KLIK: Kajian Ilmiah Informatika dan Komputer
ISSN 2723-3898 (Media Online)
Vol 4, No 1, Agustus 2023, Hal 545-552
DOI 10.30865/klik.v4i1.964
https://djournals.com/klik

Optimization of the K-Means Clustering Algorithm Using Davies Bouldin Index in Iris Data Classification

Devanta Abraham Tarigan*

Program Studi Sistem Informasi, STMIK Kristen Neumann Indonesia, Medan, Indonesia
Email: [1]abrahamtambac@gmail.com
Email Penulis Korespondensi: abrahamtambac@gmail.com

Abstract– Data grouping is done by calculating the shortest distance to the initial cluster center point as the central point in the formation of each group or cluster. The results of the K-Means optimization study with the Davies Bouldin Index k-means clustering by dividing the cluster values 3,5,7,9. In testing the K-3 cluster values, the performance value of the average centroid distance is -0.312, then the K-Means optimization results with Davies Bouldin have a performance percentage of -0.799. Testing the value of the K-5 cluster has a performance value of an average centroid distance of -0.310, then the K-Means optimization results with Davies Bouldin have a performance percentage of -0.806. Testing the value of the K-7 cluster has a performance value of an average centroid distance of -0.310, then the K-Means optimization results with Davies Bouldin have a performance percentage of -0.806. Testing the K-9 cluster values has a performance value of an average centroid distance of -0.310, then the results of the K-Means optimization with Davies Bouldin have a performance percentage of -0.806. From the test results with variations in cluster values 3,5,7,9 it can be concluded that the optimization of the K-Means method with the Davies Bouldin Index testing the K-3 cluster values has better performance with an average value of -0.312 centroid distance then the results of K-optimization Means with Davies Bouldin has a performance percentage: -0.799.

Keywords: K-Means Clustering; Davies Buoldin Index; Classification Iris

1. INTRODUCTION

In a large dataset, data mining is a form of completion process that produces several new patterns into useful information. Several settlement techniques commonly used in data mining, namely estimation, association, clustering, and classification. Clustering is usually used in finding a number of data groups, classification techniques are used to analyze data using data class models to predict labels. One of the many classification techniques that is often used is K-Means [1].

In simple terms, it can be understood that data mining, also known as knowledge discovery in databases (KDD), is a series of processes that aim to extract important or interesting patterns from very large amounts of data that cannot be recognized manually [2].

Grouping can use clustering to group data based on similarities between data, so that data with the closest similarity is in one cluster while data that is different is in another group. The process of grouping data into several clusters or groupings so that data in one cluster has the maximum level of similarity and between clusters has a minimum similarity is called Clustering [3].

Data mining is an effort to utilize data by mining important information stored in large amounts of data [4]. The K-Means clustering algorithm is a method of grouping data that is simpler to implement or easy to implement. The process flow of this algorithm is to partition or divide a number of data into K clusters through the average (Mean) of the shortest distance of all data to the data cluster which will then be calculated to average the data points in the next iteration. The advantage of the K-Means clustering algorithm is that it has a relatively fast computation time [5].

The K-Means algorithm is a relatively simple algorithm for classifying or grouping a large number of objects with certain attributes into K clusters [6]. In the k-Means algorithm, the number of K clusters has been determined beforehand. K-Means is a clustering algorithm for data mining that was created in the 70s and is useful for clustering elemental learning (unsupervised learning) in a data set based on certain parameters [7]. KMeans is an algorithm for classifying or grouping objects (in this case data) based on certain parameters into a number of groups, so that it runs faster than hierartical clustering (if small) with a large number of variables and produces denser clusters [8]

The K-Means algorithm is one of the clustering algorithms included in partitioning-based clustering. The KMeans algorithm partitions the data set into a predefined number of k clusters [9]. Partitioning of the data set is done to determine the characteristics of each cluster, so that clusters with the same characteristics are grouped into one cluster and those with different characteristics are grouped into another cluster [10]. The advantages of the K-Means algorithm are that it takes a relatively short time and is easy to implement. Even though it has advantages, the K-Means algorithm also has weaknesses. One of the weaknesses of the K-Means algorithm is the initialization of the cluster center point or initial centroid [11].

The cluster center point or centroid is the starting point for grouping within the cluster in the K-Means algorithm. Data grouping is done by calculating the shortest distance to the initial cluster center point as the central point in the formation of each group or cluster [12]. But in practice, determining the initial cluster center point is the weakness of the K-Means algorithm. This is because there is no approach used in selecting and determining the cluster center point. The cluster center point is chosen arbitrarily or randomly from a set of data. As a result, the clustering results from the K-Means algorithm are often less than optimal and not optimal in every experiment that is carried out. Therefore, it can be said that the clustering results are good or bad, very dependent on the cluster center point or initial centroid [13].

The Davies-Bouldin Index (DBI) is one of the methods used to measure cluster validity in a clustering method. This measurement with the DaviesBouldin Index maximizes the inter-cluster distance and at the same time tries to minimize the distance between points in a cluster [14]. If the inter-cluster distance is maximal, it means that the characteristics of each cluster are slightly similar, so that the differences between clusters are seen more clearly. If the intra-cluster distance is minimal, it means that each object in the cluster has a high degree of similarity of characteristics [11]. The clustering results obtained from determining the proposed cluster center point are then evaluated using the DBI method. So that it can be seen the correlation of the method of determining the cluster center point based on the Sum of Squared Error on improving cluster quality based on the DBI value obtained [15].

Research Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means After performing the k-Means clustering algorithm and performing different tests, we find that clusters are better framed at k equivalent. The main parameters on which we have decided that for what value of K, the better the DBI clustering. This work can be extended to other advanced clustering algorithms as we can carry out comparative learning of simple k-means and deep learning algorithms in the future [16].

Previous research entitled Application of the K-Modes Clustering Algorithm with the Validation of the Davies Bouldin Index on Grouping Levels of Interest in Online Shopping in the Province of the Special Region of Yogyakarta [17]. The results of his research Davies-Boulden Index (DBI) are used to determine the best number of clusters. Based on the results of the analysis, the best number of clusters was obtained, namely k = 9 with a DBI value of 1.3427. Cluster 5 is the best cluster whose members are very interested in shopping through the Marketplace and Social Media. Marketplaces of interest are Shopee, Bukalapak, and Tokopedia, while Social Media of interest are Instagram, Facebook, and Media Chatting. This cluster is dominated by young men (15-24 years) [18].

Research A Cluster Validity for Spatial Clustering Based on Davies Bouldin Index Spatial clustering is the most powerful technology for spatial data mining. One important part of spatial clustering is cluster validity.

2. RESEARCH METHODOLOGY

2.1 Research Stages

The research procedure is one of the important processes. In this research procedure step by step must be carried out in detail, structured and systematically, so that the research target is achieved as expected. As for the research framework as shown in Figure 1.

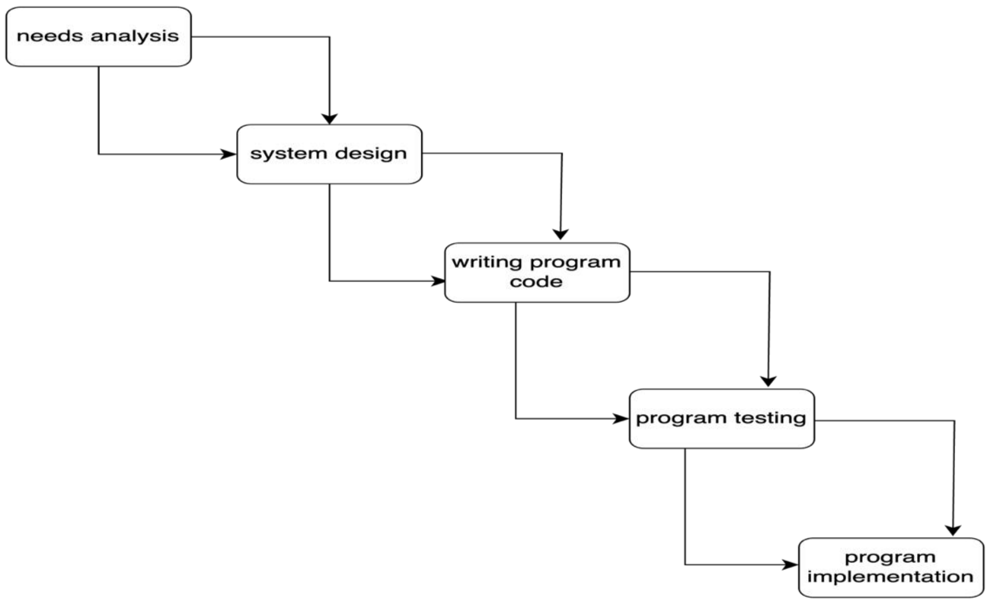

Figure 1. Research Flow Chart

1. Needs Analysis
   In carrying out this service, to support needs analysis, data and information collection is carried out at this stage to find out about the system under study. From the data and information collected, data will be obtained to support research and data collection is carried out to determine the needs of the user. The method used by the author for data collection is as follows:
   a. Observation
   Observation is useful for collecting data and observing by using methods for implementing algorithms with journals and references to previous studies
2. System Design

This stage is done before coding. This stage aims to provide an overview of what should be done and how it looks. This stage helps in specifying hardware and system requirements and defining the overall system architecture. The result is the design of the system to be built and the interface of the application to be made.

3. Application of Tools

   In this stage, rapid miner tools are used to manage data and algorithms. Determination of tools is done in order to implement the Davies Bouldin Index method. In addition, at this stage an examination is also carried out on the class created, whether it has fulfilled the desired function or not. These functions will also be adjusted to the design of the application.

4. Program Testing

   At this stage, the rapid miner tool Black Box test was carried out in testing the K-Means method and this is the final stage of the flowchart. The finished application is run and carried out maintenance. Maintenance includes fixing errors from features that are not in accordance with the application design. At this stage, the features that need to be added to the application that has been built will also be added.

The method used to obtain knowledge from existing databases. The results of the knowledge obtained can be used as a knowledge base that is used for making decisions. In more detail, the KDD process is as shown in the following figure adopted from.

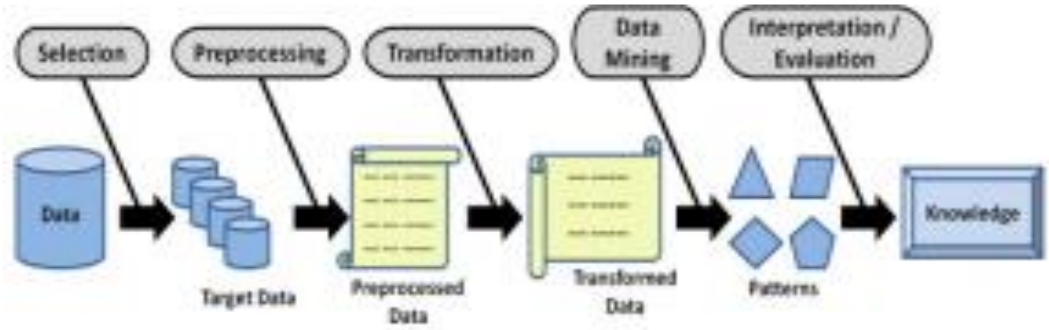

Figure 2. Knowledge Discovery In Database (KDD)

1. Selection

   Selection is used to determine the variables to be taken so that there are no similarities and unnecessary repetition occurs in data mining processing.

2. Preprocessing

   In preprocessing there are two stages, namely as follows:

   a. DataCleaning

      Eliminate unnecessary data such as handling missing values, noise data and handling inconsistent and relevant data.

   b. Data Integration

      Performed on attributes that identify unique entities.

3. Transformation

   Changing data according to the appropriate extension format in data mining processing because some methods in data mining require a special format before it can be processed in data mining.

4. Data mining

   The main process of the method applied to obtain new knowledge from the processed data. In this study, a clustering technique was applied, namely the K-Means Clustering method.

5. Evaluation/Interpretation

   Identify interesting patterns into the identified knowledge base. At this stage, it produces typical patterns and predictive models which are evaluated to assess existing studies that have met the desired target.

6. Knowledge

   The resulting patterns will be presented to the user. At this stage the new knowledge generated can be understood by everyone who will be used as a reference for decision making

2.2 K-Means Clustering

The K-Means algorithm is an algorithm used for iterative grouping, this algorithm partitions data sets into a number of K clusters that have been determined at the beginning. Partitioning of the data set is done to determine the characteristics of each cluster, so that clusters that have the same characteristics are grouped into one cluster and those with different characteristics are grouped into another cluster [19]. The following is the flow of the KMeans algorithm

1. Determine the number of K, where K is the number of clusters to be formed.
2. Set the cluster center point randomly or randomly, the cluster center point is often called the centroid.
3. Calculate each existing data distance to each centroid using the Euclidean Distance formula.

1

$$D_e = \sqrt{(xi - si)^2 + (yi - ti)^2} \quad (1)$$

Information:
De= Euclidean Distance
i = many objects,
(X,y)= are the coordinates of the object, and
(S,T)= is the centroid coordinate (cluster center point)

4. Then allocate each object into a cluster based on the minimum distance.
5. Determine the new centroid using the following equation:

$$v_{ij} = \frac{1}{Ni}\sum_{k=0}^{Ni} Xkj \qquad (2)$$

2

Information:
$\bar{v}$ij = the i-th cluster centroid/average for the jth variable
Ni = the amount of data that is a member of the i-th cluster
i, k = index of clusters
j = index of the variable
Xkj= the k-th data value in the cluster for the j-variable

6. Return to steps 3, 4 and 5.

If in the second iteration, no cluster members move to another cluster then the iteration stops but if a cluster member moves to another cluster then return to step number 3, 4 and 5. Do the next iteration until no cluster member moves to the cluster other [20].

2.2 Davies Bouldin Index

Davies Bouldin Index (DBI) is also called the index classification reliability. The Davies Bouldin Index (DBI) introduced by David L. Davies and Donald W. Bouldin in 1979 is defined as the ratio of the average distance within and between clusters for each cluster to its nearest neighbor cluster. The Davies Bouldin Index is a measure for evaluating clustering performance. DBI has a positive correlation for the "within-class" case and a negative correlation for the "between-class" case. Use DBI as a clustering metric due to the common way of clustering. Validation contains two main categories - external validation and internal validation which are used to assess the performance of clustering results [9].

The Davis Bouldin Index (DBI) is a method for evaluating clustering models. Selection of the evaluation results is done by comparing the smallest value of the cluster models that are made. The number of clusters selected is by looking at the smallest BDI value [21].

3. RESULT AND DISCUSSION

Before data processing is carried out, the data analysis stage is carried out first, which consists of the process of selecting data on a number of datasets that have been collected by sorting the data according to research needs. The selected data comes from the Uci Machine Learning dataset which consists of iris data, then proceeds to the data preprocessing stage which consists of the data cleaning stage, namely selecting the attributes and data fields to be used and removing attributes and data that are not needed in processing, then the third The dataset resulting from the cleaning process is merged or integrated data into one complete dataset.

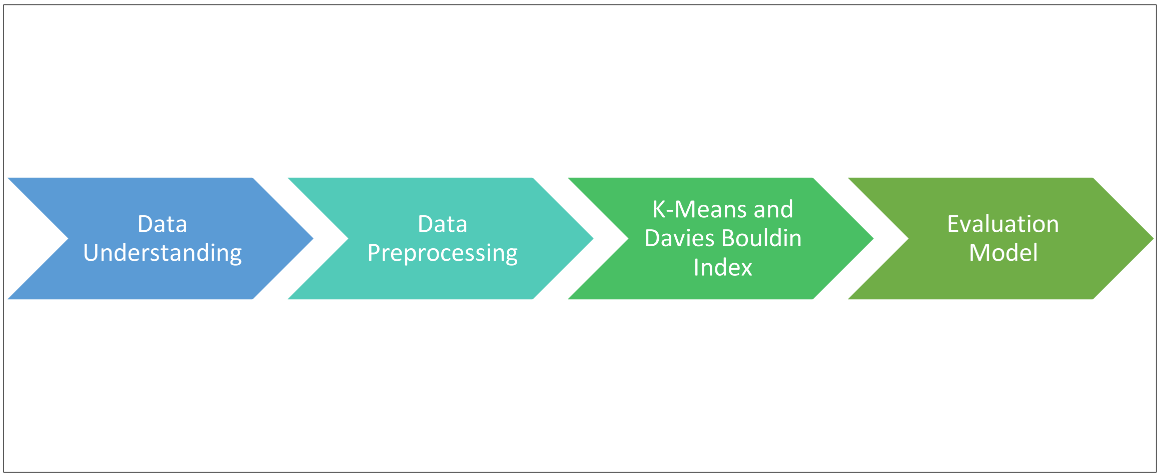

Figure 3. Research Methodology

The next step is to carry out the process of transforming the data according to the calculation requirements. At this stage, several attribute names and data entries were changed into a certain numeric code to facilitate the calculation process. The following figure presents the attributes that will be used which are the results of the integration of the three previous subdatasets and have been initialized to make calculations easier.

3.1 Data Understanding

In the data collection process, literature studies are carried out by gathering knowledge and references from various sources of literature such as references from scientific journals and scientific papers related to the topic to be studied. Then proceed with collecting some of the required data samples and will be used in the data mining processing process. The data source used comes from the Uci Machine Learning data set, namely iris data consisting of 4 attributes used with a total of 100 data. The data is the overall data which will then be carried out by a selection process to become a dataset that will be managed in research. The following is the dataset used.

Row No.	id	label	a1	a2	a3	a4
1	id_1	Iris-setosa	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	4.900	3	1.400	0.200
3	id_3	Iris-setosa	4.700	3.200	1.300	0.200
4	id_4	Iris-setosa	4.600	3.100	1.500	0.200
5	id_5	Iris-setosa	5	3.600	1.400	0.200
6	id_6	Iris-setosa	5.400	3.900	1.700	0.400
7	id_7	Iris-setosa	4.600	3.400	1.400	0.300
8	id_8	Iris-setosa	5	3.400	1.500	0.200
9	id_9	Iris-setosa	4.400	2.900	1.400	0.200
10	id_10	Iris-setosa	4.900	3.100	1.500	0.100
11	id_11	Iris-setosa	5.400	3.700	1.500	0.200
12	id_12	Iris-setosa	4.800	3.400	1.600	0.200
13	id_13	Iris-setosa	4.800	3	1.400	0.100
14	id_14	Iris-setosa	4.300	3	1.100	0.100
15	id_15	Iris-setosa	5.800	4	1.200	0.200
16	id_16	Iris-setosa	5.700	4.400	1.500	0.400
17	id_17	Iris-setosa	5.400	3.900	1.300	0.400
18	id_18	Iris-setosa	5.100	3.500	1.400	0.300
19	id_19	Iris-setosa	5.700	3.800	1.700	0.300
20	id_20	Iris-setosa	5.100	3.800	1.500	0.300
21	id_21	Iris-setosa	5.400	3.400	1.700	0.200
22	id_22	Iris-setosa	5.100	3.700	1.500	0.400

Figure 4. Data Iris

3.2 Data Preprocessing

Preprocessing is the process of cleaning and standardizing data, as user-generated raw data is usually unstructured and unanalyzable for sentiment analysis [20]. At this preprocessing stage, the data cleaning process is carried out. Data cleaning is removing attributes that will not be used and fields that are not valued (null) which are not needed in the calculation process later, so that the data to be processed is truly relevant data. Then after the cleaning stage of all the sub-datasets, the data integration process is then carried out or the merging of several sub-data sets so that they become one complete dataset that is ready for the clustering calculation process.

3.3 K-Means

The K-Means Clustering process uses the Davies Bouldin Index, which in this study uses rapid miner tools in processing data, the data used is iris data, from iris data it is necessary to select attributes by type using a subset filter the author uses 4 attributes from iris data criteria , the data that has been selected for the attribute is then linked to multiply where the author will use k-means clustering by dividing the cluster values 3,5,7,9 where each cluster has a value and to measure performance in each cluster using the Davies Bouldin Index. The results of the K-Means Clustering test using the Davies Bouldin Index have different performance percentages where each cluster has a different percentage value. The following are the test results.

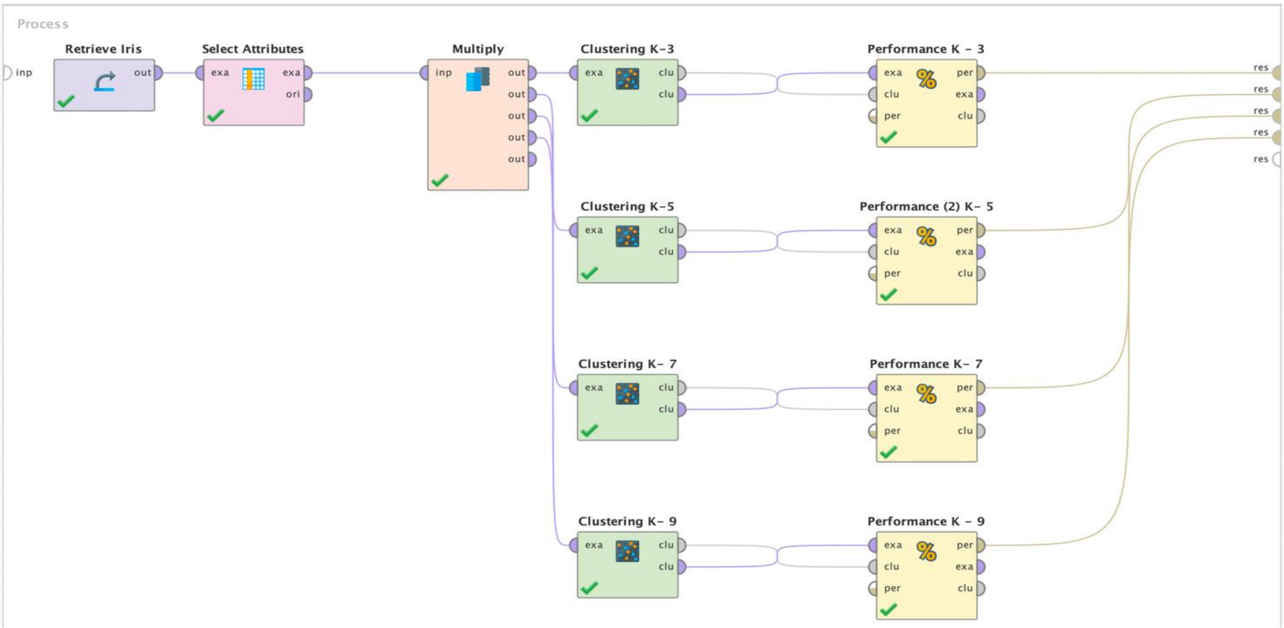

Figure 5. Rapid Miner Tools Testing

Testing using the Davies bouldin index (DBI) is one of the tests that can assist in evaluating the accuracy of the clusters in this study. DBI testing is carried out by measuring the strength or accuracy of the number of K (clusters) formed in the K-Means algorithm. The purpose of the stages of the DBI testing approach is to maximize the distance between one cluster and another and to find values to minimize the distance between document data contained in the same cluster.

Clustering Test K-3
PerformanceVector:
Avg. within centroid distance: -0.312
Avg. within centroid distance_cluster_0: -0.344
Avg. within centroid distance_cluster_1: -0.305
Avg. within centroid distance_cluster_2: -0.230
Avg. within centroid distance_cluster_3: -0.365
Avg. within centroid distance_cluster_4: -0.335
Davies Bouldin: -0.799

In testing the K-3 cluster values, the performance value of the average centroid distance is -0.312, then the K-Means optimization results with Davies Bouldin have a performance percentage of -0.799

Clustering Test K-5
PerformanceVector:
Avg. within centroid distance: -0.310
Avg. within centroid distance_cluster_0: -0.305
Avg. within centroid distance_cluster_1: -0.228
Avg. within centroid distance_cluster_2: -0.335
Avg. within centroid distance_cluster_3: -0.388
Avg. within centroid distance_cluster_4: -0.328
Davies Bouldin: -0.806

In testing the K-5 cluster values, the performance value of the average centroid distance is -0.310, then the K-Means optimization results with Davies Bouldin have a performance percentage of -0.806

Clustering Test K-7
PerformanceVector:
Avg. within centroid distance: -0.310
Avg. within centroid distance_cluster_0: -0.328
Avg. within centroid distance_cluster_1: -0.305
Avg. within centroid distance_cluster_2: -0.388
Avg. within centroid distance_cluster_3: -0.228
Avg. within centroid distance_cluster_4: -0.335
Davies Bouldin: -0.806

In testing the value of the K-7 cluster, it has a performance value of an average centroid distance of -0.310, then the results of the K-Means optimization with Davies Bouldin have a performance percentage of -0.806

Clustering Test K-9
PerformanceVector:

Avg. within centroid distance: -0.310
Avg. within centroid distance_cluster_0: -0.305
Avg. within centroid distance_cluster_1: -0.388
Avg. within centroid distance_cluster_2: -0.335
Avg. within centroid distance_cluster_3: -0.228
Avg. within centroid distance_cluster_4: -0.328
Davies Bouldin: -0.806

In testing the K-9 cluster values, the performance value of the average centroid distance is -0.310, then the K-Means optimization results with Davies Bouldin have a performance percentage of -0.806

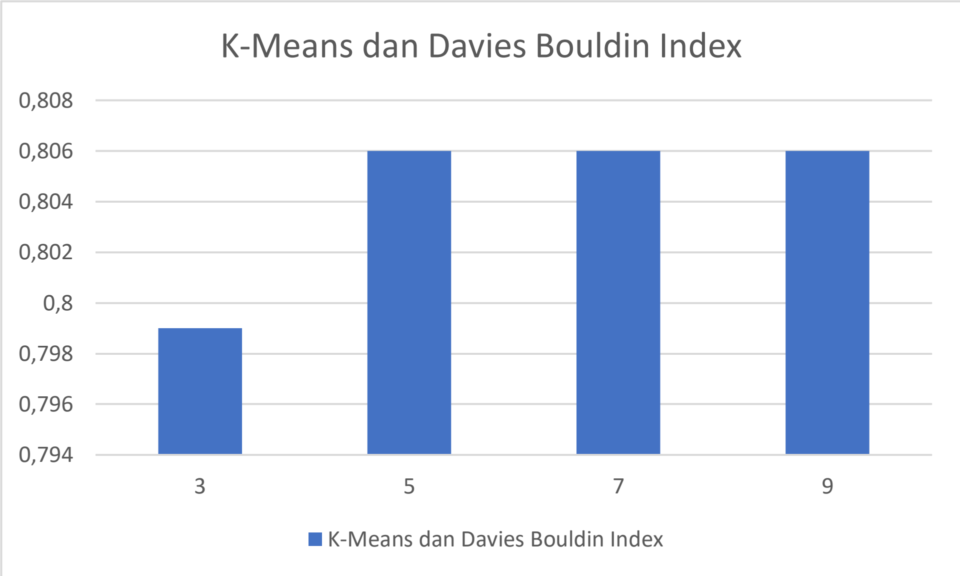

Figure 6. Results of Variation in K Values

The stages of results and discussion are the presentation of the results on the cluster and the tests carried out using the DBI test, testing the performance speed of the computational time needed in grouping iris data as well as an overview of the document membership of the two algorithms in the best cluster according to the DBI value test.

The results of the K-Means optimization study with the Davies Bouldin Index k-means clustering by dividing the cluster values 3,5,7,9. In testing the K-3 cluster values, the performance value of the average centroid distance is -0.312, then the K-Means optimization results with Davies Bouldin have a performance percentage of -0.799. Testing the value of the K-5 cluster has a performance value of an average centroid distance of -0.310, then the K-Means optimization results with Davies Bouldin have a performance percentage of -0.806. Testing the value of the K-7 cluster has a performance value of an average centroid distance of -0.310, then the K-Means optimization results with Davies Bouldin have a performance percentage of -0.806. Testing the K-9 cluster values has a performance value of an average centroid distance of -0.310, then the results of the K-Means optimization with Davies Bouldin have a performance percentage of -0.806. From the test results with variations in cluster values 3,5,7,9 it can be concluded that the optimization of the K-Means method with the Davies Bouldin Index testing the K-3 cluster values has better performance with an average value of -0.312 centroid distance then the results of K-optimization Means with Davies Bouldin has a performance percentage: -0.799.

4. CONCLUSION

The K-Means clustering algorithm is a method of grouping data that is simpler to implement or easy to implement. The process flow of this algorithm is to partition or divide a number of data into K clusters through the average (Mean) of the shortest distance of all data to the data cluster which will then be calculated to average the data points in the next iteration. The cluster center point or centroid is the starting point for grouping within the cluster in the K-Means algorithm. Data grouping is done by calculating the shortest distance to the initial cluster center point as the central point in the formation of each group or cluster. The results of the K-Means optimization study with the Davies Bouldin Index k-means clustering by dividing the cluster values 3,5,7,9. In testing the K-3 cluster values, the performance value of the average centroid distance is -0.312, then the K-Means optimization results with Davies Bouldin have a performance percentage of -0.799. Testing the value of the K-5 cluster has a performance value of an average centroid distance of -0.310, then the K-Means optimization results with Davies Bouldin have a performance percentage of -0.806. Testing the value of the K-7 cluster has a performance value of an average centroid distance of -0.310, then the K-Means optimization results with Davies Bouldin have a performance percentage of -0.806. Testing the K-9 cluster values has a performance value of an average centroid distance of -0.310, then the results of the K-Means optimization with Davies Bouldin have a performance percentage of -0.806. From the test results with variations in cluster values 3,5,7,9 it can be concluded that the optimization of the K-Means method with the Davies Bouldin Index testing the K-3 cluster values has better performance with an average value of -0.312 centroid distance then the results of K-optimization Means with Davies Bouldin has a performance percentage: -0.799.

REFERENCES

[1] DEVI TRIANA, "UNIVERSITAS SUMATERA UTARA Poliklinik UNIVERSITAS SUMATERA UTARA," *J. Pembang. Wil. Kota*, vol. 1, no. 3, pp. 82–91, 2018.

[2] D. Diy, "Analisis Data Mining Untuk Memprediksi Lama Perawatan Pasien Covid-19 Bianglala Informatika," vol. 10, no. 1, pp. 21–29, 2022.

[3] P. H. Putra, A. Hasibuan, and E. A. Marpaung, "Analisis Klasifikasi Metode X-Means Pada Minat dan Bakat Anak Dimasa Pandemi," vol. 19, no. 2, pp. 424–429, 2022.

[4] 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.

[5] S. Ramadhani, D. Azzahra, and T. Z, "Comparison of K-Means and K-Medoids Algorithms in Text Mining based on Davies Bouldin Index Testing for Classification of Student's Thesis," *Digit. Zo. J. Teknol. Inf. dan Komun.*, vol. 13, no. 1, pp. 24–33, 2022, doi: 10.31849/digitalzone.v13i1.9292.

[6] Y. Asriningtias and J. Aryanto, "K-Means Algorithm with Davies Bouldin Criteria for Clustering Provinces in Indonesia Based on Number of Events and Impacts of Natural Disasters," *Int. J. Eng.*, vol. 4, no. 1, pp. 75–80, 2022, [Online]. Available: https://dibi.bnpb.go.id/kwilayah.

[7] B. Jumadi Dehotman Sitompul, O. Salim Sitompul, and P. Sihombing, "Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm," *J. Phys. Conf. Ser.*, vol. 1235, no. 1, 2019, doi: 10.1088/1742-6596/1235/1/012015.

[8] 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.

[9] Y. Sopyan, A. D. Lesmana, and C. Juliane, "Analisis Algoritma K-Means dan Davies Bouldin Index dalam Mencari Cluster Terbaik Kasus Perceraian di Kabupaten Kuningan," *Build. Informatics, Technol. Sci.*, vol. 4, no. 3, pp. 1464–1470, 2022, doi: 10.47065/bits.v4i3.2697.

[10] A. Idrus, N. Tarihoran, U. Supriatna, A. Tohir, S. Suwarni, and R. Rahim, "Distance Analysis Measuring for Clustering using K-Means and Davies Bouldin Index Algorithm," *TEM J.*, vol. 11, no. 4, pp. 1871–1876, 2022, doi: 10.18421/TEM114-55.

[11] A. Badruttamam, S. Sudarno, and D. A. I. Maruddani, "PENERAPAN ANALISIS KLASTER K-MODES DENGAN VALIDASI DAVIES BOULDIN INDEX DALAM MENENTUKAN KARAKTERISTIK KANAL YOUTUBE DI INDONESIA (Studi Kasus: 250 Kanal YouTube Indonesia Teratas Menurut Socialblade)," *J. Gaussian*, vol. 9, no. 3, pp. 263–272, 2020, doi: 10.14710/j.gauss.v9i3.28907.

[12] D. Triyansyah and D. Fitrianah, "Analisis Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing," *J. Telekomun. dan Komput.*, vol. 8, no. 3, p. 163, 2018, doi: 10.22441/incomtech.v8i3.4174.

[13] F. N. Musid, "Implementasi Algoritma K-Means Clustering Dalam Pengelompokkan Data Jumlah Kerusakan Rumah Berdasarkan Kondisi Di Jawa Barat," vol. 1, no. 3, pp. 101–114, 2023.

[14] I. F. Ashari, R. Banjarnahor, D. R. Farida, S. P. Aisyah, A. P. Dewi, and N. Humaya, "Application of Data Mining with the K-Means Clustering Method and Davies Bouldin Index for Grouping IMDB Movies," *J. Appl. Informatics Comput.*, vol. 6, no. 1, pp. 07–15, 2022, doi: 10.30871/jaic.v6i1.3485.

[15] B. Kristanto, A. T. Zy, M. Fatchan, T. Informatika, and P. Bangsa, "Analisis Penentuan Karyawan Tetap Dengan Algoritma K-Means Dan Davies Bouldin Index," vol. 4, no. 1, pp. 112–120, 2023.

[16] M. Wahyudi and L. Pujiastuti, "Komparasi K-Means Clustering dan K-Medoids dalam Mengelompokkan Produksi Susu Segar di Indonesia Comparison of K-Means Clustering and K-Medoids in Clustering Fresh Milk Production in Indonesia," vol. 4, no. 2, pp. 243–254, 2022, doi: 10.30812/bite.v4i2.2104.

[17] A. K-means, M. Nurdayat, N. Suarna, and Y. A. Wijaya, "Analisa Clustering untuk Mengelompokan Data Penayangan Film Bioskop Menggunakan," vol. 6, no. 1, pp. 68–78, 2023.

[18] A. A. Az-zahra, A. F. Marsaoly, I. P. Lestyani, R. Salsabila, and W. O. Z. Madjida, "Penerapan Algoritma K-Modes Clustering Dengan Validasi Davies Bouldin Index Pada Pengelompokkan Tingkat Minat Belanja Online Di Provinsi Daerah Istimewa Yogyakarta," *J. MSA (Mat. dan Stat. serta Apl.)*, vol. 9, no. 1, p. 24, 2021, doi: 10.24252/msa.v9i1.18555.

[19] B. Harahap, "Penerapan Algoritma K-Means Untuk Menentukan Bahan Bangunan Laris (Studi Kasus Pada UD. Toko Bangunan YD Indarung)," *Reg. Dev. Ind. Heal. Sci. Technol. Art Life*, pp. 394–403, 2019, [Online]. Available: https://ptki.ac.id/jurnal/index.php/readystar/article/view/82.

[20] D. Rifaldi, A. Fadlil, and Herman, "Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet Mental Health," *J. Pendidik. Teknol. Inf.*, vol. 3, no. 2, pp. 161–171, 2023.

[21] P. H. Putra, B. Purba, and Y. A. Dalimunthe, "Random forest and decision tree algorithms for car price prediction," vol. 1, no. 2, pp. 81–89, 2023.