ICIC Express Letters
Volume **14**, Number **12**, December **2020**

ICIC International ©2020 ISSN 1881-803X
pp. 1225–1234

CLG CLUSTERING FOR MAPPING PATTERN ANALYSIS OF STUDENT ACADEMIC ACHIEVEMENT

Agung Triayudi[1], Wahyu Oktri Widyarto[2] and Vidila Rosalina[3]

[1]Informatics Department
Universitas Nasional
Jalan Sawo Manila, Pejaten Barat, Pasar Minggu, Jakarta 12520, Indonesia
agungtriayudi@civitas.unas.ac.id

[2]Industrial Engineering Department
[3]Informatics Department
Universitas Serang Raya
Jl. Raya Cilegon No. Km. 5, Taman, Drangong, Kota Serang, Banten 42314, Indonesia
{ woktri; vidila.suhendarsah }@gmail.com

Received March 2020; accepted June 2020

Abstract. *This paper will study student behavior patterns in online learning activities, determining student behavior patterns is generally difficult to assess because learning is not done face-to-face. This test included 815 students and selected 25 students to represent each meeting after being calculated using a random sampling sample using the Slovin formula. By looking at the activities carried out during an online meeting, the data will be separated every week and the data that will be used includes the attributes of the number of posts, meeting weeks, gender, region, age duration that will be used approximately 30 days so that the development of each week can be seen from the dendrogram generated using CLG clustering method. And the results displayed from the 1st week dendrogram are with a total post 1990 and the selected cluster is 2, with percentage (37.34%) for the first cluster and the second cluster (62.66%). For the second week dendogram the results of the post were 2223 for the first cluster received (22.45%) and for the second cluster (77.55%). Week 3 gets results (58.63%) for cluster 1 and cluster 2 (41.37%). And for the 4th week, it gets results (31.64%) for the second cluster (68.36%).*

Keywords: Student, Slovin formula, CLG clustering, Dendrogram, Result

1. **Introduction.** Smart learning systems provide relevant learning resources as pedagogical needs for students and individual preferences. One example is the Adaptive Recommendation based on Online Learning Style (AROLS), which implements adaptation of learning resources in accordance with students' online learning styles. Where the method used is Collaborative Filtering (CF) to extract the preferences and behavior patterns of each cluster [1].

Other studies [2,3,11] recommend the Single Linkage dissimilarity increment distribution method, Global cumulative score standard (SLG), and Average Linkage dissimilarity increment distribution, Global cumulative score standard (ALG) which are used to analyze student learning online interaction data. The end result is a grouping model of behavior patterns and interpersonality patterns of students.

Online learning is increasingly important for higher education institutions. However, to date the guidelines on how to best develop and implement online learning activities are still very limited [4,14]. How to connect this online learning with the selection of the right technology to support learning, and, the best way to adjust individual assessments to improve student learning experiences are discussed [5,15]. In many cases, many highlight the importance of interdependence between interaction, communication and motivation

DOI: 10.24507/icicel.14.12.1225

on student learning experiences. It identifies student-related factors and their influence on online learning activities. This study proposes an evaluation of multivariate learning models to assess student learning in an online learning environment. Data flow is divided into four categories, namely: tutoring, understanding innovation, interactive and supporting learning [6,16].

The large amount of educational data created by students who interact with digital learning devices opens up opportunities to gain insights in improving educational models [7,17]. Online learning has developed rapidly in the world. Online learning fully supports multimedia presentations and provides effective interactions [8,18]. In addition, online e-learning not only provides student competency, but also creates an independent and collaborative learning process. However, online learning must have wise management and good communication networks, if that is not done then it will have a negative impact on the students themselves [9,19,20].

The development of the Dissimilarity Increment Distribution (DID) method, namely SLG and ALG, after analysis is known to have weaknesses in measuring the distance between cluster trees in sub-clusters and main clusters, when tested in calculations for several datasets having different validation results, the characteristics of the dataset high order seen no problems because of the high validation value, but when tested on a low order dataset the validation value becomes unstable, in some cases this occurs because of the resulting gap between groups that are not included in the same cluster. This will result in limited flexibility due to joint rejection involving cluster isolation.

The main contribution of this paper is how the development of the DID method can be used in the characteristics of low-order datasets, the solution offered is to use the CLG (Complete Linkage dissimilarity increment distribution – Global cumulative score standard) method. The basic use of the complete linkage method is to be able to work better at low order dataset.

From previous studies, it was found that the limitations of previous studies have not examined how external factors can affect the level of academic achievement of online learning students. The theme of this online learning student pattern is interesting to study. Of the several methods used previously the clustering algorithm is still dominant and most suitable for use. However, the clustering algorithm also leaves the problem of limited cluster flexibility, and it is due to the choice of different dataset characteristics. From the problems presented in this paper we will propose a modified clustering method that is suitable with the characteristics of the dataset used.

2. Research Method.

In this study, a proposed modification method in the clustering algorithm is CLG (Complete Linkage dissimilarity increment distribution – Global cumulative score standard), and this algorithm is a combined algorithm between the CL (Complete Linkage) algorithm [10], the DID algorithm (Dissimilarity Increment Distribution) [11], GCSS (Global Cumulative Score Standard) algorithm [12]. The CLG algorithm works by combining elements of free graph-based parameters and model-based approaches (which are defined by combining criteria by characterizing clusters in probabilistic terms) for grouping.

$$CL = \max\{D(C_k, C_i), D(C_k, C_j)\} \tag{1}$$

$$\begin{aligned} DID = pdissinc(w;\lambda) &= \frac{\pi\beta^2}{4\lambda^2} w \exp\left(-\frac{\pi\beta^2}{4\lambda^2} w^2\right) \\ &+ \frac{\pi^2\beta^3}{8\sqrt{2}\lambda^3} X \left(\frac{4\lambda^2}{\pi\beta^2} - w^2\right) \exp\left(-\frac{\pi\beta^2}{8\lambda^2} w^2\right) erfc\left(\frac{\sqrt{\pi}\beta}{2\sqrt{2}\lambda} w\right) \end{aligned} \tag{2}$$

$$\begin{aligned} GCSS &= gcss_{th}(C_k, C_i, C_j, Y_{MIN}) = gcss_{th}(css_k, N_i, \gamma_i, \mu_i, \sigma_i, N_j, \gamma_j, \mu_j, \sigma_j, Y_{MIN}) \\ &= css_k \Upsilon(N_i, N_j) \Psi_G(N_i, \gamma_i, \mu_i, \sigma_i, N_j, \gamma_j, \mu_j, \sigma_j, Y_{MIN}) \end{aligned} \tag{3}$$

Algorithm 1. CLG Algorithm

```
Input: dataset $X$
procedure
$M_p : M_p(i, j)$
Select the most similar clusters $(C_i, C_j) \max Dist = \max\{d(x_i, x_j) : x_i \in C_i, x_j \in C_j\}$
if $|C_i| < 6$ and $|C_j| < 6$ then
        Merge clusters $C_i, C_j$ into a new cluster $C_b$ using CLDID (Equations (1) and (2))
        and GCSS (Equation (3))
end if
if $|C_i| \geq 6$ and $|C_j| < 6$ then
        if $dissinc(x_i, x_j, x_k) = |d(x_i, x_j) - d(x_j, x_k)|$ of $(C_j)$ is not in the tail then
        the $pdissinc(w; \lambda)$ (Equation (2)) then $dissinc(x_i, x_j, x_k) = |d(x_i, x_j) - d(x_j, x_k)|$
        of $(C_i)$ then
        Merge clusters $C_i, C_j$ into a new cluster $C_b$ using CLDID (Equations (1) and (2))
        and GCSS (Equation (3))
        else
        Do not merge $C_i, C_j$
        end
end
```

The CLG algorithm provides different treatment to small cluster candidate groups. Each candidate group whose size is lower than Y_{MIN} is not required to explain the merging criteria. In fact, the merger between C_i and C_j always occurs in the case of the two groups of candidates less than the value of the Y_{MIN} object. Regarding the cluster size threshold, it is important to note the difference between the H and Y_{MIN} parameters; because both values refer to group size, parameter H is the real value used in the calculation of the dynamic merge threshold, while Y_{MIN} is the integer threshold value used when directing the comparison with the required cluster size.

To calculate the amount of data to be tested will use the simple random sampling method with Slovin formula as simple random sampling [13]:

$$n = \frac{N}{N(e)^2 + 1} \tag{4}$$

n = sample, N = population, e = 80% precision value or sig = 0.2.

Figure 1 shows the methodology flow in this study, which illustrates the whole process starting from the background, design of the algorithm, implementation and construction, experimentation and analysis, result. The dataset from this study was taken from the University of Indonesia ODL (Open and Distance Learning), and the dataset taken relates to online learning student data which includes: gender, age, region. In the process of labeling the data to be processed, gender is labeled 1 male and 2 females, age will use 3 categories, namely age 15-18 labels 1, 19-21 labels 2, and > 21 labels 3, Jakarta area labels 1, Bogor labels 2, Depok labels 3, Tangerang labels 4, and Bekasi labels 5.

3. **Evaluation of Clustering Result.** This research method will use data that is data that follows online class lessons at the ODL (Open and Distance Learning) University in Indonesia with the number: 815 student data according to online results in November 2019 and which is tested using 25 students to represent the overall data obtained from the calculation of the formula Slovin simple random sampling.

From the results of the dendrogram, there are 2 clusters selected. First week total posts were 1,990 and student behavior patterns based on male gender attributes were 37.34% and women were 0% for the first cluster. Whereas the second cluster of men got 48.09% and 14.57% for women. Based on the area in the first cluster, 37.34% results were obtained for students domiciled in Jakarta while other regions did not post results, while

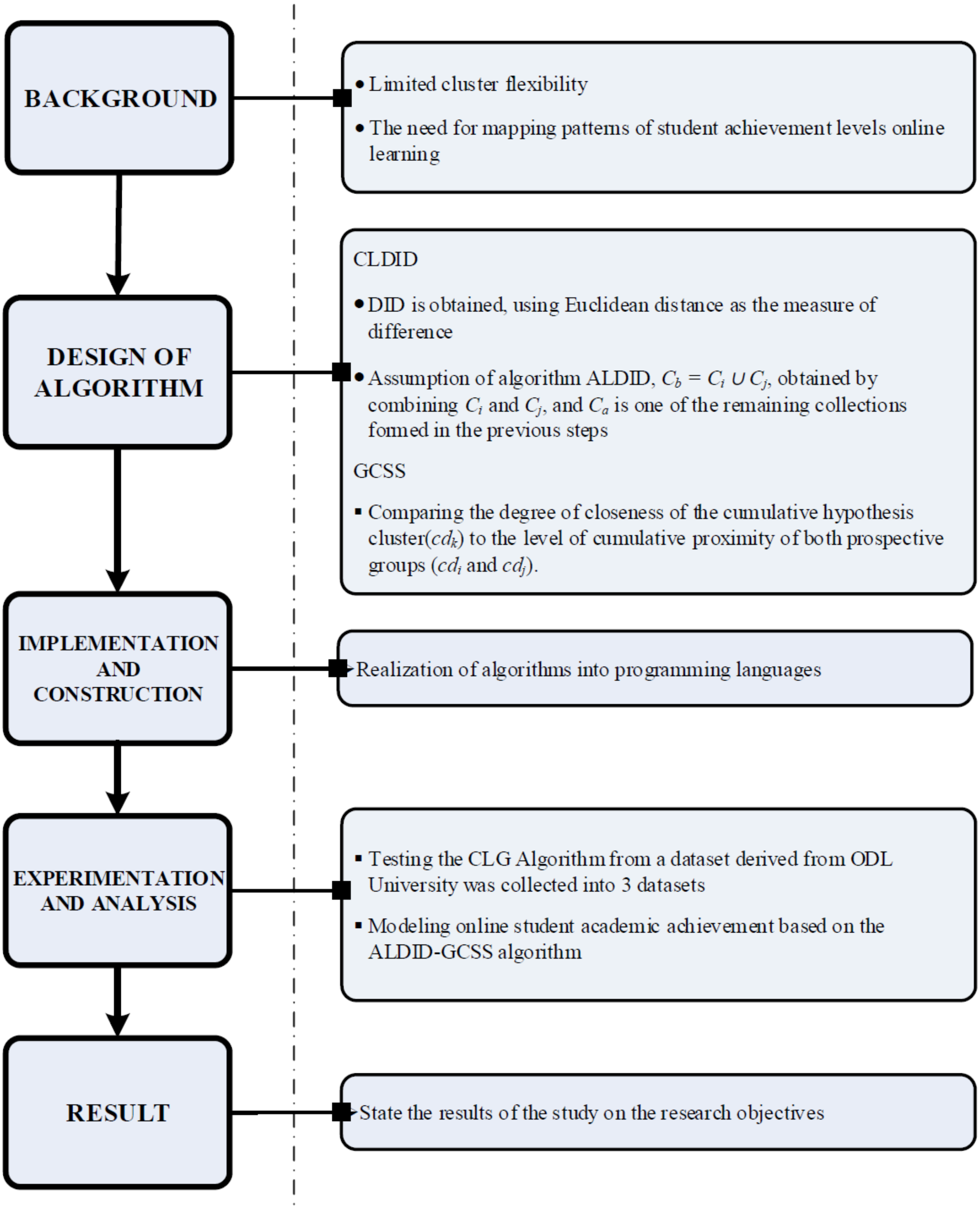
BACKGROUND
• Limited cluster flexibility
• The need for mapping patterns of student achievement levels online learning
DESIGN OF ALGORITHM
CLDID
• DID is obtained, using Euclidean distance as the measure of difference
• Assumption of algorithm ALDID, $C_b = C_i \cup C_j$, obtained by combining C_i and C_j, and C_a is one of the remaining collections formed in the previous steps
GCSS
▪ Comparing the degree of closeness of the cumulative hypothesis cluster(cd_k) to the level of cumulative proximity of both prospective groups (cd_i and cd_j).
IMPLEMENTATION AND CONSTRUCTION
Realization of algorithms into programming languages
EXPERIMENTATION AND ANALYSIS
▪ Testing the CLG Algorithm from a dataset derived from ODL University was collected into 3 datasets
▪ Modeling online student academic achievement based on the ALDID-GCSS algorithm
RESULT
State the results of the study on the research objectives

Figure 1. Research methodology

for the second cluster (Jakarta 46.28%) (Bogor 6.83%) (Depok 4.37%) (Bekasi 5.18%). Based on age the highest results obtained by the second cluster get 32.61% of the total posts for age category 3. So the student behavior patterns obtained illustrate that men are more active compared to women, for the Jakarta area more active than bodetabek areas while the more active age is age > 21 years compared to ages 20 and under.

For the second week dendogram the results of the post were 2223 for the first cluster received (22.45%) and for the second cluster (77.55%), based on gender the men got 22.41% and the women 0.04% in the first cluster, in the second cluster the men got 55.51% and women 22.04%. Based on Jakarta superior region it is with 11.29% results for

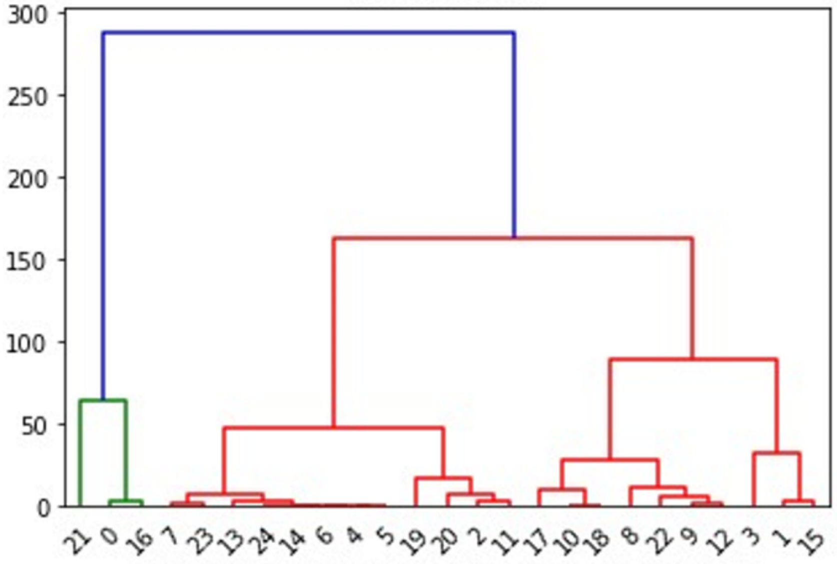

Figure 2. Cluster in the first week

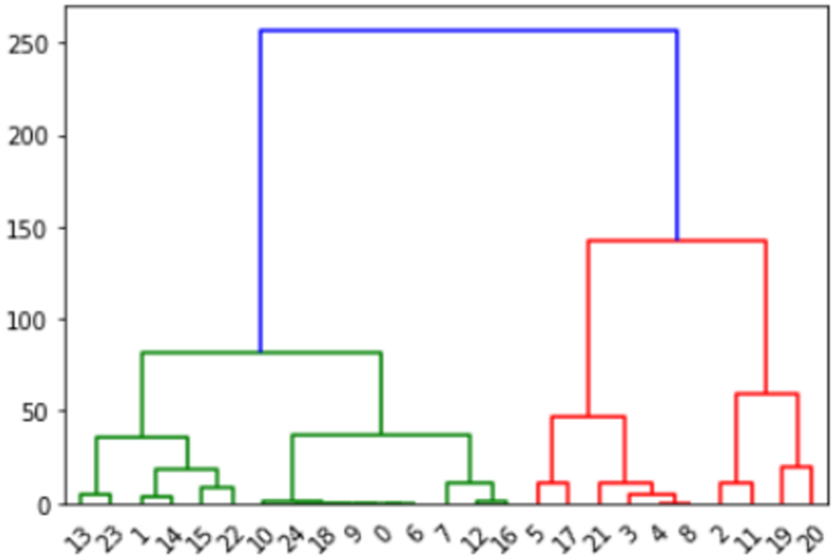

Figure 3. Cluster in the second week

the first cluster and 36.48% in the second cluster. Based on the second week age obtained > 21 years of age get the highest results in the first cluster with 13.36% results and the second cluster gets 46.56% results for the age of 19-21 years.

For the third week dendogram the results of the post were 1866 for the first cluster received (58.63%) and for the second cluster (41.37%), and based on superior male sex attributes with 44.43% results in the first cluster and the second cluster got 30.17% results. And based on the Jakarta area it still gets the highest results 51.61% for the first cluster and the second cluster 35.64%. And the highest yield age attribute obtained 31.99% for ages 19-21 years and the second cluster received 23.90% for ages > 21 years.

For the fourth week dendogram the results of the post were 1735 for the first cluster received (31.64%) and for the second cluster (68.36%). Based on the sex attributes the highest results were obtained by men with the results of 53.08% for the second cluster and 27.49% for the first cluster. The Jakarta area is still superior with a value of 22.31% for the first cluster and 22.94% for the second cluster. Based on the age of the total posts produced more students aged > 21 years in the second cluster with a result of 62.65%.

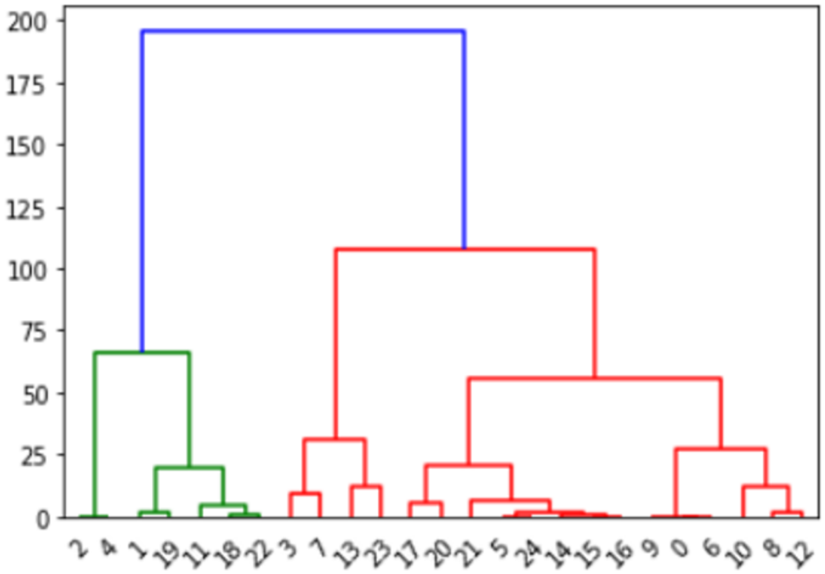

Figure 4. Cluster in the third week

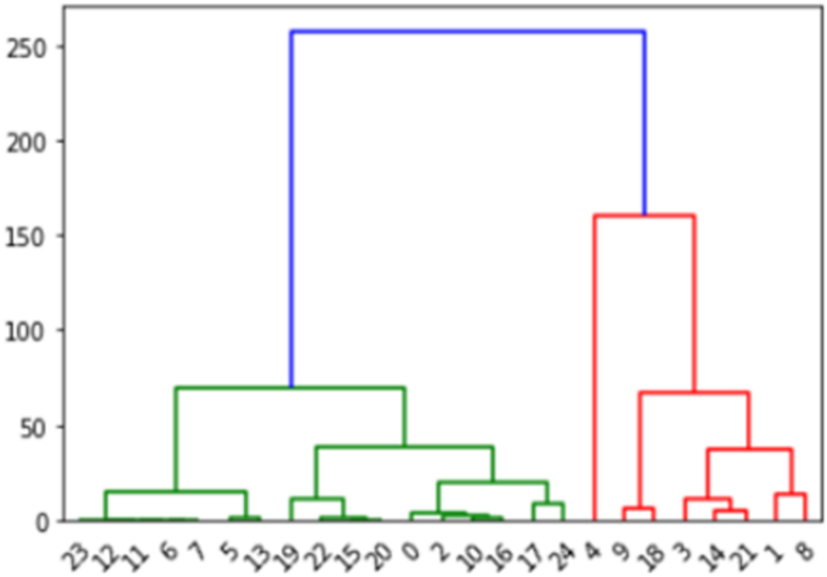

Figure 5. Cluster in the fourth week

From the 4-week dendrogram above during the online meeting the recapitulation is based on 3 attributes used as follows.

Table 1 summarizes the results of the dendrogram using data at the 1st week online meeting based on the attributes used. From the total value of the two clusters, the highest results obtained by the second cluster were 62.66%.

Table 2 summarizes the results of the dendrogram using data at the 2nd week online meeting based on the attributes used. From the total value of the two clusters, the highest result was obtained by the second cluster with a value of 77.55%.

Table 3 summarizes the results of the dendrogram using data at the 3rd week online meeting based on the attributes used. From the total value of the two clusters, the highest result was obtained by the first cluster with a value of 58.63%.

Table 4 summarizes the results of the dendrogram using data at the 4th week online meeting based on the attributes used. From the total value of the two clusters, the highest result was obtained by the second cluster with a value of 68.36%.

Table 1. The results of the percentage of the week meeting week 1

Gender	Diagram			
	Total posting	%	Total posting	%
Male	743	37.34%	957	48.09%
Female	0	0.00%	290	14.57%
Total	743	37.34%	1247	62.66%
Region	Diagram			
	Total posting	PCT	Total posting	PCT
Jakarta	743	37.34%	921	46.28%
Bogor	0	0.00%	136	6.83%
Depok	0	0.00%	87	4.37%
Tangerang	0	0.00%	0	0.00%
Bekasi	0	0.00%	103	5.18%
Total	743	37.34%	1247	62.66%
Age	Diagram			
	Total posting	PCT	Total posting	PCT
1	0	0.00%	0	0.00%
2	226	11.36%	598	30.05%
3	517	25.98%	649	32.61%
Total	743	37.34%	1247	62.66%

Table 2. The results of the percentage of the week meeting week 2

Gender	Diagram			
	Total posting	%	Total posting	%
Male	498	22.41%	1234	55.51%
Female	1	0.04%	490	22.04%
Total	499	22.45%	1724	77.55%
Region	Diagram			
	Total posting	PCT	Total posting	PCT
Jakarta	251	11.29%	811	36.48%
Bogor	65	2.92%	664	29.87%
Depok	183	8.23%	249	11.20%
Tangerang	0	0.00%	0	0.00%
Bekasi	0	0.00%	0	0.00%
Total	499	22.45%	1724	77.55%
Age	Diagram			
	Total posting	PCT	Total posting	PCT
1	0	0.00%	152	6.83%
2	202	9.09%	1035	46.56%
3	297	13.36%	537	24.16%
Total	499	22.45%	1724	77.55%

Table 3. The results of the percentage of the week meeting week 3

Gender	Diagram			
	Total posting	%	Total posting	%
Male	829	44.43%	563	30.17%
Female	265	14.20%	209	11.20%
Total	1094	58.63%	772	41.37%
Region	Diagram			
	Total posting	PCT	Total posting	PCT
Jakarta	963	51.61%	665	35.64%
Bogor	0	0.00%	0	0.00%
Depok	0	0.00%	78	4.18%
Tangerang	0	0.00%	1	0.05%
Bekasi	131	7.02%	28	1.50%
Total	1094	58.63%	772	41.37%
Age	Diagram			
	Total posting	PCT	Total posting	PCT
1	151	8.10%	0	0.00%
2	597	31.99%	326	17.47%
3	346	18.54%	446	23.90%
Total	1094	58.63%	772	41.37%

Table 4. The results of the percentage of the week meeting week 4

Gender	Diagram			
	Total posting	%	Total posting	%
Male	477	27.49%	921	53.09%
Female	72	4.15%	265	15.27%
Total	549	31.64%	1186	68.36%
Region	Diagram			
	Total posting	PCT	Total posting	PCT
Jakarta	387	22.30%	398	22.94%
Bogor	0	0.00%	129	7.44%
Depok	52	3.00%	265	15.27%
Tangerang	0	0.00%	0	0.00%
Bekasi	110	6.34%	394	22.71%
Total	549	31.64%	1186	68.36%
Age	Diagram			
	Total posting	PCT	Total posting	PCT
1	71	4.09%	0	0.00%
2	234	13.49%	99	5.71%
3	244	14.06%	1087	62.65%
Total	549	31.64%	1186	68.36%

4. **Conclusions.** From the results of the discussion it can be concluded as follows: From the dendrogram generated, from the student id data it can be calculated the value of the presentation to assess student behavior during the online meeting which takes place based on the attributes used. The purpose of this research is that the results of the second week get the highest value from weeks 1, 3 and 4 with the number of posts 2,223 with each attribute of men getting 55.51% and women getting 22.04% in the second cluster. Then based on the Jakarta area the highest yield was 36.48% compared to the 4 other regions. And for the age group aged 19-21, 46.56% of the second cluster was selected. So the results obtained that men in the Jakarta area between the ages of 19-21 are more active in carrying out online meetings. Suggestions for future research can improve data results better and use other algorithms in the use of calculations and attributes that are tested.

Acknowledgment. This research is the result of the basic research scheme of the Indonesian Dikti grant B/87/E3/RA.00/2020.

REFERENCES

[1] H. Chen, C. Yin, R. Li, W. Rong, Z. Xiong and B. David, Enhanced learning resource recommendation based on online learning style model, *Tsinghua Science and Technology*, vol.25, no.3, pp.348-356, 2019.

[2] A. Triayudi and I. Fitri, ALG clustering to analyze the behavioural patterns of online learning students, *Journal of Theoretical & Applied Information Technology*, vol.96, no.16, pp.5327-5337, 2018.

[3] A. Triayudi and I. Fitri, A new agglomerative hierarchical clustering to model student activity in online learning, *Telkomnika*, vol.17, no.3, pp.1226-1235, 2019.

[4] S. Sakulwichitsintu, D. Colbeck, L. Ellis and P. Turner, A peer learning framework for enhancing students' learning experiences in online environments, *IEEE the 18th International Conference on Advanced Learning Technologies (ICALT)*, pp.168-169, 2018.

[5] R. L. Sie, J. Delahunty, K. Bell, A. Percy, B. Rienties, T. Cao and M. de Laat, Artificial intelligence to enhance learning design in UOW online, a unified approach to fully online learning, *IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, pp.761-767, 2018.

[6] Q. Hu, Y. Huang and L. Deng, A multivariate learning evaluation model for programming course in online learning environment, *The 14th International Conference on Computer Science & Education (ICCSE)*, pp.737-740, 2019.

[7] D. Ifenthaler, D. C. Gibson and L. Zheng, The dynamics of learning engagement in challenge-based online learning, *IEEE the 18th International Conference on Advanced Learning Technologies (ICALT)*, pp.178-182, 2018.

[8] A. Anggrawan and Q. S. Jihadil, Comparative analysis of online e-learning and face to face learning: An experimental study, *The 3rd International Conference on Informatics and Computing (ICIC)*, pp.1-4, 2018.

[9] S. Mu, M. Cui, X. J. Wang, J. X. Qiao and D. M. Tang, Learners' attention preferences and learning paths on online learning content: An empirical study based on eye movement, *The 7th International Conference of Educational Innovation through Technology (EITT)*, pp.32-35, 2018.

[10] Vijaya, S. Sharma and N. Batra, Comparative study of single linkage, complete linkage, and ward method of agglomerative clustering, *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India, pp.568-573, 2019.

[11] A. Triayudi and I. Fitri, Comparison of parameter-free agglomerative hierarchical clustering methods, *ICIC Express Letters*, vol.12, no.10, pp.973-980, 2018.

[12] G. C. Rodríguez, *Parameter-Free Agglomerative Hierarchical Clustering to Model Learners' Activity in Online Discussion Forums*, Ph.D. Thesis, Universitat Oberta de Catalunya, 2014.

[13] A. D. Herlambang, Y. T. Mursityo, M. C. Saputra and L. Novianti, Criteria-based evaluation of academic information system usage at Brawijaya University based on modified technology acceptance model (TAM), *International Conference on Sustainable Information Engineering and Technology (SIET)*, Malang, Indonesia, pp.272-277, 2018.

[14] T. Ahmad and M. N. Aziz, Data preprocessing and feature selection for machine learning intrusion detection systems, *ICIC Express Letters*, vol.13, no.2, pp.93-101, 2019.

[15] F. E. Gunawan, A learning recommendation to improve electronic textbook learning experience, *ICIC Express Letters*, vol.13, no.1, pp.1-10, 2019.

[16] A. A. Saa, Educational data mining & students' performance prediction, *International Journal of Advanced Computer Science and Applications*, vol.7, no.5, pp.212-220, 2016.

[17] R. Asif, A. Merceron, S. A. Ali and N. G. Haider, Analyzing undergraduate students' performance using educational data mining, *Computers & Education*, vol.113, pp.177-194, 2017.

[18] A. Dutt, M. A. Ismail and T. Herawan, A systematic review on educational data mining, *IEEE Access*, vol.5, pp.15991-16005, 2017.

[19] B. Bakhshinategh, O. R. Zaiane, S. ElAtia and D. Ipperciel, Educational data mining applications and tasks: A survey of the last 10 years, *Education and Information Technologies*, vol.23, no.1, pp.537-553, 2018.

[20] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo and J. Rego, Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses, *Computers in Human Behavior*, vol.73, pp.247-256, 2017.