Indonesian Stock Price Prediction including Covid19 Era Using Decision Tree Regression

Kartika Maulida Hindrayani
Data Science
UPN "Veteran"Jawa Timur
Surabaya, Indonesia
kartika.maulida.ds@upnjatim.ac.id

Tresna Maulana Fahrudin
Data Science
UPN "Veteran"Jawa Timur
Surabaya, Indonesia
tresna.maulana.ds@upnjatim.ac.id

Prismahardi Aji R.
Data Science
UPN "Veteran"Jawa Timur
Surabaya, Indonesia
prismahardi.aji.ds@upnjatim.ac.id

Eristya Maya Safitri
Data Science, Information System
UPN "Veteran"Jawa Timur
Surabaya, Indonesia
maya.si@upnjatim.ac.id

2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) | 978-1-7281-8406-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ISRITI51436.2020.9315484

Predicting stock prices is an interesting field in Data Mining. There are many variables affecting stock prices. Especially in this covid19 era which impacts in economy, the stock prices become unpredictable. Telecommunications companies are observed in this research as it is one of the sectors that's still very much in demand in this pandemic situation. Fundamental data will be used to predict the Indonesian telecommunications stock price. Regression techniques will be used as the proposed model. The correlation coefficient shows that despite the covid19 era, fundamental data still play a role in stock market price. Decision Tree Regression produced competitive results compared to other methods.

Keywords— Indonesian Telecommunications Stock Prices, Prediction, Decision Tree, Regression, COVID-19

I. INTRODUCTION

Nowadays, Indonesian people already concern about investment. The common investments are buying stocks, money, land, and gold. Investment is an activity to gain profit, while shares are proof of company ownership that is attractive for investment. There are many factors influence stock prices. Rational factors and various irrational factors are the determining factors in purchasing shares. These factors include company performances, investor confidence, state conditions, government regulations, inflation rates, etc [1]. Currently, news articles and social media could also affect investor decisions.

Coronavirus Disease 2019 (COVID-19) spreads easily and deathly for some people, and the reasons are still being discovered. On March 12, 2020, (World Health Organization) WHO announced the outbreak as a pandemic. The pandemic traveled fast, China, South Korea, Italy, the UK, the USA, and other countries have been infected. They have a massive positive case before they can control the outbreak. Controlling the outbreak means controlling people movement, at first. This means, people lifestyles, consumption, and shopping are changing. Stock prices are falling globally. The global pandemic of COVID-19 caused different patterns of stock markets before and after the news hits [2].

In March 2020, the first case of COVID-19 in Indonesia was announced. The case-finding resulted in unstable conditions both in the economy and public safety. The stock situation in Indonesia at the time of COVID-19 can be seen from the Indeks Harga Saham Gabungan (IHSG), which is the joint-stock price of Indonesia.

Prediction of stock prices can help investors make decisions on stock market transactions. Prediction of stock prices is used as a consideration in deciding on whether to retain or sell its shares. There are two approaches to analyze stock price movements, namely fundamental and technical analysis.

In Indonesia, people's movements are also limited due to COVID-19, but they have to stay connected. One of the sectors that is still surviving with its performance is telecommunications. This study will observe telecommunications company data stocks to be predicted using regression techniques including the Covid19 era. The contribution of this paper is we aim to get the best model for generalizing the inference (conclusion) of each predictive decision of each company's stock and has the smallest Mean Absolute Percentage Error (MAPE).

II. LITERATURE REVIEW

A lot of research has been conducted in the fields of stock price prediction. Anwar and Deni [3] predicted the Indonesian Stock Exchange Composite Index using Optimized Fuzzy Backpropagation Neural Network using Genetic Algorithm. This method successfully predicted an 8.42% MAPE value from the composite stock price index. Zuherman, Vibranti, and Widya [4] used support vector machines and fuzzy Kernel C-Means to predict the direction of Indonesian stock price movement. The result is good because of the relatively small Normalised Mean Squared Error (NMSE) values.

Mohan et al [5] using time series forecasting models, neural networks, and a combination of sentiment analysis and neural networks, such as ARIMA, RNN, and Facebook Prophet to predict stock price. The best result achieved is with RNN. Stock price and text information from news correlation also found. Yuniar [6] also using time series analysis and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) to predict weekly stock price. Parray et al [7] analyzing stock price movement using time series analysis with machine learning technique. The accuracy increased about 2% with the proposed model.

One of the techniques used in predicting stock trends is the regression technique by Siew and Nordin [8]. The research resulted in Sequential Minimal Optimization (SMO) Regression technique outperforming the other regression techniques. Yahya and Bayu [9] analyzed a few Indonesian stock prices and their sentiment analysis using Linear Regression. Another research conducted by Zuherman and Puteri [10] predicted a few stock market prices using Support Vector Regression with Particle Warm Optimization Feature Selection, resulted in a good performance.

Ouahilal et al [11] optimizing stock market price using novel hybrid model based on Hodrick–Prescott filter and support vector regression algorithm. The results show a very good accuracy of stock market price prediction with a very minimalist execution time. Asghar et al [12] developing the stock market trend prediction system using multiple regression. In this research, regression techniques will be used to predict the stock market price.

In this research, we use four years of data to predict the stock market price. The telecommunications sector is chosen because of the role to keep the societies connect through the pandemic. Four telecommunication companies in Indonesia will be observed.

III. Data and Methodology

A. Data

Data collected in a comma-separated value (CSV) format. Collected stock price data within four years from 2016 until September 2020. Fundamental data from the company are collected from each company's Financial Report. The financial report is released every quarter in each year. Stock price data are collected from Google Finance which is released daily in the workday. The company detail can be seen in table 1.

TABLE I. Company Detail

Number	Stock Code	Company Name
1	TLKM	Telekomunikasi Indonesia
2	EXCL	XL Axiata
3	FREN	Smartfren
4	ISAT	Indosat

The fundamental data from the companies that will be used are total current assets, total liabilities, net income for the period, and Earning Per Share (EPS). Current assets including company assets that can easily be disbursed in the form of money in less than one year such as cash, marketable securities, accounts receivable, income receivables, prepaid expenses, supplies, merchandise inventory. Liability is the amount owed or the company's obligation to pay to other parties. EPS is the level of net profit for each share that the company can achieve when carrying out its operations in the current period. The EPS calculation is obtained from the profit available to ordinary shareholders divided by the number of ordinary shares.

B. Methodology

The methodology used in this research divided into seven steps as seen in Figure 1. After gathering the company financial report and stock price, the next step is determining variables. The independent variables are total current assets, total liabilities, net income for the period, and Earning Per Share (EPS). The dependent variable is the closing daily stock price.

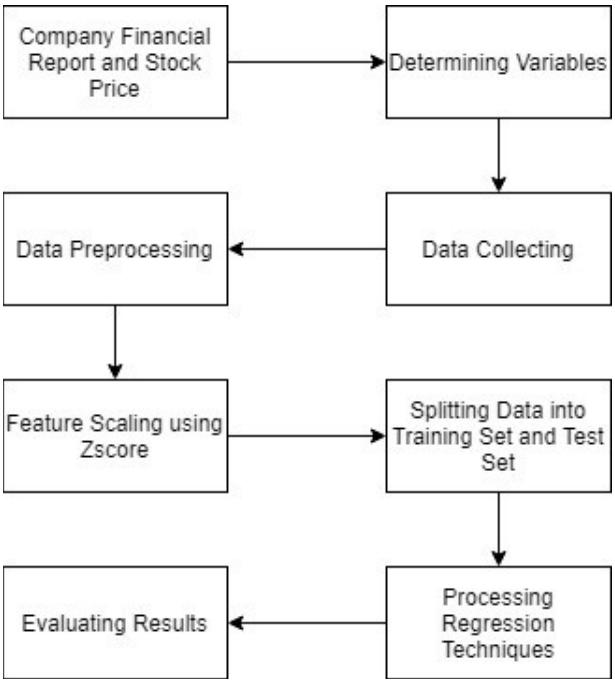

Figure 1. Research Methodology

The next step is collecting data. The financial report is released each quarter usually on December 31, March 31, June 30, and September 30. Collecting data need to be done carefully since the impact of a financial report can be seen after it has been released. For example, the company's fundamental that will guide the investor in January is the company's Q4 financial report from the previous year.

After data has been collected, cleaning data is needed in the data preprocessing. The next process is feature scaling with Zscore. Feature scaling is a step of standardization of the data. It is applied to independent variables or features of data to be normalized within a particular range. Zscore or standard score helps to calculate the probability of a score occurring within normal distribution and compare two scores from different normal distributions.

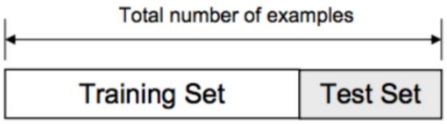

Figure 2. Train Test Split

After processing data, data should be split into a training set and test set. The training set will create the model using the regression techniques. A test set will be used to evaluate the results. A usual train test split and K-Fold Cross Validation is used to experiment with the dataset. Both

validation sampling method aims to determine the robustness of the model that was successfully made. The difference between the two splitting data is seen in Figure 2 and Figure 3.

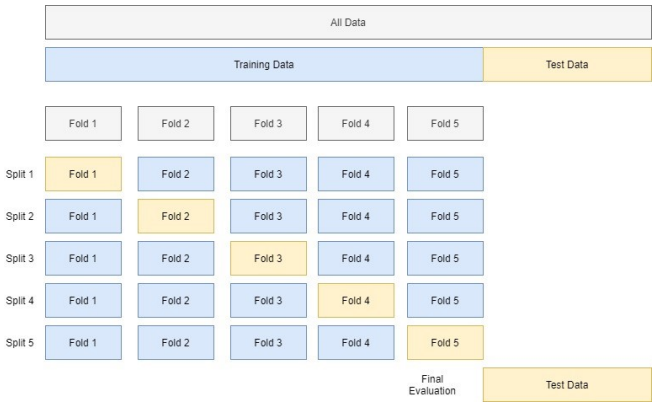

Figure 3. K-Fold Cross Validation

In this research regression techniques that will be used are Multiple Linear Regression, Support Vector Regression, Decision Tree Regression, and K-Nearest Regression. Multiple Linear Regression (MLR) consists of more than one predictor variable x and respond variable Y [13]. Both are quantitative. β is a coefficient. MLR denoted by (1).

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,n}\beta_n \quad (1)$$

Support Vector Regression (SVR) is a hyperplane model support vectors with an ε-tube, an ε-insensitive region [14]. SVR has a similar concept with Support Vector Machines (SVM). SVM can make predictions based on linear functions in a high-dimensional feature space . The algorithms are based on optimization theory by applying learning to find its origins in the study of statistical theory. The difference between SVR and SVM is in the tube. The tube contributes to optimizing solving the problem by finding the best approximates the continuous-valued function.

Decision Tree Regression is less popular than the classification trees. However, comparing to other algorithms, it is greatly competitive and usually used to solve problems in real-life [15]. Decision Trees Regression built of Root, Node, and Leaf similar to the classification trees. The difference between the regression trees and classification trees is in the output. Regression tree output is a continuous value or real number value instead of a class. For some problems, the disadvantages of the decision tree are overfitting and unstable. The advantages of a decision tree are easy to understand, interpret, useful in data exploration, and less data preparation required. A simple example of decision tree regression can be seen in Figure 4.

K-Nearest Neighbor Regression (KNN-R) is outperforming other machine learning algorithms [16]. KNN-R works similar to KNN classification, the difference is in the output. While the output of KNN classification is a class that belongs to the nearest neighbors, KNN-R output is the average of the values of k nearest neighbors. The formula of the KNN distance is shown in (2). The distance values of either 1, Manhattan distance or 2, Euclidean distance.

$$d(X,Y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \quad (2)$$

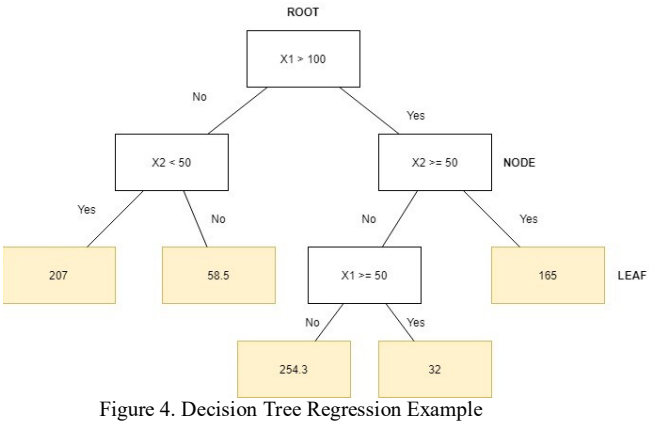

Figure 4. Decision Tree Regression Example

The last step is evaluating the results. Correlation coefficient and Mean Absolute Percentage Error (MAPE) will be used to evaluate the results. A great result with a high correlation coefficient and minimum number of MAPE is expected. MAPE formula can be seen in (3). MAPE is simply the average of the percentage errors.

$$MAPE = \frac{1}{n}\sum\frac{|e_t|}{d_t} \quad (3)$$

IV. Findings and Discussions

The research findings and discussion will be in this section. Table II shows the correlation coefficient results from the regression model on each company. Overall, it presents a good result on correlation coefficient (>0.5) except on EXCL with Multiple Linear Regression algorithm. This shows that despite the covid19 era situation, fundamental data from the company still correlate with the stock price. The highest correlation coefficient is with the Decision Tree Regression. This algorithm shows a high correlation coefficient.

TABLE I. Correlation Coefficient Result

	TLKM	EXCL	FREN	ISAT
Multiple Linear Regression	0.757	0.291	0.692	0.822
SVR	0.555	0.513	0.763	0.84
Decision Tree Regression	**0.932**	**0.86**	**0.902**	**0.977**
K-Nearest Regression	0.919	0.853	0.899	0.973

The MAPE results with the common train test split data can be seen in table III. The number hits different results with the algorithm we used. Although, decision tree regression shows an impressive number with the least MAPE results of 2.991% and the largest number is 12.859%. The second algorithm that shows good results is K-Nearest Regression. The MAPE results show a slightly different number with Decision Tree Regression and K-Nearest Regression. The algorithm that is not enough fitted with this case is multiple linear regression. This algorithm shows the

least MAPE results of 11.474% and the maximum number is 73.078%.

Table II. MAPE Results with common train test split data

	TLKM	EXCL	FREN	ISAT
Multiple Linear Regression	11.474 %	16.548 %	73.078 %	54.979 %
SVR	9.265 %	15.819 %	50.726 %	52.212 %
Decision Tree Regression	**2.991 %**	**7.745 %**	**12.859 %**	**8.314 %**
K-Nearest Regression	3.208 %	8.034 %	12.983 %	8.79 %

KFold Cross-Validation is used this time. The MAPE Results of the different splitting data can be seen in Table IV. The results show not many differences with the common train test split data. The two algorithms that predict the most minimum MAPE number are Decision Tree Regression and K-Nearest Regression. However, Multiple Linear Regression shows a slightly better result with K-Fold Cross-Validation. The least MAPE result is 11.417% and the most is 72.151%. The comparison of MAPE results of the Decision Tree Regression can be seen in Fig 5. The difference of MAPE results of Train Test Split and KFold Cross Validation is just slightly.

Table III. MAPE Results with K-Fold Cross Validation

	TLKM	EXCL	FREN	ISAT
Multiple Linear Regression	11.417 %	16.578 %	72.151 %	54.795 %
SVR	9.278 %	15.869 %	52.314 %	47.585 %
Decision Tree Regression	**2.966 %**	**7.774 %**	**12.909 %**	**8.28 %**
K-Nearest Regression	3.313 %	8.179 %	16.131 %	9.783 %

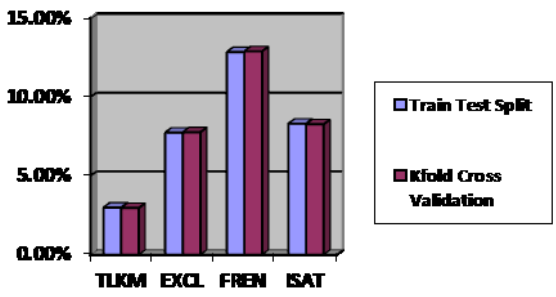

Figure 5. MAPE Results of Decision Tree Regression

V. CONCLUSSIONS AND FUTURE WORKS

In this work, we predicted the Indonesian telecommunications stock prices including covid19 era with Multiple Linear Regression, Support Vector Regression, Decision Tree Regression, and K-Nearest Regression. Historical fundamental data for 3 years used for predicting the stock prices. Although the pandemic situation greatly affects stock markets, the results show that fundamental data and the stock price has a relationship. A high correlation coefficient is produced from the Decision Tree Regression and K-Nearest Regression. Decision Tree Regression produced great results in both Train Test Split data and K-Fold Cross Validation, 2.991 % and 2.966 % repetitively. Future Works of this research will be adding more data and variables and using sentiment analysis from the news or public opinions about the stocks with a more state-of-the-art method.

REFERENCES

[1] R. B. Purnama, "Perancangan Prediksi Untuk Menentukan Indeks Harga Saham Menggunakan Jaringan Syaraf Tiruan," *Kinetik*, vol. 2, no. 2, p. 125, 2017.

[2] D. Zhang, M. Hu, and Q. Ji, "Financial markets under the global pandemic of COVID-19," *Financ. Res. Lett.*, no. April, p. 101528, 2020.

[3] A. Rifa'i and D. Mahdiana, "Optimized fuzzy backpropagation neural network using genetic algorithm for predicting indonesian stock exchange composite index," *Int. Conf. Electr. Eng. Comput. Sci. Informatics*, pp. 195–199, 2019.

[4] Z. Rustam, D. F. Vibranti, and D. Widya, "Predicting the direction of Indonesian stock price movement using support vector machines and fuzzy Kernel C-Means," *AIP Conf. Proc.*, vol. 2023, no. October, 2018.

[5] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia, and D. C. Anastasiu, "Stock price prediction using news sentiment analysis," *Proc. - 5th IEEE Int. Conf. Big Data Serv. Appl. BigDataService 2019, Work. Big Data Water Resour. Environ. Hydraul. Eng. Work. Medical, Heal. Using Big Data Technol.*, pp. 205–208, 2019.

[6] Y. Farida, "Sistem Prediksi Saham Menggunakan Adaptive Neuro Fuzzy Inference System (Studi Kasus Saham Mingguan PT Astra Agro Lestari,Tbk)," *Syst. Inf. Syst. Informatics J.*, vol. 2, no. 2, pp. 35–39, 2016.

[7] I. R. Parray, S. S. Khurana, M. Kumar, and A. A. Altalbe, "Time series data analysis of stock price movement using machine learning techniques," *Soft Comput.*, vol. 0123456789, 2020.

[8] H. L. Siew and M. J. Nordin, "Regression techniques for the prediction of stock price trend," *ICSSBE 2012 - Proceedings, 2012 Int. Conf. Stat. Sci. Bus. Eng. "Empowering Decis. Mak. with Stat. Sci.*, pp. 99–103, 2012.

[9] T. Cakra, "Stock Price Prediction using Linear Regression based on Sentiment Analysis," *Int. J. Sci. Eng. Res.*, vol. 6, no. 3, pp. 1655–1659, 2015.

[10] Z. Rustam and P. Kintandani, "Application of Support Vector Regression in Indonesian Stock Price Prediction with Feature Selection Using Particle Swarm Optimisation," *Model. Simul. Eng.*, vol. 2019, 2019.

[11] M. Ouahilal, M. El Mohajir, M. Chahhou, and B. E. El Mohajir, "A novel hybrid model based on Hodrick–Prescott filter and support vector regression algorithm for optimizing stock market price prediction," *J. Big Data*, vol. 4, no. 1, pp. 1–22, 2017.

[12] M. Z. Asghar, F. Rahman, F. M. Kundi, and S. Ahmad, *Development of stock market trend prediction system using multiple regression*, vol. 25, no. 3. Springer US, 2019.

[13] D. J. Olive, *Linear regression*. 2017.

[14] D. Mezghani, S. Boujelbene, and N. Ellouze, "Evaluation of SVM Kernels and Conventional Machine Learning Algorithms for Speaker Identification," *J. Hybrid Inf. Technol.*, vol. 3, no. 3, p. 12, 2010.

[15] M. Czajkowski and M. Kretowski, "The role of decision tree representation in regression problems – An evolutionary perspective," *Appl. Soft Comput. J.*, vol. 48, pp. 458–475, 2016.

[16] L. Vanneschi, Leonardo ; Castelli, Mauro; Manzoni, Luca; Silva, Sara; Trujillo, "Is k Nearest Neighbours Regression Better Than GP?," in *23rd European Conference, EuroGP 2020*, 2020, vol. 1, pp. 244–261.