NIVEDITA
JOURNAL INFORMATICS
e-ISSN : xxx-xxx
p-ISSN : xxx-xxx

Vol. 01, No. 1, Desember 2024

ANALISIS REGRESI *LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR* (LASSO) TERHADAP WAKTU TAHAN HIDUP PENDERITA STROKE

Putu Riska Wulandari[1], Astrid Krisdayathi[2], Ni Made Rai Kristina[3]
UHN I Gusti Bagus Sugriwa Denpasar [1,2,3]

email: riskawulandari@uhnsugriwa.ac.id

Abstract

Survival analysis is a statistical method used to study factors affecting the time until the occurrence of a specific event, such as death, disease, or relapse. In the context of stroke patients, identifying significant factors influencing survival time is crucial to support medical decision-making and clinical interventions. This study aims to apply the Least Absolute Shrinkage and Selection Operator (LASSO) method to analyze the survival time data of stroke patients. The LASSO method was chosen for its ability to perform variable selection and regularization simultaneously, enabling the generation of a simple yet accurate model. The data utilized includes clinical and demographic variables of stroke patients, with the Kaplan-Meier method used to estimate the survival function and Cox-LASSO regression employed to identify significant variables. The patient data analyzed include microarray data, where multicollinearity was observed. To address multicollinearity and overfitting, the LASSO method was applied to identify significant factors affecting the survival time of stroke patients. Additionally, using LASSO regression on microarray data revealed a lack of insight into which independent variables contribute to the dependent variable. In this study, data from seven patients were analyzed using R software version 2.12.1 with the "lars" library. The analysis employed the LASSO regression model with predefined function structures within the R package. Key metrics analyzed included parameter estimates using the beta matrix, the D value, and the t value. The conditions $D \leq tD \leq t$ and $t \geq 0t \geq 0$, indicating the LASSO constraints, were fulfilled. The results demonstrate that the LASSO method is effective in handling data with numerous predictors and capable of eliminating insignificant variables, thus improving the model's interpretability. These findings are expected to contribute to the clinical management of stroke patients $\hat{y} = -0{,}01x_2 - 0{,}13x_7 + (0{,}864)$ *and the development of data-driven health policies. The study's findings yielded a survival model for patients, identifying initial examination conditions and affected neural regions as significant influencing factors*.

Keywords: *LASSO, Overfitting, Time Survival, Microarray, Multikolinearty*

Abstrak

Analisis waktu tahan hidup (survival analysis) merupakan metode statistik yang digunakan untuk mempelajari faktor-faktor yang memengaruhi waktu hingga terjadinya suatu peristiwa tertentu, seperti kematian, penyakit, atau kekambuhan. Dalam konteks penderita stroke, identifikasi faktor signifikan yang memengaruhi waktu tahan hidup sangat penting untuk mendukung pengambilan keputusan medis dan intervensi klinis. Penelitian ini bertujuan untuk menerapkan metode Least Absolute Shrinkage and Selection Operator (LASSO) dalam menganalisis data waktu tahan hidup penderita stroke. Metode LASSO dipilih karena kemampuannya dalam melakukan seleksi variabel dan regularisasi secara simultan, sehingga mampu menghasilkan model yang sederhana namun akurat. Data yang digunakan mencakup variabel klinis dan demografis penderita stroke, dengan metode Kaplan-Meier digunakan untuk mengestimasi fungsi survival dan regresi Cox-LASSO untuk mengidentifikasi variabel-variabel signifikan. Data pasien yang ada dalam konteks data *microarray* dan terjadi multikolinearitas pada data pasien stroke. Untuk mengatasi adanya multikolinearitas dan *overfitting*, maka metode LASSO dapat digunakan untuk mengetahui faktor-faktor yang signifikan berpengaruh terhadap masa hidup penderita stroke, selain menggunakan regresi LASSO terhadap data microarray mengakibatkan tidak diketahuinya variabel bebas yang berkonstribusi terhadap variabel tak bebas. Pada penelitian ini, sebanyak tujuh data pasien digunakan dan dianalisis dengan menggunakan bantuan *software* R 2.12.1 dengan *library lars*. Data dianalisis dengan model regresi LASSO dengan struktur fungsi yang telah ada dalam paket R. Data-data yang dicari dalam analisis yaitu nilai estimasi parameter dengan matrik beta, nilai D dan nilai t. Nilai dari $D \leq t$ dan $t \geq 0$ yang berarti batasan

dari LASSO tersebut telah terpenuhi. Hasil penelitian menunjukkan bahwa metode LASSO efektif dalam menangani data dengan banyak prediktor serta mampu mengeliminasi variabel yang tidak signifikan, sehingga meningkatkan interpretabilitas model. Temuan ini diharapkan dapat memberikan kontribusi dalam pengelolaan klinis penderita stroke serta pengembangan kebijakan kesehatan berbasis data. Dari hasil penelitian diperoleh model masa tahan hidup pasien adalah $\hat{y} = -0{,}01x_2 - 0{,}13x_7 + (0{,}864)$ dan faktor yang signifikan berpengaruh adalah kondisi awal pemeriksaan dan bagian saraf yang mengalami gangguan.

Kata kunci: LASSO, *Overfitting*, Waktu Tahan Hidup, *Microarray, Multikolinearty*

Diajukan: 26 Desember 2024; Diterima: 22 Januari 2025;

PENDAHULUAN

Stroke merupakan salah satu penyebab utama kematian dan disabilitas di seluruh dunia. Data Organisasi Kesehatan Dunia (WHO) menunjukkan bahwa lebih dari 15 juta orang di seluruh dunia mengalami stroke setiap tahun, dengan sepertiganya meninggal dan sepertiganya mengalami kecacatan permanen [1] . Kondisi ini menekankan pentingnya analisis terhadap faktor-faktor yang memengaruhi waktu tahan hidup penderita stroke untuk mendukung intervensi medis yang lebih baik dan efisien.

Analisis waktu tahan hidup (*survival analysis*) merupakan metode yang sering digunakan untuk mempelajari waktu hingga terjadinya suatu peristiwa, seperti kematian atau kekambuhan penyakit. Dalam analisis ini, pemilihan variabel yang signifikan memiliki peranan penting, terutama ketika data melibatkan banyak prediktor yang berpotensi menyebabkan multikolinearitas. Multikolinearitas dapatmenyebabkan distorsi dalam estimasi parameter regresi, sehingga mengurangi keakuratan model [2] .

Banyak metode untuk melakukan analisis waktu tahan hidup seperti Metode Kaplan-Meier, Uji Long-Rank, Model Regresi Cox Proportional Hazards, Regresi Parametrik, Random Survival Forests, dan lainnya. Namun Regresi Least Absolute Shrinkage and Selection Operator (LASSO) adalah salah satu metode statistik yang dapat digunakan untuk mengatasi masalah multikolinearitas sekaligus melakukan seleksi variabel dengan klinis yang banyak . LASSO bekerja dengan menambahkan penalti terhadap koefisien regresi, sehingga hanya variabel yang signifikan yang tetap ada dalam model, sementara variabel lain dieliminasi [3]. Keunggulan metode ini membuatnya menjadi pilihan yang efektif dalam analisis data dengan jumlah variabel yang besar, seperti data klinis pasien stroke yang sering kali melibatkan banyak variabel demografis dan klinis.

Analisis tahan hidup digunakan dalam menganalisis data dengan *failure event* atau gagal pada waktu dari *time origin* (awal) hingga titik akhir. Dalam konteks analisis survival pasien stroke, metode LASSO dapat membantu mengidentifikasi faktor-faktor penting yang memengaruhi waktu tahan hidup pasien, seperti kondisi awal kesehatan, riwayat penyakit, dan area otak yang terdampak. Dengan demikian, metode ini memberikan kontribusi penting dalam memahami dinamika penyakit stroke sekaligus mendukung pengambilan keputusan klinis berbasis data [4]. Penelitian dalam dunia kesehatan, data kejadian tertentu, seperti data kematian, seringkali dijumpai.

Pada umumnya, data tersebut berupa data tahan hidup. Penelitian ini tidak hanya mengutamakan hasil kejadian, seperti kematian, namun lebih kepada masa sampai pada keadaan tertentu. Dalam menentukan masa tahan hidup, terdapat tiga faktor yang dibutuhkan yaitu *time origin/ starting point* (awal) suatu kejadian., *end-point* (masa akhir) suatu kejadian, pengukuran mengenai bagian waktu harus jelas. Jika masa akhir dari penelitian adalah kematian seorang pasien, maka data hasil tersebut dikatakan sebagai masa tahan hidup. Kematian bukanlah ujung suatu kejadian, dapat juga mengenai sembuhnya pasien dari suatu penyakit, berkurang dari gejala penyakit, atau kambuhnya pasien karena kondisi tertentu. Model tahan hidup banyak digunakan dalam menguji hubungan antara masa tahan hidup dengan satu atau lebih dari satu variabel bebas [5].

Karakteristik data dari analisis tahan hidup ada dua, yang pertama yaitu data tahan hidup secara general dengan distribusi yang nonsimetri. Pada khususnya, histogram dibentuk dari masa tahan hidup dari

kelompok homogen yang berbentuk *skewness* positif. Hal ini dilihat dari histogram yang miring ke kanan. Kedua, data masa tahan hidup dari masa awal hingga terjadinya *failure time* memang sering tersensor. Masa tahan hidup dari individu dikatakan tersensor ketika titik akhir dari pengamatan tidak jelas untuk individu [6] . Data dalam analisis tahan hidup terdiri dari data lengkap dan tidak lengkap. Data lengkap merupakan data yang tidak tersensor, sedangkan data tidak lengkap merupakan data yang tersensor. Data yang tersensor adalah data yang tahan hidupnya tidak bisa diketahui pasti dan data tidak tersensor adalah data yang masa tahan hidupnya diketahui secara pasti.

Analisis tahan hidup digunakan dalam analisis data dengan *failure event* atau gagal pada satu waktu dari kondisi *time origin* hingga akhir kejadian atau titik akhir [6]. Data tahan hidup (*survival*) sering dijumpai dalam penelitian biostatistika atau bidang kedokteran seperti penelitian mengenai penyakit stroke. Untuk meramalkan masa tahan hidup pasien stroke, data yang dianalisis merupakan data dengan jumlah variabel independen lebih banyak dari jumlah data sampel yang dipakai (*microarray*) dan adalah data yang tidak diketahui secara pasti masa tahan hidup sampelnya (data tersensor), sehingga akan muncul peluang terjadinya korelasi antara variabel bebas yang dikenal dengan istilah terjadi multikolinearitas dan terjadinya pendugaan yang bias atau *overfitting* [5].

Penggunaan data microarray dalam analisis waktu tahan hidup penderita stroke memiliki beberapa alasan penting, terutama terkait kompleksitas data biomedis dan kebutuhan untuk memahami mekanisme biologis yang mendasarinya. Data microarray digunakan karena kemampuannya untuk menangkap informasi ekspresi gen yang luas dan kompleks. Informasi ini memberikan peluang besar untuk memahami patofisiologi stroke, mengidentifikasi faktor risiko genetik, dan mengembangkan model prediksi yang lebih akurat. Teknik seperti LASSO regresi sangat cocok untuk analisis data microarray karena dapat menangani banyak variabel dan mengurangi risiko overfitting [7].

Regresi LASSO merupakan salah metode yang dipakai untuk membahas masalah multikolinearitas dan overfitting. Untuk menentukan faktor-faktor yang berpengaruh terhadap masa tahan hidup sampel dengan menggunakan metode LASSO merupakan hal yang tepat, karena jika diselesaikan dengan regresi linear berganda maka nilai signifikansinya tidak diketahui yang akan mengakibatkan tidak diperoleh variabel bebas yang berkonstribusi terhadap variabel tak bebas [8], sehingga tidak diketahui faktor-faktor yang signifikan berpengaruh terhadap masa tahan hidup penderita stroke.

Regresi LASSO hampir sama dengan regresi linear berganda, akan tetapi yang membedakan adalah variabel-variabel bebas pada regresi linear berganda tidak terjadi multikolinearitas karena tidak berkorelasi sempurna diantara variabel-variabel bebas, sedangkan pada regresi LASSO memiliki hubungan yang linear sempurna antara beberapa atau semua variabel independen dari suatu model persamaan regresi sehingga terjadi multikolinearitas [9]. Penelitian ini bertujuan untuk menerapkan analisis regresi LASSO pada data waktu tahan hidup penderita stroke, dengan menggunakan perangkat lunak statistik untuk mengestimasi model dan menganalisis faktor-faktor signifikan.

Penelitian yang dilakukan Meng et al terkait Algoritma LASSO digunakan untuk memilih variabel penting mengembangkan model prediksi baru berdasarkan data awal pemeriksaan untuk membantu identifikasi dini stroke iskemik (IS) [10]. Sehingga peneliti ingin mengetahui bagaimana model waktu tahan hidup penderita stroke sangat tepat digunakan. Sehingga nanti akan diperoleh variabel yang betpegaruh terhadap waktu tahan hidup penderita stroke.

METODE PENELITIAN

Metode penelitian yang digunakan dalam analisis Regresi Least Absolute Shrinkage and Selection Operator (LASSO) terhadap waktu tahan hidup penderita stroke adalah penelitian kuantitatif dengan pendekatan analisis survival**.** Penelitian ini menggunakan data numerik untuk mengukur hubungan antara variabel bebas (faktor klinis dan demografis) dan variabel terikat (waktu tahan hidup penderita stroke). Fokusnya adalah mempelajari waktu hingga terjadinya suatu peristiwa, seperti kematian pada pasien stroke.

Penelitian menggunakan jenis penelitian kuantitatif dengan menggunakan data primer berasal dari wawancara terhadap pasien stroke atau yang mendampingi. Untuk data sekunder diperoleh melalui rekam medik pasien tahun 2010 untuk mendukung data primer yang tidak diperoleh dari hasil wawancara.

Tabel 1. Variabel Penelitian

Variabel	Keterangan	Skala	Parameter
Variabel Tak Bebas	Waktu Tahan Hidup (Y)	Rasio	Hari
Variabel Bebas	Jenis Kelamin (X1)	Kategorik Nominal	0 = Perempuan
			1 = Laki-laki
	Kondisi Pertama kali diperiksa (X2)	Interval	<30 % = awal
			30%-50% = stroke ringan
			51%-70% = Stroke
			>70% = stroke berat (stroke berulang)
	Umur (X3)	Rasio	Tahun
	Berat Badan (X4)	Rasio	Kg
	Kebiasaan konsumsi alkohol (X5)	Kategorik Nominal	0 = Tidak
			1 = Ya
	Jumlah rokok yang dikonsumsi (X6)	Interval	berat ≥ 12 batang/hari
			sedang 2-11 batang/hari
			Ringan 1 batang/hari
	Jumlah saraf yang bermasalah (X7)	Rasio	Bagian tubuh
	Waktu pertama kali sakit sampai diteliti (X8)	Rasio	Hari

Metode LASSO menghasilkan varians kecil dari estimator dan mencapai estimasi dan prediksi yang baik dan mengkerutkan koefisien yang lebih kecil ke nol. Dua kelebihan tersebut yaitu keakuratan estimasi dan seleksi variabel konsisten dapat dicapai secara bersamaan. Pendekatan LASSO adalah sebuah alternatif untuk regresi standar dan teknik *Subset Variabel Selection*. Jika jumlah variabel bebas besar namun memiliki konstribusi yang kecil terhadap variabel tak bebas maka pendugaan dengan *Ridge Regression* adalah pilihan terbaik, sedangkan jika jumlah variabel bebas kecil memiliki konstribusi besar terhadap variabel tak bebas maka *Subset Variabel Selection* dapat melakukannya dengan baik. Jika jumlah variabel bebas sedang dan memiliki konstribusi sedang terhadap variabel tak bebas, maka LASSO adalah pilihan terbaik.

Mengkerutkan ukuran koefisien regresi dan seleksi variabel merupakan tujuan penting dalam analisis data *microarray*, dimana jumlah sampel yang dikumpulkan jauh lebih kecil daripada jumlah variansi yang diperoleh, yaitu jumlah variabel bebas jauh lebih besar dibandingkan jumlah sampel. Pendugaan dalam variansi jumlah besar dan sampel yang rendah dilakukan dengan menggunakan LASSO [5].

Misalkan $(x_i, y_i),...,(x_n, y_n)$ adalah n pasangan variabel independen atau variabel dependen dimana $y_i \in Y$ serta $x_i \in X$. Disini, Y dan X merupakan domain input dan output, Y merupakan variabel dependen dan terdapat p variabel independen $X_1, X_2,...,X_p$. Misalkan terdapat n buah pengamatan, maka model regresi linearnya yaitu:

NIVEDITA
JOURNAL INFORMATICS
e-ISSN : xxx-xxx
p-ISSN : xxx-xxx
Vol. 01, No. 1, Desember 2024

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i \quad (1)$$

Dimana ε_i berdistribusi normal dengan mean nol variansi σ^2. Estimasi $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ pada LASSO yaitu dengan meminimumkan:

$$R(\beta_0, \beta_j) = \frac{1}{n} \sum \left(Y_i - \hat{\beta}_0 - \sum_{j=1}^{p} X_{ij} \hat{\beta}_j \right)^2 \quad (2)$$

$$(\beta)_{l_1} = \sum_{i=1}^{n} \left(y_i - X_i^T \beta \right)^2 + \lambda \sum_{j=1}^{p} \left| \hat{\beta}_j \right|$$
$$= (y - X\beta)^T (y - X\beta) + \lambda \left\| \hat{\beta} \right\|_1$$

λ merupakan suatu parameter yang mengendalikan koefisien LASSO yang harus diatur melalui batasan $D = \sum_{j=1}^{p} \left| \hat{\beta}_j \right| \leq t$, dengan meminimumkan

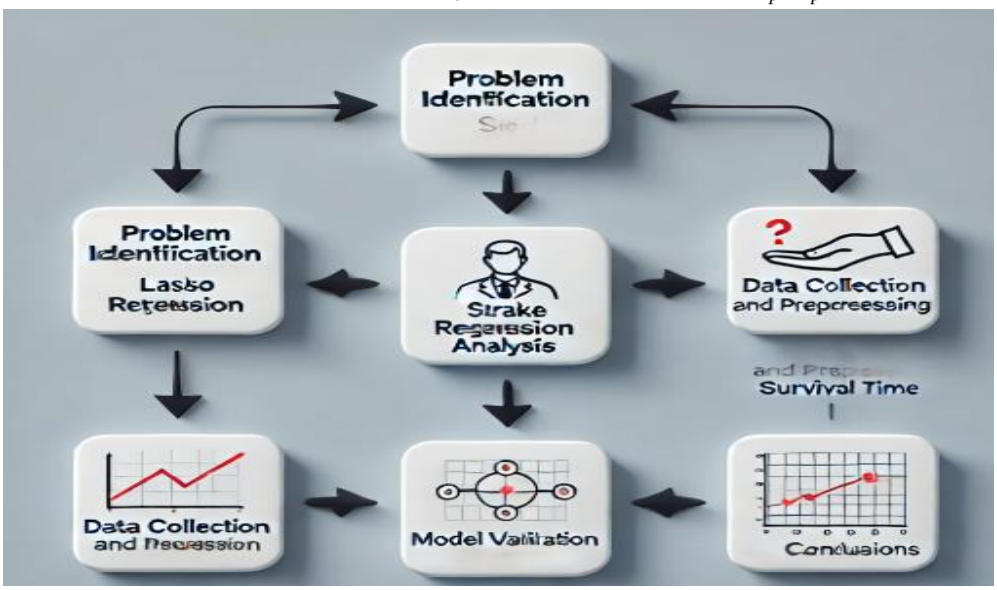
$$l_1 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \sum_{j=1}^{p} X_{ij} \hat{\beta}_j \right)_2^2 + t \sum_{j=1}^{p} \left| \hat{\beta}_j \right|_1 \quad (3)$$

dengan t adalah parameter pengkerutan dengan $t \geq 0$. Pada LASSO, peminimuman jumlah kuadrat galat dengan menggunakan metode kuadrat terkecil ini dinyatakan dengan penambahan $D = \sum_{j=1}^{p} \left| \hat{\beta}_j \right|$ sebagai koefisiennya pada vektor parameter dari masalah peminimalan l_2 dengan l_1 harus bernilai minimum [11].

$$l_2 = \sum_{i=1}^{n} \left(y_i - X_i^T \beta \right)^2 + \lambda \sum_{j=1}^{p} \left| \hat{\beta}_j \right| = \sum_{i=1}^{n} \left(y_i - X_i^T \beta \right)^2 + \sum_{j=1}^{p} \left(0 - \sqrt{\lambda} \beta_j \right)^2 \quad (4)$$

Jadi, model LASSO dapat dinyatakan sebagai berikut:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p + D \quad (5)$$



Gambar 1. Alur Penelitian

Alur penelitian pada gambar 1 dimana *Problem Identification* sebagai langkah pertama adalah mengidentifikasi masalah utama, yaitu mengetahui faktor-faktor signifikan yang memengaruhi waktu tahan hidup penderita stroke. Peneliti juga menetapkan tujuan penelitian, seperti membangun model prediktif menggunakan metode LASSO. Pada langkah kedua yaitu *Data Collection and Preprocessing* dengan mengumpulkan data pasien stroke, termasuk variabel klinis, demografis, dan waktu tahan hidup dan bersihkan data, atasi nilai hilang, dan periksa adanya multikolinearitas antar variabel. Langkah ketiga *LASSO Regression Analysis* dengan Terapkan metode LASSO untuk memilih variabel yang signifikan dan membangun model regresi survival. LASSO digunakan untuk mengatasi multikolinearitas dan menghasilkan model yang sederhana namun efektif. Langkah keempat yaitu *Model Validation* menggunakan teknik validasi silang untuk mengevaluasi performa model. Hitung metrik seperti indeks C-statistic untuk menilai akurasi dan diskriminasi model. Langkah terakhir adalah Interpretasikan hasil analisis, seperti faktor signifikan yang memengaruhi waktu tahan hidup pasien.

Pada penelitian ini, proses analisis data dengan bantuan *software* R dan SPSS. Program R merupakan bahasa pemrograman yang didukung oleh banyak paket (*library*) yang terkait dengan analisis data dan ilustrasi grafik. Untuk menganalisis data menggunakan model regresi LASSO digunakan R 2.12.1 dengan *library lars* untuk model regresi LASSO. Langkah-langkah dalam analisis data riil adalah sebagai berikut:

1. Memasukkan data pengamatan berupa matriks pada program R
2. Menguji multikolinearitas
3. Menganalisis data dengan model regresi LASSO dengan struktur fungsi yang telah ada dalam paket R, yaitu
4. Menentukan Fit model LASSO dari data untuk menentukan nilai estimasi yang sesuai dari waktu tahan hidup. Dengan struktur fungsi dalam paket R yaitu:
   predict.lars(...,X,type="fit")
5. Mencari nilai estimasi parameter

$$R\left(\hat{\beta}_0, \hat{\beta}_j\right) = \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \sum_{j=1}^{n} X_{ij}\hat{\beta}_j\right)^2 \quad (6)$$

6. Mencari nilai $D = \sum_{j=1}^{n}\left|\hat{\beta}_j\right|$ yang minimum, dengan $D \le t$, dan t adalah parameter tuning, dimana nilai t yang dipilih dari *cross-validation* yang minimum. Variabel yang memiliki parameter nol akan dihilangkan dari model.
7. Uji hipotesis dengan uji statistik t dalam program SPSS
8. Interpretasi model

HASIL DAN PEMBAHASAN

Hasil pengolahan data menunjukkan bahwa kasus data stroke dengan kondisi pertama kali diteliti dengan kondisi keparahan penyakit 60%-80% dengan jumlah saraf yang mengalami gangguan 6-8 saraf. Rata-rata pasien merupakan perokok aktif yang mengkonsumsi rokok 9-16 batang perhari dengan umur penderita stroke yaitu 55-69 tahun memiliki waktu tahan hidup rata-rata terhadap penyakit stroke kurang lebih 368 hari. Dari matrik korelasi pada gambar 2 dapat dilihat bahwa nilai korelasi terbesar terjadi pada jumlah saraf yang mengalami gangguan (X_7) dan waktu pertama kali sakit sampai diteliti (X_8) yaitu -0,82 dengan signifikansinya 0,023, jumlah saraf yang mengalami gangguan (X_7) dan kondisi pada saat pertama kali diperiksa (X_2) yaitu 0,79 dengan signifikansinya 0,035. Estimasi parameter untuk model LASSO didekati dengan menggunakan matrik dari beta pada gambar 3. Adapun hasil estimasi parameter LASSO dapat dilihat pada matrik pada gambar 3 berikut

$$R = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{array} \begin{array}{c} \begin{array}{cccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 \end{array} \\ \begin{bmatrix} 1 & 0{,}06 & 0{,}06 & 0{,}07 & 0{,}64 & 0{,}20 & -0{,}14 & -0{,}49 \\ 0{,}06 & 1 & 0{,}09 & 0{,}04 & 0{,}32 & 0{,}28 & 0{,}79 & -0{,}52 \\ 0{,}06 & 0{,}09 & 1 & 0{,}07 & 0{,}52 & -0{,}16 & 0{,}43 & -0{,}44 \\ 0{,}07 & 0{,}04 & 0{,}07 & 1 & 0{,}10 & -0{,}51 & 0{,}06 & -0{,}33 \\ 0{,}64 & 0{,}32 & 0{,}52 & 0{,}10 & 1 & 0{,}05 & 0{,}55 & -0{,}21 \\ 0{,}20 & 0{,}28 & -0{,}16 & -0{,}51 & -0{,}05 & 1 & -0{,}02 & -0{,}15 \\ -0{,}14 & 0{,}79 & 0{,}43 & 0{,}06 & 0{,}55 & -0{,}02 & 1 & -0{,}82 \\ -0{,}49 & -0{,}52 & -0{,}44 & -0{,}33 & -0{,}21 & -0{,}15 & -0{,}82 & 1 \end{bmatrix} \end{array}$$

Gambar 2. Matrik Korelasi

Matrik $\hat{\beta}_0$

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0{,}13 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0{,}15 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0{,}14 & 0 \\ 0 & -0{,}01 & 0 & 0 & 0 & 0 & -0{,}13 & 0 \\ 0 & -0{,}01 & 0 & 0 & 0 & 0 & -0{,}13 & 0 \\ 0 & -0{,}01 & -0{,}01 & 0 & -0{,}02 & -0{,}01 & -0.09 & 0 \end{bmatrix} \end{array}$$

Gambar 3. Matrik Estimasi Parameter LASSO

Kolom pada matrik beta di atas menunjukkan variabel bebas, sedangkan barisnya menunjukkan langkah (*step*) dalam proses LASSO. Dari matrik beta di atas variabel yang memiliki parameter nol akan dikerutkan dari model. Dari matrik beta dapat dilihat bahwa X_1, X_4, dan X_8 yang dikerutkan karena memiliki nilai beta sama dengan nol.

Selanjutnya, dengan menggunakan plot LASSO pada Gambar 3, maka dapat juga diketahui variabel bebas yang berpengaruh signifikan dan nilai koefisiennya dalam matrik beta berdasarkan *step* yang terpilih dari plot LASSO.

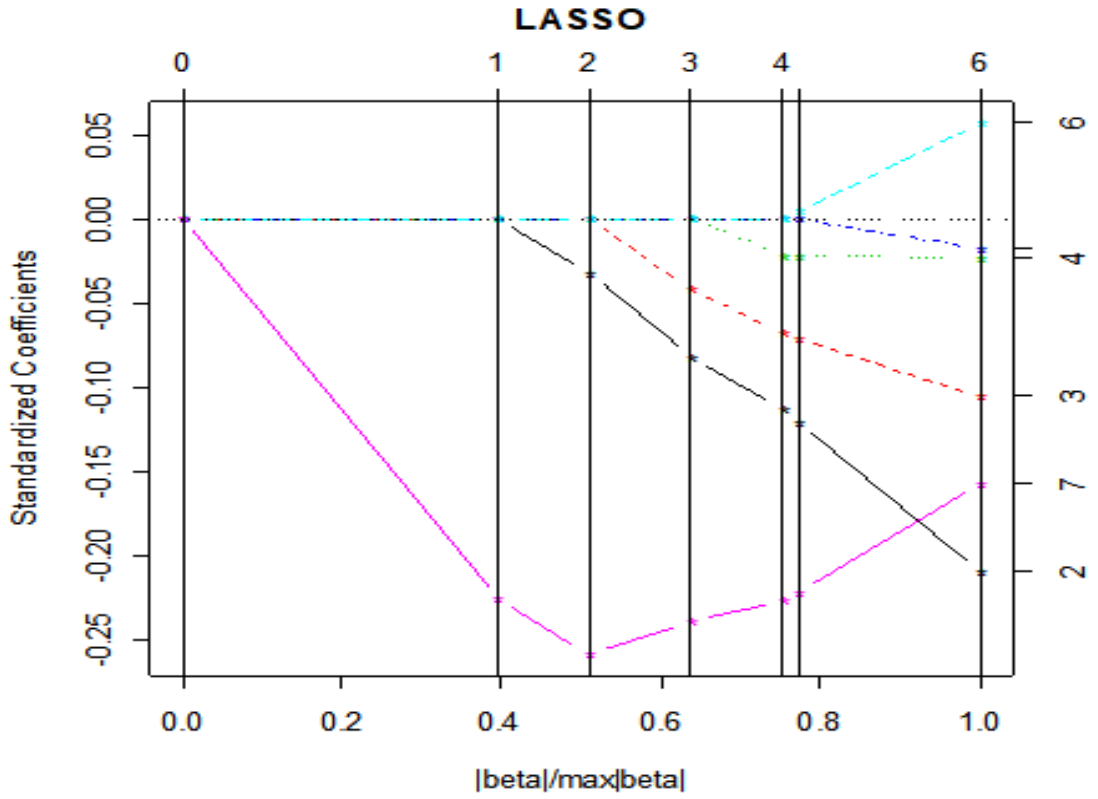
Batas dari LASSO yaitu $D = \sum_{j=1}^{p} \left|\hat{\beta}_j\right| \leq t$, dengan $t \geq 0$ adalah parameter tuning. Nilai dari D yaitu 0,864 dan nilai t yaitu 7,036 dimana nilai t ini diperoleh dari nilai *cross-validation* yang paling minimum. Karena nilai $t \geq 0$ dan D≤t, maka batasan dari LASSO tersebut telah terpenuhi. Sedangkan untuk mengetahui variabel bebas X yang mempengaruhi Y dapat dilihat pada gambar 4 plot Lasso. Plot LASSO pada variabel yang dikerutkan berdasarkan batasan LASSO.

Keidentikan LASSO dengan OLS dapat dilihat pada sumbu x, sedangkan yang dimaksudkan dengan $|beta|/\max|beta|$ adalah jumlah dari seluruh parameter LASSO dibagi dengan penduga LASSO atau rasio antara LASSO dengan OLS. Semakin identik LASSO dengan OLS maka nilainya akan mendekati 1,0. Nilai 0,1,2,…,6 pada bagian atas plot menunjukkan langkah pada proses LASSO. Langkah pada proses LASSO yang dipilih adalah yang menghasilkan nilai paling minimum.

Nilai pada bagian kanan gambar menunjukkan variabel bebas yang tersisa setelah pengerutan dengan menggunakan matrik beta. Untuk mencari nilai koefisien beta dapat dilihat dari perpotongan garis beta dengan standar koefisien.

Dari Gambar 4 dapat dilakukan pemilihan langkah yang paling dekat dengan batasan LASSSO yaitu $\sum_{j=1}^{p} \left|\hat{\beta}_j\right|$. Pada gambar tampak bahwa langkah kelima adalah langkah yang paling dekat dengan batasan LASSO, dilihat dari sisi kiri.

Setelah mengetahui bahwa pada langkah kelima yang paling dekat dengan batasan LASSO, kemudian kembali melihat matriks beta. Pilih pada langkah kelima yang memiliki nilai beta tidak sama dengan nol. Dari matrik beta pada langkah kelima variabel X_2 dan X_7 yang memiliki nilai beta tidak sama dengan nol.



Gambar 4. Plot LASSO

Jadi, dari matriks beta dapat diambil kesimpulan bahwa variabel independent yang memiliki pengaruh signifikan terhadap variabel dependen yaitu waktu tahan hidup penderita stroke adalah konstanta, variabel X_2 yaitu kondisi awal diperiksa dan X_7 yaitu banyak saraf mengalami gangguan. Sehingga diperoleh model LASSO sebagai berikut:

$$\hat{y} = -0{,}01x_2 - 0{,}13x_7 + (0{,}864) \quad (7)$$

Sehingga dapat disimpulkan bahwa kondisi pertama kali diperiksa dan jumlah saraf yang terganggu memiliki pengaruh yang signifikan terhadap masa tahan hidup penderita stroke. Dari model LASSO yaitu $\hat{y} = -0{,}01x_2 - 0{,}13x_7 + (0{,}864)$, dimana model ini setara dengan peminimalan jumlah kuadrat galat ditambahkan dengan batasan LASSO yaitu D pada koefisien regresinya sebesar 0,864.

Berdasarkan model LASSO yang diperoleh, bahwa hasil analisis model LASSO untuk delapan variabel bebas menunjukkan bahwa kondisi pertama kali diperiksa dan jumlah bagian saraf yang mengalami masalah, memiliki pengaruh signifikan terhadap masa tahan hidup penderita stroke.

KESIMPULAN

Regresi LASSO merupakan metode tepat digunakan untuk menyeesaikan masalah munculnya multikolinearitas dengan seleksi variabel independens. Regresi LASSO digunakan untuk menyelesaikan masalah *overfitting* saat data yang digunakan merupakan data *microarray* dengan cara pengerutan variabel bebas. Model yang diperoleh dengan menggunakan LASSO adalah $\hat{y} = -0{,}01x_2 - 0{,}13x_7 + (0{,}864)$, dimana model ini setara dengan peminimalan jumlah kuadrat galat ditambahkan dengan batasan LASSO yaitu D pada koefisien regresinya.

Dari model LASSO diperoleh bahwa faktor yang signifikan berpengaruh terhadap masa tahan hidup penderita stroke adalah kondisi awal diperiksa dan banyak bagian saraf yang mengalami gangguan, dimana pada kasus data penderita stroke di RSUP Sanglah yang telah melalui tahap stroke berulang rata-rata memiliki waktu tahan hidup ±368 hari.

DAFTAR PUSTAKA

[1] World Health Organization, “Stroke: Key facts,” [online] Available: https://www.who.int. [Accessed: Jan. 26, 2025].
[2] D.W. Hosmer, S. Lemeshow, and S. May, *Applied Survival Analysis: Regression Modeling of Time to Event Data*, Wiley-Interscience, 2011.
[3] R. Tibshirani, “Regression shrinkage and selection via the Lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
[4] J. Simon, H. Kohn, M. S. Y. Lee, and S. M. Yang, "Lasso regression in survival analysis," *Journal of Clinical Research*, vol. 24, no. 3, pp. 213-220, 2011.
[5] D. Recchia, G. P. Smith, and P. R. Kelly, "Model survival in healthcare research," *Biostatistics & Epidemiology*, vol. 4, no. 2, pp. 189-198, 2006.
[6] D. Lawless, *Statistical Models and Methods for Lifetime Data*, Wiley, 1982.
[7] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 58, no. 1, pp. 267–288, 1996.
[8] D. Montgomery, *Design and Analysis of Experiments*, Wiley, 1991.
[9] N. Sari, "Multikolinearitas dalam regresi LASSO," *Jurnal Statistika Terapan*, vol. 2, no. 1, pp. 45-52, 2008.
[10] Meng, et al., "Application of LASSO algorithm for developing a predictive model for early identification of ischemic stroke," *Journal of Stroke Research*, vol. 32, no. 2, pp. 112-120, 2021.

[11] Yongdai, "LASSO regression and its application in statistical modeling," *Journal of Statistical Methods*, vol. 15, no. 3, pp. 45-52, 2004.