JOIV : Int. J. Inform. Visualization, 7(2) - June 2023 542-547

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

k-Means Cluster-based Random Undersampling and Meta-Learning Approach for Village Development Status Classification

Ahmad Ilham [a,*], Luqman Assaffat [a], Laelatul Khikmah [b], Safuan [a], Suprapedi [c,d]

[a] Department of Informatics, Faculty of Engineering, University Muhammadiyah Semarang, Semarang, 50273, Indonesia
[b] Department of Statistics, Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang, Semarang, 50185, Indonesia
[c] National Research and Innovation Agency (BRIN), South Jakarta, 12710, Indonesia
[d] Graduate School, Doctoral Program of Environmental Studies, Brawijaya University, Malang, 65145, Indonesia
Corresponding author: *ahmadilham@unimus.ac.id

***Abstract*— There is a significant imbalanced class in the village development index (called IDM - *Indeks Desa Membangun*) dataset, marked by the number of self-supporting classes more than the disadvantaged class. The traditional classifiers are able to achieve high accuracy (ACC) by training all cases of the majority class but forsaking the minority class, so that possible for the classification results to be biased. In this study, a random under-sampling technique was employed based on k-means cluster (KMC) and a meta-learning approach to improving ACC of the village status classification model. Furthermore, the AdaBoost and Random Forest were used as meta technique and base learner, respectively. The proposed model has been evaluated using the area under the curve (AUC), and experimental results showed that it yielded excellent performance compared to the prior studies with the AUC, ACC, precision (PR), recall (RC), and g-mean (Gm) values of 95.50%, 95.52%, 95.5%, 95.5%, and 92.95%, respectively. Similarly, the result of the t-test also showed the proposed model yielded excellent performance compared to previous studies. It can be concluded that the AdaBoost algorithm improved misclassification and changed the distribution of data loss function in random forests. It indicates that the proposed model effectively deals with imbalanced classes in the village development status classification model.**

***Keywords*—Village development index; village development status classification; imbalanced class; meta-learning; random forest.**

Manuscript received 28 Jun. 2022; revised 30 Oct. 2022; accepted 14 Nov. 2022. Date of publication 30 Jun. 2023.

I. INTRODUCTION

The Ministry of Village, Development of Disadvantaged Regions, and Transmigration Republic of Indonesia (KEMENDESA), the Ministry of National Development Planning of the Republic of Indonesia (BAPPENAS), and the Central Bureau of Statistics Indonesia (BPS) developed a system that provides information regarding village development status. This information is compiled as a unit of analysis based on the village development index in Indonesia in line with law No. 6 of 2014. Furthermore, the information is utilized to formulate and summarize village development policies and oversight plans.

In recent decades, classification models have been used to develop policies for different functions based on class classification analysis in different fields [1]–[7]. This process generally begins with pre-processing, which deals with identifying potential problems. According to Han and Kamber [8], missing data, outliers, and imbalanced classes often provoke bias in classification results.

An imbalanced class has been identified in the IDM dataset, affecting the model's performance. Imbalanced class distribution in a dataset has caused severe difficulties for most base classifier models because they assume all data have a balanced class distribution [9], [10]. Due to two characterization classes, one is represented by a big sample and the other by a relatively tiny sample. According to Sun et al. [11], base classifier learning performs poorly on imbalanced datasets because they are designed to generalize from training data, and the results of the most straightforward hypotheses best fit the data. Besides, they assume all data have a balanced class distribution [3].

Several studies were conducted to identify the best machine learning models for determining the classification of the village development status in Indonesia, including k-prototype [12], support vector machine (SVM) [13], bootstrap sampling k-nearest neighbors (BS-KNN) [14], and decision tree (DT) [15]. Presently, performance classification of the village development status model has been the focus of further studies since the best performance of all evaluations

has not been fully achieved, and no one has been able to reconcile imbalanced class data from those utilized. The distribution of IDM data can be seen in Fig. 1.

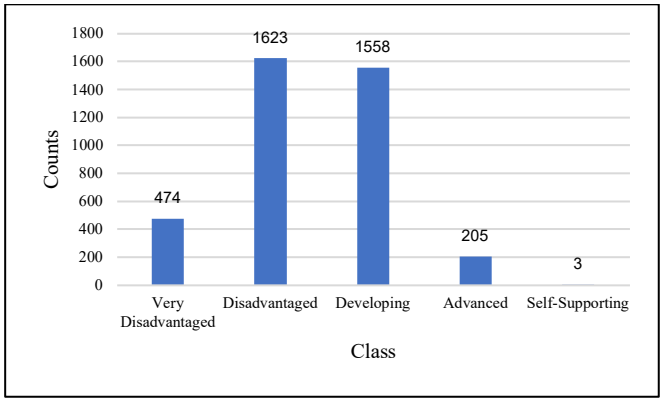

Fig. 1 Class distribution of the IDM dataset

As shown in Fig. 1, the distribution of self-supporting, advanced, and very disadvantaged classes is smaller than developing, and disadvantaged classes are identified as an imbalanced class of data problem. Therefore, a combination of random undersampling based on k-means cluster (RUS-KMC) and meta-learning (RFM) was proposed to improve the accuracy (ACC) of the village development status classification model. RUS-KMC was employed to handle imbalanced classes, while RFM was used to enhance the classifier's performance. It is important to note that RUS was selected because many previous studies reported that this method is often used to tackle the imbalanced class. Besides, KMC selected as a cluster model cause able to handle mixed-type attributes and big data sets, as well as automatically determine the clusters' ideal number and attributes that are not normally distributed [16].

II. Material and Method

A. Dataset

IDM dataset obtained from Ministry of Village, Development of Disadvantaged Regions, and Transmigration of Indonesia in 2016. The dataset includes the potential village information (PODES – *Potensi Desa*) from 16 provinces formed in three main dimensions, Social Resilience Index (SRI), Economic Resilience Index (EcRI), and Village Ecological Resilience Index (EnRI). There are 3863 villages observed in the dataset, 62 attributes, and 5 classes, including very disadvantaged, disadvantaged, developing, advanced, and self-supporting, where each attribute has a score of one to five, which indicates a score of one is very disadvantaged, and score five is self-supporting. Data types for all attributes are numeric and categorical; for the type of data, classes are categorical, as shown in Table I.

TABLE I
THE IDM DATASET INFORMATION

Dimensions	Attributes	Descriptions,	Types
Social (SRI)	a1	Doctor Scores	Numeric
	a2	Village midwife	Numeric
	a3	Another nurse	Numeric
	a4	Health facilities access	Categorical
	a5	Village health clinics (Poskesdes/Polindes) access	Categorical
	a6	Integrated healthcare center (Posyandu) scores	Numeric
	a7	BPJS membership score	Categorical
	a8	Primary school access	Categorical
	a9	Junior high school access	Categorical
	a10	Senior high school access	Categorical
	a11	PKBM	Categorical
	a12	PAUD	Categorical
	a13	Library	Categorical
	a14	Courses	Categorical
	a15	Communication diversity	Categorical
	a16	Variety of languages	Numeric
	a17	Religious diversity score	Numeric
	a18	Mutual assistance	Categorical
	a19	Mutual assistance frequency	Numeric
	a20	Public area	Categorical
	a21	JML FASI OR	Numeric
	a22	Event OR	Numeric
	a23	Score PMKS	Numeric
	a24	SLB access	Categorical
	a25	Security posts	Categorical
	a26	Citizen's Environmental Security System	Categorical
	a27	Conflict score	Categorical
	a28	Mineral water	Categorical
	a29	Latrine access	Categorical
	a30	Garbage	Categorical
	a31	Washing bath	Categorical
	a32	Electricity score	Numeric
	a33	Signal score	Numeric
	a34	Internet score	Numeric

Dimensions	Attributes	Descriptions,	Types
	a35	Citizen's internet access	Categorical
Economy (EcRI)	b1	Production diversity score	Numeric
	b2	Economy score	Numeric
	b3	Grocery store score	Numeric
	b4	Shop & lodging score	Numeric
	b5	Shop score	Numeric
	b6	Market score	Numeric
	b7	Road quality score	Numeric
	b8	Region openness score	Numeric
	b9	Mode general trans score	Numeric
	b10	Postal and logistics services score	Numeric
	b11	Credit fast score	Numeric
	b12	Bank BPR score	Numeric
Ecology/	c1	Water pollution	Categorical
Environmental	c2	Soil pollution	Categorical
(EnRI)	c3	Air pollution	Categorical
	c4	River waste pollution	Categorical
	c5	Pollution score	Numeric
	c6	Avalanche	Categorical
	c7	Flood	Categorical
	c8	Forest fires	Categorical
	c9	Disaster score	Categorical
	c10	Early warning	Categorical
	c11	Tsunami early warning	Numeric
	c12	Safety equipment	Categorical
	c13	Evacuation route	Categorical
	c14	Disaster response score	Categorical
	Class	Very disadvantaged, Disadvantaged, Developing, Advanced, Self-supporting	Categorical
The total number of attributes is **62**			

Several government regulations determine the existence of IDM dataset, 1) presidential decree (PERPRES) No. 2 of 2015 concerning the national medium-term development plan (called RPJMN); 2) village government regulations related to the development of disadvantaged areas and transmigration (PERMENDESA PDTT) No. 2 of 2016 concerning the village development index; and 3) Decision Letter of Director General of PPMD No. 30, 2016 concerning the village development status.

B. General Step

The experiments are conducted on a computer platform with the following specifications: Intel HD Graphics 4000 1536 MB, 2.5 GHz Dual-Core Intel Core i5, 8 GB RAM, and macOS Cataline Version 10.15.7 64-bit operating system, as well as the data analytics program Weka version 3.8.5. Weka will produce an AUC and a confusion matrix as computation outputs, and IBM SPSS Statistics will produce a t-test for a statistical comparison between the proposed model and prior studies.

We proposed a model called RUS-KMC+RFM, a random undersampling (RUS). It is based on the k-means cluster (KMC) and hybrid meta-learning technique (RFM) to tackle imbalanced class problems for high accuracy in the village development status classification model, as shown in Fig. 2. The KMC is a clustering technique that produces clusters of relatively uniform sizes created by Kumar et al. [16] and designed to handle very large datasets. AdaBoost was used in the meta technique, while random forest (RF) was used as the base learner. Furthermore, meta AdaBoost was employed to tackle imbalanced classes to improve RF classification performance. The model is utilized to assign different weights to misclassified samples and reduce weights correctly classified, effectively changing the data training distribution [7]. The proposed model was evaluated using the IDM dataset.

As shown in Fig. 2, the dataset was fed into the training and testing phase. The training phase was used to build the model, while the testing phase deals with testing and performance evaluation. In the pre-processing step, the KMC technique was employed in the training phase to group the dataset and its number was set to 5 to create a 5-binning or 5-quartile. The clusters are conducted randomly until the number of members in the majority, and minority classes are equal for each cluster. Consequently, those with the same proportion for each class are combined to create a new dataset.

The new dataset is later fed into a hybrid technique with a 10-fold cross-validation approach and was divided into ten pieces, in which nine serve as a training dataset, while the other one is for testing. AdaBoost and RF were employed in hybrid strategy as meta and based learners, respectively. After completing the learning process, the model is fed with test data in the testing phase, and assessment results are recorded.

This study used the area under the curve (AUC) to evaluate the proposed model. Furthermore, it is a numerical measure of differentiating the model's performance and its effectiveness in distinguishing between positive and negative observations. According to Xue and Hall [17], AUC greatly improve convergence across empirical studies in imbalanced class problems, and it is a single-measure classifier

performance that is useful for determining whether the model performs better. The general rule for categorizing ACC for the diagnostic test based on AUC, also reported by Gautheron et al. [18], divides five categories, excellent, good, fair, poor, and failure, with the respective range of 90% to 100%, 80% to 90%, 70% to 80%, 60% to 70%, and 50% to 60%. The AUC is calculated star from 1 to 8.

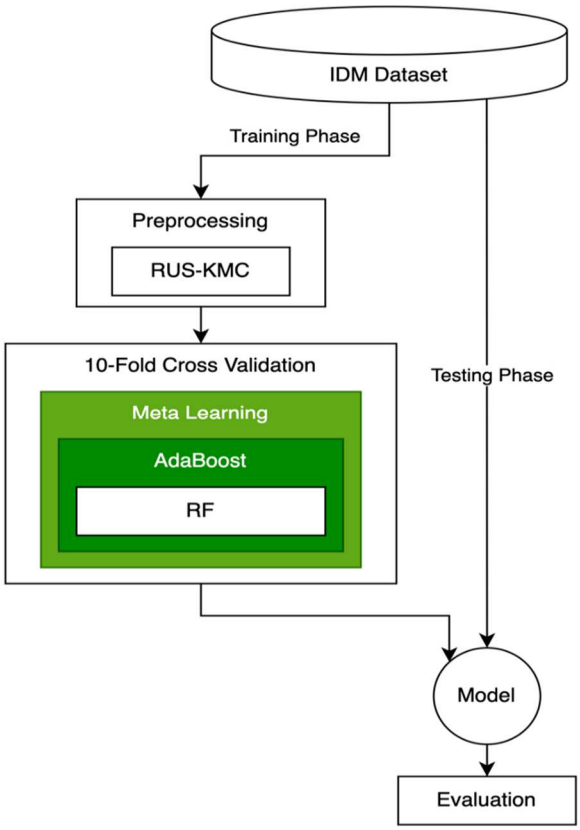

Fig. 2 Block diagram of the proposed model

According to Haixiang et al. [19] and Ri and Kim [20], recall (RC), precision (PR), and g-mean (Gm) are comprehensive predictor evaluations in an imbalanced class problem. These assessments are based on the confusion matrix with values of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), as shown in Table II. When both actual and predicted classes are in error, it is called TP. When the predicted class is flawed, but the actual class is not, this condition is called FP. It is important to note that in a non-faulty class, TN and TP are equivalent, but if it is defective, FN occurs. The calculation was conducted under the confusion matrix generated by the model.

TABLE II
CONFUSION MATRIX INTERPRETATION

		Predicted class	
		Positive	**Negative**
Actual class	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

$$Sensitivity = Recall = \frac{TP}{TP + FN} \quad (1)$$

$$Specificity = TN_{rate} = \frac{TN}{TN + FP} \quad (2)$$

$$FP_{rate} = \frac{FP}{FP + TN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (5)$$

$$G - mean = \sqrt{Sencitifity \times Specificity} \quad (6)$$

$$Error\ Classification = \frac{FP + FN}{TP + TN + FP + FN} \quad (7)$$

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \quad (8)$$

III. RESULTS AND DISCUSSION

This experiment was conducted on a computer platform with the following specifications, Intel HD Graphics 4000 1536 MB, 2.5 GHz Dual-Core Intel Core i5, 8 GB RAM, and macOS Cataline Version 10.15.7 64-bit operating system, as well as the data analytics program Weka version 3.8.5. The program produced an AUC and a confusion matrix as computational outputs, and IBM SPSS Statistics generated a t-test for a statistical comparison between the proposed model and prior studies.

TABLE III
RESULTS COMPARISON PERFORMANCE FOR RANDOM FOREST (**RF**) VS META STRATEGY (**RFM**) ONLY TECHNIQUE OF IDM DATASET

Model	**(%)**					
	AUC	**ACC**	**PR**	**RC**	**Gm**	**EC**
RF	79.42	87.24	71.54	87.30	90.30	12.71
RF+M	**90.30**	**90.15**	**90.70**	**90.50**	**90.60**	**9.51**

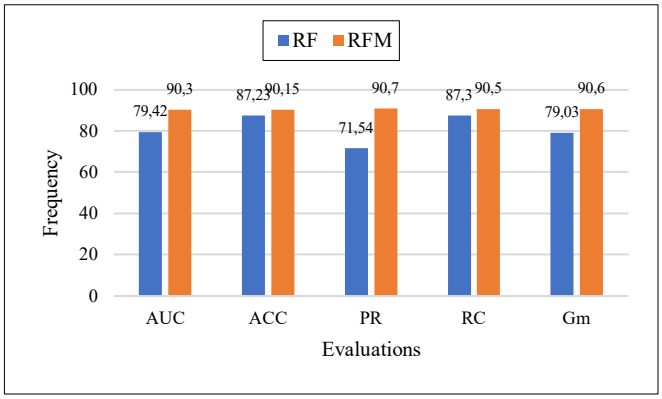

Fig. 3 Comparison of five performance evaluations for **RF** vs **RFM** only technique of IDM dataset

First, a comparison experiment was performed between RF and the meta-learning strategy (RFM) on the IDM dataset without the KMC undersampling-based filter method. To evaluate the model, Weka directly generated AUC, ACC, PR, RC, error classification (EC), and Gm, as shown in Table III, and the comparison performance can be seen in Fig. 3.

Table III shows that the hybrid strategy technique yielded AUC with an excellent classification of 90.30%, while ACC, PR, RC, and Gm values were 90.15%, 90.70%, 90.50%, and 90.60%, respectively. Meanwhile, misclassification

results are smaller than the conventional RF approach, which is 9.51% EC. This means the meta-learning strategy is promising enough for all performance evaluations but does not fully handle imbalanced class data, as stated by Wang and Sun [21]. Besides, AdaBoost algorithm is an effective solution for classification, but it still needs to improve the imbalanced class problem.

In the second experiment, the RUS-KMC+RFM technique was implemented, and the results, as shown in Table IV, and the comparison performance can be seen in Fig. 4. The ACC. PR. RC. Gm. and EC is directly calculated from Weka before calculating AUC.

TABLE IV
RESULT COMPARISON PERFORMANCE EVALUATION FOR META LEARNING ONLY TECHNIQUE (**RFM**) VS PROPOSED MODEL (**RUS-KMC+RFM**) OF IDM DATASET

Model	**(%)**					
	AUC	**ACC**	**PR**	**RC**	**Gm**	**EC**
RFM	90.30	90.15	90.70	90.50	90.60	9.51
RUS-KMC+RFM	**95.5**	**95.52**	**95.5**	**95.5**	**92.95**	**4.48**

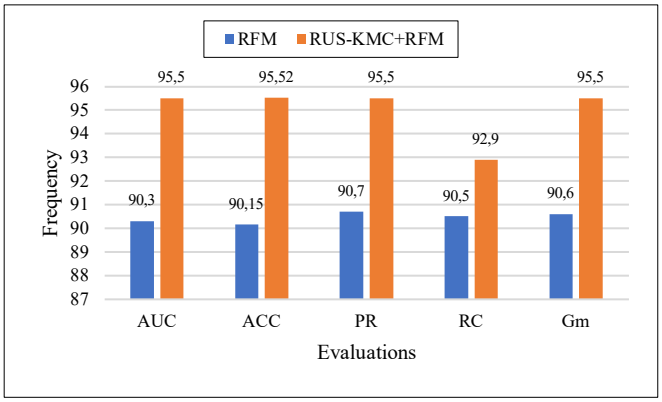

Fig. 4 Comparison of five performance evaluations for only **RFM** technique vs proposed model of IDM dataset

TABLE V
AUC COMPARISON WITH PRIOR RESEARCHES

Researchers	**Data Pre-processing**	**Learning Models**	**Model Validation**	**Evaluation**	**AUC (%)**
[14]	k-Means	SB *k*-NN	10-CV	AUC	<u>92.34</u>
Proposed	RUS-KMC	RF+M	10-CV	AUC	**95.50**

TABLE VI
TEST AUC COMPARISON WITH PRIOR STUDIES OF IDM DATASET

Schema for Comparison	***p*-Value**	**Difference**
RUS-KMC+M vs [14]	0.00001	Significant

This proposed model outperformed the first and second experiments of all evaluation performance in terms of performance. While the presentation of misclassification also gets, the best EC is smaller than the two models (RF and RFM), wherein the first experiment 4.48% < 12.71% and 9.51%, respectively. Based on this result, the overall second experiment was better than the first. On the other hand, this proposed strategy also answered what was stated by Wang and Sun [21].

Current and previous studies utilized the same private dataset; therefore, they are compared. Table V shows the k-Means+SB-kNN proposed by Siswanto, Suprapedi, and Purwanto [14] were selected for comparison. The AUC was utilized in this comparison because it is the primary evaluation in imbalanced class classification. In this comparison, a bold font means the best AUC value, and conversely underlined font represents the second best.

The proposed model outperforms prior studies and evaluations as it produced excellent AUC results and was statistically also compared to Siswanto, Suprapedi, and Purwanto [14] using the t-test can be seen in Table VI. According to this findings, although the proposed model gains the best accuracy also promising since it has a difference statistically with the best result.

In this research, specifically, the t-test model was employed to determine the difference between the proposed approach and other studies and discover which models perform better. The pair of proposed models vs. Siswanto, Suprapedi, and Purwanto [14] has a *p*-value of 0.00001, which indicates a substantial different, and that means it has a higher AUC value. Therefore, according to the t-test results, the proposed model showed an outstanding result and is competitive with the findings of the most recent study.

IV. CONCLUSION

In summary, the traditional classification model can achieve high accuracy in imbalanced class problems. It occurs because almost all the traditional classification models do only learn in the majority class and exclude the minority class, so the results are biased. In the pre-processing stage, the random under-sampling technique based on k-means cluster (RUS-KMC) was successfully used in the classification process. RUS was selected because many previous studies used this approach when data used contained imbalanced classes in classification problems. While KMC was selected as a clustering method because it promises to solve at least some of these problems, for example, (1) ability to handle mixed-type variables and large datasets, (2) the automatic determination of the number of optimal clusters, and (3) variables that may not be normally distributed. The evaluation results showed that the combination of RUS and KMC was very effective. The effectiveness is in selecting variables that might not be normally distributed and improving the performance of random forest classification based on meta-learning (RFM) in village development status classification better than previous studies in terms of AUC, ACC, PR, RC, and Gm, respectively 95.50%, 95.52%, 95.5%, 95.5%, and 92.95 %. In addition, the results of the t-test also reported a very good performance compared to previous studies. It can be concluded that this proposed model is effective in handling imbalanced classes in IDM dataset for the village development status classification model in Indonesia.

IDM dataset structure makes it difficult to study feature discretization in handling noisy attributes based on clustering techniques. Therefore, future studies need to consider comparing the suggested approach to other clustering models such DBSCAN, Fuzzy C-means, etc., as well as other meta-learning methods, including bagging and boosting.

ACKNOWLEDGMENT

The authors are grateful to the Ministry of Village. Development of Disadvantaged Regions and Transmigration (KEMENDESA) for access dataset. National Research and Innovation Agency (BRIN) for support. Institute for Research and Community Services of Universitas Muhammadiyah Semarang (LPPM UNIMUS) for the internal research grant scheme with project code: 040/UNIMUS.L/PT/PJ.INT/2021 for financial support.

REFERENCES

[1] K. Cheng, S. Gao, W. Dong, X. Yang, Q. Wang, and H. Yu. "Boosting label weighted extreme learning machine for classifying multi-label imbalanced data." *Neurocomputing*. vol. 403. pp. 360–370. Aug. 2020. doi: 10.1016/j.neucom.2020.04.098.

[2] A. Anil and S. R. Singh. "Effect of class imbalance in heterogeneous network embedding: An empirical study." *J Informetr*. vol. 14. no. 2. p. 101009. May 2020. doi: 10.1016/j.joi.2020.101009.

[3] E. Mortaz. "Imbalance accuracy metric for model selection in multi-class imbalance classification problems." *Knowl Based Syst*. vol. 210. p. 106490. Dec. 2020. doi: 10.1016/j.knosys.2020.106490.

[4] H. He and E. A. Garcia. "Learning from Imbalanced Data." *Curr Top Med Chem*. vol. 8. no. 18. pp. 1691–1709. 2008. doi: 10.2174/156802608786786589.

[5] J. M. Johnson and T. M. Khoshgoftaar. "Survey on deep learning with class imbalance." *J Big Data*. vol. 6. no. 1. pp. 1–54. Dec. 2019. doi: 10.1186/s40537-019-0192-5.

[6] A. Ali, S. M. Shamsuddin. and A. L. Ralescu. "Classification with class imbalance problem: A review." *International Journal of Advances in Soft Computing and its Applications*. vol. 7. no. 3. pp. 176–204. 2015.

[7] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas. "Handling imbalanced datasets : A review." *GETS International Transactions on Computer Science and Engineering*. vol. 30. no. 1. pp. 25–36. 2010. doi: 10.1007/978-0-387-09823-4_45.

[8] Han and Kamber. *Data Mining Concepts and Techniques Third Edition*. 3rd ed.. vol. 1. USA: Morgan Kaufmann Publishers is an imprint of Elsevier. 2012. doi: 10.1017/CBO9781107415324.004.

[9] J. Sun, H. Li, H. Fujita, B. Fu, and W. Ai. "Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting." *Information Fusion*. vol. 54. pp. 128–144. Feb. 2020. doi: 10.1016/j.inffus.2019.07.006.

[10] J. Song, X. Lu, and X. Wu. "An Improved AdaBoost Algorithm for Unbalanced Classification Data." *Sixth International Conference on Fuzzy Systems and Knowledge Discovery*. pp. 109–113. 2009. doi: 10.1109/FSKD.2009.608.

[11] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang. "Cost-sensitive boosting for classification of imbalanced data." *Pattern Recognit*. vol. 40. no. 12. pp. 3358–3378. Dec. 2007. doi: 10.1016/j.patcog.2007.04.009.

[12] H. Prasetyo and A. Purwarianti. "Comparison of distance measures for clustering data with mix attribute types for Indonesian potential-based regional grouping." in *2014 International Conference on Information Technology Systems and Innovation (ICITSI)*. Nov. 2014. pp. 13–18. doi: 10.1109/ICITSI.2014.7048230.

[13] M. S. Sari, D. Safitri, and Sugito. "Klasifikasi Wilayah Desa-Perdesaan dan Desa-Perkotaan Wilayah Kabupaten Semarang dengan Support Vector Machine." *Jurnal Gaussian*. vol. 3. no. 4. pp. 751–760. 2014. Accessed: Jun. 20. 2021. [Online]. Available: https://ejournal3.undip.ac.id/index.php/gaussian/article/view/8086

[14] E. Siswanto, Suprapedi, and Purwanto. "Metode Sample Boostraping Pada K-Nearest Neighbor Untuk Klasifikasi Status Desa." *Jurnal Teknologi Informasi*. vol. 14. pp. 13–23. 2018.

[15] A. Mahmud, A. Pangestika, A. P. Ramadhanty, G. M. Putra, G. S. N. D. S. Putri, and R. Nooraeni. "Klasifikasi Status Desa/Kelurahan DIY (Yogyakarta) Menggunakan Model Decision Tree (Studi Kasus Data Praktik Kerja Lapangan Politeknik Statistika STIS Tahun 2020)." *Engineering. MAthematics and Computer Science (EMACS) Journal*. vol. 3. no. 1. pp. 33–41. Feb. 2021. doi: 10.21512/emacsjournal.v3i1.6787.

[16] N. S. Kumar, K. N. Rao, A. Govardhan, K. S. Reddy, and A. M. Mahmood. "Undersampled K-means approach for handling imbalanced distributed data." *Progress in Artificial Intelligence*. vol. 3. no. 1. pp. 29–38. Aug. 2014. doi: 10.1007/s13748-014-0045-6.

[17] Jing-Hao Xue and P. Hall. "Why Does Rebalancing Class-Unbalanced Data Improve AUC for Linear Discriminant Analysis?." *IEEE Trans Pattern Anal Mach Intell*. vol. 37. no. 5. pp. 1109–1112. May 2015. doi: 10.1109/TPAMI.2014.2359660.

[18] L. Gautheron, A. Habrard, E. Morvant, and M. Sebban. "Metric Learning from Imbalanced Data with Generalization Guarantees." *Pattern Recognit Lett*. vol. 133. pp. 298–304. May 2020. doi: 10.1016/j.patrec.2020.03.008.

[19] 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. May 2017. doi: 10.1016/j.eswa.2016.12.035.

[20] J. Ri and H. Kim. "G-mean based extreme learning machine for imbalance learning." *Digit Signal Process*. vol. 98. p. 102637. Mar. 2020. doi: 10.1016/j.dsp.2019.102637.

[21] W. Wang and D. Sun. "The improved AdaBoost algorithms for imbalanced data classification." *Inf Sci (N Y)*. vol. 563. pp. 358–374. Jul. 2021. doi: 10.1016/j.ins.2021.03.042.