikan data medis [9]. Oleh karena itu, penelitian ini dilakukan untuk membandingkan kinerja algoritma *K-Nearest Neighbors* (KNN) dan *Random Forest* dalam mengklasifikasikan penyakit TBC berdasarkan data gejala klinis pasien. Diharapkan hasil penelitian ini dapat memberikan informasi mengenai metode terbaik yang dapat digunakan sebagai alat bantu deteksi dini TBC, khususnya di daerah dengan keterbatasan fasilitas kesehatan seperti Kabupaten Timor Tengah Selatan.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut :

1. Bagaimana penerapan metode *K-Nearest Neighbors* (KNN) dan *Random Forest* dalam melakukan klasifikasi penyakit *Tuberkulosis* (TBC) berdasarkan gejala klinis pasien?
2. Metode manakah di antara *K-Nearest Neighbors* (KNN) dan *Random Forest* yang memiliki kinerja terbaik dalam mengklasifikasikan penyakit *Tuberkulosis* berdasarkan gejala klinis pasien?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, tujuan penelitiannya adalah:

1. Untuk mengetahui penerapan metode *K-Nearest Neighbors* (KNN) dan *Random Forest* dalam melakukan klasifikasi penyakit *Tuberculosis* (TBC) berdasarkan data gejala klinis pasien.
2. Untuk membandingkan kinerja metode *K-Nearest Neighbors* (KNN) dan *Random Forest* dalam mengklasifikasikan penyakit *Tuberculosis* (TBC) berdasarkan gejala klinis pasien, sehingga dapat diketahui metode yang memiliki kinerja terbaik.

1.4 Batasan Masalah

Untuk memfokuskan ruang lingkup penelitian dan menghindari pembahasan yang terlalu luas, maka batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Penelitian ini berfokus pada perbandingan metode KNN dan *Random Forest* untuk klasifikasi penyakit TBC berdasarkan gejala klinis pasien.
2. *Dataset* diperoleh dari data rekam medis pasien TBC dan Non TBC pasien RSUD Soe, Puskesmas Siso, Puskesmas Nulle dan Puskesmas Binaus Kabupaten Timor Tengah Selatan yang berjumlah 443 *record* yang digunakan untuk menerapkan dan membandingkan Metode KNN dan *Random Forest*.
3. Proses seleksi fitur dilakukan menggunakan metode *Feature Importance* berbasis model *Random Forest* untuk mengetahui fitur-fitur gejala klinis yang paling berpengaruh dalam klasifikasi penyakit TBC.
4. Fitur yang digunakan terbatas pada gejala klinis yang dapat diobservasi secara langsung, seperti batuk lebih dari 2 minggu, penurunan berat badan, demam, keringat malam, dan gejala klinis lainnya yang tercatat dalam rekam medis.
5. Model klasifikasi yang digunakan dalam penelitian ini yaitu *K-Nearest Neighbors* (KNN) dan *Random Forest* (RF) setelah itu dilakukan perbandingan dengan tujuan memperoleh model dengan kinerja terbaik

untuk klasifikasi penyakit TBC.

6. Pengujian kinerja masing-masing model klasifikasi dilakukan menggunakan empat metrik evaluasi, yaitu *Accuracy*, *Precision*, *Recall*, dan *F1-Score*, yang umum digunakan dalam evaluasi sistem klasifikasi pada bidang kesehatan.
7. Seluruh proses eksperimen dan implementasi model *Machine Learning* dilakukan menggunakan *Jupyter Notebook* dengan bahasa pemrograman *Python*.

1.5 Manfaat Penelitian

Model klasifikasi yang dibangun dalam penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Penelitian ini diharapkan dapat memberikan manfaat bagi tenaga medis dalam membantu mendiagnosis penyakit *Tuberkulosis* (TBC) secara lebih cepat dan akurat berdasarkan gejala klinis pasien.
2. Penelitian ini diharapkan dapat memberikan manfaat bagi peneliti dalam mengimplementasikan teori dan pengetahuan yang telah diperoleh selama perkuliahan, serta memberikan kontribusi terhadap pengembangan ilmu pengetahuan dan teknologi *machine learning*, khususnya dalam bidang kesehatan.
3. Penelitian ini diharapkan dapat meningkatkan kepercayaan masyarakat terhadap kualitas pelayanan kesehatan, terutama dalam upaya deteksi dini dan penanganan penyakit TBC.

1.6 Sistematika Penulisan

Laporan Tugas Akhir ini ditulis menggunakan sistematika penulisan sebagai berikut.

BAB I PENDAHULUAN

Bab ini memuat latar belakang penelitian yang menjelaskan alasan dan pentingnya penelitian dilakukan, rumusan masalah, tujuan penelitian, batasan masalah, serta manfaat penelitian. Selain itu, pada bab ini juga disajikan

sistematika penulisan serta daftar istilah dan singkatan yang digunakan dalam laporan penelitian.

BAB II LANDASAN TEORI

Bab ini memuat landasan teori dan penelitian-penelitian terdahulu yang relevan dengan penelitian yang dilakukan. Pembahasan meliputi perbandingan penelitian sebelumnya serta teori-teori pendukung yang berkaitan dengan penyakit *Tuberkulosis* (TBC), *Machine Learning*, metode klasifikasi, seleksi fitur, algoritma *K-Nearest Neighbors* (KNN), *Random Forest*, dan evaluasi kinerja model.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan alur pelaksanaan penelitian secara sistematis, mulai dari proses pengumpulan data, pra-pemrosesan data (*pre-processing*), seleksi fitur, penerapan model klasifikasi KNN dan *Random Forest*, hingga metode evaluasi kinerja model dan tahapan memperoleh hasil penelitian.

BAB IV HASIL DAN PEMBAHASAN

Bab ini membahas hasil penerapan dan perbandingan metode *K-Nearest Neighbors* (KNN) dan *Random Forest* dalam klasifikasi penyakit *Tuberkulosis* (TBC) berdasarkan gejala klinis pasien. Pembahasan meliputi hasil implementasi masing-masing metode, analisis performa berdasarkan metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score*, serta perbandingan kinerja kedua model untuk mengetahui kelebihan dan kekurangan masing-masing algoritma. Hasil perbandingan pada bab ini digunakan sebagai dasar dalam menentukan model dengan kinerja terbaik, yang selanjutnya akan dilakukan pengujian lanjutan pada Bab V.

BAB V PENGUJIAN DAN ANALISIS HASIL

Bab ini memuat hasil proses eksperimen dan pengujian lanjutan terhadap model yang memiliki kinerja paling baik berdasarkan hasil perbandingan pada bab

sebelumnya. Pengujian dilakukan menggunakan data uji, disertai dengan analisis hasil klasifikasi berdasarkan metrik evaluasi untuk menilai keandalan dan akurasi model terpilih.

BAB VI PENUTUP

Bab ini memuat kesimpulan yang diperoleh dari hasil penelitian untuk menjawab rumusan masalah dan tujuan penelitian. Selain itu, disajikan pula saran-saran yang dapat digunakan sebagai bahan pertimbangan untuk penyempurnaan dan pengembangan penelitian selanjutnya.

1.7 Daftar Istilah

Daftar istilah dapat dilihat pada Tabel 1.1.

Tabel 1. 1 Daftar Istilah

Istilah/Singkatan	Keterangan
<i>Machine Learning</i>	Bidang kecerdasan buatan (AI) yang membuat sistem mampu belajar dari data dan membuat prediksi.
<i>Random Forest</i> (RF)	Algoritma klasifikasi berbasis kumpulan (<i>ensemble</i>) banyak pohon keputusan untuk meningkatkan akurasi prediksi.
<i>K-Nearest Neighbors</i> (KNN)	Algoritma yang menentukan kelas suatu data baru berdasarkan kedekatan jarak dengan data yang sudah diketahui kelasnya.

<i>Feature Selection</i>	Proses memilih atribut atau variabel yang paling relevan terhadap hasil prediksi.
<i>Feature Importance</i>	Nilai yang menunjukkan seberapa besar pengaruh suatu fitur terhadap hasil klasifikasi.
<i>Dataset</i>	Kumpulan data yang digunakan untuk melatih dan menguji model.
<i>Pre-processing</i>	Langkah awal pengolahan data mentah agar siap digunakan dalam proses analisis.
<i>Cleaning</i>	Proses menghapus atau memperbaiki data yang tidak lengkap, salah, atau tidak konsisten.
<i>Transformation</i>	Proses mengubah bentuk data agar sesuai dengan format yang diperlukan oleh model.
<i>Normalization / Normalisasi</i>	Proses menyamakan skala antar atribut numerik agar perbandingan menjadi adil.
<i>Z-Score Normalization</i>	Metode normalisasi berdasarkan rata-rata dan standar deviasi.
<i>Training Data</i>	Data yang digunakan untuk melatih model.

<i>Testing Data / Data Uji</i>	Data yang digunakan untuk menguji performa model setelah pelatihan.
<i>K-Fold Cross Validation</i>	Teknik evaluasi model dengan membagi data menjadi k bagian untuk pelatihan dan pengujian secara bergantian.
<i>Confusion Matrix</i>	Tabel evaluasi performa model klasifikasi yang menunjukkan jumlah prediksi benar dan salah.
<i>Accuracy</i>	Ukuran yang menunjukkan persentase prediksi benar dari seluruh data uji.
<i>Precision</i>	Ukuran ketepatan model dalam memprediksi data positif dengan benar.
<i>Recall</i>	Ukuran kemampuan model dalam menemukan semua data positif sebenarnya.
<i>F1-Score</i>	Rata-rata harmonik dari <i>precision</i> dan <i>recall</i> untuk menilai kinerja model secara seimbang.
<i>Evaluation Metrics</i>	Ukuran atau metrik yang digunakan untuk menilai performa model.
<i>Overfitting</i>	Kondisi ketika model terlalu cocok dengan data latih sehingga tidak akurat pada data baru.

<i>Classification Model</i>	Model yang digunakan untuk mengelompokkan data ke dalam kategori tertentu.
<i>Algorithm</i>	Langkah-langkah sistematis atau prosedur yang digunakan untuk menyelesaikan suatu permasalahan.
<i>Euclidean Distance</i>	Rumus jarak antara dua titik dalam ruang fitur digunakan pada algoritma KNN.
<i>Flowchart</i>	Diagram alur yang menggambarkan tahapan proses penelitian atau sistem.
<i>Python</i>	Bahasa pemrograman yang digunakan untuk membangun dan menguji model <i>machine learning</i> .
<i>Jupyter Notebook</i>	Aplikasi interaktif berbasis web untuk menjalankan kode <i>Python</i> dalam analisis data.
<i>Library</i>	Kumpulan fungsi atau modul siap pakai dalam <i>Python</i> , seperti <i>scikit-learn</i> , <i>pandas</i> , dan <i>matplotlib</i> .
<i>Model Evaluation</i>	Tahap penilaian kinerja model untuk mengetahui seberapa baik model bekerja pada data uji.
<i>Null</i>	Nilai kosong atau tidak terisi dalam <i>dataset</i> .

<i>Modus</i>	Nilai yang paling sering muncul dalam suatu data.
<i>Imputation</i>	Teknik mengisi nilai yang hilang dalam dataset.
<i>Mapping</i>	Proses mengganti nilai data sesuai aturan tertentu.
<i>Looping</i>	Perulangan perintah dalam pemrograman.
<i>DataFrame</i>	Struktur data berbentuk tabel (baris dan kolom) untuk mengelola data.
<i>Outlier</i>	Data yang nilainya menyimpang jauh dari data lainnya.
<i>Imbalance</i>	Kondisi jumlah data antar kelas tidak seimbang.
<i>Synthetic Minority Over-sampling Technique (SMOTE)</i>	mengatasi masalah ketidakseimbangan data pada proses klasifikasi.
<i>Resampling</i>	Proses pengambilan sampel ulang dari <i>dataset</i> .
<i>Macro</i>	Metode rata-rata evaluasi model dengan menghitung setiap kelas secara terpisah lalu dirata-ratakan