

Similarity Check _ Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE

By Anthony Anggrawan

Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE

Christofer Satria, Hairani Hairani, and Anthony Anggrawan*

Abstract—Student graduation accuracy is one of the indicators of the success of higher education institutions in carrying out the teaching and learning process and as a component of higher education accreditation. So it is not surprising that building a system that can predict or classify students graduating on time or not on time is necessary for universities to monitor the exact number of students graduating on time. Unfortunately, machine learning or data mining methods or approaches are less accurate in classifying classes with unbalanced data. Therefore, this research purpose is to build a machine learning system that can improve classification performance on unbalanced class data between students who graduate on time and graduate late. This study applies the Synthetic Minority Oversampling Technique (SMOTE) method to improve the classifying performance of the Support Vector Machine (SVM) data mining method. The results of the study concluded that using the Smote method increased the accuracy, precision, and sensitivity of the SVM method in classifying class data of unbalanced student graduation times. The SVM performance score rises to 3% for classification accuracy, 8% for classification precision, and 25% for classification sensitivity.

Index Terms—classification, machine learning, data mining, SVM, SMOTE

I. INTRODUCTION

Although information technology supports learning today [1][2][3][4], graduation and timeliness of graduation are different achievements for all students [5]. Many factors affect the timely completion of studies for students [6][7][8]. Statistics show that the average speed/punctuality of student graduation is not the same time [9]; specifically, there is an imbalance between students who are on time or graduating quickly and those who are not on time or late for graduation [8]. Meanwhile, the graduation rate on time is one indicator of the success of learning in higher education [6][10] and is one of the elements of the assessment of higher education accreditation in Indonesia [10], in addition to other elements that indicate the success of higher education [11][12][13]. Therefore, building a system that can predict or classify the accuracy of student graduation is one way for universities to monitor the certainty of student graduation precisely and not on time [10]. However, there are obstacles encountered in building an application system in classifying the accuracy of graduation, namely the accuracy of the system constructed especially on unbalanced class data between the number of students who graduate on time and those who do not graduate on time.

33

Manuscript received xxx yy, 2021; revised xxx yy, 20xx.

Anthony Anggrawan and Hairani Hairani are with Computer Science Department, University of Bumigora, Mataram, Indonesia (e-mail: anthony.anggrawan@universitasbumigora.ac.id, hairani@universitasbumigora.ac.id).

Christofer Satria is with Visual Design Communication, University of Bumigora, Mataram, Indonesia (e-mail:chris@universitasbumigora.ac.id).

doi: xx.xxxx/ijiet.20xx.xx.x.xxx

According to information from the Ministry [36] Higher Classifying, unbalanced class data is a significant problem in machine learning and data mining. Because, after all, causes inaccuracy in classification is the imbalance of class data [14][15]. It happened because the imbalance distribution of class data causes biased classifier performance due to misclassifying the minority class or minority classes not being considered in the overall classification results [16]. Worse, machine learning methods ignore unbalanced data, so machine learning training with unbalanced class data negatively impacts machine learning performance [17]. As a result, machine learning models perform poorly in the minority class [18]. In other words, the classification method does not achieve maximum performance when applied to unbalanced class data [18][19]. That is why the problem of unbalanced data sets gets special attention in machine learning and research related to machine learning [14][16] and has become a hot issue in data mining [20][21]. In short, classification research on unbalanced classes is essential; moreover, a class imbalance is inherent in much of the natural world [22] and not just in machine learning [17].

In essence, the classification model is a popular data mining or machine learning model [23][24][25] and has its application in various fields of science [26]. The classification model is a predictive learning model through training data on the data-set to identify the pattern of relationships between attributes and classes in the data set [27][28]. Machine learning has artificial intelligence in carrying out its jobs. Artificial intelligence is today's technology widely used for various roles [29]. Through machine learning, it is possible to uncover hidden patterns in big data and classify them [30]. Although there are several classification methods: SVM, Random Forest, Naive Bayes, Decision Tree, and others [2][31][27] however, SVM is a widely known method used for classification [32]. Each classification method has a different classification accuracy level. At the same time, inaccurate classifying of events results in errors in identifying particular patterns from the data set.

SVM is a classification method used as a training system for linear learning machines [33]. As a result, machine learning can accurately perform classification [25]. However, according to Lopez et al. (2013), SVM machine learning and decision trees are unsuitable for producing good performance on unbalanced class data [34]; therefore, it is not surprising that the imbalance of data on class attributes encourages many researchers [44] study it [19][35][36]. For this reason, this study aims to improve the performance of predictions or classification of the timeliness of graduating students by using SMOTE and SVM methods. Furthermore, to prove an increase in the accuracy of

49

XX

classifying or predicting classes on time for graduation, this study compared the performance results between the SVM method combined with the SMOTE method and the SVM method without the combination with the SMOTE method.

SMOTE is a resampling method [37] that can improve classification performance on unbalanced data, especially when combined with other methods [38]. However, the question is whether the application of SMOTE can improve the predictive performance of SVM data mining methods on unbalanced class data on the student graduation timeliness data-set? Also, how much precision/accuracy/sensitivity is the application of SMOTE in improving the classification or predictive performance of the SVM data mining method on unbalanced data from the class on the timeliness of graduation students? The research proves it.

Further discussion in this manuscript is as follows. The second subsection deals with related work. The third subsection describes the research methodology. The fourth sub-section explains the results and discussion of the research. Finally, the fifth sub-section is a sub-section of Conclusions that discusses conclusions, updates, and suggestions for further investigation.

27

II. RELATED WORK

Some of the latest related works of previous research are as follows.

Bartosz Krawczyk (2016) discusses the challenges open to researchers and future research directions for unbalanced data class [14]. The previous research differs from the research in this article not only in the research method but also in the research objectives. The previous research was a literature study review paper on unbalanced data classes. In contrast, the research in this article is an experimental study to improve the prediction performance of unbalanced class data from data on student graduation timeliness.

Dina Elreedy and Amir F. (2019) presented an analysis of the SMOTE method [39]. This last study introduced how to overcome the classification problem of unbalanced data in the minority class by generating additional data from the minority class using SMOTE. So this previous research has a different objective (focus) compared to the research in this article. The previous research describes how SMOTE makes unbalanced class data into balance class data. In contrast, this article's research improves the SVM method's performance in classifying unbalanced data on student graduation accuracy data. In the meantime, Justin M. Johnson and Taghi Khoshgoftaar (2019) surveyed the literature on using deep learning methods to address class data unbalances [22]. The previous research was survey research to overcome unbalanced class data with deep learning methods. In contrast to the research in this article, a trial study of the application of the SMOTE method to improve the accuracy of the SVM method classification in dealing with unbalanced class data.

Harshita Patel et al. (2020) reviewed the classification of unbalanced data on wireless sensor networks [16]. However, this previous research has different objectives, objects, and methods compared to the study conducted in this article. Pradeep Kumar, Roheet Bhatnagar, Kuntal Gaur, and Anurag Bhatnagar (2021) presented various approaches to classifying unbalanced data sets [17]. The main difference lies in the research methods and objectives

5

between the previous research and the research in this article. The previous research was a review study related to the unbalanced class data classification approach. In contrast, the research in this article is an experimental study to improve the classification performance of the SVM data mining method.

12

Meanwhile, Shujuan Wang, Yuntao Dai, Jihong Shen, and Jingxue Xuan (2021) proposed the use of the SMOTE method to improve the classification results of the Random Forest classification method for several data sets [20]. However, this previous study focused on enhancing classification performance using SMOTE on the Random Forest data mining method and not on student pass accuracy data. In contrast, this research focused on improving classification performance with SMOTE on the SVM data mining method on unbalanced student pass accuracy data.

Cui Yin Huang and Hong Liang Dai (2021) reviewed the class data imbalance in the Decision Tree method [26]. The difference with this article is in the research objectives and research methods. Previous research focused on discussing unbalanced class data on the Decision Tree method. In contrast, the research in this article focuses on testing classification performance to unbalanced class data on the student graduation timeliness data set on the SVM method.

In contrast, Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu (2021) proposed a scheme that can decide the composition of the training data for federated learning to reduce the impact of class data imbalance [40]. This previous study proposed a method for detecting class data imbalances in federated learning and reducing the effect of class data imbalance, in contrast to the research in this article, which focuses on applying the SMOTE method to improve prediction accuracy on unbalanced class data in the SVM method.

Wanwan Zheng and Mingzhe Jin (2022) investigated the performance effect of unbalanced class data and training data measures for classification [15, 41]. This previous research is an empirical study on the Naive Bayes, logistic regression, and Tree methods. Previous research compared balanced and unbalanced data to measure the accuracy of data mining methods; in contrast to this article's research, the mining method improves performance (accuracy, precision, and sensitivity) by applying the SMOTE method to the mining method. The research in this article then compares the performance of the data mining method between those implementing the SMOTE method and those not using the SMOTE method.

The review of several prior research-related works confirms that the study of this article differs from previous associated works. The findings of this study help reveal the impact of increasing classification accuracy arising from the application of the SMOTE method to the data set on the imbalance in the timeliness of students' graduation in the SVM method. The novelty of this study lies in improving the classification performance or prediction of unbalanced class data on student graduation timeliness which previous researchers have never done. In other words, the advantage of this research is that this research is an experimental study on unbalanced data of student graduation timeliness with SMOTE on the SVM classification method that has not been carried out by previous research. Table 1 shows the comparison between the previous related studies and this article.

TABLE I. COMPARISON OF THIS ARTICLE'S WORK WITH SOME PREVIOUS RELATED WORKS

Research By	Type of research	Method Used		Performance Testing			Research Object	Research Data / Data Set
		SVM	SMOTE	Accuracy	Precision	Sensitivity		
Bartosz Krawczyk (2016) [14]	Review	No	No	Yes	Yes	Yes	Reviewing methods for dealing with unbalanced class data problems on the Decision Tree method	Various data sets depending on the reviewed article, for example, Behavior, Cancer malignancy grading, Hyperspectral data, and others
Dina Elreedy and Amir E. (2019) [38]	Theoretical and experimental	No	Yes	Yes	No	No	Test the classification accuracy using SMOTE on K-nearest neighbors (KNN) method	Multivariate Gaussian distribution data
Justin M. Johnson and Taghi M. Khoshgoftaar (2019) [21]	Survey	No	No	No	No	No	Surveying existing deep learning techniques to overcome unbalanced class data	Various data sets depending on the surveyed article, for example, CIFAR-10, Public cameras, Building changes, and others
Harshita Patel et al. (2020) [16]	Review	No	No	No	No	No	Troubleshooting data imbalance issues of a wireless sensor network on the KNN method	No specifically mention
Pradeep Kumar et al. (2021) [17]	Review	Yes	No	No	No	No	Reviewing various data imbalance issues and learning strategies and algorithms from the Random Forest, KNN, Decision Tree, Neural Network, Naive Bayes, and SVM classification techniques.	No specifically mention (except imbalanced data)
Shujuan Wang et al. (2021) [20]	Experimental	No	Yes	Yes	No	No	Improving classification results Random Forest method for multiple data sets	Pima, WDBC, WPBC, Ionosphere, and Breast-cancer-Wisconsin
Cui Yin Huang and Hong Liang Dai (2021) [26]	Experimental	No	Yes	Yes	Yes	Yes	Reviewing the class data imbalance in the Decision Tree method	Yeast, Glass, Cleveland, and Vehicle
Lixu Wang et al. (2021) [40]	Experimental	No	No	No	No	No	Propose a scheme to decide the composition of training data to reduce the impact of class data imbalance	Clients or server data
Wanwan Zheng and Mingzhe Jin (2022) [41]	Experimental	No	No	No	No	No	Investigating the performance effects of unbalanced class data and training data measures for classifiers in the Naive Bayes, logistic regression, and Tree models	Ozone, Kc1 , Scene, Gesture, Cpu_act, Waveform-5000, Spambase, and Madelone
Our/this research	Experimental	Yes	Yes	Yes	Yes	Yes	Test the performance of the SVM method classification on the timeliness of graduating students	Student Graduation Data

III. RESEARCH METHODOLOGY

This study uses data mining stages, as shown in Figure 1.

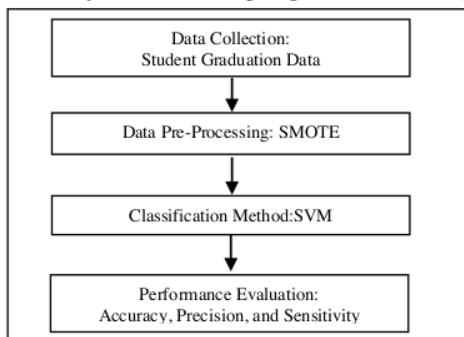


Fig. 1. Research stages

A. Data Collection

Data collection was carried out at Bumigora University. The data set was taken from graduation data for undergraduate students for the 2019-2021 academic year, totaling 265 data and having eight attributes. The attributes of this research data set are shown in Table 1. The data used as machine learning training data in this study is achievement index (IP) data from student graduation data for six semesters. Thus, machine learning has the intelligence to predict students who have completed their studies up to semester six whether these students will graduate on time or not. The sample data for students' graduation is shown in Table II.

TABLE I: STUDENT GRADUATION DATASET ATTRIBUTES

No	Attribute Name	Information	Data Type
1	JK	Gender	Nominal (Male, Female)
2	IPS 1	Semester 1 IP	Numerical
3	IPS 2	Semester 2 IP	Numerical
4	IPS 3	Semester 3 IP	Numerical
5	IPS 4	Semester 4 IP	Numerical
6	IPS 5	Semester 5 IP	Numerical
7	IPS 6	Semester 6 IP	Numerical
8	Graduation Status	Class	Nominal (On Time, Not On Time)

TABLE II: STUDENT GRADUATION DATASET

No	JK	IPS1	IPS2	...	IPS6	Status Graduation
1	F	3.06	3.16	...	3.17	On-Time
2	F	3.41	3.43	...	3.44	On-Time
3	M	2.43	2.61	...	2.67	Not On Time
4	F	3.5	3.53	...	3.53	On-Time
5	M	2.07	2.22	...	2.32	Not On Time
6	F	3.42	2.85	...	3.5	On-Time
7	M	3.33	3.28	...	3.15	Not On Time
8	F	2.83	2.05	...	2.66	Not On Time
9	M	2.94	2.21	...	3.1	Not On Time
10	M	2.56	2.0	...	2.68	Not On Time
...
264	M	2.69	1.85	...	2.5	Not On Time
265	F	2.22	1.83	...	2.21	Not On Time

B. Data Pre-Process

Data Pre-processing is one of the crucial stages in data mining to improve the quality of data sets. This study focuses on dealing with unbalanced data contained in student graduation data sets. The dataset has 171 data classes that are not on time and 94 data on time. The algorithm used to handle unbalanced data in the data-set is SMOTE (Synthetic Minority Oversampling Technique).

SMOTE is one of the most commonly used oversampling methods to solve the problem of data distribution in machine learning modeling. SMOTE aims to balance the distribution of classes by increasing the number of minority classes by synthesizing data for oversampling purposes [29]. Creating new data for the minority class uses equation (1).

$$Y^i = Y^j + (Y^j - Y^i) * \gamma \tag{1}$$

Y^i : is to hold the result of the new data. Y^j : represents the minority class. Y^j : is a randomly selected value from the k-nearest neighbors of the minority class Y^i , and γ : is a randomly selected value in a random vector with a range of 0 to 1 [42]. SMOTE generates new synthesis training data by linear interpolation for the minority class. Synthesis training data is generated by randomly selecting one or more of the k-nearest neighbors for each sample in the minority class, as shown in Figure 2.

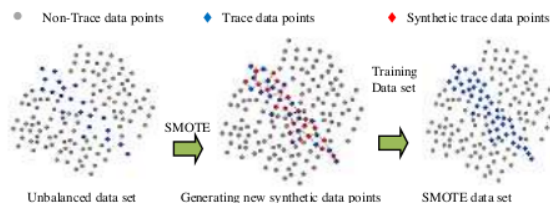


Fig. 2. Synthetic Minority Oversampling Technique (SMOTE) Algorithm Working Process [43]

C. Classification Method

The realization of classification data mining using data mining methods or machine learning algorithms involves two data sets: the first is the data-set for training, and the second is for testing. Each item set involves the attributes and categories of each training attribute with a specific target value.

This study uses the SVM data mining method to classify student graduation. Before classification, the data-set is first divided into training and testing data using 10-fold cross-validation, divided into 10 data groups using python tools.

The SVM data mining method is a supervised learning classification method aiming to find the optimal hyperplane by maximizing the distance or margin between data classes using equation (2) [44].

$$w^T x + b \tag{2}$$

$$w^T \cdot x_i + b \geq +1 \text{ when } y_i = +1 \tag{3}$$

$$w^T \cdot x_i + b \leq -1 \text{ when } y_i = -1 \tag{4}$$

w is a weight vector; x is the input vector; b is biased.

The SVM method works not only on linear data but also on nonlinear data. The technique uses two approaches to transform nonlinear data into linear data: soft margin hyperplane and feature space. The soft margin hyperplane approach in converting nonlinear data into linear ones is with the slack ξ variable formulation, as shown in equations (5) and (6).

$$x_i. w_i + b \geq 1 - \xi \text{ for } y_i = \text{class 1} \tag{5}$$

$$x_i. w_i + b \leq -1 + \xi \text{ for } y_i = \text{class 233} \tag{6}$$

D. Performance Evaluation

Evaluation (testing) of performance uses a confusion matrix. The Confusion Matrix helps calculate the amount of data classified as true and false, as shown in Table 3.

TABLE 3: CONFUSION MATRIX

Actual	Prediction	
	On-time	Not on time
On-time	TP	FN
Not On time	FP	TN

The formula for calculating accuracy, precision, and sensitivity is as follows: [28] [45]

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \tag{7}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{9}$$

True Positive (TP) is a class on time that is correctly predicted. False Positive (FP) is a class that is not on time but

11 is predicted to be on time. True Negative (TN) is an incorrectly predicted class on time. False Negative (FN) is a class that is on time but 25 predicted not to be on time.

Accuracy states the closeness of the measurement results to the actual value, while precision shows how close the difference in 10 measurement results is on repeated measurements. On the other hand, sensitivity states the level of success in retrieving information. Measurement of accuracy is based on the ratio between the correct predictions (positive and negative) with the overall data. In contrast, precision measurements are based on the percentage of true positive predictions compared to overall positive predicted outcomes. 22 Meanwhile, the recall measurement is based on the ratio of true positive predictions compared to the general actual positive data.

IV. RESULT AND DISCUSSION 28

This research starts from the stages of data collection, 10 a pre-processing, classification, and performance testing. The data used in this study is the graduation data of students. 6 pre-processing this research uses the Smote algorithm to deal with class imbalances in the graduation data used. The results of comparing the original data with the data from Smote are shown in Figure 3.

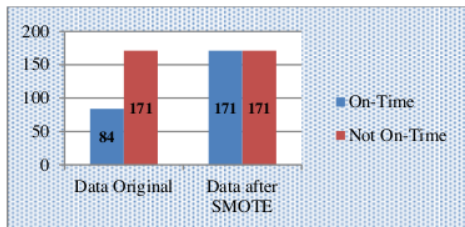


Fig. 3. The results of the comparison of the original data with the data from the Smote

35 Table 4 shows the results of testing the SVM method with a confusion matrix using 10-fold cross-validation. Meanwhile, Table 5 shows the results of the Confusion Matrix with the SVM method and the SMOTE method.

TABLE 4: CONFUSION MATRIX RESULT OF SVM METHOD

Actual	Predicted	
	On-Time	Not On Time
On-Time	50	44
Not On Time	23	148

TABLE 5: CONFUSION MATRIX RESULT OF SVM AND SMOTE METHODS

Actual	Predicted	
	On-Time	Not On Time
On-Time	134	37
Not On Time	43	128

Table 6 and Figure 2 show an increase in the performance of the SVM method with Smote based on accuracy, precision, and sensitivity. Without Smote, the SVM method has 74% accuracy, 68% precision, and 53%

sensitivity. While using Smote, the SVM method has an accuracy of 77%, 76% precision, and a sensitivity of 78%. In other words, the SVM performance score using Smote for accuracy increased by 3%, precision increased by 8%, and sensitivity increased by 25%. Thus, this 20 concludes that using the Smote method improves the accuracy, precision, and sensitivity of the SVM method in managing unbalanced class category data.

TABLE 6: PERFORMANCE RESULT OF CLASSIFICATION METHOD

Method	Accuracy	Precision	Sensitivity
SVM	74%	68%	53%
SVM with SMOTE	77%	76%	78%

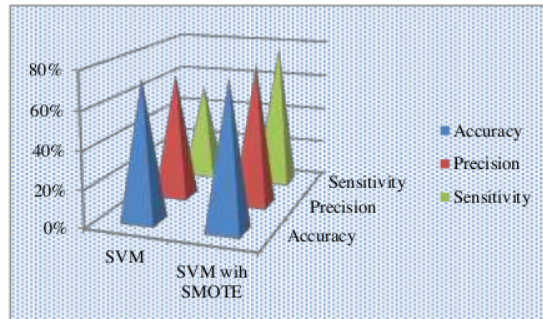


Fig. 4. Performance result of classification Method

V. CONCLUSION

The results of this study prove that the SMOTE method helps improve the performance of the accuracy, precision, and sensitivity of the SVM data mining method or the SVM machine learning algorithm in managing unbalanced student graduation time data. Furthermore, the results show the novelty of the discovery, namely the SVM performance score using SMOTE to reach 3% for the accuracy of the classification results of unbalanced class data on student graduation timeliness and up to 25% for the sensitivity of the classification results of unbalanced class data on student graduation timeliness. Meanwhile, using SMOTE, the SVM performance score increased its precision by 8% in predicting students' on-time and not on-time graduation.

47 Further research needs to conduct SMOTE testing for other machine learning algorithms. In addition, it is necessary to further develop the results of this research by building a Web or Cloud-based application program and testing its implementation on users.

19
CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

All authors participated in completing the research and writing of this article. The level of roles and tasks of research work is the basis that places each author as a correspondent author, first author, and second author.

REFERENCES

- [1] S. Smith, D. Cobham, and K. Jacques, "The Use of Data Mining and Automated Social Networking Tools in Virtual Learning Environments to Improve Student Engagement in Higher Education," *Int. J. Inf. Educ. Technol.*, vol. 12, no. 4, pp. 263–271, 2022.
- [2] C. Teoh, S. Ho, K. S. Dollmat, and C. Tan, "Ensemble-Learning Techniques for Predicting Student Performance on Video-Based Learning," *Int. J. Inf. Educ. Technol.*, vol. 12, no. 8, pp. 741–745, 2022.
- [3] A. Anggrawan, C. Satria, Mayadi, and N. G. A. Dasriani, "Reciprocity Effect between Cognitive Style and Mixed Learning Method on Computer Programming Skill," *J. Comput. Sci.*, vol. 17, no. 9, pp. 814–824, 2021.
- [4] A. Anggrawan, "Interaction between learning preferences and methods in face-to-face and online learning," *ICIC Express Lett.*, vol. 15, no. 4, pp. 319–326, 2021.
- [5] V. Bocsí *et al.*, "The discovery of the possible reasons for delayed graduation and dropout in the light of a qualitative research study," *J. Adult Learn. Knowl. Innov.*, vol. 3, no. 1, pp. 27–38, 2019.
- [6] N. Mohammad Suhaimi, S. Abdul-Rahman, S. Motalib, N. H. Abdul Hamid, and A. Md Ab Malik, "Predictive Model of Graduate-On-Time Using Machine Learning Algorithms," *Commun. Comput. Inf. Sci.*, vol. 1100, no. September, pp. 130–141, 2019.
- [7] A. Anggrawan, A. H. Yassi, C. Satria, B. Arafah, and H. M. Makka, "Comparison of Online Learning Versus Face to Face Learning in English Grammar Learning," in *The 5th International Conference on Computing Engineering and Design (ICCED)*, 2018, pp. 1–4.
- [8] T. Ojha, G. L. Heileman, M. Martinez-Ramon, and A. Slim, "Prediction of graduation delay based on student performance," in *Proceedings of the International Joint Conference on Neural Networks*, 2017, vol. 2017-May, pp. 3454–3460.
- [9] J. M. Aiken, R. de Bin, M. Hjorth-Jensen, and M. D. Caballero, "Predicting time to graduation at a large enrollment American university," *PLoS One*, vol. 15, no. 11 November, pp. 1–28, 2020.
- [10] D. Kurniawan, A. Anggrawan, and H. Hairani, "Graduation Prediction System On Students Using C4.5 Algorithm," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 19, no. 2, pp. 358–365, 2020.
- [11] A. Anggrawan, "Percentage of Effect of Blended Learning Model on Learning Outcome," in *Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019*, 2019.
- [12] M. Cheon, O. Lee, C. Mun, and H. Ha, "Factors Affecting Academic Achievement in SW Education," *Int. J. Inf. Educ. Technol.*, vol. 12, no. 4, pp. 333–338, 2022.
- [13] A. Anggrawan, Mayadi, C. Satria, and L. G. R. Putra, "Scholarship Recipients Recommendation System Using AHP and Moora Methods," *Int. J. Intell. Eng. Syst.*, vol. 15, no. 2, pp. 260–275, 2022.
- [14] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016.
- [15] D. S. Rajput, R. S. Thakur And, and G. S. Thakur, "A computational model for knowledge extraction in uncertain textual data using karnaugh map technique," *Int. J. Comput. Sci. Math.*, vol. 7, no. 2, pp. 166–176, 2016.
- [16] H. Patel, D. Singh Rajput, G. Thippa Reddy, C. Iwendi, A. Kashif Bashir, and O. Jo, "A review on classification of imbalanced data for wireless sensor networks," *Int. J. Distrib. Sens. Networks*, vol. 16, no. 4, pp. 1–15, 2020.
- [17] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of Imbalanced Data: Review of Methods and Applications," in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1099, no. 1, p. 012077.
- [18] 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.
- [19] D. Devi, S. K. Biswas, and B. Purkayastha, "A Review on Solution to Class Imbalance Problem: Undersampling Approaches," in *2020 International Conference on Computational Performance Evaluation, ComPE 2020*, 2020, pp. 626–631.
- [20] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on expansion and classification of imbalanced data based on SMOTE algorithm," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, 2021.
- [21] D. S. Rajput, "Review on recent developments in frequent itemset based document clustering, its research trends and applications," *Int. J. Data Anal. Tech. Strateg.*, vol. 11, no. 2, pp. 176–195, 2019.
- [22] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, 2019.
- [23] D. Prasad, S. Kumar Goyal, A. Sharma, A. Bindal, and V. Singh Kushwah, "System Model for Prediction Analytics Using K-Nearest Neighbors Algorithm," *J. Comput. Theor. Nanosci.*, vol. 16, no. 10, pp. 4425–4430, 2019.
- [24] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura, "Early dropout prediction using data mining: A case study with high school students," *Expert Syst.*, vol. 33, no. 1, pp. 107–124, 2016.
- [25] A. Anggrawan, C. Satria, C. K. Nuraini, Lusiana, N. G. A. Dasriani, and Mayadi, "Machine Learning for Diagnosing Drug Users and Types of Drugs Used," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, pp. 111–118, 2021.
- [26] C. Y. Huang and H. L. Dai, "Learning from class-imbalanced data: review of data driven methods and algorithm driven methods," *Data Sci. Financ. Econ.*, vol. 1, no. 1, pp. 21–36, 2021.
- [27] A. Yosipof, R. C. Guedes, and A. T. García-Sosa, "Data mining and machine learning models for predicting drug likeness and their disease or organ category," *Front. Chem.*, vol. 6, no. May, pp. 1–11, 2018.
- [28] H. Hairani, A. Anggrawan, A. I. Wathan, K. A. Latif, K. Marzuki, and M. Zulfikri, "The Abstract of Thesis Classifier by Using Naive Bayes Method," in *Proceedings - 2021 International Conference on Software Engineering and Computer Systems and 4th International Conference on Computational Science and Information Management, ICSECS-ICOCOSIM 2021*, 2021, no. August, pp. 312–315.
- [29] K. Shankar, A. R. W. Sait, D. Gupta, S. K. Lakshmanaprabu, A. Khanna, and H. M. Pandey, "Automated detection and classification of fundus diabetic retinopathy images using synergic deep learning model," *Pattern Recognit. Lett.*, vol. 133, pp. 210–216, 2020.
- [30] G. Schneider, "Automating drug discovery," *Nat. Rev. Drug Discov.*, vol. 17, no. February, pp. 97–113, 2018.
- [31] P. Kaur, M. Singh, and G. S. Josan, "Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector," *Procedia Comput. Sci.*, vol. 57, pp. 500–508, 2015.
- [32] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanes, P. Ghamisi, and S. Homayouni, "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 6308–6325, 2020.
- [33] J. C. Moreano and N. B. L. S. Palomino, "Global facial recognition using gabor wavelet, support vector machines and 3d face models," *J. Adv. Inf. Technol.*, vol. 11, no. 3, pp. 143–148, 2020.
- [34] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci. (Nijl.)*, vol. 250, pp. 113–141, 2013.
- [35] A. Onan, "Consensus Clustering-Based Undersampling Approach to Imbalanced Learning," *Sci. Program.*, vol. 2019, pp. 1–14, 2019.
- [36] X. Wang, H. Wang, D. Wu, Y. Wang, and R. Zhou, "A fuzzy consensus clustering based undersampling approach for class imbalanced learning," *ACM Int. Conf. Proceeding Ser.*, vol. December, pp. 133–137, 2019.
- [37] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, no. Sept. 28, pp. 321–357, 2002.
- [38] V. López, A. Fernández, M. J. Del Jesus, and F. Herrera, "A hierarchical genetic fuzzy system based on genetic programming for addressing classification with highly imbalanced and borderline datasets," *Knowledge-Based Syst.*, vol. 38, pp. 85–104, 2013.
- [39] D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Inf. Sci. (Nijl.)*, vol. 505, pp. 32–64, 2019.
- [40] L. Wang, S. Xu, X. Wang, and Q. Zhu, "Addressing Class Imbalance in Federated Learning," in *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, 2021, vol. 11B, pp. 10165–10173.
- [41] W. Zheng and M. Jin, "The Effects of Class Imbalance and Training Data Size on Classifier Learning: An Empirical Study," *SN Comput. Sci.*, vol. 1, no. 2, pp. 1–11, 2020.
- [42] T. Zhu, Y. Lin, and Y. Liu, "Synthetic minority oversampling technique for multiclass imbalance problems Tuanfei," *Pattern Recognit.*, vol. 72,

no. December, pp. 327–340, 2017.

- [43] J. Chen, H. Huang, A. G. Cohn, D. Zhang, and M. Zhou, "Machine learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning," *Int. J. Min. Sci. Technol.*, vol. 32, no. 2, pp. 309–322, 2022.
- [44] B. Singh and B. C. S. Rai, "Analysis of Support Vector Machine-based Intrusion Detection Techniques," *Arab. J. Sci. Eng.*, vol. 45, no. July, pp. 2371–2383, 2019.
- [45] A. Luque, A. Carrasco, A. Martín, and A. De, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, pp. 216–231, 2019.

Similarity Check _ Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE

ORIGINALITY REPORT

18%

SIMILARITY INDEX

PRIMARY SOURCES

- | | | |
|---|---|---------------|
| 1 | www.ijiet.org
Internet | 76 words — 2% |
| 2 | thesai.org
Internet | 63 words — 2% |
| 3 | Wenhao Xie, Gongqian Liang, Zhonghui Dong, Baoyu Tan, Baosheng Zhang. "An Improved Oversampling Algorithm Based on the Samples' Selection Strategy for Classifying Imbalanced Data", Mathematical Problems in Engineering, 2019
Crossref | 43 words — 1% |
| 4 | hdl.handle.net
Internet | 34 words — 1% |
| 5 | journal.universitasbumigora.ac.id
Internet | 29 words — 1% |
| 6 | 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
Crossref | 21 words — 1% |

-
- 7 Niel Ananto, Ani Budi Astuti, Achmad Efendi. "Ensemble Analysis of the Students Length of Study at University of Klabat Manado Indonesia", Journal of Physics: Conference Series, 2021
Crossref 20 words — < 1%
-
- 8 thescipub.com
Internet 20 words — < 1%
-
- 9 www.nature.com
Internet 18 words — < 1%
-
- 10 Márquez-Vera, Carlos, Alberto Cano, Cristobal Romero, Amin Yousef Mohammad Noaman, Habib Mousa Fardoun, and Sebastian Ventura. "Early dropout prediction using data mining: a case study with high school students", Expert Systems, 2015.
Crossref 17 words — < 1%
-
- 11 Prismahardi Aji Riyantoko, Sugiarto, I Gede Susrama Mas Diyasa, Kraugusteeliana. "'F.Q.A.M'" Feyn-QLattice Automation Modelling: Python Module of Machine Learning for Data Classification in Water Potability", 2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS, 2021
Crossref 17 words — < 1%
-
- 12 Shujuan Wang, Yuntao Dai, Jihong Shen, Jingxue Xuan. "Research on expansion and classification of imbalanced data based on SMOTE algorithm", Scientific Reports, 2021
Crossref 16 words — < 1%
-
- 13 Shaifu Gupta, Manpreet Singh. "Phylogenetic Method for High-Throughput Ortholog Detection", International Journal of Information Engineering and Electronic Business, 2015 15 words — < 1%

-
- 14 infocomp.dcc.ufla.br 15 words — < 1%
Internet
-
- 15 "International Conference on Innovative Computing and Communications", Springer Science and Business Media LLC, 2020 14 words — < 1%
Crossref
-
- 16 Justin M. Johnson, Taghi M. Khoshgoftaar. "Medicare fraud detection using neural networks", Journal of Big Data, 2019 14 words — < 1%
Crossref
-
- 17 easychair.org 14 words — < 1%
Internet
-
- 18 researchonline.ljmu.ac.uk 14 words — < 1%
Internet
-
- 19 worldwidescience.org 13 words — < 1%
Internet
-
- 20 Anita Desiani, Sugandi Yahdin, Annisa Kartikasari, Irmeilyana Irmeilyana. "Handling the imbalanced data with missing value elimination SMOTE in the classification of the relevance education background with graduates employment", IAES International Journal of Artificial Intelligence (IJ-AI), 2021 12 words — < 1%
Crossref
-
- 21 Díez-Pastor, José F., Juan J. Rodríguez, César I. García-Osorio, and Ludmila I. Kuncheva. "Diversity techniques improve the performance of the best imbalance learning ensembles", Information Sciences, 2015. 10 words — < 1%
Crossref

22 S A Pamungkas, I Jaya, M Iqbal. "Segmentation of *Enhalus acoroides* seagrass from underwater images using the Mask R-CNN method", IOP Conference Series: Earth and Environmental Science, 2021

Crossref

10 words — < 1%

23 Xian Shuai, Yulin Shen, Siyang Jiang, Zhihe Zhao, Zhenyu Yan, Guoliang Xing. "BalanceFL: Addressing Class Imbalance in Long-Tail Federated Learning", 2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), 2022

Crossref

10 words — < 1%

24 ijece.iaescore.com

Internet

10 words — < 1%

25 I Gede Susrama Mas Diyasa, Agus Prayogi, Intan Yuniar Purbasari, Ariyono Setiawan, Sugiarto, Prismahardi Aji Riantoko. "Data Classification of Patient Characteristics Based on Nutritional Treatment Using the K-Nearest Neighbors Algorithm", 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), 2021

Crossref

9 words — < 1%

26 core.ac.uk

Internet

9 words — < 1%

27 repository.futminna.edu.ng:8080

Internet

9 words — < 1%

28 www.icicel.org

Internet

9 words — < 1%

29 "Hybrid Artificial Intelligent Systems", Springer Science and Business Media LLC, 2020

Crossref

8 words — < 1%

30 "Proceedings of the Future Technologies Conference (FTC) 2020, Volume 1", Springer Science and Business Media LLC, 2021 8 words — < 1%
Crossref

31 "Soft Computing in Data Science", Springer Science and Business Media LLC, 2019 8 words — < 1%
Crossref

32 "Software Engineering Perspectives in Intelligent Systems", Springer Science and Business Media LLC, 2020 8 words — < 1%
Crossref

33 Dingding Li, Niyang Zhang, Mianxiong Dong, Hao Chen, Kaoru Ota, Yong Tang. "PM-AIO: An Effective Asynchronous I/O System for Persistent Memory", IEEE Transactions on Emerging Topics in Computing, 2021 8 words — < 1%
Crossref

34 [Dspace.lib.cranfield.ac.uk](https://dspace.lib.cranfield.ac.uk) 8 words — < 1%
Internet

35 Liqaa Habeb Al-Obaydi, Ragad M Tawafak, Marcel Pikhart, Wagdi Rashad Ali Bin-Hady, Erkan Yüce, Nashruddin Nashruddin. "Music as a Motivational Factor in Second Language Acquisition: Cross-cultural Insights from Iraq, Turkey, Oman, Yemen, and Indonesia", Research Square Platform LLC, 2022 8 words — < 1%
Crossref Posted Content

36 Zhi-Hua Zhou. "Training cost-sensitive neural networks with methods addressing the class imbalance problem", IEEE Transactions on Knowledge and Data Engineering, 1/2006 8 words — < 1%
Crossref

37	arxiv.org Internet	8 words — < 1%
38	discovery.researcher.life Internet	8 words — < 1%
39	dokumen.pub Internet	8 words — < 1%
40	dspace.jaist.ac.jp Internet	8 words — < 1%
41	turcomat.org Internet	8 words — < 1%
42	www.aimspress.com Internet	8 words — < 1%
43	www.ijrte.org Internet	8 words — < 1%
44	www.researchgate.net Internet	8 words — < 1%
45	www.science.gov Internet	8 words — < 1%
46	Nurafifah Mohammad Suhaimi, Shuzlina Abdul-Rahman, Sofianita Mutalib, Nurzeatul Hamimah Abdul Hamid, Ariff Md Ab Malik. "Chapter 11 Predictive Model of Graduate-On-Time Using Machine Learning Algorithms", Springer Science and Business Media LLC, 2019 Crossref	7 words — < 1%
47	Xu Du, Juan Yang, Jui-Long Hung. "An Integrated Framework Based on Latent Variational	7 words — < 1%

Autoencoder for Providing Early Warning of At-Risk Students",
IEEE Access, 2020

Crossref

48 Qinbin Li, Bingsheng He, Dawn Song. "Model-Contrastive Federated Learning", 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021

6 words — < 1%

Crossref

49 docplayer.org

Internet

5 words — < 1%

EXCLUDE QUOTES OFF

EXCLUDE SOURCES OFF

EXCLUDE BIBLIOGRAPHY ON

EXCLUDE MATCHES OFF