Comparative Analysis of Data Customer Classification with C4.5 Algorithm

Siti Aisyah, Bondang Johanes Rumapea, M. Ghifari Halwan, and Denny Hartanto Siahaan

***Abstract*—The problem that often occurs in insurance is the number of customers in arrears of paying premiums. Therefore, a system is needed to classify which prospective customers fall into the eligible group and which customers fall into the unfit group in filing as insurance customers to overcome the problem early. Self-protection is very important for both the safety of individual's life and their valuable assets in today's risky environment. The classification of prospective new insurance customers aims to facilitate the insurer in making decisions in terms of providing insurance coverage. The classification of prospective new insurance customers aims to avoid similar cases by only looking at the rules formed from the decision tree. The decision tree method using the C4.5 algorithm makes extracting information faster and more optimal with a larger data capacity. Therefore, errors caused in decision-making are greatly minimized.**

***Index Terms*— Customers, Predictions, Data**

I. INTRODUCTION

HUMANS cannot predict what will happen in the future perfectly even by using several analytical tools. The same is true for companies and individuals. Risks in the future can occur to a person's life, for example, death, illness, accident, or the risk of being fired from work, so people need insurance. In the business world, the dangers faced can be in the form of losses due to fire, damage, or loss. Therefore, every risk faced must be overcome so as not to cause even more significant losses. It can also be seen that insurance has a lot of demand among the public, but at this time, providing insurance to customers is still based on a non-objective process.

The process that is not objective here means a process whose data is not necessarily valid because one of the reasons is the surveyor who is in charge of surveying the data. There are still many mistakes in the criteria for prospective new insurance customers who will be given insurance coverage, which makes it difficult to determine the feasibility of providing insurance, which insurers often experience. The problem that often arises in insurance problems is the number of customers in arrears paying premiums. Therefore, a system is needed that can classify which prospective customers fall into the eligible group and which customers fall into the unfit group in health insurance.

Manuscript received October 9, 2020.

S. Aisyah, Faculty Technology and Computer Science, Universitas Prima Indonesia, Indonesia (email: siti_aisyah@unprimdn.ac.id).

Bondang Johanes Rumapea, Faculty Technology and Computer Science, Universitas Prima Indonesia, Indonesia (email: siti_aisyah@unprimdn.ac.id).

M. Ghifari Halwan, Faculty Technology and Computer Science, Universitas Prima Indonesia, Indonesia (email: siti_aisyah@unprimdn.ac.id).

Denny Hartanto Siahaan, Faculty Technology and Computer Science, Universitas Prima Indonesia, Indonesia (email: siti_aisyah@unprimdn.ac.id).

II. RESEARCH METHODS

A. Review Stage

This type of quantitative research is a method that is carried out based on a positive paradigm with the final result in the form of generalization. The data used is the data of prospective customers of PT Allianz Life Indonesia. The data obtained are discrete and numerical data based on each specified attribute.

B. Research Stage

The research was conducted based on the Research Work Activity Flowchart, which explains how a system that runs from the beginning to the end of the system is used. The following is a research activity diagram and a program diagram.

C. Research Variable

To determine the variables, there are two types of variables: the independent variable (X) and the dependent variable (Y). The independent variable (X) is a variable whose value does not depend on the value of other variables, while the dependent variable (Y) is a variable whose value is dependent on the value of other variables. So, the variables and attributes used in this data mining process are variables that are in accordance with the research. In this study, the independent variable (free) with four attributes, namely Name, Age, Income, Number of Dependents, and the dependent variable (bound), namely the label/output class in this study, namely Decisions based on data numbers of prospective customers with low and high labels.

This data processing uses the C4.5 algorithm method (calculations) or the modeling/series used, or the application of the algorithm to the classification of new prospective customers. The criteria used to determine the eligibility of new prospective customers are Age (X1), income (X2), and number of dependents (X3). The data used is data for new prospective customers in 2020 with 256 data records.

III. DISCUSSION

Data mining is a term used to describe knowledge discovery in databases. Data mining is a process that uses statistical,

mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and assemble knowledge from large databases [1-6].

TABLE I
DATASET

No	Name	Income	Age	Dependents	Decision
1	Dewi Pertiwi	Rp 3.500.000	30	4	Un Worthy
2	Lim Dan	Rp 4.000.000	25	2	Worthy
3	Viktor Sijabat	Rp 4.800.000	29	5	Un Worthy
4	Muhamma d Afif	Rp 5.500.000	27	0	Worthy
5	Citra Kirana	Rp 3.400.000	29	3	Worthy
6	Valecia Pasaribu	Rp 4.500.000	34	1	Worthy
7	Cindy Octavia	Rp 2.950.000	31	3	Worthy
8	Prasista Najwa	Rp 8.000.000	39	2	Worthy
9	Perdana Ismail	Rp 2.950.000	34	4	Un Worthy
10	Kurmin Jaya	Rp 4.450.000	45	1	Worthy
….	…….	……	…	….	…….
256	Perkasa Abadi	Rp 5.700.000	56	1	Un Worthy

The data in table 1 was taken from PT. Allianz Life Indonesia. Based on the data collected, a selection process and data transformation were carried out so that three criteria were used: age, income, and number of dependents. The sub-criteria of each criterion are as follows:

TABLE II
AGE

No	Age	Value
1	24-35	Good
2	36-49	Enough
3	50-70	Less

TABLE III
INCOME

No	Income	Value
1	> 6.000.000	Good
2	3.600.000-6.000.000	Enough
3	2.500.000-3.500.000	Less

TABLE IV
DEPENDENTS

No	Dependents	Value
1	0-1	Good
2	2-4	Enough
3	>5	Less

Based on the data above, the test data or testing aimed at table V: Data testing

TABLE V
DATA TESTING

No	Name	Income	Age	Dependents	Decision
1	Dewi Pertiwi	Rp 3.500.000	30	4	Un Worthy
2	Lim Dan	Rp 4.000.000	25	2	Worthy
3	Viktor Sijabat	Rp 4.800.000	29	5	Un Worthy
4	Muhamma d Afif	Rp 5.500.000	27	0	Worthy
5	Citra Kirana	Rp 3.400.000	29	3	Worthy
6	Valecia Pasaribu	Rp 4.500.000	34	1	Worthy
7	Cindy Octavia	Rp 2.950.000	31	3	Worthy
8	Prasista Najwa	Rp 8.000.000	39	2	Worthy
9	Perdana Ismail	Rp 2.950.000	34	4	Un Worthy
10	Kurmin Jaya	Rp 4.450.000	45	1	Worthy
….	…….	……	…	….	…….
256	Perkasa Abadi	Rp 5.700.000	56	1	Un Worthy

After determining the testing data, the next step is to determine the gain by calculating the gain and entropy. For the calculation of the C4.5 algorithm, it can be described as follows:

Step 1:

Counting the number of cases, the number of cases for the eligible category, and the number of cases for the unfeasible category, based on the attributes used to find entropy are as follows:

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

Description:

S : Case Set
A : Feature
n : Number of Partitions n
pi : The Proportion of Si to S

Entropy S :

$$\left(-\left(\frac{26}{64}\right)\times\log_2\left(\frac{26}{64}\right)\right)+\left(-\left(\frac{38}{64}\right)\times\log_2\left(\frac{38}{64}\right)\right)$$

$$= 0.527946 + 0.446543$$

$$= 0.974489$$

Step 2: Calculate the entropy of each attribute

a. Income entropy calculation
   Entropy (Good Income)

$$\left(-\left(\frac{7}{9}\right)\times\log_2\left(\frac{7}{9}\right)\right)+\left(-\left(\frac{2}{9}\right)\times\log_2\left(\frac{2}{9}\right)\right)$$
$$=0.281999+0.482206$$
$$=0.764205$$

Entropy (Sufficient Income)

$$\left(-\left(\frac{18}{32}\right)\times\log_2\left(\frac{18}{32}\right)\right)+\left(-\left(\frac{14}{32}\right)\times\log_2\left(\frac{14}{32}\right)\right)$$
$$=0.466917+0.521782$$
$$=0.988699$$

Entropy (Income Less)

$$\left(-\left(\frac{13}{23}\right)\times\log_2\left(\frac{13}{23}\right)\right)+\left(-\left(\frac{10}{23}\right)\times\log_2\left(\frac{10}{23}\right)\right)$$
$$=0.465243+0.52245$$
$$=0.987693$$

b. Age entropy calculation

Entropy (Good Age)

$$\left(-\left(\frac{28}{36}\right)\times\log_2\left(\frac{28}{36}\right)\right)+\left(-\left(\frac{8}{36}\right)\times\log_2\left(\frac{8}{36}\right)\right)$$
$$=0.281999+0.482206$$
$$=0.764205$$

Entropy (Sufficient Age)

$$\left(-\left(\frac{8}{15}\right)\times\log_2\left(\frac{8}{15}\right)\right)+\left(-\left(\frac{7}{15}\right)\times\log_2\left(\frac{7}{15}\right)\right)$$
$$=0.483675+0.513117$$
$$=0.996792$$

Entropy (Low Age)

$$\left(-\left(\frac{2}{13}\right)\times\log_2\left(\frac{2}{13}\right)\right)+\left(-\left(\frac{11}{13}\right)\times\log_2\left(\frac{11}{13}\right)\right)$$
$$=0.415452+0.20393$$
$$=0.619382$$

c. Dependent entropy calculation

Entropy (Good Dependent)

$$\left(-\left(\frac{23}{24}\right)\times\log_2\left(\frac{23}{24}\right)\right)+\left(-\left(\frac{1}{24}\right)\times\log_2\left(\frac{1}{24}\right)\right)$$
$$=0.0588422+0.19104$$
$$=0.2498822$$

Entropy (Enough Dependents)

$$\left(-\left(\frac{13}{24}\right)\times\log_2\left(\frac{13}{24}\right)\right)+\left(-\left(\frac{11}{24}\right)\times\log_2\left(\frac{11}{24}\right)\right)$$
$$=0.479117+0.515868$$
$$=0.994985$$

Entropy Loss Dependents)

$$\left(-\left(\frac{2}{16}\right)\times\log_2\left(\frac{2}{16}\right)\right)+\left(-\left(\frac{14}{16}\right)\times\log_2\left(\frac{14}{16}\right)\right)$$
$$=0.375+0.168564$$
$$=0.543564$$

Step 3: Calculate the gain of each attribute

a. Income entropy calculation

$$\left(\begin{array}{l}(0.974489)-\left(\left(\frac{9}{64}\right)\times(0.764205)\right)+\left(\left(\frac{32}{64}\right)\times 0.988699\right)+\\ \left(\left(\frac{23}{64}\right)\times 0.987693\right)\end{array}\right)$$
$$=0.974489-0.107466+0.494349+0.354952$$
$$=1.716324$$

b. Age gain calculation

$$\left(\begin{array}{l}(0.974489)-\left(\left(\frac{36}{64}\right)\times(0.764205)\right)+\left(\left(\frac{15}{64}\right)\times 0.9996792\right)+\\ \left(\left(\frac{13}{64}\right)\times 0.619382\right)\end{array}\right)$$
$$=0.974489-0.429865+0.233623+0.125812$$
$$=0.904059$$

c. Dependent gain calculation

$$\left(\begin{array}{l}(0.974489)-\left(\left(\frac{9}{64}\right)\times(0.764205)\right)+\left(\left(\frac{32}{64}\right)\times 0.988699\right)+\\ \left(\left(\frac{23}{64}\right)\times 0.987693\right)\end{array}\right)$$
$$=0.974489-0.107466+0.494349+0.354952$$
$$=1.716324$$

	Attribute	Amount	Worthy	Unworthy	Entropy	Gain
TOTAL		24	23	1	0,37177312	
Income	Good	4	4	0	0	0,088961351
	Enough	15	14	1	0,35335934	
	Less	5	5	0	0	
Age	Good	14	14	0	0	0,040118488
	Enough	7	7	0	0	
	Less	3	2	1	0,91829583	

TABLE VI
CALCULATION OF NODE 1

Attribute		Amount	Worthy	Unworthy	Entropy	Gain
Total		64	26	38	0,974489	
Income	Good	9	7	2	0,764205	0,017721403
	Enoug	32	18	14	0,98869	

	h				9	
	Less	23	13	10	0,98769 3	
	Good	36	28	8	0,76420 5	
Age	Enoug h	15	8	7	0,99679 2	0,1851889 97
	Less	13	2	11	0,61938 2	
	Good	24	23	1	0,24988 22	
Depende nts	Enoug h	24	13	11	0,99498 5	0,3717732 03
	Less	16	2	14	0,54356 4	

TABLE VII
NODE CALCULATION 1.1
TABLE VIII
NODE CALCULATION 1.2

	Attribute	Amount	Worthy	Un Worthy	Entropy	Gain
TOTAL		4	4	0	0	
	Good	0	0	0	0	0
Age	Enough	2	2	0	0	
	Less	2	2	0	0	

The following is the calculation using Rapid Miner.

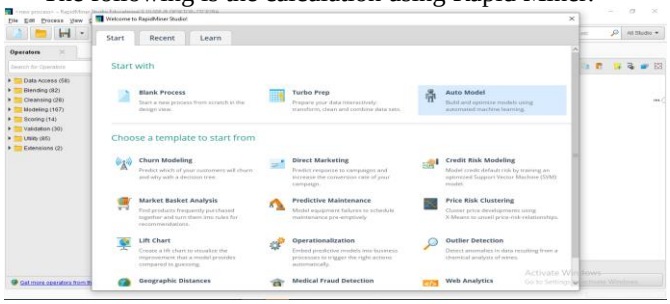

Fig. 1. Rapid Miner 9.0 Utama Main Page.

After the application is open, select Blank, select the add data menu, and look for the "testing.xlxx" data located, select it and click next. The Select the Cells to Import Data dialog box will appear, select it and click next

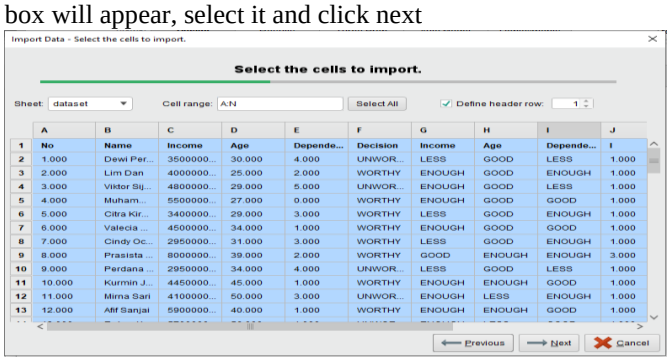

Fig. 2. Select the Cells to Import Data Halaman page

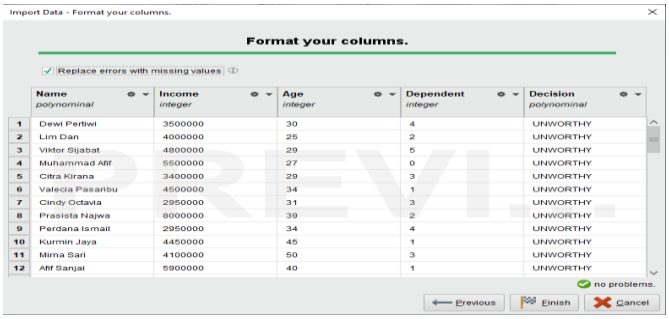

Fig. 3. Select the Cells to Import Data Halaman page

After that, in the format of your columns in the sold attribute, click the down arrow, select change role, select label, click ok, click next, and click finish.

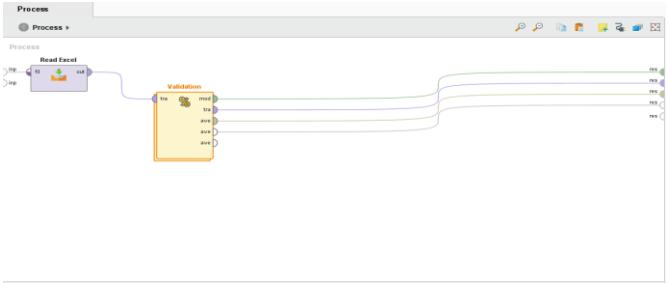

Fig. 4. Process Data Training

Drag the "testing.xlsx" data on the repository menu that we have entered into the rapid miner, the operator used is split validation, the operator is useful as a measure of the accuracy of a running model.

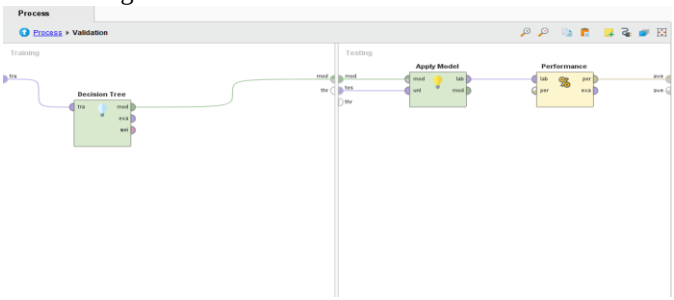

Fig. 5. Process Validation

Then the process found in the split t validation operator requires an operator that can generate test objectives. Because this study uses the C4.5 algorithm, the operator used is a decision tree operator and adds a model and performance to determine the level of accuracy by selecting and clicking run.

accuracy: 94.74%

	true UN WORTHY	true WORTHY	class precision
pred. UN WORTHY	7	0	100.00%
pred. WORTHY	1	11	91.67%
class recall	87.50%	100.00%	

Fig. 6. Table of Testing Data Accuracy Results

Class Precision is obtained with an invalid prediction of 100.00% and a proper prediction of 91.67%. It can be concluded that the level of true relevance is not feasible at 87.50%, and true relevance is feasible at 100.00%.

recall: 100.00% (positive class: WORTHY)

	true UN WORTHY	true WORTHY	class precision
pred. UN WORTHY	7	0	100.00%
pred. WORTHY	1	11	91.67%
class recall	87.50%	100.00%	

Fig. 7. Recall Data Testing Table

Class Precision is obtained with an invalid prediction result of 100.00% and a feasible prediction of 91.67%. It can be concluded that the true level is not feasible at 87.50%, and true is feasible at 100.00%.

precision: 91.67% (positive class: WORTHY)

	true UN WORTHY	true WORTHY	class precision
pred. UN WORTHY	7	0	100.00%
pred. WORTHY	1	11	91.67%
class recall	87.50%	100.00%	

Fig. 8. Precision Data Testing Table

Class Precision is obtained with an invalid prediction of 100.00% and a proper prediction of 91.67%. It can be concluded that the True level is not feasible at 87.50%, and True is feasible at 100.00%.

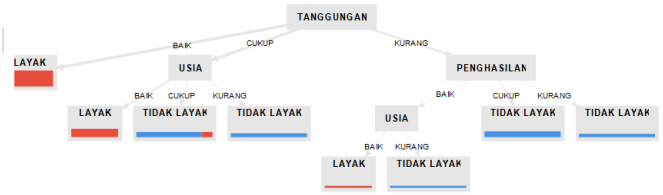

Fig. 9. Tree View RapidMiner

So the C4.5 algorithm is very effective in classifying the data of prospective health insurance customers.

IV. Conclusion and Recommendations

A. Conclusions

Based on data mining research, the classification of prospective insurance customers using the C4.5 algorithm is tested for accuracy whether the prospective customer becomes a customer or not. The test results obtained show that the precision, recall, and accuracy values are 91.67%, 100.00%, and 94.74%, respectively.

B. Recommendations

There needs to be further research by testing with other methods and comparisons such as Naïve Bayes, Neural Networks, and so on to obtain comparisons with the highest level of accuracy. In using the C4.5 algorithm to classify, it is necessary to select the right variables so that the decision tree results are more accurate or detailed.

References

[1] J. Han, et al. (2012). "Data Mining: Concepts anf Techniques Fransisco: Morgan Kaufmann Publishers.

[2] S. A. Pattekari, A. Parveen. "Prediction System for Heart Disease Using Naive Bayes," *Internasional Journal of Advanced Computer and Mathematical Sciences*, ISSN 2230-9624, Vol. 3, No 3, Hal 290-294, 2012.

[3] M. Karim, R. M. Rahman. "Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing," *J. Softw Eng Appl*, 2013; 06: 196–206.

[4] A. Naik and L. Samant. "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime," *Procedia Comput. Sci.* 85 662–628, 2016.

[5] Y. Wan and Q. Gao, "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis," *2015 IEEE International Conference on Data Mining Workshop (ICDMW) (IEEE)* 1318–1325, 2015.

[6] D. J. Hand, "Data Mining Based in part on the article "Data mining" by David Hand, which appeared in the Encyclopedia of Environmetrics, Encyclopedia of Environmetrics," (Chichester, UK: John Wiley & Sons, Ltd), 2013.

[7] D. Kurniadi, E. Abdurachman, H. L. H. S. Warnars and W. Suparta. "The prediction of scholarship recipients in higher education using k-Nearest neighbor algorithm," *IOP Conf. Ser. Mater. Sci. Eng*. 434 012039, 2018.

[8] Amin, F. A. O. S. Adnan, A. A. J. L. Babar. "Customer churn prediction in telecommunication industry using data," *Journal of Business Research*, 2018.

[9] D. Al-Nabi, L. D. and S. S. Ahmed, "Survey on Classification Algorithms for Data Mining: (Comparison and Evaluation)," *Computer Engineering and Intelligent Systems*, pp. 18-25, 2013

[10] Mobasher, Bamshad. 2005. Clasification via Decision Trees in WEKA. http://maya.cs.depaul.edu/classes/ect584/weka/classify.html.

Siti Aisyah received S.Kom degree in Information System from Universitas Prima Indonesia (UNPRI) Medan, Indonesia, in 2013 and the M.Kom degree in Information System from Universitas Putra Indonesia "YPTK" Padang, Indonesia, in 2015.Know, she worked at Universitas Prima Indonesia (UNPRI) Medan, Indonesia as a lecturer. The research area is Information System and AI.

Bondang, et all is student in Information System form Universitas Prima Indonesia (UNPRI) Medan, Indonesia.