Jurnal Smart Teknologi
Vol. 6, No. 5, Juli 2025, Halaman 722 – 733
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

# Perbandingan Hasil Penerapan Metode Algoritma C4.5 Dan *Random Forest* Pada Penyakit *Tuberculosis* Di Puskesmas Jajag

# Comparison of the Results of the Application of C4.5 and *Random Forest* Algorithm Methods on *Tuberculosis* at the Jajag Health Center

**Jesica Cahya Ningrum[1], Agung Nilogiri[2], Qurrota A'yun[3]**
[1]Mahasiswa Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jember
email: jesicacahya8@gmail.com
[2]Dosen Fakultas Teknik, Universitas Muhammadiyah Jember
email: agungnilogiri@unmuhjember.ac.id
[3]Dosen Fakultas Teknik, Universitas Muhammadiyah Jember
email: qurrota.ayun@unmuhjember.ac.id

**Abstrak**

Tuberkulosis (TBC) masih menjadi masalah kesehatan serius dan termasuk dalam sepuluh penyebab kematian utama di dunia. Indonesia menempati posisi ketiga sebagai negara dengan beban TBC tertinggi. Salah satu tantangan utama yang dihadapi adalah pertumbuhan jumlah pasien yang lebih cepat dibandingkan dengan ketersediaan dokter. Kondisi ini menjadi masalah besar karena setiap orang berhak mendapatkan pelayanan kesehatan yang memadai untuk penyakit yang mereka derita. Sistem ini tidak bertujuan untuk menggantikan peran dokter, melainkan untuk memberikan rekomendasi atau kemungkinan hasil diagnosis berdasarkan gejala yang dialami oleh pasien. Selain itu, sistem ini dapat memprediksi atau mendiagnosis penyakit TBC sejak dini, sehingga membantu mengurangi penyebaran penyakit TB di masyarakat. Penelitian ini akan menggunakan metode Algoritma C4.5 yang akan dibandingkan dengan *Random Forest*. Hasil penelitian ini adalah hasil pengujian tertinggi pada metode Algoritma C4.5 menggunakan Fold Cross Validation dengan nilai K = 2 pada Langkah Uji 1 dengan akurasi sebesar 71,2%, presisi sebesar 75% dan recall sebesar 63%. Sedangkan hasil pengujian tertinggi pada metode *Random Forest* menggunakan Fold Cross Validation dengan nilai K = 4 pada Langkah Uji 2 dengan akurasi sebesar 73%, presisi sebesar 75,8% dan recall sebesar 68,1%. Hasil dari penelitian ini adalah metode *Random Forest* memiliki hasil simulasi yang lebih baik dari metode Algoritma C4.5.

**Kata Kunci**: Algoritma C4.5; Klasifikasi; Penerapan; Random Forest; Tuberkulosis

***Abstract***

*Tuberculosis (TB) remains a serious health problem and is among the top ten causes of death worldwide. Indonesia is ranked third as the country with the highest TB burden. One of the main challenges faced is the growth in the number of patients that is faster than the availability of doctors. This condition is a big problem because everyone has the right to receive adequate health services for the disease they suffer from. This system is not intended to replace the role of doctors, but rather to provide recommendations or possible diagnosis results based on symptoms experienced by patients. In addition, this system can predict or diagnose TB disease early, thereby helping to reduce the spread of TB disease in the community. This study will use the C4.5 Algorithm method which will be compared with Random Forest. The results of this study are the highest test results on the C4.5 Algorithm method using Fold Cross Validation with a value of K = 2 in Test Step 1 with an accuracy of 71.2%, a precision of 75% and a recall of 63%. While the highest test results on the Random Forest method using Fold Cross Validation with a value of K = 4 in Test Step 2 with an accuracy of 73%, a precision of 75.8% and a recall of 68.1%. The results of this study are that the Random Forest method has better simulation results than the C4.5 Algorithm method..*

***Keywords****: C4.5 Algorithm; Classification; Application; Random Forest; Tuberculosis*

Jurnal Smart Teknologi
Vol. 6, No. 5, Juli 2025, Halaman 722 – 733
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

## 1. PENDAHULUAN

*Tuberkulosis* (TBC) merupakan suatu infeksi yang disebabkan oleh kuman *Mycobacterium tuberculosis*, yang dapat menyerang tidak hanya paru-paru tetapi juga bagian tubuh lainnya (Kemenkes, 2016). Masalah tuberkulosis (TB) tetap menjadi tantangan kesehatan yang signifikan di tingkat global, menduduki peringkat sebagai salah satu dari sepuluh penyebab kematian utama di seluruh dunia. Berdasarkan Laporan TB Global WHO 2021, Indonesia berada di urutan ketiga sebagai negara dengan kasus TBC terbanyak di dunia.

Masalah yang timbul adalah peningkatan jumlah pasien berkembang lebih cepat daripada jumlah dokter yang tersedia. Situasi ini menjadi tantangan signifikan karena setiap orang berhak menerima layanan kesehatan yang pantas sesuai dengan kondisi yang mereka alami. Sistem ini tidak dimaksudkan untuk menggantikan posisi dokter, tetapi berfungsi sebagai alat yang memberikan rekomendasi atau kemungkinan hasil diagnosis berdasarkan gejala yang dialami oleh pasien. Selain itu, sistem ini dapat berkontribusi dalam mendeteksi atau mendiagnosis penyakit TB pada tahap awal, sehingga dapat menekan penyebarannya di komunitas.

Beragam studi telah dilaksanakan di area prediksi dengan mengaplikasikan teknik klasifikasi. Salah satu contoh penelitian dilakukan oleh Kelvin dan Zakarias (2022), yang memproyeksikan penyakit Cerebrovascular dengan memanfaatkan algoritma C4.5 dan Naïve Bayes. Temuan dalam penelitian itu menunjukkan bahwa algoritma C4.5 memiliki kinerja yang lebih baik, dengan tingkat akurasi mencapai 95%. Selain itu, nilai presisi, recall, dan F1-score yang dicapai adalah masing-masing 90%, 95%, dan 93%. Penelitian lainnya yang dilakukan oleh S. Sundaramurthy dan P. Jayavel (2020) mengenai proyeksi penyakit Rheumatoid Arthritis (RA) menunjukkan bahwa rata-rata akurasi mencapai 84%.

*Random Forest* sudah banyak digunakkan untuk penelitian dengan menggunakan teknik klasifikasi karena kinerjanya yang unggul dan sederhana, diantaranya seperti penelitian yang dilakukan oleh (Sudrajat et al., 2022) dalam klasifikasi penelitiannya menggunakan metode *Random Forest* menghasilkan akurasi yang paling tinggi sebesar 93.92%.

Berdasarkan penjelasan sebelumnya, Penelitian ini akan memanfaatkan algoritma C4.5 dan Random Forest untuk memprediksi penyakit Tuberkulosis. Algoritma C4.5 adalah metode modern dalam penambangan data yang digunakan untuk membuat pohon keputusan sebagai landasan dalam klasifikasi. Pohon keputusan ini bertindak sebagai panduan dalam memprediksi penyakit Tuberkulosis. Random Forest adalah metode klasifikasi yang menghasilkan tingkat akurasi tinggi dengan cara membangun beberapa pohon keputusan secara bersamaan, lalu menggabungkan hasilnya untuk mendapatkan prediksi yang lebih konsisten dan tepat. Oleh sebab itu, hasil klasifikasi yang diperoleh dengan Algoritma C4.5 akan dibandingkan dengan hasil yang didapat dari Random Forest dan akan dicari nilai akurasi, presisi, dan recall terbaik dari perbandingan ini.

## 2. TINJAUAN PUSTAKA

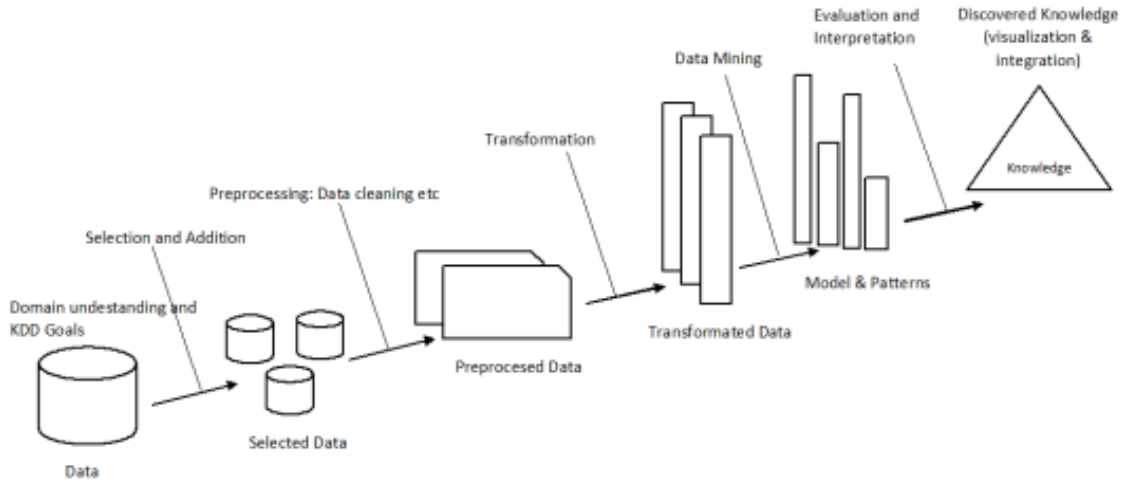
### A. Tuberkulosis

*Mycobacterium tuberculosis* adalah kuman penyebab infeksi Tuberculosis, yang menyebar melalui percikan udara (droplet). Di Indonesia, persentase kasus pada pria juga lebih tinggi (57,6%) dibandingkan wanita (Sapto et al., 2021b). Anak-anak berisiko tinggi terhadap penularan di dalam rumah, sementara kelompok lansia, khususnya mereka yang berusia 70-74 tahun, menunjukkan kerentanan yang lebih tinggi terhadap penyakit ini. Bahkan, pada usia 75 tahun ke atas, risiko penularan dapat mencapai 10,58% (Sapto et al., 2021a).

### B. Data Mining

Penggalian data memiliki sejumlah model proses yang berfungsi sebagai pedoman dalam melakukan analisis data, di antaranya Knowledge Discovery in Databases (KDD), CRISP-DM, dan SEMMA. Dalam penelitian ini, diterapkan model proses KDD yang mencakup sembilan langkah, seperti yang terlihat pada Gambar 1. (Nikmatun & Waspada, 2019).

Jurnal Smart Teknologi
Vol. 6, No. 5, Juli 2025, Halaman 722 – 733
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST




**Gambar 1.** Tahapan KDD
Sumber: Nikmatun & Waspada, 2019

Penambangan data adalah salah satu langkah dalam rangkaian Knowledge Discovery in Databases (KDD), yang mencakup beberapa tahap, seperti seleksi data, pemrosesan awal, konversi, penerapan metode penambangan data, dan penilaian hasil. Secara umum, KDD merupakan proses yang dimanfaatkan untuk menemukan pola, informasi, atau wawasan yang berguna dari sekumpulan data dalam jumlah besar. (Zai, 2022). Berikut adalah penjelasan mengenai setiap tahap dalam KDD yang tercantum di gambar 1:

a. Domain Understanding and KDD Goals, tujuannya yaitu untuk membantu dalam merancang solusi yang sesuai dengan kebutuhan pengguna dan konteks aplikasi yang relevan, serta mempertimbangkan pengetahuan yang telah ada untuk meningkatkan efektivitas dan relevansi sistem yang dikembangkan.

b. Selection and Additions, pada tahap ini, data yang akan digunakan untuk analisis atau model dipilih dengan hati-hati, memastikan bahwa data target (yang ingin diprediksi atau dianalisis) dan variabel yang mendukungnya (faktor-faktor yang mempengaruhi) ditentukan dengan tepat untuk mendapatkan hasil yang akurat dan relevan.

c. Preprocessing : Data Cleaning etc, pembersihan dan preprocessing data adalah langkah dasar yang penting untuk memastikan data yang digunakan dalam analisis atau model berada dalam kondisi yang konsisten dan bebas dari gangguan (noisy).

d. Transformation, Transformasi data adalah proses mengubah data dari satu bentuk ke bentuk lainnya sehingga data dapat diimplementasikan dan digunakan dengan lebih mudah dalam analisis atau pemodelan.

e. Data Mining (Chosing the Suitable Data Mining Task), setiap metode data mining memiliki tujuan dan aplikasi yang berbeda, dan pemilihan metode yang tepat akan membantu mencapai hasil yang optimal.

f. Data Mining (Chosing the Suitable Data Mining Algorithm), pemilihan algoritma harus disesuaikan dengan karakteristik data (seperti ukuran, tipe data, dan distribusinya) serta tujuan spesifik dari analisis data. Dengan memilih algoritma yang tepat, kita dapat memperoleh pola yang lebih akurat dan relevan sesuai dengan kebutuhan analisis.

g. Data Mining (Imploying Data Mining Algorithm), dengan implementasi yang tepat, algoritma yang dipilih akan mulai memberikan pola-pola yang relevan atau hasil prediksi yang berguna, sesuai dengan tujuan yang telah ditetapkan pada tahap awal.

h. Evaluation and Interpretation, proses ini bertujuan untuk memeriksa apakah pola atau informasi yang ditemukan sesuai dengan tujuan dan hipotesis yang telah ditetapkan

i. sebelumnya, atau apakah hasilnya memberikan wawasan baru yang perlu dianalisis lebih lanjut.
j. Discovered Knowledge, proses ini melibatkan penerapan hasil temuan untuk memberikan wawasan yang berguna dalam pengambilan keputusan atau tindakan selanjutnya. Secara keseluruhan, tahap ini adalah puncak dari proses KDD, di mana pengetahuan yang ditemukan bukan hanya menjadi informasi, tetapi juga diterjemahkan ke dalam tindakan nyata yang memberikan manfaat praktis, baik dalam konteks bisnis, medis, penelitian, maupun aplikasi lainnya.

Data mining dibagi menjadi beberapa kelompok berdaasarkan tugasnya yang dapat dilakukan, yaitu :
1. Deskripsi, Terkadang, penelitian dan analisis dilakukan dengan tujuan sederhana untuk mencari data yang dapat menggambarkan pola dan kecenderungan yang ada dalam informasi tersebut. Sebagai contoh, petugas pengumpulan suara mungkin tidak bisa memastikan atau menjelaskan fakta bahwa mereka yang kurang profesional cenderung mendapatkan dukungan lebih sedikit dalam pemilihan presiden.
2. Estimasi, estimasi mirip dengan klasifikasi, namun perbedaannya terletak pada variabel target yang lebih cenderung berbentuk numerik daripada kategori. Model dibangun dengan menggunakan data lengkap yang menyediakan nilai variabel target untuk diprediksi. Kemudian, dalam evaluasi berikutnya, nilai variabel target diperkirakan berdasarkan nilai variabel prediktor.
3. Prediksi, prediksi mirip dengan klasifikasi dan estimasi, namun perbedaannya terletak pada fokus prediksi yang berorientasi pada hasil di masa depan. Sebagai contoh, dalam bisnis dan penelitian, prediksi bisa berupa perkiraan harga beras tiga bulan ke depan. Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi juga dapat diterapkan (pada kondisi yang sesuai) untuk melakukan prediksi.
4. Klasifikasi, terdapat variabel target yang berupa kategori. Sebagai contoh, penggolongan pendapatan dapat dibagi ke dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah. Contoh lain dari klasifikasi adalah mendiagnosis penyakit pasien, di mana penyakit tersebut dikategorikan ke dalam jenis penyakit tertentu.
5. Pengklusteran (*Clustering)),* pengklusteran adalah proses pengelompokan data, pengamatan, atau objek-objek berdasarkan kemiripan tertentu. Kluster merupakan sekumpulan data yang saling mirip dan berbeda dengan data dalam kluster lain. Contoh pengklusteran adalah mengelompokkan konsumen untuk tujuan pemasaran produk, terutama bagi perusahaan yang memiliki anggaran pemasaran terbatas, sehingga mereka dapat menargetkan kelompok konsumen yang paling relevan secara efisien.
6. Asosiasi, proses ini dikenal sebagai asosiasi atau analisis pola asosiasi, yang bertujuan untuk menemukan atribut atau item yang sering muncul bersama pada suatu waktu. Dalam konteks bisnis, hal ini lebih dikenal sebagai analisis keranjang belanja. Contoh penerapannya adalah menentukan produk-produk di supermarket yang sering dibeli bersama, seperti roti dan mentega, serta produk yang jarang atau tidak pernah dibeli secara bersamaan, seperti susu dan makanan ringan.

### C. Algoritma C4.5

Algoritma C4.5 merupakan metode klasifikasi yang menggunakan struktur pohon keputusan untuk memodelkan proses klasifikasi. Algoritma ini merupakan pengembangan dari ID3, sehingga prinsip kerjanya masih sejalan dengan algoritma tersebut.
Keunggulan utamanya terletak pada kemampuannya dalam membentuk pohon keputusan yang mudah dipahami, memiliki tingkat akurasi yang baik, serta efisien dalam mengelola atribut bertipe diskret maupun numerik (Muhamad et al., 2019). Formula untuk mencari *entropy* adalah sebagai berikut (Ismanto & Novalia, 2021):

$$Entropy(S) = \sum_{i=1}^{n} - pi \ * \ log_2 pi$$

Keterangan :
S : Himpunan kasus
n : Jumlah partisi S
$P_i$ : Proporsi dari $S_i$ terhadap S

Jurnal Smart Teknologi
Vol. 6, No. 5, Juli 2025, Halaman 722 – 733
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

Kemudian lakukan perhitungan nilai *gain* dengan menggunakan formula sebagai berikut (Ismanto & Novalia, 2021):

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)$$

Keterangan :
S : Himpunan kasus
A : Atribut
n : Jumlah partisi atribut A
$|S_i|$ : Jumlah kasus pada partisi ke i

### D. Random Forest

Random Forest adalah algoritma yang sangat efektif untuk diterapkan pada masalah klasifikasi dalam *machine learning* dan *data mining* (Zailani & Hanun, 2020). Berkat arsitektur paralelnya, pengklasifikasi *Random Forest* lebih cepat dibandingkan dengan pengklasifikasi lainnya (Ramli & Sibaroni, 2022). Metode ini menggabungkan beberapa pohon klasifikasi, di mana setiap pohon dibangun secara independen dengan bergantung pada vektor sampel acak dan distribusi yang sama untuk setiap pohon dalam hutan tersebut (Syukron & Subekti, 2018).

### E. Python

Python merupakan bahasa pemrograman tingkat tinggi yang bersifat interpreted, interaktif, dan berorientasi objek. Bahasa ini dapat dijalankan di berbagai platform, seperti Linux, Windows, Mac, dan lainnya. Python dikenal karena kemudahan penggunaannya, berkat sintaks yang sederhana dan elegan. Selain itu, Python menyediakan berbagai modul dengan struktur data yang efisien dan siap pakai. Kode sumber aplikasi yang ditulis dalam Python umumnya dikompilasi terlebih dahulu menjadi bytecode sebagai format perantara sebelum dijalankan. (Ratna, 2020).

### F. Confusion Matrix

*Confusion matrix* merupakan matriks berukuran 2x2 yang digunakan untuk merepresentasikan hasil dari proses klasifikasi biner pada suatu dataset. Beberapa rumus umum dapat digunakan untuk mengukur performa suatu model klasifikasi dengan memanfaatkan nilai-nilai yang terdapat dalam confusion matrix. Hasil dari pengukuran seperti *accuracy, precision,* dan *recall* biasanya ditampilkan dalam bentuk persentase(Andika et al., 2019). Berikut merupakan tabel *Confusion matrix* (Lazuardi et al., 2020):

**Tabel 1.** Tabel Confusion Matrix

| ***Clasification*** | ***Predicted Class*** | ***Predicted Class*** |
|---|---|---|
| | **Class = *Yes*** | **Class = *Yes*** |
| Class = *Yes* | TP | FN |
| Class = *No.* | FP | TN |

Sumber: Lazuardi et al., 2020

Akurasi = $\frac{TP+TN}{TP+FP+TN+FN}$ x 100%

Precision = $\frac{TP}{TP+FP}$ x 100%

Recall = $\frac{TP}{TP+FN}$ x 100%

Keterangan :
TP = True Positives (jumlah pasien positif yang benar diklasifikasikan sebagai positif)
TN = True Negatives (jumlah pasien negatif yang benar diklasifikasikan sebagai negatif)
FP = False Positives (jumlah pasien negatif yang salah diklasifikasikan sebagai positif)
FN = False Negatives (jumlah pasien positif yang salah diklasifikasikan sebagai negatif)

## 3. METODE PENELITIAN

Metodologi penelitian ini terdiri dari beberapa tahapan, yaitu studi awal, pengumpulan data, preprocessing data, proses klasifikasi menggunakan algoritma C4.5 dan Random Forest, analisis hasil klasifikasi beserta pembahasannya, serta penarikan kesimpulan dan pemberian saran. Adapun diagram metodologi penelitian yang berisi kerangka penelitian yang akan dilakukan disajikan paga gambar 2.

### A. Studi Awal

Tahapan awal dimulai dengan mempelajari algoritma C4.5 dan *Random Forest*, kemudian dilanjutkan dengan mencari informasi mengenai penyakit Tuberkulosis, khususnya yang berkaitan dengan gejala-gejala penyakit tersebut. Selanjutnya, dilakukan pencarian dan kajian terhadap beberapa literatur, jurnal, dan paper yang relevan dengan latar belakang permasalahan yang telah dijelaskan.

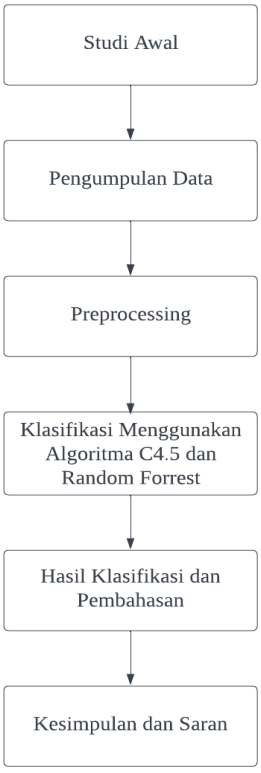
Jurnal Smart Teknologi
Vol. 6, No. 5, Juli 2025, Halaman 722 – 733
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

### B. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini diperoleh langsung dari Puskesmas Jajag. Dataset yang dipakai merupakan data Tahun 2022 bulan Januari sampai Desember dengan jumlah keseluruhan sebanyak 350 pasien. Data ini mencakup 5 *atribut*, yaitu : batuk ≥ 2 Minggu, penurunan berat badan, demam, batuk darah, dan nyeri dada yang akan diolah menggunakan algoritma C4.5 dan random forrest.

### C. Praproses Data

Setelah dilakukan analisis mendalam terhadap dataset yang telah diperoleh, beberapa atribut dipilih untuk digunakan dalam penelitian ini. Atribut-atribut tersebut meliputi batuk ≥ 2 minggu, penurunan berat badan, demam, batuk darah, dan nyeri dada. Berdasarkan pedoman yang telah ditetapkan, seluruh atribut ini dinyatakan valid dan layak digunakan dalam penelitian.




**Gambar 2.** Diagram Alur Penelitian
Sumber: Data Penelitian, 2025

### D. Smote

Data yang didapatkan dari puskesmas mengalami imbalance data, dimana jumlah data pasien positif TBC lebih kecil dari data data pasien negatif TBC atau bisa disebut data tidak seimbang, kondisi ini bisa menyebabkan metode klasifikasi mengabaikan kelas yang memiliki jumlah sampel sedikit sehingga memberi performa kurang baik. Oleh karena itu, permasalahan ketidakseimbangan data perlu ditangani dengan menerapkan teknik Synthetic Minority Oversampling Technique (SMOTE), yang berfungsi untuk menyeimbangkan distribusi kelas dalam dataset dengan cara menambahkan instance baru pada kelas minoritas. Proses oversampling pada SMOTE dilakukan dengan mengambil instance dari kelas minoritas, kemudian mencari k-nearest neighbors untuk setiap instance tersebut. Setelah itu, SMOTE menghasilkan instance sintetis baru, bukan sekadar mereplikasi instance yang ada. Dengan demikian, teknik ini dapat membantu mengurangi risiko overfitting yang berlebihan. (Sutoyo et al., 2020).

### E. Pembagian data

Pembagian data yaitu tahap pembagian dimana sebagian data disisihkan untuk data *Unseen* dan data training. Sebanyak 616 data dari proses SMOTE, akan disisihkan 11% untuk data *Unseen* sebanyak 68 data. Sisa data akan digunakan untuk data training dan data testing.

### F. K-Fold Cross Validation

*Cross-validation*, yang juga dikenal sebagai estimasi rotasi, adalah teknik untuk memvalidasi model dengan tujuan mengevaluasi sejauh mana hasil analisis statistik dapat diterapkan atau digeneralisasikan pada data lain yang tidak digunakan dalam proses pelatihan model.

Dalam penelitian ini, dataset yang digunakan terdiri dari 548 data, dan proses validasi silang dilakukan menggunakan dua jenis k-fold cross-validation: yaitu 2-fold cross validation dan 4-fold cross validation.

## 4. HASIL DAN PEMBAHASAN

### A. Implementasi Kedalam Algoritma C4.5

Dataset akan di uji sebanyak dua kali menggunakan k-fold cross validation yaitu

menggunakan 2 fold cross validation dan 4 fold cross validation, berikut adalah hasil pengujiannya:

a. 2 fold cross validation

Pada pengujian kali ini menggunakan 2 fold, dengan pembagian data training dan data testing masing masing sebanyak 274 data. Berikut adalah confusion matrix dan hasil uji data yang sudah dijalankan:

**Tabel 2.** Confusion Matrix Algoritma C4.5 Langkah Uji 1

| | Predicted Class | |
|---|---|---|
| | **Negatif** | **Positif** |
| Negatif | 108 | 29 |
| Positif | 50 | 87 |

Sumber: Hasil Penelitian, 2025

Tabel 2. Merupakan tabel confusion matrix algoritma c4.5 pada langkah uji 1 menggunakan 2 fold.

**Tabel 3.** Confusion Matrix Algoritma C4.5 Langkah Uji 2

| | Predicted Class | |
|---|---|---|
| | **Negatif** | **Positif** |
| Negatif | 108 | 29 |
| Positif | 61 | 76 |

Sumber: Hasil Penelitian, 2025

Tabel 3. Merupakan tabel confusion matrix algoritma c4.5 pada langkah uji 2 menggunakan 2 fold.

**Tabel 4.** Hasil Pengujian 2 Fold Cross Validation

| Langkah Uji | Akurasi | Presisi | Recall |
|---|---|---|---|
| Langkah Uji 1 | 71,2% | 75% | 63,5% |
| Langkah Uji 2 | 67,2% | 72,4% | 55,5% |

Sumber: Hasil Penelitian, 2025

Pada pengujian 2 fold menggunakan algoritma c4.5 dengan pembagian data training dan data testing masing-masing sebanyak 274 data, mendapatkan nilai akurasi tertinggi pada Langkah Uji 1 dengan akurasi sebesar 71,2%, presisi sebesar 75%, dan recall sebesar 63,1%.

b. 4 fold cross validation

**Tabel 5.** Confusion Matrix Algoritma C4.5 Langkah Uji 1

| | Predicted Class | |
|---|---|---|
| | **Negatif** | **Positif** |
| Negatif | 50 | 18 |
| Positif | 24 | 45 |

Sumber: Hasil Penelitian, 2025

Tabel 5. Merupakan tabel confusion matrix algoritma c4.5 pada langkah uji 1 menggunakan 4 fold.

**Tabel 6.** Confusion Matrix Algoritma C4.5 Langkah Uji 2

| | Predicted Class | |
|---|---|---|
| | **Negatif** | **Positif** |
| Negatif | 56 | 12 |
| Positif | 32 | 37 |

Sumber: Hasil Penelitian, 2025

Tabel 6. Merupakan tabel confusion matrix algoritma c4.5 pada langkah uji 2 menggunakan 4 fold.

**Tabel 7.** Confusion Matrix Algoritma C4.5 Langkah Uji 3

| | Predicted Class | |
|---|---|---|
| | **Negatif** | **Positif** |
| Negatif | 49 | 20 |
| Positif | 26 | 42 |

Sumber: Hasil Penelitian, 2025

Tabel 7. Merupakan tabel confusion matrix algoritma c4.5 pada langkah uji 3 menggunakan 4 fold.

**Tabel 8.** Confusion Matrix Algoritma C4.5 Langkah Uji 4

| | Predicted Class | |
|---|---|---|
| | **Negatif** | **Positif** |
| Negatif | 52 | 17 |
| Positif | 24 | 44 |

Sumber: Hasil Penelitian, 2025

Tabel 8. Merupakan tabel confusion matrix algoritma c4.5 pada langkah uji 4 menggunakan 4 fold.

**Tabel 9.** Hasil Pengujian 4 Fold Cross Validation

| Langkah Uji | Akurasi | Presisi | Recall |
|---|---|---|---|
| Langkah Uji 1 | 69,3% | 71,4% | 65,2% |
| Langkah Uji 2 | 67,9% | 75,5% | 53,6% |
| Langkah Uji 3 | 66,4% | 67,7% | 61,8% |
| Langkah Uji 4 | 70,1% | 72,1% | 64,7% |

Sumber: Hasil Penelitian, 2025

Pada pengujian 4 fold menggunakan algoritma c4.5 dengan pembagian data testing sebanyak 137 dan data training sebanyak 411 pada masing-masing fold, mendapatkan nilai akurasi tertinggi pada Langkah Uji 4 dengan akurasi sebesar 70,1%, presisi sebesar 72,1%, dan recall sebesar 64,7%.

Jurnal Smart Teknologi
Vol. 6, No. 5, Juli 2025, Halaman 722 – 733
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

## B. Implementasi kedalam Random Forest

Dataset akan di uji sebanyak dua kali menggunakan k-fold cross validation yaitu menggunakan 2 fold cross validation dan 4 fold cross validation, berikut adalah hasil pengujiannya:

a. 2 fold cross validation

**Tabel 10.** Confusion Matrix Random Forest Langkah Uji 1

| | Predicted Class | |
|---|---|---|
| | **Negatif** | **Positif** |
| Negatif | 105 | 32 |
| Positif | 50 | 87 |

Sumber: Hasil Penelitian, 2025

Tabel 10. Merupakan tabel confusion matrix random forest pada langkah uji 1 menggunakan 2 fold.

**Tabel 11.** Confusion Matrix Random Forest Langkah Uji 2

| | Predicted Class | |
|---|---|---|
| | **Negatif** | **Positif** |
| Negatif | 108 | 29 |
| Positif | 61 | 76 |

Sumber: Hasil Penelitian, 2025

Tabel 11. Merupakan tabel confusion matrix random forest pada langkah uji 2 menggunakan 2 fold.

**Tabel 12.** Hasil Pengujian 2 Fold Cross Validation

| Langkah Uji | Akurasi | Presisi | Recall |
|---|---|---|---|
| Langkah Uji 1 | 70,1% | 73,1% | 63,5% |
| Langkah Uji 2 | 67,2% | 72,4% | 55,5% |

Sumber: Hasil Penelitian, 2025

Pada pengujian 2 fold menggunakan random forest dengan pembagian data training dan data testing masing-masing sebanyak 274 data, mendapatkan nilai akurasi tertinggi pada Langkah Uji 1 dengan akurasi sebesar 70,1%, presisi sebesar 73,1%, dan recall sebesar 63,5%.

b. 4 fold cross validation

**Tabel 13.** Confusion Matrix Random Forest Langkah Uji 1

| | Predicted Class | |
|---|---|---|
| | **Negatif** | **Positif** |
| Negatif | 50 | 18 |
| Positif | 24 | 45 |

Sumber: Hasil Penelitian, 2025

Tabel 13. Merupakan tabel confusion matrix random forest pada langkah uji 1 menggunakan 4 fold.

**Tabel 14.** Confusion Matrix Random Forest Langkah Uji 2

| | Predicted Class | |
|---|---|---|
| | **Negatif** | **Positif** |
| Negatif | 53 | 15 |
| Positif | 22 | 47 |

Sumber: Hasil Penelitian, 2025

Tabel 14. Merupakan tabel confusion matrix random forest pada langkah uji 2 menggunakan 4 fold.

**Tabel 15.** Confusion Matrix Random Forest Langkah Uji 3

| | Predicted Class | |
|---|---|---|
| | **Negatif** | **Positif** |
| Negatif | 49 | 20 |
| Positif | 26 | 42 |

Sumber: Hasil Penelitian, 2025

Tabel 15. Merupakan tabel confusion matrix random forest pada langkah uji 3 menggunakan 4 fold.

**Tabel 16.** Confusion Matrix Random Forest Langkah Uji 4

| | Predicted Class | |
|---|---|---|
| | **Negatif** | **Positif** |
| Negatif | 53 | 16 |
| Positif | 24 | 44 |

Sumber: Hasil Penelitan, 2025

Tabel 16. Merupakan tabel confusion matrix random forest pada langkah uji 4 menggunakan 4 fold.

**Tabel 17.** Hasil Pengujian 4 Fold Cross Validation

| Langkah Uji | Akurasi | Presisi | Recall |
|---|---|---|---|
| Langkah Uji 1 | 69,3% | 71,4% | 65,2% |
| Langkah Uji 2 | 73% | 75,8% | 68,1% |
| Langkah Uji 3 | 66,4% | 67,7% | 61,8% |
| Langkah Uji 4 | 70,8% | 73,3% | 64,7% |

Sumber: Hasil Penelitian, 2025

Pada pengujian 4 fold menggunakan random forest dengan pembagian data testing sebanyak 137 dan data training sebanyak 411 pada masing-masing fold, mendapatkan nilai akurasi tertinggi pada Langkah Uji 2 dengan akurasi sebesar 73%, presisi sebesar 75,8%, dan recall sebesar 68,1%.

Jurnal Smart Teknologi
Vol. 6, No. 5, Juli 2025, Halaman 722 – 733
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

## C. Pengujian Menggunakan Unseen Data

Selanjutnya dilakukan penggujian menggunakan unseen data, unseen data yaitu data testing yang belum pernah dipelajari sebelumnya, jumlah unseen data yang dipakai sebanyak 68 data, berikut hasil pengujian unseen data menggunakan fold cross validation dengan nilai K = 2 dan K = 4.

**Tabel 18.** Hasil Pengujian 2 Fold Menggunakan Unseen Data (Algoritma C4.5)

| K-Fold | Langkah Uji | Akurasi | Presisi | Recall |
|---|---|---|---|---|
| 2 Fold | Langkah Uji 1 | 76,5% | 76,5% | 76,5% |
| | Langkah Uji 2 | 73,5% | 73,5% | 73,5% |
| 4 Fold | Langkah Uji 1 | 72,1% | 70,3% | 76,5% |
| | Langkah Uji 2 | 77,9% | 82,8% | 70,6% |
| | Langkah Uji 3 | 73,5% | 72,2% | 76,5% |
| | Langkah Uji 4 | 73,5% | 72,2% | 76,5% |

Sumber: Hasil Penelitian, 2025

Tabel 18. Merupakan hasil pengujian unseen data dengan algoritma c4.5 menggunakan 2 fold dan 4 fold yang menghasilkan nilai akurasi tertinggi pada 4 fold langkah uji 2 dengan nilai akurasi sebesar 77,9%.

**Tabel 19.** Hasil Pengujian 2 Fold Menggunakan Unseen Data (Random Forest)

| K-Fold | Langkah Uji | Akurasi | Presisi | Recall |
|---|---|---|---|---|
| 2 Fold | Langkah Uji 1 | 76,5% | 76,5% | 76,5% |
| | Langkah Uji 2 | 73,5% | 73,5% | 73,5% |
| 4 Fold | Langkah Uji 1 | 72,1% | 70,3% | 76,5% |
| | Langkah Uji 2 | 73,5% | 72,2% | 76,5% |
| | Langkah Uji 3 | 73,5% | 72,2% | 76,5% |
| | Langkah Uji 4 | 73,5% | 72,2% | 76,5% |

Sumber: Hasil Penelitian, 2025

Tabel 19. Merupakan hasil pengujian unseen data dengan random forest menggunakan 2 fold dan 4 fold yang menghasilkan nilai akurasi tertinggi pada 2 fold langkah uji 1 dengan nilai akurasi sebesar 76,5%.

**Tabel 20.** Rekapitulasi Hasil Uji Data Training dan Data Unseen (Algoritma C4.5)

| K-Fold Cross Validation | Langkah Uji | Data Training | Data Unseen |
|---|---|---|---|
| 2 Fold | Langkah Uji 1 | 71,2% | 76,5% |
| | Langkah Uji 2 | 67,2% | 73,5% |
| 4 Fold | Langkah Uji 1 | 69,3% | 72,1% |
| | Langkah Uji 2 | 67,9% | 77,9% |
| | Langkah Uji 3 | 66,4% | 73,5% |
| | Langkah Uji 4 | 70,1% | 73,5% |

Sumber: Hasil Penelitian, 2025

Pada pengujian K-Fold Cross Validation menggunakan Algoritma C4.5 mendapatkan hasil tertinggi pada K = 2 uji latih 1 dengan akurasi sebesar 71,2%. Kemudian dilakukan pengujian menggunakan unseen data mendapatkan peningkatan nilai akurasi sebesar 76,5%.

**Tabel 21.** Rekapitulasi Hasil Uji Data Training dan Data Unseen (Random Forest)

| K-Fold Cross Validation | Langkah Uji | Data Training | Data Unseen |
|---|---|---|---|
| 2 Fold | Langkah Uji 1 | 70,1% | 76,5% |
| | Langkah Uji 2 | 67,2% | 73,5% |
| 4 Fold | Langkah Uji 1 | 69,3% | 72,1% |
| | Langkah Uji 2 | 73% | 73,5% |
| | Langkah Uji 3 | 66,4% | 73,5% |
| | Langkah Uji 4 | 70,8% | 73,5% |

Sumber: Hasil Penelitian, 2025

Pada pengujian K-Fold Cross Validation menggunakan Random Forest mendapatkan hasil tertinggi pada K = 4 uji latih 2 dengan akurasi sebesar 73%. Kemudian dilakukan pengujian menggunakan unseen data mendapatkan peningkatan nilai akurasi sebesar 73,5%.

## 5. KESIMPULAN DAN SARAN

### A. Kesimpulan

Berdasarkan anlisa data dan pembahasan pada bab sebelumnya, maka di dapatkan kesimpulan sebagai berikut :

1. Hasil pengujian tertinggi pada metode Algoritma C4.5 menggunakan Fold Cross Validation dengan nilai K = 2 pada Langkah Uji 1 dengan akurasi sebesar 71,2%, presisi sebesar 75%, dan recall sebesar 63%, sedangkan hasil uji menggunakan data unseen dengan nilai K = 2 pada Langkah Uji 1 mendapatkan akurasi sebesar 76,5%.
2. Hasil pengujian tertinggi pada metode Random Forest menggunakan Fold Cross Validation dengan nilai K = 4 pada Langkah Uji 2 dengan akurasi sebesar 73%, presisi sebesar 75,8%, dan recall sebesar 68,1%, sedangkan hasil uji menggunakan data unseen dengan nilai K = 4 pada Langkah Uji 2 dengan akurasi sebesar 73,5%.
3. Pengukuran hasil penerapan metode Algoritma C4.5 dan Random Forest ini yaitu berdasarkan keakuratan prediksi akurasi. Sehingga metode Random Forest memiliki hasil penerapan yang lebih baik dari Algoritma C4.5 untuk mengklasifikasi penyakit *Tuberculosis* karena berdasarkan uji latih yang dilakukan, nilai akurasi dari metode Random Forest lebih tinggi dari pada nilai akurasi Algoritma C4.5.

### B. Saran

Saran yang dapat diberikan untuk mengembangkan sistem pada penelitian ini adalah:

1. Perlu menambah atribut yang digunakan yaitu penyakit yang bebeda dan menambah jumlah dataset yang digunakan sehingga diharapkan dapat mengasilkan model yang lebih baik.
2. Pada penelitian selanjutnya diharapkan bisa melakukan pengembangan dengan membuat aplikasi atau program yang bisa mendeteksi penyakit Tuberkulosis sesuai dengan sistem yang dibuat.

## 6. DAFTAR PUSTAKA

Adrian, M. R., Putra, M. P., Rafialdy, M. H., & Rakhmawati, N. A. 2021. Perbandingan Metode Klasifikasi Random Forest Dan Svm Pada Analisis Sentimen PSBB. *Jurnal Informatika Upgris*. 7(1). 36–40.

Aji, F., Umbara, F., & Kasyidi, F. 2023. Klasifikasi Risiko Kematian Pasien Berdasarkan Penyakit Penyerta Dan Usia Pasien Menggunakan Metode C4.5. *JIRE (Jurnal Informatika & Rekayasa Elektronika)*. 6(1). 9-17.

Andika, L. A., Azizah, P. A. N., & Wulan, R. 2019. Analisis Sentimen Masyarakat terhadap Hasil Quick Count Pemilihan Presiden Indonesia 2019 pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier. *Indonesian Journal of Applied Statistics.* 2(1). 34-41.

Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. 2021. Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest. *SISTEMASI : Jurnal Sistem Informasi.* 10(1). 163-171.

A'yun, Q., & Sujiwo., D., A., C. 2021. Analisis Keefektifan Pembelajaran Matematika Online. *Laplace: Jurnal Pendidikan Matematika.* 4(1). 88-98

Azis, H., Purnawansyah, P., Fattah, F., & Putri, I. P. 2020. Performa Klasifikasi K-NN dan Cross Validation Pada Data Pasien Pengidap Penyakit Jantung. *ILKOM Jurnal Ilmiah.* 12(2). 81–86.

Cholil, S. R., Dwijayanto, A. F., & Ardianita, T. 2020. Prediksi Penyakit Demam Berdarah di Puskesmas Ngemplak Simongan Menggunakan Algoritma C4.5. *SISTEMASI : Jurnal Sistem Informasi*. 9(3). 529-542.

Ismanto, E., & Novalia, M. 2021. Komparasi Kinerja Algoritma C4.5, Random Forest,

dan Gradient Boosting untuk Klasifikasi Komoditas. *Techno.COM*. 20(3). 400-410.

Lazuardi, M. B., Octavianto, H., & Sulistyo, H. W. 2020. Penerapan Algoritma Decision Tree C4.5 Dalam Klasifikasi Indentifikasi Pasien Penyakit Tuberkulosis (Tb) Di Puskesmas Sukorambi Jember. https://repository.unmuhjember.ac.id/8271/10/j.%20JURNAL.pdf. Diakses tanggal 15 Januari 2024.

Muhamad, M., Windarto, A. P., & Suhada, S. 2019. Penerapan Algoritma C4.5 Pada Klasifikasi Potensi Siswa Drop Out. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*. 3(1). 753-760.

Nabila, Z., Rahman Isnain, A., & Abidin, Z. 2021. Analisis Data Mining Untuk Clustering Kasus Covid-19 Di Provinsi Lampung Dengan Algoritma K-Means. *Jurnal Teknologi Dan Sistem Informasi (JTSI)*. 2(2). 100-108.

Nikmatun, I. A., & Waspada, I. 2019. Implementasi Data Mining Untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor. *Jurnal SIMETRIS*. 10(2). 421-432.

Nilogiri, A. Pegaruh Fitur Warna pada Klasifikasi Impresi Citra Batik Indonesia Menggunakan *Probabilistic Neural Network*. *JUSTINDO : Jurnal Sistem dan Teknologi Informasi Indonesia*. 1(1). 57-63.

Nurdiana, N., Nilogiri, A., & Rahman., M. Penerapan Algoritma *Fuzzy C-Means* dan Metode *Elbow* untuk Mengelompokkan Provinsi di Indonesia Berdasarkan Indeks Demokrasi Indonesia. *Jurnal Smart Teknologi*. 3(5). 544-551.

Qinanda, M., Q., Nilogiri, A., & Timur., T. Sentimen Pada Komentar Youtube Tentang Pencegahan dan Penanganan Kekerasan Seksual Pada Permendikbud Berbasis Naive Bayes dan Support Vektor Machine. 7(2). 114-121.

Ramadhani, I., R., Nilogiri, A., & A'yun, Q. 2023. Klasifikasi Jenis Tumbuhan Berdasarkan Citra Daun Menggunakan Metode Convolutional Neural Network. *Jurnal Smart Teknologi*. 3(3). 249-260.

Ramli, R. G., & Sibaroni, Y. 2022. Klasifikasi Topik Twitter Menggunakan Metode Random Forest Dan Fitur Ekspansi Word2vec. *e-Proceeding of Engineering*. 9(1). 79-92.

Ratna, S. 2020. Pengolahan Citra Digital Dan Histogram Dengan Phyton Dan Text Editor PHYCHARM. *Jurnal Ilmiah "Technologia"*. 11(3). 181-186.

Rubiah, W. R., & Ismanto, B., 2020. Evaluasi Program Bantuan Operasional Sekolah (BOS) di Sekolah Dasar. *Jurnal Kependidikan: Jurnal Hasil Penelitian dan Kajian Kepustakaan di Bidang Pendidikan, Pengajaran dan Pembelajaran*. 6(2). 220-229.

Sapto, J. 2021. Fakta Risiko Peningkatan Angka Insidensi Tuberkulosis. *Jurnal Ilmiah Pannmed (Pharmacyst, Analyst, Nutrition, Midwivery, Environment, Dental Hygiene)*. 16(1). 106-113.

Sudrajat, D., Purnamasari, A., Dikananda, A., Kurnia, D., Bahtiar, A. 2022. Klasifikasi Mutu Pembelajaran Hybrid berdasarkan Algoritma C.45, Random Forest dan Naïve Bayes dengan Optimasi Bootsrap Areggating (Bagging) pada masa COVID-19. *JURIKOM (Jurnal Riset Komputer)*. 9(6). 2227-2233.

Sujiwo, D., A., C., & A'yun., Q. 2020. Pengaruh Pemanfaatan E-learning Terhadap Motivai Belajar Mahasiswa. *JUSTINDO : Jurnal Sistem dan Teknologi Informasi Indonesia*. 4(1). 27-35.

Sutoyo, E., & Asri, F., M. 2020. Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika).* 6(3). 379-385.

Syukron, A., & Subekti, A. 2018. Penerapan Metode Random Over-Under Sampling dan Random Forest untuk Klasifikasi Penilaian Kredit. *JURNAL INFORMATIK.* 5(2). 175-185.

Zai, C. 2022. Implementasi Data Mining Sebagai Pengolahan Data. *Jurnal Portal Data* . 2(3). 1-12.

Zailani, A. U., & Hanun, N. L. 2020. Penerapan Algoritma Klasifikasi Random Forest Untuk Penentuan Kelayakan Pemberian Kredit Di Koperasi Mitra Sejahtera. *Infotech: Journal of Technology Information.* 6(1). 7–14.