E-ISSN: 2476-9606
Abstract Proceedings International Scholars Conference
Volume 7 Issue 1, October 2019, pp. 1244-1258

https://doi.org/10.35974/isc.v7i1.1872

Comparison of Accuracy Between Two Methods: Naïve Bayes Algorithm and Decision Tree-J48 to Predict The Stock Price of Pt Astra International tbk Using Data From Indonesia Stock Exchange

Joan Yuliana Hutapea[1], Yusran Timur Samuel[2], Heima Sitorus[3]
[1]Faculty of Economics, Universitas Advent Indonesia
[2,3]Faculty of Information Technolgy, Universitas Advent Indonesia
y.tarihoran@unai.edu

ABSTRACT

The ability to predict the stock prices is very important for market players, whether individual or organizational investors. The market players needs to know how to predict, that will help them in their decision making process, whether to buy or to sell its shares, so that it can maximize profits and reduce potential losses due to mistakes in decision making. In accordance to this, the authors conducted a study that aimed to analyze and to compare the accuracy of two (2) methods that is used to predict the stock prices, namely: the Naïve Bayes Method and the Decision Tree-J48 Method. The amount of data used in this study were 1,195 stock datas of PT Astra International Tbk, issued by the IDX, by the period of January 1, 2013 to November 30, 2017.
This study uses 7 attributes, namely: Previews, High, Low, Close, Volume, Value, and Frequency. By using the WEKA application the result shows that, the accuracy of the Naïve Bayes Method using 20% of testing data, is 92.0502%, the precision value is 0.920 and the value of recall is 0.961, while the accuracy of the Decision Tree J-48 method, using 20% of testing data, is 98.7448%, with precision value of 0.989 and the value of recall of 0.997.Through this results, it can be concluded that the decision tree J-48 algorithm has a better accuracy results compared to the Naive Bayes algorithm in predicting the stock price of PT. Astra Internasional Tbk.

Keywords: Predicting Stock Prices, WEKA, Naive Bayes, Decision Tree – J48.

INTRODUCTION

The development of the share prices of every companies in Indonesia that have go public can be easily traced through the capital market. The capital market which brings together market participants, such as investors and companies, has a very important function in providing financial information that is necessary in the making of investment decisions. The Indonesia Stock Exchange (BEI or IDX) is an institution that operates a trading system or buying and

selling of shares or securities in Indonesia. One company that has been listed on the Indonesian stock exchange, which has go public is PT Astra Internasional Tbk. People who want to know the movement of PT Astra International Tbk's share prices can get information through the Indonesia stock exchange.

The disclosure of information about a company's stock price on the IDX can be used by the public and investors to gain financial benefits through the buying and selling of shares on the IDX. However, to get these benefits, the public and investors must know the right time to buy or sell the shares. Analysis on stock price movements is absolutely necessary before investors make their investment decisions, whether to buy or sell shares. In practice, this analyzing process is not easy to do. Mistakes in predicting stock price movements become a frequent problem for the investors, that results in losing the opportunity to benefit from their investment decisions. As the result, often the public and investors who wants to be involved in the field of stock trading are backward and afraid to face the risk and the possibility of losing their investment due to the wrong decisions and mistakenly seized the opportunity to invest.

Based on the background description, the authors are interested in conducting a study that aims to find out how to analyze the price of shares, especially the shares of PT Astra International Tbk by utilizing information of stock price data on the IDX, using the Naïve Bayes and Decision Tree-J48 methods.

LITERATURE REVIEW

Definition of Stock Prices

Share prices that displayed on the Indonesian stock exchange or on the Indonesian capital market are generally displayed in a graph, that shows the up and down movements of the stock price. Information on stock price movements is needed by the public or investors who want to invest in certain companies. The stock price index in the Indonesian stock exchange is an indicator that shows the trend of changes in the company's stock prices in certain exchanges, as revealed by Supranto [1], which states that "Stock price index is a stock price index number that has been arranged and calculated in such a way as to produce trends . The index number is a number that is made in such a way that it can be used to compare economic activities or events, it can be a change in stock prices from time to time ".

In analyzing the stock prices, generally there are two analytical techniques that are often used, namely: fundamental analysis and technical analysis.

1. Fundamental Analysis

Fundamental analysis is a form of analysis that studies matters related to the financial condition of a company, which shows how the company is managed on its operational activities. Riswantoro [2] revealed, fundamental analysis is "an analytical method based on the economic development of a company. This analysis is also often used to conduct analysis of companies in the financial sector. This analysis is also often used by parties who want to know about stock prices in the long term or in the long term ".

2. Technical Analysis

Technical analysis is an analysis carried out using historical data about the movement of prices and volume of shares. Technical analysis is a methodology for predicting stock prices based on the conditions of supply and demand for these shares. [2] To be able to analyze the stock price, the things the writer will notice are as follows:

1) Prev or previous is the closing price of the stock on the previous stock day. For example if today is Tuesday, prev shows the closing price on the previous day, which is Monday.
2) The highest price is the highest price of the shares that occurred that day
3) The lowest price is the lowest price that occurred that day.
4) The frequency of stock trading is the number of times a sale and purchase transaction occurs on the relevant stock at a certain time.
5) Individual index is the number of stock indexes in the capital market.
6) The volume, determines the number of transactions traded in the market in a given period.
7) Value is the number of shares sold at one time.

Data collection and processing techniques

Data Mining

Data Mining (DM) is one of the fields that is currently experiencing a rapid development due to the increasing needs for value added from large-scale databases, which are increasingly accumulated in line with the growth of information technology. The general definition of data mining itself is a process undertaken to explore the added value of knowledge from a collection of data that has not been considered important by the community. [2]

Data mining, often also referred to as Knowledge Discovery in Database (KDD). KDD is an activity that includes the collection, use of data, historical to find regularities, patterns or

relationships in large data sets. [3]. The process in the data mining stage consists of three main steps, namely:

1. Data Preparation. In this step the selected data is cleaned and preprocessed.
2. The use of data mining algorithms.
3. Analysis Phase Output from data mining is evaluated to see whether knowledge domain is found in the form of rules that have been extracted.

Figure 1. Steps in the Data Mining Process

Naive Bayes Algorithm.

Naive Bayes algorithm is one of the algorithms in the classification technique. Naive Bayes is a classification with probability and statistical methods raised by British scientists Thomas Bayes, which predicts future opportunities based on past experience so that it is known as the Bayes Theorem. The theorem is combined with Naive where it is assumed that conditions between attributes are mutually independent. The Naive Bayes classification is assumed that the presence or absence of certain characteristics of a class has nothing to do with the characteristics of other classes. The equation from the Bayes theorem is [5]:

$$\text{P (s| x)} = \frac{argmax}{x \in X} \frac{P(\ y\ |x)\ P(x)}{P(X)}$$

$$= \frac{argmax}{x \in X}\ P\big((\ y\ |x\big)\ P(X)$$

Decision Tree Method J-48

The Decision Tree J-48 method is one of the classes found in the data mining program in Weka software. J48 algorithm is an implementation of the C4.5 algorithm program [6] So in general the J-48 algorithm is usually used for classification problems. A decision tree consists of several nodes, namely the root tree, internal node and leafs. The concept of entropy is used to determine which attributes a tree is divided into. The higher the entropy of a sample, the more impure the sample is.

The formula used to calculate the "entropy" sample S is;

$$Entropy\ (S) = \sum \tau - pi \log 2\ pi$$

or

Entropy(S) = -p1log2p1–p2log2p2

Waikato Environment Knowledge And Analisys (WEKA)

WEKA is a data mining / machine learning software developed by the Department of Computer Science University of Waikato in New Zealand. WEKA is able to solve problems related to data mining that exist in the real world, especially the classification that underlies the approach to machine learning. In general WEKA provides 3 facilities for data mining. The three facilities are data processing, data mining and visualization tools. [7]

Related Research

No	**Title**	***Number of Data Training***	**Number of Data *Testing***	**Naïve Bayes Accuracy**	***Decision Tree* J-48 Accuracy**
1	Comparison of Decision Tree Algorithm (C4.5) and Naive Bayes on Data Mining for Identification Growth and Development Child Toddler (Case Study PUSKESMAS KARTASURA). Mila Listiana,Drs.Sudjalwo,M. Kom., Dedi Gunawan,S.T.,M.Sc	304	304	76.97%	75.66%
2	Performance Comparison of Data Mining Classification Methods Using Naive Bayes and C4.5 Algorithm for the Accuracy Prediction of Students Graduation Time. Gian Fiastantyo, Univesitas Dian Nuswantoro, Semarang	1919	1919	74.09%	82.43%
3	Implementation of Data Mining Classification Methods For Prediction of Graduation Timeliness.	5842	891	99.89%	100%

	Asep Saefulloh, Moedjiono				
4	Comparison of Performance of C4.5 and Naive Bayes Algorithms for Determination of Student Concentration Selection. Wiwit Supriyanti, Kusrini, Armadyah Amborowati	539	539	82.01%	84,98%
5	Comparison of Performance of Naïve Bayesian, Lazy-Ibk, Zero-R, and Decision Tree J-48 Algorithms. Sulidar Fitri	1448	818	85,12%	84,23%

METHODS

Data Collection and Processing of IDX Stock Data

The research data used by the author is data that has been published through the Indonesia Stock Exchange website : (http://www.idx.co.id/id-id/beranda/unduhdata/ringkasan.aspx.)

Stock Data

Data collection of PT Astra International Tbk stock data collection were done manually. The data taken by the author are daily data starting from January 1, 2013 until November 30, 2017. However, the data need to be managed, since it is generated from the website which are the combination of stock price data of registered companies.

Daily Trading PT ASTRA INTERNASIONAL Tbk 2013 - 2017

NO	Date	Previous	High	Low	Close	Volume	Value	Frequency	Output
1	2-Jan-13	7800	7700	7450	7500	23374000	1.77427E+11	1753	Naik
2	3-Jan-13	7500	7900	7500	7850	70205000	5.42500E+11	4567	Turun
3	4-Jan-13	7850	7850	7750	7850	26485000	2.10064E+11	2069	Turun
4	7-Jan-13	7850	7850	7700	7750	20890000	1.62270E+11	1863	Turun
5	8-Jan-13	7750	7800	7600	7650	24269500	1.86656E+11	1969	Turun
6	9-Jan-13	7650	7750	7650	7700	38907000	2.99253E+11	1936	Naik
7	10-Jan-13	7700	7700	7300	7350	50290500	3.74908E+11	5175	Turun
8	11-Jan-13	7350	7500	7300	7300	33422500	2.46396E+11	3791	Naik
9	12-Jan-13	7300	7500	7350	7500	28814000	2.08141E+11	2429	Turun
10	13-Jan-13	7500	7600	7500	7600	32451000	2.44075E+11	2845	Naik
11	14-Jan-13	7600	7650	7500	7650	23595000	1.78944E+11	2810	Turun
12	17-Jan-13	7650	7650	7500	7500	27764000	2.1031E+11	2463	Turun
13	18-Jan-13	7500	7750	7550	7750	34303000	2.63287E+11	2963	Naik

Figure 2. Sample List of PT Astra International Tbk stock prices 2013-2017

RESULTS

Testing and Research Results

Testing Data

Data testing is performed using WEKA software that can be used at the same time for data processing. Stock price data that has been collected and stored in the Excel program, then saved with the CVS (Concurrent Versioning System) format, then opened in the Weka program, then saved again with the ARFF format (Andrew's Ridiculos File Format).

Testing Training Data

The amount of data to be trained is 1195 which has been saved in AFRR formats.

Table 1. **Naïve Bayes Results and Decision Tree J-48 at WEKA**

Stock Price Data	Naïve Bayes	*Decision Tree* J-48
True	1010	1182
False	185	13

Decription:

1. Naïve Bayes

There are 1010 stock price data determined by WEKA that are classified as true, and 185 stock price data that is False.

2. *Decision Tree* J-48

There are 1182 stock price data determined by WEKA that are classified as true, and 13 stock price data that is False.

Description:

1. *Correctly Classified Instance*

The formula for finding accuracy values is:

$$\text{Accuracy} = \frac{\text{number of correct s tock price prediction}}{\text{Total amount of predictive data}}$$

$$\text{Accuracy} = \frac{f_0 + f_3}{f_0 + f_1 + f_2 + f_3}$$

$$\text{Accuracy} = \frac{838 + 172}{838 + 141 + 44 + 172}$$

$$\text{Accuracy} = \frac{1010}{1195} = 0.845188$$

2. *Incorrectly Classified Instance*

$$Error\ Rate = \frac{\text{number of false stock price prediction}}{\text{Total amount of predictive data}}$$

$$Error\ Rate = \frac{f_1 + f_2}{f_0 + f_1 + f_2 + f_3}$$

$$Error\ Rate = \frac{141 + 44}{838 + 141 + 44 + 172}$$

$$\text{Accuracy} = \frac{185}{1195} = 0.154812$$

3. *Total Number Of Instance*

   Total Number of Instances is the amount of data used in the training process, which is 1,195.

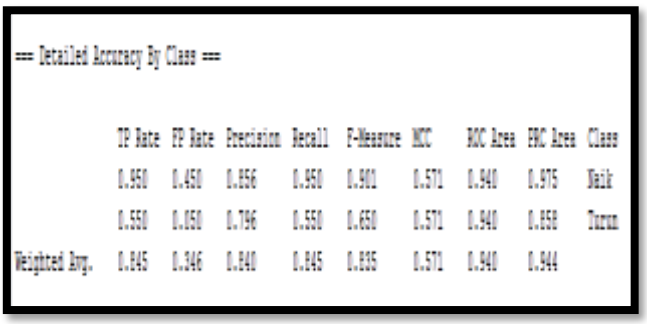
```
=== Confusion Matrix ===

   a   b   <-- classified as
 838  44 |   a = Naik
 141 172 |   b = Turun
```

Figure 3. *Confusion Matrix*

Description:

The first line explains that there are 838 stock price data from classified "Up" classified as having increased prices, and there are 44 share price data that should have increased share prices, but incorrectly classified as having decreased share prices.

The second line explains that there are 172 stock price data from classified "Down" which are classified as experiencing a decline in share prices, but there are 141 data on stock prices that should have decreased in stock prices, but which are incorrectly classified as experiencing stock price increases.



== Detailed Accuracy By Class ==

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.950	0.450	0.856	0.950	0.901	0.571	0.940	0.976	Naik
	0.550	0.050	0.796	0.550	0.650	0.571	0.940	0.858	Turun
Weighted Avg.	0.845	0.346	0.840	0.845	0.835	0.571	0.940	0.944	

Figure 4. Detailed Accuracy

Description:

1. *True positif Rate (TP Rate)*

TP Rate or True Positive Rate is the sum of the truth level of the stock price data which is classified as class x, among all the correct stock price data having class x which is equal to recall. To find the TP Rate value in the increased stock price class are:

$$True\ Positif\ Rate = \frac{\text{number of correct data}}{\text{number of correct data} + \text{number of false data}}$$

$$True\ Positif\ Rate = \frac{838}{838 + 44} = 0{,}950$$

To find the value of FN or False Negative in the decreased stock prices class are:

$$False\ Negative = \frac{\text{number of false data}}{\text{number of correct data} + \text{number of false data}}$$

$$True\ Positif\ Rate = \frac{172}{172 + 141} = 0{,}550$$

2. *False Positive (FP Rate)*

To find the value of False Positive is the value of the stock price classified as class x, but entered into a different class, and which is not from class x itself:

To find the value of False Positive for classes that have increased stock price data:

$$False\ Positive = \frac{\text{number of correct data}}{\text{number of correct data} + \text{number of false data}}$$

$$False\ Positive = \frac{141}{141 + 172} = 0{,}450$$

False Negative on the decreased class are:

$$FalseNegative = \frac{\text{number of data}}{\text{number of correct data} + \text{number of false data}}$$

$$False\ Negative = \frac{838 + 44 + 141 + 172}{141 * 172} = 0{,}050$$

3. *Precision*

Precision is the level of accuracy of information expected by the author with the answers provided by the system.

To find the Precision value from the increased class:

$$Precision = \frac{\text{number of correct data}}{\text{number of correct data} + \text{number of false data}}$$

$$False\ Negative = \frac{838}{838 + 141} = 0{,}856$$

To find the precision value from the decreased class:

$$Precision = \frac{\text{number of false data}}{\text{number of correct data} + \text{number of false data}}$$

$$False\ Negative = \frac{172}{172 + 44} = 0{,}796$$

4. *Recall*

Recall is the success rate of the system in rediscovering information. To find recall value in the increased class are:

$$Recall = \frac{\text{number of correct data}}{\text{number of correct data} + \text{number of false data}}$$

$$Recall = \frac{838}{838 + 44} = 0{,}950$$

To find recall value in the decreased class:

$$Recall = \frac{\text{Sum of incorrect clasification}}{\text{number of correct data} + \text{number of false data}}$$

$$Recall = \frac{172}{172 + 141} = 0{,}550$$

5. *F-Measure*

F-Measure is a combined calculation of Precision and Recal values..

To find the F-Measure value is:

$$F - Measure = \frac{2 * Precision * Recall}{Precision * Recall}$$

F-Measure for increasing stock price data are::

$$F - Measure = \frac{2 * 0.856 * 0.950}{0.856 + 0.950} = 0.901$$

F-Measure for decreasing stock price data are:

$$F - Measure = \frac{2 * 0{,}796 * 0.550}{0{,}796 + 0.550} = 0.650$$

Data *testing*

Data testing uses 20% of the amount of data that has been collected. The amount of data used is 1195, then the data used as testing data is 239 taken randomly. To test the data testing whose truth is unknown, a number of processes are required as follows:

1. Prepare the data, and save it as .arff file format,
2. Prepare the training data results that will be used as a reference in determining the classification of data testing
3. Perform the test using the WEKA program.

Naïve Bayes Method

The results of data testing with WEKA:

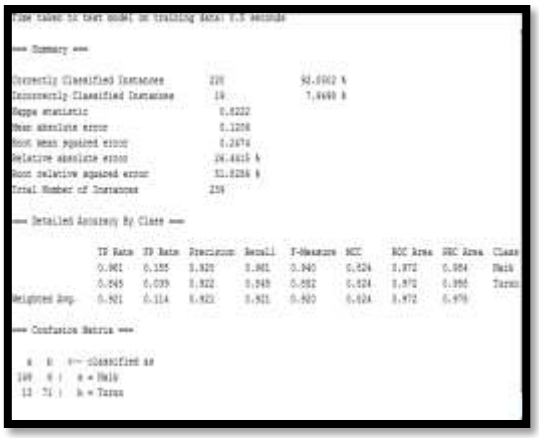
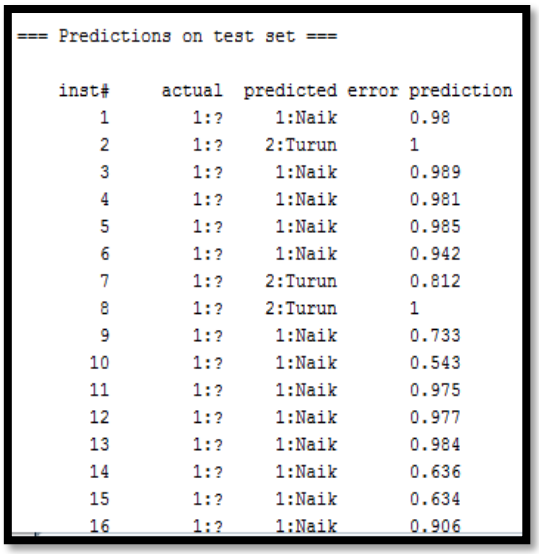


Figure 5. Prediction results of WEKA Naïve Bayes

Tabel 2**. Data *Testing* Naïve Bayes classification**

Prediction output	Frequency	Number of Data
True	**220**	**239**
False	**19**	**239**



```
=== Predictions on test set ===

    inst#     actual  predicted error prediction
        1        1:?     1:Naik       0.98
        2        1:?    2:Turun       1
        3        1:?     1:Naik       0.989
        4        1:?     1:Naik       0.981
        5        1:?     1:Naik       0.985
        6        1:?     1:Naik       0.942
        7        1:?    2:Turun       0.812
        8        1:?    2:Turun       1
        9        1:?     1:Naik       0.733
       10        1:?     1:Naik       0.543
       11        1:?     1:Naik       0.975
       12        1:?     1:Naik       0.977
       13        1:?     1:Naik       0.984
       14        1:?     1:Naik       0.636
       15        1:?     1:Naik       0.634
       16        1:?     1:Naik       0.906
```

Figure 6. Sample of Naïve Bayes Data testing of stock price prediction results

Description: Inst # is a number for every data that is tested, Actual is output that has not been filled in with the results, and is only filled with question marks (?), Predicted is the output classification generated by the WEKA program itself. Error prediction is the presentation of the number of errors that occur.

Decision Tree J-48 Method

The results of the training data with the WEKA program.

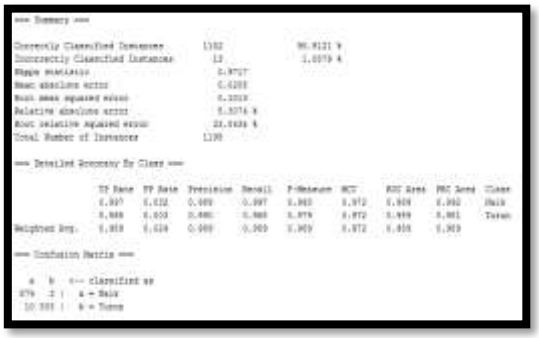


Figure 7. Data *training* J-48

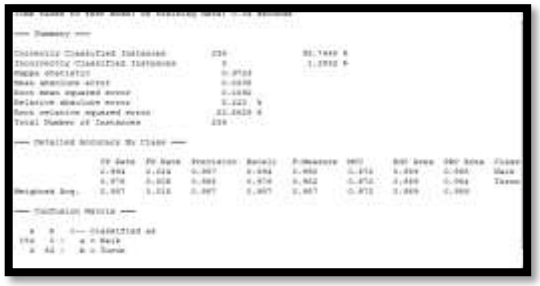


Figure 8. Classification with *Decision Tree* J-48 method

Tabel 4. **Data Testing Clasification**

Prediction output	Frequency	Number of Data
True	236	239
False	3	239

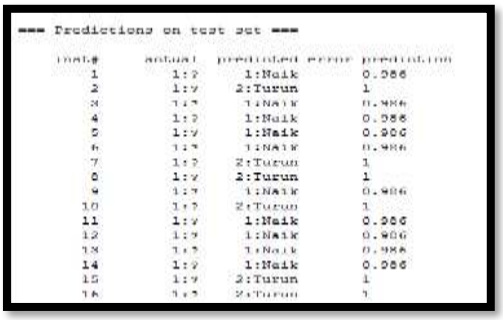


Figure 9. Sampel of Decision Tree J-48 Prediction

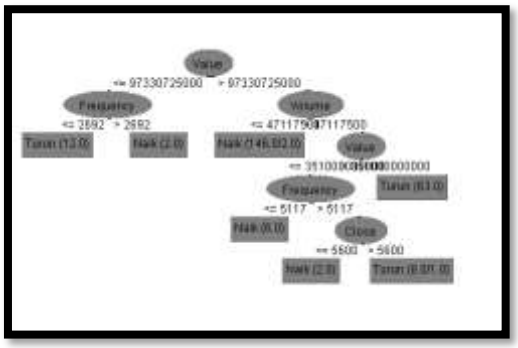


Figure 10. Classification result of *Decision Tree* J-48

Research Results

Output Prediction	**Accuracy**	***Error Rate***
Data Training	84.52%	15.48%
Data Testing	92.05%	7.95%

Tabel 5. Results of Accuracy and *Error Rate on* Naïve Bayes

Output Prediction	**Accuracy**	***Error Rate***
Data Training	98.91%	1.09%
Data Testing	98.74%	1.26%

Tabel 6. Results of Accuracy and *Error Rate* on *Decision Tree* J-48

DISCUSSION

Conclusion

Based on the results of the research, the author concluded that:

Using the Naïve Bayes method the accuracy of the data training were 84.5188% and the results of data testing were 92.0502%. There are 149 stock price data classified as having increased prices, and there are 6 share price data that should have increased share prices, but are classified as having decreased share prices. And there are 71 stock price data that are classified correctly experiencing a stock price decline, but there are 13 stock price data that should have decreased stock prices, but are classified as having stock price increases. The value of precision or the level of accuracy of information expected by the author with the answer given by the system are 0.920 and the value for recall or the success rate for information recovered by the system is 0.961.

The prediction results using the Decision Tree J-48 method, obtained the results of the accuracy for data training of 98.9121%, and the accuracy of data testing were 98.7448%. There are 154 share price data classified as having increased prices, and there is 1 share price data that should have increased prices, but is classified as decreased share prices. And there are 82 share price data that are classified as decreased share prices, but there are 2 data that should have decreased in stock prices, but are classified as increased share prices. The value of precision or the level of accuracy of information expected by the author with the answers given by the system are 0.989 and the value for recall or the success rate for information recovered by the system are 0.997.

Through this results, it can be concluded that the decision tree J-48 algorithm has a better accuracy results compared to the Naive Bayes algorithm in predicting the stock price of PT. Astra Internasional Tbk.

REFERENCES

Supranto. J, "Pasar Modal Indonesia. Rineka Cipta. Jakarta," *Pasar Modal Indones.*, 1992.

D. S. Riswantoro, "Abstrak Penerapan Data Mining Untuk Memprediksikan Fluktuasi Harga Saham Menggunakan Metode Classification Dengan Teknik Dt (Decision Tree)," pp. 4–5, 2011.

M. Ridwan, H. Suyono, and M. Sarosa, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *Eeccis*, vol. 7, no. 1, pp. 59–64, 2013.

Retno Tri Wulandari, Data Mining Teori dan Aplikasi Rapidminer, Gava Media, 2017

Sigit Adinugroho,Yuita Arum Sari, Implementsi Data Mining Menggunakan WEKA,UB Press, 2018

S. Fitri, "Perbandingan Kinerja Algoritma Klasifikasi Naïve Bayesian , Lazy-Ibk , Zero-R , Dan Decision Tree- J48," *Dasi*, vol. 15, no. 1, pp. 33–37, 2014.

D. Purnamasari, J. Henharta, Y. P. Sasmita, F. Ihsani, and I. W. S. Wicaksana, "Get Easy Using WEKA," 2013.