**Scientific Journal of Informatics**
Vol. 10, No. 3, Aug 2023

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

# Arabica Coffee Price Prediction Using the Long Short Term Memory Network (LSTM) Algorithm

**Lila Setiyani[1*], Wiranto Herry Utomo[2]**
[1,2]Department of Informatics, President University, Indonesia

**Abstract.**
**Purpose:** Arabica coffee beans have been widely cultivated in various parts of the world. The need for coffee beans is estimated to increase every year. This was followed by the rapid growth of franchised coffee shops and cafes, therefore Arabica coffee beans have been traded legally in the world, thus making the price of these Arabica coffee beans a public concern. This prediction of the price of Arabica coffee beans can be input for business actors in the coffee shop, café franchises and farmers in the decision making process. This study aims to predict the price of Arabica coffee beans in 2023 and 2024 using the long short term memory (LSTM) Algorithm.
**Methods:** The research procedure is carried out by collecting data, data analysis and preprocessing, building a forecasting model using the Long Short-Term Memory Network (LSTM) algorithm. Arabica coffee bean price datasets in this study were taken from The Pink Sheet World Bank Commodity Price Data, which presents Arabica coffee bean prices from 1960 to February 2023.
**Results:** The results of this study indicate the predicted price of Arabica coffee beans in 2023 and 2024 with Error (MAE), which is the average absolute difference between the actual value and the predicted value.
**Novelty:** What is most important and what differentiates it from previous research is in the preprocessing using two algorithms namely MinMaxScaler and Sliding Window. Meanwhile, for the training model, GridSearchCV is used. The model is evaluated using the lost function using Mean Squared Error (MSE) and Mean Absolute Error (MAE) thereby making it easy to evaluate the performance of the model.
**Keywords**: Arabica coffee beans, Prediction, LSTM, long short-term memory network
**Received** May 2023 / **Revised** May 2023 / **Accepted** July 2023



## INTRODUCTION

The concept of commodities provides a foundation for studying capitalism[1]. International commodity prices have a major impact on the farmers in developing countries who produce commodities for their livelihoods[2]. Coffee is the second largest commodity traded internationally and is a globally traded agricultural product. Additionally, coffee is one of the most consumed beverages globally, which has a significant impact on millions of people, from farmers to consumers. This has made the global coffee trade continue to grow since 2000[3]. Coffee in the World Bank Commodity Price Data (The Pink Sheet) is one of the listed commodities and is divided into two, namely Arabica Coffee and Robusta Coffee[4].

The price of coffee is an important indicator that drives welfare. So that, the price of this coffee commodity carries risks for farmers and related business actors[5]. Commodity price forecasting is a challenge for traders. This forecast had a major impact on the traders and farmers involved in maximizing profits and minimizing business losses [6]. Price forecasting is a key concern for market participant. According to Xu and Zhang, yield price forecasting can be used independently as a technical forecast that forms a price trend perspective and can be used for policy analysis[7][8]. In addition, according to Liu, et al explained that commodity prices are an important factor for investment management and assist in making better decisions. Thus, fluctuations in commodity prices in the market require satisfactory forecasting[9].

This study aims to predict the price of coffee commodities, especially Arabica types in 2023 and 2024. This research can support the business of farmers and coffee shop entrepreneurs. Currently there are many theoretical studies on commodity price forecasting algorithms including K-Nearest Neighbor[10][11][12], SVR, Prophet, XGBoost, LSTM[13][14] and so on. According to [15], chili commodity price prediction

[*]Corresponding author.
Email addresses: lila.setiyani@student.president.ac.id (Setiyani)
DOI: 10.15294/sji.v10i3.44162

can be done by applying the K-Nearest Neighbor (KNN) algorithm. In addition, according to [16], commodity price predictions for corn, gold and crude oil can be made by building a combined neural model. Chili commodity price prediction was also carried out by [17], he used an Artificial Neural Network with a MAPE value to analyze its accuracy. Research from Huang and Wu explains that Deep Multiple Learning can be used to predict energy commodities. In their research they used MAE, MAPE and Theil's U to see their performance[18]. However, in research it is often found that the dataset is limited, thus affecting accuracy. Therefore efforts are needed to improve the prediction accuracy. For time series data, the Long Short-Term Memory Network (LSTM) according to[13] promises to provide a better accuracy model. So in this study proposed LSTM as an algorithm used to build a price prediction model for Arabica coffee bean commodities. The World Bank releases data on commodity prices in real time with a time span from 1960 to the present[4]. So that the commodity price data presented by the World Bank can be used as an appropriate dataset. With the right dataset, it is possible to give the right predictions. So that the prediction results produced can be utilized by farmers and related business actors to increase their productivity. To improve the quality of predictions in the preprocessing process, two algorithms are used, namely MinMaxScaler[19] and Sliding Window[20]. Meanwhile, for the training model, GridSearchCV[21]. The model is evaluated using a missing function using Mean Squared Error (MSE)[22] and Mean Absolute Error (MAE)[23], making it easier to evaluate model performance. The remainder of this paper is organized as follows. Part II deals with the methods used for prediction and LSTM and explains the research procedure. The results of the model and evaluation model are discussed in section III. Finally, conclusions are presented and the limitations of this algorithm and future research are discussed.

## METHODS

The method used should be accompanied by references; the relevant modification should be explained. The procedure and data analysis technique should be emphasized in a literature review article. The stages and analysis of the research must be explained in detail.

### The Long Short-Term Memory Network (LSTM)

Long short-term memory network (LSTM) is a modification of the recurrent neural network (RNN). LSTM is here to complement the lack of RNN which cannot predict words based on past information stored for a long time. LSTM is able to remember information that is stored for a long period of time as well as delete irrelevant information, so that this LSTM is more efficient in processing, predicting as well as classifying data based on a certain time sequence[24]. LSTM has four gates namely forget gate, input gate, input modulation gate and output gate [25].

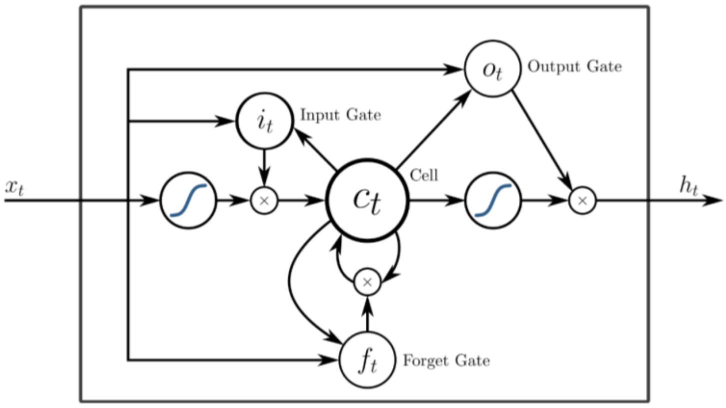


Figure 1. Typical representation of an LSTM [25]

You can see from this diagram that an LSTM unit (or layer) is composed of gates, denoted by[25]:

- *It* (the input gate: the gate that regulates the input into the unit/layer),
- *ot* (the output gate: the gate that regulates the output from the unit)
- *ft* (the forget gate: the gate that regulates what the cell should forget)

Based on the LSTM network described by[26] , we can obtain that the LSTM network enables not only extraction of information from time-series data, but also propagation of information from previous time steps with long-term dependencies[26].

The research procedure can be seen in Figure 2 below:

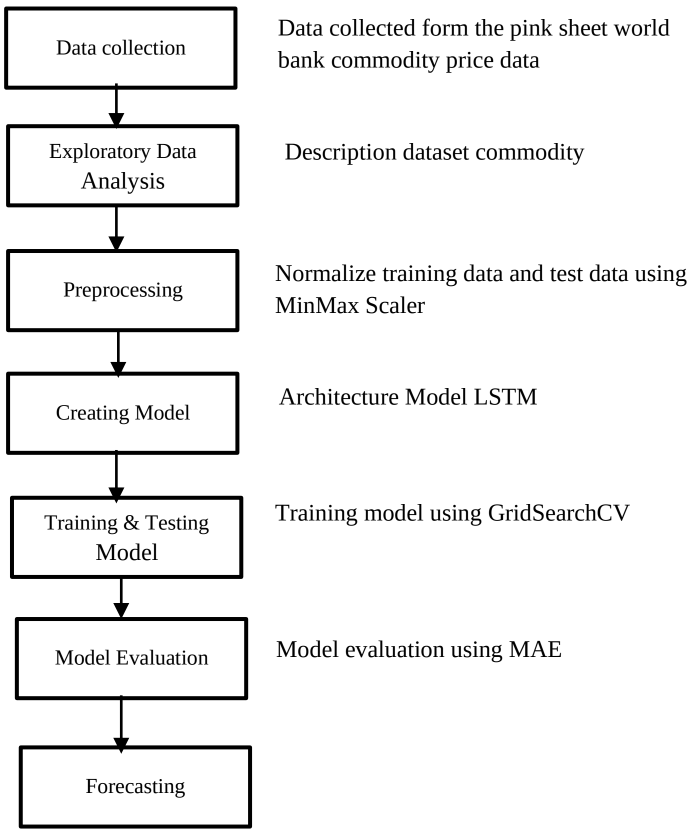


Figure 2. Research procedure

Datasets collected from the pink sheet World Bank Commodity Price Data. Exploratory data analysis do for describe datasets. At the preprocessing stage, there are two algorithms used, namely MinMaxScaler[19] and Sliding Window[20]. Meanwhile, for the training model, GridSearchCV[21] is used. The model is evaluated using the lost function using Mean Squared Error (MSE)[22] and Mean Absolute Error (MAE)[23].

**Exploratory Data Analysis**

Datasets taken from World Bank Commodity Price Data (The Pink Sheet) [4]. Arabica coffee bean commodity prices are taken from January 1960 to February 2023. Arabica coffee bean price data presented by the World Bank Commodity Price Data is in the form of monthly price data. The data is analyzed to find out if there are empty data, and to ensure the type of data contained in each attribute and to analyze the month sequence of the price data.

Table 1. Dataset description

| # | Column | Non-Null | Count | Dtype |
|---|---|---|---|---|
| **0** | Month | 758 | Non-null | Object |
| **1** | Price | 758 | Non-null | Float64 |

Data columns (total 2 columns):

The data presented has 758 non-null data with the data type for the Month attribute being object and the Price attribute being float64. So that the month attribute value can be processed, a conversion is made from the object data type to datetime.

Based on the time sequence, we found data with price movements from January 1960 to March 2023. The following figure 3 is a graph of the price movement of Arabica coffee beans:

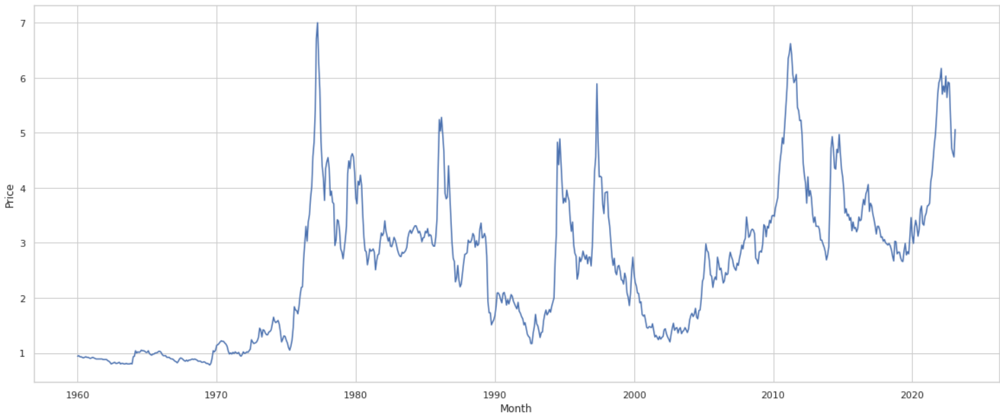

Figure 3. A graph of Arabica coffee bean price movements

Table 2 below presents a description of the datasets:

Table 2. Description of the datasets

| | Price |
|---|---|
| count | 758.000000 |
| mean | 2.582586 |
| std | 1.345676 |
| min | 0.780000 |
| 25% | 1.352500 |
| 50% | 2.700000 |
| 75% | 3.317500 |
| max | 7.000000 |

The reason why using LSTM, because the dataset obtained on the coffee commodity is in the form of a time sequence. Therefore, it is necessary to use an algorithm that has modules with repeated processing. So as to improve accuracy. Parameters used to predict the price of coffee.

The data is prepared by taking datasets from the pink sheet World Bank Commodity Price Data, then exploring and filtering the data by looking at null data and its record period. Then the datasets that have been prepared are uploaded to the drive which is then processed using Google Colab. By using the LSTM model, the researchers divided the data set into 70% training data and the rest for testing data. The iteration of the model testing process is carried out for 50 epochs and by using the MSE loss function and MAE metric.

## RESULTS AND DISCUSSIONS

### Preprocessing

At the preprocessing stage, split data is carried out, this is done to evaluate the ability of the model that has been trained. In this study, the dataset used was time series data, so the data split was not done randomly, and cross validation was carried out using train data. The train data in this study uses the first 70% of the row and the remaining 30% is used as test data. In this study, to normalize training data and test data, the MinMaxScaler feature is used. To form a structure on time series data, a Slidding Window is used, this algorithm can produce input variables (X) and target variables (y). And for the input data on the LSTM, it is ensured that it is 3D, namely [samples, timesteps, features].

**Architecture Model LSTM**

The forecasting model created using LSTM is defined by several functions including (1) input_shape = (window_size,1), (2) Dense layer with 32 neurons with the ReLu activation function, (3) Dropout between dense layer and dense output layer, (4) Dense output layer with 1 neuron, (5) lost function using Mean Squared Error (MSE), (6) Optimizer used by adam, (7) Metric used is Mean Absolute Error (MAE), (7) Parameters used as the input of the function is LSTM_unit = default 64 and dropout = default 0.2.

The model that has been made is then carried out hypertuning experiments on parameters by experimenting with combinations of LSTM units 16,32,64 and dropout opportunities of 0.1 and 0.2 and using early stopping in training.

**Training Model Using GridSearchCV**

Model training using GridSearchCV is done by entering parameters, namely the estimator (the model you want to do gridserach), param_grid (the parameter you want to test), n_jobs (the number of jobs to run in parallel, and cv (amount of k-fold cross validation), which produces a model with MSE loss function and MAE metric as shown in fig. 4 and fig. 5.

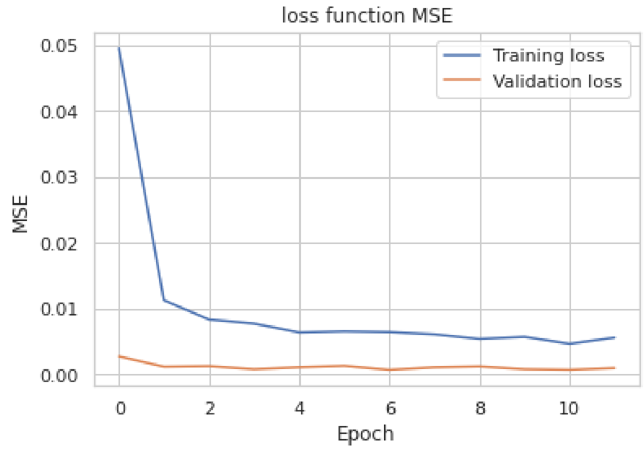


Figure 4. Loss function MSE

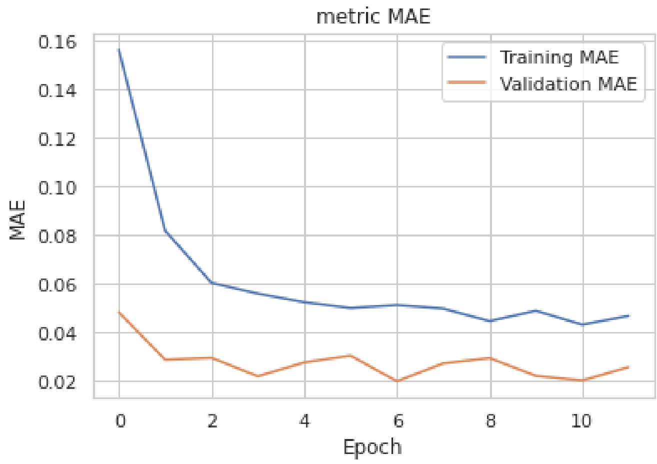


Figure 5. Metric MAE

Based on Fig. 4 and Fig 5 it can be seen that the training loss, validation loss, MAE training and MAE validation lines are moving downwards, this indicates that the model built has given good results

**Model Evaluation**

The model evaluation process is carried out by making predictions on the train data and test data which can be seen in the command below and table 3 evaluation of training data and table 4 evaluation of testing data.

Table 3. Evauation and training

| | Actual Price | Price Prediction | Error (MAE) |
|---|---|---|---|
| **0** | 0.88 | 0.971357 | 0.030765 |
| **1** | 0.88 | 0.969538 | 0.030472 |
| **2** | 0.86 | 0.968329 | 0.030278 |
| **3** | 0.85 | 0.961394 | 0.029163 |
| **4** | 0.83 | 0.953175 | 0.027842 |
| **...** | ... | ... | ... |
| **501** | 1.42 | 1.522713 | 0.119407 |
| **502** | 1.37 | 1.522402 | 0.119357 |
| **503** | 1.43 | 1.506268 | 0.116763 |
| **504** | 1.60 | 1.512726 | 0.117802 |
| **505** | 1.68 | 1.571477 | 0.127247 |

Figure 6. Training data evaluation graph

Table 4. Evaluation of testing data

| | Actual Price | Price Prediction | Error (MAE) |
|---|---|---|---|
| 0 | 2.51 | 2.658441 | 0.302000 |
| 1 | 2.54 | 2.655047 | 0.301454 |
| 2 | 2.41 | 2.660620 | 0.302350 |
| 3 | 2.27 | 2.623468 | 0.296377 |
| 4 | 2.31 | 2.551359 | 0.284784 |
| ... | ... | ... | ... |

| | | | |
|---|---|---|---|
| 199 | 5.29 | 5.313884 | 0.728920 |
| 200 | 4.72 | 5.156385 | 0.703599 |
| 201 | 4.63 | 4.894592 | 0.661510 |
| 202 | 4.56 | 4.675712 | 0.626320 |
| 203 | 5.06 | 4.492974 | 0.596941 |

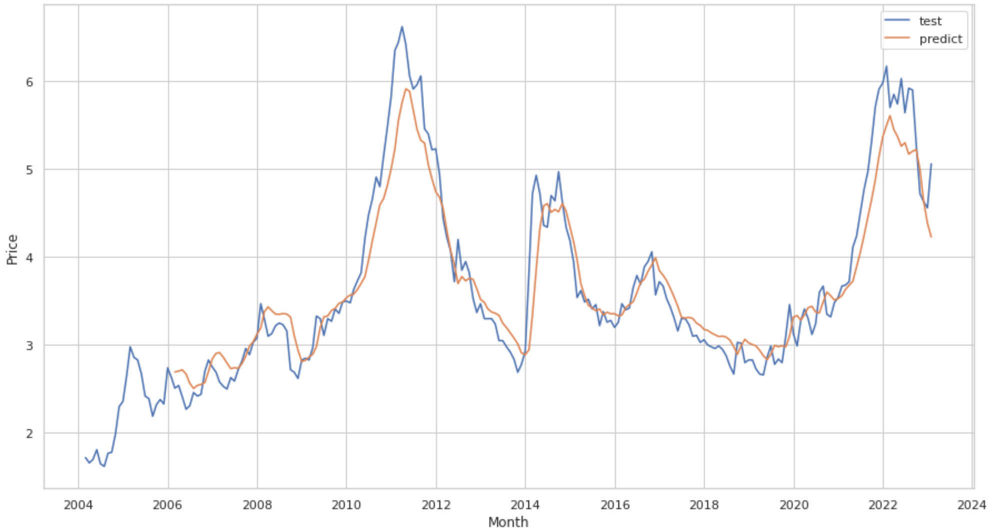


Figure 7. Testing and evaluation graph

The graph in Fig. 7 shows the compatibility between the results of testing and predictions. The blue line shows the test results while the orange line shows the predicted results, the proximity of the blue line to the orange line shows that the test results are in line with the predictions.

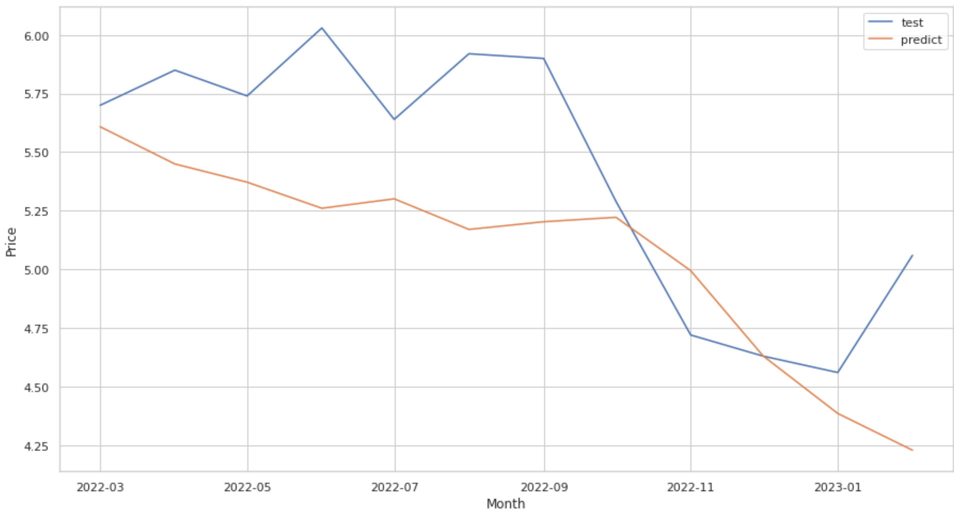


Figure 8. Plot of data for the last year

The graph in Fig. 8 shows in detail or zooms in the movement of the last year's testing and prediction results.

## Forecasting

After evaluating the train data and data testing, forecasting is carried out for the next years.

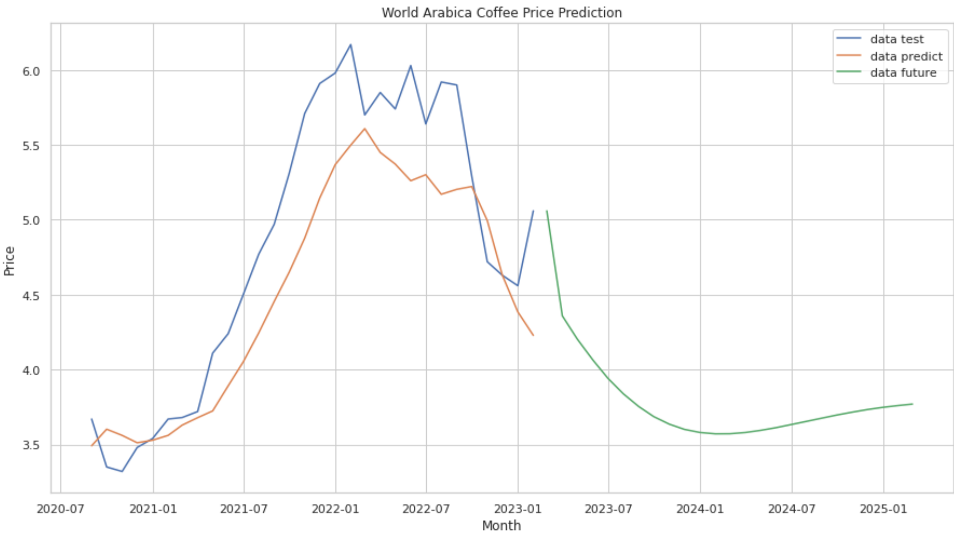


Figure 9. Plot data for the past and next

To find out the performance of forecasting results, an evaluation of the MAE error is carried out. Following are the results of the MAE Error analysis based on forecasting for the next two years accompanied by forecasting prices.

Table 5. Forecasting Arabica coffee seed with error (MAE)

| Index | Month/Date | Prediction Price | Error (MAE) |
|---|---|---|---|
| 0 | 2023-02-28 | 5.06 | 0.5753600597381592 |
| 1 | 2023-03-31 | 4.35873957157135 | 0.5503560304641724 |
| 2 | 2023-04-30 | 4.203214509487152 | 0.5280625820159912 |
| 3 | 2023-05-31 | 4.064549260139465 | 0.5084219574928284 |
| 4 | 2023-06-30 | 3.9423845756053923 | 0.491689532995224 |
| 5 | 2023-07-31 | 3.838308895230293 | 0.4779532849788666 |
| 6 | 2023-08-31 | 3.75286943256855 | 0.4671490490436554 |
| 7 | 2023-09-30 | 3.685667085051536 | 0.4590996205806732 |
| 8 | 2023-10-31 | 3.635599640011787 | 0.4535532295703888 |
| 9 | 2023-11-30 | 3.6011010879278182 | 0.4502120316028595 |
| 10 | 2023-12-31 | 3.580318836569786 | 0.4487534761428833 |
| 11 | 2024-01-31 | 3.5712466216087337 | 0.4488489627838135 |
| 12 | 2024-02-29 | 3.5718405485153197 | 0.4501781463623047 |
| 13 | 2024-03-31 | 3.580108070373535 | 0.4524375796318054 |
| 14 | 2024-04-30 | 3.5941617453098296 | 0.45534566044807434 |
| 15 | 2024-05-31 | 3.612250007987022 | 0.4586544930934906 |
| 16 | 2024-06-30 | 3.632830947041511 | 0.4620888829231262 |
| 17 | 2024-07-31 | 3.6541928517818447 | 0.4655955731868744 |
| 18 | 2024-08-31 | 3.6760044652223582 | 0.4689587950706482 |
| 19 | 2024-09-30 | 3.6969237053394313 | 0.4720494747161865 |
| 20 | 2024-10-31 | 3.71614773273468 | 0.47481033205986023 |
| 21 | 2024-11-30 | 3.7333202654123303 | 0.4771864414215088 |
| 22 | 2024-12-31 | 3.748099665641784 | 0.47914257645606995 |
| 23 | 2025-01-31 | 3.7602668255567546 | 0.48067575693130493 |

## Analysis

In this study, MinMaxScaler was able to normalize training data and testing data, this is in line with [27]which states that the MinMaxScaler technique is a feature simplification solution that goes too far and normalizes it by carrying out a linear transformation on the resulting data which has an impact on the balance of comparison values between data. In addition, the use of a sliding window is capable of producing a variable (X) and a target variable (Y), which, according to[28], the sliding window is capable of transforming time series data into cross-sectional data. The LSTM model that was formed to predict Arabica coffee prices produced good results, this can be seen from the training model shown in the MSE fan matric MAE lost function graph and the testing data evaluation graph shown in fig. 7 and fig. 8, this is in line with research conducted by[29] in predicting time series data by measuring the MSE for each epoch that is carried out. LSTM in this study proves to be able to remember a collection of information that has been stored for a long time, this is evidenced in coffee price predictions. LSTM is more efficient in predicting based on a certain time sequence.

## CONCLUSION

Based on the results of training and testing, the Arabica coffee bean prediction model using the LSTM algorithm has good performance, this can be evaluated from the MSE loss function graph and the MAE metric which drops to close to 0. In addition, the forecasting results can also look good, because it is based on the MAE error value. has a value less than 0.6. In the process of exploratory data analysis to ensure the data is ready to be modeled, it is necessary to ensure that the data type and sequence are appropriate in the sense that they do not jump, because this will affect forecasting results. This model can be used to predict other commodities presented by the World Bank Commodity Price Data (The Pink Sheet) [4]. Further research can be improved by producing better forecasting performance, which can be seen from the decrease in the MAE Error value.

## REFERENCES


[1] E. Landsteiner and E. Langthaler, “Global Commodities: Special Issue of the,” *Commod. Front.*, 2021.

[2] H. A. Kebede, “The Pass-Through of International Commodity Price Shocks to Producers’ Welfare: Evidence from Ethiopian Coffee Farmers,” *World Bank Econ. Rev.*, vol. 36, no. 2, pp. 305–328, 2022, doi: 10.1093/wber/lhab020.

[3] J. Pancsira, “International Coffee Trade: a literature review,” *J. Agric. Informatics*, vol. 13, no. 1, pp. 26–35, 2022, doi: 10.17700/jai.2022.13.1.654.

[4] World Bank Commodity Price Data (The Pink Sheet), “World Bank Commodity Price Data (The Pink Sheet).”

[5] F. Fitriani, B. Arifin, and H. Ismono, “Indonesian coffee exports and its relation to global market integration,” *J. Socioecon. Dev.*, vol. 4, no. 1, p. 120, 2021, doi: 10.31328/jsed.v4i1.2115.

[6] H. Ke, Z. Zuominyang, L. Qiumei, and L. Yin, “Predicting Chinese Commodity Futures Price: An EEMD-Hurst-LSTM Hybrid Approach,” *IEEE Access*, vol. 11, no. January, pp. 1–1, 2023, doi: 10.1109/access.2023.3239924.

[7] X. Xu and Y. Zhang, “Soybean and Soybean Oil Price Forecasting through the Nonlinear Autoregressive Neural Network (NARNN) and NARNN with Exogenous Inputs (NARNN–X),” *Intell. Syst. with Appl.*, vol. 13, p. 200061, 2022, doi: 10.1016/j.iswa.2022.200061.

[8] X. Xu and Y. Zhang, “Thermal coal price forecasting via the neural network,” *Intell. Syst. with Appl.*, vol. 14, 2022, doi: 10.1016/j.iswa.2022.200084.

[9] Y. Liu, C. Yang, K. Huang, and W. Liu, “A Multi-Factor Selection and Fusion Method through the CNN-LSTM Network for Dynamic Price Forecasting,” *Mathematics*, 2023.

[10] A. Bode, “K-NEAREST NEIGHBOR WITH FEATURE SELECTION USING BACKWARD ELIMINATION FOR PRICE PREDICTION OF ARABICA COFFEE COMMODITIES,” *Ilk. J. Ilm.*, vol. 9, no. 2, pp. 188–195, 2017, doi: 10.33096/ilkom.v9i2.139.188-195.

[11] M. E. Lasulika, “Corn Commodity Price Prediction Using K-Nn And Particle Swarm Optimization As Selection Features,” *Ilk. J. Ilm.*, vol. 9, no. 3, pp. 233–238, 2017, doi: 10.33096/ilkom.v9i3.148.233-238.

[12] M. Nanja and P. Purwanto, “Forward Selection Based K-Nearest Neighbor Method For Pepper Commodity Price Prediction,” *Pseudocode*, vol. 2, no. 1, pp. 53–64, 2015, doi: 10.33369/pseudocode.2.1.53-64.

[13] Z. Chen, H. S. Goh, K. L. Sin, K. Lim, N. K. H. Chung, and X. Y. Liew, “Automated Agriculture Commodity Price Prediction System with Machine Learning Techniques,” *Adv. Sci. Technol. Eng. Syst. J.*, vol. 6, no. 4, pp. 376–384, 2021, doi: 10.25046/aj060442.

[14] Y. H. Gu, D. Jin, H. Yin, R. Zheng, X. Piao, and S. J. Yoo, “Forecasting Agricultural Commodity Prices Using Dual Input Attention LSTM,” *Agric.*, vol. 12, no. 2, 2022, doi:

10.3390/agriculture12020256.
[15] S. Hasmita, F. Nhita, D. Saepudin, and A. Aditsania, "Chili commodity price forecasting in bandung regency using the adaptive synthetic sampling (ADASYN) and K-Nearest neighbor (KNN) algorithms," *2019 Int. Conf. Inf. Commun. Technol. ICOIACT 2019*, pp. 434–438, 2019, doi: 10.1109/ICOIACT46704.2019.8938525.
[16] J. Wang and X. Li, "A combined neural network model for commodity price forecasting with SSA," *Soft Comput.*, vol. 22, no. 16, pp. 5323–5333, 2018, doi: 10.1007/s00500-018-3023-2.
[17] W. Anggraeni *et al.*, "Agricultural strategic commodity price forecasting using artificial neural network," *2018 Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2018*, pp. 347–352, 2018, doi: 10.1109/ISRITI.2018.8864442.
[18] S. C. Huang and C. F. Wu, "Energy commodity price forecasting with deep multiple kernel learning," *Energies*, vol. 11, no. 11, pp. 1–16, 2018, doi: 10.3390/en11113029.
[19] B. Komer, J. Bergstra, and C. Eliasmith, "Hyperopt-Sklearn: Automatic Hyperparameter Configuration for Scikit-Learn," *Proc. 13th Python Sci. Conf.*, no. Scipy, pp. 32–37, 2014, doi: 10.25080/majora-14bd3278-006.
[20] V. Suresh, P. Janik, J. Rezmer, and Z. Leonowicz, "Forecasting solar PV output using convolutional neural networks with a sliding window algorithm," *Energies*, vol. 13, no. 3, 2020, doi: 10.3390/en13030723.
[21] Z. M. Alhakeem, Y. M. Jebur, S. N. Henedy, H. Imran, L. F. A. Bernardo, and H. M. Hussein, "Prediction of Ecofriendly Concrete Compressive Strength Using Gradient Boosting Regression Tree Combined with GridSearchCV Hyperparameter-Optimization Techniques," *Materials (Basel).*, vol. 15, no. 21, 2022, doi: 10.3390/ma15217432.
[22] T. O. Hodson, T. M. Over, and S. S. Foks, "Mean Squared Error, Deconstructed," *J. Adv. Model. Earth Syst.*, vol. 13, no. 12, pp. 1–10, 2021, doi: 10.1029/2021MS002681.
[23] J. Qi, J. Du, S. M. Siniscalchi, X. Ma, and C. H. Lee, "On Mean Absolute Error for Deep Neural Network Based Vector-to-Vector Regression," *IEEE Signal Process. Lett.*, vol. 27, pp. 1485–1489, 2020, doi: 10.1109/LSP.2020.3016837.
[24] Codecademy Team, "Long Short Term Memory Networks."
[25] StackExchange, "What is the difference between LSTM and RNN?"
[26] K. Qiu, J. Li, and D. Chen, "Optimized long short-term memory (LSTM) network for performance prediction in unconventional reservoirs," *Energy Reports*, vol. 8, pp. 15436–15445, 2022, doi: 10.1016/j.egyr.2022.11.130.
[27] F. T. Admojo and Y. I. Sulistya, "Stochastic Gradient Descent (SGD) Algorithm Performance Analysis in Classifying Formalin Tofu," *Indones. J. Data Sci.*, vol. 3, no. 1, pp. 1–8, 2022, doi: 10.56705/ijodas.v3i1.42.
[28] P. Arsi, T. Astuti, D. Rahmawati, and P. Subarkah, "Implementation of Sliding Window Algorithm on Neural Network-based Exchange Rate Prediction," vol. 6, no. 1, pp. 51–59, 2022.
[29] A. S. B. Karno, "Prediction of Bank BRI Stock Time Series Data Using LSTM (Long Short Term Memory) Learning Machines," *J. Inform. Inf. Secur.*, vol. 1, no. 1, pp. 1–8, 2020, doi: 10.31599/jiforty.v1i1.133.