Recursive Journal of Informatics

Vol. 1, No. 2, September 2023

p-ISSN 2963-5551 https://journal.unnes.ac.id/sju/index.php/rji e-ISSN 2986-6588

Stock Return Prediction Using Voting Regressor Ensemble Learning

Ramadhan Ridho Arrohman[1], Riza Arifudin[2]

[1,2]Department Computer Science, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang

Abstract. The value of return on stock prices is often used in predicting profits in the process of buying and selling shares based on the calculation of the return on investment. The calculation of the value of return on stock prices can be predicted automatically at certain periods, both weekly and daily
Purpose: The problem faced is determining a good algorithm for making predictions due to fluctuating data on stock prices making it difficult to predict.
Methods/Study design/approach: The stages carried out by the researcher include the data preprocessing stage and then proceed to the Exploratory Data Analysis (EDA) stage to get a pattern from the data, followed by the modeling stage on the data. This research was developed using the Python programming language where the models used to make predictions can be obtained in real-time.
Result/Findings: The results obtained in this study show that the Voting Regressor has the best model with an error rate of 0.032523 using Root Mean Square Error (RMSE).
Novelty/Originality/Value: This study can be further developed to automatically predict stock return values in the future.

Keywords: Stock Return, Ensemble Learning, Regression, Stock Market
Received April, 17 2023 / **Revised** Mei 05, 2023 / **Accepted** September 14, 2023

INTRODUCTION

The stock market is a collection of stock exchanges around the world where investors and the public buy and sell their shares publicly. Stock prices fluctuate very much because they are influenced by the law of demand and supply [1]. The stock price is a form of market analysis of the future income earned by the company. Even though the nature of stock prices fluctuates, it is possible to predict the return value of these stocks using past data [2].

Stock return prediction is carried out using a regression algorithm because the regression algorithm has the ability to produce output in the form of a numeric value. Machine learning algorithms that can be used in the process of determining stock return values are machine learning algorithms, for example the Support Vector Machine (SVM) algorithm, Artificial Neural Network (ANN), Decision Support System (DSS) as well as several boosting and bagging algorithms [3].

One of the boosting algorithms used in the regression method is the Adaboost regressor (ABR) algorithm. Adaboost regressor has the advantage of being able to overcome noise and outliers in the data. Unfortunately, the Adaboost regressor algorithm is not suitable for use on complex data because it makes the model overfitting [4]. Random forest regressor (RFR) has the advantage of generating many decisions based on a decision tree algorithm that can affect better accuracy and making, but this algorithm requires hyperparameter selection for better performance and makes the model overfitting if the model is too complex [5]. Gradient Boosting Regressor (GBR) can deal with dimensional data efficiently, produce learning algorithms and collect the results of these algorithms to have better accuracy. Unfortunately, the gradient boosting regressor requires need expensive

[1]*Corresponding author.
Email addresses: ramadhan.ridho.16@students.unnes.ac.id (Arrohman)
DOI: 10.15294/rji.v1i2.68048

computational data resources for large datasets [6]. Support Vector Regression (SVR) is a powerful algorithm that has advantages in handling data dimensionality efficiently, develops good accuracy results even with small datasets, and it is suitable for use on nonlinear data, unfortunately this algorithm requires optimal selection of hyperparameters for optimal performance and requires expensive computational power for large datasets [7].

Combining machine learning algorithms to obtain optimal results is called the ensemble learning method. One of the advantages of ensemble learning is that it can help reduce the risk of overfitting. By combining multiple models, ensembles can capture data patterns better and avoid being overly influenced by noise or outliers. Another advantage is that the ensemble is more robust against data changes, such as missing values, because different models may be able to compensate for each algorithm's weaknesses [8]. On the other hand, one of the disadvantages of ensemble learning it can increase the complexity of a model and make it more difficult to interpret. Ensembles may also require more computational resources to train, which can be a limiting factor in some applications [9].

METHODS

The method used in this study is an ensemble learning technique to get the best results and avoid overfitting the data. In this study the authors used a logarithmic transformation to obtain predictions of the value from stock returns at Microsoft companies for the last decade, namely December 30, 2012 to December 31, 2022. The Ensemble Learning algorithm used in this study is the Adaboost Regressor, Gradient Boosting Regressor, Random Forest algorithms, Support Vector Regressor and several of these algorithms will be combined using the Voting Regressor algorithm.

Stock Market Returns

In the financial sector, especially the stock market, investors often ask questions about how to define and understand the risks of financial assets such as stocks and portfolios. It is not enough just to measure the risk, but the right comparison is needed to assist investors in making financial decisions [10]. Stock return refers to the profit or loss earned by an investor on their investment in a certain stock over period of time. Usually expressed as a percentage and calculated by dividing the difference between the current share price and the original purchase price divided by the initial purchase price. Return analysis is the main analysis in finance. The return value serves as a comparison. This comparison is obtained from the relationship between securities and asset analysis. The value of returns on stocks is well distributed, while stock prices are not distributed [11].
According to Miskolczi [12], the calculation of return values can be divided into linear returns and logarithmic returns.

Linear Returns

Linear returns, also called simple returns, are linear calculations of asset values. Linear returns have a property value calculation that has asset-additive in nature, which makes it easier to combine return values . The calculation of the return value can be seen in Equation 1.

$$L_t = \frac{P_{t+1}}{P_t} - 1 \quad (1)$$

Information

L_t= Linear return at time t.

P_t= Investment price at time t.

P_{t+1}= Investment price at the next time.

If the weight value is added to the security, the portfolio return can be seen in Equation 2.

$$L_{t,p} = w_1 P_{t,1} + w_2 P_{t,2} + \cdots + w_n P_{t,n} \quad (2)$$

Information

$L_{t,p}$ = Portfolio return value using linear returns s

Linear returns are often used in risk analysis, calculation of performance values, and portfolio optimization. In linear analysis there is also a calculation using the compound return method.

p-ISSN 2963-5551 https://journal.unnes.ac.id/sju/index.php/rji e-ISSN 2986-6588

Compound return is a combination analysis for a certain period. The calculation of the compound return value can be seen in Equation 3.

$$C_t^K = \ln\left(\frac{P_{t+1}}{P_t}\right) = C_t + C_{t+1} + \cdots + C_{t+k-1} \quad (3)$$

Information

C_t^K= Combined return value in period K, at time t

P_t= Investment price at time t.

Compound Return is a time-additive. This states return method can increase in line with the existing time to get the total return in a certain period.

Logarithmic Returns

The assumption that is often used for several assets in stock prices is that the price returns are normally distributed. The main reason is because the price cannot be assumed to be a negative value or the stock price is zero, because the price value cannot be negative. Therefore, the calculation of the return value can be done so that it makes it easier for us to see positive and negative values in determining whether there is a possibility of getting profits in trading at that time.

Stock prices that change over time and the variance of a price are better off using the conditional distribution than the marginal distribution shown in Equation 4.

$$\log\left(\frac{P1}{P2}\right) \sim N(\mu, \sigma^2) \quad (4)$$

If the logarithm of stock returns is normally distributed, it can be seen in Equation 5.

$$1 + r1 = \frac{p1}{p0} = \exp^{\log\left(\frac{p1}{p0}\right)} \quad (5)$$

The advantage of Equation 5 on the normal distribution is that this method is easy to use when a variable is normally distributed depending on certain objectivity.

Raw-Log Return Equality

When the value of return on investment (ri) small, then the formula used can be seen in Equation 6.

$$\log(1 + ri) \approx ri, ri \ll 1 \quad (6)$$

Equation 7 is a good value calculation and relevant for investors who want to sell or buy shares in small amounts. This is also supported by a good programming language where researchers will later use this formula because it is easier to code than Equation 5.

Algorithm Complexity

If the shares traded have a daily or weekly amount of time, then the calculation of the share value must use the compound return method as mentioned in Equation 3 using the logarithmic calculation method. Equation 7 explains the compound return method on daily or weekly trading systems.

$$(1 + r1)(1 + r2) \ldots (1 + rn) = \prod_i (1 + ri) \quad (7)$$

This equation is more complicated to code. For this reason, it can be replaced with the log return value contained in Equation 8.

$$\log(1 + ri) = \log\frac{pi}{pj} = \log(pi) - \log(pj) \quad (8)$$

Equation 8 cannot be applied to the calculation of the return value which requires time complexity in it. Therefore, to calculate weekly or daily return values, the authors apply this concept to Equation 9.

$$\sum_{\text{ilog}} n(1 + ri) = \log(1 + r1) + \log(1 + r2) + \cdots + \log(1 + rn) \quad (9)$$

The next step is to break down the logarithmic. Equation 2.10 to a simpler form like the equation shown in Equation 10.

$$\log(p_1) - \log(p_0) + \log(p_2) - \log(p_1) + \cdots + \log(p_n) - \log(p_{n-1}) = \sum_{i=1}^{n} \log(p_n) - \log(p_{n-1}) \quad (10)$$

There are so many benefits of log returns than calculated values using simple returns. However, compounded returns are not suitable for continuous time [13]. The use of log returns is a calculation of return values that focuses on increasing interest rates.

Ensemble Learning

Ensemble learning refers to the combination of many machine learning model algorithms to improve the overall predictive results and reduce the risk of overfitting. Ensemble learning is often used for various fields of science, for example finance, healthcare, and computer vision. One of the first ensemble learning algorithms used was the boosting algorithm which was discovered by Robert Schapire in 1990. The boosting algorithm involves an iterative training process, from several poor models. For example, in the decision tree algorithm, the researcher combines the results of the predictive models to create a more robust model than before. According to Leo Breiman [5] the Random Forest algorithm involves a lot of training data by combining their predictions to make a more accurate model.

Ensemble learning has become very popular in recent years as machine learning models become more complex with increasingly large datasets. Researchers have developed several ensemble learning algorithms, for example Stacking, Bagging, and Gradient Boosting which are designed to increase the accuracy of machine learning models [9] . SVM is a machine learning algorithm based on statistical learning theory [14] . SVM is one of the supervised learning algorithms used for data analysis using regression and classification models [15]. The SVM used in the classification technique is called the Support Vector Classification (SVC) technique, while the SVM used in the regression technique is called the Support Vector Regression (SVR). The advantage of SVM is the ease in solving classification and regression problems with linear and non-linear data. SVM is used to find the best hyperplane by maximizing the distance between classes.

Adaboost Regressor (ABR) is a Machine Learning algorithm that used to perform regression on data using the boosting method [4]. The boosting method itself is a Machine Learning technique that is used to combine several weak machine learning models into one strong model. Adaboost Regressor works by developing a simple regression model, then the model will be tested again to see how accurate it is in predicting the data. After that, the incorrectly predicted data will be given greater weight so that the next model will focus more on predicting the incorrect data [16]. This process is repeated until the resulting model is sufficiently accurate.

Gradient Boosting Regressor (GBR) is one of the popular Machine Learning algorithms for performing regression on data using the boosting method. The Gradient Boosting Regressor works by combining several weak regression models into a strong model by adding new models gradually and optimizing each model based on the residual error from the previous model [17]. The Gradient Boosting method itself is a Machine Learning technique that works by gradually optimizing the model using the gradient error function. At each iteration, gradient boosting looks for a new model that can reduce errors in the previous iteration.

Random Forest Regressor (RFR) is a Machine Learning algorithm that uses ensemble learning techniques, namely combining several machine learning models into one more powerful model. In the Random Forest Regressor algorithm, several decision trees that are formed randomly will be combined to produce a prediction [5]. The Random Forest Regressor takes a random sample from the dataset for each created tree and uses the sample to construct a decision tree. Each decision tree that is formed will provide predictions independently and the results will be combined to produce more accurate predictions [18].

Voting Regressor is one of the ensemble learning techniques in Machine Learning that is used to predict the target variable by combining several different regression models into one more robust model [19] . The Voting Regressor combines several regression models by giving weights to each model, so that the prediction results are based on the results of the selection of these models. Regressor voting is used when we want to compare several different regression models and choose the model that gives the best predictive results. In this study, the authors combined several regression models, namely ABR , GBR, RFR, and SVR into one ensemble model, namely the Voting Regressor algorithm.

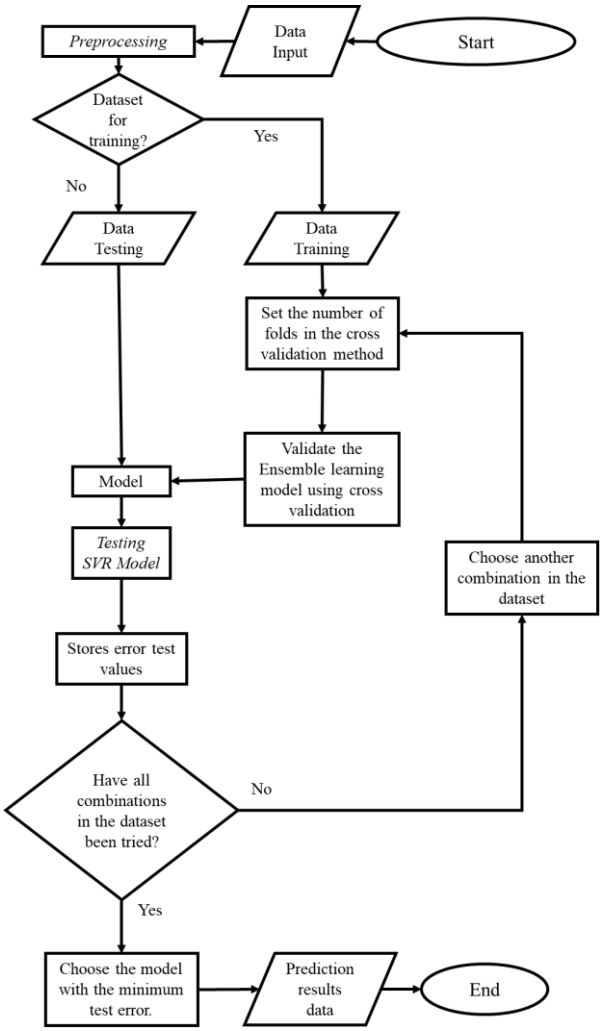

Figure 1. Flowchart of Ensemble Learning

Metrics Evaluation

In assessing a prediction model, it is necessary to evaluate the model how well the model has been made. This modeling will be assessed using the Mean Square Error (MSE). The following is a further explanation of the evaluation matrix.

1) Mean Squared Error represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals. Equation 11 show how MSE will be calculated.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}) \quad (11)$$

Information:
y_i= ActualValue
$\hat{y}$= Predictive Value

2) RMSE is the standard deviation of the error prediction value by calculating its square root because RMSE is a step in calculating the standard deviation of a residual. The RMSE formula can be seen in. Equation 12 show how RMSE will be calculated.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2} \quad (12)$$

Information:
y_i= ActualValue
$\hat{y}$= Predictive Value

3) Mean Absolute Error (MAE) represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset. Equation 13 show how MAE will be calculated.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}| \quad (13)$$

Information:
y_i= ActualValue
$\hat{y}$= Predictive Value

RESULTS AND DISCUSSION

The data obtained for this study can be found on the Yahoo Finance website. This data can be accessed by anyone, so it is necessary to do web scraping to obtain data. Web scraping can be done using a library yahoo finance and pandas_datareader library. The data collection process uses a period of 10 years with a time period from 30 December 2012 to 31 December 2022, so that 2519 data samples consist of numeric type.

The data obtained by the author consists of company stock data taken from Microsoft, IBM, and Google. Then the researcher also took data on the exchange rate of the original company's currency with the exchange rate of the foreign currency of the destination country. Companies Microsoft, IBM, and Google are companies based in the United States of America where the author is looking for the most influential foreign partnerships in the United States. Therefore, the author takes the example of 2 Japanese exchange rates (yen) to represent the Asian continent and the exchange British currency (pounds) to represent the European continent.

Data processing

Furthermore, the dataset needs to be preprocessed where the purpose of this prediction is to predict the weekly return value. The researcher counted the number of trading system working days on the stock market where the stock market is open for 5 working days. Therefore, the first step is to initialize the parameters for stock returns with the number 5. The next step is to transform the logarithm in Equation 10 to get the return value for 5 working days. The next stage is merging the data with the dataset that has been obtained from the Federal Reserve bank. The next step is to

combine the dataset taken from the exchange currency and the combined stock price index. Then delete the null value to get the overall stock price return value.

Figure 2 describes the dataset that has gone through the preprocessing stage. The dataset in Figure 2 calculates the overall return value over a period of 5 working days. For example, in the first line it is written that the value in the "MSFT_Pred" column on the date (2013-08-09T00:00:00) has a value of -0.0209, so the return value obtained on the previous five days is negative or the price has decreased after the date (2013-08-04T00:00:00) as much as 2 percent. The next stage is the analysis stage or called Exploratory Data Analysis (EDA). EDA stages are descriptive statistics stage, data visualization stage, and time series analysis.

	MSFT_pred	GOOGL	IBM	DEXJPUS	DEXUSUK	SP500	DJIA	VIXCLS
2013-08-09T00:00:00	-0.0209	-0.0180	-0.0333	-0.0287	0.0149	-0.0107	-0.0150	0.1128
2013-08-16T00:00:00	0.0887	-0.0383	-0.0133	0.0140	0.0073	-0.0213	-0.0226	0.0691
2013-08-23T00:00:00	-0.0396	0.0154	0.0004	0.0099	-0.0028	0.0046	-0.0047	-0.0275
2013-08-30T00:00:00	-0.0535	-0.0272	-0.0171	-0.0036	-0.0071	-0.0185	-0.0134	0.1962
2013-09-10T00:00:00	0.0165	0.0324	0.0142	0.0079	0.0120	0.0266	0.0238	-0.1338
2013-09-17T00:00:00	-0.0147	-0.0029	0.0294	-0.0106	0.0107	0.0123	0.0220	0.0000

Figure 2. Stock return data after pre-processing

Evaluation result

After going through the EDA process, the author performs modeling using the ensemble learning method which consists of 4 algorithms, namely Adaboost, Gradient Boosting, Random Forest, and Support Vector Regression to get better results. Table 1 is the result of an evaluation using MSE, RMSE, and MAE.

Table 1. Error rate from different algorithm

Model	MSE	MAE	RMSE
Adaboost Regressor (ABR)	0.001129	0.024235	0.032831
Gradient Boosting Regressor (GBR)	0.00119	0.024788	0.03406
Random Forest Regressor (RFR)	0.001092	0.023781	0.03254
Support Vector Regression (SVR)	0.001368	0.02823	0.036613
Voting Regressor (Proposed)	0.001099	0.02403	0.032523

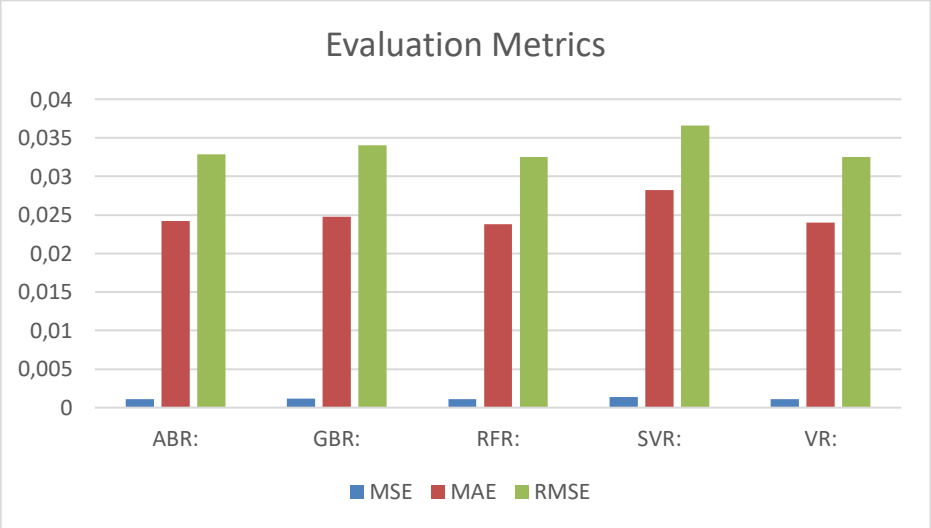

Figure 3. Evaluation metrics

In Figure 3 the evaluation using MSE and MAE has the best results were obtained using the RFR algorithm while the evaluation used RMSE the best results are obtained using the Voting Regressor algorithm.

CONCLUSION

Based on the implementation of several tests on the regression model using the algorithm above, it is obtained. The results of the data testing error rate in the experiment using the Adaboost, Gradient Boosting, Random Forest, Support Vector Regression, and Voting Regressor models so that the value with the best model error rate is obtained using the Random Forest Regressor algorithm with an MSE value of 0.001092 and an MAE value of 0.023781 for the best algorithm based on the RMSE Evaluation value obtained a value of 0.032523 for the Voting Regressor algorithm.

REFERENCES

[1] P. Chhajer, M. Shah, and A. Kshirsagar, “The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction,” Decis . Anal. J. , vol. 2, no. November 2021, p. 100015, 2022, doi: 10.1016/j.dajour.2021.100015.

[2] AM More, PU Rathod, RH Patil, DR Sarode, and B. Student, “Stock Market Prediction System using Hadoop,” Int. J.Eng. sci. Comput. , vol. 8, no. 3, pp. 16138–16140, 2018, [Online]. Available: http://ijesc.org/.

[3] DP Gandhmal and K. Kumar, “Systematic analysis and review of stock market prediction techniques,” Comput. sci. Rev. , vol. 34, p. 100190, 2019, doi: 10.1016/j.cosrev.2019.08.001.

[4] Y. Freund and RE Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," J. Comput. syst. sci. , vol. 55, no. 1, pp. 119–139, 1997, doi: 10.1006/jcss.1997.1504.

[5] L. Breiman, “A Data Mining Based System for Transaction Fraud Detection,” 2021 IEEE Int. Conf. Consum. electrons. Comput. Eng. ICCECE 2021 , pp. 542–545, 2021, doi: 10.1109/ICCECE51280.2021.9342376.

[6] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. , vol. 13-17-August, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[7] R. Gholami and N. Fakhari, Support Vector Machine: Principles, Parameters, and Applications , 1st ed. Elsevier Inc., 2017.

[8] Z. -H. Zhou, Ensemble methods: foundations and algorithms . CRC Press, 2012.

[9] TG Dietterich, “Ensemble methods in machine learning,” in Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21--23, 2000 Proceedings 1 , 2000, pp. 1–15.

[10] A. Megaritis, N. Vlastakis, and A. Triantafyllou, “Stock market volatility and jumps in times of uncertainty,” J. Int. Money Financec. , vol. 113, p. 102355, 2021, doi: 10.1016/j.jimonfin.2021.102355.

[11] RS Hudson and A. Gregoriou, “Calculating and comparing security returns is harder than you think: A comparison between logarithmic and simple returns,” Int. Rev. financec. Anal. , vol. 38, pp. 151–162, 2015, doi: 10.1016/j.irfa.2014.10.008.

[12] P. Miskolczi, “Note on simple and logarithmic returns,” Appl. Studs. Agribus. Commer. , vol. 11, no. 1–2, pp. 127–136, 2017, doi: 10.19041/apstract/2017/1-2/16.

[13] A. Meucci, “Quant Nugget 2: Linear vs. Compounded Returns,” Common pitfalls in Portfolio Management , GARP risk professional, pp. 1–5, 2010.

[14] VN Vapnik, The Nature of Statistical Learning Theory , Second. New York: Springer US, 1995.

[15] M. Ouahilal, M. El Mohajir, M. Chahhou, and BE 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, doi: 10.1186/s40537-017-0092-5.

[16] Y. CAO, Q. -G. MIAO, J. -C. LIU, and L. GAO, “Advance and Prospects of AdaBoost Algorithm,” Acta Autom. Sin. , vol. 39, no. 6, pp. 745–758, 2013, doi: 10.1016/s1874-1029(13)60052-x.

[17] JH Friedman, “Greedy function approximation: a gradient boosting machine,” Ann. Stats. , pp. 1189–1232, 2001.

[18] A. Cutler, DR Cutler, and JR Stevens, “Random forests,” Ensemble Mach. Learn. Methods Appl. , pp. 157–175, 2012.

[19] Z. Ali, I. ur Rehman, and Z. Jaan, "An Empirical Analysis on Software Development Efforts Estimation in Machine Learning Perspective," ADCAIJ Adv. District. Comput. Artif. Intell. J. , vol. 10, no. 3, pp. 227–240, 2021.