A COMPERATIVE ANALYSIS OF LINEAR REGRESSION AND RANDOM FOREST METHODS FOR PREDICTING PALM OIL PRICE SALES

[1]Surya Wijaya, [2]Fauziah
[1,2]Fakultas Teknologi Komunikasi dan Informatika, Universitas Nasional
Jl. Sawo Manila, Pejaten, Ps. Minggu, Jakarta Selatan, 12520, Daerah Khusus Ibukota Jakarta
[1]surya.wijaya2022@student.unas.ac.id, [2]fauziah@civitas.unas.ac.id

Abstrak

Dalam penelitian ini dilakukan perbandingan antara model Regresi Linier dan Random Forest dalam memprediksi harga penjualan Crude Palm Oil (CPO). Pengumpulan data dilakukan yang bersumber dari data transaksi penjualan minyak kelapa sawit dari 2018 sampai 2020, kemudian dilakukan tahap preprocessing Selanjutnya, pemisahan data dilakukan untuk data pelatihan dan pengujian, dibagi menjadi empat skema pengujian: 90:10, 80:20, 70:30, dan 60:40. Kemudian, algoritma Regresi Linier dan Random Forest diterapkan pada empat skema tersebut. Akurasi kinerja setiap model kemudian diukur untuk nilai MAE, RMSE, dan MSE. Pada setiap skema pengujian, model Regresi Linier menghasilkan nilai MAE, MSE, dan RMSE yang lebih rendah daripada model Random Forest. Skema pengujian 80:20 menghasilkan nilai MAE, MSE, dan RMSE terendah untuk Regresi Linear. Berdasarkan hasil pengujian menunjukkan bahwa Regresi Linier lebih efektif dibandingkan Random Forest.

***Kata Kunci**: harga jual CPO, prediksi, random forest, regresi linier*

Abstract

In this study, a comparative analysis was conducted between the Linear Regression and Random Forest models in predicting CPO price sales. Data collection is conducted, sourcing data from palm oil sales transaction data from 2018 to 2020, followed by data preprocessing. Subsequently, data splitting is performed for training and testing data, divided into four testing schemes: 90:10, 80:20, 70:30, and 60:40. Then, linear regression and random forest algorithms are applied to the four schemes. Next, the performance accuracy of each model is measured for MAE, RMSE, and MSE values. Notably, in each testing scheme, the Linear Regression model produced lower MAE, MSE, and RMSE values than the Random Forest model. The 80:20 testing scheme yielded the lowest MAE, MSE, and RMSE values for Linear Regression. These results suggest that Linear Regression proves to be more effective than Random Forest in predicting CPO price sales.

***Keywords**: CPO price sales, linear regression, prediction, random forest*

INTRODUCTION

Palm oil, as a plantation crop, plays a crucial role in the economic development of Indonesia, being the world's largest producer of palm oil, especially crude palm oil (CPO)[1]. The prices of CPO tend to undergo predictable fluctuations, and accurate forecasting of CPO prices holds significant importance to ensure price stability and aid decision-making. Therefore, research on the historical prices of palm oil over time is essential for effectively anticipating price fluctuations. Data mining technology is employed as a method for predicting CPO prices.

In the field of data mining[2], research related to prediction or forecasting has made significant progress. Various studies have implemented data mining techniques, including a study [3] that used Simple Linear Regression to predict rice prices with an RMSE result of 0.126. Additionally, research [4] on predicting gold values with Linear Regression yielded an MAE of 4341.140 and an RMSE of 4893.132. Meanwhile, a comparative study[5] among GBT Regression, Random Forest Regression, and Linear Regression for predicting house prices demonstrated the highest accuracy of 81.5% with Random Forest Regression. Another study [6] utilized predict house prices used Multiple Linear Regression with an accuracy rate of 66%. In the context of sales prediction at PT. Eagle Industry Indonesia, research [7] using Linear Regression resulted in an RMSE value of 36,241.241. Furthermore, a study [8] predict palm oil prices used Linear Regression and Random Forest showed an RMSE value of 30,227 for Linear Regression and 32,924 for Random Forest. A research [9] on Linear Regression for predicting house prices in Bandung City yielded an accuracy rate of 85-91%.

In this research, an analysis was conducted to compare the Linear Regression method and Random Forest in predicting the sales prices of crude palm oil (CPO). Historical data on CPO prices over the past few years were used as the dataset for train and test two models under four testing scenarios: the first test with a training-to-testing data ratio of 90:10, followed by 80:20, 70:30, and 60:40. Performance measurements were assessed using evaluation metrics such as MAE, MSE, dadn RMSE.

METHODOLOGY

The research methodology outlines the stages carried out in the study, starting from literature review to the conclusion of the research findings. The procedural stages begin with a literature review and extend to the finalization of the research results. The research methodology can also be referred to as the research framework employed in this study, as illustrated in Figure 1.

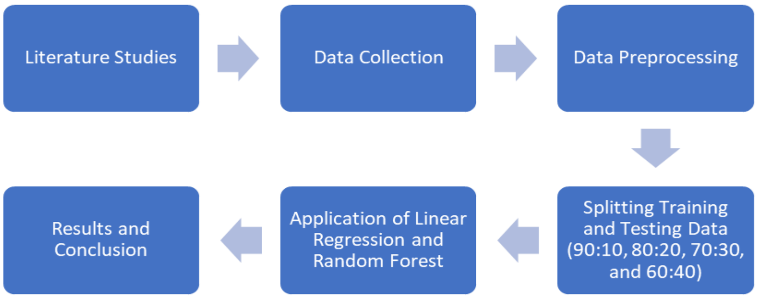

Literature Studies
Data Collection
Data Preprocessing
Splitting Training and Testing Data (90:10, 80:20, 70:30, and 60:40)
Application of Linear Regression and Random Forest
Results and Conclusion

Figure 1. Research Methodology

In Figure 1, the research begins with a literature study to find references related to our research. Then, data collection is conducted, sourcing data from palm oil sales transaction data, followed by data preprocessing to generate clean data. Subsequently, data splitting is performed for training and testing data, divided into four testing schemes: 90:10, 80:20, 70:30, and 60:40. Then, linear regression and random forest algorithms are applied to the four schemes. Next, the performance accuracy of each model is measured for MAE, RMSE, and MSE values.

Linear regression

Linear regression [10] is an approach in statistics and machine learning used to find the relationship between input values and output values. The equation for linear regression is formulated as follows:

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

where *y* represents the dependent variable, *m* is the slope of the regression line, *x* is the independent variable, and *b* is the *y*-intercept or constant term.

Random forest

Random forest [11] is an ensemble of decision trees, wherein each tree is built using a random subset of the training data and a random subset of features at each split. The ultimate prediction is derived by averaging or voting (in the case of classification problems) across all the individual tree predictions.

Evaluation of Results

The testing results are assessed using measurements of Mean Absolute Error (MAE) [12], Mean Squared Error (MSE) [13], and Root Mean Squared Error (RMSE) [14] values, with a better performance indicated by values approaching 0. The formulas for MAE, MSE, and RMSE are as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_t) \qquad (6)$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_t)^2 \qquad (7)$$

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

where y_i the actual value, $\hat{y}_t$ the predicted value of *y*, and *N* the number of periods.

RESULT AND DISCUSSION

Data Collection

The dataset utilized consists of transactional records of palm oil marketing company sales from 2018 to 2020, comprising 671 rows and 2 columns, namely date (Date) and price (Price). The sample data is presented in Table 1.

Table 1. Data Prices CPO

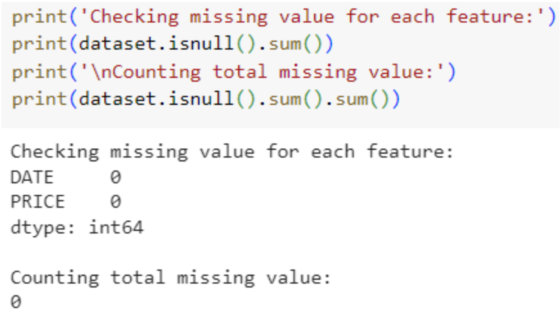
No	Date	Price
1	2018-01-04	7874
2	2018-01-05	7939
3	2018-01-08	7882
4	2018-01-09	7971
5	2018-01-10	7948
...	...	...
670	2020-10-26	9860
671	2020-10-27	9878

Data Processing

In this stage, data is checked for missing or empty values using the Python programming language, as shown in Figure 2.

In Figure 2, there is the Python programming language source code for the process of checking missing or empty data values. It can be observed from the data check that there are no missing values.

From the time series graph of CPO prices in Figure 3, it can be seen that from 2018 to 2020, there were fluctuations in CPO prices, experiencing both increases and decreases.



```
print('Checking missing value for each feature:')
print(dataset.isnull().sum())
print('\nCounting total missing value:')
print(dataset.isnull().sum().sum())
```

```
Checking missing value for each feature:
DATE     0
PRICE    0
dtype: int64

Counting total missing value:
0
```

Figure 2. Source Code Python Checking Missing Value

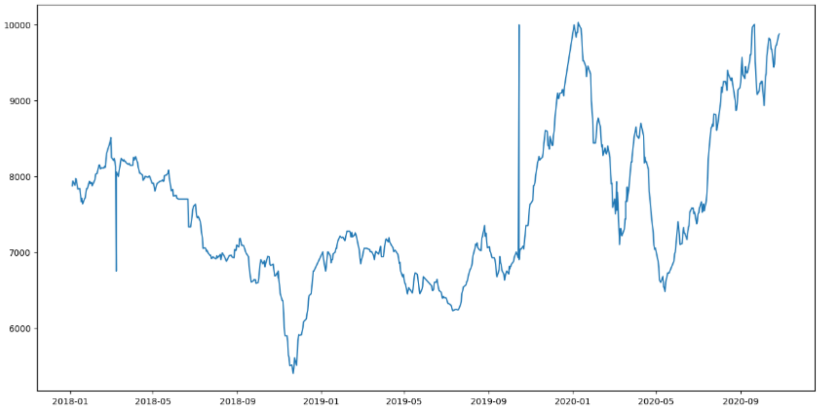


Figure 3. Plot Timeseries Price CPO

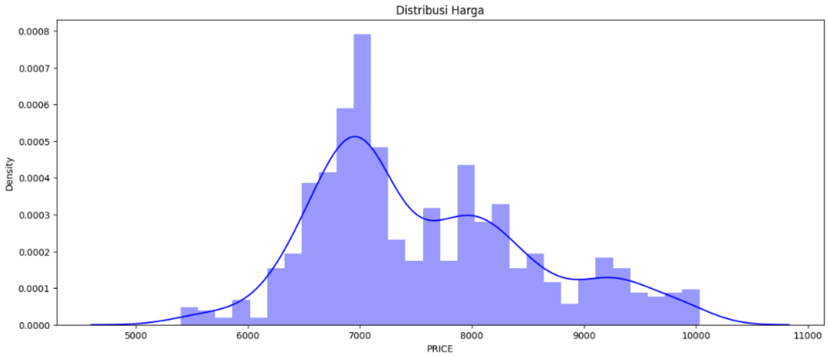


Figure 4. Plot Distribution Price CPO

Tabel 2. Ratio Test Scheme

Ratio Test Scheme	Data Train	Data Test
90:10	603	68
80:20	536	135
70:30	468	203
60:40	268	403

From the distribution price graph in Figure 4, it can be observed that from 2018 to 2020, the most common price range is around 7,000, followed by around 8,000.

Spliting Data

The dataset consists of 671 rows. In testing the model, a testing scenario scheme was used with different data train and test ratios, namely 90:10, 80:20, 70:30, and 60:40. The results of the training data can be seen in Table 2.

Based on Table 2, four testing scenarios can be observed: a 90:10 ratio with 603 training data and 67 testing data, an 80:20 ratio with 536 training data and 135 testing data, a 70:30 ratio with 468 training data and 203 testing data, and a 60:40 ratio with 268 training data and 403 testing data.

Testing Results

This stage presents the testing results of the two models, namely Random Forest and Linear Regression, for four testing scenarios with ratios of 90:10, 80:20, 70:30, and 60:40.

Based on Table 3, the index 604 shows a difference in Linear Regression prediction of 40 and a difference in Random Forest prediction of 14.

Tabel 3. Testing Result 90:10

Index	Actual Price	Linear Regression	Random Forest
604	8650	8610	8669
605	8690	8619	8601
606	8660	8658	8560
607	8823	8629	8674
608	8813	8788	8716

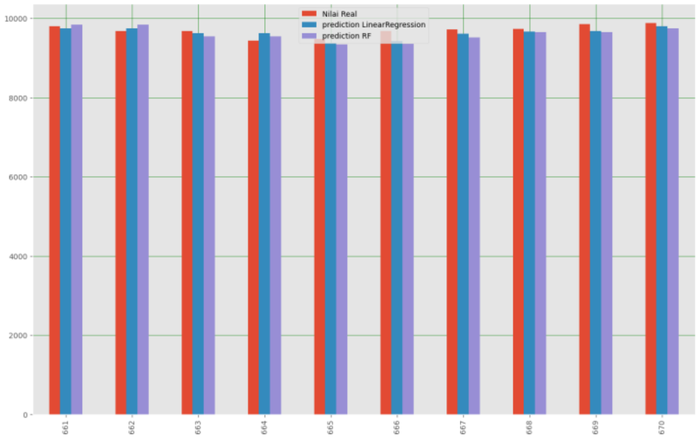


Figure 5. Plot of Test Prediction Results 90:10

Tabel 4. Result MAE, MSE, dan RMSE 90:10

Model	**MAE**	**MSE**	**RMSE**
Linear Regression	126	24740	157
Random Forest	348	326951	572

Tabel 5. Testing Result 80:20

Index	**Actual Price**	**Linear Regression**	**Random Forest**
537	8539	8473	8559
538	8637	8511	8526
539	8700	8607	8669
540	8535	8668	8588
541	8177	8507	8282

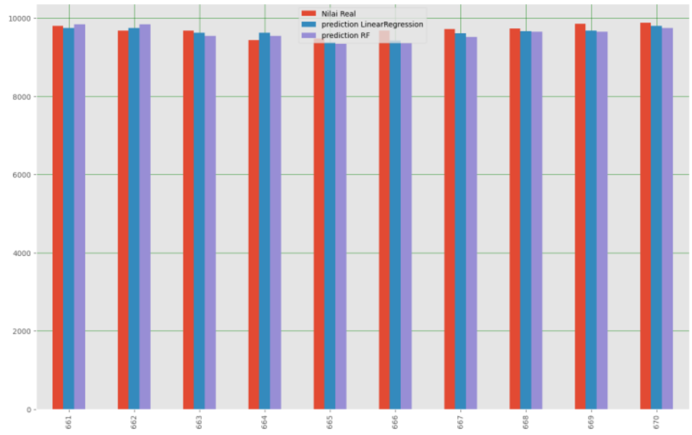


Figure 6. Plot of Test Prediction Results 80:20

Based on table 4, it can be seen that the Linear Regression produces an MAE of 126, MSE of 24740, and RMSE of 157, while Random Forest produces an MAE of 348, MSE of 326951, and RMSE of 572.

Based on Table 5, the index 537 shows a difference in Linear Regression prediction of 66 and a difference in Random Forest prediction of 20.

Tabel 6. Result MAE, MSE, dan RMSE 80:20

Model	MAE	MSE	RMSE
Linear Regression	114	20867	144
Random Forest	204	166989	409

Tabel 7. Testing Result 70:30

Index	Actual Price	Linear Regression	Random Forest
469	10032	9838	9926
470	10000	9967	9983
471	9950	9936	9959
472	9860	9887	9869
473	9650	9799	9748

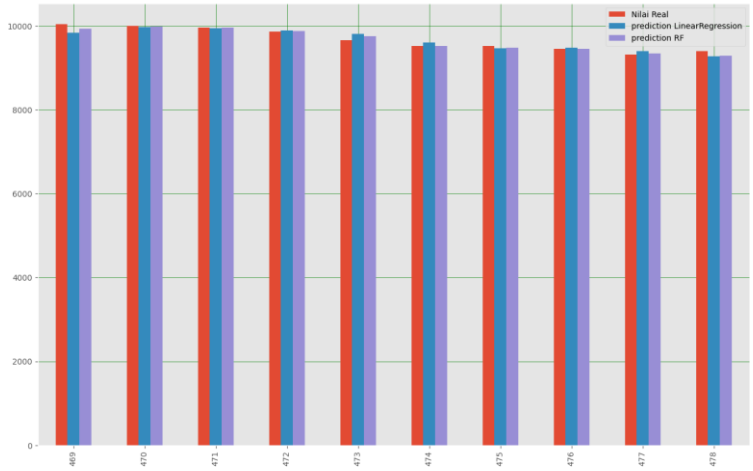


Figure 7. Plot of Test Prediction Results 70:30

Tabel 8. Result MAE, MSE, dan RMSE 70:30

Model	MAE	MSE	RMSE
Linear Regression	114	22392	150
Random Forest	157	113538	337

Based on table 6, it can be seen that the Linear Regression produces an MAE of 114, MSE of 20867, and RMSE of 144, while Random Forest produces an MAE of 204, MSE of 166989, and RMSE of 409.

Based on Table 7, the index 467 shows a difference in Linear Regression prediction of 194 and a difference in Random Forest prediction of 106.

Based on table 8, it can be seen that the Linear Regression produces an MAE of 114, MSE of 22392, and RMSE of 150, while Random Forest produces an MAE of 157, MSE of 113538, and RMSE of 337.

Based on Table 9, the index 403 shows a difference in Linear Regression prediction of 57 and a difference in Random Forest prediction of 65.

Tabel 9. Testing Result 60:40

Index	Actual Price	Linear Regression	Random Forest
403	6710	6767	6775
404	6810	6728	6727
405	6810	6825	6823
406	6810	6825	6823
407	6840	6825	6823

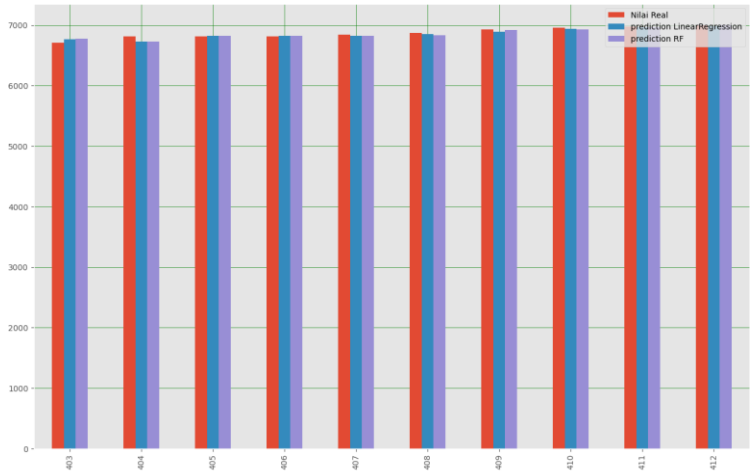


Figure 8. Plot of Test Prediction Results 60:40

Tabel 10. Result MAE, MSE, dan RMSE 60:40

Model	MAE	MSE	RMSE
Linear Regression	124	78080	279
Random Forest	136	94625	308

Tabel 11. Summary of Test Result

Model	Testing 90:10			Testing 80:20			Testing 70:30			Testing 60:40		
	MAE	MSE	RSME	MAE	MSE	RSME	MAE	MSE	RMSE	MAE	MSE	RSME
Linear Regression	126	24740	157	114	20867	144	114	22392	150	124	78080	279
Random Forest	348	326951	572	204	166989	409	157	113538	337	136	94625	308

Based on table 10, it can be seen that the Linear Regression produces an MAE of 124, MSE of 78080, and RMSE of 279, while Random Forest produces an MAE of 136, MSE of 94625, and RMSE of 308.

Table 11 shows that the Linear Regression model has lower MAE, MSE, and RMSE values than the random forest model in each testing scenario. The 80:20 testing scenario yielded the lowest MAE, MSE, and RMSE values for Linear Regression.

CONCLUSION

Linear Regression is more effective than Random Forest in predicting CPO price sales. In each testing scenario, the Linear Regression model produced lower MAE, MSE, and RMSE values than the Random Forest model. The 80:20 testing scenario yielded the lowest MAE, MSE, and RMSE values for Linear Regression.

Suggestions for this research include further development using different CPO price prediction models. Additionally, experiments with larger and more complex datasets, along with different combinations of train and test data ratios, are recommended to gain a more comprehensive understanding of model performance.

REFERENCE

[1] P. Studi Manajemen and F. Ekonomi, “Peran Aspek Tehnologi Pertanian Kelapa Sawit Untuk Meningkatkan Produktivitas Produksi Kelapa Sawit,” *J. AGRISIA*, vol. 13, no. 2, 2021.

[2] Des Suryani, Mutia Fadhilla, and Ause Labellapansa, “Indonesian Crude Oil Price (ICP) Prediction Using Multiple Linear Regression Algorithm,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi),* vol. 6, no. 6, pp. 1057–1063, 2022, doi: 10.29207/resti.v6i6.4590.

[3] I. R. Harahap, M. Z. Siambaton, and H. Santoso, “Implementasi Metode Regresi Linear Sederhana Untuk Prediksi Harga Beras Di Kota Medan,” pp. 267–273, 2013.

[4] W. Andriani, Gunawan, and A. E. Prayoga, “Prediksi Nilai Emas Menggunakan Algoritma Regresi Linear,” *J. Ilm. Inform. Komput.*, vol. 28, no. 1, pp. 27–35, 2023, doi: 10.35760/ik.2023.v28i1.8096.

[5] E. Fitri, “Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah,” *J. Appl. Comput. Sci. Technol.*, vol. 4, no. 1, pp. 2723–1453, 2023, doi: 10.52158/jacost.491.

[6] M. L. Mu’tashim, T. Muhayat, S. A. Damayanti, H. N. Zaki, and R. Wirawan, “Analisis Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Multiple Linear Regression,” *Inform. J. Ilmu Komput.*, vol. 17, no. 3, pp. 238, 2021, doi: 10.52958/iftk.v17i3.3635.

[7] Miftahuljannah, Aswan Supriyadi Sunge, and Ahmad Turmudi Zy, “Analisis Prediksi Penjualan Dengan Metode Regresi Linear Di Pt. Eagle Industry Indonesia,” *J. Inform. Teknol. dan Sains*, vol. 5, no. 3, pp. 398–403, 2023, doi: 10.51401/jinteks.v5i3.3325.

[8] U. E. Yusuf Supriyanto, M. Ilhamsyah, “Prediksi Harga Minyak Kelapa Sawit Menggunakan Linear Regression Dan Random Forest,” vol. 8, no.7, pp. 178-185, 2022.

[9] R. Mahendra Sanusi, A. Siswo, R. Ansori, and R. Wijaya, “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," *e-Proceeding of Engineering* , vol.7, no.3, pp. 9381, 2020.

[10] T. Jaelani, "Machine Learning untuk Prediksi Produksi Gula Nasional," *JMPM (Jurnal Mater. dan Proses Manufaktur),* vol. 6, no. 1, pp. 31–36, 2022, doi: 10.18196/jmpm.v6i1.14897.

[12] S. Saadah and H. Salsabila, "Prediksi Harga Bitcoin Menggunakan Metode Random Forest (Studi Kasus: Data Acak Pada Awal Masa Pandemic Covid-19)," 2021. [Online]. Available: https://jurnal.pcr.ac.id/index.php/jkt/

[13] B. Kriswantara, K. Kurniawati, and H. F. Pardede, "Prediksi Harga Mobil Bekas dengan Machine Learning," *Syntax Lit. ; J. Ilm. Indones.*, vol. 6, no. 5, pp. 2100, May 2021, doi: 10.36418/syntax-literate.v6i5.2716.

[14] A. Anggrawan, N. Azmi, U. Bumigora, and I. Anthonyangrawan, "Prediksi Penjualan Produk Unilever Menggunakan Metode Regresi Linear Sales Prediction of Unilever Products using the Linear Regression Method," *J. Bumigora Inf. Technol.*, vol. 4, no. 2, pp. 123–132, 2022, doi: 10.30812/bite.v4i2.2416.

[15] P. R. Sihombing, W. P. Lestari, M. A. Nursaskiawati, and E. Indryani, "Perbandingan Performa ETS dan ARIMA dalam Pemodelan Harga CPO," *J. Ekon. Dan Stat. Indones.*, vol. 2, no. 2, pp. 207–211, Aug. 2022, doi: 10.11594/jesi.02.02.08.