Classification of Farmer's Eligibility as Recipients of Subsidized Fertilizer Assistance with C4.5 Algorithm

IOP Conf. Series: Materials Science and Engineering **771** (2020) 012029 doi:10.1088/1757-899X/771/1/012029

Classification of Farmer's Eligibility as Recipients of Subsidized Fertilizer Assistance with C4.5 Algorithm

Diah Nur Yunita[1], Anna Hendri Soleliza Jones[2*], Dewi Soyusiawaty[3]
Department of Informatics, Universitas Ahmad Dahlan, Indonesia

[1]diahnuryunita@gmail.com, [2,*]annahendri@tif.uad.ac.id,
[3]dewi.soyusiawaty@tif.uad.ac.id,

Abstract. Farmers are the main actors in agricultural development, have a major contribution in realizing Indonesia's agricultural vision. To support the realization of agricultural production and productivity targets, the government's supports is by issuing policies to provide subsidized fertilizer assistance. The criteria for farmers who are eligible to receive subsidized fertilizer assistance are having land less than 1.9 hectares, have been members of farmer groups, and having work as a farmer or a private employee. Data collection was carried out using the method of observation and interviews with farmer groups in Gunung Kidul. Data classification is done by applying the C4.5 algorithm to 300 data of prospective farmers receiving subsidized fertilizer assistance. Application development includes interface design, making data flow diagrams, implementation and testing. From the research that has been done by applying the C4.5 Algorithm for farmer data, there are 300 data with 113 training data. Produce a data classification of farmers who are eligible to receive assistance at 95,58% accurate and targeted to the needs of farmers. This means that the C4.5 algorithm can be used as a classification for determining the eligibility of farmers to receive subsidized fertilizer.

1. Introduction

Farmers are one of the main actors in agricultural development that has a major contribution in realizing production and productivity goals to achieve the target of managing agricultural products in Indonesia. Agricultural development policies continue to be reviewed and improved, one of the government's policies to improve the quality of agricultural products is to provide subsidized fertilizer assistance to farmers. Based on the Regulation of the Minister of Agriculture No. 82 / Permentan / OT.140 / 8/2013 concerning the Development of Farmers Groups and Farmers Group Combinations, guidelines for the preparation of the Definitive Needs Group (RDKK) have been arranged. RDKK is made through the coordination of Gapoktan (Joint Farmers Group).

According to Regulation of the Minister of Agriculture No. 273 / Kpts / ot.160 / 4/2007 concerning guidelines for institutional development of farmers, Gapoktan is a collection of several farmer groups that join and work together to increase economies of scale and business efficiency. The existence of Gapoktan so that farmer groups can be more efficient and effective, and provide agricultural production facilities, improvement, capital, or expansion of farming for farmers and farmer groups from the upstream and downstream sectors, as well as increased cooperation and product marketing. And through the RDKK instructions for subsidized fertilizer are prepared as a reference for farmer groups, farmers, officers and agricultural service officials in the distribution of subsidized fertilizer. RDKK as the basis for redemption of fertilizer to retailers is not yet fully implemented as expected. Although subsidized fertilizer assistance has been implemented for a long time, the problems found in the implementation of subsidized fertilizer assistance are very complex in terms of data collection and distribution/distribution, so that often farmers who should get subsidized fertilizer products do not

receive them. RDKK is not yet fully operational. Farmers have difficulty in obtaining the subsidized fertilizer product because the price of non-subsidized fertilizer is quite expensive for farmers.

Based on the background that has been described and previous studies that have been carried out using the C4.5 algorithm in determining the classification of data such as the Khairul Zaman study in 2016 on how to determine the eligibility of beneficiaries for social rehabilitation assistance for uninhabitable homes, Muhammad Arif Rahman's 2015 study was determining scholarship recipients using the C4.5 method, Wisti Dwi Septiani's research in 2014 the application of the C4.5 algorithm for prediction of hepatitis and 2017 Hikma Widayu research on data mining to predict the types of customer transactions in savings and loan cooperatives with the C4.5 algorithm. Based on the problems that have been discussed, the author tries to find the decision tree rules using the C4.5 algorithm to the data of prospective fertilizer recipients in Gunungkidul Regency in a study entitled "Classification of Determination of Fertilizer Aid Recipients with the C4.5 Algorithm Technique". The knowledge generated is expected to help the authorities in making policies, so that in the future the recipients of fertilizer assistance are right on target.

2. Method

2.1. Preprocessing Text

The process of data classification is the process of sorting data according to similarities or differences in the data held. Before the classification is done on the data, it is necessary to preprocess the data. Preprocessing data is to take steps to clean up the data so that the results of data mining will be more accurate. The data preprocessing process has the following steps: [1]

a. Data cleaning
   Data cleaning is to get rid of inconsistent or noise data, one of the stages in data mining in the data cleaning process is very important because inconsistent data can cause invalid mining results. In general, the data obtained from the database and the results of experiments contained imperfect entries and there are data attributes that are not relevant to the data mining hypothesis that they have.
b. Data transformation (data is converted into a form suitable for mining)
   Data is modified or combined into a format suitable for processing in data mining. Data in the form of continuous numeric numbers need to be divided into several intervals. (Han, 2006).
c. Application of data mining techniques
   The main process in the method applied to find valuable and hidden knowledge from data.
d. Evaluate the patterns found
   The final stage of the data mining process is to introduce patterns to the results of the algorithm. Pattern recognition is a discipline that studies ways to classify objects into classes or categories and recognizes data trends.

2.2. Algorithm C4.5

C4.5 algorithm is a tree structure where there are nodes that describe the attributes that are tested, and each leaf represents a class. C4.5 algorithm recursively visits each decision node, selects the optimal division, until it cannot be subdivided. C4.5 algorithm uses the concept of information gain or entropy reduction to choose the optimal division. Here is the gain and entropy formula:

2.2.1. *Gain (S, A)* is the acquisition of information from attribute A relative to the data output of S. The acquisition of information is obtained from the output data or the dependent variable S which is grouped by attribute A, denoted by the gain (S, A). [4]

$$\text{Gain (S,A)} = \text{Entropy (s)} - \sum_{i=1}^{n} - \frac{|si|}{|s|} * \text{entropy (si)} \qquad \ldots\ldots\ldots\ldots (1)$$

Information:
S: case set
A: attributes
n: number of attributes A
|Si|: number of cases on the i-th partition
|S|: number of cases in S

2.2.2. *Entropy*(S) is the estimated number of bits needed to be able to extract a class (+ or -) from a random amount of data in the sample space S. Entropy can be said to be the need for bits to

express a class. the smaller the value of Entropy, the more Entropy is used in extracting a class. Entropy is used to measure the authenticity of S. [4]

$$\text{Entropy (S)} = \sum_{j=1}^{k} - pj \log 2\ pj \quad \ldots\ldots\ldots\ldots\ldots (2)$$

Information:
S: The set of cases
k: number of partitions S
Pj: Probabilities derived from the number (yes/no) divided by the total cases
After calculating the gain and entropy, the following are the steps of the decision tree construction using the C4.5 algorithm.

2.2.3. *Confusion Matrix*

Accuracy shows the closeness of the value of the measurement results with the actual value. To determine the level of accuracy, it is necessary to know the true value of the measured parameter, then the data is known to what level of accuracy. To analyze performance, a matrix is used by comparing the original class data with predictions from the input data or called a confusion matrix. [9]
To do confusion matrix calculations use the following formula:

1. *Recall* is the proportion of positive cases that are correctly identified.

$$Recall = \frac{d}{(c+d)} \quad \ldots\ldots\ldots\ldots\ldots (3)$$

2. *Precision* is the proportion of cases that are correctly identified with the positive result

$$Precision = \frac{d}{(b+d)} \quad \ldots\ldots\ldots\ldots\ldots (4)$$

3. *Accuracy* is a comparison of correctly identified cases with a total number of cases.

$$Accuracy = \frac{(a+d)}{(a+b+c+d)}$$

Attribute Description:
a: If the prediction result is positive and the data is actually positive.
b: if the prediction result is positive while the actual value is negative.
c: if the result of the prediction is negative while the actual value is positive.
d: if the predicted result is negative and the actual value is negative

3. Results and discussion

3.1. Result and Discussion

The data in this research were obtained directly from the chairman of the Gunungkidul Farmers Group. There are 300 farmers data, with 3 criteria as support in determining the decision of whether farmers are eligible or not eligible for receiving subsidized fertilizer assistance. The 3 criteria used are; Land Area (Land), Occupation and the membership of farmer groups or not. Table 1 is a sample of data obtained.

Table 1. Sample Data of The Farmers.

No	Name	Land Area (Land)	Occupation	Member of Farmers Group
1	Kirnorejo	0,1	CS	No
2	Pariyem	0,2	Farmer	Yes
3	Nurhayadi	1,9	CS	Yes
4	Pujo. W	0,3	Farmer	No
5	Pujiyanto	0,5	Employee	No
6	Lasiyanto	0,3	Farmer	Yes
7	Turiyo	0,5	Farmer	Yes
8	Sukismo	0,1	Farmer	Yes
9	Tentrem	0,1	Farmer	Yes
10	Dedi Martanto	1,7	Farmer	Yes

Data processing is done by grouping the criteria used in a more efficient form. Criteria are grouped as follows:

1. Land

Land criteria are grouped based on the wide-area of land owned by farmers. The area will be divided into 3 groups, such as S for Small at a range of 0 – 0,9 Ha, M for Medium at a range of 1 – 1,9 Ha and L for large at more than 2 Ha.

2. Occupation
   This criterion explains the variables that will be used to classify prospective recipients of subsidized fertilizer with the definition of job criteria. The variables are "Employee*"* for the farmer who has other jobs aside from a farmer, variable "Farmer" to describe a farmer and variable "CS" to describe a farmer who their main job as a civil servant.
3. Member of Farmers Group
   This criterion explains the variables that will be used to classify prospective subsidized fertilizer recipients with the definition of the criteria whether the farmer is a member of the farmer group or does not belong to the farmer group, the variable will be "Yes" for the member of farmer group and "No" if not a member of farmer group.

After defining the next variable this research will classify the data by applying the C4.5 algorithm, by finding the highest gain value of the 300 existing data using Equation 2 and Equation 1 which have been described in Chapter 3. The calculation steps to find the gain are as follows:

Step 1: Calculate the total entropy. After we get the total entropy next, do calculate the entropy for each criterion, as examples:

a. Entropy (Total) $= \left(-\frac{30}{113} * log\left(\frac{30}{113}\right)\right) + \left(-\frac{83}{113} * log\left(\frac{83}{113}\right)\right)$
   $= \left(-\frac{30}{113} * 0{,}2654\right) + \left(-\frac{83}{113} * -0{,}7345\right) = 0{,}5080 + 0{,}3269 = 0{,}8349$

b. Entropy (Land area, Large) $= \left(-\frac{5}{9} * log\left(\frac{5}{9}\right)\right) + \left(-\frac{4}{9} * log\left(\frac{4}{9}\right)\right)$
   $= \left(-\frac{5}{9} * 0{,}8481\right) + \left(-\frac{4}{9} * 1{,}1700\right) = -0{,}4711 + -0{,}52 = 0{,}9911$

c. Entropy (Land area, Medium) $= \left(-\frac{11}{20} * log\left(\frac{11}{20}\right)\right) + \left(-\frac{9}{20} * log\left(\frac{9}{20}\right)\right)$
   $= \left(-\frac{11}{20} * -0{,}8624\right) + \left(-\frac{9}{20} * -1{,}1520\right)$
   $= -0{,}47432 + -0{,}5184 = 0{,}99272$

d. Entropy (Land area, small) $= \left(-\frac{14}{84} * log\left(\frac{14}{84}\right)\right) + \left(-\frac{70}{84} * log\left(\frac{70}{84}\right)\right)$
   $= \left(-\frac{14}{84} * -0{,}7369\right) + \left(-\frac{70}{84} * -0{,}2630\right)$
   $= -0{,}1228 + -0{,}2191 = 0{,}6500$

Repeat the calculation for each of criterions.

Step 2: Calculate the total gain. Use formula 2 to get the total gain, the example can be seen below:

a. Gain (total, Land area)=Entropy (total) $\sum_{i=1}^{n} \frac{|land\ area|}{|total|}$ * Entropy (Land area)
   $= 0{,}8349 - \left(\frac{9}{113} * 0{,}9910\right) + \left(\frac{20}{113} * 0{,}9927\right) + \left(\frac{84}{113} * 0{,}6500\right)$
   $= 0{,}8349 - (-0{,}0789) + (-0.1756) + (-0{,}4831) = 0{,}0970$

b. Gain (total, member of farmer group)=Entropy (total) $\sum_{i=1}^{n} \frac{|Member\ of\ farmer\ group|}{|total|}$*Entropy (member of farmer group)
   $= 0{,}8349 - \left(-\frac{84}{113} * 0{,}2222\right) + \left(-\frac{29}{113} * 0{,}3620\right)$
   $= 0{,}8349 - (-0{,}1651) + (-0{,}0929) = 0{,}5767$

c. Gain (total, Occupation)=Entropy (total) $\sum_{i=1}^{n} \frac{|Occupation|}{|total|}$ * Entropy (Occupation)
   $= 0{,}8349 - \left(-\frac{14}{113} * 0\right) + \left(-\frac{79}{113} * 0{,}0979\right) + \left(-\frac{20}{113} * 0{,}8112\right)$
   $= 0{,}8349 - 0 + 0{,}0684 + 0{,}1435 = 0{,}6228$

The results of the calculation of entropy and gain values can be seen in Table 2.

Table 2. Calculation of algorithm C4.5

Node			Total Cases	No	Yes	Entropy	Gain
1	**Total**		300	85	215	0,8599	
	Land Area	Large	21	14	7	0,9183	0,0456
		Medium	89	30	59	0,9219	
		Small	190	41	149	0,7524	
	Member of farmer group	Yes	227	18	209	0,3997	0,5575
		No	73	68	5	0,3603	
	Occupation	CS	38	38	0	0	0,5420
		Farmer	207	8	199	0,2360	
		Employee	55	40	15	0,8453	

From the calculation of Entropy and Gain values, as shown in table 3.2, it can be seen that the attribute that has the highest Gain is the work attribute that is equal to 0.6228. The work attribute will be the root node which will be the basis of the classification of data in the form of a decision tree.
Decision tree resolution is obtained by finding the largest gain from the rest of the data that has been used in the first node decision tree search. Gain search will continue to be repeated until there is no more data that can still be broken down to look for gain. Thus, the results of the decision tree in this study can be seen in Figure 1.

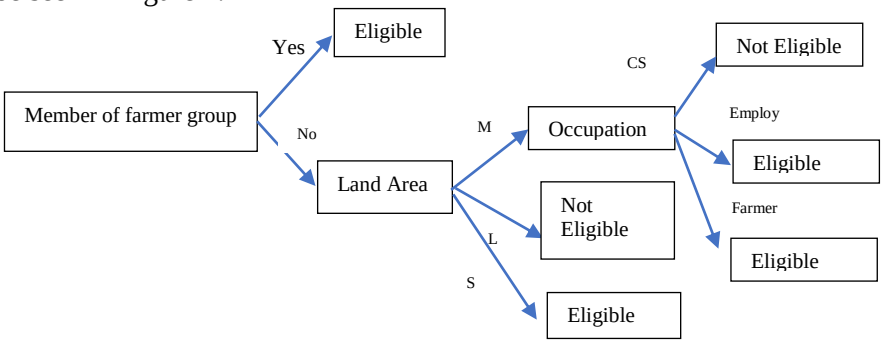

Figure 1. Decision Tree Classification of Farmers.

With the production of the classification tree of farmers, the search for the determination of farmers who are not suitable or not eligible to receive subsidized fertilizer can be done more efficiently.

3.2. System Testing

Accuracy testing is performed using 300 training data and 113 test data. Test results obtained:

Table 3. Accuracy of test result

Accuracy	True: Eligible	True: Not Eligible	Precision
Eligible Prediction	82	4	95,35%
Not Eligible Prediction	1	26	96,30%
Recall	98,80%	86,67%	

From Table 3, it was found that with 300 training data and 113 precision test data the decision of farmers who were eligible to receive subsidized fertilizer assistance was 95.35% with an accuracy value of 95.58%.

4. Conclusion

From the research that has been done, the result of C4.5 Algorithm has an effect when applied in the data of farmers to determine whether farmers are eligible or not eligible to receive subsidized fertilizer assistance. By using 300 farmer data as training data and 113 test data. The study resulted in the classification of data of farmers who met the requirements to receive assistance with the accuracy of the target accuracy of 95.58% subsidized fertilizer recipients. This means that the C4.5 algorithm can be used as a classification to determine the eligibility of farmers to receive subsidized fertilizer.

Reference

[1] Anwar, N., Pranolo, A., & Kurnaiwan, R. (2018). *Grouping the community health center patients based on the disease characteristics using C4.5 decision tree. IOP Conference Series: Materials Science and Engineering, 403, 012084.* doi:10.1088/1757-899x/403/1/012084

[2] ARIBOWO, Agus Sasmito; CAHYANA, Nur Heri. Feasibility study for banking loan using association rule mining classifier. **International Journal of Advances in Intelligent Informatics**, [S.l.], v. 1, n. 1, p. 41-47, mar. 2015. ISSN 2548-3161. Available at: <http://ijain.org/index.php/IJAIN/article/view/8%7Cto_array%3A0>. Date accessed: 09 oct. 2019. doi:https://doi.org/10.26555/ijain.v1i1.8.

[3] J. A. Suyatno, F. Nhita and A. A. Rohmawati, "Rainfall Forecasting in Bandung Regency Using C4.5 Algorithm," *2018 6th International Conference on Information and Communication Technology (ICoICT)*, Bandung, 2018, pp. 324-328. doi: 10.1109/ICoICT.2018.8528725

[4] L. Dongming, L. Yan, Y. Chao, L. Chaoran, L. Huan and Z. Lijuan, "The application of decision tree C4.5 algorithm to soil quality grade forecasting model," *2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI)*, Wuhan, 2016, pp. 552-555.doi: 10.1109/CCI.2016.7778985

[5] R. K. Amin, Indwiarti and Y. Sibaroni, "Implementation of decision tree using C4.5 algorithm in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region)," *2015 3rd International Conference on Information and Communication Technology (ICoICT)*, Nusa Dua, 2015, pp. 75-80.doi: 10.1109/ICoICT.2015.7231400

[6] S. Das, S. Dahiya and A. Bharadwaj, "An online software for decision tree classification and visualization using c4.5 algorithm (ODTC)," *2014 International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, 2014, pp. 962-965. doi: 10.1109/IndiaCom.2014.6828107

[7] Siahaan H., Mawengkang H., Efendi S., Wanto A. and Windarto A. P. 2018 Application of Classification Method C4.5 on Selection of Exemplary Teachers *IOP Conference Series* 1-6

[8] SIA, Florence; ALFRED, Rayner. Tree-based mining contrast subspace. International Journal of Advances in Intelligent Informatics, [S.l.], v. 5, n. 2, p. 169-178, july 2019. ISSN 2548-3161. Available at: <http://ijain.org/index.php/IJAIN/article/view/359%7Cto_array%3A0>.Date accessed: 09 oct. 2019. doi:https://doi.org/10.26555/ijain.v5i2.359.

[9] S. Visa, Brian. Ramsay, A. Ralescu, and E.V.D. Knaap, "Confusion matrix-based feature selection," Midwest Artificial Intelligence and Cognitive Science Conference, vol. 710, April 2011.

[10] Wahyuni, Sri & Saputra S, Kana & Iswan, Mochammad. (2017). THE IMPLEMENTATION OF DECISION TREE ALGORITHM C4.5 USING RAPIDMINER IN ANALYZING DROPOUT STUDENTS.

[11] Xuefei Wang and Yan Shi, "Design and implementation of targeting advertising system based on C4.5 algorithm," *2015 4th International Conference on Computer Science and Network Technology (ICCSNT)*, Harbin, 2015, pp. 669-672.doi: 10.1109/ICCSNT.2015.7490833