BAB I
PENDAHULUAN

1.1 Latar Belakang

Status gizi balita (bayi dibawah lima tahun) merupakan indikator utama kesehatan masyarakat yang mencerminkan keseimbangan antara asupan zat gizi dan kebutuhan tubuh untuk pertumbuhan serta perkembangan yang optimal (Burgard et al., 2024). Kekurangan gizi pada balita, terutama dalam bentuk *stunting*, wasting, dan underweight, masih menjadi permasalahan gizi secara global (Danso & Appiah, 2023; Dassie et al., 2024). Kurangnya gizi pada balita juga dapat meningkatkan risiko morbiditas dan mortalitas (Hossain et al., 2023; Vijay & Patel, 2024), dimana sekitar 45% total kematian balita disebabkan oleh gizi buruk (Khura et al., 2023). *Stunting* pada balita masih menjadi isu krusial dalam upaya pembangunan kesehatan nasional di Indonesia. Kondisi ini tidak hanya menghambat pertumbuhan fisik, tetapi juga berisiko menurunkan kemampuan kognitif serta produktivitas individu di masa depan (Rahut et al., 2024). Indonesia termasuk dalam kategori negara dengan prevalensi *stunting* yang masih cukup tinggi. Berdasarkan data Profil Kesehatan Indonesia, prevalensi *stunting* secara nasional tercatat mencapai 24,4% pada tahun 2021, kemudian turun menjadi 21,6% pada tahun 2022, dan turun 1% menjadi 21,5% pada tahun 2023. Meskipun tren penurunan ini menunjukkan kemajuan, angka tersebut masih berada di atas ambang batas yang ditetapkan oleh Organisasi Kesehatan Dunia (WHO), yaitu di bawah 20%. Pada tingkat daerah, berdasarkan Profil Kesehatan Indonesia menunjukkan prevalensi *stunting* di Provinsi Sulawesi tenggara pada tahun 2021 mencapai 30,2%, tahun 2022 sebesar 27,7%, dan tahun 2023 naik menjadi 30% (Kementrian Kesehatan RI, 2024).

Dengan meningkatnya akses terhadap data kesehatan dan kemajuan teknologi informasi, pendekatan *machine learning*, khususnya dalam metode *supervised learning*, telah membuka potensi besar untuk menerapkan model prediktif dalam berbagai bidang, termasuk di bidang kesehatan masyarakat. *Supervised learning* terbagi menjadi dua pendekatan utama, yaitu regresi dan klasifikasi, yang berfungsi penting untuk menghasilkan prediksi berdasarkan variabel historis atau prediktor (Changpetch et al., 2021; Govindarajan et al., 2022).

Namun, pada banyak kasus kesehatan masyarakat, seperti klasifikasi status gizi balita, sering ditemukan masalah ketidakseimbangan kelas (*class imbalance*) antara kelompok anak *stunting* dan normal. Kondisi ini dapat menyebabkan model klasifikasi menjadi bias terhadap kelas mayoritas sehingga menurunkan akurasi prediksi (Husain et al., 2025; Owusu-adjei et al., 2023). Salah satu metode yang efektif untuk mengatasi permasalahan tersebut adalah *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE bekerja dengan cara menghasilkan sampel sintetis baru dari kelas minoritas melalui interpolasi antara suatu titik data minoritas dan *k-nearest neighbors*-nya. Dengan demikian, distribusi kelas menjadi lebih seimbang tanpa menghapus data mayoritas, sehingga model pembelajaran mesin dapat mengenali pola minoritas dengan lebih baik (Husain et al., 2025; Wang et al., 2023; Yang et al., 2022). Sejumlah penelitian terkini menunjukkan bahwa penerapan SMOTE mampu meningkatkan akurasi model dan

kestabilan prediksi, terutama pada data kesehatan yang memiliki distribusi tidak seimbang (Ali, 2025; Elreedy et al., 2024; Husain et al., 2025; Matharaarachchi et al., 2024).

Dalam pendekatan regresi, model nonparametrik seperti fungsi *spline* banyak digunakan karena memiliki keandalan dalam menangkap hubungan *nonlinier* antar variabel, dimana hal ini sulit ditangani oleh model parametrik (Schuster et al., 2022). Beberapa varian spline seperti *smoothing spline*, *penalized spline*, dan *truncated spline* telah dikembangkan untuk memberikan fleksibilitas yang lebih tinggi saat menghadapi pola data yang berubah pada interval tertentu (Arifin et al., 2023; Islamiyati et al., 2024; Lestari et al., 2023). Sedangkan dalam klasifikasi, metode *naïve bayes* merupakan salah satu algoritma yang sederhana dalam namun tetap efektif. Dengan prinsip dasar probabilistik dan asumsi independensi antar fitur, algoritma ini menawarkan solusi klasifikasi yang cepat dan efisien (Febrian et al., 2023; Kim & Lee, 2022). Klasifikasi *naïve bayes* memiliki banyak estimator yang dapat digunakan seperti *gaussian*, *bernoulli*, dan *multinomial* yang telah banyak diterapkan pada data-data kesehatan (Mustamin et al., 2023). Namun demikian, algoritma ini memiliki keterbatasan, terutama dalam menangani fitur kontinu dan pola hubungan non-linier (Changpetch et al., 2021) yang umum ditemukan pada data antropometri anak, seperti berat badan, tinggi badan, dan usia untuk menentukan status gizi anak. Untuk mengatasi keterbatasan tersebut, diperlukan integrasi fungsi spline ke dalam model *naïve bayes*. Dengan memanfaatkan fungsi *spline* dalam proses estimasi distribusi fitur, model *naïve bayes* dapat diperluas sehingga lebih adaptif dalam menangkap pola distribusi data yang tidak mengikuti asumsi distribusi normal atau hubungan linier.

Penelitian mengenai klasifikasi dalam bidang kesehatan telah banyak dilakukan dengan memanfaatkan pendekatan pembelajaran mesin (*machine learning*) yang mampu menangani data berskala besar dan kompleks. Mustamin et al. (2023) meneliti *Classification of Maternal Health Risk* menggunakan metode *Naïve Bayes* dengan tiga varian model, yaitu *Gaussian, Multinomial, dan Bernoulli*. Penelitian tersebut memanfaatkan data kesehatan ibu hamil yang meliputi usia, tekanan darah sistolik, tekanan darah diastolik, kadar gula darah, suhu tubuh, dan denyut jantung. Data risiko ibu hamil dikategorikan menjadi tiga kelas, yaitu risiko rendah, sedang, dan tinggi. Hasil penelitian menunjukkan bahwa model *Naïve Bayes* dengan estimator *Multinomial* dan *Bernoulli* memberikan akurasi terbaik sebesar 84,8%, sedangkan *Gaussian Naïve Bayes* memiliki akurasi 82,6%.

Sementara itu, Fadil (2025) mengembangkan metode *Support Vector Machine* (SVM) dengan pendekatan fungsi *spline* dalam mengklasifikasikan status gizi balita di Kabupaten Gowa tahun 2022. Penelitian ini bertujuan untuk menangani hubungan prediktor–respon yang bersifat kompleks dan non-linier, di mana *kernel* standar pada SVM kurang optimal. Dengan memanfaatkan fungsi *spline truncated* dan pemilihan titik knot optimum menggunakan *Generalized Cross Validation* (GCV), model yang dihasilkan mampu meningkatkan fleksibilitas dalam menangkap pola data. Hasil pengujian menunjukkan akurasi klasifikasi sebesar 90,06% dengan tingkat kesalahan prediksi (*Apparent Error Rate / APER*) sebesar 9,94%.

Berdasarkan hal ini, peneliti bermaksud untuk melakukan penelitian dengan mengintegrasikan fungsi *Spline* ke dalam model klasifikasi *Naïve Bayes* dalam rangka meningkatkan akurasi klasifikasi status risiko *stunting* pada anak. Diharapkan, kombinasi kedua metode ini mampu menangkap hubungan non-linier yang kompleks antar variabel serta memberikan model prediksi yang lebih akurat.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana teknik klasifikasi melalui metode *Naïve Bayes* berbasis fungsi *Spline*?
2. Bagaimana akurasi klasifikasi data *stunting* dengan metode *Naïve Bayes* berbasis fungsi *Spline*?
3. Adakah peningkatan kinerja klasifikasi data *stunting* dengan metode *Naïve Bayes* berbasis fungsi *Spline*?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah :

1. Menganalisis teknik klasifikasi melalui metode *Naïve Bayes berbasis* fungsi *Spline*
2. Mengukur dan mengevaluasi akurasi klasifikasi data *stunting* dengan metode *Naïve Bayes* berbasis fungsi *Spline*
3. Mengukur peningkatan kinerja klasifikasi data *stunting* dengan metode *Naïve Bayes* berbasis fungsi *Spline*.

1.4 Manfaat Penelitian

Penelitian ini diharapkan memberikan manfaat sebagai berikut:

1. Memberikan kontribusi dalam pengembangan metode klasifikasi berbasis *Naïve Bayes* berbasis fungsi *spline*.
2. Memberikan model prediktif yang akurat serta mampu menangkap hubungan non-linier pada kasus *stunting*.

1.5 Teori

Sub-bab teori ini membahas mengenai beberapa teori sebagai pokok pembahasan yang membahas terkait s*tunting* dan metode *Naïve Bayes Classifier* dengan fungsi *spline*.

1.5.1 Regresi Nonparametrik

Regresi nonparametrik adalah model yang dikembangkan dari regresi parametrik. Pendekatan nonparametrik menawarkan fleksibilitas yang lebih tinggi dibandingkan pendekatan parametrik, terutama dalam konteks pemodelan data baru yang belum banyak diketahui karakteristiknya (Mahmoud, 2021). Dalam banyak kasus, informasi mengenai bentuk fungsional model yang sesuai sangat terbatas, sehingga menyulitkan penggunaan pendekatan parametrik yang bergantung pada asumsi bentuk fungsi tertentu. Meskipun tersedia sejumlah alat bantu heuristik, seperti *diagnostic plots*, untuk mengevaluasi dan menentukan bentuk model yang tepat, pendekatan nonparametrik memberikan keuntungan karena mampu menangani proses pencarian bentuk model

tersebut secara otomatis melalui struktur model yang adaptif (J. Faraway, 2016). Bentuk persamaan umum regresi nonparametrik adalah sebagai berikut :

$$y_i = f(x_i) + \varepsilon_i \quad (1)$$

dimana y_i menyatakan nilai respon pada observasi ke-i, sedangkan x_i merupakan nilai dari variabel penjelas atau prediktor. Fungsi $f(x_i)$ merepresentasikan bentuk hubungan sistematis antara x_i dan y_i yang tidak diasumsikan memiliki bentuk tertentu sebagaimana pada regresi parametrik, melainkan dipelajari secara fleksibel dari data yang tersedia. kemudian ε_i adalah komponen galat atau kesalahan acak yang mencerminkan deviasi antara nilai aktual y_i dengan nilai yang diprediksi oleh fungsi $f(x_i)$.

1.5.2 Spline truncated

Pendekatan *spline* merupakan salah satu metode dalam regresi nonparametrik yang memiliki kelebihan dari sisi interpretasi, baik secara visual maupun statistik. Teknik ini sangat cocok digunakan untuk data yang menunjukkan pola perubahan yang relatif halus (*smooth*). Selain itu, *spline* memiliki kemampuan yang baik dalam menangkap variasi perilaku data yang terjadi pada sub-interval tertentu dalam ruang pengamatan (Ramli et al., 2020). Metode ini telah banyak dimanfaatkan dan dikembangkan dalam berbagai studi. Dalam penelitian ini, digunakan pendekatan *spline truncated* untuk membangun model terhadap data *stunting*. Estimator *spline truncated* sendiri merupakan salah satu bentuk regresi nonparametrik yang didasarkan pada model polinomial yang bersifat tersegmentasi (Sifriyani et al., 2023). jika orde *spline* dinyatakan sebagai q dan jumlah titik simpul (τ) adalah m, maka model *spline* dapat dituliskan sebagai berikut (Islamiyati et al., 2023):

$$f(x_i) = \beta_0 + \sum_{j=1}^{q} \beta_j x_i^j + \sum_{k=1}^{m} \beta_{q+k}(x_i - \tau_k)_+^q \quad (2)$$

Di mana x_i adalah variabel prediktor ke-i, β_0 merupakan intersep, β_j adalah koefisien dari regresi *spline truncated*, τ_k adalah titik simpul dengan $k = 1,2,\dots,m$. dan $(x_i - \tau_k)_+^q$ adalah fungsi polinomial *truncated* yang didefinisikan sebagai berikut :

$$(x_i - \tau_k)_+^q = \begin{cases} (x_i - \tau_k)^q & , \text{jika } x_i > \tau_k \\ 0 & , \text{jika } x_i \le \tau_k \end{cases}$$

Model pada persamaan 2 dapat dituliskan dalam bentuk berikut :

$$f(x_i) = \beta_0 + \beta_1 x_i + \cdots + \beta_q x_i^q + \beta_{q+1}(x_i - \tau_1)_+^q + \cdots + \beta_{q+2}(x_i - \tau_2)_+^q + \cdots + \beta_{q+m}(x_i - \tau_m)_+^q \quad (3)$$

Selanjutnya, dapat dinyatakan dalam bentuk matriks seperti pada persamaan 4 berikut :

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad (4)$$

dimana :

$$\boldsymbol{y} = [\, y_1,\, y_2,\, y_3 \dots y_n]^T$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^q & (x_1-\tau_1)_+^q & \cdots & (x_1-\tau_m)_+^q \\ 1 & x_2 & x_2^2 & \cdots & x_2^q & (x_2-\tau_1)_+^q & \cdots & (x_2-\tau_m)_+^q \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^q & (x_n-\tau_1)_+^q & \cdots & (x_n-\tau_m)_+^q \end{bmatrix}$$

$$\boldsymbol{\beta} = \left[\beta_0, \beta_1, \ldots \beta_q \;\; \beta_{q+1} \ldots \; \beta_{q+m}\right]^T$$

$$\boldsymbol{\varepsilon} = \left[\varepsilon_0, \varepsilon_1, \ldots \varepsilon_n\right]^T$$

1.5.3 *Knot* Optimal

Salah satu pendekatan yang efektif dalam menentukan titik simpul (*knot point*) yang optimal adalah melalui metode *Generalized Cross Validation* (GCV). Teknik ini mengevaluasi kualitas estimasi model berdasarkan keseimbangan antara kesesuaian model dan kompleksitasnya (Arifin et al., 2023). Persamaan GCV sebagai berikut :

$$\text{GCV}(k_1, k_2, \ldots, k_r) = \frac{\text{MSE}(k_1,k_2,\ldots,k_r)}{\left[n - \text{trace}\left(\mathbf{I} - \mathbf{A}(k_1,k_2,\ldots,k_r)\right)\right]^2} \qquad (5)$$

dengan nilai rata-rata kuadrat error didefinisikan sebagai :

$$\text{MSE}(k_1, k_2, \ldots, k_r) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (6)$$

Dalam hal ini, k merepresentasikan titik knot, $\boldsymbol{A} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ dan $\boldsymbol{I}$ merupakan matriks identitas. Nilai GCV terkecil menunjukkan konfigurasi titik *knot* yang memberikan kinerja terbaik untuk model (Islamiyati et al., 2023).

1.5.4 *Synthetic Minority Oversampling Technique* (SMOTE)

Dalam *machine learning*, ketidakseimbangan kelas (*class imbalance*) merupakan kondisi di mana proporsi jumlah observasi pada satu kelas jauh lebih besar dibandingkan kelas lainnya (Elreedy et al., 2024). Ketidakseimbangan ini sering muncul pada data kesehatan, seperti kasus gizi buruk, penyakit langka, atau deteksi risiko kesehatan anak, di mana jumlah individu dengan kondisi normal jauh lebih banyak dibandingkan dengan kelompok berisiko. Situasi ini dapat menyebabkan model klasifikasi cenderung bias terhadap kelas mayoritas, karena proses pelatihan lebih banyak mengenali pola dari kelompok tersebut. Akibatnya, model memiliki akurasi tinggi secara keseluruhan tetapi gagal mengidentifikasi kelas minoritas yang justru menjadi fokus analisis (Ali, 2025).

Untuk mengatasi masalah ini, dikembangkan berbagai pendekatan penyeimbangan data pada tahap pra-pemrosesan (*data-level methods*), salah satunya adalah teknik *Oversampling* melalui *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE pertama kali diperkenalkan oleh Chawla dkk pada tahun 2002 sebagai metode untuk menyeimbangkan distribusi kelas melalui pembangkitan data sintetis dari kelas minoritas (Elreedy et al., 2024). Tidak seperti *random oversampling* yang hanya menduplikasi data minoritas, SMOTE membentuk observasi baru melalui interpolasi linier antara satu titik data minoritas dan k-tetangga terdekatnya (*k-nearest neighbors*). Secara matematis, proses pembentukan sampel sintetis dijelaskan sebagai:

$$x_{new} = x_i + \alpha(x_j - x_i) \quad (7)$$

dimana x_i merupakan observasi dari kelas minoritas, x_j adalah salah satu dari k tetangga terdekatnya (*k-nearest neighbors*), dan $\alpha \in \{0,1\}$ merupakan bilangan acak yang mengatur posisi interpolasi. Proses ini diulangi hingga jumlah data minoritas mendekati atau sama dengan jumlah data mayoritas.

Pendekatan ini memperluas ruang representasi kelas minoritas di dalam ruang fitur, sehingga model pembelajaran dapat mengenali batas keputusan (*decision boundary*) dengan lebih seimbang. Dalam konteks prediksi kesehatan, misalnya klasifikasi status gizi atau deteksi risiko penyakit, SMOTE membantu model mengenali pola unik dari kelompok anak yang mengalami *stunting* yang jumlahnya relatif sedikit dibandingkan anak normal. Proses interpolasi SMOTE dapat dijelaskan secara teoretis melalui analisis distribusi probabilitas hasil sintesis (Elreedy, Atiya, dan Kamalov 2024). Namun Salah satu kelemahan SMOTE yang perlu diantisipasi adalah sensitivitas terhadap outlier atau abnormal instances yang terdapat pada kelas minoritas, karena titik data abnormal dapat menghasilkan sampel sintetis yang tidak representatif (Matharaarachchi, Domaratzki, & Muthukumarana, 2024).

1.5.5 Standarisasi Data

Standarisasi data merupakan salah satu proses penting dalam tahap prapemrosesan, terutama dalam konteks analisis statistik dan implementasi algoritma pembelajaran mesin. Prosedur ini bertujuan untuk menyelaraskan skala antar variabel dalam dataset, dengan cara mentransformasi setiap variabel agar memiliki nilai rata-rata nol dan simpangan baku satu. Dengan standarisasi, kontribusi setiap fitur menjadi seimbang, sehingga mencegah terjadinya distorsi dalam analisis akibat perbedaan skala antar atribut. Terdapat dua pendekatan umum dalam standarisasi, yakni *z-score standardization* dan *min-max scaling*. Pada teknik z-score standardization, nilai setiap data dikurangi dengan rata-rata (*mean*) dan dibagi dengan simpangan baku (*standard deviation*), sehingga menghasilkan distribusi dengan nilai tengah nol dan deviasi standar satu (Sujon et al., 2024). Proses ini dilakukan menggunakan persamaan berikut :

$$z_i = \frac{x_i - \bar{x}}{sd} \quad (8)$$

di mana x_i merupakan nilai asli suatu atribut, $\bar{x}$ adalah rata-rata keseluruhan, dan sd adalah simpangan baku (Muhammad, 2024).

Penerapan teknik ini telah terbukti efektif dalam meningkatkan kinerja model-model pembelajaran mesin berbasis linier, seperti *Support Vector Machine* (SVM) dan *Logistic Regression*, khususnya ketika diterapkan pada dataset berskala menengah hingga besar. Selain itu, standarisasi juga memberikan manfaat dalam menyederhanakan proses komputasi serta meningkatkan stabilitas model, karena mampu mengeliminasi dominasi fitur-fitur dengan skala nilai yang besar (Sujon et al., 2024). Lebih lanjut, *z-score* tidak hanya digunakan dalam proses penskalaan, tetapi juga efektif dalam mendeteksi nilai-nilai ekstrem yang menyimpang dari distribusi umum data (*outlier*). Nilai

z-score yang tinggi menunjukkan jarak yang signifikan dari rata-rata, sehingga titik data tersebut dapat diidentifikasi sebagai *outlier* atau anomali dalam *dataset* (Chikodili et al., 2021).

1.5.6 Klasifikasi *Naïve Bayes*

Klasifikasi merupakan salah satu metode dalam *supervised learning* yang berfungsi untuk mengelompokkan data ke dalam kelas-kelas tertentu yang telah ditentukan sebelumnya. Proses ini bersifat iteratif, di mana sistem secara bertahap mengidentifikasi pola dan mengorganisasi objek data ke dalam kategori atau label yang telah didefinisikan. Klasifikasi bertujuan memprediksi keluaran dari suatu permasalahan berdasarkan serangkaian fitur input yang tersedia. Klasifikasi dapat diterapkan baik pada data yang terstruktur maupun tidak terstruktur. Kelas dalam klasifikasi biasanya dikenal dengan istilah label, target, atau kategori. Tujuan akhir dari klasifikasi adalah menetapkan suatu pola data yang tidak diketahui ke dalam salah satu kelas yang telah dikenali sebelumnya (Alnuaimi & Albaldawi, 2024).

Dalam *supervised learning* khususnya pada klasifikasi, *naïve bayes* memiliki komputasi yang sederhana namun tetap efisien dibanding metode lain seperti K-NN dan SVM (Febrian et al., 2023). *Naïve Bayes* bekerja berdasarkan teorema bayes dengan asumsi independensi antar fitur (Glucina,Matko; Lorencin,Ariana; Andelic,Nikola; Lorencin, 2023; Uddin et al., 2024). Teori keputusan Bayes merupakan pendekatan statistik yang mendasar dalam bidang pengenalan pola (*pattern recognition*). Pendekatan ini bekerja dengan menyeimbangkan berbagai kemungkinan keputusan klasifikasi berdasarkan probabilitas serta mempertimbangkan risiko yang mungkin timbul dari setiap keputusan yang diambil (Santoso B. & Umam A., 2018). Metode ini mengklasifikasikan data ke dalam kelas-kelas yang telah ditentukan sebelumnya dengan memanfaatkan konsep probabilitas bersyarat, yaitu peluang suatu kejadian terjadi dengan asumsi bahwa kejadian lain telah terjadi sebelumnya. Pendekatan aturan *Bayes* digunakan untuk memperkirakan kemungkinan suatu atribut muncul berdasarkan data input yang tersedia. Istilah *"naive"* pada algoritma ini menunjukkan bahwa model mengasumsikan adanya kemandirian (independensi) antar nilai atribut (Occhipinti et al., 2022), meskipun dalam praktiknya asumsi ini sering kali tidak sepenuhnya terpenuhi (Chang et al., 2023).

Metode klasifikasi Bayes didasarkan pada probabilitas *posterior* untuk menentukan kelas dari suatu observasi. Dalam konteks klasifikasi kasus *stunting*, metode ini digunakan untuk memperkirakan peluang seorang anak tergolong dalam kategori "*stunting*" atau "tidak *stunting*" berdasarkan sejumlah variabel pengamatan (Bai et al., 2022). Persamaan umum dari klasifikasi *bayes* ditunjukkan pada Persamaan (9):

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)} \tag{9}$$

Dimana Y menyatakan kategori status *stunting* anak (*stunting* atau tidak *stunting*), dan X menyatakan atribut atau variabel pengamatan (tinggi badan, usia, berat badan, dan status gizi) yang diperoleh dari data. $P(Y|X)$ merupakan probabilitas *posterior* bahwa seorang anak termasuk ke dalam kategori status gizi Y, berdasarkan vektor fitur X.

Dalam penelitian ini $Y = \{y = 0, y = 1\}$, dengan $y = 0$ menunjukkan anak *tidak stunting*, dan $y = 1$ menunjukkan anak *stunting*. Selanjutnya, $P(X \mid Y)$ adalah fungsi *likelihood*, yang menggambarkan probabilitas munculnya karakteristik atau atribut X (tinggi badan, berat badan, dan status gizi) jika anak tergolong ke dalam kategori tertentu. $P(Y)$ merupakan probabilitas awal (*prior probability*) seorang anak berada dalam suatu kategori, sedangkan $P(X)$ adalah probabilitas marginal dari fitur X secara keseluruhan.

Dalam penelitian ini, data anak-anak dengan status gizi yang telah diketahui (tidak *stunting* dan *stunting*) digunakan sebagai vektor fitur atau *evidence*, yaitu $X = \{x_1, x_2, \dots, x_n\}$. Ini mencerminkan bahwa pola historis dari atribut anak yang tergolong *stunting* maupun tidak *stunting* turut dipertimbangkan dalam proses klasifikasi. Dengan demikian, klasifikasi *bayes* untuk kasus *stunting* dapat dinyatakan dalam Persamaan (10) (Hosein & Baboolal, 2024):

$$P(Y|X) = \frac{P(Y)P(x_1, x_2, x_3, \dots, x_n|Y)}{P(x_1, x_2, x_3, \dots, x_n)} \tag{10}$$

Fitur-fitur $x_1, x_2, x_3, \dots, x_n$ merepresentasikan nilai atribut (misalnya 0 atau 1) dari anak-anak dengan status gizi yang telah diketahui dan tidak terkontaminasi oleh data tidak lengkap atau *outlier*, dan n merupakan jumlah data anak yang dipilih sebagai sampel. Apabila fitur-fitur tersebut diasumsikan bersifat independen, artinya variabel berat badan, atau tinggi badan tidak memengaruhi variabel lainnya maka metode yang digunakan adalah *naive bayes*, yang menghasilkan Persamaan (10) (Gohari et al., 2023):

$$P(Y|X) = \frac{P(Y)P(x_1|Y)P(x_2|Y)P(x_3|Y) \dots P(x_n|Y)}{P(x_1)P(x_2)P(x_3) \dots P(x_n)} \tag{11}$$

Karena penyebut pada kategori *stunting* dan tidak *stunting* adalah sama, maka probabilitas *posterior* sebanding dengan pembilangnya (Maciura et al., 2023), sebagaimana ditunjukkan pada Persamaan (12). Probabilitas *prior* dari Y dihitung menggunakan Persamaan (13), sedangkan probabilitas kondisional dari fitur dihitung menggunakan Persamaan (14).

$$P(Y|X) \propto P(Y) \prod_{i=0}^{n} P(x_i|Y) \tag{12}$$

$$P(Y) = \frac{n_y}{N} \tag{13}$$

$$P(x_i|Y) = \frac{\sum(x_i|Y)}{n_y} \tag{14}$$

Dimana N adalah jumlah data historis anak dengan status gizi yang valid (tidak mengandung data hilang atau keliru), dan n_y menyatakan jumlah kasus anak yang diklasifikasikan sebagai *stunting* atau tidak *stunting* dalam seluruh data historis tersebut. Sementara itu, $\sum(x_i|Y)$ merujuk pada jumlah kemunculan atribut x_i (seperti berat badan, tinggi badan, atau usia tertentu) dalam kategori *stunting* atau tidak *stunting*, berdasarkan data historis yang telah divalidasi.

Pada suatu kategori tertentu Y (misalnya *stunting* atau tidak *stunting*), probabilitas kondisional suatu fitur dapat bernilai nol akibat keterbatasan jumlah sampel, kesalahan dalam pengukuran, kekeliruan data, atau ketidaksempurnaan dalam faktor-faktor pendukung (evidence). Jika hal ini terjadi, maka hasil perkalian pada sisi kanan Persamaan (13) akan menghasilkan nilai nol. Apabila fitur tersebut muncul pada kedua kategori, maka probabilitas a posteriori yang diperoleh juga akan bernilai nol untuk keduanya, yakni $P(y = 1 \mid X) = P(y = 0 \mid X) = 0$, sehingga proses klasifikasi tidak dapat dilakukan (Bai et al., 2022). Untuk mengatasi permasalahan tersebut, dilakukan modifikasi terhadap perhitungan probabilitas kondisional pada Persamaan (14), yang kemudian direformulasikan dalam Persamaan (15).

$$P_k(x_i|Y) = \begin{cases} \frac{1}{N}, & P(x_i|Y) = 0 \\ P(x_i|Y), & P(x_i|Y) \neq 0 \end{cases} \quad (15)$$

Untuk menghindari pengulangan informasi atau perhitungan yang tidak menambah nilai baru atau tidak memengaruhi hasil akhir (*redundant*), fitur-fitur x_i yang memiliki nilai $P(x_i \mid Y)$ yang sama pada berbagai kategori Y (*stunting* dan tidak *stunting*) disaring terlebih dahulu. Setelah itu, proses dilanjutkan dengan menentukan kategori Y yang memberikan nilai probabilitas maksimum, sebagaimana ditunjukkan pada Persamaan (16).

$$y = argmax_y P(Y) \prod_{i=0}^{n} P_k(x_i|Y), kecuali\ \ i\ \ jika\ P(x_i|y = 1) = P(x_i|y = 0) \quad (16)$$

Terdapat sejumlah besar fitur yang digunakan, sehingga proses perkalian dalam perhitungan probabilitas dapat menghasilkan nilai yang sangat kecil (mendekati nol). Oleh karena itu, digunakan transformasi logaritma untuk mengubah bentuk perhitungan menjadi penjumlahan, sebagaimana ditunjukkan pada Persamaan (17).

$$P(Y|X) \propto \ln\left\{P(Y) \prod_{i=0}^{n} [P_k(x_i|Y)]\right\} = \ln[P(Y)] + \sum_{i=0}^{n} \ln[P_k(x_i|Y)] \quad (17)$$

Dengan demikian, akhir dalam proses klasifikasi *stunting* ini adalah menentukan kategori Y (*stunting* atau tidak *stunting*) yang menghasilkan nilai maksimum dari perhitungan probabilitas, sebagaimana ditunjukkan pada Persamaan (18).

$$y = argmax_y \left\{\ln[P(Y)] + \sum_{i=0}^{n} \ln[P_k(x_i|Y)]\right\} \quad (18)$$

Algoritma Naïve bayes memiliki beberapa estimator seperti *Multinomial*, *Bernoulli* dan *Gaussian* (Singh et al., 2022). Dalam penelitian ini estimator yang digunakan adalah *Bernoulli.* Pada model *Bernoulli Naive Bayes*, setiap fitur diperlakukan sebagai variabel biner yang saling independen (Boolean) dan berfungsi untuk merepresentasikan karakteristik dari data masukan (Occhipinti et al., 2022).

Perhitungan probabilitas pada distribusi *Bernoulli* dilakukan dengan rumus berikut (Mustamin et al., 2023) :

$$P(X = x_i|Y = y) = P(x_i = 1|y)x_i + \big(1 - P(x_i = 1|y)\big)(1 - x_i) \qquad (19)$$

Dalam konteks klasifikasi *stunting*, model *Bernoulli Naive Bayes* (NB) berperan sebagai pengklasifikasi untuk data yang memiliki karakteristik biner, yaitu fitur-fitur yang hanya bernilai 0 atau 1. Artinya, setiap variabel prediktor menggambarkan apakah suatu kondisi terjadi (*Stunting*) atau tidak terjadi (normal) pada seorang anak. Misalkan v adalah vektor fitur yang merepresentasikan kondisi seorang anak, dan $\tilde{p} = P[v = 1|y]$ menunjukkan probabilitas bersyarat bahwa fitur tersebut muncul pada kelas tertentu (misalnya, kelas *stunting*). Maka fungsi probabilitas bersyarat model *Bernoulli Naïve Bayes* dapat dituliskan sebagai berikut (Peretz et al., 2024):

$$P[v|y] = \tilde{p}v + (1 - \tilde{p})(1 - v) \qquad (20)$$

1.5.7 *Cross Validation*

Cross Validation adalah metode untuk mengevaluasi kinerja model dengan cara membagi data secara acak ke dalam beberapa bagian, di mana masing-masing bagian akan digunakan dalam proses klasifikasi. Proses validasi bertujuan untuk mengukur performa hasil prediksi model, sedangkan evaluasi digunakan untuk mengamati serta menganalisis hasil kerja model secara keseluruhan.

Akurasi menggambarkan seberapa dekat hasil prediksi model dengan nilai aktual. Presisi menunjukkan tingkat ketepatan model dalam melakukan klasifikasi, sedangkan recall digunakan untuk menilai seberapa baik model dalam mengidentifikasi data positif yang benar dari keseluruhan data positif aktual yang tersedia (Wibawa *et al.*, 2022).

Dalam proses evaluasi untuk menilai akurasi, salah satu metode yang umum digunakan adalah *confusion matrix*. *Confusion matrix* berfungsi sebagai alat ukur berbentuk matriks yang digunakan untuk mengetahui seberapa tepat hasil klasifikasi terhadap setiap kelas berdasarkan algoritma yang diterapkan (Fadil, 2025). Tabel *confusion matrix* dapat dilihat pada **Tabel.1** berikut:

Tabel 1. *Confusion matrix*

Klasifikasi		**Aktual**	
		True	**False**
Prediksi	**True**	*True Positif (TP)*	*False Negatif (FN)*
	False	*False Positif (FP)*	*True Negatif (TN)*

Sumber (Fadil, 2025)

Dimana *True Positive* (TP) menunjukkan jumlah data yang secara aktual bernilai positif dan berhasil dikenali dengan tepat sebagai positif oleh model. *True Negative* (TN) mengacu pada jumlah data yang benar-benar negatif dan diprediksi negatif secara akurat. *False Positive* (FP) terjadi ketika data yang sebenarnya negatif justru diklasifikasikan sebagai positif. Sementara itu, *False Negative* (FN) merupakan kondisi di mana data yang seharusnya positif salah diprediksi sebagai negatif oleh model.

Nilai dari akurasi, presisi, dan recall dapat diperoleh melalui perhitungan menggunakan rumus sebagai berikut (Fadil et al., 2025):

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (21)$$

$$precison = \frac{TP}{TP + FP} \quad (22)$$

$$recall = \frac{TP}{TP + FN} \quad (23)$$

Kemudian untuk melihat peluang kesalahan klasifikasi atau *Apparent Error Rate* (APER) dapat dihitung seperti pada Persamaan (24) berikut ini (Fadil et al., 2025):

$$APER = 1 - accuracy \quad (24)$$

1.5.8 Stunting

Stunting adalah kondisi di mana pertumbuhan anak balita terhambat akibat kekurangan gizi yang berlangsung lama, terutama pada masa penting 1000 hari pertama kehidupan, yaitu dari kehamilan hingga anak berusia dua tahun. Anak dengan *stunting* memiliki tinggi badan yang lebih rendah dibandingkan dengan standar pertumbuhan anak seusianya (Kementrian Kesehatan RI, 2023). Menurut Kementerian Kesehatan Republik Indonesia, *stunting* didefinisikan sebagai gangguan pertumbuhan pada anak balita yang ditandai dengan tinggi badan kurang dari -2 standar deviasi di bawah median standar pertumbuhan yang ditetapkan oleh *World Health Organization* (WHO). *Stunting* bukan hanya soal tinggi badan yang pendek, Kondisi ini juga berdampak pada perkembangan kemampuan berpikir, kesehatan secara keseluruhan, serta potensi anak dalam menjalani kehidupan dan produktivitas di masa depan (Kementrian Kesehatan RI, 2022). Anak yang mengalami *stunting* juga berisiko lebih tinggi terhadap penyakit kronis seperti hipertensi dan diabetes (Fitriana et al., 2026).

Anak dikategorikan mengalami *stunting* apabila tinggi badan menurut umur (TB/U) berada di bawah -2 standar deviasi (SD) dari median standar WHO, yang menunjukkan gangguan pertumbuhan tinggi badan. Jika TB/U kurang dari -3 SD, anak tersebut diklasifikasikan sebagai sangat *stunting*, yang berarti kondisi pertumbuhannya lebih parah. Sedangkan anak dengan TB/U sama dengan atau lebih besar dari -2 SD dianggap memiliki pertumbuhan tinggi badan yang normal dan tidak mengalami *stunting* (Fitriana et al., 2026). Klasifikasi status *stunting* pada balita disajikan pada Tabel 2.

Tabel 2. Klasifikasi Kasus Gizi *Stunting*

	Status	***Z-score***
Panjan Badan atau tinggi Badan menurut Umur (PB/U atau TB/U) anak usia 0-60 bulan	Sangat Pendek	< -3
	Pendek (*stunting*)	-3 sd < -2
	Normal	-2 sd < +3
	Tinggi	> +3

Sumber : Permenkes No 2 Tahun 2020

Adapun standar *Z-score* Panjang Badan menurut Usia (PB/U) untuk balita berdasarkan standar pertumbuhan yang dikeluarkan oleh *World Health Organization* (WHO) dapat dilihat pada **Tabel 3** dan **Tabel 4.**

Tabel 3. *Z-score* Panjang badan/Usia Balita Perempuan

Usia (Bulan)	-3 SD	-2 SD	-1 SD	Median	1 SD	2 SD	3 SD
0	43,6	45,4	47,3	49,1	51,0	52,9	54,7
1	47,8	49,8	51,7	53,7	55,6	57,6	59,5
2	51.0	53,0	55,0	57,1	59,1	61,1	63,2
3	53,5	55,6	57,7	59,8	61,9	64,0	66,1
4	55,6	57,8	59,9	62,1	64,3	66,4	68,6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
56	93,4	98,1	102,7	107,3	111,9	116,5	121,1
57	93,9	98,5	103,2	107,8	112,5	117,1	121,8
58	94,3	99,0	103,7	108,4	113,0	117,7	122,4
59	94,7	99,5	104,2	108,9	113,6	118,3	123,1

Sumber : (WHO, 2021)

Tabel 4. *Z-score* Panjang badan/Usia Balita Laki-Laki

Usia (Bulan)	-3 SD	-2 SD	-1 SD	Median	1 SD	2 SD	3 SD
0	44,2	46,1	48,0	49,9	51,8	53,7	55,6
1	48,9	50,8	52,8	54,7	56,7	58,6	60,6
2	52,4	54,4	56,4	58,4	60,4	62,4	64,4
3	55,3	57,3	59,4	61,4	63,5	65,5	67,6
4	57,6	59,7	61,8	63,9	66,0	68,0	70,1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
56	94,3	98,8	103,3	107,8	112,3	116,7	121,2
57	94,7	99,3	103,8	108,3	112,8	117,4	121,9
58	95,2	99,7	104,3	108,9	113,4	118,0	122,6
59	95,6	100,2	104,8	109,4	114,0	118,6	123,2

Sumber : (WHO, 2021)

1.5.9 Hubungan Karakteristik Anak dengan Kejadian *Stunting*

Karakteristik anak memiliki peranan penting dalam menentukan risiko terjadinya *stunting*. Beberapa studi menunjukkan bahwa faktor-faktor seperti jenis kelamin, usia, berat badan lahir, dan urutan kelahiran anak berpengaruh signifikan terhadap status gizi dan pertumbuhan anak (Niragire et al., 2025). Anak laki-laki cenderung memiliki risiko lebih tinggi mengalami *stunting* dibandingkan anak perempuan (Guo et al., 2025; Karlsson et al., 2023; Samuel et al., 2022) sebagaimana ditunjukkan oleh nilai *Prevalence Odds Ratio* (POR) (Niragire et al., 2025). Hal ini dapat dikaitkan dengan kebutuhan gizi yang lebih tinggi pada anak laki-laki serta perbedaan fisiologis dalam respons terhadap stres lingkungan dan infeksi.

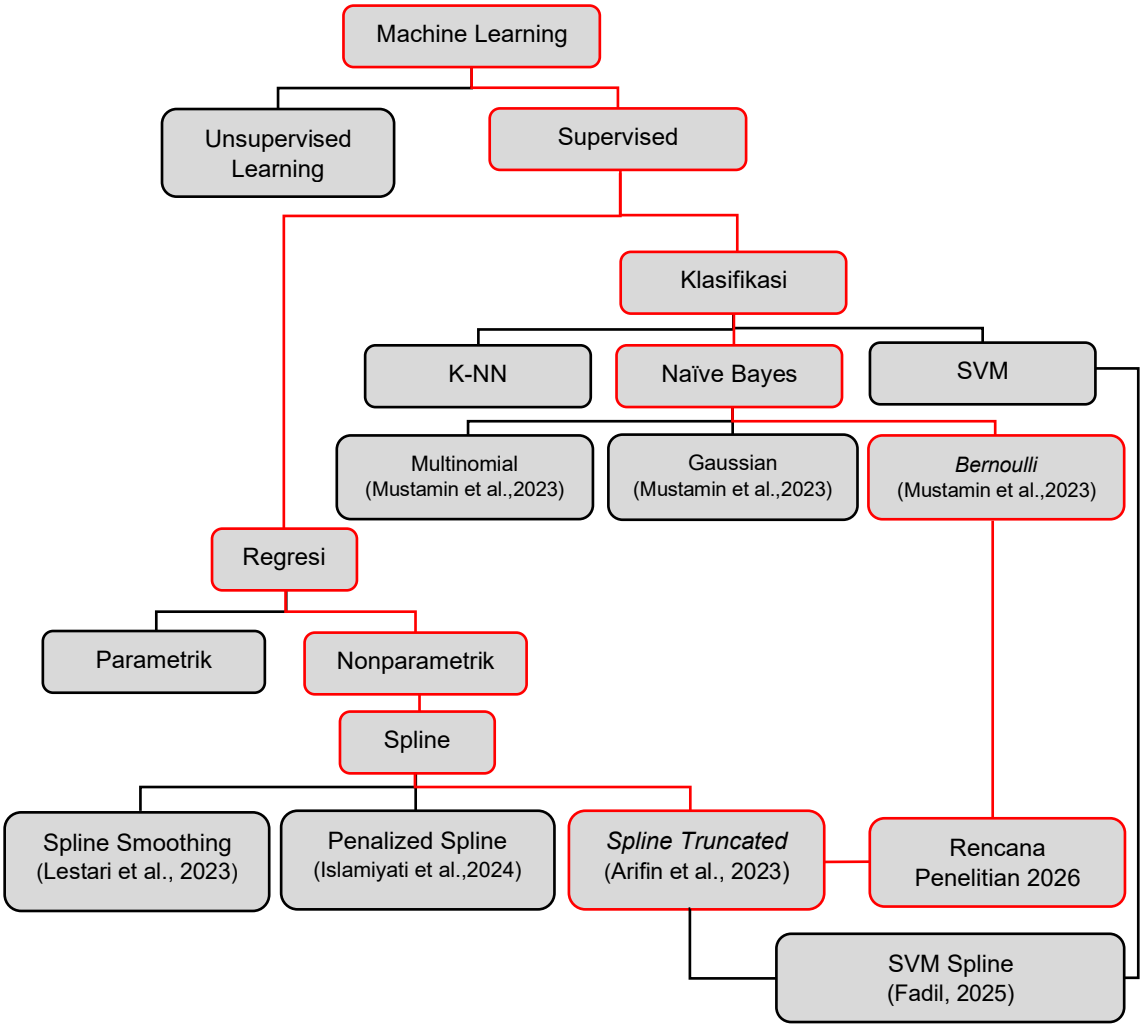
Selain itu, usia anak juga merupakan faktor yang berpengaruh kuat, di mana prevalensi *stunting* paling tinggi terjadi pada kelompok usia 24–35 bulan (Niragire et al., 2025).

Periode ini merupakan fase transisi penting dari pemberian ASI eksklusif menuju konsumsi makanan pendamping, sehingga ketidakseimbangan asupan gizi dan paparan infeksi dapat meningkat. Di sisi lain, berat badan lahir juga menjadi indikator awal status gizi anak (Cvorovi, 2022; Id et al., 2021; Prasetyo et al., 2023). Anak yang lahir dengan berat badan normal (≥2,5 kg) memiliki risiko lebih rendah mengalami *stunting* dibandingkan anak dengan berat badan lahir rendah (Niragire et al., 2025). Kondisi berat lahir rendah mencerminkan gangguan pertumbuhan janin akibat kekurangan gizi ibu selama kehamilan, yang berdampak jangka panjang terhadap pertumbuhan linier anak (Bilal et al., 2025; Leroy et al., 2020).

1.5.10 Kerangka Konseptual

Penelitian ini memanfaatkan pendekatan Supervised *Learning* untuk membangun model klasifikasi status *stunting* pada balita berdasarkan berbagai variabel prediktor seperti usia, berat badan, serta panjang badan balita. *Supervised learning* dipilih karena menyediakan kerangka kerja yang sesuai untuk membangun model prediktif berdasarkan data berlabel, di mana status *stunting* (ya/tidak) menjadi label kelas yang ingin diprediksi. Pendekatan ini mencakup dua teknik utama, yakni regresi nonparametrik dan klasifikasi probabilistik. Pada tahap awal, digunakan metode *Spline Truncated,* untuk memodelkan hubungan non-linier antara variabel prediktor dengan probabilitas terjadinya *stunting*.

Hasil dari regresi spline ini digunakan untuk mengekstraksi pola atau fitur baru yang lebih representatif terhadap dinamika status pertumbuhan anak. Fitur-fitur ini selanjutnya dimasukkan ke dalam algoritma klasifikasi *Naive Bayes*, dengan pendekatan *Bernoulli Naïve Bayes* yang mampu menghitung probabilitas seorang anak mengalami *stunting* berdasarkan distribusi nilai dari masing-masing fitur. Metode *Naive Bayes Spline* ini diharapkan dapat menghasilkan model prediktif yang tidak hanya akurat, tetapi juga mampu menangkap hubungan non-linier antar variabel yang umum terjadi dalam konteks kesehatan masyarakat. Dengan demikian, model ini diharapkan dapat menjadi alat bantu yang efektif dalam proses skrining dan pengambilan keputusan intervensi dini terhadap risiko *stunting* pada anak balita.



Gambar 1. Kerangka Konseptual

BAB II
METODE PENELITIAN

2.1 Sumber Data

Penelitian ini menggunakan data sekunder berupa informasi mengenai status gizi balita di Kota Kendari tahun 2024. Data tersebut merupakan hasil pengukuran status gizi berdasarkan indikator berat badan menurut panjang badan (BB/PB) yang dilakukan di 10 Puskesmas yang ada di wilayah Kota Kendari.

2.2 Variabel Penelitian

Penelitian ini menggunakan satu variabel respon (y) dan tiga variabel prediktor (x). Variabel respon yang dianalisis adalah status gizi balita berdasarkan indikator berat badan menurut panjang badan (BB/PB), yang diklasifikasikan ke dalam dua kategori, *stunting* dan normal. Sementara itu, variabel prediktor mencakup sejumlah faktor yang memengaruhi status gizi balita, sebagaimana ditampilkan pada **Tabel 5**.

Tabel 5. Variabel Penelitian

Variabel Penelitian	Kode	Nama Variabel	Skala Pengukuran
Variabel Respon	y	Status Gizi Balita Berdasarkan Berat Badan Menurut Panjang Badan (BB/PB)	0 = Normal 1 = *Stunting*
Variabel Prediktor	x_1	Usia (Bulan)	Rasio
	x_2	Berat Badan (Kg)	Rasio
	x_3	Panjang Badan (Cm)	Rasio

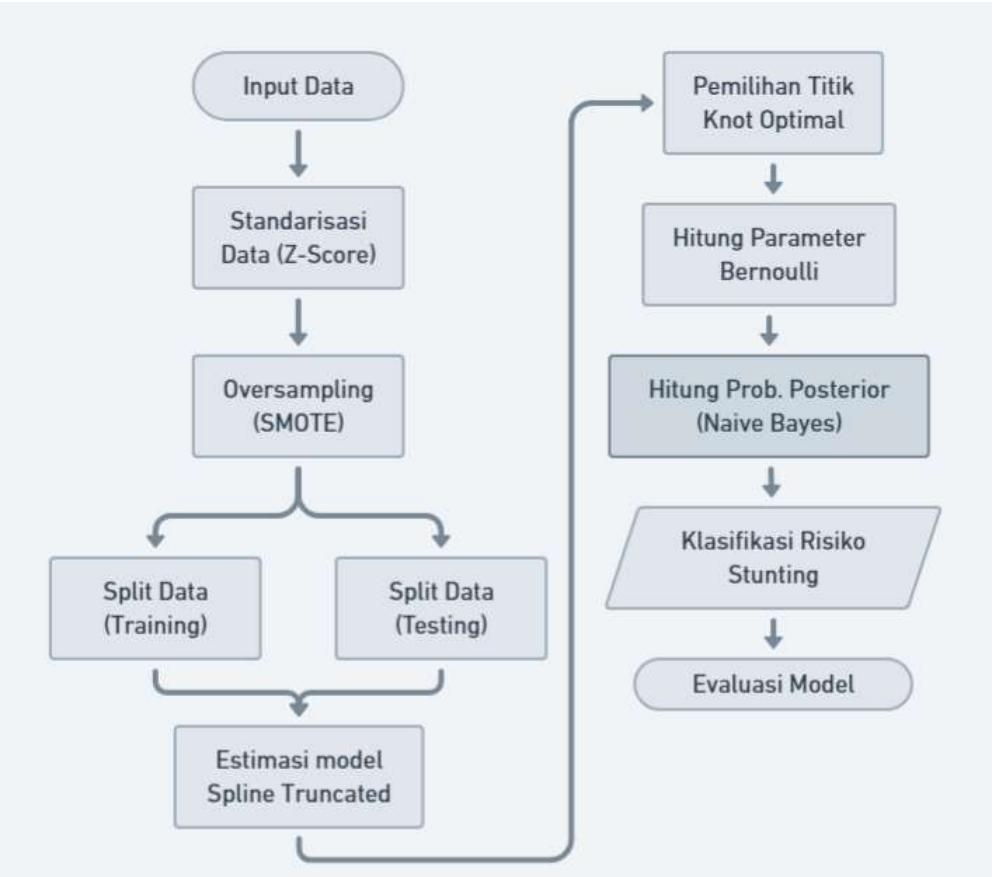
2.3 Metode Penelitian

Langkah-langkah yang akan dilakukan dalam penelitian ini adalah sebagai berikut :

1. Melakukan standarisasi variabel prediktor untuk menyamakan skala antar fitur menggunakan persamaan (7).
2. Melakukan *Oversampling* dengan metode SMOTE.
3. Membangun model *spline truncated* :
   a. Menentukan titik simpul (*knot*) optimal menggunakan persamaan (5).
   b. Menghasilkan basis spline sebagai representasi baru dari fitur.
4. Mengintegrasikan hasil spline ke dalam *Bernoulli Naïve Bayes classifier*.
5. Melatih model menggunakan data training.
6. Menguji model pada data testing untuk mendapatkan hasil klasifikasi status gizi balita.
7. Evaluasi Akurasi Model :
   a. Mengukur performa model menggunakan *confusion matrix* untuk menghitung (Akurasi, presisi*, recall,* dan *error rate/APER).*
   b. Menggunakan teknik *cross-validation* untuk validasi model.
8. Perbandingan dengan Naïve Bayes Standar :
   a. Membangun model *Naïve Bayes* (tanpa spline).

b. Melakukan proses klasifikasi dan evaluasi performa dengan data yang sama.
c. Membandingkan hasil akurasi, presisi, recall, dan error rate antara *naïve bayes spline* dengan *naïve bayes.*
d. Membandingkan kurva ROC berdasarkan nilai AUC antara *naïve bayes spline* dengan *naïve bayes.*

9. Kesimpulan.



Gambar 2. Diagram Alir Metode Penelitian