IMCOM 2022

by Yoga Pristyanto

Submission date: 27-Sep-2021 10:24AM (UTC+0700)
Submission ID: 1658355064
File name: IMCOM_202-_Yoga_Pristyanto.docx (153.27K)
Word count: 3992
Character count: 22476

Multiclass Imbalanced Handling using ADASYN Oversampling and Stacking Algorithm

1 Yoga Pristyanto
Information System
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
yoga.pristyanto@amikom.ac.id

2nd Anggit Ferdita Nugraha
Computer Engineering
Univeristas Amikom Yogyakarta
Yogyakarta, Indonesia
anggitferdita@amikom.ac.id

3rd Akhmad Dahlan
Information Management
Univeristas Amikom Yogyakarta
Yogyakarta, Indonesia
alland@amikom.ac.id

4th Lucky Adhikrisna Wirasakti
Informatics
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
lucky.30@students.amikom.ac.id

5th Aditya Ahmad Zein
Informatics
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
aditya.z@students.amikom.ac.id

6th Irfan Pratama
Information System
Universitas Mercu Buana Yogyakarta
Yogyakarta, Indonesia
irfanp@mercubuana-yogya.ac.id

***Abstract*—Class imbalance conditions in datasets are common in real-world problems. Class imbalance is a condition where the number of classes in the dataset used in the classification process has a significant difference in number. In theory, most single classifiers have a weakness against class imbalance conditions in datasets, especially those with multiclass types, so their performance cannot be maximized. This study proposes two approaches to overcome the problem of multiclass imbalanced, namely the use of ADASYN (Adaptive Synthetic) Sampling and the Stacking Algorithm. As confirmed by testing on five multiclass datasets, the proposed method outperforms other methods in terms of accuracy values, true positive rate, true negative rate, and geometric mean values. As a result, the method proposed in this study can solve class imbalance problems in multiclass-type datasets. However, this study has limitations. Namely, the dataset used is a multiclass category with a maximum number of six classes. For this reason, further research will suggest testing using imbalanced class datasets in the category of multiclass datasets with more than six classes.**

Keywords*— *Multiclass, Imbalanced Class, ADASYN, Ensemble, Stacking.

I. INTRODUCTION

Class imbalance in the dataset is a common occurrence in a real-world problem. Class imbalance is a condition where the number of classes in the dataset used in the classification process has a significant difference [1]. Often overlooked by researchers and practitioners in machine learning, imbalance problems are particularly prevalent when applying machine learning to classification techniques. Its presence frequently causes significant problems for the classification model when it is used in the context of machine learning applications [2]. In theory, the majority of single classifiers have a weakness against class imbalance conditions in datasets. This condition caused by the number of majority classes often appears compared to the minority class during the model training process. The single classifier tends only to be able to recognize the pattern of the majority class. As a result, the performance of the single classifier is not optimal both in accuracy and other measurement metrics [3]. In addition, the existence of class imbalance is also a challenge for researchers and practitioners, especially in machine learning. Therefore, we need a solution to overcome these problems [4].

Various researches have been conducted on how the dataset's class imbalance conditions are handled. Gongzhu Hu [5], and Jishan [6] researched the effect of class imbalances on classification algorithm performance. Both studies demonstrate that by incorporating class imbalance handling, the classification algorithm's performance can be enhanced. They used the SMOTE (Synthetic Minority Over-Sampling Technique) algorithm in their research. Additionally, Imran [7] compared two oversampling methods, namely SMOTE and ROS (Random Over Sampling). The study's findings indicate that both can help improve the classification algorithm's performance. While Rashu [8] and Thammasiri [9] employed one of the undersampling methods, RUS (Under Random Sampling), the research findings indicated that the RUS method degrades the performance of classification algorithms.

On the other hand, Kubat's research employs a technique known as OSS (One-Sided Selection) [10]. The results indicated that by incorporating the OSS method, the classification algorithm's performance could be enhanced. Noorhalim [11] and Zhihao [12] also use SMOTE to treat a similar class imbalance approach. Both studies demonstrate that applying class imbalance handling to a dataset can significantly improve the performance of several classification algorithms. Additionally, Sajid Ahmed [13] researched how to deal with datasets that contain class imbalances. Ensemble resampling was used in this study. Simultaneously, SMOTE-Bagging, RUS-Bagging, ADASYN-Bagging, and RYSIN-Bagging were evaluated. The study's findings indicate that all four methods successfully enhanced the performance of the classification algorithms. Yingze Yang [14] used one of the ensemble resampling methods, SMOTE-Boosting, in his research. While the findings of this research show that the proposed approach enhances the effectiveness of the classification algorithm, it has not been tested on a range of datasets with different degrees of unbalanced ratios.

To resolve class imbalances, the majority of researchers use resampling methods. On the other hand, the resampling method has a disadvantage: it increases the potential of duplicated occurrences, which results in the loss of data and patterns in the dataset. This circumstance definitely has an effect on a single classifier's performance. Additionally, using a data-level method may change the makeup of the dataset. Simultaneously, the algorithmic method has a flaw that

renders it ineffective for datasets with a reasonably large class imbalance ratio.

This study proposes to use two approaches, namely ADASYN (Adaptive Synthetic) Sampling and Stacking Algorithm. Previous research [15] [16] using the SMOTE-OSS method still has limitations that only use a binary class imbalanced dataset. Therefore, the study focused on testing the imbalanced multiclass datasets using the proposed method. The aim is to find out whether the proposed method is effective and significant in dealing with multiclass imbalanced datasets.

II. METHODS

This research consists of four main stages: data acquisition, pre-processing, classification, and ensemble processes. The last stage is the evaluation process. The following Figure 1 is a diagram of the stages in this research.

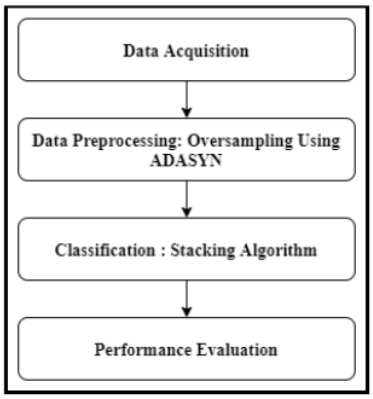

Figure 1. Research Flow

A. Data Acquisition

The dataset used is public. The dataset source is from the OpenML Datasets Repository website page[17]. In this study, five types of multiclass datasets were used. All datasets have an unbalanced number of classes. The following Table 1 is information from the five datasets used in this study.

TABLE I. DATASETS INFORMATION

Datasets	Codes	Instances	Class Proportions
Web-phishing-multiclass	D1	1353	702 : 548 : 103
xAPI-Edu-Dataset	D2	480	211 : 42 : 127
Vehicles	D3	846	218 : 217 : 212 : 199
Cars	D4	1728	1210 : 384 : 69 : 65
Dermatology	D5	366	112 : 72 : 61 : 52 : 49 : 20

Datasets can be downloaded on the following page: https://bit.ly/dataset-used

B. Data Preprocessing : Oversampling Using ADASYN

At the data preprocessing stage, the class balancing process is carried out using a synthetic oversampling approach. The algorithm used is ADASYN (Adaptive Synthetics). The ADASYN algorithm works by determining the k-nearest neighbor value. The number of neighbors between instances in the minority class will be increased. Following that, the data synthesis procedure is repeated for as many occasions as the desired percentage of minor data and adaptively picked k-nearest neighbors [18]. Data synthesis aims to increase the number of minority classes synthetically to reduce the risk of data duplication during the oversampling process. In this study, the value of beta = 1 and the value of k = 5 was used in each test of each dataset. The process of implementing the ADASYN algorithm uses python 3.8. In addition, the library used is imbalanced-learn version 0.8.0.

C. Classification : Stacking Algorithm

As illustrated in Figure 2, stacking is a technique used in ensemble algorithms. The training dataset is split into N separate subsets using stratified sampling with replacements, while maintaining the relative proportions of various classes across all subsets. Each subset of the training set is utilized to evaluate the performance of the classifier on the training set. The meta-classifier is built in terms of relative weights for each classifier by weighting classifiers according to their performance. Meta-classifiers may be defined in steps in a simple meta-learning scenario [19].

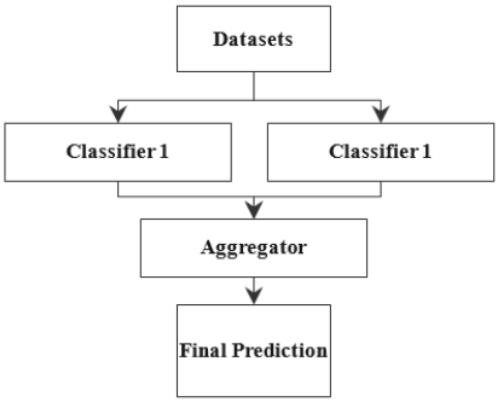

Figure 2. Stacking Algorithm Illustration

According to Figure 2, the stacking algorithm consists of the following steps:

- The initial training set is used to train basic classifiers (basic level).
- Predictions are made using classifiers that have been trained on distinct validation sets.
- The meta-level training set is constructed from the validation and prediction sets generated by the validation set's classifier.
- The meta- or final classifier is trained using the meta-level training set.

In this study, stacking was carried out using SVM (Support Vector Machine) and RF (Random Forest) as the base learner. While the aggregator algorithm is used Logistic Regression.

D. Performance Evaluation

Evaluation is a technique for determining the effectiveness of the resulting model. The confusion matrix, also known as

the error matrix is the most frequently used evaluation technique. The error matrix is a special type of table that facilitates visualizing an algorithm [20]. The following Table 2 is a confusion matrix table.

TABLE II. CONFUSION MATRIX

Actual Classification	Prediction Classification	
	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

True Positive (TP) is defined in Table 2 as the number of occurrences of a positive class that are properly predicted to be positive. The number of occurrences of negative classes correlates to the number of estimated positive False positive (FP) classes. The number of occurrences of the negative class equals the number of estimated positive False negative (FN) classes. The True Negative (TN) count is the number of occurrences of the negative class that were correctly predicted despite the fact that.

The accuracy, true negative rate (TNR), true positive rate (TPR), and geometric mean (G-Mean) were utilized to evaluate the model's performance in this research. The four indicators are comprehensive enough to conduct the assessment procedure, particularly when the dataset includes class imbalances. Accuracy is the number or percentage of properly categorized data records after the classification findings have been validated. True Positive rate measures the percentage of positives that are properly recognized. True Negative Rate is a metric that measures the percentage of properly recognized negatives. A geometric mean is a performance indicator whose value is calculated by multiplying their square root's sensitivity and specificity values. The following equation is used to determine the values of the four indicators.

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

$$True\ Positive\ Rate = \frac{TP}{TP+FN} \quad (2)$$

$$True\ Positive\ Rate = \frac{TN}{TN+FP} \quad (3)$$

$$Geometric\ Mean = \sqrt{\text{TNR}\ x\ \text{TPR}} \quad (4)$$

III. RESULTS AND DISCUSSION

Discussion of the results in this study focused on two main parts. The first is the result of the oversampling process using the ADASYN algorithm. The second is the evaluation of the performance of the classification process using the stacking algorithm.

A. *Oversampling Using ADASYN Algorithm*

At this stage, the oversampling process is carried out using ADASYN, as previously mentioned. The ADASYN parameter used is the value of k = 5 and the value of beta = 1 in each tested dataset. The following figure 3 to figure 7 is the distribution of the results of the oversampling process using ADASYN for each dataset.

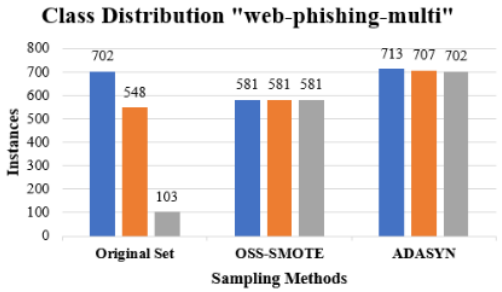

Figure 3. Oversampling Results on Web Phising Multiclass Dataset

In Figure 3, the results of the oversampling process using ADASYN on multiclass phishing web dataset show an increase in the number of synthetic minority classes so that the proportion of class distribution becomes 713 instances: 707 instances: 702 instances. So the class becomes a relatively balanced distribution. This is due to a decrease in the imbalance ratio in the dataset.

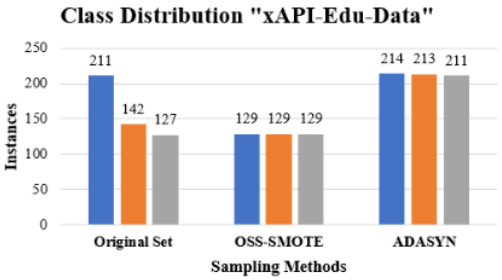

Figure 4. Oversampling Results on xAPI-Edu-Dataset

Based on Figure 4 shows an increase in the number of synthetic minority classes in the xAPI-Edu dataset. so the proportion of the class distribution becomes 214 instances: 213 instances: 211 instances. So the class becomes a relatively balanced distribution. This is due to a decrease in the imbalance ratio in the dataset.

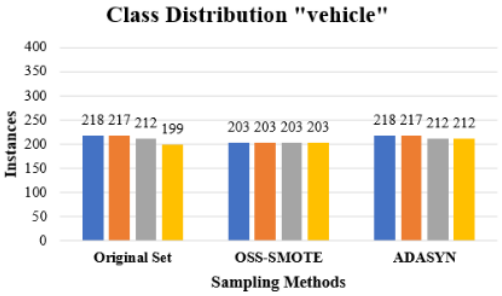

Figure 5. Oversampling Results on Vehicles Dataset

Figure 5 is the result of the oversampling process using ADASYN on the Vehicles dataset. This dataset also shows an increase in the number of synthetic minority classes. So the distribution proportion of the class is 218 instances: 217

instances: 212 instances: 212 instances. So the condition becomes relatively balanced.

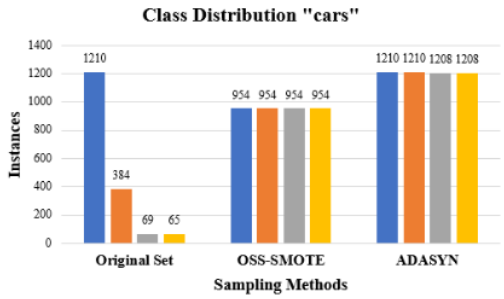

Figure 6. Oversampling Results on Cars Dataset

Figure 6 shows an increase in the number of synthetic minority classes in the Cars dataset so that the proportion of the class distribution becomes 1210 instances: 1210 instances: 1208 instances: 1208 instances. So the condition becomes relatively balanced.

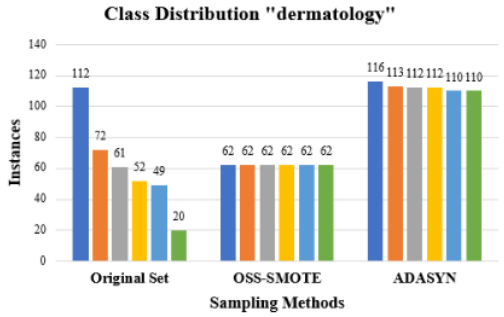

Figure 7. Oversampling Results on Dermatology Dataset

Based on Figure 7, the results of the oversampling process using ADASYN on the Dermatology dataset show an increase in the number of synthetic minority classes so that the proportion of the class distribution becomes 116 instances: 113 instances: 112: 112 instances: 110 instances: 110 instances. So the class becomes a relatively balanced distribution. This is due to a decrease in the imbalance ratio in the dataset.

Based on the five graphs above. After the oversampling process using ADASYN on all datasets. The proportion of class distribution has not shown a perfectly balanced condition compared to the oversampling process using SMOTE and OSS. This is due to the working principle of ADASYN in making samples completely synthetically. While in SMOTE OSS [15] [16], several samples of the majority class are removed, and there is also an increasing number of minority classes. However, the working principle of SMOTE itself is not completely synthetic because there is still some duplication of instances during the oversampling process. Therefore, with no duplication of instances in the improved minority class, the ADASYN algorithm should give better results. These results can be known at the next stage, namely the classification process and evaluation of the classification results from the model that has been tested.

B. Classification And Evaluation Results

The dataset is divided into two sections in this study, namely training data and testing data. The dataset's percentage distribution is as follows: 80% is used for model training, while 20% is used for model validation or testing. SVM (Support Vector Machine) and RF (Random Forest) classification algorithms were used in the testing process, while Stacking was proposed (SVM-RF). This study examines three scenarios for the testing and evaluation process. The first scenario (S1) is the test without going through the resampling process or the original set. The second scenario (S2) is tested by resampling process using OSS SMOTE [15] [16]. While the third scenario (S3) is the proposed method in this research, the test uses the ADASYN algorithm as the resampling technique used. All scenarios will be evaluated and their performance compared in terms of accuracy, true positive rate (TPR), true negative rate (TNR), and the resulting Geometric Mean (G-Mean) value.

TABLE III. AVERAGE ACCURACY SCORE OF THE CLASSIFICATION ALGORITHM

Datasets	Classifier	Scenarios		
		S1	S2	S3
D1	SVM	82%	84%	87%
	Random Forest	80%	70%	86%
	Stacking	**87%**	**93%**	**94%**
D2	SVM	65%	**82%**	78%
	Random Forest	**82%**	70%	75%
	Stacking	78%	**82%**	**82%**
D3	SVM	71%	79%	82%
	Random Forest	70%	71%	78%
	Stacking	**82%**	**85%**	**87%**
D4	SVM	80%	93%	92%
	Random Forest	71%	90%	90%
	Stacking	**96%**	**95%**	**97%**
D5	SVM	**91%**	94%	91%
	Random Forest	87%	86%	85%
	Stacking	89%	**95%**	**95%**

Following testing, Table 3 shows that the third scenario produces a greater overall accuracy value than the first and second scenarios. This required the use of ADASYN to improve the overall classification algorithm's accuracy score. Additionally, when compared to other classification algorithms, the suggested method, which involves stacking the SVM and Random Forest algorithms, may achieve the best accuracy value, especially in the third scenario.

TABLE IV. AVERAGE TRUE POSITIVE RATE SCORE OF THE CLASSIFICATION ALGORITHM

Datasets	Classifier	Scenarios		
		S1	S2	S3
D1	SVM	83%	85%	87%
	Random Forest	81%	71%	87%
	Stacking	**88%**	**93%**	**94%**
D2	SVM	66%	**82%**	**90%**
	Random Forest	**82%**	69%	76%
	Stacking	78%	**82%**	82%
D3	SVM	72%	80%	83%
	Random Forest	71%	72%	79%
	Stacking	**83%**	**86%**	**88%**
D4	SVM	81%	93%	93%
	Random Forest	72%	90%	91%
	Stacking	**97%**	**95%**	**97%**
D5	SVM	**92%**	94%	91%
	Random Forest	86%	87%	85%
	Stacking	90%	**96%**	**94%**

Table 4 demonstrates that the third scenario generates a higher overall true positive rate after testing than either scenario one or scenario two. This resulted in the resampling process using ADASYN increasing the overall classification algorithm's true positive rate value. Additionally, compared to other classification algorithms, the proposed algorithm, namely stacking the SVM and Random Forest algorithms, can produce the highest true positive rate, particularly in scenario three.

TABLE V. AVERAGE TRUE NEGATIVE RATE SCORE OF THE CLASSIFICATION ALGORITHM

Datasets	Classifier	Scenarios		
		S1	S2	S3
D1	SVM	86%	92%	94%
	Random Forest	84%	85%	93%
	Stacking	**90%**	**97%**	**98%**
D2	SVM	70%	90%	89%
	Random Forest	**88%**	86%	88%
	Stacking	87%	**91%**	**92%**
D3	SVM	91%	92%	96%
	Random Forest	90%	92%	94%
	Stacking	**95%**	**96%**	**96%**
D4	SVM	73%	98%	97%
	Random Forest	75%	97%	97%
	Stacking	**97%**	**98%**	**99%**
D5	SVM	**96%**	96%	**96%**
	Random Forest	94%	93%	90%
	Stacking	**96%**	**97%**	**96%**

After testing, Table 5 demonstrates that the third scenario's overall true negative rate value is superior to those generated by scenarios one and two. This resulted in the resampling process using ADASYN increasing the overall classification algorithm's true negative rate value. Additionally, compared to other classification algorithms, the proposed algorithm, namely stacking the SVM and Random Forest algorithms, can produce the highest true negative rate, particularly in scenario three.

TABLE VI. AVERAGE GEOMETRIC MEAN SCORE OF THE CLASSIFICATION ALGORITHM

Datasets	Classifier	Scenarios		
		S1	S2	S3
D1	SVM	0.85	0.89	0.91
	Random Forest	0.83	0.78	0.90
	Stacking	**0.89**	**0.95**	**0.96**
D2	SVM	0.68	**0.86**	0.765
	Random Forest	**0.85**	0.77	0.811
	Stacking	0.82	**0.86**	**0.844**
D3	SVM	0.81	0.86	0.89
	Random Forest	0.80	0.81	0.86
	Stacking	**0.89**	**0.91**	**0.92**
D4	SVM	0.769	0.955	0.95
	Random Forest	0.735	0.935	0.94
	Stacking	**0.97**	**0.965**	**0.98**
D5	SVM	**0.94**	0.95	0.935
	Random Forest	0.9	0.9	0.875
	Stacking	0.93	**0.965**	**0.95**

Table 6 demonstrates that the third scenario's overall geometric mean value is superior to those produced by scenarios one and two. As a result, resampling with ADASYN can be used to raise the geometric mean value of the classification algorithm utilized in its entirety. Additionally, when the suggested classification algorithm is stacked with the SVM and Random Forest algorithms, it produces the greatest geometric mean value, notably in scenario three, when compared to other classification techniques.

Generally, scenario three generates more accurate numbers for accuracy, true positive rate, true negative rate, and geometric mean than scenarios one and two. True positive rate of return, true negative rate of return, and geometric mean rate of return As a result, the strategy described in this paper, namely the stacking of the SVM and Random Forest algorithms, as well as the addition of ADASYN to the resampling process, may be a viable alternative for resolving the problem of class imbalance in multiclass datasets. This is indicated by the accuracy value, true positive rate value, true negative rate value, and the resulting geometric mean value, which is better than scenario two [15] [16] and scenario one. The four evaluation indicators are comprehensive indicators to evaluate the model in class imbalance case

IV. CONCLUSION

On the basis of the study done, it can be stated that resolving class imbalances in multiclass type datasets is important, even more so in machine learning classification. According tests performed on five multiclass datasets with imbalance conditions using the method proposed in this study, namely stacking the SVM and Random Forest algorithms and adding ADASYN during the resampling process, scenario three outperforms scenarios two and one on all measurement parameters. The approach described in this work for resolving class imbalance problems in multiclass datasets can be applied to addressing class imbalance problems in multiclass datasets. This study, however, has limitations. Specifically, the dataset used is a multiclass classification with a maximum of six classes. As a result, additional research will suggest testing with datasets with imbalanced classes within the category of multiclass datasets with more than six classes..

ACKNOWLEDGMENT

Thanks to the Department of Research and Community Service, Universitas Amikom Yogyakarta for funding and supporting this research.

REFERENCES

[1] K. Yang *et al.*, "Hybrid Classifier Ensemble for Imbalanced Data," *IEEE Trans. Neural Networks Learn. Syst.*, vol. PP, pp. 1–14, 2019, doi: 10.1109/tnnls.2019.2920246.

[2] T. Alam, C. F. Ahmed, S. A. Zahin, M. A. H. Khan, and M. T. Islam, "An effective recursive technique for multi-class classification and regression for imbalanced data," *IEEE Access*, vol. 7, pp. 127615–127630, 2019, doi: 10.1109/ACCESS.2019.2939755.

[3] Z. Yuan and P. Zhao, "An improved ensemble learning for imbalanced data classification," *Proc. 2019 IEEE 8th Jt. Int. Inf. Technol. Artif. Intell. Conf. ITAIC 2019*, no. Itaic, pp. 408–411, 2019, doi: 10.1109/ITAIC.2019.8785887.

[4] 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.

[5] G. Hu, T. Xi, F. Mohammed, and H. Miao, "Classification of wine quality with imbalanced data," *Proc. IEEE Int. Conf. Ind. Technol.*, pp. 1712–1717, 2016, doi: 10.1109/ICIT.2016.7475021.

[6] 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.

[7] M. Imran, M. Afroze, S. K. Sanampudi, A. Abdul, and M. Qyser,

"Data Mining of Imbalanced Dataset in Educational Data Using Weka Tool," *Int. J. Eng. Sci. Comput.*, vol. 6, no. 6, pp. 7666–7669, 2016, doi: 10.4010/2016.1809.

[8] R. I. Rashu, N. Haq, and R. M. Rahman, "Data mining approaches to predict final grade by overcoming class imbalance problem," *2014 17th Int. Conf. Comput. Inf. Technol. ICCIT 2014*, pp. 14–19, 2014, doi: 10.1109/ICCITechn.2014.7073095.

[9] D. Thammasiri, D. Delen, P. Meesad, and N. Kasap, "A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition," *Expert Syst. Appl.*, vol. 41, no. 2, pp. 321–330, 2014, doi: 10.1016/j.eswa.2013.07.046.

[10] M. Kubat and S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One Sided Selection," in *International Conference on Machine Learning*, 1997, vol. 97, pp. 179–186, doi: 10.1007/s13398-014-0173-7.2.

[11] N. Noorhalim, A. Ali, and S. M. Shamsuddin, "Handling Imbalanced Ratio for Class Imbalance Problem Using SMOTE," in *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*, 2017, pp. 19–29, doi: 10.1007/978-981-13-7279-7.

[12] Z. Peng, F. Yan, and X. Li, "Comparison of the different sampling techniques for imbalanced classification problems in machine learning," *Proc. - 2019 11th Int. Conf. Meas. Technol. Mechatronics Autom. ICMTMA 2019*, pp. 431–434, 2019, doi: 10.1109/ICMTMA.2019.00101.

[13] 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.

[14] Y. Yang, P. Xiao, Y. Cheng, W. Liu, and Z. Huang, "Ensemble Strategy for Hard Classifying Samples in Class-Imbalanced Data Set," *Proc. - 2018 IEEE Int. Conf. Big Data Smart Comput. BigComp 2018*, pp. 170–175, 2018, doi: 10.1109/BigComp.2018.00033.

[15] 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, pp. 310–314, doi: 10.1109/ICOIACT.2018.8350792.

[16] 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.

[17] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo, "OpenML: Networked Science in Machine Learning," *SIGKDD Explor.*, vol. 15, no. 2, pp. 49–60, 2013, doi: 10.1145/2641190.2641198.

[18] S. He, H., Bai, Y., Garcia, E., & Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In IEEE International Joint Conference on Neural Networks, 2008," in *IJCNN 2008.(IEEE World Congress on Computational Intelligence) (pp. 1322– 1328)*, 2008, no. 3, pp. 1322– 1328.

[19] N. Chanamarn, K. Tamee, and P. Sittidech, "Stacking technique for academic achievement prediction," *Int. Work. Smart Info-Media Syst. Asia (SISA 2016)*, no. Sisa 2016, pp. 14–17, 2016.

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

IMCOM 2022

ORIGINALITY REPORT

22%	%	22%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	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 Publication	19%
2	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 Publication	3%

Exclude quotes	On	Exclude matches	< 50 words
Exclude bibliography	Off		