Advances in Social Science, Education and Humanities Research, volume 545
Proceedings of the International Conference on Educational Assessment and Policy (ICEAP 2020)

Building a Classification Model to Predict School Quality in Indonesia

Satriyo Wibowo*
Directorate General of Early Childhood Education, Primary Education, and Secondary Education
Ministry of Education and Culture
Jakarta, Indonesia
*satriyo.wibowo@kemdikbud.go.id

***Abstract*—The aim of the present study was to explore the best classification model based on the predictive performance of decision tree (DT), random forest (RF), and boosted tree (BT) models used for school quality status detection in Indonesia. In order to get a better insight into the predictive abilities of these three models, 213,536 records from Basic Education Data (Dapodik), an education management information system managed by Ministry of Education and Culture, were utilized. A total of 23 predictors were extracted from Dapodik. School quality status (two classes: 'met national education standards' and 'improvement required') was an output (dependent) variable. Accuracy of the DT, RF, and BT models evaluated on the training dataset and the testing dataset was 71.51%, 99.91%, and 72.66% as well as 71.53%, 73.56%, and 72.82%, respectively. The findings indicate that the BT model had better performance than the DT dan RF models.**

Keywords—school quality, decision tree, random forest, boosted tree, classification

I. INTRODUCTION

Ministry of Education and Culture has developed an education management information system called Basic Education Data (*Dapodik*) which collects data from schools on school facilities, students, teachers and education personnel, and educational substances. *Dapodik* as the only data collecting system is continuously updated each semester and has been widely used in assessing the quality of schools.

Many researchers have studied the effect of school quality using measures such as school leadership and management, student and teacher characteristics, school facilities, and community involvement on student outcomes [1–3]. The study of school quality and factors associated with are growing fields of research in the education sciences. Furthermore, extensive progress has been made in the development of algorithms for modeling and identifying information from big databases. This study proposes data mining techniques to identify the key factors that classify and differentiate high- and low-quality schools.

Classification modeling is one of the analytical tools that can be utilized to distinguish the most influential factors. Classification is an operation that places each object from the population in one of a number of specified classes, according to the characteristics of the object which are identified as independent variables. An object is generally assigned to a class by using a formula, an algorithm, or a set of rules, which forms a model [4]. Classification modeling can be used to build predictive models and identify factors that affect school quality. Previous studies use classification modeling such as decision tree (DT), random forest (RF), and boosted tree (BT) to identify main factors of school accreditation rank and quality of education [5,6].

The DT model is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification researches. The DT model uses the C4.5 algorithm to form a decision tree. The trees are generated when the response variable is categorical data [7] This C4.5 algorithm is a development of the earlier ID3 algorithm and has the advantage of being able to process numeric and discrete data and can handle missing attribute values. In the decision tree, each leaf node is assigned as a class label. The root node and internal node contain attribute test conditions to separate records that have different characteristics [8].

The RF model is a type of ensemble machine learning algorithm called bootstrap aggregating (bagging) method and random feature selection [9]. Unlike the classification and regression tree (CART) method, the RF model is a compound tree method. In the RF model, many classification trees are built to form a forest, then the analysis is carried out based on voting from the tree clusters of the classification.

Meanwhile, the BT model is an iterative algorithm that also uses a combination of several classification trees that give different weights to the distribution of training data for each iteration. Each boosting iteration adds weight to the examples of misclassification and reduces the weight to the correct classification example so that it can effectively change the distribution of training data [10]. The adaptive boosting method (AdaBoost) can be a more effective solution to class imbalance problems [11]. In this study, 3 methods of classification modeling (DT, RF, and BT models) were

analyzed to find the most accurate model to predict school quality status.

II. METHODS

A. Data

School quality status and Basic Education Data (*Dapodik*) datasets were used in this study. The school quality status dataset was extracted from the 2019 school quality index generated by the Center of Data and Statistics for Education and Culture, Ministry of Education and Culture in building a zoning-based education quality index [12]. The school quality status can be categorized into two classes namely 'met national education standards' (class '1') and 'improvement required' (class '0').

TABLE I. LIST OF PREDICTORS

Variable Name	Description
X1	School has headmaster
X2	School has headmaster with bachelor degree
X3	School has certified headmaster
X4	School has trained headmaster
X5	School-based management applied
X6	School operational assistance fund program (BOS) allocated
X7	School operating time
X8	School has ISO certification
X9	School has accreditation rank
X10	Student teacher ratio
X11	Student class groups ratio
X12	Student classroom ratio
X13	Percentage of teachers with bachelor degree
X14	Percentage of certified teachers
X15	Number of proposed students to get Smart Indonesia Program (PIP)
X16	Number of dropout students
X17	School ownership status
X18	Source of electricity
X19	School has internet connection
X20	Number of good condition classrooms
X21	School has library
X22	Number of computers
X23	School has parent teacher committee

A total of 23 variables were generated and used as predictors to classify school quality status. The list of those predictors is shown in Table 1. An initial dataset contained 215,866 schools extracted from *Dapodik* but was subsequently reduced after editing for missing, incorrect values and outliers. The final dataset had 213,536 schools and will be used to build the model. As seen in Table 2, the data consists of two classes of school quality status namely 'met national education standards' and 'improvement required'.

TABLE II. DISTRIBUTION OF SCHOOLS BY QUALITY STATUS

Status	Frequency	Percent
Met national education standards	65,104	30.5%
Improvement required	148,432	69.5%
Total	213,536	100.0%

B. Method

In this paper, the analysis was carried out using R software by comparing the performance of DT, RF, and BT models as follows:

- Splitting the data into a training dataset and testing dataset (70% vs 30%)
- Create DT, RF, and BT models using the training dataset
- Evaluate the predictive performance of the DT, RF, and BT models. The difference between predicted and expected binary school quality status was described by the number of True Positives (*TP*), True Negatives (*TN*), False Positives (*FP*), False Negatives (*FN*), where the sum of $TP+TN+FP+FN = n$ is the total number of schools. To assess the model predictive accuracy of DT, RF and BT model for school quality status, the following performance metrics were used: correct classification rate ($C=(TP+TN)/n$), sensitivity ($S_n=TP/(TP+FN)$), specificity ($S_p=TN/(TN+FP)$), positive predictive value ($PPV=TP/(TP+FP)$), negative predictive value ($NPV=TN/(TN+FN)$). Moreover, to illustrate the relationship between sensitivity and specificity for the two-class category, the receiver operating characteristic (ROC) curve was plotted and the area under the ROC curves (*AUC*) was calculated [13].

III. RESULTS AND DISCUSSION

The confusion matrices from the DT model is shown in Fig. 1. The DT model was able to correctly predict school quality status in C=71.51% for the training dataset and C=71.53% for the testing dataset (Fig. 1). The results indicate that the DT model provided a fairly stable prediction and no overfitting occurred. Fig. 2 shows the classifier ROC curves with the AUC = 0.6264 on training samples and AUC = 0.6245 for the DT prediction on test samples. When evaluated, the DT model has a specificity of 88.34% for the training dataset and 88.01% for testing dataset which means it is good at predicting 'improvement required' schools.

A RF of 100 trees achieved C = 99.91% during the training phase and correctly classified C = 73.56% of the testing dataset. The results indicate that the RF model provided unstable prediction and overfitting occurred. The confusion matrices from the RF model is shown in Fig. 3. Fig. 4 shows the ROC curve with the AUC = 0.9996 for RF prediction on the training dataset and AUC = 0.7508 on the testing dataset. When evaluated, the RF model has a specificity of 99.98% for the training dataset and 89.89% for the testing dataset. This means that the RF model is good at predicting poor schools.

The BT model provided an accuracy rate of C = 72.66% for the training dataset and C = 72.82% for the testing dataset. The results indicate that the BT model provided a fairly stable prediction and no overfitting occurred. The confusion matrices of the BT model are shown in Fig. 5 while Fig. 6 shows the

ROC curve with the AUC = 0.7423 on training samples and AUC = 0.7390 on test samples. When further evaluated, the BT model has a specificity of 92.93% for the training dataset and 92.74% for the testing dataset which higher than those DT and RF models. This BT model is the best for predicting non-good performing schools.

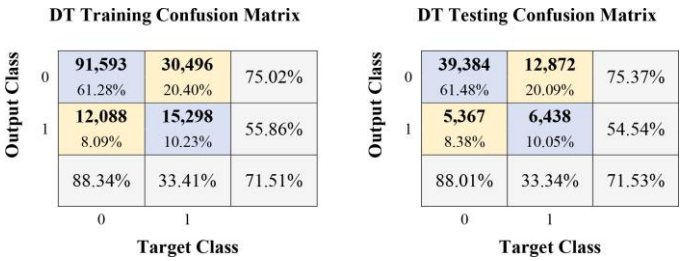

Fig. 1. Confusion matrices for the training dataset (left) and the test samples (right) for DT model.

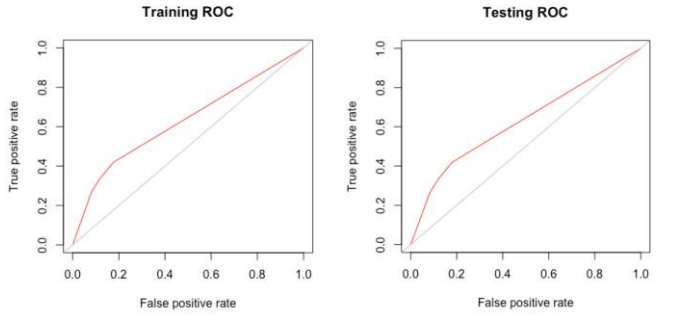

Fig. 2. ROC curves for DT classification accuracy on the training (left) and test (right) samples matrices for the training dataset (left) and the test samples (right) for DT model.

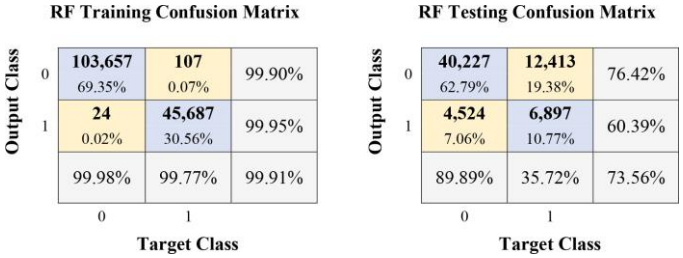

Fig. 3. Confusion matrices for the training dataset (left) and the test samples (right) for RF model.

Based on the findings of this study, the RF model gave an overfitting result. The RF model was poorly performed in making predictions, while the DT and BT models provided fairly stable predictive results. The BT model gave a slightly higher accuracy than the DT model. In addition, the BT model also has the advantage of predicting class imbalance data. This result was quite different from the study conducted by Sadiq and Ahmed [14] that concluded Decision Tree model shows better results in predicting student performance in the final exam. Meanwhile, Cinaroglu [15] stated that Random Forest provides better AUC score than the Decision Tree in predicting health expenditures.

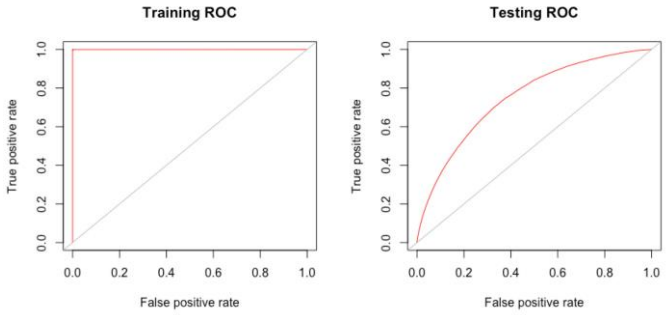

Fig. 4. ROC curves for RF classification accuracy on the training (left) and test (right) samples matrices for the training dataset (left) and the test samples (right) for RF model.

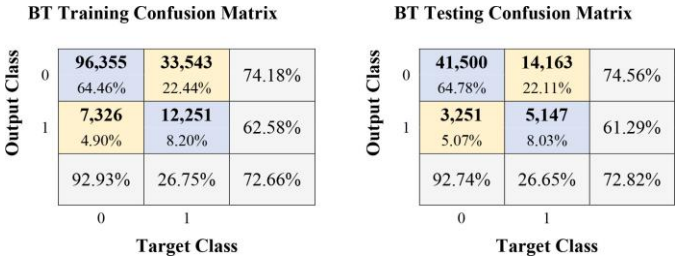

Fig. 5. Confusion matrices for the training dataset (left) and the test samples (right) for BT model.

TABLE III. PREDICTIVE PERFORMANCE OF THE THREE MACHINE LEARNING MODELS

Performance	DT		RF		Boosting	
	Training	Testing	Training	Testing	Training	Testing
Correct classification rate, C (%)	71.51%	71.53%	99.91%	73.56%	72.66%	72.82%
Sensitivity, Sn (%)	33.41%	33.34%	99.77%	35.72%	26.75%	26.66%
Specificity, Sp (%)	88.34%	88.01%	99.98%	89.89%	92.93%	92.74%
Positive Predictive Value, PPV (%)	55.86%	54.54%	99.95%	60.39%	62.58%	61.29%
Negative Predictive Value, NPV (%)	75.02%	75.37%	99.90%	76.42%	74.18%	74.56%
Area under the ROC curve, AUC	0.6264	0.6245	0.9996	0.7508	0.7423	0.7390

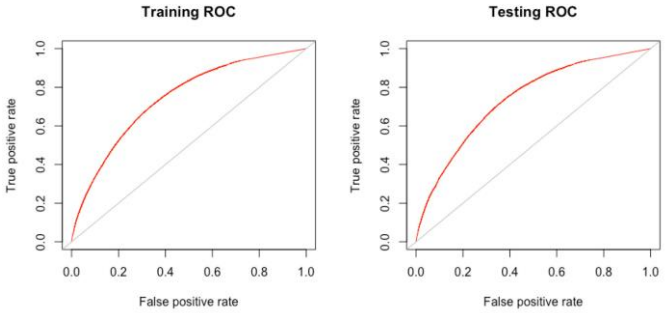

Fig. 6. ROC curves for BT classification accuracy on the training (left) and test (right) samples matrices for the training dataset (left) and the test samples (right) for BT model.

The findings about BT model performance in this study was supported from the research conducted by Esseiva et al. [16] that examined the classification modeling on feet fidgeting detection based on accelerometers and concluded that the Boosting model provides better accuracy than the Decision Tree and Random Forest. Miao and Heaton [17] study also showed that Boosting provides higher classification accuracy than Random Forest in predicting ecosystem classification data in the east Mojave Desert (see table 3).

IV. Conclusion

This paper presents preliminary research and describes the importance of tree-based machine learning algorithm and their application in predicting school quality status. The results show that the BT model emerged best compared to DT and RF models. The proposed tree-based algorithms have limitations. They require correct parameter tuning, the BT model may be overfitting if imprecise parameters are used. The existing dataset didn't have most of the school features such as learning materials, laboratories, and academic performance. In the future, we consider expanding the scope of *Dapodik* utilization to extract insights that may be valuable to school stakeholders.

Acknowledgment

This work was supported by the Secretariat of Directorate General of Early Childhood Education, Primary Education, and Secondary Education.

References

[1] M. Shodiq, Suyata, and S. Wibawa, “Developing quality evaluation instrument for islamic senior high school,” Jurnal Penelitian dan Evaluasi Pendidikan, vol. 21, no. 2, pp. 189–205, December 2017.

[2] S. Wu, C. Lin, S. Wu, C. Chuang, and H. Kuan, "Factors affecting quality of service in schools in Hualien, Taiwan," in 5th World Conference on Educational Sciences (WCES), 2013, pp. 1160–1164.

[3] D.D. Goldhaber and D.J. Brewer, “Does teacher certification matter? High school teacher certification status and student achievement,” Educ. Eval. Policy Anal, vol. 22, no. 2, pp. 129–145, 2000.

[4] S. Tufféry, Data Mining and Statistics for Decision Making. West Sussex: John Wiley & Sons Ltd, 2011.

[5] Y. Nindahayati, “Building a model to predict school accreditation rank using boosted classification tree,” Master’s thesis (unpublished), IPB University, 2015.

[6] A. Ramadhan, B. Susetyo, and Indahwati, “Penerapan metode klasifikasi random forest dalam mengidentifikasi faktor penting penilaian mutu pendidikan,” Jurnal Pendidikan dan Kebudayaan, vol. 4, no. 2, pp. 169–182, December 2019.

[7] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Olsen, Classification and Regression Trees. Boca Raton: Chapman & Hall, 1984.

[8] P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Essex: Pearson Education Limited, 2014.

[9] L. Breiman, “Random forests,” Machine Learning, Vol. 45, pp. 5–32, October 2001.

[10] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, “Handling imbalanced datasets: a review,” GESTS International Transactions on Computer Science and Engineering, vol. 30, pp. 25–36, 2006.

[11] S.B. Kotsiantis and P.E. Pintelas, “Selective costing ensemble for handling imbalanced data sets,” International Journal of Hybrid Intelligent System, vol. 6, pp. 123–133, 2009.

[12] Ministry of Education and Culture, Indeks Mutu Pendidikan Berbasis Zonasi. Jakarta: PDSPK, 2019.

[13] S. Rogers and M. Girolami, A First Course in Machine Learning, 2nd edition. Boca Raton: CRC Press, 2017.

[14] M.H. Sadiq and N.S. Ahmed, "Classifying and predicting students’ performance using improved decision tree C4.5 in higher education institutes," Journal of Computer Science, vol. 15, no. 9, pp. 1291–1306, October 2019.

[15] S. Cinaroglu, "Comparison of performance of decision tree algorithms and random forest: an application on OECD countries health expenditures," International Journal of Computer Applications, vol. 138, no. 1, pp. 37–41, March 2016.

[16] J. Esseiva, M. Caon, E. Mugellini, O.A. Khaled, and K. Aminian, "Feet Fidgeting Detection Based on Accelerometers Using Decision Tree Learning and Gradient Boosting," In: I. Rojas and F. Ortuño (Eds.) IWBBIO 2018. Lecture Notes in Computer Science, vol. 10814, pp. 75–84, 2018.

[17] X. Miao and J.S. Heaton, "A comparison of random forest and Adaboost tree in ecosystem classification in east Mojave Desert," 2010 18th International Conference on Geoinformatics, pp. 1–6, 2010.