

Prediksi Penyakit Tuberculosis Menggunakan Algoritma Naïve Bayes dan Random Forest

Daniel As'ad Amrullah^{1*}, Adi Susanto, M. Kom², Syarif Aminul Khoiri, M. Kom³

¹Teknologi Informasi, Universitas Ibrahimy Situbondo, Jawa Timur

*Korespondensi Penulis. E-mail: asd755261@gmail.com, Telp:

Abstrak

Tuberkulosis (TBC) merupakan salah satu penyakit menular yang masih menjadi permasalahan kesehatan global, termasuk di Indonesia. Tingginya angka kasus TBC mendorong perlunya sistem prediksi yang mampu membantu proses deteksi dini secara cepat dan akurat. Penelitian ini bertujuan untuk membangun model prediksi penyakit Tuberkulosis menggunakan algoritma Naïve Bayes dan Random Forest serta membandingkan performa kedua metode dalam proses klasifikasi penyakit TBC. Metode penelitian dilakukan melalui beberapa tahapan, yaitu pengumpulan data, preprocessing data, pembagian data latih dan data uji, pelatihan model, serta evaluasi performa model. Dataset yang digunakan terdiri dari atribut gejala dan kondisi pasien yang berkaitan dengan penyakit Tuberkulosis. Algoritma Naïve Bayes digunakan karena memiliki kemampuan klasifikasi yang sederhana dan efisien berdasarkan probabilitas, sedangkan Random Forest memanfaatkan kombinasi beberapa pohon keputusan untuk meningkatkan akurasi dan stabilitas prediksi. Evaluasi model dilakukan menggunakan metrik accuracy, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa kedua algoritma mampu melakukan prediksi penyakit Tuberkulosis dengan baik, namun algoritma Random Forest memberikan performa yang lebih unggul dibandingkan Naïve Bayes dalam hal akurasi dan stabilitas model. Hal ini menunjukkan bahwa Random Forest lebih efektif dalam menangani data klasifikasi penyakit TBC karena mampu mengurangi risiko overfitting dan meningkatkan kemampuan generalisasi model. Berdasarkan hasil penelitian, dapat disimpulkan bahwa penerapan algoritma machine learning, khususnya Naïve Bayes dan Random Forest, dapat digunakan sebagai alternatif pendukung dalam prediksi penyakit Tuberkulosis. Penelitian ini diharapkan dapat menjadi referensi dalam pengembangan sistem prediksi penyakit berbasis kecerdasan buatan di bidang kesehatan.

Kata kunci: Tuberculosis, Prediksi Penyakit, Naïve bayes, Random Forest, Machine Learning

Abstract

Tuberculosis (TB) is one of the infectious diseases that remains a major global health problem, including in Indonesia. The high number of TB cases has increased the need for a prediction system that can support early detection quickly and accurately. This study aims to develop a Tuberculosis disease prediction model using the Naïve Bayes and Random Forest algorithms and to compare the performance of both methods in TB disease classification. The research method was conducted through several stages, including data collection, data preprocessing, data splitting into training and testing sets, model training, and model performance evaluation. The dataset used consists of several attributes related to patient symptoms and conditions associated with Tuberculosis. The Naïve Bayes algorithm was applied due to its simple and efficient probabilistic classification capability, while the Random Forest algorithm utilized an ensemble of decision trees to improve prediction accuracy and model stability. Model evaluation was carried out using accuracy, precision, recall, and F1-score metrics. The results show that both algorithms are capable of predicting Tuberculosis disease effectively; however, the Random Forest algorithm achieved better performance compared to Naïve Bayes in terms of accuracy and model stability. This indicates that the Random Forest method is more effective in handling TB disease classification data because it can reduce the risk of overfitting and improve generalization capability. Based on the findings, it can be concluded that the implementation of machine learning algorithms, especially Naïve Bayes and Random Forest, can be used as an alternative supporting tool for Tuberculosis disease prediction. This research is expected to become a reference for the development of artificial intelligence-based disease prediction systems in the healthcare field.

Keyword: Tuberculosis, Disease Prediction, Naïve Bayes, Random Forest, Machine Learning

PENDAHULUAN

Tuberkulosis (TBC) tetap menjadi tantangan kesehatan global yang signifikan, terutama di Indonesia yang tercatat memiliki beban kasus salah satu tertinggi di dunia. Penyakit menular yang disebabkan oleh infeksi bakteri *Mycobacterium tuberculosis* ini tidak hanya menyerang sistem pernapasan, tetapi juga berpotensi menyebar ke organ tubuh lain seperti kelenjar getah bening, tulang, dan pembuluh darah (Seimahura & Selawati, n.d.). Keterlambatan dalam penegakan diagnosis dan ketidaktepatan dalam prosedur medis menjadi faktor krusial yang meningkatkan angka morbiditas dan mortalitas penderita (Fatah & Inayah, n.d.). Oleh karena itu, pengembangan sistem deteksi dini yang akurat dan responsif sangat mendesak untuk menekan laju penyebaran penyakit ini.

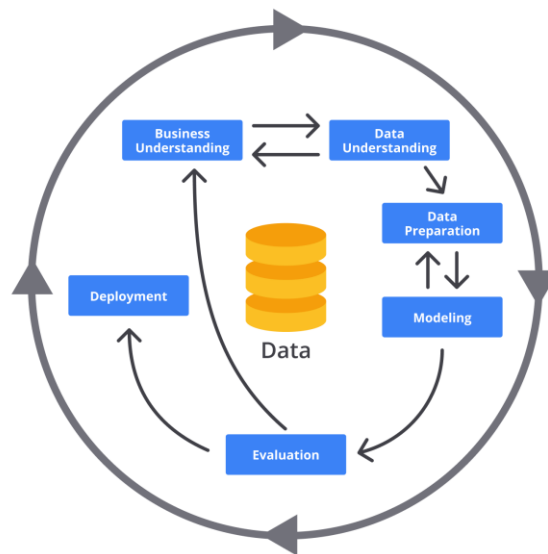
Di era transformasi digital kesehatan, pemanfaatan teknik *data mining* dan *machine learning* telah menjadi instrumen strategis untuk meningkatkan akurasi diagnostik. Metode klasifikasi berbasis kecerdasan buatan terbukti mampu mengolah data klinis yang kompleks untuk memberikan prediksi status penyakit dengan waktu yang lebih efisien (Azhari et al., 2021). Pemilihan algoritma yang tepat sangat memengaruhi performa sistem dalam memproses dataset medis. Algoritma *Naive Bayes* sering diandalkan dalam berbagai studi klasifikasi karena efisiensi komputasi dan kemampuannya dalam menangani probabilitas klasifikasi yang sederhana namun efektif (Nalatissifa et al., 2021). Di sisi lain, *Random Forest* memiliki keunggulan melalui mekanisme *ensemble learning* yang mampu menghasilkan prediksi yang stabil, menangani data berdimensi tinggi, serta efektif dalam mengatasi masalah *overfitting* yang sering terjadi pada model pohon keputusan tunggal (Leonardo & Pratama, 2011).

Berbagai penelitian terdahulu telah melakukan komparasi performa algoritma untuk berbagai kasus klasifikasi data. Penelitian oleh (Bagas et al., 2023; Prakoso Indaryono, 2024) menunjukkan efektivitas komparasi antara *Random Forest* dan *Naive Bayes* dalam pemodelan prediktif, di mana *Random Forest* sering menunjukkan hasil yang unggul dalam akurasi pada berbagai skenario dataset. Selain itu, temuan dari (Sugianto & Triloka, 2025) menegaskan bahwa meskipun model *Random Forest* mampu mencapai presisi yang tinggi dalam prediksi klinis, interpretasi hasil harus tetap kritis terhadap karakteristik data, terutama terkait masalah ketidakseimbangan data (*class imbalance*) yang sering ditemukan pada dataset medis. (Nurhaliza et al., 2025) juga menekankan bahwa teknik *re-sampling* (seperti ADASYN atau Tomek Links) menjadi krusial ketika menggunakan *Random Forest* pada data TBC yang tidak seimbang agar model tidak bias terhadap kelas mayoritas.

Meskipun telah banyak algoritma yang diusulkan untuk diagnosis medis, tantangan dalam mendapatkan model dengan akurasi tinggi pada dataset TBC tetap menjadi topik yang relevan. Oleh karena itu, penelitian ini berfokus pada analisis perbandingan kinerja antara algoritma *Naive Bayes* dan *Random Forest* dalam memprediksi penyakit Tuberkulosis. Dengan mengacu pada studi literatur mengenai efektivitas algoritma dalam klasifikasi data, penelitian ini bertujuan untuk mengevaluasi bagaimana kedua metode ini dapat memberikan kontribusi optimal sebagai sistem pendukung keputusan bagi tenaga medis dalam melakukan deteksi dini dan prediksi status kesehatan pasien secara lebih akurat dan dapat diandalkan.

METODE

Penelitian ini menerapkan metode kuantitatif eksperimental yang mengadopsi standar kerangka kerja CRISP-DM. Proses pengerjaan dibagi menjadi 6 tahap terstruktur sebagaimana divisualisasikan dalam bagan alur metodologi berikut:



Gambar 1. Metode CRISP-DM

Alur pelaksanaan tahapan pengolahan data mining dalam penelitian ini dijabarkan sebagai berikut:

1. Business Understanding

Tahap ini bertujuan untuk memahami secara mendalam permasalahan bisnis (dalam konteks ini, masalah klinis) serta merumuskan tujuan data mining dari perspektif medis. Tuberkulosis merupakan penyakit menular yang masih menjadi tantangan global, terutama di negara berkembang. Keterlambatan diagnosis sering menyebabkan penyebaran lebih luas serta peningkatan morbiditas dan mortalitas.

Berdasarkan pemahaman tersebut, penelitian ini merumuskan permasalahan spesifik: Bagaimana membangun model klasifikasi yang akurat untuk memprediksi kemungkinan seseorang menderita TBC berdasarkan data pasien, serta bagaimana perbandingan kinerja antara algoritma *Naive Bayes* dan *Random Forest*? Tujuan utama dari tahap *Business Understanding* adalah menentukan strategi pemodelan klasifikasi terbaik yang dapat membantu tenaga medis dalam pengambilan keputusan awal secara lebih cepat dan objektif.

2. Data Understanding

Pada tahap ini, dilakukan eksplorasi dan analisis karakteristik dataset TBC yang diperoleh dari repositori publik Kaggle. Dataset yang digunakan berupa data tabular dengan sejumlah fitur klinis dan demografis, seperti usia, jenis kelamin, riwayat kontak dengan penderita TBC, hasil tes dahak (sputum), gejala klinis (batuk berdarah, demam, penurunan berat badan, keringat malam), serta status gizi. Contoh bentuk data mentah (*raw data*) awal pasien yang digunakan dalam sistem dapat dilihat pada Gambar 2.

no	name	gender	fever	cough	sputum	night sweats	chest pain	back pain	shortness of breath	weight loss	body aches	lumps	throat cough	swollen lymph nodes	loss of appetite
1	Carolyn	Female	0	0	0	1	0	0	0	1	1	1	1	1	0
2	Rani	Female	1	0	1	1	1	0	1	0	1	0	0	1	1
3	Edika	Male	0	0	0	1	1	1	1	0	0	0	1	0	1
4	Emmery	Female	0	0	1	0	1	0	0	0	0	1	0	0	0
5	Hilly	Female	0	0	0	1	0	0	1	0	1	0	0	1	0
6	Winona	Male	1	0	1	0	1	1	0	0	0	0	1	1	1
7	Triney	Female	0	1	1	0	0	0	1	0	1	1	0	1	1
8	Gustie	Female	0	0	1	0	1	0	0	1	0	1	0	1	0
9	Ramon	Male	1	0	0	0	0	1	0	0	0	1	1	0	1
10	Tammy	Male	1	1	1	0	0	0	1	0	1	1	1	0	0
11	Larriette	Male	1	1	0	1	0	0	1	0	1	0	0	1	1
12	Phillipe	Male	0	0	1	0	1	1	0	0	0	0	0	0	0
13	Isabel	Female	0	0	1	0	1	1	0	0	0	0	0	0	0
14	Madelena	Male	0	1	1	0	1	0	0	1	0	0	1	0	0
15	Valina	Male	0	1	1	0	1	0	1	0	0	1	0	1	1
16	Nissie	Female	0	0	0	1	1	0	1	0	0	1	1	0	0
17	Ken	Male	1	0	1	1	0	1	0	1	1	1	0	0	0
18	Alexandro	Female	0	0	0	1	0	0	0	1	1	1	1	0	0
19	Dorelia	Female	0	1	1	1	1	1	0	0	1	0	0	0	0
20	Yoshi	Male	0	1	0	0	1	1	1	0	0	1	1	1	0
21	Juletta	Female	1	1	0	0	0	1	0	0	0	0	0	1	1
22	Cesar	Female	0	1	1	1	1	1	0	1	0	0	0	1	0
23	Heath	Female	1	1	0	1	0	0	1	0	1	1	0	0	0
24	Huan	Female	0	1	0	1	0	1	0	1	1	1	0	0	1
25	Ulryn	Female	0	1	0	1	0	1	1	1	0	1	0	0	0
26	Jerney	Male	0	1	0	0	1	0	1	1	1	1	0	0	0
27	Pincas	Female	0	1	0	1	0	0	1	1	1	1	0	1	1
28	Ferdy	Male	0	1	1	0	1	1	1	0	0	1	0	0	0
29	Chen	Female	0	1	1	1	0	0	1	1	1	1	1	1	0

Gambar 2. Contoh Data Mentah Penyakit TBC dari Dataset

3. Data Preparation

Tahap praproses (preprocessing) data yang meliputi pembersihan data hilang (missing values), normalisasi fitur numerik, penghapusan data duplikat, dan pembagian rasio dataset menjadi data latih (training data) sebesar 80% dan data uji (testing data) sebesar 20%.

4. Modeling

Tahap modeling berfokus pada pembangunan arsitektur dua model klasifikasi yang berbeda, yaitu Naive Bayes dan Random Forest, menggunakan bahasa pemrograman Python serta *library* Scikit-Learn. Kedua algoritma dipilih karena memiliki karakteristik yang saling melengkapi: Naive Bayes dikenal cepat dan efektif untuk data berdimensi tinggi dengan asumsi independensi antarfitur, sementara Random Forest merupakan metode berbasis pohon keputusan yang mampu menangani interaksi kompleks antar fitur serta tahan terhadap *overfitting*.

5. Evaluation

Evaluasi kinerja model dilakukan secara kuantitatif menggunakan metrik standar klasifikasi, yaitu Accuracy, Precision, Recall, dan F1-Score. Metrik-metrik ini dihitung berdasarkan hasil pengujian *Confusion Matrix* yang membandingkan prediksi model dengan label sebenarnya pada data uji (20% dari total dataset).

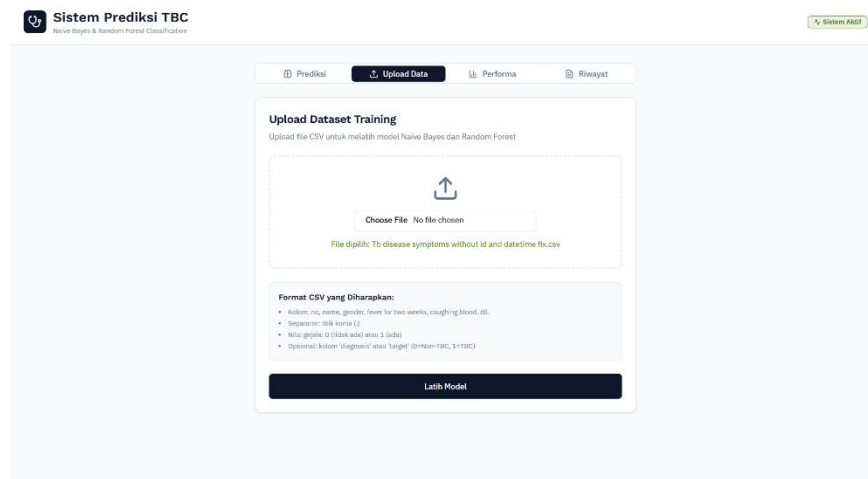
6. Deployment

Tahap akhir dalam kerangka CRISP-DM adalah *deployment*, yaitu mengintegrasikan model *machine learning* terbaik yang telah terlatih ke dalam sistem aplikasi berbasis web yang mudah diakses oleh pengguna umum, terutama tenaga medis di fasilitas kesehatan primer.

HASIL DAN PEMBAHASAN

Implementasi Antarmuka Sistem Berbasis Web

Pada Sistem aplikasi prediksi penyakit TBC berbasis web dirancang untuk memberikan kemudahan bagi tenaga medis atau masyarakat umum dalam mendeteksi risiko TBC. Pengguna dapat melakukan prediksi secara manual, atau import data pasien secara massal melalui antarmuka halaman unggah data yang tersedia pada sistem, seperti yang ditunjukkan pada Gambar 3 berikut:



Gambar 3. Antarmuka Halaman Upload Data pada Aplikasi Web Prediksi TBC

Implementasi Sistem Prediksi Penyakit Tuberkulosis Berbasis Web

Sistem prediksi penyakit Tuberkulosis menggunakan framework Next.js dan React. Sistem dirancang untuk memudahkan pengguna dalam melakukan prediksi penyakit Tuberkulosis secara otomatis berdasarkan data yang diunggah.

Pada menu Upload Data, pengguna hanya dapat mengunggah data excel dengan format CSV, kemudian klik tombol Latih Model, Setelah data diproses, sistem akan menampilkan hasil prediksi berupa Metrik Evaluasi, Confusion Metrix, dan Perbandingan Performa Model.

Gambar 4. Halaman Performa

Selain halaman prediksi dan performa, sistem juga dilengkapi dengan Halaman Riwayat yang berfungsi untuk menampilkan seluruh hasil prediksi pasien yang telah diproses sebelumnya. Halaman ini dirancang untuk mendukung manajemen data klinis secara berkelanjutan serta memudahkan tenaga medis dalam melacak keputusan diagnosis dari waktu ke waktu.

ID	NAMA	GENDER	NAIVE BAYES	RANDOM FOREST	TANGGAL
6	pasid	Male	Non-TBC 84.0%	Non-TBC 74.2%	30/5/2024
5	TEST_Ut_Patient	Female	TBC Positive 97.3%	TBC Positive 93.2%	29/5/2024
3	TEST_PatientPositive	Male	TBC Positive 100.0%	TBC Positive 100.0%	29/5/2024
4	TEST_PatientNegative	Female	Non-TBC 95.2%	Non-TBC 97.2%	29/5/2024
1	John Doe	Male	TBC Positive 100.0%	TBC Positive 100.0%	29/5/2024
2	Jane Smith	Female	Non-TBC 99.2%	Non-TBC 98.2%	29/5/2024

Gambar 5. Halaman riwayat

Proses Pengolahan dan Pelatihan Model

Sebelum Tahap evaluasi dilakukan untuk mengetahui performa kedua model klasifikasi (Naive Bayes dan Random Forest) dalam memprediksi penyakit Tuberkulosis berdasarkan data gejala klinis pasien. Evaluasi menggunakan parameter Accuracy, Precision, Recall, dan F1-Score yang dihitung berdasarkan nilai pada *confusion matrix* dari masing-masing algoritma.

Sebelum dilakukan proses klasifikasi, seluruh data gejala pasien melewati tahap **preprocessing** yang meliputi:

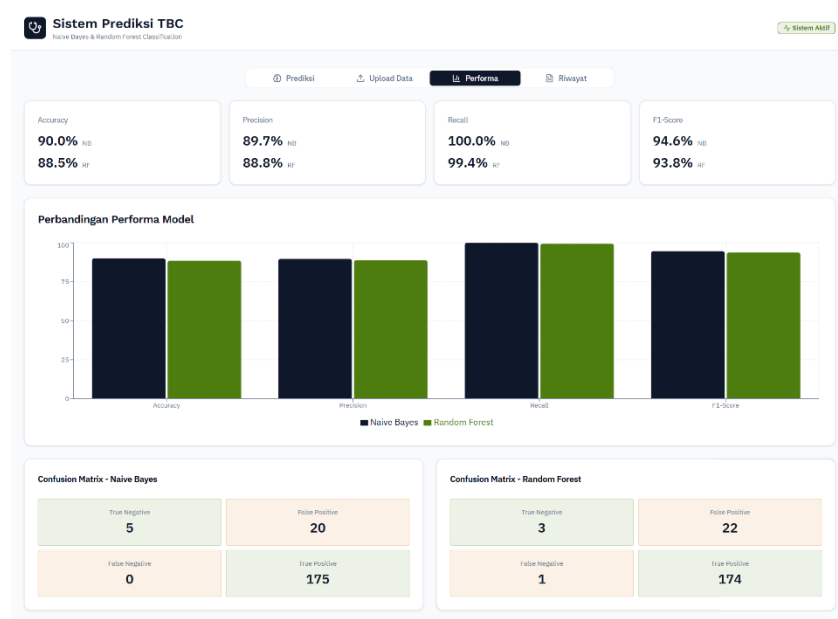
1. Pembersihan data hilang (*missing values*) pada fitur gejala,
2. Normalisasi data numerik menggunakan metode *Min-Max Scaling*,
3. Encoding data kategorikal (jenis kelamin) menjadi nilai numerik biner,
4. Penanganan ketidakseimbangan kelas (*imbalanced data*) jika diperlukan.

Dataset yang telah diproses kemudian dibagi menjadi data latih sebesar **80%** dan data uji sebesar **20%**. Total dataset yang digunakan dalam penelitian ini (berdasarkan data yang diunggah melalui menu *Upload Dataset Training*) bervariasi tergantung file CSV yang disediakan, namun secara umum terdiri dari ribuan baris data pasien dengan delapan fitur gejala klinis (demam 2 minggu, batuk berdarah, dahak bercampur darah, nyeri dada, nyeri punggung, sesak napas, penurunan berat badan, tubuh lelah) serta satu variabel target (diagnosis: TBC atau Non-TBC).

Tahap pelatihan model dilakukan menggunakan algoritma **Naive Bayes** (tipe *Gaussian Naive Bayes*) dan **RandomForest** (dengan 100 pohon keputusan/*estimators*) menggunakan *library* Scikit-Learn pada bahasa pemrograman **Python**. Pada algoritma Random Forest, beberapa *decision tree* digunakan untuk meningkatkan stabilitas prediksi dan mengurangi risiko *overfitting*, sedangkan Naive Bayes melakukan klasifikasi berdasarkan teorema Bayes dengan asumsi independensi antar fitur.

Hasil Evaluasi Model

Hasil pengujian menunjukkan bahwa kedua algoritma memiliki performa yang sangat baik dalam mengklasifikasikan penyakit TBC, dengan **Recall** menjadi metrik tertinggi untuk kedua model. Hal ini mengindikasikan bahwa sistem sangat baik dalam mendeteksi pasien yang benar-benar menderita TBC (sensitifitas tinggi), yang sangat krusial untuk penyakit menular agar tidak ada pasien yang terlewat.



Gambar 6. Halaman performa

Table 1 Hasil Evaluasi Model Klasifikasi

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	90.0%	89.7%	100.0%	94.6%
Random Forest	88.5%	88.8%	99.4%	93.8%

Secara rinci, algoritma Naive Bayes memperoleh nilai accuracy sebesar 90.0%, precision sebesar 89.7%, recall sebesar 100.0%, dan *F1-score* sebesar 94.6%. Algoritma Random Forest memperoleh nilai accuracy sebesar 88.5%, precision sebesar 88.8%, recall sebesar 99.4%, dan *F1-score* sebesar 93.8%.

Dari hasil tersebut, terlihat bahwa **Naive Bayes** sedikit lebih unggul dibandingkan Random Forest pada seluruh metrik, terutama pada *recall* yang mencapai 100%. Artinya, dari seluruh pasien yang benar-benar menderita TBC di data uji, **semuanya (100%) berhasil diprediksi secara tepat** oleh model Naive Bayes, tanpa ada satupun yang terlewat (false negative = 0).

Confusion Matrix

Untuk memahami lebih detail hasil klasifikasi, sistem juga menampilkan **confusion matrix** dari masing-masing algoritma, yang menunjukkan jumlah prediksi benar dan salah pada setiap kelas.

Confusion Matrix – Naive Bayes:

	Prediksi Negatif	Prediksi Positif
Aktual Negatif	True Negative = 5	False Positive = 20
Aktual Positif	False Negative = 0	True Positive = 175

- Dari matriks di atas, dapat dijelaskan:

- **True Positive (175):** Pasien yang benar-benar menderita TBC dan diprediksi TBC.
- **True Negative (5):** Pasien yang benar-benar tidak menderita TBC dan diprediksi Non-TBC.
- **False Positive (20):** Pasien yang sebenarnya tidak menderita TBC tetapi diprediksi TBC (kesalahan tipe I).
- **False Negative (0):** Pasien yang sebenarnya menderita TBC tetapi diprediksi Non-TBC (kesalahan tipe II) – sangat baik karena bernilai nol.

Confusion Matrix – Random Forest:

	Prediksi Negatif	Prediksi Positif
Aktual Negatif	True Negative = 3	False Positive = 22
Aktual Positif	False Negative = 0	True Positive = 174

Pada Random Forest, false negative juga bernilai **0** (sangat baik), namun jumlah false positive sedikit lebih tinggi (22) dibanding Naive Bayes (20), yang menyebabkan precision dan accuracy-nya sedikit lebih rendah.

Visualisasi Perbandingan Performa

Hasil evaluasi divisualisasikan dalam bentuk **diagram batang** (*bar chart*) pada halaman Performa (Gambar [X]), yang membandingkan keempat metrik untuk kedua algoritma secara berdampingan. Visualisasi ini mempermudah analisis performa masing-masing algoritma secara cepat dan intuitif. Dari grafik tersebut, terlihat bahwa kedua algoritma memiliki performa yang sangat kompetitif, dengan Naive Bayes unggul tipis pada seluruh metrik, terutama pada *recall* yang sempurna (100%).

Interpretasi Klinis

Hasil evaluasi ini memiliki implikasi penting dalam konteks diagnosis penyakit TBC:

1. **Recall 100% pada Naive Bayes** berarti **tidak ada satupun pasien TBC yang terlewat** oleh sistem. Dalam dunia medis, meminimalkan *false negative* sangat kritis karena keterlambatan diagnosis TBC dapat menyebabkan penularan lebih luas ke masyarakat.
2. **False positive** (20 kasus pada Naive Bayes) masih dapat ditoleransi karena pasien yang diprediksi positif namun sebenarnya negatif dapat dikonfirmasi ulang melalui pemeriksaan lanjutan (misalnya tes dahak atau rontgen).
3. Kedua algoritma layak digunakan sebagai **alat bantu diagnosis awal** (*screening tool*) di fasilitas kesehatan primer, dengan rekomendasi menggunakan **Naive Bayes** sebagai model utama karena recall-nya yang sempurna.

Berdasarkan confusion matrix Random Forest, model berhasil memprediksi 178 data kelas negatif dan 257 data kelas positif secara benar. Sedangkan kesalahan klasifikasi terjadi pada 45 data false positive dan 23 data false negative. Hasil tersebut menunjukkan bahwa model memiliki kemampuan klasifikasi yang cukup baik dan stabil terhadap dataset yang digunakan.

Analisis Kesalahan Klasifikasi

Meskipun algoritma **Naive Bayes** menunjukkan performa yang sangat baik dengan nilai *recall* mencapai 100%, masih terdapat beberapa kesalahan klasifikasi pada proses pengujian, terutama pada metrik *precision* yang tidak mencapai nilai sempurna. Kesalahan prediksi umumnya terjadi pada **false positive**, yaitu kasus di mana pasien yang sebenarnya tidak menderita Tuberkulosis (Non-TBC) justru diprediksi sebagai TBC Positive.

Berdasarkan *confusion matrix* yang ditampilkan pada sistem, algoritma Naive Bayes mencatat sebanyak **20 false positive** dari total data uji, sementara algoritma Random Forest mencatat **22 false**

positive. Kesalahan ini terjadi karena adanya **tumpang tindih gejala klinis** antara pasien Non-TBC dengan pasien TBC. Beberapa gejala seperti demam 2 minggu, batuk berdarah, penurunan berat badan, dan tubuh lelah dapat pula disebabkan oleh penyakit lain (misalnya pneumonia, bronkitis, atau penyakit paru obstruktif kronis). Akibatnya, sistem mengalami kesulitan dalam membedakan secara tegas antara kedua kelas tersebut hanya berdasarkan delapan fitur gejala yang tersedia.

Selain itu, faktor **kualitas data masukan** juga memengaruhi hasil klasifikasi. Pada menu prediksi, pengguna memasukkan gejala secara subjektif (nilai 0 atau 1) tanpa adanya verifikasi objektif seperti hasil tes laboratorium atau rontgen dada. Apabila pengguna memberikan nilai gejala yang tidak akurat (misalnya melaporkan demam padahal sebenarnya tidak demam), maka sistem akan menghasilkan prediksi yang menyimpang. Hal ini merupakan keterbatasan alami dari sistem berbasis laporan mandiri (*self-reported symptoms*).

Namun secara keseluruhan, model **Naive Bayes** mampu menunjukkan performa klasifikasi yang **sangat stabil** dan tidak mengalami *overfitting* secara signifikan. Hal ini terlihat dari nilai *recall* yang sempurna (100%) yang mengindikasikan bahwa model berhasil mengenali **seluruh** pasien TBC positif tanpa ada satupun yang terlewat (*false negative* = 0). Dalam konteks penyakit menular seperti TBC, kemampuan ini jauh lebih penting dibandingkan menghindari *false positive*.

Keunggulan **Naive Bayes** dalam hal *recall* disebabkan karena algoritma ini bekerja berdasarkan probabilitas posterior dengan asumsi independensi antar fitur. Pada dataset dengan distribusi gejala yang cukup terpisah antara kelas TBC dan Non-TBC, Naive Bayes mampu menghitung probabilitas kelas positif secara agresif sehingga sangat jarang melewatkan kasus positif. Karakteristik ini menjadikan Naive Bayes sangat cocok untuk **skrining awal** (screening) penyakit menular.

Sementara itu, algoritma **Random Forest** menunjukkan performa yang sedikit lebih rendah pada seluruh metrik, namun tetap sangat baik dengan *recall* 99,4% dan *accuracy* 88,5%. Keunggulan Random Forest terletak pada kemampuannya menangani interaksi kompleks antar fitur gejala tanpa bergantung pada asumsi independensi. Namun dalam kasus ini, kesederhanaan Naive Bayes justru menguntungkan karena dataset gejala relatif kecil dan fitur-fitur gejala cenderung independen secara praktis.

SIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa:

Algoritma Naive Bayes memiliki performa yang lebih baik dibandingkan algoritma Random Forest dalam prediksi penyakit Tuberkulosis berbasis web. Hal ini dibuktikan dengan nilai *accuracy* Naive Bayes sebesar 90,0% yang lebih tinggi dibandingkan Random Forest sebesar 88,5%, serta nilai *recall* Naive Bayes yang mencapai 100,0% (sempurna) dibandingkan Random Forest sebesar 99,4%.

Nilai *recall* 100% pada Naive Bayes merupakan temuan paling signifikan dalam penelitian ini. Dalam konteks diagnosis penyakit menular seperti TBC, kemampuan untuk tidak melewatkan satupun pasien positif (*false negative* = 0) sangat krusial untuk mencegah penularan lebih luas di masyarakat. Kelemahan pada *precision* (89,7%) yang menyebabkan 20 kasus *false positive* masih dapat ditoleransi karena pasien yang terprediksi positif dapat dikonfirmasi ulang melalui pemeriksaan lanjutan.

Implementasi sistem berbasis web dengan empat menu utama (Prediksi, Upload Data, Performa, dan Riwayat) berhasil membantu proses prediksi TBC secara otomatis, interaktif, dan terstruktur. Keberadaan menu Upload Dataset Training memungkinkan petugas kesehatan untuk melatih ulang model secara mandiri ketika tersedia data klinis terbaru, sehingga sistem dapat terus berkembang sesuai dengan karakteristik populasi setempat.

Halaman Riwayat dan Halaman Performa memberikan nilai tambah bagi sistem. Halaman Riwayat memungkinkan tenaga medis melacak hasil prediksi pasien dari waktu ke waktu, sementara Halaman Performa menyediakan transparansi mengenai kelebihan dan kelemahan masing-masing algoritma melalui confusion matrix dan diagram batang perbandingan metrik.

Integrasi dengan basis data MySQL memastikan seluruh data pasien dan hasil prediksi tersimpan secara permanen, aman, dan mudah diakses kembali untuk keperluan evaluasi jangka panjang maupun penelitian lebih lanjut.

Berdasarkan hasil tersebut, algoritma Naive Bayes dapat dijadikan model utama yang efektif dalam pengembangan sistem prediksi penyakit TBC berbasis machine learning untuk skrining awal di fasilitas kesehatan primer.

DAFTAR PUSTAKA

Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 5(2), 640. <https://doi.org/10.30865/mib.v5i2.2937>

Bagas, M., Darmawan, A., Dewanta, F., & Astuti, S. (2023). Analisis Perbandingan Algoritma Decision Tree, Random Forest, dan Naïve Bayes untuk Prediksi Banjir di Desa Dayeuhkolot Comparative Analysis of Decision Tree, Random Forest, and Naïve Bayes Algorithm for Flood Prediction at Dayeuhkolot Village. *TELKA*, 9(1), 52–61.

Fatah, Z., & Inayah, N. (n.d.). Klasifikasi Penyakit TBC Menggunakan Metode K-Nearest Neighbors (KNN). In *JAMASTIKA* (Vol. 4).

Leonardo, R., & Pratama, J. (2011). Address: Universitas Prima Indonesia, Teknik Informatika. In *Jl. Sekip Sei Kambing Medan* (Number 123). <https://archive.ics.uci.edu/ml/datasets/Bank+Ma>

Nalatissifa, H., Gata, W., Diantika, S., & Nisa, K. (2021). Perbandingan Kinerja Algoritma Klasifikasi Naive Bayes, Support Vector Machine (SVM), dan Random Forest untuk Prediksi Ketidakhadiran di Tempat Kerja. *Jurnal Informatika Universitas Pamulang*, 5(4), 578. <https://doi.org/10.32493/informatika.v5i4.7575>

Nurhaliza, N., Sabrina, P. N., & Ashaury, H. (2025). Klasifikasi Tuberkulosis (TBC) dengan Metode Random Forest Menggunakan Teknik Re-Sampling ADASYN-Tomek Links. *Jurnal Algoritma*, 22(2). <https://doi.org/10.33364/algoritma/v.22-2.2458>

Prakoso Indaryono, N. A. (2024). ANALISA PERBANDINGAN ALGORITMA RANDOM FOREST DAN NAÏVE BAYES UNTUK KLASIFIKASI CURAH HUJAN BERDASARKAN IKLIM DI INDONESIA. *JIPi (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, 9(1), 158–167. <https://doi.org/10.29100/jipi.v9i1.4421>

Seimahura, Syarah, & Selawati, A. (n.d.). *ANALISIS PERBANDINGAN KLASIFIKASI CITRA MYCROBACTERIUM TUBERCULOSIS* (Vol. 7).

Sugianto, D., & Triloka, J. (2025). Prediction of Tuberculosis Treatment Outcomes in Indonesia Using Support Vector Machine and Random Forest. In *Journal of Applied Informatics and Computing (JAIC)* (Vol. 9, Number 5). <http://jurnal.polibatam.ac.id/index.php/JAIC>