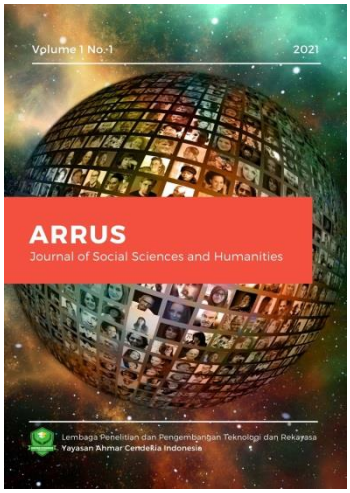
https://doi.org/10.35877/soshum2449

OPEN ACCESS

ISSN: 2776-7930 (Print) / 2807-3010 (Online)



RESEARCH ARTICLE

Evaluating Random Forest Algorithm in Educational Data Mining: Optimizing Graduation on-time prediction using Imbalance Methods

Rizal Bakri[1,2,*], Niken Probondani Astuti [3], & Ansari Saleh Ahmar[4]

[1]Statistics Research Group, STIEM Bongaya Makassar, Indonesia
[2]Department of Digital Business, Universitas Negeri Makassar, Indonesia
[3]Department of Management, STIEM Bongaya Makassar, Indonesia
[4]Department of Statistics, Universitas Negeri Makassar, Indonesia

*Corresponding author: Rizal Bakri, Statistics Research Group of STIEM Bongaya Makassar, Indonesia

E-mail: rizal.bakri@stiem-bongaya.ac.id

Abstract: The study aims to evaluate the performance of Random Forest algorithms in data mining education by optimizing graduation on-time (GOT) predictions using imbalanced data methods. Methods used to handle imbalanced data include random under-sampling (RUS), random over-sampling (ROS), hybrids of RUS and ROS, synthetic minority over-sampling techniques for nominal classes (SMOTE-NC), and hybrids of SMOTE-NC and RUS. After applying these methods, studies analyze their performance on training and testing data. The research findings show that on training data, the RUS-ROS hybrid showed the best performance compared to other methods, while the SMOTENC and RUS hybrid techniques showed the best performance on testing data based on AUC values. The research showed that the use of an imbalanced data method significantly improved the ability of Random Forest algorithms to predict graduation on time (GOT) in the context of educational data. We discuss the implications for educational data mining applications and provide suggestions for future research.

Keywords: educational data mining, random forest algorithm, imbalance data methods

1. Introduction

Educational data mining (EDM) is a rapidly expanding interdisciplinary field that combines techniques from education, computer science, and statistics to extract valuable insights from educational data. By leveraging data mining, machine learning, and statistical analysis, EDM seeks to improve educational practices, improve learning outcomes, and provide information for decision-making processes in educational institutions (Rabelo et al., 2023). Prediction plays an important role in the EDM field, offering the ability to predict a variety of educational outcomes and phenomena. Among these predictive tasks, the prediction of Graduation on time (GOT) becomes an important undertaking, offering valuable insights into student success and academic achievement. Accurate predictive models allow educators and administrators to identify students at risk (Sorensen, 2018), manage resources effectively (Arcinas et al.), and implement timely interventions to support student progress and retention (Dawson et al., 2017).

The Random Forest algorithm, which is an ensemble learning technique, has gained significant interest in predictive modeling tasks in the field of educational data mining (EDM) because of its strong performance, capacity to handle large datasets, and capability to handle

https://doi.org/10.35877/soshum2449

ISSN: 2776-7930 (Print) / 2807-3010 (Online)

intricate data structures. The inherent capacity of this method to handle non-linear relationships, assess the significance of features, and handle data with a high number of dimensions makes it a match for forecasting educational outcomes (Bakri et al., 2022). Despite this, Random Forest has trouble with unbalanced datasets that have unequal class distribution, which causes model performance to be biased (Yao et al., 2013). The disparities in educational data, such as the uneven allocation of students across different academic outcomes, pose a substantial challenge for Random Forest. This might potentially lead to suboptimal prediction performance. The presence of class imbalance, if not addressed, can compromise the accuracy of the prediction model and hinder its ability to facilitate results-oriented decision-making (Barros et al., 2019).

Because of Random Forest's limitations in handling unbalanced data, researchers are increasingly using imbalanced learning strategies to overcome this difficulty. To interpret and extract information from data with very blurry distributions, they created unbalanced learning (Chau & Phung, 2013; Ma & He, 2013; Utari et al., 2020). Random undersampling (RUS), random oversampling (ROS), the synthetic minority oversampling technique for nominal and continuous (SMOTENC), and hybrid approaches have all been looked into as ways to improve Random Forest's performance in educational prediction tasks (Ghorbani & Ghousi, 2020; Wongvorachan et al., 2023).

Although there has been significant growth in research on Random Forest and learning methods for imbalances in the EDM domain, there is still a shortfall in a comprehensive analysis that compares the effectiveness of these techniques. Furthermore, some studies tend to compare strategies using simulated data rather than actual data or use imbalance learning techniques as a means of building predictive models without explaining the underlying mechanisms or challenges during the modeling process. Although some research has been done on the application of imbalanced learning techniques, very few studies have specifically focused on imbalanced learning in the context of education. To address this gap, our study aims to perform a detailed comparison of imbalanced learning techniques to optimize Random Forest's performance in predicting timely graduation based on our previous research results that did not apply imbalanced learning techniques (Bakri et al., 2022). We also tested the models, showing how well they worked with training data and testing data as predictions. This helped us learn more about how to use predictive modeling effectively in an educational setting. These illustrative examples serve as didactic guidelines to help EDM researchers alleviate the problem of class imbalances, while also highlighting the right techniques for dealing with data with different imbalance ratios.

2. Research Method and Materials

2.1. Dataset and Data Preprocessing

This study is based on previously collected research data sets focusing on timely graduation at STIEM Bongaya (Bakri et al., 2022). The data set consists of 15 variables and includes 4093 data entries. Of these variables, six are continuous and nine are categorical, including target variables as described in Appendix 1. Out of the total number of data entries, 3071 students graduated on time (majority class), while 1022 others did not (minority class). Then we divided the data into two sub-sets, which are training data with an 80% ratio and testing data with a 20% ratio of both the original datasets as baseline and the data sets that were conditioned with imbalance methods. After building the model with training data, we evaluated it using both testing data and training data. We utilized the R programming language (R Core Team, 2020) with the readr (Wickham et al., 2018) and ggplot2 (Wickham, 2016) packages for data preprocessing.

2.2. Imbalance Data Methods

We evaluate the Random Forest method's performance by simulating various conditions in the data imbalance resistance method. The first method used is Random undersampling

https://doi.org/10.35877/soshum2449

ISSN: 2776-7930 (Print) / 2807-3010 (Online)

(RUS), which reduces the size of the majority class by random sampling to match the dimension of the minority class (Punlumjeak et al., 2017). Second, the Random Oversamplings (ROS) method involves adding the dimensions of minority classes by random sampling until they match the size and size of majority classes (Rachburee & Punlumjeak, 2021). Third, the hybrid method of RUS and ROS is to use random sampling to make the minority groups bigger and the majority groups smaller until both groups are equal (Hassan et al., 2020). Fourthly, the SMOTE-NC method is to increase the size of the minority class by generating new data based on the k-nearest neighborhood method so that the sample size matches the majority class (Mukherjee & Khushi, 2021). Finally, make a hybrid between the SMOTE-NC and RUS methods, i.e., increase the size of the minority class using the SMOTE-NC method and then reduce the magnitude of the majority class with the RUS technique (Wongvorachan et al., 2023). For the SMOTE-NC configuration, this study synthesized the minority data points based on their five nearest neighbors and 0.8 resampling ratios. We utilized the ROSE (Lunardon et al., 2014) and themis (Hvitfeldt, 2023) packages in R to implement resampling techniques

2.3. Machine Learning Algorithm and Matrics

In this study, the machine learning algorithm that we have applied is based on the best results of machine learning methods in our previous research, namely, Random Forest algorithms with tuning parameters (Bakri et al., 2022). We adopted this by applying the same procedure as in the previous study, by determining mtry values 2, 3, 4, and through 10-fold cross-validation. We utilized the randomForest package in R to implement the Random Forest algorithm (Liaw & Wiener, 2022). We use different metrics, like precision, recall, F1, area under the receiver operating characteristic curve (ROC-AUC), and accuracy values, to compare how well the Random Forest algorithm works with different methods for dealing with uneven data. We selected the best imbalance method based on the highest ROC-AUC value on both training data and testing data. We utilized the pROC (Robin et al., 2011) and caret (Kuhn, 2020) packages in R to measure performance.

3. Results

3.1. Resampling Results

Before evaluating the performance of Random Forest algorithms, we implemented data imbalance management techniques using various resampling methods. The resampling results of various imbalance methods are presented in Table 1. To compare the ratio between the majority class and the minority, we also used a non-resampling approach as a baseline to compare the ratio between the majority class and the minority. In the whole data set, the number "1" indicates a student who graduated on time (majority class), while the number "0" indicates students who did not graduate on time (minority class).

Table 1. Class size of full dataset, Training, and testing for each imbalance method

Dataset	Baseline	Undersampling (RUS)	Oversampling (ROS)	Hybrid 1 (RUS+ROS)	SMOTE-NC	Hybrid 2 (SMOTE-NC+RUS)
Full Dataset	1:3071 0:1022	1:1022 0:1022	1:3071 0:3071	1:2115 0:1978	1:3071 0:2456	1:2456 0:2456
Training (80%)	1:2457 0:818	1:818 0:818	1:2457 0:2457	1:1692 0:1583	1:2457 0:1965	1:1965 0:1965
Testing (20%)	1:614 0:204	1:212 0:212	1:614 0:614	1:423 0:395	1:614 0:491	1:491 0:491

In Random undersampling (RUS), the size of the majority class was decreased by randomly resampling to align with the minority class, resulting in 1022 rows of data. In random oversampling (ROS), the minority size increased to 3071 with randomly re-sampling to match the majority class. As for the hybrid approach (RUS+ROS), the minority class has been

https://doi.org/10.35877/soshum2449

ISSN: 2776-7930 (Print) / 2807-3010 (Online)

increasing while the majority has been decreasing, with random resampling to prevent extreme data imbalances. Then, the SMOTE-NC method increases the size of the minority class by generating new data based on the k-nearest neighborhood method, so that the sample size matches the majority class. In the hybrid approach (SMOTENC+RUS), the minority class is increasing with the SMOTENC technique, and in parallel, the majority class is decreasing with random resampling.

3.2. Classification Results

We evaluate the performance of the Random Forest algorithm based on the prediction results using both the training data itself and the testing data. Table 2 presents the evaluation of the prediction results for each resampling condition. This study indicates that using imbalance methods significantly improves the Random Forest algorithm's performance, although each imbalance method yields different metric values. In the training dataset, the hybrid method of RUS and ROS shows the best metric results (ROC-AUC = 0.998, accuracy = 0.971, F1 = 0.972, recall = 0.974, and precision = 0.971) compared to the other five conditions. Meanwhile, the lowest metric values are found in the baseline dataset (ROC-AUC = 0.888, accuracy = 0.960, F1 = 0.915, recall = 0.858, and precision = 0.979), despite having the highest precision value among the others. Subsequently, the second-best ranking is held by the ROS method (ROC-AUC = 0.997, accuracy = 0.967, F1 = 0.966, recall = 0.961, and precision = 0.972). The third-best ranking is the hybrid method of SMOTENC and RUS (ROC-AUC = 0.996, accuracy = 0.963, F1 = 0.963, recall = 0.960, and precision = 0.966). Following that, the fourth-best ranking is SMOTENC (ROC-AUC = 0.996, Accuracy = 0.966, F1 = 0.961, Recall = 0.957, and Precision = 0.966). Lastly, the fifth-best ranking is RUS (ROC-AUC = 0.996, accuracy = 0.961, F1 = 0.961, recall = 0.969, and precision = 0.953).

Table 2. Evaluate prediction of RF using training and testing data for each resampling condition

Dataset	Method	Precision	Recall	F1-score	AUC	Accuracy
Training	RUS-ROS	0.971	0.974	0.972	0.998	0.971
	ROS	0.972	0.961	0.966	0.997	0.967
	SMOTENC-RUS	0.966	0.960	0.963	0.996	0.963
	SMOTENC	0.966	0.957	0.961	0.996	0.966
	RUS	0.953	0.969	0.961	0.996	0.961
	Baseline	0.979	0.858	0.915	0.888	0.960
Testing	SMOTENC-RUS	0.797	0.941	0.863	0.978	0.925
	ROS	0.919	0.904	0.911	0.971	0.912
	RUS-ROS	0.919	0.917	0.918	0.970	0.916
	SMOTENC	0.854	0.821	0.837	0.926	0.858
	Baseline	0.854	0.574	0.686	0.888	0.869
	RUS	0.740	0.824	0.780	0.847	0.767

In the testing dataset, the performance metrics of the imbalance methods exhibit a different ranking compared to the training data. The hybrid method of SMOTENC and RUS demonstrates the most superior metric results (ROC-AUC = 0.978, accuracy = 0.925, F1 = 0.863, recall = 0.941, and precision = 0.797) compared to the other five conditions. Meanwhile, the lowest metric values are observed in RUS (ROC-AUC = 0.847, accuracy = 0.767, F1 = 0.780, recall = 0.824, and precision = 0.740), which ranks below the baseline data (ROC-AUC = 0.888, accuracy = 0.869, F1 = 0.686, recall = 0.574, and precision = 0.854) as the fifth-best condition. Subsequently, the second-best ranking is held by the ROS method

https://doi.org/10.35877/soshum2449

ISSN: 2776-7930 (Print) / 2807-3010 (Online)

(ROC-AUC = 0.971, accuracy = 0.912, F1-score = 0.911, recall = 0.904, and precision = 0.919). The third-best ranking is the hybrid RUS and ROS (ROC-AUC = 0.970, accuracy = 0.916, F1 = 0.918, recall = 0.917, and precision = 0.919). Next, the fourth-best ranking is SMOTENC (ROC-AUC = 0.926, accuracy = 0.858, F1 = 0.837, recall = 0.821, and precision = 0.854).

The performance of the Random Forest algorithm can also be evaluated by examining the Receiver Operating Characteristic (ROC) curve, as depicted in Figure 1. It is evident that significantly, with the handling of data imbalances, the performance of the Random Forest algorithm improves in making predictions, both on training and testing data. The ROC curve analysis of the training data indicates that the Random Forest algorithm provides consistent prediction results under various resampling conditions. Meanwhile, evaluation of the testing data shows better variability than the baseline, indicating an improvement in the prediction quality of the Random Forest algorithm after handling the data imbalance.

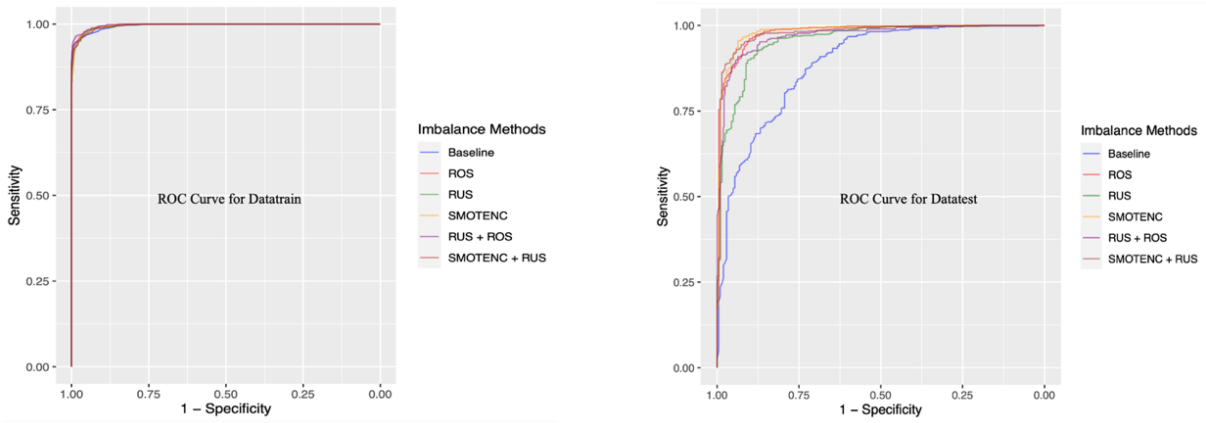


Figure 1. Receiver Operating Characteristic (ROC) curve for data training dan data testing

4. Discussion and Conclussions

This study presents a comprehensive evaluation of the performance of the Random Forest algorithm following the implementation of four data imbalance handling techniques (oversampling, undersampling, SMOTENC, and hybrid approach) within the context of educational data mining (EDM) classification. Our primary focus is to deliver an analysis of the Random Forest algorithm's performance in prediction, utilizing balanced training and testing data. The findings of this research aim to offer insights to researchers in the EDM field regarding the effectiveness of the applied resampling techniques while identifying their respective strengths and weaknesses. Additionally, the comparative results presented among the variations of imbalance handling techniques can provide practical guidance for researchers in selecting suitable strategies for handling their educational data

Our research shows that the performance of Random Forest algorithms varies when making predictions using both the training data itself and the testing data. It was found that the use of resampling techniques such as Random undersampling (RUS), Random Over Sampler (ROS), Synthetic Minority Over-sampling Techniques for nominal and continuous (SMOTENC), and hybrid methods could significantly improve model performance. The RUS-ROS method showed excellent performance in distinguishing positive and negative classes on the training data, with the highest AUC value. However, the performance declined slightly on test data, although it was still pretty good. The ROS method also provides consistent results on both training and test data, demonstrating its ability to improve model performance by striking a balance between precision and recall. On the other hand, the RUS method showed a significant decrease in both test data and training data, indicating the possibility of overfitting. These findings suggest that, when the target variable is mediumly

ISSN: 2776-7930 (Print) / 2807-3010 (Online)

imbalanced, ROS may be more advantageous for classification than RUS (Mohammed et al., 2020). Thus, RUS is not recommended as a first option as it leads to the loss of data that is potentially useful for classifiers (He & Ma, 2013). Furthermore, most of our variables are category-shaped, which poses a challenge to some learning techniques of imbalance because they only support sustainable variables such as SMOTE or ENN (Lemaˆıtre et al., 2017). We use SMOTE-NC as a new solution that can be applied to real-life situations because some variables related to education are either category or ordinal (Wibowo & Dewi Ratih, 2021). Therefore, we conducted experiments combining the method with RUS, and the results were excellently improved. The performance of the model was significantly improved on the testing data, which is the third sequence in the data training. In addition, the SMOTENC method is also implemented without hybridizing with other methods and provides consistent results on both training data and test data. Finally, the baseline method showed the lowest performance among all methods, suggesting that class imbalance handling was necessary for better Random Forest algorithm performance. This point is supported by class imbalance literature detrimental to the performance of the predictive model because it makes the algorithm learn more from the majority class than from the minority class, thus producing bias in the prediction model (He & Ma, 2013; Johnson & Khoshgoftaar, 2019). Therefore, reducing class imbalances can be beneficial to the classifiers in general, as it is the primary goal of the imbalances learning technique (Johnson & Khoshgoftaar, 2019).

This study has its limitations. We use a single dataset (the graduation on-time of STIEM Bongaya) in evaluating Random Forest algorithms after using the data imbalance method. However, each dataset has its own characteristics, such as variable patterns, imbalance ratios, and the number of minority classes. The results of the imbalance learning algorithm may vary depending on the uniqueness of each data set. Thus, researchers are encouraged to explore the characteristics of their datasets carefully before choosing imbalance methods to address the class imbalance problem when using Random Forest algorithms. Finally, future research is needed to explore different combinations of sampling techniques (e.g., application of ensemble techniques) and other classification algorithms because the effectiveness of each re-sampling technique may partly depend on the classifiers (Chakravarthy et al., 2019).

References

Arcinas, M. M., Sekhar Sajja, G., Asif, S., Gour, S., Okoronkwo, E., & Naved, M. (n.d.). ROLE OF DATA MINING IN EDUCATION FOR IMPROVING STUDENTS PERFORMANCE FOR SOCIAL CHANGE. Turkish Journal of Physiotherapy and Rehabilitation, 32(3). Retrieved March 7, 2024, from www.turkjphysiotherrehabil.org

Bakri, R., Astuti, N. P., & Ahmar, A. S. (2022). Machine Learning Algorithms with Parameter Tuning to Predict Students' Graduation-on-time: A Case Study in Higher Education. Journal of Applied Science, Engineering, Technology, and Education, 4(2), 259–265. https://doi.org/10.35877/454RI.ASCI1581

Barros, T. M., Neto, P. A. S., Silva, I., & Guedes, L. A. (2019). Predictive Models for Imbalanced Data: A School Dropout Perspective. Education Sciences 2019, Vol. 9, Page 275, 9(4), 275. https://doi.org/10.3390/EDUCSCI9040275

Chakravarthy, A. D., Bonthu, S., Chen, Z., & Zhu, Q. (2019). Predictive models with resampling: A comparative study of machine learning algorithms and their performances on handling imbalanced datasets. Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019, 1492–1495. https://doi.org/10.1109/ICMLA.2019.00245

Chau, V. T. N., & Phung, N. H. (2013). Imbalanced educational data classification: An effective approach with resampling and random forest. Proceedings - 2013 RIVF International Conference on Computing and Communication Technologies: Research, Innovation, and Vision for Future, RIVF 2013, 135–140. https://doi.org/10.1109/RIVF.2013.6719882

ISSN: 2776-7930 (Print) / 2807-3010 (Online)

Dawson, S., Jovanovic, J., Gašević, D., & Pardo, A. (2017). From prediction to impact: Evaluation of a learning analytics retention program. ACM International Conference Proceeding Series, 474–478. https://doi.org/10.1145/3027385.3027405

Ghorbani, R., & Ghousi, R. (2020). Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques. IEEE Access, 8, 67899–67911. https://doi.org/10.1109/ACCESS.2020.2986809

Hassan, H., Ahmad, N. B., & Anuar, S. (2020). Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining. Journal of Physics: Conference Series, 1529(5), 052041. https://doi.org/10.1088/1742-6596/1529/5/052041

He, H., & Ma, Y. (2013). Imbalanced Learning: Foundations, Algorithms, and Applications. John Wiley & Sons: Hoboken.

Hvitfeldt, E. (2023). themis: Extra Recipes Steps for Dealing with Unbalanced Data. https://cran.r-project.org/package=themis

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 1–54. https://doi.org/10.1186/S40537-019-0192-5/TABLES/18

Kuhn, M. (2020). caret: Classification and Regression Training. https://github.com/topepo/caret/

LemaˆıtreLemaˆıtre, G., Nogueira, F., & Aridas char, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research, 18(17), 1–5. http://jmlr.org/papers/v18/16-365.html

Liaw, A., & Wiener, M. (2022). Classification and Regression by randomForest. R News, 2(3), 18–22. https://cran.r-project.org/package=randomForest

Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: a Package for Binary Imbalanced Learning. R Journal, 6(1), 82–92.

Ma, Y., & He, H. (2013). Imbalanced Learning: Foundations, Algorithms, and Applications. In Imbalanced Learning: Foundations, Algorithms, and Applications. John Wiley & Sons Inc.

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. 2020 11th International Conference on Information and Communication Systems, ICICS 2020, 243–248. https://doi.org/10.1109/ICICS49469.2020.239556

Mukherjee, M., & Khushi, M. (2021). SMOTE-ENC: A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features. Applied System Innovation 2021, Vol. 4, Page 18, 4(1), 18. https://doi.org/10.3390/ASI4010018

Punlumjeak, W., Rugtanom, S., Jantarat, S., & Rachburee, N. (2017). Improving classification of imbalanced student dataset using ensemble method of voting, bagging, and adaboost with under-sampling technique. Lecture Notes in Electrical Engineering, 449, 27–34. https://doi.org/10.1007/978-981-10-6451-7_4/COVER

R Core Team. (2020). R: A Language and Environment for Statistical Computing. https://www.R-project.org/

Rabelo, A., Rodrigues, M. W., Nobre, C., Isotani, S., & Zárate, L. (2023). Educational data mining and learning analytics: a review of educational management in e-learning. Information Discovery and Delivery, ahead-of-print(ahead-of-print). https://doi.org/10.1108/IDD-10-2022-0099/FULL/XML

Rachburee, N., & Punlumjeak, W. (2021). Oversampling technique in student performance classification from engineering course. International Journal of Electrical and Computer Engineering (IJECE), 11(4), 3567–3574. https://doi.org/10.11591/IJECE.V11I4.PP3567-3574

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, 77. https://cran.r-project.org/package=pROC

ISSN: 2776-7930 (Print) / 2807-3010 (Online)

Sorensen, L. C. (2018). "Big Data" in Educational Administration: An Application for Predicting School Dropout Risk. 55(3), 404–446. https://doi.org/10.1177/0013161X18799439

Utari, M., Warsito, B., & Kusumaningrum, R. (2020). Implementation of Data Mining for Drop-Out Prediction using Random Forest Method. 2020 8th International Conference on Information and Communication Technology, ICoICT 2020. https://doi.org/10.1109/ICOICT49345.2020.9166276

Wibowo, W., & Dewi Ratih, I. (2021). Classification of Non-Performing Financing Using Logistic Regression and Synthetic Minority Over-sampling Technique-Nominal Continuous (SMOTE-NC). Int. J. Advance Soft Compu. Appl, 13(3). https://doi.org/10.15849/IJASCA.211128.09

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., Hester, J., & Francois, R. (2018). readr: Read Rectangular Text Data. https://CRAN.R-project.org/package=readr

Wongvorachan, T., He, S., & Bulut, O. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. Information 2023, Vol. 14, Page 54, 14(1), 54. https://doi.org/10.3390/INFO14010054

Yao, D., Yang, J., & Zhan, X. (2013). An improved random forest algorithm for class-imbalanced data classification and its application in PAD risk factors analysis. Open Electrical and Electronic Engineering Journal, 7(SPEC ISS 1), 62–70. https://doi.org/10.2174/1874129001307010062

https://doi.org/10.35877/soshum2449

ISSN: 2776-7930 (Print) / 2807-3010 (Online)

Appendix 1. Attributes of Graduation on time (GOT)

Atribute ID	Value	Description
NCP	0 - 4	Student's Number Credit Passed
SMT4	0 - 4	Student's GPA Semester 4
SMT3	0 - 4	Student's GPA Semester 3
SMT2	0 - 4	Student's GPA Semester 2
SMT1	0 - 4	Student's GPA Semester 1
AA	16 - 46	Student's Age Admission
FS	Accounting, Financial, Marketing, Human Resource	Student's Focus Study
FI	IDR 1 - IDR 499.999	Student's Father Income
	IDR 500.000 - IDR 999.999	
	IDR 1.000.000 - IDR 1.999.999	
	IDR 2.000.000 - IDR 4.999.999	
	IDR 5.000.000 - IDR 20.000.000	
	More than IDR 20.000.000	
	Nil Income	
MI	IDR 1 - IDR 500.000	Student's Mother Income
	IDR 500.000 - IDR 999.999	
	IDR 1.000.000 - IDR 1.999.999	
	IDR 2.000.000 - IDR 4.999.999	
	IDR 5.000.000 - IDR 20.000.000	
	More than IDR 20.000.000	
	Nil Income	
SEX	Male, Female	Student's Sex
RE	with Parents, with Guardian, Boarding House, Dormitory, Others	Student's Residence
TR	Public transportation, Private Car, Private Motorcycle Walk to campus	Student's Transportation
DEP	Management, Accounting	Department taken by student
CT	Regular Class, Executive Class	Class type taken by student
GOT Status	1, 0	Student's graduation status (1 is GOT and 0 is not GOT)