Maliki Interdisciplinary Journal (MIJ) eISSN: 3024-8140
Volume 3, Issue Mei, 2025 pp. 428-437 http://urj.uin-malang.ac.id/index.php/mij/index

Implementasi Metode Regresi dalam Pengolahan Data *Platfrom Rotten Tomatoes Movies*

Hanik Muafiyah[1], Riza Mar'atus Sholihah[2]
Program Studi Matematika, Universitas Islam Negeri Maulana Malik Ibrahim Malang
e-mail: 210601110098@student.uin-malang.ac.id

Kata Kunci:
Regresi Linier; Big Data; Data Mining, Rotten Tomatoes; Analisis Data Film.

Keywords:
Linear Regression; Big Data; Data Mining, Rotten Tomatoes; Movie Data Analysis

A B S T R A K

Penelitian ini bertujuan untuk mengimplementasikan metode regresi linier dalam menganalisis data dari platform Rotten Tomatoes Movies, yang menyediakan informasi film global mencakup rating kritikus, durasi, dan skor audiens. Metode regresi linier berganda diterapkan untuk menganalisis pengaruh variabel independen, yaitu durasi film dan rating kritikus (tomatoMeter), terhadap variabel dependen, yaitu skor penonton (audienceScore). Hasil analisis menunjukkan bahwa kedua variabel independen tersebut memiliki pengaruh signifikan terhadap skor audiens, dengan koefisien determinasi (R^2) sebesar 23,51%, yang berarti bahwa 23,51% variasi skor penonton dapat dijelaskan oleh durasi dan rating kritikus. Model regresi yang diperoleh adalah $Y = 12{,}900 + 0{,}48109X_1 + 0{,}09546X_2 + \varepsilon$ dengan Y (skor penonton), X_1 (durasi film), X_2 (rating kritikus). Meskipun demikian, nilai R^2 yang relatif rendah menunjukkan adanya variabel lain yang mungkin memengaruhi skor penonton. Studi ini mengilustrasikan potensi regresi linier dalam mengolah big data di industri film untuk keperluan prediksi, serta mengindikasikan pentingnya variabel tambahan guna meningkatkan akurasi model prediksi skor penonton.

A B S T R A C T

This research aims to implement the linear regression method in analyzing data from the Rotten Tomatoes Movies platform, which provides global film information including critic ratings, duration and audience scores. The multiple linear regression method is applied to analyze the influence of independent variables, namely film duration and critic rating (tomatoMeter), on the dependent variable, namely audience score (audienceScore). The results of the analysis show that these two independent variables have a significant influence on audience scores, with a coefficient of determination (R^2) of 23.51%, which means that 23.51% of the variation in audience scores can be explained by duration and critic ratings. The regression model obtained is $Y = 12{,}900 + 0{,}48109X_1 + 0{,}09546X_2 + \varepsilon$ with Y (audience score), X_1 (film duration), X_2 (critic rating). However, the relatively low R^2 value indicates the presence of other variables that may influence viewer scores. This study illustrates the potential of linear regression in processing big data in the film industry for prediction purposes, and indicates the importance of additional variables to improve the accuracy of audience score prediction models.

Pendahuluan

Big data adalah istilah yang digunakan untuk menggambarkan sumber data yang sangat besar, beragam, dan rumit yang sulit ditangani dengan menggunakan teknik pengolahan data konvensional. Salah satu teknik untuk menangani *big data* adalah penambangan data, yang mencari koneksi, pola, atau wawasan baru dalam sejumlah besar data yang rumit. Teknologi ini menyediakan alat untuk mengeksploitasi data

dalam basis data yang memiliki tingkat kompleksitas tinggi, sehingga memungkinkan pengolahan dan analisis data yang lebih efisien dan efektif (Lestari, 2019).

Salah satu metode *data mining* terbaik untuk membuat prediksi dalam pemrosesan data besar adalah regresi linier. Metode statistik yang digunakan untuk mensimulasikan hubungan antara satu atau lebih variabel independen dan variabel dependen adalah analisis regresi (Sholeh dkk., 2023). Regresi linier membantu memahami arah dan besarnya pengaruh variabel independen terhadap variabel dependen, di mana variabel dependen adalah variabel yang dipengaruhi, sedangkan variabel independen adalah variabel yang memberikan pengaruh. Film diproduksi dengan tujuan mencapai kesuksesan, seperti menembus pasar global, menghibur penonton, memperoleh rating tinggi, serta meraih keuntungan besar (Ilmi dkk., 2023). Industri film terus berkembang pesat berkat kemajuan teknologi informasi, yang memudahkan akses dan analisis data dalam skala besar. Salah satu *platform* populer yang menyediakan data terkait film adalah Rotten Tomatoes Movies. *Platform* ini menawarkan informasi tentang berbagai film dari seluruh dunia, termasuk para aktor, sutradara, penulis, penata rias, hingga musisi yang terlibat dalam pembuatan film tersebut.

Rotten Tomatoes Movies adalah salah satu sumber data terkaya di industri film. Data yang tersedia mencakup rating film, ulasan kritikus, komentar penonton, genre, dan tanggal rilis. Dengan jumlah data yang terus berkembang, *Rotten Tomatoes* menghasilkan *big data* yang memerlukan metode pemrosesan canggih untuk pengolahan datanya.Penelitian ini akan menggunakan metode regresi linier dengan studi kasus pengolahan big data dari *platform Rotten Tomatoes Movies*. Tujuan penelitian ini adalah untuk mengeksplorasi penerapan teknik *clustering* dalam pengolahan data *Rotten Tomatoes Movies*. Dengan menggunakan *clustering*, diharapkan dapat ditemukan wawasan yang lebih mendalam mengenai pola dan preferensi penilaian audiens, serta memberikan rekomendasi yang lebih akurat dan relevan.

Pembahasan

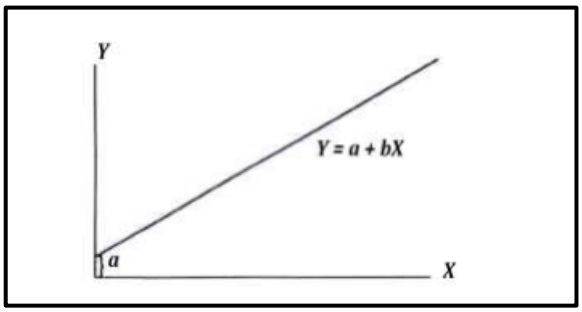
Menurut Sudjana (2005) dalam (Ningsih dkk., 2019) menjelaskan bahwa analisis regresi dalam statistika adalah salah satu metode yang digunakan untuk menentukan hubungan sebab-akibat antara satu variabel dengan variabel lainnya. Utama dan Hajarisman (2021) dalam (Baskoro dkk., 2022) menjelaskan bahwa metode regresi bertujuan untuk menghasilkan model prediksi yang baik, yang ditunjukkan dengan tingkat akurasi tinggi dan kesalahan yang rendah. Dalam analisis regresi, variabel yang saling berkaitan dikelompokkan menjadi dua jenis. Pertama, variabel penyebab atau prediktor yang disebut juga variabel independen, dilambangkan dengan "x" karena dalam grafik sering digambarkan pada sumbu x. Kedua, variabel akibat atau respons yang disebut variabel dependen, dilambangkan dengan "y." Kedua jenis variabel ini dapat berupa variabel acak, tetapi variabel yang dipengaruhi harus selalu merupakan variabel acak.

Jenis-Jenis Regresi

Regresi linier dibagi menjadi dua kelompok berdasarkan pada jumlah variabel prediktor. Pada dasarnya dua kelompok regresi linier tersebut sama-sama melibatkan variabel pemberi pengaruh (Arif dkk., 2023).

Regresi Linier Sederhana

Regresi linier sederhana merupakan suatu model analisis yang menggambarkan hubungan antara satu variabel penyebab (variaibel prediktor) yang dilambangkan dengan "x" dengan variabel akibatnya (variabel respon) yang dilambangkan dengan "y". Regresi linier sederhana sering disingkat dengan SLR (*simple linier regression*) yang digunakan dalam statistika untuk meramalkan atau prediksi tentang karakteristik kualitas dan kuantitas. Regresi linier sederhana biasanya digambarkan dalam garis lurus, seperti gambar 1.1



Gambar 1.1 Ilustrasi Garis Linear Sederhana

Secara matematik, persamaan regresi linier sederhana dapat diekspresikan oleh (Sudrajat & Tjuju, 2010):

$$\hat{Y} = a + bx$$

di mana:

$\hat{Y}$: garis regresi atau variabel respon,

a: konstanta (intersep), perpotongan dengan sumbu vertikal,

b: konstanta regresi (*slope*),

x: variabel penyebab atau variabel prediktor.

Sudjana (2005) dalam (Zuhri, 2020) menjelaskan bahwa untuk mencari nilai a dan b dapat menggunakan metode *Least Square*. Metode *Least Square* ialah sebagai berikut:

$$b = \frac{n \sum XY - \sum X . \sum Y}{n \sum X^2 - (\sum X)^2}$$

$$a = \frac{\sum Y - b \sum X}{n}$$

dengan:

Y : garis regresi atau variabel respon,

a : konstanta (intersep), perpotongan dengan sumbu vertikal,

b : *slope* atau kemiringan garis yaitu perubahan rata-rata pada y untuk setiap unit perubahan pada variabel x,

x : variabel penyebab atau variabel prediktor,

n : jumlah sampel.

Regresi Linier Berganda

Secara matematik, persamaan regresi linier berganda dapat diekspresikan oleh (Yamin, 2011):

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

dengan:

Y : variabel dependen (nilai yang diprediksikan),

X_1X_2 : variabel independen,

a : konstanta,

b : koefisien regresi (nilai peningkatan atau penurunan).

Pada penelitian ini terdapat dua variabel independen, yaitu *runtimeMinutes* (durasi film) dan *tomatoMeter* (rating kritikus), yang digunakan untuk memprediksi satu variabel dependen, yaitu *audienceScore* (skor penonton).

Regresi linier berganda digunakan ketika ada lebih dari satu variabel independen yang mempengaruhi variabel dependen, seperti yang terlihat pada model regresi berikut:

$$Y = 12{,}900 + 0{,}48109X_1 + 0{,}09546X_2 + \varepsilon$$

dengan:

Y : *audienceScore* (skor penonton),

X_1 : *runtimeMinutes* (durasi film),

X_2 : *tomatoMeter* (rating kritikus).

Artinya penilaian dan pemringkatan suatu film dalam situs Rotten Tomatoes oleh penonton secara konstan yaitu sebesar $12{,}900$. Selanjutnya, setiap peningkatan penilaian film dari kritikus film yang dinilai dalam variabel *tomatoMeter* akan meningkat sebesar $0{,}48109$. Setiap durasi meningkat maka terjadi peningkatan sebesar 1 maka penilaian dan pemringkatan suatu film akan meningkat sebesar $0{,}09546$.

Bentuk Uji Analisis Regresi

Penelitian ini menggunakan metode regresi linier untuk menganalisis hubungan antara variabel independen (durasi film dan *tomatoMeter*) terhadap variabel dependen (skor penonton). Berikut ini adalah uji-uji yang digunakan dalam analisis regresi untuk memastikan model yang akurat dan dapat diandalkan (Yamin, 2011).

Uji Simultan (Uji F)

Uji simultan ialah suatu uji yang digunakan untuk menunjukkan apakah semua variabel independen atau variabel prediktor secara bersama-sama berpengaruh signifikan terhadap variabel dependen atau variabel respon dengan menggunakan uji *Analysis of Variance* (ANOVA) (Ghozali, 2012).Untuk menguji hipotesis ini digunakan statistik F dengan kriteria pengambilan keputuasan. Dalam ANOVA jika $F_{hitung} > F_{tabel}$ atau $P_{value} < \alpha$, maka dapat dikatakan secara bersama-sama variabel independen berpengaruh signifikan terhadap variabel dependen.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	36460424	18230212	22010,51	0,000
tomatoMeter	1	31258440	31258440	37740,32	0,000
durasi	1	1886157	1886157	2277,28	0,000
Error	143255	118651017	828		
Lack-of-Fit	7100	7953010	1120	1,38	0,000
Pure Error	136155	110698007	813		
Total	143257	155111441			

Gambar 1.2 *Output Analysis Of Variance*

Dari output *Analysis of Variance* di atas dapat dilakukan uji kesesuaian model yang dilihat dari nilai *p-value* yaitu $0.000 < 0.05$, maka dapat disimpulkan bahwa seluruh variabel independen berpengaruh signifikan terhadap skor audiens

Uji Parsial (Uji T)

Uji parsial adalah uji yang digunakan untuk melihat sejauh mana setiap variabel independen secara individual mempengaruhi variabel dependen dalam penelitian (Ghozali, 2012). Uji parsial dikatakan berpengaruh signifikan apabila $T_{hitung} > T_{tabel}$ atau $P_{value} > \alpha$.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	12,900	0,184	70,07	0,000	
tomatoMeter	0,48109	0,00248	194,27	0,000	1,03
durasi	0,09546	0,00200	47,72	0,000	1,03

Gambar 1.3 *Output of Coefficients*

Dari output *Coefficient* di atas dapat dilakukan uji parameter model menggunakan uji t yaitu melihat matriks koefisien di atas. Nilai dari *p-value* suhu yaitu $0.000 < 0.05$ artinya variabel tomatoMeter berpengaruh signifikan terhadap penilaian penonton dalam suatu film. Sedangkan untuk *p-value* durasi film yaitu $0.000 < 0.05$ artinya

variabel durasi lamanya film berpengaruh signifikan terhadap penilaian penonton. Maka dari itu dapat disimpulkan bahwa kedua memiliki pengaruh yang signifikan terhadap penilaian penonton dalam suatu film di situs Rotten Tomatoes.

Uji Koefisien Determinasi

Uji koefisien determinasi (R^2) adalah alat yang digunakan untuk mengukur seberapa baik model mampu menjelaskan variasi-variasi dependen (Ghozali, 2012). Nilai *R-Square* yang bertujuan untuk menunjukkan seberapa besar kontribusi dari variabel prediktor yang digunakan untuk menjelaskan variabel respon, dengan *R-Square* yang memiliki nilai berada diatas 50%.

Model Summary

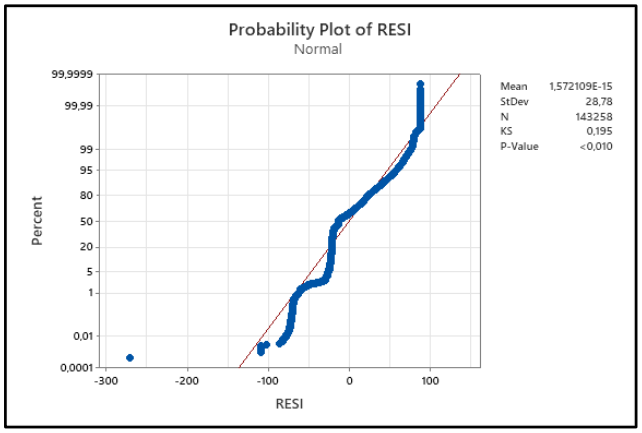
S	R-sq	R-sq(adj)	R-sq(pred)
28,7793	23,51%	23,50%	23,50%

Gambar 1.4 *Output model summary*

Dari *output model summary* di atas didapatkan nilai *R-square* adalah koefisien determinasi yaitu 23,51% yang artinya penilaian dari kritikus dan lamanya durasi dalam film mempengaruhi penilaian penonton sebesar 23,51%, sedangkan sekelilingnya dipengaruhi oleh variabel lain.

Uji Asumsi Normalitas Galat/Error/Residual

Tujuan uji normalitas ialah untuk menguji apakah dalam modelregresi variabel dependen dan variabel independen berkontribusi atau tidak dengan data yang berdistribusi normal atau mendekati normal (Ghozali, 2012). Beberapa metode untuk menguji data berdistribusi normal antara lain: Anderson-Darling, Ryan-Joiner, dan Kolmogorov-Smirnov. Cara tersebut bisa dipilih salah satu. Alternatif lain yaitu dengan Q-Q plot, namun netode ini tidak direkomendasikan. data dianggao berdistribusi normal jika $P_{value} > \alpha$.



Gambar 1.4 *Output Uji Kolmogorov-Smirnov(KS)*

Didapatkan nilai $p - value = 0.010 < 0.05$, yang artinya residualnya tidak berdistribusi normal. Nilai 0,05 adalah nilai standar alpha pada umumnya. Berdasarkan *scatterplot* yang menunjukkan bahwa titik-titik tersebut membentuk garis linier

menunjukkan data residual berdistribusi normal, akan tetapi terdapat beberapa *outlier* sehingga nilai residual menjadi tidak menyebar normal atau dikatakan tidak berdistribusi normal.

Uji Non-autokorelasi

Uji non-autokorelasi bertujuan untuk mendeteksi korelasi antara kesalahan pada periode-t dengan kesalahan pada periode t-1 (sebelumnya). Pengujian non-autokorelasi dilakukan dengan membandingkan nilai durbin watson hitung (d) dengan nilai durbin watson tabel, yaitu batas atas (du) dengan batas bawah (dL). Hipotesisnya adalah:

H_0: $\rho = 0$, (tidak terjadi autokorelasi),

H_1: $\rho \neq 0$, (terjadi autokorelasi).

Dapat dinyatakan gagal tolak H_0 jika $du < d < 4 - du$ (Ghozali, 2012).

Durbin-Watson Statistic

Durbin-Watson Statistic = 1,98890

Gambar 1.5 *Output Uji Durbin Watson*

Pada output di atas didapatkan nilai Durbin-Watson: $d = 1{,}98890$

Tingkat Signifikasi: $\alpha = 0{,}05$.

Jumlah sampel $(n) = 143.258$

Jumlah variabel bebas $(k) = 2$

Mencari nilai dL menggunakan tabel Durbin-Watson didapatkan nilai nya 1.6475. Karena nilai Durbin-Watson berada antara 0 < d = $1{,}98890$ < dL = 1.6475. Maka terjadi autokorelasi positif, artinya terdapat korelasi antar kesalahan pengganggu (residual) pada periode t dengan kesalahan pada periode sebelumnya.

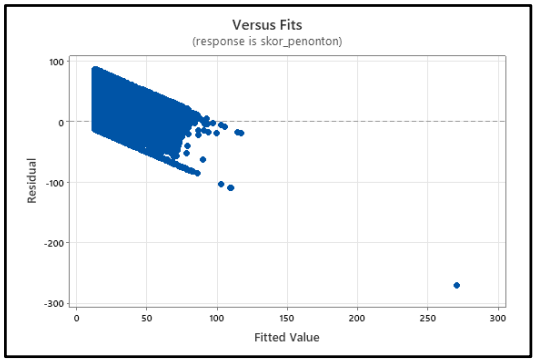
Uji Asumsi Non-Heterokedastisitas

Uji ini bertujuan untuk memastikan varians residual antara pengamatan tetap konstan (homoskedastisitas). Jika varians dari residual berubah antar pengamatan, maka disebut heteroskedastisitas (Ghozali, 2012).Salah satu metode yang digunakan adalah uji glejser, yaitu dengan cara meregresikan variabel-variabel prediktor terhadap nilai absolut residual dan menganalisis hasil uji-t.

Tests

Method	Test Statistic	P-Value
Bartlett	37345,28	0,000

Gambar 1.6 Uji Bartlett



Gambar 1.7 *Output Uji Bartlett*

Didapatkan nilai $p - value$ $0.000 < 0.05$, yang artinya residualnya identik Nilai 0,05 adalah nilai standar alpha pada umumnya. Berdasarkan *scatterplot* yang menunjukkan bahwa titik-titik tersebut tidak menyebar secara acak, maka dapat diindikasikan bahwa secara visual data hasil diatas residual bersifat identik. Hasil menunjukkan bahwa residual tidak memenuhi homoskedastisitas.

Uji Non-Multikolinearitas

Uji non-multikolinearitas digunakan untuk memastikan tidak ada korelasi antar variabel bebas (independen). Model regresi yang baik ialah regresi yang bebas dari multikolinearitas, yang diuji dengan melihat nilai VIF (*Variance Inflation Factor*) dan *tolerance*. *Tolerance* yang rendah sama dengan nilai VIF yang tinggi (karena VIF = 1/*tolerance*) menunjukkan adanya multikolinearitas, dan umumnya *cutoff* yang digunakan adalah *tolerance*≥ 0,01 atau VIF ≤ 10 (Ghozali, 2012).

Kesimpulan dan Saran

Kesimpulan

Penelitian ini telah mengimplementasikan metode regresi linier berganda dalam pengolahan data dari platform Rotten Tomatoes Movies. Berdasarkan hasil analisis, ditemukan bahwa kedua variabel independen, yaitu durasi film (*runtimeMinutes*) dan rating kritikus (*tomatoMeter*), memiliki pengaruh signifikan terhadap variabel dependen, yaitu skor audiens (*audienceScore*). Hal ini dibuktikan dengan nilai koefisien determinasi (R^2) sebesar 23,51%, yang menunjukkan bahwa model regresi linier yang dibangun dapat menjelaskan 23,51% dari variabilitas skor audiens. Persamaan regresi yang diperoleh adalah:

$$Y = 12{,}900 + 0{,}48109X_1 + 0{,}09546X_2 + \varepsilon$$

dengan:

Y : *audienceScore* (skor penonton),

X_1 : *runtimeMinutes* (durasi film),

X_2 : *tomatoMeter* (rating kritikus).

Hasil ini menunjukkan bahwa variabel rating kritikus (*tomatoMeter*) dan durasi film berkontribusi signifikan terhadap penilaian audiens terhadap film. Namun, koefisien determinasi yang relatif rendah mengindikasikan bahwa masih terdapat variabel lain di luar model yang turut memengaruhi penilaian audiens, seperti genre, aktor, atau popularitas film.

Saran

1. Penambahan Variabel Lain: ntuk penelitian lanjutan, disarankan untuk menambahkan variabel independen lainnya, seperti genre film, popularitas aktor, atau tahun rilis, yang kemungkinan juga memengaruhi skor audiens. Hal ini diharapkan dapat meningkatkan nilai koefisien determinasi model.
2. Penggunaan Metode Alternatif: Metode regresi linier berganda memiliki keterbatasan dalam mengungkap hubungan non-linear. Oleh karena itu, disarankan untuk mempertimbangkan metode lain, seperti regresi non-linear atau model *machine learning* yang lebih canggih, untuk prediksi yang lebih akurat.
3. Peningkatan Kualitas Data: Pengolahan data *Rotten Tomatoes* perlu lebih ditingkatkan untuk mengatasi masalah data kosong dan *outlier* yang dapat memengaruhi hasil analisis. Teknik data cleaning dan imputation dapat digunakan untuk mengisi data yang hilang atau mengatasi outlier.
4. Pengujian Asumsi Model yang Lebih Ketat: Hasil uji menunjukkan adanya beberapa pelanggaran asumsi, seperti non-normalitas residual dan heteroskedastisitas. Untuk mendapatkan hasil yang lebih robust, sebaiknya asumsi model diuji lebih lanjut, serta menggunakan teknik yang dapat menangani pelanggaran asumsi seperti regresi robust atau transformasi variabel.

Daftar Pustaka

Arif, M., Syukur, A., & Faisal, M. (2023). Model Regresi Linear Untuk Estimasi Mobil Bekas Menggunakan Bahasa Python. *Jurnal Ilmiah Matematika, Sains Dan Teknologi*, *11*(2), 182–191. https://doi.org/10.34312/euler.v11i2.20698

Baskoro, S., Chamidy, T., & Zaman, S. (2022). Pengujian akurasi model regresi logistik multinomial untuk memprediksi keberhasilan mahasiswa di perguruan tinggi menggunakan r. *Jurnal Ilmiah Akuntansi Dan Keuangan*, *5*(3), 1551–1565. https://journal.ikopin.ac.id/index.php/fairvalue

Ghozali, I. (2012). *Aplikasi Analisis Multivariate dengan Program IBM SPSS 20*. Badan Penerbit Universitas Diponegoro.

Ilmi, R. R., Kurniawan, F., & Harini, S. (2023). Prediksi Rating Film IMDb MenggunakanDesition Tree. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, *10*, 791–798. https://doi.org/10.25126/jtiik.2023106615

Lestari, W. (2019). Clustering Data Mahasiswa Menggunakan Algoritma K-Means Untuk Menunjang Strategi Promosi (Studi Kasus : STMIK Bina Bangsa Kendari). In *SIMKOM* (Vol. 4, Issue 2). http://e-jurnal.stmikbinsa.ac.id/index.php/simkom35

Ningsih, T., Herrhyanto, N., & Rachmatin, D. (2019). *ANALISIS REGRESI LINEAR PIECEWISE DUA SEGMEN DENGAN MENGGUNAKAN METODE KUADRAT TERKECIL*.

Sholeh, M., Kumalasari Nurnawati, E., & Lestari, U. (2023). Penerapan Data Mining dengan Metode Regresi Linear untuk Memprediksi Data Nilai Hasil Ujian Menggunakan RapidMiner. In *Jurnal Informatika Sunan Kalijaga)* (Vol. 8, Issue 1). https://archive.ics.uci.edu/ml/datasheets.php.

Sudrajat, M., & Tjuju S. A. (2010). *Statistika-Konsep Dasar Pengumpulan dan Pengolahan Data*. Widya Padjajaran.

Yamin, S. (2011). *Generasi Baru Mengolah Data Penelitian dengan Partial Least Square Path Modeling*. Penerbit Salemba Infotek.

Zuhri. (2020). Analisis Regresi Linier dan Korelasi menggunakan Pemrograman Visual Basic. *Jurnal Ilman: Jurnal Ilmu Manajemen*, *8*(2), 42–50. http://journals.synthesispublication.org/index.php/ilman