TIERS Information Technology Journal
Vol.4, No.2, December 2023, pp. 122~128
ISSN: 2723-4533 / E-ISSN: 2723-4541
DOI: 10.38043/tiers.v4i2.5069

Resampling Techniques in Rainfall Classification of Banjarbaru using Decision Tree Method

Selvi Annisa[1], Yeni Rahkmawati[2]
selvi.annisa@ulm.ac.id [1], yeni.rahkmawati@ulm.ac.id [2]
[1,2]Department of Statistics, Universitas Lambung Mangkurat, South Kalimantan

ABSTRACT

Continuous heavy rains, such as in 2021, can cause flood emergencies in various areas of Banjarbaru. Therefore, classification modeling is needed to predict rainfall classes based on climate parameters. The problem faced in the classification case is the unbalanced class distribution. Class imbalance occurs when the minority class is much smaller than the majority class. This research aims to compare three resampling techniques in handling imbalanced rainfall data in Banjarbaru using the Decision Tree model. The comparison methods used were sensitivity, specificity, and G-Mean values. In this research, the method used is a decision tree model with Random undersampling, Random Oversampling, and SMOTE. The result shows that the best model is the Decision tree model with the Random Undersampling technique because it provides the highest G-Mean value and sensitivity and specificity values above 70%. Based on this model, the variables that can separate the Rainy and Cloudy classes are Minimum temperature, Maximum temperature, and Sunshine duration, with the best separator being Maximum Temperature.

Keywords: Random Undersampling; Random Oversampling; SMOTE; Decision Tree; Imbalanced Dataset

Article Info

Accepted : 11-12-2023
Revised : 08-09-2023
Published Online : 25-12-2023

Correspondence Author:

Selvi Annisa
Department of Statistics,
Universitas Lambung Mangkurat,
Komplek Griya Pemurus Indah Blok J No 21, Kabupaten Banjar, South Kalimantan, 70654.
Email: selvi.annisa@ulm.ac.id

1. INTRODUCTION

Banjarbaru has been the capital and center of government of South Kalimantan Province since March 16, 2022. With this new status, various activities ranging from government, education, economics, plantations, and development are focused on this city. Bad climate changes can affect various activities in Banjarbaru City. Moreover, continuous heavy rains, such as in 2021, have caused flood emergencies in various areas of Banjarbaru, disrupting the activities of the entire community.

Climate change can be seen in changes in indicators of atmospheric conditions, one of which is rainfall. Rainfall is the cause that most influences the emergence of natural disasters such as floods and landslides [1]. Several factors influencing rainfall include minimum temperature, maximum temperature, average humidity, exposure time, and wind speed [2].

This research categorizes rainfall into two classes, namely Rainy and Cloudy. Therefore, this case is included in the classification case [3]. One of the problems faced in classification cases is unbalanced class distribution. Class imbalance occurs when the minority class is much smaller than the majority class [4].

Building a classification model using imbalanced data will result in low minority prediction accuracy. The majority class information dominates the minority class, causing biased decision boundaries in the classification system [5]. Handling class imbalance in this research is divided into three resampling methods. The first is the Random Undersampling technique, which produces random subsamples from observations from

the majority class to have comparable proportions to the minority class [6]. This technique lowers the cost of learning by offering a well-balanced dataset [7]. The second technique is Random Oversampling, which is increasing the minority class sample until it has a proportion comparable to the majority class [8, 9]. This technique selects samples at random from the data with replacements. As a result, throughout model training, the number of samples in the majority and minority classes is balanced, preventing the majority class from dominating the minority class [10]. The three Synthetic Minority Oversampling Technique (SMOTE) techniques, if the oversampling method has the principle of increasing observations randomly, then the SMOTE method proposed by [11] increases the amount of minority class data so that it is equal to the majority class by generating artificial data. The artificial data is created based on k-nearest neighbors.

In this research, the classification modeling used is a Decision Tree. The classification modeling used is a Decision Tree formed in three stages. The first stage involves selecting and separating variables, with each split depending only on the value of one independent variable. The second stage is the development of the tree by searching for all possible separators to provide the highest heterogeneity reduction value. The third stage labels each terminal node based on the rule of the most significant number of class members [12].

Several studies that use the Decision model to classify rainfall data include rainfall estimates in Central Java and East Java using the C4.5 algorithm with an accuracy of 89.4% [13]. Then research by [14] to model flood-prone areas in Karawang Regency, West Java used the C4.5 algorithm with an accuracy of 84.385%. Furthermore, research by [15] to estimate the potential for rainfall to impact regional flooding used the C4.5 algorithm with an accuracy of 83.33%. The difference between previous research and this research is the addition of resampling techniques before modeling. Based on the previous description, this research aims to compare three resampling techniques, namely Random Undersampling, Random Oversampling, and SMOTE in handling imbalanced rainfall data in Banjarbaru using the Decision Tree model. Apart from that, we will see what predictor variables can separate the Rainy and Cloudy classes based on the model that gives the highest G-Mean value.

2. RESEARCH METHOD

2.1. Data Source

This quantitative study uses daily data from the online data website, the database center of the Meteorology, Climatology and Geophysics Agency (BMKG) from January 2021 to May 2023 which can be accessed at https://dataonline.bmkg.go.id [16]. This data includes rainfall, and the independent variables are Minimum Temperature, Maximum Temperature, Average Humidity, Sunshine Duration, and Wind Speed. Information regarding the data used is in Table 1.

Table 1. Data Description

Variable	Unit
Minimum temperature (X1)	$^{\circ}C$
Maximum temperature (X2)	$^{\circ}C$
Rainfall	mm
Sunshine duration (X3)	hour
Maximum wind speed (X4)	m/s
Wind direction at maximum speed (X5)	°
Average wind speed (X6)	m/s

2.2. Data Analysis

Data analysis in this research follows the Sample, Explore, Modify, Model, and Assess (SEMMA) data mining process with the software used is R version 4.3.2. These include sample datasets from the BMKG website, data exploration, modified datasets with three resampling techniques, data modeling with a Decision Tree, and model evaluation with sensitivity, specificity, and G-Mean. The following are the analysis stages:

1. Preprocessing rainfall data and other variables.
   a. Delete empty data on each variable.
   b. Delete data containing the value 8888 because this means the data is not measurable.
   c. Delete data containing the value 9999 because this means no measurements were taken on that day.
2. Classify rainfall into two classes: rainy and cloudy. If the amount of rainfall is equal to 0, it is classified as Sunny class. If the amount of rainfall is more than 0, it is classified as Rainy class. [17].
3. Explore rainfall data in Banjarbaru.
   a. View the distribution of rainfall using a dot plot [18].
   b. Create a rainfall percentage table to show the proportion between rainy and sunny days and detect imbalances in the number of observations in the two classes.

4. Divide the training data and test data with a ratio of 80% training data and 20% test data.
5. Apply random undersampling to majority classess in the training data [19].
6. Apply random oversampling techniques to minority classes in the training data [7].
7. Apply the SMOTE technique to the majority class in the training data based on the five nearest neighbors [20].
8. Build a decision tree model from training data based on steps 5, 6, and 7 [21].
9. Apply the model from step 8 to the test data.
10. Calculate sensitivity, specificity, and G-Mean values based on steps 8 and 9 [22].
11. Select the best model with the highest G-Mean value and interpret the model [23].

3. RESULTS AND DISCUSSION

3.1. Data Overview

Descriptive analysis of rainfall data in Banjarbaru shows that there are 677 observation data, with the average daily rainfall during the period 01 January 2021 to 31 May 2023 being 11.63 mm. The lowest rainfall was 0 mm, and the highest rainfall was 255.3 mm. The deviation of rainfall data from the average is shown by the standard deviation value of 20.71 mm.

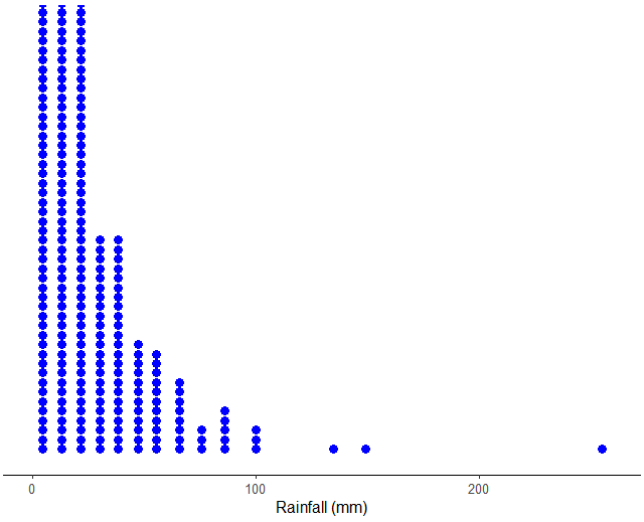

Figure 1. Distribution of Rainfall Data in Banjarbaru

The distribution shows that the rainfall in Banjarbaru is clustered from 0 mm to 20 mm, as shown in Figure 1. One observation, 255.3 mm, has much higher rainfall than the others. Next, the rainfall data is divided into two classes: rainy (rainfall > 0 mm) and sunny (rainfall = 0 mm).

Table 2. Rainfall Class Percentage

Class	Frequency	Percentage (%)
Sunny	160	24.8
Rainy	509	75.2
Total	677	100

Dividing rainfall classes is used to obtain response variables, which will later be used in classification modeling using decision trees. Table 2 shows that the Rainy class has a much greater number of observations than the Sunny class. In this case, the Rainy class is also referred to as the majority class. Meanwhile, the Sunny class is called a minority class.

3.2. Handling Imbalanced Classes

The disproportionate percentage between the Rainy and Sunny classes shows class imbalance in rainfall data. Table 2 shows an imbalance in data classes: only 24.8% of observations are included in the Sunny class. In comparison, the remaining 75.2% of observations are included in the Rainy class. This causes problems when classification modeling is carried out, which is the low classification accuracy value for observations with the Sunny class. The Rainy and Sunny classes have a ratio of 1:3. To balance the data, it is necessary to add data to the Sunny class or subtract data from the Rainy class. In the Random undersampling

technique, data in the majority class is randomly removed around 200% of the minority data, whereas in this case, 272 data are removed from the Rainy class. In the Random oversampling technique, data in the minority class is randomly duplicated around 200% of the minority data, whereas in this case, 270 data are added in the Sunny class. Meanwhile, in the SMOTE technique, the data added is artificial data based on data from the five nearest neighbors.

Based on Table 3, the number of observations in the minority class (Sunny) in the training data is almost comparable to the majority class (Rainy). The next stage is to create a Decision Tree model using training data from the three resampling techniques.

Table 3. Percentage of Rainfall Classes with Resampling Technique in Training Data

Class	Before	Random Undersampling	Random Oversampling	SMOTE
Sunny	135	135	405	405
Rainy	408	136	408	408
Total	543	271	813	813

3.3. Decision Tree Model

Classification modeling using Decision Trees is carried out on four groups of training data: initial training data, training data using the Random Undersampling technique, training data using the Random Oversampling technique, and training data using the SMOTE technique. The model formed is then validated using test data. The Sensitivity, Specificity, and G-Mean obtained in Table 4 are the values from the validation results.

Table 4. Measurement of the Goodness of the Decision Tree Model

Resampling Technique	Sensitivity	Specificity	G-Mean
None	0.5758	0.8812	0.5074
Random Undersampling	0.7273	0.7030	0.5113
Random Oversampling	0.7879	0.5842	0.4603
SMOTE	0.6970	0.7129	0.4969

Based on Table 4, Decision Tree modeling without handling data imbalances delivers the lowest sensitivity and highest specificity values. Decision Tree modeling with Random Oversampling produces the highest sensitivity values and the lowest specificity and G-Mean. On the other hand, Decision Tree with Random Undersampling gives the highest G-Mean value. The smaller the error rate in each class for imbalanced data, the greater the G-mean value [24]. So the best model in this case is Decision Tree modeling with Random Undersampling, apart from providing the highest G-mean value, it also gives sensitivity and specificity values above 70%.

3.4. Model Interpretation

Decision Tree modeling with Random Undersampling provides a tree diagram (Figure 2).

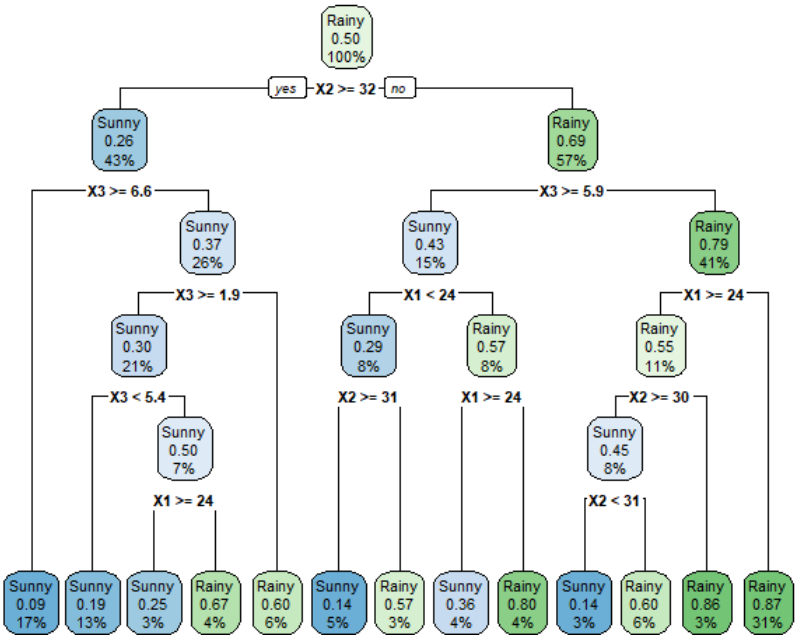

Figure 2. Decision Tree Model Plot

The tree diagram in Figure 2 provides information that the variables that can separate the Rainy and Cloudy classes are Minimum temperature (X1), Maximum temperature (X2), and Sunshine duration (X3), with the best separator being Maximum Temperature (X2). This is in line with research conducted by [25], which found that temperature is one of the factors that influences rainfall. Research by [26] also provides results where sunshine duration is one factor that influences rainfall. Based on Figure 2, the weather is at risk of rain if:

- Maximum temperature at least 32 $°C$, sunshine duration at least 5.4 hours, and minimum temperature less than 24 $°C$
- Maximum temperature at least 32 $°C$ and sunshine duration less than 1.9 hours
- Maximum temperature less than 31 $°C$ and sunshine duration at least 5.9 hours
- Maximum temperature at least 31 $°C$ and less than 32 $°C$, sunshine duration less than 5.9 hours, and minimum temperature at least 24 $°C$
- Maximum temperature less than 32 $°C$, sunshine duration less than 5.9 hours, and minimum temperature less than 24 $°C$

4. CONCLUSION

Classification modeling using a Decision Tree was carried out on four types of training data, with the results that the Random Undersampling technique was the best because it had the highest G-Mean value. Based on this model, the variables that can separate the Rainy and Cloudy classes are Minimum temperature (X1), Maximum temperature (X2), and Sunshine duration (X3), with the best separator being Maximum Temperature (X2).

Other machine learning models, such as Random Forest, SVM, Gradient Boosting, Stacking, etc., are available for further research.

REFERENCES

[1] Siregar, D. C., Ardah, V. P., and Ninggar, R. D, "Identifikasi Kenyamanan Kota Tanjungpinang Berdasarkan Indeks Panas Humidex," *Jurnal Ilmu Lingkungan*, vol. 17, no. 2, Sep., pp. 316-322, 2019. https://doi.org/ 10.14710/jil.17.2.316-322

[2] Gunadi, I. G. A., and Dewi, A. A. K, "Klasifikasi Curah Hujan di Provinsi Bali Berdasarkan Metode Naïve Bayesian," *Wahana Matematika dan Sains: Jurnal Matematika, Sains, dan Pembelajarannya*, vol. 12, no. 1, Apr., pp. 14-25, 2018. https://doi.org/10.23887/wms.v12i1.13843

[3] Wanto, A., et al, *Data Mining: Algoritma dan Implementasi*. Medan: Yayasan Kita Menulis, 2020.

[4] Ren, F., et. al., "Ensemble Based Adaptive over-sampling method for imbalanced data Learning aided detection of microaneurysm," *Computerized Medical Imaging and Graphics*, vol. 55, Jan., pp. 54-67, 2017. https://doi.org/10.1016/j.compmedimag.2016.07.011
[5] Jian, C., Gao, J., and Ao, Y, "A New Sampling Method for Classifying Imbalanced Data Based on Support Vector Machine Ensemble," *Neurocomputing*, vol. 193, June, pp. 115-122, 2016. https://doi.org/10.1016/j.neucom.2016.02.006
[6] Rajesh, K. N. V. P. S., and Dhuli, R, "Classification of imbalanced ECG beats using re-sampling techniques and AdaBoost ensemble classifier," *Biomedical Signal Processing and Control*, vol. 41, Mar., pp. 242–254, 2018. https://doi.org/10.1016/j.bspc.2017.12.004
[7] Holte, R. C, Acker, L., and Porter, B.W, "Concept Learning and the Problem of Small Disjuncts". *InIJCAI*, vol 89, pp. 813-818, 1989.
[8] Gosain, A., and Sardana, S, "Handling class imbalance problem using oversampling techniques: A review". *In 2017 international conference on advances in computing, communications and informatics (ICACCI)*, Sep., pp. 79-85, 2017.
[9] He, H., Zhang, W., and Zhang, S, "A novel ensemble method for credit scoring: adaption of different imbalance ratios," *Expert Systems with Applications*, vol. 98, May, pp. 105-117, 2018. https://doi.org/10.1016/j.eswa.2018.01.012
[10] Kim, A., and Jung, I., "Optimal selection of resampling methods for imbalanced data with high complexity," *PLoS One*, vol. 18, no. 7, Jul, 2023. https://doi.org/10.1371/journal.pone.0288540
[11] Chawla, V. N., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P, "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.
[12] Prasetya, R, "Penerapan Teknik Data Mining dengan Algoritma Classification Tree untuk Prediksi Hujan," *Jurnal Widya Climago*, vol. 2, no. 2, Nov., pp. 13-23, 2020.
[13] Hasanah, M. A, Soim, S., and Handayani, A. S., "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," *Journal of Applied Informatics and Computing*, vol. 5, no. 2, Dec., pp. 103-108, 2021. https://doi.org/10.30871/jaic.v5i2.3200
[14] Khusaeri, A., et al., "Algoritma C4.5 untuk Pemodelan Daerah Rawan Banjir Studi Kasus Kabupaten Karawang Jawa Barat," *ILKOM Jurnal Ilmiah*, vol. 9, no. 2, pp. 132-136, 2017. https://doi.org/10.33096/ilkom.v9i2.128.132-136
[15] Risnawati, I, et al., "Klasifikasi Data Mining Untuk Mengestimasi Potensi Curah Hujan Berdampak Banjir Daerah Menggunakan Algoritma C4.5," *Jurnal INSAN*, vol. 3, no. 2, pp. 78-84, 2023. https://doi.org/10.31294/jinsan.v3i2.3050
[16] Meteorological, Climatological, and Geophysical Agency (BMKG), Onlie Data – Database Center – BMKG, 2023. Available: https://dataonline.bmkg.go.id. [Accessed: June 01, 2023]
[17] Meteorological, Climatological, and Geophysical Agency (BMKG), "Probabilistik Curah Hujan 24 Jam", 2023. Available: https://www.bmkg.go.id/cuaca/probabilistik-curah-hujan.bmkg. [Accessed: June 01, 2023]
[18] Zhao, F., and Gaschler, R, "Best Graph Type to Compare Discrete Groups: Bar, Dot, and Tally," *Frontiers in Psychology*, vol. 12, Dec., 2021. https://doi.org/ 10.3389/fpsyg.2021.775721
[19] Rajesh, K., and Dhuli, R, "Classification Of Imbalanced ECG beats using re-sampling techniques And AdaBoost ensemble classifier," *Biomedical Signal Processing and Control*, vol. 41, Mar., pp. 242-254, 2018. https://doi.org/10.1016/j.bspc.2017.12.004
[20] Elreedy, D., and Atiya, A. F, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Information Sciences*, vol. 505, Dec., pp. 32-64, 2019. https://doi.org/10.1016/j.ins.2019.07.070
[21] Charisma, R. A., et al, "Analisis Penerapan Metode Ensembled Learning Decision Tree Pada Klasifikasi Virus Hepatitis C," *Journal of Computer System and Informatics (JoSYC)*, vol. 3, no. 4, pp. 405-409, 2022. https://doi.org/10.47065/josyc.v3i4.2064
[22] Wegier, W., and Ksieniewicz, P, "Application of Imbalanced Data Classification Quality Metrics as Weighting Methods of the Ensemble Data Stream Classification Algorithms," *Entropy (Basel)*, vol. 22, no. 8, Aug., pp. 849, 2020. https://doi.org/10.3390/e22080849
[23] Sofyan, S., and Prasetyo, A, "Penerapan Synthetic Minority Oversampling Technique (SMOTE) Terhadap Data Tidak Seimbang Pada Tingkat Pendapatan Pekerja Informal Di Provinsi D.I. Yogyakarta Tahun 2019," In Proc. Seminar Nasional Official Statistics, 2021, pp. 868-877.
[24] Ri, J. H., and Kim, H., "G-Mean Based Extreme Learning Machine for Imbalance Learning," *Digital Signal Processing*, vol. 98, March, 2020.
[25] Rohmana, S. F., Rusgiyono, A., and Sugito, "Penentuan Faktor-Faktor Yang Mempengaruhi Intensitas Curah Hujan Dengan Analisis Diskriminan Ganda Dan Regresi Logistik Multinomial (Studi Kasus: Data

Curah Hujan Kota Semarang dari Stasiun Meteorologi Maritim Tanjung Emas Periode Oktober 2018 – Maret 2019)," *Jurnal Gaussian*, vol. 8, no. 3, pp. 398-406, 2019.
[26] Sunarmi, N., et al, "Analisis Faktor Unsur Cuaca terhadap Perubahan Iklim di Kabupaten Pasuruan pada Tahun 2021 dengan Metode Principal Component Analysis," *Newton-Maxwell Journal of Physics*, vol. 3, no. 2, pp. 56-64, 2022.

.