Available online at http://iorajournal.org/indx.php/orics /index

Operations Research: International Conference Series

e-ISSN: 2722-0974
p-ISSN: 2723-1739

Vol. 4, No. 3, pp. 109-116, 2023

Ensemble Rock Application for Classification of Slb in Riau Province Based on Infrastructure Facilities

Munzhiroh Rizki Minallah[1*]

[1] *Statistics Study Program, Mathematics Department, Faculty of Mathematics and Natural Sciences, University of Riau*

**Corresponding author email: munzhiroh.rizki3262@student.unri.ac.id*

Abstract

The 2020 school participation figure states that 20.56% of children in the disability category have the status of not/never been to school (BPS, 2020). This shows that there are still many children with disabilities who have not received adequate education. Therefore, attention to the availability of facilities and access to education for children with disabilities needs to be increased so that there is no inequality of school participation between children with disabilities and non-disabled children. On extraordinary school data statistical methods can be applied for various purposes. The method that can be used to group mixed-type data is ensemble. In this study, the ensemble ROCK (Robust Clustering using links) method was used at 47 extraordinary schools in Riau Province. Using the value θ of 0.22 in the ROCK ensemble method, we get 3 optimal clusters with a ratio of 0.08177794. It was found that cluster 3 is a cluster that does not have adequate facilities such as a laboratory, library and internet network than other clusters. It can be said that cluster 3 needs more attention than other clusters.

Keywords: Disability, extraordinary school, mixed type data, ensemble ROCK.

1. Introduction

Based on the law of the Republic of Indonesia (2003) number 20 of 2003 article 14, there are several types of education, including special education. Special education known to the public is special schools (SLB) and inclusive schools. SLB is an educational institution specifically to run educational programs for children with special needs (Pramartha, 2015). Children with special needs have the same rights as other normal children in receiving education, this is written in the 1945 constitution article 31 paragraph 1.

According to the Central Statistics Agency (2020), school enrollment figures in 2020 stated that 20.56% of children with disabilities had not/never attended school. This shows that there are still many children with disabilities who have not received adequate education. Therefore, attention to the availability of infrastructure and access to education for children with disabilities needs to be increased so that there is no inequality in school participation between children with disabilities and non-disabled children.

To resolve the problem regarding the lack of availability of facilities and infrastructure which was mentioned previously, statistical analysis can be applied including clustering or grouping. There are grouping methods specifically designed to group mixed variable problems, namely cluster ensemble, two step clustering and latent class cluster (LCC) (Pelaez, 2019). There are several methods for ensemble clusters, including ROCK ensemble clusters and Fuzzy ensemble clusters.

There are several previous studies that have used the ROCK ensemble cluster to group various cases. Tyagi, A and Sharma, S (2012) used ROCK to group national and international journal documents. Alvionita et al (2017) compared the ROCK and SWFM ensemble methods for grouping citrus accessions. Wulandari et al (2020) used the ROCK ensemble to group disadvantaged areas in East Java. Based on the background and previous research, the author is interested in using the ROCK ensemble method to group special schools in Riau Province.

2. Theory

Agglomerative nesting (AGNES) is a grouping method that is included in hierarchical clustering. The AGNES method is used to group numeric data. In grouping numerical data, a measure of dissimilarity is used to measure the closeness between data. The most commonly used measure of dissimilarity is the Euclidean distance and the Manhattan distance. The Euclidean distance is generated from the square root of the difference between the two data (Johnson & Wichern, 2007). The Euclidean distance is formulated as follows (Johnson & Wichern, 2007).

$$d_{ij} = \sqrt{\sum_{m=1}^{p}\left(X_{im} - X_{jm}\right)^2}\,, \tag{1}$$

The grouping step using agglomerative nesting (AGNES) follows the following algorithm (Hair et al, 2010). First, all data is considered as a single unit (cluster) so that there are N clusters and calculate the distance matrix NxN using a distance measurement scale between them, namely Euclidean, Manhattan and so on. The second step is to find the smallest distance in the distance matrix. If cluster u and cluster v have the closest distance then it is denoted by duv. The third step is to combine the clusters u and v and then give a new label to the cluster uv. Replace the entries in the distance matrix by deleting the row and column corresponding to cluster uv then adding a row and column containing the distances between cluster uv and the remaining clusters. The distance calculation uses the distance formula according to the method used, namely single linkage using equation (2), average linkage using equation (3) and complete linkage using equation (4). Repeat steps 2 and 3 $N - 1$ times.

$$d_{(uv)w} = min\{d_{uw}, d_{vw}\}, \tag{2}$$

$$d_{(uv)w} = max\{d_{uw}, d_{vw}\}, \tag{3}$$

$$d_{(uv)w} = \frac{\sum_{i}^{n(uv)}\sum_{j}^{n(w)} d_{iuvjw}}{n(uv)n(w)}, \tag{4}$$

where d_{uw} is the distance between cluster u and w and dvw is the distance between cluster v to w.

To calculate validation, the group size of the AGNES method can be measured using the Dunn index. The way the Dunn validity index works is to calculate the minimum value from the comparison between the dissimilarity function value between two clusters as separation and the maximum value of the cluster diameter as compactness (Khairati et al, 2019). Suppose there is a data set that has k clusters in which there are cluster p, cluster q, and cluster r. For example, x_i is the ith point in cluster p, y_i is the ith point in cluster q, and z_i and z_j are the ith point and j point respectively in the cluster r, $d(c_p, c_q)$ is a function of dissimilarity between clusters p and q, $d(c_p, c_q) = \min_{\mathrm{x_i \in c_p, y_i \in c_q}} d(x_i, y_i), diam(c_r)$ is the cluster diameter r, $diam(c_r) = \max_{\mathrm{z_i, z_j \in c_r}} d(z_i, z_j)$ so the calculation of the Dunn index (DN) is formulated in the following equation.

$$DN = min_{p\,=\,1,\dots,k}\left\{min_{q\,=\,i+1,\dots,k}\left(\frac{d(c_p, c_q)}{max_{r\,=\,1,\dots,k}diam(c_r)}\right)\right\},$$

The greater the Dunn index value, the better the cluster size (Dunn, 1974). To calculate the validation of the AGNES method grouping method, it is done by calculating the ratio between $sum\,within\,(S_w)$ with $sum\,between\,(S_b)$ which is denoted by $\frac{S_w}{S_b}$. Groups that show high homogeneity within groups while having high heterogeneity between groups are good groups (Hair et al, 2010) The S_w value can be formulated as in the following equation (Bunkers, Miller, & DeGaetano, 1996).

$$S_w = \frac{1}{K}\sum_{k=1}^{K} S_k, \tag{6}$$

where K represents the number of clusters formed, S_k represents the standard deviation of the k cluster.

If given cluster c_k, where $k = 1, \dots, K$ and members in the k cluster are denoted by x_{kik} , when $ik = 1, \dots, nk$ and nk represents the number of members of each cluster and $\bar{x}k$ represents the average of the k cluster, then the standard deviation value of the k cluster (Sk) can be calculated using the formula in the following equation:

$$S_k = \sqrt{\frac{1}{n_k-1}\sum_{i_k=1}^{n_k}\left(x_{ki_k} - \bar{x}_k\right)^2} \tag{7}$$

where $\bar{x}_k$ represents the k cluster average $\bar{x}_k = \frac{\sum_{i_k=1}^{n_k} x_{ki_k}}{n_k}$ The calculation of S_b can be formulated as the following equation.

$$S_b = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(\bar{x}_k - \bar{x})^2},$$

where $\bar{x}$ represents the average of the entire cluster $\bar{x} = \frac{\sum_{k=1}^{K}\sum_{i_k=1}^{n_k} x_{ki_k}}{\sum_{k=1}^{K} n_k}$. The calculation of S_b can be formulated as the following equation. The performance of a clustering method is said to be good if the value of the ratio S_w and S_b is smaller, meaning that there is maximum homogeneity within the cluster and maximum heterogeneity between clusters (Bunkers, Miller, & DeGaetano, 1996). Grouping based on the distance concept is considered inappropriate for categorical data, so Sudipto Guha and friends developed a new method, namely robust clustering using linked or ROCK, which uses the link concept in the grouping process. The ROCK method uses the link concept in forming clusters as a measure of similarity (Guha et al, 2000). There are 4 grouping steps using ROCK as follows (Dutta, Mahanta, & Pujari, 2005). First, calculate the similarity between data using the Jaccard coefficient. The measure of similarity between the i data and the j data can be calculated using the formula in the following equation.

$$sim(X_i, X_j) = \frac{|X_i \cap X_j|}{|X_i \cup X_j|}, \tag{9}$$

with X_i: the i set of observations with $X_i = \{x_{i1}, x_{i2}, x_{i3}, \dots, x_{im}\};$ X_j: j set of observations with $X_j = \{x_{j1}, x_{j2}, x_{j3}, \dots, x_{jm}\}$ and $|X|$: Cardinal numbers or the number of members of the set X. Then the similarity between the i data and the j data is arranged into an S matrix as follows.

$$\mathbf{S} = \begin{pmatrix} sim(X_1,X_1) & sim(X_1,X_2) & \cdots & sim(X_1,X_{n-1}) & sim(X_1,X_n) \\ sim(X_2,X_1) & sim(X_2,X_2) & \cdots & sim(X_2,X_{n-1}) & sim(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ sim(X_{n-1},X_1) & sim(X_{n-1},X_2) & \cdots & sim(X_{n-1},X_{n-1}) & sim(X_{n-1},X_n) \\ sim(X_n,X_1) & sim(X_n,X_2) & \cdots & sim(X_n,X_{n-1}) & sim(X_n,X_n) \end{pmatrix}$$

The second step determines the neighbors between data. If the value of $sim(X_i, X_j) \geq \theta$ then $X_i$and $X_j$are said to be neighbors. The θ or threshold value is determined by the researcher to control how close the relationship between objects of observation is, ranging from $0 < \theta < 1$. If X_i and X_j are neighbors, they are labeled 1 and if they are not neighbors, they are labeled 0. The neighbor status labels are arranged in the form of an nxn **T** matrix where n is the amount of data.

$$\mathbf{T} = \begin{pmatrix} T(X_1,X_1) & T(X_1,X_2) & \cdots & T(X_1,X_{n-1}) & T(X_1,X_n) \\ T(X_2,X_1) & T(X_2,X_2) & \cdots & T(X_2,X_{n-1}) & T(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ T(X_{n-1},X_1) & T(X_{n-1},X_2) & \cdots & T(X_{n-1},X_{n-1}) & T(X_{n-1},X_n) \\ T(X_n,X_1) & T(X_n,X_2) & \cdots & T(X_n,X_{n-1}) & T(X_n,X_n) \end{pmatrix}$$

Then calculate the link between X_i and X_j which is denoted by $link(X_i, X_j)$. For $link(X_i, X_j)$ it is calculated by multiplying the neighboring matrices X_i and X_j with itself which is formulated by the following equation.

$$link(X_i, X_j) = \mathbf{T}^2, \tag{10}$$

The bigger the $link(X_i, X_j)$, the greater the possibility that X_i and $X_j$are in the same group.

The third step is to calculate the goodness measure (GM) value. GM is an equation calculated by dividing the number of links by the probability of the link being formed based on the group size (Tyagi & Sharma, 2012). Goodness measure can be calculated using the following formula.

$$(C_i, C_j) = \frac{link\ (C_i, C_j)}{\left[(n_i+n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}\right]} \tag{11}$$

with $link\left(C_i, C_j\right) = \sum_{X_{i\in C_i}, X_{j\in C_j}} link(X_i, X_j)$ which states the number of all possible pairs of data that exist in C_i and C_j, and $f(\theta) = (1-\theta)/(1+\theta)$.

The validity of the performance of grouping results using the ROCK algorithm is calculated from the ratio S_{wc} $(sum\ within\ categorical)$with S_{bc} $(sum\ between\ categorical)$, $\left(\frac{S_{wc}}{S_{bc}}\right)$. . Better grouping performance has the smallest ratio value (Wulandari et al, 2020). The values of S_{wc} and S_{bc} are respectively formulated using equations (12) and (13).

$$S_{wc} = \left(\frac{\sum_{k=1}^{K}\left(\frac{n_{.k}}{2} - \frac{1}{2n_{.k}}\sum_{h=1}^{H} n_{hk}^2\right)}{(n-K)}\right)^{\frac{1}{2}}, \tag{12}$$

$$S_{bc} = \left(\frac{\frac{1}{2}\left(\sum_{k=1}^{K}\frac{1}{n_{\cdot k}}\sum_{h=1}^{H} n_{hk}^2\right) - \frac{1}{2n}\sum_{h=1}^{H} n_{h.}^2}{(K-1)}\right)^{\frac{1}{2}} \tag{13}$$

with n_{hk} representing the number of observations in category h in the group $k,\ h = 1,2,3,\ldots,H;$ $n_{\cdot k} = \sum_{h=1}^{H} n_{hk}$ states the number of observations in the group k; $n_{h\cdot} = \sum_{k=1}^{K} n_{hk}$ states the number of observations in category h.

Cluster ensemble is a method of combining several results from different grouping algorithms so that a combined solution is obtained as the final solution. This grouping is also known as the cluster ensemble based mixed data clustering (algCEBMDC) algorithm. Cluster ensemble combines partitions of different data sets into a final result and this method is superior to single clustering methods (Topchy et al, 2004).

According to (He et al, 2005) the algCEBMDC algorithm is as follows. First, the available data sets are separated into categorical and numerical data. The second step is to group categorical data using a clustering algorithm for categorical data, including the ROCK, K-Modes, K-Prototype algorithms. The third step is to group numerical data using clustering algorithms for numerical data, including hierarchical agglomerative nesting (AGNES), K-Means and K-Medois. The fourth step combines the cluster results from steps 2 and 3. This combination is called an ensemble. The final step, to get the final cluster results, carry out ensemble grouping using grouping methods for categorical data including the ROCK and K-Modes algorithms.

3. Research Methodology

This research uses secondary data, namely SLB data in Riau Province in 2021. Data was taken from the official website of the Ministry of Education and Culture, namely the basic education data website (DAPODIK). The data consists of 47 special schools with categorical variables used, namely school accreditation, school district/city, school status, library availability, lab availability and internet network availability, while the numerical variables used are number of teachers, number of students and number of study groups (groups). The analysis steps carried out in this research are as follows:

1. Collect data from the DAPODIK website then tidy up the data in Excel.
2. Separate data between numerical variable data and categorical variables.
3. Group the data, as follows:

(i) To group numeric type data, an agglomerative nesting algorithm is used with three methods, namely single linkage, complete linkage and average linkage, then the best group is determined by looking at the minimum S_w/S_b.

(ii) In grouping categorical data, the ROCK algorithm is used, then determining the best group by calculating the minimum S_{wc}/S_{bc}.

4. The results from steps 3(i) and 3(ii) are combined and used as categorical data, then a combined clustering (cluster ensemble) is carried out using the ROCK ensemble.
5. Determine the optimum group of the ROCK ensemble method by looking at the minimum S_{wc}/S_{bc} .
6. Describe the characteristics of each cluster formed.

4. Ensemble Rock to Group Slb in Riau Province Based on Infrastructure

In SLB numerical data, namely many teachers, many students, many groups are grouped into 3 to 5 groups using the AGNES method which includes single linkage, complete linkage and average linkage. The first step in this grouping is to assume that each data is a group that has a single member, namely itself. Next, the distance between data is calculated using the Euclidean distance measure as in equation (1). The results of calculating the Euclidean distance from all data combinations are entered in a matrix **D** measuring 47x47.

$$\mathbf{D} = \begin{pmatrix} 0 & 15.03 & 54.03 & 238.34 & \cdots & 6.78 & 36.85 & 91.26 & 101.98 \\ 15.03 & 0 & 41.34 & 224.15 & \cdots & 21.21 & 50.47 & 76.41 & 87.46 \\ 54.85 & 41.34 & 0 & 184.43 & \cdots & 61.13 & 90.33 & 40.46 & 49.73 \\ 238.34 & 224.15 & 184.83 & 0 & \cdots & 371.01 & 270.83 & 151.23 & 139.81 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 6.78 & 21.21 & 61.13 & 371.01 & \cdots & 0 & 32.28 & 97.37 & 108.47 \\ 101.57 & 86.72 & 51.04 & 139.67 & \cdots & 107.9 & 0 & 126 & 135.46 \\ 91.26 & 76.41 & 40.46 & 151.23 & \cdots & 97.37 & 14.45 & 0 & 19.03 \\ 101.98 & 87.46 & 49.73 & 139.81 & \cdots & 108.47 & 135.46 & 19.03 & 0 \end{pmatrix}$$

The distance between the observation object and itself will have a value of 0. Based on matrix **D**, it can be seen that the distance between the 1st and 2nd data is 15.03 as well as the distance between other observation objects.

To determine the optimum group size, look at the Dunn index value.

Tabel 1: Dunn Index AGNES Grouping

Group size	Index Dunn		
	Single Linkage	Complete Linkage	Average Linkage
3	**0.3005292**	**0.1822005**	**0.3005292**
4	0.2678318	0.09570696	0.1302651
5	0.2252422	0.1594249	0.1302651

Based on Table 1, the largest Dunn index value for the single linkage method is located in 3 groups, namely 0.3005292. The complete linkage and average linkage methods have the largest Dunn index in 3 groups, respectively, namely 0.1822005 and 0.3005292. So the optimal number of clusters for each method is 3 groups.

The ratio of S_w to S_b grouping SLB numerical data as in Table 2.

Tabel 2: Ratio of s_w to s_b AGNES Grouping

Method	Group Size	Value S_w	Value S_b	$\frac{s_w}{s_b}$
Single linkage	3	26.3346	20.7558	1.268783
Complete linkage	3	9.538534	20.04942	**0.475751**
Average linkage	3	26.3346	20.7558	1.268783

Based on Table 2, the complete linkage grouping method has the smallest ratio, namely 0.475751. This means that complete linkage with 3 groups is the best grouping in AGNES for SLB cases in Riau Province. Members of the complete linkage method group are presented in Table 3.

Table 3: Complete Linkage Method Group Members

Group	Member
1	Data to- 1, 2 ,5, 6, 7, 9, 10, 11, 13, 14, 15, 16, 18, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 36,37, 40, 41, 42, 43, 44.
2	Data to- 3, 8, 12, 17, 20, 22, 32, 35, 38, 39, 45, 46, 47
3	Data to- 4

Based on Table 3, the characteristics of each group are that group 1 is an SLB which has the least number of teachers, students and study groups compared to the other two groups and has an average number of teachers of 8.67 people, an average number of students of 39.56 people and The average number of study groups is 11,02941 groups. Group 2 is an SLB which has more teachers, students and study groups than group 1 but fewer than group 3 and has an average number of teachers of 17.08 people, an average number of students of 121.33 people and an average number of study groups amounting to 24.17 rombels. Group 3 is the SLB which has the most teachers, students and study groups among the other two groups and has an average number of teachers of 59, an average of 262 students and an average of 55 study groups.

Categorical data, namely accreditation, school status, lab availability, library availability, internet network availability, special school locations are grouped using the ROCK method. At the beginning of grouping using the ROCK method, consider each data as a group that only has its own members. Next, the similarity between the data is calculated using the formula in equation (9). The similarity of the 47 data will be arranged in an S matrix measuring 47x47.

$$\mathbf{S} = \begin{pmatrix} 1.00 & 0.71 & 0.5 & 0.33 & \cdots & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.71 & 1.00 & 0.33 & 0.2 & \cdots & 0.33 & 0.09 & 0.33 & 0.33 \\ 0.5 & 0.33 & 1.00 & 0.33 & \cdots & 0.2 & 0.09 & 0.33 & 0.09 \\ 0.33 & 0.2 & 0.33 & 1.00 & \cdots & 0.09 & 0.5 & 0 & 0.2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0.2 & 0.33 & 0.2 & 0.09 & \cdots & 1.00 & 0.2 & 0.5 & 0.5 \\ 0.2 & 0.09 & 0.09 & 0.5 & \cdots & 0.2 & 1.00 & 0.2 & 0.5 \\ 0.2 & 0.33 & 0.33 & 0 & \cdots & 0.5 & 0.2 & 1.00 & 0.5 \\ 0.2 & 0.33 & 0.09 & 0.2 & \cdots & 0.5 & 0.5 & 0.5 & 1.00 \end{pmatrix}$$

The θ value used in this research is 0.05; 0.07; 0.10; 0.12; 0.15; 0.17; 0.20; 0.22; 0.25; 0.27; 0.30. Based on (*) the neighbor matrix for $\theta = 0.05$ is presented in the following matrix $T_\theta = 0.05$

$$\mathbf{T}_{\theta\,=\,0.05} = \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & \cdots & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & \cdots & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & \cdots & 1 & 1 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 & \cdots & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & \cdots & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & \cdots & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & \cdots & 1 & 1 & 1 & 1 \end{pmatrix}$$

After obtaining the neighborhood status, then $link(X_i, X_j)$ is calculated using equation (10) and goodness measure using equation (11). The greater the $link(X_i, X_j)$ the greater the possibility of observation data in the same group.

The best grouping results using the ROCK method are seen from the smallest $\frac{S_{wc}}{S_{bc}}$.

Table 4: Ratio of S_{wc} to S_{bc} ROCK Grouping

Value θ	$\frac{S_{wc}}{S_{bc}}$
0.05	0.2983234
0.07	0.45961
0.10	0.3487696
0.12	0.5796813
0.15	0.3499557
0.17	0.1078205
0.20	0.1077693
0.22	0.1124353
0.25	**0.09026779**
0.27	0.2722354
0.30	0.1354518

It can be seen from Table 4. The smallest S_{wc}/S_{bc} is $\theta = 0.25$ with a ratio of 0.09026779. So the best grouping method for ROCK for categorical SLB is to use the value $\theta = 0.25$. The ROCK method grouping using $\theta = 0.25$ produces 3 groups, each member of which is presented in Table 5.

Table 5: ROCK Method Group Members $\theta = 0.25$

Group	Member
1	Data to-: 1-3, 5, 6, 8-15, 30.
2	Data to-: 7, 25, 33, 42-44, 46
3	Data to-: 4,16-24, 26-29, 31, 32, 34-41, 45, 47.

Based on Table 5, the characteristics of each group are that group 1 SLB is on average located in the capital city with private status, has laboratory facilities, a library and an adequate internet network. This is directly proportional to fairly good accreditation. In group 2, the average SLB is located outside the capital city with private status, has inadequate facilities. This is directly proportional to the average SLB having poor accreditation. In group 3, SLBs are on average located outside the capital with state status, have laboratory facilities, libraries and good internet networks. This is directly proportional to good SLB accreditation.

After each data type is grouped according to the algorithm, grouping is then carried out for mixed data using the ROCK ensemble. The results of grouping each data type are used as categorical data, then grouping is carried out using the ROCK ensemble method. The mixed data structure that will be analyzed in this research is presented in Table 6.

Table 6: Mixed Ensemble ROCK Data Structure

Data to-	Numerical Data Cluster Results (AGNES)	Categorical Data Cluster Results (ROCK)
1	1	1
2	1	1
3	2	1
⋮	⋮	⋮
46	2	2
47	2	3

The ratio of S_{wc} and S_{bc} for the ROCK ensemble method can be seen in Table 7. Based on Table 7, grouping using θ= 0.10; 0.15; 0.20; 0.22 has the same ratio, namely 0.08177794 which is the smallest ratio, so the best ROCK ensemble grouping is using the value $\theta = 0.22$.

The grouping results of the ROCK ensemble method for mixed data using a value of $\theta = 0.22$ produced 3 groups, each member of which is presented in Table 8.

Table 7: Ratio of S_{wc} to S_{bc} Ensemble ROCK Method

Value θ	$\frac{S_{wc}}{S_{bc}}$
0.05	0.09005792
0.07	0.09005792
0.10	0.08177794
0.15	0.08177794
0.20	0.08177794
0.22	**0.08177794**
0.25	0.1052279
0.27	0.1052279
0.30	0.1052279

Table 8: ROCK Ensemble Group Members $\theta = 0.22$

Group	Member
1	Data to-: 1, 2, 3, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 17, 20, 22, 30, 32, 35, 38, 39, 45, 46, 47.
2	Data to-: 4, 16, 18, 19, 21, 23, 24, 26, 27, 28, 29, 31, 34, 36, 37, 40, 41.
3	Data to-: 7, 25, 33, 42, 43, 44.

Based on Table 8, the characteristics of the grouping results of the ROCK ensemble method are that group 1 is a high quality SLB with the following characteristics, private and public status, but the dominant private sector is located outside or inside the capital city of Riau Province. Facilities such as labs, libraries and internet networks are adequate, this has an impact on the fact that this SLB has a large number of teachers, students and study groups and has many accreditations, the rest have B and C accreditations. Group 2 is a medium quality SLB with the following characteristics, status public and private, but the majority are public with the average location being outside the capital city of Riau province. Facilities such as labs, libraries and internet networks are quite adequate, this has an impact on the fact that the SLBs in this group have quite a lot of teachers, students and study groups and the accreditation is quite good, only 1 SLB has a C accreditation, the rest have A and B accreditations. Group 3 is a low quality SLB with the following characteristics, private status and located outside the capital city of Riau province. Facilities such as labs, libraries and internet access are not available, this has an impact on the fact that this SLB has few teachers, students and study groups and low accreditation, namely B, C and even D.

5. Conclusion

Based on the results of the grouping that has been carried out, it can be concluded that grouping the SLB data using the ROCK ensemble method with θ=0.22 produces 3 groups of SLB. Group 1 is a high quality SLB, group 2 is a medium quality SLB and group 3 is a low quality SLB.

References

Alvionita, Sutikno, & Suharsono, A. (2017). Ensemble ROCK methods and ensemble SWFM methods for clustering of cross citrus accessions based on mixed numerical and categorical dataset. *In IOP Conf*, *58*(1), 1-10.

Bunkers, M. J., Miller, J. R., & DeGaetano, A. T. (1996). Definition of climate regions in the northern plains using an objective cluster modification technique. *Journal of Climate*, *9*(1), 130–146.

Pelaez, K. (2019). Latent class analysis and random forest ensemble to identify at-risk students in higher education. Journal of Educational Data Mining, 11.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, *4*(1), 95–104.

Dutta, M., Mahanta, A. K., & Pujari, A. K. (2005). QROCK: A quick version of the ROCK algorithm for clustering of categorical data. *Pattern Recognition Letters*, *26*(15), 2364–2373.

Guha, S., Rastogi, R., & Shim, K. (2000). Rock: a robust clustering algorithm for categorical attributes. *Information Systems*, *25*(5), 345–366.

Hair, J. F. J., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis (7th Edition)*. New Jersey: Prentice Hall inc.

He, Z., Xu, X. i, & Deng, S. (2005a). A cluster ensemble method for clustering categorical data. *Information Fusion*, *6*(1), 143–151.

He, Z., Xu, X. i, & Deng, S. (2005b). Clustering mixed numeric and categorical data: A cluster ensemble approach. *High Technology Letters*, *1*(1), 1–14.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6 th). New Jersey: Person Prentice Hall.

Khairati, A. F., Adlina, A. A., Hertono, G. F., & Handari, B. D. (2019). Study of the validity index of the K-Means enhanced and K-Means MMCA algorithms. *PRISMA Prosiding Seminar Nasional Matematika*, *2*(1), 161–170.

Pramartha, I. N. B. (2015). History and education system of special schools in the state of Denpasar, Bali. *Historia*, *3*(2), 13–14.

Topchy, A., Jain, A. K., & Punch, W. (2004). A mixture model for clustering ensembles. *SIAM Proceedings Series*, *1*(1), 379–390.

Tyagi, A., & Sharma, S. (2012). Implementation of ROCK clustering algorithm for the optimization of query searching time. *International Journal on Computer Science and Engineering (IJCSE)*, *4*(05), 809–815.

Wulandari, L., Farida, Y., Fanani, A., Ulinnuha, N., & Using, T. (2020). Evaluation of disadvantaged regions in East Java based-on the 33 indicators of the ministry of villages , development of disadvantaged regions , and transmigration using the Ensemble ROCK (Robust Clustering Using Link) method. *ASTES*, *5*(5), 193–200.