JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 8, No. 1, JANUARI, 2023, Pp. 10 – 21
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Penerapan *Data Mining* dengan Metode Regresi Linear untuk Memprediksi Data Nilai Hasil Ujian Menggunakan RapidMiner

Muhammad Sholeh [(1)*], Erna Kumalasari Nurnawati [(2)], Uning Lestari [(3)]
Informatika, Fakultas Teknologi Informasi dan Bisnis, Institut Sains & Teknologi AKPRIND, Yogyakarta
e-mail : {muhash,ernakumala,uning}@akprind.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 7 Juni 2022, direvisi 1 September 2022, diterima 3 September 2022, dan dipublikasikan 30 Januari 2023.

Abstract

Prediction is one of the methods in data mining. One of the models that can be used in prediction is using linear regression. Linear regression is used to make predictions on the data that has been provided. In this study, a linear regression model was made with a datasheet containing data that affected student achievement in achieving final exam scores. The linear regression model developed can be used to predict student test scores. The linear regression model developed can be used to predict student test scores. The datasheet used in the test uses a public datasheet, namely student_performance.csv. The datasheet consists of 395 records and 33 attributes. The attributes used are selected that influence the label. The selection of attributes is based on the results of the weighting in the process of checking the correlation matrix. Based on the weighting, the attributes used are seven attributes and one attribute becomes a label. The research method uses CRISP DM which consists of business understanding, data understanding, data preparation, model making, evaluation, and deploying. The data mining process uses the Rapid Miner application. The results of the study resulted in a linear regression model $y = 0.729 - (0.024 \times Medu) - (0.020 \times Fedu) + (0.053 \times failures) - (0.077 \times goout) - (0.012 \times absences) + (0.126 \times G1) + (0.862 \times G2)$*. The result of evaluating the performance of the RMSE value was 0.675. Based on these results, it can be concluded that the resulting model can be recommended for use in predicting student test scores.*

Keywords: Model, Data Mining, Linear Regression, RapidMiner, Datasheet

Abstrak

Salah satu metode dalam *data mining* adalah prediksi. Salah satu model yang dapat digunakan dalam prediksi adalah menggunakan regresi linear. Regresi linear digunakan untuk melakukan prediksi pada data yang sudah disediakan. Pada penelitian ini dilakukan pembuatan model regresi linear dengan *datasheet* berisi data-data yang mempengaruhi prestasi siswa dalam meraih nilai ujian akhir. Model regresi linear yang dikembangkan dapat digunakan untuk melakukan prediksi hasil nilai ujian siswa. *Datasheet* yang digunakan dalam pengujian menggunakan *datasheet* publik yaitu student_performance.csv. *Datasheet* terdiri dari 395 data dan 33 atribut. Atribut yang digunakan dipilih yang mempunyai pengaruh pada label. Pemilihan atribut berdasar hasil pembobotan pada proses pengecekan korelasi matriks. Berdasarkan pada pembobotan, atribut yang digunakan adalah tujuh atribut dan satu atribut menjadi label. Metode penelitian menggunakan CRISP-DM yang terdiri dari *business understanding, data understanding, data preparation*, pembuatan model, evaluasi dan *deploying*. Proses *data mining* menggunakan aplikasi RapidMiner. Hasil penelitian menghasilkan model regresi linear $y = 0{,}729 - (0{,}024 \times Medu) - (0{,}020 \times Fedu) + (0{,}053 \times failures) - (0{,}077 \times goout) - (0{,}012 \times absences) + (0{,}126 \times G1) + (0{,}862 \times G2)$. Hasil evaluasi performance nilai RMSE adalah 0,675. Berdasar hasil tersebut dapat disimpulkan bahwa model yang dihasilkan dapat direkomendasikan untuk digunakan dalam memprediksi nilai ujian siswa.

Kata Kunci: Model, *Data Mining*, Regresi Linear, RapidMiner, *Datasheet*

1. PENDAHULUAN

Perkembangan *data mining* tumbuh dengan sangat pesat. Hal ini seiring dengan tumbuhnya data yang semakin banyak dan digunakan dalam proses pengambilan kebijakan. Kebijakan yang diambil dengan menggunakan data dilakukan dengan pembuatan model *data mining*. Saat ini, data menjadi unsur penting dalam suatu perusahaan. Data menjadi aset yang dapat digunakan untuk mencari pola yang dapat digunakan dalam pengambilan kebijakan. Informasi dari model yang diolah dapat digunakan dalam memproyeksikan strategi atau kebijakan yang dilakukan untuk proses pengembangan bisnis. Proses penelusuran data yang besar harus dilakukan secara cermat. Semakin besar data semakin besar proses untuk melakukan pemilahan data yang sesuai dengan keperluan.

Data mining merupakan proses menemukan informasi dari suatu data yang tersimpan dalam suatu *database* atau *datasheet*. Pembuatan model dilakukan dengan proses menggunakan algoritma atau rumus tertentu. Proses *data mining* menggunakan berbagai teknik seperti teknik dalam proses statistik, matematika, dan *machine learning* yang digunakan dalam melakukan identifikasi dan mengolah berbagai data menjadi informasi yang bermanfaat (Arhami & Nasir, 2020; Jollyta et al., 2020).

Salah satu model prediksi yang digunakan pada *data mining* adalah regresi linear. Analisis regresi dapat digunakan untuk melihat pengaruh antara variabel bebas (*independen*) dan variabel tidak bebas (*dependen*). Regresi linear dibedakan menjadi regresi sederhana dan regresi linear berganda. Regresi linear sederhana hanya terdapat satu variabel bebas dan satu variabel yang menjadi variabel tidak bebas dan regresi linear berganda apabila terdapat lebih dari satu variabel bebas. Kegunaan dari analisis regresi linear adalah untuk mengetahui arah dan seberapa besar pengaruh variabel independen terhadap variabel dependen. Variabel yang dapat mempengaruhi sering disebut variabel dependen atau tidak bebas dan variabel yang mempengaruhi variabel lain disebut variabel independen atau variabel bebas.. Model persamaannya matematika ditampilkan pada Pers. (1).

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

Di mana y adalah variabel *dependen* dan $x_1, x_2, \ldots, x_n$ merupakan variabel *independen*, a merupakan nilai konstanta, dan b adalah nilai koefisien regresi (Kurniawan, 2016).

Model regresi linear dalam *data mining* menjadi salah satu model yang banyak digunakan. Penggunaan regresi linear dalam *data mining* menggunakan berbagai *datasheet* baik *datasheet* milik sendiri ataupun *datasheet* publik. Salah satu *datasheet* publik yang sering dugunakan dalam pembuatan model *data mining* adalah *datasheet* student_rerformance.csv yang diambil dari https://archive.ics.uci.edu/ml/datasheets.php. *Datasheet* ini berisi data yang diolah dari data siswa dan data keluarga pada sekolah menengah di Portugal. *Datasheet* tersebut menjadi salah satu *datasheet* yang digunakan dalam penelitian pembuatan model *data mining*. Setiyorini & Asmono (2020) menggunakan *Student Performance dataset* untuk pembuatan model klasifikasi terutama dengan menggunakan K-Nearest Neighbour, Ünal (2021) membuat model klasifikasi dengan menggunakan Decision Tree, Random Forest dan Naive Bayes, Deepika & Sathyanarayana (2018) menggunakan Decision Tree, dan Oyedeji et al. (2020) yang melakukan analisis kinerja akademik siswa untuk mengetahui cara-cara meningkatkan kinerja individu siswa.

Model *data mining* untuk memprediksi keberhasilan siswa dengan *datasheet* yang selain *student performance* dan menggunakan *datasheet* yang bersifat privat dilakukan oleh Ofori et al. (2020) yang melakukan identifikasi model *machine learning* untuk melakukan prediksi kinerja siswa dan model *machine learning* yang tepat dalam meningkatkan pembelajaran bagi siswa. Bahri et al. (2022), membuat model yang diharapkan dapat membantu siswa dalam menentukan jurusan yang akan diambil dalam menempuh jenjang pendidikan tinggi. Hasil pengujian, faktor yang paling mempengaruhi kesalahan dalam mengambil jurusan di perguruan tinggi adalah variabel informasi jurusan dan pengujian yang dilakukan dengan menggunakan tiga algoritma, algoritma

Decision Tree merupakan algoritma yang menghasilkan nilai akurasi tertinggi dengan tingkat akurasi tinggi 75,38%. Penelitian lain yang masih terkait *data mining* yang menggunakan data siswa dilakukan oleh Hendrian (2018), Putro et al. (2021), dan Ramadhani & Hendriyani (2021).

Penelitian *data mining* yang menggunakan berbagai *datasheet* dan menggunakan model regresi linear sudah dilakukan dengan menggunakan berbagai aplikasi. Ariesanto & Ekka (2020) melakukan penelitian dengan menerapkan *data mining* dengan regresi linear. Regresi linear digunakan untuk melakukan prediksi harga suatu saham pada perusahaan pelayaran. Evaluasi dilakukan dengan nilai *Root Mean Square Error*. Nilai RMSE menunjukkan angka plus 7,522 dari data aktual harga penutupan saham. Gaol et al. (2019) menggunakan regresi linear berganda untuk melakukan prediksi data dalam ketersediaan buku. Penelitian serupa juga dilakukan oleh Rahayu et al. (2022), Sinaga et al. (2022), dan Siregar (2021).

Pembuatan model *data mining* dapat diimplementasikan dengan berbagai aplikasi baik yang berbasis dengan membuat program dan menggunakan aplikasi Visual Programming. Pengembangan model *data mining* yang menggunakan bahasa pemrograman sudah dilakukan oleh Sholeh et al. (2022). Penelitian dilakukan dengan menggunakan *datasheet* asuransi kesehatan untuk memprediksi biaya asuransi. Pembuatan model menggunakan pemrograman Python dengan memanfaatkan berbagai *library* yang mendukung proses pembuatan model seperti pandas dan sklearn. Penelitian sejenis yang menggunakan pemrograman Python untuk *data mining* dilakukan oleh Kurniatullah & Pramudi (2017), N. et al. (2019), Nishadi (2019), dan Prabha et al. (2020).

Selain menggunakan bahasa pemrograman, pembuatan model *data mining* dapat menggunakan Visual Programming seperti RapidMiner. RapidMiner dapat digunakan untuk melakukan proses analisis pada *data mining*, *text mining*, dan analisis prediksi. RapidMiner menggunakan berbagai cara dan teknik deskriptif serta prediksi dalam pembuatan model yang dapat digunakan dalam pengambilan keputusan (Chisholm, 2013). Penggunaan RapidMiner dalam pembuatan model tidak memerlukan program dan semua yang digunakan dalam pembuatan model sudah tersedia dalam bentuk operator. Pembuatan model menggunakan berbagai operator yang sesuai dan saling dikaitkan dalam membentuk suatu model. Proses pembuatan *data mining* dengan RapidMiner sudah dilakukan beberapa peneliti dengan berbagai *datasheet* dan model serta algoritma yang digunakan. Penelitian tersebut di antaranya dilakukan oleh Chisholm (2013), Prasetyo et al. (2021), dan Sudarsono et al. (2021).

Berdasar latar belakang, tinjauan pustaka, dan studi literatur, proses *data mining* menjadi salah satu cara dalam memberikan informasi dalam bentuk model dalam pengambilan keputusan. Salah satu model yang dapat digunakan dalam pembuatan *data mining* adalah regresi linear. Batasan dalam penelitian ini adalah membuat model regresi linear dengan *datasheet student performance* dan tidak semua atribut digunakan. *Datasheet student performance* merupakan *datasheet* publik, yang terdiri dari 395 data dan 33 atribut. *Datasheet* tersebut dapat diunduh pada laman www.archive.ics.uci.edu/ml/datasheets/student+performance. Pertimbangan penggunaan *datasheet* ini karena memiliki jumlah data dan atribut cukup banyak sehingga dapat dilakukan berbagai pengujian sehingga dapat ditentukan atribut apa saja yang mempengaruhi dalam pembuatan model. Dengan demikian, model yang dihasilkan diharapkan dapat digunakan dalam memprediksi nilai siswa.

2. METODE PENELITIAN

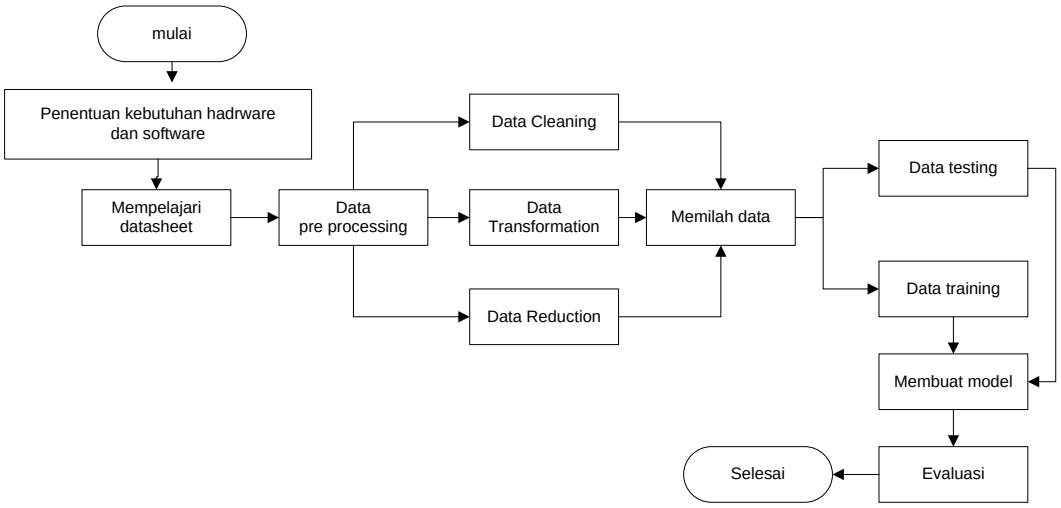
Metodologi penelitian dalam membuat model *data mining* regresi linear, menggunakan metodologi CRISP-DM. Pembuatan model dengan metodologi CRISP-DM terdapat enam tahapan, yaitu *business understanding*, *data understanding, data preparation, modeling, evaluation,* dan *deployment* (Hidayati et al., 2021).

2.1 *Datasheet*

Datasheet yang digunakan adalah StudentPerformance.csv. Sumber *datasheet* ada pada laman https://archive.ics.uci.edu/ml/datasheets/student%2Bperformance. *Datasheet* terdiri dari 395 baris dan 33 atribut.

2.2 Tahapan Penelitian

Tahapan penelitian dengan berdasar pada model CRISP-DM, langkah–langkah yang dilakukan di antaranya mempelajari *datasheet* dan membersihkan *datasheet* seperti memeriksa data kosong, memeriksa batasan nilai dan lainnya. Dalam pembuatan model, *datasheet* dibagi menjadi dua terdiri dari 80% *datasheet* digunakan untuk *data training* dan 20% digunakan untuk *data testing*. Gambar 1 merupakan tahapan dalam pembuatan model *data mining.*



Gambar 1 Tahapan Pembuatan *Data Mining*

3. HASIL DAN PEMBAHASAN

3.1 *Business Understanding*

Langkah awal dalam penelitian adalah mengidentifikasi manfaat dan kegunaan dari model yang dikembangkan. *Datashee*t dilakukan identifikasi keterkaitan antar atribut terutama dengan atribut yang menjadi label. Hasil dari model, model diharapkan dapat digunakan untuk melakukan identifikasi faktor-faktor yang berkontribusi terhadap kegagalan siswa dalam menempuh ujian dan dapat digunakan untuk melakukan prediksi nilai ujian akhir.

3.2 *Data Understanding*

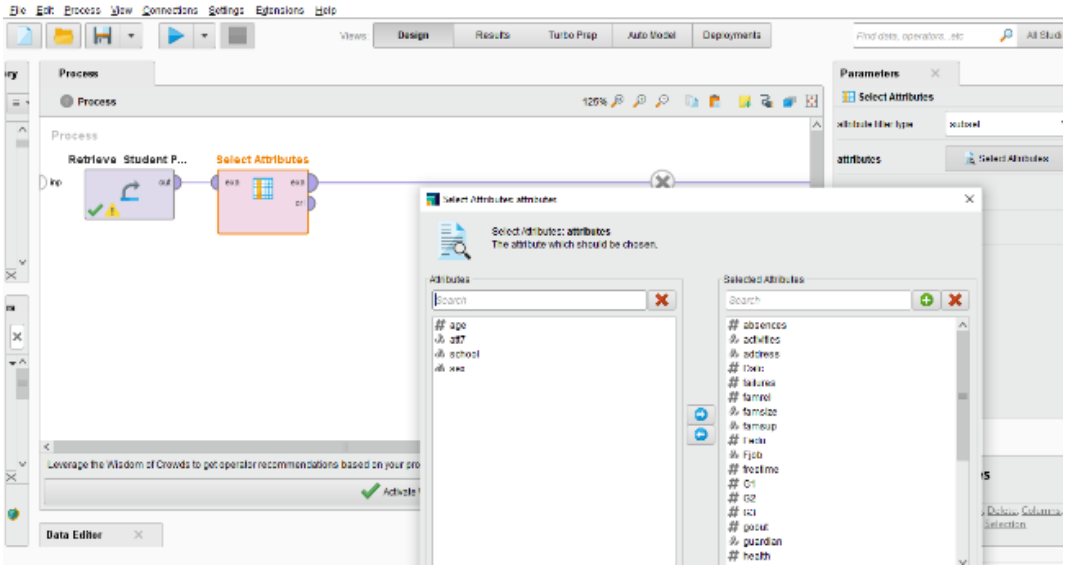
Proses ini dilakukan dengan mengidentifikasi atribut yang ada dalam *datasheet*. Atribut yang ada dilakukan proses pemilahan awal, di antaranya adalah menentukan atau memilih atribut yang tidak mempunyai keterkaitan dalam pembuatan model. Atribut yang tidak mempengaruhi dalam pembuatan model tidak akan digunakan.

3.3 *Data Preparation*

Data preparing sangat diperlukan dan bertujuan untuk mengolah data agar data tidak mengandung kesalahan. Pemeriksaan data yang dilakukan diantaranya memeriksa data kosong, data yang di luar ambang batas dan tipe data yang tidak sesuai. Proses *data preparing* yang dilakukan sebagai berikut.

1) Memilih atribut yang akan digunakan.

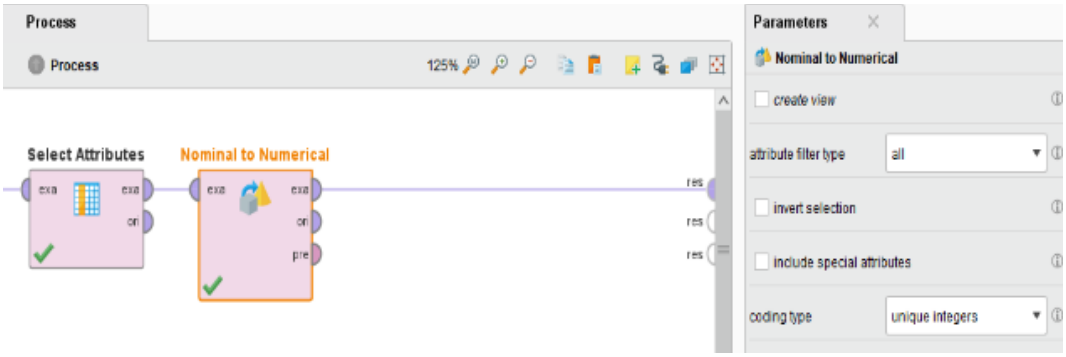
Tidak semua atribut digunakan, sehingga akan dipilih atribut yang mempengaruhi model. Dari 33 atribut yang ada, atribut att, sex, age, dan school tidak dipilih dalam proses pembuatan model. Atribut tersebut tidak saling mempengaruhi dalam pembuatan model. Gambar 2 menunjukkan penggunaan operator *select atribut* dalam proses pemilihan atribut.



Gambar 2 Penggunaan Operator *Select Atribut* dalam Proses Pemilihan Atribut

2) Mengubah tipe data nominal menjadi numerik.

Atribut yang berisi data selain numerik akan diubah menjadi atribut yang berisi data numerik. Atribut sex yang berisi nominal F dan M akan diubah menjadi 0 dan 1, atribut Mjob yang berisi other, services, at_home, teacher, dan health akan diubah menjadi numerik 0-4. Gambar 3 merupakan penggunaan operator *nominal to numerical* dalam proses mengubah tipe nominal menjadi numerik.

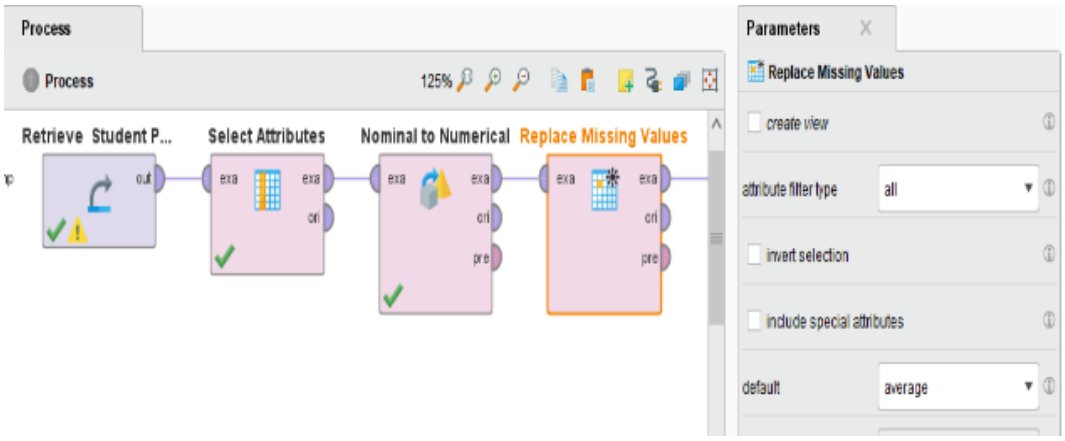


Gambar 3 Penggunaan Operator *Nominal to Numerical*

3) Mengganti data *missing value* dengan data tertentu.

Semua isi *datasheet* yang digunakan harus terisi data atau tidak boleh mengandung data kosong. Agar data tidak mengadung data kosong, data yang berisi nilai kosong akan diganti dengan data rata-rata pada masing-masing atribut. Proses untuk mengganti data kosong dengan data tertentu

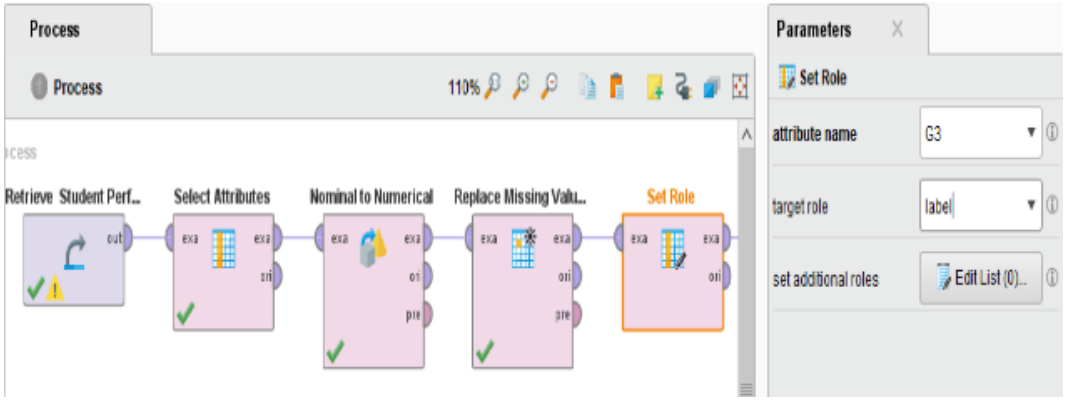
dapat menggunakan operator *replace missing value*. Gambar 4 memperlihatkan proses penggantian data kosong dengan nilai rata-rata.



Gambar 4 Proses Mengganti Data Kosong dengan Nilai Rata-Rata

4) Menentukan atribut yang menjadi label.

Label yang digunakan dalam pembuatan model adalah G3. Gambar 5 menunjukkan penggunaan operator *set role* untuk menentukan atribut G3 menjadi label.



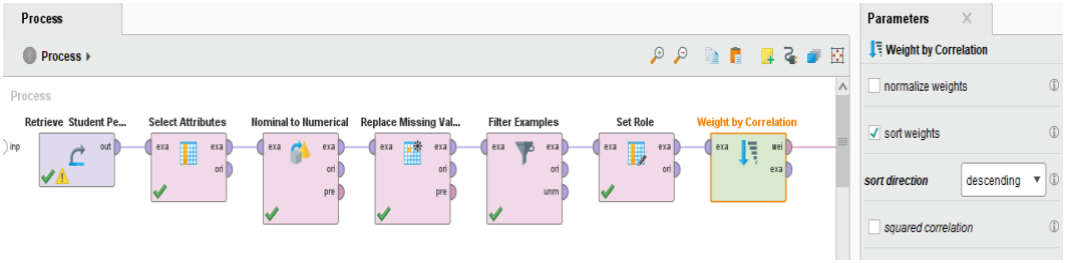
Gambar 5 Penggunaan Operator *Set Role* untuk Menentukan Label

3.4 Pembuatan model

Tahapan berikutnya adalah membuat model regresi linear dengan atribut yang sudah ditentukan dengan tahapan pemuatan model sebagai berikut.

1) Menampilkan bobot hasil matriks korelasi.

Matriks korelasi merupakan matriks yang memuat koefisien korelasi dari semua atribut yang digunakan. Hasil dari matriks dapat digunakan untuk memperoleh nilai kedekatan hubungan antar atribut. Gambar 6 menunjukkan penggunaan operator *weight by Correlation* untuk mengetahui keterkaitan antar atribut dan Gambar 7 memperlihatkan hasil pembobotan keterkaitan antar atribut terutama dengan atribut label yaitu G3.



Gambar 6 Penggunaan Operator *Weight by Correlation*

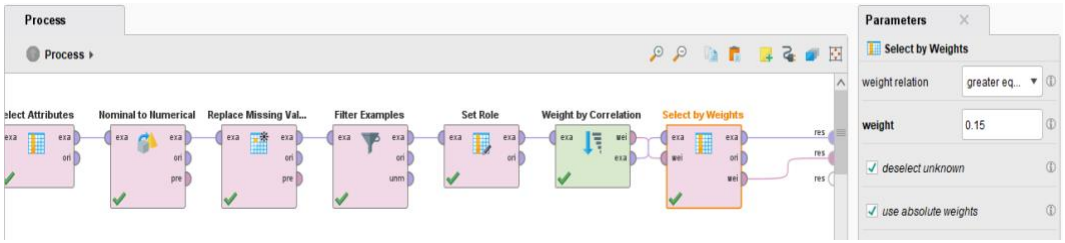
attribute	weight	attribute	weight
G2	0.966	traveltime	0.102
G1	0.892	Mjob	0.084
failures	0.294	health	0.082
schoolsup	0.238	famsup	0.067
absences	0.213	reason	0.061
Walc	0.190	activities	0.059
Medu	0.188	romantic	0.050
goout	0.177	famsize	0.040
Fedu	0.163	famrel	0.038
Dalc	0.141	guardian	0.035
address	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 7 Hasil Pembobotan Keterkaitan Atribut Terutama dengan Atribut Label G3

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

Hasil pembobotan pada Gambar 7, ditentukan batas nilai bobot yang digunakan. Dalam penelitian ini, batas pembobotan yang akan digunakan adalah pembobotan di atas 0,15. Batas ini dipilih dengan pertimbangan batas 0,15 masih mendekati batas korelasi cukup (>0,25 – 0,5) dan akan dilakukan pengujian, apakah atribut yang dibawah batas cukup mempengaruhi model. Atribut yang digunakan dalam membuat model adalah G2 (nilai ujian kedua), G1 (nilai ujian pertama), failures (berapa kali pernah tinggal kelas), schoolsup (tambahan pelajaran di luar sekolah) , absences (kehadiran di kelas), Walc (konsumsi alkohol), Medu (tingkat pendidikan orang tua (ibu)), goout (berapa banyak bermain/keluar rumah dengan teman sebaya), dan Fedu (tingkat pendidikan orang tua (ayah)).

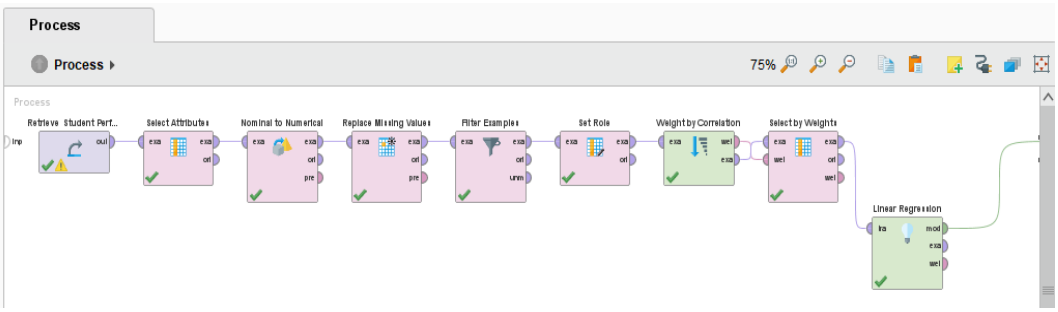
Proses untuk pemilihan bobot menggunakan operator *select by weight dengan* parameter pemilihan bobot di atas 0,15. Gambar 8 menunjukkan proses penggunaan operator *select by weight*.



Gambar 8 Proses Pemberian Bobot di Atas 0,15

3) Proses pembuatan model.

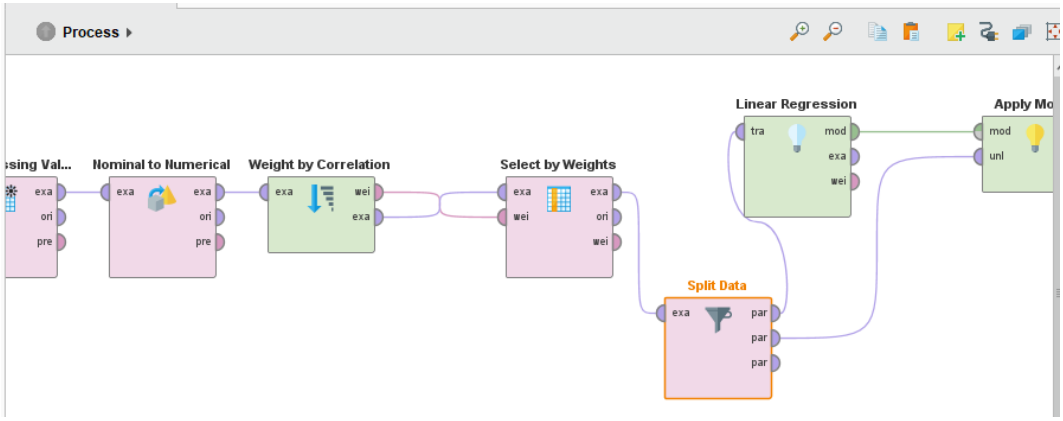
Berdasar pada pembobotan, pembuatan model *linear regresi* dapat dilakukan. Gambar 9 merupakan penggunaan operator *linear regresi* dalam proses pembuatan model.



Gambar 9 Penggunaan Operator *Linear Regresi*

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

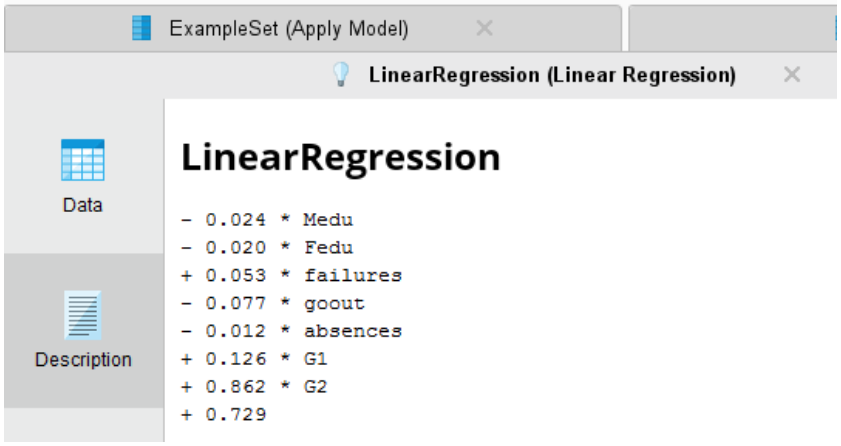
Model yang dihasilkan pada Gambar 9, menggunakan semua data yang ada di *datasheet* dan belum melakukan pengujian. Proses pembuatan model akan menggunakan *data training* dan proses evaluasi menggunakan *data testing*. Pembagian data dilakukan dengan membagi *data training* sebanyak 80% dan *data testing* sebanyak 20%. Proses pembagian data menggunakan operator *split data* dan proses untuk melakukan pengujian menggunakan operator *apply data*. Gambar 10 merupakan proses pembuatan data dengan *data training* dan pengujian dengan *data testing*.



Gambar 10 Proses Pembagian Data untuk *Data Training* dan *Data Testing*

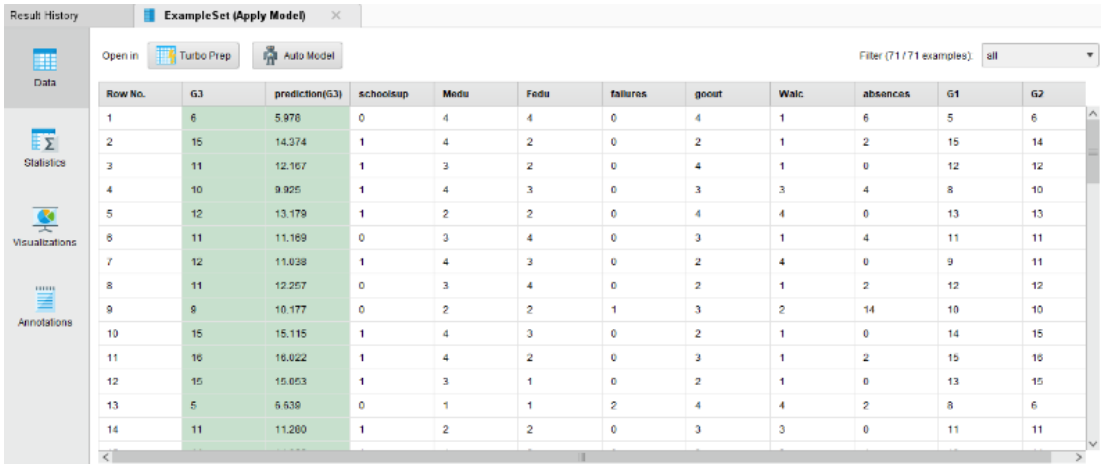
Hasil dari proses pembuatan model menghasilkan model linear regresi yang hasilnya ditunjukkan pada Pers. (2) dan Gambar 11.

$$y = 0{,}729 - (0{,}024 \times Medu) - (0{,}020 \times Fedu) + (0{,}053 \times failures) - (0{,}077 \times goout) - (0{,}012 \times absences) + (0{,}126 \times G1) + (0{,}862 \times G2) \quad (2)$$



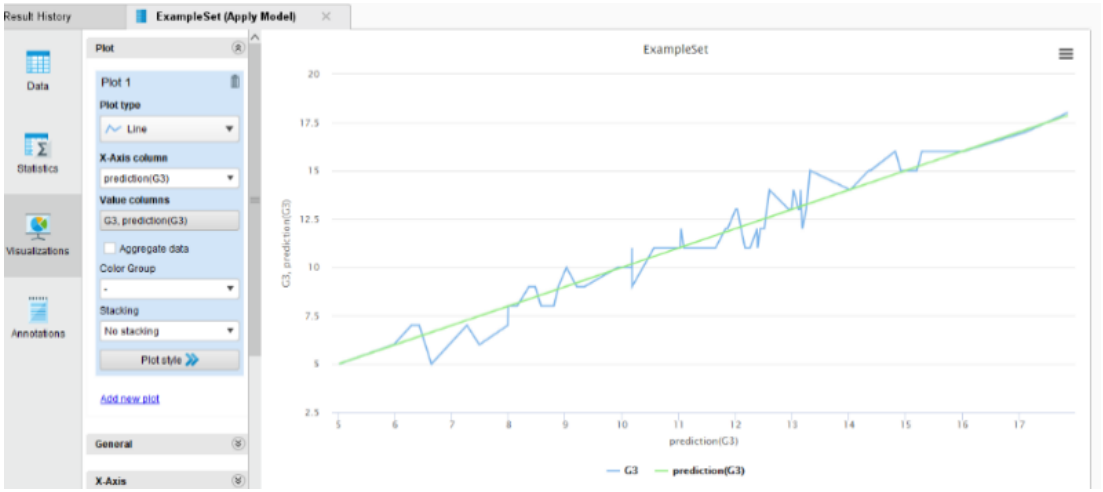
Gambar 11 Hasil Model Linear Regresi

Model yang dihasilkan pada Gambar 11, dilakukan pengujian dengan menggunakan *data training*. Hasil pengujian model dapat dilihat dengan melakukan perbandingan antara nilai G3 dengan nilai prediksi. Gambar 12 memperlihatkan hasil pengujian model dengan melihat perbandingan antara nilai G3 dengan nilai hasil prediksi. Sesuai hasil pada Gambar 12, pada baris 1, nilai G3 asli sebesar 6 dan hasil nilai prediksi G3 sebesar 5,978. Hasil ini menunjukkan ada selisih, hasil prediksi masih di bawah nilai G3 asli. Nilai prediksi baris 14, nilai G3 asli sebesar 11 dan hasil nilai prediksi G3 sebesar 11,280. Hasil ini menunjukkan ada selisih, hasil prediksi di atas dari nilai G3 asli. Hasil nilai prediksi dengan nilai G3 (nilai asli) dapat disajikan dalam bentuk grafik. Nilai prediksi ditunjukkan dengan garis lurus dan nilai asi ada di sekitar garis lurus hasil prediksi. Gambar 13 merupakan penyajian perbandingan antara nilai G3 dengan nilai hasil prediksi dalam bentuk grafik.



Row No.	G3	prediction(G3)	schoolsup	Medu	Fedu	failures	goout	Walc	absences	G1	G2
1	6	5.978	0	4	4	0	4	1	6	5	6
2	15	14.374	1	4	2	0	2	1	2	15	14
3	11	12.167	1	3	2	0	4	1	0	12	12
4	10	9.925	1	4	3	0	3	3	4	8	10
5	12	13.179	1	2	2	0	4	4	0	13	13
6	11	11.169	0	3	4	0	3	1	4	11	11
7	12	11.038	1	4	3	0	2	4	0	9	11
8	11	12.257	0	3	4	0	2	1	2	12	12
9	9	10.177	0	2	2	1	3	2	14	10	10
10	15	15.115	1	4	3	0	2	1	0	14	15
11	16	16.022	1	4	2	0	3	1	2	15	16
12	15	15.053	1	3	1	0	2	1	0	13	15
13	5	6.539	0	1	1	2	4	4	2	8	6
14	11	11.280	1	2	2	0	3	3	0	11	11

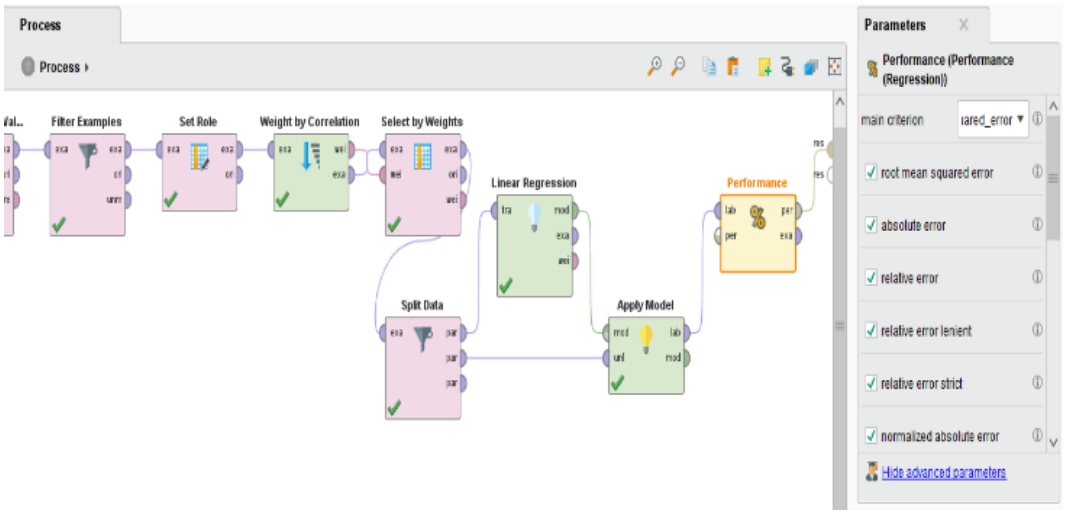
Gambar 12 Perbandingan Nilai G3 dengan Nilai Hasil Prediksi



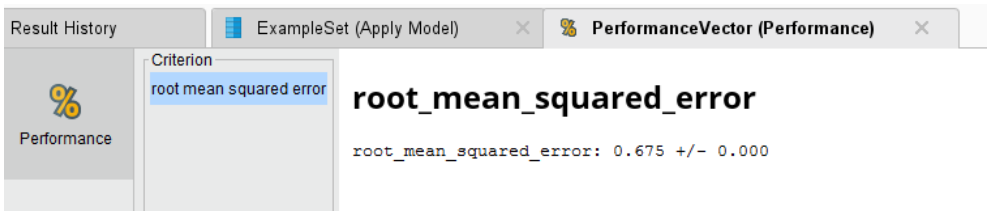
Gambar 13 Grafik Perbandingan Nilai G3 dengan Nilai Hasil Prediksi

3.5 Evaluasi Model

Model yang sudah selesai dibangun, dilakukan proses evaluasi. Proses evaluasi dilakukan untuk melihat *performance*. Proses evaluasi dilakukan dengan menggunakan *data testing* yang sudah ditentukan sebanyak 20% dari *datasheet*. Proses pengujian menggunakan operator *performance*. Gambar 14 merupakan proses pengujian dengan menggunakan operator *performance* dan Gambar 15 adalah hasil luaran dari *performance.*



Gambar 14 Proses Pengujian dengan Melihat *Performance*



Gambar 15 Hasil Luaran *Performance*

Berdasar pada Gambar 15, hasil luaran *performance* dengan menggunakan evaluasi *root mean squared error* menunjukkan nilai 0,675. Hasil RMSE 0,675 dapat disimpulkan hasil model mempunyai nilai kesalahan yang kecil.

4. KESIMPULAN

Data mining dapat digunakan dalam membantu proses pengambilan kebijakan. Salah satu model yang dapat digunakan adalah dengan melakukan prediksi dengan regresi linear. Hasil penelitian dari 33 atribut yang ada pada *datasheet* student_performance.csv, tidak semua atribut mempunyai berpengaruh yang signifikan pada atribut yang menjadi prediktor. Atribut yang sangat kuat mempengaruhi adalah G2 (nilai ujian ke 2) dan G1 (nilai ujian ke 1). Bobot dari G2 sebesar 0,966 dan G1 sebesar 0,892. Atribut lain dengan bobot di atas 0,2 adalah failures (berapa kali pernah tinggal kelas), schoolsup (tambahan pelajaran di luar sekolah) dan absences (kehadiran di kelas) sedangkan atribut yang lain di bawah 0,2. Bobot di bawah 0,2 ini tidak mempunyai pengaruh pada prediktor. Hasil model adalah $y = 0{,}729 - (0{,}024 \times Medu) - (0{,}020 \times Fedu) + (0{,}053 \times failures) - (0{,}077 \times goout) - (0{,}012 \times absences) + (0{,}126 \times G1) + (0{,}862 \times G2)$. Hasil pengujian nilai RMSE adalah 0,675. Semakin kecil nilai RMSE berarti nilai yang diprediksi dekat dengan nilai yang diamati atau observasi. Berdasar hasil RMSE tersebut, hasil model mempunyai nilai kesalahan yang kecil dan model yang dihasilkan dapat direkomendasikan untuk dapat digunakan dalam melakukan prediksi nilai siswa.

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=AtcCEAAAQBAJ&redir_esc=y

Ariesanto, A., & Ekka, P. (2020). Data Mining Menggunakan Regresi Linear untuk Prediksi Harga Saham Perusahaan Pelayaran. *Jurnal Aplikasi Pelayaran Dan Kepelabuhanan*, *10*(2), 120. https://doi.org/10.30649/japk.v10i2.83

Bahri, S., Itb, A., & Dahlan, 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

Chisholm, A. (2013). *Exploring Data with RapidMiner* (Vol. 1). Packt Publishing. 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. http://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 MINING DENGAN 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-dasar-dan-penerapannya-dengan-r/

Nishadi, A. S. T. (2019). Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab. *International Journal 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-Techn/article/view/480

Oyedeji, 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://sersc.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.v7i1.2021.8-17

Putro, M. F., Prayitno, E., Siregar, J., & Muharrom, M. (2021). PENERAPAN DATA MINING DENGAN NAÏVE BAYES UNTUK KLASIFIKASI SISWA SEKOLAH MENENGAH ATAS DALAM PENENTUAN 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 for Motorcycle Sales Estimation. *JOMLAI: 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

Setiyorini, T., & Asmono, R. T. (2020). IMPLEMENTATION OF GAIN RATIO AND K-NEAREST NEIGHBOR FOR CLASSIFICATION OF STUDENT PERFORMANCE. *Jurnal Pilar Nusa Mandiri*, *16*(1), 19–24. https://doi.org/10.33480/pilar.v16i1.813

Sholeh, M., Suraya, S., & Andayati, D. (2022). Machine Linear untuk Analisis Regresi Linier Biaya Asuransi Kesehatan dengan Menggunakan Python Jupyter Notebook. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)*, *8*(1), 20–27. https://doi.org/10.26418/JP.V8I1.48822

Sinaga, W. A. L., Sumarno, S., & Sari, I. P. (2022). The Application of Multiple Linear Regression Method for Population Estimation Gunung Malela District. *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, *1*(1), 55–64. https://doi.org/10.55123/jomlai.v1i1.143

Siregar, A. Z. (2021). Implementasi Metode Regresi Linier Berganda Dalam Estimasi Tingkat Pendaftaran Mahasiswa Baru. *Kesatria : Jurnal Penerapan Sistem Informasi (Komputer Dan Manajemen)*, *2*(3), 133–137. https://doi.org/10.30645/KESATRIA.V2I3.73

Sudarsono, B. G., Leo, M. I., Santoso, A., & Hendrawan, F. (2021). ANALISIS DATA MINING DATA NETFLIX MENGGUNAKAN APLIKASI RAPID MINER. *JBASE - Journal of Business and Audit Information Systems*, *4*(1), 13–21. https://doi.org/10.30813/jbase.v4i1.2729

Ünal, F. (2021). Data Mining for Student Performance Prediction in Education. In *Data Mining - Methods, Applications and Systems*. IntechOpen. https://doi.org/10.5772/intechopen.91449