Decision Science Letters 12 (2023) 617–628

Contents lists available at GrowingScience

Decision Science Letters

homepage: www.GrowingScience.com/dsl

Modeling of citizen science cluster in making decision for readiness towards bogor smart village: An application of fuzzy c-means algorithm

Eneng Tita Tosida[a*], Riko Setiawan[a], Irma Anggraeni[a], Roni Jayawinangun[b], Sukono[c] and Jumadil Saputra[d]

[a]*Department of Computer Science, Universitas Pakuan, Bogor, Indonesia*
[b]*Department of Communication Science, Universitas Pakuan, Bogor, Indonesia*
[c]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Jatinangor, Indonesia*
[d]*Faculty of Business, Economic and Social Development, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia*

C H R O N I C L E

Article history:
Received: November 21, 2022
Received in revised format: December 28, 2022
Accepted: April 4, 2023
Available online: April 4, 2023

Keywords:
Fuzzy C-means
Information Gain
Citizen Science
Clustering
Smart Village

A B S T R A C T

The construction of smart villages has begun in many Indonesian villages, along with the advancement of technology and local economic growth. Villagers must participate in constructing the smart economy-smart village by becoming familiar with the characteristics of the village's inhabitants using the citizen science model. This study intends to categorize villagers so that researchers can assess and decide their level of readiness for a smart economy in an ecosystem based on a smart village. Clustering is required to find communities of residents who are ready based on their traits. Using fuzzy C-Means with a Davied Bouldin Index value of 0.129, the data were divided into 4 clusters. The most important variables were chosen using information from the test's 300 responders, and the Kaiser Mayer Olkin assumption of 0.975 was used to validate the results. Our paper provides new information on how smart village readiness is assessed by the citizen science cluster. It was decided to divide residents into four groups: those who are less prepared (24.33%), those who are somewhat prepared (29.33%), those who are ready (25.67%) %), those who are ready (level of participatory knowledge), and those who are very ready for the smart economy (20.67%) based on the cluster model.

1. Introduction

Smart villages and rural communities are built on the strengths and resources of the area. Additionally, efforts are being made to develop new opportunities where traditional and modern networks and services are improved through digital technology, telecommunications, innovation, and knowledge-based smart use. Citizen science in village development can be seen as villager participation in data collection for the scientifically based study. Smart village research and citizen science research, in which villagers work with scientists to analyze and collect data that will eventually be useful for better resource management, are indistinguishable from one another (Maja et al., 2020; Tosida et al., 2020). The ease with which anyone can now get and distribute data due to technological improvements is one of the causes. A rise in citizen scientists' active participation can be seen in the growth of scientific communication. Analyzing the inhabitants' personalities is necessary to create a strong economic ecosystem dependent on the villagers' maturity. Social science concepts like citizen science and citizen science lend credence to this. In the developing discipline of "citizen science," scientists and citizens collaborate to produce new knowledge that advances both science and society (Beza et al., 2017; Shamir et al., 2016).

* Corresponding author.
E-mail address: enengtitatosida@unpak.ac.id (E. T. Tosida)

doi: 10.5267/dsl.2023.4.003

At present, the village is considered capable of developing and innovating in alleviating the problems in the village. Villages are encouraged to be advanced and independent, hoping to develop their various potentials for their village development. Particularly with strong backing from the government, in the form of Village Funds (DD), and local governments to finance its growth (Article 72 of the Village Law), which genuinely intends to enhance the standard of living in rural communities and lower poverty (Article 78 of the Village Law).

Poverty in rural areas is caused by the development gap between regions, which negatively impacts the community's social life and becomes a serious problem. Based on the results of the Developing Village Index (IDM) data collection, the level of village development is categorized as being behind, developing, advanced, and independent. The IPD (Village Potential Index) classification was then used to identify 14.461 underdeveloped villages (19.17%), 55.369 developing villages (73.40%), and 5.606 independent villages (7.43 percent). In 2015-2019, the number of underdeveloped villages greatly increased to their status as developing villages. It assumes that village development has been carried out properly as the Village Law mandates. Still, if we examine further, the number of independent villages is only about 7%, meaning only a few once-developed villages that rose to their status as independent villages (Ella & Andari, 2018). Efforts to reduce rural farmer poverty are a high priority in many countries (Tosida et al., 2022b). One notion that has been successful in reducing poverty among farmers in rural regions is the smart village concept. The success of smart people in village projects depends on inhabitants and stakeholders working together to achieve those goals (Tosida, Herdiyeni, et al., 2020).

Analyzing the residents' personalities is required to actualize an economic ecosystem and develop a decent smart village, both of which depend on the maturity of the villagers. It is important to strengthen social science concepts like citizen science or citizen science to analyze the characteristics of citizens. Collaboration between scientists and the general public to produce new information for science and society is known as citizen science.

This study uses clustering to analyze and decide on ecosystem readiness in smart villages for a smart economy. Next, the author's research will use an unsupervised method using fuzzy c-means (FCM). FCM algorithms are the most popular clustering algorithms because of their ease of use and speed, according to Hendalianpour et al. (2017). FCM aims to classify data points into various clusters, allowing for precise and reliable data recovery without losing data accuracy due to algorithm blurring (Reddy et al., 2019). Furthermore, FCM is more robust to noisy data and outliers than K-means (Wiharto & Suryani, 2020), and more flexible in handling complex data structures than other clustering algorithms (Zhang & Shen, 2014).

2. Literature Review

2.1. Citizen Science

There are growing computational technologies and citizen science initiatives (Ponti & Seredko, 2022). The study of science and technology by society is known as citizen science or science in a different sense. Whereas citizen science is defined by The Oxford English Dictionary as a scientific study carried out by members of the general public, frequently in collaboration with professional scientists or under the supervision of professional scientists or scientific institutions. Actors who engage in citizen science are citizen scientists (Assumpcao et al., 2019). Clustering and citizen research have been combined in various scientific domains, including environment and health (Ozyigit et al., 2019; Kirschke et al., 2022). They also cover this strategy's possible advantages and difficulties and advise on developing and carrying out fruitful citizen science and clustering projects. A citizen scientist is a volunteer who gathers and/or processes data as part of a scientific investigation. Although citizen science has its roots in the early beginnings of modern science itself, projects involving citizen scientists are expanding, especially in ecology and the environmental sciences (Silvertown, 2009).

Clear objectives, reliable data, citizen empowerment, effective communication, scientific input, and the capacity to serve as a resource and advance alongside research are essential to successful citizen science. The four levels of citizen science can be divided into three categories: 1) Crowdsourcing, where citizens act as sensors to aid in data collection and input; 2) Distributed Intelligence, where citizens act as interpreter bases; and 3) Participatory Science, where citizens actively engage in the process. In 3) Extreme / Collaborative Science, citizens formulate problems, gather data, and even analyze that data. So, a citizen science study may involve participation from professional scientists, credentialed scientists, university scientists, residents, amateurs (hobbyists), community members, volunteers, indigenous people, and human sensors (Tosida, Herdiyeni, et al., 2022).

2.2. Information Gain

To determine the upper bounds of an attribute's significance, researchers commonly use the feature selection technique known as "information gain" (Azhagusundari & Thanamani, 2013; Deng & Runger, 2016), which is the difference between an object's entropy value before and after separation. The attributes that will ultimately be used or eliminated are presented only in the first step of this value measurement, qualities that meet the weighting criteria for the categorization step of the algorithm. The feature selection and data gathering methods are divided into three parts (Maulana & Karomi, 2015), i.e.:

a. The first step is to determine the information gain acquisition value for each attribute in the dataset that must be processed.
b. Specify the desired threshold. As a result, attributes with weights equal to or higher than the limit can be preserved, whereas attributes with weights lower than the limit can be removed.
c. After determining the dataset's highest information gain value, the attributes will be reduced.
Claude Shannon developed the concept of this attribute's measurement for the first time in information theory (Maulana & Karomi, 2015; Tangirala, 2020) and written as:

$$info(D) = -\sum_{i=1}^{m} pi\ log_2(pi) \quad (1)$$

where:
D : Case set
M : Number of partitions D
pi : Proportion of Di to D

While pi is the probability of a tuple in D that falls into class Ci and is estimated by |Ci,D| / |D|. The log function, in this case, uses log-based 2 because the information is encoded bit-based. The calculation of the entropy value after separation can be done using the following formula:

$$info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times\ I\ (D_j) \quad (2)$$

where:
D : Case set
A : Attribute
v : Number of attribute partitions A
|Dj| : Number of cases on j partition
|D| : Number of cases in D
I (Dj) : Total entropy in partition

Meanwhile, to find the information gain attribute A, the following formula can be used:

$$Gain\ (A) = I\ (D) - I\ (A) \quad (3)$$

where:

Gain (A) : A attribute information
I (D) : amount of entropy
I (A) : *entropy* A

2.3. Kaiser Mayer Olkin

Before entering the factor analysis stage, several assumptions need to be made, namely the assumption of data adequacy and correlation between variables. The Kaiser Mayer Olkin (KMO) test aims to determine whether all the data taken are sufficient to be factored (Prasetyo et al., 2020; Shrestha, 2021). The hypothesis of KMO is as follows:

H_0: The amount of data is sufficient for factor analysis
H_1: The amount of data is not sufficient for factor analysis

Test Statistics:

$$KMO = \frac{\Sigma_{i=1}^{p}\Sigma_{j=1}^{p} {r_{ij}}^2}{\Sigma_{i=1}^{p}\Sigma_{j=1}^{p} {r_{ij}}^2 + \Sigma_{i=1}^{p}\Sigma_{j=1}^{p} {a_{ij}}^2} \quad (4)$$

where:

i = 1,2,3, …, p and j = 1,2, …,p with i ≠ j

r_{ij}: correlation coefficient (relationship between 2 variables) between variables *i* and *j*
a_{ij}: partial correlation coefficient (relationship between 2 variables controlling other variables) between variables *i* and *j*

2.4. Fuzzy C Means

FCM is a flexible clustering technique that may be utilized in various research fields to cluster data and extract insightful information. It consists of market segmentation (Phuc & Chi, 2021; Shi et al., 2015), health (Christyanti et al., 2022; Setiawan et al., 2023), environmental monitoring (Lusiana et al., 2023; Rajput & Kumaravelu, 2021), etc. The Hard C-Mean clustering approach gave rise to the Fuzzy C-Means clustering method in 1981. Unsupervised clustering algorithms such as FCM are used to solve problems with feature analysis, clustering, and classifier building. It is widely applied in fields such as agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis and target recognition. The FCM clustering technique, which is actually based on Ruspini Fuzzy clustering theory, was proposed in the 1980s along with the development of fuzzy theory. The cluster centres are generated for each cluster, and the clusters are formed based on the spacing between the data points (Ghosh & Dubey, 2013). The FCM method assigns a degree of membership to each class based on a fuzzy membership. Like the pixel probability in a mixture modeling assumption, the degree of membership in fuzzy clustering is crucial. FCM helps create new clusters from data points with close membership values to existing classes. Fundamentally, the fuzzy membership function, partition matrix, and objective function are the three fundamental operators of the FCM technique (Nayak et al., 2015). The clustering approach involves grouping data and their parameters into categories based on the tenacity of each data type (similarity of properties). Using a method known as fuzzy c-means clustering, the optimal cluster in a vector space is chosen using the Euclidean normal form for the distance between vectors. Fuzzy grouping is beneficial for locating fuzzy rules in fuzzy modeling. The Fuzzy C-Means approach starts by locating the cluster center, which will act as the average position for each generated cluster. Under the initial conditions established by the first iteration computation, the center of this cluster is still not exact. Each cluster has a different level of membership for each data point. By continuously improving the cluster center and the degree of membership of each data point, it can be shown that the cluster center will move to the correct location with each iteration. This iteration is based on the cluster's center. It is based on the minimization of the objective function, which calculates the distance between a specific data point and the cluster's center and is weighted by the degree of the data point's membership (Nugraha & Riyandari, 2020). The Fuzzy C-means method's algorithmic stages for calculation are as follows:

a. The input data to be clustered, *X*, is a matrix measuring $n \times m$ (n = number of data samples, m = attributes of each data). Xij the i sample data (i = 1,2, ,n), the j attribute (j = 1,2,…,m).
b. Define:
   - Number of clusters = c
   - Rank = w
   - Maximum iterations = MaxIter
   - Smallest error expected = ξ
   - Initial objective function = $P_0 = 0$
   - Early iteration = $t = 1$
c. Generate random number μ_{ik}, $i = 1,2,\ldots,c$; as elements of the initial partition matrix U.
   Count the number of each column

$$Qi = \sum_{k=1}^{c} \mu\, ik \text{ with } j = 1,2,\ldots,n \tag{5}$$

   Count: $\mu\, ik = \frac{\mu\, ik}{Qi}$

d. Cluster centroid calculation formula to -k, V_{kj} with k=1,2,⋯,c and j=1,2,⋯,m

$$V_{kj} = \frac{\sum_{i=1}^{n}((\mu ik)^w * x_{ij})}{\sum_{i=1}^{n}(\mu_{ik})^w} \tag{6}$$

e. Partition matrix change formula

f.
$$\mu_{ik} = \frac{[\sum_{j=1}^{m}(X_{ij} - V_{ij})^2]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}[\sum_{j=1}^{m}(X_{ij} - V_{ij})^2]^{\frac{-1}{w-1}}} \tag{7}$$

   with: i=1,2,⋯,n and $k = 1,2,\cdots,c$

g. The objective function formula in the th iteration -t, Pt

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{k=1}^{m}(Xij - Vkj)^2\right](\mu_{ik})^w\right) \tag{8}$$

h. Check stop condition
   - If: ($|Pt - Pt{-}1| < \xi$) or ($t >$ MaxIter) then stop;
   - If not: $t = t{+}1$, repeat step to-d.

2.5. Davies-Bouldin Index

The method for evaluating the results of the clusters formed is using a certain algorithm, namely, the Davies-Bouldin Index (DBI), used to evaluate clusters. The idea method was introduced by David L. Davies and Donald W. Bouldin, and this method uses both names so that the Davies-Bouldin index method appears (Radius et al., 2020).

1) The sum of within-cluster squares (SSW) Calculating the Sum of the Square Within-Cluster value will reveal the cohesion in the i-th cluster (SSW). The proximity of the data to the cluster center point is added together to form cohesion. The following equation was used to find the sum of squares within the cluster.

$$SSW_i = \frac{1}{mi}\sum_{j=i}^{mi} d(xj.ci) \quad (9)$$

2) The cluster separation will be ascertained using the Sum of Square Between-cluster (SSB) calculation. The following equation is used to determine the Sum of Square Between clusters.

$$SSB_{i,j} = d(c_i,c_j) \quad (10)$$

3) Ratio (Ratio) aims to determine the value of comparing the i-cluster and j-cluster. The following equation is used to calculate the value of the ratio owned by each cluster.

$$R_{i,j} = \frac{SSWi+SSj}{SSBi,j} \quad (11)$$

4) Davies Bouldin Index The ratio value obtained from the equation is used to find the Davies-Bouldin Index (DBI) value using the following equation:

$$DBI = \frac{1}{K}\sum_{i=1}^{K} max_{i\neq j}\,(R_{i,j}) \quad (12)$$

From the equation, k is the number of clusters. The smaller the Davies Bouldin Index (DBI) value obtained (non-negative >= 0), the better the cluster obtained from grouping using the clustering algorithm (Surarso & Gernowo, 2020).

3. Materials and Methods

Knowledge Discovery in Databases (KDD) is a series of procedures for extracting meaningful knowledge from data, the research methodology used in this study. Data pretreatment and postprocessing are two transformation phases of Knowledge Discovery in Databases. The process of converting raw data into a format suitable for further analysis is known as data preparation. In order to discover features and data segments important to data mining processes, data pre-processing is also performed. Data mining and knowledge discovery in databases (KDD) are frequently used interchangeably to extract hidden information from a huge database. Although the two names have separate concepts, they are related. Another step in the whole process of finding knowledge in databases is data mining (Karsito & Monika Sari, 2018). This database knowledge discovery technique seeks to reveal the potential of the data gathered from the database and afterward examined for patterns, evaluated, and clarified through visualization (Watulangkouw, 2022). The stages of the database's knowledge discovery process are shown in Fig. 1.

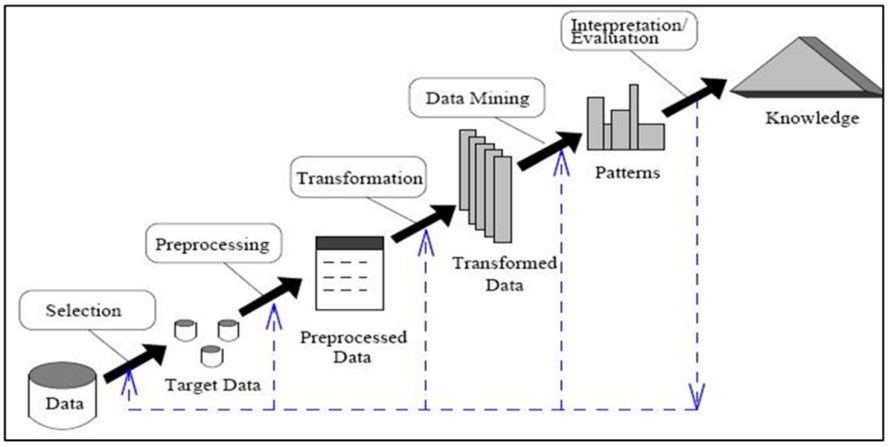

Fig. 1. Database Knowledge Discovery Process Stages

The stages of the knowledge discovery process in the database based on Fig. 1 can be explained as follows:

a. Data Selection

A new operational data set must be selected before the data mining phase of the information extraction process can begin. Then, the data selected for use in the data mining process is stored in a separate file of the operations database.

b. Pre-processing/Cleaning

Furthermore, the data that is the focus of KDD must first undergo a cleaning process used for data mining. The cleaning process includes eliminating data duplication, searching for conflicting data, and fixing problems like typos.

c. Transformation

The data that has been selected undergo a processing process called coding to make it suitable for data mining.

d. Data Mining

Data mining involves specific tools or approaches to hunting for intriguing patterns or information in chosen data.

e. Interpretation/Evaluation

It is necessary to present the data mining process's pattern of information in a way that is understandable to interested parties.

f. Knowledge

The process's final stage is how to create conclusions or take actions based on the analysis's findings.

The hard k-means approach includes a clustering technique called fuzzy c-means that uses a fuzzy grouping model to allow data to belong to any class or cluster created with membership levels ranging from 0 to 1. Fig. 2 depicts the Fuzzy c-means algorithm's steps.

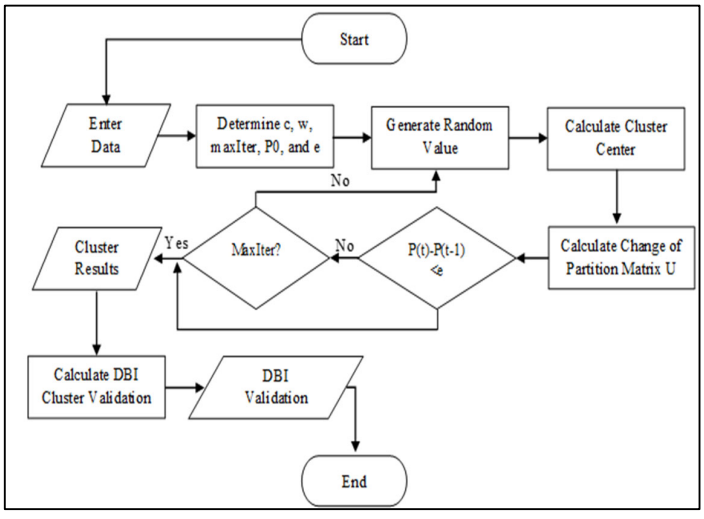

Fig. 2. Flowchart Fuzzy c-means

4. Results and discussion

4.1 Results of information gain

On the basis of Eq. (1), Eq. (2), and Eq. (3), the results of the calculation of information gain are obtained, as shown in Fig. 3. Fig. 3 shows the gain variable information value. The attribute that is reduced is the attribute with the smallest information gain value. As many as 221 attributes are reduced by 171, and only 50 attributes are taken. Irrelevant attributes will reduce machine learning performance. Meanwhile, redundant attributes will make machine learning work longer (Astuti, 2017). Based on the equation described previously for research with the largest information gain value with a value of 0.78 on the X224 variable (Ability to report internet problems to the village/regional government (after the smart village) up to 0.19 on the X167 variable (The ability level of the village/regional government in transforming based on ICT to support business development). Access to reliable and fast internet connectivity is essential for various aspects of modern life, including business, education, healthcare, and communication. In rural areas and villages, where internet connectivity can be limited,

addressing internet problems can help improve residents' quality of life and support economic development (Ruiz-Martínez & Esparcia, 2020).

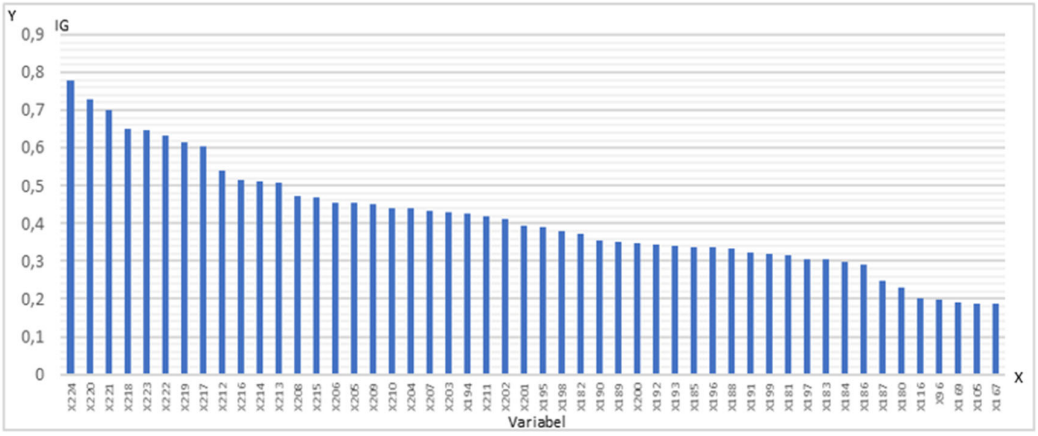

Fig. 3. Gain Variable Information Value

4.2. Result of Assumption Test for Kaiser Mayer Olkin (KMO)

The computational process obtained the KMO value of 0.975263 on all the reduced variables. Therefore, the data can be said to represent the population or representative.

Clustering

The clustering technique is used for the data mining process with fuzzy c-means algorithm. The FCM method in this study uses four clusters. Data processing in this study uses the help of the Rstudio program with the e1071 library. So the results obtained can be explained as follows.

1) Maximum Iterations

To calculate the first iteration function at P1 to P73, you can use equation (8). Then the results of the objective function are shown in Table 1.

Table 1
Objective Function Result

Iteration to	Objective Function
1	11.48
2	11.01
3	10.82
...	...
72	10.73
73	10.73

The clusters generated in Table 1 finish at the 73rd iteration with an objective function value of 10.73; the appendix shows the objective function's overall value. The fault found using this objective function is as follows.

$$|P_{73} - P_{72}| = 10.72912311 - 10.72912324| = 0.00000013 < \xi$$

2) Cluster Center

With the objective function obtained previously, the cluster center is shown in Fig. 4. Based on the results of calculating the cluster center using equation (6) can be seen in Figure 4. It can be explained by the movement of variables in cluster 1 (blue) flexibly at centroid -1. Still, at variable X116, the centroid increases at -0.5, so the next movement is at the centroid -0.5. For variables in cluster 2 (yellow), the graph movement is stable at centroid 0; for variable X193 the centroid increases to 0.5. However, it decreased back to centroid 0. The variables in cluster 3 (gray) show that the graph's movement is stable for each variable. Furthermore, the variables in cluster 4 (red) indicate that the graph movement is stable, namely moving in centroid 1, then in the variable X116 cluster 4 the centroid decreases, namely 0.5. So that the cluster 1 centroid point is

found in variable X96, the cluster 2 centroid point is found in variable X193, the cluster 3 centroid point is found in variable X193 found in variable X169, and the centroid point of cluster 4 is found in variable X194.

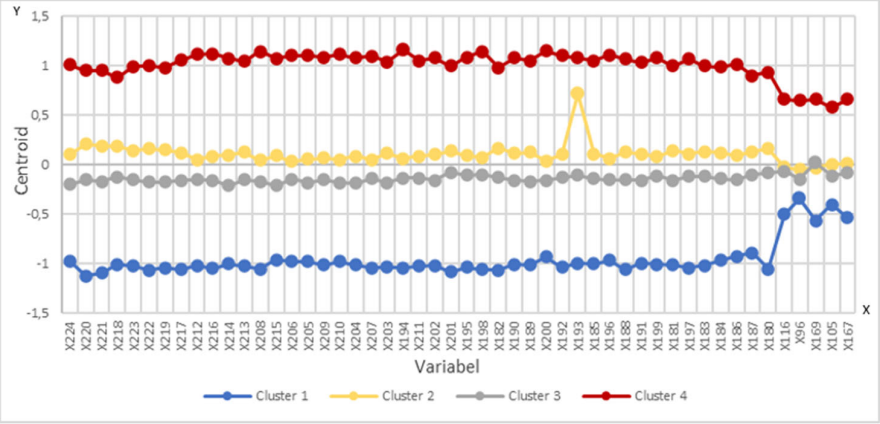

Fig. 4. Cluster Center Results

3) Membership of Each Cluster

To determine each villager's membership in the Kemang subdistrict, which contained up to 300 respondents, the degree of membership of each cluster is used, as indicated in Table 2. Using the same level of membership as in the previous iteration.

Table 2
Cluster membership degree

Respondent	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	0.299139408	0.26344983	0.35855009	0.078860664
2	0.073735045	0.28318018	0.19864521	0.444439567
3	0.041861622	0.22686507	0.13648157	0.594791741
4	0.079680221	0.44095412	0.40615227	0.073213387
5	0.097906104	0.34549766	0.50505940	0.051536833
6	0.100536661	0.36061848	0.47486736	0.063977500
7	0.086764320	0.29012896	0.21753947	0.405567252
8	0.021858430	0.19090685	0.77601516	0.011219558
9	0.065523665	0.37029966	0.52617620	0.038000469
10	0.007749029	0.08035143	0.90770880	0.004190749
...	...	...	...	...
300	...	...	...	...

Table 2 shows the output from the fuzzy c-means process using the Rstudio program. The tendency of the villagers to enter which cluster is determined by the degree of membership of each village respondents. The villagers have the strongest propensity to join the cluster, as indicated by their highest degree of membership. Fig. 5 displays the full outcomes of dividing the villages into 4 groups.

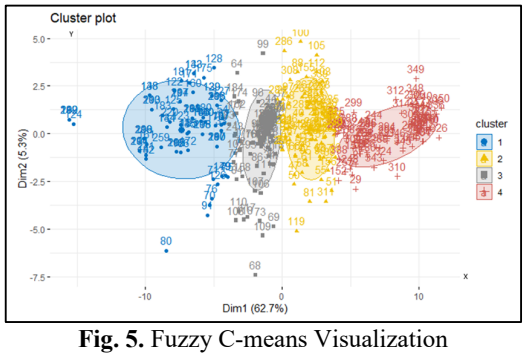

Fig. 5. Fuzzy C-means Visualization

On the basis of the fuzzy c-means method using the Rstudio program, the tendency of the villagers to join which cluster is inferred from the degree of membership of each village response. According to the highest degree of membership, villagers

have the greatest propensity to join clusters. The visualization of Figure 5 shows the full outcomes of clustering the villagers into 4 groups.

4) Cluster Validity

The Davies-Bouldin Index approach assesses the clustering method's outcomes. The principles of coherence and separation form the cornerstone of this approach. Cohesion in the clustering process is the total of the data's distance from the cluster's centroid. The separation is determined by the separation between the cluster's centroids (Dinata et al., 2020). The cluster produced by the fuzzy C-Means algorithm is better the smaller the DBI value obtained (non-negative >= 0). Table 3 provides the findings for the smallest DBI value that was obtained.

Table 3
Validity Davies-Bouldin Index (DBI)

Cluster	DBI
K=4	0.129

Referring to the DBI value in Table 3 and based on the previous analysis, the researchers found that dividing the data into 4 clusters was sufficient to explain the diversity and characteristics of the data groups and to classify the characteristics of villagers in Kemang sub-district. Figure 5 shows the population distribution of each cluster in the sub-district. Cluster 1 has the smallest population of 73 residents, primarily Kemang and Pabuaran Village residents. Cluster 2 consists of 77 residents and has the largest population of residents from Atang Sanjaya, Jampang, and West Semplak villages. Cluster 3 comprises 88 residents and has the largest population from Parakan Jaya and Pondok Udik villages. Finally, Cluster 4 has 62 residents and the largest population from Bojong and Tegal villages. As shown in Figure 6, these population distributions provide insight into the characteristics and readiness of different groups of villagers in Kemang sub-district for smart economy-smart village development.

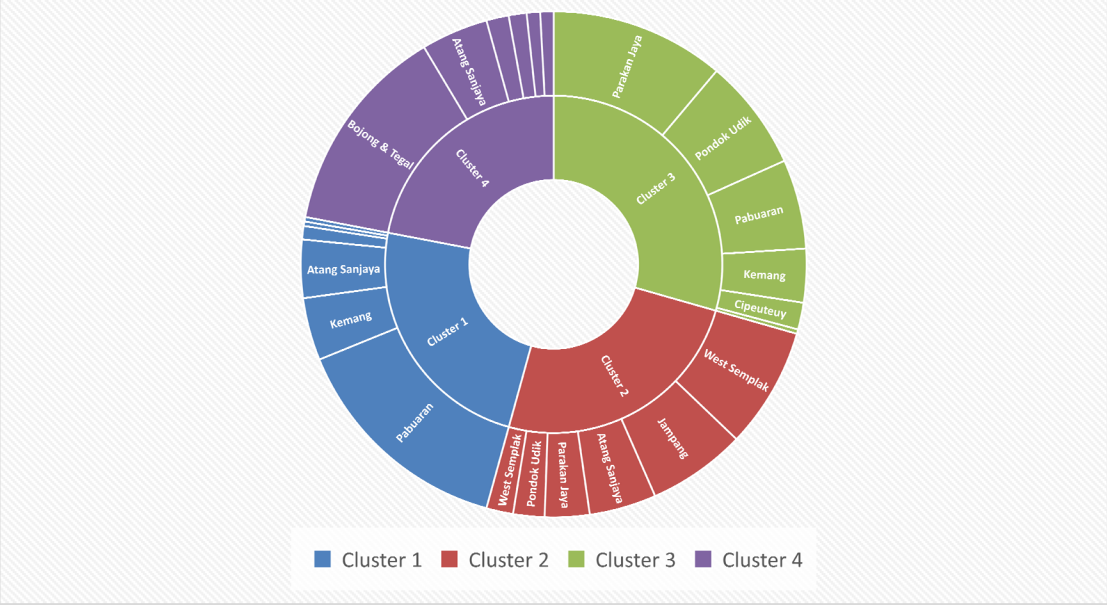

Fig. 6. population of villagers by cluster

Fig. 6 captures the outcomes of the cluster center or centroid sequence of variables: the growth of the smart economy, innovation, entrepreneurship, and empowerment, all of which directly contribute to the development of smart villages in the Kemang District. The cluster center and the cluster average can describe the group's characteristics. Based on the characteristics of the group, the average variables are sorted based on their group ranking from the highest group to the lowest ranking, which is calculated using the RANK function (Number, Ref, Order). After being sorted, group 1 is the lowest, while group 4 ranks highest. The arrangement of group rankings based on the center of the group or centroid according to Eq. (7) is as follows:

$$\textit{Centroid}\ (V_i) = \begin{matrix} \text{Cluster 1} \rightarrow \text{Ranking 4} \\ \text{Cluster 2} \rightarrow \text{Ranking 2} \\ \text{Cluster 3} \rightarrow \text{Ranking 3} \\ \text{Cluster 4} \rightarrow \text{Ranking 1} \end{matrix}$$

The clustering groups ranking indicates that the interpretation of each of the existing groups is as follows:

1) Cluster 1 consists of villagers with the lowest readiness characteristics for smart economy-smart villages compared to other groups. Villagers in this cluster have a minimum contribution value on almost all variables. This cluster consists of villagers who are least prepared for smart economy-smart village readiness compared to the other 3 clusters of villagers. In cluster 1, the centroid point is found in the village residents' decision-making during the socialization process through the empowerment variable of the Kemang Smart Village beneficiaries. Cluster 1 belongs to the crowdsourcing level.
2) Cluster 3 is the villagers who are not ready for a smart economy-smart and viewed from several aspects, only the variable contribution of the level of the ability of business owners to utilize ICT in accessing research funds has the highest value. The readiness of villagers in this group is almost equal to cluster 1, but based on its characteristics, it is still above the cluster. In the cluster of 3 centroid points, business owners can utilize ICT in accessing research funds for villagers through innovation variables, with an innovation dimension that contributes to the ability of research funds. Cluster 3 belongs to the distributed intelligence level.
3) Cluster 2 consists of villagers who are quite ready regarding several aspects of variable contribution. Only a few have a minimum value. The readiness of the villagers in this group is almost the same as cluster 4, but based on its characteristics, it is still below that group. In cluster 2, the centroid point is found in the ability to access highway infrastructure (after the smart village) through the smart economic development variable. Cluster 2 belongs to the participatory science level.
4) Cluster 4 consists of villagers with the highest readiness for smart economy-smart villages. This cluster has the maximum contribution value in almost all variables. This cluster consists of the most prepared villagers compared to the other 3 clusters of villagers. In the 4-point cluster, the centroid is found in the strength of the community being able to push the technological constraints of the villagers through the smart economic development variable. Cluster 4 belongs to the extreme citizen science level.

It is possible to distinguish between the role and participation of its citizens to determine the level of citizen science using the signs and the cluster form from the application of the fuzzy c-means algorithm, which is expected to be able to measure the readiness of citizens toward the smart economy village sages.

The smart economy-smart village will function at its best if citizen science has advanced to the point of extreme citizen science, where the majority of citizens are capable of participating at their level of involvement and may be involved in the analysis, publication, or use of results, necessitating that scientists act as facilitators in addition to their expert role. Cluster 4, which has a higher level than the other clusters but with a smaller population, it was decided that it would not have much impact on the expansion of smart villages in Kemang Regency even though it has a higher level. If a smart city or smart economy is to flourish, it must involve the public and incorporate people, institutions, and technology. Citizens' involvement is recognized as crucial to the success of a smart city or economic initiatives since it is thought to be a method to drive innovation and to develop more responsive and effective government processes (Andria et al., 2022; Ardiansyah et al., 2022; Hadian & Susanto, 2022; Tosida, Solihin, et al., 2022).

5. Conclusion

In order to create a smart economy-smart village based on citizen science, this study uses the fuzzy c-means algorithm to measure the villagers' readiness level. According to the information gain value, the 50 variables with the highest number are the most informative, which means they are the most pertinent to the target class, according to data on the variable. The influence on the process of grouping the preparedness of villagers and the contribution of variables to the smart economy-smart village increases in direct proportion to the value of information gained on a characteristic. The data can therefore be stated to represent the population or be representative since the Kaiser Mayer Olkin value of 0.975 is obtained on the entire variable, which is considered extremely good.

Using the fuzzy c-means technique, 73 entities were placed in cluster 1, 77 entities in cluster 2, 88 entities in cluster 3, and 62 entities in cluster 4, with the clustering results. On the basis of the interpretation of the results of the clustering or centroid analysis, it was decided that 24.33% of the population was very unprepared for the smart economy-smart village and were part of cluster 1, which was included in the crowdsourcing level with relatively low readiness of residents for the smart economy-smart village. This cluster comprises residents who are least prepared for smart economy-smart readiness villages. As many as 29.33% of the population not ready to go to a smart economy-smart village are included in cluster 3. It is classified as a distributed level of intelligence with the readiness of villagers not being ready to go to a smart economy-smart village. Residents fully prepared for the smart economy-smart village comprise 25.67% of the population and are included in cluster 2. As a result, the level of participatory science is determined by the villagers' readiness for the smart economy-smart village. People who are extremely prepared for a smart economy-smart village make up 20,67% of the population and are found in cluster 4, rated as having exceptional citizen science preparation.

In the Kemang District of Bogor Regency, West Java, Indonesia, where villagers' willingness to move to a smart economy-smart village is concerned, they occupy the distributed intelligence level with the greatest population dominating. Although

the comparison of the population of residents in cluster 4 with that of the entire cluster does not dominate, the degree of extreme citizen science in cluster 4 has not been able to substantially impact the readiness of the global smart economy-smart village in the sub-districts. The development of a smart economy, innovation, entrepreneurship, and empowerment are the factors that immediately follow in creating a smart economy-smart village in Kemang District. The varying degrees of villager readiness may be assessed and distinguished depending on numerous factors. Villagers with similar preparation levels are grouped by the clusters, which may then be utilized to construct targeted interventions and policies to increase readiness and further smart village development.

Acknowledgment

We would like to thank the Ministry of Education and Research and Technology Directorate General of Higher Education, Research, and Technology (DRTPM) for the Higher Education Excellence Basic Research grant (PDUPT 2022) Num.023.17.1.690523/2022 and agreement Num.111/SP2H/RT-MONO/KK4/2022.

References

Andria, F., Rahmi, A., Sunarzi, M., Nuramanah, S., Selatan, A. I., Salmah, S., Tosida, E. T., & Harsani, P. (2022). Community-Based Local Wisdom Development: Strengthening Accounting and Production Management Skills "Batik Village New Normal Bogor." *International Journal of Research in Community Services*, *3*(2), 63–70. https://doi.org/10.46336/ijrcs.v3i2.268

Ardiansyah, D., Harsani, P., Tosida, E. T., Saputra, A. O., & Bhayangkari, A. (2022). Development of A Village Information System for Acceleration of Village Services in Desa Tegal Kecamatan Kemang Bogor. *Jurnal Informatika dan Sains*, *5*(1), 54–57. https://doi.org/10.31326/jisa.v5i1.1113

Assumpcao, T. H., Jonoski, A., Theona, I., Tsiakos, C., Krommyda, M., Tamascelli, S., Kallioras, A., Mierla, M., Georgiou, H. V., Miska, M., Pouliaris, C., Trifanov, C., Cimpan, K. T., Tsertou, A., Marin, E., Diakakis, M., Nichersu, I., Amditis, A. J., & Popescu, I. (2019). Citizens' campaigns for environmental water monitoring: Lessons from field experiments. *IEEE Access*, *7*, 134601–134620. https://doi.org/10.1109/ACCESS.2019.2939471

Astuti, F. D. (2017). Seleksi Atribut Menggunakan Information Gain Untuk Clustering Penduduk Miskin Dengan Validity Index Xie Beni. *Teknika*, *6*(1), 61–65. https://doi.org/10.34148/teknika.v6i1.58

Azhagusundari, B., & Thanamani, A. S. (2013). *Feature Selection Based on Information Gain*. *2*, 18–21.

Beza, E., Steinke, J., Van Etten, J., Reidsma, P., Fadda, C., & Mittra, S. (2017). What are the prospects for citizen science in agriculture? Evidence from three continents on motivation and mobile telephone use of resource-poor farmers. *PLoS ONE*, *12*(5), 1–26. https://doi.org/10.1371/journal.pone.0175700

Christyanti, R. D., Sulaiman, D., Utomo, A. P., & Ayyub, M. (2022). Implementation of Fuzzy C-Means in Clustering Stunting Prone Areas. *International Journal of Natural Science and Engineering*, *6*(3), Article 3. https://doi.org/10.23887/ijnse.v6i3.53048

Deng, H., & Runger, G. (n.d.). *Feature Selection via Regularized Trees*.

Dinata, R. K., Novriando, H., Hasdyna, N., & Retno, S. (2020). Reduksi Atribut Menggunakan Information Gain untuk Optimasi Cluster Algoritma K-Means. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *6*(1), 48. https://doi.org/10.26418/jp.v6i1.37606

Ella, S., & Andari, R. N. (2018). Developing a Smart Village Model for Village Development in Indonesia. *Proceeding - 2018 International Conference on ICT for Smart Society: Innovation Toward Smart Society and Society 5.0, ICISS 2018*. https://doi.org/10.1109/ICTSS.2018.8549973

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).

Hadian, N., & Susanto, T. D. (2022). Pengembangan Model Smart Village Indonesia: Systematic Literature Review. *Journal of Information System,Graphics, Hospitality and Technology*, *4*(2), Article 2. https://doi.org/10.37823/insight.v4i2.234

Hendalianpour, A., Razmi, J., & Gheitasi, M. (2017). Comparing clustering models in bank customers: Based on Fuzzy relational clustering approach. *Accounting*, *3*(2), 81–94. https://doi.org/10.5267/j.ac.2016.8.003

Karsito, & Monika Sari, W. (2018). Prediksi Potensi Penjualan Produk Delifrance Dengan Metode Naive Bayes Di Pt. Pangan Lestari. *Jurnal Teknologi Pelita Bangsa*, *9*(1), 67–78.

Kirschke, S., Bennett, C., Bigham Ghazani, A., Franke, C., Kirschke, D., Lee, Y., Loghmani Khouzani, S. T., & Nath, S. (2022). Citizen science projects in freshwater monitoring. From individual design to clusters? *Journal of Environmental Management*, *309*, 114714. https://doi.org/10.1016/j.jenvman.2022.114714

Lusiana, E. D., Astutik, S., Nurjannah, N., & Sambah, A. B. (2023). Spatial delineation on marine environmental characteristics using fuzzy c-means clustering method. *Global Journal of Environmental Science and Management*, *9*(3). https://doi.org/10.22034/gjesm.2023.03.07

Maja, P. W., Meyer, J., Member, S., & Solms, S. V. O. N. (2020). Development of Smart Rural Village Indicators in Line With Industry 4. 0. *IEEE Access*, *8*(152017), 152017–152033. https://doi.org/10.1109/ACCESS.2020.3017441

Maulana, M. R., & Karomi, M. A. Al. (2015). Information Gain Untuk Mengetahui Pengaruh Atribut. *Jurnal Litbang Kota Pekalongan*, *9*, 113–123.

Nayak, J., Naik, B., & Behera, H. (2015). Fuzzy C-means (FCM) clustering algorithm: a decade review from 2000 to 2014. In *Computational Intelligence in Data Mining-Volume 2: Proceedings of the International Conference on CIDM, 20-21 December 2014* (pp. 133-149). Springer India.

Nugraha, G. S., & Riyandari, B. A. (2020). Implementasi Fuzzy C-Means Untuk Pengelompokan Daerah Berdasarkan Indikator Kesehatan. *Jurnal Teknologi Informasi*, *4*(1), 52–62. https://doi.org/10.36294/jurti.v4i1.1222

Ozyigit, T., Yavuz, C., Egi, S. M., Pieri, M., Balestra, C., & Marroni, A. (2019). Clustering of recreational divers by their health conditions in a database of a citizen science project. *Undersea & Hyperbaric Medicine: Journal of the Undersea and Hyperbaric Medical Society, Inc*, *46*, 171–183.

Phuc, N. H. T., & Chi, H. T. X. (2021). *Customer Segmentation Based on Fuzzy C-Means and Weighted Interval-Valued Dual Hesitant Fuzzy Sets*.

Ponti, M., & Seredko, A. (2022). Human-machine-learning integration and task allocation in citizen science. *Humanities and Social Sciences Communications*, *9*(1), 48. https://doi.org/10.1057/s41599-022-01049-z

Prasetyo, S. S., Mustafid, M., & Hakim, A. R. (2020). Penerapan Fuzzy C-Means Kluster Untuk Segmentasi Pelanggan E-Commerce Dengan Metode Recency Frequency Monetary (Rfm). *Jurnal Gaussian*, *9*(4), 421–433. https://doi.org/10.14710/j.gauss.v9i4.29445

Radius, T. V., Song, J., Li, F., Li, R., Plants, A., Particle, U., Algorithm, K., Anam, S., Jumadi, B., Sitompul, D., Sitompul, S., & Sihombing, P. (2020). *Analysis of determining centroid clustering x- means algorithm with davies-bouldin index evaluation*. https://doi.org/10.1088/1757-899X/725/1/012128

Rajput, A., & Kumaravelu, V. B. (2021). FCM clustering and FLS based CH selection to enhance sustainability of wireless sensor networks for environmental monitoring applications. *Journal of Ambient Intelligence and Humanized Computing*, *12*(1), 1139–1159. https://doi.org/10.1007/s12652-020-02159-9

Reddy, B. R., Vijay Kumar, Y., & Prabhakar, M. (2019). Clustering large amounts of healthcare datasets using fuzzy c-means algorithm. *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, 93–97. https://doi.org/10.1109/ICACCS.2019.8728503

Ruiz-Martínez, I., & Esparcia, J. (2020). Internet Access in Rural Areas: Brake or Stimulus as Post-Covid-19 Opportunity? *Sustainability*, *12*(22), 9619. https://doi.org/10.3390/su12229619

Setiawan, K. E., Kurniawan, A., Chowanda, A., & Suhartono, D. (2023). Clustering models for hospitals in Jakarta using fuzzy c-means and k-means. *Procedia Computer Science*, *216*, 356–363. https://doi.org/10.1016/j.procs.2022.12.146

Shamir, L., Diamond, D., & Wallin, J. (2016). Leveraging Pattern Recognition Consistency Estimation for Crowdsourcing Data Analysis. *IEEE Transactions on Human-Machine Systems*, *46*(3), 474–480. https://doi.org/10.1109/THMS.2015.2463082

Shi, D., Guan, J., Zurada, J., & Levitan, A. S. (2015). An Innovative Clustering Approach to Market Segmentation for Improved Price Prediction. *Journal of International Technology and Information Management*, *24*(1). https://doi.org/10.58729/1941-6679.1033

Shrestha, N. (2021). Factor Analysis as a Tool for Survey Analysis. *American Journal of Applied Mathematics and Statistics*, *9*(1), 4–11. https://doi.org/10.12691/ajams-9-1-2

Silvertown, J. (2009). A new dawn for citizen science. *Trends in ecology & evolution*, *24*(9), 467-471.

Surarso, B., & Gernowo, R. (2020). *IMPLEMENTATION OF K-MEDOIDS CLUSTERING FOR HIGH*. *10*(3), 119–128.

Tangirala, S. (2020). *Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm *. 11*(2), 612–619.

Tosida, E. T., Herdiyeni, Y., Marimin, & Suprehatin, S. (2022). Investigating the effect of technology-based village development towards smart economy: An application of variance-based structural equation modeling. *International Journal of Data and Network Science*, *6*(3), 787–804. https://doi.org/10.5267/j.ijdns.2022.3.002

Tosida, E. T., Herdiyeni, Y., Suprehatin, S., & Marimin. (2020, September 16). The potential for implementing a big data analytic-based smart village in Indonesia. *2020 International Conference on Computer Science and Its Application in Agriculture, ICOSICA 2020*. https://doi.org/10.1109/ICOSICA49951.2020.9243265

Tosida, E. T., Solihin, I. P., Jayawinangun, R., & Ardiansyah, D. (2022). Implementation of Multiple Discriminant Analysis (MDA) for Clustering Smart Village in West Java Based Podes (Potensi Desa) Database. *2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, 451–456. https://doi.org/10.1109/ICIMCIS56303.2022.10017815

Tosida, E. T., Suprehatin, S., Herdiyeni, Y., Marimin, & Solihin, I. P. (2020). Clustering of Citizen Science Prospect to Construct Big Data-based Smart Village in Indonesia. *Proceedings - 2nd International Conference on Informatics, Multimedia, Cyber, and Information System, ICIMCIS 2020*, 58–63. https://doi.org/10.1109/ICIMCIS51567.2020.9354323

Watulangkouw, J. (2022). Application of Data Mining to Determine Promotion Strategy Using Algorithm Clustering at SMK Yadika 1. *JISA(Jurnal Informatika Dan Sains)*, *5*(1), 35–49. https://doi.org/10.31326/jisa.v5i1.1107

Wiharto, W., & Suryani, E. (2020). The Comparison of Clustering Algorithms K-Means and Fuzzy C-Means for Segmentation Retinal Blood Vessels. *Acta Informatica Medica*, *28*(1), 42–47. https://doi.org/10.5455/aim.2020.28.42-47

Zhang, J., & Shen, L. (2014). An Improved Fuzzy *c* -Means Clustering Algorithm Based on Shadowed Sets and PSO. *Computational Intelligence and Neuroscience*, *2014*, 1–10. https://doi.org/10.1155/2014/368628