

BAB I PENDAHULUAN

1.1 Latar Belakang

Analisis regresi merupakan salah satu metode statistik yang paling umum digunakan untuk memahami hubungan antara satu variabel respon dengan satu atau lebih variabel prediktor. Estimasi parameter dalam regresi klasik umumnya menggunakan metode *Ordinary Least Squares* (OLS) yang bertujuan meminimalkan jumlah kuadrat galat (Augino dkk., 2019). Namun, OLS memiliki keterbatasan ketika terjadi multikolinearitas, yaitu kondisi di mana antar variabel prediktor saling berkorelasi tinggi. Multikolinearitas menyebabkan varians estimasi parameter meningkat, membuat hasil estimasi menjadi tidak stabil diinterpretasikan (Yanuar dkk., 2021).

Permasalahan multikolinearitas sering muncul pada penelitian sosial dan ekonomi, termasuk studi mengenai pembangunan manusia. Salah satu indikator utama dalam mengukur kualitas pembangunan adalah Indeks Pembangunan Manusia (IPM) yang mencakup dimensi kesehatan, pendidikan, dan standar hidup. Ketiga dimensi ini saling berkaitan secara struktural—misalnya, peningkatan pendidikan dapat mendorong pendapatan yang lebih tinggi, sedangkan pendapatan yang tinggi mendukung akses terhadap layanan kesehatan. Hubungan kuat semacam ini berpotensi menimbulkan multikolinearitas dalam pemodelan IPM (Hendriana, 2024).

Sebagai solusi, teknik regresi regularisasi seperti *Ridge Regression* dan *Least Absolute Shrinkage and Selection Operator* (LASSO) telah banyak diterapkan untuk menangani multikolinearitas antarvariabel prediktor. Metode LASSO memiliki keunggulan melakukan seleksi variabel melalui penalti l_1 , yaitu penjumlahan nilai absolut dari seluruh koefisien regresi yang dikalikan dengan parameter penalti λ . Penalti l_1 ini bekerja dengan cara mendekatkan nilai koefisien regresi menjadi nol, dan pada tingkat penalti tertentu, beberapa koefisien dapat menjadi nol secara tepat, sehingga variabel tersebut secara efektif dikeluarkan dari model. Dengan demikian, LASSO tidak hanya mampu mengendalikan varians model akibat multikolinearitas, tetapi juga menghasilkan model yang lebih sederhana, efisien, dan mudah diinterpretasikan (Hastie dkk., 2009). Namun, dalam situasi di mana data memiliki struktur lebih rumit, seperti data spasial atau temporal, LASSO standar mengalami keterbatasan yakni karena tidak dapat menangani keterkaitan antar lokasi atau waktu. Untuk mengatasi hal tersebut, dikembangkan metode *Generalized LASSO*, yaitu bentuk perluasan dari LASSO yang memungkinkan penggunaan matriks penalti yang lebih fleksibel dan dapat disesuaikan dengan struktur data yang diamati, seperti



spasial, atau hierarkis. Pendekatan ini menjadikan *Generalized LASSO* dalam menganalisis data multidimensi yang kompleks, karena mengatasi multikolinearitas, tetapi juga mempertahankan informasi yang relevan (Mutaqin dkk., 2024).

Untuk konteks data spasial yang memiliki struktur ketetanggaan, *Generalized LASSO* standar masih belum sepenuhnya memanfaatkan

informasi spasial secara eksplisit. Oleh karena itu, pendekatan yang lebih adaptif seperti *Generalized Fused LASSO* menjadi solusi yang lebih tepat. *Generalized Fused LASSO* tidak hanya mengintegrasikan seleksi variabel melalui penalti l_1 seperti pada LASSO, tetapi juga menambahkan penalti *fused* yang mendorong koefisien dari wilayah-wilayah yang bertetangga agar memiliki nilai serupa (Tibshirani dkk., 2005).

Efektivitas *Generalized LASSO* dalam mengatasi permasalahan kompleks telah ditunjukkan dalam berbagai penelitian. Misalnya, Rahardiantoro dan Sakamoto (2024) menerapkan *Generalized LASSO* untuk mengungkap pola spasio-temporal penyebaran kasus COVID-19 di Jepang. Penelitian ini menunjukkan bahwa metode tersebut mampu mengidentifikasi kluster penyebaran dengan akurasi tinggi dan stabilitas model yang baik. Studi lain oleh Mutaqin dkk. (2024) menunjukkan bahwa *Generalized LASSO* efektif dalam mengelompokkan wilayah berdasarkan kemiripan sosial-ekonomi di Provinsi Papua menggunakan indikator IPM dan Indeks Pembangunan Gender (IPG), menghasilkan model yang mampu menangkap struktur spasial antar wilayah. Lebih lanjut, metode *Generalized Fused LASSO* telah diterapkan dalam penelitian oleh Petersen dan Müller (2019) untuk menganalisis data spasial dengan mempertimbangkan keterkaitan geografis antar area. Hasilnya menunjukkan bahwa metode ini secara efektif mendorong koefisien wilayah bertetangga untuk memiliki nilai yang serupa, sehingga mampu mempertahankan kontinuitas spasial dan meningkatkan interpretabilitas model. Pendekatan ini sangat bermanfaat dalam konteks pemodelan wilayah yang saling terhubung secara geografis, seperti dalam studi kebijakan pembangunan berbasis wilayah.

Metode *Generalized Fused LASSO* dianggap tepat untuk pemodelan IPM karena mampu menangkap pola spasial yang halus, mempertahankan kontinuitas antar wilayah, serta mengatasi multikolinearitas dan seleksi variabel. Dalam konteks Provinsi Sulawesi Selatan yang memiliki dinamika sosial ekonomi tersebar secara spasial, pendekatan ini menghasilkan model yang lebih stabil, interpretatif, dan relevan untuk mendukung kebijakan pembangunan manusia berbasis data spasial.

Indeks Pembangunan Manusia (IPM) merupakan indikator penting dalam mengukur kualitas hidup masyarakat. IPM mencakup tiga dimensi utama pembangunan: kesehatan (diukur dari angka harapan hidup), pendidikan (diukur dari rata-rata lama sekolah dan harapan lama sekolah), serta standar hidup (diukur dari pengeluaran per kapita yang disesuaikan). Perubahan IPM dipengaruhi oleh berbagai faktor sosial dan ekonomi yang saling berkaitan, seperti tingkat kemiskinan, ketersediaan fasilitas pendidikan, dan distribusi pendapatan (Yanuar dkk., 2021). Di Provinsi Sulawesi Selatan, perkembangan IPM dalam beberapa tahun terakhir



yang fluktuatif, mencerminkan adanya dinamika dalam faktor-kondisi ini menghadirkan tantangan dalam pemodelan statistik, variabel penentu IPM sering kali memiliki hubungan korelasional sama lain. Hal ini menyebabkan munculnya masalah yang berdampak pada instabilitas estimasi parameter regresi jika metode konvensional seperti *Ordinary Least Squares* (OLS). Hal ini menyebabkan koefisien regresi menjadi bias, meningkatkan

residual standar, dan melemahkan kekuatan prediktif model (Badan Pusat Statistik, 2023).

Sejumlah penelitian telah membuktikan efektivitas *Generalized LASSO* dalam pemodelan data pembangunan manusia, terutama dalam konteks yang melibatkan variabel-variabel dengan korelasi tinggi dan struktur spasial yang kompleks. Mengingat karakteristik data Indeks Pembangunan Manusia (IPM) di Provinsi Sulawesi Selatan yang dipengaruhi oleh berbagai faktor sosial dan ekonomi yang saling berkaitan, penggunaan metode ini dipandang sebagai pendekatan yang tepat. Dalam analisis ini, IPM diklasifikasikan berdasarkan kriteria Badan Pusat Statistik (BPS) ke dalam empat kategori, yaitu rendah ($IPM < 60$), sedang ($60 < IPM < 70$), tinggi ($70 < IPM < 80$), sangat tinggi ($IPM > 80$). Klasifikasi ini memberikan dasar interpretasi yang jelas terhadap tingkat pembangunan manusia antar wilayah, sekaligus menjadi acuan dalam mengidentifikasi pola spasial ketimpangan pembangunan.

Lebih spesifik, penelitian ini menerapkan *Generalized Fused LASSO*, yaitu pengembangan dari *Generalized LASSO* yang tidak hanya mengintegrasikan seleksi variabel melalui penalti l_1 , tetapi juga menambahkan penalti *fused* yang mendorong keseragaman koefisien antar wilayah yang bertetangga. Dengan demikian, metode ini mampu menangkap pola spasial secara lebih halus dan menjaga kontinuitas antar wilayah. *Generalized Fused LASSO* tidak hanya menghasilkan estimasi parameter yang lebih stabil, tetapi juga meningkatkan interpretabilitas model dalam konteks spasial. Oleh karena itu, hasil penelitian ini diharapkan dapat memberikan kontribusi nyata dalam pengembangan teknik analisis spasial yang lebih akurat, serta menjadi dasar yang kuat bagi pengambilan kebijakan pembangunan manusia yang berbasis data dan kontekstual di tingkat provinsi.

1.2 Batasan Masalah

Berdasarkan permasalahan yang telah diuraikan sebelumnya, batasan masalah pada penelitian ini adalah sebagai berikut.

1. Data yang digunakan terbatas pada Indeks Pembangunan Manusia (IPM) tahun 2023 beserta variabel-variabel prediktor yang relevan dengan dimensi pendidikan, kesehatan, ekonomi, dan sosial di 24 kabupaten/kota di Provinsi Sulawesi Selatan.
2. Validasi model dilakukan menggunakan *K-fold Cross Validation*, dengan evaluasi model difokuskan pada dua ukuran performa, yaitu *Root Mean Squared Error* (RMSE) dan Koefisien Determinasi (R^2).
3. Interpretasi hasil difokuskan pada identifikasi variabel signifikan, efek spasial dan pembentukan kluster IPM, tanpa membahas implementasi operasional di tingkat makro.



1.3 Tujuan dan Manfaat Penelitian

Tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut.

1. Memperoleh model *Generalized Fused Least Absolut Shrinkage and Selection Operator Regression* pada Indeks Pembangunan Manusia di Provinsi Sulawesi Selatan tahun 2023.
2. Mengklasterkan Kabupaten/Kota di Provinsi Sulawesi Selatan berdasarkan kemiripan karakteristik pembangunan manusia menggunakan model GF-LASSO.

Manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1. Penelitian ini memperkaya literatur mengenai penerapan metode *Generalized LASSO* dan *Generalized Fused LASSO* (GF-LASSO) dalam bidang statistika terapan, khususnya pada pemodelan Indeks Pembangunan Manusia (IPM) yang melibatkan aspek spasial. Hasil penelitian ini dapat menjadi referensi bagi penelitian selanjutnya yang ingin mengembangkan metode regularisasi dengan mempertimbangkan keterkaitan antarwilayah.
2. Hasil penelitian memberikan gambaran mengenai variabel-variabel signifikan yang memengaruhi IPM serta pola klasterisasi spasial antar kabupaten/kota di Sulawesi Selatan. Informasi ini dapat dimanfaatkan oleh pemerintah daerah dalam merumuskan kebijakan pembangunan yang lebih inklusif, terutama dalam mengurangi ketimpangan pembangunan antarwilayah.

1.4 Landasan Teori

1.4.1 Analisis Regresi

Analisis regresi merupakan salah satu metode statistika yang paling umum digunakan untuk menjelaskan hubungan antara satu variabel respon dengan satu atau lebih variabel prediktor. Tujuan utama dari analisis regresi adalah untuk membentuk model prediksi serta memahami kekuatan dan arah hubungan antarvariabel. Model regresi yang paling dasar adalah regresi linier, yang digunakan untuk mengetahui seberapa besar pengaruh variabel prediktor terhadap variabel respon (Gujarati & Porter, 2009). Bentuk umum persamaan regresi linier sederhana dapat dituliskan sebagai berikut (Montgomery dkk., 2013):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (1)$$

Dengan:

Y_i : variabel respon pada pengamatan ke- i

β_0, β_1 : parameter regresi

X : variabel prediktor



dan dari regresi linier sederhana adalah regresi linier berganda. Analisis regresi linier berganda digunakan untuk mempelajari hubungan antara satu variabel respon dengan lebih dari satu variabel prediktor. Bentuk umum model regresi linier berganda dapat dituliskan sebagai berikut :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i, \quad i = 1, 2, \dots, n$$

$$Y_i = \beta_0 + \sum_{j=1}^k \beta_j X_{ij} + \varepsilon_i \quad (2)$$

dengan:

i : 1, 2, ..., n

n : banyaknya pengamatan

k : banyaknya variabel prediktor

Y_i : variabel respon untuk pengamatan ke- i

X_{ij} : variabel prediktor untuk pengamatan ke- i variabel ke- j

$\beta_1, \beta_2, \dots, \beta_k$: parameter regresi

ε_i : residual pada pengamatan ke- i

Bentuk notasi matriks di atas tersebut bisa dituliskan secara sederhana yaitu:

$$\begin{aligned} Y &= X\beta + \varepsilon \\ \varepsilon &= (Y - \hat{Y}) \\ &= (Y - X\hat{\beta}) \end{aligned} \quad (3)$$

Persamaan (2) dapat diubah dituliskan dalam bentuk matriks sebagai berikut:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (4)$$

Keterangan:

Y : vektor variabel respon ($n \times 1$)

X : matriks dari variabel prediktor ($n \times (k + 1)$)

β : vektor parameter model regresi ($(k + 1) \times 1$)

ε : vektor *error* ($n \times 1$)

1.4.2 Ordinary Least Square

Ordinary Least Square (OLS) merupakan salah satu metode untuk mengestimasi parameter regresi dalam analisis regresi linier berganda yang bertujuan untuk meminimumkan jumlah kuadrat galat (Zulkarnain dkk., 2020). Jika model yang akan diestimasi yaitu parameter dari Persamaan (3) dalam menentukan penduga Y yaitu $\hat{Y} = X\hat{\beta}$, diperlukan nilai penduga untuk parameter β yaitu $\hat{\beta}$ dengan parameter $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ dan dalam bentuk vektor dilambangkan sebagai berikut:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$



parameter regresi OLS dilakukan dengan cara meminimumkan $\sum_{i=1}^n \varepsilon_i^2$. Nilai estimasi parameter model regresi dinyatakan pada

$$\begin{aligned} &X\beta) \\ & - X\beta) \\ & - X\beta) \end{aligned}$$

$$= Y^T Y - Y^T X \beta - \beta^T X^T Y + \beta^T X^T X \beta$$

Sesuai dengan sifat matriks *transpose*, $Y^T X \beta = (\beta^T X^T Y)^T$, maka

$$= Y^T Y - \beta^T X^T Y - \beta^T X^T Y + \beta^T X^T X \beta$$

$$= Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta$$

dengan menurunkan $(\epsilon^T \epsilon)$ secara parsial terhadap β dan kemudian menyamakan dengan 0, maka diperoleh:

$$\begin{aligned} \frac{\partial(\epsilon^T \epsilon)}{\partial \beta} \bigg|_{\hat{\beta}} &= 0 - 2X^T Y + 2X^T X \hat{\beta} = 0 \\ -X^T Y + X^T X \hat{\beta} &= 0 \\ \hat{\beta}_{OLS} &= (X^T X)^{-1} X^T Y \end{aligned} \quad (5)$$

1.4.3 Multikolinearitas

Multikolinearitas merupakan kondisi di mana dua atau lebih variabel prediktor dalam model regresi memiliki hubungan linier yang kuat satu sama lain. Keberadaan multikolinearitas dapat menyebabkan ketidakstabilan dalam estimasi koefisien regresi, meningkatnya varians, serta kesulitan dalam menginterpretasikan pengaruh masing-masing variabel (Shrestha, 2020; Tay, 2017). Salah satu metode yang umum digunakan untuk mendeteksi gejala multikolinearitas adalah uji koefisien korelasi *Pearson* antar variabel prediktor. Koefisien korelasi *Pearson* antara variabel X_j dan X_k sebagai berikut:

$$r_{jk} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}} \quad (6)$$

dengan:

x_{ij} : nilai variabel prediktor ke- j pada observasi ke- i

$\bar{x}_j$: rata-rata variabel prediktor ke- j

n : jumlah observasi

Matriks korelasi digunakan untuk mendeteksi adanya multikolinearitas pada variabel prediktor dengan menghitung nilai koefisien korelasi sederhana (*simple correlation*). Terdapat beberapa klasifikasi nilai koefisien korelasi yaitu sebagai berikut (Tichy, 2012).

- Nilai koefisien korelasi 0 – 0,3 artinya memiliki korelasi rendah
- Nilai koefisien korelasi 0,3 – 0,5 artinya memiliki korelasi sedang
- Nilai koefisien korelasi 0,5 – 0,7 artinya memiliki korelasi yang dianggap penting



orelasi 0,7 – 0,9 artinya memiliki korelasi yang kuat

orelasi 0,9 – 1 artinya memiliki korelasi sangat kuat

; 0,7 umumnya dianggap menunjukkan adanya indikasi, sehingga variabel terkait perlu dipertimbangkan untuk menangani menggunakan metode regulasi seperti LASSO atau (Tibshirani, 1996; Hastie dkk., 2009).

1.4.4 Uji Normalitas Residual

Uji normalitas residual bertujuan untuk memastikan bahwa galat (residual) model regresi mengikuti distribusi normal. Kenormalan residual diperlukan untuk menjamin validitas inferensi statistik dan kestabilan estimasi model (Montgomery dkk. 2021).

Residual didefinisikan sebagai berikut:

$$e_i = y_i - \hat{y}_i \quad (7)$$

dengan y_i adalah nilai aktual observasi ke- i , dan $\hat{y}_i$ adalah nilai prediksi model. Salah satu metode yang paling banyak digunakan untuk menguji kenormalan residual adalah Uji *Shapiro–Wilk*. Uji ini diperkenalkan oleh Shapiro dan Wilk (1965) dan dianggap memiliki kekuatan (*power*) yang tinggi dalam mendeteksi penyimpangan dari distribusi normal, khususnya untuk ukuran sampel kecil hingga menengah ($n < 50$). Prinsip dasar uji ini adalah membandingkan urutan data sampel dengan nilai ekspektasi dari distribusi normal. Jika perbedaan antara keduanya kecil, maka data dianggap berdistribusi normal. Secara sistematis, statistik uji Shapiro-Wilk dirumuskan sebagai berikut:

$$W = \frac{(\sum_{i=1}^n a_i e_{(i)})^2}{\sum_{i=1}^n (e_i - \bar{e})^2} \quad (8)$$

dengan:

W = nilai statistik Shapiro-Wilk

$e_{(i)}$ = residual yang telah diurutkan,

$\bar{e}$ = rata-rata residual,

a_i = koefisien konstan yang ditentukan berdasarkan kovarians dan varian dari distribusi normal,

n = jumlah observasi

Nilai W berada pada rentang 0 hingga 1. Semakin besar nilai W (mendekati 1), semakin besar kemungkinan bahwa data mengikuti distribusi normal. Uji hipotesis pada Shapiro-Wilk dapat dinyatakan sebagai berikut:

H_0 : Data berdistribusi normal

H_1 : Data tidak berdistribusi normal

Kriteria pengambilan keputusan dilakukan dengan melihat nilai p-value dari uji. Jika p-value > 0.05 maka H_0 diterima, artinya residual berdistribusi normal; sebaliknya, jika p-value ≤ 0.05 maka H_0 ditolak yang berarti residual tidak berdistribusi normal (Ghasemi & Zahediasl, 2012). Dalam konteks penelitian ini, Uji Shapiro-Wilk digunakan untuk menguji kenormalan residual sebelum dan sesudah transformasi

an uji ini bertujuan memastikan bahwa distribusi residual lask regresi, sehingga hasil estimasi parameter pada model neralized Fused LASSO dapat diinterpretasikan secara valid.



1.4.5 Least Absolute Shrinkage and Selection Operator (LASSO) Regression

Least Absolute Shrinkage and Selection Operator (LASSO) Regression merupakan metode regresi yang digunakan untuk mengatasi permasalahan multikolinearitas dan melakukan seleksi variabel secara simultan. Metode ini bekerja dengan memberikan penalti terhadap besar kecilnya koefisien parameter β sehingga nilainya dapat mendekati nol bahkan menjadi nol. Proses penyusutan (*shrinkage*) ini menghasilkan sejumlah parameter yang lebih sedikit, sehingga model yang diperoleh menjadi lebih sederhana dan mudah diinterpretasikan dibandingkan dengan model regresi biasa (Hastie dkk., 2009). Pada *LASSO Regression*, sebuah penalti ditambahkan ke dalam Persamaan (3). Adapun estimasi koefisien *LASSO Regression* dituliskan pada Persamaan (9) berikut:

$$\hat{\beta}_{lasso} = \arg \min \left\{ \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \quad (9)$$

dengan $\lambda \|\beta\|_1$ merupakan penalti *LASSO Regression* (L_1), dan λ merupakan nilai parameter *tuning* optimal. Semakin besar nilai λ , semakin kuat penalti yang diberikan dan semakin banyak koefisien yang disusutkan menuju nol (Rusdy dkk., 2025).

Model *LASSO Regression* mendorong koefisien regresi menjadi lebih kecil bahkan nol, sehingga hanya mempertahankan variabel prediktor yang memberikan kontribusi signifikan terhadap model. Dengan demikian, *LASSO* berfungsi tidak hanya sebagai model estimasi, tetapi sebagai metode seleksi variabel otomatis yang efektif.

1.4.6 Generalized Least Absolute Shrinkage and Selection Operator Regression

Generalized Least Absolute Shrinkage and Selection Operator (Generalized LASSO) merupakan pengembangan dari metode *LASSO* klasik yang memungkinkan penggunaan penalti terhadap transformasi linier dari koefisien regresi. Metode ini diperkenalkan oleh Tibshirani & Taylor (2011) sebagai perluasan dari kerangka regularisasi *LASSO* dengan menambahkan fleksibilitas pada struktur penalti, yang direpresentasikan oleh matriks D . Dalam metode *LASSO*, penalti L_1 diterapkan langsung terhadap koefisien β , sehingga mendorong beberapa koefisien menjadi nol untuk melakukan seleksi variabel (Hastie dkk., 2009). Namun pada *Generalized LASSO*, penalti tersebut tidak lagi diterapkan langsung terhadap β , melainkan terhadap hasil transformasi liniernya yaitu $D\beta$. Pendekatan ini memungkinkan penalti untuk memodelkan hubungan struktur tertentu di antara koefisien regresi, seperti keterkaitan spasial, temporal atau blok variabel yang saling bergantung (Tibshirani & Taylor 2011; Petersen & Müller, 2019).



omatis, fungsi objektif dari *Generalized LASSO* dirumuskan

$$genlass = \arg \min_{\beta} \left\{ \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \|D\beta\|_1 \right\} \quad (10)$$

di mana matriks penalti berukuran $m \times p$. Matriks D dirancang sesuai dengan analisis seperti hubungan spasial. Jika matriks penalti D tidak ada, maka *Generalized LASSO* akan menghasilkan *LASSO*

standar (Hastie dkk., 2009). Namun, ketika D diambil sebagai operator deferensiasi (seperti matriks selisih orde pertama atau kedua), maka metode dapat termasuk konteks fused LASSO atau trend filtering, yang sangat berguna dalam menganalisis data dengan keterkaitan spasial maupun temporal (Wang dkk., 2016).

Keunggulan utama *Generalized LASSO* adalah kemampuannya dalam menyesuaikan struktur penalti dengan karakteristik data. Hal ini menjadikan metode ini lebih fleksibel dibandingkan LASSO standar, terutama dalam menangani data kompleks yang memiliki struktur spasial maupun temporal. Dengan demikian, *Generalized LASSO* tidak hanya efektif untuk mengatasi multikolinearitas dan melakukan seleksi variabel, tetapi juga mampu mempertahankan pola atau struktur alami yang terdapat dalam data, sehingga hasil estimasi lebih stabil, interpretatif, dan sesuai dengan konteks penelitian (Tibshirani & Taylor, 2011; Petersen & Müller, 2019).

1.4.7 Generalized Fused LASSO

Generalized Fused Least Absolute Shrinkage and Selection Operator (GF-LASSO) merupakan pengembangan dari metode *Least Absolute Shrinkage and Selection Operator* (LASSO) yang diperkenalkan oleh Tibshirani (1996). Metode ini menggabungkan dua tujuan utama, yaitu melakukan seleksi variabel melalui penalti norma L_1 dan menjaga keseragaman antar koefisien yang memiliki hubungan struktural (Tibshirani & Taylor, 2011). Secara intuitif, GF-LASSO tidak hanya berfungsi untuk menyusutkan atau mengeleminasi koefisien regresi yang tidak signifikan, tetapi juga mendorong agar koefisien dua unit yang berdekatan memiliki nilai serupa, sehingga terbentuk klaster spasial pada koefisien (Bleakley & Vert, 2011). Secara matematis bentuk umum fungsi objektif GF-LASSO sebagai berikut:

$$\hat{\beta}_{GFL} = \arg \min_{\beta} \left\{ \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|D_{spatial}\beta\|_1 \right\} \quad (11)$$

Di mana:

Y : vektor respon,

X : matriks prediktor,

β : vektor koefisien regresi yang akan diestimasi,

$\lambda_1 \geq 0$: parameter penalti seleksi variabel,

$\lambda_2 \geq 0$: parameter penalti *fused* yang menjaga keseragaman koefisien,

$D_{spatial}$: matriks penalty yang merepresentasikan struktur ketetanggaan antar wilayah.

Setiap baris pada D merepresentasikan sebuah sisi (i, j) dengan nilai $+1$ pada kolom i dan -1 pada kolom j . Bentuk eksplisit penalti fused dinyatakan



$$\|D_{spatial}\beta\|_1 = \sum_{(i,j) \in E} |\beta_i - \beta_j| \quad (12)$$

it dari penalti fused ditunjukkan pada Persamaan (12), yang ilai absolut selisih koefisien regresi antar pasangan wilayah ulan metode GF-LASSO terletak pada kemampuannya untuk

melakukan seleksi variabel dan pembentukan pola klaster spasial secara simultan (Arnold & Tibshirani, 2020). Dalam konteks analisis Indeks Pembangunan Manusia (IPM), metode ini tidak hanya mengidentifikasi faktor-faktor yang berpengaruh signifikan terhadap IPM, tetapi juga memvisualisasikan keterkaitan antar kabupaten/kota melalui klasterisasi koefisien spasial. Dengan demikian, GF-LASSO menjadi metode yang sangat relevan dan informatif dalam penelitian pembangunan wilayah, karena mampu menghasilkan model yang efisien sekaligus mendukung interpretasi kebijakan berbasis spasial.

1.4.8 Matriks Keterhubungan (*Adjacency Matrix*) dalam Struktur Spasial

Adjacency matrix atau matriks keterhubungan merupakan salah satu representasi paling fundamental dalam teori graf dan statistik spasial. Matriks ini digunakan untuk menggambarkan hubungan ketetanggaan antar unit observasi, baik dalam konteks graf matematis, jaringan, maupun data berbasis wilayah. Menurut Cressie (2015), adjacency matrix adalah struktur dasar yang mengkodekan apakah dua lokasi saling berdekatan atau tidak melalui indikator biner 0 atau 1. Representasi ini menjadi sangat penting dalam analisis spasial karena memungkinkan model statistik menangkap pola dependensi antar wilayah. Secara matematis, *adjacency matrix* berbentuk matriks persegi $n \times n$ dengan elemen w_{ij} yang didefinisikan:

$$w_{ij} = \begin{cases} 1, & \text{wilayah } i \text{ dan } j \text{ bertetanggan} \\ 0, & \text{jika tidak bertetanggan} \end{cases} \quad (13)$$

Dalam konteks geografis, konsep ketetanggaan ini dapat didefinisikan berdasarkan *rook contiguity* (berbagi garis batas), *queen contiguity* (berbagi titik atau garis), atau kedekatan radius tertentu (Anselin, 1988). Struktur adjacency matrix membantu menggambarkan pola interaksi, pengaruh ruang, dan kemiripan karakteristik antar lokasi. Anselin (2013) menekankan bahwa matriks ketetanggaan memungkinkan pengukuran autokorelasi spasial, yang merupakan fenomena penting dalam data wilayah seperti demografi, ekonomi, atau indeks pembangunan. Dalam pemodelan berbasis graf, adjacency matrix digunakan sebagai fondasi untuk membangun struktur penalti dalam berbagai metode regularisasi. Höfling (2010) menunjukkan bahwa penalti *fusion* dalam metode Fused LASSO dan Generalized Fused LASSO sangat bergantung pada struktur adjacency, karena penalti tersebut bekerja dengan menghaluskan atau menyatukan nilai koefisien di antara node-node yang terhubung. Dengan kata lain, adjacency matrix menentukan pasangan wilayah mana yang diberi penalti perbedaan koefisiennya. Lebih lanjut, Tibshirani dkk. (2008) menyatakan bahwa penggunaan adjacency dalam regularisasi graf menghasilkan model yang lebih stabil pada data yang memiliki struktur topologis, seperti data genomik, maupun data spasial. Adjacency matrix juga membuat model menjadi lebih efisien karena penalti hanya diterapkan pada pasangan yang relevan secara struktural. Zhu dan Hastie (2004) memperkuat pernyataan bahwa penalti berbasis graf mampu meningkatkan kinerja model pada data berstruktur, terutama ketika hubungan antar unit observasi penting.



Dalam penelitian ini, adjacency matrix digunakan untuk merepresentasikan hubungan ketetanggaan antar kabupaten/kota di Provinsi Sulawesi Selatan. Dengan menggunakan struktur graf berbasis ketetanggaan ini, model Generalized Fused LASSO mampu mempertimbangkan keterkaitan spasial antar wilayah sehingga menghasilkan estimasi parameter yang lebih stabil, halus, dan relevan secara geografis.

1.4.9 K-Fold Cross Validation

Pemilihan parameter penalti λ merupakan aspek penting dalam model regresi regularisasi, seperti *Least Absolute Shrinkage and Selection Operator* (LASSO) maupun *Generalized LASSO*. Parameter ini berfungsi sebagai pengontrol tingkat penyusutan terhadap koefisien regresi. Nilai λ yang terlalu kecil dapat menyebabkan model menjadi kompleks dan *overfitting*, sedangkan nilai yang terlalu besar dapat menyebabkan *underfitting* dan hilangnya informasi penting. Oleh karena itu, diperlukan metode seleksi parameter yang sistematis untuk menentukan nilai λ yang optimal (Hastie dkk., 2009).

Dalam penelitian ini, pemilihan nilai dilakukan menggunakan *k-fold cross validation*. Metode ini membagi data sebanyak k subset atau *fold* yang kurang lebih sama besar. Pada setiap iterasi, $k - 1$ *fold* digunakan sebagai data latih untuk membangun model, sedangkan satu *fold* sisanya digunakan sebagai data uji untuk mengevaluasi performa model. Proses ini diulang sebanyak k kali hingga setiap *fold* digunakan sekali sebagai data uji (James dkk., 2021). Untuk setiap nilai λ yang diuji, dihitung rata-rata residual prediksi pada seluruh *fold*. Nilai λ yang memberikan rata-rata residual terkecil dipilih sebagai nilai optimal. Residual prediksi yang digunakan dalam evaluasi adalah Nilai CV untuk masing-masing λ yang dirumuskan sebagai berikut:

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{y}_i^{(-k(i))} \right)^2 \quad (14)$$

dengan:

n : jumlah seluruh observasi dalam data,

k : jumlah fold (umumnya 5 atau 10)

y_i : nilai aktual variabel respon pada observasi ke- i

$\hat{y}_i^{(-k(i))}$: nilai prediksi untuk observasi ke- i yang diperoleh dari model.

Nilai $CV(\lambda)$ menggambarkan rata-rata kesalahan prediksi model terhadap data uji. Nilai λ optima dipilih pada titik di mana $CV(\lambda)$ mencapai nilai minimum. Pemilihan k yang digunakan pada penelitian ini adalah $k = 10$ karena jumlah observasi relatif banyak dan memberikan keseimbangan terbaik antara *bias* dan *variance*, serta berukuran kecil hingga menengah. Dengan membagi data ke k bagian, proporsi data latih mencapai 90% dari total data, sehingga model akan lebih akurat untuk membangun model menjadi lebih maksimal. Hal ini



menghasilkan estimasi parameter yang lebih stabil dan galat prediksi yang lebih presisi dibandingkan 5-fold.

(Hastie dkk., 2009) juga menegaskan bahwa pada ukuran sampel kecil, 10 - fold memiliki *bias* lebih rendah dan performa validasi lebih baik karena setiap model dilatih menggunakan hampir seluruh data yang tersedia. Sebaliknya, 5-fold menyebabkan proporsi data latih hanya 80%, sehingga risiko fluktuasi galat lebih tinggi dan estimasi penalti menjadi kurang stabil, terutama pada model regularisasi seperti LASSO, Generalized LASSO, dan Generalized Fused LASSO yang sensitif terhadap pemilihan parameter penalti. Oleh karena itu, penelitian ini menggunakan $k = 10$ karena lebih sesuai untuk jumlah data yang terbatas dan mampu memberikan estimasi galat yang lebih stabil.

1.4.10 Uji Kelayakan Model

Uji kelayak model dilakukan untuk menilai sejauh mana model regresi yang dibangun sesuai dengan data yang dianalisis. Tujuan utama uji ini adalah memastikan bahwa model mampu menjelaskan variabel data secara signifikan serta layak digunakan untuk prediksi maupun interpretasi. Beberapa ukuran yang umum digunakan dalam menilai kelayakan model adalah *Root Mean Squared Error (RMSE)* dan koefisien determinasi (R^2) merupakan dua ukuran evaluasi utama yang umum digunakan untuk menilai kelayakan dan performa model regresi, termasuk model penalized seperti *Generalized LASSO Regression*.

RMSE mengukur rata-rata akar kuadrat dari selisih antara nilai yang diprediksi dan nilai aktual. Metrik ini berguna untuk mengevaluasi ketepatan prediksi secara absolut, di mana semakin kecil nilai RMSE, maka semakin baik model dalam merepresentasikan data aktual (Kuhn & Johnson, 2013; Chai & Draxler, 2014). Sementara itu, koefisien determinasi (R^2) digunakan untuk mengukur proporsi variansi variabel respon yang dapat dijelaskan oleh variabel prediktor dalam model. Nilai R^2 berada dalam rentang $[0, 1]$, dengan nilai yang mendekati 1 menunjukkan bahwa model memiliki kemampuan prediktif yang tinggi dan mampu menjelaskan sebagian besar variasi data (James dkk., 2013).

Kedua metrik ini sering digunakan secara bersamaan untuk menilai performa model, terutama ketika pemilihan parameter *tuning* terbaik seperti penalti λ pada persamaan (10) dan (12). Berikut model koefisien determinasi (R^2) (Kuhn & Johnson, 2013).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (15)$$

dengan:



y_i pengamatan ke- i
 $\hat{y}_i$ pengamatan ke- i
 y ai aktual
 $\bar{y}$ amatan

Sementara itu berikut rumus RMSE sebagai berikut:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (16)$$

Nilai RMSE yang lebih kecil menunjukkan bahwa model memiliki Tingkat kesalahan prediksi yang rendah, sehingga lebih akurat dalam menggambarkan data. Dalam konteks regresi dengan penalti seperti LASSO, *Generalized LASSO*, dan *Generalized Fused LASSO*, uji kelayakan model biasanya dipadukan dengan prosedur *cross validation* untuk memilih parameter penalti optimal. Dengan demikian, model yang terpilih bukan hanya sesuai dengan data latih, tetapi juga memiliki kemampuan generalisasi yang baik pada data baru (Hastie dkk., 2009).

1.4.11 Indeks Pembangunan Manusia

Indeks Pembangunan Manusia (IPM) adalah indikator komposit yang digunakan untuk mengukur capaian pembangunan manusia dari tiga dimensi dasar, yaitu kesehatan, pendidikan, dan standar hidup layak. IPM dikembangkan oleh United Nations Development Programme (UNDP) sebagai alternatif dari ukuran ekonomi semata, guna merefleksikan keberhasilan pembangunan yang berorientasi pada manusia. Di Indonesia, IPM menjadi indikator strategis dalam evaluasi pembangunan daerah dan nasional karena mampu menggambarkan kualitas hidup penduduk secara lebih holistik (Badan Pusat Statistik, 2023).

Dimensi kesehatan diukur melalui indikator Umur Harapan Hidup (UHH), yaitu estimasi rata-rata usia yang diharapkan dapat dicapai oleh seseorang sejak lahir, berdasarkan kondisi lingkungan dan layanan kesehatan yang tersedia (Badan Pusat Statistik, 2022). Dimensi pendidikan mencakup dua indikator utama, yakni Rata-rata Lama Sekolah (RLS), yang menggambarkan jumlah tahun pendidikan formal yang telah diselesaikan oleh penduduk usia 15 tahun ke atas, serta Harapan Lama Sekolah (HLS), yang menunjukkan potensi lama pendidikan yang akan ditempuh anak usia sekolah (Kemendikbud, 2021). Sementara itu, dimensi standar hidup layak direpresentasikan oleh indikator Pengeluaran per Kapita Disesuaikan, yang mencerminkan kemampuan masyarakat dalam memenuhi kebutuhan hidup dasar dan konsumsi yang layak (UNDP, 2020; Bappenas, 2021).

Selain ketiga indikator utama, beberapa variabel pendukung juga berkontribusi dalam menganalisis ketimpangan pembangunan manusia antar wilayah. Angka Melek Huruf (AML) menjadi indikator penting dalam menilai tingkat s dasar penduduk, khususnya usia 15 tahun ke atas, yang gembangan kualitas sumber daya manusia (OECD, 2019). k (ASL) mencerminkan ketersediaan fasilitas sanitasi sehat, jsung dengan penurunan angka penyakit berbasis lingkungan ijat kesehatan (WHO, 2021) Tingkat Partisipasi Angkatan Kerja proporsi penduduk usia kerja yang aktif secara ekonomi, Pengangguran Terbuka (TPT) mencerminkan efisiensi pasar



tenaga kerja. Ketidakseimbangan antara TPAK dan ketersediaan lapangan kerja dapat menjadi hambatan dalam pembangunan manusia yang inklusif (ILO, 2022).

Di Provinsi Sulawesi Selatan, IPM mengalami tren peningkatan selama beberapa tahun terakhir. Namun demikian, masih terdapat disparitas antar wilayah kabupaten/kota yang disebabkan oleh ketimpangan dalam akses terhadap pendidikan, layanan kesehatan, dan kesempatan ekonomi. Berdasarkan data BPS (2023), indikator seperti RLS dan pengeluaran per kapita memberikan kontribusi signifikan terhadap kenaikan IPM, terutama dalam mencerminkan kualitas SDM dan kesejahteraan masyarakat. Meski Umur Harapan Hidup juga menunjukkan kemajuan, tantangan utama masih berfokus pada pemerataan akses sanitasi, literasi dasar, serta penyediaan lapangan kerja, khususnya di daerah pedesaan. Oleh karena itu, upaya peningkatan IPM di Sulawesi Selatan perlu diarahkan pada penguatan layanan dasar, pembangunan infrastruktur pendukung, serta kebijakan ekonomi yang menasar penciptaan kerja berkelanjutan (Bappenas, 2021; BPS Sulsel, 2023).



BAB II METODE PENELITIAN

2.1 Sumber Data

Penelitian ini menggunakan data sekunder yang diperoleh dari situs resmi Badan Pusat Statistik (BPS) melalui laman <https://www.bps.go.id/id>. Data yang digunakan adalah Indeks Pembangunan Manusia (IPM) dan faktor-faktor yang mempengaruhi. Data penelitian mencakup 24 kabupaten/kota di Provinsi Sulawesi Sulawesi tahun 2023. Rincian data lengkap disajikan pada Lampiran 1.

2.2 Variabel Penelitian

Variabel yang digunakan dalam penelitian ini terdiri dua jenis variabel, yaitu variabel respon dan variabel prediktor. Variabel respon berupa Indeks Pembangunan Manusia (IPM), sedangkan variabel prediktor terdiri atas sepuluh indikator yang diduga berpengaruh terhadap IPM pada 24 kabupaten/kota di Provinsi Sulawesi Selatan. Daftar lengkap variabel dapat dilihat pada Lampiran 2.

Tabel 1. Variabel Penelitian

Variabel	Keterangan
Y	Indeks Pembangunan Manusia (IPM)
X_1	Angka Harapan Hidup (AHH)
X_2	Air Minum Layak (AML)
X_3	Sanitasi Layak (SL)
X_4	Rata-rata Lama Sekolah (RLS)
X_5	Harapan Lama Sekolah (HLS)
X_6	Pengeluaran Perkapita (PPK)
X_7	Produk Domestik Regional Bruto (PDRB)
X_8	Tingkat Pengangguran Terbuka (TPT)
X_9	Persentase Penduduk Miskin (PPM)
X_{10}	Tingkat Partisipasi Angkatan Kerja (TPAK)

2.3 Metode Analisis

Analisis dalam penelitian ini dilakukan dengan bantuan *software* Rstudio. Untuk mencapai tujuan penelitian, data akan dianalisis melalui tahapan-tahapan berikut.

1. Melakukan eksplorasi data untuk memperoleh gambaran umum mengenai karakteristik variabel-variabel yang digunakan.
2. Mengidentifikasi multikolinearitas pada data menggunakan Korelasi *Pearson* dan (6).
 a. Menormalkan residual menggunakan uji Shapiro-Wilk pada distribusi variabel prediktor agar berada pada skala yang sama.
 b. Struktur data dan matriks ketetanggaan (*adjacency matrix*) sebagai input spasial pada model *Generalized Fused LASSO*.



6. Membentuk model regresi *Generalized* LASSO untuk seleksi variabel. Pemilihan parameter penalti (λ) yang optimal dilakukan menggunakan *k-fold cross validation* ($k = 5$ atau 10) untuk menghasilkan nilai residual prediksi terkecil.
7. Mengestimasi koefisien dari model *Generalized* LASSO untuk mengidentifikasi variabel-variabel signifikan yang memengaruhi respon.
8. Mengembangkan model *Generalized* LASSO dengan menambahkan penalti *fused* yang memaksa koefisien variabel dari wilayah bertetangga agar seragam. Ini bertujuan untuk menangkap efek keterkaitan spasial dan membentuk kluster-kluster alami.
9. Menyusun representasi spasial melalui *adjacency matrix* untuk memastikan hubungan antarwilayah terintegrasi dalam penalti *fused*.
10. Menentukan nilai λ terbaik pada model GF-LASSO menggunakan *k-fold cross validation*.
11. Mengestimasi koefisien dari model GF-LASSO untuk memperoleh pola keterkaitan antarwilayah.
12. Menganalisis pengaruh penalti *fused* terhadap kestabilan estimasi dan terbentuknya kluster spasial.
13. Melakukan interpretasi spasial dan menarik implikasi kebijakan dari hasil model GF-LASSO.
14. Mengevaluasi dan membandingkan kinerja model *Generalized* LASSO dan GF-LASSO menggunakan metrik *Root Mean Squared Error* (RMSE) dan Koefisien Determinasi (R^2). Model dengan RMSE terkecil dan (R^2) terbesar dianggap sebagai model terbaik.
15. Melakukan interpretasi terhadap hasil estimasi koefisien, mengidentifikasi pola kluster spasial yang terbentuk, dan menarik kesimpulan berdasarkan temuan yang ada.

