Forum Analisis Statistik
Desember 2025, 5 (2): 113-124

Perbandingan Metode *Decision Tree* dan *Naïve Bayes* dalam *Data Mining* untuk Klasifikasi Data Penyakit *Tuberculosis* di Indonesia

Comparison of Decision Tree and Naïve Bayes in Data Mining for Classifying Tuberculosis Disease Data in Indonesia

Putri Dewi Febrianti[1*], Alissa Chintyana[2], Ristu Haiban Hirzi[3]

[1,2,3] Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hamzanwadi
Jl. Cut Nyak Dien No.85, Nusa Tenggara Barat
*Penulis Korespondensi. e-mail: putridf.210304021@student.hamzanwadi.ac.id

ABSTRACT

Tuberculosis (TB) remains a serious health problem in Indonesia, with high case and death rates. Indonesia has the second-highest number of TB cases in the world after India, accounting for approximately 10% of global cases. Data from the Indonesian Ministry of Health indicate that the number of TB cases is estimated to exceed one million per year, with hundreds of thousands of deaths. Therefore, better data analysis is needed to support TB control and prevention strategies. This study aims to classify the spread of TB in various districts/cities in five provinces with the highest TB cases in Indonesia by comparing the Decision Tree and Naïve Bayes methods using TB data from 2021 to 2024. The dataset includes variables such as age, population density, and health services. The research stages include data pre-processing, converting numerical data to categorical data, and validation using K-Fold Cross Validation. Model performance was evaluated using a Confusion Matrix. The results showed that the Decision Tree method had a higher accuracy rate (76.63%) than the Naïve Bayes method (63.27%). The superiority of Decision Tree in identifying age as the main predictor shows its ability to model complex TB data patterns.

Keywords: Data Mining, Classification, Decision tree, Naïve Bayes, Tuberculosis.

ABSTRAK

Tuberkulosis (TB) tetap menjadi masalah kesehatan serius di Indonesia, dengan angka kasus dan kematian yang tinggi. Indonesia memiliki jumlah kasus TB tertinggi kedua di dunia setelah India, menyumbang sekitar 10% dari kasus global. Data dari Kementerian Kesehatan Indonesia menunjukkan bahwa jumlah kasus TB diperkirakan melebihi satu juta per tahun, dengan ratusan ribu kematian. Oleh karena itu, analisis data yang lebih baik diperlukan untuk mendukung strategi pengendalian dan pencegahan TB. Studi ini bertujuan untuk mengklasifikasikan penyebaran TB di berbagai kabupaten/kota di lima provinsi dengan kasus TB tertinggi di Indonesia menggunakan perbandingan metode *Decision Tree* dan *Naïve Bayes* dengan menggunakan data TB dari tahun 2021 hingga 2024. Dataset mencakup variabel seperti usia, kepadatan penduduk, dan layanan kesehatan. Tahapan penelitian meliputi pra-pemrosesan data, konversi data numerik ke data kategorikal, dan validasi menggunakan *K-Fold Cross Validation*. Kinerja model dievaluasi menggunakan Matriks Konfusi. Hasil penelitian menunjukkan bahwa metode *Decision Tree* memiliki tingkat akurasi yang lebih tinggi (76,63%) dibandingkan metode *Naïve Bayes* (63,27%), Keunggulan *Decision Tree* dalam mengidentifikasi usia sebagai prediktor utama menunjukkan kemampuannya dalam memodelkan pola data TB yang kompleks.

Kata kunci: Data Mining, Decision Tree, Klasifikasi, Naïve Bayes, Tuberkulosis.

PENDAHULUAN

Kesehatan merupakan salah satu parameter utama dalam menilai kualitas sumber daya manusia di suatu negara. Saat ini, Indonesia masih menghadapi berbagai permasalahan kesehatan, baik penyakit menular maupun tidak menular. Perubahan gaya hidup menyebabkan penyakit tidak menular seperti diabetes melitus, penyakit jantung, dan kanker menjadi penyumbang utama biaya kesehatan nasional. Namun demikian, penyakit menular seperti Tuberkulosis (TB) masih memerlukan perhatian serius karena prevalensinya yang tinggi serta dampaknya yang luas terhadap kesehatan masyarakat, khususnya pada kelompok usia produktif. TB bersifat menular dan berpotensi menyebar dengan cepat, terutama di wilayah dengan kepadatan penduduk tinggi, sehingga menjadi ancaman kesehatan masyarakat yang signifikan (Muntasir & Weraman, 2025) .

Tuberkulosis (TB) adalah infeksi menular yang disebabkan oleh bakteri *Mycobacterium tuberculosis yang* umumnya menyerang paru-paru dan dapat menyebar ke organ lain melalui aliran darah (Lina Yunita *et al*., 2023). Data dari Kementerian Kesehatan RI menunjukkan bahwa jumlah kasus baru tuberkulosis (TB) di Indonesia diperkirakan mencapai lebih dari satu juta kasus per tahun dengan angka kematian yang masih tinggi. Pada tahun 2024, penemuan kasus TB mencapai 81% dari total estimasi kasus, dengan sebagian besar penderita telah mendapatkan pengobatan. Provinsi dengan jumlah kasus tertinggi meliputi Jawa Barat, Jawa Timur, Jawa Tengah, Sumatera Utara, dan DKI Jakarta (Kemenkes RI, 2024;2025).

Tingginya angka kasus TB di Indonesia menunjukkan perlunya analisis lebih mendalam terhadap faktor-faktor yang memengaruhi penyebaran penyakit ini. Beberapa variabel yang diduga berhubungan dengan kejadian TB antara lain usia, akses terhadap layanan kesehatan, dan kepadatan penduduk. Kelompok usia lanjut cenderung lebih rentan akibat penurunan sistem imun, sementara keterbatasan akses layanan kesehatan serta kepadatan penduduk yang tinggi dapat mempercepat penularan penyakit di masyarakat (Laporan Tuberkulosis Global, 2023). Berdasarkan permasalahan tersebut, diperlukan metode klasifikasi untuk membantu pengambilan keputusan yang lebih tepat dalam upaya pencegahan dan pengendalian TB. Klasifikasi merupakan proses pengelompokan data ke dalam kategori tertentu berdasarkan model yang dibangun dari data pelatihan (Depari *et al*., 2022) .

Klasifikasi *Decision Tree* adalah algoritma efektif yang digunakan dalam klasifikasi atau prediksi. Algoritma ini berupaya meningkatkan akurasi dengan menghilangkan cabang pohon yang mencerminkan *noise* dalam data. *Decision Tree* menawarkan berbagai kemampuan untuk memecahkan masalah, dengan mempertimbangkan elemen-elemen yang dapat memengaruhi pilihan dan memberikan hasil estimasi untuk setiap alternatif (Rokhman et al., 2021). Pengklasifikasi *Naïve Bayes* adalah metode dalam *Data Mining* untuk mengelompokkan informasi. Mekanisme metode Pengklasifikasi *Naïve Bayes* bergantung pada perhitungan probabilitas. *Naïve Bayes* adalah algoritma yang ada dalam teknik klasifikasi (Zulfauzi & Alamsyah, 2020) .

Beberapa penelitian sebelumnya menunjukkan bahwa *Decision Tree* dan *Naïve Bayes* memiliki tingkat akurasi yang baik pada berbagai studi kasus, khususnya pada analisis sentimen. Namun, kajian yang secara khusus membandingkan kedua metode tersebut pada kasus tuberkulosis di Indonesia masih terbatas terutama pada tingkat kabupaten/kota dan menggunakan data terbaru. Oleh karena itu, tujuan penelitian ini adalah untuk membandingkan dan mengklasifikasikan penyebaran tuberkulosis (TB) di kabupaten/kota pada lima provinsi dengan jumlah kasus tertinggi menggunakan metode *Decision Tree* dan *Naïve Bayes* berdasarkan data tahun 2021–2024. Hasil penelitian ini diharapkan dapat memberikan kontribusi bagi pemerintah dan sektor kesehatan dalam mendukung pengambilan keputusan serta meningkatkan strategi pencegahan dan pengendalian tuberkulosis di Indonesia.

METODOLOGI

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari situs resmi Badan Pusat Statistik (BPS) dengan rentang waktu 2021–2024. Unit analisis penelitian adalah kabupaten/kota pada lima provinsi yaitu Jawa Barat, Jawa Timur, Jawa Tengah, Sumatera Utara, dan DKI Jakarta dengan jumlah kasus tuberkulosis tertinggi di Indonesia. Variabel respon berupa status tuberkulosis (TB), yang didefinisikan sebagai hasil kategorisasi tingkat kejadian TB pada setiap kabupaten/kota sehingga memiliki skala nominal. Variabel penjelas meliputi usia penduduk, layanan kesehatan, dan kepadatan penduduk yang pada awalnya berskala numerik. Untuk menyesuaikan dengan metode klasifikasi *Decision Tree* dan *Naïve Bayes*, variabel numerik tersebut dikonversi ke dalam bentuk kategorik menggunakan nilai *cut-off* berdasarkan distribusi data, seperti pembagian kuartil. Dengan demikian, unit analisis, proses konversi data, dan definisi operasional variabel telah dijelaskan secara jelas dan konsisten, sebagaimana disajikan pada Tabel 1.

Tabel 1. Deskripsi Variabel

Variabel	Keterangan	Satuan
Y	Status Tuberkulosis (hasil kategorisasi positif negatif tingkat kejadian TB pada kabupaten/kota)	Nominal
X_1	Usia Penduduk	Tahun
X_2	Layanan Kesehatan (Jumlah Fasilitas kesehatan)	Unit
X_3	Kepadatan Penduduk	Jiwa/km^2

Sumber: Hasil Penelitian (2025)

2.1. Data Mining

Data Mining, juga dikenal sebagai Penemuan Pengetahuan dalam Basis Data (*Knowledge Discovery in Databases*/KDD), adalah proses yang melibatkan pengumpulan informasi dan penggunaan data historis untuk menemukan wawasan, informasi, struktur, pola, atau hubungan dalam kumpulan data yang sangat besar (Listanto & Kristania, 2022) . Secara umum, ada lima peran dalam *Data Mining*: estimasi, prediksi, klasifikasi, pengelompokan, dan asosiasi (Widaningsih et al., 2022). Beberapa tahapan terlibat dalam proses *Data Mining*, termasuk:

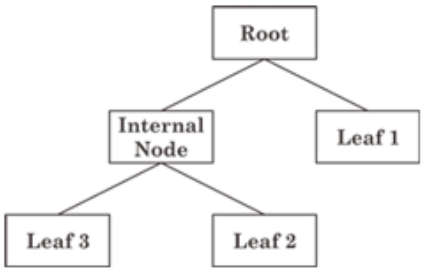
a. Seleksi data: Proses memilih data yang akan digunakan dalam proses *Data Mining*.
b. Praproses data: Melakukan praproses data, seperti membersihkan data dari *noise* dan *outlier*, mengisi nilai yang hilang, dan melakukan transformasi data.
c. Pemodelan: Membangun model untuk menemukan pola atau informasi dari data.
d. Evaluasi Model: Mengevaluasi model yang telah dibangun untuk melihat seberapa baik model tersebut dapat menemukan pola atau informasi dari data.
e. Interpretasi Hasil: Menginterpretasikan hasil dari proses *Data Mining* dan mengekstrak pengetahuan baru dari data tersebut.

2.2. Algoritma *Decision Tree* C4.5

Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang diperkenalkan oleh Quinlan pada tahun 1996. Keunggulan utama C4.5 dibandingkan ID3 terletak pada kemampuannya dalam menangani atribut bertipe numerik, melakukan proses pemangkasan (*pruning*) pada pohon keputusan, serta menghasilkan aturan keputusan yang lebih optimal. *Decision Tree* sendiri merupakan metode klasifikasi yang membangun model prediksi dalam bentuk struktur pohon berdasarkan atribut-atribut yang memiliki kemampuan terbaik dalam membedakan kelas data (Riyadi *et al*., 2022).

Struktur *Decision Tree* terdiri atas tiga jenis node, yaitu node akar sebagai simpul awal pohon, node internal sebagai titik percabangan berdasarkan atribut tertentu, dan node daun sebagai simpul

terminal yang merepresentasikan hasil klasifikasi (Kalimah, 2022) . Setiap percabangan dalam pohon menunjukkan hasil pengujian nilai atribut yang digunakan dalam proses pengambilan keputusan, di mana jawaban ya dan tidak mewakili hasil positif dan negatif (Prasitpuriprecha dkk., 2023) . Berikut adalah contoh visual *Decision Tree* (Nti, 2021) .



Gambar 1. Struktur *Decision Tree*
Sumber: Nti (2021)

1) Node akar adalah akar atau node teratas dari pohon, tanpa input dan mungkin tidak memiliki output atau memiliki lebih dari satu output.
2) Node internal adalah node yang berfungsi sebagai titik percabangan, dengan hanya satu input dan setidaknya dua output.
3) Node daun, atau node terminal, adalah node terakhir dalam *decision tree*, dengan hanya satu input dan tanpa output.

Saat menerapkan algoritma C4.5, ada beberapa langkah yang perlu dilakukan untuk mengubah tabel data menjadi *Decision Tree* (Suntoro, 2019). Langkah-langkah tersebut adalah sebagai berikut:

a) Hitung jumlah titik data.
b) Pilih atribut yang akan digunakan sebagai node.
c) Buat cabang untuk setiap anggota node.
d) Periksa apakah ada nilai entropi anggota node yang bernilai nol. Jika ditemukan, tetapkan daun yang dihasilkan. Jika semua nilai entropi anggota node bernilai nol, proses dianggap selesai.
e) Jika ada anggota node yang memiliki entropi lebih besar dari nol, ulangi proses dari awal dengan node yang dibutuhkan hingga semua anggota node mencapai nol.

Dalam algoritma C4.5, beberapa perhitungan diperlukan untuk membuat *Decision Tree*. Algoritma C4.5 menggunakan perhitungan entropi dan rasio perolehan dalam proses pemilihan atribut sebagai node (Suntoro , 2019).

1. Rumus entropi adalah sebagai berikut: di mana S adalah himpunan kasus, n adalah jumlah partisi S, dan pi adalah proporsi himpunan kasus ke-i terhadap himpunan (Suntoro, 2019). Rumus untuk menghitung entropi adalah sebagai berikut. Dengan persamaan di bawah ini:

$$Entrophy\ (S) = -\sum_{i=0}^{n} - pi\ * log^2\ pi \quad (1)$$

2. Selanjutnya adalah rumus untuk perolehan informasi, dapat dilihat bahwa S adalah himpunan kasus, A adalah atribut, n adalah jumlah partisi atribut A, Si adalah proporsi Si terhadap S, |S | adalah jumlah kasus dalam S, dan entropi (Si) adalah entropi untuk sampel yang memiliki nilai ke-i (Suntoro, 2019).

$$Gain\ (S, A) = Entrophy\ (S) - \sum |si|\,|s| * Entrophy\ (Si) \quad (2)$$

2.3. Pengklasifikasi *Naïve Bayes*

Pengklasifikasi *Naïve Bayes* adalah metode klasifikasi yang didasarkan pada teorema Bayes. Teknik ini bergantung pada prinsip-prinsip probabilitas dan statistik yang diperkenalkan oleh ilmuwan Inggris Thomas Bayes, yaitu memperkirakan kemungkinan masa depan berdasarkan pengalaman sebelumnya, sehingga dikenal sebagai Teorema Bayes (Rayuwati *et al.,* 2022). Menurut Han & Kamber (2000), proses yang dilakukan dalam algoritma *Naïve Bayes* adalah sebagai berikut:

1) Hitung jumlah kelas dalam data pelatihan.
2) Hitung jumlah kemunculan di setiap kelas.
3) Hitung probabilitas setiap fitur untuk setiap kelas (*likelihood*).
4) Kalikan nilai probabilitas fitur-fitur dalam suatu kelas.
5) Bandingkan hasil probabilitas antar kelas dan pilih kelas dengan nilai probabilitas tertinggi sebagai hasil klasifikasi.

Rumus Teorema *Naïve Bayes* adalah sebagai berikut (Han & Kamber , 2000):

$$P(C_k|x)\frac{P(C_k)P(x|C_k)}{P(x)} \quad (3)$$

Di mana:

P($Ck \mid x$) : probabilitas posterior, probabilitas kelas x (misalnya, positif/negatif) berdasarkan fitur x.

P(Ck) : probabilitas awal, probabilitas awal kelas x sebelum melihat data fitur. Dihitung dari proporsi data di setiap kelas.

P($x \mid Ck$) : Kemungkinan, probabilitas melihat fitur x dengan asumsi bahwa 14 data berasal dari kelas x. Dihitung dari frekuensi fitur tertentu dalam kelas tersebut.

P(x) : Bukti (Probabilitas Marjinal), probabilitas melihat data x secara keseluruhan.

Biasanya diabaikan karena sama untuk semua kelas. Algoritma *Naïve Bayes* menyatakan bahwa setiap fitur dalam data independen dari fitur lainnya, sehingga kemungkinan dapat dihitung sebagai hasil perkalian probabilitas setiap fitur terhadap kelas tersebut, teorema Bayes $P\,(x \mid Ck)$ disederhanakan menjadi (Han & Kamber, 2000):

$$P(x|C_k) = \prod_{i=1}^{n} P(x_i|C_k) \quad (4)$$

Di mana:

x : x 1, x 2,…, xn data dengan n fitur.

C k : kelas ke k.

P($xi \mid Ck$) : probabilitas fitur ke-i diberi kelas Ck.

Π : mewakili perkalian semua fitur.

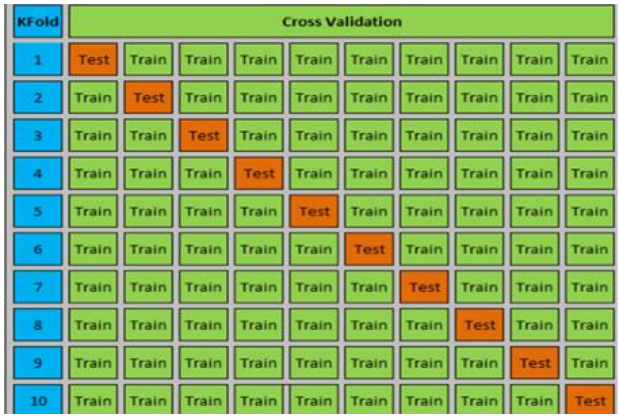
Pada penelitian ini, probabilitas posterior P($Ck|x$) pada algoritma *Naïve Bayes* diinterpretasikan sebagai peluang suatu kabupaten/kota termasuk ke dalam kelas status tuberkulosis tertentu (misalnya tingkat kejadian TB tinggi atau rendah) berdasarkan kombinasi variabel penjelas yang digunakan, yaitu usia penduduk, ketersediaan layanan kesehatan, dan kepadatan penduduk. Probabilitas ini dihitung dengan mengalikan probabilitas awal kelas TB dengan probabilitas kemunculan masing-masing variabel penjelas pada kelas tersebut. Kelas dengan nilai probabilitas posterior tertinggi kemudian dipilih sebagai hasil klasifikasi status TB pada setiap kabupaten/kota.

Algoritma *Naïve Bayes* mengasumsikan bahwa setiap variabel penjelas bersifat independen secara kondisional terhadap kelas TB. Artinya, pengaruh usia penduduk, layanan kesehatan, dan kepadatan penduduk terhadap status TB diasumsikan tidak saling bergantung satu sama lain. Asumsi ini merupakan keterbatasan metode *Naïve Bayes*, mengingat dalam konteks epidemiologi tuberkulosis, variabel-variabel tersebut secara empiris berpotensi saling berkaitan. Meskipun demikian, *Naïve Bayes* tetap banyak digunakan karena kesederhanaannya, efisiensi komputasi yang tinggi, serta

kemampuannya menghasilkan kinerja klasifikasi yang baik pada berbagai studi kasus, termasuk pada data kesehatan.

2.4. K-Fold Cross-Validation

K-Fold Cross-Validation adalah metode statistik yang diterapkan untuk menilai kinerja model yang akan dibuat. Dalam metode ini, dataset dibagi menjadi k bagian (*fold*) yang memiliki ukuran yang sama (Fuadah *et al.,* 2022). Selain itu, *K-Fold Cross-Validation* juga merupakan cara yang efisien untuk mengintegrasikan atribut parameter dari pembelajaran mesin dalam proses pelatihan model prediksi yang lebih baik dan rata-rata dari semua hasil iterasi digunakan sebagai estimasi kinerja akhir (Nugroho *et al*., 2023).



Gambar 2. Skema 10 *Fold Cross-Validation*
Sumber: Nugoroho et.al (2021)

2.5. Confusion Matrix

Confusion matrix adalah tabel yang digunakan untuk menganalisis seberapa akurat metode klasifikasi memprediksi kelas data (Gifari *et al*., 2022)s. *Confusion matrix* ini menggunakan tabel matriks seperti pada Tabel 2, yang menunjukkan perbandingan hasil klasifikasi dari data pengujian yang dihasilkan berdasarkan data pelatihan dengan data aktual (Supriyadi, 2023). Tabel *confusion matrix* dan rumus perhitungannya dapat dilihat di bawah ini:

Tabel 2. *Confusion matrix*

Prediksi Kelas		Kelas Hasil Prediksi (j)	
		Negatif	Positif
Kelas Asli (i)	Negatif	TN	FN
	Positif	FP	TP

Sumber: Supriyadi (2023)

Untuk mengukur performa suatu model pada penelitian ini menggunakan nilai akurasi dengan rumus sebagai berikut:

$$Akurasi = \frac{jumlah\ data\ yang\ diprediksi\ benar}{jumlah\ prediksi\ yang\ dilakukan} = \frac{(TP+TN)}{(TP+TN+FN+FP)} \times 100\% \quad (5)$$

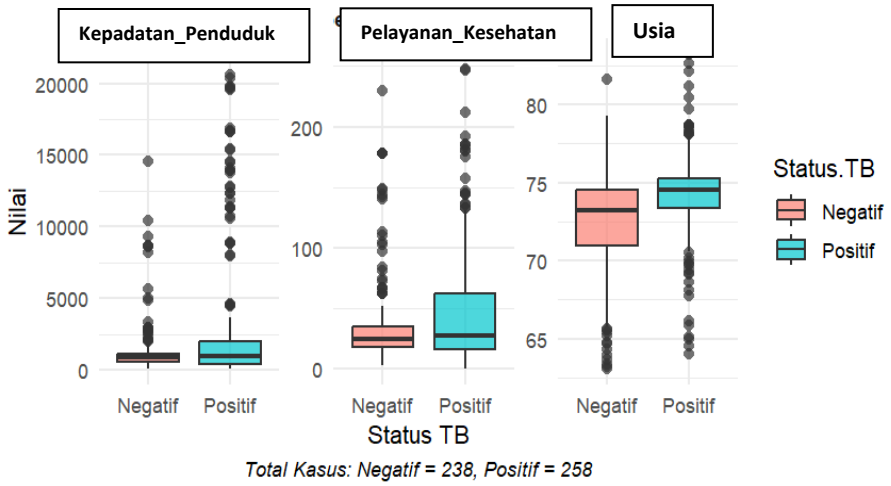
Berdasarkan tabel 2, *true positive* adalah jumlah dari *record* positif yang diklasifikasikan sebagai positif, *false positive* adalah jumlah dari *record* negatif yang diklasifikasikan sebagai positif,

false negative adalah jumlah dari *record* positif yang diklasifikasikan sebagai negatif, *true negative* adalah jumlah dari *record* negatif yang diklasifikasikan sebagai negatif.

HASIL DAN PEMBAHASAN

3.1. Analisis Deskriptif

Distribusi nilai dari tiga variabel prediktor utama, yaitu kepadatan penduduk, layanan kesehatan, dan usia, berdasarkan klasifikasi status TB (Negatif dan Positif).



Gambar 3. *Boxplot* dari Setiap Variabel Prediktor
Sumber: Hasil Penelitian (2025)

Gambar 3 memperlihatkan perbedaan distribusi tiga variabel prediktor utama, yaitu kepadatan penduduk, layanan kesehatan, dan usia, terhadap status TB. Kelompok dengan status TB positif cenderung memiliki kepadatan penduduk yang lebih tinggi, yang mengindikasikan meningkatnya risiko penularan di lingkungan padat (WHO, 2023). Sebaliknya, kelompok TB negatif menunjukkan tingkat layanan kesehatan yang lebih baik, menandakan peran akses layanan dalam pencegahan dan deteksi dini TB. Selain itu, distribusi usia menunjukkan bahwa kasus TB lebih banyak terjadi pada kelompok usia lanjut, yang berkaitan dengan penurunan sistem imun dan peningkatan komorbiditas.

3.2. Klasifikasi Status Penyakit Tuberkulosis dengan *Decision Tree*

Tabel 3. CP Nilai (Parameter Kompleksitas)

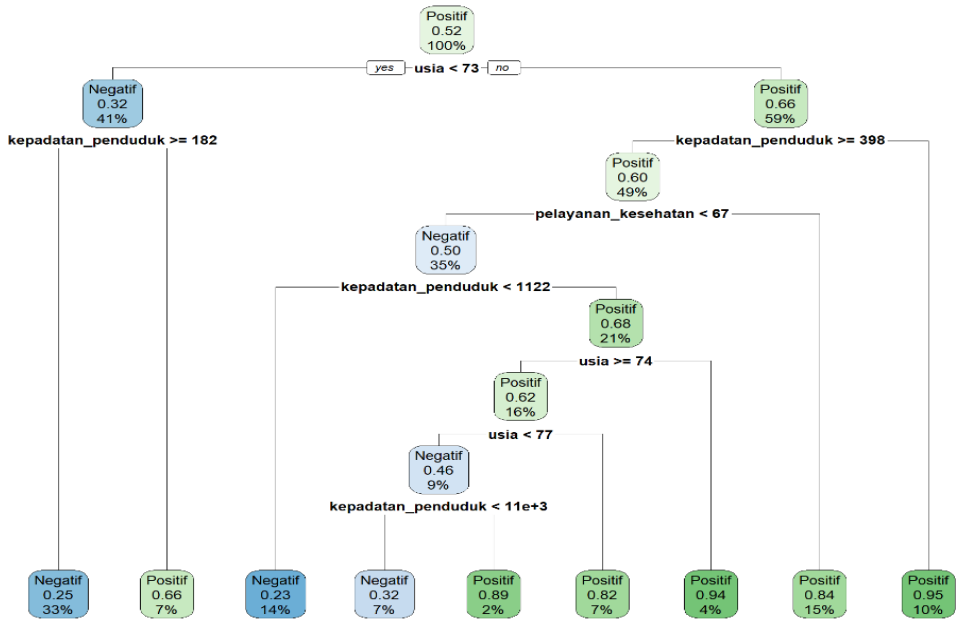
CP	Jumlah Pemisahan (nsplit)	*Kesalahan* Relatif	Xerror (*Kesalahan Validasi Silang*)	Deviasi Standar (XSTD)
0,303665	0	1,00000	1,00000	0,052183
0,054101	1	0,69634	0,70681	0,049450
0,047120	4	0,53403	0,68063	0,048985
0,017452	5	0,48691	0,58639	0,046970
0,015707	8	0,43455	0,62304	0,047819
0,010000	12	0,40314	0,62304	0,047819

Sumber: Hasil Penelitian (2025)

Proses klasifikasi menggunakan metode *Decision Tree* dimulai dengan menghitung nilai entropi dan perolehan informasi untuk menentukan atribut akar. Data pelatihan yang digunakan berjumlah 398 sampel, terdiri dari 207 kasus TB positif dan 191 kasus TB negatif. Nilai entropi total

(S) yang diperoleh adalah 0,998. Hasil perhitungan perolehan informasi menunjukkan bahwa variabel usia memiliki nilai tertinggi yaitu 0,041, diikuti oleh layanan kesehatan sebesar 0,011 dan kepadatan penduduk sebesar 0,005. Dengan demikian, variabel usia dipilih sebagai atribut utama (akar) dalam pembentukan *Decision Tree*. Selanjutnya, dilakukan proses pemangkasan untuk menentukan nilai parameter kompleksitas (CP) yang optimal.

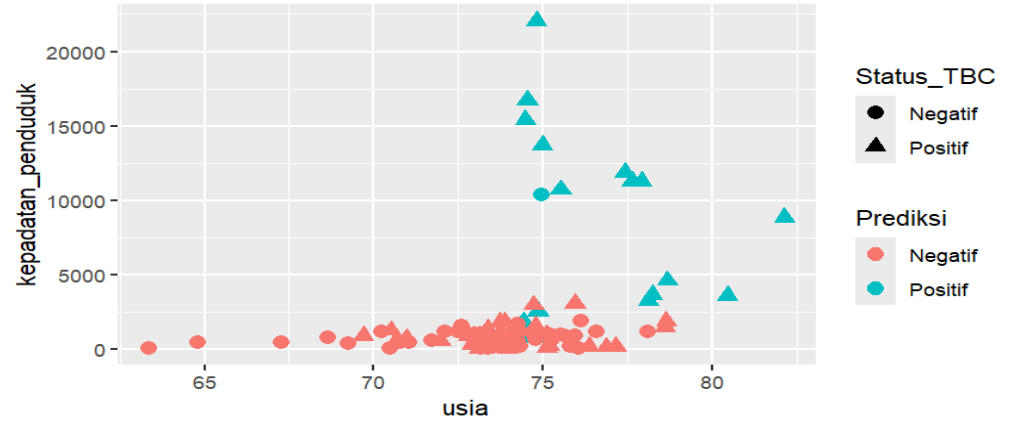
Tabel 3 menyajikan hasil analisis nilai CP berdasarkan validasi silang. Nilai CP sebesar 0,015707 menghasilkan 8 pembagian dengan kesalahan relatif 0,43455, kesalahan validasi silang 0,62304, dan deviasi standar 0,047819. Hasil ini menunjukkan keseimbangan antara kompleksitas dan kemampuan generalisasi model, sehingga menghasilkan *Decision Tree* yang lebih sederhana dan stabil.



Gambar 4. *Decision Tree* Setelah Pemangkasan
Sumber: Hasil Penelitian (2025)

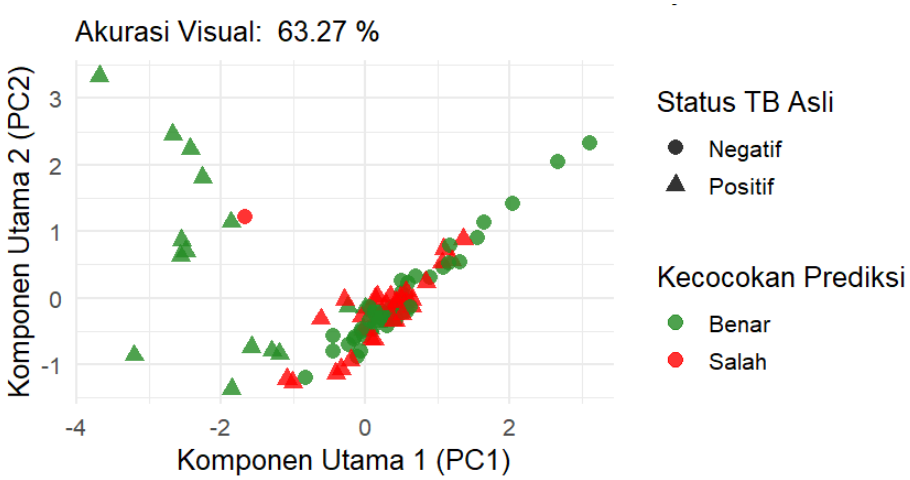
3.3. Klasifikasi Pengklasifikasi *Naïve Bayes*

Dalam metode Naïve Bayes, perhitungan probabilitas awal (prior) menghasilkan nilai P(Negatif) = 0,4799 dan P(Positif) = 0,5201. Berdasarkan perhitungan kemungkinan setiap atribut, nilai probabilitas posterior diperoleh sebagai P(Negatif | x) = 0,0488 dan P(Positif | x) = 0,0896, sehingga data diklasifikasikan ke dalam kelas TB Positif.



Gambar 5. Visualisasi prediksi model *Naïve Bayes*
Sumber: Hasil Penelitian (2025)

Gambar 5 menunjukkan visualisasi prediksi model *Naïve Bayes*, yang bertujuan untuk memperjelas evaluasi kinerja model untuk setiap kategori. Pendekatan faset digunakan dalam visualisasi. Faset adalah metode dalam visualisasi data yang memecah satu grafik utama menjadi beberapa subgraf berdasarkan nilai variabel kategorikal tertentu.



Gambar 6. Visualisasi PCA dari Prediksi *Naïve Bayes*
Sumber: Hasil Penelitian (2025)

Gambar 6 menunjukkan visualisasi hasil klasifikasi yang dilakukan menggunakan metode *facet* dan Analisis Komponen Utama (PCA). PCA digunakan untuk mereduksi variabel numerik utama usia, kepadatan penduduk, dan layanan kesehatan menjadi dua komponen utama untuk memfasilitasi analisis pola dari hasil klasifikasi.

3.4. Hasil Klasifikasi Metode Terbaik

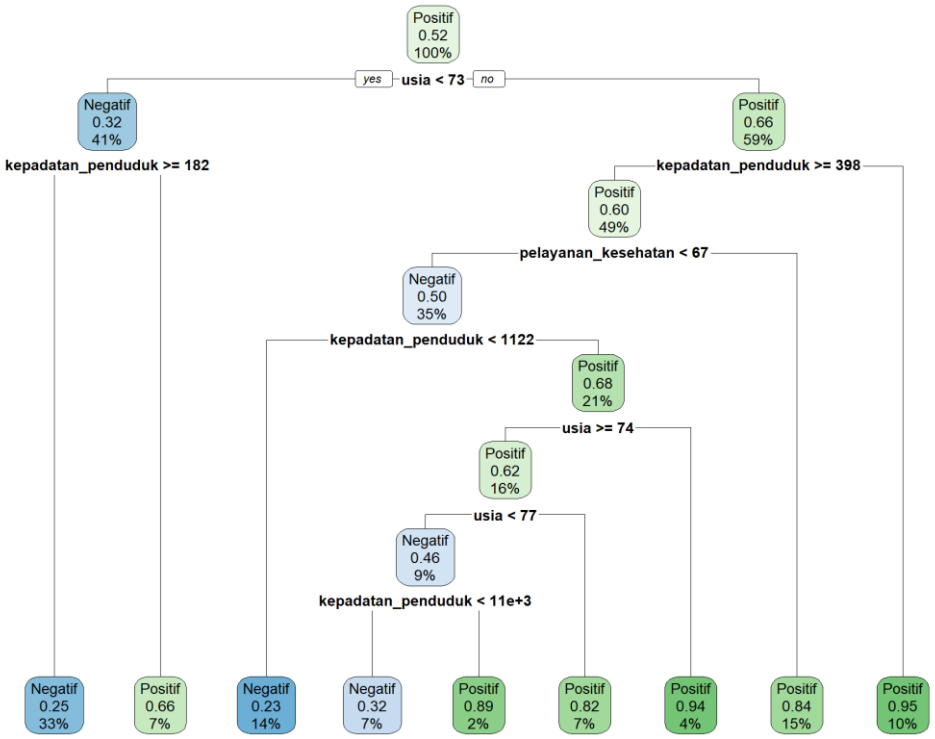
Untuk menentukan metode klasifikasi mana, *Naïve Bayes* atau *Decision Tree*, yang lebih efektif dalam mengklasifikasikan tuberkulosis (TB), dilakukan analisis berdasarkan tingkat akurasi. Penilaian ini dilakukan menggunakan Matriks Konfusi yang dihitung pada langkah sebelumnya. Hasil evaluasi akurasi untuk kedua metode dapat dilihat pada Tabel 4 di bawah ini:

Tabel 4. Matriks Konfusi Nilai Akurasi Komparatif Dua Metode

Metode	Akurasi
Decision Tree	76,63%
Naïve Bayes	63,27%

Sumber: Hasil Penelitian (2025)

Dalam penelitian ini, evaluasi kinerja model dilakukan menggunakan *confusion matrix* dengan indikator akurasi. Distribusi kelas pada data tuberkulosis dalam penelitian ini tidak sepenuhnya seimbang antara kelas TB positif dan TB negatif, sehingga berpotensi memengaruhi nilai akurasi sebagai indikator kinerja model. Meskipun demikian, perbandingan kinerja *Decision Tree* dan *Naïve Bayes* tetap valid karena kedua metode diuji pada dataset dan skema evaluasi yang sama.



Gambar 7. Visualisasi Metode Terbaik
Sumber: Hasil Penelitian (2025)

Gambar 7 menunjukkan struktur *Decision Tree* yang telah disederhanakan melalui pemangkasan. Model ini dibangun untuk mengklasifikasikan status TB menjadi positif dan negatif berdasarkan jumlah variabel prediktor, seperti usia, kepadatan penduduk, dan layanan kesehatan. Proses pemangkasan bertujuan untuk mengurangi kompleksitas model dengan menghilangkan elemen yang tidak penting untuk meningkatkan akurasi. Hasilnya adalah model yang lebih sederhana dan mudah diinterpretasikan, namun tetap memiliki kemampuan prediksi yang baik. Model yang dipangkas ini menyeimbangkan akurasi prediksi dan interpretasi, sekaligus menghindari *overfitting* pada data pelatihan. Struktur akhir ini mencerminkan pola hubungan yang paling bermakna antara variabel prediktor dan status TB dalam dataset yang digunakan.

KESIMPULAN DAN SARAN

1. KESIMPULAN

Hasil penelitian ini menunjukkan bahwa metode *Decision Tree* memiliki kinerja klasifikasi yang lebih unggul dibandingkan *Naïve Bayes* dalam mengklasifikasikan data tuberkulosis di Indonesia. *Decision Tree* mencapai tingkat akurasi sebesar 76,63%, sedangkan *Naïve Bayes* hanya mencapai 63,27%, dengan selisih akurasi sebesar 13,36%. Keunggulan *Decision Tree* dalam mengidentifikasi usia sebagai prediktor utama menunjukkan kemampuannya dalam memodelkan pola data TB yang kompleks. Secara kebijakan, hasil ini dapat dimanfaatkan oleh Kementerian Kesehatan dan Dinas Kesehatan daerah untuk mendukung pemetaan wilayah dan kelompok usia berisiko tinggi. Dengan demikian, program skrining, pencegahan, serta alokasi sumber daya kesehatan dapat dilakukan secara lebih tepat sasaran dan efektif.

2. SARAN

Penelitian selanjutnya disarankan menggunakan dataset yang lebih besar dan mencakup lebih banyak kabupaten/kota di seluruh Indonesia dengan rentang waktu yang lebih panjang agar variasi pola penyebaran TB dapat dianalisis secara lebih komprehensif. Selain itu, penelitian lanjutan dapat membandingkan kinerja *Decision Tree* dan *Naïve Bayes* dengan metode klasifikasi lain seperti *Random Forest*, *Support Vector Machine* (SVM), serta algoritma pembelajaran *ensemble*, disertai evaluasi menggunakan metrik tambahan seperti *precision*, *recall*, dan *F1-score* untuk mengatasi potensi ketidakseimbangan kelas. Penambahan variabel kontekstual, seperti tingkat kemiskinan, kepadatan hunian, dan akses sanitasi, juga disarankan guna meningkatkan kemampuan model dalam mendukung perumusan strategi pencegahan dan pengendalian TB yang lebih tepat sasaran oleh instansi kesehatan dan pemerintah daerah.

DAFTAR PUSTAKA

Asy'ari, T. S., Hidayat, F., & Lasari, H. H. (2024). *Analisis Epidemiologi Tuberkulosis Di Nusa Tenggara Barat Berdasarkan Data Sitk Tahun 2021–2024*. Konferensi Nasional.

Depari, D. H., Widiastiwi, Y., Santoni, M. M., Komputer, F. I., Pembangunan, U., Veteran, N., & Control, D. (2022). *Perbandingan Model Decision Tree , Naïve Bayes Dan Random Forest Untuk Prediksi Klasifikasi Penyakit Jantung*. *4221*, 239–248.

Fuadah, Y. N., Ubaidullah, I. D., Ibrahim, N., Taliningsing, F. F., Sy, N. K., & Pramuditho, M. A. (2022). Optimasi Convolutional Neural Network Dan K-Fold Cross Validation Pada Sistem Klasifikasi Glaukoma. *Elkomika: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika*, *10*(3), 728. Https://doi.org/10.26760/Elkomika.V10i3.728.

Gifari, O. I., Adha, M., Freddy, F., & Durrand, F. F. S. (2022). Film Review Sentiment Analysis Using Tf-Idf And Support Vector Machine. *Journal Of Information Technology*, *2*(1), 36–40.

Han, J., & Kamber, M. (2000). *Data Mining: Concepts And Techniques*. San Francisco, Ca: Morgan Kaufmann.

Kalimah, N. (2022). *Decision Tree Dan Struktur Node Dalam Klasifikasi Data*. Jurnal Ilmu Komputer Dan Informatika, 7(2), 45–52.

Kementerian Kesehatan Republik Indonesia. (2023). *Laporan Situasi Tuberkulosis Di Indonesia Tahun 2023*. Jakarta: Direktorat Jenderal Pencegahan Dan Pengendalian Penyakit, Kementerian Kesehatan Republik Indonesia.

Kementerian Kesehatan Republik Indonesia. (2024). *Profil Kesehatan Indonesia Tahun 2024*. Jakarta: Kementerian Kesehatan Republik Indonesia.

Kementerian Kesehatan Republik Indonesia. (2025). *Konferensi Pers Perkembangan Kasus Tb 2024–2025*. Jakarta: Kementerian Kesehatan Republik Indonesia.

Listanto, S., & Kristania, Y. M. (2022). *Implementasi Data Mining Terhadap Data Penjualan Dengan Algoritma Apriori Pada Pt . Duta Kencana Swaguna*. *16*, 364–372.

Muntasir, M., & Weraman, P. (2025). *Inovasi Penanggulangan Tuberkulosis Dengan Tcm*. *February*.

Muntiari, Nr, & Hanif, Kh (2022). *Klasifikasi Penyakit Kanker Payudara Menggunakan Perbandingan Algoritma Machine Learning* . *3* (1), 1–6.

Nti, Ik (2021). *Kinerja Algoritma Pembelajaran Mesin Dengan Nilai K Yang Berbeda Dalam Validasi Silang K-Fold* . *Desember* . Https://doi.org/10.5815/Ijitcs.2021.06.05

Nugroho, Agung & Agit Amrulloh. (2023). *Evaluasi Kinerja Algoritma K-Nn Menggunakan K-Fold Cross Validation Pada Data Debitur Ksp Galih Manunggal* . *5* (2), 294–300.

Prasitpuriprecha, C., Jantama, Ss, Preeprem, T., Pitakaso, R., Srichok, T., Khonjun, S., Weerayuth, N., Gonwirat, S., Enkvetchakul, P., Kaewta, C., & Nanthasamroeng, N. (2023). Rekomendasi Pengobatan Tuberkulosis Resisten Obat, Dan Deteksi Dan Klasifikasi Tuberkulosis Multikelas Menggunakan Sistem Berbasis Pembelajaran Mendalam Ensemble. *Farmasi* , *16* (1).

Quinlan, J. R. (1993). *C4.5: Programs For Machine Learning*. Morgan Kaufmann.

Rayuwati, Husna Gemasih, & Irma Nizar. (2022). Implementasi Algoritma *Naïve Bayes* Untuk Memprediksi Tingkat Penyebaran Covid. *Jurnal Riset Rumpun Ilmu Teknik*, *1*(1), 38–46. Https://doi.org/10.55606/Jurritek.V1i1.127

Riyadi, S., Siregar, M. M., Margolang, K. Fadhli F., & Andriani, K. (2022). Analysis Of Svm And *Naïve Bayes* Algorithm In Classification Of Nad Loans In Save And Loan Cooperatives. *Jurteksi (Jurnal Teknologi Dan Sistem Informasi)*, *8*(3), 261–270. Https://doi.org/10.33330/Jurteksi.V8i3.1483

Rokhman, K. A., Berlilana, B., & Arsi, P. (2021). Perbandingan Metode Support Vector Machine Dan Decision Tree Untuk Analisis Sentimen Review Komentar Pada Aplikasi Transportasi Online. *Journal Of Information System Management (Joism)*, *3*(1), 1–7. Https://doi.org/10.24076/Joism.2021v3i1.341

Sari, C. A., Sukmawati, A., Aprilli, R. P., Kayaningtias, P. S., & N. Y. (2022). Perbandingan Metode Naïve Bayes, Support Vector Machine Dan Decision Tree Dalam Klasifikasi Konsumsi Obat. *Jurnal Litbang Edusaintech (Jle) Http://Journal.Pwmjateng.Com/Index.Php/Jle Perbandingan*, *3*(1), 33–41.

Suntoro, J. (2019). *Data Mining: Algoritma Dan Implementasi Dengan Pemrograman Php*. Jakarta: Elex Media Komputindo.

Supriyadi, A. (2023). *Perbandingan Algoritma Naïve Bayes Dan Decision Tree (C4 . 5) Dalam Klasifikasi Dosen Berprestasi*. *7*(1), 39–49.

World Health Organization (2023). *Global Tuberculosis Report 2023*. Geneva: WHO.

Widaningsih, S., Yusuf, S., Informatika, J. T., Teknik, F., & Suryakancana, U. (2022). *Penerapan Data Mining Untuk Memprediksi Siswa Berprestasi Dengan Menggunakan Algoritma K Nearest Neighbor*. *9*(3), 2598–2611.

Yunita, L., Rahagia, R., Tambuala, F. H., Musrah, A. S., Sainal, A. A., & Suprapto. (2023). Efektif Pengetahuan Dan Sikap Masyarakat Dalam Upaya Pencegahan Tuberkulosis. *Journal Of Health (Joh)*, 186–193. Https://doi.org/10.30590/Joh.V10n2.619

Zulfauzi, Z., & Alamsyah, M. N. (2020). Penerapan Algoritma *Naïve Bayes* Untuk Prediksi Penerimaan Mahasiswa Baru Studi Kasus Universitas Bina Insan Fakultas Komputer. *Jurnal Teknologi Informasi Mura*, *12*(02), 156–165.