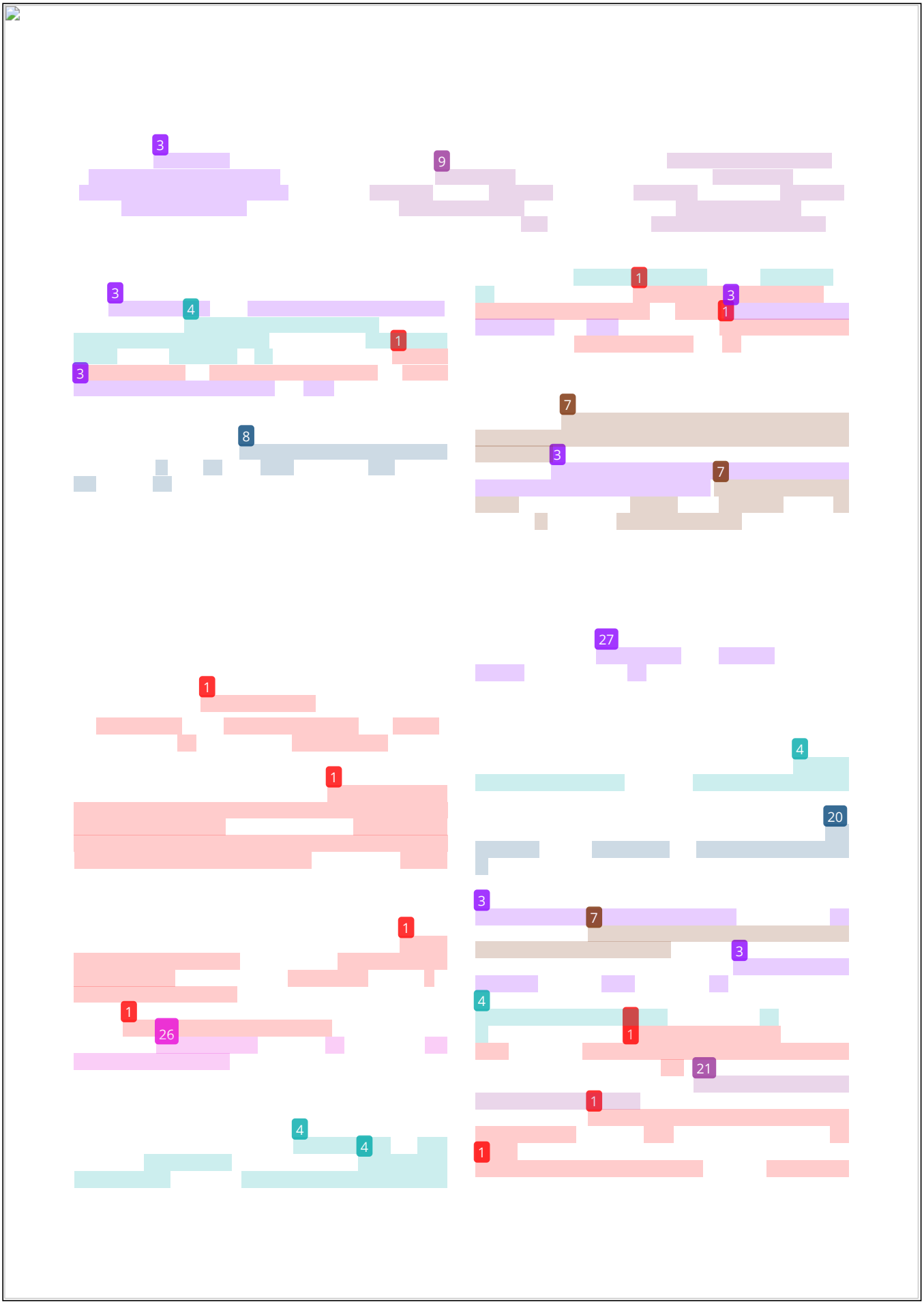
Imbalanced Class handling and Classification on Educational Dataset

WORD COUNT 4390
TIME SUBMITTED 29-AUG-2021 08:36PM
PAPER ID 76055030


3
9
3
4
1
3
1
3
1
7
8
3
7
27
1
4
1
20
3
7
1
3
4
1
1
26
21
1
4
4
1

study proposes a combination of Tomek Links (TL) [18] and Synthetic Minority Over-Sampling Technique (SMOTE) [18] to solve the problem of class imbalance in datasets, especially educational datasets.

There are three contributions of this research. First, the method that we propose can be a solution for dealing with class imbalances in the student performance classification, especially in dealing with classification problems with binary class and multiclass on student performance datasets. Second, the method we propose can increase the performance value of the classification algorithm used in this study. Third, it can be a reference for further research related to handling class imbalances in datasets in data mining research, especially in the field of education.

II. Materials and Methods

A. Materials

In this study, the dataset used is public. This dataset is a dataset related to the education process taken from kaggle.com [19], [20]. There are 480 student records and 16 characteristics in the dataset. The characteristics are divided into three groups: (1) Gender and nationality are examples of demographic characteristics. (2) Features of the academic background, such as educational stage, grade level, and section. (3) Behavioral characteristics such as raised hand-on class, resource opening, parent survey responses, and school satisfaction.

There are 305 males and 175 females in the sample. 179 are from Kuwait, 172 are from Jordan, 28 are from Palestine, 22 are from Iraq, 17 are from Lebanon, 12 are from Tunis, 11 are from Saudi Arabia, 9 are from Egypt, 7 are from Syria, 6 are from the United States, Iran, and Libya, 4 students are from Morocco, and one student is from Venezuela. The data was gathered over the course of two academic semesters: During the first semester, 245 student records are gathered, and during the second semester, 235 student records are collected. The data set also includes a school attendance feature in which students are divided into two groups depending on their absence days: 191 students have more than 7 absence days, and 289 students have fewer than 7. This dataset also includes a brand-new element called parent parturition in the educational process. Parent Answering Survey and Parent School Satisfaction are two subfeatures of the parent participation feature. 270 parents responded to the poll, while 210 did not; 292 parents are pleased with the school, while 188 are not..

The dataset is a classification dataset consisting of 3 classes, namely Low-Level, Middle-Level, and High-Level. The Low-Level consists of 126 instances, the Middle-Level consists of 211 instances, and the High-Level consists of 142 instances. Based on the number of instances in each class, it indicates a class imbalance in the dataset.

B. Methods

Generally, this research includes the following four main phases of Data Acquisition, Data Pre-processing, Classification, Evaluation, and Comparison.

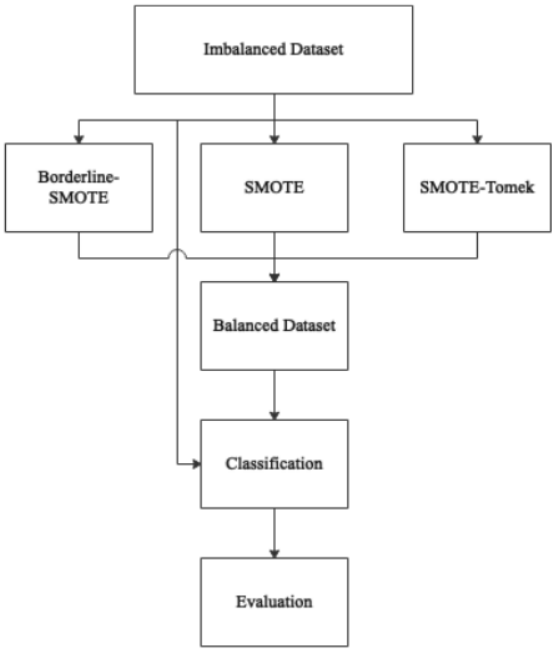


Figure 1. Research Workflow

The first stage of this research is collecting a dataset. In this study, the dataset used has been describing in the materials subsection. The second stage is data pre-processing. In this study, the resampling process was carried out using several resampling methods including combination of two resampling methods. The single resampling method is SMOTE while the combination used is SMOTE + Tomek Links and Borderline + SMOTE. SMOTE is used to increase the number of minority class instances. Meanwhile, Tomek Link and Borderline is used to reducing noise samples in the majority class.

1) SMOTE

The working principle of the method SMOTE finds the value of k-nearest neighbors, namely adjacency between data for each record in the minor class, once it is made of synthetic as much data as the percentage of the desired duplication between data minor and k-nearest neighbors are chosen randomly. The goal is to increase the amount of data on the minority class. The following figure 2 is the illustration of the SMOTE mechanism.

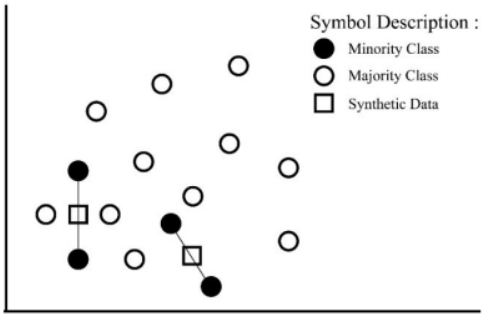


Figure 2. SMOTE Mechanism

2) Tomek Links

The Tomek Links' principle is to eliminate noisy samples and to focus on the majority class to ensure that the reduction is not excessive. In theory, noise samples are one of the factors that contribute to misclassification and a decrease in the classifier's performance. Therefore, Tomek Links is chosen as the under-sampling method for this study. Figure 3 below shows the mechanism of Tomek Links.

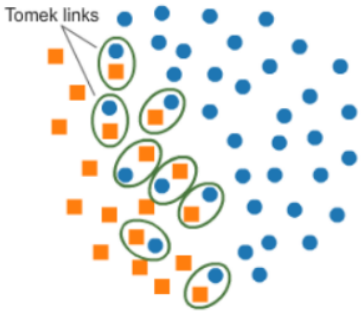


Figure 3. Tomek Links Mechanism

After pre-processing the data, the third step is to classify using several algorithms for the testing process. In this study, the classification algorithms used for testing include Logistic regression, k-Nearest Neighbor (K-NN), CART, Random Forest, Support Vector Machine (SVM), and Stacking. The testing process in this study is divided into two scenarios. The first scenario uses a dataset splitting process with a composition of 80% training data and 20% test data. The second scenario is to use 10-fold cross-validation. The main objective of both scenarios is to test the consistency of the classification algorithm model performance. Also, factors the limited number of the dataset used is another reason for the use of both scenarios.

The last stage in this research is the evaluation and comparison process. In this step, the classification methods will be compared to each other through several stages. The first is a comparison of classification methods that do not use resampling, the second is a comparison of classification methods that use Tomek Links, the third is a comparison of classification methods that use SMOTE, and the fourth is a comparison of classification methods that use Tomek Links and SMOTE. The accuracy and geometric mean (g-mean) are used as assessment indicators in this study because they are the most complete in terms of the imbalance class context. [21]. Accuracy on classification is a score indicating the level of certainty of a classed data record followed evaluation. As with G-mean, the score is calculated by multiplying the true positive rate (TPR) and the true negative rate (TNR). The G-mean statistic indicates the overall accuracy for either a minority or a majority class [22]. The accuracy and G-mean formula can be seen in the equation below.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$G-Mean = \sqrt{TPR \; x \; TNR} \quad (2)$$

In addition, other indicators used are precision, recall, and F1-Score. Precision is defined as the ratio of true positive predictions to the total number of true positive predictions. The recall statistic measures the proportion of true positive predictions to all true positive data. The F1 score, on the other hand, is a weighted average of accuracy and recall.

III. Result and Discussion

In this part of the paper, the result of the data balancing process and classification will be discussed, analyzed and compared to each other. This paper tries to present the effect of data imbalance to the classification's result and resampling methods as a way to overcome the imbalanced data problems. Additionally, different evaluation metrics of the classification also carried on to give an insight of the effect of the evaluation mechanism and so on. From those result's then the best resampling method and the classification method for all models can be determined. All presented models are built in Python using Google Collaboration platform. The classifier first executed on the original data which is imbalanced to show the performance of the classifier on the imbalanced data. After that, the data is undergoing a balancing process using several oversampling methods and then the classifiers will be implemented.

A. Original Dataset Classifications' result

Table III shows the performance of the classifiers on the original dataset. The evaluations done under two condition, using the 10-fold cross validation which resulting average accuracy scores from the iterations of the k-fold mechanism. While 80 – 20 data splitting ratios which produced several measurements such as, Precision, Recall, and F1 Scores. Table III provided the information of how the models performs under the imbalanced data condition.

TABLE I. CLASSIFICATION RESULT'S ON ORIGINAL DATASET

Classifier	Imbalanced Dataset						
	Data Splitting						*10-fold cross validation*
	Precision	Recall	Geometric Mean	F1			Average Accuracy
				Class 0	Class 1	Class 2	
Logistic regression	70%	70%	0.76	65%	79%	65%	70.9%
K-NN	66%	66%	0.73	58%	74%	63%	62.4%
CART	73%	73%	0.79	71%	83%	66%	71.3%
Random Forest	77%	76%	0.80	71%	83%	73%	**79.2%**
SVM	62%	59%	0.80	45%	71%	58%	62.5%
Stacking	**79%**	**77%**	**0.81**	**75%**	**82%**	**75%**	75.2%

Accuracy is one of the most popular evaluation metrics to measure a classifier's performance. As for the 10-fold cross validation evaluation model, Random Forest is the most accurate method to classify such imbalanced dataset. besides, the data splitting evaluation model stated that Stacking is the most precise method to classify the dataset. Recall is the probability of correctly detected items. It means that Stacking method identifies 77% of all different students in the dataset. While SVM produced the lowest Recall score with just 59%. Precision is the portion of the relevant results. Stacking has a decent score on Precision with 79%, with SVM has the lowest score. Although the dataset is imbalanced, the score of all classifier can reach over 50% which is a good sign already. Yet it is not enough to say it is a good classification's result and it is possible for some improvement.

B. Classification with Resampling Method

As stated, the class distribution is imbalanced, means the numbers of each class member is not in relatively same proportion, the eff[14] of the imbalanced class classification can be seen on the F1 score. F1 score is the harmonic average of Precision and Recall, includes an important result about classifiers' performance on each class respectively. For example, K-NN produced a relatively bad result on the F1 Score for Class 0, while the other 2 class produced score that are above 50%.

As this paper working on a dataset with multi-classification problem, the effect of the imbalanced distribution of the class is bad for the classifiers' performance. Shown from the obtained result on Table III as the accuracy is still low. Therefore, solving the imbalanc[25]ata is necessary to produce a better classification's result. Table IV, Table V, and Table VI represents the results of each machine learning technique on balanced data, which also resampled using several different resampling models.

TABLE II. SMOTE RESAMPLING METHOD'S RESULT

Classifier	**SMOTE**				
	Data Splitting				10-Cross Validation
	Recall	Specificity	Geometric Mean	F1 Score	
Logistic regression	78%	89%	0.83	78%	74.9%
K-NN	72%	82%	0.79	72%	71.5%
CART	81%	90%	0.86	81%	78.5%
Random Forest	**83%**	**92%**	**0.87**	**83%**	**82.7%**
SVM	70%	85%	0.76	69%	69.3%
Stacking	79%	89%	0.84	78%	79.6%

TABLE III. BORDERLINESMOTE RESAMPLING METHOD'S RESULT

Classifier	**Borderline-SMOTE**				
	Data Splitting				10-Cross Validation
	Recall	Specificity	Geometric Mean	F1 Score	
Logistic regression	74%	87%	0.8	73%	73.5%
K-NN	65%	82%	0.71	62%	70.8%
CART	77%	88%	0.82	77%	76%
Random Forest	**83%**	**92%**	**0.87**	**83%**	**84.8%**
SVM	67%	83%	0.72	63%	66.4%
Stacking	83%	92%	0.87	83%	76.4%

TABLE IV. SMOTE-TOMEK RESAMPLING METHOD'S RESULT

Classifier	**SMOTE-Tomek**				
	Data Splitting				10-Cross Validation
	Recall	Specificity	Geometric Mean	F1 Score	
Logistic regression	74%	88%	0.8	75%	75.6%
K-NN	68%	84%	0.75	68%	74%
CART	79%	89%	0.83	78%	77.8%
Random Forest	**86%**	**93%**	**0.89**	**86%**	**85.8%**
SVM	65%	82%	0.72	64%	70.6%
Stacking	84%	92%	0.88	84%	83.7%

Table IV represents the classifications results using SMOTE resampling technique. The SMOTE resampling method can improve the overall accuracy of all classifier with Random Forest is the strongest classifier in terms of accuracy and Geometric Mean score. Geometric Mean is a scalar which a squared-root of Recall and Specificity multiplication. The Geometric Mean is a measurement to the accuracy of any classification with resampled data. The higher the Geometric Mean score means the better the classification performance. In this case, Random Forest is the highest wielding score in both of evaluation models with 0.87 Geometric Mean and 82.7% Accuracy in 10-fold Cross Validation.

Table V represents the result of classification which the data is resampled beforehand using Borderline-SMOTE resampling technique. As shown in the table, Random Forest classification method produced the highest results on both of the evaluation models with 0.87 Geometric Mean and 84.8% accuracy. Even though there are no differences between Smote resampled and Borderline-Smote resampled data in terms of Geometric mean, but the accuracy of the Borderline-Smote resampled data is better.

Table VI represent the classification result using SMOTE-Tomek resampled dataset. Smote-Tomek itself is a hybrid resampling method which is a combination of an over-sampling and under-sampling method. As indicated by the bold font on the Table VI that the Random Forest gives the highest score on Geometric Mean and accuracy on both evaluation method. Random Forest comes out as the best classifier from each resampling models. Even challenging the Stacking method which in theory is an enhancement of a single classifier method, yet it cannot beat Random Forest in this case.

As for the resampling method, the comparison of the three resampling method will be provided in Table VII. The comparison made by taking out each best score from any classifier of each resampling method which is all of it won by Random Forest.

TABLE V. RESAMPLING METHOD COMPARISON

Resampling method + Classifier	Data Splitting		10-Cross Validation
	Geometric Mean	F1 Score	Average Accuracy
SMOTE + Random Forest	0.87	83%	82.7%
Borderline-SMOTE + Random Forest	0.87	83%	84.8%
SMOTE-Tomek + Random Forest	**0.89**	**86%**	**85.8%**

Table VII represents the comparisons of Random Forest result on each resampling method. Can be seen that SMOTE-Tomek resampling method helps Random Forest classifier to get highest score in Geometric Mean and 10-fold Cross Validation evaluation models. The results confirm that SMOTE-Tomek might be better than the other resampling methods.

IV. CONCLUSION

This study intends to show the effect of the imbalanced data problem and find out the better resampling method to be implemented into the machine learning process. The resampling method used in this study are SMOTE, Borderline SMOTE, SMOTE -Tomek. The students' performance dataset used as the data source and classify using several classifiers namely, Logistic Regression, K-NN, CART, Random Forest, SVM, Stacking ensemble method. The initial result of classifications on imbalanced dataset do not produce a good performance as the F1 score margin on each class is too big. The results of each classifier measured by two evaluation metrics to give a clearer picture of the performance about not only the classification but also the performance of the resampling method. The Geometric Mean indicated how good the resampling method provide a data for the classifier, and the accuracy provide the general picture of how good the classification's performance on each resampled dataset. It seems that SMOTE-Tomek work better on the dataset as it produced the highest Geometric Mean on the best classifier on this study which is Random Forest. Hence, the SMOTE-Tomek and Random Forest is the best pair to work with the dataset.

This study can be developed in so many ways and possible to perform future work in the following directions. A deeper experiment on ensemble classification can be introduced to give a better comparison also hopefully achieve better results. Take more resampling models can also become a good study to enrich the analysis from this study, and a better resample method may be found on the midst of the study. As from the dataset perspective, more students' performance related dataset can be added to give better understanding about the results with several adjustment such as feature selection which may give better performance.

REFERENCES

[1] 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.

[2] Presiden RI, *UU No 12 Thn 2012 ttg Pendidikan Tinggi*. 2012.

[3] Y. Pristyanto, N. A. Setiawan, and I. Ardiyanto, "Hybrid Resampling to Handle Imbalanced Class on Classification of Student Performance in Classroom," in *The First International Conference on Informatics and Computational Sciences (ICICoS 2017)*, 2017, pp. 215–220, doi: 10.1109/ICICOS.2017.8276363.

[4] I. Hidayah, A. E. Permanasari, and N. Ratwastuti, "Student classification for academic performance prediction using neuro fuzzy in a conventional classroom," *Inf. Technol. Electr. Eng. (ICITEE), 2013 Int. Conf.*, pp. 221–225, 2013, doi: 10.1109/ICITEED.2013.6676242.

[5] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 40, no. 6, pp. 601–618, 2010, doi: 10.1109/TSMCC.2010.2053532.

[6] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, "Implementing autoML in educational data mining for prediction tasks," *Appl. Sci.*, vol. 10, no. 1, pp. 1–27, 2020, doi: 10.3390/app10010090.

[7] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," *Decis. Anal.*, vol. 2, no. 1, pp. 1–25, 2015, doi: 10.1186/s40165-014-0010-2.

[8] Y. Pristyanto, N. A. Setiawan, and I. Ardiyanto, "Hybrid

resampling to handle imbalanced class on classification of student performance in classroom," in *Proceedings - 2017 1st International Conference on Informatics and Computational Sciences, ICICoS 2017*, 2018, vol. 2018-Janua, doi: 10.1109/ICICOS.2017.8276363.

[9] S. Sawangarreerak and P. Thanathamathee, "Random forest with sampling techniques for handling imbalanced prediction of university student depression," *Inf.*, vol. 11, no. 11, pp. 1–13, 2020, doi: 10.3390/info11110519.

[10] F. Shakeel, A. S. Sabhitha, and S. Sharma, "Exploratory review on class imbalance problem: An overview," 2017, doi: 10.1109/ICCCNT.2017.8204150.

[11] V. S. Spelmen and R. Porkodi, "A Review on Handling Imbalanced Data," *Proc. 2018 Int. Conf. Curr. Trends Towar. Converging Technol. ICCTCT 2018*, pp. 1–11, 2018, doi: 10.1109/ICCTCT.2018.8551020.

[12] H. Guo, J. Zhou, and C. A. Wu, "Imbalanced learning based on data-partition and SMOTE," *Inf.*, vol. 9, no. 9, 2018, doi: 10.3390/info9090238.

[13] M. S. Islam, M. Arifuzzaman, and M. S. Islam, "SMOTE Approach for Predicting the Success of Bank Telemarketing," *TIMES-iCON 2019 - 2019 4th Technol. Innov. Manag. Eng. Sci. Int. Conf.*, 2019, doi: 10.1109/TIMES-iCON47539.2019.9024630.

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

[15] E. Kurniawan, F. Nhita, A. Aditsania, and D. Saepudin, "C5.0 algorithm and synthetic minority oversampling technique (SMOTE) for rainfall forecasting in bandung regency," in *2019 7th International Conference on Information and Communication Technology, ICoICT 2019*, 2019, vol. 4, pp. 1–5, doi: 10.1109/ICoICT.2019.8835324.

[16] E. AT, A. M, A.-M. F, and S. M, "Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method," *Glob. J. Technol. Optim.*, vol. 01, no. S1, 2016, doi: 10.4172/2229-8711.s1111.

[17] M. Bach, A. Werner, and M. Palt, "The proposal of undersampling method for learning from imbalanced datasets," in *Procedia Computer Science*, 2019, vol. 159, pp. 125–134, doi: 10.1016/j.procs.2019.09.167.

[18] B. Jonathan, P. H. Putra, and Y. Ruldeviyani, "Observation Imbalanced Data Text to Predict Users Selling Products on Female Daily with SMOTE, Tomek, and SMOTE-Tomek," in *Proceedings - 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2020*, 2020, pp. 81–85, doi: 10.1109/IAICT50021.2020.9172033.

[19] 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.

[20] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Preprocessing and analyzing educational data set using X-API for improving student's performance," 2015, doi: 10.1109/AEECT.2015.7360581.

[21] M. Han, J., & Kamber, *Data Mining: Concepts and Techniques Second*, Second Edi., vol. 12. San Fransisco: Morgan Kauffman, 2006.

[22] S. Wang and X. Yao, "Using class imbalance learning for software defect prediction," *IEEE Trans. Reliab.*, vol. 62, no. 2, pp. 434–443, 2013, doi: 10.1109/TR.2013.2259203.

Imbalanced Class handling and Classification on Educational Dataset

ORIGINALITY REPORT

22%
SIMILARITY INDEX

PRIMARY SOURCES

1	Yoga Pristyanto, Noor Akhmad Setiawan, Igi Ardiyanto. "Hybrid resampling to handle imbalanced class on classification of student performance in classroom", 2017 1st International Conference on Informatics and Computational Sciences (ICICoS), 2017 Crossref	137 words — 4%
2	Yoga Pristyanto, Akhmad Dahlan. "Hybrid Resampling for Imbalanced Class Handling on Web Phishing Classification Dataset", 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2019 Crossref	132 words — 4%
3	Yoga Pristyanto, Anggit Ferdita Nugraha, Irfan Pratama, Akhmad Dahlan, Lucky Adhikrisna Wirasakti. "Dual Approach to Handling Imbalanced Class in Datasets Using Oversampling and Ensemble Learning Techniques", 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM), 2021 Crossref	96 words — 3%
4	Yoga Pristyanto, Anggit Ferdita Nugraha, Irfan Pratama, Akhmad Dahlan. "Ensemble Model Approach For Imbalanced Class Handling on Dataset", 2020 3rd International Conference on Information and Communications Technology (ICOIACT), 2020	71 words — 2%

Crossref

5	scholarworks.sjsu.edu Internet	47 words — 1%
6	www.kaggle.com Internet	45 words — 1%
7	Yoga Pristyanto, Irfan Pratama, Anggit Ferdita Nugraha. "Data level approach for imbalanced class handling on educational data mining multiclass classification", 2018 International Conference on Information and Communications Technology (ICOIACT), 2018 Crossref	41 words — 1%
8	Yoga Pristyanto, Sumarni Adi, Andi Sunyoto. "The Effect of Feature Selection on Classification Algorithms in Credit Approval", 2019 International Conference on Information and Communications Technology (ICOIACT), 2019 Crossref	31 words — 1%
9	Irfan Pratama, Putri Taqwa Prasetyaningrum, Putry Wahyu Setyaningsih. "Time-Series Data Forecasting and Approximation with Smoothing Technique", 2019 International Conference on Information and Communications Technology (ICOIACT), 2019 Crossref	22 words — 1%
10	Amjad Alowaigl, Khalil H. A. Al-Shqeerat, Mohammed Hadwan. "A multi-criteria assessment of decision support systems in educational environments", Indonesian Journal of Electrical Engineering and Computer Science, 2021 Crossref	21 words — 1%
11	Yuguo Zha, Cheng Chen, Qihong Jiao, Xiaomei Zeng, Xuefeng Cui, Kang Ning. "Ontology-Aware Deep	20 words — 1%

Learning Enables Novel Antibiotic Resistance Gene Discovery Towards Comprehensive Profiling of ARGs", Cold Spring Harbor Laboratory, 2021
Crossref Posted Content

12 Kimberly J. Petersen, Pamela Qualter, Neil Humphrey. "The Application of Latent Class Analysis for Investigating Population Child Mental Health: A Systematic Review", Frontiers in Psychology, 2019 — 18 words — < 1%
Crossref

13 Lecture Notes in Electrical Engineering, 2015. — 13 words — < 1%
Crossref

14 muabusalah.medium.com — 13 words — < 1%
Internet

15 link.springer.com — 12 words — < 1%
Internet

16 "Proceedings of the International Conference on Data Engineering and Communication Technology", Springer Science and Business Media LLC, 2017 — 10 words — < 1%
Crossref

17 Amsal Pardamean, Hilman F. Pardede. "Tuned bidirectional encoder representations from transformers for fake news detection", Indonesian Journal of Electrical Engineering and Computer Science, 2021 — 9 words — < 1%
Crossref

18 Harsurinder Kaur, Husanbir Singh Pannu, Avleen Kaur Malhi. "A Systematic Review on Imbalanced Data Challenges in Machine Learning", ACM Computing Surveys, 2019 — 9 words — < 1%
Crossref

19	Mahmudul Hasan Popel, Khan Md. Hasib, Syed Ahsan Habib, Faisal Muhammad Shah. "A Hybrid Under-Sampling Method (HUSBoost) to Classify Imbalanced Data", 2018 21st International Conference of Computer and Information Technology (ICCIT), 2018 Crossref	9 words — < 1%
20	Utomo Pujianto, Ilham Ari Elbaith Zaeni, Ninon Oktaviani Irawan. "SVM Method for Classification of Primary School Teacher Education Journal Articles", 2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE), 2019 Crossref	9 words — < 1%
21	www.mdpi.com Internet	9 words — < 1%
22	"International Conference on Innovative Computing and Communications", Springer Science and Business Media LLC, 2021 Crossref	8 words — < 1%
23	Anggit Ferdita Nugraha, Luthfia Rahman. "Meta-Algorithms for Improving Classification Performance in the Web-phishing Detection Process", 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2019 Crossref	8 words — < 1%
24	Jue Zhang, Li Chen, Jian-xue Tian, Fazeel Abid, Wusi Yang, Xiao-fen Tang. "Breast Cancer Diagnosis Using Cluster-based Undersampling and Boosted C5.0 Algorithm", International Journal of Control, Automation and Systems, 2021 Crossref	8 words — < 1%

25 Rizal Broer Bahaweres, Arif Imam Suroso, Alam Wahyu Hutomo, Indra Permana Solihin, Irman Hermadi, Yandra Arkeman. "Tackling Feature Selection Problems with Genetic Algorithms in Software Defect Prediction for Optimization", 2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), 2020
Crossref

8 words — < 1%

26 eprints.utm.my
Internet

8 words — < 1%

27 Nhlakanipho Mqadi, Nalindren Naicker, Timothy Adeliyi. "A SMOTe based Oversampling Data-Point Approach to Solving the Credit Card Data Imbalance Problem in Financial Fraud Detection", International Journal of Computing and Digital Systems, 2021
Crossref

7 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	OFF