Jurnal of Data Science Methods and Applications
Vol. 01, No. 01, April 2025, pp. 33-41
E-ISSN: xxxx-xxxx, DOI: 10.30873/jodmapps.v1i1.pp33-41

Prediksi Penyakit Paru-Paru MenggunakanAlgoritma Naïve Bayes

Selvida Widi Audria[1], Izzah Farikhah[2], Reza Maulana Saputra[3], Neni Purwati*[4]

[1,2,3,4]Sistem Informasi, Universitas YPPI Rembang, Jl. Raya Rembang-Pamotan KM. 4, Rembang, 59261
email: [1]selvidawidiaudria@gmail.com; [2] izzahfarikhah495@gmail.com; [3]rezamalnas@gmail.com
[4]nenipurwati@uyr.ac.id

Article History:
Received: DD-MM-20XX
Revised : DD-MM-20XX
Accepted: DD-MM-20XX
Online : DD-MM-20XX

A B S T R A K

Paru-paru memiliki peran penting dalam tubuh manusia, yaitu sebagai organ utama dalam sistem pernapasan, berfungsi mengolah karbondioksida yang dibawa oleh darah menjadi oksigen dari udara yang dihirup, yang kemudian disebarkan ke seluruh tubuh untuk memenuhi kebutuhan oksigen. Gangguan paru-paru juga berisiko terhadap Kesehatan hingga kematian. Diperlukan metode yang akurat untuk mendiagnosis penyakit paru-paru agar penanganan dapat dilakukan dengan tepat. *Dataset* yang digunakan sebanyak 10000 dengan 10 atribut yakni usia, jenis kelamin, kebiasaan merokok, status pekerjaan, kondisi rumah tangga, aktivitas begadang, aktivitas rumah tangga, kepemilikan asuransi, riwayat penyakit bawaan, dan label hasil. Tujuan penelitian ini adalah untuk memprediksi penyakit paru-paru menggunakan model *naïve bayes* dengan hasil validasi yang *robust* dan *reliable* dengan penerapan *dual-validation framework* yakni *split validation* dan *k-fold cross validation.* Metode yang digunakan adalah pengumpulan data, pengelolaan data (seleksi dan pembersihan data), penerapan metode, pengujian metode dan kesimpulan. Hasil dari penerapan model *naïve bayes* dari *data testing* sebanyak 2000 menunjukkan nilai *accuracy* tertinggi sebesar 86.90%, *precision* tertinggi sebesar 87.65% diperoleh dari iterasi sebanyak *10 Fold Cross Validation*, sedangkan nilai *recall* tertinggi diperoleh dari penerapan *split validation* sebesar 87.75%, sehingga hasil tersebut termasuk klasifikasi yang sangat baik diterapkan untuk melakukan prediksi penyakit paru-paru yang di derita masyarakat.

Kata Kunci: Naïve Bayes; Split Validation; Cross Validation; Paru-paru

A B S T R A C T

The lungs have an important role in the human body, namely as the main organ in the respiratory system, functioning to process carbon dioxide carried by the blood into oxygen from the inhaled air, which is then distributed throughout the body to meet oxygen needs. Lung disorders are also a risk to health and even death. An accurate method is needed to diagnose lung disease so that treatment can be carried out properly. The dataset used was 10,000 with 10 attributes, namely age, gender, smoking habits, employment status, household conditions, staying up late, household activities, insurance ownership, history of congenital diseases, and result labels. The purpose of this study was to predict lung disease using the naïve Bayes model with robust and reliable validation results by applying a dual-validation framework, namely Split validation and k-Fold Cross Validation. The methods used are data collection, data management (data selection and cleaning), method application, method testing and conclusions. The results of the application of the naïve Bayes model from 2000 testing data showed the highest accuracy value of 86.90%, the highest precision of 87.65% obtained from 10 iterations of Fold Cross Validation, while the highest recall value was obtained from the application of Split Validation of 87.75%, so that these results include a very good classification applied to predict lung diseases suffered by the community.

Keywords: *Naïve Bayes; Split Validation; Cross Validation; Lungs*

1. PENDAHULUAN

Teknologi informasi telah memberikan kontribusi yang signifikan dalam berbagai aspek seperti bidang kesehatan, khususnya dalam penanganan penyakit [1]. Salah satu penyakit yang perlu perhatian serius adalah penyakit paru-paru sebagai bagian dari sistem pernapasan manusia. Paru-paru memiliki peran penting dalam tubuh manusia, yaitu sebagai organ utama dalam sistem pernapasan. Fungsinya adalah mengolah karbon dioksida yang dibawa oleh darah menjadi oksigen dari udara yang dihirup, yang kemudian disebarkan ke seluruh tubuh untuk memenuhi kebutuhan oksigen. Organ vital bernama paru-paru ini rentan terhadap berbagai masalah kesehatan yang dapat terjadi baik dalam jangka pendek maupun

jangka panjang, dan risiko gangguan paru-paru dapat dialami oleh siapa saja tanpa memandang usia atau jenis kelamin [2], hingga risiko kematian. Kelompok yang berisiko meliputi pria, wanita, anak-anak, perokok aktif, perokok pasif, bahkan mantan perokok. Salah satu faktor utama yang meningkatkan risiko penyakit paru-paru di Indonesia adalah tingginya angka konsumsi tembakau dan kebiasaan merokok [3]. Penyakit Paru-paru Obstruktif Kronis (PPOK) disebabkan oleh peradangan paru-paru yang kemudian berkembang dalam jangka waktu yang lama, menghalangi aliran udara dari paru-paru, akibatnya mengalami kesulitan bernapas karena pembengkakan dan produksi lendir atau dahak. Risiko PPOK meningkat pada orang yang sudah lanjut usia dan yang merokok aktif dalam jangka waktu yang lama, dan menyerang orang yang merokok secara aktif maupun pasif sebagai akibat dari paparan asap rokok [4]. Gejala gangguan kesehatan paru-paru yang mirip seperti asma, bronkitis, dan tuberkulosis, menyulitkan diagnosis bagi petugas medis, sehingga diperlukan metode yang akurat untuk mendiagnosis penyakit paru-paru agar penanganan dapat dilakukan dengan tepat [5].

Beberapa penelitian yang telah dilakukan tentang penyakit paru-paru dengan berbagai macam algoritma *machine learning* diantaranya oleh Shafara Rahmaeda, dan Rastri Prathivi berjudul Komparasi Metode SVM dan Logistic Regression untuk Klasifikasi Hipotesa Penyakit Kanker Paru Paru Berdasarkan Gejala Awal, ditemukan *performa* yang digunakan dalam memprediksi model adalah *Accuracy*, *Precission*, *Recall*, dan *F1-Score*, lalu memperoleh nilai *Accuarcy* pada Algoritma Logistic Regression sebesar 97.85%, hal ini menjelaskan bahwa algoritma logistic regression memiliki performa lebih baik dalam melakukan klasifikasi penyakit kanker paru-paru dibandingkan algoritma SVM [6].

Penelitian selanjutnya dilakukan oleh Gifthera Dwilestari, dan Turfa Azmi Afifah dengan judul Perbandingan Kinerja Algoritma Naive Bayes dan Decision Tree Dalam Klasifikasi Kanker Paru-Paru, menunjukkan bahwa Naive Bayes memiliki akurasi lebih tinggi sebesar 92,47% dan unggul dalam mendeteksi kelas positif (recall sebesar 98,77%), tetapi kurang optimal pada kelas negatif. Sementara itu, Decision Tree menunjukkan akurasi 88,17% dengan deteksi keseimbangan antar kelas yang lebih baik (recall kelas negative sebesar 66,67%). Sehingga Naive Bayes lebih cocok untuk aplikasi yang membutuhkan deteksi cepat dan efisien, sedangkan Decision Tree lebih sesuai untuk skenario dengan kebutuhan deteksi keseimbangan antar kelas [7].

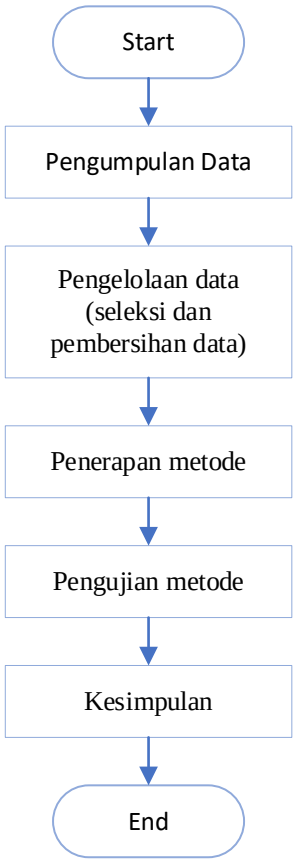
Penelitian lain yang telah dilakukan oleh Pungkas Budiyono, Suwanda Aditya Saputra, Beni Rahmatullah, dan Dikdik Permana Wigandi tentang Penerapan Algoritma Naive Bayes Untuk Prediksi Penyakit Paru-Paru Pada Sumber Kaggle Menggunakan Aplikasi Rapid Miner, diperoleh hasil bahwa dalam *Performance Vector* hasil akurasi yang dihasilkan adalah 87.10%, Class Recall yang dihasilkan adalah 87,09% dan class presisi yang dihasilkan adalah 87,06%. Nilai akurasi sebesar 87.10% menunjukkan kinerja yang baik dalam memprediksi sejumlah kasus positif penyakit paru-paru[8].

Penelitian yang relevan juga dilakukan oleh Ade Christian, Hariyanto, Ahmad Yani, Sumanto dengan judul Analisis *Machine Learning* Untuk Prediksi Penyakit Paru-Paru Menggunakan *Random Forest* menghasilkan kinerja yang sangat baik dengan tingkat akurasi 94,7%, *F1-score* 0.946, *Precision* 0.952, dan *Recall* 0.947. Hal ini membuktikan bahwa *Random Forest* mampu menggeneralisasi data dengan baik dan memberikan prediksi yang andal, menjadikannya metode yang layak digunakan dalam pengembangan model prediksi penyakit paru-paru atau kasus serupa [9].

Metode Naïve Bayes untuk melakukan penambangan data telah banyak digunakan karena mempertimbangkan beberapa keunggulannya yakni mencapai akurasi yang tinggi walau hanya pada data *testing* yang sedikit [10], dapat menangani kompleksitas dataset dan memberikan prediksi yang kredibel [11], sebuah teknik klasifikasi yang efisien untuk data kategori [12], metode klasifikasi sederhana yang menghitung peluang berdasarkan kombinasi data yang ada [13], mampu melakukan klasifikasi dengan menampilkan hasilnya secara otomatis dan akurat [14], kepadatan dan kecepatan dalam penambangan data [15], serta mampu memprediksi kemungkinan masa depan berdasarkan pengalaman periode sebelumnya [16]. Naïve Bayes merupakan algoritma yang sering digunakan untuk klasifikasi dan prediksi menggunakan peluang dengan memilih hasil akhir yang memiliki peluang terbesar untuk memasukan data baru ke dalam kelasnya [17]. Metode Naïve Bayes dan Naïve Bayes *Updateable* dalam *statistical learning* telah terbukti memberikan hasil yang efisien dengan akurasi yang tinggi dibandingkan dengan model pembelajaran fungsional lainnya [18], dan algoritma ini cukup populer hingga masuk ke dalam kategori sepuluh algoritma teratas dalam penambangan data[19]. Algoritma Nave Bayes adalah salah satu algoritma pembelajaran induktif yang paling efektif dan efisien untuk pembelajaran mesin dan penambangan data, kinerjanya kompetitif dalam proses klasifikasi meskipun menggunakan asumsi independensi atribut (tidak ada hubungan antar atribut) [20].

2. METHODS

Metode yang digunakan dalam penelitian ini ada lima tahapan yakni pengumpulan data, pengelolaan data (seleksi dan pembersihan data), penerapan metode, pengujian metode dan kesimpulan, dapat dilihat pada gambar *flowchart* berikut:



Gambar 1. *Flowchart* Tahapan Penelitian

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

1. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini berasal dari situs *Kaggle* [21] dengan jumlah data sebanyak 10.000 entri. Atribut-atribut yang digunakan sebanyak 10 meliputi usia, jenis kelamin, kebiasaan merokok, status pekerjaan, kondisi rumah tangga, aktivitas begadang, aktivitas rumah tangga, kepemilikan asuransi, riwayat penyakit bawaan, serta target atau label hasil.

2. Pengelolaan Data (seleksi dan pembersihan data)

Data yang berasal dari *Kaggle* akan digunakan semua untuk analisis data, tahap pengelolaan data menggunakan aplikasi RapidMiner, dengan memanfaatkan operator yang ada untuk mengecek obyek agar siap digunakan.

3. Penerapan Metode

Metode yang digunakan dalam penelitian ini adalah *Naive Bayes*, dan diproses menggunakan data *training* dan data *testing* yang sudah ditentukan untuk mengetahui hasil prediksi dari penerapan metode. Proses dimulai dengan melakukan pembagian data dengan menentukan data *training* dan data *testing*, kemudian diproses dengan menggunakan metode *Naive Bayes*. Data *training* digunakan untuk melatih model, memperkirakan berbagai parameter atau membandingkan kinerja berbagai model; data *testing*

digunakan untuk mengevaluasi atau menguji data, dari data pelatihan dan pengujian dibandingkan untuk memeriksa bahwa model akhir berfungsi dengan benar [22].

4. Pengujian Metode

Tahap pengujian metode akan digunakan *tools* RapidMiner dimana hasil dari pengujian data dengan *split validation* dan *k-fold cross validation*. Proses pengujian juga dilakukan dengan menilai kinerja metode Naïve Bayes yang telah dilakukan, dengan menggunakan *confusion matriks* dan *performa metode* (Akurasi, Presisi dan Recall). *Confusion Matrix* adalah suatu tabel yang menyatakan tentang jumlah data pengujian yang terklasifikasi dengan benar dan jumlah data pengujian yang terklasifikasi salah, dengan parameter TP(*true positive*), FN(*false negative*), TN(*true negative*) dan FP(*false positive*)[23]. Proses evaluasi selanjutnya menggunakan metrik seperti akurasi, presisi, recall, untuk memastikan bahwa model memiliki tingkat keakuratan yang memadai serta mampu menghasilkan prediksi yang efektif[24].

5. Kesimpulan

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

3. HASIL DAN PEMBAHASAN

3.1. Pengumpulan Data

Dataset yang digunakan sebanyak 10.000, dengan data label positif (tidak) sebanyak 5227 dan data label negatif (ya) sebanyak 4773, dengan detail data sebagai berikut:

No	Usia	Jenis_Kelamin	Merokok	Bekerja	Rumah_Tangga	Aktivitas_Begadang	Aktivitas_Olahraga	Asuransi	Penyakit_Bawaan	Hasil
1	Tua	Pria	Pasif	Tidak	Ya	Ya	Sering	Ada	Tidak	Ya
2	Tua	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Ada	Tidak
3	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak	Tidak
4	Tua	Pria	Aktif	Ya	Tidak	Tidak	Jarang	Ada	Ada	Tidak
5	Muda	Wanita	Pasif	Ya	Tidak	Tidak	Sering	Tidak	Ada	Ya
6	Muda	Wanita	Pasif	Ya	Tidak	Tidak	Sering	Tidak	Ada	Tidak
7	Tua	Wanita	Pasif	Tidak	Ya	Tidak	Sering	Tidak	Tidak	Ya
8	Muda	Pria	Aktif	Tidak	Ya	Ya	Sering	Tidak	Tidak	Tidak
9	Tua	Wanita	Aktif	Ya	Ya	Ya	Jarang	Ada	Ada	Ya
10	Muda	Wanita	Pasif	Ya	Tidak	Ya	Jarang	Ada	Ada	Ya
11	Tua	Wanita	Pasif	Ya	Ya	Tidak	Sering	Ada	Ada	Ya
12	Tua	Wanita	Aktif	Tidak	Ya	Tidak	Jarang	Ada	Tidak	Tidak
13	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak	Tidak
14	Tua	Wanita	Aktif	Ya	Tidak	Ya	Jarang	Ada	Ada	Tidak
15	Muda	Wanita	Pasif	Ya	Tidak	Ya	Sering	Tidak	Ada	Ya
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
9995	Tua	Pria	Aktif	Ya	Tidak	Tidak	Jarang	Ada	Ada	Tidak
9996	Muda	Wanita	Pasif	Ya	Tidak	Tidak	Sering	Tidak	Ada	Ya
9997	Muda	Wanita	Pasif	Ya	Tidak	Ya	Jarang	Ada	Ada	Ya
9998	Tua	Wanita	Pasif	Ya	Ya	Tidak	Sering	Ada	Ada	Ya
9999	Tua	Wanita	Aktif	Tidak	Ya	Tidak	Jarang	Ada	Tidak	Tidak
10000	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak	Tidak

Gambar 2. *Dataset* (Sumber: Kaggle.com)

Seluruh attribut berjenis data binomial, kecuali Aktivitas Olahraga berjenis kategori. Attribut Usia (Tua dan Muda), Jenis Kelamin (Pria dan Wanita), Merokok (Aktif dan Pasif), Bekerja (Ya dan Tidak), Rumah Tangga (Ya dan Tidak), Aktivitas Begadang (Ya dan Tidak), Aktivitas Olahraga (Sering dan Jarang), Asuransi (Ada dan Tidak), Penyakit Bawaan (Ada dan Tidak), serta label Hasil (Ya dan Tidak).

3.2. Pengelolaan Data

Proses yang dilakukan pada tahap ini diawali dengan melakukan pengecekan data yang mengalami *missing value* dengan menggunakan operator yang disediakan oleh RapidMiner dan untuk memastikan kondisi data dapat melihat *result history* setelah dilakukan *run*, pada operator *read csv* sesuai dengan tipe file sumbernya berjenis .csv pada RapidMiner. Hasil pengecekan *dataset* yang digunakan menunjukkan bahwa tidak ditemukan data yang *missing value*, tidak ada data yang inkonsisten dan tidak ada data yang *invalid*, sehingga data telah siap untuk diproses selanjutnya. Adapun hasil pengecekan kondisi data dapat dilihat pada Gambar 3.

Name	Type	Missing	Statistics		
Metadata **Hasil**	Binominal	0	Negative: Ya	Positive: Tidak	Values: Tidak (5227), Ya (4773)
Usia	Binominal	0	Negative: Tua	Positive: Muda	Values: Muda (5114), Tua (4886)
Jenis_Kelamin	Binominal	0	Negative: Pria	Positive: Wanita	Values: Wanita (7384), Pria (2616)
Merokok	Binominal	0	Negative: Pasif	Positive: Aktif	Values: Aktif (5087), Pasif (4913)
Bekerja	Binominal	0	Negative: Tidak	Positive: Ya	Values: Ya (6301), Tidak (3699)
Rumah_Tangga	Binominal	0	Negative: Ya	Positive: Tidak	Values: Ya (5159), Tidak (4841)
Aktivitas_Begadang	Binominal	0	Negative: Ya	Positive: Tidak	Values: Ya (5869), Tidak (4131)
Aktivitas_Olahraga	Binominal	0	Negative: Sering	Positive: Jarang	Values: Jarang (6004), Sering (3996)
Asuransi	Binominal	0	Negative: Ada	Positive: Tidak	Values: Ada (7090), Tidak (2910)
Penyakit_Bawaan	Binominal	0	Negative: Tidak	Positive: Ada	Values: Ada (6438), Tidak (3562)

Filter (10 / 10 attributes): Search for Attributes

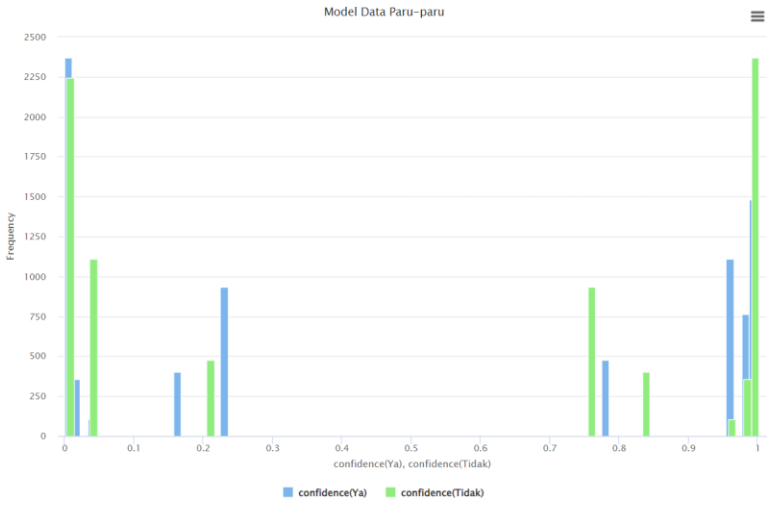
Showing attributes 1 - 10 Examples: 10,000 Special Attributes: 1 Regular Attributes: 9

Gambar 3. Hasil Pengecekan *Dataset*

Jumlah data pada gambar di atas menjelaskan bahwa attribut Usia (Tua 4886 dan Muda 5114), Jenis Kelamin (Pria 2616 dan Wanita 7384), Merokok (Aktif 5087 dan Pasif 4913), Bekerja (Ya 6301 dan Tidak 3699), Rumah Tangga (Ya 5159 dan Tidak 4841), Aktivitas Begadang (Ya 5869 dan Tidak 4131), Aktivitas Olahraga (Sering 3996 dan Jarang 6004), Asuransi (Ada 7090 dan Tidak 2910), Penyakit Bawaan (Ada 6438 dan Tidak 3562), serta label Hasil (Ya 4773 dan Tidak 5227).

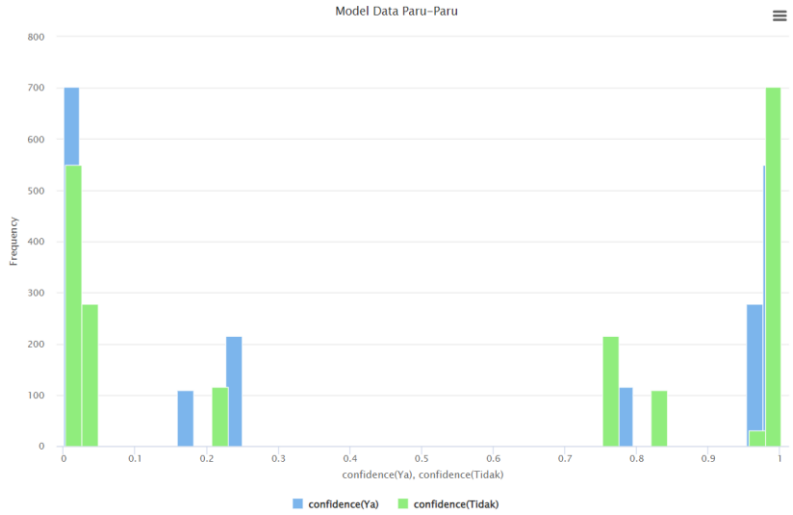
3.3. Penerapan Metode

Tahap ini operator yang digunakan adalah *split data* yang berfungsi untuk membagi data menjadi dua yakni *data training* sebesar 80% sebanyak 8000 data, dan *data testing* sebesar 20% sebanyak 2000 data. Model *Naïve Bayes* yang diterapkan pada *data training* sebanyak 8000 data menghasilkan *value label* hasil positif (tidak) sebanyak 4182 dan negatif (ya) sebanyak 3818, sedangkan pada *value prediction* hasil label positif (tidak) 4167 dan negatif (ya) sebanyak 3833. Visualisasi hasil penerapan model pada *data training* dapat dilihat pada Gambar 4.



Gambar 4. Visualisasi *Data Training*

Selanjutnya dilakukan penerapan model terhadap *data testing* sebesar 20% sebanyak 2000 data. Pengujian yang telah dilakukan dengan operator *apply model* menunjukan bahwa hasil *data testing* 2000 menghasilkan *value* label hasil positif (tidak) sebanyak 1.045 dan negatif (ya) sebanyak 955, sedangkan pada *value prediction* hasil label positif (tidak) 1.057 dan negatif (ya) sebanyak 943. Visualisasi hasil pengujian terhadap *data testing* dapat dilihat pada Gambar 5.



Gambar 5. Visualisasi *Data Testing*

3.4. Pengujian Metode

Proses selanjutnya dilakukan pengujian hasil dari penerapan model *Naïve Bayes* dengan *cross validation* yang berfungsi untuk memvalidasi hasil prediksi. Penerapan *dual-validation framework* berfungsi untuk menghubungkan pengujian antara *Split Validation* dengan *K-Fold Cross Validation* agar dapat memberikan evaluasi model yang lebih *robust* dan *reliable*, dengan meminimalisasi bias dalam pengukuran performa model [25],[26].

3.4.1. Split Validation

Validasi model dengan cara *split validation* yang dilakukan menghasilkan *confusion matrix* pengujian metode dari data testing dengan rincian sebagai berikut:

	true Ya	true Tidak
pred. Ya	815	128
pred. Tidak	140	917

Gambar 6. Confusion Matrix (*Split Validation*)

Pada gambar di atas menjelaskan bahwa TP (*true positive*) jumlah data sebanyak 815 yang diklasifikasikan pada kelas positif dan kelas sebenarnya positif, TN (*true negative*) sebanyak 917 jumlah data yang diklasifikasikan pada kelas Negatif dan kelas sebenarnya negatif, FN (*false negative*) sebanyak 128 jumlah data yang diklasifikasikan pada kelas negatif dan kelas sebenarnya positif, dan FP (*false positive*) sebanyak 140 jumlah data yang diklasifikasikan pada kelas positif dan kelas sebenarnya negatif.

Hasil *confusion matrix* di atas selanjutnya digunakan sebagai nilai untuk menemukan akurasi, presisi, recall seperti dalam tabel berikut:

Tabel 1. *Performance* Hasil *Split Validation*

Performance	**Nilai**
Accuracy	86.60%
Precision	86.75% (*positive class*: Tidak)
Recall	87.75% (*positive class*: Tidak)

3.4.2. K-Fold Cross Validation

K-fold merupakan parameter metode *cross validation*, yang bekerja tidak hanya membuat beberapa sampel data uji berulang kali, tetapi membagi dataset menjadi bagian terpisah dengan ukuran yang sama; model dilatih oleh subset data latih dan divalidasi oleh subset validasi (data uji) sebanyak *k*. Hasil akurasi *cross validation* memiliki standar deviasi atau simpangan baku yaitu ukuran penyebaran data yang menunjukan jarak rata-rata dari nilai tengah ke suatu titik nilai; semakin besar simpangan baku yang dihasilkan, maka penyebaran dari nilai tengahnya juga besar, begitu pula sebaliknya; tujuannya untuk melihat jarak antara rata-rata akurasi dengan akurasi setiap percobaan (iterasi) [27]. Pada penelitian ini ditentukan k sebanyak 10 yang berarti bahwa dilakukan iterasi sebanyak 10 kali pada *cross validation* dari data testing, sehingga menghasilkan *confusion matrix* dari model sebagai berikut:

	true Ya	true Tidak
pred. Ya	825	132
pred. Tidak	130	913

Gambar 6. Confusion Matrix (*k-Fold Cross Validation*)

Gambar 6. Confusion matrik diatas menjelaskan bahwa ditemukan nilai TP (*true positive*) sebanyak 825 sebagai data yang diklasifikasikan pada kelas positif dan kelas sebenarnya positif, TN(*true negative*) sebanyak 913 sebagai data yang diklasifikasikan pada kelas negatif dan kelas sebenarnya negatif, FN (*false negative*) sebanyak 132 sebagai data yang diklasifikasikan pada kelas negatif dan kelas sebenarnya positif, dan FP (*false positive*) sebanyak 130 sebagai data yang diklasifikasikan pada kelas positif dan kelas sebenarnya negatif.

Hasil dari pengecekan validasi menggunakan *cross validation* dari *confusion matrix* di atas sebagai nilai untuk menghasilkan nilai rata-rata *performance* metode pada tabel berikut:

Tabel 2. Performance Hasil Cross Validation

Performance	**Nilai *Performance***	**Nilai Standar Deviasi**
Accuracy	86.90%	+/- 1.63% (*micro average:* 86.90%)
Precision	87.65%	+/- 2.88% (*micro average:* 87.54%) (*positive class:* Tidak)
Recall	87.37%	+/- 2.63% (*micro average:* 87.37%) (*positive class:* Tidak)

4. KESIMPULAN

Pada penerapan model dapat ditemukan pola atau aturan yang dapat digunakan, dan nilai *accuracy*, *precision* dan *recall* digunakan untuk mengukur kinerja algoritma atau model Naïve Bayes dari *dataset* sebanyak 2000 *data testing*. Hasil dari percobaan yang telah dilakukan terhadap model Naïve Bayes pada data Paru-Paru dengan penerapan *dual-validation framework* yakni *split validation* dan *k-fold cross validation* menghasilkan nilai *accuracy* tertinggi sebesar 86.90%, *precision* tertinggi sebesar 87.65% diperoleh dari iterasi sebanyak *10 fold cross validation*, sedangkan nilai *recall* tertinggi diperoleh dari penerapan *split validation* sebesar 87.75%, sehingga hasil tersebut termasuk klasifikasi yang sangat baik diterapkan untuk melakukan prediksi penyakit paru-paru yang diderita masyarakat.

DAFTAR PUSTAKA

[1] F. Ramadhana, Fauziah, and Winarsih, "Aplikasi Sistem Pakar untuk Mendiagnosa Penyakit ISPA Menggunakan Metode Naive Bayes Berbasis Website," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.*, vol. 4, no. 3, pp. 320–329, 2020, doi: 10.30998/string.v4i3.5441.

[2] F. M. A. Sofyan, A. Voutama, and Y. Umaidah, "Penerapan Algoritma C4.5 Untuk Prediksi Penyakit Paru-paru Menggunakan Rapidminer," *JATI(Jurnal Mhs. Tek. Inform.*, vol. 7, no. 2, pp. 1409–1415, 2023, doi: https://doi.org/10.36040/jati.v7i2.6810.

[3] D. B. Dinova and B. Prasetiyo, "Implementasi Random Forest dalam Klasifikasi Kanker Paru-Paru," *JOINTER–JOURNAL INFORMATICS Eng.*, vol. 5, no. 1, pp. 27–31, 2024, doi: https://doi.org/10.53682/jointer.v5i01.331.

[4] Rumah Sakit Islam Ahmad Yani, "Merokok Pencetus Utama Penyakit Paru Obstruktif Kronis," Rumah Sakit Islam Surabaya. [Online]. Available: https://rsisurabaya.com/merokok-pencetus-

utama-penyakit-paru-obstruktif-kronis/
[5] R. A. Karima and F. Zaehol, "Implementasi Metode K-Nearest Neighbor Untuk Klasifikasi Penyakit Paru-paru Pada Anak," *J. Ilm. MULTIDISIPLIN ILMU*, vol. 1, no. 6, pp. 10–17, 2024, doi: https://doi.org/10.69714/yx4smf68.
[6] S. Rahmaeda and R. Prathivi, "Komparasi Metode SVM dan Logistic Regression untuk Klasifikasi Hipotesa Penyakit Kanker Paru Paru Berdasarkan Gejala Awal," *KESATRIA J. Penerapan Sist. Inf. (Komputer Manajemen)*, vol. 6, no. 1, pp. 210–218, 2025, doi: https://doi.org/10.30645/kesatria.v6i1.562.g557.
[7] G. Dwilestari and T. A. Afifah, "Perbandingan Kinerja Algoritma Naive Bayes dan Decision Tree Dalam Klasifikasi Kanker Paru-Paru," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 1, pp. 801–807, 2025, doi: https://doi.org/10.36040/jati.v9i1.12463.
[8] P. Budiyono, S. A. Saputra, B. Rahmatullah, and D. P. Wigandi, "Penerapan Algoritma Naive Bayes Untuk Prediksi Penyakit Paru-Paru Pada Sumber Kaggle Menggunakan Aplikasi Rapid Miner," *J. Tek. Inform. dan Sist. Inf.*, vol. 11, no. 4, pp. 549–557, 2024, doi: https://doi.org/10.35957/jatisi.v11i4.9365.
[9] A. Christian, Hariyanto, A. Yani, and Sumanto, "Analisis Machine Learning Untuk Prediksi Penyakit Paru-Paru Menggunakan Random Forest," *J. Innov. Futur. Technol.*, vol. 7, no. 1, pp. 122–131, 2025, doi: https://doi.org/10.47080/iftech.v7i1.3906.
[10] D. Juliani and M. Soleh, "Implementasi Machine Learning untuk Klasifikasi Penyakit Kanker Paru Menggunakna Metode Naïve Bayes dengan Tambahan Fitur Chatbot," *J. IPTEK*, vol. 8, no. 2, pp. 12–17, 2024, [Online]. Available: https://jurnaliptek.iti.ac.id/index.php/jii/article/view/351/120
[11] P. C. Pradhani, A. S. Indrayani, N. Azzahra, S. E. Aflikha, F. Zahra, and A. D. Kalifia, "Prediksi Diabetes Mellitus Berdasarkan Data Pasien SYLHET Diabetes Hospital Dengan Metode Naive Bayes," *Sci. (Jurnal Ilm. Sain dan Teknol.*, vol. 3, no. 3, pp. 342–353, 2025, [Online]. Available: https://jurnal.researchideas.org/index.php/scientica/article/view/163/151
[12] R. Rusli and H. Harmayani, "Prediksi Hasil Pemilihan Umum Berdasarkan Data Media Sosial Menggunakan Teknik Data Mining Naive Bayes," *J. Sci. Soc. Res.*, vol. 8, no. 1, pp. 357 – 362, 2025, doi: https://doi.org/10.54314/jssr.v8i1.2259.
[13] E. D. K. Wardani, F. F. Yo, and W. N. Meylugita, "Implementasi Algoritma Naive Bayes Untuk Analisis Ulasan Pengguna Untuk Aplikasi SEABANK di Google Play Store," *J. Kecerdasan Buatan dan Teknol. Inf.*, vol. 4, no. 1, pp. 13–24, 2025, doi: https://doi.org/10.69916/jkbti.v4i1.193.
[14] N. Nurwati and Y. Santoso, "Analisis Perbandingan K-Nearest Neighbors Dan Naive Bayes Untuk Rekomendasi Pilihan Program Studi Bagi Mahasiswa," *Idealis Indones. J. Inf. Syst.*, vol. 8, no. 1, pp. 117–126, 2025, doi: https://doi.org/10.36080/idealis.v8i1.3344.
[15] B. Ravinder, S. K. S. Srinivasan, V. S. Prabhu, P. Asha, S. P. Maniraj, and C. Srinivasan, "Web Data Mining with Organized Contents Using Naive Bayes Algorithm," in *2024 2nd International Conference on Computer, Communication and Control (IC4)*, Indore, 2024. doi: 10.1109/IC457434.2024.10486403.
[16] J. F. Idris, R. Ramadhani, and M. M. Mutoffar, "Klasifikasi Penyakit Kanker Paru Menggunakan Perbandingan Algoritma Machine Learning," *J. MEDIA Akad.*, vol. 2, no. 2, pp. 1981–2000, 2024, doi: https://doi.org/10.62281/v2i2.145.
[17] A. Gunawan and Z. Fatah, "Klasifikasi Penerima Bantuan Sktm Menggunakan Algoritma Naive Bayes: Studi Kasus Desa Pesanggrahan," *J. Ris. Sist. Inf.*, vol. 2, no. 1, pp. 45–52, 2025, doi: https://doi.org/10.69714/6w34wq73.
[18] M. Yuichi and Y. A. Susetyo, "Klasifikasi Penyakit Migrain dengan Metode Naïve Bayespada Dataset Kaggle," *J. Indones. Manaj. Inform. dan Komunikasi(JIMIK)*, vol. 6, no. 1, pp. 139–151, 2025, doi: https://doi.org/10.35870/jimik.v6i1.1150.
[19] O. S. D. Fadhillah, J. H. Jaman, and C. Carudin, "No TitlePerbandingan Naive Bayes, Support Vector Machine, Logistic Regression dan Random Forest dalam Menganalisis Sentimen Mengenai Tiktokshop," *JITET (Jurnal Inform. dan Tek. Elektro Ter.*, vol. 13, no. 1, pp. 840–847, 2025, doi: http://dx.doi.org/10.23960/jitet.v13i1.5746.
[20] Jefria and Z. Fatah, "Klasifikasi Data Mining Untuk Memprediksi Kelulusan Mahasiswa Menggunakan Metode Naive Bayes," *J. Ilm. MULTIDISIPLIN ILMU*, vol. 2, no. 1, pp. 29–37, 2025, doi: https://doi.org/10.69714/mhjq1v85.
[21] Andot03 Bsrc, "Dataset Predic Terkena Penyakit Paru-Paru," kaggle.com. [Online]. Available: https://www.kaggle.com/datasets/andot03bsrc/dataset-predic-terkena-penyakit-paruparu
[22] A. S. Gillis, "Data Splitting," TechTarget. [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/data-splitting

[23] S. Adiningsi, B. Pramono, A. M. Sajiah, and R. A. Saputra, "Identifikasi Kualitas Ikan Cakalang Segar Berbasis Citra Mata Menggunakan Metode Support Vector Machine (SVM) Dengan Fungsi Kernel Radial Basis Function," *JATI(Jurnal Mhs. Tek. Inform.*, vol. 9, no. 1, pp. 1522–1530, 2025, doi: https://doi.org/10.36040/jati.v9i1.12641.

[24] D. I. Imron, R. Astuti, and W. Prihartono, "Prediksi Churn Pelanggan Pada Layanan Desain Grafis Home Desain Menggunakan Algoritma Naïve Bayes," *JITET (Jurnal Inform. dan Tek. Elektro Ter.*, vol. 13, no. 1, pp. 1022–1028, 2025, doi: http://dx.doi.org/10.23960/jitet.v13i1.5811.

[25] D. M. U. Atmaja, A. R. Hakim, D. Haryadi, and N. Suwaryo, "Penerapan Algoritma K-Nearest Neighbor Untuk Prediksi Pengelompokkan Tingkat Risiko Penyebaran Covid-19 Jawa Barat," in *Prosiding Seminar Nasional Teknologi Energi dan Mineral*, 2021, pp. 1218–1226.

[26] A. Desiani, D. A. Zayanti, I. Ramayanti, F. F. Ramadhan, and G. Giovillando, "Perbandingan Algoritma Support Vector Machine (SVM)Dan Logistic Regression dalam Klasifikasi Kanker Payudara," *J. Kecerdasan Buatan dan Teknol. Inf.*, vol. 4, no. 1, pp. 33–42, 2025, doi: https://doi.org/10.69916/jkbti.v4i1.191.

[27] K. S. Nugroho, "Validasi Model Klasifikasi Machine Learning pada RapidMiner : Split Validation (Training Error & Test Error) dan Cross Validation," Medium. [Online]. Available: https://ksnugroho.medium.com/validasi-model-machine-learning-pada-rapidminer-50be0080df14