International Journal of Artificial Intelligence & Robotics (IJAIR)
Vol.4, No.1, 2022, pp.39-47

E-ISSN : 2686-6269

Optimization of the Number of Clusters of the K-Means Method in Grouping Egg Production Data in Indonesia

Solikhun[1], Verdi Yasin[2]*, Donni Nasution[3]
[1]AMIK Tunas Bangsa, Pematangsiantar, Sumatera Utara, Indonesia
[2]STMIK Jayakarta, Jakarta, Indonesia
[3]Universitas Prima Indonesia, Medan, Indonesia
[1]solikhun@amiktunasbangsa.ac.id; [2]verdiyasin29@gmail.com*, [3]nasution.donni@gmail.com
*Corresponding author

ABSTRACT

The need for eggs that continues to increase will not increase with large egg production so that there is a shortage of egg supplies which results in high egg prices. It is necessary to group egg production in Indonesia to find out which areas fall into the high cluster and which areas fall into the low cluster. This study aims to classify the egg production of laying hens in Indonesia. The method used is the K-Means Clustering method which is a popular clustering method. To find out how optimal the number of clusters in the K-Means method is for grouping egg production in Indonesia, the researcher evaluates the DBI value of each number of existing clusters. In this study, 8 clusters were used, namely 2 clusters, 3 clusters, 4 clusters, 5 clusters, 6 clusters, 7 clusters, 8 clusters, and 9 clusters. The results of measuring the DBI value are the number of clusters 2 = 0.215, the number of clusters 3 = 0.149, the number of clusters 4 = 0.146, the number of clusters 5 = 0.157, the number of clusters 6 = 0.180, the number of clusters 7 = 0.205, the number of clusters 8 = 0.192 and the number of clusters 9 = 0.154. This study shows that the best number of clusters is the number of clusters 4 with the smallest DBI value of 0.146.

Keywords : Data mining, Clustering, K Means, Egg Production, DBI Value

Article History

Received :	Revised :	Accepted :

I. INTRODUCTION

Eggs are a food ingredient that is a source of protein. Whole milk that has not been added or reduced by anything has natural ingredients except the cooling process, which does not reduce purity. Whole milk is good for the health of the body because it has good nutritional content, and pure milk has not had any processing.

Clustering is a frequent unattended machine learning technique in which data sets are automatically partitioned into clusters so that objects in the same cluster are more similar and those in separate clusters are more dissimilar [1]. K-Means is a non-hierarchical data clustering approach for dividing data into one or more clusters or groups. Data with similar features are grouped together in one cluster, while data with distinct aspects are divided into separate groups [2]–[4].

Literature study on previous research [5]–[8]. The proposes a method for grouping or classifying distinct sorts of eggs utilizing the connected components method of analysis applied to an object in a Fig. of an egg, allowing computer applications to solve Egg classification challenges [9]. The connected component analysis method was successfully applied to the classification of eggs with black backgrounds and grouping results obtained processing chicken eggs and quail eggs with a 100% success rate, as well as counting the number of egg classification results with 100% accuracy, based on test results on 10 data images. The need for eggs that continues to increase is not accompanied by large egg production, so there is a shortage of egg supply which results in high egg prices. The problem in this study is that there is no research on grouping egg production in Indonesia specifically to recommend to the government information about grouping egg production in high or low or medium clusters as material for government evaluation in determining egg production policies in Indonesia.

This study focuses on grouping egg production in Indonesia using the K-Means method. In this study, a trial was used with 8 clusters, starting from the number of clusters 2 to the number of clusters 9. To get the best number of clusters from grouping egg production in Indonesia, researchers used an evaluation based on the DBI value. A good clustering result is the one with the smallest DBI value. The best clustering results are used as recommendations to the government to increase egg production in Indonesia.

DOI: http://dx.doi.org/10.25139/ijair.v4i1.4328

II. LITERATURE REVIEW

A. Data Mining

Data mining is also a technique used in the processing of enormous amounts of data. As a result, data mining is important in a variety of fields, including industry, finance, weather, research, and technology [10]–[12].

B. Analysis Cluster

The popular Partitioning Around Medoids (PAM) algorithm, also known as K-Medoids clustering, is one of the most extensively used algorithms for grouping non-Euclidean data, aside from hierarchical clustering. The mean, as used in K-means, is an excellent estimator for cluster centers in Euclidean geometry, but it is absent for arbitrary dissimilarities.[13].

C. K-Means

This non-hierarchical approach is known as the K-Means clustering method. Grouping in this method is done based on the smallest distance between the object and the centre of the cluster metode non-hierarki ini dikenal dengan metode K-Means clustering. In simple terms the K-Means algorithm starts from the following stages:

- Select *K* centroid points.
- Calculate the distance data with the centroid using Equation (1).

$$D_{(i,f)} = \sqrt{(X_{1i} - X_{1j})^2 + \left(X_{2i} - X_{2j}\right)^2 + \cdots + (X_{ki} - X_{kj})^2} \quad (1)$$

- Update the value of the centroid point.
- Repeat steps 2 and 3 until the value of the centroid point no longer changes [14].

D. Related Research

Study[15], The outcome is information on stunting data derived from the ideal number of clusters, with values of 23,403 and 1,178 for the highest SSE and smallest DBI, respectively, for k=5. Study [16], The number of villages or sub-districts based on the presence and kind of small and micro industries was utilized as the data. The test is carried out by determining the smallest value from the DBI, where the smallest value in the number of clusters is 0.175 after the data has been processed. Study [17], the DBI value for the conventional K-Means was 2.1972, while the DBI for the K-Means which had been reduced from 1 attribute to 5 attributes obtained an average DBI value of 2.0290, 1.8771, 1.8641, 1.8389, and 1.8117. Study [14], The method used in this study is the K- method. To overcome the weakness of the K-Means method in determining the number of clusters, the Elbow method is used. This method obtains a comparison of the number of clusters added by calculating the SSE of each cluster value. From the computed SSE value, it is found that the best number of clusters is 3, where the number of members of cluster 1 is 47 customers, cluster 2 has 18 customers, and cluster 3 has eight customers.

III. RESULT AND DISCUSSION

The research data was taken from the Indonesian National Statistics Center in the form of egg production data in Indonesia from 2018 to 2020. The data consists of 34 provinces in Indonesia. The following is data on egg production in Indonesia in Indonesia as seen in Fig. 1.

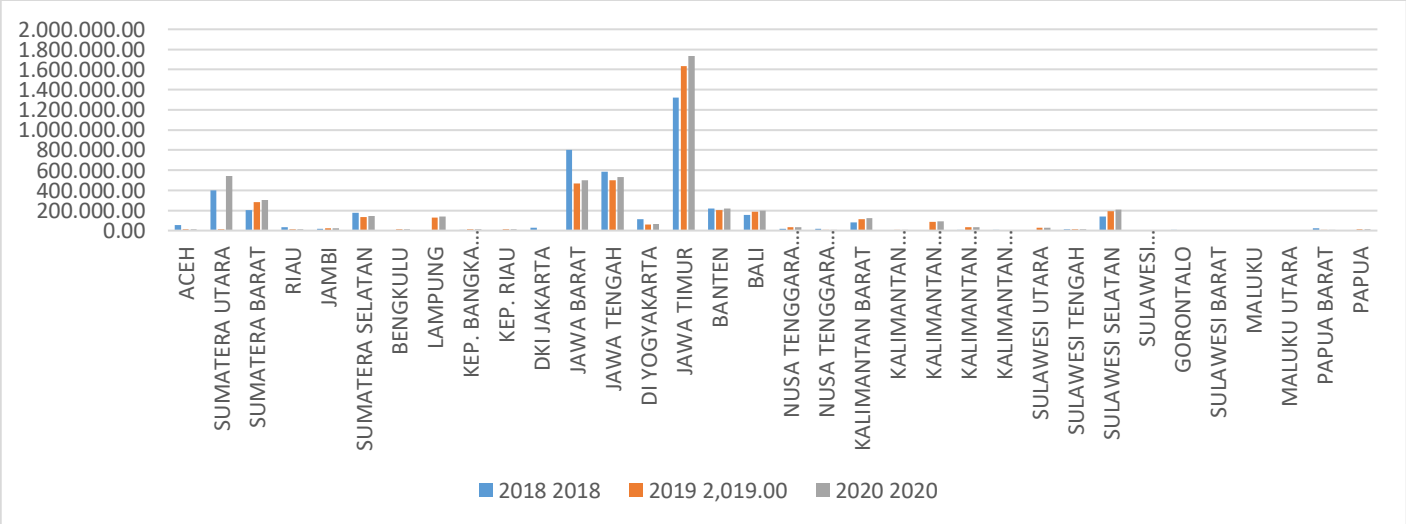


Fig. 1. Egg Production Data in Indonesia

A. Normalization Data

Egg production data in Indonesia is normalized first as shown in Table I.

TABLE I
NORMALIZATION DATA

2018	2019	2020
0.043223	0.007542	0.007542
0.304412	0.007615	0.313895
0.152299	0.174050	0.174050
0.026515	0.007488	0.007488
0.013254	0.013906	0.013906
0.135917	0.083802	0.083802
0.001031	0.006641	0.006641
0.085805	0.080138	0.080138
0.004397	0.006481	0.006481
0.002929	0.008568	0.008568
0.021890	0.000000	0.000000
0.608037	0.287212	0.287212
0.444103	0.306054	0.306054
0.084529	0.038497	0.038497
1.000000	1.000000	1.000000
0.165141	0.125193	0.125193
0.117509	0.114126	0.114126
0.015481	0.020794	0.020794
0.014365	0.005628	0.005628
0.060882	0.070691	0.070691
0.001175	0.004527	0.004527
0.081876	0.053156	0.053156
0.006507	0.020427	0.020427
0.005555	0.000372	0.000372
0.010805	0.016287	0.016287
0.009140	0.008475	0.008475
0.106529	0.119235	0.119253
0.003467	0.001735	0.001735
0.004680	0.002339	0.002339
0.000000	0.001520	0.001503
0.000004	0.000379	0.000379
0.000000	0.000054	0.000054
0.018914	0.005039	0.005039
0.000122	0.007482	0.007482

After the data is normalized, it is continued with the grouping process using calculation of the K-Means Algorithm with the Number of Clusters 3 with the following steps:

- Determination of Centroid value with the result in Table II.

DOI: http://dx.doi.org/10.25139/ijair.v4i1.4328

TABLE II
CENTROID VALUE

Centroid	2018	2019	2020
C1	0.021890	0.000000	0.000000
C2	0.444103	0.306054	0.306054
C3	1.000000	1.000000	1.000000

- Calculate the distance from centroid. Calculation of distance using Euclidian Distance iteration-1 as the results can be seen in Table III. The K-Means algorithm stops when the cluster members do not change.
The calculation on iteration-1 and so on up to Dx34,c1, namely:

$$Dx1, c1 = \sqrt{(0.043223 - 0.021890)^2 + (0.007542 - 0.00000)^2 + (0.007542 - 0.00000)^2} = 0.043223$$

$$Dx2, c1 = \sqrt{(0.304412 - 0.021890)^2 + (0.007615 - 0.00000)^2 + (0.313895 - 0.00000)^2} = 0.304412$$

Calculate the distance from the 2nd Centroid and so on up to Dx34,c2, namely:

$$Dx1, c2 = \sqrt{(0.043223 - 0.444103 + (0.007615 - 0.00000)^2 + (0.007542 - 0.00000)^2} = 0.007542$$

$$Dx2,c2 = \sqrt{(0.304412 - 0.444103)^2 + (0.007615 - 0.306054)^2 + (0.313895 - 0.306054)^2} = 0.007615$$

Calculate the distance from the 3rd Centroid and the calculation is continued until Dx34,c3, namely

$$Dx1,c3 = \sqrt{(0.043223 - 1.00000 + (0.007542 - 1.00000)^2 + (0.007542 - 1.00000)^2} = 0.007542$$

$$Dx2, c3 = \sqrt{(0.304412 - 1.00000 + (0.007615 - 1.00000)^2 + (0.3138995 - 1.00000)^2} = 0.313895$$

TABLE III
ITERATION CENTROID DISTANCE TO -1

CLUSTER 1	CLUSTER 2	CLUSTER 3	DISTANCE SHORTEST
0.023850	0.582172	1.698637	0.023850
0.422382	0.329607	1.392627	0.329607
0.278555	0.346411	1.443255	0.278555
0.011555	0.593855	1.708167	0.011555
0.021480	0.596936	1.708341	0.021480
0.164462	0.440194	1.557394	0.164462
0.022876	0.612870	1.723793	0.022876
0.130112	0.480056	1.589983	0.130112
0.019749	0.610599	1.722030	0.019749
0.022502	0.609615	1.720472	0.022502
0.000000	0.604651	1.719505	0.000000
0.713127	0.166085	1.081558	0.166085

DOI: http://dx.doi.org/10.25139/ijair.v4i1.4328

CLUSTER 1	CLUSTER 2	CLUSTER 3	DISTANCE SHORTEST
0.604651	0.000000	1.127893	0.000000
0.082991	0.521985	1.639227	0.082991
1.719505	1.127893	0.000000	0.000000
0.227743	0.378473	1.492503	0.227743
0.187597	0.424660	1.532427	0.187597
0.030098	0.588612	1.699107	0.030098
0.010953	0.604307	1.717274	0.010953
0.107307	0.507593	1.615294	0.107307
0.021683	0.614836	1.726147	0.021683
0.096174	0.509042	1.623569	0.096174
0.032728	0.595531	1.704745	0.032728
0.016344	0.615799	1.728419	0.016344
0.025562	0.596388	1.707012	0.025562
0.017499	0.605228	1.716988	0.017499
0.188685	0.428664	1.532890	0.188685
0.018586	0.615939	1.728047	0.018586
0.017525	0.614474	1.726648	0.017525
0.021995	0.618644	1.730306	0.021995
0.021893	0.619757	1.731611	0.021893
0.021891	0.620081	1.731988	0.021891
0.007722	0.601670	1.715350	0.007722
0.024204	0.612708	1.723352	0.024204

B. Evaluation of the Number of Clusters Based on the DBI Value

Researchers evaluated the DBI value. The better the clustering outcomes, the lower the DBI value. The number of clusters 2, the number of clusters 3, the number of clusters 4, the number of clusters 5, the number of clusters 6, the number of clusters 7, the number of clusters 8, and the number of clusters 9 were all used by the researchers. The DBI value is calculated using the formula below.:

- Finding SSW using the Equation (1). Where, the m_i variable is the amount of data in the i-th cluster, the *X* variable is data in cluster, the *D(x,c)* variable is distance data to centroid, the X_j variable is data on the cluster, and the C_i variable is centroid cluster i.

$$SSW_i = \frac{1}{m_i} \sum_{j=i}^{m_i} d(x_j, c_i) \quad (2)$$

- Finding SSB with using Equation (3). Where, the c_i variable is cluster 1, the c_j variable is another cluster, and the $d(c_i c_j)$ variable is distance from centroid sat to other.

$$SSB_{ij} = d(c_i c_j) \quad (3)$$

- Finding ratio using Equation (4). Where, the R_{ij} variable is the ratio between clusters, the SSW_i variable is cluster 1, the SSW_j variable is cluster 2, and the SSB_{ij} variable is separation of cluster 1 and 2.

$$R_{ij} = \frac{SSW_i + SSW_j}{SSB_{ij}} \quad (4)$$

- Finding DBI using Equation (5). Where, the K variable is existing cluster, the $R_{i,j}$ variable is ratio betwen cluster i and j, and the Max variable is finding the largest inter-cluster ratio

DOI: http://dx.doi.org/10.25139/ijair.v4i1.4328

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

The results of each number of clusters in the K-Means algorithm starting from the number of clusters 2, the number of clusters 3, the number of clusters 4, the number of clusters 5, the number of clusters 6, the number of clusters 7, the number of clusters 8, the number of clusters 9 can be seen in the value graph. DBI of each number of clusters as seen in Fig. 2

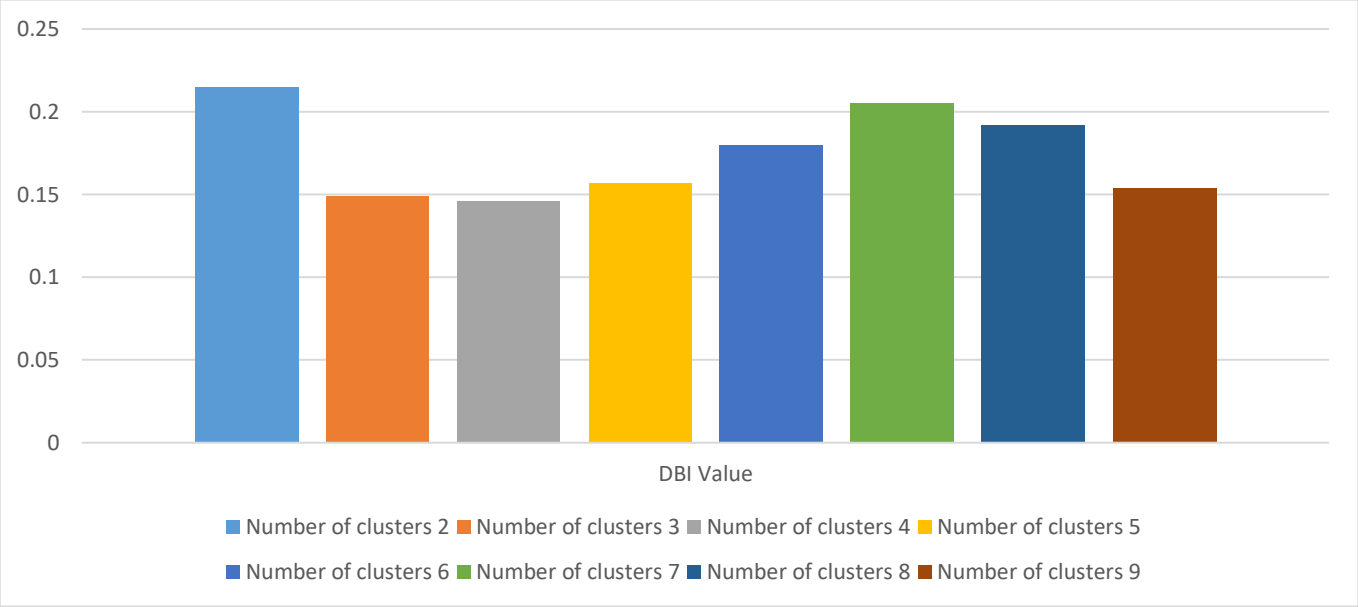


Fig. 2. The results of the measurement of the DBI value from each number of clusters

C. Processing, Grouping Results and DBI Values From Several forms of the number of clusters K-Means Clustering

From the evaluation results based on the DBI value from the form of each number of clusters, the DBI values obtained are as follows in Table IV:

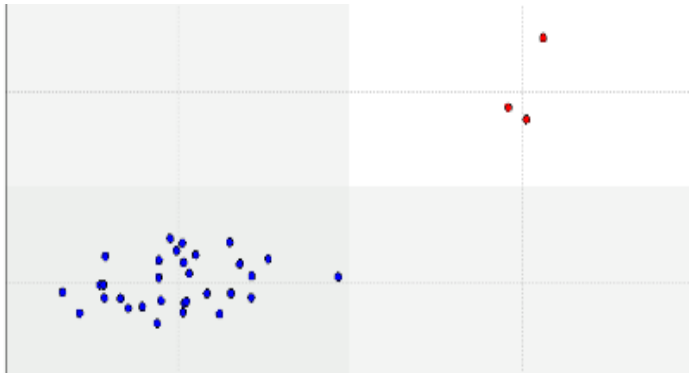
TABLE IV
COMPARISON OF DBI VALUES FROM SEVERAL FORMATS OF NUMBER OF CLUSTER

No	Number Of Clusters	DBI Value
1	2	0.215
2	3	0.149
3	4	0.146
4	5	0.157
5	6	0.180
6	7	0.205
7	8	0.192
8	9	0.154

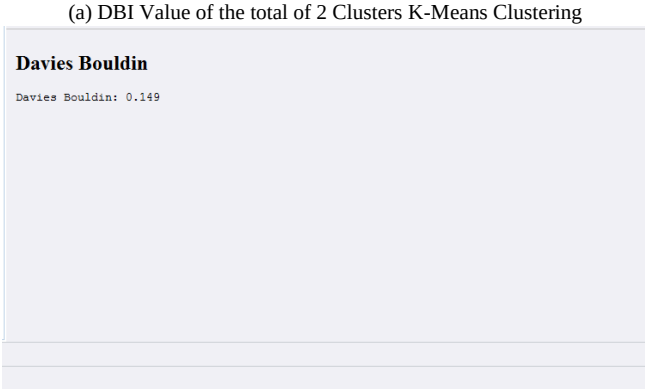
From the Table IV, the best form of cluster is cluster 4, consisting of very high, high, low and very low. From the condition of the best number of clusters, we get the results of grouping egg production in Indonesia, namely very high clusters is East Java, high clusters is North Sumatra, West Java and Central Java, low clusters 22 provinces, and very low clusters 8 provinces. Fig. 3 is the result of the DBI K-Means Clustering value with a total of 2 to 9 clusters.

DOI: http://dx.doi.org/10.25139/ijair.v4i1.4328

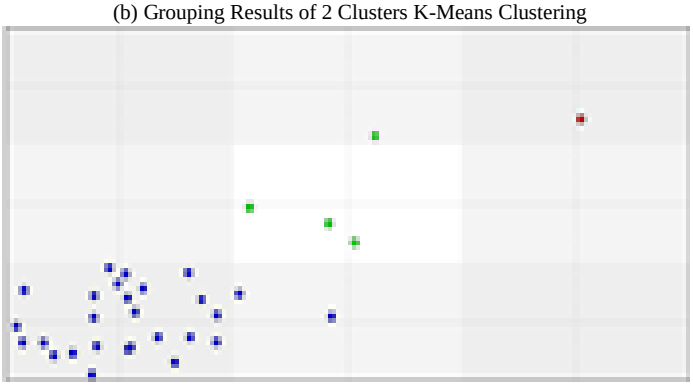
(a) DBI Value of the total of 2 Clusters K-Means Clustering



(b) Grouping Results of 2 Clusters K-Means Clustering

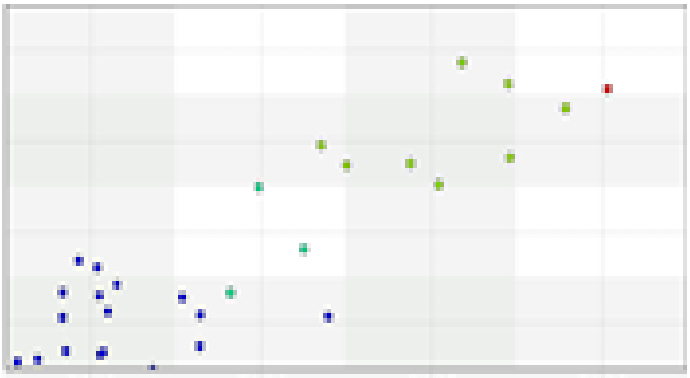


(c) DBI Value of the total of 3 Clusters K-Means Clustering



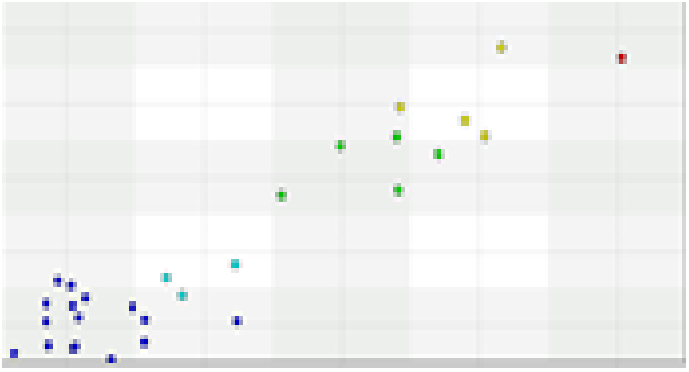
(d) Grouping Results of 3 Clusters K-Means Clustering

(e) DBI Value of the total of 4 Clusters K-Means Clustering



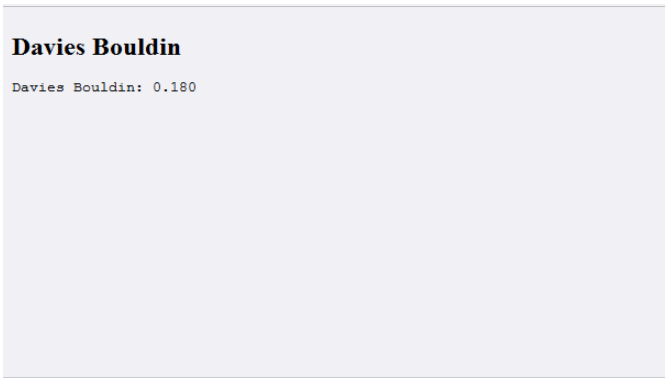
(f) Grouping Results of 4 Clusters K-Means Clustering

(g) DBI Value of the total of 5 Clusters K-Means Clustering

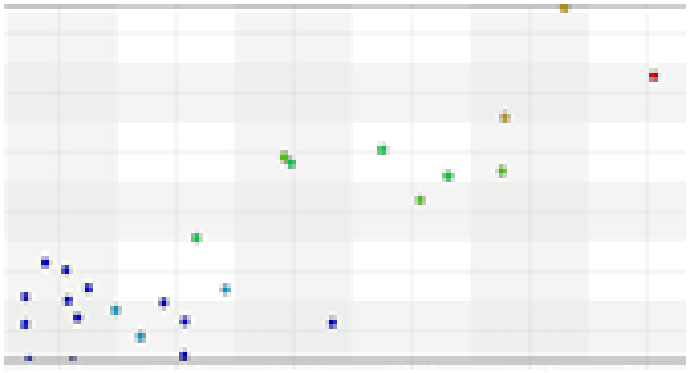


(h) Grouping Results of 5 Clusters K-Means Clustering

DOI: http://dx.doi.org/10.25139/ijair.v4i1.4328

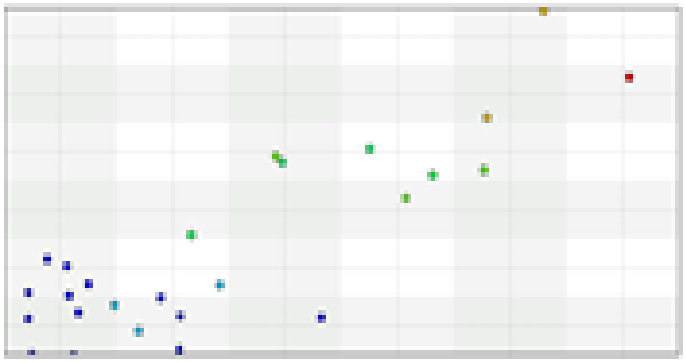


(i) DBI Value of the total of 6 Clusters K-Means Clustering



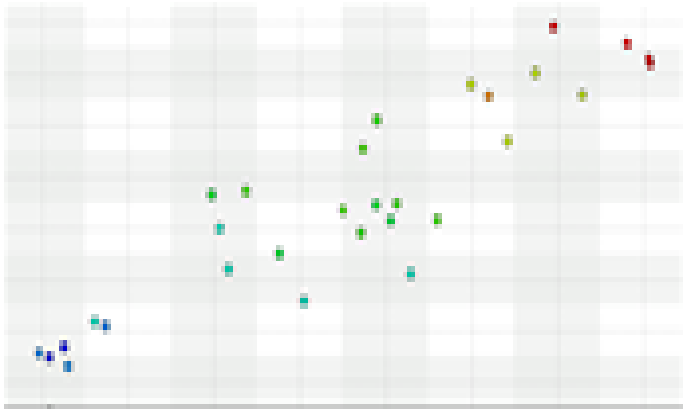
(j) Grouping Results of 6 Clusters K-Means Clustering

(k) DBI Value of the total of 7 Clusters K-Means Clustering

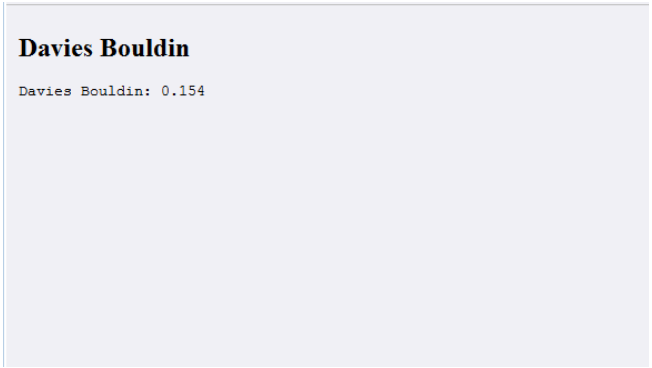


(l) Grouping Results of 7 Clusters K-Means Clustering

(m) DBI Value of the total of 8 Clusters K-Means Clustering



(n) Grouping Results of 8 Clusters K-Means Clustering



(o) DBI Value of the total of 9 Clusters K-Means Clustering

(p) Grouping Results of 9 Clusters K-Means Clustering

Fig. 3. The Result Of The DBI K-Means Clustering Value With A Total Of 2 To 9 Clusters

DOI: http://dx.doi.org/10.25139/ijair.v4i1.4328

IV. CONCLUSION

The study's findings include a comparison of the number of clusters in the grouping of egg production in Indonesia based on the DBI value. The number of clusters 4 with the K-Means Clustering algorithm has the greatest DBI value, which is 0.146. The findings of applying the K-Means Clustering method to group the number of four clusters in Indonesian egg production are very high clusters: East Java has three high clusters: North Sumatra, West Java, and Central Java; 22 low clusters; and eight extremely low clusters: West Sumatra, South Sumatra, Lampung, Banten, Bali, West Kalimantan, South Kalimantan, and South Sulawesi.

REFERENCES

[1] R. Setiawan, "PENERAPAN DATA MINING MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING UNTUK MENENTUKAN STRATEGI PROMOSI MAHASISWA BARU (Studi Kasus : Politeknik LP3I Jakarta)," vol. 3, no. 1, pp. 76–92, 2016.
[2] N. H. Kristanto, A. C. L. A, and H. B. S, "Implemantasi K-Means Clustering untuk Pengelompokan Analisis Rasio Profitabilitas dalam Working Capital," *Juisi*, vol. 02, no. 01, pp. 9–15, 2016.
[3] H. Effendi, A. Syahrial, S. Prayoga, and W. D. Hidayat, "Penerapan Metode K-Means Clustering Untuk Pengelompokan Lahan Sawit Produktif Pada PT Kasih Agro Mandiri," *Teknomatika*, vol. 11, no. 02, pp. 117–126, 2021.
[4] Y. H. Susanti and E. Widodo, "Perbandingan K-Means dan K-Medoids Clustering terhadap Kelayakan Puskesmas di DIY Tahun 2015," *Pros. SI MaNIs (Seminar Nas. Integr. Mat. dan Nilai Islam.*, vol. 1, no. 1, pp. 116–122, 2017.
[5] R. R. Muhima, M. Kurniawan, and O. T. Pambudi, "A LOF K-Means Clustering on Hotspot Data," *Int. J. Artif. Intell. Robot.*, vol. 2, no. 1, p. 29, 2020.
[6] M. Kurniawan, R. R. Muhima, and S. Agustini, "Comparison of Clustering K-Means, Fuzzy C-Means, and Linkage for Nasa Active Fire Dataset," *Int. J. Artif. Intell. Robot.*, vol. 2, no. 2, p. 34, 2020.
[7] M. Z. Sarwani, D. A. Sani, and F. C. Fakhrini, "Personality Classification through Social Media Using Probabilistic Neural Network Algorithms," *Int. J. Artif. Intell. Robot.*, vol. 1, no. 1, p. 9, 2019.
[8] R. Hariyanto and M. Z. Sarwani, "Optimizing K-Means Algorithm by Using Particle Swarm Optimization in Clustering for Students Learning Process," *Inf. J. Ilm. Bid. Teknol. Inf. dan Komun.*, vol. 6, no. 1, pp. 65–68, 2021.
[9] I. Ruslianto, "PUYUH MENGGUNAKAN METODE CONNECTED," vol. 3, no. 1, pp. 41–50, 2013.
[10] E. G. Sihombing, "KLASIFIKASI DATA MINING PADA RUMAH TANGGA MENURUT PROVINSI DAN STATUS KEPEMILIKAN RUMAH KONTRAK / SEWA MENGGUNAKAN K-MEANS CLUSTERING METHOD," vol. 2, no. 2, pp. 74–82, 2017.
[11] E. Nanda, Solikun, and Irawan, "PENERAPAN DATA MINING DALAM MENGELOMPOKAN PRODUKSI JAGUNG MENURUT PROVINSI MENGGUNAKAN ALGORITMA K-MEANS," vol. 3, pp. 702–709, 2019.
[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] N. Suryana, "Penggunaan metode statistik K-Means clustering pada analisis peruntukan lahan usaha tambang berbasis sistem informasi geografi," *J. Teknol. Miner. dan Batubara*, vol. 7, no. 1, pp. 42–53, 2011.
[14] E. Muningsih, "Optimasi Jumlah Cluster K-Means Dengan Metode Elbow Untuk Pemetaan Pelanggan," *Pros. Semin. Nas. ELINVO*, no. September, pp. 105–114, 2017.
[15] D. Jollyta, S. Efendi, M. Zarlis, and H. Mawengkang, "Optimasi Cluster Pada Data Stunting: Teknik Evaluasi Cluster Sum of Square Error dan Davies Bouldin Index," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, no. September, p. 918, 2019.
[16] E. Muningsih, I. Maryani, and V. R. Handayani, "Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi Berdasarkan Potensi Desa," *J. Sains dan Manaj.*, vol. 9, no. 1, pp. 95–100, 2021.
[17] R. K. Dinata, H. Novriando, N. Hasdyna, and S. Retno, "Reduksi Atribut Menggunakan Information Gain untuk Optimasi Cluster Algoritma K-Means," *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 1, p. 48, 2020.

DOI: http://dx.doi.org/10.25139/ijair.v4i1.4328