Distance Analysis Measuring for Clustering using K-Means and Davies Bouldin Index Algorithm

Ali Idrus [1], Nafan Tarihoran [2], Ucup Supriatna [3], Ahmad Tohir [4], Suwarni Suwarni [5], Robbi Rahim [6]

[1] *Universitas Jambi, Jambi, Indonesia*
[2] *Universitas Islam Negeri Sultan Maulana Hasanuddin Banten, Indonesia*
[3] *STAI Al Haudl Ketapang, Indonesia*
[4] *STKIP Al Islam Tunas Bangsa, Bandar Lampung, Indonesia*
[5] *Universitas Dehasen Bengkulu, Bengkulu, Indonesia*
[6] *Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia*

Abstract **– The purpose of this research is to analyze mapping results in the form of clusters formed using clustering method measures. This is done to determine the connections that the existing clusters create. Some of the measurements used are mixed measurements, Bregman differences, and number measurements (Mixed Euclidean Distance, Generalized Divergence, Squared Euclidean Distance, Mahalanobis Distance, and Euclidean Distance). Distance measurement shall be applied on number with primary school facilities in Indonesia. The Davies Bouldin Index (DBI) is different from the cluster number test (k = 2-10) for each Distance Measure. The average DBI value in the type of measure (mixed measure) and numerical measurement (Mixed Euclidean Distance) is 0.54. The average DBI value in the type of measure (Bregman divergences) and numeric measurements (generalized IDivergence) is 0.66. The average DBI value is 0.77 for the measurement type (Bregman divergences) and numerical measurement (Squared Euclidean Distance). From the results, the measurement of distance with mixed measurement and the mixed Euclidean distance with the cluster number (k = 2), namely 0.269, have the best DBI value.**

Keywords **– Distance Measure, K-Means, Davies Bouldin Index, Clustering.**

DOI: 10.18421/TEM114-55
https://doi.org/10.18421/TEM114-55

Corresponding author: Robbi Rahim,
Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia.
Email: usurobbi85@zoho.com

Received: 20 July 2022.
Revised: 09 September 2022.
Accepted: 27 September 2022.
Published: 25 November 2022.

1. Introduction

Clustering is a data segmentation approach implemented in several areas including marketing, market segmentation and business problem analysis, computer vision patterns, object zoning and image processing [1], [2], [3], [4]. The aim of the cluster is to search for groups of objects in the same (or associated) group and various objects in other groups [5], [6], [7], [8], [9]. The k-mean method used in the clustering generally employs the Euclidean Distance measurement method [10], [11], [12]. The aim of this study is to analyze the number of clusters (k) formed by examining the parameter of distance measurement because the measurement of a distance plays an important role in determining the similarity or regularity between data and items. The number of villages in Indonesia with primary schools was used as the data set[13]. The instrument factor such as schools [14] is one of the factors that can support learning results. School facilities are directly used tools or equipment for supporting the education process, especially the teaching, scoring and learning process [14], [15], [16].

The study topic on the number of villages with school facilities [17] was carried out using the k-medoids method. The k-medoids method is proposed as a mapping solution. This contrasts with ongoing research in which we use the k-means method. In this case, the number of clusters is also determined by comparing several distance measures which have not been found in previous studies. In the previous study,

three clusters (k = 3) were used directly in mapping areas with school facilities. The study is carried out in order to see the optimal value of the clusters formed by using the Davies Bouldin Index (DBI) algorithm. The smaller the DBI, the better the cluster was formed.

Distance measurement or similar research has been done much in the resolution of data mining problems [10], [12], [18], [19]. As was done [11] with the effect of different distance measurements on the performance of the K-means algorithm with a Mathematics Laboratory (MATLAB) application on the iris and grape dataset. The results indicate the distance measurement "correlation" indicates a better interpretation of the cluster data. Further research on categorical and numerical data sets has been conducted [20], [21], [22], [23]. The paper presents a new clustering technique for large data sets. It works with large data sets and small data sets efficiently. The main idea behind this procedure is to divide the entire process into two stages. The first step involves cost-effective measured approximate distance, which divides the data into overlapping subsets, which we call stubs. Then the grouping will be conducted in the second step to measure the correct distance between the points in general stubs only. The pilot-based clustering approach reduces the calculation time and increases its efficiency during conventional clustering.

2. Methodology

2.1. Dataset

The data set used in the study is the number of villages with primary school facilities in Indonesia. The data set comes from https://www.bps.go.id/ accessible from the https://osf.io/msk6a connection. The data set includes 34 provinces (2011, 2014, 2018). Analysis supported by software Rapid Miner. The number of clusters (k) with multiple distance measurement values to obtain the optimal Davies Bouldin Index (DBI) value was determined. The following is the research design which includes a block diagram that explains the research workflow:

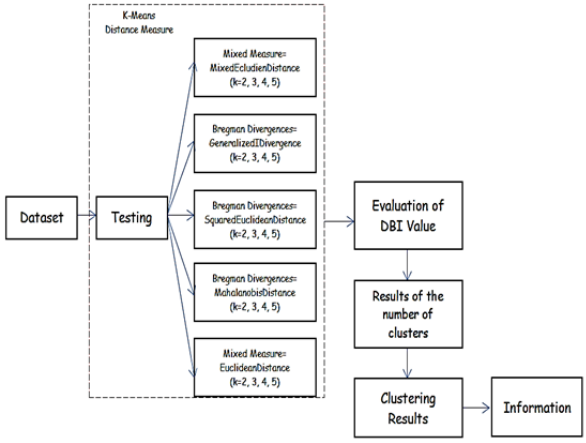


Figure 1. The proposed research flow

2.2. Distance Measure

Clustering is a way to group similar pieces of data [2], [24]. We need multiple measurements for two similar or unlike objects. The measurement is called a distance measurement [25], [26], [27] and it can be used to determine the similarity.

2.3. Clustering

Distance Clustering, so called distance-based clustering, is an extremely popular method for clustering objects and has produced good results [28]. The clusters are created so that two data objects in a group have a minimum distance value and a maximum distance value for two data objects across different clusters [27], [29].

2.4. Model Evaluation

For an accuracy indicator, the area under the curve (AUC) should increase the convergence between experiments. The following are the guidelines for AUC accuracy classification as shown in Table 1 [30].

Table 1. AUC accuracy classification

AUC	Meaning
0.90 - 1.00	Excellent grading.
0.80 - 0.90	Well-classified
0.70 - 0.80	Classification Fair
0.60 - 0.70	Bad Classification
< 0.60	Failure

3. Results and Discussion

The steps of this analysis process use Rapid Miner software to analyze distances and produce the best cluster of villages with elementary school facilities in Indonesia. The design model for a distance measurement using the K-means method using the software Rapid Miner in Figure 2 below is as follows:

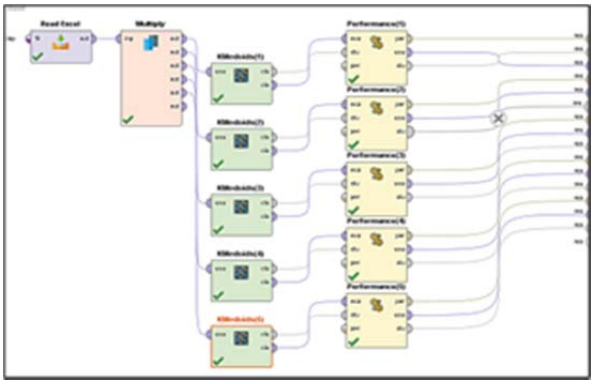


Figure 2. Design of the K-means distance measurement model

The input data set uses the "read excel" (xls) operator in Figure 2. The data is connected to several distance measurements via the operator "multiply". A multiply operator can connect several distance measurements with the k-means method. The testing is done by entering cluster values (k) from 2 to 10. The results are then recorded with the Davies Bouldin Index for comparison (DBI). The optimum number of clusters is shown with accuracy, precision, recalculation and AUC values (the smallest DBI value). The following is the recapitulation of the calculation of the analysis.

Table 2. The results of the analysis method with different Distance Measure

Measure Type	Numerical Measure	K=2	K=3	K=4	K=5
Mixed Measure	Mixed Euclidean Distance	0.269	0.294	0.478	0.37
Bregman Divergences	GeneralizedI Divergence	1.607	0.588	0.548	0.37
Bregman Divergences	Squared Euclidean Distance	0.269	1.785	1.597	0.37
Bregman Divergences	Mahalanobis Distance	0.269	0.658	0.548	1.495
Numerical Measure	Euclidean Distance	0.269	0.658	0.548	0.543

Measure Type	Numerical Measure	K=6	K=7	K=8	K=9	K=10
Mixed Measure	Mixed Euclidean Distance	1.134	0.377	1.095	0.421	0.437
Bregman Divergences	GeneralizedI Divergence	0.386	0.784	0.43	0.532	0.677
Bregman Divergences	Squared Euclidean Distance	0.349	0.774	1.015	0.409	0.4
Bregman Divergences	Mahalanobis Distance	0.816	0.359	0.502	0.778	0.437
Numerical Measure	Euclidean Distance	0.778	0.821	2.688	0.783	0.381

In table 2, it is explained that the test results for different Distance Measure for the number of clusters (k = 2 to 10) have different evaluation values (different DBI values). For measure type (Numerical Measure) and Numerical Measure (Euclidean Distance), the DBI value is not optimal, namely 2.688 (k = 8). Meanwhile, the result that the DBI value is less than optimal also occurs in the measure type (Bregman Divergences) and Numerical Measure (Squared Euclidean Distance) for k = 3 and k = 4, namely 1,785 and 1,597. In addition, the measure type (Bregman Divergences) and Numerical Measure (Generalized Divergence) also have a less optimal DBI value, namely 1.607 (k = 2). The measure type (Bregman Divergences) and Numerical Measure (Mahalanobis Distance) also have a less optimal DBI value, namely 1.495 (k = 5). The following are the results of the bar chart of each measure type as shown in Figure 3.

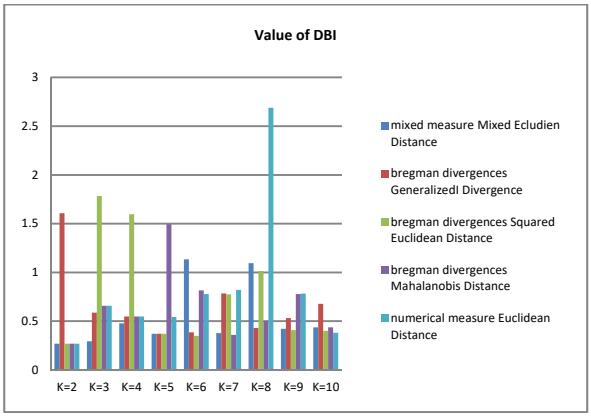


Figure 3. Diagram of the results of the analysis method with different Distance Measure

In addition, there are several results from measure types that have optimal DBI values. This is obtained by taking the DBI average value for each measure type. So that for measure type (mixed measure) and numerical measure (Mixed Euclidean Distance), the average DBI value is 0.54. This result is much better than other measure types. For measure type (mixed measure) and numerical measure (Mixed Euclidean Distance), the number of clusters (k = 2) has the most optimal DBI value of 0.269. Here are the measure types that have the optimal DBI value as shown in the following figure:

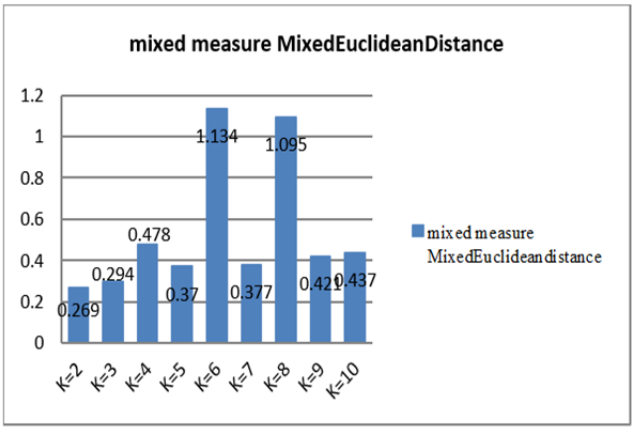


Figure 4. The measure types that have the optimal DBI value

By using a measure type (mixed measure) and a numerical measure (Mixed Euclidean Distance) at k = 2, the clustering results formed will be tested by looking at the accuracy, recall, precision and AUC values. The following are the results of the clustering formed for k = 2.

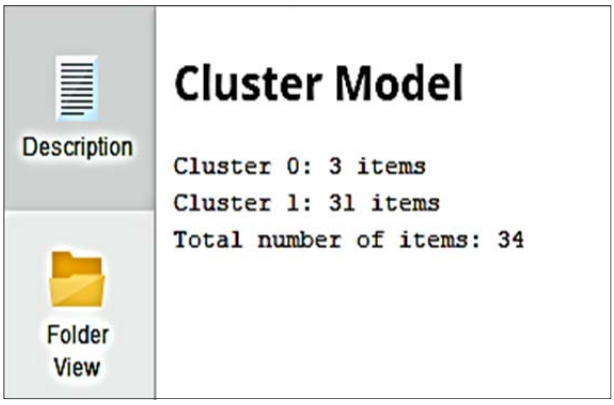


Figure 5. Clustering formed

Figure 5 shows the clusters of 34 datasets, namely 3 high cluster provinces (cluster 0), and 31 low cluster provinces (cluster 1).

This demonstrates that more than 90% of Indonesia's areas still have a minimum number of schools, particularly primary schools. The following is the exactness of the formed cluster, namely 94.17 percent.

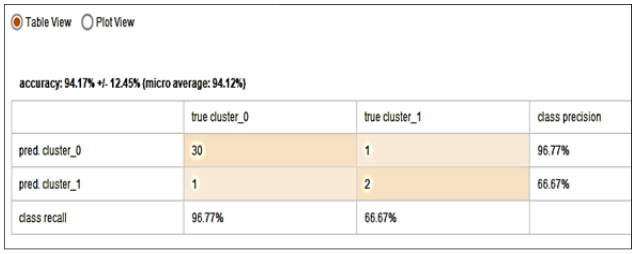


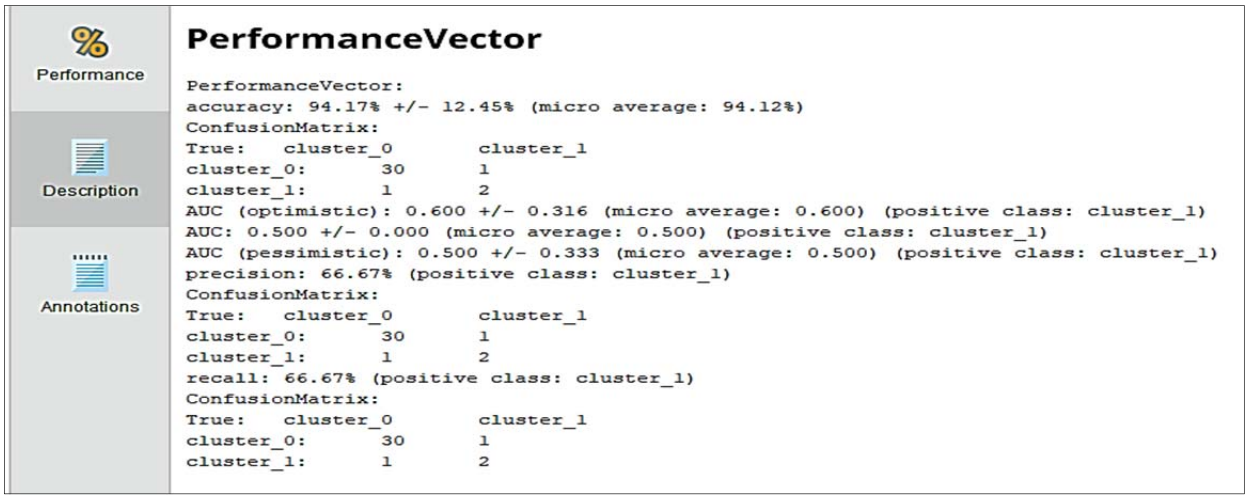
Table View Plot View

accuracy: 94.17% +/- 12.45% (micro average: 94.12%)

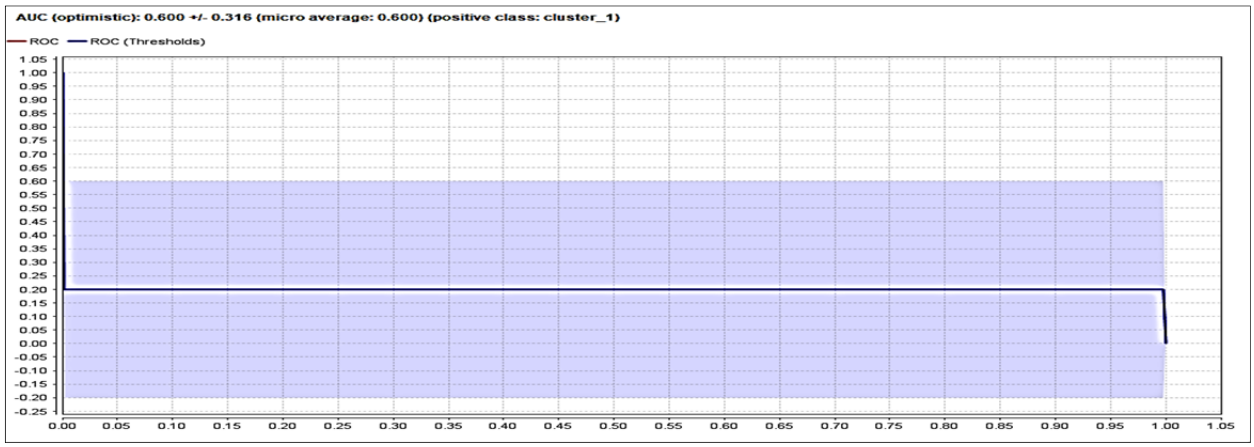
	true cluster_0	true cluster_1	class precision
pred. cluster_0	30	1	96.77%
pred. cluster_1	1	2	66.67%
class recall	96.77%	66.67%	

Figure 6. Accuracy value on the cluster formed

Moreover, from the results of the confusion matrix and the ROC curve testing the value k-means algorithm produces an precision value of 66.67, a recall value of 66.67% and an AUC value of 0.6%. The k-means algorithm is therefore an algorithm and technique to map the number of villages in Indonesia with have elementary school facilities. Then there are statistical data and AUC graphs, as shown in the figure below.



(a)



(b)

Figure 7. Recall, Precision and AUC values on the cluster formed

4. Conclusion

In this study, it can be explained that selecting a measure type will produce a different cluster. Where in the study the number of villages that have potentials for elementary schools in Indonesia can be applied and the appropriate clustering is produced. The clustering process is carried out by looking at the most optimal DBI value from a series of measure type tests. In addition, the results of the clustering that are formed will be evaluated by looking at the accuracy, recall, precision and AUC values. For the mapping of the clusters that were formed, it was found that 3 provinces were in the high cluster and 31 provinces were in the low cluster.

Data availability

The data in this study can be accessed at: https://www.bps.go.id/ accessible from the https://osf.io/msk6a

References

[1]. Feng, Z. K., Niu, W. J., Zhang, R., Wang, S., & Cheng, C. T. (2019). Operation rule derivation of hydropower reservoir by k-means clustering method and extreme learning machine based on particle swarm optimization. *Journal of hydrology*, *576*, 229-238. doi:10.1016/j.jhydrol.2019.06.045

[2]. Singh, A. K., Mittal, S., Malhotra, P., & Srivastava, Y. V. (2020, March). Clustering Evaluation by Davies-Bouldin Index (DBI) in Cereal data using K-Means. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 306-310). IEEE. doi:10.1109/ICCMC48092.2020.ICCMC-00057

[3]. Ma'arif, A., Mahendra, K. W., Ferbriyanto, S., Nuriman, A., & Habibillah, A. (2021). Training on installing solar water pump for resident of singkar 1 wareng wonosari gunungkidul yogyakarta indonesia. *Jurnal Pengabdian dan Pemberdayaan Masyarakat Indonesia*, *1*(1), 31-38.

[4]. Handana, M. A. P., & Karolina, R. (2018, February). Performance evaluation of existing building structure with pushover analysis. In *IOP Conference Series: Materials Science and Engineering* (Vol. 309, No. 1, p. 012039). IOP Publishing. doi:10.1088/1757-899X/309/1/012039

[5]. Koren, O., Hallin, C. A., Perel, N., & Bendet, D. (2019). Decision-making enhancement in a big data environment: application of the k-means algorithm to mixed data. *Journal of Artificial Intelligence and Soft Computing Research*, *9*(4), 293-302. doi:10.2478/jaiscr-2019-0010

[6]. Xu, H., Ma, C., Lian, J., Xu, K., & Chaima, E. (2018). Urban flooding risk assessment based on an integrated k-means cluster algorithm and improved entropy weight method in the region of Haikou, China. *Journal of hydrology*, *563*, 975-986. doi:10.1016/j.jhydrol.2018.06.060

[7]. Supriyadi, B., Windarto, A. P., & Soemartono, T. (2018). Classification of natural disaster prone areas in Indonesia using K-means. *International Journal of Grid and Distributed Computing*, *11*(8), 87-98.

[8]. Handrina, E. (2021). Creativity of De-Ka kopertis X canteen empowerment through various oyster mushroom processing. *Jurnal Pengabdian dan Pemberdayaan Masyarakat Indonesia*, *1*(2), 67-74.

[9]. Syafar, F.; Husain, H.; Ridwansyah; Harun, S.; Sokku, S. (2017). *Key data and information quality requirements for asset management in higher education: A case study*, Proceedings of the 30th International Business Information Management Association Conference, IBIMA 2017 - Vision 2020: Sustainable Economic Development, Innovation Management, and Global Growth, Vols 2017-January, 1670–1677

[10]. Bisandu, D. B., Prasad, R., & Liman, M. M. (2019). Data clustering using efficient similarity measures. *Journal of Statistics and Management Systems*, *22*(5), 901-922. doi:10.1080/09720510.2019.1565443

[11]. Bora, M., Jyoti, D., Gupta, D., & Kumar, A. (2014). Effect of different distance measures on the performance of K-means algorithm: an experimental study in Matlab. *International Journal of Computer Science and Information Technologies*, *5*(2), *2501–2506.*

[12]. Diogo, A. C., & Martins, A. F. (1981). Thermal behaviour of the twist viscosity in a series of homologous nematic liquid crystals. *Molecular Crystals and Liquid Crystals*, *66*(1), 133-146. doi:10.1080/00268948108072666

[13]. Susanto, R., Agustina, N., & Rozali, Y. A. (2020). Analysis of the Application of the Pedagogical Competency Model Case study of Public and Private Primary Schools in West Jakarta Municipality, DKI Jakarta Province. *Elementary Education Online*, *19*(3), 167-167. doi:10.17051/ILKONLINE.2020.03.114

[14]. Lina, E. O. (2019). Pengaruh Jumlah Desa Yang Memiliki Fasilitas Sekolah Terhadap Penduduk Buta Huruf Di Provinsi Kepulauan Bangka Belitung. *AL-ISHLAH: Jurnal Pendidikan*, *11*(1), 71-81.

[15]. Wahyuningsih, H. P., Kusmiyati, Y., & Khasanah, F. (2020). Scoring model using stunting cards for toddlers. *Pakistan Journal of Medical and Health Sciences*, *14*(2), 1419-1424.

[16]. Khairuldin, W. M. K. F. W., Anas, W. N. I. W. N., Embong, A. H., Hassan, S. A., Hanapi, M. S., & Ismail, D. (2019). Ethics of mufti in the declaration of fatwa according to islam. *Journal of Legal, Ethical and Regulatory Issues*, *22*(5), 1-6.

[17]. Damanik, I. I. P., Solikhun, S., Saragih, I. S., Parlina, I., Suhendro, D., & Wanto, A. (2019, September). Algoritma K-Medoids untuk Mengelompokkan Desa yang Memiliki Fasilitas Sekolah di Indonesia. In *Prosiding Seminar Nasional Riset Information Science (SENARIS)* (Vol. 1, pp. 520-527). doi:10.30645/senaris.v1i0.58

[18]. Patel, V. R., & Mehta, R. G. (2012). Data clustering: Integrating different distance measures with modified k-means algorithm. In *Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2011) December 20-22, 2011* (pp. 691-700). Springer, New Delhi. doi:10.1007/978-81-322-0491-6_63

[19]. Mujanah, S., Ardiana, I. D. K. R., Nugroho, R., Candraningrat, C., Fianto, A., & Arif, D. (2022). Critical thinking and creativity of MSMEs in improving business performance during the covid-19 pandemic. *Uncertain Supply Chain Management*, *10*(1), 19-28. doi:10.5267/J.USCM.2021.10.014

[20]. Bagga, S., & Singh, G. N. (211). Clustering method for categorical and numeric data sets. *Global Journal of Computer Science and Technology*, *11*(18).

[21]. Syafar, F., Gao, J., & Du, J. T. (2013). Applying the international Delphi technique in a study of mobile collaborative maintenance requirements. In *PACIS* (p. 221).

[22]. Sucipto, A., Khasanah, F., Fadililah, S., Setiawan, D. I., & Rahil, N. H. (2020). Short Message System media as an alternative education in reducing blood sugar levels of type 2 diabetes mellitus patients. *Pakistan Journal of Medical and Health Sciences*, *14*(2), 1413-1418.

[23]. Karolina, R., & Sianipar, Y. G. C. (2018, February). The utilization of stone ash on cellular lightweight concrete. In *IOP Conference Series: Materials Science and Engineering* (Vol. 309, No. 1, p. 012084). IOP Publishing. doi:10.1088/1757-899X/309/1/012084

[24]. Coelho, G. P., Barbante, C. C., Boccato, L., Attux, R. R., Oliveira, J. R., & Von Zuben, F. J. (2012, June). Automatic feature selection for BCI: an analysis using the davies-bouldin index and extreme learning machines. In *The 2012 international joint conference on neural networks (IJCNN)* (pp. 1-8). IEEE. doi:10.1109/IJCNN.2012.6252500

[25]. Javadi, S., Hashemy, S. M., Mohammadi, K., Howard, K. W. F., & Neshat, A. (2017). Classification of aquifer vulnerability using K-means cluster analysis. *Journal of hydrology*, *549*, 27-37. doi:10.1016/j.jhydrol.2017.03.060

[26]. Vergani, A. A., & Binaghi, E. (2018, July). A soft davies-bouldin separation measure. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1-8). IEEE. doi:10.1109/FUZZ-IEEE.2018.8491581

[27]. Yi, J., Zhang, Y., Yin, M., & Zhao, X. (2017, August). A novel user-interest model based on mixed measure. In *Journal of Physics: Conference Series* (Vol. 887, No. 1, p. 012061). IOP Publishing. doi:10.1088/1742-6596/887/1/012061

[28]. Ma, J., Gong, M., & Jiao, L. (2011). Evolutionary clustering algorithm based on mixed measures. *International Journal of Intelligent Computing and Cybernetics*, *4*(4), 511-526. doi:10.1108/17563781111186770

[29]. Tol, W. A., Komproe, I. H., Jordans, M. J., Susanty, D., & De Jong, J. T. (2011). Developing a function impairment measure for children affected by political violence: a mixed methods approach in Indonesia. *International journal for quality in health care*, *23*(4), 375-383. doi:10.1093/intqhc/mzr032

[30]. Wahono, R. S., Herman, N. S., & Ahmad, S. (2014). A comparison framework of classification models for software defect prediction. *Advanced Science Letters*, *20*(10-11), 1945-1950. doi:10.1166/asl.2014.5640