Jurnal Mantik, 6 (3) (2022), ISSN 2685-4236 (Online)

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik

Prediction of the Number of Course Participants Using Random Forest Regression Algorithm

Rina Septiriana[1], Anggi Perwitasari[2], Tursina[3]
[1,2,3]Teknik/Informatika, Universitas Tanjungpura, Indonesia

A R T I C L E I N F O

Article history:

Received Sep 28, 2022
Revised Oct 5, 2022
Accepted Oct 26, 2022

Keywords:

Course Participant
Ensemble Learning
Prediction
Random Forest Regression
Regression

ABSTRACT

Classroom management is the process of using resources effectively to achieve goals. Planning the schedule is not easy because sometimes, after a schedule designed and when the schedule is published and used, there are problems where the class division of a course is not right on target. This study used the random forest regression method to predict the number of class participants. Data pattern affects the accuracy of calculating the predicted value. The best RMSE and MAE results in the Matematika Dasar Course are 6,51 for RMSE and 2,12 for MAE. At the same time, the prediction of course participant number is 73,18.

Corresponding Author:

Rina Septiriana,
Teknik/Informatika,
Universitas Tanjungpura,
Profesor Dokter H. Hadari Nawawi Street, Pontianak, Kalimantan Barat, 78124, Indonesia.
Email: rinaseptiriana@informatika.untan.ac.id

1. INTRODUCTION

Classroom management is the process of using resources effectively to achieve goals. One way to achieve these goals is by creating an effective classroom atmosphere by carrying out quality and well-implemented learning plans (Astuti, 2019). Setting the course schedule is one form of learning planning where a quality schedule arrangement requires a good design. Planning the schedule is not easy because sometimes, after a schedule is designed and when the schedule is published and used, there are problems where the class division of a course is not right on target. So when a class is opened, it turns out that the number of enthusiasts with the number of classes is not balanced where several classes are crowded or some classes are quiet. This can make it difficult for related parties where the time for filling out the KRS becomes longer to accommodate students who run out of class.

To assist in designing the proper course schedule is to predict the number of teaching participants who take the course so that the provision of the number of classes opened for each particular course can be predicted well. In order to reduce error—the difference between what really occurs and the projected outcome—prediction is the systematic estimation of what is most likely to occur in the future based on known past and present events. The goal of prediction is to come up with responses that are as close to what will happen as feasible rather than having to give a definitive answer to what will happen (Herdianto,2017). Ensemble Learning is one of the methods in machine learning,

where it works by combining several learning algorithms to produce predictions that can improve accuracy [Syahputra et al., 2022]. There are several Ensemble Learning types: Bagging, Boosting, Stacking, and Voting. Random Forest Regression is a form of ensemble learning in bagging and a supervised learning algorithm for regression. The ensemble learning method is a technique that combines several predictions from various machine learning algorithms to produce predictions that are more accurate than a single model (Indahyanti et al., 2022).

Random Forest Regression is a powerful and accurate model. Random Forest Regression is suitable for many problems, including features with nonlinear relationships. Therefore, the random forest regression method was used in this study to predict the number of class participant based on the method's advantages in a simple model parameter for sequential data (Rianto & Wahono, 2015).

2. RESEARCH METHOD

2.1 Random Forest Regression

The end outcome of random forest regression is the average of all the decision trees, which are created by running hundreds to thousands of decision trees through their respective regression functions. Breiman et al. developed the statistical model known as the decision tree (1984). The decision tree is a non-parametric model that operates on the assumption that it is not constrained in its ability to learn, hence the tree expands in response to the complexity of the input (Alpaydin, 2014). Decision nodes and leaf nodes make up each decision tree. Each sample is assessed by the decision node using the test function, and depending on the characteristics of the sample, it passes on a different branch. Y is the scalar output, and X represents the input vector containing m features with X=x1,x1,…,xm (Shu et al, 2022).

$$S_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}, \; X \in \mathbb{R}^m, \; Y \in \mathbb{R}. \quad (1)$$

The input data is divided using an algorithm at each node during training in order to maximize the split function's parameters with the set Sn. The decision tree must do the best split possible among all the variables in the initial stage. Each node applies its own split function to the new input x as the division process moves from the root to the nodes. The final node is reached by continually doing this.

The decision tree can be extended to use random forest regression, which can result in more accurate predictions. Random Forest (RF) gathers numerous uncorrelated decision trees during the training phase. It is known as a random forest because each tree in the RF is created from a random subset of the predictor.

The bagging method, an ensemble learning technique used by RF, integrates all the decision trees that were produced. A technique called bagging can be applied to regression algorithms to lower the variance of predictions and increase the performance of predictions (Breiman, 1996). In each decision tree, RF gathers samples through a process known as bootstrapping. Numerous bootstrap samples $S_n^{\Theta_1}, \ldots, S_n^{\Theta_q}$ with probability 1/n are used in the bagging technique. To create a series of prediction trees q, namely $\hat{h}(X, S_n^{\Theta_1}), \ldots, \hat{h}\left(X, S_n^{\Theta_q}\right)$, the preceding decision tree is combined with each sample. For each tree, the ensemble generates an output with the formula $\hat{Y_1} = \hat{h}(X, S_n^{\Theta_1}), \ldots, \hat{Y_q} = \hat{h}\left(X, S_n^{\Theta_q}\right)$. The average output of each tree is calculated as part of the aggregation process, and its equation is in Equation (2) (Rodriguez-Galiano et al.,2012; Lahouar & Slama, 2017).

$$\hat{Y} = \frac{1}{q} \sum_{l=1}^{q} \hat{Y_l} = \frac{1}{q} \sum_{l=1}^{q} \hat{h}\left(X, S_n^{\Theta_l}\right), \quad (2)$$

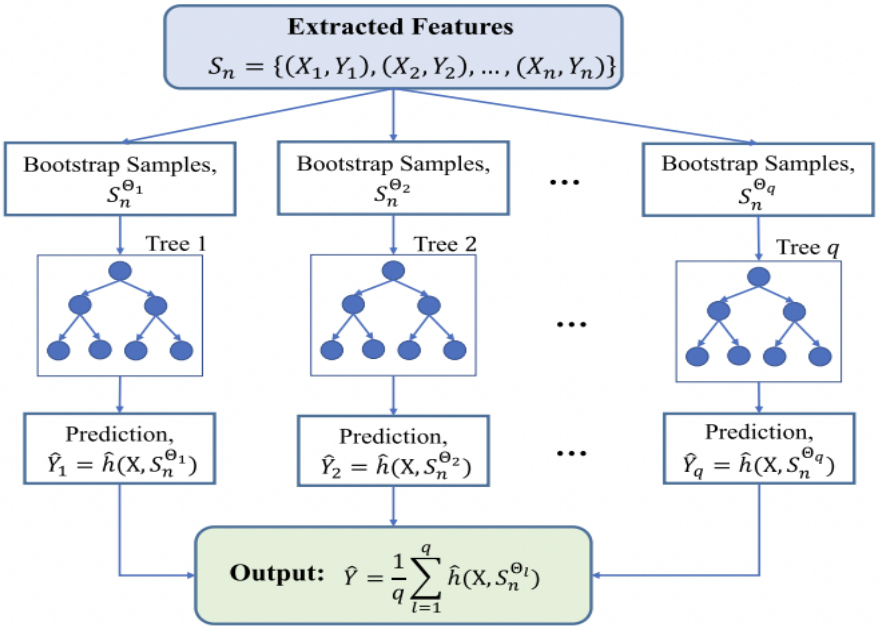

Figure 1 Illustration of Random Forest Construction (y.Li,2018)

Based on Figure 1, the Random Forest construction, according to Liaw & Wiener (2002), is as follows:

a. Create a bootstrap n treesample from the original data, with each bootstrap subset containing about two-thirds of the original dataset's elements.
b. Use a bootstrap sample with alterations to each node, such as the random sample m try on the predictor, and select the optimal split between these variables to build an untrimmed regression tree.
c. By combining forecasts from the ntree tree, forecast fresh data (e.g., most votes for classification and average for regression)

2.2 Evaluation Metrics

After the model training is finished, the model evaluation phase can begin. The evaluation data used to assess the model cannot be the same as the training data. The results of testing a trained model against this evaluation data will reveal its true accuracy value. The performance metrics employed in the research below include: (y.Li, 2018; Chai & Gunawan, 2022).

a. Mean Absolute Error (MAE)
   MAE is an average absolute difference between the results tested and the predicted value. Each fault has the same weight on the MAW. The smaller the MAE value, the more accurate the prediction results. Where n indicates the number of observations, y represents the predicted value which can be seen in Equation (3). (Na et al.,2015; Khaleghi et al., 2019)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (3)$$

b. Root Mean Square Error (RMSE)
c. RMSE is used to measure the difference between the predicted value with the model and the observed value. MSE is more popular than MAE

because it focuses on large errors because they are squared so that they have a greater impact on larger errors than smaller errors which can be seen in Equation (4)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \quad (4)$$

3 RESULTS AND DISCUSSIONS

3.1 Data Analysis Results

The data used is the academic data of students in the Department of Informatics from 2014 to 2022. There are 41 compulsory courses and 11 elective courses, which are divided into eight semesters. Where there are 23 courses in odd semesters and 18 courses in even semesters. In this study, four courses were used in the odd semester, namely Basic Mathematics, Artificial Intelligence, Mathematical Logic, and Basic Programming. It is used to see the prediction results usiang Random Forest Regression. The description of the data used can be seen in Table 1.

Table 1 Description of Student Academic Data

	idmahasiswa	sks	absen	tugas	uts	uas	nilaitotal	nilaimutu	ips	ipk
count	7390.000000	7390.000000	5857.000000	5856.000000	5860.000000	5856.000000	5872.000000	6046.000000	7390.000000	7390.000000
mean	146941.475778	2.815968	19.094439	81.541995	76.048439	75.827156	79.220445	3.535891	2.860437	3.489398
std	9072.875483	0.700880	17.592921	14.346250	18.789904	18.899456	13.870884	0.868739	1.458437	0.473143
min	136307.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	139903.000000	2.000000	12.000000	80.000000	70.000000	70.000000	76.000000	3.500000	2.870000	3.360000
50%	144857.000000	3.000000	14.000000	80.000000	80.000000	80.000000	82.000000	4.000000	3.570000	3.590000
75%	149334.000000	3.000000	16.000000	90.000000	86.000000	85.000000	85.900000	4.000000	3.800000	3.780000
max	167023.000000	4.000000	100.000000	100.000000	100.000000	100.000000	100.000000	4.000000	4.000000	4.000000

In the description in Table 1, the highest standard deviation values are in the UTS and UAS attributes, which are used as the basis for data processing for calculating the number of course participants for each semester. The data is then prepared according to the data needs, and the number of participants is determined for each subject to be observed. The number of participants in the courses used in this study can be seen in Figure 2.

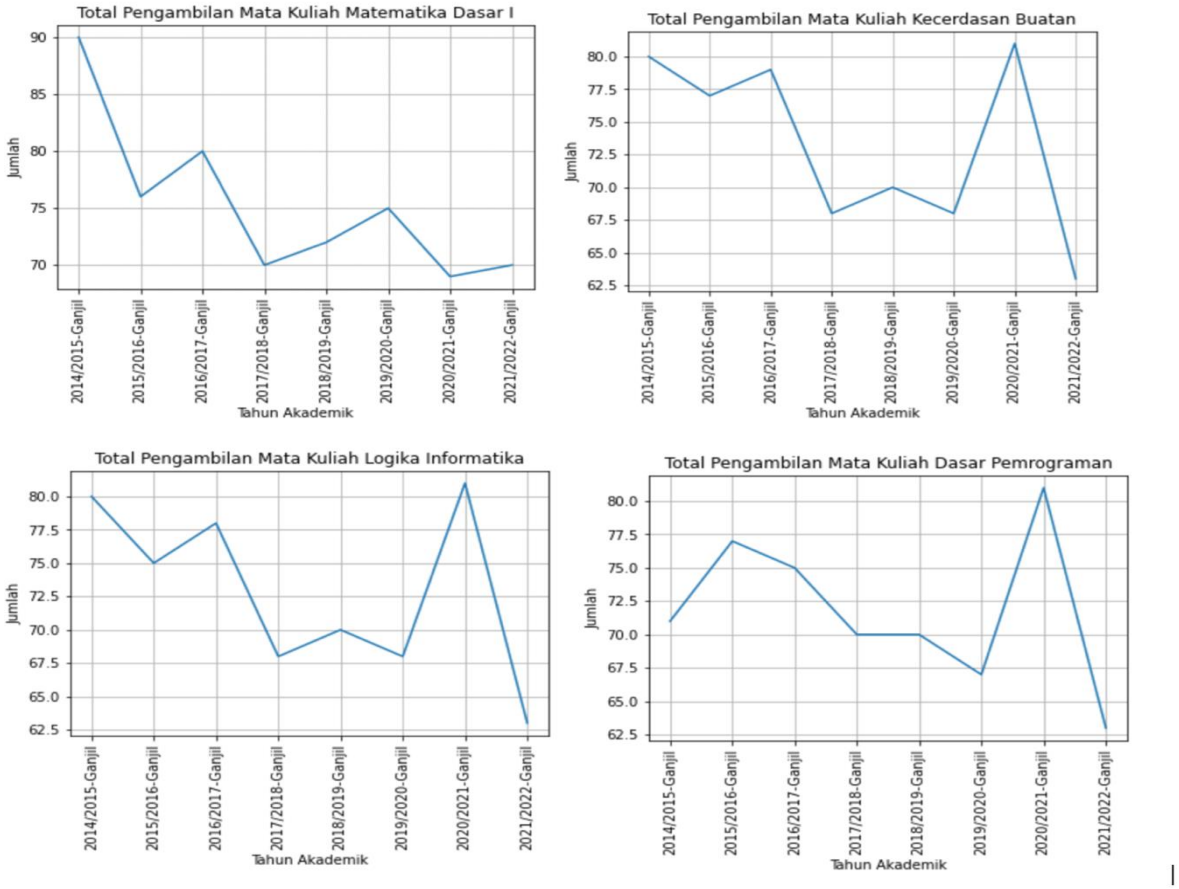

Figure 2. Graph of the number of courses taken each semester

3.2 Data Prediction Results

Existing data were tested using the Random Forest Regression algorithm using a distribution of 80:20, namely 80% training data and 20% testing data. The maximum number of trees used is 100, which aims to provide the best performance in determining predictions. After the prediction results are obtained, evaluation metrics are calculated using MAE and RMSE. The prediction results can be seen in Table 2. In comparison, one of the data prediction processes using Random Forest Regression can be seen in Figure 3.

Table 2. Prediction Results and Evaluation Metrics

Course	Prediction Result	RMSE	MAE
Informatics Logic	70,95	9,52	2,52
Artificial intelligence	71,1	7,05	2,21
Basic mathematic	73,18	6,51	2,12
Basic Programming	70,14	9,21	2,45

In Table 2, it can be seen the prediction results of the four courses tested. The best RMSE results are in the Matematika Dasar course. Based on Figure 2, it can be seen that the data pattern in the "Matematika Dasar" course has a linear pattern that looks decreasing without an increasing pattern so that it gets the best results. The results of the visualization can be seen in Figure 4, where it can be seen that there is a decreasing pattern even though there are some increasing patterns, but the increase is not significant enough so that it can get the best prediction results compared to the prediction results for other courses.

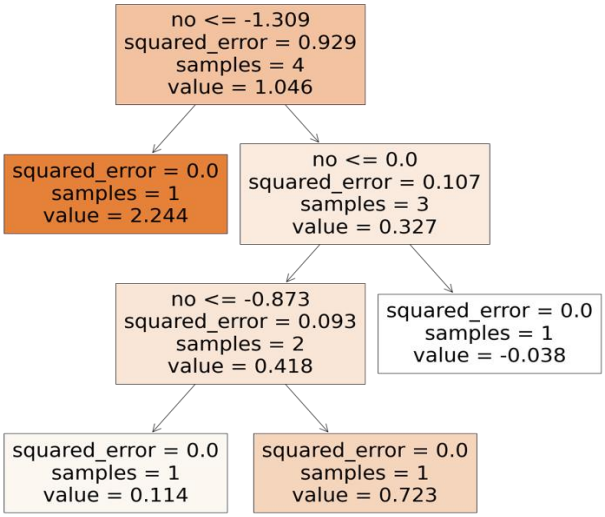

Figure 3. Illustration of Random Forest Regression Process

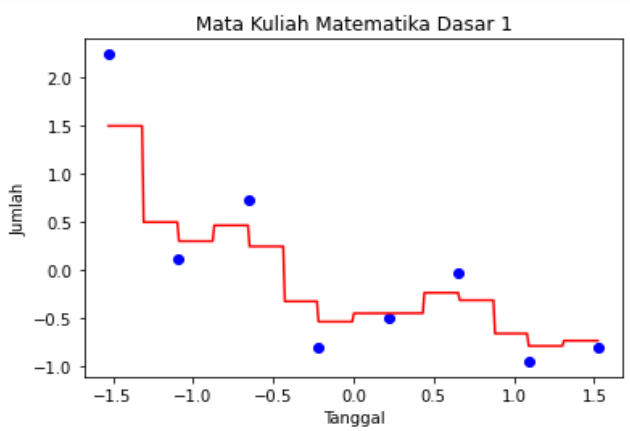

Figure 4. Visualization of Prediction Result

In Figure 3, it can be seen that the stages of random forest regression were carried out regarding the Random Forest construction by Liaw et al. (2002). The initial process is carried out by selecting 4 data samples and dividing them into 1 and 3. The branch with the minor error is used as one of the prediction results, while the enormous error value reduces to a new leaf by dividing the existing sample. The number of branches in the RF does not depend on the amount of data used because the amount of random data taken as a sample is less than the overall data.

4 CONCLUSION

Based on the results of research that has been done, there are several conclusions, namely: Random Forest Regression can be used to find the predicted number of course participants. The data pattern affects the accuracy of calculating the predicted value. The amount of data in Random Forest Regression affects the number of existing branches when calculating the predictive value.Random Forest Regression can reduce the overfitting that tends to occur when using the Decision Tree algorithm.

ACKNOWLEDGEMENTS

This research was funded through DIPA research activities at Teknik Faculty Universitas Tanjungpura for Fiscal Year 2022. Acknowledgments we conveyed this to UPT TIK Tanjungpura University, who assisted in collecting academic data used in this study.

REFERENCES

Alpaydin, E. (2014). Kernel Machines.

Astuti, A. S. T. U. T. I. (2019). Manajemen Kelas yang Efektif. Adaara: Jurnal Manajemen Pendidikan Islam, 9(2), 892-907.

Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

Chai, R., & Gunawan, D. (2022). Optimizing CNN Hyperparameters for Blastocyst Quality Assessment in Small Datasets. IEEE Access, 10, 88621-88631.

Herdianto. (2013). Prediksi Kerusakan Motor Induksi Menggunakan Metode. Jaringan Saraf Tiruan Backpropagation, Tesis, Universitas Sumatera Utara.

Indahyanti, U., Azizah, N. L., & Setiawan, H. (2022). Educational Data Mining on Student Academic Performance Prediction: A Survey. Procedia of Social Sciences and Humanities, 3, 1442-1447.

Khaleghi, S., Firouz, Y., Van Mierlo, J., & Van Den Bossche, P. (2019). Developing a real-time data-driven battery health diagnosis method, using time and frequency domain condition indicators. Applied Energy, 255, 113813.

Lahouar, A., & Slama, J. B. H. (2017). Hour-ahead wind power forecast based on random forests. Renewable energy, 109, 529-541.

Li, Y., Zou, C., Berecibar, M., Nanini-Maury, E., Chan, J. C. W., Van den Bossche, P., ... & Omar, N. (2018). Random forest regression for online capacity estimation of lithium-ion batteries. Applied energy, 232, 197-210.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Na, X. D., Zang, S. Y., Wu, C. S., & Li, W. L. (2015). Mapping forested wetlands in the Great Zhan River Basin through integrating optical, radar, and topographical data classification techniques. Environmental monitoring and assessment, 187(11), 1-17.

Rianto, H & Wahono, R. S.. (2015). Resampling Logistic Regression untuk Penanganan Ketidakseimbangan Class pada Prediksi Cacat Software. Journal of Software Engineering, 1(1).

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS journal of photogrammetry and remote sensing, 67, 93-104.

Saputra, T. A. N., Arizona, K. I., Andrian, M. R., Kurniadi, F. I., & Juarto, B. (2022, August). Random Forest in Detecting Hepatitis C. In 2022 9th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE) (pp. 299-302). IEEE.

Shu, X., Chen, Z., Shen, J., Guo, F., Zhang, Y., & Liu, Y. (2022). State of Charge Estimation for Lithium-ion Battery Based on Hybrid Compensation Modeling and Adaptive H-Infinity Filter. IEEE Transactions on Transportation Electrification.