Bit-Tech

Vol.5, No.2, December 2022
Available online at: http://jurnal.kdi.or.id/index.php/bt

# Implementation of Linear Regression Algorithm to Predict Stock Prices Based on Historical Data

**Jelvin Putra Halawa[1)*], Aditiya Hermawan[2)], Junaedi[3)]**

*[1)2)3)]Universitas Buddhi Dharma*
*Jl.Imam Bonjol No. 41, Tangerang, Indonesia*
[1)]jelvin99.halawa@gmail.com
[2)]aditiya.hermawan@ubd.ac.id
[3)]junaedi@ubd.ac.id

*Article history:*

Received 22 Nov 2022;
Revised 25 Nov 2022;
Accepted 13 Dec 2022;
Available online 14 Dec 2022

*Keywords:*

Historical Data
Investment
Linear Regression
Prediction
Stock Price

***Abstract***

Stock investment is in great demand by investors because it can provide large profits with large risks or losses, in accordance with the investment principle of low risk low return, high risk high return. Stock prices that fluctuate in a very short time make it difficult for investors to predict stock prices in the future, so investors must pay more attention and gather as much information as possible regarding the shares to be bought or sold. This study aims to create a data mining model using a Linear Regression algorithm that can predict daily stock closing prices to provide information that supports investors in stock transactions. The data used is historical data on daily stock prices for 10 companies in the last 8 years for the period 25 February 2013 – 25 February 2021. Historical stock price data will be prepared using the Noving Average method and create a data mining model using the linear regression method to generate stock price prediction models. The resulting model can be used to predict stock prices well enough to assist investors in making investment decisions to obtain large profits with low risk.

## I. Introduction

In Indonesia, investment is currently experiencing very rapid development. Indonesian people's interest in the capital market increased in 1 year, where the Indonesian people became one of the most active in the capital market within 1 year, with a total Single Investor Identification (SID) as of March 2020 reaching 2,679,039 SID, growing 44% from last year (YoY) [1].

One of the investments that are in great demand by investors in the capital market is stock investment. Shares are certificates proof of ownership of a company with a nominal value listed, with the holder being given rights and obligations in accordance with what has been explained by the company [2].

One of the things that need to be considered in investing in stocks is the stock price. The share price is the share price when the market is in progress as well as when the market closes [3]. However, investors still have doubts about investing in stocks because stock prices fluctuate which can go up or down in a very short time. To assist investors in seeing investment prospects in the future, as well as making decisions to buy or sell shares according to the application of forecasting or prediction science.

Stock price prediction means seeing and predicting the value of stock prices in the future by taking into account the value of stock prices in the past and present. Prediction of stock prices can be done by utilizing past data or time series data (time series).

In the field of information technology, there is one technique that can be used to make predictions with the help of past data, namely Data Mining. Stock price time series data will be mined to produce a model that is used to predict stock prices in the future. The model or pattern that is obtained must provide benefits, especially economic benefits [4].

[*] Corresponding author

ISSN 2622-2728 (online) 2622-271X (print)
doi: 10.32877/bt.v5i2.616

In this study the method used to predict the data is linear regression. The linear regression method was chosen because it is able to estimate simple model parameters and the stock data used has several appropriate data conditions, namely the data used are historical data or certain time series, the form of the data is numerical and can be estimated to have a past pattern that may continue in the future [5].

Based on the background of the problems discussed above, the authors conducted research by making "Implementation of a Linear Regression Algorithm to Predict Stock Prices Based on Historical Data" which is expected to be able to help investors make the right decisions in stock investment.

## II. METHODS

### Metode Penelitian

The data mining method used in this research is *CRISP-DM (Cross Industry Standart Process for Data Mining)* with the following stages:

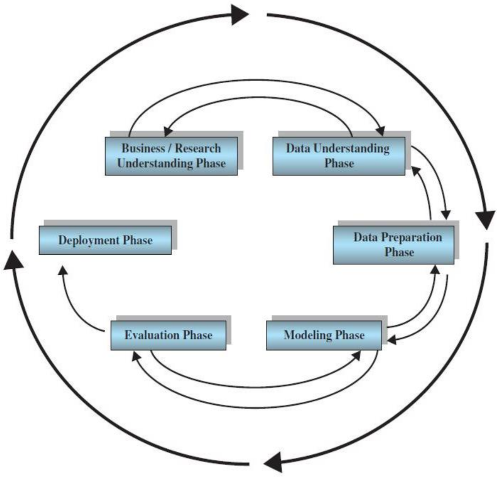


Fig 1. Process Stages *CRISP-DM* [6]

The data mining method used in this study is CRISP-DM (Cross Industry Standard Process for Data Mining) with the following stages :

a. *Business Understanding*
The stage of understanding the objectives from a business perspective, then the information is converted into a problem definition that will be answered by data mining. The goal is to predict stock prices to get big profits and reduce the risk of loss and to help make decisions to buy or sell stocks. By utilizing data mining, stock price equation patterns are searched using the Linear Regression algorithm. The pattern that can be used to predict stock prices in the future.

b. *Understanding Data*
The initial data collection stage, data exploration, and identifying problems related to data quality. In this study, the data used is stock price time series data sourced from the website https://www.investing.com. The data taken is already in excel form.

c. *Preparation Data*
The stage of compiling the final dataset that will be used as input in the modeling stage. Stock price time series data is transformed into a simple moving average of 5, then transformed into a percentage. The attributes used are open, high, low, and close as labels.

d. *Modeling*
The stage of selecting and implementing the data mining modeling techniques to be used. This stage is the stage of using the Linear Regression algorithm to find patterns of equations according to predetermined data.

e. *Evaluation*
The results evaluation stage of the data mining process. At the end of this stage, a decision is made whether the results of the data mining process will be used or not. In this study, testing was carried out by calculating the RMSE, MAPE, Correlation, Confusion Matrix, and Back Testing values.

f. *Deployment*

The stage of using the results of the data mining process. Compile and present patterns or models of the results of the data mining process. The stock price equation pattern that has been prepared can be compiled and inputted into a web-based application that is created so that it can be used to predict future stock prices.

**Data Collection Technique**

a. Secondary Data
The type of data used in this study is secondary data, which is obtained indirectly through intermediary media. The author obtains data from the website https://www.investing.com.

b. Literature Review
Collect information that supports or is relevant to research from various sources such as books, journals, articles, similar literature, internet sites and other sources needed in this research.

**Linear Regression**

Linear regression algorithm is a statistical method used to determine the effect of one or several variables on one variable. Variables are quantities that change in value. Variables that influence are called independent variables, independent variables, or explanatory variables. The variable that is affected is called the dependent variable or the dependent variable [7].

In the regression analysis, there are 2 types of variables, namely [8]:

1. The dependent variable is the dependent variable because the variable whose existence is influenced by other variables has the nature of not being able to stand alone and is expressed by the symbol Y.
2. The independent variable is a variable that is independent because it is not affected by other variables that have independent characteristics and is represented by the symbol X.

Linear regression can be divided into 2 forms :

1. Simple analysis regression or simple regression analysis. The general form is:
   $Y = a + bX$ .................. i
2. Multiple regression analysis or multiple regression analysis The general form is:
   $Y' = \boldsymbol{a + b_1X_1 + b_2X_2 + \ldots + b_nX_n}$ ... ii

**Moving Average**

Moving average is one of the methods in the time series method. The moving average method is often used in forecasting by determining the number of periods (T) to be used in observations, then using the results of the average value as a forecast in the future. The use of moving averages aims to eliminate or reduce randomness in the time series by averaging several data values so that the possibility of positive and negative errors can be eliminated or removed [9].

Simple Moving Average

The Simple Moving Average or abbreviated as SMA is a Moving Average that does not use weights in calculating data values so that it can be said to be the simplest Moving Average. Despite its simplicity, SMA is very effective in identifying trends that occur from a time series. The way to read it is quite simple [10]

## III. Results

### *A. Analysis*

**Business Understanding**

This study aims to find patterns of stock price equations using Linear Regression algorithms assisted by past data. The pattern that can be used to predict stock prices in the future so that it can help investors get big profits and reduce the risk of loss and to assist decision making and become a benchmark for investors whether they want to buy or sell shares of a particular company.

**Understanding Data**

The initial data collection stage is sourced from the website https://www.investing.com. The data used is time series data on the stock prices of 10 companies with the largest market capitalization in the Jakarta Composite Index (IHSG) for the period 25 February 2013 – 25 February 2021, including the following :

1. Astra International Tbk. by stock code ASII

2. Bank Central Asia Tbk. by stock code BBCA
3. Bank Negara Indonesia (Persero) Tbk. by stock code BBNI
4. Bank Rakyat Indonesia (Persero) Tbk. by stock code BBRI
5. Bank Mandiri (Persero) Tbk. by stock code BMRI
6. PT. Chandra Asri Petrochemical Tbk. by stock code TPIA
7. H.M. Sampoerna Tbk. by stock code HMSP
8. Indofood CBP Sukses Makmur Tbk. by stock code ICBP
9. Telkom Indonesia (Persero) Tbk. by stock code TLKM
10. Unilever Indonesia Tbk. by stock code UNVR

From the stock price dataset, several variables are obtained, including:

TABLE 1
EXPLANATION OF VARIABLES

| No. | Variable | Information |
|---|---|---|
| 1 | Date | Stock exchange transaction date |
| 2 | Open | Stock price at the time the stock exchange is open |
| 3 | High | The highest price that occurs on exchange days |
| 4 | Low | The lowest price that occurs on exchange days |
| 5 | Close | Stock price at the end of the trading day |
| 6 | Volume | The number of shares traded on an exchange day |

**Preparation Data**
The stage of compiling the final dataset that will be used as input in the modeling stage.

**a. *Selection Data***

At this stage, the attributes that were not used in the study were removed, namely the Date, Volume, and % Change attributes. The attributes used are open, high, low, and close and sort the data from newest to oldest.

TABLE 2
THE DATASET AFTER BEING SELECTED

| Open | High | Low | Close |
|---|---|---|---|
| 5700 | 5700 | 5475 | 5575 |
| 5775 | 5850 | 5600 | 5625 |
| 5700 | 5825 | 5700 | 5725 |
| 5825 | 5850 | 5700 | 5700 |
| 5725 | 5850 | 5725 | 5775 |
| 5750 | 5875 | 5725 | 5725 |
| 5900 | 5925 | 5750 | 5800 |

**b. *Simple Moving Average* (5)**

Data open, high, low, and close is transformed into a simple moving average of 5 by taking the average value of 5 days from these data.

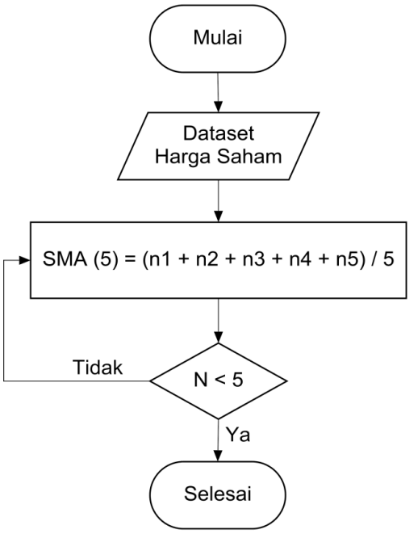


Fig 2. Flowchart Simple Moving Average 5

### c. *Transformation Data*

SMA(5) data that has been obtained from the stock price time series data is transformed into a percentage as follows:

1. *Open SMA(5)* : Percentage of the opening price of the SMA(5) and the opening price of SMA(5) the previous day's.
2. *High SMA(5)* : Percentage of the highest SMA(5) and the highest price of SMA(5) the previous day
3. *Low SMA(5)* : Percentage of the previous day's SMA(5) low and the previous day's SMA(5) low
4. *Close SMA(5)* : Percentage of closing price of SMA(5) and closing price of SMA(5) the next day

### d. *Integration Data*

At this stage data integration is carried out, namely bringing together 10 high school (5) excel data files from each issuer that have been transformed into percentages into one excel file so that it can be used as input in the modeling stage.

TABLE 3
TOTAL OF DATA IN THE DATASET FILE

| **File** | **Data** |
|---|---|
| ASII | 1938 |
| BBCA | 1938 |
| BBNI | 1938 |
| BBRI | 1938 |
| BMRI | 1938 |
| HMSP | 1892 |
| ICBP | 1938 |
| TLKM | 1938 |
| TPIA | 1601 |
| UNVR | 1938 |
| **Total** | 18997 |

### e. *Cleaning Data*

At this stage, outlier data is removed to avoid inconsistent data. The outlier data referred to is the percentage of Close SMA(5) which increases and decreases by more than 10% which will be identified as outlier data.

TABLE 4
STOCK PRICE DATASETS

| Open SMA(5) % | High SMA(5) % | Low SMA(5) % | Close SMA(5) % |
|---|---|---|---|
| -0.4325 | -0.2557 | -0.5244 | -0.5253 |
| -0.9425 | -0.5932 | -0.6944 | -0.6092 |
| -0.9337 | -1.3377 | -0.8605 | -0.6914 |
| -0.5067 | -0.3333 | -0.3430 | -0.8568 |
| -0.3367 | -0.3322 | -0.3418 | -0.2564 |

## *B. Design*

### Linear Regression Algorithm

In the linear regression algorithm, a model is formed to perform data processing which is described using a flowchart. In simple terms, here is a description of the Linear Regression flowchart :

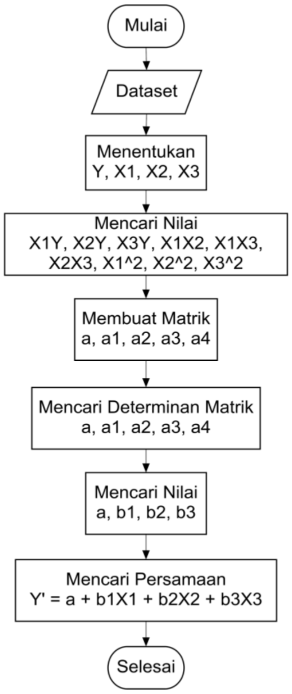

Mulai
Dataset
Menentukan Y, X1, X2, X3
Mencari Nilai X1Y, X2Y, X3Y, X1X2, X1X3, X2X3, X1^2, X2^2, X3^2
Membuat Matrik a, a1, a2, a3, a4
Mencari Determinan Matrik a, a1, a2, a3, a4
Mencari Nilai a, b1, b2, b3
Mencari Persamaan Y' = a + b1X1 + b2X2 + b3X3
Selesai

Fig 3. *Linear Regression Flowchart*

### Modelling

The stage uses the Linear Regression algorithm to find equation patterns according to predetermined data. Linear regression calculations get the following results:

b0 = 0.010
b1 = -0.629
b2 = 0.518
b3 = 0.607

So the linear regression equation is :

$$Y' = 0.010 - 0.629X_1 + 0.518X_2 + 0.607X_3 \text{ ..................iii}$$

To find out whether the results of manual calculations are appropriate or not, calculations are performed using the RapidMiner application. The following are the results obtained using RapidMiner :

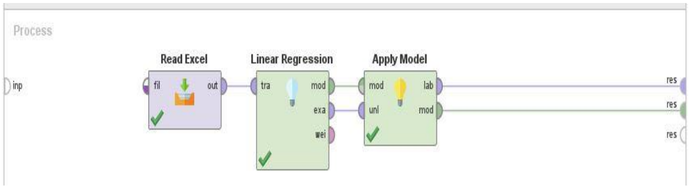
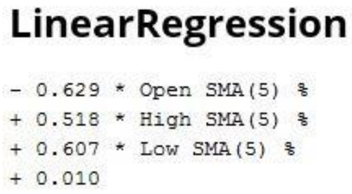


Fig 4. Linear Regression Process on RapidMiner



LinearRegression

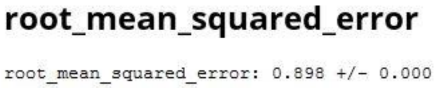
```
- 0.629 * Open SMA(5) %
+ 0.518 * High SMA(5) %
+ 0.607 * Low SMA(5) %
+ 0.010
```

Fig 5. Linear Regression Model on RapidMiner

### *C. Evaluation*

At this stage, testing is done by calculating the Root Mean Square Error (RMSE), which is the difference or the sum of the squared prediction errors, namely between the predicted value and the actual (actual) value, which then divides that number by the number of times the prediction data is, then draws the roots and Correlation Coefficient which is a value to measure the relationship between the actual (actual) value [11] and the predicted value. The stronger the correlation, the better the results, which means the higher the level of prediction accuracy [12].

**RMSE**



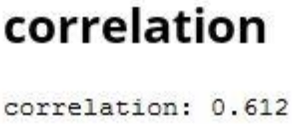
root_mean_squared_error

```
root_mean_squared_error: 0.898 +/- 0.000
```

Fig 6. Model Evaluation with RMSE

The RMSE value of the model obtained is 0.898. The lower the RMSE value, the better the predictions are made.

**Correlation**



correlation

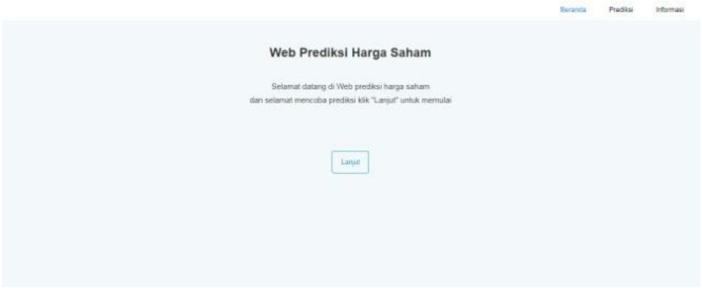
```
correlation: 0.612
```

Fig 7. Model Evaluation with Correlation

The correlation value of the model obtained is 0.612. According to [13], this value indicates that the predicted result with the actual value has a strong relationship or correlation.

### *D. Implementation*

Implementation of Linear Regression models into web-based applications.

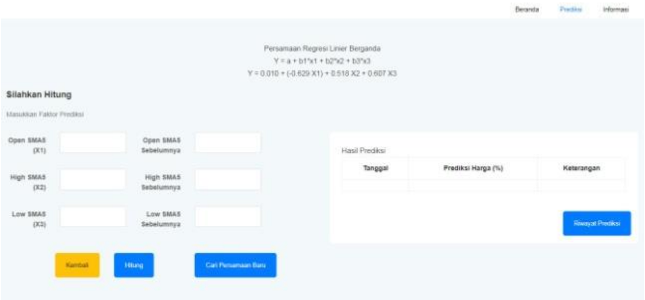


Web Prediksi Harga Saham

Selamat datang di Web prediksi harga saham
dan selamat mencoba prediksi klik "Lanjut" untuk memulai

Fig 8. Home App Display

The image above is the initial view of the application. On this page we are told to click the "Continue" button to enter the stock price prediction page.



Persamaan Regresi Linier Berganda
Y = a + b1*x1 + b2*x2 + b3*x3
Y = 0.010 + (-0.629 X1) + 0.518 X2 + 0.607 X3

Silahkan Hitung

Masukkan Faktor Prediksi

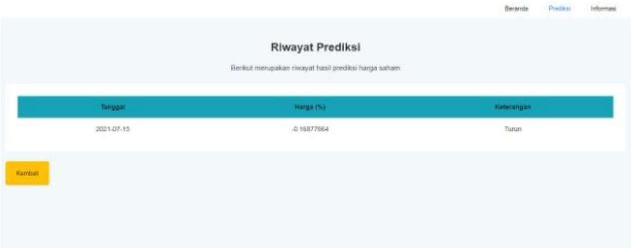
Open SMA5 (X1)
Open SMA5 Sebelumnya
High SMA5 (X2)
High SMA5 Sebelumnya
Low SMA5 (X3)
Low SMA5 Sebelumnya

Hasil Prediksi

| Tanggal | Prediksi Harga (%) | Keterangan |
|---|---|---|

Fig 9. Application Count Display

The image above is a display of the stock price prediction page that appears after we click the "Continue" button on the start page. On this page we can see the linear regression equation that has been obtained. Also on this page, we enter the prediction factor, namely the stock price data that we want to predict. There are several buttons available, namely, the "Calculate" button to calculate the predicted results, the "Back" button to return to the start page, the "Input Data Excel" button to go to the stock price prediction page with the new equation and the "Prediction History" button to go to stock price prediction history page.



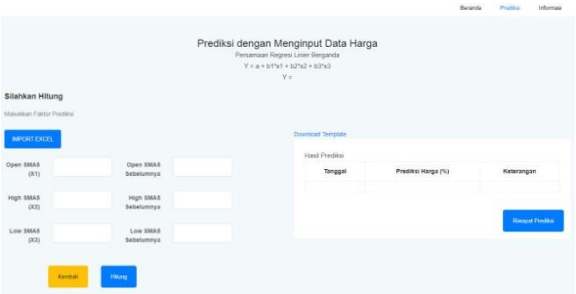
Riwayat Prediksi

Berikut merupakan riwayat hasil prediksi harga saham

| Tanggal | Harga (%) | Keterangan |
|---|---|---|
| 2021-07-15 | -0.16877664 | Turun |

Fig 10. Prediction History View

The picture above is a display of the prediction history page that appears after we click the "Prediction History" button on the calculation page. On this page, we can see the results of the predictions that have been made. There is a "Back" button to return to the calculation page.



Prediksi dengan Menginput Data Harga
Persamaan Regresi Linier Berganda
Y = a + b1*x1 + b2*x2 + b3*x3
Y =

Silahkan Hitung

Masukkan Faktor Prediksi

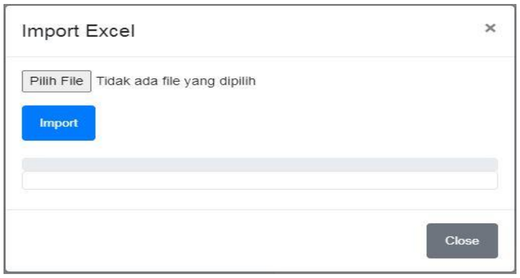
Open SMA5 (X1)
Open SMA5 Sebelumnya
High SMA5 (X2)
High SMA5 Sebelumnya
Low SMA5 (X3)
Low SMA5 Sebelumnya

Hasil Prediksi

| Tanggal | Prediksi Harga (%) | Keterangan |
|---|---|---|

Fig 11. Page Display with Excel Data Import

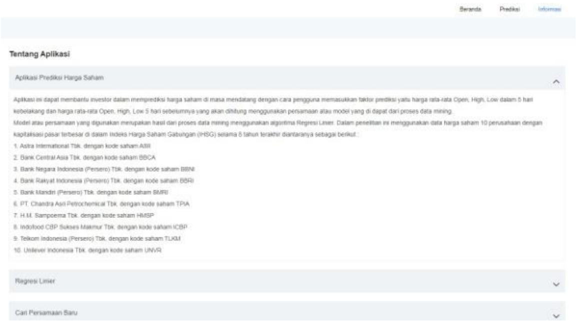
The image above is a stock price prediction page display, where users enter their own data. Users download an excel template to fill in stock price data which will be used as a dataset. After finishing entering price data, the user imports the excel file by clicking the "Import Excel" button, selecting the file and then clicking the "import" button. After the price data is inputted by the user, the program will look for a new linear regression equation according to the available data. On this page we can see the linear regression equation that has been obtained. Also on this page, we enter the prediction factor, namely the stock price data that we want to predict. There are several buttons available, namely the "Calculate" button to calculate the prediction results, the "Back" button to return to the calculation page and the "Prediction History" button to go to the stock price prediction history page.



Tidak ada file yang dipilih

Fig 12. Excel Data Import view

The picture above is a display when the user wants to input the stock price data excel file that appears after we click the "Import Excel" button. Select the excel file that has been filled with stock price data, then click "Import".



Tentang Aplikasi

Aplikasi Prediksi Harga Saham

Aplikasi ini dapat membantu investor dalam memprediksi harga saham di masa mendatang dengan cara pengguna memasukkan faktor prediksi yaitu harga rata-rata Open, High, Low dalam 5 hari kebelakang dan harga rata-rata Open, High, Low 5 hari sebelumnya yang akan dihitung menggunakan persamaan atau model yang di dapat dari proses data mining.

Model atau persamaan yang digunakan merupakan hasil dari proses data mining menggunakan algoritma Regresi Linier. Dalam penelitian ini menggunakan data harga saham 10 perusahaan dengan kapitalisasi pasar terbesar di dalam Indeks Harga Saham Gabungan (IHSG) selama 5 tahun terakhir diantaranya sebagai berikut :

1. Astra International Tbk. dengan kode saham ASII
2. Bank Central Asia Tbk. dengan kode saham BBCA
3. Bank Negara Indonesia (Persero) Tbk. dengan kode saham BBNI
4. Bank Rakyat Indonesia (Persero) Tbk. dengan kode saham BBRI
5. Bank Mandiri (Persero) Tbk. dengan kode saham BMRI
6. PT. Chandra Asri Petrochemical Tbk. dengan kode saham TPIA
7. H.M. Sampoerna Tbk. dengan kode saham HMSP
8. Indofood CBP Sukses Makmur Tbk. dengan kode saham ICBP
9. Telkom Indonesia (Persero) Tbk. dengan kode saham TLKM
10. Unilever Indonesia Tbk. dengan kode saham UNVR

Regresi Linier

Cari Persamaan Baru

Fig 13. Information Page Display

The page above contains information about the application, the source of the dataset used, the algorithm used, namely the Linear Regression algorithm, an explanation of the model or equation obtained and an explanation of the Find New Equations feature.

## IV. Conclusions

Based on the creation of a new model using the Moving Average method and the Linear Regression algorithm that has been created and explained in this study, the authors conclude that the model created can assist investors in making decisions to buy or sell stocks where the predicted results show ups and downs, where ups are used as buy and down recommendations are used as sell recommendations with an evaluation using the RMSE of 0.898, and a correlation value of 0.612.

## References

[1] kseinews, “Ksei Menjadi Kustodian Sentral Terbaik Di Asia Tenggara Ketiga Kalinya,” Edisi 01, p. 10, 04 Juni 2020.

[2] I. Fahmi, dalam Analisis Laporan Keuangan, Bandung, Alfabeta, 2014, p. 324.

[3] M. Azis, S. Mintarti dan M. Nadir, dalam Manajemen Investasi Fundamental, Teknikal, Perilaku Investor dan Return Saham, Yogyakarta, Deepublish (Grup Penerbitan CV Budi Utama), 2015, p. 80.

[4] R. Hidayat, “Prediksi Harga Saham Menggunakan Neural Network,” Jurnal Gema Aktualita, p. 65, 2016.

[5] F. S. Gharehchopogh, T. H. Bonab dan S. R. Khaze, “A Linear Regression Approach To Prediction of Stock Market Trading Volume : A Case Study,” International Journal of Managing Value and Supply Chains (IJMVSC), 2013.

[6] D. T. Larose, Discovering Knowledge in Data : An Introduction to Data Mining, New Jersey: John Wiley & Sons, Inc, 2014.

[7] Kamal, I. M., P, T. H., & Ilyas, R. (2017). Prediksi Penjualan Buku Menggunakan Data Mining. Seminar Nasional Teknologi Informasi Dan Multimedia, 49–54.

[8] E. S. Tataming, T. K. Sendow, O. H. Kaseke och S. Diantje, ”Analisis Besar Kontribusi Hambatan Samping Terhadap Kecepatan dengan Menggunakan Model Regresi Linier Berganda,” Jurnal Sipil Statik, pp. 29-36, 2014.

[9] A. Nurlifa Och S. Kusumadewi, ”Sistem Peramalan Jumlah Penjualan Menggunakan Metode Moving Average Pada Rumah Jilbab Zaky,” Jurnal Inovtek Polbeng, Vol. Ii, Pp. 18-25, 2017.

[10] H. Utari, M. Och N. Silalahi, ”Perancangan Aplikasi Peramalan Permintaan Kebutuhan Tenaga Kerja Pada Perusahaan Outsourcing Menggunakan Algoritma Simple Moving Average,” Jurnal Times, Vol. V, Pp. 1-5, 2016.

[11] H. Budiman, ”Analisis Dan Perbandingan Akurasi Model Prediksi Rentet Waktu Support Vector Machines Dengan Support Vector Machines Particle Swarm Optimization Untuk Arus Lalu Lintas Jangka Pendek,” Systemic, Vol. %1 Av %22, No. 01, Pp. 19-24, 2016.

[12] A. Fadholi, ”Pemanfaatan Suhu Udara Dan Kelembapan Udara Dalam Persamaan Regresi Untuk Simulasi Prediksi Total Hujan Bulanan Di Pangkalpinang,” Jurnal Cauchy, Vol. %1 Av %2iii, No. 1, Pp. 1-9, 2013.

[13] Riduwan, Dasar-Dasar Statistika, Bandung: Alfabeta, 2010.