ICIC Express Letters
Volume **17**, Number **6**, June **2023**

ICIC International ©2023 ISSN 1881-803X
pp. **615–624**

EDUCATIONAL DATA MINING IN CLUSTERIZATION WITH DATA ON STUDENT HABITS IN ONLINE LEARNING

Agung Triayudi* and Iskandar Fitri

Department of Information Technology
Universitas Nasional
Jl. Sawo Manila, RT.14/RW.3, Ps. Minggu, Kec. Ps. Minggu, Kota Jakarta Selatan
Jakarta 12540, Indonesia
iskandar.fitri@civitas.unas.ac.id
*Corresponding author: agungtriayudi@civitas.unas.ac.id

Received July 2022; accepted October 2022

Abstract. *As technology develops in the world, technology also develops in various sectors such as agriculture and government. The education sector is no exception. One of the technological developments in the education sector is the application of Blended Learning (BL). BL is a learning method that combines face-to-face meetings with online material in a harmonious and mutually integrated manner. There are 5 main keys in implementing this BL learning, namely live events, self-paced learning, collaboration, assessment, and performance support materials. To identify what BL actually looks like, we can use a data-driven approach. One of the clustering approaches of Educational Data Mining (EDM) is the latent class analysis method. With this method, we can extract common activity features from hundreds of courses at universities by using a dataset of student behavior in these online classes. For that, the authors will divide the existing types into 4 clusters. With the existing cluster division, decision makers at the university can take actions in accordance with the data generated in this paper.*
Keywords: Blended learning, Clustering, Online course, EDM, Analysis

1. **Introduction.** The use of Learning Management Systems (LMS) in online learning is currently evenly distributed throughout the world, as an effect of the pandemic [1-3]. With online learning, it can help anyone to learn regardless of time and place. However, some students still need face-to-face meetings in class to discuss and complete the learning process that has been carried out at LMS [4,5]. It is called blended learning. Blended learning is a learning paradigm that can be applied in the use of technology-based media. Blended learning is also defined as a combination of online learning and multimedia technologies, such as video streaming, virtual classes, online text animations combined with conventional forms of classroom instruction. Blended learning can also be interpreted more simply as learning that combines online learning with face-to-face [6,7]. Universities in Indonesia have started to apply the blended learning model into their learning activities. Usually, colleges apply this blended learning using Moodle LMS media [8-10].

In the application of blended learning, many education actors feel comfortable because they experience a very personal and interesting learning experience. One example is the Learning Management Systems (LMS) as a learning environment, which features real-time feedback about student learning growth through the existing dashboard [7,11,12]. The college in this case only provides material to the LMS using only primary characteristics such as uploading their syllabus and lecture notes. This results in a lack of understanding regarding which instructional adopt blended learning correctly [12-14]. In answer to these problems, many study attempts have been made to develop blended learning. However, this is still lacking because there are still a lot of dependencies on traditional perceptions

DOI: 10.24507/icicel.17.06.615

that hinder evaluation to be more methodical and synchronous from current sets. In this study, we will apply the latent class analysis method to cluster each individual from various groups into more petite and relatively homogeneous subgroups [15-17].

Although in theory blended learning can be said to be simple, its implementation is quite complicated and challenging because the existing designs are almost not limited to the number of online instructions that are inherent in the blended learning. One of the studies [16] that has been conducted shows during the last few decades, approximately 41% of research on blended learning [18]. The taxonomy of learning experiences adopted in this study suggests 4 main varieties of learning. In other studies, as well, there have been findings regarding new pedagogical methods. This pedagogical method uses non-real-time (asynchronous) material exposure videos and problem-solving disciplines as a study for group-based students who are active in the course [13,19,20].

One of the researchers explained the spectrum of blended learning appropriation with 3 levels. In addition, other researchers also included three modes of e-learning in their involvement with blended learning. The three modes are 1) basic course management and support, 2) mixed learning which leads to meaningful improvements to the lecture process, and 3) a mixture of the two modes to a personalized level of teaching through various existing courses. In connection with these two categories, there is also a categorization based on the adoption phase which is based on the theory diffusion of innovation. In this theory, the entanglements show that higher education administration must be more aware of several aspects from the aspects of strategy, structure, and support [15,21]. Although the implementation of the terminology is inconsistent with the identification of blended learning between changes in F2F timing by online instruction and additional actors in online instruction can hinder further discussion, it can be seen that most methods have identified different types of blending [12,13].

Several studies have concentrated on the status of technology appropriation in an individual college, with a try to identify how several LMSs are applied in the habit. The research was conducted by identifying the perceptions and experiences of students at a university in Portugal. The results of this study indicate the level of basic utilization includes dominant perceptions in Mode 1 [24], and a few examples in Mode 2 that have previously been shown in other studies [12,14,22].

Academic analytics was originally introduced by [18]. Furthermore, academic analytics is experiencing developments. It was developed to be the newest method in increasing the accountability of higher education and to develop intelligence that can increase the success rate of students in learning. One of the implementations of academic analytics that has been well implemented is The Early Warning Systems developed by Purdue University, namely Course Signal. This signal works by retrieving student data from the Students Information System and Course Management Systems. After the data is collected, the data cleansing and transformation stages are carried out, and then the algorithm determines the risk level of each student in the college with indicators, namely red, yellow, and green. As a result, it can be determined that the administrator can consider academic analytics as a solution that can be used as a parameter of student success in learning and training students to use campus facilities. In making the Course Signal, academic analysis, educational data mining techniques, and audit course management systems can be used to design new courses. With this in mind, researchers specifically distinguish between academic analytics and educational data mining [16].

From several previous studies discussed, there are main problems, namely online learning which is commonly used today, has several shortcomings, among others, differences in student behavior when meeting face to face and when learning online, from a lot of data knowledge can be taken about student behavior to improvements and learning methods in the future. One of the Educational Data Mining (EDM) clustering approaches is the latent class analysis method. With this method, we can extract common activity features from hundreds of courses at the university by using the dataset of student behavior in these online classes. With the existing cluster division, university decision makers can take action according to the data generated in this paper.

2. **Research Method.** The dataset used in this study is data taken from a private university in Jakarta, Indonesia. In 2020, this private tertiary institution is expected to have around 12367 students (Diploma to Bachelor level). According to the 2020 annual report, this college has 2759 classes of 3 sizes, namely small class (less than 30 students), middle class (less than 60 students), and large class (more than 60 students). For small classes, there are 1020 small classes (37% of total class), 1352 middle class (49% of total class), and 387 large classes (14% of total class). Referring to the blended learning adoption framework initiated by [22], this college is at level 2 or the early adoption and implementation stages. With Moodle, teachers can also manage a variety of taught classes, discussion forums for each class, and support for project-based and collaborative teaching methods. According to [23], Moodle contains three types of data elements, namely 1) determined by a table to describe each existing object, 2) interaction with learning objects, and 3) association table to describe relationships between learning objects. In addition, Moodle also has a feature that can store all user logs including timestamps and types of interactions performed by the user while in the system.

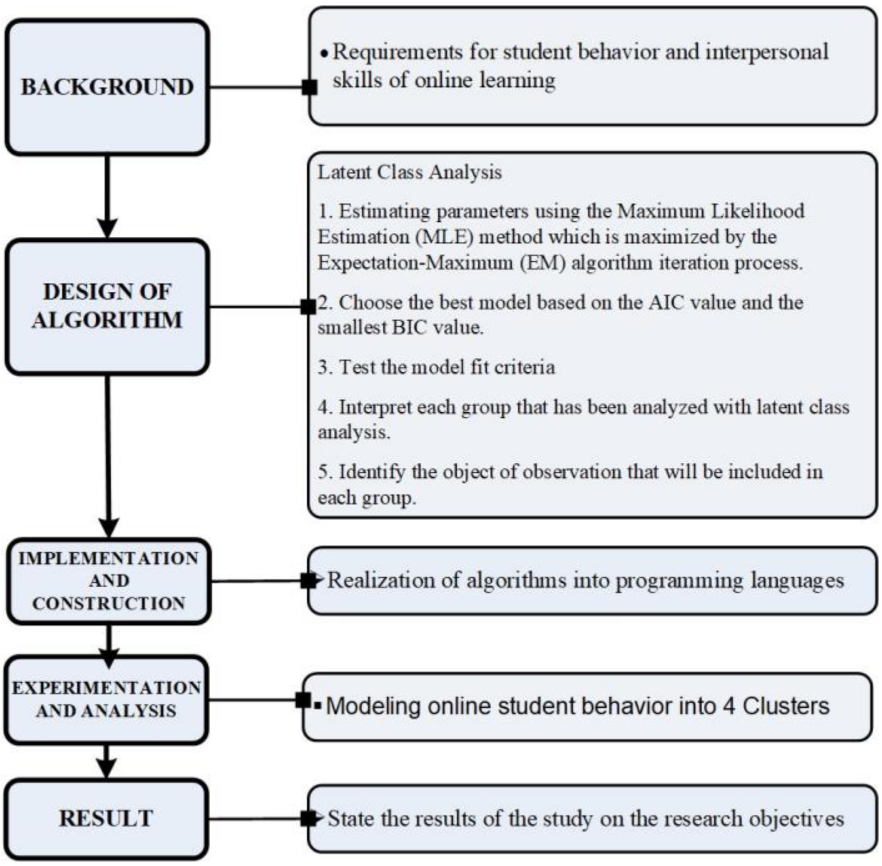


Figure 1. Methodology

In the data mining process, pre-processing the dataset, applying algorithms for data analysis, and post-process results is required [18,23]. Pre-processing the dataset is a first step and also very crucial because at that stage, the existing data set must be converted into data that has a certain pattern and meaning. In the first step, namely data extraction, we must collect the necessary student data that is in the database (LMS). Data taken from the course management system is in the form of course-related information that shows the hierarchical categorization of each specific class such as undergraduate vs diploma, which is affiliated with the campus and department. While the data taken from the learning management system are data about the total students in the class, student log data, and class log data.

In the second step, the data that has been taken are then combined using the course ID as the key value. To minimize the possibility of cheating, the course ID which was originally the name of each teacher was then converted into a number. In the third step, namely the elimination of missing values, there are many considerations that must be considered carefully because many data values are zero or empty in the log data. This is because some of the missing marks are caused by little or no specific activity in a course. One case in point is that the number of users logged into the system is less than the total number of users it should be. Therefore, 864 courses had a value of 0 and were deleted from the dataset. As a result, there were 1895 courses remaining in the dataset.

Before analyzing data, the researcher must filter the existing data again, whether the data is in good condition to enter cluster analysis or not. The first stage is an exploration of the pattern of classroom activities carried out on the remaining 1895 courses. Apart from these three variables, ten other activities were also reviewed, namely announcements, links, resources, quizzes, group assignments, class notes, questions and answers, assignment submissions, discussion forums, and wikis. After exploring the pattern of class activities, the remaining data will be determined for inclusion and exclusion criteria. The remaining 921 classes were randomized according to clusters that have a high degree of similarity to one another in an instructional intervention using latent class analysis in XLSTAT.

In the latent class analysis, it states that the latent class model must be local independence, which means that each variable in one group/latent class must be independent of each other. The assumption of local freedom only applies to latent variables, not every observed variable must be independent. If this assumption is not met, the analysis process will be much more complicated, because the observed object must be conditioned not only on latent class membership, but also on others.

$$P_{i_1,i_2,\ldots,i_n} \approx \sum_{t}^{T} P_t \prod_{n}^{N} P_{i_n,t}^{n} \tag{1}$$

where T is the number of latent classes or the unconditional probability.

For the two-way latent class model, the formula is

$$P_{ij} \approx \sum_{t}^{T} P_t P_{it} P_{jt} \tag{2}$$

3. **Results and Discussion.** Table 1 shows various aspects of the value of Class Size (MEM), Login Frequency per Course (LF), Average Login Frequency per Person (ALF), and Number of Activity Items (NAI). Class size (N of Member) is a variable whose value is taken from the number of students in each class. In this case, the MEM variable has a minimum value of 7, a maximum of 85, a mean of 33.57, and an SD of 18.59. Then the Login Frequency per Course (LF) variable is a variable whose value is taken from the basic log to understand all the activities in the class. Furthermore, the variable Average Login Frequency per Person (ALF) is a variable whose value is the average value of each user's login history for one semester (16 weeks). Finally, the variable Number of Activity Items (NAI) is a variable used to estimate the overall pattern of user online activities.

The first three activity items, namely an announcement, lecture note, and link, are activity items that have a teacher-led interaction type. The teacher-led interaction type is an interaction where teaching activities are carried out by the way the teacher provides material and then students download and read it. Furthermore, the fourth activity item

TABLE 1. Statistics of 1895 courses

Variables	Min	Med	Max	Mean	Mode	Elementary
Class Size (N)	7	28	85	(MEM)	19	18.59
Login Frequency per Course (LF)	33	12342	24848	12450.69	21285	7022.08
Average Login Frequency per Person (ALF)	37	66	94	65.57	91	17.06
Number of Activity Items (NAI)	1	5	10	5.414775726	2	2.898638228

TABLE 2. Segmentation of variables

Activities	Segment	Range	N	Min	Max	Mean	SD	Skewness	Kurtosis
Announcement	1	7∼12	316	7	64	18.66884	12.22606	2.07654	4.121562
	2	13∼20	378						
	3	21∼28	98						
	4	29∼64	129						
Lecture note	1	0∼	506	0	59	15.31162	19.32278	0.827729	−0.79642
	2	1∼59	415						
Link	1	0∼	884	0	1	0.040174 0.196366	4.690	991	20.04893
	2	1∼	37						
Resource	1	7∼12	456	7	396	15.924	14.7298	18.79478	481.1374
	2	13∼20	250						
	3	21∼28	125						
	4	29∼396	90						
Q&A	1	0∼	421	0	310	13.962	39.00753	6.729101	47.41481
	2	1∼310	500						
Quiz	1	0∼	877	0	1	0.047774	0.213288	4.247438	16.07564
	2	1∼	44						
Group project	1	0∼	734	0	1	0.20304	0.402262	1.478859	0.187426
	2	1∼	187						
Assignment submission	1	0∼	665	0	32	4.713355	8.894137	1.726974	1.568766
	2	1∼32	256						
Discussion forum	1	0∼	803	0	1	0.128122	0.334225	2.228948	2.974666
	2	1∼	118						
Wiki	1	0∼	593	0	1	0.356135	0.478856	0.601852	−1.64134
	2	1∼	328						

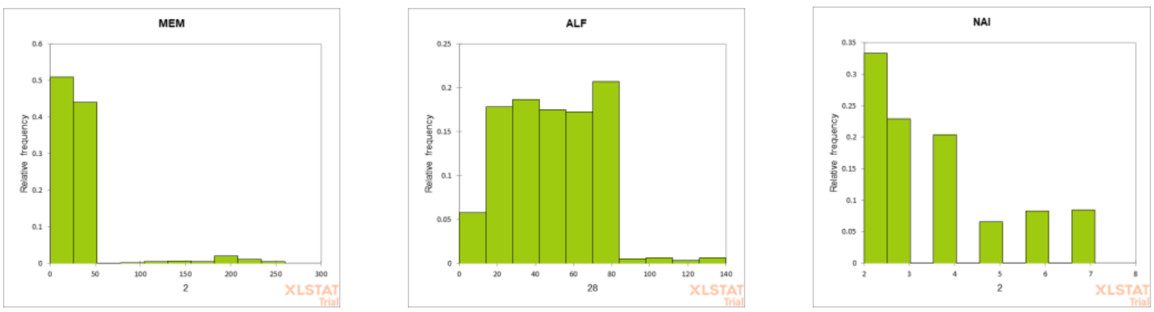
(resource), usually the type of interaction that is carried out in that class, is interactive activities. This type of interaction means that all participants in the class including teachers and students can upload and share information. This fourth activity item is also usually associated with the fifth activity item, namely Question and Answer (Q&A). Q&A is a question and answer activity that can be carried out between students and students and teachers. Then the last five activity items, namely quizzes, group projects, assignment submissions, discussion forums and wikis, have teacher-guided, student-centered activities. This type of interaction means the type where the facility must be opened by the teacher, but still requires student participation.

Descriptive statistics of ten variables have been presented in Table 2. In the dataset, most of the existing data tended to have a value of "0", and this resulted in many classes that did not include many activity items. In addition, this data with a value of 0 can also be evidence that in courses, students are rarely given the opportunity to learn actively online. In this case, the announcement has a maximum value of 129, the lecture note has a maximum value of 163, the link has a maximum value of 64, the resource has a maximum value of 559, the Q&A has a maximum value of 312, the quiz has a maximum value of 325, the project group has a maximum value of 1031, assignment submission has a maximum value of 44, discussion forum has a maximum value of 4197, and wiki has a maximum value of 47. Even though in Table 2 all activity items have a high enough value and prove that online activities are running well, there is still a value of 0 in this data which may hinder further analysis such as classification or prediction due to distorted estimates. Therefore, the authors combine two strategies, namely data reduction and segmentation for the next stage.

In the classification of the blended course phase, the first step that must be done is to segment the data and take samples to determine the most representative samples and the blended course profile. In Table 2, the variable segmentation of all activities in the online course is presented. There are 2 activity items that are most frequently used, i.e., resources and announcements, which are converted using a quantitative score from each variable into 4 segments. Meanwhile, activity items that are rarely used, such as Q&A, lecture notes, assignment submissions, group projects, links, discussion forums and wikis, are divided into 2 segments (0 and 1). Based on the table, it can be seen that the activity of delivering instructions is the most frequently used activity item, while student-centered learning activities are rarely used.

After doing the segmentation stage data, then the next step is to determine inclusion and exclusion criteria. At that stage, the courses used are only courses that show the top 25% of use of one of the main learning activities and contain one minor activity. As a result, 1296 courses were selected to manually check the patterns of their online activities. We also found that the majority of the existing courses still had very low levels of use which resulted in ambiguous interpretation of the data. Therefore, as many as 1151 courses that are included in the low 50% segmentation such as resources and announcements are excluded.

To review the basic features in selecting 921 courses, therefore the existing descriptive statistics are carried out again. As a result, the majority of the undergraduate level (n = 737, 80%) classes were designed as major-specific (n = 277, 30%) or health sciences (n = 369, 40%) courses. In the figure below, there is a graph of the MEM, ALF, and NAI variables which have a normal distribution level. On the median score, it is known



Median = 25

Course type	Members	Case
Small	2∼20	288
Medium	21∼30	391
M to L	31∼40	105
Large	41∼255	137

Median = 47

Course type	Members	Case
Rarely	10∼30	253
Occasionally	31∼50	242
Frequently	51∼70	231
Very frequently	71∼134	195

Median = 3

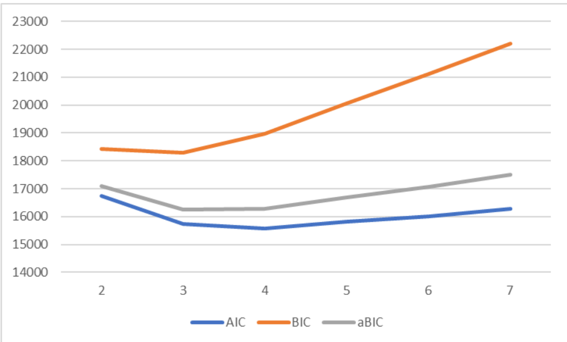
Course type	Members	Case
Not diverse	2	308
Little diverse	3	211
Somewhat diverse	4	188
Very diverse	5∼7	214

Figure 2. Chart of MEM, ALF, and NAI (n = 921)

that the distribution is divided into four quartiles. The data set from the 921 blended courses shows that about 25 students who logged in, with an average of 47 points, also participated in three online activities.

Furthermore, to create a classification model in the second stage, the determination of the number of latent classes must be done carefully. The AIC value obtained shows a decreasing pattern, while the BIC value shows an increasing pattern. Therefore, the LMR-LRT values declared with P-values with 2 and 4 classes, indicate that the two models are more correct solutions than the 5 and 6 classes. These results prove that the 4-class model has a lower BIC value and has the best fit with the highest level and unique patterns.

From the 4 existing class models, then the next step is to compare each class using the profile of each activity item as shown in Figures 3(a) and 3(b). In Table 3, the clustered data has been presented. There are 4 clusters in this case and each of them is divided into 3 levels. At level 1, the cluster is divided into two parts, namely Undergraduate and Graduate. Then at level 2, the cluster is further divided into 2 to 3 groups (divided according to their majors). Finally, at level 3, the cluster is divided into several more specific sections (according to the direction). Furthermore, in Table 4, the general features of the 4 clusters discussed previously are also presented.



(a)

(b)

FIGURE 3. (a) Class plot (AIC = 16800-16150), (BIC = 18300-22100), (aBIC = 17050-17500); (b) probability plots (Class 1 = 0.8-0.2), (Class 2 = 0.9-0.1), (Class 3 = 0.75-0.3), (Class 4 = 0.2-0.45)

4. **Conclusions.** In this study, the authors used the academic analytics approach, which is a method of classifying blended learning courses. This approach is also included in Educational Data Mining (EDM) because the dataset used is a dataset of Learning Management System (LMS) from a university in Indonesia. The result of this research is that the latent class analysis method can analyze the behavior of students and teachers while in the Learning Management System (LMS). In this study, the dataset used initially amounted to 2759 courses. However, after the data mining process was carried out several times, the dataset used shrank to 1895, then finally to 921. When doing the first data mining process, where the shrinkage of the number of datasets occurred, it was seen that the pattern of the 1895 courses that existed tended to be low in many aspects. Therefore, the researcher conducted a second data mining process to combine several classes to make it more effective. Researchers recommend in-depth analysis to examine the planned instructional methods of the course from two sides, namely online and offline.

In the second data mining process, researchers used 921 sample courses that present relatively active online activities. With latent class analysis, 4 clusters are generated to divide 4 types of courses. In cluster 1, as many as 355 classes with a percentage of 38.55%, are considered as the group that has the highest level of communication and collaboration. The majors that are included in cluster 1 are Social Science and Technology. Furthermore,

Table 3. Demographic characteristic of 4 clusters (n = 921)

Cluster	Level 1	Level 2	Level 3	n (%)
Cluster 1: 355 (38.55%)	Undergraduate: 213 (60%)	Major specific: 97 (27.3%)	Social Sciences	44 (12.4%)
			Others	53 (14.9%)
		Health sciences: 86 (24.2%)	Mandatory	41 (11.5%)
			Selective	45 (12.7%)
		Major foundation: 30 (8.5%)		
	Graduate: 142 (40%)	General: 36 (10.1%)		
		Special	Technology	64 (18%)
			Others	42 (11.8%)
Cluster 2: 179 (19.44%)	Undergraduate: 128 (71.5%)	Major specific: 63 (49.2%)	English Literature	38 (29.7%)
			Others	25 (19.5%)
		Health sciences: 46 (35.9%)	Mandatory	27 (21.1%)
			Selective	19 (14.8%)
		Major foundation: 19 (14.9%)		
	Graduate: 51 (28.5%)	General: 13 (7.3%)		
		Special	Sociology	17 (9.5%)
			Others	21 (11.7%)
Cluster 3: 265 (28.77%)	Undergraduate: 166 (46.8%)	Major specific: 73 (20.6%)	Business Management	39 (11%)
			Others	34 (9.6%)
		Health sciences: 48 (13.5%)	Mandatory	29 (8.2%)
			Selective	19 (5.4%)
		Major foundation: 45 (12.7%)		
	Graduate: 99 (27.9%)	General: 42 (11.8%)		
		Special	Accounting	21 (5.9%)
			Others	36 (10.1%)
Cluster 4: 122 (13.25%)	Undergraduate: 63 (17.7%)	Major specific: 12 (3.4%)	Hospitality	8 (2.3%)
			Others	4 (1.1%)
		Health sciences: 26 (7.3%)	Mandatory	12 (3.4%)
			Selective	14 (3.9%)
		Major foundation: 25 (7%)		
	Graduate: 59 (16.6%)	General: 26 (7.3%)		
		Special	Politics Science	13 (7.3%)
			Others	20 (5.6%)

Table 4. General features of 4 clusters (n = 921)

Variables	Cluster 1 (n = 355, 38.55%)	Cluster 2 (n = 179, 19.44%)	Cluster 3 (n = 265, 28.77%)	Cluster 4 (n = 122, 13.25%)
MEM				
2∼20	67 (18.9%)	28 (15.6%)	33 (12.5%)	28 (23%)
21∼30	97 (27.3%)	65 (36.3%)	51 (19.2%)	43 (35.2%)
31∼40	86 (24.2%)	27 (15.1%)	79 (29.8%)	33 (27%)
41∼255	85 (23.9%)	59 (33%)	102 (38.5%)	18 (14.8%)
ALF				
10∼30	76 (21.4%)	76 (42.5%)	13 (4.9%)	71 (58.2%)
31∼50	81 (22.8%)	32 (17.9%)	89 (33.6%)	23 (18.9%)
51∼70	87 (24.5%)	18 (10.1%)	96 (36.2%)	17 (13.9%)
71∼134	91 (25.6%)	53 (29.6%)	67 (36.2%)	11 (9%)
NAI				
2	27 (7.6%)	66 (36.9%)	0 (0%)	27 (22.1%)
3	162 (45.6%)	76 (42.5%)	90 (34%)	20 (16.4%)
4	42 (11.8%)	21 (11.7%)	89 (33.6%)	44 (36.1%)
5∼7	104 (29.3%)	16 (8.9%)	86 (32.5%)	31 (25.4%)

in cluster 2, 179 classes with a percentage of 19.44% were considered as the superior group in the field of group discussion. Programs that are included in cluster 2 are English Literature and Sociology. Then in cluster 3, as many as 265 classes with a percentage of 28.77%, are considered as the superior group in the submission or quiz field. In cluster 3, the majors included in it are Business Management and Accounting. Finally, in cluster 4, as many as 122 classes with a percentage of 13.25%, are considered as a group that has not yet matured in the application of blended learning or is even inactive. Programs included in cluster 4 are Hospitality and Political Science. In order to increase the effectiveness of blended learning, it is hoped that the university can invest more in developing LMS. In addition, good communication between teachers and students is needed in order to create online courses that are informative, communicative, and effective.

Acknowledgment. This research is the result of the basic research scheme of the Indonesian Dikti grant 0054/E5/AK.04/2022.

REFERENCES

[1] A. Akram, C. Fu, Y. Li, Y. Javed, R. Lin, Y. Jiang and Y. Tang, Predicting students' academic procrastination in blended learning course using homework submission data, *IEEE Access*, vol.7, pp.102487-102498, 2019.

[2] K. I. M. Ramaphosa, T. Zuva and R. Kwuimi, Educational data mining to improve learner performance in Gauteng primary schools, *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, pp.1-6, 2018.

[3] F. Harrak, F. Bouchet and V. Luengo, PHS profiling students from their questions in a blended learning environment, *LAK'18: International Conference on Learning Analytics and Knowledge*, pp.102-110, 2018.

[4] F. Harrak, F. Bouchet and V. Luengo, From student questions to student profiles in a blended learning environment, *Journal of Learning Analytics*, vol.6, no.1, pp.54-84, 2019.

[5] O. H. Lu, A. Y. Huang, J. C. Huang, A. J. Lin, H. Ogata and S. J. Yang, Applying learning analytics for the early prediction of students' academic performance in blended learning, *Educational Technology and Society*, vol.21, no.2, pp.220-232, 2018.

[6] C. Silva and J. Fonseca, Educational data mining: A literature review, in *Europe and MENA Cooperation Advances in Information and Communication Technologies*, Springer, Cham, pp.87-94, 2017.

[7] M. Nakayama, K. Mutsuura and H. Yamamoto, Using note taking instructions to reform student's note taking activities and improve learning performance in a blended learning course, *2018 22nd International Conference Information Visualisation*, pp.326-331, 2018.

[8] M. A. Rentroia-Bonito, D. Goncalves and J. A. Jorge, Clustering students based on motivation to learn: A blended learning approach, *International Journal of Mobile and Blended Learning*, vol.7, no.3, pp.18-39, 2015.

[9] M. Zaffar, M. A. Hashmani and K. S. Savita, Performance analysis of feature selection algorithm for educational data mining, *2017 IEEE Conference on Big Data and Analytics (ICBDA)*, 2017.

[10] E. B. Costa et al., 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.

[11] 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.

[12] A. Triayudi, W. O. Widyarto and V. Rosalina, CLG clustering for mapping pattern analysis of student academic achievement, *ICIC Express Letters*, vol.14, no.12, pp.1225-1234, 2020.

[13] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho and G. V. Erven, Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil, *Journal of Business Research*, vol.94, pp.335-343, 2019.

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

[15] T. Devasia, T. P. Vinushree and V. Hegde, Prediction of students performance using Educational Data Mining, *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, pp.91-95, 2016.

[16] A. A. Saa, Educational data mining and 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 and Education*, vol.113, pp.177-194, 2017.

[18] A. D. Kumar, R. P. Selvam and K. S. Kumar, Review on prediction algorithms in educational data mining, *International Journal of Pure and Applied Mathematics*, vol.118, no.8, pp.531-537, 2018.

[19] Y. K. Salal, S. M. Abdullaev and M. Kumar, Educational data mining: Student performance prediction in academic, *International Journal of Engineering and Advanced Technology*, vol.8, no.4C, pp.54-59, 2019.

[20] S. T. Ahmed, R. Al-Hamdani and M. S. Croock, Developed third iterative dichotomizer based on feature decisive values for educational data mining, *Indonesian Journal of Electrical Engineering and Computer Science*, vol.18, no.1, pp.209-217, 2020.

[21] C. Jalota and R. Agrawal, Analysis of educational data mining using classification, *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India, pp.243-247, 2019.

[22] A. I. Adekitan and O. Salau, The impact of engineering students' performance in the first three years on their graduation result using educational data mining, *Heliyon*, vol.5, no.2, e01250, 2019.

[23] H. Aldowah, H. Al-Samarraie and W. M. Fauzy, Educational data mining and learning analytics for 21st century higher education: A review and synthesis, *Telematics and Informatics*, vol.37, pp.13-49, 2019.

[24] M. C. Sáiz-Manzanares, J. J. Rodríguez-Díez, J. F. Díez-Pastor, S. Rodríguez-Arribas, R. Marticorena-Sánchez and Y. P. Ji, Monitoring of student learning in learning management systems: An application of educational data mining techniques, *Applied Sciences*, vol.11, no.6, 2677, 2021.