Comparison Of K-Means, K-Medoids, and Fuzzy C-Means Algorithms for Clustering Drug User's Addiction Levels

Annisa Nadaa Shabrina
Information System, Faculty of Science and Technology
UIN Suska Riau
Pekanbaru, Riau
11950324528@students.uin-suska.ac.id

M. Afdal
Information System, Faculty of Science and Technology
UIN Suska Riau
Pekanbaru, Riau
m.afdal@uin-suska.ac.id

Siti Monalisa
Information System, Faculty of Science and Technology
UIN Suska Riau
Pekanbaru, Riau
siti.monalisa@uin-suska.ac.id

Abstract—Narcotics, psychotropics, and addictive substances are drugs that can activate brain systems, affect dopamine levels, and cause addiction. In Indonesia, a law requires drug addicts to receive treatment and care. To properly treat a drug addict, it is first necessary to determine the level of addiction. Data mining methods such as clustering can be used to assess a user's level of drug addiction. This study uses the clustering algorithms Fuzzy C-means, K-Medoids, and K-means. The performance of the three clustering algorithms will then be evaluated based on the average similarity of clusters. Data such as how many types of drugs that used, the length of time they were used, the psychiatric status, and the physical condition status are used. Clustering was accomplished using the data mining software RStudio. The clustering algorithms were then evaluated with the Davies Bouldin Index (DBI). Based on the analysis results, the K-Medoids algorithm was found to have the best average similarity value of cluster for determining drug users' addiction levels.

Keywords—Drugs, Clustering, Fuzzy C-Means, K-Means, K-Medoids, DBI, RStudio

I. Introduction

The number of drug cases in 2021 reached 41.084 cases, with a total of 53.405 suspects, indicating abuse and illicit drug trafficking [1]. BNN (National Narcotics Agency of Indonesia) seized at least 20 different types of drugs in these cases. The number of drug cases can also be seen in the number of people assisted in drug cases in various correctional institutions in Indonesia, with 1.296 drug cases as producers, 18.579 as dealers, 3.790 as intermediaries, and 21.313 as drug users [2].

In Southeast Asia, Indonesia has some of the strictest drug laws. To begin with, Indonesian law enforcement is allowed to use a "shoot-on-sight" policy against drug distributors and traffickers. Despite this, drug abuse rates are rising, and drug users are being imprisoned. Fortunately, judges can provide drug abusers with rehabilitation programs after their trials. Unsurprisingly, going to jail does not help those struggling with substance abuse; however, rehab can provide much-needed help.

According to Indonesian Law No. 35 of 2009, drug users must complete rehabilitation. Treatment options for rehabilitation include outpatient care, inpatient care, and referrals to additional rehab facilities. To choose the appropriate type of rehabilitation, drug users' addiction level must first be determined [3].

From the given amount of information on drug use, it is crucial to use efficient presentation techniques to ensure that the data recipients receive highly accurate information suitable for their needs. Therefore, creating a data mining algorithm is extremely precise and effective for huge data sets [4].

Clustering analysis is an important data mining technique for locating new pattern data [5]. After the data has been clustered, it can be further analyzed to find any specific data. Cluster analysis serves as the pre-processing technique for all other data mining operations, according to Guedalia et al. [6]. Data objects with similar characteristics are grouped through a process called clustering. Clustering is the process of grouping data objects that have similar characteristics. After clustering, each group will contain objects similar to one another [7]. Using clustering techniques, homogeneous drug users are identified and grouped together. These grouped drug users are then used to determine their level of drug addiction.

K-means is one of the methods used to cluster data. Existing data is divided into one or more clusters or groups, each with its unique characteristics, using this non-hierarchical data clustering technique. Existing data is divided into one or more clusters or groups, each with its unique characteristics, using this non-hierarchical data clustering technique [8]. K-means clustering has several advantages, including low complexity, fast calculations, the capacity to manage huge quantities of data, and the ability to adjust cluster members [9].

A clustering algorithm similar to the K-means algorithm is the K-medoids algorithm. Both partitional algorithms aim to reduce the squared error—the separation between a point labeled as the cluster center and a point labeled as being in the cluster—between two points. Unlike the K-means algorithm, K-medoids choose data points as centers. It is more resistant to noise and outliers than K-means. The point in the data set that is most centrally located is called a "medoid," which is an object in a cluster with the lowest dissimilarity to all the other objects in the cluster [10].

Another clustering technique that is frequently used in numerous clustering studies is Fuzzy C-Means. Due to the requirement that the number of clusters to be formed be predetermined, fuzzy C-Means is a supervised clustering algorithm. The fundamental idea is to identify a cluster center, which symbolizes the typical location of each cluster [11]. Fuzzy C-Means have the advantage of providing more realistic data that is more likely to be included in a cluster and allowing for greater flexibility in cluster assignment [9].

This study used the K-means, K-Medoids, and Fuzzy C-Means algorithms for the clustering process to be compared so that the algorithm with the best average similarity of clusters could be identified based on the results of the cluster validity values with drug users as the source of the dataset in the hope that the cluster results could gauge the level of addiction of drug users.

II. METHODOLOGY

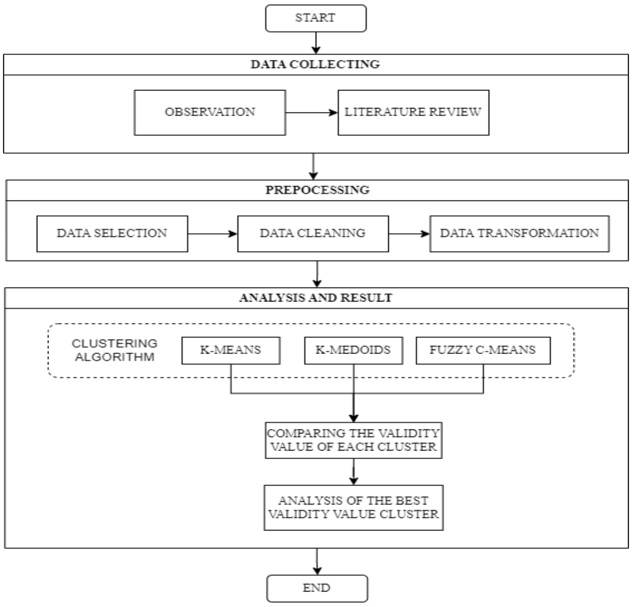

Figure 1. Proposed Methodology for Comparative Analysis

This study employed a quantitative research methodology. The quantitative method is a type of research where each step is completed using numerical data. Data collection and analysis are needed to fully understand the stages of this study.

A. Data Collection

The data set used was obtained from the Riau Province National Narcotics Agency. The data was collected from 357 drug users between 2021 and 2022. The criteria used include; how many types of drugs used, the length of time they were used, the user's psychiatric status, and the user's physical condition status.

Table 1. The Data Set of The Drug Users

No	The Number of Drug Types Used	Length of Use	Psychiatric Status	Physic Status
1	1 Type	1 Year	1	1
2	2 Types	2 Years	3	3
3	2 Types	5 Months	1	1
4	2 Types	3 Years	4	5
5	1 Type	7 Years	7	6
6	2 Types	7 Months	1	1
7	2 Types	6 Years	6	5
		…		
357	1 Type	2 Months	2	0

The following steps were used in the data collection process:

1) *Observation*

Specifically, it is one of the data collection steps that entails making in-person observations of research subjects and examining issues in the field that are directly related to the subject of study, specifically by visiting the Riau Province National Narcotics Agency.

2) *Literature Review*

It’s a step in which sources for data, books, and other materials relevant to writing research reports are found and studied.

B. Pre-processing

The stage of data pre-processing entails preparing and cleaning raw data to eliminate redundancy, incompleteness, and inconsistencies. The procedure used during the pre-processing stage is as follows [11]:

1) *Data Selection*

At this point, the tasks will be completed, including choosing the data used. An item that serves as a benchmark during the grouping calculation is the data type used.

2) *Data Cleaning,*

At this stage, activities include determining whether any data is unclear, lacking, or empty; the issue will be resolved by deleting the data.

3) *Data Transformation*

Data normalization is a step in the data transformation process that aims to use a common scale for the numerical values in the data set without distorting the range of values' variations or losing information.

C. Implementation of Clustering Algorithm

1) *K-Means Algorithm*

The following are the steps taken in implementing the K-Means Algorithm [12]:

a) K points should be added to the space that a cluster of objects represents. These points represent the centroid of the initial group.

b) Each item should be placed in the group with the closest centroid.
c) Recalculate the K centroids' locations once every object has been assigned.
d) Until the centroids stop moving, Steps 2 and 3 must be repeated. To determine the metric that needs to be minimized, the objects are thus divided into groups.

2) K-Medoids Algorithm

The following are the steps taken in implementing the K-Medoids Algorithm [10]:

a) The algorithm starts by randomly choosing k objects as medoids point from n data points (n>k).
b) Choose the K-Medoids point that most accurately represents each data object in the supplied data set. This step defines the similarity using the distance measure, which can be Euclidean, Manhattan, or Minkowski.
c) Pick an object that is not a medoid at random.
d) Calculate the total cost, S, of changing the starting medoids object to O'
e) Replace the original medoids with the new ones if S <0; otherwise, a new set of medoids will be created.
f) Steps 2-5 should be repeated until the medoids show no change

3) Fuzzy C-Means Algorithm

The following are the steps taken in implementing the Fuzzy C-Means Algorithm [13]:

a) Initialize U= $\left[u_{ij}\right] matrix, U^{(0)}$
b) At k-step: calculate the centers vectors $C^{(k)} = [c_j]$ with $U^{(k)}$
c) Update $U^{(k)}, U^{(k+1)}$

$$c_j = \frac{\sum_{i-1}^{N} u_{ij}^{m} . x_i}{\sum_{i-1}^{N} u_{ij}^{m}} \qquad (1)$$

d) $If \left\| U^{(k+1)} - U^{(k)} \right\| < \varepsilon$ then STOP; otherwise return to step 2.

D. Comparing the Validity of Each Cluster

The next step is to compare the K-means, K-Medoids, and Fuzzy C-Means algorithms by running a cluster validation test to determine which of the three algorithms has the highest cluster validity value.

The Davies-Bouldin index (DBI) is used to evaluate clusters [14]. The system-wide average similarity of each cluster is the separation metric suggested by Davies and Bouldin for crisp clustering. The index is formally:

$$SSW = \frac{1}{m_i} \sum_{j-i}^{m_i} d(x_j, c_i) \qquad (2)$$

The cluster's centroid is represented by c_i, the number of data in the cluster is represented by m_i in cluster $k = e - i$, and $d(x_j, c_i)$ is the Euclidean distance between each data point and the centroid. The sum of squares between clusters (SSB) equation is used to calculate the distance between clusters, and it is calculated as follows [15]:

A represents the cluster $k = e - i$ centroid, $k = e - i$ represents the number of data points in cluster $k = e - i$, and d represents the Euclidean distance between each data point and the centroid. To determine the distance between clusters, the sum of squares between clusters (SSB) equation is used and is calculated as follows:

$$SSB_i = d(x_j, c_i) \quad (3)$$

The ra good cluster. The ratio value is calculated using the equation below:

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \quad (4)$$

Using the following equation, the obtained ratio value is used to determine the Davies Bouldin index (DBI) value:

$$DBI = \frac{1}{K_i} \sum_{i-1}^{k} max_{i \neq j}(R_{i,j}) \quad (5)$$

In the equation above, the variable K represents the number of clusters used. The smaller the DBI value obtained (non-negative >= 0), the better the cluster obtained from the clustering method used [16].

The variable K indicates the number of clusters used in the equation. The better the cluster obtained using the employed clustering method, the smaller the DBI value obtained (non-negative >= 0).

E. Analysis of the Best Validity Value Cluster

Results analysis is carried out in the form of interpretation to review the data with the best cluster results to give meaning to the data, explain descriptive patterns, and produce relevant conclusions. Interpretation is the process of interpreting the results of data analysis, which can be defined as sorting, categorizing, and summarizing data to draw conclusions.

III. RESULT AND DISCUSSION

Pre-processing is done on the research data to raise the data quality. The sample data underwent pre-processing, including data selection, cleaning, and transformation. The results of pre-processing data are as follows:

Table 2. Result of the Data Pre-Processing

No	The Number of Drug Types Used	Length of Use	Psychiatric Status	Physic Status
1	0	0,033333	0,111111	0,111111
2	0,5	0,066667	0,333333	0,333333
3	0,5	0,013667	0,111111	0,111111
4	0,5	0,1	0,444444	0,555556
5	0	0,233333	0,777778	0,666667
6	0,5	0,019333	0,111111	0,111111
7	0,5	0,2	0,666667	0,555556
		…		
357	0	0,005333	0,222222	0

Using the RStudio tool and the following syntax, the clustering process is carried out using the three clustering algorithms K-Means, K-Medoids, and Fuzzy C-Means.

The following is the syntax used in the RStudio to apply the K-means, K-Medoids, and Fuzzy C-Means Algorithm:

K-Means

```
  library(cluster)

  Drug_data.km  <-  eclust(drugs_data,  "kmeans",  k  =  2,
nstart = 1, graph = FALSE)
```

K-Medoids

```
  library(cluster)

  Drugs_data.kmed<- pam(drugs_data, k=2)
```

Fuzzy C-Means

```
library(ppclust)

  Drugs_data.fcm  =  fcm(drugs_data,  centers=2,  iter.max  =
1000, con.val = 1e-05, nstart = 1)
```

The data processed using RStudio also show the Davies-Bouldin Index (DBI) result value as the ideal cluster grouping reference. The clusters with 2 clusters to 5 clusters were used for the test. The outcomes of the data processing are shown in Figure 2.

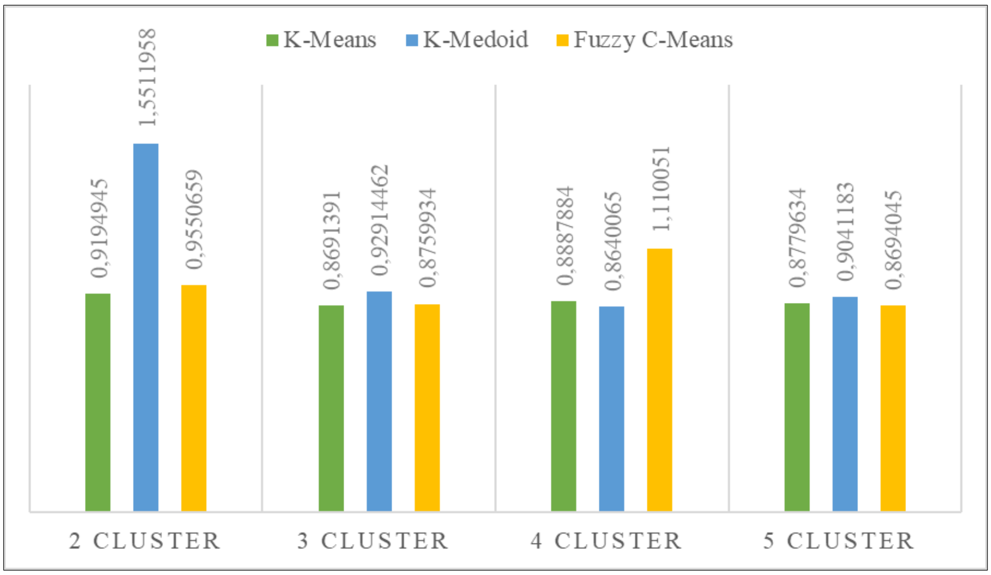

Figure 2. Comparison of Data Set DBI Value Results

The details of the testing of the DBI value result of each cluster against each algorithm are shown in the following table.

Table 3. Comparison of Data Set DBI Value Results

No	Algoritma/Cluster	K-Means	K-Medoids	Fuzzy C-Means
1	2 Clusters	0.9194945	1.551198	0.9550659
2	3 Clusters	0.8691391	0.9291462	0.8759934
3	4 Clusters	0.8887884	0.8640065	1.110051
4	5 Clusters	0.8779634	0.9041183	0.8694045

According to Fig. 2, the 4 Clusters of the K-Medoids algorithms have the lowest DBI value, the best indicator of cluster validity, with a DBI value of 0.8640065. This indicates that the data is homogeneous if grouped into 4 groups in 357 data sets. The clustering results reveal that the user data, which originally included 357 data sets of drug users, was divided into 4 groups.

Table 4. The Total of Data Sets in Each Best Cluster

Cluster	**Items**
1	105
2	142
3	56
4	54
TOTAL	**357**

The K-Medoids algorithm plot on the data with 4 clusters is shown in Fig. 3.

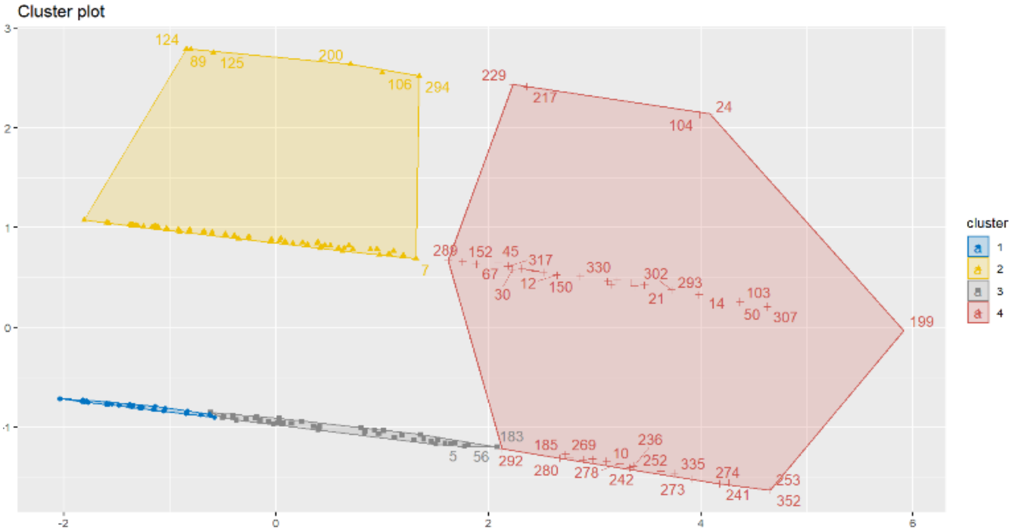

Figure 3. Plot Data from Cluster 4

The tables below display the interpretation of the findings from the data clustering of drug users into 4 clusters.

Table 5. Interpretation of The Result

Cluster	Criteria	Description of The Drug User
1	Types of Drugs	Only 1 type of drug is used
	Length of Used Time	From first-time users to those users who have been using drugs for 2 years
	Psychiatric Status	Dominated by a scale 0-4 scale
	Physical Status	Dominated by a scale 0-3 scale
2	Types of Drugs	Mostly 2 Types and a few three types are used
	Length of Used Time	From first-time users to those users who have been using drugs for six years
	Psychiatric Status	Dominated by a scale of 0-7
	Physical Status	Dominated by a scale of 0-6
3	Types of Drugs	Only 1 type of drug is used
	Length of Used Time	From one year to eight years
	Psychiatric Status	Dominated by a scale of 2-9
	Physical Status	Dominated by a scale of 2-8
4	Types of Drugs	Mostly two Types and a few three types and one type are used
	Length of Used Time	From two years to 30 years
	Psychiatric Status	Dominated by a scale of 6-9
	Physical Status	Dominated by a scale of 6-9

In the first group of Cluster 4, 105 data sets are dominated by drug users who use only 1 type of drug. Users in this group range is from first-time users to those who have been using drugs for 2 years. A scale of 0-4 dominates the psychiatric status of users in this cluster, and the physical status of users in this cluster is also on a scale of 0-3, so the first group in in this cluster can be referred to as the data group with low-risk drug addiction.

In the second group of Cluster 4 there are 142 data sets and most of the data belong to drug users by drug users who use 2 types of drugs. Users in this group range is from first-time users to those who have been using drugs for 6 years with the psychiatric status of users in this cluster is dominated by a scale of 0-7, whereas the physical status of users in this cluster is on a scale of 0-6, so the second group in this cluster can be referred to as the data group with mild risk drug addiction.

In the third group of Cluster 4, are 56 datasets dominated by drug users who use 2 types of drugs and have been using them for 1 to 8 years. A scale of 2-9 dominates the psychiatric status of users in this cluster whereas the physical status of users in this cluster is on a scale of 2-8, so the second group in cluster 4 can be referred to as the data group with moderate risk drug addiction.

In the fourth group of Cluster 4, 54 data sets are dominated by drug users who use 2 types of drugs, with a few users who also use 1 and 3 types of drugs and have been using them for 2 to 30 years. A scale of 6-9 dominates the psychiatric status of users in this cluster, and the physical status of users in this cluster is also on a scale of 6-9, so the fourth group in Cluster 4 can be referred to as the data group with high-risk drug addiction

IV. CONCLUSION

Based on the results of the average similarity of clusters of the K-Means, K-Medoids, and Fuzzy C-Means clustering algorithm on drug user data to determine the level of addiction, the K-Medoids method is found to be better than the K-Means and Fuzzy C-Means algorithm. The lowest DBI value is obtained in the K-Medoids algorithm with a value of 0.8640065 in Cluster 4, with the data in each cluster being 105, 142, 56, and 54 data on drug users.

According to the analysis's findings on the level of drug users in Cluster 4, which is the best cluster, drug users with low-risk drug addiction are represented in the first group, drug users with mild-risk drug addiction are represented in the second group, drug users with moderate risk drug addiction are represented in the third group and drug users with high-risk drug addiction are represented in the fourth group.

REFERENCES

[1] W. Utami Putri, “INDONESIA DRUGS REPORT 2022 Pusat Penelitian, Data, Dan Informasi Badan Narkotika Nasional (PUSLITDATIN BNN).” p. 24, 2022.

[2] BNN, *Survei Prevalensi Narkoba 2019*. 2019.

[3] R. Indonesia, *Undang-undang dasar negara republik indonesia Tahun 1945*. Sekretariat Jenderal MPR RI, 2002.

[4] A. Winarta and W. J. Kurniawan, “Optimasi Cluster K-means Menggunakan Metode Elbow pada Data Pengguna Narkoba dengan Pemrograman Python,” *J. Tek. Inform. Kaputama*, vol. 5, no. 1, pp. 113–119, 2021.

[5] S. S. Tandel, A. Jamadar, and S. Dudugu, “A survey on text mining techniques,” in *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, 2019, pp. 1022–1026.

[6] I. D. Guedalia, M. London, and M. Werman, “An on-line agglomerative clustering method for nonstationary data,” *Neural Comput.*, vol. 11, no. 2, pp. 521–540, 1999.

[7] D. Barbará and P. Chen, “Using self-similarity to cluster large data sets,” *Data Min. Knowl. Discov.*, vol. 7, pp. 123–152, 2003.

[8] R. Helilintar and I. NUR FARIDA, “Penerapan Algoritma K-Means Clustering Untuk Prediksi Prestasi Nilai Akademik Mahasiwa,” *J. sains dan Inform.*, vol. 4, no. 2, pp. 80–87, 2018.

[9] P. Govender and V. Sivakumar, “Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019),” *Atmos. Pollut. Res.*, vol. 11, no. 1, pp. 40–56, 2020.

[10] T. S. Madhulatha, “Comparison between k-means and k-medoids clustering algorithms,” in *Advances in Computing and Information Technology: First International Conference, ACITY 2011, Chennai, India, July 15-17, 2011. Proceedings*, 2011, pp. 472–481.

[11] K. Maharana, S. Mondal, and B. Nemade, “A review: Data pre-processing and data augmentation techniques,” *Glob. Transitions Proc.*, 2022.

[12] L. Morissette and S. Chartier, “The k-means clustering technique: General considerations and implementation in Mathematica,” *Tutor. Quant. Methods Psychol.*, vol. 9, no. 1, pp. 15–24, 2013.

[13] S. Araki, H. Nomura, and N. Wakami, “Segmentation of thermal images using the fuzzy c-means algorithm,” in *[Proceedings 1993] Second IEEE International Conference on Fuzzy Systems*, 1993, pp. 719–724.

[14] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 2, pp. 224–227, 1979.

[15] T. Wahyudi and T. Silfia, “Implementation of Data Mining Using K-Means Clustering Method To Determine Sales Strategy in S&R Baby Store,” *J. Appl.*

Eng. Technol. Sci., vol. 4, no. 1, pp. 93–103, 2022, doi: 10.37385/jaets.v4i1.913.

[16] S. Sukamto, I. D. Id, and T. R. Angraini, “Penentuan Daerah Rawan Titik Api di Provinsi Riau Menggunakan Clustering Algoritma K-Means,” *JUITA J. Inform.*, vol. 6, no. 2, p. 137, 2018, doi: 10.30595/juita.v6i2.3172.