TIFDA JOURNAL
Technology Information and Data Analytic
Website Jurnal : http://jeptira.id/index.php/tifda
e-ISSN : 3064-0660

Comparison of Decision Tree and K-Means Clustering Algorithms to Determine Awards for Customer Loyalty

Bagus Tri Mahardika [1]*, Donnie Varyasetya Prastowo [2]
[1,2] Information Technology Program, Darma Persada University
[1]Jl. Taman Malaka Selatan, East Jakarta, DKI Jakarta 13450, Indonesia
*bagusunsada@gmail.com

Abstract — PT Tangguh Buana Roda Indonesia has difficulty in retaining loyal customers due to less than optimal customer management. This research proposes the use of a data mining-based system to categorize loyal customers using the K-Means and Decision Tree methods. The evaluation shows that the combination of K-Means and Decision Tree algorithms provides a higher average accuracy of 93.7175%. Compared to using Decision Tree alone which reached 92.8525% and K-Means which was only 91.667%. With the combination of these two algorithms, it is expected to support the awarding of loyal customers and strengthen the relationship between customers and companies. The system that has been created is web-based which will facilitate strategic planning to increase customer loyalty.

Keywords - Customer Loyalty, Data Mining, K-Means Clustering, Decision Tree, Classification.

I. INTRODUCTION

Data mining is a method for identifying patterns in large amounts of data. Although this technique is mostly studied in the fields of computer science and statistics, data mining can be used in various aspects to facilitate work. Common techniques used in data mining include clustering and classification. Many companies use clustering and classification analysis to group and label their data. This helps optimize the sales process and strengthen customer relationships.

One of the commonly used clustering methods is K-Means. This method groups data with similar characteristics into one cluster, while data with different characteristics are grouped into other clusters. For classification methods, Decision Tree is an algorithm that can predict the label or class of data. It maps out possible outcomes based on a set of choices or rules and has a tree-like structure, with each branch representing a choice or condition that leads to a final decision.

In an increasingly fierce competition, retaining customers becomes more challenging. PT Tangguh Buana Roda Indonesia often faces a situation where loyal customers who often buy their tire products move to other companies, which is caused by less than optimal loyal customer management in the company. Therefore, the researcher argues that PT Tangguh Buana Roda Indonesia needs a data mining system to process data and categorize loyal customers, so that it can identify which customers deserve an award.

In the inventory system of PT Tangguh Buana Roda Indonesia, there is data on outgoing goods every month. However, so far the data has not been utilized by the company. The data can actually be processed further to provide benefits and increase customer loyalty. Therefore, researchers will use Decision Tree and K-Means clustering algorithms to analyze the data in order to determine awards for loyal customers.

This research will analyze the inventory data of PT Tangguh Buana Roda Indonesia for the last few months. This data will be used to categorize customer loyalty. The results of this research are expected to be useful for PT. Tangguh Buana Roda Indonesia in making more informed decisions regarding the awarding of loyal customers.

II. THEORITICAL FOUNDATION

A. Customer Loyality

Reporting from the journal [1] Loyalty in this context means when customers spend a lot of money to buy products from the company, in other words, customers

do not mind the amount of costs incurred. Therefore, customer loyalty in a company is very important to pay attention to.

Basically, customers who are loyal to buy products from the company will indirectly help marketing by recommending the product to those closest to them. This makes the company's products a priority in the hearts of loyal customers. However, if a company cannot pay attention to its customers, it is likely that customers will switch their hearts to competing companies.

B. Data Mining

Reporting from the journal [2] Data mining is a method used to analyze large amounts of data with the aim of finding relationships between data and presenting them in an easy-to-understand format. So it can be concluded that data mining is a very good method used to analyze data in a company, and can also help provide decision recommendations for company strategies so that they can develop in the future.

C. Data Processing Method

CRISP- DM (Cross Industry Standard Process for Data Mining) is a data mining process designed to ensure data goes through each phase structured, clearly defined, and efficient [3]. This process consists of six stages, including Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

D. K-Means Clustering

K-Means is a simple iterative method for dividing data into a number of clusters set by the user. The purpose of this data clustering is to reduce the objective function set in the clustering process, with an effort to reduce variation within a cluster and increase variation between clusters [4].

The K-Means algorithm uses clustering analysis with non-hierarchical clusters. This method begins by determining the desired cluster, which can consist of two, three, or more to form into a cluster.

E. Decision Tree (CART)

Decision tree is an algorithm used to predict a class. this method works by dividing data into smaller sized groups. And the Decision Tree algorithm can process attributes that are discrete or numeric [5].

Classification and regression tree (CART) is a method included in data exploration techniques, namely decision tree techniques. CART aims to obtain an accurate group of data as a characteristic of a classification, besides that CART is used to explain the relationship between response variables and one or more predictor variables [6].

CART is a nonparametric statistical method used for classification and regression. This method produces a decision tree that explains the relationship between the response variable (dependent) and one or more predictor variables (independent). In the context of classification, CART creates a classification tree if the response variable has a categorical scale, and produces a regression tree if the response variable is in the form of continuous data [7].

III. METHODOLOGY

A. Data Processing CRISP-DM

The CRISP-DM method is a systematic approach used in the data analysis process. The following is a design using the K-Means Clustering and Decision Tree algorithms for customer loyalty grouping

1.Business Understanding Stage Analysis.

Objective: Grouping customers based on their loyalty to the company to help more effective marketing strategies.

Expected output:

a. Customer segmentation into groups (Loyal (1), Semi-Loyal (2) and Non-Loyal (3)).
b. Decision rules from the decision tree model to understand the attributes that affect loyalty.

2.Data Understanding Stage Analysis.

Data source: The data used comes from the Inventory System database that records outgoing goods transactions. This data reflects customer purchase history which is the basis for analyzing customer loyalty.Data Processing (CRISP-DM)

Outgoing goods data consists of various attributes, including:

a. Outgoing Goods ID.
b. Outgoing Date.
c. Purchasing Order Number.
d. Item Code.
e. Name of Goods.
f. Brand.
g. Size.
h. Price per unit.
i. Purchase Quantity.
j. Total purchase price.
k. Customer Code.

Attributes to be used:

a. Customer: Customer code or name as unique identification.
b. Quantity of goods purchased : Total units of goods that have been purchased by the customer in a certain period.
c. Total price of goods purchased: Total value of customer transactions in a certain period.

3.Analyze the Data Preparation stage.

Sort data: At the time of sorting to get the customer code, number of items, total price of goods purchased. To get the dataset according to the attributes that will be used the following SQL query 1.

Query 1.

```
SELECT
    kode_customer,
    SUM(qty_keluar) AS total_qty,
    SUM(total_harga) AS total_harga
  FROM
    tbl_bout
  WHERE
    tanggal_out BETWEEN (
      SELECT
DATE_SUB(LAST_DAY(MAX(tanggal_out)),
INTERVAL 3 MONTH)
      FROM tbl_bout
    ) AND (
      SELECT LAST_DAY(MAX(tanggal_out))
      FROM tbl_bout
    )
  GROUP BY
    kode_customer;
```

Addition of loyalty labels:

Supplement the data with loyalty labeling based on the evaluation of the staff at the company.

Loyalty labels:

a. Loyal (1): Customers who make frequent purchases with high transaction values.
b. Semi-Loyal (2): Customers who make purchases with medium transaction values.
c. Non-Loyal (3): Customers who make purchases with a small transaction value.

Split data for the model (Split Data):

Split Data will be conducted with 4 trials, including:

a. 60:40 : 60% Training Data and 40% Test Data.
b. 70:30 : 70% Training Data and 30% Test Data.
c. 80:20 : 80% Training Data and 20% Test Data.
d. 90:10 : 90% Training Data and 10% Test Data.

4.Modeling stage analysis.

K-means Clustering: Grouping customers based on the number of total purchases to understand customer loyalty to the company.

Steps:

a. Apply the K-Means Clustering algorithm to group customers based on the number of items purchased and the total price of the items purchased.
b. Analyze the clustering results by looking at the characteristics of each cluster.
c. Comparing clustering results with loyalty labels determined for validation using confusion matrix.

Decision Tree: Determine the decision rules that explain the main factors that affect customer loyalty.

Steps:

a. Compile a dataset based on the results of the Data Preparation stage.
b. Apply the Decision Tree algorithm to find the smallest gini index to become the root node.
c. Analyze the classification results in the form of a decision tree.
d. Compare the classification results with the specified loyalty label for validation using confusion matrix.

Combination of K-Means and Decision Tree: Classify customers using the K-Means and Decision Tree models where the results of K-Means will be continued with the Decision Tree model.

Steps:

a. Apply the K-Means Clustering algorithm to group customers based on the number of items purchased and the total price of the items purchased.
b. After getting the clustering results, then apply the Decision Tree algorithm to find the smallest gini index to become the root node.
c. Analyze the classification results in the form of a decision tree.
d. Comparing the results of the combination model with the specified loyalty label for validation using confusion matrix.

5.Analyze the business understanding stage.

Evaluation of K-Means Clustering:

a. Cluster visualization using scatter plot diagram.
b. Validation of cluster results with Confusion Matrix.

Decision Tree Evaluation:

a. Validation of classification results using Confusion Matrix.
b. Interpret the decision rules of the decision tree and evaluate the influential attributes.

6.Analyze the Deploy stage.

Implementation Strategy:

a. Integrate the segmentation results into the analysis system.
b. Use the results of k-means and decision tree to prioritize customer interventions.

Loyalty-based Marketing Strategy:

a. Loyal customers (1): Rewarded with discount coupons of Rp.150,000 per unit and free shipping.
b. Semi-Loyal Customers (2): Rewarded with a discount coupon of Rp.75,000 per unit and 50% off shipping costs.
c. Non-Loyal Customers (3): Rewarded with a 50% discount on shipping costs.

Maintenance Monitoring:

a. Periodically retraining the model to ensure the results remain relevant.

B. CART Decision Tree

The following are the stages of the CART Decision Tree. [8]

1. Calculate the Gini Index for each subset.

Calculate the gini index in each subset with the following formulation.

$$Gini(D) = 1 - \sum_{i=1}^{m} P_i^2 \quad (1)$$

Description:

G(D) = Gini Index.

k = Number of classes.

P_i^2 = Probability of a data in dataset D belonging to class i against impurity.

2. Calculate the Gini Index for each attribute.

After getting the probability of each partition of each attribute, then calculate the gini index for each attribute.

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \quad (2)$$

Description:

$\frac{|D_1|}{|D|}$ = Proportion of subset size D1 to total D.

$\frac{|D_2|}{|D|}$ = Proportion of subset size D2 to total D.

$G(D_1)\ dan\ G(D_2)$ = Gini index of each subset after splitting.

$Gini_A(D)$ = Gini index of an attribute.

3. Selecting the Root Node

After knowing the results of the gini index of each attribute, select the feature with the lowest result to be used as the root node. Then the process is repeated for each sub-dataset resulting from the splitting, until it reaches the leaf node (all samples in the node have the same label or there are no features left to split).

C. K-Means Clustering

To perform the K-Means method in general is done using the basic algorithm as follows. [9]

1. Determine the number of clusters (k) in the dataset.
2. Determine the center value (centroid). To find the center value at the beginning of the stage, the method is to find a value randomly, the formula determines the initial target of k-means, this formula is used to find the target value of data or distance between clusters, which serves as the initial center point in the calculation of the k-means algorithm iteration 0, such as the following formulation.

$$\frac{Total\ data}{Number\ of\ classes + 1} \quad (3)$$

Description:

Total data = The total amount of data used.

Number of classes = Number of predetermined groups such as high and low.

The average formula is used to find the iteration value. The calculation in the formula aims to find the average value as stated in the following equation.

$$V_{ij} = \frac{1}{N_i} \sum_{K=0}^{Ni} Xkj \quad (4)$$

Description:

Vij = The average centroid of the i-th cluster for the j-th variable.

ni = Number of members of the i-th cluster.

i,k = Index of the cluster.

j = index of the variable.

Xkj = The kth data value of the j-th variable for the cluster.

3. For each record, calculate the closest distance to the centroid. The formula used is euclidean distance, as shown in the following equation.

$$De = \sqrt{(xi - si)^2 + (yi - ti)^2} \quad (5)$$

Description:

De = Euclidean Distance.

i = Number of objects.

(x,y) = Object coordinates.

(s,t) = Centroid coordinates.

4. Group objects based on the distance to the closest centroid.

Repeatedly iterate from step 2 until the data remains in the same cluster.

D. Confussion Matrix

[10] Confusion matrix is an evaluation method used to measure the performance of Machine Learning algorithms. This method works by comparing the predictions made by the algorithm with the actual data that is known beforehand. It provides a clear visualization of the extent to which the algorithm can correctly predict each class or label in the dataset, and helps identify various evaluation metrics such as accuracy.

Table 1. Confusion Matrix Explanation

	Predicted Class		
Actual Class		True	False
	True	TP	FP
	False	FN	TN

Table 1 is a metric used as an evaluation tool in classifying an algorithm. The metric consists of four main elements, including True Positive (TP), True Negative (TN), False Negative (FP), False Negative (FN). The following is an explanation of the four elements:

TP = Number of positive values that are correctly predicted as positive.

FP = Number of negative values that are incorrectly predicted as positive.

TN = Number of data with negative values that are correctly predicted as negative values.

FN = number of positive values that are incorrectly predicted as negative values.

The following is a formulation to compare the predictions generated with the actual facts.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

IV. RESULTS AND DISCUSSION

The results and discussion section will discuss the results of the Decision Tree method, K-Means, and a combination of the two methods. The following is a dataset that has been labeled with loyalty obtained from related company staff.

Table 2. Labeled Dataset

No	Customer Code	Total Qty	Total Price	label
1	CUST348	90	315000000	1
2	CUST738	8	68000000	3
3	CUST323	85	272000000	1
4	CUST303	11	37700000	3
….	…….	……..	…….	…….
33	CUST259	60	210000000	2
34	CUST685	8	25600000	3
35	CUST238	60	192000000	2
36	CUST108	8	29600000	3

In Table 2, the dataset comes from outgoing goods that have been labeled within 3 months, namely September, August, July 2024 to predict the loyalty label for each customer.

A. Decision Tree Process

The first step is to separate the dataset into training data and test data with a ratio of 60% training data and 40% test data. The following are the results of the split data.

Table 3. Training Data (Decision Tree)

No	Customer Code	Total Qty	Total Price	label
1	CUST748	18	63200000	3
2	CUST738	8	68000000	3
3	CUST454	10	32000000	3
4	CUST682	10	48000000	3
….	…….	……..	…….	…….
18	CUST310	8	68000000	3
19	CUST303	11	37700000	3
20	CUST271	18	63000000	3
21	CUST238	60	192000000	2

Table 3 is data from split data results intended for training data with a ratio of 60%.

Table 4. Test Data (Decision Tree)

No	Kode Customer	Total Qty	Total Harga	label
1	CUST396	85	322500000	1
2	CUST258	16	56000000	3
3	CUST166	17	160000000	2
4	CUST240	66	211200000	1
….	…….	……..	…….	…….
12	CUST348	90	315000000	1
13	CUST546	2	17000000	3
14	CUST108	8	29600000	3
15	CUST259	60	210000000	2

Table 4 presents the data from the split results allocated for testing data with a 40% ratio. Below is the categorization of each attribute for the predictor variables derived from the training data.

Table 5. Category for each Attribute

Atribute	Category
Total qty	Total qty is divided into two categories: 1.<=68 2.>68
Total Price	Total price is divided into two categories: 1.<=141800000 2.>141800000

In Table 5 are the categories for each attribute to be used in Decision Tree modeling. Next is calculating the Gini Index for each subset.

Table 6. Proportion of Each Subset (Decision Tree)

Attribute	Split point	Proportion (1)	Proportion (2)	Proportion (3)	Total Proportion(|D|)
Total qty	<=68	0	3	16	19
	>68	2	0	0	2
Total Price	<=141800000	0	0	16	16
	>141800000	2	3	0	5
Total Proportion from Training Data		2	3	16	21

After obtaining the proportion of each subset, the next step is to calculate the Gini index. Below is the formula for calculating the Gini index using equation (1). Here is the calculation:

$$Gini(D) = 1 - \left(\frac{2}{21}\right)^2 + \left(\frac{3}{21}\right)^2 + \left(\frac{16}{21}\right)^2 = 0{,}390022676$$

$$Gini(\leq 68) = 1 - \left(\frac{0}{19}\right)^2 + \left(\frac{3}{19}\right)^2 + \left(\frac{16}{19}\right)^2 = 0{,}265927978$$

$$Gini(> 68) = 1 - \left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2 + \left(\frac{0}{2}\right)^2 = 0$$

$$Gini(\leq 141800000) = 1 - \left(\frac{0}{16}\right)^2 + \left(\frac{0}{16}\right)^2 + \left(\frac{16}{16}\right)^2 = 0$$

$$Gini(> 141800000) = 1 - \left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2 + \left(\frac{0}{5}\right)^2 = 0{,}48$$

After obtaining the Gini Index for each subset, the next step is to calculate the Gini Index for each attribute using formula (2). Here's the calculation:

$$G(\text{Total qty } 68) = \left(\frac{19}{21}\right) * 0{,}265927978 + \left(\frac{2}{21}\right) * 0 = 0{,}240601504$$

$$G(\text{Total price } 141800000) = \left(\frac{16}{21}\right) * 0 + \left(\frac{5}{21}\right) * 0{,}48 = 0{,}114285714$$

The result with the smallest Gini Index is selected as the root node because that attribute can better separate the data, resulting in more homogeneous partitions. Here's the Decision Tree.

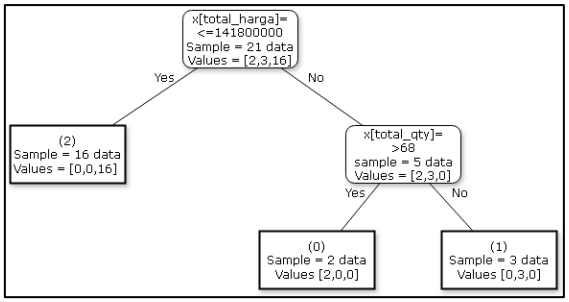

Figure 1. Decision Tree

Here is the description of Figure 1.

1.Root Node :
Condition: x[total_harga] <=141800000.
If true, the data belongs to class 3.
if false, the data is further split based on x[total_qty].

2.Right Branch:
Condition: x[total_qty] >68.
Decision: class 1.
Condition: x[total_qty] <=68.
Decision: class 2.

Final Decision:
Class 3: If x[total_harga] <= 141800000.
Class 1: If x[total_harga] > 141800000 and x[total_qty] >68.
Class 2: If x[total_harga] > 141800000 dan x[total_qty] <= 68.

After obtaining the decision tree results, the following are the predicted outcomes for the test data listed in table 6.

Table 7. Final Prediction for Decision Tree

No	Kode Customer	Total Qty	Total Harga	actual label	predict label
1	CUST396	85	322500000	1	1
2	CUST258	16	56000000	3	3
3	CUST166	17	160000000	2	2
4	CUST240	66	211200000	1	2
….	…….	……..	…….	…….	
12	CUST348	90	315000000	1	1
13	CUST546	2	17000000	3	3
14	CUST108	8	29600000	3	3
15	CUST259	60	210000000	2	2

After obtaining the prediction results from the model, the next step is to calculate the accuracy.

Table 8. Confusion Matrix Decision Tree

Confusion Matrix				
	Predict			
Actual		1	2	3
	1	3	1	0

	2	0	2	0
	3	0	0	9

After obtaining the confusion matrix table, the next step is to calculate the accuracy.
Accuracy = (3+2+9) / (3+1+0+0+2+0+0+0+9) = 0,93 x 100 = 93%.
The following are the results of the 70:30, 80:20, and 90:10 data splits.

Table 9. The average accuracy of the decision tree

Split Data	**Accuracy**
60:40	93%.
70:30	90,91%.
80:20	87,5%.
90:10	100%.
Average Accuracy	**92,8525%**

B. K-Means Process

In the initial stage of performing the K-Means model, the first step is to select the centroids or initial cluster centers.

Table 10. Centroids or Initial Cluster Centers

Centroid	Total qty	Total price
C1	90	315000000
C2	48	153600000
C3	12	38400000

After obtaining the initial cluster centers in Table 10, the next step is to calculate the distance to the centroid centers using the Euclidean distance formula. Below is the distance calculation from the first data point to the cluster center.

$$C1 = \sqrt{(90 - 90)^2 + (315000000 - 315000000)^2} = 0$$

$$C2 = \sqrt{(90 - 48)^2 + (315000000 - 153600000)^2} = 161400000$$

$$C3 = \sqrt{(90 - 12)^2 + (315000000 - 38400000)^2} = 276600000$$

After calculating the distance from the 1st data point to the cluster center, it is found that the result for C0 with a distance of 43000000 to the center is the closest. Therefore, the 1st data point belongs to cluster 2. Next, repeat step 2 for the 2nd data point and so on to determine the closest distance for each cluster. Below are the results of iteration 1.

Table 11. Iteration 1

Data i-th	**C1**	**C2**	**C3**	**Cluster**
1	0	161400000	276600000	1
2	247000000	85600000	29600000	3
3	43000000	118400000	233600000	1
4	277300000	115900000	700000	3
…..	……	……	…….	……
33	105000000	56400000	171600000	2
34	289400000	128000000	12800000	3
35	123000000	38400000	153600000	2
36	285400000	124000000	8800000	3

After obtaining the results from the first iteration, the next step is to calculate the average of the data points within the same cluster center to determine the new cluster centers.

Table 12. Averages for the new clusters

		Total	Number of C	Average
C1	x	336	4	84
	y	1175500000	4	293875000
C2	x	358	8	44,75
	y	1463300000	8	182912500
C3	x	294	24	12,25
	y	1185100000	24	49379166,67

Next, perform iterative calculations using the new centroid from each iteration until no data points move to another cluster. Below are the results of the k-means model compared with the actual labels from Table 1.

Table 13. Comparison of Clustering Results with Actual Labels

No	**Customer Code**	***Actual label***	***Cluster label***
1	CUST348	1	1
2	CUST738	3	3
3	CUST323	1	1
4	CUST303	3	3
…….	……..	…….	…….
33	CUST259	2	2
34	CUST685	3	3
35	CUST238	2	2
36	CUST108	3	3

Here is the calculation of the accuracy of the k-means model against the actual labels

Table 14. Confusion Matrix K-Means

Confusion Matrix				
	Predict			
Actual		1	2	3
	1	4	2	0
	2	0	5	0
	3	0	1	24

After obtaining the confusion matrix table, the next step is to formulate the accuracy.
Accuracy = (4+5+24) / (4+2+0+0+5+0+0+1+24) = 0.916 × 100 = 91.667%.

C. Combination Process

After obtaining the results from the K-Means algorithm in the previous step, the process continues with the Decision Tree model. In this phase, the clustering labels generated by the K-Means algorithm are used as new features or labels for prediction in the Decision Tree algorithm. This combination aims to evaluate the effectiveness of the data processing approach, where the resulting accuracy will later be compared to the performance of each algorithm individually. The following is the training data, using a 60% split ratio.

Table 15. Training Data For Combination Model

No	**Customer Code**	**total qty**	**Total price**	***Cluster label***
1	CUST748	18	63200000	3
2	CUST738	8	68000000	3
3	CUST454	10	32000000	3
4	CUST682	10	48000000	3
….	……	…	……..	…..
18	CUST310	8	68000000	3
19	CUST303	11	37700000	3
20	CUST271	18	63000000	3
21	CUST238	60	192000000	2

Here is the test data with a 40% ratio.

Table 16. Test Data For Combination Model

No	Customer Code	Total qty	Total Price	*Cluster label*
1	CUST396	85	322500000	1
2	CUST258	16	56000000	3
3	CUST166	17	160000000	2
4	CUST240	66	211200000	2
…	……..	….	…….	….
12	CUST348	90	315000000	1
13	CUST546	2	17000000	3
14	CUST108	8	29600000	3
15	CUST259	60	210000000	2

After dividing the dataset into training data and test data, the following are the categories of the "total qty" and "total price" attributes derived from the training data in Table 15. Below are the categories for each attribute.

Table 17. Categories of Each Attribute (Combination)

Atribute	Category
Total qty	Total qty is divided into two categories: 1.<=68 2.>68
Total price	Total price is divided into two categories: 1.<=108200000 2.>108200000

Next is calculating the Gini Index for each subset.

Table 18. Proportion of Each Subset (Combination)

Attribute	**Split point**	**Proportion (0)**	**Proportion (1)**	**Proportion (2)**	**Total Proportion (|D|)**
Total qty	<=68	0	4	15	19
	>68	2	0	0	2
Total harga	<=108200000	0	0	15	15
	>108200000	2	4	0	6
Total Proportion from Training Data		2	4	15	21

After obtaining the proportion of each subset, the next step is to calculate the Gini index using formula (1). The calculation is as follows:

$$Gini(D) = 1 - \left(\frac{2}{21}\right)^2 + \left(\frac{4}{21}\right)^2 + \left(\frac{15}{21}\right)^2 = 0{,}444444444$$

$$Gini(\leq 68) = 1 - \left(\frac{0}{19}\right)^2 + \left(\frac{4}{19}\right)^2 + \left(\frac{15}{19}\right)^2 = 0{,}332409972$$

$$Gini(> 68) = 1 - \left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2 + \left(\frac{0}{2}\right)^2 = 0$$

$$Gini(\leq 108200000) = 1 - \left(\frac{0}{15}\right)^2 + \left(\frac{0}{15}\right)^2 + \left(\frac{15}{15}\right)^2 = 0$$

$$Gini(> 108200000) = 1 - \left(\frac{2}{6}\right)^2 + \left(\frac{4}{6}\right)^2 + \left(\frac{0}{6}\right)^2 = 0{,}444444444$$

After obtaining the Gini Index for each subset, the next step is to calculate the Gini Index for each attribute using formula (2). Application:

$$G(Total\ qty\ 68) = \left(\frac{19}{21}\right) * 0{,}332409972 + \left(\frac{2}{21}\right) * 0 = 0{,}30075188$$

$$G(Total\ price\ 108200000) = \left(\frac{16}{21}\right) * 0 + \left(\frac{5}{21}\right) * 0{,}444444 = 0{,}126984127$$

The smallest Gini Index calculation result is selected as the root node because that attribute can better separate the data, resulting in more homogeneous partitions. Here is the decision tree.

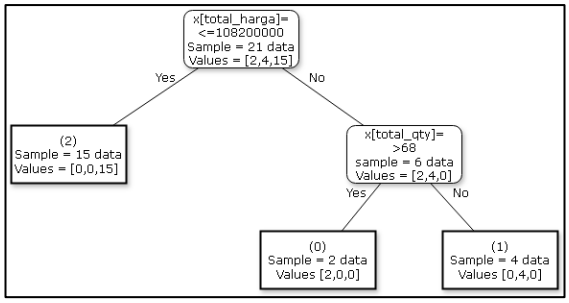

Figure 2. Decision tree (Combined)

The following is the description of Figure 2:

1.Root Node :

Condition: x[total_harga] <=108200000.

If true, the data belongs to class 3.

if false, the data is further split based on x[total_qty].

2.Right Branch:

Condition: x[total_qty] >68.

Decision: class 1.

Condition: x[total_qty] <=68.

Decision: class 2.

Final Decision:

Class 3: If x[total_harga] <= 108200000.

Class 1: If x[total_harga] > 108200000 and x[total_qty] >68.

Class 2: If x[total_harga] > 108200000 dan x[total_qty] <= 68.

After obtaining the results from the decision tree, the following are the prediction outcomes from the test data as shown in Table 19.

Table 19. The Result of Combined Algorithm

No	Customer Code	Total Qty	Total Harga	actual label	predict label
1	CUST396	85	322500000	1	1
2	CUST258	16	56000000	3	3
3	CUST166	17	160000000	2	2
4	CUST240	66	211200000	2	2
….	…….	……..	……..	…….	
12	CUST348	90	315000000	1	1
13	CUST546	2	17000000	3	3
14	CUST108	8	29600000	3	3
15	CUST259	60	210000000	2	2

After obtaining the prediction results from the model, the next step is to calculate the accuracy of the overall results from the combined algorithm.

Table 20. Overall Results of the K-Means and Decision Tree Model

No	Customer Code	Total Qty	Total Harga	actual label	predict label
1	CUST748	18	63200000	3	3
2	CUST738	8	68000000	3	3
3	CUST454	10	32000000	3	3
4	CUST682	10	48000000	3	3
….	…….	……..	……..	…….	
33	CUST348	90	315000000	1	1
34	CUST546	2	17000000	3	3
35	CUST108	8	29600000	3	3
36	CUST259	60	210000000	2	2

Table 21. Confusion Matrix Decision Tree

Confusion Matrix				
	Predict			
Actual		1	2	3
	1	2	0	0
	2	1	3	0
	3	0	0	9

After obtaining the confusion matrix table, the next step is to formulate the accuracy.

Accuracy = (5+5+24) / (5+1+0+0+5+0+0+1+24) = 0.944 × 100 = 94.4%

The following are the results of the 70:30, 80:20, and 90:10 data splits.

Table 22. Average Accuracy Combination

Split Data	**Accuracy**
60:40	94,4%
70:30	94,4%
80:20	94,4%
90:10	91,67%
Average	**93,7175%**

After obtaining the accuracy results from the three experiments Decision Tree, K-Means, and the combination of both algorithms the following is a comparison of the accuracy scores from each method.

Table 23. Average accuracy results of the three methods

Methods	**Accuracy**
Decision Tree	92,8525%
K-Means Clustering	91,667%
Combination Methods (*Decision Tree* dan *K-Means*)	93,7175%

Based on Table 23, it can be concluded that although the data does not include customer loyalty labels, the combination of the Decision Tree and K-Means Clustering algorithms achieved the highest

accuracy average at 93.7175%, compared to using each algorithm individually.

After completing all the manual calculations, the researcher proceeded to implement the algorithms in a web-based application. This implementation utilized the Python framework Flask, along with native PHP as the core programming language for the website.

V. CONCLUSION

Based on the analysis conducted in the study, the following conclusions can be drawn:

1. Based on the Accuracy calculation results, it can be concluded that the combination of K-Means and Decision Tree provides a higher average accuracy rate (93.7175%) compared to the average accuracy of Decision Tree alone (92.8525%) and K-Means alone (91.667%). Therefore, overall, the combination of K-Means and Decision Tree is more effective in categorizing customer loyalty at PT. Tangguh Buana Roda Indonesia. By using the K-Means algorithm and the combination of K-Means and Decision Tree, researchers can obtain more accurate customer loyalty clustering results that closely reflect the true values. The combined K-Means and Decision Tree algorithm is recommended for use in supporting customer loyalty rewards to enhance customer loyalty to the company.
2. Based on the output of the K-Means and Decision Tree algorithms, there are 5 loyal customers (1) who are eligible for a discount coupon of Rp. 150,000 per unit and free shipping. There are 7 semi-loyal customers (2) who are eligible for a discount coupon of Rp. 75,000 per unit and a 50% shipping discount. Meanwhile, 24 non-loyal customers (3) are eligible for a 50% shipping discount.
3. The implementation of the Decision Tree and K-Means Clustering algorithms offers an effective and efficient solution for determining customer loyalty rewards. With accurate modeling, the company can develop better future planning for customer engagement. This system combines the power of data clustering analysis with the ease of a web-based interface, providing a robust tool for users in awarding customer incentives.
4. Data limitations posed a major challenge in this study, as the research site is a loader tire retailer with a relatively low purchase frequency (approximately once every 4 to 8 months). This condition makes the transaction data insufficient for adding features such as purchase frequency. The limited amount of data also increases the risk of overfitting. Therefore, future research on customer segmentation is advised to use a larger dataset to obtain more accurate results.

REFFERENCES

[1] A. Prasetya and R. Utari, "Analisis Customer Relationship Management (Crm) Terhadap Loyalitas Pelanggan Dengan Kepuasan Pelanggan Sebagai Variabel Intervening Pada Pelanggan Cv. Cipta Adhi Nugraha Creative," *J. Ekon. Bisnis dan Akunt.*, vol. 2, no. 2, pp. 88–98, 2022, doi: 10.55606/jebaku.v2i2.547.

[2] E. Gunia, A. Irma Purnamasari, and I. Ali, "Penerapan Datamining Dalam Menentukan Pola Penjualan Produk Menggunakan Algoritma Fp-Growth," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 2, pp. 2417–2422, 2024, doi: 10.36040/jati.v8i2.9506.

[3] M. A. Hasanah, S. Soim, and A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," *J. Appl. Informatics Comput.*, vol. 5, no. 2, pp. 103–108, 2021, doi: 10.30871/jaic.v5i2.3200.

[4] I. S. Hidayat, S. Defit, and G. W. Nurcahyo, "Simulasi dalam Optimalisasi Pengadaan Barang menggunakan Metode K-Mean Clustering," *J. Sistim Inf. dan Teknol.*, vol. 3, pp. 281–286, 2021, doi: 10.37034/jsisfotek.v3i4.79.

[5] S. B. Bulkisah, R. Astuti, and A. Bahtiar, "Implementasi Data Mining Algoritma Decision Tree Untuk Klasifikasi Status Gizi Balita Di Kecamatan Ciledug," *J. Ilm. Inform. Komput.*, vol. 29, no. 1, pp. 1–12, 2024, doi: 10.35760/ik.2024.v29i1.10346.

[6] N. Erliani, K. Suryowati, and M. T. Jatipaningrum, "Klasifikasi Tingkat Penjualan Laptop Di E-Commerce Menggunakan Algoritma Classification and Regression Tree (Cart)," *J. Stat. Ind. dan Komputasi*, vol. 8, no. 2, pp. 40–47, 2023, doi: 10.34151/statistika.v8i2.4455.

[7] F. A. Oktavirahani and R. Maharesi, "Implementasi Algoritma Decision Tree Cart Untuk Merekomendasikan Ukuran Baju," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 1, p. 138, 2022, doi: 10.30865/jurikom.v9i1.3838.

[8] R. Irmanita, Sri Suryani Prasetiyowati, and Yuliant Sibaroni, "Classification of Malaria Complication Using CART (Classification and Regression Tree) and Naïve Bayes," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 10–16, 2021, doi: 10.29207/resti.v5i1.2770.

[9] A. T. Sipayung., Saifullah, and R. Winanjaya, "Penerapan Metode K-Means Dalam Mengelompokkan Banyaknya Desa/ Kelurahan Menurut Keberadaan Permukiman Di Bantaran Sungai Berdasarkan Provinsi," *Brahmana J. Penerapan Kecerdasan Buatan*, vol. 2, no. 1, pp. 49–56, 2020, doi: 10.30645/brahmana.v2i1.48.

[10] R. Antika, A. Rifa'I, F. Dikananda, D. Indriya Efendi, and R. Narasati, "Penerapan Algoritma Decision Tree Berbasis Pohon Keputusan Dalam Klasifikasi Penyakit Jantung," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 6, pp. 3688–3692, 2024, doi: 10.36040/jati.v7i6.8264.