Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta

Ilham Firman Ashari [1]*, Eko Dwi Nugroho[2]*, Randi Baraku [3]*, Ilham Novri Yanda [4*], Ridho Liwardana [5*]
* Teknik Informatika, Institut Teknologi Sumatera
firman.ashari@if.itera.ac.id [1], eko.nugroho@if.itera.ac.id [2]

Article Info

Article history:

Received 2022-12-13
Revised 2023-07-25
Accepted 2023-07-26

Keyword:

K-means,
Banjir,
Jakarta,
Klasifikasi,
Rand-Index

ABSTRACT

Jakarta is the capital city of Indonesia, which has a high population density, and is an area that is frequently hit by floods. This study aims to determine the classification of flood-affected areas in Jakarta between severe, moderate, and low. Design/method/approach: The study was conducted using the elbow, Silhouette, Davidson-Bouldin, and Calinski-Harabasz methods on the K-means algorithm, as well as the Rand method. index for evaluation. Grouping with 3 and 6 groups is the best grouping value based on Calinski-Harabasz. By using the davies bouldin index from the observations, the K value with a value of 6 has the smallest Davies-Bouldin value with a value of 0.2737. By using sillhoute, the experimental results obtained the best values sequentially, namely K=2, K=3, and K=6 with silhouette values of 0.866, 0.854, and 0.803. In this experiment, based on the elbow method, it was found that the best K value was K=3. This was obtained because it was based on observations on the appearance of the SSE data compared to the value of K. In the graph above, it can be seen that the largest decrease in data occurred at K=3 and after this decrease, the decline began to slope. The rand index is a method used to compare several cluster methods. If the value is >= 90 it is a very good result, if the value is in the range 80 to 90 it identifies a good index, whereas if it is below 80 it indicates a bad index. The results show that cluster three is verified as the best cluster with a value of 1, followed by a second alternative with cluster 2 of 0.9182. From several validation and evaluation methods it can be concluded that the best grouping can be done using 3 clusters. The results of the study yielded a value of 75.4% in low areas, 21.1% in moderate areas, and 3.5% in severe areas.

This is an open access article under the CC–BY-SA license.

I. INTRODUCTION

Jakarta is the country's center of economy, industry and government. This region is the center of the national economy, with economic activity reaching 80 percent of the entire territory of Indonesia. In addition, the money circulating in the capital region reaches 60 percent of the national scale [1]. As the capital city of Indonesia, Jakarta is the center of all activities and activities. This status also makes it one of the most densely populated areas in Indonesia. An increase in the population of a region causes an increase in population density in that region. The high population density has resulted in the area becoming increasingly denser and difficult to balance with the capacity of the water catchment area it needs. As a result, this region has become one of the areas in Indonesia with the largest flood disaster cases.

The DKI Jakarta Province area consists of several cities/regencies which have certain areas that often experience flooding. This research collects detailed data based on the kelurahan areas in DKI Jakarta Province which are the points with the highest frequency of flood events in the last 3 months. According to BMKG data, there are at least 93 points that are vulnerable to flooding in the DKI Jakarta area, with a minimum air level of 10 cm to 80 cm [2]. The data processed is regional data that does not have a flood severity level label. Processing and classification of data is done using a clustering algorithm. Clustering algorithm is an algorithm used to classify data objects based on their similarity. This algorithm is included in the category of unsupervised learning, which can be used to classify data without labels [3].

Classification of data without labels using the clustering method has a variety of available algorithms. One of the popular algorithms in clustering is K-Means. The K-Means algorithm is a clustering method in the unsupervised learning technique, where data has no labels. This algorithm performs data grouping based on convenience and results in the separation of the initial dataset into several clusters [4]. K-means clustering can be interpreted as data segmentation by increasing the similarity of data within one cluster and reducing the level of similarity of data between different clusters [5]. However, this algorithm has a weakness in determining the number of clusters needed [4]. Cluster is the center which is the center point of a group. In grouping data, the distance is calculated based on the closest distance to the centroid. Distance measurement is generally carried out using several distance measurement methods, one of which is the Euclidean distance. Euclidean distance is a method of measuring the distance between two points based on a straight line connecting them [6][7].

By using this method, we can determine the cluster of each data we have against the centroid. The specified number of centroids is affected by the type of data being processed. The number of centroids in this algorithm is a determining factor in classification because it measures the highest level of increase of each data [8]. Determining the number of centroids can be done by various methods including the Elbow Method, Silhouette Index, David-Bouldin, and Calinski Method [9][10][11][12]. Elbow method is one method that is often used [13]. This method involves observing the angles of the SSE (Sum of Squared Errors) values for each number of clusters tested. However, a comparison is needed with various other methods to strengthen the decision in determining the number of centroids. The Silhouette Index method is one of the methods used to determine the best number of clusters, because it is able to evaluate the best clusters from a point that is among many clusters [14]. The Calinski Method is the determination of clusters based on the average cluster dispersion level [15], the higher the average value, the better the number of clusters. Then the researchers used the David-Bouldin method, this method evaluates the number of clusters based on the quantity and proximity [16]. The smaller the closeness value, the better the number of clusters obtained. This study uses 3 supporting methods in the search for clusters to produce cluster values that are more optimal in its classification. In addition, evaluation was also carried out using the Rand Index measurement method which compared the clustering results of the two methods. The evaluation results will support the best cluster management in this study [17].

This research was conducted to process and process the data obtained based on the case. The method and evaluation used will support the research results so that they can be used as information for future needs. Given the problems faced by the capital region, this research is considered important to do.

II. Methods

A. Overall research flow

The overall research flow can be seen in Figure 1.

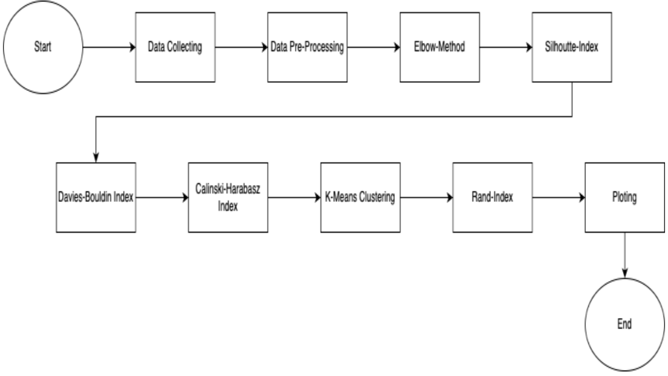


Figure 1. Overall Research Flow

This research began by collecting data from the official website of the DKI Jakarta government. The data used is flood data for the first 2 months of 2022. The data consists of separate flood reports for that period. Furthermore, in the data preprocessing stage, the separate data will be combined and processed so that they have the same size and dimensions. At this stage, the data will also be cleaned and adjusted according to clustering needs. After the data is clean, the data will be used in the modeling process.

For modeling with the K-means algorithm, setting the number of clusters is necessary initially. Therefore, in this study several methods were used to determine the optimal number of clusters. The methods used include the Elbow Method, Silhouette Index, David-Bouldin Index, and Calinski-Harabasz Index. Each method will produce the best cluster value according to each method. To compare the results of each method, an evaluation was carried out using the Rand Index. The evaluation results will provide the best cluster value for modeling. After obtaining the best number of clusters, the data will be clustered using the K-means algorithm using Euclidean Distance to determine groups based on data similarity. Thus, the next step is plotting to visualize regions based on the type of cluster.

B. Data Collecting

Researchers conducted a data search with a focus on relevant and urgent topics, namely the problem of floods in the Capital region. After determining the research topic, the researcher searched for data sources and managed to get data from the official website of the DKI Jakarta government. The data source obtained by the researcher is a valid and accountable data source. The data is in the form of a separate flood report for each sub-district area. Researchers chose data from January and February as the data to be used in this study.

C. Preprocessing Data

Data Preprocessing is done by making data month by month. Data for the first month (January) will be loaded and data dimensions will be reduced to take only the features needed. The features that researchers need are affected victims and flood heights. Next loads the second data and performs the same dimension reduction for this data. Dimensionally reduced data will be combined into one data file (csv). This data will be normalized in value to suit the needs of the model. The feature that receives normalization is the "flood height" flood height is a range data that has no valid value. Normalization is carried out by taking the average of the initial range and the final range of each data so that the average water level is obtained from each flood case. This data is ready to be used for further processing. The next process is a modeling process using the K-means algorithm and a series of methods for determining the number of centroids.

D. Elbow Method

The elbow method process flow can be seen in Figure 2.

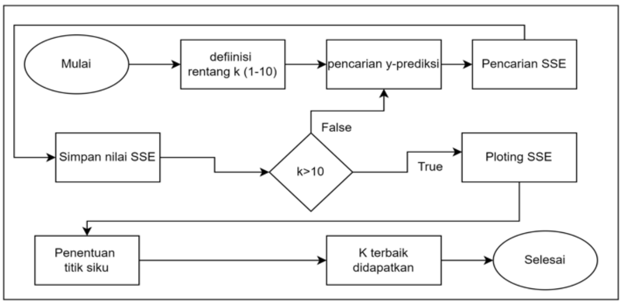


Figure 2. Elbow Method Process Flow

Initially, the value of K that you want to see the elbow points will be determined in advance, in this case 1 to 10. The data the researcher has (2 dimensions) will search for the Y-prediction, the Y-prediction is a temporary prediction of clusterization based on the value of k has been set in advance. Then calculate the SSE (Sum of Square Error) value, this value is the error value of the centroid distance that has been set for each K. The distance calculation is done using the euclidean distance. The calculation of the SSE value is performed for each K value from 1 to 10. Furthermore, after all K values are obtained for each SSE value, plotting of these values will be carried out. Plotting is done to see the angle points of the SSE graph. This elbow point will be the best cluster value in this method.

E. Silhouette Method

The silhouette method process flow can be seen in Figure 3.

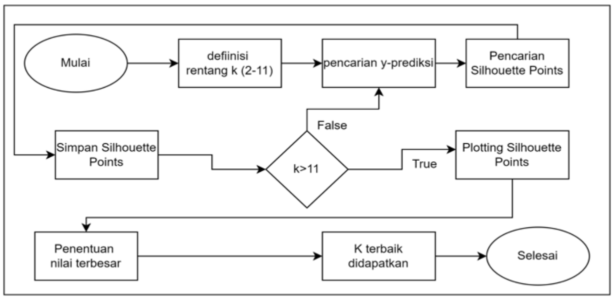


Figure 3. Silhouette Method Process Flow

This method begins by determining the range of clusters whose silhouette values you want to find, the researcher takes the range 2 to 11. After that, a y-prediction search is performed for each cluster value (K) for each K value, the silhouette value is searched for each data using the Euclidean distance. . Furthermore, after the values of all K values are obtained for each sillhouete value, then plotting of these values will be carried out. Plotting is done to see which value is closest to 1, the value closest to 1 (the highest) is the value of the best cluster based on the SSE graph method. this elbow point will be the best cluster value in this method.

F. David-Bouldin Index

The David bouldin index process flow can be seen in Figure 4.

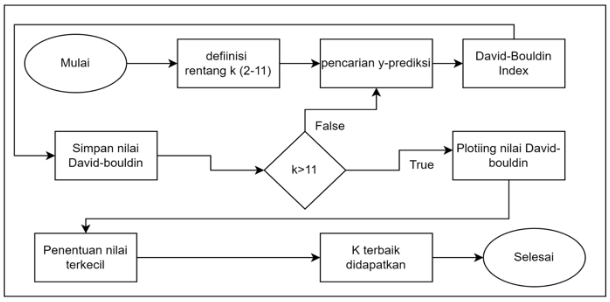


Figure 4. Process Flow of the David-Bouldin Index

As before, this method requires the initialization of the k range values (2-11). Then for each value of K the value of the david-bouldin index will be calculated. The researcher will save each value from David Bouldin for mapping when the k value reaches 11. After the K value reaches 11, the mapping is carried out and the lowest value is the best result of this clustering.

G. Calinski-Harabasz Index

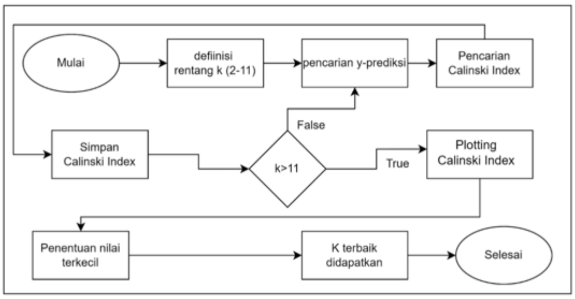


Figure 5. Calinski-Harabasz Index Process Flow

The Calinski Harabasz Index is obtained by first defining the value of the K range you want to search for. Then we do cluster predictions based on the value of each K. Each value of K will have its own Calinski value. We'll take this value to map. The smallest value of the Calinski index is the best cluster based on this method.

H. K-Means Algorithm

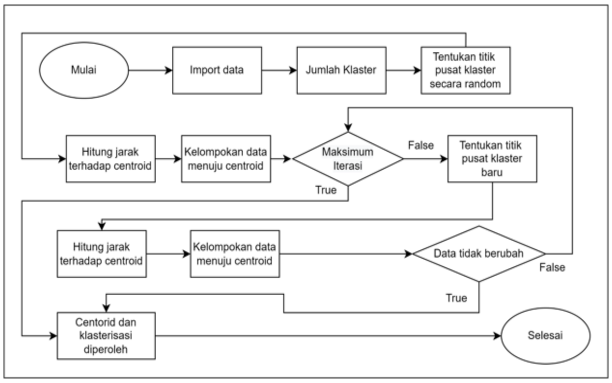

Mulai
Import data
Jumlah Klaster
Tentukan titik pusat klaster secara random
Hitung jarak terhadap centroid
Kelompokan data menuju centroid
Maksimum Iterasi
False
Tentukan titik pusat klaster baru
True
Hitung jarak terhadap centroid
Kelompokan data menuju centroid
Data tidak berubah
False
True
Centorid dan klasterisasi diperoleh
Selesai

Figure 6. Process Flow K-means Algorithm

The method adopted in this algorithm is as follows. The data that we have prepared previously will be used to be grouped based on the number of clusters that have been obtained in several previous methods. The researcher determines one of the number of clusters that feels the best in this case. After the value of K is determined, this modeling algorithm can be carried out. K-means starts by determining one point randomly as many as the number of clusters. Then all data is calculated based on the distance to the nearest point. Data will be grouped into clusters based on that distance, distance calculation is done using the Euclidean distance. After that we will check whether all iterations have been done, iterations are initiated as many times as the number of data. If it is the maximum iteration then clustering is found. If the iteration has not reached the maximum iteration, then a new centroid will be randomly re-selected. Then the distance from all points will be measured on the centorid, the data will then be regrouped. If the grouped data changes with the previous grouping, the process will take place again by checking the number of iterations, but if the grouping data does not change, then the clustering process ends. The grouping process will end when the data does not change or the iteration ends.

I. Rand Index Evaluation

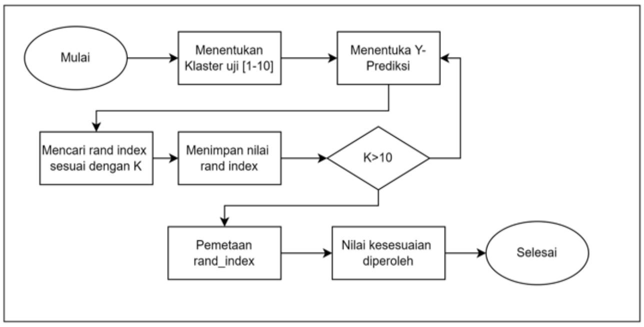

Mulai
Menentukan Klaster uji [1-10]
Menentuka Y-Prediksi
Mencari rand index sesuai dengan K
Menimpan nilai rand index
K>10
Pemetaan rand_index
Nilai kesesuaian diperoleh
Selesai

Figure 7. Rand Index Process Flow

This method is used to measure the suitability of the percentage of the number of clusters with the number of other clusters. The higher the percentage value, the more appropriate the data grouping is. At first we tested the cluster range from 1 to 10. Then the rand index will be calculated and stored. The rand index search is carried out until K reaches 10. The result is that the value with the highest suitability will be the best cluster in this evaluation. This can be seen from the mapping and numerical results received.

J. Plotting

The plotting flow can be seen in Figure 8.

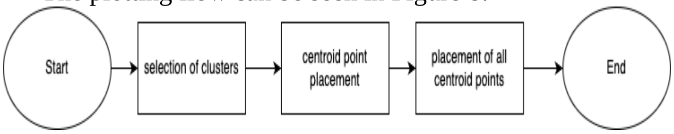


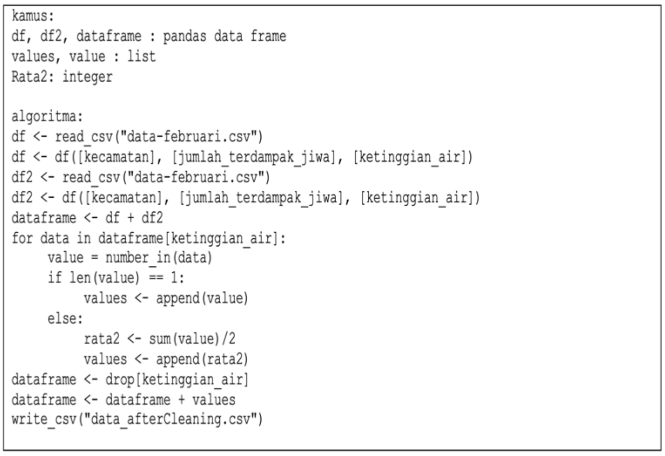
Figure 8. Plotting Process Flow

Plotting is the final stage of the research method that researchers do. This is done by first determining the centroid that has been obtained, then mapping all points based on the grouping results.

III. RESULT AND DISCUSSION

A. Data Preprocessing

Pseudocode of the data preprocessing program can be seen in Figure 9 below. At this stage, the raw data of type csv from Open Data Jakarta is cleaned of unnecessary features (data columns), such as rt, rw, and flood_dates. And with the deletion, only 3 columns of data are obtained, namely, the number of residents affected by the flood, the height of the flood and the sub-district where the flood hit.



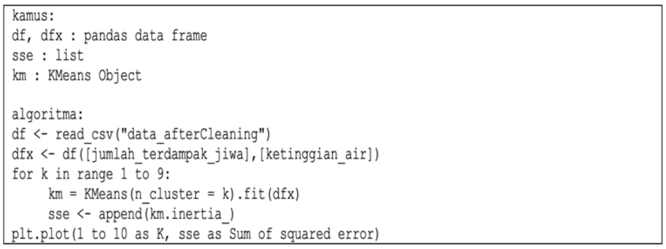
```
kamus:
df, df2, dataframe : pandas data frame
values, value : list
Rata2: integer

algoritma:
df <- read_csv("data-februari.csv")
df <- df([kecamatan], [jumlah_terdampak_jiwa], [ketinggian_air])
df2 <- read_csv("data-februari.csv")
df2 <- df([kecamatan], [jumlah_terdampak_jiwa], [ketinggian_air])
dataframe <- df + df2
for data in dataframe[ketinggian_air]:
    value = number_in(data)
    if len(value) == 1:
        values <- append(value)
    else:
        rata2 <- sum(value)/2
        values <- append(rata2)
dataframe <- drop[ketinggian_air]
dataframe <- dataframe + values
write_csv("data_afterCleaning.csv")
```

Figure 9. Pseudocode Data Preprocessing

B. Elbow Method

Pseudocode of the data preprocessing program can be seen in Figure 10 below



```
kamus:
df, dfx : pandas data frame
sse : list
km : KMeans Object

algoritma:
df <- read_csv("data_afterCleaning")
dfx <- df([jumlah_terdampak_jiwa],[ketinggian_air])
for k in range 1 to 9:
    km = KMeans(n_cluster = k).fit(dfx)
    sse <- append(km.inertia_)
plt.plot(1 to 10 as K, sse as Sum of squared error)
```

Figure 10. Pseudocode Elbow Method

The use of the Elbow method begins with the initialization of the iteration value of the K value (which is a reference in the K-Means algorithm), for this experiment the K values from 1 to 10 are used. nearest centroid, or commonly also called the sum of squared errors (Sum of Squared Error). The value is then saved to the SSE list. And by using the plot, the data is displayed. In this experiment, for the elbow method, the results are shown in Figure 11.

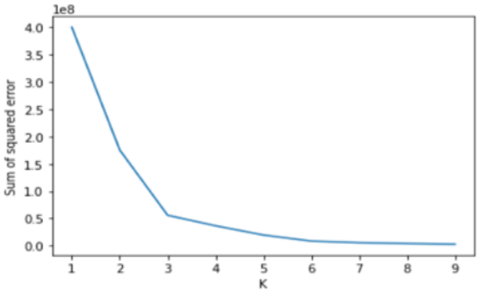
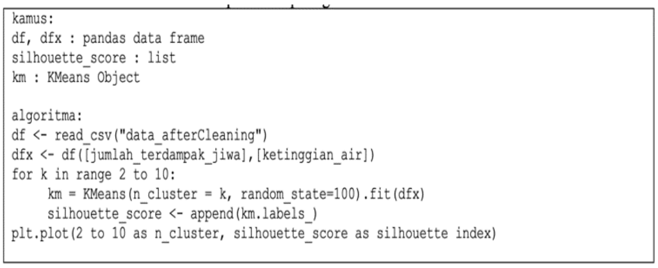


Figure 11. Elbow Graph

In this experiment, based on the elbow method, it was found that the best K value was K=3. This was obtained because it was based on observations on the appearance of the SSE data compared to the value of K. In the graph above, it can be seen that the largest decrease in data occurs at K=3 and after this decrease, the decline begins to slope.

C. Silhouette Index

The pseudocode of the silhouette index can be seen in Figure 12.



```
kamus:
df, dfx : pandas data frame
silhouette_score : list
km : KMeans Object

algoritma:
df <- read_csv("data_afterCleaning")
dfx <- df([jumlah_terdampak_jiwa],[ketinggian_air])
for k in range 2 to 10:
    km = KMeans(n_cluster = k, random_state=100).fit(dfx)
    silhouette_score <- append(km.labels_)
plt.plot(2 to 10 as n_cluster, silhouette_score as silhouette index)
```

Figure 12. Pseudocode Silhouette Index

In silhouette analysis, we compare the results of experiments using K-Means with several K values. The K values used start from 2 to 10. And later the Silhouette values are analyzed based on the closeness between the values of fellow clusters and their dissimilarity with different clusters, which value is obtained from the label property on the clustering results and the value of each data. Later, based on the calculation results, the silhouette values are obtained as in table 1.

TABEL I
SILHOUETTE CALCULATION RESULTS

The Number of Cluster	Silhouette Score
2	0,866465
3	0,854246
4	0,796296
5	0,793692
6	0,803351
7	0,719845
8	0,715878
9	0,716153
10	0,719677

Or it can be displayed in graphical form in Figure 13.

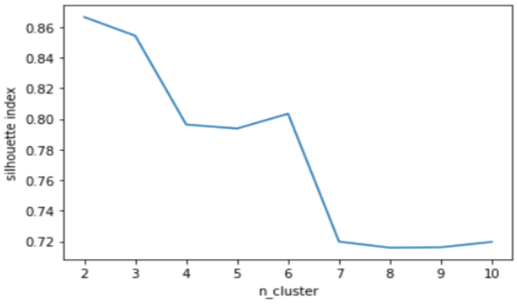


Figure 13. Silhouette Index Graph

In the analysis using the silhouette method, the closest value to 1 is considered the best number of groupings. And from the experimental results, the best values were obtained sequentially, namely K=2, K=3, and K=6 with silhouette values of 0.866, 0.854 and 0.803. And the three groupings if visualized can look like in pictures 14,15 and 16.

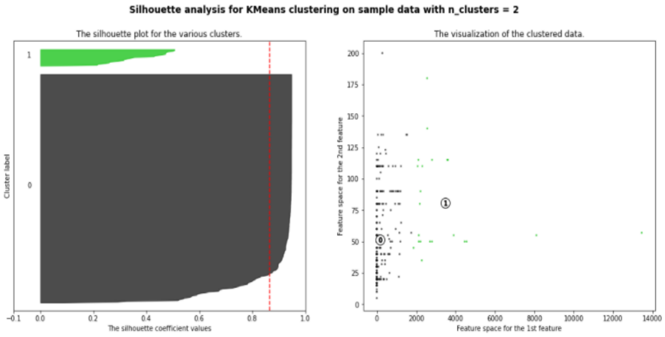


Figure 14. Results of Silhouette Analysis with 2 Clusters

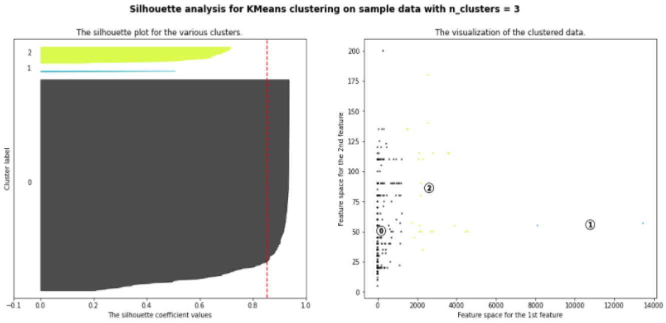


Figure 15. Results of Silhouette Analysis with 3 Clusters

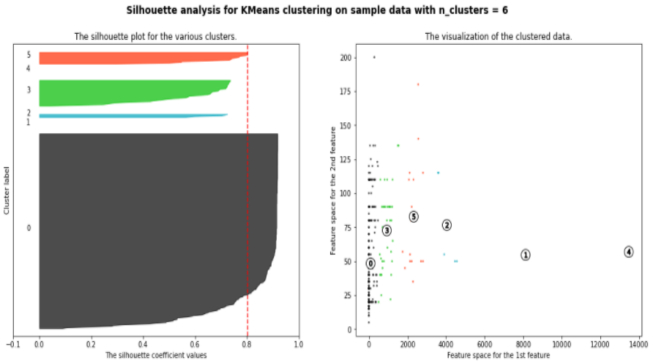
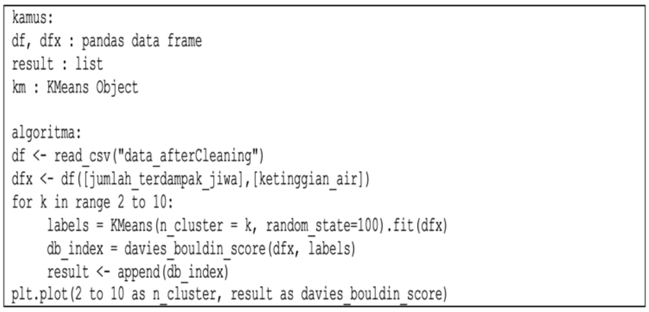


Figure 17. Results of Silhouette Analysis with 6 Clusters

Based on the results of the evaluation using the silhouette, it can be seen that grouping with 2 clusters is considered the best grouping, with cluster 0 dominating the data grouping and in cluster 1 the lowest silhouette value reaches a value of around 0.2 which is quite low compared to the silhouette value owned by cluster members at other K values. Furthermore, by grouping with 3 clusters it was found that the members of one of the clusters at K = 2 were broken down and a better average silhouette per cluster was obtained, where cluster 1 could accommodate outlier values in the grouping method with only 2 clusters. Finally, there is grouping with 6 clusters (K=6). It can be seen that one of the clusters in this grouping has only one member, namely cluster 4, this causes the silhouette method to be unable to calculate its silhouette value.

D. Davies-Bouldin Method

The pseudocode of the davies-bouldin index can be seen in Figure 18.



```
kamus:
df, dfx : pandas data frame
result : list
km : KMeans Object

algoritma:
df <- read_csv("data_afterCleaning")
dfx <- df([jumlah_terdampak_jiwa],[ketinggian_air])
for k in range 2 to 10:
    labels = KMeans(n_cluster = k, random_state=100).fit(dfx)
    db_index = davies_bouldin_score(dfx, labels)
    result <- append(db_index)
plt.plot(2 to 10 as n_cluster, result as davies_bouldin_score)
```

Figure 18. Pseudocode Davies-Bouldin Method

In the analysis using the David-Bouldin method, iterations in the range of 2 to 10 are used, where this range is the experimental range of K values that will be used in grouping using the K-Means algorithm. Then the results of each grouping experiment are calculated by the Davies Bouldin score, which is based on the value of closeness between cluster members and the distance between clusters, where the smaller the value of the Davies-Bouldin calculation results, the more accurate the K value is used. Then the calculation results are displayed using a plot as shown in Figure 19.

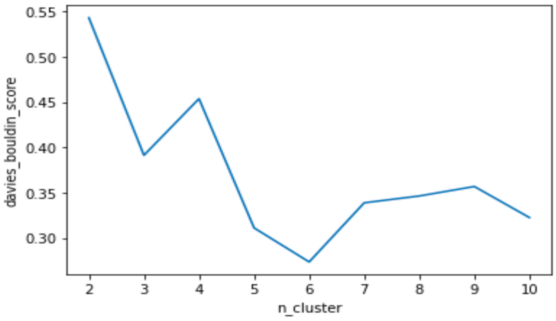
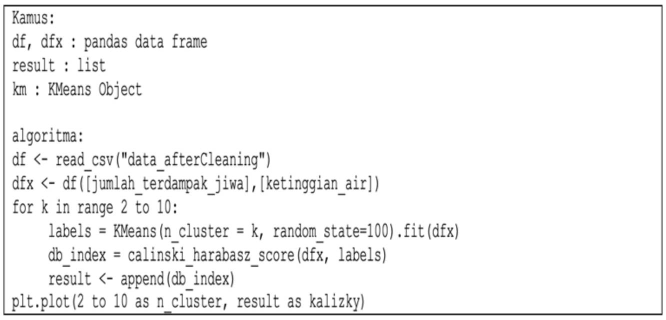


Figure 19. Davies-Bouldin Method Graph

It can be seen in the picture above that K with a value of 6 has the smallest Davies-Bouldin value with a value of 0.2737. This shows that based on Davies-Bouldin calculations on flood data clustering in the capital city of Jakarta, there are 6 data groupings. In addition to grouping with 6 clusters, there are also several values that can be considered here, namely grouping data with 5 and 3 clusters where these two values are values that have the highest decreasing value compared to the previous value. This causes the two values to form a lower valley than the grouping data with the previous value.

E. Calinski-Harabasz Index

The pseudocode of the Calinski-Harabasz Index can be seen in Figure 20.



```
Kamus:
df, dfx : pandas data frame
result : list
km : KMeans Object

algoritma:
df <- read_csv("data_afterCleaning")
dfx <- df([jumlah_terdampak_jiwa],[ketinggian_air])
for k in range 2 to 10:
     labels = KMeans(n_cluster = k, random_state=100).fit(dfx)
     db_index = calinski_harabasz_score(dfx, labels)
     result <- append(db_index)
plt.plot(2 to 10 as n_cluster, result as kalizky)
```

Figure 20. Calinski-Harabasz Index

In the analysis using the Calinski-Harabasz Index, loops starting from 2 to 10 are used, which repetitions are used as the K value in calculating the K-Means algorithm. Then the results of these calculations are used to calculate the Calinski-Harabasz Index. In calculations using Calinski-Harabasz the similarity of each data in the cluster and the distribution of data between clusters. In Calinski-Harabasz calculations, the higher the value indicates the better the grouping is. The following are the results of an experiment using Calinski-Harabasz in classifying flood data for the capital city of Jakarta, which can be seen in Figure 21.

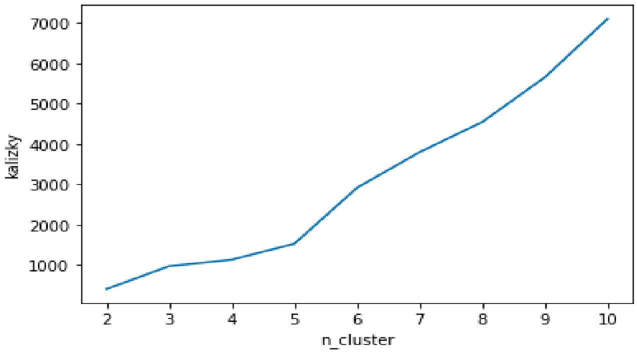
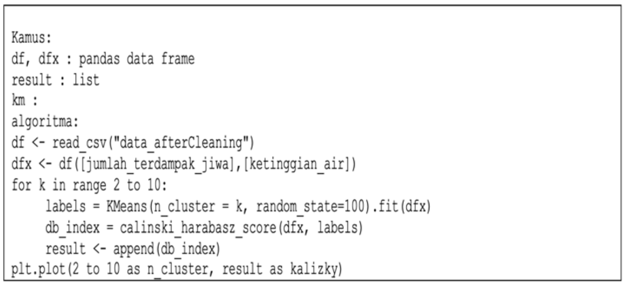


Figure 21. Graph of Calinski-Harabasz Method

In the picture above it can be observed that the increase that the highest percentage increase occurred from grouping 5 groups to grouping 6 groups. And after grouping with 3 groups the value of the increase is sloping. Based on this, the researcher concluded that grouping with 3 and 6 groups is the best grouping value based on Calinski-Harabasz.

F. K-Means Clustering

Pseudocode of K-Means Clustering can be seen in Figure 22.



```
Kamus:
df, dfx : pandas data frame
result : list
km :
algoritma:
df <- read_csv("data_afterCleaning")
dfx <- df([jumlah_terdampak_jiwa],[ketinggian_air])
for k in range 2 to 10:
     labels = KMeans(n_cluster = k, random_state=100).fit(dfx)
     db_index = calinski_harabasz_score(dfx, labels)
     result <- append(db_index)
plt.plot(2 to 10 as n_cluster, result as kalizky)
```

Figure 22. Pseudocode K-Means Clustering

Based on the several validation and evaluation methods above, it can be concluded that the best grouping can be done using 3 clusters. And with the value of the number of clusters, a clustering experiment was carried out using the KNN-Means algorithm. The results of the k-means method are cluster predictions for each sub-district area with the severity due to flooding that occurred in Jakarta, which can be seen in Figure 23.

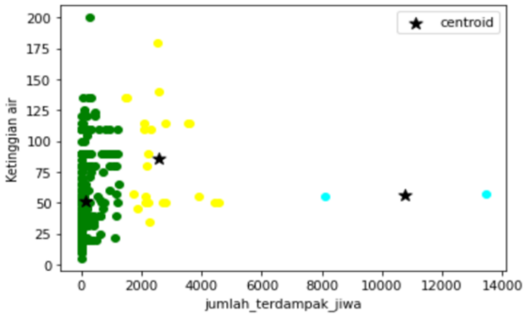


Figure 23. Kmeans Plotting Results with 3 Clusters

From the data above, it is obtained that cluster 1 (green) is a flood area with a low impact and level of damage, for cluster 2 (yellow) is an area with a moderate level of impact, and cluster 3 (blue) is an area with a moderate level of damage. critical. The results of grouping regions with 3 clusters can be seen in table 2.

TABEL II
REGIONAL GROUPING

Regional Grouping	
Low Damage	Johar Baru, Kemayoran, Sawah Besar, Tanah Abang, Senen, Menteng, Cempaka Putih, Gambir, Tanjung Priok, Kelapa Gading, Koja, Penjaringan, Pademangan, Cilincing, Cengkareng, Grogol Petamburan, Kalideres, Kebon Jeruk, Kembangan, Palmerah, Taman Sari, Tambora, Cilandak, Jagakarsa, Kebayoran Baru, Kebayoran Lama, Mampang Prapatan, Pancoran, Pasar Minggu, Pesanggrahan, Setiabudi, Tebet, Cakung, Cipayung, Ciracas, Duren Sawit, Jatinegara, Kramat Jati, Makassar, Matraman, Pasar Rebo, Pulogadung, Cilincing
Moderate Damage	Cakung, Ciracas, Duren Sawit, Jatinegara, Tanah Abang, Cengkareng, Kalideres, Pancoran, Tebet, Kramat Jati
High Damage	Makassar, Pulogadung

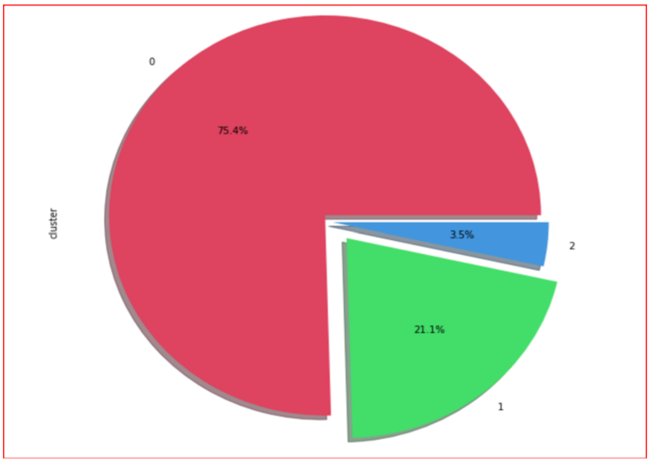


Figure 24. Pie Chart Data Clustering

From Figure 24, it can be concluded that 75.4% of the area is an area with a low level of damage, 21.1% of the area is an area with a moderate level of damage, and the remaining 3.5% is an area with a high level of damage.

G. Rand Index

The results of the rand index obtained for each cluster are as follows shown in table 3.

TABEL III
RAND INDEX CALCULATION RESULTS

Cluster	Rand Index
1	0.0
2	0.9182048844698869
3	1.0
4	0.5362315076071391
5	0.5390899107517388
6	0.5307590092051565
7	0.3307479803929599
8	0.24303668358641659

The Rand Index method is used to compare several cluster methods. Values >= 90 indicate very good results, ranges 80-90 indicate good results, while values below 80 indicate poor results. The Rand Index graph obtained can be seen in Figure 25.

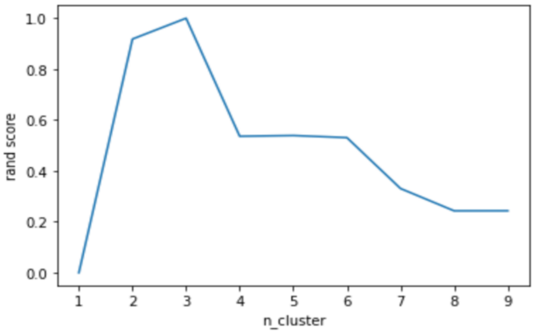


Figure 25. Graph of Rand Index

The results show that cluster three is verified as the best cluster with a value of 1, followed by a second alternative with cluster 2 of 0.9182048844698869. The evaluation above proves that determining the number of clusters as many as 3 is the best choice based on various validation methods that have been used previously.

V. CONCLUSIONS

Within the scope of sub-districts in DKI Jakarta Province, using the validation method of the elbow method, silhouette index, davies-bouldin method, and calinski-harabasz index, it was determined that cases of flood severity were divided into 3 cluster areas. The handling of the sub-district area is based on the parameters of the flood height level and the impact on fatalities in high, medium and low flood severity levels. The results show areas with a high level of damage of 3.5%, areas with a moderate level of damage of 21.1%, and areas with a low level of damage of 75.4%.

REFERENCES

[1] Rahmatulloh, “DINAMIKA KEPENDUDUKAN DI IBUKOTA JAKARTA (Deskripsi Perkembangan Kuantitas, Kualitas dan Kesejahteraan Penduduk di DKI Jakarta),” *Genta Mulia*, vol. VIII, no. 2, pp. 54–67, 2017.

[2] Eldi, “Analisis Penyebab Banjir di DKI Jakarta,” *J. Inov. Penelit.*, vol. 1, no. 6, pp. 1057–1065, 2020.

[3] 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.

[4] H. Firdaus and A. Sofro, “Analisa Cluster Menggunakan K-Means Dan Fuzzy C-Means Dalam Pengelompokan Provinsi Menurut Data Intesitas Bencana Alam Di Indonesia Tahun 2017-2021,” *MATHunesa J. Ilm. Mat.*, vol. 10, no. 1, pp. 50–60, 2022, doi: 10.26740/mathunesa.v10n1.p50-60.

[5] M. Nishom, “Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square,” *J. Inform. J. Pengemb. IT*, vol. 4, no. 1, pp. 20–24, 2019, doi: 10.30591/jpit.v4i1.1253.

[6] 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.

[7] M. Hoffmann and F. Noé, “Generating valid Euclidean distance matrices,” 2019, [Online]. Available: http://arxiv.org/abs/1910.03131.

[8] C. Yuan and H. Yang, “Research on K-Value Selection Method of

K-Means Clustering Algorithm," *J*, vol. 2, no. 2, pp. 226–235, 2019, doi: 10.3390/j2020016.

[9] 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.

[10] X. Wang and Y. Xu, "An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 569, no. 5, 2019, doi: 10.1088/1757-899X/569/5/052024.

[11] Y. Yu, Y. Wang, G. Zhang, and J. Wang, "Research of Fault Feature Extraction and Analysis Method Based on Aeroengine Fault Data," *Proc. - 2020 Chinese Autom. Congr. CAC 2020*, pp. 2960–2965, 2020, doi: 10.1109/CAC51589.2020.9327519.

[12] I. F. Ashari, R. Banjarnahor, and D. R. Farida, "Application of Data Mining with the K-Means Clustering Method and Davies Bouldin Index for Grouping IMDB Movies," vol. 6, no. 1, pp. 7–15, 2022.

[13] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 336, no. 1, 2018, doi: 10.1088/1757-899X/336/1/012017.

[14] A. R. Mamat, F. S. Mohamed, M. A. Mohamed, N. M. Rawi, and M. I. Awang, "Silhouette index for determining optimal k-means clustering on images in different color models," *Int. J. Eng. Technol.*, vol. 7, pp. 105–109, 2018, doi: 10.14419/ijet.v7i2.14.11464.

[15] S. P. Lima and M. D. Cruz, "A genetic algorithm using Calinski-Harabasz index for automatic clustering problem," *Rev. Bras. Comput. Apl.*, vol. 12, no. 3, pp. 97–106, 2020, doi: 10.5335/rbca.v12i3.11117.

[16] S. I. Murpratiwi, I. G. Agung Indrawan, and A. Aranta, "Analisis Pemilihan Cluster Optimal Dalam Segmentasi Pelanggan Toko Retail," *J. Pendidik. Teknol. dan Kejuru.*, vol. 18, no. 2, p. 152, 2021, doi: 10.23887/jptk-undiksha.v18i2.37426.

[17] J. E. Chacón and A. I. Rastrojo, "Minimum adjusted Rand index for two clusterings of a given size," *Adv. Data Anal. Classif.*, 2022, doi: 10.1007/s11634-022-00491-w.