BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terkait

Penelitian terdahulu dengan judul Klasifikasi Status Stunting pada Balita Menggunakan Metode Naif Bayes di Kota Madiun Berdasarkan Website. Penelitian tersebut menggunakan *Naive Bayes* untuk mengklasifikasikan status stunting pada balita di Kota Madiun. Tingkat model evaluasi menunjukkan rata-rata akurasi 58%, presisi 68%, dan *recall* 58%, berdasarkan uji *confusion matrix* dengan data pengujian sebesar 30% dan data pelatihan 70%. Penelitian bertujuan untuk menghasilkan sistem yang dapat klasifikasi berbasis website untuk membantu tenaga kesehatan dalam mengidentifikasi status stunting pada balita [7].

Selanjutnya penelitian dengan judul Penerapan Algoritma K-Means dalam Klasterisasi Kasus Stunting Balita Desa Tegalwangi. Penelitian tersebut menggunakan K-Means untuk mengelompokkan balita menjadi dua *cluster* berdasarkan usia, tinggi badan, dan berat badan. Tingkat model evaluasi memanfaatkan indeks Davies-Bouldin (DBI) menghasilkan nilai optimal sebesar 0.007 pada cluster terbaik. Penelitian ini bertujuan untuk menghasilkan informasi status balita, yaitu normal atau stunting, guna membantu petugas Posyandu dalam mengolah data dan mendukung intervensi kesehatan yang lebih efektif [8].

Selanjutnya pada penelitian dengan judul Implementasi Algoritma *Naïve Bayes* dan K-NN untuk Diagnosis Stunting pada Anak. Dalam penelitian ini, algoritma Naive Bayes dan K-Nearest Neighbors (K-NN) digunakan untuk mengklasifikasikan status gizi anak berdasarkan data antropometri. Parameter yang digunakan meliputi usia, berat badan, dan tinggi badan. Model *Naïve Bayes* menghasilkan akurasi 71% dengan F1-score 73%. Tujuan penelitian ini adalah untuk membandingkan efektivitas dua algoritma klasifikasi dalam membantu petugas kesehatan mendiagnosis kasus stunting dengan lebih cepat dan akurat [9].

Kemudian penelitian dengan judul Analisis Prediksi Stunting Menggunakan *Naïve Bayes* dan Penyeimbangan Data ADASYN Penelitian ini membandingkan algoritma *Naïve Bayes Gaussian* dan *Bernoulli* dalam klasifikasi data stunting pada

balita di Kabupaten Bekasi. Data tidak seimbang diatasi menggunakan metode *Adaptive Synthetic Sampling* (ADASYN), dan evaluasi model *menggunakan Stratified K-Fold Cross Validation*. Hasil menunjukkan akurasi 67,83% pada model *Bernoulli*. Tujuan penelitian ini adalah untuk mengevaluasi dampak teknik penyeimbangan data terhadap performa klasifikasi, khususnya dalam konteks prediksi stunting [10].

Selanjutnya pada penelitian dengan judul Integrasi *Naïve Bayes* dan SMOTE untuk Klasifikasi Status Kredit sebagai Studi Pendukung Meskipun tidak langsung membahas kasus stunting, penelitian mengombinasikan algoritma *Naïve Bayes* dan metode *Synthetic Minority Over-sampling Technique* (SMOTE) pada klasifikasi status kredit pelanggan koperasi. Model menunjukkan peningkatan klasifikasi dari 818 menjadi 1131 data setelah penerapan SMOTE. Studi ini bermanfaat dalam menunjukkan efektivitas SMOTE dalam mengatasi ketidakseimbangan data, yang dapat diterapkan pada domain stunting dengan karakteristik data serupa [11].

Kemudian penelitian dengan judul Implementasi *Naïve Bayes* dan SVM untuk Klasifikasi Stunting di Kabupaten Lima Puluh Kota. Penelitian mengaplikasikan algoritma *Naïve Bayes* dan *Support Vector Machine* untuk mengklasifikasikan status gizi balita berdasarkan 2018 data survei. Meskipun tidak dijelaskan secara eksplisit akurasi yang diperoleh, penelitian ini digunakan untuk mengevaluasi efektivitas program pencegahan stunting. Tujuan utamanya adalah menghasilkan sistem klasifikasi berbasis data untuk mendukung program intervensi gizi di tingkat daerah [12].

2.2 Landasan Teori

2.2.1 Stunting

Stunting adalah kondisi gagal tumbuh pada anak yang disebabkan oleh kekurangan gizi kronis dalam jangka waktu yang lama, terutama pada 1.000 hari pertama, yang mencakup kehamilan sampai usia dua tahun. Anak yang mengalami stunting memiliki tinggi badan yang lebih pendek dari rata-rata, dan hal ini bisa memengaruhi perkembangan fisik, kognitif, dan sosial mereka.

Faktor utama penyebab stunting meliputi kekurangan asupan gizi yang cukup, infeksi berulang, serta faktor lingkungan yang tidak mendukung pertumbuhan optimal, seperti sanitasi yang buruk dan akses terbatas terhadap perawatan kesehatan yang berkualitas [13].

Stunting berdampak pada kesehatan fisik anak, tetapi juga memiliki konsekuensi pada jangka panjang dalam kualitas hidup mereka. Anak yang stunting berisiko tinggi mengalami keterlambatan dalam perkembangan otak, kemampuan belajar, dan daya tahan tubuh [13]. Selain itu, stunting dapat mempengaruhi produktivitas dan potensi ekonomi di masa depan, karena mereka mungkin kesulitan dalam mencapai kecakapan kerja yang optimal. Oleh karena itu, Untuk menangani stunting secara efektif, diperlukan pendekatan yang holistik yang mencakup peningkatan gizi dan sanitasi serta pelatihan masyarakat [14].

Stunting pada anak dapat diklasifikasikan berdasarkan standar WHO menggunakan pengukuran tinggi badan (TB/U) menggunakan Z-score. Anak dikategorikan normal memiliki nilai Z-score ≥ -2 SD, artinya tinggi badannya sesuai standar usia [15]. Anak masuk kategori stunting jika Z-score-nya < -2 SD, menunjukkan tinggi badan di bawah standar akibat gagal tumbuh kronis. Sementara itu, anak dengan Z-score < -3 SD tergolong severely stunted (stunting berat), di mana tinggi badannya sangat jauh di bawah normal, mengindikasikan gangguan pertumbuhan yang parah. Sebagai contoh, pada balita laki-laki dengan usia 2 tahun, dengan tinggi badan normal adalah adalah ≥ 85 cm, stunting di bawah 85 cm (misalnya 80 cm), dan severely stunted di bawah 80 cm (misalnya 75 cm). Kondisi stunting, terutama severely stunted, berdampak serius pada perkembangan fisik, kognitif, dan kesehatan jangka panjang. Oleh karena itu, pencegahan melalui intervensi gizi, sanitasi, dan layanan kesehatan pada 1.000 hari pertama kehidupan sangat penting untuk menghindari dampak permanen [16].

Tabel 2. 1 Indikator Stunting WHO

Kategori	Indikator	Kriteria Z-Score	Dekripsi
Normal	Tinggi Badan/Usia (TB/U)	Z-score ≥ -2 SD	Tinggi badan sesuai standar usia
Stunting	Tinggi Badan/Usia (TB/U)	Z-score < -2 SD	Tinggi badan di bawah standar akibat gagal tumbuh kronis
Severely Stunted	Tinggi Badan/Usia (TB/U)	Z-score < -3 SD	Tinggi badan sangat jauh di bawah normal, gangguan pertumbuhan parah

2.2.2 *Naive Bayes Gaussian*

Metode klasifikasi yang termasuk dalam kategori probabilistik. Metode ini berdasarkan pada *Teorema Bayes*, yang digunakan untuk menghitung nilai probabilitas pada suatu kelas berdasarkan atribut-atribut yang digunakan [17]. Prinsip utama dari Naive Bayes adalah mengasumsikan bahwa setiap atribut dalam dataset bersifat independen (*naive assumption*), meskipun dalam kenyataannya atribut-atribut tersebut sering kali saling bergantung. Dalam varian *Gaussian*, asumsi independensi atribut tetap diterapkan, namun distribusi data untuk setiap kelas diasumsikan mengikuti distribusi normal (*Gaussian*). Asumsi ini memungkinkan model untuk melakukan klasifikasi secara efisien meskipun data yang digunakan tidak selalu memenuhi asumsi distribusi normal secara sempurna.

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)} \quad \text{.....(1)}$$

Keterangan:

X : Data input atau vektor fitur (atribut)

H : Hipotesis kelas (class label)

P(*H*) : Prior probabilitas, yaitu probabilitas awal dari kelas *H*

P(X|*H*) : Likelihood, yaitu probabilitas data X diberikan bahwa kelasnya adalah *H*

P(X) : Evidence, yaitu probabalitas keseluruhan data

P(H|X) : Posteriori probabilitas, yaitu probabilitas suatu data berasal dari kelas *H* setelah memperhatikan X

Karena *P*(X) konstan untuk semua kelas saat klasifikasi, maka dapat diabaikan dalam perhitungan perbandingan antar kelas. Sehingga disederhanakan:

$$P(H \mid X) \propto P(X \mid H) \cdot P(H) \quad \text{.....(2)}$$

Untuk Naive Bayes Gaussian, distribusi dari setiap fitur $x_1$diasumsikan mengikuti distribusi normal menggunakan parameter rata-rata (μ) dan simpangan baku (σ). Fungsi distribusi Gaussian dinyatakan dengan:

$$P(x_i \mid H) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \quad \text{.....(3)}$$

Keterangan:

x_1 : Nilai fitur ke- i

μ : Rata-rata nilai fitur ke- i pada kelas *H*

σ : Simpangan baku fitur ke- i pada kelas *H*

Karena semua fitur dianggap independen, probabilitas gabungan $P(X|H)$ dihitung sebagai hasil kali dari seluruh probabilitas fitur terhadap kelas tersebut:

$$P(X \mid H) = \prod_{i=1}^{n} P(x_i \mid H) \quad \text{.....(4)}$$

Sehingga, probabilitas posterior menjadi:

$$P(H \mid X) \propto P(H) \cdot \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right) \right) \quad \text{.....(5)}$$

Pada akhirnya, data X diklasifikasikan ke dalam kelas H yang memberikan nilai $P\ (H|X)$ paling tinggi.

Penerapan Naive Bayes Gaussian dalam klasifikasi data sangat berguna dalam berbagai bidang, terutama dalam analisis teks, deteksi spam, dan pengenalan pola. Kelebihan utama dari metode ini adalah kesederhanaan, kecepatan, dan kemampuannya untuk menangani data jumlah atribut yang besar. Mengasumsikan independensi antara atribut yang sering kali tidak realistis, Naive Bayes tetap dapat memberikan hasil yang cukup akurat dalam banyak kasus. Model ini juga sangat efisien dalam menangani data yang hilang atau tidak lengkap. Dengan pendekatan ini, prediksi kelas dapat dilakukan dengan memaksimalkan fungsi kemungkinan berdasarkan distribusi normal yang diasumsikan, membuatnya menjadi alat yang berguna untuk klasifikasi di banyak aplikasi praktis. Kekurangan Naive Bayes Gaussian adalah asumsi independensi antar fitur yang jarang terpenuhi, sensitif terhadap distribusi data yang tidak normal, dan kurang mampu menangani hubungan non-linear yang kompleks [18].

2.2.3 Metrik Evaluasi

Model *Naive Bayes Gaussian* adalah salah satu algoritma klasifikasi yang mengasumsikan bahwa setiap fitur dalam dataset mengikuti distribusi normal (*Gaussian*). Untuk menilai performa model ini, beberapa metrik evaluasi

digunakan, seperti akurasi, *recall, presisi*, F1-*Score*, serta kurva ROC dan AUC (*Area Under Curve*).

Akurasi untuk mengukur persentase prediksi benar dibandingkan dengan total prediksi. Rumusnya ditunjukkan dalam Persamaan (6):

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad \text{.....(6)}$$

Di mana Number of Correct Predictions adalah jumlah data yang diprediksi dengan benar, sedangkan Total Number of Predictions adalah seluruh data yang diuji.

Selain akurasi, presisi dan recall juga penting, terutama ketika dataset tidak seimbang. Presisi (Persamaan 7) untuk menilai akurat model dalam memprediksi kelas positif, sementara recall (Persamaan 8) mengevaluasi positif yang berhasil diidentifikasi model.

$$Precision = \frac{TP}{(TP + FP)} \quad \text{.....(7)}$$

$$Recall = \frac{TP}{(TP + Fn)} \quad \text{.....(8)}$$

Dalam kedua rumus ini:

1. TP (*True Positive*) = Jumlah sampel positif yang diprediksi benar.
2. FP (*False Positive*) = Jumlah sampel negatif yang salah diprediksi sebagai positif.
3. FN (*False Negative*) = Jumlah sampel positif yang salah diprediksi sebagai negatif.

Karena presisi dan recall sering kali bertolak belakang, F1-Score (Persamaan 9) digunakan sebagai metrik yang menyeimbangkan keduanya:

$$F1 - Score\ = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \text{.....(9)}$$

Selain itu, kurva ROC dan AUC memberikan gambaran kinerja model dalam membedakan kelas pada berbagai ambang batas klasifikasi. Dengan mempertimbangkan berbagai metrik ini, kita dapat memahami kelebihan dan kelemahan model Naive Bayes Gaussian secara lebih komprehensif dalam berbagai konteks aplikasi.

2.2.4 Machine Learning

Machine learning bagian kecerdasan buatan (Artificial Intelligence) berfokus pada pengembangan sistem yang dapat belajar dan membuat keputusan berdasarkan data. Dalam machine learning, komputer tidak diprogram secara eksplisit, melainkan dilatih untuk mengenali pola dari data yang diberikan. Salah satu dalam machine learning adalah supervised learning, di mana model dilatih dengan data yang memiliki label. Metode Naive Bayes Gaussian termasuk dalam kategori ini dan sering digunakan untuk tugas klasifikasi. Dengan pendekatan ini, machine learning memberikan kontribusi penting dalam pengolahan data skala besar dan pengambilan keputusan berbasis data.

3.3.1 *Supervised Learning*

Pada jenis *Machine learning* ini, model dilatih dengan menggunakan data berlabel. Model belajar dari pasangan *input-output* untuk memprediksi nilai atau klasifikasi baru. Contohnya adalah *Naive Bayes Gaussian* untuk klasifikasi probabilistik berdasarkan distribusi *Gaussian*, serta *regresi linier* untuk prediksi nilai kontinu.

3.3.2 *Unsupervised Learning*

Dalam jenis *Machine learning* ini, data yang digunakan tidak memiliki label. Tujuan utamanya adalah menemukan struktur atau pola tersembunyi dalam data. Contohnya adalah algoritma *clustering* seperti K-Means dan analisis komponen utama (PCA).

3.3.3 *Semi-Supervised Learning*

Jenis *Machine learning* ini menggabungkan sejumlah kecil data yang berlabel dengan sejumlah besar data yang tanpa label. digunakan ketika saat pelabelan data membutuhkan biaya atau waktu .

3.3.4 Reinforcement Learning

Machine learning jenis ini melibatkan agen yang belajar melalui interaksi lingkungan. Agen menerima umpan balik *reward* atau *punishment* dan bertujuan memaksimalkan total *reward* dalam jangka panjang.

2.2.5 *Python*

Python salah satu bahasa pemrograman populer dan banyak digunakan dalam bidang analisis data, statistika, dan pengembangan model prediktif berkat sintaksisnya yang sederhana dan mudah dipahami. Bahasa ini didukung oleh berbagai pustaka canggih yang memudahkan proses pengolahan data dan pembangunan model machine learning. Untuk kebutuhan manipulasi dan pembersihan data, Python menyediakan library Pandas yang powerful, sedangkan komputasi numerik dapat dilakukan secara efisien menggunakan NumPy. Dalam hal visualisasi data, Python menawarkan Matplotlib dan Seaborn yang memungkinkan pembuatan grafik dan plot informatif. Salah satu library terpenting adalah Scikit-learn yang menyediakan berbagai algoritma machine learning siap pakai, termasuk implementasi Naive Bayes Gaussian (GaussianNB) untuk keperluan klasifikasi probabilistik. Dalam konteks penelitian ini, Python berperan sebagai alat utama untuk membangun model Naive Bayes Gaussian, melakukan praproses data historis seperti normalisasi fitur kontinu, serta mengevaluasi performa model melalui berbagai metrik klasifikasi seperti akurasi, presisi, dan recall untuk memastikan keandalan prediksi yang dihasilkan [19].

2.2.6 Klasifikasi

Klasifikasi adalah proses pengelompokan atau pengategorian suatu objek, data, atau informasi berdasarkan kriteria atau karakteristik tertentu yang serupa. Tujuannya adalah untuk menciptakan susunan yang sistematis dan terstruktur sehingga memudahkan identifikasi, analisis, dan pengelolaan. Proses ini melibatkan pemilihan ciri-ciri pembeda yang relevan untuk membagi suatu kesatuan ke dalam kelompok-kelompok yang lebih spesifik dan teratur. Klasifikasi digunakan dalam berbagai bidang untuk meningkatkan efisiensi dan kejelasan dalam pengorganisasian [20].

2.2.7 Synthetic Minority Over-sampling Technique (SMOTE)

Teknik Pengambilan Sampel Berlebih Minoritas Sintetis (Synthetic Minority Over-sampling Technique (SMOTE) adalah metode yang banyak digunakan untuk mengatasi ketidakseimbangan pada kelas. Prosedur ini mensintesis sampel baru dari kelas minoritas dengan menggabungkan data baru, sehingga menyeimbangkan kumpulan data dengan mengambil sampel berlebih dari kelas minoritas. kelas minoritas. Oversampling, dalam metode SMOTE, melibatkan pengambilan sampel dari kelas minoritas dalam dataset dan menemukan tetangga terdekat dari setiap contoh tersebut. Hal ini untuk mencegah masalah overfitting disebabkan jika pengulangan hanya dari contoh minoritas yang tersedia. Kelebihan SMOTE adalah mampu menyeimbangkan distribusi kelas dan meningkatkan kinerja model pada data tidak seimbang tanpa menimbulkan overfitting seperti oversampling biasa. Kekurangannya, SMOTE bisa menambahkan noise jika data minoritas berkualitas rendah dan berisiko menurunkan akurasi jika sampel sintetis terlalu dekat dengan kelas mayoritas. [21].