JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 348-358

MODEL RFGS-CS UNTUK MENGATASI *HIGH DIMENSIONAL* DATA STUNTING KOTA SAMARINDA

Lidya Sari[1)], Taghfirul Azhima Yoga Siswa*[2)], Wawan Joko Pranoto[3)]
1. Teknik Informatika, Sains Dan Teknologi, Universitas Muhammadiyah Kalimantan Timur , Indonesia
2. Teknik Informatika, Sains Dan Teknologi, Universitas Muhammadiyah Kalimantan Timur , Indonesia
3. Teknik Informatika, Sains Dan Teknologi, Universitas Muhammadiyah Kalimantan Timur , Indonesia

Article Info
Kata Kunci: *Chi-Square; Grid Search; High Dimensional*; Klasifikasi; *Random Forest*.

Keywords: *Chi-Square; Classification; Grid Search; High Dimensional; Random Forest*.

Article history:
Received 18 October 2024
Revised 12 November 2024
Accepted 13 December 2025
Available online 1 March 2025

DOI :
https://doi.org/10.29100/jipi.v10i1.5997

* Corresponding author.
Corresponding Author
E-mail address:
tay758@umkt.ac.id

ABSTRAK
Di Samarinda, Kalimantan Timur, prevalensi stunting terus meningkat, dengan angka mencapai 23,9% pada tahun 2022. Kondisi ini menunjukkan perlunya intervensi lebih efektif untuk mengatasi masalah gizi di wilayah tersebut. Metode klasifikasi data mining dapat memprediksi risiko stunting, namun penelitian sebelumnya menghadapi tantangan dengan dataset berdimensi tinggi yang dapat mempengaruhi akurasi. Tujuan dari penelitian ini adalah untuk meningkatkan akurasi klasifikasi stunting di Kota Samarinda menggunakan algoritma *Random Forest* (RF) yang dioptimalkan dengan seleksi fitur *Chi-Square* dan optimasi parameter *Grid Search*. Dataset yang digunakan adalah data stunting dari 26 puskesmas di Kota Samarinda tahun 2023 dari Dinas Kesehatan Kota Samarinda. Metode validasi yang digunakan yaitu *cross-validation* dengan k=10. Hasil penelitian menunjukkan bahwa fitur-fitur seperti BB/U, Tinggi, ZS BB/U, ZS TB/U adalah yang paling signifikan dalam mempengaruhi performa model RF. Model RF dengan seleksi fitur *Chi-Square* mencapai akurasi sebesar 99.11%, tidak ada peningkatan akurasi setelah penambahan metode optimasi *Grid Search*. Hasil penelitian ini menunjukkan bahwa model *Random Forest* (RF), baik dengan maupun tanpa optimasi, efektif dalam mengklasifikasikan data stunting. Keefektifan model ini dalam menangani dataset yang rumit dan kompleks, sehingga diharapkan dapat mendukung kebijakan serta intervensi kesehatan.

ABSTRACT
In Samarinda, East Kalimantan, the prevalence of stunting continues to rise, reaching 23.9% in 2022. This condition highlights the need for more effective interventions to address nutritional issues in the region. Data mining classification methods can predict the risk of stunting; however, previous research faced challenges with high-dimensional datasets that could affect accuracy. The aim of this study is to improve the accuracy of stunting classification in Samarinda using the Random Forest (RF) algorithm optimized with Chi-Square feature selection and Grid Search parameter optimization. The dataset used comprises stunting data from 26 health centers in Samarinda in 2023, provided by the Samarinda City Health Office. The validation method used is cross-validation with k=10. The results show that features such as weight-for-age (BB/U), height, weight-for-age Z-score (ZS BB/U), and height-for-age Z-score (ZS TB/U) are the most significant in influencing the performance of the RF model. The RF model with Chi-Square feature selection achieved an accuracy of 99.11%, with no further accuracy improvement after adding the Grid Search optimization method. These findings indicate that the Random Forest (RF) model, both with and without optimization, is effective in classifying stunting data. The effectiveness of this model in handling complex and high-dimensional datasets is expected to support health policies and interventions.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 348-358

I. PENDAHULUAN

Di Indonesia, stunting telah menjadi masalah kesehatan utama yang harus ditangani segera. Indonesia adalah salah satu negara dengan masalah gizi paling parah. Negara ini menghadapi masalah gizi yang sangat pelik dan memerlukan perhatian segera. Stunting adalah masalah gizi penting di Indonesia[1]. Menurut data dari Survei Status Gizi Indonesia (SSGI), persentase anak-anak balita yang menderita stunting di Indonesia pada tahun 2022 adalah sebesar 21,6%[2]. Kalimantan Timur sendiri memiliki tingkat stunting pada anak usia dini di bawah rata-rata nasional, yaitu sebesar 29,2% berbeda dengan rata-rata nasional berjumlah 30,8%. Namun, perlu dicatat bahwa Kalimantan Timur adalah satu-satunya provinsi di mana frekuensi stunting meningkat sejak tahun 2013[3]. Menurut hasil Survei Status Gizi Indonesia (SSGI) yang dirilis oleh Kementerian Kesehatan untuk tahun 2021 dan 2022, tercatat adanya peningkatan prevalensi stunting pada balita di setiap kabupaten/kota di Kalimantan Timur. Angkanya naik dari 22,8% di tahun 2021 menjadi 23,9% di tahun 2022[4]. Bahaya stunting bagi perkembangan anak sangatlah serius, mengancam kesehatan fisik, kognitif, dan sosial mereka. Oleh karena itu, penting untuk memiliki pendekatan yang efektif dalam menangani masalah ini. Dengan menggunakan metode klasifikasi, data mining dapat membantu memprediksi risiko stunting pada anak-anak dan mengidentifikasi kelompok yang rentan terhadap kondisi ini. Informasi ini dapat digunakan untuk merancang program intervensi yang lebih efektif dan tepat sasaran untuk mencegah atau mengurangi kasus stunting.

Machine Learning telah terbukti efektif dalam memprediksi berbagai masalah kesehatan seperti diabetes melitus, stroke dan stunting. Misalnya penelitian oleh [5]–[7] menunjukan bahwa model *machine learning* dapat meningkatkan akurasi prediksi hingga 99,14%, yang pada akhirnya membantu dalam pengambilan keputusan klinis dan intervensi dini. Berdasarkan keberhasilan ini, model *machine learning* diterapkan untuk memprediksi risiko stunting di Samarinda. Tetapi untuk menerapkan metode ini secara optimal, penting untuk memahami kelebihan dan kekurangan dari berbagai algoritma *machine learning*. Contohnya seperti *Naive Bayes* cepat dan efisien, namun asumsinya sering tidak terpenuhi, *Support Vector Machine* kinerja tinggi pada data berdimensi tinggi, namun membutuhkan waktu pelatihan lama, *K-Nearest Neighbor* mudah diimplementasikan tapi lambat pada dataset besar, *Decision Tree* mudah dipahami tetapi rentan *overfitting* dan *Random Forest* dipilih karena lebih baik dalam mengurangi *overfitting*, dan *robust* terhadap *noise*. Selain itu, *Random Forest* dapat menangani data yang tidak seimbang, menyediakan validasi *internal* dengan *out-of-bag error*, memiliki skalabilitas yang baik untuk dataset besar.

Berdasarkan penelitian sebelumnya yang memanfaatkan metode klasifikasi seperti *naïve Bayes* (NB), *Latent Dirichlet Allocation* (LDA), *Support Vector Machine* (SVM), *K-Nearest Neighbor* (kNN), *Decision Tree* (DT), *Random Forest* (RF), dan lain-lain, secara keseluruhan sudah tercapai tingkat akurasi yang signifikan, di mana rata-rata persentase akurasi melebihi 85%[8], [9]. Namun dalam penelitian tersebut perhatian terfokus pada data yang memiliki sedikit fitur. Jumlah fitur yang terlalu rendah juga dapat mempengaruhi hasil klasifikasi. Ketika fitur yang digunakan terbatas, informasi yang tersedia untuk proses klasifikasi berkurang sehingga meningkatkan risiko kesalahan pada hasil akhir klasifikasi[10]. Dalam dataset yang memiliki banyak fitur, jumlahnya bisa sangat besar, termasuk fitur yang penting, ada yang saling terkait, dan ada juga yang tidak berpengaruh pada aplikasi yang digunakan. Banyaknya fitur ini menyebabkan komputer memerlukan waktu lebih lama untuk memprosesnya. Masalahnya, fitur yang saling terkait dan yang tidak berpengaruh bisa merusak hasil analisis data[11]. Sebagai contoh, dalam penelitian oleh [12], [13] menghasilkan tingkat akurasi yang rendah, di mana rata-rata persentase akurasi dibawah 62%. Maka dari itu, untuk mengatasi data dengan banyak fitur, beberapa contoh pendekatan yang sering digunakan yaitu pengelompokan fitur, regularisasi dan seleksi fitur.

Dalam penelitian ini metode klasifikasi yang digunakan yaitu *Random Forest*(RF), dalam kasus klasifikasi dan regresi yang membutuhkan banyak dataset, RF sangat efektif dan sering digunakan[14] merujuk pada penelitian sebelumnya oleh [7], [8] menunjukan bahwa RF lebih unggul dibandingkan dengan algoritma lain seperti *decision tree* (DT), *naïve Bayes* (NB), *k-nearest neighbor* (kNN), *logistic regression*, *Support Vector Machine* (SVM) dan *Neural Networks*. Namun, pada studi yang dikerjakan oleh [12] RF mengalami penurunan akurasi karena penggunaan dataset yang memiliki banyak fitur yang mengandung informasi tidak relevan atau redundan, yang mengakibatkan *overfitting*. Maka dari itu, diperlukan proses seleksi fitur untuk memperbaiki kualitas analisis data. *Chi-Square* adalah salah satu metode yang dapat meningkatkan kualitas analisis data. *Chi-Square* digunakan untuk menilai independensi antara setiap fitur dan kelas target. Ini membantu mengidentifikasi fitur yang paling relevan dengan mengukur kekuatan hubungan antara fitur individu dan variabel target.

Dengan memilih fitur yang signifikan, model dapat fokus pada informasi yang paling penting, mengurangi kompleksitas model, dan meningkatkan kinerja prediksi, seperti penelitian yang dilakukan oleh [15], [16] menunjukan bahwa RF menghasilkan akurasi diatas 80% dengan menggunakan seleksi fitur *Chi-Square* pada dataset yang memiliki banyak fitur. Selain itu, untuk meningkatkan akurasi model lebih lanjut, penelitian ini menerapkan metode

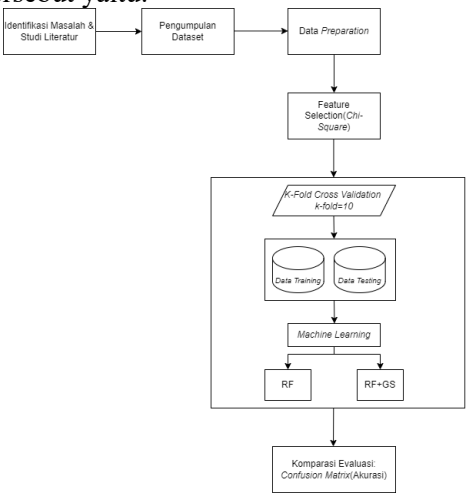
JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 348-358

optimasi *Grid Search*. *Grid Search* adalah teknik pencarian hiperparameter yang sistematis, di mana berbagai kombinasi parameter diuji untuk menemukan konfigurasi yang paling optimal. Dengan *Grid Search*, parameter seperti jumlah pohon dalam hutan (*number of trees*), kedalaman *maximum depth*, dan jumlah fitur yang dipertimbangkan untuk setiap *split* (*max features*) dapat disesuaikan untuk memaksimalkan kinerja model. Berdasarkan penelitian oleh [17], [18] penggabungan RF dengan *Grid Search* bisa meningkatkan akurasi hingga 13.7%.

Penelitian sebelumnya yang menggunakan teknik *machine learning* untuk memprediksi stunting telah menunjukkan hasil yang menjanjikan, namun menghadapi beberapa keterbatasan, seperti kurangnya seleksi fitur yang relevan, yang mengakibatkan *overfitting* dan penurunan akurasi pada dataset besar[12], [13] .Beberapa penelitian juga tidak mengoptimalkan hiperparameter secara sistematis, sehingga model tidak mencapai performa optimal. Untuk mengatasi hal ini, penelitian ini menggunakan seleksi fitur *Chi-Square* untuk mengidentifikasi fitur yang paling relevan, meningkatkan efisiensi model[15], [16] . Selain itu, optimasi parameter dengan *Grid Search* memastikan konfigurasi optimal untuk *Random Forest*, meningkatkan akurasi hingga 13.7%[17], [18]. Dengan pendekatan ini, penelitian ini diharapkan dapat mengatasi keterbatasan penelitian sebelumnya, menghasilkan model yang lebih akurat dan efisien dalam memprediksi risiko stunting. Kombinasi algoritma *Random Forest* dengan seleksi fitur *Chi Square* dan optimasi *Grid Search* diharapkan menghasilkan model klasifikasi yang lebih akurat dan efektif dalam memprediksi risiko stunting, membantu perancangan program intervensi yang tepat sasaran untuk mencegah dan mengurangi kasus stunting.

II. METODE PENELITIAN

Dalam penelitian ini, akan ada serangkaian tahapan pelaksanaan yang akan menjadi panduan untuk menjalankan proses penelitian. Beberapa tahapan tersebut yaitu:



Gambar. 1. Tahapan Penelitian

A. Pengumpulan Data

Langkah awal dalam penelitian ini adalah tahap pengumpulan data. Dataset yang digunakan dalam penelitian ini merupakan data dari Dinas Kesehatan Kota Samarinda . Data yang diperoleh terdiri dari 20 kolom dan memiliki jumlah data sebanyak 150.465.

B. Data Pre-Processing

Selanjutnya, data yang di dapat akan diolah untuk mengoptimalkan kualitas dan struktur dataset sehingga memungkinkan model pembelajaran mesin untuk mempelajari pola-pola yang tepat dari data dan menghasilkan hasil yang akurat. Beberapa tahapan dalam mengolah data yaitu seleksi data, pembersihan data dan transformasi data. Dengan mengaplikasikan tahapan seleksi data, pembersihan data, dan transformasi data, penelitian ini memastikan bahwa dataset yang digunakan memiliki kualitas tinggi dan siap untuk dianalisis oleh model pembelajaran mesin, yang pada akhirnya akan meningkatkan akurasi prediksi risiko stunting. Berikut tahapan pengolahan data yang lebih rinci:

1. Seleksi Data

   Pada tahap ini, data dikumpulkan dengan memilih atribut atau kolom yang relevan untuk dicari, sedangkan atribut yang dianggap tidak relevan dihilangkan. Data awal yang diperoleh dari Dinas Kesehatan Kota Samarinda memiliki 20 kolom. Setelah proses seleksi data dilakukan, 6 kolom dianggap tidak relevan untuk prediksi stunting pada anak, sehingga jumlah kolom berkurang menjadi 14 atribut yang digunakan. Seleksi

data penting karena mengurangi kompleksitas model, meningkatkan kecepatan pemrosesan, dan menghindari *overfitting*.

2. Pembersihan data
   Tahap pembersihan data bertujuan untuk menghilangkan data yang tidak konsisten, data yang mengandung *noise*, dan data yang memiliki nilai kosong (*missing value*). Hal ini dilakukan untuk memastikan bahwa data yang digunakan dalam tahap pemodelan adalah akurat. Pembersihan data penting untuk memastikan kualitas data yang tinggi, yang merupakan dasar dari model prediksi yang akurat.
3. Transformasi data
   Langkah transformasi data dilakukan dengan mengubah nilai-nilai atribut yang bersifat kategorikal menjadi bentuk numerik menggunakan `*LabelEncoder*` dan `*OrdinalEncoder*` dari `*sklearn*`. `*LabelEncoder*` mengubah setiap kategori unik dalam kolom menjadi angka, cocok untuk data tanpa urutan tertentu. `*OrdinalEncoder*`, di sisi lain, digunakan untuk fitur input dengan urutan logis, bekerja pada *array* 2D untuk mengubah beberapa kolom kategori sekaligus. Transformasi ini penting karena memastikan data kompatibel dengan algoritma *machine learning* yang hanya dapat mengolah atribut numerik, sehingga meningkatkan efisiensi dan akurasi model.

C. Pembagian Data

Sebelum dilakukan pemodelan, dataset harus dibagi menjadi dua komponen utama: data latih dan data uji. Data uji digunakan untuk mengevaluasi kinerja model yang telah dibuat, sementara data latih berfungsi sebagai dasar untuk pembuatan model. Penelitian ini menggunakan teknik *K-Fold Cross Validation* yang diimplementasikan dengan *library* `*sklearn.model_selection*` dan fungsi `*cross_val_score*` pada *Python*. *K-Fold Cross Validation* adalah teknik validasi model yang membagi dataset menjadi k bagian atau "*folds*". Dengan k=10, data dibagi menjadi 10 bagian. Dalam setiap iterasi, satu *fold* digunakan sebagai data uji, sementara k-1 *fold* lainnya digunakan sebagai data latih. Proses ini diulang 10 kali sehingga setiap *fold* digunakan sekali sebagai data uji. Rata-rata hasil dari semua iterasi kemudian dihitung untuk memberikan estimasi kinerja model yang lebih stabil dan mengurangi variabilitas hasil yang mungkin terjadi jika hanya satu pembagian data latih dan uji yang digunakan. Metode ini dipilih karena memberikan estimasi kinerja model yang lebih stabil dengan menggunakan beberapa pembagian data yang berbeda, memanfaatkan data secara efisien karena setiap data *point* digunakan baik sebagai data latih maupun data uji, serta mengurangi risiko *overfitting* dengan melakukan validasi pada beberapa *fold*.

D. Permodelan

Pada studi ini, data dibagi menggunakan pendekatan *K-Fold Cross Validation* dengan skema 10-*fold*, di mana dataset dibagi menjadi data *training* dan data *testing*. Bagian data *training* digunakan untuk melatih model, sedangkan data *testing* digunakan untuk mengevaluasi kinerja model yang telah dilatih. *Random Forest* dipilih sebagai model dalam klasifikasi data stunting di Kota Samarinda karena dapat menangani data yang tidak seimbang, menyediakan validasi *internal* dengan *out-of-bag error*, memiliki skalabilitas yang baik untuk dataset besar. Langkah-langkah dalam menerapkannya meliputi:

$$\hat{y} = argmax_c\left(\sum_{i=1}^{N} I(h_i(x) = c)\right) \quad (1)$$

Keterangan:
$\hat{y}$: Prediksi akhir untuk input x.
$argmax_c$: Operator yang mencari nilai c (kelas) yang memaksimalkan jumlah prediksi.
$I(h_i(x) = c)$: Menghasilkan 1 jika prediksi dari pohon h_i untuk input x adalah kelas c, dan menghasilkan 0 jika prediksi dari pohon h_i untuk input x bukan kelas c.
N : Jumlah total pohon dalam *Random Forest*.

1. Model *Random Forest*
   a. Model Random Forest dibuat menggunakan `*RandomForestClassifier*` dari *library scikit-learn*. Algoritma ini menggabungkan banyak pohon keputusan, yang dilatih pada subset acak dari data, untuk meningkatkan akurasi dan stabilitas klasifikasi. Setiap pohon memberikan prediksi, dan hasil akhir ditentukan melalui voting mayoritas, yang membantu mengurangi *overfitting* dan meningkatkan kemampuan model untuk menangani data baru. Parameter *random_state*=42 digunakan untuk memastikan hasil yang konsisten dan dapat direproduksi.
   b. Dataset dibagi menjadi 10 subset menggunakan metode *K-Fold Cross Validation* dengan parameter *n_splits*=10, *shuffle*=True, dan *random_state*=42. Teknik ini mengacak dan membagi data secara merata untuk memastikan evaluasi yang konsisten. Setiap iterasi menggunakan satu subset sebagai data uji,

sementara subset lainnya digunakan sebagai data latih. Dengan k=10, setiap subset menjadi data uji sekali.

c. Dataset dibagi untuk menyimpan akurasi setiap fold dalam *K-Fold Cross Validation*. Pada setiap iterasi, satu subset digunakan sebagai data uji dan sisanya sebagai data latih. Model dilatih dengan data latih dan diuji dengan data uji. Akurasi model untuk data uji dicatat pada setiap iterasi. Proses ini diulang untuk semua *fold*, sehingga setiap subset menjadi data uji sekali. Akurasi dari semua iterasi disimpan dan kemudian dirata-rata untuk memberikan gambaran keseluruhan tentang performa model. Pendekatan ini memastikan evaluasi yang konsisten dan mengurangi bias.

d. *Confusion matrix* dihitung untuk setiap *fold* menggunakan *confusion_matrix* dari *scikit-learn* untuk mengevaluasi performa model pada setiap subset data. *Confusion matrix* memberikan informasi detail tentang jumlah prediksi benar (*true positives* dan *true negatives*) serta kesalahan (*false positives* dan *false negatives*). Rata-rata akurasi dari setiap *fold* dihitung dan ditampilkan, bersama dengan rata-rata *confusion matrix* dari setiap *fold*, yang menunjukkan distribusi prediksi benar dan salah untuk seluruh dataset.

2. Seleksi fitur *Chi-Square*

a. Dataset dikumpulkan dan diorganisir ke dalam tabel kontingensi, yang menunjukkan frekuensi observasi untuk kombinasi kategori dari dua variabel kategorikal.

b. Nilai yang diharapkan untuk setiap sel dalam tabel kontingensi dihitung berdasarkan asumsi bahwa tidak ada hubungan antara kedua variabel, menggunakan rumus

$$E_{ij} = \frac{R_i \times C_j}{N} \qquad (2)$$

Keterangan:
E_{ij} : Nilai yang diharapkan untuk sel pada baris i dan kolom j dalam tabel kontingensi.
R_i : Total frekuensi observasi pada baris i.
C_j : Total frekuensi observasi pada kolom j.
N: Total keseluruhan observasi dalam dataset.

c. Statistik *Chi-Square* dihitung dengan rumus

$$X^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad (3)$$

Keterangan:
X^2 : Nilai statistik *Chi-Square*.
O_{ij} : Nilai observasi sebenarnya untuk sel pada baris i dan kolom j dalam tabel kontingensi.
E_{ij}: Nilai yang diharapkan untuk sel pada baris i dan kolom j.

d. Derajat kebebasan dihitung dengan rumus

$$df = (r - 1) \times (k - 1) \qquad (4)$$

Keterangan:
df: Derajat kebebasan.
r: Jumlah baris dalam tabel kontingensi.
k: Jumlah kolom dalam tabel kontingensi.

3. Optimasi *Grid Search* dengan *Random Forest* melibatkan pembagian dataset menjadi data latih dan uji, pemilihan model *Random Forest*, dan definisi *grid* parameter untuk `*n_estimators*`, `*max_depth*`, dan `*min_samples_split*`. Kombinasi terbaik dipilih berdasarkan akurasi, dan model optimal digunakan untuk prediksi pada data uji untuk memastikan kinerja yang baik dan generalisasi optimal.

E. Komparasi Evaluasi

Dalam tahap ini, dilakukan pengukuran akurasi untuk memvalidasi hasil dari model yang telah dibangun menggunakan algoritma *Random Forest*. Pengukuran ini menggunakan *confusion matrix* untuk mengevaluasi seberapa akurat hasil yang diperoleh. *Confusion matrix* adalah alat evaluasi yang memperlihatkan jumlah prediksi benar (*true positives* dan *true negatives*) serta kesalahan (*false positives* dan *false negatives*) dalam bentuk matriks. Selain *confusion matrix*, metrik evaluasi lainnya seperti *Precision, Recall, dan F1-Score* juga digunakan.

Confusion Matrix memberikan gambaran detail tentang performa model, termasuk jumlah prediksi benar dan salah. *Precision* mengukur proporsi prediksi positif yang benar, *Recall* mengukur proporsi kasus positif yang benar-benar terdeteksi, dan *F1-Score* adalah harmonik rata-rata *Precision* dan *Recall*. Metrik ini dipilih karena mereka memberikan perspektif yang lebih lengkap tentang kinerja model daripada akurasi saja. *Precision* dan *Recall* sangat berguna dalam kasus dengan distribusi kelas yang tidak seimbang, di mana salah satu kelas jauh lebih banyak daripada yang lain. *F1-Score* memberikan keseimbangan antara *Precision* dan *Recall*, terutama ketika ada *trade-off* antara keduanya. Dengan menggunakan *confusion matrix* dan metrik evaluasi tambahan ini dapat mengevaluasi

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 348-358

kinerja model secara lebih komprehensif dan memastikan bahwa model tidak hanya akurat tetapi juga efektif dalam mendeteksi dan mengklasifikasikan kasus yang benar-benar positif serta menghindari kesalahan prediksi yang bisa berpengaruh negatif pada hasil akhir.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Keterangan:
TP (*True Positive*) : merupakan jumlah data yang memiliki label "*yes*" dan berhasil diidentifikasi dengan benar.
TN (*True Negative*) : merupakan jumlah data yang memiliki label "*no*" dan berhasil diidentifikasi dengan benar.
FP (*False Positive*) : merupakan jumlah data yang memiliki label "*yes*" tetapi diidentifikasi sebagai "*no*".
FN (*False Negative*): merupakan jumlah data yang memiliki label "*no*" tetapi diidentifikasi sebagai "*yes*".

III. HASIL DAN PEMBAHASAN

Penelitian ini menginvestigasi klasifikasi stunting menggunakan algoritma *Random Forest* dengan teknik seleksi fitur *Chi Square* dan optimasi *Grid Search*. Metode ini diimplementasikan untuk memprediksi status gizi anak-anak berdasarkan fitur yang relevan. Hasil penelitian menunjukkan bahwa model yang dihasilkan memiliki akurasi yang baik, dengan nilai yang terverifikasi melalui *confusion matrix*. Penggunaan teknik seleksi fitur *Chi Square* membantu mengidentifikasi fitur-fitur yang paling signifikan dalam prediksi stunting, sementara optimasi *Grid Search* memperbaiki kinerja model dengan menemukan parameter terbaik.

A. Seleksi Data

Untuk memastikan analisis yang akurat dan relevan, perlu dilakukan seleksi data guna memilih atribut yang paling berpengaruh dalam prediksi stunting. Proses ini melibatkan identifikasi dan pemisahan atribut yang signifikan dari yang tidak relevan, sehingga hanya data yang relevan digunakan dalam analisis lebih lanjut.

TABEL 1
SELEKSI DATA

	Nama	JK	Berat	Tinggi	LILA	BB/U	ZS BB/U	Tanggal Pengukuran	ZS TB/U	BB/TB	ZS BB/TB	Naik Berat Badan	Jml Vit A	TB/U
1	DIMAS ADITYA	L	9.01		0	Kurang	-0.39	2023-01-02	-0.21	Gizi Baik	-0.39	O		Normal
2	SITI AISYAH	P	12	94	0		-2.25	2023-01-02	-2.09	Gizi Baik	-1.46	O		Pendek
3	M AL FATIH ALMAHIRA	L	8.01	69	0	Berat Badan Normal	-0.53	2023-01-02	-0.65	Gizi Baik	-0.14	O		Normal
4	AKIRA AKBAR GUINIA	P	6.03		0	Berat Badan Normal	-0.31	2023-01-02	0.42	Gizi Baik	-0.74	O		Normal
5	QAMELA	P	10.06		0	Risiko Lebih	0.09097222	2023-01-02	0.23	Gizi Lebih	0.09	O		Normal
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
150464	AFIZAH KHAIRINA	P	2.05	45		Kurang	-2.03	2023-12-09	-2.63	Gizi Baik	-0.35	-		Normal
150465	M ARSYA KHOLIF	P	3	49		Berat Badan Normal	-1.56	2023-12-19	-1.3	Gizi Baik	-1.04	-		Pendek
150466	MUHAMMAD IQBAL	L	2.09	49		Kurang	-2.97	2023-12-29	-2.48	Gizi Baik	-1.37	-		Normal

Tabel 1 adalah hasil dari proses seleksi data dilakukan, 6 kolom dianggap tidak relevan untuk prediksi stunting pada anak, sehingga jumlah kolom berkurang menjadi 14 atribut yang digunakan dengan 1 target atau kelas berada di atribut TB/U.

B. Pembersihan Data

Setelah tahap seleksi data, langkah selanjutnya adalah melakukan pembersihan data. Pembersihan data adalah proses penting untuk memastikan bahwa data yang akan dianalisis bebas dari kesalahan, inkonsistensi, dan nilai yang hilang. Tahap ini melibatkan identifikasi dan koreksi data yang tidak valid, menghapus duplikasi, serta menangani nilai yang hilang atau outliers. Dengan pembersihan data yang tepat, kualitas data dapat ditingkatkan, sehingga analisis dan prediksi yang dilakukan menjadi lebih akurat dan dapat diandalkan.

TABEL 2
PEMBERSIHAN DATA

	Nama	JK	Berat	Tinggi	LILA	BB/U	ZS BB/U	Tanggal Pengukuran	ZS TB/U	BB/TB	ZS BB/TB	Naik Berat Badan	Jml Vit A	TB/U
0	A ALVIN	L	18.06	NaN	18.0	Risiko Lebih	1.190000	2023-01-02	0.41	Risiko Gizi Lebih	0.07	N	NaN	Normal

1	A FADLAN	L	11.06	83.0	0.0	Berat Badan Normal	-0.80000	2023-01-02	-0.97	Gizi Baik	0.04	-	NaN	Normal
2	A FARIS WICAKSONO	L	9.07	78.0	16.0	Berat Badan Normal	-1.750000	2023-01-02	-2.84	Gizi Baik	-0.47	O	1.0	Pendek
3	A FATHAN	L	15.00	107.0	17.0	Berat Badan Normal	-1.080000	2023-01-02	0.14	Gizi Baik	-1.83	T	1.0	Normal
4	A FAUJAN	L	14.00	100.0	0.0	Berat Badan Normal	-0.850000	2023-01-02	-0.16	Gizi Baik	-1.14	O	1.0	Normal
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
34196	Zulaikha mina chandra	P	14.00	104.0	NaN	Berat Badan Normal	-1.430000	2023-12-09	-0.42	Gizi Baik	-1.81	N	1.0	Normal
34197	Zulkifli abdi	L	15.06	102.0	0.0	Berat Badan Normal	-0.590000	2023-12-19	-0.69	Gizi Baik	-0.25	N	1.0	Normal
34198	NaN	L	8.0	71.0	NaN	Berat Badan Normal	-1.570000	2023-12-29	-1.40	Gizi Baik	-1.18	T	NaN	Normal

Tabel 2 menampilkan data setelah melalui proses pembersihan. Setelah mengidentifikasi dan mengoreksi data yang tidak valid, menghapus duplikasi, serta menangani nilai yang hilang dan outliers, dataset yang dihasilkan menjadi lebih bersih dan siap untuk dianalisis. Proses pembersihan ini memastikan bahwa hanya data yang berkualitas tinggi yang digunakan dalam analisis, sehingga hasil prediksi stunting pada anak di Kota Samarinda menjadi lebih akurat.

C. *Transformasi Data*

Setelah melalui proses seleksi dan pembersihan data, tahap berikutnya adalah transformasi data. Langkah transformasi data dilakukan dengan mengubah nilai-nilai atribut yang bersifat kategorikal menjadi bentuk numerik. Beberapa contoh data yang ditransformasi meliputi jenis kelamin, pertambahan berat badan, berat badan menurut usia, dan berat badan menurut tinggi badan. Selain itu, kelas untuk hasil stunting dan tidak stunting juga ditambahkan. Transformasi data ini memastikan bahwa semua fitur dalam dataset berada dalam format yang sesuai dan siap untuk dianalisis.

TABEL 3
DATA SEBELUM PROSES TRANSFORMASI

No	JK	BB/U	BB/TB	Naik Berat Badan
0	L	Berat Badan Normal	Gizi Baik	O
1	L	Berat Badan Normal	Gizi Baik	T
2	L	Berat Badan Normal	Gizi Baik	O
...	...	...	...	...
8056	L	Berat Badan Normal	Gizi Baik	T
8057	L	Berat Badan Normal	Gizi Baik	O
8058	L	Berat Badan Normal	Gizi Baik	N

Tabel 3 menunjukkan data sebelum proses transformasi. Data ini masih dalam bentuk aslinya setelah melalui tahap pembersihan, namun belum diubah ke dalam format numerik. Atribut masih dalam bentuk kategorikal, sehingga belum dapat diolah oleh algoritma *machine learning* yang akan digunakan.

TABEL 4
DATA SETELAH PROSES TRANSFORMASI

No	JK	BB/U	BB/TB	Naik Berat Badan
0	0	0	0	2
1	0	0	0	3
2	0	0	0	2
...	...	...	...	...
8056	0	0	0	3
8057	0	0	0	2
8058	0	0	0	1

Tabel 4 menampilkan data setelah proses transformasi. Setelah melalui tahap ini, nilai-nilai atribut yang bersifat kategorikal telah diubah menjadi bentuk numerik. Contoh atribut yang ditransformasi termasuk jenis kelamin, pertambahan berat badan, berat badan menurut usia, dan berat badan menurut tinggi badan.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 348-358

D. *Seleksi Fitur Chi Square*

Setelah melakukan proses transformasi data, langkah berikutnya adalah seleksi fitur menggunakan metode *Chi-Square*. Metode ini digunakan untuk mengidentifikasi atribut-atribut yang memiliki pengaruh signifikan terhadap klasifikasi stunting.

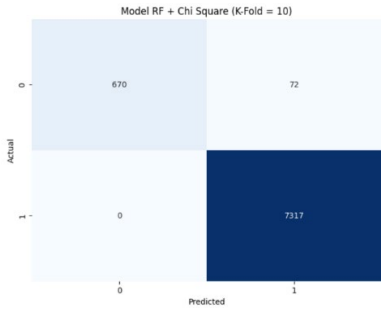
TABEL 5
HASIL SELEKSI FITUR CS

Rangking	Atribut	Nilai *Chi Square*
1	ZS TB/U	1783
2	BB/U	1086
3	Tinggi	652
4	ZS BB/U	562
5	LiLA	355
6	Berat	288
7	BB/TB	15
8	JK	4
9	ZS BB/TB	4
10	Naik Berat Badan	0

Tabel 5 menampilkan hasil seleksi fitur menggunakan *Chi-Square*. Dari hasil tersebut, terlihat bahwa enam fitur dengan nilai *Chi-Square* tertinggi adalah ZS TB/U, BB/U, Tinggi, ZS BB/U, LiLA, dan Berat. Fitur-fitur ini dianggap signifikan dan memiliki pengaruh besar terhadap klasifikasi stunting. Dengan mengeliminasi fitur-fitur yang tidak signifikan, model yang dihasilkan diharapkan dapat menjadi lebih akurat dan efisien dalam memprediksi stunting pada anak-anak di Kota Samarinda.

E. *Implementasi Random Forest*

Pada tahap awal pengujian, model *Random Forest* diterapkan tanpa menggunakan optimasi *Grid Search* untuk mengklasifikasikan data stunting di Kota Samarinda. Pada tahap awal pengujian, model *Random Forest* diterapkan tanpa menggunakan optimasi *Grid Search* untuk mengklasifikasikan data stunting di Kota Samarinda. Pengujian dilakukan dengan menggunakan teknik validasi silang *K-Fold* dengan 10 lipatan (10-*fold cross-validation*). Teknik validasi silang ini dipilih untuk memastikan bahwa model diuji secara menyeluruh dan memberikan gambaran yang akurat tentang kinerjanya pada data yang berbeda. Proses ini diulang sebanyak 10 kali sehingga setiap *fold* berfungsi sebagai data uji satu kali. Hasil pengujian dari setiap *fold* dicatat untuk menilai variasi kinerja model pada setiap subset data.



Gambar. 2. *Confusion Matrix Random Forest* Tanpa GS

$$Akurasi = \frac{670 + 7317}{670 + 7317 + 0 + 72} = \frac{7987}{8059} = 0.9911$$

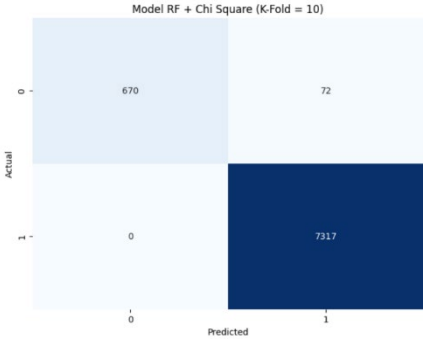
Hasil permodelan menunjukkan bahwa *Random Forest* mampu melakukan klasifikasi data stunting dengan baik pada berbagai subset data, sehingga memiliki kemampuan generalisasi yang kuat. Rata-rata akurasi yang diperoleh sebesar 99.11%, yang mengindikasikan performa model yang sangat tinggi dalam memprediksi data stunting di Kota Samarinda.

F. *Implementasi Random Forest Dengan Grid Search*

Setelah menguji model *Random Forest* tanpa optimasi parameter, langkah berikutnya adalah melakukan optimasi untuk meningkatkan kinerja model. Optimasi dilakukan dengan menggunakan teknik *Grid Search*, yang bertujuan

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 348-358

untuk menemukan kombinasi parameter terbaik yang menghasilkan kinerja model optimal. Dalam penelitian ini, parameter yang dioptimasi adalah *max_depth* dari model *Random Forest* dengan nilai (2). Proses *Grid Search* melibatkan pengujian setiap kombinasi parameter melalui validasi silang *K-Fold* dengan 10 lipatan (10-*fold cross-validation*) yang sama seperti pada tahap sebelumnya. Teknik ini memastikan bahwa setiap parameter diuji secara menyeluruh dan kinerja model dioptimalkan berdasarkan hasil terbaik dari kombinasi parameter tersebut.



Gambar. 3. *Confusion Matrix Random Forest* Dengan GS

$$Akurasi = \frac{670 + 7317}{670 + 7317 + 0 + 72} = \frac{7987}{8059} = 0.9911$$

Hasil permodelan *Random Forest* mampu mempertahankan kinerjanya secara konsisten pada berbagai subset data. Selain itu, rata-rata akurasi dari semua *fold* juga sangat baik, dengan nilai yang mendekati 99.11%. Ini mengindikasikan bahwa model tidak hanya bekerja baik pada satu subset data, tetapi juga secara keseluruhan memiliki performa yang solid dan andal dalam mengklasifikasikan data stunting.

G. Pembahasan

Hasil dari penelitian ini menunjukkan bahwa optimasi parameter menggunakan *Grid Search* tidak memberikan peningkatan yang signifikan dalam akurasi model *Random Forest*.

TABEL 6
PERBANDINGAN AKURASI *RANDOM FOREST*

Fold	**RF Tanpa *Grid Search***	**RF Dengan *Grid Search***	**Status**
1	99.01%	99.01%	Tetap
2	98.76%	98.76%	Tetap
3	99.75%	99.75%	Tetap
4	98.39%	98.39%	Tetap
5	98.88%	98.88%	Tetap
6	98.64%	98.64%	Tetap
7	99.01%	99.01%	Tetap
8	99.26%	99.26%	Tetap
9	99.50%	99.50%	Tetap
10	99.88%	99.88%	Tetap

TABEL 7
PERBANDINGAN RATA- RATA AKURASI *RANDOM FOREST*

RF Tanpa *Grid Search*	**RF Dengan *Grid Search***
99.11%	99.11%

ss.

Penelitian oleh [19] di Indonesia menggunakan kombinasi algoritma *Random Forest* dengan *Support Vector Machine* (SVM) dan seleksi fitur *algoritma genetika* menunjukkan bahwa fitur berat lahir dan tinggi lahir merupakan prediktor penting dalam model stunting, mencapai akurasi tertinggi menggunakan *Random Forest* dan SVM dengan akurasi 98.5%. Metode ini membantu mengidentifikasi faktor-faktor utama yang mempengaruhi

stunting pada anak-anak. Penelitian lain oleh [20] di Bangladesh menggunakan berbagai algoritma *machine learning* untuk memprediksi malnutrisi, menunjukkan bahwa fitur-fitur seperti tinggi menurut usia dan berat menurut usia merupakan prediktor penting dalam model mereka, hal ini menunjukkan bahwa penggunaan fitur *Chi Square* meningkatkan kinerja model *Random Forest* dalam klasifikasi stunting.

Penelitian di Zambia menggunakan Random Forest yang menunjukkan akurasi tertinggi sebesar 79% dalam memprediksi stunting, dengan faktor penting seperti status ekonomi keluarga dan usia anak[12]. Sedangkan di Papua New Guinea, model LASSO-XGBoost dengan AUC 0.765 mengidentifikasi tempat tinggal dan status ekonomi sebagai faktor kunci[21]. Konsistensi ditemukan dalam efektivitas model machine learning untuk prediksi stunting di berbagai studi, sementara perbedaan terletak pada algoritma dan variabel yang digunakan serta konteks geografis yang berbeda.

IV. KESIMPULAN

Penelitian ini menunjukkan bahwa fitur-fitur yang paling mempengaruhi performa model *Random Forest* dalam klasifikasi data stunting di Kota Samarinda adalah atribut terkait kondisi fisik dan kesehatan anak, seperti berat badan menurut umur (BB/U), tinggi badan (Tinggi), berat badan (Berat), lingkar lengan atas (LiLA), serta beberapa indikator gizi seperti Z *Score* berat badan menurut umur (ZS BB/U) dan Z *Score* tinggi badan menurut umur (ZS TB/U). Penerapan metode *Random Forest*, baik tanpa optimasi maupun dengan optimasi *Grid Search*, menghasilkan akurasi tinggi sebesar 99.11%. Optimasi parameter melalui *Grid Search* tidak memberikan peningkatan signifikan pada kinerja model, yang menunjukkan bahwa model *Random Forest* tanpa optimasi sudah cukup optimal untuk klasifikasi data stunting di Kota Samarinda.

DAFTAR PUSTAKA

[1] D. A. P. Ramadhan and M. J. Ahmad, "Pertanggungjawaban Negara Terhadap Permasalahan Anak Stunting Di Indonesia," *Civilia J. Kaji. Huk. dan Pendidik. Kewarganegaraan*, vol. 3, no. 1, pp. 14–26, 2024, [Online]. Available: http://jurnal.anfa.co.id/index.php/civilia/article/view/1650/1532

[2] S. Handayani, "Save the Nation's Generation From the Dangers of Stunting," *J. Midwifery Sci. Women's Heal.*, vol. 3, no. 2, pp. 87–92, 2023, doi: 10.36082/jmswh.v3i2.1082.

[3] F. Noviasty, R., Mega I., Fadillah R., "EDUWHAP Remaja Siap Cegah Stunting Dalam Wadah Kumpul Sharing Remaja," *J. Ilm. Pengabdi. Kpd. Masy.*, vol. 4, no. 2, pp. 494–501, 2020, [Online]. Available: file:///C:/Users/HP/Downloads/Documents/458-1-1543-1-10-20210127.pdf

[4] Nety, "OPTIMIS 2024 STUNTING DI KALTIM TURUN HINGGA 12,83%," *PEMPROV KALTIM*, 2023. https://www.kaltimprov.go.id/berita/optimis-2024-stunting-di-kaltim-turun-hingga-1283#:~:text=Hasil Survei Status Gizi Indonesia,2021 sebesar 22%2C8%25.

[5] O. Shobayo, O. Zachariah, M. O. Odusami, and B. Ogunleye, "Prediction of Stroke Disease with Demographic and Behavioural Data Using Random Forest Algorithm," *Analytics*, vol. 2, no. 3, pp. 604–617, 2023, doi: 10.3390/analytics2030034.

[6] S. M. Ganie and M. B. Malik, "An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators," *Healthc. Anal.*, vol. 2, no. January, p. 100092, 2022, doi: 10.1016/j.health.2022.100092.

[7] N. F. Sahamony, T. Terttiaavini, and H. Rianto, "Analysis of Performance Comparison of Machine Learning Models for Predicting Stunting Risk in Children ' s Growth," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. April, pp. 413–422, 2024.

[8] R. Gustriansyah, N. Suhandi, S. Puspasari, and A. Sanmorino, "Machine Learning Method to Predict the Toddlers' Nutritional Status," *Infotel*, vol. 16, no. 1, pp. 1–6, 2024, [Online]. Available: https://ejournal.ittelkom-pwt.ac.id/index.php/infotel/article/view/988

[9] I. P. Putri, T. Terttiaavini, and N. Arminarahmah, "Analisis Perbandingan Algoritma Machine Learning untuk Prediksi Stunting pada Anak," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 257–265, 2024, doi: 10.57152/malcom.v4i1.1078.

[10] R. A. Azizah, F. Bachtiar, and S. Adinugroho, "Klasifikasi Kinerja Akademik Siswa Menggunakan Neighbor Weighted K-Nearest Neighbor dengan Seleksi Fitur Information Gain," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 3, pp. 605–614, 2022, doi: 10.25126/jtiik.2022935751.

[11] M. S. H. Bhuiyan, N. Al Raian, S. I. Leon, and M. Khan, "Study of Influence of Dimension Reduction of High Dimensional Datasets in Classification Problem," *Proc. 4th Int. Conf. Comput. Methodol. Commun. ICCMC 2020*, no. Iccmc, pp. 147–151, 2020, doi: 10.1109/ICCMC48092.2020.ICCMC-00030.

[12] O. N. Chilyabanyama *et al.*, "Performance of Machine Learning Classifiers in Classifying Stunting among Under-Five Children in Zambia," *Children*, vol. 9, no. 7, 2022, doi: 10.3390/children9071082.

[13] E. A. Turjo and M. H. Rahman, "Assessing risk factors for malnutrition among women in Bangladesh and forecasting malnutrition using machine learning approaches," *BMC Nutr.*, vol. 10, no. 1, pp. 1–25, 2024, doi: 10.1186/s40795-023-00808-8.

[14] T. A. Yoga, "DATA MINING - MENGUPAS TUNTAS ANALISIS DATA DENGAN METODE KLASIFIKASI HINGGA DEPLOYMENT APLIKASI MENGGUNAKAN PYTHON," 2023.

[15] M. M. Islam *et al.*, "Application of machine learning based algorithm for prediction of malnutrition among women in Bangladesh," *Int. J. Cogn. Comput. Eng.*, vol. 3, no. January, pp. 46–57, 2022, doi: 10.1016/j.ijcce.2022.02.002.

[16] S. M. J. Rahman *et al.*, "Investigate the risk factors of stunting, wasting, and underweight among under-five Bangladeshi children and its prediction based on machine learning approach," *PLoS One*, vol. 16, no. 6 June 2021, pp. 1–11, 2021, doi: 10.1371/journal.pone.0253172.

[17] P. Dewi, P. Purwono, and S. Kurniawan Dwi, "Pemanfaatan Teknologi Machine Learning pada Klasifikasi Jenis Hipertensi Berdasarkan Fitur Pribadi," *Smart Comp Jurnalnya Orang Pint. Komput.*, vol. 11, no. 3, pp. 377–387, 2022, doi: 10.30591/smartcomp.v11i3.3721.

[18] M. M. Sugiman and H. D. Purnomo, "Prediksi Kegagalan Transformator Daya dengan Metode DGA (Dissolved Gas Analysis) Menggunakan Random Forest Berbasis TDCG," vol. 8, pp. 441–449, 2024, doi: 10.30865/mib.v8i1.7036.

[19] S. Sutarmi, W. Warijan, T. Indrayana, D. P. P. B, and I. Gunawan, "Machine Learning Model For Stunting Prediction," *J. Heal. Sains*, vol. 4, no. 9, pp. 10–23, 2023, doi: 10.46799/jhs.v4i9.1073.

[20] A. Talukder and B. Ahammed, "Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh," *Nutrition*, vol. 78, 2020, doi: 10.1016/j.nut.2020.110861.

[21] H. Shen, H. Zhao, and Y. Jiang, "Machine Learning Algorithms for Predicting Stunting among Under-Five Children in Papua New Guinea," *Children*, vol. 10, no. 10, 2023, doi: 10.3390/children10101638.