III.B.1.a3_TURNITIN_IEEE_On The Comparison: Random Forest, SMOTE-Bagging, and Bernoulli Mixture to Classify Bidikmisi Dataset in East Java

by W Suryaningtyas

Submission date: 23-Nov-2022 01:25PM (UTC+0700)

Submission ID: 1961875187

File name: Bernoulli_Mixture_to_Classify_Bidikmisi_Dataset_in_East_Java.pdf (153.7K)

Word count: 3563

Character count: 20094

On The Comparison: Random Forest, SMOTE-Bagging, and Bernoulli Mixture to Classify Bidikmisi Dataset in East Java

Nur Iriawan
Department of Statistics
Faculty of Mathematics, Computing and Data Science
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
nur_i@statistika.its.ac.id

Kartika Fithriasari
Department of Statistics
Faculty of Mathematics, Computing and Data Science
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
kartika_f@statistika.its.ac.id

Brodjol Sutijo Suprih Ulama
Department of Statistics
Faculty of Mathematics, Computing and Data Science
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
brodjol_su@statistika.its.ac.id

Wahyuni Suryaningtyas
Department of StatisticsFaculty of Mathematics, Computing and Data Science
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
wahyuni.pendmat@fkip.um-surabaya.ac.id

Sinta Septi Pangastuti
Department of Statistics
Faculty of Mathematics, Computing and Data Science
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
sisepti22@gmail.com

Nita Cahyani
Department of Statistics
Faculty of Mathematics, Computing and Data Science
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
cahyani.nc@gmail.com

Laila Qadrini
Department of Statistics
Faculty of Mathematics, Computing and Data Science
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
qadrini.laila@gmail.com

***Abstract*—The Bidikmisi is a scholarship program from the Indonesian government that intended for students who are not economically capable, but they have good academic performance. In the implementation of the Bidikmisi scholarship program, there are indications of a problem, namely the condition of inaccurate allocation in the Bidikmisi scholarship that is accepted or unaccepted. The purpose of this study was to examine several comparison methods that were used to get the accuracy allocation of the Bidikmisi scholarship in East Java. These methods include random forest, SMOTE-Bagging, and Bernoulli mixture model. Based on the AUC and g-mean values, the Bernoulli mixture method has a better proficiency than the random forest and SMOTE-Bagging.**

***Keywords*— AUC, bagging, Bernoulli mixture, g-mean, random forest, SMOTE.**

I. Introduction

Bidikmisi is a scholarship program [1] that intended for students who are not economically capable with good academic performance. Economically incapacitated one of them is a student with middle and lower economic conditions, namely parental income divided by the number of dependents not more than Rp. 750,000 [2]. However, in the implementation of the Bidikmisi scholarship program there are indications of a problem, namely the inaccurate allocation conditions in receiving the Bidikmisi scholarship that is accepted or unaccepted, so in this case, an in-depth evaluation is needed. The use of classification method in data mining is an appropriate tools to be proposed here.

Data mining is a method that is often used to find out hidden relationships between variables. One of the popular data mining methods is classification, where classification is a supervised approach that classifies students into known classes. Classification techniques have the purpose of finding a decision function that accurately predicts the class of testing data that comes from the same distribution function as the data for training. Therefore there are two conditions in the data class set, namely the balance and imbalance of the data. When a class exceeds the number of other classes, there is an imbalance data. The field of machine learning and data mining having the imbalances data have been identified as important issues. Bidikmisi data based on pre-processing results indicate the problem of imbalance class. Addressing the problem of imbalance data there are several approaches; i.e. the random forest which is a combination of bootstrap aggregating (bagging) methods [3] and random feature selection. It's because almost all classifiers including the random forest assume an even distribution between observation classes, so when a dataset that has imbalance classes, the performance tends to give less than optimal result. Pangastuti [4] had conducted a study on imbalance dataset by combining boosting and bagging algorithms with the Synthetic Minority Oversampling Technique (SMOTE) algorithm. The analysis results shown that the ensemble algorithm SMOTE-Boosting [5] and SMOTE-Bagging [6] gives better performance.

There were two works that had already succeeded to studying the selection framework of the Bidikmisi dataset. These two works had been done by Latumakulita, et. al. [7] who had implemented Fuzzy Inferences System Approach, and Latumakulita and Usagawa [8] who had demonstrated the

978-1-5386-7509-0/18/$31.00 ©2018 IEEE

work of the Combination of Back-Propagation Neural Network and Fuzzy Inference System Approaches.

Another method that was recently used is Bayesian Bernoulli Mixture Regression Model carried out by Iriawan, et. el. [9], on Bidikmisi Scholarship acceptance in the Indonesian provinces. This modelling made by arranging clusters of Bidikmisi scholarship applicants that are accepted and not accepted and through the Bernoulli mixture regression model estimation is carried out for each cluster. This study had cucceeded to compare methods of Bayesian Bernoulli mixture regression models, dummy regression models, and polytomous regression models. The comparison results had shown that the Bayesian mixture Bernoulli regression model provides better classification accuracy than the dummy regression model and polytomous regression model.

II. Background

In this research, we performed ensemble methods of the SMOTE-Bagging random forest and Bayesian Bernoulli mixture model to classify the Bidikmisi dataset. The Bidikmisi dataset was collected from the Bidikmisi database of Kemenristekdikti - Indonesia. This dataset evaluated by dividing data into two parts with stratified 10-fold cross-validation. Model classification performance is obtained by the performance value of each iteration [10].

A. Methods

This section showed the methods background about random forest, SMOTE-Bagging, and Bernoulli Mixture Model. The algorithm of each model can be seen in Fig.1, Fig.2, and Fig.3.

- Random Forest

In this research, we used random forest as a classical classification method. This method was begining to be widely discussed since the writing of Breiman [11] appeared in the Machine Learning journal. The random forest method is the development of the CART method, which is by applying the bootstrap aggregating (bagging) and random feature selection methods. Simply put, random forest formation algorithms can be mentioned as follows. Suppose we have n training data and p predictors variables. The stages of preparation and estimation using random forest is shown in Fig.1.

```
Random Forest Algorithm:
1. Pull with the replacement of a random sample size $n$ from training data (bootstrap stage).
2. Using the bootstrap example, the tree is built to reach the maximum size (without replacement).
3. Arrange the tree based on the data, but at each separation process select randomly $m < p$ explanatory variables, and do the best separation (random sub-setting stage).
4. Repeat steps 1-3 until the forest consisting of only one tree is formed.
5. Conduct a combined estimation based on the one tree (for example using majority vote in the case of classification or average for regression cases).
```

Fig. 1. Random Forest Algorithm

- SMOTE-Bagging

SMOTE-Bagging is a blended algorithm between SMOTE and Bagging algorithms. The construction of a subset in the SMOTE-Bagging method involves the generation of composite data [12]. SMOTE was first introduced by [5], is one of the oversampling methods. Synthetic data is generated by SMOTE until the amount of minor data is equivalent with the major data. Based on SMOTE-Bagging, before the model is formed, SMOTE does its duty by bootstrapping process balances for each subset. Based on two parameters, namely the number of oversampling (N) minority classes and closest neighbors can be made as synthetic data. Total oversampling is decided in such a way that the number of major classes and minor classes is balanced. The SMOTE-Bagging algorithm is displayed in Fig.2.

```
SMOTE-Bagging Algorithm:
1. Initiation of training data $D$
2. For $t = 1, \ldots, T$
    a. Creating a $D_t$ dataset using minor class ($N$) resample with returns, where $N$ is a multiple of 100%;
    b. Generalizing new data with SMOTE;
    c. Get a special classifier: $H : D_t \rightarrow R$ with the algorithm that has been given based on the original training data $D_t$;
3. The combination of the H classifier is made as a special classification aggregation: $t = 1, \ldots, T$ and an example are classified into $c_j$ class according to the number of votes obtained from the specific classifications;
```

$$H\left(d_i, c_j\right) = sign\left(\sum_{t=1}^{T} \alpha_t H_t\left(d_i, c_j\right)\right) \qquad (1)$$

where α_t is .... and d_i is ...

Fig. 2. SMOTE-Bagging Algorithm

- Bernoulli Mixture Model (BMM)

BMM was first performed by [13]. On Cancer and Schizophrenia [14-16] and Machine Learning research [17,18]. The Gibbs sampler is employed to generate the estimated parameters in BMM by using univariate conditional distribution by the steps given in Fig. 2.

```
Gibbs Sampler for BMM Algorithm:
Given $\mathbf{\Theta} = (\Theta_1, \Theta_2, \ldots, \Theta_d)' = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_L, \boldsymbol{\pi})'$
1. Determine the initial values $\mathbf{\Theta}^{(0)}$.
2. For $b = 1, 2, \ldots, B$ repeat the following steps
    a. $\mathbf{\Theta} = \mathbf{\Theta}^{(b-1)}$
    b. For $j = 1, \ldots, d$ renew $\Theta_j$ from
       $\Theta_j \sim f(\Theta_j \mid \mathbf{\Theta}_{\setminus j}, \mathbf{y})$
       where $f(\Theta_j \mid \mathbf{\Theta}_{\setminus j}, \mathbf{y})$ is a full conditional posterior distribution and
       $\mathbf{\Theta}_{\setminus j} = (\Theta_1, \ldots, \Theta_{j-1}, \Theta_{j+1}, \ldots, \Theta_d)^B$    (2)
    c. Set $\mathbf{\Theta}^{(b)} = \mathbf{\Theta}$ and save it as the generated set of values at $b + 1$ iteration.
```

Fig. 3. Gibbs Sampler Algorithm for BMM

Fig.4. shows the prediction parameter of BMM with two mixture components. The formation of the BMM model is depict using a WinBUGS doodle.

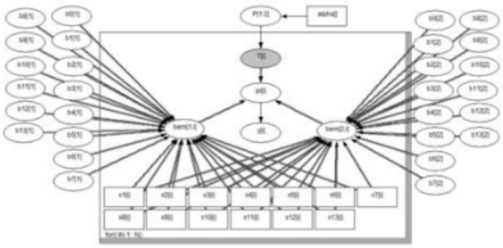

Fig. 4. Bernoulli Mixture Model Doodle

B. Performance Evaluation Classification Methods

Actual data and predictive data from the classification model are presented using a Confusion matrix containing information about the actual data class represented as the matrix row and the prediction data class as the column [20]. In the case of an imbalance class where the majority class is 98-99% of the total population, the results of the classification will achieve high accuracy because they only see the majority class. It is clear that for imbalance cases, the standard criteria measurement cannot be fulfilled by the classification accuracy. Area Under Curve (AUC) and metrics as precision are displayed, as well as using F-values to determine the proficiency of algorithms in minority classes. To evaluate the performance of the method as a whole, geometric mean (G_mean) and AUC analysis could be used. G_mean represents a geometric mean of Sensitivity and Specificity.

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN+FP})} \tag{3}$$

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP+FN})} \tag{4}$$

$$\text{G\_mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \tag{5}$$

AUC provides a single size classifier performance for evaluating which models are better on average. AUC value is obtained by calculating the value of the true positive rate (TPR) that is the number of objects in the positive class that is classified correctly and the false positive rate (FPR) is the number of objects in the positive class that is incorrectly classified.

$$\text{TPR} = \frac{\text{TP}}{(\text{TP+FN})} \tag{6}$$

$$\text{FPR} = 1 - \frac{\text{TN}}{(\text{TN+FP})} \tag{7}$$

$$\text{AUC} = \frac{1 + \text{TPR - FPR}}{2} \tag{8}$$

C. Pre-Processing

Identification of the formation of the Bernoulli Mixture distribution is given through the pre-processing stage as follows:

Step 1. Use the Y variable

Step 2. Pick covariate "father's income", "mother's income", and "family dependent"

Step 3. Define covariate "Characterization Category (CC)" by joining the amount of "father's income" and "mother's income" divided by "the number of family dependents"

Step 4. Provide coding for CC based on the criteria as follows:

0 = category of wealthy family with CC> Rp. 750,000 per head in the family

1 = category of poor families with CC <Rp. 750,000 per head in the family

Step 5. Fit the variable Y with CC in Step 4 to the Acceptance Condition (AC) with the Bidikmisi acceptance classification table given in TABLE I.

TABLE I. COMPONENTS MIXTURE OF BIDIKMISI SCHOLARSHIP

Y	CC	AC	Condition	Interpretation
1	0	0	Wrong	Acceptance Condition is wrong (Y = 1 and CC=0)
0	1	0	Wrong	Acceptance Condition is wrong (Y = 0 and CC=1)
1	1	1	Right	Acceptance Condition is Right (Y = 1 and CC=1)
0	0	1	Right	Acceptance Condition is Right (Y = 0 and CC=0)

The pre-processing show that *Y* is Bernoulli mixed distribution with two components, namely the component of the wrong acceptance conditions and the components of the correct acceptance conditions.

D. Research Variables and Flowchart

Response variable (Y) and the predictor variable (X) used in this research are shown in TABLE II.

TABLE II. SUMMARY OF THE BIDIKMISI DATASET

Variable & Description	Data Scale	Dummy Variables
Y (acceptance status)	Category	1 = Accepted, 0 = Not accepted.
X_1 (Father's job)	Nominal	4 Dummy Variables: d_{11}, d_{12}, d_{13} and d_{14}, where d_{11} = 1, if father's job is farmer, fisherman or others job which relate with agriculture or others job which relate with agriculture, and d_{11} = 0, otherwise. d_{12} = 1, if father's job is a civil servant, police, or army. d_{12} = 0, otherwise. d_{13} = 1, if father's job is an entrepreneur. d_{13} = 0, otherwise. d_{14} = 1, if father's job is a private employee. d_{14} = 0, otherwise
X_2 (Mother's Job)	Nominal	4 Dummy Variables: d_{21}, d_{22}, d_{23} and d_{24}, where d_{21} = 1, if mother's job is a farmer, fisherman or others job which relate with agriculture and d_{21} = 0, otherwise.

Variable & Description	Data Scale	Dummy Variables
		d_{22}= 1, if mother's job is a civil servant, police, or army. d_{22} = 0, otherwise. d_{23} = 1, if mother's job is an entrepreneur. d_{23} = 0, otherwise. d_{24} = 1, if mother's job is a private employee. d_{24} = 0, otherwise.
X_3 (Father's Education)	Nominal	3 Dummy Variables: d_{31}, d_{32}, and d_{33}, where d_{31} = 1, if father's education is not continue to school. d_{31} = 0, otherwise. d_{32} = 1, if father's education is an elementary, junior high or senior high school graduate level. d_{32} = 0, otherwise. d_{33} = 1, if father's education is a higher education level. d_{33} = 0, otherwise.
X_4 (Mother's Education)	Nominal	3 Dummy Variables: d_{41}, d_{42}, and d_{43}, where d_{41} = 1, if mother's education is not continue to school. d_{41} = 0, otherwise. d_{42} = 1, if mother's education is an elementary, junior high or senior high school graduate level. d_{42} = 0, otherwise. d_{43} = 1, if mother's education is a higher education level. d_{43} = 0, otherwise.
X_5 (Ownership of Family Homes)	Nominal	2 Dummy Variables: d_{51} and d_{52}, where d_{51} = Homeless d_{52} = Rent (Annually, Monthly) and hitchhike
X_6 (Land Area of Family Homes)	Nominal	2 Dummy Variables: d_{61} and d_{62}, where d_{61} = 25-50 m^2 d_{62} = 50-99 m^2
X_7 (The Extent of Family Residential Buildings)	Nominal	2 Dummy Variables: d_{71} = 25-50 m^2 d_{72} = 50-99 m^2
X_8 (Ownership of Toilet Washing Facilities)	Nominal	1 Dummy Variable: d_8 = Sharing
X_9 (Water source used by the family)	Nominal	1 Dummy Variable: d_9 = Wells, Rivers / Springs
X_{10}(Number of Families in the Household (Per person))	Ratio	
X_{11} (City Distance (Kilometer))	Ratio	
X_{12} (4th-semester ranking)	Nominal	2 Dummy Variables: d_{12_1} = Rangking 1 - 20 d_{12_2} = Rangking 21 - 40
X_{13} (5th-semester ranking)	Nominal	2 Dummy Variables: d_{13_1} = Rangking 1 - 20 d_{13_2} = Rangking 21 - 40

The classification analysis steps using random forest, SMOTE-Bagging, and Bernoulli mixture model, are given as the following research flowchart shown Fig.5.

III. RESULTS AND DISCUSSION

In this passage, we implement the Bidikmisi dataset which has binary classes. Table III summarizes the characteristics of the Bidikmisi dataset: the number of examples (Example), quantity of attributes (Attribute), the number of each class, and the imbalance-ratio (IR). We have gained the AUC and G_mean metric estimates by averaging the stratified cross-validation 10-folds.

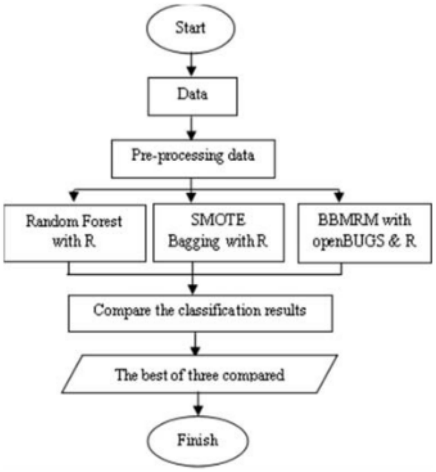

Fig. 5. Research Flowchat

TABLE III. SUMMARY OF THE BIDIKMISI DATASET

Data Set	*Example*	*Attribute*	*Class(Min;Max)*	*IR*
Bidikmisi	41937	29	(10654;31283)	2.94

In a binary classification problem, on the other hand, the labels are either positive or negative. The decision made by the classifier can be represented as a 2x2 confusion matrix. In general, use evaluation metrics to determine accuracy results, but this measurement is not suitable for evaluating the imbalanced dataset. Performance evaluation for Bidikmisi classification, therefore, was performed with several criteria. The AUCs in TABLE IV show that for dataset that combined with SMOTE is able to improve over under bagging with random forest as the base classifier.

The same conclusion obtained from G_mean, that combines true negative rate and true positive rate, where both the errors are considered equal. As we can see in TABLE V that the G-mean of SMOTE-Bagging method has higher value than other methods. Thus, the SMOTE approach provides an improvement in performance. It could be said that the method was quite successful in taking advantage of bagging algorithm couple with SMOTE. It can be said that the bagging affect the accuracy of random forest by focusing on all data classes, and the SMOTE algorithm changes the performance value of random forest only in the minority classes.

TABLE IV. AUC OF THE BIDIKMISI DATASET WITH THREE METHODS

Model	**Mean of Performance Classification with Tree = 50**	
	G_mean (%)	***AUC (%)***
Random Forest	10.889	53.728
SMOTE-Bagging	32.320	56.265

TABLE V. G-MEAN OF THE BIDIKMISI DATASET WITH THREE METHODS

Model	Mean of Performance Classification with Tree = 100	
	G_mean (%)	*AUC (%)*
Random Forest	9.524	52.857
SMOTE-Bagging	27.622	56.441

The comparison of several methods; i.e. random forest, under bagging, SMOTE-Bagging, and Bernoulli mixture model are then done on their optimum model. The performance of the classification method which includes G_mean and AUC shown in TABLE VI. The result demonstrates that the Bernoulli mixture model gives higher value of G_mean and AUC than other methods, with 44.748% and 75.094%. SMOTE-Bagging as an ensemble method gives 32.02% of G_mean and AUC 56.441%. It can be said that over bagging is quite successful than classic and under bagging methods, because the SMOTE algorithm improves the proficiency of the classifier only on the minority classes of the data observation.

TABLE VI. PERFORMANCE OF THE METHODS

Model	Mean of Performance Classification	
	G_mean (%)	*AUC (%)*
Random Forest	10.889	53.728
SMOTE-Bagging	32.320	56.441
Bernoulli Mixture	44.748	75.094

This study is an in-depth evaluation carried out to analyze Bidikmisi dataset, which had imbalance class problems. Therefore, a number of classification methods had been compared, namely random forest which is a combination of bootstrap aggregating (bagging) methods with the SMOTE-Bagging method and the Bayesian approach method on the Bernoulli Mixture distribution data. Based on the AUC and G_mean values, the Bernoulli mixture method has a better performance compared to the random forest and SMOTE-Bagging.

ACKNOWLEDGMENT

This research was funded by DPRM-DIKTI under scheme PTUPT. The author thanks to the Kemenristekdikti Republic of Indonesia for funding which support this research under PUPT research grant no. 1049/PKS/ITS/2018.

REFERENCES

[1] Ministry of Education, Bidik Misi Scholarship Program: Educational Scholarships for Prospective Students with Achievement from Underprivileged Families, Jakarta: Directorate General of Higher Education, 2010.

[2] Ministry of Technology & Education Research, Guidance for Providing Assistance for Bidikmisi Education Costs in 2015, Jakarta: Director General of Learning and Student Affairs, 2015.

[3] Breiman, L. "Random Forests". Machine Learning, vol. 45, 2001, pp. 5-32.

[4] Pangastuti, S.S. "Comparison of random forest ensemble method with SMOTE-bosting and SMOTE-bagging in the classification of data mining for inbalance classes (case study: Bidikmisi Scholarship Data in 2017 in East Java)," unpublished. 2018.

[5] Chawla, N.V., Lazarevic, A., Hall, L.O and Bowyer, K.W., "SMOTE Boost: Improving prediction of the minority class in boosting," Proc. Knowl.Discov, Databases, 2002, pp. 107–119.

[6] Wang, S and Yao, X., "Diversity analysis on imbalanced datasets by using ensemble models," IEEE Symp. Comput. Intell. Data Mining, 2009, pp. 324–331.

[7] Latumakulita, L. A., Purnama, F., Usagawa, T., Paturusi, S., and Prima, D. A. "Indonesia Scholarship Selection Framework Using Fuzzy Inferences System Approach. Case Study: "Bidik Misi" Scholarship Selection", The 2016 International Conference on Information, Communication Technology and System (ICTS), 978-1-5090-1381-4/16/$31.00 ©2016 IEEE, DOI: 10.1109/ICTS.2016.7910282, pp 107-113, 2016

[8] Latumakulita, L. A. and Usagawa, T., "Indonesia Scholarship Selection Model Using a Combination of Back- Propagation Neural Network and Fuzzy Inference System Approaches", International Journal of Intelligent Engineering and Systems, Vol.11, No.3, 2018 DOI: 10.22266/ijies2018.0630.09, pp. 79-90, 2018.

[9] Iriawan, N., Fihriasari, K., Ulama, B, S, S., Suryaningtyas, W., Susanto, I., and Pravitasari, A. A. "Bayesian Bernoulli mixture regression model for Bidikmisi scholarship classification," Journal of Computer Science and Information, vol. 11, pp. 67 – 76. UI Depok, 2018.

[10] Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A Feature selection for high-dimensional data, Springer, 2015.

[11] Breiman, L. Bagging Predictors. Machine Learning, 1996, pp. 123-140.

[12] Wang, S and Yao, X., "Diversity analysis on imbalanced datasets by using ensemble models," IEEE Symp. Comput. Intell, Data Mining, 2009, pp. 324–331.

[13] Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N, "The BUGS project: evolution, critique and future directions (with discussion)," Statistics in Medicine, vol. 28, DOI : 10.1002/sim.3680, 2009. pp 3049-3082.

[14] Sun, Z., Rosen, O., and Sampson, A, "Multivariate Bernoulli Mixture Models with application to Postmortem Tissue Studies in Schizophrenia," Biometrics, vol. 63, pp. 901-909, 2007.

[15] Tikka, J., Hollmen, J., and Myllykangas, S, "Mixture Modelling of DNA Copy Number Amplification Patterns in Cancer," in Proceedings of the 9th International Work-Conference on Artificial Neural Networks (IWANN), Springer-Verlag, Berlin, Heidelberg, 2007.

[16] Myllykangas, S., Tikka, J., Böhling, T., Knuutila, S., and Hollmén, J, "Classification of Human Cancers Based on DNA Copy Number Amplification Modelling," BMC Med. Genomics, vol. 1, pp. 1-13, 2008.

[17] Saeed, M., Javed, K., and Babri, H. A, "Machine Learning Using Bernoulli Mixture Models: Clustering, Rule Extraction and Dimensionality Reduction," Neurocomputing, vol. 119, pp. 366–374, 2013.

[18] Bouguila, N, "On Multivariate Binary Data Clustering and Feature Weighting," Comput. Stat.Data Anal, vol. 54, pp. 120-134, 2010.

[19] McLachlan, G., and Peel, D, Finite Mixture Models. New York: John Wiley and Sons, Inc., 2000.

[20] Kubat, M., and Matwin, S. "Addressing the curse of imbalanced training sets: one-sided selection". 14th International Conference on Machine Learning, Conference Proceedings, pp. 179–186, 1997.

III.B.1.a3_TURNITIN_IEEE_On The Comparison: Random Forest, SMOTE-Bagging, and Bernoulli Mixture to Classify Bidikmisi Dataset in East Java

ORIGINALITY REPORT

10%	%	10%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)", Springer Science and Business Media LLC, 2019 Publication	2%
2	N I Asrori, N Iriawan, W S Winahju. "Vehicle's Density Prediction Based on History Data of e-Toll in PT Jasamarga Pandaan Tol Using Hidden Markov Model", Journal of Physics: Conference Series, 2020 Publication	2%
3	D Rantini, N Iriawan, Irhamah. "Bayesian Mixture Generalized Extreme Value Regression with Double-Exponential CAR Frailty for Dengue Haemorrhagic Fever in Pamekasan, East Java, Indonesia", Journal of Physics: Conference Series, 2021 Publication	2%

4	"Soft Computing in Data Science", Springer Science and Business Media LLC, 2019 Publication	1%
5	Fikri Budiman, Irwan Agus Saputro, Purwanto Purwanto, Pulung Nurtantio Andono. "Optimization Of Classification Results By Minimizing Class Imbalance On Decision Tree Algorithm", 2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE), 2022 Publication	1%
6	Siti Qomariyah, Nur Iriawan, Kartika Fithriasari. "Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis", AIP Publishing, 2019 Publication	1%
7	Ria Arafiyah, Zainal A. Hasibuan, Harry Budi Santoso. "Monitoring online learners' performance based on learning progress prediction", AIP Publishing, 2021 Publication	1%
8	Heri Kuswanto, Achmad Naufal. "Evaluation of performance of drought prediction in Indonesia based on TRMM and MERRA-2 using machine learning methods", MethodsX, 2019 Publication	1%

Exclude quotes	On	Exclude matches	< 20 words
Exclude bibliography	On		