Clustering of Potency of Shrimp in Indonesia with K-Means Algorithm and Validation of Davies-Bouldin Index

Firman Tempola[1], Achmad Fuad Assagaf[2]
Department of Informatics
Universitas Khairun
Ternate, Indonesia
[1]firman.tempola@unkhair.ac.id

***Abstract*—Among ASEAN countries, Indonesia is first ranked shrimp producers. Although shrimp from Indonesia is one of the main producers in the ASEAN region, in fact, the potential of each province has not been explained in detail, so it needs to be studied further related to shrimp production in every province in Indonesia. This research will do mapping the potential of shrimp per province in Indonesia. The method applied in doing the mapping is by using K-Means Clustering algorithm, with clustering validation used is Davies-Bouldin index. The result showed that clustering position was in the sixth iteration for two clusters, cluster 1 became the best cluster with the value of SSW 0.13507 and the value of DBI 0,96. As for the division of 3 clusters only occur once an iteration, the first cluster becomes the best cluster with the value of SSW 0.095 and the value of DBI 0.76. Thus, from the distribution of the number of clusters formed then the division of 3 clusters is better than the division of 2 clusters.**

Keywords—shrimp; clustering K-Means; Davies Bouldin Index

I. INTRODUCTION

The fisheries and maritime sector are also one of the sectors that contribute greatly to Gross Domestic Product (GDP) in the Ministry of Fisheries report in the last 3 years GDP in the fisheries sector continues to increase In 2012, Indonesia's fisheries GDP is Rp. 184.25 trillion and contributes 2, 14 percent of national GDP. In 2013, its contribution increased to 2.21 per cent of national GDP. This figure continues to increase in 2014 with a value of Rp 247.09 trillion or contributing 2.34 per cent of national GDP. Whereas in 2015, the fisheries sector contributed IDR 288.92 trillion to GDP with a contribution of 2.51 per cent and 2016 amounting to IDR 317.09 trillion rupiahs with a contribution of 2.56 per cent [1].

Not only that, but other Indonesian territorial waters also have abundant fisheries potential. Namely the Arafura Sea as much as 855 thousand tons, Java Sea 836 thousand tons, Tomini Bay 595 thousand tons, the Indian Ocean in Sumatra west 565 thousand tons, Indian Ocean south side of Java Island as much as 491 thousand tons. And other waters such as the Sea, the Banda Sea, the Malacca Straits of the Pacific Ocean have a fishery potential of around 300 thousand tons.

Indonesian fisheries resources do not stop there. Not to mention superior commodities such as shrimp, crab, and seaweed. Like shrimp, for example, Indonesia ranks first as the largest shrimp producer in ASEAN. Even dubbed as the king of ASEAN shrimp with total production reaching 645 thousand tons in 2014 and donating foreign exchange worth up to billions of US $.

Although shrimp from Indonesia is one of the main producers in the ASEAN region, in fact, the potential of each province has not been explained in detail, so it needs to be studied further related to shrimp production in every province in Indonesia. In carrying out the process of mapping the potential of shrimp in Indonesia, a method or method is needed that can map the potential of each province by applying the clustering method.

The application of clustering has also been carried out in various fields, for example, clustering in the biometric field [2]. Clustering determination of potential crime areas in the city of Banjarbaru with the k-means method [3]. Whereas in the field of fisheries, it has been carried out by [4] where objects in the cluster are regencies in East Java. Whereas, in this study, a cluster will be conducted by the province in Indonesia with the potential of fishery commodities in the cluster are shrimp.

There are two shrimp commodities that are the mainstay, namely Windu Shrimp and Vaname Shrimp. In addition, currently other shrimp such as Prawn Shrimp are also being promoted even though the production is still small compared to the commodity of Shrimp and Vaname Shrimp, but the development is quite good in recent years. In this study, the clustering process used the K-Means method and validation using the Davies Bouldin Index.

II. METHODOLOGY

A. Clustering Data Mining

In any case at the moment, it always produces data. Whether in the fields of economics, health, astronomy, fisheries, agriculture, education, technology and law. Data is a reference in the decision-making process. In the process of giving recommendations or decisions at this time, one of the techniques carried out is by using data mining techniques. Large companies today require data mining techniques to improve efficiency and effectiveness in their business processes.

Clustering is one part of data mining techniques by learning without guidelines or usually referred to as unsupervised learning. Clustering is used to group objects into groups based on similarities between objects, wherein one cluster must contain objects that are similar to each other. This clustering without using training data as in the classification must use training data [5]. In Figure 1 is a step in the implementation of the Clustering K-Means Algorithm on the production of shrimp potential in Indonesia with data validation of Davies-Bouldin index.

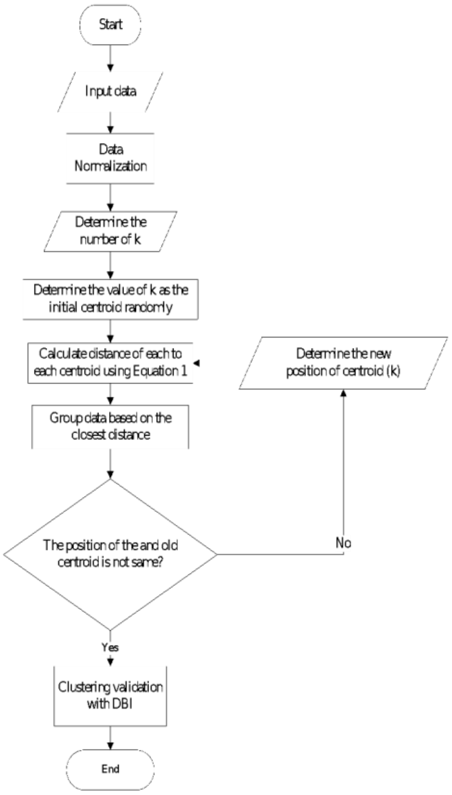
Start
Input data
Data Normalization
Determine the number of k
Determine the value of k as the initial centroid randomly
Calculate distance of each to each centroid using Equation 1
Determine the new position of centroid (k)
Group data based on the closest distance
The position of the and old centroid is not same?
No
Yes
Clustering validation with DBI
End

Fig. 1. *Steps by K-Means Algorithm*

B. K-Means Algorithm

From several clustering techniques, the simplest and most commonly known is clustering k-means. In this technique, objects are grouped into k or clusters. To do this clustering, the current value of the k is determined first, but previously the object is studied and how many clusters are the most appropriate. K-Means method tries group existing data into several, where data is in one group have the same characteristics with each other and has different characteristics with data who is in another group. The basic K-means algorithm is as follows [6].

1. Determine the value of k as the number of clusters to be formed.
2. Initialize k as a centroid which can be generated randomly.
3. Calculate the distance of each data to each centroid using the Euclidean Distance Equation 1 as follows:

$$d(x_i, y_i) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

4. Group each data based on the closest distance between the data and the centroid.
5. Determine the position of the new centroid (k)
6. Go back to step 3 if the new centroid by centroid position is not the same old.

C. Davies-Bouldin Index

The Davies-Bouldin index (DBI) metric was introduced by [7] used to evaluate clusters. The internal validity that is done is how well clustering has been done by calculating the quantity and the derivative features of the data set. Essentially the DBI value is closer to non-negative zero to be able to judge the goodness of the cluster obtained. The equation used is as in Equation 1 below.

$$DBI = \frac{1}{k}\sum_{i=1,\, i=j}^{k}\max(R_{i,j}) \tag{2}$$

k is the number of used clusters Whereas $R_{i,j}$ is the ratio of the ratio between the i-cluster and the j-cluster. The value is obtained from the cohesion and separation components. A good cluster is one that has the smallest possible cohesion and the largest possible separation. $R_{i,j}$ formulated in the following Equation 3.

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \tag{3}$$

SSB (sum of square between cluster) is a metric for separating between two clusters, for example, cluster *i* and *j*, by measuring the distance between centroid c_i and c_j as in Equation 4 below.

$$SSB_{i,j} = d(c_i, c_j) \tag{4}$$

For SSW (Sum of Square within a cluster) is a cohesion metric in an i-cluster which is formulated as in Equation 5 [7].

$$SSW_i = \frac{1}{m}\sum_{j=1}^{k} d(x_j, c_i) \tag{5}$$

mi is the amount of data that is in the *i* cluster, while c_i is the centroid of cluster the-*i*. For the value d in equation 5 can use euclidean distance or Equation 1.

D. Dataset

The dataset used in this study is data obtained from the Ministry of Maritime Affairs and Fisheries, namely data on aquaculture production in 2013 which included shrimp production data divided by province and type of shrimp. There are 3 types of shrimp selected as criteria in conducting clustering, namely Tiger shrimp, Vaname shrimp and other shrimp. Dataset in the form of shrimp types are shown in Table 1.

TABLE I. PRODUCE TYPE SHRIMP BASED ON PROVINCE IN INDONESIA

No	Province	Windu Shrimp (ton)	Vaname Shrimp (ton)	Shrimp others (Ton)
1	Aceh	5621	1244	748
2	North Sumatra	9627	19791	0
3	West Sumatra	2	3	23
4	Riau	27	32	0
5	Riau Islands	0	32	0
6	Jambi	0	0	1
7	South Sumatra	5641	40016	1
8	Bangka Belitung	0	710	0
9	Bengkulu	278	945	3
10	Lampung	2791	72051	129
11	D.I. Yogyakarta	0	809	219
12	Central Java	33580	13872	16506
13	West Java	27860	57678	16270
14	Banten	404	1407	1093
15	DKI Jakarta	201	0	0
16	East Java	9842	47150	7302
17	Bali	0	2932	449
18	NTB	4299	56960	168
19	NTT	0	13	0
20	West Kalimantan	1865	39092	4663
21	Central Kalimantan	52	0	2157
22	South Kalimantan	4758	0	1
23	Kalimantan Timur	10758	0	12145
24	North Sulawesi	7390	272	634
25	Gorontalo	143	996	10
26	Central Sulawesi	22403	91	1180
27	West Sulawesi	1898	1138	370
28	South Sulawesi	15319	8542	10566
29	Southeast Sulawesi	13275	18369	6
30	Maluku	526	2065	1
31	North Maluku	1	111	20
32	Papua	17	0	0
33	West Papua	4	0	28

III. RESULT AND DISCUSSION

In accordance with the algorithm or steps of K-means clustering that the initial stage is determining the number of k or how many clusters will be formed. First is to form 2 clusters while the second is to determine 3 clusters. The goal is to see the difference between how many iterations and the best clusters of shrimp potential production in Indonesia. For clustering with k-means in the production of shrimp species in Indonesia, the initial stage is to normalize.

This normalization applies to the formation of 2 clusters and 3 clusters. The goal is to normalize data to minimize the distance of each data. Normalization for shrimp production data is made with a range [1 0]. The result of normalization is then carried out the clustering process. Normalization results are shown in Table II.

TABLE II. NORMALIZATION DATA

No	Province	Windu Shrimp	Vaname Shrimp	Shrimp others
1	Aceh	0.16739	0.01727	0.04532
2	North Sumatra	0.28669	0.27468	0
3	West Sumatra	6×10^{-5}	4×10^{-5}	0.00139
4	Riau	0.0008	0.00044	0
5	Riau Islands	0	0.00044	0
6	Jambi	0	0	6×10^{-5}
7	South Sumatra	0.16799	0.55538	6×10^{-5}
8	Bangka Belitung	0	0.00985	0
9	Bengkulu	0.00828	0.01312	0.00018
10	Lampung	0.08311	1	0.00782
11	D.I. Yogyakarta	0	0.01123	0.01327
12	Central Java	1	0.19523	1
13	West Java	0.82966	0.80052	0.9857
14	Banten	0.01203	0.01953	0.06622
15	DKI Jakarta	0.00599	0	0
16	East Java	0.29309	0.6544	0.44238
17	Bali	0	0.04069	0.0272
18	NTB	0.12802	0.79055	0.01018
19	NTT	0	0.0018	0
20	West Kalimantan	0.05554	0.54256	0.2825
21	Central Kalimantan	0.00155	0	0.13068
22	South Kalimantan	0.14169	0	6×10^{-5}
23	Kalimantan Timur	0.32037	0	0.73579
24	North Sulawesi	0.22007	0.00378	0.03841
25	Gorontalo	0.00426	0.01382	0.00061
26	Central Sulawesi	0.66715	0.00126	0.07149
27	West Sulawesi	0.05652	0.01579	0.02242
28	South Sulawesi	0.45619	0.11855	0.64013
29	Southeast Sulawesi	0.39532	0.25494	0.00036
30	Maluku	0.01566	0.02866	6×10^{-5}
31	North Maluku	3×10^{-5}	0.00154	0.00121
32	Papua	0.00051	0	0
33	West Papua	0.00012	0	0.0017

A. Testing with 2 clusters

Test results with 2 cluster iterations process stopped at the sixth iteration. This is because in the sixth iteration the position of the centroid is the same as the previous iteration, with the final centroid value being:

C1 = {0.08267166667; 0.029469166667; 0.017526666667}

C2 = {0.37044111111; 0.51716555556; 0.45606222222}.

Whereas for provinces that are divided based on the results of clustering with k-means are.

Cluster 1: (Aceh, North Sumatra, West Sumatra, Riau, Riau Islands, Jambi, Bangka Belitung, Bengkulu, DI Yogyakarta, Banten, DKI Jakarta, Bali, East Nusa

Tenggara, Central Kalimantan, South Kalimantan North Sulawesi, Gorontalo, Central Sulawesi, West Sulawesi, Southeast Sulawesi, Maluku, North Maluku, Papua, West Papua)

Cluster 2: (South Sumatra, Lampung, Central Java, West Java, East Java, West Nusa Tenggara, West Kalimantan, East Kalimantan, South Sulawesi) .

The results of the clustering were then validated with Davies-Bouldin index. To validate the Davies Bouldin index, the initial step is to calculate the Sum of square within a cluster (SSW) using Equation 5.

B. Testing with 3 cluster

1) *The test results with 3 clusters:* the iteration stops at the tenth iteration. As for the final centroid of each cluster are as follows,

- C1 = {0.651555; 0.2779; 0.840405}
- C2 = {0.059186818;0.00807454545;0.0191036363}
- C3 = {0.2013942857; 0.5817871428; 0.106185714}

2) *For the division of clusters of each province as follows:*

- C1 = {Central Java, West Java, East Kalimantan, South Sulawesi}
- C2 = {Aceh, West Sumatra, Riau, Kepulauan Riau, Jambi, Bangka Belitung, Bengkulu, Yogyakarta, Banten, Jakarta, Bali, East Nusa Tenggara, Central Kalimantan, South Kalimantan, North Sulawesi, Gorontalo, Central Sulawesi, West Sulawesi, Maluku, North Maluku, Papua, West Papua}
- C3 = {North Sumatra, South Sumatra, Lampung, East Java, West Nusa Tenggara, Kalimantan West, Southeast Sulawesi}.

3) *Then cluster validation is done as in testing with 2 clusters:* Validation results with Davies-Bouldin index are shown in Table 3. From the validation results the smallest SSW value is in cluster C2, so C2 is selected as the best cluster.

TABLE III. VALIDATION RESULTS WITH DAVIES-BOULDIN INDEX

Clusters	SSW	SSB	Ratio	DBI
1	0.432592	1.047970	0.50307	0.75857
2	0.094617	0.913274	0.80525	
3	0.302821	0.597454	0.66522	

IV. CONCLUSION

From the results of the application of methods and system testing it can be concluded that the k-means method that is applied is effective in clustering shrimp production data in Indonesia. This is because the literacy process does not take a long time either with two clusters or 3 clusters. In the division of 2 clusters the value of Davies Bouldinn index is 0.958. The best cluster is in cluster C1 where Sum of Square within cluster (SSW) is 0.131 smaller than C2 cluster is 0.556. Different in the distribution of 3 clusters, the DBI value is 0.759. For the best cluster there is C2 with a value of 0.095 SSW smaller than C1 and C3. While the comparison between the division of 2 clusters and 3 clusters, the best clusters are found in the division of 3 clusters, this is because the DBI value of the division of 3 clusters is closer to zero compared to the division of 2 clusters.

REFERENCES

[1] Report: Ministry of Maritime Affairs and Fisheries, 2016.

[2] M. Arunachalam and K. Subramanian, “Finger Knuckle Print Authentication Using AES and K-Means Algorithm,” Int. Arab J. of Inf. Technol., vol. 12, no. 6A, pp. 642-649, 2015.

[3] S. Rahayu, D.T. Nugrahadi, and F. Indriani, “Clustering Determination of Potential Crimes Areas in the City of Banjarbaru with the Method K-Means,” Collect. Comp. Sci. J., vol. 1, no. 1, pp. 33-45, 2015.

[4] Rahman and Yuniati, "Analysis of the Cluster of the Marine Fisheries Sector with using Fuzzy K-Means," unpublished.

[5] B. Santosa and A. Umam, Data Mining and Big Data Analytics: Theories and Implementation Using Python and Apache Spark, Yogyakarta: Spreader Media Library, 2018.

[6] P. Harrington, Machine Learning In Action, New York: Manning Publications, 2012.

[7] D.L. Davies and D.W. Bouldin, "Cluster Separation Measure," IEEE Trans. on Pattern, Anal. and Mach. Intell., vol. 1, no. 2, pp. 95-104, 1979.