Advances in Economics, Business and Management Research, volume 175
Proceedings of the 2nd International Conference on Business and Management of Technology (ICONBMT 2020)

Implementation of C4.5 Algorithm for Analysis of Service Quality in Companies of PT. XYZ

Supangat [1], Aji Pratama[2], Titasari Rahmawati[3]

*[1*2]Program Studi Teknik Informatika, Universitas 17 Agustus 1945, Surabaya, Jawa Timur*
[3]Program Studi Sistem Informasi, Institut Informatika Indonesia, Surabaya, Jawa Timur

Email: [1] supangat@untag-sby.ac.id*
[2] ajipp69@gmail.com,
[3]tita@ikado.ac.id

Abstract. Currently the world's automotive companies are increasing and the development of these companies is very fast, various car companies compete in releasing their products. Resulting in competition in the automotive business world. One of the automotive companies examined in this study was PT. XYZ In the company PT. The XYZ strategy used to attract customers is to consider the quality of their services. If the quality of PT. XYZ is good, so customers will be interested in buying Suzuki products and are loyal to the services provided by Suzuki. Because the customer is an assessor of service quality, in this study researchers used the C4.5 algorithm to measure the quality of services at PT. XYZ The method of collecting datasets is to use surveys. Then the data will be processed using the C.45 algorithm, the company can find out what attributes can determine the quality of good or bad service. The dataset is 500 records, in which researchers use a percentage of 60% for training data and 40% for testing data. From the trial results there is obtained an accuracy of 90%, Precision of 94%, Recall of 95%, and F-Measure of 94%. From the results of the study using the C4.5 algorithm in the existing dataset obtained 10 rules consisting of 3 rules with good service quality conclusions and 7 rules with poor service quality conclusions. This rule will be used to analyze the company's service quality to customers, which are the attributes that influence and the most dominant determines the quality of service.

Keywords: *automotive companies, customer, quality of service, C4.5 algorithm, dataset, rule, accuracy*.

1. INTRODUCTION

Automotive companies in the world are now increasing rapidly, various automotive companies are creating innovations in their products. This has caused intense competition in the automotive companies in the world. The producers innovate with each other but still meet customer demand. Some of their newest products are the result of innovations from their old products that display attractive attributes and still meet market demand, in addition to attractive shapes, some advanced features are also offered, as well as comfort for drivers and passengers, features that are not forgotten by consumers are safety features. In addition to the features and technologies offered, the affordable financing system and down payment are highly considered by consumers.

Automotive industry products are very competitive in shape, colour, technological sophistication, and brand. The car brands that are on the automotive market in Indonesia today come from European and Asian made such as Mitsubishi, KIA, Honda, Daihatsu, Suzuki, Ford, Proton, Nissan, Hyundai, and Toyota which are much in demand by Indonesian people. Each brand of product offers its own services to its customers, in the form of after-sales services, services, spare parts, to pricing quite

competitive in accordance with the type and market segment. One company that introduced automobiles and at the same time became the object of research was PT. XYZ, with Suzuki car products. In the company PT. XYZ strategy used to attract customers is to consider the quality of services, product quality and the quality of their services.

The purpose of this study is to determine the level of customer satisfaction with the services provided by PT.XYZ and determine the level of accuracy of customer satisfaction predictions. The quality of services that are targeted by P.XYZ must be viewed from the customer's point of view, because the customer is a service assessor so the top priority in quality assurance is the customer's assessment of the quality of PT.XYZ's services. One way to achieve satisfactory service results from customers is to provide excellent and efficient service [1]. In this study, researchers used the decision tree method with the C4.5 algorithm to get a predictive classification decision tree for the Company's service quality. Decision tree is one of the classification methods that represent rules and rules are very easy to understand, therefore decision trees are trusted to explore data and find hidden relationships between variables [2]

1.1. Related Work

Some of the research that participated as references in this study include:

1.1.1. C4.5 Algorithm to Get Model Classification

The method used in this research is C4.5 Algorithm decision tree algorithm group. This algorithm has an input form training samples and samples. Training samples in the form of examples used for a tree that has been tested for correctness. While samples are field-filed data which we will use as parameters in do data classification [3].

In general the C4.5 algorithm for constructing a decision tree is as following:

a. Select the attribute as root
b. Create a branch for each value
c. Divide cases in branches
d. Repeat the process for each branch until all cases are on the branch have the same class.

The results of this study are a model for conducting systems classification. The results of the entropy and gain calculation process are used to form a tree a decision where the largest gain value is used to determine the root. After the decision has been made, the decision will be used as a reference for process if there is a next data input.

1.1.2. C4.5 Algorithm to Analyze The Performance of an Attribute

As in previous studies using the C4.5 algorithm. But in this study the results of the resulting rule are used to analyze the performance of each attribute [4]. If the data is tested using Matlab, a decision tree will be formed from the data. Then the accuracy, precision, recall and f- measure will be calculated, then evaluation and validation of the results will be calculated using the formula of accuracy, precision recall and f-measure.

Testing is done twice, namely by distinguishing the amount of training data and testing data for each test. In the first test training data 60% and testing data 40% and the second test training data 80% and testing data 20%. The ratio of training data used affects the accuracy value in each trial.

1.1.3. The Application of the Decision Tree Algorithm (C4.5) Uses the Adaboost Method.

As in research with the C4.5 algorithm in general, but in this study after conducting the stages in making a decision tree with C4.5 algorithm, weights are given to a single tree such that produce a new hypothesis and a new decision tree with the following steps [5]. In testing the K-Fold Cross Validation Algorithm C4.5 and Algorithms Adaboost-based C4.5, researchers also used 10 trials with stratified sampling type (stratified) using random local use seed because the accuracy is also higher.

In this study conducted using the C4.5 algorithm method C4.5 algorithm is based on the Adaboost method. Try reducing some attribute and retry with another algorithm with optimize other than adaboost which results in a high degree of accuracy.

1.1.4. C4.5 Algorithm in Prediction Case

C4.5 algorithm has the main advantage that it can produce a model in the form of a tree or a rule easy to interpret, has a level of accuracy which is acceptable, can handle type attributes discrete and numerical. In the C4.5 algorithm, the model generated by the process of "learning" from data training in the form of a decision tree. Decision tree this can then be used to predict class of new cases [6]. At the decision tree design stage, a process which includes 3 the stages of the process, the first stage is cleaning data, task-relevant data that is doing data selection has relevant attributes, and the third stage, i.e. data transformation. Data that has been transformed, then analyzed using the C4.5 algorithm by calculating entropy and gain for produce a decision tree. Implementation is done using one Data Mining software, Rapid Miner 5.3.015. All input indicator attributes and destination attributes saved in xlsx format, then imported to Rapid Miner software 5.3.015. From the results of data analysis, there are 4 attributes that are used as a predictor attribute and can determine the classification needed according to the problem.

1.2. Implementation C4.5 Algorithm

In this research, C4.5 algorithm will be implemented to get a classification of good and bad service quality. In

addition, the rules generated from the decision tree will be analyzed the attributes that influence in determining the quality of company services. The accuracy of the prediction will also be proven by testing twice by differentiating the ratio of training data and testing data for each test.

2. BACKGROUND

2.1. C4.5 Algorithm

C4.5 algorithm or commonly known as decision tree C4.5 has a structure similar to a flowchart. The flowchart has a node that represents the attribute value and has a branch that represents the test results and a branch that represents the class. C4.5 algorithm is an algorithm with a classification method and is a development of the ID3 algorithm. C4.5 algorithm has advantages compared to the ID3 and CART algorithms because of its ability to not limit branches in binary and separate forms

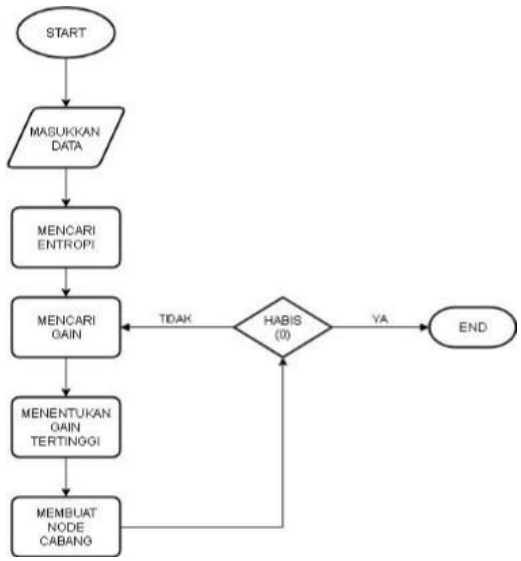

Figure 1 Flowchart C4.5 Algorithm

To select attributes with roots, based on the highest gain value of the existing attributes. To calculate the gain the following formula is used:

$$Gain\ (A) = Entropi\ (S)) - \sum_{i=1}^{k} \frac{|Si|}{|S|} \times Entropi\ (Si)$$

1)

Definition (1):
S: case set A: attribute
N: number of attribute attributes A
| Si |: number of cases on the i partition
| S |: number of cases in S

So that the gain value of the highest attribute will be obtained. Gain is one of the selection measure attributes used to select the test attribute for each node in the tree [8]. The attribute with the highest information gain is chosen as the test attribute of a node. Meanwhile, the calculation of entropy values can be seen in the equation:

$$\text{Entropi}\ (S) = -\sum_{i=1}^{k} \frac{|Si|}{|S|} \times (p_j\ log_2 p_j)$$

2)

Definition (2):
S: case set A:
attribute
N: number of partitions S
Pj: proportion of S

2.2. Customer Satisfaction Analysis

The assessment data used to assess the quality of company services is to use a survey that will be distributed to 500 customers. The survey contains questions about assessing the quality of company services. In this question there are 6 variables that will be used to form a decision tree with the C4.5 algorithm.

Survey data that has been inputted is then calculated entropy and gain to make a decision tree. The decision tree can later find out the results of service quality in the company PT. XYZ is good or bad. After getting the results of entropy and gain calculation, the highest gain value is selected, which is found in the waiting room attribute which is used as the root node in the decision tree and then the calculation is resumed with the remaining attributes and the highest gain is speed.

The process of calculating the waiting room attribute is not comfortable with the remaining attributes, the highest gain value is obtained in the speed attribute so that the decision tree becomes like in Figure 2.

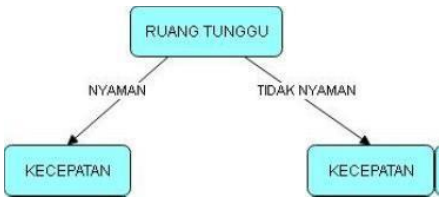

Figure 2 Decision Tree Node 2

Because the entropy yield of fast speeds is 1, fast speeds are good. So, for branches of fast speed is good as in Figure 3.

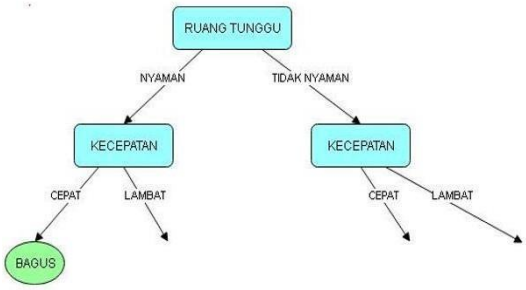

Figure 3 Decision Tree Node 3

While the slow entropy is still 0, it is necessary to recalculate the entropy and its gain. Then the highest gain is the price for the slow speed branch with a comfortable waiting room is the price. And also, the entropy yield from expensive is 0, then the branch of expensive is bad as in figure 4

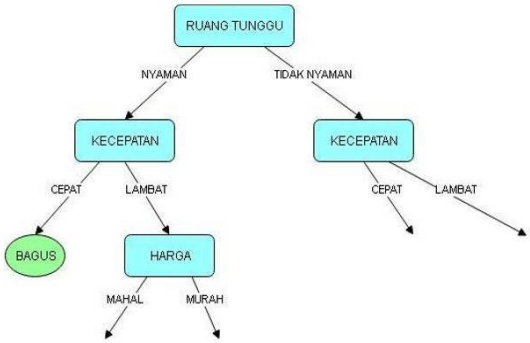

Figure 4 Decision Tree Node 4

To find branches from an uncomfortable room and fast and slow speeds are calculated back in entropy and gain, entropy of fast speed is 1, then branches of slow speed are bad. The highest gain is from the price attribute so the branch of fast is the price like Figure 5

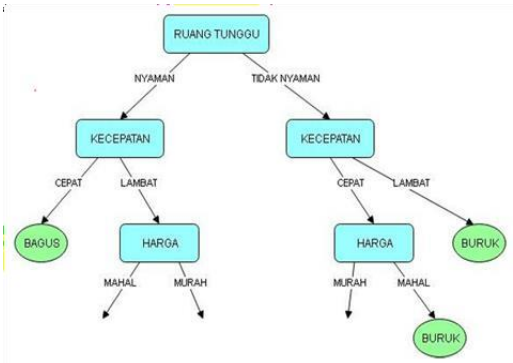

Figure 5 Decision Tree Node 5

From the tree above, to find out the branches of expensive and cheap slow speeds the comfortable room is re- calculated entropy and gain. From the above calculation, the highest gain in the spare part attribute is obtained. So the branch for a cheap price is a spare part then the branch from the expensive one is bad like picture 6.

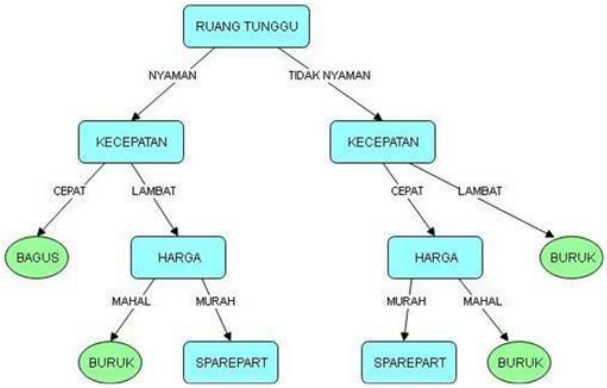

Figure 6 Decision Tree Node 6

From the tree above, to find out the branch of the spare part, the entropy and the gain of the spare part are recalculated. From the calculation, the service gain is worth 1. Then the branch for complete spare parts is service while incomplete is bad, then the branch of friendly service is good and the branch of service that is not friendly is bad as shown in Figure 7.

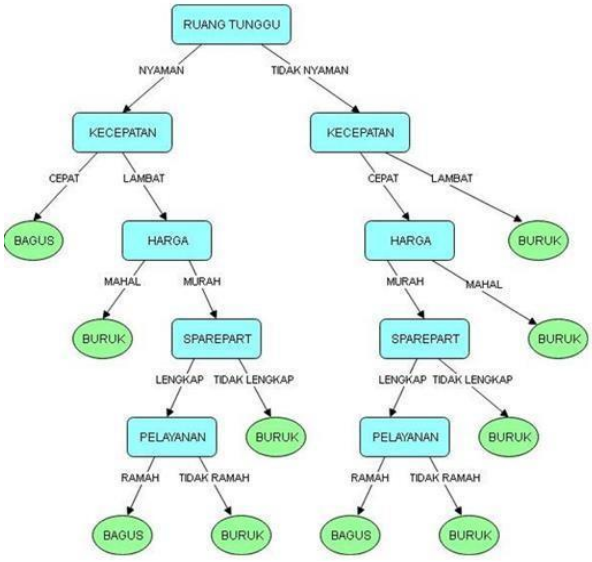

Figure 7 Decision Tree Node 7

In the picture above, it can be concluded that the waiting room attribute has a positive influence on the

quality of service to the customer. But the waiting room becomes uncomfortable if the speed of service is slow, and what affects the speed of service is fast in terms of price. What influences low prices is from spare parts while those that affect complete spare parts are from services. The service quality of a company is said to be good can be summarized as follows:

- comfortable waiting room and fast service speed
- Comfortable waiting room and slow service speed but cheap prices and complete spare parts and friendly company service
- The waiting room is uncomfortable and the service speed is fast but cheap prices and complete spare parts and friendly company service

Whereas the quality of company services is said to be bad if:

- The waiting room is comfortable and the service speed is slow and the price is expensive
- The waiting room is comfortable and the service speed is slow but the prices are cheap and the spare parts are incomplete
- The waiting room is comfortable and the service speed is slow but the prices are cheap and the spare parts are complete but the company service is not friendly
- The waiting room is uncomfortable and the speed is slow
- The waiting room is uncomfortable and the speed is fast but the price is expensive
- The waiting room is uncomfortable and the service speed is fast but the price is cheap but the spare parts are incomplete
- The waiting room is uncomfortable and the service speed is fast but the prices are cheap but the spare parts are complete but the service is not friendly

3. RESULTS AND DISCUSSION

a) Testing Rules

After getting the rules, the rules will then be tested. Because C4.5 Algorithm comes from a decision tree that has several vertices that describe the attributes and each branch describes the results of the attributes being tested [9]. The test is conducted to determine whether the solutions or rules are valid by the decision tree. Rules are said to be valid if the number and satisfied customers match the dataset. The dataset is divided into two: 60% training data and 40% testing data. The total dataset is 500 data. After that training and testing data is obtained:

Table 1 Testing Rules Results

The number of data testing	Error	Positif Benar (TP)	Positif Salah (FP)	Negatif Salah (FN)
200	19	162	11	8

Then it can be seen from the results of the rules by dividing the amount of data that is classified fully by the total test data testing as follows,

$$\text{Akurasi} = \frac{\text{Jumlah Data testing} - \text{Error Jumlah Data testing}}{\textit{Jumlah Data Testing}}$$

$$\textbf{Akurasi} = \frac{\mathbf{200 - 19}}{\mathbf{200}}$$

$$\text{Akurasi} = 0{,}905$$

In addition to accuracy, the level of precision is also calculated or also called precision which is the ratio of true positive predictions compared to the overall positive predicted results. Precision answers the question "How many customers rate good service from all customers predicted to rate good service?" As in the following calculation,

$$\textit{Precission} = \frac{TP}{TP+FP}$$

$$\text{Akurasi} = \frac{162}{162+11} = 0{,}936$$

To guarantee the predicted results of the C4.5 algorithm above, the recall value is calculated where recall is the success rate of the system in rediscovering information or it can be said that the recall is a positive predictive ratio compared with overall positive true data. Recall answers the question "What percentage of customers are predicted to rate good service compared to overall customers who actually rate good service?" So we get the recall value as follows.

$$\textit{Recall} = \frac{TP}{TP+FN}$$

$$\text{Akurasi} = \frac{162}{162+8}$$

$$\text{Akurasi} = 0{,}952$$

Then calculate the value of F-Measure F-Measure is the harmonic mean of precision and recall. The F-Measure value is obtained from the calculation of the division of results from precision and recall multiplication with the sum of precision and recall, then multiplied by two.

$$F - Measure = \frac{2\,.\,precision\,.recall}{precision\,+\,recall}$$

$$Akurasi\ = \frac{2.0.94.0.94}{0.94 + 0.95}$$

$$Akurasi = 0{,}935$$

So it can be concluded the percentage of each value of accuracy, precision and sensitivity are presented in the following table.

Table 2 Percentage of Accuracy, Precission and Recall

raining Data	esting Data	Accuracy	Precision	Recall	F-Measure
300	200	90%	94%	95%	94%

b) Start Page

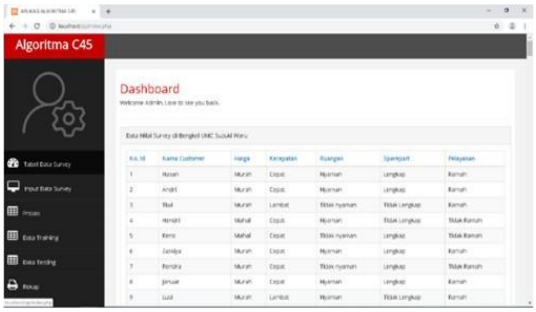

Figure 8 Main Page

The initial menu of the application can be seen in Figure 8 where there are 6 menus namely survey data table menu, survey data input menu, process menu, training data menu, data testing menu, and recap menu. This dashboard menu serves to display data that has been inputted from various survey data filled out by customers.

c) Form Fill in Customer Survey Data

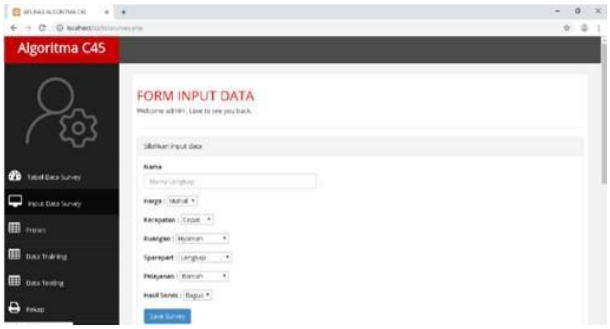

Figure 9 Form Fill in Page

The contents of the user survey data menu can be seen by the admin in Figure 9 where this menu is for inputting all survey results conducted by the customer. Which will be carried out the calculation process to determine the level of satisfaction with the customer.

d) Calculation process page

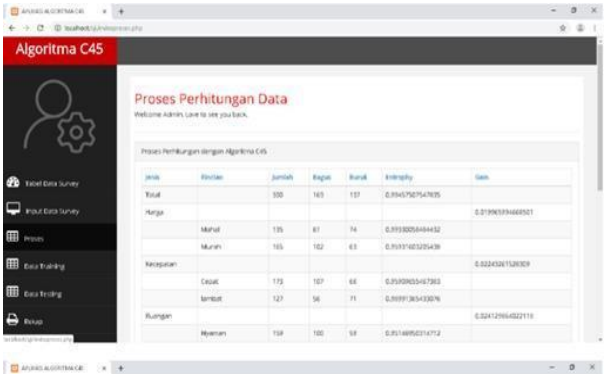

Figure 10 Calculation Page

The user calculation process menu can be seen in Figure 10 where this menu is to perform the calculation process which will later calculate entropy and gain from survey data at PT XYZ Company that has been inputted.

e) History of Calculation Results

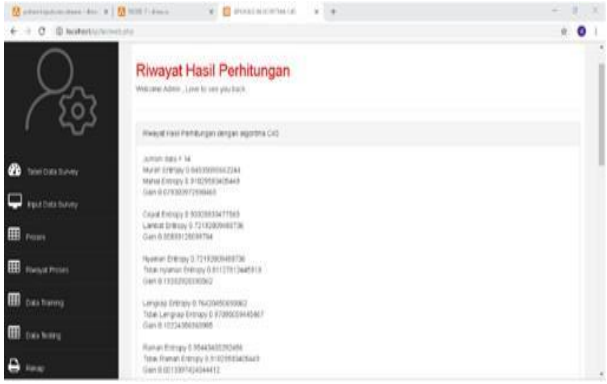

Figure 11 Calculation History Page

This page displays the history of the entropy and gain calculation process from start to finish to make a decision tree.

f) Training Data and Testing Data Trial Page

Training data menu and testing data can be seen in Figure 12 and Figure 13 where this menu is to display training data from 300 customer data and 200 PT testing data. XYZ.

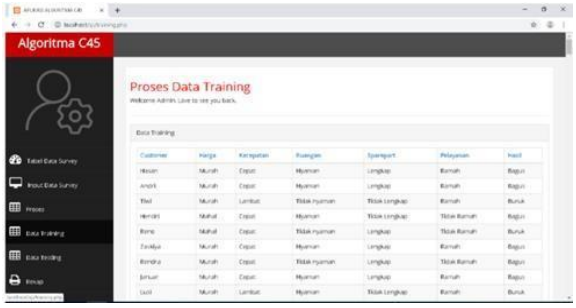

Figure 12 Training Process

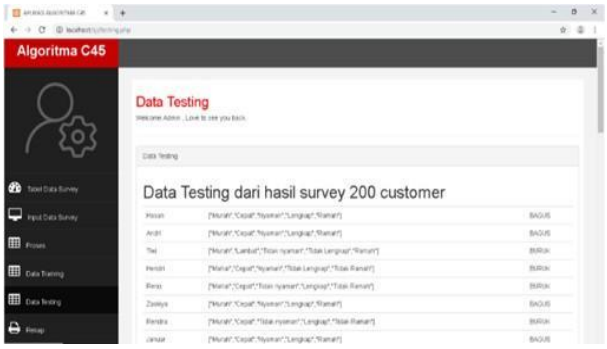

Figure 13 Testing Process

g) Trial Test page for Calculation and Accuracy Results

Figure 14 Trial Test Page

In the figure 14 there is a conclusion of all the rules produced that will be a company record to improve service quality. There are 10 rules with the conclusion that the quality of services is good and bad.

4. CONCLUSION

The conclusion of the results obtained from the calculation of training data 60% of 500 data and testing data 40% of 500 \ data, there are 19 data errors, then true data that is positive as many as 162 data, the total result of false data which is positive is 11 data, the results of total wrong data are negative as many as 8 data. So that the accuracy is obtained at 90%, Precision at 94%, Recall at 95%, and F- Measure at 94%.

With the analysis of PT. XYZ's service quality to customers using the C4.5 method, the level of PT. XYZ's service can be clearly measured. After being analyzed from several aspects, the benchmark for assessing service quality is the waiting room. The waiting room aspect is the most dominant aspect of several aspects of the quality of services provided by PT. XYZ.

REFERENCES

[1] Shiddiq, A., Niswatin, R. K., & Farida3, I. N. (2018). Analisa Kepuasan Konsumen Menggunakan Klasifikasi Decision tree Di Restoran Dapur Solo (Cabang Kediri). Generation Journal, 2(1), 9-18.

[2] Febriyanto, D. B., Handoko, L., Wahyuli, Aisyah, H., & Rumini. (2018). Implementasi Algoritma C4.5 Untuk Klasifikasi Tingkat Kepuasan Pembeli Online Shop. Jurnal Riset Komputer , 5(6), 569-575.

[3] Lestari, S. (2014). Model Klasifikasi Kinerja Dan Seleksi Dosen Berprestasi Dengan Algoritma C4.5. Proseding Seminar Bisnis & Teknologi, 340-350.

[4] Wiratama, F. R., & Astuti, S. (2016). Implementasi Algoritma C4.5 untuk Analisa Performa Pelayanan Bank Terhadap Nasabah. Eksplora Informatika, 127-135.

[5] Rohman, A., Suhartono, V., & Supriyanto, C. (2017). Penerapan Algoritma C4.5 Berbasis Adaboost Untuk Prediksi Penyakit Jantung. Jurnal Teknologi Informasi, 13(1), pp. 13–19.

[6] Rismayanti. (2016). Implementasi Algoritma C4.5 Untuk Menentukan Penerima Beasiswa Di Stt Harapan Medan. Jurnal Media Infotama, 12(2), 116-120.

[7] Supangat, Amna, A. R., & Rahmawati, T. (2019). Implementasi Decision tree C4.5 untuk Menentukan Status Berat Badan dan Kebutuhan Energi pada Anak Usia 7-12 Tahun. Teknika, 7(2), 73-78.

[8] Rani, L. N. (2016). Klasifikasi Nasabah Menggunakan Algoritma C4.5 Sebagai Dasar Pemberian Kredit. Jurnal Inovtek Polbeng - Seri Informatika, 1(2), 126-132.

[9] Nasrullah, H. A. (2018). Penerapan Metode C4.5 Untuk Klasifikasi Mahasiswa Berpotensi Drop Out. ILKOM-Jurnal Ilmiah, 10(2), 244-250