

BAB I PENDAHULUAN

1.1 Latar Belakang

Analisis regresi merupakan metode yang umum digunakan untuk mengubah suatu permasalahan nyata menjadi persamaan matematis yang menjelaskan pola hubungan antara variabel respon dan variabel prediktor. Ketika data mengikuti pola tertentu, analisis regresi parametrik dapat digunakan. Namun, sering kali dijumpai permasalahan bahwa data tidak mengikuti pola linear maupun bentuk kurva tertentu. Ketidaksesuaian ini dapat menyebabkan hasil estimasi menjadi bias atau kurang akurat, sehingga regresi nonparametrik dipilih karena mampu memodelkan hubungan yang kompleks dan fleksibel antara variabel respon dan prediktor.

Model regresi nonparametrik memuat asumsi bahwa bentuk kurva regresi f tidak diketahui, tetapi hanya diasumsikan termuat dalam satu ruang fungsi tertentu (Islamiyati, 2017). Pengembangan analisis regresi nonparametrik telah banyak dilakukan oleh para peneliti untuk meningkatkan akurasi dan efisiensi model, antara lain, Regresi Nonparametrik Menggunakan Fungsi Kernel (Widiardi, 2014), Estimasi Model Regresi Nonparametrik Dengan Metode B-Spline (Salam, et al. 2022), *Regresi Adaptif Multivariat Splines* (Lewis dan Stevens, 1990 ; Friedman, 1991; Otok, 2010) B-spline (Budiatara, 2006; Lisnaini, 2014; Suwarno, 2023), Spline Polynomial Truncated (Islamiyati, 2017; Dani, 2017), Truncated Spline (Islamiyati, 2022; Andrianzah, 2023; Khotijah, 2020, Wongkar, 2023; Syamsualam, 2022), Smoothing spline (Jayanti, 2018; Wati, 2014; Khoiurnissa, 2016), Deret Fourier (Mariati, 2015). Meskipun metode-metode tersebut mampu menangani data dengan pola nonlinear, sebagian masih memiliki keterbatasan, misalnya kurang mampu memodelkan hubungan saling memengaruhi antarvariabel, membutuhkan parameter penghalus pada smoothing parameter yang sulit ditentukan secara optimal, atau pemilihan titik knot yang dilakukan secara manual. Kondisi-kondisi inilah yang mendorong pendekatan regresi nonparametrik secara adaptive banyak diminati. Salah satunya adalah model Multivariate Adaptive Regression Spline (MARS) yang menghasilkan prediksi variabel respon yang akurat dan menghasilkan model yang kontinu pada knot (Friedman, 1991).

Metode MARS merupakan pengembangan dari *Recursive Partition Regression* (RPR) yang dipadukan dengan teknik spline. Metode RPR awalnya masih memiliki kelemahan karena model yang dihasilkan tidak bersifat kontinu pada titik knot (Tamonob, 2020). Melalui Penggabungan Pengembangan RPR dengan metode *Spline*, dihasilkan model yang lebih baik, yaitu model yang tetap kontinu



ingga garis regresi selalu menyambung dan setiap titik knot ng dengan fungsi basisnya (Friedman, 1991). Salah satu ARS adalah proses penentuan knot dapat dilakukan secara oritma stepwise. Tahap Forward stepwise digunakan untuk dengan jumlah fungsi basis maksimum, sedangkan tahap digunakan untuk mengurangi jumlah fungsi basis yang

kontribusinya kecil terhadap estimasi variabel respons, dengan mempertimbangkan nilai Generalized Cross Validation (GCV)(Friedman, 1991). MARS merupakan metode regresi nonparametrik multivariat yang dikembangkan untuk menangani permasalahan berdimensi tinggi, yaitu data yang memiliki jumlah variabel prediktor $3 \leq x \leq 20$, dimana x adalah banyaknya variabel. Metode ini juga sesuai untuk ukuran sampel antara $50 \leq n \leq 1000$ dimana n adalah ukuran sampel, sehingga mampu menghasilkan prediksi variabel respons yang lebih akurat (Friedman, 1991).

Pada penelitian sebelumnya yang menggunakan metode MARS diperoleh hasil bahwa ketepatan klasifikasi pada model MARS menghasilkan 99,55% pada rata-rata data training dan 99,33% pada rata-rata data testing. Pada penelitian ini juga membandingkan metode MARS dengan metode Analisis Diskriminan, Neural Network, Diskriminan Kernel dan Regresi Logistik. Dimana diperoleh kesimpulan bahwa metode MARS adalah metode yang paling baik karena ketepatan klasifikasinya paling tinggi diantara metode yang lain pada data penelitian tersebut.

Berbicara tentang regresi nonparametrik khususnya metode MARS yang melibatkan banyak prediktor, seringkali dijumpai kondisi dimana suatu variabel prediktor saling berkorelasi dan multikolinearitas di antara prediktor tersebut meningkat. Multikolinearitas merupakan kondisi ketika variabel-variabel independen dalam suatu model regresi saling berkorelasi atau memiliki hubungan linier satu sama lain. Istilah ini pertama kali diperkenalkan oleh Ragnar Frisch untuk menggambarkan adanya hubungan linier yang sempurna atau hampir sempurna di antara variabel-variabel prediktor(Gunawan et al. 2024). Salah satu metode statistik yang umum digunakan untuk mengatasi permasalahan ini adalah Principal Component Analysis(Islamiyati et al., 2022). Dalam penelitiannya, Islamiyati et.al menerapkan PCA untuk mengatasi multikolinearitas pada memodelkan regresi studi longitudinal dalam analisis data gula darah pasien Diabetes Melitus.

Metode ini berfungsi untuk mengidentifikasi dan menampilkan tren dengan variasi maksimum dalam data pada subruang baru yang memiliki dimensi lebih kecil daripada dimensi asli. Dengan menyertakan komponen utama yang mencakup sekitar 75% variasi data, akan dapat lebih mudah menginterpretasikan dan menganalisis data (Sriningsih, et al., 2018). Pengolahan data yang kompleks ini sering kali dihadapkan pada tantangan seperti multikolinearitas dan overfitting serta metode analisis tradisional seperti regresi linier sering kali tidak mampu menangkap hubungan non-linier dan interaksi kompleks antar variabel, yang dapat secara signifikan memengaruhi hasil analisis. Dalam konteks ini, *Principal Component Analysis* muncul sebagai teknik statistik yang efektif untuk mereduksi data yang berdimensi tinggi tanpa kehilangan informasi penting.



n metode PCA-MARS telah dilakukan oleh beberapa peneliti. dilakukan oleh Azid et al. (2014) menerapkan PCA-ANN untuk Jara di Malaysia. Sadra et al. (2023) menerapkan PCA-MARS iruan untuk memprediksi efisiensi sistem desulfurisasi, dan n kinerja yang lebih baik dibandingkan metode regresi biasa. al. (2022) menggunakan PCA-MARS untuk memprediksi hasil nemukan bahwa metode ini lebih cocok untuk menganalisis

hubungan non-linear yang kompleks. Penelitian oleh Mohammadi et al. (2023) menunjukkan bahwa rPCA-MARS lebih efektif dalam analisis data spektral dibandingkan model PLS-R dan SVM-R. Dengan demikian, PCA-MARS merupakan metode yang menjanjikan untuk menangani berbagai tantangan dalam analisis data.

Pendapatan perkapita pada awalnya dijadikan sebagai indikator penting dari makro ekonomi suatu negara oleh beberapa pakar sosial ekonomi, akan tetapi hal itu tidak menggambarkan secara keseluruhan tingkat kesejahteraan dan kemakmuran masyarakat. Tingginya pendapatan per kapita tidak selalu mencerminkan bahwa penduduk suatu negara hidup sejahtera, karena kesejahteraan tidak semata-mata diukur dari aspek ekonomi. Sehingga hal itu dinilai kurang tepat dalam menilai kemajuan suatu bangsa. Hal inilah yang mendorong munculnya indikator-indikator baru yang lebih menitikberatkan pada aspek pembangunan manusia (Azfirmawarman, et al, 2023). Indeks Pembangunan Manusia (IPM) merupakan indikator penting yang digunakan untuk menilai keberhasilan pembangunan dalam meningkatkan kualitas hidup masyarakat. IPM menunjukkan sejauh mana penduduk dapat memperoleh manfaat pembangunan, khususnya dalam aspek pendapatan, kesehatan, pendidikan, dan aspek sosial lainnya. Konsep IPM pertama kali diperkenalkan oleh UNDP pada tahun 1990 dan sejak itu dipublikasikan secara rutin melalui laporan tahunan *Human Development Report* (HDR) (Patiung & Ernawati, 2024).

IPM adalah indikator yang menunjukkan tingkat kualitas hidup masyarakat melalui tiga dimensi utama, yaitu kesehatan, pendidikan dan standar hidup yang layak. Nilai IPM kerap digunakan sebagai acuan untuk menilai keberhasilan pembangunan suatu daerah, termasuk dalam melihat tingkat kesejahteraan serta pemerataan antarprovinsi di Indonesia. Namun, dinamika faktor-faktor yang memengaruhi IPM sangat kompleks dan saling berkaitan satu sama lain, seperti rata-rata lama sekolah, pengeluaran perkapita, dan angka harapan hidup. Selain itu, ketiga dimensi dalam Indeks Pembangunan Manusia saling memengaruhi satu sama lain. Kondisi tersebut dipengaruhi oleh berbagai faktor, seperti ketersediaan lapangan kerja yang berperan penting dalam mendorong pertumbuhan ekonomi di masa mendatang, serta dukungan sarana, prasarana, dan program pembangunan yang direncanakan oleh pemerintah. Dengan demikian, semakin berhasil suatu daerah dalam meningkatkan ketiga dimensi tersebut, maka semakin tinggi pula nilai IPM yang dicapai, yang pada akhirnya tercermin melalui pertumbuhan sektor ekonominya.

Berdasarkan uraian diatas, maka penulis menyusunnya dalam sebuah penelitian dengan judul Analisis Efektivitas Metode *Principal Component Analysis-*



Regression Splines Pada Data Indeks Pembangunan Manusia. Penelitian yang dilakukan oleh Nayana et al. (2022) yang menggunakan PCA-MARS untuk memprediksi hasil panen gandum yang menggunakan data Time Series, penulis akan menggunakan data Cross Section pembangunan Manusia di Indonesia tahun 2023.

1.2 Rumusan Masalah

Berikut rumusan masalah berdasarkan latar belakang yang telah diuraikan:

1. Bagaimana mengestimasi parameter model *Principal Component Analysis-Multivariate Adaptive Regression Spline* (PCA-MARS) ?
2. Bagaimana memodelkan data Indeks Pembangunan Manusia dengan *Principal Component Analysis-Multivariate Adaptive Regression Spline* (PCA-MARS)?

1.3 Batasan Masalah

Penelitian ini hanya menggunakan data data Indeks Pembangunan Manusia (IPM) yang dikumpulkan dari situs Badan Pusat Statistik pada tahun 2023. Penelitian hanya berfokus pada bagaimana memodelkan faktor-faktor penyusun Indeks Pembangunan Manusia yang di optimasi dengan *Principal Component Analysis*.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah diuraikan, maka tujuan dari penelitian ini adalah sebagai berikut:

1. Mengetahui estimasi parameter yang dihasilkan pada model *Principal Component Analysis-Multivariate Adaptive Regression Spline* (PCA-MARS)
2. Mengevaluasi memodelkan data Indeks Pembangunan Manusia dengan *Principal Component Analysis-Multivariate Adaptive Regression Spline* (PCA-MARS)

1.5 Manfaat Penelitian

Penelitian ini diharapkan mampu memberikan kontribusi berarti dalam memperluas pemahaman terkait penerapan metode *Principal Component Analysis - Multivariate Adaptive Regression Spline* (PCA-MARS), sehingga dapat menjadi rujukan bagi penelitian selanjutnya. Selain itu, studi ini juga memberikan gambaran yang lebih komprehensif mengenai Indeks Pembangunan Manusia (IPM), yang dapat dimanfaatkan sebagai dasar dalam penyusunan kebijakan untuk meningkatkan kualitas pembangunan manusia di Indonesia.

1.6 Kajian Teori

1.6.1 Multikolinearitas

Multikolinearitas merupakan kondisi ketika variabel-variabel independen dalam suatu model regresi saling berkorelasi atau memiliki hubungan linier satu sama lain. Istilah ini pertama kali diperkenalkan oleh Ragnar Frisch untuk menggambarkan adanya hubungan linier yang sempurna atau hampir sempurna di antara variabel prediktor dalam model. Multikolinearitas mencerminkan keberadaan



gan linier antarvariabel, sedangkan kolinearitas hanya merujuk a linier tunggal. Kondisi ini dapat memengaruhi ketepatan regresi (Gunawan et al., 2024).

erapa metode yang dapat digunakan untuk mendeteksi ni dengan nilai faktor inflasi ragam (Variance Inflation Factor / nghitung nilai koefisien korelasi sederhana (*simple correlation*) iktor. VIF digunakan sebagai kriteria untuk mendeteksi

multikolinearitas pada regresi linier yang melibatkan lebih dari dua variabel bebas. Nilai VIF lebih besar dari 10 mengidentifikasi adanya masalah multikolinearitas yang serius (Ryan, 1997). Nilai VIF dapat diperoleh dengan rumus:

$$VIF_j = \frac{1}{1 - R_j^2} \quad (1)$$

dengan R_j^2 adalah koefisien determinasi yang diperoleh jika X_j diregresikan dengan $(p - 1)$ variabel lain.

Nilai koefisien korelasi sederhana (*simple correlation*) antar variabel prediktor Berdasarkan matriks korelasi R sebagai berikut:

$$R = \frac{1}{n - 1} X^T X \quad (2)$$

dengan X merupakan variabel prediktor yang telah di standarisasi dan n adalah banyaknya sample. Jika terdapat koefisien korelasi sederhana yang mencapai atau melebihi 0,8 dan atau mencapai kurang dari -0,8, maka hal tersebut menunjukkan terjadinya masalah multikolinearitas dalam regresi (Ridwan & Sunendiari, 2021).

1.6.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) yang pertama kali diperkenalkan oleh Pearson pada tahun 1901, merupakan teknik analisis multivariat yang mengubah variabel-variabel awal yang saling berkorelasi menjadi seperangkat variabel baru yang saling tidak berkorelasi. Metode ini bertujuan untuk mereduksi jumlah variabel sehingga menghasilkan dimensi yang lebih sederhana namun tetap mampu mempresentasikan sebagian besar keragaman dari variabel aslinya (Susilowati, 2021). PCA memungkinkan kita untuk melihat struktur laten yang umum dalam kumpulan data yang besar. Inti dari PCA adalah mengidentifikasi sejumlah kecil komponen utama yang secara efektif merangkum sebagian besar variasi dalam data, sehingga mengurangi dimensi tanpa kehilangan informasi (Larasati et al 2021).

PCA akan mereduksi sekumpulan variabel yang saling berkorelasi menjadi variabel baru yang saling bebas atau tidak saling berkorelasi. Variabel baru tersebutlah yang disebut dengan komponen utama. Misalkan vektor acak $X' = [X_1, X_2, \dots, X_p]$ dengan matriks varians kovarians Σ dengan nilai eigen $\lambda_1 \geq \lambda_2 \geq \lambda_p \geq 0$ dan vektor eigen $e' = (e_1, e_2, \dots, e_p)$ maka kombinasi liniernya yaitu:

$$\begin{aligned} W_1 &= e_{11}X_1 + e_{12}X_2 + \dots + a_{1p}X_p \\ W_2 &= e_{21}X_1 + e_{22}X_2 + \dots + a_{2p}X_p \\ &\vdots \\ &= e_{p1}X_1 + e_{p2}X_2 + \dots + a_{pp}X_p \end{aligned} \quad (3)$$



into (1988) dalam Sriningsih, et al. (2018) terdapat beberapa ketentuan jumlah komponen utama, yaitu jumlah komponen utama yang membentuk model ditentukan dengan memilih nilai eigen dari satu karena nilai eigen yang hampir nol dianggap tidak berpengaruh signifikan. Kriteria ini digunakan ketika analisis faktor

didasarkan pada matriks korelasi. Selain itu Sriningsih, et al. (2018) menyarankan untuk memilih komponen-komponen utama yang memiliki keragaman kumulatif sebesar 75%.

1.6.7 *Multivariate Adaptive Regression Splines (MARS)*

Multivariate Adaptive Regression Splines (MARS) adalah metode regresi nonparametrik multivariat yang dikembangkan oleh Jerome Friedman pada tahun 1991. Metode ini dirancang untuk menangani permasalahan berdimensi tinggi serta mengatasi diskontinuitas pada data, sehingga mampu memberikan prediksi variabel respons yang lebih akurat (Friedman, 1991). MARS merupakan pengembangan dari metode *Recursive Partition Regression (RPR)*, yang sebelumnya memiliki kelemahan karena model yang dihasilkan tidak bersifat kontinu pada titik knot (Tamonob, 2020).

Dengan pendekatan ini, MARS mampu menggabungkan segmen linear untuk menciptakan model regresi piecewise yang lebih akurat, mencerminkan struktur data yang mendasarinya. Selain itu, MARS menggunakan serangkaian segmen linear atau kubik untuk merepresentasikan hubungan non-linear antara variabel input dan output. Metode ini memungkinkan eksplorasi hubungan non-linear dengan fleksibilitas lebih, serta mempertimbangkan berbagai tingkat interaksi antar variabel, sehingga mampu mengidentifikasi pola tersembunyi dalam data berdimensi tinggi (Putra, et al., 2021).

Menurut Nash dan Brandford (2001), terdapat beberapa aspek penting yang perlu diperhatikan dalam proses pembangunan model MARS, yaitu:

- a. *Knot*, yaitu titik yang menandai berakhirnya satu segmen garis regresi dan dimulainya segmen berikutnya. Pada setiap knot, diharapkan terjadi kontinuitas sehingga fungsi basis dari satu region dapat tersambung dengan region lainnya.
- b. *Basis Function*, yaitu fungsi yang digunakan untuk menggambarkan hubungan antara variabel respons dan variabel prediktor. Fungsi ini dapat dipahami sebagai sekumpulan fungsi yang mewakili informasi dari satu atau lebih variabel, termasuk kemungkinan interaksi antarvariabel. Fungsi basis terbentuk dari jarak antara knot-knot yang berurutan (Cherkassky & Mulier, 1998).
- c. *Maximum interaction (MI)* yang umumnya ditetapkan pada tingkat 1,2 dan 3 dengan pertimbangan jika dari 3 akan menghasilkan bentuk model yang semakin kompleks.
- d. *Minimum Observation (MO)* jarak antara knots atau minimum observasi antara knots sebesar 0,1,2,3, ... sampai maksimum jumlah observasi (Fatmawati, 2017).

Model *Multivariate Adaptive Regression Splines (MARS)*, yang merupakan hasil pengembangan dari *Recursive Partitioning Regression (RPR)* dengan kombinasi (Friedman, 1991), dapat dijelaskan sebagai berikut.



$$x_0 + \sum_{m=1}^M a_m \prod_{k=1}^{K_m} [S_{km}(x_{v(k,m)} - t_{km})]_+ + \varepsilon \quad (4)$$

dan (4), model MARS dapat dituliskan sebagai berikut:

$$\hat{f}(x) = a_0 + \sum_{m=1}^M a_m B_m(x) + \varepsilon \quad (5)$$

dengan Bentuk fungsi basis ke- m adalah sebagai berikut:

$$B_m(x) = \prod_{k=1}^{K_m} [S_{km}(x_{v(k,m)} - t_{km})]_+ \quad (6)$$

Jika $S_{km} = +1$, maka $[(x_{v(k,m)} - t_{km})]_+ = \begin{cases} x_{v(k,m)} - t_{km}, & \text{jika } x_{v(k,m)} > t_{km} \\ 0, & \text{sebaliknya} \end{cases}$

Jika $S_{km} = -1$, maka $[-(x_{v(k,m)} - t_{km})]_+ = \begin{cases} t_{km} - x_{v(k,m)}, & \text{jika } t_{km} > x_{v(k,m)} \\ 0, & \text{sebaliknya} \end{cases}$

K_m menunjukkan jumlah fungsi linier terpotong yang dikalikan dalam setiap fungsi basis ke- m . $x_{v(k,m)}$ adalah vektor dari variabel independen, sedangkan $t_{k,m}$ merupakan nilai knot (titik pemisah), dan $S_{km} = \pm 1$ (Friedman, 1991).

Estimasi untuk $\{a_m\}_{m=0}^M$ ditentukan dengan menggunakan metode kuadrat terkecil, mengingat masing-masing basis fungsi sudah ortogonal dan saling bebas. Dengan menggunakan estimator MARS, didapatkan persamaan regresi:

$$y_i = a_0 + \sum_{m=1}^M a_m \prod_{k=1}^{K_m} [S_{km}(x_{v(k,m)} - t_{km})]_+ + \varepsilon_i \quad (4)$$

dengan,

a_0 : konstanta regresi dari fungsi basis

a_m : koefisien dari fungsi basis ke- m

M : maksimum fungsi basis (nonconstant fungsi basis)

k_m : derajat interaksi

S_{km} : bernilai +1 jika knot terletak di kanan titik knot, dan bernilai -1 jika knot terletak di kiri titik knot

$x_{v(k,m)}$: variabel prediktor

v : banyaknya variabel prediktor

m : banyaknya basis fungsi

t_{km} : nilai knot dari variabel prediktor $x_{v(k,m)}$

k : banyaknya interaksi

Sehingga dapat dituliskan dalam bentuk matriks menjadi:

$$Y = B_a + \varepsilon \quad (5)$$



on

$_{km}^K]$

s regresi

dengan,

$$Y = (y_1, \dots, y_n)^T, \quad a = (a_1, \dots, a_M)^T, \quad \varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$$

$$B = \begin{bmatrix} 1 & \prod_{k=1}^{K_1} [S_{1m}(x_{1(1,m)} - t_{1m})]_+ & \dots & \prod_{k=1}^{K_M} [S_{Mm}(x_{v(M,m)} - t_{Mm})]_+ \\ 1 & \prod_{k=1}^{K_1} [S_{1m}(x_{2(1,m)} - t_{1m})]_+ & \dots & \prod_{k=1}^{K_M} [S_{Mm}(x_{2(M,m)} - t_{Mm})]_+ \\ \vdots & \vdots & \dots & \vdots \\ 1 & \prod_{k=1}^{K_1} [S_{1m}(x_{n(1,m)} - t_{1m})]_+ & \dots & \prod_{k=1}^{K_M} [S_{Mm}(x_{n(M,m)} - t_{Mm})]_+ \end{bmatrix}$$

- Y : vektor variabel respon berukuran $(n \times 1)$
 a : vektor koefisien regresi yang berukuran $((M + 1) \times 1)$
 B : matriks basis bungsi yang berukuran $(n \times (M + 1))$
 ε : vektor error yang berukuran $(n \times 1)$

Untuk memperoleh estimator $\hat{a}$ digunakan dengan emnggunakan metode kuadrat terkecil yang pada dasarnya dari upaya untuk meminimumkan jumlah kuadrat errornya atau Sum Square Error (SSE). Yaitu dengan cara mengkuadratkan persamaan berikut:

$$\varepsilon = Y - B\alpha$$

$$\begin{aligned} SSE &= \sum_{i=1}^n \varepsilon_i^2 = \varepsilon^T \varepsilon = (Y - B\alpha)^T (Y - B\alpha) \\ &= (Y^T - \alpha^T B^T)(Y - B\alpha) \\ &= Y^T Y - Y^T B\alpha - \alpha^T B^T Y + \alpha^T B^T B\alpha \\ &= Y^T Y - 2\alpha^T B^T Y + \alpha^T B^T B\alpha \end{aligned}$$

Untuk mendapatkan persamaan normal dilakukan dengan menurunkan atau melakukan diferensial SSE secara parsial terhadap α . Diperoleh hasil sebagai berikut:



$$\partial^T B\alpha = 0$$

$$\hat{\alpha} = (B^T B)^{-1} B^T Y$$

Metode MARS memulai dengan model yang hanya berisi nilai intersep, yaitu rata-rata dari nilai respon. Kemudian, metode ini mengevaluasi setiap prediktor untuk menemukan pasangan fungsi basis yang terdiri dari sisi-sisi berlawanan dari fungsi engsel cermin yang paling mengurangi kesalahan model. Prosedur ini diulang hingga jumlah istilah atau perubahan dalam tingkat kesalahan melampaui batas maksimum yang telah ditentukan sebelumnya.

1.6.8 Evaluasi Model Regresi

Evaluasi model regresi dilakukan untuk menilai sejauh mana model yang dihasilkan mampu menjelaskan variasi pada variabel respons serta menilai tingkat kecocokan model terhadap data. Evaluasi model dilakukan menggunakan dua ukuran utama, yaitu *Generalized Cross Validation* (GCV) dan Koefisien Determinasi R^2 , yang dijelaskan sebagai berikut:

a) *Generalized Cross Validation*

Pada model MARS, pemilihan model menggunakan metode stepwise yang terdiri dari forward dan backward. Forward stepwise dilakukan untuk mendapatkan jumlah basis fungsi maksimum dengan kriteria pemilihan basis fungsi adalah meminimumkan Average Sum Of Square Residual (ASR). Sedangkan backward stepwise digunakan untuk memenuhi prinsip parsimoni (model yang sederhana) dengan cara mengeliminasi fungsi basis yang berkontribusi kecil. Ukuran kontribusi pada backward stepwise ditentukan dengan kriteria *Generalized Cross Validation* (GCV) (Pintowati & Otok, 2012). Bentuk fungsi GCV minimum menurut Friedman (1991) didefinisikan pada persamaan berikut:

$$GCV = \sum_{i=1}^N \frac{(y_i - f(x_i))^2}{(1 - \frac{C}{N})^2} \quad (8)$$

dimana $C = 1 + cd$, N adalah jumlah observasi dalam dataset, d adalah derajat kebebasan, dan c adalah penalti untuk penambahan fungsi dasar. y_i merupakan variabel dependen, sedangkan $f(x_i)$ adalah nilai prediksi model untuk observasi ke- i (Nayana, 2022).

b) Koefisien Determinasi (R^2)

Koefisien determinasi merupakan salah satu patokan yang biasa digunakan untuk mengukur seberapa baik suatu model regresi yang dicocokkan belum atau sudah disesuaikan dengan (R^2). Koefisien determinasi (R^2) adalah persentase variasi dalam variabel respons yang dijelaskan oleh garis regresinya adalah

$$R^2 = 1 - \frac{SSE}{SST} \quad (9)$$



dimana $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ adalah jumlah kuadrat dari selisih antara nilai prediksi dan nilai yang diamati, sedangkan $SST = \sum_{i=1}^n (y_i - \bar{y})^2$ adalah jumlah kuadrat dari selisih antara nilai yang diamati dengan nilai rata-rata keseluruhan.

1.6.9 Indeks Pembangunan Manusia

Gagasan tentang Indeks Pembangunan Manusia (IPM) telah berkembang menjadi tolok ukur yang sangat diperlukan untuk memahami kondisi sosial-ekonomi di berbagai negara dan wilayah. Indeks Pembangunan Manusia (IPM) diperkenalkan pada tahun 1990 oleh Program United Nations Development Programme (UNDP) dengan memasukkan semua aspek pembangunan ekonomi ke dalam satu atap. Sementara indikator ekonomi tradisional seperti Produk Domestik Bruto (PDB) hanya berfokus pada kinerja ekonomi. IPM berfokus pada kualitas hidup. IPM adalah indeks gabungan yang menggabungkan tiga indikator dasar yakni angka harapan hidup, pencapaian pendidikan (rata-rata lama sekolah dan harapan lama sekolah), dan GNI per kapita ke dalam ukuran untuk menangkap kualitas hidup (Dey., et al., 2024).

Indeks ini terbentuk dari rata-rata ukur capaian tiga dimensi utama pembangunan manusia, yaitu umur panjang dan hidup sehat, pengetahuan, dan standar hidup layak. Dimensi umur panjang dan hidup sehat diukur dengan umur harapan hidup saat lahir. Dimensi pengetahuan diukur dengan rata-rata lama sekolah penduduk berusia 25 tahun ke atas dan harapan lama sekolah penduduk yang berumur 7 tahun. Sementara itu, dimensi standar hidup layak diukur dengan pengeluaran riil per kapita yang disesuaikan. Sejak tahun 2010 IPM Indonesia terus tumbuh secara konsisten. Perlambatan pertumbuhan IPM umumnya disebabkan oleh melambatnya pertumbuhan umur harapan hidup dan pendidikan, serta menurunnya pengeluaran riil per kapita sebagai akibat dari kontraksi pertumbuhan ekonomi (BPS, 2024).

a. Umur Panjang dan Hidup Sehat

Variabel Angka Harapan Hidup, Angka Kematian Ibu, Angka Kelahiran Total, dan Angka Kematian Bayi dimasukkan sebagai komponen penting dalam analisis Indeks Pembangunan Manusia (IPM), khususnya pada indikator umur panjang dan hidup sehat. Angka Harapan Hidup mencerminkan kualitas kesehatan secara keseluruhan, sedangkan Angka Kematian Ibu dan Angka Kematian Bayi menjadi indikator penting untuk menilai layanan kesehatan ibu dan anak. Angka Kelahiran Total menggambarkan tren demografi dan dampaknya terhadap kesehatan masyarakat, sementara Jumlah Tenaga Medis menjadi parameter yang menunjukkan aksesibilitas dan ketersediaan layanan kesehatan yang berkualitas. Dengan mempertimbangkan kelima variabel ini, analisis IPM dapat memberikan gambaran



Melek Huruf, APM SD, APM SMP, APM SMA, Harapan Lama rata Lama Sekolah digunakan sebagai penyusun Indeks Manusia (IPM) pada indikator pengetahuan. Angka Melek Huruf kemampuan dasar literasi masyarakat, sementara Angka APM untuk jenjang SD, SMP, dan SMA menunjukkan tingkat

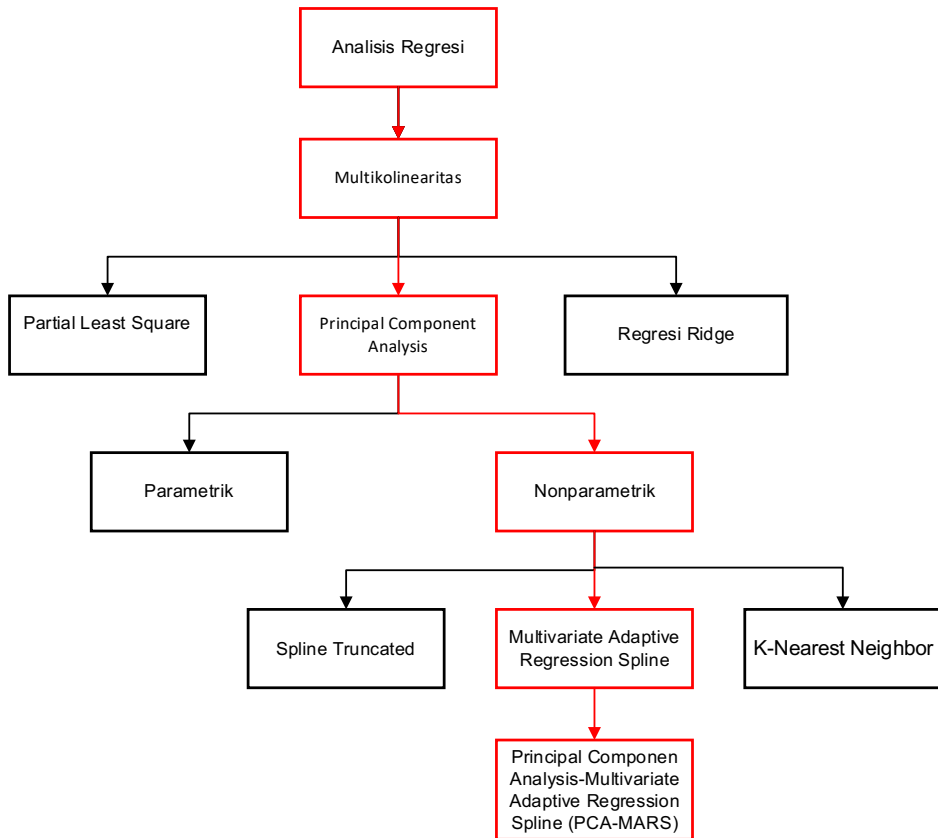
partisipasi pendidikan formal pada berbagai tahapan usia sekolah. Harapan Lama Sekolah mencerminkan ekspektasi jumlah tahun pendidikan yang akan diikuti oleh seorang anak, sedangkan Rata-rata Lama Sekolah mengukur pencapaian rata-rata pendidikan penduduk dewasa. Dengan memasukkan variabel-variabel ini, analisis IPM memberikan gambaran komprehensif tentang tingkat pengetahuan dan akses terhadap pendidikan dalam suatu masyarakat.

c. Standar Hidup Layak

Variabel Kepadatan Penduduk, Pengeluaran per Kapita, Jumlah Penduduk Miskin, dan Tingkat Pengangguran Terbuka digunakan sebagai komponen penting dalam menyusun Indeks Pembangunan Manusia (IPM) pada indikator standar hidup layak. Kepadatan Penduduk mencerminkan distribusi dan tekanan demografis terhadap sumber daya, sedangkan Pengeluaran per Kapita menjadi indikator utama tingkat kesejahteraan ekonomi masyarakat. Jumlah Penduduk Miskin menunjukkan sejauh mana kemiskinan menjadi tantangan dalam pembangunan, dan Tingkat Pengangguran Terbuka menggambarkan peluang akses terhadap pekerjaan yang layak. Dengan memadukan variabel-variabel ini, analisis IPM memberikan gambaran menyeluruh mengenai kualitas standar hidup dan kesejahteraan ekonomi suatu wilayah.



1.6.11 Kerangka Konseptual



Gambar 1. Kerangka Konseptual



BAB II METODE PENELITIAN

2.1 Sumber Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari Badan Pusat Statistik (BPS) selama tahun 2023. Dataset ini mencakup variabel-variabel yang merepresentasikan rata-rata ukuran capaian dari tiga dimensi utama pembangunan manusia, yaitu umur panjang dan hidup sehat, pengetahuan, serta standar hidup layak. Data ini dianalisis menggunakan perangkat lunak R-Studio.

2.2 Variabel Penelitian

Variabel penelitian yang digunakan terdiri dari satu variabel respon (Y) dan 15 variabel (X). Variabel respon dalam penelitian ini adalah Indeks Pembangunan Manusia di Indonesia Tahun 2023, sedangkan untuk variabel prediktornya adalah faktor-faktor yang mempengaruhi Indeks Pembangunan Manusia yang diukur berdasarkan 3 kategori kriteria yakni Umur panjang dan sehat, pengetahuan dan standar hidup layak, seperti yang tertera pada tabel 1 berikut ini:

Tabel 1. Variabel Penelitian

Jenis Variabel	Kode	Keterangan	Satuan
Respon		Indeks Pembangunan Manusia	
	X_1	Angka Melek Huruf	Persen
	X_2	APM SD	Persen
	X_3	APM SMP	Persen
	X_4	APM SMA	Persen
	X_5	Harapan Lama Sekolah	Tahun
	X_6	Rata-rata Lama Sekolah	Tahun
	X_7	Angka Harapan Hidup	Tahun
Prediktor	X_8	Angka Kematian Ibu	Jiwa/1000 kematian
	X_9	Angka Kelahiran Total	Persen
	X_{10}	Angka Kematian Bayi	Persen
	X_{11}	Jumlah Tenaga Medis	Jiwa
	X_{12}	Kepadatan Penduduk	Jumlah penduduk/km ²
	X_{13}	Pengeluaran Per Kapita	Rupiah
	X_{14}	Jumlah Penduduk Miskin	Jiwa
	X_{15}	Tingkat Pengangguran Terbuka	Persen



Sumber Data

Sumber data yang digunakan untuk mencapai tujuan penelitian

1. Mengestimasi parameter model regresi *Principal Component Analysis Multivariate Adaptive Regression Spline*
2. Melakukan uji korelasi pearson antar variabel penjelas untuk mengetahui tingkat korelasinya
3. Melakukan adanya multikolinearitas pada variabel prediktor melalui matriks korelasi dan nilai VIF
4. Standarisasi data dengan menghitung mean dan standar deviasi dari setiap variabel pada data.

$$Z_{ij} = \frac{(X_j - \bar{X}_i)}{S_{ij}}$$

5. Menghitung *matriks korelasi*
6. Menghitung *eigen-value* dan *eigen-vector* dari matriks korelasi
7. Memilih nilai komponen utama
8. Membentuk dataset baru dengan menggunakan nilai komponen utama yang dipilih dari hasil PCA
9. Membangun model MARS dengan dataset baru menggunakan nilai komponen utama yang terpilih dengan persamaan:

$$y_i = a_0 + \sum_{m=1}^M a_m \prod_{k=1}^{K_m} [S_{km}(PC_k - t_{km})]_+ + \varepsilon_i$$

10. Menentukan maksimum jumlah basis fungsi (BF), dalam pemodelan MARS digunakan 4 variabel prediktor yang merupakan nilai komponen utama, sehingga maksimum jumlah basis fungsinya adalah 8, 12, 16 dan 20 fungsi basis.
11. Menentukan jumlah maksimum interaksi. Dalam penelitian ini digunakan interaksi 1,2, dan 3.
12. Menentukan minimum observasi yakni 0,1,2, dan 3.
13. Evaluasi model dengan menggunakan dataset pengujian untuk memastikan bahwa model MARS memiliki performa yang baik:

- $GCV = \sum_{i=1}^N \frac{(y_i - f(x_i))^2}{(1 + \frac{C}{N})^2}$, untuk menentukan jumlah fungsi basis M yang optimal
- Menghitung nilai R^2

14. Menarik kesimpulan.

