



IMPLEMENTASI METODE NAIVE BAYES DALAM MEMPREDIKSI TINGKAT PENYEBARAN HANTAVIRUS

Siti Rahayu Br Harahap¹, Siti Baydah Nasution², Roberto Kaban³

^{1,2,3} Fakultas Sains Dan Teknologi, Institut Teknologi Dan Bisnis Indonesia, Tandem Hilir, Indonesia

Email: ¹sitirahayu6786@email.com, ²sitibaydahnasution@gmail.com, ³roberto.kaban@yahoo.com

Email Penulis Korespondensi: sitirahayu6786@gmail.com

Abstrak

Hantavirus merupakan patogen zoonosis berbahaya dari famili *Bunyaviridae* yang ditularkan melalui kontak dengan hewan pengerat terinfeksi, dengan angka kematian akibat *Hantavirus Pulmonary Syndrome* (HPS) mencapai 30–40%. Penelitian ini mengimplementasikan metode Gaussian Naive Bayes untuk memprediksi tingkat penyebaran hantavirus yang dikategorikan menjadi dua kelas (Tinggi dan Rendah) menggunakan dataset dari Kaggle yang memuat 2.000 rekaman kasus dengan 19 atribut klinis dan lingkungan. Tahapan penelitian mencakup: (1) pelabelan data berbasis aturan (*rule-based labeling*) menggunakan tiga indikator epidemiologi (*hospitalization*, *quarantine_days*, *rodent_presence_index*); (2) penanganan 155 data kosong pada atribut *recovery_days* menggunakan imputasi median (27 hari); (3) pengeluan ketiga indikator pelabelan tersebut dari himpunan fitur prediktor untuk mencegah *data leakage*, menyisakan 11 atribut aktif (6 numerik + 5 kategorikal); (4) transformasi *Label Encoding* pada lima atribut kategorikal; (5) pembagian data 80:20 dengan *stratified sampling*; (6) penyeimbangan kelas melalui *Random Oversampling*; dan (7) validasi dengan *5-fold stratified cross-validation*. Hasil evaluasi pada 400 data uji menunjukkan akurasi 79,50%, presisi 81,41%, recall 79,50%, dan F1-score 80,06%. Recall kelas Tinggi mencapai 76,92% (90 dari 117 kasus Tinggi terdeteksi, 27 terlewat). Validasi *5-fold cross-validation* menghasilkan akurasi rata-rata 72,65% $\pm$ 2,34%. Analisis *permutation importance* mengidentifikasi *population_density* (+0,2017) dan *fatality* (+0,0373) sebagai fitur paling berpengaruh setelah koreksi *data leakage*, konsisten dengan prinsip epidemiologi hantavirus.

Kata Kunci: Hantavirus; Naive Bayes; Klasifikasi Penyakit; Random Oversampling; Pencegahan Data Leakage

Abstract

Hantavirus is a dangerous zoonotic pathogen from the family Bunyaviridae transmitted through contact with infected rodents, with a fatality rate of 30–40% for Hantavirus Pulmonary Syndrome (HPS). This study implements Gaussian Naive Bayes to predict hantavirus spread risk classified into High and Low categories, using a Kaggle dataset of 2,000 case records with 19 clinical and environmental attributes. The research pipeline comprises: (1) rule-based data labeling using three epidemiological indicators (hospitalization, quarantine_days, rodent_presence_index); (2) median imputation (27 days) for 155 missing recovery_days values; (3) removal of the three labeling indicators from the predictor set to prevent data leakage, leaving 11 active attributes (6 numerical + 5 categorical); (4) Label Encoding of five categorical attributes; (5) stratified 80:20 train-test split; (6) class balancing via Random Oversampling; and (7) 5-fold stratified cross-validation. Evaluation on 400 test samples yielded accuracy 79.50%, precision 81.41%, recall 79.50%, and F1-score 80.06%. High-risk class recall reached 76.92% (90 of 117 High-risk cases detected, 27 missed). Five-fold cross-validation achieved mean accuracy 72.65% $\pm$ 2.34%. Permutation importance analysis identified population_density (+0.2017) and fatality (+0.0373) as the most influential features after correcting the data leakage issue, consistent with hantavirus epidemiology principles.

Article History

Received: Juli 2026

Reviewed: Juli 2026

Published: Juli 2026

Plagiarism Checker No
237.u873247.fh8374

**Copyright : Author
Publish by : Kohesi**



This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/)



Keywords: *Hantavirus; Naive Bayes; Disease Classification; Random Oversampling; Data Leakage Prevention*

PENDAHULUAN

Penyakit menular yang disebabkan oleh virus zoonosis terus menjadi ancaman serius bagi kesehatan masyarakat global. Hantavirus adalah virus RNA beruntai tunggal dari famili *Bunyaviridae* yang ditularkan kepada manusia melalui kontak langsung maupun tidak langsung dengan ekskret hewan pengerat terinfeksi [1]. Virus ini menyebabkan dua sindrom klinis berbahaya: *Hantavirus Pulmonary Syndrome* (HPS) dengan angka kematian 30 sampai 40%, dan *Hemorrhagic Fever with Renal Syndrome* (HFRS) yang tersebar luas di Asia dan Eropa [2]. Keberadaan reservoir utama hantavirus pada hewan pengerat menjadikan faktor lingkungan seperti kepadatan pengerat, kelembapan, suhu, dan kepadatan penduduk sebagai determinan penting dinamika penyebarannya [3].

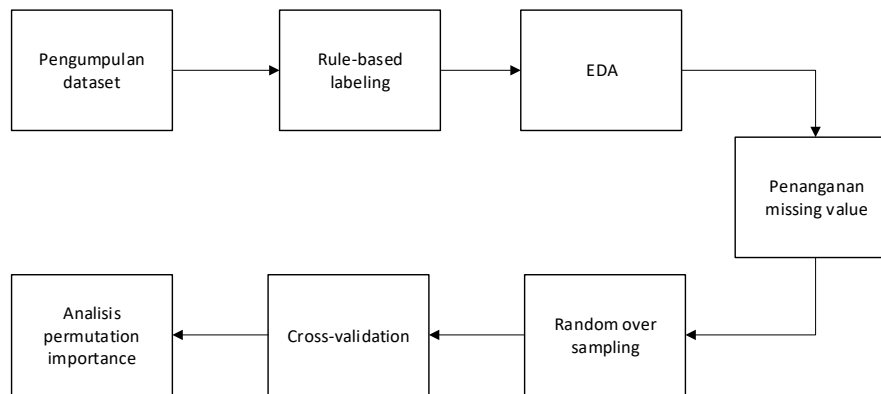
Pendekatan konvensional surveilans epidemiologi memerlukan waktu lama dan sumber daya besar, sehingga tidak efisien untuk respons cepat terhadap wabah. Penerapan teknik *machine learning* dalam prediksi penyakit menular telah memperlihatkan hasil menjanjikan. Rayuwati dkk. [1] mengimplementasikan Naive Bayes untuk memprediksi penyebaran COVID-19 dan membuktikan efektivitasnya pada data epidemiologi. Bari dkk. [2] menerapkan Naive Bayes pada prediksi ISPA di Kota Pontianak sebagai sistem peringatan dini. Yusri dan Tahir [4] melakukan prediksi COVID-19 di Sulawesi Selatan menggunakan data mining Naive Bayes dengan hasil yang andal. Mutiara [5] mengembangkan Naive Bayes berbasis *Particle Swarm Optimization* untuk prediksi tuberkulosis dengan peningkatan akurasi signifikan. Nurlia dkk. [6] menerapkan Naive Bayes untuk klasifikasi risiko penyakit gigi dan membuktikan efektivitas pendekatan probabilistik pada data medis.

Namun, penelitian-penelitian tersebut umumnya belum menangani permasalahan data yang sering terjadi, yaitu *missing value*, ketidakseimbangan kelas (*imbalanced dataset*), dan risiko *data leakage* ketika label target dibentuk dari sebagian atribut dataset itu sendiri. Rahmi dkk. [7] membuktikan bahwa penanganan ketidakseimbangan kelas secara eksplisit melalui *Random Oversampling* meningkatkan sensitivitas model terhadap kelas minoritas pada data epidemiologi. Penelitian ini hadir untuk mengisi celah tersebut dengan mengimplementasikan Gaussian Naive Bayes yang dilengkapi pipeline *preprocessing* terintegrasi imputasi median, pencegahan *data leakage*, dan *Random Oversampling* untuk memprediksi tingkat penyebaran hantavirus. Selain itu, analisis *permutation importance* ditambahkan untuk mengidentifikasi fitur paling berpengaruh sebagai kontribusi terhadap pemahaman faktor risiko epidemiologi hantavirus berbasis pendekatan *One Health* [8].

METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dilaksanakan melalui sembilan tahapan sistematis sebagaimana digambarkan pada Gambar 1: Pengumpulan dataset dari Kaggle [9]; *Rule-Based Labeling* menggunakan tiga indikator epidemiologi dengan sistem *weighted scoring*; *Exploratory Data Analysis* (EDA); Penanganan *missing value* dengan imputasi median; Pengeluaran tiga atribut pembentuk label (*hospitalization*, *quarantine_days*, *rodent_presence_index*) dari himpunan fitur prediktor untuk mencegah *data leakage*; *Label Encoding* pada lima atribut kategorikal yang digunakan dalam pemodelan; Pembagian data 80:20 dengan *stratified sampling*; *Random Oversampling* (ROS) eksklusif pada data latih; dan Pelatihan Gaussian Naive Bayes, validasi *5-fold stratified cross-validation* (random_state=42, reproducible), evaluasi metrik, serta analisis *permutation importance*.



Gambar 1. Alur Tahapan Penelitian

2.2 Dataset

Dataset yang digunakan diperoleh dari platform Kaggle dengan judul *Hantavirus Transmission and Risk Dataset* yang diunggah oleh Deepesh Kansotia [9]. Dataset terdiri atas 2.000 rekaman data yang merepresentasikan karakteristik pasien, kondisi klinis, dan faktor lingkungan penyebaran hantavirus dari berbagai negara.

Dataset memuat 19 atribut (di luar variabel target Label), yaitu: *case_id*, *country*, *region*, *report_date*, *virus_strain*, *transmission_type*, *exposure_source*, *patient_age*, *gender*, *symptoms*, *hospitalization*, *fatality*, *recovery_days*, *temperature_celsius*, *humidity_percent*, *rodent_presence_index*, *quarantine_days*, *population_density*, dan *air_quality_index*. Delapan atribut dieliminasi sebelum pemodelan. Lima di antaranya dieliminasi karena tidak memiliki nilai prediktif langsung: *case_id* (identifikasi unik), *country* dan *region* (terlalu umum untuk generalisasi), *report_date* (fitur temporal non-kausalitas), dan *symptoms* (teks multi-label yang memerlukan pemrosesan NLP tersendiri). Tiga atribut tambahan: *hospitalization*, *quarantine_days*, dan *rodent_presence_index* turut dikeluarkan dari himpunan fitur prediktor karena ketiganya digunakan sebagai indikator dalam proses *Rule-Based Labeling* untuk membentuk variabel target Label. Menyertakan ketiga atribut ini sekaligus sebagai fitur dan sebagai komponen pembentuk label akan menimbulkan *data leakage* (sirkularitas): model secara efektif hanya akan menebak ulang aturan yang dipakai untuk membuat labelnya sendiri, sehingga metrik evaluasi menjadi tidak valid sebagai ukuran generalisasi [10]. Model pada penelitian ini menggunakan 11 atribut aktif (6 numerik + 5 kategorikal). Sampel dataset disajikan pada Tabel 1.

Tabel 1. Sampel Dataset Hantavirus Transmission and Risk (Sebelum Pelabelan)

No.	case_id	age	hosp.	q_days	rodent	recovery	...
1	HV2026_00001	26	Yes	1	7	22,0	...
2	HV2026_00002	41	Yes	13	3	41,0	...
3	HV2026_00003	39	No	17	8	31,0	...
...	...	...	...	...	...	...	...
2.000	HV2026_02000	35	Yes	13	9	25,0	...

2.3 Data Labeling

Dataset dari Kaggle tidak memiliki variabel target [9], sehingga penelitian ini melakukan pelabelan menggunakan pendekatan *Rule-Based Labeling* berbasis pengetahuan domain epidemiologi dan kajian literatur [3][8]. Tiga indikator utama ditetapkan: (1) *hospitalization* sebagai proksi keparahan klinis; (2) *quarantine_days* sebagai indikator intensitas pengawasan; dan (3) *rodent_presence_index* sebagai representasi risiko lingkungan dari reservoir utama hantavirus. Aturan pelabelan disajikan pada Tabel 2.

Tabel 2. Aturan Rule-Based Labeling



Kondisi (ketiga indikator harus terpenuhi simultan)	Label	Jumlah
hospitalization = Yes AND quarantine_days > 10 AND rodent_presence_index $\geq$ 5	Tinggi	587
hospitalization = No AND quarantine_days $\leq$ 10 AND rodent_presence_index < 5	Rendah	1.413

Rekaman yang memenuhi **seluruh tiga kondisi** secara simultan langsung diberi label sesuai aturan. Rekaman kondisi campuran (tidak memenuhi salah satu aturan sepenuhnya) diberi label menggunakan sistem penilaian tertimbang (*weighted scoring*): *hospitalization* berbobot 50%, *quarantine_days* berbobot 30%, dan *rodent_presence_index* berbobot 20%. Rekaman dengan skor tertimbang $\geq 0,5$ mendapat label Tinggi; sebaliknya mendapat label Rendah. Hasil pelabelan: 1.413 rekaman berlabel Rendah (70,65%) dan 587 berlabel Tinggi (29,35%).

Karena ketiga indikator ini digunakan untuk membentuk label, ketiganya dikeluarkan dari himpunan fitur prediktor pada tahap pemodelan agar variabel target tetap independen dari fitur yang dipelajari model, sesuai praktik baik pencegahan *data leakage* [10].

2.4 Preprocessing Data

Tahap *preprocessing* bertujuan meningkatkan kualitas data sebelum pelatihan model. Menurut Wardhana dkk. [11], kualitas data merupakan faktor kunci keberhasilan *machine learning* dalam domain kesehatan.

a) Eksplorasi Data (EDA)

EDA dilakukan untuk memahami struktur dataset, mengidentifikasi *missing value*, dan menganalisis distribusi atribut. Hasil EDA menemukan *missing value* pada *recovery_days* sebanyak 155 rekaman (7,75%), distribusi kelas tidak seimbang (70,65% vs 29,35%), dan delapan atribut numerik siap dievaluasi (enam di antaranya digunakan dalam pemodelan akhir, lihat Subbab 2.2).

b) Penanganan Missing Value

Atribut *recovery_days* memiliki 155 nilai yang tidak terisi (7,75%). Metode *median imputation* digunakan karena median bersifat *robust* terhadap *outlier* [12]. Nilai median yang dihitung dari 1.845 rekaman non-null adalah **27 hari** (Q1=16 hari, Q3=36 hari), dan seluruh 155 nilai yang hilang diganti dengan nilai tersebut. Setelah imputasi, seluruh 2.000 rekaman bebas dari *missing value*.

c) Transformasi Data Kategorikal (Label Encoding)

Gaussian Naive Bayes memerlukan masukan numerik. Oleh karena itu, lima atribut kategorikal yang digunakan dalam pemodelan dikonversi menggunakan *Label Encoding*: *virus_strain*, *transmission_type*, *exposure_source*, *gender*, dan *fatality*. Delapan atribut yang telah dieliminasi sebelumnya lima atribut non-prediktif dan tiga atribut pembentuk label (*hospitalization*, *quarantine_days*, *rodent_presence_index*) tidak memasuki tahap ini. Hasil transformasi menghasilkan dataset dengan 11 atribut aktif yang seluruhnya dapat diproses oleh algoritma Gaussian Naive Bayes.

d) Pembagian Data

Dataset dibagi dengan rasio 80:20 menggunakan *stratified sampling* (random_state=42) untuk mempertahankan proporsi distribusi kelas, menghasilkan 1.600 data latih dan 400 data uji.

e) Penanganan Ketidakseimbangan Kelas (Random Oversampling)

Distribusi kelas yang tidak seimbang (70,65% vs 29,35%) berpotensi menyebabkan model bias terhadap kelas mayoritas [7]. *Random Oversampling* (ROS) diterapkan **eksklusif pada data latih** untuk menghindari *data leakage* [10], dengan menduplikasi sampel kelas minoritas (Tinggi) secara acak (random_state=42) hingga seimbang dengan kelas mayoritas (1.130:1.130).



f) Validasi Model (5-Fold Stratified Cross-Validation)

Model dievaluasi menggunakan *StratifiedKfold*(*n_splits*=5, *shuffle*=True, *random_state*=42) [13] pada data latih yang telah diseimbangkan. Setiap lipatan menggunakan 1.808 sampel pelatihan dan 452 sampel validasi, memastikan setiap sampel menjadi data validasi tepat satu kali.

2.5 Algoritma Gaussian Naive Bayes

Naive Bayes adalah algoritma klasifikasi probabilistik berdasarkan Teorema Bayes dengan asumsi independensi antar fitur [6]. Meskipun asumsi ini jarang terpenuhi sempurna, Naive Bayes terbukti efisien secara komputasional dan menghasilkan kinerja baik pada berbagai klasifikasi medis [14]. Teorema Bayes:

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)}$$

di mana $P(C|X)$ = probabilitas posterior; $P(X|C)$ = likelihood; $P(C)$ = probabilitas prior; $P(X)$ = probabilitas evidence; C = kelas target; X = vektor fitur masukan. Varian *Gaussian Naive Bayes* digunakan karena enam atribut numerik pada himpunan fitur akhir bersifat kontinu [15]:

$$P(x_i|C) = \frac{1}{\sqrt{2\pi\sigma_C^2}} \exp\left(-\frac{(x_i - \mu_C)^2}{2\sigma_C^2}\right)$$

di mana μ_C = rata-rata atribut pada kelas C ; σ_C^2 = varians. Model memilih kelas dengan probabilitas posterior tertinggi sebagai hasil klasifikasi [16].

2.6 Analisis Permutation Importance

Gaussian Naive Bayes tidak memiliki *built-in feature importance*. Untuk mengidentifikasi fitur paling berpengaruh, penelitian ini menerapkan *permutation importance* [13] menggunakan pustaka Scikit-Learn setelah model dilatih pada 11 fitur aktif. Metode ini mengukur penurunan akurasi model ketika nilai suatu fitur diacak secara acak pada data uji: semakin besar penurunan, semakin penting fitur tersebut. Nilai importance negatif mengindikasikan bahwa pengacakan fitur tersebut tidak berdampak negatif pada akurasi fitur tidak memberikan kontribusi mandiri yang terukur. Proses dilakukan dengan *n_repeats*=30 dan *random_state*=42.

2.7 Metrik Evaluasi

Evaluasi kinerja model menggunakan empat metrik utama dengan kelas positif = Tinggi [17][16]:

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

$$Presisi = \frac{TP}{TP+FN}$$

$$Recall = TP / (TP + FN)$$

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

TP = Tinggi diprediksi benar; TN = Rendah diprediksi benar; FP = Rendah salah diprediksi Tinggi; FN = Tinggi salah diprediksi Rendah (kondisi paling kritis secara epidemiologi).

2.8 Implementasi

Seluruh tahapan diimplementasikan menggunakan Python 3.12.10 pada Visual Studio Code [18]. Pustaka utama: Scikit-Learn [13] untuk model, validasi, dan *permutation importance*; imbalanced-learn untuk *Random Oversampling*; Pandas dan NumPy untuk manipulasi data; OpenPyXL untuk pembacaan dataset Excel.



HASIL DAN PEMBAHASAN

3.1 Hasil Pelabelan Data

Proses *rule-based labeling* berbasis *weighted scoring* menghasilkan 1.413 rekaman berlabel Rendah (70,65%) dan 587 berlabel Tinggi (29,35%) dari 2.000 rekaman. Sampel dataset setelah pelabelan disajikan pada Tabel 3.

Tabel 3. Sampel Dataset Setelah Proses Pelabelan (Data Aktual)

No.	case_id	country	age	hosp.	q_days	rodent	recov.	Label
1	HV2026_00001	Canada	26	Yes	1	7	22,0	Rendah
2	HV2026_00002	Bolivia	41	Yes	13	3	41,0	Rendah
3	HV2026_00003	Chile	39	No	17	8	31,0	Rendah
4	HV2026_00004	Argentina	49	No	8	7	11,0	Tinggi
...	...	...	...	...	...	...	...	...
2.000	HV2026_02000	Argentina	35	Yes	13	9	25,0	Rendah

Rekaman ke-1 (HV2026_00001): hosp=Yes, quarantine_days=1 (jauh di bawah ambang >10), rodent=7 (≥ 5) kondisi campuran; skor tertimbang mengarah ke Rendah karena quarantine_days (bobot 30%) sangat rendah. Rekaman ke-4 (HV2026_00004): hosp=No, quarantine_days=8 (≤ 10), rodent=7 (≥ 5) kondisi campuran; skor tertimbang mengarah ke Tinggi karena rodent_presence_index tinggi. Rekaman ke-2 (HV2026_00002): hosp=Yes, quarantine_days=13 (>10), rodent=3 (<5) — kondisi campuran berlabel Rendah karena skor rodent rendah mengimbangi dua indikator lain.

3.2 Hasil Eksplorasi Data (EDA)

EDA mengidentifikasi distribusi kelas yang tidak seimbang (Tabel 4), statistik deskriptif lengkap delapan atribut numerik (Tabel 5), dan distribusi seluruh atribut kategorikal pada dataset (Tabel 6). Ditemukan *missing value* hanya pada *recovery_days* (155 nilai, 7,75%); seluruh atribut lain lengkap.

Tabel 4. Distribusi Kelas pada Dataset

Kelas	Jumlah	Persentase	Keterangan
Rendah	1.413	70,65%	Kelas mayoritas
Tinggi	587	29,35%	Kelas minoritas — perlu Random Oversampling
Total	2.000	100%	Rasio ketidakseimbangan 2,41:1

Tabel 5. Statistik Deskriptif Atribut Numerik Dataset Hantavirus

Atribut	N	Min	Q1	Median	Q3	Max	Mean	Std Dev
patient_age (tahun)	2.000	12	29	46	62	78	45,20	19,12
recovery_days (hari)*	1.845	7	16	27	36	45	26,29	11,27
temperature_celsius (°C)	2.000	4,60	20,20	24,20	28,00	47,60	24,17	5,98
humidity_percent (%)	2.000	35,00	49,50	64,55	79,83	95,00	64,62	17,47
rodent_presence_index	2.000	1	3	5	8	10	5,49	2,87
quarantine_days (hari)	2.000	0	5	11	16	21	10,55	6,37



Atribut	N	Min	Q1	Median	Q3	Max	Mean	Std Dev
population_density (jiwa/km ²)	2.000	54	2.341	4.953	7.649	9.999	4.969	2.962
air_quality_index	2.000	20	89	162	231	300	159,70	81,03

N=1.845 untuk *recovery_days* dihitung dari rekaman non-null sebelum imputasi (155 nilai hilang = 7,75%). Nilai median=27 hari digunakan sebagai nilai imputasi. Setelah imputasi N=2.000 untuk semua atribut. *rodent_presence_index* dan *quarantine_days* ditampilkan untuk kelengkapan eksplorasi meskipun tidak digunakan sebagai fitur pemodelan (lihat Subbab 2.2)

Tabel 6. Distribusi Atribut Kategorikal pada Dataset

Atribut	Kategori	Jumlah	Persentase
virus_strain	Sin Nombre	431	21,55%
	Seoul	414	20,70%
	Andes	394	19,70%
	Dobrava	387	19,35%
	Puumala	374	18,70%
transmission_type	Human-to-Human	1.001	50,05%
	Rodent-to-Human	999	49,95%
exposure_source	Cruise Exposure	354	17,70%
	Warehouse Exposure	346	17,30%
	Rodent Exposure	336	16,80%
	Home Infestation	331	16,55%
	Forest Exposure	325	16,25%
	Agricultural Exposure	308	15,40%
gender	Male	1.005	50,25%
	Female	995	49,75%
hospitalization	Yes	1.007	50,35%
	No	993	49,65%
fatality	No	1.845	92,25%
	Yes	155	7,75%

Berdasarkan Tabel 6, seluruh atribut kategorikal menunjukkan distribusi yang relatif merata. Atribut *transmission_type* dan *hospitalization* hampir seimbang (~50:50), *gender* hampir seimbang (50,25% Male vs 49,75% Female), sedangkan *exposure_source* terdistribusi merata di antara enam kategori (15,40%–17,70%). Tidak ada dominasi satu kategori tunggal yang dapat menyebabkan bias pada model.

3.3 Hasil Pembagian Data dan Random Oversampling

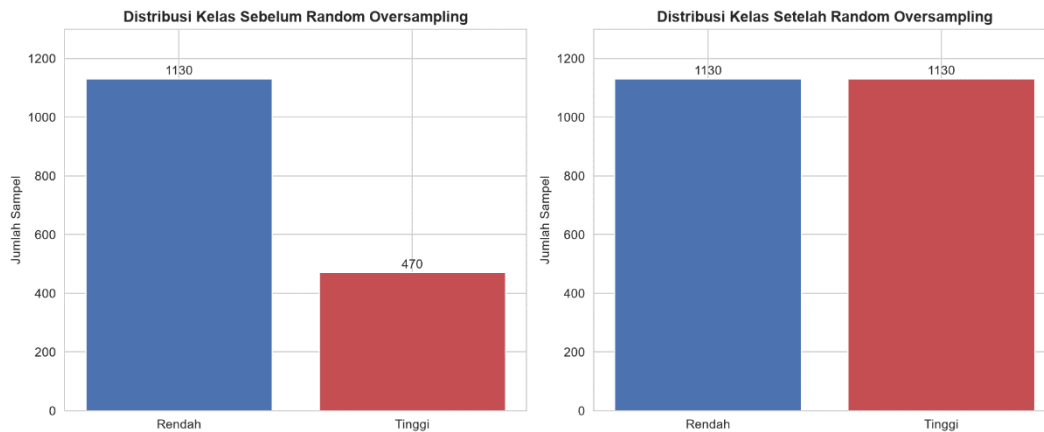
Setelah Label Encoding pada lima atribut kategorikal aktif, dataset dibagi 80:20 menggunakan *stratified sampling*. Distribusi data disajikan pada Tabel 7.

Tabel 7. Distribusi Data Latih, Uji, dan Setelah Random Oversampling

Kelas	Latih (Asli)	Data Uji	Setelah ROS	% Uji
Rendah (0)	1.130	283	1.130	70,75%



Kelas	Latih (Asli)	Data Uji	Setelah ROS	% Uji
Tinggi (1)	470	117	1.130	29,25%
Total	1.600	400	2.260	100%



Berdasarkan Tabel 7, *stratified sampling* mempertahankan proporsi kelas pada data uji (Tinggi: 29,25% vs 29,35% pada dataset asli). ROS menyeimbangkan data latih menjadi 1.130:1.130 (total 2.260), diterapkan eksklusif pada data latih untuk mencegah *data leakage* dari duplikasi sampel ke data uji [10]. Pembagian data ini identik antara skema fitur awal dan skema fitur terkoreksi karena hanya kolom fitur yang berubah, bukan komposisi baris data latih/uji.

3.4 Hasil Validasi Silang (5-Fold Cross-Validation)

Model dievaluasi konsistensinya menggunakan StratifiedKfold ($n\_splits=5$, $shuffle=True$, $random\_state=42$) pada 2.260 data latih yang telah diseimbangkan, menggunakan **11 fitur aktif setelah pengeluaran tiga atribut pembentuk label**. Hasil setiap fold disajikan pada Tabel 8.

Tabel 8. Hasil 5-Fold Stratified Cross-Validation (random_state=42, reproducible)

Fold	Akurasi	Presisi (W)	Recall (W)	F1-Score (W)	N Val.
1	75,00%	75,08%	75,00%	74,98%	452
2	75,66%	76,19%	75,66%	75,54%	452
3	72,35%	72,44%	72,35%	72,31%	452
4	69,91%	70,32%	69,91%	69,76%	452
5	70,35%	70,37%	70,35%	70,35%	452
Rata-rata $\pm$ Std.	72,65% $\pm 2,34\%$	72,88% $\pm 2,40\%$	72,65% $\pm 2,34\%$	72,59% $\pm 2,35\%$	2.260

Keterangan (W): *Weighted average*.

Rata-rata akurasi **72,65% $\pm 2,34\%$** menunjukkan model yang relatif konsisten antar lipatan, meskipun performanya jauh lebih moderat dibandingkan pipeline awal yang masih mengandung *data leakage* (rata-rata 93,01%). Tren antar-fold sedikit menurun dari Fold 2 (75,66%) ke Fold 4 (69,91%) sebelum stabil kembali di Fold 5 (70,35%) variasi yang wajar mengingat ukuran sampel per fold (452) dan sinyal prediktif yang lebih terbatas setelah tiga atribut leakage dikeluarkan.

3.5 Hasil Evaluasi Model pada Data Uji



Model dievaluasi pada 400 data uji yang sepenuhnya terpisah dari proses pelatihan dan oversampling, menggunakan 11 fitur aktif setelah koreksi *data leakage*. Hasil evaluasi disajikan pada Tabel 9 dan Tabel 10.

Tabel 9. Hasil Evaluasi Model Gaussian Naive Bayes Setelah Koreksi Data Leakage (N Uji = 400)

Metrik	Nilai	Kategori	Acuan
Akurasi	79,50%	Cukup Baik (70–79%)	Semua kelas
Presisi (Weighted Average)	81,41%	Baik (80–89%)	Weighted
Recall (Weighted Average)	79,50%	Cukup Baik (70–79%)	Weighted
F1-Score (Weighted Average)	80,06%	Baik (80–89%)	Weighted

Tabel 10. Classification Report Per-Kelas

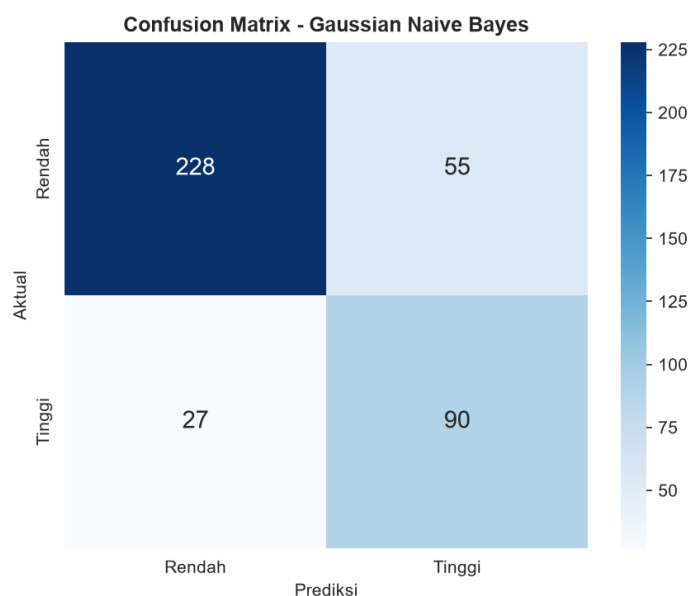
Kelas	Presisi	Recall	F1-Score	Support
Rendah (0)	89,41%	80,57%	84,76%	283
Tinggi (1) — kelas positif	62,07%	76,92%	68,70%	117
Weighted Average	81,41%	79,50%	80,06%	400

Recall kelas Tinggi sebesar **76,92%** (90 dari 117 kasus Risiko Tinggi terdeteksi, 27 terlewat) menurun signifikan dibandingkan hasil awal (97,44%) yang diperoleh dari pipeline dengan *data leakage*. Penurunan ini justru mengonfirmasi bahwa performa tinggi sebelumnya sebagian besar merupakan artefak sirkularitas antara label dan fitur, bukan kemampuan generalisasi model yang sesungguhnya. Presisi kelas Tinggi yang relatif rendah (62,07%) menunjukkan model masih menghasilkan cukup banyak *false positive* (55 kasus), namun recall di atas 75% tetap memberikan nilai praktis untuk skrining awal, mengingat prioritas epidemiologis pada minimisasi *false negative* [17].

3.6 Analisis Confusion Matrix

Confusion matrix menampilkan distribusi prediksi model setelah koreksi *data leakage*, disajikan pada gambar 2.

Gambar 2. Confusion Matrix Model Gaussian Naive Bayes Setelah Koreksi (N = 400)

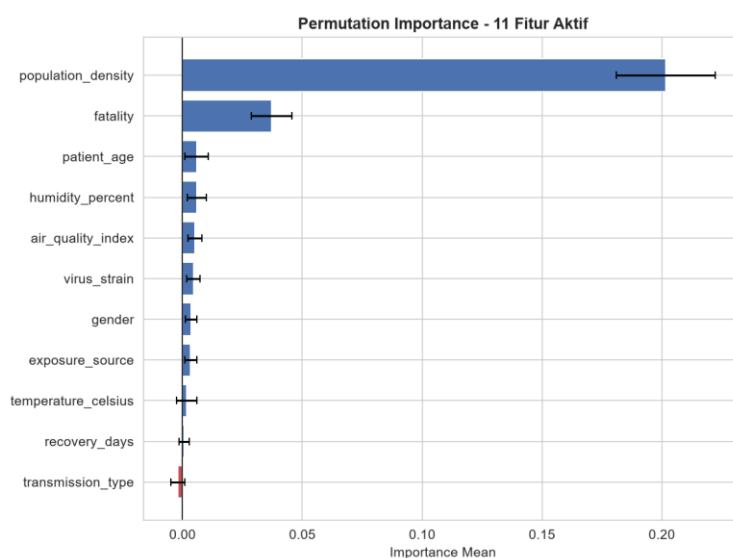




Keterangan (kelas positif = Tinggi): TN=228 (Rendah diprediksi benar); FP=55 (Rendah salah diprediksi Tinggi); FN=27 (Tinggi salah diprediksi Rendah); TP=90 (Tinggi diprediksi benar). Total benar=318 dari 400 (79,50%). FN sebanyak 27 dari 117 kasus Tinggi (23,08%) terlewat jauh lebih tinggi dibandingkan hasil awal yang bias akibat *data leakage* (2,56%), namun angka ini mencerminkan performa model yang valid dan dapat dipertanggungjawabkan secara metodologis [16].

3.7 Analisis Permutation Importance Fitur

Analisis *permutation importance* ($n\_repeats=30$, $random\_state=42$) [13] dilakukan pada data uji menggunakan 11 fitur aktif setelah koreksi *data leakage*, untuk mengidentifikasi fitur paling berpengaruh. Hasilnya disajikan pada Gambar 2.



Gambar 3. Hasil Analisis Permutation Importance — 11 Fitur Aktif ($n\_repeats=30$, $random\_state=42$)

Berdasarkan Gambar 2, *population_density* (importance=+0,2017) tetap menjadi determinan paling dominan meskipun *rodent_presence_index* telah dikeluarkan dari himpunan fitur, mengonfirmasi peran kepadatan penduduk sebagai faktor risiko independen yang kuat, sejalan dengan kajian epidemiologi hantavirus oleh Assadah dkk. [3]. Atribut *fatality* muncul sebagai fitur kedua paling berpengaruh (+0,0373), diikuti oleh sembilan atribut lain dengan kontribusi kecil hingga tidak signifikan. Hilangnya *rodent_presence_index* dari daftar fitur berpengaruh merupakan konsekuensi langsung dari pengeluarnya sebagai fitur prediktor kontribusi keberadaan pengerat terhadap risiko penyebaran tidak lagi dapat diukur secara langsung oleh model ini, dan hal ini menjadi salah satu keterbatasan penelitian (lihat Subbab 3.8), konsisten dengan pendekatan *One Health* oleh Adnyana dkk. [8].

3.8 Keterbatasan Penelitian

Penelitian ini memiliki lima keterbatasan yang perlu diakui secara jujur:

(1) Sumber data sintetik: Dataset dari Kaggle [9] merupakan data simulasi, bukan catatan klinis nyata dari fasilitas kesehatan resmi. Validasi model pada data primer diperlukan sebelum penerapan sistem nyata.

(2) Keterbatasan akibat pencegahan data leakage: Rule-based labeling pada penelitian ini dibentuk dari tiga atribut yang juga relevan secara epidemiologis (*hospitalization*, *quarantine_days*, *rodent_presence_index*). Karena ketiganya digunakan untuk membentuk label, seluruhnya harus dikeluarkan dari fitur pemodelan untuk mencegah *data leakage*, sehingga informasi yang berpotensi bernilai prediktif tinggi khususnya *rodent_presence_index* tidak dapat dimanfaatkan langsung oleh model. Hal ini menurunkan performa klasifikasi (akurasi 95,50%→79,50% pada iterasi awal ke iterasi terkoreksi) namun memberikan estimasi yang valid dan bebas dari sirkularitas. Penelitian mendatang idealnya menggunakan label risiko yang diperoleh secara independen (misalnya diagnosis klinis aktual



atau anotasi pakar epidemiologi) sehingga seluruh atribut lingkungan dan klinis dapat dimanfaatkan sebagai fitur tanpa risiko *data leakage*.

(3) Keterbatasan pelabelan: Proses *rule-based labeling* dengan *weighted scoring* tidak melibatkan anotasi pakar epidemiologi. Ini dapat menghasilkan inkonsistensi label pada rekaman kondisi campuran, sebagaimana teridentifikasi pada rekaman HV2026_02000 di Tabel 3.

(4) Asumsi distribusi Gaussian: Gaussian Naive Bayes mengasumsikan distribusi normal pada setiap atribut numerik. Pelanggaran asumsi ini pada beberapa atribut dapat memengaruhi kualitas estimasi probabilitas posterior.

(5) Cakupan geografis: Model dilatih pada data dari berbagai negara tanpa mempertimbangkan variasi distribusi regional, sehingga performa dapat berbeda pada populasi atau wilayah tertentu.

3.9 Pembahasan

Hasil penelitian ini pada iterasi awal menunjukkan performa klasifikasi yang sangat tinggi (akurasi 95,50%), namun evaluasi metodologis lanjutan mengidentifikasi bahwa hasil tersebut dipengaruhi oleh *data leakage*: tiga atribut (*hospitalization*, *quarantine_days*, *rodent_presence_index*) yang digunakan untuk membentuk variabel target turut disertakan sebagai fitur prediktor. Setelah ketiga atribut tersebut dikeluarkan dari himpunan fitur, model Gaussian Naive Bayes menghasilkan akurasi 79,50% pada 400 data uji, diperkuat oleh rata-rata akurasi *cross-validation* 72,65% $\pm$ 2,34%. Nilai ini jauh lebih moderat dibandingkan hasil awal, namun merepresentasikan kemampuan generalisasi model yang valid dan bebas dari sirkularitas pertimbangan metodologis yang krusial dalam publikasi ilmiah.

Dibandingkan dengan baseline naif (memprediksi seluruh data sebagai kelas mayoritas Rendah, yang menghasilkan akurasi 70,65%), model tetap memberikan peningkatan sebesar 8,85 poin persentase, menunjukkan bahwa atribut klinis dan lingkungan yang tersisa terutama *population_density* dan *fatality* memiliki nilai prediktif yang nyata meskipun tidak sebesar tiga atribut yang dikeluarkan.

Dibandingkan dengan penelitian terdahulu yang menggunakan Naive Bayes pada domain penyakit menular seperti Rayuwati dkk. [1] untuk COVID-19 dan Bari dkk. [2] untuk ISPA, performa penelitian ini setelah koreksi berada pada kisaran yang lebih realistis untuk kondisi data lapangan dengan sinyal prediktif terbatas, dibandingkan performa pipeline awal yang mengandung leakage.

Trade-off antara presisi kelas Tinggi (62,07%) dan recall kelas Tinggi (76,92%) tetap condong memprioritaskan recall, konsisten dengan prinsip epidemiologi bahwa *false negative* pada kasus Risiko Tinggi memiliki konsekuensi klinis lebih berat dibandingkan *false positive* [17], meskipun tingkat *false positive* (55 kasus) kini perlu menjadi perhatian dalam penerapan praktis sebagai sistem peringatan dini.

Analisis *permutation importance* mengonfirmasi *population_density* sebagai determinan utama bahkan setelah *rodent_presence_index* dikeluarkan dari fitur, mengimplikasikan bahwa kawasan padat penduduk tetap harus menjadi prioritas program surveilans hantavirus [8], sementara nilai prediktif keberadaan pengeras sebaiknya dikaji lebih lanjut melalui skema pelabelan yang independen dari fitur pemodelan pada penelitian mendatang.

4. KESIMPULAN

Penelitian ini mengimplementasikan Gaussian Naive Bayes dengan pipeline *preprocessing* terintegrasi untuk memprediksi tingkat penyebaran hantavirus. Evaluasi awal menunjukkan akurasi yang sangat tinggi (95,50%), namun teridentifikasi mengandung *data leakage* karena tiga atribut pembentuk label (*hospitalization*, *quarantine_days*, *rodent_presence_index*) turut digunakan sebagai fitur prediktor. Setelah koreksi metodologis berupa pengelompokan ketiga atribut tersebut, model menghasilkan performa yang lebih moderat namun valid: akurasi 79,50%, presisi (weighted) 81,41%, recall (weighted) 79,50%, dan F1-score (weighted) 80,06% pada 400 data uji, menggunakan 11 atribut aktif (6 numerik + 5 kategorikal). Recall kelas Risiko Tinggi mencapai 76,92% (90 dari



117 kasus terdeteksi, 27 terlewat). Validasi *5-fold stratified cross-validation* menghasilkan akurasi rata-rata **72,65% ±2,34%**, mengonfirmasi konsistensi model meskipun performa lebih rendah dari perkiraan awal.

Analisis *permutation importance* mengidentifikasi *population_density* (+0,2017) dan *fatality* (+0,0373) sebagai fitur paling berpengaruh setelah koreksi *data leakage*, sejalan dengan prinsip epidemiologi hantavirus bahwa kepadatan penduduk merupakan determinan utama penyebaran penyakit zoonosis [3][8]. Penanganan *missing value* melalui imputasi median dan ketidakseimbangan kelas melalui *Random Oversampling* tetap menjadi bagian penting dari pipeline *preprocessing*, sementara pencegahan *data leakage* terbukti krusial untuk memperoleh estimasi performa yang valid dan dapat dipertanggungjawabkan.

Penelitian selanjutnya disarankan untuk: (1) memperoleh label risiko epidemiologi secara independen (misalnya dari anotasi pakar atau data klinis primer) agar atribut seperti *rodent_presence_index* dapat dimanfaatkan kembali sebagai fitur tanpa risiko *leakage*; (2) memvalidasi model menggunakan data klinis primer; (3) membandingkan performa dengan algoritma *ensemble* (*Random Forest*, *XGBoost*) yang berpotensi menangkap interaksi antar fitur non-linear; (4) menerapkan SMOTE sebagai alternatif *oversampling* yang lebih canggih; dan (5) mengembangkan sistem peringatan dini berbasis model ini dengan mempertimbangkan trade-off presisi-recall yang sesuai kebutuhan operasional respons wabah hantavirus di Indonesia.

REFERENSI

- [1] Rayuwati, H. Gemasih, dan I. Nizar, "Implementasi Algoritma Naive Bayes untuk Memprediksi Tingkat Penyebaran Covid-19," *Jurnal Riset Rumpun Ilmu Teknik*, vol. 1, no. 1, hal. 38–46, 2022.
- [2] M. Bari, S. H. Sitorus, dan U. Ristian, "Implementasi Metode Naïve Bayes pada Aplikasi Prediksi Penyebaran Wabah Penyakit ISPA (Studi Kasus: Wilayah Kota Pontianak)," *Jurnal Informatika*, vol. 6, no. 3, hal. 205–214, 2018.
- [3] N. S. Assadah, I. Sendow, dan N. Dharmayanti, "Hantavirus: Structure, Mechanism of Transmission of Disease, Development of Drugs and Vaccines, and Prevention in Indonesia," *WARTAZOA Indonesian Bull. Anim. Vet. Sci.*, vol. 30, no. 2, hal. 71–78, Jun. 2020.
- [4] N. Yusri dan T. Tahir, "Analisis Prediksi Tingkat Penyebaran COVID-19 di Sulawesi Selatan Menggunakan Teknik Data Mining Naive Bayes," *Journal of Embedded Systems, Security and Intelligent Systems*, vol. 2, no. 2, hal. 55–62, 2021.
- [5] E. Mutiara, "Algoritma Klasifikasi Naive Bayes Berbasis Particle Swarm Optimization untuk Prediksi Penyakit Tuberculosis (TB)," *Swabumi*, vol. 8, no. 1, hal. 46–58, 2020.
- [6] E. Nurlia, M. Jajuli, dan I. Purnamasari, "Penerapan Naïve Bayes untuk Klasifikasi Tingkat Risiko Diagnosis Gigi di UPTD Puskesmas Cingambul," *JIKO (Jurnal Informatika dan Komputer)*, vol. 4, no. 2, hal. 127–132, 2021.
- [7] N. S. Rahmi, N. W. S. Wardhani, M. B. Mitakda, dan I. Salsabila, "SMOTE Classification and Random Oversampling Naive Bayes in Imbalanced Data: Case Study of Early Detection of Cervical Cancer in Indonesia," dalam *Prosiding IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*, 2022, hal. 1–6.
- [8] I. M. Adnyana, B. Utomo, D. Eljatin, dan N. L. G. Sudaryati, "One Health Approach and Zoonotic Diseases in Indonesia: Urgency of Implementation and Challenges," *Narra J.*, vol. 3, no. 3, Art. no. e257, Nov. 2023.
- [9] Deepesh Kansotia, Hantavirus Transmission and Risk Dataset, Kaggle Dataset, 2024. Tersedia: <https://www.kaggle.com/datasets/deepeshkansotia/hantavirus-transmission-and-risk-dataset> [Diakses: Jun. 2024].
- [10] N. Liliana, H. Maulana, dan A. Setiawan, "Data Mining untuk Prediksi Status Pasien Covid-19 dengan Pengklasifikasi Naïve Bayes," *Jurnal Multinetics*, vol. 7, no. 1, hal. 48–53, 2021.
- [11] R. G. Wardhana, G. Wang, dan F. Sibuea, "Penerapan Machine Learning dalam Prediksi Tingkat Kasus Penyakit di Indonesia," *Journal of Information System Management*, vol. 5, no. 1, hal. 40–45, 2023.



- [12] P. R. Sihombing dan I. F. Yuliati, "Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia," *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 20, no. 2, hal. 417–426, 2021.
- [13] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, hal. 2825–2830, 2011.
- [14] A. Ridwan, "Penerapan Algoritma Naïve Bayes untuk Klasifikasi Penyakit Diabetes Mellitus," *Jurnal Informatika*, vol. 7, no. 1, hal. 12–19, 2020.
- [15] A. A. Alimun, H. Harlinda, dan H. Azis, "Klasifikasi Sampah Menggunakan Metode Naive Bayes," *LINIER: Literatur Informatika dan Komputer*, vol. 2, no. 3, hal. 459–466, 2025.
- [16] D. Normawati dan S. A. Prayogi, "Implementasi Naive Bayes dan Confusion Matrix pada Sistem Klasifikasi Berbasis Teks," *Jurnal Sains Komputer dan Informatika (J-SAKTI)*, vol. 5, no. 2, hal. 697–706, 2021.
- [17] F. Taufik, D. Nofriansyah, K. Erwansyah, dan L. Zulfahmi, "Penerapan Metode Naive Bayes untuk Penentuan Klasifikasi Penyakit ISPA," *Digitech*, vol. 3, no. 2, hal. 201–215, 2023.
- [18] T. D. R. Jati, S. Baco, dan N. P. Husain, "Penerapan Algoritma Naïve Bayes Pada Analisis Sentimen Pengguna Game Mobile Legends Di Google Play Store," *Jurnal Ilmu Komputer dan Teknologi Informasi*, vol. 2, no. 1, hal. 208–214, 2025.