E-learning in the Transformation of Education in Digital Society
Scientific Editor Eugenia Smyrnova-Trybulska
"E-learning", 14, Katowice–Cieszyn 2022, pp. 257–267
https://doi.org/10.34916/el.2022.14.19

IDENTIFICATION OF STUDENTS WITH SIMILAR BEHAVIOURAL PATTERNS USING CLUSTERING TECHNIQUES

Janka Pecuchova[1] **& Martin Drlik**[2]
[1,2] Constantine the Philosopher University in Nitra
[1] janka.pecuchova@ukf.sk, ORCID 0000-0002-9225-7449
[2] mdrlik@ukf.sk, ORCID 0000-0002-5958-7147

Abstract: *A log file is used by most e-learning platforms to hold the data collected concerning the learning and teaching activity of the stakeholders. The data gathered needs to be transformed into a helpful form before any decisions can be made. Using educational data mining techniques, one may analyse this data to understand the learning process better and predict their outcomes or early dropout. Simultaneously, it is possible to identify students with similar behavioural patterns. However, students should not be categorized only according to their grades. Instead, their engagement with the educational resources and activities presented within e-learning courses should also be considered. Therefore, this paper presents research in which an unsupervised clustering algorithm is applied to logs to identify groups of students with similar characteristics based on their actions within the context of course data. The most widely used metrics were used to evaluate the proposed clustering model (Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index). A Silhouette score of 0.762 suggests that the clusters are partitioned more effectively, and a Davies-Bouldin score of 0.39 reveals less variance between the clusters. The research results revealed that the k-means clustering algorithm is suitable for identifying students with similar behavioural patterns.*

Keywords: Educational Data Mining, Clustering, Log files, k-means algorithm.

INTRODUCTION

The advent of the information and technology era has resulted in harvesting of vast amounts of data. The information gathered needs to be transformed into a useful form before any decisions can be made. There are reams of student records kept at any university or other institution of higher education. The diversity of educational data is increasing, but the extent to which this growth varies from one e-learning platform to the next is inconsistent. The task of effectively managing this vast amount of

varied educational data is not easy and brings a number of challenges. However, it is not possible to directly apply traditional data mining methods to educational data because of the volume, nature, and structure of the data. As a result, educational data mining (EDM) has just come into existence to help solve this issue. Additionally, in order to get better results from the EDM process, one of the most crucial things to do is to prepare the data for further analysis (Dutt et al., 2017).

Discovering student behavioural patterns and taking the appropriate actions to maximize the educational process is an important effort of contemporary education. According to (Aljohani et al., 2019), educational data from learning management systems (LMS) provides chances to analyse students' behaviour patterns and improve the effectiveness of teaching and learning behaviours. One example would be identifying numerous behavioural characteristics that have strong correlations with academic success, assessing the learning behaviours of students to enable educators to modify their lesson plans to get better results and providing early warnings to students who are at risk of not passing examinations and identify any unusual behaviours exhibited by students. According to (Hussain et al., 2019), the instructor's primary resource for monitoring a student's level of online engagement and ensuring that they receive a high-quality education is the data obtained from the virtual learning environment (VLE) in the form of logs. However, it might be challenging for teachers to keep track of and access each student's data on the online platform, making it impossible for them to evaluate the degree of student participation in their classes.

Applying educational data mining techniques to this data allows qualitative and quantitative analysis to be carried out. Performance evaluation plays an essential function in educational settings at a higher level. In other words, understanding the student's behaviour or learning experience within the course is an important aspect. Grades are the essential factor that is considered when determining a student's overall performance in college. However, a student's potential for employment is also affected by various other factors, including the completion of projects, participation in joint activities, and interaction with other stakeholders within the course. Therefore, students should not be categorized solely according to the grades they have been receiving, with their participation in extracurricular activities being ignored in the process. In order to acquire a full view of the student's performance and simultaneously ascertain information from their performance from time to time, it is important to know groups of students based on these factors (Palani et al., 2021).

Therefore, this paper presents research in which an unsupervised clustering algorithm is applied to logs to identify groups of students with similar characteristics based on their actions within the context of course data. The most important contribution of this study is that it assists educators in keeping track of their students' online activities and better understanding their profiles. This may help forecast the future consequences of student performance, which can then be used to adjust the teaching content. It also helps to optimize the learning environment within the learning management system.

1. RELATED WORKS

In the framework of EDM, researchers examine the practicability and viability of clustering algorithms from an application standpoint. Clustering algorithms have been recently developed as an effective method for extracting information from large amounts of data in various sectors. Because there are so many kinds of data to examine, several different clustering techniques have been created simultaneously to accommodate this diversity. These algorithms aim to group similar objects into the same cluster. These algorithms can be roughly classified into the following five categories: partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods. Each method of clustering has benefits and drawbacks that should be considered during the data clustering process. Therefore, it is natural that a single universal clustering method is still not available to manage all forms of data, including text, numbers, photos, and videos (Hervianti, 2018). Agnihotri et al. (2015) used data-driven clustering techniques K-means to determine which students in online courses had the highest and lowest levels of success. K-means clustering is a technique used to divide students into classes based on factors such as their login history and the number of times they have attempted to pass the course. Unfortunately, the data aggregations used in this study were handled incorrectly. During the training phase of the model, there was a large number of null values. Simultaneously, fewer factors were used to form the model during this phase. In order to investigate the study habits of the students, Preidys and Sakalauskas (2010) collected a large quantity of data from the BlackBoard Vista platform often used for online education. The dataset was partitioned into three groups using K-means clustering: Important, Unimportant, and Average importance. However, the final dataset also contained many outliers, which could negatively influence a final clustering model.

The challenges related to the correct clustering of educational data were researched further in (Navarro & Moreno-Ger, 2018). In this study, a large dataset with no outliers was used to evaluate which clustering technique could more effectively predict a low proficiency level of students in the LMS. This work used seven different clustering models. Moreover, various evaluation metrics such as the Dunn Index, the Silhouette score, and the Davies-Bouldin score were compared to benchmark their performance. This was done to determine which algorithm performs better. However, a notable disadvantage of this research was that missing data in instances in the factors were removed, which might include helpful information and provide more insights.

In contrast to these related works, the study presented in this paper directly investigates the LMS Moodle logs, where data is stored in xAPI format. The logs are stored in LMS Moodle database tables. The data represents the LMS stakeholders' interaction with different instructional materials and activities.

The dataset used in this study was created as an export from the database. It includes the following input variables: *component*, *action*, *target*, *objectid*, *contextid*, *contextlevel*, *contextinstanceid*, *userid*, *courseid* and *timecreated*.

As was mentioned earlier, this study aims to deploy the clustering model on the log and determine the groups of students with the same level of learning engagement exhibited by the students enrolled in the chosen course.
The K-means clustering algorithm is utilized for this purpose. The algorithm possesses two significant aspects. The first is selecting the ideal centre point and the second is detecting and eliminating disturbances. In addition, choosing the correct number of potential clusters *k* is essential when using the K-means clustering algorithm. This is because establishing an inappropriate *k* value may alter the data characteristics that are exhibited by the clusters.
Canopy, Elbow, Dunn Index, Silhouette Coefficient, and Gap Statistic are some of the various ways that can be used to understand the identified clusters better. Other options include Gap Statistic. The Elbow method was chosen in this study to determine the initial number *k* of clusters. The Elbow method proved to be effective in identifying the optimal number of clusters for the partitioning-based clustering algorithms (Moodle 4.0 Documentation, 2022).

2. METHODOLOGY

In this section, we will discuss the EDM process that has been selected to manage the knowledge discovery process. The proposed method can be broken down into four main phases as shown in Figure 1.

Figure 1. An outline of the EDM framework that has been adopted

Source: Own work.

Collecting relevant data is the first and most crucial step in every educational data mining approach. We started with collecting activity logs of the Moodle LMS. The raw dataset consisted of 1,529,342 log records produced by 4623 users participating in 894 courses and were accessible between February 1, 2019, and December 31, 2019.
The distribution of the logs between the courses was very diverse. For example, there were courses with few participants or minimal activity. Therefore, it did make sense to identify groups of students with similar behaviour on the dataset containing these outliers.
Therefore, we suggested to create a subset of the courses with higher activity during the knowledge discovery process pre-processing phase. The course selection was based on the number of activities, interactions, and the number of students who enrolled on the course.
As a result, an intensively used course was selected to extract information from the log data and show if the clustering could uncover some interesting insights into the course. The final dataset consisted of the students of one science course from 2019.

This course consisted of 82 users and 96 277 actions. Clustering analysis was performed on this dataset.
Figure 2 shows the list of features labelled as *component*, *action*, *target*, *objectid*, *contextid*, *contextlevel*, *contextinstanceid*, *userid*, *courseid* and *timecreated*, which were included in the input dataset.
The extracted data were pre-processed, and the detected outliers and missing values were removed.
Then, data were normalised, and relevant features of the data were selected.
Because we worked with the Moodle log file, it was vital to understand the individual attributes to pre-process the data correctly.
Mainly, it was necessary to take a closer look at the *edulevel* attribute since we attempted to classify students into individual clusters based on the same criteria. This feature provides information on the dates on which particular actions were carried out, mostly the educational significance of those acts. This feature can take on the values 0 through 2, specifically.
The records assigned the value 2 are all pertinent activities or events undertaken by a user. They are all connected to the user's learning experience.
According to the official documentation for Moodle logs, entries with a value of 1 represent teaching actions and have a teaching value. Therefore, entries with a value of 1 should not be mixed with participating events marked with a value of 2, as this is not recommended. The last group of entries were entries with a value of 0, representing any other action related to the administration of the website or any other users. Because these actions also did not have any educational value, they were not considered. As a result, entries with *edulevel* attribute values of 1 or 0 were not considered and were removed by forming a final pre-processed dataset.

	component	action	target	objectid	contextid	contextlevel	contextinstanceid	userid	courseid	timecreated
0	mod_page	viewed	course_module	7027	212936	70	95231	17	1489	1550343718
1	mod_resource	viewed	course_module	32534	213483	70	95760	10746	1489	1550349523
2	mod_resource	viewed	course_module	32532	213481	70	95758	10746	1489	1550349602
3	mod_resource	viewed	course_module	32534	213483	70	95760	10746	1489	1550349773
4	mod_resource	viewed	course_module	32280	212944	70	95239	10746	1489	1550352803

F i g u r e 2. Pre-processed final dataset

S o u r c e: Own work.

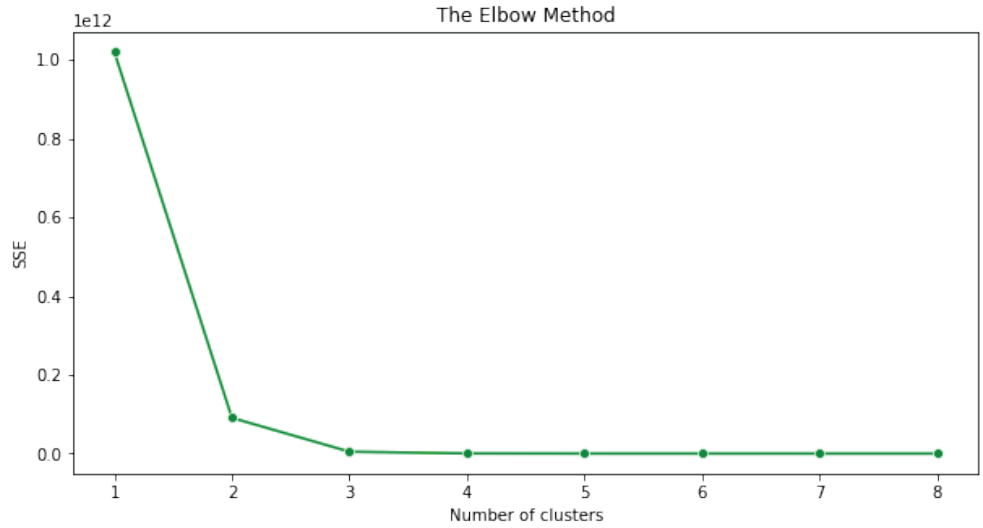
Also, according to the specification, the *ghost* role records can take on the values 0 or 1. It is advised that event records with a value of 1 should be ignored. Finally, after these pre-processing tasks, we were able to obtain a pre-processed final dataset that was suitable for the clustering algorithm. The revised dataset had ten attributes and comprised 44,285 records created by the activity of 77 different users.
The third step (Figure 1), in which we analysed data through data clustering, is one of the most critical steps in our proposed methodology. K-means algorithm was applied to the pre-processed data set for obtaining the clusters. Before clustering our education data, it was quite challenging to determine which clustering algorithm was the

most suitable because every algorithm has its unique set of characteristics that make it suitable for a specific application. Despite this, we investigated the hierarchical clustering method, which, compared to the partitioned clustering techniques, requires a more significant amount of CPU and memory resources. In addition, a hierarchical approach might quickly become expensive when used with extensive datasets. The primary areas where hierarchical and partitioned approaches are differentiated are computing time, prior assumptions, data sets, and clustering goals and results.
On the other hand, by applying certain mathematical techniques, the unsupervised clustering method is used to group the data from the vast dataset that are the most comparable to one another.
The k-means clustering technique is a method that is both straightforward and efficient when it comes to dividing up data. K-means clustering is an excellent method, although it does have a few drawbacks, the most notable of which are the inability to determine the number of clusters and the absence of an ideal centre. The total number of k values used in the k-means clustering technique is one factor that significantly impacts the data analysis findings.
In order to determine the correct value of k for the k-means clustering algorithm, the Elbow approach was applied. In its most basic form, the Elbow method works by computing the value of the sum of squared errors (SSE) in each cluster to determine the optimal number of clusters for a dataset. A low SSE value indicates that the cluster in question will be of high quality (Hussain et al., 2018).
Following its deployment, the k-means algorithm should be analysed for its effectiveness. The performance of a clustering method can be evaluated using a process known as cluster validation, consisting of internal and external evaluations. The findings and the ground truth are compared in order to carry out the external evaluation. For instance, the results of clustering are compared with public class labels in order to ascertain how well the cluster approach works with actual data groups on the dataset that is now available (Dobashi, 2017). Techniques used internally evaluate the degree to which one object in a cluster is similar to other objects in the cluster, as well as the degree to which one cluster differs from another.
Silhouette Coefficient (SC) is an internal validity measure that evaluates clustering performance based on the pairwise difference between and within cluster distance (Estacio & Raga, 2017). The findings fall somewhere between –1 and 1. The closer they are to 1, the more similar the entities are to one another. Vice versa, the further they are from 1, the less similar the entities are to each other (Kadoić & Oreški, 2018).

3. EXPERIMENTAL RESULTS

In this section, we discuss the various results obtained from the experiments. In the first step of this process, we clustered the data and then extracted the fundamental statistical characteristics from each cluster. Next, we used the Elbow method on the dataset to figure out the appropriate number of clusters to use, denoted by the letter k. To begin, we randomly selected a number of clusters and then calculated the SSE value for each cluster. As shown in Figure 3, the SSE value converges once the number of clusters equals 3. In accordance with the findings of the Elbow technique,

we were able to determine the suitable number ($k = 3$) of clusters to use for the primary clustering operation when applied to the dataset using the k-means algorithm.



F i g u r e 3. Optimal *k* value selection using the Elbow method

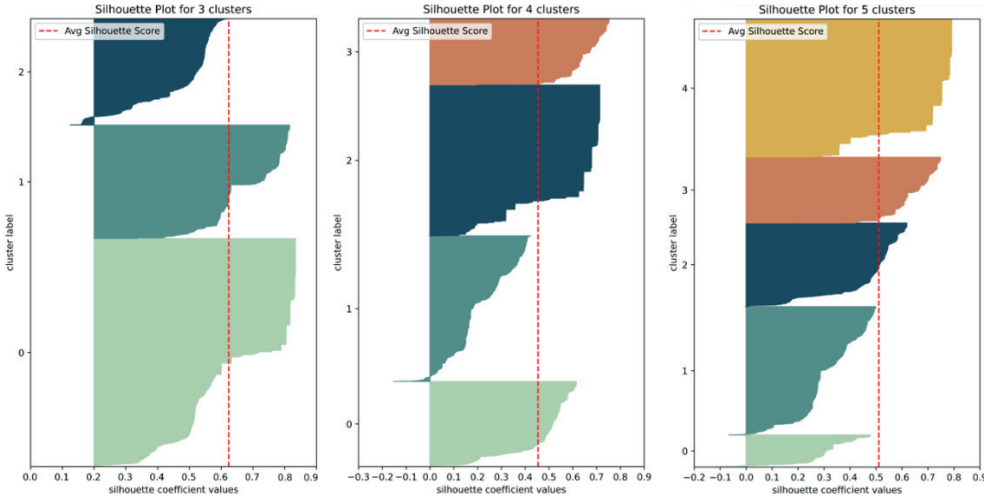
S o u r c e: Own work.

In order to investigate the distance between the generated clusters, silhouette analysis can be utilized. Because it measures how close each point in one cluster is to points in the neighbouring clusters, the silhouette plot offers a means to evaluate factors such as the number of clusters visually. This measurement has a range from –1 to 1. When the silhouette coefficients (another name for these values) are close to 1, the sample is located at a significant distance from the nearby clusters. Contrary, a score of 0 indicates that the sample is either on the decision border between two neighbouring clusters or extremely close to it. Negative values indicate that such samples may have been incorrectly assigned to the cluster they belong to.

In this case as shown in Figure 4, silhouette analysis is utilized to determine the best possible value for n clusters. The silhouette plot demonstrates that n clusters equal to 3 or 4 were relatively good choices for the proposed data due to the presence of clusters with above-average silhouette scores in the case for these clusters. Furthermore, we can see that there are relatively small fluctuations in the size of the silhouette plots in each case.

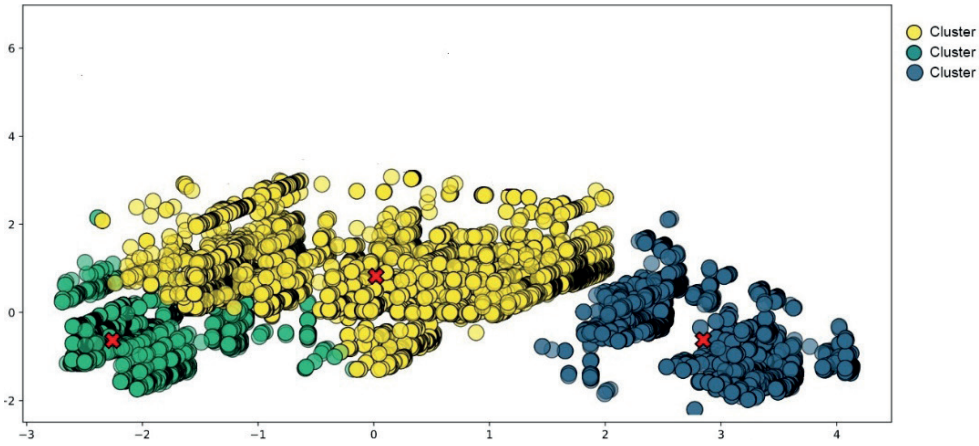
Furthermore, there are no spaces between the various groups of data, which is another indication that they are accurate. As can be seen, the silhouette analysis shows that the option with $n = 3$ would produce more accurate coefficient values. Therefore, it was chosen over $n = 4$.

The dataset was then subjected to random initial centre selection (RCS), followed by k-means clustering techniques. Figure 5 depicts the cluster data visualization created using RCS in conjunction with K-means. Finally, the data were organized into neat clusters by an algorithm.



F i g u r e 4. Silhouette plots for 3,4 and 5 clusters

S o u r c e: Own work.



F i g u r e 5. The result of K-means clustering with randomly selected initial centres

S o u r c e: Own work.

Following the application of the clustering technique, we discovered that 50.92 per cent, 25.35 per cent, and 23.73 per cent of the students, respectively, belonged to clusters 1, 2, and 3. In order to evaluate the k-means model performance as well as the hyperparameters tuning, we used the most common metrics such as Silhouette Coefficient, Davies-Bouldin Index and Calinski-Harabasz Index according to the Table 1. Because categorical variables are discrete and do not have any natural origin, the K-means algorithm cannot be applied to categorical data. As a result, we transformed them into numeric data by employing label encoding, which might have resulted in a decrease in the performance of the model.

Table 1. Results for evaluation of metrics on the K-means model

Silhouette Coefficient	Davies-Bouldin Index	Calinski-Harabasz Index
0.627	0.39	91715

Source: Own work.

However, the overall performance was not skewed, and according to the findings, the k-mean clustering model was accurately distributed. Table 1 shows that the silhouette coefficient score for the k-mean model was 0.627, indicating a high separation level as the value is close to 1. Therefore, samples are not placed in the incorrect clusters. In order to evaluate the scaled data, the Davies-Bouldin score was computed. The lower the value of the Davies-Bouldin score, the more clearly the clusters may be distinguished. The score, according to the k-means model, is 0.39.

Next, the Calinski-Harabasz score was applied to determine the extent of the variation in the data points between the clusters. If the value of the score is higher than expected, then the cluster is dense and has clear boundaries between its members. The Calinski-Harabasz score for k-means was 91,715, which is higher than each cluster distribution because there are 22,549 records in cluster 1, 11,227 in cluster 2, and 10,509 in the third cluster.

CONCLUSION

Identifying students in an online learning environment with similar behaviour, which can lead to a low level of involvement, is vital since this allows the instructor to influence or target the student's behaviour (Kabathova & Drlik, 2021).

Considering the study results, we can confirm the findings of other researchers that data mining in education can assist in extracting useful information from large datasets. Moreover, EDM can help to build a model of e-environment of the university, which would take into account the needs of today's students and today's market and would ensure a high level of university competitiveness (Morze et al., 2015). Cluster analysis is a method for performing data analysis that does not require prior information from the analyst and knowledge about the internal structure of the dataset. Understanding the data structure, the kind of analysis that is wanted, and the quantity of the dataset being analysed are all critical factors to consider when selecting the appropriate clustering technique (Munk et al., 2010).

Data used in this study were extracted from LMS Moodle. The next thing needed was to transform the data into a format that could be used as input into the clustering model. After that, the k-mean technique was applied to the data to identify students with similar characteristics. Finally, the most common metrics were used to evaluate the performance of the proposed model (Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index). As a result, three clusters have been identified. The findings demonstrated that the k-mean clustering algorithm is an appropriate algorithm model in this particular case. The Silhouette score of 0.627 indicates that

the clusters are partitioned more effectively, and a Davies-Bouldin score of 0.39 demonstrates less variance across the clusters.

In this study, all experimental findings were derived from the information gathered during a particular class. Therefore, the findings of experiments might differ from the data obtained in other classes. Therefore, the number of clusters produced by the k-means method is subject to change depending on the dataset.

As a result, the EDM framework that was proposed may produce satisfactory or unsatisfactory results when applied to additional data sets. Another limitation of this study is that the used model demonstrates a few spots in which clusters are overlapping. This fact could lead to failing to find students struggling academically with high performance.

For future work, we would like to apply association and pattern mining algorithms to each cluster to achieve a more comprehensive study of student performance, better understand their behaviour and compare results with the findings obtained in our previously proposed approach (Drlik et al., 2021).

ACKNOWLEDGEMENTS

This research was funded by the Science grant agency of the Ministry of Education of the Slovak Republic, grant number: 1/0490/22 "Design and Evaluation of Selected Learning Analytics Methods for Understanding and Improving the Learning Process and Virtual Learning Environment," European Commission under the ERASMUS+ Programme 2021, KA2, grant number: 2021-1-SK01-KA220-HED-000032095 "Future IT Professionals EDucation in Artificial Intelligence," and the Ministry of Education of Slovakia, grant number 004UKF-2-1/2021 "Preparation and development of teaching courses in English with a focus on artificial intelligence in the form of blended-learning".

REFERENCES

A g n i h o t r i, L. et al. (2015). 'Mining Login Data For Actionable Student Insight', Proceedings of the 8th International Conference on Educational Data Mining (EDM) (pp. 472–475). ISBN 978-84-606-9425-0.

A l j o h a n i, N.R., F a y o u m i, A., & H a s s a n, S.U. (2019). Predicting at-risk students using2 clickstream data in the virtual learning environment. *Sustainability*, *11*(24), 7238. https://doi.org/10.3390/su11247238.

D o b a s h i, K. (2017). Automatic data integration from Moodle course logs to pivot tables for time series cross section analysis. *Procedia computer science*, *112*, 1835–1844. https://doi.org/10.1016/j.procs.2017.08.222.

D r l i k, M., M u n k, M., & S k a l k a, J. (2021). Identification of Changes in VLE Stakeholders' Behavior Over Time Using Frequent Patterns Mining. *IEEE Access,* 9, 23795–23813. https://doi.org/10.1109/ACCESS.2021.3056191.

D u t t, A., I s m a i l, M.A., & H e r a w a n, T. (2017). A systematic review on educational data mining. *Ieee Access*, *5*, 15991–16005. https://doi.org/10.1109/ACCESS.2017.2654247.

Estacio, R.R. & Raga Jr, R.C. (2017). Analyzing students online learning behavior in blended courses using Moodle. *Asian Association of Open Universities Journal*. ISSN 2414-6994.

Hassan, S.U., Waheed, H., Aljohani, N.R., Ali, M., Ventura, S., & Herrera, F. (2019). Virtual learning environment to predict withdrawal by leveraging deep learning. *International Journal of Intelligent Systems*, *34*(8), 1935–1952. https://doi.org/10.1002/int.22129.

Hervianti, S.P. (2018). Analisis Pengelompokan Penyebaran Lulusan Mahasiswa Universitas Gunadarma Menggunakan Metode Clustering. *ikraith-informatika*, *2*(2), 42–48. e-ISSN 2654-8054.

Hussain, S., Atallah, R., Kamsin, A., & Hazarika, J. (2018, April). Classification, clustering and association rule mining in educational datasets using data mining tools: A case study. In *Computer Science On-line Conference* (pp. 196–211). Cham: Springer. https://doi.org/10.1007/978-3-319-91192-2_21.

Kabathova, J. & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. *Applied Sciences*, *11*(7), 3130. https://doi.org/10.3390/app11073130.

Kadoić, N. & Oreški, D. (2018, May). Analysis of student behavior and success based on logs in Moodle. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 0654–0659). IEEE. https://doi.org/10.23919/MIPRO.2018.8400123.

Moodle 4.0 Documentation. Retrieved from https://docs.moodle.org/dev/Events_API.

Morze, N., Smyrnova-Trybulska, E., & Umryk, M. (2015). Designing an e-university environment based on the needs of net-generation students. *International Journal of Continuing Engineering Education and Life-Long Learning, 25*(4), 466–486. https://doi.org/10.1504/IJCEELL.2015.074230.

Munk, M., Kapusta, J., & Švec, P. (2010). Data pre-processing evaluation for web log mining: Reconstruction of activities of a web visitor. *Procedia Computer Science, 1*, 2273–2280. https://doi.org/10.1016/j.procs.2010.04.255.

Navarro, Á.M. & Moreno-Ger, P. (2018) 'Comparison of Clustering Algorithms for Learning Analytics with Educational Datasets', International Journal of Interactive Multimedia and Artificial Intelligence, *5*(2), p. 9. https://doi.org/10.9781/ijimai.2018.02.003.

Palani, K., Stynes, P., & Pathak, P. (2021). Clustering Techniques to Identify Low-engagement Student Levels. In *CSEDU (2)* (pp. 248–257). https://doi.org/10.5220/0010456802480257.

Preidys, S. & Sakalauskas, L. (2010). 'Analysis of students' study activities in virtual learning environments using data mining methods'. *Technological and Economic Development of Economy, 16*(1), 94–108. https://doi.org/10.3846/tede.2010.06.