International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN 2964-3139 e-ISSN (on Process)
Vol. 1, No. 1, page 51 - 57
Submitted 30/12/2022; Accepted 01/01/2023; Published 14/01/2023

Exclusive Clustering Technique for Customer Segmentation in National Telecommunications Companies

[1]Jhon Kristian Vieri, [2]Tb Ai Munandar and [3]Dwi Budi Srisulistiowati
[1,2,3]Informatics Department, Universitas Bhayangkara Jakarta Raya, INDONESIA
e-mail : [1]johnchristianvieri@gmail.com, [2]tbaimunandar@gmail.com

Corresponding Autor: John Kristian Vieri

Abstract
This study aims to empirically examine consumer behavior based on customer transaction history. Analyzing consumer behavior can provide very useful information for businesses in making decisions, particularly business decisions toward customers, in order to survive in such intense competition.Companies are becoming faster and more precise in reading environmental conditions and predicting what conditions may occur as a result of machine learning technology.This technology can also assist companies in making decisions that are more targeted according to actual secondary data provided for research. One of the machine learning methods, unsupervised learning, can help explicitly identify hidden structures or patterns in data and determine correlations. This method uses the Exclusive Clustering method, using two algorithms, namely, K-Means and K-Medoids, to use the comparison method to get optimal segmentation results. The results obtained are expected to be a reference for making a change in the company's marketing policy in order to retain and gain customers who are constantly decreasing.

Keywords— exclusive clustering, machine learning, K-means, K-medoids, and unsupervised learning.

1 Introduction

National Telecommunications Company, is a large company engaged in the information and communication sector that provides complete telecommunication network services. Currently, they are trying to increase customer satisfaction in the country, especially in the Bekasi city area. One of the service products offered to consumers is Indihome.

The development of internet services in Indonesia is increasing rapidly, marked by the emergence of many internet service providers. This has turned National Telecommunications Company into an indirect competitor.They compete with each other to get consumers to subscribe to their products. Some of these competitors include MNC Play, Biznet, and ICONNet. These three providers compete seriously with National Telecommunications Company to get consumer attention through the products they offer. No kidding, the products offered are also competitive in terms of price and facilities. In terms of price, National Telecommunications Company currently has a higher rate compared to its competitors. The high price of this service is often not accompanied by the maximum service and quality of Indihome products, which customers often complain about. The results of observations show that Indihome customer trend data in Bekasi City from 2016 to 2021 shows an indication that many customers choose to stop subscribing rather than continue subscribing; this can be seen from the data on the percentage of customers who are still subscribed and who are not (churn). If this trend of decreasing numbers of subscribers continues, it will disrupt existing business processes at National Telecommunications Company. Therefore, a segmentation analysis is needed to obtain information about customers that can support the company in making fast and accurate business decisions.

There are many techniques that can be used to make business decisions, especially data-driven ones, one of which is the unsupervised learning technique, namely clustering. Clustering is a data mining processing technique that is included in the unsupervised learning category, namely machine learning technology or artificial

intelligence, which is also used for business intelligence. This technology is used because it is faster and better at grouping data that doesn't have a label and figuring out what's wrong with it.

Clustering is a way of grouping data that must be understood; data mining is part of clustering. That is, it extracts patterns of interest from a large number of data sets. Clustering is commonly used in business intelligence, image pattern recognition, web search, life sciences, and security. Clustering groups data into several clusters so that the data within the clusters is more similar. It is also possible to identify and retrieve information between different clusters that have minimal similarities. Therefore, the objects in a cluster have the same characteristics and are different from those in other clusters. As previously mentioned, clustering is a method of grouping data. Because the similarity that underlies the grouping is not universal, the similarity measure must be explained in advance by the researcher or analyst. Therefore, clustering is the process of grouping data into several clusters or groups to maximize data similarity within a cluster and minimize data similarity between different clusters.

There are many studies that use cluster techniques, especially for data segmentation needs. Several studies related to segmentation use clustering, for example, for grouping customers for better business decisions as well as product recommendations and seeing customer loyalty [1], [2], [3], [4], [5], market segmentation [6], grouping documents [7], [8], [9] provinces based on development indicators [10], consumption of cosmetic products [11], grouping of health profiles [12], detection of vegetable diseases [13], fruit grouping based on digital images [14], as well as image segmentation in general [15], [16], [17]. Several studies that have been carried out provide indications that research on customer segmentation for telecommunications companies can be carried out.

This research is an attempt to segment customers based on transaction history using a clustering approach. The expected results can be used as an alternative for making business decisions, especially at National Telecommunications Company. As a result, National Telecommunications Company can prepare for inactive and inactive customers. The main contribution of research is to provide customer segmentation information to companies as material for future business decisions..

2 Research methods

2.1 Research design and stages

The design of this research will be based on the flow of the research being conducted. The first thing to do after starting the research is to collect historical data from customers and then do a descriptive analysis on the data to find out each type of variable and detailed information about the variable, then do handlers for missing data (or missing values) and outlier data (or outliers for data) to be able to perform cleaning and transformation of the data so that it can be modeled. After handling missing and outlier data, as well as transforming the data, the next step is implementing the algorithm model and evaluating each output result and the model's performance. After getting the evaluation results of each model, do a comparison to determine the best method.

2.2 2.2 Data and Analysis Tools

The dataset is taken from existing data in the company's data storage with 4 variables, and the dataset will contain information such as customer status (whether they are still subscribed or not), monthly payment history, total payment, tenor of the subscription period, and the types of services that the customer uses. To perform data analysis, we used the Python library.

2.3 K-means

K-means is a clustering algorithm that groups data based on the proximity distance between data points. The data point with the closest distance is grouped, while the data point with the farthest distance is grouped into another group. The process will be as follows:

1. Determine the number of clusters (k).
2. Initialization of the cluster center point k can be done randomly.
3. Using the Euclidean distance equation, calculate the distance between the two points as follows:

$$Distance(p,q) = \left(\sum_{k}^{n} u\,k|Pk - qk|r\right)1/r \quad (1)$$

4. Allocate all data or objects to the nearest cluster.
5. The cluster center is updated with the most recent cluster membership.Use equation (2) to update the cluster center.

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q \quad (2)$$

6. Repeat steps 3 to 5 so that the cluster center does not change or the cluster members do not change positions.

2.4 Partitioning Around Medoid (PAM)

The PAM algorithm is a classic clustering technique that clusters n object datasets into k clusters known as a priori (Abhishek & Purnima, 2013). This algorithm operates on the principle of minimizing the number of similarities between each object and its corresponding reference point. The grouping stage with PAM is as follows:

a. Initialize k cluster centers (number of clusters).
b. Count each object to the nearest cluster using the Euclidian distance measurement equation.
c. Calculation of Euclidian Distance using Equation 1.
d. After calculating the Euclidian distance, initialize a new cluster center randomly on each object as a non medoid candidate.
e. Calculate the distance between objects in each cluster that have non-medoids as candidates.
f. Calculate the total deviation (S) by calculating the new total distance minus the old total distance. If S < 0 then swap objects with non medoids cluster data to form a new set of k objects as medoids.
g. Repeat steps c through e until there is no change in the medoid. This will give you the clusters and the members of each cluster.

3 Results and Discussion

The initial stage of grouping data is to carry out analysis for each data variable; this is done to determine the readiness and relationship between variables before being trained with the exclusive clustering model. Figure 1 visualizes the distribution of data for each numeric variable, which explains if the monthly and tenor payment variables have a normal distribution because the shape is not sideways and tends to be in the middle position, while for the total payment variable it has a distribution with a negative slope because it is clear that the shape of the distribution tends to be skewed towards the left.

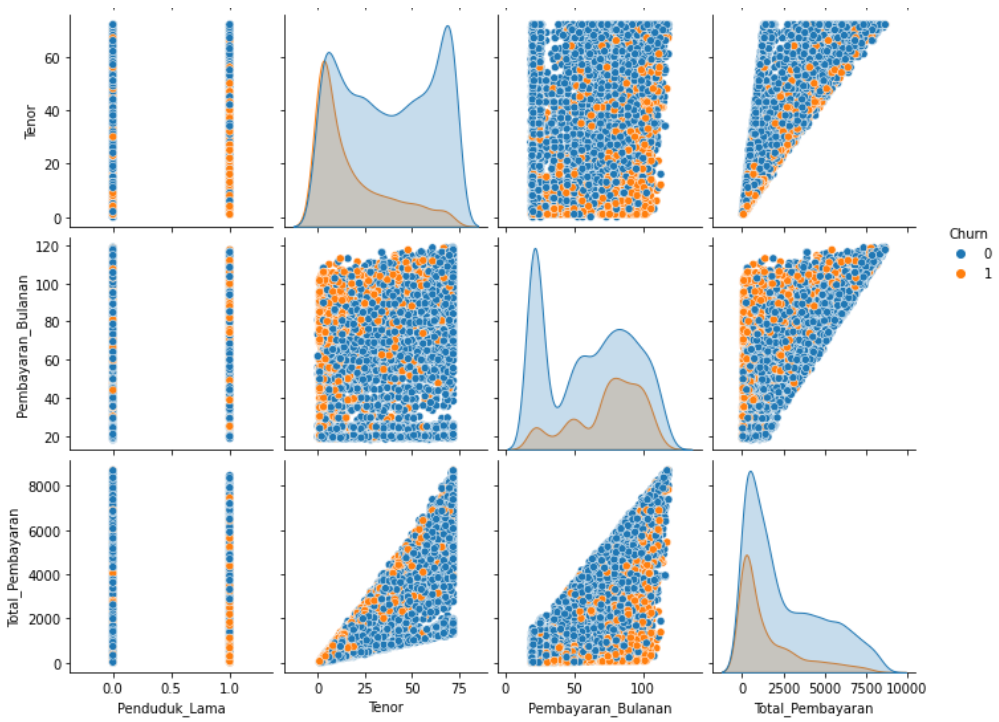

Figure 1. Data distribution for each variable

The next stage is handling missing values. At this stage, the missing values will be identified in each data variable because missing values can cause errors during the process. If there are missing values, the values will be imputed according to the missing value handling technique, but at this stage no missing values have been found.

In addition to handling missing data, it also handles outlier data or extreme data that is far from the normal majority of the variables. This is very important to do because the clustering method is very sensitive to extreme values. This will cause errors when grouping clusters or reduce accuracy and error in calculating distance between clusters.At this stage, identification of the extreme values that are far from the normal limits for each variable (outliers) will be carried out. Identification is done by creating a function that can identify outlier data and the distribution of the data. The first step in identifying outlier data is to find the 1st and 3rd quartiles of the data. The results at this stage did not find any data outliers.

After carrying out the data cleaning stages, such as handling missing values and extreme values, data transformation is also carried out, this stage is the process where the data will be converted into an array matrix. This process has several terms of use, such as numerical variables that have a normal distribution or skewed distribution and categorical variables that have nominal or ordinal types and have different handling. At this stage, it is found that if the numeric variables have a normal distribution, they can use "standardscaler()" to transform the matrix scale of the SKLEARN module, and categorical variables, which are almost all nominal, can use "onehotencoder" to transform data containing letters into numeric so that it can be transmitted. The next step is to define the model for the clustering process.

After defining the model and training the model with data, the next step is to determine the optimal number of clusters. The determination of the number of clusters is carried out using two methods, namely the elbow method. The elbow method is a method for determining the number of clusters that are accurate based on the percentage comparison results between the number of clusters that make an angled line at the cluster point. Figure 2 shows the results of determining the number of clusters with the elbow method.

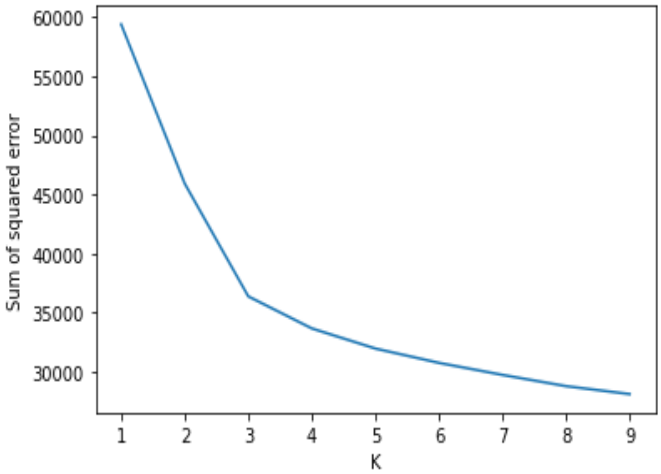

Figure 2. Visualization of the results of the elbow method

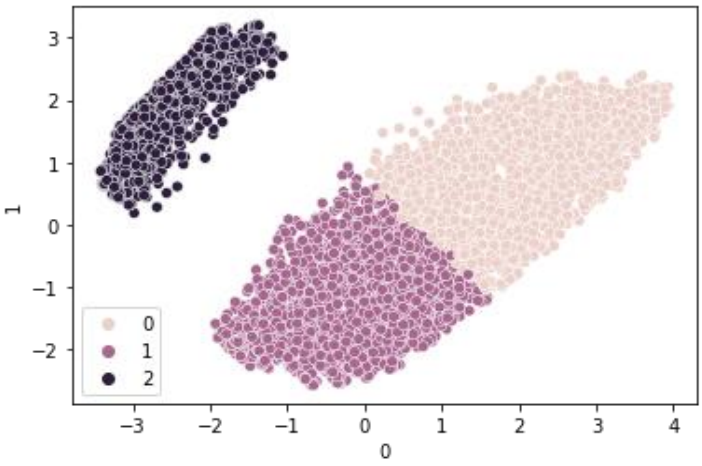

Figure 3. Visualization of K-Means Segmentation Results

Based on Figure 2, if the results obtained from visualization can be clustered up to three or four clusters, this is because the fault lines are at points three and four, in accordance with the theory of applying the elbow method. In this study, 4 clusters were used to cluster data using both k-means and PAM.

After determining the optimal number of clusters, the next step is to implement the optimal number of clusters with the K-Means model on the data to obtain segmentation results. The results of this segmentation show that the cluster group that has the highest population is cluster 3, and the cluster with the lowest population is cluster 1. The highest total income is generated by cluster 3, and the lowest is generated by cluster 1. The highest monthly income is generated by cluster 3, and the lowest is produced by cluster 1. Customers with the highest subscription times are customers from cluster 3, and the lowest are owned by cluster 1. Figure 3 is a visualization of cluster results with k-means.

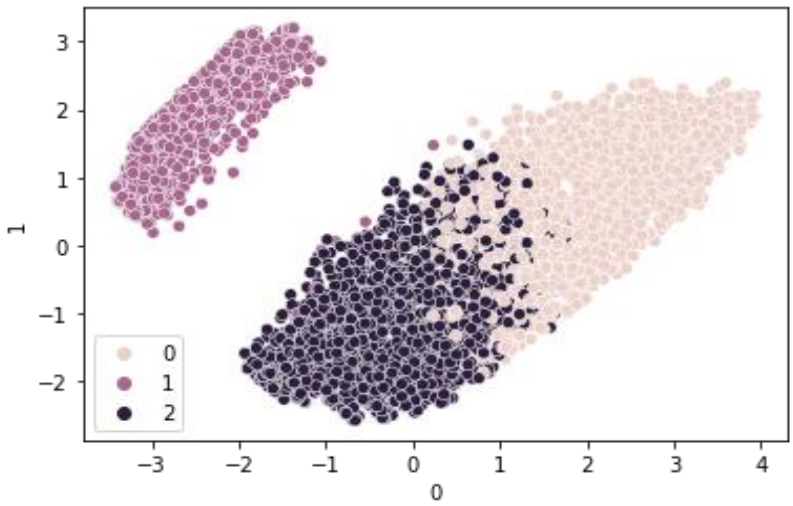

Figure 4. Visualization of PAM Cluster Results

After implementing one of the clustering exclusive models with the K-Means algorithm, the next step is implementing the PAM algorithm so that we can compare the right exclusive algorithms. The results of segmentation with PAM showed that the cluster group that had the highest population was cluster 2, and the cluster with the lowest population was cluster 0. The highest total income was generated by cluster 2, and the lowest was generated by cluster 0. The highest monthly income was generated by cluster 2, and the lowest was produced by cluster 0. Customers from cluster 2 have the longest subscription periods, while those from cluster 0 have the shortest. The standard length of a customer subscription is around 24 months, or 2 years of subscription in the last 6 years. Figure 4 is a visualization of the PAM segmentation results.

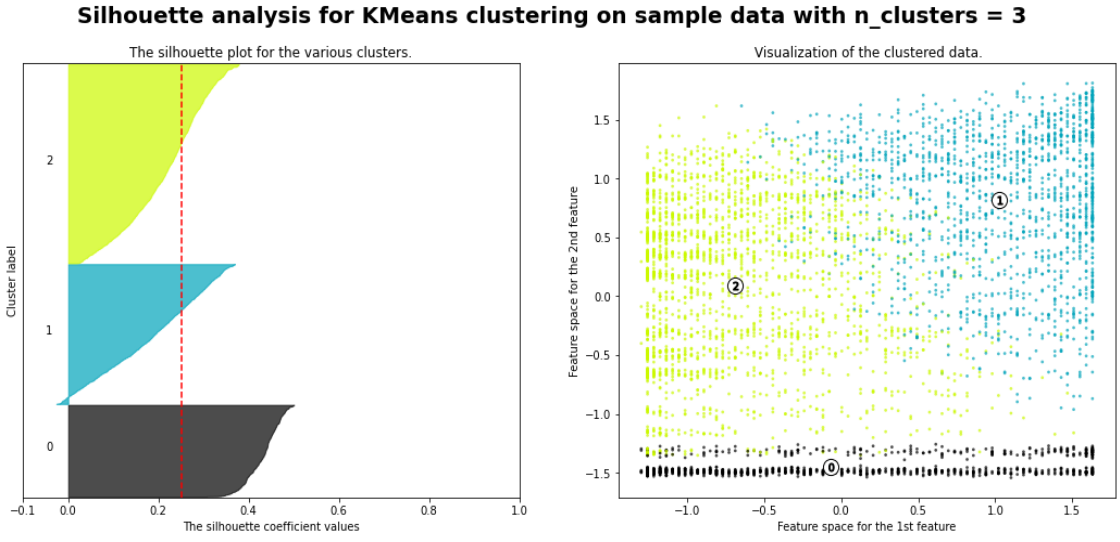

Figure 5. Evaluation visualization with three clusters

After getting the next result, it is necessary to conduct a discussion that explains the results of the model evaluation to determine the appropriate method. Based on the evaluation results, it was found that the model could segment up to 3 clusters because the results of the silhouette coefficient score for cluster 3 had the highest value compared to the other clusters with a value of 0.2514768478887445, and from the results of the analysis on the visualization results, it was found that using 3 clusters, the data could be grouped properly and there are no fatal errors in the division of cluster areas rather than clusters on cluster 3. This evaluation uses the silhouette coefficient method. This method functions to measure how far a cluster is separated from other clusters.

The results of the evaluation of the model with the number of clusters show that there are very different cluster groups. Cluster 2 has a very large majority compared to cluster 1, with a distance value, or "silhouette score," or closeness between clusters of 0.2499573832965281. While the results of the evaluation of the model with the number of clusters 3 show that clusters 3 and 2 can be grouped well, these two clusters get a balanced group as the majority, as well as cluster 1, which has a minority group but can be grouped properly. With a silhouette score or closeness between clusters of 0.2514768478887445, it is only 0.01 different from the grouping with 3 clusters. The final cluster evaluation results have a number of clusters of 4. The results show that clusters can be grouped properly, but clusters 1, 2, and 3 occupy other cluster areas with a silhouette score or closeness between clusters of 0.2014768478887445. Figure 5 is an index silhouette visualization for the evaluation of the second scenario.

Cluster evaluation was also carried out on the PAM algorithm to see how valid the cluster results were. A comparison of cluster results for k-means and PAM is shown in Table 1. Based on the comparison of the silhouette coefficient score, the optimal model for segmentation is the K-Means algorithm. This is because K-Means can group and separate each cluster properly and precisely, where only a few clusters overlap or overlap each other compared to the PAM algorithm model.

Table 1. Silhouette Coefficient Score Comparison Table

Number of cluster	K-Means	K-Medoids
2	0.2499573832965281	0.1303734179992566
3	0.2514768478887445	0.1303734179992566
4	0.2042978397035282	0.1303734179992566

4 Conclusion

Based on the research conducted, it can be concluded that the optimal exclusive cluster algorithm in this study is K-Means with three cluster groups. The results of clustering with k-means show that each cluster group only has a difference of 1 month in deciding not to subscribe anymore. Cluster groups with low monthly payments subscribe longer than customers from other cluster groups. The highest total income is obtained from the cluster group with the lowest payout. The cluster group that decided to unsubscribe is not from the cluster group with the highest monthly payments.

BIBLIOGRAPHY

[1] . A.R. Danurisa, and J. Heikal, Customer Clustering Using the K-Means Clustering Algorithm in the Top 5 Online Marketplaces in Indonesia, Budapest International Research and Critics Institute-Journal (BIRCI-Journal) Vol. 5, No. 3, DOI: https://doi.org/10.33258/birci.v5i3.6450

[2] . B. Mulyawan, M.V. Christanti, and R. Wenas, Recommendation Product Based on Customer Categorization with K-Means Clustering Method, IOP Conf. Series: Materials Science and Engineering 508, 2019, doi:10.1088/1757-899X/508/1/012123

[3] . H. Kilari, S. Edara, G.R.S. Yarra, and D.V. Gadhiraju, Customer Segmentation using K-Means Clustering, International Journal of Engineering Research & Technology (IJERT), Vol. 11, Issue 03, pp. 303 – 208

[4] . C.D.O Soleman1, N. Pramaita, and M. Sudarma, Classification Of Loyality Customer Using K-Means Clustering, Studi Case : PT. Sucofindo (Persero) Denpasar Branch, International Journal of Engineering and Emerging Technology, Vol.5, No.2, pp. 160 - 167, 2020

[5] . K. Tabianan, S. Velu, and V. Ravi, K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data, Sustainability 2022, 14, 7243, https://doi.org/10.3390/su14127243

[6] . S. H. Shihab, S. Afroge and S. Z. Mishu, "RFM Based Market Segmentation Approach Using Advanced K-means and Agglomerative Clustering: A Comparative Study," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1-4, doi: 10.1109/ECACE.2019.8679376.

[7] . R.C. Balabantaray, C. Sarma, and M. Jha, Document Clustering using K-Means and K-Medoids, International Journal of Knowledge Based Computer System, Vol. 1, No. 1, 2013

[8] . P. Gurung, and R. Wagh, A study on Topic Identification using K means clustering algorithm: Big vs. Small Documents, Advances in Computational Sciences and Technology, Volume 10, Number 2, 2017, pp. 221-233

[9] . V. K. Singh, N. Tiwari and S. Garg, "Document Clustering Using K-Means, Heuristic K-Means and Fuzzy C-Means," 2011 International Conference on Computational Intelligence and Communication Networks, 2011, pp. 297-301, doi: 10.1109/CICN.2011.62.

[10] . A.S. Ahmar, D. Napitupulu, R. Rahim, R. Hidayat, Y. Sonatha, and M. Azmi, Using K-Means Clustering to Cluster Provinces in Indonesia, Journal of Physics: Conference Series, Volume 1028, 2nd International Conference on Statistics, Mathematics, Teaching, and Research, 2017, DOI 10.1088/1742-6596/1028/1/012006

[11] . Amanda and M.V. Sitorus, Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Konsumsi Produk Kosmetik milik PT Cedefindo, Jurnal Ilmiah MIKA AMIK Al Muslim, Volume V No. 2, pp. 63 - 68, 2021

[12] . M.A.W. Saputra1, and S. Harini, Java Island Health Profile Clustering using K-Means Data Mining, Intl. Journal on ICTVol. 8, No. 1, pp. 1-9, 2022, doi:10.21108/ijoict.v8i1.606

[13] . U. Rahamathunnisa, M. K. Nallakaruppan, A. Anith and S. Kumar K.S., "Vegetable Disease Detection Using K-Means Clustering And Svm," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 1308-1311, doi: 10.1109/ICACCS48705.2020.9074434.

[14] . S.R. Dubey, P. Dixit, N.Singh, and J.P. Gupta, Infected Fruit Part Detection using K-Means Clustering Segmentation Technique, International Journal of Artificial Intelligence and Interactive Multimedia, Vol. 2, No. 2, pp. 65 - 72, DOI: 10.9781/ijimai.2013.229

[15] . N. Dhanachandra, K. Manglem, and Y.J. Chanu, Image Segmentation using K-means Clustering Algorithm and Subtractive Clustering Algorithm, Procedia Computer Science 54 (2015) 764 – 771, Eleventh International Multi-Conference on Information Processing, 2015.

[16] . V. K. Dehariya, S. K. Shrivastava and R. C. Jain, "Clustering of Image Data Set Using K-Means and Fuzzy K-Means Algorithms," 2010 International Conference on Computational Intelligence and Communication Networks, 2010, pp. 386-391, doi: 10.1109/CICN.2010.80.

[17] . K. Venkatachalam, V. P. Reddy, M. Amudhan, A. Raguraman and E. Mohan, "An Implementation of K-Means Clustering for Efficient Image Segmentation," 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), 2021, pp. 224-229, doi: 10.1109/CSNT51715.2021.9509680.