METODE PENILAIAN KEKUATAN GEMPA MENGGUNAKAN MODEL FEATURE SELECTION M5-PRIME DAN LINEAR REGRESSION

Oman Somantri[1], Ratih Hafsarah Maharrani[2]

[1,2] Jurusan Teknik Informatika, Politeknik Negeri Cilacap
[1]oman.somantri@pnc.ac.id, [2] ratih.hafsarah@pnc.ac.id

Abstrak

Pemetaan suatu wilayah khususnya daerah rawan bencana bagi pemangku kepentingan sangat penting sekali karena hal ini akan memberikan pengaruh terhadap kebijakan yang nantinya akan ditetapkan, terlebih dalam upaya penanggulangan bencana alam gempa bumi. Memprediksi kekuatan gempa menjadi permasalahan yang sampai saat ini belum dapat dipastikan dan diprediksikan sehingga yang bisa dilakukan adalah sebatas memprediksi dari dari kejadian-kejadian sebelumnya. Sebuah pendukung keputusan yang tepat dalam upaya mengatasi dan menanggulangi persiapan menghadapi kejadian bencana gempa yang tidak diinginkan menjadi sebuah kebutuhan dan sangat penting untuk didapatkan bagi setiap pemangku keputusan disuatu daerah. Pada penelitian ini mengusulkan sebuah algoritma linear regression untuk diterapkan dan digunakan sebagai model dalam prediksi gempa, selain itu sebagai upaya dalam peningkatan akurasi yang dihasilkan maka diusulkan model feature selection dengan menggunakan algoritma M5-Prime agar terjadinya peningkatan nilai root mean square error (RMSE) terbaik pada model yang dihasilkan. Tahapan penelitian yang dilakukan pada eksperimen ini adalah dengan melakukan proses praprocessing data, menerapkan model yang diusulkan, pencarian nilai parameter terbaik model yang diusulkan, proses validasi model, dan tahapan terakhir yang dilakukan adalah evaluasi model. Hasil penelitian memperlihatkan algoritma linear regression berbasis feature selection M5-Prime mampu dijadikan sebagai model prediksi gempa dengan nilai RMSE terbaik sebesar 0,707. Berdasarkan penelitian yang telah dilakukan maka sebuah model terbaik telah dihasilkan untuk dapat digunakan sebagai pendukung keputusan dalam penilaian kekuatan gempa sebagai upaya mitigasi penanggulangan bencana gempa bumi.

Kata kunci : gempa, feature selection, M5-Prime, linear regression

1. Pendahuluan

Pendataan dan pemetaan wilayah sebuah daerah memiliki peranan penting yang tidak bisa diabaikan karena merupakan upaya strategis suatu daerah untuk mengetahui keberadaan wilayahnya. Potensi-potensi yang dimiliki oleh setiap daerah memiliki perbedaan yang tidak sama antara satu daerah dengan lainnya. Sumber daya alam dan keadaan topografi suatu daerah menjadi sumber data yang akan digunakan oleh pihak pemangku kebijakan dalam hal ini pemerintah daerah dalam upaya memutuskan sebuah kebijakan tertentu. Potensi bencana dimana salah satunya adalan gempa bumi menjadi prioritas sebuah daerah untuk membuat skema kebijakan dalam mitigasi bencana daerah.

Memastikan dan menilai sebuah kekuatan gempa yang akan terjadi di sebuah daerah menjadi sangat penting karena ini akan memberikan sebuah data untuk memutuskan kebijakan sebagai upaya penanggulangan bencana daerah. Data menunjukan pada tahun 2022 selama bulan Januari 2022 terjadi kejadian gempa sebanyak 729 kali di Indonesia, hal ini menunjukan potensi gempa di Indonesia sangat tinggi. Upaya-upaya yang dilakukan salah satunya adalah dengan melakukan penilaian prediksi kekuatan gempa bumi sehingga dapat diketahui potensi kedepannya apabila terjadi kejadian gempa, selain itu pihak pemerintah sudah mempunyai persiapan yang lebih matang dalam mempersiapkan segala sesuatunya karena berkaitan dengan masyarakat di sekitar lokasi terjadinya gempa.

Linear Regression (LR) saat ini dengan kemampuan yang dimilikinya digunakan untuk model prediksi data khususnya berbentuk *time series* seperti yang telah dilakukan oleh beberapa peneliti terdahulu (Afrifa-Yamoah et al., 2020; Benjamin & Konstantinos, 2002; Ewusie et al., 2020; Shaikh et al., 2021). Algoritma yang telah banyak digunakan untuk berbagai kasus data *time series* seperti digunakan untuk beberapa prediksi seperti prediksi saham, data cuaca, peramalan jumlah demam berdarah (Khotimah & Sari Rochman, 2021), kedatangan wisata turis (Yijun & Zhang, 2012), prediksi *smart shoping* (Nastiti et al., 2019), prediksi penyakit covid-19 (Mandayam et al., 2020), dan

lainnya baik itu model yang diterapkan berdiri sendiri maupun berbasis *hybrid*.

Berkaitan dengan bencana, beberapa penelitian yang dilakukan sebelumnya telah menerapkan algoritma *Linear Regression* dan beberapa metode lainnya untuk memprediksi data bencana. Penelitian dilakukan dengan menerapkan *data mining* untuk menganlisis data bencana BNPB dengan menggunakan algoritma *liner regression* dan *k-Means* (Muhamad Iqbal & Prihandoko, 2017). Pada penelitian lainnya melakukan penelitian untuk mengelompokan dampak gema bumi dan kerusakan pada wilayah berpotensi gempa di provinsi sumatera Barat (Sugiyarto et al., 2021).

Penelitian yang dilakukan pada artikel ini adalah mengusulkan sebuah model baru yang digunakan untuk memprediksi kekuatan gempa bumi. Penelitian sejenis sudah dilakukan dengan mengusulkan algoritma *Support Vector Machine* (SVM) yang digunakan sebagai model prediksi gempa, pada penelitian ini dihasilkan model menghasilkan nilai RMSE sebesar 0,712 (Somantri et al., 2022). Penelitian lain menggunakan metode *Neural Network* (NN) untuk menghasilkan model prediksi gempa, NN pada penelitian ini telah dilakukan optimasi dengan menggunakan algoritma genetika sehingga menghasilkan nilai RMSE sebesar 0,708 (Somantri, 2021). Berdasarkan hasil dari penelitian yang telah dilakukan sebelumnya memerlukan upaya peningkatan akurasi dengan menggunakan algoritma prediksi lainnya yang mempunyai hasil tingkat akurasi lebih tinggi.

Tujuan dari penelitian ini adalah menerapkan algoritma *linear regression* untuk mendapatkan model dalam melakukan penilaian prediksi kekuatan gempa bumi. Kontribusi utama pada penelitian ini adalah untuk optimalisasi model yang sudah diperoleh kemudian diterapkan metode *feature selection* dengan menggunakan metode *M5-Prime* untuk optimalisasi model sehingga mendapatkan peningkatan akurasi prediksi. Algoritma *M5-Prime* merupakan sebuah algoritma pengembangan dari algoritma M5, dan merupakan sebuah kombinasi dari pohon keputusan dan digunakan sebagai analisis *regresi linier* sederhana (Yildiz & Yandi, 2021). *Feature selection* digunakan untuk mendapatkan atribut terbaik sehingga dapat berpengaruh terhadap tingkat akurasi model yang diusulkan.

2. Metode

2.1 Tools dan Data

Proses dalam menemukan model terbaik pada penelitian ini dilakukan dengan metode eksperimen dengan menggunakan Rapidminer studio memakai komputer processor i7 dan memori RAM 8Gb. *Dataset* pada penelitian ini adalah menggunakan data yang dikeluarkan BMKG Indonesia yang berisi data kejadian gempa bumi di Indonesia. Jenis data yang digunakan adalah berbentuk *timeseries univariate* sebanyak 272 *record* data pada bulan Januari 2021. *Dataset* penelitian yang digunakan diperlihatkan pada Tabel 1.

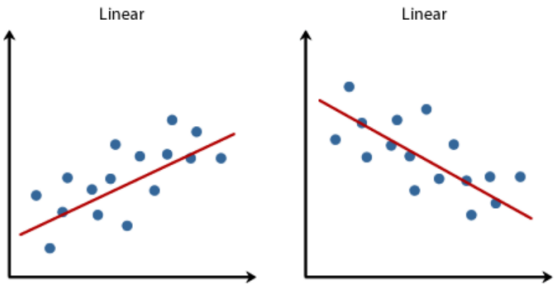
Tabel 1. Data penelitian

No.	Tanggal (GMT)	Magnitudo (SR)
1	2021-01-01 23:24:25	3,06
2	2021-01-01 20:35:22	2,55
3	2021-01-01 11:32:18	3,77
4	2021-01-01 11:13:23	4,21
5	2021-01-01 08:57:43	2,77
…	…	…
…	…	…
….	….	….
270	2021-01-11 01:32:30	3,57
271	2021-01-11 01:20:38	5,55
272	2021-01-11 01:13:38	3,28

Proses tahapan penelitian dilakukan mulai dari tahapan pencarian data penelitian, *praprocessing* data, analisis data, validasi model dan evaluasi model.

2.2 Linear Regression

Metode *linear regression* (LR) merupakan sebuah metode yang digunakan dan diterapkan untuk memprediksi dan atau mengestimasi hubungan antara dua variabel (Jung & Choi, 2021). Metode LR mampu menghubungkan antara dua variabel dimana diantaranya adalah variabel yang *dependent* dan *independent* melalui garis yang sesuai diantara garis lurus pada kurva (Permai & Tanty, 2018), penggambaran kurva seperti pada Gambar 1 (Laerd Statistics, n.d.).



Gambar 1. Kurva linear regression

Persamaan untuk mendapatkan nilai variabel *dependent* dan variabel *independent* dapat dihitung dengan menggunakan persamaan (1) dan persamaan (2) (Draper & Harry, 2014).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \quad (1)$$

$$Y = X\beta + \varepsilon \quad (2)$$

dimana,
Y = variabel dependen
X = variabel independen
β = vektor parameter model regression
ε = vektor error

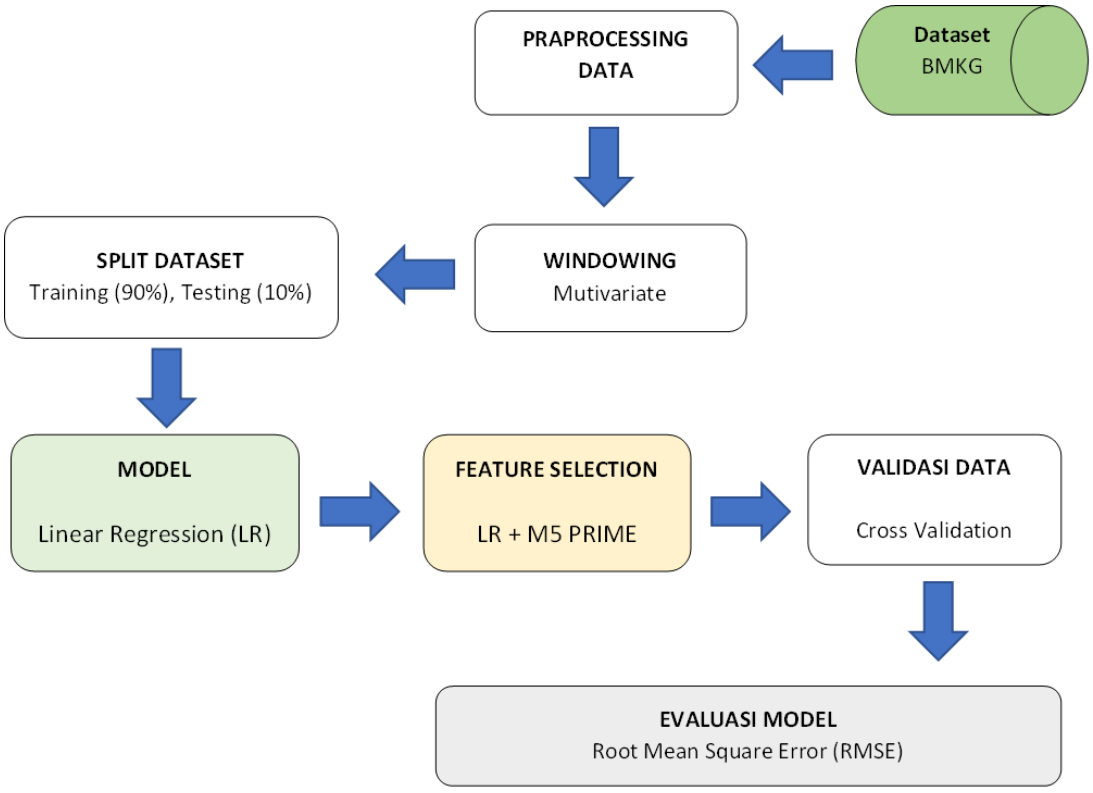
2.3 Framework Penelitian yang diusulkan

Pada penelitian ini tahapan penelitian dilakukan setelah mendapatkan *dataset* yang diinginkan tahapan selanjutnya adalah dilakuakn tahapan *praprocessing* data. Tahapan *praprocessing* data dilakukan dengan melakukan process pencarian missing data, dan pemilihan atribut yang menjadi target untuk analisis data, pada tahapan ini terdapat dua atribut utama yang digunakan yaitu "tanggal (GMT)" dan "Magnitudo (SR)". Proses tahapan selanjutnya adalah melakukan *split* antara data *training* dan data *testing* dengan rasio 90% untuk data *training* dan 10% untuk data *testing*.

Tahapan penerapan algoritma *Linear Regression* dilakukan setelah *dataset* yang diperoleh dilakukan terlebih dahulu proses *windowing*. *Windowing* adalah proses dimana dilakukanya perubahan atribut data *univariate* menjadi *multivariate*, dimana pengaturan nilai parameter *windowing* diatur tersendiri pada proses eksperimen untuk mendapatkan model dengan tingkat akurasi prediksi yang terbaik. Tahapan perapan penggunaan *feature selection* diberikan setelah didapatkannya model dengan menggunakan LR yang kemudian dioptimalisasi melalui proses *feature selection* ini. Validasi data pada penelitian ini menggunakan cross validation, dan tahapan terakhir penelitian adalah melakukan evaluasi model. *Framework* tahapan penelitian yang dilakukan diperlihatkan pada Gambar 2.

Berdasarkan pada Gambar 2, pada tahapan evaluasi model untuk indikator penilaiannya diukur dengan menggunakan *Root Mean Square Error* (RMSE) yaitu dimana model yang mempunyai nilai RMSE paling kecil adalah model yang terbaik (Chai & Draxler, 2014) (Chicco et al., 2021)(Ćalasan et al., 2020). Pada evaluasi ini dilakukan komparasi model antara model yang menerapkan LR klasik dibandingkan dengan LR yang telah dioptimasi dengan *feature selection* (LR + M5-Prime).

Setiap data yang digunakan pada model yang dicari merupakan *dataset* yang telah terlebih dahulu dilakukan *filtering* atribut serta telah dilakukan proses penentuan nilai parameter *windows* pada proses *praprocessing* data dilakukan sebelum pembagian data *training* dan data *testing*.



Gambar 2. Framework metode yang diusulkan

3. Hasil dan Pembahasan

3.1 Penerapan Model LR

Eksperimen yang dilakukan pada tahapan ini adalah menerapkan algoritma LR untuk prediksi gempa dengan terlebih dahulu menetapkan nilai parameter *windows*. Pada penentuan penetapan parameter *windows* ini dilakukan dengan cara tanpa dilakukan berdasarkan ketetapan tertentu, akan tetapi model yang diharapkan dapat menghasilkan model prediksi dengan RMSE yang terbaik.

Ekperimen yang dilakukan dengan merapapkan parameter *cross validation* ditentukan k-*fold*=10 menghasilkan model dengan beberapa nilai RMSE diperlihatan pada Tabel 1. Penelitian selanjutnya dilakukan dengan menentukan *fold*=5, dari hasil eksperimen menghasilkan beberapa model dengan tingkat prediksi akurasi RMSE yang berbeda. Hasil eksperimen diperlihatkan pada Tabel 2.

Pada Tabel 1 diperlihatkan bahwa nilai RMSE terbaik adalah sebesar 0,712 dimana model tersebut menggunakan *sample linear* dan *windows*=2. Berbeda dengan apa yang diperlihatkan pada Tabel 3 menunjukan untuk model terbaik adalah menunjukan nilai RMSE sebesar 0,732.

Tabel. 1 Model Linear Regression dengan fold=10

eksperimen	sampling	RMSE	windows
1	Linear	0.717	4
2	Shuffled	0.730	4
3	Linear	0.717	3
4	Shuffled	0.723	3
5	Linear	0.712	2
6	Shuffled	0.723	2

Tabel. 2 Model LR menggunakan fold=5

eksperimen	sampling	RMSE	windows
1	Linear	0.732	4
2	Linear	0.728	3
3	Linear	0.723	2
4	Shuffled	0.737	4
5	Shuffled	0.729	3
6	Shuffled	0.728	2

Perbedaan nilai RMSE yang didapatkan pada eksperimen Tabel 1 dan Tabel 2 ini ditentukan oleh beberapa parameter yang mempengaruhinya seperti metode *sampling*, jumlah *windows*, dan jumlah *fold* yang digunakan. Penentuan nilai parameter ini ditentukan berdasarkan intuisi sehingga tidak selamanya nilai parameter tersebut sesuai dengan yang diharapkan karena nilai RMSE yang dihasilkan masih belum yang terbaik. Hasil yang didapatkan dari model yang didapatkan masih memerlukan upaya peningkatan akurasi dan hal ini akan dilakukan dengan menggunakan metode *feature selection*. Upaya yang dilakukan untuk peningkatan akurasi memerlukan algoritma yang tepat, pada penelitian ini algoritma *M5-Prime* akan diterapkan.

3.2 Model LR dan Feature Selection

Pada tahapan optimalisasi model yang sudah didapatkan pada tahapan ini menerapkan metode *M5-Prime* yang kemudian menghasilkan nilai RMSE yang lebih baik. Hasil optimalisasi model diperlihatkan pada Tabel 3 dan Tabel 4. Hasil ekperimen terhadap model yang dioptimasi tersebut terjadi perubahan nilai RMSE yang lebih baik dibandingan dengan nilai RMSE sebelumnya. Terlihat untuk nilai prediksi RMSE terbaik adalah sebesar 0,707, hasil ini merupakan nilai terkecil yang dihasilkan dan merupakan model yang diusulkan pada penelitian ini.

Tabel 3. Model LR + M5-Prime dengan fold=10

No	sampling	RMSE	windows
1	Linear	0.705	4
2	Shuffled	0.717	4
3	Linear	0.707	3
4	Shuffled	0.720	3
5	Linear	0.704	2
6	Shuffled	0.717	2

Tabel 4. Model LR + M5-Prime dengan fold=5

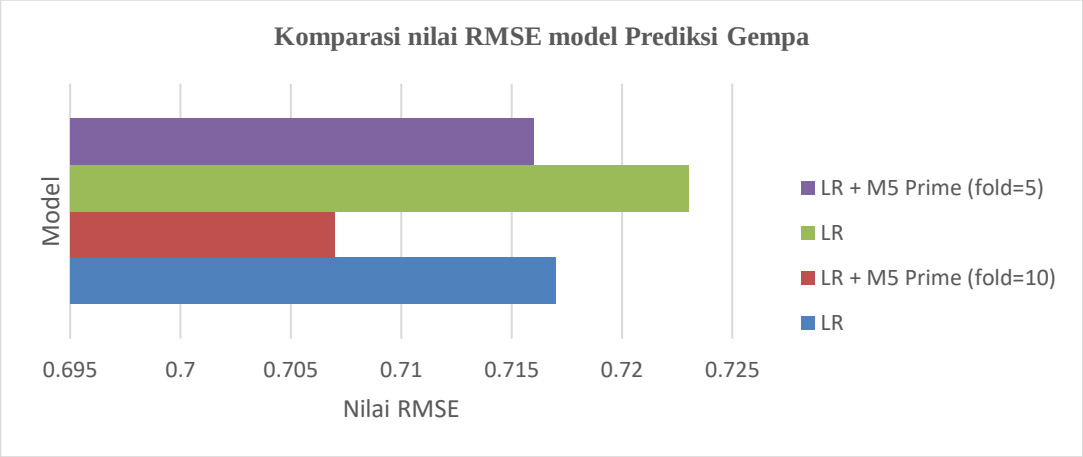
No	sampling	RMSE	windows
1	Linear	0.716	4
2	Linear	0.718	3
3	Linear	0.718	2
4	Shuffled	0.722	4
5	Shuffled	0.726	3
6	Shuffled	0.721	2

3.3 Evaluasi Model

Berdasarkan penelitian yang telah dilakukan, model yang diusulkan memberikan perbedaan khususnya tingkat akurasi prediksi yang dihasilkan. Optimalisasi model dengan mengusulkan *feature selection* dengan menggunakan M5-Prime memiliki nilai RMSE yang lebih baik dibandingkan dengan model LR biasa. Pada Tabel 5 memberikan perbedaan dari model-model yang telah didapatkan.

Tabel 5. Komparasi nilai RMSE model

fold	Metode	sampling	RMSE	windows
10	LR	Linear	0.717	4
10	LR + M5-Prime (diusulkan)	Linear	0.707	3
5	LR	Linear	0.723	2
5	LR + M5-Prime	Linear	0.716	4



Gambar 3. Grafik komparasi nilai RMSE model LR yang diusulkan

Pada penelitian ini terdapat beberapa faktor yang mempengaruhi hasil eksperimen yang dilakukan diantaranya adalah nilai dan jenis parameter dari model yang ditentukan seperti jenis penentuan *sampling* pada saat eksperimen dan jumlah nilai *window*. Selain itu dalam model ini jumlah *dataset* yang digunakan sangat berpengaruh juga terhadap nilai RMSE yang dihasilkan.

4. Kesimpulan dan Saran

Prediksi sebuah kekuatan gempa bumi yang saat hampir sulit untuk dapat diprediksikan melalui model *data mining* dengan menggunakan *linear regression* dapat memberikan gambaran prediksi gempa kedepannya sehingga hal ini memberikan sebuah pendukung keputusan bagi pemegang kebijakan dalam hal ini pemerintah untuk dapat melakukan upaya-upaya pencegahan meminimalisisr dampak dari bencana gempa tersebut. Algoritma *linier regression* dengan optimalisasinya dengan menggunakan *feature selection* masih memiliki kekurangan khususnya dalam hal tingkat akurasi prediksi sehingga memerlukan eksperimen lanjutan dalam pencarian model terbaik. Untuk penelitian selanjutnya diusulkan untuk menggunakan algoritma *machine learning* lain yang disertai dengan optimasi model sehingga niai RMSE yang dihasilkan semakin kecil dan semakin lebih baik.

Daftar Pustaka:

Afrifa-Yamoah, E., Mueller, U. A., Taylor, S. M., & Fisher, A. J. (2020). Missing data imputation of high-resolution temporal climate time series data. *Meteorological Applications*, *27*(1). https://doi.org/10.1002/met.1873

Benjamin, K., & Konstantinos, F. (2002). *Regression Models for Time Series Analysis*. Wiley.

Ćalasan, M., Abdel Aleem, S. H. E., & Zobaa, A. F. (2020). On the root mean square error (RMSE) calculation for parameter estimation of photovoltaic models: A novel exact analytical solution based on Lambert W function. *Energy Conversion and Management*, *210*, 112716. https://doi.org/10.1016/j.enconman.2020.112716

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623. https://doi.org/10.7717/peerj-cs.623

Draper, N. R., & Harry, S. (2014). *Applied Regression Analysis, 3rd Edition*. John Willey

& Sons, Inc.

Ewusie, J. E., Soobiah, C., Blondal, E., Beyene, J., Thabane, L., & Hamid, J. S. (2020). Methods, Applications and Challenges in the Analysis of Interrupted Time Series Data: A Scoping Review. *Journal of Multidisciplinary Healthcare*, *Volume 13*, 411–423. https://doi.org/10.2147/JMDH.S241085

Jung, D., & Choi, Y. (2021). Systematic Review of Machine Learning Applications in Mining: Exploration, Exploitation, and Reclamation. *Minerals*, *11*(2), 148. https://doi.org/10.3390/min11020148

Khotimah, B. K., & Sari Rochman, E. M. (2021). MODEL PERAMALAN JUMLAH PENYAKIT DEMAM BERDARAH DENGAN PENDEKATAN METODE FUZZY LINEAR REGRESSION (FLR). *Network Engineering Research Operation*, *6*(1), 49. https://doi.org/10.21107/nero.v6i1.215

Laerd Statistics. (n.d.). *Linear Regression Analysis using SPSS Statistics*. Retrieved April 1, 2022, from https://statistics.laerd.com

Mandayam, A. U., A.C, R., Siddesha, S., & Niranjan, S. K. (2020). Prediction of Covid-19 pandemic based on Regression. *2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 1–5. https://doi.org/10.1109/ICRCICN50933.2020.9296175

Muhamad Iqbal, R., & Prihandoko, P. (2017). PENERAPAN DATA MINING UNTUK ANALISIS DATA BENCANA MILIK BNPB MENGGUNAKAN ALGORITMA K-MEANS DAN LINEAR REGRESSION. *Jurnal Ilmiah Informatika Komputer*, *22*(1). https://ejournal.gunadarma.ac.id/index.php/infokom/article/view/1535

Nastiti, M. D., Abdurohman, M., & Putrada, A. G. (2019). Smart Shopping Prediction on Smart Shopping With Linear Regression Method. *2019 7th International Conference on Information and Communication Technology (ICoICT)*, 1–6. https://doi.org/10.1109/ICoICT.2019.8835271

Permai, S. D., & Tanty, H. (2018). Linear regression model using bayesian approach for energy performance of residential building. *Procedia Computer Science*, *135*, 671–677. https://doi.org/10.1016/j.procs.2018.08.219

Shaikh, S., Gala, J., Jain, A., Advani, S., Jaidhara, S., & Roja Edinburgh, M. (2021). Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting. *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 989–995. https://doi.org/10.1109/Confluence51648.2021.9377137

Somantri, O. (2021). PREDIKSI KEKUATAN GEMPA BUMI INDONESIA BERDASARKAN NILAI MAGNITUDO MENGGUNAKAN NEURAL NETWORK. *Prosiding Seminar Nasional Informatika Bela Negara*, *2*, 203–207. https://doi.org/10.33005/santika.v2i0.124

Somantri, O., Purwaningrum, S., & Riyanto, R. (2022). MODEL SUPPORT VEKTOR MACHINE (SVM) BERDASARKAN PARAMETER WINDOWS UNTUK PREDIKSI KEKUATAN GEMPA BUMI. *JTT (Jurnal Teknologi Terapan)*, *8*(1), 17–24. https://doi.org/10.31884/JTT.V8I1.352

Sugiyarto, I., Irawan, R., & Rosiyadi, D. (2021). Pengelompokan Dampak Gempa Bumi Dan Kerusakan Pada Wilayah Berpotensi Gempa Di Provinsi Sumatra Barat. *Journal of Students' Research in Computer Science*, *2*(2), 211–222. https://doi.org/10.31599/jsrcs.v2i2.850

Yijun, X., & Zhang, Y. (2012). Analysis and prediction of the total number of ice-snow tourism in Heilongjiang based on times series: A case study of Harbin. *2012 IEEE Symposium on Robotics and Applications (ISRA)*, 624–626. https://doi.org/10.1109/ISRA.2012.6219266

YILDIZ, H., & YANDI, A. (2021). Comparison of M5-Prime and Linear Regression Methods in Various Relationship Types between Variables. *İnönü Üniversitesi Eğitim Fakültesi Dergisi*, *22*(1), 744–771. https://doi.org/10.17679/inuefd.758378