Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022
DOI : https://doi.org/10.33395/sinkron.v7i2.11371
e-ISSN : 2541-2019
p-ISSN : 2541-044X

K-Means Performance Optimization Using Rank Order Centroid (ROC) And Braycurtis Distance

Hafiz Irwandi[1)*], Opim Salim Sitompul[2)], Sutarman[3)]
[1,2,3)]Universitas Sumatera Utara, Medan, Indonesia
[1]hafizirwan@gmail.com, [2)]opim@usu.ac.id, [3)]sutarman@usu.ac.id

Submitted : Apr 6, 2022 | **Accepted** : Apr 10, 2022 | **Published** : Apr 14, 2022

Abstract: K-Means is a clustering algorithm that groups data based on similarities between data. Some of the problems that arise from this algorithm are when determining the center point of the cluster randomly. This will certainly affect the final result of a clustering process. To anticipate the poor accuracy value, a process is needed to determine the initial centroid in the initialization process. The second problem is when calculating the Euclidean distance on the distance between data. However, this method only gives the same impact on each data attribute. From some of these problems, this study proposes the Rank Order Centroid (ROC) method for initializing the cluster center point and using the Braycurtis distance method to calculate the distance between data. With the experiment K=2 to K=10, the results obtained in this study are the proposed method obtains an iteration reduction of 6.6% on the Student Performance Exams dataset and 19.3% on the Body Fat Prediction dataset. However, there was an increase in iterations on the Heart Failure dataset by 24.2%. In testing the cluster results using the Silhouette Coefficient, this method shows an increase in the evaluation value of 5.9% in the Student Performance Exams dataset. However, the evaluation value decreased by 8.3% in the Body Fat Prediction dataset and 3.3% in the Heart Failure dataset.

Keywords: Clustering, K-Mans, Rank Order Centroid , Braycurtis Distance, Silhouette Coefficient

INTRODUCTION

Clustering is a method for identifying data groups based on similarity measures (Tan et al., 2006). Clustering aims so that data in the same cluster are related to each other and not related to data from other clusters (Vashistha & Nagar, 2017). If the similarity in a data is getting closer , the further the difference is to other data , so that the grouping becomes more effective . Clustering is widely used in several fields including artificial intelligence, machine learning and pattern recognition (Capó et al., 2017). There are many techniques in clustering, one of which is K-Means.

K-Means is one of the ten most popular clustering algorithms and simplest method in clustering (Syakur et al., 2018). K-Means is included in the category of partitioning clustering (Sitompul et al., 2019) where each data in the cluster has the closest mean value. The process of the K-Means method begins with determining the desired number of clusters, then selecting the initial centroid at random as much as the number of clusters that have been determined previously. After that, it is continued by calculating the Euclidean distance data to the center point of the cluster. This method is carried out continuously so that no data is shifted to other clusters.

Determining the initial centroid with a random method will only give different outputs. This will certainly affect the final result of a clustering process. To anticipate the poor accuracy value, a process is needed to determine the initial centroid in the initialization process. Therefore, many researchers use various methods to determine the initial centroid of the K-Means with the aim of increasing the accuracy of the K-Means algorithm.

Based on past studies, problems were found related to the determination of the initial centroid at random and the distance between the data to the center of the centroid. The process has not been maximized and will have an influence on the results of clustering. Then several studies have been applied by adapting the initial weighting method to determine the centroid and managed to get good accuracy. So this needs to be developed in order to produce a more effective output in increasing the accuracy of the K-Means algorithm.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022
DOI : https://doi.org/10.33395/sinkron.v7i2.11371
e-ISSN : 2541-2019
p-ISSN : 2541-044X

After finishing determining the initial centroid, the next step is to determine the Euclidean Distance to the centroid. However, the traditional weighting model used such as Euclidean Distance will only have the same impact on each data attribute (Faisal & Zamzami, 2020). This certainly causes performance in data grouping to be less than optimal (Kumar & Vashistha, 2017). One solution to overcome this problem is to use the Bray Curtis Distance method (Thakur et al., 2019).

Based on past studies, problems were found related to the determination of the initial centroid at random and the distance between the data to the center of the centroid. The process has not been maximized and will have an influence on the results of clustering. Then several studies have been applied by adapting the initial weighting method to determine the centroid and managed to get good accuracy. So this needs to be developed in order to produce a more effective output in increasing the accuracy of the K-Means algorithm.

LITERATURE REVIEW

Retno's research, 2019 which determines the starting point of the centroid by calculating the Purity value. Of the four datasets tested, all four experienced an increase in accuracy and converged faster with fewer iterations than using a random starting point of the cluster center (centroid). Research from (Rahim & Ahmed, 2017) uses a new approach model for initial centroid initialization using radial and angular coordinates. The results show that in most cases the proposed method dominates in terms of processing time and iterations.

Selvida's research, 2019 used Rapid Estimation Centroid (RCE) to determine the initial centroid on K-Means. From the test results, the RCE method obtained the 7th iteration, on the contrary without RCE obtained the 9th iteration.

Several studies related to initial weighting using Rank Order Centroid (ROC) were investigated by Ahn, 2011. In his research, Ahn tested several weighting methods such as Rank Sum (RS), Rank RecipROCal (RR) and Rank Order Centroid (ROC) using maximum entropy ordered weighted averaging (MEOWA). . The results of this study, ROC is in first place with an average of 87%.

Research by Waruwu & Mesran, 2021 uses the Weighted Aggregated Sum Product Assessment (WASPAS) and ROC methods in determining young lecturers. In this study the WASPAS + ROC ranking method showed the best results than the method without ROC.

According to research conducted by (Pulungan et al., 2020) on the KNN algorithm, the Bray Curtis Distance method has a more effective performance than the Canberra Distance and Euclidean Distance methods at values of K = 6, K = 7, K = 8 and K = 10. with an accuracy value of 96%, sensitivity of 96.8% and specificity of 98.2%.

Subsequent studies classify satellite images using several distance calculation methods, namely Braycurtis Distance, Manhanttan Distance and Euclidean Distance (Alamri et al., 2016). The best accuracy when using the Braycurtis method with a value of 85% .

METHOD

Dataset

In this study, the authors used a dataset sourced from the Kaggle Machine Learning and Data Science Community. The dataset is a dataset that has been normalized, tested and valid so that it can be trusted as a data source in this study. The information from the dataset used is in the table 1.

Table 1. Dataset

No.	Dataset	Datas	Attributs	Data Type
1.	Students Performance in Exams	100	8	String,Integer
2.	Body Fat Prediction Dataset	251	15	Real,Integer,
3.	Heart Failure Prediction Dataset	918	12	String,Integer,Real

Research Stages

This study analyzes K-Means by using a combination of Rank Order Centroid (ROC) to determine the initial cluster center value and Bray Curtis Distance in the process of assigning weights to K-Means on large-dimensional datasets. The results of the cluster will be analyzed using the Silhoutte Coefficient method.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022
DOI : https://doi.org/10.33395/sinkron.v7i2.11371
e-ISSN : 2541-2019
p-ISSN : 2541-044X

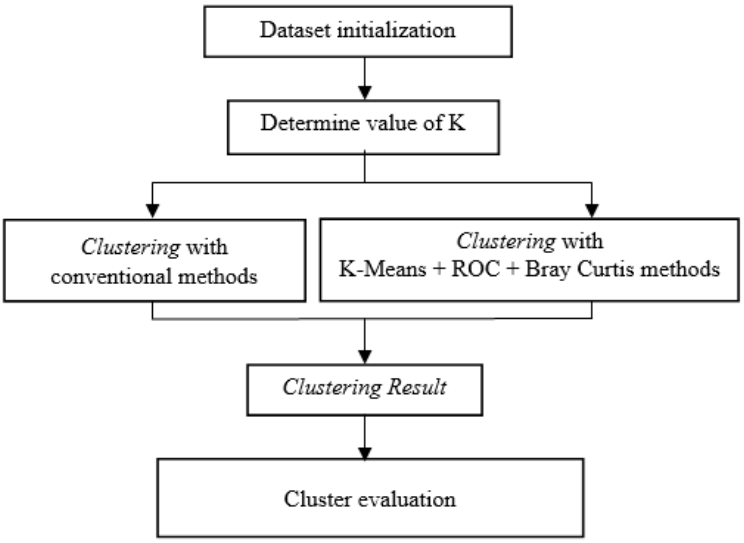

Fig. 1 Research Design

Fig. 1 shows that the initial process of this research is dataset initialization. Then proceed with determining the number of K or the desired number of data clusters. Calculate the weighting of each data using ROC and the min and max values of ROC will be the initial centroid using the formula in (1).

Calculate the distance between the data to the center of the centroid using Bray Curtis Distance using formula contained in (2). Perform the clustering process of a predetermined number of K, then calculate the average or mean of each cluster and determine the new centroid of the results of the means with the formula listed in (3). This process is referred to as one iteration. Then do the process repeatedly until you get a convergent cluster result. Do the same for classic k-means but the initial centroid is determined randomly. After the results and clusters are obtained and converge, calculate the validity of the cluster using the Silhoutte Coefficient. Silhouette coefficient is a method for evaluating the accuracy of a cluster that has been formed when clustering is carried out (Mamat et al., 2018). The silhouette coefficient method is a combination of the separation and cohesion methods (Wang & Xu, 2019), where separation serves to measure how far a cluster is separated from other clusters and cohesion serves to measure how close the relationship between data and other data in a cluster is. Then the Silhouette Coefficient value is used as a reference in comparing the accuracy between conventional K-Means and K-Means Plus.

$$W_k = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{i}\right) \quad (1)$$

$$d_{ij} = \sum_{k=1}^{n}\frac{|x_{ik}-x_{jk}|}{|x_{ik}|-|x_{jk}|} \quad (2)$$

$$\overline{v}_{ij} = \frac{1}{N_i}\sum_{k=0}^{n} x_{kj} \quad (3)$$

RESULT

Testing on Student Performance Exam Dataset

In table 2 , K-Means algorithm has an average processing time of 0.534s to execute dataset 1 starting from k=2 to k=10. Then followed by K-Means ROC with a value of 0.531s and K-Means ROC Braycurtis with a value of 0.826s. In this case, K-Means ROC and Braycurtis are considered slower than other models. The fastest time of K-Means ROC and Braycurtis when processing data with K=2 and the longest when processing data with K=10. K-Means ROC and Braycurtis iterations have an average of 7.4 followed by Conventional K-Means 7.5 and K-Means ROC 8.4. In K-Measn ROC and Braycurtis the smallest iteration is at K=2 and K=5 where the iterations are 3 iterations. Then the largest iteration is when K = 6 which is 13 iterations.In this case, K-Means occupies the top position with an average SC score of 0.360, followed by K-Means ROC of 0.352 and 0.341 for K-Means ROC and Braycurtis. In the Silhouette Coefficient method, a value close to 1 is the cluster with the strongest structure. In K-Means ROC and Braycurtis cluster the best is at K=2 with a value of 0.525 and the worst is at K=8 with a value of 0.251.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022
DOI : https://doi.org/10.33395/sinkron.v7i2.11371
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 2.Result execution of Student Performance Exam datasets

K	K-Means I			K-Means II			K-Means III		
	Time(s)	Iter	SC	Time(s)	Iter	SC	Time(s)	Iter	SC
2	0.171	6	0.639	0.175	6	0.639	0.276	3	0.604
3	0.246	9	0.707	0.250	8	0.707	0.377	9	0.722
4	0.327	14	0.801	0.367	5	0.869	0.571	5	0.907
5	0.464	4	0.954	0.427	4	0.723	0.663	3	0.720
6	0.504	4	0.922	0.489	17	0.962	0.766	13	1.137
7	0.618	5	1.054	0.696	5	1.033	1.015	7	1.089
8	0.705	12	1.117	0.701	16	1.062	1.123	8	1.373
9	0.808	6	0.965	0.782	11	1.095	1.26	12	1.105
10	0.964	8	0.913	0.892	4	1.093	1.380	7	0.957
Mean	0.534	7.5	0.897	0.531	8.4	0.909	0.826	7.4	0.957

Testing on Body Fat Prediction Dataset

The Body Fat Prediction dataset is real-type data with 15 attributes and a total of 251 data. From several studies, the results of the processing time were obtained where K-Means ROC became the fastest model with a time of 1.588s and followed by Conventional K-Means with a time of 1.668s and 2,471s for K-Means ROC and Braycurtis. In K-Means ROC and Braycurtis the fastest study was at K=2 and the longest was at K=10. The results of the research on the number of K for iterations can be seen in table 4.20. From the table, it can be seen that K-Means ROC and Braycurtis have fewer iterations, namely 10.44 compared to K-Means ROC which are worth 13.44 and 12.44 Conventional K-Means. In K-Means ROC and Braycurtis the best iteration is at K=2 and K-3, while the worst iteration is at K=6. K-Means ROC has a fairly strong structure because it is in the top rank with a value of 0.273 and Conventional K-Means with a value of 0.270 then 0.249 K-Means ROC and Braycurtis. In K-Means ROC and Braycurtis, the best SC value is when K=2 is 0.409 and the worst is K=10 with a value of 0.178.

Table 3. Result execution of Body Fat Prediction datasets

K	K-Means I			K-Means II			K-Means III		
	Time(s)	Iter	SC	Time(s)	Iter	SC	Time(s)	Iter	SC
2	0.625	10	0.430	0.522	7	0.430	0.770	5	0.409
3	0.774	8	0.311	0.829	5	0.318	1.157	5	0.308
4	0.996	10	0.244	1.160	10	0.270	1.656	7	0.253
5	1.423	8	0.275	1.328	17	0.254	2.0268	11	0.214
6	1.599	16	0.226	1.568	12	0.248	2.328	18	0.214
7	1.736	13	0.249	1.762	20	0.239	2.616	17	0.220
8	2.126	17	0.216	2.060	13	0.232	3.675	8	0.225
9	3.095	11	0.234	2.455	10	0.229	3.719	12	0.219
10	2.634	19	0.247	2.608	26	0.235	4.294	11	0.178
Mean	1.668	12.44	0.270	1.588	13.44	0.273	2.471	10.44	0.249

Testing on Heart Failure Dataset

In the experiment using the Heart Failure Prediction dataset, K-Means ROC and Braycurtis took the 3rd position where the average processing time was 12,453s. In K-Means ROC and Braycurtis the fastest processing time is at K=2 with a processing time of 3.181s and the longest is at K=9 with a processing time of 10,927. For iteration K-Means ROC occupies the 2nd position with an average iteration of 15.88 and followed by Conventional K-Means with iterations of 19.5 and finally 21.7 by K-Means ROC Braycurtis. In Braycurtis Distance the best iteration is at K=2 with 7 iterations and the worst is at K=10 with 37 iterations. The results of

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022
DOI : https://doi.org/10.33395/sinkron.v7i2.11371
e-ISSN : 2541-2019
p-ISSN : 2541-044X

the cluster validity test using the Silhouette Coefficient in table 4.25 are K-Means ROC and Braycurtis occupies the 3rd position with an average of 0.248. In the K-Means ROC and Braycurtis cluster the best is at K=2 with an SC value of 0.696 and the worst is at K=10 with a value of 0.306.

Table 4. Result execution of Heart Failure datasets

K	K-Means I			K-Means II			K-Means III		
	Time(s)	Iter	SC	Time(s)	Iter	SC	Time(s)	Iter	SC
2	2.180	7	0.693	2.070	7	0.693	3.181	7	0.696
3	3.192	8	0.465	3.246	13	0.465	4.764	15	0.455
4	4.151	29	0.420	4.722	10	0.420	5.839	29	0.364
5	5.330	18	0.361	5.297	17	0.355	7.539	22	0.358
6	5.572	22	0.362	5.878	24	0.363	8.862	24	0.360
7	6.723	24	0.276	6.341	21	0.343	10.120	21	0.338
8	7.542	24	0.349	7.492	19	0.350	10.705	22	0.339
9	8.636	22	0.345	7.972	15	0.324	10.927	19	0.306
10	8.681	22	0.314	8.225	17	0.257	12.453	37	0.248
Mean	5.779	19.556	0,398	5.693	15.889	0,396	8.265	21.778	0,384

DISCUSSION

Fig. 5,6 and 7 are iteration comparison charts of the proposed method against several datasets. In the Student Performance Test and Body Fat Prediction dataset, the proposed method shows the lowest score. This is certainly good news where this method is proven to be able to reduce the iteration value in the data set. However, things are different in the Heart Failure dataset where iteration shows the highest value. The author's initial hypothesis was that this was due to the characteristics of the data set itself. In the Heart Failure dataset many data are worth 0 . This causes a cluster difficult to achieve convergence which will then increase iterations. In Fig 8, 9 and 10 are graphs of the comparison of the Silhoute Coefficient values to the clusters in the dataset. The average increase occurred in the Student Performance Exam dataset compared to other datasets.

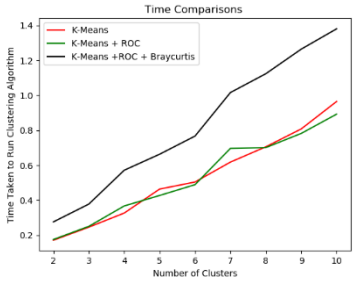

Fig. 2 Time Comparison Dataset Student Performance Exam

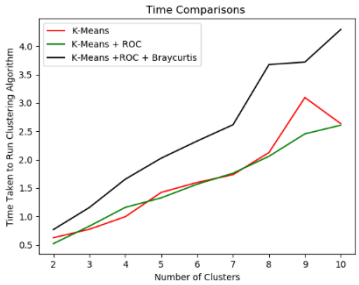

Fig. 3 Time Comparison Dataset Body Fat Prediction

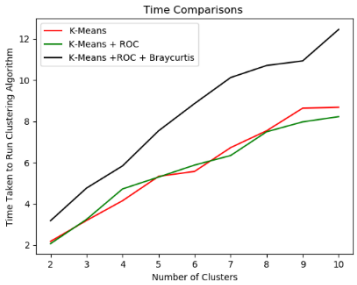

Fig. 4 Time Comparison Dataset Heart Failure

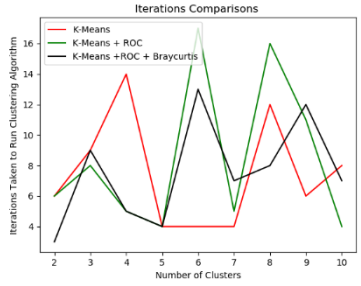

Fig. 5 Iteration Comparison Dataset Student Performance Exam

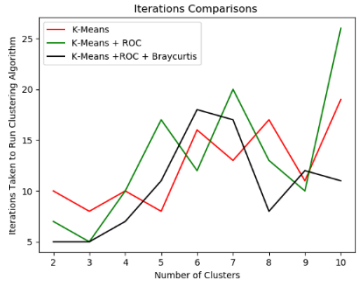

Fig. 6 Iteration Comparison Dataset Body Fat Prediction

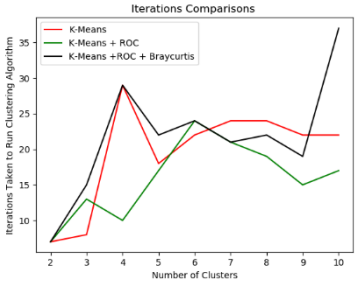

Fig. 7 Iteration Comparison Dataset Heart Failure

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022
DOI : https://doi.org/10.33395/sinkron.v7i2.11371
e-ISSN : 2541-2019
p-ISSN : 2541-044X

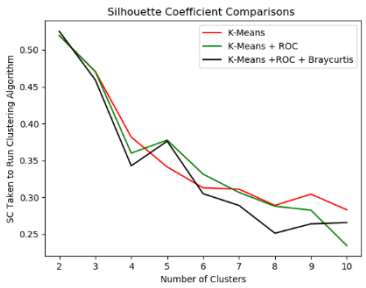

Fig. 8 SC Comparison Dataset Student Performance Exam

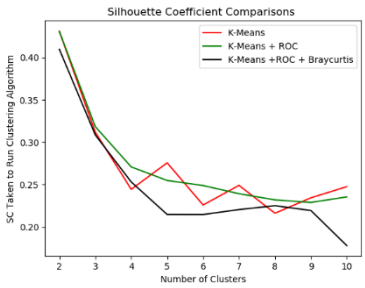

Fig. 9 SC Comparison Dataset Body Fat Prediction

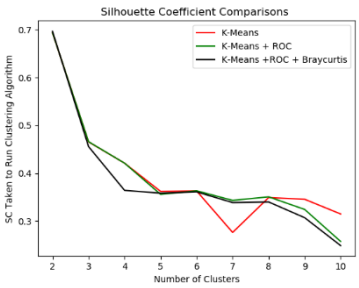

Fig. 10 SC Comparison Dataset Heart Failure

CONCLUSION AND SUGGESTION

From the research that has been done, The Student Performance Exams and Body Fat Prediction K-Means datasets ROC and Braycurtis obtained the smallest iterations compared to other models, while the Heart Failure Prediction K-Means ROC and Braycurtis datasets obtained the largest iterations compared to other models.In the Student Performance Exams K-Means ROC and Braycurtis datasets, the Silhouette Coefficient value is closest to 1, meaning that the resulting cluster has a good structure compared to trials using the Body Fat Prediction and Heart Failure dataset. Trials using other more specific datasets should be carried out in order to obtain a more detailed hypothesis on the use of ROC and Braycurtis methods on K-Means.

REFERENCES

Ahn, B. S. (2011). Compatible weighting method with rank order centroid: Maximum entropy ordered weighted averaging approach. *European Journal of Operational Research*, *212*(3), 552–559.

Alamri, S. S. A., Bin-Sama, A. S. A., & Bin-Habtoor, A. S. Y. (2016). Satellite image classification by using distance metric. *International Journal of Computer Science And Information Security*.

Capó, M., Pérez, A., & Lozano, J. A. (2017). An efficient approximation to the K-means clustering for massive data. *Knowledge-Based Systems*, *117*, 56–69.

Faisal, M., & Zamzami, E. M. (2020). Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance. *Journal of Physics: Conference Series*, *1566*(1), 012112.

Kumar, J., & Vashistha, R. (2017). Estimation of inter-centroid distance quality in data clustering problem using hybridized K-means algorithm. *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 1–7.

Mamat, A. R., Mohamed, F. S., Mohamed, M. A., Rawi, N. M., & Awang, M. I. (2018). Silhouette index for determining optimal k-means clustering on images in different color models. *International Journal of Engineering and Technology*, *7*(2.14), 105–109.

Nawrin, S., Rahman, M. R., & Akhter, S. (2017). Exploreing k-means with internal validity indexes for data clustering in traffic management system. *International Journal of Advanced Computer Science and Applications*, *8*(3), 264–272.

Pulungan, A. F., Zarlis, M., & Suwilo, S. (2020). *Performance Analysis of Distance Measures in K-Nearest Neighbor*.

Rahim, M. S., & Ahmed, T. (2017). An initial centroid selection method based on radial and angular coordinates for K-means algorithm. *2017 20th International Conference of Computer and Information Technology (ICCIT)*, 1–6.

Retno, S. (2019). *Peningkatan Akurasi Algoritma K-Means dengan Clustering Purity Sebagai Titik Pusat Cluster Awal (Centroid)*.

Selvida, D. (2019). *Analisis Klasifikasi Data dengan Kombinasi Metode K-Means dan Rapid Centroid Estimation (RCE)*.

Sitompul, B. J. D., Sitompul, O. S., & Sihombing, P. (2019). Enhancement clustering evaluation result of davies-bouldin index with determining initial centroid of k-means algorithm. *Journal of Physics: Conference Series*, *1235*(1), 012015.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP Conference Series: Materials Science and Engineering*, *336*(1), 012017.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Data mining introduction*. Bei Jing: The people post and Telecommunications Press.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Thakur, N., Mehrotra, D., Bansal, A., & Bala, M. (2019). Analysis and Implementation of the Bray–Curtis Distance-Based Similarity Measure for Retrieving Information from the Medical Repository. *International Conference on Innovative Computing and Communications*, 117–125.

Vashistha, R., & Nagar, S. (2017). An intelligent system for clustering using hybridization of distance function in learning vector quantization algorithm. *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 1–7.

Wang, X., & Xu, Y. (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. *IOP Conference Series: Materials Science and Engineering*, *569*(5), 052024.

Waruwu, F. T., & Mesran, M. (2021). Comparative Analysis of Ranking Methods of WASPAS+ ROC with Preference Selection Index (PSI) in Determining the Performance of Young Lecturers. *IJISTECH (International Journal of Information System & Technology)*, *5*(2), 207–214.

*name of corresponding author