Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 4 No. 2 2024, 76 - 82
DOI: *https://doi.org/10.22487/27765660.2024.v4.i2.17279*
eISSN: 2776-5660

ADAPTIVE SYNTHETIC IMPLEMENTATION ON RANDOM FOREST IN ARCHIPELAGIC FISHING PORT OF PEMANGKAT

Naomi Nessyana Debataraja [1*], Dadan Kusnandar [2], Joannes Fregis Philosovio Anugrahnu [3]

[1,2,3]Department of Statistics, Tanjungpura University

***e-mail**: *naominessyana@math.untan.ac.id*

ABSTRACT

Random Forest is one of the classification methods employed in data mining. One of the problems in data mining classification is the problem of unbalanced class data This phenomenon arises when the data classes utilized do not have identical instances. Imbalance class data causes the classification results to be biased towards the majority class. Adaptive Synthetic (ADASYN) can be used to deal with this problem. ADASYN generates synthetic data by assigning different importance of minority class samples and then producing synthetic data with similar characteristics. The implementation of ADASYN is suitable for fishery production data, which will experience the problem of unbalanced class data. Fish production is part of the measured fishery. This study aims to classify the value of measured fishery production at PPN Pemangkat through Random Forest Classification using ADASYN to handle the imbalance class data problem and compare the results with those without ADASYN implementation. This study uses four predictor variables which include fishing gear types (X_1), number of trip days (X_2), number of crew (X_3), and the total weight of fish (X_4) with production value as response variable (Y). Accuracy, precision, recall, specificity, and G-mean are the model performance indicators used. The results showed that ADASYN successfully handles the problem of unbalanced class data in Random Forest classification. Accuracy is increased from 78.9% to 79.19%, Specificity is increased from 0.68% to 4.11%, Precision from 78.98% to 79.47%, and G-Mean from 8.23% to 20.19%. The 0.55% decrease in recall is negligible due to the small amount, so the Random Forest classification with ADASYN is better than without ADASYN.

Keywords: *Classification, Minority Class, Fishing Production, Accuracy*

Cite: Debataraja, N. N., Kusnandar, D., & Anugrahnu, P. F. J, (2024). *Adaptive Synthetic Implementation on Random Forest in Archipelagic Fishing Port of Pemangkat*. *Parameter: Journal of Statistics*, *4*(2), 76-82, https://doi.org/10.22487/27765660.2024.v4.i2.17279.

[*]Received: Jul 2024; Reviewed: Jul 2024; Published: Dec 2024

INTRODUCTION

Random forest is an algorithm built from a collection of structured classification trees (Genuer & Poggi, 2020). Random Forest uses randomization and the ensemble learning process to reduce the correlation between trees, handle large and diverse amounts of data, and be able to handle missing data problems (Ali, Khan, Ahmad, & Maqsood, 2012). One of the problems in data mining classification is the imbalance class data problem. Imbalance class data occurs when the data classes used do not have the same number of data (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Classification results on data with imbalanced class data problems will be biased towards the majority class so that almost all classification results can safely always produce the majority class.

One method that can be used to deal with the problem of imbalanced class data is the Adaptive Synthetic (ADASYN) Sampling Approach. ADASYN can improve classification accuracy by creating synthetic data on minority class data (Aqsha, Thamrin, & Lawi, 2021). The synthetic data created by ADASYN has a wide variety of data with data characteristics that are close to the original data achieved from assigning weights to minority classes based on its difficulty of understanding. This characteristic is an advantage of ADASYN over the usual undersampling and oversampling methods so that its use can be applied to various types of data. Data on fisheries production is one of the data that does not escape the problem of imbalance class data. In practice, not all fisheries production can have the same results in a certain period. This targeted synthetic data generation provides an advantage over traditional under-sampling and over-sampling, making it adaptable across varied data types, including fisheries production data where imbalance issues frequently arise. Fisheries production data are therefore suitable for the implementation of ADASYN.

The Ministry of Maritime Affairs and Fisheries (MMAF) is a government agency that oversees marine affairs and fisheries. There are several types of ports under the regulation of MMAF, including the Pelabuhan Perikanan Samudera (PPS) and the Pelabuhan Perikanan Nusantara (PPN). There are seven PPSs and 18 PPNs throughout Indonesia. Pemangkat PPN in Sambas Regency is one of these PPNs. Pemangkat PPN is one of the centers of fisheries activities in West Kalimantan (Safitri & Magdalena, 2018). Things related to fisheries production such as annual data on fisheries production is one aspect that is considered by Pemangkat PPN. This study aims to classify Pemangkat PPN's fisheries production data using Random Forest, comparing results with and without ADASYN integration to determine the effectiveness of ADASYN in addressing data imbalance issues.

MATERIALS AND METHODS

Random Forest was developed by Leo Breiman from the Bootstrap Aggregating (Bagging) process. If Bagging randomizes the sample data, Random Forest does the same thing and also randomizes the independent variables used. The results of the decision tree formed then have different sizes and shapes (Jatmiko, Padmadisastra, & Chadidjah, 2019). There are several steps that need to be taken to construct Random Forest classification model (Lee, Ullah, & Wang, 2020).

1. Divide the data into training and testing data.
2. In the training data, a number of B sample sets are formed through the bootstrapping process.
3. After forming a number of B sample sets, a Decision Tree is formed on each sample set formed.
4. In each Decision Tree formed, only m variables are used. To select m variables in each Decision Tree, the criterion $1 \leq m \leq \sqrt{p}$ is used where p is the total number of variables.
5. Perform majority voting on all Decision Tree to determine the final classification result (aggregating process).

According to Breimann, the next step is to determine the relative importance of each variable in the Decision Tree from the bootstrapping sample set. The relative importance values are then averaged to see the relative variable importance in Random Forest as a whole. Given each sample set b of B population with the jth variable, error improvement of e_t, and total number of nodes T_b with each of internal node t, relative variable importance level can be calculated from the following equation.

$$I_j^2(b) = \sum_{t=1}^{T_b} e_t^2 I(v(t)_b = j) \quad (1)$$

The overall relative variable importance level then can be calculated through the equation below given the population B for each jth variable in each b tree.

$$I_j^2 = \frac{1}{B}\sum_{b=1}^{B} I_j^2(b) \quad (2)$$

Classification models in Data Mining, especially machine learning algorithms, uses confusion matrix to evaluate their model's accuracy through performance model indicator. Performance model incicators that are widely used include Accuracy, Recall, Specificity, and Precision. In handling imbalance class data, thus G-mean score is used. To calculate G-mean score, Recall and Specificity are required simultaneously. The following table shows the confusion matrix with its components (Syukron, Sasntoso, & Widiharih, 2020).

Table 1. Confusion Matrix

Class Data	Actual Positive (Majority)	Actual Negative (Minority)
Positive Prediction (Majority)	True Positive (TP)	False Positive (FP)
Negative Prediction (Minority)	False Negative (FN)	True Negative (TN)

Accuracy is a simple performance model indicator by calculating the correct prediction compared to the total sample. Recall is used to measure the accuracy of classification for positive data classes. In the case of imbalance class data, Accuracy and Recall cannot be directly interpreted to explains the model because there is a bias that occurs in the majority class so that the correct prediction in the calculation cannot represent the actual class. Thus, we need to mainly use Precision, Specificity, and G-Mean. Precision and Specificity are used to measure the accuracy of the model in predicting positive classes and negative classes respectively. G-Mean then is used to measure the balance of classification in classifying the majority and minority classes from Precision and Recall. G-Mean score is always positive. The following equations are used to calculate Accuracy, Recall, Specificity, Precision, and G-Mean. All the components to calculate these equations are generated from confusion matrix.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$Specificity = \frac{TN}{TN+FP} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$G\text{–}Mean = \sqrt{(Recall \times Specificity)} \quad (7)$$

Adaptive Synthetic (ADASYN) is an oversampling technique to increase minority class data to overcome the problem of imbalance class data. Its core concept involves assigning weights to minority class distributions based on their complexity level, aiming to address comprehension challenges. The process of increasing data to minority class data is done by generating synthetic data in the training data. Pseudo code for ADASYN can be shown below (Chen, Zhou, & Yu, 2021).

Table 2. Pseudo Code for ADASYN

Algorithm: ADASYN
Input: Training dataset X_T, Hyper parameter $\beta \in [0,1]$, K=5 The ith sample in the minority sample x_i $(i = 1, 2, 3, \dots, m_s)$ A random minority sample x_{zi} in K-nearest neighbors of x_i Output: Synthetic minority samples s_i, oversampled training dataset X_{ADASYN}
1 Calculate the number of majority samples m_l and the number of minority samples m_s in Training dataset X_T
2 According to the formula $G = (ml - ms) \times \beta$ calculates the number of samples to be synthesized for minority class
3 **For each example** $xi \in$ **minorityclass:**
4 Calculate Δi //the number of majority samples in K-nearest neighbors of minority Sample x_i

5 Calculate $r_i = \Delta i/K$ //the ratio of majority samples in K-nearest neighbors of minority Sample x_i

6 Standardize ri through the formula $\hat{r}_i = r_i / \sum_{i=1}^{m_s} r_i$

7 Calculate $gi = \hat{r}_i \times G$ //the number of new samples to be generated for each minority x_i

8 **Do the loop from 1 to** g_i

9 Using the formula $s_i = x_i + (x_{zi} - x_i) \times \gamma$ to synthesize data samples //γ is a random number: $\gamma \in [0,1]$

10 **End**

11 **End**

12 **Return** Oversampled training datasets X_{ADASYN}

RESULTS AND DISCUSSION

The data is obtained from Pemangkat PPN and is the data of 2021 fishery production value consists of five variables with four predictor and one response variable. Fishing gear types (X_1), fishing day (X_2), number of crew (X_3), and total weight of fish (X_4) are the predictor variables while production value (Y) is the response variable. The data scoring with each detail is shown in Table 2 (Irnawati, Simbolon, Wiryawan, Murdianto, & Nurani, 2011). The data is then analysed using random forest classification, followed by the comparison between before and after the implementation ofAdaptive Synthetic (ADASYN).

The total data is 3445 and Table 2 shows that there is an imbalance class data problem in the data used. In the production value response variable, it is obtained that the class is divided into two and the production value of more than or equal to 30 million rupiah is the minority class. It is also noted that the ratio of majority and minority classes in the data used is 20:80. Only 20% of all measured fishery production values in 2021 were able to reach a number equal to or above IDR 30 million. Therefore, the implementation of Adaptive Synthetic (ADASYN) is the right step to take.

The amount of data in each class of variables used is also uneven. In the predictor variable of fishing gear types, small purse seine is the class that has the largest amount of data compared to other classes in the same variable. This is also the case for the classes on the variables fishing day, number of crew, and total weight of fish. This information reflects that the measured fishing conditions at Pemangkat PPN in 2021 are skewed towards certain characteristics. Therefore, the results obtained from the Random Forest classification method can help the process of increasing the effectiveness of measured fishing at Pemangkat PPN due to its ability to generate different models of tree simultaneously.

Table 3. Scoring and Data Description

Variables	Categories	Information	Total
Production Value	1	< IDR 30 Million	2730
	2	≥ IDR 30 Million	715
Fishing Gear Types	1	Small purse seine	3337
	2	Stick held dip net	16
	3	Cast net	64
	4	Bottom long line	28
Day of Trip	1	< 21 days	713
	2	21 to 40 days	2648
	3	> 40 days	84
Number of Crew	1	< 11 people	44
	2	11 to 20 people	1369
	3	> 20 people	2032
Total Weight of Fish	1	< 10 Tons	336
	2	≥ 10 Tons	3109

In this research, the data split is on the ratio of 80:20 with 80% for training data and 20% for testing data. The data split is followed by the ADASYN implementation on increasing the minority class data as the data is already confirmed as having imbalance class data problem. Figure 1 is showing the initial split data side by side with the final split data after ADASYN already implemented.

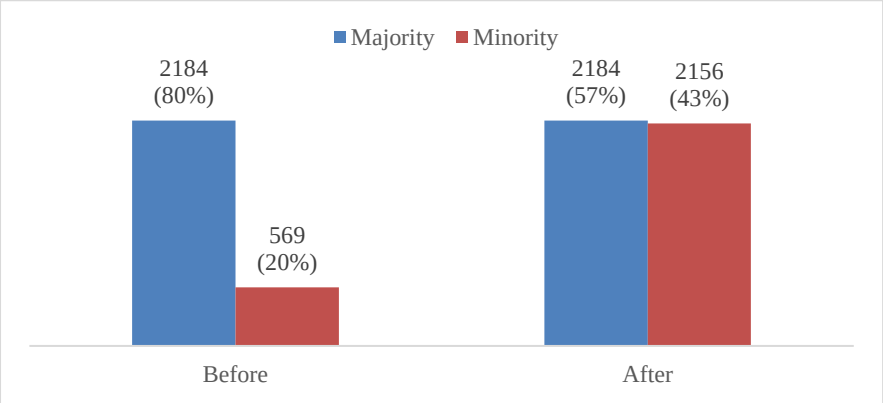

Figure .1 Split Data Before and After Implementing ADASYN

The next step is to use Random Forest to do classification on both split data. Number of trees created is 500 with all variables availavle in each tree's split node. After acquiring the model, the result of average relative variable importance is also generated. Figure 2 is showing the side-by-side comparison of average relative variable importance for both Random Forest model with and without implementation of ADASYN. The results are similar thus proving that ADASYN implementation does not affect the characteristics of the original dataset.

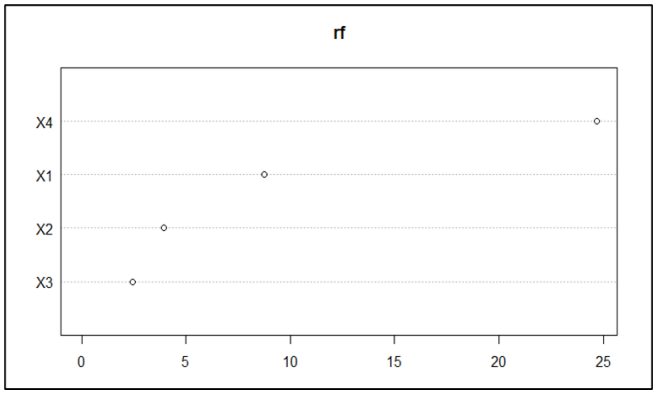

Figure 2. Average Variable Importance of Random Forest

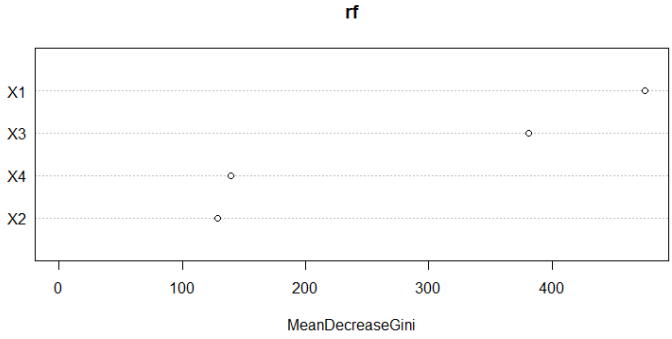

Figure 3. Average Variable Importance of Random Forest with ADASYN

Each score in performance model indicator of Accuracy, Recall, Specificity, Precision, and G-Mean have been obtained for Random Forest classification with and without ADASYN from confusion matrix. The direct comparison of the performance model indicator is shown on Table 3. Random Forest classification with ADASYN provides better results than Random Forest classification without

ADASYN. The main difference lies in the Accuracy, Specificity, Precision, and G-Mean scores which have increased by 0.29%, 3.43%, 0.49%, and 11.96% respectively. Random Forest classification with ADASYN has better classification accuracy for minority classes. This difference proves that the bias arising in the problem of imbalance class data towards the majority data class can be mitigated by the implementation of ADASYN. In addition, the decreased Accuracy and Recall values are negligible in the case of imbalance class data considering that the bias in the majority class greatly affects the scores generated by Accuracy and Recall.

Table 4. Confusion Matrix

Performance Model Indicator	Random Forest	Random Forest with ADASYN
Accuracy	78.9%	79.19%
Recall	99.81%	99.26%
Specificity	0.68%	4.11%
Precision	78.98%	79.47%
G-Mean	8.23%	20.19%

CONCLUSION

Adaptive Synthetic (ADASYN) can be implemented well in the classification of measured fishing production data from the Pemangkat Pelabuhan Perikanan Nusantara (PPN) in 2021 with the Random Forest classification. ADASYN adds a layer of analysis that is useful in identifying and being a solution to the problem of imbalance class data in Random Forest classification conducted in the research. The results of Random Forest classification with the implementation of ADASYN also significantly help on handling imbalance class data problem. There were increases in Accuracy performance indicators by 0.29%, Specificity performance indicators by 3.43%, Precision by 0.49%, and G-Mean by 11.96% Therefore, ADASYN is proven to be able to improve the performance of Random Forest classification in dealing with data class imbalance problems. Recall value were decreased but were negligible due to the bias that occurred in the majority class. Nevertheless, further research is anticipated to yield alternative methodologies for enhancing the value of performance model indicators, particularly in terms of accuracy and precision.

REFERENCES

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. *IJCSI International Journal of Computer Science Issues*, *9*(5). 272-278

Aqsha, M., Thamrin, S., & Lawi, A. (2021). Combination of ADASYN-N and Random Forest in Predicting of Obesity Status in Indonesia: A Case Study of Indonesian Basic Health Research 2013. *Journal of Physics: Conference Series*, *2123*(1), 012039. https://doi.org/10.1088/1742-6596/2123/1/012039

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, Z., Zhou, L., & Yu, W. (2021). ADASYN−Random Forest Based Intrusion Detection Model. *2021 4th International Conference on Signal Processing and Machine Learning*, 152–159. https://doi.org/10.1145/3483207.3483232.

Genuer, R., Poggi, JM. (2020). *Random Forests. In: Random Forests with R. Use R!*. Springer, Cham. https://doi.org/10.1007/978-3-030-56485-8_3

Irnawati, R., Simbolon, D., Wiryawan, B., Murdianto, B., & Nurani, T. W. (2011). Leading commodity analysis of capture fisheries in Karimunjawa National Park. *Jurnal Perikanan dan Kelautan*, *1*(1), 11-17 (2011).

Jatmiko, Y. A., Padmadisastra, S., & Chadidjah, A. (2020). Analisis Perbandingan Kinerja Cart Konvensional, Bagging Dan Random Forest Pada Klasifikasi Objek: Hasil Dari Dua Simulasi. *Media Statistika*, *12*(1), 1-12.

Lee, T.-H., Ullah, A., & Wang, R. (2020). Bootstrap Aggregating and Random Forest. In P. Fuleky (Ed.). *Macroeconomic Forecasting in the Era of Big Data,* 52, 389–429.

Safitri, I., & Magdalena, W. (2018). Perikanan Tangkap Purse Seine di Pelabuhan Perikanan Nusantara (PPN) Pemangkat Kalimantan Barat. *Jurnal Laut Khatulistiwa, 1*(3), 89–96.

Syukron, M., Santoso, R., & Widiharih, T. (2020). Perbandingan Metode Smote Random Forest dan Smote Xgboost untuk Klasifikasi Tingkat Penyakit Hepatitis C pada Imbalance Class Data. *Jurnal Gaussian*, *9*(3), 227–236.