

HYBRID FEATURE SELECTION DAN ENSEMBLE LEARNING UNTUK KLASIFIKASI RISIKO STUNTING ANAK DI INDONESIA

¹Ommi Alfina[✉], ¹Nita Syahputri, ²M. Safii, ³Muhammad Taufiq Rustam

¹Sistem Informasi, Universitas Pembinaan Masyarakat Indonesia, Medan, Indonesia

²Informatika, STIKOM Tunas Bangsa, Pematangsiantar, Indonesia

³Informatika, Universitas Potensi Utama, Medan, Indonesia

Email: ny.aeroen@gmail.com

DOI: <https://doi.org/10.46880/jmika.Vol10No1.pp107-111>

ABSTRACT

Stunting is a chronic nutritional problem that remains a major public health issue in Indonesia. This study aims to develop a classification model for stunting risk in children using a combination of hybrid feature selection and ensemble learning methods. The dataset used is derived from socio-economic and health data obtained from the Central Statistics Agency and open datasets. The research method includes data preprocessing, feature selection, model development using Random Forest and Gradient Boosting combined with a Voting Classifier, and evaluation using accuracy, precision, recall, F1-score, and ROC-AUC metrics. The results show that the proposed model achieves high performance with accuracy reaching 98% and ROC-AUC close to 1. The hybrid feature selection successfully improves model efficiency by selecting relevant features. This study demonstrates that the integration of feature selection and ensemble learning can produce an accurate and interpretable model for early detection of stunting risk.

Keyword: *Stunting, Machine Learning, Feature Selection, Ensemble Learning, Classification.*

ABSTRAK

Stunting merupakan permasalahan gizi kronis yang masih menjadi isu utama dalam pembangunan kesehatan di Indonesia. Penelitian ini bertujuan untuk mengembangkan model klasifikasi risiko stunting pada anak dengan mengintegrasikan metode hybrid feature selection dan ensemble learning. Data yang digunakan berasal dari data sosial ekonomi dan kesehatan yang diperoleh dari Badan Pusat Statistik serta dataset terbuka. Metode penelitian meliputi preprocessing data, seleksi fitur, pembangunan model menggunakan Random Forest dan Gradient Boosting yang dikombinasikan dengan Voting Classifier, serta evaluasi menggunakan metrik accuracy, precision, recall, F1-score, dan ROC-AUC. Hasil penelitian menunjukkan bahwa model yang diusulkan memiliki performa tinggi dengan akurasi mencapai 98% dan nilai ROC-AUC mendekati 1. Hybrid feature selection terbukti mampu meningkatkan efisiensi model dengan memilih fitur yang relevan. Penelitian ini menunjukkan bahwa kombinasi feature selection dan ensemble learning mampu menghasilkan model yang akurat dan interpretatif untuk deteksi dini risiko stunting.

Kata Kunci: *Stunting, Machine Learning, Seleksi Fitur, Ensemble Learning, Klasifikasi.*

PENDAHULUAN

Stunting adalah salah satu isu gizi jangka panjang yang terus menjadi tantangan besar dalam pengembangan kesehatan di Indonesia. Keadaan ini ditandai dengan masalah pertumbuhan pada anak yang disebabkan oleh kurangnya nutrisi dalam waktu yang lama. Dampak stunting tidak hanya terbatas pada aspek fisik juga berpengaruh terhadap perkembangan kognitif, produktivitas di masa depan, serta peningkatan risiko penyakit degeneratif. Oleh karena itu, penanganan stunting menjadi prioritas strategis dalam agenda pembangunan nasional. Data yang dirilis oleh Badan Kebijakan Pembangunan Kesehatan (BKPK) menunjukkan bahwa prevalensi stunting di

Indonesia telah mengalami penurunan sebesar 19,8% dan angka ini lebih rendah 0,3% poin dibandingkan dengan sasaran prevalensi stunting yang ditentukan untuk tahun 2024 yaitu 20,1% (Afifah et al., 2025). Distribusi kasus stunting masih menunjukkan ketimpangan antar wilayah terutama antara daerah perkotaan dan perdesaan, serta antar kelompok sosial ekonomi (Al-Maraghi et al., 2025; Ismayanti, 2024). Hal ini menunjukkan bahwa permasalahan stunting bersifat kompleks dan dipengaruhi oleh berbagai faktor multidimensional seperti kondisi ekonomi keluarga, tingkat pendidikan orang tua, akses terhadap layanan kesehatan, sanitasi lingkungan, serta pola asuh anak (Maku, 2026; Rahman et al., 2024).



Pendekatan konvensional dalam identifikasi risiko stunting umumnya masih menggunakan metode statistik deskriptif atau analisis regresi sederhana (Pangestu & Mujiyono 2025). Metode tersebut dapat memberikan gambaran umum namun belum mampu menangkap hubungan non-linear dan interaksi kompleks antar variabel yang mempengaruhi stunting. Perkembangan teknologi dalam bidang data mining dan machine learning memberikan peluang baru dalam mengatasi keterbatasan tersebut. Algoritma machine learning memiliki kemampuan untuk mengolah data dalam jumlah besar dan mengidentifikasi pola tersembunyi yang tidak dapat ditemukan dengan metode konvensional (Ardhitha et al., 2025; Maulani et al., 2024; Waruwu et al., 2024). Penerapan machine learning dalam klasifikasi risiko stunting masih menghadapi beberapa tantangan. Permasalahan utama adalah tingginya jumlah fitur yang digunakan yang sering kali mengandung redundansi dan informasi yang tidak relevan. Hal ini dapat menyebabkan penurunan performa model, meningkatkan kompleksitas komputasi, serta berpotensi menimbulkan overfitting (Arifuddin et al., 2025; Prayoga et al., 2025; Siswanto et al., 2023). Dibutuhkan teknik feature selection yang efektif untuk memilih subset fitur yang paling relevan dan signifikan terhadap target prediksi (Huizen et al., 2025; Salsabilla et al., 2025). Dalam konteks ini pendekatan ensemble learning menjadi solusi karena mampu menggabungkan beberapa model untuk meningkatkan akurasi dan stabilitas prediksi seperti Random Forest, Gradient Boosting, dan stacking ensemble. Penelitian yang mengintegrasikan hybrid feature selection dengan ensemble learning dalam konteks klasifikasi risiko stunting di Indonesia masih relatif terbatas. Sebagian besar penelitian hanya menggunakan salah satu pendekatan, baik feature selection maupun ensemble learning tanpa mengkombinasikan keduanya secara sistematis. Kombinasi kedua pendekatan tersebut berpotensi menghasilkan model yang lebih efisien dan akurat.

Pemanfaatan data kesehatan nasional yang bersumber dari Badan Pusat Statistik maupun dataset terbuka lainnya masih belum dioptimalkan secara maksimal dalam penelitian berbasis machine learning. Tantangan dalam preprocessing data, penanganan missing value, serta ketidakseimbangan kelas/ imbalanced data juga menjadi faktor yang perlu diperhatikan dalam pengembangan model klasifikasi. Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengembangkan model klasifikasi risiko stunting anak di Indonesia dengan mengintegrasikan pendekatan hybrid feature selection dan ensemble learning. Hybrid feature selection

digunakan untuk meningkatkan kualitas fitur yang digunakan dalam model, sedangkan ensemble learning bertujuan untuk meningkatkan performa prediksi melalui kombinasi beberapa algoritma. Dengan demikian diharapkan model yang dihasilkan tidak hanya memiliki tingkat akurasi yang tinggi, juga mampu memberikan interpretasi yang lebih baik terhadap faktor-faktor yang mempengaruhi stunting. Kontribusi dari penelitian ini adalah pengembangan model klasifikasi yang optimal dan berbasis data yang dapat digunakan sebagai alat bantu dalam deteksi dini risiko stunting. Temuan penelitian ini diharapkan dapat memberikan wawasan yang lebih mendalam tentang faktor-faktor utama yang memengaruhi stunting di Indonesia sehingga dapat mendukung perumusan kebijakan yang lebih efektif.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan metode kuantitatif dengan fokus pada optimasi model klasifikasi untuk memprediksi risiko stunting pada anak di Indonesia. Data yang digunakan bersumber dari data sekunder yang mencakup berbagai indikator kesehatan, gizi, dan sosio-ekonomi anak. Proses penelitian yang dilakukan tertuang pada gambar 1 berikut:



Gambar 1. Alur Riset

Pengumpulan Data

Tahapan awal penelitian dimulai dengan proses pengumpulan dan integrasi data dari berbagai sumber untuk menghasilkan dataset yang komprehensif bersumber dari data Badan Pusat Statistik (BPS) dan Sumber data terbuka kaggle sebanyak 10.000 record data.

Preprocessing Data

Preprocessing data bertujuan untuk meningkatkan kualitas data sebelum digunakan dalam proses pemodelan. Tahapan ini meliputi penanganan missing value menggunakan teknik imputasi, transformasi data kategorikal melalui metode encoding, serta normalisasi data menggunakan pendekatan Min-Max Scaling. Untuk menangani isu ketidakseimbangan kelas yang sering muncul pada dataset kesehatan,

diterapkan metode SMOTE guna memperbaiki keberadaan kelas yang kurang terwakili.

Feature Selection

Tahap feature selection untuk memperoleh subset fitur yang paling relevan dan signifikan terhadap variabel target. Hasil dari proses ini adalah dataset dengan fitur optimal yang diharapkan dapat meningkatkan performa model sekaligus mengurangi kompleksitas komputasi.

Model Klasifikasi Ensemble Learning

Tahap berikutnya pembangunan model klasifikasi menggunakan pendekatan ensemble learning. Menggunakan beberapa algoritma dasar Random Forest dan Gradient Boosting. Model-model tersebut kemudian dikombinasikan menggunakan teknik ensemble Voting Classifier untuk meningkatkan akurasi dan stabilitas prediksi. Untuk mencapai hasil terbaik, dilakukan pengaturan hiperparameter melalui metode Grid Search yang digabungkan dengan teknik validasi silang k-fold untuk menjamin kemampuan generalisasi model.

Evaluasi Model

Evaluasi model dilakukan secara menyeluruh dengan memanfaatkan beberapa ukuran kinerja seperti akurasi, presisi, recall, F1-score, dan ROC-AUC. Pemakaian berbagai ukuran ini bertujuan untuk memberikan wawasan yang lebih luas mengenai kinerja model, terutama saat menghadapi data yang tidak seimbang.

Interpretasi Hasil

Untuk memperbaiki pemahaman terhadap model, diterapkan metode Artificial Intelligence yang dapat dijelaskan (XAI) yang dikenal dengan SHAP (SHapley Additive Explanations) guna menentukan peran setiap fitur dalam hasil prediksi. Karena itu, model yang dihasilkan tidak hanya mencapai akurasi yang tinggi, tetapi juga mampu memberikan wawasan mendalam tentang faktor-faktor utama yang berpengaruh pada risiko stunting.

HASIL DAN PEMBAHASAN

Hasil Eksperimen

Pengujian model dilakukan menggunakan dataset stunting yang telah melalui tahapan preprocessing, penanganan data tidak seimbang menggunakan SMOTE serta seleksi fitur menggunakan metode Chi-Square. Dari proses feature selection, diperoleh sejumlah fitur yang paling relevan dalam mempengaruhi status stunting, seperti faktor kondisi

sanitasi, tingkat pendidikan orang tua, pendapatan keluarga, serta pola konsumsi gizi. Model klasifikasi yang digunakan dalam penelitian ini adalah pendekatan *ensemble learning* dengan teknik *Voting Classifier* yang mengkombinasikan algoritma Random Forest dan Gradient Boosting. Hasil pengujian model terhadap data uji menunjukkan performa yang cukup baik dalam mengklasifikasikan risiko stunting pada gambar berikut:

```
Accuracy: 0.980466724286949

Classification Report:

```

	precision	recall	f1-score	support
0	0.99	0.95	0.97	14384
1	0.98	0.99	0.99	14530
2	0.97	0.98	0.97	14300
3	0.98	1.00	0.99	14636
accuracy			0.98	57850
macro avg	0.98	0.98	0.98	57850
weighted avg	0.98	0.98	0.98	57850

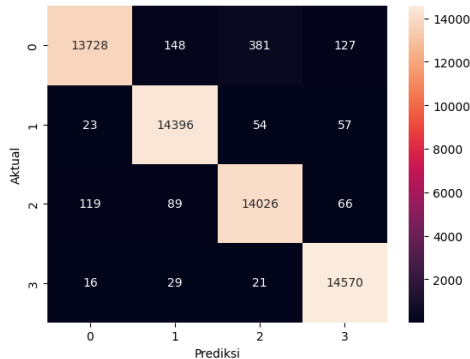
Gambar 2. Classification Report

Berdasarkan hasil laporan klasifikasi yang terlihat pada gambar di atas, model yang dibuat menunjukkan kinerja yang sangat baik dalam mengklasifikasikan data mengenai stunting. Nilai accuracy mencapai sekitar 98% yang mengindikasikan bahwa sebagian besar data berhasil diprediksi dengan benar. Selain itu, Nilai precision, recall, dan f1-score untuk setiap kelas (0, 1, 2, dan 3) berada dalam rentang 0.97 sampai 1.00, menandakan bahwa model berfungsi seimbang dalam mendeteksi tiap kategori tanpa adanya bias yang berarti terhadap kelas tertentu. Kelas 3 bahkan memiliki nilai recall sempurna (1.00) yang berarti seluruh data pada kelas tersebut berhasil diidentifikasi dengan tepat. Nilai macro average dan weighted average yang sama-sama tinggi (0.98) menunjukkan bahwa model memiliki kemampuan generalisasi yang baik meskipun jumlah data di setiap kelas bervariasi. Secara keseluruhan hasil ini membuktikan bahwa pendekatan ensemble learning yang digunakan mampu menghasilkan model klasifikasi yang akurat, stabil, dan efektif dalam mengidentifikasi berbagai kategori status pertumbuhan anak.

Confusion Matrix

Confusion matrix menunjukkan bahwa model memiliki kinerja klasifikasi yang sangat baik, terlihat dari nilai tinggi pada diagonal utama yang menandakan sebagian besar data berhasil diprediksi dengan benar pada masing-masing kelas. Sementara itu, jumlah kesalahan prediksi pada luar diagonal relatif kecil dan tidak signifikan dibandingkan dengan prediksi yang benar. Hal ini menunjukkan bahwa model mampu membedakan antar kelas dengan akurat dan memiliki

tingkat kesalahan yang rendah, sehingga dapat dianggap andal dalam mengklasifikasikan status stunting. Seperti pada hasil berikut:

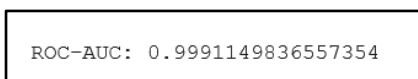


Gambar 3. Confusion Matrix

Gambar diatas menunjukan kinerja model klasifikasi dalam memprediksi empat kelas (0, 1, 2, dan 3). Terlihat bahwa nilai pada diagonal utama (13728, 14396, 14026, dan 14570) sangat tinggi, yang menandakan bahwa sebagian besar data berhasil diklasifikasikan dengan benar sesuai kelas aslinya. Sementara itu nilai di luar diagonal relatif kecil, yang menunjukkan bahwa kesalahan prediksi terjadi dalam jumlah yang minimal. Pada kelas 0 terdapat sebagian kecil data yang salah diklasifikasikan ke kelas 2 sebanyak 381 data, namun jumlah ini tetap jauh lebih kecil dibandingkan prediksi yang benar. Pola serupa juga terlihat pada kelas lainnya, di mana kesalahan tersebar tetapi tidak signifikan. Secara keseluruhan, confusion matrix ini mengindikasikan bahwa model memiliki kemampuan yang sangat baik dalam membedakan antar kelas, dengan tingkat akurasi yang tinggi dan kesalahan prediksi yang rendah, sehingga model dapat dikatakan cukup andal untuk digunakan dalam klasifikasi status stunting.

ROC-AUC

Nilai ROC-AUC yang besar menunjukkan bahwa model mampu dengan efektif membedakan antara kategori positif dan negatif. Semakin mendekati nilai 1, semakin baik performa model dalam mengklasifikasikan data secara tepat pada berbagai threshold. Hal ini mengindikasikan bahwa model tidak hanya akurat tetapi juga memiliki kemampuan diskriminatif yang kuat sehingga dapat diandalkan dalam mendeteksi risiko stunting secara efektif. Seperti pengujian yang sudah dilakukan berikut:

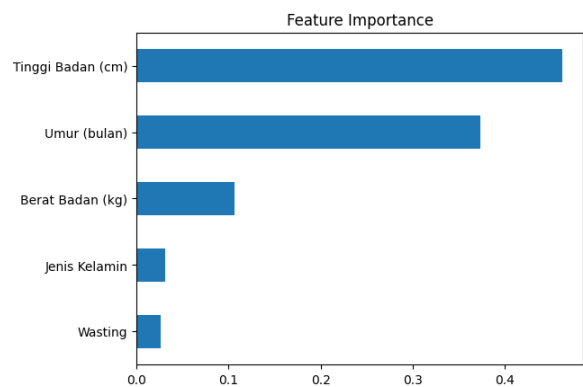


Gambar 4. Nilai ROC-AUC

Nilai ROC-AUC yang mencapai 0.9991 menunjukkan bahwa model ini hampir sepenuhnya berhasil dalam membedakan antara kelas-kelas yang ada. Angka yang sangat mendekati 1 ini menunjukkan bahwa model dapat mengklasifikasi data dengan tingkat keakuratan yang luar biasa tinggi pada berbagai ambang batas, serta memiliki kesalahan klasifikasi yang sangat kecil. Dengan demikian model dapat dikatakan sangat andal dan efektif dalam mendeteksi serta membedakan kondisi stunting secara akurat.

Interpretasi Sederhana (Feature Importance)

Berdasarkan analisis feature importance terlihat bahwa tidak semua variabel memiliki kontribusi yang sama dalam menentukan status stunting. Beberapa fitur utama memberikan pengaruh yang lebih dominan terhadap hasil prediksi, sehingga dapat dianggap sebagai faktor kunci dalam klasifikasi. Hal ini menunjukkan bahwa model mampu mengidentifikasi variabel yang paling relevan, sehingga tidak hanya menghasilkan prediksi yang akurat, tetapi juga memberikan insight yang berguna dalam memahami faktor-faktor yang mempengaruhi stunting yang dapat dilihat pada hasil berikut:



Gambar 5. Feature Importance

Grafik tersebut menunjukkan bahwa Tinggi Badan (cm) merupakan faktor yang paling dominan dan kritis dalam menentukan status stunting, disusul erat oleh variabel Umur (bulan) sebagai dasar pembandingan pertumbuhan yang esensial. Faktor Berat Badan (kg) menempati posisi ketiga dengan pengaruh yang jauh lebih moderat, sementara variabel Jenis Kelamin dan Wasting memiliki kontribusi yang sangat kecil atau minimal terhadap model prediksi ini. Secara keseluruhan, interpretasi ini menegaskan bahwa penilaian stunting pada grafik tersebut sangat bergantung pada korelasi antara dimensi fisik tinggi badan terhadap usia kronologis anak dibandingkan indikator kesehatan lainnya.

KESIMPULAN

Menurut hasil studi, dapat disampaikan bahwa implementasi metode Hybrid Feature Selection yang dikombinasikan dengan Ensemble Learning mampu menghasilkan model klasifikasi yang sangat baik dalam memprediksi risiko stunting anak. Model yang dibangun menunjukkan tingkat akurasi, precision, recall, dan F1-score yang tinggi, serta didukung oleh nilai ROC-AUC yang mendekati sempurna sehingga menunjukkan kemampuan diskriminatif yang sangat kuat. Proses seleksi fitur berhasil mengidentifikasi variabel-variabel penting yang berpengaruh signifikan terhadap status stunting sehingga tidak hanya meningkatkan performa model tetapi juga memberikan pemahaman yang lebih mendalam terkait faktor penyebab stunting. Dengan cara ini, metode yang diterapkan dalam studi ini terbukti efektif dan berpotensi untuk diterapkan sebagai sistem pendukung keputusan dalam upaya deteksi dini dan penanganan stunting di Indonesia.

DAFTAR PUSTAKA

- Afifah, A., Studi, P., Administrasi, I., & Dumai, L. K. (2025). *Analisis implementasi kebijakan percepatan penurunan stunting di indonesia: studi literatur*. 5, 1–14.
- Al-Maraghi, D., Manurung, S., & Mubarak, M. H. (2025). Pengaruh Ketimpangan Pendapatan dan Kemiskinan terhadap Stunting di Pulau Sumatera. *Kajian Ekonomi Dan Akuntansi Terapan*, 2, 251–266.
<https://doi.org/10.61132/keat.v2i4.1919>
- Ardhitha, R., Anugerah, R., & Sutabri, T. (2025). *Analisis Penerapan Machine Learning dan Algoritma Anomali untuk Deteksi Penipuan pada Transaksi Digital*. (1), 80–90.
- Arifuddin, N., Andriyani, W., Dahlan, A., Insani, C., Nur, N., Arifin, N., Situju, S., Setiawan, H., Sutanto, A., Kristanto, T., Wulan, D., & Pamungkas, T. (2025). *Machine Learning*.
- Huizen, L. M., A, M. B., & Idris, M. (2025). *Meningkatkan kinerja SVM: Dampak berbagai teknik seleksi fitur pada akurasi prediksi*. 22(1), 1–14.
- Ismayanti, I. (2024). *Pengembangan Kebijakan: Mendorong Pemerintah Kota Makassar Dalam Meningkatkan Aksesibilitas Layanan Kesehatan Pencegahan Stunting*. 6(2), 162–174.
- Maku, R. M. (2026). *Analisis faktor penyebab stunting pada anak usia dini di puskesmas suri sina*. 9(3), 68–76.
- Maulani, G., Sigitta, R., Tania, K., Purbaya, M., Wahidin, A., Ayuningtyas, A., Santoso, C., Pujiarini, E., Lestari, S., Sari, W., & Rifai, A. (2024). *Penerapan data mining di berbagai bidang*.
- Salsabilla, A. R., Sani, R. R., & Dewi, I. N. (2025). Perbandingan Metode Seleksi Fitur Chi-Square dan Information Gain untuk Peningkatan Interpretabilitas dan Optimasi Kinerja Model TabNet. *Jurnal Nasional Teknologi dan Sistem Informasi*, 11(3), 253-262..
- Pangestu, A., & Mujiyono, S. (2025). Implementasi Algoritma XGBoost Untuk Prediksi Status Gizi Balita Berbasis Website. *Jurnal Algoritma*, 22(2), 176-187.
- Prayoga, A., Hasanuddin, M., Khodijah, S., & Rizki, C. (2025). Analisis Penerapan Machine Learning dalam Sistem Prediksi dan Pengambilan Keputusan. *Journal of Electrical Engineering Research*, 1.
<https://doi.org/10.64803/joecer.v1i3.19>
- Rahman, A., Mas, B., Janur, N. A., & Daimun, B. I. (2024). *Strategi Penurunan Prevalensi Stunting di Kampung KB Desa Palipi Soreang Kabupaten Majene: Studi Analisis Kebijakan dan Intervensi Komunitas Berbasis Hukum Keluarga*. 5(2), 214–224.
<https://doi.org/10.46870/jhki.v5i2.1649>
- Siswanto, S., Dewi, M. U., Kholifah, S., Widhiati, G., & Sains, U. (2023). Penggunaan Model Deep Learning Untuk Meningkatkan Efisiensi Dalam Aplikasi Machine Learning. *Jurnal Penelitian Sistem Informasi (JPSI)*, 1(4), 215-238.
- Waruwu, M. N., Zega, Y., Mendrofa, R. N., & Telaumbanua, Y. N. (2024). Implementasi algoritma machine learning untuk deteksi performa akademik mahasiswa. *TEKNIMEDIA: Teknologi Informasi dan Multimedia*, 5(2), 181-186.