International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Using of R Software for GGRM Daily Stock Price Forecasting Through ARIMA Model

Edwin Setiawan Nugraha[1], Celine Alvina[2], Samsul Arifin[3*], Suwarno[4], Agus Eka Sopian Hidayat[5], Fauziah Nur Fahira Sudding[6]

Submitted: 28/04/2023 **Revised**: 29/06/2023 **Accepted**: 09/07/2023

Abstract: Stocks are one of the attractive investment instruments for companies and individuals to raise their finances. However, there are volatility of stocks made risk for the investor. Statistical tools offer a approach to predict the stocks price to minimize the risk. ARIMA(p,d,q) technical analysis will be used in this study to predict the stock price of PT. Gudang Garam Tbk. for 8 days from March 1, 2022, to March 8, 2022. This forecasting uses PT. Gudang Garam Tbk. historical stock price data from December 2021 to February 2022 was obtained from the Yahoo Finance website. Based on the test results of 24 ARIMA models, the researcher got model 2 ARIMA (4,1,3) as the best model with the equation $Y_t = -0.1389\, Y_{t-1} - 0.8033\, Y_{t-2} - 0.0575\, Y_{t-3} - 0.4440\, Y_{t-4} + e_t + 0.2217\, e_{t-1} + 0.9025\, e_{t-2} - 0.2205 e_{t-3}$. This model is the best because it has the second smallest AIC value, which is 874.42, and the smallest MAPE value, which is 3.14%. This study shows that the stock price is predicted to fall for the next five days from March 1 to March 5, 2022 and will rise again from March 6 to March 8, 2022.

Keywords: *Stocks, forecasting, arima, time series, R*

1. Introduction

Stock is a symbol of a person's or a party's (business entity's) capital investment in a corporation or limited liability company. When selecting how to fund a business, one of the options available to the company is to issue shares [1]. Investors get advantage by buying or owning shares which called dividends. Recently, many Indonesians, especially the young generation, learn about investment whether in stock, bonds, or the money market. This is very good, considering that stock investment can overcome inflation. Cited from BI Website (2020), Inflation interprets an increase in the price of goods and services in general and continuously within a certain period. For this reason, it is better not only to save in a bank that only provides 2% annual interest profit, while the annual inflation rate can be 3% to 5% [2].

Stock investment is indeed more profitable, but an investor still must be selective in choosing which stock portfolio has the potential to provide positive results. Blue-chip stocks are usually recommended because they come from well-known companies, have a good reputation, and are easy to trade on the stock exchange [3]. Shares of cigarette issuers with the largest market capitalization on the stock exchange, PT. Gudang Garam Tbk., is one of them. Gudang Garam shares are listed and traded on the Indonesia Stock Exchange (IDX) with the code GGRM. Until now, Gudang Garam has been widely known both domestically and abroad as a producer of high-quality kretek cigarettes [4]. However, lately, the government's policy to increase excise on tobacco products has become a sentiment for cigarette issuers such as PT Gudang Garam Tbk. (GGRM). The increase in cigarette excise, which took effect on February 1, 2021, made cigarette stocks even more depressed [5]. This situation also contributes to price fluctuations that shape stock volatility. Further analysis of stock price is needed so that investors can make wise decisions in their investments. Therefore, in this study, stock price fluctuations from March 1, 2022, to March 8, 2022, will be predicted using the ARIMA (Autoregressive Integrated Moving Average) model using historical stock price data from December 2021 to February 2022 [6].

Some researchers have carried out several related studies. One of the studies is the daily GGRM price per share prediction in July and August 2020 by Maulani, et al. with the best model ARIMA (2,1,2). They conducted the research in 2020 using the daily stock data from January 2019 to the first semester of 2020 and has concluded that stock prices are estimated to decline from the previous 20 days due to economic turmoil during the COVID-19 pandemic and other issues [7]. A similar study was also conducted by H. Winata and Y. D. Hapsari in 2016 using the Treshold Garch method for 2 weeks (4 January 2016–15 January 2016) with TGARCH (1,1) as the most suitable model. The empirical results of this study proved to be accurate, with a 4% level of the average forecasting error [8].

1,2,5,6 Study Program of Actuarial Science, School of Business, President University, 17550, Indonesia

3 Statistics Department, Faculty of Humanities, Bina Nusantara University. Jakarta. Indonesia.

4Primary Teacher Education Department, Faculty of Humanities, Bina Nusantara University, Jakarta, 11480, Indonesia

** Corresponding Author Email: samsul.arifin@binus.edu*

2. Methods

First, we'll go through the basics of time series. Time series analysis is a method of analyzing rows of data points collected over a period. In time series analysis, analysts don't capture data points at random or infrequent intervals, but rather at regular intervals over a set length of time [9]. Many statistical methods deal with data that are either uncorrelated or independent. Data correlation can be useful in a variety of contexts. This is especially true when multiple observations of the same system are made in a sequential order across time [10]. The study of these experimental data collected at various times in time leads to new and distinct statistical modeling and inference issues. The visible association generated by sampling neighboring points in time can significantly limit the applicability of many standard statistical methods that are based on the assumption that adjacent data are independent and uniformly distributed [11].

Time series analysis is a systematic technique to answering the mathematical and statistical concerns raised by these time connections. Time series approaches have traditionally been used to solve problems in the physical and environmental sciences. This explains why the vocabulary of time series analysis has a strong technical flavor. A detailed examination of the collected data plotted across time is always the first step in any time series analysis. Before delving more into the specific statistical procedures, it's worth noting that there are two distinct, although not necessarily mutually exclusive, approaches to time series analysis: the time domain approach and the frequency domain approach [12].

Following our discussion of Time Series, we shall move on to the topic of Iid Noise. The most basic model for a time series is one with no trend or seasonal component and data that are simply independent and identically distributed (iid) random variables with zero mean [13]. The concept of Stationary Models and the Autocorrelation Function will be discussed next. A time series $\{X_t, t = 0, \pm 1, \dots\}$ is said to be stationary if, for each integer h, it has statistical features identical to those of the "time-shifted" series $\{X_{t+h}, t = 0, \pm 1, \dots\}$. Let $\{X_t\}$ be a time series with $E(X_t^2) < \infty$. The mean function of $\{X_t\}$ is:

$$\mu_X(t) = E(X_t). \qquad (1)$$

The covariance function of $\{X_t\}$ is:

$$\begin{aligned}\gamma_X(r, s) &= \mathrm{Cov}(X_r, X_s) \qquad (2)\\ &= E[(X_r - \mu_X(r))(X_s - \mu_X(s))]\end{aligned}$$

for all integers r and s. $\{X_t\}$ is (weakly) stationary if $\mu_X(t)$ is independent of t and $\gamma_X(t + h, t)$ is independent of t for each h. Strict stationarity of a time series $\{X_t, t = 0, \pm 1, \dots\}$ is defined by the condition that $(X_1, \dots, X_n)(X_{1+h}, \dots, X_{n+h})$ have the same joint distributions for all integers h and $n > 0$. It is easy to check that if $\{X_t\}$ is strictly stationary and $E(X_t^2) < \infty$ for all t, then $\{X_t\}$ is also weakly stationary. Iid noise with finite second moment is stationary. We shall use the notation $\{X_t\} \sim \mathrm{IID}\,(0, \sigma^2)$ to indicate that the random variables X_t are independent and identically distributed random variables, each with mean 0 and variance σ^2. If $\{X_t\}$ is a sequence of uncorrelated random variables, each with zero mean and variance σ^2, then clearly $\{X_t\}$ is stationary with the same covariance function as the iid noise [14]. The concept of the Moving Average Process will be presented next. We have what is known as a moving average process when only a small proportion of the ψ-weights are nonzero.

$$Y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \qquad (3)$$

We call such a series a moving average of order q and abbreviate the name to MA(q) [7]. The topic of Autoregressive Processes will be addressed next. Autoregressive processes are regressions on themselves, as the name implies. The equation is satisfied by a pth-order autoregressive process $\{Y_t\}$ [15].

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \qquad (4)$$

The notion of ARIMA Models will be explored next. If the dth difference $W_t = \nabla^d Y_t$ is a stationary ARMA process, a time series $\{Y_t\}$ is said to follow an integrated autoregressive moving average model. If $\{W_t\}$ follows an ARMA (p, q) model, we say that $\{Y_t\}$ is an ARIMA(p, d, q) [7], [16]. With $W_t = Y_t - Y_{t-1}$, we have the equation of ARIMA (p, 1, q)

$$\begin{aligned}W_t = \phi_1 W_{t-1} + \phi_2 W_{t-2} + \cdots + \phi_p W_{t-p} + e_t\\ - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \qquad (5)\end{aligned}$$

Because ARIMA is an attempt to determine the best data pattern from a collection of data, it requires both history and current data to provide short-term forecasts. The ARIMA notation is used to express the Box-Jenkins model in general ARIMA (p, d, q). The order or degree of AR in this example is p. (Autoregressive). The order or degree of differentiation is denoted by the letter d (Differencing). MA's order or degree is q (Moving Average) [6], [15]. Note that Figure 1 below illustrates the ARIMA modeling flowchart.

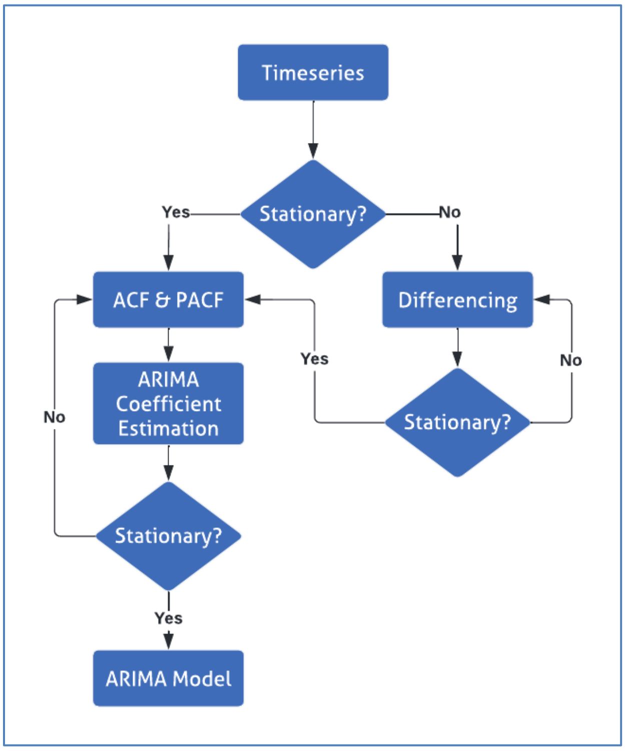


Fig 1. ARIMA modeling flowchart

(Source: private document, 2022)

The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) will be discussed next. According to the Jagostat website, the ACF and PACF graphs can help you analyze sequences and determine which model is best for time series data. The ACF plot for the AR process is expected to steadily diminish, but the PACF plot will undergo a dramatic reduction after a considerable p lag. The ARIMA process, on the other hand, should be considered for modeling if the ACF and PACF plots indicate a steady diminishing pattern. The ACF and PACF for some stationary time series data are shown in Figures 2 and 3 [17], [18]. After that, we'll go through the idea of Root Mean Square Error (RMSE). The root mean square error measures how far the forecast deviates from the true value using the Euclidean distance, which is stated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}} \tag{6}$$

where N is the number of data points, $y(i)$ is the i-th measurement, and $\hat{y}(i)$ is the corresponding prediction [19]. Next will be discussed about the concept of Mean Square Error (MSE). Mean squared error (MSE) measures the amount of error in the statistical model. The mean squared error is also known as the mean squared deviation (MSD) [20]. The formula for MSE is as follows.

$$\mathrm{MSE} = \frac{\sum(\mathrm{y(i)} - \hat{\mathrm{y}}(\mathrm{i}))^2}{\mathrm{n}} \tag{7}$$

Where $y(i)$ is the value of the i-th observation, i = the corresponding predictive value, and n = number of observations [19], [21]. Next will be discussed about the concept of Mean Absolute Error (MAE). MAE measures the absolute error between the predicted and observed values. In practice, MAE is a popular error metric because it is intuitive and easy to calculate. Mathematically, MAE is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |Actual_i - Predition_i|$$
$$= \frac{1}{n}\sum_{i=1}^{n} |y(i) - \hat{y}(i)| \tag{8}$$

where y(i) is the value of the i-th observation, i = the corresponding predictive value, and n = number of observations [22]. Next, we will discuss the concept of Mean Absolute Percent Error (MAPE). MAPE is often used in practice because of its very intuitive interpretation in terms of relative error. The use of MAPE is relevant in finance [23]. Mathematically, MAPE is defined as:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \frac{|Actual_i - Prediction_i|}{|Actual_i|}$$
$$= \frac{1}{n}\sum_{i=1}^{n} \frac{|y(i) - \hat{y}(i)|}{|y(i)|} \tag{9}$$

Where y(i) is the value of the i-th observation, i = the corresponding predictive value, and n = number of observations [24], [25].

3. Results and Discussion

In this session, we will talk about data preparation first. The data used is the daily stock price per share of PT. Gudang Garam Tbk (GGRM) from December 1, 2021, to February 25, 2022, with a total of 61 data [26]. The data is processed using R Studio. Table 1 below presents the dataset.

Table 1. PT. Gudang Garam (GGRM) Tbk's daily stock price per share.

Date	Price	Date	Price	Date	Price	Date	Price
2021-12-01	31700	2021-12-22	31075	2022-01-13	32250	2022-02-04	30550
2021-12-02	31675	2021-12-23	30825	2022-01-14	32250	2022-02-07	30650
2021-12-03	32300	2021-12-24	30625	2022-01-17	32200	2022-02-08	31000
2021-12-06	31950	2021-12-27	30525	2022-01-18	31950	2022-02-09	31000
2021-12-07	32200	2021-12-28	30675	2022-01-19	30725	2022-02-10	30625
2021-12-08	32000	2021-12-29	30800	2022-01-20	31000	2022-02-11	30525
2021-12-09	31800	2021-12-30	30675	2022-01-21	31200	2022-02-14	30500
2021-12-10	32050	2022-01-03	30600	2022-01-24	31750	2022-02-15	30800
2021-12-13	32050	2022-01-04	30675	2022-01-25	31750	2022-02-16	31000
2021-12-14	32000	2022-01-05	31175	2022-01-26	31375	2022-02-17	31000
2021-12-15	31400	2022-01-06	31350	2022-01-27	30650	2022-02-18	31400
2021-12-16	31575	2022-01-07	30800	2022-01-28	30575	2022-02-21	31800
2021-12-17	31175	2022-01-10	31125	2022-01-31	30550	2022-02-22	31500
2021-12-20	31000	2022-01-11	30950	2022-02-02	30625	2022-02-23	31150
2021-12-21	31100	2022-01-12	31150	2022-02-03	30600	2022-02-24	31100
						2022-02-25	32050

Source: https://finance.yahoo.com/quote/GGRM.JK/

Some of the packages needed in processing the data are forecast, TSA, tseries, readxl, and ggplot. First, the package forecast provides methods and tools to show univariate analysis of time series predictions including exponential smoothing through state-space models and ARIMA automatic modeling. Second, TSA is a package to call ACF, PACF, and Arima functions. Third, we use tseries to calculate the Augmented Dickey-Fuller Test (ADF Test). Fourth, readxl enable us to import xlsx data to R Studio. The last one, ggplot, gives us the ability to plot the data. The following is a data plot of the daily stock price per share of GGRM with the x-axis as the date and the y-axis as the stock price. The Figure 3 below is created using the "ggplot" function [27].

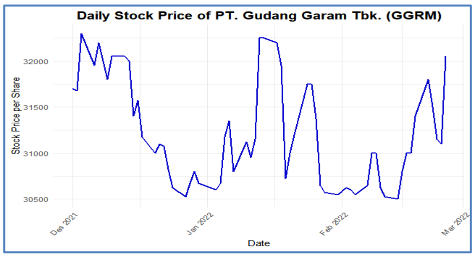
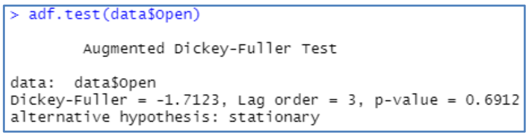


Fig 3. Graph of Stock Price Dataset of PT. Gudang Garam Tbk.

(Source: Private document, 2022)

Next will be discussed about the concept of Stationarity Check. To check the stationarity of the data, we use the Augmented Dickey-Fuller Test (ADF Test) in R Studio by calling the "adf.test(data)" function [28]. The following Figure 4 is the result of executing the code in the console.

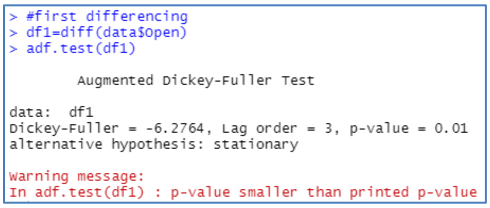


```
> adf.test(data$Open)

        Augmented Dickey-Fuller Test

data:  data$Open
Dickey-Fuller = -1.7123, Lag order = 3, p-value = 0.6912
alternative hypothesis: stationary
```

Fig 4. Console on R Studio

(Source: private document, 2022)

The ADF Test shows a p-value of 0.6912, so this data is declared not stationary because the value is greater than 0.05. We need to perform differencing to make this data stationary by calling the "diff(data)" function, then check again the stationarity using "adf.test(data)" function [29]. The following Figure 5 is the result of executing the code in the console.



```
> #first differencing
> df1=diff(data$Open)
> adf.test(df1)

        Augmented Dickey-Fuller Test

data:  df1
Dickey-Fuller = -6.2764, Lag order = 3, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(df1) : p-value smaller than printed p-value
```

Fig 5. Console on R Studio

(Source: Private document, 2022)

The first derivative has succeeded in making the p-value less than 0.05, which is 0.01. As such, this data can be said to be stationary [30]. Thus, the value of d used is 1. Here is the data plot after the first derivative in Figure 6.

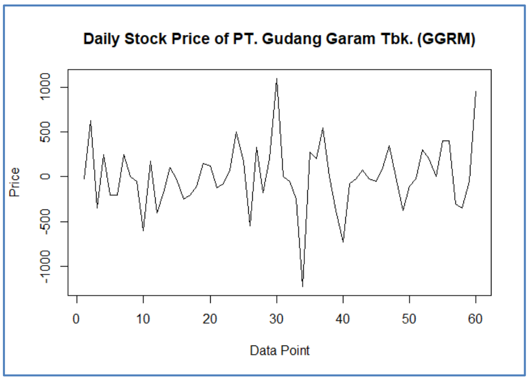


Fig 6. The First Derivative Graph of the Stock Price Dataset of PT. Gudang Garam Tbk.

(Source: Private document, 2022)

Furthermore, the concept of ARIMA Model Specifications will be discussed. The autoregressive (AR) function order value, lag p, is found using the “pacf(first derived data)” function. Meanwhile, the moving average (MA) function value can be found using the “acf(first derived data)” function. The image on the left shows the PACF plot experiencing a sharp increase after the lag p is 4, so the p value is 4. The image on the right shows the ACF plot also experienced a significant increase after the lag q, so the value of q is 4. Please see Figure 7 for details.

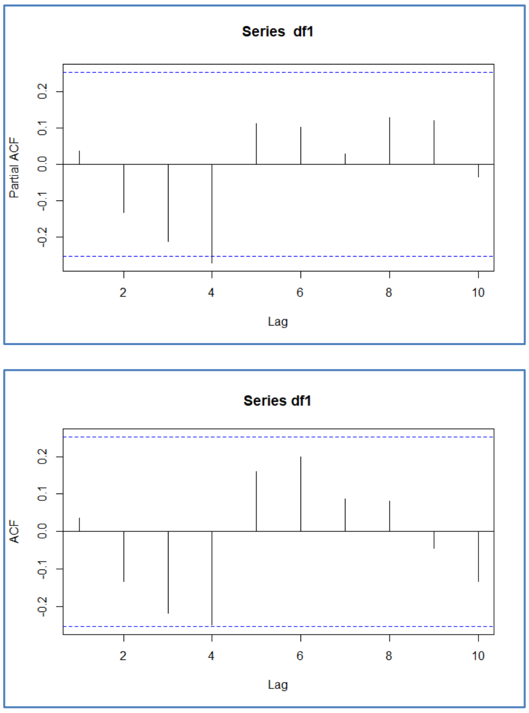


Fig 7. PACF and ACF Plot

(Source: Private document, 2022)

Therefore, this data has ARIMA(4,1,4) with the equation

$$\textbf{1. } \boldsymbol{Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \phi_4 Y_{t-4} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \theta_3 e_{t-3} - \theta_4 e_{t-4}} \tag{7}$$

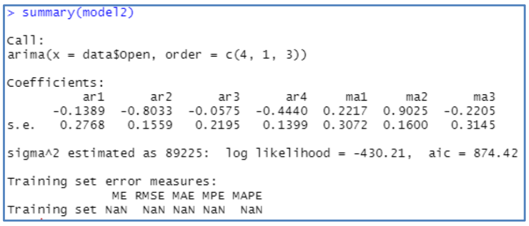
Here are some of the ARIMA models that can be built. Please see Table 2 for more details.

Table 2. ARIMA (4,1,4) model

ARIMA Model	p	d	q	ARIMA Model	p	d	q	ARIMA Model	p	d	q
ARIMA(4,1,4)	4	1	4	ARIMA(3,1,1)	3	1	1	ARIMA(1,1,3)	1	1	3
ARIMA(4,1,3)	4	1	3	ARIMA(3,1,0)	3	1	0	ARIMA(1,1,2)	1	1	2
ARIMA(4,1,2)	4	1	2	ARIMA(2,1,4)	2	1	4	ARIMA(1,1,1)	1	1	1
ARIMA(4,1,1)	4	1	1	ARIMA(2,1,3)	2	1	3	ARIMA(1,1,0)	1	1	0
ARIMA(4,1,0)	4	1	0	ARIMA(2,1,2)	2	1	2	ARIMA(0,1,4)	0	1	4
ARIMA(3,1,4)	3	1	4	ARIMA(2,1,1)	2	1	1	ARIMA(0,1,3)	0	1	3
ARIMA(3,1,3)	3	1	3	ARIMA(2,1,0)	2	1	0	ARIMA(0,1,2)	0	1	2
ARIMA(3,1,2)	3	1	2	ARIMA(1,1,4)	1	1	4	ARIMA(0,1,1)	0	1	1

Source: Private document

Next, the concept of Parameter Estimation will be discussed. Estimation of parameters ϕ_p of autoregressive (AR) function and θ_q of moving average (MA) function in each model can be found by looking at the summary of each model using “summary(model)” function. By executing this function, we can find the values of AR1 to AR4, MA1 to MA4, Means Squared Error (MSE), Log Likelihood, and AIC that will be considered in choosing the best ARIMA model. These results can be seen on the console shown in the following Figure 8.



```
> summary(model2)

Call:
arima(x = data$Open, order = c(4, 1, 3))

Coefficients:
          ar1      ar2      ar3      ar4     ma1     ma2      ma3
      -0.1389  -0.8033  -0.0575  -0.4440  0.2217  0.9025  -0.2205
s.e.   0.2768   0.1559   0.2195   0.1399  0.3072  0.1600   0.3145

sigma^2 estimated as 89225:  log likelihood = -430.21,  aic = 874.42

Training set error measures:
              ME RMSE MAE MPE MAPE
Training set NaN  NaN NaN NaN  NaN
```

Fig 8. Console on R Studio

(Source: private document, 2022)

The following is the result of parameter estimation for each model. Please see Table 3 for more details.

ARIMA Model	Parameter Estimation Results										
	AR1	AR2	AR3	AR4	MA1	MA2	MA3	MA4	MSE	Log Likelihood	AIC
ARIMA(4,1,4)	-0.1776	-0.7783	-0.0761	-0.4211	0.2649	0.8755	-0.1931	-0.0496	89178	-430.2	876.39
ARIMA(4,1,3)	-0.1389	-0.8033	-0.0575	-0.4440	0.2217	0.9025	-0.2205		89225	-430.21	874.42
ARIMA(4,1,2)	-0.3085	-0.8576	-0.1770	-0.4517	0.4285	1.0000			89422	-430.4	872.79
ARIMA(4,1,1)	-0.2880	-0.1747	-0.2530	-0.3783	0.2755				110679	-433.91	877.83
ARIMA(4,1,0)	-0.0438	-0.1761	-0.2199	-0.3067					112300	-434.31	876.61
ARIMA(3,1,4)	-0.3545	-0.4050	-0.0820		0.4432	0.5256	-0.1493	-0.4927	94550	-431.39	876.78
ARIMA(3,1,3)	0.6316	0.1397	-0.5816		-0.8528	-0.3140	0.7958		98466	-432.28	876.56
ARIMA(3,1,2)	1.4894	-1.0277	0.1605		-1.6693	1.0000			97430	-432.07	874.14
ARIMA(3,1,1)	0.3521	-0.1489	-0.2265		-0.3702				119165	-435.95	879.9
ARIMA(3,1,0)	0.0188	-0.1340	-0.2414						122994	-436.85	879.69
ARIMA(2,1,4)	-0.2870	-0.3784			0.3834	0.5012	-0.2210	-0.4897	94785	-431.44	874.88
ARIMA(2,1,3)	-0.1928	0.7434			0.1416	-0.9999	-0.1417		116294	-436.49	882.98
ARIMA(2,1,2)	0.1787	0.4976			-0.2583	-0.7417			118788	-436.72	881.45
ARIMA(2,1,1)	-0.6909	-0.0527			0.7437				131105	-438.67	883.35
ARIMA(2,1,0)	0.0464	-0.1550							129872	-438.39	880.78
ARIMA(1,1,4)	-0.1595				0.2868	0.2077	-0.4108	-0.7474	97890	-433.18	876.36
ARIMA(1,1,3)	0.7127				-0.7766	-0.1245	-0.0988		118416	-436.74	881.48
ARIMA(1,1,2)	0.7662				-0.8331	-0.1669			118753	-436.76	879.52
ARIMA(1,1,1)	-0.9559				0.9999				129481	-438.7	881.4
ARIMA(1,1,0)	0.0388								132754	-439.02	880.05
ARIMA(0,1,4)					0.1579	0.1406	-0.5277	-0.7705	95328	-433.39	874.79
ARIMA(0,1,3)					-0.0504	-0.1705	-0.1390		125290	-437.35	880.71
ARIMA(0,1,2)					-0.0425	-0.2706			127209	-437.82	879.64
ARIMA(0,1,1)					0.0526				132683	-439.01	880.02

Table 3. Parameter estimation results

Source: Private document

Next will be discussed about the concept of Residual Analysis. In determining the ARIMA model that gives the best prediction, the normality test with Saphiro and white noise test with Ljung-Box are needed. We use the function “saphiro.test(residuals(model))” and “Box.test(residuals(model),type="Ljung-Box")” to find the p-value. The model that has a p-value of more than 0.05 from the two tests will be chosen as the best model. Here is the code in the R Studio. The following Table 4 is each model value summary.

Table 4. Residual analysis results

ARIMA Model	Saphiro Test	Ljung-Box Test	AIC	Accepted	ARIMA Model	Saphiro Test	Ljung-Box Test	AIC	Accepted
ARIMA(4,1,4)	0.4315	0.9924	876.39	Yes	ARIMA(2,1,2)	0.02395	0.9718	881.45	No
ARIMA(4,1,3)	0.399	0.9881	874.42	Yes	ARIMA(2,1,1)	0.03242	0.9725	883.35	No
ARIMA(4,1,2)	0.5238	0.8383	872.79	Yes	ARIMA(2,1,0)	0.03218	0.7973	880.78	No
ARIMA(4,1,1)	0.06306	0.9179	877.83	Yes	ARIMA(1,1,4)	0.2364	0.8632	876.36	Yes
ARIMA(4,1,0)	0.08107	0.7372	876.61	Yes	ARIMA(1,1,3)	0.01782	0.8619	881.48	No
ARIMA(3,1,4)	0.208	0.9671	876.78	Yes	ARIMA(1,1,2)	0.01986	0.889	879.52	No
ARIMA(3,1,3)	0.004112	0.7549	876.56	No	ARIMA(1,1,1)	0.03708	0.8996	881.4	No
ARIMA(3,1,2)	0.002407	0.8795	874.14	No	ARIMA(1,1,0)	0.02305	0.9601	880.05	No
ARIMA(3,1,1)	0.01644	0.8521	879.9	No	ARIMA(0,1,4)	0.2392	0.4889	874.79	Yes
ARIMA(3,1,0)	0.00911	0.6124	879.69	No	ARIMA(0,1,3)	0.01381	0.8625	880.71	No

ARIMA(2,1,4)	0.2091	0.9564	874.88	Yes	ARIMA(0,1,2)	0.03569	0.8134	879.64	No
ARIMA(2,1,3)	0.02816	0.9033	882.98	No	ARIMA(0,1,1)	0.02452	0.97	880.02	No

Source: Private document

Models that satisfy the Saphiro and Ljung-Box tests are Model 1 ARIMA(4,1,4), Model 2 ARIMA(4,1,3), Model 3 ARIMA(4,1,2), Model 4 ARIMA(4,1 ,1), Model 5 ARIMA(4,1,0), Model 6 ARIMA(3,1,4), Model 11 ARIMA(2,1,4), Model 16 ARIMA(1,1,4), and Model 21 ARIMA(0,1,4). To find the best model, we need to compare the four error value estimation parameters. Here is the comparison. Looking at the AIC value, model 3 and model 2 have the smallest value, thus they will be the candidates to be compared [31]. Please see Table 5 for more details.

Table 5. Error value estimation parameters comparison between model 1 and model 2

ARIMA Model	MSE	RMSE	MAE	MAPE
Model 2 ARIMA(4,1,3)	1,355,189.32	1,164.13	1,018.04	3.30%
Model 3ARIMA(4,1,2)	1,244,219.69	1,115.45	970.37	3.14%

Source: Private document

From these models, the best model is Model 2 because it has the smallest value on all error value estimation parameters and the second smallest AIC compared to other models. Hence, the equation of model 2 ARIMA(4,1,3) is

$$\textbf{1. } Y_t = -0.1389\, Y_{t-1} - 0.8033\, Y_{t-2} - 0.0575\, Y_{t-3} - 0.4440\, Y_{t-4} + e_t + 0.2217\, e_{t-1} + 0.9025\, e_{t-2} - 0.2205 e_{t-3} \quad (34)$$

Here is the plot of Model 2. Please see Figure 9 for details.

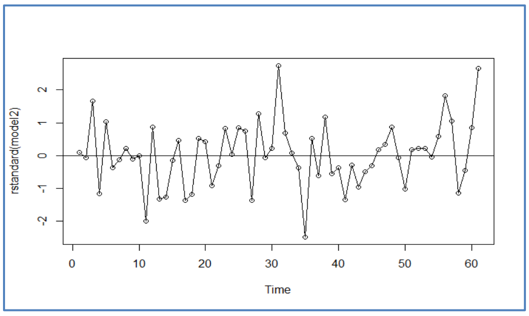


Fig 9. Model 2 plot
(Source: Private document, 2022)

To visually prove that the residual model 2 follows a normal distribution, we use "qqnorm(residuals(model))" and "qqline(residuals(model2))" function. Here is the visualization in Figure 10 below.

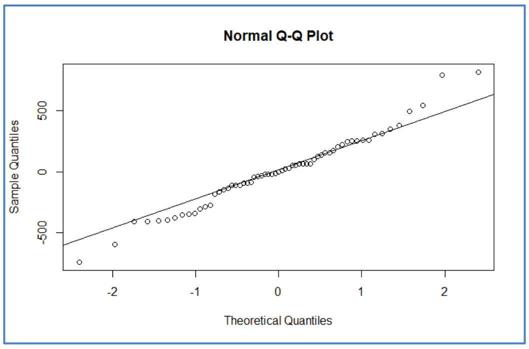


Fig 10. Model 2 Q-Q Normal Plot

(Source: Private document, 2022)

The Figure 10 above can prove that the residual model 2 follows a normal distribution because the points on the graph approach a straight line. Figure 11 below is a plot of the residual ACF and PACF model 2.

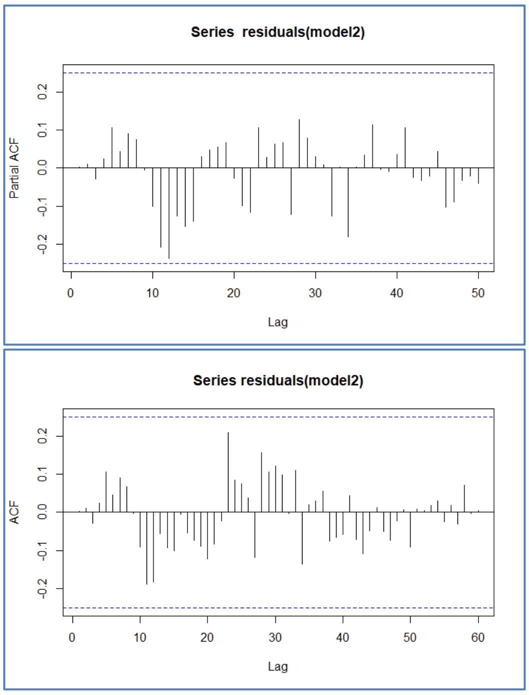


Fig 11. Model 2 PACF and ACF Plot
(Source: Private document, 2022)

Next will be discussed about the concept of Forecasting. The following Figure 12 is the plot of the forecasted daily stock price per share of GGRM for 8 days, starting from March 1, 2022, to March 8, 2022. The blue line represents the predicted data, while the gray line represents the actual data.

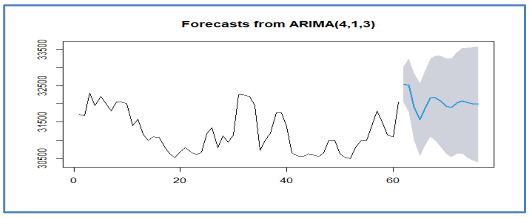


Fig 12. Plot of Stock Price Prediction Results of PT. Gudang Garam Tbk.

(Source: Private document, 2022)

Next will be discussed about Comparison. The Table 6 below shows the comparison between the actual data and the predicted data, and the error value is calculated using four parameters: MSE, RMSE, MAE, and MAPE [32]. Table 6 will close this session.

Table 6. Comparison of actual data and predicted data of model 2 ARIMA(4,1,3)

Date	Lower Bound	Upper Bound	Predicted Data	Actual Data	MSE	RMSE	MAE	MAPE
2022-03-01	32042,18	33040,47	32541,33	31700	707.836,17	707.836,17	841,33	2,65%
2022-03-02	31785,23	33251,55	32518,39	31400	1.250.796,19	1.250.796,19	1.118,39	3,56%
2022-03-03	30995,33	32848,03	31921,68	31900	470,02	470,02	21,68	0,07%
2022-03-04	30565,34	32580,65	31573,00	31300	74.529,00	74.529,00	273,00	0,87%
2022-03-05	30861,60	32906,33	31883,96	30650	1.522.657,28	1.522.657,28	1.233,96	4,03%
2022-03-06	31093,54	33237,15	32165,34	30525	2.690.715,32	2.690.715,32	1.640,34	5,37%
2022-03-07	30988,03	33334,81	32161,42	30500	2.760.316,42	2.760.316,42	1.661,42	5,45%
2022-03-08	30825,04	33320,66	32072,85	31100	946.437,12	946.437,12	972,85	3,13%
				Results	1.244.219,69	1.115,45	970,37	3,14%

(Source: Private document, 2022)

4. Conclusion

Based on the analysis from the previous chapter on the historical data of the daily stock prices of PT. Gudang Garam Tbk. from December 2021 to February 2022 shows that the best model is model 2 ARIMA (4,1,3) with the equation

$$1.\ \boldsymbol{Y_t = -0.1389\, Y_{t-1} - 0.8033\, Y_{t-2} - 0.0575\, Y_{t-3} - 0.4440\, Y_{t-4} + e_t + 0.2217\, e_{t-1} + 0.9025\, e_{t-2} - 0.2205 e_{t-3}} \quad (34)$$

The model is considered the best because it has the second smallest AIC value, which is 874.42, and the smallest MAPE value, which is 3.14%. The stock price is predicted to fall for the next 5 days from March 1 to March 5, 2022, and will rise again from March 6 to March 8, 2022.

Acknowledgements

This study is supported and funded by Bina Nusantara University Research and Technology Transfer Office under the terms of the university's International Research Grant (PIB 2023) under contract number 029/VRRTT/III/2023.

Author contributions

Edwin Setiawan Nugraha, Celine Alvina: Conceptualization, Methodology, Software, Field study; **Samsul Arifin, Suwarno:** Data curation, Writing-Original draft preparation, Software, Validation., Field study; **Agus Eka Sopian Hidayat, Fauziah Nur Fahira Sudding:** Visualization, Investigation, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] B. E. Indonesia, “Saham,” 2018. https://www.idx.co.id/produk/saham/

[2] Revina, “Alasan Investasi Saham Penting untuk Dilakukan Sejak Dini,” *Ajaib*, 2020. https://ajaib.co.id/alasan-investasi-saham-penting-untuk-dilakukan-sejak-dini/

[3] Z. Afdika, “14 Saham Blue Chip Terbaik di Indonesia pada 2021 Yang Layak dibeli,” *Qoala*, 2021. https://www.qoala.app/id/blog/keuangan/investasi/saham-blue-chip-terbaik/

[4] Web Team, “PT Surya Madistrindo,” 2002. https://www.gudanggaramtbk.com/suryamadistrindo/ (accessed Jun. 15, 2020).

[5] E. Asror, “Lis Lengkap Saham Emiten Rokok Berikut Profil Mininya,” 2021. https://bigalpha.id/news/lis-lengkap-saham-emiten-rokok-berikut-profil-mininya

[6] G. P. Zhang, “Time series forecasting using a hybrid ARIMA and neural network model,” vol. 50, pp. 159–175, 2003.

[7] D. Maulani and D. Riani, “Autoregressive Integrated Moving Average (ARIMA) pada Industri Manufaktur di Bursa Efek Indonesia (STUDI KASUS PT.

GUDANG GARAM Tbk.),” *Pros. LPPM UIKA BOGOR; 2020*, 2020, [Online]. Available: http://pkm.uika-bogor.ac.id/index.php/prosiding/article/view/624

[8] H. Winata and Y. D. Hapsari, “PENGGUNAAN METODE TRESHOLD GARCH DALAM MEMPREDIKSI HARGA SAHAM PT. GUDANG GARAM, Tbk.,” *Optim. J. Ekon. dan Pembang.*, vol. 7, no. 1, p. 59, 2017, doi: 10.12928/optimum.v7i1.7893.

[9] Tableau, “Time Series Analysis: Definition, Types, Techniques, and When It’s Used.” https://www.tableau.com/learn/articles/time-series-analysis

[10] S. Arifin, I. B. Muktyas, and J. M. Mandei, “Graph coloring program for variation of exam scheduling modeling at Binus University based on Welsh and Powell algorithm,” in *International Conference Advanced in Applied Mathematics (ICAAM 2021)*, 2022, pp. 1–5. doi: 10.1088/1742-6596/2279/1/012005.

[11] M. Žáček, “Introduction to Time Series,” 2017, pp. 32–52. doi: 10.4018/978-1-5225-0565-5.ch002.

[12] H. Song, “Review of Time Series Analysis and Its Applications With R Examples (3rd Edition), by Robert H. Shumway & David S. Stoffer,” *Struct. Equ. Model. A Multidiscip. J.*, vol. 24, no. 5, pp. 800–802, Sep. 2017, doi: 10.1080/10705511.2017.1299578.

[13] H. Syafwan, M. Syafwan, E. Syafwan, and A. F. Hadi, “Forecasting Unemployment in North Sumatra Using Double Exponential Smoothing Method Forecasting Unemployment in North Sumatra Using Double Exponential Smoothing Method,” 2021, doi: 10.1088/1742-6596/1783/1/012008.

[14] H. Wang and K. Li, “Resistance of IID Noise in Differentially Private Schemes for Trajectory Publishing,” *Comput. J.*, vol. 63, no. 4, pp. 549–566, Apr. 2020, doi: 10.1093/comjnl/bxz097.

[15] M. Lefebvre, “The homing problem for autoregressive processes,” *IMA J. Math. Control Inf.*, vol. 39, no. 1, pp. 322–344, Mar. 2022, doi: 10.1093/imamci/dnab047.

[16] J. Li, J. Cai, R. Li, Q. Li, and L. Zheng, “Wavelet transforms based ARIMA-XGBoost hybrid method for layer actions response time prediction of cloud GIS services,” *J. Cloud Comput.*, vol. 12, no. 1, pp. 1–17, 2023.

[17] S. Bocquet, “Ocean wave autocorrelation function,” *Appl. Math. Comput.*, vol. 426, p. 127114, 2022, doi: https://doi.org/10.1016/j.amc.2022.127114.

[18] Y. Zhang, X. Wang, S. Cui, J. Zhang, and J. Su, “Study on growth prediction of Haematococcus pluvialis based on ARIMA model,” in *2022 7th International Conference on Multimedia and Image Processing*, 2022, pp. 177–181.

[19] W. Wang and Y. Lu, “Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 324, p. 12049, 2018, doi: 10.1088/1757-899x/324/1/012049.

[20] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, “Applications of support vector machine (SVM) learning in cancer genomics,” *Cancer genomics and proteomics*, vol. 15, no. 1, pp. 41–51, 2018.

[21] D. Fan, H. Sun, J. Yao, K. Zhang, X. Yan, and Z. Sun, “Well production forecasting based on ARIMA-LSTM model considering manual operations,” *Energy*, vol. 220, p. 119708, 2021.

[22] B. H. Farizan, A. G. Putrada, and R. R. Pahlevi, “Analysis of Support Vector Regression Performance in Prediction of Lettuce Growth for Aeroponic IoT Systems,” in *2021 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS)*, 2021, pp. 1–6. doi: 10.1109/ICADEIS52521.2021.9702093.

[23] A. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, “Mean Absolute Percentage Error for regression models,” *Neurocomputing*, vol. 192, pp. 38–48, 2016, doi: https://doi.org/10.1016/j.neucom.2015.12.114.

[24] N. M. Vural, F. Ilhan, S. F. Yilmaz, S. Ergüt, and S. S. Kozat, “Achieving Online Regression Performance of LSTMs With Simple RNNs,” *IEEE Trans. Neural Networks Learn. Syst.*, pp. 1–12, 2021, doi: 10.1109/TNNLS.2021.3086029.

[25] A. M. Elshewey, M. Y. Shams, Z. Tarek, M. Megahed, E.-S. M. El-Kenawy, and M. A. El-Dosuky, “Weight Prediction Using the Hybrid Stacked-LSTM Food Selection Model,” *Comput. Syst. Sci. Eng.*, vol. 46, no. 1, pp. 765–781, 2023, doi: 10.32604/csse.2023.034324.

[26] S. Arifin and I. B. Muktyas, “Generate a system of linear equation through unimodular matrix using Python and Latex,” in *AIP Conference Proceedings*, American Institute of Physics Inc., Apr. 2021. doi: 10.1063/5.0041651.

[27] A. Ebert, P. Wu, K. Mengersen, and F. Ruggeri, “Computationally Efficient Simulation of Queues: The R Package queuecomputer,” *J. Stat. Softw.*, vol. 95, no. 5 SE-Articles, pp. 1–29, Oct. 2020, doi: 10.18637/jss.v095.i05.

[28] W. N. Venables, D. M. Smith, and R. C. Team, “An introduction to R, Notes on R: A Programming Environment for Data Analysis and Graphics Version 3.6. 3.” R Foundation for Statistical Computing Vienna, Austria, 2020.

[29] I. G. A. A. Yudistira, “Pengembangan Simulasi Kejadian Diskret Berbasis Paket Simmer pada R,”

Eng. Math. Comput. Sci. J., vol. 3, no. 2, pp. 79–85, 2021, doi: 10.21512/emacsjournal.v3i2.7386.

[30] K. M. Ramachandran and C. P. Tsokos, *Mathematical statistics with applications in R*. Academic Press, 2020.

[31] S. Tarigan, N. P. Murnaka, and S. Arifin, “Development of teaching material in mathematics ‘Sapta Maino Education’ on topics of plane geometry,” in *AIP Conference Proceedings*, American Institute of Physics Inc., Apr. 2021, p. 020003. doi: 10.1063/5.0041650.

[32] A. A. Abdillah, Azwardi, S. Permana, I. Susanto, F. Zainuri, and S. Arifin, “Performance Evaluation Of Linear Discriminant Analysis And Support Vector Machines To Classify Cesarean Section,” *Eastern-European J. Enterp. Technol.*, vol. 5, no. 2–113, pp. 37–43, 2021, doi: 10.15587/1729-4061.2021.242798.

[33] Babu, D. R. ., & Sathyanarayana, B. . (2023). Design and Implementation of Technical Analysis Based LSTM Model for Stock Price Prediction. International Journal on Recent and Innovation Trends in Computing and Communication, 11(4s), 01–07. https://doi.org/10.17762/ijritcc.v11i4s.6301

[34] Dhingra, M., Dhabliya, D., Dubey, M. K., Gupta, A., & Reddy, D. H. (2022). A Review on Comparison of Machine Learning Algorithms for Text Classification. 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), 1818–1823. IEEE.

[35] Ms. Mohini Dadhe, Ms. Sneha Miskin. (2015). Optimized Wireless Stethoscope Using Butterworth Filter. International Journal of New Practices in Management and Engineering, 4(03), 01 - 05. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/37