



Contents lists available at [JESA](https://jesa.aks.or.id/index.php/jesa/index)

Journal of Engineering and Science Application

journal homepage:
<https://jesa.aks.or.id/index.php/jesa/index>



Application of the K-Means Algorithm for the Grouping of Regional Income Patterns in Kudus Regency

Alif Miftachul Nasikhah^a, Ade Ima Afifa Himayati^b, Findasari^c

¹²³S1 Mathematics Study Program, Faculty of Science and Technology, University of Muhammadiyah Kudus

Article Info

Keywords:

K-Means
clustering
Regional Revenue
RapidMiner
Kudus Regency

ABSTRACT (10 PT)

Regional revenue is one of the indicators of regional financial ability that needs to be analyzed to determine the realization of revenue. Analyses that are still descriptive have not been able to group data based on similar characteristics. This study aims to apply the K-Means algorithm to classify the regional income pattern of Kudus Regency based on monthly income realization data. The research uses a quantitative approach with secondary data in the form of the realization of regional revenue in Kudus Regency in 2020–2024 obtained from the Regional Revenue, Finance, and Asset Management Agency (BPPKAD) of Kudus Regency. Data processing is carried out using the RapidMiner application through the Read Excel, Set Role, Normalize, K-Means Clustering, and Performance stages. The number of clusters is set to three ($k = 3$). The results of the study showed that the K-Means algorithm succeeded in grouping data into three *Cluster 0* clusters consisting of 4 data, namely September, October, November, and December. *Cluster 1* consists of 1 data, namely August, while *Cluster 2* consists of 7 data, namely January, February, March, April, May, June, and July. Centroid analysis showed that each cluster had different characteristics, while evaluation using Performance Vector yielded a Davies-Bouldin Index value of -0.444 which showed good grouping results based on the RapidMiner evaluation. The results of this study are expected to be supporting information in the evaluation and planning of regional revenue management in Kudus Regency.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Alif Miftachul Nasikhah
Department of Mathematics
University of Muhammadiyah Kudus
Kudus, Central Java
Email: 42022130009@std.umku.ac.id

1. INTRODUCTION

Regional revenue is one of the important signs that shows the financial ability and level of independence of local governments in carrying out public service tasks and economic development [1]. Regional income that has been achieved each period has different characteristics, both in terms of the amount and the way it is achieved in each region. These differences can be seen in an orderly way through

grouping data based on their properties or characteristics. For example, some regions can be divided into low, medium, or high income categories based on the amount of income they earn [2].

In its management, regional revenues have different characteristics, both in terms of the amount and the way they are realized. Based on the level of realization and the share given to the total regional revenue, regional revenue is divided into three categories, namely low, medium, and high [3]. Low regional income is shown by a realization value that is not so large and often changes, while moderate regional income shows a fairly fixed realization with a medium role. High regional income has a large realization value and plays a dominant role in total regional income [4]. The difference in these characteristics needs to be looked at carefully so that regional income patterns can be understood more clearly.

Kudus Regency, which is one of the districts in Central Java Province, has a diverse regional revenue structure and continues to change over time, especially when viewed from the monthly revenue data that has been achieved. Until now, the analysis of regional revenues is still descriptive, so it has not been possible to systematically classify revenues based on the same pattern. This causes information about income groups that have similar characteristics to not be optimally utilized in regional financial evaluation and planning efforts.

Analysis of regional revenues that only describes in a descriptive manner often cannot reveal hidden patterns in complex data. Therefore, a better way of analyzing data is needed to classify regional revenue data based on the characteristics of the value that has been achieved. *Clustering*, as part of data mining, is an analysis technique that is suitable for use because it can group data based on its similarity of properties without the need for initial label information [5].

Clustering is a method of data mining analysis that aims to group data into several small groups (*Clusters*). The goal is for the data within one group to have a high similarity, while the data between the groups has a marked difference [6]. Clustering techniques are often used in various fields, such as regional economics and finance, because they can find patterns hidden in numerical data. *Clustering* is very suitable for analyzing regional income data that has a considerable variation in value and is variable [7]. *Clustering* is *unsupervised learning*, meaning that grouping is done without using predetermined labels or categories. This technique is very suitable for grouping regional income data that has diverse values and has not been divided into previous categories.

The *K-Means algorithm* is one of the data grouping techniques that is often used in scientific research because it has good efficiency and is able to group data in the form of large numbers based on their similarity [1]. The *K-Means* algorithm is often used in regional financial analysis, for example to segment regional tax revenues, segment regional budgets, and map economic conditions in various regions [8]. Studies on the application of *K-Means* to financial data show that this method is able to recognize similar patterns and data groups, such as in the grouping of tax revenue types in regions [9]. In this study, the *K-Means* algorithm was used as a method to group the regional income pattern of Kudus Regency based on monthly income data into three groups, namely low, medium, and high income.

Previous research has shown that the *K-Means* algorithm is used to map economic conditions in various regions by grouping districts or cities based on the value of economic indicators such as income or Gross Regional Domestic Product (GDP), resulting in a structured mapping of regional conditions [10]. This grouping helps to recognize differences in economic characteristics in different regions and can support better policy-making.

In addition, the *K-Means* algorithm is also used in other types of regional data, such as in taxpayer classification to help with regional revenue management strategies [11] and regional budget data grouping to make the budget planning process easier [12]. These findings show that the *K-Means* algorithm is very relevant to be used in the research on the grouping of regional income patterns in Kudus Regency, which is the main focus of this study.

Based on this explanation, this research was carried out because there is a need to recognize regional income patterns in a more structured manner using data mining techniques, especially the *K-Means* algorithm. This study uses the *K-Means* algorithm to group the regional income pattern of Kudus Regency into three categories, namely low, medium, and high, based on monthly income realization data that has been recorded. Therefore, the results of this study are expected to provide an overview of regional revenue patterns, as well as become a reference in the management and planning of regional revenue in Kudus Regency.

2. RESEARCH METHOD

This research is carried out through several stages that are systematically arranged. Each stage is carried out to support the achievement of research objectives, ranging from data collection to analysis of results. For more details the flow is depicted in the following image:

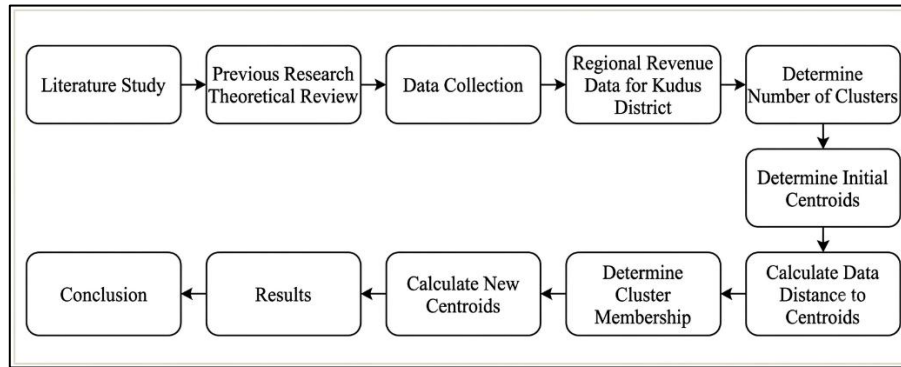


Figure 1. Research Stages

This research was carried out in Kudus Regency, Central Java Province, using secondary data in the form of monthly regional revenue realization data for the period January 2020 to December 2024 obtained from the Regional Revenue, Finance, and Asset Management Agency (BPPKAD) of Kudus Regency. The population in this study is all data on the realization of regional income in Kudus Regency during that period. The sampling technique used is a saturated sample (total sampling), so that all data is used as a research sample. Thus, all regional revenue realization data for the 2020–2024 period is used as research data for the *clustering* process using the K-Means algorithm to group regional revenue patterns into three categories, namely low, medium, and high. Here is an overview of the data that has been collected:

Table 1. Revenue Realization Data

Revenue Realization of BPPKAD Kudus Regency					
month	Year				
	2020	2021	2022	2023	2024
1	8.450.357.684	7.119.391.634	8.617.864.922	11.834.812.594	13.654.347.478
2	16.319.875.107	9.604.164.074	8.816.450.563	12.098.598.327	25.654.836.176
3	24.442.466.758	11.684.495.316	13.727.321.066	13.007.949.446	41.449.489.228
4	30.633.136.472	17.188.554.579	12.619.123.514	11.122.835.621	52.814.036.508
5	39.187.163.459	8.602.222.131	16.179.372.786	15.474.280.546	68.273.610.000
6	47.337.999.688	11.933.400.093	14.272.846.099	19.759.345.323	85.445.201.536
7	61.220.247.211	8.313.950.276	16.454.192.238	17.772.773.621	112.592.557.371
8	38.721.685.912	106.604.150.865	24.187.388.431	29.378.424.163	141.936.437.571
9	62.746.522.451	117.018.652.820	15.159.016.440	14.572.457.211	155.749.983.633
10	113.936.387.157	128.671.986.003	12.232.026.166	12.863.050.442	170.327.873.446
11	113.844.484.473	138.766.930.199	14.057.050.430	12.984.137.767	183.767.618.642
12	126.240.575.956	114.330.256.306	11.339.784.384	12.915.003.098	200.191.740.795

The data analysis method in this study uses a *clustering* technique with the *K-Means* algorithm to group the pattern of regional revenue realization in Kudus Regency based on monthly income data for the period January 2020–December 2024. The analysis stage starts from collecting regional revenue realization data obtained from the local government's official report, then preprocessing the data is carried out through completeness checks, format adjustments, and data processing so that it is ready for analysis. Next, the number of *clusters* in three categories is determined, namely low, medium, and high income, then the initial *centroid* is determined as the central point of each *cluster*. Each data is calculated in its distance to the *centroid* using the Euclidean Distance method, then the data is grouped into *clusters* that have the closest distance. After the grouping process is complete, new *centroids* are calculated based on the average value of the members in each *cluster*. This process is carried out iteratively until no more changes in *cluster* members or *centroids* have reached a convergent state [13].

Determination of the number of Cluster (k) [14]:

$$WCSS = \sum_{j=1}^k \sum_{i=1}^{n_j} \|x_i - C_j\| \quad (1)$$

Where:

x_i = data i , with $i = 1, 2, 3, \dots, n$

C_j = Centroid cluster to J , with $j = 1, 2, 3, \dots, k$

n_j = amount of data in the J cluster, with $j = 1, 2, 3, \dots, k$

k = number of clusters

$\|x_i - C_j\|$ = The distance between the data and *the centroid cluster*

Determining *the initial* centroid [15]:

$$C = \{c_1, c_2, c_3, \dots, c_k\} \quad (2)$$

Where:

C = Initial *centroid* assembly

C_j = *Centroid cluster* Ke-J

k = Number of clusters

Calculating the distance of data to *the centroid* [16]:

$$d(x_i, C_j) = \sqrt{\sum_{k=1}^p (x_{ik} - C_{jk})^2} \quad (3)$$

Where:

x_i = Data to i

C_j = *Centroid cluster* Ke-J

k = Number of variables (from 1, 2, 3, ..., p)

x_{ik} = The data value on the k th variable

C_{jk} = centroid value on the k variable

If the data has only 1 variable (e.g. income value), then the formula becomes:

$$d(x_i, C_j) = |x_i - C_j| \quad (4)$$

Defining *cluster membership* [17]:

$$d(x_i, C_j) = \min\{d(x_i, C_1), d(x_i, C_2), \dots, d(x_i, C_k)\} \quad (5)$$

Where:

x_i = Data to i

C_j = *Centroid cluster* Ke-J

$C_1, C_2, \dots, C_k$ = *the first, second, to last cluster*

Calculating *the new* centroid [18]:

$$C_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i \quad (6)$$

Where:

C_j = *New centroids* in the J th cluster

n_j = The amount of data in the J cluster

x_i = The I-I data in the J-Cluster

3. RESULTS AND DISCUSSION

This section presents the results of the application *of the K-Means algorithm* in grouping the regional revenue pattern of Kudus Regency based on monthly income realization data for the 2020–2024 period. The results obtained include the data processing process, cluster formation, *centroid value* of each *cluster*, and interpretation of the characteristics of each income group. Furthermore, the results of the grouping are explained to describe regional revenue patterns and discuss their effectiveness in supporting decision-making in the field of regional financial management.

3.1 Research Data

This study uses secondary data obtained from the Regional Revenue, Finance, and Asset Management Agency (BPPKAD) of Kudus Regency. The data used is data on the regional income of Kudus Regency which has been realized during the period from 2020 to 2024. The data was chosen because it was able to show the pattern of changes in regional income over five years, so that it was in accordance with the purpose of the study to group regional income patterns using the *K-Means method*.

The research data was made based on the month period, from January to December. Thus, the total objects analyzed were 12 observation data, while the variables used consisted of five numerical variables, namely the realization of regional revenue for 2020, 2021, 2022, 2023, and 2024. In addition, there is one variable on the Moon that acts as an identity to distinguish each data object.

Before the grouping process is carried out, the data first goes through the data preparation stage. This stage aims to ensure that the data used is of high quality, so that it can produce the best grouping. The check is done to find out if there is incomplete data, repetitive data, or inappropriate data types. Based on the results of the examination, all the data provided is considered complete and can be used for research purposes without having to delete or add anything.

The research data used in this study is presented in table 2:

Table 2. Regional Revenue Data for 2020-2024

Revenue Realization of BPPKAD Kudus Regency					
Month	Year				
	2020	2021	...	...	2024
January	8.450.357.684	7.119.391.634	...	...	13.654.347.478
February	16.319.875.107	9.604.164.074	...	...	25.654.836.176
March	24.442.466.758	11.684.495.316	...	...	41.449.489.228
April	30.633.136.472	17.188.554.579	...	...	52.814.036.508
May	39.187.163.459	8.602.222.131	...	...	68.273.610.000
June	47.337.999.688	11.933.400.093	...	...	85.445.201.536
July	61.220.247.211	8.313.950.276	...	...	112.592.557.371
August	38.721.685.912	106.604.150.865	...	...	141.936.437.571
September	62.746.522.451	117.018.652.820	...	...	155.749.983.633
October	113.936.387.157	128.671.986.003	...	...	170.327.873.446
November	113.844.484.473	138.766.930.199	...	...	183.767.618.642
December	126.240.575.956	114.330.256.306	...	...	200.191.740.795

Based on Table 1, it can be seen that the value of regional income achieved every month varies during the research period. These differences show that the characteristics of regional income vary from month to month. Some months result in higher revenues, while others are lower. These differences in properties are the basis for using the *K-Means* algorithm to group data based on the degree of similarity between objects.

In this study, each month is considered as one object, while the value of income achieved from 2020 to 2024 is used as a variable in the grouping process. The *K-Means* algorithm calculates how similar two objects are by comparing their variable values using Euclidean Distance, so that objects that have similar income patterns will be grouped into the same *cluster*.

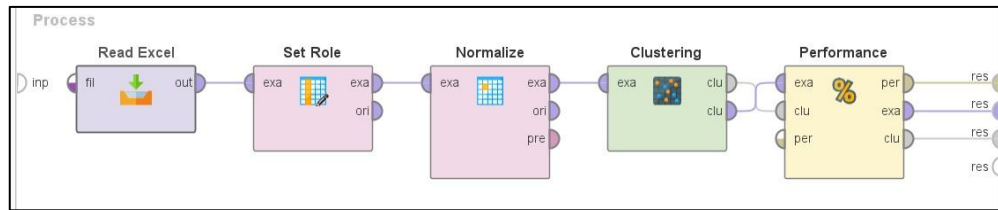
Before the grouping process begins, the data also goes through the normalization stage using the Z-Transformation method. This stage aims for all variables to have the same scale, so that no one particular variable is too dominant in the distance calculation. Therefore, all variables make a balanced contribution to the grouping results.

Based on the purpose of the study, the number of *clusters* used was three *clusters* with a value of k equal to 3. The value of $k = 3$ was chosen because the purpose of the study wanted to divide the regional income pattern into three groups that have certain characteristics. The results of the grouping will then be analyzed by paying attention to the value of the group center and the characteristics of each group, so that it can provide information about the regional income pattern of Kudus Regency during the research period.

3.2 Implementation of the *K-Means* Algorithm Using RapidMiner

After the research data has been obtained and prepared, the next step is to apply the *K-Means* algorithm through the Altair AI Studio (*RapidMiner*) application. This implementation aims to group data on the realization of regional income in Kudus Regency based on the similarity of data characteristics during the period from 2020 to 2024. Grouping is carried out automatically by the application using the *K-Means* algorithm, so that the analysis process is faster and produces data groups that have similar characteristics.

Before the grouping process begins, the data must go through a pre-processing stage so that the data can be properly processed by the system. These stages include the process of importing data, determining the role of each variable (*Set Role*), normalizing data (*Normalize*), grouping process using *K-Means* operators, and evaluating results with *Cluster Distance Performance* operators. All of these stages are organized in a *workflow* in Altair AI Studio (*RapidMiner*) so that each operator is connected to each other and forms an organized data processing process.

Figure 2. *K-Means Algorithm Implementation Workflow*

The process begins by using the Read Excel operator which aims to import research data into Altair AI Studio, also known as RapidMiner. Next, it uses the Set Role operator to assign the Moon variable as the ID, so that the variable only functions as a data identity and is not involved in the calculation process of the *K-Means* algorithm.

The next step is *Normalize*, which is the process of converting data into a more standard form using the Z-Transformation method. Normalization is carried out because the amount of regional income varies every year. If the data is not normalized, variables with larger values will further affect the distance calculation, which can change the grouping results. With normalization, all variables have the same scale so that each variable makes a balanced contribution to the formation of the group.

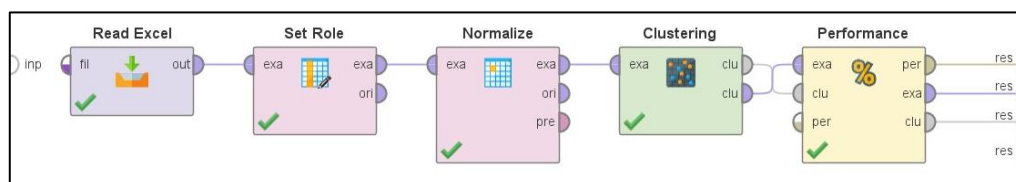
After the normalization process is completed, the data is then processed using the *Clustering* method, namely *K-Means*. In this study, the number of *clusters* was set as three ($k = 3$). The selection of the number of groups was carried out based on the purpose of the research, which was to group regional income patterns into three categories, then interpreted as low, middle, and high income groups based on the results of the group center analysis and the characteristics of the data in each group.

In the *clustering* stage, the *K-Means* algorithm works by determining the initial central point, then calculating the distance of each data to that central point using Euclidean distances. Next, each data is placed in the group that is the closest distance. Once all the data is grouped into clusters, the algorithm calculates the new *cluster* centers by taking the average of all members in each *cluster*. The process is repeated continuously (iteration) until the position of the center of the group no longer changes or until it reaches the maximum limit of the specified number of iterations.

The last stage is to evaluate the *clustering* results using the *Cluster Distance Performance* operator. This operator is used to evaluate the quality of grouping by paying attention to the distance between members in one group as well as the distance between different groups. The results of the evaluation are used as a reference to determine whether the grouping process has succeeded in creating data groups that have high similarities in them and clear differences from other groups.

3.3 Import Research Data

The first step in running the *K-Means* algorithm is to feed the research data into the *Altair AI Studio (RapidMiner)* application. The research data that has been compiled in *Microsoft Excel (.xlsx)* format is imported using the Read Excel operator. This operator is used to read information from Excel files so that it can proceed to the next stage. The process of importing data is carried out by choosing a place to store the file, choosing the worksheet to use, and ensuring that the data type in each column is correct. After the import process is complete, all the research data successfully appears in the form of an *ExampleSet* and is ready to be used for the next processing step.

Figure 3. *Import Data Using the Read Excel Operator*

After the research data is obtained, the next stage is to implement the *K-Means* algorithm using the Altair AI Studio (*RapidMiner*) application. The implementation begins with importing data using the *Read Excel* operator. This stage is carried out so that the regional revenue realization data that has been compiled in Microsoft Excel format can be read and recognized by the system as a dataset that is ready to be processed. The data import process also aims to ensure that all variables and their data types are appropriate so that there are no errors in the next analysis process. The success of this process is shown

by all the data that has been successfully displayed in the form of an *ExampleSet*, so that the data is ready to be used as input in the processing process.

Next, variable settings are carried out using the *Set Role* operator. This stage aims to determine the function of each variable in the *clustering* process. In this study, the *Moon* variable was designated as the *ID variable (Identifier)* because it only functions as the identity of each research object and has no influence on the distance calculation process. Meanwhile, the regional revenue realization variable from 2020 to 2024 will continue to be used as a numerical variable because it is the basis for the grouping process. This variable is set so that the *K-Means* algorithm only processes data that truly represents the characteristics of regional income so that the grouping results become more accurate.

The next stage is to normalize the data using the *Normalize* operator with the *Z-Transformation* method. Normalization is carried out because regional revenue realization data has a different range of values for each variable. The difference in the range of values can cause variables with larger values to dominate the distance calculation process, thus affecting the *clustering* results. Therefore, normalization is necessary to equalize the scale of all variables so that each variable makes a balanced contribution in the process of *cluster formation*. With the data that has been normalized, the measurement of similarity between objects becomes more objective because it is not affected by the difference in the magnitude of the value in each variable.

After the normalization process is completed, data grouping is carried out using the *K-Means* operator. In this study, the number of *clusters* was determined as many as three ($k = 3$) in accordance with the purpose of the study, which was to group the regional income pattern of Kudus Regency into three groups based on the level of similarity of data characteristics. The *clustering process* is carried out by calculating the distance of each object to the *centroid* using the *Euclidean Distance* method. Next, each object will be placed in the *cluster* that has the closest distance. Once all objects acquire *clusters*, *centroids* are updated based on the average value of the members in each *cluster*. The process is carried out repeatedly until the position of the *centroid* has not changed again or has reached a convergent state. This stage is at the heart of the research because it produces data groups that have similar regional income characteristics.

After the *clustering* process is complete, the last stage is to evaluate the grouping results using the *Cluster Distance Performance operator*. Evaluation was carried out to determine the quality of the *clusters* generated by the *K-Means algorithm*. The assessment was carried out based on the proximity of each member to their *respective centroid cluster*. The smaller the average distance value produced, the higher the level of similarity between the members in the *cluster*, thus indicating that the grouping process has gone well. The results of this evaluation are then used to support the interpretation of the results of the *Cluster Model*, *ExampleSet*, and *Centroid Table*, so that it can provide a more comprehensive picture of the regional income pattern of Kudus Regency formed from the *clustering process*.

3.4 Clustering Results

The *clustering* process using the *K-Means* algorithm produces several *outputs* that are the basis for the analysis of research results. The following are the results of *clustering* using the Altair AI Studio (RapidMiner) application, which consists of *Cluster Model*, *ExampleSet*, *Centroid Table*, and *Performance Vector*. Each *output* provides different information about the results of data grouping, starting from the number of clusters formed, the membership of each data in each *cluster*, the value of the *centroid* as the center of the *cluster*, to the evaluation of the quality of the grouping results. Analysis of all outputs was carried out to find out the characteristics of each *cluster* so that it could provide an overview of the regional income pattern of Kudus Regency based on the level of similarity of data.

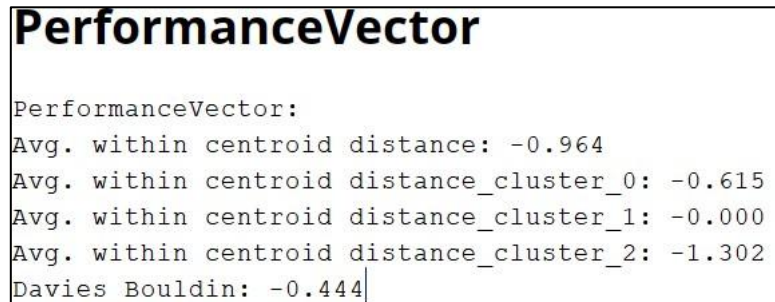
The process of grouping the regional income data of Kudus Regency using the *K-Means* algorithm resulted in a group model that shows how many members there are in each group. These results were obtained through a grouping process using the *RapidMiner* application with a specified number of groups (k) of three. The *Cluster Output* model serves to provide information about the amount of data that has been successfully grouped into each *cluster*.

Cluster Model

```
Cluster 0: 4 items  
Cluster 1: 1 items  
Cluster 2: 7 items  
Total number of items: 12
```

Figure 3. *Cluster Model*

Figure 3 shows that *Cluster 2* has the largest number of members, namely 7 data (58.33%), while *Cluster 0* consists of 4 data (33.33%) and *Cluster 1* only 1 data (8.33%). The difference in the number of members shows that most of the regional income data of Kudus Regency has a similar pattern so that they are in the same group. Meanwhile, clusters with fewer members show different data characteristics. However, the naming of *Clusters 0, 1, and 2* is only an automatic label from the *K-Means* algorithm, not an indication of the level of income. The low, medium, or high category will be determined based on the *centroid* value of each *cluster*.



Gambar 4. Performance Vector

The results of the *Performance Vector* evaluation showed that the *K-Means algorithm* was able to group the regional income data of Kudus Regency based on the level of similarity of patterns. The mean value of the distance of the data to the *centroid* of -0.964 indicates that most of the data has proximity to the center of the respective cluster. *Cluster 2* has the smallest average distance, which is -1.302, so its members have the closest pattern to the *centroid* than the other clusters. *Cluster 0* has a value of -0.615, while *Cluster 1* is -0.000 because it consists of only one data. In addition, the Davies-Bouldin Index (DBI) value of -0.444 shows that the grouping results are of fairly good quality because they can be clearly distinguished between clusters. Thus, the *K-Means method* succeeded in dividing the regional income data of Kudus Regency into groups based on the similarity of their characteristics.

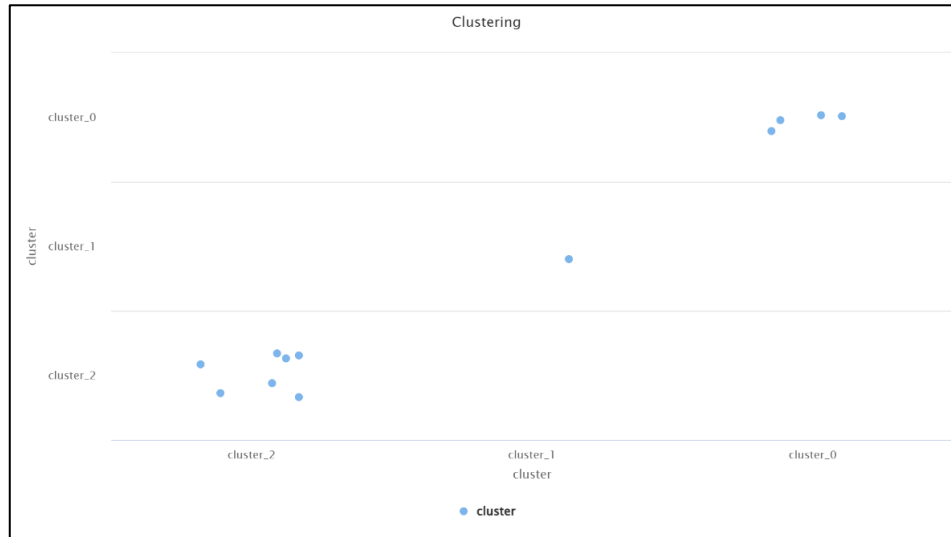


Figure 5. Plot Clustering

The visualization of the clustering results shows that the *K-Means algorithm* has succeeded in forming three data groups, namely *cluster 0*, *cluster 1*, and *cluster 2*. Each point on the graph depicts observation data based on the month, with the position of the points indicating proximity to a particular cluster. *Cluster 2* has the largest number of members, namely 7 data, so it looks like a denser group because it has relatively similar income characteristics. Meanwhile, *cluster 0* consists of 4 data with different patterns than *cluster 2*. *Cluster 1* only has 1 data that is visible as a separate point, indicating the existence of periods with the most different income characteristics. Overall, the visualization shows a fairly clear separation between clusters without much overlap, thus reinforcing the results that the *K-Means algorithm* is able to group the revenue data of the Kudus Regency region based on the similarity of patterns.

The results of the study show that *the K-Means algorithm* is able to group the regional income data of Kudus Regency into three clusters based on the level of similarity of data characteristics. The grouping process is carried out after the normalization stage so that each variable has the same scale, so that *the clustering results* are not affected by large differences in values between variables. Based on the results of processing using *RapidMiner*, *Cluster 2* has the largest number of members, namely seven data, followed by *Cluster 0* with four data, and *Cluster 1* with one data. The dominance of *Cluster 2* shows that most regional income data has almost the same pattern, while *Cluster 1* shows the existence of a single period with significantly different income characteristics. The difference in the number of members in each *cluster* illustrates the variation in regional income patterns that can be separated based on their level of similarity.

Centroid value analysis shows that each *cluster* has different revenue characteristics. *Cluster 0* has a relatively higher central value on several variables so that it shows a greater income tendency than the other groups. *Cluster 2* describes the most common revenue patterns because it has the most members, while *Cluster 1* is a special group because it consists of only one data with different characteristics. The results of the evaluation using *Performance Vector* with a *Davies-Bouldin Index* (DBI) value of -0.444 showed that the grouping process had resulted in a fairly good *cluster* separation. The visualization of *the clustering results* also shows that each group is clearly formed. Thus, *the K-Means algorithm* can be used as an analysis method to recognize the pattern of regional income in Kudus Regency and provide information that can support the evaluation and planning of regional revenue management.

4. CONCLUSION

Based on the results of the study, *the K-Means algorithm* was successfully applied to classify the regional income patterns of Kudus Regency using *the RapidMiner* application. The *clustering process* with a number of 3 *clusters* ($k = 3$) resulted in three *clusters of Cluster 0* consisting of 4 data, namely September, October, November, and December. *Cluster 1* consists of 1 data, namely August, while *Cluster 2* consists of 7 data, namely January, February, March, April, May, June, and July. data. The results of *centroid* analysis showed that each *cluster* had different characteristics, while evaluation using *Performance Vector* showed that the grouping results were of good quality. This study shows that *the K-Means algorithm* is able to group regional income data based on the level of similarity of its characteristics. The results of the grouping can be used as supporting information for the Kudus Regency Government in understanding regional revenue patterns and as a consideration in evaluating and planning regional revenue management in a more targeted manner.

REFERENCES

- [1] G. R. Wahyudi, Rahmaddeni, E. Dini, S. Adrianto, and R. Fadila, "Clustering Regencies in Indonesia for Regional Mapping Using the K-Means Algorithm Pengelompokan Kabupaten di Indonesia untuk Pemetaan Daerah Menggunakan Algoritma K-Means," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. 3, pp. 1143–1151, 2025.
- [2] A. D. Sofia, "Pengelompokan Kabupaten / Kota berdasarkan Indikator Indeks Pembangunan Manusia 2022 Menggunakan K-Harmonic Means Clustering," *J. Ris. Statistika*, vol. 3, no. 2, pp. 163–172, 2023.
- [3] H. Sutrisno, J. Junaedi, and A. Ekowati, "Analisis Kontribusi Pajak Daerah terhadap Pendapatan Asli Daerah Kabupaten Nganjuk," *J. Public Power*, vol. 7, no. 2, pp. 101–117, 2023, doi: 10.32492/jpp.v7i2.7206.
- [4] N. W. Tyas, L. Kristiyanti, and Darmanto, "Analisis Kontribusi dan Efektivitas Pajak dan Retribusi Daerah terhadap Pendapatan Asli Daerah: Studi Kasus di BPKPAD Kota Sukoharjo," *J. Ilm. Keuang. Akunt. Bisnis*, vol. 4, no. 1, pp. 684–690, 2025, doi: 10.53088/jikab.v4i1.117.
- [5] M. A. Hakim, A. B. Prasetyo, and D. Eridani, "Penerapan Data Mining Dengan Algoritma K-Means Clustering Penyewaan Alat-Alat Event Pada Studi Kasus Cv. Dipo Rental Creativindo Implementation of Data Mining Using the K-Means Clustering Event Rental Algorithm in a Case Study of Cv. Dipo Rental Creativindo," *J. Tek. Komput.*, vol. 1, no. 4, pp. 148–155, 2023, doi: 10.14710/jtk.v1i4.37011.
- [6] N. A. Maori, "METODE ELBOW DALAM OPTIMASI JUMLAH CLUSTER PADA K-MEANS CLUSTERING," *J. SIMETRIS*, vol. 14, no. 2, pp. 277–287, 2023.
- [7] M. Faisal, J. I. Saputro, W. S. Utami, and S. Suharmanto, "Model Analisis Clustering KMeans Untuk Pengelompokan Pendapatan Daerah Berdasarkan Jenis Pekerjaan," *ICIT J.*, vol. 12, no. 1, pp. 13–24, 2026, doi: 10.33050/icit.v12i1.4139.
- [8] A. Hasibuan, D. R. Kembuan, C. T. M. Manoppo, and M. T. Hermanto, "Optimization of k-means algorithm in grouping data using the statistical gap method," *J. Intell. Decis. Support Syst.*, vol. 6, no. 3, pp. 112–120, 2023.
- [9] M. Zaki, N. Ibrahim, R. R. Marlina, P. Statistika, F. Matematika, and P. Alam, "Perbandingan K-Means Clustering dan Fuzzy C-Means Clustering untuk Mengelompokkan Kabupaten / Kota Berdasarkan Faktor Tingkat Pengangguran di Jawa Barat Tahun 2023," *Bandung Conf. Ser.*, vol. 1, no. 5, p. 2, 2023.
- [10] M. Amelia, A. Faqih, and A. R. Rinaldi, "Penerapan Metode K-Means Clustering Dalam Pemetaan Kemiskinan Kabupaten/Kota Di Indonesia Untuk Perencanaan Kebijakan Yang Tepat," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 2, p. 6231, 2025, doi: 10.23960/jitet.v13i2.6231.
- [11] N. N. Saragih and R. Kurniawan, "(Journal of Computer Engineering, System and Science)," *J. Comput. Eng. Syst. Sci.*, vol. 10, no. 1, pp. 299–311, 2025.
- [12] G. M. M. Sujak, H. N. Rofiq, and F. I. Tawakal, "Implementation of K-Means Clustering for Optimizing Non-Communicable Disease Budgets," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. 1, pp. 67–74, 2025.
- [13] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
- [14] D. J. Tambunan et al., "Analisis Algoritma K-Means untuk Pengelompokan Tingkat Kemiskinan di Kota Medan," *J. Pendidik.*

- Tambusai*, vol. 10, no. 1, pp. 3990–3996, 2026.
- [15] M. Ahmed, R. Seraj, and S. M. S. Islam, “The k-means algorithm: A comprehensive survey and performance evaluation,” *Electron.*, vol. 9, no. 8, pp. 1–12, 2020, doi: 10.3390/electronics9081295.
- [16] B. Hartono, S. Eniyati, and K. Hadiono, “Perbandingan Metode Perhitungan Jarak pada Nilai Centroid dan Pengelompokan Data Menggunakan K-Means Clustering,” *J. Sist. Komput. dan Inform.*, vol. 4, no. 3, p. 503, 2023, doi: 10.30865/json.v4i3.6021.
- [17] A. Vouros, S. Langdell, M. Croucher, and E. Vasilaki, “An empirical comparison between stochastic and deterministic centroid initialisation for K - means variations,” *Mach. Learn.*, vol. 110, no. 8, pp. 1975–2003, 2021, doi: 10.1007/s10994-021-06021-7.
- [18] M. F. Akbar and L. Zahrotun, “K-Means Centroid Optimization with Genetic Algorithm for Clustering Micro , Small , Medium Enterprises in Yogyakarta,” *JUITA J. Inform.*, vol. 13, no. 2, pp. 87–97, 2025.