Indonesian Journal of Applied Statistics

Journal homepage: https://jurnal.uns.ac.id/ijas/index
E-ISSN 2621–086X
Volume X No. X, Month Year

Perbandingan Metode Optimasi Coordinate Descent dan Stochastic Gradient Descent pada Regresi LASSO Kasus TBC di Jawa Barat Tahun 2022

Ria Yunita[1*], Johannes Pande Manurung [2], Tiara Ayu Pertiwi[3] , R Mugni Chairil Arbi[4] , Ali Uraidly[5], Sachnaz Desta Oktarina[6] , Rahma Anisa[7]
IPB University

*Corresponding author: riayunita8965@gmail.com

Abstract. Tuberculosis (TB) is a contagious disease that remains a major public health concern in Indonesia, with the highest number of cases found in West Java Province. Various factors influence TB incidence, including population density, access to healthcare services, environmental conditions, and socioeconomic status. However, the intercorrelation among these factors often leads to multicollinearity problems, complicating conventional regression analysis. LASSO regression addresses this issue by performing automatic variable selection and regularization, yet its performance depends on the optimization method used. Coordinate Descent (CD) is known for its stability in updating coefficients one at a time, while Stochastic Gradient Descent (SGD) offers computational efficiency through updates based on randomly selected data, although it is sensitive to parameter tuning. This study aims to compare the two optimization methods in building LASSO regression models and to identify factors affecting TB cases in West Java. The analysis shows that the LASSO model using Coordinate Descent achieved an RMSE of 808.421 and an adjusted R^2 of 87.665% with five selected variables, whereas the model using Stochastic Gradient Descent resulted in an RMSE of 819.481 and an adjusted R^2 of 85.089% with seven selected variables. The best-performing model is the LASSO with Coordinate Descent optimization, as it yields the lowest RMSE and highest adjusted R^2, and all five variables in the model have a positive relationship with TB incidence. These findings suggest that LASSO regression with Coordinate Descent optimization can be an effective approach to support TB control policies in high-burden areas such as West Java.

Keywords: *coordinate descent: lasso regression; multicollinearity; stochastic gradient descent; tuberculosis*

1. PENDAHULUAN

Tuberkulosis (TBC) merupakan penyakit infeksi menular yang disebabkan oleh bakteri *Mycobacterium tuberculosis*, yang umumnya menyerang paru-paru dan bagian tubuh lain. Penyakit ini menyebar melalui udara saat penderita batuk atau berinteraksi dengan orang lain. Menurut World Health Organization pada Tahun 2022, Indonesia menempati posisi urutan kedua dengan jumlah kasus TBC tertinggi setelah India [1]. Jumlah kasus TBC di Indonesia pada tahun tersebut melonjak 61,8% menjadi 717.941 dibandingkan Tahun 2021 [2]. Salah satu provinsi yang memiliki kasus TBC terbanyak yaitu Provinsi Jawa Barat dengan jumlah kasus sebesar 85.681 kasus. Hal ini menjadi perhatian serius, sehingga perlu upaya untuk mengetahui faktor-faktor yang menyebabkan peningkatan kasus TBC, khususnya di Jawa Barat.

Jumlah kasus TBC dipengaruhi oleh beberapa faktor, yaitu kepadatan penduduk, kemiskinan, akses layanan kesehatan yang terbatas, serta tantangan dalam diagnosis pengobatan [3]. Faktor lingkungan seperti seperti imunisasi BCG, kelembaban, ventilasi,

jenis kelamin, riwayat kontak, kepemilikan aset, dan tingkat kepadatan penduduk juga memengaruhi kerentanan terhadap TBC [4]. Faktor-faktor tersebut dapat diidentifikasi dengan menggunakan regresi linear berganda, tetapi identifikasi faktor yang akurat sulit dilakukan karena multikolinearitas antar peubah sehingga hasil estimasi regresi ols tidak stabil [5].

Regresi LASSO menawarkan solusi inovatif melalui kemampuan seleksi variabel otomatis. Metode ini secara simultan dapat melakukan seleksi peubah dan regularisasi untuk meningkatkan akurasi dan interpretabilitas model statistik [6]. Namun, efektivitas metode ini sangat bergantung pada proses optimasi untuk menentukan parameter regulasi (λ). Penelitian ini berfokus pada dua optimasi utama, yaitu *Coordinate Descent* dan *Stochastic Gradient Descent*. Algoritma *Coordinate Descent* melakukan pendekatan sistematis dengan memperbarui satu koefisien pada setiap iterasi. Metode ini dikenal stabil, namun kurang efisien untuk dataset berukuran besar [7]. Sebaliknya, *Stochastic Gradient Descent* menggunakan sampel acak untuk mempercepat proses komputasi, sehingga lebih ideal untuk dataset besar, meskipun memerlukan penyesuaian parameter lainnya agar menghasilkan konvergensi yang optimal [8].

Penelitian terdahulu mengenai pendekatan regresi dalam analisis faktor-faktor yang memengaruhi kasus TBC di Jawa Barat telah banyak dilakukan. Penelitian terdahulu yang menggunakan metode regresi binomial negatif untuk memodelkan kasus penyakit TBC di Jawa Barat mendapatkan nilai *Pseudo R-square sebesar* 88,04% [9]. Kemudian, penerapan regresi LASSO dan Group LASSO pada kasus yang sama telah dilakukan dan diperoleh metode Group LASSO memberikan performa terbaik dengan nilai *R-square* sebesar 85,27% [10]. Sementara itu, penelitian terkait regresi LASSO dengan optimasi *Coordinate Descent* dilakukan pada model regresi logistik faktor *public speaking anxiety level* dan hasilnya menunjukkan bahwa regresi lasso dapat menyeleksi faktor-faktor yang signifikan secara otomatis [11].

Namun, penelitian yang secara khusus mengunakan metode regresi LASSO dengan optimasi *Coordinate Descent* dan *Stochastic Gradient Descent* untuk memodelkan kasus TBC di Jawa Barat belum pernah dilakukan. Oleh karena itu, penelitian ini bertujuan untuk membandingkan performa dua algoritma optimasi, yaitu *Coordinate Descent* dan *Stochastic Gradient Descent* dalam estimasi parameter model regresi LASSO untuk kasus TBC di Jawa Barat. Selain itu, penelitian ini juga bertujuan untuk mengidentifikasi peubah-peubah penjelas yang memengaruhi peningkatan jumlah kasus TBC di Jawa Barat. Dengan demikian, penelitian ini diharapkan dapat memberikan metode yang lebih efektif dan efisien serta menghasilkan wawasan komprehensif untuk mendukung kebijakan penanggulangan TBC yang tepat sasaran.

2. TINJAUAN PUSTAKA

2.1 Regresi LASSO.

Regresi *Least Shrinkage and Selection Operator* (LASSO) merupakan salah satu metode regresi yang mulai diperkenalkan oleh Tibshirani pada Tahun 1996 yang mampu mengatasi permasalahan multikolineritas dalam suatu model regresi [12]. LASSO dapat menyusutkan nilai koefisien regresi hingga bernilai nol dengan menambahkan penalti *l1-norm.* Penalti penyusutan ini menyebabkan nilai penduga koefisien parameter menyusut

sehingga peubah penjelas yang berpengaruh dimasukkan ke dalam model, sedangkan peubah yang tidak berpengaruh disusutkan sampai nol sehingga model menjadi lebih efisien [13]. Bentuk persamaan umum dari fungsi objektif LASSO sebagai berikut.

$$\beta_{lasso} = argmin \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 + \lambda\sum_{j}^{k}|\beta_j|$$

n : jumlah pengamatan y_i : nilai peubah respon untuk pengamatan ke-i

k : jumlah peubah penjelas x_{ij} : nilai peubah penjelas ke-j untuk pengamatan ke-i

β_0 : intersep β_0 : koefisien regresi untuk peubah penjelas ke-j

λ : parameter tuning yang menentukan tingkat penalti

Jika nilai λ semakin besar, maka akan semakin besar nilai penalti yang digunakan sehingga menyebabkan semakin sedikit peubah yang dipertahankan dalam model. Disamping itu, nilai λ juga berfungsi sebagai penyeimbang antara bias dan ragam [14]. Kebiasan akan semakin besar jika λ semakin besar, sedangkan ragam akan semakin besar jika nilai λ semakin kecil.

2.2 *Coordinate Descent*

Algoritma *Coordinate Descent* adalah salah satu metode optimasi yang dapat membantu untuk menyelesaikan permasalahan optimasi dengan meminimalkan fungsi objektif yang kompleks. Algoritma ini dijalankan dengan membentuk suatu nilai parameter baru (B_j^{New}) yang mana akan dapat mengoptimasi nilai dari salah satu peubah dan nilai peubah lainnya akan konstan [11]. Secara umum, fungsi objektif yang digunakan *Coordinate Descent* dalam membagi optimasi menjadi beberapa pecahan bagian masalah dengan satu peubah, yaitu [15] :

$$\min_{\beta_j} f(\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)$$

Metode ini umum digunakan dalam permasalahan yang menggunakan banyak peubah. Tujuan utama dari algoritma ini adalah untuk menyelesaikan permasalahan optimasi yang berdimensi tinggi secara efisien dalam kasus regularisasi seperti *LASSO regression*. Fungsi objektif yang diminimalkan oleh algoritma *Coordinate Descent* dalam regresi LASSO ditunjukkan oleh persamaan berikut :

$$\min_{\beta}\left\{\frac{1}{2n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

Dalam persamaan tersebut, terdapat penalti l1 ($|\beta_j|$), sehingga digunakan pendekatan *soft-thresholding* dalam algoritma ini untuk menyelesaikan permasalahan optimasi. Langkah-langkah Algoritma *Coordinate Descent* sebagai berikut [16]:

1. Inisiasi semua koefisien $\beta_j = 0$.
2. Untuk semua parameter β_j , ulangi :
    a. Hitung residual dengan persamaan: $r_j = y - \sum_{k=j} x_k\beta_k$

b. Hitung korelasi dengan persamaan : $\rho_j = \frac{1}{n} X_j^T , r_j = \frac{1}{n} \sum_{i-1}^{n} x_{ij} (y_i - \hat{y}_i^{(-j)})$

c. Lakukan pembaruan untuk *soft-thresholding* dengan persamaan :

$$\beta_j^{new} \leftarrow \frac{S(\rho_j, \lambda)}{\frac{1}{n}\sum_{i=l}^{n} x_{ij}^2}$$

dimana fungsi soft-thresholding yang digunakan yaitu :

$$S(\rho, \lambda) = \begin{cases} \rho - \lambda \, jika \, \rho > \lambda \\ 0 \, jika \, |\rho| \leq \lambda \\ \rho + \lambda \, jika \, \rho < \lambda \end{cases}$$

3. Ulangi langkah ke-2 untuk seluruh β_j hingga mencapai konvergen.

2.3 *Stochastic Gradient Descent*

Algoritma *Stochastic Gradient Descent* (SGD) dapat digunakan sebagai pendekatan dalam mengaproksimasikan gradien dari fungsi objektif LASSO [17]. Algoritma ini mengambil satu contoh secara acak dari set pelatihan pada setiap iterasi serta menghitung gradien berdasarkan contoh tersebut [18]. Secara umum, pembaruan parameter pada optimasi SGD mengikuti persamaan berikut :

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla f_i(\theta^{(t)})$$

Nilai θ adalah parameter model, η adalah *learning rate,* dan $\nabla f_i(\theta)$ merupakan gradien dari fungsi objektif f terhadap contoh ke-i yang diambil secara acak. Dalam konteks ini, f merupakan fungsi objektif dari regresi LASSO, yaitu gabungan *squared loss* dan regularisasi L_1. Persamaan SGD dalam regresi LASSO dapat dituliskan secara khusus sebagai berikut :

$$\beta^{(t+1)} = \beta^{(t)} - \eta\left[-\left(y_i - \beta^{(t)T} x_i\right) x_{ij} + \lambda . sign\left(\beta_j^t\right)\right]$$

Proses algoritma SGD tidak mengingat contoh-contoh yang digunakan pada iterasi sebelumnya sehingga dapat memproses data secara langsung dalam sistem yang sedang berjalan [19]. Namun, sifat stokastik SGD menyebabkan algoritma menjadi tidak teratur sehingga nilai parameter yang diperoleh tidak stabil. Hal ini menyebabkan nilai parameter yang diperoleh sudah baik, tetapi tidak optimal [18]. Proses algoritma *Stochastic Gradient Descent* yang digunakan pada penelitian ini adalah sebagai berikut:

1. Inisiasi koefisien regresi $\beta_0 = 0$ dan $\beta_j = 0$ untuk $j = 1,2,..,k$
2. Inisiasi hyperparameter yang digunakan, yaitu parameter penyusutan (λ), *learning rate* (η), jumlah iterasi, dan skema *learning rate*.
3. Untuk setiap iterasi $t = 1,2,..,T$, lakukan:
    a) Mengambil satu contoh acak (x_i, y_i)
    b) Menghitung nilai *learning rate* pada iterasi ke-t (η_t) sesuai dengan skema *learning rate* yang digunakan
    c) Menghitung nilai prediksi $\widehat{y_i} = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$
    d) Menghitung nilai sisaan $e_i = y_i - \widehat{y_i}$
    e) Menghitung nilai intersep $\beta_0 = \beta_0 + \eta_t \cdot e_i$
    f) Untuk setiap $j = 1,2,..,k$, lakukan:

1) Menghitung nilai gradien $\nabla f = \frac{\partial f}{\partial \beta_j} = -e_i x_{ij}$
2) Menghitung nilai koefisien $\beta_j = \beta_j - \eta_t\left[\left(-e_i x_{ij}\right) + \lambda \cdot sign\left(\beta_j\right)\right]$, dengan $sign(\beta_j) = 1$ jika $\beta_j > 0$, $sign(\beta_j) = -1$ jika $\beta_j < 0$, dan $sign(\beta_j) = 0$ jika $\beta_j = 0$.

2.4 *K-Fold Cross Validation*

Validasi silang (*Cross Validation*) merupakan metode untuk mengevaluasi kinerja model [20]. Terdapat beberapa jenis validasi silang, salah satunya yaitu *K-Fold Cross Validation*. *K-Fold Cross Validation* membagi data menjadi k bagian (*folds*) yang sama besar [21]. Sebanyak k-1 gugus data secara acak dipilih untuk melatih model sebagai data latih dan 1 gugus data yang tersisa digunakan sebagai data uji, kemudian menghitung *Root Mean Square Error* (RMSE) sebagai validasi model. Proses ini dilakukan hingga semua bagian (*folds*) menjadi data *testing*. Model terbaik ditentukan berdasarkan nilai rataan RMSE terkecil. *Root Mean Square Error* (RMSE) dapat dihitung melalui persamaan berikut:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\widehat{y_1} - y_i)^2}{n}}.$$

3. Metode Penelitian

3.1 Data

Penelitian ini menggunakan data sekunder yang diperoleh dari laman resmi Badan Pusat Statistik (www.bps.go.id) dan laman *Open* Data Jawa Barat (opendata.jabarprov.go.id). Data ini terdiri dari jumlah kasus penyakit TBC sebagai peubah respon dan 20 faktor yang mempengaruhi terjadinya kasus tersebut sebagai peubah penjelas dengan terdapat 27 amatan. Peubah-peubah yang digunakan dalam analisis disajikan pada Tabel 1.

Tabel 1. Peubah Penelitian

Peubah	Keterangan	Satuan
Y	Jumlah Kasus Penyakit TBC	Orang
X1	Jumlah Penduduk Disabilitas	Orang
X2	Jumlah Rumah Sakit Umum	Unit
X3	Jumlah Puskesmas	Unit
X4	Jumlah Tenaga Medis	Orang
X5	Jumlah Tenaga Keafirmasian	Orang
X6	Jumlah Balita dengan Gizi Kurang	Orang
X7	Jumlah HIV	Orang
X8	Jumlah Penduduk Miskin	Orang
X9	Produk Domestik Regional Bruto	Ribu Rupiah
X10	Jumlah Kecamatan	Kecamatan
X11	Jumlah Desa	Desa
X12	Jumlah Sampah Ditangani	Ton/Hari

Peubah	Keterangan	Satuan
X13	Jumlah Sampah TPA	Ton/Hari
X14	Proyeksi Penduduk	Ribu Orang
X15	Luas Daerah	Kilometer Persegi
X16	Jumlah Orang Bekerja	Orang
X17	Jumlah Orang Pengangguran Terbuka	Orang
X18	Indeks Pembangunan Manusia	-
X19	Jumlah Angkatan Kerja	Orang
X20	Harapan Lama Sekolah	Tahun

3.3 Tahapan Analisis

Penelitian ini membandingkan performa dua algoritma optimasi yaitu *Coordinate Descent* dan *Stochastic Gradient Descent* (SGD) pada regresi LASSO dalam mengidentifikasi peubah yang yang berpengaruh pada jumlah kasus TBC di Jawa Barat. Penelitian ini menggunakan perangkat lunak python dan R 4.3.1 dengan tahapan yang dilakukan adalah:

1. Melakukan eksplorasi data untuk melihat karakteristik dan pola data jumlah kasus TBC melalui visualisasi dengan boxplot dan barchart. Selanjutnya, melakukan eksplorasi dengan matriks korelasi untuk mengetahui hubungan antara jumlah kasus TBC dengan peubah-peubah penjelas yang digunakan.
2. Melakukan standarisasi z-*score* pada peubah penjelas dengan persamaan sebagai berikut:
$$z = \frac{x_i - \bar{x}}{s}$$
3. Melakukan pemodelan jumlah kasus TBC dengan pendekatan *Ordinary Least Squares* (OLS)
4. Melakukan uji asumsi sisaan klasik dan deteksi multikolinearitas pada model OLS.
5. Melakukan pemodelan jumlah kasus TBC dengan pendekatan regresi LASSO dengan optimasi *Coordinate Descent* dan *Stochastic Gradient Descent*.
    a) *Coordinate Descent*
    Pemodelan dengan *Coordinate Descent* dilakukan dengan membangkitkan maksimal 100 kemungkinan nilai λ dari sebaran logaritma. Semua nilai λ dilakukan validasi silang 10-*fold* untuk mencari λ optimum yang meminimumkan nilai rataan *Root Mean Square Error* (RMSE).
    b) *Stochastic Gradient Descent*
    Pemodelan dengan *Stochastic Gradient Descent* dilakukan dengan menginisiasi *hyperparameter tuning* yang digunakan. Semua kombinasi *hyperparameter tuning* dilakukan validasi silang 10-*fold* untuk mencari *hyperparameter tuning* optimum yang meminimumkan nilai rataan *Root Mean Square Error* (RMSE). *Hyperparameter tuning* yang digunakan adalah:
        1) Lambda $\lambda = 1,2,3,\ldots,300$
        2) *Learning rate* $\eta = 0{,}001;\ 0{,}01;\ 0{,}1$

3) Jumlah iterasi $t = 500, 1000, 2000$
4) Skema *learning rate*, yaitu *constant* dan *invscaling*. Skema *learning rate constant* adalah sebagai berikut:

$$\eta_t = \eta$$

Skema *learning rate invscaling* adalah sebagai berikut:

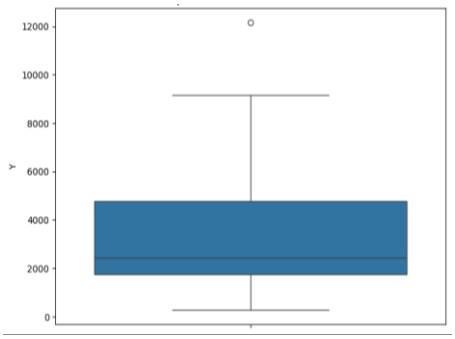
$$\eta_t = \frac{\eta}{t^{0,25}}$$

6. Melakukan transformasi balik nilai koefisien model regresi LASSO.
7. Mengevaluasi model regresi LASSO dengan RMSE dan *Adjusted* R^2. Selanjutnya, dilakukan pemilihan model terbaik dengan kriteria nilai RMSE terkecil dan *Adjusted* R^2 terbesar.
8. Melakukan interpretasi model terbaik regresi LASSO.

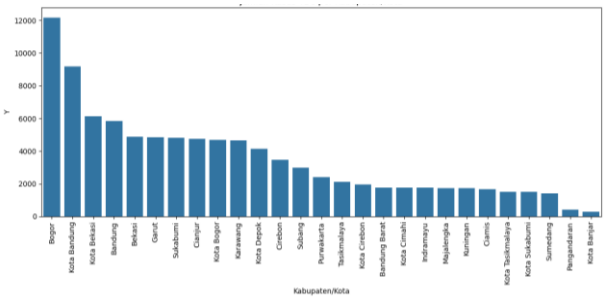
4. Hasil dan Pembahasan

4.1 Eksplorasi Data

Eksplorasi terhadap data kasus Tuberkulosis (TBC) di Provinsi Jawa Barat Tahun 2022 dilakukan untuk mengetahui karakteristik distribusi data sebelum dilakukan pemodelan. Gambar 1 menyajikan boxplot yang menggambarkan sebaran jumlah kasus TBC di Provinsi Jawa Barat. Berdasarkan hasil visualisasi tersebut, dapat dilihat bahwa distribusi jumlah kasus cenderung miring ke kanan (*right-skewed*). Distribusi tersebut menunjukkan bahwa nilai rata-rata kasus TBC lebih besar dari median karena keberadaan nilai yang sangat tinggi. Berdasarkan boxplot tersebut dapat dilihat bahwa terdapat satu outlier yang memiliki nilai jumlah kasus yang melebihi 12.000 kasus, dimana nilai ini berada jauh di atas ambang batas kuartil atas. Hal ini mengindikasikan adanya ketimpangan dalam penyebaran kasus TBC di provinsi tersebut.

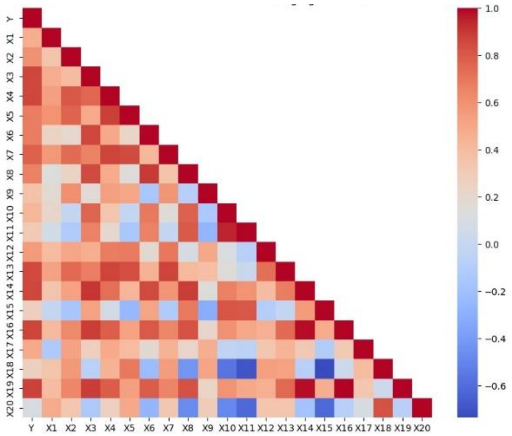


Gambar 1. Boxplot sebaran kasus TBC



Gambar 2. Jumlah kasus TBC per kabupaten/kota

Ketimpangan kasus TBC dapat dilihat dengan jelas pada Gambar 2 yang menampilkan jumlah kasus TBC di masing-masing kabupaten/kota. Kabupaten Bogor menjadi wilayah dengan jumlah kasus tertinggi, diikuti oleh Kota Bandung dan Kota Bekasi. Sebagian besar kabupaten/kota menunjukkan jumlah kasus di bawah 5.000, tetapi terdapat beberapa daerah seperti Kota Banjar dan Pangandaran dengan angka yang sangat rendah. Pola ini menunjukkan distribusi yang tidak merata dalam kasus TBC di Jawa Barat.



Gambar 3. Matriks korelasi

Kemudian, analisis korelasi antar peubah dilakukan untuk mengetahui hubungan antara jumlah kasus TBC dengan peubah-peubah penjelas yang digunakan. Gambar 3 menunjukkan matriks korelasi yang menggambarkan hubungan antara peubah respon (Y) dengan 20 peubah penjelas (X). Berdasarkan Gambar 3 tersebut, dapat dilihat bahwa sebagian besar peubah penjelas memiliki korelasi yang positif dengan peubah respon, hanya peubah harapan lama sekolah (X20) saja yang memiliki korelasi negatif dengan peubah respon. Selain itu, terlihat juga bahwa terdapat korelasi tinggi antar beberapa peubah penjelas, misalnya antara peubah X6 dan X8 yang memiliki nilai korelasi sebesar 0,90 serta peubah X3 dan X14 yang memiliki nilai korelasi sebesar 0,91. Hal ini mengindikasikan adanya potensi multikolinearitas yang tinggi ketika akan dilakukan pemodelan regresi linear berganda.

4.2 Regresi Linear Berganda (OLS)

Tahapan awal dalam penelitian ini yaitu membangun model regresi linear berganda untuk mengetahui karakteristik dan hubungan antar peubah respon dan peubah-peubah penjelas. Pemodelan regresi dilakukan dengan menggunakan pendekatan Ordinary Least Squares (OLS). Hasil estimasi koefisien regresi pada peubah penjelas ditunjukkan oleh Tabel 2.

Tabel 2. Estimasi koefisien regresi OLS

Peubah	Koefisien	Peubah	Koefisien	Peubah	Koefisien	Peubah	Koefisien
X1	-0,193	X6	-0,393	X11	0,544	X16	0,0676
X2	-120,233	X7	-0,403	X12	-3,772	X17	692,896
X3	225,963	X8	-33,335	X13	3,385	X18	-568,717
X4	1,240	X9	-0,005	X14	-0,281	X19	-0,055
X5	-2,733	X10	-205,755	X15	-0,864	X20	144,386
Intercept	$3{,}48 \times 10^4$						

Berdasarkan Tabel 2, estimasi awal dari pemodelan regresi menunjukkan adanya perbedaan arah hubungan antara koefisien regresi dengan korelasi awal yang diperoleh dari matriks korelasi antar peubah. Hal ini ditunjukkan seperti pada peubah X1,X2,X5,X6, X7, X8, X9, X10,X12, X14, X18, dan X19 menunjukkan perubahan arah koefisien dari positif ke negatif. Sedangkan untuk peubah X20 menunjukkan hubungan yang bernilai positif setelah pemodelan regresi, tetapi hasil korelasi sebelumnya menunjukkan nilai yang

negatif. Ketidakkonsistenan ini mengindikasikan adanya potensi masalah dalam struktur hubungan antar peubah sehingga perlu dilakukan pengujian asumsi-asumsi klasik regresi *Ordinary Least Squares* (OLS) dan pengecekan multikolinearitas.

4.3 Uji Asumsi Klasik dan Deteksi Multikolinearitas

Pengujian terhadap asumsi klasik regresi OLS meliputi uji normalitas residual, uji heteroskedastisitas, uji autokorelasi, multikolinearitas, dan uji nilai harapan sisaan sama dengan nol. Uji asumsi klasik dilakukan untuk memperoleh model regresi yang menghasilkan ketepatan estimasi, ketiadaan bias, dan konsisten [22]. Tabel 3 berikut merupakan hasil uji asumsi klasik pada regresi OLS.

Tabel 3. Hasil uji asumsi klasik

Asumsi	Uji yang digunakan	Nilai p	Keputusan
Normalitas sisaan	*Shapiro-Wilks*	0,782	Sisaan menyebar normal
Heteroskedastisitas sisaan	*Breusch-Pagan*	0,727	Ragam sisaan homogen
Autokorelasi sisaan	*Durbin-Watson*	0,343	Sisaan saling bebas
Nilai harapan sisaan sama dengan nol	*t-test*	1	Nilai harapan sisaan sama dengan nol

Keputusan agar semua asumsi klasik terpenuhi terjadi ketika nilai p dari hasil uji asumsi lebih besar dari 0,05 karena asumsi regresi berada dalam hipotesis nol, sehingga dalam uji asumsi diharapkan untuk menerima hipotesis nol. Berdasarkan Tabel 3, seluruh hasil uji asumsi menunjukkan nilai *p* yang lebih besar dari 0,05, sehingga dapat disimpulkan bahwa semua asumsi regresi telah terpenuhi. Meskipun demikian, untuk memastikan kualitas model secara menyeluruh, perlu dilakukan pemeriksaan terhadap multikolinearitas, yang ditampilkan dalam Tabel 4 berikut.

Tabel 4. Deteksi multikolineritas

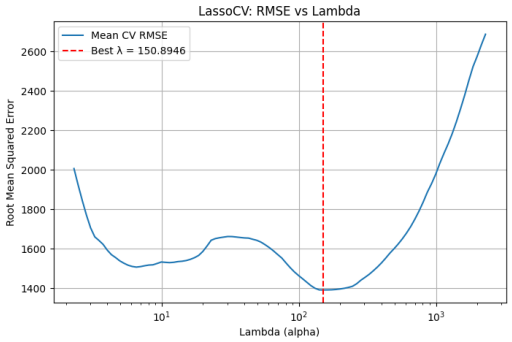
Peubah	VIF	Peubah	VIF	Peubah	VIF	Peubah	VIF
X1	5,674	X6	26,037	X11	41,087	X16	25960,246
X2	18,232	X7	26,118	X12	10,995	X17	27,003
X3	176,566	X8	123,361	X13	36,016	X18	109,116
X4	41,671	X9	8,682	X14	297,818	X19	23183,085
X5	22,351	X10	84,148	X15	23,907	X20	7,509

Pengecekan multikolinearitas dilakukan dengan menggunakan *Variance Inflation Factor* (VIF). Peubah penjelas yang memiliki nilai VIF > 10 maka terjadi multikolineritas di peubah tersebut. Berdasarkan Tabel 4, hanya peubah penjelas X1, X9, dan X20 yang memiliki nilai VIF < 10. Hal ini berarti semua peubah penjelas kecuali X1,X9, dan X20 dalam model regresi tersebut mengalami multikolinearitas. Beberapa peubah penjelas juga memiliki nilai VIF yang sangat besar seperti X16 dan X19 memiliki nilai VIF sebesar 25960,24 dan 23183,08 sehingga terjadi multikolinearitas yang ekstrim. Multikolinearitas ini dapat mengakibatkan koefisien regresi menjadi tidak stabil, *standard error* membesar, dan interpretasi hubungan antar peubah menjadi tidak akurat. Dampak tersebut dapat terlihat dari hasil estimasi koefisien regresi OLS pada Tabel 2 yang mengalami perubahan dengan nilai korelasi awal. Oleh karena itu model OLS tidak dapat digunakan dalam kasus ini sehingga akan dilakukan pemodelan regresi LASSO dengan dua algoritma optimasi.

4.3 Pemodelan Regresi LASSO

4.3.1 Regresi LASSO dengan Optimasi *Coordinate Descent*

Analisis regresi LASSO yang pertama dilakukan dengan menggunakan metode optimasi *Coordinate Descent*. Dalam analisis ini, dilakukan *tuning hyperparameter* untuk mencari nilai λ terbaik. T*uning hyperparameter* dengan validasi silang 10-fold menghasilkan nilai lamda (λ) terbaik sebesar 150,89, seperti yang ditunjukkan oleh Gambar 4.



Gambar 4. Kurva validasi silang setiap parameter (λ)

Berdasarkan Gambar 4 garis vertikal bewarna merah yang menunjukkan nilai lamda (λ) terbaik yang dipilih karena memberikan nilai *Root Mean Squared Error* (RMSE) terkecil pada saat validasi. Nilai RMSE yang dihasilkan dari model ini yaitu sebesar 808,421. Selain itu, model LASSO ini menghasilkan performa nilai R^2 sebesar 90,512% dan $R^2{}_{adj}$ sebesar 87,665%. Nilai tersebut menunjukkan bahwa proporsi variabilitas data yang mampu dijelaskan oleh model cukup tinggi. Dari hasil kebaikan model ini, nilai estimasi koefisien peubah penjelas ditunjukkan oleh Tabel 5.

Tabel 5. Estimasi koefisien regresi LASSO *Coordinate Descent*

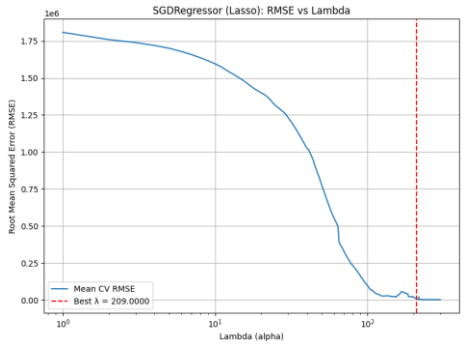
Peubah	Koefisien	Peubah	Koefisien	Peubah	Koefisien	Peubah	Koefisien
X1	0	X6	0	X11	0	X16	0
X2	0	X7	0	X12	0	X17	105,038
X3	37,355	X8	0	X13	2,292	X18	0
X4	0,121	X9	0	X14	0,464	X19	0
X5	0	X10	0	X15	0	X20	0
Intercept	−1048,141						

Berdasarkan hasil estimasi koefisien pada Tabel 5, model regresi LASSO menunjukkan kemampuannya dalam mereduksi sejumlah besar koefisien regresi menjadi nol. Dari total dua puluh peubah penjelas, hanya tersisa lima peubah yang memiliki koefisien tidak nol, yaitu X3, X4, X13, X14, dan X17. Nilai koefisien yang tidak nol tersebut menunjukkan bahwa kelima peubah tersebut memiliki pengaruh signifikan dalam menjelaskan peubah respon. Sedangkan, peubah lain yang memiliki koefisien nol tidak memiliki pengaruh signifikan dalam menjelaskan peubah respon.

4.3.2 Regresi LASSO dengan Optimasi *Stochastic Gradient Descent* (SGD)

Analisis regresi LASSO selanjutnya dilakukan dengan metode *Stochastic Gradient Descent*. Dalam analisis ini, dilakukan *tuning hyperparameter* untuk menentukan nilai parameter terbaik. Hasil *tuning hyperparameter* memberikan kombinasi parameter terbaik

yaitu lamda (λ) sebesar 209, *learning rate* sebesar 0,1, jumlah iterations sebesar 500, dan skema *learning rate invscaling*. Hasil tuning hyperparameter menggunakan validasi silang ditunjukkan oleh Gambar 5.



Gambar 5. Kurva validasi silang setiap parameter (λ) bagian SGD

Berdasarkan Gambar 5, garis biru menunjukkan nilai rata-rata RMSE dari validasi silang untuk berbagai nilai lambda, sedangkan garis merah vertikal menunjukkan titik lambda terbaik yang menghasilkan nilai RMSE terkecil yaitu sebesar 819,481. Model LASSO yang dibangun menunjukkan performa yang baik dengan menghasilkan nilai R^2 sebesar 90,250% dan $R^2{}_{adj}$ sebesar 85,089%. Hal tersebut menunjukkan bahwa model cukup representatif dalam menjelaskan variabilitas data, dan kontribusi masing-masing peubah penjelas dapat dilihat melalui estimasi koefisien regresi pada Tabel 6.

Tabel 6. Estimasi koefisien regresi LASSO *Stochastic Gradient Descent*

Peubah	Koefisien	Peubah	Koefisien	Peubah	Koefisien	Peubah	Koefisien
X1	0	X6	0,015	X11	0	X16	0
X2	0	X7	0	X12	0	X17	124,622
X3	49,143	X8	0	X13	2,748	X18	0
X4	0,006	X9	0,001	X14	0,219	X19	0
X5	0	X10	0	X15	0	X20	0
Intercept	−1577,368						

Model regresi LASSO dengan pendekatan SGD menunjukkan kemampuan untuk melakukan seleksi variabel secara otomatis dengan mereduksi sebagian besar koefisien menjadi nol, seperti yang ditunjukkan oleh hasil estimasi koefisien yang ditunjukkan dalam Tabel 6. Terdapat tujuh peubah dari dua puluh peubah penjelas memiliki koefisien tidak nol, yaitu X3, X4, X6, X9, X13, X14, dan X17. Koefisien yang tidak nol menunjukkan bahwa peubah-peubah ini memainkan peran yang signifikan dalam menjelaskan variabilitas peubah respon.

4.3 Pemilihan Model Terbaik

Pemilihan model terbaik berdasarkan nilai RMSE dan $R^2{}_{adj}$ yang dihasilkan pada setiap model optimasi LASSO. Hasil analisis LASSO dengan *Coordinate Descent* dan *Stochastic Gradient Descent* memiliki perbedaan. Salah satunya yaitu banyak peubah yang dihasilkan di *Coordinate Descent* lebih sedikit dibandingkan *Stochastic Gradient Descent*. Hal ini akan memengaruhi nilai R^2 dan $R^2{}_{adj}$, dimana kedua nilai tersebut bergantung pada banyaknya peubah dalam model, tetapi nilai $R^2{}_{adj}$ lebih stabil untuk membandingkan

model dengan jumlah peubah yang berbeda. Evaluasi kebaikan kedua model tersebut ditunjukkan dalam Tabel 7.

Tabel 7. Ukuran kebaikan analisis LASSO *Coordinate Descent* dan SDG

Metode	RMSE	$R^2{}_{adj}$
LASSO *Coordinate Descent*	808,421	87,665%
LASSO *Stochastic Gradient Descent*	819,481	85,089%

Nilai RMSE menunjukkan ukuran error, sehingga semakin kecil nilai RMSE makan model yang dihasilkan akan semakin baik. Sedangkan nilai $R^2{}_{adj}$ menunjukkan persentase keragaman peubah respon yang dapat dijelaskan oleh peubah penjelas, semakin besar nilai $R^2{}_{adj}$, maka model yang dihasilkan semakin baik. Berdasarkan Tabel 7, model LASSO dengan *Coordinate Descent* memiliki nilai RMSE terkecil, dan $R^2{}_{adj}$ terbesar. Oleh karena itu, dapat disimpulkan bahwa metode *Coordinate Descent* memberikan performa terbaik untuk membangun model regresi LASSO pada data yang digunakan. Model terbaik yang diperoleh menggunakan metode *Coordinate Descent* menghasilkan persamaan regresi sebagai berikut:

$$Y_{lasso} = -1048{,}141 + 37{,}355X_3 + 0{,}121X_4 + 2{,}292X_{13} + 0{,}464X_{14} + 105{,}038X_{17}$$

Model LASSO dengan *Coordinate Descent* menghasilkan persamaan regresi yang menunjukkan lima peubah yang berpengaruh terhadap jumlah kasus TBC, yaitu jumlah puskesmas (X3), jumlah tenaga medis (X4), jumlah sampah TPA (X13), proyeksi penduduk (X14), dan jumlah pengangguran terbuka (X17). Jumlah puskesmas memiliki koefisien positif sebesar 37,355, sehingga setiap penambahan satu puskemas menyebabkan peningkatan dugaan rataan jumlah kasus TBC sebesar 37,355 orang. Jumlah tenaga medis juga memiliki pengaruh positif, yang mana setiap peningkatan satu tenaga medis menyebabkan peningkatan dugaan rataan jumlah kasus TBC sebesar 0,121 orang. Proyeksi penduduk menunjukkan hubungan positif dengan jumlah kasus TBC dan memiliki koefisien sebesar 0,464. Oleh sebab itu, nilai proyesi penduduk yang meningkat sebesar seribu orang menyebabkan peningkatan dugaan rataan jumlah kasus TBC sebesar 0,464 orang. Hal ini sesuai dengan penelitian Banapon *et al.* yang menunjukkan hubungan positif antara peubah jumlah puskemas dan jumlah penduduk terhadap jumlah kasus TBC[9]. Kemudian, jumlah sampah TPA berpengaruh positif dengan jumlah kasus TBC yang mana peningkatan satu ton sampah per hari menyebabkan peningkatan dugaan rataan jumlah kasus TBC sebesar 2,292 kasus. Axmalia dan Mulasari menjelaskan bahwa kondisi TPA yang tidak memenuhi standar akan berdampak pada timbulnya penyakit salah satunya TBC [23]. Sementara itu, jumlah pengangguran terbuka memiliki hubungan positif yang paling besar yaitu kenaikan satu orang pengangguran menyebabkan kenaikan dugaan rataan jumlah kasus sebesar 105,038 kasus.

5. Kesimpulan

Model regresi LASSO dengan *Coordinate Descent* terbukti menjadi model terbaik, menunjukkan RMSE 808.421 dan $R^2{}_{adj}$ sebesar 87,665%. Model ini berhasil mereduksi jumlah peubah penjelas dari 20 menjadi hanya 5 peubah yang signifikan, yaitu X3 (jumlah puskesmas), X4 (jumlah tenaga medis), X13 (jumlah sampah TPA), X14 (proyeksi penduduk), dan X17 (jumlah orang pengangguran terbuka). Meskipun model LASSO

dengan SGD juga efektif, namun menghasilkan RMSE yang sedikit lebih tinggi (819.48) dan mempertahankan lebih banyak peubah (7 peubah). Temuan ini menunjukkan bahwa ketersediaan fasilitas kesehatan, tenaga medis, pengelolaan sampah, proyeksi penduduk, dan tingkat pengangguran memiliki peran krusial dalam menjelaskan variabilitas kasus TBC di Jawa Barat. Dengan demikian, model LASSO berbasis *Coordinate Descent* tidak hanya memberikan prediksi yang akurat, tetapi juga menyajikan model yang lebih ringkas dan mudah diinterpretasikan, yang sangat bermanfaat untuk mendukung perumusan kebijakan penanggulangan TBC yang tepat sasaran.

Daftar Pustaka

[1] World Health Organization, "Global tuberculosis report 2022." [Online]. Available: https://www.who.int/publications/i/item/9789240061729

[2] R. Kemenkes, "Tuberkulosis." [Online]. Available: https://www.tbindonesia.or.id/

[3] I. Onozaki, I. Law, C. Sismanidis, M. Zignol, P. Glaziou, and K. Floyd, "National tuberculosis prevalence surveys in Asia, 1990-2012: An overview of results and lessons learned," *Trop. Med. Int. Heal.*, vol. 20, no. 9, pp. 1128–1145, 2015, doi: 10.1111/tmi.12534.

[4] S. F. Febrilia, B. Lapau, K. Zaman, M. Mitra, and M. Rustam, "Hubungan Faktor Manusia dan Lingkungan Rumah Terhadap Kejadian Tuberkulosis di Wilayah Kerja Puskesmas Rejosari Kota Pekanbaru," *J. Kesehat. Komunitas*, vol. 8, no. 3, pp. 436–442, 2022, doi: 10.25311/keskom.vol8.iss3.618.

[5] T. Kyriazos and M. Poga, "Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions," *Open J. Stat.*, vol. 13, no. 03, pp. 404–424, 2023, doi: 10.4236/ojs.2023.133020.

[6] G. A. Kesse, "Variable selection using LASSO," no. January, 2025.

[7] C. J. Hsieh and I. S. Dhillon, "Fast coordinate descent methods with variable selection for non-negative matrix factorization," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 1064–1072, 2011, doi: 10.1145/2020408.2020577.

[8] J. Yang and G. Yang, "Modified convolutional neural network based on dropout and the stochastic gradient descent optimizer," *Algorithms*, vol. 11, no. 3, 2018, doi: 10.3390/a11030028.

[9] A. Banapon, M. L. P. Putra, and E. Widodo, "Penerapan Regresi Binomial Negatif Untuk Mengatasi Pelanggaran Overdispersi Pada Regresi Poisson (Studi Kasus Penderita Tuberculosis Di Provinsi Jawa Barat Tahun 2017)," *J. Stat. dan Apl.*, vol. 14, no. 1, pp. 39–52, 2020.

[10] S. Chen, K. A. Notodiputro, and S. Rahardiantoro, "Penerapan Analisis Lasso Dan Group Lasso Dalam Mengidentifikasi Faktor-Faktor Yang Berhubungan Dengan Tuberkulosis Di Jawa Barat," *Indones. J. Stat. Its Appl.*, vol. 4, no. 1, pp. 39–54, 2020, doi: 10.29244/ijsa.v4i1.510.

[11] N. D. Ovalingga, N. Amalita, Y. Kurniawati, and Z. Martha, "Regularized Ordinal Regression with LASSO : Identifying Factors in Students ' Public Speaking Anxiety at Universitas Negeri Padang," vol. 2, no. 1999, pp. 475–482, 2024.

[12] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 58, no. 1, pp. 267–288, 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[13] F. K. H. Prabowo, Y. Wilandari, and A. Rusgiyono, "Pemodelan Pertumbuhan Ekonomi Jawa Tengah Menggunakan Pendekatan Least Absolute Shrinkage and Selection Operator (LASSO)," *J. Gaussian*, vol. 4, no. 1996, pp. 855–864, 2015, [Online]. Available: http://ejournal-s1.undip.ac.id/index.php/gaussian

[14] Y. A. Mait, D. Tineke Salaki, and H. A. H. Komalig, "Kajian Model Prediksi Metode Least Absolute Shrinkage and Selection Operator (LASSO) pada Data Mengandung Multikolinearitas LASSO Metode kuadrat terkecil Multikolinearitas," *J. Mat. dan Apl.*, vol. 10, no. 2, pp. 69–75, 2021, [Online]. Available: https://ejournal.unsrat.ac.id/index.php/decartesian

[15] H.-J. M. Shi, S. Tu, Y. Xu, and W. Yin, "A Primer on Coordinate Descent Algorithms," 2016, [Online]. Available: http://arxiv.org/abs/1610.00040

[16] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *J. Stat. Softw.*, vol. 33, no. 1, pp. 1–22, 2010.

[17] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "Stochastic gradient descent training for L1-regularized log-linear models with cumulative penalty," *ACL-IJCNLP 2009 - Jt. Conf. 47th Annu. Meet. Assoc. Comput. Linguist. 4th Int. Jt. Conf. Nat. Lang. Process. AFNLP, Proc. Conf.*, pp. 477–485, 2009, doi: 10.3115/1687878.1687946.

[18] A. Géron, *Hands-on Machine Learning whith Scikit-Learing, Keras and Tensorfow*. 2019.

[19] L. Bottou, "Large-scale machine learning with stochastic gradient descent," *Proc. COMPSTAT 2010 - 19th Int. Conf. Comput. Stat. Keynote, Invit. Contrib. Pap.*, pp. 177–186, 2010, doi: 10.1007/978-3-7908-2604-3_16.

[20] Nurhayati, I. Soekarno, I. K. Hadihardaja, and M. Cahyono, "for Accuracy of Groundwater Modeling in Tidal Lowland Reclamation Using Extreme Learning," *2014 2nd Int. Conf. Technol. Informatics, Manag. Eng. Environ.*, pp. 228–233, 2014.

[21] Y. Widyaningsih, G. P. Arum, and K. Prawira, "Aplikasi K-Fold Cross Validation Dalam Penentuan Model Regresi Binomial Negatif Terbaik," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 15, no. 2, pp. 315–322, 2021, doi: 10.30598/barekengvol15iss2pp315-322.

[22] S. M. Sholihah, N. Y. Aditiya, E. S. Evani, and S. Maghfiroh, "Konsep Uji Asumsi Klasik Pada Regresi Linier Berganda," *J. Ris. Akunt. Soedirman*, vol. 2, no. 2, pp. 102–110, 2023, doi: 10.32424/1.jras.2023.2.2.10792.

[23] A. Axmalia and S. A. Mulasari, "Dampak Tempat Pembuangan Akhir Sampah (TPA) Terhadap Gangguan Kesehatan Masyarakat," *J. Kesehat. Komunitas*, vol. 6, no. 2, pp. 171–176, 2020, doi: 10.25311/keskom.vol6.iss2.536.