

Implementasi *Data Mining* menggunakan Metode Regresi Linear untuk Memprediksi Nilai Hasil Ujian Akhir

R. Wahyudi Darmawan, Amelia Sholikhaq

¹Program Studi Teknologi Informasi Universitas Tangerang Raya, Tangerang, Indonesia

Sejarah Artikel:

Diterima **Agustus 2024**

Disetujui **Agustus 2024**

Dipublikasi **September 2024**

Kata Kunci: Model, Data Mining, Regresi Linear, Rapid Miner, Datasheet.

Abstrak: Prediksi dalam data mining dapat dilakukan menggunakan regresi linear, yang berfungsi untuk memperkirakan nilai berdasarkan faktor-faktor relevan. Penelitian ini mengembangkan model regresi linear guna memprediksi nilai ujian mahasiswa dengan menggunakan dataset publik *student_performance.csv*, yang terdiri dari 395 data dan 33 atribut. Pemilihan atribut dilakukan melalui analisis korelasi matriks dengan metode pembobotan, menghasilkan tujuh atribut utama sebagai variabel prediktor. Penelitian ini menerapkan metode CRISP-DM, yang mencakup *business understanding*, *data preparation*, *model creation*, *evaluation*, dan *deployment*, dengan RapidMiner sebagai alat analisis. Model regresi linear yang diperoleh adalah:

$$y = 0,729 - (0,024 \times \text{Medu}) - (0,020 \times \text{Fedu}) + (0,053 \times \text{kegagalan}) - (0,077 \times \text{gout}) - (0,012 \times \text{absen}) + (0,126 \times G1) + (0,862 \times G2)$$
$$y = 0,729 - (0,024 \times \text{Medu}) - (0,020 \times \text{Fedu}) + (0,053 \times \text{kegagalan}) - (0,077 \times \text{gout}) - (0,012 \times \text{absen}) + (0,126 \times G1) + (0,862 \times G2)$$

Evaluasi model menggunakan *Root Mean Square Error* (RMSE) menghasilkan nilai 0,675, yang menunjukkan tingkat kesalahan rendah. Dengan demikian, model ini dapat direkomendasikan untuk memprediksi nilai ujian mahasiswa secara akurat.

Abstract: Prediction in data mining can be done using linear regression, which functions to estimate values based on relevant factors. This study develops a linear regression model to predict student exam scores using the public dataset *student_performance.csv*, which consists of 395 data and 33 attributes. Attribute selection is done through matrix correlation analysis with a weighting method, resulting in seven main attributes as predictor variables. This study applies the CRISP-DM method, which includes *business understanding*, *data preparation*, *model creation*, *evaluation*, and *deployment*, with RapidMiner as an analysis tool. The linear regression model obtained is:

$$y = 0.729 - (0.024 \times \text{Medu}) - (0.020 \times \text{Fedu}) + (0.053 \times \text{failure}) - (0.077 \times \text{gout}) - (0.012 \times \text{absent}) + (0.126 \times G1) + (0.862 \times G2)$$
$$y = 0.729 - (0.024 \times \text{Medu}) - (0.020 \times \text{Fedu}) + (0.053 \times \text{kegagalan}) - (0.077 \times \text{gout}) - (0.012 \times \text{absen}) + (0.126 \times G1) + (0.862 \times G2)$$

Model evaluation using *Root Mean Square Error* (RMSE) yields a value of 0.675, indicating a low error rate. Thus, this model can be recommended to predict student exam scores accurately.

PENDAHULUAN

Perkembangan data mining mengalami pertumbuhan yang pesat, seiring dengan meningkatnya jumlah data yang digunakan dalam proses pengambilan keputusan. Model berbasis data mining digunakan untuk mendukung kebijakan yang dibuat berdasarkan analisis data. Dalam dunia bisnis, data menjadi aset berharga yang dapat mengungkap pola-pola penting guna mendukung pengambilan keputusan strategis. Model yang dihasilkan dari analisis data digunakan untuk memproyeksikan strategi atau kebijakan yang diterapkan dalam pengembangan bisnis. Oleh karena itu, pengelolaan data dalam skala besar harus dilakukan dengan cermat, karena semakin besar volume data, semakin kompleks pula proses evaluasi yang diperlukan untuk memastikan relevansi dan akurasi informasi.

Data *mining* merupakan proses ekstraksi informasi dari data yang tersimpan dalam database atau datasheet. Model dalam data mining dikembangkan melalui proses yang melibatkan algoritma atau formula tertentu. Teknik yang digunakan dalam proses ini mencakup metode statistik, matematika, serta *machine learning*, yang berperan dalam mengidentifikasi dan mengolah data menjadi informasi yang bernilai (Arhami & Nasir, 2020; Jolly et al., 2020).

Salah satu model prediksi yang digunakan dalam data mining adalah regresi linear. Analisis regresi berfungsi untuk mengukur pengaruh antara variabel independen terhadap variabel dependen. Regresi linear terbagi menjadi dua jenis, yaitu regresi linear sederhana dan regresi linear berganda. Regresi linear sederhana melibatkan satu variabel independen yang mempengaruhi satu variabel dependen, sedangkan regresi linear berganda melibatkan lebih dari satu variabel independen. Analisis regresi digunakan untuk menentukan sejauh mana variabel independen berpengaruh terhadap variabel dependen. Variabel yang dipengaruhi disebut sebagai variabel dependen, sementara variabel yang memberikan pengaruh disebut sebagai variabel independen. Model persamaan matematis ditampilkan pada Persamaan (1)

$$y = a_0 + A_1x_1 + A_2x_2 + \dots = A_nx_n \dots \dots \dots (1)$$

Dimana y adalah variabel *dependent* dan x_1, x_2, \dots, x_n merupakan variabel *independent*, a merupakan nilai konstanta, dan b adalah nilai koefisien regresi (Kurniawan, 2016).

Regresi linear adalah salah satu metode yang paling umum digunakan dalam data mining. Teknik ini dapat diterapkan pada berbagai jenis data, baik dari sumber pribadi maupun publik. Salah satu dataset publik yang sering digunakan dalam pengembangan model data mining adalah *student_performance.csv*, yang berisi informasi mengenai siswa dan latar belakang keluarga dari SMKS XYZ di Kabupaten Tangerang. Dalam penelitian ini, dataset tersebut digunakan sebagai sumber utama untuk pengembangan model data mining.

Setoyprini dan Asmono (2020) memanfaatkan dataset Kinerja Siswa untuk membangun model klasifikasi menggunakan metode K-Nearest Neighbour. Unal (2021) mengembangkan model klasifikasi dengan menerapkan algoritma Decision Tree, Random Forest, dan Naïve Bayes. Sementara itu, Deepika dan Sathyanarayana (2018) juga menerapkan Decision Tree, sedangkan Oyedeji et al. (2020) menganalisis kinerja akademik siswa guna menemukan strategi yang dapat meningkatkan hasil belajar individu.

Model data mining untuk memprediksi keberhasilan akademik siswa menggunakan datasheet dengan data kinerja siswa yang bersifat privat telah diteliti oleh Ofori et al. (2020). Penelitian tersebut mengidentifikasi model pembelajaran mesin yang mampu memprediksi hasil belajar siswa serta menemukan teknik *machine learning* yang efektif dalam meningkatkan pengalaman belajar. Sementara itu, Bahri et al. (2022) mengembangkan model yang bertujuan membantu siswa dalam memilih jurusan di perguruan tinggi. Berdasarkan hasil analisis, faktor utama yang menyebabkan kesalahan dalam pemilihan jurusan adalah kurangnya informasi mengenai jurusan tersebut. Pengujian dilakukan dengan menerapkan tiga algoritma, di mana algoritma *Decision Tree* menunjukkan akurasi tertinggi sebesar 75,38%. Penelitian lain yang membahas penggalian data dengan melibatkan data siswa juga telah dilakukan oleh Hendrian (2018), Putro et al. (2021), serta Ramadhani dan Hendriyani (2021).

Berbagai penelitian mengenai data mining dengan beragam datasheet dan metode regresi linear telah banyak dilakukan dalam berbagai bidang aplikasi. Ariesanto dan Ekka (2020) meneliti penggunaan regresi linear dalam data mining untuk memprediksi harga saham pada perusahaan layanan, dengan evaluasi model menggunakan Root Mean Square Error (RMSE), yang menghasilkan nilai sebesar 7,522 dibandingkan dengan harga penutupan saham sebenarnya. Sementara itu, Goal et al. (2019) menerapkan regresi linear berganda untuk memperkirakan jumlah ketersediaan buku dalam suatu sistem. Penelitian serupa juga dilakukan oleh Rahayu et al. (2022), Sinaga et al. (2022), serta Siregar (2021), yang menunjukkan berbagai penerapan regresi linear dalam analisis data.

Pengembangan model data mining dapat dilakukan melalui berbagai aplikasi, baik dengan menulis kode pemrograman maupun menggunakan *Visual Programming*. Sholeh et al. (2022) menerapkan pendekatan berbasis pemrograman dalam penelitian mereka yang berfokus pada prediksi biaya asuransi menggunakan dataset asuransi kesehatan. Model tersebut dikembangkan dengan bahasa pemrograman Python, memanfaatkan pustaka seperti *pandas* dan *sklearn* untuk membangun serta mengoptimalkan model. Penelitian lain yang juga menggunakan Python dalam penerapan data mining telah dilakukan oleh Kurniatullah & Pramudi (2017), N. et al. (2019), Nishadi (2019), serta Prabha et al. (2020).

Selain menggunakan bahasa pemrograman, pembuatan model *data mining* juga dapat dilakukan dengan menggunakan Visual Programming seperti RapidMiner. RapidMiner memungkinkan pengguna untuk melakukan analisis *data mining*, *text mining*, serta analisis prediktif tanpa perlu menulis kode pemrograman. Aplikasi ini menerapkan berbagai teknik deskriptif dan prediktif dalam pengembangan model yang dapat digunakan untuk mendukung pengambilan keputusan (Chisholm, 2013). Keunggulan RapidMiner terletak pada kemudahan penggunaannya, karena seluruh proses pembuatan model dilakukan melalui operator yang sudah tersedia, tanpa memerlukan penulisan kode. Pembuatan model dilakukan dengan menghubungkan berbagai operator yang sesuai untuk membentuk alur analisis data yang optimal. Beberapa penelitian telah menerapkan RapidMiner dalam pembuatan model data mining dengan berbagai datasheet, model, dan algoritma yang digunakan. Penelitian tersebut antara lain dilakukan oleh Chisholm (2013), Prasetyo et al. (2021), dan Sudarsono et al. (2021).

Berdasarkan latar belakang, tinjauan pustaka, dan studi literatur, data mining merupakan metode yang dapat digunakan untuk membangun model dalam mendukung proses pengambilan keputusan. Salah satu pendekatan yang umum diterapkan dalam data mining adalah regresi linear. Penelitian ini berfokus pada pengembangan model regresi linear dengan memanfaatkan datasheet *Student Performance* yang tersedia secara publik. Datasheet tersebut berisi 395 data dengan 33 atribut. Pemilihan dataset ini didasarkan pada jumlah data dan atribut yang cukup besar, sehingga memungkinkan berbagai pengujian untuk mengidentifikasi faktor-faktor yang memengaruhi pembentukan model. Dengan demikian, model yang dihasilkan diharapkan mampu memprediksi nilai siswa dengan tingkat akurasi yang optimal.

METODOLOGI PENELITIAN

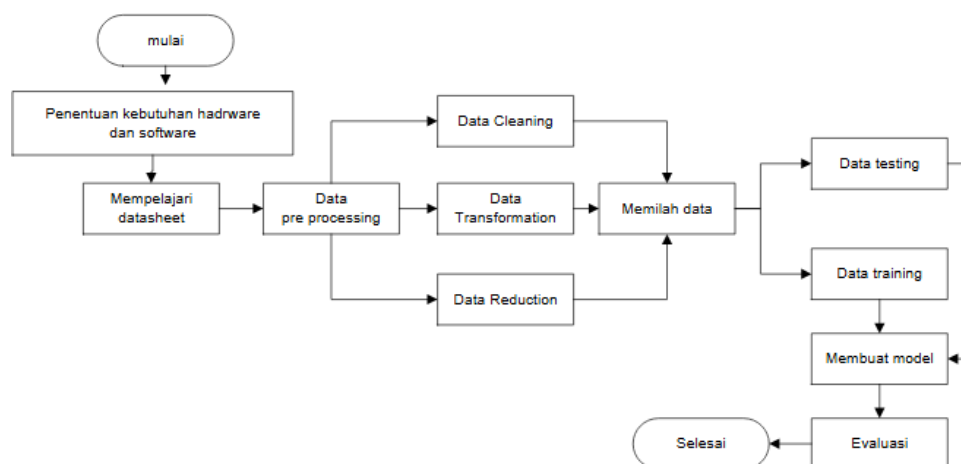
Penelitian ini menggunakan metodologi CRISP-DM dalam pengembangan model data mining dengan pendekatan regresi linear. Proses pembuatan model berdasarkan CRISP-DM terdiri dari enam tahapan utama, yaitu pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan penerapan (Hidayat et al., 2021)

Datasheet

Datasheet yang digunakan adalah StudentPerformance.csv. Datasheet terdiri dari 395 baris dan 33 atribut.

Tahapan Penelitian

Penelitian ini mengikuti tahapan model CRISP-DM, yang mencakup beberapa langkah, antara lain mempelajari dan membersihkan datasheet, seperti memeriksa data kosong, memastikan batas nilai, dan melakukan validasi lainnya. Dalam proses pembuatan model, datasheet dibagi menjadi dua bagian, di mana 80% digunakan sebagai data *training* dan 20% sisanya sebagai data *testing*.



Gambar 1. Tahapan pembuatan Data Mining

HASIL DAN PEMBAHASAN

Business Understanding

Tahap awal dalam penelitian ini mencakup identifikasi manfaat serta penerapan model yang dikembangkan. Analisis dilakukan terhadap dataset guna memahami keterkaitan antaratribut, terutama yang berkaitan dengan atribut berlabel. Model yang dihasilkan diharapkan mampu mengungkap faktor-faktor yang memengaruhi kegagalan siswa dalam ujian serta digunakan untuk memprediksi skor ujian akhir.

Data Understanding

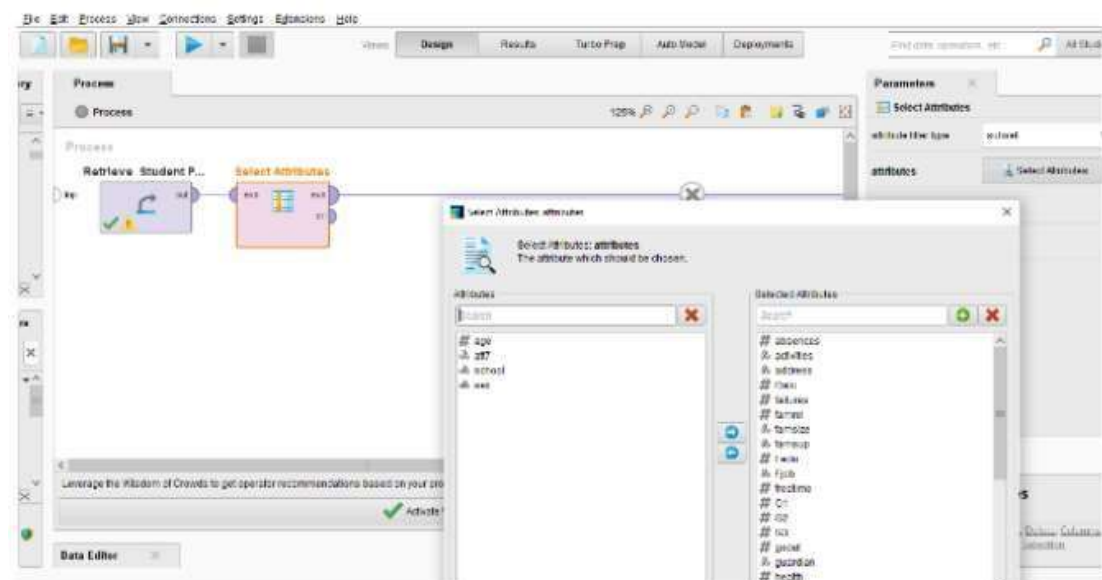
Pada tahap ini, proses dilakukan dengan mengidentifikasi atribut yang terdapat dalam datasheet. Langkah pertama dalam penyortiran adalah menentukan atau memilih atribut yang tidak memiliki keterkaitan atau hubungan dalam pembentukan model. Atribut yang tidak berpengaruh terhadap pemodelan akan dikeluarkan dan tidak digunakan dalam proses lebih lanjut.

Data Preparation

Data *preparing* merupakan langkah penting yang bertujuan untuk memastikan bahwa data yang digunakan bebas dari kesalahan. Proses ini mencakup pemeriksaan terhadap data kosong, nilai yang berada di luar ambang batas, serta ketidaksesuaian tipe data. Tahapan data *preparing* yang dilakukan dalam penelitian ini adalah sebagai berikut.

1. Memilih atribut yang akan digunakan

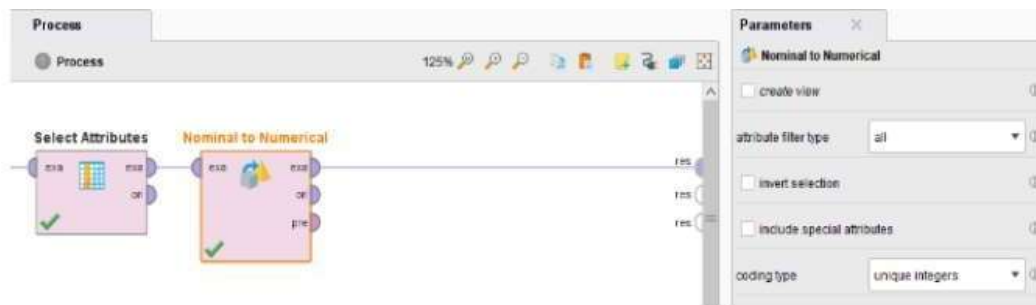
Tidak semua atribut digunakan dalam pembuatan model, sehingga hanya atribut yang berpengaruh yang dipilih. Dari 33 atribut yang tersedia, *att*, *sex*, *age*, dan *school* tidak disertakan



Gambar 2. Penggunaan Operator Select Atribut dalam proses Pemilihan Atribut

2. Mengubah tipe data nominal menjadi numerik

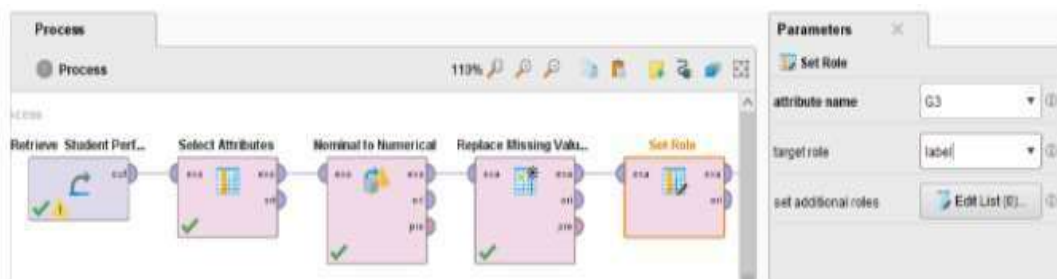
Atribut dengan data non-numerik dikonversi ke dalam format numerik. Sebagai contoh, atribut *sex*, yang awalnya memiliki nilai *F* dan *M*, diubah menjadi 0 dan 1. Sementara itu, atribut *Mjob*, yang terdiri dari kategori *other*, *services*, *at_home*, *teacher*, dan *health*, dikonversi ke nilai numerik dalam rentang 0–4. Gambar 3 menunjukkan penggunaan operator *nominal to numerical* dalam proses konversi data nominal menjadi numerik.



Gambar 3. Penggunaan Operator Nominal to Numerical

3. Mengganti data *missing value* dengan data tertentu

Label yang digunakan untuk membuat model G3. Pada gambar 4 menggambarkan penggunaan operator *set role* untuk menentukan atribut G3 menjadi label.



Gambar 4. Penggunaan Operator Set Role untuk menentukan label

Pembuatan Model

Langka selanjutnya adalah membuat model regresi linier menggunakan atribut yang ditentukan oleh tahapan pemrosesan model sebagai berikut :

1. Menampilkan bobot hasil matriks kolerasi.

Matriks korelasi merupakan tabel yang berisi koefisien korelasi dari seluruh atribut yang digunakan dalam analisis. Matriks ini digunakan untuk menentukan tingkat hubungan antara setiap atribut. Gambar 5 menunjukkan penerapan operator *Weight by Correlation* dalam mengukur keterkaitan antar atribut, sedangkan Gambar 6 menampilkan hasil bobot keterkaitan tersebut, khususnya terhadap atribut label, yaitu G3.



Gambar 5. Penggunaan Operator *Weight by Correlation*

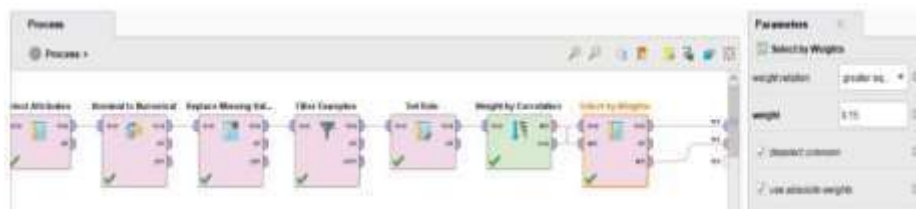
attribute	weight	attribute	weight
G2	0.956	travelltime	0.102
G1	0.892	Mjob	0.084
failures	0.294	health	0.082
schoolsup	0.238	famsup	0.067
absences	0.213	reason	0.061
Walc	0.190	activities	0.059
Medu	0.188	romantic	0.050
goout	0.177	famsize	0.040
Fedu	0.163	famrel	0.038
Dalc	0.141	guardian	0.035
address	0.130	Fjob	0.032
studytime	0.121	paid	0.029
higher	0.113	Pstatus	0.027
internet	0.112	nursery	0.027
		freetime	0.022

Gambar 6. Hasil Pembobotan Keterkaitan Atribut Terutama dengan Atribut Label G3

2. Menentukan nilai bobot yang digunakan dalam proses pembuatan model.

Berdasarkan analisis bobot yang ditampilkan pada Gambar 6, ditentukan ambang batas bobot yang berlaku dalam studi ini, yakni di atas 0,15. Pemilihan nilai ini mempertimbangkan bahwa 0,15 masih berada dalam rentang kategori korelasi yang cukup baik ($>0,25 - 0,5$) dan juga memberikan kesempatan untuk menganalisis atribut yang berada di bawah batas tersebut guna mengevaluasi dampaknya terhadap model.

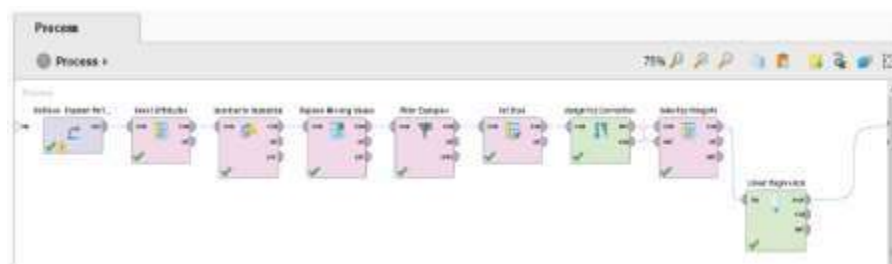
Pengembangan model ini melibatkan beberapa atribut utama, di antaranya *G2* (nilai ujian kedua), *G1* (nilai ujian pertama), *failures* (jumlah kegagalan dalam kelulusan), *schoolsup* (dukungan akademik tambahan di luar sekolah), *absences* (jumlah ketidakhadiran dalam kelas), *Walc* (tingkat konsumsi alkohol), *Medu* (tingkat pendidikan ibu), *gout* (frekuensi interaksi atau bermain dengan teman sebaya), dan *Fedu* (tingkat pendidikan ayah). Pemilihan atribut dilakukan menggunakan operator *Select by Weight* dengan kriteria bobot di atas 0,15. Gambar 7 menunjukkan proses penggunaan operator ini.



Gambar 7. Proses Pemebrian Bobot di Atas 0,15

3. Proses pembuatan model.

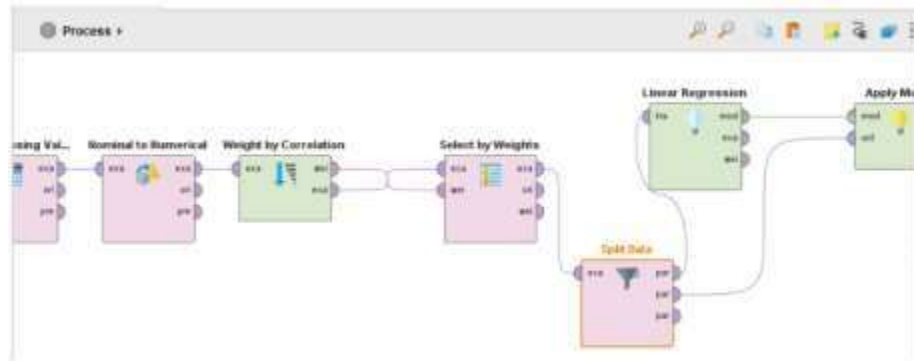
Model *linier regresi* dapat dijalankan berdasarkan bobot. Pada gambar 8 menunjukkan penggunaan operator *linier regresi* dalam proses pembuatan model.



Gambar 8. Penggunaan Operator *Linear Regresi*

4. Pembuatan model dengan *data training* dan pengujian dengan *data testing*.

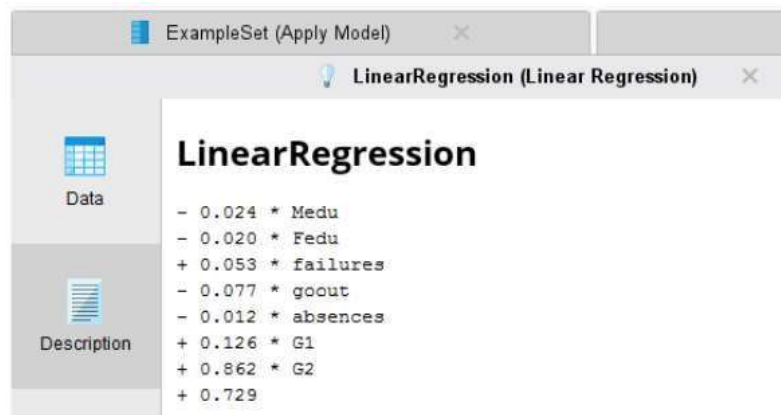
Model yang ditampilkan pada Gambar 8 dikembangkan menggunakan seluruh data dalam dataset tanpa melalui tahap pengujian. Dalam proses pengembangan, data dibagi menjadi dua bagian: *training* dan *testing*, di mana 80% data digunakan untuk melatih model, sementara 20% sisanya digunakan untuk evaluasi. Pembagian data dilakukan dengan operator *split data*, sedangkan pengujian model dilakukan menggunakan operator *apply model*. Gambar 9 menunjukkan tahapan pembuatan model dengan data *training* serta proses evaluasi menggunakan data *testing*.



Gambar 9. Proses Pembagian Data untuk *Data Training* dan *Data Testing*

Hasil dari proses pemodelan menghasilkan model linear regresi. Hasilnya ditunjukkan pada Pers. (2) dan Gambar 10.

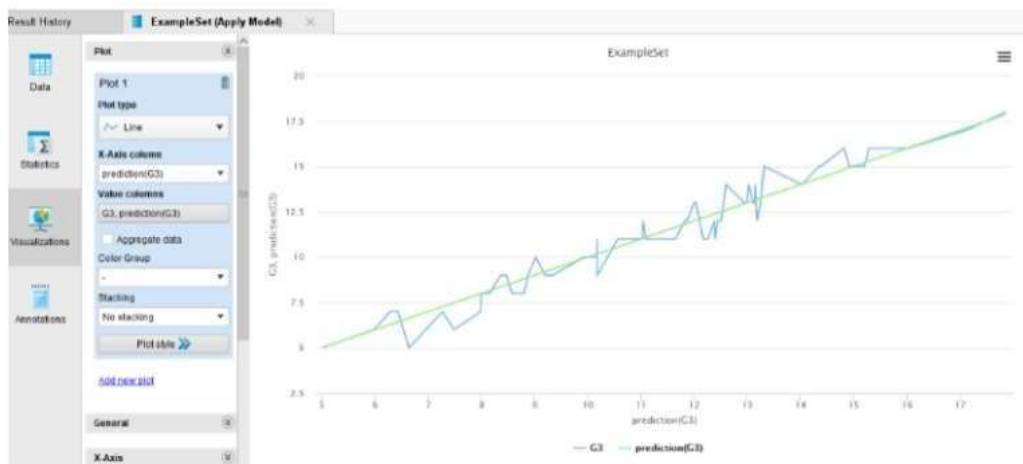
$$y = 0,729 - (0,024 \times Medu) - (0,020 \times Fedu) + (0,053 \times failures) - (0,077 \times goout) - (0,012 \times absences) + (0,126 \times G1) + (0,862 \times G2)$$



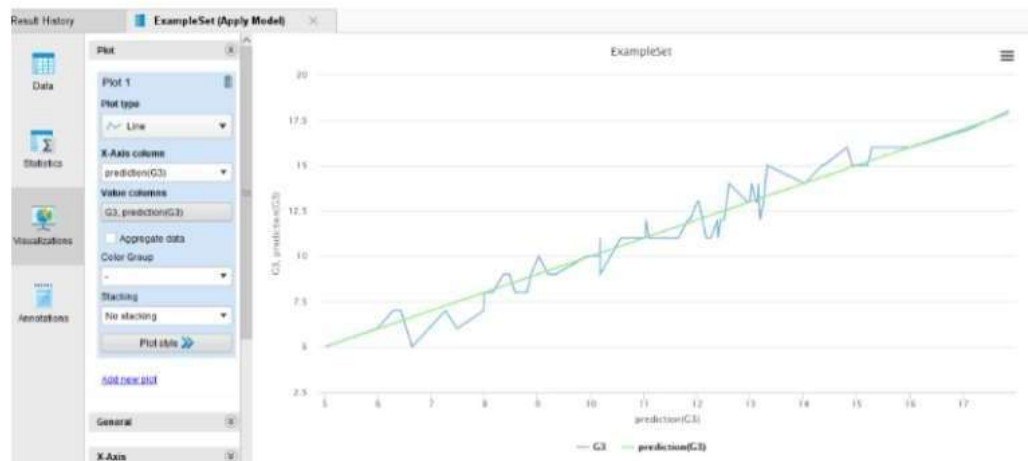
Gambar 10. Hasil Model Linear Regresi

Model yang ditampilkan pada Gambar 10 diuji menggunakan data *training*. Hasil pengujian dianalisis dengan membandingkan nilai asli G3 dengan nilai prediksi. Gambar 11 memperlihatkan hasil perbandingan tersebut. Pada baris pertama, nilai asli G3 adalah 6, sedangkan nilai prediksi G3 sebesar 5,978, menunjukkan sedikit selisih dengan hasil prediksi yang berada di bawah nilai asli. Sebaliknya, pada baris ke-14, nilai asli G3 adalah 11, sementara nilai prediksi G3 sebesar 11,280, yang menunjukkan selisih dengan hasil prediksi di atas nilai asli.

Perbandingan antara nilai prediksi dan nilai asli G3 juga divisualisasikan dalam bentuk grafik. Dalam visualisasi ini, nilai prediksi ditampilkan sebagai garis lurus, sementara nilai asli tersebar di sekitar garis tersebut. Gambar 12 menyajikan grafik perbandingan ini untuk mempermudah analisis hasil model.



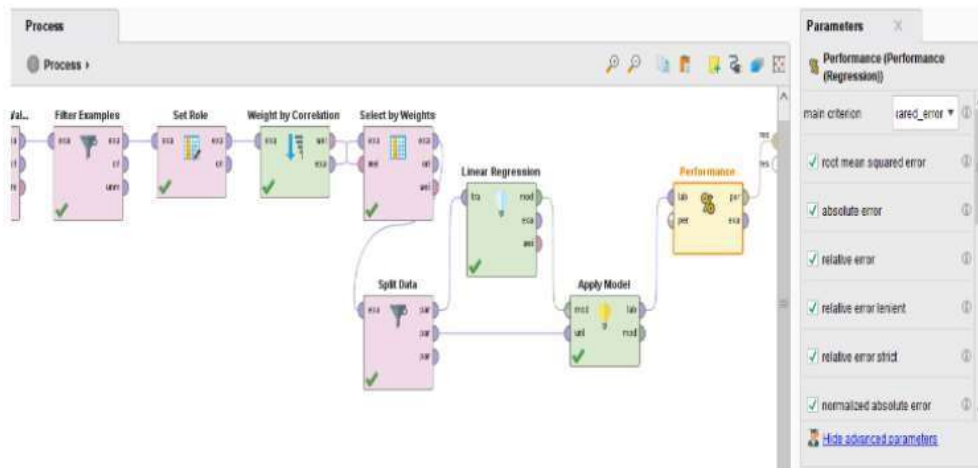
Gambar 11. Perbandingan Nilai G3 dengan Nilai Hasil Prediksi



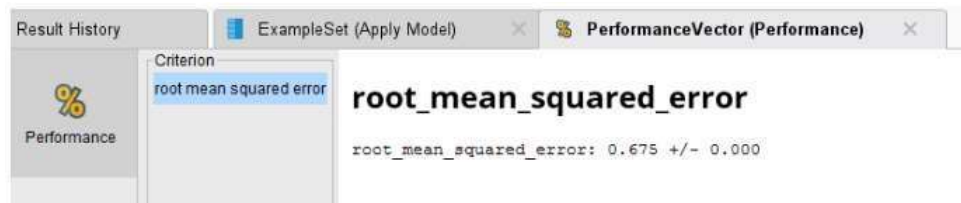
Gambar 12. Grafik Perbandingan Nilai G3 dengan Nilai Hasil Prediksi

Evaluasi Model

Evaluasi terhadap model yang telah selesai dilakukan bertujuan untuk menilai kinerjanya. Tahap evaluasi ini menggunakan data uji yang diambil sebanyak 20% dari keseluruhan datasheet. Pengujian dilakukan dengan menerapkan operator performance. Gambar 13 menunjukkan proses pengujian dengan menggunakan *operator performance* dan Gambar 14 adalah luaran dari *performance*.



Gambar 13. Proses Pengujian dengan Melihat Performance



Gambar 14. Hasil Luaran Performance

SIMPULAN

Data mining dapat digunakan untuk mendukung proses pengambilan keputusan, salah satunya melalui metode regresi linear dalam melakukan prediksi. Dalam penelitian ini, dari 33 atribut yang terdapat dalam dataset *student_performance.csv*, tidak semua memiliki pengaruh signifikan terhadap variabel prediktor. Atribut dengan pengaruh terkuat adalah *G2* (nilai ujian kedua) dengan bobot 0,966 dan *G1* (nilai ujian pertama) dengan bobot 0,892. Selain itu, beberapa atribut lain dengan bobot di atas 0,2, seperti *failures* (jumlah kegagalan atau pernah tinggal kelas), *schoolsup* (dukungan pembelajaran tambahan di luar sekolah), dan *absences* (jumlah ketidakhadiran di kelas), juga memberikan kontribusi dalam prediksi. Sementara itu, atribut dengan bobot di bawah 0,2 tidak memiliki pengaruh signifikan terhadap variabel prediktor. Model regresi linear yang diperoleh adalah:

$$y = 0,729 - (0,024 \times \text{Medu}) - (0,020 \times \text{Fedu}) + (0,053 \times \text{failures}) - (0,077 \times \text{gout}) - (0,012 \times \text{absences}) + (0,126 \times \text{G1}) + (0,862 \times \text{G2})$$

Evaluasi model menggunakan *Root Mean Square Error* (RMSE) menghasilkan nilai 0,675. Nilai RMSE yang lebih rendah menunjukkan bahwa prediksi semakin mendekati nilai yang diamati. Berdasarkan hasil ini, model yang dihasilkan memiliki tingkat kesalahan yang rendah dan dapat direkomendasikan untuk memprediksi nilai siswa dengan akurasi yang baik.

DAFTAR PUSTAKA

- Arhami, M., & Nasir, M. (2020). *Data Mining – Algoritma dan Implementasi*. Penerbit Andi. https://books.google.co.id/books/about/Data_Mining_Algoritma_dan_Implementasi.html?id=AtcCEAAQBAJ&redir_esc=y
- Ariesanto, A., & Ekka, P.(2020). Data Mining Menggunakan Regresi Linear untuk Prediksi Harga Saham Perusahaan Pelayanan. *Jurnal Alikasi Pelayanan Dan Kepelabuhan*, 10(2), 120. <https://doi.org/10.30649/japk.v10i2.83>
- Bahri, S., Itb, A., & Dalan, J. (2022). Implementasi Data Mining Untuk Menentukan Minat Siswa Dalam Menentukan Jurusan Pada Perguruan Tinggi. *Jurnal Sistem Informasi (JUSIN)*, 3(1), 23-33. <https://ojs.itb-ad.ac.id/index.php/JUSIN/article/view/1644>
- Chrisholm, A. (2013). *Exploring Data with RapidMiner* (Vol. 1). Packt Publishing. [Htts://www.perlego.com/book/390375/exploring-data-with-rapidminer-pdf](https://www.perlego.com/book/390375/exploring-data-with-rapidminer-pdf)
- Deepika, K., & Sathyanarayana, N. (2018). Comparison Of Student Academic Performance On Different Educational Datasets Using Different Data Mining Techniques. *International Journal of Computational Engineering Research (IJCER)*, 8(9), 28–38. https://www.ijceronline.com/papers/Vol8_issue9?Version-2/E0809022838.pdf
- N., A. G., Singh, B. P., Sah, B., & Tiwari, D. (2019). Air Quality Index Prediction using Linear Regression. *International Journal of Recent Technology and Engineering (IJRTE)*, 8(2), 4247-4252. <https://doi.org/10.35940/ijrte.B2437.078219>
- Gaol, I. L. L., Sinurat, S., & Siagian, E. R. (2019). IMPLEMENTASI DATA MININGDENGAN METODE REGRESI LINEAR BERGANDA UNTUK MEMPREDIKSI DATA PERSEDIAAN BUKU PADA PT. YUDHISTIRA GHALIA INDONESIA AREA SUMATERA UTARA. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, 3(1). <https://doi.org/10.30865/komik.v3i1.1579>
- Hendrian, S. (2018). Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan. *Faktor Exacta*, 11(3). <https://doi.org/10.30998/faktorexacta.v11i3.2777>
- Hidayati, N, Suntoro, J., & Setiaji, G. G. (2021). Perbandingan Algoritma Klasifikasi untuk Prediksi Cacat Software dengan Pendekatan CRISP-DM. *Jurnal Sains Dan Informatika*, 7(2), 117-126. <https://doi.org/10.34128/jsi.v7i2.313>
- Jollyta, D., Ramdhan, W., & Zarlis, M. (2020). Konsep Data Mining Dan Penerapan. In *Konsep Data Mining Dan Penerapan*. Deepublish. <https://deepublishstore.com/shop/buku-konsep-data-mining-dan-penerapan/>
- Kurniatullah, B. D. F., & Pramudi, Y. T. C. (2017). Estimation of Students' Graduation Using Multiple Linear Regression Method. *Journal of Applied Intelligent System*, 2(1), 29-36. <https://doi.org/10.33633/jais.v2i1.1415>
- Kurniawan, R. (2016). *Analisis Regresi*. Dasar dan Penerapannya dengan R. Prenada Media. <https://prenadamedia.com/product/analisis-regresi-dasarOdan-penerapannya-dengan-r/>

- Nishadi, A. S. T. (2019). Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab. *International of Advanced Research and Publications*, 3(8), 69-74. <https://www.kaggle.com>
- Ofori, F., Maina, E, & Gitonga, R. (2020). Using Machine Learning Algorithms to Predict Students' Performance and Improve Learning Outcome: A Literature Based Review. *Journal of Information and Technology*, 4(1), 2616-3573, <https://stratfordjournals.org/journals/index.php/Journal-of-Information-and-Tech/article/view/480>
- Oyediji, A. O., Salami, A. M., Folorunsho, O., & Abolade, O. R. (2020). Analysis and Prediction of Student Academic Performance Using Machine Learning. *JITCE (Journal of Information Technology and Computer Engineering)*, 4(01), 10-15. <https://doi.org/10.25077/jitce.4.01.10-15.2020>
- Prabha, D., Anindhitha, A., Archana, A., & Balaji, N. M. v. (2020). Predicting House Price Values Using Linear Regression with Ridge Regularization Approach. *International Journal of Advanced Science and Technology*, 29(9s), 5489-5495. <http://serc.org/journals/index.php/IJAST/article/view/18069>
- Prasetyo, V. R., Lazuardi, H., Mulyono, A. A., & Lauw, C. (2021). Penerapan Aplikasi RapidMiner Untuk Prediksi Nilai Tukar Rupiah Terhadap US Dollar Dengan Metode Linear Regression. *Jurnal Nasional Teknologi Dan Sistem Informasi*, 7(1), 8-17. <https://doi.org/10.25077/TEKNOSI.v7i1.202.8-17>
- Putro, M. F., Prayitno, E., Siregar, J., & Muharrom, M. (2021). PENERAPAN DATA MINING DALAM PENETUAN PERGURUAN TINGGI. *Akrab Juara : Jurnal Ilmu-Ilmu Sosial*, 6(2), 306-312. <https://doi.org/10.58487/AKRABJUARA.V6I2.1473>
- Rahayu, E., Parlina, I., & Siregar, Z. A. (2022). Application of Multiple Linear Regression Algorithm fo Motorcycle Sales Estimation. *JOMLA: Journal of Machine Learning and Artificial Intelligence*, 1(1), 1-10. <https://doi.org/10.55123/jomlai.v1i1.142>
- Ramadhani, R., & Hendriyani, Y. (2021). Prediksi Prestasi Siswa Berbasis Data Mining Menggunakan Algoritma Decision Tree (Studi Kasus: SMKN 2 Padang). *Voteteknika (Vocational Teknik Elektronika Dan Informatika)*, 9(3), 11. <https://doi.org/10.24036/voteteknika.v9i3.112633>