Handling the imbalanced data with missing value elimination SMOTE in the classification of the relevance education background with graduates employment

Anita Desiani, Sugandi Yahdin, Annisa Kartikasari, Irmeilyana
Department of Mathematics and Natural Science, Universitas Sriwijaya, Indonesia

Article Info

Article history:

Received Feb 22, 2020
Revised Nov 25, 2020
Accepted Mar 22, 2021

Keywords:

Graduates employment
Imbalanced data
Minority class
Missing value
SMOTE

ABSTRACT

The imbalanced data affect the accuracy of models, especially for precision and sensitivity, it makes difficult to find information on minority class. The problem is identified in the tracer study dataset Universitas Sriwijaya that has 2934 data. The label attribute is divided into several label classes, namely not tight, somewhat-tight, tight, very tight, and tightest. The number of the tightest and very tight is 27% and 38.6% of the number majority classes. In the study, the SMOTE is combined with eliminating the missing value of data to handle the imbalanced data. The method was evaluated by the classification methods KNN, ANN, and C4.5. The results of these methods show a significant increase in accuracy as a whole and a significant increase in the precision and sensitivity of minority classes. The precision and sensitivity of both the majority and minority are not too different, although the number of the minority is very less compared to the majority class. the information on minority classes can be obtained with quite high precision and sensitivity. As a conclusion, the proposed method is passably to improve accuracy and greatly affects the increase in sensitivity and precision.

Corresponding Author:

Anita Desiani
Department of Mathematic and Natural Science
Universitas Sriwijaya
Palembang-Prabumulih Street, km.32, South Sumatera, Indonesia
Email: Anita_desiani@unsri.ac.id

1. INTRODUCTION

Educational data mining (EDM) is technique of data mining used in educational data and it is usually used for classification [1]. Aldowah *et al.* [2] stated that EDM and learning analytics in education institutions can develop educational institutions as well as continuous improvement. It shows that EDM and LA can provide probability and solve to various learning problems such as learning analysis, predictive analysis, behavior analysis, and visualization analysis. Research by Natek and Zwilling [3] that uses student's data shows that data mining contributes effectively to planning strategies and policies even before the student registration process. Tracer study is a form of educational data that provides valuable information for evaluating educational outcomes at tertiary institutions which are used as further development of institutions in guaranteeing quality [4]. Research using tracer study was also carried out by Johnson *et al.* [5], namely research on graduates of pediatric degree option programs (PDOP) to study the pattern of confidence with pediatric pharmacotherapy and categorize the initial their employment after graduation. Universitas Sriwijaya has tracer studies every year conducted by the career development center as CDC-UNSRI. One of CDC's tasks is to trace the extent of graduate employment in the world of work through tracer study questionnaires.

This tracer study questionnaire provides 11 questions covering the ability of graduate employment both in the college major, the research projects experience and the internships experience.

The problems in imbalanced dataset often occur after data retrieval, for example, rare events or rarely experienced by respondents [6]. Some studies have a focus on the imbalanced data problem. Crone and Finlay [7] handled Imbalanced data by the sample size method, research by Louzada *et al.* [8] used sample selection models, and Zhang *et al.* [9] used the deep generative adversarial networks method to handle imbalanced data. In the CDC's tracer study dataset, there is an imbalanced data on the attributes of the relevance education background with graduate employment. The most number of data in attribute the relevance educational background with graduates employment is owned by the class "not tight" namely 787 data, while there is number of the class less than 50% is "tightest" as much as 214 data, and "very tight" class as much as 304 data. The imbalanced data in each class can affect the accuracy, precision and sensitivity [10], [11]. The accuracy is the level of the success of the correct classification of all classifications for whole data regardless of the class of the data [12], so the accuracy is not enough for imbalanced data because the accuracy does not take the accuracy on each class of labels in output attribute. There are number of indicators to measure the success in classifications such as accuracy, precision, sensitivity, and so on, but for imbalanced data is recommended to use accuracy, precision, and sensitivity [13]. The imbalanced data problem can affect the machine's ability to classify. the accuracy of imbalanced data can be very high, it may be due to the training data and test data selected by the majority class, so the minority class is automatically recognized as the majority class. On imbalanced data that has high accuracy is not a guarantee of the precision and sensitivity of one or several classes are also high, especially for a class whose numbers are very few, it is difficult to obtain high precision and sensitivity. If the majority class is better recognized than information about the minority class then the minority class cannot be found and called back by the machine. This means that the precision and sensitivity values of the class are poor. It is necessary to consider the precision and sensitivity are used to measure the reliability of the machine to find and call back data in each class on the output label attribute, so the level accuracy for each class between requests and predictive answers can be measured [13]–[15].

The synthetic minority oversampling technique (SMOTE) is a simple method to get over the imbalanced data [10]. Several studies have shown that the SMOTE method can improve the accuracy of classifications as mentioned in studies [11], [16]–[18]. The imbalanced data is done at the pre-processing stage. Chawla *et al.* [10] states that the SMOTE method works by multiplying the number of data in the minor class to be equivalent to the major class by generating artificial data based on the k-nearest neighbor (KNN). Several studies on the imbalanced data do not only use the SMOTE method in general but also combine it with other methods or modify the SMOTE method itself. Research by Zhang *et al.* [9] combines SMOTE with hybrid features to deal with the imbalanced data before classifications are made. Sun *et al.* [19] conducted the SMOTE method for oversampling on several labels with a small number of data and undersampling techniques on several labels. Both of these techniques are only used at the data training stage, while at the testing stage the original data is used without oversampling or undersampling. Pan *et al.* [20] developed the SMOTE method to be adaptive-SMOTE by selecting the data groups in the minority classes before they were reproduced, so they could strengthen the characteristics of the original data distribution without crossing the category boundaries. The studies did not pay attention to the missing value of data on their research. Incomplete data can influence the succesful of classification methods [21]. The handling of missing value in a data can be overcome by predicting the missing value or can delete data directly that has the missing value. This technique of removing data can be done more easily than predicting the missing value. This research uses a combination of missing value elimination techniques for handling missing data and the SMOTE method for handling the imbalanced data at the preprocessing stage of the tracer study dataset in Universitas Sriwijaya to improve the success of classification the relevance educational background with graduates employment.

2. RESEARCH METHOD

2.1. Data collection

The dataset was taken from Carrier Development Center Universitas Sriwijaya (CDC-UNSRI) that graduates in 2014-2016. Data were obtained from 2934 graduates who filled out online questionnaires on the CDC-UNSRI website. The number of attributes was 11 attributes, 1 attribute as a label output and 10 other attributes as label inputs. The attributes used in the study were presented in Table 1.

The attribute of the relevance educational background with graduates employment classified into 5 classes based on the label attribute on the 11th attribute with successive values of 1-5 meaning, namely tightest (TT), very tight (VT), tight (T), somewhat tight (ST), not tight (NT). TT has 214 data, VT has 304 data, T class has 475 data, ST class has 463 data and NT class has 787 data. It can be seen as graphically in Figure 1.

Table 1. The attributes and information about data

No.	Attributes	Value of attributes	Missing data
1	The College Major	1: Excellent 2: Very Good 3: Good Low 4: Fair 5: Poor	Null
2	The research projects experience	1: Excellent 2: Very Good 3: Good Low 4: Fair 5: Poor	Null
3	The internships experience	1: Excellent 2: Very Good 3: Good Low 4: Fair 5: Poor	Null
4	The extracurricular activities	1: Excellent 2: Very Good 3: Good Low 4: Fair 5: Poor	Null
5	The competence of education background	1: Excellent 2: Very Good 3: Good Low 4: Fair 5: Poor	Null
6	The relevant coursework	1: Excellent 2: Very Good 3: Good Low 4: Fair 5: Poor	Null
7	The English skill	1: Excellent 2: Very Good 3: Good Low 4: Fair 5: Poor	Null
8	Internet and computer knowledge	1: Excellent 2: Very Good 3: Good Low 4: Fair 5: Poor	Null
9	The research skill level	1: Excellent 2: Very Good 3: Good Low 4: Fair 5: Poor	Null
10	The communication skill	1: Excellent 2: Very Good 3: Good Low 4: Fair 5: Poor	Null
11	The relevance educational background with graduates employment	1: Tightest (TT) 2: VeryTight (VT) 3: Tight (T) 4: somewhat tight (ST) 5: not Tight (NT)	There are 691 missing value

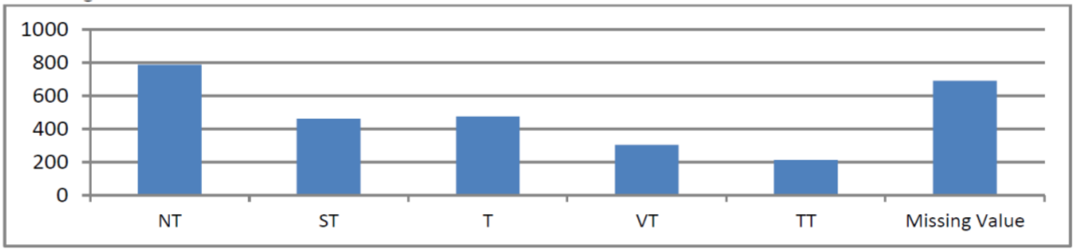

Figure 1. The original overall data of the attribute of the relevance educational background with graduates employment based on five classes

2.2. Eliminating missing value

In this step, the data would be separated the missing value data with the complete data by eliminating missing value. The eliminating missing value was done by deleting data that was carried out to handle missing value on the attribute of the relevance educational background with graduates' employment. The attribute had 691 missing data values. The data was deleted from the data set. The total data before SMOTE in this study was 2243 data. The number of data on each class is presented in Figure 2.

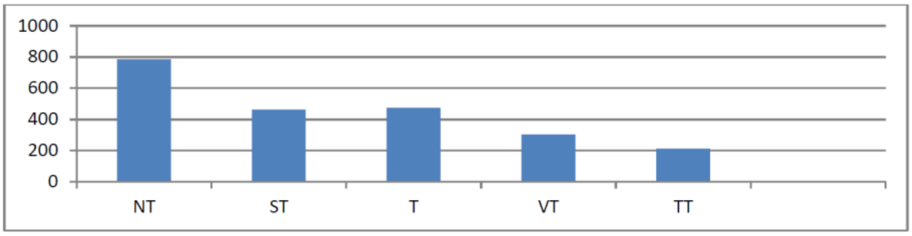

Figure 2. The total data of the attribute of the relevance educational background with graduates employment after missing value elimination

2.3. One hot-encoding transformation

The dataset consist of categorical variabels. The SMOTE just works for numerical variables so the variables in dataset should be transform to integer values. A one hot encoding is a representation of categorical variables as binary vectors, but it requires to be mapped to integer values [22]. The integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1. In the dataset the value of data has used integer value between 1 and 5.

2.4. SMOTE implementation

The SMOTE method would be applied to these steps [10]:

a. Determine the minority labels to be replicated by looking at the number of data on each label.

b. Calculate Euclidean distances between minority data with the (1):

$$d(x_a, x_b) = \sqrt{\sum_{i=1}^{n}(x_{ai} - x_{bi})^2} \quad (1)$$

where d(x_a , x_b): distance between x_{ai} is a value of the i-th data in test data set and x_{bi} is a value of the i-th data in training data set. n is the number of attributes or variables in the data set.

c. Based on the smallest distance d, generate artificial data with the (2),

$$x_{syn} = \left((x_{knn} - x_i)\ \delta\right) + x_i \quad (2)$$

where x_{syn} is artificial created data. x_i is the i-th replicated and x_{knn} is the data that has the smallest distance from the x_i data. δ is a random number between 0 and 1 if it

d. Repeat the process for every data in minority data class until data is balanced with the majority data class.

2.5. Evaluation

In this process, an evaluation of the classification of the field of study with work is carried out based on the accuracy of each test. The evaluation is carried out using test data by three classifications methods:

a. K-nearest neighbor (KNN) classification
KNN classification works by finding the closest distance between new data or test data with data that is already known to its class (training data) using (1) [23].

b. Artificial neural network (ANN) classification
In [24] states that the ANN classification works by adjusting the weights with each connection. In this study, it uses multilayer perceptron with 3 layers that consist of input layer (x_i), one hidden layer (z_i) and output layer (Y). The sigmoid bipolar is applied as activation function for output in the study in (3).

$$f(x) = \frac{1-e^{-x}}{1+e^{-x}} \quad (3)$$

c. C4.5 classification
C4.5 models can be easily integrated and also can be applied with continuous values and discrete values. C4.5 works by finding the best and right attributes to divide the data into classes by recognizing the attributes in the training data to be the root node of a tree. This algorithm finds the value of gain with entropy reduction to gets the optimal branching giving by [25]:

$$Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{k}\frac{|S|_i}{|S|} \times Entropy(S_i) \quad (4)$$

Where S is all possible value of attribute A, |S| is case number of set and |S_i| is number of the-ith case in the set. The entropy (S) is calculated by:

$$Entropy(S) = \sum_{i=1}^{n} -p_i \times ln_2 p_i \quad (5)$$

Where pi is probability distribution of the-ith p.

2.6. Analysis of results

In this paper, evaluation calculations include accuracy, precision, and sensitivity based on the results of the classification test KNN, ANN, and C4.5. To evaluate the accuracy of the proposed method, a confusion matrix is applied. The confusion matrix is a matrix containing information about the actual classification results that can be predicted by a classification system. Rows in the matrix represent the actual class in the test data and columns in the matrix represent classifications in each class [26]. The accuracy, precision dan sensitivity can be calculated by (7), (8) and (9), where TP is true positive predictions (true positive). FP is wrong positive predictions (false positive). TN is correct negative predictions (true negative), and FN is wrong negative predictions (false negative).

$$\text{Accuracy} = \left(\frac{TP+TN}{TP+FP+FN+TN}\right) \cdot 100\% \quad (6)$$

$$\text{Precision} = \frac{TP}{TP+FP}.100\% \quad (7)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}.100\% \quad (8)$$

In general, the proposed method is explained in Figure 3.

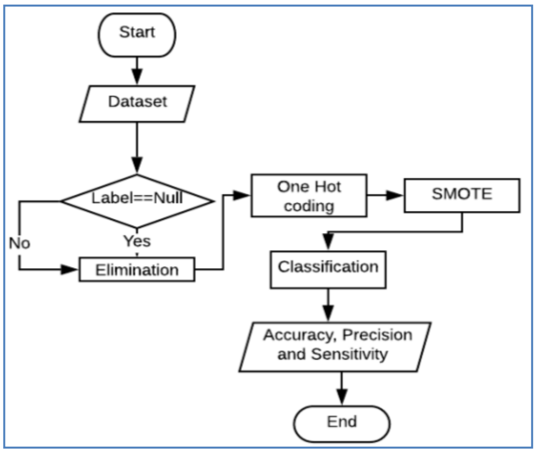

Figure 3. The proposed method in the study

3. RESULTS AND ANALYSIS

Before the SMOTE process, the data was tranformed by one hot-encoding. The replicated data started from the TT class because TT was a minority class in the initial data. The number of TT data which was originally 214 data was replicated to as much as 787 data so that it was balanced with the majority data. The number of data after the SMOTE method on the TT class became 2816 data. Then the SMOTE process was continued on the VT, T, and ST classes so that the final total number of data became 3935 data with each class totaling 787 data. Every class had 787 data. The total data for this study was 3935 data. The classification of the relationship between the relevance educational background with graduates employment was done by 3 test classifications namely KNN, ANN, and C4.5. Testing is done by comparing the results on research data before SMOTE and the results on research data after using SMOTE to handle imbalanced data. The testing results of the SMOTE method was done by comparing the values of accuracy, precision, and sensitivity of the data before SMOTE and after SMOTE. The performance measure of the classification success rate was based on the confusion matrix for each test classification method. The accuracies of the methods before and after handling the imbalanced with missing value elimination were increased. The Accuracies and the differences results from classifications on research data using the KNN, ANN and C4.5 methods were presented in Table 2.

Table 2. Accuracy results from classifications on research data using the KNN, ANN, and C4.5

Number of data/method	Before SMOTE (%)	After SMOTE (%)	Difference (%)
KNN	62.15	83.71	21.56
ANN	60.81	80.41	19.6
C4.5	61.79	81.02	19.23
Average	61.58	81.71	20.13

In the imblanced data problem, the other important things are the ability to find information in each class on the label attribute. According to Beckmann *et al.* [14], the accuracy is not always relevant to evaluate the Imbalanced data problem, because it did not take into account the number of samples distributed among classes. There were other evaluations for imbalanced datasets namely precision, and sensitivity.

According to Ali *et al.* [12] sensitivity is a description of how well classifying positive classes and precision is a measure of accuracy in the proportion of observations from positive classes correctly classified as positive.

In the study, the minority classes were owned by TT and VT classes where their number were less than 50% of majority class number. The Precisions and the sensitivity of the classes were very low compared to other labels. The precision and the sensitivity for each class label from KNN, ANN and C4.5 was showed on Tables 3-5. From Tables 3 and 5, for data before SMOTE, the precisions for TT and VT classes was only ranged around 53% and sensitivity was only around 52%. In the ANN method as shown in Table 4, for data before SMOTE, the results of precisions and sensitivities were smaller compared to other classification methods. The smallest precision was obtained on the VT class which was around 50%, while the smallest sensitivity is obtained by the TT class with 48%. However, after SMOTE the results of precision and sensitivity increased and the result was not much different from the other classes. Even though the ST class was not a minority that was less than 50%, the precision result produced before SMOTE was higher than the sensitivity. The precisions and the sensitivity for each class on the label attribute after handling the imbalanced data with missing value elimination were significantly increased. The highest precision increase was experienced by the smallest minority labels namely TT and VT, while the highest sensitivity increase was explained by the ST class as the second minority label which was more than 50% of majority class number. Graphically the increasing in accuracy, precision and sensitivity of each class on the label attribute before and after the SMOTE for each classification method was in Figure 4. Figure 4 showed that SMOTE has increased the ability to obtain and call information on minority classes. The SMOTE has given increased the precision and the sensitivity results on each minority class and the results are not much different from the results of the precision and the sensitivity in majority classes such as NT and T.

Table 3. The precision and the sensitivity of KNN classification method for each class in label attribute

Output labels	Before SMOTE		After SMOTE		Difference	
	Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity
NT	68.8	75.8	93.4	92	24.6	13.2
ST	62.5	44.4	87.3	81.5	24.8	37.1
T	61.9	66.6	76.6	80.7	14.7	14.1
VT	53.3	52.9	83.7	82.6	30.4	29.7
TT	47.3	50.5	76.5	80.7	29.2	30.2
Average	58.76	58.04	83.5	83.5	24.74	24.86

Table 4. The precision and the sensitivity of ANN classification method for each class in label attribute

Output labels	Before SMOTE		After SMOTE		Difference	
	Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity
NT	69.8	72.8	74.1	91.3	4.3	18.5
ST	63.4	46.1	89	77.2	25.6	31.1
T	59.2	67.8	76.4	76.3	17.2	8.5
VT	50	46.2	85.5	92.5	35.5	46.3
TT	39.4	48.3	79.1	64.8	39.7	16.5
Average	46.36	56.24	80.82	80.42	21.26	24.18

Table 5. The precision and the sensitivity of C4.5 classification method for each class in label attribute

Output Labels	Before SMOTE		After SMOTE		Difference	
	Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity
NT	69.1	74.4	91.3	88.58	22.2	14.18
ST	61.7	44.6	82.4	80.2	20.7	35.6
T	60.4	66.6	75.4	75.2	15	8.6
VT	53.1	52.6	80.6	81.1	27.5	28.5
TT	48.2	51.9	74.2	78.5	26	26.6
Average	58.5	58.02	80.78	80.72	22.28	22.7

Based on Tables 6 and 7, it showed there was an increase of classifications both accuracy, precision, and sensitivity occurred in UNSRI's tracer study dataset compared to other studies. The highest increase of accuracy, precision and sensitivity was in the proposed method of KNN. The proposed method of ANN when it was before SMOTE, it had the lowest sensitivity, but after SMOTE it increased the average sensitivity value for each class better than C4.5, but for the presicion on C4.5 was a higher increase than ANN. From the results obtained, it could be seen that the precision and sensitivity of the minority classes significantly increased. This described that the proposed method was feasible to improve the accuracy especially the precision and the sensitivity of each class in the tracer study dataset for classifications of the relevance of education background with graduate employment. In Table 7, the accuracy of data in [17], [18], [27]-[28] was more than 90% but the increase of accuracy was less than other studies. Combination of SMOTE with hybrid scheme [18] increase accuracy than original SMOTE that applied in [11], [17], [27] and SMOTE that combine with axiomatic fuzzy SMOTE [28].

Based on Figure 4, it was shown that on a class that was propagated the result of its precision and sensitivity increase and it is balanced between all class in label attribute. By increasing precision and sensitivity values, the accuracy of the classifications on each class was also increased and the information or pattern of each class obtained by the machine was more visible. Based on the results, it could be explained that the SMOTE method could provide opportunities to find information on minority data. The SMOTE method has been widely carried out in the research and study for a variety of data and classifications. In Table 7, there were several studies for handling imbalanced data by SMOTE and combine with classification methods like KNN, naive bayes and SVM, in many subjects. The studies' results would be compared with the result on this study in Table 4 to find out how well and precisely the proposed method in this study.

Table 6. Several studies on handling imbalanced data by SMOTE modified and classification methods

No.	Kind of dataset	Enhancement of SMOTE	Classification
1.	Educational Data [11]/2019	Original SMOTE	K-Star
2.	Educational Data [11]/2019	Original SMOTE	K-NN
3,	Real pedagogical Data [17]/2018	Original SMOTE	J48
4.	Undergraduate student data [18]/2013	Rebalancing with a hybrid scheme SMOTE	Random Forest
5.	Cancer data [20]/2019	SMOTE and Gaussian distribution	SVM
6.	Freshmen Student Data [27]/2014	Original SMOTE	SVM
7.	Cryotheraphy Data [28]/2020	SMOTE Axiomatic Fuzzy Set theory	Adaboost
8.	Academic Data [29]/2019	Attribute Weighted SMOTE	KNN
9.	German Credit Approval Scoring Data [30]/2017	Original SMOTE	Random Forrest
10.	Proposed method with KNN	Proposed method	C4.5
11.	Proposed Method with ANN	Proposed method	ANN
12.	Proposed method with C4.5	Proposed method	KNN

Table 7. The comparison on several studies for imbalanced data by SMOTE and classification methods

No	Kind of dataset	Accuracy after SMOTE (%)	The increase		
			The accuracy (%)	The precision (%)	The sensitivity (%)
1.	Educational data [11]	82.21	5.42	Unknown	Unknown
2.	Educational data [11]	87.84	7.93	Unknown	Unknown
3.	Data *real pedagogical* [17]	92.98	7,8	4	8
4.	Undergraduate student data [18]	93.33	14.09	Unknown	Unknown
5.	Cancer data [20]	Unknown	Unknown	9	10
6.	Freshmen Student Data [27]	90.22	3.8	-23	-18
7.	Cryotheraphy Data [28]	90.33	1.5	Unknown	Unknown
8.	Academic Data [29]	Unknown	Unknown	12	27
9.	German Credit Approval Scoring Data [30]	Unknown	Unknown	8	6
10.	Proposed Method with KNN	83.71	21.56	24.74	24.86
11.	Proposed Method with ANN	80.41	19.6	21.26	24.18
12.	Proposed Method with C4.5	81.02	19.23	22.28	22.7

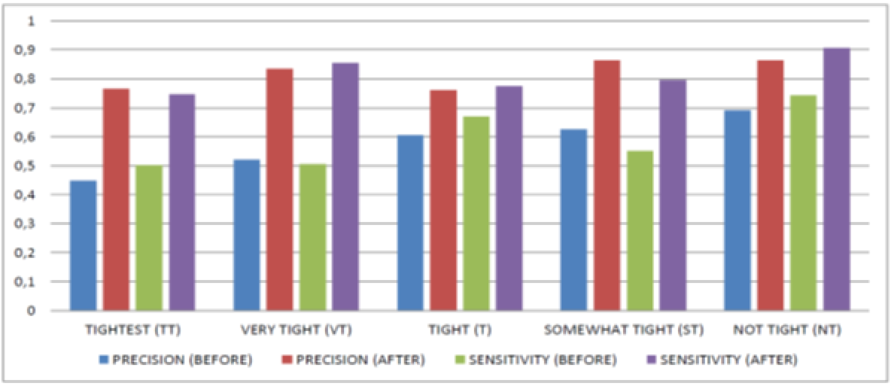

Figure 4. The precision and the sensitivity average before and after SMOTE of KNN, ANN and C4.5

On the previous studies, some literature only compared the accuracy but not for precision and sensitivity. On the other hand, the important thing in imbalanced data problem was the probability to find and call back the information for each class on label attribute, even though number of the class was very less than other classes. To measure the performance for each label, it was needed to measure the precision and the sensitivity of each class, because the accuracy is just total the right prediction of all classes in dataset. Thammasiri *et al.* [27] applied original SMOTE with KNN as the classification method, the accuracy of the study was increased but for the precision and the sensitivity of the study actually dropped considerably compared to the results of the precision and sensitivity before SMOTE. The improving SMOTE with attribute weighted SMOTE [29] has given better the increase of precision and sensitivity than the increase precision and sensitivity in other studies that applied original SMOTE [17], [30] and combined SMOTE with gaussian distribution [20]. The proposed method that handles the imbalanced data with the missing value elimination SMOTE had the highest increases of accuracy, precision and sensitivity even though the accuracy was not the highest compared with other studies. It concludes that the proposed method was significantly robust to use in classification with imbalanced data problem.

4. CONCLUSION

The SMOTE implementation that combines with handling the imbalanced data with missing value elimination for this study, shows that the proposed method can increase the accuracy, precision, and sensitivity for the classifications of the relevance of education background with graduates employement by using KNN, ANN, and C4.5. It describes that the proposed method not only improves the accuracy, but also improves the precision, and sensitivity on minority classes in the label attribute. The high precision and sensitivity of a class means that the model provides high probabilities to find and call back information on the class. Minority classes in a dataset can cause loss of information about the data in that class. In this study the problem of class imbalance in the dataset can be overcome by removing all data that has a missing value and doing SMOTE handling to overcome the imbalanced data. The outcome from the proposed method produces dataset that has balanced data for every class in label attribute. In addition, the precision and sensitivity values between minority classes and majority classes can be not much different. As result, the information in the minority class can still be obtained. Based on comparisons in several studies, the proposed method, which combined between handling the imbalanced data using missing value elimination and the SMOTE produces a robust method to increase or improve the accuracy, precision, and sensitivity of the classification on the imbalanced dataset.

REFERENCES

[1] 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, doi: 10.1016/j.procs.2015.07.372.

[2] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, “Educational data mining and learning analytics for 21st century higher education: A review and synthesis,” *Telemat. Informatics*, vol. 37, no. January, pp. 13–49, 2019, doi: 10.1016/j.tele.2019.01.007.

[3] S. Natek and M. Zwilling, “Student data mining solution-knowledge management system related to higher education institutions,” *Expert Syst. Appl.*, vol. 41, no. 14, pp. 6400–6407, 2014, doi: 10.1016/j.eswa.2014.04.024.

[4] H. Schomburg, *Handbook for Graduate Tracer Studies*, no. July. Bonn, 2003.

[5] P. N. Johnson *et al.*, “A survey of pediatric degree option program graduates in a doctor of pharmacy curriculum: Confidence and initial employment position,” *Curr. Pharm. Teach. Learn.*, vol. 11, no. 12, pp. 1296–1302, 2019, doi: 10.1016/j.cptl.2019.09.013.

[6] G. Haixiang, L. Yijing, L. Yanan, L. Xiao, and L. Jinling, “BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification,” *Eng. Appl. Artif. Intell.*, vol. 49, pp. 176–193, 2016, doi: 10.1016/j.engappai.2015.09.011.

[7] S. F. Crone and S. Finlay, “Instance sampling in credit scoring: An empirical study of sample size and balancing,” *Int. J. Forecast.*, vol. 28, no. 1, pp. 224–238, 2012, doi: 10.1016/j.ijforecast.2011.07.006.

[8] F. Louzada, P. H. Ferreira-Silva, and C. A. R. Diniz, “On the impact of disproportional samples in credit scoring models: An application to a Brazilian bank data,” *Expert Syst. Appl.*, vol. 39, no. 9, pp. 8071–8078, 2012, doi: 10.1016/j.eswa.2012.01.134.

[9] L. Zhang, C. Zhang, R. Gao, R. Yang, and Q. Song, “Using the SMOTE technique and hybrid features to predict the types of ion channel-targeted conotoxins,” *J. Theor. Biol.*, vol. 403, no. 17923, pp. 75–84, 2016, doi: 10.1016/j.jtbi.2016.04.034.

[10] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, no. Sept. 28, pp. 321–357, 2002, doi: 10.1613/jair.953.

[11] Y. Ünal, A. Sağlam, and O. Kayhan, “Improving classification performance for an imbalanced educational dataset example using SMOTE,” *Eur. J. Sci. Technol.*, vol. Special Is, no. October, pp. 485–489, 2019, doi:

10.31590/ejosat.638608.
[12] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, “Classification with class imbalance problem: A review,” *Int. J. Adv. Soft Comput. its Appl.*, vol. 5, no. 3, pp. 176–204, 2013.
[13] I. Ha, K.-J. Oh, M.-D. Hong, and G.-S. Jo, “Social Filtering Using Social Relationship for Movie Recommendation,” 2012, vol. 7653, pp. 395–404, doi: 10.1007/978-3-642-34630-9_41.
[14] M. Beckmann, N. F. F. Ebecken, and B. S. L. Pires de Lima, “A KNN Undersampling Approach for Data Balancing,” *J. Intell. Learn. Syst. Appl.*, vol. 07, no. 04, pp. 104–116, 2015, doi: 10.4236/jilsa.2015.74010.
[15] T. L. Octaviani, Z. Rustam, and T. Siswantining, “Ovarian Cancer Classification using Bayesian Logistic Regression,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 546, no. 5, 2019, doi: 10.1088/1757-899X/546/5/052049.
[16] B. Gong and J. Ordieres-Meré, “Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques: Case study of Hong Kong,” *Environ. Model. Softw.*, vol. 84, pp. 290–303, 2016, doi: 10.1016/j.envsoft.2016.06.020.
[17] M. Ashraf, M. Zaman, and M. Ahmed, “Using Ensemble StackingC Method and Base Classifiers to Ameliorate Prediction Accuracy of Pedagogical Data,” *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1021–1040, 2018, doi: 10.1016/j.procs.2018.05.018.
[18] V. Thi, N. Chau, and A. P. Definition, “Imbalanced educational data classification : an effective approach with resampling and random forest,” in *International Conference on Computing & Communication Technologies - Research, Innovation, and Vision for the Future (RIVF)*, 2013, pp. 135–140.
[19] J. Sun, J. Lang, H. Fujita, and H. Li, “Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates,” *Inf. Sci. (Ny).*, vol. 425, pp. 76–91, 2018, doi: 10.1016/j.ins.2017.10.017.
[20] T. Pan, J. Zhao, W. Wu, and J. Yang, “Learning imbalanced datasets based on SMOTE and Gaussian distribution,” *Inf. Sci. (Ny).*, vol. 512, pp. 1214–1233, 2020, doi: 10.1016/j.ins.2019.10.048.
[21] O. F. Ayilara, L. Zhang, T. T. Sajobi, R. Sawatzky, E. Bohm, and L. M. Lix, “Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry,” *Health Qual. Life Outcomes*, vol. 17, no. 1, pp. 1–9, 2019, doi: 10.1186/s12955-019-1181-2.
[22] P. Cerda and G. Varoquaux, “Encoding high-cardinality string categorical variables,” *CoRR*, vol. abs/1907.0, 2019, [Online]. Available: http://arxiv.org/abs/1907.01860.
[23] Z. E. Rasjid and R. Setiawan, “Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques,” *Procedia Comput. Sci.*, vol. 116, pp. 107–112, 2017, doi: 10.1016/j.procs.2017.10.017.
[24] R. Shankar, K. R. Balasubramanian, S. P. Sivapirakasam, and K. Ravikumar, “Materials Today : Proceedings ANN and RSM models approach for optimization of HVOF coating,” *Mater. Today Proc.*, no. xxxx, pp. 1–6, 2020, doi: 10.1016/j.matpr.2020.01.211.
[25] S. Yahdin, A. Desiani, A. Amran, and D. Rodiah, “Pattern recognation for study period of student in Mathematics Department with C4.5 algorithm data mining technique at the Faculty of Mathematics and Natural Science Universitas Sriwijaya,” in *Journal of Physics: Conference Series*, 2019, doi: 10.1088/1742-6596/1282/1/012014.
[26] R. Zacharski, “Evaluating algorithms and kNN,” in *A Programmer’s Guide to Data Mining*, 2015.
[27] D. Thammasiri, D. Delen, P. Meesad, and N. Kasap, “Expert Systems with Applications A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition,” *Expert Syst. Appl.*, no. August, 2014, doi: 10.1016/j.eswa.2013.07.046.
[28] W. Jia, H. Xia, L. Jia, Y. Deng, and X. Liu, “The selection of wart treatment method based on Synthetic Minority Over-sampling Technique and Axiomatic Fuzzy Set theory,” *Biocybern. Biomed. Eng.*, pp. 1–10, 2020, doi: 10.1016/j.bbe.2020.01.002.
[29] T. Fahrudin, J. L. Buliali, and C. Fatichah, “Enhancing the performance of SMOTE algorithm by using attribute weighting scheme and new selective sampling method for imbalanced data set,” *Int. J. Innov. Comput. Inf. Control*, vol. 15, no. 2, pp. 423–444, 2019, doi: 10.24507/ijicic.15.02.423.
[30] M. Anis and M. Ali, “Investigating the Performance of SMOTE for Class Imbalanced Learning: A Case Study of Credit Scoring Datasets,” *Eur. Sci. Journal, ESJ*, vol. 13, no. 33, p. 340, 2017, doi: 10.19044/esj.2017.v13n33p340.