Variance: Journal of Statistics and Its Applications
P-ISSN: 2685-8738 E-ISSN: 2685-872X
Variance: Journal of Statistics and Its Applications
Volume 7 Issue 1 | April 2025
10.30598/variancevol7iss1page31-38

PERFORMANCE ANALYSIS OF RANDOM FOREST CLASSIFICATION ON UNEMPLOYMENT RATE IN MALUKU PROVINCE BASED ON DATA BALANCING METHOD

Mahdayani Putri Yunizar[1], Lexy Janzen Sinay[2], Yudistira[3*]
[1,2,3]Statistics Study Program, Faculty of Science and Technology, Pattimura University
Jalan Ir. M. Putuhena, Kampus Unpatti Poka, Ambon, 97233, Maluku, Indonesia

Corresponding Author's Email*: *yudistira@lecturer.unpatti.ac.id

Abstract: *In 2023, the number of unemployed people in Maluku will reach 59,800 or 6.08% of the total population. To reduce unemployment in Maluku, it is essential to understand the unemployment situation of the Moluccan population based on socioeconomic factors immediately. Therefore, applying classification methods such as random forests is the right step, but it is recommended that the data be balanced to get accurate results. However, the unemployment rate in Maluku is much lower than that of the unemployed, so data imbalance affects the accuracy of the classification results. Therefore, a data balancing process is needed, among others, using the Random Oversampling of Sample (ROSE), Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN) methods. This study uses data from the 2023 National Labor Force Survey (SAKERNAS) conducted in February by the Central Statistics Agency (BPS) of Maluku. The number of unemployed people is smaller than the number of unemployed residents. Therefore, action needs to be taken to address data inequality. The results of this study show that the random forest classification model with SMOTE has the best performance with a combination of 90% training data and 10% testing data, with a higher AUC value than other methods, and age variables are the most essential variables built into the model.*

Keywords: *ADASYN, Classification, Random Forest, ROSE, SMOTE, Unemployment.*

1. INTRODUCTION

Unemployment is one of the problems faced by developing countries, including Indonesia. Many factors interact and form sometimes unclear patterns, making unemployment a complex issue. Unemployment can cause social vulnerability and ultimately lead to poverty if not addressed immediately [1].

Nationally, Maluku Province ranks seventh out of 38 provinces in February 2023 with a TPT of 6.08%, higher than the national TPT of 5.45%. With a TPT of 6.08%, the government must work hard to tackle unemployment. To reduce the still-high unemployment rate, the regional government must consider a development master plan with employment-based policies [2].

In efforts to reduce unemployment in Maluku Province, understanding the unemployment status of the population based on socioeconomic factors becomes crucial. Factors such as education level, geographic location, and employment sector are relevant indicators in determining the unemployment rate in the region. Therefore, classification methods like random forest are the right choice for identifying complex patterns from various variables affecting unemployment risk. Previous studies have shown that random forests can predict the Not in Education, Employment, or Training (NEET) status of young people in Indonesia with an accuracy of 82.94% [3].

Accurate random forest classification requires balanced data, meaning the number of data points between the majority class (employed) and the minority class (unemployed) should not differ significantly. However, based on factual data, the number of unemployed is much smaller than the employed, causing the model to pay more attention to the majority class and ignore the minority class, posing a challenge in applying random forest. Therefore, data-balancing methods are needed to address this class imbalance.

Data balancing methods such as Random Over-sampling Examples (ROSE), Synthetic Minority Over-sampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN) can be used to address this imbalance issue. ROSE randomly duplicates samples from the minority class to match the number of the majority class. SMOTE creates new synthetic samples among existing minority samples, while ADASYN adaptively performs synthetic oversampling by considering data density around minority samples. By applying

Copyright© 2025: Author(s)
https://ojs3.unpatti.ac.id/index.php/variance/ *Submitted: 2024-07-18* *Accepted: 2025-05-06*
This work is licensed with Lisensi Internasional 4.0 Atribusi Creative Commons-NonCommercial-ShareAlike 4.0.

these data balancing methods, the resulting classification model is expected to recognize patterns from the minority class better and produce more accurate classifications overall.

In this study, the use of data balancing methods such as Random Over-sampling Examples (ROSE), Synthetic Minority Over-sampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN) aims to improve the balance between the majority and minority classes in the unemployment dataset in Maluku Province. By balancing the number of samples between classes, the built classification model is expected better to recognize patterns from the minority class, namely unemployment, and produce more accurate classification results. The results of this study will then be compared to evaluate the performance of the classification model using metrics such as accuracy, sensitivity, specificity, and Area Under the Curve (AUC). Thus, this study not only attempts to identify and classify unemployment rates in Maluku Province but also contributes to addressing data imbalances to improve classification model accuracy [4].

2. METHODOLOGY

2.1. Data Source and Research Variables

The data used is the National Labor Force Survey (SAKERNAS) data for Maluku Province in February 2023, obtained from BPS Maluku Province. The data points in this study are 3,669 observations with nine variables consisting of 8 independent variables and one dependent variable. The variables in this study are shown in Table 1.

Table 1. Research Variables

No.	Variable Code	Variable Name	Notes	Type of Data
1	X1	Area Classification	1: Urban; 2: Rural	Nominal
2	X2	Marital Status	1: Single; 2: Married; 3: Divorced Alive; 4: Widowed	Nominal
3	X3	Age	1: 15-19 years; 2: 20-24 years; 3: 25-29 years; 4: 30-34 years; 5: 35-39 years; 6: 40-44 years; 7: 45-49 years; 8: 50-54 years; 9: 55-59 years; 10: 60+ years	Nominal
4	X4	Gender	1: Male; 2: Female	Nominal
5	X5	School Participation	1: Not yet in school; 2: Still in school; 3: No longer in school	Nominal
6	X6	Education Level	1: Not/Uncompleted Elementary; 2: Elementary/MI/SDLB/Package A; 3: Junior High/MTS/SMPLB/Package B; 4: Senior High/MA/SMLB/Package C; 5: Vocational School; 6: MAK; 7: Diploma I/II/III; 8: Diploma IV; 9: Bachelor; 10: Master; 11: Applied Master; 12: Doctor	Nominal
7	X7	Training	1: Yes; 2: No	Nominal
8	X8	Work Experience	1: Yes; 2: No	Nominal
9	Y	Unemployment Status	1: Yes; 2: No	Nominal

a. Research Analysis Step

The data analysis used in the performance analysis of random forest classification on the unemployment rate in Maluku Province, based on data balancing methods, consists of several stages, as follows:

1. Input data.
2. Conduct descriptive and exploratory data analysis to get a general overview of the variables to be analyzed.
3. Examine data proportions based on categories (unemployed and employed).
4. Balance data with the ROSE, SMOTE, and ADASYN methods to address data imbalances.
5. Split the data into training and testing data with combinations of 70% training: 30% testing, 80% training: 20% testing, and 90% training: 10% testing.
6. Model the random forest algorithm with training data.
7. Test the model with testing data.

8. Calculate accuracy, sensitivity, specificity, and AUC values with the confusion matrix.
9. Determine the best model by looking at the highest accuracy, sensitivity, specificity, and AUC values.
10. Identify the most important variables with a mean decrease Gini to explain the importance of independent variables in the model.

2.2. Random Forest

Random forest is an extension of the CART method in data mining and does not require assumptions. This method uses the concept of decision trees. This model is built from many trees to form a forest using bootstrap aggregating (bagging) and random feature selection methods [5]. This method can perform classification and regression. Compared to CART, this method uses more than one tree. The classification or regression results of each tree will be voted on to find the most frequent class produced. The term "random" in this method's name means the training data used in building the trees is randomly shuffled [6]. In a random forest, many trees are grown to form a forest; then, a collection of trees is analyzed.

The ntree, mtry, and a_n nodesize parameters are four important parameters to know when building a random forest algorithm [7]. Before constructing the decision tree to be used, the observations are randomly selected with or without the return of the original set. Then, for each cell in the decision tree, a split is performed a_n by maximizing the characters in CART for a uniform mtry between the p-native variables. When each cell contains less than a nodesize point, decision tree formation is stopped. This process will continue until the number of decision trees built reaches a certain number of trees [8].

There are four important parameters in this algorithm:

1. $a_n \epsilon \{1, \ldots., n\}$: The number of data entries sampled in each decision tree.
2. $mtry \epsilon \{1, \ldots., p\}$: The number of variables that are randomly drawn as candidates on each split.
3. $nodesize \epsilon \{1, \ldots., a_n\}$: Minimum size on node terminals (setting a larger number causes a smaller decision tree to be built).
4. $ntree \epsilon N$: Number of decision trees built.

2.3. Imbalance Dataset

An imbalanced dataset, which can be translated as unbalanced data, is a condition in which the distribution of classes in a dataset is unbalanced, and the amount of data in one class is much more or less than in another. The larger group of data classes is called the major class, while the smaller group is called the minor class. Since almost all data will be generalized into major classes, imbalanced data will be challenging to use for classification in machine learning. If tested with the level of accuracy, then the accuracy for the major class will be very high, but the accuracy for the minor class is very poor in some cases, and even nothing is classified as a minor class [9]. This can make the classification performance not good. The method of oversampling in minority classes can improve classification performance. The level of imbalance in the dataset can be measured by calculating the Imbalance Ratio (IR) with the following formula [10]:

$$IR = \frac{m_{majority}}{m_{minority}} \quad (2)$$

with

$m_{majority}$: The amount of data in the major class

$m_{minority}$: The amount of data in the minor class

In this study, the major class is data that is not unemployed, and the minor class is unemployed. Research on unbalanced classes typically assumes the average size of a minority class is less than 40% (or the value of $IR > 1{,}5$). Class imbalance can affect the accuracy of the classification results, so several data balancing methods are needed to overcome this, including the ROSE, SMOTE, and ADASYN methods used in this study.

2.4. Classification Accuracy Measures

The classification methods use test sets to evaluate the quality of the resulting classification. If the label of a row in the test set is the same as the classification result produced by the model, it is called a correct classification. If the label of a row in the test set is different from the classification result produced by the model, it is called a misclassification. Thus, the greater the number of correct classifications, the more accurate the classification, and conversely, the greater the number of misclassifications, the less precise the classification [11].

One tool used to assess the classification's effectiveness is the confusion matrix generated by applying the model to the test set.

Table 2. Confusion Matrix

Actual	Predictions	
	Positive	**Negative**
Positive	True Positive (a)	False Negative (b)
Negative	False Positive (c)	True Negative (d)

Source: [7]

From the confusion matrix, various evaluation measures can be derived, including:

1. Accuracy is the proportion of actual data that is correctly classified overall.

$$Accuracy = \frac{a + d}{a + b + c + d} \tag{3}$$

2. Sensitivity is the proportion of negative actual data that is correctly classified.

$$Sensitivity = \frac{a}{a + b} \tag{4}$$

3. Specificity is the proportion of actual negative data that is correctly classified.

$$Sensitivity = \frac{a}{a + b} \tag{5}$$

4. AUC or *Balanced Accuracy* as an indicator of the performance of the classification model.

$$\text{AUC}/Balance\ Accuracy = \frac{Sensivity + Specificity}{2} \tag{6}$$

In model accuracy assessment, specificity is the ability to predict events with a positive value, while sensitivity is the ability to predict events with a negative value. In addition, accuracy describes the overall level of structural accuracy of models, which is most used to assess classification performance. However, for the case of imbalanced datasets, accuracy places more weight on the majority class, which usually has no significant impact to concern. Therefore, accuracy is less suitable as a measure of classification performance in the case of imbalanced data [12].

In addition, another evaluation measure often used in measuring the performance of classification models, namely AUC [13], which is the probability that a classifier ranks a positive random sample higher than a negative random sample. The curve referred to here is the ROC (Receiver Operating Characteristics) curve, which shows the relationship between sensitivity and specificity. The AUC value ranges from 0 to 1, so the wider it is (closer to 1), the better the classification performance.

Tabel 3. Classification of AUC Values

AUC value	Classification
≥ 0.90	Very good
0.80 – 0.89	Good
0.70 – 0.79	Simply
0.60 – 0.69	Bad
≤ 0.59	Very Bad

Source: [14]

2.5. Variable Importance Level

An important feature of the random forest algorithm is the calculation of variable importance. This algorithm analyzes each attribute and reveals its importance in predicting the correct classification of the random forest machine learner. The researcher can then filter out unnecessary attributes, saving time during data collection and experimental process time. The algorithm first calculates the number of correct untouchables and the number of correct classifications using the dataset as its test. This new dataset is then tested for accurate classification.

Mean Decrease Gini (MDG) is one of the measures of segregation criteria used in the random forest and CART. To calculate MDG, the following formula is used:

$$MDG\ (x_h) = \frac{1}{k}[1 - \sum_k Gini\ (h)^k] \quad (7)$$

$$\sum_k Gini\ (h)^k = 1 - \sum_{j=1}^{p} p_j^2 \quad (8)$$

with

$Gini\ (h)^k$: Gini index for explanatory variables x_h in k[th] tree

k : number of trees in the random forest (random forest size)

p_j : relative frequency of class j in h

p : number of variables

At each split, one of the classes is used to form the split, and a Gini-decline occurs. The sum of all decreases in the forest due to a particular variable, normalized by the number of trees, forms the Gini measure [15]. The MDG number is a number that explains how important an independent variable is in its contribution to the model. The higher the number, the higher the contribution to the model [16].

3. RESULTS AND DISCUSSION

Based on the following SAKERNAS data for February 2023, the number of people with jobs is greater than the number of those who do not have jobs (unemployed) for Maluku Province. This can be seen in Figure 1, where the percentage of the unemployed population is 6.08% while the rate of the working population is 93.92% of the total labor force population.

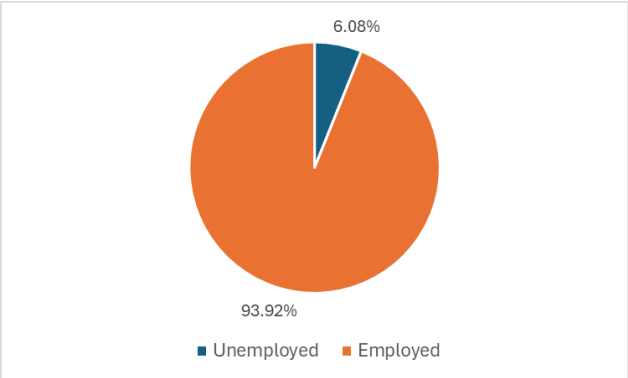

Figure 1. Percentage of Unemployment Status in Maluku Province February 2023

This significant difference indicates that there is unequal data, so the classification model built will be ineffective. So, this research must use data balancing methods. The methods used are the ROSE, SMOTE, and ADASYN methods.

3.1. Unbalance Data Handling

To identify data imbalance, the IR (Imbalance Ratio) formula shown in Equation (2) is used. Thus, the results of the IR calculation based on the proportion of data classes are obtained as follows:

$$IR = \frac{m_{majority}}{m_{minority}} = \frac{3{,}532}{137} = 25.78$$

As reflected in a high IR, significant data imbalance indicates a glaring imbalance. This phenomenon harms the construction of the classification model, and the classification model constructed is ineffective.

To balance the number of samples between minority and majority classes in the dataset, some methods that can be used are ROSE, SMOTE, and ADASYN. ROSE, SMOTE, and ADASYN are oversampling methods used to balance the data by adding synthetic data to the minority class. In this case, the minority class is the smaller unemployment data.

Table 4. Number of data before and after balancing

Description	Before Data Balancing	ROSE	SMOTE	ADASYN
Unemployed	137	1,826	3,425	3,557
Employed	3,532	1,843	3,532	3,532

After balancing the data with ROSE, SMOTE, and ADASYN, the condition of unemployment data in Maluku Province becomes more balanced.

3.2. Modeling with Random Forest

In this section, we will analyze the performance of the random forest classification model without using the data balancing method and the data balancing method, with the proportion of data. The training data is 70% to 90% of the total data, and the testing data used is 30% to 10% of the remaining data. The accuracy, sensitivity, specificity, and AUC values generated from each proportion are presented in Tables 5 to 7.

Table 5. Performance Comparison of the Best Random Forest Classification Model Based on 70% Training Data Share and 30% Testing Data Share

n	Before DataBalancing	ROSE	SMOTE	ADASYN
Sensitivity	0.9568	1.0000	**0.9888**	0.9968
Specificity	0.1667	0.8822	**0.9032**	0.8861
Accuracy	0.9482	0.9346	**0.9435**	0.9342
AUC	0.5157	0.9359	**0.9450**	0.9339

Based on the model performance measure with 70% training data and 30% testing data in Table 5, the results show that the random forest classification model with data balancing using the SMOTE method has the most significant AUC value compared to the random forest classification model with data balancing using other methods. This model can predict the minority class in the classification dataset very well. So that for a division of 70% training data and 30% testing data, the results of accuracy and AUC values with SMOTE are higher than ROSE and ADASYN, thus illustrating that the model after data balancing with SMOTE is better than ROSE and ADASYN in classifying the working status of individuals on SAKERNAS 2023 data in Maluku Province.

Table 6. Performance Comparison of the Best Random Forest Classification Model Based on 80% Training Data Share and 20% Testing Data Share

n	Before databalancing	ROSE	SMOTE	ADASYN
Sensitivity	0.9515	1.0000	**0.9951**	0.9871
Specificity	0.3846	0.8867	**0.8928**	0.8883
Accuracy	0.9414	0.9360	**0.9382**	0.9316
AUC	0.5567	0.9358	**0.9384**	0.9310

Based on the model performance measure with 80% training data and 20% testing data in Table 6, the results show that the random forest classification model with data balancing using the SMOTE method has the largest AUC value compared to the random forest classification model with data balancing using other methods. This model shows its ability to predict the minority class in the classification dataset very well. So that for a division of 80% training data and 20% testing data, the results of the AUC value with SMOTE are higher than those of ROSE and ADASYN, thus illustrating that the model after data balancing with SMOTE is better than ROSE and ADASYN in classifying the working status of individuals in the 2023 SAKERNAS data in Maluku Province.

Table 7. Performance Comparison of the Best Random Forest Classification Model Based on 90% Training Data Share and 10% Testing Data Share

n	Before databalancing	ROSE	SMOTE	ADASYN
Sensitivity	0.9577	1.0000	**0.9968**	0.9935
Specificity	0.3333	0.8850	**0.8974**	0.8947

n	Before databalancing	ROSE	SMOTE	ADASYN
Accuracy	0.9373	0.9373	**0.9425**	0.9379
AUC	0.5938	0.9395	**0.9435**	0.9372

Based on the model performance measure with 90% training data and 10% testing data in Table 7, the results show that the random forest classification model with data balancing using the SMOTE method has the largest AUC value compared to the random forest classification model with data balancing using other methods. This model shows its ability to predict the minority class in the classification dataset very well. So that for a division of 90% training data and 10% testing data, the results of accuracy and AUC values with SMOTE are higher than ROSE and ADASYN, thus illustrating that the model after data balancing with SMOTE is better than ROSE and ADASYN in classifying the working status of individuals on SAKERNAS 2023 data in Maluku Province.

3.3. Variable Importance

One of the outputs of the random forest is the most essential variable. The most important variable is obtained from the average decrease in the Gini index, or Gini decrease index, obtained during the forest formation. This is the most critical variable, which strongly influences the unemployment rate.

The model's Mean Decrease Gini (MDG) value is used to determine the most influential variable in distinguishing the response categories, namely unemployment and non-unemployment. The MDG value determines the order of importance of the independent variables in determining unemployment status in Maluku Province; a higher MDG value indicates that the variable is more important [17].

Table 8. Level of Importance of Independent Variables for Maluku Province in the Best Random Forest Model

Variable Code	Variables	MDG Value
X_6	Education Level	**446.70135**
X_3	Age	442.17618
X_1	Region Classification	362.82259
X_2	Marital Status	356.88711
X_5	School Participation	133.52121
X_4	Gender	104.70148
X_8	Work Experience	90.44749
X_7	Training	67.54829

Table 8 shows that the age variable has the highest MDG value of all the variables. The results align with the results from SAKERNAS 2023 in Maluku Province, which shows that the education level has a significant role in determining who is employed and who is unemployed. Furthermore, the age variable contributes the second-largest difference between employed and unemployed individuals.

4. CONCLUSION

Based on the performance analysis of *random forest* classification on the unemployment rate in Maluku Province based on the data balancing method, the following conclusions are obtained:

1. After balancing the data with Random Over Sampling Example (ROSE), Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN). The condition of unemployment data in Maluku Province has become more balanced. The data balancing method works by overcoming class imbalance in the classification dataset by increasing the number of samples in the minority class.
2. Based on the comparison of data balancing methods, it is concluded that the SMOTE balancing method is better than ROSE and ADASYN. Generally, the data balancing methods using ROSE, SMOTE, and ADASYN make the prediction results more accurate and unbiased towards the majority, thus improving model performance and reducing overfitting. This point is consistent with previous research that SMOTE's more uniform approach to synthetic sample generation, consistent performance across varying imbalance ratios, and strong results combined with popular classification algorithms make it a preferred choice in many scenarios [16].
3. Based on the best *random forest* model produced, the factor that most influences the unemployment rate in Maluku Province is the education level variable, which has the highest MDG value.

ACKNOWLEGEMENT

A big thank you to the Statistics Study Program, Department of Mathematics, Pattimura University, for supporting this research and to Statistics Indonesia (BPS) Maluku Province, which has provided raw data as the primary source to conduct this research.

REFERENCES

[1] Badan Pusat Statistika, *Profil Pengangguran Maluku Agustus 2022*. Badan Pusat Statistika, 2022.

[2] Badan Pusat Statistika, *Booklet Survei Angkatan Kerja Nasional Februari 2023*. Badan Pusat Statistika, 2023.

[3] H. D. Ramadhanti, "Klasifikasi Status NEET pada Penduduk Usia Muda di Indonesia dengan SVM dan Random Forest," *Journal of System and Computer Engineering (JSCE)*, vol. 2, no. 1, pp. 95–105, 2021.

[4] F. Gorunescu, *Data Mining: Concepts, Models and Techniques*, vol. 12. in Intelligent Systems Reference Library, vol. 12. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. doi: 10.1007/978-3-642-19721-5.

[5] T. Purwa, "Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Data Imbalanced (Studi Kasus: Klasifikasi Rumah Tangga Miskin di Kabupaten Karangasem, Bali Tahun 2017)," *Jurnal Matematika, Statistika dan Komputasi*, vol. 16, no. 1, pp. 58–73, 2019.

[6] L. Binarwati, I. Mukhlash, and S. Soetrisno, "Implementasi algoritma genetika untuk optimalisasi random forest dalam proses klasifikasi penerimaan tenaga kerja baru: Studi kasus PT. XYZ," *Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Bachelor Degree Final Project of Department of Mathematics*, 2017.

[7] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, pp. 197–227, 2016.

[8] F. L. Damamain, "Klasifikasi Status Kemiskinan Rumah Tangga di Provinsi Maluku Menggunakan Random Forest," Univeristas Pattimura, 2023.

[9] R. Siringoringo, "Klasifikasi data tidak seimbang menggunakan algoritma SMOTE dan k-nearest neighbor," *Journal Information System Development (ISD)*, vol. 3, no. 1, 2018.

[10] H. Feng, M. Li, X. Hou, and Z. Xu, "Study of network intrusion detection method based on SMOTE sand GBDT," *Application Research of Computers*, vol. 34, no. 12, pp. 3745–3748, 2017.

[11] S. Pramana, B. Yuniarto, S. Mariyah, I. Santoso, and R. Nooraeni, "Data mining dengan R konsep serta implementasi," *Bogor: In Media*, p. 206, 2018.

[12] M. Maalouf and T. B. Trafalis, "Rare events and imbalanced datasets: an overview," *International Journal of Data Mining, Modelling and Management*, vol. 3, no. 4, pp. 375–388, 2011.

[13] W. Nugraha and R. Sabaruddin, "Teknik Resampling untuk Mengatasi Ketidakseimbangan Kelas pada Klasifikasi Penyakit Diabetes Menggunakan C4. 5, Random Forest, dan SVM," *Techno. Com*, vol. 20, no. 3, pp. 352–361, 2021.

[14] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010950718922.

[15] A. Fauzi, M. Rizki, R. Rendi, R. Nurul, T. Novitasari, and R. Nooraeni, "Workforce Classification in West Java 2018 With Random Forest," *Jurnal Matematika, Statistika dan Komputasi*, vol. 17, pp. 240–251, Dec. 2020, doi: 10.20956/jmsk.v17i2.11680.

[16] J. Brandt and E. Lanzén, "A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification," *2021*, p. 42, 2020.