Performance Analysis Of C4.5 And Naïve Bayes Algorithm On Customer Relationship Management (CRM) In Jatimas Furniture Ltd.

Beni Apriliyawan[1], Sri Lestari[2]
[1,2] Faculty of Computer Science, Informatic and Business Institute Darmajaya,
Jl. Z.A. Pagar Alam, Bandar Lampung 35142, Indonesia
[1]beni.apriliyawan.1721210002@mail.darmajaya.ac.id
[2]srilestari@darmajaya.ac.id

Abstract: Customer relationship management (CRM) was a part of the marketing strategies to maintain customer satisfaction. Customer satisfaction was an important factor in a business to stay afloat or continued to grow so that business managers needed an evaluation from the customers to find out whether customers were satisfied or not satisfied with the services provided by the company under customer expectations. The objective of this study was to measure the customer satisfaction in the Jatimas Furniture Ltd. using C4.5 and Naïve Bayes Algorithm. The variables of this study were service, price, promotion, product quality, and facilities. This study expected to assist the company in increasing the level of customer satisfaction to defend the company from its competitors. After that, the performance analysis of the C4.5 and Naïve Bayes Algorithm done to find out the algorithm that had higher accuracy for evaluating customer satisfaction survey results. The tests carried out through the Split Validation technique. First test used six variables with the Naïve Bayes Algorithm by generating 81.48% accuracy; AUC (Area under Curve) was 0.871 through 90% training data and 10% testing data; while test 2 used four variables of decision trees C4.5 Algorithm by generating 76.19 % accuracy and AUC (Area under Curve) was 0.500 through 70% training data and 30% testing data. The conclusion of this study was that the accuracy and AUC (Area under Curve) of the Naïve Bayes Algorithm were higher than the decision tree C4.5 Algorithm.

Keywords: Customer Relationship Management (CRM), Decision tree C.45 Algorithm, Naïve Bayes Algorithm

1. INTRODUCTION

The increasing competition in the business world today required entrepreneurs to be quick and responsive in making decisions so that the established companies were able to survive. Business doers needed to innovate and to be able to provide customer satisfaction. One of the steps to implement the customer satisfaction was by providing maximum satisfaction to customers because the goal of a business was to create a sense of satisfaction in customers. Customer relationship management (CRM) was the strategy to optimize profitability through good relationships with customers and creating customer satisfaction [1]. The problem on the customer satisfaction was no customer satisfaction indicator on furniture sales shop visit based on the number of visits or sales turnover. For this reason, the company must have a strong marketing strategy in its marketing services so that it was able to survive in business competition. One of the actions that must been taken to achieve competitive advantage was by focusing on improving the service area to meet customer satisfaction so that the approach was needed for generating information so that it was able to meet customer satisfaction in the Jatimas Furniture Ltd. One of the ways to receive information or patterns from data sets was through data mining. The algorithm used in this study was the C4.5 and Naïve Bayes Algorithm. This study measured the customer satisfaction used to provide information on the factors that needed to improve. Besides, it was necessary to evaluate the customer side to find out whether the customer was satisfied or

dissatisfied with the services provided by the company by customer expectations and to bring information to the researchers who were interested in Performance Analysis of C4.5 and Naïve Bayes Algorithm on Customer Relationship Management (CRM) in Jatimas Furniture Ltd.

2. LITERATURE REVIEW

C4.5 and Naïve Bayes Algorithm used in measuring the result of the customer satisfaction with the association rules. The data was able to been processed as information and was useful for companies. Several previous studies used as a reference for the author to conduct the study related to customer satisfaction. The study on [2] comparing the accuracy of the K-Nearest Neighbour method and the C4.5 Algorithm for Customer Relationship Management in financing companies in the classification of loyalty customer results from the comparison showed that the C4.5 accuracy was 93.10% and the KKN Algorithm accuracy value was 90.52% with an accuracy difference of 2.58%, using 122 sample data testing. Furthermore, the study of [3] the Determination of Extra-Curricular Students of Integrated High School Students in Lampang Subang with the C4.5 Algorithm used 20 data from evaluation and validation results with confusion matrix by dividing 70% training data and 30% testing data indicating that the level of accuracy in the algorithm C4.5 was 81.94%. Moreover, the study conducted through the Split Validation technique by dividing the training data and testing data on the data set by comparing the C4.5 and Naïve Bayes Algorithms to determine the level of data accuracy using the two algorithms to measure the results of customer satisfaction on the CV. Jatimas Furniture.

2.1 Data Mining

Data mining was the process of finding meaningful relationships, patterns, and trends by examining large amounts of data stored in storage using pattern recognition techniques such as statistical and mathematical techniques. Data mining was a process that employed one or more computer learning techniques (machine learning) to analyse and extract knowledge automatically [4]

2.2 Customer Relationship Management (CRM)

CRM was the business strategy for selecting and managing customers to optimize long-term value. Customer relationship management required a customer-centred business philosophy and culture to support effective marketing, sales, and service processes [5].

2.3 Algorithm C 4.5

C 4.5 Algorithm was the algorithm used to form a decision tree. The decision tree of large data sets became smaller record sets by applying a series of decision rules [6]
In general, to build the C4.5 algorithm is as follows:

a. The first thing to do was preparing a dataset used as a test material.
b. Afterwards, the total entropy value and information gain were obtained from the sum of all data, using an equation like this:
   Entropy (S) = $\sum_i^n 0 - pi * log2(pi)$ ……..…….. (1)
   Calculation of the entropy value of the attributes in the dataset done through the equation (1)
   Gain (S, A) = *Entropy* (S) $-\sum_i^n 1 - \frac{si}{s} * Entropy\ (S)$ ……..……..(2)
   Calculation of the gain value of the attributes in the dataset done through the equation (2)

c. Furthermore, the decision tree designed by looking at the highest *gain* value of all existing attributes, after which the highest gain ratio value of the attribute, was used as the root in the decision tree.
d. Then, a branch made from each of the remaining attribute values according to the highest gain ratio variable value through the equations (1) and (2).
e. In the last step, the process for each branch repeated until all cases in the branch had the same class and a process re-carried out.

1.4 Decision tree

A decision tree was the decision tree structure that could been used to divide large data sets into smaller record sets by applying a set of decision rules. With each set of dividers, the members of the result set were like each other. The data in the decision tree was usually expressed in a table with attributes and records [7].

2.5 Customer Satisfaction

The customer satisfaction was the customer response to the comparison between performance and expectations according to the evaluation of the non-conformity after the customer used the product. The customers felt this satisfaction on condition that they had consumed a product or service. If the customers liked the products they consumed, the customers were already satisfied. On the contrary, if the products they consumed did not match their needs and bought the other products, the customer did not feel satisfied [8]

2. METHOD

The data used in this study were the data based on the criteria used to measure the customer satisfaction test of Jatimas Furniture Ltd. with the C4.5 Algorithm method and to find out the patterns and methods of classification in data mining. They examined through the rapid miner tool to compare the performance of the Naïve Bayes Algorithm and decision tree C4.5 Algorithm. The steps seen in Figure 3.1

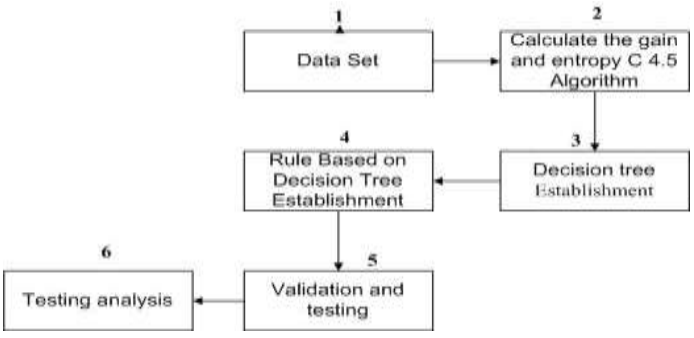

Figure 3.1 Flow diagram

1. It was the data set from the results of questionnaire conducted in Jatimas Furniture Ltd.
2. The value of entropy and information gain obtained from the sum of all data.
3. The decision tree designed from the calculation C4.5algorithm to observe the highest gain of all the variables used as the root in decision tree.
4. The rule established based on decision tree establishment of C4.5Algorithm.
5. The stages of validation and testing was determined through tools rapid miner with split validation technique.

6. The last stage of testing analysis was through the accuracy comparison and AUC (Area under Curve) performance C4.5 and Naive Bayes algorithm on Customer Relationship Management (CRM) in Jatimas Furniture Ltd.

3. RESULT AND DISCUSSION

The data used in this study sourced from questionnaire data. The google form was share throughhttps://docs.google.com/forms/d/e/1FAIpQLScbFwiKDqjw1ZJX_-Kadxde1Usf5y6viFgeZzFzarc69kxImA/viewform. Moreover, there were 6 variables and 30 data used for testing as it was seen on Table 4.1

Table 4.1 Questionnaire result data variables.

No	Price	Service	Promotion	Product Quality	Facilities	Results
1	Good	Good	Good	Very Good	Good	Satisfied
2	Very Good	Very Good	Good	Good	Bad	Satisfied
3	Good	Very Bad	Bad	Very Bad	Very Bad	Dissatisfied
4	Bad	Very Good	Bad	Good	Very Good	Satisfied
5	Good	Good	Good	Bad	Good	Satisfied
6	Good	Bad	Bad	Very Bad	Very Bad	Dissatisfied
7	Very Good	Good	Very Good	Very Good	Very Good	Satisfied
8	Good	Very Good	Good	Bad	Very Good	Satisfied
9	Very Good	Good	Good	Very Good	Good	Satisfied
10	Very Good	Good	Very Good	Very Good	Good	Satisfied
11	Very Bad	Bad	Good	Very Good	Bad	Dissatisfied
12	Good	Good	Very Good	Bad	Good	Satisfied
13	Very Bad	Bad	Good	Good	Good	Dissatisfied
14	Good	Good	Bad	Good	Good	Satisfied
15	Bad	Good	Bad	Very Bad	Very Bad	Dissatisfied
16	Very Good	Very Good	Good	Good	Bad	Satisfied
17	Very Good	Bad	Good	Very Good	Good	Satisfied
18	Good	Good	Very Good	Very Good	Very Good	Satisfied
19	Good	Good	Good	Very Good	Good	Satisfied
20	Good	Bad	Good	Good	Good	Satisfied
21	Bad	Good	Very Bad	Very Bad	Bad	Dissatisfied
22	Good	Good	Good	Good	Very Good	Satisfied
23	Good	Bad	Good	Good	Good	Satisfied
24	Good	Bad	Good	Good	Good	Satisfied
25	Bad	Bad	Good	Very Good	Bad	Satisfied
26	Very Good	Bad	Good	Good	Good	Satisfied
27	Bad	Good	Good	Bad	Bad	Dissatisfied
28	Very Bad	Good	Very Bad	Bad	Very Bad	Dissatisfied
29	Good	Good	Bad	Good	Good	Satisfied
30	Bad	Good	Very Bad	Good	Good	Satisfied

4.1 Data Classification C 4.5 Algorithm Process

The classification stage used in this study was entropy and information gains. The first step was by counting the number of cases for each of obtained case variables. Afterwards, the number of cases classified for satisfied and dissatisfied decisions. The entropy value calculated after the entropy value obtained. The gain value of each variable calculated so that the variable with the highest gain value taken as a root. It repeated until there were no more nodes. Each of calculation of the C4.5 Algorithm seen on the Table 4.2.

Table 4.2 Calculation C.45Algorithm.

Variables		Many Cases	Satisfied	Dissatisfied	ENTROPY	GAIN
Total Cases		30	22	8	0,836641	
PRICE						1,270128
	VERY GOOD	7	7	0	0	
	GOOD	14	12	2	0,500329	
	BAD	6	3	3	1	
	VERY BAD	3	0	3	0	
PROMOTION						1,295163
	VERY GOOD	5	5	0	0	
	GOOD	18	14	4	0,764205	
	BAD	4	2	2	0	
	VERY BAD	3	1	2	0	
SERVICE						1,544811
	VERY GOOD	4	4	0	0	
	GOOD	16	12	4	0,811278	
	BAD	9	6	3	0,918296	
	VERY BAD	1	0	1	0	
PRODUCT QUALITY						1,013015
	VERY GOOD	9	8	1	0,503258	
	GOOD	12	11	1	0,413817	
	BAD	5	3	2	0,970951	
	VERY BAD	4	0	4	0	
FACILITIES						1,428066
	VERY GOOD	5	5	0	0	
	GOOD	15	13	2	0,56651	
	BAD	6	3	3	1	
	VERY BAD	4	1	3	0,811278	

4.2 IF-THEN Rule Based on Decision Tree Establishment

The result of the process of this decision tree was that 12 rules already formed. Seven rules successfully classified with satisfied values and the remaining rules after classified were five rules with dissatisfied values. The results of the calculation process of the C4.5 Algorithm forming a decision tree shown on Figure 4.1.

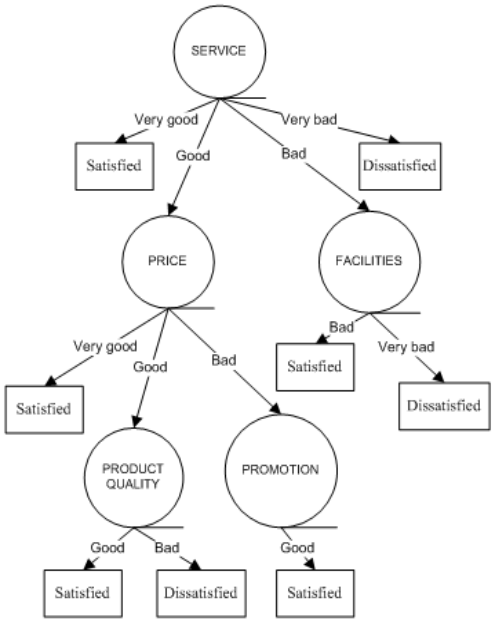

Figure 4.1 Decision Tree Establishment

1. IF Service= Good AND Price = Good AND Product Quality = Good THEN Value_ Inf= Satisfied
2. IF Service= Good AND Price =Good AND Product Quality = Bad THEN Value_ Inf =Dissatisfied
3. IF Service = Good AND Price = Bad AND Promotion = Good THEN Value_ Inf =Satisfied
4. IF Service = Bad AND Facilities = Bad THEN Value_ Inf =Satisfied
5. IF Service = Bad AND Facilities = Very Bad THEN Value_ Inf =Dissatisfied
6. IF Service = Very Good THEN Value_ Inf = Satisfied
7. IF Service = Very Bad THEN Value_ Inf = Dissatisfied
8. IF Price = Very good THEN Value_ Inf = Satisfied
9. IF Price = Very Bad THEN Value_ Inf = Dissatisfied
10. IF Product Quality = Good THEN Value_ Inf = Satisfied
11. IF Product Quality = Bad THEN Value_ Inf = Dissatisfied
12. IF Promotion = Good THEN Value_ Inf = Satisfied

4.3 C4.5 Algorithm and Naïve Bayes Testing Model

This examination was carried out to determine the performance of the C4.5 and Naïve Bayes Algorithm in classifying, the researcher used the split validation technique through the rapid miner tool to determine whether the effect of the variables emerge and was used to measure customer satisfaction results. Moreover, the validity test carried out. The data set divided into 2 parts e.g., the training data and testing data. The total of the data sets were 30 data sets with 6 data variables e.g., service, product quality, promotion, price, facilities, and decision results. To find out the value of accuracy and AUC (Area under Curve), six variables were determined to find out whether they were appropriate or not. A comparison made through four variables e.g., service, product quality, facilities, and decision results through which the split validation with different comparisons between training data and testing data carried out. The following was a description of the 3-time tests.

1. 70% of training data and 30% of testing data (21 and 9)
2. 80% of training data and 20% of testing data (24 and 6)
3. 90% of training data and 10% of testing data (27 and 3)

The testing process at this stage was a test process based on the data trained and the testing data seen on Figure 4.2.

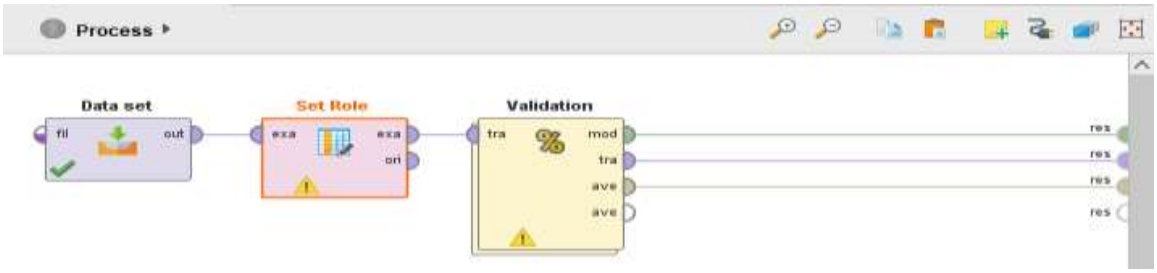

Figure 4.2. Classification Design Training and Testing

4.4 Test Results Discussion

This test was carried out through 6 data variables and 4 data variables used to determine the classification process of C4.5 Algorithm into a decision tree and to find out the accuracy so that it was necessary to conduct the examination through the split validation technique. The rapid miner

tool used as a tool in the data testing process with different sample ratios. The following was a description of the tests carried out by six variables and four variables as they seen on the Table 4.3.

Table 4.3 Results of testing six data variables and testing four data variables

Testing 1.6 variables data	Ratio Data Training (%)					
	70		80		90	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Algorithm C 4.5 Decision tree	76.19%	0.706	79.17%	0.764	74.07%	0.500
Algorithm Naïve bayes	76.19%	0.650	79.17%	0.880	81.48%	0.871
Testing 2 .4 variables data	Ratio Data Training (%)					
	70		80		90	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
C 4.5 decision tree Algorithms	76.19%	0.500	75.00%	0.500	74.07%	0.500
Naïve bayes Algorithms	66.67%	0.600	70.83%	0.718	66.67%	0.600

It seen that the result of testing one of C4.5 Algorithm decision tree by involving 90% of the training data and 10% of testing data was 81.48% of accuracy and AUC was 0.871. The result of testing two of the C4.5 Algorithm decision tree by involving 70% of training data and 30 of testing data was 76.19% of accuracy and AUC (Area under Curve) 0.500. This test was carried out through 6 data variables and 4 data variables used to determine the classification process of C4.5 Algorithm into a decision tree whether or not it was appropriate or not through 6 data variables so it was necessary to test with the split validation technique with different sample ratios. The performance analysis of the C4.5 Algorithm and Naïve Bayes showed the six data variables had a higher accuracy and AUC (Area under Curve) than four data variables so that the performance analysis of C4.5 and Naïve Bayes Algorithms was able to strengthen the use of six data variables previously used for the classification process of C4.5 Algorithms.

4. CONCLUSIONS

From the description that has stated, several conclusions can been drawn as follows.

a. Product quality, service, promotion, price, and facilities are important factors to increase customer satisfaction, especially in furniture sales companies, based on the results of calculations made using C4.5Algorithm is formed decision tree rule that can be used to measure customer satisfaction results. So that it can, been use by Jatimas Furniture Ltd.
b. Performance analysis of C4.5 and Naïve Bayes Algorithm on Customer Relationship Management (CRM) inthe CV. Jatimas Furniture, using a split validation technique with different sample ratios shows that test 1 uses 90% training data and 10% testing data. The Naïve Bayes Algorithm is superior with an accuracy difference of 2.31% and AUC (Area Under Curve) 0.107 compared to the decision treeC4.5 Algorithm, while test 2 using 70% training data and 30% testing data. The decision tree C4.5 Algorithm is superior with a difference of 4.14% accuracy and 0.282 AUC (Area under Curve) compared to the Naïve Bayes Algorithm.
c. Performance analysis of C.45 decision tree Algorithm and Naïve Bayes to measure the results of customer satisfaction inthe CV. Jatimas Furniture, testing with different sample ratios, obtained the accuracy value and AUC (Area under Curve) shows that the Naïve Bayes Algorithm is superior to the decision tree C.45 Algorithm.

Based on the conclusions above, the researcher proposes that some recommendations for this study can be develop with other classification data mining methods to make comparisons such as the genetic Algorithm and the K-Nearest Neighbor Algorithm to find the rules or model approach to been achieved.

REFERENCES

[1] Sirait, "PENGARUH CUSTOMER RELATIONSHIP MANAGEMENTDAN KUALITAS PELAYANAN TERHADAP KEPUASAN PELANGGAN (The Effect of Relationship Management and Service Quality on Customer Satisfaction)," vol. 3, no. 1, pp. 79–85, 2018.

[2] Iskandar, "KOMPARASI AKURASI METODE K-NEAREST NEIGHBOR DAN C4.5 UNTUK CUSTOMER RELATIONSHIP MANAGEMENT PADA PERUSAHAAN PEMBIAYAAN," vol. 1, pp. 1–14, 2020.

[3] Gupitha, "SISWA SMK TERPADU LAMPANG SUBANG Pendahuluan Kajian Teori Penelitian Terkait Ekstrakurikuler," vol. IV, no. 2, pp. 31–37, 2018.

[4] Widayu, "Data Mining Untuk Memprediksi Jenis Transaksi Nasabah Pada Koperasi Simpan Pinjam Dengan Algoritma C4.5," *Issn 2548-8368*, vol. Vol 1, No, no. June, p. 7, 2017.

[5] Imasari, "Pengaruh Customer Relationship Management Terhadap Loyalitas Pelanggan Pada PT BCA Tbk," *Fokus Ekon.*, vol. 10, no. 3, pp. 183–192, 2011.

[6] Utari, "Implementasi Metode C4 . 5 Untuk Menentukan Guru Terbaik Pada Smk 1 Percut Sei Tuan Medan," no. April, pp. 82–86, 2015.

[7] Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *J. Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2017.

[8] Endrawati, "Hubungan Kualitas Layanan dengan Kepuasan Pelanggan PT. Indoguna Utama Semarang," 2018.