Lattice Journal : *Journal of Mathematics Education and Applied* e-ISSN: 2798-4621 p-ISSN: 2798-6470
Vol. 5 No. 2, Juli-Desember 2025 pp. 189-202 DOI:http://dx.doi.org/10.30983/lattice.v5i2.10376

# Optimisasi Model Regresi Linier Menggunakan Pendekatan Teori Rough Set

**Lita Lovia[1], Yessy Yusnita[2], Izzati Rahmi[3], Widdya Rahmalina[4]**
[1]Universitas Dharma Andalas, Padang, Indonesia
[2]Institut Teknologi Padang, Padang, Indonesia
[3]Universitas Andalas, Padang, Indonesia
[4]Universitas Adzkia, Padang, Indonesia

**Informasi Artikel**

Diterima Redaksi: 11 November 2025
Revisi Akhir: 21 Desember 2025
Diterbitkan *Online*: 31 Desember 2025

**Kata Kunci**

Rough Set Theory
Regresi Linier Berganda
Data Reduction

**Korespondensi**

E-mail: litalovia80@gmail.com

A B S T R A C T

*Linear regression is widely used to model student performance data; however, its effectiveness can decrease when applied to datasets containing inconsistent samples, which affects the clarity and stability of the model. This study explores the use of Rough Set Theory (RST) as a data reduction approach to improve the quality of linear regression modeling. RST is applied in the pre-modeling stage to identify and reduce inconsistent samples through two schemes: majority-keep reduction and strict reduction. Linear regression models are then built using the reduced datasets and compared with the initial model based on the coefficient of determination ($R^2$) and classical regression assumption tests. The results show an increase in $R^2$ from 0.624 in the initial model to 0.741 with RST majority-keep and to 0.862 with RST strict reduction, indicating improved model fit after data reduction, and the classical regression assumptions are satisfied. These findings suggest that integrating RST improves the diagnostic quality and stability of linear regression, with majority-keep reduction providing an optimal balance between enhancing model and maintaining a representative sample size.*

Regresi linier banyak digunakan untuk memodelkan data student performance. Namun, efektivitasnya dapat menurun ketika diterapkan pada data yang mengandung sampel inkonsisten sehingga memengaruhi kejelasan dan kestabilan model. Penelitian ini mengkaji penggunaan Rough Set Theory (RST) sebagai pendekatan reduksi data untuk meningkatkan kualitas pemodelan regresi linier pada data student performance. RST diterapkan pada tahap pra-pemodelan untuk mengidentifikasi dan mereduksi sampel yang tidak konsisten melalui dua skema reduksi, yaitu majority-keep reduction dan strict reduction. Model regresi linier kemudian dibangun menggunakan dataset hasil reduksi dan dibandingkan dengan model awal berdasarkan nilai koefisien determinasi ($R^2$) dan hasil pengujian asumsi klasik regresi. Hasil penelitian menunjukkan bahwa nilai $R^2$ meningkat dari 0,624 pada model awal menjadi 0,741 pada model dengan RST majority-keep dan 0,862 pada model dengan RST strict reduction, yang menunjukkan peningkatan kecocokan model pada data yang dianalisis setelah dilakukan reduksi data dan uji asumsi klasik terpenuhi. Analisis ini mengindikasikan bahwa integrasi RST berkontribusi pada peningkatan kualitas diagnostik dan stabilitas model regresi linier melalui reduksi data. Di antara kedua skema reduksi, RST majority-keep memberikan keseimbangan yang lebih baik antara perbaikan model dan mempertahankan ukuran sampel yang representatif.
.



## 1. Pendahuluan

Model regresi linier merupakan salah satu metode paling umum dan mudah digunakan dalam analisis prediktif. Keunggulannya terletak pada kesederhanaan, kecepatan komputasi, serta kemudahan interpretasi [1][2][3], sehingga banyak diterapkan di berbagai bidang seperti ekonomi, psikologi, kedokteran, dan kesehatan Masyarakat [4][5]. Regresi linier memungkinkan peneliti mengidentifikasi serta memodelkan hubungan antara variabel independen dan dependen, sekaligus memprediksi bagaimana perubahan pada variabel independen memengaruhi variabel dependen.

Namun, performa regresi linier sangat bergantung pada pemenuhan sejumlah asumsi statistik, antara lain linearitas antarvariabel, normalitas residual, homoskedastisitas, dan tidak adanya multikolinearitas [6]. Dalam praktiknya, data yang digunakan sering kali mengandung

ketidakpastian, ketidaktepatan, dan inkonsistensi, misalnya adanya sampel yang tidak konsisten atau atribut yang tidak relevan. Kondisi ini dapat menurunkan akurasi prediksi dan validitas model. Lebih jauh, regresi linier tidak memiliki mekanisme internal untuk mendeteksi maupun menangani inkonsistensi sampel, yaitu situasi ketika dua atau lebih observasi memiliki atribut identik tetapi menghasilkan keluaran berbeda [7].

Sampel yang tidak konsisten dan atribut yang tidak relevan dapat memengaruhi nilai *adjusted R-square*, kemiringan (*slope*) dan intersep, serta meningkatkan *mean square error* pada model regresi. Selain itu, model regresi linier klasik umumnya tidak menyediakan mekanisme khusus untuk mendeteksi dan menangani inkonsistensi sampel, sehingga hasil analisis bisa menjadi bias atau kurang dapat diandalkan.

RST yang diperkenalkan oleh Pawlak, menawarkan kerangka matematis untuk menangani data yang mengandung ketidakpastian dan inkonsistensi tanpa memerlukan asumsi tambahan seperti distribusi probabilitas. RST dapat digunakan untuk melakukan *data reduction*, yaitu mengidentifikasi dan mengeliminasi sampel yang tidak konsisten serta atribut yang tidak relevan, sehingga menghasilkan dataset yang lebih bersih dan konsisten. Pendekatan ini terbukti dapat meningkatkan performa model regresi linier, baik dari segi nilai *adjusted R-square*, kemiringan dan intersep, maupun menurunkan *mean square error* [7][8].

Dengan demikian, integrasi antara model regresi linier dan pendekatan *Rough Set* diterapkan untuk meningkatkan kualitas model regresi linier pada dataset *Student Performance Score*. Tahap pertama dilakukan dengan reduksi inkonsistensi sampel menggunakan RST untuk memastikan konsistensi dan keandalan data sebelum proses pemodelan. Selanjutnya, tahap kedua berupa pembangunan model regresi linier berganda menggunakan data hasil reduksi untuk menganalisis pengaruh variabel-variabel terhadap *Student Performance Score*. Melalui integrasi kedua tahap ini, model regresi yang dihasilkan terbukti lebih akurat, stabil, dan tahan terhadap noise, serta mampu memberikan pemahaman yang lebih komprehensif mengenai hubungan antara faktor akademik, perilaku belajar, dan lingkungan terhadap variasi *Student Performance Score*.

## 2. Metode Penelitian

### Jenis Penelitian

Penelitian ini merupakan studi kuantitatif dengan pendekatan perbandingan model. Tujuan penelitian adalah mengevaluasi pengaruh penanganan inkonsistensi sampel serta identifikasi outlier terhadap kualitas model regresi linier berganda. Pemodelan dilakukan menggunakan analisis regresi linier berganda pada dua kondisi dataset, yaitu: (1) data asli tanpa perlakuan dan (2) data yang telah direduksi inkonsistensinya menggunakan Rough Set Theory (RST). Setiap model kemudian dievaluasi melalui serangkaian uji asumsi klasik dan dibandingkan berdasarkan nilai koefisien regresi, signifikansi parameter, nilai $R^2$, serta pola residual untuk menilai stabilitas struktur model dan akurasi prediksi.

.

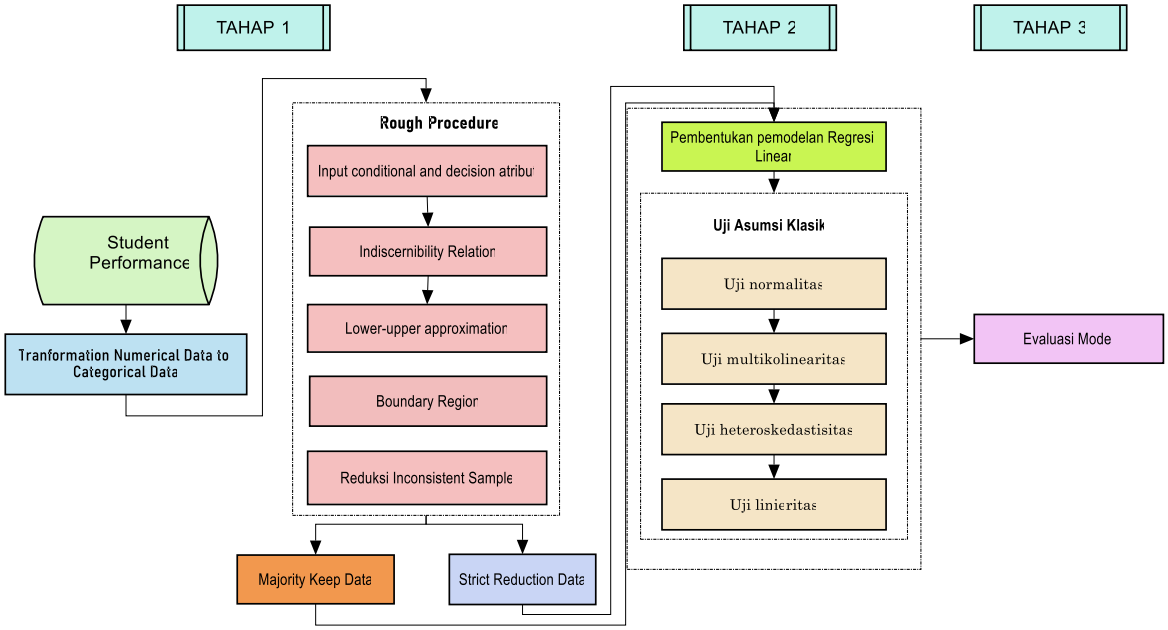
### Jenis dan Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder diperoleh dari dataset *students performance* yang tersedia pada platform Kaggle (https://www.kaggle.com). Dataset ini memuat informasi mengenai berbagai faktor yang memengaruhi kinerja belajar siswa, dengan lima variabel utama yang digunakan dalam penelitian ini, yaitu *Sleep Hours, Stress Level, Previous Exam Scores, Study Hours*, sebagai variabel independen, dan *Performance Score* sebagai variabel dependen. Dalam teori rough set, variabel independen disebut sebagai *condition variables*, sedangkan variabel dependen disebut sebagai *decision variable*.

http://dx.doi.org/10.30983/10376

## Prosedur Penelitian

Prosedur penelitian ini terdiri dari tiga tahap utama, yaitu: (1) Tahap *pre-processing* data, (2) Tahap Uji Asumsi Klasik dan Pembentukan Model Regresi, dan (3) Tahap evaluasi kinerja model.




**Gambar 1.** Prosedur Penelitian

**Tahap 1**. Data *Pre-Processing*

Tahap pre-processing dilakukan untuk menyiapkan data sebelum proses pemodelan, termasuk penyesuaian format variabel agar sesuai dengan kebutuhan algoritma RST, yang bekerja pada representasi data kategorikal. Oleh karena itu, variabel numerik dikonversi menjadi variabel kategorik melalui proses kategorisasi berbasis distribusi data.

Step 1. Transformasi data numerik ke data kategorik

Kriteria pengelompokan data ke dalam tiga kategori berdasarkan karakteristik distribusi, baik berdistribusi normal maupun tidak berdistribusi normal [9], disajikan pada Tabel 1.

**Tabel 1.** Kategorisasi untuk 3 kriteria

| Kategori | *Normality Attributes* | *Normality not Attributes* |
|---|---|---|
| *Low* | $X < \text{Mean} - SD$ | $X < Q_1$ |
| *Medium* | $\text{Mean} - SD < X < \text{Mean} + SD$ | $Q_1 \leq X \leq Q_3$ |
| *High* | $X > \text{Mean} + SD$ | $X > Q_3$ |

Step 2. Reduksi inkonsistensi sampel dengan menggunakan pendekatan RST

Pendekatan RST digunakan untuk mengidentifikasi dan mengeliminasi sampel-sampel yang tidak konsisten, yaitu sampel yang memiliki kombinasi atribut sama tetapi menghasilkan keputusan berbeda. Dengan mempertahankan hanya sampel yang memiliki kejelasan keputusan (*decision clarity*), kualitas basis data meningkat sehingga model regresi yang dibangun menjadi lebih stabil dan akurat[10]. Pendekatan ini sejalan dengan konsep dasar RST yang diperkenalkan [11] dan diperluas dalam oleh Komorowski et al [12]. Tahapan reduksi dengan RST dilakukan sebagai berikut:

http://dx.doi.org/10.30983/10376

a. Penyusunan Tabel Informasi (*Information System*)
b. Penentuan Relasi Ketakterbedaan (*Indiscernibility Relation*)
c. Penetapan *Lower Approximation, Upper Approximation*, dan *Boundary Region*
d. Reduksi Inkonsistensi Sampel
   Tahap ini merupakan inti dari proses RST. Dua pendekatan reduksi digunakan:
   1. *Majority-Keep*
      Sampel dalam boundary region dipertahankan hanya jika nilai keputusannya sesuai dengan nilai mayoritas dalam kelompok ketakterbedaan. Pendekatan ini menyeimbangkan jumlah data dan tingkat konsistensi [13].
   2. *Strict Reduction*
      Semua inkonsistensi sampel dihapus tanpa mempertimbangkan mayoritas. Hanya *lower approximation* yang dipertahankan.

   Tahap reduksi ini menghasilkan dua versi dataset yang kemudian digunakan untuk membangun model regresi baru yang lebih stabil.
e. Pembangkitan Aturan Keputusan (*Decision Rules*)

**Tahap 2**. Uji Asumsi dan Pembentukan Model Regresi

Pada tahap ini dilakukan serangkaian uji asumsi klasik guna memastikan bahwa data yang digunakan memenuhi prasyarat kelayakan untuk dianalisis lebih lanjut. Setelah seluruh asumsi terpenuhi, dilakukan analisis regresi linier berganda dengan tujuan untuk mengidentifikasi dan mengukur hubungan antara variabel independen terhadap variabel dependen.

a. Uji Asumsi Klasik
   Langkah-langkah yang perlu dilakukan dalam uji asumsi klasik meliputi uji linearitas, normalitas, multikolinearitas, dan heteroskedastisitas.
b. Regresi Linear berganda
   Bentuk umum persamaan linear regresi berganda sebagai berikut :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \quad (1)$$

Proses estimasi dilakukan dengan menggunakan metode *Ordinary Least Squares* (OLS), yaitu teknik yang meminimalkan jumlah kuadrat selisih antara nilai aktual dan nilai prediksi. Hasil estimasi parameter ini kemudian digunakan untuk membentuk model regresi yang mewakili hubungan linier antara variabel-variabel penelitian [14].

**Tahap 3.** Evaluasi kinerja model

Tahap ini bertujuan untuk menilai sejauh mana model regresi linier mampu menjelaskan variasi variabel dependen dan menghasilkan prediksi yang akurat, baik pada kondisi data asli maupun setelah perlakuan (reduksi inkonsistensi). Tahap evaluasi model bertujuan menilai dampak reduksi inkonsistensi terhadap kualitas regresi. Evaluasi dilakukan dengan mengamati tiga aspek utama [7], yaitu: (1) perubahan nilai *Adjusted R-Square* sebagai ukuran kemampuan penjelasan model, (2) perubahan dan stabilitas koefisien regresi (*intercept* dan *slope*), serta (3) peningkatan akurasi prediksi yang dinilai melalui *Mean Square Error* (MSE).

## 3. Hasil dan Pembahasan

Bagian ini membahas hasil analisis regresi linier berganda pada dataset sebelum dan sesudah penerapan *Rough Set Theory* (RST). Pembahasan difokuskan pada pengaruh reduksi sampel inkonsisten melalui strategi *majority-keep* dan *strict-reduction* terhadap struktur data, kestabilan

http://dx.doi.org/10.30983/10376

koefisien regresi, dan akurasi model. Selanjutnya, dihitung nilai rata-rata (*mean*) dan simpangan baku (*standard deviation*) untuk variabel *Sleep Hours, Stress Level, Previous Exam Scores, Study Hours,* dan *Performance Score* pada dataset *student performance* dari Kaggle (https://www.kaggle.com). Nilai tersebut digunakan sebagai dasar kategorisasi data ke dalam tiga kelas, yaitu Low, Medium, dan High. Tabel 2 menyajikan nilai *mean* dan simpangan baku masing-masing variabel.

**Tabel 2.** Nilai Rata-rata (Mean) dan Simpangan Baku (SD) untuk Kategorisasi Data

| Mean dan SD | *Sleep Hours* | *Stress Level* | *Previous Exam Scores* | *Study Hours* | *Performance Score* |
|---|---|---|---|---|---|
| Mean | 7,0799 | 5,07584 | 63,32518 | 4,03618 | 30,96612 |
| SD | 1,4795 | 1,94709 | 14,54454 | 1,99945 | 8,34671 |
| Mean - SD | 5,6004 | 3,12875 | 48,78064 | 2,03673 | 22,61941 |
| Mean + SD | 8,5594 | 7,02293 | 77,86973 | 6,03563 | 39,31283 |

Berdasarkan Tabel 2, nilai numerik data *student performance* ditransformasi menjadi nilai kategorik yang disajikan pada Tabel 3.

**Tabel 3.** Data kategorik student performance

| *Observasi* | *Sleep Hours* | *Stress Level* | *Previous Exam Scores* | *Study Hours* | *Performance Score* |
|---|---|---|---|---|---|
| O1 | *Low* | *High* | *High* | *High* | *High* |
| ⋮ | | | | | |
| O34 | *Low* | *High* | *High* | *High* | *Medium* |
| ⋮ | | | | | |
| O323 | *High* | *High* | *Low* | *Medium* | *Medium* |

Data kategorik pada Tabel 3 selanjutnya digunakan untuk mengidentifikasi kelas ekivalensi dan sampel inkonsisten dalam proses analisis RST.

**Reduksi inkonsistensi sampel dengan menggunakan pendekatan RST**

Setelah proses transformasi nilai numerik menjadi nilai kategorik sebagaimana disajikan pada Tabel 4, data selanjutnya disusun dalam bentuk *decision table* yang terdiri atas atribut kondisi (*condition attributes*) dan atribut keputusan (*decision attribute*). Dalam penelitian ini, atribut kondisi meliputi *Sleep Hours*, *Stress Level*, *Previous Exam Scores*, dan *Study Hours*, sedangkan *Performance Score* digunakan sebagai atribut keputusan.

Pada tahap ini dilakukan identifikasi terhadap inkonsistensi data. Inkonsistensi terjadi ketika terdapat lebih dari satu observasi yang memiliki kombinasi nilai atribut kondisi yang sama, namun menghasilkan nilai atribut keputusan yang berbeda. Kondisi tersebut menunjukkan adanya ketidakpastian dalam hubungan antara atribut kondisi dan keputusan, yang berpotensi menurunkan kualitas pemodelan regresi apabila tidak ditangani.

Oleh karena itu, dilakukan proses reduksi inkonsistensi sampel menggunakan pendekatan *Rough Set Theory* (RST). Dalam penelitian ini digunakan dua strategi reduksi, yaitu *majority-keep reduction* dan *strict reduction* [15]. Pada pendekatan *majority-keep*, apabila dalam satu kelompok sampel dengan atribut kondisi identik ditemukan lebih dari satu nilai keputusan, maka hanya keputusan yang paling dominan (mayoritas) yang dipertahankan. Sampel dengan keputusan minoritas dikategorikan sebagai inkonsisten dan dieliminasi. Pendekatan ini bertujuan untuk meningkatkan kejelasan keputusan dengan tetap mempertahankan ukuran sampel yang relatif besar. Dari total 322 data awal, jumlah data yang tersisa setelah penerapan *majority-keep reduction* adalah 250 data.

Sementara itu, pendekatan *strict reduction* mengeliminasi seluruh sampel dalam kelompok yang mengandung nilai keputusan berbeda. Dengan kata lain, apabila terdapat ketidakkonsistenan pada suatu

http://dx.doi.org/10.30983/10376

kombinasi atribut kondisi, maka seluruh sampel dalam kelompok tersebut dikeluarkan dari dataset. Strategi ini menghasilkan dataset dengan tingkat konsistensi paling tinggi, namun dengan ukuran sampel yang jauh lebih kecil. Setelah penerapan *strict reduction*, jumlah data berkurang menjadi 60 data dari total awal 322. Dataset hasil reduksi dari kedua pendekatan tersebut selanjutnya digunakan sebagai dasar dalam pembentukan model regresi linier dan evaluasi kinerja model.

**Uji Asumsi Klasik dan Pembentukan Model Regresi**

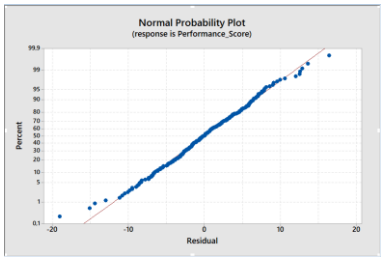
Regresi linier berganda mensyaratkan terpenuhinya sejumlah asumsi klasik agar koefisien yang diperoleh bersifat tidak bias, efisien, dan konsisten. Karena itu, sebelum melakukan interpretasi koefisien, uji asumsi klasik. Uji ini meliputi normalitas residual, multikolinearitas, heteroskedastisitas, dan linearitas, yang seluruhnya diperlukan agar parameter regresi dapat diinterpretasikan secara akurat.

a. **Regresi linier berganda sebelum reduksi data**

Pada tahap awal, dilakukan uji asumsi menggunakan dataset asli sebelum prosedur reduksi data dengan pendekatan RST. Estimasi ini bertujuan untuk memperoleh nilai konstanta (intercept) dan koefisien regresi masing-masing variabel independen terhadap variabel dependen.

Sebelum interpretasi hasil dilakukan, terlebih dahulu diuji asumsi klasik untuk memastikan validitas model. Uji asumsi yang diterapkan meliputi:

1. Uji normalitas




**Gambar 1.** Plot probabilitas normal residual sebelum reduksi

Berdasarkan Plot probabilitas normal residual pada Gambar 1, titik-titik residual mengikuti garis diagonal secara relatif konsisten, menunjukkan bahwa distribusi residual mendekati normal. Meskipun terdapat sedikit penyimpangan pada bagian ekor, deviasi tersebut masih dalam batas toleransi dan tidak menunjukkan pola sistematis yang signifikan, dengan demikian asumsi normalitas terpenuhi[16].

**Tabel 4.** Hasil Uji normalitas residual regresi sebelum reduksi

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| Unstandardized Residual | .035 | 322 | .200* | .995 | 322 | .412 |

*. This is a lower bound of the true significance.

Berdasarkan hasil uji normalitas terhadap unstandardized residual, diperoleh nilai signifikansi uji Kolmogorov–Smirnov sebesar 0,200 dan uji Shapiro–Wilk sebesar 0,412. Kedua nilai signifikansi tersebut lebih besar dari 0,05, sehingga $H_0$ tidak ditolak. Dengan demikian, residual regresi dapat dinyatakan berdistribusi normal.

http://dx.doi.org/10.30983/10376

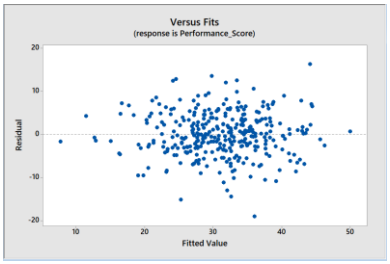
2. Uji multikolinearitas

**Tabel 5.** Uji multikolinearitas sebelum reduksi

| *Term* | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| *Constant* | 6,17 | 2,33 | 2,65 | 0,009 | |
| *Sleep_Hours* | 1,257 | 0,197 | 6,37 | 0,000 | 1,03 |
| *Stress_Level* | -1,642 | 0,150 | -10,97 | 0,000 | 1,03 |
| *Previous_Exam_Scores* | 0,2815 | 0,0199 | 14,14 | 0,000 | 1,02 |
| *Study_Hours* | 1,587 | 0,144 | 11,00 | 0,000 | 1,01 |

Nilai *Variance Inflation Factor* (VIF) pada Tabel 5 berada pada kisaran 1,01–1,03, jauh di bawah ambang batas multikolinearitas (VIF > 10). Hal ini menunjukkan tidak adanya tidak ada indikasi multikolinearitas [17]. Kondisi ini mengindikasikan bahwa keempat variabel bebas tidak saling berkorelasi satu sama lain.

3. Uji heteroskedastisitas




**Gambar 2.** Plot residual terhadap nilai prediksi data sebelum reduksi

Gambar 2 menunjukkan plot residual terhadap nilai prediksi. Titik-titik residual tampak tersebar secara acak di sekitar garis horizontal nol tanpa membentuk pola sistematis. Penyebaran residual yang homogen menunjukkan bahwa asumsi homoskedastisitas terpenuhi [18].

4. Uji linieritas

Berdasarkan Gambar 2 yang menampilkan Plot residual terhadap nilai prediksi, terlihat bahwa titik-titik residual tersebar secara acak di sekitar garis horizontal nol tanpa menunjukkan pola tertentu seperti lengkungan atau arah kecenderungan tertentu. Kondisi tersebut mengindikasikan bahwa hubungan antara variabel independen dan variabel dependen dalam model bersifat linier.

**Tabel 6.** Uji Linieritas sebelum reduksi

| Variabel Independen | Sig. Linearity | Sig. Deviation from Linearity | Kesimpulan |
|---|---|---|---|
| Sleep Hours | 0.000 | 0.688 | Linier |
| Stress Level | 0.000 | 0.555 | Linier |
| Previous Exam Scores | 0.000 | 0.632 | Linier |
| Study Hours | 0.000 | 0.992 | Linier |

Berdasarkan hasil uji linearitas sebelum reduksi data sebagaimana disajikan pada Tabel 6, seluruh variabel independen menunjukkan nilai signifikansi pada baris *Linearity* lebih kecil dari 0,05 dan nilai signifikansi pada baris *Deviation from Linearity* lebih besar dari 0,05. Hasil

http://dx.doi.org/10.30983/10376

ini mengindikasikan bahwa hubungan antara masing-masing variabel independent bersifat linear. Dengan demikian, dapat disimpulkan bahwa asumsi linearitas pada model regresi sebelum reduksi data telah terpenuhi, sehingga model regresi linier layak digunakan pada tahap analisis selanjutnya.

Model regresi linier dibuat dengan menggunakan dataset asli yang terdiri dari 322 observasi secara matematis, model regresi dapat dituliskan sebagai berikut:

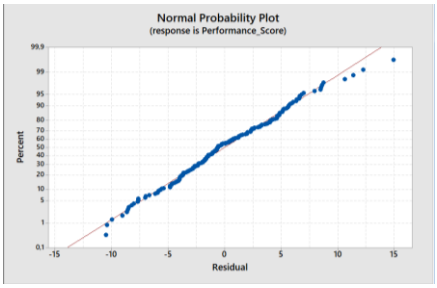
$$\textit{Performance Score} = 6{,}17 + 1{,}257\ \textit{Sleep Hours} - 1{,}642\ \textit{Stress Level} + 0{,}2815\ \textit{Previous Exam Scores} + 1{,}587\ \textit{Study Hours} \quad (2)$$

Persamaan tersebut menunjukkan bahwa variabel Sleep Hours, Previous Exam Scores, dan Study Hours berpengaruh positif terhadap *Performance Score*, sedangkan *Stress Level* berpengaruh negatif. Hal ini berarti bahwa semakin tinggi waktu tidur, nilai ujian sebelumnya, dan waktu belajar, maka semakin tinggi pula Performance Score yang dicapai siswa. Sebaliknya, semakin tinggi tingkat stres, semakin rendah performa akademik yang dihasilkan

**b. Regresi Linier Berganda setelah data reduksi dengan menggunakan pendekatan RST *majority-Keep***

Setelah penerapan RST diperoleh subset data yang merepresentasikan *lower approximation*, yaitu sampel-sampel dengan pola keputusan yang sepenuhnya konsisten. Pada tahap ini digunakan prinsip *majority-keep*, yaitu mempertahankan nilai keputusan yang dominan pada kelompok inkonsisten. Model regresi linier berganda kemudian diestimasi ulang menggunakan data yang sudah direduksi untuk menilai dampak peningkatan konsistensi data terhadap akurasi prediksi model. Seluruh uji asumsi klasik kembali diterapkan dengan prosedur yang sama dalam pemenuhan asumsi statistik setelah reduksi data.

1. Uji normalitas




**Gambar 3.** Plot probabilitas normal residual setelah reduksi *majority-keep*

Sebagaimana ditunjukkan pada Gambar 3, titik-titik residual secara relatif konsisten mengikuti garis diagonal merah, dengan penyimpangan yang kecil pada bagian ekor artinya, setelah data direduksi sebaran error model menjadi lebih simetris. Pola ini mengindikasikan bahwa distribusi residual mendekati normal.

**Tabel 7.** Uji normalitas setelah reduksi *majority-keep*

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| Unstandardized Residual | .028 | 250 | .200* | .991 | 250 | .108 |

*. This is a lower bound of the true significance.

http://dx.doi.org/10.30983/10376

Berdasarkan hasil uji normalitas terhadap unstandardized residual, diperoleh nilai signifikansi uji Kolmogorov–Smirnov sebesar 0,200 dan uji Shapiro–Wilk sebesar 0,108. Kedua nilai signifikansi tersebut lebih besar dari 0,05, (p > 0,05) yang mengindikasikan bahwa asumsi normalitas terpenuhi. Dengan demikian, residual regresi dapat dinyatakan berdistribusi normal.

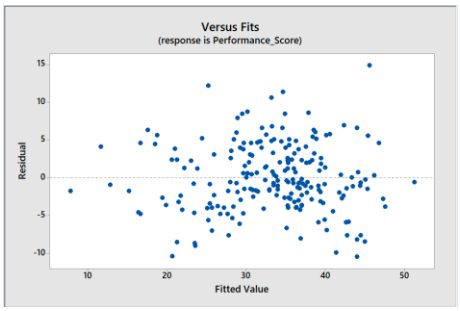
2. Uji multikolinearitas

**Tabel 8.** Uji multikolinearitas setelah reduksi *majority-keep*

| *Term* | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| *Constant* | 3,32 | 2,58 | 1,29 | 0,200 | |
| *Sleep_Hours* | 1,521 | 0,218 | 6,97 | 0,000 | 1,03 |
| *Stress_Level* | -1,574 | 0,164 | -9,60 | 0,000 | 1,07 |
| *Previous_Exam_Scores* | 0,3044 | 0,0226 | 13,45 | 0,000 | 1,04 |
| *Study_Hours* | 1,633 | 0,153 | 10,69 | 0,000 | 1,06 |

Nilai *Variance Inflation Factor* (VIF) pada hasil regresi menunjukkan bahwa seluruh variabel independen memiliki nilai yang sangat rendah, yaitu antara 1,03 hingga 1,07. Nilai VIFnya kurang dari 10, hal ini menunjukkan bahwa tidak ada indikasi multikolinearitas antarvariabel independen. Kondisi ini mengindikasikan bahwa keempat variabel bebas tidak saling berkorelasi satu sama lain.

3. Uji heteroskedastisitas




**Gambar 4.** Plot residual terhadap nilai prediksi data setelah reduksi

Sebagaimana ditunjukkan pada *Residual versus Fitted Value Plot*, sebaran titik residual tampak acak di sekitar garis horizontal nol tanpa membentuk pola, sehingga asumsi homoskedastisitas terpenuhi.

4. Uji linieritas

Sebaran titik pada Gambar 4 juga menunjukkan bahwa tidak terdapat pola lengkung, gelombang, atau kecenderungan arah tertentu, melainkan tersebar secara acak di sekitar garis horizontal nol. Pola acak tersebut mengindikasikan bahwa hubungan antara variabel independen dan variabel dependen dalam model bersifat linier.

**Tabel 9.** Uji linieritas setelah reduksi *majority-keep*

| Variabel Independen | Sig. Linearity | Sig. Deviation from Linearity | Kesimpulan |
|---|---|---|---|
| Sleep Hours | 0.001 | 0.111 | Linear |
| Stress Level | 0.000 | 0.562 | Linear |

http://dx.doi.org/10.30983/10376

| | | | |
|---|---|---|---|
| Previous Exam Scores | 0.000 | 0.808 | Linear |
| Study Hours | 0.000 | 0.989 | Linear |

Hasil uji linearitas menunjukkan bahwa seluruh variabel independen, yaitu Sleep Hours, Stress Level, Previous Exam Scores, dan Study Hours, memiliki hubungan linear yang signifikan dengan Performance Score (Sig. Linearity < 0,05) serta tidak menunjukkan penyimpangan dari linearitas (Sig. Deviation from Linearity > 0,05). Dengan demikian, asumsi linearitas dalam model regresi terpenuhi.

Setelah menerapkan strategi *majority-keep*, diperoleh model regresi linear berganda dengan persamaan berikut:

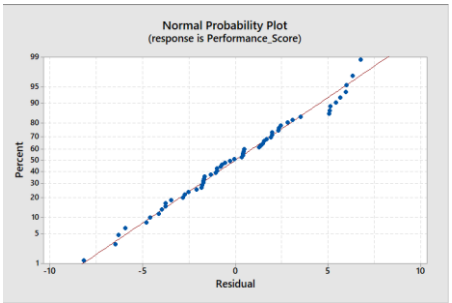
$$Performance\_Score = 3{,}32 + 1{,}521\ Sleep\ Hours - 1{,}574\ Stress\_Level + 0{,}3044\ Previous\ Exam\ Scores + 1{,}633\ Study\ Hours \quad (3)$$

Persamaan tersebut menunjukkan bahwa variabel *Sleep Hours, Previous Exam Scores*, dan *Study Hours* berpengaruh positif terhadap *Performance Score*, sedangkan *Stress Level* berpengaruh negatif. Hal ini berarti bahwa semakin tinggi waktu tidur, nilai ujian sebelumnya, dan waktu belajar, maka semakin tinggi pula *Performance Score* yang dicapai siswa. Sebaliknya, semakin tinggi tingkat stres, semakin rendah performa akademik yang dihasilkan.

**c. Regresi linier berganda sesudah data reduksi dengan menggunakan pendekatan RST *Strict Reduction***

Setelah penerapan pendekatan RST *strict reduction*, diperoleh subset data yang hanya berisi sampel-sampel dengan pola keputusan yang sepenuhnya konsisten. Model regresi linier berganda kemudian diestimasi ulang menggunakan data yang sudah direduksi dengan RST *strict reduction*. Seluruh prosedur uji asumsi klasik diterapkan kembali dengan metode yang sama seperti pada dua model sebelumnya. Tujuannya adalah membandingkan perubahan karakteristik residual, tingkat homoskedastisitas, distribusi normalitas, serta linearitas.

1. Uji normalitas




**Gambar 5.** Plot probabilitas normal residual setelah reduksi RST Strict Reduction

Setelah penerapan RST strict reduction, Normal Probability Plot pada Gambar 5, residual cenderung mengikuti garis diagonal, yang menunjukkan bahwa residual terdistribusi mendekati normal. Meskipun terdapat sedikit penyimpangan pada bagian ekor, penyimpangan tersebut tidak signifikan sehingga asumsi normalitas residual dapat dinyatakan terpenuhi.

http://dx.doi.org/10.30983/10376

**Tabel 10.** Uji normalitas setelah reduksi *strict-reduction*

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| Unstandardized Residual | .112 | 60 | .058 | .959 | 60 | .043 |

*. This is a lower bound of the true significance.

Hasil uji Kolmogorov–Smirnov dan Shapiro–Wilk menunjukkan bahwa residual mendekati normal.

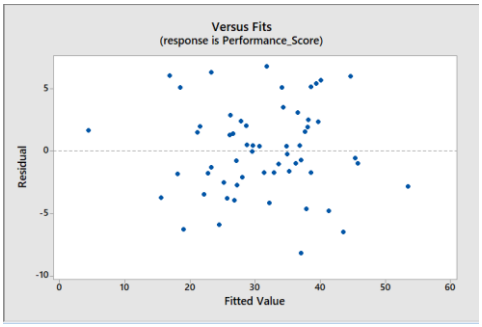
2. Uji multikolinearitas

**Tabel 11.** Uji multikolinearitas setelah RST *Strict Reduction*

| *Term* | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| *Constant* | 6,30 | 5,29 | 1,19 | 0,239 | |
| *Sleep_Hours* | 0,948 | 0,466 | 2,04 | 0,047 | 1,25 |
| *Stress_Level* | -1,978 | 0,313 | -6,32 | 0,000 | 1,50 |
| *Previous_Exam_Scores* | 0,3251 | 0,0458 | 7,10 | 0,000 | 1,23 |
| *Study_Hours* | 1,811 | 0,297 | 6,10 | 0,000 | 1,57 |

Nilai *Variance Inflation Factor* (VIF) pada hasil regresi menunjukkan bahwa seluruh variabel independen memiliki nilai yang sangat rendah, yaitu antara 1,25 hingga 1,57. Nilai VIFnya kurang dari 10, hal ini menunjukkan bahwa tidak ada indikasi multikolinearitas antarvariabel independen. Kondisi ini mengindikasikan bahwa keempat variabel bebas tidak saling berkorelasi satu sama lain.

3. Uji heteroskedastisitas




**Gambar 6.** Plot residual terhadap nilai prediksi data setelah RST *Strict Reduction*

Plot residual terhadap nilai *fitted* pada model setelah penerapan RST *Strict Reduction* menunjukkan pola penyebaran residual yang acak dan tidak membentuk struktur tertentu di sepanjang sumbu nilai prediksi. Titik-titik residual tersebar baik di atas maupun di bawah garis nol tanpa membentuk pola lengkung, gelombang, atau pola sistematis lainnya. Penyebaran acak ini mengindikasikan bahwa asumsi homoskedastisitas terpenuhi secara lebih baik.

4. Uji Linieritas

Hasil uji linearitas melalui plot *Residual versus Fitted* pada model setelah penerapan RST *Strict Reduction* menunjukkan bahwa hubungan antara variabel prediktor dan *Performance Score* dapat dikatakan linear. Penyebaran residual tampak acak dan tersebar di sekitar garis nol tanpa menunjukkan kecenderungan bergerak pada arah tertentu mengikuti nilai *fitted*.

http://dx.doi.org/10.30983/10376

Eliminasi penuh terhadap inkonsistensi sampel melalui *strict reduction* turut memperkuat linearitas ini, karena data yang tersisa memiliki pola hubungan yang lebih konsisten. Dengan demikian, model *strict reduction* menunjukkan pemenuhan asumsi linearitas yang baik dan dapat dipercaya dalam menggambarkan hubungan linear antara variabel independen dan *Performance Score*.

**Tabel 12.** Uji linieritas setelah RST *Strict Reduction*

| Variabel Independen | Sig. Linearity | Sig. Deviation from Linearity | Kesimpulan |
|---|---|---|---|
| Sleep Hours | 0.053 | 0.240 | Cenderung linear (marginal) |
| Stress Level | 0.000 | 0.289 | Linear |
| Previous Exam Scores | 0.076 | 0.609 | Cenderung linear (marginal) |
| Study Hours | 0.027 | 0.918 | Linear |

Hasil uji linearitas setelah penerapan skema *strict reduction* menunjukkan bahwa variabel *Stress Level* dan *Study Hours* memiliki hubungan linear yang signifikan dengan Performance Score (Sig. Linearity < 0,05) serta tidak menunjukkan penyimpangan dari linearitas (Sig. Deviation from Linearity > 0,05). Sementara itu, variabel *Sleep Hours* dan *Previous Exam Scores* menunjukkan kecenderungan hubungan linear (Sig. Linearity sedikit di atas 0,05), namun tetap tidak menunjukkan penyimpangan dari linearitas. Secara umum, asumsi linearitas masih dapat diterima.

Setelah menerapkan RST *strict reduction*, diperoleh model regresi linear berganda dengan persamaan berikut:

$$\textit{Performance Score} = 6{,}30 + 0{,}948\ \textit{Sleep Hours} - 1{,}978\ \textit{Stress Level} + 0{,}3251\ \textit{Previous Exam Scores} + 1{,}811\ \textit{Study Hours} \quad (4)$$

## Evaluasi Model

Untuk menilai dampak reduksi inkonsistensi terhadap kualitas pemodelan, Tabel 12 menyajikan perbandingan komprehensif kinerja tiga model regresi pada kondisi data yang berbeda model awal tanpa reduksi, model dengan reduksi RST menggunakan pendekatan *majority-keep*, dan model dengan *strict reduction*. Perbandingan ini mencakup ukuran kecocokan model ($R^2$, *Adjusted* $R^2$, *Predicted* $R^2$), akurasi prediksi (MSE dan *Standar Error*) dan perubahan koefisien regresi. Penyajian ini memungkinkan evaluasi menyeluruh mengenai bagaimana peningkatan konsistensi data melalui RST memengaruhi performa dan reliabilitas model regresi linier berganda.

**Tabel 13.** Perbandingan kinerja tiga model regresi sebelum dan sesudah reduksi RST

| Komponen | Sebelum RST | RST *Majority-Keep* | RST *Strict Reduction* |
|---|---|---|---|
| *Number of Data* | 322 | 209 | 58 |
| *R-squared* ($R^2$) | 62.41% | 74.10% | 86.16% |
| *Adjusted R-squared* | 61.94% | 73.59% | 85.12% |
| *Predicted R-squared* | 61.22% | 72.79% | 83.10% |
| *Standard Error (S)* | 5.14935 | 4.51864 | 3.68176 |
| *MSE* | 26.52 | 20.42 | 13.56 |
| *Intercept* | 6.17 | 3.32 | 6.30 |
| *Sleep Hours (β1)* | +1.257 | +1.521 | +0.948 |
| *Stress Level (β2)* | -1.642 | -1.574 | -1.978 |
| *Previous Exam Scores (β3)* | +0.2815 | +0.3044 | +0.3251 |
| *Study Hours (β4)* | +1.587 | +1.633 | +1.811 |

http://dx.doi.org/10.30983/10376

1. Evaluasi *Adjusted R-Square (Adjusted $R^2$)*
   Nilai *Adjusted R-Square* mengalami peningkatan yang signifikan pada kedua model setelah penerapan *Rough Set Theory* (RST). Pada model awal tanpa reduksi, nilai *Adjusted $R^2$* sebesar 61,94%, yang menunjukkan bahwa model hanya mampu menjelaskan sekitar 62% variasi *Performance Score*. Setelah dilakukan reduksi dengan metode *majority-keep*, nilai *Adjusted $R^2$* meningkat menjadi 73,59%, menandakan bahwa inkonsistensi yang dihilangkan pada tahap ini memberikan pengaruh positif terhadap kemampuan model dalam menjelaskan variabel dependen. Peningkatan paling besar terlihat pada model *strict reduction* dengan nilai *Adjusted $R^2$* mencapai 85,12%, menunjukkan bahwa model ini memiliki kemampuan penjelasan terbaik dibandingkan model lainnya. Hasil ini konsisten dengan teori bahwa penghapusan sampel data yang inkonsisten dapat memperbaiki struktur relasi linear sehingga model menjadi lebih stabil dan informatif [19].
2. Perbandingan Koefisien Regresi (*Slope & Intercept*)
   Koefisien regresi menunjukkan pola perubahan yang menarik setelah diterapkan metode RST. Pada model awal, intercept dan slope beberapa variabel menunjukkan nilai yang kurang stabil dan cenderung dipengaruhi oleh keberadaan data tidak konsisten. Setelah penerapan *majority-keep*, terjadi perubahan pada *intercept* (dari 6,17 menjadi 3,32) dan peningkatan pada *slope* variabel *Sleep Hours* dan *Study Hours*, menandakan bahwa model mulai mengoreksi ketidakteraturan data. Sementara itu, *strict reduction* menghasilkan koefisien yang paling stabil dan konsisten secara teoretis. Misalnya, pengaruh negatif *Stress Level* menjadi lebih kuat (−1,978), pengaruh *Previous Exam Scores* semakin meningkat (+0,3251), dan *Study Hours* menjadi lebih dominan (+1,811). Stabilitas koefisien meningkat dari kategori rendah pada model awal, sedang pada *majority-keep*, hingga tinggi pada *strict reduction*.
3. Evaluasi Akurasi Model (MSE)
   Nilai MSE menunjukkan penurunan yang jelas pada kedua model hasil reduksi RST, yang menandakan adanya peningkatan akurasi prediksi. Pada model awal, nilai MSE sebesar 26,52, menunjukkan tingkat error yang relatif tinggi dalam memprediksi *Performance Score*. Setelah penerapan *majority-keep*, MSE turun menjadi 20,42, yang berarti prediksi model menjadi lebih baik dengan berkurangnya inkonsistensi sampel. Penurunan terbesar terjadi pada *strict reduction* dengan nilai MSE mencapai 13,56, menjadikannya model dengan akurasi terbaik di antara ketiga model.

Berdasarkan hasil analisis, pendekatan *majority-keep* dan *strict reduction* menghasilkan karakteristik model yang berbeda namun saling melengkapi. Pendekatan *majority-keep* mempertahankan jumlah data yang lebih besar sehingga menghasilkan estimasi yang lebih stabil dan representatif terhadap pola umum data, dengan pemenuhan asumsi statistik yang konsisten. Sebaliknya, *strict reduction* menekankan kualitas data melalui seleksi yang lebih ketat terhadap observasi inkonsisten, yang berdampak pada peningkatan kecocokan model pada data yang dianalisis serta perbaikan model regresi. Meskipun ukuran sampel pada *strict reduction* lebih terbatas, hasil pengujian menunjukkan bahwa asumsi statistik utama tetap terpenuhi pada kedua pendekatan, sehingga hasil analisis masih layak untuk diinterpretasikan, dengan tetap mempertimbangkan trade-off antara konsistensi data dan representativitas sampel [7][15].

## 4. Kesimpulan

Penelitian ini menunjukkan bahwa pendekatan *majority-keep* dan *strict reduction* menghasilkan karakteristik model regresi linier yang berbeda namun saling melengkapi. Pendekatan *majority-keep* mempertahankan ukuran data yang lebih besar sehingga memberikan

http://dx.doi.org/10.30983/10376

estimasi yang stabil dan representatif terhadap pola umum data, sementara *strict reduction* meningkatkan konsistensi data dan kualitas model regresi melalui seleksi sampel yang lebih ketat. Meskipun ukuran sampel pada *strict reduction* lebih terbatas, pemenuhan asumsi statistik utama tetap terjaga pada kedua pendekatan. Oleh karena itu, pemilihan strategi reduksi perlu mempertimbangkan trade-off antara konsistensi data dan representativitas sampel sesuai dengan tujuan analisis.

## Daftar Pustaka


[1] D. H. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," vol. 01, no. 02, pp. 140–147, 2020, doi: 10.38094/jastt1457.

[2] K. Qu, "Research on linear regression algorithm," vol. 01046, 2024.

[3] M. A. Iqbal, "Aplication of Regression Techniques with their Advantages and Disadvantages." 2021.

[4] N. Roustaei, "Application and interpretation of linear-regression analysis," 2024.

[5] Y. Abdulraheem, "Journal of Public Health Issues and Practices The Role of Regression Analysis in Preventive Research Modalities : A Medical-Focused Comprehensive Review," vol. 9, pp. 1–5, 2025.

[6] S. Chung, Y. W. Park, and T. Cheong, "A mathematical programming approach for integrated multiple linear regression subset selection and validation," *Pattern Recognit.*, vol. 108, 2020, doi: 10.1016/j.patcog.2020.107565.

[7] Rasyidah, R. Efendi, N. M. Nawi, M. M. Deris, and S. M. A. Burney, "Cleansing of inconsistent sample in linear regression model based on rough sets theory," *Syst. Soft Comput.*, vol. 5, no. December 2022, 2023, doi: 10.1016/j.sasc.2022.200046.

[8] Z. Pawlak, "RoughSets," *Inst. Theor. Appl. Informatics, Polish Acad. Sci. ul. Bałtycka 5, 44 100 Gliwice, Pol.*, vol. 2, pp. 1–51, 1982.

[9] S.Santoso, *Statistik non-parametrik*. Elex Media Kompotindo, 2001.

[10] R. Słowiński, J. Stefanowski, S. Greco, and B. Matarazzo, "Rough set based processing of inconsistent information in decision analysis," *Control and Cybernetics*, vol. 29, no. 1. pp. 378–404, 2000.

[11] D. Dubois and H. Prade, "Rough fuzzy sets and fuzzy rough sets," *Int. J. Gen. Syst.*, vol. 17, no. 2–3, pp. 191–209, 1990, doi: 10.1080/03081079008935107.

[12] J. Komorowski, L. Polkowski, and A. Skowron, "Rough sets: A tutorial," *Rough fuzzy Hybrid. A new trend Decis.*, pp. 3–98, 1999, [Online]. Available: http://secs.ceas.uc.edu/~mazlack/dbm.w2011/Komorowski.RoughSets.tutor.pdf

[13] I. Rahmi, R. Efendi, and N. A. Samat, "Examining Risk Factors of Anemia in Pregnancy Using," *BAREKENG J. Math. App*, vol. 18, no. 1, pp. 537–552, 2024.

[14] X. Su, X. Yan, and C. L. Tsai, "Linear regression," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 4, no. 3, pp. 275–294, 2012, doi: 10.1002/wics.1198.

[15] I. Rahmi, R. Efendi, N. A. Samat, H. Yozza, and M. Wahyudi, "The Effects of Data Reduction Using Rough Set Theory on Logistic Regression Model," *Lecture Notes in Networks and Systems*, vol. 1078 LNNS. pp. 64–73, 2024. doi: 10.1007/978-3-031-66965-1_7.

[16] H. Kim and H. Kim, "Statistical notes for clinical researchers : simple linear regression 3 – residual analysis," vol. 44, no. 1, pp. 1–8, 2019.

[17] N. Shrestha, "Detecting Multicollinearity in Regression Analysis," no. June, pp. 1–5, 2020, doi: 10.12691/ajams-8-2-1.

[18] G. Vining, E. A. Peck, and D. Montgomery, *Introduction to Linear Regression Analysis*. 2012.

[19] R. Efendi, S. Mu'at, V. A. Dewi, N. Arisandy, N. A. Samsudin, and D. S. S. Sahid, "Rough-regression for categorical data prediction based on case study," *Proceeding - 2019 Int. Conf. Artif. Intell. Inf. Technol. ICAIIT 2019*, pp. 277–281, 2019, doi: 10.1109/ICAIIT.2019.8834584.


http://dx.doi.org/10.30983/10376