ISSN 2714-7533

The Effect of Adaptive Synthetic and Information Gain on C4.5 and Naive Bayes in Imbalance Class Dataset

Mulia Sulistiyono [a], Lucky Adhikrisna Wirasakti [a], Yoga Pristyanto [b,1,*]

[a] Department of Informatics, Universitas Amikom Yogyakarta, Indonesia
[b] Department of Information System, Universitas Amikom Yogyakarta, Indonesia
[1] yoga.pristyanto@amikom.ac.id
* corresponding author

ARTICLE INFO

Article history
Received November 10, 2021
Revised December 7, 2021
Accepted January 8, 2022

Keywords
ADASYN
information gain
imbalanced class
feature selection
high dimensional dataset

ABSTRACT

Class imbalance is a severe problem in classification due to the deep slope on the class axis. The dataset is dominated by the majority class, which has the potential for misclassification. Another problem in classification and clustering is that high-dimensional datasets are found that have the potential to affect the performance of classification algorithms in terms of computation and accuracy. In this study, the class imbalance was handled using the ADASYN k - NN resampling technique and the selection feature using Information Gain. Based on the evaluation results, the sampling contribution matrix can improve the classification model by improving the geometric mean value. The selection feature helps interpret data with more simple features but can reduce the accuracy of the results. The results showed that the implementation of ADASYN k-NN and Information Gain could increase the accuracy score and geometric mean score of Decision Tree C4.5 and Naive Bayes. For further work, this proposed method will be tested on multiclass imbalanced datasets.

1. Introduction

The classification method is one of the popular methods used in machine learning. Classification is a method for arranging data systematically or according to predetermined rules or rules. A rule or rule is obtained from studying a data set [1], [2]. The class distribution conditions in the dataset are crucial in applying the classification algorithm to build a successful model. The presence of a class imbalance in a dataset is frequently overlooked by machine learning researchers, particularly in using the classification approach. [3]. We need a methodology or strategy for resolving the class imbalance problem in the dataset.

Based on the research that has been done regarding the handling of class imbalances in the dataset, one approach that can be used is the approach at the data level. The data level approach is usually carried out at the pre-data processing stage by changing or correcting the class distribution bias contained in the dataset. The method that is often used in the data-level approach is to apply resampling and data synthesis techniques. In the following, studies have been carried out regarding the handling of class imbalances, in some of these studies using data-level approaches as a solution. Gongzhu Hu [4], and Jishan [5] performed a study on addressing class imbalance on the performance of the classification algorithm. Both research' findings suggest that using class imbalance management can improve the classification algorithm's performance. Both employed the SMOTE (Synthetic Minority Over-Sampling Technique) technique in their study. Furthermore, Imran [6]

compared two oversampling approaches, namely SMOTE (Synthetic Minority Over-Sampling Technique) and ROS (Random Over Sampling). The findings of these experiments indicate that both can increase the classification algorithm's performance. While Rashu [7] and Thammasiri [8] implemented one of the under sampling techniques, RUS (Random Under Sampling), the results of their research indicated that the RUS technique reduced the classification algorithm's performance..

Also, the class imbalance condition in the dataset is a challenge for researchers and practitioners in the field of machine learning. In addition to the existence of class imbalance, datasets that have high dimensions are often found, this is marked by a large number of features, of course, it will have an influence on the process of implementing machine learning itself both classifications, clustering and prediction, the problem that often arises is the performance of the classification algorithm both in terms of computation time and accuracy as a result of several attributes features that have no relevance to attribute class [9]. Therefore, the feature selection process is needed to deal with high dimensional problems in the dataset. Several study on feature selection have been conducted. Several approaches, including information gain ratio (Gain Ratio) [10], information gain (IG) [11], and correlation-based feature selection (CBFS) [12] are often applied in feature selection. The following are studies that have been conducted regarding the application of feature selection to increase the accuracy value of the classification algorithm. Such as research conducted by [11], dan [10]. They proved that applying information gain to implement feature selection can improve the classification algorithm's accuracy. Meanwhile, [13] dan [14] showed that using the correlation-based feature selection (CBFS) technique can improve the classification algorithm's accuracy. In a study conducted by [10] applying the Gain Ratio method to the feature selection process, the results showed an increase in the accuracy value of the classification algorithm used. Of the three algorithms that on average can improve classifier performance is IG (Information Gain).

Based on the description above, the majority of the resampling technique used is the SMOTE algorithm, but this algorithm has a weakness, namely, when doing the data synthesis process, it sometimes causes class overlapping. Therefore, in this study, the ADASYN (Adaptive Synthetic) algorithm will be used as a resampling method to deal with class imbalances in the dataset. The ADASYN algorithm was chosen because it was able to complement the shortcomings of the SMOTE algorithm in terms of the synthesis data replication process. Meanwhile, for the selected feature process, the IG (Information Gain) method will be used. The following are the contributions to this research firstly the proposed resampling method can improve the performance of the classification algorithm. Secondly, the proposed method can be a solution for dealing with class imbalances in the dataset. Last, the proposed method can be a solution to find out whether the feature selection process can always increase the classification algorithm accuracy value or not

2. Method

At this stage the dataset for classification is applied, the methodology and tools are described. The methods used are resampling, data transformation, and feature selection.

A. Materials

In this study, secondary data was used in the form of numerical data in the CSV extension sourced from the KEEL repositories (http://tiny.cc/imbalanced-dataset) which consisted of:

a. Ecoli Datasets with the Imbalance Ratio (IR) is 8.6
b. Yeast Ml8 with the Imbalance Ratio (IR) 13
c. Libras Move with the Imbalance Ratio (IR) 14
d. Wine Quality with the Imbalance Ratio (IR) 26

e. Mammography with the Imbalance Ratio (IR) 42

B. Method

Figure 1. The description of each step is presented as follows. All fundamental theories are explained briefly in every sub section. The method is based on fundamental theories and is designed to meet the purpose of the research.

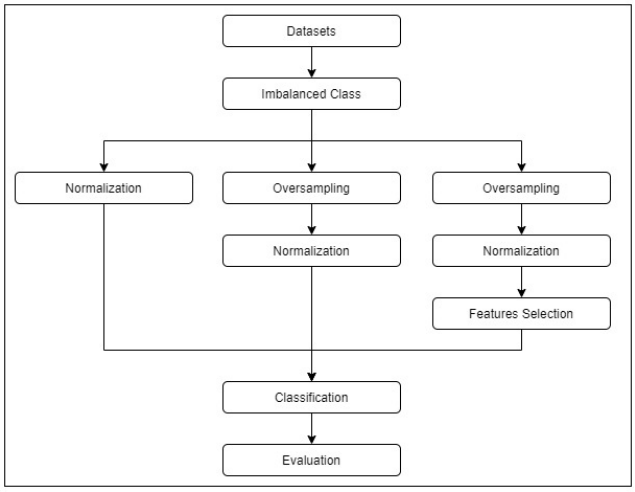

Figure 1. Research Workflow

1) Data Normalization

Normalization is the initial stage that aims to transform from Raw Data (Real World Data) to a form that can be processed, a common problem that often arises in a dataset of conditions where the scale values of each attribute are not aligned, normalization is a process of scaling the attributes so that they have a range the specified value, the range of values commonly used (-1.1) and (0.0-1.0) so that each attribute has the same weight normalization of the input value will accelerate the learning phase, in a distance-based method will help the initial range difference (for example: salary against age) [15]. In this study using the min-max method to normalize. Min-Max performs a linear transformation of the original data to balance comparison values between data before and after processing it becomes balanced or standardized. Here is the equation for performing the min-max normalization.

$$V'i = (\frac{Vi-minA}{maxA-minB})(new\_maxA - new\_minA) + new\_minA \quad (1)$$

where,

V'i: the new value of each entry in the data.

Vi: the old value of each entry in the data

minA: minimum value in data.

maxA: the maximum value in the data.

new_maxA: the new maximum value from the data range.

new_minA: the new minimum value of the data range.

2) Oversampling using ADASYN

Adaptive Synthetic (ADASYN) was introduced by [16], moving on from the success of sampling approaches like SMOTE, SMOTEBoost, and DataBoostIM in overcoming learning from category imbalances by completing the most plan exploitation distribution weight. The ADASYN approach enhances learning regarding information distribution in two ways: reducing bias caused by category imbalances and reconciling shifts the boundary of classification selections towards challenging examples. Figure 2. following is ADASYN Algorithm.

Input
Training data set Dtr with m samples {xi, yi}, i = 1, ..., m, where xi is an instance in the n dimensional feature space X and yi ∈ Y = {1, −1} is the class identity label associated with xi. Define ms and ml as the number of minority class examples and the number of majority class examples, respectively. Therefore, ms ≤ ml and ms + ml = m.
Procedure
1. Calculate the degree of class imbalance
$$d = ms/ml$$
where d ∈ [0, 1]
2. If d<dth then (dth is a preset threshold for the maximum tolerated degree of class imbalance ratio):
a. Calculate the number of synthetic data examples that need to be generated for the minority class:
$$G = (ml - ms) * \beta$$
Where β ∈ [0, 1] is a parameter used to specify the desired balance level after generation of the synthetic data. β = 1 means a fully balanced data set is created after the generalization process.
b. For each example xi ∈ minority class, find K nearest neighbors based on the Euclidean distance in n dimensional space, and calculate the ratio ri defined as:
ri = Δi/K, i = 1,...,ms
where Δi is the number of examples in the K nearest neighbors of xi that belong to the majority class, therefore ri ∈ [0, 1];
c. Normalize ri according to r^i = ri/ms i=1 ri, so that r^i is a density distribution.
d. Calculate the number of synthetic data examples that need to be generated for each minority example xi:
$$gi = \hat{}rixG$$
where G is the total number of synthetic data examples that need to be generated for the minority class as defined in Equation in point (2) (a).
e. For each minority class data example xi, generate gi synthetic data examples according to the following steps:
Do the **Loop** from 1 to gi:
i. Randomly choose one minority data example, xzi, from the K nearest neighbors for data xi.
ii. Generate the synthetic data example:
$$si = xi + (xzi - xi) \times \lambda$$
where (xzi − xi) is the difference vector in n dimensional spaces, and λ is a random number: λ ∈ [0, 1].
End Loop
Output Synthetic data from minority class data

Figure 2. ADASYN Pseudocode [16]

3) Information Gain Feature Selection

Information Gain relies on previous analysis by Claude E. Shannon on scientific theory. That studies the worth or data content of the message. The attribute with the best data retrieval is hand-picked because of the apparatus attribute. This attribute minimizes the data required to classify tuples within the next partition. This sort of approach minimizes the number of tests needed to classify a given tuple. The stages in calculating information gain are as follows. Calculate the entropy value for each feature as follows:

- Determine the value of information gain for each attribute in the original dataset.

- Establish the desired threshold. This step allows attributes with weights equal to or greater than the limit to be retained while discarding attributes with weights more minor than the limit.
- The dataset is improved by removing attributes based on information gain value ranking.
- Determine the threshold (threshold). Features with a weight greater than equal to the threshold will be retained. Dataset with features selected by weight.

4) Naïve Bayes

The Bayes theorem is performed with high independent assumptions in Nave Bayes, a primary probabilistic classifier. The general process phases for the Naive Bayes algorithm are as follows:

- Determine the number of classes/labels.
- Count the number of instances in each class.
- Multiply all of the class variables.
- Compare the outcomes of each class.

5) Decision Tree C4.5

C4.5 was chosen as the decision tree method for this investigation since it is well-known and often used in classification approaches. The steps for computing the C4.5 algorithm are as follows. Create a branch for each value in the root node, then choose an attribute as the root. The case is divided into branches. Then, repeat the procedure for each branch until all cases in that branch have the same result. The selection of attributes with roots is based on the highest acquisition value of existing attributes.

6) Evaluation

The evaluation method is used to test the performance of the generated model. The confusion matrix, often known as the error matrix, is the most widely used evaluation methodology. The error matrix is a special table arrangement that allows an algorithm to be shown [17]. A confusion matrix table is shown in Table I.

Table 1. Confusion Matrix

Actual Class	**Predicted Class**	
	Positive	**Negative**
Positive	True Positive	False Positive
Negative	False Negative	True Negative

Accuracy score in classification is the value of accuracy of data records that are classified correctly after testing the classification results.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Geometric mean score (g-mean) is one of the performance evaluation indicators whose value comes from the square root of the multiplication of the true positive rate and true negative rate.

$$G - Mean = \sqrt{TPR\ x\ TNR} \quad (5)$$

3. Result and discussion

In this section, the result will be presented in tables and figures and will be explained in four sections namely normalization, resampling, features selection and evaluation results.

A. Normalization

The dataset used in the study is not in normal conditions, each feature has a range of values by the range of values to be interpreted, so a normalization process is required using Min-Max Normalization so that there is harmony between the range of values for each feature, according to [15] the value range (0) – (1) is a general range commonly used in research. The process of implementing the Min-Max Normalization on the dataset by carrying out a linear transformation of the original dataset into a new dataset is based on the value factor min = 0 and max = 1. Figure 3. The following is an illustration of the Min-Max Normalization implementation process.

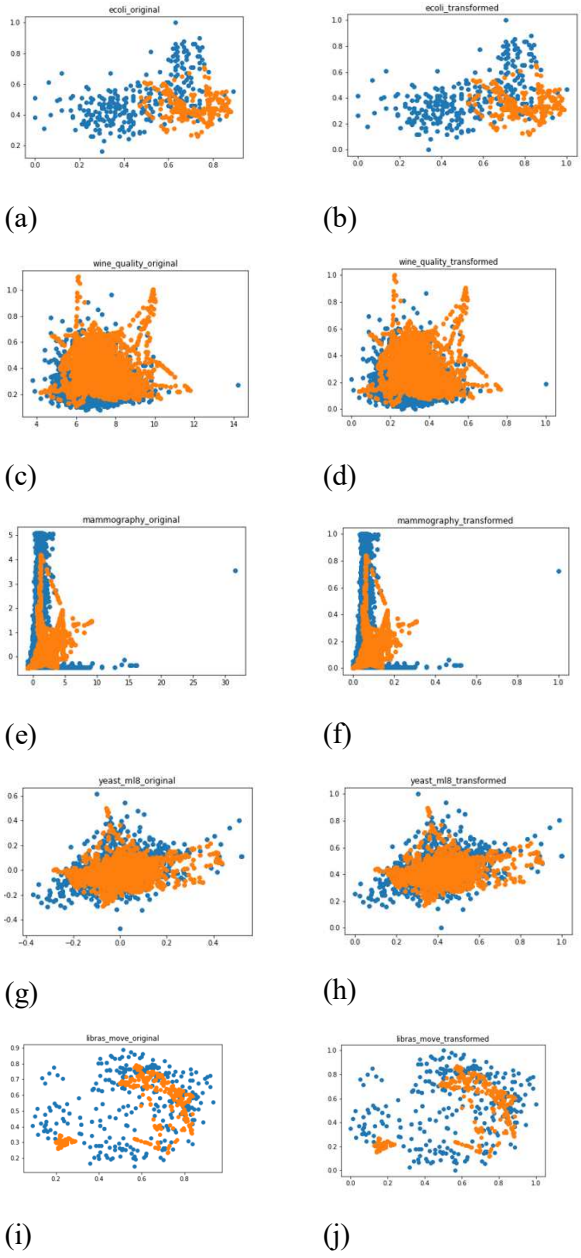

Figure 3. Illustration of the results of the normalization process

(a) ecoli original set, (b) ecoli normalized, (c) wine original set, (d) wine normalized, (e) mammography original set, (f) mammography normalized, (g) yeast_ml8 original set, (h) yeast_ml8 normalized, (i) libras_move original set, (j) libras_move normalized

B. Resampling using ADASYN

At this stage, the ADASYN algorithm is implemented to handle the unbalanced class distribution in the dataset. For the balanced class distribution in this study, the value of beta = 1 and k = 5 are used in each test. Below, figure 4 illustrates before and after synthesis with ADASYN on each dataset used in this study.

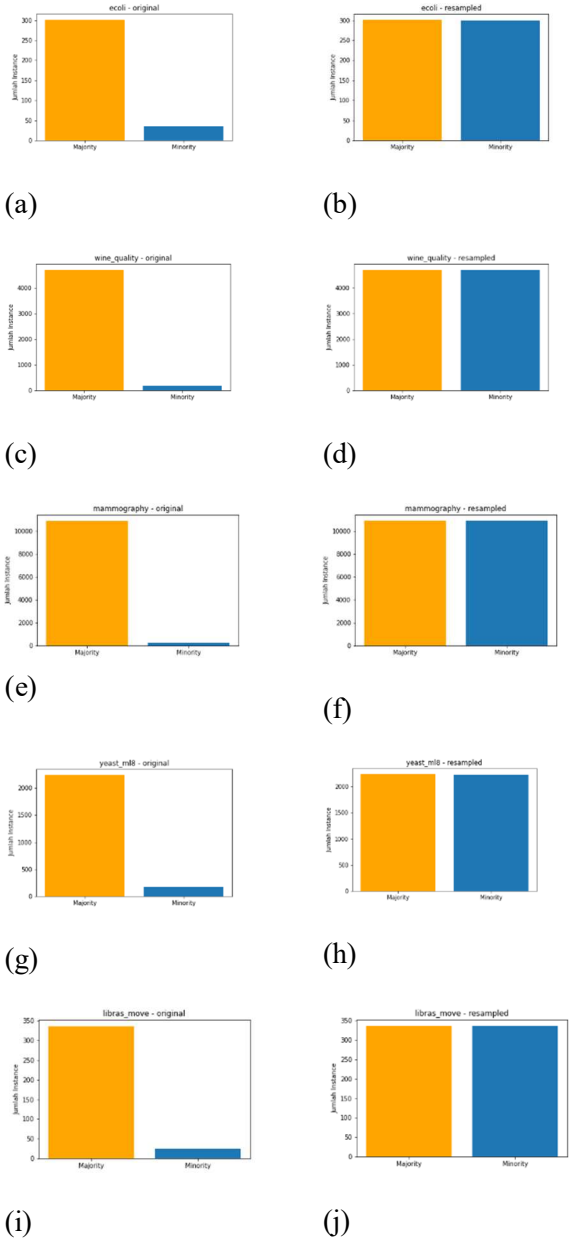

Figure 4. Illustration of the results of the normalization process

(a) ecoli original set, (b) ecoli resampled, (c) wine original set, (d) wine resampled, (e) mammography original set, (f) mammography resampled, (g) yeast_ml8 original set, (h) yeast_ml8 resampled, (i) libras_move original set, (j) libras_move resampled

C. Information Gain Feature Selection

Before the training process is carried out, the dataset will be ranked using the scoring method (Information Gain), so that only the correlated features will be maintained, this is very dependent on the threshold value, in this study the threshold value (threshold = 0.8) is used. The process of implementing Information Gain on the dataset is to calculate the entropy value and the difference in entropy values for each feature, then rank the resulting gain values. Table II is the process of implementing feature selection using information gain on the wine quality, mammography, and ecoli dataset.

Table 2. Features Selection Process of wine quality dataset

Wine Quality Dataset		
Descriptive Feature	**Information Gain**	**Ranking**
7	0.895	1
3	0.831	2
2	0.825	3
6	0.819	4
8	0.814	5
1	0.813	6
9	0.807	-
4	0.805	-
10	0.799	-
0	0.787	-
5	0.778	-

Table 3. Features Selection Process of mammography dataset

Mammography Dataset		
Descriptive Feature	**Information Gain**	**Ranking**
0	0.883	1
1	0.806	2
3	0.713	3
2	0.696	4
5	0.671	5
4	0.471	6
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-

Table 4. Features Selection Process of ecoli dataset

Ecoli Dataset		
Descriptive Feature	**Information Gain**	**Ranking**
0	0.876	1
6	0.855	2
5	0.838	3
4	0.78	4
1	0.751	5
2	0.037	6
3	0.002	7
-	-	-

Ecoli Dataset		
Descriptive Feature	**Information Gain**	**Ranking**
-	-	-
-	-	-
-	-	-

Based on table II, III and IV above, the features taken in the Ecoli dataset have 3 features, namely 7, 3 and 2. In the wine quality dataset, the features taken are 8 features, namely 7, 3, 2, 6, 8, 1, 9 and 4. Meanwhile, In the mammography dataset, there are 2 features taken, namely 0 and 1.

The yeast ml8 dataset consists of 103 features with the highest gain value = 1 and the lowest gain value = 0.999 with an average gain value = 0.999971, so based on the threshold value all features in this dataset are maintained. Whereas the Libras Move dataset consists of 90 features with the highest gain value = 0.982 and the lowest gain value = 0.918 with an average gain value = 0.483, so based on the threshold value all features in this dataset are retained.

D. Evaluation Results

The dataset in this study is separated into two parts: training data and testing data. The dataset is divided into two parts: 80 percent is utilized for training the model, and 20 percent is utilized for model validation or model testing. In addition, there are three scenarios of the testing method in this study. The first example involves simply the data normalization procedure, with no resampling or selection characteristics. The second scenario is the testing procedure without a feature selection method, which employs data normalization and resampling. Data normalization, resampling, and feature selection are all used in the three-scenario testing method. While the Decision Tree C4.5 and Nave Bayes classification algorithms are utilized, the accuracy and geometric mean metrics are used to evaluate the performance of each scenario. The results of the performance evaluation of each scenario using the Decision Tree C4.5 algorithm and Nave Bayes are shown in table V and table VI.

Table 5. Results of Performance Evaluation of Each Scenario in the Naïve Bayes Algorithm

Datasets		**Naïve Bayes**		
		First Scenario	**Second Scenario**	**Third Scenario**
Ecoli	**Accuracy**	0,882353	0,842975	0,900826
	G-Mean	0,736788	0,843047	0,899735
Libras move	**Accuracy**	0,944444	0,851852	0,851852
	G-Mean	0,696631	0,840323	0,840323
Mammography	**Accuracy**	0,948145	0,770481	0,554691
	G-Mean	0,860329	0,770112	0,387627
Wine Quality	**Accuracy**	0,956122	0,732095	0,697082
	G-Mean	0,532821	0,728741	0,688963
Yeast Ml8	**Accuracy**	0,739669	0,725644	0,725644
	G-Mean	0,556187	0,726526	0,726481

Table 6. Results of Performance Evaluation of Each Scenario in the Decision Tree C 4.5 Algorithm

Datasets		Decision Tree C4.5		
		First Scenario	Second Scenario	Third Scenario
Ecoli	Accuracy	0,941176	0,884298	0,942149
	G-Mean	0,762202	0,880364	0,939319
Libras move	Accuracy	0,958333	0,992593	0,985185
	G-Mean	0,701888	0,993127	0,986206
Mammography	Accuracy	0,982566	0,949886	0,881922
	G-Mean	0,727292	0,949251	0,880930
Wine Quality	Accuracy	0,956122	0,937931	0,922016
	G-Mean	0,561340	0,937913	0,921882
Yeast Ml8	Accuracy	0,878099	0,853303	0,849944
	G-Mean	0,261846	0,854461	0,851064

Table V and table VI above shows that after testing the Decision Tree C4.5 algorithm in the second and third scenarios, it always produces a better geometric mean value than the first scenario in the entire dataset used. These results show that the resampling process has an essential role in dealing with class imbalances in the datasets, especially when tested using the C4.5 Decision Tree algorithm. Meanwhile, in terms of accuracy, the second scenario and third scenario can only increase in the Ecoli and Libras Move dataset. Furthermore, after testing the Naïve Bayes algorithm second scenario and third scenario result in better geometric mean values than the first scenario in the Ecoli and Libras Move datasets. Meanwhile, the first scenario is still better in terms of the accuracy value than the second scenario and third scenario. This result can happen because the characteristics of the Naïve Bayes algorithm tend to have less influence on the class distribution in the datasets.

4. Conclusion

Based on the results of the research that has been done, several points can be concluded that. First, the Decision Tree algorithm by handling class imbalances with ADASYN resampling can improve classification skills in training and testing and the use of the selection feature has the potential to reduce the value of the evaluation results for the dataset Libras move, mammography, wine quality, and yeast ml8, inversely proportional to the implementation of the selected feature in the ecoli dataset. can increase the accuracy and geometric mean values, this condition depends on the characteristics of the data and the presence of noise samples in each dataset. Second, the implementation of ADASYN resampling on the Naïve Bayes Algorithm can increase the geometric mean value and the selection feature tends to decrease the performance of the geometric mean value. Based on the results of the evaluation of the Naïve Bayes algorithm, it achieves the best ability in the second scenario, namely resampling without selection features. Third, handling the imbalance dataset using the ADASYN oversampling resampling method in this study does not always provide an increase in the accuracy value, but there is an improvement in the geometric mean value both in training and testing so that the minority class is easier to recognize even with a lower accuracy value. Fourth, the contribution of information gain makes data interpretation easier for researchers because the features presented are simpler, another advantage provides an increase in computational acceleration. Finally, ADASYN resampling has poor performance in synthesizing datasets with the involvement of noise samples in the dataset, because the noise will also be duplicated in the process of synthesizing new data, in this study, the direct impact that can be seen is a decrease in accuracy results from before and after the synthesis process.

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] S. Datta and A. Arputharaj, "An Analysis of Several Machine Learning Algorithms for Imbalanced Classes," *5th Int. Conf. Soft Comput. Mach. Intell. ISCMI 2018*, pp. 22–27, 2018, doi: 10.1109/ISCMI.2018.8703244.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Y. Pristyanto, S. Adi, and A. Sunyoto, "The effect of feature selection on classification algorithms in credit approval," *2019 Int. Conf. Inf. Commun. Technol. ICOIACT 2019*, pp. 451–456, 2019, doi: 10.1109/ICOIACT46704.2019.8938523.

[10] R. S. Ramya and S. Kumaresan, "Analysis of feature selection techniques in credit risk assessment," in *ICACCS 2015 - Proceedings of the 2nd International Conference on Advanced Computing and Communication Systems*, 2015, pp. 1–6, doi: 10.1109/ICACCS.2015.7324139.

[11] W. Punlumjeak and N. Rachburee, "A Comparative Study of Feature Selection Techniques for Classify Student Performance," in *2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2015, pp. 425–429, doi: 10.1109/ICMLA.2010.27.

[12] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[13] Y. Khaokaew and T. Anusas-Amornkul, "A performance comparison of feature selection techniques with SVM for network anomaly detection," in *2016 4th International Symposium on Computational and Business Intelligence, ISCBI 2016*, 2016, pp. 85–89, doi: 10.1109/ISCBI.2016.7743263.

[14] K. R. Pushpalatha and A. G. Karegowda, "CFS Based Feature Subset Selection for Enhancing Classification of Similar Looking Food Grains-A Filter Approach," in *2017 2nd International Conference On Emerging Computation and Information Technologies, ICECIT 2017*, 2018, pp. 1–6, doi: 10.1109/ICECIT.2017.8453403.

[15] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[16] 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.

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