

Analisis Curah Hujan Menggunakan *Machine Learning* Metode Regresi Linier Berganda Berbasis Python dan Jupyter Notebook

Rainfall Analysis using Machine Learning-Multiple Linear Regression Method Based on Python and Jupyter Notebook

Jesi Pebralia^{1*}

^{1*}Program Studi Fisika, Universitas Jambi, Jambi, Indonesia

Email: jesipebralia@unja.ac.id

ABSTRAK

Indonesia merupakan negara yang berada di garis khatulistiwa. Akibatnya Indonesia memiliki musim kemarau dan musim penghujan. Prediksi curah hujan sangat bermanfaat dalam berbagai bidang. Metode prediksi yang sedang berkembang dengan pesat pada saat ini yaitu metode prediksi menggunakan teknik kecerdasan buatan (*Artificial Intelligent/AI*). *Machine learning* adalah bagian dari AI. Penerapan algoritma regresi linier berganda pada *machine learning* dapat digunakan untuk memprediksi suatu variable terikat dengan berbagai jenis variable bebas yang mempengaruhinya. Pada penelitian ini, telah dilakukan prediksi curah hujan dengan melibatkan tiga variable bebas yaitu kecepatan angin, suhu udara maksimum, dan suhu udara minimum dengan dataset diperoleh dari situs kaggle.com. Dataset yang digunakan berjumlah 6.574 data, dimana data tersebut dikelompokkan ke dalam data training sebanyak 80% dan data test sebanyak 20%. Algoritma regresi linier berganda dibuat dalam Bahasa pemrograman python dan diimplementasikan menggunakan jupyter notebook. Pada penelitian ini dihasilkan model regresi linier berganda dengan persamaan $y = 1.23 + 0.1x_1 - 0.06x_2 + 0.07x_3$, nilai MSE sebesar 14.02, RMSE sebesar 3.74, dan MAE sebesar 2.27.

Kata Kunci: curah hujan; jupyter notebook; machine learning; python; regresi linier berganda

ABSTRACT

Indonesia is a country located on the equator. As a result, Indonesia has a dry season and a rainy season. Rainfall prediction is very useful in various fields. The prediction method that is currently developing rapidly is the prediction method using artificial intelligence (AI) techniques. Machine learning is a subset of AI. The application of multiple linear regression algorithms in machine learning can be used to predict a dependent variable with various types of independent variables that affect it. In this study, rainfall prediction has been carried out involving three independent variables, namely wind speed, maximum air temperature, and minimum air temperature with dataset obtained from the kaggle.com site. The dataset used is 6,574 data, where the data is grouped into training data as much as 80% and test data as much as 20%. Multiple linear regression algorithm is written in Python programming language and implemented using jupyter notebook. In this study, a multiple linear regression model was produced with the equation $y = 1.23 + 0.1x_1 - 0.06x_2 + 0.07x_3$, MSE value was 14.02, RMSE was 3.74, and MAE was 2.27.

Keyword: rainfall; jupyter notebooks; machine learning; multiple linear regression; pythons

PENDAHULUAN

Indonesia merupakan negara yang berada di garis khatulistiwa. Akibatnya Indonesia memiliki musim kemarau dan musim penghujan. Dua musim ini sangat berpengaruh terhadap kelangsungan hidup makhluk hidup di negara Indonesia. Salah satu faktor yang berpengaruh terhadap musim-musim tersebut adalah curah

hujan. Curah hujan merupakan jumlah air yang jatuh di permukaan tanah datar selama periode tertentu yang diukur dengan satuan tinggi milimeter (mm) di atas permukaan horizontal. Curah hujan juga dapat diartikan sebagai ketinggian air hujan yang terkumpul dalam tempat yang datar, tidak menguap, tidak meresap dan tidak mengalir.

Prediksi curah hujan sangat bermanfaat dalam berbagai bidang, misalnya di bidang pertanian, prediksi curah hujan dapat digunakan sebagai informasi di sector pertanian untuk menghadapi perubahan iklim (Surmaini, dkk., 2011), informasi mengenai curah hujan juga dapat digunakan sebagai upaya deteksi bencana kekeringan (Laksono dan Nurgiyatna, 2020), di bidang pembudidayaan informasi curah hujan juga sangat berguna untuk meningkatkan produktivitas hasil budidaya, misalnya pada budidaya lebah penghasil madu, informasi curah hujan dapat digunakan untuk menghitung besarnya pembiayaan budidaya (Yuda, 2011), melalui informasi indeks curah hujan dapat diperoleh informasi prediksi harga premi asuransi pertanian (Erfiana, dkk., 2020). Oleh sebab itu, metode prediksi curah hujan sangat penting sekali untuk dikembangkan.

Metode prediksi yang sedang berkembang dengan pesat pada saat ini yaitu metode prediksi menggunakan teknik kecerdasan buatan (*Artificial Intelligent/AI*). Pada dasarnya kecerdasan buatan adalah suatu pengetahuan yang membuat computer dapat meniru kecerdasan manusia, misalnya melakukan analisis penalaran untuk mendapatkan suatu kesimpulan (Subakti, dkk., 2022). Machine learning adalah bagian dari AI yang berfungsi untuk membuat computer memiliki kemampuan untuk belajar tentang data baru tanpa harus deprogram secara eksplisit. Fokus utamanya adalah untuk membangun sebuah aplikasi computer yang dapat mempelajari data, kemudian membuat sebuah model yang siap digunakan untuk menyelesaikan kasus tertentu (Id, Ibnu Daqiqil, 2021).

Machine learning dapat dibedakan menjadi dua tipe, yaitu *supervised learning* dan *unsupervised learning*. Pada pembelajaran tipe *supervised learning*, telah diketahui apa yang akan diprediksi atau target variabelnya. Terdapat berbagai macam algoritma pada pembelajaran tipe supervised

learning, yaitu linear regression, neural network, decision trees, support vector machine (SVM), naïve Bayes, dan K-nearest neighbors. Sedangkan pada tipe pembelajaran unsupervised learning tidak terdapat target variable (target feature) pada dataset. Contoh algoritma pada pembelajaran unsupervised learning yaitu association rule dan K-mean clustering (Faisal dan Nugrahadi, 2017).

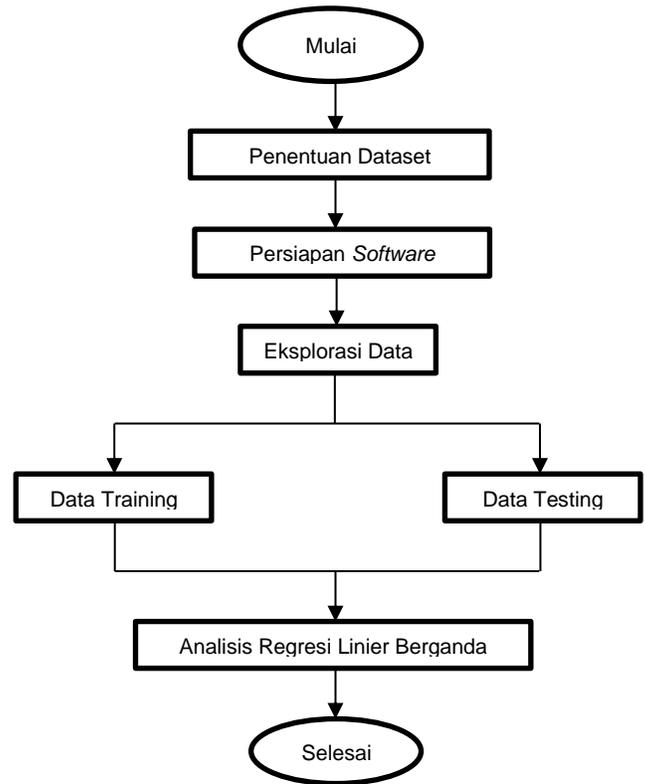
Penerapan algoritma regresi linier berganda pada machine learning telah banyak dilakukan oleh peneliti terdahulu. Algoritma regresi linier berganda dapat digunakan untuk melakukan prediksi biaya asuransi kesehatan berdasarkan kriteria variable bebas seperti umur, jenis kelamin, berat badan, jumlah anak, kebiasaan merokok, dan wilayah (Sholeh, dkk., 2022). Penelitian lain yang menggunakan algoritma regresi linier dilakukan oleh Maulidi dan Nafiiyah tahun 2022 yang memprediksi presentasi potongan harga penjualan dengan dataset berjumlah 1.525 data. Selain itu, algoritma regresi linier juga dapat digunakan untuk memprediksi jumlah peminat mata kuliah pilihan (Afkarina, dkk., 2019), prediksi penyebaran Covid-19 (Putri, dkk., 2021), prediksi tingkat inflasi (Amrin, 2016), prediksi omset penjualan suatu produk ((Adiguno, dkk., 2022), (Nafiiyah, 2019)), dan prediksi hujan bulanan menggunakan data suhu dan kelembaban udara (Fadholi, 2013), (Budiman dan Akhlakulkarimah, 2016)). Penelitian-penelitian yang telah dilakukan tersebut menunjukkan bahwa algoritma regresi linier berganda dapat digunakan untuk memprediksi suatu variable terikat dengan berbagai jenis variable bebas yang mempengaruhinya.

Penelitian untuk memprediksi curah hujan menggunakan metode regresi linier pernah dilakukan oleh Fadholi pada tahun 2013. Pada penelitian tersebut digunakan data curah hujan, suhu udara, dan

kelembaban udara bulanan yang diperoleh dari Stasiun Meteorologi Sultan Baabullah Ternate. Selain itu, Budiman dan Akhlakulkarimah pada tahun 2016 juga telah melakukan penelitian prediksi curah hujan menggunakan metode regresi linier berganda dengan melibatkan dua variable bebas berupa jumlah hari hujan dan lama penyinaran matahari dalam sebulan, yang diperoleh dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) Stasiun Klimatologi Kls I Banjarbaru. Pada penelitian ini, peneliti melakukan prediksi curah hujan dengan melibatkan tiga variable bebas yaitu kecepatan angin, suhu udara maksimum, dan suhu udara minimum dengan dataset diperoleh dari situs kaggle.com. Peneliti menggunakan algoritma regresi linier berganda pada *machine learning* yang dijalankan menggunakan Bahasa pemrograman python dan jupyter notebook.

METODE PENELITIAN

Penelitian ini menggunakan metode regresi linier berganda untuk memprediksi curah hujan. Terdapat tiga variable bebas yang digunakan yaitu kecepatan angin, suhu maksimum, dan suhu minimum. Dataset yang digunakan diperoleh dari situs www.kaggle.com. Diagram alir penelitian ditampilkan pada gambar 1.



Gambar 1. Diagram alir penelitian

Dataset Penelitian

Jumlah dataset yaitu sebesar 6.574 data. Terdapat dua tahapan yang harus dilakukan dalam proses pembelajaran algoritma di machine learning, yaitu tahap training dan tahap pengujian (testing). Pada penelitian ini, dataset dibagi menjadi 80% data training dan 20% data testing.

Eksplorasi Data

Langkah eksplorasi data bertujuan untuk lebih mengenal data dan untuk meningkatkan kualitas data yang akan digunakan. Langkah pertama dalam proses eksplorasi data yaitu menyesuaikan data mentah dengan kebutuhan data yang akan dianalisis.

Analisis Data

Analisis regresi linier adalah metode statistic yang dapat digunakan untuk mempelajari hubungan antar sifat permasalahan yang sedang diselidiki. Model regresi linier berganda merupakan model regresi yang melibatkan lebih dari satu variable bebas. Analisis regresi linier berganda bertujuan untuk mengetahui arah dan seberapa besar pengaruh variable bebas terhadap variable terikat.

$$y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n \quad (1)$$

Pada persamaan 1 variabel y merupakan variable terikat, a merupakan intersept atau konstanta regresi, x merupakan variable bebas, dan b merupakan koefisien regresi pada masing-masing variable bebas. Besarnya keterkaitan antara variable terikat dengan variable bebas dinyatakan oleh koefisien korelasi,

$$r = \frac{N \sum_{i=1}^N (x_i y_i) - \sum_{i=1}^N (x_i) \sum_{i=1}^N (y_i)}{\sqrt{[N \sum_{i=1}^N (x_i^2) - (\sum_{i=1}^N (x_i))^2][N \sum_{i=1}^N (y_i^2) - (\sum_{i=1}^N (y_i))^2]}} \quad (2)$$

Nilai r yang positif mengindikasikan bahwa meningkatnya nilai x akan menyebabkan meningkatnya nilai y , sebaliknya untuk nilai r yang negative mengindikasikan bahwa meningkatnya nilai x akan menyebabkan menurunnya nilai y , dan jika nilai r sama dengan nol mengindikasikan bahwa tidak ada korelasi antara x dan y . Pengujian model algoritma yang digunakan dapat dilakukan dengan cara mencari nilai *Mean Square Error* (MSE), *Root Mean Square Error* (RMSE), dan *Mean Absolute Error* (MAE). MSE merupakan kuadrat rata-rata dari nilai error (Kusuma, 2020),

$$MSE = \sum \frac{(y' - y)^2}{n} \quad (3)$$

dimana y' merupakan nilai variable terikat hasil prediksi, y merupakan nilai variable terikat sebenarnya, dan n merupakan jumlah data. *Root Mean Square Error* (RMSE) adalah jumlah dari kesalahan kuadrat atau selisih antara nilai

sebenarnya dengan nilai prediksi yang telah ditentukan.

$$RMSE = \sqrt{\sum \frac{(y' - y)^2}{n}} \quad (4)$$

Mean Absolute Error (MAE) menunjukkan nilai kesalahan rata-rata yang error dari nilai sebenarnya dengan nilai prediksi. MAE sendiri secara umum digunakan untuk pengukuran prediksi error pada analisis time series.

$$MAE = \sum \frac{|y' - y|}{n} \quad (5)$$

HASIL DAN PEMBAHASAN

Proses prediksi curah hujan dibuat menggunakan bahasa pemrograman python dan diimplementasikan menggunakan *software* jupyter notebook. Jupyter notebook memiliki protocol kernel yang memungkinkan server untuk melimpahkan tugas menjalankan kode menjadi berbagai bahasa.



Gambar 2. Tampilan awal *software* jupyter notebook

Terdapat lima jenis modul pada python yang digunakan untuk membuat prediksi curah hujan berdasarkan metode regresi linier berganda yaitu modul *pandas*, *numpy*, *matplotlib*, *scikit-Learn*, dan *seaborn*.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
```

Gambar 3. Modul-modul python yang digunakan untuk prediksi curah hujan berbasis metode regresi linier berganda

Pengolahan Dataset

Data mentah yang diperoleh dari situs www.kaggle.com pada awalnya terdiri dari sembilan kolom, yaitu kolom DATE, WIND, IND, RAIN, IND.1, T.MAX, IND.2, T.MIN, dan T.MIN.G. Pada penelitian ini, peneliti hanya mengambil tiga kolom yang akan digunakan sebagai variable bebas, yaitu kolom WIND, T.MAX, dan T.MIN. Adapun variable yang berfungsi sebagai variable terikat yaitu terdapat pada kolom RAIN.

	DATE	WIND	IND	RAIN	IND.1	T.MAX	IND.2	T.MIN	T.MIN.G
0	1961-01-01	13.67	0	0.2	0.0	9.5	0.0	3.7	-1.0
1	1961-01-02	11.50	0	5.1	0.0	7.2	0.0	4.2	1.1
2	1961-01-03	11.25	0	0.4	0.0	5.5	0.0	0.5	-0.5
3	1961-01-04	8.63	0	0.2	0.0	5.6	0.0	0.4	-3.2
4	1961-01-05	11.92	0	10.4	0.0	7.2	1.0	-1.5	-7.5

(a)

	WIND	TMAX	TMIN	RAIN
0	13.67	9.5	3.7	0.2
1	11.50	7.2	4.2	5.1
2	11.25	5.5	0.5	0.4
3	8.63	5.6	0.4	0.2
4	11.92	7.2	-1.5	10.4
...
6569	14.46	9.8	4.0	16.8
6570	14.33	9.1	8.5	16.0
6571	19.17	5.0	3.5	14.7
6572	18.08	2.9	0.3	4.9
6573	19.25	1.2	-1.5	0.5

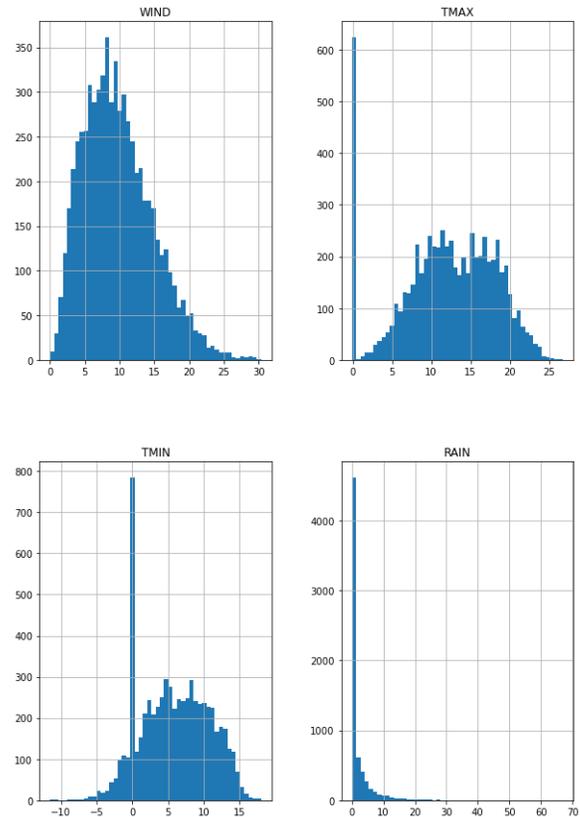
[6574 rows x 4 columns]

(b)

Gambar 4. a) Data mentah yang digunakan sebagai dataset penelitian, terdiri dari Sembilan kolom dan 6.574 baris, b) Data yang dipilih untuk dijadikan sebagai variable bebas dan variable terikat pada penelitian.

Variable-variabel bebas berfungsi sebagai variable input, dan variable terikat berfungsi sebagai variable output. Nilai dari variable-variabel input akan bertanggung jawab terhadap nilai keluaran variable output. Berikut

grafik histogram dari variable-variabel input dan variable output.



Gambar 5. Grafik histogram variable-variabel penelitian

Analisis Regresi Linier Berganda

Pada penelitian ini terdapat tiga variable bebas yang dilambangkan dengan x_1 , x_2 , dan x_3 , dan mengikuti persamaan berikut,

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 \quad (6)$$

Dengan mencari nilai a , b_1 , b_2 , dan b_3 melalui perintah pada modul scikit-Learn, maka persamaan regresi linier berganda menjadi,

$$y = 1.23 + 0.1x_1 - 0.06x_2 + 0.07x_3 \quad (7)$$

Pada persamaan tersebut, y merupakan variable curah hujan, x_1 merupakan variable kecepatan angin, x_2 merupakan variable

suhu maksimum perhari, dan x_3 merupakan variable suhu minimum perhari. Nilai korelasi antara variable bebas dan variable terikat ditampilkan pada table 1,

Tabel 1. Nilai korelasi antara variable bebas dan variable terikat

	WIND	TMAX	TMIN	RAIN
WIND	1.00	-0.21	-0.10	0.12
TMAX	-0.21	1.00	0.80	-0.05
TMIN	-0.10	0.80	1.00	0.01
RAIN	0.12	-0.05	0.01	1.00

Berdasarkan table 1, korelasi paling kuat terdapat pada korelasi antara variable kecepatan angin dengan variable curah hujan. Nilai korelasi antara kedua variable tersebut bernilai positif, yang mengindikasikan bahwa semakin besar nilai kecepatan angin, maka curah hujan juga akan meningkat. Korelasi positif juga terdapat antara variable suhu minimum dan curah hujan, yang mengindikasikan bahwa semakin tinggi suhu minimum, maka curah hujan juga semakin meningkat. Adapun nilai korelasi antara variable suhu maksimum dengan curah hujan bernilai negative, hal ini mengindikasikan bahwa semakin tinggi suhu maksimum, maka curah hujan akan semakin menurun.

Untuk melihat seberapa baik kemampuan algoritma dalam melakukan prediksi, digunakan tiga indikator statistic yaitu MSE, RMSE, dan MAE. Perhitungan ketiga indikator tersebut masing-masing mengikuti persamaan 3, 4, dan 5. Nilai dari MSE, RMSE, dan MAE ditampilkan pada table 2.

Tabel 2. Nilai MSE, RMSE, dan MAE dari analisis regresi linier berganda

Indikator	Nilai
MSE	14.02
RMSE	3.74
MAE	2.23

Berdasarkan tabel 2, dapat diketahui bahwa implementasi algoritma regresi linier berganda

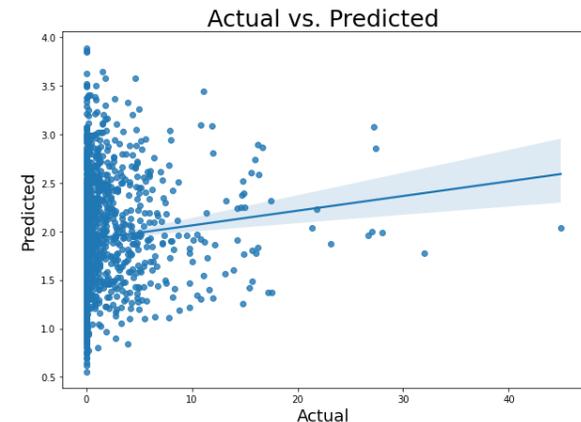
untuk memprediksi curah hujan mempunyai performansi yang baik.

Setelah proses pelatihan data selesai dijalankan, tahap selanjutnya adalah melakukan uji coba terhadap data uji. Data uji yang digunakan pada penelitian ini yaitu sebanyak 1.314 data.

Tabel 3. Sampel perbandingan nilai y sebenarnya dengan y prediksi

y_{actual}	$y_{prediksi}$
16.1	1.8
0.0	0.7
5.5	1.6
0.0	1.1
4.3	1.3
1.9	2.1
0.2	2.9
6.4	2.1
0.2	1.9

Tabel 3 menampilkan 10 sampel perbandingan data y sebenarnya dengan y prediksi. Perbandingan nilai antara nilai y sebenarnya dengan y prediksi dapat dilihat pada gambar 6,



Gambar 6. Grafik antara y sebenarnya dengan y prediksi untuk semua data uji.

KESIMPULAN

Prediksi curah hujan dapat dilakukan menggunakan algoritma linier berganda dengan Bahasa pemrograman python dan

diimplementasikan menggunakan jupyter notebook. Terdapat lima jenis modul python yang digunakan yaitu modul pandas, numpy, matplotlib, scikit-Learn, dan seaborn. Kelima modul tersebut dapat berfungsi dengan baik. Dataset yang digunakan mempunyai 6.574 data, yang dibagi menjadi 80% untuk data training, dan 20% untuk data testing. Pada penelitian ini dihasilkan persamaan regresi linier berganda yaitu $y = 1.23 + 0.1x_1 - 0.06x_2 + 0.07x_3$. Berdasarkan perhitungan nilai korelasi, dapat disimpulkan bahwa korelasi antara kecepatan angin dan curah hujan mempunyai hubungan yang paling kuat, dibandingkan dengan variable suhu maksimum dan suhu minimum. Terdapat tiga analisis statistic yang digunakan untuk melihat performansi dari algoritma regresi linier berganda, yaitu MSE, RMSE, dan MAE. Pada penelitian ini, dihasilkan nilai MSE sebesar 14.02, RMSE sebesar 3.74, dan MAE sebesar 2.23.

DAFTAR PUSTAKA

- Adiguno, S., Syahra, Y., & Yetri, M. (2022). Prediksi Peningkatan Omset Penjualan Menggunakan Metode Regresi Linier Berganda. *Jurnal Sistem Informasi Triguna Dharma (JURSI TGD)*, 1(4), 275-281.
- Afkarina, N. K., Widodo, A. W., & Furqon, M. T. (2019). Implementasi Regresi Linier Berganda Untuk Prediksi Jumlah Peminat Mata Kuliah Pilihan. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN, 2548, 964X*.
- Amrin, A. (2016). Data Mining Dengan Regresi Linier Berganda Untuk Peramalan Tingkat Inflasi. *Techno Nusa Mandiri: Journal of Computing and Information Technology*, 13(1), 74-79.
- Budiman, I., & Akhlakulkarimah, A. N. (2016). Aplikasi Data Mining Menggunakan Multiple Linear Regression Untuk Pengenalan Pola Curah Hujan. *Klik-Kumpulan Jurnal Ilmu Komputer*, 2(1), 34-33.
- Erfiana, D., Prabowo, A., Tripena, A., & Riyadi, S. (2020). Penentuan Harga Premi Asuransi Pertanian Berbasis Indek Curah Hujan Dengan Model Black-Scholes.
- Fadholi, A. (2013). Persamaan regresi prediksi curah hujan bulanan menggunakan data suhu dan kelembapan udara di Ternate. *Statistika*, 13(1).
- Faisal, M.R, dan Nugrahadi, D.T. 2019. Belajar Data Science: Klasifikasi dengan Bahasa Pemrograman R. Banjarbaru: Scripta Cendekia.
- Id, Ibnu Daqiqil. 2021. Machine Learning: Teori, Studi Kasus, dan Implementasi Menggunakan Python. Riau: UR Press.
- Kusuma, Purba Daru. 2020. Machine Learning Teori, Program, dan Studi Kasus. Yogyakarta: Deepublish.
- Laksono, S. S., & Nurgiyatna, N. (2020). Sistem Pengukur Curah Hujan sebagai Deteksi Dini Kekeringan pada Pertanian Berbasis Internet of Things (IoT). *Emitor: Jurnal Teknik Elektro*, 20(2), 117-121.
- Nafi'iyah, N., & Maulidi, N. F. (2022). Linear regression for discounting presentation recommendations (kaggle dataset). *Jurnal teknologi informasi dan komunikasi*, 13(2), 67-73.
- Nafi'iyah, N. (2019). Prediksi jumlah penjualan pada toko makmur jaya elektronik dengan regresi linier. *RESEARCH: Journal of Computer, Information System & Technology Management*, 2(2), 47-50.
- Putri, E. R. S., Novianti, F., Yasmin, Y. R. A., & Novitasari, D. C. R. (2021). prediksi kasus aktif kumulatif covid-19 di indonesia menggunakan model regresi linier berganda. *Transformasi: Jurnal Pendidikan Matematika Dan Matematika*, 5(2), 567-577.

Subakti, dkk. 2022. Artificial Intelligence. Bandung: Media Sains Indonesia.

Surmaini, E., Runtunuwu, E., & Las, I. (2011). Upaya sektor pertanian dalam menghadapi perubahan iklim. *Jurnal Litbang Pertanian*, 30(1), 1-7.

YUDA, A. (2011). *Analisa Pembiayaan Budidaya Lebah Madu Apis Mellifera pada Periode Musim Hujan di Kecamatan Tumpang Kabupaten Malang* (Doctoral dissertation, University of Muhammadiyah Malang).