International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 02, Issue 03, August 2021
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

Implementation of Data Mining Using C4.5 Algorithm for Predicting Customer Loyalty of PT. Pegadaian (Persero) Pati Area Office

1st Ridlo Muttaqien, 2nd Musthofa Galih P, 3rd Andri Pramuntadi
[1,2,3]Computer Science/Informatics
[1,2,3]Alma Ata University Yogyakarta, Indonesia
173200015@almaata.ac.id, mgalihpradana@almaata.ac.id,andripramuntadi@almaata.ac.id.

Abstract*—PT Pegadaian (Persero) is engaged in the business of providing credit services with pawn, non-pawning and gold investment products. One of the right marketing strategies to survive today's high competition is to maintain customer loyalty, researchers use several data variables available in the MIS (Management Information System) in the form of customer transaction frequency, how many products are taken by customers, customer satisfaction and direct interviews. to predict customer loyalty of PT Pegadaian (Persero) by implementing the c4.5 algorithm. The c4.5 algorithm is the algorithm used to create a decision tree. Decision trees are a very powerful and well-known method of classification and prediction. The decision tree method converts very large facts into a decision tree that represents the rule. Rules can be easily understood in natural language. This study aims to determine the accuracy of the C4.5 algorithm to predict customer loyalty of PT Pegadaian (Persero) and the most influential factors in loyalty. The results of the experimental application of the c4.5 algorithm show that the level of accuracy generated in predicting customer loyalty is quite high, namely 89.94% in data testing 1 and 94% in data testing 2. The application of the c4.5 algorithm in predicting customer loyalty of PT Pegadaian (Persero) can be well applied.

Keywords : *C4.5 algorithm, Loyalty, Rapidminer*.

I. INTRODUCTION

PT Pegadaian (Persero) is one of the many financial institutions managed by BUMN. PT Pegadaian (Persero) is engaged in the business of providing credit services to Indonesian citizens with pawn, non-pawning and gold investment products. Currently, PT Pegadaian (Persero) is experiencing very tight challenges. These challenges arise as many credit lender services appear that offer lower interest rates and easy terms. PT Pegadaian (Persero) has challenges on how to retain customers.

Loyalty is a deeply held commitment to repurchase or subscribe to products and services in the future despite situational influences. According to [1] there are six reasons why companies need to maintain customer loyalty. First: existing customers are more prospective, meaning that loyal customers will provide large benefits to the institution. Second: the cost of acquiring new customers is much greater than keeping and retaining existing customers. Third: customers who already believe in the company in one matter will believe in other matters. Fourth: the company's operating costs will be efficient if it has many loyal customers. Fifth: the company can reduce psychological and social costs because old customers have had many positive experiences with institutions. Sixth: loyal customers will always defend the company and even try to attract and advise others to become customers.

PT Pegadaian (Persero) has a web-based information system, namely MIS (Management Information System) in which the company can monitor information used in planning, controlling, and continuous improvement such as increasing and decreasing customer status with several requests and reasons, and customer non-performing credit records. , and customer transaction records. The MIS (Management Information System) that is running is already computerized and already has a database to store company management data.

To find out the factors that influence customer loyalty from PT Pegadaian (Persero) customers based on the data in the MIS (Management Information System), the researcher uses several data variables available in the MIS (Management Information System) in the form of customer transaction frequency, how many products are available. taken by customers and conduct interviews directly to predict customer loyalty of PT Pegadaian (Persero) by implementing the c.45 algorithm to calculate the data.

II. RESEARCH METHODS

The author uses the C 4.5 algorithm on the basis of some literature from previous research from Teguh Budi Santoso with the title "Analisa Dan Penerapan Metode C4.5 Untuk Prediksi Loyalitas Pelanggan" with the results showing that classification with the C4.5 algorithm gets an accuracy of up to 97.5%, which means that the exact C4.5 algorithm is used to calculate the level of customer loyalty [1]. The second study entitled "Penerapan Algoritma C4.5 Untuk Prediksi Loyalitas Nasabah PT Erdika Elit Jakarta" by Khotibul Umam, based on the decision tree that has been made the attribute that has the most influence on customer loyalty is educational background because it has the highest gain value, namely 1.545292721 and as the root of the decision tree while the gender of the customer does not have much effect on customer loyalty because it is always at the last node with a gain value of 0.623919119 [2]. The third study entitled "Penerapan algortima c4.5 untuk penentuan kelayakan kredit" by Siti Nur Khasanah, Based on the results of the application of the C4.5 classifier algorithm, it can be concluded that to determine credit worthiness whether a prospective customer will become a customer with smooth or problematic payments using the C4.5 algorithm classifier and the accuracy of the C4.5 classifier algorithm using training data for 270 customers is 88.52% [3]. The fourth study entitled "Pengambilan Keputusan Pegawai Tidak Tetap Menjadi Pegawai Tetap

Dengan Decission Tree" by Febryantahanuji, By using the C4.5 algorithm and testing methods using X-Fold Cross Validation can help companies to create a decision support system for the appointment of non-permanent employees to become permanent employees , because the main indicators are obtained in determining permanent employees [4]. The fifth study entitled "Penerapan Algoritma C4.5 untuk Penentuan Kelayakan Pemberian Kredit" by Teguh Budi Santoso, that the results of the classification using the C4.5 Algorithm show that an accuracy of 97.5% is obtained, based on the results obtained. The results obtained indicate that the c4.5 algorithm is suitable to be used to determine the feasibility of giving [5]. The sixth study entitled "Implementasi Algoritma C4. 5 terhadap Kepuasan Pelanggan, by Harry Dika, the results of this study resulted in the formation of a fast food restaurant satisfaction model, the level of accuracy produced was quite high and included in the very good category [6].

Data mining is a method for finding data in a large database. Data mining is the process of analyzing and extracting data in the database to obtain useful new data, become new patterns or patterns, and to calculatepredictions. Classification and prediction are two forms of data analysis that can be used to extract models from data containing classes or to predict future data trends. Classification predicts data in the form of categories, while prediction models functions of continuous values. For example, a classification model can be made to classify loan applications to banks as risky or safe, while a predictive model can be made to predict expenses for buying computer equipment from potential customers based on their income and location of residence.[7]

C4.5 algorithm is an algorithm used to generate a decision tree. The workflow of the C4.5 algorithm is making a decision tree based on the selection of the attribute that has the highest priority or that has the highest gain value based on the value of the entropy attribute as the axis of the classification attribute. At this stage the C4.5 algorithm has 2 working rules, namely: Making a decision tree, and making rules (rule model). The rules formed from the decision tree will form a condition in the form of if then.[8]

Calculate *entropy* :

$$Entropy(S) = \sum_{i=1}^{n}(-p_i) * log_2(p_i)$$

description :

S : case set.

N : number of partitions S.

Pi : Proportion of Si to S.

Calculate the Gain value with the information gain method:

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

description :

S : Case set.

A : Attribute

n : Number of attribute partitions A.

|Si| : Number of cases on partition-i

|S| : Number of cases in S.

Figure 1. C4.5 Algoritm Formula

2.1 Design

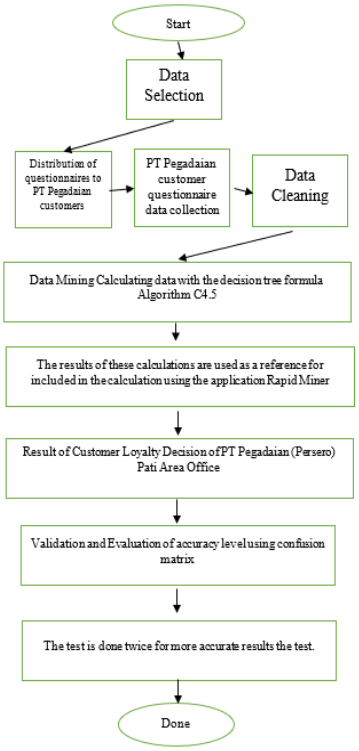

Figure 2. Research Design

2.2 Research Variables

In this study the authors use research variables to measure customer loyalty of PT Pegadaian (Persero) which is divided into six categories, namely: frequency of transactions, how many products are taken, customer satisfaction, age and gender of customers. The following variables are used and their descriptions :

Table 1. Research Variables Table

Category	Information
Transaction frequency	How often do customers transact
Product taken	The number of products taken by customers
Customer satisfaction	Customer satisfaction level
Gender	Customer gender
Age	Age of customer

III. RESULT AND ANALYSIS

3.1 Data Selection

The data used in this study is the source of data from MIS Pegadaian and questionnaires. Data were obtained from the distribution of 166 questionnaires. Questionnaire data is data distributed to customers of PT Pegadaian (Persero) Pati area office in March – June 2021 with variable frequency of transactions, how many products are taken, customer satisfaction level, age and gender of the customer.

3.2 Data Cleaning

The data obtained from the respondents will be selected first, in order to obtain data that is in accordance with the

International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 02, Issue 03, August 2021
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

needs of the researcher. Of the 200 data obtained, 170 data were complete and correct, the remaining 166 data were not included because there was incomplete data when filling out the questionnaire and it was not filled in according to the existing instructions.

3.3 Data Training

The results of the training data are used to obtain the results of the level of customer loyalty in the form of a decision tree.

Table 2. Data Training

Age Of Customer	Gender Of Customer	Product Taken	Transaction Frequency	Customer Satisfication	Result Of loyalty
Adult	Male	1	Often	Satisfied	Yes
Adult	Male	1	Often	Satisfied	Yes
Old	Female	1	not often	Netral	No
Old	Male	2	not often	Netral	Yes
Adult	Female	1	not often	Satisfied	Yes
Teenager	Male	2	Often	Satisfied	Yes
Teenager	Male	2	not often	Not satisfied	No
Teenager	Female	1	Often	Not satisfied	No
Old	Female	2	Often	satisfied	Yes
....	.....	....	.....	.....	.....

Then calculate the entropy of each attribute and gainvalue with the C 4.5 algorithm formula which results:

Table 3. Result of Decision Tree

Attribute		Total Case	No	Yes	Entropy	Gain
Total		66	28	38	0,98337619	
Umur						0,02148474
	Teenager	16	7	9	0,988699408	
	Adult	39	14	25	0,941828535	
	Old	11	5	6	0,994030211	
Customer Gender						0,00442794
	Male	36	14	22	0,964078765	
	Female	30	14	16	0,996791632	
Product Taken						0,247643891
	1	41	26	15	0,947435136	
	2	18	1	17	0,309543429	
	3	7	1	6	0,591672779	
Frequency Transaction						0,414352159
	Often	32	2	30	0,337290067	
	Not Often	34	26	8	0,787126586	
Customer Satisfication						0,479549994
	<=2	32	26	6	0,69621226	
	=>2	34	2	32	0,322756959	

From the results of the manual calculation above, it is found that the customer satisfaction variable gets the highest gain value and is used as the root node of the decision tree and is calculated again with the rapidminer application.

Figure 3. Decision Tree

3.4 Data Testing 1

This test is done by dividing as much as 2 parts of the testing data to be tested. Test 1 uses 50 training data, and in this test will produce values of accuracy, *precision,*and *recall.*

Table 4. Data Testing 1

Age Of Customer	Gender Of Customer	Product Taken	Transaction Frequency	Customer Satisfication	Result Of loyalty
Adult	Male	1	Often	Satisfied	Yes
Adult	Male	1	Often	Satisfied	Yes
Old	Female	1	not often	Netral	No
Old	Male	2	not often	Netral	Yes
Adult	Female	1	not often	Satisfied	Yes
Teenager	Male	2	Often	Satisfied	Yes

From the decision tree implementation table above, then enter the validation process with x validation in rapidminer. And the result is as follows:

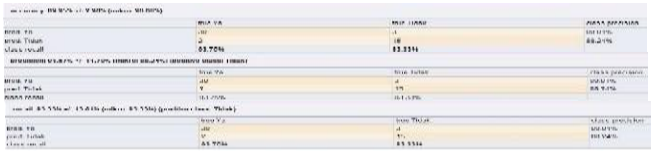

Figure 4. Calculation of Rapid Miner Data Testing 1

From the data validation process testing 1, the results obtained are: True Positive (TP) totaling 30 data, True Negative (TN) 15 data, False Positive (FP) 3 data, False Negative (FP) 2 data, which means that TP has 30 data with the result loyal and loyal prediction results, FP 3 data with loyal results but disloyal predictions, FN 2 data with disloyal results but loyal predictions, TN with disloyal results and disloyal predictions. And obtained 89.95% accuracy, 91.67% precision, and 83.33% recall.

Table 5. Result Rapid Miner Data Testing 1

N = 50	Aktual : Yes	Aktual : No
Prediction : Yes	TP: 30	FP: 3
Prediction: No	FN: 2	TN: 15

3.5 Data Testing 2

This test is done by dividing as much as 2 parts of the testing data to be tested. Test 2 uses 50 training data, and in this test will produce values of accuracy, *precision,*and *recall.*

Table 6. Data Testing 2

Age Of Customer	Gender Of Customer	Product Taken	Transaction Frequency	Customer Satisfication	Result Of loyalty
Adult	Male	1	Often	Satisfied	Yes
Adult	Male	1	Often	Satisfied	Yes
Old	Female	1	not often	Netral	No
Old	Male	2	not often	Netral	Yes
Adult	Female	1	not often	Satisfied	Yes
Teenager	Male	2	Often	Satisfied	Yes

From the decision tree implementation table above, then enter the validation process with x validation in rapidminer. And the result is as follows:

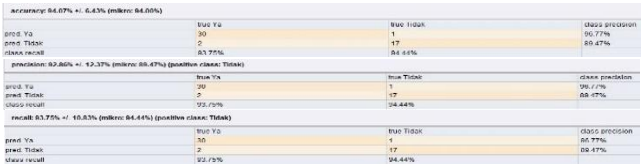

Figure 5. Calculation of Rapid Miner Data Testing2

From the data validation process testing 2, the results obtained: True Positive (TP) totaling 30 data, True Negative (TN) 17 data, False Positive (FP) 1 data, False Negative (FP) 2 data, which means that TP has 30 data with results loyal and loyal prediction results, FP 1 data with loyal results but disloyal predictions, FN 2 data with disloyal results but loyal predictions, TN 17 data with disloyal results and disloyal predictions. And obtained accuracy of 94.07%, precision of 92.86%, and recall of 94.44%.

Table 7. Result Rapid Miner Data Testing 2

N = 50	Aktual : Yes	Aktual : No
Prediction : Yes	TP: 30	FP: 1
Prediction : No	FN: 2	TN: 17

3.6 Result Customer Loyalty

Of all the sample data there are 166 data, showing that there are 102 data with loyal results and 64 data with disloyal results. And from the hypothesis that has been tested, it shows that the variable customer satisfaction is very influential in customer loyalty at PT Pegadaian (Persero) Pati Area Office.

Table 8. Result Customer Loyalty

Result	Total	Presentation
Loyal Customer	102	61,5%
Disloyal Customers	64	38,5%

IV.CONCLUSION

Based on trials, evaluations and research results that researchers have done from the data collected regarding the use of the c4.5 algorithm method for predicting customer loyalty at PT Pegadaian (Persero) Pati area offices, it can be concluded that:

- The use of the C4.5 algorithm method can be used to predict customer loyalty of PT Pegadaian (Persero) area office.
- The formation of a customer loyalty model with a decision tree and customer satisfaction variable has the highest gain value and greatly influences customer loyalty.
- From the results of the evaluation using the confusion matrix, a high level of accuracy was obtained, namely 90% in testing data testing 1 and 94% in testing data testing 2.

Based on the analysis that has been carried out and to improve performance and perfect the research that has been made, the researchers provide the following suggestions:

- Further research needs to be done by testing with other methods such as bayesian methods, clustering, neural networks, etc. to compare which results are more accurate.
- Adding more specific variables to customer loyalty for more accurate prediction results.

THANK-YOU NOTE

A big thank you to Alma Ata University which provided facilities in completing this research and to PT. Pegadaian (Persero) Kantor area Pati who provided theopportunity to conduct research and collect data.

REFERENCES

[1] T. B. Santoso, "ANALISA DAN PENERAPAN METODE C4.5 UNTUK PREDIKSI LOYALITAS PELANGGAN," *J. Ilm. Fak. Tek. LIMIT'S*, vol. 10, no. 1, 2011.

[2] K. Umam, D. Puspitasari, and A. Nurhadi, "Penerapan Algoritma C4.5 Untuk Prediksi Loyalitas Nasabah PT Erdika Elit Jakarta," *J. Media Inform. Budidarma*, vol. 4, no. 1, p. 65, 2020, doi: 10.30865/mib.v4i1.1652.

[3] S. N. Khasanah, "Penerapan algortima c4.5 untuk penentuan kelayakan kredit," vol. XIV, no. 1, pp. 9–14, 2017.

[4] H. D. P. Febryantahanuji, Irwan Sembiring, "Pengambilan Keputusan Pegawai Tidak Tetap Menjadi Pegawai Tetap Dengan Decission Tree," *Joural Informatics Educ.*, vol. 1, no. 2, pp. 26–37, 2018.

[5] T. B. Santoso and D. Sekardiana, "Penerapan

Algoritma C4.5 untuk Penentuan Kelayakan Pemberian Kredit," *J. Algoritm. Log. dan Komputasi*, vol. II, no. 1, pp. 130–137, 2019.

[6] H. Dhika, F. Destiawati, and A. Fitriansyah, "Implementasi Algoritma C4. 5 terhadap Kepuasan Pelanggan," pp. 80–86, 2018, doi: 10.31227/osf.io/fgc7a.

[7] S. Bayu, "DATA MINING KLASIFIKASI," *Conv. Cent. Di Kota Tegal*, vol. 4, no. 80, p. 4, 2018.

[8] A. R. Sukma, R. Halfis, and A. Hermawan, "Klasifikasi Channel Youtube Indonesia Menggunakan Algoritma C4.5," *J. Tek. Komput.*, vol. V, no. 1, pp. 21–28, 2019, doi: 10.31294/jtk.v4i2.