https://doi.org/10.48048/tis.2022.2167

Rule Formation Application based on C4.5 Algorithm for Household Electricity Usage Prediction

Firman Tempola[*], Miftah Muhammad,
Abdul Kadir Maswara and Rosihan Rosihan

Department of Informatics, Khairun University, North Maluku 97719, Indonesia

([*]Corresponding author's e-mail: firman.tempola@unkhair.ac.id)

Received: 16 September 2020, Revised: 20 June 2021, Accepted: 30 June 2021

Abstract

Electrical energy is one of the essential energies for the world due to the fast growth of the world population and houses. In Indonesia, 95 % of the energy used in the household in 2017 is electrical energy. Therefore, reducing the use of electricity in the household is crucial. In the past decades, customers have carried out several approaches to reduce the use of their electricity. One of the widely used methodologies is the EMS. However, the PDCA model has not been implemented in electricity consumption. Subsequently, this study applies such an approach focusing more on the planning stage and is implemented in Ternate City, North Maluku, Indonesia. The C4.5 algorithm is applied at the planning stage to form a rule in predicting household electricity consumption. Moreover, the system performance is tested using the confusion matrix. The data of electricity consumption is collected and the data is treated with a varying amount depending on the number of the training data applied. The results of the system performance test by applying the confusion matrix are 76.22, 90.3, and 74.4 % for the highest accuracy value, precision, recall, respectively with the number of rules formed by 14.

Keywords: C4.5, Household electricity, Prediction, Electrical energy, Energy management system

Introduction

The growth of the population, coupled with the rapid technological advances in all fields, causes an increase in energy consumption, specifically electrical energy. Electricity is the most needed energy because it is easily to distribute and convert into energy and hence, is vital energy for modern society nowadays [1].

Based on the segmentation conducted by the State Electricity Company/ Perusahaan Listrik Negara (PLN), electricity customers divided into 5 sectors, i.e. the social activities, household, business, industrial and public sectors. In Indonesia, the use of household electricity has increased significantly from 52 % in 2001 to 95 % in 2017 [2].

The increasing consumption of electrical energy certainly affects the raw materials to produce electrical energy. Currently, the government continues to campaign for an energy-saving culture among the communities so that it can reduce electricity use in particular for household electricity [3]. In addition, to save electrical energy, an energy management system (EMS) applied with the Plan-Do-Check-Act (PDCA) model has been applied to optimise the energy used in small and medium-sized enterprises [4]. An evaluation in one of the most important stages in the EMS, the planning stage, has also been carried out by applying the auto-encoder algorithm in deep learning [5]. No EMS-PDCA model has been implemented for the management of electricity consumption in the household in Indonesia. Subsequently, this study applies such an approach focusing only on the plan stage.

Several studies have been done to predict the electricity consumption with the data applied, namely from the UCI Machine Learning Repository; University of California, School of Information and Computer Science [6,7]. Similar works have been also carried out in Indonesia to predict electrical energy consumption by applying the Naive Bayes algorithm [8] and classified the electricity consumption for different types of houses [3]. On the other hand, this study develops a rule-forming application based on a C4.5 algorithm to predict the electricity of households in the city. The formation of rules with the C4.5 algorithm has also been built to assess the voltage stability [9]. In this study, the criteria to predict household electricity usage are the number of families, the monthly income and electrical power.

Moreover, 2 predicted target is classified: (1) electricity saving and non-elecricity saving. From the predicted results, the accuracy system is calculated and then validated by the holdout validation.

Related work

Energy consumption for buildings is quite high, reaching 76 % in the U.S [10]. In Indonesia, the use of household electricity continues to increase every year [2] and it may continue to soar because by 2026 all regions of Indonesia will use electricity based on the general plan for electricity providers. In general, household electricity is used for various needs, e.g. cooking, refrigerator, entertainment, lighting and cooling. In Indonesia, especially in Bandung, it mostly used electricity for cooking, amounting to 48 % [11,12] compared household electricity usage between Bandung and Jakarta clustering with k-means algorithms. In addition, [13] predicted general electricity consumption in household Indonesia. Meanwhile [14] looked at the characteristics of electricity consumption in Indonesia based on the power used in households for electricity use. The results in [14] showed that the power cluster iv with monthly use was reached 248.1 kWh with the most electricity consumption is for air conditioner by 43.66 %.

High electricity consumption needs to be estimated to save energy. One approach is with energy management system with a plan-do-check-act (PDCA) concept. The PDCA concept utilizes existing models in data mining or machine learning [5,15]. Several studies adopted a model in data mining [1] to predict the constation of electrical energy with artificial neural networks [16], to build a decision support system to measure power automatically on household electricity [17], to estimate the use of electricity in air conditioning [18], to apply the Support vector regression algorithm to the data used by existing electricity users in Canada [19] and to predict what type of electrical interference [20].

The C4.5 algorithm, which is part of an existing algorithm in data mining or machine learning, has been implemented in various cases. For instance, [21] conducted sentiment analysis in English. [22] compared the C4.5 algorithm with the fuzzy one showing the highest accuracy in the C4.5 algorithm at 93.9 %, [23] studied data mining in identifying determinant factor related to customer satisfaction in fast-food restaurant with system accuracy results of over 80 %. [24] selected logistic algorithm as the best predicted model with a high accuracy rate of 74.8 % to perform the classification of diabetic patients. [25] used classification methods (Naïve Bayes, C4.5, SVM) in classifying text in Arabic. In addition, [26] measured student performance through mining the information hiding inside the student scores.

Materials and methods

The application of rule formation by applying the C4.5 algorithm for household electricity prediction comprises several steps: data acquisition, rule formation construction, system design, and testing the performance of the prediction algorithm.

Data aquisition

The data acquisition in this study is real data taken from household electricity users in the city of Ternate, North Maluku, Indonesia. The number of households in the city of Ternate is 37,735 residents' houses, so the sample data applied in this study is 3800 data. This data consists several features including the number of electronic equipment (X5 in **Table 1**), the number of family members (X1 in **Table 1**), electric power (X4 in **Table 1**), house area (X2 in **Table 1**) and monthly income (X3 in **Table 1**). Features of the research based on several studies related to household electricity in Indonesia and interviews with household electricity users. Moreover, the class or label is the amount of household electricity kWh usage. The type of data is numerical data as shown in **Table 1**. However, this rule data will be transformed into categorical data.

Table 1 Data acquisition.

No	Criteria					Class
	X1	X2	X3	X4	X5	
1	2	8×6	500000	1300	5	183 kWh
2	7	10×8	4000000	900	9	101 kWh
3	4	8.5×7.5	2500000	900	14	189 kWh
4	3	15×10	500000	450	7	123 kWh
5	5	15×10	2000000	900	10	222 kWh
6	3	9×6.15	500000	900	5	77 kWh
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
70	4	15×18	800000	900	13	165 kWh

Construction rule formation

This research forms a rule by applying the C4.5 Algorithm. The C4.5 algorithm is one of the machine learning algorithms [27] developed from the ID3 algorithm to accommodate both numerical and categorical data. In addition, the C4.5 Algorithm can also handle noise and missing data problems. The formation of rules in the C4.5 algorithm comprises several stages which are the following:

Find the entropy value

The entropy that is the entire feature in the dataset is calculated using Eq. (1). If in 2 classes, the number of class 2 and class 3 samples is the same then the entropy = 1. However, if one of the classes is class 0 then entropy = 0. Entropy is instrumental in determining which nodes will be the solution to the next data training. High entropy values have a greater chance of improving performance than the predicted one. The entropy of each feature is calculated iteratively for all the data.

$$entropy\ (s) = \textstyle\sum_{j=1}^{k} -p_j\ log_2\, p_j \tag{1}$$

Find the gain value

The calculated gain value is the gain value of all features in household electricity usage data. The feature with the highest gain value is defined as the root node. Eq. (2) is used to get the gain value.

$$gain\ (f) = entropy\ (s) - \textstyle\sum_{i=1}^{k} \frac{|s_i|}{|s|} \times entropy\ (s_i) \tag{2}$$

Calculate splitinfo and gain ratio

After the root node is determined by calculating the overall gain of the features, the splitinfo and gain ratio are further estimated to define the branches of the tree. To get the splitinfo and ratio values, Eqs. (3) - (4) are used, respectively.;

$$SplitInfo\ (S,S_i) = -\sum_{i=1}^{k} p\ (v_i \mid S)\ log_2\, p\ (v_i \mid S) \tag{3}$$

$$Rasio\ Gain\ (S,S_i) = \frac{Gain\ (S,\ S_i)}{SplitInfo\ (S,\ S_i)} \tag{4}$$

System design

This designed system consists of 2 types of users: (1) admin and (2) electricity customers. Each type of user has its own access rights (**Figure 1**).

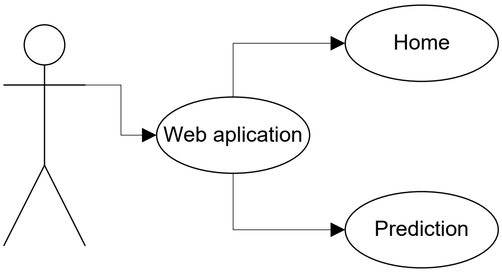


Figure 1 Design UML admin.

Testing the performance of the predicition algorithm

This system, after forming a rule by applying the C4.5 algorithm, is continued by making predictions through utilizing household electricity customer data. The next step is to test the algorithm to see the performance of the C4.5 algorithm. The performance of the algorithm known by calculating precision, recall and accuracy. **Table 2** shows the results of algorithm testing [28].

Table 2 Confusion matrix

Classification		Predicted class	
		True	False
Observed class	True	True Positive (TP)	False Negative (FN)
	False	False Positive (FP)	True Negative (TN)

True Positive in the above table means that the predicted data of saving energy in class is the same as the actual data. False Positive is classified as a non-saving electricity even though the actual data is a saving-electricity. False negative is predicted as a saving-elextricity with an actual data of non-saving electricity. In addition, true negative is predicted to have a non-saving electricity as same as the actual data;

$$Precision = \frac{\sum true\ positive\ (TP)}{\sum true\ positive\ (TP) + \sum false\ positive\ (FP)} \times 100\% \tag{5}$$

$$Recall = \frac{\sum True\ Postifive\ s\ (TP)}{\sum true\ positives\ (TP) + \sum False\ Negatives\ (FN)} \times 100\% \tag{6}$$

$$Accuracy = \frac{\sum TP + \sum TN}{\sum TP + \sum TN + \sum FP + \sum FN} \times 100\% \tag{7}$$

Results and discussion

Transformation of data

The first step is to know the criteria data in numerical form and transformed the data into categories. The transformed data are not only the criteria data but also on class or label from the household electricity user dataset. **Table 3** shows the ranges of numerical data which will be transformed into categories.

Table 3 Transformed of data.

No.	Criteria	Criteria value	Transformation results
1	X1	≤ 3	Little
		4 – 6	Medium
		> 6	Many
2	X2	≤ 80	Small
		81 – 150	Are
		> 150	Great
3	X3	$\leq Rp. 1.500.000$	Low
		$Rp. 1.500.001 - Rp. 2.500.000$	Are
		$Rp. 2.500.001 - Rp. 3.500.000$	High
		$> 3.500.000$	Very high
4	X4	450 VA	Low
		900 VA	Are
		1300 VA	High
		2200 VA	Very high
5	X5	≤ 3	Little
		4 – 6	Medium
		> 6	Many

Besides the criteria data that are transformed into categories, label or class is also transformed into a category (i.e. saving or non-saving electricity) as shown in **Table 4**.

Table 4 Transformed class or target.

Class/label	Electrical Power	Class grades	Transformation results
Electricity Usage	450 VA	≤ 75kWh	electricity saving
		> 75 kWh	non-electricity saving
	900 VA	≤ 115 kWh	electricity saving
		> 115 kWh	non-electricity saving
	1300	≤ 201 kWh	electricity saving
		> 201 kWh	non-electricity saving
	2200	≤ 358 kWh	electricity saving
		> 358 kWh	non-electricity saving

Rule formed

Based on the data that has transformed, the application of the c4.5 algorithm then carried out and the results in terms of the overall entropy value, the entropy value of each criterion in each category, the gain value of each feature, the splitinfo value of each feature and the gain ratio value of each feature is shown in the **Table 5**.

Table 5 Results of counting training data.

Node	Attribute	Attribute value	Number of cases	Save	Not frugal	Entropy	Gain	Splitinfo	Gain ratio
1	Total		55	23	32	0.981			
	X1						0.04	1.405	0.028
		Little	7	5	2	0.863			
		medium	28	10	18	0.94			
		Many	20	8	12	0.971			
	X2						0.026	1.423	0.018
		Small	16	9	7	0.989			
		Are	30	11	19	0.948			
		Great	9	3	6	0.918			
	X3						0.038	1.887	0.02
		Low	23	12	11	0.999			
		Are	11	3	8	0.845			
		High	8	4	4	1			
		Very High	5	1	4	0.89			
	X4						0.08	1.638	0.049
		Low	7	5	2	0.863			
		Are	31	10	21	0.907			
		High	12	7	5	0.98			
		Very high	5	1	4	0.722			
	X5						0.153	1.524	0.1
		Little	12	7	5	0.98			
		Medium	25	14	11	0.99			
		Many	18	2	16	0.503			

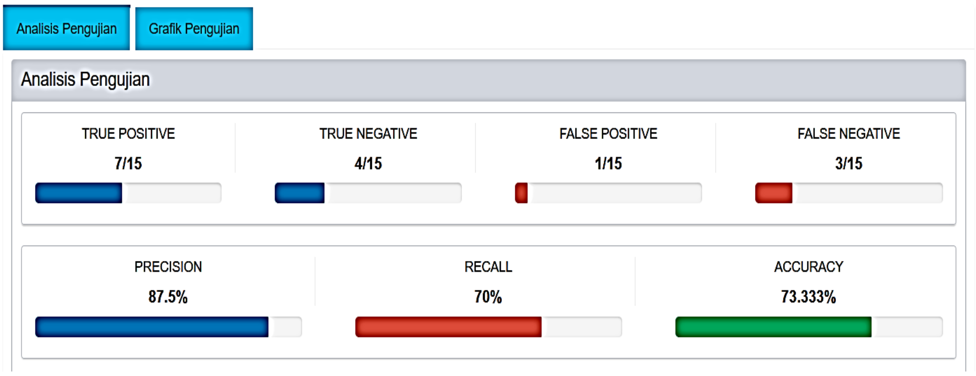
The results of the calculations in **Table 5** are the first step to determine the root node if you want to form a decision tree, to determine the root node, the highest gain value of all features sought. Based on the results of the calculation, the highest gain value is on the X5 criterion (i.e. the number of electronic equipment) with a value of 0.153 and hence, the root node is electronic equipment. Such a root node determined the 3 branches of the root node through the calculation of the value of the features of electronic equipment: few, medium and many. From the formed branches, there is no known rule to determine the classification or prediction of the data. Therefore, the calculation is done from the beginning once more to find the next value or node from the remaining data. The results of calculating the overall training data in the application in the formation of rules are shown in **Table 6.**

Table 6 Formation of rules.

Number	Rule
1	If (Number of electronics = little and number of family members = little) Then electricity saving
2	If (Number of Electronics = little and number of family members = medium and electric power = low) Then electricity saving
3	If (Number of Electronics = medium and electric power = very high) Then non-electricity saving
4	If (Number of Electronics = medium and electric power = low and number of family members = little) Then non-electricity saving
5	If (Number of Electronics = medium and electric power = low and number of family members = medium) Then electricity saving
6	If (Number of Electronics = medium and electric power = low and number of family members = many) Then electricity saving
7	If (Number of Electronics = medium and electric power = high and monthly income = are) Then electricity saving
8	If (Number of Electronics = medium and electric power = high and monthly income = high) Then non-electricity saving
9	If (Number of Electronics = medium and electric power = high and monthly income = very high) Then electricity saving
10	If (Number of Electronics = medium and electric power = high and monthly income = low and number of family members = low) Then non-electricity saving
11	If (Number of Electronics = medium and electric power = high and monthly income = low and number of family members = medium) Then electricity saving
12	If (Number of Electronics = many and electric power = low) Then non-electricity saving
13.	If (Number of Electronics = many and electric power = are) Then non-electricity saving
14.	If (Number of Electronics = many and electric power = high) Then non-electricity saving

Algorithm performance results

After a rule formed calculating by the C4.5 algorithm, this system also conducts tests to see the performance of the C4.5 algorithm by implementing a confusion matrix as shown in **Table 1**. The results of the system performance test are presented in **Figure 3**.



Analisis Pengujian

TRUE POSITIVE 7/15

TRUE NEGATIVE 4/15

FALSE POSITIVE 1/15

FALSE NEGATIVE 3/15

PRECISION 87.5%

RECALL 70%

ACCURACY 73.333%

Figure 3 Performance C4.5 algorithm.

In this test, the test carried out 5 times to see the differences and performance of the test model. The test model applied in this test is to create changing training data and its results are shown in **Table 6**.

Table 6 system performance test comparison results.

No	Amount of training data	Number of rules	Number of test data	System performance results (%)		
				Precision	Recall	Accuracy
1	15	6	55	69.44 %	67.56 %	58.2 %
2	25	13	45	64.29 %	69.231%	60 %
3	35	18	35	72.73 %	69.57 %	62.86 %
4	45	10	25	65 %	86,675	64 %
5	55	14	15	87.5 %	70 %	73.33 %
6	1000	14	2800	90.3 %	74.4 %	76.22 %

The results show that the number of formed rules varies because of the variation of the training data number. In addition, the value for precision, recall, and accuracy of each test also has different values. In this study, the highest accuracy value is the 6th test with 1000 training data and 2800 test data.

Conclusions

The amount of applied training data strongly influenced the number of formed rules through the application of the C4.5 algorithm. The highest accuracy is in testing with the amount of training data as much as 1000 and test data 2800 producing a precision of 90.3 %, a recall of 74.4 % with an accuracy of 76.22 %. In this research, it has not carried out algorithm validation, so for the future work, it is necessary to validate the algorithm and increase the number of datasets.

References

[1] T Jasiński. Modeling electricity consumption using nighttime light images and artificial neural networks. *Energy* 2019; **179**, 831-42.

[2] Ministry of Energy and Material Resources. *Handbook of energy & economic statistics of Indonesia.* Ministry of Energy and Material Resources, Jakarta, Indonesia, 2018, p. 21-40.

[3] CY Lee, S Kaneko and A Sharifi. Effects of building types and materials on household electricity consumption in Indonesia. *Sustain. Cities Soc.* 2020; **54**, 101999.

[4] A Prashar. Adopting PDCA (Plan-Do-Check-Act) cycle for energy optimization in energy-intensive SMEs. *J. Clean. Prod.* 2017; **145**, 277-93

[5] JY Kim and SB Cho. Electric energy consumption prediction by deep learning with state explainable autoencoder. *Energies* 2019; **12**, 739.

[6] Z Zheng, H Chen and X Luo. Spatial granularity analysis on electricity consumption prediction using LSTM recurrent neural network. *Energ. Proc.* 2019; **158**, 2713-8.

[7] Z Chang, Y Zhang and W Chen. Electricity price prediction based on hybrid model of adam optimized LSTM neural network and wavelet transform. *Energy* 2019; **187**, 115804.

[8] A Saleh. Implementasi metode klasifikasi naïve bayes dalam memprediksi besarnya penggunaan listrik rumah tangga. *Creativ. Inform. Tech. J.* 2015; **2**, 207-17.

[9] X Meng, P Zhang, Y Xu and H Xie. Construction of decision tree based on C4.5 algorithm for online voltage stability assessment. *Int. J. Electr. Power Energ. Syst.* 2020; **118**, 105793.

[10] U.S. Department of Energy and SF Baldwin. *Quadrennial techonology review: An assesment of energy techonolgies and research oppurtunities*. U.S. Department of Energy, Washington DC, 2015.

[11] U Surahman, J Maknun and E Krisnanto. Survey on household energy consumption of public apartments in Bandung City , Indonesia. *In*: Proceedings of the 8th International Conference on Architecture Research and Design, Surabaya, Indonesia. 2016, p. 181-7.

[12] T Kubota, U Surahman and O Higashi. A comparative analysis of household energy consumption in Jakarta and Bandung. *In*: Proceedings of the 30th International Conference on Passive and Low Energy Architecture, Ahmedabad, India. 2014, p. 260-7

[13] MA McNeil, N Karali and V Letschert. Forecasting Indonesia's electricity load through 2030 and peak demand reductions from appliance and lighting efficiency. *Energ. Sustain. Dev.* 2019; **49**, 65-

77.

[14] H Batih and C Sorapipatana. Characteristics of urban households' electrical energy consumption in Indonesia and its saving potentials. *Renew. Sustain. Energ. Rev.* 2016; **57**, 1160-73.

[15] M Molina-Solana, M Ros, MD Ruiz, J Gómez-Romero and MJ Martin-Bautista. Data science for building energy management: A review. *Renew. Sustain. Energ. Rev.* 2017; **70**, 598-609.

[16] C Li, Z Ding, D hao, J Yi and G Zhang. Building energy consumption prediction: An extreme deep learning approach. *Energies*. 2017; **10**, 1-20.

[17] AZ Dobaev, MP Maslakov and AA Dedegkaeva. Development of decision support system for data analysis of electric power systems. *In*: Proceedings of the 2nd International Conference on Industrial Engineering, Applications and Manufacturing. Chelyabinsk, Russia, 2016, p. 1-4.

[18] H Murata and T Onoda. Estimation of power consumption for household electric appliances. *In*: Proceedings of the 9th International Conference on Neural Information Processing. Singapore, 2002, p. 2299-303.

[19] XM Zhang, K Grolinger, MAM Capretz and L Seewald. Forecasting residential energy consumption: Single household perspective. *In*: Proceedings of the 17th IEEE International Conference on Machine Learning and Applications. Orlando, USA, 2019, p. 110-7.

[20] S Omran and EMF El-Houby. Prediction of electrical power disturbances using machine learning techniques *J. Ambient Intell. Humaniz. Comput.* 2020; **11**, 2987-3003.

[21] PV Ngoc, CVT Ngoc, TVT Ngoc and DN Duy. A C4.5 algorithm for english emotional classification. *Evol. Syst.* 2019; **10**, 425-51.

[22] CP Balasubramaniam and R Gunasundari. Improved C4.5: An agent-based supply chain management system. *J. Theor. Appl. Inf. Technol.* 2018; **96**, 555-567.

[23] BA Tama. Data mining for predicting customer satisfaction. *J. Theor. Appl. Inf. Technol.* 2015; **75**, 3-7.

[24] TM Ahmed. Using data mining to develop model for classifying diabetic patient control level based on historical medical records. *J. Theor. Appl. Inf. Tech.* 2016; **87**, 316-23.

[25] AH Mohammad. Comparing two feature selections methods (Information gain and gain ratio) on three different classification algorithms using arabic dataset. *J. Theor. Appl. Inf. Tech.* 2018; **96**, 1561-9.

[26] P Gu and Q Zhou. Student performances prediction based. *Emerg. Comput. Inf. Technol. Educ.* 2012; **146**, 1-8.

[27] J Quinlan. *C4.5 : Programs for machine learning*. Morgan Kaufmann, San Franscisco, 1993.

[28] F Gorunescu. *Data mining: Concepts, models and techniques*. Springer, London, 2011, p. 319-30.