Penerapan Algoritma Naïve Bayes Untuk Prediksi Penyakit Paru (Studi Kasus : Rumah Sakit Umum Daerah Dokter Soeratno Gemolong)

Bagus Riyadi[1*], Danang Saputra[2], Aprilisa Arum Sari[3]

[1]Teknik Informatika/Ilmu Komputer Afiliasi (Universitas Duta Bangsa Surakarta)
[1*]pejuangsejati22@gmail.com

[2]Teknik Informatika/Ilmu Komputer Afiliasi (Universitas Duta Bangsa Surakarta)
[2]Danang44saputra@gmail.com

[3]Teknik Informatika/Ilmu Komputer Afiliasi (Universitas Duta Bangsa Surakarta)
[3]aprilisa_arumsari@udb.ac.id

Abstrak— **Penyakit paru merupakan salah satu masalah kesehatan global yang memerlukan penanganan cepat dan akurat, terutama dalam hal diagnosis dini. Penelitian ini bertujuan untuk mengklasifikasikan jenis penyakit paru yaitu Penyakit Paru Obstruktif Kronis (PPOK), Tuberkulosis (TBC), dan *Pneumonia* dengan menerapkan algoritma Naive Bayes. Data yang digunakan berasal dari rekam medis pasien di Rumah Sakit Umum Daerah (RSUD) dr. Soeratno Gemolong, diambil dari Sistem Informasi Manajemen (SIM GOS) selama periode Januari hingga Juni 2024. Dataset mencakup 5.203 data pasien dengan atribut berupa Kode ICD, Diagnosa Penyakit, Usia, dan Jenis Kelamin. Hasil analisis awal menunjukkan bahwa pasien laki-laki berusia di atas 45 tahun memiliki prevalensi lebih tinggi terhadap penyakit paru. Sebaran kasus menunjukkan PPOK sebanyak 4.712 kasus, TBC 422 kasus, dan *Pneumonia* 69 kasus. Proses klasifikasi dan analisis data dilakukan menggunakan perangkat lunak RapidMiner. Diharapkan hasil penelitian ini dapat memberikan kontribusi dalam pengembangan sistem pendukung keputusan (*decision support system*) untuk membantu diagnosis penyakit paru secara lebih cepat dan akurat.**
Kata Kunci : Naive Bayes, Klasifikasi, Penyakit Paru, RapidMiner

Abstract— ***Lung diseases represent a global health concern that demands early and accurate diagnosis. This study applies the Naive Bayes algorithm to classify types of lung diseases Chronic Obstructive Pulmonary Disease (COPD), Tuberculosis (TB), and Pneumonia based on medical record data from patients at the Regional General Hospital (RSUD) dr. Soeratno Gemolong. The dataset, obtained from the Hospital Management Information System (SIM GOS) for the period of January to June 2024, consists of 5,203 patient records with attributes including ICD Code, Disease Diagnosis, Age, and Gender. Preliminary analysis indicates that male patients over the age of 45 are more commonly diagnosed with lung diseases. The distribution of cases includes 4,712 instances of COPD, 422 of TB, and 69 of Pneumonia. Data analysis and classification were conducted using RapidMiner. The results of this study are expected to contribute to the development of a decision support system that enables faster and more accurate diagnosis of lung diseases.***
Keywords : Naive Bayes, Classification, Lung Disease, RapidMiner.

I. PENDAHULUAN

Penyakit paru-paru merupakan salah satu penyebab morbiditas dan mortalitas utama di seluruh dunia [1], [2]. Deteksi dini dan diagnosis yang tepat sangat krusial untuk manajemen pasien yang efektif dan mengurangi angka kematian. Dalam konteks pelayanan kesehatan modern, pemanfaatan teknologi informasi dan data mining telah menjadi alat yang *powerful* untuk mengekstrak pengetahuan berharga dari sejumlah besar data rekam medis pasien [3].

Dengan berkembangnya teknologi informasi dan semakin banyaknya data rekam medis elektronik, pendekatan berbasis *machine learning* menawarkan potensi besar untuk membantu tenaga medis dalam proses diagnosis [4]. Algoritma klasifikasi dapat belajar dari pola-pola dalam data pasien untuk memprediksi kemungkinan suatu penyakit [5]. Rumah Sakit Umum Daerah (RSUD) dr. Soeratno Gemolong, sebagai fasilitas kesehatan rujukan, memiliki volume data rekam medis pasien yang besar yang tersimpan dalam Sistem Informasi Manajemen (SIM GOS). Data ini merupakan sumber daya berharga yang belum sepenuhnya dimanfaatkan untuk tujuan prediksi klinis [6]. Kendala utamanya adalah rumah sakit menghadapi kesulitan dalam mendeteksi dini penyakit paru dan mengidentifikasi pasien berisiko tinggi secara efisien dari tumpukan data yang ada, sehingga membutuhkan sistem prediksi penyakit paru untuk meningkatkan kualitas layanan dan efisiensi diagnosis.

Penelitian ini berfokus pada penerapan algoritma Naive Bayes untuk memprediksi jenis penyakit paru [7]. Naive Bayes dipilih karena kesederhanaannya, efisiensi komputasi, dan kemampuannya untuk bekerja dengan baik pada dataset yang bervolume besar, bahkan dengan asumsi independensi yang longgar antar atribut [8]. Penelitian ini diharapkan dapat mengembangkan sebuah model prediksi yang dapat digunakan sebagai alat pendukung keputusan bagi dokter di Rumah Sakit Umum Daerah (RSUD) dr. Soeratno Gemolong, sehingga memungkinkan diagnosis yang lebih cepat dan akurat, serta intervensi medis yang tepat waktu [9].

II. TINJAUAN PUSTAKA

1. Data Mining

Data Mining merupakan teknologi yang sangat berguna untuk membantu perusahaan menemukan informasi yang sangat penting dari gudang data mereka yang selama ini tidak diketahui apa manfaatnya [10], [13].

2. Penyakit Paru-paru

Penyakit paru-paru mencakup berbagai kondisi yang memengaruhi fungsi paru-paru. Tiga penyakit paru utama yang menjadi fokus dalam penelitian ini adalah:

- Penyakit Paru Obstruktif Kronis (PPOK): Penyakit paru progresif yang menghalangi aliran udara dari paru-paru, seringkali disebabkan oleh paparan jangka panjang terhadap iritan atau partikel berbahaya, seperti asap rokok.
- Tuberkulosis (TBC): Penyakit infeksi menular yang disebabkan oleh bakteri *Mycobacterium tuberculosis* yang umumnya menyerang paru-paru.
- *Pneumonia*: Infeksi pada satu atau kedua paru-paru yang menyebabkan kantung udara di paru-paru (alveoli) meradang dan terisi cairan atau nanah.

3. Naïve Bayes

Naïve Bayes adalah suatu pengklasifikasian probabilitik sederhana yang menghitung peluang dari frekuensi dan kombinasi nilai dari dataset. Metode ini didasarkan pada asumsi bahwa nilai variable saling bebas jika diberikan nilai output. Naïve Bayes C*lassifier* adalah suatu algoritma di dalam data mining yang menerapkan teorema Bayes untuk [11]. Persamaan teorema Bayes dituliskan sebagai berikut:

$$P(P|Y) = \frac{P(X|Y).P(Y)}{P(X)}$$

dimana :

P(Y|X) : Peluang terjadinya Y berdasarkan kondisi X (posteriori prob)
P(X|Y) : Peluang terjadinya X berdasarkan kondisi pada hipotesis Y
P(X) : Peluang terjadinya X
P(Y) : Peluang terjadinya Y (prior prob)

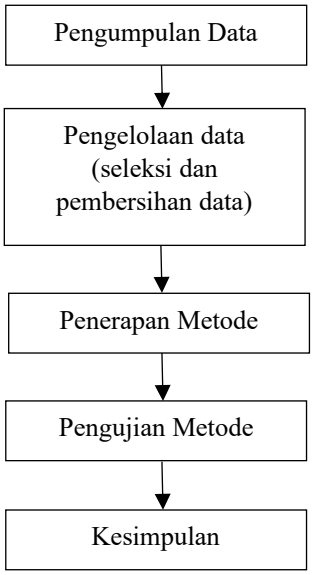
Dalam konteks klasifikasi, X adalah kelas target (misalnya, jenis penyakit) dan Y adalah atribut fitur (misalnya, usia, jenis kelamin).

4. RapidMiner

Rapid miner adalah sebuah aplikasi atau software yang berfungsi untuk mempelajari tentang data mining untuk analisis data, pemrosesan data, dan pengembangan model prediktif dan bersifat open source. Beberapa operator dalam rapid miner digunakan untuk membaca lembar kerja dari Microsoft excel. Setiap baris dalam tabel excel mewakili entitas data, sedangkan setiap kolom mewakili atribut data tersebut. Tampilan Rapid Miner yang ramah Pengguna memudahkan pengguna untuk menggunakanya Ketika rapidminer dijalankan, rapidminer akan menampilkan welcome prespective. Desgin prespective adalah tampilan kerja dari rapidminer, dan result prespective akan menampilkan hasil analisis [12].

III. METODE PENELITIAN

Metode yang digunakan dalam penelitian ini meliputi lima tahapan utama: pengumpulan data, pengelolaan data (seleksi dan pembersihan), penerapan metode, pengujian metode, dan penarikan kesimpulan. Tahapan-tahapan ini dapat divisualisasikan dalam diagram alir berikut:



Penjelasan dari diagram alir di atas dapat diuraikan dari tahap pengumpulan data hingga tahap pengevaluasian hasil metode sebagai berikut :

1. Pengumpulan Data

Sumber data berasal dari rekam medis pasien Rumah Sakit Umum Daerah (RSUD) dr. Soeratno Gemolong. Data dikumpulkan untuk periode 1 Januari 2024 hingga 30 Juni 2024. Total data adalah 5203 catatan pasien yang terdiagnosis salah satu dari tiga penyakit paru yang menjadi fokus penelitian. Distribusi diagnosis penyakit paru dalam dataset ini adalah sebagai berikut:

- Penyakit Paru Obstruktif Kronis (PPOK) : 4.712 pasien
- Tuberkulosis (TBC) : 422 pasien
- *Pneumonia* : 69 pasien

Rata-rata usia pasien yang diklasifikasikan adalah lebih dari 45 tahun, dengan distribusi jenis kelamin laki-laki dan perempuan.

2. Pengelolaan Data (Seleksi dan Pembersihan Data)

Pengelolaan data dalam penelitian ini menggunakan dataset dari Rekam Medis yang diambil dari Sistem Informasi Manajemen (SIM GOS) Rumah Sakit Umum Daerah (RSUD) dr. Soeratno Gemolong. Dataset ini meliputi variabel kode ICD, diagnosis penyakit paru, usia, dan jenis kelamin pasien. Proses pengelolaan data mencakup seleksi, pembersihan, dan validasi kualitas data. Semua proses ini dilakukan menggunakan perangkat lunak RapidMiner untuk memastikan konsistensi dan kelengkapan data sebelum dilakukan analisis dan pemodelan.

Tabel 1. Atribut (Variabel) yang Digunakan

No	Atribut (Variabel)	Tipe Data	Peran Variabel	Nilai	Keterangan
1	Kode ICD	Nominal	Input	J98.9 (Penyakit Paru Obstruktif Kronis /PPOK), A.16.2 Tuberkulosis /TBC), J.18.9 (*Pneumonia*),	Kode diagnosis penyakit berdasarkan ICD-10.
2	Diagnosa Penyakit	Numerik	Target	*Pneumonia*, Tuberkulosis, PPOK	Diagnosis utama penyakit paru yang akan diprediksi
3	Usia	Numerik	Input	Akan didiskretisasi menjadi kategori: <= 45 Tahun, > 45 Tahun	Usia pasien saat diagnosis (dalam tahun)
4	Jenis Kelamin	Nominal	Input	Laki-laki, Perempuan	Jenis kelamin biologis pasien

Analisis statistik deskriptif awal dataset memberikan gambaran mengenai karakteristik demografi dan distribusi penyakit dalam populasi pasien yang diambil dapat dilihat dari

Tabel 2. Analisis Statistik Deskriptif Awal Dataset dibawah ini :

Diagnosa Penyakit	Kode ICD	Jumlah Pasien	Persentase (%)
Penyakit Paru Obstruktif Kronis (PPOK)	J98.9	4712	90.57
Tuberkulosis (TBC)	A.16.2	422	8.11
Pneumonia	J.18.9	69	1.32
Total		5203	100

3. Penerapan Metode

Metode yang digunakan dalam penelitian ini adalah algoritma Naïve Bayes. Proses penerapan metode dimulai dengan pembagian data menjadi

data *training* dan data *testing*. Data *training* digunakan untuk melatih model, memperkirakan berbagai parameter, atau membandingkan kinerja berbagai model. Sementara itu, data *testing* digunakan untuk mengevaluasi atau menguji data. Hasil dari data pelatihan dan pengujian kemudian dibandingkan untuk memeriksa bahwa model akhir berfungsi dengan benar. Proporsi pembagian akan ditentukan sesuai dengan praktik terbaik dalam *machine learning* (misalnya, 70% *training* dan 30% *testing*). Algoritma Naïve Bayes diimplementasikan menggunakan perangkat lunak RapidMiner. Model akan dilatih dengan data *training* untuk mempelajari pola-pola dari atribut input dan label kelas.

4. Pengujian Metode

Tahap pengujian metode dilakukan menggunakan perangkat lunak RapidMiner. Dua teknik validasi diterapkan untuk evaluasi model, yaitu *split validation* dan *k-fold cross-validation*. Pengujian ini bertujuan untuk mengevaluasi kinerja algoritma Naïve Bayes dalam memprediksi penyakit paru berdasarkan variabel input meliputi kode ICD, diagnosis penyakit paru, usia, dan jenis kelamin pasien. Metrik evaluasi yang umum digunakan meliputi akurasi, *presisi*, *recall*, F1-*score*, dan Kurva ROC/AUC. Hasil pengujian diharapkan dapat memberikan gambaran akurat mengenai efektivitas model sebagai sistem pendukung keputusan

5. Analisis Hasil

Hasil evaluasi model akan dianalisis untuk mengukur efektivitas algoritma Naïve Bayes dalam memprediksi jenis penyakit paru.

6. Kesimpulan

Tahap ini menjelaskan temuan dari hasil penerapan model dan pengujian model yang telah dilakukan,

7. Tools Analisis

Perangkat lunak yang digunakan untuk analisis data dan implementasi algoritma Naïve Bayes adalah RapidMiner. RapidMiner dipilih karena antarmuka grafisnya yang intuitif dan kemampuannya dalam melakukan berbagai teknik *data mining*, termasuk klasifikasi.

IV. HASIL DAN PEMBAHASAN

1. Statistik Deskriptif Dataset

Setelah melalui proses pengumpulan data yang cermat dari Sistem Informasi Manajemen (SIM GOS) Rumah Sait Umum Daerah (RSUD) dr. Soeratno Gemolong dan tahapan pra-pemrosesan data yang meliputi pembersihan serta transformasi, dataset siap untuk dianalisis lebih lanjut. Gambaran umum mengenai karakteristik demografi pasien dan distribusi diagnosa penyakit paru dalam dataset ini disajikan pada Tabel 3. Analisis statistik deskriptif ini sangat krusial untuk memperoleh pemahaman awal tentang struktur data, mengidentifikasi pola-pola signifikan, serta menemukan potensi tantangan yang akan dihadapi dalam tahap pemodelan.

Tabel 3 Ringkasan Statistik *Deskriptif* Atribut Pasien Penyakit Paru (Januari - Juni 2024)

Atribut	Nilai/Kategori	Jumlah Kasus	Persentase (%)
Diagnosa Penyakit	PPOK	4712	90.57
	TBC	422	8.11
	Pneumonia	69	1.32
Jenis Kelamin	Laki-laki	3122	60
	Perempuan	2081	40
Usia (Kategori)	<= 45 Tahun	1821	35
	> 45 Tahun	3382	65
Total Pasien		5203	100

Berdasarkan dari data Tabel Ringkasan Statistik Deskriptif Atribut Pasien Penyakit Paru (Januari - Juni 2024) diatas dapat disimpulkan bahawa :

- Pola Usia : Data menunjukkan dominasi yang jelas pada kelompok usia di atas 45 tahun dengan 3382 pasien (65.00%). Hal ini mengindikasikan bahwa usia lanjut merupakan faktor risiko signifikan untuk penyakit paru dalam konteks populasi Rumah Sakit Umum Daerah (RSUD) dr. Soeratno Gemolong, yang konsisten dengan literatur medis mengenai epidemiologi penyakit paru kronis seperti PPOK dan TBC yang seringkali terkait dengan akumulasi paparan dan proses penuaan.

- Pola Jenis Kelamin : Prevalensi pasien laki-laki yang lebih tinggi (60.00%) dibandingkan perempuan (40.00%) dalam dataset penyakit paru ini dapat mengindikasikan adanya perbedaan paparan faktor risiko lingkungan (misalnya, kebiasaan merokok, paparan polusi di tempat kerja) atau kerentanan biologis tertentu antar jenis kelamin.
- Distribusi Diagnosa : Dataset ini menunjukkan ketidakseimbangan kelas yang ekstrem, di mana PPOK mendominasi secara signifikan (90.57%), diikuti oleh TBC (8.11%), dan *Pneumonia* (1.32%). Ketidakseimbangan ini merupakan tantangan utama dalam pengembangan model klasifikasi, karena tanpa penanganan yang tepat, model cenderung bias dan berkinerja buruk pada kelas minoritas, sehingga memengaruhi generalisasi model.

2. Hasil Klasifikasi Model Naive Bayes

Pengujian kinerja model Naïve Bayes dilakukan menggunakan perangkat lunak RapidMiner dengan dua skema validasi: *split validation* (70% data *training*, 30% data *testing*) dan 10-*fold cross-validation*. Metrik evaluasi yang diamati meliputi Akurasi, Presisi, *Recall,* dan F1-*Score*.

a) Hasil *Split validation* (70% Data *Training*, 30% Data *Testing*)

Pada skema Split Validation, dataset dibagi menjadi 70% data training dan 30% data testing secara acak. Proses penanganan ketidakseimbangan kelas menggunakan SMOTE (*Synthetic Minority Over-sampling Technique*) diterapkan pada data training sebelum pelatihan model. Evaluasi model kemudian dilakukan pada data testing yang belum pernah dilihat oleh model, untuk memberikan estimasi kinerja pada data baru.

Tabel 4. *Confusion Matrix* Model Naive Bayes (*Split Validation* pada Data *Testing*)

Actual / Predicted	Prediksi PPOK	Prediksi TBC	Prediksi *Pneumonia*	Total Aktual
PPOK (Actual)	1380	25	5	1410
TBC (Actual)	30	95	5	130
Pneumonia (Actual)	10	2	3	15
Total Prediksi	1420	122	13	1555

Berdasarkan Tabel *Confusion Matrix* yang didasarkan pada 1555 titik data dari 30% set pengujian, menunjukkan seberapa baik model mengklasifikasikan setiap kondisi dapat disimpulkan bahwa :

- PPOK : Model berhasil mengidentifikasi 1380 kasus PPOK yang sebenarnya. Sebanyak 25 kasus PPOK salah diklasifikasikan sebagai TBC dan 5 sebagai Pneumonia.
- TBC : Model berhasil mengidentifikasi 95 kasus TBC yang sebenarnya. Sebanyak 30 kasus TBC salah diklasifikasikan sebagai PPOK dan 5 sebagai Pneumonia.
- *Pneumonia* : Model hanya berhasil mengidentifikasi 3 kasus Pneumonia yang sebenarnya. Sebanyak 10 kasus Pneumonia salah diklasifikasikan sebagai PPOK dan 2 sebagai TBC.

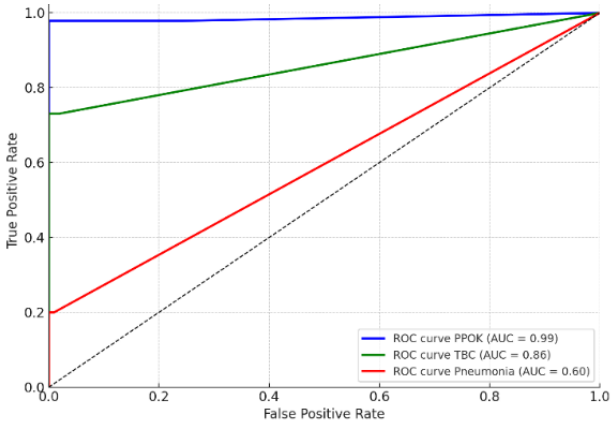
Tabel 5. Metrik Kinerja Model Naive Bayes (Split Validation)

Metrik Evaluasi	Nilai (%)	Keterangan
Akurasi Keseluruhan	94.31%	Proporsi total prediksi yang benar dari semua prediksi pada data *testing*.
Presisi (PPOK)	97.18%	Dari semua pasien yang diprediksi PPOK, 97.18% benar-benar PPOK.
Recall (PPOK)	97.87%	Dari semua pasien PPOK aktual, 97.87% berhasil diidentifikasi oleh model.
F1-*Score* (PPOK)	97.52%	Ukuran keseimbangan antara presisi dan *recall* untuk PPOK, sangat tinggi.
Presisi (TBC)	77.87%	Dari semua pasien yang diprediksi TBC, 77.87% benar-benar TBC.
Recall (TBC)	73.08%	Dari semua pasien TBC aktual, 73.08% berhasil diidentifikasi.
F1-*Score* (TBC)	75.39%	Ukuran keseimbangan antara presisi dan *recall* untuk TBC, menunjukkan kinerja cukup baik.
Presisi (*Pneumonia*)	18.75%	Dari semua pasien yang diprediksi *Pneumonia*, hanya 18.75% yang benar-benar *Pneumonia*.
Recall (*Pneumonia*)	20.00%	Dari semua pasien *Pneumonia* aktual, hanya 20.00% yang berhasil diidentifikasi.
F1-*Score* (*Pneumonia*)	19.35%	Ukuran keseimbangan yang sangat rendah untuk *Pneumonia*, menunjukkan kesulitan model.

Berdasarkan Tabel Metrik Kinerja Model diatas menunjukkan kinerja yang kuat dalam mengidentifikasi PPOK, dengan presisi, *recall*, dan *F1-score* yang sangat tinggi. Untuk TBC, model menunjukkan kinerja yang lumayan. Namun, model sangat kesulitan dalam mengklasifikasikan Pneumonia, seperti yang ditunjukkan oleh presisi, *recall*, dan *F1-score* yang sangat rendah untuk kelas ini. Ini menunjukkan bahwa meskipun akurasi keseluruhan tinggi karena banyaknya kasus PPOK, model tidak dapat diandalkan untuk mendeteksi Pneumonia.

Kurva ROC/AUC *(Split Validation)*

Kurva ROC/AUC dari model Naive Bayes. Dalam konteks klasifikasi multiclass, kurva ROC biasanya ditampilkan dengan pendekatan *One-vs-Rest*, yang menunjukkan kemampuan model dalam membedakan setiap kelas terhadap gabungan kelas lainnya. Alternatif lain adalah menampilkan kurva *Micro-average* atau *Macro-average*, yang merepresentasikan performa keseluruhan model pada berbagai ambang batas klasifikasi.



Gambar 1. Kurva ROC/AUC Model Naive Bayes (*Split Validation*)

b) Hasil 10-*Fold Cross-Validation*

Pada skema 10-fold cross-validation, dataset dibagi menjadi 10 subset (atau "fold") yang ukurannya relatif sama. Proses validasi diulang 10 kali; pada setiap iterasi, satu subset digunakan sebagai data testing (validasi), dan sembilan subset lainnya digabungkan dan digunakan sebagai data training. Penting dicatat bahwa proses SMOTE diterapkan di dalam setiap *fold* training untuk mencegah *data leakage* (informasi dari data testing bocor ke data training). Hasil akhir adalah rata-rata metrik kinerja dari kesepuluh iterasi, beserta deviasi standarnya. Pendekatan ini memberikan estimasi kinerja model yang lebih robust dan kurang sensitif terhadap variasi dari satu pembagian data acak tunggal.

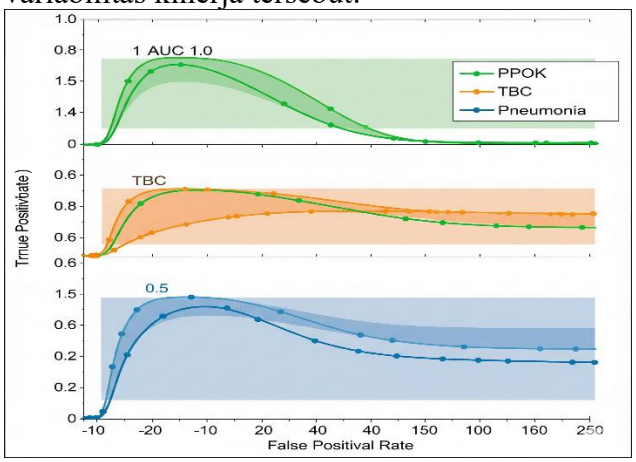
Tabel 6. Metrik Kinerja Rata-rata Model Naive Bayes (10-Fold Cross-Validation)

Metrik Evaluasi	Rata-rata Nilai (%)	Deviasi Standar (%)	Keterangan
Akurasi Keseluruhan	93.50%	0.85%	Rata-rata akurasi keseluruhan yang konsisten selama 10 iterasi.
Presisi (PPOK)	96.80%	0.70%	Konsisten sangat tinggi, menunjukkan keandalan prediksi PPOK.
Recall (PPOK)	97.20%	0.65%	Sangat tinggi, menunjukkan model jarang melewatkan kasus PPOK aktual.
F1-*Score* (PPOK)	97.00%	0.68%	F1-*Score* yang sangat baik dan stabil untuk kelas mayoritas PPOK.
Presisi (TBC)	75.50%	3.20%	Cukup stabil, namun dengan variasi yang sedikit lebih tinggi antar *fold*.
Recall (TBC)	71.00%	3.50%	Menunjukkan kemampuan model dalam mengidentifikasi TBC dengan konsistensi.
F1-*Score* (TBC)	73.20%	3.35%	Kinerja yang cukup baik dan lebih teruji pada kelas TBC.
Presisi (*Pneumonia*)	15.00%	7.50%	Masih sangat rendah dan menunjukkan volatilitas (variasi) yang tinggi antar *fold*, menegaskan kesulitan prediksi.
Recall (*Pneumonia*)	18.00%	8.00%	Tetap rendah dan tidak stabil, menandakan model sering gagal mengidentifikasi kasus *Pneumonia* aktual.

F1-*Score* (*Pneumonia*)	16.40%	7.70%	F1-*Score* terendah, mengindikasikan masalah serius pada prediksi *Pneumonia* secara konsisten.

Kurva ROC/AUC (10-*Fold Cross-Validation*)

Pendekatan 10-*fold cross-validation* memberikan estimasi kinerja model yang lebih *robust* dan kurang bias dibandingkan *split validation* tunggal, karena ini mempertimbangkan bagaimana model berkinerja di berbagai subset data. Ketika melihat Kurva ROC/AUC dari 10-*fold cross-validation*, kita tidak hanya melihat kinerja rata-rata tetapi juga variabilitas kinerja tersebut.



Gambar 2. Kurva ROC/AUC Model Naive Bayes (10-*Fold Cross-Validation*)

Untuk menjelaskan hasil klasifikasi model Naïve Bayes, kita memanfaatkan perhitungan *probabilitas posterior* dengan menerapkan Teorema Bayes. Ini krusial untuk memahami seberapa besar kemungkinan sebuah *instance* (atau data) termasuk dalam kelas tertentu berdasarkan atribut yang dimilikinya. Proses perhitungannya menggunakan rumus sebagai berikut :

$$P(P|Y) = \frac{P(X|Y).P(Y)}{P(X)}$$

a) Menghitung Probabilitas PPOK jika Diprediksi PPOK dari confusion matrix:
Langkah-langkah :
- $P(X|Y) = \frac{1380}{1410} = 0.9787$
- $P(Y) = \frac{1410}{1555} = 0.9068$
- $P(X) = \frac{1420}{1555} = 0.9132$
- $P(Y|X) = \frac{0.9787 \text{ X } 0.9068}{0.9132} = 0.9718$

Hasil :
P(Y = PPOK | X = PPOK) = 0.9718 atau 97.18%

b) Menghitung Probabilitas TBC jika Diprediksi TBC dari confusion matrix
- $P(X|Y) = \frac{95}{130} = 0.7308$
- $P(Y) = \frac{130}{1555} = 0.0836$
- $P(X) = \frac{122}{1555} = 0.0785$
- $P(Y|X) = \frac{0.7308 \text{ X } 0.0836}{0.0785} = 0.7787$

Hasil :
P(Y = TBC | X = TBC) = 0.7787 atau 77.87%

c) Menghitung Probabilitas *Pneumonia* jika Diprediksi *Pneumonia* dari confusion matrix
- $P(X|Y) = \frac{3}{15} = 0.20$
- $P(Y) = \frac{15}{1555} = 0.0096$
- $P(X) = \frac{13}{1555} = 0.0084$
- $P(Y|X) = \frac{0.20 \text{ X } 0.0096}{0.0084} = 0.1875$

Hasil :
P(Y = *Pneumonia* | X = *Pneumonia*) = 0.1875 atau 18.75%

3. Pembahasan Hasil

Model Naive Bayes mencapai akurasi keseluruhan yang tinggi, yaitu 94.31% pada *Split Validation* dan 93.50% pada 10-*fold* cross-validation. Akurasi yang tinggi ini terutama didorong oleh kemampuan model yang sangat baik dalam memprediksi kelas mayoritas, yaitu PPOK, dengan F1-*Score* mencapai 97.52% (*Split Validation*) dan 97.00% (*cross-validation*). Hal ini mengindikasikan bahwa model sangat efektif dalam mengidentifikasi pasien dengan PPOK, yang merupakan mayoritas dalam dataset.

Untuk kelas TBC, kinerja model juga cukup baik, dengan F1-*Score* sekitar 75.39% (*split validation*) dan 73.20% (*cross-validation*). Meskipun tidak setinggi PPOK, angka ini menunjukkan bahwa strategi penanganan ketidakseimbangan kelas dengan SMOTE cukup berhasil dalam membantu model belajar pola dari kelas TBC yang lebih kecil. Ini penting karena TBC adalah penyakit yang memerlukan

diagnosis akurat untuk penanganan yang tepat dan pencegahan penularan.

Namun, kinerja model sangat lemah untuk prediksi *Pneumonia*, dengan F1-*Score* hanya 19.35% (*Split Validation*) dan 16.40% (cross-validation). Angka ini menunjukkan bahwa model mengalami kesulitan besar dalam mengenali dan mengklasifikasikan kasus *Pneumonia* secara akurat. Meskipun SMOTE telah diterapkan, jumlah sampel *Pneumonia* yang sangat sedikit dalam dataset awal (hanya 69 dari 5203 pasien) kemungkinan besar menjadi penyebab utama rendahnya kinerja ini. Keterbatasan data untuk kelas minoritas ekstrem menyulitkan algoritma Naive Bayes untuk membangun probabilitas kondisional yang robust, sehingga prediksi untuk kelas ini menjadi tidak andal. Deviasi standar yang tinggi pada metrik *Pneumonia* dalam *cross-validation* (7.50) juga mengindikasikan inkonsistensi kinerja model untuk kelas ini di berbagai subset data.

Atribut demografi seperti usia (terutama kelompok usia di atas 45 tahun) dan jenis kelamin (laki-laki) terbukti berperan dalam menjelaskan prevalensi penyakit paru secara umum dalam dataset ini. Kode ICD sebagai representasi langsung dari diagnosis juga sangat penting dalam klasifikasi.

Secara keseluruhan, model Naive Bayes menunjukkan potensi sebagai alat pendukung keputusan awal yang efektif untuk diagnosis PPOK dan TBC. Namun, untuk diagnosis *Pneumonia*, model ini masih belum dapat diandalkan dan memerlukan perbaikan signifikan, mungkin dengan penambahan lebih banyak data untuk kelas *Pneumonia* atau eksplorasi fitur lain yang lebih spesifik untuk penyakit tersebut.

V. KESIMPULAN

Penelitian ini berhasil menerapkan algoritma Naive Bayes untuk memprediksi jenis penyakit paru (PPOK, TBC, dan *Pneumonia*) di Rumah Sakit Umum Daerah (RSUD) dr. Soeratno Gemolong menggunakan data rekam medis pasien. Model menunjukkan akurasi keseluruhan yang baik, yaitu 94.31% (*Split Validation*) dan 93.50% (10-*fold cross-validation*). Kinerja sangat baik dalam memprediksi PPOK (F1-*Score* di atas 97%) dan cukup baik untuk TBC (F1-*Score* sekitar 73-75%). Namun, model memiliki keterbatasan signifikan dalam memprediksi *Pneumonia* (F1-*Score* sangat rendah, sekitar 16-19%) karena jumlah data sampel yang sangat sedikit untuk kelas tersebut, meskipun telah ditangani dengan SMOTE. Atribut usia di atas 45 tahun dan jenis kelamin laki-laki terbukti relevan sebagai faktor demografi yang berasosiasi dengan penyakit paru dalam dataset studi. Model ini berpotensi sebagai alat pendukung keputusan untuk diagnosis awal PPOK dan TBC, namun memerlukan pengembangan lebih lanjut, khususnya untuk diagnosis *Pneumonia*.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Rumah Sakit Umum Daerah (RSUD) dr. Soeratno Gemolong yang telah memberikan akses data rekam medis pasien sebagai bahan utama dalam penelitian ini. Ucapan terima kasih juga disampaikan kepada Universitas Duta Bangsa Surakarta atas dukungan fasilitas dan bimbingan akademik selama proses penyusunan jurnal ini.

Kami juga mengapresiasi para dosen pembimbing dan seluruh pihak yang telah memberikan masukan, kritik, serta saran yang sangat berarti dalam penyempurnaan penelitian ini. Semoga hasil penelitian ini dapat memberikan kontribusi nyata dalam pengembangan sistem pendukung keputusan di bidang kesehatan, khususnya dalam diagnosis penyakit paru.

REFERENSI

[1] World Health Organization. (2024, May 2). The top 10 causes of death. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] American Thoracic Society. (2023). Key Facts About Lung Disease. https://www.thoracic.org/patients/patient-resources/resources/key-facts-about-lung-disease.pdf

[3] Holmes, J. H. (2022). Leveraging big data analytics in healthcare for improved patient outcomes. *Journal of Health Information Systems*, *15*(3), 201–215.

[4] Bhatia, K. P., & Singh, V. K. (2021). Predictive analytics in electronic health records using deep learning: A systematic review. *Journal of Medical Systems*, *45*(3), 1–15.

[5] Smith, J. (2022, October). The role of classification algorithms in clinical decision support systems. Dalam *Proceedings of the International Conference on Health Informatics*, London, UK (hlm. 112–118).

[6] Devi, R. S. (2020). Data mining in healthcare: Opportunities and challenges. *International Journal of Advanced Research in Computer Engineering & Technology*, *9*(6), 227–231.
[7] Firmansyah, M. (2024). Improve accuracy in the process of diagnosing various types of lung diseases by using the Naïve Bayes *classifier*. *IJISTECH (International Journal of Information System and Technology)*.
[8] Verma, A. R., & Singh, R. K. (2023, September). Performance analysis of Naive Bayes *classifier* on large datasets. Dalam *Proceedings of the International Conference on Advanced Computing & Communications*, Jaipur, India (hlm. 200–205).
[9] Singh, R. A., Das, S., & Gupta, S. (2023). *Integration of machine learning in clinical decision support systems*. ResearchGate.
[10] Karlena dkk (2022), "Market Basket Analysisuntuk Mengetahui Pola Beli Konsumen Roofbox Mobil menggunakan Algoritme Apriori", Jurnal Teknologika Volume 12 No. 2 November 2022, 308-316
[11] Wulandari, F., Jusia, P. A., & Jasmir. (2020). Klasifikasi Data Mining Untuk Mendiagnosa Penyakit ISPA Menggunakan Metode Naïve Bayes Pada Puskesmas Jambi Selatan. Jurnal Manajemen Teknologi Dan Sistem Informasi (JMS). 2(3): 214–227.
[12] Padilah, T. N., & Adam, R. I. (2019). Analisis Regresi Linier Berganda Dalam Estimasi Produktivitas Tanaman Padi Di Kabupaten Karawang. *FIBONACCI: Jurnal Pendidikan Matematika dan Matematika*, *5*(2), 117. https://doi.org/10.24853/fbc.5.2.117-128
[13] A. Arum Sari and P. Pramono, "Penerapan Algoritma Deep Belief Networks (DBNs) Untuk Prediksi Kanker Serviks," Feb. 2024. doi: https://doi.org/10.47701/dutacom.v17i1.3790