ANALYSIS OF ONLINE STORE CONSUMER BEHAVIOUR WITH K-MEANS AND AGGLOMERATIVE CLUSTERING ALGORITHMS

[1]Venansius Fortunatus Wijaya, [2]Yulianto Tejo Putranto
[1,2]Program Studi Teknik Informatika Fakultas Ilmu Komputer,
Universitas Katolik Soegijapranata
[2]yulianto@unika.ac.id

ABSTRACT

An online store or e-commerce is a business system that allows buyers and sellers to conduct electronic transactions over the internet. Along with the increasing use of the internet around the world, the e-commerce business is growing rapidly and becoming a very large industry. To find out consumer behaviour, data analysis can be done using clustering algorithms, namely K-Means compared with Agglomerative Clustering and K-Means compared with K-Means++. Through these three algorithms, the data is reduced in dimension by PCA and t-SNE methods. It is then also analysed using the attributes provided by the dataset to further determine its performance measures. The parameters that will be used for clustering are K values 4, 5, and 6 of K-Means algorithm compared to Agglomerative Clustering and the optimal K value of 4 with the initialised centroid center value for K-Means++. Then determining the distance between data using the Euclidean Distance method, while for data grouping using the Average Linkage method in the Agglomerative Clustering algorithm. Through the method, the results are Agglomerative with a K value of 4 and the best t-SNE data type because the K value is good and the data type used is very good so that the results are also good and K-Means++ with an optimal K value of 4 and the best t-SNE data type because the centroid value is well initialised so that this algorithm speeds up the clustering process rather than the usual K-Means which takes longer in the process of clustering online store consumer behaviour data.

Keywords: consumer behaviour, online store, clustering, k-means, agglomerative clustering, k-means++

BACKGROUND

Online shop or e-commerce is a business system that allows buyers and sellers to conduct electronic transactions via the internet. Along with the increasing use of the internet around the world, business in e-commerce is growing rapidly and becoming a very large industry. In the e-commerce business, understanding customers is very important, because it can help in developing the right marketing strategy and increasing customer satisfaction.

One way to understand consumer behaviour is through data analysis using clustering algorithms. Clustering algorithm is a method in machine learning that is used to group data that has similarities or similarities. In analysing online store customer behaviour data, there are several types of clustering algorithms that are often used, but in this research the author uses the K-Means and Agglomerative Clustering algorithms.

The purpose of this research is to compare the performance measures of K-Means and Agglomerative Clustering algorithms in clustering online store consumer behaviour data. In addition to comparing K-Means and Agglomerative Clustering, researchers also do a separate comparison for K-Means and K-Means++ in order to know also the performance measures between K-Means and K-Means++. This research will use a data set with a total of 2240 data and also perform clustering with several attributes in the datasets. By comparing the performance measures results of the two clustering algorithms, it is expected to provide results that can be used to determine which algorithm is better and suitable for use in analysing online store consumer behaviour. Then, in the journal [1] which has the same problem of analysing customer behaviour but with different algorithms, namely FP-Growth, Apriori, and Squeezer. This journal will be used to compare the results of the journal with the research to be carried out to determine the performance measures of each algorithm with the same problem. So that with the results of my research in addition to the results of the journal that I found, it is hoped that it can find out which algorithm is suitable for customer behaviour data.

LITERATURE STUDY

Research by Qomariyah [1] uses three algorithms to analyse consumer behaviour, namely FP-Growth, Apriori, and Squeezer, to find out consumer purchasing patterns at K1mart minimarkets and help make marketing strategies more effective and efficient. This research will be the basis for comparison with the K-Means and Agglomerative Clustering algorithms and compare which of the three consumer behaviour analysis algorithms, namely FP-Growth, Apriori, and Squeezer is better by raising similar problems, namely consumer behaviour. However, the limitations of this journal are that the Squeezer algorithm is only able to generate a few purchase patterns compared to the FP-Growth and Apriori algorithms because the data size is too large.

Research by Setyorini et al [2] aims to analyse customer purchasing behaviour and identify frequently purchased items at PT Citra Mustika Pandawa by using the Association Rule method that combines FP-Growth and K-Means algorithms. However, the journal has limitations because it only shows that K-Means produces good results, while the suitability of the FP-Growth algorithm is not explained and this algorithm should be more suitable for clustering than clustering.

Research by Ruchjana et all [3] aims to group areas based on high and low rainfall as a reference for the government in handling disasters using Agglomerative and K-Means. The data used is aggregate data from December - January - February from 24 rainfall stations. The result of both clusters is 49.4%. The limitations of this journal are that more data is needed to know which algorithm is better. This journal will be used for comparison.

Research by Musdalifah and Jananto [4] This study aims to compare the performance of Apriori and FP-Growth algorithms in forming customer shopping cart association patterns to improve the company's sales strategy. However, this research has limitations as it requires larger data collection and higher minimum parameter settings to improve system performance.

Research by Sibarani [5] aims to produce an analysis to cluster data on graduating students in 2016 and 2017 to make promotions more targeted and right on target. Limitations of this journal are that the marketing department can consider areas or cities that are potential areas to visit.

Research by Rachman [6] aims to analyse products that are often purchased together in one transaction at the Cerdas-Sehat Online Store. The limitation of this journal is the lack of comparison between FP-Growth and Apriori algorithms. This journal will be used as a basis for comparative research between the speed and performance measures of K-Means and Agglomerative Clustering.

Research by Murpratiwi et al [7] aims to find the best method and k value resulting from the three clustering methods combined with the RFM model. The limitations of this journal are that clusters 2, 3, and 5 produce the same number of members in the three methods implemented, this is a consideration for using a method with a value of k = 5, because it is able to provide an even and equal distribution of values.

Research Priambodo and Jananto [8] The purpose of inventory prediction for future sales is done by interpreting the cluster results formed by the clustering algorithm used. The limitations of this journal need further study on which algorithm produces more accurate inventory predictions, based on the real sales results that have been done in the past.

Research Wulandari et al [9] aims to determine the best method on Twitter digital marketing data. Limitations are less recommended to use linkage to be used on digital marketing data, then research can be done with other clustering methods or combined clustering methods such as combined K-means and Agglomerative Nesting methods to get a higher performance measures value.

Research by Rachman and Hunaifi [10] used the Apriori algorithm to analyse drug purchases at Farma Kimia in Jakarta with the aim of identifying associative rules useful for product and drug purchase information. The Apriori algorithm is preferred over FP-growth because it has a higher minimum support level for combinations of items in an associative agreement with all the data associated with the item part transactions. However, the FP-growth algorithm is faster in performance as it has a higher lift ratio than the Apriori algorithm.

A book authored by Junjie Wu [11] k-means is one of the oldest and most widely used clustering algorithms which basically aims to find K non-overlapping clusters in the data. The algorithm is simple, efficient, straightforward, and is often used on a wide range of data. Although it has some disadvantages, such as poor performance for non-globular data clusters and sensitivity to extreme data, K-Means has dominating advantages and has inspired new variations. It remains a popular and important clustering algorithm to use in data mining research and practice.

A book authored by Sadaaki Miyamoto [12] Agglomerative hierarchical clustering is an algorithm for clustering objects based on their level of similarity or difference. This algorithm starts with each object as its own cluster, then sequentially combines the most similar clusters with each other. The end result is a hierarchy of clusters that can be depicted in the form of a dendrogram. This algorithm can be used to classify data without the need for external rules, so it is often referred to as "unsupervised classification."

RESEARCH METHODOLOGY

1. Analysis Method

   The analysis method used is descriptive method. This method is used to explain the characteristics of the data and analyse the results of the clustering process of online store consumer behaviour data using K-Means algorithm compared to Agglomerative Clustering and K-Means algorithm compared to K-Means++. The analysis is done by drawing the clustering comparison results in a table.

2. Comparison of K-means versus Agglomerative Clustering and K-Means versus K-Means++ Algorithms

   The comparison method used is the experimental method. This method is done by testing the two algorithms on store consumer behaviour data that has been obtained. The testing process includes data preprocessing, parameter selection, clustering process, and evaluating the results of each clustering using Silhouette Score.

3. Steps of the Problem-Solving Method

   The method steps taken in this project are as follows:

a. Search for online store consumer behaviour datasets for analysis from the Kaggle platform.

b. Preprocessing data by cleaning data from noise, missing values, and outliers.

c. Performing dimension reduction with PCA and t-SNE.

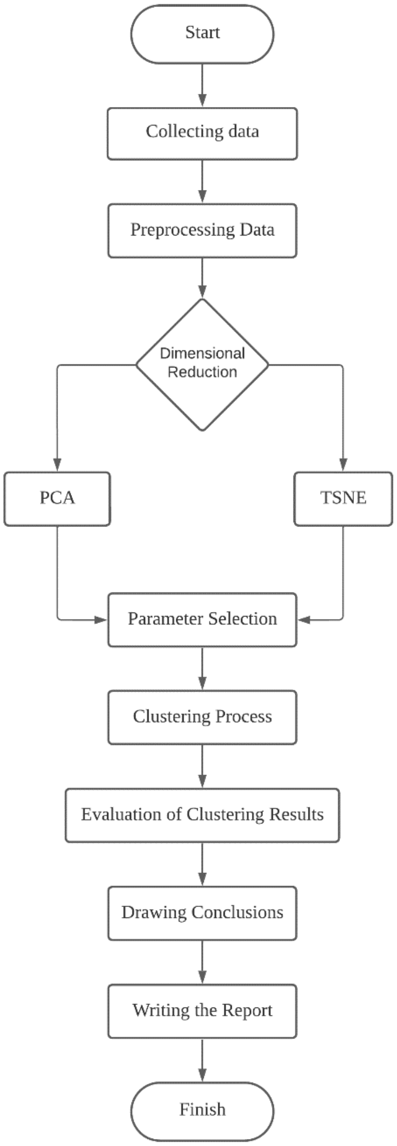
d. Parameter selection for each algorithm which includes the number of clusters, distance formula, and linkage method in Agglomerative Clustering.

e. Clustering process using K-means, Agglomerative Clustering and K-Means++ algorithms on pre-processed consumer behaviour data.

f. Evaluation of clustering results using the Silhouette Score table.

g. Drawing conclusions from the evaluation results and clustering interpretation.

h. Writing the project report.



Gambar 1. Research Methodology

Dataset Collection

In the Dataset Collection section, the data used in this research comes from Kaggle, namely marketing_campaign. This dataset contains information about the store's customer behaviour, including the customer's unique ID, Year Birth, Education, Marital Status, Income, Kidhome, Teenhome, Dt_Customer, Recency, complaints, and many more for its attributes. This dataset consists of 2240 data and has a file size of 220.19byte with csv format.

Data Cleaning

Datasets cleaning is a process to remove noise and inconsistencies. The data obtained from Kaggle is not completely perfect because there are some incomplete or missing data. In addition, if there are unnecessary attributes, they will also be removed. Data cleaning will also affect the performance of the analysis to compare K-Means algorithm with Agglomerative Clustering and K-Means with K-Means++, with complete data will make the comparison results more accurate.

Data Augmentation

Augmentation

The first step is to get the training and testing datasets, then preprocessing the training dataset where this dataset processes raw data into data that has been cleaned. After performing dimension reduction, then selecting parameters for each algorithm which includes the number of clusters, distance type, and linkage method in Agglomerative Clustering. The next step is the clustering process using K-means, Agglomerative Clustering, and K-Means++ algorithms on consumer behaviour data that has been pre-processed and reduced dimensions. Then evaluate the clustering results using the Silhouette Score table to compare the performance of the K-means algorithm compared to Agglomerative Clustering and K-Means compared to K-Means++. The Silhouette Score table provides a Silhouette Score value for each clustering model, where the higher the Silhouette Score value but the results are not only seen from the Silhouette Score value results alone, but from the visual results produced as well, so the Silhouette Score value and the corresponding visual results can be said to be good results. After obtaining the Silhouette Score value and the corresponding visuals, the clustering results are interpreted and conclusions are drawn to understand consumer behaviour in each cluster. The conclusions are used to determine a better clustering algorithm and presented in the project report.

Function

a. K-Means Formula

1. Specify k(free value) as the number of clusters to be formed.
2. Generate a random value for the initial center of each cluster (centroid).
3. Calculate the distance between each input data and each centroid using the Euclidean distance formula. here for the formula:

$$d(xi, \mu j) = \sqrt{\sum (xi - \mu j)^2} \quad (1)$$

In function (1), for each dimension we will calculate the difference between data point xi and centroid μi in that dimension. Then, the difference is squared and summed up for all dimensions. After that, the square root of the sum is taken to get the distance between the data point and the centroid. By using this formula, we can determine the closest centroid for each data point and classify the data into the appropriate cluster.

4. Classification of each data is done by finding the closest distance between the data and the centroid in each cluster. Each data will be classified into the cluster with the closest centroid.
5. Update the new centroid value of each cluster based on the average data value in the cluster using the formulas:

$$\mu j\ (t+1) = \frac{1}{Nsj} \sum j \in sj^{xj} \tag{2}$$

In function (2), to find the new centroid value μ at the next iteration t-1 in the k-means algorithm. The method is to use the average of all data in a cluster sj for each dimension xj, then the average value is used as the new centroid value for the cluster. Ns is the amount of data contained in cluster sj. This calculation is done to update the centroid position at each iteration until a normal cluster is found.
6. Repeating from step 3 to 5 until the members of each cluster have not changed or have reached the predetermined iteration limit.

b. Agglomerative Clustering Formula

1. Calculate the distance matrix using the following Euclidean distance formula:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2} \tag{3}$$

In function (3), d_ij is the distance between object i and j, then x_ij is the value of object i in the kth variable, for x_jk is the value of object j in the kth variable. And p is the number of variables observed. Each variable of the first data with the second data variable is subtracted, then the result is raised by 2, then all the power results are summed, and the result is the square root of the sum result.
2. If the distance between object a and b has the smallest distance value compared to the distance between other objects in the Euclidian distance matrix, then the combination of the two clusters in the first stage is d_ab.
3. If d_ab is the closest distance from Euclidian distance, then the formula for Agglomerative Clustering agglomerative method is:

$$d_{(ab)c} = min\{d_{a,c}\} \tag{4}$$

$$d_{(ab)c} = average\{d_{a,c}\} \tag{5}$$

$$d_{(ab)c} = max\{d_{a,c}\} \tag{6}$$

In function (4), the distance between clusters a and b to c is the closest distance between points in cluster a or b and points in cluster c. Function (4) is a calculation for single linkage.

In function (5), take the average distance between point a and all points in cluster c, then repeat with point b and add it up. Furthermore, the result is divided by the total number of points in cluster c. Function (5) is the calculation for average linkage.

In function (6), by finding the maximum distance between data points a and c through data point b. Function (6) is a calculation for complete linkage.

4. Repeat steps 2 and 3 until only one cluster remains.

c. Silhouette Score

1. Calculate the average distance of the i-th data with all data in the same cluster.

with the following formula:

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i,j) \tag{7}$$

In function 7, a(i) is the average distance of the i-th data to all data in the same cluster, then i is the index of the i-th data, C_i is the cluster containing the i-th data, d(i, j) is the distance between the i-th data and the j-th data.

2. Calculate the minimum distance of the i-th data with all data in other clusters.

With the following formula:

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j) \tag{8}$$

In formula b(i) is the minimum distance of the i-th data to all data in other clusters, i is the index of the i-th data, C_j is the other cluster containing the i-th data, and d(i,j) is the distance between the i-th data and the j-th data.

3. Calculate the Silhouette Score with the following formula:

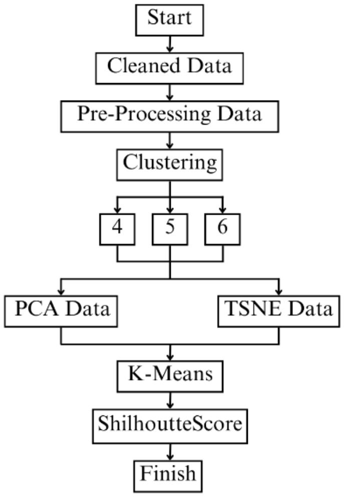
$$s(i) = \frac{\left(b(i) - a(i)\right)}{\max\{a(i), b(i)\}}, if |C_I| > 1 \tag{9}$$

In formula 9, s(i) is the silhouette score of the i-th data, b(i) is the minimum distance of the i-th data to all data in other clusters, and a(i) is the average distance of the i-th data to all data in the same cluster.

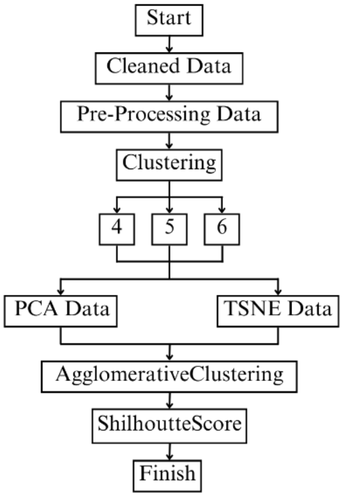
The Silhouette score is a measure that ranges from -1 to 1. A positive score indicates that the i-th data point is closer to the data in its own cluster than to the data in other clusters. A negative value indicates that the i-th data is closer to the data in another cluster than to the data in its own cluster. A score of 0 indicates that the i-th data point is in between the two clusters. The higher the Silhouette score, the better the clustering.

RESULT

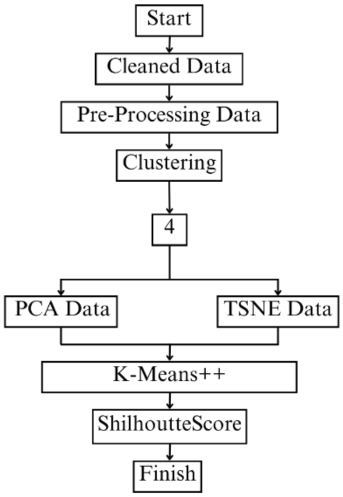
Test Scenario



Gambar 2. Test Scenario with K-Means



Gambar 3. Test Scenario with AgglomerativeClustering



Gambar 4. Test Scenario withK-Means++

The analysis was conducted with two scenarios, namely:

1. Analysis the data that has been reduced in dimension with PCA and t-SNE then clustering process with K-Means and tested using several different clusters namely 4, 5, and 6 shown in Figure 4.1.
2. Analysis of data that has been reduced in dimensions with PCA and t-SNE and then clustering process with Agglomerative Clustering and tested using several different clusters of 4, 5, and 6 shown in Figure 4.2.
3. Analysis the data that has been reduced in dimension with PCA and t-SNE is then carried out the clustering process with K-Means++ and tested using the optimal cluster of 4 shown in Figure 4.3.

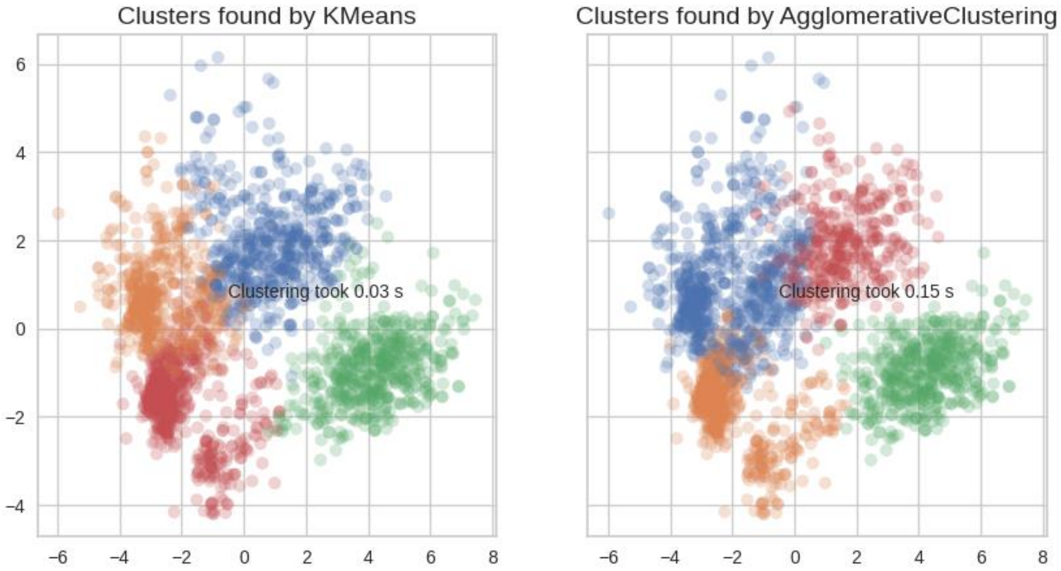
Both analysis scenarios were conducted using data that had been cleaned, pre-processed, and dimension reduced. There are 2 data with different data reduction methods, namely:

1. Data is tested with the K-Means algorithm.
2. Data is tested with the AgglomerativeClustering algorithm.
3. Data is tested with the K-Means++ algorithm.
4. Dimension-reduced data with PCA method.
5. Dimension-reduced data with t-SNE method.
6. Data tested with cluster value 4.
7. Data tested with cluster value 5.
8. Data tested with cluster value 6.

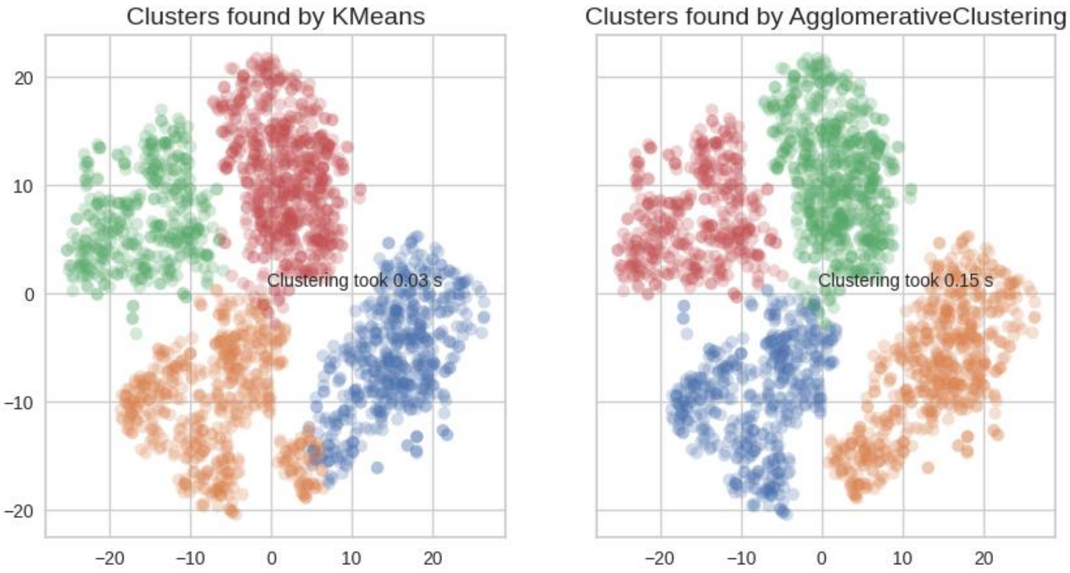
Evaluation is done by looking at the results of each Silhouette Score by looking at the effect of performance measures to see the results of the clustering comparison of the two algorithms used.

Tabel 1. Analysis Result using Silhouette Score with cluster value 4

No.	Clustering Method	Silhouette Score	Data Type
1.	K-Means	0.373734	PCA
2.	AgglomerativeClustering	0.332647	PCA
3.	K-Means	0.391825	t-SNE



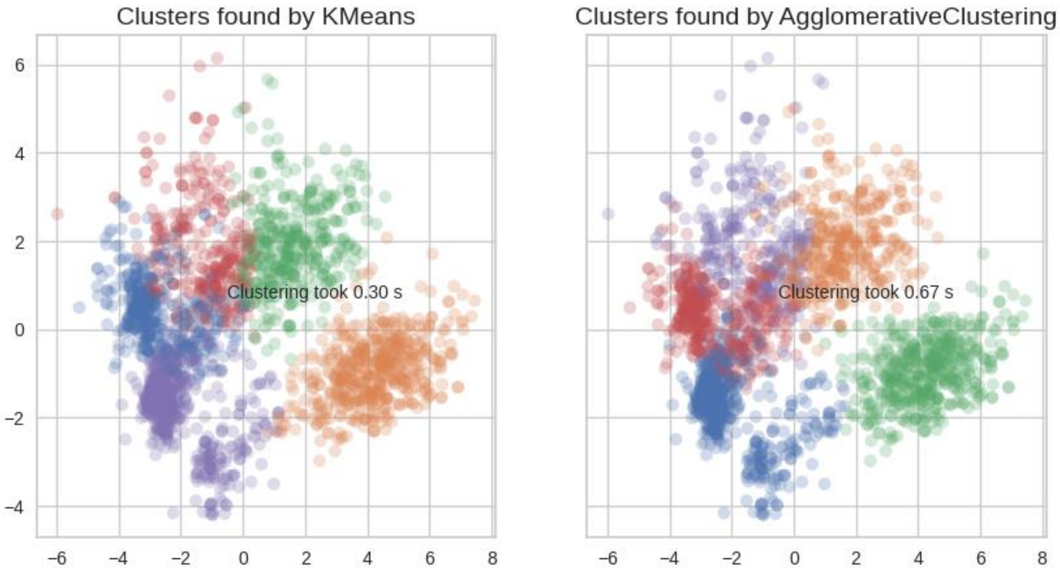
Gambar 5. Cluster 4 test result with PCA data Type



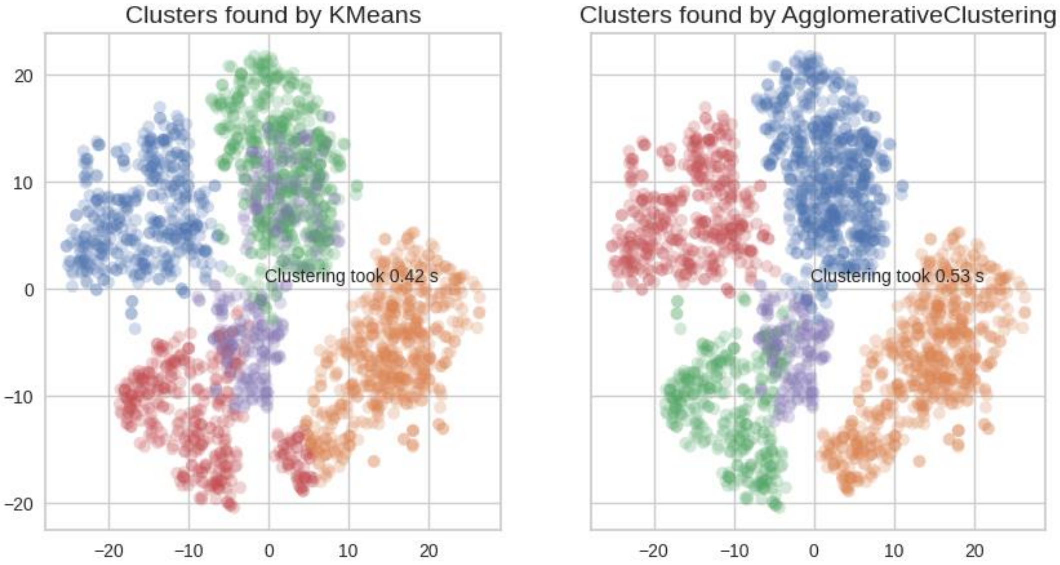
Gambar 6. Cluster 4 test result with t-SNE data Type

Tabel 2. Analysis Result using Silhouette Score with cluster value 5

No.	Clustering Method	Silhouette Score	Data Type
1.	K-Means	0.350334	PCA
2.	AgglomerativeClustering	0.331357	PCA
3.	K-Means	0.406119	t-SNE
4.	AgglomerativeClustering	0.404798	t-SNE



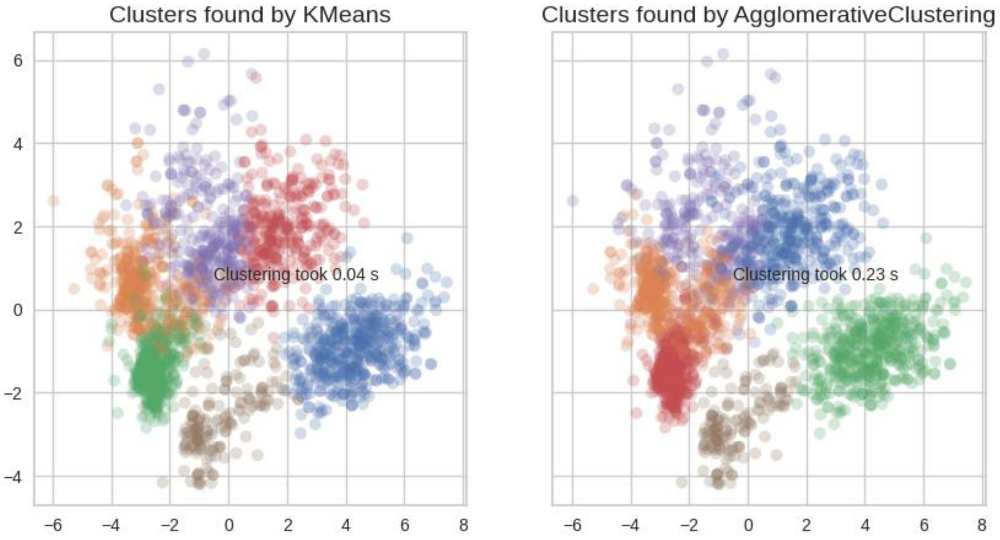
Gambar 7. Cluster 5 test result with PCA data Type



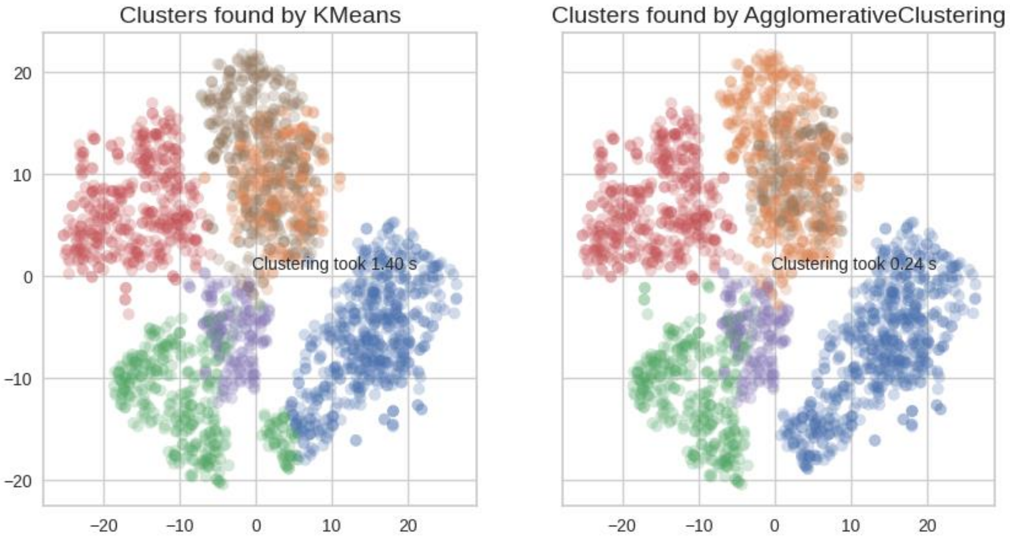
Gambar 8. Cluster 5 test result with t-SNE data Type

Tabel 3. Analysis Result using Silhouette Score with cluster value 6

No.	Clustering Method	Silhouette Score	Data Type
1.	K-Means	0.350372	PCA
2.	AgglomerativeClustering	0.334229	PCA
3.	K-Means	0.414760	t-SNE
4.	AgglomerativeClustering	0.410416	t-SNE



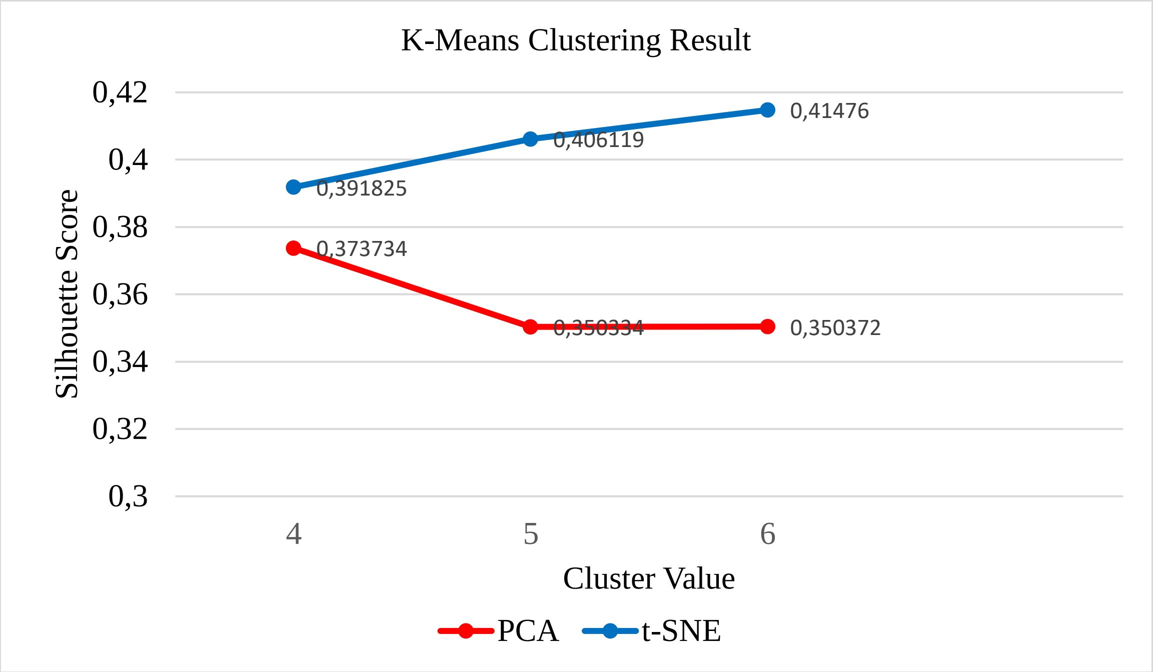
Gambar 9. Cluster 6 test result with PCA data Type



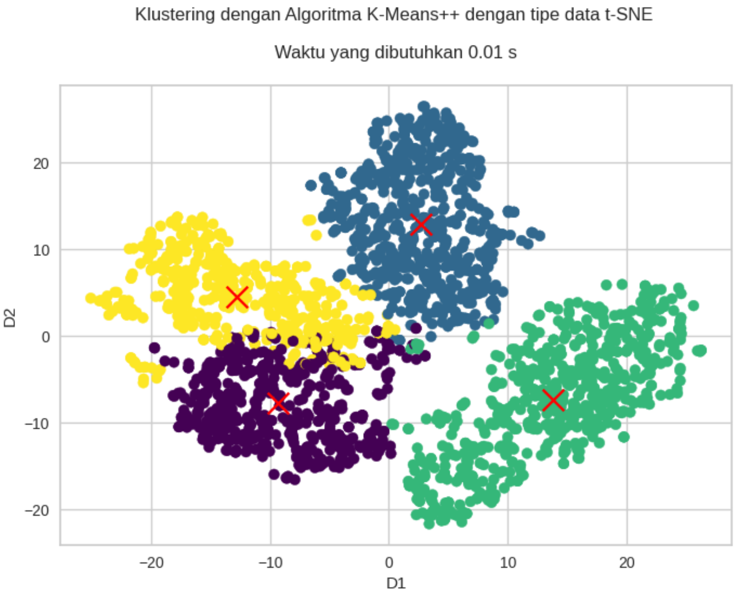
Gambar 10. Cluster 6 test result with t-SNE data Type

Tabel 4. Analysis Result using Silhouette Score and Finishing Time

No.	Clustering Method	Silhouette Score	Data Type	Finishing Time
1.	K-Means++	0.373734	PCA	0.02 s
2.	K-Means++	0.391825	t-SNE	0.01s
3.	K-Means	0.373734	PCA	0.03s
4.	K-Means	0.391825	t-SNE	0.03s

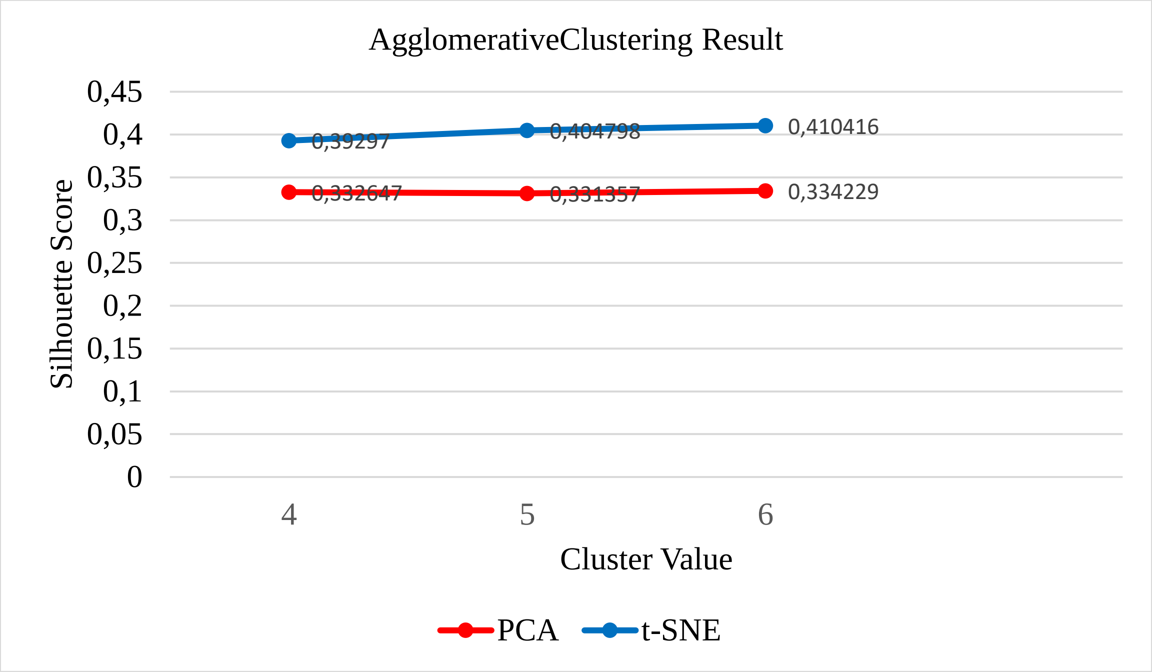


Gambar 11. K-Means++ Clustering test result with PCA data type

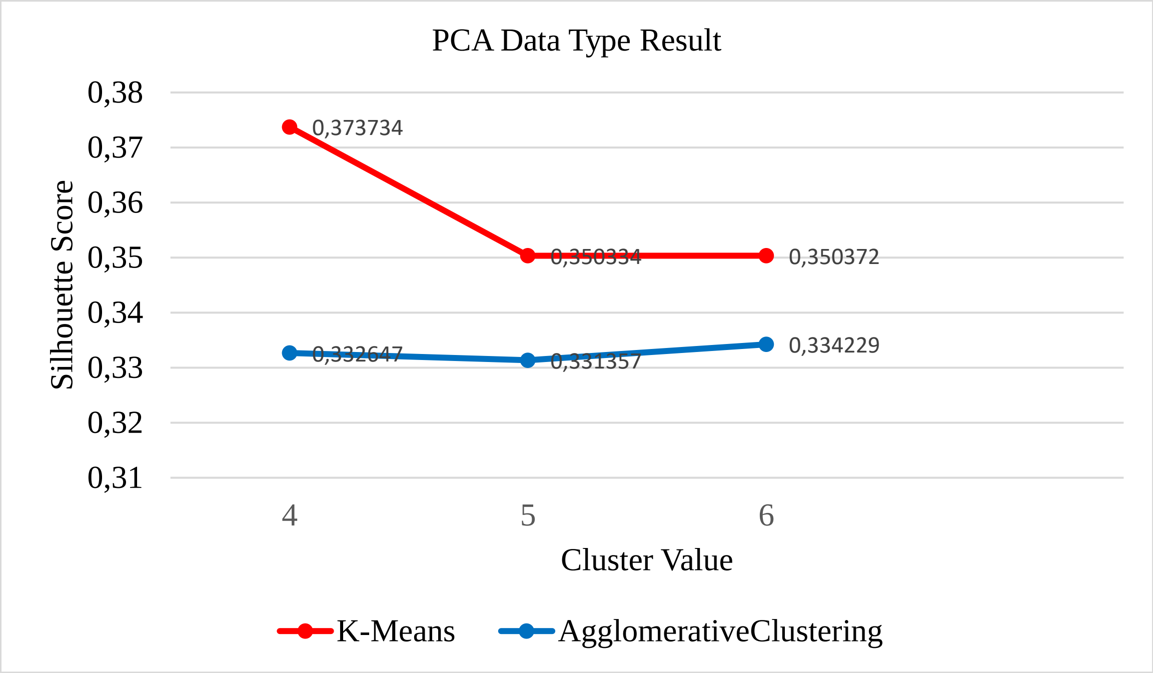


Gambar 12. K-Means++ Clustering test result with PCA data type

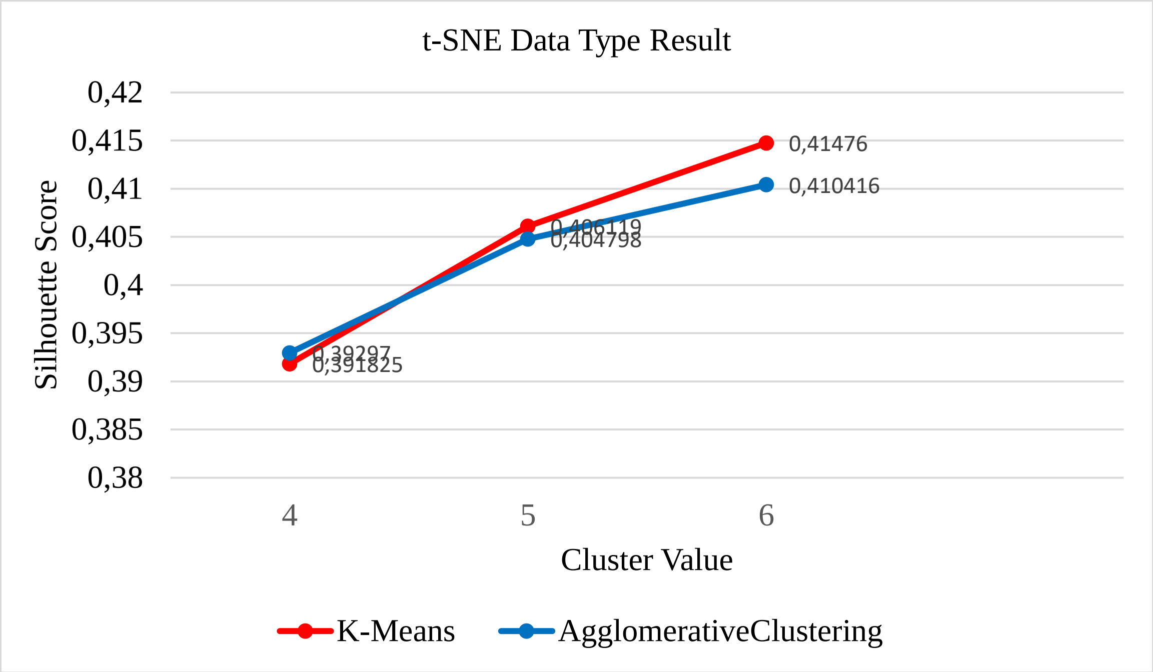
Gambar 13. Graph of Clustering Result with K-Means Algorithm



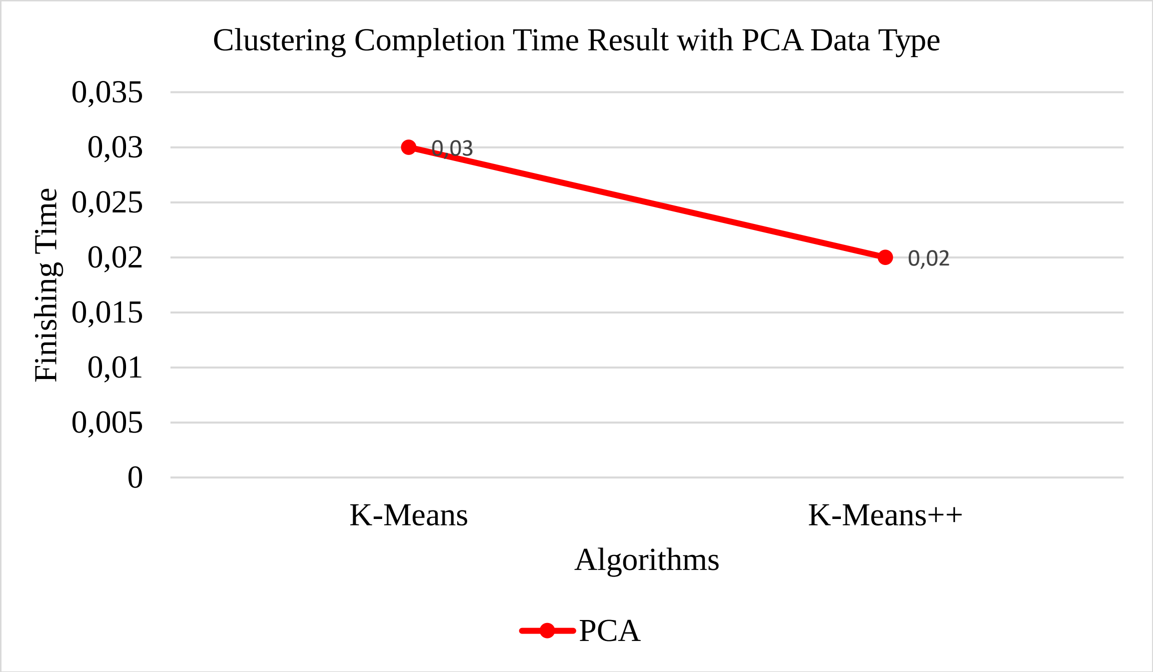
Gambar 14. Graph of Clustering Result with AgglomerativeClustering Algorithm



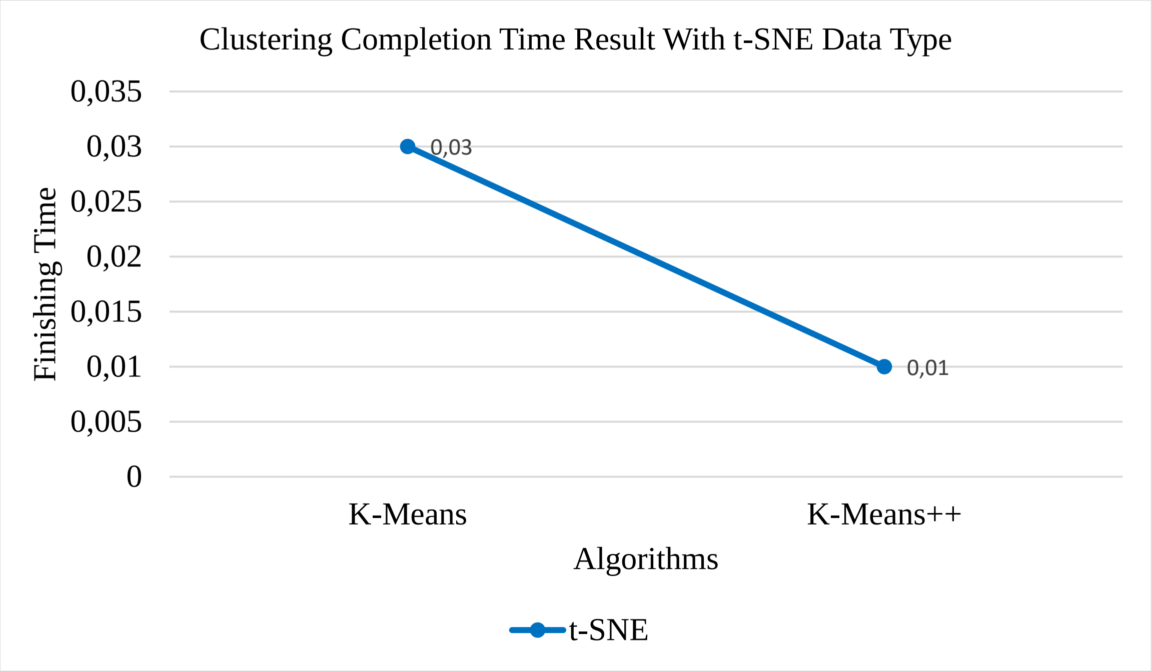
Gambar 15. Graph of Clustering Result with PCA data type



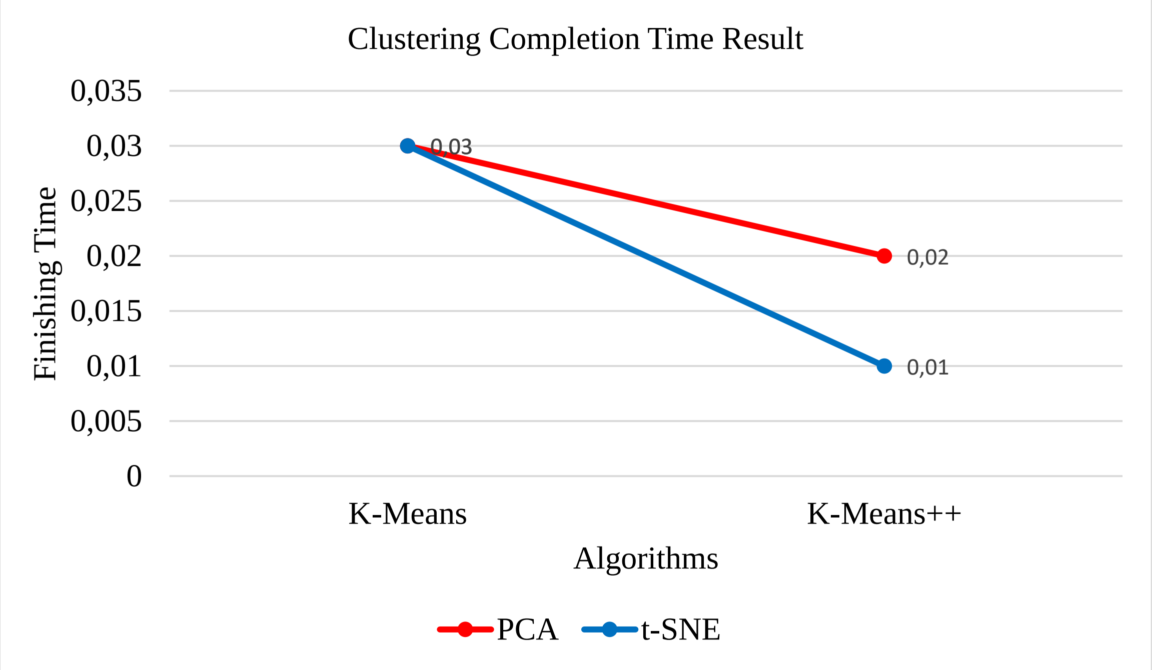
Gambar 16. Graph of Clustering Result with t-SNE data type



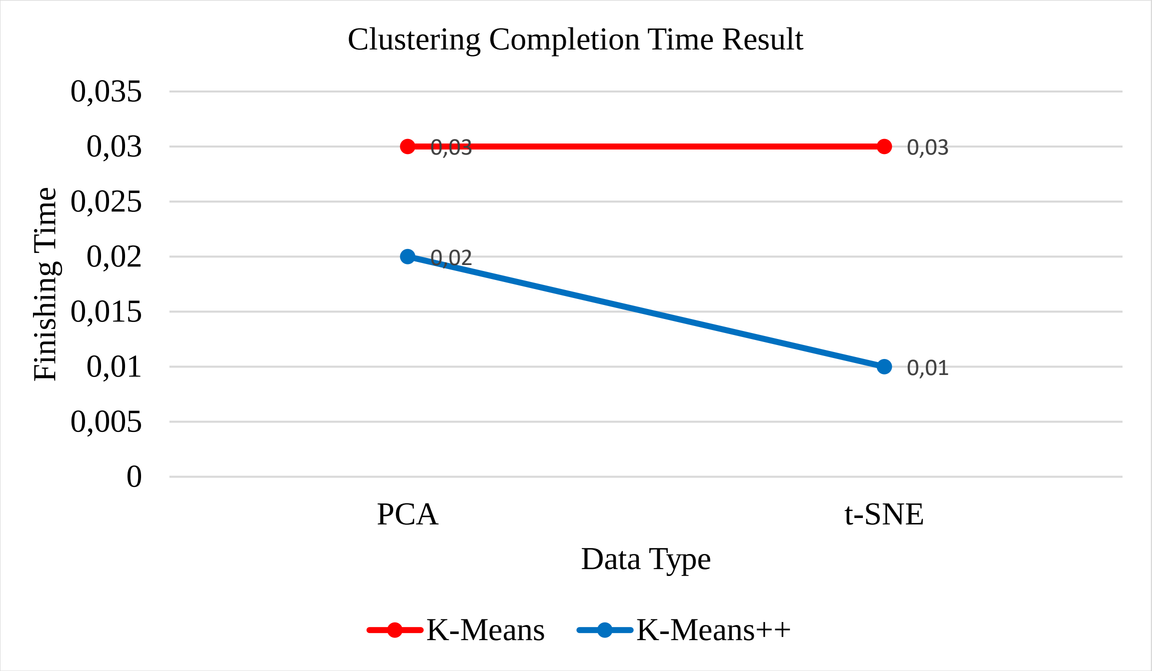
Gambar 17. Graph of Completion Time with PCA Data Type



Gambar 18. Graph of Completion Time with t-SNE Data Type



Gambar 19. Graph of Completion Time



Gambar 20. Graph of Completion Time with K-Means and K-Means++ Algorithms

From the clustering results above, it can be seen that the data is clustered with k values of 4, 5, and 6, then also using PCA and t-SNE dimension reduction. At a value of k 4, the Silhouette Score results are obtained, namely in the K-Means algorithm 0.373734 with PCA data type and 0.391825 with t-SNE data type, then with AgglomerativeClustering 0.332647 with PCA data type and 0.392970 t-SNE. Here the superior is the AgglomerativeClustering algorithm with t-SNE data type and the lowest is the Agglomerative Clustering algorithm with PCA data type. Then on the graph it can be seen that the clustering results with type data are still less than optimal because many clusters are still together, then in the t-SNE data type the clustering results look better because the clusters look far away even though some are still close together and for good cluster results visually is Agglomerative with t-SNE data type because the clusters do not overlap but only close together while K-Means with t-SNE data type there is still a little piling up. So, it can be concluded for a good k value of 4 in the AgglomerativeClustering algorithm with t-SNE data type because both the Silhouette Score or graph show good results. At a value of k 5, the results obtained are in the K-Means algorithm getting a Silhouette Score value of 0.350344 with type data and 0.406119 with t-SNE data type, then with the AgglomerativeClustering algorithm getting a Silhouette Score value of 0.331357 with PCA data type and 0.404798 with t-SNE data type.

At a value of k 5 the best Silhouette Score results are in K-Means with t-SNE data type and the lowest in AgglomerativeClustering with PCA data type. Then when viewed on a three-dimensional graph the results with PCA data type are very ugly because the cluster results overlap. Then in the graph results with the t-SNE data type the results show better than PCA but there are still some that overlap both K-Means and Agglomerative but the better results are on Agglomerative when viewed from the graph results. It can be concluded that the best Silhouette is in K-Means with t-SNE data type and three-dimensional graph on AgglomerativeClustering with t-SNE data type with different results on the Silhouette Score and three-dimensional graph. At a value of k 6, the Silhouette Score results are obtained, namely in the K-Means algorithm getting a value of 0.350372 with PCA data type and 0.414760 with t-SNE data type, then in AgglomerativeClustering getting a result of 0.334229 with PCA data type and 0.410416 with t-SNE data type.

In the value of k 6 the best Silhouette Score results are in K-Means with t-SNE data type and the lowest value is in AgglomerativeClustering with t-SNE data type. Then the results of the three-dimensional graph with PCA data type are very bad because the cluster results still stack each other. The t-SNE data type is better than PCA but there are still many clusters and if you look at the good results there is Agglomerative Clustering. From these two results it can be concluded that in the Silhouette Score the best results are in K-Means with the t-SNE data type and in the three-dimensional graph the different results are with the AgglomerativeClustering algorithm with the t-SNE data type.

In testing data with PCA data type using the optimal K value of 4 with the K-Means++ algorithm produces visual cluster results that still look less good. It can be seen that the clusters are still close together and piled up. Then the middle centroid value is initialized. The benefit of the center centroid value is to display the average of all points in the cluster. So that we can easily see more accurate results whether it can be said to be good or not. In addition to accurate results, it also speeds up the completion time in the cluster, because the center centroid value is used as the center in iteration so it speeds up the cluster process. In testing the K-Means++ algorithm with PCA data type, the Silhouette Score result is 0.373734 with a completion time of 0.02 seconds. This result is faster than K-Means with regular PCA data type which is 0.03 seconds. For the Silhouette Score and visual results, all that distinguishes is the completion time and the centroid center value that has been initialized properly at each cluster point so as to speed up the cluster process. Then with the same algorithm with the t-SNE data type with the optimal K value of 4, the results obtained are even better. The Silhouette Score result obtained is 0.391825, this result is the same as the usual K-Means with t-SNE data type which is 0.391825. For visual results, it is also the same, there is not much accumulation, the only difference is in the centroid center value in the K-Means++ algorithm. Because there is a centroid center value, the time is also accelerated which is 0.01 seconds faster than the usual K-Means which is 0.03 seconds. The results show that both with PCA and t-SNE data types with the K-Means++ algorithm, the time for completion is accelerated because there is an initialized centroid value. The best result is K-Means++ with t-SNE data type because in terms of visual, Silhouette Score value, and completion time are all the best. There is not much piling up, the centroid center value is also initialized properly, the time is also very fast only 0.01 seconds.

From the graphical results of the comparison results on each algorithm, it can be seen in the K-Means table with the t-SNE data type that the graph is getting longer and the Silhouette Score value is increasing, while in PCA it dropped at a value of k 5 then at a value of k 6 it rose slightly, then in AgglomerativeClustering with the t-SNE data type it is increasing and in the PCA data type it dropped at a value of k 5 and rose again slightly at a value of k 6. When viewed as a whole K-Means algorithm with t-SNE data type the graph is increasing and is at the top and AgglomerativeClustering algorithm with PCA data type is at the bottom. So, the conclusion is that the best k value is at a value of k 4 because both the Silhouette Score value and the three-dimensional graph are in the same result, namely on Agglomerative with t-SNE data type, while at k values 5 and 6 there are no similar results on the Silhouette Score and three-dimensional graph.

From the graphical results on each type of data, several results are obtained. In the PCA K-Means data type has higher results. The Silhouette Score result obtained at k value 4 is 0. 373734, k value 5 is 0.350334, and value 6 is 0.350372. The graph shows that under the highest Silhouette Score value there is a value of k 4 and the lowest at value 5. There are ups and downs on the graph with the order from the high value of K 4, then the value of K 6, the lowest at the value of K 5. Then for AgglomerativeClustering is far below K-Means. For the Silhouette Score results, the K 4 value is 0.332647, the K 5 value is 0.331357, the K 6 value is 0.334229. The highest Silhouette Score result is at the value of K 6 then in the middle is at the value of K 4 and the lowest is at the value of K 5. Overall, for PCA data type, the highest Silhouette Score value is in K-Means with a value of K 4 and the lowest is in AgglomerativeClustering with a value of K 5. So, the conclusion for the best PCA data type is at the value of K 4 in the K-Means algorithm. In the result graph with t-SNE data type, several results are obtained. For the K-Means algorithm, there are several results for the Silhouettes Score. With a value of K 4, the Silhouette Score result is 0.391825, for a value of K 5, the result is 0.406119, for a value of K 6, the result is 0.414476. for this result sequentially increases from the value of K 4 to the value of K 5 then to the value of K 6. The highest is at the value of K 6 and the lowest is at the value of K 4 for the K-Means Algorithm. Then using the AgglomerativeClustering algorithm has different results for the Silhouette Score results. With the value of K 4, the Silhouette Score result is 0.391825, the value of K 5 is 0.404798, the value of K 6 is 0.410416. From these results, the higher the K value, the higher the Silhouette Score result, so the order from smallest to largest is the value of K 4 to the value of K 5 and to the value of K 6. For the largest result at the value of K 6 and the smallest at the value of K 4. So, for the highest Silhouette result is at the value of K 6. Then for the conclusion on the t-SNE data type graphically initially at the top is the K 4 value of AgglomerativeClustering but when it goes to the K 5 value it starts to change, namely K-Means is at the top and settles to the K 6 value. So, the conclusion is that the highest is at the K 6 value of K-Means and the lowest is at the K 4 value of K-Means.

From the graph of the results of the completion time with the data type to compare between the K-Means and K-Means++ algorithms, it shows that the graph that was originally at the top has dropped due to the good results of the K-Means++ completion time. In the PCA data type with the K-Means algorithm, the completion time is 0.03 seconds then decreases when in K-Means++

which is only 0.02 seconds. Good results because it speeds up the cluster process. Then in the t-SNE data type with the K-Means algorithm has a completion time of 0.03 then decreases a lot when in K-Means++ which is only 0.01 seconds. A very good result because it greatly speeds up the cluster process. By looking at this graph, it can be seen that with K-Means++ the cluster process is accelerated both with PCA and t-SNE data types. The best results are also obtained from K-Means++ with t-SNE data type because the completion time is very fast at 0.01 seconds only. This result can happen because K-Means++ initializes the center centroid value well which is very useful for the clustering analysis process and speeds up the completion time, because the center centroid value has been initialized first. The conclusion from this graph is that K-Means++ algorithm with t-SNE data type is the best in performing clustering process for consumer behaviour data.

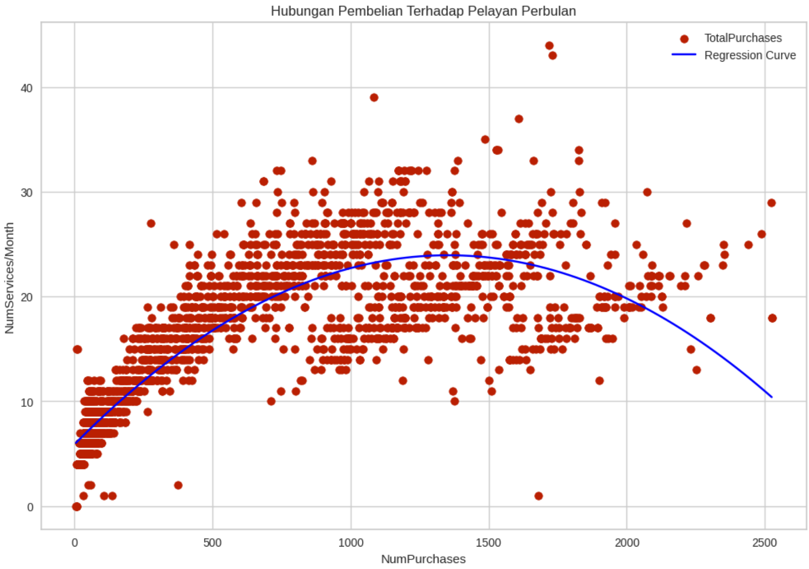
From the results of the graph of completion time against the K-Means and K-Means++ algorithms, several results are shown. In the PCA data type with the K-Means algorithm the time required is 0.03 seconds then decreases in K-Means++ with the time required 0.02 seconds only. Then with the t-SNE data type with the K-Means algorithm the time required is 0.03 seconds then decreases in K-Means++ with the time required 0.01 seconds only. It can be seen that the PCA data type is still not good at clustering. This is shown by the PCA data type which is above t-SNE. t-SNE is below because with K-Means++, the time required is short at 0.01 seconds only. This result can be considered very good. This is because K-Means initializes the center centroid value better than K-Means, especially the data type used by t-SNE which is better than PCA data type. The conclusion from this graph is that K-Means++ with t-SNE data type can perform the process of clustering consumer behaviour data very quickly and very well, which is only 0.01 seconds.

From the results of the graph of completion time against data type, several results are shown. In the K-Means Algorithm with PCA and t-SNE data types, the time required is the same at 0.03 seconds. Then in the K-Means++ algorithm with PCA data type the time required is only 0.02 then decreases in the t-SNE data type with the time required only 0.01 seconds. In K-Means there is no decrease in 0.03 seconds even though the data type is changed, then the results are above K-Means++. While K-Means++ has decreased from PCA to t-SNE which means it is getting faster by changing the data type and the graph is far below K-Means. This is due to the center centroid value that facilitates the analysis in the cluster process and also speeds up the time required to perform the cluster process. So, in this graph, it can be concluded that K-Means++ is best for clustering consumer behaviour data because the time required is faster than regular K-Means.

DISCUSSION

From the above analysis, it can be seen that there are several factors that affect the clustering results. First, the clustering method used. The clustering method determines the grouping of the data. Different clustering methods will give different Silhouette Score results. For example, the Agglomerative method tends to produce a higher Silhouette Score than K-Means. Secondly, the type of data used can also affect the Silhouette Score results. More evenly distributed data will produce a higher Silhouette Score than uneven data. Third, the number of clusters. The number of clusters can also affect the Silhouette Score. Too few or too many clusters can cause a low Silhouette Score. Trials were conducted at several K values, namely 4, 5, and 6 and there were also 2 types of data used, namely PCA and t-SNE. The results obtained are different both from the k value used and the results in the Silhouette Score and three-dimensional graphs. At the best K value of 4 there is AgglomerativeClustering with t-SNE data type, then the best k value of 5 is K-Means with t-SNE data type, then the best k value of 6 is K-means with t-SNE data type as well. Then for graph three produces the same results, namely in AgglomerativeClustering with t-SNE data type. Then in the line graph it also goes up and down too. The results are different in the Silhouette Score and the graph this happens because the type of data used is different and the k value is also different, causing different results. The reason t-SNE can be superior is because it is made to handle non-linearity and the relative distance between close points is maintained in high dimensions. This makes t-SNE very effective for mapping data that has non-linear relationships, resulting in better visualization for this kind of data. In this case, the change in the value of one variable is not always proportional to the change in the value of another variable. Based on the Figure 4.20 it can be seen that the number of purchases (NumPurchases) and the number of services (NumServices/Month) are not linearly relationships. The blue curve shows the regression of the data. For example, if the number of purchases increases from 10 to 20, then the number of services does not increase from 2 to 4, but rather increases from 2 to 3. There are also other examples such as the data points (10, 2) and (20, 3). An increase in the number of purchases from 10 to 20 does not result in an increase in the number of services by 2, but by 1. then at data point (30, 4). An increase in the number of purchases from 20 to 30 does not result in an increase in the number of services by 2, but by 1. PCA then works by defining principal axes (principal components) that describe most of the variability in the data. While effective in reducing dimensionality, PCA can be less effective in dealing with non-linear structures or complex manifolds in data. This results in poor or even very poor visualization results when there are significant non-linear patterns. So, from this explanation, it can be seen that t-SNE type data can handle consumer behaviour data well because the data is non-linear. Then for the value of k it is best at a value of 4 so it is concluded for good results at a value of k 4 with the AgglomerativeClustering algorithm with the t-SNE data type. because both the Silhouette Score and the three-dimensional graph produce the same results in contrast to the k values of 5 and 6 producing different results between the Silhouette Score and the three-dimensional graph. Furthermore, comparing K-Means and K-Means + + has the same Silhouette Score results, namely in the PCA data type both have results of 0.373734 and for t-SNE 0.391825. So, for the Silhouette Score results both have the same results there is no difference. Then what is compared here is the different completion times. In K-Means both PCA and t-SNE data types have the same completion time of 0.03 seconds. Then in K-Means with PCA data type has a slightly faster time of 0.02 seconds. This shows there is an increasing improvement in K-Means++ which previously with K-Means can be with PCA data type the result is 0.03 seconds increased to 0.02 seconds with K-

Means. Then when viewed on K-Means++ with t-SNE data type, it gets the best and fastest time from among all of them both K-Means with PCA and t-SNE data types and K-Means++ with PCA data type. K-Means++ with t-SNE data type only takes 0.01 seconds, a very fast time to do the cluster process. This can happen because K-Means++ initializes the centroid value well. This center centroid value serves to help the process of analysing clusters and also with the center centroid value initialized, it is more efficient in carrying out the process so that the completion time is well accelerated. So, the center centroid value is very influential for doing clusters to be faster. In running K-Means++ sometimes I find there are differences in the completion time for the PCA data type I tested five times and got the results, namely the initial results were at 0.03 seconds then the second test decreased to 0.02 seconds, the third 0.03 seconds again then the fourth 0.02 seconds again and the fifth remained at 0.02 seconds so if the average result is 0.02 seconds and this PCA is never less than 0.02 so at least only at 0.02 seconds. Then with the t-SNE data type in testing five times the results also vary the first time is 0.01 seconds, the second time is 0.02 seconds, the third time is 0.02 seconds, the fourth time is 0.01 seconds and the fifth is 0.01 seconds so if the average time generated is 0.01 seconds and this t-SNE never exceeds 0.02 at most at 0.02 only. In conclusion, the comparison for K-Means and K-Means++ is that K-Means++ can be a good algorithm in the cluster process besides being fast, the initialization of the center centroid is good, also the visual results for the t-SNE data type are also good, not many are piled up or close together. The conclusion for the whole is the comparison for the best algorithm using K-Means and AgglomerativeClustering with PCA and t-SNE data types, with K values of 4, 5, and 6 is AgglomerativeClustering with t-SNE data types using a K value of 4 is the best with a good Silhouette value besides that the visuals are not much piled up and close together. Then the comparison for the best algorithm using K-Means and K-Means++ with PCA and t-SNE data types with an optimal K value of 4 is the best K-Means++ algorithm with t-SNE data type because when viewed from the visual and Silhouette Score values the results are the same but the fastest completion time than others because there is a well-initialized center centroid value. Thus, the best algorithms are Agglomerative Clustering and K-Means++ with t-SNE data type to perform the cluster process on consumer behaviour data.



Gambar 21. Graph of Numpurchase Relationship to NumServices/Month

CONCLUSION

The conclusion of this study is the performance measures of the K-Means and Agglomerative Clustering algorithms in clustering online store consumer behaviour data by reducing dimensions using PCA and t-SNE which also uses several attributes in the data with tests carried out using 3 K values, K values 4, 5, and 6 obtained several results. The results obtained are different, both from the Silhouette Score results and also seeing through the visualization results. The results for the Silhouette Score are different for each K value and each type of data tested. At a value of K 4 the K-Means algorithm with PCA data type obtained a result of 0.373734 and with t-SNE data type the result was 0.391824, then for the AgglomerativeClustering algorithm with PCA data type the result was 0.332647 and with t-SNE data type the result was 0.392970. At K value 5, the K-Means algorithm with PCA data type results in 0.350344 and with t-SNE data type results in 0.406119, then with the AgglomerativeClustering algorithm with PCA data type results in 0.331357 and with t-SNE data type results in 0.404798. At K value 6, the K-Means algorithm with PCA data type results 0.35372 and t-SNE data type results 0.414760, then the AgglomerativeClustering algorithm with PCA data type results 0.334229 and with t-SNE data type results 0.410416. Then for the results visually it can be seen that at K values 5 and 6 there are still many close together and piled up so the results are not good, then at K value 4 in the K-Means algorithm with PCA data type looks close together and piled up so it is not good then in the t-SNE data type it is a bit far away but still fairly close but not much piled up so in this visual result it is still not good. In the AgglomerativeClustering algorithm with PCA data type the results are still piled up and close together so it is still not good, then in t-SNE it is fairly good because the results are not too close together and nothing is piled up. So, the conclusion is that for the performance measures results of the comparison of the K-Means and AgglomerativeClustering algorithms for clustering online store consumer behaviour data, the best is AgglomerativeClustering with the t-SNE data type using a K value of 4 because in terms of the Silhouette Score results, namely 0.392970 and also visuals that are not piled up or close together are fairly good.

The next conclusion is to compare the level of speed in clustering the K-Means and K-Means++ algorithms for clustering online store consumer behaviour data that has been reduced using PCA and t-SNE along with several components tested with the optimal K value of 4, the results are only in the completion time and visual results are slightly different. In K-Means and K-Means++ with PCA and t-SNE data types have exactly the same Silhouette Score results for PCA the result is 0.373734 and with t-SNE the result is 0.391825 so there is no difference for Silhouette Score results. In the visual results there is a slight difference, namely for K-Means ++ there is an initialized center centroid value. With the center centroid value initialized, the cluster process will run faster and facilitate the analysis process. The comparison is on the completion time. In the K-Means algorithm with PCA and t-SNE data types the completion time is the same which is 0.03 seconds, then for K-Means ++ with PCA data type the completion time is 0.02 seconds with the center centroid value initialized quite well and with t-SNE data the completion time is 0.01 seconds with the center centroid value initialized very well. The fastest completion time is the K-Means++ algorithm with t-SNE data type with a completion time of 0.01 seconds with a very well initialized center centroid value.

And suggestions for future research are as follows. First, try adding other algorithms so that you can compare the results of other accuracies for researching online store consumer behaviour. Second, try research by reducing dimensions with types other than PCA and t-SNE to find out better performance measures.

DAFTAR PUSTAKA

[1] S. Qomariyah, “Perbandingan Algoritma FPGrowth, Apriori dan Squeezer pada Analisis Perilaku Konsumen di Minimarket K1mart ITS,” p. 103.

[2] S. G. Setyorini, E. K. Sari, L. R. Elita, and S. A. Putri, “Analisis Keranjang Pasar Menggunakan Algoritma K-Means dan FP-Growth pada PT. Citra Mustika Pandawa: Market Basket Analysis with K-Means and FP-Growth Algorithm as Citra Mustika Pandawa Company,” *MALCOM*, vol. 1, no. 1, pp. 41–46, Mar. 2021, doi: 10.57152/malcom.v1i1.62.

[3] B. N. Ruchjana, H. Khoirunnisa, I. Irianingsih, and B. Suhandi, “Perbandingan Penerapan Metode Agglomerative dengan Metode K-Means pada Data Curah Hujan di Wilayah Bogor,” *Kubik*, vol. 5, no. 2, pp. 71–82, Nov. 2020, doi: 10.15575/kubik.v5i2.7581.

[4] I. Musdalifah and A. Jananto, “Analisis Perbandingan Algoritma Apriori Dan FP-Growth Dalam Pembentukan Pola Asosiasi Keranjang Belanja Pelanggan,” *Progresif J. Ilmi. Kom*, vol. 18, no. 2, p. 175, Jul. 2022, doi: 10.35889/progresif.v18i2.878.

[5] R. Sibarani, “ALGORITHMA K-MEANS CLUSTERING STRATEGI PEMASARAN PENERIMAAN MAHASISSWA BARU UNIVERSITAS SATYA NEGARA

INDONDESIA [ALGORITHMA K-MEANS CLUSTERING STRATEGY MARKETING ADMISSION UNIVERSITAS SATYA NEGARA INDONESIA],” 2018.

[6] R. Rachman, “Penentuan Pola Penjualan Media Edukasi dengan Menggunakan Metode Algoritme Apriori dan FP-Growth,” *Jurnal Sistem Informasi, Teknik Informatika, Software Engineering, dan Multimedia*, vol. 23, no. 1, Mar. 2021, doi: 10.31294/p.v23i1.9884.

[7] S. I. Murpratiwi, I. G. Agung Indrawan, and A. Aranta, “ANALISIS PEMILIHAN CLUSTER OPTIMAL DALAM SEGMENTASI PELANGGAN TOKO RETAIL,” *j. pendidik. teknologi. kejuruan.*, vol. 18, no. 2, p. 152, Sep. 2021, doi: 10.23887/jptk-undiksha.v18i2.37426.

[8] E. P. Priambodo and A. Jananto, “Perbandingan Analisis Cluster Algoritma K-Means Dan AHC Dalam Perencanaan Persediaan Barang Pada Perusahaan Manufaktur,” vol. 18, no. 2.

[9] N. A. Wulandari, H. Pratiwi, and S. S. Handayani, “Perbandingan Metode K-means and Agglomerative Nesting untuk Clustering Data Digital Marketing di Twitter,” vol. 2, 2023.

[10] R. Rachman and N. Hunaifi, “Penerapan Metode Algoritma Apriori dan FP-Tree Pada Penentuan Pola Pembelian Obat,” *Jurnal Sistem Informasi, Teknik Informatika, Software Engineering, dan Multimedia*, vol. 22, no. 2, pp. 175–182, Sep. 2020, doi: 10.31294/p.v22i2.8258.

[11] J. Wu, *Advances in K-means clustering: a data mining thinking*. in Springer theses. Heidelberg Berlin: Springer, 2012.

[12] S. Miyamoto, *Theory of agglomerative hierarchical clustering*. Singapore: Springer, 2022.