Application of data mining for Indonesian products export in South Korea using clustering: Indonesia Trade Promotion Center Busan

C R M Pandin¹ and Fahrudin²
¹,²Department of Postgraduate Information System, Universitas Komputer Indonesia
Email: catheriniin93@gmail.com

Abstract. This research purpose is to classify data export products of Indonesia against South Korea. The clustering method used in this research was the K-Means method, in K-Means cluster had a good degree of accuracy. This research examines how the use of the K-means method in case studies export products Indonesia against South Korea. From the results of clustering using K-Means method based on the value of the USD and Kg, then the generated 3 (three) clusters with cluster values are high, medium and low. Based on those results suggest that K-Means method can be used for inventory control of product reference or as Indonesia Indonesia Trade Promotion Center (ITPC) in Busan to view any export products should be retained and what are the export products should be improved to be promoted in South Korea so as to increase the number of export products of Indonesia against South Korea.

1. Introduction
The importance of the value of the export of South Korea against Indonesia became the main focus in this research, as it is known that the value of exports can add foreign exchange of the country. In addition, the value of exports of a country can be one of these values to keep the stability of foreign exchange rates and keep the price of domestic products. Indonesia's relations with South Korea are bilateral relations; the bilateral ties through Indonesia have to ease in conducting import and export toward South Korea.

The purpose of this research was to increase the value of exports of Indonesia against South Korea, Indonesia Trade Promotion Center (ITPC) Busan aims to bridge the trade relations between Indonesia and South Korea. The forms of assistance provided ITPC Busan in bridging the trade relations are realized in the form of services in the form of the granting of information on business opportunities in South Korea to its exporters Indonesia and vice versa the granting information about the potential of Indonesia's exports products to importers in South Korea. Other forms of support are to provide media Indonesia export product promotion in South Korea. The products exported by Indonesia are of two kinds, namely petroleum and natural gas (oil and gas) as well as non-Migas. The data used is product data export of Indonesia against South Korea 2017-2018 period, the source of the data obtained from the ITPC Busan. This research uses data mining [1] with cluster analysis [2]. Cluster analysis is a multivariate technique with the main purpose of grouping objects based on characteristics that are owned. The issue has been used in a size that can explain the closeness between the data for describing the structure of a simple group of complex data, i.e. the size of the range. Measure the distance [3] that is often used is the Euclidean distance [4] measure. In this case, the researchers raised the topic of export products of Indonesia against South Korea based on the
level of exports, where the process method does clustering. The study tried to look at the accuracy of the results of the K-Means [5] clustering method in the case of a grouping of Indonesia exports against South Korea. The K-Means method was chosen because it is considered as a method that can produce accurate grouping the data export products Indonesia against South Korea.

The result of the K-Means cluster method can be used as input for the ITPC Busan as a mapping of product export Indonesia against South Korea. The process of mapping can be in a cluster into three clusters. The export product export products are high, medium, and low export products.

2. Methods

The methods used in this study was the analysis of the cluster. The cluster was the process of grouping similar objects into different groups, or rather partitions from a data set into subsets so that the data in each subset had a useful meaning [5]. Clustering algorithm consists of two parts, namely a hierarchical basis and partition [6]. A hierarchical algorithm [7] for discovering clusters in a sequence where the cluster defined previously, while the algorithm determines all partition [8] group at any given time. The current grouping techniques could be classified into three categories namely partition, hierarchical and locality-based algorithm. There was a set of objects and the criteria of clustering or grouping, grouping partition won partition objects into clusters so that objects in the cluster would be more akin to the objects that were in the cluster of objects contained in a different cluster [9].

2.1 K-Means Method

The algorithms used to process cluster was K-Means. The k-means algorithm was one with partition because of K-Means based on the determination of the amount of the early groups by defining the value of centroid first. K-means clustering is one method of data clustering of non-hierarchical group data in the form of one or more cluster/groups. Data that has the same characteristics are grouped in one cluster/groups and data have different characteristics are grouped with other groups/clusters so that the data is in a single cluster/groups have the level of variation is small [5].

The steps do clustering with the K-means method is as follows: 1) choose the number of clusters; the first step is to choose the number of clusters. For example, we choose to divide it into 3 clusters (of course the determination K free). 2) Specify centroid value (at the rate of random values). Place any data/object to the cluster. The proximity of the two objects in the specified based on the two objects. Distance the distance most near between one record with a particular cluster will determine an incoming data in a cluster. 3) calculate the back Center of the cluster with the cluster member now. The center of the cluster is the average of all data/object in the cluster. 4) Transferred again every object wore the new cluster Center. If the center of the cluster has not changed again cluster, then the process is complete. 5) go back to step 2 Center cluster does not change anymore [10] (see Figure 1).
The process of grouping data into a cluster can be done by calculating the closest distance from a point to a centroid data. Calculate distance for all the data to any point in the center of the cluster can use Euclidean distance theory formulated as the following equation:

$$D(i,j) = \sqrt{(x_{i1} - x_{1j})^2 + (x_{i2} - x_{2j})^2 + \cdots + (x_{ik} - x_{kj})^2}$$

(1)

With:
- \(D(i,j)\) = The distance data \(i\) to the center of the cluster \(j\)
- \(x_{ki}\) = Data \(i\) to attribute data to in \(k\)
- \(X_{kj}\) = The central point to the \(j\) on attributes into \(k\)

The distance of the cluster Center is calculated again with cluster membership now. The center of the cluster is the average of all data/objects in a particular cluster. If you want to be able to also use the median of the cluster. So the average (mean) is not the only measure that can be used. Each object then assigned back to wear the new cluster Center[5] [10].

2.2 Research Phase
The stage of research that will be undertaken are as follows: 1) an export history data collection products that carried South Korea to Indonesia. 2) selection of data, see the data being used and not used. 3) manipulate data using an algorithm for K-Means Clustering. 4) results of the cluster using the K-Means method[5] [10] (see Figure 2).
3. Results and Discussion

Data export of Indonesia's products get from ITPC Busan, from product data collected on each export period features, then the resulting data from the results of the data export of 50 major products Indonesia to South Korea. The data can be cluster using K-means Clustering algorithm. Those results as shown in Figure 3 and Figure 4 below:

![Figure 3](image-url)  
Figure 3. Best 50 Product Export by USD

![Figure 4](image-url)  
Figure 4. Best 50 Product by KG

Figure 3 and 4 are data images of 50 export products based on the number of export values in units of USD and Kg. Data in the form of export data for the period 2017 - 2018 with units of USD and Kg, then the data is averaged based on units of USD and Kg.

3.1 K-Means Based on USD

The data that has been collected will be included in the Rapid Miner tool. Then the selection process is carried out by removing the outlier, determined by 3 outliers which are eliminated from the other data clusters[5]. Furthermore, the data is selected by taking data that is not an outlier that has the potential to
enter into a particular cluster[11]. Because all data is numeric data, the data transformation process is not needed (Figure 5).

Figure 5. The Process of Using Rapid Miner

Figure 5 is the design of the process of clustering on a rapid miner, where data will be normalized before doing the data cluster process first. So that data with numbers that are too large do not dominate the data grouping.

Figure 6. Normalize Data Product by USD

Figure 6 is a graph of the normalization of the data of 50 products based on the USD value, after which the data can actually be clustered into 3 clusters. In the grouping process itself, 3 centroid points are randomly assigned. The centroid results are in the Table 1 below:

Table 1. Centroid Table Based on USD

| Attribute     | C1  | C2  | C3  |
|---------------|-----|-----|-----|
| Average USD   | 4.52| 0.69| -0.27|

Table 1 shows the results of applying the K-Means algorithm resulting in the midpoint or centroid values of the data obtained provided that the desired cluster is 3 clusters.

Place each data in the cluster, in this study used the hard k-means method to allocate each data into a cluster so that data will be entered in a cluster that has the distance closest to the center point of each cluster[10]. To find out which cluster is closest to the data, it is necessary to calculate the distance of each data with the center point of each cluster. For example, the distance from the first data to the first cluster center will be calculated, the data is on average and the results of normalization.

\[
D(1,1) = \sqrt{(0.39 - 4.52)^2} = 4.91
\]

From the results of the above calculations, the results show that the first data distance with the first cluster center is 4.91

\[
D(1,2) = \sqrt{(0.39 - 0.69)^2} = 1.08
\]
From the results of the above calculations, it is found that the first data distance with the second cluster center is 1.08

$$D(1,3) = \sqrt{(0.39 - (-0.27))^2} = 0.12$$

From the results of the above calculations, the results show that the distance of the first data with the third cluster center is 0.12. From the results of the above calculations, the results show that the closest distance of the first data is the third cluster center with a value of 0.12.

In applying the K-Means algorithm a midpoint or centroid value is generated from the data obtained provided that the desired cluster is 3 clusters[5]. Cluster determination is divided into three parts: high export product cluster, medium export product cluster, and low export product cluster. The following are the results of cluster division (see Figure 7):

![Figure 7. Clustering Based on USD](image)

Figure 7 is the result of the K-means clustering based on the USD value. It is known that the results of data clustering show 50 data grouped into 3 clusters. The first cluster is 2 export products, the second cluster is 4 export products, and the third cluster is 44 export products (see Figure 8).

![Figure 8. Clustering 50 Product by USD](image)

In figure 8 above, we can know the USD value of each product. In the first cluster with the number 2 products having a value in USD worth 1,351,358,874 - 1,938,960,495 the second cluster with the number
4 products having a value in USD worth 268,095,937 - 5019,424,020 and finally in the third cluster having 44 products having a value in USD worth 19,459,183 - 201,374,821.

3.2 K-Means Based on KG

Like the previous process, the selected data is removed from the outlier and selects data that is not an outlier. Then the data is normalized first, after that the data can actually be clustered into 3 clusters. The following is data on 50 export products based on normalized Kg (see Figure 9).

| Attribute  | C1   | C2   | C3   |
|------------|------|------|------|
| Average USD| 6.88 | 0.37 | -0.16|

Table 2 shows the results of applying the K-Means algorithm resulting in the midpoint or centroid values of the data obtained provided that the desired cluster is 3 clusters.

Next place each data on the cluster in this study, used the hard k-means method to allocate each data into a cluster so that the data will be entered in a cluster that has the distance closest to the center point of each cluster[10]. To find out which cluster is closest to the data, it is necessary to calculate the distance of each data with the center point of each cluster. For example, the distance from the first data to the first cluster center will be calculated, the data is on average and the results of normalization.

\[ D(1,1) = \sqrt{((0.119) - 6.88)^2} = 6.99 \]

From the results of the above calculations, the results show that the first data distance with the first cluster center is 6.99

\[ D(1,2) = \sqrt{((0.119) - 0.37)^2} = 0.49 \]

From the results of the calculation above, it is obtained that the first data distance with the second cluster center is 0.49

\[ D(1,3) = \sqrt{((0.119) - (-0.16))^2} = 0.04 \]

From the results of the above calculations, the results show that the first data distance with the third cluster center is 0.04. From the results of the above calculations, the results show that the closest distance of the first data is to the third cluster center with a value of 0.04.
In applying the K-Means algorithm a midpoint or centroid value is generated from the data obtained provided that the desired cluster is 3 clusters. Cluster determination is divided into three parts: high export product cluster, medium export product cluster, and low export product cluster. The following are the results of cluster division (see Figure 10):

Figure 10. Clustering Based on Kg

Figure 10 is the result of K-means clustering based on Kg values; it is known that the results of data clustering show 50 data grouped into 3 clusters. The first cluster is 1 export product, the second cluster is 2 export products and the zero clusters are 47 export products (see Figure 11).

Figure 11. Clustering 50 Product by Kg

In Figure 11 above shows clustering based on Kg values. The first cluster consists of 1 product weighing 37,613,021,774 Kg. Then the second cluster of 2 products weighs 2,169,835,000 Kg and 3,750,142,758 Kg respectively. In the zero clusters, there are 47 products that have a weight of 78,168 Kg to 995,603,851 Kg.

Similar research was conducted by [12], Clustering using the K-means method for export value has been done, but the subject of the study is only for fruit exports. In this study, it focuses more on all Indonesian products exported to South Korea, both in migas and non-migas products. In this study, the K-mean method successfully grouped Indonesian's export products into three clusters, namely high export clusters, medium exports, and low exports. From these results, it is hoped that ITPC Busan can be used as a reference for making any product policy that needs to be maintained and what products need to be promoted.
4. Conclusion
The clustering technique with the K-Means algorithm can help group data on Indonesian export products against South Korea. From 50 numbers of data processed determined 3 centroid points randomly in 3 clusters. So that the assessment is based on UDS values and KG units. In high-export products, there are 2 products based on USD values and 1 product based on KG units. Export products are currently having 4 products based on USD values and 2 products based on KG units. And in low export products, there are 44 products based on USD values and 47 products based on KG units. From these results, it can be seen which export products must be maintained and what export products should be increased to be promoted in South Korea so that it can increase the number of Indonesian export products to South Korea.

Acknowledgments
We would like to express our sincere gratitude to Assoc. Prof. Dr. Ir. Eddy Soeryanto Soegoto, MT as Rector of Universitas Komputer Indonesia (UNIKOM), and Dr. Yeffry Handoko Putra, ST., M.T as Head of Master of Information Systems.

Reference
[1] Tan, P. N. 2018. Introduction to data mining. Pearson Education India.
[2] Anderberg, M. R.(201). Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks. Academic press, 19(1), pp. 1-10
[3] Ross, A. J., Samushia, L., Howlett, C., Percival, W. J., Burden, A., & Manera, M. 2015. The clustering of the SDSS DR7 main Galaxy sample–I. A 4 per cent distance measure at z= 0.15. Monthly Notices of the Royal Astronomical Society, 449(1), pp.835-847.
[4] Huang, A. 2008. Similarity measures for text document clustering. In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, 4, pp. 9-56.
[5] Jain, A. K. 2010. Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), pp.651-666.
[6] Antonopoulos, P., Nikolaidis, N., & Pitas, I. 2007. Hierarchical face clustering using sift image features. In 2007 IEEE Symposium on Computational Intelligence in Image and Signal Processing (pp. 325-329). IEEE.
[7] Gil-García, R., & Pons-Porrata, A. 2010. Dynamic hierarchical algorithms for document clustering. Pattern Recognition Letters, 31(6), pp. 469-477.
[8] Nanda, S. J., & Panda, G. 2014. A survey on nature inspired metaheuristic algorithms for partitional clustering. Swarm and Evolutionary computation, 16, pp.1-18.
[9] Zhu, A., Wang, G., & Dong, Y. 2015. Detecting natural scenes text via auto image partition, two-stage grouping and two-layer classification. Pattern Recognition Letters, 67, pp.153-162.
[10] Ghosh, S., & Dubey, S. K. 2013. Comparative analysis of k-means and fuzzy c-means algorithms. International Journal of Advanced Computer Science and Applications, 4(4), pp. 3-10
[11] Manzoni, G. C., Micieli, G., Granella, F., Tassorelli, C., Zanferrari, C., & Cavallini, A. 1991. Cluster headache course over ten years in 189 patients. Cephalalgia, 11(4), pp.169-174.
[12] Windarto, A. P. 2017. Penerapan Datamining Pada Ekspor Buah-Buahan Menurut Negara Tujuan Menggunakan K-Means Clustering Method. Techno. Com, 16(4), pp.348-357.