Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 4 Iss. 2 April 2024, pp: 413-422
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Analysis of Performance Comparison of Machine Learning Models for Predicting Stunting Risk in Children's Growth

Analisis Perbandingan Kinerja Model Machine Learning untuk Memprediksi Risiko Stunting pada Pertumbuhan Anak

Nur Fitriyani Sahamony[1*], Terttiaavini[2], Harsih Rianto[3]

[1]Program Studi Bisnis Digital, Fakultas Bisnis dan Ilmu Sosial, Universitas Binawan, Indonesia
[2]Program Studi Sistem Informasi, Fakultas Ilmu Komputer dan Sains,
Universitas Indo Global Mandiri, Indonesia
[3] Program Studi Teknologi Informasi, Fakultas Teknik dan Informatika,
Universitas Bina Sarana Informatika, Indonesia

E-Mail: [1]mony@binawan.ac.id, [2]avini.saputra@uigm.ac.id, [3]harsih.hhr@bsi.ac.id

Received Nov 12th 2023; Revised Jan 20th 2024; Accepted Feb 15th 2024
Corresponding Author: Nur Fitriyani Sahamony

Abstract

Stunting is a serious problem in child growth in Indonesia, prompting this research to develop a prediction model using Machine Learning. The research objective is to compare the performance of five algorithms namely Random Forest, Logistic Regression, Naïve Bayes, SVM and Neural Networks to predict child stunting. Child stunting data for 2023 from Lubuk Linggau City was used with a total of 400 samples. The research methodology involved initiation steps, linear model development, model testing results comparison, and prediction analysis using the KNIME platform. The test results showed that Naïve Bayes had the highest performance with accuracy = 98.57%, F1-Score = 0.99, and very high recall and precision. Random Forest also gave good results with accuracy = 98.29%, but Naïve Bayes was identified as the best model. This research makes a significant contribution to stunting prevention by combining Machine Learning technology and health dataset analysis. Developing prediction models using various machine learning algorithms, is expected to help health practitioners in identifying the risk of stunting early. The optimal model can be used as a decision-support tool to provide appropriate and effective interventions.

Keyword: KNIME, Machine Learning, Naïve Bayes, Prediction Model

Abstrak

Stunting menjadi masalah serius dalam pertumbuhan anak di Indonesia, mendorong penelitian ini untuk mengembangkan model prediksi menggunakan *Machine Learning*. Tujuan penelitian adalah membandingkan performa dari lima algoritma yaitu *Random Forest*, *Logistic Regression*, *Naïve Bayes*, SVM dan *Neural Networks* untuk memprediksi stunting anak. Data stunting anak tahun 2023 dari Kota Lubuk Linggau yang digunakan dengan total 400 sampel. Metodologi penelitian melibatkan langkah inisiasi, pengembangan model linier, pembandingan hasil pengujian model, dan analisis prediksi menggunakan platform KNIME. Hasil uji menunjukkan bahwa *Naïve Bayes* memiliki performa tertinggi dengan akurasi = 98,57%, *F1-Score* = 0,99, serta *recall* dan *precision* yang sangat tinggi. *Random Forest* juga memberikan hasil baik dengan akurasi = 98,29%, namun *Naïve Bayes* diidentifikasi sebagai model terbaik. Penelitian ini memberikan kontribusi signifikan dalam upaya untuk pencegahan stunting dengan menggabungkan teknologi *Machine Learning* dan analisis dataset kesehatan. Dengan mengembangkan model prediksi menggunakan berbagai algoritma *machine learning*, diharapkan dapat membantu praktisi kesehatan dalam mengidentifikasi risiko stunting secara dini. Model yang optimal dapat digunakan sebagai alat pendukung keputusan untuk memberikan intervensi yang tepat dan efektif.

Kata Kunci : KNIME, Machine Learning, Naïve Bayes, Stunting, Prediction Model

1. PENDAHULUAN

Stunting, sebuah kondisi yang menandai kegagalan pertumbuhan anak akibat kekurangan gizi kronis yang telah menjadi permasalahan serius di Indonesia [1]. Secara global, menurut data dari UNICEF dan WHO,

Indonesia berada pada peringkat ke-27 dari 154 negara, sedangkan di kawasan Asia Indonesia berada di peringkat ke-5 dalam tingkat prevalensi stunting [2]. Dampak dari stunting menyebabkan pertumbukan anak mengalami keterlabatan dalam tumbuh kembang [3]. Pada masa Covid-19 berdampak pada penurunan ekonomi [4], hal ini juga menyumbangkan peningkatan jumlah stunting. Meskipun berbagai upaya telah dilakukan, angka stunting sampai saat ini masih tinggi [5]. Oleh karena itu, diperlukan upaya preventif untuk menangani masalah tersebut.

Penelitian ini bertujuan untuk mengembangkan model prediksi pertumbuhan stunting pada anak dengan membandingkan beberapa algoritma *Machine learning* untuk menemukan model yang paling optimal. Model ini diharapkan dapat membantu praktisi kesehatan dalam mengidentifikasi risiko stunting pada anak untuk memberikan intervensi dini. Penelitian ini akan melakukan identifikasi faktor-faktor yang signifikan sebagai penyebab stunting pada anak-anak. Faktor-faktor ini diharapkan dapat memberikan wawasan mendalam untuk pengembangan strategi pencegahan yang lebih efektif

Untuk menentukan model prediksi yang paling optimal, dilakukan pengujian menggunakan lima algoritma ML, yaitu *Random Forest*, *Logistic Regression*, *Naïve Bayes*, SVM, dan *Neural Networks*. Setiap algoritma memiliki keungulannya masing-masing, dan kinerja algoritma dapat dipengaruhi oleh karakteristik *dataset*. Sebagai contoh, dalam penelitian oleh M Syauqi dkk (2022) tentang prediksi prevalensi stunting di Provinsi Jawa Timur, algoritma *support vector regression* terbukti menjadi yang terbaik dengan nilai MAE = 0,91 dan MSE = 1,30 [6]. Penelitian lain, seperti yang dilakukan oleh Putri dkk (2024) dalam memprediksi stunting pada anak, menunjukkan bahwa *Random Forest* memiliki nilai akurasi tertinggi sebesar 87,75% [7]. Begitu pula dalam penelitian oleh Amirudin dan Wowor (2023), algoritma SVM menjadi yang terbaik dengan nilai akurasi sebesar 83% [8]. Hasil penelitian ini menegaskan bahwa pemilihan algoritma prediksi yang optimal sangat tergantung pada jenis data yang diolah. Oleh karena itu, untuk mencapai tujuan penelitian ini, yaitu menghasilkan model yang paling optimal, perlu dilakukan perbandingan beberapa algoritma *machine learning*.

Untuk meningkatkan akurasi proses perhitungan *machine learning*, penelitian ini memanfaatkan KNIME sebagai platform analisis data yang lebih akurat. KNIME, singkatan dari *Konstanz Information Miner*, merupakan platform analisis data sumber terbuka yang menyediakan lingkungan grafis untuk pengolahan data, pemodelan, dan analisis. Dengan memanfaatkan fitur-fitur KNIME, penelitian ini dapat menjalankan berbagai tahapan eksperimen, termasuk *pre-processing* data, pemodelan *machine learning*, dan evaluasi hasil secara efisien. Penggunaan KNIME dalam penelitian ini diharapkan memberikan kontribusi pada kemudahan penggunaan dan efektivitas analisis data dalam penelitian ini.

Hasil dari penelitian ini dapat memberikan kontribusi yang signifikan dalam upaya pencegahan stunting pada anak-anak. Diharapkan agar model yang dibangun dapat diterapkan untuk mendukung implementasi intervensi yang lebih tepat dan efektif dalam mengatasi masalah stunting pada anak.

2. TINJAUAN PUSTAKA

2.1. Parameter Stunting

Di berbagai negara, perbedaan dalam indikator penilaian stunting disebabkan oleh perbedaan penggunaan metode pengukuran dan parameter untuk menilai status stunting. Perbedaan ini terkait dengan penggunaan standar pertumbuhan anak yang berbeda, seperti di Rwanda, faktor-faktor seperti Indeks Massa Tubuh (BMI) ibu, durasi menyusui, usia, berat badan lahir, jenis kelamin, urutan kelahiran, dan faktor penyakit, diidentifikasi sebagai faktor yang berperan dalam stunting [9], di Bangladesh, faktor yang mempengaruhi stunting meliputi berat badan ibu, tingkat pendidikan ibu, kerawanan pangan, akses terhadap nutrisi, pemberian ASI eksklusif, dan kasus diare [10], di Kenya faktor yang mempengaruhi adalah berat badan, wilayah, usia anak, etnis, dan usia ibu [11], sedangkan di Indonesia, risiko stunting disebabkan oleh ekonomi, jenis kelamin, berat badan bayi lahir rendah (BBLR), tingkat pendidikan orang tua, tinggi badan orang tua, usia anak, pemberian ASI eksklusif, riwayat infeksi, serta pemberian makanan pendamping ASI (MPASI) [12]. Perbedaan dalam indikator tersebut menggambarkan kompleksitas masalah di setiap negara. Variasi ini mencerminkan perbedaan dalam konteks sosial, ekonomi, gizi, dan kesehatan antara negara-negara tersebut.

2.2. *Machine learning* (ML)

ML adalah metode yang digunakan untuk memprediksi hasil atau perilaku masa depan berdasarkan pola dan informasi yang diberikan kepada komputer. Teknik ini melibatkan penggunaan algoritma dan model statistik untuk mengidentifikasi pola dalam data dan membuat prediksi atau keputusan tanpa instruksi yang eksplisit. Implementasi ML telah banyak di terapkan untuk melakukan prediksi dalam berbagai kasus [13], [14]. Pada kasus stunting, beberapa metode ML yang telah digunaakan adalah metode DT [15], *Naïve Bayes* [16], [17], SVM [18], *k-Nearest Neighbors* [19], *Neural Networks* [20], [21], *Random Forest* [22], [23] dan *Logistic Regression* [24][25]. Selain pada kasus stunting, ML juga banyak di terapkan pada bidang Kesehatan untuk memprediksi penyakit jantung, seperti *Random forest* [26], *Logistic Regression* [27], *Naïve Bayes* [28], SVM [29], dan *Neural Networks* [29], Penyakit diabetes : *Random Forest* [30], *Neural Network* [31], Penyakit

kutil: *Random Forest* [30], penyakit kulit : *Neural Networks* [32], Kanker payudara : *Logistic Regression* [33], *Naïve Bayes* [34] dan Gejala covid-19 : SVM [30]. Hasil dari prediksi tersebut menunjukkan bahwa implementasi ML memberikan kontribusi yang signifikan dalam melakukan prediksi pada berbagai bidang Kesehatan. Penerapan metode ML dapat meningkatkan akurasi prediksi dan memberikan dampak positif dalam upaya pencegahan dan penanganan berbagai penyakit.

Setiap metode memiliki keunggulan dan kelemahan dalam pemodelan dan kemampuannya dalam mengklasifikasikan kasus stunting berdasarkan data yang tersedia. Penelitian ini akan mengeksplorasi efektivitas berbagai metode ML dalam memprediksi kasus stunting berdasarkan faktor-faktor risiko yang di identifikasi. Beberapa metode ML yang digunakan dalam penelitian akan di terapkan untuk menentukan model yang paling optimal untuk memprediksi stunting pada anak.

2.3. *Random Forest (RF)*

Metode *Random Forest* dapat meningkatkan akurasi dengan membangun banyak pohon keputusan secara acak. Proses pembangunan pohon keputusan dimulai dengan simpul akar, dilanjutkan dengan simpul percabangan yang memiliki minimal dua output, dan akhirnya simpul daun yang memiliki satu input. Pengambilan keputusan pada setiap simpul dilakukan dengan mengukur tingkat ketidakmurnian atribut menggunakan entropy dan menghitung information gain untuk menentukan atribut yang paling berguna dalam membuat keputusan. Dengan menggabungkan nilai *information gain* dan penggunaan atribut secara acak, *Random Forest* membangun pohon keputusan yang beragam, menghasilkan prediksi yang lebih kuat dan akurat. Untuk menghitung nilai entropy digunakan persamaan 1 dan 2.

$$Entropy\ (Y) = -\sum_i \mathrm{p(c|Y)}\ \log_2\ \mathrm{p(c|Y)} \tag{1}$$

Dimana : Y adalah himpunan kasus, p(c|Y) merupakan proporsi nilai Y terhadap kelas c.

$$Information\ Gain\ (Y, a) = Entropy\ (Y) - \sum_{ve} values\ \frac{Y_v}{Y_a}\ Entropy\ (Y_v) \tag{2}$$

Dimana : values $\frac{Y_v}{Y_a}$ merupakan semua nilai yang mungkin dalam himpunan kasus a, Y_v adalah subkelas dari Y dengan kelas v yang berhubungan dengan kelas a. Ya adalah semua nilai yang sesuai dengan a. Pemilihan atribut untuk menjadi simpul, baik sebagai akar (*root*) maupun simpul internal, tergantung pada *information gain* tertinggi yang dimiliki oleh atribut-atribut yang tersedia. *Gain ratio* dihitung dengan membagi hasil dari perhitungan *information gain* dengan nilai *split information*. *Split information* yang digunakan dapat ditemukan dalam persamaan 3, sementara nilai gain ratio terlihat dalam persamaan 4.

$$Split\ Information\ (S, A) = \sum_i^c \frac{|S_i|}{|S|}\ \log_2\ \frac{|S_i|}{|S|} \tag{3}$$

Dimana: $Split\ Information\ (S, A)$ adalah perkiraan nilai entropy dari variabel input S yang memiliki kelas c, sementara $\frac{|S_i|}{|S|}$ merepresentasikan probabilitas dari kelas i dalam atribut tersebut.

$$Gain\ ratio\ (S, A) = \frac{Information\ (S, A)}{Split\ information\ (S, A)} \tag{4}$$

2.4. *Logistic Regression*

Logistic Regression adalah sebuah metode dalam statistik yang digunakan untuk memodelkan hubungan antara variabel dependen kategorikal dengan satu atau lebih variabel independen. Meskipun namanya mengandung istilah "regresi," namun sebenarnya *Logistic Regression* merupakan algoritma klasifikasi yang digunakan untuk memprediksi probabilitas terjadinya suatu peristiwa dengan output berupa kelas diskrit, umumnya dalam bentuk biner (misalnya, ya/tidak, 0/1). Algoritma ini menggunakan fungsi logistik untuk memetakan variabel input ke dalam nilai probabilitas dalam rentang 0 hingga 1, yang kemudian diinterpretasikan sebagai prediksi kelas target. Rumus *Logistic Regression* untuk kasus biner (dua kelas) ditunjukkan pada persamaan 5.

$$P(Y = 1|X) = \frac{1}{1 + e^{(\beta_{0+}\ \beta_1 X_1 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}} \tag{5}$$

Dimana : $P(Y = 1|X)$ adalah probabilitas bahwa variabel target Y adalah 1 (kemungkinan terjadinya peristiwa yang diprediksi) berdasarkan variabel input X, e adalah konstanta logaritma yang bernilai ± 2.71828, $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ adalah koefisien yang harus diestimasi dari data latih, $X_1, X_2, \dots, X_n$ adalah variable input. Rumus ini menggunakan fungsi logistik (atau sigmoid) untuk mengubah nilai hasil dari regresi linear ke dalam rentang 0 hingga 1, yang diinterpretasikan sebagai probabilitas kelas positif (dalam kasus biner, kelas 1 atau "Ya"). Estimasi koefisien (β) dilakukan dengan menggunakan teknik seperti metode maksimum *likelihood*.

2.5. *Naïve Bayes*

Naive Bayes merupakan algoritma klasifikasi yang sederhana namun efisien dengan keunggulan dalam implementasi yang mudah, kinerja yang baik pada dataset besar, dan kemampuan menangani *multi-class classification*. Algoritma ini terbukti efektif terutama pada data dengan jumlah kategori yang besar dan dapat memberikan hasil yang cukup baik meskipun asumsi dasar independensi fitur yang sederhana. Meskipun tidak selalu cocok untuk semua jenis data, *Naive Bayes* tetap menjadi pilihan yang menarik terutama ketika diterapkan pada konteks di mana asumsi dasarnya dapat dipertahankan dan interpretabilitas model menjadi prioritas. Rumus dasar *Naïve Bayes* dapat dijelaskan menggunakan teorema *Bayes*, ditunjukkan pada persamaan 6.

$$P(A|B) = \frac{P(B|A)\ .\ P(A)}{P(B)} \tag{6}$$

Dalam konteks klasifikasi, kita memiliki kelas target (A) dan fitur-fitur (B). Naïve Bayes menghitung probabilitas kelas (C_k) untuk suatu data (X) berdasarkan fitur-fiturnya pada persamaan 7.

$$P(C_k|X) = \frac{P(X|C_k)\ .\ P(C_k)}{P(X)} \tag{7}$$

Dimana $(C_k|X)$ adalah probabilitas kelas C_k mengingat data X, $P(X|C_k)$ adalah likelihood dari data C_k, $P(C_k)$ adalah prior probability dari kelas C_k, $P(X)$ adalah probabilitas margin dari data X. Asumsi utama Naïve Bayes adalah bahwa fitur-fitur X bersifat independen jika diketahui kelasnya. Sehingga, likelihood dapat dihitung sebagai perkalian dari probabilitas individu fitur, ditunjukkan pada persamaan 8.

$$P(C_k|X) = P(x_1|C_k).(x_{21}|C_k) \dots (x_{1n}|C_n) \tag{8}$$

2.6. *Support Vector Machine* (SVM)

SVM adalah algoritma pembelajaran mesin yang digunakan untuk tugas klasifikasi dan regresi. Algoritma ini mencari garis atau permukaan pemisah optimal di antara kelas-kelas data dengan memanfaatkan vektor pendukung. SVM memiliki keunggulan dalam menangani ruang berdimensi tinggi dan ketidaklinieran data, serta dapat diterapkan pada masalah klasifikasi multi-kelas. Meskipun kompleks, SVM umumnya diandalkan dalam berbagai konteks, termasuk pengenalan pola, bioinformatika, dan analisis citra.

Dalam konteks klasifikasi, SVM mencari *hyperplane* terbaik yang dapat memisahkan dua kelas data. *Hyperplane* ini dipilih sedemikian rupa sehingga margin (jarak) antara dua kelas maksimal. Rumus dasar hyperplane dalam SVM dapat dinyatakan sebagai:

$$f(x) = \omega.x + b \tag{9}$$

Dimana : ω adalah vektor bobot yang tegak lurus terhadap *hyperplane*, x adalah vektor fitur input b adalah bias atau pergeseran. Dalam konteks klasifikasi, SVM berusaha memaksimalkan margin, yaitu jarak antara *hyperplane* dan sampel data terdekat dari setiap kelas. Margin ini dihitung sebagai $\frac{2}{\|\omega\|}$ di mana $\|\omega\|$ adalah norma Euclidean dari vektor bobot. SVM juga dapat memanfaatkan fungsi kernel untuk menangani kasus di mana data tidak dapat dipisahkan secara linear di ruang asli. Salah satu formulasi SVM dengan kernel ditunjukkan pada persamaan 10.

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b \tag{10}$$

Dimana: K adalah fungsi kernel, α adalah bobot dari vektor dukungan, y adalah label kelas, dan b adalah bias. Untuk klasifikasi, prediksi kelas dilakukan dengan menghitung $f(x)$ dan menentukan kelas

berdasarkan tanda dari $f(x)$ (positif atau negatif). Parameter SVM, seperti ω. α dan b ditentukan melalui proses pelatihan yang melibatkan optimasi untuk mencapai margin maksimal dan mengatasi kesalahan klasifikasi.

2.7. *Neural Networks* (NNK)

Neural Networks adalah model matematis yang terinspirasi oleh struktur dan fungsi jaringan saraf biologis. Jaringan Saraf Tiruan terdiri dari lapisan-lapisan neuron atau unit pemrosesan informasi yang saling terhubung. Salah satu jenis *Neural Networks* yang umum digunakan adalah Jaringan Saraf Tiruan Multilayer (*Multilayer Neural Network*) yang terdiri dari lapisan input, lapisan tersembunyi, dan lapisan output**.** Rumus umum untuk menghitung output dari suatu neuron pada lapisan *Neural Networks* pada persamaan 11.

$$a = \sigma(\omega.x + b) \tag{11}$$

Dimana: a adalah output dari neuron, σ adalah fungsi aktivasi, ω adalah vektor bobot, x adalah input, b adalah bias. Fungsi aktivasi diperlukan untuk memperkenalkan non-linearitas ke dalam model, memungkinkan *Neural Networks* untuk memodelkan hubungan yang kompleks. Proses ini diulang melalui lapisan-lapisan hingga mencapai lapisan output. Selama proses pelatihan, bobot dan bias disesuaikan menggunakan algoritma pembelajaran, seperti algoritma penurunan gradien stokastik (*Stochastic Gradient Descent*).

2.8. *Konstanz Information Miner* (KNIME)

KNIME merupakan platform analisis data yang sangat berguna dalam konteks ML prediktif. Penggunaan KNIME dalam ML prediksi stunting melibatkan beberapa tahap, mulai dari pembersihan dan persiapan data hingga pengembangan model prediktif dan evaluasinya. Antarmuka grafis pada KNIME memungkinkan pengguna untuk dengan mudah menentukan alur kerja analisis data (*workflow*) dan mengintegrasikan berbagai algoritma ML yang disediakannya

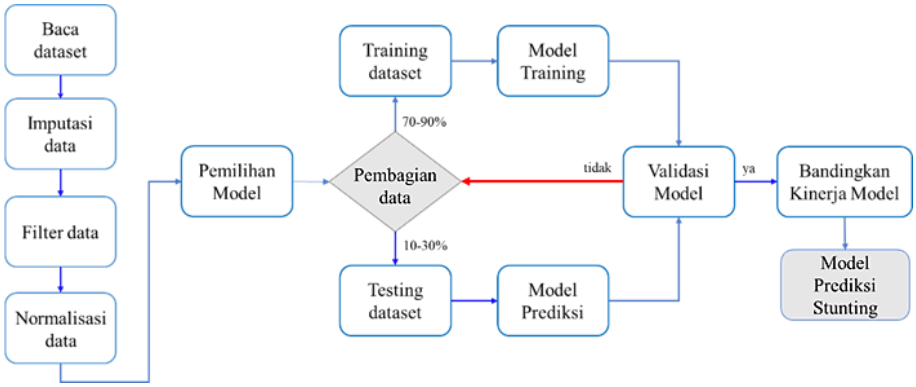
KNIME memiliki kelebihan signifikan dalam konteks ML prediksi stunting. Fleksibilitas dan keterbukaan KNIME memungkinkan integrasi mudah dengan berbagai sumber data dan algoritma ML, sementara proses otomatisasinya membantu meningkatkan efisiensi pengolahan data dan pembuatan model. Selain itu, dukungan aktif dari komunitas dan fitur visualisasi data yang kuat memberikan nilai tambah dalam pemahaman pola dan interpretasi hasil.

3. METODOLOGI PENELITIAN

Metodologi Penelitian merupakan pendekatan sistematik yang digunakan untuk merancang, melaksanakan, dan menganalisis data dalam suatu penelitian. Metodologi pada penelitian ini terdiri dari dari empat tahap utama, yaitu inisiasi, pengembangan model linier, pembandingan hasil pengujian model, dan analisis hasil prediksi. Langkah inisiasi tidak hanya mencakup pembersihan dan penyiapan data, tetapi juga identifikasi variabel yang paling berpengaruh terhadap prediksi stunting.

Model yang dikembangkan terdiri dari lima model dengan menggunakan parameter yang sama untuk memastikan konsistensi dalam perbandingan. Hal ini dilakukan agar hasil evaluasi dan performa masing-masing model agar dapat mengidentifikasi model yang paling optimal dalam konteks penelitian ini. Penggunaan parameter yang seragam juga membantu mengurangi faktor variabilitas yang dapat mempengaruhi hasil, sehingga mempermudah interpretasi dan generalisasi model.

Setelah model prediksi dikembangkan, tahap selanjutnya adalah membandingkan hasil pengujian model menggunakan metrik *Accuracy*, *Recall*, *Precision*, *F1-Score* dan *Cohens's Kappa*. Evaluasi ini memberikan wawasan tentang sejauh mana model dapat memprediksi kejadian stunting dengan tepat. Tahapan dalam pengembangan model di jelaskan dalam bentuk alur kerja pada Gambar 1.



Gambar 1. Alur Kerja Membangun Model Prediksi Stunting Pada Anak

3.1. Dataset Stunting Anak

Penelitian ini memanfaatkan data stunting anak yang diperoleh dari dinas Kesehatan kota Lubuk Linggau tahun 2023. *Dataset* tersebut mencakup informasi tentang stunting, berat badan anak, tinggi badan anak, umur ibu, tinggi badan ibu, pendidikan ibu, jumlah anak, jumlah anggota keluarga, pemberian ASI, pendapatan keluarga, pemberian bantuan, dan kepemilikan rumah yang merupakan faktor -faktor indikasi stunting pada anak. *Dataset* tersebut selajutnya di transformasi dari ordinal menjadi data numerik untuk dapat di proses pada *machine learning*. Jumlah data stunting adalah n = 298 data [35]. Data tersebut merupakan data sampel dari populasi N = 400. Informasi rinci dari dataset ditampilkan pada tabel 1.

Tabel 1. Data data stunting anak tahun 2023

Variabel	Frekuensi	Persentase
Stunting		
Stunting	298	74,5
Tidak Stunting	102	25,5
Berat badan anak		
Normal	182	45,75
Kurang	161	40,5
Sangat Kurang	50	13,75
Tinggi badan anak		
Normal	277	69,25
Pendek	4	2,75
Sangat Pendek	112	28
Umur Ibu		
Beresiko (<20 atau > 35)	97	24,25
Tidak beresiko (20-35)	303	75,75
Tinggi badan Ibu		
< 150	87	21,75
≥ 150	313	78,25
Pendidikan Ibu		
Tidak sekolah	14	3,5
SMA	229	57,25
Sarjana (S1/S2)	157	39,25
Jumlah anak		
1 anak	65	16,25
≥ 2 anak	335	83,75
Jumlah anggota Keluarga		
3-4 orang	249	62,25
5-6 orang	124	31,00
≥ 7 orang	27	6,75
Asi eksklusif		
Ya	273	68,25
Tidak	127	31,75
Pendapatan Keluarga		
< 3,5 jt	265	66,25
≥ 3,5 jt	135	33,75
Dapat bantuan		
Tidak pernah	119	29,75
Sering	281	70,25
Kepemilikan rumah		
Sewa / milik orangtua	298	74,5
Milik sendiri	102	25,5

3.2. Pra-pemrosesan data (*data preprocessing*)

Tahap pra-pemrosesan data merupakan langkah krusial dalam mempersiapkan *dataset* sebelum digunakan untuk analisis atau pembangunan model. Proses ini melibatkan sejumlah kegiatan, termasuk membaca data mentah dari sumbernya, mengisi nilai yang hilang melalui imputasi data, menerapkan filter untuk mengecilkan *dataset* dengan menghilangkan kolom-kolom yang tidak diperlukan seperti NIK, Nama anak, Tgl lahir, Alamat, Desa, Kecamatan, Kabupaten, Posyandu, dan Tanggal ukur. Selanjutnya, *dataset* dinormalisasi untuk menghasilkan representasi data biner, memungkinkan efisiensi dalam analisis atau pelatihan model ML. Langkah-langkah ini bertujuan untuk memastikan kebersihan, kelengkapan, dan kesiapan data tanpa mengorbankan informasi esensial, yang pada akhirnya meningkatkan kualitas hasil analisis atau model yang dikembangkan.

3.3. Pemilihan Model

ML memiliki beragam algoritma untuk memprediksi stunting pada anak. Setiap algoritma memiliki kelebihan dan kelemahan tersendiri. Untuk memperoleh model terbaik, perbandingan antar beberapa algoritma perlu dilakukan. Pada tahap pemilihan model ini, algoritma yang digunakan untuk pengujian adalah *Random Forest*, *Logistic Regression*, *Naïve Bayes*, SVM, dan *Neural Networks*. Penilaian model terbaik diperoleh dari nilai performa yang diukur menggunakan berbagai metrik evaluasi yaitu *Accuracy*, *Recall*, *Precision*, *F1-Score* dan *Cohens's Kappa*. Proses perbandingan ini memungkinkan untuk menentukan algoritma yang paling sesuai dan efektif untuk memodelkan masalah prediksi stunting anak dengan akurasi dan kinerja yang optimal.

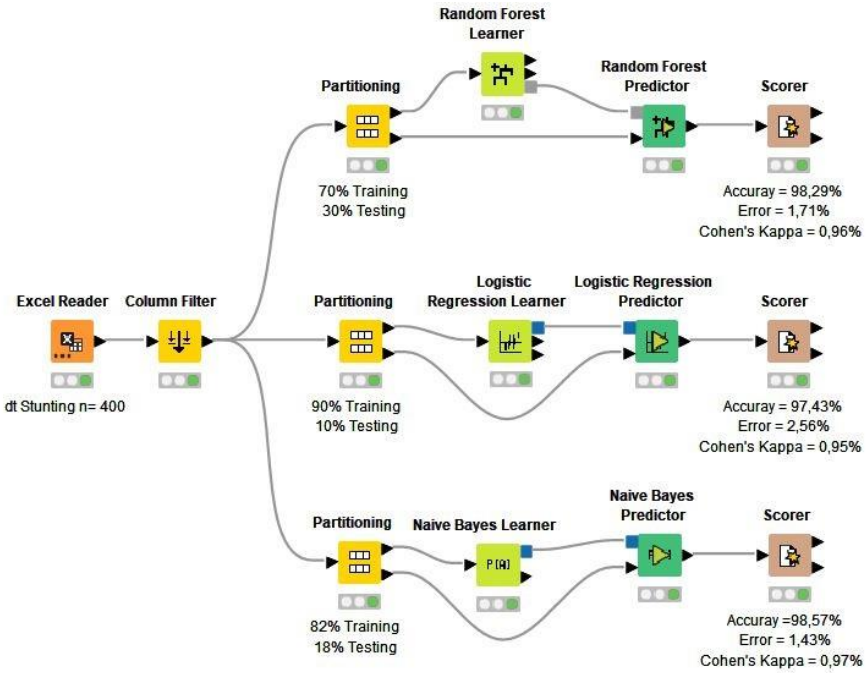
3.4. Pembagian Data (*Data Splitting*)

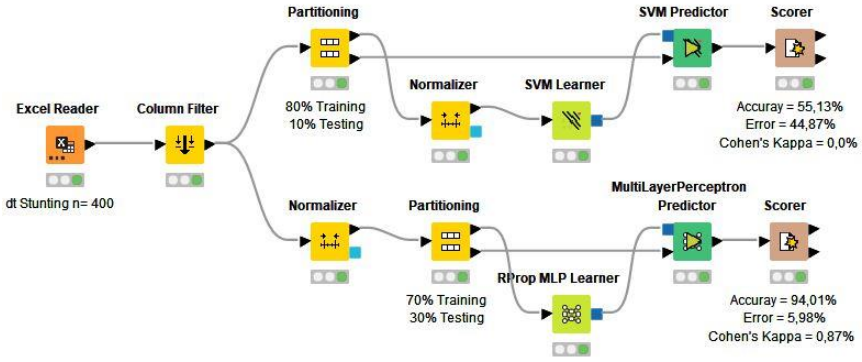
Setelah tahap pra-pemrosesan data, langkah berikutnya dalam pengembangan model *machine learning* adalah membagi dataset menjadi dua subset utama: *training* dan *testing*. Subset *training* digunakan untuk melatih model, sedangkan subset *testing* berfungsi untuk menguji performa model yang telah dilatih. Pembagian ini bertujuan untuk mengevaluasi sejauh mana model dapat melakukan prediksi yang akurat pada data baru yang belum pernah dilihat sebelumnya.

Praktik umum dalam pembagian *dataset* adalah mengalokasikan sebagian besar data untuk *training*, biasanya berkisar 70-90%, dan sisanya dialokasikan untuk *testing*, berkisar 10-30%. Proporsi ini dapat bervariasi tergantung pada kompleksitas masalah dan jumlah data yang tersedia. Dengan cara ini, model dapat dilatih dengan cukup data untuk memahami pola umum dan diuji pada subset yang terpisah untuk mengukur kemampuannya dalam melakukan prediksi pada data baru. Pembagian *dataset* ini membantu mencegah *overfitting* dan memastikan bahwa model memiliki generalisasi yang baik pada data yang belum pernah dilihat sebelumnya.

3.5. Pengembangan Model

Pengembangan model menggunakan KNIME melibatkan berbagai algoritma ML yang masing-masing memiliki node atau komponen yang spesifik dalam platform tersebut. Algoritma seperti *Random Forest*, *Logistic Regression*, *Naïve Bayes*, SVM, dan *Neural Networks* memiliki node yang berbeda untuk setiap tahap dalam proses pengembangan model. Proses melatih model pada subset data *training* dengan suatu algoritma melibatkan penyesuaian *hiperparameter*, *tuning parameter*, dan iterasi model untuk meningkatkan performa dan akurasi prediksi. Gambar 2 menampilkan 5 (lima) alur kerja model prediksi stunting pada anak menggunakan KNIME.





Gambar 2. Lima Alur Kerja Model Prediksi Stunting pada Anak menggunakan KNIME

Pada alur kerja, data stunting dibaca menggunakan *node excel reader*. Selanjutnya, *dataset* tersebut di filter untuk memisahkan data yang akan digunakan. Data yang digunakan mencakup informasi tentang stunting, berat badan anak, tinggi badan anak, umur ibu, dan tinggi badan ibu. Parameter ini memiliki sifat kualitatif dan kuantitatif yang beragam. Parameter-parameter tersebut diubah menjadi data numerik melalui node *normalizer* sehingga nilai-nilainya di normalisasi ke rentang nilai antara 0 hingga 1.

Selanjutnya, ke tahap pengujian algoritma. Setiap algoritma menggunakan node spesifik, yaitu, *Random Forest* menggunakan *node Random Forest Learner* dan *Predictor*, *Logistic Regression* menggunakan *node Logistic Regression Learner* dan *Predictor*, *Naïve Bayes* menggunakan *node Naïve Bayes Learner* dan *Predictor*, SVM menggunakan *node SVM Learner* dan *Predictor*, serta *Neural Networks* menggunakan *node RProp MLP Learner* dan *Predictor*. Hasil dari prediksi dapat diperoleh dengan menambahkan *node scorer* pada *node Predictor*.

3.6. Validasi Model

Model ML dilatih berulang kali dengan variasi parameter untuk mencapai nilai akurasi tertinggi. Hal ini membantu mengoptimalkan model serta mencegah *overfitting*. Proses ini bertujuan memastikan bahwa model yang dihasilkan memiliki performa optimal dan mampu memberikan prediksi yang akurat pada data baru. Tabel 2 memuat hasil perbandingan performa lima algoritma pada ML.

Tabel 2. Hasil Perbandingan Performa Lima Algoritma pada *Machine Learning*

Algoritma	Accuracy	*Recall*	*Precision*	*F1-Score*	Cohens's Kappa
Random Forest	98,29%	0,96	0,97	0,98	0,96
Logistic Regression	97,43%	0,94	0,96	0,97	0,95
Naïve Bayes	98,57%	0,96	0,97	0,99	0,97
SVM	55,13%	1,00	0,55	0,71	0,00
Neural Networks	94,01%	0,86	0,90	0,95	0,87

Hasil pengujian algoritma ML menunjukkan performa yang berbeda dari algoritma yang digunakan. Hasil pengujian tersebut dijabarkan sebagai berikut:

1. *Random Forest* memiliki tingkat akurasi sebesar 98,29%, yang menunjukkan tingkat keakuratan tinggi dalam melakukan klasifikasi. F1-Score yang tinggi (0,98) menunjukkan keseimbangan antara *presisi* (0,97) dan *recall* (0,96).
2. *Logistic Regression* juga memberikan hasil yang baik dengan akurasi sebesar 97,43%, serta *F1-Score*, presisi, dan recall yang mendekati hasil *Random Forest*.
3. *Naïve Bayes* menunjukkan hasil akurasi yang paling tinggi, mencapai 98,57%, serta *F1-Score* yang sangat tinggi (0,99). Hal ini menandakan performa yang baik dalam klasifikasi data.
4. SVM menunjukkan akurasi yang lebih rendah, hanya sebesar 55,13%, namun memiliki *recall* (1,00) yang sangat tinggi, menunjukkan kemampuan SVM dalam mengidentifikasi secara sempurna stunting anak, namun dengan nilai presisi yang rendah (0,55).
5. *Neural Networks* menunjukkan akurasi sebesar 94,01% dengan *F1-Score* yang tinggi (0,95), menunjukkan keseimbangan antara *presisi* (0,90) dan *recall* (0,86).

4. KESIMPULAN

Berdasarkan hasil pengujian maka, dapat disimpulkan hasil peneliitan ini diantaranya adalah, Hasil pengujian, dari penerapan lima algoritma ML, menghasilkan bahwa *Naïve Bayes* memiliki performa terbaik dengan akurasi tertinggi (98,57%) serta nilai-nilai *Recall*, *Precision*, *F1-Score*, dan *Cohens's Kappa* yang

sangat tinggi. Meskipun *Random Forest* juga memiliki performa yang sangat baik, *Naïve Bayes* memiliki nilai *F1-Score* yang sedikit lebih tinggi, menandakan kesesuaian yang lebih baik antara *presisi* dan *recall* dalam memprediksi stunting pada anak. Oleh karena itu, *Naïve Bayes* bisa dianggap sebagai model terbaik dari kelima algoritma yang diuji dalam kasus ini.

Dari hasil pengujian lima algoritma *Machine Learning*, dapat disimpulkan bahwa setiap algoritma memiliki performa yang berbeda dalam memprediksi stunting pada anak-anak. *Random Forest* menunjukkan akurasi yang tinggi (98,29%) dengan *F1-Score* yang seimbang antara *presisi* dan *recall*. *Logistic Regression* juga memberikan hasil yang baik dengan akurasi yang tinggi (97,43%), mendekati *performa Random Forest*. *Naïve Bayes* menonjol dengan akurasi tertinggi (98,57%) *dan F1-Score* sangat tinggi (0,99), menandakan performa yang luar biasa dalam klasifikasi data. SVM memiliki *recall* yang tinggi (1,00), menunjukkan kemampuan SVM dalam mengidentifikasi stunting, meskipun dengan presisi yang rendah (0,55). Sementara *Neural Networks* memiliki akurasi yang baik (94,01%) dengan F1-Score yang seimbang.

REFERENCES

[1] Eko, "149 Juta Anak di Dunia Alami Stunting Sebanyak 6,3 Juta di Indonesia, Wapres Minta Keluarga Prioritaskan Kebutuhan Gizi," *Direktorat Pendidikan Anak Usia Dini*, 2023. https://paudpedia.kemdikbud.go.id/berita/149-juta-anak-di-dunia-alami-stunting-sebanyak-63-juta-di-indonesia-wapres-minta-keluarga-prioritaskan-kebutuhan-gizi?do=MTY2NC01YjRhOGZkNA==&ix=MTEtYmJkNjQ3YzA= (accessed Jan. 06, 2024).

[2] U. M. Alam, "Perlu Terobosan dan Intervensi Tepat Sasaran Lintas Sektor untuk Atasi Stunting," *Kemenko PKM.go.id*, 2023. https://www.kemenkopmk.go.id/perlu-terobosan-dan-intervensi-tepat-sasaran-lintas-sektor-untuk-atasi-stunting (accessed Jan. 03, 2024).

[3] M. Rosyidah, Y. L. R. Dewi, and I. Qadrijati, "Effects of Stunting on Child Development: A Meta-Analysis," *J. Matern. Child Heal.*, vol. 6, no. 1, pp. 25–34, 2021, doi: 10.26911/thejmch.2021.06.01.03.

[4] N. F. Sahamony, R. Meliyani, and S. Idaman, "Analisa Resiko dampak Ekonomi pada saat Covid-19," *Media Bina Ilm.*, vol. 16, no. 19, pp. 64–71, 2021.

[5] M. Wahid and Mujib Rahman, "Rakornas 2023: Pastikan Prevalensi Stunting Turun Menjadi 14% Pada Tahun 2024," *Kementerian Sekretariat Negara RI*, 2023. https://stunting.go.id/rakornas-2023-pastikan-prevalensi-stunting-turun-menjadi-14-pada-tahun-2024/ (accessed Jan. 07, 2024).

[6] M. S. Haris, A. N. Khudori, and W. T. Kusuma, "Perbandingan Metode Supervised Machine Learning untuk Prediksi Prevalensi Stunting di Provisi Jawa Timur," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 7, p. 1571, 2022, doi: 10.25126/jtiik.2022976744.

[7] I. P. Putri, T. Terttiaavini, and N. Arminarahmah, "Comparative Analysis of Machine Learning Algorithms for Predicting Child Stunting," *MALCOM Indones. J. Mach. Learn. Comput. Sci. J.*, vol. 4, no. January, pp. 257–265, 2024.

[8] M. Amirudin and A. D. Wowor, "Analisis Perbandingan Klasifikasi Balita Beresiko Stunting Menggunakan Metode Support Vector Machine dan Decission Tree," in *Conference on Electrical Engineering, Informatics, Industrial Technology, and Creative Media 2023*, 2023, pp. 1–11.

[9] C. Scheffler and M. Hermanussen, "Stunting is the natural condition of human height," *Am. J. Hum. Biol.*, vol. 34, no. 5, pp. 1–13, 2022, doi: 10.1002/ajhb.23693.

[10] M. S. Islam, A. N. Zafar Ullah, S. Mainali, M. A. Imam, and M. I. Hasan, "Determinants of stunting during the first 1,000 days of life in Bangladesh: A review," *Food Sci. Nutr.*, vol. 8, no. 9, pp. 4685–4695, 2020, doi: 10.1002/fsn3.1795.

[11] R. Mburu, "Comparison of Elastic Net and Random Forest in identifying risk factors of stunting in children under rve years of age in Kenya," 2020.

[12] I. M. Apriliani, N. P. Purba, L. P. Dewanti, H. Herawati, and I. Faizal, "Stunting Risk Factors in Children Under Five in Indonesia: A Scoping Review," *Indones. J. Heal. Promot.*, vol. 5, no. 6, pp. 654–661, 2022.

[13] F. H. Bitew, C. S. Sparks, and S. H. Nyarko, "Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia," *Public Health Nutr.*, vol. 25, no. 2, pp. 269–280, 2022, doi: 10.1017/S1368980021004262.

[14] Mambang, F. D. Marleny, and M. Zulfadhilah, "Prediction of linear model on stunting prevalence with machine learning approach," *Bull. Electr. Eng. Informatics*, vol. 12, no. 1, pp. 483–492, 2023, doi: 10.11591/eei.v12i1.4028.

[15] A. Nugroho, H. L. H. S. Warnars, F. L. Gaol, and T. Matsuo, "Trend of Stunting Weight for Infants and Toddlers Using Decision Tree," *IAENG Int. J. Appl. Math.*, vol. 52, no. 1, 2022.

[16] E. R. Arumi, Sumarno Adi Subrata, and Anisa Rahmawati, "Implementation of Naïve bayes Method for Predictor Prevalence Level for Malnutrition Toddlers in Magelang City," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 2, pp. 201–207, 2023, doi: 10.29207/resti.v7i2.4438.

[17] Terttiaavini, T. S. Saputra, and A. Fitriani, "Classification of the final project utilized a modified naïve

bayes algorithm," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 5, 2020, doi: 10.30534/ijatcse/2020/48952020.

[18] A. W. M. Gaffar, Sugiarti, Dewi Widyawati, Andi Muhammad Kemai Arief Hidayat Paharuddin, and Andi Vania Anastasia, "Spatial Prediction of Stunting Incidents Prevalence Using Support Vector Regression Method," *Indones. J. Data Sci.*, vol. 4, no. 2, pp. 70–76, 2023, doi: 10.56705/ijodas.v4i2.68.

[19] M. M. Khudri, K. K. Rhee, M. S. Hasan, and K. Z. Ahsan, *Predicting nutritional status for women of childbearing age from their economic, health, and demographic features: A supervised machine learning approach*, vol. 18, no. 5 May. 2023. doi: 10.1371/journal.pone.0277738.

[20] A. R. Lakshminarayanan, B. Pavani, V Rajeswari, S. Parthasarathy, A. A. Azeez Khan, and K. Javubar Sathick, "Malnutrition Detection using Convolutional Neural Network," *Proc. 2021 IEEE 7th Int. Conf. Bio Signals, Images Instrumentation, ICBSII 2021*, no. March, 2021, doi: 10.1109/ICBSII51839.2021.9445188.

[21] A. Heryati, Erduandi, and Terttiaavini, "Penerapan Jaringan Saraf Tiruan Untuk Memprediksi Pencapaian Prestasi Mahasiswa," in *Konferensi Nasional Sistem Informasi 2018 STMIK Atma Luhur Pangkalpinang, 8 – 9 Maret 2018*, 2018, pp. 8–9.

[22] 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.

[23] 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.

[24] F. K. Alam, Y. Widyaningsih, and S. Nurrohmah, "Geographically weighted logistic regression modeling on stunting cases in Indonesia," *J. Phys. Conf. Ser.*, vol. 1722, no. 1, 2021, doi: 10.1088/1742-6596/1722/1/012085.

[25] H. Rianto and R. S. Wahono, "Resampling Logistic Regression untuk Penanganan Ketidakseimbangan Class pada Prediksi Cacat Software," *J. Softw. Eng.*, vol. 1, no. 1, pp. 46–53, 2015.

[26] P. Wahyu, S. Aji, R. Dijaya, F. Sains, and U. Muhammadiyah, "Prediksi Penyakit Stroke Menggunakan Metode Random Forest," *KESATRIA J. Penerapan Sist. Inf. (Komputer Manajemen)*, vol. 4, no. 4, pp. 916–924, 2023.

[27] P. P. Jantung, F. Handayani, K. S. Kusuma, H. L. Asbudi, R. G. Purnasiwi, and R. Kusuma, "Komparasi Support Vector Machine , Logistic Regression Dan Artificial Neural Network dalam," *JEPIN (Jurnal Edukasi dan Penelit. Inform.*, vol. 7, no. 3, pp. 329–334, 2021.

[28] U. Erdiansyah, A. I. Lubis, and K. Erwansyah, "Komparasi Metode K-Nearest Neighbor dan Random Forest Dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil," vol. 6, pp. 208–214, 2022, doi: 10.30865/mib.v6i1.3373.

[29] P. Butarbutar *et al.*, "Implementasi Jaringan Syaraf Tiruan Menggunakan Metode Elman Recurrent Neural Network Untuk Prediksi Penyakit Jantung Koroner," *Coding J. Komput. dan Apl.*, vol. 10, no. 01, pp. 103–113, 2022.

[30] A. Primajaya *et al.*, "Random Forest Algorithm for Prediction of Precipitation," *Indones. J. Artif. Intell. Data Min.*, vol. 1, no. 1, pp. 27–31, 2018.

[31] B. Sivasakthi and D. Selvanayagi, "A comparison of machine learning algorithms for osteoporosis prediction," *2022 1st Int. Conf. Electr. Electron. Inf. Commun. Technol. ICEEICT 2022*, vol. 7, pp. 432–439, 2022, doi: 10.1109/ICEEICT53079.2022.9768568.

[32] I. W. Prastika *et al.*, "Deteksi penyakit kulit wajah menggunakan tensorflow dengan metode convolutional neural network," vol. 4, no. 2, pp. 84–91, 2021.

[33] C. W. Cahyana and A. Nurlayli, "Analisis Performa Logistic Regression, Naïve Bayes, Dan Random Forest Sebagai Algoritma Pendeteksi Kanker Payudara," *Inser. Inf. Syst. Emerg. Technol.*, vol. 4, no. 1, pp. 51–64, 2023.

[34] C. Journal, I. Mubarog, A. Setyanto, H. Sismoro, and U. A. Yoyakarta, "Sistem Klasifikasi pada Penyakit Breast Cancer dengan Menggunakan Metode Naïve Bayes," *Citec J.*, vol. 6, no. 2, pp. 109–118, 2019.

[35] D. Febriansyah, "298 Anak di Kota Lubuklinggau Terdeteksi Stunting," *SINDOnews.com*, 2023. https://daerah.sindonews.com/read/1130961/720/298-anak-di-kota-lubuklinggau-terdeteksi-stunting-1687151172 (accessed Jan. 10, 2024).