Jurnal Teknik Industri: Jurnal Hasil Penelitian dan Karya Ilmiah dalam Bidang Teknik Industri
Vol. 11, No. 1, 2025

Implementation of Data Mining to Classify Potential Customers Using the C5.0 Algorithm

Muhammad Rizki[1*], Cintya Nil Maghfirah[2], Fitra Lestari Norhiza[3], Nofirza[4], Fitriani Surayya Lubis[5]
Industrial Engineering Department, Universitas Islam Negeri Sultan Syarif Kasim Riau
Email: Muhammad.rizki@uin-suska.ac.id

ABSTRACT

PT Pegadaian, as listed on its official website, is a fast-growing financial company. One of its key challenges is late installment payments, which can lead to financial losses. Using pawn customer data from 2013 to 2021, this study found that out of 534 customers, 68 were late in paying installments, and 10 did not pay. To address this issue, this research applies customer classification to identify borrowers who are more likely to pay on time. The classification model is developed using data mining with the C5.0 algorithm to generate decision-tree rules. Prior to modeling, the dataset is processed through the Knowledge Discovery in Databases (KDD) stages, including data selection, cleaning, and transformation. The proposed model produces 26 classification rules and achieves an accuracy of 87.04%. All data processing, modeling, and validation are conducted using RapidMiner Studio.

Keywords: *Classification, Decision Tree, C5.0 Algorithm, Data Mining*

Introduction

The times we live in, along with increasing human needs, lead people to do whatever it takes to meet them, especially during the Covid-19 pandemic, which resulted in a decline in the community's economy. Financial institutions utilize this by providing savings and loan services in the form of credit. These services are not only used by the lower middle class, but also utilized by the upper middle class. This service is usually used to meet human needs, such as buying daily necessities, or to serve as business capital. One of the financial institutions engaged in this service is Pegadaian[1], [2].

The main challenge for pawnshops is customers who are late in paying installments. The arrears rate increased due to the customer's negligence in paying installments, causing the company to be unable to provide loans to other customers due to the lockdown in the system to disburse customer credit. Before the payment is made, the company cannot provide a loan. Determining potential customers is very important to know which customers make installment payments on time. Determining potential customers can be done using classification [3], [4].

Data Mining is a technique for finding patterns hidden in large amounts of data (Big data). The pattern found will be used to gain knowledge (Knowledge) that may be used. However, the quality of the knowledge produced depends on the data processing techniques and the quality of the data used. Each type of data requires different processing techniques, depending on the data and the knowledge you want to find [5].

Classification is the process of assessing data objects to assign them in a class from the available classes[6]. Among the several classification methods available is the decision tree method [7]. The decision tree method can convert a large set of facts into a decision tree that represents rules. Rules can be easily understood in natural language[8], [9]. A decision tree is a predictive model that uses a tree structure to generate decision rules from data[10][11].

The C5.0 algorithm is an algorithm used to generate decision trees. This method is used to classify data with both categorical and numerical attributes. Calculation using the C5.0 Algorithm, there are several steps as follows:

1. Create a decision system that includes condition attributes and decision attributes. Then describe the decision system consisting only of n objects.
2. Calculates the number of columns of data, where the amount of data must be based on members of a particular attribute whose results are based on certain conditions.
3. Select the attribute used as a Node.
4. Create a branch for each node member.

5. Checks whether the entropy value of each Node member is zero. If the value is 0, determine which leaves have formed. If the entropy of each node member is zero, the process stops.

The formula is shown in the following equation. To calculate entropy, the equation used equation :

$$\textit{Entropy}\ (S) = \sum_{j=1}^{k} -\ pj\ x \log 2\ pj \tag{1}$$

Where;
S = set of cases
A = Attribute
k = number of classes in variable A
pj = proportion of Sj
While the calculation of the Gain value can be seen in the following formula:

$$\textit{Gain}\ (S, A) = \textit{Entropy}\ (S) - \sum_{i=1}^{m} \frac{|Si|}{|S|}\ x\ \textit{Entropy}\ (Si) \tag{2}$$

Where:
S = case set
A = attribute
m = number of partitions attribute A
|Si| = number of cases on the i-th partition
|S| = number of cases in S
After obtaining the Entropy and Gain values, the next step is to calculate the Gain Ratio value. The basic formula of the Gain Ratio calculation is as follows:

$$\textit{Gain Ratio} = \frac{\textit{Gain}\ (S, A)}{\sum_{i=1}^{m} \textit{Entropy}\ Si} \tag{3}$$

Where
Gain (S, A) = gain value of a variable
$\sum_{i=1}^{m}$ *Entropy* Si = number of Entropy values in a variable
The process is repeated for each branch until each branch has its own classes.

Research Methods

Before the raw data is processed, it will undergo the KDD (Knowledge Discovery in Databases) process: data selection, data preprocessing, and data transformation. Data selection aims to eliminate unnecessary data. Pre-cleaning involves cleaning selected data, such as removing inconsistent, incomplete, or duplicate entries. Data transformation is the process of converting cleaned, ready-to-process data into data that suits your needs in accordance with the data mining process and using data mining methods[12], [13], [14].

Split Validation is a validation technique that randomly divides data into two parts. The data is divided into training and test sets. Usually, the training-to-testing split is 90:10 (90% training, 10% testing) or 50:50 (50% training, 50% testing). The amount of training data and testing data can be calculated using the following equation:

1. Amount of Data Training

$$\text{Data training} = \text{Proportion} \times N \tag{4}$$

2. Amount of Data Testing

$$\text{Data testing} = N - \text{Data training} \tag{5}$$

Where:
N = represents the amount of data to be used as a sample.

Training data will be processed using the C5.0 algorithm using the formula previously presented. Manual data processing is not error-free. Therefore, it is necessary to test using software that is usually used to solve problems. Testing is carried out using Rapid Miner in process data. It aims to test whether the manual calculations are correct and in accordance with the C5.0 algorithm.

The data processed using the C5.0 Algorithm generates a decision tree. From the decision tree results, a rule will be derived and applied to the next dataset. The rules created are the result of the decision tree's narrative. The rules obtained classify potential customers[15], [16], [17]. The classification results for potential customers are useful for grouping them based on customer attributes.

Data Collection

The data collected in this study is secondary data, namely a dataset of customers who took out loans from 2013 to 2021, with a total of 534 data points. The dataset contains customer borrowing data with the following attributes: branch, customer name, total loan, and address. Phone, Age, Opening date, Gender, Marital status, Education, Occupation, Source of funds, Company Name, Religion, and Payment description. After passing the KDD (Knowledge Discovery in Database) process[18], [19], [20], customer data is as follows:

Table 1. Lending Customer Data 2013-2021

No	**Total Up**	**Age**	**Open credit**	**Gender**	**Mating Status**	**Education**	**Occupation as**	**Source of funds**	**Information**
1	D	Adult	2018	P	Married	Senior High School	Housewife	From salary	Good
2	B	Adult	2014	P	Married	Diploma	Housewife	Results of efforts	Good
3	B	Elderly	2014	L	Married	Senior High School	Other	From salary	Late
4	B	Adult	2014	L	Unmarried	Senior High School	Private employees	From salary	Good
5	B	Adult	2019	P	Unmarried	Senior High School	Student	Results of efforts	Late
6	B	Elderly	2019	P	Married	Senior High School	Housewife	From salary	Good
7	C	Adult	2016	P	Married	Senior High School	Other	From salary	Late
8	B	Adult	2014	P	Married	Senior High School	Housewife	From salary	Good
9	B	Adult	2015	L	Married	Senior High School	Other	Loan	Late
.									
534	C	Adolescent	2020	P	Unmarried	Senior High School	Student	From salary	Late

Results and Discussion

Split Validation

Before carrying out the classification process, the first step is to split the training and testing data. The data split is 90:10, with 90% for training and 10% for testing[21], [22]. The calculation to determine the amount of data entered into the training data uses a proportion of 90:10:

Amount of Data Training = 90% x 534

$= \frac{90}{100}$ x 534
= 480

The following is a calculation to determine the amount of data that goes into the testing data:

Amount of Data Testing = 534 - 480
= 54

Based on the calculation results above, the 90:10 training data set contained 480 data, and the remaining 54 were used for the test set. The division process is carried out using RapidMiner.

Algorithm C5.0

The data processed using the C5.0 algorithm is training data totaling 480 data. In building the classification tree, the first stage of the C5.0 algorithm is to determine the number of attribute values. Determination is performed using Microsoft Excel 2010.

Table 2. Manual Recapitulation

Node	Attribute	Value	Number of cases	Good	Late	Bad	Entropy	Gain	Gain Ratio
1	Total		480	410	61	9	0,6800		
2	Total Up	A	14	10	3	1	1.0949	0,0080	0,0041
		B	274	236	33	5	0.6586		
		C	137	116	18	3	0.7086		
		D	55	48	7	0	0.5499		
3	Age	Adolescent	28	18	9	1	1,1077	0,0184	0,0056
		Adult	259	222	34	3	0,6496		
		Elderly	178	158	16	4	0,5881		
		Seniors	15	12	2	1	0,9055		
4	Open credit	2013	2	2	0	0	0	0,0752	0,0151
		2014	193	174	16	3	0,5259		
		2015	43	34	9	0	0,7401		
		2016	36	34	2	0	0,3095		
		2017	31	28	2	1	0,5475		
		2018	32	31	1	0	0,2006		
		2019	43	23	15	5	1,3738		
		2020	41	35	6	0	0,6006		
		2021	59	49	10	0	0,6565		
5	sex	W	379	327	48	4	0,6305	0,007924	0,0054
		M	101	83	13	5	0,8280		
6	Marital status	Single	102	80	19	3	0,8761	0,009073	0,0049
		Widow	18	17	1	0	0,3095		
		Married	360	313	41	6	0,6308		
7	Education	Primary school	3	3	0	0	0	0,0127	0,0053
		Junior high school	5	4	1	0	0,7219		
		Senior high school	382	320	55	7	0,7223		
		Diploma	55	52	2	1	0,3554		
		Bachelor’s degree	35	31	3	1	0,6054		

8	Occupation	Housewife	255	224	31	0	0,5338	0,1029	0,0228
		Private employees	58	57	1	0	0,1256		
		BUMN / BUMD	10	10	0	0	0		
		Student	47	32	13	2	1,0842		
		GOVERN MENT OFFICER	12	11	1	0	0,4138		
		Profession al	27	24	3	0	0,5032		
		Merchant	8	3	0	5	0,9544		
		Other	63	49	12	2	0,8956		
9	Source of income	Salary	280	263	14	3	0,3710	0,1163	0,0547
		Business income	163	132	25	6	0,8366		
		Grant	4	4	0	0	0		
		Loan	33	11	22	0	0,9182		

Based on Table 2, the Fund Source variable has the highest gain ratio, so it is used as the root node (node 1). Then there are four branches of the root node: From Salary, Business Results, Grants, and Loans. Recapitulation of the entire decision tree using Rapid Miner software, with the results of the decision tree as follows:

Validation

The Decision Tree obtained will then be evaluated by comparing test data with predictions from the training data. This is done to assess the accuracy of the decision tree in predicting the data. Testing was carried out using RapidMiner. The accuracy of the measurement data obtained from the test data reached 87.04.

Classification potential customer

Based on the data processing that has been carried out, the final result is obtained in the form of 26 classifications of potential customers with an If ... Then as follows:

1 If the Source of income is a grant, then Good
2 If the Source of income is a loan, then Late
3 If the Source of income is from a salary, Occupation as BUMN/BUMD, Then Good
4 If the source of income is from salary, Occupation as Housewife, then Good
5 If the source of income is from a salary, Occupation as Private employees, then Good
6 If the source of income is from salary, Occupation as Other, then Good
7 If the source of income is from a salary, Occupation as Government Officer, then Good
8 If the source of income is from a salary, Occupation as a Professional, then Good
9 If the source of income is from a salary, Occupation as a student, Education Diploma, then Good
10 If the source of income is from a salary, Occupation as a student, Education Bachelor’s degree, then Good
11 If the source of income is from a salary, Occupation as a student, Education Senior High School, Age Adult, then Good
12 If the source of income is from salary, Occupation as Student, Education Senior High School, Age Adolescent, Total Loan Gol B, then Good
13 If the source of income is from a salary, Occupation as Student, Education Senior High School, Age Adolescent, Total Loan Gol C, then Late
14 If the Source of income is business income, Occupation as Private employees, Then Good
15 If Source of income, Business income, Occupation as Other, then Good
16 If the Source of income is business income, Occupation as a student, Then Good
17 If the Source of income is business income, Occupation as a professional, Then Good
18 If the source of income is Business income, Occupation as Merchant, sex is Men, then Good
19 If the source of income is Business income, Occupation as Merchant, sex is women, then Bad
20 If the source of income is Business income, Occupation as Housewife, single, then Good
21 If the source of income is business income, Occupation as a housewife, or widow, then good

22 If Source of income Business income, Occupation as Housewife, Married, Total Loan Gol A, Then Good
23 If Source of income Business income, Occupation as Housewife, Married, Total Loan Gol C, Then Good
24 If Source of income Business income, Occupation as Housewife, Married, Total Loan Gol D, Then Good
25 If Source of income Business income, Occupation as Housewife, Married, Total Loan Gol B, Then Good
26 If Source of income Business income, Occupation as Housewife, Married, Total Loan Gol B, Age Adult, Then Late

Conclusion

Data Mining is applied in the data processing process by passing through the KDD (Knowledge Discovery in Databases) process, namely data selection, data pre-cleaning, and data transformation. Then the data is processed using the C5.0 algorithm to obtain a decision tree. The decision tree will be converted into in set of rules, yielding 26 rules to classify potential customers.

References

[1] M. K.Jayaswal, “Learning Eoq Model With Trade-Credit Financing Policy For Imperfect Quality Items Under Cloudy Fuzzy Environment,” *Mathematics*, Vol. 10, No. 2, 2022, Doi: 10.3390/Math10020246.

[2] M.Kumari, “An Eoq Model For Deteriorating Items Analyzing Retailer’s Optimal Strategy Under Trade Credit And Return Policy With Nonlinear Demand And Resalable Returns,” *International Journal Of Optimization And Control: Theories And Applications*, Vol. 12, No. 1, Pp. 47–55, 2022, Doi: 10.11121/Ijocta.2022.1025.

[3] J.Gong, “Determination Of Key Components In Automobile Braking Systems Based On Abc Classification And Fmeca,” *Journal Of Traffic And Transportation Engineering (English Edition)*, Vol. 9, No. 1, Pp. 69–77, 2022, Doi: 10.1016/J.Jtte.2019.01.008.

[4] Y. C.Hsu, “Abc-Norm Regularization For Fine-Grained And Long-Tailed Image Classification,” *Ieee Transactions On Image Processing*, Vol. 32, Pp. 3885–3896, 2023, Doi: 10.1109/Tip.2023.3273455.

[5] G.Houge, “Stepwise Abc System For Classification Of Any Type Of Genetic Variant,” *European Journal Of Human Genetics*, Vol. 30, No. 2, Pp. 150–159, 2022, Doi: 10.1038/S41431-021-00903-Z.

[6] A.Lefteh, “Optimization Of Modified Adaptive Neuro-Fuzzy Inference System (Manfis) With Artificial Bee Colony (Abc) Algorithm For Classification Of Bone Cancer,” 2022. Doi: 10.1109/Dchpc55044.2022.9731840.

[7] H. K.Dreiner, “The Abc Of Rpv: Classification Of R-Parity Violating Signatures At The Lhc For Small Couplings,” *Journal Of High Energy Physics*, Vol. 2023, No. 7, 2023, Doi: 10.1007/Jhep07(2023)215.

[8] K. S.Dorman Andr.Maitra, “An Efficient *K*-Modes Algorithm For Clustering Categorical Datasets,” *Statistical Analysis And Data Mining*, Vol. 15, No. 1, Pp. 83–97, 2022, Doi: 10.1002/Sam.11546.

[9] K.Kim, “A Weighted *K*-Modes Clustering Using New Weighting Method Based On Within-Cluster And Between-Cluster Impurity Measures,” *Journal Of Intelligent & Fuzzy Systems*, Vol. 32, No. 1, Pp. 979–990, 2017, Doi: 10.3233/Jifs-16157.

[10] Y.Son, “Development Of Methodology For Classification Of User Experience (Ux) In Online Customer Review,” *Journal Of Retailing And Consumer Services*, Vol. 71, 2023, Doi: 10.1016/J.Jretconser.2022.103210.

[11] R.Krittayaphong, “Clinical Phenotype Classification To Predict Risk And Optimize The Management Of Patients With Atrial Fibrillation Using The Atrial Fibrillation Better Care (Abc) Pathway: A Report From The Cool-Af Registry,” *Qjm: An International Journal Of Medicine*, Vol. 117, No. 1, Pp. 16–23, 2024, Doi: 10.1093/Qjmed/Hcad219.

[12] P.Dobra, “Cumulative And Rolling Horizon Prediction Of Overall Equipment Effectiveness (Oee) With Machine Learning,” *Big Data And Cognitive Computing*, Vol. 7, No. 3, 2023, Doi: 10.3390/Bdcc7030138.

[13] A.Belhadi, “The Integrated Effect Of Big Data Analytics, Lean Six Sigma And Green Manufacturing On The Environmental Performance Of Manufacturing Companies: The Case Of North Africa,” *Journal Of Cleaner Production*, Vol. 252, 2020, Doi: 10.1016/J.Jclepro.2019.119903.

[14] H.Yao, “Identification Of Encrypted Traffic Through Attention Mechanism Based Long Short Term

Memory," *Ieee Transactions On Big Data*, Vol. 8, No. 1, Pp. 241–252, 2022, Doi: 10.1109/Tbdata.2019.2940675.

[15] F.Yuan, Y. L.Yang, Andt. T.Yuan, "A Dissimilarity Measure For Mixed Nominal And Ordinal Attribute Data In K-Modes Algorithm," *Applied Intelligence*, Vol. 50, No. 5, Pp. 1498–1509, 2020, Doi: 10.1007/S10489-019-01583-5.

[16] D. T.Dinh Andv. N.Huynh, "*K*-Pbc: An Improved Cluster Center Initialization For Categorical Data Clustering," *Applied Intelligence*, Vol. 50, No. 8, Pp. 2610–2632, 2020, Doi: 10.1007/S10489-020-01677-5.

[17] S.Bensalem, S.Naouali, Andz.Chtourou, "Rough Mode: A Generalized Centroid Proposal For Clustering Categorical Data Using The Rough Set Theory," In *Big Data And Smart Digital Environment*, Vol. 53, 2019, P. 225. Doi: 10.1007/978-3-030-12048-1_24.

[18] F. Y.Cao *Et Al.*, "An Algorithm For Clustering Categorical Data With Set-Valued Features," *Ieee Transactions On Neural Networks And Learning Systems*, Vol. 29, No. 10, Pp. 4593–4606, 2018, Doi: 10.1109/Tnnls.2017.2770167.

[19] I.Saha, J. P.Sarkar, Andu.Maulik, "Integrated Rough Fuzzy Clustering For Categorical Data Analysis," *Fuzzy Sets And Systems*, Vol. 361, Pp. 1–32, 2019, Doi: 10.1016/J.Fss.2018.02.007.

[20] F. O.Defrança, "A Hash-Based Co-Clustering Algorithm For Categorical Data," *Expert Systems With Applications*, Vol. 64, Pp. 24–35, 2016, Doi: 10.1016/J.Eswa.2016.07.024.

[21] L.Bai Andj. Y.Liang, "Cluster Validity Functions For Categorical Data: A Solution-Space Perspective," *Data Mining And Knowledge Discovery*, Vol. 29, No. 6, Pp. 1560–1597, 2015, Doi: 10.1007/S10618-014-0387-5.

[22] S.Bensalem, S.Naouali, Andz.Chtourou, "Rough Mode: A Generalized Centroid Proposal For Clustering Categorical Data Using The Rough Set Theory," In *Big Data And Smart Digital Environment*, Vol. 53, Y.Farhaoui Andl.Moussaid, Eds., Polytech Sch Tunisia, Bp 743,Rue El Khawarizmi, Tunis 2078, Tunisia Mil Acad Fondouk Jedid, Virtual Real & Informat Technol, Tunis, Tunisia Digital Res Ctr Sfax, Bp 275, Sfax 3021, Tunisia, 2019, Pp. 225–236. Doi: 10.1007/978-3-030-12048-1_24 10.1007/978-3-030-12048-1.