Asian Stock Index Price Prediction Analysis Using Comparison of Split Data Training and Data Testing

Baharman Supri
Prodi Manajemen, Universitas Andi Djemma, Indonesia
baharmansupri02@gmail.com

Rudianto
Prodi Manajemen, Universitas Andi Djemma, Indonesia
rudianto.unanda@gmail.com

Abdurohim
Prodi Manajemen, Universitas Jendral Ahmad Yani, Indonesia
abdurrohim@mn.unjani.ac.id

Badriatul Mawadah
Prodi Manajemen, Universitas Brawijaya, Indonesia
badriatulmawadah@gmail.com

Helmi Ali
Prodi Manajemen & Magister Manajemen, Institut Teknologi dan Bisnis Haji Agus Salim (ITBHAS) Bukittinggi, Indonesia
helmi_akbary@yahoo.com

Article's History:

Received 8 Juli 2023; *Received in revised form* 19 Juli 2023; *Accepted* 24 Juli 2023; *Published* 1 Agustus 2023. All rights reserved to the Lembaga Otonom Lembaga Informasi dan Riset Indonesia (KITA INFO dan RISET).

Suggested Citation:

Supri, B., Rudianto., Abdurrohim., Mawadah, B., & Ali, H. (2023). Asian Stock Index Price Prediction Analysis Using Comparison of Split Data Training and Data Testing. JEMSI (Jurnal Ekonomi, Manajemen, Dan Akuntansi), 9 (4). 1403-1408. https://doi.org/10.35870/jemsi.v9i4.1339

Abstract:

This study implements stock index price predictions using the LSTM method, where one of the processes in data management before running with the LSTM method is data split. This study also looks for the most appropriate split data ratio in predicting stock index prices to minimize error rates and differences in forecasted prices and original prices because in previous studies there were several rules of thumb in dividing data, so it is necessary to compare the most appropriate ratios in this research. Based on the evaluation process, the error value was found from nine split data ratios that were run by five ratios which produced a predictive graph line shape that resembled the validation line. Three datasets, namely split data ratios of 80:20, 70:30, and 60:40, are the ratios that get the lowest error values based on the RMSE, MSE, MAPE, and MAE values in the five stock index datasets. The three ratios are then compared again by looking at the average percentage difference between the validation price and the predicted price for the next working day, and it is found that the ratio of 80:20 is the most suitable split data ratio for predicting the stock index price for the next working day, with a level of difference in the average value between the original price and the predicted price on the stock index of 1.3%. While the ratio of 70:30 has an average predicted value of five stock index datasets of 1.9% and a ratio of 60:40 of 1.8%.

Keywords: data split, stock index, datasets, original price

Introduction

Investment is now starting to become a necessity that some people are starting to take into account. Currently, the number of capital market investors is 8,022,386, which has increased from the position on January 21, 2022, where the number of capital market investors reached 7.75 million investors, which figure has increased by more than 260 thousand investors since the end of December 2021. There are various kinds of capital market investment instruments, such as stocks, mutual funds, crowdfunding, and many others. Stocks are the most popular money market instruments because they are able to provide attractive profit levels but also have large potential losses. According to general information, the stock exchange is a sign of the equity participation of a person or business entity in a company or limited liability company. The uncertain nature of the stock market is a concern for stock investors, who want to avoid the potential for large losses. Stock price predictions are necessary for investors to determine the movement of a stock price in order to avoid losses and maximize profits in the stock market. To maximize profits and reduce losses, many methods have been carried out, starting with technical analysis such as moving averages, ARIMA, and the stochastic optimization method, but the results are not satisfactory (Hansun & Young, 2021).

One of the benchmarks in the stock market is the stock index value of a country's stock market, because this index value reflects the market condition of that country at that time. As explained in the explanation, an equity index is a statistical indicator of changes in the market value of a particular group of stocks or stocks, sometimes referred to as a stock index (singular) or a stock index or indices (plural) (Samsul, 2006). The JCI is an example of a stock index that represents the overall value of the Indonesian stock market index; besides that, there are other examples of a country's stock index, such as the HSI in Hong Kong, the Nikkei 225 in Japan, the KOSPI in South Korea, and the SSE Index in China. The stock itself is an important part of the stock index because stocks are the basis for its formation. The stock index can also be interpreted as a combination of stock prices expressed in an index value and becomes a consideration for stock price movements and the economic conditions of a country for investors investing in the stock market. Generally, the movement of the stock index will depend on the stock with the highest price or the stock with the largest market capitalization. There are several methods that can be used to predict stock prices, such as long- and short-Term memory (Darmawan et al., 2021).

In previous research, the researchers compared the Decision Tree, Bagging, Random Forest, Adaptive Boosting (Adaboost) models, Gradient Boosting, eXtreme Gradient Boosting (XGBoost), Artificial Neural Networks (ANN), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM). This study divides the dataset into 70% train data and 30% test data, with LSTM results being superior in predicting the overall price of the test stock based on the lowest error value and the ability to test using the Mean Absolute Percentage Error (MAPE) value. In another journal, when CNN, LSTM, and RNN were compared using data every minute, which was then made into a window size per 100 minutes to predict prices in the next 10 minutes, it was found that CNN was superior to LSTM and RNN in predicting stock prices based on window size model data (Hiransha et al., 2018). One of the stages in this research is the split-data train and test. In theory and in figures, there is no definitive decision regarding the split data ratio (Ardiansyah & Rudianto, 2018). Some references or rules of thumb that are often used in dividing train and test data First, according to professor Andrew Ng in an online class, the rule of thumb in dividing split data, namely 70:30, is possible on data 100, 1000, or 10000. However, on the amount of data 1000000, it becomes incompatible with a ratio of 70:30, so split the data 99:1. Besides that, there is another rule of thumb in data distribution, namely 80:20, also known as the Pareto principle, which is often used in mathematics, economics, and computers. Picard and Berk journals recommend 25%–50% of the data used for data testing, and Afendras and Markatou journals recommend 50%.

Literature Review

There are many investment media in Indonesia; one of the most popular is stocks. The stock index is a statistical indicator of changes in the market value of a group of stocks or certain stocks (Supri & Rudianto, 2017). The stock itself is an important part of the stock index because stocks are the basis for its formation. The stock index can also be interpreted as a combination of stock prices expressed in an index value and becomes a consideration for stock price movements and the economic conditions of a country for investors investing in the stock market. LSTM is a better form of Recurrent Neural Network (RNN), where there are additional memory cells that can store information longer. The fundamental difference between LSTM and RNN is that LSTM completes the deficiency of its predecessor, the recurrent neural network, which cannot predict data based on information that has been stored for a long time (Lestari et al., 2018). Systems that implement LSTM can process, predict, and classify information based on time series data (Benedicto, 2014). In accordance with the concept, LSTM can remember and delete old data that is no longer relevant. Technically, RNN has a simple structure with a feedback loop, while LSTM has memory blocks or cells. Data Splitting is dividing data into two or more subsets; we are usually familiar with train data and test data (Madyatmadja et al., 2021). Data splitting is important in data science, especially in the data modeling section (Nabipour et al., 2020). In the division of two sets of data, a data train will generally be formed, namely the data used to build the model (Tannady et al., 2020). Test data is data used after the model training process is complete. The data sharing ratio is divided based on the size of the dataset owned; this is because there are no guidelines or metrics in the data split.

Methodology

The stages of research progressed from the stock index stage to the evaluation stage. We first started by fetching datasets from Yahoo Finance, which is a website that provides data about the economy. We can download the data or use the yfinance library in Python. Furthermore, after the dataset is obtained, there are three stages in data preparation: cleansing, filtering, and scaling. Cleansing is doing data cleaning if there is data in the form of null. Filtering is the process of filtering data that will be used for research because data from Yahoo! Finance is still common. Scaling is scaling the data with the aim of ensuring that the prediction results are not too large. The next process after data preparation, namely data split, divides the dataset into two forms, namely train data and test data. After the data is divided, the next step is to reshape it. The goal is that the model to be used is run on a 3D model, while the current data model is 2D, so it needs to add 1D. After the data is ready, the next step is the data processing stage, namely data management to predict where this study uses the LSTM model. The LSTM model will later be formed, and the data will be processed at this stage. The results of this data processing will produce predictive values or predictive models. Furthermore, these results are still in the form of a scale; therefore, it is necessary to have an inverse process from the results of the scale to a form like the data before being scaled. After the results are converted back into their original form, the next process is an evaluation, namely calculating values based on 4 parameters, namely RMSE, MSE, MAPE, and MAE. From these 4 parameters, the results of the research with the smallest error are obtained.

Case studies

The research was started by collecting data on five stock indices from Yahoo! Finance. After the dataset is collected, it will be managed, starting with filtering the closing price or data labeled closed. After the data filtering process, scaling will be carried out using the MinMax scaling feature. The dataset that has been scaled is then split into two, known as split data. This process is an important part of this research, namely finding the best split data distribution ratio. The split data ratios studied are 90:10, 80:20, 70:30, 60:40, 50:50, 40:60, 30:70, 20:80, and 10:90, which will be run on five stock index datasets in Asia, namely JCI, N225, KOSPI, HSI, and SSE Composite. The next step after conducting data splits is to detect and test the data to predict the value of the stock price, or in this study, the stock index, using the LSTM method. The use of machine learning models requires the sum of batch sizes and epochs; therefore, this research also looks for the best batch size and epoch combinations that can be used in making predictions. After the warning dataset is predicted using LSTM, the next step is evaluating errors using RMSE, MSE, MAPE, and MAE values. To get the best research results, it is necessary to know the best number of batch sizes and epochs that match the LSTM model used in this study. There are no definite parameters or provisions for determining the batch size and research epoch; therefore, the authors conducted a running test program with a different number of batch sizes and epochs. The batch size range used is 1, 2, 4, 8, 16, 32, 64, and 128. This is supported based on number 2 because the computer works with a binary system. While the number of epochs starts from 1 overall process up to 35 overall processes. There are three datasets used in the batch size and epoch replacement processes, namely the JCI, N225, and SSE Composite index datasets.

Based on the test running program, namely batch size 2 with epoch program 11, the lowest RMSE value is

361.96, so in the split data comparison study on the N225 and KOSPI datasets, batch size 2 and epoch program 11 will be used. After finding epoch 11 and batch size 2, it will use the continuing research process to conduct training and test data to find the most suitable dataset split ratio in this study. A comparison of the split data ratio will be carried out on five stock index datasets, namely JCI, N225, KOSPI, HSI, and SSE Composite, to predict stock prices using the LSTM method. The ratios to be compared are 90:10, 80:20, 70:30, 60:40, 50:50, 40:60, 30:70, 20:80, and 10:90 and are run on five research datasets. The results of the evaluation of the experimental process for determining the most suitable split data for predicting stock prices can be seen in the following elaboration: the results of the model prediction with a ratio of 90:10, where the blue line represents train data and the red and orange lines are test data. The predicted results in the model with a ratio of 70:30 is the same as in the previous graph. The predicted results on the model with a ratio of 60:40 It can be seen that at the beginning, the predicted value does not really follow the original value because the index price at the beginning of the test data experiences a very dynamic up and down pattern change so that the model continues to catch new patterns in the predicted value. After that, the predicted value looks like it can resemble the original value line even though the predicted value looks only like a slanted line when the data experiences a pattern of high increases. the predicted results in the model with a ratio of 50:50. The predicted value is slightly below the original value but forms a similar pattern. When the index price experienced a deep fall due to the COVID-19 pandemic, the predictive value line pattern experienced quite a large difference in the shape of the prediction, but after that, the predictive value line looked similar in shape with a lower difference in values when dynamic changes occurred, such as at the end of the chart. The predicted value only resembles the pattern with the original value line, but the predicted value given is below the original value. The predicted value of the model on the JCI dataset with ratios of 40:60, 30:70, 20:80, and 10:90. These four show that the predicted value does not appear to provide a detailed pattern of changes to the original value, so for a ratio of 40:60, a ratio of 10:90 in this dataset is not recommended to be the ratio of train data and test data because the dataset used is not too large. This causes this ratio to experience a lack of information in predicting the value of the stock index.

Based on testing the N225 or Nikkei 225 dataset, it can be seen that the division ratio of 80:20 is the appropriate ratio in the N225 dataset. Based on the research that the author did, it can be seen that the RMSE (374), MSE (139924), MAPE (1.04), and MAE (291.7) values found the lowest error values. In addition, very high price changes make the prediction value miss, and on the graph of the data sharing ratio below 50:50, it can be seen that the orange prediction line is not shaped like the red line, namely the validation line. the predicted value of the model on the N225 dataset with ratios of 40:60, 30:70, 20:80, and 10:90. These four show that the predicted value does not appear to provide a detailed pattern of change and does not even form a pattern of change that resembles the original value, so for a ratio of 40:60, a ratio of 10:90 in this dataset is not recommended to be the ratio of train data and test data because the dataset used is not too big or too small. This causes this ratio to experience a lack of information in predicting the value of the stock index. Based on test data that was run on the KOSPI dataset, it was found that a ratio of 60:40 is more suitable than a ratio of 70:30 for RMSE and MSE values because using a ratio of 60:40 has lower RMSE and MSE values, but the difference is slight. The 70:30 ratio has lower MAPE and MAE values, so both are suitable for use in the KOSPI dataset based on this study. the predicted value of the model on the KOSPI dataset with ratios of 40:60, 30:70, 20:80, and 10:90. These four show that the predicted value does not appear to provide a detailed pattern of change and does not even form a pattern of change that resembles the original value, so for a ratio of 40:60, a ratio of 10:90 in this dataset is not recommended to be the ratio of train data and test data because the dataset used is not too big or too small. This causes this ratio to experience a lack of information in predicting the value of the stock index.

Based on model testing on the HSI dataset, the ratio of 80:20 is more suitable than 70:30 for the RMSE, MSE, and MAE values. Even though the MAPE value of the 70:30 ratio is about 0.03 lower than the 80:20 ratio, with a fairly slight difference of about 6 points in the RMSE value and about 12 points in the MAE value, for testing on the HSI dataset, the authors found a ratio of 80:20 is more suitable than the ratio of 70:30. In the graphical form of this dataset, it can be seen that the LSTM looks less predictable with the orange line, which is the prediction line that does not follow the shape of the red validation line. the predicted value of the model on the HSI dataset with ratios of 40:60, 30:70, 20:80, and 10:90. These four graphs show that the predicted value does not appear to provide a detailed pattern of change and does not even form a pattern of change that resembles the original value, so for a ratio of 40:60, a ratio of 10:90 in this dataset is not recommended to be the ratio of train data and test data because the dataset used is not too big or too small. This causes this ratio to experience a lack of information in predicting the value of the stock index.

Based on experiments on the SSE Composite dataset, it was found that the ratio of 70:30 was more suitable based on the error values of RMSE (42.07), MSE (1770), and MAPE (0.976), but based on the MAE value (33.88), the ratio of 70:30 was slightly less than the ratio of 90:10, which is about 0.5, whereas 90:10 has the MAE value (33.31). The difference in values Even so, the 70:30 ratio is more appropriate overall in testing on the SSE Composite dataset because it is superior on three other error parameters, namely RMSE, by around 3.6 out of 90:10 at a value of 45.96. MSE excels by about 400 values, where 90:10 gets a value of 2113, and MAE is around 0.02 percent, where 90:10 gets a value of 0.99, the predicted value of the model on the SSE dataset with ratios of 40:60, 30:70, 20:80, and 10:90. These four graphs show that the predicted value does not appear to provide a detailed pattern of change and does not even form a pattern of change that resembles the original value, so for a ratio of 40:60, a ratio of 10:90 in this dataset is not recommended to be the ratio of train data and test data because the dataset used is not too big or too small. This causes this ratio to experience a lack of information in predicting the value of the stock index.

Based on the prediction results outside the dataset with the actual price, it can be seen that the 80:20 ratio has an average percentage difference between the original price and the lower predicted price of around 1.26%. This makes the division of the data split accordingly overall. Based on the research that the writer did in predicting stock prices using the LSTM, the rule of thumb based on the Pareto principle is 80:20. In addition, based on this research, it can be seen that when the original value changes quickly, the accuracy of the prediction is slightly off, according to the research conducted by previous by researcher, where in this study, when the stock price changed suddenly, LSTM failed to catch the price change. This is also seen. In previous studies that explained LSTM using information from previous lags in predicting, dynamic stock markets and stock patterns were not always the same, which made it fail to capture dynamic changes accurately. In addition, this study also shows that when the data train ratio used is less than a 50:50 ratio, the predicted value of LSTM does not look too similar to this, as stated in the journals Afendras and Markatou, where the distribution of train data should be closer to 50% of the entire dataset.

Conclusion

The distribution ratio of the split train and test data is most suitable for predicting stock prices with the LSTM model on the time series data type. LSTM was found to be able to predict stock prices with a fairly low error value. In the search for the ratio of the 5 datasets used, the ratio of 60:40 and the ratio of 80:20 have the lowest RMSE values, each of which excels in 2 of the 5 test datasets. The 60:40 ratio is superior if the last test stock price data is above tens of thousands, and in this case, the two datasets where the 60:40 ratio is superior, namely HSI and N225, have a final price above 20,000 based on their respective currencies. Meanwhile, the 80:20 ratio excels in two other datasets, namely the JCI and KOSPI, which have prices below 10,000 based on their respective currencies. Whereas based on the MAPE value, the 70:30 ratio is better because this ratio gets the lowest error value in 4 out of 5 datasets, which makes the 70:30 ratio more consistent in giving the percentage error value in predicting prices. Meanwhile, based on the MAE value, the ratio of 80:20 is superior in 2 datasets to 1 with ratios of 90:10, 70:30, and 60:40; this means that the ratio of 80:20 has the lowest absolute error value compared to other ratios based on this study. Based on data forecasting 1 day after the train data and test data, it was found that the ratio of 80:20 is the most suitable ratio. This is because the ratio of 80:20 has the lowest average error rate for the difference between the original data and forecasting data compared to the ratios 70:30 and 60:40, with a price difference percentage level of 1.26% compared to a 70:30 ratio of 1.91% and a 60:40 ratio of 1.81%. Therefore, the ratio of 80:20 is the most suitable ratio for predicting stock index prices based on this research for forecasting stock prices one day later or the price the next day. Meanwhile, in the epoch and batch size tests, it was found that epoch 11 and batch size 2 were the best combinations in this study. It should be noted that this combination will not necessarily be the same on different models or hardware because this determination does not have an exact measurement tool and the values will generally be different even when run in the same place. Therefore, care needs to be taken in determining the best combination of epoch and batch size to produce the best accuracy. From this study, it can be concluded that the LSTM method can predict stock prices on stock indexes. As well as the use of a split data ratio of 80:20 to be suitable in this study based on the lowest error results in data collection for the past 5 years in the stock index dataset.

References

Hansun, S., & Young, Y. C. (2021). Predicting LQ45 Financial Sector Indices Using RNN-LSTM. *Journal of Big Data,* 8(1).

Darmawan, J., Wijaya, A. H., Hakim, L., & Tannady, H. (2021, February). Comparing Freeman Chain Code 4 Adjacency Algorithm and LZMA Algorithm in Binary Image Compression. In *Journal of Physics: Conference Series* (Vol. 1783, No. 1, p. 012045). IOP Publishing.

Supri, B., & Rudianto, R. (2017). Peranan Pajak Hotel Terhadap Pertumbuhan Ekonomi Kota Palopo. *Jurnal Ekonomi Pembangunan STIE Muhammadiyah Palopo*, *3*(2).

Hiransha, M., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. (2018). NSE Stock Market Prediction Using Deep-Learning Models. *Computer Science*, (132), 1351-1362.

Ardiansyah, M. A., & Rudianto, R. (2018). Pengembangan Dan Penerapan Teknologi Tepat Guna Pada Industri Rumahan Pembuat Produk Lokal Berbahan Dasar Sagu Di Kota Palopo. *To Maega: Jurnal Pengabdian Masyarakat*, *1*(1), 29-34.

Lestari, A., Tannady, H., & Nurprihatin, F. (2018, December). Analisis Produktivitas Kasir Guna Menentukan Beban Kerja Menggunakan Work Sampling pada Gerai Makanan Cepat Saji. In *Prosiding Seminar Rekayasa Teknologi* (pp. 578-587).

Benedicto, L. (2014). Encyclopedia of Quality of Life and Well-Being Research. *Encyclopedia of Quality of Life and Well-Being Research,* 1956–57.

Madyatmadja, E. D., Marvell, M., Andry, J. F., Tannady, H., & Chakir, A. (2021, August). Implementation of big data in hospital using cluster analytics. In *2021 International Conference on Information Management and Technology (ICIMTech)* (Vol. 1, pp. 496-500). IEEE.

Nabipour, M. P., Nayyeri, H., Jabani, A., Mosavi, E., Salwana., & Shahab, S. (2020). Deep Learning for Stock Market Prediction. *Entropy*, 22(8).

Tannady, H., Andry, J. F., Suyoto, Y. T., & Herlian, A. (2020). Business Architecture of Public Guest Service for University Using TOGAF ADM Framework. *Technology Reports of Kansai University*, *62*(5), 2421-2428.

Samsul, M. (2006). *Capital Markets and Portfolio Management*. Jakarta: Erlangga.