Contents lists available at www.journal.unipdu.ac.id

Register

Journal Page is available to www.journal.unipdu.ac.id/index.php/register

Research article

Measuring Resampling Methods on Imbalanced Educational Dataset's Classification Performance

Irfan Pratama [a*], *Putri Taqwa Prasetyaningrum* [b], *Albert Yakobus Chandra* [c], *Ozzi Suria* [d]

[a,b,c,d] *Faculty of Information Technology, Universitas Mercu Buana Yogyakarta,Yogyakarta, Indonesia*
email: [a*] *irfanp@mercubuana-yogya.ac.id*, [b] *putri@mercubuana-yogya.ac.id*, [c] *albert.ch@mercubuana-yogya.ac.id*, [d] *ozzisuria@mercubuana-yogya.ac.id*
* *Correspondence*

ARTICLE INFO

Article history:
Received 27 February 2023
Revised 17 June 2023
Accepted 25 January 2024
Available online 25 February 2024

Keywords:
Classification
Educational data mining
Imbalanced Data
Resampling

IEEE style in citing this article:
I. Pratama, P. T. Prasetyaningrum, A. Y. Chandra, O. Suria, "Measuring Resampling Methods on Educational Dataset's Classification Performance," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 10, no. 1, pp. 1-11, 2024.

ABSTRACT

Imbalanced data refers to a condition that there is a different size of samples between one class with another class(es). It made the term "majority" class that represents the class with more instances number on the dataset and "minority" classes that represent the class with fewer instances number on the dataset. Under the target of educational data mining which demands accurate measurement of the student's performance analysis, data mining requires an appropriate dataset to produce good accuracy. This study aims to measure the resampling method's performance through the classification process on the student's performance dataset, which is also a multi-class dataset. Thus, this study also measures how the method performs on a multi-class classification problem. Utilizing four public educational datasets, which consist of the result of an educational process, this study aims to get a better picture of which resampling methods are suitable for that kind of dataset. This research uses more than twenty resampling methods from the SMOTE variants library. as a comparison; this study implements nine classification methods to measure the performance of the resampled data with the non-resampled data. According to the results, SMOTE-ENN is generally the better resampling method since it produces a 0,97 F1 score under the Stacking classification method and the highest among others. However, the resampling method performs relatively low on the dataset with wider label variations. The future work of this study is to dig deeper into why the resampling method cannot handle the enormous class variation since the F1 score on the student dataset is lower than the other dataset.

1. Introduction

Imbalanced data classification is a common problem in data mining [1]. Imbalanced data refers to a condition that there is a different size of samples between one class with another class(es). The term "majority" class represents the class with more instances number on the dataset, and "minority" classes represent the class with fewer instances number on the dataset [2]. Imbalanced data can be encountered in various fields of applications such as financial services [3], [4], healthcare[5], [6], blockchain[7], [8], and educational data mining [9]–[14]. Any imbalanced problem will affect the classification's performance because the model will only be well-trained toward the majority class on the dataset as its data amount is more significant than the other class(es). The behavior will produce a biased result and favor the majority class over the minority class(es). The aim is to provide the classification model with a fair condition of a dataset with relatively the same amount of data for each class(es). Under the assumption that the same amount of data of each class(es) will give an unbiased result because the amount of the data supplied toward the model is the same.

Particularly, educational data mining is emerging as it can provide quality education for students to enhance academic performance according to their study records that are processed using machine learning methods [15]. Research on educational data mining is increasing due to the benefits obtained from the acquired knowledge of the machine learning processes. It improves the institution's decision-making toward learning outcomes and planning [16]. Under the target of educational data mining which demands accurate measurement of the student's performance analysis, data mining requires appropriate datasets to produce good accuracy [11]. However, getting a good classification accuracy with the imbalanced data for each class is impossible as it decreases the effectiveness of the classifier's performance [9], [11]. The stated potential problem then triggered several studies toward the imbalanced class handling on educational datamining fields. The studies mainly focused on how the data should be treated to make the classification produced fair result. There are two mechanisms to overcome the imbalanced class problem: the data-level approach and the algorithmic approach [9]. According to the reseach by [9], algorithmic level approach works less efficiently toward the high ratio of imbalanced data and that alone makes the data level approach more popular to handle imbalanced dataset. The data-level approach is mainly focusing on resampling data mechanisms. It is about how to make a balanced amount of data for each class by utilizing the deletion mechanism (undersampling) or data synthesis mechanism (oversampling). The current development of imbalanced class handling is proposing a hybrid resampling mechanism that combines any undersampling method and any oversampling method to negate each other drawbacks.

As the data from the educational fields cannot be ensured to be in good condition whether it is caused by imbalanced data or other dataset problem, this research want to answer several question such as: (1) How would the resampling mechanism affects the classification performance; (2) How would any resampling method affected with the imbalanced ratio of a dataset; (3) How would resampling methods affects the multi-classification problem.

2. Materials and Methods

Several studies conducted regarding imbalanced problem in educational dataset. Imbalanced distribution of a dataset is a crucial problem that affects the classification performance. Hence, there are various study about imbalanced class or imbalanced data handling.

The algorithmic level approach relies on specific classifiers to classify the imbalanced dataset with preferable classification output. Cost-sensitive learning, bagging, boosting, and stacking are well-known for algorithmic-level approaches [17]. The data-level approaches favor solving the imbalanced class problem more than the Algorithmic level approaches. The data level approach becomes more valuable as there are studies conducted utilizing such an approach to overcome imbalanced class problems [2], [10],[18]–[23].

The data level approach has two unique mechanisms: random under-sampling and random oversampling [24]. Random oversampling is randomly generating duplicates from the minority class, while random under-sampling is deleting random instances from the majority class. Even though the excessive use of oversampling would lead to overfitting, excessive use of the under-sampling would lead to information loss [25]. Hence, there are several hybrid resampling methods or enhancements to the previous oversampling methods to prevent the overfitting problem, such as SMOTE-Tomek [26], SMOTE-ENN [27], BorderlineSMOTE [28], Geometric SMOTE [29] and other SMOTE variants and derivatives [8].

This research consists of two primary processes, which are the data preparation process and the classification process. In the data preparation stage, several steps are needed, such as data acquisition, data preprocessing (resampling), and training-testing data split. In the classification stage, the dataset is divided into two scenarios which are the classification process on the train-test split dataset, and the classification process on the full dataset using cross-validation.

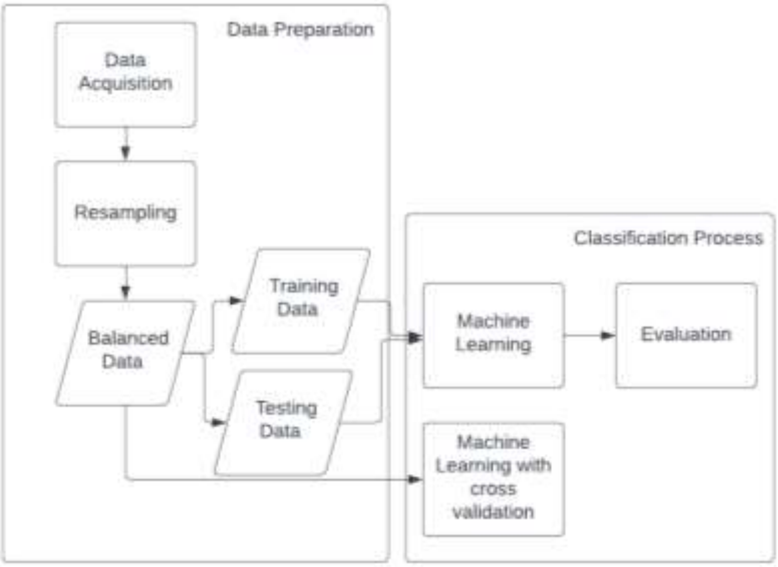


Fig. 1. Research Flowchart

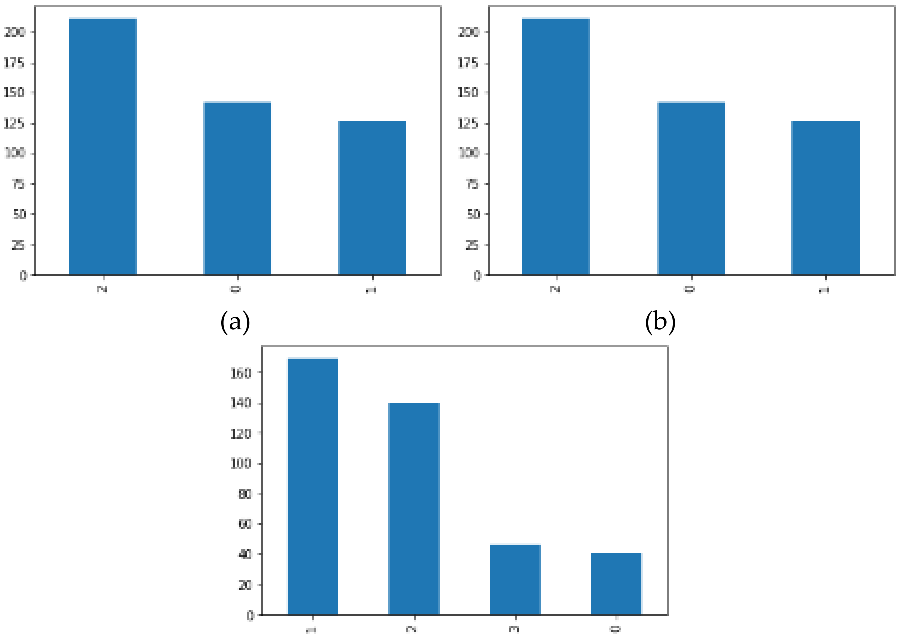
2.1. Data Acquisition

This research utilizes four public datasets retrieved from either Kaggle or UCI repository. Those datasets are students' performance data classified into several labels (class) and have different imbalance ratios. Table 1 Shows the dataset description used in this study.

Table 1. Dataset Description

	Dataset			
	xAPI-edu-data [30]	Student Performance Dataset Por [31]	Student Performance Dataset Math [31]	Student's Grade Dataset [32]
Number of Attribute	16	33	33	22
Class variation	3	4	4	7
Missing Values	No	No	No	No
Number of Instance	480	649	649	1203
Imbalance ratio	1 : 1.6	1 : 21.8	1 : 4.2	1 : 10.7

The imbalanced ratio is shown in Table 1 and is calculated by dividing the majority class instance number and (lowest) minority instance number since it is a multi-class dataset. The following figures show the class distribution on each dataset.



(a)

(b)

(c)

Fig. 2. Class Distribution: (a) X-API dataset; (b) Por dataset and (c) Math dataset

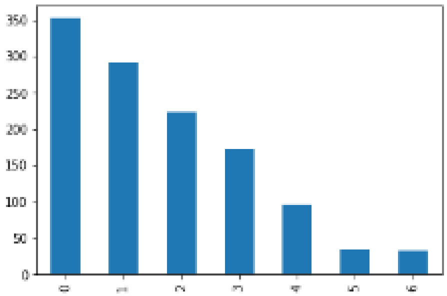


Fig 3. Student dataset

Fig 1 and Fig 2 shows the visual distribution of each dataset based on class(es). Can be seen that the data is suffered an imbalanced class. The chart description is the x-axis is the encoded class of the dataset and the y-axis is the number of instances for each class(es).

2.2. Resampling

The imbalanced condition is found at the very start of the data mining process, which is in the data acquisition step. When the dataset is known to be imbalanced, one of the preprocessing steps is handling the imbalanced problem using the resampling mechanism. Any classification procedure should be carried out to know the quality of the resampled data. Therefore, it needs a whole data mining process to implement and evaluate the resampling method on a dataset. This study applies several scenarios of the resampling–classification process to get a complete and more comprehensive measurement result. It will also help determine which resampling method is generally better according to the dataset's imbalance state and the classification method's result.

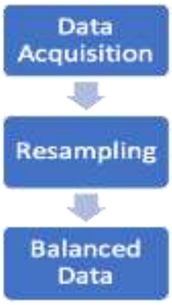


Fig 4. Resampling Stage

As shown in Fig. 4 the collected dataset which are already explained in the previous section that have imbalanced class problem will be resampled. The resampling mechanism used in this study are replicating the methods from previous study and state of the art research product which are either oversampling method and hybrid method. Based on the current existing hybrid model SMOTE-Tomek [26], this study do an experiment on a hybrid method that combining ADASYN oversampling method and Tomek-Links undersampling method. Under the knowledge that ADASYN is developed from SMOTE oversampling method and said to be better than SMOTE. Based on that assumption, the hypothesis are ADASYN-Tomek will likely performs better than SMOTE-Tomek because the ADASYN method is a better version of SMOTE. The mechanism of the experimental resampling method are as follows:

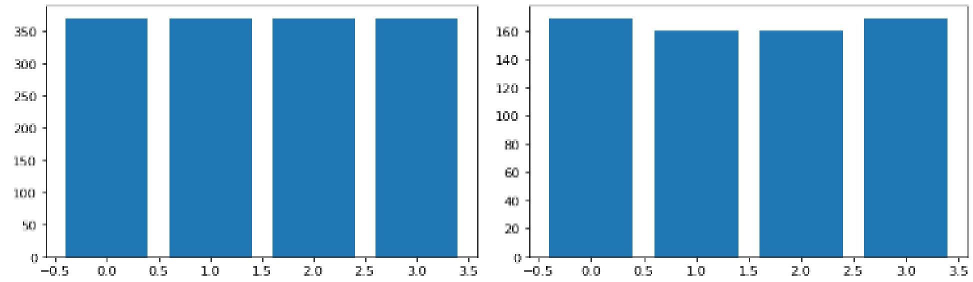
1. The imbalanced dataset will be oversampled using ADASYN oversampling method.
2. The oversampled dataset will be the input to the Tomek undersampling method.
3. The balanced dataset produced.

3. Results and Discussion

3.1. Resampling Result

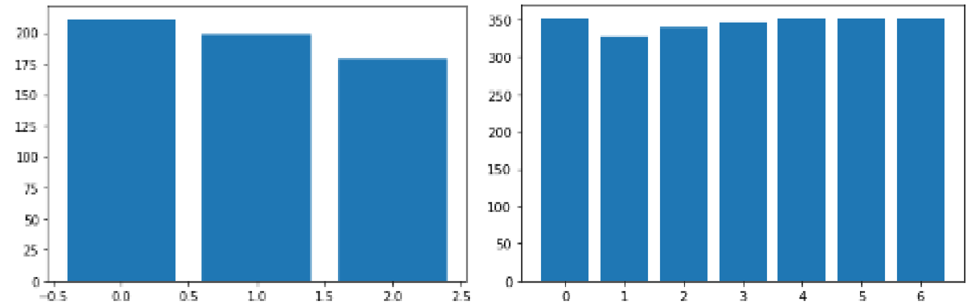
In this section, the result of each process will be provided, along with an explanation of the results. The first step of the research, which is resampling steps, the resampling visualization of each dataset using one of the resampling methods used in this study (ADASYN-Tomek), can be seen in Fig 6. to Fig 10.

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
10 (1) January 2024 1-11



(a) (b)

Fig 5. Resampling Result: (a) Por dataset and (b) Math dataset



(a) (b)

Fig 6. Resampling Result: (a) X-Api dataset and (b) Student dataset

The sample resampling method in the figure is ADASYN-Tomek. The dataset is oversampled using ADASYN in the first process, then resampled the second time using the TomekLinks under-sampling method. The results of the resampling method made the dataset not strictly equal in class instances number but still relatively balanced.

3.2. Classification Result

After each resampling process, the dataset is classified using the mentioned classification algorithm. To determine how good the classification with imbalanced dataset can be seen from the F1-score. F1-score can be interpreted as weighted average or harmonic mean of precision and recall. The best value of F1-score is 1 and the worst will be at 0. Table 2 shows the highest F1-score among all datasets' resampling and classification methods using both splitting and cross-validation evaluation metrics.

Table 2. Experiment Result

80:20 splitting	Dataset			
Classifiers	**xAPI**	**Por**	**Math**	**Student**
Logistic Regression	SMOTEENN (0,86)	DEAGO CURE-SMOTE Trim-SMOTE MWMOTE (0,93)	Trim-SMOTE (0,92)	Trim-SMOTE (0,79)
K-NN	ENN (0,95)	SMOTEENN (0,9)	SMOTEENN (0,9)	SMOTEENN (0,86)
CART	SMOTEENN (0,87)	SMOTEENN (0,96)	SMOTEENN (0,93)	SMOTEENN (0,84)
Random Forest	SMOTEENN (0,95)	MCT (0,97)	SMOTEENN Polynom-fit-SMOTE (0,96)	SMOTEENN (0,9)
SVM	SMOTEENN (0,88)	Borderline-SMOTE1 LLE-SMOTE (0,92)	Trim-SMOTE (0,9)	SMOTEENN (0,86)
STACKING	ENN (0,93)	SMOTEENN (0,97)	SMOTEENN (0,96)	SMOTEENN (0,91)

http://doi.org/10.26594/register.v10i1.3397

80:20 splitting	Dataset			
Classifiers	xAPI	Por	Math	Student
XGBoost	SMOTEENN (0,97)	SMOTEENN Borderline-SMOTE1 Cure-SMOTE SMOTE-D (0,95)	SMOTE-IPF (0,94)	ENN (0,93)
XGBRF	SMOTEENN (0,88)	ENN (0,95)	Polynom-fit-SMOTE (0,94)	Trim-SMOTE (0,78)
AdaBoost	Trim-SMOTE (0,88)	ANS (0,86)	LLE-SMOTE Cure-SMOTE (0,89)	Trim-SMOTE (0,68)

K-Fold	Dataset			
Classifiers	xAPI	Por	Math	Student
Logistic Regression	SMOTEENN (0,88)	SMOTENN Trim – SMOTE (0,92)	SMOTEENN (0,89)	Trim-SMOTE (0,77)
K-NN	SMOTE (0,97)	SMOTENN (0,94)	SMOTEENN (0,9)	SMOTEENN (0,91)
CART	SMOTEENN (0,89)	SMOTENN (0,94)	SMOTEENN (0,93)	SMOTEENN (0,94)
Random Forest	SMOTEENN (0,91)	SMOTENN (0,97)	SMOTEENN (0,95)	SMOTEENN (0,91)
SVM	Trim-SMOTE (0,90)	LLE-SMOTE (0,9)	SMOTE-IPF (0,85)	SMOTEENN (0,88)
STACKING	SMOTEENN (0,94)	SMOTEENN (0,98)	SMOTEENN (0,96)	SMOTEENN (0,93)
XGBoost	SMOTEENN (0,93)	SMOTEENN (0,96)	SMOTEENN (0,95)	SMOTEENN (0,83)
XGBRF	Trim-SMOTE (0,89)	Borderline-SMOTE1 Polynom-fit-SMOTE LLE-SMOTE DEAGO Cure-SMOTE (0,93)	Polynom-fit-SMOTE (0,93)	Trim-SMOTE (0,79)
AdaBoost	SMOTE-Tomek (0,97)	ENN (0,86)	Cure-SMOTE (0,9)	Trim-SMOTE (0,71)

Each resampling method preserved its default parameter as the experiment was carried out. The same goes for the classification methods. In the Table 2, each scenario on each dataset performs differently. In most scenarios, SMOTEENN performs better than any other resampling method with good accuracy results on the classifier performance proven on both evaluation metrics. The SMOTEENN resampling method is a good pair with any classification method to handle the imbalanced class problem. However, other resampling methods include ENN, Trim-SMOTE, Polynom-fit-SMOTE, LLE-SMOTE, DEAGO, Cure-SMOTE, Borderline-SMOTE1, MCT, MWMOTE, SMOTE, and SMOTE-Tomek can outperform SMOTEENN in some cases, among all of the results on Table 2, the highest F1 score produced by Stacking classification method and SMOTEENN resampling with 0,98 under k-fold evaluation metric, which slightly higher than the splitting evaluation metric with 0,97 on the same pair plus XGBoost and SMOTEENN pair with the same F1 score.

3.3. Influence of resampling method on classification's performance

This research utilizes many resampling methods to measure the resampled dataset during the classification process. On the first dataset (x-API), the average of all classification methods on every resampling technique produced a 0,78 F1 score which is a decent score. When broken down to each classification method's average F1 scores on every resampling technique, Random Forest produced the highest with 0.86 and is measured using 80:20 splitting evaluation.

Table 3. Average F1 Score Of Each Classifier On X-api Data

Classifier	Average F1 Score on Every Resampling technique	No- resampling classification result
Logistic Regression	0,75	0,69
K-NN	0,72	0,63
CART	0,78	0,71
Random Forest	**0,86**	**0,8**
SVM	0,69	0,60
Stacking	0,81	0,77
XGboost	0,84	0,75
XGBRF	0,78	0,74
Adaboost	0,73	0,68

Table 4. Average F1 Score Of Each Classifier On Por Data

Classifier	Average F1 Score on Every Resampling technique	No- resampling classification result
Logistic Regression	0,87	0,78
K-NN	0,76	0,51
CART	0,87	0,71
Random Forest	0,92	0,8
SVM	0,84	0,59
Stacking	0,93	0,81
XGboost	0,92	0,81
XGBRF	0,87	0,83
Adaboost	0,59	0,70

Table 5. Average F1 Score Of Each Classifier On Math Data

Classifier	Average F1Score on Every Resampling technique	No- resampling classification result
Logistic Regression	0,82	0,78
K-NN	0,69	0,57
CART	0,83	0,78
Random Forest	0,90	0,83
SVM	0,74	0,54
Stacking	0,89	0,82
XGboost	0,89	0,82
XGBRF	0,87	0,83
Adaboost	0,78	0,90

Table 6. Average F1 Score Of Each Classifier On Student Data

Classifier	Average F1Score on Every Resampling technique	No- resampling classification result
Logistic Regression	0,43	0,24
K-NN	0,61	0,25
CART	0,57	0,28
Random Forest	0,67	0,26
SVM	0,60	0,22
Stacking	0,64	0,28
XGboost	0,59	0,31
XGBRF	0,48	0,28
Adaboost	0,40	0,30

From Table 3 to 6, it can be seen that the resampling method improves most of the classification's performances, with Random Forest producing the best average F1 Score among all of the classification results. Some differences occur on the Por dataset, which produced the best F1 score average under the Stacking classification; Adaboost produced the highest F1 Score (0,90) among all the no-resampling results on the Math dataset. Although all differences happened in the experiment result, the goal of this study has been fulfilled in that most of the resampling methods enhance the classification result compared to the no-resampling scenario.

3.4. Influence of dataset condition toward classification's performance

According to Table 2, each resampling and classification method's performance differs on each dataset. Especially the student dataset, which is compared to the other dataset. The dataset has more class variation than the other, which may cause differences in the method's performance. With the broader variation of class and imbalance scale of each class, the methods may find it hard to synthesize the minority sample. There are way too many classes that exist. Compared to the Por and Math datasets with the relatively same imbalanced condition, the resampling and classification method can perform well and produce a better score when compared head-to-head on each classification method's result with the student dataset.

From this study can be infered that the class variation may influence the method's performance, but it must also be solidly proven. The highest result (accumulatively) was produced by most of the classification and resampling methods on the Por and Math datasets, which have four label variations. The imbalanced degree between the majority and the minority class are 1:21 and 1: 4,2 imbalanced ratio, respectively.

The results show varying results from all the resampling scenarios of the imbalanced dataset with different imbalance ratios. It is mainly affected by the different imbalanced ratios, which lead to the quality of the synthesized data from the resampling mechanism. However, the classification method shows that the resampled dataset produced better results than the original dataset.

According to the earlier-stated research question, the answer based on the finding of this research will be concluded. First, the resampling method's impact on classification results according to Table 3 to 6 in general. In a specific way, Random Forest dominates almost every scenario on every dataset. The best resampling method that can improve the results according to the experiment is the SMOTEENN resampling method, even though it excels on different classifiers on every dataset. It can be inferred that the resampling method cannot be steady on different dataset conditions or may cause the evaluation metrics (splitting and k-fold).

The different condition of imbalance between datasets (Imbalance Ratio) affects each classifier's performance. It may be caused by the performance of the resampling method in synthesizing the data. For example, on the Por dataset, SMOTE-ENN classified using Stacking produces a 0,97 F1 score, while with the same resampling and classifier on the student dataset, it produces only a 0,13 F1 score. Although all differences happened in the experiment result, the goal of this study has been fulfilled in that most of the resampling methods enhance the classification result compared to the no-resampling scenario.

Third, according to the result in Table 3 to 6, all of the resampling methods can handle the multi-classification problem well, as the classification's method performance are great too. The resampling method can give a synthetic dataset that can be used for a well-performed classification.

4. Conclusions

Data mining can help in any way possible in education, yet the interaction between data and methods can differ. Hence, experimental research in searching for the most suitable methods for a specific dataset must be done first. Multiclass classification with an imbalanced class is a specific task that needs a specific sequence of methods to work with to produce a good amount of accuracy. The best sequence found from this study uses SMOTEENN as the resampling method and utilizes the Random Forest or Stacking classification method. The future work of this study is to dig deeper into why the resampling method cannot handle the enormous class variation since the F1 score on the student dataset is lower than the other dataset.

Author Contributions

I. Pratama: Conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing – original draft, and writing - review & editing. P. T. Prasetyaningrum: Conceptualization, formal analysis, funding acquisition, methodology, resources, supervision, writing - original draft, and writing – review & editing. A. Y. Chandra: Conceptualization, data curation, funding acquisition, methodology, project administration, resources, software, validation, visualization, writing - original draft, and writing - review & editing. O. Suria: Data curation, formal analysis, funding acquisition, resources, software, writing - original draft, and writing - review & editing.

Declaration of Competing Interest

We declare that we have no conflict of interest.

References

[1] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, 2017, doi: 10.1016/j.eswa.2016.12.035.

[2] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data," *J. Biomed. Inform.*, vol. 107, no. May 2019, p. 103465, 2020, doi: 10.1016/j.jbi.2020.103465.

[3] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M. S. Hacid, and H. Zeineddine, "An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection," *IEEE Access*, vol. 7, pp. 93010–93022, 2019, doi: 10.1109/ACCESS.2019.2927266.

[4] Y. Zhang and P. Trubey, "Machine Learning and Sampling Scheme: An Empirical Study of Money Laundering Detection," *Comput. Econ.*, vol. 54, no. 3, pp. 1043–1063, 2019, doi: 10.1007/s10614-018-9864-z.

[5] B. A. Akinnuwesi *et al.*, "Application of intelligence-based computational techniques for classification and early differential diagnosis of COVID-19 disease," *Data Sci. Manag.*, vol. 4, pp. 10–18, 2021, doi: https://doi.org/10.1016/j.dsm.2021.12.001.

[6] G. Fan, Z. Deng, Q. Ye, and B. Wang, "Machine learning-based prediction models for patients no-show in online outpatient appointments," *Data Sci. Manag.*, vol. 2, pp. 45–52, 2021, doi: https://doi.org/10.1016/j.dsm.2021.06.002.

[7] M. A. Harlev, H. S. Yin, K. C. Langenheldt, R. R. Mukkamala, and R. Vatrapu, "Breaking bad: De-anonymising entity types on the bitcoin blockchain using supervised machine learning," *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, vol. 2018-Janua, pp. 3497–3506, 2018, doi: 10.24251/hicss.2018.443.

[8] I. Alarab and S. Prakoonwit, "Effect of data resampling on feature importance in imbalanced blockchain data: Comparison studies of resampling techniques," *Data Sci. Manag.*, vol. 5, no. 2, pp. 66–76, 2022, doi: 10.1016/j.dsm.2022.04.003.

[9] Y. Pristyanto, I. Pratama, and A. F. Nugraha, "Data level approach for imbalanced class handling on educational data mining multiclass classification," in *2018 International Conference on Information and Communications Technology, ICOIACT 2018*, 2018, vol. 2018-Janua. doi: 10.1109/ICOIACT.2018.8350792.

[10] E. Buraimoh, R. Ajoodha, and K. Padayachee, "Importance of Data Re-Sampling and Dimensionality Reduction in Predicting Students' Success," *3rd Int. Conf. Electr. Commun. Comput. Eng. ICECCE 2021*, no. June, pp. 12–13, 2021, doi: 10.1109/ICECCE52056.2021.9514123.

[11] D. Jahin, I. J. Emu, S. Akter, M. J. A. Patwary, M. A. S. Bhuiyan, and M. H. Miraz, "A Novel Oversampling Technique to Solve Class Imbalance Problem: A Case Study of Students&#x2019; Grades Evaluation," in *2021 International Conference on Computing, Networking, Telecommunications & Engineering Sciences Applications (CoNTESA)*, 2021, pp. 69–75. doi: 10.1109/CoNTESA52813.2021.9657151.

[12] M. Utari, B. Warsito, and R. Kusumaningrum, "Implementation of Data Mining for Drop-Out Prediction using Random Forest Method," *2020 8th Int. Conf. Inf. Commun. Technol. ICoICT 2020*, 2020, doi: 10.1109/ICoICT49345.2020.9166276.

[13] R. Ghorbani and R. Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," *IEEE Access*, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.

[14] M. Revathy, S. Kamalakkannan, and P. Kavitha, "Machine Learning based Prediction of Dropout Students from the Education University using SMOTE," in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2022, pp. 1750–1758. doi: 10.1109/ICSSIT53264.2022.9716450.

[15] P. Dabhade, R. Agarwal, K. P. Alameen, A. T. Fathima, R. Sridharan, and G. Gopakumar, "Educational data mining for predicting students' academic performance using machine learning algorithms," *Mater. Today Proc.*, vol. 47, no. xxxx, pp. 5260–5267, 2021, doi: 10.1016/j.matpr.2021.05.646.

[16] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," *Heliyon*, vol. 5, no. 2, p. e01250, 2019, doi: 10.1016/j.heliyon.2019.e01250.

[17] D. Zhang, W. Liu, X. Gong, and H. Jin, "A Novel Improved SMOTE Resampling Algorithm Based on Fractal," *J. Comput. Inf. Syst.*, vol. 7, Jun. 2011.

[18] A. Aditsania, Adiwijaya, and A. L. Saonard, "Handling imbalanced data in churn prediction using ADASYN and backpropagation algorithm," *Proceeding - 2017 3rd Int. Conf. Sci. Inf. Technol. Theory Appl. IT Educ. Ind. Soc. Big Data Era, ICSITech 2017*, vol. 2018-Janua, pp. 533–536, 2017, doi: 10.1109/ICSITech.2017.8257170.

[19] S. Ahmed, A. Mahbub, F. Rayhan, R. Jani, S. Shatabda, and D. M. Farid, "Hybrid Methods for Class Imbalance Learning Employing Bagging with Sampling Techniques," *2nd Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solut. CSITSS 2017*, pp. 1–5, 2018, doi: 10.1109/CSITSS.2017.8447799.

[20] B. S. Raghuwanshi and S. Shukla, "SMOTE based class-specific extreme learning machine for imbalanced learning," *Knowledge-Based Syst.*, vol. 187, p. 104814, 2020, doi: 10.1016/j.knosys.2019.06.022.

[21] D. Bajer, B. Zonć, M. Dudjak, and G. Martinović, "Performance Analysis of SMOTE-based Oversampling Techniques When Dealing with Data Imbalance," in *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2019, pp. 265–271. doi: 10.1109/IWSSIP.2019.8787306.

[22] F. Sağlam and M. A. Cengiz, "A novel SMOTE-based resampling technique trough noise detection and the boosting procedure," *Expert Syst. Appl.*, vol. 200, no. April 2020, pp. 1–12, 2022, doi: 10.1016/j.eswa.2022.117023.

[23] H. Guo, J. Zhou, and C.-A. Wu, "Imbalanced Learning Based on Data-Partition and SMOTE," *Information* , vol. 9, no. 9. 2018. doi: 10.3390/info9090238.

[24] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Data Level Preprocessing Methods*. 2018. doi: 10.1007/978-3-319-98074-4_5.

[25] Y. Pristyanto, N. A. Setiawan, and I. Ardiyanto, "Hybrid resampling to handle imbalanced class on classification of student performance in classroom," *Proc. - 2017 1st Int. Conf. Informatics Comput. Sci. ICICoS 2017*, vol. 2018-Janua, pp. 207–212, 2017, doi: 10.1109/ICICOS.2017.8276363.

[26] M. Zeng, B. Zou, F. Wei, X. Liu, and L. Wang, "Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data," in *2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS)*, 2016, pp. 225–228. doi: 10.1109/ICOACS.2016.7563084.

[27] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.

[28] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," *Lect. Notes Comput. Sci.*, vol. 3644, no. PART I, pp. 878–887, 2005, doi: 10.1007/11538059_91.

[29] G. Douzas and F. Bacao, "Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE," *Inf. Sci. (Ny).*, vol. 501, pp. 118–135, 2019, doi: 10.1016/j.ins.2019.06.007.

[30] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining Educational Data to Predict Student's academic Performance using Ensemble Methods," *Int. J. Database Theory Appl.*, vol. 9, no. 8, pp. 119–136, 2016, doi: 10.14257/ijdta.2016.9.8.13.

[31] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," *15th Eur. Concurr. Eng. Conf. 2008, ECEC 2008 - 5th Futur. Bus. Technol. Conf. FUBUTEC 2008*, vol. 2003, no. 2000, pp. 5–12, 2008.

[32] J. Asiya, "Student Performance Prediction." [Online]. Available: https://www.kaggle.com/datasets/asiyajan001/student-performance-prediction