

Analysis of the Comparison Between Linear Regression, Random Forest, and Logistic Regression Methods in Predicting Crude Palm Oil (CPO) Price

Surya Wijaya^{1*}, Fauziah²

^{1,2}Universitas Nasional, Indonesia

¹surya.wijaya2022@student.unas.ac.id, ²fauziah@civitas.unas.ac.id



ABSTRACT

This study compared the predictive capabilities of linear regression, random forest, and logistic regression models for forecasting crude palm oil prices. Utilizing historical data from 02/11/2020 to 13/11/2023, the dataset underwent training and testing with three scenarios: 90:10, 80:20, and 70:30. Evaluation metrics, including RMSE, MSE, and MAPE, assessed model performance. Each method had unique strengths and weaknesses, and the choice depended on application needs. The goal was to improve decision accuracy in predicting crude palm oil price trends. In the 90:10 scenario, random forest outperformed linear and logistic regression, yielding smaller MSE (43948.56), MAE (80.37), and RMSE (209.64). Similarly, in the 80:20 scenario, random forest had smaller MSE (137787.61), MAE (106.38), and RMSE (371.20). In the 70:30 scenario, random forest showed smaller MSE (107582.32), MAE (104.13), and RMSE (328). Overall, random forest consistently demonstrated better performance than linear and logistic regression.

*Corresponding Author

Article History:

Submitted: 07-12-2023

Accepted: 20-12-2023

Published: 22-12-2023

Keywords:

Random Forest, Linear Regression, Logistic Regression

Brilliance: Research of Artificial Intelligence is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

INTRODUCTION

The swift progress of Information and Communication Technology (ICT) is influencing diverse sectors globally, encompassing the economic, health, educational, and other domains. In the Industry 4.0 era, the business world is filled with competition, and companies are demanded to make accurate decisions in determining sales strategies. With the help of Information Technology, existing sales transaction data is processed and analyzed to understand buyer patterns, thus predicting the selling price. The information technology used in the data processing process is data mining. Data mining is the process of collecting and processing data aimed at obtaining essential information for decision-making (Mai et al., 2022).

In data mining, research on prediction or forecasting has been growing rapidly. Many studies have implemented data mining, including research on Linear Regression and Random Forest to predict palm oil prices (Yusuf Supriyanto, M. Ilhamsyah, 2022). Linear Regression research on house prices in the city of Bandung (Mahendra Sanusi et al., 2020). Linear Regression research on Unilever sales (Anggrawan et al., 2022). Linear Regression research on rice prices in the city of Padang (Harianti Hasibuan et al., 2022). Research using the Random Forest method on LQ45 stock prices produced (Valiant et al., 2019). Research using the random forest and decision tree to predict car prices (Kriswantara et al., 2021). Research using Linear Regression to predict sales at company PT. Eagle Industry Indonesia (Miftahuljannah et al., 2023). Investigations employing Linear Regression and LSTM for Bitcoin price prediction (Saadah & Salsabila, 2021). A study focused on monitoring and forecasting prices in an e-commerce platform utilizing Linear Regression and LSVM-ABC (Rahmawati, 2022). Research dedicated to predicting stock prices through the application of KNN and Linear (Emioma & Edeki, 2021).

LITERATURE REVIEW

(Yusuf Supriyanto, M. Ilhamsyah, 2022) conducted a comparative analysis between random forest and linear regression methods for palm oil. In the 90:10 test data ratio, the RMSE for linear regression was 27,778, while for random forest, it was 25,106. With an 80:20 ratio, the RMSE for linear regression was 31,174, and for random forest, it was 32,850. In the 70:30 ratio, the RMSE for linear regression was 31,174, and for random forest, it was 32,850. (Mahendra Sanusi et al., 2020) conducted a study using Linear Regression to analyze house prices in the city of Bandung, achieving accuracy levels ranging from 85% to 91% and error rates between 9% and 15%. Linear Regression research on Unilever sales was made by (Anggrawan et al., 2022) yielded the smallest MAPE of 1%. Linear Regression research on rice prices in the city of Padang was made by (Harianti Hasibuan et al., 2022) resulted in an RMSE value of 0.126. Research using Linear Regression to predict sales at company PT. Eagle Industry Indonesia resulted in RMSE 36,241.241 was by made (Miftahuljannah et al., 2023). Research using Multiple Linear Regression to predict Indonesia Crude Oil Prices was made by (Des Suryani et al., 2022) resulted in MAPE values of 9% for SLC, 45% for Attaka, 126% for Duri, 33% for Belida, 150% for Banyu, and 50% for SC. Research on the stock market resulted in Linear



Regression was made by (Kumar, 2022) accuracy of 85% and Support Vector Machine accuracy of 73%. (Poornima et al., 2019) conducted a study involving KNN and Linear Regression for predicting stock prices, yielding R2 values of 63% for the KNN algorithm and 98% for Linear Regression. Research on price monitoring and prediction for an e-commerce platform using Linear Regression and LSVM-ABC was made by (M Zaim Shahrel, Sofianita Mutalib, 2021) resulted in an accuracy of 62% for LR and 84% for LSVM-ABC. Research using the random forest and decision tree to predict car prices was made by (Putra et al., 2023) resulted in an accuracy for random forest of 72.13% and decision tree of 67.21%. Random Forest research on used car prices was made by (Kriswantara et al., 2021) generated an MAE of 501,232 and $R^2=0.88$.

In this research, an analysis was conducted to compare the Linear Regression, Random Forest, and Logistic Regression algorithm in predicting crude palm oil sales prices. Historical data on crude palm oil prices from 02/11/2022 to 13/11/2023 were used as the dataset for training and testing linear regression, logistic regression, and random forest models. The performance of these three methods is tested using MAE, MSE, and RMSE in each dataset split scenario of 90:10, 80:20, and 70:30.

METHOD

The approach employed in this study involves the utilization of historical data on the selling prices of crude palm oil spanning from 02/11/2020 to 13/11/2023. This dataset was utilized to both train and evaluate the linear regression, random forest, and logistic regression models. The research methodology is depicted in Figure 1.

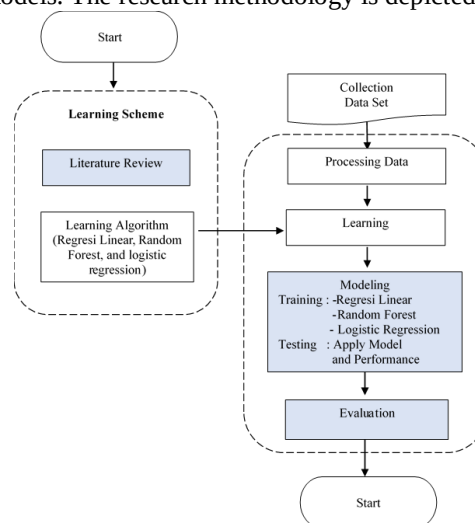


Figure 1. Purpose Research Method

Figure 1. illustrates the systematic organization of this research into several stages, including data collection (datasets), data processing and analysis, modeling, experimentation, model testing, as well as comprehensive assessment and validation.

Data Collection (dataset)

The data used in this research consists of the 02/11/2020 to 13/11/2023 sales transactions of the palm marketing company, comprising a total of 728 entries and 5 columns, including date, rott_price, soy_oil_price, kpbn_price, and price_bid.

	date	rott_price	soy_oil_price	kpbn_price	price_bid
0	2020-11-02	750	741	9526.0	9526.0
1	2020-11-03	755	732	9700.0	9700.0
2	2020-11-04	800	745	9820.0	9797.0
3	2020-11-05	805	758	9828.0	9828.0
4	2020-11-06	835	782	9806.0	9806.0
...	...	...	...	...	...
723	2023-11-07	925	1120	11190.0	11190.0
724	2023-11-08	920	1091	11235.0	11235.0
725	2023-11-09	920	1101	11194.0	11194.0
726	2023-11-10	940	1112	11280.0	65000.0
727	2023-11-13	915	1129	11270.0	11270.0

728 rows × 5 columns

Figure 2. Dataset

Processing Data

During this phase, data processing is conducted, encompassing the examination for missing values through the removal of empty or absent entries. Subsequently, data normalization is implemented, involving the transformation of data to standardize values within each record, accounting for potential variations in range across different datasets. Ultimately, the data is segregated into training and testing sets.

Modeling

During the modeling phase aimed at predicting the preprocessed data, forecasts are generated by comparing three prediction models: Linear Regression, Random Forest, and Logistic Regression. Linear regression is one of the algorithms used in data science and falls under the category of supervised learning algorithms (Januardi Ababil et al., 2022). The formula for simple linear regression can be articulated as follows:

$$y = mx + b \quad (1)$$

where:

y is the dependent variable (response variable),

m is the slope of the regression line,

x is the independent variable,

b is the y-intercept or constant term.

Random Forest is a machine learning algorithm that combines the outputs of multiple decision trees to achieve a single result. The ultimate prediction is derived through averaging or taking a vote (in the case of classification problems) across all individual tree predictions (Fitri, 2023). Logistic regression, on the other hand, is a statistical analysis technique employed to elucidate the connection between a dependent variable with two or more categories and one or more independent variables that can be either categorical or continuous (Paliling et al., 2023). The assessment phase is carried out to evaluate the performance of the algorithms on the employed data. This assessment entails a comparison of MAE, MSE, and RMSE test results (Rizkya et al., 2023) generated by the linear regression, random forest, and logistic regression algorithms.

$$MAE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_t) \quad (2)$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_t)^2 \quad (3)$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_t)^2} \quad (4)$$

where:

y_i = actual value

$\hat{y}_t$ = predicted value of y

N = number of periods

RESULT

The test model in this research was conducted using a laptop with the following specifications: AMD Ryzen processor clocked at 2.38 GHz, 8 GB of memory, and a 256 GB SSD, running on Windows 11 Pro 64-bit operating system. The testing of Linear Regression, Random Forest, and Logistic Regression models for data processing was performed using Google Colab as the tool. The training and testing data were divided using a random split with three divisions: 90:10, 80:20, and 70:30. The table below shows the testing results for MAE, MSE, and RMSE from the testing of Linear Regression, Random Forest, and Logistic Regression models.

Testing 90:10

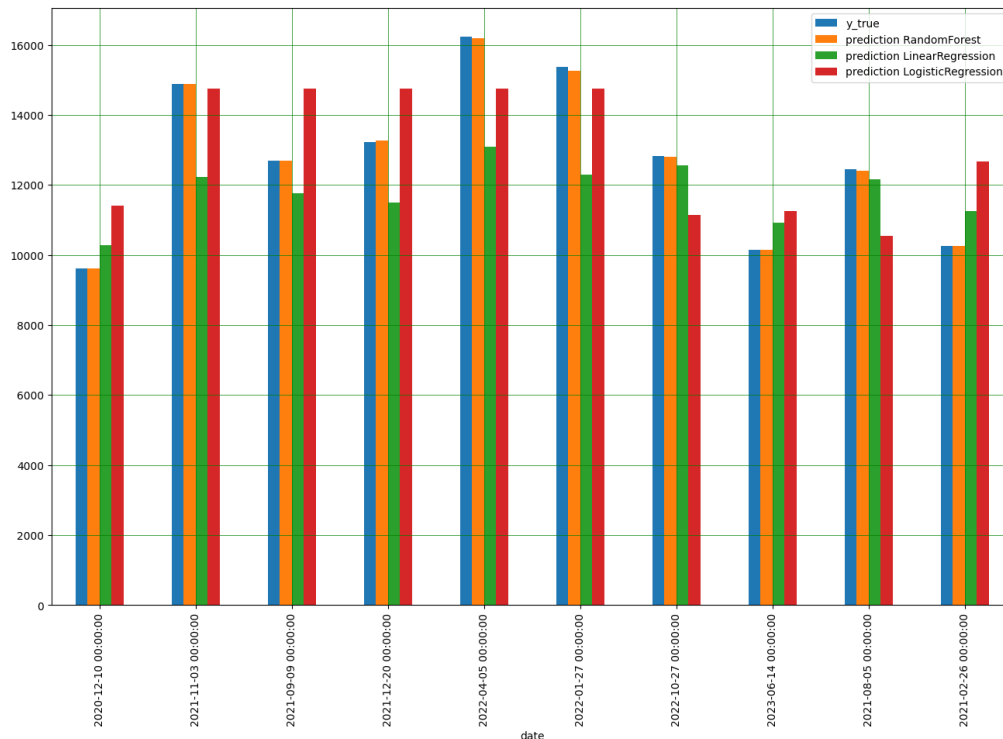


Figure 3. Result Prediction Linear Regression, Random Forest, and Logistic Regression (90:10)

After analyzing the results presented in Figure 3, it is evident that the prediction outcomes from the three models, using a 90:10 data split, demonstrate impressive performance. In 2020-12-10 00:00:00, where y_true is 9625, the Random Forest prediction is 9625.74, the Linear Regression prediction is 10274.08, and the Logistic Regression 11411. In 2021-11-03 00:00:00, where y_true is 14893, the Random Forest prediction is 14877.33, the Linear Regression prediction is 12224.97, and the Logistic Regression 14750. In 2021-09-09 00:00:00, where y_true is 12701, the Random Forest prediction is 12697.86, the Linear Regression prediction is 11756.95, and the Logistic Regression 14750.

Table 1. Result Test 90:10

MODEL	RESULT		
	MSE	MAE	RSME
Random Forest	43948.56	80.37	209.64
Linear Regression	4057263.71	1516.66	2014.27
Logistic Regression	2241867.71	1083.16	1497.29

From the testing results of the linear regression, random forest, and logistic regression models with a 90:10 data split. The Mean Squared Error (MSE) for Random Forest is 43948.56, Linear Regression is 4057263.71, and Logistic Regression is 2241867.71. The Mean Absolute Error (MAE) for Random Forest is 80.37, Linear Regression is 1516.66, and Logistic Regression is 1083.16. The Root Mean Squared Error (RSME) for Random Forest is 209.64, Linear Regression is 2014.27, and Logistic Regression is 1497.29.

Testing 80:20

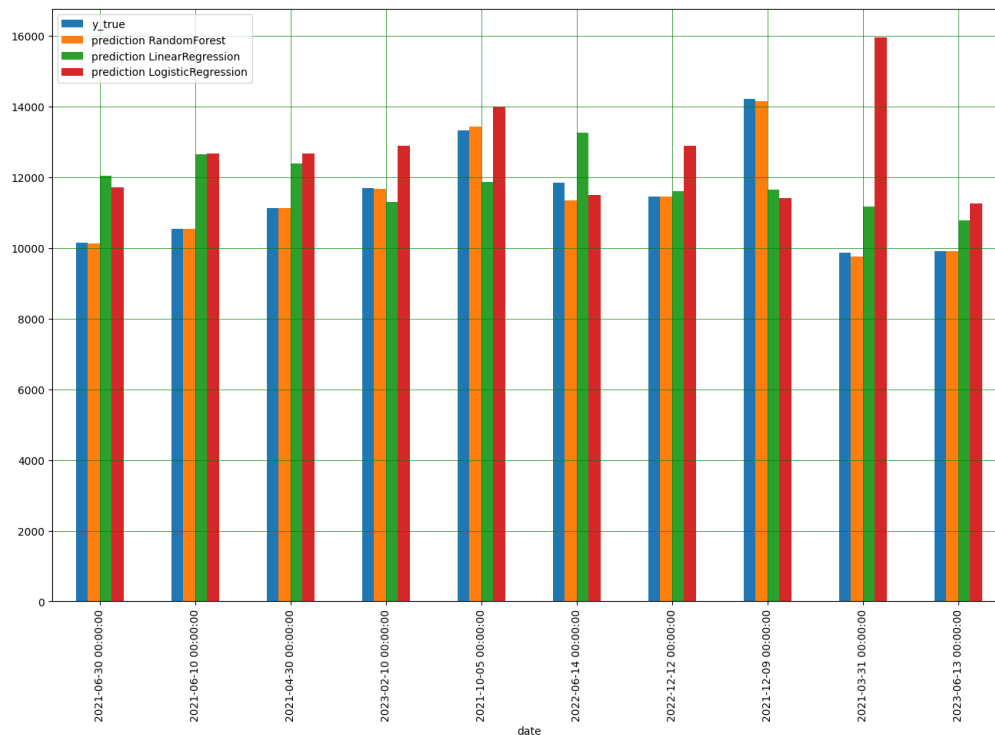


Figure 4. Result Prediction Linear Regression, Random Forest, and Logistic Regression (80:20)

After analyzing the results presented in Figure 4, it is evident that the prediction outcomes from three models, using a 80:20 data split, demonstrate impressive performance. In 2021-06-30 00:00:00, where y_{true} is 10141, the Random Forest prediction is 10121.66, the Linear Regression prediction is 12054.39, and the Logistic Regression 11711. In 2021-06-10 00:00:00, where y_{true} is 10541, the Random Forest prediction is 10540.55, the Linear Regression prediction is 12663.39, and the Logistic Regression 12682. In 2021-04-30 00:00:00, where y_{true} is 11128, the Random Forest prediction is 11127.08, the Linear Regression prediction is 12386.76, and the Logistic Regression 12682.

Table 2. Result Test 80:20

MODEL	RESULT		
	MSE	MAE	RSME
Random Forest	137787.61	106.38	371.20
Linear Regression	3411402.45	1420.79	1847
Logistic Regression	4848540.39	1597.83	2201.94

From the testing results of the linear regression, random forest, and logistic regression models with a 80:20 data split. The Mean Squared Error (MSE) for Random Forest is 137787.61 Linear Regression is 3411402.45, and Logistic Regression is 4848540.39. The Mean Absolute Error (MAE) for Random Forest is 106.38, Linear Regression is 1420.79, and Logistic Regression is 1597.83. The Root Mean Squared Error (RSME) for Random Forest is 371.20 Linear Regression is 1847, and Logistic Regression is 2201.94.

Testing 70:30

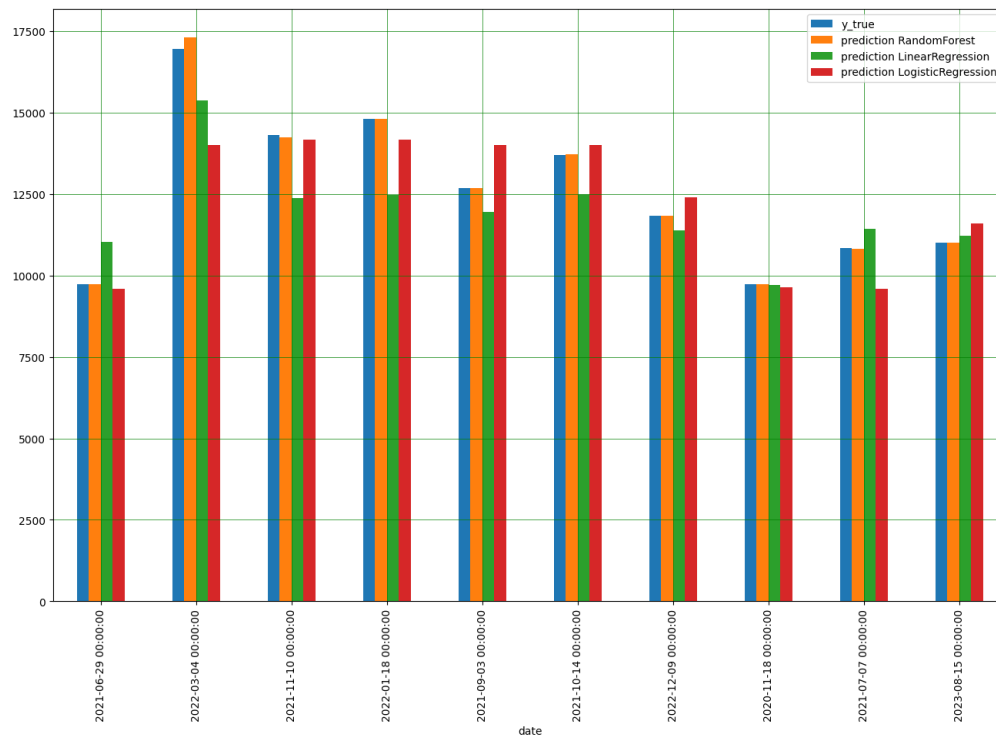


Figure 5. Result Prediction Linear Regression, Random Forest, and Logistic Regression (70:30)

After analyzing the results presented in Figure 5, it is evident that the prediction outcomes from three models, using a 70:30 data split, demonstrate impressive performance. In 2021-06-29 00:00:00, where y_{true} is 9743, the Random Forest prediction is 9736.54, the Linear Regression prediction is 11028.90, and the Logistic Regression 9580. In 2022-03-04 00:00:00, where y_{true} is 16953, the Random Forest prediction is 17308.09, the Linear Regression prediction is 15381.11, and the Logistic Regression 14010. In 2021-11-10 00:00:00, where y_{true} is 14320, the Random Forest prediction is 14236.37, the Linear Regression prediction is 12377.85, and the Logistic Regression 14170.

Table 3. Result Test 70:30

MODEL	RESULT		
	MSE	MAE	RSME
Random Forest	107582.32	104.13	328
Linear Regression	11808343.17	1339.82	3436.33
Logistic Regression	2876623.66	1208.50	1696.06

From the testing results of the linear regression, random forest, and logistic regression models with a 70:30 data split. The Mean Squared Error (MSE) for Random Forest is 107582.32 Linear Regression is 11808343.17, and Logistic Regression is 2876623.66. The Mean Absolute Error (MAE) for Random Forest is 104.13, Linear Regression is 1339.82, and Logistic Regression is 1208.50. The Root Mean Squared Error (RSME) for Random Forest is 328 Linear Regression is 3436.33, and Logistic Regression is 1696.06.

DISCUSSION

Table 4. Result Summary

MODEL	RESULT 70:30			RESULT 80:20			RESULT 90:10		
	MSE	MAE	RSME	MSE	MAE	RSME	MSE	MAE	RSME
Random Forest	107582.32	104.13	328	137787.61	106.38	371.20	43948.56	80.37	209.64
Linear Regression	11808343.17	1339.82	3436.33	3411402.45	1420.79	1847	4057263.71	1516.66	2014.27
Logistic Regression	2876623.66	1208.50	1696.06	4848540.39	1597.83	2201.94	2241867.71	1083.16	1497.29

In the evaluation of linear regression, random forest, and logistic regression, a comprehensive understanding of the models' predictive capabilities is obtained by comparing MSE, MAE, and RMSE. Reduced MAE, RMSE, and MSE values indicate enhanced model performance, reflecting lower error levels in predictions.

In the 90:10 testing scenario, the random forest algorithm successfully outperformed both linear regression and logistic regression algorithms, yielding smaller values for MSE (43948.56), MAE (80.37), and RMSE (209.64). Similarly, in the 80:20 testing scenario, the random forest algorithm produced smaller values for MSE (137787.61), MAE (106.38), and RMSE (371.20). In the 70:30 testing scenario, the random forest algorithm also demonstrated smaller results for MSE (107582.32), MAE (104.13), and RMSE (328).

CONCLUSION

The data used in this study consists of sales transactions from the palm marketing company from 02/11/2020 to 13/11/2023, comprising a total of 728 entries and 5 columns. Random Forest, Linear Regression, and Logistic Regression are used to predict the price of Crude Palm Oil (CPO). The performance of these three methods is tested using MAE, MSE, and RMSE in each dataset split scenario of 90:10, 80:20, and 70:30. Based on the test results, it is found that Random Forest performs well compared to Linear Regression and Logistic Regression.

REFERENCES

- Anggrawan, A., Azmi, N., Bumigora, U., & Anthonyangrawan, I. (2022). Prediksi Penjualan Produk Unilever Menggunakan Metode Regresi Linear Sales Prediction of Unilever Products using the Linear Regression Method. *Jurnal Bumigora Information Technology (BITE)*, 4(2), 123–132. <https://doi.org/10.30812/bite.v4i2.2416>
- Des Suryani, Mutia Fadhillah, & Ause Labellapansa. (2022). Indonesian Crude Oil Price (ICP) Prediction Using Multiple Linear Regression Algorithm. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 6(6), 1057–1063. <https://doi.org/10.29207/resti.v6i6.4590>
- Emioma, C. C., & Edeki, S. O. (2021). Stock price prediction using machine learning on least-squares linear regression basis. *Journal of Physics: Conference Series*, 1734(1). <https://doi.org/10.1088/1742-6596/1734/1/012058>
- Fitri, E. (2023). Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah. *JOURNAL OF APPLIED COMPUTER SCIENCE AND TECHNOLOGY (JACOST)*, 4(1), 2723–1453. <https://doi.org/10.52158/jacost.491>
- Harianti Hasibuan, L., Musthofa, S., Studi Matematika, P., & Imam Bonjol Padang, U. (2022). Journal of Science and Technology Penerapan Metode Regresi Linear Sederhana Untuk Prediksi Harga Beras di Kota Padang. In *Journal of Science and Technology* (Vol. 2, Issue 1).
- Januardi Ababil, O., Adi Wibowo, S., & Zulfia Zahro, H. (2022). PENERAPAN METODE REGRESI LINIER DALAM PREDIKSI PENJUALAN LIQUID VAPE DI TOKO VAPOR PANDAAN BERBASIS WEBSITE. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 6, Issue 1).
- Kriswantara, B., Kurniawati, K., & Pardede, H. F. (2021). Prediksi Harga Mobil Bekas dengan Machine Learning. *Syntax Literate ; Jurnal Ilmiah Indonesia*, 6(5), 2100. <https://doi.org/10.36418/syntax-literate.v6i5.2716>
- Kumar, S. A. (2022). STOCK MARKET PREDICTION USING LINEAR REGRESSION ALGORITHM BY COMPARING WITH ARTIFICIAL NEURAL NETWORK ALGORITHM TO IMPROVE ACCURACY. 12, 4741–4747.
- M Zaim Shahrel, Sofianita Mutalib, S. A.-R. (2021). PriceCop – Price Monitor and Prediction Using Linear Regression and LSVM-ABC Methods for E-commerce Platform. *International Journal of Information Engineering and Electronic Business*, 13(1), 1–14. <https://doi.org/10.5815/ijieeb.2021.01.01>
- Mahendra Sanusi, R., Siswo, A., Ansori, R., Wijaya, R., & Si, S. (2020). PREDIKSI HARGA RUMAH DI KOTA BANDUNG BAGIAN TIMUR DENGAN MENGGUNAKAN METODE REGRESI PREDICTION OF HOUSE PRICES IN THE EAST BANDUNG CITY USING THE REGRESSION METHOD.
- Mai, P., Tarigan, S., Tata Hardinata, J., Qurniawan, H., Safii, M., Winanjaya, R., Studi, P., Informasi, S., Tunas, S., & Pematangsiantar, B. (2022). IMPLEMENTASI DATA MINING MENGGUNAKAN ALGORITMA APRIORI DALAM MENENTUKAN PERSEDIAAN BARANG (STUDI KASUS : TOKO SINAR HARAHAH) (Vol. 12, Issue 2). <https://jurnal.umj.ac.id/index.php/just-it/index>
- Miftahuljannah, Aswan Supriyadi Sunge, & Ahmad Turmudi Zy. (2023). Analisis Prediksi Penjualan Dengan Metode Regresi Linear Di Pt. Eagle Industry Indonesia. *Jurnal Informatika Teknologi Dan Sains (Jinteks)*, 5(3), 398–403. <https://doi.org/10.51401/jinteks.v5i3.3325>
- Paliling, Y., Fathurahman, M., & Wahyuningsih, S. (2023). Multinomial Logistic Regression To Model The Combination Of Phdi Status And Hdi Status Of Districts/Cities In Kalimantan Island. *Jurnal Matematika, Statistika Dan Komputasi*, 19(3), 460–472. <https://doi.org/10.20956/j.v19i3.22299>
- Poornima, S. P., Priyanka, C. N., Reshma, P., Jaiswal, S. K., & Surendra Babu, K. N. (2019). Stock price prediction using KNN and linear regression. *International Journal of Engineering and Advanced Technology*, 8(5 SpecialIssue), 142–145. <https://doi.org/10.35940/ijeat.e1030.0585s19>

- Putra, P. H., Azanuddin, A., Purba, B., & Dalimunthe, Y. A. (2023). Random forest and decision tree algorithms for car price prediction. *Jurnal Matematika Dan Ilmu Pengetahuan Alam LLDikti Wilayah 1 (JUMPA)*, 3(2), 81–89. <https://doi.org/10.54076/jumpa.v3i2.305>
- Rahmawati, E. (2022). *IMPLEMENTASI POWER BI UNTUK DASHBOARD ANALISIS HASIL RATING DAN REVIEW PRODUK ELEKTRONIK DI LAZADA.COM* (Vol. 14, Issue 1).
- Rizkya, D., Roosaputri, H., & Dewi, C. (2023). *Perbandingan Algoritma ARIMA, Prophet, dan LSTM dalam Prediksi Penjualan Tiket Wisata Taman Hiburan (Studi Kasus: Saloka Theme Park)*. 4(3), 507–517.
- Saadah, S., & Salsabila, H. (2021). Jurnal Politeknik Caltex Riau Prediksi Harga Bitcoin Menggunakan Metode Random Forest (Studi Kasus: Data Acak Pada Awal Masa Pandemic Covid-19). In *Jurnal Komputer Terapan* (Vol. 7, Issue 1). <https://jurnal.pcr.ac.id/index.php/jkt/>
- Valiant, K., Lukito, Y., & Santosa, R. G. (2019). *Sistem Prediksi Harga Saham LQ45 Dengan Random Forest Classifier*. 2. <https://doi.org/10.21460/jutei.2019.32.187>
- Yusuf Supriyanto, M. Ilhamsyah, U. E. (2022). *Prediksi Harga Minyak Kelapa Sawit Menggunakan Linear Regression Dan Random Forest*. Vol. 8, No.