

PERBANDINGAN REGRESI NONPARAMETRIK *SPLINE TRUNCATED* DAN *KERNEL GAUSSIAN* DALAM PEMODELAN PENDERITA TUBERKULOSIS (TBC) DI INDONESIA TAHUN 2025

Fauziah Syafitri

Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Bengkulu

email: fauziah.syafitri@mhs.unib.ac.id

Reva Amelia

Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Bengkulu

email: reva43665@gmail.com

Chesa Nabilla Azzahra

Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Bengkulu

email: chesaazzahra.cna@gmail.com

Idhia Sriliana

Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Bengkulu

email: idhiasriliana@unib.ac.id*

Etis Sunandi

Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Bengkulu

email: esunandi@unib.ac.id

Abstrak

Tuberkulosis (TBC) masih menjadi masalah kesehatan utama di Indonesia dengan tingkat prevalensi yang bervariasi antar wilayah. Hubungan antara prevalensi TBC dan faktor-faktor yang memengaruhinya cenderung kompleks serta tidak selalu mengikuti pola linier. Penelitian ini membandingkan kinerja regresi nonparametrik *Spline Truncated* dan *Kernel Gaussian* dalam memodelkan prevalensi TBC di Indonesia tahun 2025, menggunakan variabel jumlah rumah sakit umum, jumlah kasus HIV/AIDS, persentase akses sanitasi layak, dan persentase akses air minum layak. Pemilihan *knot* optimal pada *Spline Truncated* dan *bandwidth* optimal pada *Kernel Gaussian* dilakukan menggunakan GCV. Pada metode *Spline Truncated* diperoleh model terbaik dengan satu titik *knot* optimal pada masing-masing variabel prediktor, nilai MSE sebesar 2,2365 dan R^2 sebesar 83,89%. Sementara itu, metode *Kernel Gaussian* menghasilkan MSE sebesar 8,311203 dan R^2 sebesar 40,13%. Berdasarkan perbandingan tersebut, regresi *Spline Truncated* menunjukkan kinerja yang lebih baik dibandingkan *Kernel Gaussian* karena memiliki MSE lebih kecil dan R^2 lebih tinggi. Temuan ini dapat digunakan sebagai alternatif statistik untuk mendukung kebijakan kesehatan masyarakat di Indonesia.

Kata Kunci: Tuberkulosis, Regresi Nonparametrik, *Spline Truncated*, *Kernel Gaussian*, *Generalized Cross Validation* (GCV)

Abstract

Tuberculosis (TB) remains a major public health problem in Indonesia, with prevalence rates varying across regions. The relationship between TB prevalence and the factors influencing it tends to be complex and does not always follow a linear pattern. This study compares the performance of Truncated Spline and Gaussian Kernel nonparametric regression in modeling TB prevalence in Indonesia in 2025, using the number of general hospitals, the number of HIV/AIDS cases, the percentage of access to adequate sanitation, and the percentage of access to safe drinking water as variables. Optimal knot selection for the Truncated Spline and optimal bandwidth for the Gaussian Kernel were determined using GCV. For the Truncated Spline method, the best model was obtained with one optimal knot for each predictor variable, an MSE of 2,2365 and an R^2 of 83,89%. Meanwhile, the Gaussian Kernel method produced an MSE of 8,311203 and an R^2 of 40,13%. Based on this comparison, Truncated Spline regression demonstrated better performance than the Gaussian Kernel method because it had a smaller MSE and a higher R^2 . These findings can serve as statistical evidence to support public health policies in Indonesia.

Keywords: Tuberculosis, nonparametric regression, truncated spline, Gaussian kernel, Generalized Cross Validation (GCV).

PENDAHULUAN

Tuberkulosis (TBC) adalah penyakit infeksi menular yang disebabkan oleh bakteri *Mycobacterium tuberculosis* dan hingga saat ini masih menjadi salah satu permasalahan kesehatan dengan jumlah kasus yang tinggi di berbagai negara, terutama di negara-negara berkembang (Meliasari, 2021). *World Health Organization* melaporkan bahwa jumlah kasus TBC global mencapai 10,6 juta kasus pada tahun 2022 dan menjadikan TBC sebagai salah satu penyebab kematian tertinggi akibat penyakit menular (World Health Organization, 2023). Di Indonesia, TBC masih menjadi fokus utama dalam program pengendalian penyakit karena tingginya prevalensi kasus yang tersebar di berbagai provinsi (Sudirman dkk., 2024). Tingginya prevalensi tersebut menunjukkan bahwa penyebaran TBC tidak hanya dipengaruhi oleh faktor kesehatan, tetapi juga berkaitan dengan kondisi lingkungan dan sosial masyarakat (Kusumaningrum dkk., 2023). Perbedaan karakteristik antarwilayah di Indonesia menyebabkan tingkat prevalensi TBC pada setiap daerah memiliki pola yang beragam sehingga diperlukan pendekatan analisis yang mampu menggambarkan hubungan data secara lebih fleksibel.

Ketersediaan fasilitas pelayanan kesehatan seperti rumah sakit dan tenaga kesehatan memiliki peranan penting dalam mendukung proses pencegahan, diagnosis, dan penanganan TBC. Selain itu, peningkatan jumlah kasus *Human Immunodeficiency Virus* (HIV) diketahui dapat memperbesar risiko infeksi TBC akibat menurunnya sistem kekebalan tubuh. Faktor lingkungan seperti akses sanitasi layak dan air minum layak juga menjadi indikator penting yang berkaitan dengan kualitas kesehatan masyarakat dan potensi penyebaran penyakit menular (Kusumaningrum dkk., 2023; Rahmah, 2020). Daerah dengan kualitas sanitasi yang rendah dan keterbatasan akses air minum layak cenderung memiliki risiko penyebaran penyakit yang lebih tinggi dibandingkan daerah dengan kondisi lingkungan yang lebih baik. Oleh karena itu, analisis terhadap faktor-faktor tersebut menjadi penting untuk memahami pola prevalensi TBC di Indonesia secara lebih komprehensif.

Hubungan antara prevalensi TBC dan faktor-faktor yang memengaruhinya sering kali membentuk pola yang kompleks dan tidak mengikuti asumsi

linear tertentu sehingga pendekatan regresi parametrik menjadi kurang fleksibel dalam menggambarkan karakteristik data (Naufal dkk., 2025). Pada kondisi seperti ini, penggunaan metode statistik yang mampu menyesuaikan bentuk kurva berdasarkan pola data aktual menjadi lebih relevan. Regresi nonparametrik merupakan salah satu pendekatan statistik yang mampu memodelkan hubungan antarvariabel tanpa mensyaratkan bentuk fungsi tertentu sehingga lebih adaptif terhadap pola data yang bersifat *non-linear* (Härdle, 1990). Pendekatan ini dapat memberikan hasil estimasi yang lebih baik ketika pola hubungan antarvariabel berubah-ubah pada interval tertentu atau tidak diketahui secara pasti. Selain itu, regresi nonparametrik juga dinilai lebih fleksibel dalam menangani data dengan pola hubungan yang tidak stabil dibandingkan pendekatan parametrik (Abdy, 2019).

Metode *Spline Truncated* bekerja melalui pembentukan titik *knot* untuk menangkap perubahan pola hubungan pada interval tertentu, sedangkan *Kernel Gaussian* menggunakan pendekatan pembobotan lokal untuk menghasilkan estimasi kurva yang halus dan kontinu (Eubank, 1999; Härdle dkk., 2004). Pemilihan parameter optimal pada kedua metode menjadi tahapan penting karena akan memengaruhi tingkat keakuratan model yang dihasilkan. Pada *Spline Truncated*, pemilihan titik *knot* yang optimal mampu menghasilkan kurva yang lebih representatif terhadap pola data, sedangkan pada *Kernel Gaussian* pemilihan *bandwidth* yang tepat dapat mengurangi kesalahan estimasi akibat kurva yang terlalu halus maupun terlalu kasar. Oleh karena itu, metode *Generalized Cross Validation* (GCV) digunakan dalam penelitian ini untuk memperoleh parameter optimal pada masing-masing metode. Penggunaan GCV dinilai efektif dalam menentukan parameter optimal pada model nonparametrik karena mampu meminimumkan kesalahan prediksi model (Ratnasari dkk., 2021).

Beberapa penelitian sebelumnya menunjukkan bahwa regresi nonparametrik mampu memberikan hasil pemodelan yang baik pada data dengan pola hubungan yang kompleks. Penelitian mengenai *Spline Truncated* pada data longitudinal menunjukkan bahwa metode tersebut mampu memberikan hasil estimasi yang baik pada pemodelan persentase penduduk miskin (Dani dkk.,

2021). Selain itu, penelitian mengenai regresi *Kernel Gaussian* juga menunjukkan performa yang baik dalam memodelkan indeks pembangunan manusia berdasarkan faktor sosial-ekonomi di Indonesia (Anwar dkk., 2025). Penelitian lain terkait pemodelan kasus tuberkulosis dengan pendekatan *Spline Truncated* menunjukkan bahwa metode nonparametrik mampu menangkap pola hubungan data secara lebih fleksibel (Pratama dkk., 2021). Namun, penelitian yang secara khusus membandingkan performa *Spline Truncated* dan *Kernel Gaussian* dalam pemodelan prevalensi TBC di Indonesia masih relatif terbatas. Selain itu, sebagian besar penelitian terdahulu lebih berfokus pada pendekatan parametrik sehingga belum banyak mengkaji fleksibilitas model nonparametrik dalam menangkap pola hubungan prevalensi TBC dengan faktor-faktor kesehatan dan lingkungan.

Berdasarkan penjelasan tersebut, penelitian ini dilakukan untuk membandingkan metode regresi nonparametrik *Spline Truncated* dan *Kernel Gaussian* dalam memodelkan prevalensi TBC di Indonesia pada tahun 2025 berdasarkan variabel jumlah rumah sakit, jumlah kasus HIV/AIDS, persentase akses sanitasi layak, serta persentase akses air minum layak. Penelitian ini bertujuan untuk menentukan metode dengan kinerja terbaik berdasarkan nilai *Generalized Cross Validation* (GCV), *Mean Squared Error* (MSE), dan koefisien determinasi (R^2). Hasil penelitian diharapkan dapat memberikan informasi mengenai metode yang paling sesuai dalam memodelkan prevalensi TBC serta menjadi referensi dalam pengembangan analisis statistik di bidang kesehatan masyarakat. Selain itu, hasil penelitian juga diharapkan dapat membantu pemerintah dan pihak terkait dalam memahami faktor-faktor yang berhubungan dengan prevalensi TBC sebagai dasar penyusunan kebijakan pengendalian penyakit yang lebih tepat sasaran.

KAJIAN TEORI

Tuberkulosis

Tuberkulosis (TBC) merupakan penyakit infeksi pada sistem pernapasan manusia yang disebabkan oleh bakteri berbentuk basil yang memiliki daya tahan tinggi, sehingga proses pengobatannya memerlukan waktu relatif lama. Penyakit ini ditimbulkan oleh bakteri *Mycobacterium tuberculosis* yang pertama kali diidentifikasi oleh Robert Koch

pada 24 Maret 1882. Bakteri tersebut bersifat patogen bagi manusia, terutama pada kelompok masyarakat miskin dan penderita *Human Immunodeficiency Virus* (HIV) yang tinggal di wilayah perkotaan padat penduduk, karena tingginya potensi penularan melalui udara serta adanya kasus yang tidak mendapatkan pengobatan. Penularan terjadi melalui saluran pernapasan, sehingga bakteri *M. tuberculosis* dapat dengan mudah berpindah dari satu individu ke individu lain hanya melalui hembusan napas atau percikan batuk.

Bakteri *M. tuberculosis* atau basil tuberkel berbentuk batang lurus dengan ukuran sekitar $0,4 - 3 \mu_m$. Gejala yang umum muncul pada penderita TBC antara lain batuk berdahak yang berlangsung lebih dari dua minggu, nyeri dada, dan sesak napas, serta disertai gejala sistemik seperti demam, malaise, keringat malam, dan anoreksia. Dalam beberapa kasus, batuk juga dapat disertai keluarnya darah. Selain itu, penderita TBC kerap mengalami penurunan nafsu makan yang berujung pada berkurangnya berat badan dan timbulnya rasa lelah yang berkepanjangan (Rahmah, 2020).

Regresi Nonparametrik

Statistika nonparametrik merupakan metode statistika yang digunakan ketika bentuk distribusi data tidak diketahui secara pasti atau tidak memenuhi asumsi parametrik tertentu. Suatu metode statistika dikatakan nonparametrik apabila dapat digunakan pada data dengan skala pengukuran nominal, ordinal, maupun interval dan rasio tanpa memerlukan asumsi distribusi tertentu pada data. Pada data berskala interval atau rasio, metode nonparametrik digunakan ketika fungsi distribusi dari peubah acak yang menghasilkan data tidak diketahui atau memiliki parameter yang sangat banyak dan tidak diketahui. Terminologi lain dari statistika nonparametrik (*nonparametric statistics*) adalah statistika bebas sebaran (*distribution-free statistics*) karena metode ini tidak bergantung pada asumsi distribusi tertentu (Nugroho, 2008).

Regresi nonparametrik adalah metode regresi yang digunakan untuk memodelkan hubungan antara variabel respon dan variabel prediktor tanpa menetapkan asumsi bentuk fungsi tertentu, seperti fungsi linear maupun kuadrat. Metode ini bersifat

fleksibel karena bentuk kurva regresi mengikuti pola data pengamatan (Abdy, 2019).

Regresi Nonparametrik *Spline Truncated*

Regresi *Spline Truncated* merupakan salah satu metode regresi nonparametrik yang digunakan untuk memodelkan hubungan antara variabel respon dan variabel prediktor ketika pola hubungan antarvariabel belum diketahui secara jelas atau pasti. Metode *Spline Truncated* memiliki fleksibilitas yang tinggi karena mampu menyesuaikan karakteristik lokal data secara efektif melalui pembentukan titik *knot* pada interval tertentu (Dani dkk., 2021).

Pada regresi *Spline Truncated*, fungsi *spline* dibentuk menggunakan potongan-potongan polinomial yang kontinu pada titik *knot* sehingga mampu menangkap perubahan pola data pada sub-interval tertentu. Titik *knot* merupakan titik perpindahan pola kurva yang menyebabkan model menjadi lebih fleksibel dibandingkan model regresi parametrik biasa. Model regresi *Spline Truncated* secara umum dituliskan sebagai berikut (Pratama dkk., 2021):

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_i^j + \sum_{k=1}^K \beta_{p+k} (x_i - K_k)_+^p + \varepsilon_i \quad (1)$$

dengan:

K_k merupakan titik *knot*,

p merupakan orde *spline*,

$(x_i - K_k)_+^p$ merupakan fungsi *truncated*.

Fungsi *truncated* didefinisikan sebagai berikut (Dewi & Ahadiyah, 2023):

$$(x_i - K_k)_+^p = \begin{cases} (x_i - K_k)^p, & x_i \geq K_k \\ 0, & x_i < K_k \end{cases} \quad (2)$$

Pemilihan titik *knot* optimal sangat penting dalam regresi *Spline Truncated* karena memengaruhi tingkat kehalusan kurva yang terbentuk. Salah satu metode yang umum digunakan dalam pemilihan titik *knot* optimal adalah *Generalized Cross Validation* (GCV), di mana model terbaik diperoleh berdasarkan nilai GCV minimum (Dani & Ni'matuzzahroh, 2021).

Regresi Nonparametrik *Kernel Gaussian*

Regresi *Kernel Gaussian* merupakan salah satu metode regresi nonparametrik yang digunakan untuk mengestimasi hubungan antara variabel respon dan variabel prediktor tanpa mengasumsikan bentuk fungsi tertentu. Metode ini bekerja dengan memberikan pembobot pada data di sekitar titik pengamatan menggunakan fungsi kernel

sehingga menghasilkan kurva regresi yang lebih halus dan fleksibel. Salah satu fungsi kernel yang sering digunakan adalah *Kernel Gaussian* karena memiliki bentuk yang kontinu dan mampu menghasilkan estimasi yang stabil (Ratnasari dkk., 2021).

Estimator regresi *Kernel Gaussian* secara umum dinyatakan sebagai berikut:

$$\hat{m}(x) = \frac{\sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) y_i}{\sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)} \quad (3)$$

dengan:

$K\left(\frac{x - x_i}{h}\right)$ merupakan fungsi kernel,

h merupakan *bandwidth*,

y_i merupakan variabel respon.

Fungsi *Kernel Gaussian* dirumuskan sebagai:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}, -\infty < u < \infty \quad (4)$$

Parameter *bandwidth* (h) memiliki peranan penting dalam regresi *Kernel Gaussian* karena memengaruhi tingkat kehalusan kurva estimasi yang dihasilkan. Nilai *bandwidth* yang terlalu kecil menyebabkan kurva menjadi sangat berfluktuasi, sedangkan *bandwidth* yang terlalu besar menyebabkan kurva menjadi terlalu halus sehingga pola data tidak dapat ditangkap dengan baik. Oleh karena itu, diperlukan pemilihan *bandwidth* optimal untuk memperoleh model terbaik (Efendi dkk., 2025).

Generalized Cross Validation (GCV)

Generalized Cross Validation (GCV) adalah salah satu metode yang digunakan untuk menentukan model terbaik dalam regresi nonparametrik, khususnya dalam pemilihan titik *knot* pada *spline truncated* dan *bandwidth* pada *Kernel Gaussian*. Metode GCV digunakan karena mampu mengukur tingkat kesalahan model tanpa harus membagi data menjadi data pelatihan dan data pengujian secara terpisah (Dani & Ni'matuzzahroh, 2021).

Nilai GCV diperoleh berdasarkan keseimbangan antara tingkat kecocokan model terhadap data dan kompleksitas model yang digunakan. Model terbaik ditentukan berdasarkan nilai GCV minimum, karena nilai GCV yang lebih kecil menunjukkan bahwa model memiliki kemampuan prediksi yang lebih baik (Dani dkk., 2021).

Rumus *Generalized Cross Validation* (GCV) secara umum dinyatakan sebagai berikut:

$$GCV = \frac{MSE}{\left(1 - \frac{\text{trace}(H)}{n}\right)^2} \quad (5)$$

dengan:

MSE merupakan *Mean Square Error*,

H merupakan matriks hat dengan elemen

$$h_{ij} = \frac{K\left(\frac{x_i - x_j}{h}\right)}{\sum_{k=1}^n K\left(\frac{x_i - x_k}{h}\right)} \quad (6)$$

n merupakan banyaknya pengamatan.

Dalam penelitian ini, metode GCV digunakan untuk menentukan titik *knot* optimal pada regresi *Spline Truncated* dan bandwidth optimal pada regresi *Kernel Gaussian* sehingga diperoleh model terbaik dalam memodelkan jumlah penderita tuberkulosis di Indonesia tahun 2025.

Mean Square Error (MSE)

Mean Square Error (MSE) merupakan salah satu ukuran yang digunakan untuk mengevaluasi tingkat kesalahan suatu model regresi. MSE diperoleh dari rata-rata kuadrat selisih antara nilai aktual dengan nilai prediksi yang dihasilkan model. Nilai MSE yang kecil menunjukkan bahwa model memiliki kemampuan estimasi yang baik karena nilai prediksi mendekati nilai aktual. Oleh karena itu, MSE sering digunakan sebagai kriteria dalam pemilihan model terbaik pada analisis regresi nonparametrik (Hodson, 2022).

Rumus *Mean Square Error* (MSE) dapat dinyatakan sebagai berikut:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (7)$$

Koefisien Determinasi (R^2)

Koefisien determinasi (R^2) merupakan ukuran yang digunakan untuk mengetahui besarnya kemampuan variabel prediktor dalam menjelaskan variasi variabel respon pada model regresi. Nilai R^2 berada pada interval 0 sampai 1. Semakin besar nilai R^2 , maka semakin baik kemampuan model dalam menjelaskan variasi data pengamatan. Sebaliknya, nilai R^2 yang kecil menunjukkan bahwa kemampuan model dalam menjelaskan variasi data masih rendah (Chicco dkk., 2021). Koefisien determinasi (R^2) secara umum dapat dirumuskan sebagai berikut:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (8)$$

METODE

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari data Badan Pusat Statistik tahun 2025. Data yang digunakan terdiri dari prevalensi tuberkulosis (Y), jumlah rumah sakit umum (X_1), jumlah kasus HIV/AIDS (X_2), persentase akses sanitasi layak (X_3), dan persentase akses air minum layak (X_4). Berikut adalah langkah-langkah yang dilakukan dalam penelitian ini:

1. Mengumpulkan data dari laman Badan Pusat Statistik (BPS).
2. Menentukan variabel independen dan dependen.
3. Melakukan analisis statistika deskriptif untuk mengetahui gambaran umum data penelitian.
4. Membuat *scatter plot* antara variabel respon dan setiap variabel prediktor untuk mengetahui pola hubungan antar variabel.
5. Menyusun model regresi nonparametrik *Spline Truncated* dengan menentukan titik *knot* optimum menggunakan metode *Generalized Cross Validation* (GCV).
6. Menyusun model regresi nonparametrik *Kernel Gaussian* dengan menentukan nilai *bandwidth* yang optimal.
7. Membandingkan kinerja model *Spline Truncated* dan *Kernel Gaussian* menggunakan nilai GCV, *Mean Square Error* (MSE), dan koefisien determinasi (R^2).
8. Menentukan model terbaik berdasarkan nilai GCV dan MSE terkecil dan nilai R^2 terbesar.

HASIL DAN PEMBAHASAN

Statistik Deskriptif

Eksplorasi data bertujuan untuk memahami karakteristik mendasar dari data, sebaran setiap variabel, serta hubungan awal antara variabel respons dan variabel prediktor, sebelum dilanjutkan ke tahap pemodelan regresi nonparametrik dengan metode *spline* maupun *kernel*.

Tabel 1. Statistik Deskriptif Data

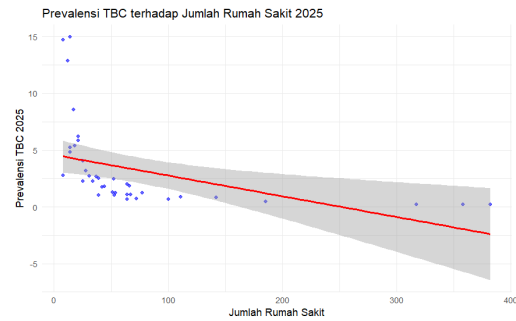
	Y	X ₁	X ₂	X ₃	X ₄
Min.	0,183	8	130	16,34	32,89
1st Qu	10,47	22	534,5	81,15	83,29
Median	19,22	43	839,5	85,99	89,94
Mean	32,54	72,89	1665,7	82,96	87,61
3rd Qu.	38,46	66,75	1789	90,12	94,9
Max.	149,8	382	10612	98,2	99,98

Tabel 1. menyajikan distribusi variabel penelitian. Statistik deskriptif dari data yang digunakan menunjukkan variasi yang signifikan pada prevalensi TBC dan faktor-faktor yang diduga memengaruhinya. Nilai prevalensi TBC (Y) memiliki rentang antara 0,1829 hingga 14,9762, dengan rata-rata 3,2543. Hal ini mencerminkan adanya perbedaan beban kasus TBC yang cukup lebar antar daerah. Median Y sebesar 1,9219 yang lebih kecil dari rata-rata mengindikasikan distribusi yang condong ke kanan, artinya sebagian besar daerah memiliki prevalensi TBC relatif rendah, namun ada beberapa daerah dengan nilai ekstrem tinggi. Variabel jumlah rumah sakit umum (X_1) menunjukkan variasi sangat tinggi, dari minimum 8 hingga maksimum 382 unit, dengan rata-rata 72,89 dan median 43. Jumlah kasus HIV/AIDS (X_2) memiliki sebaran paling ekstrem, dengan minimum 130 kasus dan maksimum 10612 kasus. Rata-rata berada jauh di atas median menandakan distribusi sangat menceng kanan akibat daerah-daerah dengan kasus HIV/AIDS yang sangat tinggi.

Sementara itu, persentase akses sanitasi layak (X_3) menunjukkan variasi yang relatif lebih sempit, nilai rata-rata yang hampir sama dengan median menunjukkan distribusi yang cenderung simetris. Adapun persentase akses air minum layak (X_4) memiliki nilai *mean* yang sedikit lebih rendah dari median mengindikasikan distribusi yang condong ke kiri, artinya sebagian besar daerah sudah memiliki akses air minum layak yang sangat tinggi, namun masih terdapat beberapa daerah dengan akses yang sangat rendah. Secara keseluruhan, distribusi variabel ini mencerminkan hubungan potensial antara ketersediaan fasilitas kesehatan, beban HIV/AIDS, akses sanitasi, serta akses air minum layak terhadap prevalensi TBC, di mana daerah dengan akses sanitasi dan air minum yang lebih baik serta jumlah kasus HIV/AIDS yang lebih rendah cenderung memiliki prevalensi TBC yang lebih kecil.

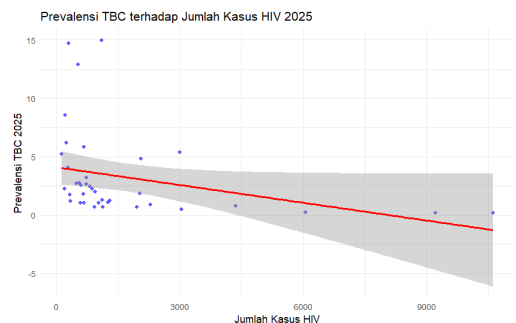
Pola Hubungan Antar Variabel

Berikut ini merupakan *scatter plot* hubungan variabel dependen dengan masing-masing variabel independen.



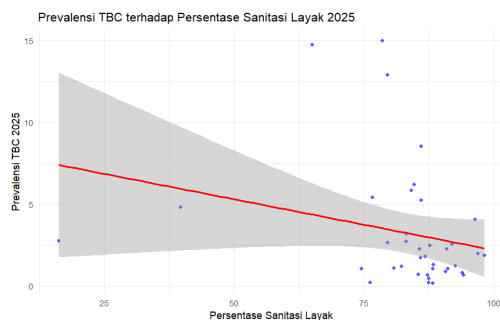
Gambar 1. Scatter Plot Variabel Y dan X_1

Pada Gambar 1, *scatter plot* menunjukkan hubungan antara Jumlah Rumah Sakit dengan prevalensi TBC. Pada sebaran titik-titik terlihat pola yang cenderung positif, yaitu semakin banyak jumlah rumah sakit, maka prevalensi TBC juga semakin tinggi. Hal ini dapat dijelaskan oleh beberapa kemungkinan. Pertama, rumah sakit sebagai fasilitas kesehatan dapat menjadi tempat akumulasi pasien TBC aktif, sehingga risiko penularan meningkat, baik antar pasien maupun antara pasien dan petugas kesehatan. Kedua, daerah dengan prevalensi TBC yang tinggi cenderung membutuhkan lebih banyak rumah sakit untuk menangani kasus, sehingga hubungan yang teramati bisa juga disebabkan oleh peningkatan akses dan deteksi. Namun, perlu diingat bahwa korelasi positif ini tidak serta-merta menunjukkan sebab akibat; faktor lain.



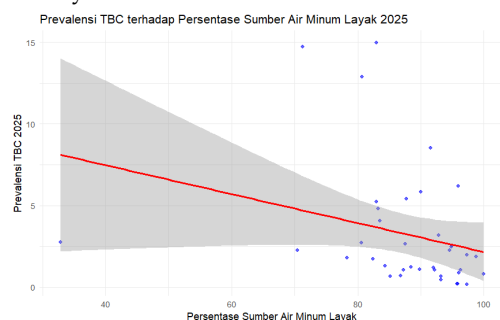
Gambar 2. Scatter Plot Variabel Y dan X_2

Pada Gambar 2, *scatter plot* menunjukkan hubungan antara jumlah kasus HIV dengan prevalensi TBC. Dari sebaran titik-titik tidak terlihat pola yang jelas, cenderung acak dan bahkan menunjukkan kecenderungan menurun pada jumlah kasus HIV yang tinggi. Sebagian besar titik berada pada prevalensi TBC yang rendah ketika jumlah kasus HIV melebihi 500. Ketidakhadiran pola positif yang jelas mengindikasikan bahwa jumlah kasus HIV saja tidak cukup untuk menjelaskan tingginya prevalensi TBC. Hal ini dapat dijelaskan karena ko-infeksi TB-HIV memang meningkatkan risiko TBC aktif, tetapi tidak bersifat linier sederhana.

Gambar 3. Scatter Plot Variabel Y dan X_3

Pada Gambar 3, *scatter plot* menunjukkan hubungan antara persentase sanitasi layak dengan prevalensi TBC. Dari sebaran titik-titik terlihat pola yang sangat jelas dan negatif, yaitu semakin tinggi persentase sanitasi layak, maka prevalensi TBC semakin rendah.

Pola negatif yang kuat ini dapat dijelaskan secara epidemiologis. Sanitasi yang layak secara signifikan mengurangi risiko penularan berbagai penyakit menular, termasuk TBC. Sanitasi yang buruk menciptakan lingkungan padat, lembap, dan tidak higienis yang mempermudah penularan bakteri *Mycobacterium Tuberculosis*. Hal ini mengindikasikan bahwa peningkatan akses terhadap sanitasi layak merupakan intervensi yang sangat potensial dalam upaya menurunkan angka TBC di suatu wilayah.

Gambar 4. Scatter Plot Variabel Y dan X_4

Pada Gambar 4, *scatter plot* menunjukkan hubungan antara persentase sumber air minum layak dengan prevalensi TBC. Titik menunjukkan pola yang sangat jelas dan negatif, yaitu semakin tinggi persentase sumber air minum layak, maka prevalensi TBC semakin rendah. Akses terhadap sumber air minum yang layak mengurangi risiko berbagai penyakit infeksi, termasuk TBC.

Pemodelan Tuberkulosis dengan *Spline Truncated*

Dilakukan pemodelan prevalensi Tuberkulosis (TBC) menggunakan regresi nonparametrik *Spline Truncated*. Diperoleh nilai GCV minimum yaitu

sebesar 3,588313 dengan lokasi titik *knot* yang optimal untuk masing-masing variabel prediktor, yaitu variabel jumlah rumah sakit umum (X_1) memiliki $K_1 = 23,5833$; variabel Jumlah kasus HIV/AIDS (X_2) memiliki $K_1 = 566,75$; variabel persentase akses sanitasi layak (X_3) memiliki $K_1 = 19,75083$; dan variabel persentase akses air minum layak (X_4) memiliki $K_1 = 35,68542$. Sehingga dapat dituliskan model regresi nonparametrik *Spline Truncated* dibatasi dengan 1 titik *knot* yang merupakan titik *knot* terbaik adalah:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 (x_{i1} - 23,5833)_+ + \beta_3 x_{i2} + \beta_4 (x_{i2} - 566,75)_+ + \beta_5 x_{i3} + \beta_6 (x_{i3} - 19,75083)_+ + \beta_7 x_{i4} + \beta_8 (x_{i4} - 35,68542)_+$$

Berikut merupakan hasil estimasi parameter model regresi nonparametrik *Spline Truncated* yang terbaik, yaitu model regresi nonparametrik *Spline Truncated* dengan 1 titik *knot*.

Tabel 2. Estimasi Parameter *Spline Truncated*

Variabel	Estimasi Parameter
Konstanta	-0,5351
X_1	-0,9169
	0,0017
X_2	4,2221
	-1,8028
X_3	0,9114
	-0,0017
X_4	-4,1399
	1,7784

Berdasarkan hasil estimasi parameter pada tabel di atas, maka dapat dituliskan model regresi nonparametrik *Spline Truncated* yang terbaik dengan 1 titik *knot* sebagai berikut.

$$\hat{y}_i = -0,5351 - 0,9169x_{i1} + 0,0017(x_{i1} - 23,5833)_+ + 4,2221x_{i2} - 1,8028(x_{i2} - 566,75)_+ + 0,9114x_{i3} - 0,0017(x_{i3} - 19,75083)_+ - 4,1399x_{i4} + 1,7784(x_{i4} - 35,68542)_+$$

Evaluasi kinerja model regresi dilakukan dengan menggunakan dua indikator utama, yaitu *Mean Squared Error* (MSE) dan koefisien determinasi (R^2). Hasil evaluasi menunjukkan bahwa nilai MSE yang diperoleh sebesar 2,2365, yang mencerminkan tingkat galat prediksi yang relatif rendah. Nilai MSE ini mengindikasikan bahwa model memiliki kemampuan estimasi yang baik dalam merepresentasikan hubungan antara variabel-variabel yang dianalisis. Selanjutnya, nilai koefisien

determinasi (R^2) yang diperoleh sebesar 83,89% menunjukkan bahwa proporsi variabilitas data yang dapat dijelaskan oleh model mencapai 83,89%, sementara sisanya sebesar 16,11% kemungkinan disebabkan oleh faktor-faktor lain yang tidak dimodelkan atau oleh keberadaan *noise* dalam data. Dengan demikian, dapat disimpulkan bahwa model regresi yang digunakan menunjukkan performa yang baik dan mampu memberikan representasi yang akurat terhadap struktur hubungan antar variabel prevalensi TBC dengan faktor-faktor yang mempengaruhinya.

Pemodelan Tuberkulosis dengan *Kernel Gaussian*

Selanjutnya berikut ini adalah hasil pemodelan prevelensi tuberkulosis menggunakan regresi nonparametrik *Kernel Gaussian*. Pemilihan parameter optimal sangat penting dalam *Kernel Gaussian* untuk memastikan model yang akurat dan efisien. Metode *Generalized Cross Validation* (GCV) digunakan untuk menentukan parameter *bandwidth* optimal dalam estimasi kernel.

Tabel 3. Estimasi Parameter *Spline Truncated*

Indeks	1	2	3	4	5
<i>Bandwith</i> 1	23.288	37.464	51.639	65.815	79.990
<i>Bandwith</i> 2	592.61	953.32	1314.0	1674.8	2035.5
<i>Bandwith</i> 3	3.8232	6.1503	8.4774	10.805	13.132
<i>Bandwith</i> 4	2.9682	4.7750	6.5818	8.3885	10.195
MSE	8.3112	9.9417	10.924	11.549	11.975
GCV	11.469	12.475	13.079	13.462	13.711

Tabel 3 menunjukkan hasil perhitungan *Generalized Cross Validation* (GCV) dengan menampilkan lima nilai parameter GCV pertama yang diperoleh. Berdasarkan tabel tersebut, nilai GCV minimum yang diperoleh adalah sebesar 11,46943. Penentuan *bandwidth* optimal dilakukan berdasarkan nilai GCV terkecil yang dihasilkan. Berdasarkan hasil perhitungan diperoleh *bandwidth* optimal yaitu 23,28834; 592,6081; 3,823165; 2,968246.

Bentuk umum dari estimator *Kernel Gaussian* adalah sebagai berikut.

$$\begin{aligned}
 k(x, x') = & \exp\left(-\frac{\|x_1, x'_1\|^2}{2 \times 23,28834^2}\right) \\
 & \times \exp\left(-\frac{\|x_2, x'_2\|^2}{2 \times 592,6081^2}\right) \\
 & \times \exp\left(-\frac{\|x_3, x'_3\|^2}{2 \times 3,823165^2}\right)
 \end{aligned}$$

$$\times \exp\left(-\frac{\|x_4, x'_4\|^2}{2 \times 2,968246^2}\right)$$

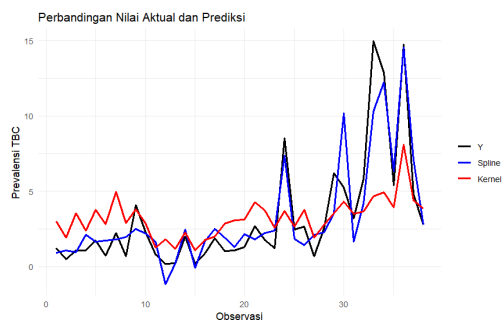
Kernel Gaussian digunakan untuk mengestimasi kepadatan probabilitas pada ruang multivariat. Dalam proses estimasi tersebut, pemilihan *bandwidth* memiliki peranan penting karena memengaruhi bentuk estimasi fungsi distribusi yang dihasilkan. Melalui pendekatan *Generalized Cross Validation* (GCV), diperoleh parameter *bandwidth* optimal yang mampu menghasilkan estimasi dengan keseimbangan antara bias dan variansi.

Evaluasi model regresi kernel berdasarkan nilai *Mean Squared Error* (MSE) dan koefisien determinasi (R^2) menunjukkan hasil yang kurang baik. Nilai MSE sebesar 8,311203 menunjukkan bahwa galat prediksi model relatif kecil, sedangkan nilai R^2 sebesar 40,13039% mengindikasikan bahwa model hanya mampu menjelaskan sebagian keragaman data. Dengan demikian, regresi kernel dinilai kurang efektif dalam memodelkan hubungan antarvariabel.

Model Terbaik

Berdasarkan nilai *Mean Squared Error* (MSE) dan koefisien determinasi (R^2), regresi *Spline Truncated* memiliki performa yang lebih baik dibandingkan regresi *Kernel Gaussian* dalam memodelkan prevalensi tuberkulosis. Model regresi *Spline Truncated* menghasilkan nilai MSE yang lebih rendah, yaitu 2,2365, serta nilai R^2 yang lebih tinggi sebesar 83,89%. Hasil ini menunjukkan bahwa model *Spline Truncated* memiliki tingkat kesalahan prediksi yang lebih kecil dan mampu menjelaskan keragaman data dengan lebih baik.

Sementara itu, regresi kernel menghasilkan nilai MSE sebesar 8,311203 dan nilai R^2 sebesar 40,13039%. Nilai tersebut menunjukkan bahwa kemampuan regresi kernel dalam menjelaskan variabilitas data prevalensi TBC masih lebih rendah dibandingkan regresi *Spline Truncated*. Dengan demikian, regresi *Spline Truncated* dapat dikatakan lebih unggul dalam memodelkan hubungan antarvariabel pada kasus prevalensi tuberkulosis. Ditampilkan visualisasi nilai $\hat{y}$ (prediksi) dengan y (nilai aktual) pada gambar berikut.



Gambar 5. Plot perbandingan nilai aktual dan prediksi kedua metode

Berdasarkan perbandingan visual antara nilai aktual dan nilai prediksi dari kedua metode, regresi *Spline Truncated* juga menunjukkan pola prediksi yang lebih stabil dan lebih mendekati data aktual. Sebaliknya, regresi kernel cenderung menghasilkan pola yang lebih berfluktuasi sehingga pada beberapa bagian kurang mampu mengikuti pola data sebenarnya. Oleh karena itu, regresi *Spline Truncated* dinilai lebih baik dalam menggambarkan pola hubungan data prevalensi TBC pada penelitian ini.

PENUTUP

SIMPULAN

Berdasarkan hasil analisis yang telah dilakukan, dapat ditarik kesimpulan bahwa secara deskriptif, prevalensi TBC di Indonesia tahun 2025 menunjukkan distribusi menceng ke kanan dengan variasi antar daerah yang cukup lebar. Variabel jumlah kasus HIV/AIDS memiliki sebaran paling ekstrem, sementara persentase akses sanitasi layak dan air minum layak cenderung lebih simetris. Pola hubungan antara variabel prediktor dengan prevalensi TBC menunjukkan bahwa jumlah rumah sakit berkorelasi positif, sedangkan persentase sanitasi layak dan air minum layak berkorelasi negatif yang kuat. Adapun jumlah kasus HIV tidak menunjukkan pola yang jelas.

Pemodelan regresi nonparametrik, metode *Spline Truncated* dengan satu titik *knot* optimal pada masing-masing prediktor menghasilkan nilai GCV minimum sebesar 3,588313, MSE sebesar 2,2365, dan R^2 sebesar 83,89%. Sementara itu, metode *Kernel Gaussian* dengan *bandwidth* optimal menghasilkan MSE sebesar 8,311203 dan R^2 sebesar 40,13%. Berdasarkan kedua indikator tersebut, metode *Spline Truncated* terbukti lebih unggul karena memiliki tingkat kesalahan prediksi yang lebih rendah dan kemampuan menjelaskan keragaman data yang jauh

lebih tinggi dibandingkan *Kernel Gaussian*. Plot perbandingan nilai aktual dan prediksi menunjukkan bahwa model *Spline Truncated* menghasilkan prediksi yang lebih stabil dan mendekati data aktual, sedangkan model *Kernel Gaussian* cenderung berfluktuasi dan kurang mampu mengikuti pola data sebenarnya.

SARAN

Bagi peneliti selanjutnya, disarankan untuk menguji penggunaan lebih dari satu titik *knot* pada metode *Spline Truncated* guna mengeksplorasi kemungkinan peningkatan akurasi model, serta mempertimbangkan penambahan variabel prediktor lain yang relevan seperti kepadatan penduduk, cakupan imunisasi, atau status gizi masyarakat. Bagi pengambil kebijakan kesehatan, hasil penelitian ini mengindikasikan bahwa peningkatan akses terhadap sanitasi layak dan air minum layak merupakan intervensi prioritas yang potensial dalam upaya menurunkan prevalensi TBC, sedangkan pendekatan statistik seperti *Spline Truncated* dapat digunakan sebagai alat bantu dalam memonitor dan mengevaluasi faktor-faktor risiko TBC secara lebih fleksibel.

DAFTAR PUSTAKA

- Abdy, M. (2019). Tinjauan Singkat Tentang Regresi Parametrik dan Regresi Non Parametrik. *SAINTIFIK*, 5 (1).
- Anwar, R., Rasiyanti, L., & Pitri, R. (2025). Analisis Regresi Kernel Gaussian untuk Memprediksi Indeks Pembangunan Manusia Berdasarkan Faktor Sosial-Ekonomi Provinsi di Indonesia. *Jurnal Riset Rumpun Matematika dan Ilmu Pengetahuan Alam*, 4(3), 44–53.
- Badan Pusat Statistik (BPS). (2025). Kasus Penyakit Menurut Provinsi dan Jenis Penyakit. <https://www.bps.go.id/id/statistics-table/3/YTA1Q1ptRmhUMEpXWTBsOmQyZzBjVzgwUzB4aVp6MDkjMyMwMDAw/kasus-penyakit-menurut-provinsi-dan-jenis-penyakit.html?year=2025>
- Badan Pusat Statistik (BPS). (2025). Jumlah Rumah Sakit Umum, Rumah Sakit Khusus, Puskesmas, Klinik Pratama, dan Posyandu Menurut Provinsi. <https://www.bps.go.id/assets/statistics-table/3/YmlzemNGUkNVblZLVVhObIREWnZXbkEzWld0eVVUMDkjMw==/jumlah-rumah-sakit-umum-rumah-sakit-khusus-puskesmas--klinik-pratama--dan-posyandu-menurut-provinsi--2018.html>

- Badan Pusat Statistik (BPS). (2025). Kasus Penyakit Menurut Provinsi dan Jenis Penyakit. <https://www.bps.go.id/id/statistics-table/3/YTA1Q1ptRmhUMEpXWTBsQmQyZzBjVzgwUzB4aVp6MDkjMyMwMDAw/Kasus-Penyakit-Menurut-Provinsi-Dan-Jenis-Penyakit.Html>.
- Badan Pusat Statistik (BPS). (2025). Persentase Rumah Tangga Menurut Provinsi dan Memiliki Akses Terhadap Sanitasi Layak. <https://www.bps.go.id/id/statistics-table/2/ODQ3lzl=/persentase-rumah-tangga-menurut-provinsi-dan-memiliki-akses-terhadap-sanitasi-layak.html>
- Badan Pusat Statistik (BPS). (2025). Persentase Rumah Tangga Menurut Provinsi dan Sumber Air Minum Layak. <https://www.bps.go.id/id/statistics-table/2/ODQ1lzl=/persentase-rumah-tangga-menurut-provinsi-dan-sumber-air-minum-layak.html>
- Chicco, D., Warrens, M. J., & Jurman, G. (2021). *The Coefficient of Determination R-Squared is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation*. *PeerJ Computer Science*, 7, e623.
- Dani, A. T. R., & Ni'matuzzahroh, L. (2021). Pemodelan persentase Penduduk Miskin Kabupaten/Kota di Provinsi Jawa Barat dengan Pendekatan Regresi Nonparametrik Spline Truncated. *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika*, 14(1).
- Dani, A. T. R., Ni'matuzzahroh, L., Ratnasari, V., & Budiantara, I. N. (2021). Pemodelan Regresi Nonparametrik Spline Truncated pada Data Longitudinal (Studi Kasus: Persentase Penduduk Miskin Provinsi Papua tahun 2016-2019). *Jurnal Inferensi*, 4(1).
- Dewi, A. F., & Ahadiyah, K. (2023). Analisis Regresi Spline Truncated pada Indeks Pembangunan Manusia Jawa Timur tahun 2021. *INFERENSI*, 6(1), 1-10.
- Efendi, M., Budiantara, I. N., & Ratnasari, V. (2025). Pemilihan Bandwidth Optimal pada Regresi Nonparametrik Kernel Menggunakan Metode *Unbiased Risk* (UBR). *Jurnal Gaussian*, 14(2), 619-630.
- Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. Marcel Dekker.
- Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.
- Härdle, W., Müller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. Springer.
- Hodson, T. O. (2022). *Root-Mean-Square Error (RMSE) or Mean Absolute Error (MAE): When to Use Them or Not*. *Geoscientific Model Development*, 15, 5481-5487.
- Kusumaningrum, A., Wulandari, G., & Kautsar, A. (2023). Tuberkulosis di Indonesia: Apakah status Sosial Ekonomi dan Faktor Lingkungan Penting?. *Jurnal Ekonomi dan Pembangunan Indonesia*, 23(1), 1-14.
- Meliasari. (2021). Terapi Tuberkulosis. *Jurnal Medika Utama*, 3(1), 1571-1575.
- Naufal, M. J., Ompusunggu, D. P., Sinaga, R. A., Sitohang, M. D. A., Gunawan, T. N., Simatupang, M., Salsabila, N. S., Simanullang, T., & Hutasoit, B. T. (2025). *A Theoretical Study Of Multicollinearity and Linearity in Econometric Models for Economic Research*. *Balance Jurnal Ekonomi*, 21(1), 43-52.
- Nugroho, S. (2008). *Statistika Nonparametrik*. UNIB Press.
- Pratama, M. H., Sifriyani, & Nohe, D. A. (2021). Regresi Nonparametrik Multivariabel dengan pendekatan spline truncated pada Kasus Tuberculosis. *Statistika*, 22(1).
- Rahmah, C. (2020). Studi Literatur Pengaruh Kondisi Rumah yang Kumuh Terhadap Kejadian Penderita Tuberculosis Paru. *Jurnal Ilmu Kedokteran dan Kesehatan*, 7(4), 593-598.
- Ratnasari, V., Budiantara, I. N., & Dani, A. T. R. (2021). *Nonparametric Regression Mixed Estimators of Truncated Spline and Gaussian Kernel Based on Cross-Validation (CV), Generalized Cross-Validation (GCV), and Unbiased Risk (UBR) Methods*. *International Journal on Advanced Science, Engineering and Information Technology*, 11(6). [\[https://doi.org/10.18517/ijaseit.11.6.14464\]](https://doi.org/10.18517/ijaseit.11.6.14464) (<https://doi.org/10.18517/ijaseit.11.6.14464>)
- Sudirman, Kolupe, V. M., Assa, A. A. Y., Purwiningsih, S., & Susianawati, D. E. (2024). Implementasi Program Pengendalian TB Paru: Studi Kualitatif di Puskesmas Parigi. *Jurnal Promotif Preventif*, 7(4), 694-703.
- World Health Organization. (2023). *Global Tuberculosis Report 2023*.