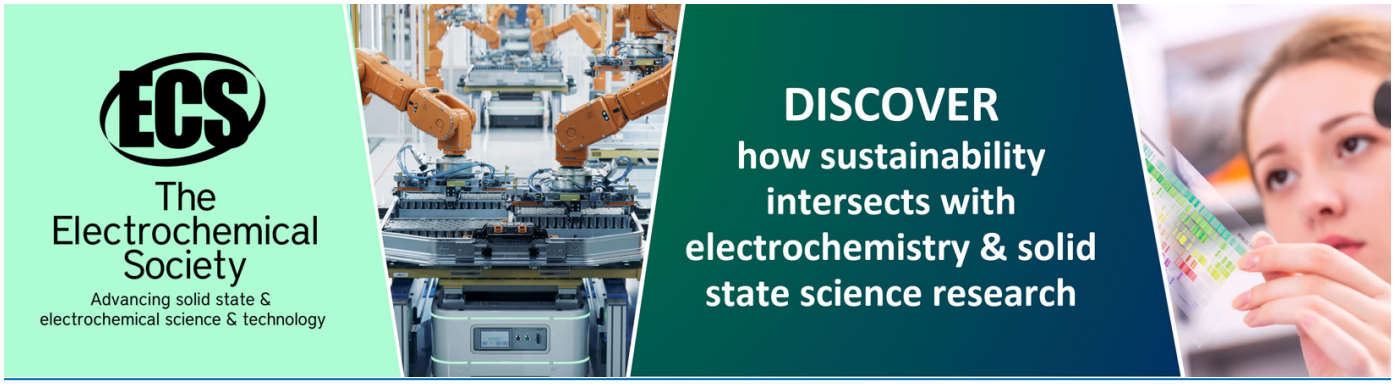
# FCM using squared euclidean distance for e-commerce classification in Indonesia





Journal of Physics: Conference Series **1613** (2020) 012071 doi:10.1088/1742-6596/1613/1/012071

# FCM using squared euclidean distance for e-commerce classification in Indonesia

**E Z Khulaidah and N Irsalinda**
Department of Mathematics, Universitas Ahmad Dahlan, Ring Road Selatan Street, Bantul, Yogyakarta, Indonesia

E-mail: ema110397@gmail.com

**Abstract**. Clustering is a method of grouping data into several clusters so that the data in one cluster has a high level of similarity while the data between other clusters have a low level of similarity. One method used in clustering is Fuzzy C-Means (FCM) which is a data clustering technique in which the existence of each data point in a cluster is determined by the degree of membership in each cluster. The FCM algorithm has an objective function that requires distance. The distance used in this study is Squared Euclidean distance. The clustering conducted is the clustering of the popularity of e-commerce in Indonesia in 2019 using the variable average number of monthly visitors, number of website visitors, number of social media followers (Twitter, Instagram, and Facebook) as well as the number of workers. The result of this method is the level of popularity of e-commerce in Indonesia, which is divided into gold, silver, and bronze. Clustering results were tested with the Partition Entropy Index (PEI) and Classification Entropy (CE) if the results are getting closer to 0, the results are getting better. The result of PEI is 2.9697e-0, and CE is 2.5710e-04. So, based on the two indexes It can be concluded that FCM using Squared Euclidean distance is good to clustering.

## 1. Introduction

Fuzzy logic was introduced by Prof. Lotfi A. Zadeh in 1965. The basis of fuzzy logic is the fuzzy set theory. In the fuzzy set theory, the role of the degree of membership as a determinant of the existence of elements in a set is very important and becomes the main characteristic of fuzzy logic reasoning [1].

According to Cox in 1994 there were several reasons for using fuzzy logic, among others: the concept of fuzzy logic was easy to understand, very flexible, had tolerance of inaccurate data, was able to model nonlinear functions that were very complex, could build and apply the experiences of the experts directly without having to go through a training process, can cooperate with conventional control techniques, and are based on natural or everyday language.

Fuzzy Clustering is one of the techniques to determine the optimal cluster in a vector space. One data clustering algorithm is Fuzzy C-Means (FCM) which is a data clustering technique where the existence of each data point in a cluster is determined by the degree of membership for each cluster. This technique was first introduced by Jim Bedzek in 1981 [1].

The FCM algorithm, which requires a distance function, is generally used to calculate the similarity or similarity between two objects. There are several distances that can be used in FCM, including Euclidean distance, Euclidean Square, Manhattan, Canberra, and others.

The results of the FCM algorithm in the form of a row of cluster centers and some degree of membership which can then be obtained by groups from these clusters. In real life, many cases can be solved using FCM, one of which is to classify the popularity of e-commerce among Indonesian consumers.

E-commerce (electronic commerce) is the distribution, purchase, sale, marketing of goods and services through electronic systems such as the internet, television or other computer networks. E-commerce business in Indonesia has been going on for a long time. In e-commerce there are merchants, merchants in Indonesia including Tokopedia, Bukalapak, Shopee, Lazada, Zalora, Blibli, and others. The number of these merchants makes the competition tighter among Indonesian consumers. This increasingly fierce competition makes merchants develop themselves to attract more consumers. In developing themselves, the merchant must know that how much the merchant is interested in consumers. To find out how much consumers are interested in the merchant, it is necessary to group or cluster the merchants according to the level of popularity among consumers. Clustering the popularity of merchants in this study is divided into three classes, namely gold, silver and bronze.

Based on previous research from Arwan Ahmad Khoiruddin (2007), the cluster results from the FCM method are more natural because they are based on the tendency of each data in the clusters [2]. Furthermore, previous research by Wulan Anggraeni (2015) states that in the FCM algorithm the determination of a rank number of 2 will produce a higher level of accuracy compared to using other numbers, so it is recommended in the FCM algorithm to use a rank of 2 [3]. According to research from Fajar Agustini (2017) The FCM algorithm can be used in grouping applications and the results of clustering can help companies to increase effectiveness by increasing potential products and minimizing potential products [4]. Because FCM can work well, this study will also use the FCM method to classify the popularity of e-commerce among Indonesian consumers by using Squared Euclidean distance. The results of this study are expected to help merchants know their popularity among consumers.

## 2. Research Method

Several FCM methods for data classification are available in the literature. This section deals with the FCM method that used in this research.

### *2.1. Fuzzyfication*

The growth S-curve membership function is as follows:

$$\mu_{X_j}(X) = \begin{cases} 0, & X \le a \\ 2\left(\frac{X-a}{f-a}\right)^2, & a < X < \frac{a+f}{2} \\ 1-2\left(\frac{X-a}{f-a}\right)^2, & \frac{a+f}{2} \le X < f \\ 1, & X \ge f \end{cases} \quad (1)$$

where $X$ is the data, $a$ and $f$ are the interval of $X$ [5].

*2.2. Fuzzy C-Means Clustering*

The algorithm in Fuzzy C-Means (FCM) are follows:

1) Input data to be clustered $X$, in the form of $n \times m$ matrix ($n$ is number of sample data, $m$ is attribute of each data). $X_{ij}$= $i$-th sample data ($i = 1,2,\ldots,n$) , $j$-attribute ($j = 1,2,\ldots,m$) .
2) Determine the number of clusters ($C$), rank ($w$), maximum iteration ($MaxIter$), smallest expected error ($\xi$), initial objective function ($P_0 = 0$), and initial iteration ($t = 1$).
3) Generating random numbers $v$ . Count the numer of each column :

$$Q_k = \sum_{k=1}^{c} u_{ik}$$

4) Calculates $\mu_{ik}$ as matrix elements of the initial partition $U$:

$$\mu_{ik} = \frac{u_{ik}}{Q_k}$$

5) Calculate the centroid of the $k$-cluster: $V_{kj}$, where $k = 1,2, \ldots , c$ ; and $j = 1,2, \ldots , m$

$$V_{kj} = \frac{\sum_{i=1}^{n}\left((\mu_{ik})^w X_{ij}\right)}{\sum_{i=1}^{n}(\mu_{ik})^w}$$

6) Calculating an objective function on the $t$-iteration, $P_t$:

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum\nolimits_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right](\mu_{ik})^w\right)$$

7) Calculate the change in the partition matrix:

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}}$$

where $i = 1,2, \ldots, n$; and $k = 1,2, \ldots, c$.

8) Checking the stooping criteria, if ( $| P_t - P_{t-1}| < \xi$ ) or ( $t > MaxIter$ ) . If not, $t = t + 1$ and then repeat from step 4 [6].

*2.3. Squared Euclidean Distance*

The Squared Euclidean distance metric uses the same equation as the Euclidean distance metric, but does not take the squared root [7].

$$d = \sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2$$

*2.4. Index Validity*

To show the accuracy of the method, we need to compares the results of the Euclidean and Squared Euclidean using the validity index.

1. Partition Entropy Index (PEI)
   PEI value evaluates the randomness of the data in the cluster. The range values are on [0,1], the smallest value (close to 0) means that the cluster obtained is getting better. The following formula for calculating PEI [8]:

$$PEI = -\frac{1}{n}\left(\sum_{i=1}^{n}\sum_{k=1}^{c}\mu_{ik}\ {}^{2}\log\mu_{ik}\right)$$

2. Classification Entropy (CE)
   CE only measures the fuzziness of group partitions. The index equation can be written as follows:

$$CE = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c}\mu_{ik}\log\mu_{ik}$$

   where $n$ is a lot of research objects, $c$ is a lot of groups, and $\mu_{ik}$ is the membership value of the $i$-th object with the center of the $k$-th group. This index has a range from 0 to $\ln(c)$. The smaller CE index indicates better grouping. [11]

## 3. Study and result

### *3.1. Data Description*

The data that used are e-commerce ranking data in Indonesia in the first four months of 2019 obtained from the I price website is show in Table 1. The data to be processed is as follows [1]:

a. Merchant is a store that is at the top of e-commerce in Indonesia. taken that has complete data only.
b. Monthly Website Visitors is an average monthly web visits.
c. The numbers of the social media followers the source is from Facebook, Twitter, and Instagram. The number of Facebook followers is taken from country-specific pages except for regional players where the number of country-specific followers is not publicly available.
d. Number of Workers

**Table 1.** E-Commerce Data

| Merchant | Web Visitors | Twitter | Instagram | Facebook | Number of workers |
|---|---|---|---|---|---|
| Tokopedia | 137200900 | 192100 | 1148500 | 6049900 | 2677 |
| Bukalapak | 115256600 | 161500 | 711700 | 2423200 | 2575 |
| Shopee | 74995300 | 69300 | 2164100 | 14409600 | 2748 |
| Lazada | 52044500 | 365300 | 1173200 | 28245000 | 2212 |
| Blibli | 32597200 | 483300 | 627400 | 8244800 | 1217 |
| JD ID | 10656900 | 22800 | 406300 | 778300 | 1021 |
| Orami | 8380600 | 6200 | 72900 | 357600 | 106 |
| Sociolla | 4838300 | 700 | 462800 | 4900 | 157 |
| Zalora | 4343000 | 66800 | 337500 | 7719000 | 466 |
| Bhinneka | 3446500 | 71400 | 40200 | 1029400 | 547 |
| Elevenia | 3394400 | 120800 | 118800 | 1181100 | 297 |
| Jakarta Notebook | 2450100 | 10900 | 24100 | 45600 | 44 |
| Shopie Paris | 1967700 | 34800 | 611900 | 1966900 | 612 |
| Alfacart | 1364700 | 6300 | 38100 | 911200 | 131 |
| Jakmall | 977700 | 3500 | 40300 | 103500 | 50 |
| Sorabell | 922100 | 14300 | 300 | 4349700 | 552 |
| Fabelio | 877600 | 500 | 108000 | 86000 | 156 |
| Matahari | 769100 | 101000 | 385500 | 1645900 | 656 |
| Otten Coffe | 506500 | 8100 | 337500 | 640000 | 37 |

| | | | | | |
|---|---|---|---|---|---|
| Asmaraku | 388100 | 600 | 14800 | 8700 | 25 |
| Mothercare | 366300 | 28300 | 459300 | 153400 | 410 |
| Oriori | 340200 | 2500 | 43300 | 237200 | 33 |
| Pemmz | 283400 | 1400 | 22000 | 30500 | 14 |
| Hijup | 282300 | 57600 | 890300 | 317800 | 157 |
| Berrybenka | 272100 | 16100 | 287400 | 962400 | 206 |
| Hijabenka | 253800 | 2600 | 393300 | 776200 | 206 |
| Bobobobo | 217400 | 3700 | 104100 | 230700 | 73 |
| Bukupedia | 208200 | 129200 | 11200 | 17000 | 3 |
| VIP Plaza | 204900 | 2700 | 24300 | 100300 | 89 |
| Bro.do | 185800 | 19800 | 441800 | 1242600 | 70 |
| Sephora | 173000 | 3700 | 321900 | 18460100 | 59 |
| Electronic City | 169400 | 46500 | 21300 | 214300 | 478 |
| Dinomarket | 162100 | 36200 | 38900 | 43500 | 27 |
| Tees | 132000 | 9800 | 5600 | 57100 | 18 |
| Maskoolin | 121900 | 6700 | 20000 | 109400 | 8 |
| Muslimarket | 85700 | 800 | 24700 | 221500 | 9 |
| 8Wood | 13900 | 3300 | 501200 | 14900 | 14 |
| Electronic Solution | 2000 | 20200 | 7900 | 184500 | 223 |
| Mamaway | 800 | 200 | 200 | 287000 | 2 |

*3.2. Data Clustering Process*

The process of the clustering are follows:

a. Input the data to be clustered $X$, in the form of $n \times m$ ($n$ is the number of data samples, in this case the number of merchants =39, $m$= data variable is 5). $X_{ij}$= -$i$-th sample data ($i = 1, 2, \ldots, 39$) , $j$-variable ($j = 1,2, \ldots, 5$) .

b. Determine number of clusters $(C) = 3$, rank $(w) = 2$, maximum iteration $(MaxIter) = 100$, smallest expected error $(\xi) = 10^{-5}$, initial objective function $(P_0 = 0)$, and iteration initial $(t = 1)$.

c. Generating random numbers $u_{ik}$, $i = 1,2, \ldots ,39$ ; $k = 1,2,3$. The initial partition matrix U is obtained from a random matrix measuring 39 × 5 which we then normalize so that the sum of each row of the U matrix equals one.

d. Calculates $\mu_{ik}$ as elements of the initial partition matrix $U$:

$$\mu_{ik} = \frac{u_{ik}}{Q_k}$$

e. Calculates the center of the $k$-cluster: $V_{kj}$, with $k = 1,2,3$ ; and $j = 1,2, \ldots ,5$

$$V_{kj} = \frac{\sum_{i=1}^{39}\left((\mu_{ik})^{39} X_{ij}\right)}{\sum_{i=1}^{39}(\mu_{ik})^2}$$

f. Calculating an objective function on the $t$-iteration, $P_t$:

$$P_t = \sum_{i=1}^{39}\sum_{k=1}^{3}\left(\left[\sum_{j=1}^{5}\left(X_{ij} - V_{kj}\right)^2\right](\mu_{ik})^2\right)$$

g. Calculates the partition matrix:

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{5}\left(X_{ij} - V_{kj}\right)^2\right]^{-1}}{\sum_{k=1}^{3}\left[\sum_{j=1}^{5}\left(X_{ij} - V_{kj}\right)^2\right]^{-1}}$$

with $i = 1,2, \ldots ,39$; and $k = 1,2,3$.

h. Checking the stopping criteria, if ( $|P_t - P_{t-1}| < \xi$ ) or ( $t > MaxIter$ ) . If not, $t = t + 1$ and repeat from d.

In this case, the new process will stop after the 15th iteration. In this 15th iteration, 3 cluster centers are $V_{kj}$ with $k = 1,2,3$; and $j = 1,2,3,4,5$ as follows:

$$V = \begin{pmatrix} 0{,}0015 & 0{,}0184 & 0{,}0363 & 0{,}0618 & 0{,}0274 \\ 0{,}8860 & 0{,}2508 & 0{,}4825 & 0{,}1368 & 0{,}9712 \\ 0{,}2669 & 0{,}8026 & 0{,}5120 & 0{,}7985 & 0{,}7927 \end{pmatrix}.$$

The results of the e-commerce clustering can be shown from the $\mu_{ik}$ in Table 2 and Figure 1.

**Table 2.** Result of E-Commerce Clustering using FCM

| E-commerce ($i$) | $\mu_{i1}$ | $\mu_{i2}$ | $\mu_{i3}$ | c1 | c2 | c3 | E-commerce ($i$) | $\mu_{i1}$ | $\mu_{i2}$ | $\mu_{i3}$ | c1 | c2 | c3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.0109 | 0.9700 | 0.0191 | | * | | 21 | 0.9964 | 0.0019 | 0.0017 | * | | |
| 2 | 0.0432 | 0.9035 | 0.0533 | | * | | 22 | 0.9976 | 0.0013 | 0.0012 | * | | |
| 3 | 0.1268 | 0.5724 | 0.3008 | | * | | 23 | 0.9975 | 0.0013 | 0.0012 | * | | |
| 4 | 0.0223 | 0.0427 | 0.9350 | | | * | 24 | 0.9107 | 0.0465 | 0.0429 | * | | |
| 5 | 0.3048 | 0.2196 | 0.4756 | | | * | 25 | 0.9993 | 0.0004 | 0.0003 | * | | |
| 6 | 0.9282 | 0.0392 | 0.0326 | * | | | 26 | 0.9983 | 0.0009 | 0.0008 | * | | |
| 7 | 0.9977 | 0.0012 | 0.0011 | * | | | 27 | 0.9979 | 0.0011 | 0.0010 | * | | |
| 8 | 0.9960 | 0.0021 | 0.0019 | * | | | 28 | 0.9826 | 0.0088 | 0.0086 | * | | |
| 9 | 0.9798 | 0.0100 | 0.0101 | * | | | 29 | 0.9976 | 0.0013 | 0.0011 | * | | |
| 10 | 0.9952 | 0.0025 | 0.0023 | * | | | 30 | 0.9968 | 0.0017 | 0.0015 | * | | |
| 11 | 0.9876 | 0.0063 | 0.0061 | * | | | 31 | 0.6299 | 0.1471 | 0.2231 | * | | |
| 12 | 0.9976 | 0.0013 | 0.0012 | * | | | 32 | 0.9974 | 0.0014 | 0.0012 | * | | |
| 13 | 0.9780 | 0.0116 | 0.0104 | * | | | 33 | 0.9978 | 0.0012 | 0.0011 | * | | |
| 14 | 0.9978 | 0.0012 | 0.0010 | * | | | 34 | 0.9975 | 0.0013 | 0.0012 | * | | |
| 15 | 0.9976 | 0.0013 | 0.0012 | * | | | 35 | 0.9975 | 0.0013 | 0.0012 | * | | |
| 16 | 0.9946 | 0.0028 | 0.0026 | * | | | 36 | 0.9975 | 0.0013 | 0.0012 | * | | |
| 17 | 0.9981 | 0.0010 | 0.0009 | * | | | 37 | 0.9938 | 0.0033 | 0.0030 | * | | |
| 18 | 0.9858 | 0.0073 | 0.0068 | * | | | 38 | 0.9981 | 0.0010 | 0.0009 | * | | |
| 19 | 0.9986 | 0.0007 | 0.0007 | * | | | 39 | 0.9975 | 0.0013 | 0.0012 | * | | |
| 20 | 0.9975 | 0.0013 | 0.0012 | * | | | | | | | | | |

Cluster 1, 2 and 3 determination can be seen from the value of $V$ that we have obtained before. $V_{1j}$ as the cluster 1 (bronze popularity), $V_{2j}$ as the cluster 2 (golden popularity) and $V_{3j}$ as cluster 3 (silver popularity). Determination of golden, silver and bronze popularity are obtained from the average values of $V_{1j}, V_{2j}$ and $V_{3j}$.

From the Table 2, to decide a merchant include in cluster 1, 2 or 3 is show from the maximum value of $\mu_{ik}$ . Merchant 1 (Tokopedia) include in cluster 2 because it has $\mu_{ik}$ maximum value in $\mu_{i2}$. As well as other merchants until 39. So we can get all of clustering that show in Figure 1.

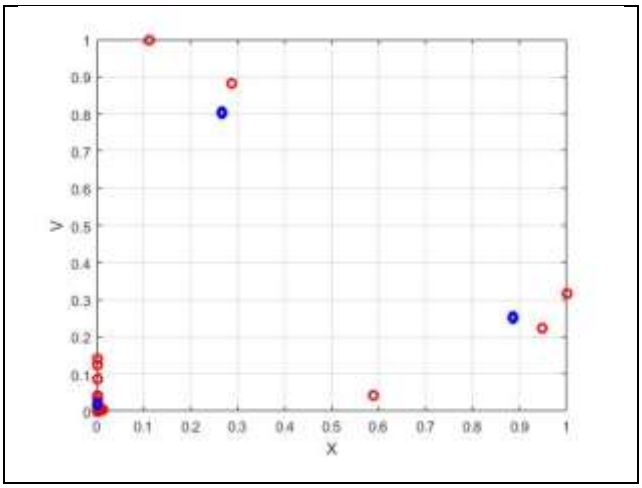



**Figure 1.** Clustering FCM with Squared Euclidean distance

In Figure 1 the blue points show the distribution of the cluster centers. namely gold. silver. and bronze. while the red circle shows the distribution of merchants. So it can be concluded that merchants that gained gold popularity were Lazada and Blibli. Merchants who gained popularity in Silver were Tokopedia, Bukalapak, and Shopee. Merchants that gained Bronze popularity were JD ID, Orami, Socioll, Zalora, Bhinneka, Elevenia, Blanja, Laku6, Jakarta Notebook, Ralal, Shopie Paris, iLotte, AliExpress, Alfacart, Jakmall, Sorabell, Fabelio, Matahari, PlazaKamera, Otten Coffe, Otto Coffe, Weshop, Asmaraku, Mothercare, Oriori, Qoo10, Pasarwarga, Mapemall, Pemmz, Hijup, Berrybenka, Hijabenka, Bobobobo, Bukupedia, VIP Plaza, Bro, Do,Sephora, Electronic City, Dinomarket, Tees, Maskoolin, Muslimarket, 8Wood, Electronic Solution, and Mamaway.

### *3.3. Validity Index*

To show the accuracy of the clustering results, we need to calculates the partition entropy and classification entropy as follows:

1. Partition Entropy Index (PEI)

$$PEI = -\frac{1}{39}\left(\sum_{i=1}^{39}\sum_{k=1}^{3}\mu_{ik}\ {}^{2}\log\mu_{ik}\right) = 2.9697\text{e}-04.$$

2. Classification Entropy (CE)

$$CE = -\frac{1}{39}\sum_{i=1}^{39}\sum_{k=1}^{3}\mu_{ik}\log\mu_{ik} = 2.5710\text{e}-04.$$

## 4. Conclusion

Based on the results of this research about FCM clustering e-commers in Indonesia the Squared Euclidean distance is good because after checking with two index the result are close to zero. Clustering using PEI is 2.9697e-04. and using CE is 2.5710e-04.

## References

[1] Kusumadewi S and Purnomo H 2010 *Aplikasi Logika Fuzzy untuk Pendukung Keputusan Edisi 2* (Yogyakarta : Graha Ilmu)
[2] Khoiruddin A 2007 Menentukan Nilai Akhir Kuliah dengan Fuzzy C-Means *SNSI* **7** pp 232-38
[3] Anggraeni W 2015 Penentuan Nilai Pangkat Pada Algoritma Fuzzy C-Means *Faktor Exacta* **3** pp 266-78

[4] Agustini F 2017 Implementasi Algoritma Fuzzy C-Means Studi Kasus Penjualan di Sushigroove Restaurant *Jurnal Ilmu Pengetahuan dan Teknologi Komputer* **3** pp 127-32.
[5] Kusumadewi S 2001 *Analisis dan Desain Sistem Fuzzy* (Yogyakarta: UII)
[6] Kusumadewi S and Purnomo H 2004 *Aplikasi Logika Fuzzy untuk Mendukung Keputusan* (Yogyakarta: Graha Ilmu)
[7] Sakthivel E and Kannan K S 2013 Clustering Algorithms using Different Distance Measures *CiiT International Journal Data Mining and Knowledge Engineering* **4** pp 140-43
[8] Prasetyo 2014 *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab* (Yogyakarta: ANDI)
[9] Iprice 2018 *The Map of E-commerce in Indonesia*
[10] Kusumadewi Sri 2002 *Analisis dan Desain Sistem Fuzzy Menggunakan Tool Box Matlab* (Yogyakarta: Graha Ilmu)
[11] Mashfuufah S and Istiawan D 2018 *The 7th University Research Colloqium 2018 STIKES PKU Muhammadiyah Surakarta* **7** pp 51-60