

Comparative Analysis of Horticulture Crop Farmer Index Price during COVID-19 Pandemic using ARIMA and LSTM

Jhordy Wong Abuhasan¹, Sigit Widiyanto², Teddy Oswari³, Reni Diah Kusumawati⁴

¹Faculty of Computer Science, Gunadarma University, Jakarta, Indonesia, Country-16421

²Faculty of Computer Science, Gunadarma University, Jakarta, Indonesia, 16424

³Faculty of Economic, Gunadarma University, Jakarta, Indonesia, 16424

⁴Faculty of Economic, Gunadarma University, Jakarta, Indonesia, 16424

Abstract— Indonesia is a country that is rich in agricultural products so that it is called an agrarian country, coupled with the advantages of natural resources such as agricultural land, abundant diversity and tropical climate, making Indonesia able to plant and harvest various types of agricultural products throughout the year. . The price index received by farmers is a value that shows the level of development of farmers' production. This value can be used to see fluctuations in the price of products produced by farmers. In addition, the value of IT is also used as supporting data in calculating income in the agricultural sector. IT is based on the selling value of agricultural products produced by all farmer sectors, such as the food crop sector, the agricultural sector, to the fish and fishermen cultivation sector. The COVID-19 pandemic that occurs throughout the year certainly has an impact on the Indonesian economy, the Covid-19 pandemic has caused instability in all sectors of the economy without the agricultural sector. Various efforts must be made to stay focused on the sector in the midst of a pandemic like this, one way that can be done is by making predictions or forecasts of this pandemic period, so that the government can draw up plans according to the prediction results whether to complete or estimate to increase production output. The forecasting is using LSTM and ARIMA algorithm. The ARIMA used in this paper is using parameter (3,1,3), meanwhile the LSTM is using parameter (11,200,0.3,100). The results shows that LSTM is better than ARIMA because the MSE value in the LSTM model is lower (0.0015) than the ARIMA model (0.031).

Keywords— Index price, horticulture crop farmer, forecasting, time-series data, ARIMA, LSTM

I. INTRODUCTION

Indonesia is a country that is rich in agricultural products so that it is called an agrarian country, coupled with the advantages of natural resources such as agricultural land, abundant diversity and tropical climate, making Indonesia able to plant and harvest various types of agricultural products throughout the year. Based on data obtained from BPS in August 2020, agriculture is the sector that absorbs the most labor with 29.76% of the total workforce in Indonesia, presenting the agricultural sector as one of the most important sectors in Indonesia so that it can have a positive impact on the macro economy in Indonesia. The price index received by farmers is a value that shows the level of development of farmers' production. This value can be used to see fluctuations in the price of products produced by farmers. In addition, the

value of IT is also used as supporting data in calculating income in the agricultural sector. IT is based on the selling value of agricultural products produced by all farmer sectors, such as the food crop sector, the agricultural sector, to the fish and fishermen cultivation sector. The COVID-19 pandemic that occurs throughout the year certainly has an impact on the Indonesian economy, the Covid-19 pandemic has caused instability in all sectors of the economy without the agricultural sector. The sector became the last sector to survive any shocks to agriculture. But that doesn't mean the Covid-19 pandemic doesn't have an effect on farming activities. Various efforts must be made to stay focused on the sector in the midst of a pandemic like this, one way that can be done is by making predictions or forecasts of this pandemic period, so that the government can draw up plans according to the prediction results whether to complete or estimate to increase production output. forecasting is a machine learning implementation that uses data in the form of time series data. time series data is data that consists of a certain period of time.

In this study, predictions were made using the LSTM and ARIMA algorithms using time series data on agricultural price indexes from before and before the COVID-19 pandemic. The data used on this study is horticulture crop farmer index price (IT) collected from Badan Pusat Statistika Indonesia. The ARIMA and LSTM models developed in this research have several limitations, such as:

1. The research only use ARIMA and LSTM as an algorithm for prediction IT.
2. The dataset used is IT data received by horticulture plants obtained from the Badan Pusat Statistik (BPS) from April 2020 to Maret 2021.
3. The model was build using the python programming language along with jupyter notebook
4. Model analysis is done by comparing the Mean Squared Error value from the output of the two models.

The ARIMA algorithm was selected as the algorithm in this research because ARIMA can perform forecasting on stable time series dataset. While LSTM algorithm was selected because LSTM is a machine learning algorithm that can perform forecasting on time series dataset with a long period of time. In this study, the training process will using data in a .csv format which contain 2 column, namely date and index

and consisting 12 pairs of months and index

II. RELATED WORK

To support the research conducted by the author, several previous journals are needed that will be used as a reference for the preparation of this paper. References to the journals used have similarities with the research conducted by the author, therefore the journals used are also a consideration for the author in conducting research and making this paper.

The first paper is Comparison of ARIMA, ANN and LSTM for stock prediction. In this paper the author (Ma, 2020) compare 3 prediction models on stock market price data, the models are ARIMA, ANN, and LSTM. ARMA used on ARIMA model. While the ANN model is constructed by 1 hidden feed forward network layer along with 3 simple processing unit. The author comparing the ARIMA with ANN and THE LSTM with ANN model. The result is, the ANN is better than ARIMA, meanwhile the LSTM is better than ANN.

The second paper is A Comparative study between LSTM and ARIMA for sales forecasting in retail. The author (Elmasdotter, 2018) in this study comparing LSTM and ARIMA in order to forecasting the sales time series data. The study using 2 scenario for forecasting. First scenario is to forecasting on one day ahead meanwhile the second scenario is to forecasting on week a head for each dat. The results of the comparison in this study showed that LSTM tends to be superior to ARIMA, but in the scenario of predicting the next 1 day, LSTM implementation is not really needed, because the p value does not have a significant difference compared to ARIMA.

The third paper is Forecasting Economics and Financial Time Series: ARIMA vs. LSTM. The author (Namin, 2018) forecasting economic and financial time series data using ARIMA and LSTM. The ARIMA is using p,d,q parameter equals to 5,1,0 and the LSTM is constructed by 4 layer. The result of this study shows that LSTM is better then ARIMA.

III. LITERATURE REVIEW

A. Agriculture

Agriculture is a human activity in harvesting solar energy to produce food and fiber. The results of this role will fulfill human life, both food and clothing. and boards. Therefore, the progress of a nation is largely determined by the fulfillment of food, clothing and housing needs, where the level of power/ability of the nation and state in meeting these needs will determine it as a rich country or a poor country. Agricultural business is the activity of tapping solar energy into chemical energy of photosynthesis. the end product becomes parts of plants and animals which in turn become food, clothing, shelter, energy sources and industrial raw materials. The ability of an agricultural community can be relied on as a feeder of the nation (feed the nation). Even countries that are advanced in agriculture such as the United States have the belief to feed the world "feed the world". The First President of the Republic of Indonesia, I r. Soekarno emphasized that the issue of food ingredients as part of development is a matter of life or death for the Indonesian

nation. He said food development should not only rely on paddy fields, but also dry land which is Indonesia's most extensive natural resource. He conveyed this at the groundbreaking ceremony for the Agricultural development of the University of Indonesia in Bogor in 1952 (Baharsjah et al. 2014).

B. Price Index Received by Farmers (I_t)

Price Index received by farmers is an index measuring the average price change in a period of a package of types of goods produced by agriculture at the producer price level in farmers on the basis of a certain period. This index is useful for seeing fluctuations in the price of goods produced by farmers and also as supporting data in calculating agricultural sector income. The value of I_t is directly proportional to the value of production, meaning that the higher the value of I_t , the higher the value of production produced by farmers, whereas if it decreases, the income received by farmers will be lower. For example, National I_t September 2019 = 120 .02, meaning that the price level of agricultural products in September 2019 increased by an average of 1.20 times compared to the same product in the base year (2018).

C. Auto Regressive Integrated Moving Average(ARIMA)

ARIMA (Autoregressive Integrated Moving Average) is one of the forecasting techniques with a time series approach that uses correlation techniques between a time series. The rationale of the ARIMA model is that the current observation (z_t) depends on one or more previous observations (z_{t-k}). In other words, this model is made because there is a static correlation (dependent) between the series of observations. To see the existence of dependencies between observations, you can perform a correlation test between observations which is often known as the autocorrelation function (ACF) (Iriawan, 2006: 341).

D. Long Short Term Memory(LSTM)

Long Short Term Memory (LSTM) is a method of Recurrent Neural Network (RNN). LSTM addresses the problem of long-term dependence on RNNs. In the iterative RNN model only uses one simple single layer, namely the tanh layer as shown in Figure 1. The tanh layer aims to make the input into a number with a range of -1 to 1. X_{t-1} is the previous input, h_{t-1} is the previous output which will be entered as input along with the new input. H_{t+1} is output after order t and X_{t+1} is input after order

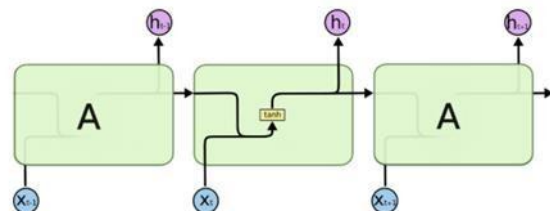


Fig. 1. Loop On 1 Layer in RNN

In contrast to RNN, LSTM has four layers, namely forget gate (1), input gate (2), new cell state candidate (3), and output gate (4) in the model loop as shown in Figure 2. The forget

gate is the gate that decides which information to delete from the cell. The input gate is the gate that decides the value of the input to be updated in the state memory. The output gate is a gate that decides what output will be produced according to the input and memory in the cell.

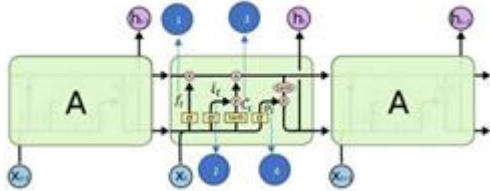


Fig. 2. Loop On 4 Layer in LSTM (Olah, 2014)

LSTM has the ability to add and remove information from the cell state. This ability is called gates. Gates as a regulator of whether the information will be forwarded or dismissed. Gates consists of a sigmoid layer and a pointwise multiplication operation. The output of the sigmoid layer is between the numbers 1 to 0 which indicates whether the information will be forwarded or dismissed. The number 0 indicates that no information will be forwarded, while the number 1 indicates that all information will be forwarded

IV. RESEARCH METHOD

The overall flow of the system on the application of the ARIMA and LSTM algorithms in the case of IT prediction for horticultural farmers can be seen in Figure 3.

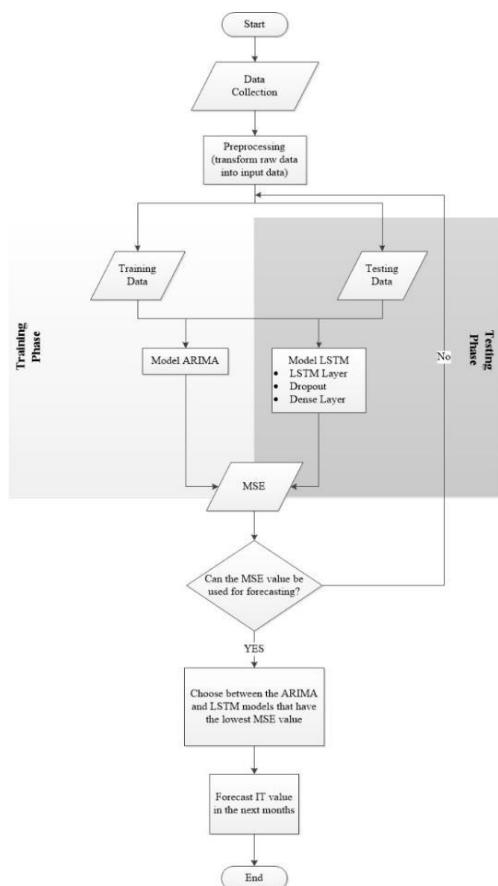


Fig. 3. System Flowchart

The following is an explanation of the flowchart in Figure 3:

1. Analysis The first step in the analysis of the time series forecasting model with the ARIMA and LSTM algorithms on the IT value of horticultural farmers begins with data collection. The data collected in the form of IT values obtained by horticultural farmers every month from April 2020 to March 2021 obtained from the BPS Indonesia website.
2. Next is preprocessing, the preprocessing stage is done by converting the raw data obtained from BPS into time series data that can be used by the designed ARIMA and LSTM models, this data will then be divided into training data and test data
3. After the raw data has been converted in to training data and test data, then the training data will be used as input data in the training phase. In the training phase, the training data will be trained using the ARIMA model and the LSTM model.
4. Furthermore, the ARIMA and LSTM models that have been successfully trained will be tested using test data obtained from split data input, the results of this test are MSE values
5. After that, the MSE value generated by the model must be considered with the condition that the MSE value generated by the model must be less than 1, if the MSE value is more than 1 then the ARIMA or LSTM model must be changed back to the architecture and parameters so that the MSE value is less than 1
6. Both ARIMA and LSTM models are compared to the MSE value obtained from the test, the model with the lowest MSE value will be chosen to predict the IT value in the next few months

A. Method of Collecting data

The data used in this study is the average value of the price index received by horticultural farmers (IT) per month obtained from the BPS Indonesia website. This research is devoted to predicting the value of IT in the period before and after the COVID-19 pandemic, the month before and after the pandemic according to BPS is April 2020 to March 2021, therefore the data taken from the BPS Indonesia website is only data from April 2020 to March 2021. The data consisting of 12 months will then be used as time series data by going through the preprocessing step. This step is carried out to modify the raw data from BPS into a dataset that can be used in the ARIMA and LSTM models which later this dataset will be divided into training data and test data.

B. ARIMA Model

The ARIMA model used in this study was prepared in several stages before training and testing the model. Some of these stages are data stability testing, differencing data, training and test data split, data training, data testing, and data forecasting.

There are two ways that will be carried out in research to test the stability of the existing dataset. The first is by analyzing the lines on the Auto-Correlation Function (ACF)

plot of the data. Second, a statistical method called the Augmented Dickey-Fuller (ADF) test will be carried out. In the ADF test, the initial hypothesis (H_0) is time series data or the existing input data is considered unstable. Then H_0 can be ignored if the p-value of the test is smaller than the significance level ($\alpha = 0.05$) and concludes that the data is stable, but if the p-value is $> \alpha$, then the value of d is required by differencing the data. The ACF plot is a graph plot of the correlation coefficient, from looking at the existing graph we can actually notice that the data we use is unstable because it has trends and seasonality. The ACF plot is only a plot (graph) of the correlation coefficient. Actually, just by observing the original graph of the time series data that we will use, we already know that the series is unstable because it has trends and seasonality.

C. LSTM Model

The LSTM model is a sequential architecture that runs the process sequentially from layer to layer that has been designed previously. The LSTM model used in this study consists of LSTM, dropout, and dense layer. Dense Layer is a layer that aims to connect all neurons in one layer with neurons in other layers. In addition, dense layers are also useful for determining the loss and MSE values of an LSTM model. Loss is a parameter that contains the value of forecasting results that are inappropriate or bad. While MSE is error metrics used to see the performance of a model, the lower these two parameters, the LSTM model designed is able to predict well.

V. DISCUSSION

A. ARIMA Preprocessing

The preprocessing applied to the data for the ARIMA model is a stationarity test or data stability test, this test is needed in order to get optimal results when training data with the ARIMA model. The data is said to be stable if the data does not have a trend and seasonality. If the data is not yet stable, it is necessary to do differencing so that the data becomes stable. The differencing stage is the preprocessing stage of the data in the ARIMA model. The data stability test is carried out in two ways, namely by paying attention to the lines on the ACF (Auto-Correlation Function) plot of the data, and the second is done by performing the ADF (Augmented Dickey-Fuller) test on the existing data. In the ADF test, the data is said to be stable if the data is less than 0.05 and vice versa.

Figure 4 below is a line chart and ACF of existing data, it can be seen that the value on the ACF chart shows that the data has a trend which is marked by changes in value from month to month starting from the first month which decreases to the sixth month and continues with an increase until the month 12, besides the results from the ADF test stated that the data had to pass the differencing stage because it had a p-value equal to 0.97.

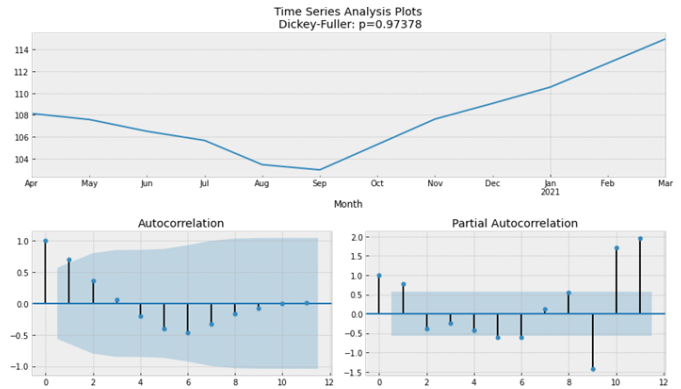


Fig. 4. ACF Plotting

Figure 5 below shows the results of the data that has gone through differencing, which shows that the data no longer has a trend compared to the initial data, besides the p-value is 0.03 which means < 0.05 therefore it can be said that differencing succeeded in making the data stable, because differencing which is done only 1 time then the value of d in the ARIMA parameter is 1.

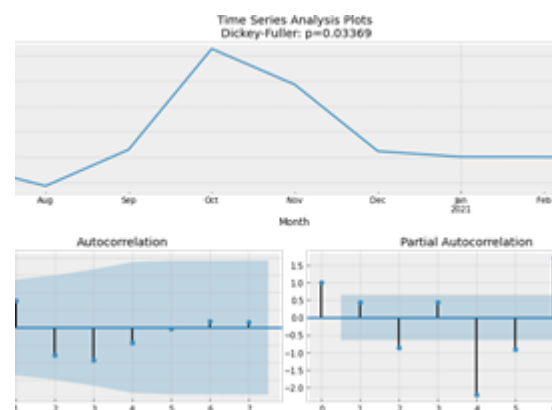


Fig. 5. Data Differencing

B. ARIMA Parameter Determination

The parameters used in the ARIMA model are p, d, q . This parameter is obtained at the preprocessing stage, where the value of d is obtained from the number of times differencing is carried out which in this case is 1, then the value of d is equal to 1. Then for the values of p and q seen from the last significant gap in the ACF shown in Figure 5 is at 3 then p and q are equal to 3. Based on the parameter values above, we will test with the parameters (3,1,3).

C. ARIMA Model Output

After the parameters are successfully set, then the ARIMA model training is carried out with the selected parameters. After the model has been trained then the model will be tested with the existing test data. In this study of a total of 12 months of data. 10 months of data will be used as training data and 2 months of data will be used as test data.

Figure 6 is a graph of the prediction results of the ARIMA model that has been designed, where the blue line is the input data and the orange line is the prediction of the model that has been trained, it can be seen that the model initially

experienced errors when making predictions but finally the prediction of the model was accurate starting from month 12 to a few months ahead. From this graph, the government as the person in charge of the agricultural sector can be said to have taken the right action because the data seems to have increased steadily, if the predicted data for the next few months has decreased, the government should take preventive action. However, the government must continue to control the horticultural agricultural processes and activities so that the monthly IT value produces a stable value or increases. The MSE of this ARIMA Model is 0.031

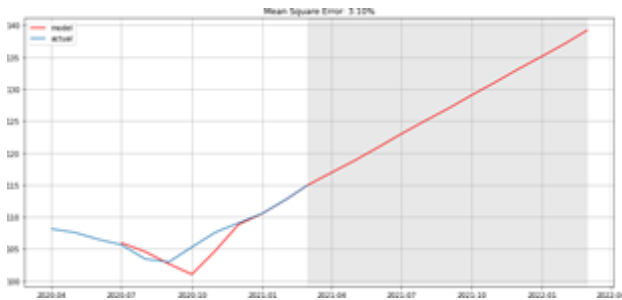


Fig. 6. ARIMA Model Prediction Result Graph

D. LSTM Parameter Determination

The determination of the parameters of the LSTM model is different from the determination of the ARIMA model where the parameters can be determined from the analysis of existing data, in the LSTM model the parameters used must undergo testing for each parameter, the parameter that has the lowest MSE value is the parameter that will be used to predict the IT value of horticultural farmers a few months ahead. The parameters used by the LSTM model in this study are:

- n_input : number of months to be forecasted (forecast)
- LSTM : value on LSTM layer layer
- dropout : value on dropout
- epoch : number of model training

From the four existing parameters, we will test each of the existing parameters. We will test 4 parameters and see the MSE value of each selected parameter. The selected parameter is the parameter that produces the lowest MSE value compared to the MSE value generated by other parameter scenarios

TABLE I. n_input Parameter Testing.

n_input	LSTM	dropout	epoch	MSE	loss
9	200	0.2	100	0.0113	0.0057
10	200	0.2	100	0.0090	0.0045
11	200	0.2	100	0.0015	0.0031

Table I shows that n_input which produces the most optimal MSE value is n_input equal to 11, this is because n_input is worth 9 and 10 produces a larger MSE than n_input 11, then we will test the LSTM parameter with a value of 100,150,200, from this result the LSTM parameter (9,200,0.2,100) we hypothesize that this parameter is temporarily the best parameter

TABLE II. LSTM Parameter Testing.

n_input	LSTM	dropout	epoch	MSE	loss
11	200	0.2	100	0.0015	0.0031
11	150	0.2	100	0.0097	0.0048
11	100	0.2	100	0.0091	0.0046

Table II shows that the LSTM 100 parameter is the most optimal parameter with the lowest MSE result compared to the LSTM 100 or 150 parameter, our first hypothesis is still fairly true. Next, we will test the epoch parameters with 100, 150 and 200 scenarios using the n_input and LSTM parameters which are considered optimal, namely 11 and 200.

TABLE III. Epoch Parameter Testing

n_input	LSTM	dropout	epoch	MSE	loss
11	100	0.2	100	0.0015	0.0031
11	100	0.2	150	0.0072	0.0036
11	100	0.2	200	0.0065	0.0032

In Table III from three scenarios, it can be seen that epoch 100 is the most optimal epoch compared to 150 and 200 where the resulting MSE value is higher than epoch 100. From this result, our initial hypothesis is still considered correct. The last parameter we want to test is dropout with 0.1, 0.2, and 0.3. scenarios

TABLE IV. Dropout Parameter Testing

n_input	LSTM	dropout	epoch	MSE	loss
11	200	0.1	100	0.0037	0.0019
11	200	0.2	100	0.0015	0.0031
11	200	0.3	100	0.0231	0.0115

Table IV shows that dropout 0.2 is the dropout parameter with the smallest MSE. Based on the test data obtained from tables I to IV, it can be concluded that our initial hypothesis is that the LSTM model has the most optimal MSE value if it is trained with the n_input, LSTM, dropout, and epoch parameters of 11, 200, 0.2, 100. Figure 7 below is error metrics diagram of the results of the most optimal parameters for the designed LSTM model, where the blue line shows the loss value from epoch to epoch and the orange line shows the MSE value from epoch to epoch

E. LSTM Model Output

After testing the most optimal parameter for LSTM model, we found the most optimal parameter is using 100 epoch for training with LSTM 100, dropout 0.2, and n_input 11 with MSE 0.0015. After trained, the LSTM model is used to predict the IT value of horticulture farmers in 11 months start from April 2021. To see the prediction results (forecasting) in more detail, please see figure 7.

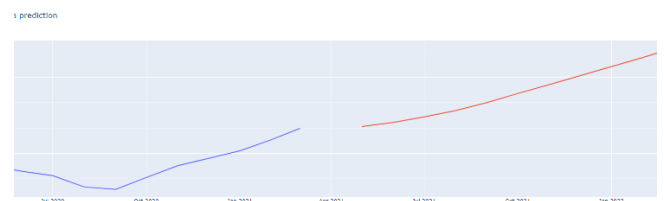


Fig. 7. LSTM Model Prediction Result Graph

In figure 6, the blue line is a line flow that represents the actual IT data flow for food crop farmers, while the red line is a line flow that represents predictive data in the form of food crop farmer IT values generated by the LSTM model starting from April 2021

F. Analysis between ARIMA and LSTM Models

Based on the prediction results from the designed model, the most optimal model for forecasting is the LSTM model. Which can be seen that the MSE value on the LSTM, which is 0.0015, is much smaller than the MSE value on ARIMA, which is 0.0095.

VI. CONCLUSIONS

The ARIMA and LSTM models that have been designed is working properly. When viewed from the test results, the prediction model has been able to predict the value of IT received by horticultural farmers. The outputs of the ARIMA and LSTM models are MSE values and predictive data for the next few months. Details of the results of this study can be seen below:

- The dataset used in this study is a dataset consisting of the value of IT received by horticultural farmers in the period April 2020 to March 2021
- The dataset used consists of 10 months of training data and 2 months of test data for the ARIMA model, while in the LSTM model, the dataset used consists of 12 months of training data.
- The ARIMA model has an MSE value of 0.0095, while the LSTM model has an MSE value of 0.0015.
- Predictions generated by the ARIMA and LSTM models in the form of predictive IT values in April 2020 to March 2021

- The results of the prediction analysis (forecasting) of IT values received by horticultural crop farmers during the COVID19 pandemic period using the ARIMA and LSTM models that have been designed show more optimal prediction results in the LSTM model where this is seen from the MSE value of the LSTM model which is smaller than the ARIMA model

For future work, other researchers can improve the accuracy of the prediction results. The author recommends further research to use more technic for preprocessing data, using different kind of ARIMA such as ARMA or SARIMAX and using more scenario for parameter determination on LSTM

REFERENCES

- [1] Badan Pusat Statistik, 2021. <https://sirusa.bps.go.id/sirusa/index.php/indikator/65> and https://daps.bps.go.id/file_artikel/77/arima.pdf. Accessed on: 1:26, 24/4/2021
- [2] Baharsjah S, Kasryno F, Pasandaran E. 2014. *Reposisi politik pertanian meretas arah baru pembangunan pertanian*. Jakarta (ID): Yayasan Pertanian Mandiri.
- [3] Elmasdotter, A., & Nyströmer, C. (2018). A comparative study between LSTM and ARIMA for sales forecasting in retail.
- [4] Iriawan, Nur, Astuti, Septin Puji, 2006, *Mengolah Data Statistik dengan mudah menggunakan Minitab 14*, Yogyakarta: ANDI, 2006
- [5] Ma, Qihang. (2020). Comparison of ARIMA, ANN and LSTM for Stock Price Prediction. *E3S Web of Conferences*. 218. 01026. 10.1051/e3sconf/202021801026.
- [6] Siami Namini, Sima & Siami Namin, Akbar. (2018). *Forecasting Economics and Financial Time Series: ARIMA vs. LSTM*.
- [7] Sima Siami-Namini & Akbar Siami Namin, 2018. "Forecasting Economics and Financial Time Series: ARIMA vs. LSTM," *Papers* 1803.06386, arXiv.org.
- [8] Christopher Olah, 2014, <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>. Accessed on 1:26, 26/4/2021