Information System Journal (INFOS)
Vol. 9, No. 2, November 2026, pp. 109-118
E-ISSN : 2655-142X , P-ISSN : 2655-190X, DOI:https://doi.org/10.24076/infosjournal.2026v9i02.2956

KLASIFIKASI KASUS GIZI BALITA PADA DATA AGREGAT DENGAN NAIVE BAYES BERBASIS PARTICLE SWARM OPTIMIZATION

Ridwansyah [1)], **Muhammad Dwi Cahyo** [2)], **Indah Ariyati** [3)], **Nastiti Tri Permatasari** [4)]
[1)2)3)4)] *Informatika, Universitas Bina Sarana Informatika*

email: ridwansyah.rid@bsi.ac.id [1)], mdwicahyo059@gmail.com [2)], indah.ayi@bsi.ac.id [3)], nastiti.ntp@bsi.ac.id [4)]

INFO ARTIKEL

Riwayat Artikel:

Diterima Juli, 2026
Revisi November, 2026
Terbit November, 2026

Penulis Korespondensi:

Ridwansyah
Informatika,
Universitas Bina Sarana Informatika

Email:
ridwansyah.rid@bsi.ac.id

ABSTRAK

Permasalahan gizi balita perlu dipantau melalui pengolahan data yang tepat. Penelitian ini bertujuan mengklasifikasikan keberadaan kasus gizi balita pada data agregat di DKI Jakarta menggunakan *Categorical Naive Bayes* dan seleksi fitur *Particle Swarm Optimization* (PSO). Dataset terdiri atas 2.403 catatan tahun 2025 dengan target berdasarkan jumlah kasus, yaitu tidak ada kasus (0) dan ada kasus (1). Lima atribut kategorikal digunakan sebagai prediktor. Evaluasi dilakukan menggunakan *5-fold Stratified Group Cross-Validation* dengan kelurahan sebagai kelompok. Hasil menunjukkan bahwa *Naive Bayes* tanpa PSO memperoleh akurasi 77,32%, *precision* 40,57%, *recall* 68,77%, *F1-score* 51,03%, *ROC-AUC* 0,8060, dan *PR-AUC* 0,5039. Sementara itu, *Naive Bayes* dengan PSO memperoleh akurasi 81,27%, *precision* 45,82%, *recall* 49,15%, *F1-score* 47,43%, *ROC-AUC* 0,7943, dan *PR-AUC* 0,4865. PSO meningkatkan akurasi, tetapi menurunkan *recall* dan *F1-score*. Dengan demikian, PSO belum meningkatkan seluruh aspek kinerja klasifikasi. Penelitian ini berfokus pada klasifikasi catatan agregat, bukan diagnosis status gizi individu.

Kata Kunci :
Data Agregat; Gizi Balita; Klasifikasi; *Naive Bayes; Particle Swarm Optimization*

ABSTRACT

Toddler nutritional problems require continuous monitoring through appropriate data analysis. This study aims to classify the presence of toddler nutritional problem cases in aggregate data from DKI Jakarta using Categorical Naive Bayes and Particle Swarm Optimization (PSO)-based feature selection. The dataset consists of 2,403 records from 2025, with the target defined as no reported cases (0) or reported cases (1) based on the recorded count. Five categorical attributes are used as predictors. Model evaluation employs 5-fold Stratified Group Cross-Validation, with kelurahan as the grouping variable. The results show that Naive Bayes without PSO achieves 77.32% accuracy, 40.57% precision, 68.77% recall, a 51.03% F1-score, 0.8060 ROC-AUC, and 0.5039 PR-AUC. Meanwhile, Naive Bayes with PSO achieves 81.27% accuracy, 45.82% precision, 49.15% recall, a 47.43% F1-score, 0.7943 ROC-AUC, and 0.4865 PR-AUC. Although PSO improves accuracy, it reduces recall and F1-score. Therefore, PSO does not improve all aspects of classification performance. This study focuses on classifying aggregate records rather than diagnosing individual toddlers' nutritional status.

***Keywords*:**
Aggregate Data; Toddler Nutrition; Classification; Naive Bayes; Particle Swarm Optimization

1. PENDAHULUAN

Gizi merupakan salah satu faktor penting yang mendukung pertumbuhan, perkembangan, dan kesehatan anak, khususnya pada masa balita. Pada periode tersebut, pemenuhan kebutuhan gizi berkaitan dengan perkembangan fisik dan kemampuan kognitif anak. Penilaian status gizi secara individual umumnya dilakukan menggunakan indikator antropometri, seperti berat badan, tinggi badan, dan umur. Informasi tersebut diperlukan untuk mengetahui kondisi pertumbuhan anak dan membantu menentukan tindak lanjut yang sesuai [1].

Permasalahan gizi balita masih menjadi perhatian dalam pembangunan kesehatan masyarakat, termasuk di Provinsi DKI Jakarta. Informasi mengenai permasalahan gizi dapat disajikan dalam bentuk jumlah kasus berdasarkan wilayah administratif, kelompok usia, dan kategori masalah gizi. Data tersebut memberikan gambaran mengenai distribusi catatan kasus pada tingkat wilayah. Meskipun demikian, jumlah kasus tidak dapat langsung disamakan dengan prevalensi karena prevalensi memerlukan informasi jumlah populasi sebagai pembanding [2]. Informasi mengenai stunting di DKI Jakarta yang diberitakan oleh ANTARA dan penanganan kasus di Jakarta Selatan juga menunjukkan bahwa permasalahan gizi masih menjadi perhatian pemerintah daerah [3]. Namun, stunting memiliki pengertian dan indikator yang berbeda dari kategori gizi buruk, gizi kurang, dan obesitas yang digunakan dalam dataset penelitian ini.

Pemanfaatan data mining dapat membantu mengolah data kesehatan dalam jumlah besar untuk menemukan pola dan menghasilkan klasifikasi berdasarkan atribut yang tersedia. Beberapa penelitian telah menggunakan *Naive Bayes* untuk klasifikasi dan pemantauan status gizi balita [4], identifikasi gizi balita [5] , serta klasifikasi stunting dengan mengombinasikan *Naive Bayes* dan *K-Nearest Neighbor* [6]. Penelitian lainnya menerapkan *Naive Bayes* untuk klasifikasi stunting maupun status gizi berdasarkan data antropometri [9], [10], [11], [12]. Penelitian tersebut umumnya menggunakan data yang merepresentasikan status atau kondisi individu, sedangkan penelitian ini menggunakan data agregat yang disusun berdasarkan wilayah, kelompok usia, dan kategori masalah gizi.

Naive Bayes merupakan metode klasifikasi probabilistik yang menggunakan *Teorema Bayes* dengan asumsi independensi bersyarat antarfitur. Metode ini relatif sederhana dan dapat digunakan untuk membangun model klasifikasi dengan kebutuhan komputasi yang tidak terlalu besar. Akan tetapi, pemilihan varian *Naive Bayes* perlu memperhatikan karakteristik data. Fitur kategorikal yang telah dikodekan menjadi angka tidak otomatis berubah menjadi data kontinu. Oleh karena itu, penggunaan *Gaussian Naive Bayes* pada fitur nominal berpotensi tidak sesuai dengan karakteristik atribut yang digunakan. *Categorical Naive Bayes* dapat digunakan ketika prediktor berupa kategori yang direpresentasikan dengan kode bilangan bulat.

Salah satu pendekatan yang dapat digunakan untuk memilih kombinasi atribut adalah *Particle Swarm Optimization* (PSO). PSO bekerja dengan memperbarui posisi partikel berdasarkan pengalaman terbaik individu dan pengalaman terbaik kelompok [7]. Dalam konteks seleksi fitur, setiap partikel merepresentasikan kombinasi fitur yang digunakan dalam proses klasifikasi. Penelitian sebelumnya telah menerapkan PSO untuk optimasi *Naive Bayes* pada klasifikasi status stunting [8], sedangkan penelitian lain membandingkan penerapan PSO pada beberapa algoritma klasifikasi [13]. Penggunaan PSO pada penelitian ini diarahkan untuk mencari kombinasi atribut yang memberikan nilai evaluasi terbaik berdasarkan PR-AUC pada data latih.

Karakteristik dataset yang digunakan memerlukan perhatian khusus dalam penentuan target. Dataset terdiri atas catatan jumlah permasalahan gizi berdasarkan wilayah, kelurahan, kelompok usia, dan kategori masalah gizi. Setiap kombinasi wilayah, kelurahan, kelompok usia, dan kategori memiliki nilai jumlah yang menunjukkan banyaknya kasus yang tercatat. Dengan demikian, tujuan penelitian tidak diarahkan untuk menentukan apakah seorang balita mengalami gizi buruk, gizi kurang, atau obesitas. Penelitian ini mengklasifikasikan apakah suatu catatan agregat menunjukkan keberadaan kasus atau tidak berdasarkan nilai jumlah yang tercatat.

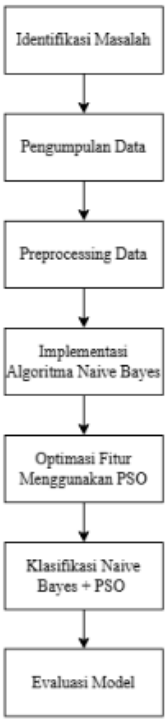
Berdasarkan pertimbangan tersebut, penelitian ini bertujuan membangun dan membandingkan model *Categorical Naive Bayes* tanpa seleksi fitur dan *Categorical Naive Bayes* dengan seleksi fitur berbasis PSO untuk mengklasifikasikan keberadaan kasus pada data agregat permasalahan gizi balita di DKI Jakarta. Kontribusi penelitian terletak pada penerapan klasifikasi biner terhadap catatan agregat, penggunaan validasi berbasis kelompok kelurahan untuk mengurangi kebocoran data, serta evaluasi pengaruh seleksi fitur PSO terhadap akurasi dan kemampuan model mengenali kelas positif. Hasil penelitian diharapkan dapat

Information System Journal (INFOS)
Vol. 9, No. 2, November 2026, pp. 109-118
E-ISSN : 2655-142X , P-ISSN : 2655-190X, DOI:https://doi.org/10.24076/infosjournal.2026v9i02.2956

memberikan gambaran mengenai pemanfaatan data agregat untuk analisis pola keberadaan kasus pada tingkat administratif, tanpa dimaksudkan sebagai pengganti pemeriksaan atau diagnosis gizi individu.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komputasional pada bidang data mining. Eksperimen dilakukan dengan membandingkan kinerja algoritma *Categorical Naive Bayes* sebagai model dasar dan *Categorical Naive Bayes* yang dikombinasikan dengan *Particle Swarm Optimization* sebagai metode seleksi fitur. Tahapan penelitian meliputi pengumpulan data, pemeriksaan karakteristik dataset, pembentukan variabel target, *preprocessing*, pemodelan, optimasi fitur, dan evaluasi model, sebagaimana ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.1 Pengumpulan Data

Dataset yang digunakan merupakan data sekunder mengenai permasalahan gizi balita di Provinsi DKI Jakarta tahun 2025. Data diperoleh dalam bentuk berkas *Excel* yang memuat 2.403 catatan dan tujuh atribut, yaitu *periode_data*, wilayah, kecamatan, kelurahan, usia, *kategori_masalah_gizi*, dan jumlah.

Atribut wilayah, kecamatan, dan kelurahan menunjukkan lokasi administratif pencatatan. Atribut usia menunjukkan kelompok usia yang digunakan dalam penyajian data, sedangkan *kategori_masalah_gizi* terdiri atas “GIZI BURUK”, “GIZI KURANG”, dan “OBESITAS”. Atribut jumlah menunjukkan banyaknya kasus yang tercatat pada kombinasi wilayah, kelompok usia, dan kategori masalah gizi tertentu.

Dataset memiliki 2.403 catatan yang terbentuk dari 801 kombinasi wilayah, kelurahan, dan kelompok usia, dengan tiga kategori masalah gizi pada setiap kombinasi. Dengan demikian, unit analisis dalam penelitian ini adalah catatan agregat, bukan individu balita. Seluruh catatan memiliki nilai pada setiap atribut yang digunakan sehingga tidak ditemukan nilai kosong pada pemeriksaan awal.

Tabel 1. Struktur Atribut Dataset

No.	Atribut	Keterangan	Perlakuan
1	periode_data	Tahun pencatatan	Dikeluarkan karena konstan
2	wilayah	Wilayah administratif	Kandidat prediktor
3	kecamatan	kecamatan	Kandidat prediktor

No.	Atribut	Keterangan	Perlakuan
4	kelurahan	kelurahan	Kandidat prediktor
5	usia	Kelompok usia balita	Kandidat prediktor
6	kategori_masalah_gizi	Kategori masalah gizi	Kandidat prediktor
7	jumlah	Jumlah kasus tercatat	Digunakan membentuk target, bukan prediktor

Variabel target dalam penelitian ini dibentuk dari atribut jumlah. Pembentukan target dilakukan untuk membedakan catatan agregat yang menunjukkan adanya kasus dan catatan yang tidak menunjukkan kasus. Apabila jumlah bernilai nol, catatan dimasukkan ke kelas 0 atau Tidak Ada Kasus. Apabila jumlah lebih besar dari nol, catatan dimasukkan ke kelas 1 atau Ada Kasus. Atribut jumlah tidak disertakan dalam variabel prediktor karena nilai tersebut digunakan secara langsung untuk membentuk target. Penggunaan jumlah sebagai fitur akan menyebabkan kebocoran informasi karena model dapat memperoleh informasi target dari atribut yang menjadi dasar pembentukan label. Berdasarkan hasil pemeriksaan, terdapat 1.990 catatan dengan jumlah nol dan 413 catatan dengan jumlah lebih dari nol. Distribusi kelas tersebut menunjukkan bahwa dataset memiliki ketidakseimbangan kelas, dengan kelas Ada Kasus sebagai kelas minoritas.

Tabel 2. Distribusi Kelas Target

Kelas	Jumlah Catatan	Persentase
Tidak Ada Kasus (0)	1.990	82,81%
Ada Kasus (1)	413	17,19%
Total	2.403	100%

Persentase pada Tabel 2 merupakan proporsi catatan agregat dalam dataset, bukan prevalensi masalah gizi pada seluruh populasi balita di DKI Jakarta.

2.2 *Preprocessing* Data

Tahap *preprocessing* dilakukan untuk menyiapkan atribut agar dapat digunakan oleh algoritma klasifikasi. Proses pengolahan data dilakukan menggunakan bahasa pemrograman *Python* dengan bantuan pustaka *Pandas, NumPy*, dan *Scikit-learn*. Tahapan *preprocessing* meliputi pemeriksaan nilai kosong, penghapusan atribut yang tidak memberikan informasi pembeda, pemisahan fitur dan target, serta pengkodean atribut kategorikal. Atribut *periode_data* dikeluarkan karena seluruh catatan memiliki nilai tahun yang sama, yaitu 2025. Atribut jumlah juga dikeluarkan dari fitur prediktor karena telah digunakan dalam pembentukan variabel target.

Lima atribut yang digunakan sebagai kandidat prediktor adalah wilayah, kecamatan, kelurahan, usia, dan *kategori_masalah_gizi*. Atribut tersebut merupakan data kategorikal. Pengkodean dilakukan dengan mengubah setiap kategori menjadi kode bilangan bulat yang berfungsi sebagai identitas kategori, bukan sebagai nilai numerik yang menunjukkan urutan atau jarak antar kategori. Proses pengkodean dilakukan berdasarkan data latih pada setiap pembagian data. Kategori yang tidak ditemukan pada data latih tetapi muncul pada data uji diperlakukan sebagai kategori yang belum dikenal. Langkah ini diterapkan agar informasi dari data uji tidak digunakan dalam pembentukan representasi kategori pada data latih.

2.3 Algoritma *Categorical Naïve Bayes*

Naive Bayes merupakan algoritma klasifikasi probabilistik yang menggunakan *Teorema Bayes* untuk menghitung probabilitas suatu data termasuk ke dalam kelas tertentu. Persamaan *Teorema Bayes* dinyatakan menggunakan (1).

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)}, \tag{1}$$

Keterangan:

$(P(C|X))$ adalah probabilitas posterior kelas (C) berdasarkan data (X).

$(P(X|C))$ adalah probabilitas data (X) jika diketahui kelas (C).

$(P(C))$ adalah probabilitas awal kelas (C).

$(P(X))$ adalah probabilitas data (X).

Penelitian ini menggunakan *Categorical Naive Bayes* karena seluruh fitur prediktor merupakan atribut kategorikal. Berbeda dengan *Gaussian Naive Bayes* yang menggunakan asumsi distribusi *Gaussian* pada fitur

Information System Journal (INFOS)
Vol. 9, No. 2, November 2026, pp. 109-118
E-ISSN : 2655-142X , P-ISSN : 2655-190X, DOI:https://doi.org/10.24076/infosjournal.2026v9i02.2956

numerik kontinu, *Categorical Naive Bayes* menghitung probabilitas kemunculan setiap kategori pada setiap kelas. Probabilitas kategori dihitung dengan pendekatan *Laplace smoothing* sebagaimana persamaan (2).

$$P(X_j = v|C = c) = \frac{N_{c,j,v} + \alpha}{N_c + \alpha K_j}, \quad (2)$$

Keterangan:

$N_{c,j,v}$	adalah jumlah data kelas (c) dengan fitur (j) bernilai kategori (v).
N_c	adalah jumlah data pada kelas (C).
K_j	adalah jumlah kategori yang diperhitungkan pada fitur (j).
(α)	adalah parameter *smoothing*.

Pada eksperimen ini digunakan nilai α=1. Model dilatih menggunakan data latih dan menghasilkan probabilitas kelas untuk setiap catatan pada data uji. Prediksi kelas ditentukan menggunakan ambang probabilitas 0,5.

2.4 Seleksi Fitur Menggunakan *Particle Swarm Optimization* (PSO)

Particle Swarm Optimization merupakan algoritma optimasi yang meniru perilaku kawanan dalam mencari solusi terbaik. Setiap partikel merepresentasikan satu kandidat solusi. Dalam penelitian ini, kandidat solusi berupa kombinasi fitur yang digunakan pada model *Categorical Naive Bayes*. Setiap posisi partikel direpresentasikan dalam bentuk biner. Nilai 1 menunjukkan fitur dipilih, sedangkan nilai 0 menunjukkan fitur tidak digunakan. Perubahan posisi partikel dipengaruhi oleh kecepatan partikel, pengalaman terbaik individu (personal *best*), dan pengalaman terbaik seluruh partikel (global *best*) [15]. Pembaruan kecepatan partikel dinyatakan sebagaimana persamaan (3).

$$v_{id}^{t+1} = wv_{id}^{t} + c_1 r_1 (pbest_{id} - x_{id}^{t}) + c_2 r_2 (gbest_d - x_{id}^{t}) . \quad (3)$$

Pembaruan posisi biner dilakukan menggunakan fungsi *sigmoid*. Nilai posisi selanjutnya ditentukan berdasarkan probabilitas yang dihasilkan fungsi *sigmoid*. Kombinasi fitur yang terbentuk dievaluasi menggunakan model *Categorical Naive Bayes*. Fungsi *fitness* pada penelitian ini menggunakan nilai *PR-AUC* atau *Average Precision* dari validasi silang pada data latih. Pemilihan *PR-AUC* dilakukan karena distribusi kelas target tidak seimbang dan perhatian penelitian diarahkan pada kemampuan model mengenali kelas Ada Kasus. Semakin tinggi nilai *PR-AUC*, semakin baik kemampuan model dalam memberikan peringkat probabilitas pada kelas positif.

Pengaturan PSO menggunakan 20 partikel dan 30 iterasi. Parameter inersia ditetapkan sebesar 0,7, sedangkan koefisien kognitif dan sosial masing-masing sebesar 1,4. Nilai *fitness* setiap kombinasi fitur dihitung menggunakan 3-*fold Stratified Group Cross-Validation* pada data latih. Fitur yang terpilih kemudian digunakan untuk membentuk model *Categorical Naive Bayes* pada tahap evaluasi.

2.5 Evaluasi Model

Evaluasi model menggunakan 5-*fold Stratified Group Cross-Validation* dengan kelurahan sebagai variabel kelompok (*group*). Pemilihan kelurahan sebagai kelompok dilakukan karena beberapa catatan berasal dari kelurahan yang sama. Dengan pembagian berbasis kelompok, seluruh catatan dari satu kelurahan ditempatkan pada *fold* yang sama sehingga tidak tersebar secara bersamaan ke data latih dan data uji. Pendekatan tersebut digunakan untuk menguji kemampuan model dalam mengklasifikasikan catatan dari kelurahan yang tidak disertakan pada proses pelatihan di *fold* terkait. Pengaturan pembagian data menggunakan pengacakan dengan *random state* 42. Pada setiap *fold*, proses seleksi fitur PSO hanya dilakukan menggunakan bagian data latih. Data uji tidak digunakan dalam pencarian kombinasi fitur. Setelah kombinasi fitur diperoleh, model dilatih kembali menggunakan data latih dan dievaluasi pada data uji. Prediksi dari seluruh *fold* kemudian digabungkan untuk menghitung metrik evaluasi secara keseluruhan.

Evaluasi dilakukan menggunakan *confusion matrix, accuracy, precision, recall, F1-score, ROC-AUC*, dan *PR-AUC*. Penggunaan beberapa metrik diperlukan karena akurasi saja belum cukup untuk menggambarkan kinerja model pada dataset dengan distribusi kelas tidak seimbang.

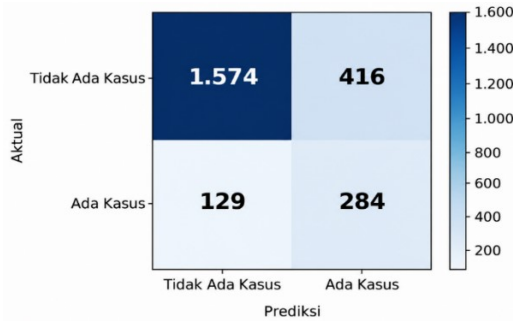
3. HASIL DAN PEMBAHASAN

3.1. Hasil Pemeriksaan Dataset

Dataset penelitian terdiri atas 2.403 catatan dengan tujuh atribut. Pemeriksaan awal menunjukkan tidak terdapat nilai kosong pada atribut yang tersedia. Atribut *periode_data* memiliki satu nilai unik, yaitu 2025, sehingga tidak digunakan dalam pemodelan. Atribut jumlah memiliki nilai yang bervariasi dan digunakan untuk membentuk variabel target. Distribusi target menghasilkan 1.990 catatan Tidak Ada Kasus dan 413 catatan Ada Kasus. Dengan demikian, proporsi kelas positif sebesar 17,19%, sedangkan kelas negatif sebesar 82,81%. Ketidakseimbangan tersebut menjadi pertimbangan dalam pemilihan metrik evaluasi dan fungsi *fitness* PSO. Perlu ditegaskan bahwa distribusi tersebut menggambarkan jumlah catatan agregat, bukan jumlah individu balita yang mengalami atau tidak mengalami masalah gizi. Setiap catatan merepresentasikan kombinasi wilayah, kelompok usia, dan kategori masalah gizi.

3.2. Hasil Klasifikasi *Categorical Naive Bayes*

Pengujian pertama dilakukan menggunakan *Categorical Naive Bayes* tanpa seleksi fitur. Model menggunakan lima atribut prediktor, yaitu wilayah, kecamatan, kelurahan, usia, dan *kategori_masalah_gizi*. Evaluasi dilakukan menggunakan 5-*fold Stratified Group Cross-Validation* dengan kelurahan sebagai kelompok. Hasil prediksi dari seluruh *fold* menghasilkan *confusion matrix* sebagaimana ditunjukkan pada Gambar 2.



Gambar 2. *Confusion Matrix Categorical Naive Bayes*

Berdasarkan Gambar 2, model berhasil mengklasifikasikan 1.574 catatan Tidak Ada Kasus dan 284 catatan Ada Kasus dengan benar. Sebanyak 416 catatan Tidak Ada Kasus salah diprediksi sebagai Ada Kasus, sedangkan 129 catatan Ada Kasus salah diprediksi sebagai Tidak Ada Kasus. Hasil evaluasi model dasar menunjukkan *accuracy* sebesar 77,32%, *precision* 40,57%, *recall* 68,77%, *F1-score* 51,03%, *ROC-AUC* 0,8060, dan *PR-AUC* 0,5039. Nilai *recall* sebesar 68,77% menunjukkan bahwa model dapat mengenali sebagian besar catatan positif, meskipun masih terdapat catatan Ada Kasus yang tidak berhasil terdeteksi.

Tabel 3. Hasil Evaluasi *Categorical Naive Bayes*

Metrik	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*ROC-AUC*	*PR-AUC*
Nilai	77,32%	40,57%	68,77%	51,03%	0,8060	0,5039

3.3. Hasil Seleksi Fitur menggunakan PSO

Seleksi fitur menggunakan PSO dilakukan terhadap lima atribut kandidat prediktor. Setiap partikel merepresentasikan kombinasi fitur yang berbeda. Nilai *fitness* dihitung berdasarkan *PR-AUC* hasil validasi silang pada data latih. Hasil seleksi fitur pada lima *fold* menunjukkan bahwa kecamatan, usia, dan *kategori_masalah_gizi* terpilih pada seluruh *fold*. Fitur kelurahan terpilih pada tiga *fold*, sedangkan wilayah tidak terpilih pada *fold* pengujian.

Tabel 4. Frekuensi Fitur Terpilih oleh PSO

No	Fitur	Frekuensi Terpilih
1	kecamatan	5 dari 5 *fold*
2	usia	5 dari 5 *fold*
3	kategori_masalah_gizi	5 dari 5 *fold*
4	kelurahan	3 dari 5 *fold*
5	wilayah	0 dari 5 *fold*

Information System Journal (INFOS)
Vol. 9, No. 2, November 2026, pp. 109-118
E-ISSN : 2655-142X , P-ISSN : 2655-190X, DOI:https://doi.org/10.24076/infosjournal.2026v9i02.2956

Hasil tersebut menunjukkan bahwa kecamatan, usia, dan *kategori_masalah_gizi* merupakan fitur yang paling konsisten dipilih oleh PSO pada eksperimen ini. Sementara itu, pemilihan kelurahan berubah pada beberapa *fold*. Perbedaan tersebut menunjukkan bahwa kombinasi fitur yang dianggap optimal dapat dipengaruhi oleh komposisi data latih. Fitur yang terpilih tidak dapat langsung ditafsirkan sebagai faktor penyebab terjadinya masalah gizi. Seleksi fitur dalam penelitian ini hanya menunjukkan atribut yang membantu model membedakan catatan Ada Kasus dan Tidak Ada Kasus berdasarkan pola pada dataset.

3.4. Hasil Klasifikasi Categorical Naive Bayes + PSO

Setelah proses seleksi fitur dilakukan, model *Categorical Naive Bayes* dibentuk kembali menggunakan fitur terpilih pada masing-masing *fold*. Hasil prediksi seluruh *fold* menghasilkan *confusion matrix* seperti ditunjukkan pada Gambar 3.



Gambar 3. *Confusion Matrix Categorical Naive Bayes* + PSO

Berdasarkan Gambar 3, model berhasil memprediksi 1.750 catatan Tidak Ada Kasus dan 203 catatan Ada Kasus secara benar. Sebanyak 240 catatan Tidak Ada Kasus salah diklasifikasikan sebagai Ada Kasus, sedangkan 210 catatan Ada Kasus tidak berhasil dikenali oleh model. Hasil evaluasi menunjukkan *accuracy* sebesar 81,27%, *precision* 45,82%, *recall* 49,15%, *F1-score* 47,43%, *ROC-AUC* 0,7943, dan *PR-AUC* 0,4865.

Tabel 5. Hasil Klasifikasi *Naive Bayes* + PSO

Metrik	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*ROC-AUC*	*PR-AUC*
Nilai	81,27%	45,82%	49,15%	47,43%	0,7943	0,4865

3.5. Perbandingan Akurasi *Naive Bayes* dan *Naive Bayes* + PSO

Perbandingan dilakukan untuk mengetahui pengaruh seleksi fitur PSO terhadap kinerja *Categorical Naive Bayes*. Hasil pengujian kedua model ditampilkan pada Tabel 6.

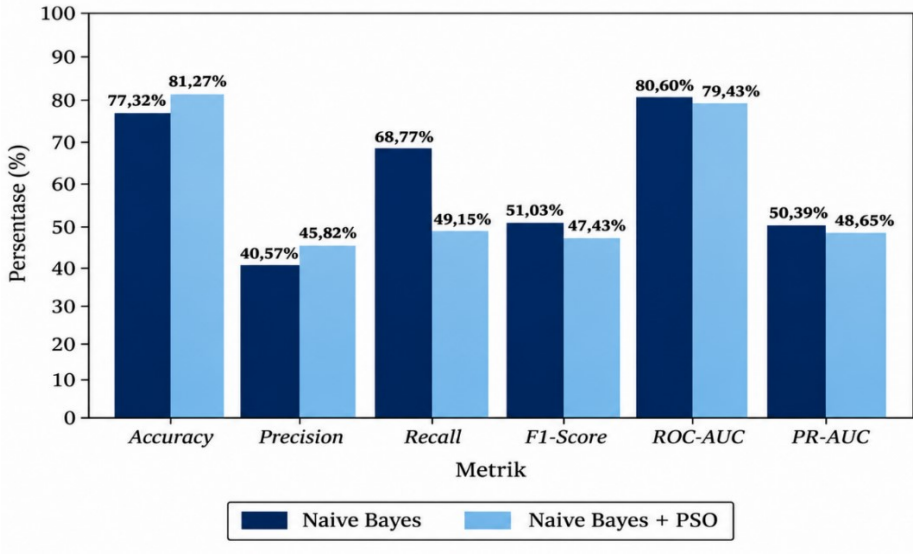
Tabel 6. Perbandingan Kinerja *Categorical Naive Bayes*

Metrik	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*ROC-AUC*	*PR-AUC*
Naive Bayes	77,32%	40,57%	68,77%	51,03%	0,8060	0,5039
***Naive Bayes* + PSO**	81,27%	45,82%	49,15%	47,43%	0,7943	0,4865

Hasil perbandingan menunjukkan bahwa penerapan PSO meningkatkan *accuracy* dari 77,32% menjadi 81,27%, atau sebesar 3,95 poin persentase. *Precision* juga meningkat sebesar 5,25 poin persentase. Peningkatan tersebut menunjukkan bahwa model dengan PSO menghasilkan proporsi prediksi positif yang benar lebih tinggi dibandingkan model dasar. Namun, peningkatan *accuracy* dan *precision* tersebut diikuti penurunan *recall* dari 68,77% menjadi 49,15%. *F1-score* juga menurun dari 51,03% menjadi 47,43%. Selain itu, *ROC-AUC* turun dari 0,8060 menjadi 0,7943, sedangkan PR-AUC turun dari 0,5039 menjadi 0,4865. Penurunan *recall* menunjukkan bahwa model dengan PSO lebih banyak melewatkan catatan yang sebenarnya termasuk kelas Ada Kasus. Hal ini terlihat dari jumlah *false negative* yang meningkat dari 129 menjadi 210 catatan. Sebaliknya, jumlah *false positive* menurun dari 416 menjadi 240 catatan. Dengan demikian, model PSO cenderung menghasilkan lebih sedikit prediksi positif yang keliru, tetapi juga kehilangan sebagian kemampuan dalam mengenali catatan positif.

Temuan tersebut memperlihatkan adanya pertukaran kinerja (*trade-off*) antara ketepatan prediksi positif dan kemampuan menemukan seluruh catatan positif. Oleh karena itu, peningkatan akurasi setelah PSO tidak dapat diartikan sebagai peningkatan kinerja model secara menyeluruh. Mengingat tujuan penelitian berkaitan

dengan identifikasi keberadaan kasus, *recall* dan *PR-AUC* tetap perlu diperhatikan dalam menentukan model yang lebih sesuai.



Gambar 4. Visualisasi Perbandingan Kinerja *Categorical Naive Bayes*

3.6. Pembahasan

Hasil eksperimen menunjukkan bahwa *Categorical Naive Bayes* dapat digunakan untuk mengklasifikasikan keberadaan kasus pada data agregat dengan atribut kategorikal. Penggunaan *Categorical Naive Bayes* juga disesuaikan dengan karakteristik fitur yang berupa kategori wilayah, kelompok usia, dan jenis masalah gizi. Pendekatan tersebut lebih sesuai dibandingkan memperlakukan kode kategori sebagai nilai numerik kontinu pada *Gaussian Naive Bayes*. Penerapan validasi berbasis kelompok kelurahan menjadi bagian penting dalam pengujian. Data yang berasal dari kelurahan yang sama tidak ditempatkan secara bersamaan pada data latih dan data uji dalam satu *fold*. Dengan demikian, evaluasi diarahkan untuk mengukur kemampuan model ketika menghadapi catatan dari kelurahan yang tidak digunakan dalam pelatihan pada *fold* terkait. Hasil pengujian ini memberikan gambaran yang lebih relevan dibandingkan pembagian acak biasa ketika data memiliki keterkaitan berdasarkan wilayah.

PSO menghasilkan kombinasi fitur yang tidak sepenuhnya sama pada setiap *fold*. Meskipun demikian, kecamatan, usia, dan *kategori_masalah_gizi* terpilih secara konsisten. Hal tersebut menunjukkan bahwa ketiga atribut tersebut memberikan kontribusi terhadap proses klasifikasi pada berbagai pembagian data. Fitur kelurahan hanya terpilih pada sebagian *fold*, sedangkan wilayah tidak terpilih. Perbedaan ini menunjukkan bahwa hasil seleksi fitur perlu dilihat berdasarkan konsistensinya, bukan hanya satu kombinasi fitur dari satu proses pengujian. Dari sisi kinerja, PSO meningkatkan *accuracy* dan *precision*, tetapi menurunkan *recall*, *F1-score*, *ROC-AUC*, dan *PR-AUC*. Hasil ini memperlihatkan bahwa pengurangan jumlah fitur tidak selalu meningkatkan seluruh aspek performa klasifikasi. Model dengan PSO lebih selektif dalam memberikan prediksi Ada Kasus, tetapi lebih banyak gagal mengenali catatan positif. Karena kelas positif hanya mencakup 17,19% dari keseluruhan catatan, penggunaan *accuracy* sebagai satu-satunya dasar pemilihan model dapat memberikan gambaran yang kurang lengkap.

Penelitian ini memiliki batasan yang perlu diperhatikan. Pertama, data yang digunakan merupakan data agregat tahun 2025 sehingga hasil klasifikasi tidak dapat ditafsirkan sebagai diagnosis status gizi individu. Kedua, dataset tidak menyediakan informasi jumlah seluruh balita pada setiap wilayah dan kelompok usia, sehingga hasil penelitian tidak dapat digunakan untuk menghitung prevalensi. Ketiga, penggunaan data satu tahun belum memungkinkan analisis perubahan atau prediksi tren antarperiode. Keempat, model hanya menggunakan atribut administratif dan kategori masalah gizi yang tersedia dalam dataset. Penelitian lanjutan dapat mengevaluasi pengaturan ambang klasifikasi, metode seleksi fitur lain, dan penggunaan data dari beberapa periode untuk mengetahui kestabilan kinerja model.

4. KESIMPULAN

Penelitian ini menerapkan *Categorical Naive Bayes* dan *Particle Swarm Optimization* untuk mengklasifikasikan keberadaan kasus permasalahan gizi balita pada data agregat di Provinsi DKI Jakarta. Variabel target dibentuk berdasarkan nilai jumlah, yaitu Tidak Ada Kasus untuk catatan dengan jumlah nol dan Ada Kasus untuk catatan dengan jumlah lebih dari nol. Atribut jumlah tidak digunakan sebagai prediktor, sedangkan periode_data dikeluarkan karena memiliki nilai konstan. Hasil pengujian menggunakan 5-*fold Stratified Group Cross-Validation* menunjukkan bahwa model *Categorical Naive Bayes* tanpa seleksi fitur memperoleh *accuracy* 77,32%, *precision* 40,57%, *recall* 68,77%, *F1-score* 51,03%, *ROC-AUC* 0,8060, dan *PR-AUC* 0,5039. Sementara itu, model *Categorical Naive Bayes* dengan seleksi fitur PSO memperoleh *accuracy* 81,27%, *precision* 45,82%, *recall* 49,15%, *F1-score* 47,43%, *ROC-AUC* 0,7943, dan *PR-AUC* 0,4865. Penerapan PSO meningkatkan *accuracy* sebesar 3,95 poin persentase dan *precision* sebesar 5,25 poin persentase. Akan tetapi, *recall*, *F1-score*, *ROC-AUC*, dan *PR-AUC* mengalami penurunan. Dengan demikian, PSO belum menunjukkan peningkatan kinerja secara menyeluruh, khususnya dalam mengenali catatan yang termasuk kelas Ada Kasus. Fitur kecamatan, usia, dan kategori_masalah_gizi menjadi atribut yang paling konsisten dipilih pada lima fold pengujian.

Penelitian ini terbatas pada klasifikasi catatan agregat dan tidak ditujukan untuk mendiagnosis status gizi individu maupun menghitung prevalensi masalah gizi balita. Penelitian selanjutnya dapat menguji optimasi ambang klasifikasi, membandingkan PSO dengan metode seleksi fitur lainnya, serta menggunakan data dari beberapa tahun agar pola keberadaan kasus dapat dianalisis secara lebih luas.

DAFTAR PUSTAKA

[1] Harliana, H., Rusdiyan Yusron, R. D., & Machfud, I. (2022). Klasifikasi dan Monitoring Status Gizi Balita Melalui Penerapan Metode Naïve Bayes Classification Berbasis GIS. Jurnal Ilmiah Intech : Information Technology Journal of UMUS, 4(02), 161–168. https://doi.org/10.46772/intech.v4i02.869

[2] Santosa, L. W. (2025). Dinkes sebut prevalensi stunting di Jakarta turun menjadi 17,2 persen. ANTARA. https://www.antaranews.com/berita/5055921/dinkes-sebut-prevalensi-stunting-di-jakarta-turun-menjadi-172-persen

[3] Fazli, A. Z. (2025). Pemkot Jakarta Selatan Tangani 794 Kasus Stunting pada 2025. METROTVNEWS. https://www.metrotvnews.com/read/k8oCQZ95-pemkot-jakarta-selatan-tangani-794-kasus-stunting-pada-2025

[4] Viransyah, V., & Sugiarto, B. (2023). Sistem Pakar Diagnosis Gizi Buruk Pada Balita Menggunakan Metode Dempster-Shafer Berbasis Website. 3(2), 569–576. https://sipora.polije.ac.id/id/eprint/25289

[5] Mutiah, M., & Widyaningsih, T. W. (2026). Machine Learning untuk Identifikasi Gizi Balita menggunakan Algoritma Naïve Bayes. Sistem Komputer Dan Kecerdasan Buatan, 9(2). https://doi.org/10.47970/siskom-kb.v9i2.993

[6] Meriyana, P., Rizky Pratama, A., Nurlaelasari, E., & Ratna Juwita, A. (2025). Penerapan Algoritma KNN dan Naive Bayes untuk Klasifikasi Stunting pada Balita di Desa Pasirjengkol. PETISI (Pendidikan Teknologi Informasi), 6(2), 156–168. https://doi.org/10.36232/jurnalpetisi.v6i2.2020

[7] Suprapto, S. (2025). Improvement Naive Bayes Menggunakan Forward Selection, Information Gain dan Gain Ratio untuk Penanganan Independensi Fitur. Jurnal Sosial Teknologi, 5(4), 925–934. https://doi.org/10.59188/jurnalsostech.v5i4.32084

[8] Pahlevi, O., Amrin, A., & Handrianto, Y. (2024). Optimasi Algoritma Naïve Bayes Berbasis Particle Swarm Optimization Untuk Klasifikasi Status Stunting. Computer Science (CO-SCIENCE), 4(1), 37–43. https://doi.org/10.31294/coscience.v4i1.2963

[9] Hardiani, T., & Putri, R. N. (2024). Implementasi Metode Naïve Bayes Classifier Untuk Klasifikasi Stunting Pada Balita. Digital Transformation Technology, 4(1), 621–627. https://doi.org/10.47709/digitech.v4i1.4481

[10] Rahim, H. K. A., & Yoga, T. A. (2024). Implementasi Algoritma Gaussian Naïve Bayes Dalam Klasifikasi Status Gizi Pada Balita. Building of Informatics, Technology and Science (BITS), 6(2), 627–635. https://doi.org/10.47065/bits.v6i2.5493

[11] Muchtar, M. F., Laila, R., Shinta, D., & Pusadan, Y. (2024). Perbandingan Algoritma Naïve Bayes dan Support Vector Machine untuk Klasifikasi Status Stunting pada Balita. Indonesian Journal of Computer Science, 13(1), 3056–3068. https://doi.org/10.33022/ijcs.v13i4.4055

[12] Kurniawan, W., Muriyatmoko, D., & Taufiq, S. A. (2025). Klasifikasi Status Gizi Balita Menggunakan Algoritma Naive Bayes Berdasarkan Pengukuran Antropometri Classification of Toddler Nutrition Status Using Naive Bayes Algorithm Based on Anthropometric Measurements. Jurnal IPTEK, 9(1), 1–5. https://doi.org/10.31543/jii.v9i1.379

[13] Dopo, F. B. D. P., Mau, S. D. B., Tedy, F., & Tengangatu, P. F. M. (2026). IMPLEMENTASI ALGORITMA NAIVE BAYES PADA APLIKASI WEB. JIP(Jurnal Informatika Polinema, 12, 333–340. https://doi.org/10.33795/jip.v12i2.8898

[14] Nurliani, N., Ulfi, M., Saudah, S., Ayu, T. S. P., & Adyanatalubis, A. (2024). Implementasi Deep Learning dengan Convolutional Neural Network untuk Tingkat Akurasi Citra Image Hama Sawi Hijau Menggunakan Google Colab. Riau Journal of Computer Science, 10(2), 116–125. https://doi.org/10.30606/rjocs.v10i2.2854

[15] Saputra, N. R., & Aziz, R. A. (2024). Komparasi Penerapan Particle Swarm Optimization (PSO) Pada Algoritma Naïve Bayes, K-Nearest Neighbor Dan Random Forest Untuk Prediksi Penyakit Hati. Explore: Jurnal Sistem Informasi Dan Telematika, 15(2), 179. https://doi.org/10.36448/jsit.v15i2.3851