ENTHUSIASTIC
INTERNATIONAL JOURNAL OF APPLIED STATISTICS AND DATA SCIENCE
Volume 3, Issue 1, April 2023, pp. 111-124

MSME Sales Clustering Based on Business Aid Distribution Priority Using K-Affinity Propagation

Tarisya Qurrota A'yuni [a,1,*], Baiq Nina Febriati [a,2], Lazuardy Ilham Effendie [a,3], Muhammad Muhajir [a,4], Rahmadi Yotenka [a,5]

[a] Department of Statistics, Universitas Islam Indonesia, Jl. Kaliurang, Sleman and 55584, Indonesia
[1] 19611145@students.uii.ac.id*; [2] 19611148@students.uii.ac.id; [3] 19611141@students.uii.ac.id; [4] mmuhajir@uii.ac.id; [5] rahmadi.yotenka@uii.ac.id
* Corresponding author

ARTICLE INFO

Keywords
MSMEs
K-affinity propagation
Commodity

ABSTRACT

In rural areas of Indonesia, micro, small, and medium enterprises (MSMEs) are often isolated; however, they have been proven to play an important role as the economic backbone of millions of communities. In fact, the sluggish development of MSMEs in Indonesia become a severe problem for the community welfare. The government continues to strive for the welfare of the local communities, one of which is by supporting the existing MSMEs. However, the provision of government assistance may not be optimal for the incorrect target of the MSMEs. This study informs the government and other related parties regarding subdistrict groups whose MSMEs are considered to be their target. The k-affinity propagation method was used to find a set of representative examples, called exemplars, that best summarize the data. The result shows that sub-districts clusters based on general welfare in five commodities. K-affinity propagation algorithm clusters vary by commodity. Data fluctuation from each commodity's three factors causes this. From this research, it can be determined which subdistricts have the most or least prosperous MSMEs in each of the five commodities analyzed.

1. Introduction

The Ministry of Cooperatives and Small and Medium Enterprises (SMEs) noted that the number of micro, small, and medium enterprises (MSMEs) in Indonesia reached 65.47 million business units in 2019. This number indicates an increase compared to the previous year, showing an increase of 1.98%. This amount is equivalent to 99.99% of the total business units in Indonesia, of which the remaining 0.01% are large-scale business units. Data on the growth of the number of MSMEs in Indonesia can be seen in Fig. 1 [1].

p-ISSN 2798-253X
e-ISSN 2798-3153

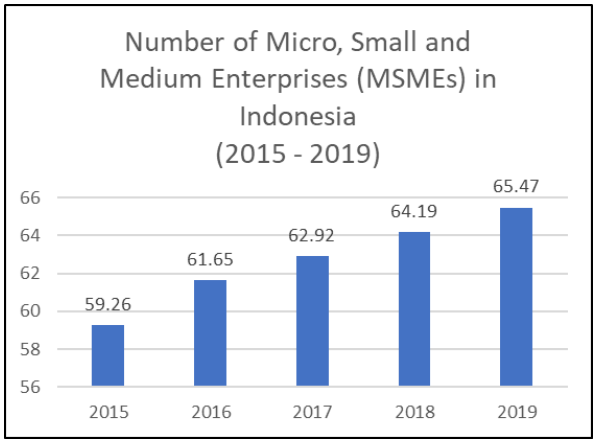

Fig. 1 Bar chart of MSMEs in Indonesia.

Various real roles in the Indonesian economy have made MSMEs recognized as a very important business sector. Based on Statistics Indonesia (Badan Pusat Statistik, BPS) data, the ability of MSMEs to absorb labor was always increased, from around 12 million in 1980 to 74.5 million in 2001. In 2001, the number of MSMEs increased significantly to 40 million from around 7,000 in 1980. Small businesses with a capital of less than 1 billion rupiah are able to absorb 88.59% of the total workforce in one year. Then medium-sized businesses with a capital value of between 1 billion up to 50 billion rupiah are able to absorb 10.83% of the workforce. Meanwhile, large-scale businesses (0.01% of total business units) with a capital of over 54 billion rupiah can only absorb 0.56% of the workforce. With the development of MSMEs, known from the data, namely their ability to provide employment, the growth and role of MSMEs in Indonesia should be continuously raised [2].

At the "Grebeg UMKM" event in 2019, Adipati Aryo (KGPAA) Paku Alam X stated that in the last ten years, MSMEs were the engine of the economy of the Special Region of Yogyakarta. The economic character of the Special Region of Yogyakarta is that it is dominated with micro and small industries, which are 98.4% and labor absorption is 79%. MSMEs in Yogyakarta have advantages due to the high level of vocational education, culture, and creativity of their human resources. The infrastructure not only support human resources, but also the availability of relatively cheap raw materials.

Based on dataumkm.slemankab.go.id, one of the regencies the Special Region of Yogyakarta favoring MSMEs in its economy is Sleman Regency. The total number of MSMEs in Sleman Regency is 90,441 units, with 90,419 micro-enterprises, 19 small-scale businesses, and 3 medium-sized businesses. Then, based on one data recorded at the Sleman Cooperatives and MSME Service, the number of MSMEs in Sleman Regency has increased significantly in the last two years. In 2019, there were 48,000 business units and in December 2020 it increased to 68,000 business units, and reached 90,182 units in 2022 [3]. Then, the Head of the Cooperatives and SMEs Service of Sleman Regency Rr. Mae Rusmi Suryaningsih stated that around 56% of business units are engaged in the food or beverage sector and culinary business. She added that the sectors can survive during the pandemic. The government must take seriously the phenomenon of the emergence of new food products, because food products must of course be safe, quality, and nutritious [4].

MSMEs are business sectors whose existence are very essential because they have a role in the economy such as the share of Gross Domestic Product (GDP) formation, the ability to accommodate workers or has the involvement of a very large number of business units. A number of studies confirm that MSMEs have a large contribution to the GDP, which occurs because of their ability to drive economic depth, strengthen the domestic economy, and strengthen industrialization [5].

In addition, empowering MSMEs is one of the efforts that has been proven can contribute to overcoming the problem of poverty. In 2005, the MSME sector absorbed more than 99.45% of the workforce. The existence of the development of MSMEs will absorb even more labor so that it can

p-ISSN 2798-253X
e-ISSN 2798-3153

reduce the unemployment rate, which makes people who previously had no income at all become productive and have income [6].

In carrying out development, there are many things that become obstacles, both financial and non-financial problems. Financial problems that often occur are the mismatch between the available funds that can be accessed by MSMEs, access to formal sources of funds is still lacking, usually caused by the absence of banks in remote areas or the unavailability of adequate information, and high transaction costs and credit interest. While nonfinancial problems involve limitations of human resources such as low technological capabilities because they do not keep up with developments and low marketing knowledge, which result in limited ability of small business to develop [6].

Business enthusiasm in the national economy will be much better if the government has stronger commitment to providing support to MSMEs [2]. Strong and consistent support will encourage MSMEs so that they can become the foundation of the Indonesian economy, even in times of crisis. In an effort to encourage equitable distribution of community welfare through the MSMEs development program in Sleman Regency, subdistricts in Sleman Regency can be grouped using variables related to the welfare of MSMEs.

Based on the description of the problems above, researchers are interested in knowing the results of grouping subdistricts based on the priority of MSME assistance in Sleman Regency. It will be a novelty in the Department of Industry and Trade of Sleman Regency. The data used in this study were secondary data obtained from the aforementioned department. The amount of data used was 11,067 community-owned business units in Sleman Regency. The initial data were the data on ownership of business units in seventeen subdistricts in Sleman Regency. The data were cross-sectional data that formed from subdistrict variables as the object to be grouped, then there was a business welfare variable as a determining variable for the grouping carried out in this study. The business wealth was interpreted in three variables, namely assets, average sales, and total manpower. That variables were chosen because based on previous studies these three variables were sufficient to describe the entire population and to achieve the research objectives. Then, the clustering method was used since this research aims to group subdistricts based on several variables.

There are many kinds of clustering methods, in this study the method used was k-affinity propagation clustering. On a similar topic, namely about SMEs, a cluster analysis was carried out using the Fuzzy C-Means and K-Means methods. The research conducted by Erni Raouza and Luth Fimawahib, using cross-sectional data with variables in the form of turnover, assets, and total manpower, showed that the Fuzzy C-Means Clustering method had average validation value of close to 1, indicating that the method had a high accuracy rate of 90%. The implementation and testing had been done, it is concluded that the method is able to classify the types of SMEs in accordance with the Law Number 20 of 2008 [7].

Whereas the research of Dicky Jordan A.P., Dwi Remawati, and Tri Irawati in 2021, with almost the same data on MSMEs and same variables, implemented the K-Means method for mapping distribution of MSMEs in Sragen Regency. The K-Means algorithm performed grouping based on the cluster center point (centroid) closest to the data. K-Means maximizes data similarity between clusters. Based on the results of this study, the grouping of business levels was successfully carried out in accordance with Law Number 20 of 2008. This results proves that the simple K-Means method is able to classify business levels so that MSME assistance can be distributed according to the mapping on the subdistrict map [8].

The two studies discussed almost the same topic, regarding the grouping of MSMEs in a district, using different methods. Then, there are also the other studies that compare Fuzzy C-Means and K-Means clustering. The researcher argues that the K-Means method is faster that Fuzzy C-Means because, in terms of clustering time, Fuzzy C-Means requires more time due to the need to update the cluster center and its membership degree based on the smaller number of iterations or objective functions. Updating the center and degree of membership of each cluster takes as much time as the

p-ISSN 2798-253X
e-ISSN 2798-3153

square of the amount of data, weight values, and certain iterations with c being the number of clusters [9].

In previous studies, there are also comparison of the k clustering method, namely between k-means, k-medoids, and k-affinity propagation [10]. The standard deviation was used to determine the ideal number of clusters. The smaller of the standard deviation value, the greater the object similarity. It was found that the k-affinity propagation was the best method in the way of grouping part-time workers who used the internet to carry out their main job. This study determined the standard deviation of k-means and k-medoid. The three clustering methods produced significantly different standard deviation values, with the order from smallest to largest being k-affinity propagation, k-medoids, and k-means. It is concluded that the k-affinity propagation method proved to be the best with the smallest standard deviation [10].

The method used for grouping each subdistrict in Sleman Regency is clustering using k-affinity propagation, this method is a modification of affinity propagation to optimize the k value to obtain optimal copies. This method is used because it has a low error rate, and can find clusters fairly quickly [10]. Cluster method identifies copies among all data points and then forms a cluster of data points around the copies. The way it works is by assuming that all data points have the same potential to become copies, then messages of genuine value are exchanged between data points until a good set of copies and clusters emerges.

Furthermore, clustering was carried out based on the commodities in the MSMEs in Sleman Regency, namely buildings, handicrafts, electronic metal, clothing, and food. This method was selected because we want to know which business wealth variable plays an important factor in each of the existing commodities. The results of this study can be used as material for evaluating government programs, especially the Sleman Regency Industry and Trade Service in determining targets for MSME assistance in the form of financial assistance and training. With the hope that MSMEs can develop better.

2. Method

This study was conducted to determine the number of clusters formed using k-affinity Propagation clustering method to group subdistrict in Sleman Regency based on business aid distribution priority of MSME using the RStudio software. The grouping was done with three variables that were suspected to be a factor in the welfare of an MSME, namely business assets, average sales, and total manpower.

2.1. Clustering

Clustering is a technique for grouping data based on data similarity. Clustering differs from classification, in a way that grouping is done without based on a particular class or group. The absence of reference variables in clustering process makes it can be used to assign labels to data group whose class is not known beforehand [9]. The distance between objects is carrying out the grouping and the shape is only affected by the size of the distances [10]. This process is calculated using the Euclidean distance, as follows.

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

where $D(x,y)$ is an euclidean distance between objects; x_i, y_i are the object coordinate; n is the number of variables [11], [12].

2.2. Non-Multicollinearity Assumption

Multicollinearity is a condition where there is a strong relationship or correlation between objects. It is better that there is no correlation between objects in grouping; if there is a correlation, it is recommended to eliminate several variables that have a large correlation. Multicollinearity assumptions can be tested by plotting correlations or by Variance Inflation Factor (VIF) values. If the VIF value < 10, then there is no multicollinearity. The VIF formula is presented in (2) [9].

p-ISSN 2798-253X
e-ISSN 2798-3153

$$VIF = \frac{1}{1-R^2_{yx1x2}} \quad (2)$$

$$R^2{}_{yx1x2} = \frac{\left(r_{yx1}-r_{yx2}\right)^2}{1-r^2_{x1x2}} \quad (3)$$

$$r_{xy} = \frac{(n\sum x_i y_i - \sum x_i \sum y_i)}{\sqrt{(n\sum x_i^2 - (\sum x_i)\text{^}2\,)(n\sum y_i - (\sum y_i)^2}} \quad (4)$$

where *VIF* is a value that we need to know about data's multicollinearity assumption; R^2_{yx1x2} is coefficient of determination; $R_{yx_1x_2}$ is multiple correlation value between variables y, x1, and x2; r_{xy} is correlation value between variables x and y; *n* is amount of data used [9].

The non-multicollinearity assumption does not apply in the clustering context. This assumption is related to the regression analysis involving the dependent variable and independent variable. The main purpose of clustering is to group objects or data based on their similarities or relatedness, to find structures or patterns in data without considering the dependent or independent variables specifically. Therefore, the assumption of no multicollinearity does not apply in clustering. However, it is still important to consider the linkages and correlations between variables in the broader context of analysis when preparing data prior to clustering.

2.3. K-Affinity Propagation

The k-affinity propagation method is a modified method, which was originally only affinity propagation which is then added k in the research of Zhang, et al [13]. The affinity propagation method was first proposed by Frey and Dueck [14], which is one of the best and most recent partitioning clustering algorithms. By passing two types of messages between data points iteratively, the algorithm selects exemplars of each data point [15]. To get desired number of clusters, it is not easy to set an appropriate preference parameter for AP algorithm. This problem was solved by the k-affinity propagation algorithm very well [16].

The k-affinity propagation method aims to produce the optimal number of copies. This new method identifies copies, which form clusters of data points. K is compared with several indices to determine the optimal one from Jia et al [10], [17]. The number of clusters can be generated according to what a user specifies by adding one constraint in the message passing process to restrict the number of clusters to K [15]. While the AP is a parameter set by its users, another advantage of this method is the belief in an object to serve as an example which is automatically adapted by k-affinity propagation [10], [18].

2.4. Cluster Validity Index

K-affinity propagation is a modified method of affinity propagation, which is the original method, that adopted to obtain the optimal number of exemplars and objects. The number of *k* is one advantage of this approach because it does not need to be entered at the beginning. Besides, when large datasets are used relatively small errors tend to occur [19]. Compared to K-means, both have similarities in the approach to the number of *k*, but k-affinity propagation is more stable. C-index, Davies Bouldin, and McClain Rao are used to obtained the optimal number of cluster [10], [20].

a. C-Index

Impeding a robust automatic determination of the optimal number of clusters, the C-index is hampered by the fact of showing optimal index values for different numbers of clusters. The C-index is defined as follow.

$$C = \frac{S-S_{min}}{S_{max}-S_{min}} \quad (5)$$

where S is the sum of distances over all pairs of objects from the same cluster, S_{min} is the sum of the n smallest distances if all pairs of objects are considered. Likewise, S_{max} is the sum of the *n* largest

p-ISSN 2798-253X
e-ISSN 2798-3153

distances out of all pairs. The C-index is limited to the interval [0, 1] and should be minimized [20], [21].

b. Davies Bouldin

The Davies-Bouldin index cannot achieved by average determination when dealing with binary data, requires the computation of the cluster center. It is defined as follow.

$$DB = \frac{1}{n}\sum_{i=1,i\neq j}^{n} max\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right) \quad (6)$$

where n is the number of clusters, σ_i is the average distance of all patterns in cluster I to their cluster center c_i; σ_j is the average distance of all patterns in cluster j to their cluster center c_j; and $d(c_i, c_j)$ is the distance of the cluster c_i and c_j. The number of clusters that minimizes DB is taken as the optimal number of clusters, because of small values of DB correspond to clusters that are compact and whose centers are far away from each other [20], [22].

c. McClain Rao

McClain and Rao are defined as the average of the individual cluster ratios. The minimum value of the index denotes the best partition. The ratio between the average within-cluster distance and the average between-cluster distance is computed for each cluster [23], [24].

The McClain Rao index is defined as the quotient between the mean within-cluster and between-cluster distances as follow.

$$C = \frac{S_w/N_w}{S_B/N_B} = \frac{N_B}{N_w}.\frac{S_w}{S_B} \quad (7)$$

where the S_w is same with S_{min} at C-index, which is it the sum of the within-cluster distances as follow.

$$S_w = \sum_{(i,j)\in I_w} d(M_i, M_j) = \sum_{(i,j)\in I_w}^{K} \sum_{\substack{i,j\in I_k \\ i<j}} d(M_i, M_j) \quad (8)$$

That the total number of distances between pairs of points belonging to a same cluster is N_w. Besides that, S_B is the sum of the between-cluster distances that defined as follow.

$$S_B = \sum_{(i,j)\in I_w} d(M_i, M_j) = \sum_{k<k'} \sum_{\substack{i,j\in I_{k'} \\ i<j}} d(M_i, M_j) \quad (9)$$

The total number of distances between pairs of points which do not belong to the same cluster is $N_B = \frac{N(N-1)}{2-N_w}$ [25].

2.5. Determine Goodness of the Cluster Method

Standard deviation is a statistical measure that measures how far the data are spread out from the average value. In the clustering method, the standard deviation can provide an overview of the goodness of the clustering method by indicating how well the clusters formed in it are close together and homogeneous. In the field of study, the metric of standard deviation is employed as a means to assess the outcomes of clustering. The authors elucidated that a decrease in standard deviation signifies improved clustering outcomes in relation to the compactness and homogeneity of the data.

If the standard deviation between the data in the clusters is very small, it indicates that the data points in each cluster are very close to each other and have similar characteristics. It indicates that the clustering method used is successful in forming cohesive and compact clusters. Conversely, if the standard deviation between the data in the clusters that are formed is large, it indicates that the data points in each cluster are more spread out and have different characteristics. It indicates that the clustering method may not be effective in forming adjacent and homogeneous clusters [26].

The standard deviation in the group can be calculated as follow.

$$S_w = K^{-1}\sum_{k=1}^{K} S_k \quad (10)$$

p-ISSN 2798-253X
e-ISSN 2798-3153

where K is number cluster formed; S_k is the standard deviation of k cluster that can be calculated as follow.

$$S_b = [(K-1)^{-1} \sum_k^K (\bar{X}_k - \bar{X})^2]^{1/2} \quad (11)$$

where S_b is standard deviation between cluster; $\bar{X}_k$ is mean of k cluster; $\bar{X}_k$ is overall mean of cluster.

The method that has lower ratio of S_w/S_b is the best method [20], [27]. However, it simply means that it is appropriately formed by assuming there are high homogeneity and heterogeneity values among members belonging to the same cluster [10], [28].

3. Methodology

3.1. Data

Data were collected from the Industry and Trade Office of Sleman Regency through a census of seventeen business in Sleman Regency's subdistricts, including Berbah, Cangkringan, Depok, Gamping, Godean, Kalasan, Minggir, Mlati, Moyudan, Ngaglik, Ngemplak, Pakem, Prambanan Districts, Seyegan, Sleman, Paste, and Turi. The raw data collected were the result of data collection of all business owners in Sleman Regency. Then, using the raw data, the researcher collected data according to commodity, and calculated the average for each subdistrict. Therefore, each district had only one row of data. Since there were seventeen subdistricts, there were seventeen rows of data in one dataset that the researcher used. Of the seventeen datasets, there were five datasets that the researchers combined according to their commodities, namely chemical building materials, electronic metals, crafts, clothing, and business food. For each commodity, the analysis variables were subdistricts, number of workers, and average sales.

3.2. Research Stages

The researcher describes it using a flow chart in order to explain the algorithm for using the k-affinity propagation. The flow chart is given in Fig. 2.

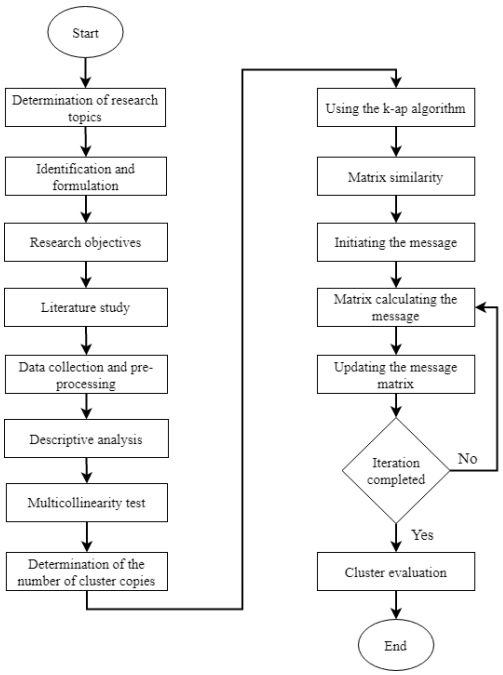

Fig. 2 Flow chart of the research stages.

4. Results and Discussion

4.1. Descriptive Analysis

Before entering the cluster analysis, the researcher made a summary of the quantitative data to determine the size of the data concentration. The size of the concentration of each variable used is as follows.

Table 1. Data Summarize

Summary	Asset	Average Sales	Total Manpower
Minimum	3,255,124	12,408,165	2.000
Mean	20,740,191	83,776,763	2.776
Maximum	61,438,330	299,192,143	6.000

Table 1 shows that the minimum value of all variables is quite far from the average value, except for the total manpower variable, which averages at 2,776 while the minimum value is 2.00. These results means that in this variable it is known that the number of workers mostly owned by MSMEs in Sleman Regency is in the range of 2 until 3 people and there are at most 6 workers. Then the average assets owned by MSMEs in Sleman Regency are 20,740,191 rupiahs, but there are still MSMEs whose assets are only 3,255,124 rupiahs and some whose assets are quite much larger in value, namely 61,438,330. The average sales made by most MSMEs in Sleman Regency amounted to Rp83,776,763.00 with the lowest average sales value of 12,408,165 and the highest reaching two hundreds of millions, namely Rp299,192,143.

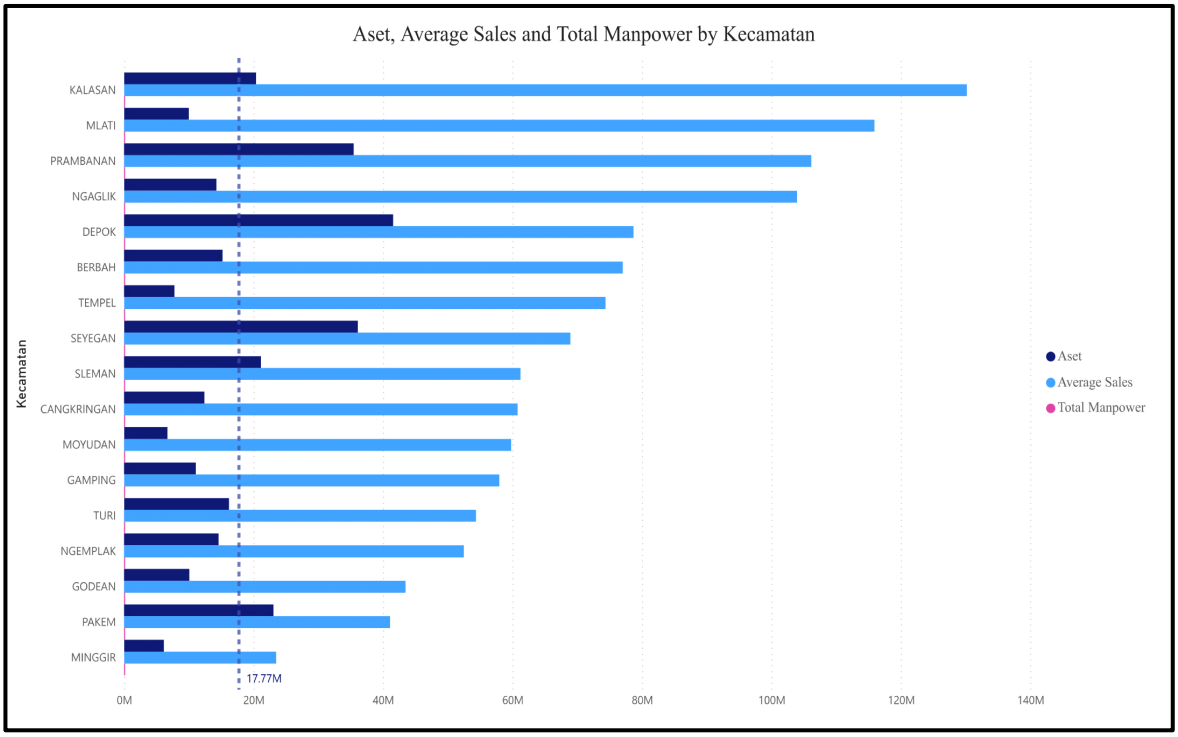

Fig 3. Bar chart of asset, average sales, and total manpower.

Across all seventeen subdistricts, Asset ranged from 6,129,252.32 to 41,562,993.50, Average Sales ranged from 23,485,690.27 to 130,157,075.08, and Total Manpower ranged from 2.33 to 3.19. At 41,562,993.50, Depok had the highest Asset and was 578.11% higher than Minggir, which had the lowest Asset at 6,129,252.32. From this visualization, it can be seen that the lowest and highest values for each variable. However, it is difficult to conclude which districts have the most

p-ISSN 2798-253X
e-ISSN 2798-3153

prosperous SMEs and which are less prosperous. Therefore, a grouping method is needed to find out which districts are very prosperous, and which are less prosperous.

Furthermore, before grouping or clustering, an assumption test is carried out first. The non-multicollinearity assumption test is used to determine the relationship between variables used. The non-multicollinearity assumption test can be seen through the VIF value. If the VIF value < 10, there is no multicollinearity. This is the correlation value (r_{xy}) of the data as follow.

Table 2. Correlation Value of the Variables

	Asset	Average Sales	Total Manpower
Asset	1.00000000	0.1540887	0.09320926
Average Sales	0.15408867	1.0000000	0.56727969
Total Manpower	0.09320926	0.5672797	1.00000000

From the correlation value, it is known that the coefficient of determination and its VIF value are as follow.

Table 3. Multicollinearity Value

	r_{xy}	R^2_{yx1x2}	VIF
Asset and Average Sales	0.15408867	0.0237433	1.024321
Asset and Total Manpower	0.09320926	0.0086880	1.008764
Average Sales and Total Manpower	0.5672797	0.3218062	1.474505

All of VIF values are less than 10 (VIF < 10), do not exist multicollinearity. It means, the variables in the data do not show a multicollinearity or in other words the assumption of no multicollinearity is met. So, the researchers continued the analysis stage to clustering using the k-affinity propagation algorithm.

4.2. K-Affinity Propagation Algorithm

The result of this research show that there are differences in the number of clusters of each MSME data in each commodity. Based on the MSME Commodities in Sleman Regency, they are divided into five namely chemical building materials, crafts, electronic metals, clothing, and food. the following is a description of the results of the cluster based on these commodities.

4.3. Chemical Building Materials

Cluster validity test was conducted to determine the number of clusters to be used, the following are the results of the cluster validity test:

Table 4. Index for Chemical Building Materials

Number of Cluster	C-Index	Davies Bouldin	McClain Rao
2	0.009297708	0.344781	1.152505e-08
3	0.02143737	0.5677957	0.1937311
4	0.013316	0.4879844	0.1781169
5	0.00938117	0.3054457	NaN

Based on Table 4, it is known that the validity test uses 4 indices, namely C-Index, Davies Boulding, and McClain Rao. From the table it can be seen that the number of clusters formed is 2. Hence, to determine which number of clusters is the best, two clusters were used to group subdistricts in Sleman Regency based on priority assistance. The following are the cluster results obtained.

Table 5. Member of Cluster and Profiling of Chemical Building Materials

Cluster	Exemplar	Members	Profiling
1	Minggir	Berbah, Godean, Minggir, Mlati, Ngemplak	All variable values are low compared to other clusters.

p-ISSN 2798-253X
e-ISSN 2798-3153

2	Sleman	Cangkringan, Depok, Gamping, Kalasan, Moyudan, Ngaglik, Pakem, Prambanan, Seyegan, Sleman, Tempel, Turi	The assets and average sales variables show a larger number than the first cluster.

4.4. Craft

To determine the number of clusters to be used, following are the results of the cluster validity test:

Table 6. Index for Craft

Number of Cluster	C-Index	Davies Bouldin	McClain Rao
2	0.006219132	0.4089967	0.2075118
3	**0.00585489**	0.3502248	0.2038678
4	0.2284667	0.4571429	0.200374
5	0.2079457	**0.2940136**	**0.1975789**

Based on the table above, it is known that the validity test uses 4 indices, namely the C-Index, Davies Boulding, McClain Rao, and Silhqoutte. from the table it can be seen that the number of clusters formed is 5. So to determine which number of clusters is the best, 5 clusters will be used to group subdistricts in Sleman Regency based on aid priorities. Here are the cluster results obtained.

Table 7. Member of Cluster and Profiling of Craft

Cluster	Exemplar	Members	Profiling
1	Cangkringan	Cangkringan, Minggir	All variable values are low compared to other clusters.
2	Depok	Depok	Has the same value of total manpower with cluster 3 and 5, this is the wealthiest cluster compared to others. Due to having assets of more than 70 million, and average sales of more than 200 million rupiah, this values are greater than in other clusters.
3	Kalasan	Kalasan	Has the same value of total manpower too, which is with cluster 2 and 5. This cluster has an average asset value that is the second largest compared to other clusters in this commodity. But this cannot be called the second wealthiest cluster because the average sales is quite much lower than the second and fourth cluster.
4	Mlati	Mlati	Average sales are the second largest compared to other clusters, which is around 240 million rupiahs, but the average of assets are still lower than the third and fifth clusters.
5	Prambanan	Berbah, Gamping, Godean, Moyudan, Ngaglik, Ngemplak, Pakem, Prambanan, Seyegan, Sleman, Tempel, Turi	There is a significant difference in numbers compared to other clusters, this cluster still has a greater value than the first cluster.

4.5. Electronic Metals

To determine the number of clusters to be used, the following are the results of the cluster validity test:

Table 8. Index for Electronic Metals

Number of Cluster	C-Index	Davies Bouldin	McClain Rao
2	0.02868687	0.5477596	0.3447231
3	0.04322443	0.623095	0.3103491

p-ISSN 2798-253X
e-ISSN 2798-3153

4	**0.01550926**	**0.505364**	**0.265922**
5	0.3874971	0.5219176	0.2475283

Based on the table above, it is known that the validity test uses 4 indices, namely the C-Index, Davies Boulding, and McClain Rao. From the table it can be seen that the number of clusters formed is 4. So, to determine which number of clusters is the best, 4 clusters will be used to group subdistricts in Sleman Regency based on aid priorities. Here are the cluster results obtained.

Table 9. Member of Cluster and Profiling of Electronic Metals

Cluster	Exemplar	Members	Profiling
1	Berbah	Berbah, Godean, Minggir, Prambanan, Turi	All variables in this cluster have the smallest average value compared to other clusters.
2	Depok	Depok, Kalasan	The average sales of this cluster are the highest compared to the others. However, the average number of MSMEs assets is still quite far, around 28 million lower than the fourth cluster.
3	Moyudan	Cangkringan, Gamping, Mlati, Moyudan, Pakem, Seyegan, Sleman, Tempel	With a slightly larger number of workers compared to the first cluster, MSMEs in the third cluster can have an average sale almost double that of the first cluster.
4	Ngemplak	Ngaglik, Ngemplak	With average sales and a high average number of assets, MSMEs in this cluster are the most prosperous compared to other clusters.

4.6. Clothing

To determine the number of clusters to be used, following are the results of the cluster validity test:

Table 10. Index for Clothing

Number of Cluster	C-Index	Davies Bouldin	McClain Rao
2	**0.0122839**	**0.1605984**	0.2396513
3	0.03229665	0.4640659	0.2757027
4	0.03417213	0.4976575	0.2625766
5	0.01860541	0.4275896	**0.2322461**

Based on the table above, it is known that the validity test uses 4 indices, namely the C-Index, Davies Boulding, and McClain Rao. From the table it can be seen that the number of clusters formed is 2. Hence, to determine which number of clusters is the best, two clusters were used to group subdistricts in Sleman Regency based on priority assistance. Here are the cluster results obtained.

Table 11. Member of Cluster and Profiling of Clothing

Cluster	Examplar	Members	Profiling
1	Kalasan	Kalasan	All variables in this cluster have a higher average value than other clusters. Especially in the variable average sales, this cluster is in the range of 274 million rupiahs, while the second cluster is only at 52 million rupiahs.
2	Sleman	Berbah, Cangkringan, Depok, Gamping, Godean, Minggir, Mlati, Moyudan, Ngaglik, Ngemplak, Pakem, Prambanan, Seyegan, Sleman, Tempel, Turi.	All variables in this cluster have lower average value than other clusters.

4.7. Food

To determine the number of clusters to be used, following are the results of the cluster validity test:

Table 12. Index for Food

p-ISSN 2798-253X
e-ISSN 2798-3153

Number of Cluster	C-Index	Davies Bouldin	McClain Rai
2	0.1746721	0.8076914	0.570021
3	0.1554957	0.8379969	0.5186185
4	0.04996724	0.5852072	0.3533367
5	**0.02238693**	**0.4803103**	**0.2976215**

Based on the table above, it is known that the validity test uses 4 indices, namely the C-Index, Davies Boulding, and McClain Rao. From the table it can be seen that the number of clusters formed is 5. So to determine which number of clusters is the best, five clusters were used to group subdistricts in Sleman Regency based on aid priorities. Here are the cluster results obtained.

Table 13. Member of Cluster and Profiling of Food

Cluster	Exemplar	Members	Profiling
1	Cangkringan	Cangkringan, Minggir, Ngemplak, Pakem	Average sales and total manpower have the lowest values compared to other clusters. While the average number of assets is lower than second and fifth clusters.
2	Depok	Depok	Average sales and total manpower are the lowest compared to other clusters. But the average number of assets is quite high compared to others.
3	Kalasan	Kalasan, Moyudan	With average sales and high total manpower, MSMEs in this cluster are the most prosperous compared to other clusters.
4	Mlati	Berbah, Gamping, Godean, Mlati, Ngaglik	The average asset in this cluster has the lowest value compared to other clusters. But the average sales value is lower than third cluster. Meanwhile, the total manpower value has a higher value than other clusters.
5	Tempel	Prambanan, Seyegan, Sleman, Tempel, Turi	The average asset in this cluster has a lower value than second cluster. The average total sales has a lower value than third cluster while the total manpower has a higher value.

5. Conclusion

From the results of the analysis, it is known that subdistrict clusters are based on their general welfare in five different commodities. In each commodity, there are differences in the number of clusters obtained from the k-affinity propagation algorithm. This condition is due to differences in data variations from the three variables in each commodity.

From this research, it cannot be known which commodity is the most prosperous, but it can be known which majority of MSMEs in the subdistrict are the most prosperous or the least prosperous in each of the five commodities studied. By knowing the subdistrict groups with their MSMEs welfare levels in each commodity, the government or related parties can directly know which subdistricts need more assistance for the welfare of their MSMEs than other subdistricts and can adjust the form of business assistance according to the MSME commodities.

In addition, the researcher suggests that for further research, only one index can be used. The difference in the number of clusters in each commodity could be due to the selected index. In the future, use k optimization without using index.

Acknowledgment

The authors are grateful to all related parties, especially to Department of Industry and Trade of Sleman Regency for allowing the researchers to use the data to be processed in this study. We know that there are still many our lack of understanding. However, we very thanks to the lecturers and friend who have supported the work of this research.

References

[1] M.I. Mahdi, "Berapa Jumlah UMKM di Indonesia?," *dataindonesia.id*, 2022. https://dataindonesia.id/sektor-riil/detail/berapa-jumlah-umkm-di-indonesia

[2] I.Y. Niode, "Sektor UMKM di Indonesia: profil, Masalah dan Strategi Pemberdayaan," *Jurnal Kajian Ekonomi dan Bisnis OIKOS-NOMOS*, vol. 2, no. 1, pp. 1–10, 2019, [Online]. Available: https://repository.ung.ac.id/kategori/show/uncategorized/9446/jurnal-sektor-umkm-di-indonesia-profil-masalah-dan-strategi-pemberdayaan.html

[3] A. Syarifudin, "Jumlah UMKM di Sleman Meningkat Hingga 90 Ribu Selama Pandemi Covid-19," *jogja.tribunnews.com*, 2022. https://jogja.tribunnews.com/2022/01/15/jumlah-umkm-di-sleman-meningkat-hingga-90-ribu-selama-pandemi-covid-19

[4] M. Kriesdinar, "Jumlah UMKM di Sleman Meningkat Signifikan di Masa Pandemi," *jogja.tribunnews.com*, 2021. https://jogja.tribunnews.com/2021/06/06/jumlah-umkm-di-sleman-meningkat-signifikan-di-masa-pandemi

[5] T. Nenova, C.T. Niang, and A. Ahmad, *Bringing Finance to Pakistan's Poor: Access to Finance for SME and the Unserved*. Washington D.C., USA: The World Bank, 2009.

[6] Supriyanto, "Pemberdayaan Usaha Mikro, Kecil dan Menengah (UMKM) di Kota Malang Berbasis Webgis.5," *Jurnal Ekonomi dan Pendidikan*, vol. 3 No.1, pp. 1–16, 2012, doi: 10.21831/jep.v3i1.627.

[7] E. Rouza and L. Fimawahib, "Implementasi Fuzzy C-Means Clustering dalam Pengelompokan UKM Di Kabupaten Rokan Hulu," *Techno.Com: Jurnal Teknologi Informasi*, vol. 19, no. 4, pp. 481–495, 2020, doi: 10.33633/tc.v19i4.4101.

[8] D. Remawati, D.J.A. Putra, and T. Irawati, "Metode K-Means untuk Pemetaan Persebaran Usaha Mikro Kecil dan Menengah," *Jurnal TIKomSiN*, vol. 9, no. 2, p. 39, 2021, doi: 10.30646/tikomsin.v9i2.574.

[9] Y.K. Siregar, "Analisis Perbandingan Algoritma Fuzzy C-Means dan K-Means," *Annual Research Seminar 2016,* vol. 2, no. 1, pp. 151–155, 2016.

[10] N.I. Asriny, M. Muhajir, and D. Andrian, "K-Affinity Propagation Clustering Algorithm for the Classification of Part-Time Workers Using the Internet," *Indonesian Journal Electrical Engineering and Computer Science*, vol. 24, no. 1, pp. 464–472, 2021, doi: 10.11591/ijeecs.v24.i1.pp464-472.

[11] K. Maheswari, "Finding Best Possible Number of Clusters using K-Means Algorithm," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 1S4, pp. 533–538, 2019, doi: 10.35940/ijeat.a1119.1291s419.

[12] M. Charrad, N. Ghazzali, and A.N. Boiteau, "NbClust: an R package for Determining the Relevant Number of Clusters in a Data Set," *Journal of Statistical Software*, vol. 61, no. 6, pp. 1–36, 2014, doi: 10.18637/jss.v061.i06.

[13] X. Zhang, W. Wang, K. Nørvåg, and M. Sebag, "K-AP: Generating Specified K Clusters by Efficient Affinity Propagation," *2010 IEEE International Conference on Data Mining*, 2010, pp. 1187–1192, doi: 10.1109/ICDM.2010.107.

[14] B.J. Frey and D. Dueck, "Clustering by Passing Messages between Data Points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007, doi: 10.1126/science.1136800.

[15] A.M. Serdah and W.M. Ashour, "Clustering Large-Scale Data Based on Modified Affinity Propagation Algorithm," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 6, no. 1, pp. 23–33, 2016, doi: 10.1515/jaiscr-2016-0003.

[16] N.M. Arzeno and H. Vikalo, "Semi-Supervised Affinity Propagation with Soft Instance-Level Constraints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 1041–1052, 2015, doi: 10.1109/TPAMI.2014.2359454.

[17] H. Jia, L. Wang, H. Song, Q. Mao, and S. Ding, "A K-AP Clustering Algorithm Based on Manifold Similarity Measure," in *Intelligent Information Processing IX*, Z. Shi, E. Mercier-Laurent, and J.Z. Li Eds. New York, USA: Springer Cham, 2018, doi: 10.1007/978-3-030-00828-4_3.

[18] A.F. Moiane and Á.M.L. Machado, "Evaluation of the Clustering Performance of Affinity Propagation Algorithm Considering the Influence of Preference Parameter and Damping Factor," *Boletim de Ciências Geodésicas*, vol. 24, no. 4, pp. 426–441, 2018, doi: 10.1590/S1982-21702018000400027.

[19] X. Zhang, C. Furtlehner, C. Germain-Renaud, and M. Sebag, "Data Stream Clustering with Affinity Propagation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 7, pp. 1644–1656, 2014, doi: 10.1109/TKDE.2013.146.

[20] M. Muhajir and N.N. Sari, "K-Affinity Propagation (K-AP) and K-Means Clustering for Classification of Earthquakes in Indonesia," *2018 International Symposium on Advanced Intelligent Informatics (SAIN)*, 2019, pp. 6–10, doi: 10.1109/SAIN.2018.8673344.

[21] L. Hubert and J. Schultz, "Quadratic Assignment as a General Data-Analysis Strategy," *British Journal of Mathematical and Statistical Psychology*, vol. 29, pp. 190–241, 1976, doi: 10.1111/j.2044-8317.1976.tb00714.x.

[22] D.L. Davies and D.W. Bouldin, "A Cluster Separation Measure," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 1, no. 4, pp. 224–227, 2000, doi: 10.1109/TPAMI.1979.4766909.

[23] I. Gurrutxaga, J. Muguerza, O. Arbelaitz, J. M. Pérez, and J.I. Martín, "Towards a Standard Methodology to Evaluate Internal Cluster Validity Indices," *Pattern Recognition Letters*, vol. 32, no. 3, pp. 505–515, 2011, doi: 10.1016/j.patrec.2010.11.006.

[24] J.O. McClain and V.R. Rao, "CLUSTISZ: A Program to Test For The Quality of Clustering of A Set of Objects," *Journal of Marketing Research*, vol. 12, pp. 456–460, 1975.

[25] B. Desgraupes, "Clustering Indices ClusterCrit," *CRAN Packag.*, no. April, pp. 1–10, 2013, [Online]. Available: cran.r-project.org/web/packages/clusterCrit

[26] A. Banerjee, I.S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the Unit Hypersphere Using Von Mises-Fisher Distributions," *Journal of Machine Learning Research*, vol. 6, pp. 1345–1382, 2005.

[27] M.J. Bunkers , J.R. Miller Jr., and A.T. DeGaetano, "Definition of Climate Regions in the Northern Plains Using an Objective Cluster Modification Technique," *Journal of Climate*, vol. 9, no. 1, pp. 130–146, 1996, doi: 10.1175/1520-0442(1996)009<0130:DOCRIT>2.0.CO;2.

[28] A.R. Barakbah and K. Arai, "Identifying Moving Variance to Make Automatic Clustering for Normal Data Set," in *Proceedings of the IECI Japan Workshop*, 2004, pp. 26–30.

p-ISSN 2798-253X
e-ISSN 2798-3153