JOSCEX
Journal of Soft Computing Exploration

Classification of potential customers using C4.5 and k-means algorithms to determine customer service priorities to maintain loyalty

Nur Hazimah Syani Harahap[1], Afif Amirullah[2], Meidika Bagus Saputro[3], Ilham Alzahdi Tamaroh[4],
[1,2,3]Department of Computer Science, Universitas Negeri Semarang, Indonesia
[4]Department of Computer Science, Universitas Negeri Jakarta, Indonesia

Article Info

Article history:

Received Sep 21, 2022
Revised Sep 24, 2022
Accepted Sep 29, 2022

Keywords:

Small company
Potential customer
Customer service
Maintenance
Loyalty

ABSTRACT

The increasing competition among Middle-Class Micro Enterprises (MSMEs) is a problem because business actors must improve techniques and strategies to maintain customer satisfaction, and the number of customers continues to increase. Customers are an essential asset for the company. To maintain customer loyalty with promising prospects for the company, a strategy is needed to support this. Strategies such as service prioritization can be used to maintain customer loyalty. This research was conducted to classify customers who are estimated to have good prospects for the company so that service priorities are not mistargeted by utilizing 1683 data from store By.SIRR, a fashion store in Semarang, Indonesia contains five attributes, and customers are classified and are estimated to have promising prospects for the company. Data mining methods use the C4.5 and K-Means algorithms to classify the classification process. The research resulted in the grouping of customers into four categories: potential lover, flirting, faithful lover, and spiritual friend. From the validation test conducted using the Confusion Matrix Validation method, the classification results get an Accuracy of 97.70%.

Corresponding Author:

Nur Hazimah Syani Harahap,
Department of Computer Science,
Universitas Negeri Semarang,
Sekaran, Gunungpati, Semarang, Indonesia
Email: nurhazimahsyani@students.unnes.ac.id

1. INTRODUCTION

The magnitude of the growth of the MSME sector in Indonesia certainly raises the existence of tight business competition among MSME actors. Increased competition among MSMEs requires efforts to improve techniques and strategies to maintain customer satisfaction levels to continue to increase [1]. Maintaining customer satisfaction is one way that customers are not lost because losing customers reduces sales results and increases the cost of the need to attract new customers. Research shows that the cost of acquiring new customers is estimated to be 5 to 6 times greater than retaining existing customers [2].

Customers are assets for the company, and of course, each customer has different preferences [3]. To maintain customer loyalty to the ongoing MSME business, business actors can provide priority services according to customer needs. The application of priority service certainly cannot be given to all customers because high costs are needed to reduce costs if it is applied to all customers. This implementation will only target potentially profitable customers, so it is necessary to identify customers who have the potential to provide benefits for MSME businesses so that they are loyal and loyal. Loyal to the SME business.

A model is needed to analyze the level of potential customers that becomes a reference for the implementation of service priorities to maintain their loyalty [4]. In this study, the data mining algorithm used is the Decision Tree algorithm. K-Means algorithm to segment customers so that their potential level can be measured. The level of potential customers will be added as an attribute to help classify loyalty with the C4.5 Algorithm so that the accuracy of the C4.5 Algorithm will be better [5][6].

The Decision Tree algorithm used in this study is the C4.5 Algorithm. The C4.5 Algorithm is one of the algorithms used to classify data with numeric and categorical attributes [7]. The C4.5 Algorithm is considered easy to understand because it is a derivative of an interpretation algorithm that is very easy to find parameter settings to build a model accurately [8]. In this study, customers will be segmented with the K-Means algorithm based on customer payment information to measure the level of potential customers. The next step is that the C4.5 Algorithm will classify each segment and attribute so that the accuracy of the C4.5 Algorithm will be better and provide predictions at a better level of accuracy than before segmentation [9].

C4.5 is often applied with the K-Means algorithm [10]. K-Means is an optimization algorithm used to improve the quality of the parameters that will be used in the clustering process [11]. A dynamic function is required for a set of parameters that affect clustering performance improvement [12]. In previous research, the K-Means algorithm had a weakness, namely in determining the number of clusters [13]. One good method for determining the number of clusters is the Elbow Method [14]. This method is used in cluster analysis for interpretation and performance testing of the consistency of the right number of clusters by testing the SSE value. At some point, the graph will descend significantly into a curve called the angular criterion. This value then becomes the best value of k or the number of clusters [15].

Therefore, this study aims to classify the potential value of customers by segmenting customers using the K-Means algorithm based on the LRFM model (Length, Recency, Frequency, and Monetary). This study focuses on classifying potential customers at store By.SIRR uses the C45 Algorithm, more widely known as the Decision Tree, by utilizing the LRFM model in determining attributes and using the K-Means algorithm, and using the Elbow method to determine the best cluster in its segmentation so that business owners can provide priority of service to them to maintain their loyalty.

2. METHOD

The research method is the steps used as a reference in researching so that the implementation of research can be justified scientifically. The following are the methods used in compiling this research:

1. Create Labels on data so that data is easy to classify [16].
2. Select data according to attributes L, R, F, and M

   LRFM model is one of the most common segmentation methods that can identify customer value in a company with 3 variables: novelty, communication, and monetary [17]. In this study, L, R, F, and M attributes have different meanings. L is the Length, in this case, it represents the length of the interval between the initial transaction and the customer's final transaction in a certain period. R is Recency, which is the last date the customer made the transaction to the company. The period taken in this study is data from January to April. F is the Frequency which in this case means how often customers make transactions at the By.SIRR store in the period from January to April. Furthermore, M is monetary, which in this case means how much money customers have spent from January to April.
3. Calculating customer's L, R, F, and M values.
4. Calculating LRFM *score*

   The LRFM score is obtained by adding up the values of L, R, F, and M
5. Standardization of data because the difference in data between L, R, F, and M is very far, so it needs to be standardized to the same scale. Standardization is converting the values of a feature so that these values have the same scale. This standardization process uses the standard scaler in the SKLearn library. In the standard scaler, there is a fit function to calculate each attribute column's average and standard deviation to be used in the transform function. The transform function is used to apply a standard scaler to the data.
6. Segmentation with K-Means clustering for *k=1* to *k=n*

   K-Means is one of the clustering algorithms commonly used to group data according to similar characteristics, and grouping the data can be called a cluster [18]. K-Means training is carried out on each input vector to be mapped to the nearest weighted cluster [19]. K-Means that use the winning weights and all their neighbors to be dynamically updated by the objective function require optimization techniques based on the input data. [20]. The stages in the K-Means segmentation method are as follows:

1) Determine the number of clusters,
   Choose the best cluster based on the Elbow method. The Elbow method is used to determine the number of clusters of a data set, identifying the clusters in such a way that the total internal variation (the total number of variances in the cluster or the sum of the squares of the clusters) is minimized [15]. This method is a visual method that starts with k = 2 and increases at each step by adding 1 to the value of k. At the value of k = 3, if there is a strong chance that is inversely proportional to the previous value, the value before the change is considered the most appropriate number of clusters [21].
2) Choose the initial centroid at random according to the number of clusters,
3) Calculate the distance of the data to the centroid with the following Euclidean distance formula (1):

$$dxy\ = \sqrt{\sum} \left(xi\ - yi\right) n\ 2\ i = 1 \qquad (1)$$

4) Update the centroid by calculating the average value in each cluster,
5) Return to stage 3 if there is still data that moves clusters or changes in the centroid value[22].
6) Create customer segmentation with K-Means Clustering based on L, R, F, and M values. In creating customer segmentation, the L, R, F, and M values are processed into the same scale and then normalized to 0 so that the value can be maintained for the next process. After that, check the skewness of each value of L, R, F, and M. If so, then a transformation is carried out to get a normally distributed variable. After that, clustering was carried out using K-Means.
7) Classification using decision tree
   In classifying using the decision tree algorithm. The remaining attributes that will be used for the classification of customer loyalty are Loyalty, StockCode, Description, Country, and Cluster. Then classification without k-means, the attributes are Customer ID, Loyalty, StockCode, Description, Quantity, InvoiceDate, Price, Country, TotalPrice, and Date.
8) Classification results were tested with a confusion matrix using the RapidMiner tools.
   Confusion Matrix can be used to measure algorithm performance in the world of data mining. This is a common method for calculating accuracy [23].

3. RESULTS AND DISCUSSIONS

This study uses sales data at the By.SIRR Store, which contained 1683 data with 8 attributes in it. The attributes are Invoice, Stock Code, Description, Quantity, Invoice Date, Price, Customer ID, and Country.

Before clustering, data preprocessing needs to be done by carrying out several stages, including cleansing and data transformation with a standardization process used to transform the range of values for each variable to smaller values or the same scale [24]. The preprocessing process until the clustering process is carried out using Google Collab IDE [25].

Based on the available data, segmentation and classification will be carried out. To test the classification results, this study uses a confusion matrix.

3.2.1. Segmentation using the K-m-Means Algorithm

The first process in customer segmentation is to take certain data attributes to be segmented using the K-Means algorithm. These attributes are selected from the LRFM method: Date, Invoice Date, Invoice, and Total Price. Table 1 shows the results of the LRFM scores obtained.

Table 1. LRFM score

Username (Pembeli)	Length	Recency	Frequency	Monetary
.tiaraa_	0	13	1	217316
012ratnawati	0	37	1	165868
divaniameliaputri	46	6	3	525264
09xx366bl_	0	12	1	155817
0f25q52619	0	103	1	158007
…	…	…	…	…
0mrbtt5aa0	0	112	1	139407
10ika_chacha	0	110	1	156240
rasya_aca	49	13	2	294856
141101.	0	8	1	165868
nins. official	60	59	4	1912327

In the segmentation process with K-Means, 9 clusters have been formed. In this test, the performance of each number of clusters is adjusted to the range of values in the Elbow method. Figure 1 is the result of testing with the Elbow method.

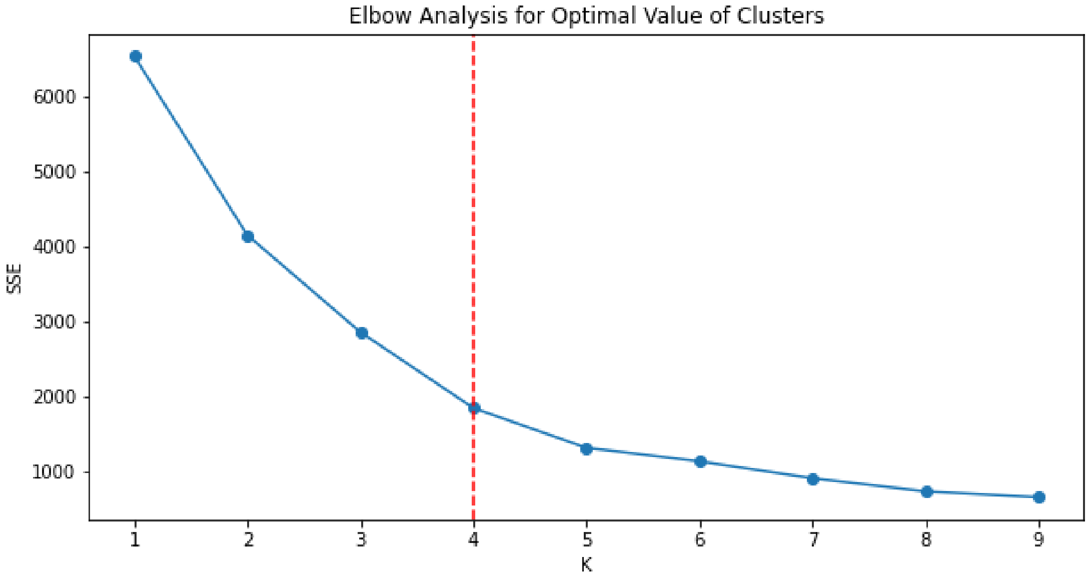

Figure 1. Elbow graphic result

Based on Figure 1, a decrease is seen in cluster-0 and cluster-1. cluster-2 and cluster-3 are marked by the most graph declines, while at the next point, there is a steady decline. Then the value of k used is 4.

Segmentation is done based on the LRFM attributes that we have specified. Table 2 shows the results of segmentation which are grouped into 4 clusters.

Table 2. Segmentation result

Username	Length	Recency	Frequency	Monetary	Cluster
.tiaraa_	0	13	1	217316	0
012ratnawati	0	37	1	165868	0
divaniameliaputri	46	6	3	525264	1
09xx366bl_	0	12	1	155817	0
0f25q52619	0	103	1	158007	3
…	…	…	…	…	…
0mrbtt5aa0	0	112	1	139407	3
10ika_chacha	0	110	1	156240	3
rasya_aca	49	13	2	294856	1
141101.	0	8	1	165868	0
nins.official	60	59	4	1912327	2

3.2.2. Classification with C4.5 Algorithm

The following process after segmentation using the K-Means algorithm is classification using the C4.5 Algorithm. The decision tree shows that customers are classified into 2 groups. as in Figure 2, this process produces a decision tree visualization that describes the customer grouping measured based on the LRFM score obtained from sales data that has been previously processed using the K-Means algorithm. while Figure 3 illustrates the grouping of customers measured based on the LRFM score obtained from sales data that is not processed using the K-Means algorithm.

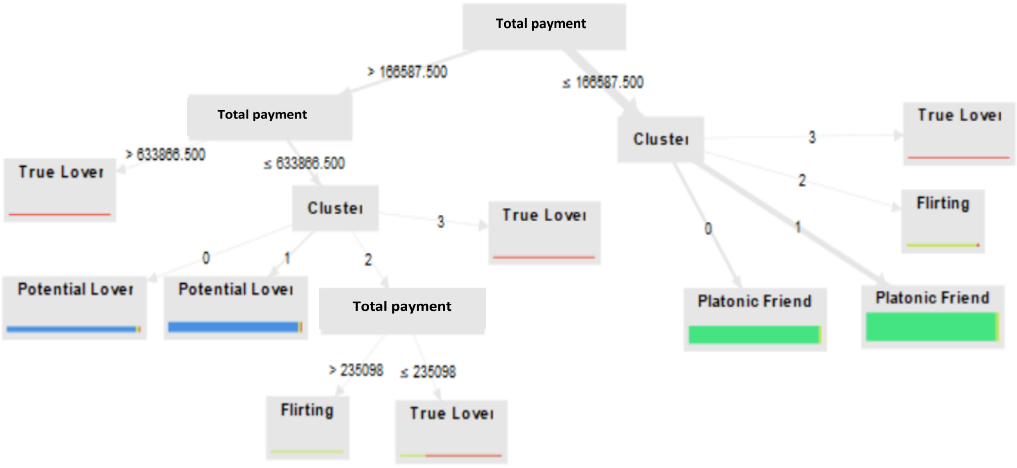

Figure 2. Classification with C4.5 after segmentation using k-means

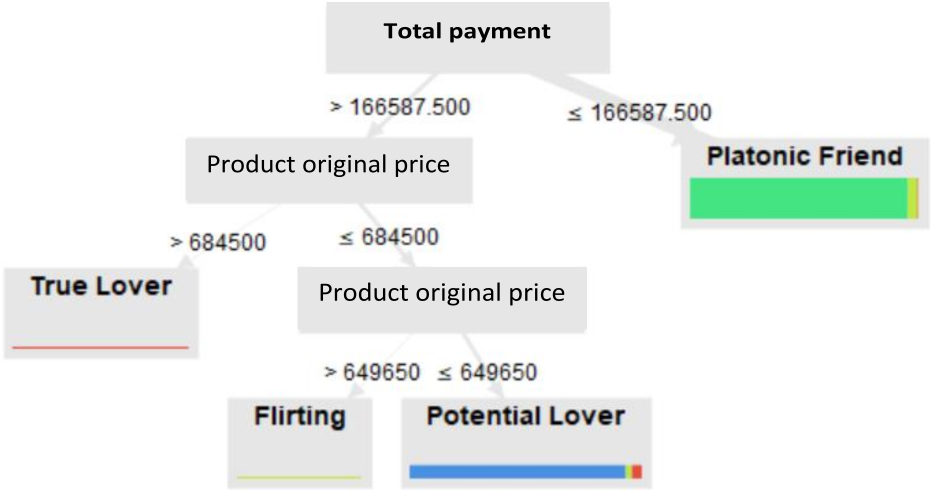

Figure 3. Classification with c4.5 algorithm without segmentation using k-means

3.2.3. Data Validation by Method *Confusion Matrix*

Next, the classification results were tested using the confusion matrix. Table 3 shows that the results of the confusion matrix validation test with K-Means segmentation first using the rapid miner tools get an accuracy result of 97.70%. In contrast, Table 4 shows that the results of the confusion matrix validation test without k-means segmentation first using the rapid miner tools get accuracy results 94.95%.

Table 3. Confusion matrix test results after k-means segmentation

Accuracy: 97.70%					
	true Potential Lover	true Platonic Friend	true Flirting	true True Lover	class precision
Pred. Potential Lover	288	0	5	3	97.30%
Pred. Platonic Friend	0	987	19	0	98.11%
Pred.Flirting	0	0	29	4	87.88%
Pred.True Lover	0	0	0	11	100.00%
Class recall	100.00%	100.00%	54.72%	61.11%	

Table 4. Confusion matrix test results without k-means segmentation

Accuracy: 94.95%					
	true Potential Lover	true Platonic Friend	true Flirting	true True Lover	class precision
Pred. Potential Lover	288	0	8	13	93.20%
Pred. Platonic Friend	0	987	44	3	95.45%
Pred.Flirting	0	0	1	0	100.00%
Pred.True Lover	0	0	0	2	100.00%
Class recall	100.00%	100.00%	1,89%	11.11%	

4. CONCLUSION

The progress of MSMEs that are currently advancing is causing competition between business actors in maintaining their customers. Customers are assets for companies where customers themselves have different preferences. Management of the implementation of service priorities to customers must be adjusted to the preferences of each customer. Therefore, we need a model that can analyze the level of potential customers as a reference for implementing service priorities to maintain their loyalty. The K-Means and C4.5 algorithms were chosen in this study. The K-Means algorithm is used to segment payment behavior so that the level of potential customers can be measured. While the C4.5 algorithm is used to help classify loyalty with the C4.5 algorithm. In this classification, they are divided into four groups, namely true lover, flirting, potential lover, and platonic friend. This classification shows very good results as indicated by the results of the validation test which obtained an accuracy of 97.70%. From the results of this classification, business actors can provide the right service priorities to maintain customer loyalty.

REFERENCES

[1] D. Lee and K. Hosanagar, "How do product attributes and reviews moderate the impact of recommender systems through purchase stages?," *Manage. Sci.*, vol. 67, no. 1, pp. 524–546, 2021.

[2] W. Hengliang and Z. Weiwei, "A customer churn analysis model in e-business environment," *China Int. J. Digit. Content Its Appl.*, 2012.

[3] T. Hong and E. Kim, "Segmenting customers in online stores based on factors that affect the customer's intention to purchase," *Expert Syst. Appl.*, vol. 39, no. 2, pp. 2127–2131, 2012.

[4] W. Abd and A. Al Manhawy, "TQM critical success factors in hospitality Industry and their impact on Customer Loyalty, a theoretical Model," *Int. J. Sci. Eng. Res.*, vol. 4, no. 1, pp. 1–15, 2013.

[5] E. Sugiharti and M. MUSLIM, "On-Line Clustering of Lecturers Performance of Computer Science Department of Semarang State University using K-Means algorithm.," *J. Theor. Appl. Inf. Technol.*, vol. 83, no. 1, 2016.

[6] S. Moedjiono, Y. R. Isak, and A. Kusdaryono, "Customer loyalty prediction in multimedia Service Provider Company with K-Means segmentation and C4. 5 algorithm," in *2016 Int. Conf. Inform. Comput. (ICIC)*, 2016, pp. 210–215.

[7] J.-S. Lee, "AUC4. 5: AUC-based C4. 5 decision tree algorithm for imbalanced data classification," *IEEE Access*, vol. 7, pp. 106034–106042, 2019.

[8] J. S. Mapa, A. Sison, and R. P. Medina, "A Modified C4.5 Classification Algorithm: With the Discretization Method in Calculating the Goodness Score Equivalent," *ICETAS 2019 - 2019 6th IEEE Int. Conf. Eng. Technol. Appl. Sci.*, pp. 4–7, 2019.

[9] I. S. Damanik, A. P. Windarto, A. Wanto, S. R. Andani, and W. Saputra, "Decision tree optimization in C4. 5 algorithm using genetic algorithm," in *J. Phys.: Conf. Ser.*, 2019, vol. 1255, no. 1, p. 12012.

[10] A. P. Muniyandi, R. Rajeswari, and R. Rajaram, "Network anomaly detection by cascading k-Means clustering and C4. 5 decision tree algorithm," *Procedia Eng.*, vol. 30, pp. 174–182, 2012.

[11] A. Nazir, A. Akhyar, Y. Yusra, and E. Budianita, "Toddler Nutritional Status Classification Using C4. 5 and Particle Swarm Optimization," *Sci. J. Informatics*, vol. 9, no. 1, pp. 32–41, 2022.

[12] M. A. Farag, M. A. El-Shorbagy, I. M. El-Desoky, A. A. El-Sawy, and A. A. Mousa, "Genetic algorithm based on k-means-clustering technique for multi-objective resource allocation problems," *Br. J. Appl. Sci. Technol.*, vol. 8, no. 1, pp. 80–96, 2015.

[13] F. Marisa, S. S. S. Ahmad, Z. I. M. Yusof, F. Hunaini, and T. M. A. Aziz, "Segmentation model of customer lifetime value in small and medium enterprise (SMEs) using K-means clustering and LRFM model," *Int. J. Integr. Eng.*, vol. 11, no. 3, 2019.

[14] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," in *IOP conf. ser.: mater. sci. eng.*, 2018, vol. 336, no. 1, p. 12017.

[15] P. Bholowalia and A. Kumar, "EBK-means: A clustering technique based on elbow method and k-means in WSN," *Int. J. Comput. Appl.*, vol. 105, no. 9, 2014.

[16] S. Singhal and M. Jena, "A study on WEKA tool for data preprocessing, classification and clustering," *Int. J. Innov. Technol. Explor. Eng.*, vol. 2, no. 6, pp. 250–253, 2013.

[17] A. Amine, B. Bouikhalene, and R. Lbibb, “Customer segmentation model in e-commerce using clustering techniques and LRFM model: The case of online stores in Morocco,” *Int. J. Comput. Inf. Eng.*, vol. 9, no. 8, pp. 1993–2003, 2015.

[18] N. Dhanachandra, K. Manglem, and Y. J. Chanu, “Image segmentation using K-means clustering algorithm and subtractive clustering algorithm,” *Procedia Comput. Sci.*, vol. 54, pp. 764–771, 2015.

[19] K. P. Sinaga and M.-S. Yang, “Unsupervised K-means clustering algorithm,” *IEEE access*, vol. 8, pp. 80716–80727, 2020.

[20] D. Binu, “Cluster analysis using optimization algorithms with newly designed objective functions,” *Expert Syst. Appl.*, vol. 42, no. 14, pp. 5848–5859, 2015.

[21] T. M. Kodinariya and P. R. Makwana, “Review on determining number of Cluster in K-Means Clustering,” *Int. J.*, vol. 1, no. 6, pp. 90–95, 2013.

[22] S. Ghosh and S. K. Dubey, “Comparative analysis of k-means and fuzzy c-means algorithms,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 4, 2013.

[23] B. Prasetiyo and M. A. Muslim, “Analysis of building energy efficiency dataset using naive bayes classification classifier,” in *J. Phys.: Conf. Ser.*, 2019, vol. 1321, no. 3, p. 32016.

[24] X. Peng, H. Zhu, J. Feng, C. Shen, H. Zhang, and J. T. Zhou, “Deep clustering with sample-assignment invariance prior,” *IEEE Trans. neural netw. Learn. Syst.*, vol. 31, no. 11, pp. 4857–4868, 2019.

[25] Y. Tokuyama, Y. Furusawa, H. Ide, A. Yasui, and H. Terato, “Role of isolated and clustered DNA damage and the post-irradiating repair process in the effects of heavy ion beam irradiation,” *J. Radiat. Res.*, vol. 56, no. 3, pp. 446–455, 2015.