Bulletin of Applied Mathematics and Mathematics Education
Volume 5, Number 1, April 2025 pp. 55-70
https://doi.org/10.12928/bamme.v5i1.13122
e-ISSN 2776-1029, p-ISSN 2776-1002

Comparing the performance of DTID3 and DTID3-Smote methods in predicting the rain events with unbalanced classes

Des A. Zayanti, Ning Eliyati, Yulia Resti*, Sajiril Hoiri, Endang S. Kresnawati, Novi R. Dewi, Ali Amran, Irsyadi Yani

Universitas Sriwijaya, Jl. Raya Palembang-Prabumulih Km. 32, Inderalaya, Ogan Ilir 30662 Indonesia

*Corresponding e-mail: yulia_resti@mipa.unsri.ac.id

ARTICLE INFO

Article History
Received 29 April 2025
Revised 24 May 2025
Accepted 28 May 2025

Keywords
Decision tree method
Unbalanced classes
Weather events

How to cite this article:
Zayanti, D. A., Eliyati, N., Resti, Y., Hoiri, S., Kresnawati, E. S., Dewi, N. R., Amran, A., & Yani, I. (2025). Comparing the performance of DTID3 and DTID3-Smote methods in predicting the rain events with unbalanced classes. *Bulletin of Applied Mathematics and Mathematics Education*, *5*(1), 55-70.

ABSTRACT

Prediction of rainfall events in a region is important for many aspects of life. However, the majority of datasets that predict rainfall events have an unbalanced distribution of observations in their classes, including the Prabumulih city dataset, South Sumatra. DTID3 provides very satisfactory performance in many cases of prediction, while the Smote technique is useful for balancing the distribution of data classes. This study aims to compare the performance of the DTID3 and DTID3-Smote methods in predicting rainfall events in Prabumulih City. The main contribution of this study compared to previous studies is that the DTID3 and Smote methods are used together to predict rainfall events, especially in Prabumulih City. Using training data from 2017-2022 and test data from 2023, the results show that the DTID3-Smote method has a better performance measure than the decision tree method in predicting rainfall events in Prabumulih City. In the decision tree method, the accuracy, precision, recall, specificity, and f1-score metrics are 73.56%, 81.91%, 50.94%, 91.22%, and 62.81%, respectively. In the decision tree-SMOTE method, the values are respectively 74.66%, 82.61%, 53.44%, 91.22%, and 64.9%.

Introduction

DTID3 is a decision tree method based on the Iterative Dichotomiser 3 (ID3) algorithm (Huang & Wang, 2024). DTID3 requires categorical data to determine the probability of determining entropy. Discretization is a technique that transforms numeric data into categorical types (García et al., 2015), (Dougherty et al., 1995). DTID3 is a nonparametric classification method that uses a tree structure representation in making decisions (Matzavela & Alepis, 2021). The branching at each node represents the decision of a variable being the first category or another category (Taha Jijo & Mohsin Abdulazeez, 2021). Each decision is determined based on the gain (Mienye & Jere, 2024). The application of this method in many cases provides very satisfactory performance with metric values of more than 90%. Some examples are the research of Prasad et al., 2025 (rainfall prediction), Kresnawati et al., 2024 (degenerative diseases), Yani & Resti, 2024 (plastic types), Price et al., 2025, Noeman et al., 2022, Deng, 2020 (weather events), Resti et al., 2022 (corn diseases

and pests), Chandra et al., 2023 (air quality), Nicholas et al., 2025, Sondos Jameel Mukhyber, 2025, Amokun et al., 2024, Vijaya Saraswathi et al., 2022, (heart disease), Xiang et al., 2020 (precipitation prediction).

In the case of classification or prediction of observations into certain classes, cases are often encountered where the distribution of observations is not balanced between classes, this imbalance between classes can affect the prediction results (Thölke et al., 2023). Ignoring the problem of class imbalance results in prediction results skewing towards the majority class (Kumar et al., 2021). The accuracy of predictions for minority classes is very important because inaccurate predictions can be fatal or result in very expensive costs (Abdulazeez et al., 2023). Imbalanced data can be found in many cases such as coronary heart disease (Mondal et al., 2025), credit card fraud (Breskuvienė & Dzemyda, 2024), air quality (Chandra et al., 2023), breast cancer (Walsh & Tardy, 2023), educational data mining (Wongvorachan et al., 2023), child occupant crash injury severity (Abdulazeez et al., 2023), and financial statement fraud (Cheng et al., 2022). The Smote method is a resampling method that balances the minority class by adding observations so that they are the same as the number of observations in the majority class (Husain et al., 2025). This method synthesizes new samples from the minority class by identifying vectors between samples from the minority class and samples from selected neighbors (Zhang et al., 2024). Several studies have shown that applying Smote to imbalanced data can improve classification performance, including decision tree methods (Chandra et al., 2023).

Rain event data in Prabumulih City is one of the datasets that have an imbalanced class distribution. Predicting rain events in Prabumulih City is important. It considering that Prabumulih City is one of four cities in South Sumatra Province where most of the land is used for agriculture and plantations (Pratiwi et al., 2021). Although many studies use DTID3 or SMOTE, but are used separately, very few studies combine both for weather prediction in a local context such as Prabumulih City. Therefore, this study aims to compare the performance of the decision tree method and the decision tree method combined with the Smote method (decision tree-Smote) in predicting rain events in Prabumulih City.

Method

The data used in this study is secondary data in the period from January 1, 2017, to December 31, 2023, obtained from https://www.visualcrossing.com/weather/weather-data-services. The data is data on rainfall events in Prabumulih City with sixteen predictor variables as presented in Table 1. These variables are meteorological factors that commonly influence rain events (Sasanya et al., 2022).

Table 1. Predictor and target variables

Variable	Name	Information	Name	Information
Predictor	Max. Temperature (X_1)	25 - 43 °C	Visibility (X_9)	1 - 10 Nm
	Min. Temperature (X_2)	19.3 - 27 °C	UV Index (X_{10})	1.2 – 27.2 Wp
	Ave. Temperature (X_3)	24 – 30.2 °C	Solar energy (X_{11})	61 – 96.9 %
	Max. Feels like (X_4)	25 - 50.9 °C	Humidity (X_{12})	5.4 – 177.8 Km/h
	Min. Feels like (X_5)	19.3 – 30.6 °C	Wind Speed (X_{13})	24.1 – 35.8 °C
	Ave. Feels like (X_6)	17.6 – 25.9 °C	Cloud Layer (X_{14})	0 - 100 Octa
	Dew (X_7)	0 – 359.6 °C	Solar Radiation (X_{15})	14.7 – 315.1 Nm
	Wind Direction (X_8)	1.8 – 18.7 Nm	Moon Phase (X_{16})	0 – 0.98 %
Response	Rain Event (Y)	No (69.6%)		
		Yes (30.4%)		

10.12928/bamme.v5i1.13122

The stages carried out in this research are as follows:

(1) Discretize the predictor variables using the equation (Resti et al., 2022):

$$X_d = {X_d}^o + \text{Range}(X_d) \quad (1)$$

For ${X_d}^o$ be the d-th original predictor variable. X_d is variable ${X_d}^o$ which is discretized as much as $k(X_d)$, and:

$$\text{Range}(X_d) = \frac{\max({X_d}^o) - \min({X_d}^o)}{k(X_d)} \quad (2)$$

(2) Divide the data into training data and test data. Training data for modeling is 70% (January 1, 2017 – December 31, 2022) and test data is 30% (January 1, 2023 - December 31, 2023). The selection of 2023 data as test data is based on the fact that this data is the latest data in the dataset.

(3) Balancing the distribution of observations between the No Rain and Rain classes on the training data using the Smooth technique. This technique replicates the data in the minority class so that it is balanced with the data in the majority class using equation (3) where for each X_d apply (Zhang et al., 2024).

$$X_{new} = X_i + \left(\hat{X}_i - X_i\right).\delta \quad (3)$$

(4) Membangun model prediksi menggunakan metode decision tree dan metode decision tree-Smote. The decisions in these methods are obtained using equations (4), (5), and (6) which are the gain $G(Y, X_d)$, entropy for target and predictor variables, $E\big(S(Y)\big)$ and $E(X_d^m)$, respectively. The p_j and p_m are the probabilities of the j-th class and the m-th category respectively.

$$G(Y, X_d) = E\big(S(Y)\big) - \sum_{m=1}^{k(X_d)} \frac{S(X_d^m)}{S(Y)} E(X_d^m) \quad (4)$$

$$E\big(S(Y)\big) = -\sum_{j=1}^{k(Y)} p_j * \log_2 p_j \quad (5)$$

$$E(X_d^m) = -\sum_{m=1}^{k(X_d)} p_m * \log_2 p_m \quad (6)$$

(5) Make predictions using test data based on each model formed in Step 4.

(6) Form a confusion matrix for each prediction result as in Table 2.

Table 2. Confusion matrix

		Prediction	
	Class	No Rain	Rain
Actual	No Rain	True Negatif	False Positive
	Rain	False Negative	True Positive

(7) Calculate and compare the metric values for each confusion matrix: accuracy, precision, recall, specificity, and F1-score.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (7)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (8)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (9)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (10)$$

$$\text{F1} - \text{Score} = \frac{2(\text{PrecisionRecall})}{\text{Precision} + \text{Recall}} \quad (11)$$

Results and discussion

Data discretization

Discretization uses Equation (1), with the number of categories for each predictor variable varying according to the classifications commonly used by experts. For example, the variables temperature and feels like are typically categorized into three groups: cold, moderate, and hot.

Table 3. Discretization of predictor variables

Variabel	Category	Interval	Variabel	Category	Interval
X_1	1	$25 \leq X_1 \leq 31$	X_{10}	1	$1 \leq X_9 \leq 2.8$
	2	$31.1 \leq X_1 \leq 37$		2	$2.9 \leq X_9 \leq 4.6$
	3	$37.1 \leq X_1 \leq 43$		3	$4.7 \leq X_9 \leq 6.4$
X_2	1	$19.3 \leq X_2 \leq 21.9$		4	$6.5 \leq X_9 \leq 8.2$
	2	$22 \leq X_2 \leq 24.4$		5	$8.3 \leq X_9 \leq 10$
	3	$24.5 \leq X_2 \leq 27$	X_{11}	1	$1.2 \leq X_{10} \leq 9.9$
X_3	1	$24 \leq X_3 \leq 26.1$		2	$10 \leq X_{10} \leq 18.5$
	2	$26.2 \leq X_3 \leq 28.1$		3	$18.6 \leq X_{10} \leq 27.2$
	3	$28.2 \leq X_3 \leq 30.2$	X_{12}	1	$61 \leq X_{11} \leq 73$
X_4	1	$25 \leq X_4 \leq 33.6$		2	$73.1 \leq X_{11} \leq 84.9$
	2	$33.7 \leq X_4 \leq 42.3$		3	$85 \leq X_{11} \leq 96.9$
	3	$42.4 \leq X_4 \leq 50.9$	X_{13}	1	$5.4 \leq X_{12} \leq 24.6$
X_5	1	$19.3 \leq X_5 \leq 23.1$		2	$18.7 \leq X_{12} \leq 31.9$
	2	$23.2 \leq X_5 \leq 26.8$		3	$32 \leq X_{12} \leq 45.18$
	3	$26.9 \leq X_5 \leq 30.6$		4	$45.19 \leq X_{12} \leq 58.44$
X_6	1	$24.1 \leq X_6 \leq 28$		5	$58.45 \leq X_{12} \leq 71.7$
	2	$28.1 \leq X_6 \leq 31.9$		6	$71.8 \leq X_{12} \leq 84.9$
	3	$31.91 \leq X_6 \leq 35.8$		7	$85 \leq X_{12} \leq 98.2$
X_7	1	$17.6 \leq X_6 \leq 18.64$		8	$98.3 \leq X_{12} \leq 111.4$
	2	$18.65 \leq X_6 \leq 19.68$		9	$111.45 \leq X_{12} \leq 124.75$
	3	$19.69 \leq X_6 \leq 20.71$		10	$124.76 \leq X_{12} \leq 138$
	4	$20.72 \leq X_6 \leq 21.75$		11	$138.1 \leq X_{12} \leq 151.27$
	5	$21.76 \leq X_6 \leq 22.79$		12	$151.28 \leq X_{12} \leq 164.53$
	6	$22.8 \leq X_6 \leq 23.83$		13	$164.54 \leq X_{12} \leq 177.8$
	7	$23.84 \leq X_6 \leq 24.86$	X_{14}	1	$27.4 \leq X_{14} \leq 50.13$
	8	$24.87 \leq X_6 \leq 25.9$		2	$50.14 \leq X_{14} \leq 72.86$
X_8	1	$0 \leq X_7 \leq 45$		3	$72.87 \leq X_{14} \leq 95.6$
	2	$45.1 \leq X_7 \leq 89.9$	X_{15}	1	$14.7 \leq X_{15} \leq 114.83$
	3	$90 \leq X_7 \leq 134.9$		2	$114.84 \leq X_{15} \leq 215$
	4	$135 \leq X_7 \leq 179.8$		3	$215.1 \leq X_{15} \leq 315.1$
	5	$179.9 \leq X_7 \leq 224.8$	X_{16}	1	$0 \leq X_{16} \leq 0.245$
	6	$224.9 \leq X_7 \leq 269.7$		2	$0.246 \leq X_{16} \leq 0.49$
	7	$269.8 \leq X_7 \leq 314.7$		3	$0.491 \leq X_{16} \leq 0.734$
	8	$314.8 \leq X_7 \leq 359.6$		4	$0.736 \leq X_{16} \leq 0.98$
X_9	1	$1.8 \leq X_8 \leq 6$			
	2	$6.1 \leq X_8 \leq 10.3$			
	3	$10.4 \leq X_8 \leq 14.5$			
	4	$14.6 \leq X_8 \leq 18.7$			

10.12928/bamme.v5i1.13122

Another example is the wind speed variable, which can be categorized into thirteen levels such as calm, light air, light breeze, gentle breeze, moderate breeze, fresh breeze, strong breeze, near gale, gale, strong gale, storm, violent storm, and hurricane. Discretization of predictor variables in this work is given in Table 3, while the data distribution for each category in ech predictor variable mentioned in Table 3 is presented in Table 4.

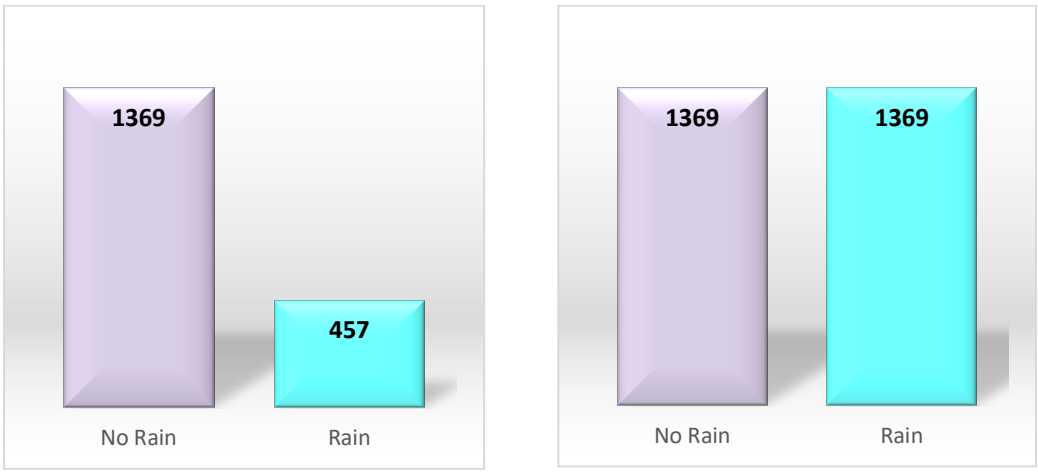
Tabel 4. Distribution of categories on predictor variables in training data

Var.	Cat.	Class of target variable		Total	Var.	Cat.	Class of target variable		Total
		No Rain	Rain				No Rain	Rain	
Y		1369	457	1826	X_{10}	1	7	0	7
X_1	1	497	45	542		2	170	11	181
	2	871	409	1280		3	456	55	511
	3	1	3	4		4	533	184	717
X_2	1	25	89	114		5	203	207	410
	2	1270	340	1610	X_{11}	1	112	3	115
	3	74	28	102		2	832	138	970
X_3	1	175	16	191		3	425	316	741
	2	1090	320	1410	X_{12}	1	7	34	41
	3	104	121	225		2	115	203	318
X_4	1	70	6	76		3	908	214	1122
	2	1293	446	1739		4	339	6	345
	3	6	5	11	X_{13}	1	899	246	1145
X_5	1	452	275	727		2	444	206	650
	2	703	281	1099		3	19	3	22
	3	0	0	0		4	4	0	4
X_6	1	624	140	764		5	0	1	1
	2	703	281	984		6	0	1	1
	3	42	36	78		7	0	0	0
X_7	1	0	2	2		8	1	0	1
	2	0	0	0		9	0	0	0
	3	1	19	20		10	0	0	0
	4	5	42	47		11	1	0	1
	5	47	101	148		12	0	0	0
	6	492	152	644		13	1	0	1
	7	734	124	858	X_{14}	1	35	96	131
	8	90	17	107		2	901	325	1226
X_8	1	78	9	87		3	433	36	469
	2	48	14	62	X_{15}	1	119	3	122
	3	327	247	574		2	832	138	970
	4	259	112	371		3	418	316	734
	5	55	6	61	X_{16}	1	341	105	446
	6	66	8	74		2	339	122	461
	7	213	37	250		3	350	99	449
	8	323	24	347		4	339	131	470
X_9	1	79	35	114					
	2	1290	422	1712					
	3	0	0	0					
	4	0	0	0					

The probability of each predictor variable for each 'No Rain' and 'Rain' event can be obtained from Table 4.

Class balancing using smote technique

The distribution between the No Rain and Rain classes in the training data is 74.97% and 25.03% with a total of 1826 observations. After being balanced using the Smote technique to 50% each, the total observations become 2738. Changes in the distribution of each class in the training data that has been balanced using the Smote formula in equation (1) are presented in Figure 1.



(a) (b)

Figure 1. The before (a) and the after class balancing (b) for No Rain and Rain events.

Modeling of decision tree

Modeling using the decision tree model begins with determining the predictor variables that become the root node. Determination of the root node of predictor variables uses the entropy and gain values. Both values from each variable in the training data before balancing are presented in Table 5.

Table 5. Entropy and gain of predictor variable for root node

Variable	Category of predictor variable	Class of target variable		Total	p(No Rain)	p(Rain)	Entropy	Gain
		No Rain	Rain					
Y		1369	457	1826	0.750	0.25	0.812	
X_1	1	497	45	542	0.917	0.083	0.413	0.054
	2	871	409	1280	0.681	0.320	0.904	
	3	1	3	4	0.250	0.750	0.811	
X_2	1	25	89	114	0.219	0.781	0.759	0.061
	2	1270	340	1610	0.789	0.211	0.744	
	3	74	28	102	0.726	0.275	0.848	
X_3	1	175	16	191	0.916	0.084	0.415	0.049
	2	1090	320	1410	0.773	0.227	0.773	
	3	104	121	225	0.462	0.538	0.996	
X_4	1	70	6	76	0.921	0.079	0.399	0.007
	2	1293	446	1739	0.744	0.257	0.821	
	3	6	5	11	0.546	0.455	0.994	

10.12928/bamme.v5i1.13122

Variable	Category of predictor variable	Class of target variable		Total	p(No Rain)	p(Rain)	Entropy	Gain
		No Rain	Rain					
X_5	1	452	275	727	0.622	0.378	0.957	0.041
	2	917	182	1099	0.834	0.166	0.648	
	3	0	0	0	0	0	0	
X_6	1	624	140	764	0.817	0.183	0.687	0.017
	2	703	281	984	0.714	0.286	0.863	
	3	42	36	78	0.539	0.462	0.996	
X_7	1	0	2	2	0	1	0	0.128
	2	0	0	0	0	0	0	
	3	1	19	20	0.050	0.950	0.286	
	4	5	42	47	0.106	0.894	0.489	
	5	47	101	148	0.318	0.682	0.902	
	6	492	152	644	0.764	0.236	0.788	
	7	734	124	858	0.856	0.145	0.596	
	8	90	17	107	0.841	0.159	0.632	
X_8	1	78	9	87	0.897	0.103	0.480	0.087
	2	48	14	62	0.774	0.226	0.771	
	3	327	247	574	0.570	0.430	0.986	
	4	259	112	371	0.698	0.302	0.884	
	5	55	6	61	0.902	0.098	0.464	
	6	66	8	74	0.892	0.108	0.494	
	7	213	37	250	0.852	0.148	0.605	
	8	323	24	347	0.931	0.069	0.363	
X_9	1	79	35	114	0.693	0.307	0.890	0.001
	2	1290	422	1712	0.754	0.247	0.806	
	3	0	0	0	0	0	0	
	4	0	0	0	0	0	0	
X_{10}	1	7	0	7	1	0	0	0.094
	2	170	11	181	0.939	0.061	0.331	
	3	456	55	511	0.892	0.108	0.493	
	4	533	184	717	0.743	0.257	0.822	
	5	203	207	410	0.495	0.505	0.999	
X_{11}	1	112	3	115	0.974	0.026	0.174	0.088
	2	832	138	970	0.858	0.142	0.590	
	3	425	316	741	0.574	0.427	0.984	
X_{12}	1	7	34	41	0.171	0.829	0.659	0.177
	2	115	203	318	0.362	0.638	0.944	
	3	908	214	1122	0.809	0.191	0.703	
	4	339	6	345	0.983	0.017	0.127	
X_{13}	1	899	246	1145	0.785	0.215	0.751	0.013
	2	444	206	650	0.683	0.317	0.901	
	3	19	3	22	0.864	0.136	0.575	
	4	4	0	4	1	0	0	
	5	0	1	1	0	1	0	
	6	0	1	1	0	1	0	
	7	0	0	0	0	0	0	
	8	1	0	1	1	0	0	
	9	0	0	0	0	0	0	
	10	0	0	0	0	0	0	

Variable	Category of predictor variable	Class of target variable		Total	p(No Rain)	p(Rain)	Entropy	Gain
		No Rain	Rain					
	11	1	0	1	1	0	0	
	12	0	0	0	0	0	0	
	13	1	0	1	1	0	0	
X_{14}	1	35	96	131	0.267	0.733	0.837	0.091
	2	901	325	1226	0.735	0.265	0.833	
	3	433	36	469	0.923	0.077	0.391	
X_{15}	1	119	3	122	0.975	0.025	0.167	0.091
	2	832	138	970	0.858	0.142	0.590	
	3	418	316	734	0.570	0.431	0.986	
X_{16}	1	341	105	446	0.765	0.235	0.787	0.002
	2	339	122	461	0.735	0.265	0.834	
	3	350	99	449	0.780	0.221	0.761	
	4	339	131	470	0.721	0.279	0.854	

In Table 5, it can be seen that the highest gain value is the variable X_{12} (humidity) of 0.177 so that the variable X_{12} is the root node. This variable has 4 categories so that its node has 4 branches that become nodes for the next decision (See Figure 2).

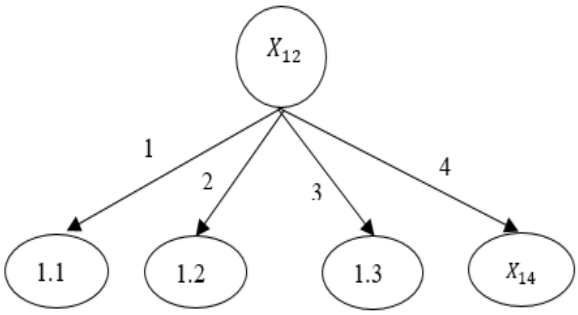


Figure 2. Branching of node X_{12} as root

Next, determine the branches for each first node after the root node. The following describes the determination of branches for the 4th category at the first node after the root node (node 1.4). The determination of branches for the other three categories is obtained in the same way. In this 4th category, there are 345 events consisting of 339 No Rain events and 6 Rain events. The entropy and gain values for node 1.4 can be seen in Table 6.

Table 6. Entropy and gain of node X_{14} for category 2

Variable	Category of predictor variable	Class of target variable		Total	p(No Rain)	p(Rain)	Entropy	Gain
		No Rain	Rain					
Y		339	6	345	0.983	0.017	0.127	
X_1	1	265	6	271	0.978	0.022	0.153	0.006
	2	74	0	74	1	0	0	
	3	0	0	0	0	0	0	
X_2	1	5	1	6	0.833	0.167	0.650	0.008
	2	309	4	313	0.987	0.013	0.099	
	3	25	1	26	0.962	0.039	0.235	

10.12928/bamme.v5i1.13122

Variable	Category of predictor variable	Class of target variable		Total	p(No Rain)	p(Rain)	Entropy	Gain
		No Rain	Rain					
X_3	1	137	2	139	0.986	0.014	0.109	0.003
	2	202	4	206	0.981	0.019	0.138	
	3	0	0	0	0	0	0	
X_4	1	54	2	56	0.964	0.036	0.222	0.023
	2	284	4	288	0.986	0.014	0.106	
	3	1	0	1	1	0	0	
X_5	1	94	2	96	0.979	0.021	0.146	0.002
	2	245	4	249	0.984	0.016	0.119	
	3	0	0	0	0	0	0	
X_6	1	287	6	293	0.980	0.021	0.144	0.004
	2	51	0	51	1	0	0	
	3	1	0	1	1	0	0	
X_7	1	0	0	0	0	0	0	0.002
	2	0	0	0	0	0	0	
	3	0	0	0	0	0	0	
	4	0	0	0	0	0	0	
	5	3	0	3	1	0	0	
	6	71	2	73	0.973	0.027	0.181	
	7	219	3	222	0.987	0.014	0.103	
	8	46	1	47	0.979	0.021	0.149	
X_8	1	13	0	13	1	0	0	0.005
	2	6	0	6	1	0	0	
	3	70	1	71	0.986	0.014	0.107	
	4	60	1	61	0.984	0.016	0.121	
	5	16	0	16	1	0	0	
	6	17	1	18	0.944	0.056	0.310	
	7	61	1	62	0.984	0.016	0.119	
	8	96	2	98	0.980	0.020	0.144	
X_9	1	47	0	47	1	0	0	0.004
	2	292	6	298	0.980	0.020	0.142	
	3	0	0	0	0	0	0	
	4	0	0	0	0	0	0	
X_{10}	1	4	0	4	1	0	0	0.012
	2	85	0	85	1	0	0	
	3	135	2	137	0.9854	0.015	0.110	
	4	101	4	105	0.9619	0.038	0.234	
	5	14	0	14	1	0	0	
X_{11}	1	64	0	64	0	0	0	0.001
	2	225	5	230	0.9783	0.022	0.151	
	3	50	1	51	0.9804	0.020	0.139	
X_{13}	1	218	4	222	0.982	0.018	0.130	0.001
	2	113	2	115	0.9826	0.017	0.127	
	3	6	0	6	1	0	0	
	4	2	0	2	1	0	0	
	5	0	0	0	0	0	0	
	6	0	0	0	0	0	0	
	7	0	0	0	0	0	0	
	8	0	0	0	0	0	0	

Variable	Category of predictor variable	Class of target variable		Total	p(No Rain)	p(Rain)	Entropy	Gain
		No Rain	Rain					
	9	0	0	0	0	0	0	
	10	0	0	0	0	0	0	
	11	0	0	0	0	0	0	
	12	0	0	0	0	0	0	
	13	0	0	0	0	0	0	
X_{14}	1	1	1	2	0.5	0.5	1	0.026
	2	168	5	173	0.9711	0.029	0.189	
	3	170	0	170	1	0	0	
X_{15}	1	66	0	66	1	0	0	0.005
	2	223	5	228	0.9781	0.022	0.152	
	3	50	1	51	0.9804	0.020	0.139	
X_{16}	1	93	1	94	0.9894	0.011	0.085	0.008
	2	83	3	86	0.9651	0.035	0.218	
	3	68	0	68	1	0	0	
	4	95	2	97	0.9794	0.021	0.145	

In Table 6, it is known that the highest gain value is owned by the variable X_{14} (Cloud Layer) of 0.026. This variable has 3 categories expressed as 3 branches on the Decision tree. In category 3, the entropy value is zero, meaning that this node is at the final node (terminal node) which indicates that a final decision has been formed, and because the probability of a No Rain event is greater than the probability of a Rain event, the final decision is a No Rain event (See Figure 3).

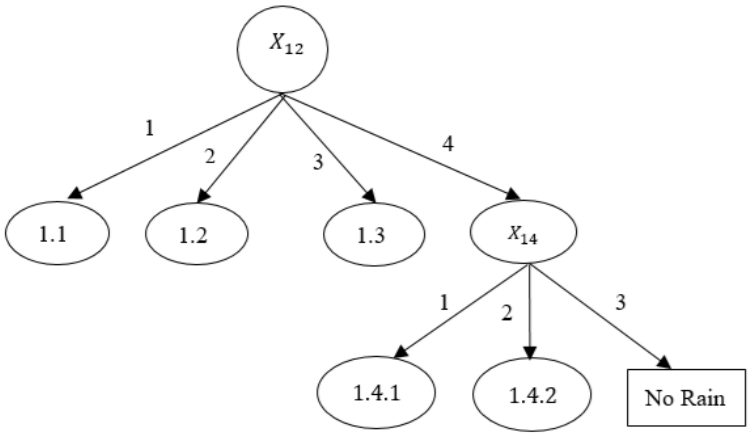


Figure 3. Branching of node X_{14} for category 3.

Table 7 presents the entropy and gain for the second category of variable X_{14} which has a total number of events of 173, distributed into No Rain and Rain events of 168 and 5 respectively.

Table 7. Entropy and gain of node X_{14} for category 2

Variable	Category of predictor variable	Class of target variable		Total	p(No Rain)	p(Rain)	Entropy	Gain
		No Rain	Rain					
Y		168	5	173	0.9711	0.0289	0.1889	
X_1	1	127	5	132	0.9621	0.0379	0.2325	0.0115
	2	41	0	41	1	0	0	
	3	0	0	0	0	0	0	

10.12928/bamme.v5i1.13122

Variable	Category of predictor variable	Class of target variable		Total	p(No Rain)	p(Rain)	Entropy	Gain
		No Rain	Rain					
X_2	1	4	1	5	0.8	0.2	0.7219	0.0135
	2	149	4	153	0.9739	0.0262	0.1747	
	3	15	0	15	1	0	0	
X_3	1	57	2	59	0.9661	0.0339	0.2136	0.0003
	2	111	3	114	0.9737	0.0263	0.1756	
	3	0	0	0	0	0	0	
X_4	1	26	2	28	0.9286	0.0714	0.3712	0.0071
	2	141	3	144	0.9792	0.0208	0.1461	
	3	1	0	1	1	0	0	
X_5	1	44	2	46	0.9565	0.0435	0.2580	0.0018
	2	124	3	127	0.9764	0.0236	0.1613	
	3	0	0	0	0	0	0	
X_6	1	137	5	142	0.9648	0.0352	0.2199	0.0084
	2	30	0	30	1	0	0	
	3	1	0	1	1	0	0	
X_7	1	0	0	0	0	0	0	0.0113
	2	0	0	0	0	0	0	
	3	0	0	0	0	0	0	
	4	0	0	0	0	0	0	
	5	2	0	2	1	0	0	
	6	32	2	34	0.9412	0.0588	0.3228	
	7	104	3	107	0.9720	0.0280	0.1845	
	8	30	0	30	1	0	0	
X_8	1	6	0	1	1	0	0	0.0094
	2	5	0	11	1	0	0	
	3	27	1	28	0.9643	0.0357	0.2223	
	4	34	1	35	0.9714	0.0286	0.1872	
	5	11	0	11	1	0	0	
	6	10	0	10	1	0	0	
	7	32	1	33	0.9697	0.0303	0.1959	
	8	43	2	45	0.9556	0.0444	0.2623	
X_9	1	26	0	26	1	0	0	0.0069
	2	142	5	150	0.966	0.034	0.2141	
	3	0	0	0	0	0	0	
	4	0	0	0	0	0	0	
X_{10}	1	1	0	1	1	0	0	0.0141
	2	34	0	34	1	0	0	
	3	65	2	67	0.9701	0.0299	0.1936	
	4	58	3	61	0.9508	0.0492	0.2829	
	5	10	0	10	1	0	0	
X_{11}	1	23	0	23	1	0	0	0.00002
	2	113	4	117	0.9658	0.0342	0.2150	
	3	32	1	33	0.9697	0.0303	0.1959	
X_{13}	1	91	3	94	0.9681	0.0319	0.2039	0.0018
	2	70	2	72	0.9722	0.0278	0.1831	
	3	5	0	5	1	0	0	

Variable	Category of predictor variable	Class of target variable		Total	p(No Rain)	p(Rain)	Entropy	Gain
		No Rain	Rain					
	4	2	0	2	1	0	0	
	5	0	0	0	0	0	0	
	6	0	0	0	0	0	0	
	7	0	0	0	0	0	0	
	8	0	0	0	0	0	0	
	9	0	0	0	0	0	0	
	10	0	0	0	0	0	0	
	11	0	0	0	0	0	0	
	12	0	0	0	0	0	0	
	13	0	0	0	0	0	0	
X_{15}	1	24	0	24	1	0	0	0.0064
	2	112	4	116	0.9655	0.0345	0.2164	
	3	32	1	33	0.9697	0.0303	0.1959	
X_{16}	1	46	1	47	0.9787	0.0213	0.1485	0.0170
	2	41	3	44	0.9318	0.0682	0.3591	
	3	36	0	36	1	0	0	
	4	45	1	46	0.9783	0.0217	0.1511	

Table 7 shows that the variable X16 with four categories has the highest Gain of 0.017. In this variable, the third category has an entropy of 0 which indicates that a final decision has been formed. The probability of a No Rain event is greater than the probability of a Rain event indicating that the final decision is a No Rain event (Figure 4).

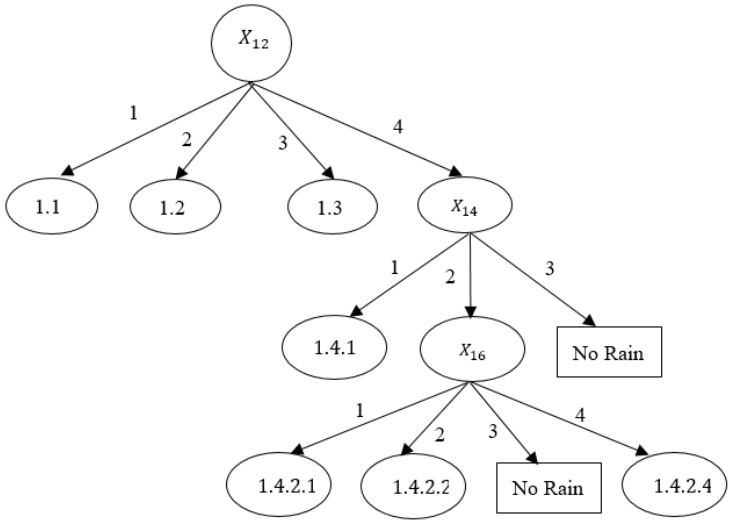


Figure 4. Branching of node X_{14} for category 2.

Branching of each node continues in the same way until all nodes end at the terminal node which is the decision of Rain or No Rain event, and forms a decision tree. Furthermore, the decision tree that is formed produces If-Then rules. These rules are used to predict test data. The prediction results are then tabulated into a confusion matrix as presented in Table 8.

10.12928/bamme.v5i1.13122

Table 8. Confusion matrix of decision tree model

		Prediction		
	Class	No Rain	Rain	Total
Actual	No Rain	374	36	410
	Rain	157	163	320
	Total	531	199	730

From Table 8, it is obtained that 374 No Rain events were predicted correctly from 410 actual No Rain events while 36 No Rain events were predicted as Rain events. 163 Rain events were predicted correctly from 320 actual Rain events while 157 Rain events were predicted as No Rain events. The prediction results using this decision tree model have accuracy, precision, recall, specificity, and f1-score performance of 73.56%, 81.91%, 50.94%, 91.22%, and 62.81%, respectively. Accuracy indicates the percentage of Rain and No Rain events that are predicted correctly from all events. Precision is the percentage of events that are correctly Rain from all events that are predicted as Rain. The percentage of Rain events that are predicted correctly compared to all events that are Rain is represented by the Recall metric. Conversely, the percentage of No Rain events that are predicted correctly compared to all events that are No Rain is represented by the Specificity metric.

Modeling using the decision tree-Smote model is obtained in the same way but using data that has been balanced in class distribution using the Smote technique. In this model, the predictor variable that becomes the root node is the variable X12. However, the final prediction results are slightly different and the difference can be seen in the confusion matrix as presented in Table 9.

Table 9. Confusion matrix of decision tree model with Smote

		Prediction		
	Class	No Rain	Rain	Total
Actual	No Rain	374	36	410
	Rain	149	171	320
	Total	523	207	730

From Table 9, it is obtained that 379 No Rain events were predicted correctly from 410 actual No Rain events while 31 No Rain events were predicted as Rain events. 164 Rain events were predicted correctly from 320 actual Rain events while 156 Rain events were predicted as No Rain events. The prediction results using the decision tree-SMOTE method, the accuracy, precision, recall, specificity, and f1-score values were 74.66%, 82.61%, 53.44%, 91.22%, and 64.9% respectively. In predicting No Rain event, DTID3 and DTID3-Smote methods have the same performance of 91.22%. However, in predicting Rain events, DTID3-Smote has a high performance of 82.61%. Generally, the decision tree-Smote method has better performance than the decision tree method in predicting rainfall events in Prabumulih City. However, the use of Smote can result in overfitting due to synthetic data (Bunkhumpornpat et al., 2024).

The existence of this research is expected to help related parties in this case the Meteorology, Climatology, and Geophysics Agency in predicting rain events. This agency is a Non-Ministerial Government Institution (LPNK) of Indonesia which has the task of carrying out government duties in the fields of meteorology, climatology, and geophysics, including rain events.

Conclusion

Rainfall prediction is important for Prabumulih City considering that most of its area is used for agriculture and plantations. The Prabumulih City rainfall dataset has an unbalanced distribution of observations in its classes. DTID3 provides very satisfactory performance in many cases of prediction, while the Smote technique is useful for balancing the distribution of data classes. The main contribution of this study compared to previous studies is that the DTID3 and Smote methods are used together to predict rainfall events, especially in Prabumulih City. The results show that the DTID3-SMOTE method has better performance than the DTID3 method in predicting rainfall events in Prabumulih City for training data is 2017-2022 and test data is 2023. This conclusion refers to the metric measurements of accuracy, precision, recall, specificity, and f1-score. In the DTID3 method, the values of these metric measurements are 73.56%, 81.91%, 50.94%, 91.22%, and 62.81%, respectively. In the DTID3-SMOTE method, the values are 74.66%, 82.61%, 53.44%, 91.22%, and 64.9% respectively. Several other prediction methods can be used together with the Smote method, such as naïve Bayes, support vector machines, logistic regression and so on, which are expected to provide better prediction performance.

Acknowledgment

This work was supported by the DIPA of Universitas Sriwijaya 2024 Public Service Agency, SP DIPA-023.17.2.677515/2024, on November 24, 2023, and the Rector's Decree 0012/UN9/SK.LP2M. PT/2024, May 20, was issued.

References

Abdulazeez, M. U., Khan, W., & Abdullah, K. A. (2023). Predicting child occupant crash injury severity in the United Arab Emirates using machine learning models for imbalanced dataset. *International Association of Traffic and Safety Sciences*, *47*(2), 134–159. https://doi.org/10.1016/j.iatssr.2023.05.003

Amokun, R., Arowolo, O. T., & Eke, J. (2024). Comparative analysis of machine learning algorithms for heart disease prediction. *The International Conference on Artificial Intelligence and Robotics (MIRG-ICAIR 2024)*, *November*, 107–117. https://doi.org/10.56726/irjmets59893

Bunkhumpornpat, C., Boonchieng, E., Chouvatut, V., & Lipsky, D. (2024). FLEX-SMOTE: Synthetic over-sampling technique that flexibly adjusts to different minority class distributions

Breskuvienė, D., & Dzemyda, G. (2024). Enhancing credit card fraud detection: highly imbalanced data case. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-024-01059-5

Chandra, W., Suprihatin, B., & Resti, Y. (2023). Median-KNN Regressor-SMOTE-Tomek Links for Handling Missing and Imbalanced Data in Air Quality Prediction. *Symmetry*, *15*(4), 887. https://doi.org/10.3390/sym15040887

Cheng, Q., Xu, H., Fei, S., Li, Z., & Chen, Z. (2022). Estimation of Maize LAI using ensemble learning and UAV multispectral imagery under different water and fertilizer treatments. *Agriculture*, *12*(8), 1267. https://doi.org/10.3390/agriculture12081267

Deng, F. (2020). Research on the Applicability of weather forecast model—based on logistic regression and decision tree. *Journal of Physics: Conference Series*, *1678*(1), 012110. https://doi.org/10.1088/1742-6596/1678/1/012110

Dougherty, J., Kohavi, R., & Mehran, S. (1995). Supervised and unsupervised discretization of continuous features. *Machine Learning? Proceedings of the Twelfth International Conference*.

García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining. In J. Kacprzyk & L. C. Jain (Eds.), *Intelligent Systems Reference Library* (72nd ed., Vol. 72). Springer Cham Heidelberg New York Dordrecht London. https://doi.org/10.1007/978-3-319-10247-4

Huang, K., & Wang, T. (2024). Optimized application of the decision tree ID3 algorithm based on big data in sports performance management. *International Journal of E-Collaboration*, *20*(1), 1–20. https://doi.org/10.4018/IJeC.350022

 10.12928/bamme.v5i1.13122

Husain, G., Nasef, D., Jose, R., Mayer, J., Bekbolatova, M., Devine, T., & Toma, M. (2025). SMOTE vs. SMOTEENN: A study on the performance of resampling algorithms for addressing class imbalance in regression models. *Algorithms*, *18*(1), 1–16. https://doi.org/10.3390/a18010037

Kresnawati, E. S., Suprihatin, B., & Resti, Y. (2024). The combinations of fuzzy membership functions on discretization in the decision tree-ID3 to predict degenerative disease status. *Symmetry*, *16*(12). https://doi.org/10.3390/sym16121560

Kumar, P., Bhatnagar, R., Gaur, K., & Bhatnagar, A. (2021). Classification of imbalanced data:review of methods and applications. *IOP Conference Series: Materials Science and Engineering*, *1099*(1), 012077. https://doi.org/10.1088/1757-899x/1099/1/012077

Matzavela, V., & Alepis, E. (2021). Decision tree learning through a predictive model for student academic performance in intelligent M-Learning environments. *Computers and Education: Artificial Intelligence*, *2*, 100035. https://doi.org/10.1016/j.caeai.2021.100035

Mienye, I. D., & Jere, N. (2024). A survey of decision trees: Concepts, Algorithms, and applications. *IEEE Access*, *12*, 86716–86727. https://doi.org/10.1109/ACCESS.2024.3416838

Mondal, S., Maity, R., & Nag, A. (2025). An efficient artificial neural network-based optimization techniques for the early prediction of coronary heart disease: comprehensive analysis. *Scientific Reports*, *15Mondal,*(1), 1–24. https://doi.org/10.1038/s41598-025-85765-x

Nicholas, Hoendarto, G., & Tjen, J. (2025). Heart disease prediction with decision tree. *Social Science and Humanities Journal*, *9*(01), 6451–6457. https://doi.org/10.18535/sshj.v9i01.1444

Noeman, A., Handayani, D., & Hiswara, A. (2022). Decision tree-based weather prediction. *PIKSEL : Penelitian Ilmu Komputer Sistem Embedded and Logic*, *10*(1), 67–78. https://doi.org/10.33558/piksel.v10i1.4418

Prasad, B. K., Uddinlb, M. Z., Nithinc, P., Goudd, T. A., & Subbaiahe, H. V. (2025). Rainfall prediction using machine learning Bikan. *International Journal of Research Publication and Reviews*, *6*(4), 2364–2369. https://doi.org/10.2139/ssrn.4909110

Pratiwi, Y., Rejo, A., Fariani, A., & Faizal, M. (2021). Monitoring and prediction land cover in Prabumulih City, South Sumatera Province, Indonesia using land change modeler and multi-temporal satellite data. *Ecology, Environment and Conservation Paper*, *27*(2021), S334–S340.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R., & Willson, M. (2025). Probabilistic weather forecasting with machine learning. *Nature*, *637*(8044), 84–90. https://doi.org/10.1038/s41586-024-08252-9

Resti, Y., Irsan, C., Amini, M., Yani, I., Passarella, R., & Zayanti, D. A. (2022). Performance improvement of decision tree model using fuzzy membership function for classification of corn plant diseases and pests. *Science and Technology Indonesia*, *7*(3), 284–290. https://doi.org/10.26554/sti.2022.7.3.284-290

Sasanya, B. F., Awodutire, P. O., Ufuoma, O. G., & Balogun, O. S. (2022). Modelling the effects of meteorological factors on maximum rainfall intensities using exponentiated standardized half logistic distribution. *Journal of Applied Mathematics, 2022*(1), 3250954.

Sondos Jameel Mukhyber. (2025). Classification of heart disease using feature selection and machine learning techniques. *Physical Sciences, Life Science and Engineering*, *2*(3), 9. https://doi.org/10.47134/pslse.v2i3.386

Taha Jijo, B., & Mohsin Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

Thölke, P., Mantilla-Ramos, Y. J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Mekki Berrada, L., Sahraoui, M., Young, T., Bellemare Pépin, A., El Khantour, C., Landry, M., Pascarella, A., Hadid, V., Combrisson, E., O'Byrne, J., & Jerbi, K. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage*, *277*(April).

https://doi.org/10.1016/j.neuroimage.2023.120253

Vijaya Saraswathi, R., Gajavelly, K., Kousar Nikath, A., Vasavi, R., & Reddy Anumasula, R. (2022). Heart disease prediction using decision tree and SVM. In *Algorithms for Intelligent Systems* (Issue March, pp. 69–78). Springer Nature Singapore Pte Ltd. https://doi.org/10.1007/978-981-16-7389-4_7

Walsh, R., & Tardy, M. (2023). A comparison of techniques for class imbalance in deep learning classification of breast cancer. *Diagnostics*, *13*(1), 1–19. https://doi.org/10.3390/diagnostics13010067

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and smote methods for dealing with imbalanced classification in educational data mining. *Information (Switzerland)*, *14*(1). https://doi.org/10.3390/info14010054

Xiang, B., Zeng, C., Dong, X., & Wang, J. (2020). The Application of a decision tree and stochastic forest model in summer precipitation prediction in Chongqing. *Atmosphere*, *11*(5), 508. https://doi.org/10.3390/atmos11050508

Yani, I., & Resti, Y. (2024). Plastic-type prediction based on digital image using multinomial Naïve Bayes method. *AIP Conference Proceedings*, *2920*(1), 040005. https://doi.org/10.1063/5.0179636

Zhang, Y., Deng, L., & Wei, B. (2024). Imbalanced data classification based on improved random-SMOTE and feature standard deviation. *Mathematics*, *12*(11), 1709. https://doi.org/10.3390/math12111709

10.12928/bamme.v5i1.13122