E-ISSN : 2963-4946
Vol. 1 No. 01 October 2022

Asian Journal of
Social and Humanities
https://ajosh.org/

Mapping Student Log Files With K-Means Clustering

Indra Maulana[1], Moh. Agri Triansyah[2]
[1] Universitas Negeri Yogyakarta, Indonesia
[2] Institut Pendidikan dan Bahasa Invada, Indonesia
Email: indra@stkipinvada.ac.id, agritriansyah@ipbcirebon.ac.id

* Correspondence: indra@stkipinvada.ac.id

KEYWORDS

Mapping Student; Log Files; K-Means Clustering.

ABSTRACT

One of the important characteristics of e-learning platforms is that students can take courses at any time, and they are not required to complete all available learning activities at one time. In Moodle, log data is valuable information that contains the activities of course users and course teachers. The data recorded in the Moodle data log can be in the form of activity data, assignment collection time (assignment timestamp), and ranking or final grades (grade). Exploration of data logs on educational data or educational data mining can be used to facilitate monitoring and see what activities are often carried out by course participants on the Moodle platform. One of the techniques used in data mining log data analysis is cluster analysis. Cluster analysis is the process of grouping data into groups whose members have similar characteristics. K-means Clustering is one of the algorithms of cluster analysis that is often used. Based on the output, it can be seen that the members of cluster 1 are students with id 1,3,4,5, and 9. Then for cluster 2 are students with id 2,8,10,12 which in cluster 2 the average student clicks is the highest. ,. and the last cluster 3 is filled by students with id 6, 7, and 11. It can be concluded that the second cluster is a collection of students who are active in accessing the LMS during learning it can be seen that the members of cluster 1 are students with id 1,3,4,5, and 9. Then for cluster 2 are students with id 2,8,10,12 which in cluster 2 the average student clicks is the highest. and the last cluster 3 is filled by students with id 6, 7, and 11. It can be concluded that the second cluster is a collection of students who are active in accessing the LMS during learning it can be seen that the members of cluster 1 are students with id 1,3,4,5, and 9. Then for cluster 2 are students with id 2,8,10,12 which in cluster 2 the average student clicks is the highest. and the last cluster 3 is filled by students with id 6, 7, and 11. It can be concluded that the second cluster is a collection of students who are active in accessing the LMS during learning.

Introduction

The implementation of e-Learning in the world of education continues to grow and even becomes the key to implementing learning during the Covid-19 pandemic. Currently, e-learning platforms are emerging as the main platform for distance learning, especially in integrating teaching and learning activities with technology, e-learning platforms not only provide various teaching materials and learning resources but also help students in independent learning (Lee et al. , 2021). One of the important characteristics of e-learning platforms is that students can take courses at any time, and they are not required to complete all available learning activities at one time (Y. Lee, 2019). Most MOOC systems i.e. e-learning store data about teaching and learning actions in log files, which gives us detailed information about learner behavior (Cocea & Weibelzahl, 2009). Log data or even log files are collections or lists of various actions that have been carried out by users (Jong et al., 2007).

The increasing use of Learning Management Systems (LMS), Massive Open Online Courses (MOOC), and other Online Learning Environments (OLE), which has significantly increased the volume of log data, takes advantage of the opportunity to examine student activity during their online learning. Nonetheless, as evidenced by a number of studies on this issue (Alario-Hoyos et al., 2020). Along with the increasing popularity of online learning, more in-depth monitoring is needed to determine its success in the educational environment. The main challenge is figuring out how to "translate" the data obtained into useful knowledge for education.

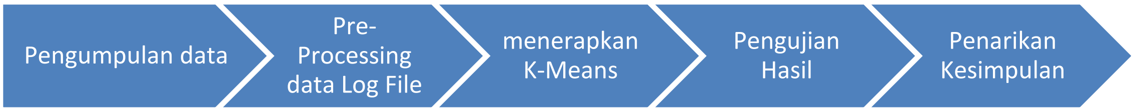
Moodle As an LMS that has many advanced features and functions, various transactions from the use of this feature are recorded and recorded by Moodle which are then stored in the data log. Log data or even log files are collections or lists of various actions that have been carried out by users (Herrero et al., 2014). In Moodle, log data is valuable information that contains the activities of course users and course teachers. The data recorded in the moodle data log can be in the form of activity data, task collection time (assignment timestamp), and ranking values or final grades (Młynarska et al., 2016).

Exploration of data logs on educational data or educational data mining can be used to facilitate monitoring and see what activities are often carried out by course participants on the Moodle platform. Data logs record various participant activities and store them in the system database. In addition, the results from the analysis of mining log data can be used to determine what learning strategies can be used to increase the participation and activity of course participants.

One of the techniques used in data mining log data analysis is cluster analysis. Cluster analysis is the process of grouping data into groups whose members have similar characteristics (Liu, 2015). Cluster analysis aims that objects in one group have similarities with each other while with objects of different groups there are differences. Cluster analysis has several advantages, namely it can group large amounts of data and many variables and can be used on ordinal, interval and ratio data scales. K-means Clustering is one of the algorithms of cluster analysis that is often used. This is because this method is quite easy to use and interpret (Liu, 2015).

There are several cluster analysis studies with k-means, namely the Cyber Profiling in Criminal Investigation (Yu, 2020) research which performs a profile analysis of digital traces associated with a person whose identity may not be known. Then there is also research conducted (Zhang et al., 2020), conducting log file research to find out student activities in the LMS on student learning outcomes.

Research Methods



This study uses a quantitative approach to view the description of the log data and see the pattern of grouping of trainees based on training activities. The data used in this study were derived from user participant data and the data log of the "Data Visualization with Tableau" course which was held on June 29 – July 1, 2020. The Moodle environment not only provides students with easy access to educational services, but also provides higher education institutions the ability to collect large amounts of data about student behavior. This data provides many opportunities to analyze student behavior and can also help determine whether there are trends that are contributing to increased learning success

The number of participants who took this course was 313 participants other than the instructor and course admin which was obtained from the user participants table on the Moodle LMS. The log data obtained from the Moodle LMS contains 13027 rows and 9 columns. The following are the attributes contained in the log data used:

Table 1.Dimensions of Logs Files

Data Dimensions	**Description**
Time	Date and time when the Action is performed
User Full Name	The name of the user who performed the action.
Affected users	Users affected by participant actions.
Event Context	General information about learning activities.
Element	Where are the elements in the Moodle system
Event Name	Clicked Event Name
Description	Viewed Event ID
Origin	What application is used to login to Moodle.
IP address	IP Address

In processing this data, it is not simple and indirect because of its large volume and high speed. Prepocssing data is needed to extract useful information for tracking and assessing activities carried out by students. Log files are accessed from Their Moodle, describing what actions students have taken over time. Data can be exported to Microsoft Excel (.xlsx), open documents (.ods) or comma-separated values (.csv).

Table 2 shows a typical log file layout taken from e-learning Moodle and used in this study. It shows how the data is organized: the leftmost column is the date and time of access. Each line of the log file measures the amount of time participants spend in Moodle minute by minute. The next two columns are the name of the logged in participant and who was affected by that participant's actions. The last three columns hold the data of interest. The middle two columns indicate the context, which is classified as "Event Context", that is, where the action is performed in the LMS. "Elements" contains all the elements used by students. The far right column shows what participants have done on the Moodle site classified as "Event Name." other information,

Table 2. Example of log data from e-learning course

Time	Full name	Affected users	Event context	Component	Event name	Description	Origin	IP Address
03/29/22, 19:52	Muhammad Nuralim	-	Forum: discussion of e-learning tools	System	View courses	The user with id '374' viewed the course with id '8'.	web	114.5.215.197
03/29/22, 5:50pm	Suradi Fauzi	-	File: meeting materials for 3 eLearning tools	Forum	View discussion	The user with id '394' has viewed the discussion with id '178' in the forum with course module id '1214'.	web	103.143.98.129
03/29/22, 12:57	Novi Andini	-	Forum: Task Forum	System	View courses	The user with id '381' viewed the course with id '8'.	web	139.194.27.252

03/29/22, 11:36	Amine	-	Forum: discussion of e-learning tools	System	View courses	The user with id '390' viewed the course with id '8'.	web	114.122.104.116
….	…..	…..	…..	…..	…..	…..	….	….

In this study, a cluster analysis will be conducted using Moodle logs to track the behavior of two types of students: students with high scores on both exams are referred to as "High Involvement Students" and students with low scores on both tests are referred to as "Low Involvement Students".

K-means Clustering

This study uses the K-Means algorithm method, to create clusters of groups of students who are active in learning in e-learning at SMK AlWashliyah Cirebon students by using log files. K-Means Clustering method aims to minimize the objective function set in the clustering process by minimizing variations between data in a cluster and maximizing variations with data in other clusters.

Results and Discussions

Before entering the data mining analysis using the k-means clustering method, an exploratory analysis of the data was carried out in the form of descriptive statistics. The results of the analysis through descriptive statistics discuss the statistical summary of actions taken by students in e-learning. Actions are actions taken by users (users) in the LMS or you could say the number of clicks made by students on activities on the material in the LMS

In this paper, the author uses the log data for the "E-Learning" course and focuses on "Description", because the Description of the Event Name stores information on the menu name clicked by the student. The research was conducted using the following parameters:

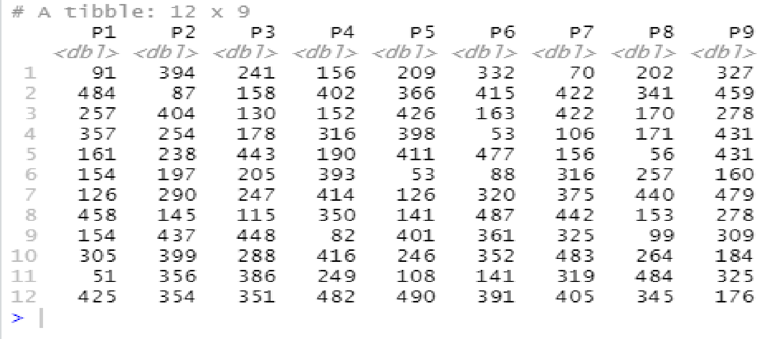
Number of clusters : 3
Number of data : 10
Number of attributes : 9

Determine the criteria that will be used as value attributes for each alternative, with the following criteria:

Table 3. Attribute Description of Event Name

P1 =	Material External Link 1
P2 =	Meeting Material 1
P3 =	Meeting Material 2
P4 =	Video Material 2
P5 =	Material External Link 2
P6 =	Task Forum
P7 =	Discussion Session
P8 =	Meeting Material 3

Table 4 Summary of Action Statistics Based on Clicks Made



```
# A tibble: 12 x 9
      P1    P2    P3    P4    P5    P6    P7    P8    P9
   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
 1    91   394   241   156   209   332    70   202   327
 2   484    87   158   402   366   415   422   341   459
 3   257   404   130   152   426   163   422   170   278
 4   357   254   178   316   398    53   106   171   431
 5   161   238   443   190   411   477   156    56   431
 6   154   197   205   393    53    88   316   257   160
 7   126   290   247   414   126   320   375   440   479
 8   458   145   115   350   141   487   442   153   278
 9   154   437   448    82   401   361   325    99   309
10   305   399   288   416   246   352   483   264   184
11    51   356   386   249   108   141   319   484   325
12   425   354   351   482   490   391   405   345   176
> |
```

There are several ways that are used to determine the number of clusters, namely using the Elbow, Silhouette method. Figures 3 and 4 show how to determine clusters with two methods, namely wss/Elbow and Silhouette in RStudio.

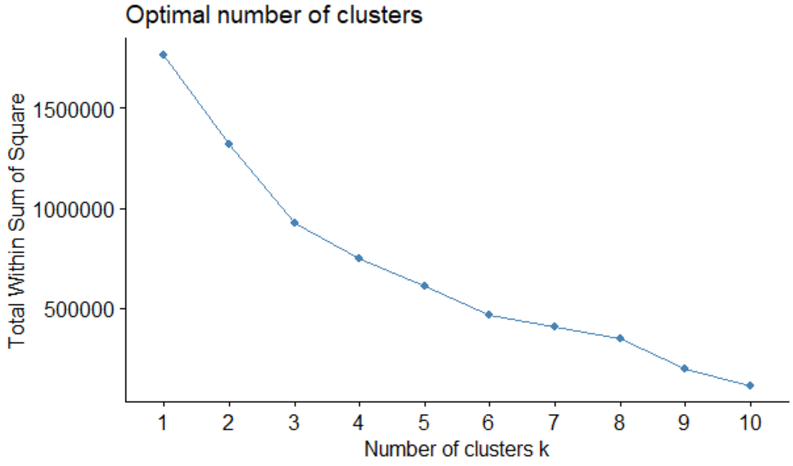


Figure 3. Elbow method

The first method uses the Elbow method. Figure 3 shows a line diagram of the results of the Elbow method. In the picture there is a blunt point formed between points two and four, after point three there is no longer too deep a decline but at number eight there is a slight increase which then drops again, so the number of clusters from the Elbow method is three clusters,=3.

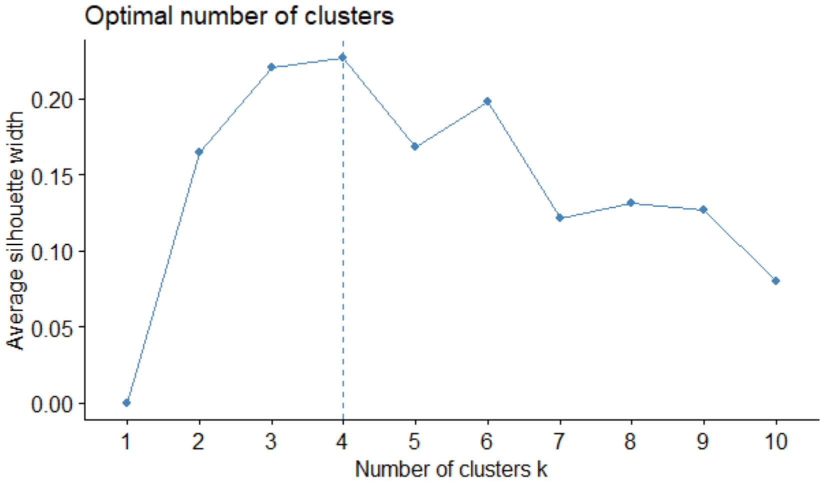
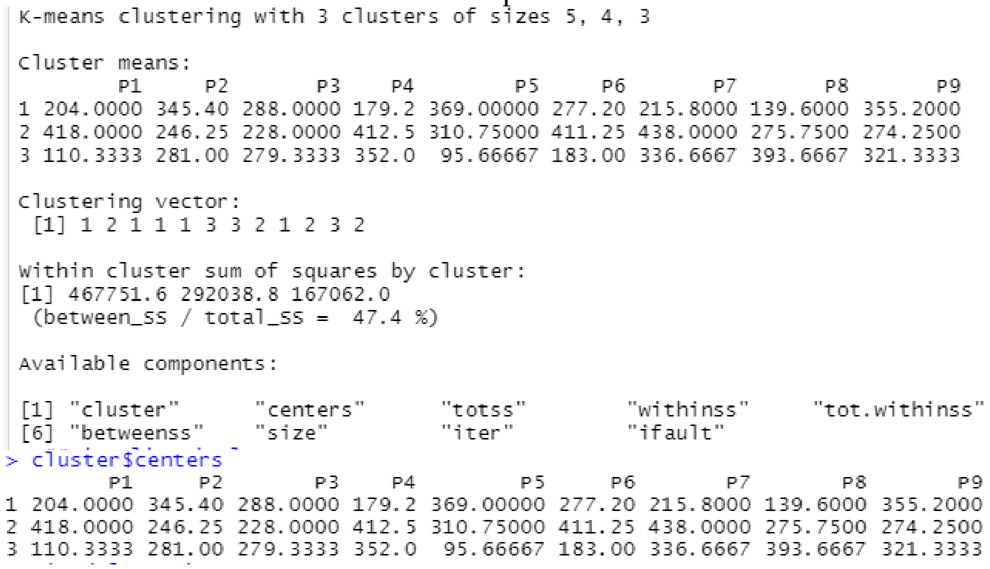


Figure 4. Determination of k with Silhouette

The second way is to use the Silhouette method. Figure 6 shows a line diagram of the results of the Silhouette method. In the picture there is the highest point formed between points three and five, after point three a too deep decline did not occur but in figures eight and six there was a slight increase which then fell again, so that the number of clusters according to the results of the Silhouette method is 4 clusters. Because 4 clusters are too many and for this research, 3 clusters are enough, so 3 clusters are used. Furthermore, the results of the analysis using the k-means clustering method are presented in Figure 3. Based on Figure 3, it can be seen that there are 3 groups (clusters) formed. Coloring for each cluster in Figure 3 is done to make it easier to visualize the results of the cluster analysis

Table 5. Output k-means



```
K-means clustering with 3 clusters of sizes 5, 4, 3

Cluster means:
        P1     P2       P3    P4        P5     P6       P7       P8       P9
1 204.0000 345.40 288.0000 179.2 369.00000 277.20 215.8000 139.6000 355.2000
2 418.0000 246.25 228.0000 412.5 310.75000 411.25 438.0000 275.7500 274.2500
3 110.3333 281.00 279.3333 352.0  95.66667 183.00 336.6667 393.6667 321.3333

Clustering vector:
 [1] 1 2 1 1 1 3 3 2 1 2 3 2

Within cluster sum of squares by cluster:
[1] 467751.6 292038.8 167062.0
 (between_SS / total_SS =  47.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
> cluster$centers
        P1     P2       P3    P4        P5     P6       P7       P8       P9
1 204.0000 345.40 288.0000 179.2 369.00000 277.20 215.8000 139.6000 355.2000
2 418.0000 246.25 228.0000 412.5 310.75000 411.25 438.0000 275.7500 274.2500
3 110.3333 281.00 279.3333 352.0  95.66667 183.00 336.6667 393.6667 321.3333
```

based on the output obtained by cluster 1 members there are 5, cluster 2 there are 4, and cluster 3 there are 3. Next, we visualize the clustering with K = 3

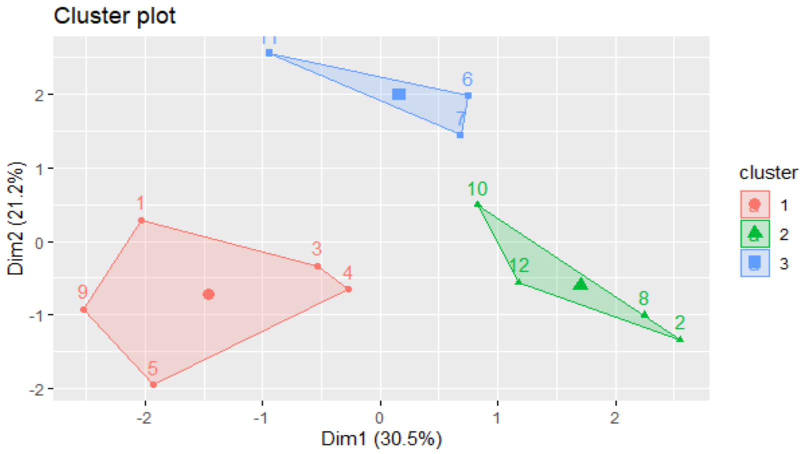
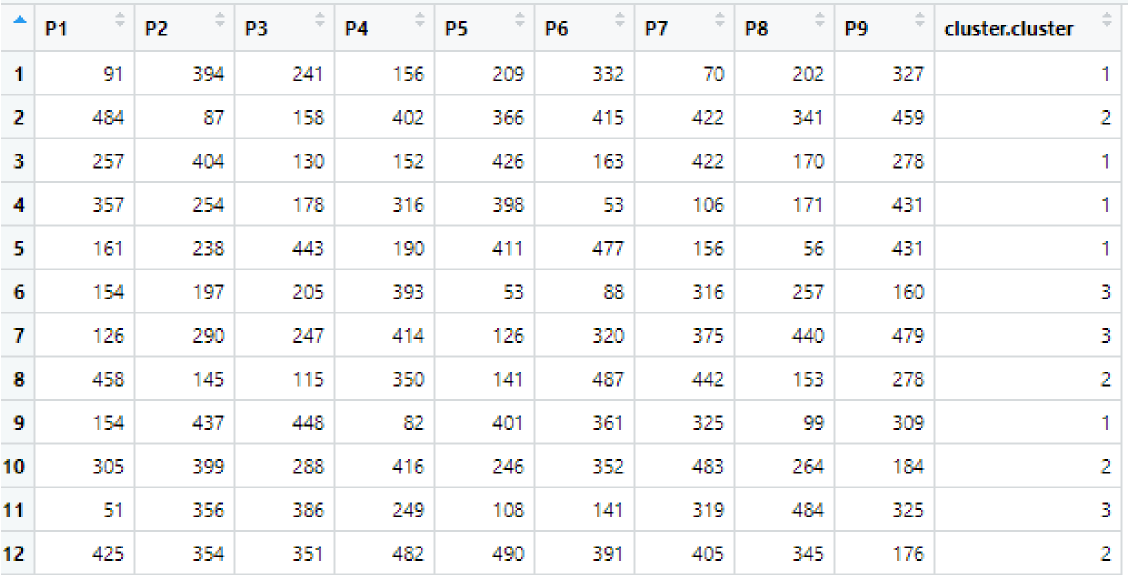


Figure 5. Cluster Visualization

Based on the output obtained, it can be seen that members of cluster 1 are red, members of cluster 2 are green, and members of cluster 3 are blue.

Table 6. cluster results for each data



	P1	P2	P3	P4	P5	P6	P7	P8	P9	cluster.cluster
1	91	394	241	156	209	332	70	202	327	1
2	484	87	158	402	366	415	422	341	459	2
3	257	404	130	152	426	163	422	170	278	1
4	357	254	178	316	398	53	106	171	431	1
5	161	238	443	190	411	477	156	56	431	1
6	154	197	205	393	53	88	316	257	160	3
7	126	290	247	414	126	320	375	440	479	3
8	458	145	115	350	141	487	442	153	278	2
9	154	437	448	82	401	361	325	99	309	1
10	305	399	288	416	246	352	483	264	184	2
11	51	356	386	249	108	141	319	484	325	3
12	425	354	351	482	490	391	405	345	176	2

Conclusion

Based on the output, it can be seen that the members of cluster 1 are students with id 1,3,4,5, and 9. Then for cluster 2 are students with id 2,8,10,12 which in cluster 2 the average student clicks is the highest. ,. and the last cluster 3 is filled by students with id 6, 7, and 11. It can be concluded that the second cluster is a collection of students who are active in accessing the LMS during learning.

References

Alario-Hoyos, C., Rodríguez-Triana, MJ, Scheffel, M., Arnedillo-Sánchez, I., & Dennerlein, SM (2020). Addressing Global Challenges and Quality Education. 15th European Conference on Technology Enhanced Learning, EC-TEL 2020 Heidelberg, Germany, September 14–18, 2020 Proceedings.

Cocoa, M., & Weibelzahl, S. (2009). Log file analysis for disengagement detection in e-Learning environments. In User Modeling and User-Adapted Interaction (Vol. 19, Issue 4). https://doi.org/10.1007/s11257-009-9065-5

Herrero, ., Baruque, B., Klett, F., Abraham, A., Snášel, V., De Carvalho, ACPLF, Bringas, PG, Zelinka, I., Quintián, H., & Corchado, E. (2014). Preface. Advances in Intelligent Systems and Computing, 239, v–vi. https://doi.org/10.1007/978-3-319-01854-6

Jong, BS, Chan, TY, & Wu, YL (2007). Learning log explorer in e-learning diagnosis. IEEE Transactions on Education, 50(3), 216–228. https://doi.org/10.1109/TE.2007.900023

Lee, CA, Tzeng, JW, Huang, NF, & Su, YS (2021). Prediction of Student Performance in Massive Open Online Courses Using Deep Learning System Based on Learning Behaviors. Educational Technology and Society, 24(3), 130–146.

Liu, B. (2015). Web Data Mining Exploring Hyperlinks, Contents, and Usage Data. In Global Journal of Pure and Applied Mathematics (Vol. 11, Issue 5).

Młynarska, E., Greene, D., & Cunningham, P. (2016). Time series clustering of Moodle activity data. CEUR Workshop Proceedings, 1751, 104–115.

Yu, S. (2020). Cyber Profiling in Criminal Investigation. 333–343. https://doi.org/10.4018/978-1-7998-3479-3.ch024

Zhang, Y., Ghandour, A., & Shestak, V. (2020). Using Learning Analytics to Predict Students Performance in Moodle LMS. International Journal of Emerging Technologies in Learning, 15(20), 102–114. https://doi.org/10.3991/ijet.v15i20.15915

Alario-Hoyos, C., Rodríguez-Triana, MJ, Scheffel, M., Arnedillo-Sánchez, I., & Dennerlein, SM (2020). Addressing Global Challenges and Quality Education. 15th European Conference on Technology Enhanced Learning, EC-TEL 2020 Heidelberg, Germany, September 14–18, 2020 Proceedings.

Cocoa, M., & Weibelzahl, S. (2009). Log file analysis for disengagement detection in e-Learning environments. In User Modeling and User-Adapted Interaction (Vol. 19, Issue 4). https://doi.org/10.1007/s11257-009-9065-5

Herrero, ., Baruque, B., Klett, F., Abraham, A., Snášel, V., De Carvalho, ACPLF, Bringas, PG, Zelinka, I., Quintián, H., & Corchado, E. (2014). Preface. Advances in Intelligent Systems and Computing, 239, v–vi. https://doi.org/10.1007/978-3-319-01854-6

Jong, BS, Chan, TY, & Wu, YL (2007). Learning log explorer in e-learning diagnosis. IEEE Transactions on Education, 50(3), 216–228. https://doi.org/10.1109/TE.2007.900023

Lee, CA, Tzeng, JW, Huang, NF, & Su, YS (2021). Prediction of Student Performance in Massive Open Online Courses Using Deep Learning System Based on Learning Behaviors. Educational Technology and Society, 24(3), 130–146.

Liu, B. (2015). Web Data Mining Exploring Hyperlinks, Contents, and Usage Data. In Global Journal of Pure and Applied Mathematics (Vol. 11, Issue 5).

Młynarska, E., Greene, D., & Cunningham, P. (2016). Time series clustering of Moodle activity data. CEUR Workshop Proceedings, 1751, 104–115.

Yu, S. (2020). Cyber Profiling in Criminal Investigation. 333–343. https://doi.org/10.4018/978-1-7998-3479-3.ch024

Zhang, Y., Ghandour, A., & Shestak, V. (2020). Using Learning Analytics to Predict Students Performance in Moodle LMS. International Journal of Emerging Technologies in Learning, 15(20), 102–114. https://doi.org/10.3991/ijet.v15i20.15915