Journal of Information Systems and Informatics

Vol. 6, No. 1, March 2024 e-ISSN: **2656-4882** p-ISSN: **2656-5935**

DOI: 10.51519/journalisi.v6i1.666 Published By DRPM-UBD

Dynamic Segmentation Analysis for Expedition Services: Integrating K-Means and Decision Tree

Dwi Himatul Khoiriyah[1], Rita Ambarwati Sukmono[2*]

[1,2]Departement of Management, Muhammadiyah Sidoarjo University, Sidoarjo, Indonesia
Email: [1]dwihima95@gmail.com, [2*]ritaambarwati@umsida.ac.id

Abstract

Technological developments have an impact on increasing the level of competition between companies in acquiring and retaining customers. With this competition, companies must maximise efforts to reach consumers and understand customer service needs so that the business can continue to survive and experience development. In this effort, a segmentation analysis was carried out on marketplace accounts and expedition services commonly used by consumers to make transactions. The first step is to correct the dataset obtained to avoid errors in the final results. Next, data processing was done using rapidminer with the k-means clustering and decision tree methods. The research results show that k-means clustering achieved the lowest Davies Bouldin Index (DBI) accuracy, namely -0.943 in cluster_8. In the results of research using the decision tree method, accuracy results were obtained at 49.83%. The results obtained with the decision tree method cannot be said to be good because the results are below the 50% value; however, the decision tree method shows that a good cluster is cluster_7. In this case, better accuracy values can be achieved by using the k-means clustering method. This research can illustrate the importance of utilizing the k-means and decision tree algorithms in classifying sales data as a tool for optimizing marketing and service efforts.

Keywords: K-Means Clustering, Decision Tree, Expedition Services, Marketplace

1. INTRODUCTION

Technological developments affect retail and logistics operations, where consumers can now purchase transactions only through online media. Companies that can effectively meet consumer needs through online media will make consumers willing to be loyal and continue to interact with the company [1]. These technological developments are in line with increasing competition in the business world. Competition means that companies must be able to maximize their capabilities as best as possible in order to continue to survive in the face of competition [2]. One way to maximize business capabilities is to take advantage of online media such as marketplaces. This marketing activity makes it easier to reach consumers and prov. It providesities to maximize profits because it can reduce marketing costs so that business people can expand the market and provide easy access and service to consumers [3], [4]. Online sales platforms can optimize sales efforts by easily distributing promotions through electronic systems or a series of computer networks, so that consumers can carry out transactions simply via online media quickly [5].

Optimal customer service will produce company output in the form of profits, which is the main goal of business activities. To maintain business and expand it, companies need to undertake a series of customer-focused efforts. In this case the company can provide the best service to provide a good transaction experience with the company. A bad transaction experience will greatly influence customer loyalty, the causes of a bad transaction experience include the company's inability to provide solutions to problems experienced by customers, the company's inability to identify problems often experienced by customers which can pose a risk of losing customers owned by the company [2].

Companies can take advantage of the ease of marketing today to expand consumer reach and continue to strive to retain customers so they can continue to survive business competition. Through the marketplace, companies can gain benefits in the form of flexibility and operational efficiency, so they can provide easy interaction services in the form of questions and answers with consumers and make it easier for consumers to get access to information related to existing products and services because marketing via online media is not complicated [6], [7], [8].

Retail companies generally carry out two marketing activities, namely offline and online. Therefore, it is necessary to optimize the company's marketing activities in the form of mapping groups of transactions carried out by customers based on the frequency, ratio and diversity contained in them. In this case, managing sales media within the company will have a big impact on the success and sustainability of the business. To be able to optimize this, this research uses a data mining method in the form of grouping marketplace accounts and expedition services using the k-means clustering method, as well as using the decision tree method to simply describe the results of the grouping that has been obtained.

Clustering is able to divide data based on the characteristics of the data itself, then combine it into similar clusters and separate it into different clusters [9]. In other words, clustering is used to group data based on similarities between each other and those that do not have similarities with other cluster objects, in order to be able to divide information from existing datasets based on their similarities [10], [11]. The main aim of the k-means method is to obtain the most stable number of clusters or groups based on the lowest value of the Davies Bouldin Index (DBI) [11], [12].

Decision trees are one of the tools used in this research. Where decision trees can produce decisions that are deemed relevant to the research objectives. We use the decision tree method to get final information from the existing cluster results. Decision trees are modelling with a simple appearance and can provide good results for managing big data or large amounts of data. This modeling will produce output in the form of instance classification based on data probability values [13]. In a previous study conducted by Indivar, et al [11] regarding grouping similar objects into clusters, it weans found that k-means succeeded in forming clusters from e-commerce bid data. According to Yi Lei and Xiandong [13], the decision tree method is used to evaluate cross-border e-commerce, and the prediction accuracy rate is above 95%. The k-means and decision tree methods are used in this research to avoid random errors and obtain optimal results in accordance with the research objectives, namely segmentation to optimize marketplace accounts and can have an impact on the efficiency and effectiveness of the performance of marketing employees in retail companies.

2. METHODS

This research uses secondary data from sales records of medical equipment retail companies, which will then be processed using modeling contained in data mining, namely k-means and decision trees. Data mining is a process carried out to obtain useful information from large and complex data including algorithmic techniques, statistics, and functions to find hidden patterns in data by analyzing the data and then concluding it into useful information [14][15]. According to Han and Kamber, data mining is divided into two main categories, namely [15]:

a. Predictive category, which is useful as a tool for predicting the value of a particular attribute based on the values of other attributes. The predicted attribute is the target or dependent variable, and other attributes used to make predictions are independent or explanatory variables.
b. Descriptive categories aim to determine patterns which usually consist of correlations, trends, territories, anomalies and clusters. In this category, descriptive data mining often requires post-preprocessing techniques for validation or testing accuracy and explanation of results.

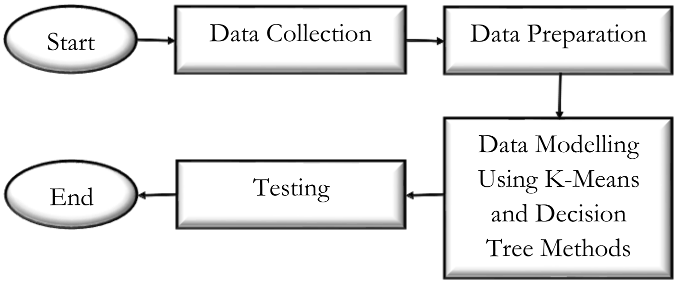


Figure 1. Stages of Research

In this research, the data mining used is descriptive categories. Where in this category, this research is included in the cluster formation pattern. The research stages are presented in the flow of research stages as in Figure 1.

2.1 Data Collection

The processed dataset is a dataset presented in the form of an Excel report. We use data obtained from sales records of medical equipment retail companies during the period from early 2020 to early 2023, namely from January to June 9. The total dataset processed is 15,274 data consisting of 15 attributes including date/working day, number, customer name, telephone number, city of origin, transaction origin, item name, quantity, dropship, expedition service, item price, shipping costs, insurance , additional fees, and total price.

2.2 Data Preparation

The existing dataset is prepared so that it can then be entered into machine learning. The data preparation in question is carrying out checks regarding the presence of data that does not match the writing and spelling, as well as the presence of missing values, which will make the learning machine not optimal in producing an output. The first thing to do is data cleaning to ensure that there is no data that is inconsistent in writing and spelling, data transformation to change the form of the data according to the input desired by the learning machine but still not changing the value of the existing data, and carrying out data selection to select attributes. which will be researched in machine learning.

At the data transformation stage, the initial data attributes are changed into a form that is more uniform and in accordance with the input requirements of the Rapidminer learning machine. At this stage, the city origin attribute is changed to province manually using Microsoft Excel. After the data is entered into the learning machine, the data is then transformed into numerical form. The attributes entered into the learning machine are the characteristics of the province, the origin of the transaction, and the expedition service, as shown in Table 1 and Table 2.

Table 1. Data set Attribute

No	Province	Origin Of The Transaction	Expedition Services
1	Jawa Tengah	Shopee Mastha Medika Surabaya	JNE Reguler
2	Sulawesi Selatan	Bukalapak Mastha Surabaya	J&T Reguler
3	Surabaya	Shopee Mastha Medika Surabaya	J&T Express
4	Surabaya	Shopee Mastha Medica	JNE Reguler

No	Province	Origin Of The Transaction	Expedition Services
5	Jawa Barat	Tokopedia Mastha Medica Jakarta	Wahana
....			
15270	Surabaya	Tokopedia Fanmed sby	JNE Reguler
15271	Surabaya	Tokopedia Fanmed sby	SiCepat Reguler
15272	Jambi	Tokopedia Fanmed sby	SiCepat Reguler
15273	Banten	Tokopedia Fanmed (Jakarta)	JNE Reguler
15274	Kalimantan Barat	Tokopedia Ama Medica Sby	JNE Reguler

Table 2. Attribute Description

Attribute	Explanation
Provinces	Information on the region of origin of consumers who purchase products
Origin of the Transactions	Information on the origin of the transaction account chosen by the consumer to purchase products
Expedition Services	Information on the delivery service or distributor chosen by the consumer to deliver product purchases

2.3 Data modeling

The data modelling was carried out using two data mining modelling methods, namely k-means clustering and decision trees. In this modeling we use the rapidminer application because the output results can be easily displayed and easy to understand.

1) K-Means

K-means is a learning technique used to group data sets into separate groups so that each group can be divided according to similar characteristics [16]. To be able to group with accurate results, experiments on varying the number of clusters need to be carried out. In this case, k=10 experiments were carried out to see which cluster had the best Davies Bouldin Index (DBI) value, namely the one with the lowest value compared to the other cluster values.

Cluster results using the K-Means method are greatly influenced by the k value or number of k specified [17]. Experiment as many as k=10, namely by taking samples of values k=2 up to a value of k=10 to see which cluster has the lowest value compared to the values of other clusters.

This algorithm is rated as one of the most powerful and popular cluster modeling [16]. The disadvantage of k-means is that it can only accept input in the form of numeric data, and is sensitive to the selection of the centroid starting point [11][14]. For this reason, it is necessary to transform the data and experiment by varying the number of k to get the best value. To validate the method, it can be done by measuring clustering and iteration in the following way.

a) Choose the number of clusters K
b) Selecting the number of clusters can generally be done in various ways, but what is often done is the random method, namely by entering the initial value with random numbers
c) Allocate all data/objects to the nearest cluster

In the process of allocating the proximity of objects to other objects, the distance between each data and each cluster center is calculated. The resulting distance determines whether data will fall into a certain cluster. Calculation of the data distance to each cluster point center can be done using Euclidean distance theory with Equation 1 and 2 [18].

$$Vij= \frac{1}{Ni} \sum_{k=0}^{Ni} Xkj \tag{1}$$

$$D=\sqrt{(Xi-Si)^2+(yi-ti)^2} \tag{2}$$

2) Decision Tree

Decision trees are one of the algorithms used for decision making. Decision trees change criteria into nodes that are connected to each other to form a tree-like structure. Each tree has branches that represent attributes that must be met to reach the next branch until there are no more branches in the decision tree structure that is formed. The decision tree algorithm processes decisions by reformatting tabular data into a tree model that produces simplified rules. This algorithm represents a simple classification method for many classes by marking with attribute names the internal nodes and root nodes formed, labeling attribute values at the edges, and marking different classes at the leaf nodes [19].

A decision tree is a modeling used to classify data based on input data in the form of a probability distribution with a simple display that can provide good results for large amounts of data. This modeling produces output in the form of instance classification based on probability values from the data [13]. Decision trees were chosen as the second model because classification using decision trees is able to handle large amounts of data based on the training set and labels from the data [20]. Modeling using this method can simplify the description of the results of distance calculations and cluster determination which have been carried out using the k-means method.

2.4 Testing

The test carried out was to see the accuracy value of the data processing results using the k-means and decision tree methods. In this process, measuring the quality of cluster results is carried out using the cluster distance performance feature to obtain clusters that have the lowest/best Davies Bouldin Index (DBI) values. After getting the best cluster, modeling is carried out again using the decision tree method by testing the cluster data. We use classification performance measurements to obtain an accuracy value for the system's level of success in forming classifications of previously obtained input data.

3. RESULTS AND DISCUSSION

In this research, the dataset processed was 15,274 total data. The rapidminer application is used to model data using the k-means and decision tree methods. To carry out modeling, data preparation must be carried out on the dataset which will be processed through a series of preprocessing processes by correcting spelling, writing, eliminating missing values, and changing the form of the data into an acceptable form and in accordance with the modeling method and research objectives.

3.1 K-Means

The first modeling carried out is using the k-means method which will be varied by the number of clusters to get the lowest/best Davies Bouldin Index (DBI) value. At this stage, data transformation is carried out on the attributes from the city to province, and the entire data has been transformed into numerical form. The attribute chosen as the label is the expedition with the transaction origin id.

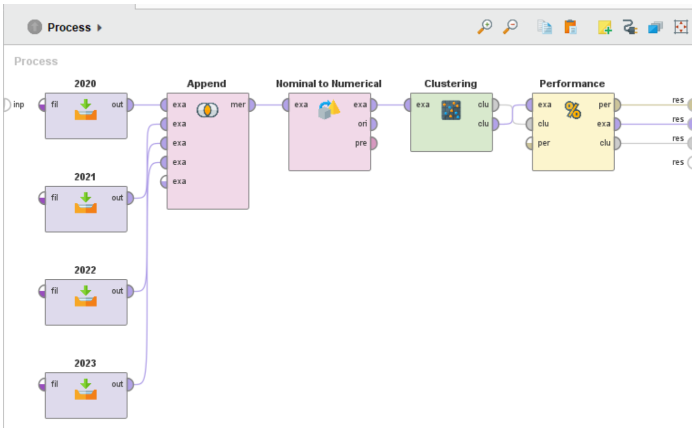


Figure 2. K-Means Method Modeling

In Figure 2, the data processing flow in Rapidminer is presented so that processing can be carried out using the k-means method. Separate datasets are placed in order of year to make it easier to identify data that is still experiencing defects. The data is then combined and transformed into numerical form to meet the data input requirements of the k-means method. Performance is used to determine the value of the Davies Bouldin Index (DBI), where we use the cluster distance performance assessment as a benchmark for the value of the cluster being tested.

Table 3. Hasil DBI Metode K-Means

Cluster	Davies Bouldin Index
cluster_2	-0,856
cluster_3	-0,842
cluster_4	-0,806
cluster_5	-0,778
cluster_6	-0.718
cluster_7	-0,888
cluster_8	-0,943
cluster_9	-0,939
cluster_10	-0,938

We conducted k=10 trials to find the lowest Davies Bouldin Index (DBI) value as the best cluster formed by the k-means method. Testing is carried out randomly from k=2 to k=10 as presented in Table 3. In this table, among the 9 clusters tested, cluster_8 shows the lowest value compared to the other clusters, namely -0.943, where the lowest number indicates that this cluster is the best cluster among other clusters that were also tested previously.

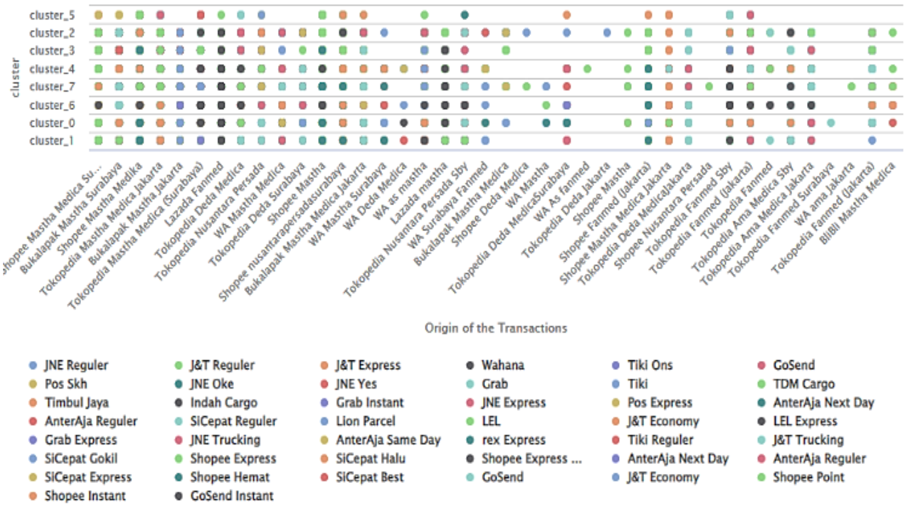
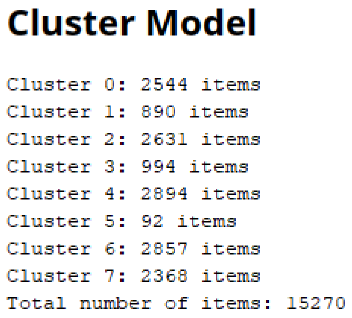


Figure 3. Cluster Result Visualisation

After finding the best Davies Bouldin Index value, we visualized the results of cluster_8 as presented in Figure 3. and found that there were clusters that did not contain several accounts and expedition services that consumers usually choose. Some clusters had almost all accounts and expedition services. Cluster_8 has 8 groups with different total items, number of accounts, and number of expedition services, as in Figure 4 and Table 4.



Cluster Model

```
Cluster 0: 2544 items
Cluster 1: 890 items
Cluster 2: 2631 items
Cluster 3: 994 items
Cluster 4: 2894 items
Cluster 5: 92 items
Cluster 6: 2857 items
Cluster 7: 2368 items
Total number of items: 15270
```

Figure 4. Total Items K-Means

Table 4. Results of Cluster Findings in K-Means

Cluster	Account	Expedition
cluster_0	34	30
cluster_1	30	23
cluster_2	34	26
cluster_3	26	24
cluster_4	33	30
cluster_5	18	16
cluster_6	32	31
cluster_7	35	30

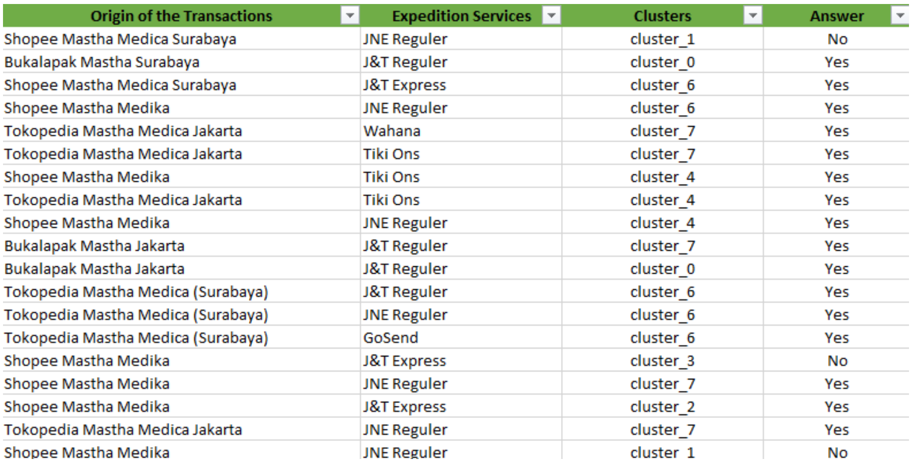
The results of data processing carried out using the k-means method found the following analysis results:

1) 1. Cluster_0, cluster_2, cluster_4, cluster_6, and cluster_7 are groups with more than 2000 total transactions with similar customer habits in selecting the account from which the transaction originates and the expedition service chosen for making purchases. Namely there are at least ≥ 30 total accounts from which the transaction originates and ≥ 25 selected expedition services. This group is a group containing general or majority things that consumers usually do. The more general it is, the more options to choose from, and the more ratios there are in this group, which becomes the basis for more supervision and control because there are so many concentrations that must be divided and regulated.
2) 2. Cluster_1, cluster_3, and cluster_5 are purchasing groups with a number of similar transactions < 1000. This group is a group of

consumers who have narrower purchasing habits compared to the first group, namely there are ≤30 accounts from which transactions originate and have no more than 25 delivery services provided by companies, and there are fewer transaction origin accounts from which consumers choose to make purchases. This group is a group that has a lower ratio in all aspects compared to the first group.

3.2 Decision Tree

The results of cluster formation by the k-means method are then exported into Excel and can then be modelled again using the decision tree method. This modelling was carried out with the same application, namely rapidminer, by adding a new attribute, namely answer, as in Figure 5. This is done to simplify the form of the cluster results that have been obtained.



Origin of the Transactions	Expedition Services	Clusters	Answer
Shopee Mastha Medica Surabaya	JNE Reguler	cluster_1	No
Bukalapak Mastha Surabaya	J&T Reguler	cluster_0	Yes
Shopee Mastha Medica Surabaya	J&T Express	cluster_6	Yes
Shopee Mastha Medika	JNE Reguler	cluster_6	Yes
Tokopedia Mastha Medica Jakarta	Wahana	cluster_7	Yes
Tokopedia Mastha Medica Jakarta	Tiki Ons	cluster_7	Yes
Shopee Mastha Medika	Tiki Ons	cluster_4	Yes
Tokopedia Mastha Medica Jakarta	Tiki Ons	cluster_4	Yes
Shopee Mastha Medika	JNE Reguler	cluster_4	Yes
Bukalapak Mastha Jakarta	J&T Reguler	cluster_7	Yes
Bukalapak Mastha Jakarta	J&T Reguler	cluster_0	Yes
Tokopedia Mastha Medica (Surabaya)	J&T Reguler	cluster_6	Yes
Tokopedia Mastha Medica (Surabaya)	JNE Reguler	cluster_6	Yes
Tokopedia Mastha Medica (Surabaya)	GoSend	cluster_6	Yes
Shopee Mastha Medika	J&T Express	cluster_3	No
Shopee Mastha Medika	JNE Reguler	cluster_7	Yes
Shopee Mastha Medika	J&T Express	cluster_2	Yes
Tokopedia Mastha Medica Jakarta	JNE Reguler	cluster_7	Yes
Shopee Mastha Medika	JNE Reguler	cluster_1	No

Figure 5. Cluster Answers

The answer attribute is created using the Microsoft Excel application with the vlookup formula, namely with the provision that clusters that require more supervision because they have a total of > 2000 items with a total of ≥ 30 have the answer Yes. Meanwhile, the No answer is intended for clusters that do not require lower supervision than clusters that have a Yes answer, where in that cluster, there are fewer total items, namely <1000, with a number of accounts ≤30. After adding the answer attribute using the vlookup formula in Microsoft Excel, the dataset is then entered into the rapidminer learning machine with the flow as in Figure 6.

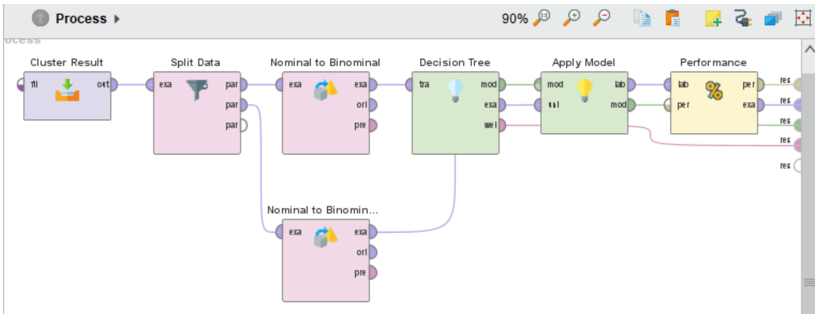
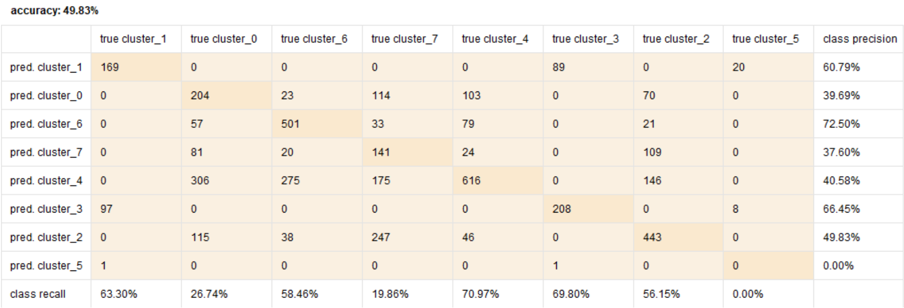


Figure 6. Decision Tree Method Modeling

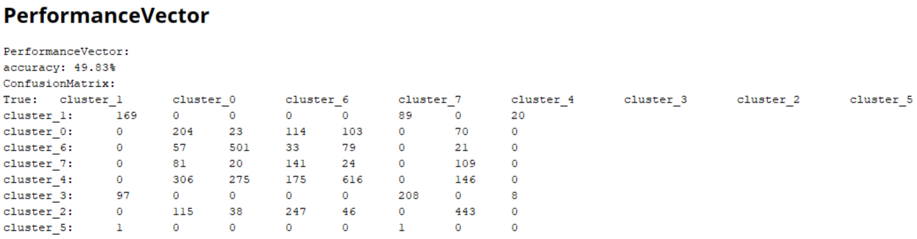
We set the cluster attribute as a label and use split data with a division ratio of 7:3, where 70% is training data, and 30% is testing data. The data is then transformed into a binominal form so that it can be accepted by the decision tree method feature without changing the values in the data. The decision tree used compares the information_gain criteria to get the highest feature selection accuracy results. All account types contained in the transaction origin attribute are included in the maximum depth, namely a total of 35. In this process, we do not use apply pruning and apply prepruning to get maximum results and have clearer and more comprehensive analysis results. The next thing to do is to test and measure the level of accuracy obtained from using the information_gain criteria in modelling cluster data results, as in Figure 7 and Figure 8.



accuracy: 49.83%

	true cluster_1	true cluster_0	true cluster_6	true cluster_7	true cluster_4	true cluster_3	true cluster_2	true cluster_5	class precision
pred. cluster_1	169	0	0	0	0	89	0	20	60.79%
pred. cluster_0	0	204	23	114	103	0	70	0	39.69%
pred. cluster_6	0	57	501	33	79	0	21	0	72.50%
pred. cluster_7	0	81	20	141	24	0	109	0	37.60%
pred. cluster_4	0	306	275	175	616	0	146	0	40.58%
pred. cluster_3	97	0	0	0	0	208	0	8	66.45%
pred. cluster_2	0	115	38	247	46	0	443	0	49.83%
pred. cluster_5	1	0	0	0	0	1	0	0	0.00%
class recall	63.30%	26.74%	58.46%	19.86%	70.97%	69.80%	56.15%	0.00%	

Figure 7. Decision tree accuracy view table



PerformanceVector

```
PerformanceVector:
accuracy: 49.83%
ConfusionMatrix:
True:   cluster_1       cluster_0       cluster_6       cluster_7       cluster_4       cluster_3       cluster_2       cluster_5
cluster_1:      169     0       0       0       0       89      0       20
cluster_0:      0       204     23      114     103     0       70      0
cluster_6:      0       57      501     33      79      0       21      0
cluster_7:      0       81      20      141     24      0       109     0
cluster_4:      0       306     275     175     616     0       146     0
cluster_3:      97      0       0       0       0       208     0       8
cluster_2:      0       115     38      247     46      0       443     0
cluster_5:      1       0       0       0       0       1       0       0
```

Figure 8. Confusion Matrix

Information_Gain is a criterion commonly used in classification using the decision tree method [21]. In this modelling using the decision tree method with information_gain criteria and a maximum depth of 35, the accuracy results were 49.83%; the confusion matrix compared the total True Positive (TP) of 2,282 and False Positive (FP) of 2,296 as shown in Figure 7 and Figure 8. After modelling using the decision tree method above, it is known that the accuracy value using the decision tree method is 49.83%. This value cannot be said to be good because the resulting value is less than 50%, and it cannot be said that the decision tree method is successful in forming classes if the data has many attributes and features. The accuracy results in decision tree modelling are greatly influenced by the depth of the tree, namely the longest distance measured from the roots to the leaves of the tree [22]. Apart from that, the amount of data, the high dimension of the data, and the consistency of the data in the dataset also greatly influence the accuracy of the results of data mining [23]. Based on the confusion matrix results in Figure 8. the only cluster that does not have a True Positive (TP) value and has a False Positive (FP) value is cluster_5.

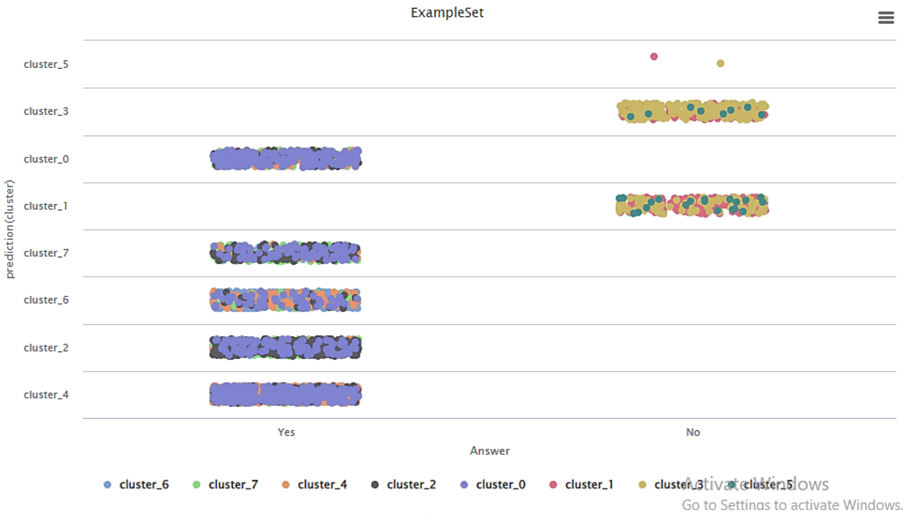


Figure 9. Scatter / Bubble visualisations

In Figure 9, the decision tree method shows that for answer No, it is concluded that there are only 2 clusters, namely cluster_3 with the highest ratio and followed by cluster_1, while the True Positive cluster_5 is divided into cluster_3 and cluster_1. In other clusters that show the answer Yes in them, there is cluster_4 with the highest ratio, followed by cluster_6, cluster_2, cluster_0, and cluster_7. Based on the results obtained using the decision tree method, there are only 7 clusters from the prediction results of the decision tree method from the cluster dataset by k-means. The total items with a True Positive (TP) value produced by the decision tree are briefly presented in Table 5.

Table 5. Total Items True Value (TP) Decision Tree

Cluster	Total Items True Positive (TP)
cluster_0	204
cluster_1	169
cluster_2	443
cluster_3	208
cluster_4	616
cluster_5	0
cluster_6	501
cluster_7	141

The data processing results using the decision tree method are presented in Figure 9 and Table 5. The analysis results were found as follows:

1) Clusters labeled Yes which include cluster_0, cluster_2, cluster_4, cluster_7, and cluster_6 have an even distribution on the same label. Where the yes label cluster has a high number of true positives, and their existence is spread among clusters with the same label.
2) Clusters labelled No, which include cluster_1 and cluster_3, have many true positives, and their presence is evenly distributed in clusters that both have the label No.
3) Cluster_5, which is the result of the k-means method, was removed or deleted because it did not have a positive trus value, and there were only two false positive distributions in the cluster. So, there are only 7 clusters that are considered successful.

4. CONCLUSION

Based on the results of the processing and trials, it was concluded that the k-means method produced the best group pattern, namely 8 clusters with a Davies Bouldin Index (DBI) value of -0.943. K-means succeeded in forming a group with the Yes label, namely, cluster_4, then cluster_6, cluster_2, cluster_0, cluster_7, as well as groups with the label No, which included cluster_3, cluster_1, and cluster_5. Meanwhile, analysis using the decision tree method shows that based on its accuracy value, the decision tree method cannot produce maximum output for managing big data, which has many varied features and attributes. This is proven by the low accuracy value of 49.83%. The decision tree shows the results of 7 clusters, namely the Yes label in sequence based on the number of items in cluster_4, cluster_2, cluster_6, cluster_7, and cluster_0, and the No label in cluster_3 and cluster_1. The Yes and No labels describe the need for supervision so that marketing operational activities can run smoothly. Based on the accuracy results of the two methods, it can be concluded that k-means has a higher accuracy level than the decision tree method, even though the results or output of both are almost in the same number of supervisory divisions.

REFERENCES

[1] L. T. T. Tran, “Managing the effectiveness of e-commerce platforms in a pandemic,” *J. Retail. Consum. Serv.*, vol. 58, no. September 2020, p. 102287, 2021, doi: 10.1016/j.jretconser.2020.102287.

[2] J. Teknologi, E. Febrianty, L. Awalina, and W. I. Rahayu, “Optimalisasi Strategi Pemasaran dengan Segmentasi Pelanggan Menggunakan Penerapan K-Means Clustering pada Transaksi Online Retail,” *Jurnal Teknologi dan Informasi,* vol. 13, no. September, pp. 122–137, 2023, doi: 10.34010/jati.v13i2.

[3] I. P. Artaya and T. Purworusmiardi, “Efektifitas Marketplace Dalam Meningkatkan Konsentrasi,” *Ekon. Dan Bisnis, Univ. Narotama Surabaya*, no. April, pp. 1–10, 2019, doi: 10.13140/RG.2.2.10157.95206.

[4] W. Novita Sari., Achmad Hizazi., “Effect of Good Corporate Governance and Leverage on Profitability-Mediated Tax Avoidance (Study on Mining Companies listed on the Indonesia Stock Exchange 2016 – 2019),” *Int. J. Acad. Res. Account. Financ. Manag. Sci.*, vol. 11, no. 2, pp. 202–221, 2021, doi: 10.6007/IJARAFMS.

[5] P. N. I. Sari, “Pengaruh Brand ambassador,kepercayaan dan resiko terhadap keputusan pembelian di e-commerce Shopee oleh mahasiswa di Pekanbaru,” *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2020.

[6] Ismai, “E-commerce dorong perekonomian Indonesia, selama pandemi covid 19 sebagai entrepreneur,” *J. Manaj. dan Bisnis Prodi Kewirausahaan*, vol. 2, no. 2, pp. 111–124, 2020.

[7] B. Algifari and A. Ariesta, “Penerapan E-Commerce Untuk Meningkatkan Penjualan Sepatu Pada Toko Garasi Spokat,” *Prosiding SISFOTEK*, vol. 4, no. 1, pp.99-105, 2020.

[8] V. No and Z. Kedah, “Startupreneur Bisnis Digital (SABDA) Use of E-Commerce in The World of Business,” vol. 2, no. 1, pp. 51–60, 2023.

[9] B. Zhang, L. Wang, and Y. Li, “Precision Marketing Method of E-Commerce Platform Based on Clustering Algorithm,” *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/5538677.

[10] F. A. Dewa and M. T. Jatipaningrum, “Segmentasi E-Commerce Dengan Cluster K-Means Dan Fuzzy C-Means (Studi Kasus : Media Sosial di Indonesia yang diunduh di Play Store),” *Jurnal Statistika Industri dan Komputasi*, vol. 4, no. 1, pp. 53–67, 2019.

[11] I. Shaik, S. S. Nittela, T. Hiwarkar, and S. Nalla, “K-means Clustering Algorithm Based on E-Commerce Big Data,” *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 11, pp.1910-1914, 2019.

[12] E. Muningsih, I. Maryani, and V. R. Handayani, “Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi Berdasarkan Potensi Desa,” *Evolusi: Jurnal Sains dan Manajemen,* vol. 9, no. 1, pp. 95–100, 2021.

[13] Y. Lei and X. Qiu, "Research on the Evaluation of Overseas Strategic Climate Based on Decision Tree and Adaptive Boosting Classification Models," *Frontiers in Psychology,* vol. 12, no. December, pp. 1–10, 2021, doi: 10.3389/fpsyg.2021.803989.

[14] M. R. Nahjan, N. Heryana, A. Voutama, F. I. Komputer, U. S. Karawang, and R. Miner, "Implementasi Rapidminer Dengan Metode Clustering K-Means Untuk Analisa Penjualan Pada Toko Oj Cell," *JATI (Jurnal Mahasiswa Teknik Informatika),* vol. 7, no. 1, pp. 101–104, 2023.

[15] G. Indrawan, G. R. Dantes, P. Studi, I. Komputer, P. Pascasarjana, and U. P. Ganesha, "Data Mining Rekomendasi Calon Mahasiswa Technique For Others Reference By Similarity To Ideal," *JST (Jurnal Sains Dan Teknologi)*, no. 1, pp. 11–21, 2019.

[16] M. Ahmed, R. Seraj, S. Mohammed, and S. Islam, "The k-means Algorithm : A Comprehensive Survey and Performance Evaluation," *Electronics*, vol. 9, no. 8, pp. 1–12, 2020, doi: 10.3390/electronics9081295.

[17] B. H. Prakoso, E. Rachmawati, D. Rachmatta, and P. Mudiono, "Klasterisasi Puskesmas dengan K-Means Berdasarkan Data Kualitas Kesehatan Keluarga dan Gizi Masyarakat," *Jurnal Buana Informatika*, vol. 14, no. April, pp. 60–68, 2023.

[18] N. Suwaryo, A. Rahman, D. Marini, U. Atmaja, and A. Basri, "Klasterisasi Stok Produk Retail Untuk Menetukan Pergerakan Kebutuhan Konsumen Dengan Algoritma K-Means," *Bulletin of Information Technology (BIT)*, vol. 4, no. 3, pp. 306–312, 2023.

[19] M. Rizal *et al.*, "Algoritma Decision Tree Untuk Analisis Sentimen Public Terhadap Marketplace," *Naratif: Jurnal Nasional Riset, Aplikasi dan Teknik Informatika*, vol. 05, no. 01, pp. 18–25, 2023.

[20] B. T. Jijo and A. M. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 02, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[21] A. Setyawan and D. N. Fauzi, "Implementasi Fungsi Dispersion Ratio Pada Proses Spliting Atribut Algoritma Decision Tree," *Madani: Jurnal Ilmiah Multidisiplin,* vol. 2, no. 2, pp. 86–91, 2022, doi: 10.5281/zenodo.7782439.

[22] V. M. Member, P. Casari, and S. Member, "A Novel Hyperparameter-free Approach to Decision Tree Construction that Avoids Overfitting by Design," *IEEE Access*, vol. 7, pp.99978-99987, 2019.

[23] A. Laturiuw and Y. Singgalen, "Sentiment Analysis of Raja Ampat Tourism Destination Using CRISP-DM: SVM, NBC, DT, and k-NN Algorithm", *J. Inf. Syst. Informatics*, vol. 5, no. 2, pp. 518-535, May 2023.