the_C5.0_AlgorithmCase_Study_Indonesia_Family_Survey_Data.pdf

by

Submission date: 08-Nov-2022 04:53PM (UTC+0700)
Submission ID: 1948101712
File name: the_C5.0_AlgorithmCase_Study_Indonesia_Family_Survey_Data.pdf (642.15K)
Word count: 4742
Character count: 22954

THE 1st INTERNATIONAL CONFERENCE ON DATA SCIENCE AND OFFICIAL STATISTICS
ICDSOS
"Strengthening the Contribution of Data Science and Official Statistics to the Society in the Distruption Era"
2021

Application of Resampling and Boosting Methods Using the C5.0 Algorithm

Case Study Indonesia Family Survey Data

H Kuswanto, N Sunusi, Siswanto, Nirwan

Department of Statistics, Faculty of Mathematics and Natural Sciences Hasanuddin University, Makassar, Indonesia, 90245

*Corresponding author's e-mail: hedikuswanto454@gmail.com

Abstract. Hypertension is a non-communicable disease that is characterized by an increase in systolic and diastolic blood pressure of more than 140 mmHg and or 90 mmHg. Hypertension needs to get more attention the condition is because hypertension will cause complications in the target organs and this disease does not appear to show significant symptoms at the beginning of the disease because it is called "silent disease". The study discusses the integration method of resampling and boosting in predicting hypertension status using the C5.0 algorithm. Classification of the C5.0 Algorithm by applying to resample increases performance specificity and AUC. Random oversampling (ROS) increased the specificity by 95.67% and AUC increased by 91.11%. Random over-under sampling (ROUS) increased specificity by 88.84% and AUC increased by 87.13%. In addition, applying boosting to the C5.0 algorithm that has been reapplied increases the accuracy performance. Random oversampling (ROS) increased accuracy by 93.86% and random over-under sampling (ROUS) increased accuracy by 89.98%. The response variables that contributed the most were high cholesterol and heart problems. The application of resampling and boosting to the contribution of high cholesterol and heart problems always topped the list.

1. Introduction

Non-communicable diseases are the main cause of death and physical dysfunction suffered by people throughout the world, especially in heart and blood vessel disease. Riskesdas data in 2013, diagnoses made to see the symptoms of hypertension and hypertension drug consumption only reached 9.50% [1]. Most hypertension does not show any initial symptoms. Hypertension can trigger a stroke and sudden cardiac arrest resulting in death. This is what causes hypertension is considered a deadly disease [2].

Hypertension is a non-communicable disease that is characterized by an increase in systolic and diastolic blood pressure of more than 140 mmHg and or 90 mmHg. Symptoms of hypertension that are not detected early and do not get better care can cause damage to organs [3]. Hypertension needs to get more attention the condition is because hypertension will cause complications in the target organs and this disease does not appear to show significant symptoms at the beginning of the disease because it is called "silent disease" [4].

Hypertension is a disease defined as a persistent increase in blood pressure [5]. The World Health Organization (WHO) estimates that currently, the global prevalence of hypertension is 22% of the world's total population. The results of Riskesdas 2018 show the prevalence of hypertension in the

population aged over 18 years based on national measurements of 34.11%. Nationally, the prevalence of hypertension shows an increasing trend from Riskesdas in 2007 [6]. Risk factors for hypertension can be divided into two, namely uncontrolled such as heredity, gender, and age. The controls are obesity, lack of exercise, smoking, and consumption of alcohol and salt [7]. Therefore, a model is needed to find the right formulation to determine a person's hypertension status using machine learning.

There are several studies that h[24] discussed the implementation of machine learning in the scope of Health, namely regarding the diagnosis of diabetes using classification mining techniques, the results of this study [22] that diabetes detection in the early stages is the key to treatment, things that need to be detected are plasma glucose concentration, body mass index, age, and diabetes pedigree function [8]. In [3]ddition, regarding the classification of factors causing diabetes mellitus using the C4.5 algorithm, the results of this study indicate that the factors that substantially affect the status of diabetes mellitus are fasting blood glucose, LDL cholesterol, age, and weight [9].

This study uses the C5.0 algorithm for the calculation process. In previous studies using the C5.0 algorithm the accuracy of 84.49% for buy accuracy and 83.69% for sale accuracy in the forex market forecasting [10]. In other studies concerning individual evaluation credit at the Bank using the C5.0 algorithm, an accuracy of 85.36% was obtained [11] and regarding the classification of child developmental deviations obtained the highest accuracy of 95.99% [12].

One of the things that need to be considered in evaluating the C5.0 algorithm model is the accuracy of a model in predicting responses correctly. Based on hypertension data obtained from the Indonesia Family Life Survey in 2014, it is known that there is a small proportion of people with hypertension status. This indicates that there is an imbalance of data between not being exposed to hypertension (majority) and affected by hypertension (minority). This imbalance will have an impact on the results of classification predictions because almost all classification analyzes produce much higher accuracy for the majority class than the minority class when there is an imbalance of data [13].

The resampling method is one method that can be used in handling the existence of data imbalance. The resampling method in classification is effective in handling class imbalance. However, the application of the resampling method only increases the minority class, so that misclassification can still occur. A good classification method will produce a few misclassifications. One [9]f the developments of machine learning to improve model accuracy is the ensemble method [14]. One of the ensemble methods of classification is boosting which is more popular to use compared to bagging. One of the commonly used boosting techniques is the Adaptive Boosting algorithm [15].

This study discusses the integration meth[21]of resampling and boosting in predicting hypertension status using the C5.0 algorithm. In addition, the data preparation stage (preprocessing) will be carried out on the data to improve performance and adjust input data in the classification analysis used.

2. Methodology

The following is a systematic scheme in researching the application of resampling and boosting methods
using the C5.0 algorithm

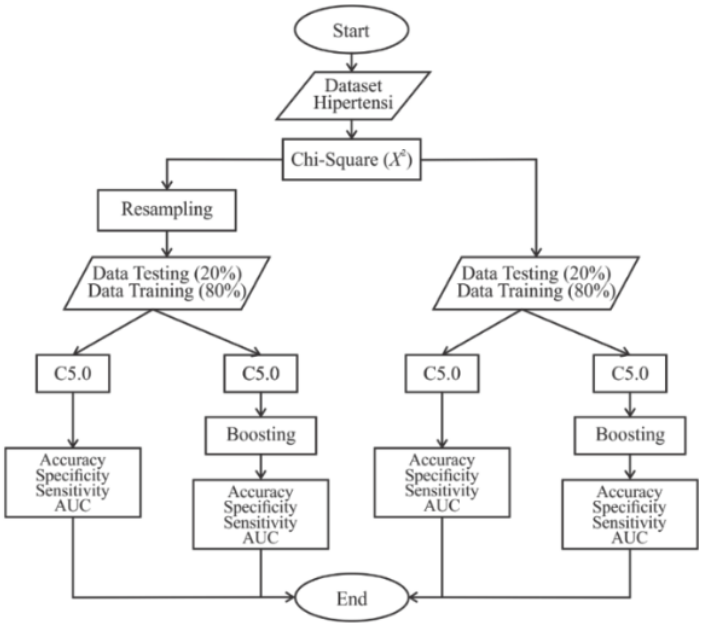

Figure 1. Research system scheme.

2.1. Data Source

The data used in this study are secondary data, namely Indonesian hypertension data obtained through the 2014 Indonesia Family Life Survey with a total of 31397 data. The applications used in processing this data are software R 3.6.2 and R-Studio. The variables that will be used in this study are:

Table 1. Model building variables.

Response variable	Variable	Scale	Information
Y	Hypertension status	Nominal	1: Yes 2: No
X1	Hearts problem	Nominal	1: Yes 2: No
X2	High Cholesterol	Nominal	1: Yes 2: No
X3	Kidney illness	Nominal	1: Yes 2: No
X4	Psychic problem	Nominal	1: Yes 2: No
X5	Smoking habit	Nominal	1: Yes 2: No
X6	Vision is not perfect	Nominal	1: Yes 2: No
X7	Find it difficult to concentrate on doing something	Ordinal	1: Rarely (< 1 day) 2: A little (1-2 days) 3: Sometimes (3-4 days) 4: Often (5-7 days)

Response variable	Variable	Scale	Information
X8	Feel depressed	Ordinal	1: Rarely (< 1 day) 2: A little (1-2 days) 3: Sometimes (3-4 days) 4: Often (5-7 days)
X9	Feeling requires a lot of effort in doing something	Ordinal	1: Rarely (< 1 day) 2: A little (1-2 days) 3: Sometimes (3-4 days) 4: Often (5-7 days)
X10	Feel worried	Ordinal	1: Very unsuitable 2: Not suitable 3: Neutral 4: Sufficiently Suitable 5: Very appropriate
X11	Tend to be lazy	Ordinal	1: Very unsuitable 2: Not suitable 3: Neutral 4: Sufficiently Suitable 5: Very appropriate
X12	Sometimes it's rude to others	Ordinal	1: Very unsuitable 2: Not suitable 3: Neutral 4: Sufficiently Suitable 5: Very appropriate
X13	Sleep quality	Ordinal	1: Very bad 2: Bad 3: Enough 4: Good 5: Very good
X14	Feel tired	Ordinal	1: Not at all 2: A little 3: Somewhat 4: Enough 5: Very much
X15	Headache	Nominal	1: Yes 2: No
X16	Out of breath	Nominal	1: Yes 2: No
X17	Nauseous vomit	Nominal	1: Yes 2: No
X18	Eat egg	Ordinal	0: Never 1: 1 day 2: 2 days 3: 3 days 4: 4 days 5: 5 days 6: 6 days 7: Every day

Response variable	Variable	Scale	Information
X19	Eat fish	Ordinal	0: Never 1: 1 day 2: 2 days 3: 3 days 4: 4 days 5: 5 days 6: 6 days 7: Every day
X20	Eat meat (beef, chicken, fork, etc)	Ordinal	0: Never 1: 1 day 2: 2 days 3: 3 days 4: 4 days 5: 5 days 6: 6 days 7: Every day
X21	Eat green vegetables	Ordinal	0: Never 1: 1 day 2: 2 days 3: 3 days 4: 4 days 5: 5 days 6: 6 days 7: Every day
X22	Eat instant noodles	Ordinal	0: Never 1: 1 day 2: 2 days 3: 3 days 4: 4 days 5: 5 days 6: 6 days 7: Every day
X23	Eat fast food	Ordinal	0: Never 1: 1 day 2: 2 days 3: 3 days 4: 4 days 5: 5 days 6: 6 days 7: Every day
X24	Drink soft drink	Ordinal	0: Never 1: 1 day 2: 2 days 3: 3 days 4: 4 days 5: 5 days

Response variable	Variable	Scale	Information
X25	Eat the chili sauce	Ordinal	6: 6 days 7: Every day 0: Never 1: 1 day 2: 2 days 3: 3 days 4: 4 days 5: 5 days 6: 6 days 7: Every day
X26	Eat fried food	Ordinal	0: Never 1: 1 day 2: 2 days 3: 3 days 4: 4 days 5: 5 days 6: 6 days 7: Every day
X27	Eat sweet foods	Ordinal	0: Never 1: 1 day 2: 2 days 3: 3 days 4: 4 days 5: 5 days 6: 6 days 7: Every day
X28	Do heavy physical activity	Nominal	1: Yes 2: No
X29	Doing moderate physical activity	Nominal	1: Yes 2: No
X30	On foot	Nominal	1: Yes 2: No

2.2. Research Methods

The steps of data analysis carried out in this study are as follows:

1. Data preparation phase
   1) Overcoming the problem of high data dimensions, can be done by selecting variables. In this study, the filter approach uses *Chi-Square* (χ^2) test. This aims to get the best classification results with several important variables, with the formula:
   $$\chi^2 = \frac{N(o_{11}o_{22} - o_{12}o_{21})}{n_1 n_2 C_1 C_2} \quad (1)$$
   2) The most informative variables will be identified by sorting each variable based on the p-value, which is p-value <0.05
   3) Exploring data to find out the general description of the data obtained
   4) Divide data groups into training data and test data. In this study the distribution of data by comparison of training data (80%) and testing data (20%)
2. Perform unbalanced data handling with the Random Oversampling (ROS) and Random Over – Under Sampling (ROUS) resampling methods, using the ROSE package by determining the

proportion (p) 0.5 or 50% so that the amount of class data is not affected by hypertension and is affected hypertension becomes more balanced.

3. Algorithm phase C5.0
   1. Calculate the entropy value for each response variable, using the formula:

$$Entropy(S) = \sum_{j=1}^{k} - p_j \log_2 p_j \quad (2)$$

   where:
   S: the set of cases on variables
   k: the number of partitions S
   p_j: the probability of a variable case
   2. Calculates the information gain value, using the formula:

$$Gain(A) = Entropy(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (3)$$

   where
   A : the response variable
   $|S_i|$: the amount of data in each category *i*
   $|S|$: the sum of all data *S*
   (S_i) : *entropy* value in each category data *i*
   3. Calculate *split information* value,using the formula:

$$SplitInfo(S, A) = -\sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S} \quad (4)$$

   4. Calculate *gain ratio value*, using the formula:

$$Gain\ Ratio(A) = \frac{Gain(A)}{SplitInfo(S, A)} \quad (5)$$

   5. Response variable which has the highest gain ratio will be selected as the main node. The process is carried out until it reaches the last node
4. *Boosting* phase
   1) Set the number of iterations or *trial* (t)
   2) Calculate initial weight W, with the formula:

$$W_i^t = \frac{1}{N}$$

   where:
   W_i^t : Sample weight *I* at *trial t*
   N : sum of all data
   Calculate the normalized value of weights in each sample with the formula:

$$P_i^t = \frac{W_i^t}{\sum_{i=0}^{n} W_i^t} \quad (6)$$

   Where:
   P_i^t : Sample weight -*I* at *trial t* which has been normalized
   Calculate the error rate value, using the formula:

$$\varepsilon^t = \sum_{i=0}^{n} (P_i^t \theta_i^t) \quad (7)$$

   where:
   ε^t : *Error rate* decision tree at *trial* for $i = 1, \dots, t$
   θ_i^t : Indicator function of sample for $i = 1, \dots, t$
5. Evaluation of the model
   1) Evaluate the models built by the CART method by calculating the value of accuracy, sensitivity, and specificity.

2) Look at the value of Area Under Curve (AUC) goodness of the results of the classification method
3) Knowing the percentage contribution variable in the model.

Overcoming the problem of high data dimensions, can be done by selecting variables. In this study, the filter approach uses the chi-square. This aims to get the best classification results with several important variables

3. Result and Discussion

3.1. Data description

Description of the response variable hypertension data in the 2014 Indonesia Family Life Survey questionnaire can be seen from Table 2:

Table 2. General description of hypertension data.

Hypertension Status	Frequency	Percentage
Normal	27.664	88,11%
Hypertension	3.732	11,89%
Total	31.396	100%

Based on Table 2, from 31,396 Hypertension status in Indonesia, 3,732 (11.89%) people experienced Hypertension, as many as 27,664 (88.11%) people did not experience Hypertension (normal). *Chi-Square*.

Determination of variable selection, i.e. p-value <0.05. To do Chi-Square (χ^2), the *chi-square* test (χ^2) function is used on each variable. The selection results are given in Table 3:

Table 3. Chi-Square variable selection.

Response variable	*p-value*	Information	Response variable	*p-value*	Information
X1	2.20E-16	Significant	X16	4.49E-15	Significant
X2	2.20E-16	Significant	X17	1.52E-05	Significant
X3	2.20E-16	Significant	X18	0.2597	Not Significant
X4	0.1787	Not Significant	X19	0.4405	Not Significant
X5	2.20E-16	Significant	X20	0.1668	Not Significant
X6	2.20E-16	Significant	X21	0.03882	Significant
X7	0.0003844	Significant	X22	2.56E-08	Significant
X8	0.05743	Not Significant	X23	0.05584	Not Significant
X9	0.02158	Significant	X24	0.002307	Significant
X10	3.06E-05	Significant	X25	0.02014	Significant
X11	0.009716	Significant	X26	0.3861	Not Significant
X12	0.0009435	Significant	X27	0.002931	Significant
X13	0.1215	Not Significant	X28	5.44E-05	Significant
X14	0.1164	Not Significant	X29	0.4699	Not Significant
X15	2.20E-16	Significant	X30	0.1235	Not Significant

In the variable selection stage that has been outlined in Table 3, selection of variables there are 19 influential variables and 11 variables that have not significant.

3.2. Training data and testing data

The distribution of training data and test data was carried out randomly on the available data, based on the magnitude of the difference in observations between classes on the Hypertension Status variable which indicated an imbalance between Hypertension and Normal classes, it is necessary to balance the data to minimize classification errors because it is dominated by class classification results. The majority so that efforts were made to balance the data by applying three types of resampling methods to the training data, using the Random Oversampling (ROS) and Random Over – Under Sampling (ROUS) methods, with details as shown in Figure 2

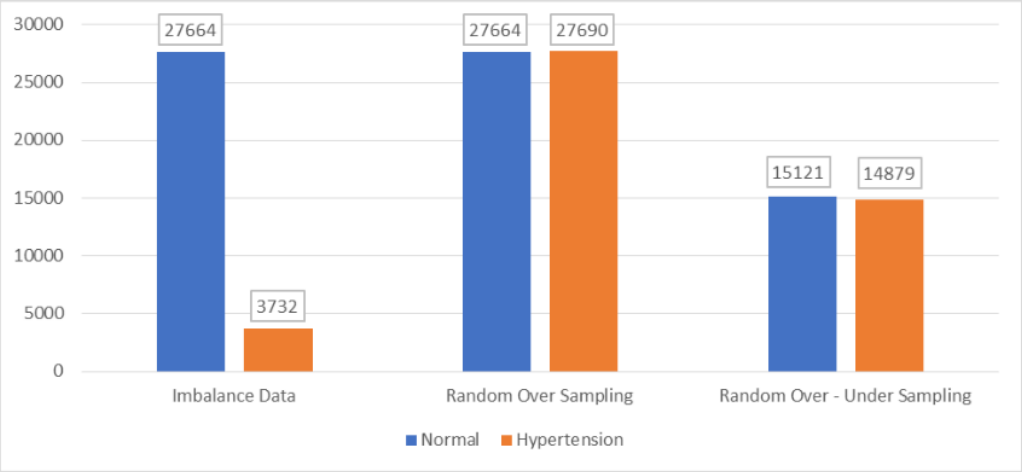

Figure 2. The dataset before and after resampling.

Based on the bar diagram in Figure 2, it can be seen that there are changes in the number of observations in the minority or majority classes when balancing the data using the ROS and ROUS methods in the proportion of 0.5 or 50% so that the data are almost balanced. The ROS method manages the imbalance of data by adding observations to the Hypertension class to as many as 27,690 observations, close to the number of normal class observations. The ROUS method handles data imbalance by reducing the number of normal class observations to 15,121 and increasing the number of Hypertension class observations to 14,879 observations.

3.3. Algorithm C5.0

Table 4. Gain ratio value.

C5.0		ROS + C5.0		ROUS + C5.0	
Response variable	*Gain Ratio*	Response variable	*Gain Ratio*	Response variable	*Gain Ratio*
X1	0.0456	X1	0.0487	X1	0.0485
X2	0.0731	X2	0.0836	X2	0.0793
X3	0.0152	X3	0.0248	X3	0.0208
X5	0.0023	X5	0.0069	X5	0.0066
X6	0.0099	X6	0.0193	X6	0.0165
X7	0.0005	X7	0.0015	X7	0.0012
X9	0.0001	X9	0.0008	X9	0.0009
X10	0.0010	X10	0.0018	X10	0.0023

C5.0		ROS + C5.0		ROUS + C5.0	
Response variable	*Gain Ratio*	Response variable	*Gain Ratio*	Response variable	*Gain Ratio*
X11	0.0006	X11	0.0014	X11	0.0022
X12	0.0009	X12	0.0018	X12	0.0028
X15	0.0022	X15	0.0056	X15	0.0077
X16	0.0029	X16	0.0061	X16	0.0100
X17	0.0006	X17	0.0015	X17	0.0019
X21	0.0004	X21	0.0007	X21	0.0013
X22	0.0011	X22	0.0035	X22	0.0030
X24	0.0009	X24	0.0022	X24	0.0012
X25	0.0005	X25	0.0011	X25	0.0014
X27	0.0006	X27	0.0017	X27	0.0022
X28	0.0004	X28	0.0011	X28	0.0015

The node used as the main node is the variable X2 or high cholesterol which divides the population into two nodes, namely the left node for the Yes category and the right node for the No category. The X2 variable produces the highest gain ratio compared to other variables, namely 0.0731 for C5.0, 0.0836 for ROS + C5.0, and 0.0793 for ROUS + C5.0. Returns the contribution variable response as follows:

Table 5. Response variable contribution.

C5.0		ROS + C5.0		ROUS + C5.0	
Response variable	Contribution (%)	Response variable	Contribution (%)	Response variable	Contribution (%)
X2	100.00	X2	100.00	X2	100.00
X1	4.08	X1	99.48	X1	99.03
X11	3.79	X3	96.94	X3	91.62
X15	3.46	X22	90.92	X16	90.48
X12	2.39	X27	90.03	X6	89.93
X22	1.80	X24	89.40	X5	87.35
X16	1.49	X6	88.00	X24	82.67
X27	1.31	X5	85.82	X11	79.79
X6	1.18	X21	82.42	X27	79.73
X24	0.60	X10	82.15	X25	74.70
X25	0.48	X25	80.51	X21	73.87
X21	0.33	X15	80.19	X15	71.19
X7	0.30	X11	80.13	X22	66.85
X5	0.20	X12	78.19	X12	61.25
X10	0.12	X7	75.11	X9	55.77
X28	0.10	X16	68.26	X10	53.14
X9	0.04	X9	63.65	X7	52.52
X17	0.03	X17	41.01	X17	36.00
		X28	35.71	X28	33.01

Variables X2 (high cholesterol) and X1 (heart problems) are the variables that have the highest contribution, so it can be said that high cholesterol and heart problems are the main factors in determining hypertension status.

Confusion matrix for the classification accuracy of the C5.0 algorithm model:

Table 6. Confusion matrix algorithm C5.0.

Prediction	C5.0		ROS + C5.0		ROUS + C5.0	
	Reference		Reference		Reference	
	Normal	Hypertension	Normal	Hypertension	Normal	Hypertension
Normal	5449 (99.09%)	678 (94.04%)	4378 (76.46%)	235 (4.33%)	2313 (77.20%)	334 (11.64%)
Hypertension	50 (0.91%)	43 (5.96%)	1132 (20.54%)	5195 (95.67%)	683 (22.80%)	2659 (88.84%)

Table 6 shows the classification accuracy of hypertension class is only 43 (5.96%). Classification error data 678 (94.04%) data, and normal class shows classification accuracy 5449 (99.09%), classification error data 50 (0.91%) data for C5.0 model. This caused by unbalanced data, this condition occurs because the amount of data from the normal class is far more than the hypertension class. The application of the over-sampling hypertension class in the C5.0 model increased the classification accuracy of 5195 (95.67%) data of misclassification 235 (4.33%) data, but the normal class of classification accuracy was reduced to 4378 (76.46%). Classification error data of 1132 (20.54%). The application of over-sampling hypertension class and under-sampling normal class on C5.0 model increased the classification accuracy of 2659 (88.84%), 334 (11.64%) misclassification data, but the normal class classification accuracy was reduced to 2313 (77.20%), classification error data 683 (22.80%).

Applying resampling to the C5.0 algorithm results in a greater classification accuracy of the Hypertension class than without applying to resample

Table 7. Classification performance algorithm C5.0.

C5.0		ROS + C5.0		ROUS + C5.0	
Criteria	Data Test 20% (%)	Criteria	Data Test 20% (%)	Criteria	Data Test 20% (%)
Accuracy	88.30	Accuracy	87.50	Accuracy	83.02
Sensitivity	99.09	Sensitivity	79.46	Sensitivity	77.20
Specificity	5.96	Specificity	95.67	Specificity	88.84
AUC	56.17	AUC	91.11	AUC	87.13

Applying to resample obtained better specificity performance. However, the performance of accuracy and sensitivity is reduced because it causes a decrease in classification accuracy in the normal class. After resampling, the proportion of data is almost the same Hypertension class information becomes more so that it will affect the decrease in performance sensitivity. Therefore a boosting method is used to reduce misclassification.

3.4. Boosting

The number of iterations or trials used is 6, then produces error rate:

Table 8. Boosting error rate value.

C5.0 + Boosting		ROS + C5.0 + Boosting		ROUS + C5.0 + Boosting	
Trial	Error rate	Trial	Error rate	Trial	Error rate
0	2836(11.3%)	0	2155(4.9%)	0	1465(6.1%)

C5.0 + Boosting		ROS + C5.0 + Boosting		ROUS + C5.0 + Boosting	
Trial	Error rate	Trial	Error rate	Trial	Error rate
1	3310(13.1%)	1	6092(13.7%)	1	3808(15.9%)
2	4503(17.9%)	2	5829(13.1%)	2	4091(17.0%)
3	4253(16.9%)	3	6875(15.5%)	3	3933(16.4%)
4	3664(14.6%)	4	6660(15.0%)	4	4010(16.7%)
5	3085(12.3%)	5	6782(15.3%)	5	4025(16.8%)
6	2862(11.4%)	6	916(2.1%)	6	513(2.1%)

By applying boosting on the data that has been sampled can reduce the classification error to 2.1%. As for the data that is not resampling results in an increase in the classification error of 0.1%. This is due to the determination of the trial that is less precise. In addition, the contribution variable response also increased:

Table 9. Response variable contribution.

C5.0		ROS + C5.0 + Boosting		ROUS + C5.0 + Boosting	
Response variable	Contribution (%)	Response variable	Contribution (%)	Response variable	Contribution (%)
X1	100.00%	X1	100.00%	X1	100.00%
X2	100.00%	X2	100.00%	X2	100.00%
X3	100.00%	X10	100.00%	X24	99.97%
X6	100.00%	X24	100.00%	X21	99.39%
X5	99.59%	X25	99.91%	X25	99.38%
X10	98.69%	X21	99.78%	X27	99.01%
X16	91.35%	X12	99.76%	X22	98.92%
X12	88.62%	X27	99.64%	X3	98.73%
X24	64.80%	X22	99.53%	X11	98.51%
X22	63.23%	X7	99.20%	X12	98.18%
X15	62.23%	X11	99.16%	X6	98.00%
X21	40.64%	X3	98.51%	X16	97.72%
X11	34.85%	X16	97.67%	X10	96.61%
X25	30.66%	X6	97.46%	X9	96.54%
X7	28.76%	X5	97.18%	X15	95.91%
X27	26.48%	X9	96.87%	X5	95.60%
X9	7.63%	X15	95.35%	X7	94.95%
X28	5.27%	X28	92.93%	X17	84.47%
X17	3.94%	X17	91.16%	X28	83.72%

Variables X2 (high cholesterol) and X1 (heart problems) are the variables that have the highest contribution, so it can be said that high cholesterol and heart problems are the main factors in determining hypertension status.

Confusion matrix for the classification accuracy of the C5.0 algorithm model using boosting:

Table 10. Confusion matrix algorithm C5.0 using boosting.

Prediction	C5.0 + Boosting		ROS + C5.0 + Boosting		ROUS + C5.0 + Boosting	
	Reference		Reference		Reference	
	Normal	Hypertension	Normal	Hypertension	Normal	Hypertension
Normal	5467 (99.42%)	695 (96.39%)	4942 (89.69%)	104 (1.92%)	2566 (85.65%)	170 (5.68%)
Hypertension	32 (0.58%)	26 (3.61%)	568 (10.31%)	5326 (98.08%)	430 (14.35%)	2823 (94.32%)

Applying boosting to the C5.0 algorithm that has been sampled results in a far less classification error than without applying to boost.

Table 11. Classification performance algorithm C5.0 using boosting.

C5.0 + Boosting		ROS + C5.0 + Boosting		ROUS + C5.0 + Boosting	
Criteria	Data Test 20% (%)	Criteria	Data Test 20% (%)	Criteria	Data Test 20% (%)
Accuracy	88.31	Accuracy	93.86	Accuracy	89.98
Sensitivity	99.42	Sensitivity	89.69	Sensitivity	85.65
Specificity	3.61	Specificity	98.08	Specificity	93.05
AUC	62.62	AUC	98.53	AUC	94.32

Applying to boost obtained all classification performance for the better, this can be seen from the accuracy performance.

4. Conclusion

Algorithm Classification of C5.0 by applying to resample increases the specificity and AUC performance. Random oversampling (ROS) increased the specificity by 95.67% and AUC increased 91.11%. Random over-under sampling (ROUS) increased specificity by 88.84% and AUC increased by 87.13%.

In addition, applying boosting to the resampling C5.0 algorithm increases accuracy performance. Random oversampling (ROS) increases accuracy by 93.86% and random over-under sampling (ROUS) increases accuracy by 89.98%. The most contributing response variables are high cholesterol and heart problems. The application of resampling and boosting contribution to cholesterol is high and heart problems are always at the top level. Thus the cause of hypertension problems in Indonesia which often occurs is high cholesterol and heart problems

Acknowledgment

Thank you to our team and the Head of the Department of Statistics, Faculty of Mathematics and Natural Sciences Hasanuddin University for their moral support so that this research can be carried out well.

References

[1] Kemenkes RI. (2013). Laporan riset Indonesia dasar (Riskesdas) tahun 2013. Kementerian Kesehatan RI. Jakarta.

[2] Sihombing, M. (2017). Faktor yang berhubungan dengan Hypertension pada penduduk Indonesia yang menderita diabetes melitus (data Riskesdas 2013). Buletin Penelitian Kesehatan, 45(1), 53–64. https://doi.org/10.22435/bpk.v45i1.5730.53- 64
[3] Joint National Committee. (2004). The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. National Institutes of Health.
[4] Feryadi, R., Sulastri, D., & Kadri, H. (2014). Hubungan kadar profil lipid dengan kejadian Hypertension pada masyarakat etnik minangkabau di Kota Padang tahun 2012. Forbes. Working From Home During The Coronavirus Pandemic: What You Need To Know. Available: http://www.forbes.com/. Jurnal Kesehatan Andalas, 3(2), 206–211.
[5] Dipiro, J. T., Talbert, R. L., Yee, G. C., Matzke, G. R., Wells, B. G., & Posey, L. M. (2011). Pharmacoterapy : A Pathophysiologic Approach. United States: The McGraw-Hill Companies, Inc.
[6] Kementerian Kesehatan RI. (2019). Hipertensi. Jakarta: Pusat Data dan Informasi Kementerian Kesehatan RI.
[7] Manurung, W. P., & Wibowo, A. (2016, Desember). Pengaruh Konsumsi Semangka (Citrullus vulgaris) untuk Menurunkan Tekanan Darah pada Penderita Hipertensi. MAJORITY, 102-107.
[8] Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv:1502.03774.
[9] Ente, D. R., Thamrin, S. A., Arifin, S., Kuswanto, H., & Andreza, A. (2020). Klasifikasi Faktor-Faktor Penyebab Penyakit Diabetes Melitus Di Rumah Sakit Unhas Menggunakan Algoritma C4.5. Indonesian Journal of Statistics and Its Applications, 4(1), 80-88.
[10] Wirdhaningsih, K. P. (2013). Penerapan Algoritma Decision tree C5.0 untuk Peramalan Forex. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 2, 8.
[11] Lin, P. S., & Zhang, G. J. (2009). C5.0 Classification Algorithm and Application on Individual Credit. Evaluation of Banks. Systems Engineering - Theory & Practice, 29(12).
[12] Dewi, D. A. W., Cholissodin, I., & Sutrisno. (2019). Klasifikasi Penyimpangan Tumbuh Kembang Anak Menggunakan Algoritma C5.0 . Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, Vol. 3, No. 10, Oktober 2019. 10258-10265.
[13] Gu Q, Wang XM, Wu Z, Ning B, Xin CS. 2016. An improved SMOTE Algorithm Based On Genetic Algorithm for Imbalanced Data Classification. Journal of Digital Information Management. 14(2): 92–103.
[14] Kurniawan, D., Supriyanto. 2013. Optimasi Algoritma Support Vector Machine (SVM) Menggunakan AdaBoost untuk Penilaian Risiko Kredit. Jurnal Teknologi Informasi.
[15] Feng, G., Zhang, J.-D. & Shaoyi Liao, S. 2014. A novel method for combining Bayesian networks, theoretical analysis, and its applications. Pattern Recognition.

the_C5.0_AlgorithmCase_Study_Indonesia_Family_Survey_Da...

ORIGINALITY REPORT

11%	7%	7%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	P. Little. "Randomised controlled trial of self management leaflets and booklets for minor illness provided by post", BMJ, 2001 Publication	2%
2	Submitted to University of KwaZulu-Natal Student Paper	1%
3	eudl.eu Internet Source	1%
4	scholarhub.ui.ac.id Internet Source	1%
5	sipora.polije.ac.id Internet Source	1%
6	www.sciencegate.app Internet Source	<1%
7	"Advances in Ergonomics Modeling, Usability & Special Populations", Springer Science and Business Media LLC, 2017 Publication	<1%
8	portal.research.lu.se Internet Source	<1%

9	Armin Lawi, Ali Akbar Velayaty, Zahir Zainuddin. "On identifying potential direct marketing consumers using adaptive boosted support vector machine", 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT), 2017 Publication	<1%
10	www.scilit.net Internet Source	<1%
11	Harleen Kaur, Ritu Chauhan, M. Alam. "chapter 5 An Optimal Categorization of Feature Selection Methods for Knowledge Discovery", IGI Global, 2013 Publication	<1%
12	Tiara, , Mira Kania Sabariah, and Veronikha Effendy. "Sentiment analysis on Twitter using the combination of lexicon-based and support vector machine for assessing the performance of a television program", 2015 3rd International Conference on Information and Communication Technology (ICoICT), 2015. Publication	<1%
13	Submitted to Universiti Teknologi MARA Student Paper	<1%
14	jurnal.undhirabali.ac.id	

Internet Source

<1%

---

15 M. Aldiki Febriantono, Sholeh Hadi Pramono, Rahmadwati Rahmadwati, Golshah Naghdy. "Classification of multiclass imbalanced data using cost-sensitive decision tree C5.0", IAES International Journal of Artificial Intelligence (IJ-AI), 2020

Publication

<1%

---

16 Xin Ma. "School experiences influence personal health and interpersonal relationships of adolescents: The Canadian case", School Effectiveness and School Improvement, 2007

Publication

<1%

---

17 journal.unhas.ac.id

Internet Source

<1%

---

18 cbs.aw

Internet Source

<1%

---

19 mafiadoc.com

Internet Source

<1%

---

20 Erwin Kurniawan, Fhira Nhita, Annisa Aditsania, Deni Saepudin. "C5.0 Algorithm and Synthetic Minority Oversampling Technique (SMOTE) for Rainfall Forecasting in Bandung Regency", 2019 7th International Conference

<1%

on Information and Communication Technology (ICoICT), 2019
Publication

21	Faizal Sudrajat, Rachmadita Andreswari, Nia Ambarsari. "Simulation of A Decision Support System Using Data Mining Method with C4.5 Algorithm: A Case Study", 2021 International Conference on Artificial Intelligence and Big Data Analytics, 2021 Publication	<1 %
22	dokumen.pub Internet Source	<1 %
23	www.ncbi.nlm.nih.gov Internet Source	<1 %
24	www.scirp.org Internet Source	<1 %
25	António Almeida, Beatriz García Fernández, Penelope Papadopoulou. "Animals as performers or exhibits: A study of their relevance to Portuguese, Spanish, and Greek pre-service teachers", Journal of Applied Animal Welfare Science, 2022 Publication	<1 %

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		